Max Planck Institute for
Marine Microbiology

School of Engineering and Science

Exploring the Marine Virosphere

From Genome Context to Content

by

Melissa Beth Duhaime, M.Sc.

A thesis submitted in partial fulfillment
of requirements for the degree of

Doctor of Philosophy

Approved Thesis Committee

Prof. Dr. Frank Oliver Glöckner *(chair)*
Max Planck Institute for Marine Microbiology and Jacobs University

Dr. Antje Wichels
Alfred Wegener Institute for Marine and Polar Research

Prof. Dr. Matthias Ullrich
Jacobs University

Statement of Sources

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from published or unpublished scientific work has been cited in the text and listed in the references.

Signature **Date**

Thesis Abstract

Over the last decade, we have witnessed the dawn of marine phage genomics. Since the first genome was sequenced just over ten years ago, nearly thirty additional marine phage genomes and eleven marine virus metagenomes have offered a small glimpse into the genomic underpinnings of phage in the oceans.

Genomics has revealed their role in host metabolism, as phage genomes harbor environmentally relevant functional genes, such as those involved in photosynthesis, nutrient stress response, nucleotide scavenging, and vitamin biosynthesis, some of which are expressed during infection. Furthermore, due to evidence of phage activity in host genomic islands, phages are now thought to be important drivers of microbial niche adaptation and diversification in the oceans.

In recent years, advances in sequencing technologies and falling costs have led to sequence generation at an unprecedented rate. However, with this input, our ability to analyze and interpret new sequence data in attempt to garner biological insights (as those described above) is under serious threat. The role of bioinformatics in the current age is to keep this data accessible, or we risk serious loss in its value. For this reason the development of contextual data standards becomes crucial. The power of *context* to a biologist touches upon the core tenets of comparative genomics by expanding the dimensions among which comparisons and inferences can be made.

This thesis addresses bioinformatic themes of contextual data development and implementation as a means to enhance marine phage genomics. This puts into practice the ultimate role of bioinformatics: to facilitate the capture and collection of various data sources, enabling *in silico* biological predictions that lead to acute, solvable laboratory experiments.

The primary accomplishments of this work include: *a*) contributing to the development of the megx.net portal for marine genomics, supplementing the database with the addition of marine phages, *b*) generation of the first set of manually curated MIGS-compliant contextual data reports describing the marine phage genome collection, and first pass exploratory statistics, which lead to hypotheses that are experimentally addressed in a later chapter, *c*) the first study to make use of megx.net interpolated environmental data paired with viral genomic data to make ecological inferences ("ecogenomics") about a marine phage from Helgoland, *d*) a novel application of genome signatures (tetranucleotide frequencies and codon adaptation), as an approach to infer biological knowledge from phage genomes, and *e*) investigation of the role of environmental conditions (nutrient starvation) on the infection dynamics of a marine phage host system.

Keywords: marine, phage, genomics, ecogenomics, evolution, contextual data, genome standards

Contents

1 An Introduction to Exploring the Marine Virosphere 1
1.1 Elements of the Virosphere: Phage Evolution, Diversity, and Ecology 1
1.1.1 Phage Genome Evolution: The mosaic mobilome 2
1.1.2 Marine Phage Diversity: In search of patterns in the particles 5
1.1.3 Influence of phages on host microbial communities at sea: Pirates or Partners? 10
1.1.4 Phage Ecology and 'Ecogenomics' 17
1.2 Theory Behind the Bioinformatic and Genomic Approach 25
1.2.1 Bioinformatics as it pertains to us 25
1.3 Motivation and Research Aims 30
1.4 Content Overview 32

2 Megx.net: integrated database resource for marine ecological genomics 35
2.1 Introduction 36
2.2 New Database Structure and Content 37
2.2.1 Contextual and Sequence Data Content 37
2.2.2 Environmental Data Content 39
2.3 User Access 40
2.3.1 Genes Mapserver 40
2.3.2 Geographic-BLAST 40
2.3.3 Software Extensions to the Portal 40
2.3.4 User Test Case 42
2.3.5 Web Services 42
2.4 Other Changes 42
2.5 Summary 43

3 Enriching public descriptions of marine phages using the MIGS standard: A case study assessing the contextual data frontier 45
3.1 Introduction 46
3.2 Methods 48
3.2.1 Genomes and contextual data sources: setting up the comparison 48
3.2.2 MIGS-compliance 48
3.2.3 GCDML files 52
3.2.4 Exploratory contextual data analyses 52
3.3 Results 53
3.3.1 A comparison of INSDC and manually curated GCDML reports 53
3.3.2 Exploratory Data Analysis and Interpolated Metadata 56
3.4 Discussion 58

3.4.1 Towards consistency and persistence of contextual data 62
3.4.2 Biological lessons learned 63
3.4.3 GDCML 66
3.4.4 Outlook 66

4 Ecogenomics and Genome Landscapes of Marine Pseudoalteromonas Phage H105/1 69

4.1 Introduction 70
4.2 Materials and Methods 72
4.2.1 Phage Harvesting, DNA Isolation and Sequencing 72
4.2.2 Virion Structural Proteome Analysis 73
4.2.3 Genome Annotation 73
4.2.4 Ecogenomic Analysis 74
4.2.5 Host Phylogenetic Analysis and Tree Reconstruction 75
4.2.6 Genome Signature Analysis 75
4.3 Results and Discussion 76
4.3.1 Pseudoalteromonas Phage H105/1 Biology 76
4.3.2 Genome Features and Annotations 77
4.3.3 Ecogenomics: H105/1 in GOS and five marine virus metagenomes 86
4.3.4 Painting a genome landscape: codon adaptation and tetranucleotide usage 88
4.4 Summary: Phage diversity and evolution in light of marine Phage H105/1 91

5 Insights into infection dynamics of ocean cyanophage P-SS2: to integrate, or not to integrate? 93

5.1 Introduction 93
5.2 Methods 96
5.2.1 Host growth 96
5.2.2 MIT9313 infection with P-SS2 96
5.2.3 Quantitative PCR (QPCR): qINT/qEXT 99
5.2.4 Quantifying host nutrient stress through gene expression . . . 100
5.3 Results and troubleshooting 100
5.3.1 MIT9313 Growth, Infection and P-SS2 Integration 100
5.3.2 Quantifying host and phage through single-copy gene detection 101
5.3.3 Manipulating host MIT9313 stress response 102
5.4 Outlook and Future Work 103
5.4.1 Integration 103
5.4.2 Towards understanding the influence of nutrients on infection and integration 104
5.4.3 The influence of AMG expression in phage-host systems . . . 105
5.4.4 Underlying phage biology and other considerations 106

6 Summary and Outlook 109
6.1 A platform for marine virus genomics: megx.net and standards development . . . 110
6.1.1 Future virus extensions to megx.net . . . 110
6.1.2 Maintaining robust marine virus contextual data . . . 112
6.2 Transforming sequence data into knowledge through ecogenomics . 112
6.2.1 Contextual insights into genome content: Phage H105/1 . . . 113
6.2.2 Linking *bio*– with –*informatics*: The case of cyanophage P-SS2 116
6.3 The road ahead . . . 116
6.3.1 Ecogenomic principles to tackle "regionality" versus "ubiquity"117
6.3.2 Towards understanding microbial ecosystems through the viral component . . . 118

7 Additional Analyses and Supplementary Material 121
7.1 Supplement to Chapter 1: Introduction and The Marine Phage (Meta)Genome Collection . . . 121
7.2 Supplement to Chapter 3: Phage Contextual Data . . . 126
7.3 Supplement to Chapter 4: Phage H105/1 Genome . . . 128
7.4 Supplement to Chapter 5: Infection and Integration Dynamics of P-SS2 . . . 138

Bibliography 143

List of Figures

1.1 Conserved 'Lambdoid' Genome Order 4
1.2 Method to model viral community structure 7
1.3 Virome community structure estimates 8
1.4 Marine Food Web 11
1.5 "Kill the Winner" Model 13
1.6 Hypervariable genomic islands in microbial genomes 16
1.7 Phage Replication Strategies 19
1.8 Phage "one-step" growth curve 20
1.9 GenBank Growth, 1982-2009 26
1.10 The Bioinformatic Funnel 29
1.11 Addressing phage biology through bioinformatic development and genomic discoveries 31

2.1 Megx.net conceptual schema and general architectual 38
2.2 Megx.net platform use case 41

3.1 Model of flow of contextual data into biological knowledge 49
3.2 Summary of MIGS compliance among marine phage genomes 54
3.3 Statistics describing marine phage isolation/sequencing year and genome properties inferred from GCDML reports 57
3.4 Overview of phage taxonomic data 59
3.5 Network of 'genome pairs' and interactions between sequenced marine phages and sequenced hosts 60
3.6 Marine phage genomes "one the map" 61

4.1 Phylogenetic characterization of host *Pseudoalteromonas* sp. H105 16S rRNA gene 71
4.2 Phage H105/1 genome, synteny with other siphoviruses, and phylogenetic and marine signatures 77
4.3 Neighbor joining tree of ORF 8, single-stranded binding protein . . . 84
4.4 Pseudoalteromonas phage H105/1 proteome 85
4.5 Presence of Pseudoalteromonas phage H105/1 proteins in GOS microbial metagenomes and marine viromes 86
4.6 Pseudoalteromonas H105/1 Genome Signatures 91

5.1 Model of lysogenic/lytic infection and assay to detect integration . . 97
5.2 MIT9313 infection with P-SS2 101
5.3 PCR-based detection of P-SS2 integration 102
5.4 Prochlorococcus marinus MIT9313 response to nutrient limitation . . 103
5.5 MIT9313 RNA detection 104

6.1 Helgoland phage-host network. 115

7.1 Marine prophage host taxonomy and length descriptions 123
7.2 Overview sequenced viral metagenomes 126
7.3 Screenshot GCDML Report revealing the GCDML schema 127
7.4 Influence of varying e-value and % query identity thresholds on raw hit counts of GOS and virome 135
7.5 Effect of genome window size on minimum tetra Pearson correlation coefficient cut-off 136
7.6 TEM of Pseudoalteromonas phage H105/1 virion and plaque morphology 137
7.7 P-SS2 isolation location mapped onto environmental parameters and host MIT9313 abundance 139

List of Tables

1.1 MazG in marine phage genomes . . . 23

3.1 Sequenced marine phages and the extent of their georeferencing . . . 50

4.1 Pseudoalteromonas phage H105/1 ORF and annotation summary based on homology to NCBI-nr, Pfam, and MeGo databases . . . 78

5.1 MIT9313 and P-SS2 att sites . . . 96
5.2 Primers designed . . . 98

6.1 Prophage statistics . . . 111

7.1 Sequenced Viromes . . . 124
7.2 Web Map Service query to generate megx.net map . . . 129
7.3 GOS sites queried by Pseudoalteromonas Phage H105/1 proteins, and their respective metadata . . . 130
7.4 Marine viromes and their megx.net interpolated metadata . . . 133
7.5 Data accompanying Phage H105/1 proteome . . . 138

CHAPTER 1

An Introduction to Exploring the Marine Virosphere

On the tip of your very finger, there are viruses raging a war against the microbes that call your skin home. There is likely no living ecosystem on Earth lacking of viruses, or 'phages', as they are called when they infect bacteria. They are the most abundant biological entity on the planet (Suttle, 2005). We are only beginning to appreciate their "megascale" global influence as they flourish in our planet's "*nanoscale microcosm where visible light washes over phages the way ocean swells move through plankton*" (Steven McQuinn).

This introductory chapter represents both the biological basis (section 1.1) and the bioinformatic approaches (section 1.2) relevant to the research themes of this thesis, as they are ultimately applied to interpret phage biology from the perspective of their genomes (Chapters 2-5). It is intended to guide a reader not well versed in phage biology, evolution, and ecology towards an understanding that will fully appreciate the motivation of the research aims and the studies conducted.

1.1 Elements of the Virosphere: Phage Evolution, Diversity, and Ecology

The global phage population is immense: at an estimated 10^{31} tailed phages on the planet, if assembled head to tail, the population would extend over 100 million light years into space (Suttle, 2005), further than the next 60 galaxies! In sterile, saline buffer, marine phage particles are quite stable, and can be stored for years to decades at 4° C (Angly et al., 2006; Sullivan et al., 2003). However, in the wild, they decay quite rapidly (Suttle and Chen, 1992), due to a variety of abiotic (e.g., UV radiation, interaction with labile particles) and biotic factors (e.g., phytoplankton grazing has been implicated) (Suttle and Chen, 1992). Due to this natural instability, it can be assumed that wild phage populations turn over rapidly to maintain their observed abundances. In the marine system, it is estimated that viruses kill

10-40% of all bacteria every day (Suttle, 2005). This would require 10^{24} successful infections per second (Hendrix, 2008), with virus-mediated gene transfer occurring up to 10^{15} times per second (Bushman, 2002)! This section explores our enormous global 'virosphere' by considering the consequences of these quantitative impressions on the evolution, diversity, and ecology of marine phages.

1.1.1 Phage Genome Evolution: The mosaic mobilome

"*[The size and turn over rate of the global virome] are vastly greater than anyone had imagined before the measurements were made, and they have a wonderfully liberating effect on how we can think about the mechanisms of phage evolution.*" Roger Hendrix (Hendrix, 2008)

Before venturing into a discussion of phage diversity (Section 1.1.2) and forays into how the environment may influence the distribution of phages and their genes, concepts to "quantify" (or at least track) the *units* of phage diversity[1] and mechanisms of this diversification will be discussed.

As phage genomes have been sequenced and comparative genomics made possible, patterns have begun to emerge that shed light on the evolutionary forces behind phage diversification. One of the most striking (and at times troublesome) observations is the extreme diversity contained in a virus genome. There is such rampant mosaicism in their genomes due to horizontal transfer, it is not possible to establish a linear ancestry connecting phages to their predecessors. Once we understand and appreciate this feature, "the complexity introduced into the phages' history by horizontal exchange *enriches rather than confuses*" our concept of how phages evolve (Hendrix, 2008).

Towards a functional phage taxonomy. Lacking a conserved marker gene universal among all phage genomes (Rohwer and Edwards, 2002) and with more reticulate than linear decent (Lima-Mendez et al., 2008b), viruses can not be fit into the traditional species concept definition[2]. On a molecular/sequence level, how does one group something that shares nothing? Yet, in order to understand how mechanisms of evolution differ through space and time, it is essential to resolve some concept of phage relatedness. Phage taxonomy (including that installed at NCBI) has classically depended on the definitions outlined by the International Committee on the Taxonomy of Viruses (ICTV) (Büchen-Osmond, 1996), which

[1] For microbial diversity, this is often a 'species' or operational taxonomic unit (OTU).

[2] It can also (plausibly) be argued that the same problem exists for microbes (Doolittle and Zhaxybayeva, 2009).

groups phages based on morphological and behavioral phenotypes (i.e., capsid size, shape, structure and resistance to organic solvents, host range, genome size and molecule type). This approach lacks a direct connection to the phage genome sequence, and it is useless when attempting to classify phages for which such phenotypic traits are unknown. To examine the sequence-based structure among 105 sequenced phage genomes, Rohwer and Edwards (2002) calculated protein distances between whole genome amino acid sequences to build a "Phage Proteomic Tree" (PPT). This tree showed remarkable (though not absolute) resemblance to the ICTV-based taxonomy and provides a mechanism to track phage "groups" based on gene sequences (Rohwer and Edwards, 2002). This has proven particularly useful in the classification of viral metagenomes and comparisons between virus metagenomes (Angly et al., 2006; Bench et al., 2007; Breitbart et al., 2004; Desnues et al., 2008), as discussed in Section 1.1.2.

However, considering their promiscuous nature, the extent to which we are creating artificial delineations between phages should be considered. "*Are there discrete phage types or is there a continuum in which different types grade into one another?*" (Casjens, 2005). Each sequenced phage genome will bring us closer to truly understanding the "phage diversity landscape."[3].

The most influential "creative force" in phage evolution is recombination (Hendrix, 2008), the effects of which are some of the only constants that have emerged from comparative phage genomics (Casjens, 2005; Pedulla et al., 2003).

Inter-module recombination and 'Modular Evolution' Perhaps the most evident footprint left on a phage genome is that of *non-homologous* recombination.This mechanism creates easily identified "novel joints", which describes any point where two previously unassociated groups of genes are now joined.[4] We can observe non-homologous recombination when two genomes are compared, as they could be conserved gene by gene until a certain point, when sequence identity immediately and drastically diminishes (Casjens, 2005). Intriguingly, the location of these joints is not random. They occur predominately at gene boundaries (rather than within genes), and often at boundaries of functional gene modules (groups of neighboring genes involved in related tasks). This observation led to the devel-

[3] It is recognized that more sequencing is not the absolute the solution to understanding nature. However, it is justified in the case of sequence-based questions of evolution and diversity. We will never have the complete picture, but deep sequencing gets us closer. Some of the most remarkable studies of microbial evolution have come from sequencing low diversity systems where the "complete picture" was nearly attained (Tyson et al., 2004).

[4] As with most awe-inspiring feats of our natural world, the ability to observe this seemingly simple process so acutely may be the reason for its beauty: K.I.S.S. - *Keep It Simple 'Sweety'*

opment of the 'Modular Theory of Evolution' for bacteriophages[5] (Botstein, 1980), whereby functional modules are frequently swapped between phages infecting diverse hosts (Filée et al., 2006; Lucchini et al., 1999). Yet, in most phages, even those completely lacking sequence homology, the order of the modules is conserved to a remarkable degree (Figure 1.1), especially in 'lambdoid' phages (phages which closely resemble the organization and replication of *E. coli* phage λ[6]. Such organization ensures that functional modules *a*) "travel" with related genes that work together towards a specific phage task or structure, e.g., 'head formation genes', and *b*) are inserted at the essential genome position. In this way, fully functional 'experimental' phages are formed. Yet, is this the origin and selective pressure for their conserved order? 'Historical accidents' have also been implicated (Casjens, 2005), but are highly unlikely to generate or maintain such order (perhaps to destroy it). The most plausible explanation is that the conserved order reflects the molecular life cycle of most tailed phages, such that genes are transcribed as they are temporally needed: genes involved in recombination and DNA replication (traditionally described as "early genes") are needed before those required for head and then tail formation ("late genes") (Ptashne, 2004).

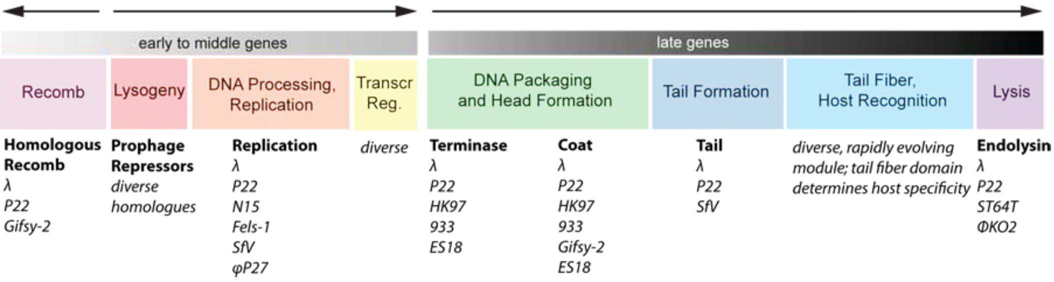


Figure 1.1: Phage functional modules conserved in 'lambdoid' phages and many integrated prophages. Phages with identifiable homologues are listed under each protein or function. Note the logical congruence of time of transcription and appearance of each gene product as it required for phage assembly, i.e., replication happens before phage capsid and tail production: "replicate the genome before you are prepared to package it." Image modified from concepts of (Canchaya et al., 2003; Casjens, 2005).

Inter-genic recombination and functional plasticity. Non-homologous recombination also leads to novel joints at domain boundaries *within* genes encoding multi-domain proteins. For instance, one of the longest genes in a tailed phage

[5] Modular evolution and the conserved order ('synteny') of these modules is further discussed and complemented by experimental evidence in Chapter 4, with respect to marine Phage H105/1 from Helgoland (Figure 4.2a)

[6] λ is the most thoroughly studied phage to date, and the basis for molecular biology advancements over the last half century.

is that of its multi-domain tail fiber gene, which contains the domain responsible for recognition of host cell receptors (Haggard-Ljungquist et al., 1992). As this is the primary mechanism by which hosts attempt to avoid phage predation, while phage evolution leads to renewed recognition (Comeau and Krisch, 2005), this domain is one of the most rapidly evolving components of the coevolutionary phage-host "arms race." In fact, by exchanging solely the tail fiber of T2 phage, it is shifted from infecting *E. coli* K12 to infecting *E. coli* O157:H7 (Yoichi et al., 2005). The tail fiber is also mechanistically diverse: some phages recognize multiple hosts with a single tail fiber protein, while some rely on an invertible DNA switch to alternate between multiple sets of tail fiber genes to alter its host range (Haggard-Ljungquist et al., 1992)[7]. This lack of conservation is further observed in the marine virus metagenomes. In one such study from Tampa Bay, nearly the entire 252 kb of the Cyanophage P-SSM2 genome was covered by metagenome reads, except four regions, two of which are tail fiber proteins (McDaniel et al., 2008).

In essence, every new phage genome is an experiment, and most are sub-optimal and imperfect. But such modular swapping of functional components accelerates and facilitates successful combinations of proteins and domains, and has direct implications for phage ecology (i.e., host range and how it "behaves", see section 1.1.4.1).

1.1.2 Marine Phage Diversity: In search of patterns in the particles

As "*[it] is impossible to understand evolutionary processes without having parallel knowledge of diversity*" (Casjens, 2003), the following section will not only complement the picture of phage evolution, but pave the way towards understanding what forces (be they environmental, stochastic, etc.) are driving the diversity and distribution of phages and their genes.

Phage diversity can be examined at the level of:

1. infection phenotype: host range, plaque morphology, differences in latent period and burst sizes (as discussed in section 1.1.4.1, Figure 1.8), etc.

2. morphology: virion shape, nucleic acid properties (i.e., ssDNA, dsDNA, ssRNA, and dsRNA)

3. sequence diversity: as evidenced by genome and metagenome sequencing

All three fields of phage diversity merit attention and are actively pursued. However, tackling phage diversity from the context of their sequences is the most

[7] Incredible.

high-throughput and exhaustive approach. Further, it is not limited by our lack of host knowledge, nor by our inability to culture a vast majority of the possible microbial hosts. Thus, phage sequence diversity, and its implications about the global phage protein pool, will be the focus of the following section. Morphology and host range are addressed in Chapter 3, in the context of the diversity captured by our current marine phage genome collection.

Sequence and modeled community diversity In the absence of a tractable marker gene, a virus "species inventory," as is typical of microbial diversity studies, is not possible. Nonetheless, using sequence data from viral metagenomes (viromes), we can estimate the number of viruses in an environment based on "contig spectra" modeling of the sequence reads (Angly et al., 2005). Using stringent[8] assembly parameters, this approach considers the relative occurrence of individual sequence reads (the *in silico* "contig spectrum", Figure 1.2) (Angly et al., 2005). The *in silico* contig spectrum based on read data is compared to a predicted contig sprectrum calculated using a modified Lander-Waterman algorithm based on values describing: genome length[9], number of reads, average read length, and assembly parameters. The read-based *in silico* spectrum and predicted spectrum are compared and the parameters altered until the best fit model is found to describe the community structure (Figure 1.2, Angly et al., 2005).

Based on the indices modeled by the contig spectra approach, patterns in virus community structure begin to emerge. For instance, based on metagenomic data of viruses found in one-week old human infant feces, the diversity was estimated to be far less than the average virome, including that of *adult* feces (Shannon Index of 1.69 nats, as compared to 8 nats; Appendix Table 7.1). The single most abundant genotype of this "extreme" virome represented 43.6% of the entire infant feces community, though the most abundant genotypes in other samples comprised, on average, close to 4% of the virus community (Figure 1.3). The explanation offered for this finding is that infant intestines are sterile at birth, and as microbiota colonize (via, i.e., mother's milk, the birth canal, the environment), the viruses will simultaneously follow suit (Breitbart et al., 2008).

The first modeled estimates reported over 7000 viral genotypes in a 200 liter seawater sample (Breitbart et al., 2002) and 10^4 genotypes per kilogram of marine sediment (Breitbart et al., 2004), with similar estimates observed since (Figure 1.3, see Appendix Table 7.1) (Angly et al., 2006; Bench et al., 2007). To maintain a Shan-

[8] Angly et al. (2005) based their assembly on 98% similarity of at least 20 bp of (on average) 670 bp reads. The stringency can then be tested by assembling genomes of closely related phages.

[9] Average or dominant virus genome length can be determined from any sample experimentally using genome finger-printing methods, such as pulsed-field electrophoresis.

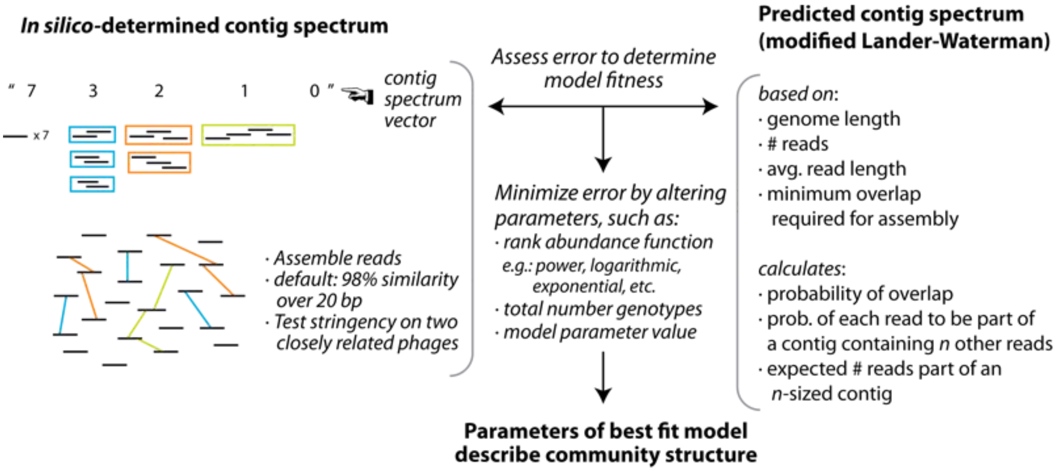

In silico-determined contig spectrum
" 7 3 2 1 0 "
contig spectrum vector
x 7
· Assemble reads
· default: 98% similarity over 20 bp
· Test stringency on two closely related phages
Assess error to determine model fitness
Minimize error by altering parameters, such as:
· rank abundance function
e.g.: power, logarithmic, exponential, etc.
· total number genotypes
· model parameter value
Parameters of best fit model describe community structure
Predicted contig spectrum (modified Lander-Waterman)
based on:
· genome length
· # reads
· avg. read length
· minimum overlap required for assembly
calculates:
· probability of overlap
· prob. of each read to be part of a contig containing *n* other reads
· expected # reads part of an *n*-sized contig

Figure 1.2: Overview of method used to model virus community structure based on metagenome data, as developed by Angly et al. (2005). *In silico*-determined (left) and predicted (right) contig spectra are compared and parameters altered until the best fit model is found to describe the community structure.

non Index greater than 9 nats (as calculated by Breitbart et al. (2004)), there must be over 8000 'species' (or phage genotypes) in one kilogram of marine sediment, which is impressive considering the 4200 amphibians and 6300 reptile species predicted to inhabit the *entire planet* (Breitbart et al., 2004). On a global scale, there are thought to be hundreds of thousands of viral genotypes in our oceans (Angly et al., 2006). As a confounding factor, our current sequence databases do not grasp the diversity observed in nature. As such, 60-80% of the open reading frames (ORFs) of sequenced phage genomes lack homologues (Paul and Sullivan, 2005), and 65-95% of viromes contain novel sequences (Appendix Table 7.1).

Based on this estimated viral sequence diversity, combined with the overall lack of conservation to proteins in our sequence databases, we can infer that virus diversity is exceedingly high.

Method to the Madness? Within this great diversity, patterns do emerge. Relying on the structure of the sequence-based Phage Proteomic Tree (PPT), proposed by Rohwer and Edwards (see section 1.1.1), metagenome reads can be mapped to their most similar protein homologues and classified according to the phage group in which they are found. Applying this approach, there is a "regionalization" in the representation of phage groups in the ocean (Angly et al., 2006; Bench et al., 2007; McDaniel et al., 2008). In a comparative virome study across four marine regions, according to PPT assignment of the reads, proteins of 84 phage species were found specific to one region (Angly et al., 2006). Angly et al. also found, for

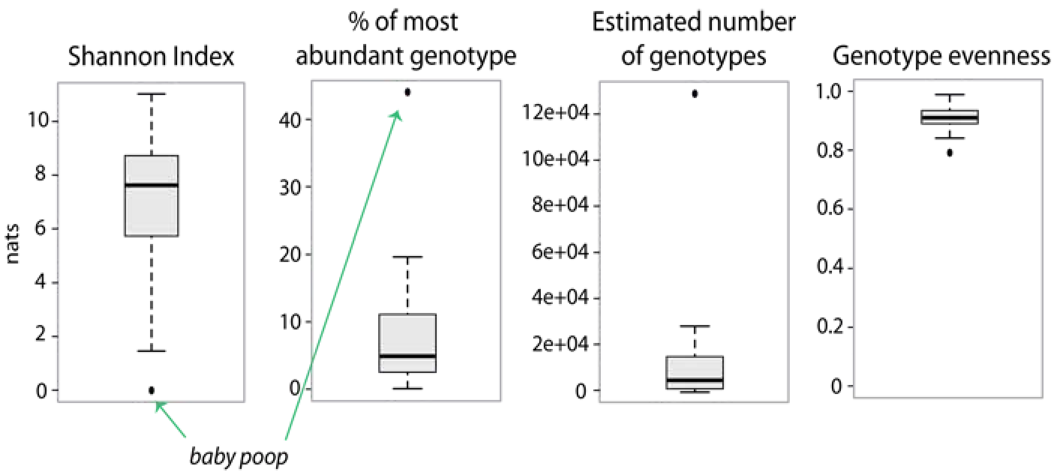


Figure 1.3: Overview of virome community structure estimates. Structure is modeled using the "contig spectra" approach, as described in the text. The outlier virome labeled is that derived from one-week old infant feces, a community known to have extremely low microbial diversity. For data describing, all viromes as they were published by authors of each study, see Appendix Table 7.1.

instance, prophage[10] sequences to be more prevalent in Arctic viromes, whereas cyanophages were more prevalent in the Sargasso Sea.

In addition to such location-specific phage protein representation, there is an overall "marine-ness" among all viromes of marine origin. While there is no significant difference in the representation of phage groups between marine water column and sediment viromes, there are significant differences between the shared marine profile when compared to phage genomes isolated from all habitats (Breitbart et al., 2004). These observations imply a *marine phylogenetic quality* (Breitbart et al., 2004), further supported by equivalent observations in the four marine viromes (Angly et al., 2006) and an ancient stromatolite virome (with an associated "marine" microbial community) (Desnues et al., 2008). This "marine" trend was also observed in a sequenced Pseudoalteromonas phage genome isolated from the North Sea (Figure 4.2c, Duhaime et al., 2010b), the focus of Chapter 4.

Does this seemingly biogeographical pattern (on both regional and global habitat scales) suggest a restricted dispersal of marine phages or their proteins? Limited by the extent to which we can link a viral sequence in a virome to a specific host, this trend likely suggests what can be expected: the most abundant phages in a given environment reflect the most abundant microbes. For instance, within the oligotrophic surface Sargasso Sea virome, *Prochlorococcus marinus* cyanophages

[10] Prophages refer to phages integrated in their host chromosome. Thus, "prophage sequences" are similar to prophages previously identified in sequenced microbial genomes (Angly et al., 2006)

significantly dominated the dataset (Angly et al., 2006). This likely reflects the high abundance of their *Prochlorococcus* hosts in the Sargasso Sea (Venter et al., 2004). This pattern extends beyond the marine environment. Phage groups typical of the ocean, e.g., T7-like podophage, λ-like siphophage and T4-like myophage, are known to infect Gram-negative hosts (prevalent in the oceans); whereas phage groups abundantly found in fecal matter, e.g., SFI21-like and TP901-siphophage, are known to infect Gram-positive bacteria, diagnostic of enteric habitats (Breitbart and Rohwer, 2005).

It is likely that our inability to make such connections elsewhere is due to a lack of whole genome sequences of phages infecting other dominant marine organisms. To this avail, the continued pursuit of phage whole genome sequencing will increase the value of the virome data and offer profuse insights into the diversity (and mechanisms influencing the diversity) of environmental phages.

Most viral proteins are still ubiquitous. Despite the detectable degree of ocean "regionalization" (best explained by phage populations mirroring their hosts), the majority of the viral sequences are ubiquitous across marine biomes. Forty-five phage species are common to all marine regions, and another 102 species are dispersed among several viromes (Angly et al., 2006). While ubiquitous, it is thought that they can differ in their relative abundances in different viral communities (Angly et al., 2006; Breitbart and Rohwer, 2005). Through time, the most abundant phage may swap "rank", jumping from low background abundance to most abundant, analogous to the classic "everything is everywhere, but, the environment selects" dogma of Baas Becking (de Wit and Bouvier, 2006), now so commonly[11] used to explain the rare microbial biosphere (Sogin et al., 2006). But in the case of phages, non-living entities that experience environmental "selection" only during infection[12], *is* this entirely analogous? The relative abundance of distinct phage "genotypes" is a consequence of their dependence on changes in host density (and thus, also on their ability to alter host ranges), as will be further discussed in the next section.

[11] though controversially

[12] not considering physical/chemical influences on decay

1.1.3 Influence of phages on host microbial communities at sea: Pirates or Partners?

"The bacteriophage is everywhere...in a ceaseless struggle which goes on in nature between bacteriophage and bacteria...In nature every time that bacteria do something, a bacteriophage interferes and destroys the bacteria, or provokes a modification of their action."
Félix d'Hérelle, in a 30-year retrospective (1949)

Over the last few decades, marine phage research has built upon this insight of the grandfather of phage biology, Félix d'Hérelle[13], revealing key mechanisms through which phages influence their host microbial communities, best summarized as:

1. host mortality and the ensuing consequences on carbon flow, the microbial loop, and host community diversity ("Kill the Winner")
2. contributing to host metabolism through their own host-like "auxiliary metabolic genes"
3. driving host genome evolution: phage-laden genomic islands and CRISPRs

1.1.3.1 "Death by phage it shall be!"

Influence of mortality on microbial loop and aquatic food webs. Half of the world's carbon is fixed every day by photosynthesizing marine microbes, transforming CO_2 into organic carbon, a significant portion of it by *Prochlorococcus* in the oligotrophic regions of the ocean (Cavicchioli et al., 2003). Through this process, carbon enters the "microbial loop" of the marine food web. Now in particulate form, this bacterially assimilated carbon moves 'up' the web through grazing by eukaryotes (i.e., protozoa, then metazoa, and so on), making it available to higher trophic levels (Figure 1.4 Thingstad et al., 2008). Marine phage interfere directly with the efficiency of this process. As the microbial component of the marine ecosystem is completely lysed every two days due to viral infection (Suttle, 2005), this photosynthetically fixed organic carbon is recycled (or "shunted") back to the dissolved organic carbon (DOC) pool. Some portion of the DOC is taken

[13] Félix d'Hérelle (in 1917) and Frederick Twort (in 1915) were the first to discover and describe the effects of bacteriophages in microbial cultures. At the time, they both immediately recognized the value of administering phage to treat disease and infection, in what we now call "phage therapy." This concept was soon overshadowed by the discovery of antibiotics. However, with the current increased awareness and threat of antibiotic resistant bacteria, the Western world is taking a second look at the application of phage in both the food and pharmaceutical industries. Since their discovery nearly 100 years ago, many parts of the Eastern world, led by researchers in Georgia, have continued to develop and successfully administer phage therapy to treat both human and animal disease. (*Phages in Interaction Symposium III*, Leuven, Belgium, December 2009)

up by heterotrophic bacteria that either (i) pass it up the food web (again as prey to grazers) or (ii) are lysed by their own phages, again returning the carbon to the DOC pool (Thingstad et al., 2008). This action of "short-circuiting" the movement of organic carbon through the microbial loop and preventing it from moving up the aquatic food chain, is coined the "viral shunt" (Suttle, 2005). This is expected to be the fate of as much as a quarter of all primary production in the ocean (Suttle, 2007). Further, as some dissolved carbon escapes the microbial loop altogether, it settles to the deep ocean where it becomes recalcitrant dissolved organic matter, which is less accessible to heterotrophs and thus excluded from the entire system.

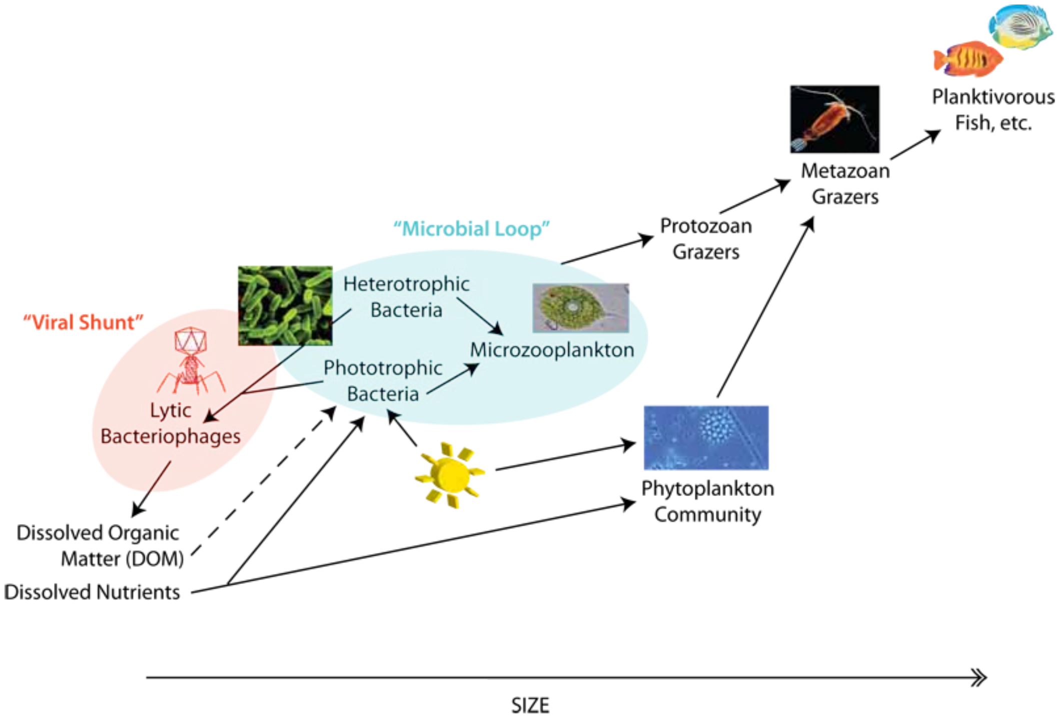


Figure 1.4: Marine food web depicting the flow of carbon, nutrients and energy up the chain, including the role of viruses in "shunting" dissolved organic matter away from the higher trophic levels. Dashed arrow indicates that DOM goes to only the heterotrophic members of the microbial loop, while phototrophs fix carbon through photosynthesis. Image modified from (Thingstad et al., 2008). For simplicity, this depiction omits the role of chemoautotrophs, who have an analogous role as phototrophs, when sun is replaced by energy derived from chemical reactions. It can be assumed they are equally as susceptible to phages as other microbes.

This is the current, rather simplistic, model of how phages contribute to the microbial loop. However, they may also impose a number of other provocative feats. For instance, they may be directly grazed upon by protozoa (Suttle and Chen, 1992) or their nucleic acids, when infection fails, may serve as sustenance to microbes, which is especially important when resources are scant. Suttle (2007) has also implicated viruses in influencing the efficiency of the Biological Pump, the process by which carbon is sequestered from our atmosphere and moves to the deep ocean, a vital process in our current global climate awareness. Yet, we lack significant information about the rates of key processes and the fate of carbon and nutrients released due to lysis (Suttle, 2007).

"Killing the Winner" maintains diversity. "Hutchinson's paradox" addresses a classic problem in, e.g., phytoplankton ecology:

> *"In a homogenous aquatic environment, how do diverse phyto- and bacterioplankton assemblages coexist, though competition for similar resources would predict only a few dominant species?"*
> Hutchinson (1962)

By "killing the winner", viruses have been proposed as a mechanisms to resolve this paradox (Thingstad and Lignell, 1997). Should one species benefit from a selective force (Figure 1.5), such as drastic input of carbon from an algal bloom, it will come to dominate the microbial community. As phage infection is a consequence of Brownian-like random encounters between particles (e.g., host and phage), any increase in either component will lead to increased incidence of infection. Thus, in a density dependent fashion, following a host "bloom", its specific virus population will infect at a higher rate and simultaneously become more abundant itself. The host bloom will thus be modulated, bringing it back to baseline abundances and preventing it from outcompeting other members of the community (Figure 1.5). Through this mechanism of selective lysis, phages maintain bacterial species diversity through predation[14].

1.1.3.2 Auxiliary Metabolic Genes: Mucking with Metabolism

As more marine phage genomes are sequenced, it becomes increasingly common to find host-like metabolic genes in their genomes (Mann et al., 2005, 2003; Rohwer et al., 2000; Sullivan et al., 2005, 2009), referred to as auxiliary metabolic genes ("AMGs") (Breitbart et al., 2007). Environmentally significant functional genes involved in photosynthesis, phosphate stress response, vitamin biosynthesis, antibiotic resistance, and nitrogen fixation, have also been identified on the 'viral' clas-

[14] This model also explains how the most abundant viral populations are likely to reflect the most abundant host population a any given time and space, with implications on viral diversity, as previously discussed 1.1.2.

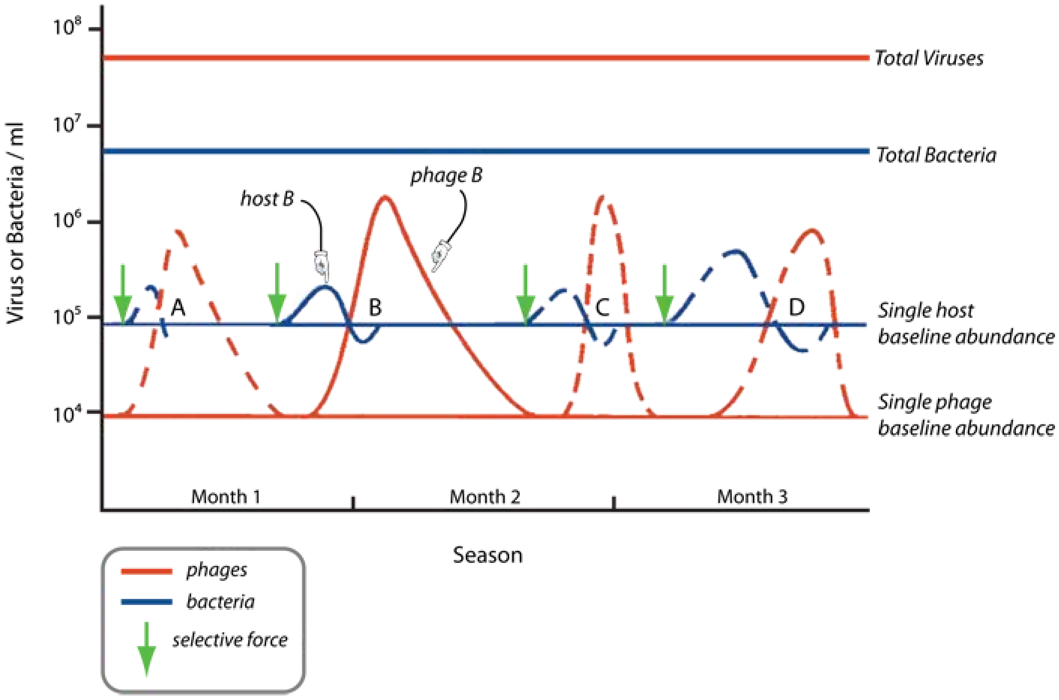


Figure 1.5: "Kill the Winner" model proposed (and well accepted) to resolve Hutchinson's paradox as a mechanism by which host density-dependent phage predation maintains high species diversity in a community of organisms competing for the same limited resources. Image modified from (Wommack and Colwell, 2000).

sified Global Ocean Survey (GOS) scaffolds (Williamson et al., 2008), suggesting AMGs to be a common feature of marine virus genomes. It is believed that, when expressed, these phage-carried genes supplement host metabolism at key limiting steps where successful production of their virus progeny is threatened by bottlenecks in, i.e., nucleotide metabolism, replication and phage production (Breitbart et al., 2007).

The best studied AMG system is that of the photosynthesis genes commonly found in cyanophages (Mann et al., 2005, 2003; Sullivan et al., 2005). These genes are not only present, but are expressed (Clokie et al., 2006; Lindell et al., 2007) and produce proteins (Lindell et al., 2005) during infection. It is likely that this is not a trivial process. Cyanobacteria, namely *Prochlorococcus*, are the most abundant photosynthesizers producing our planet's oxygen, and many of their highly abundant phages contain photosynthesis-related AMGs of their own. Some of the very air we breath is due to the direct contribution of phage photosynthesis genes expressed during infection.

Such observations beg to ask: does the viral shunt divert carbon away from the microbial loop comparable to the degree phages contribute to host metabolism (and production) through critical metabolic pathways? The relative impact of viruses of the two aforementioned processes has yet to be determined. Single-cell analysis of phage infection may provide useful metrics by which we may be able to scale up and extrapolate the impact of phage-mediated metabolism on a global scale.

1.1.3.3 Impact on host genomes: the phage footprint

Genomic islands When closely related strains of bacteria are compared, distinct regions of dissimilarity, or 'Genomic Islands', emerge (Coleman et al., 2006). These islands are recombination hotspots, showing signs of prolific lateral gene transfer, rearrangements, gene gain/loss, repeat elements, and association with tRNA genes (Coleman et al., 2006; Kettler et al., 2007; Sullivan et al., 2009). In *Prochlorococcus*), these islands can contain up to 80% noncyanobacterial genes, including those from phages, Eukarya, and Archaea (Coleman et al., 2006). These regions are enriched in strain-specific genes related to physiological stress (phosphate and nitrogen limitation, light) and nutrient uptake (cyanate transporter, lyase, amino acid and trace metal transporters), which likely drive the development and success of niche-specific ecotypes in the ocean (Coleman et al., 2006; Sullivan et al., 2009). Based on typical island content, such as phage integrases, DNA methylases, transcriptional regulators, endonucleases, and potential phage insertion sites (Sullivan et al., 2009), there is now convincing evidence that phage, to a significant extent, mediate the lateral transfer of genes to and away from these genomic islands.

In addition to genes with an explicit role in environmental response, the islands also encode cell surface proteins (Kettler et al., 2007). The distinguishing feature differentiating the most closely related *Prochlorococcus* isolates are genes related to outer membrane synthesis (Kettler et al., 2007; Venter et al., 2004), which not only play a role in selective substrate uptake, but are common phage receptors. This pattern, first described in *Prochlorococcus* (Coleman et al., 2006; Kettler et al., 2007), is consistent among numerous other marine and non-marine microbes, including three strains of *Candidatus* Pelagibacter ubique, *Shewanella* spp., and *Synechococcus* (Figure 1.6, Rodriguez-Valera et al., 2009). The hypervariable islands of these genomes are dominated by genes encoding proteins for nutrient transport, environmental sensing, and phage recognition sites. All islands contained extracellularly exposed lipopolysaccharides, while exopolysaccharides, pili, flagellar components, and extracellular giant proteins were also present, all of which are putative phage recognition and docking sites (Rodriguez-Valera et al., 2009). Rodriguez-Valera et al. proposed that selection to avoid phage predation by altering such proteins involved in selective substrate uptake can lead to altered substrate usage. When selective sweeps in a strain population occur due to, for instance, an adaptive mutant with improved substrate uptake, the fitter mutant increases in number. Simultaneously, the phage population that recognizes this new receptor will increase and thereby keep the invasive clonal population in check, thus unable to outcompete other members of the community. In this manner, functional diversity and a community more efficient at total resource exploitation is maintained (Rodriguez-Valera et al., 2009). In essence, this a is more resolved "Kill the Winner" model (influencing strain/ecotype-level, rather than species level, diversity) with direct consequences on ecosystem functioning.

