MOLECULAR AND ECOLOGICAL ASPECTS OF THE INTERACTIONS BETWEEN *AUREOCOCCUS ANOPHAGEFFERENS* AND ITS GIANT VIRUS

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Mohammad Moniruzzaman

December 2016

Copyright © 2016 by Mohammad Moniruzzaman

All rights reserved

To Ma, Abba, and Boro Mama

Acknowledgements

The body of work presented in this dissertation reflects the dedication of an incredible team of researchers who assisted me in every way from the beginning. My sincere thanks to my supervisor, Professor Steven W. Wilhelm, for his unflagging enthusiasm through the entire course of my research. His deep passion about research is contagious and has always inspired me to keep exploring, no matter how complex the scientific problem was. Choosing Dr. Wilhelm as my PhD supervisor was one of the best decisions I ever made. I also must thank Dr. Gary LeCleir, who supported me from the very beginning by teaching me new skills and providing valuable suggestions on the projects that I undertook. Thank you, Big Poppa!

Without my collaborators, it would be impossible to accomplish a number of projects. My sincere thanks to Dr. Christopher Gobler, Dr. Christopher M. Brown, Dr. Sonya Dyhrman, Dr. Harriet Alexander and Dr. Louie Wurch, who provided me with environmental samples and access to datasets.

My sincere thanks to my committee members who gave valuable suggestions and have been patient with my throughout these years. They provided specific guidelines on the directions I took, saving me a lot of time. Specifically, I have to mention Dr. Willie Wilson from Sir Alister Hardy Foundation for Ocean Sciences who flew to Knoxville every year and attended my committee meetings.

It has been an incredible journey of friendship and scientific camaraderie. Over the years, I have interacted with an amazing team of graduate students in the Wilhelm Lab. Morgan Wurch constantly inspired me with her hard work, dedication and great discussion on science and life's goals. Jackson, Lauren, Robbie, Sam and Josh – I am grateful to you for the wonderful time and discussion we had together in and outside the lab. I learned a lot as I mentored Eric Gann and Mark McDonald, who received training as undergraduates in Wilhelm lab. Thank you, for always being there during the long experiments and whenever I needed to vent out my frustration or crazy ideas!

Over the years, I have had the opportunity to travel to different parts of the USA because of my incredible photographer friends living all over the country. I am grateful to you for your hospitality and the memorable trips we took together.

Abstract

Viruses are increasingly being recognized as an important biotic component of all ecosystems including agents that control the rapid ecological events that are harmful algal blooms (HABS). *Aureococcus anophagefferens* is a pelagophyte which causes recurrent ecosystem devastating brown tide blooms along the east coast of the USA and has recently spread to China and South Africa. It has been suggested that a large virus (AaV) is possibly an important agent for demise of brown tide blooms. This observation is consistent with the recognition of a number of other giant viruses modulating algal blooms in marine systems. In this dissertation, we investigated both the molecular underpinnings of *Aureococcus*-AaV interactions and the dynamics of AaV and the associated viral community *in situ*. We determined the genome sequence and phylogenetic history of AaV using high throughput sequencing approach and revealed it's intertwined evolutionary history with the host and other organisms. Building upon the available genome of AaV and its host, we took an RNA-seq approach to provide insights on the physiological state of the AaV-infected *Aureococcus* 'virocell' that is geared towards virus production. *In situ* activity of AaV was detected by targeted amplicon and high throughput community RNA sequencing (metatranscriptomics) from Quantuck Bay, NY, a site with recurrent brown tide blooms. *AaV* and associated giant algal viruses in the Mimiviridae clade were found to respond to environmental changes, indicating that this newly recognized phylogenetic group is an important contributor to the eukaryotic phytoplankton dynamics. Analyzing time series metatranscriptomics from two distinct coastal sites recovered diverse viruses infecting microeukaryotes (including AaV) as part of interacting networks of viruses and microeukaryotes. Results from these studies testify AaV as an important factor for brown tide bloom demise, reveals the molecular underpinnings of AaV-host interactions and establishes the ecological relevance of Mimivirus-like algal viruses. We also provide foundation for using metatranscriptomics as an important tool in marine virus ecology – capable of recovering associations among coexisting marine microeukaryotes and viruses.