Taken together, this implicates phages as important drivers of host niche adaptation and diversification by way of genomic island modification, particularly among genes responsible for selective substrate uptake.

CRISPRs Another remarkable system whereby phages leave a literal footprint in their host (both bacterial and archaeal) genomes, is that of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs). This genetic system confers nucleic acid-based phage immunity to microbes (Barrangou et al., 2007). During infection, hosts acquire short fragments of phage genomic material (or plasmid or transposons DNA), which become spacers separating direct repeats concatenated to other previously acquired phage fragments in identifiable loci (Banfield and

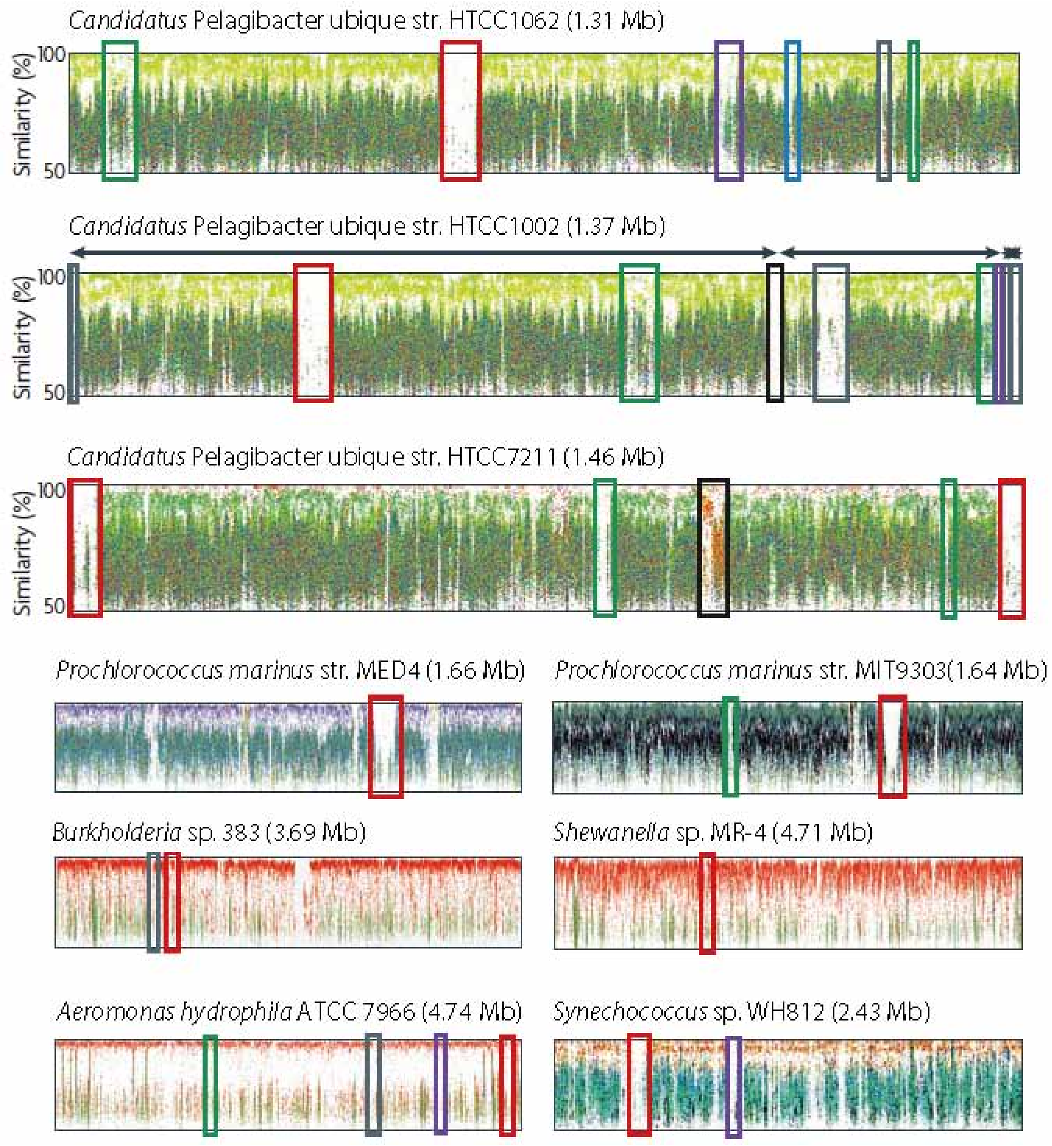


Figure 1.6: Hypervariable genomic islands in microbial genomes are indicated in boxes. The regions emerge when Global Ocean Sampling (GOS, Rusch et al., 2007) metagenome reads are mapped to highly represented whole genomes (>90% nucleotide sequence identity over 80% of the genome). Areas of low read coverage are thought to be hypervariable and strain specific, thus are poorly represented among all members of the community. In these, and other, genomes, the hypervariable islands typically contain large, extracellular proteins involved in both substrate uptake, as well as phage recognition and docking. Image from Rodriguez-Valera et al. (2009).

Young, 2009). This creates a veritable historical account of previous phage exposure in host genomes. When these phage-derived spacers are transcribed during infection by a phage the host has seen in its "clonal past", the spacers act as small RNAs that interfere with identical nucleic acids of the infecting phage. With the help of CRISPR-associated (Cas) proteins, the phage-like spacer interferes with phage protein production and infection is disrupted (Banfield and Young, 2009). However, viral populations can respond with their own slew of tractable counter mutations (Barrangou et al., 2007) in an effort to overcome this RNA interference. And so goes the CRISPR dance.

1.1.4 Phage Ecology and 'Ecogenomics'

"*T4 [phage] thrives in a domain of scale where matter and energy function very differently from the way they do in the everyday world of our visual intuition.*" Stephen T. Abedon

This interplay between phage, matter, and energy underlies the very concept of phage ecology. According to the traditional definition of Ernst Haeckel (1866), ecology is "*the comprehensive science of the relationship of the organism to the environment.*" Whereas, evolutionary ecology addresses the evolution of organisms as they adapt to their environment, encompassing both biotic and abiotic components of the ecosystem. As such, though there is no lack of evidence supporting their ability to undergo Darwinian evolution, is there even such a thing as *phage ecology*?

An immediate exception to the definition of ecology must be made, in that phages are not *organisms* (which implies life[15]). Rather, in the absence of other living entities (typically hosts), viruses are ″nothing but inanimate complex organic matter" (Moreira and López-García, 2009). How does a non-living entity interact with its environment? A closer look at phage biology and replication strategies reveals the great extent to which environmental factors do indeed influence phage activity, justifying the, at first seemingly nonsensical, field of phage ecology.

1.1.4.1 Phage biology primer

In order to appreciate the dearth of factors involved in the "relationship between phage and their environment", a basic understanding of phage biology and replication strategies is required (Figure 1.7).

[15] The definition of "Life", in the eyes of the author and for the purpose of this thesis, will be borrowed from (G. Joyce, 1994) "*a self-sustained chemical system capable of undergoing Darwinian evolution*" (also the official definition adopted by NASA). The underlying assumption being that "capable" refers to when this system is in a suitable environment for its "chemistry" to be favorable.

Upon adsorption to a recognizable phage receptor on a host cell, infection begins. The phage genome is injected and a molecular "decision" is made: the phage will either:

- take over host cellular machinery and immediate begin processes of replication (Figure 1.7a): LYSIS
- integrate into the host genome as a prophage; the host is now called a "lysogen" and phages with this capability are called "temperate" (Figure 1.7b): LYSOGENY
- neither replicate nor integrate, remaining in a stalled state referred to as "pseudolysogeny" (Figure 1.7c): PSEUDOLYSOGENY

In most double stranded DNA phages, the final outcome of phage infection is production of progeny and host lysis. Lysis can be broken down into three stages (Figure 1.8):

- latent period: time from phage adsorption to first host lysis
- eclipse period: time from initial infection to production of first intracellular phage; during this time, the phage takes over host replication machinery and begins protein production
- rise: "period [begins] when populations of synchronously [infected] bacteria begin lysing and ends when these populations finish lysing" (Max Delbrück, 1942); essentially, (if infection is synchronized across the host population) this is the period between release of first and last phage of a new generation

In its most basic form, phage ecology is the study of environmental influences on these parameters, which constrain phage replication strategies (Figure 1.7) and growth rates (Figure 1.8). The following section describes systems evidencing phage ecology as manifested through observable changes in phage "behavior", such as latent period, burst size, decay rates, host ranges, the decision between lysis, lysogeny, and pseudolysogeny, etc.

1.1.4.2 Examples and mechanisms of phage ecology: responding to the biotic (host) environment

Assessing phage "decision making" and growth rates Though the factors underlying the first molecular "decision" (options a-c Figure 1.7) are not well known, experimental evidence supports that it is heavily reliant on the nutritional/energetic status of the host (Miller and Day, 2008; Ptashne, 2004).

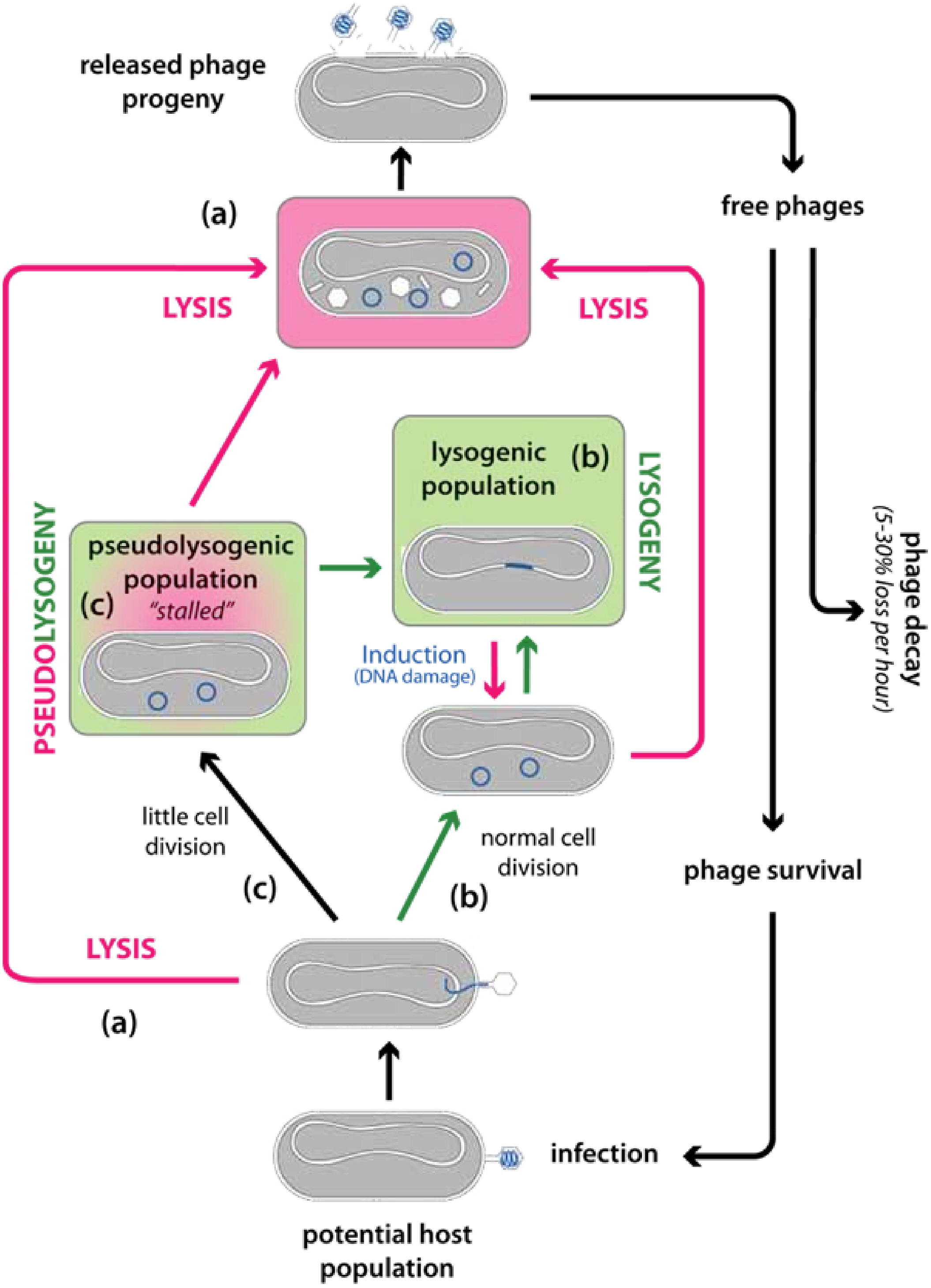


Figure 1.7: Phage replication strategies. For most dsDNA phages. all strategies end in host cell lysis, as is modeled in Figure 1.8. Little is known of the factors that influence the molecular "decision" at each juncture, especially in natural environments. Image modified from (Miller and Day, 2008).

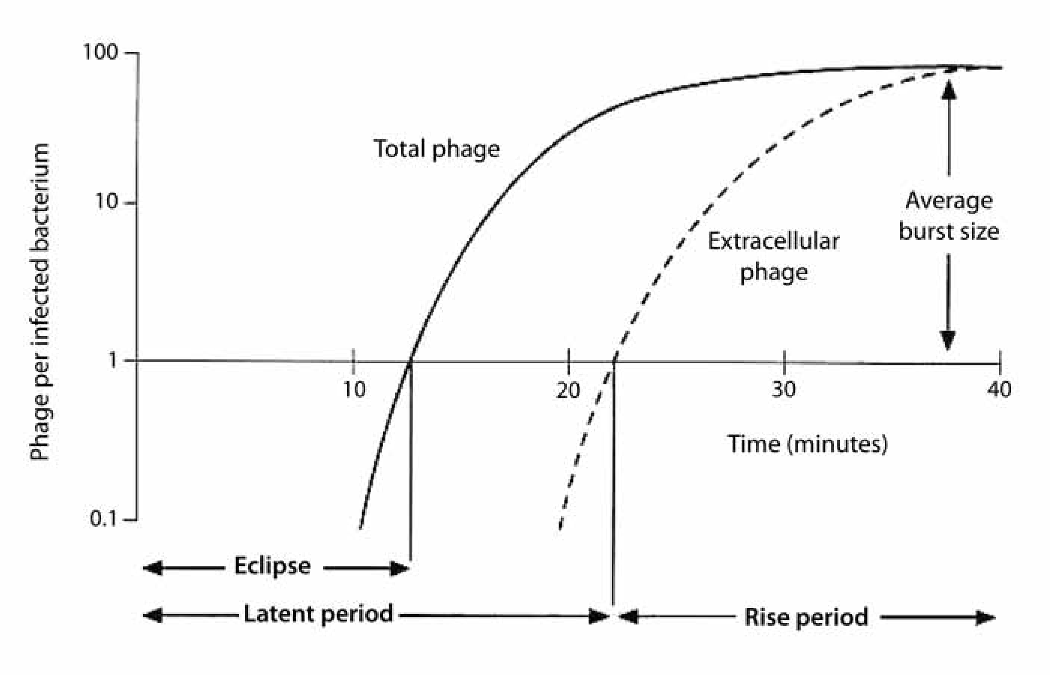


Figure 1.8: Model of phage one-step growth curve, representing different stages of the lytic cycle following adsorption of the phage during a successful infection. The relative rates and magnitudes of these parameters can vary in response to environmental stimuli. Image from (Baron, 1986).

As far as the impact on components of the lytic process, experimental work has shown that as *E. coli* growth rate increases (due to an environmental stimuli), the T4 phage infecting it will increase its rate of production and burst size, while decreasing both its eclipse and latent periods (Figure 1.8, Adams, 1959).

In a lysogenic system, when *E. coli* is infected with temperate phage λ and grown in rich media, it is more likely to proceed immediately to the lytic cycle; whereas growth on minimal media promotes integration (Ptashne, 2004). It is thought that the higher concentration of intracellular proteases in rapidly growing cells will degrade a λ protein (cII) essential for the cascade of integration processes (Ptashne, 2004). Support for this, outside of a few well-studied model systems, is lacking, though it is one of the major avenues by which phage can "respond" to their environment through an altered "behavior"[16].

In the environment, carbon and energy sources are scarce and bacteria are generally starved, especially in the marine system where oligotrophic species are thought to dominate (Lauro et al., 2009). Phage "behavior" is also modulated under such conditions, as phage in the environment are known to have longer latent periods, smaller burst sizes, and greater rates of decay (5-30% loss per hour, Miller and Day, 2008). Yet, free phages are still able to maintain high abundances (generally an order of magnitude more abundant that their hosts). It is believed that wild phages have evolved strategies to weather such low energy conditions, namely "pseudolysogeny" (Figure 1.7). This may enable them to persist in a "dormant" (non-replicating, non-lysing, non-integrating) state, despite slow growing host populations (Miller and Day, 2008), until more energetically favorable conditions prevail.

1.1.4.3 Phage 'Ecogenomics'

As a logical offshoot of ecology in the current "Age of Genomics"[17], 'ecogenomics' is an emerging field that can be described as the application of ecological theory and concepts to garner insights into genomic studies. In this way, ecology complements observations traditionally made through sequence analyses. The manner in which the environment affects the evolution and distribution of phage groups and genes, as observable through sequence data, is phage ecogenomics. The mechanism whereby phage experience habitat-specific selective *environmental* pressure on their gene content, is during gene expression while infecting their hosts. Thus,

[16] Quotations imply anthropomorphism. Phages neither respond nor behave; their actions are consequences of a cascade of molecular events driven by thermodynamics. A result of the most simple natural forces, these actions appear inherently "logical" to a human observer.

[17] as introduced in section 1.2

the "information" pertaining to selection on phage environmental interaction persists in the phage genomes.

MazG: 'ecogenomics' in action? In the phage MazG system (described below), phage may react to their hosts' physiological response to the environment via a mechanism that could be traced, not on the scale of infection rates and processes (as discussed thus far), but, through the influence of a single identifiable gene.

In *Escherichia coli*, MazG is a pyrophosphohydrolase known to play a role in programmed cell death under conditions of amino acid starvation, as a counter-component of the toxin-antitoxin mazEF "suicide" module (Gross et al., 2006). By causing a reduction in the cellular pool of ppGpp (a cell-death effector molecule), MazG activity leads to longer cell life, even when starved (Lee et al., 2008). In cyanobacteria under nitrogen starvation, the pool of ppGpp increases as the cell prepares for metabolic arrest or death (Friga et al., 1981). However, when infected by phage carrying a MazG homologue, cells do not decline (Borbély et al., 1980). Though the mechanism remains to be shown, the phage-encoded MazG may be maintaining the host metabolism long enough to ensure phage propagation (Clokie and Mann, 2006). Intriguingly, MazG is found prolifically among marine phages, particularly cyanophages (Bryan et al., 2008). Of the twelve phages in the MazG Pfam domain family (PF03819), six are marine (Table 1.1). This becomes significant when one considers that of the 580 sequenced phages, only 28 are marine: 21% of marine phages contain MazG, whereas 1% on non-marine phages contain it.

This MazG model may represent a process whereby the phage itself is adapted to succeed in low nutrient environments, yet, only as mediated through its hosts' experiences. The marine bias of the MazG domain further lends itself to hypotheses about the importance of the protein in phage-host systems of the low nutrient oceans. A similar marine-specificity is seen in the ribonucleotide reductase (RNR) gene, which is found in all marine T4-like and T7-like phages, as well as marine siphoviruses PSS2 and phi-JL001, though not found in non-marine T7-like phages (Sullivan et al., 2009). During phage infection, RNR-encoding genes are thought to supplement host DNA synthesis by scavenging nucleic acids, important in the low nitrogen and phosphorus marine environment (Sullivan et al., 2009). Such processes should leave tractable remnants in phage genomes and metagenomes as they adapt to certain environments, which is thus identifiable through 'ecogenomics'.

Unraveling the influences driving such phage ecogenomic patterns is an emerging field and the main motivation of this thesis, as will be outlined at the end of

Table 1.1: Phage genomes containing the MazG protein domain (PF03819), including their taxonomic classification and at least one host (when known, this is the isolation host). Also included is Pseudoalteromonas phage H105/1, recently sequenced and discussed in Chapter 4, which contains the MazG domain. For more information on the isolation location of the marine phages, see Figure 3.6)

.

Phage	**Phage Family**	**Host**
Marine **(of 28)**		
Roseobacter phage SIO1	Myoviridae	*Roseobacter* sp. SIO1
Synechococcus phage syn9	Myoviridae	*Synechococcus* sp. WH 8109
Synechococcus phage S-PM2	Myoviridae	*Synechococcus* sp.
Prochloroccocus phage P-SSM2	Myoviridae	*Prochlorococcus marinus* str. NATL1A
Prochloroccocus phage P-SSM4	Myoviridae	*Prochlorococcus marinus* str. NATL2A
Pseudoalteormonas phage H105/1	Siphoviridae	*Pseudoalteromonas* sp. H105
Non-marine **(of 552)**		
Mycobacterium phage Che12	Siphoviridae	*Mycobacterium tuberculosis*
Mycobacterium phage L5	Siphoviridae	*Mycobacterium smegmatis*, *Mycobacterium tuberculosis*
Staphylococcus phage G1	Myoviridae	*Staphylococcus aureus*
Staphylococcus phage K	Myoviridae	*Staphylococcus aureus DPC5246*
Haemophilus phage HP1	Myoviridae	*Haemophilus influenzae* L-10
Myxococcus phage Mx8	Myoviridae	*Myxococcus xanthus* strain DK883
Listeria phage P35	Myoviridae	*Listeria monocytogenes*

this chapter (Section 1.3).

1.2 Theory Behind the Bioinformatic and Genomic Approach

This section introduces the platform needed to address the phage "ecogenomic" themes previously proposed.

1.2.1 Bioinformatics as it pertains to us

At its most basic, bioinformatics, first coined by Pauline Hogeweg (University of Utrecht) in 1978 (Hogeweg, 1978), is "Information technology as applied to the life sciences, especially the technology used for the *collection*, *storage*, and *retrieval* of genomic data." (Dictionary, 1991).

1.2.1.1 Bioinformatics in the "*Coming of* Age of Genomics"

Since its colloquial inception over 30 years ago, when it was initially used to describe computer simulations modeling cell growth, "bioinformatics" has grown hand in hand with advances in molecular biology, namely those pertaining to genomics and sequencing technology. Encompassing the "collection, storage, and retrieval" components, the International Nucleotide Sequence Database Collaboration (INSDC) has become a bioinformatic hub, as the umbrella organization of three more commonly known sequence databases: GenBank (USA), EMBL Nucleotide Sequence Database (Europe), and DDBJ (Japan), which exchange new and updated data daily. Since its inception in 1982, "the number of bases in GenBank has doubled approximately every 18 months" (Figure 1.9; GenBank release notes 162.0), and now houses over 110 Gbp (nearly 113 million entries), requiring roughly 422 GB storage space (GenBank release 175.0, December 2009).

In fact, any tome highlighting currents in bioinformatics can not go without impressing the fact that, due to remarkable advances in sequencing technologies, we are now at a point where the amount of data we amass may soon surpass our ability to analyze it (editorial, 2009). Though understandably attractive to biologists seeking to decipher the genomic bases of biological patterns, the ever-popular "second generation sequencing platforms" [18] generate much shorter (50-90%) reads, which require new algorithms and pipelines to analyze, and generate exponentially increasing amounts of data (editorial, 2009). The first "second-generation" metagenome generated 40 Mbp, while, depending on the method, newly sequenced metagenomes can range from 200 Mbp (pyrosequencing) to 20,000

[18] "second-generation" platforms refer to pyrosequencing (Margulies et al., 2005), Illumina (Oliphant et al., 2002), and SOLiD (ABI Life Sciences) platforms, as opposed to the "first-generation" microfluidics-based Sanger sequencing (Sanger et al., 1977), which initiated the "Age of Genomics." Frederick Sanger was awarded the Nobel prize for his contributions to this advancement in 1980.

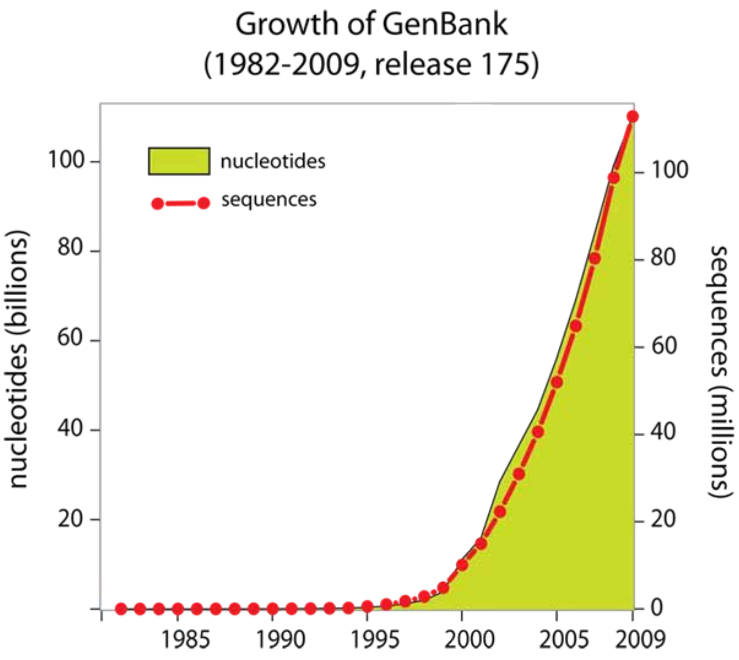


Figure 1.9: Growth of the GenBank database (non-WGS) from 1982 to the last release, 15 December 2009 (release 175.0). This release was comprised of 422 GB of sequence data. Data retrieved from NCBI 29 January 2010.

Mbp (Illumina-based sequencing) (editorial, 2009). Today, we are limited primarily by our computing power, which can be alleviated only through (i) increased funding dedicated to data analysis, (ii) centralization of data sharing and computing efforts, and (iii) and, particularly, *data standardization* (editorial, 2009). The role of bioinformatics in the "Coming of Age of Genomics" is to keep this data accessible, or we risk a serious loss in data value as well as our ability to do science efficiently.

1.2.1.2 The power of standards

The role of contextual data standards In the most simple sense, data standards mean more data. Contextual data standards, the most relevant to the topic of this thesis, enforce raw sequence data to be accompanied by a slew of contextual data describing the sequences: data about the data. The power of contextual data to an end-user biologist touches upon the core tenets of comparative genomics. For instance, given two sequences, one can do an alignment and assess how similar they are, even make hypotheses about their evolution. However, when these sequences are accompanied by contextual information describing, e.g., the organism from which they were derived, the habitat of this organism, the exact time and coordinates of the organism collection, etc., comparisons can extend beyond sequence analysis to include context. *Ecogenomic* hypotheses are then possible. ″Are these sequences more closely related because they are from the marine environment?" "or because they are derived from the same species?" "or because they are found only at hydrothermal vents below Antarctic sea ice?"

A greater number of descriptors associated with each sequence allows for more complex comparisons. The opportunity to consider such complexities proffers a greater chance of finding the most relevant explanation for observed sequence associations, or, in other words, the most significant source of selective pressure leading to evolution of a given system. It is through this process, facilitated by the implementation of contextual data standards, that we may, at the very least, begin to grasp the complexities of the systems with which we flirt.

Progress: the Genomic Standards Consortium (GSC) The value of genome standards is becoming increasingly appreciated on an international scale. The Genomic Standards Consortium (GSC) is a grassroots initiative of scientists working towards the common goal of achieving richer descriptors of the public collection of genomes, metagenomes, and marker genes. In 2008, the Minimal Information about a Genome Sequence (MIGS) standards were established (Field et al., 2008), and progress has continued. The relevance to and influence of these standards on marine phage genomics is further discussed in Chapter 3.

1.2.1.3 The Funnel Concept

Bioinformatics will never deliver a *Biological Truth* about, i.e., the holitic functioning of a system, nor is it expected or intended to. It is, however, an essential field to embrace to approach the *Truth*[19], especially considering our current, and ever-increasing, information over-saturation. As it applies to biological understanding, one can think of the application of our suite of bioinformatic tools as a funnel (Figure 1.10). A successful bioinformatic platform, in its ability to "collect, store, and retrieve" genomic data, will accelerate us beyond our human limitations by making connections and abstractions in the data where we are unable to see them.

It is through bioinformatic applications and the development and implementation of data standards, that the "structure" of the ′bioinformatics funnel′ is imposed (Figure 1.10). Such structure makes it possible to get from 113 million GenBank entries of nucleotide and amino acid sequences to, for instance,

[19] ...or a close approximation to the *Truth*, i.e., how we perceive it and justify it through our own experiences.

1. a single phage genome (or subset of, i.e., marine phage genomes), in order to
2. make an ecology-based hypothesis ("ecogenomics"), with the end goal of
3. designing an intriguing lab-based experiment (Figure 1.10).

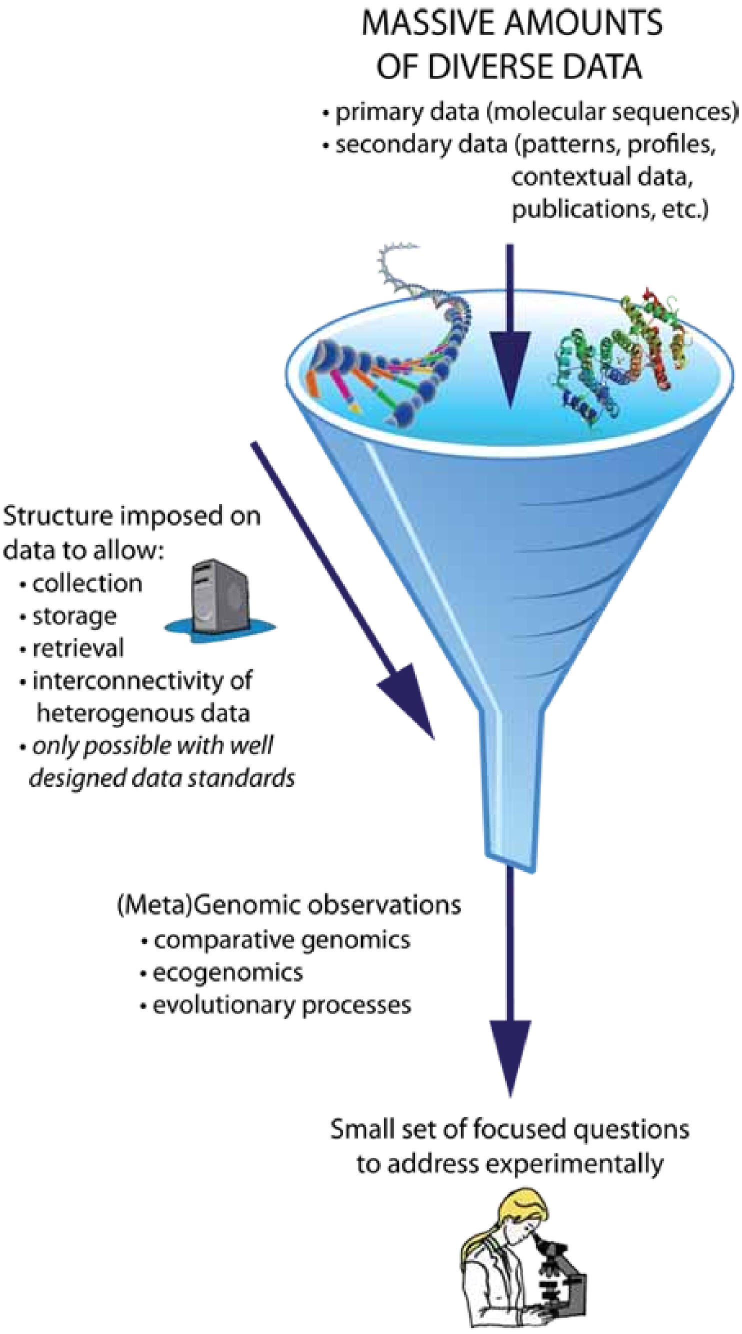


Figure 1.10: Modeling the application of our suite of bioinformatic tools to answer biological questions.

1.3 Motivation and Research Aims

Outstanding questions in marine phage ecology and genomics

1. What insights can we garner about *phage ecogenomics*?
2. Is there structure in the distribution of environmentally relevant phage genes (i.e., auxiliary metabolic genes to supplement host metabolism, genes of other mechanisms whereby phage can respond to their environment, etc.)?
3. Will this structure correlate to environmental parameters?
4. Can ecogenomic findings be translated to "wet lab" experimental design and validation?

Research Aims

Before hypotheses can be generated and tested pertaining to phage ecology and ecogenomics, a platform for consistently collecting and storing contextual data, such as phage taxonomy, host range, and (most importantly) habitat and isolation location (including associated descriptive data), must be established. The development, assessment, and initial application of such a framework are the main objectives of this thesis.

Through contributions to all components of the "bioinformatic funnel" (Figure 1.11), this work sets the stage to address questions pertaining to phage biology through bioinformatic development (Chapter 2) and assessment (Chapter 3), and genomic discoveries (Chapter 4), for a final foray into experimental phage biology to test hypotheses birthed by bioinformatic and genomic experiences (Chapter 5).

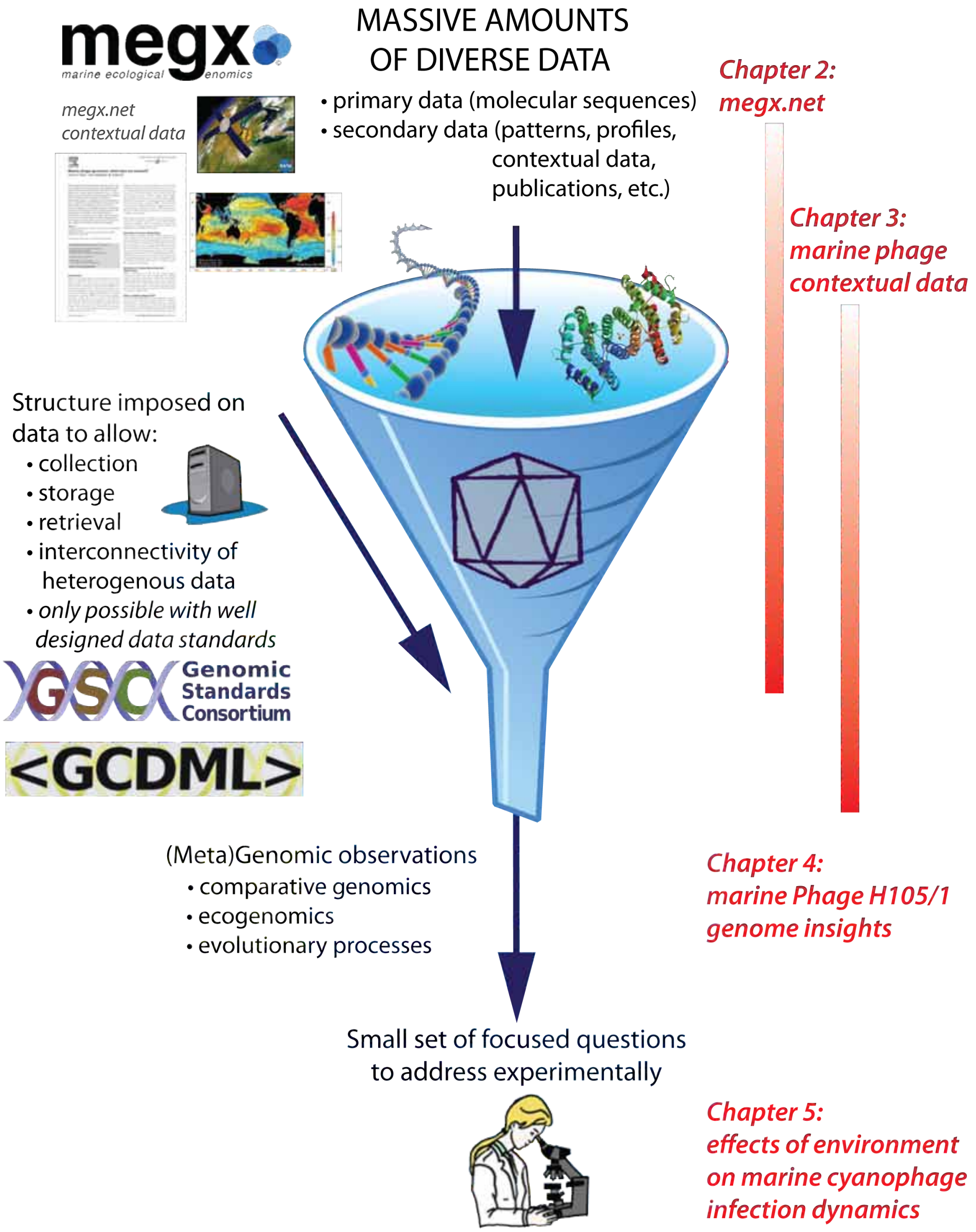

megx
marine ecological genomics
megx.net
contextual data
MASSIVE AMOUNTS
OF DIVERSE DATA
• primary data (molecular sequences)
• secondary data (patterns, profiles, contextual data, publications, etc.)
Chapter 2:
megx.net
Chapter 3:
marine phage
contextual data
Structure imposed on data to allow:
• collection
• storage
• retrieval
• interconnectivity of heterogenous data
• only possible with well designed data standards
GSC Genomic Standards Consortium
<GCDML>
(Meta)Genomic observations
• comparative genomics
• ecogenomics
• evolutionary processes
Chapter 4:
marine Phage H105/1
genome insights
Small set of focused questions
to address experimentally
Chapter 5:
effects of environment
on marine cyanophage
infection dynamics

Figure 1.11: Components of this thesis as they pertain to the "bioinformatic funnel."

1.4 Content Overview

Overview of work published, submitted, and in progress within the confines of this thesis.

Chapter 2

MEGX.NET: INTEGRATED DATABASE RESOURCE FOR MARINE ECOLOGICAL GENOMICS

Authors: Renzo Kottmann, Ivaylo Kostadinov, Melissa Beth Duhaime, Pier Luigi Buttigieg, Pelin Yilmaz, Wolfgang Hankeln, Frank Oliver Glöckner

Published in: *Nucleic Acids Research*. 2010

Personal Contribution: assembled and curated the marine phage collection, contributed biological knowledge to database design concept (implemented by RK and IK), web portal design and programming (with RK, IK, PLB, PY, and WH), wrote the core manuscript

Relevance: To introduce the core megx.net concept and technological back-end, and how the phages fit into the research platform of the Microbial Genomics Group at Max Planck Institute for Marine Microbiology.

Chapter 3

ENRICHING PUBLIC DESCRIPTIONS OF MARINE PHAGES USING THE MIGS STANDARD: A CASE STUDY ASSESSING THE CONTEXTUAL DATA FRONTIER

Authors: Melissa Beth Duhaime, Renzo Kottmann, Dawn Field, Frank Oliver Glöckner

Submitted to: *Standards in Genome Sciences*, 1 December 2009.

Personal Contribution: conducted all analyses, designed experiment (with RK), wrote the manuscript

Relevance: Holistic overview of contextual metadata for the sequenced marine phages. Introduces and highlights the benefits of contextual data (especially x, y, z, t and habitat descriptors), as well as contributions to GCDML development (the language that manages contextual data to be stored with sequence data; written by Kottmann et al.) through real-life application of the MIGS (Minimal Information about a Genome Sequence) standards.

Chapter 4

ECOGENOMICS AND GENOME LANDSCAPES OF MARINE PSEUDOALTEROMONAS PHAGE H105/1

Authors: Melissa Beth Duhaime, Antje Wichels, Jost Waldmann, Hanno Teeling, Frank Oliver Glöckner

Submitted: 14 January 2010

Personal Contribution: designed experiments, conducted all analyses, wrote the manuscript

Relevance: To apply, first-hand, the strengths of using contextual data for *in silico* hypothesis testing, as well as represent core phage biological knowledge, comparative genomes, and the evolution of genome signatures, as they pertain to a novel marine phage from the North Sea.

Chapter 5

INSIGHTS INTO INFECTION DYNAMICS OF OCEAN CYANOPHAGE PSS2: TO INTEGRATE, OR NOT TO INTEGRATE?

Authors: Melissa Beth Duhaime, Frank Oliver Glöckner, Matthew B. Sullivan

in progress, to be completed during post-doc

Personal Contribution: designed experiments (with MBS), conducted analyses, wrote preliminary manuscript concept

Relevance: This is a short, highly conceptual chapter, representing work in progress to investigate the effect of host nutritional status on components of a temperate marine *Prochlorococcus* host-phage system.

CHAPTER 2

Megx.net: integrated database resource for marine ecological genomics

Renzo Kottmann[a,†,*], Ivalyo Kostadinov[a,b,†], Melissa Beth Duhaime[a,b], Pier Luigi Buttigieg[a,b], Pelin Yilmaz[a,b], Wolfgang Hankeln[a,b], Jost Waldmann[a], Frank Oliver Glöckner[a,b,*]

[a]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

[b]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

[†]authors contributed equally to this work

[*]To whom correspondence should be addressed.

Tel: +49 421 2028974; Fax: +49 421 2028580; Email: rkottman@mpi-bremen.de

Received September 15, 2009. *Accepted* October 8, 2009; *Published* October 25, 2009

Nucleic Acids Research, 2010, Vol. 38 / *For direct online access:* doi:10.1093/nar/gkp918

Keywords: genomics / metagenomics / environmental data / database / mapserver

Abstract: Megx.net is a database and portal that provides integrated access to georeferenced diversity, environment, and marine genome and metagenome projects for microbial ecological genomics. All data are stored in the Microbial Ecological Genomics DataBase (MegDB), which is subdivided to hold both sequence and habitat data, and global environmental data layers. The extended system provides access to several hundreds of genomes and metagenomes from prokaryotes and phages, as well as over a million small and large subunit ribosomal RNA sequences. With the refined Genes Mapserver, all data can be interactively visualized on a world map and statistics describing environmental parameters can be calculated. Sequence entries have been curated to comply with the proposed minimal standards for genomes and metagenomes (MIGS/MIMS) of the Genomic Standards Consortium. Access to data is facilitated by Web Services. The updated megx.net portal offers microbial ecologists greatly enhanced database content, and new features and tools for data analysis, all of which are freely accessible from our webpage: http://www.megx.net.

2.1 Introduction

Over the last years, molecular biology has undergone a paradigm shift, moving from a single experiment science to a high throughput endeavour. Although the genomic revolution is rooted in medicine and biotechnology, it is currently the environmental sector, specifically the marine that delivers the greatest quantity of data. Marine ecosystems, covering more than 70% of the earth's surface, host the majority of biomass and significantly contribute to global organic matter and energy cycling. Microorganisms are known to be the "gatekeepers" of these processes and insights into their life style and fitness will enhance our ability to monitor, model and predict future changes. Recent developments in sequencing technology have made routine sequencing of whole microbial communities from natural environments possible. Prominent examples in the marine field are the ongoing Global Ocean Sampling (GOS) campaign (Rusch et al., 2007; Venter et al., 2004) and Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project (http://www.moore.org/microgenome/). Notably, the GOS resulted in a major input of new sequence data with unprecedented functional diversity (Yooseph et al., 2007). The resulting flood of sequence data available in public databases is an extraordinary resource with which to explore microbial diversity and metabolic functions at the molecular level. These large-scale sequencing projects bring new challenges to data management and software tools for assembly, gene prediction, and annotation —fundamental steps in genomic analysis. Several new dedicated database resources have emerged recently to tackle the current needs for large scale metagenomic data management, namely, CAMERA (Seshadri et al., 2007), IMG/M (Markowitz et al., 2008), and MG-RAST (Meyer et al., 2008). Nevertheless, it is increasingly apparent that the full potential of comparative genome and metagenome analysis can be achieved only if the geographic and environmental context of the sequence data is considered (Field, 2008; Field et al., 2008). The metadata describing a sample's geographic location and habitat, the details of its processing, from the time of sampling up to sequencing, and subsequent analyses, are important for, e.g., modeling species' responses to environmental change or the spread and niche adaptation of bacteria and viruses. This suite of metadata is collectively referred as contextual data (Kottmann et al., 2008). Megx.net offered the first database to integrate curated contextual data with their respective genes, genomes and metagenomes in the marine environment (Lombardot et al., 2006). Now, the extended megx.net database resource allows post factum retrieval of interpolated environmental parameters, such as temperature, nitrate, phosphate, etc., for any location in the ocean waters based on profile and remote sensing data.

Furthermore, the content has been significantly updated to include prokaryote and marine phage genomes, metagenomes from the GOS project (Rusch et al., 2007), and all georeferenced small and large subunit ribosomal RNA sequences from the SILVA database project (Pruesse et al., 2007). The extended megx.net portal is the first resource of its kind to offer access to this unique combination of data, with manually curated habitat descriptors for all genomes, metagenomes, and marker genes, their respective contextual data, and additionally integrated environmental data. See the megx.net online video tutorial for a guided introduction and overview at http://www.megx.net/portal/tutorial.html (Supplementary Material).

2.2 New Database Structure and Content

The Microbial Ecological Genomics DataBase (MegDB), the backbone of megx.net, is a centralized database based on the PostgreSQL database management system. The georeferenced data concerning geographic coordinates and time are managed with the PostGIS extension to PostgreSQL. PostGIS implements the "Simple Features Specification for SQL" standard recommended by the Open Geospatial Consortium (OGC; http://www.opengeospatial.org/), and therefore offers hundreds of geospatial manipulation functions. MegDB is comprised of (a) MetaStorage, which stores georeferenced DNA sequence data from a collection of genomes, metagenomes, and genes of molecular environmental surveys, with their contextual data, and (b) OceaniaDB, which stores georeferenced quantitative environmental data (Figure 2.1).

2.2.1 Contextual and Sequence Data Content

Sequences in MetaStorage are retrieved from the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/). However, as of September 2009, GOLD reported 5776 genome projects, of which, only 1095 were finished and published (http://www.genomesonline.org/gold.cgi). As most of the sequenced functional diversity is contained in these draft and shotgun datasets, megx.net was extended to host draft genomes and whole genome shotgun (WGS) data. Currently, MegDB contains 1832 prokaryote genomes (940 incomplete or draft) and 80 marine shotgun metagenomes from the GOS microbial dataset.

Marine viruses are a missing link in the correlation of microbial sequence data with contextual information to elucidate diversity and function. Consequently, megx.net now incorporates all sequenced marine phage genomes in MegDB, the

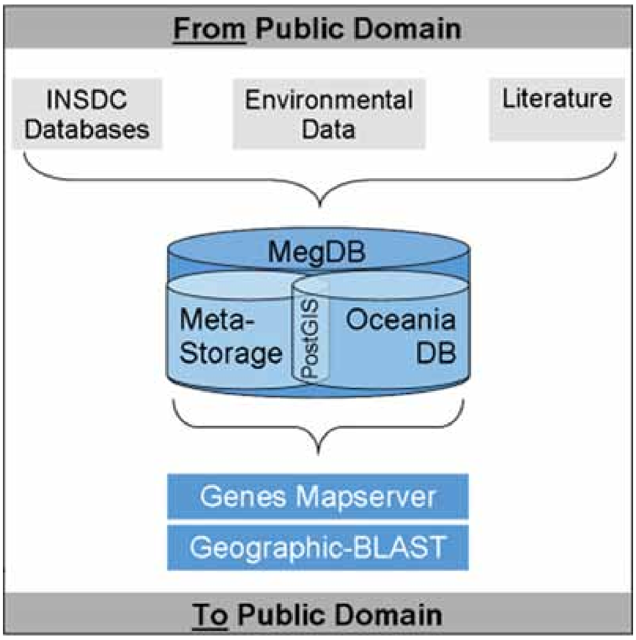


Figure 2.1: General Architecture of megx.net: DNA sequence data (from INSDC) is integrated with contextual data from diverse resources (i.e., manual literature mining, the GOLD database) and interpolated environmental data. MegDB integrates the data conforming to OGC standards and MIGS/MIMS specification. The core megx.net tools, Genes Mapserver and Geographic-BLAST, access the MegDB content.

first step towards a community call for integration of viral genomic and biogeochemical data (Brussaard et al., 2008). In an effort towards integrating microbial diversity with specific sampling sites, megx.net has been extended to include georeferenced small and large subunit ribosomal RNA (rRNA) sequences from the SILVA ribosomal RNA databases project (Pruesse et al., 2007). Currently, only 9% (16S/18S5) and 2% (23S/28S) of over one million sequences in SILVA SSUParc (16S/18S) and LSUParc (23S/28S) databases are georeferenced. With the implementation of the Minimal Information about an Environmental Sequence (MIENS) standard for marker gene sequences (http://gensc.org/gc_wiki/index.php/MIENS), efforts are ongoing to significantly improve this situation.

All genomic sequences in megx.net are supplemented by contextual data from GOLD (Liolios et al., 2008) and NCBI Genome Projects (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). The database is designed to store all contextual data recommended by the Genomics Standards Consortium (GSC), and is thus compliant with the Minimum Information about a Genome Sequence (MIGS) standard, and its extension, Minimum Information about a Metagenome Sequence (MIMS) (Field et al., 2008; Kottmann et al., 2008).

Furthermore, megx.net is the first resource to provide a manually annotated collection of genomes using terms from EnvO-Lite (Rev. 1.4), a subset of the Environment Ontology (EnvO) (Hirschman et al., 2008). An EnvO-Lite term was assigned to each genome project, identifying the environment where its original sample material was obtained. The annotation can be browsed on the megx.net portal using, e.g., tag clouds, and may be used as a categorical variable in comparative analyses.

2.2.2 Environmental Data Content

OceaniaDB was added to MegDB to supplement the georeferenced molecular data of MetaStorage with interpolated environmental parameters. When sufficient date, depth, and location measurements are provided, any 'on site' contextual data taken at a sampling site can be supplemented by environmental data describing physical, chemical, geological and biological parameters, such as ocean water temperature and salinity, nutrient concentrations, organic matter and chlorophyll. The environmental data is retrieved from three sources:

1. World Ocean Atlas: a set of objectively analyzed (one decimal degree spatial resolution) climatological fields of in situ measurements (http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html);

2. World Ocean Database: a collection of scientific, quality-controlled ocean profiles (http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html);

3. SeaWIFS chlorophyll a data (http://seawifs.gsfc.nasa.gov).

These data are described at 33 standard depths for annual, seasonal and monthly intervals. Together, the location and time data (x, y z, t) serve as a universal anchor, and link environmental data to the sequence and contextual data in MetaStorage (Figure 2.1). As such, megx.net integrates biologist-supplied sequence and contextual data (measured at the time of sampling) with oceanographic data provided by third party databases. All environmental data are compatible with OGC standards (http://www.opengeospatial.org/standards) and are described with exhaustive meta-information consistent with the ISO 19115 standard. Moreover, based on the integrated environmental data, megx.net provides information to aid biologists in grasping the ocean stability, on both global and local scales. For all environmental parameters, the yearly standard deviation of the monthly values can be viewed on a world map, for easy visualization of high and low variation sample sites. Furthermore, for each sample site, users can view trends in numerous parameters.

2.3 User Access

2.3.1 Genes Mapserver

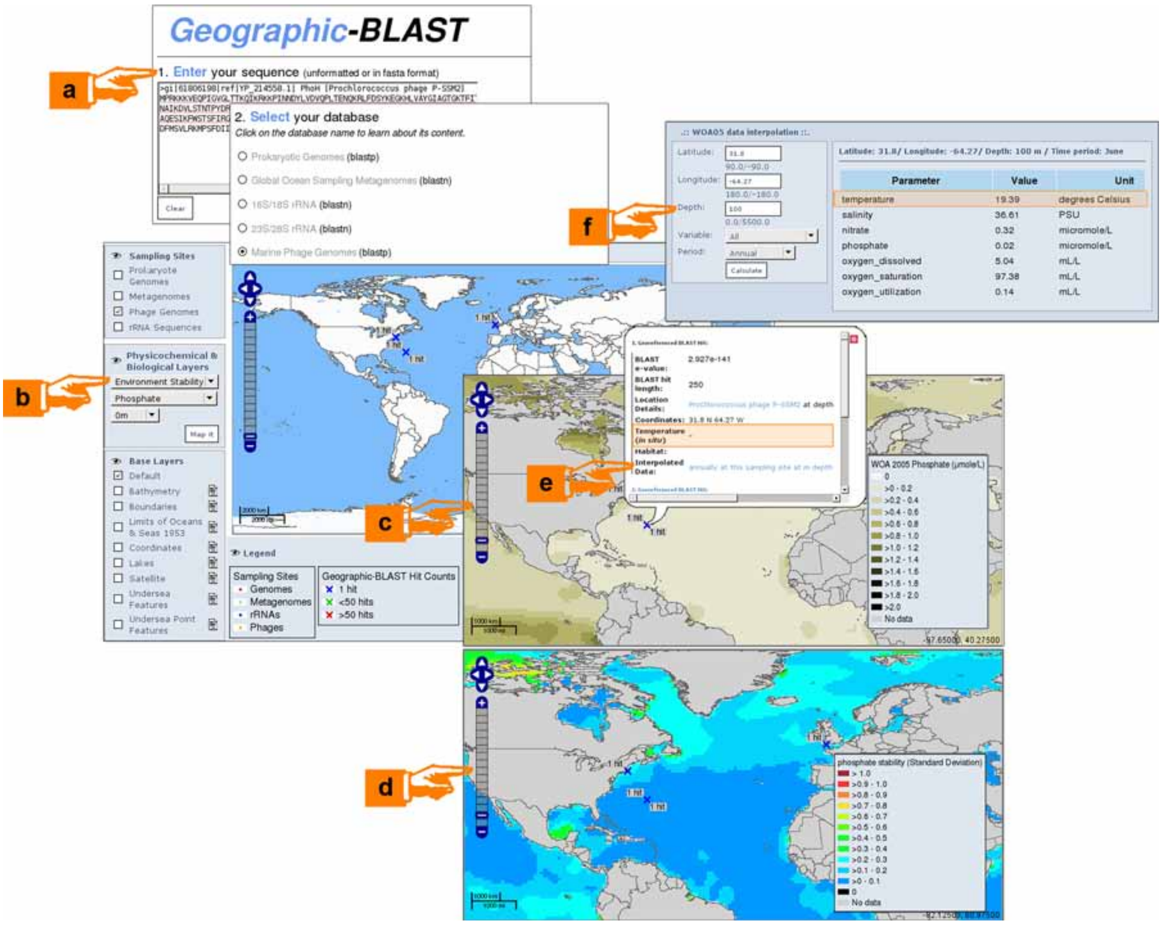
The Genes Mapserver (formerly Metagenomes Mapserver) offers a sample-centric view of the georeferenced MetaStorage content. Substantial improvements to the underlying Geographic Information System (GIS) and web view have been made. The website is now interactive, offering user-friendly navigation and an overlay of the OceaniaDB environmental data layers to display sampling sites on a world map in their environmental context. Sample site details and interpolated data can be retrieved by clicking the sampling points on the map (Figure 2.2). The GIS Tools of the Genes Mapserver allow extraction of interpolated values for several physicochemical and biological parameters, such as temperature, dissolved oxygen, nitrate and chlorophyll concentrations, over specified monthly, seasonally, or annually intervals (Figure 2.2f).

2.3.2 Geographic-BLAST

The Geographic-BLAST tool queries the MegDB genome, metagenome, marine phages, and ribosomal RNA sequence data using the BLAST algorithm (Altschul et al., 1990). The results are reported according to the sample locations (when provided) of the database hits. With the updated Geographic-BLAST, results are plotted on the Genes Mapserver world map, where they are labeled by number of hits per site (Figure 2.2). Standard BLAST results are shown in a table, which also provides direct access to the associated contextual data of the hits.

2.3.3 Software Extensions to the Portal

In addition to the services directly provided by megx.net, the project serves as a portal to software for general data analysis in microbial genomics. MetaBar (http://www.megx.net/metabar) is a tool developed with the aim to help investigators efficiently capture, store, and submit contextual data gathered in the field. It is designed to support the complete workflow from the sampling event up to the metadata enriched sequence submission to an INSDC database. MicHanThi (http://www.megx.net/michanthi) is a software tool designed to facilitate the genome annotation process through rapid, high quality prediction of gene functions. It clearly out-performs the human annotator in terms of accuracy and reproducibility. JCoast (http://www.megx.net/jcoast; (Richter et al., 2008)) is a desktop application primarily designed to analyze and compare (meta)genome sequences of prokaryotes. JCoast offers a flexible graphical user interface (GUI), as well as



Geographic-BLAST

1. Enter your sequence (unformatted or in fasta format)

```
>gi|61806198|ref|YP_214558.1| PhoH [Prochlorococcus phage P-SSM2]
MPRKKKVEQPIGVGLTTKQIKRKKPINNDYLVDVQPLTENQKRLFDSYKEGKHLVAYGIAGTGKTFI
NAIKDVLSTNTPYDR
AQESIKPWSTSFIRG
DFMSVLRKMPSFDII
```

2. Select your database

Click on the database name to learn about its content.

○ Prokaryotic Genomes (blastp)
○ Global Ocean Sampling Metagenomes (blastn)
○ 16S/18S rRNA (blastn)
○ 23S/28S rRNA (blastn)
● Marine Phage Genomes (blastp)

.:: WOA05 data interpolation ::.

Latitude: 31.8 (90.0/-90.0)
Longitude: -64.27 (180.0/-180.0)
Depth: 100 (0.0/5500.0)
Variable: All
Period: Annual

Latitude: 31.8/ Longitude: -64.27/ Depth: 100 m / Time period: June

Parameter	Value	Unit
temperature	19.39	degrees Celsius
salinity	36.61	PSU
nitrate	0.32	micromole/L
phosphate	0.02	micromole/L
oxygen_dissolved	5.04	mL/L
oxygen_saturation	97.38	mL/L
oxygen_utilization	0.14	mL/L

BLAST e-value:	2.927e-141
BLAST hit length:	250
Location Details:	Prochlorococcus phage P-SSM2 at depth
Coordinates:	31.8 N 64.27 W
Temperature (in situ):	-
Habitat:	
Interpolated Data:	annually at this sampling site at in depth

Figure 2.2: User Test Case: (a) BLAST a sequence against the marine phage genomes to see the results on the Genes Mapserver. (b) View the BLAST hits with underlying environmental data, such as (c) average annual phosphate values, or (d) stability of phosphate concentrations in terms of monthly standard deviation. (e) BLAST result information can be displayed in a pop-up window, (f) where you can link out to megx.net's GIS data interpolator.

an application programming interface (API) that facilitates back-end data access to GenDB projects (Meyer et al., 2003). JCoast offers individual, cross genome and metagenome analysis, including access to Geographic-BLAST.

2.3.4 User Test Case

To demonstrate the interpretation of genomic content in environmental context, consider a test case with the marine phages. Marine phage genomes (Sullivan et al., 2005) and 'viral' classified GOS scaffolds (Williamson et al., 2008) have revealed host-related metabolic genes involved in, i.e., photosynthesis, phosphate stress, antibiotic resistance, nitrogen fixation, and vitamin biosynthesis. Geographic-BLAST can be used to investigate the presence of PhoH (accession YP_214558), a phosphate stress response gene, among the sequenced marine phages. The search results can then be interpreted in their environmental context, either as (i) average annual phosphate measurements, or (ii) stability of phosphate concentrations in terms of monthly standard deviation (Figure 2.2c-d). A closer look at a single genome sample site reveals that in situ temperature was not originally reported (Figure 2.2e), whereas the interpolated data supplements this parameter (Figure 2.2f).

2.3.5 Web Services

The newly extended version of megx.net offers programmatic access to MegDB content via Web Services, a powerful feature for experienced users and developers. All geographical maps can be retrieved via simple web requests, as specified by the Web Map Service (WMS) standard. The base URL for WMS requests is http://www.megx.net/wms/gms, where more detailed information on how to use this service can be found. megx.net also provides access to MIGS/MIMS reports in Genomic Contextual Data Markup Language (GCDML) XML files for all marine phage genomes through similar HTTP queries, e.g., http://www.megx.net/gcdml/Prochlorococcus_phage_P-SSP7.xml (Field et al., 2008; Kottmann et al., 2008).

2.4 Other Changes

The massive influx of sequence data in the last years will out-compete the ability of scientists to analyze it (editorial, 2009). This development already pushes megx.net's capability to provide comprehensive pre-computed data to the limit.

To better focus on integration of molecular sequence, contextual, and environmental data, megx.net no longer offers pre-computed analyses, especially considering that other facilities, such as MG-RAST and CAMERA have emerged. Furthermore, the 'EasyGenomes Browser' has been replaced with links to the NCBI Genome Projects.

2.5 Summary

Since the first publication (Lombardot et al., 2006), megx.net has undergone extensive development. The web design has been revamped for better user experience, and the database content greatly enhanced, providing considerably more genomes and metagenomes, marine phages, and rRNA sequence data.

Megx.net's unique integration of environmental and sequence data allows microbial ecologists and marine scientists to better contextualize and compare biological data, using, e.g., the Genes Mapserver and GIS Tools. The integrated datasets facilitate a holistic approach to understanding the complex interplay between organisms, genes, and their environment. As such, megx.net serves as a fundamental resource in the emerging field of ecosystem biology, and paves the road to a better understanding of the complex responses and adaptations of organisms to environmental change.

Database Access The database and all described resources are freely available at http://www.megx.net/. Continuously updated statistics of the content are available at http://www.megx.net/content. A web feed for news related to megx.net is available at http://www.megx.net/portal/news/. Feedback and comments, the most effective springboard for further improvements, are welcome at http://www.megx.net/portal/contact.html and via email to megx@mpi-bremen.de. Overall, it is important to note that the megx.net website does not fully reflect the content and search functionalities of MegDB. For any specialized data request, contact the corresponding author.

Funding This work was supported by the FP6 EU project MetaFunctions [CT 511784]; the Network of Excellence "Marine Genomics Europe"; and the Max Planck Society.

Acknowledgments We would like to acknowledge Timmy Schweer, Thierry Lombardot, Magdalena Golden, and Laura Sandrine for their valuable input to megx.net, as well as David E. Todd for redesigning the web page.

CHAPTER 3

Enriching public descriptions of marine phages using the MIGS standard: A case study assessing the contextual data frontier

Melissa Beth Duhaime[a,b,*], Renzo Kottmann[a], Dawn Field[c], Frank Oliver Glöckner[a,b,*]

[a]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany
[b]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany
[b]Oxford Centre for Ecology and Hydrology, Mansfield Road, Oxford OX1 3SR, UK
[*]To whom correspondence should be addressed.
Tel: +49 421 2028974; Fax: +49 421 2028580; Email: mduhaime@mpi-bremen.de

Submitted December 1, 2009 *to*
Standards in Genomic Sciences

Keywords: marine phages / contextual data / genome standards / markup language

Abstract: In any sequencing project, the depth of comparative analysis possible is dictated by the degree of available accompanying contextual data. The structure, content, and storage of this contextual data must be standardized to ensure consistent coverage of all sequenced entities and facilitate comparisons. Such standardization efforts are now emerging as a rapidly developing field through a range of grass-roots efforts, most notably by the Genomic Standards Consortium (GSC), which has published the "Minimum Information about Genome/Metagenome Sequences (MIGS/MIMS)" checklist. In the present study, a subset of genomes, the thirty sequenced marine bacteriophages, is collected and its compliance with the MIGS specification is checked. This analysis confirms that current International Nucleotide Sequence Database Collaboration (INSDC) submissions are deficient in most MIGS fields. To counter the missing data, contextual information was manually curated and stored in MIGS-compliant reports for each of the sequenced marine phages. These reports are available in Genomic Contextual Data Markup Language (GCDML) format, where they will persist in structured machine-readable form for future re-use. This case study presents several descriptive exploratory data analyses based on contextual data automatically extracted from the machine-readable reports, including trends in isolation and sequencing dates, distribution of genome sizes and %G+C, phage and host taxonomy, and patterns in sample site environmental parameters. Furthermore, manual curation of data necessary for georeferencing made it possible to put 27 of the 30 phage genomes "on the map", facilitating future analysis based on their environmental context. This work emphasizes how the integration of contextual data at the time of sequencing or database submission will minimize downstream manual curation, maximize integration of data (e.g. environmental parameters) from emerging resources, and streamline biological knowledge generation.

3.1 Introduction

At an estimated 10^{30} viruses in the ocean (Suttle, 2007), viruses are increasingly recognized as the most abundant biological entity on the planet, the majority of which are bacteriophages, viruses that infect bacteria. Some of the first marine phage genomes to be sequenced revealed photosynthesis genes (Sullivan et al., 2005), some of which are not only rampantly transferred between phage and host (Lindell et al., 2004), but are also expressed during host infection (Lindell et al., 2005). Furthermore, analysis of the scaffolds classified as 'viral' from the Global Ocean Survey metagenomic dataset has identified a plethora of host-specific, environmentally significant functional genes, including genes involved in photosynthesis, phosphate stress response, vitamin biosynthesis, antibiotic resistance, and nitrogen fixation (Williamson et al., 2008). Marine virus experts have acknowledged the dire need to link viral genomic data with both biogeochemical contextual data and host sequence data in order to investigate environmentally relevant virus-host systems (Brussaard et al., 2008). This study maximally enriches the contextual data accompanying the currently sequenced marine phage genomes and, in light of compliance with recently proposed genome standards (Field et al., 2008), reinforces the need for persisting information from diverse sources in a single location in a structured form to enable robust comparative genomics.

The power to gain knowledge from any genomic venture depends heavily on the a priori sequence content of public databases with which to compare new sequences to, via, i.e., sequence alignment approaches (Chain et al., 2003). This is the main tenet of comparative genomics. With nothing similar, new sequences can only be labeled as unknown, with no 'handle' by which to base functional or evolutionary hypotheses.

The same 'context-mining' principle extends to sequence-associated contextual data. Sequences can only be grouped by contextual parameters and then interpreted in a comparative context when this data is available and stored in an accurate, structured and accessible fashion. This allows for interpretation in light of other organisms (or communities, in the case of metagenomes), the habitat, isolation location, biological features, the molecular procedures applied to obtain genomic material, sequencing and post-sequencing methods, etc. Given the vast number of sequences already available, these contextual descriptors are becoming as valuable as the very nucleotides that make up the sequences. When present and correct, they expand the number of dimensions available in the realm of comparative genomics and downstream hypothesis testing (Martiny and Field, 2005).

To promote better descriptions of our complete collection of genomes and metagenomes, the Genomic Standards Consortium (GSC) has published the "Minimum Information about Genome/Metagenome Sequences" (MIGS/MIMS) checklist, which recommends a required set of contextual data, e.g., sample site latitude (x), longitude (y), depth (z), and time (t), to accompany all genomic sequence submissions to the public domain (Field et al., 2008). To facilitate the implementation of this standard, and promote the capture, exchange, and downstream comparison of MIGS contextual data, an XML exchange language has also been developed: the Genomic Contextual Data Markup Language (GCDML, Kottmann et al., 2008).

To underscore the importance of consistently integrating contextual and sequence data, this study compares contextual data accompanying marine phage genomes pre-MIGS compliance (INSDC reports) with manually curated information available in a parallel MIGS-compliant data set (GCDML reports) ready for comparative analysis. This is the first concerted effort to bring legacy data up to date with the GSC-recommended MIGS checklist and to (i) determine the effort required to make legacy data comply with the MIGS standard, (ii) determine the degree to which compliance is possible using public annotations and associated literature, and (iii) pave the way for the use of automatically retrievable contextual data in first-pass exploratory analyses to inspire hypothesis generation and subsequent genomic and lab studies.

3.2 Methods

3.2.1 Genomes and contextual data sources: setting up the comparison

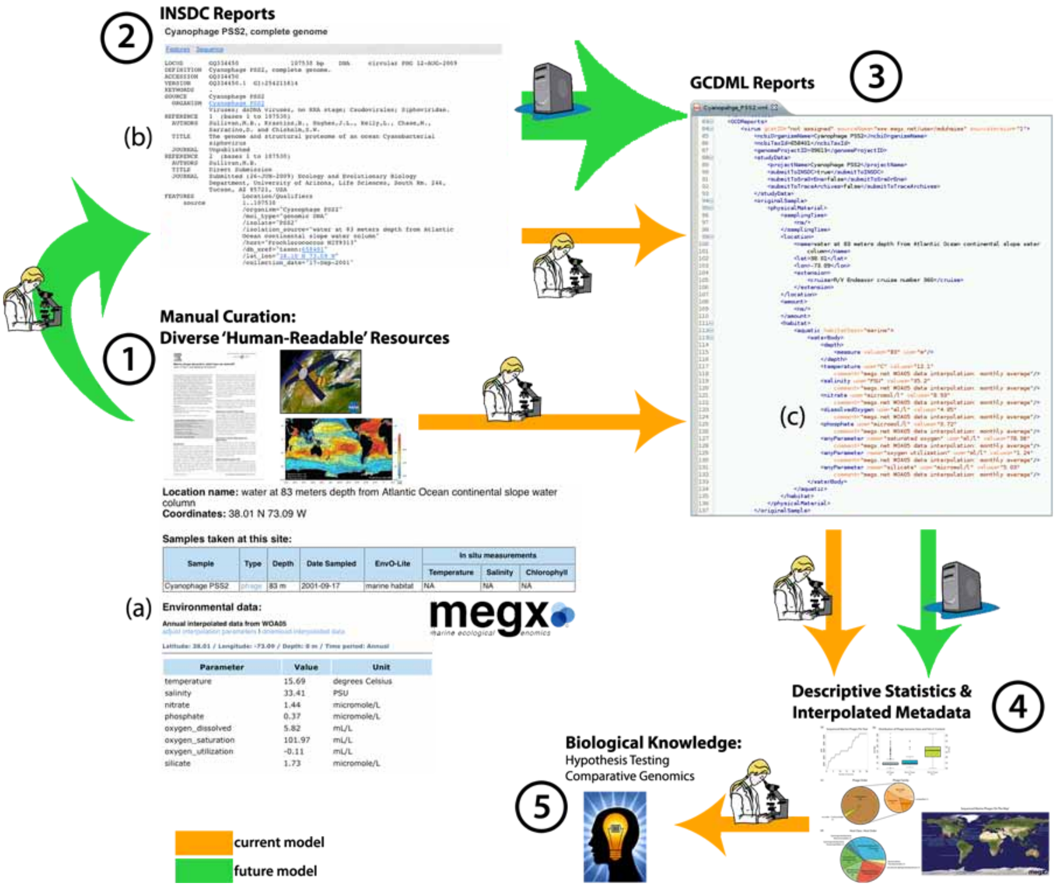
The complete set of all phage genomes reportedly isolated from marine habitats was identified through literature (Paul et al., 2005) and text searches of PubMed. Associated genome files were collected in GenBank format (hereafter referred to as 'INSDC reports', as they represent data stored by the International Nucleotide Sequence Database Collaboration: INSDC) along with publications describing the organism isolation and sequencing. Two datasets were then generated and compared:

1. a set of reports fulfilling MIGS criteria based solely on contextual data available in the structured 'machine-readable' INSDC reports (Figure 3.1 part 2), and
2. a set of reports fulfilling MIGS criteria based on both INSDC reports and manual curation of diverse 'human-readable' resources (Figure 3.1 part 1).

These diverse resources include literature, correspondence with authors, culture collections, and specialized databases, e.g., the Félix d'Hérelle Reference Center for Bacterial Viruses (FHRCBV), a highly curated reference catalog, which bases its taxonomy on morphology evident through their collection of high quality electron microscopy (EM) images of each phage (http://www.phage.ulaval.ca/). Interpolated environmental parameters (temperature, salinity, nitrate, phosphate, dissolved oxygen, oxygen saturation, oxygen utilization, and silicate) describing the organisms' sample sites were also assembled for all possible phage genomes (Table 3.1), using the megx.net GIS Tools (http://www.megx.net/gms/tools/tools.html). This megx.net resource employs oceanographic data from large-scale datasets, such as the World Ocean Atlas (http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html), to interpolate data for single points in the world's oceans at one decimal degree resolution (Kottmann et al., 2010).

3.2.2 MIGS-compliance

MIGS-compliance was achieved through extensive manual curation efforts. Among other essential information, this compliance imposes the critical storage of:



Cyanophage PSS2, complete genome

Features Sequence

```
LOCUS       GQ334450               107530 bp    DNA     circular PHG 12-AUG-2009
DEFINITION  Cyanophage PSS2, complete genome.
ACCESSION   GQ334450
VERSION     GQ334450.1  GI:254211614
KEYWORDS    .
SOURCE      Cyanophage PSS2
  ORGANISM  Cyanophage PSS2
            Viruses; dsDNA viruses, no RNA stage; Caudovirales; Siphoviridae.
REFERENCE   1  (bases 1 to 107530)
  AUTHORS   Sullivan,M.B., Krastins,B., Hughes,J.L., Kelly,L., Chase,M.,
            Sarracino,D. and Chisholm,S.W.
  TITLE     The genome and structural proteome of an ocean Cyanobacterial
            siphovirus
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 107530)
  AUTHORS   Sullivan,M.B.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-JUN-2009) Ecology and Evolutionary Biology
            Department, University of Arizona, Life Sciences, South Rm. 246,
            Tucson, AZ 85721, USA
FEATURES             Location/Qualifiers
     source          1..107530
                     /organism="Cyanophage PSS2"
                     /mol_type="genomic DNA"
                     /isolate="PSS2"
                     /isolation_source="water at 83 meters depth from Atlantic
                     Ocean continental slope water column"
                     /host="Prochlorococcus MIT9313"
                     /db_xref="taxon:658401"
                     /lat_lon="38.01 N 73.09 W"
                     /collection_date="17-Sep-2001"
```

Cyanophage_PSS2.xml

```
<GCDReports>
    <virus gcatID="not assigned" sourceName="www.megx.net/user/mdubhaise" sourceVersion="1">
        <ncbiOrganismName>Cyanophage PSS2</ncbiOrganismName>
        <ncbiTaxId>658401</ncbiTaxId>
        <genomeProjectID>39813</genomeProjectID>
        <studyData>
            <projectName>Cyanophage PSS2</projectName>
            <submitToINSDC>true</submitToINSDC>
            <submitToSraOrEna>false</submitToSraOrEna>
            <submitToTraceArchives>false</submitToTraceArchives>
        </studyData>
        <originalSample>
            <physicalMaterial>
                <samplingTime>
                    <na/>
                </samplingTime>
                <location>
                    <name>water at 83 meters depth from Atlantic Ocean continental slope water
                        column</name>
                    <lat>38.01</lat>
                    <lon>-73.09</lon>
                    <extension>
                        <cruise>R/V Endeavor cruise number 360</cruise>
                    </extension>
                </location>
                <amount>
                    <na/>
                </amount>
                <habitat>
                    <aquatic habitatClass="marine">
                        <waterBody>
                            <depth>
                                <measure value="83" unit="m"/>
                            </depth>
                            <temperature unit="C" value="12.1"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <salinity unit="PSU" value="35.2"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <nitrate unit="micromol/l" value="8.59"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <dissolvedOxygen unit="ml/l" value="4.85"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <phosphate unit="micromol/l" value="0.72"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <anyParameter name="saturated oxygen" unit="ml/l" value="78.38"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <anyParameter name="oxygen utilization" unit="ml/l" value="1.24"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <anyParameter name="silicate" unit="micromol/l" value="5.03"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                        </waterBody>
                    </aquatic>
                </habitat>
            </physicalMaterial>
        </originalSample>
```

Location name: water at 83 meters depth from Atlantic Ocean continental slope water column
Coordinates: 38.01 N 73.09 W

Samples taken at this site:

Sample	Type	Depth	Date Sampled	EnvO-Lite	In situ measurements		
					Temperature	Salinity	Chlorophyll
Cyanophage PSS2	phage	83 m	2001-09-17	marine habitat	NA	NA	NA

Environmental data:

megx marine ecological genomics

Annual interpolated data from WOA05
adjust interpolation parameters | download interpolated data

Latitude: 38.01 / Longitude: -73.09 / Depth: 0 m / Time period: Annual

Parameter	Value	Unit
temperature	15.69	degrees Celsius
salinity	33.41	PSU
nitrate	1.44	micromole/L
phosphate	0.37	micromole/L
oxygen_dissolved	5.82	mL/L
oxygen_saturation	101.97	mL/L
oxygen_utilization	-0.11	mL/L
silicate	1.73	micromole/L

Figure 3.1: Model of flow of contextual data into biological knowledge. (a) screenshot of interpolated data for Cyanophage P-SS2 from megx.net website (b) screenshot of Cyanophage P-SS2 GenBank file, the only INSDC report to store x, y, z, t data, (c) section of GCD report showing GCDML structure, highlighting the storage of cruise information and interpolated data from megx.net GIS tools.

Table 3.1: Phages that have been identified as, or suggested to be, from a marine habitat based on literature searches, including their INSDC accession numbers. Phages that were not isolated from marine habitats are noted and alternatively described according to EnvO-Lite (v1.4). Genomes were assessed as to (i) whether they are linked to enough contextual data to retrieve interpolated data from megx.net GIS Tools (www.megx.net) and (ii) which elements are missing for precise georeferencing (x, y, z) with the most representative interpolated data (which requires t).

NCBI Organism Name	**INSDC identifier**	**Interpolated data?**[I]	**Missing Elements**
Cyanophage P-SS2	GQ334450	Yes	None
Flavobacterium phage 11b[II]	AJ842011	No - insufficient data	x, y, z, t
Halomonas phage phiHAP-1	EU399241	Yes	None
Listonella phage phiHSIC	AY772740	Yes	x, y
Phage phiJL001	AY576273	Yes	x, y
Pseudoalteromonas phage PM2	AF155037	No - insufficient data	x, y, z, t
Prochlorococcus phage P-SSP7	AY939843	Yes	None
Prochlorococcus phage P-SSM2	AY939844	Yes	None
Prochlorococcus phage P-SSM4	AY940168	Yes	None
Roseobacter phage SIO1	AF189021	No - insufficient data	x, y, z
Roseobacter phage SIO1-2001	FJ867910	No - insufficient data	x, y, z, t
Roseobacter phage SBRSIO67-2001	FJ867912	No - insufficient data	x, y, z, t
Roseobacter phage OS-2001	FJ867913	No - insufficient data	x, y, z, t
Roseobacter phage MB-2001	FJ867914	No - insufficient data	x, y, z, t
Silicibacter phage DSS3phi2	FJ591093	No - insufficient data	x, y

Table con't from previous page

NCBI Organism Name	INSDC identifier	Interpolated data?[I]	Missing Elements
Sulfitobacter phage EE36phi1	FJ591094	No - insufficient data	x, y
Synechococcus phage P60	AF338467	No - insufficient data	x, y, z
Synechococcus phage S-PM2	AJ630128	No - insufficient data	t
Synechococcus phage S-RSM4	FM207411	No - insufficient data	x, y, z, t
Synechococcus phage syn9	DQ149023	No - too close to coast	x, y, t
Synechococcus phage Syn5	EF372997	Yes	t
Vibrio phage VP2	AY505112	No - insufficient data	x, y, z, t
Vibriophage VP4	DQ029335	No - insufficient data	x, y, z, t
Vibrio phage VP5	AY510084	No - insufficient data	x, y, z, t
Vibrio phage VP16T	AY328852	No - too close to coast	x, y, t
Vibrio phage VP16C	AY328853	No, *t*oo close to coast	x, y, t
Vibriophage VpV262	AY095314	No - insufficient data	x, y, z, t
Vibrio phage VHML[III]	AY133112	No - insufficient data	x, y, z, t
Vibrio phage KVP40	AY283928	No - insufficient data	x, y, z, t
Vibrio phage K139[IV]	AF125163	No - insufficient data	x, y, z, t

[I] this can be as minimal as a "fuzzy" habitat descriptor (rather than precise x, y), requires a depth (or 'surface sample' description), and does not require a date (as yearly averages can be taken). However, if the sample site is too close to the shore, data interpolation is not possible.

[II] isolated from sea ice (aquatic habitat)

[III] isolated from aquacultured shrimp (organism-associated habitat)

[IV] isolated from human (organism-associated habitat)

1. data describing the isolation location, depth, and time (x, y, z, t), frequently overlooked parameters (editorial, 2008), but ones which allow for more comprehensive analysis (Field, 2008), such as proper annotation of the organism's habitat and post hoc interpolation of environmental data describing the habitat (Kottmann et al., 2010),

2. and detailed biological parameters, namely host range and expert phage taxonomy.

Compliance to the 'habitat' descriptor of MIGS was achieved using terms from the EnvO-Lite (v1.4) controlled vocabulary (Hirschman et al., 2008). There is currently no structured grammar by which INSDC reports define habitat in a 'machine-readable' format. However, for the purpose of this study, when the INSDC location name contained a known marine habitat, the phage was labeled as 'marine' according to INSDC.

3.2.3 GCDML files

MIGS-compliant reports were rendered in GCDML, version 1.7 (Figure 3.1 part 3, Supplementary Figure 7.3) (Kottmann et al., 2008). GCDML reports were manually created using the oXygen XML editor (version 11). GCDML reports consist of two core components: (i) Minimal Information about a Genome Sequence (MIGS) reports, which implement the GSC checklist by strictly requiring components of the MIGS list for report validation, and (ii) Genomic Contextual Data (GCD) reports, which contain all required components of the MIGS reports, as well as flexible, yet structured, extensions for additional contextual data (Figure 3.1 part 3c, Supplementary Figure 7.3). In this study, these extensions allowed for consistent storage of genome size and %G+C content, latitude and longitude for 'manually determined' locations based on verbose geographic descriptors (rather than precise numeric reports), cruise ship name and number (allowing coordination with other samples collected on this cruise), and environmental metadata, either collected in situ or interpolated using, i.e., megx.net GIS tools (Figure 3.1 part 1a) (http://www.megx.net/gms/tools/tools.html). All GCDML files are available at http://www.megx.net/genomes/viruses/phages/phages.html.

3.2.4 Exploratory contextual data analyses

Once all contextual data was collected and made available in fully 'machine-readable' GCDML reports, exploratory data analyses were performed to describe features of the marine sequenced phages, their genomes and their sample sites (Figure 3.1

part 4). Data describing all phages (size and taxonomy) were extracted from their respective GenBank files from NCBI (19 November 2009) with Perl scripts. A dendrogram clustering phages by sample site physical-chemical parameters (salinity, nitrate, dissolved oxygen, phosphate, oxygen saturation, oxygen utilization, and silicate) was derived from a distance matrix (Euclidean distance coefficient) of z-score transformed data using average linkage clustering. Phages were displayed on the megx.net map (Kottmann et al., 2010) using its integrated Web Map Service technology (Supplementary Table 7.2).

3.3 Results

3.3.1 A comparison of INSDC and manually curated GCDML reports

Surveying the literature and the public databases revealed a set of 27 phages isolated from a ′marine′ habitat (Table 3.1). Figure 3.2 compares the number of MIGS-compliant fields fulfilled by INSDC documents to those fulfilled after manual curation of the literature and other resources. Nearly half of the fields examined held no information in INSDC reports (especially pertaining to documentation of ′Sequencing′ components), but following curation this rose to one hundred percent compliance, in that the field was no longer empty (Figure 3.2). Unknown MIGS fields are filled with either an ′inapplicable′ or missing′ qualifier; this acknowledges the presence/absence of this information and therefore is more valuable than its complete absence from the report (Figure 3.2).

Overall, when the minimum required date resolution is ′year′, only 21% of the components recommended by the MIGS checklist are reported in the current marine phage INSDC reports (Figure 3.2). Intensive manual curation, which is informative but not practical for large legacy datasets, was able to satisfy 66% of all MIGS components when all ′mappable′ locations are considered (63% when only precise reports of latitude and longitude were considered). Of the unknown components of the GCDML reports that still evaded manual curation (34%), 25% are due to fields deemed ′inapplicable′ for phages (though still apply to other virus types), such as ′Subspecific genetic lineage′ and ′Health or disease status of host′, while the remaining 75% is unknown due to missing information. Of the manually curated data, 1% of the fields could be confirmed only through personal communication with authors (e.g., to confirm habitat) or other experts in the field (e.g., to confirm taxonomy). Below, compliance with core MIGS descriptors is further described and discussed.

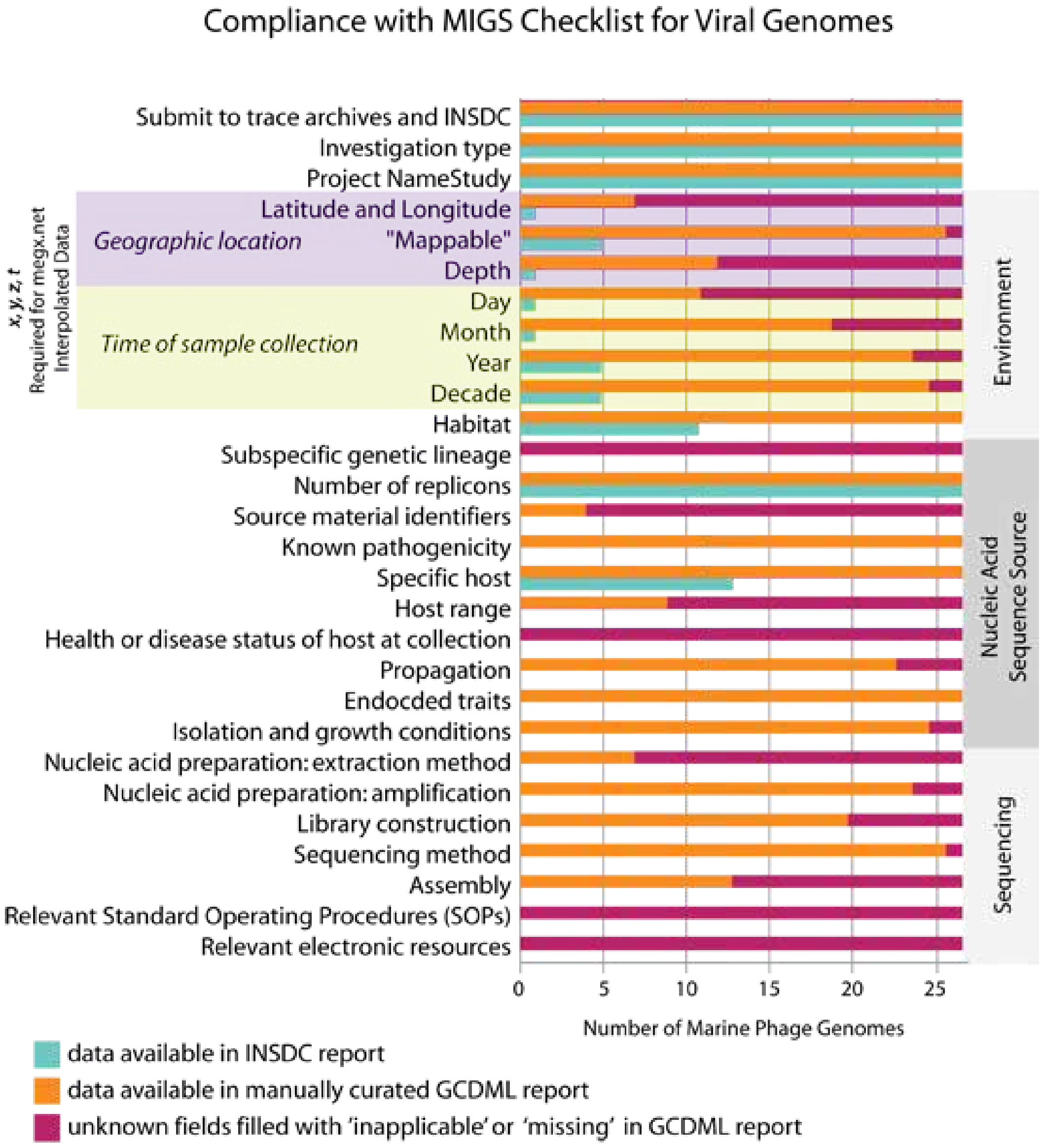


Figure 3.2: Comparison of compliance with viral components of the MIGS checklist between data available in INSDC reports and that in MIGS/GDC reports that have been supplemented with extensive manual curation. List modified from (Field et al., 2008).

3.3.1.1 Habitat.

An essential piece of information about any genome is the habitat from which the genome (i.e., organism or sample) originated. To date, this information has not been captured systematically in public databases, yet is core to the MIGS specification due to its biological importance (Field, 2008; Hirschman et al., 2008). Information in INSDC files made it possible to classify 41% of the phages as ’marine’, meaning isolated from "A habitat that is in or on a sea or ocean containing high concentrations of dissolved salts and other total dissolved solids (typically >35 grams dissolved salts per litre)" (per Envo-Lite v1.4).

Following manual curation, three of the phages still could not be classified definitively as marine: Vibrio phage K139, Vibrio phage VHML, and Flavobacterium phage 11b (Table 3.1). The vibriophages are now annotated as ’organism-associated’, having originated from "A habitat that is in or on a living thing" (per Envo-Lite v1.4). Kapfhammer et al. report that Vibrio phage K139 was isolated from its host lysogen, *Vibrio cholerae* O139 strain M010 (Kapfhammer et al., 2002), which is a clinical strain isolated in 1992 from the tenth *V. cholerae* O139 victim in Madras, India (Matthew Waldor, personal communication). *Vibrio phage* VHML was isolated from its host lysogen cultured from prawn larvae (*Penaeus monodon*) from an aquaculture pond in Australia (Oakey and Owens, 2000). Flavobacterium phage 11b is now reported as ’aquatic’, originating from "A habitat that is in or on water" (Envo-Lite v1.4). This phage was isolated from melted Arctic sea ice, a term which itself can not be classified as definitively marine, as sea ice has variable salinity (the distinguishing element between ’marine’ and ’aquatic’) depending on the ice growth stage or local structure, i.e., high-salinity brine chamber or low-salinity melt pool. In all, habitat curation (guided by an accepted habitat ontology) resulted in 27 ’marine genomes’, which are those considered in the remaining analyses.

3.3.1.2 Bringing Molecules to the Map.

In addition to habitat, georeferencing of sample collection sites is increasingly essential. Unsurprisingly (editorial, 2008; Field, 2008), only a single marine phage, Cyanophage P-SS2, contained sufficient latitude, longitude, and depth data (x, y, and z) in the INSDC report place conclusively on a map (Figure 3.1 part 2b). This was also the only INSDC report to contain depth.

After manual curation, precise x and y coordinates were determined for only seven (26%) of the genomes. However, all but one phage (96%) were ’mappable’, in that they described imprecise sample site descriptors, such as ’Scripps Pier, La

Jolla California, USA' (Figures 3.2 and 3.6). Depth could be added to 12 (44%); most manually curated depths were due to literature reports of "surface samples", rather than exact depth measurements and reports. The union of x, y, z, and t (time) allows for extraction of interpolated environmental parameters; after manual curation, this data was available for only 11 (41%) of the phage genomes using megx.net GIS tools (http://www.megx.net; Table 3.1). However, due to the inaccuracy of environmental data interpolation near land, the three sample sites too close to the coast are missing this data (Table 3.1).

3.3.1.3 Phage Host Range and Taxonomy.

Information on host-range and host taxonomy provides essential information on the biological and ecological impact of phages. INSDC reports stored information about host taxonomy in 48% of the reports. Information regarding host range was completely lacking from all INSDC reports.

After manual curation, information about host taxonomy was expanded to 100% through manual curation ('Specific Host', Figure 3.2) and alternate hosts were manually determined for nine (33%) phages ('Host Range', Figure 3.2). The phage taxonomies documented in INSDC reports were compared to taxonomies documented in the phage isolation and sequencing publications, as well as to the Félix d'Hérelle Reference Center for Bacterial Viruses (FHRCBV). When conflicts occur, the FHRCBV is considered the expert taxonomy. For instance, Vibrio phage VP5 (NCBI taxid: 260827) is classified as Podovirdae in its INSDC report, whereas, according to its long non-contractile tail evident in the EM image in FHRCBV (accession: HER 169), it has been expertly classified as Siphoviridae (Sylvain Moineau, personal communication).

3.3.2 Exploratory Data Analysis and Interpolated Metadata

Once a full set of MIGS-compliant records was compiled by merging INSDC information with manually curated data, all information was place into 'machine readable' GCDML reports for each genome. This allowed for the automatic extraction of information for exploratory data analysis used to reveal patterns in the contextual data of the sequenced marine phages. This included trends in the number of isolated and submitted genomes per year for sequenced marine phages (Figure 3.3a), their range and distribution of genome size and %G+C content (Figure 3.3b), the taxonomic classification of all phages versus marine phages (Figure 3.4a), the marine hosts (Figure 3.4b), and the sequenced 'genome pairs' (Figure 3.5).

Marine phages tend to be larger than phages sequenced from other habitats

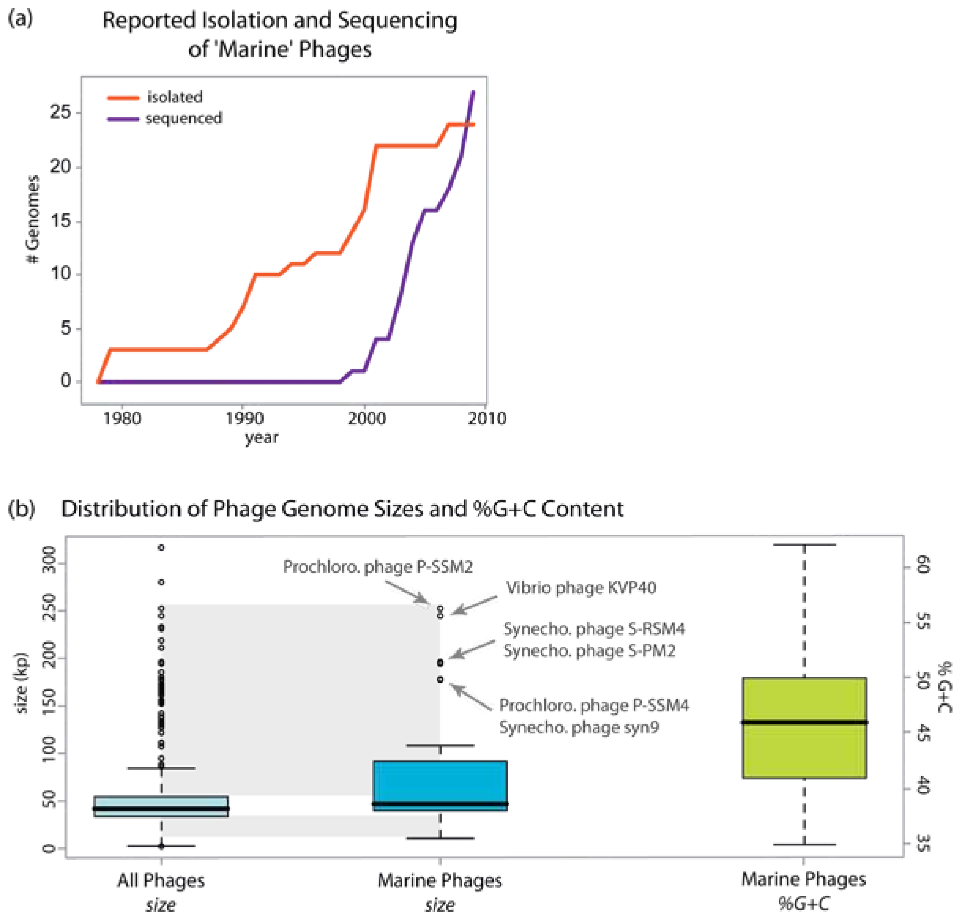


Figure 3.3: Overview of marine phage isolation/sequencing year and genome properties stored in GCDML reports. (a) Trends of isolation and sequencing of the sequenced 'marine' phages over the last two decades. (b) Box and whisker plots showing range and distribution of genome sizes for all versus marine phages and %G+C content for marine phages. The box shows the interquartile range (middle 50% of the data); the thick black line demarcates the median, the dotted line extends to the minimum and maximum values; outliers are shown by empty circles. Data for genome sizes of "All Phages" were retrieved from NCBI.

(Figure 3.3b). Based on genome size, one-third of the sequenced marine phages are in the 75th percentile of all sequenced phages (Figure 3.3).

The taxonomic diversity of sequenced marine phages is quite low as compared to the diversity of the sequenced phages from all habitats (Figure 3.4). Of the 27 marine phages sequenced, all are double-stranded DNA phages, with no RNA stage; 96% are of the viral order *Caudovirales* (Pseudoalteromonas phage PM2 has an unclassified order and belongs to the *Corticoviridae* family), as opposed to 76% of all sequenced *Caudovirales* phages (123 phages with no order span 13 different Classes).