Table of Contents

Chapter I: Literature review **1**

Objectives **19**

List of references **21**

Chapter II: Genome of Brown Tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution **32**

Publication note **33**

Abstract **34**

Introduction **35**

Materials and Methods **36**

Results and discussion **40**

Conclusion **53**

List of references **55**

Chapter II Appendix **64**

Chapter III: The transcriptomic landscape of *Aureococcus* virocell **104**

Publication note **105**

Abstract **106**

Introduction **107**

Materials and Methods **108**

Results **110**

Discussion **132**

Conclusion **138**

List of references **140**

Chapter III Appendix **148**

Chapter IV: Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, *Aureococcus anophagefferens* **155**

Publication note **156**

Abstract **157**

Introduction **158**

Materials and Methods **160**

Results **163**

Discussion **168**

List of references **176**

Chapter IV Appendix **182**

Chapter V: Virus-host infection dynamics for marine microeukaryotes resolved from metatranscriptomics **185**

Publication note **186**

Abstract **187**

Introduction **188**

Materials and Methods **189**

Results and discussion **194**

Conclusion **215**

List of references **217**

Chapter VI: Conclusion **223**

List of references **227**

Vita **228**

List of Tables

Table 2.1: Mimivirus cluster of orthologous groups (MimiCOGs) present in AaV. The universal NCLDV core genes present in AaV are also listed. **78**

Table 2.2: CDS name, location, length, functional annotation and probable phylogenetic origin of putative AaV coding sequences. **80**

Table 2.3: Paralogous genes in AaV clustered in groups. **100**

Table 2.4: tRNAs in the AaV genome. **101**

Table 2.5: Genes uniquely present in AaV among the nucleocytoplasmic large DNA viruses (NCLDVs) **102**

Table 2.6: Putative AaV CDSs having homologs only in the NCLDVs (NCLDV specific ORFans). **103**

Table 3.1: Over and underexpressed genes encoding components of photosystem I and II across the time course compared to uninfected control. **127**

Table 3.2: AaV genes with COG/NCVOG assignments and time point when their expressions were first observed. Genes with no COG/NCVOG assignments are not shown. **148**

Table 4.1: Spearman's correlation co-efficient (ρ) indicating significant relationship between cell counts vs. environmental variables. (PC_Cyano: Phycocyanin rich cyanobacteria.). **184**

Table 5.1: RNA–seq library sizes after quality trimming. The Quantuck Bay samples were sequenced as single reads, while paired-end sequencing was carried out for the Narragansett Bay samples (denoted by X2). **192**

List of Figures

Figure 1.1: Maximum likelihood phylogenetic reconstruction of NCLDV major capsid protein sequences from environmental metatranscriptomic datasets. **10**

Figure 1.2: A number of ultrastructure analyses of *Aureococcus* and associated virus from different studies. **14**

Figure 2.1: Best BLASTp hits of AaV proteome against NCLDVs and three domains of life. **41**

Figure 2.2: Genome architecture of AaV AaV showing possible phylogenetic origins of the genes. **46**

Figure 2.3: Maximum likelihood phylogenetic trees of (A) B family DNA polymerase (AaV_141), (B) Major capsid protein (AaV_096) and (C) A32-like virion packaging ATPase (AaV_165) with other NCLDV members. **48**

Figure 2.4: BLASTn hits of whole genomes of diverse NCLDV members recruited against the genome of AaV showing the level of nucleotide similarity. **50**

Figure 2.5: Genome map of AaV depicting forward and reverse strand genes, COG and NCVOG categories, repetitive regions, GC content, GC skew and hypothetical genes with no known functions. **64**

Figure 2.6: Number of AaV genes classified in different cluster of orthologous groups (COG) categories. **65**

Figure 2.7: Maximum likelihood phylogenetic trees of selected AaV genes **66**

Figure 2.8: Whole genome dot plot of AaV. The DUF285 domain containing repetitive regions are marked with black arrowheads. Forward repeats are in green and inverted repeats are in red. **75**

Figure 2.9: MEME sequence conservation logo of intergenic motifs present at the upstream regions of the DUF_ORFs present in the AaV genome. **76**

Figure 2.10: Sequence conservation at the immediate upstream and downstream regions of ‘AaV and Phycodnavirus specific highly conserved genetic elements (AP_HGEs). **77**

Figure 3.1: *Aureococcus* and AaV counts over the course of infection. **111**

Figure 3.2: Percentage of reads mapped from each sequence library on host and virus genomes. **113**

Figure 3.3: Average number of reads mapped from the virus treated samples to AaV genome over time. **114**

Figure 3.4: Trend in expression of individual AaV genes over time. The read mapping from each timepoint was converted to coverage graphs with a 100bp sliding window. **115**