The distribution of their hosts is also biased (Figures 3.4 and 3.5). Two thirds of the sequenced marine phages infect *Proteobacteria*. Furthermore, most hosts are restricted to three major sets; 30% infect *Vibrio* spp. (likely due to interest in pathogenicity), 33% infect *Cyanobacteria* (either *Chroococcales* or *Prochlorales*), and another 30% infect Alphaproteobacteria (all but one infect *Rhodobacterales*; Figure 3.4b).

Additionally, the 27 'mappable' genomes can be further analyzed in their environmental context using emerging resources, such as megx.net, to (i) 'put them on the map' (Figure 3.6a; http://www.megx.net), and (ii) extract interpolated environmental data, though only possible for the eight genomes where depth is reported and which are not too close to the coast (Table 1). Preliminary analysis of the megx.net interpolated data available in the GCD reports revealed that, based on physical-chemical parameters across sample sites, e.g., the four phages isolated from the Sargasso Sea cluster together, while Cyanophage P-SS2 appears to be an outlier (Figure 3.6b). Further examination of the range and distribution of each parameter show the Cyanophage P-SS2 sample site to have quite distinct interpolated nitrate, phosphate, and dissolved oxygen values (Figure 3.6c).

3.4 Discussion

The manual curation and generation of MIGS-compliant GCDML reports for 30 phage genomes have transferred extensive amounts of contextual data from human-to machine- readable form, making it now programmatically accessible and ensuring it is explicitly linked to its original sequenced entity (Figures 3.1 and 3.2). This is the first study to (i) assess the realistic implementation of recently imposed contextual data standards (ii) assess the ability of current technologies to manage and store contextual data, and (iii) establish ecogenomic trends within the sequenced marine phage genome collection using solely contextual data, following up on similar ecological considerations and biological patterns observed by Martiny and

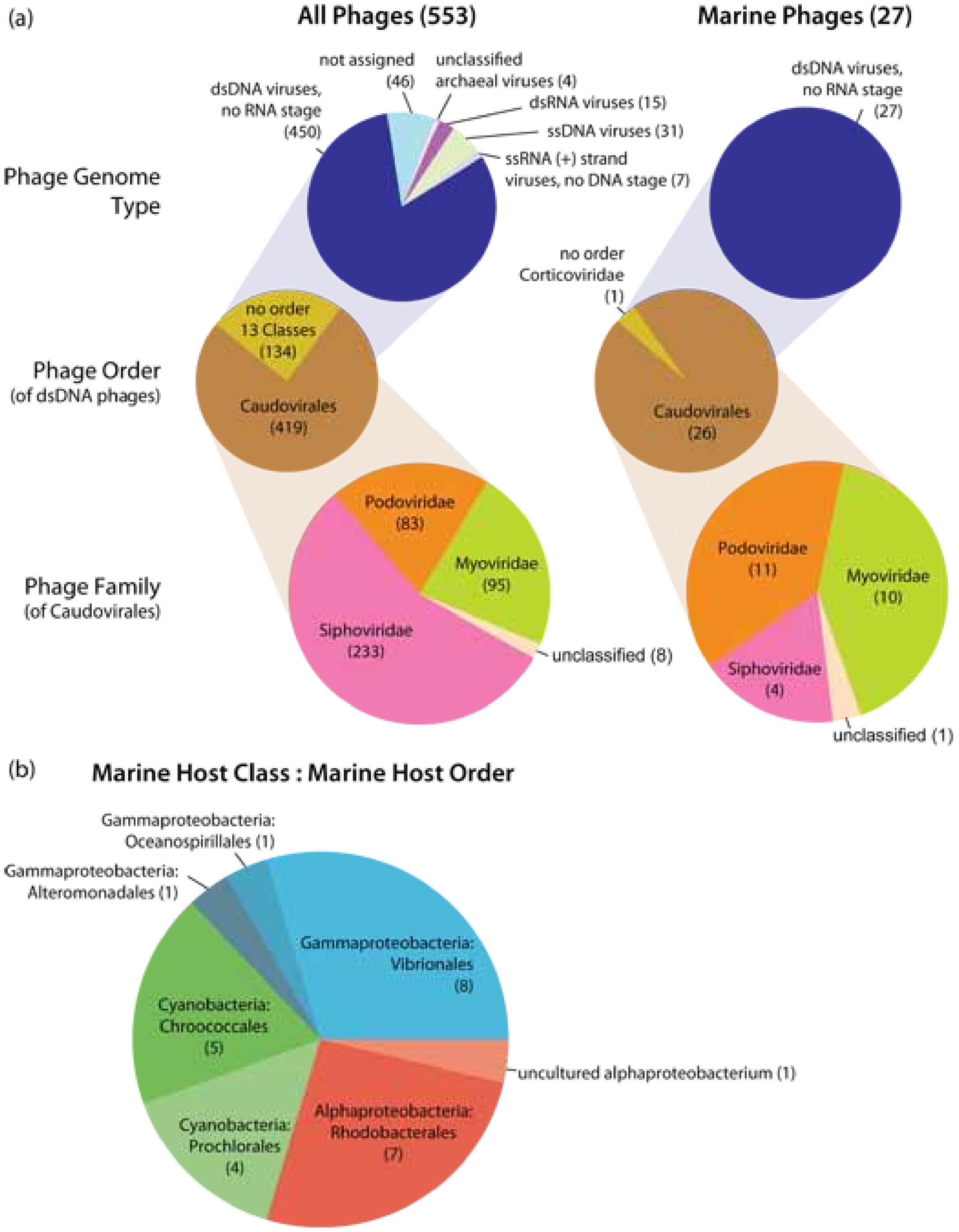


Figure 3.4: Overview of phage taxonomic data. (a) The taxonomic distribution of all sequenced phages versus all sequenced marine phages and (b) the hosts of all sequenced marine phages. All information describing marine phages and their hosts is accessible via GCDML reports.

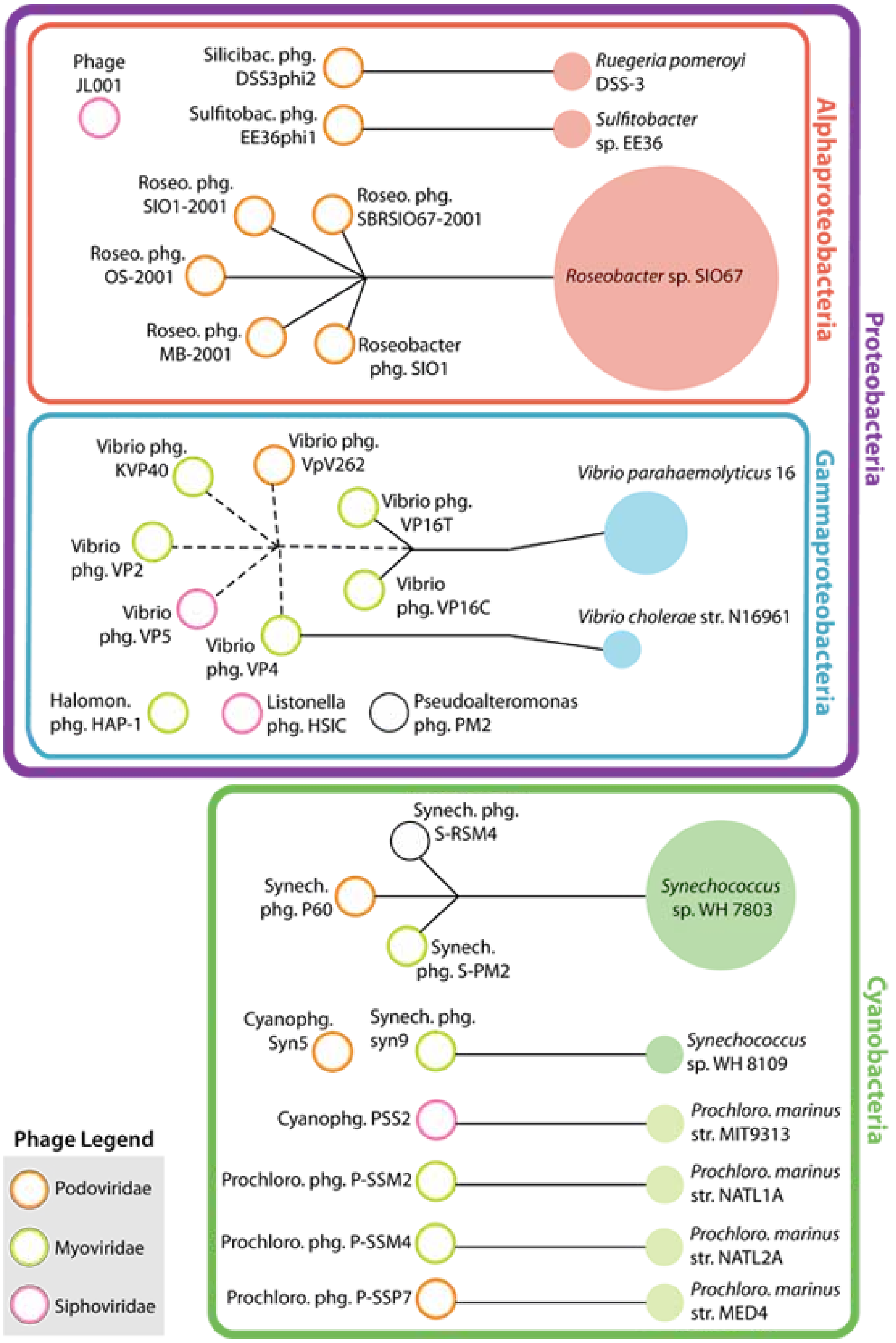


Figure 3.5: Network of ′genome pairs′ and interactions between sequenced marine phages and sequenced hosts. Solid lines link phages (empty circles) to the host strain (solid circles) they infect; dashed lines connect phages to the host species (but not necessarily strain) they infect. Phages with no sequenced host are grouped by host Class (or Subclass for *Cyanobacteria*). Phage taxonomy is reflected by the color of the empty phage circle. Number of phages infecting a sequenced host is reflected by the size of the solid host circles.

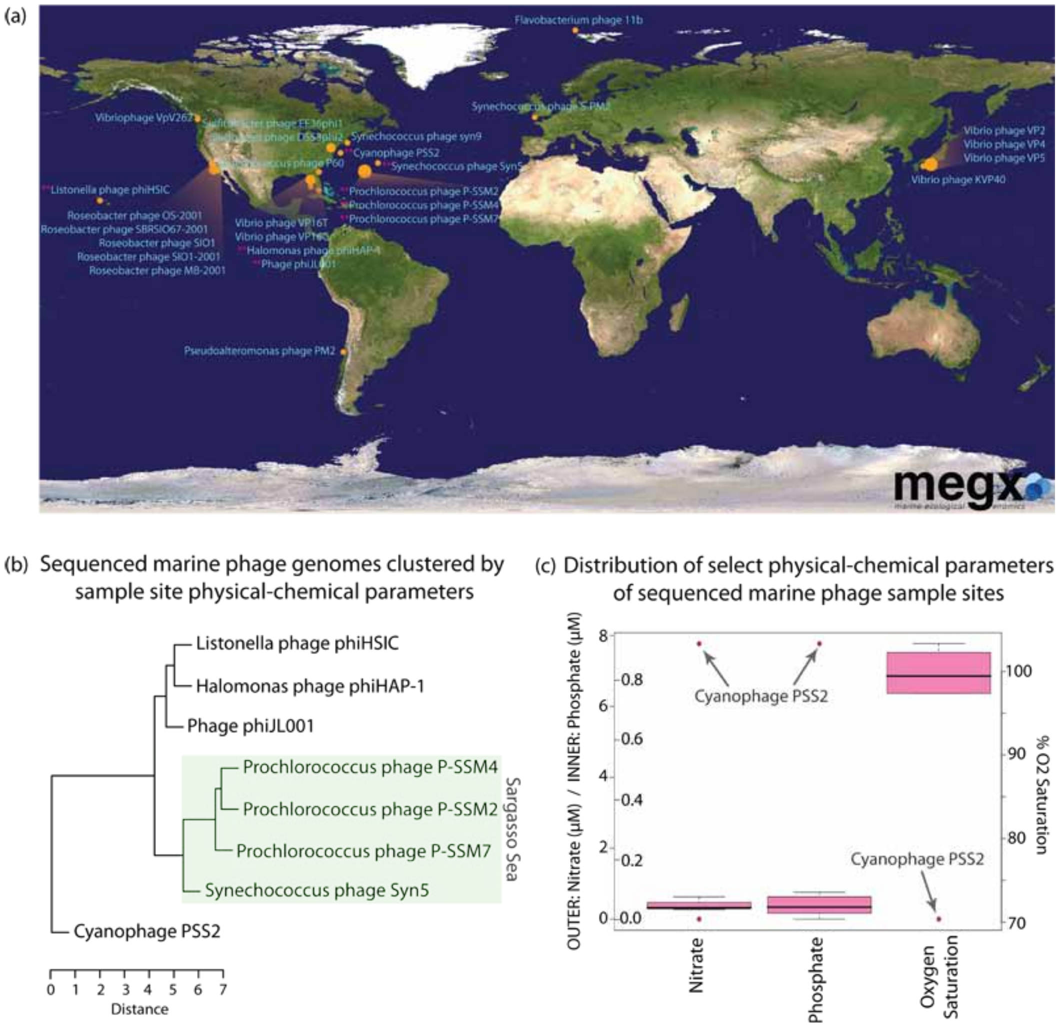


Figure 3.6: (a) The 26 ’marine’ phage genomes (plus ’aquatic’ Flavobacterium phage 11b) able to be mapped based on data in their GCDML reports. The map is modified from that available from megx.net (see Supplementary Information 7.2 for exact webserver query); (b) sample sites of marine phages clustered by interpolated environmental data; (c) distribution of three of the interpolated environmental parameters (nitrate, phosphate, and oxygen saturation) demonstrating the Cyanophage P-SS2 outlier.

Field (2005) in their landmark study of contextual data describing the first several hundred genomes sequenced.

3.4.1 Towards consistency and persistence of contextual data

It was difficult, often impossible, to backtrack and generate MIGS-compliant reports using legacy data, while it will be absolutely feasible if MIGS specifications are stored at the time of sequencing and submission in the future.

This study found the most overlooked components to be sample site location (x, y, z), sample collection date (t), host range, and whether the organism exists in a culture collection (Figure 3.2). Notably, nearly all of the 'Sequencing' components (Figure 3.2) are missing or filled with a 'not available' placeholder in the MIGS report; in a world of rapidly evolving technologies, this component is critical in order to evaluate quality measures of the sequence data as techniques change through time.

Implementing standards, such as those of the GSC, is thus an invaluable means to encourage sequence submitters to carry contextual data over to the public databases. As nearly 60% of the data missing from INSDC reports can be supplemented by manual curation (Figure 3.2), it is not the case that this data is too difficult to collect or that MIGS is not possible to comply with. Rather, there simply needs to be a structured and inviting system in place for its storage.

As a case study of how pertinent biological information can be lost, consider the Vibrio phages VP2, VP4, and VP5. The first element of inconsistency arises with taxonomy. All three phages are reported as belonging to the Podoviridae in their INSDC genome reports. However, according to the Félix d'Hérelle Reference Center for Bacterial Viruses, VP5 belongs to the Siphoviridae (as confirmed by expert electron micrography), and VP2 and VP4 are described, with accompanying EM images, as myoviruses by Koga et al. (1982) in the description of their initial isolation. Furthermore, the INSDC reports for Vibrio phages VP2, VP4, and VP5 report their host as *Vibrio cholerae*. This may be true for the phages used in the sequencing project in 2003 (though this can not be confirmed, as their genomes were directly submitted with no accompanying publication), however the phages were reportedly collected from seawater near Tokushima, Japan and isolated on *Vibrio parahaemolyticus* in 1982 (Koga et al., 1982). If different hosts were used in isolation and sequencing, were genetically identical phages propagated in both cases? Without manually effort, and now electronic persistence in GCDML, it could be soon forgotten that these phages were isolated on different hosts. These issues further emphasize the need for (i) more explicit MIGS-compliant comment fields in INSDC reports to differentiate ’isolation hosts’ from ’sequencing hosts’, and (ii) i.e., ’Short Genome Reports’ in journals such as the Standards in Genomic Sciences to verbosely explain the biological history of sequenced entities.

3.4.2 Biological lessons learned

Through these manual curation efforts, we begin to appreciate where our gaps in biological knowledge exist, and how they may be limiting our ability to establish accurate "rules and exceptions" (Martiny and Field, 2005) to describe the impact of viruses in the marine realm.

3.4.2.1 Large marine phage genomes

Genome size has been implicated as being diagnostic of biological properties of the phage; size is directly correlated with virion complexity and interference with host cellular activities (Brüssow and Hendrix, 2002). As more are sequenced, an emerging property of marine phage genomes is that they are among the largest known (Sullivan et al., 2005, 2009) (Figure 3.3b). A closer look at the gene content of marine versus non-marine phages could suggest whether this size is due to the great number of host-related genes carried by marine phages (Lindell et al., 2004; Mann et al., 2005, 2003; Sullivan et al., 2005), or some other underlying evolutionary process.

3.4.2.2 Biases in represented phage and host types

Among all sequenced phages, there is general bias towards double-stranded DNA (dsDNA) viruses lacking an RNA stage (possibly influenced by, e.g., cloning biases in sequencing efforts, chloroform extractions that disrupt lipid-membranes of, i.e., dsRNA viruses, the difficulty in culturing archaeal hosts, etc.), despite the fact that, from an epidemiological perspective, over 75% of all viral diseases are the result of RNA viruses (Makeyev and Bamford, 2004). Consider the intriguing biology of five Pseudomonas phages (known to infect plant pathogens Mindich, 1988), all of which belong to the dsRNA genome group, yet to be represented by any sequenced marine phages. These odd dsRNA phages have segmented genomes, whereby three ’chromosomes’ exist in each virion and are often reassorted during co-infection of the same host (Abedon, 2008), where phages can exist in a ’carrier state’, reproducing without killing their host (Onodera et al., 1992). This feature, combined with the intrinsically low fidelity of RNA replication due to the apparent lack of proof-reading in RNA-dependent polymerases, allows for RNA viruses to rapidly adapt to new environments, offering insights into modeling of viral population genetics and evolutionary theory we can not yet consider in the marine realm (Makeyev and Bamford, 2004). ssDNA phages are also one of the major 'odd' phages groups not yet represented in the marine phage genome collection (Figure 3.4a), and are also under selective pressure quite unique from their dsDNA counterparts (Xia and Yuen, 2005).

However, the greatest bias is the incredibly restricted host taxonomic distribution. All sequenced marine phages infect only two of the twenty-four Bacteria phyla (*Proteobacteria* and *Cyanobacteria*) and no Archaea (Figure 3.4b). Of these, only four families are represented, which also reflects metabolic/niche biases towards interest in: pathogenicity (namely phages of *Vibrio parahaemolyticus* infecting the *Vibrionales*), marine phototrophs (*Chroococcales* and *Prochlorales*), and ubiquitous coastal microbes essential to global carbon and sulfur cycles (*Rhodobacterales*) (Wagner-Dőbler and Biebl, 2006). A similar pattern of habitat-driven taxonomic bias was seen in the first ecogenomic survey of sequenced microbial genomes, whereby 67% of the sequenced marine microbes were phototrophs (Martiny and Field, 2005).

3.4.2.3 Genome Pairs

The study of phages and hosts intrinsically lends itself to taking advantage of what Martiny and Field describe as "one of the most exciting and underutilized aspects of the genome collection" (Martiny and Field, 2005): genome pairs. A genome pair occurs when organisms with potential natural interactions are both sequenced, e.g., a phage and host. These associations have revealed patterns in genome biology, such as how well pairs correlate based on GC content or tetranucleotide genome signatures (Martiny and Field, 2005; Pride et al., 2006). Such pairs can (and soon will) rapidly evolve to complex networks as multiple phages infecting the same host are sequenced, or multiple hosts infected by the same phage are sequenced. This complexity obviates the need for the basic units, the pairs, to be explicitly documented (as called for by MIGS) in a structured form. This is possible through the GCDML 'original host' and 'alternate host' fields, where they can be stored for automated retrieval and network visualization. This process was just barely possible by hand with the 27 marine phage genomes, and reveals interesting trends (Figure 3.5). Thus far, most cyanophage-cyanobacteria associations are one-to-one pairs, though many cyanophages are known with broad host ranges (Sullivan et al., 2003). Furthermore, such visualization leads to hypotheses about the 'lone phages', such as Phage phiJL001, Halomonas phage HAP-1, and

Cyanophage Syn5, which lack a sequenced host, but which exist in phylogenetic groups with related sequenced hosts (Figure 3.5). The current map is useful in designing future sequencing ventures to answer targeted questions, such as "What drives phage host range and what are the genomic consequences of all members belonging to the same network?"

3.4.2.4 The added value of interpolated environmental parameters comes at a price: *x*, *y*, *z*, *t*

The lack of explicit sample site geographic location and time (x, y, z, t) is apparent (Figure 3.2), though, for environmental isolates, this may be the most 'value-added' component of MIGS compliance. These elements allow for genomes to be "put on the map" (Field, 2008), thus reaping the benefits of, for example, comparisons using environmental data, either collected *in situ*, or interpolated using, i.e., the megx.net GIS Tools (Kottmann et al., 2010).

Using megx.net's emerging resources, any sample site in the ocean where x, y, z, and t are known can be supplemented by interpolated environmental data, such as temperature, salinity, phosphate, silicate, nitrate, dissolved oxygen, Apparent Oxygen Utilization (AOU), oxygen saturation, chlorophyll, etc., at standard depth levels for various time periods (Kottmann et al., 2010). The most resolved time unit for this interpolated data is by month, but if precise enough time is not known, average values at varying resolutions (seasonally, yearly, etc.) can be substituted. Furthermore, georeferenced genomes can be viewed in their environmental context on a world map (Figure 3.6a), and can be overlaid on numerous map data layers, such as nitrate, phosphate, silicate, and chlorophyll, or the environmental stability (expressed as standard deviations) of a parameter. Having such environmental data easily accessible and integrated with sequenced entities via GCDML reports allows for a rapid, automated "first pass" evaluation of environmental/ecological clusters and outliers (Figure 3.6b-c). This process greatly facilitates hypothesis and research question generation, such as: "what are the functional implications of Cyanophage P-SS2[1] being isolated from such a comparatively high nutrient, low oxygen site?" "What genomic features might isolates from similar habitats, such as the Sargasso Sea cluster, share?" Having such data accessible narrows down the search time and space as researchers design comparative, and even laboratory, studies. This clearly speaks to the power of coupling sequence, contextual, and environmental data, further accentuating the need to

[1] In Chapter 5, this question is taken to the lab, as infection dynamics of P-SS2 and its host, *Prochlorococcus* MIT9313 dynamics are investigated under varying nutrient conditions.

keep them stringently linked and stored at our 'digital fingertips.'

3.4.3 GDCML

This is the first successful implementation of GCDML (Kottmann et al., 2008). The curation of real biological data has tested the pre-release version of this XML language, and has helped guide the development of GCDML through v1.7, confirming its ease of use and strength in validating MIGS compliance. Furthermore, all marine phage GCDML reports are hosted by the database of the megx.net marine ecological genomics portal (Kottmann et al., 2010). This is the first database specifically designed to store all contextual data recommended by the GSC, representing a gold standard in the emerging first generation of contextual data databases. The marine phage GCDML reports are the first set of fully MIGS compliant data to be stored there, thus validating the megx.net MIGS-compliant structure.

3.4.4 Outlook

The degree of in-depth manual curation of legacy data presented here is possible with 30 genomes ('current model', Figure 3.1). Using GCDML to generate MIGS compliant reports was straightforward and easy to learn. Should all information be readily accessible, it would take no more than 30 minutes per report. However, the manual collection and validation of the diverse contextual data took hours to days per report. Thus, considering the massive influx of marine virus genomic data expected in the next year (200 marine virus genomes and 50 metagenomes; http://www.broadinstitute.org/annotation/viral/Phage), the overall increase in sequenced genomes, and the increasing number of sequences submitted without an accompanying manuscript, this manual component is unsustainable. Contextual data needs to be stored at the time of sequence submission in a machine-readable form, independent of publications ('future model', Figure 3.1).

As such, we need to work towards an automated process. This comes by (i) recognizing and appreciating the need for storing contextual data, (ii) designing and implementing a centralized system to address the need, (iii) using it. The seed has been sowed for recognition, but it relies on positive feedback; the more contextual data that exists, the more the end-users appreciate its value for their own comparative work, and the more willing they are to collect and submit their own data. The design and implementation of a standardized system have also taken great leaps forward in the last years (Field et al., 2008). Computer scientists and database specialists have designed the tools to consistently store this information in 'machine readable' form (Hirschman et al., 2008; Kottmann et al., 2010). The final steps for its more universal use, already underway, include the partnership between expert and normal users, computer scientists, and the world's major sequence databases to integrate these required contextual elements in the already established databases in the form of, i.e., 'structured comments' (unpublished, NCBI), such that they will be permanently stored with their sequenced entity. This approach can be complemented by proposed efforts, namely the Genomes and Metagenomes (GEM) Catalogue, to centralize the submission, editing, browsing and storage of a rich set of metadata describing the complete genome and metagenome collection, bringing us to new frontiers in contextual data management

CHAPTER 4

Ecogenomics and Genome Landscapes of Marine Pseudoalteromonas Phage H105/1

Melissa Beth Duhaime[a,b,*], Antje Wichels[c], Jost Waldmann[a], Hanno Teeling[a], Frank Oliver Glöckner[a,b,*]

[a]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

[b]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

[c]Alfred Wegener Institute for Polar and Marine Research, Kurpromenade, D-27498 Helgoland, Germany

[*]To whom correspondence should be addressed.

Tel: +49 421 2028974; Fax: +49 421 2028580; Email: mduhaime@mpi-bremen.de

Submitted January 14, 2010

Keywords: ecogenomics / genome signatures / genomics / marine / phage / *Pseudoalteromonas*

Abstract: Marine phages have an astounding global abundance and ecological impact. However, little knowledge is derived from phage genomes, as most of the open reading frames (ORFs) in their small genomes are unknown, novel proteins. To infer potential functional and ecological relevance of sequenced marine Pseudoalteromonas phage H105/1, two strategies were applied. First, similarity searches were extended to include six viral and bacterial metagenomes paired with their respective environmental contextual data. This approach revealed "ecogenomic" patterns of Pseudoalteromonas phage H105/1, such as its estuarine origin and temperate replication strategy. Second, intrinsic genome signatures (phylogenetic, codon adaptation, and tetranucleotide frequencies) were evaluated to shed light on the evolution of the phage functional modules. Based on differen-

tial codon adaptation of Phage H105/1 proteins to the sequenced *Pseudoalteromonas* spp., regions of the phage genome with the most 'host'-adapted proteins have the strongest 'bacterial' tetranucleotide signature; whereas the least 'host'-adapted proteins have the strongest 'phage' tetranucleotide signature. Such a pattern may reflect the evolutionary history of the respective phage proteins and functional modules. Finally, structural proteomics identified seven proteins that make up the mature virion, four of which were previously unknown. This integrated approach combines both novel and classical strategies and serves as a model to elucidate ecological inferences and evolutionary relationships from phage genomes that typically abound with unknown gene content.

4.1 Introduction

Viruses are the most abundant biological entity and largest source of genetic material on the planet (Suttle, 2007), and are likely the major vehicle for gene transfer in the ocean. Considering the global volume of seawater, the worldwide abundance of marine phages and bacteria, and the frequency of gene transfers per infection, virus mediated transfers occur up to 10^{15} times per second in the ocean (Bushman, 2002), with an extrapolated 10^{28} bp of DNA transduced by phages per year (Paul et al., 2002). Evidence shows that these transfers include host-derived metabolic genes central to the metabolism of the world's oceans (Lindell et al., 2004), carried by the virus and expressed during infection(Lindell et al., 2005).

Pseudoalteromonads are ubiquitous heterotrophic members of marine bacterial communities, which, as with most microbial life, are ecologically and evolutionarily influenced by phages (Médigue et al., 2005; Mãnnistő et al., 1999; Moebus, 1992; Thomas et al., 2008; Wichels et al., 1998, 2002). Of the three sequenced Pseudoalteromonads, all contain integrated prophages, two of which are dominated by P2-like myovirus proteins (Prophinder, Lima-Mendez et al., 2008a). The only sequenced Pseudoalteromonas viral isolate, Pseudoalteromonas phage PM2, is a lytic marine phage and the only member of the Corticoviridae (Mãnnistő et al., 1999). Pseudoalteromonas phage H105/1, the focus of the present study, is a member of the Siphoviridae family isolated off Helgoland, in the North Sea, on *Pseudoalteromonas* sp. H105 (Figure 4.1). Phage H105/1 also infects *Pseudoalteromonas* spp. H103 and H108, which were isolated with H105 from the same water sample (Wichels et al., 1998). The host, *Pseudoalteromonas* sp. H105, is susceptible to lysis/growth inhibition by other Helgoland and North Sea phages, including members of both the Myovridae and Siphoviridae families (Wichels et al., 1998, 2002).

Phage genomes are small (3-300 kb) as compared to the Bacteria and Archaea

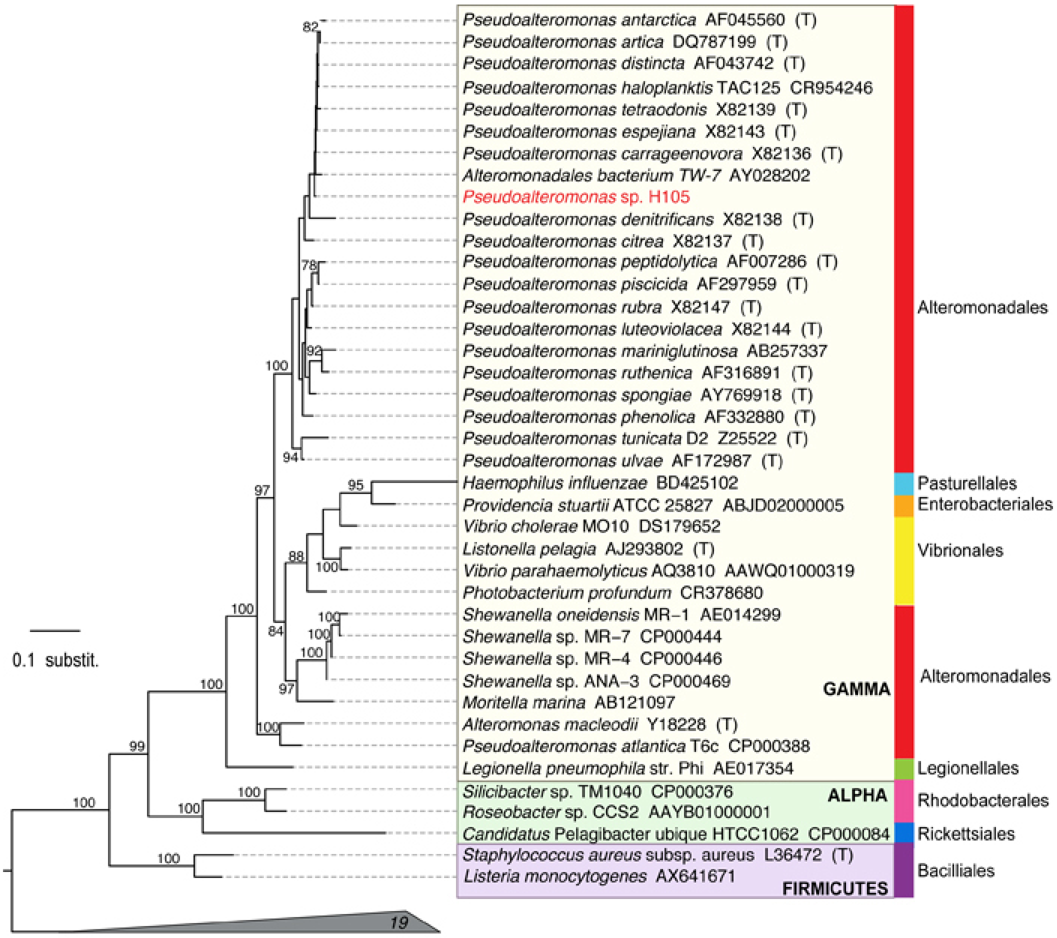

Pseudoalteromonas antarctica AF045560 (T)
Pseudoalteromonas artica DQ787199 (T)
Pseudoalteromonas distincta AF043742 (T)
Pseudoalteromonas haloplanktis TAC125 CR954246
Pseudoalteromonas tetraodonis X82139 (T)
Pseudoalteromonas espejiana X82143 (T)
Pseudoalteromonas carrageenovora X82136 (T)
Alteromonadales bacterium TW-7 AY028202
Pseudoalteromonas sp. H105
Pseudoalteromonas denitrificans X82138 (T)
Pseudoalteromonas citrea X82137 (T)
Pseudoalteromonas peptidolytica AF007286 (T)
Pseudoalteromonas piscicida AF297959 (T)
Pseudoalteromonas rubra X82147 (T)
Pseudoalteromonas luteoviolacea X82144 (T)
Pseudoalteromonas mariniglutinosa AB257337
Pseudoalteromonas ruthenica AF316891 (T)
Pseudoalteromonas spongiae AY769918 (T)
Pseudoalteromonas phenolica AF332880 (T)
Pseudoalteromonas tunicata D2 Z25522 (T)
Pseudoalteromonas ulvae AF172987 (T)
Haemophilus influenzae BD425102
Providencia stuartii ATCC 25827 ABJD02000005
Vibrio cholerae MO10 DS179652
Listonella pelagia AJ293802 (T)
Vibrio parahaemolyticus AQ3810 AAWQ01000319
Photobacterium profundum CR378680
Shewanella oneidensis MR-1 AE014299
Shewanella sp. MR-7 CP000444
Shewanella sp. MR-4 CP000446
Shewanella sp. ANA-3 CP000469
Moritella marina AB121097
Alteromonas macleodii Y18228 (T)
Pseudoalteromonas atlantica T6c CP000388
Legionella pneumophila str. Phi AE017354
Silicibacter sp. TM1040 CP000376
Roseobacter sp. CCS2 AAYB01000001
Candidatus Pelagibacter ubique HTCC1062 CP000084
Staphylococcus aureus subsp. aureus L36472 (T)
Listeria monocytogenes AX641671
GAMMA
ALPHA
FIRMICUTES
Alteromonadales
Pasturellales
Enterobacteriales
Vibrionales
Alteromonadales
Legionellales
Rhodobacterales
Rickettsiales
Bacilliales
82
78
92
100
94
97
95
88
100
100
84
100
100
100
97
100
100
99
100
100
100
100
0.1 substit.
19

Figure 4.1: Phylogenetic characterization of host *Pseudoalteromonas* sp. H105 16S rRNA gene. Maximum likelihood tree calculated with 1000 bootstraps using RAxML. RefSeq accession numbers follow the organism name; (T) indicates a type strain. Background shading highlights the taxonomic classification of the organism. Bootstrap values greater than 75 are shown on the branches; 19 sequences were used as an outgroup. Bar represents 10 percent estimated sequence change. ALPHA and GAMMA denote the class of *Proteobacteria* per cluster.

they infect (1000-13 000 kb) and typically abound with unknown gene content. Of the sequenced marine phage genomes, the majority of their open reading frames (over 60%) are hypothetical proteins (unique in public sequence databases) or conserved hypothetical proteins (similar only to other unknown proteins). As the public sequence databases are insufficient to grasp phage protein diversity, traditional approaches to genome analysis, which rely on similarity searches, e.g., blast (Altschul et al., 1990) against NCBI-nr, or protein family classification, e.g., Pfam (Finn et al., 2010), reveal little about phage evolution. We expanded the similarity searches of Phage H105/1 to include the Global Ocean Sampling dataset (GOS) (Rusch et al., 2007) and five publicly available marine viral metagenomes missing from NCBI-nr and -env (Angly et al., 2006; McDaniel et al., 2008). Such an approach lends itself to various "ecogenomic" interpretations, whereby ecological inferences about Phage H105/1 can be made based on genomic patterns and their respective environmental contextual data (Kottmann et al., 2010). Further complicating phage genomics, their evolution is driven by the rampant exchange of functional genome modules (Botstein, 1980; Hendrix et al., 1999; Pedulla et al., 2003), which are frequently swapped between phages infecting diverse hosts (Filée et al., 2006; Lucchini et al., 1999). Phage genomes can be thought of as veritable "concatenated metagenomes", in that consecutive fragments have very dissimilar origins and evolutionary pasts. Tetranucleotide usage frequencies, a feature increasingly used to cluster sequence fragments originating from discrete organisms, i.e., metagenomes (Andersson and Banfield, 2008; Dick et al., 2009; Woyke et al., 2006), were considered in this study as a tool to differentiate and shed light on the evolutionary history of Phage H105/1 'functional modules'. These novel approaches, integrated with experimental characterization of the phage's infection dynamics and structural proteome, offer strategies to elucidate ecological and evolutionary patterns and understand genomic features of Pseudoalteromonas phage H105/1, a temperate marine siphovirus.

4.2 Materials and Methods

4.2.1 Phage Harvesting, DNA Isolation and Sequencing

Pseudoalteromonas sp. strain H105 and Pseudoalteromonas phage H105/1 were isolated at 1 m, in September 1990 (Moebus, 1992), off the coast of Helgoland (10° 11'3 N, 7° 54' W) in the North Sea. Host was stored in liquid nitrogen, and the phage at 4° C in SM Buffer [100 mM NaCl, 81.2 mM MgSO4·7H2O, 50 mM Tris-HCl (pH 7.5), 0.01% gelatin]. Host was reconstituted in marine media and infected with

H105/1 (Oct 2006) using the agar overlay method (Wichels et al., 1998). Phages were harvested from plates with SM, precipitated (PEG/NaCl method) (Sambrook and Russell, 2001), and recovered in SM. Purified lysates were incubated (1 hr, 65° C) with Proteinase K (100 μg/ml final) and SDS (0.5% final). DNA was phenol:chloroform extracted, ethanol precipitated, and re-suspended in TE (1X). The genome was sequenced by Agowa GmbH (Berlin, Germany) using a linear Escherichia coli vector, pJAZZ-KA (BigEasy-pTEL, Lucigen; Middleton, WI, USA).

4.2.2 Virion Structural Proteome Analysis

Lysates were purified via CsCl centrifugation (Sambrook and Russell, 2001). Briefly, debris was extracted from PEG purified lysates with chloroform (1:1), vortexed, centrifuged (3000 x g, 15 min, 4° C), and the aqueous phase used in CsCl purification. The gradient tube (Ultra-Clear™, Beckman, Fullerton, CA, USA) was layered with 1.125 ml each of (1) 1.7 g CsCl/ml, (2) 1.5 g CsCl/ml, (3) 1.45 CsCl/ml, (4) topped with 1.15 g CsCl/ml, and centrifuged (87,000 x g, 2 hrs, 4 ° C). A blue-white band containing the phage was removed with a sterile syringe (2 ml) and dialyzed (Pierce Slide-A-Lyzer 10K MWCO, Rockford, IL, USA) twice in one liter buffer [10 mM NaCl, 50 mM Tris-HCl (pH 8), 10 mM MgCl2] to remove CsCl. Phages were concentrated 10x (Microcon 30 kD; Millipore) and proteins denatured by five freeze-thaw (96° C) cycles and 12% SDS, then separated on 12% SDS-polyacrylamide gel electrophoresis (PAGE), as described by (Paul et al., 2005). A MALDI-TOF-MS peptide mass fingerprint spectra was generated from trypsin-digested bands excised from PAGE gel (TOPLAB GmbH; Martinsried, Germany). Each peptide fingerprint was best matched to its original protein in Phage H105/1 genome using probability-based Mowse Score (-10*Log(P), where P is probability that observed match is random). Protein scores greater than 29 were considered significant ($p < 0.05$).

4.2.3 Genome Annotation

Genes were predicted based on (i) GeneMark.hmm (prokaryotic version using bacterial/archaeal genetic code, precomputed *Pseudoalteromonas haloplanktis* chromosome 1 model, and default settings, Besemer et al., 2001) and (ii) FGENESB (generic bacterial model, default settings; Softberry, Mount Kisco, NY, USA), also used to predict operons. Rho-independent bacterial transcriptional terminators were predicted using FindTerm (energy threshold -11, default settings; Softberry). Promoters were predicted searching regions 150 bp upstream of predicted starts (BPROM, threshold 0.2, default settings; Softberry) with custom Perl wrappers. Annotation and comparative genomics of Phage H105/1 employed JCoast (Richter et al., 2008), streamlining annotation protocols and results of: Blastp (low complexity filter) against NCBI-nr, Pfam (Finn et al., 2010), SignalP (Emanuelsson et al., 2007), and TMHMM (Krogh et al., 2001). Predicted ORFs were searched against ACLAME MeGO (Mobile Element Gene Ontology) database, which provides functional annotations based on manually curated database of viruses and mobile genetic elements (Toussaint et al., 2007). When Phage H105/1 proteins were most similar to proteins of bacterial or archaeal genome, hits were classified as 'prophage' if (a) they have a 'phage-like neighborhood' (i.e., phage-like proteins 10 genes up/downstream of a 100-kb range), or (b) they lie in a Profinder-predicted prophage (Lima-Mendez et al., 2008a).

4.2.4 Ecogenomic Analysis

Reads from Global Ocean Sampling ('GOS', 0.1-0.8 μm fraction, plus two 0.8-3.0 μm fraction samples, Rusch et al., 2007) were retrieved from CAMERA database (Seshadri et al., 2007) (Supplementary Material 7.3). Five marine virus metagenomes were retrieved from NCBI, representing pooled viromes from the Arctic, British Columbia (Straight of Georgia estuary), Gulf of Mexico, Sargasso Sea, and a coastal Tampa Bay community whose integrated prophages were induced to undergo lysis (gpids: 18225 and 28619; Supplementary Material 7.4, Angly et al., 2006; McDaniel et al., 2008). Specialized tblastn (BLOSUM62 substitution matrix) of predicted Phage H105/1 proteins was performed against all six reading frames of nucleotide reads using Decypher (TimeLogic, Inc., Carlsbad, CA, USA) hardware. Hits with e-values $<10^{-4}$ and 20% query coverage (GOS) or 10% query coverage (viromes) were accepted to minimize false positives as determined by the behavior over a range of thresholds (Supplementary Material 7.4). Raw hit counts were normalized by gene size, number of reads/site, and (for GOS) number of sites/habitat to reduce the effects of uneven sampling of different habitats; values were multiplied by a constant (10^8 for GOS, 10^7 for virome hits) to bring final counts near whole numbers. To determine the extent to which habitat influences the distribution of Phage H105/1 proteins, environmental physiochemical parameters (temperature, salinity, nitrate, phosphate, silicate, dissolved oxygen, oxygen saturation, oxygen utilization) for the marine viromes were collected using megx.net GIS tools (Kottmann et al., 2010) based on published location, depth and time parameters (Supplementary Material 7.4), except for the Arctic sites, where interpolation is not possible. Unfortunately for the purposes of this study, many of the viromes were pooled samples collected over a range of locations and depths. When a depth range was reported, data for minimum and maximum (and for BBC, intermediate) depths were collected. Sample sites were clustered based on a distance matrix (Euclidean distance coefficient) of z-score transformed environmental data using average linkage clustering.

4.2.5 Host Phylogenetic Analysis and Tree Reconstruction

16S rRNA gene tree. A 500 μl aliquot of *Pseudoalteromonas* sp. H105 culture was subjected to 4x freeze-thaws and used in polymerase chain reaction under standard conditions with GM3F and GM4R primers (Muyzer et al., 1995). Products were gel extracted, purified, used to construct clone libraries with the pGEM-T Easy Vector System I (Promega; Madison, WI, USA), and inserts sequenced. Host 16S rRNA gene sequence was imported into ARB (Ludwig et al., 2004) with the SILVA 98 SSU Ref dataset (Pruesse et al., 2007), from which additional sequences for the tree were selected and exported using a 5% similarity filter to remove highly variable positions. A maximum likelihood tree was calculated using a randomly seeded rapid bootstrap analysis (n = 1000) and search for best-scoring tree using RAxML (Stamatakis, 2006), version 7.0.4 (MPI Master process), with the GTRGAMMA model, which optimizes substitution rates and uses a GTR model of nucleotide substitution and a gamma model of rate heterogeneity.

ssb protein tree. Sequences were retrieved from GenBank and aligned (ClustalW, default parameters, Larkin et al., 2007). Maximum likelihood tree was calculated using the JTT matrix model for protein distance and empirical base frequencies and bootstrapped (n = 1000).

4.2.6 Genome Signature Analysis

Codon Adaptation Index (CAI). As its specific host is not sequenced, the relative 'host' codon adaptation of the Pseudoalteromonas Phage H105/1 proteins was modeled using *Pseudoalteromonas atlantica* and *Pseudoalteromonas haloplanktis* genomes as a reference. Codon adaptation index (CAI) of the phage to these bacteria was calculated using Jcat (Grote et al., 2005), which implements the algorithm proposed by Carbone et al. to distinguish highly expressed genes based on internal codon biases.

Tetranucleotide frequencies. To examine tetranucleotide correlations, all large sequence fragments (>25 kb, n = 347 886) were retrieved from GenBank (July 2008).

H105/1 genome was split into 30 fragments (1 kb steps, 10 kb window). Fragments were extended by their reverse complement to account for strand biases. Observed and expected frequencies for the 256 possible tetrads were computed by a maximal-order Markov model; differences between observed and expected frequencies were transformed into z-scores (Teeling et al., 2004). To determine which GenBank sequences are most similar to Phage H105/1 genome fragments, the squared z-scores were correlated and Pearson coefficient of all pair-wise correlations calculated (Waldmann, 2010). The coefficient cut-off was determined as the minimum value resulting in a score in each of the 30 windows. A balance between highest possible correlation coefficients (large window size) and most granular resolution of tetra signal along the genome (small window size) was achieved using a 10 kb window and 0.61 Pearson coefficient cut-off (Supplementary Material 7.5). Correlation scores for each 10 kb genome fragment were normalized, such that the sum of all coefficients of each GenBank fragment type recruited ('Bacteria', 'Phage', or 'Unassigned') was divided by the total sum and cumulatively mapped onto the Phage H105/1 genome for each window. Considering the window and step sizes, each 1 kb portion of the genome is represented by 10 overlapping 10 kb fragments all normalized to one, thus the sum of Bacteria, Phage, and Unassigned scores will equal 10 at all points along the genome.

4.3 Results and Discussion

4.3.1 Pseudoalteromonas Phage H105/1 Biology

Phage H105/1 has a long, non-contractile tail (characteristic of siphoviruses), with unique knobs (Supplementary Material 7.6A Wichels et al., 1998). Infection with Phage H105/1 led to rapid lysis, as evidenced by plaque formation in 10-12 hours. Intriguingly, the plaques had clear centers surrounded by fuzzy haloes (Supplementary Material 7.6B). Haloed plaques are thought to indicate either (i) the presence of a phage-encoded polysaccharide depolymerase (Erskine, 1973; Vandenbergh and Cole, 1986), or (ii) 'pseudolysogeny', a poorly understood condition used to describe the sustained coexistence of a high number of both virus and host. Phage H105/1 does not appear to carry a polysaccharide depolymerase (Figures 4.2a, Table 4.1). The latter case, pseudolysogeny, is thought to be caused by stalled/incomplete lysis of the host population, as the phage upon infection passively resides in its host, neither integrating, nor lysing, nor replicating as a plasmid in host progeny (Miller and Day, 2008). The condition is hypothesized to be a phage strategy to weather starved, low energy conditions (Ripp and Miller, 1997), and, though growth on nutrient-rich agar can hardly be considered oligotrophic, it may represent sub-optimal conditions for the phage-host system. Haloed plaques have been observed in other marine phages, where pseudolysogeny has been implicated: Pseudoalteromonas phages H24/1 and H24/2 (isolated from Helgoland on *Pseudoalteromonas* sp. H24 (Moebus, 1997)) and Listonella phage HSIC (Williamson et al., 2001). Its plaque behavior and lambda-like genome content and architecture (Figure 4.2a) suggest Pseudoalteromonas phage H105/1 is a temperate phage, able to (or with the past ability to) integrate into its host genome.

4.3.2 Genome Features and Annotations

Pseudoalteromonas phage H105/1 is 30.7 kb with 52 predicted ORFs. The total G+C% content is 40.85% and the genome-wide coding density is 91%, which is comparable to the average coding density of all marine phages: 89% (data not shown). Over 60% of Phage H105/1 ORFs are unknown (Figure 4.2b), though the genome organization shows remarkable functional synteny with other lambda-like siphoviruses (Figure 4.2a), which is likely preserved by the temporal control under which phage genes are transcribed (Calendar, 1970). Phage H105/1 has two distinct functional supermodules, whereby the proteins that require direct interaction with the host genome, replicative machinery, metabolic/stress response processes, and cell lysis (Host Interaction Module) are physically separated from those involved in structure and assembly (Phage Structural Module; Figure 4.2a). Intriguingly, of the six most similar sequenced phages (those sharing the greatest number of proteins), three are marine (Figure 4.2a), suggesting an overall 'marine' character (Figure 4.2c).

Phylogenetic Signature. Though 33 proteins have no homologues in GenBank (hypothetical proteins), the bacterial homologues show a distinct trend towards the host class (*Gammaproteobacteria*), and the phage homologues are dominated by either Siphoviridae (Phage H105/1 class) or prophages (Figure 4.2b), providing further bioinformatic support that Phage H105/1 integrates as well. Of the 11 phage hits, eight have gammaproteobacteria hosts. The phylogenetic signature of Phage H105/1 suggests that a majority of its proteins come from a common pool of *Gammaproteobacteria* or phages that infect *Gammaproteobacteria*. Such a host phylogenetic trend has been seen previously in phage genomes (Sullivan et al., 2005) and supports the view that phages are a mobile genomic extension of the hosts they infect (Siefert, 2009).

Table 4.1: Pseudoalteromonas phage H105/1 ORF and annotation summary based on homology to NCBI-nr, Pfam, and MeGo databases.

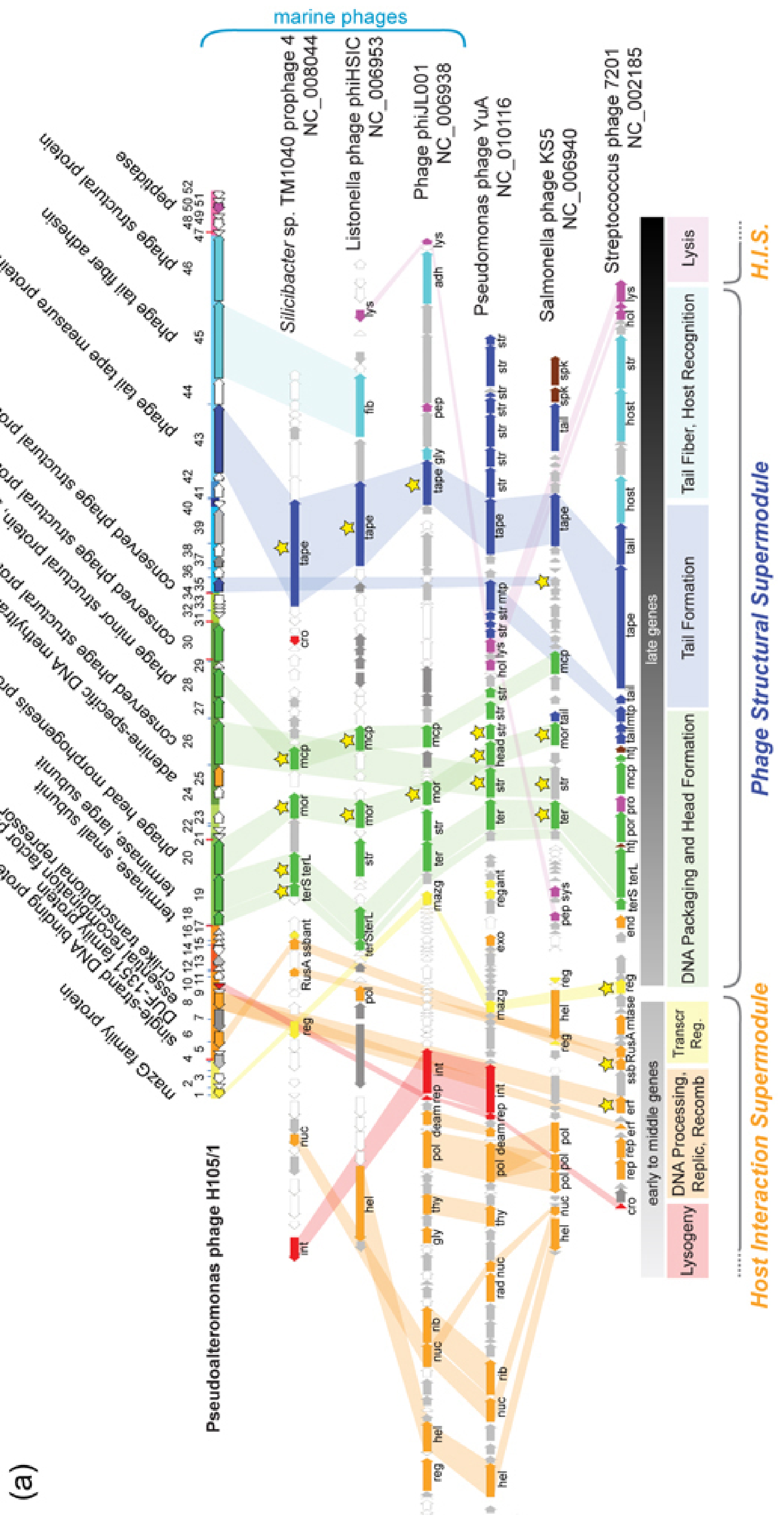
ORF	AA	Annotation	e-value (% id / % orf cov.) BlastP (tax); accn	Pfam; MeGO
1 / -	103	transcriptional repressor, MazG family protein	4.0E-12 (46/83) *Nitratiruptor* sp. SB155-2 prophage (*Epsilonproteobacteria*); YP_001355920	PF03819: MazG, 3.7E-2; MeGO: transcriptional repressor activity (phi:0000127), maintenance of prophage immunity (phi:0000057)
2 / -	64	hypothetical protein	no sig.	
3 / -	109	hypothetical protein	no sig.	
4 / -	156	conserved hypothetical protein	3.0E-19 (38/96) *Vibrio cholerae* V51 (*Gammaproteobacteria*; *Vibrionales*); ZP_01487122	
5 / -	59	hypothetical protein		
6 / -	184	single-stranded DNA binding protein	9.0E-43 (71/61) Alteromonadales bacterium TW-7 (*Gammaproteobacteria*; *Alteromonadales*); ZP_01613173	PF00436: SSB, 1.6E-40; MeGO: single-stranded DNA binding (GO:0003697)
7 / -	239	conserved hypothetical phage protein, DUF1351 family	9e-05 (47) Iodobacteriophage phiPLPE (Myovirus)	PF07083: DUF1351 protein family of unknown function, 1E-1; MeGO: phage function unknown (phi:0000326)
8 / -	191	essential recombination function protein	2.0E-40 (63/72) Enterobacteria phage P22 (Podovirus), host: *Salmonella enterica* serovar *typhimurium*; NP_059596	PF04404: ERF superfamily, 2.2E-28; MeGO: DNA strand annealing activity (GO:0000739), DNA recombination (phi:0000130)
9 / -	34	hypothetical protein	no sig.	
10 / -	62	transcriptional repressor	7.7E-2 (46) Bacteriophage APSE-2, host: *Candidatus* Hamiltonella defensa; ACJ10163	PF01381: HTH_3, 9.7E-2; MeGO: transcriptional repressor activity (phi:0000127), maintenance of prophage immunity (phi:0000057)
11 / +	91	hypothetical protein	no sig.	
12 / +	65	hypothetical protein	no sig.	
13 / +	101	conserved hypothetical protein	5.0E-4 (24/77) *Pseudomonas fluorescens* Pf0-1; YP_349064	
14 / -	89	hypothetical protein	no sig.	
15 / +	55	hypothetical protein	no sig.	
16 / +	63	hypothetical protein	no sig.	
17 / +	56	hypothetical protein	no sig.	

table continued on the following page

ORF	AA	Annotation	e-value (% id / % orf cov.) BlastP (tax); accn	Pfam; MeGO
18 / +	153	phage terminase, small subunit	8.0E-17 (52/62) *Yersinia enterocolitica* subsp. Enterocolitica 8081 prophage (*Gammaproteobacteria*; *Enterobacteriales*); YP_001006550	PF03592: Terminase small subunit; 1.9E-13; MeGO: phage terminase small subunit (phi:0000074), DNA binding activity (phi:0000109), phage DNA maturation (phi:0000019)
19 / +	415	phage terminase, large subunit	9.0E-135 (57/97) *Silicibacter* sp. TM1040 prophage (*Alphaproteobacteria*; *Rhodobacterales*); YP_612796	PF03237: Terminase-like family, 1.9E-31; MeGO: phage terminase large subunit (phi:0000073); phage DNA maturation (phi:0000019)
20 / +	388	phage head morphogenesis protein	Pseudomonas phage YuA (Siphovirus); YP_001595877	PF04233: Phage Mu protein F like, 1.2E-2; MeGO: phage head/capsid minor protein (phi:0000185)
21 / -	59	hypothetical protein	no sig.	
22 / -	81	hypothetical protein	no sig.	
23 / +	114	hypothetical protein	no sig.	
24 / +	127	hypothetical protein	no sig.	
25 / +	229	adenine-specific DNA methyltransferase	9.0E-31 (40/97) *Spiroplasma citri* poss. degenerate prophage (*Tenericutes*; *Mollicutes*; *Entomoplasmatales*); CAK98777	PF01555: N6_N4_Mtase; 1.3E-30; MeGO: DNA methyltransferase activity (phi:0000117)
26 / +	489	conserved phage structural protein	1.0E-33 (27/94) *Pseudomonas fluorescens* Pf-5 prophage (*Gammaproteobacteria*; *Pseudomonadales*); YP_260866	MeGO: phage function unknown (phi:0000326)
27 / +	237	phage minor structural protein GP20 family	no sig.	PF06810: Phage minor structural protein GP20 family; 3.2E-3
28 / +	320	conserved phage structural protein	*Delftia acidovorans* SPH-1 prophage (*Betaproteobacteria*; *Burkholderiales*); YP_00156426	
29 / +	54	hypothetical protein	no sig.	
30 / +	411	conserved hypothetical phage protein	1.0E-4 (21/89) Vibrio phage KVP40 (Myovirus); NP_899611	
31 / -	77	hypothetical protein	no sig.	
32 / -	61	hypothetical protein	no sig.	
33 / -	61	hypothetical protein	no sig.	

table continued on the following page

ORF	AA	Annotation	e-value (% id / % orf cov.) BlastP (tax); accn	Pfam; MeGO
34 / -	134	hypothetical protein	no sig.	
35 / +	161	conserved phage structural protein	5.1E-2 (31/78) Pseudomonas phage M6 (Siphovirus); YP_001294532	MeGO: phage function unknown (phi:0000326)
36 / +	119	hypothetical protein	no sig.	
37 / +	127	conserved hypothetical phage protein	4.0E-6 (33/83) Salmonella phage KS7 (Siphovirus)	MeGO: phage function unknown (phi:0000326)
38 / +	140	hypothetical protein	na	
39 / +	391	conserved hypothetical protein	3.0E-4 (25/74) alpha proteobacterium BAL199 (*Proteobacteria*; *Alphaproteobacteria*); ZP_02186593	
40 / -	80	hypothetical protein	no sig.	
41 / +	154	hypothetical protein	no sig.	
42 / +	91	hypothetical protein	no sig.	
43 / +	767	phage tail tape measure protein	9.0E-19 (33/31) *Verminephrobacter eiseniae* EF01-2 poss. degenerate prophage (*Burkholderiales*); YP_999425	MeGO: phage tail tape measure protein (phi:0000086)
44 / +	292	hypothetical protein	no sig.	
45 / +	867	phage tail fibre adhesin Gp38 family protein	no sig.	PF05268: Phage tail fibre adhesin Gp38, 3.E-2
46 / +	747	phage structural protein	no sig.	
47 / -	53	hypothetical protein	no sig.	
48 / -	126	hypothetical protein	no sig.	
49 / -	61	hypothetical protein	no sig.	
50 / -	114	carboxypeptidase, Peptidase M15 family protein	3.0E-15 (36/99) *Magnetococcus* sp. MC-1 poss. degenerate prophage (*Proteobacteria*); YP_865602	PF08291: Peptidase M15, 3.4E-20; MeGO: carboxypeptidase activity (GO:0004180)
51 / -	51	hypothetical protein	no sig.	
52 / -	83	hypothetical protein	no sig.	


(a)
Pseudoalteromonas phage H105/1
mazG family protein
single-strand DNA binding protein
DUF-1351 family protein
essential recombination factor protein
cI-like transcriptional repressor
terminase, small subunit
terminase, large subunit
phage head morphogenesis protein
adenine-specific DNA methyltransferase
conserved phage structural protein
phage minor structural protein, gp20 family
conserved phage structural protein
conserved phage structural protein
phage tail tape measure protein
phage tail fiber adhesin
phage structural protein
peptidase
marine phages
Silicibacter sp. TM1040 prophage 4
NC_008044
Listonella phage phiHSIC
NC_006953
Phage phiJL001
NC_006938
Pseudomonas phage YuA
NC_010116
Salmonella phage KS5
NC_006940
Streptococcus phage 7201
NC_002185
early to middle genes
late genes
Lysogeny
DNA Processing, Replic, Recomb
Transcr Reg.
DNA Packaging and Head Formation
Tail Formation
Tail Fiber, Host Recognition
Lysis
Host Interaction Supermodule
Phage Structural Supermodule
H.I.S.

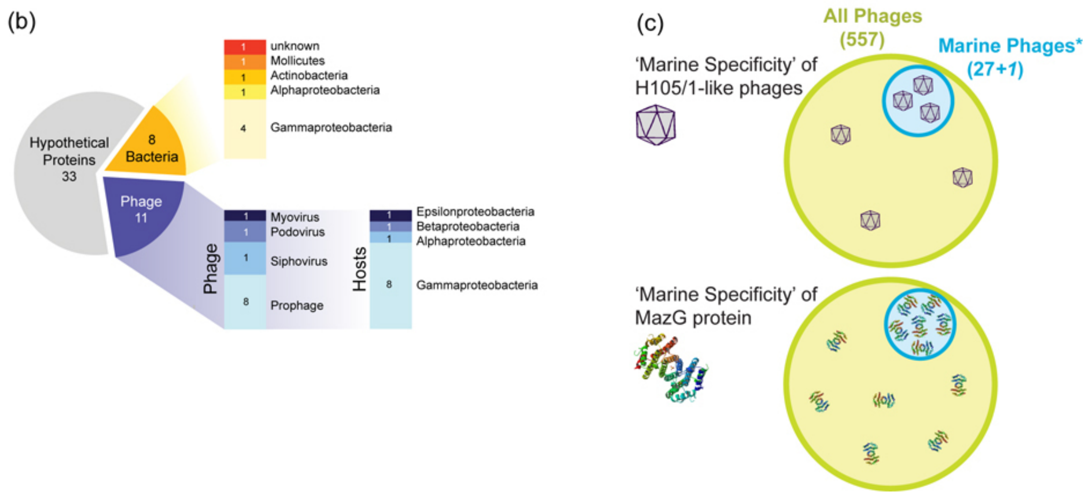


Figure 4.2: Pseudoalteromonas phage H105/1 genome and synteny with other siphoviruses. Conserved ORFs with homologues in GenBank or those in the phage proteome (Figure 4.4) are color-coded based on their functional module (labeled in the bottom row); hypothetical proteins, with no homologous in GenBank (white); conserved hypothetical bacterial proteins (light grey); conserved hypothetical phage proteins (dark grey); promoters (blue lines), transcription terminators (red squiggles); operons are delineated by color blocks behind the Phage H105/1 genome. Also shown are the six phages sharing the greatest number of homologues with Phage H105/1. Background vertical color blocks connect phage proteins of similar function, while stars indicate explicit blastp-based sequence similarity to Phage H105/1 ($e<10^{-5}$). The three 'marine phages' are indicated. Early, middle, and late genes describe the temporal transcription of the different modules typical of lambda-like phages. The 'Host Interaction Supermodule' and 'Phage Structural Module' are indicated as defined in this study. int - integrase; nuc - exodeoxyribonuclease; reg - transcriptional regulator; RusA; ssb - single-stranded binding protein; ant - antirepressor; terS - terminase, small subunit; terL - terminase, large subunit; ter - terminase; mor - morphogenesis protein; mcp - major capsid protein; cro - repressor; tape - tape measure protein; hel - helicase; pol - DNA polymerase; str - structural protein; fib - tail fiber; lys - lysin; rib - ribonucleotide reductase; gly - glycosyl transferase; thy - thymidylate synthase; deam - deaminase; rep - repressor; mazg; pep - peptidase; adh - adhesin; rad - ; exo - exonuclease; hol - holin; mtp - minor tail protein; spk - tail spike protein; erf - essential recombination function protein; end - endonuclease; mtase - methyltransferase; htj - head-tail joining protein; por - portal; tail - tail protein; host - host recognition protein. (b) Best Blastp hits to Pseudoalteromonas phage H105/1 are characterized as Hypothetical Proteins (no similarity to proteins in NCBI-nr), Phage (including manually determined prophages), or Bacteria. Phage hits are classified by virus family and host class; Bacteria hits are classified by class. (c) Overrepresentation of marine phages (of 27) among the six most similar to Phage H105/1, relative to all available phage genomes (557), and overrepresentation of marine phages among all phages containing the MazG protein domain. *one additional marine prophage is considered among the marine phages.

Host Interaction: *Recombination and Replication.* Containing a MazG pyrophosphohydrolase domain, it is likely that ORF 1 is involved in transcriptional repression (Table 4.1). In Escherichia coli, MazG is known to interfere with (or reverse) starvation-induced programmed cell death by decreasing the cellular pool of effector nucleotide, guanosine 3′,5′-bispyrophosphate (ppGpp) (Gross et al., 2006). When cyanobacteria are subjected to nitrate starvation, their pool of ppGpp increases and amino acid levels drop (Friga et al., 1981), but this process can be impeded by phage infection (Borbély et al., 1980). Thus, if functional, a phage MazG protein may help maintain the metabolism of a starving host (Bryan et al., 2008; Clokie and Mann, 2006) long enough for the phage to propagate. Of the twelve phage proteins in this domain family, six are marine (Figure 4.2c): Phage H105/1, Roseobacter phage SIO1, and Cyanophages P-SSM2, P-SSM4, S-PM2, and Syn9, suggesting a unique marine signature to this protein family not seen in any other Phage H105/1 proteins, and implicating an important role for MazG in marine phage systems.

ORF 6, a single-stranded binding (ssb) protein is often found in an operon with essential recombination function (erf) proteins (ORF 8; Figure 4.2a); they are known to interact as erf specifically binds single-stranded DNA to facilitate phage genome circularization (Iyer et al., 2002; Poteete et al., 1983). The ssb protein has a strong ’host phylogenetic signature’. Of all similar ssb proteins in GenBank, it clusters most closely with host-like homologues, while plasmids form there own ssb cluster (Figure 4.3), none of which are from integrated prophages, suggesting the ssb is of host origin. Considering that such a strong host phylogenetic association is not seen in homologues of any other Phage H105/1 ORF (Table 4.1), and that phage ssb proteins cluster most closely with ssb proteins of their host, or host affiliation (Figure 4.3), single-stranded binding proteins may serve as an informative diagnostic of phage-host associations, especially for temperate phages that could benefit from host-like recombination proteins.

Triggered by (host) stress-inducing environmental conditions, temperate phages rely on a ’genetic switch’ to initiate the lytic replication cycle (Dodd and Egan, 1996; Ptashne, 2004). The lysogenic state of integrated prophages is maintained by the binding of a repressor protein, which prevents the expression of phage genes needed for lytic replication. Containing a helix-turn-helix domain found in phage and plasmid transcription control proteins, and with homology to putative phage cI proteins, ORF 10 may be involved in cI repressor-like activity (Table 4.1).

Phage Structure and DNA Packaging. Structural proteomics was used to investigate the proteins of the mature Phage H105/1 virion. Seven proteins of the

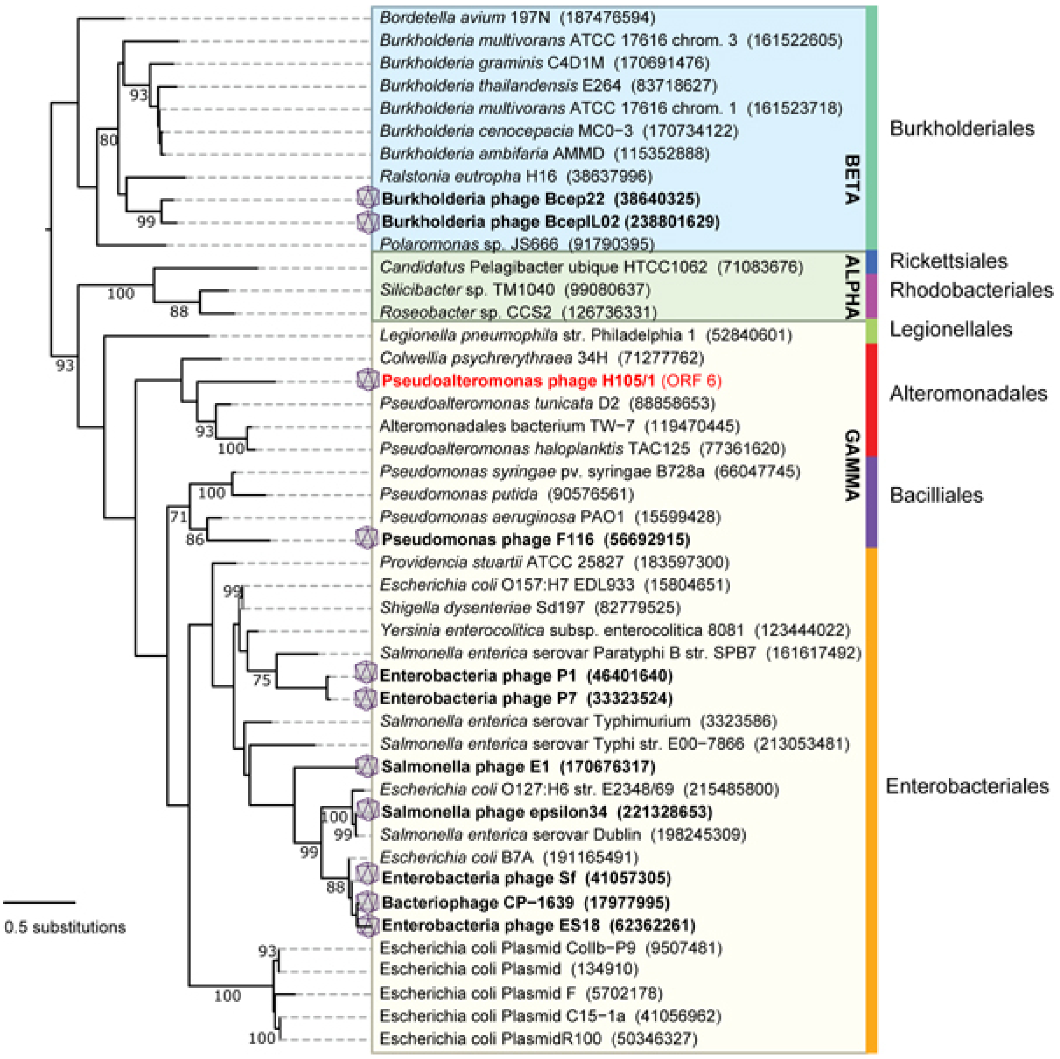

Bordetella avium 197N (187476594)
Burkholderia multivorans ATCC 17616 chrom. 3 (161522605)
Burkholderia graminis C4D1M (170691476)
Burkholderia thailandensis E264 (83718627)
Burkholderia multivorans ATCC 17616 chrom. 1 (161523718)
Burkholderia cenocepacia MC0-3 (170734122)
Burkholderia ambifaria AMMD (115352888)
Ralstonia eutropha H16 (38637996)
Burkholderia phage Bcep22 (38640325)
Burkholderia phage BcepIL02 (238801629)
Polaromonas sp. JS666 (91790395)
Candidatus Pelagibacter ubique HTCC1062 (71083676)
Silicibacter sp. TM1040 (99080637)
Roseobacter sp. CCS2 (126736331)
Legionella pneumophila str. Philadelphia 1 (52840601)
Colwellia psychrerythraea 34H (71277762)
Pseudoalteromonas phage H105/1 (ORF 6)
Pseudoalteromonas tunicata D2 (88858653)
Alteromonadales bacterium TW-7 (119470445)
Pseudoalteromonas haloplanktis TAC125 (77361620)
Pseudomonas syringae pv. syringae B728a (66047745)
Pseudomonas putida (90576561)
Pseudomonas aeruginosa PAO1 (15599428)
Pseudomonas phage F116 (56692915)
Providencia stuartii ATCC 25827 (183597300)
Escherichia coli O157:H7 EDL933 (15804651)
Shigella dysenteriae Sd197 (82779525)
Yersinia enterocolitica subsp. enterocolitica 8081 (123444022)
Salmonella enterica serovar Paratyphi B str. SPB7 (161617492)
Enterobacteria phage P1 (46401640)
Enterobacteria phage P7 (33323524)
Salmonella enterica serovar Typhimurium (3323586)
Salmonella enterica serovar Typhi str. E00-7866 (213053481)
Salmonella phage E1 (170676317)
Escherichia coli O127:H6 str. E2348/69 (215485800)
Salmonella phage epsilon34 (221328653)
Salmonella enterica serovar Dublin (198245309)
Escherichia coli B7A (191165491)
Enterobacteria phage Sf (41057305)
Bacteriophage CP-1639 (17977995)
Enterobacteria phage ES18 (62362261)
Escherichia coli Plasmid CoIIb-P9 (9507481)
Escherichia coli Plasmid (134910)
Escherichia coli Plasmid F (5702178)
Escherichia coli Plasmid C15-1a (41056962)
Escherichia coli PlasmidR100 (50346327)
93
80
99
100
88
93
93
100
100
71
86
99
75
100
99
99
88
93
100
100
BETA
ALPHA
GAMMA
Burkholderiales
Rickettsiales
Rhodobacteriales
Legionellales
Alteromonadales
Bacilliales
Enterobacteriales
0.5 substitutions

Figure 4.3: Neighbor joining tree of ORF 8, single-stranded binding protein. Consensus tree generated from 1000 bootstrapped re-sampled versions of the original dataset using the JTT matrix model for protein distance measures. Bootstraps values greater than 75 displayed on the branches. A phage capsid symbol and bold print denote phage sequences. Note that ssb ORF 6 of Pseudoalteromonas phage H105/1 groups with non-prophage single-stranded binding proteins of *Pseudoalteromonas tunicata* D2, *Pseudoalteromonas haloplanktis* TAC125, and *Alteromonadales bacterium* TW-7 (manually determined by examining gene neighborhood in the bacterial genome). All other phages group with their host, or a closely related organism, while plasmids cluster independently. ALPHA, BETA, GAMMA denotes the class of *Proteobacteria* per cluster.

'Phage Structural Module' were identified in the phage proteome (Figure 4.4, Supplementary Material 7.5): a phage head morphogenesis protein (ORF 20), a phage tail tape measure protein (ORF 43), a phage tail fiber adhesin (ORF 45), and four novel proteins (ORFs 26, 28, 35, and 46) that are verified experimentally now as structural proteins.

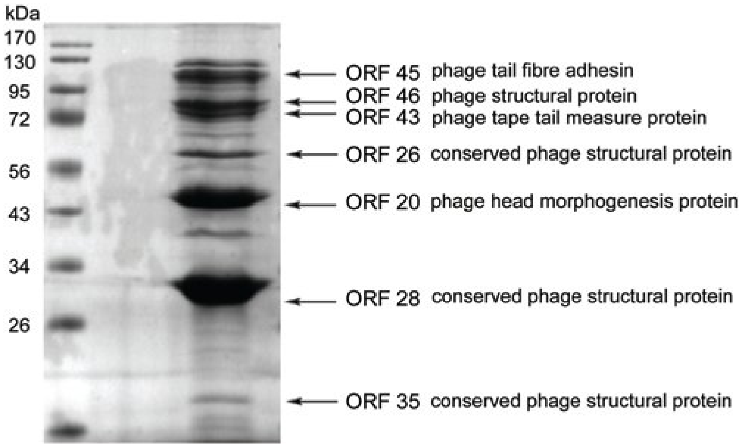

kDa
170
130
95
72
56
43
34
26
ORF 45 phage tail fibre adhesin
ORF 46 phage structural protein
ORF 43 phage tape tail measure protein
ORF 26 conserved phage structural protein
ORF 20 phage head morphogenesis protein
ORF 28 conserved phage structural protein
ORF 35 conserved phage structural protein

Figure 4.4: Pseudoalteromonas phage H105/1 proteome. SDS-PAGE gel image of Pseudoalteromonas phage H105/1 lysate containing mature phage particles. ORFS 26, 28, 35, 43, 45, and 46, which were previously unknown proteins (hypothetical or conserved hypothetical), are annotated as phage structural proteins based on this proteomic confirmation.

The 'Phage Structural Supermodule' of Phage H105/1, responsible for phage assembly and structure, is syntenous with the morphogenetic operon of other temperate phages and prophages (Figure 4.2a) (Botstein and Matz, 1970; Canchaya et al., 2003). Typical of a lambda-like morphogenetic operon (Casjens, 2003), ORF 20, a putative head morphogenesis protein, is found in the 'DNA Packaging and Head Formation' module with the large and small terminases (ORFs 18 and 19), ATP-binding proteins that cut the concatenated phage DNA and connect it to a portal protein, such that "headful" packaging can proceed (Black, 1989) (Figure 4.2a). ORF 27 shares a domain with the Staphylococcus phage-dominated minor structural protein Gp20 family (PF06810). Among the 'Tail Formation' genes, ORF 43, a phage tail length tape measure protein, is involved in the regulation of the phage tail length (Abuladze et al., 1994). In the 'Tail Fiber, Host Recognition' module, ORF 45 contains a domain of the Phage tail fiber adhesin Gp38 Pfam family (Table 4.1). In T2-like phages, gp38 is responsible for recognition of host cell receptors (Haggard-Ljungquist et al., 1992), thus is under great selection for change and one of the most rapidly evolving components of a phage-host system. The presence of ORF 25 (Table 4.1), a methyltransferase, among the Phage H105/1 'late genes' involved in phage structure (rather than in a DNA modification module of an 'early' operon (Figure 4.2a) (Mobberley et al., 2008)) suggests that the enzyme does not methylate incoming phage DNA at the time of infection/insertion in an attempt to mask itself from host restriction enzymes. An alternative strategy, also proposed in Bacteriophage N15 (Ravin et al., 2000), may exist: as new virions are assembled during the lytic phase, the replicated DNA is methylated prior to packaging.

Host lysis. Host lysis requires both a phage lysin and holin to dissolve the membrane potential and permeabilize the cell wall, respectively (Wang et al., 2000). ORF 50 contains a conserved domain of the Peptidase M15 Pfam family of metallopeptidases (Table 4.1), a lysin likely involved in host cell lysis.

4.3.3 Ecogenomics: H105/1 in GOS and five marine virus metagenomes

Of the 52 ORFs, 14 have homologues in samples from the GOS dataset (Figure 4.5A). These genes, many of which are found in the 'Host Interaction Supermodule' (ORFs 1, 6, 8, 50), are seen proportionately more in the GOS 'Estuary' sites, with the most hits (63) to Delaware Bay (NJ, USA). Among the marine virus metagenomes, there are proportionally more hits to the British Columbia samples ('BBC', Figure 4.5b), a trend again strongest in the 'Host Interaction Supermodule'. The British Columbia surface site (low than average salinity) clusters with Helgoland, a low-salinity, turbid region of the North Sea, influenced by the Elbe river plume (Becker et al., 1992) (Figure 4.5d; Supplementary Material 7.4). The BBC surface parameters are most diagnostic of the BBC virome, as nearly all of the 86 pooled samples were from the upper water column (C. Suttle, personal communication). This ecogenomic trend may reflect the original habitat of Phage H105/1 and further intimates the importance of temperate phages in offering genome plasticity to lysogens (hosts with integrated prophages) in unstable habitats, a concept also suggested by a high prevalence of lysogens among microbial populations in a low salinity, high turbidity Mississippi River plume (Long et al., 2008). In light of the assumption that there are site-specific differences in the relative abundances of similar virus sequences between viromes (Angly et al., 2006), these biases may be influenced by environmental parameters. The estuary-enriched hits tend to be found in the 'Host Interaction Module', which may represent the mechanism through which "phage organismal ecology" can exist. The adaptation of a phage to its environment can happen only through close association with its host, whose metabolic state it can respond to via, i.e., phage-mediated transcriptional regula-

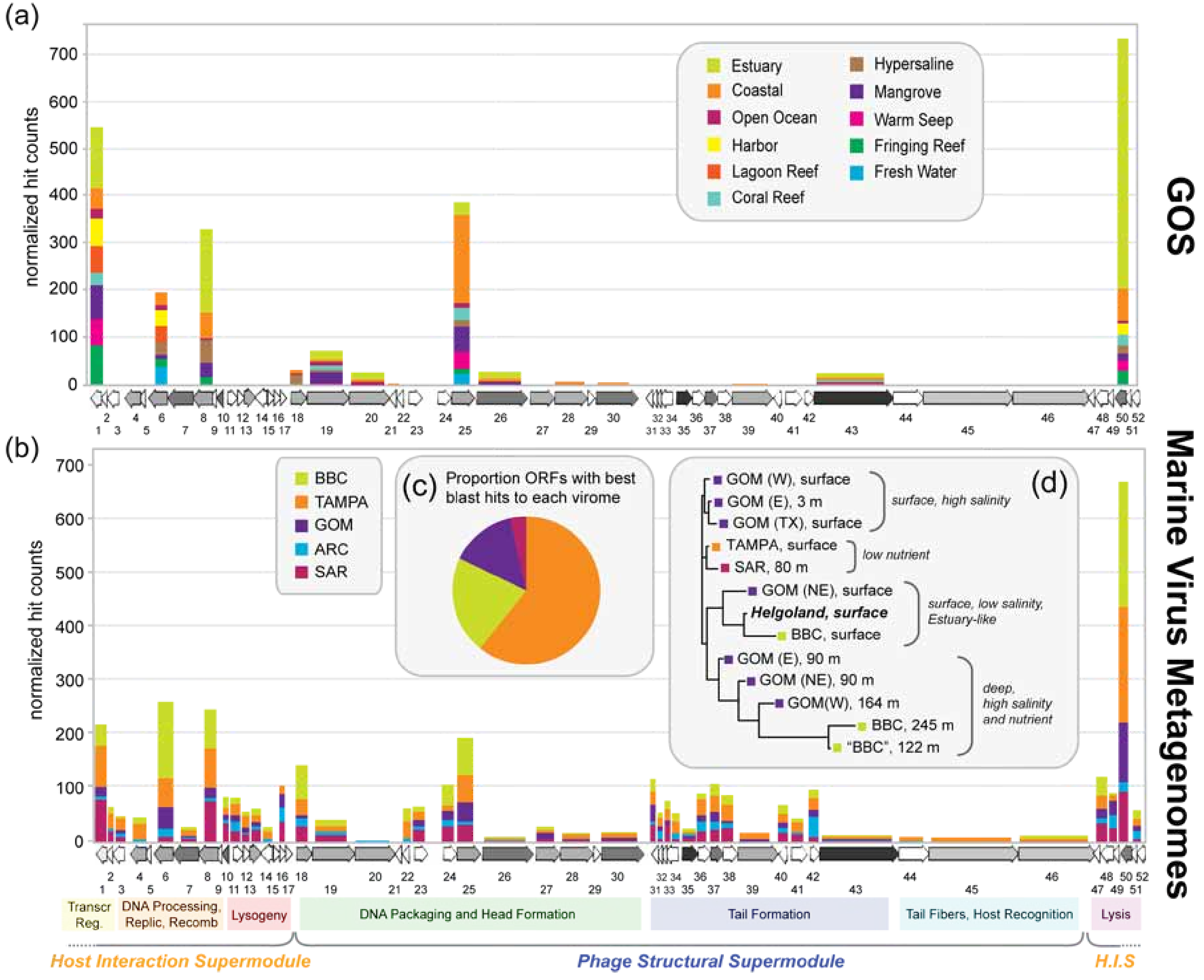

GOS
(a)
normalized hit counts
0
100
200
300
400
500
600
700
Estuary
Coastal
Open Ocean
Harbor
Lagoon Reef
Coral Reef
Hypersaline
Mangrove
Warm Seep
Fringing Reef
Fresh Water
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
Marine Virus Metagenomes
(b)
normalized hit counts
BBC
TAMPA
GOM
ARC
SAR
(c)
Proportion ORFs with best blast hits to each virome
(d)
GOM (W), surface
GOM (E), 3 m
GOM (TX), surface
surface, high salinity
TAMPA, surface
SAR, 80 m
low nutrient
GOM (NE), surface
Helgoland, surface
BBC, surface
surface, low salinity, Estuary-like
GOM (E), 90 m
GOM (NE), 90 m
GOM(W), 164 m
BBC, 245 m
"BBC", 122 m
deep, high salinity and nutrient
Transcr Reg.
DNA Processing, Replic, Recomb
Lysogeny
DNA Packaging and Head Formation
Tail Formation
Tail Fibers, Host Recognition
Lysis
Host Interaction Supermodule
Phage Structural Supermodule
H.I.S

Figure 4.5: Presence of Pseudoalteromonas phage H105/1 proteins in Global Ocean Sampling (GOS) microbial metagenomes and marine virus metagenomes (viromes). (a) Normalized tblastn hit counts to the GOS metagenomes grouped by habitat type, as defined in the original dataset. (b) Normalized tblastn hit counts to five marine virus metagenomes. (c) For each of the 44 ORFs with virome hits, the origin of the best blast hit (lowest e-value, greatest ORF coverage, and highest domain score) was tallied and depicted in the pie chart. (d) Neighbor-joining tree clustering virome sample sites based on their environmental parameters as interpolated by the megx.net GIS tools. Pooled samples that represent a range of depths were treated as independent sites; depths used for data interpolation are indicated with the sample name. Overall, sites cluster most strongly by salinity and depth; as such, Helgoland clusters with BBC and northeastern GOM sites, all of which have lower than the average salinity and are influenced by major river outflows. Note that data interpolation for Arctic sites is not possible, thus ARC is missing from this depiction. See Supplementary Material 7.4 for precise coordinates and depth used to retrieve interpolated data.

tion, as the host directly responds to its environment.

Also highly represented among the marine virome hits, in terms of a high number of hits per gene, the greatest number of genes along the entire genome, and the greatest overall similarity to Phage H105/1 ORFS (Figure 4.5c), is the Tampa Bay metagenome (McDaniel et al., 2008), which, through chemical induction, is enriched in temperate phages of its respective marine microbial community. This pattern extends evenly across the whole genome and reflects the intrinsic Phage H105/1 temperate/prophage signature also evident among NCBI-nr homologues (Figure 4.2b).

4.3.4 Painting a genome landscape: codon adaptation and tetranucleotide usage

As a large portion of the Phage H105/1 genes are novel, or have only very distant homology, few evolutionary relationships based on sequence similarity alone can be established to describe the history of the phage proteins and supermodules. Thus, taking an alignment-free approach, we investigate patterns of codon adaptation and tetranucleotide frequencies across the genome to look into the evolutionary history of Phage H105/1.

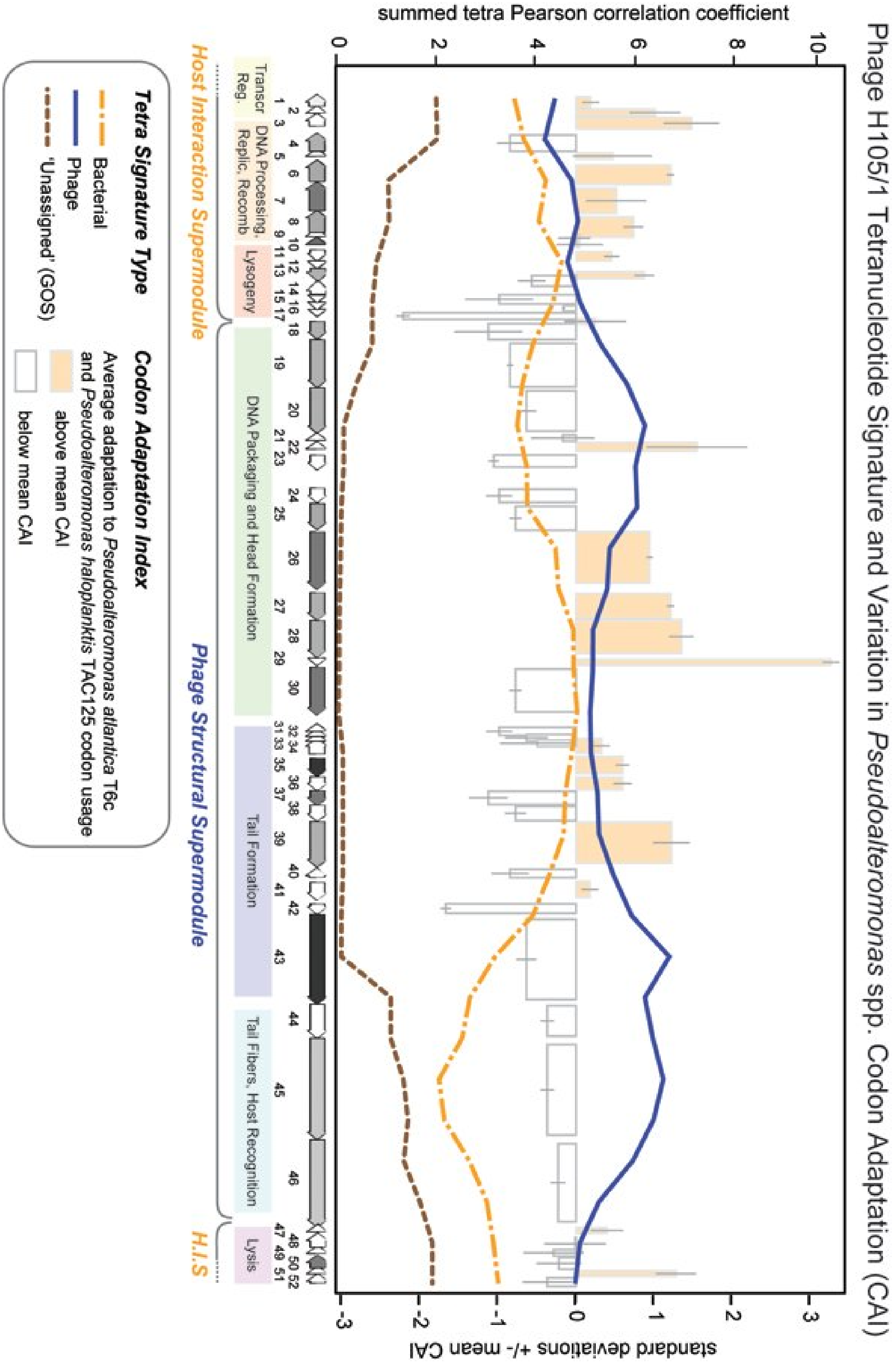
4.3.4.1 Host-Indexed CAI

Phage genomes are under codon-selective pressure imposed by the translational biases of their microbial hosts (Bahir et al., 2009; Carbone, 2008; Lucks et al., 2008). The Lambda phage genome ’landscape’ is sub-divided into peaks and valleys based on its Codon Adaptation Index (CAI), which reflects the adaptation of each gene to the codon bias of its host, *Escherichia coli* (Lucks et al., 2008). In the absence of a host genome sequence, Phage H105/1 codon adaptation index was calculated based on the preferred codon usage of two sequenced *Pseudoalteromonas* spp. (Figure 4.6). Previous studies have found that genes with the greatest host-indexed CAI (genes most resembling host codon bias) encode phage structural proteins, i.e., capsid (Carbone, 2008) and tail genes (Lucks et al., 2008). They presume that proteins made rapidly en mass during lytic growth most resemble codon usage of their host due to selection for translational efficiency, the fundamental force thought to drive codon bias in single-cell organisms (Sharp and Li, 1987). A similar pattern was found in select structural proteins of the Phage H105/1 head (ORFs 26-28) and tail (ORFs 34-36 and 39) formation modules (Figure 4.6). Carbone also found genes responsible for host interaction, inhibition of host functions, ssDNA binding, and transcriptional regulation to be strongly host biased (Carbone, 2008), which is also seen in the respective proteins of Phage H105/1's 'Host Interaction Supermodule' (Figure 4.6).

4.3.4.2 Tetranucleotide Frequency Correlation

We ask: "*given a portion of the H105/1 genome, which sequences (of the roughly 350,000 large fragments in GenBank) are most correlated based on their respective tetranucleotide frequencies?*" Tetranucleotide frequencies of phages have been shown to correlate with those of their hosts (Pride et al., 2006). However, when examined on a finer scale, it may be only certain portions of a phage genome that retain tetranucleotide frequencies of their host.

We found a predominately phage tetranucleotide (tetra) signature across the entire genome, with peaks and valleys that coincide with the differential codon adaptation (Figure 4.6). Regions with greatest 'host' codon adaptation have the greatest bacterial tetra signal, whereas regions of low adaptation peak in phage tetra signature (Figure 4.6; see Supplementary Information 7.3 for a description of the unassigned fragments). These patterns likely reflect alternative, mutually inclusive selective forces acting on different signatures. When codon usage is biased, codon adaptation reflects selection on mechanistic properties of efficient translation. Whereas tetranucleotide frequencies, though poorly understood, are likely influenced by (a) stochastic processes that accumulate through time (Pride et al., 2006), and (b) restriction-modification-related processes through the avoidance of restriction sites (Pride et al., 2003). However, codon usage and tetra frequency are inevitably intertwined through their coupled reliance on the same nucleotides. A convincing correlation between tetranucleotide frequencies and preferential codon usage has been observed in genomes assembled from environmental communities


Phage H105/1 Tetranucleotide Signature and Variation in *Pseudoalteromonas* spp. Codon Adaptation (CAI)
summed tetra Pearson correlation coefficient
0
2
4
6
8
10
standard deviations +/- mean CAI
-3
-2
-1
0
1
2
3
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 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
Transcr Reg.
DNA Processing, Replic. Recomb
Lysogeny
DNA Packaging and Head Formation
Tail Formation
Tail Fibers, Host Recognition
Lysis
Host Interaction Supermodule
Phage Structural Supermodule
H.I.S
Tetra Signature Type
Bacterial
Phage
'Unassigned' (GOS)
Codon Adaptation Index
Average adaptation to *Pseudoalteromonas atlantica* T6c and *Pseudoalteromonas haloplanktis* TAC125 codon usage
above mean CAI
below mean CAI

as well (Dick et al., 2009).

Figure 4.6: Pseudoalteromonas H105/1 Genome Signatures. The left axis denotes cumulative tetranucleotide correlations between 10 kb genome fragments of Phage H105/1 and large GenBank fragments, including the origin (bacterial, phage, or unassigned) of the GenBank fragment. The right axis quantifies the degree of codon adaptation around the mean for Phage H105/1 genes indexed to the sequenced Pseudoalteromonas spp. Proteins with codon usage more adapted to Pseudoalteromonas spp. bias have positive values and are labeled in orange. Error bars indicate the standard deviation of the two averaged Pseudoalteromonas spp. CAI. The relevance of UNA is described in Supplementary Information 1.

Phage proteins will remain 'associated with' a certain host by persisting among the new combinations of genes that make up phages infecting it, as gene flow predominately occurs between phages that infect the same host (Duffy and Turner, 2008). Guided by selection, some proteins of the combination will differ; some will remain the same. Though phage fitness is influenced by several factors on many levels (Duffy and Turner, 2008), we take advantage of the fact that codon adaptation is a selective force observed at the sequence level. As such, the Phage H105/1 proteins of greatest codon adaptation may be the proteins that are selected to remain 'associated' with its *Pseudoalteromonas* sp. host. As such, they have longer residence time with their host, and thus have the time to ameliorate a host/bacteria-like tetra signature (Pride et al., 2006). Whereas the non-adapted proteins are under less selective pressure to remain associated with a specific host, and, as many are structural proteins highly conserved in other phages (Table 4.1), may be more mobile components of a greater phage protein pool. In the absence of host amelioration, these proteins retain a phage tetra signature common to the phage pool. Thus, the bacterial/phage tetranucleotide pattern could reflect different stages of amelioration. However, little is known of how mutation rates differ in different portions of phage genomes (Duffy and Turner, 2008), nor how rates of swapping may differ between the phage functional modules.