Figure 3.5: Heatmap showing temporal pattern in AaV gene expression. The read mapping data over individual genes was hierarchically clustered. **116**

Figure 3.6: Sequence logo of putative early promoter motif of AaV. **117**

Figure 3.7: Sequence logo of the motif overrepresented in the 'late' class of genes compared to the 'early to intermediate' gene class in AaV. **118**

Figure 3.8: Total number of genes up- and downregulated in the host across different time points compared to the healthy cultures. **121**

Figure 3.9: nMDS plot of the TMM normalized gene expression data from all replicate samples. **122**

Figure 3.10: Notable GO categories and KEGG pathways over- or underrepresented in the virus treated samples over different time points. **123**

Figure 3.11: The expression pattern of all the light harvesting complex (LHC) protein encoding genes in virus treated *Aureococcus* cultures compared to the control. **125**

Figure 3.12: The expression pattern of 35 selenoproteins in *Aureococcus* showing significant fold change (FDR $p<0.1$) in the virus infected culture at least at one time point. **129**

Figure 3.13: KEGG pathway "Selenocompound metabolism" with *Aureococcus* genes mapped on it and their expression pattern over time presented as heatmaps. **130**

Figure 3.14: Up- or downregulation of six annotated phosphate transporters in *Aureococcus* over six time points. **131**

Figure 3.15: Differential expression of *Aureococcus* genes involved in spermidine biosynthesis pathway. **133**

Figure 4.1: Dynamics of *Aureococcus*, other picoeukaryotes, phycocyanin-rich cyanobacteria and phycoerythrin-rich cyanobacteria during the sampling period in Quantuck Bay and Weesuck Creek . **161**

Figure 4.2: Heatmap of the distribution of amplicons representing the brown tide OTUs across different weekly samples in Quantuck Bay and Weesuck Creek (normalized to the library size). **165**

Figure 4.3: Maximum likelihood phylogenetic analysis of MCP Operational Taxonomic Units (OTU) from this study and from previous studies using the *mcp* primers (Larsen et al, 2008). **166**

Figure 4.4: Principal coordinate analysis of the pairwise Unifrac distances between the weekly samples from Quantuck Bay and Weesuck Creek. **169**

Figure 4.5: Alignment of Major Capsid Protein (MCP) sequences from the Megaviridae and Phycodnaviridae clade used to design the degenerate primers specific to algal Megaviridae. **182**

Figure 4.6: Distribution of singleton OTUs in each of the weekly samples from both the sampling locations. QB – Quantuck Bay, WC – Weesuck Creek. **183**

Figure 5.1: *Aureococcus* cell counts and nutrient dynamics during a 2011 brown tide bloom in Quantuck Bay, NY. **195**

Figure 5.2: Abundance of 9 NCLDV core genes including major capsid protein (MCP) in terms of normalized read counts and number of contigs recovered (up to 100bp length) from Quantuck Bay and Narragansett Bay . **197**

Figure 5.3 (A & B): Phylogenetic placement of major capsid protein contigs from Quantuck Bay (prefix: Q MC-) and Narragansett Bay (prefix N MC-) on a reference tree of NCLDVs with icosahedral capsids. **199**

Figure 5.4: Normalized read mapping statistics to AaV genome from the three Quantuck Bay samples. **201**

Figure 5.5 (A &B): Phylogenetic placement of (+) ssRNA virus contigs harboring RdRP motifs from Quantuck Bay (prefix: Q_sR_OV_) and Narragansett Bay (prefix: N_OV_) on reference trees. **203**

Figure 5.6: Phylogenetic placement of ssDNA virus contigs (prefix: N_OV_ and Q_sR_OV_) harboring viral replicase motifs from Quantuck Bay and Narragansett Bay on a reference tree. **205**

Figure 5.7: Proportional abundance of reads (in percentage) assigned to different viral groups in the Narragansett Bay study site over the five *in situ* samples. **207**

Figure 5.8: Proportional abundance of reads (in percentage) assigned to different viral groups in the Quantuck Bay study site over the three *in situ* samples. **208**

Figure 5.9: Complete or near-complete Picornavirales genomes recovered from both Quantuck Bay and Narragansett Bay study sites. **210**

Figure 5.10: Maximum likelihood phylogenetic tree of eukaryotic RPB1 (RNA polymerase II large subunit) used as reference to classify the RPB1 contigs assisted by pplacer. **212**

Figure 5.11: Representative SIMPROF clusters containing both viral and eukaryotic members from Quantuck Bay (Panel A) and Narragansett Bay (Panel B). **213**