4.4 Summary: Phage diversity and evolution in light of marine Phage H105/1

The ecogenomic and evolutionary influences on the Phage H105/1 genome content are highlighted by the phylogenetic signature of its functional annotations, the global distribution of its protein-coding genes, codon-adaptation, and tetranucleotide frequency correlations. These approaches (i) extend beyond the commonly searched public databases, (ii) take advantage of the invaluable environmental context of the sequenced organisms, a frequently neglected asset (Field, 2008) that, through integration with marine ecology, will shed light on the hidden pool of phage functional diversity, and (iii) are not restricted by limitations of sequence similarity. When integrated with experimental approaches (i.e., proteomic validation), such analyses will further enrich our ecological and evolutionary understanding of phage genomics, especially valid considering the drastic increase in marine phage genome sequences that will soon be available (Institute, 2010).

Acknowledgements MBD would like to thank Amelia Rotaru for SDS-PAGE assistance, Marianne Jacob for contributions to tetra analysis, Matthew B. Sullivan and Vincent Denef for critically reading the paper, and Matthias Ullrich, Renzo Kottmann and Ivaylo Kostadinov for fruitful discussions and support. A Marie Curie Early Stage Training Fellowship to MBD supports funding for this project (MEST-CT-2004-007776) through the MarMic program of the MPI for Marine Microbiology.

The authors declare no competing interests.

CHAPTER 5

Insights into infection dynamics of ocean cyanophage P-SS2: to integrate, or not to integrate?

Melissa Beth Duhaime[a,b], Frank Oliver Glöckner[a,b,*], Matthew B. Sullivan[c]

[a]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany
[b]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany
[c]University of Arizona, Ecology and Evolutionary Biology, Tucson, Arizona 85721, USA

Keywords: cyanophage / marine / Prochlorococcus / temperate

Abstract: Phages are implicated as being a major force in the "activity" of hypervariable regions and genomic islands in their hosts. The regions, which often contain genes mediating nitrogen, phosphate, and light stress response, are also thought to be important in providing niche-specific traits leading to the establishment and success of host ecotypes. This study investigates the infection dynamics of temperate marine siphovirus, P-SS2, showing its integration at five sites in such a genomic island of its host, low-light adapted *Prochlorococcus marinus* str. MIT9313. This is the first report of a marine phage integrating in *Prochlorococcus marinus*, despite numerous cyanophage and host isolates, including 12 sequenced strains (all of which lack prophages). As integration is thought to be influenced by environmental conditions, and the hypervariable region of P-SS2 integration is littered with domains related to nitrogen-stress response, the influence of varying environmental conditions (i.e., N and P-limitation) will be investigated through changes in integration dynamics and whole genome expression analysis of the phage and host during infection.

5.1 Introduction

Phages, virus that infect bacteria, are the most abundant biological entities on the planet and are responsible for an incredible degree of lateral gene transfer to and from their hosts (Canchaya et al., 2003), with the marine environment as no exception (Bryan et al., 2008; Coleman et al., 2006; Lindell et al., 2004; Sullivan et al., 2009). *Prochlorococcus* are dominant primary producers of the low nutrient, mid-latitude photic zone, and thought to be the most abundant photosynthetic organism on the planet (Partensky et al., 1999). If they are outnumbered by their viruses by an order of magnitude (Rohwer, 2003), *Prochlorococcus* phages are likely to be the most dominant cyanophages on the planet. These cyanophages carry metabolically relevant host-like *Auxiliary Metabolic Genes* ("AMGs"; reviewed in Breitbart et al., 2007), such as those central to photosynthesis, nucleotide scavenging, and vitamin synthesis, including many host homologues of unknown function (Mann et al., 2005, 2003; Rohwer et al., 2000; Sullivan et al., 2005, 2009). Through whole-genome expression analysis of both host and phage, Lindell et al. showed the expression of such phage-carried photosynthesis genes during infection of *Prochlorococcus* with a lytic cyanomyovirus, a process thought to supplement the compromised host photosystem, ensuring sustained host fitness for phage propagation.

Cyanophage P-SS2 (Sullivan et al., 2009), and its host, *Prochlorococcus marinus* str. MIT9313, a low-light ecotype, make up the first fully sequenced phage-host system of a marine siphovirus. Though P-SS2 lacks photosynthesis-related genes, it does contain several cyanobacterial homologues that are similarly thought to improve host fitness and efficient phage production during infection. These genes include those common to other phages (including DNA primase, rubonucleotide reductase), as well as unique to P-SS2 (*cobO* required for cobalamin synthesis gene, three conserved cyanobacterial hypothetical proteins, and two cyanobacterial ORFan genes, Sullivan et al., 2009). P-SS2 may further impose its influence on MIT9313, as there is convincing evidence it may integrate into this host chromosome, establishing lysogeny. The pair shares a 53-bp region of exact match, unique in all public sequence databases, including the Global Ocean Sampling (GOS) dataset, suggestive of a P-SS2 integration site (Sullivan et al., 2009). Within the same roughly 13 kb region of the MIT9313 genomic island where the large match is found, there are four additional 36 to 37 bp exact (or near exact) matches between the phage and host genomes. However, despite a great number of isolates, *Prochlorococcus* cyanophages are not known to integrate and, of the 12 sequenced *Prochlorococcus* strains, none are shown to have integrated prophages (Lima-Mendez et al., 2008a; Sullivan et al., 2009).

Little is know, even among model systems, about what influences a phage to either integrate of go immediately to the lytic cycle upon infection. In the model λ/*Escherichia coli* phage-host system, integration is dependent on host nutritional and energy status. When grown in high nutrient media, there is a greater incidence of λ integration than when *E. coli* is grown in minimal media, which contrarily promotes integration (Ptashne, 2004). This is thought to be a due to a high concentration of concomitant intracellular proteases in the fast growing cells, which degrade cII, a critical integration protein (Ptashne, 2004). In the absence of cII, the cascade of events leading to integration is not possible. Through such a mechanism, the phage ensures that the lytic cycle is pursued only when its host has sufficient resources to expend on the production of progeny phage. This energy/nutrient dependent establishment of lysogeny may extend to natural systems as a mechanism for phages to persist in low nutrient conditions. When phophate is limited in cyanophage-infected *Synecchococcus*, lysis can be delayed up to 18 hours, with an 80% decrease in host cell lysis, while phage adsorption kinetics remains constant (Wilson et al., 1998). Though indicating an environmental influence on virus activity, the mechanism of this delay is not resolved. Under nutrient restriction, are the phages more prone to integrating into a lysogenic state? Or are they persisting in stalled lysis (i.e., pseudolysogeny) until the phosphate status of their host improves (Miller and Day, 2008)? Though most abundant in the low light, high nutrient water of the deep photic zone, MIT9313 can maintain their abundance at midlatitudes where nutrient concentrations are comparatively lower (Johnson et al., 2006), and where P-SS2 was isolated (Supplementary Figure 7.7, Sullivan et al., 2009). Thus, nutrient fluctuations may be an important component of this system, especially considering the over-representation of mobile nitrogen stress-related domains and related promoters in the region of putative phage integration (Sullivan et al., 2009).

The driving questions of this study ask: does phage P-SS2 integrate into MIT9313 during infection, and if so, what is the affect of various host stressors, i.e., nutrient and energy (light) fluctuations, on the decision between lysis and integration? The current understanding of factors influencing the "molecular decision" to integrate is based solely upon a handful of well-characterized model systems, and certainly none as integral to the functioning of our oceans as a cyanobacterial phage-host system. As this is the first fully sequenced, potentially temperate, marine phage-host system, experimental manipulation combined with whole-genome expression analysis can offer insights into the environmental and physiological influences on lysogeny in one of most important primary producers of the world's oceans.

5.2 Methods

5.2.1 Host growth

Normal growth *Prochlorococcus marinus* str. MIT9313 was grown in acid washed flasks using Pro99 media, according to (Moore et al., 2007). Cultures were grown under constant light at 12-14 μmol Q/m^2·s and their growth monitored based on relative fluorescence (675 nm emission, 485 nm absorption).

5.2.2 MIT9313 infection with P-SS2

Infection conditions To monitor infection of MIT9313 by P-SS2, 5 ml of phage lysate (approximately 10^8 phage/ml) was added to 25 ml of freshly transferred (1:5) cells, allowed to adsorb for 1 hour, then diluted to 150 ml with additional Pro99 media. To ensure the phage was the agent preventing growth, in a control, the phage lysate was replaced by 0.02 μm-filtered lysate; in a second control, the phage lysate was replaced by Pro99 media. Growth was monitored, even past the point of culture decline.

Detection of Insertion Five putative insertion sites were predicted based on sequence similarity between MIT9313 and P-SS2 (Table 5.1; Sullivan et al., 2009). Based on the genome neighborhood of these sites in both phage and host, the hybrid sequences that would result from homologous recombination were modeled (Figure 5.1) and primers designed to detect these products by polymerase chain reaction (PCR; Table 5.2)

Table 5.1: The three longest regions of identical sequence stretches between host MIT9313 and phage P-SS2. Regions in host are hypothesized to be the phage integration sites (*att*). Integration was monitored by PCR-probing this region for hybrid junctures made at these sites due to homologous recombination, as described below.

Organism	*att* site name	Sequence (5′ to 3′)
MIT9313	AttB	GAT CTG CCC CTG TCT ATT CGG GCT CAT AAC CCG AAG GTC GGG AGT TCA AAT CT]
MIT9313	36A and 36B	TCG GGC TCA TAA CCC GAA GGT CGG GAG TTC AAA TCT
MIT9313	37A and 37B	TTC GTG CTC ATA ACC CGA AGG TCG GGA GTT CAA ATC T
P-SS2	attp	GAT CTG CCC CTG TCT ATT CGG GCT CAT AAC CCG AAG GTC GGG AGT TCA AAT CT

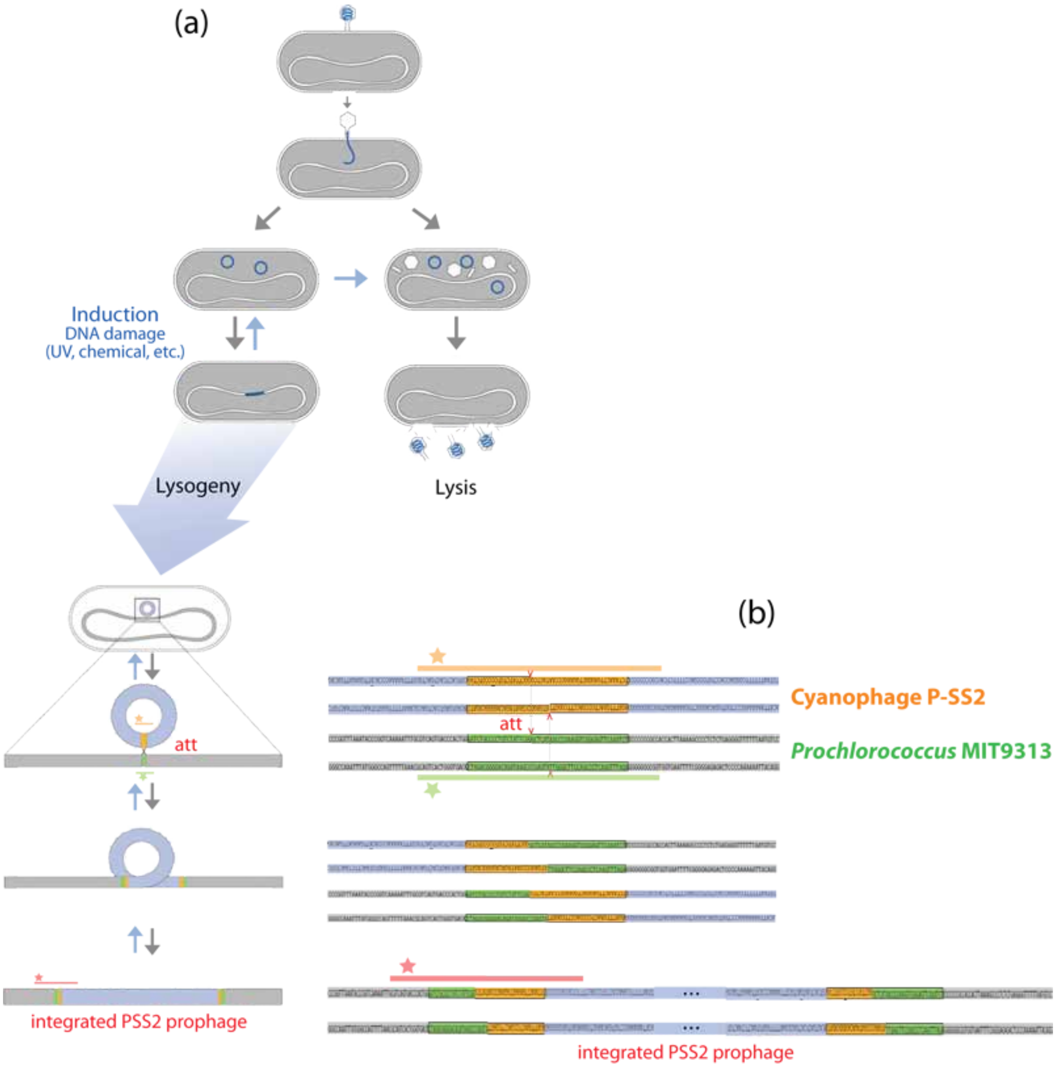


Figure 5.1: (a) Model of the molecular "decision" of a temperate phage to either integrate (lysogeny) or proceed with the lytic cycle (lysis); figure edited from Ptashne. (b) PCR-based assay to detect integration of phage P-SS2 in its MIT9313 host chromosome during lysogenic infection. *att* sites (red stars: AttB, 36A, and 36B), no integration across one host site (green star), and no integration across phage *att* (orange star); primers used are indicated above each lane and are as in Table 5.2. att: identical region between phage and host genome where homologous recombination facilitates phage integration (five such regions were identified in MIT9313: ’AttB’, ’36A’, ’36B’, ’37A’, ’37B’; one in P-SS2: ’attp’); orange line and star: PCR product indicating continuous phage *att* region (no phage integration); green line and star: PCR product indicating continuous host *att* region (no phage integration at this position); red line and star: PCR product indicating phage integration at targeted host *att* region (identifiable as hybrid of both phage and host *att* region)

The PCR reaction for this assay (and all subsequent non-quantitative PCR reactions) was carried out in standard 25 μl reactions with the following cycler parameters: 95° C, 3 min; 30x(95° C, 15 sec; 60.8° C, 30 sec; 72° C, 30 sec); 72° C, 5 min. Primers used are as indicated in Table 5.2 for "P-SS2 Integration assay."

Table 5.2: Primer sequences used for the P-SS2 integration assay, quantitative PCR (to monitor host and phage abundances via the qINT/qEXT assay), RT-PCR (to monitor expression of host stress response genes under N and P limitation). terL: large terminase gene; tf, tail fiber gene; mcp, major capsid protein gene; rbcL, ribulose-bisphosphate carboxylase gene; ntcA, nitrogen stress gene; pstS2, phosphate binding protein gene (1 of 2 copies in MIT9313); phoE, outer membrane phosphoporin protein gene

Primer	Target Organism	Primer Sequence (5′ to 3′)
P-SS2 INTEGRATION ASSAY, see Table 5.1		
attp_F265	P-SS2	AAG GTG CCA GGT ATC TGT GG
attp_qF	P-SS2	TTC TCT TTA GAT TCC CCA GA
attp_qR	P-SS2	AAT GGA AAC CCA TTT ACA GA
AttB_R355	MIT9313	GTA AGG GTG TAG CCA TTA GG
36A_R288	MIT9313	AGC GAT CAA AGT AGA AGA CG
36B_R299	MIT9313	CAA TAA CCC TCA GGA TCA GA
37A_R	MIT9313	CAG CCT CAG AGC TGT TAT TC
37B_R	MIT9313	CGA CCT TTG ATC ACT TCA AT
AttB_qF	MIT9313	GCC TTC CTG TCT GTT TTA GA
AttB_qR	MIT9313	GAC TGC GTG AAA ATA GGA AT
qINT/qEXT ASSAY		
terL_qF	P-SS2	CTT TTG GGA GAC ATA CGA AC
terL_qR	P-SS2	CTT GTA CCA ACC ATG GAG AG
tf_qF	P-SS2	AGA AGG TTG ACA TCC AAA CA
tf_qR	P-SS2	GTC GAG ACA CAA ACA GAA GC
mcp_qF	P-SS2	GAA TAC ACG ACA GGT TCA GG
mcp_qR	P-SS2	GAC ATC GGT CTT TAG TTT GG
rbcL_qF	MIT9313	AAC CTG TGT TAG CGG TAA AT
rbcL_qR	MIT9313	GCA AGA AAC AGG AGA AGT CA
NUTRIENT STRESS ASSAY		
ntcA_qF	MIT9313	GCT GCA GGG TCT CTC AAG TC
ntcA_qR	MIT9313	GCA GGT CAA TGG TGA TTC CT
phoE_qF	MIT9313	CTG AGC TTG AAA TAC CTG CT
phoE_qR	MIT9313	GAG GGC ATG TTC AAT AAT TC
pstS1_qF	MIT9313	TTC TGC TTT CTT CAC TGG TC
pstS1_qR	MIT9313	ACG GGT ATA GAT CTT TGC A
rbcL_qF	MIT9313	AAC CTG TGT TAG CGG TAA AT
rbcL_qR	MIT9313	GCA AGA AAC AGG AGA AGT CA

Nutrient-limited growth In an attempt to compromise the host metabolism (mimicing "minimal" conditions) and attain a range of P- and N-stress response, MIT9313 was grown under varying PO_4^{3-} and NH_4^+ concentrations. The range was designed to capture fine scale changes in lower concentrations relevant to MIT9313's ocean habitat (Supplementary Figure 7.7), rather than the 10-20x excess it sees in culture in Pro99 media. MIT9313 was grown normally (as detailed above) in full Pro99 media. At mid-exponential growth (typically day 7–8), cultures were pelletted (15 min, 9000 g, 22° C), washed twice with nutrient-free autoclaved seawater, subdivided and resuspended in a range of five variably amended P [50 (normal Pro99), 10, 2, 1, and 0 μM PO_4^{3-}] and N [800 (normal Pro99), 64, 32, 16, and 0 μM NH_4^+] Pro99 media.

5.2.3 Quantitative PCR (QPCR): qINT/qEXT

In order to track the parameters of phage P-SS2 infection (eclipse, latent, and rise periods, burst size), a QPCR assay to quantify intracellular (qINT) and extracellular (qEXT) viruses was designed[1], according to Lindell et al. Briefly, primers were designed to track increases in copy number of single copy genes of the phage (terL, tf, and mcp) and host (rbcL). qINT was determined by quantifying phage genes in an infected culture after 0.2 μm filtration and 2x rinsing of the filter; whereas qEXT was determined by quantifying phage genes present in this filtrate. QPCR was performed with the 2X QuantiTect SYBR Green kit (Qiagen, Valencia, CA; cat#204143) in 10 μl triplicate reactions and primers at a final concentration of 500 nM. QPCR was performed on a BioRad iCycler (Bio-Rad Laboratories, Hercules, CA), followed by an 80-cycle melt-curve analysis: [95° C, 15 min; 40x(94° C, 15 sec; 60.8° C, 30 sec; 72° C, 30 sec); 72° C, 5 min; 80x(55° C, 10 sec) increasing 0.5° with each cycle]. Plasmid clones of all target genes were generated to serve as QPCR standards (see Table 5.2 for primer sequences). Products were PCR-amplified from MIT9313 cultures or P-SS2 lysates, ligated into pGEM-T Easy vectors, and transformed into competent *Escherichia coli* JM109 cells according to manufacturer's protocol (Promega Corp., Madison, WI). Plasmids of successful clones were isolated with the BioRad MiniPrep Kit and digested with single-cut restriction endonuclease, PstI (Roche, cat#10621625001) in 25 μl reactions according to manufacturer's instructions, with the exception of not diluting the enzyme. For complete digestion, 10 U/μl enzyme for 1 hour at 37° C was needed, as evidenced

[1] These parameters are *essential* to understand the biology underlying any phage-host system. However, lack of virus filters (due to severe manufacturer issues, still not resolved) during the course of this experiment prevented conclusive measurements. Though time prevented conclusive quantitative PCR results as of yet, the experimental design and current progress will be discussed.

by a single band observed on a 1% (wt/vol) agarose gel. Digested plasmids were quantified with the Quant-iT™PicoGreen dsDNA assay (Invitrogen Corp., Carlsbad, CA; cat#18064-022) and standards ranging from one to 1x10^6 copies per reaction were prepared in triplicate for each QPCR run. Samples were incubated at 95° C for 15 minutes to deactivate DNases and lyse the cells prior to QPCR reactions (per Zinser et al., 2006).

5.2.4 Quantifying host nutrient stress through gene expression

To monitor acute physiological response of MIT9313 to nutrient stress, expression analysis was designed.[2]

RNA extraction and Reverse Transcriptase Quantitative PCR (RT-QPCR) Based on the genes significantly upregulated in the expression profile of MIT9313 under nutrient stress, the ntcA gene was chosen as a proxy for nitrogen stress (Tolonen et al., 2006), and pstS2 and phoE genes for phosphate stress (Martiny et al., 2006), to be monitored in RT-QPCR. The rbcL gene served as a single copy gene to normalize the expression. Cells harvested during the nutrient-limited growth experiments were flash-frozen and stored at -80° C until RNA extraction. RNA was extracted using the *mir*Vana RNA isolation kit (Applied Biosystems/Ambion, Austin, TX; cat#1560) and DNA removed (DNase I) with the Turbo DNA-*free* kit (Applied Biosystems/Ambion, Austin, TX; cat#1907). The reverse transcriptase reaction was carried out using Superscript II reverse transcriptase (Invitrogen Corp., Carlsbad, CA; cat#18064-022), including reactions lacking SuperScript II as no RT controls. RNA was detected on an Agilent 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA). QPCR of the cDNA was performed as described above.

5.3 Results and troubleshooting

5.3.1 MIT9313 Growth, Infection and P-SS2 Integration

A typical uninfected culture reached exponential growth after a five to seven day lag phase, with fluorescence decreasing after 12–14 days, as the photosystems of the compromised cells lost efficacy (red curve, Figure 5.2). Infection of MIT9313 cultures with P-SS2 prevented normal host growth (gold curve, Figure 5.2). However, after nearly 20 days, infected cultures showed evidence of growth. This is

[2] Time prohibited conclusive quantitative RT-PCR results; experimental concept, design and progress are presented.

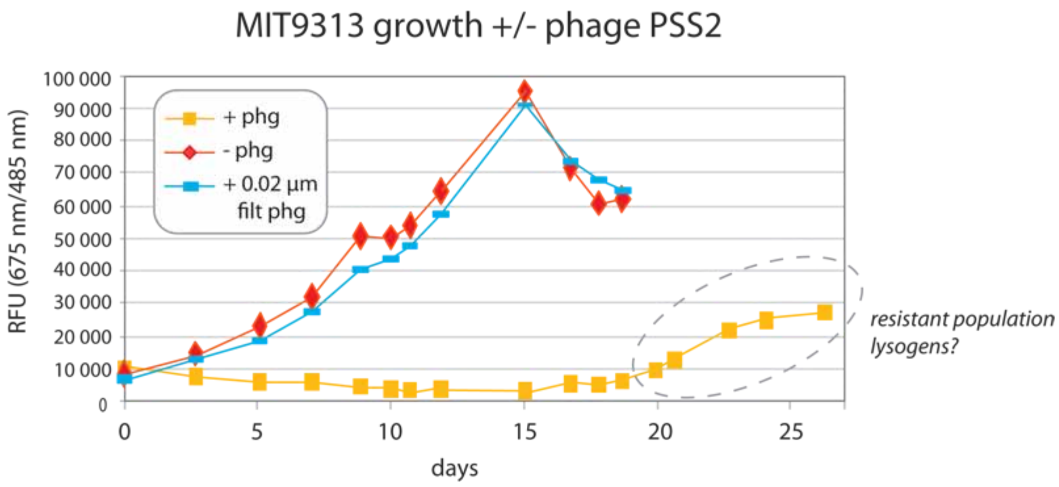


Figure 5.2: Growth of *Prochlorococcus marinus* str. MIT9313 infected with phage P-SS2 and filtered P-SS2 (indicating the phage is the responsible agent for lack of growth). The seemingly resistant population that develops 20 days post-infection is thought to be enriched in lysogens; this culture is used to screen for P-SS2 integration.

presumed to be a P-SS2-resistent population, some portion of which may be experiencing homoimmunity caused by phage integration. As such, this population was probed with the integration assay.

Primers designed to detect phage P-SS2 integration in infected MIT9313 cultures indicate integration products at five positions in the host genome (red stars, Figure 5.3).

5.3.2 Quantifying host and phage through single-copy gene detection

MIT9313 gene quantification QPCR amplification of the host rbcL gene as a proxy for MIT9313 abundance was, thus far, the most robust. Based on standard curve generation of plasmids with known gene copy number, a 0.996 correlation coefficient was achieved, 95.9% PCR efficiency, and melt-curve analysis showed the absence of unspecific amplification. However, cultures were calculated to be on average 2–6x10^6 cells/ml, which is low for a culture (typically 10^{7-8} cells/ml), but this measurement was made at a time when the culture had inadvertently been exposed to long period of low light. Obvious validation of this figure with flow cytometry and microscopy-based counts is required.

P-SS2 gene quantification Currently there is no amplification of phage genes (neither large terminase, tail fiber nor major capsid protein) from lysates in QPCR, though they are expected to be present at 2x10^8/ml, based on microscopy counts (data provided by J. Brum). However, amplification of terL, tf, and mcp plas-

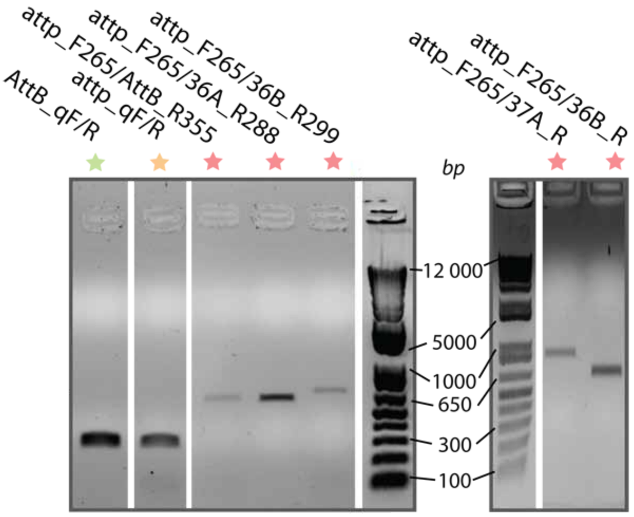


Figure 5.3: PCR products detecting integration across five host *att* sites (red stars: AttB, 36A, 36B, 37A, 37B), no integration across host *AttB* (green star), and no integration across phage *attp* (orange star); primers used are indicated above each lane, and are described in Table 5.2. Stars indicate PCR products, as modeled in Figure 5.1.

mid standards is successful, though not entirely efficient. PCR efficiencies were on the range of 95%, though 55-63% when quantification of the lowest concentration standards was not possible. To test the source of this poor standard amplification and lack of lysate amplification, the QPCR reaction (identical master mix, template, and reaction cycle) was performed in a standard, non-quantitative cycler; again, only the the standard amplified, and even this resulted in a relatively weak band, as visualized on an 1% (wt/vol) agarose gel. The reaction was again performed with the addition of Takara Taq™polymerase (TaKaRa Bio Inc., Otsu, Japan), which led to successful amplification of all samples. This implicates the Quanti-Tect SYBR Taq as being the cause of no amplification from viral lysates, rather than an issue with the viruses themselves. The supplementation of Takara Taq to QPCR reactions, as well as alternatives to SYBR Taq, will be considered in future trials.

5.3.3 Manipulating host MIT9313 stress response

Induction of host nutrient stress response There is a resolved biological stress response by MIT9313 to both PO_4^{3-} and NH_4^+ limitation (Figure 5.4). The culture crashes at different time points depending upon the concentration of available P and N.

Furthermore, this response was remarkably resolved on a scale relevant to nutrient levels seen in the ocean, with differences observed at even the lowest concentrations (Figure 5.4). In the upper photic zone of the Eastern US, in the proximity of where the phage and host were isolated, PO_4^{3-} can range from 0.4–1.4 μM (as interpolated by megx.net GIS tools, see Supplementary Figure 7.7; Kottmann et al., 2010); unfortunately, NH_4^+ data is not available for interpolations. However, in

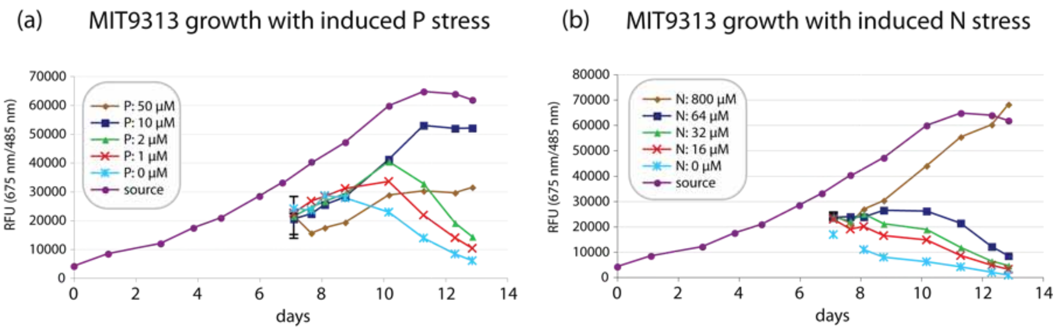


Figure 5.4: Growth curves of MIT9313 grown under varying PO_4^{3-} and NH_4^+ concentrations.

the open ocean, including shelf and coastal waters, ammonium is typically found at nannomolar concentrations, with the exception being at the base of the photic zone, where MIT9313 is typically found. In this zone, NH_4^+ can be 4–5 μM in highly productive areas, though a 1 μM ammonium max is more typical of meso- and oligotrophic areas (G. Lavik, personal communication)[3].

Detection of mRNA: on the way to "qStress" For a molecular approach to monitor this stress at the transcriptional level, a "qStress" assay was designed based on the results of Martiny et al. (2006) and Tolonen et al. (2006). Isolation of RNA following lysozyme treatment of MIT9313 cultures proved more efficient than without lysozyme (lanes 1 and 6, Figure 5.5). However, prior to amplification, no RNA could be detected from experimental samples (lanes 2-5, Figure 5.5), which may be due to the low cell concentration (10^{5-6} cells total); fluorometric RNA detection, i.e., RiboGreen RNA Quantitation (Invitrogen Corp., Carlsbad, CA), may be more successful at quantifying these low concentrations. Though, even in the absence of visualization, subsequent amplification during QPCR may be possible nonetheless.

5.4 Outlook and Future Work

5.4.1 Integration

Though more than half of the sequenced microbial genomes contain prophages, and of those, they harbor an average of almost three per genome (Supplementary Table 6.1), prophages are lacking from nearly all cyanobacterial genomes (Supplementary Figure 7.1). Of the 12 sequenced *Prochlorococcus* strains, none contain

[3] This suggests ammonium concentrations in the nutrient stress experiments could be even lower in order to model natural conditions.

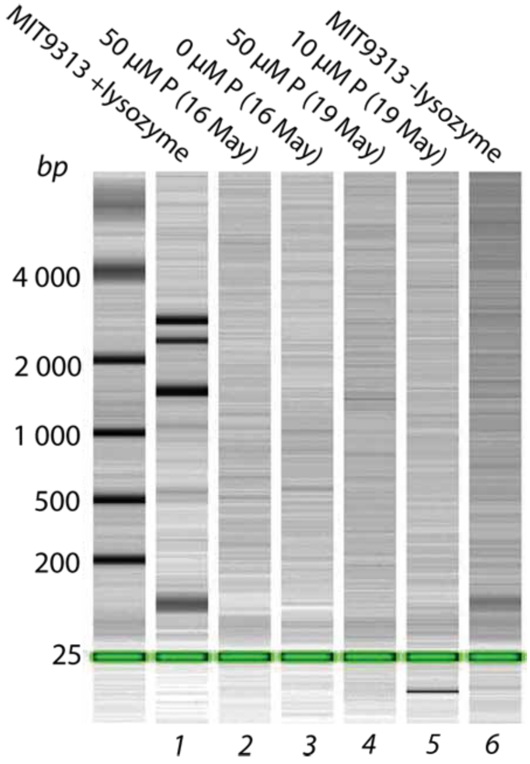


Figure 5.5: RNA as detected by Agilent Bio-Analyzer following *mir*Vana isolation and Superscript II reverse transcription. Lanes 1 and 6 represent contain MIT9313 sample with and without lysozyme, showing the effectiveness of this treatment in isolating RNA.

integrated prophages, and only three of the 11 sequenced *Synechococcus* genomes, *Synechococcus elongatus* strains PCC7942 and PCC6301 and *Synechococcus* sp. RCC307, are predicted to contain prophages (Lima-Mendez et al., 2008a). This is the first report of a marine phage integrating in a *Prochlorococcus* host, and its validation inspires continued investigation of the more complex dynamics of this systems, as is embedded in the future of the current study.

5.4.2 Towards understanding the influence of nutrients on infection and integration

Whole-genome transcriptional response of *Prochlorococcus marinus* str. MIT9313 to phosphate and nitrogen deprivation have been investigated, and the findings are crucial to the design of this experiment, particularly related to the timing of stress response. For instance, 24 hours post-infection, 33 MIT9313 genes were up-regulated and 143 down-regulated (1.4% and 6.1 of all genes in the genome; Martiny et al., 2006). Specifically, transcript levels of pstS2 (a periplasmic P-binding protein expressed under P-starvation in cyanobacteria), a gene chosen to monitor P-stress in this study, increased 50-fold in 24 hours, then diminished (Martiny et al., 2006).

Tolonen et al. (2006) found that MIT9313's transcriptional response to N-starvation (imposed as NH_4^+ starvation) occurred more rapidly than the physiological response, with differential expression occurring six hours post-starvation. This merits the use of the "qStress" assay to monitor host stress, rather than simply changes in culture growth, in search of influences on phage infection dynamics. The number of differentially expressed genes (120 up-regulated and 251 down-regulated) continued to increase through the entire 48 hour experiment (Tolonen et al., 2006). This is a nearly 4-fold increase in the number of up-regulated genes, and double the number of down-regluated, as compared to P-stress. Among these are four nif11-domain-containing genes found in the hypervariable, insertion sequence (IS)-rich region where P-SS2 is now known to integrate (Figure 5.3, Sullivan et al., 2009). Whether P-SS2 integration "behavior" correlates to the activity of this region, especially under nitrogen stress, is not yet known.

These studies investigated the response of *Prochlorococcus marinus* ecotypes to nutrient stress in a "binary" fashion: phosphate (Martiny et al., 2006) or nitrogen (Tolonen et al., 2006) available in excess, or none at all. Whereas, the nutrient concentration-dependent temporal resolution seen in the current study (Figure 5.4) is crucial as a "handle" for strategic manipulation of MIT9313 along a stress gradient.

5.4.3 The influence of AMG expression in phage-host systems

The role of cyanophage auxiliary metabolic genes, with particular focus on those involved in photosynthesis, has provided intriguing insights. During lytic infection of *Prochlorococcus* MED4 with T7-like cyanophage P-SSP7, Lindell et al. (2007) found four AMGs–those encoding photosystem II D1, high-light inducible protein, transaldolase, and ribnucleotide reductase–to be expressed in a cluster with genes typical of phage DNA replication. Based on this observation, this mechanism is thought to supplement host energy and nucleotide production during phage replication (Lindell et al., 2007). The expression cluster profile of P-SS2 through infection will offer insights into the function and influence of its own AMGs, with particular interest in *cobO* (required for cobalamin/B12 synthesis, an often limiting marine micronutrient with a role in ribonucleotide reduction during DNA synthesis; Panzeca et al., 2009) and P-SS2's five conserved cyanobacterial hypothetical proteins of unknown function.

5.4.4 Underlying phage biology and other considerations

One-step growth curves Progress thus far was hindered by the consequential inability to reliably quantify viruses[4]. Three methods at hand, microscopy, QPCR (i.e., qINT/qEXT assay), and flow cytometry will continue to be optimized, as they have been successfully applied to count phages in a reliable and correlated manner (Lindell et al., 2007). Quantification will grant the ability to assess the dynamics of P-SS2 infection with a "one-step" growth curve, discerning burst size and rates of different stages of the lytic cycle (Figure 1.8). Such knowledge will allow necessary fine-tuning of the infection experiments.

Benefits of axenic culture Microscopy results of MIT9313 culture strongly suggest the culture is not axenic (data not shown). Previous reports have indicated that *Prochlorococcus* cultures frequently contain "helper" heterotrophic bacteria, hypothesized to reduce *Prochlorococcus* oxidative stress, which can be eliminated by antibiotics (Morris et al., 2008). As the effects of a contaminating organism can be neither modulated nor quantified in the proposed expression and infection experiments, methods of obtaining an axenic MIT9313 culture may be requisite.

Light The cell cycle of cyanobacteria, including *Prochlorococcus*, are known to follow a diel cycle in phase with daily light patterns. In the environment, *Prochlorococcus* DNA replication occurs in the afternoon, while cell division takes place in the early evening (Vaulot et al., 1995). In *Synechococcus*, this mechanism relies on a finely tuned "circadian clock", based on the interaction of the *kaiA*, *kaiB*, and *kaiC* gene products (Mullineaux and Stanewsky, 2009). A similar system, perhaps simpler, is thought to influence the cell cycle of *Prochlorococcus* (Axmann et al., 2009). Infection of cyanobacteria also cycles in a diel fashion. However, myovirus infection in *kai* mutants does not differ from infected wild type *Synechococcus* (Kao et al., 2005). How is the diel infection thus regulated or established? Adsorption of phage to *Synechococcus* cells appears to be greater under light conditions than dark (Kao et al., 2005), and in *Anacytis nidulans*, phage adsorption is know to be regulated by both light and sodium ions (Cséke and Farkas, 1979). Thus, infection is affected more by the *availability* of light, than by the mechanisms of the host circadian cycle. This suggests a direct reliance on the primary effects of light on their host, such as initiation of photosynthesis (which may cause altered cell surface potential affecting phage adsorption), or discreet stages of the cell cycle (e.g., phage infection occurs only during host DNA replication), rather than the "pre-

[4] As mentioned previously, this was due to methodological limitations now partially resolved.

programmed" internal clock of their host. Such a synchronized host cell cycle is lost during growth under continuous light. However, it may *a*) serve as a useful mechanism to synchronize MIT9313 infection for a maximally resolved response to P-SS2, and *b*) be an additional influence on the molecular "decision" of P-SS2 to integrate.

Outlook

Evidence for P-SS2 integration and the resolved response of MIT9313 to a range of nutrient stress mark the current success of this project, thus far. Barring a better grasp of P-SS2/MIT9313 infection dynamics (namely, one-step phage "growth" curves by microscopy and the qINT/qEXT assay), the influence of nutrient stress on promoting prophage integration can now be investigated.

Acknowledgements

MBD wishes to acknowledge the Tucson Marine Phage Lab for endless knowledge-bestowing conversations and enthusiastic support, as well as an incredible opportunity to be immersed in an expert phage lab. Particularly important was the guiding hand of Bonnie Poulos in offering expertise in many of the methods presented and for maintaining the MIT9313 culture in periods of "project hibernation." A Marie Curie Early Stage Training Fellowship to MBD supports funding for this project (MEST-CT-2004-007776) through the MarMic program of the MPI for Marine Microbiology.

CHAPTER 6

Summary and Outlook

SUMMARY

As introduced in Chapter 1, marine viruses are known to influence the global cycling of matter, the diversity of their host communities, and the transfer of genomic material between microbes. This thesis took a bioinformatic approach to expand and address the extent to which genomics offers insights into processes of phage ecology and evolution in the marine system by:

1. contributing to the development of a portal for marine ecological genomics (http://www.megx.net) and supplementing the megx.net database with the addition of marine phages (Chapter 2; Kottmann et al., 2010),

2. working with members of the Genome Standards Consortium (GSC) to develop appropriate standards and platforms to describe phages, such as the viral component of the MIGS/MIMS standards (Field et al., 2008) and validation of GCDML (Kottmann et al., 2008) through its use in

3. creating the first set of MIGS-compliant genomes: the marine phage genome collection (Chapter 3; Duhaime et al., 2010a),

4. being the first study to make use of megx.net interpolated environmental data paired with (viral) genomic data to make ecological inferences ("ecogenomics") about a marine phage from Helgoland (Chapter 4; Duhaime et al., 2010b),

5. applying a novel application of genome signatures (tetranucleotide frequencies and codon adaptation) as an approach to better infer biological knowledge from phage genomes, a non-trivial pursuit considering the the diverse and dynamic evolution of phages (Chapter 4; Duhaime et al., 2010b), and

6. setting the stage to approach phage ecology in the lab system through the development of a marine host-phage model system to investigate the role of the environment in phage infection dynamics (Chapter 5).

Thesis in the Big Picture and Future Perspectives

"There are many hypotheses in science which are wrong. That's perfectly alright; they're the aperture to finding out what's right." - Carl Sagan, astronomer and science writer

6.1 A platform for marine virus genomics: megx.net and standards development

The work of this thesis was integral towards the development of a platform focused on marine virus genomics, and is the first of its kind. This platform took shape from two major efforts: the integration of marine phage genomes in megx.net (Chapter 2; Kottmann et al., 2010) and extensive manual curation of marine phage contextual data (Chapter 3; Duhaime et al., 2010a). The commitment to these efforts and their dire relevance to the greater research community is reflected in the continued development of these initiatives.

6.1.1 Future virus extensions to megx.net

Phages represent the mobile genomic extensions of their microbial hosts. They drastically influence both host genomic diversity and function in the oceans (as introduced in section 1.1.3). In order to facilitate marine microbial genomics, the addition of marine phages was an essential measure in the development of megx.net (Chapter 2; Kottmann et al., 2010). In order for megx.net to provide the most holistic marine virus genome collection to the scientific community, two additional sources of sequence data will be considered in future development: the marine virus metagenomes and prophages.

Marine metagenomes At last count, there were estimates of 11 sequenced marine phage metagenomes, and another eight non-marine or organism-associated habitats (Appendix section 7.1). These viromes represent over 500 Mb of viral diversity from diverse habitats that is currently missing from the megx.net portal. Future efforts will extend the megx.net viral component by including these viromes, including the generation of MIMS-compliant GCDML reports describing their contextual data (Field et al., 2008), analogous to the MIGS reports generated for marine phage genomes in Chapter 3 (Duhaime et al., 2010a). Such a feature will allow megx.net users to conduct gene-centric queries (e.g., Blast) of the marine virus metagenome collection in light of their habitat and environmental parameters, such as those performed to shed light on the ecogenomic patterns of

Pseudoalteromonas phage H105/1 (Chapter 4; Duhaime et al., 2010b).

Marine prophages There is strong evidence that the ocean environment exerts selective pressure on the gene content of marine phages, some of which is evidenced by a degree of "marine-ness" in their gene content (Figure 4.2; Duhaime et al., 2010b). Our ability to make such deductions, however, is limited by the small number of marine phages currently sequenced (Chapter 3; Duhaime et al., 2010a). There are nearly 2000 sequenced microbes in megx.net. It is likely that more than half of these contain at least one integrated prophage, and probably several. These sequenced prophages contain information pertaining to the adaptation of phage to the marine system, which, if made more accessible, could be used to generate ecogenomic hypotheses. For instance, the most similar sequenced entity to marine Pseudoalteromonas phage H105/1 from Helgoland, was an integrated prophage from a sequenced marine *Silicibacter* sp. (Chapter 4; Duhaime et al., 2010b). Casting a wider net in our collection of marine phage proteins will enhance our quest for such patterns that reflect phage adaptation to the ocean environment. Thus, to maximally expand the megx.net marine virus sequence collection, the prophage genome sequences will be included in future releases of megx.net.

Preliminary surveys of the marine prophages are already underway. In its last release (Lima-Mendez et al., 2008a), the Prophinder prophage database contained predictions of 1107 prophages, 56 of which are marine (nearly twice the number of marine phages sequenced; Table 6.1). The taxonomic classification of hosts containing prophages was similar to that of hosts of the sequenced marine phages (Figures 7.1a and 3.3, respectively; Duhaime et al., 2010a). Prophages tend to be smaller than sequenced phage isolates (Figure 7.1b), which may be due to muta-

Table 6.1: Statistics describing the prophages, including those identified in marine microbial sequences, found in the Profinder database.

All prophages	
Total number of prophages predicted by Prophinder:	1107
Number of microbial genomes searched:	727 (51.7 %)
Number of microbial genomes containing at least one prophage:	376
Avg. number of prophages per ’prophage-containing’ microbe:	2.94
Marine prophages	
Number of marine microbes containing prophages:	27 (41.5 %)
Number of marine prophages predicted:	56
Avg. number of prophage per marine ’prophage-containing’ microbe:	2.07

tions and partial degradation of their genomes, rendering them no longer able to excise as a functional phage.

An additional benefit of including prophages within the auspices of megx.net is that the host taxonomy, genome sequence, and habitats are known *a priori*. This allows further speculation about the role of context in phage gene content and habitat specificity (with the caveat that genes they contain may no longer be functional, or their own). A fresh prediction and identification of prophages in the megx.net microbial genomes will make this possible.

6.1.2 Maintaining robust marine virus contextual data

The development and implementation of genome standards in an ongoing task. Yet, as our sequence databases grow, the need for standards becomes more acute. The efforts invested to create MIGS-compliant marine phage reports (Chapter 3; Duhaime et al., 2010a) will be continued to ensure future phage sequencing initiatives uphold this indispensable standard. In 2009, the Gordon and Betty Moore Foundation, together with the Broad Institute, announced the sequencing of 200 marine phage genomes and 50 viral metagenomes (http://www.broadinstitute.org/annotation/viral/Phage/Home.html). This will increase the current collection of marine phage genomes 7-fold, and viromes nearly 5-fold. This large input of sequence data from a centralized source extends itself to further imposing the MIGS/MIMS standards. Some degree of manual effort will be required (and is planned) to ensure these standards are able to fully capture the contextual data of these projects, but such an investment will make MIGS/MIMS more robust for future phage sequence data.

6.2 Transforming sequence data into knowledge through ecogenomics

The development of a platform to examine the collection of marine virus sequence data in its environmental context has enabled the generation of ecogenomic hypotheses (Chapter 4; Duhaime et al., 2010b) and the development of lab-based experiments (Chapter 5), for future translation into biological knowledge (Figure 1.11).

6.2.1 Contextual insights into genome content: Phage H105/1

The analysis of marine Pseudoalteromonas phage H105/1 was greatly enhanced by the ability to consider the habitat and environmental context (via megx.net interpolated physicochemical data) of both Phage H105/1 and the marine phage genomes and virus metagenomes to which it was compared. Through this process, ecogenomic patterns of Pseudoalteromonas phage H105/1 emerged, reflecting, for instance, its low-salinity habitat and temperate replication strategy (Chapter 4; Duhaime et al., 2010b). With the addition of marine phage and prophage genomes and viral metagenomes, the megx.net platform serves as a resource enabling future researchers to answer such questions as new phage genomes are sequenced.

Additional results of this single genome analysis further serve as a springboard for questions central to phage ecology and evolution that can be enhanced through comparative analyses.

Single genome speculations. The fluctuation in codon adaptation and tetranucleotide signatures across the Pseudoalteromonas phage H105/1 genome gave rise to the hypothesis of a 'host-associated' gene pool, which, in its mobility, operates differently than the 'phage-specific' gene pool (Section 4.3.4; Duhaime et al., 2010b). In this model, *Pseudoalteromonas* codon adaptation bias and 'bacterial' tetranucleotide signatures were used to identify genes hypothesized to be more 'host-associated.' The *least* 'host' codon-adapted genes, in regions where the 'phage' tetranucleotide signature peaks, are conversely thought to represent a 'phage' gene pool. For such a mechanism to exist, there must be differential exchange of genes and modules. This represents one of the greatest outstanding questions concerning phage evolution: at what rates are new recombinant alleles created and shuffled within phage groups (Casjens, 2005)?