CHAPTER I

Literature Review

Literature review

In the quest of defining life, viruses have always been neglected as byproducts of cellular biological entities: bags of escaped genes from the cells (Forterre, 2010). However, viruses have recently enjoyed renewed interest from the scientific community owing to advancements in both molecular and evolutionary biology. The recent availability of genomic data from a wide number of cellular life forms and viruses has provided unprecedented insights on the possible contribution of viruses in cellular evolution. There are several hypotheses (reviewed in (Forterre, 2010)) that put viruses at center stage of cellular evolution. For example, it has been proposed that viruses existed even before the Last Universal Common Ancestor (LUCA) (Bamford, 2003), that cellular replication machineries for informational molecules (DNA and RNA) originated in the viral world (Forterre, 1999) and even viruses induced the transition of cells from RNA to DNA genomes – which triggered the emergence of three distinct domains of life (Forterre, 2006). Adding to these virocentric perspectives on evolution, it was recently proposed that a large, dsDNA virus might be the origin of the eukaryotic nucleus (Takemura, 2001), while a protein encoded by a retrovirus integrated in the mammalian genome was shown to initiate placentation in mammals (Prudhomme *et al.*, 2005). Despite all these discoveries and propositions on the profound role of viruses in the evolution of the 'living' organisms, a question remains open 'Are viruses alive themselves?' The recent discovery of some unusually large 'giant viruses' or 'giruses' infecting unicellular eukaryotes has renewed interest in this question (Raoult *et al.*, 2004, Boyer *et al.*, 2010). Observing the complexities in their genomic content and replication strategies inside the host, some scientists have gone as far as to propose a new domain of life (Boyer *et al.*, 2010). Some even argue that the intracellular stage of viruses should be considered the actual 'virus organism' (Forterre, 2013), with free virion particles no more analogous to virus than fish eggs to actual fish. While these giant viruses fascinate scientists seeking answers to questions pertaining to evolution of life, they have also caught the attention from the microbial ecologists, due to their possible contribution to the Earth's biogeochemistry as manifested through their interaction with aquatic microeukaryotes (Short, 2012). In this chapter we review the current knowledge regarding the origin, evolution and ecology of giant viruses infecting microeukaryotes, emphasizing their ecological dynamics in the marine system.

Giant viruses: Origin and evolution

Unusually large viruses infecting diverse eukaryotes have caught widespread attention of scientists recently, but the observation of large viruses inside different eukaryotic cells goes back to several decades. For example, Granoff et al reported Frog virus 3, the type species in the Iridoviridae family in in 1966 (Granoff *et al.*, 1966). The ultrastructure of certain algae, for example, *Micromonas pusilla* (Mayer

& Taylor, 1979), *Oedogonium sp.* (Pickett-Heaps, 1972), and *Uronema gigas* (Dodds & Cole, 1980) revealed large virus like icosahedral particles as early as the 70's. Drawing a dichotomy between 'large/giant viruses' and 'regular viruses' strictly based on particle or genome size is difficult – it is likely that no clear limit exists on the genome size or complexity of a giant virus that is yet to be isolated (Claverie *et al.*, 2006). A recently proposed set of criteria that seem to be reasonable are a genome size ≥200,000 bp and/or particle size ≥0.2 μM (Wilhelm *et al.*, 2016). These limits allow inclusion of majority of the giant viruses isolated till date.

In 2001, Iyer et al showed through phylogenetic analysis of homologous genes that a number of large dsDNA viruses infecting diverse eukaryotes are evolutionarily related, having a common ancestor (Iyer *et al.*, 2001). In that paper, they first coined the phrase 'Nucleocytoplasmic Large DNA Viruses (NCLDVs)' to refer to such viruses. The rationale behind this assignment was the fact that Poxviruses complete their replication cycle entirely in the cytoplasm of the host, while other viruses, namely Iridoviruses, African Swine fever virus and Phycodnaviruses (infecting eukaryotic algae) complete part of their life cycle in the cytoplasm and/or nucleus of the host. The fact that these viruses are rather large compared to their bacteriophage or RNA virus counterparts was indeed recognized by the scientific community. However, things took a dramatic turn when the Mimivirus was officially reported (Raoult *et al.*, 2004). With its 1.2 Mbp genome, rivalling some parasitic bacteria, Mimivirus was a true giant in the virus world in terms of genomic content and size. What's interesting is Mimivirus was actually isolated in 1995, but was mistaken as a 'gram-positive' bacteria because of its large size and staining characteristics (Raoult *et al.*, 2007). A follow up phylogenetic study by Iyer et al (2006) revealed that Mimivirus shares common ancestry with other NCLDVs, thereby making it the biggest 'giant' in the realm of NCLDVs at that time. However, Mimivirus quickly lost its crown to *Megavirus chilensis*, a virus isolated from the marine system capable of infecting *Acanthamobea,* having a genome size of 1.26 Mbp and 440 nm capsid. The scientific community was left awestruck by the discovery of Pandoravirus (~2.5 MBp genome) (Philippe *et al.*, 2013) and *Pithovirus sibericum*, a virus resurrected from a 30,000 years old permafrost soil (Legendre *et al.*, 2014). As we will discuss in Chapters I and III, the genomics and phylogenetics of giant viruses is a rapidly evolving discipline, with new findings frequently reshaping our ideas on their ecology and evolution. One outcome of this rapid unfolding insight is the proposals to restructure the taxonomy of NCLDVs. Recently, an order 'Megavirales' was proposed for the NCLDVs which pass a number of 'inclusion criteria' which include viral particle size, presence of jelly roll capsid, the presence of NCLDV core genes and common ancestral origin (Colson *et al.*, 2012). As will be discussed in Chapter III, discovery of the Mimiviridae family was proposed to reassign as 'Megaviridae', that includes viruses infecting both algal and non-algal hosts (Arslan *et al.*, 2011, Santini *et al.*, 2013). None of these proposals

are approved by International Committee of Taxonomy of Viruses (ICTV) yet. Throughout this dissertation, we used the family name Mimiviridae and Megaviridae interchangeably, which reflects our view on the classification of these viruses as it has evolved over time.

A question central to the evolution of giant viruses is their origin. The first attempt to decipher the origin of NCLDVs put forward the 4th domain hypothesis – that giant viruses represent a possibly extinct domain of life. Phylogenetic analysis of a number of informational genes in NCLDVs demonstrated their deep monophylatic branching within Eukarya and Archaea, congruent with the three domain of life as reconstructed using the ribosomal gene based phylogeny (Boyer *et al.*, 2010). This idea has been subjected to a lot of debate and rigorous testing by other research groups. The first rebuttal of this hypothesis came from a reanalysis of the original data, where the data was fit to better models taking homoplasy and compositional heterogeneity of the sequences into account (Williams *et al.*, 2011). The reanalyzed trees couldn't reject the hypothesis that these genes could be acquired through horizontal gene transfer, and thus a forth domain was not necessary to explain the data. A comprehensive analysis on Mimi-, Pandora- and Pithovirus by Yutin *et al.* (2014) which belong to the proposed order 'Megavirales' revealed further insights on the origin of giant viruses. In these viruses, all the nearly universal cellular genes except one were found to branch within the eukaryotic tree, indicating that these genes were possibly acquired from the hosts during the course of evolution. Only one gene was found to be compatible with the fourth domain hypothesis, which formed a weakly supported sister-clade with the eukaryotic homologs. Detailed phylogenomic analysis of the genes in these viruses having cellular homologs found majority of these genes to be eukaryotic and bacterial origin. Additionally, reconstruction of gene loss and gain in these viruses tracked their likely origin in viruses with smaller and simpler gene repertoires (Yutin *et al.*, 2014). Polintons are large eukaryotic transposons, which were found to share blocks of homologous genes with diverse viruses, virophages, transposons and plasmids (Koonin *et al.*, 2015). The presence of jelly roll capsid proteins similar to Phycodnaviridae members and two Megavirales homologs of virus morphogenesis protein in Polintons suggests that Polintons possibly produce functional viruses under still unknown circumstances, and share evolutionary link with giant viruses (Krupovic *et al.*, 2014, Krupovic & Koonin, 2015). Together, this evidence supported a hypothesis that 'Megavirales' members inherited the virus morphogenesis module from Polintons (Koonin *et al.*, 2015).

The genetic potential of giant viruses

The large number of genes harbored by giant viruses gives rise to questions regarding their potential functions and how carrying these genes can provide a fitness benefit in terms of resources required for

replicating such large genomes (Wilhelm *et al.*, 2016). While the core genes shared across different NCLDV families mostly are involved in genome replication, transcription and virion morphogenesis (Koonin & Yutin, 2010), much is still unknown regarding the genes putatively acquired from diverse organismal sources *via* horizontal gene transfer (HGT) that dominate the genomic space of giant viruses. One of the very first insights regarding the cost and fitness benefits of harboring such a massive genome came from an experiment where Mimivirus was propagated within *Acanthamoeba* for 150 generations in an axenic condition (Boyer *et al.*, 2011). This resulted in the emergence of a Mimivirus strain which showed dramatic genome reduction at the terminal regions. Interestingly, this new strain lacked the typical surface fibers and was protected from virophage attack. It was suggested that the allopatric lifestyle without co-occurring bacteria and other organisms might let Mimivirus lose the genes that could otherwise be useful in competition for cellular niche and host internalization (Boyer *et al.*, 2011).