As 'experimental' phage are assembled by novel combinations of genes through recombination, some of the proteins they encode may be better adapted to a specific host (through codon adaptation, nucleotide signatures, supplementing hosts with metabolic genes, etc.). Such proteins will remain 'associated with' a certain host by persisting among the new combinations of genes that make up phages infecting it, as gene flow predominately occurs between phages that infect the same host (Duffy and Turner, 2008). Similarly, proteins not host-adapted (e.g., lacking host codon adaptation, or other traits a specific host association may influence) will appear as more promiscuous components of the 'phage-specific' gene pool. In this way, they will be seen in new recombinants at higher rates than 'host-associated' phage genes.

These conjectures are based on observations seen in a single genome from Helgoland. As such, it is not possible to speculate about rates of exchange or any variation in sequence amelioration or codon adaptation through time. A larger dataset of associated phage genomes, covering a range of genome and host range-relatedness (from nearly identical to more divergent), would be a gold-mine in providing a more robust view on this mechanism, or (equally as valuable) invalidating it altogether.

Towards comparative genomics of the Helgoland phages. In fact, such a "gold-mine" collection of marine phage-host systems already exists, and the dynamic infection ranges of these 22 phages and over 50 host strains has been determined (Wichels et al., 1998). This historic collection of Helgoland phages, well known to the marine phage research community, is part of a thirty year exploration of North Sea and Atlantic marine phages (Frank and Moebus, 1987; Moebus, 1980, 1983, 1991, 1992, 1997; Moebus and Nattkemper, 1981; Wichels et al., 2002), the first sequenced genome of which was described in this thesis (Chapter 4; Duhaime et al., 2010b).

Efforts are already underway to determine, after nearly a decade of senescence, which phages are still infective and which hosts are still viable. With such information, phages, and perhaps hosts, will be strategically selected for sequencing in order to examine the extent of lateral transfer between (Figure 6.1):

1. phages of different families (phages 13-15b, 11-68c, and 10-94a) infecting the same host (strain 13-15)
2. phages of the same family (phages H103/1, H105/1, H108/1) infecting the the same host (strain H105)
3. phages of the same family able to infect different hosts (numerous examples)
4. and all permutations of the above.

The collection also could offer insights into the genomic basis of host ranges among closely related *Pseudoalteromonas* spp. strains

1. within phage families (e.g., within myophages, such as the wide host-range H71/1 versus narrow range H106/1)
2. and between phage families (e.g., the large number of wide range myophages versus the narrow range sipho- and podophages),

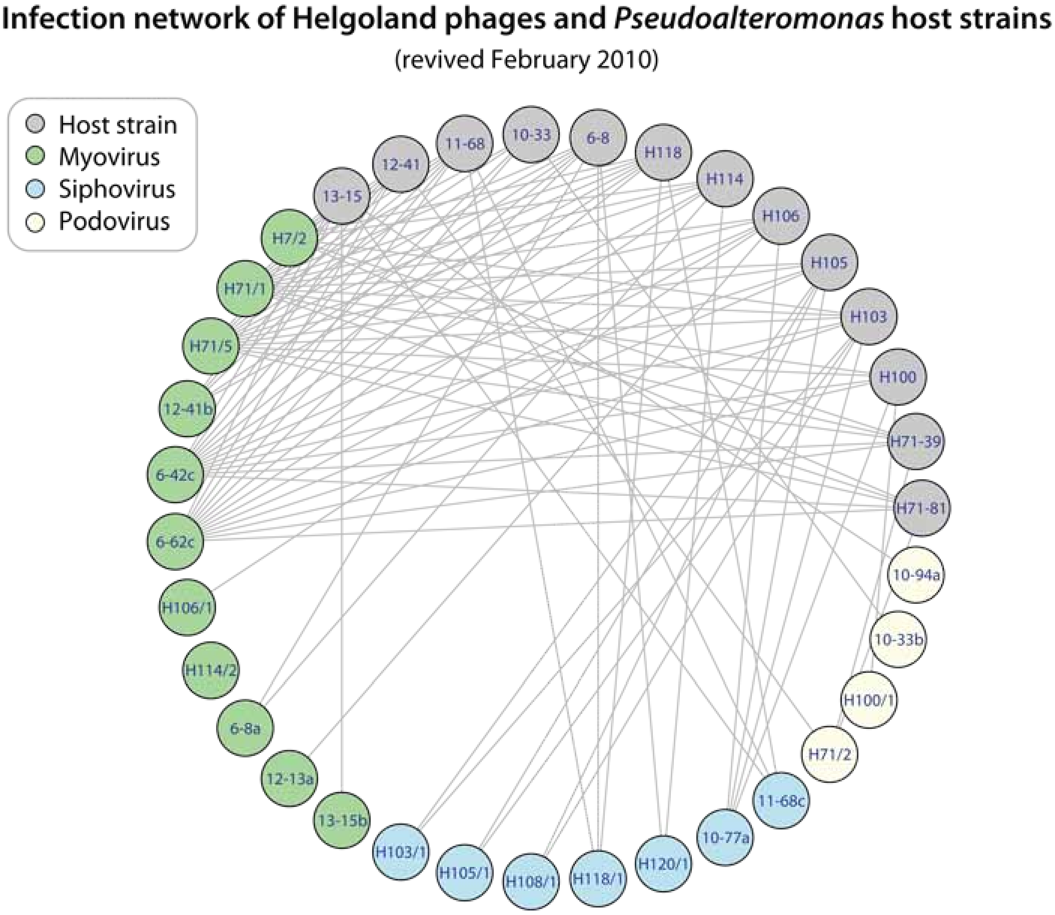


Figure 6.1: Helgoland phage-host network. Phages are those investigated by Wichels et al. (1998). A subset of available hosts was chosen to test infectivity in February 2010, nearly 10 years (and often more) after the last time the phages were shown to be infective and the hosts shown to be viable.

be they due to structural adaptations (e.g., tail fiber mutations) or genomic 'fine-tuning' (e.g. restriction-modification systems or CRISPRs, as introduced in section 1.1.3.3).

6.2.2 Linking *bio*– with –*informatics*: The case of cyanophage P-SS2

The exhaustive manual curation of metadata describing the sequenced marine phages (including megx.net interpolated environmental parameters) revealed that the habitat of one phage, Cyanophage P-SS2, is unique (Figure 3.6; Duhaime et al., 2010a). Such a bioinformatic-enabled observation lends itself to various ecogenomic hypotheses and questions about the influence of environment on phage gene content. As compared to other sequenced marine phages, is P-SS2 specially adapted to, i.e., persistence at specific nutrient conditions? Do P-SS2 phage proteins benefit its host in response to changing nutrient conditions? Such speculations are best tested experimentally, rather than *in silico*. Phage infection experiments were designed and conducted to investigate the influence of nutrient concentrations (phosphate and nitrogen) on the infection dynamics of P-SS2 and its host, *Prochlorococcus* MIT9313 (Chapter 5). Future plans to monitor global genome expression of both phage and host are likely to implicate specific genes involved in modulating nutrient stress during phage infection. Furthermore, prior to this work, there had been speculation that Cyanophage P-SS2 was capable of integrating into its host chromosome (Sullivan et al., 2009), which has yet to be observed in any marine cyanophage systems. Results of this thesis confirm that phage P-SS2 indeed integrates at several positions in the MIT9313 chromosome (Figure 5.3). As environmental conditions (and ensuing host stress) are suspected to drive the decision of whether, upon infection, a phage immediately lyses its host or integrates in a lysogenic state (Figure 1.7), the influence of nutrient stress on integration will also be investigated.

6.3 The road ahead

The marine virus genomics platform, and its planned future developments, will serve to address additional questions of marine phage ecology and evolution not addressed in the core components of this thesis. Furthermore, strategic sequencing and increased focus on the viral component of microbial systems will shed light on the role of viruses in our world's oceans.

6.3.1 Ecogenomic principles to tackle "regionality" versus "ubiquity"

As discussed in the introductory section on phage diversity (Section 1.1.2), there is evidence for both *a*) regionalization of phage types specific to certain biomes, as well as *b*) a large degree of ubiquity, whereby most genes are pervasive across all habitats. These patterns emerged as metagenome reads were mapped to homologous genes in phages of the phage Proteomic Tree (Rohwer and Edwards, 2002). Contradictions between regionalization and ubiquity are likely influenced by the inherent problem with mapping *reads* to the Phage Proteomic Tree to make inferences about phage distribution. Due to their "Modular Evolution", reads of genome fragments represent neither genotypes nor species nor phage groups, but simply genes or modules, which are known to rampantly shuffle between phages, even those infecting different hosts.

If biased recombination plays a role in phage evolution, such that 'host-associated' and 'phage-specific' patterns exist (as discussed in the previous section), this would be evident in the large-scale distribution of phage genes. Should some genes be more prone to, i.e., phage-phage swapping, the presence of these promiscuous phage genes in an environment would give the impression of a "ubiquitous" phage genotype or group when mapped to the Phage Proteomic Tree (PPT).

Contrarily, should certain phage genes be more susceptible to 'host-associated' swapping, these genes (mirroring their host global distribution) will be more prone to exhibiting patterns of regionalization. A single phage genotype can be a concatenated mix of both wide- spread 'phage-specific' genes, and restricted 'host-associated' genes, thereby causing a PPT-based phage distribution profile to appear both regional and ubiquitous.

Thus, rather than contemplate whether phage genotypes/'species' are widespread or localized, future efforts should investigate the global distribution of phage genes (or perhaps entire modules). In the case of genes with true regionalization, should ecogenomics reveal them to have a non-random distribution with respect to environmental physicochemical parameters (while ubiquitous genes have a random distribution), this may further suggest a 'host-associated' phage gene pool. With the marine phage genomics platform developed through this thesis, partnered with interpolated environmental parameters available through megx.net, these speculative ventures are now possible.

6.3.2 Towards understanding microbial ecosystems through the viral component

There are now many clues supporting the view that, as far as their influence on microbes, phages are mere genomic extensions of the hosts they infect. They likely mediate much of the novel gene transfer that leads to strain-specific diversity, allowing ecotypes to become established and thrive in certain niches (Coleman et al., 2006; Kettler et al., 2007). As host strains come in and out of a given microenvironment, they are exposed to an amalgam of mobile genomic material, with transduction happening once every one million infections (Kenzaka et al., 2010; Miller and Day, 2008). In this way, hypervariable islands are "dynamic reservoirs for recent and local adaptation" (Kettler et al., 2007). Hosts then have a chance to put novel acquired genome content to the test in their new environment. Though there is strong indication that much interstrain genomic variation confers no selective advantage (Denef et al., 2010), when selection *does* favor the novel content, a new successful ecotype is born. The local virome thus holds the key to micro-scale success of cosmopolitan species.

As such, any attempt at a holistic understanding of microbial ecosystem functioning and evolution without considering the impact of phages and their gene content, is fruitless.

In 2008, an inter-institute "–omics" project, MIMAS (Microbial Interactions in MArine Systems), was initiated. One of the central aims of MIMAS is to investigate changes in microbial diversity and gene expression in response to seasonal dynamics (i.e., during and after the spring bloom) off the coast of Helgoland, an open ocean site in the North Sea and origin of Phage H105/1 (Chapter 4; Duhaime et al., 2010b). Inspired by the recognition of viruses as a critical component of microbial systems, an application was written (M. Duhaime), and successfully funded, to sequence the marine virus communities (viral metagenomes) accompanying microbial communities (microbial metagenomes and 'meta-transcriptomes') investigated in MIMAS. Availability of the coincident virome will allow the possibility to "map" virome sequences to microbial transcripts from the same water sample. Previous studies suggest that microbial metatranscriptomes contain a high degree of virus-like transcripts (Frias-Lopez et al., 2008; Gilbert et al., 2008). However, the degree of *local* viral gene expression is not known. The upcoming project is the first of its kind to examine what fraction of microbial community gene expression is comprised of virus sequences.

Finale

In light of currents in marine phage genomics at the inception of this thesis three years ago, it was obvious that a platform for consistently collecting and storing contextual data was needed to facilitate the generation and testing of hypotheses pertaining to phage ecology and ecogenomics. The core accomplishments of this thesis include the development, assessment, and initial application of such a framework. This bioinformatic platform was then successfully employed to offer insights into processes of phage ecology and evolution in the worlds oceans, as it will continue to accomplish in the years to come.

CHAPTER 7

Additional Analyses and Supplementary Material

7.1 Supplement to Chapter 1: Introduction and The Marine Phage (Meta)Genome Collection

This section provides minimal background information of the marine phage genome and metagenome collection, including the addition of prophages identified in sequenced marine microbes.

Marine Phages Genomes

There are currently 27 phage genomes sequenced from habitats classified as 'marine', according to EnvoLite terminology (an ontology of habitat descriptors. This work is described in detail in Chapter 3, with an aside mentioned below.

Still not grasping the diversity of genome types and concomitant viral life strategies There is now significant metagenomic evidence that important viral types are thus far completely missing from the marine phage genome collection, which currently only contains double-stranded DNA phages. There is mounting evidence that single-stranded DNA phages are highly abundant in a number of marine ecosystems (Angly et al., 2006; Desnues et al., 2008), though, primarily due to methodological biases, the sequenced marine phage isolates consist of only double-stranded DNA viruses. To appreciate the breadth of functional diversity held in the marine phages, efforts should be focused on isolating and sequencing the ssDNA phages.

Marine Prophages 'Genomes'

The host taxonomic breakdown of the identified marine prophages (Figure 7.1a) is quite similar to the current collection of marine phages (Figure 3.3), with the major exception pertaining to the relative number of *Cyanobacteria* phages and prophages. For instance, to date, of the 12 *Prochlorococcus* spp. genomes sequenced, none contain integrated prophages. It has been speculated that this may be due to selection for reduced genome sizes as a mechanism for these small organisms to succeed in the oligotrophic marine regions where they are typically found (Kettler et al., 2007).

Furthermore, the average size of marine prophages tends to be smaller than both marine phages and phages from all environments (Figure 7.1b). This is likely due to the fact that prophage detection, based on phage-specific proteins, foreign genome signatures, and positions near, i.e., tRNA genes (common insertion sites for mobile elements), does not ensure that the prophage is still a viable phage. Prophages can reside in their host chromosomes from minutes to millions of years (Casjens, 2003). During this time, they run the risk of losing their own gene function, including the ability to pull out of the host chromosome, due to mutation (amelioration processes) and illegitimate recombination. As such, many identified prophages may be non-functional phage remnants that have already experienced consequences of host genome reduction, and are thus smaller than their functional counterparts. Nonetheless, experimental attempts have confirmed that marine prophages found during microbe sequencing are fully functional by inducing the phage lytic cycle though host cell stress (e.g., mitomycin C, UV radiation, etc.) and producing viable progeny (Chen et al., 2006).

Sequenced Viromes (marine and non-marine)

There are currently 11 marine virus metagenomes (or viromes) sequenced, two of which target RNA viruses, while the remainder are enriched for DNA (Table 7.1). There are another eight from non-marine or marine ’organism-associated’ habitats.

Of these 19 viromes, eight were generated with short-read pyrosequencing technology (average roughly 100 bp per read). Depending on the sequencing method, there are marked differences in the information content contained therein. With the emergence of pyrosequencing, viromes got larger (by approximately two orders of magnitude), but saw a greater drop in information (Figure 7.2). This will improve as "second generation" sequencing technologies produce longer length reads.

Though they generally produce orders of magnitude more sequences, the pyrosequence-based viromes are consistently at the lower extreme of the "percent ’known’ sequence" range, with only 4-11% known; Appendix Table 7.1). It has already been shown that this short-read sequence data can not be expected to deliver reliable functional classification of proteins, such as of viruses, with distant homologues

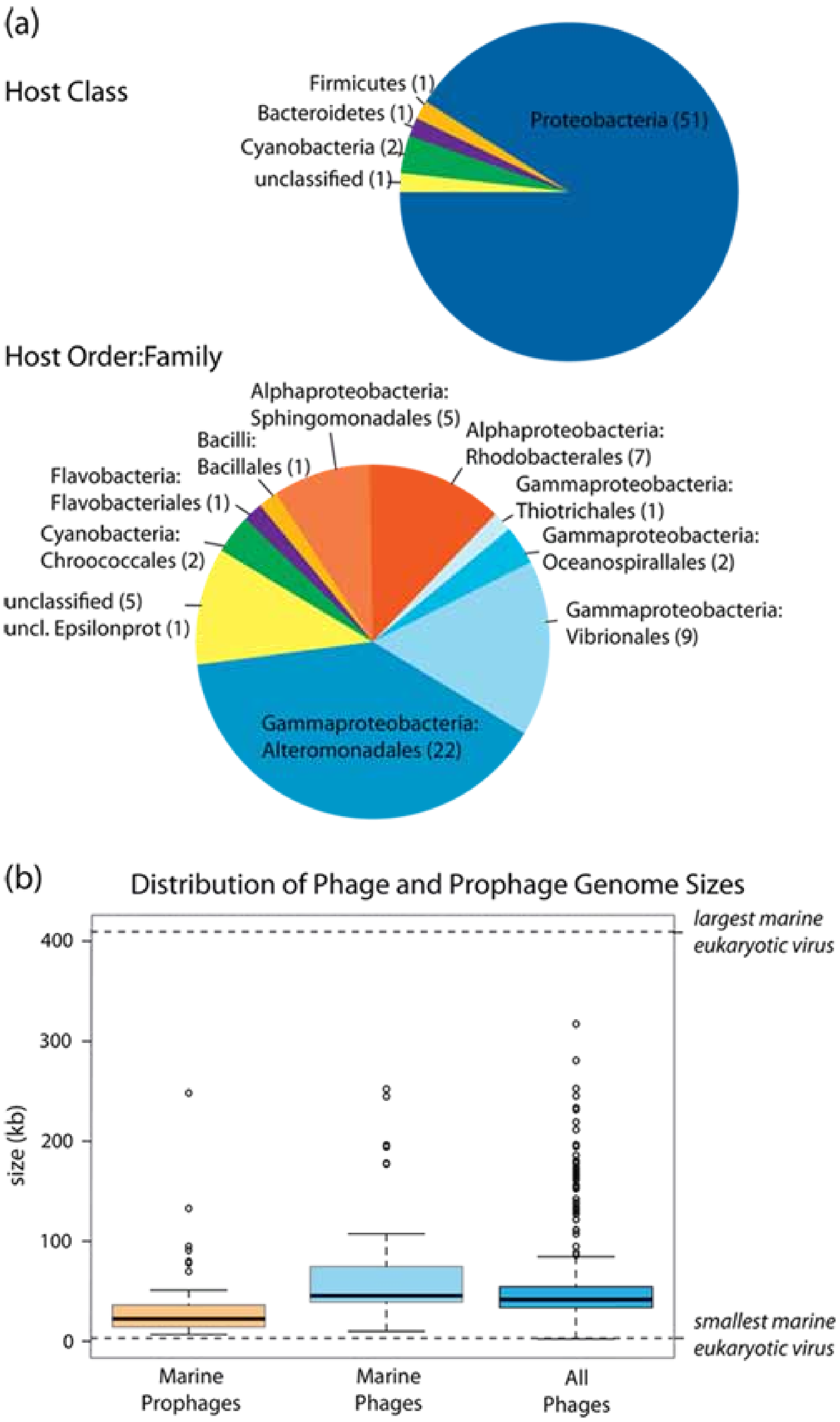


Figure 7.1: Taxonomy of sequenced marine microbes containing integrated prophages, and the distribution of prophage genome lengths.

Table 7.1: Marine and non-marine virus metagenomes grouped according to sequencing method. Number of reads, e-value cut-off applied, % unknown reads, % known reads, read "information", % reads in environmental databases (other microbial or viral metagenomes), % known hits to phage, viral, prophage, bacterial, archaeal, and eukaryotic sequences, the Shannon diversity index, % of the sample comprised of the most abundant genotype, estimated number of total genotypes ('g.typs'), and evenness values are as reported in the literature. Parameters describing community structure are based on power-log models (unless otherwise noted), and in most cases are generated by PHACCS (Angly et al., 2005), %*Read "Info"* is the information contained in the reads: number of reads classified as 'known'; *env* refers to a database of environmental microbial or viral metagenome reads; *vir* indicates phage reads are combined in the 'viral' count; no value designates that the value was not reported; *n.a.* designates the parameter is not applicable

virome	**reads**	**e-val**	**unk**	**read "info"**	**env**	**phg**	**vir**	**pro**	**bac**	**arc**	**euk**	**Shan. Index (nats)**	**% most abun g.typ**	**num g.typs**	**even**
pyrosequencing-derived reads															
Arctic^{i}	688 590	E-5^{ii}	86.9	77 811	1.9	*vir*	0.1	3.9	95.4	0.5	0.5	6.05	2.27	532	0.964
Sargasso Sea^{i}	399 343	E-5^{ii}	89.5	3993	9.5	*vir*	51.4	32.7	14	0	1.6	7.74	8.45	5140	0.905
British Columbia^{i}	416 456	E-5^{ii}	95.7	8329	23	*vir*	3.6	7.0	87.7	0.2	1.5	10.80	7.28	129 000	0.918
Gulf of Mexico^{i}	263 908	E-5^{ii}	93.3	11 876	2.2	*vir*	1.5	4.9	92.7	0.1	0.8	8.21	13.30	15 400	0.851
stromatolites (Highborne Cay, Bahamas)	150 223	E-2^{ii}	98.8	1803								4.1	16.5	161	0.8
stromatolites (Rio Mesquites, Mexico)	328 656	E-2^{ii}	97.7	7559								8.9^{iii}	8.6	19 520	0.9
stromatolites (Pozas Azules II, Mexico)	302 987	E-2^{ii}	99.3	2121								2.9^{iii}	19.4	23	0.92
Tampa Bay, FL, induced lysogens	294 068	E-3	93.4	19 408			30.5					9.13^{iii}	4.43	15 400	

table continued on next page

table continued from previous page

virome	reads	e-val	unk	read "info"	env	phg	vir	pro	bac	arc	euk	Shan. Index (nats)	% most abun g.typ	num g.typs	even
Sanger-derived reads															
Chesapeake Bay, MD, surface seawater	5641	E-3	30	2200	31	*vir*	17		18	0.4	2.6	8.31[iii]	0.065	4100	0.999
D. strigosa, healthy coral	1580	E-3	56 (35)	695	60	12	6.2					8.96	2.6	28 600	n.a.
D. strigosa, bleached coral	930	E-3	41 (18)	549	77	16	9.1					modeling not possible			
Scripps Pier, CA, surface seawater	1064	E-3	73.8	279	n.a.	*vir*	37.8	5.0	32.0	1.8	4.7	7.57	2.03	3350	0.932
Mission Bay, CA, surface seawater	873	E-3	65.2	304	n.a.	*vir*	31.25	4.3	23.4	0.7	11.8	7.99	2.63	7180	0.9
Mission Bay sediment, CA	1156	E-3	74.7	292	n.a.	43.8	2.7	15.1	3.4	30.5	8.6	8.90	0.01	7340	1
Equine Faeces	268		68	86	n.a.	*vir*	63	n.a.	20	7	6	not modeled			
human feces, adult (500 g)	532	E-3	58.6	220	n.a.	24.1	2.7	6.8	49.5	6.4	4.1	6.8	4.8	2390	
human feces, newborn (4 g)	477	E-3	65.8	163	n.a.	50.6	0.6	12.8	20.7	1.2	13.4	1.69	43.6	8	
Marine RNA viromes															
Jericho Pier, BC	247	E-3	63	91	not applicable										
Straight of Georgia, BC	108	E-3	81	21	not applicable										

(i) Arctic through Gulf of Mexico viromes are described in more details in Chapter 4, Table 7.4

(ii) homology search was done against the SEED database, not only NCBI-nr

(iii) community structure estimates based on logarithmic model (rather than power-law)

(Wommack et al., 2008).

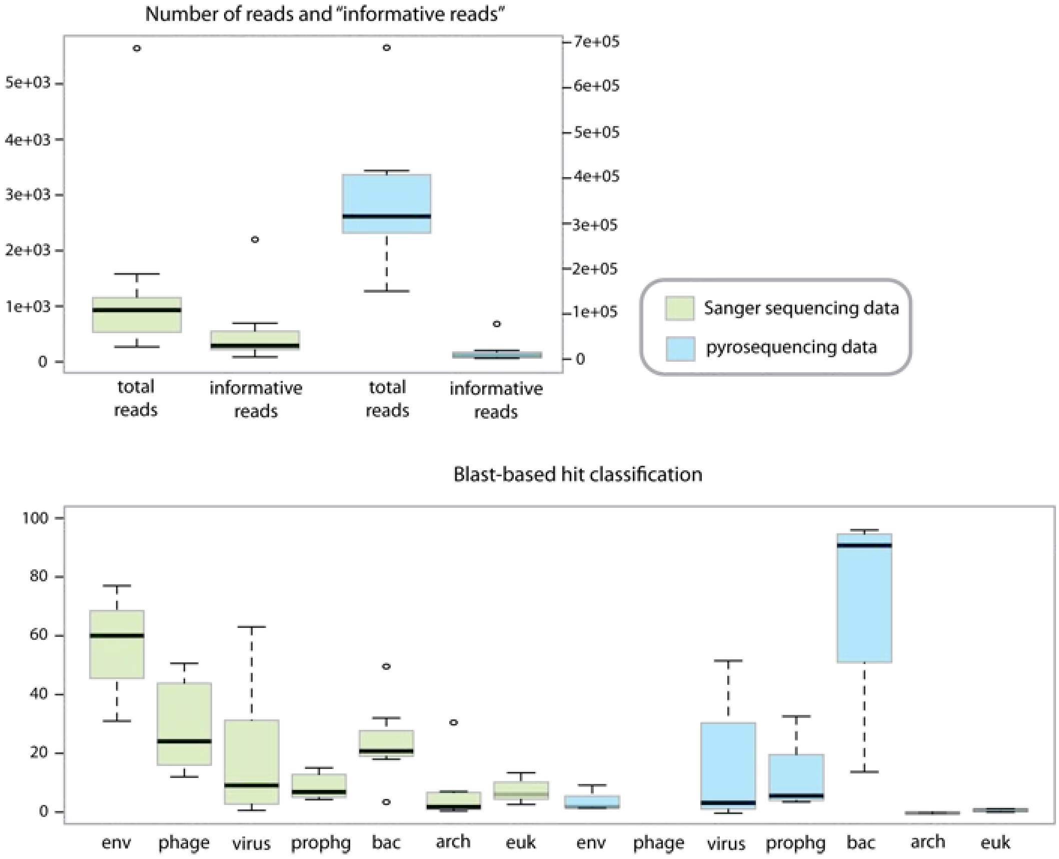
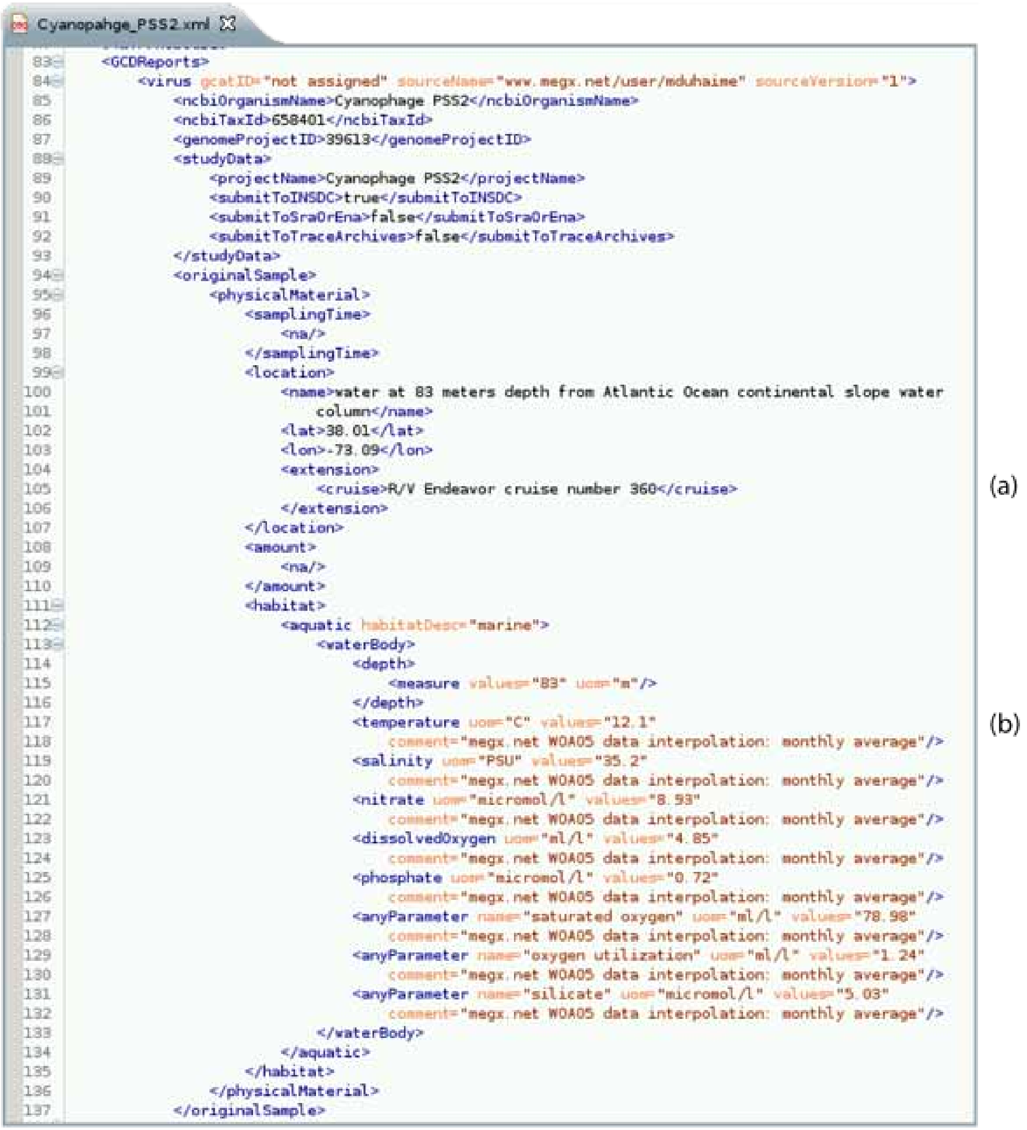


Figure 7.2: Summary of reads, "information", and classification of known Blast hits in the sequenced viral metagenomes. Informative reads are considered the fraction of known reads. Note the greater loss of information due to short read (approx. 100 bp) pyrosequencing.

7.2 Supplement to Chapter 3: Phage Contextual Data

MIGS-compliant contextual data of Cyanophage PSS2 captured in GCDML

Screenshot GCDML Report revealing the GCDML schema using the Eclipse plug-in, oXygen. Note the (a) cruise data and (b) interpolated environmental parameters retrieved from megx.net for this genome can be added through the flexible GCDML ’extensions’ (Figure 7.3).



```
<GCDReports>
    <virus gcatID="not assigned" sourceName="www.megx.net/user/mduhaime" sourceVersion="1">
        <ncbiOrganismName>Cyanophage PSS2</ncbiOrganismName>
        <ncbiTaxId>658401</ncbiTaxId>
        <genomeProjectID>39613</genomeProjectID>
        <studyData>
            <projectName>Cyanophage PSS2</projectName>
            <submitToINSDC>true</submitToINSDC>
            <submitToSraOrEna>false</submitToSraOrEna>
            <submitToTraceArchives>false</submitToTraceArchives>
        </studyData>
        <originalSample>
            <physicalMaterial>
                <samplingTime>
                    <na/>
                </samplingTime>
                <location>
                    <name>water at 83 meters depth from Atlantic Ocean continental slope water
                        column</name>
                    <lat>38.01</lat>
                    <lon>-73.09</lon>
                    <extension>
                        <cruise>R/V Endeavor cruise number 360</cruise>
                    </extension>
                </location>
                <amount>
                    <na/>
                </amount>
                <habitat>
                    <aquatic habitatDesc="marine">
                        <waterBody>
                            <depth>
                                <measure values="83" uom="m"/>
                            </depth>
                            <temperature uom="C" values="12.1"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <salinity uom="PSU" values="35.2"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <nitrate uom="micromol/l" values="8.93"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <dissolvedOxygen uom="ml/l" values="4.85"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <phosphate uom="micromol/l" values="0.72"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <anyParameter name="saturated oxygen" uom="ml/l" values="78.98"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <anyParameter name="oxygen utilization" uom="ml/l" values="1.24"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                            <anyParameter name="silicate" uom="micromol/l" values="5.03"
                                comment="megx.net WOA05 data interpolation: monthly average"/>
                        </waterBody>
                    </aquatic>
                </habitat>
            </physicalMaterial>
        </originalSample>
```

Figure 7.3: Screenshot GCDML Report revealing the GCDML schema using the Eclipse plug-in, oXygen. Note the (a) cruise data and (b) interpolated environmental parameters retrieved from megx.net for this genome can be added through the flexible GCDML 'extensions.'

Persistence of url to visualize the marine phages using megx.net Web Service

Web Map Service query to generate megx.net map in Figure 3.6a. For more information about the mapserver technology used by megx.net, see: http://www.megx.net/portal/tutorials/web_services.html (Table 7.2).

7.3 Supplement to Chapter 4: Phage H105/1 Genome

GOS metadata and hit counts

GOS sites queried by Pseudoalteromonas Phage H105/1 proteins, and their respective metadata, including habitat type, number of reads, and normalized hit counts to Phage H105/1. Data retrieved from CAMERA web portal (Seshadri et al., 2007) (Table 7.3).

Marine virome metadata and hit counts

Marine virus metagenomes used in this analysis. Reads were translated in all six reading frames for similarity searches against Phage H105/1 predicted ORFs. Environmental metadata describing the sites were interpolated using the megx.net GIS tools (Kottmann et al., 2010) then used to cluster the viromes by their ambient conditions. Numbers in parentheses are (# samples pooled, # sites, # reads) per virome, and (# samples pooled, # sites) per sample location, if samples had been pooled (Table 7.4).

Influence of varying e-value and % query identity thresholds on raw hit counts

(a) The GOS dataset (longer, Sanger sequence reads) resulted in more robust hits, with a less severe drop in hit counts as e-value and percent identity were changed in the less strict ranges. The GOS cut-off was set at e-values less than 10-4 and sequence identity greater than 20%. (b) The viromes dataset (short, average <100

Table 7.2: Web Map Service query to generate megx.net map in Figure 3.6a. For more information about the mapserver technology used by megx.net, see: http://www.megx.net/portal/tutorials/web_services.html.

HTTP request	http://www.megx.net/wms/?gmsSERVICE=WMS&VERSION=1.1.1&REQUEST=GetMap&LAYERS=satellite, samplingsites&PHAGES=phage&SRS=EPSG3A4326&BBOX=-180,-90,180,90&FORMAT=image%2Fpng&width =1600&height=800

Table 7.3: GOS sites queried by Pseudoalteromonas Phage H105/1 proteins, and their respective metadata, including habitat type, number of reads, and normalized hit counts to Phage H105/1. Data retrieved from CAMERA web portal (Seshadri et al., 2007).

Sample Site	Coordinates	Location	Fraction (μm)	Depth (m)	No. Reads	Habitat	Hit Count
GS 01c	32.33 N 64.56 W	Sargasso Sea	0.1-0.8	5.0	644551	Open Ocean	2.0
GS 02	45.50 N 67.24 W	Gulf of Maine	0.1-0.8	1.0	317180	Coastal	2.9
GS 03	42.85 N 66.22 W	Browns Bank, Gulf of Maine	0.1-0.8	1.0	368835	Coastal	1.4
GS 04	44.13 N 63.64 W	Outside Halifax, Nova Scotia	0.1-0.8	2.0	332240	Coastal	1.5
GS 05	44.69 N 63.64 W	Bedford Basin, Nova Scotia	0.1-0.8	1.0	142352	Coastal	1.4
GS 07	43.63 N 66.84 W	Northern Gulf of Maine	0.1-0.8	1.0	92351	Coastal	6.8
GS 08	41.49 N 71.35 W	Newport Harbor, RI	0.1-0.8	1.0	121590	Coastal	26.3
GS 09	41.09 N 71.60 W	Block Island, NY	0.1-0.8	1.0	61605	Coastal	23.4
GS 10	38.94 N 74.69 W	Cape May, NJ	0.1-0.8	1.0	52959	Coastal	22.4
GS 11	39.42 N 75.50 W	Delaware Bay, NJ	0.1-0.8	1.0	61131	Estuary	823.8
GS 12	38.94 N 76.41 W	Chesapeake Bay, MD	0.1-0.8	13.2	59679	Estuary	49.7
GS 13	36.00 N 75.39 W	Off Nags Head, NC	0.1-0.8	2.1	50980	Coastal	13.2
GS 14	32.50 N 79.26 W	South of Charleston, SC	0.1-0.8	1.0	129655	Coastal	8.2
GS 15	24.49 N 83.07 W	Off Key West, FL	0.1-0.8	1.7	79303	Coastal	8.6
GS 16	24.17 N 84.34 W	Gulf of Mexico	0.1-0.8	2.0	78304	Coastal	8.7
GS 17	20.52 N 85.41 W	Yucatan Channel, Mexico	0.1-0.8	2.0	124435	Open Ocean	2.7
GS 18	18.03 N 83.78 W	Rosario Bank, Honduras	0.1-0.8	1.7	126162	Open Ocean	1.4
GS 19	10.72 N 80.25 W	Northeast of Colon, Panama	0.1-0.8	1.7	138033	Coastal	36.3
GS 20	9.16 N 79.84 W	Lake Gatun, Panama	0.1-0.8	2.0	128885	Fresh Water	64.8
GS 21	8.13 N 79.69 W	Gulf of Panama	0.1-0.8	1.6	127362	Coastal	6.5
GS 22	6.49 N 82.90 W	250 miles from Panama City	0.1-0.8	2.0	127122	Open Ocean	4.6
GS 23	5.64 N 86.56 W	30 miles from Cocos Island, Costa Rica	0.1-0.8	2.0	257581	Open Ocean	4.4
GS 25	5.55 N 87.09 W	Dirty Rock, Cocos Island, Costa Rica	0.8-3.0	1.1	135325	Fringing Reef	140.9

table continued on the following page

table continued from previous page

Sample Site	Coordinates	Location	Fraction (μm)	Depth (m)	No. Reads	Habitat	Hit Count
GS 26	1.26 N 90.29 W	134 miles NE of Galapagos	0.1-0.8	2.0	296355	Open Ocean	1.3
GS 27	1.22 S 90.42 W	Devil's Crown, Floreana Island, Galapagos	0.1-0.8	2.2	131798	Coastal	6.7
GS 28	1.22 S 90.32 W	Coastal Floreana, Galapagos	0.1-0.8	2.0	121662	Coastal	2.4
GS 29	0.2 S 90.84 W	North James Bay, Santigo Island, Galapagos	0.1-0.8	2.1	133051	Coastal	0.8
GS 30	0.27 S 91.63 W	Warm seep, Roca Redonda, Galapagos	0.1-0.8	19.0	120671	Warm Seep	120.7
GS 31	0.30 S 91.65 W	Upwelling, Fernandina Island, Galapagos	0.1-0.8	12.0	102708	Coastal upwelling	3.4
GS 32	0.59 S 91.07 W	Mangrove on Isabella Island, Galapagos	0.1-0.8	0.1	222080	Mangrove	199.2
GS 33	1.22 S 90.43 W	Punta Cormorant Lagoon, Galapagos	0.1-0.8	0.2	189052	Hypersaline	141.6
GS 34	0.38 S 90.28 W	North Seamore, Galapagos	0.1-0.8	2.1	131529	Coastal	0.7
GS 35	1.39 N 91.82 W	Wolf Island, Galapagos	0.1-0.8	1.7	359152	Coastal	2.6
GS 36	0.02 S 91.20 W	Cabo Marshall, Isabella Island, Galapagos	0.1-0.8	2.1	436401	Coastal	4.0
GS 37	1.97 S 95.01 W	Equatorial Pacific TAO Buoy	0.1-0.8	1.8	148018	Open Ocean	6.7
GS 38	2.58 S 91.85 W	Tropical South Pacific	0.1-0.8	1.8	692255	Open Ocean	7.7
GS 39	3.34 S 101.37 W	Tropical South Pacific	0.1-0.8	2.0	134347	Open Ocean	2.2
GS 40	4.50 S 105.07 W	Tropical South Pacific	0.1-0.8	2.2	140814	Open Ocean	3.9
GS 41	5.93 S 108.69 W	Tropical South Pacific	0.1-0.8	2.0	77538	Open Ocean	1.3
GS 42	7.11 S 116.12 W	Tropical South Pacific	0.1-0.8	1.7	65670	Open Ocean	2.1
GS 48a	17.47 S 149.81 W	French Polynesia	0.8-3.0	1.4	47692	Coral Reef	77.7
GS 108a	12.09 S 96.88 E	Coccos Keeling	0.1-0.8	1.8	51788	Lagoon Reef	97.5
GS 110a	10.44 S 88.30 E	Indian Ocean	0.1-0.8	1.5	99288	Open Ocean	1.6

table continued on the following page

table continued from previous page

Sample Site	**Coordinates**	**Location**	**Fraction (μm)**	**Depth (m)**	**No. Reads**	**Habitat**	**Hit Count**
GS 111	9.59 S 84.19 E	Indian Ocean	0.1-0.8	1.8	59080	Open Ocean	1.0
GS 112a	8.50 S 80.37 E	Indian Ocean	0.1-0.8	1.8	99781	Open Ocean	1.2
GS 114	4.99 S 64.97 E	500 Miles west of the Seychelles	0.1-0.8	1.5	348823	Open Ocean	1.1
GS 115	4.66 S 60.52 E	Indian Ocean	0.1-0.8	1.5	61020	Open Ocean	0.7
GS 116	4.63 S 56.83 E	Outside Seychelles, Indian Ocean	0.1-0.8	1.5	60932	Open Ocean	2.6
GS 117a	4.61 S 55.50 E	St. Anne Island, Seychelles	0.1-0.8	1.8	346952	Coastal	2.0
GS 119	23.21S 52.30 E	International Water Outside Reunion Island	0.1-0.8	2.0	60987	Open Ocean	1.5
GS 120	26.03 S 50.12 E	Madagascar Waters	0.1-0.8	2.8	46052	Open Ocean	1.2
GS 121	29.34 S 43.21 E	between Madagascar and South Africa	0.1-0.8	1.5	110720	Open Ocean	2.7
GS 122a	30.89 S 40.02 E	between Madagascar and South Africa	0.1-0.8	1.9	101559	Open Ocean	2.0
GS 123	32.39 S 36.59 E	between Madagascar and South Africa	0.1-0.8	2.2	107966	Open Ocean	1.5
GS 148	6.31 S 39.00 E	East coast Zanzibar (Tanzania)	0.1-0.8	0.3	107741	Fringing Reef	23.4
‘ GS 149	6.11 S 39.11 E	West coast Zanzibar (Tanzania)	0.1-0.8	1.5	110985	Harbor	117.3

Table 7.4: Marine virus metagenomes used in this analysis. Reads were translated in all six reading frames for similarity searches against Phage H105/1 predicted ORFs. Environmental metadata describing the sites were interpolated using the megx.net GIS tools (Kottmann et al., 2010) then used to cluster the viromes by their ambient conditions. Numbers in parentheses are (# samples pooled, # sites, # reads) per virome, and (# samples pooled, # sites) per sample location, if samples had been pooled. *Lat/Lon and depth are as input to retrieve interpolated data from the GIS Tools available at http://www.megx.net/gms/tools/woa.html.

				megx.net interpolated data								
Virome	**Location**	**Lat/Lon***	**Depth (m)***	**Temp (C)**	**Sal (psu)**	**Nit (μM)**	**Phos (μM)**	**Silic (μM)**	**dis O_2 (ml/l)**	**O_2 sat (ml/l)**	**O_2 util (ml/l)**	**hits**
ARC	Chukchi Sea	69.69 N 171.46 W	10-3246									
(42 / 23 / 688,590)	Canadian Arctic	70.69 N 136.46 W	2-968	data interpolation not possible for Arctic sample sites								313
SAR (1 / 1 / 399,343)	Sargasso Sea	32.17 N 64.5 W	80	20.37	36.59	0.34	0.04	0.92	5.08	100.04	0.00	669
BBC	Georgia Straight	48.47 N	0	11.14	31.56	4.68	0.80	14.04	6.58	103.81	-0.23	929
(85 / 38 / 416,456)		125.03 W	122	7.83	33.64	25.86	2.14	39.47	3.10	46.37	3.61	
			245	7.18	33.91	30.73	2.46	47.34	2.42	36.01	4.34	

table continued on following page

table continued from previous page

				megx.net interpolated data								
Virome	**Location**	**Lat/Lon***	**Depth (m)***	**Temp (C)**	**Sal (psu)**	**Nit (μM)**	**Phos (μM)**	**Silic (μM)**	**dis O_2 (ml/l)**	**O_2 sat (ml/l)**	**O_2 util (ml/l)**	**hits**
GOM (41 / 15 / 263,908)	off Texas Coast (13 / 5)	27.62 N 96.81 W	0	28.88	35.90	0.53	0.06	0.97	4.42	100.22	-0.01	460
	Northeastern Gulf of Mexico (14 / 6)	28.26 N 87.67 W	1	29.37	32.34	0.93	0.12	2.78	4.72	103.28	-0.15	
			120	19.03	36.38	9.02	0.51	3.07	3.63	72.34	1.45	
	Eastern Gulf of Mexico (8 / 2)	25.90 N 85.18 W	3	26.40	35.98	0.62	0.12	1.43	4.70	101.92	-0.08	
			90	22.25	36.36	3.77	0.25	1.86	4.17	83.84	0.84	
	Western Gulf of Mexico (6 / 2)	26.21 N 93.94 W	0	25.20	35.25	0.74	0.20	1.72	4.70	100.51	-0.01	
			164	17.22	36.20	17.88	0.76	3.89	3.13	56.94	2.37	
TAMPA (1 / 1 / 294,068)	St. Petersburg	27.76 N 82.55 W	0	22.99	35.93	0	0.2	2.76	4.96	99.79	0.01	933
Phage H105/1	Helgoland	54.17 N 7.88 E	0	16.42	32.96	4.68	0.6	5.64	5.65	101.48	-0.08	n.a.

bp, pyrosequencing reads) resulted in poor similarity results, with the number of hits at higher, less stringent e-values dropping off drastically. Thresholds were chosen to minimize these likely false-positive hits at e-values less than 10-4 and sequence identity greater than 10% (Figure 7.4).

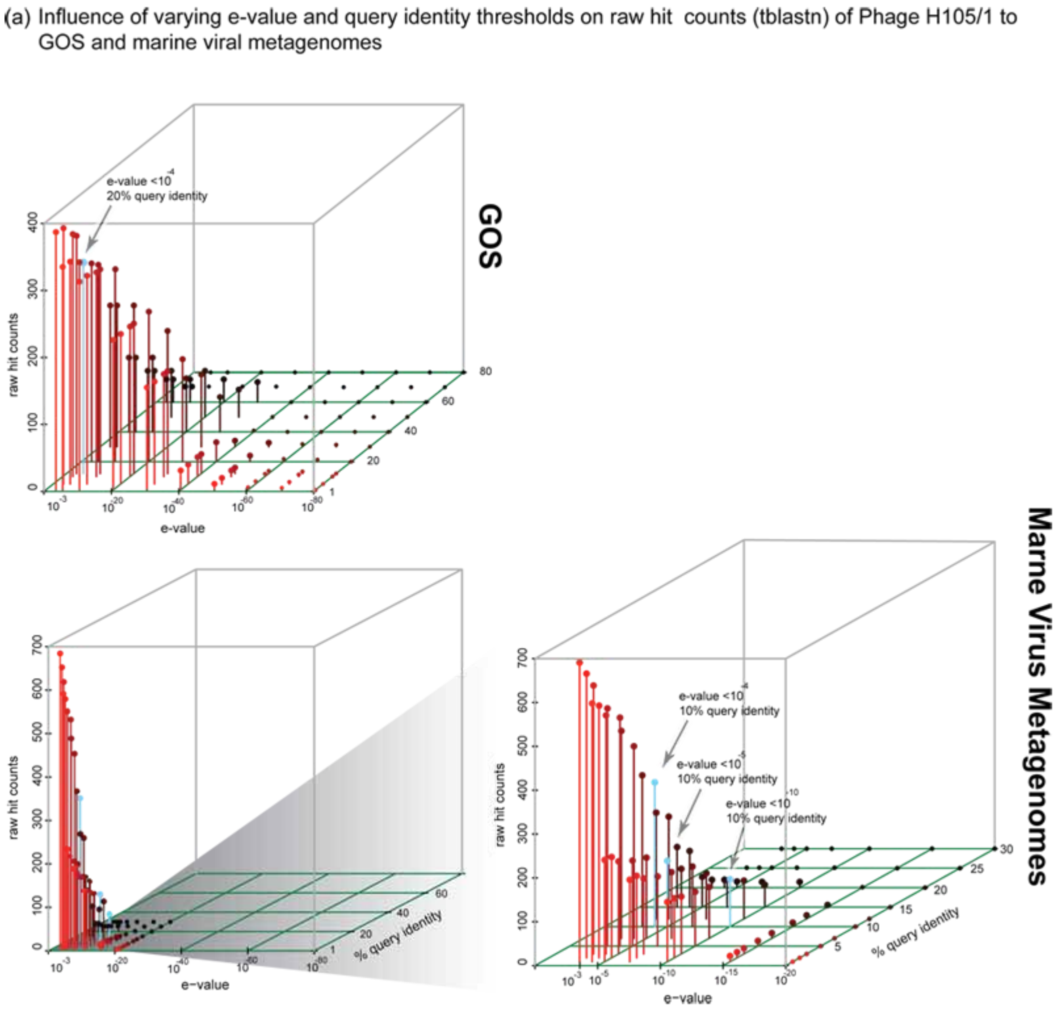

(a) Influence of varying e-value and query identity thresholds on raw hit counts (tblastn) of Phage H105/1 to GOS and marine viral metagenomes

Figure 7.4: (a) The GOS dataset (longer, Sanger sequence reads) resulted in more robust hits, with a less severe drop in hit counts as e-value and percent identity were changed in the less strict ranges. The GOS cut-off was set at e-values less than 10-4 and sequence identity greater than 20%. (b) The viromes dataset (short, average <100 bp, pyrosequencing reads) resulted in poor similarity results, with the number of hits at higher, less stringent e-values dropping off drastically. Thresholds were chosen to minimize these likely false-positive hits at e-values less than 10-4 and sequence identity greater than 10%.

Effect of genome window size on minimum tetra Pearson correlation coefficient cut-off

To reach a balance between highest possible correlation coefficients (long genome fragments) and most resolved tetra signature (short genome fragments), multiple lengths were tested. The final window size used was 10 kb, which allowed for a 0.61 Pearson coefficient cut-off (Figure 7.5).

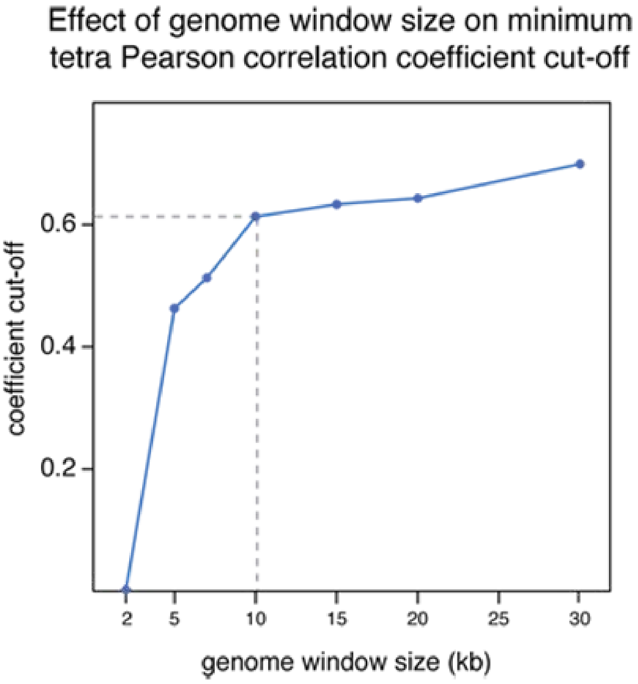


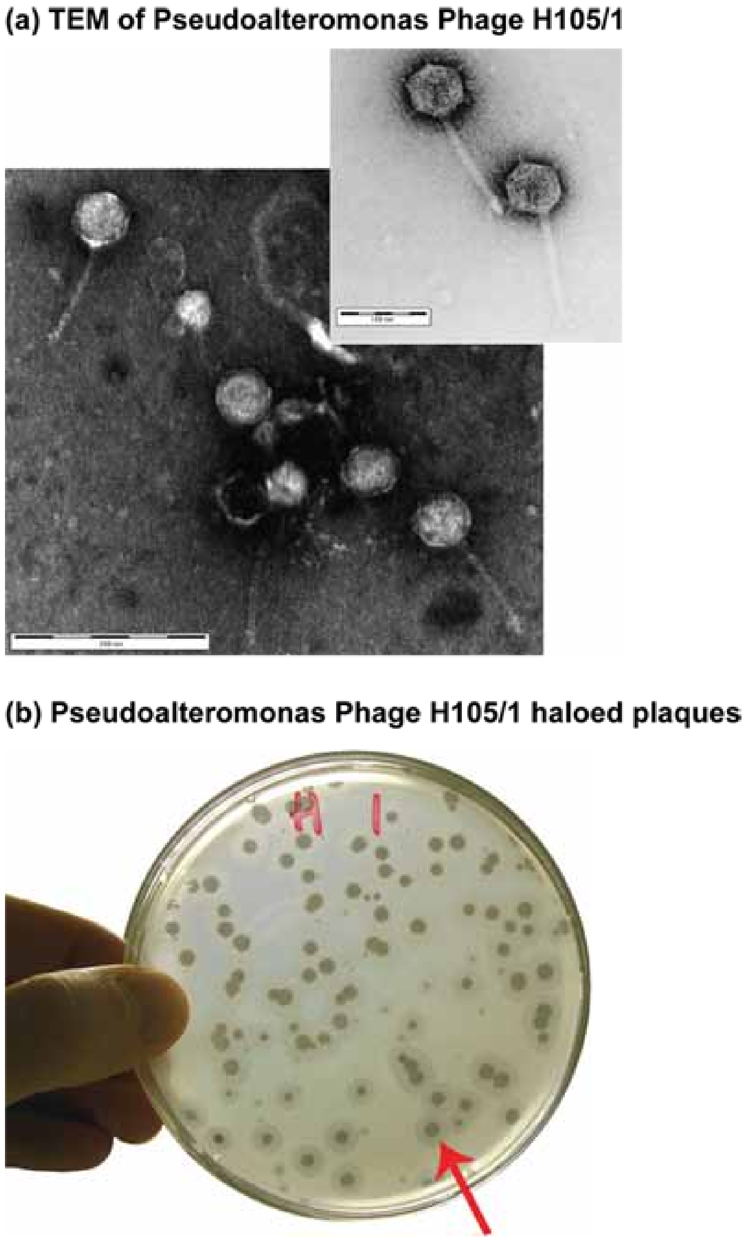
Figure 7.5: To reach a balance between highest possible correlation coefficients (long genome fragments) and most resolved tetra signature (short genome fragments), multiple lengths were tested. The final window size used was 10 kb, which allowed for a 0.61 Pearson coefficient cut-off.

Pseudoalteromonas phage H105/1 virion and plaque morphology

(a) Transmission electron micrograph of Pseudoalteromonas phage H105/1 (Wichels et al., 1998). (b) Plaque formation of Pseudoalteromonas phage H105/1 on *Pseudoalteromonas* sp. H105. Arrow indicates area of hazy halo around clear plaque (Figure 7.6).

Phage H105/1 structural proteome

Protein data and mapping to predicted ORFs of Phage H105/1 genome (Table 7.5).



(a) TEM of Pseudoalteromonas Phage H105/1

(b) Pseudoalteromonas Phage H105/1 haloed plaques

Figure 7.6: (a) Transmission electron micrograph of Pseudoalteromonas phage H105/1 (Wichels et al., 1998). (b) Plaque formation of Pseudoalteromonas phage H105/1 on *Pseudoalteromonas* sp. H105. Arrow indicates area of hazy halo around clear plaque.

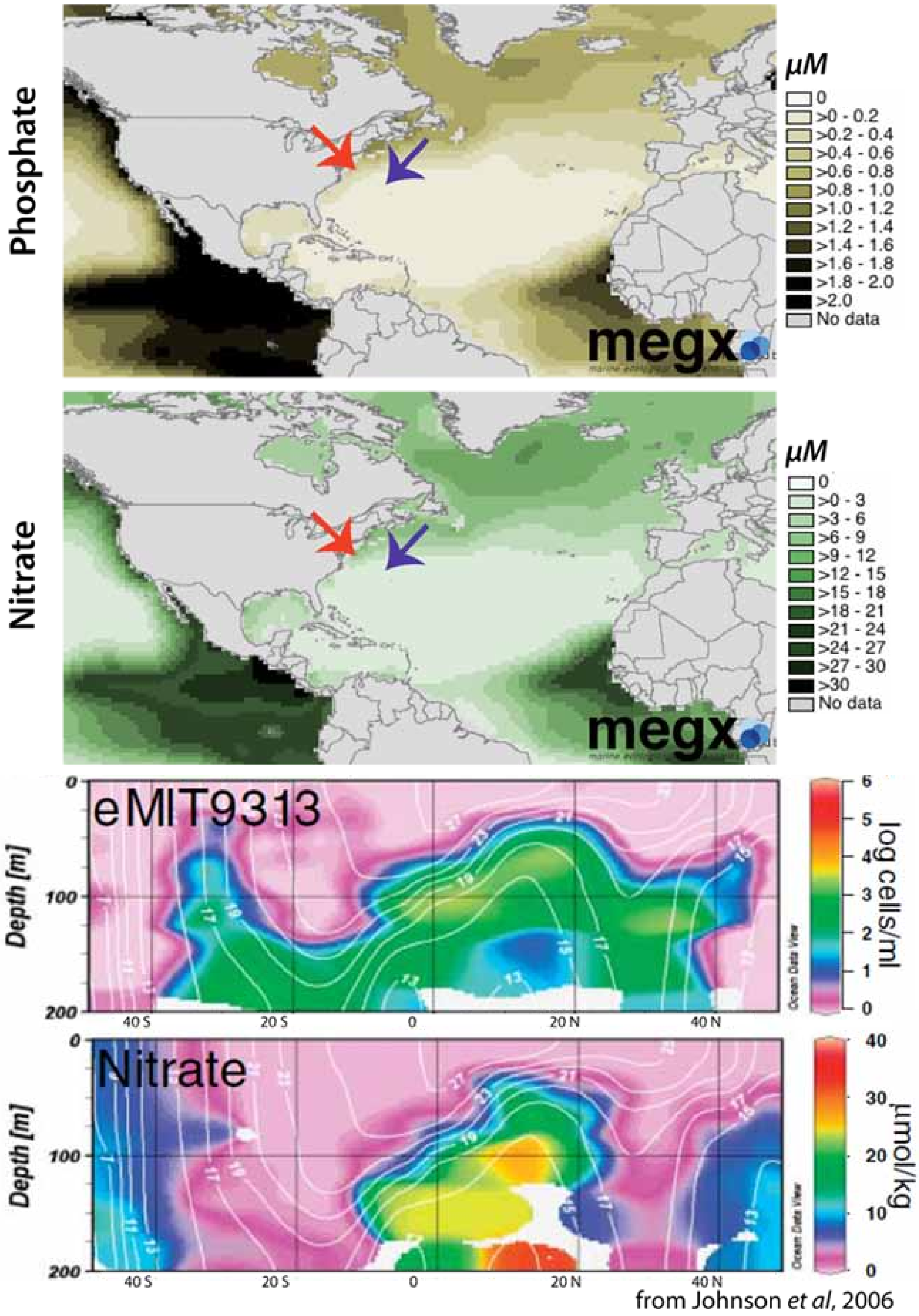
Table 7.5: Protein data and mapping to predicted ORFs of Phage H105/1 genome.

ORF	Sequence coverage	Calculated pI	Mowse Score	Nominal mass (Mr)	Annotation
20	36%	8.98	32	43656	phage head morphogenesis protein
26	76%	4.55	202	55195	conserved phage structural protein
28	95%	5.1	205	33681	conserved phage structural protein
35	91%	4.58	43	17666	conserved phage structural protein
43	66%	5.34	216	81159	phage tape tail measure protein
45	26%	4.35	45	92174	phage tail fibre adhesin Gp38 family protein
46	50%	4.34	86	78868	phage structural protein

Additional observations in light of the taxonomic power of tetranucleotide usage frequencies in phage genomics

In light of the rapid evolution of phage proteins and host ranges, it is highly doubtful that tetranucleotide frequencies will be successful in pairing a phage with its host in all cases. However, some associations can be made. In addition to bacterial and phage fragments in GenBank, Phage H105/1 consistently correlated with three unidentified fragments in GenBank (Figure 4.6). These belong to an assembled scaffold from the marine GOS dataset, which, once annotated, were found to be from Shewanella spp. (*Gammaproteobacteria*, *Alteromonadales*; Figure 4.1), demonstrating the ability for tetranucleotide frequency correlations to establish at least distant phage-host taxonomic relationships. Alternative methods of correlating tetranucleotide frequencies, such as those that employ emergent self-organizing maps (Dick et al., 2009), rather than tetranucleotide-derived z-score correlations, may be more successful in garnering taxonomic associations. Such approaches will also benefit from requiring shorter fragments, i.e., 2 kb (Dick et al., 2009) rather than 10 kb (Supplementary Material 7.5), for significant correlations, making a more granular examination of tetra-based genome landscapes possible.

7.4 Supplement to Chapter 5: Infection and Integration Dynamics of P-SS2



from Johnson et al, 2006

Figure 7.7: MIT9313 is a low light *Prochlorococcus marinus* ecotype most abundant in the higher nutrient, low latitude water of the deep photic zone. P-SS2, a siphovirus infecting it, was isolated at the northern, relatively lower nutrient, extent of its range (at least as depicted in this transect). megx.net generated maps of interpolated (a) phophate and (b) nitrate concentrations at 100 m depth, yearly average; MIT9313 (purple arrow, 135 m) and P-SS2 (red arrow, 83 m) isolation locations are depicted; (c) MIT9313 ecotype abundance and (d) nitrate measurements through a N-S transect, directly from Johnson et al. (2006);

ACKNOWLEDGEMENTS

This work would not have been possible without the invitation of Frank Oliver to join his group, as an advisor with the great trust to let me venture into a new topic to explore and develop my own passion. Thank you for taking a chance.

Similarly, on Helgoland I was lucky to find a phage accomplice, as, without Antje, the initial seed of this thesis may never have been sown. Thank you for passing on the BAH phage tradition and for rekindling your own interest in the viruses of the sea. The island also brought me great friends, Antje, Gunnar, Sonja, and all the other Heligolanders whose smiles, even if unknowingly, helped me reach the end.

Thanks to Christiane and Karl-Heinz for being the guiding hands and counseling minds of Marmic. Over the last four years I've seen such great changes and new traditions form. Your smiles will likely become the timeless Marmic mascots.

Many thanks to the entire Microbial Genomics Group for their help and guidance through the years, especially to Michi, Renzo, Ivo, Elmar, Jost and Hanno, who have invested extra effort to share their knowledge and appreciation for bioinformatics, such that it has become one of my own fascinations. And yes, a very special thanks to my wise and handsome officemates, thanks for all the great music, timely coffee breaks, and essential laughter. Until we meet again...

The kinks of this thesis were worked out through to the great advice and proof reading of Petra, Angelique, and Vincent. Thank you guys for your encouragement in the final push.

My great Bremen family, from wading through the mudflats on Sylt to a handful of entertaining Julefrokosts and Marmic retreats, we've held each other up in the tougher moments and laughed and cooked through the fun. And to my dear Amelia, Miss S, Jonnie, Lu, Pablo, Mia, Sonnie and Elmar, all of you whom this vagabond (with Lola in tow) has shared a flat with, thank you. No days have been sweeter than those in which you were there for me to say goodnight to. So, no goodbyes, only thanks and prospects of next meetings.

From thousands of miles, too many time zones and another continent away, I owe the strength of my independence to my Mamma, and the warm Virgil home to revive my soul to Hannah, James, Kate and Dad. Thank you for keeping me close. I miss you, but laugh with you often, and am so proud of you always.

And to Vini, the man of Antarctic exhibition expeditions, thank you for holding my hand close and my chin up, and lending your gloves when the wind swirled. Vallen in liefde is een understatement. Mijn leven is veranderd, en zij heeft maar begonnen. Ik kom nu *thuis*, samen ontroerd en in vrede. To cry all our laughter, and laugh all our tears. Dr. DuDes. Ik hou van jou.

Bibliography

Abedon, S. T. (2008). *Bacteriophage Ecology: Population Growth, Evolution, and Impact of Bacterial Viruses*. Cambridge University Press, ed. S.T. Abedon. Cambridge, UK. 17, 63

Abuladze, N. K., Gingery, M., Tsai, J., and Eiserling, F. A. (1994). Tail length determination in bacteriophage t4. *Virology*, 199(2):301–10. 85

Adams, M. H. (1959). *Bacteriophages*. Interscience Publishers, New York. 21

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10. 40, 72

Andersson, A. F. and Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, 320(5879):1047–50. 72

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., Felts, B., Nulton, J., Mahaffy, J., and Rohwer, F. (2005). Phaccs, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*, 6:41. 6, 7, 124

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A., and Rohwer, F. (2006). The marine viromes of four oceanic regions. *PLoS Biol*, 4(11):e368. 1, 3, 6, 7, 8, 9, 72, 74, 86, 121

Axmann, I. M., Dühring, U., Seeliger, L., Arnold, A., Vanselow, J. T., Kramer, A., and Wilde, A. (2009). Biochemical evidence for a timing mechanism in prochlorococcus. *J Bacteriol*, 191(17):5342–7. 106

Bahir, I., Fromer, M., Prat, Y., and Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol*, 5:311. 88

Banfield, J. F. and Young, M. (2009). Microbiology. variety–the splice of life–in microbial communities. *Science*, 326(5957):1198–9. 15, 17

Baron, S. (1986). *Medical microbiology*. Addison Wesley Publishing Company. 20

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–12. 15, 17

Becker, G. A. B., Dick, S. D., and Dippner, J. W. D. (1992). Hydrography of the german bight. *Marine Ecology Progress Series*, 91:9–19. 86

Bench, S. R., Hanson, T. E., Williamson, K. E., Ghosh, D., Radosovich, M., Wang, K., and Wommack, K. E. (2007). Metagenomic characterization of chesapeake bay virioplankton. *Appl Environ Microbiol*, 73(23):7629–41. 3, 6, 7

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. 73

Black, L. W. (1989). Dna packaging in dsdna bacteriophages. *Annu. Rev. Microbiol.*, 43:267–292. 85

Borbély, G., Kaki, C., Gulyás, A., and Farkas, G. L. (1980). Bacteriophage infection interferes with guanosine 3′-diphosphate-5′-diphosphate accumulation induced by energy and nitrogen starvation in the cyanobacterium anacystis nidulans. *J Bacteriol*, 144(3):859–64. 22, 83

Botstein, D. (1980). A theory of modular evolution for bacteriophages. *Annal of the New York Academy of Sciences*, 354:484–491. 4, 72

Botstein, D. and Matz, M. J. (1970). A recombination function essential to the growth of bacteriophage p22. *J Mol Biol*, 54(3):417–40. 85

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J. M., Nulton, J., Salamon, P., and Rohwer, F. (2004). Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci*, 271(1539):565–74. 3, 6, 7, 8

Breitbart, M., Haynes, M., Kelley, S., Angly, F., Edwards, R. A., Felts, B., Mahaffy, J. M., Mueller, J., Nulton, J., Rayhawk, S., Rodriguez-Brito, B., Salamon, P., and Rohwer, F. (2008). Viral diversity and dynamics in an infant gut. *Res Microbiol*, 159(5):367–73. 6

Breitbart, M. and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *TRENDS in Microbiology*, 13(6):278–284. 9

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. 6

Breitbart, M., Thompson, L. R., Suttle, C. A., and Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. *Special issue on a sea of microbes*, 1:135. 12, 14, 94

Brussaard, C. P. D., Wilhelm, S. W., Thingstad, F., Weinbauer, M. G., Bratbak, G., Heldal, M., Kimmance, S. A., Middelboe, M., Nagasaki, K., Paul, J. H., Schroeder, D. C., Suttle, C. A., Vaqué, D., and Wommack, K. E. (2008). Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J*, 2(6):575–8. 38, 46

Brüssow, H. and Hendrix, R. W. (2002). Phage genomics: Small is beautiful. *Cell*, 108:13–16. 63

Bryan, M. J., Burroughs, N. J., Spence, E. M., Clokie, M. R. J., Mann, N. H., and Bryan, S. J. (2008). Evidence for the intense exchange of mazg in marine cyanophages by horizontal gene transfer. *PLoS One*, 3(4):e2048. 22, 83, 94

Büchen-Osmond, C. (1996). Ictvdb-the universal virus database. *Data and Knowledge in a Changing World: The Quest for a Healthier Environment; Chambery*, 94. 2

Bushman, F. (2002). *Lateral DNA transfer: mechanisms and consequences*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. 2, 70

Calendar, R. (1970). The regulation of phage development. *Annu Rev Microbiol*, 24:241–96. 77

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L., and Brüssow, H. (2003). Phage as agents of lateral gene transfer. *Curr Opin Microbiol*, 6(4):417–24. 4, 85, 93

Carbone, A. (2008). Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol*, 66(3):210–23. 88, 89

Carbone, A., Zinovyev, A., and Képès, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*, 19(16):2005–15. 75

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol*, 49(2):277–300. 5, 85, 122

Casjens, S. R. (2005). Comparative genomics and evolution of the tailed-bacteriophages. *Curr Opin Microbiol*, 8(4):451–8. 3, 4, 113

Cavicchioli, R., Ostrowski, M., Fegatella, F., Goodchild, A., and Guixa-Boixereu, N. (2003). Life under nutrient limitation in oligotrophic marine environments: an eco/physiological perspective of sphingopyxis alaskensis (formerly sphingomonas alaskensis). *Microb Ecol*, 45(3):203–17. 10

Chain, P., Kurtz, S., Ohlebusch, E., and Slezak, T. (2003). An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief Bioinform*, 4(2):105–23. 47

Chen, F., Wang, K., Stewart, J., and Belas, R. (2006). Induction of multiple prophages from a marine bacterium: a genomic approach. *Appl Environ Microbiol*, 72(7):4995–5001. 122

Clokie, M. R. J. and Mann, N. H. (2006). Marine cyanophages and light. *Environ Microbiol*, 8(12):2074–82. 22, 83

Clokie, M. R. J., Shan, J., Bailey, S., Jia, Y., Krisch, H. M., West, S., and Mann, N. H. (2006). Transcription of a 'photosynthetic' t4-type phage during infection of a marine cyanobacterium. *Environ Microbiol*, 8(5):827–35. 14

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., Delong, E. F., and Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of prochlorococcus. *Science*, 311(5768):1768–70. 14, 15, 94, 118

Comeau, A. M. and Krisch, H. M. (2005). War is peace–dispatches from the bacterial and phage killing fields. *Curr Opin Microbiol*, 8(4):488–94. 5

Cséke, C. S. and Farkas, G. L. (1979). Effect of light on the attachment of cyanophage as-1 to anacystis nidulans. *J Bacteriol*, 137(1):667–9. 106

de Wit, R. and Bouvier, T. (2006). ’everything is everywhere, but, the environment selects’; what did baas becking and beijerinck really say? *Environ Microbiol*, 8(4):755–8. 9

Denef, V. J., Kalnejais, L. H., Mueller, R. S., Wilmes, P., Baker, B. J., Thomas, B. C., VerBerkmoes, N. C., Hettich, R. L., and Banfield, J. F. (2010). Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *PNAS*, 107:1–8. 118

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., Liu, H., Furlan, M., Wegley, L., Chau, B., Ruan, Y., Hall, D., Angly, F. E., Edwards, R. A., Li, L., Thurber, R. V., Reid, R. P., Siefert, J., Souza, V., Valentine, D. L., Swan, B. K., Breitbart, M., and Rohwer, F. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, 452(7185):340–3. 3, 8, 121

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., and Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol*, 10(8):R85. 72, 91, 138

Dictionary, W. N. N. C. (1991). Springfield: Merriam-webster. 25

Dodd, I. B. and Egan, J. B. (1996). Dna binding by the coliphage 186 repressor protein ci. *J Biol Chem*, 271(19):11532–40. 83

Doolittle, W. F. and Zhaxybayeva, O. (2009). On the origin of prokaryotic species. *Genome Research*, 19(5):744. 2

Duffy, S. and Turner, P. E. (2008). *Phage evolutionary biology*, pages 147–176. Cambridge University Press, Cambridge, UK, 1st edition. 91, 113

Duhaime, M. B., Kottmann, R., Field, D., and Glőckner, F. O. (2010a). Enriching public descriptions of marine phages using the migs standard: A case study assessing the contextual data frontier. *submitted*. 109, 110, 111, 112, 116

Duhaime, M. B., Wichels, A., Waldmann, J., Teeling, H., and Glőckner, F. O. (2010b). Ecogenomics and genome landscapes of marine pseudoalteromonas phage h105/1. *submitted*. 8, 109, 111, 112, 113, 114, 118

editorial (2008). A place for everything. *Nature*, 453(7191):2. 52, 55 editorial (2009). Metagenomics versus moore's law. *Nature Methods*, 6:623. 25, 26, 42

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using targetp, signalp and related tools. *Nat Protoc*, 2(4):953–71. 74

Erskine, J. M. (1973). Characteristics of erwinia amylovora bacteriophage and its possible role in the epidemology of fire blight. *Can J Microbiol*, 19(7):837–45. 76

Field, D. (2008). Working together to put molecules on the map. *Nature*, 453(7198):978. 36, 52, 55, 65, 92

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., DePamphilis, C., Edwards, R., Faruque, N., Feldman, R., Gilbert, J., Gilna, P., Glőckner, F. O., Goldstein, P., Guralnick, R., Haft, D., Hancock, D., Hermjakob, H., Hertz-Fowler, C., Hugenholtz, P., Joint, I., Kagan, L., Kane, M., Kennedy, J., Kowalchuk, G., Kottmann, R., Kolker, E., Kravitz, S., Kyrpides, N., Leebens-Mack, J., Lewis, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.-A. A., Spiers, A., Stevens, R., Swift, P., Taylor, C., Tateno, Y., Tett, A., Turner, S., Ussery, D., Vaughan, B., Ward, N., Whetzel, T., San Gil, I., Wilson, G., and Wipat, A. (2008). The minimum information about a genome sequence (migs) specification. *Nat Biotechnol*, 26(5):541–7. 27, 36, 38, 42, 46, 47, 54, 66, 109, 110

Filée, J., Bapteste, E., Susko, E., and Krisch, H. M. (2006). A selective barrier to horizontal gene transfer in the t4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol Biol Evol*, 23(9):1688–96. 4, 72

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–22. 72, 74

Frank, H. and Moebus, K. (1987). An electron microscopic study of bacteriophages from marine waters. *Helgoland Marine Research*, 41:385–414. 114

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and Delong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A*, 105(10):3805–10. 118

Friga, G. M., Borbély, G., and Farkas, G. L. (1981). Accumulation of guanosine tetraphosphate (ppgpp) under nitrogen starvation in anacystis nidulans, a cyanobacterium. *Arch Microbiol*, 129(5):341–3. 22, 83

Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., and Joint, I. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, 3(8):e3042. 118

Gross, M., Marianovsky, I., and Glaser, G. (2006). Mazg – a regulator of programmed cell death in escherichia coli. *Mol Microbiol*, 59(2):590–601. 22, 83

Grote, A., Hiller, K., Scheer, M., Münch, R., Nőrtemann, B., Hempel, D. C., and Jahn, D. (2005). Jcat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res*, 33(Web Server issue):W526–31. 75

Haggard-Ljungquist, E., Halling, C., and Calendar, R. (1992). Dna sequences of the tail fiber genes of bacteriophage p2: evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. *Journal of bacteriology*, 174(5):1462. 5, 85

Hendrix, R. W. (2008). *Phage evolution*, pages 177–194. Cambridge University Press, ed. S.T. Abedon. Cambridge, UK, 1st edition. 2, 3

Hendrix, R. W., Smith, M., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the worlds a phage. *Proceedings of the National Academy of Sciences*, 96(5):2192. 72

Hirschman, L., Clark, C., Cohen, K. B., Mardis, S., Luciano, J., Kottmann, R., Cole, J., Markowitz, V., Kyrpides, N., Morrison, N., Schriml, L. M., Field, D., and Project, N. (2008). Habitat-lite: a gsc case study based on free text terms for environmental metadata. *OMICS*, 12(2):129–36. 38, 52, 55, 66

Hogeweg, P. (1978). Simulating the growth of cellular forms. *Simulation*, 31(3):90–96. 25

Institute, B. (2010). New reference. 92

Iyer, L. M., Koonin, E. V., and Aravind, L. (2002). Classification and evolutionary history of the single-strand annealing proteins, rect, redbeta, erf and rad52. *BMC Genomics*, 3(1):8. 83

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., and Chisholm, S. W. (2006). Niche partitioning among prochlorococcus ecotypes along ocean-scale environmental gradients. *Science*, 311(5768):1737–40. 95, 139

Joyce, G. F. (1994). In origins of life: The central concepts, eds. dw deamer, gr fleischaker. 17

Kao, C. C., Green, S., Stein, B., and Golden, S. S. (2005). Diel infection of a cyanobacterium by a contractile bacteriophage. *Appl Environ Microbiol*, 71(8):4276–9. 106

Kapfhammer, D., Blass, J., Evers, S., and Reidl, J. (2002). Vibrio cholerae phage k139: complete genome sequence and comparative genomics of related phages. *J Bacteriol*, 184(23):6592–601. 55

Kenzaka, T., Tani, K., and Nasu, M. (2010). High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level. *ISME J*. 118

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., and Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of prochlorococcus. *PLoS Genet*, 3(12):e231. 14, 15, 118, 122

Koga, T., Toyoshima, S., and Kawata, T. (1982). Morphological varieties and host ranges of vibrio parahaemolyticus bacteriophages isolated from seawater. *Appl Environ Microbiol*, 44(2):466–70. 62

Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., Glőckner, F. O., and Consortium, G. S. (2008). A standard migs/mims compliant xml schema: toward the development of the genomic contextual data markup language (gcdml). *OMICS*, 12(2):115–21. 32, 36, 38, 42, 47, 52, 66, 109

Kottmann, R., Kostadinov, I., Duhaime, M. B., Buttigieg, P. L., Yilmaz, P., Hankeln, W., Waldmann, J., and Glőckner, F. O. (2010). Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res*, 38(Database issue):D391–5. 48, 52, 53, 65, 66, 72, 74, 102, 109, 110, 128, 133

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–80. 74

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–8. 75

Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., DeMaere, M. Z., Ting, L., Ertan, H., Johnson, J., Ferriera, S., Lapidus, A., Anderson, I., Kyrpides, N., Munk, A. C., Detter, C., Han, C. S., Brown, M. V., Robb, F. T., Kjelleberg, S., and Cavicchioli, R. (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci U S A*, 106(37):15527–33. 21

Lee, S., Kim, M. H., Kang, B. S., Kim, J.-S. S., Kim, G.-H. H., Kim, Y.-G. G., and Kim, K. J. (2008). Crystal structure of escherichia coli mazg, the regulator of nutritional stress response. *J Biol Chem*, 283(22):15232–40. 22

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008a). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, 24(6):863–5. 70, 74, 94, 104, 111

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008b). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol*, 25(4):762–77. 2

Lindell, D., Jaffe, J. D., Coleman, M. L., Futschik, M. E., Axmann, I. M., Rector, T., Kettler, G., Sullivan, M. B., Steen, R., Hess, W. R., Church, G. M., and Chisholm, S. W. (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature*, 449(7158):83–6. 14, 94, 99, 105, 106

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., and Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–9. 14, 46, 70

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., and Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from prochlorococcus viruses. *PNAS*, 101(30):11013. 46, 63, 70, 94

Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N. C. (2008). The genomes on line database (gold) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 36(Database issue):D475. 38

Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C., and Glőckner, F. O. (2006). Megx.net–database resources for marine ecological genomics. *Nucleic Acids Res*, 34(Database issue):D390–3. 36, 43

Long, A., McDaniel, L. D., Mobberley, J., and Paul, J. H. (2008). Comparison of lysogeny (prophage induction) in heterotrophic bacterial and synechococcus populations in the gulf of mexico and mississippi river plume. *ISME J*, 2(2):132–44. 86

Lucchini, S., Desiere, F., and Brüssow, H. (1999). Comparative genomics of streptococcus thermophilus phage species supports a modular evolution theory. *J Virol*, 73(10):8647–56. 4, 72

Lucks, J. B., Nelson, D. R., Kudla, G. R., and Plotkin, J. B. (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol*, 4(2):e1000001. 88, 89

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Főrster, W., Brettske, I., Gerber, S., Ginhart, A. W., Gross, O., Grumann, S., Hermann, S., Jost, R., Kőnig, A., Liss, T., Lüssmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., and Schleifer, K.-H. H. (2004). Arb: a software environment for sequence data. *Nucleic Acids Res*, 32(4):1363–71. 75

Makeyev, E. V. and Bamford, D. H. (2004). Evolutionary potential of an rna virus. *The Journal of Virology*, 78(4):2114. 63, 64

Mann, N. H., Clokie, M. R. J., Millard, A., Cook, A., Wilson, W. H., Wheatley, P. J., Letarov, A., and Krisch, H. M. (2005). The genome of s-pm2, a "photosynthetic" t4-type bacteriophage that infects marine synechococcus strains. *J Bacteriol*, 187(9):3188–200. 12, 14, 63, 94

Mann, N. H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature*, 424(6950):741. 12, 14, 63, 94

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80. 25

Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M. A. M., Grechkin, Y., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K., Hugenholtz, P., and Kyrpides, N. C. (2008). Img/m: a data management and analysis system for metagenomes. *Nucleic Acids Res*, 36(Database issue):D534–8. 36

Martiny, A. C., Coleman, M. L., and Chisholm, S. W. (2006). Phosphate acquisition genes in prochlorococcus ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci U S A*, 103(33):12552–7. 100, 103, 104, 105

Martiny, J. B. H. and Field, D. (2005). Ecological perspectives on the sequenced genome collection. *Ecology Letters*, 8(12):1334–1345. 47, 58, 63, 64

McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., and Paul, J. H. (2008). Metagenomic analysis of lysogeny in tampa bay: implications for prophage gene expression. *PLoS One*, 3(9):e3263. 5, 7, 72, 74, 88

Médigue, C., Krin, E., Pascal, G., Barbe, V., Bernsel, A., Bertin, P. N., Cheung, F., Cruveiller, S., D'Amico, S., Duilio, A., Fang, G., Feller, G., Ho, C., Mangenot, S., Marino, G., Nilsson, J., Parrilli, E., Rocha, E. P. C., Rouy, Z., Sekowska, A., Tutino, M. L., Vallenet, D., von Heijne, G., and Danchin, A. (2005). Coping with cold: the genome of the versatile marine antarctica bacterium pseudoalteromonas haloplanktis tac125. *Genome Res*, 15(10):1325–35. 70

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., et al. (2003). Gendb–an open source genome annotation system for prokaryote genomes. *Nucleic acids research*, 31(8):2187. 42

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. A. (2008). The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386. 36

Mãnnistő, R. H., Kivelã, H. M., Paulin, L., Bamford, D. H., and Bamford, J. K. (1999). The complete genome sequence of pm2, the first lipid-containing bacterial virus to be isolated. *Virology*, 262(2):355–63. 70

Miller, R. V. and Day, M. J. (2008). *Contribution of lysogeny, pseudolysogeny, and starvation to phage ecology*, pages 114–143. Cambridge University Press, ed. S.T. Abedon. Cambridge, UK, 1st edition. 18, 19, 21, 76, 95, 118

Mindich, L. (1988). Bacteriophage phi 6: a unique virus having a lipid-containing membrane and a genome composed of three dsrna segments. *Adv Virus Res*, 35:137–76. 63

Mobberley, J. M., Authement, R. N., Segall, A. M., and Paul, J. H. (2008). The temperate marine phage phihap-1 of halomonas aquamarina possesses a linear plasmid-like prophage genome. *J Virol*, 82(13):6618–30. 86

Moebus, K. (1980). A method for the detection of bacteriophages from ocean water. *Helgoland Marine Research*, 34(1):1–14. 114

Moebus, K. (1983). Lytic and inhibition responses to bacteriophages among marine bacteria, with special reference to the origin of phage-host systems. *Helgoland Marine Research*, 36(4):375–391. 114

Moebus, K. (1991). Preliminary observations on the concentration of marine bacteriophages in the water around helgoland. *Helgoland Marine Research*, 45(4):411–422. 114

Moebus, K. (1992). Further investigations on the concentration of marine bacteriophages in the water around helgoland, with reference to the phage-host systems encountered. *Helgoland Marine Research*, 46(3):275–292. 70, 72, 114

Moebus, K. (1997). Investigations of the marine lysogenic bacterium h24. 2. development of pseudolysogeny in nutrient rich broth. *Mar. Ecol. Prog. Ser*, 148:229–240. 77, 114

Moebus, K. and Nattkemper, H. (1981). Bacteriophage sensitivity patterns among bacteria isolated from marine waters. *Helgoland Marine Research*, 34(3):375–385. 114

Moore, L. R., Coe, A., Zinser, E. R., Saito, M. A., Sullivan, M. B., Lindell, D., Frois-Moniz, K., Waterbury, J., and Chisholm, S. W. (2007). Culturing the marine cyanobacterium prochlorococcus. *Limnology and Oceanography: Methods*, 5:353–362. 96

Moreira, D. and López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol*, 7(4):306–11. 17

Morris, J. J., Kirkegaard, R., Szul, M. J., Johnson, Z. I., and Zinser, E. R. (2008). Facilitation of robust growth of prochlorococcus colonies and dilute liquid cultures by "helper" heterotrophic bacteria. *Appl Environ Microbiol*, 74(14):4530–4. 106

Mullineaux, C. W. and Stanewsky, R. (2009). The rolex and the hourglass: a simplified circadian clock in prochlorococcus? *J Bacteriol*, 191(17):5333–5. 106

Muyzer, G., Teske, A., Wirsen, C. O., and Jannasch, H. W. (1995). Phylogenetic relationships of thiomicrospira species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16s rdna fragments. *Arch Microbiol*, 164(3):165–72. 75

Oakey, H. J. and Owens, L. (2000). A new bacteriophage, vhml, isolated from a toxin-producing strain of vibrio harveyi in tropical australia. *J Appl Microbiol*, 89(4):702–9. 55

Oliphant, A., Barker, D. L., Stuelpnagel, J. R., and Chee, M. S. (2002). Beadarray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques*, Suppl:56–8, 60–1. 25

Onodera, S., Olkkonen, V. M., Gottlieb, P., Strassman, J., Qiao, X. Y., Bamford, D. H., and Mindich, L. (1992). Construction of a transducing virus from double-stranded rna bacteriophage phi6: establishment of carrier states in host cells. *J Virol*, 66(1):190–6. 63

Panzeca, C., Beck, A. J., Tovar-Sanchez, A., Segovia-Zavala, J., Taylor, G. T., Gobler, C. J., and Sañudo Wilhelmy, S. A. (2009). Distributions of dissolved vitamin b12 and co in caostal and open-ocean environments. *Estuarine, Coastal and Shelf Science*, 85:223–230. 105

Partensky, F., Hess, W. R., and Vaulot, D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev*, 63(1):106–27. 94

Paul, J. H. and Sullivan, M. B. (2005). Marine phage genomics: what have we learned? *Curr Opin Biotechnol*, 16(3):299–307. 7

Paul, J. H., Sullivan, M. B., Segal, A. M., and Rohwer, F. (2002). Marine phage genomics. *Comparative Biochemistry and Physiology*, 133:463–476. 70

Paul, J. H., Williamson, S. J., Long, A., Authement, R. N., John, D., Segall, A. M., Rohwer, F. L., Androlewicz, M., and Patterson, S. (2005). Complete genome sequence of phihsic, a pseudotemperate marine phage of listonella pelagia. *Appl Environ Microbiol*, 71(6):3311–20. 48, 73

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. 3, 72

Poteete, A. R., Sauer, R. T., and Hendrix, R. W. (1983). Domain structure and quaternary organization of the bacteriophage p22 erf protein. *J Mol Biol*, 171(4):401–18. 83

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M., and Blaser, M. J. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res*, 13(2):145–58. 89

Pride, D. T., Wassenaar, T. M., Ghose, C., and Blaser, M. J. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, 7:8. 64, 89, 91

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glőckner, F. O. (2007). Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic Acids Res*, 35(21):7188–96. 37, 38, 75

Ptashne, M. (2004). *A genetic switch*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 3rd edition. 4, 18, 21, 83, 95, 97

Ravin, V., Ravin, N., Casjens, S., Ford, M. E., Hatfull, G. F., and Hendrix, R. W. (2000). Genomic sequence and analysis of the atypical temperate bacteriophage n15. *J Mol Biol*, 299(1):53–73. 86

Richter, M., Lombardot, T., Kostadinov, I., Kottmann, R., Duhaime, M. B., Peplies, J., and Glőckner, F. O. (2008). Jcoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. *BMC Bioinformatics*, 9:177. 40, 74

Ripp, S. and Miller, R. V. (1997). The role of pseudolysogeny in bacteriophage-host interactions in a natural freshwater environment. *Microbiology*, 143(6):2065. 76

Rodriguez-Valera, F., Martin-Cuadrado, A.-B. B., Rodriguez-Brito, B., Pasić, L., Thingstad, T. F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nat Rev Microbiol*, 7(11):828–36. 15, 16

Rohwer, F. (2003). Global phage diversity. *Cell*, 113(2):141–141. 94

Rohwer, F. and Edwards, R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. *Journal of bacteriology*, 184(16):4529. 2, 3, 7, 117

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. 12, 94

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H. H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M., and

Venter, J. C. (2007). The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLoS Biol*, 5(3):e77. 16, 36, 37, 72, 74

Sambrook, J. and Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual, 3rd ed.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 3rd edition. 73

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463. 25

Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., and Frazier, M. (2007). Camera: a community resource for metagenomics. *PLoS Biol*, 5(3):e75. 36, 74, 128, 130

Sharp, P. M. and Li, W. H. (1987). The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15(3):1281–95. 89

Siefert, J. L. (2009). Defining the mobilome. *Methods Mol Biol*, 532:13–27. 77

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, 103(32):12115–20. 9

Stamatakis, A. (2006). Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–90. 75

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., and Chisholm, S. W. (2005). Three prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol*, 3(5):e144. 12, 14, 42, 46, 63, 77, 94

Sullivan, M. B., Krastins, B., Hughes, J. L., Kelly, L., Chase, M., Sarracino, D., and Chisholm, S. W. (2009). The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. *Environ Microbiol*, 11(11):2935–51. 12, 14, 22, 63, 94, 95, 96, 105, 116

Sullivan, M. B., Waterbury, J. B., and Chisholm, S. W. (2003). Cyanophages infecting the oceanic cyanobacterium prochlorococcus. *Nature*, 424(6952):1047–51. 1, 64

Suttle, C. A. (2005). Viruses in the sea. *Nature*, 437(7057):356–61. 1, 2, 10, 11

Suttle, C. A. (2007). Marine viruses–major players in the global ecosystem. *Nat Rev Microbiol*, 5(10):801–12. 11, 12, 46, 70

Suttle, C. A. and Chen, F. (1992). Mechanisms and rates of decay of marine viruses in seawater. *Appl Environ Microbiol*, 58(11):3721–3729. 1, 11

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glőckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol*, 6(9):938–47. 76

Thingstad, T. F., Bratbak, G., and Heldal, M. (2008). *Aquatic phage ecology*, pages 251–280. Cambridge University Press, Cambridge, UK. 10, 11

Thingstad, T. F. and Lignell, R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology*, 13:19–27. 12

Thomas, T., Evans, F. F., Schleheck, D., Mai-Prochnow, A., Burke, C., Penesyan, A., Dalisay, D. S., Stelzer-Braid, S., Saunders, N., Johnson, J., Ferriera, S., Kjelleberg, S., and Egan, S. (2008). Analysis of the pseudoalteromonas tunicata genome reveals properties of a surface-associated life style in the marine environment. *PLoS One*, 3(9):e3252. 70

Tolonen, A. C., Aach, J., Lindell, D., Johnson, Z. I., Rector, T., Steen, R., Church, G. M., and Chisholm, S. W. (2006). Global gene expression of prochlorococcus ecotypes in response to changes in nitrogen availability. *Mol Syst Biol*, 2:53. 100, 103, 104, 105

Toussaint, A., Lima-Mendez, G., and Leplae, R. (2007). Phigo, a phage ontology associated with the aclame database. *Res Microbiol*, 158(7):567–71. 74

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43. 3

Vandenbergh, P. A. and Cole, R. L. (1986). Cloning and expression in escherichia coli of the polysaccharide depolymerase associated with bacteriophage-infected erwinia amylovora. *Appl Environ Microbiol*, 51(4):862–864. 76

Vaulot, D., Marie, D., Olson, R. J., and Chisholm, S. W. (1995). Growth of prochlorococcus, a photosynthetic prokaryote, in the equatorial pacific ocean. *Science*, 268(5216):1480–1482. 106

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66. 9, 15, 36

Wagner-Dőbler, I. and Biebl, H. (2006). Environmental biology of the marine roseobacter lineage. *Annu Rev Microbiol*, 60:255–80. 64

Waldmann, J. (2010). ocount2: a library for the calculation and comparison of oligonucleotide patterns. *Promedici®Software Labs*, online access. 76

Wang, I. N., Smith, D. L., and Young, R. (2000). Holins: the protein clocks of bacteriophage infections. *Annu Rev Microbiol*, 54:799–825. 86

Wichels, A., Biel, S. S., Gelderblom, H. R., Brinkhoff, T., Muyzer, G., and Schütt, C. (1998). Bacteriophage diversity in the north sea. *Appl Environ Microbiol*, 64(11):4128–33. 70, 73, 76, 114, 115, 136, 137

Wichels, A., Gerdts, G., and Schütt, C. (2002). Pseudoalteromonas spp. phages, a significant group of marine bacteriophages in the north sea. *Aquatic microbial ecology*, 27(3):233–239. 70, 114

Williamson, S. J., McLaughlin, M. R., and Paul, J. H. (2001). Interaction of the phihsic virus with its host: lysogeny or pseudolysogeny? *Appl Environ Microbiol*, 67(4):1682–8. 77

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C. S., Sutton, G., Frazier, M., and Venter, J. C. (2008). The sorcerer ii global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One*, 3(1):e1456. 14, 42, 46

Wilson, W. H., Turner, S., and Mann, N. H. (1998). Population dynamics of phytoplankton and viruses in a phosphate-limited mesocosm and their effect on dmsp and dms production. *Estuarine, Coastal and Shelf Science*, 46(2):49–59. 95

Wommack, K. E., Bhavsar, J., and Ravel, J. (2008). Metagenomics: read length matters. *Appl Environ Microbiol*, 74(5):1453–63. 126

Wommack, K. E. and Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev*, 64(1):69–114. 13

Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., Boffelli, D., Anderson, I. J., Barry, K. W., Shapiro, H. J., Szeto, E., Kyrpides, N. C., Mussmann, M., Amann, R., Bergin, C., Ruehland, C., Rubin, E. M., and Dubilier, N. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, 443(7114):950–5. 72

Xia, X. and Yuen, K. Y. (2005). Differential selection and mutation between dsdna and ssdna phages shape the evolution of their genomic at percentage. *BMC Genet*, 6(1):20. 64

Yoichi, M., Abe, M., Miyanaga, K., Unno, H., and Tanji, Y. (2005). Alteration of tail fiber protein gp38 enables t2 phage to infect escherichia coli o157:h7. *J Biotechnol*, 115(1):101–7. 5

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., et al. (2007). The sorcerer ii global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol*, 5(3):e16. 36

Zinser, E. R., Coe, A., Johnson, Z. I., Martiny, A. C., Fuller, N. J., Scanlan, D. J., and Chisholm, S. W. (2006). Prochlorococcus ecotype abundances in the north atlantic ocean as revealed by an improved quantitative pcr method. *Appl Environ Microbiol*, 72(1):723–32. 100