Functional insights into some of the giant virus specific genes have been obtained through *in vitro* and *in silico* approaches. It has been suggested that viruses with large genomes possibly encode proteins that can modulate host-virus interactions at different levels rather than having a direct role in virus reproduction (Yutin & Koonin, 2012). Perhaps one of the most notable examples is the acquisition of genes encoding sphingoplipids by the *E. huxleyi* viruses from the host (Monier *et al.*, 2009). As we will discuss later in this chapter, these genes play a profound role in modulating processes related to viral morphogenesis and apoptosis in the *E. huxleyi* 'virocell'(Rosenwasser *et al.*, 2016).

The function of a number of genes have been elucidated in the *Chlorella* viruses using recombinant protein expression approach, giving insights into possible role of these genes in giant virus – host interaction. PBCV-1 encodes a functional hyaluronan synthase, and cell surface hyaluronan is produced on the surface of the infected hosts (DeAngelis *et al.*, 1997). It was hypothesized that production of hyaluronan might prevent the uptake of the infected algae by *Paramacium*, or even facilitate attachment/consumption by an alternative host, if there is any. *Chlorella* virus CVK-2 encodes a functional chitin synthase and chitin fibers were found to be expressed on the infected cells, just like hyaluronan (Kawasaki *et al.*, 2002). It was proposed that chitin and hyaluronan both might play a similar function in *Chlorella* viruses. Chloroviruses also encode a number of transporter proteins, namely, aqualglyceroporin, calcium transporting ATPase and potassium ion channel protein (Van Etten & Dunigan, 2012). All these proteins were found to be functional by recombinant protein expression approach. It is important to mention that some of the chlorovirus proteins are the smallest or among the smallest in their family (Van Etten & Dunigan, 2012). Phylogenetic analyses have suggested that some of these proteins might be precursors of more complex cellular homologs. The small size of these proteins make them suitable for mechanistic and structural studies in the laboratory (Van Etten & Dunigan, 2012).

*Phaeocystis globosa viruses (*PgV) and Organic lake phycodnaviruses (OLPV) encode proteorhodopsins putatively acquired via HGT from proteorhodopsin encoding protists (Yutin & Koonin, 2012). Both these viruses infect photosynthetic protists, and based on amino acid conservation pattern it was hypothesized that viral proteorhodopsins might be involved in light-dependent signaling in infected hosts (Yutin & Koonin, 2012). Marseillevirus, an NCLDV infecting *Acanthamoeba,* contains three histone like protein coding genes. Interestingly, the histone like proteins were present in the viral particles and it was suggested that they might play a role in DNA condensation during packaging into the virions (Boyer *et al.*, 2009). In CroV, a 38 kb genomic region was putatively acquired from bacteria, which is enriched in carbohydrate biosynthesis genes (Fischer *et al.*, 2010). Among them were genes involved in 3-deoxy-D-manno-octulosonate (KDO) biosynthesis. Presence of KDO biosynthesis and other genes involved in carbohydrate metabolism might point to a role for these genes in viral surface glycoprotein synthesis (Fischer *et al.*, 2010). In line with this observation, Mimivirus was found to encode a gene involved in UDP-viosamine, a glycan that was demonstrated to be part of the external fibers lining the capsid along with rhamnose, glucose and N-acetylglucosamine (Piacente *et al.*, 2012). Phylogenetic analysis indicated that the gene encoding viosamine was acquired from bacteria early in the evolution. *Megavirus chilensis* encode a functional Cu,Zn superoxide dismutase, an enzyme known to detoxify reactive oxygen species which might be released as part of host defense mechanism (Lartigue *et al.*, 2015).Together, these findings indicate that the functional landscape of horizontally acquired genes might be highly diverse in giant viruses, a large number of which might be specific for virion structure or modulating the host environment.

A distinct feature of some of the giant viruses is their own 'mobilome'- a collection of mobile genetic elements (MGE) that further set them apart from 'regular' viruses. The genome of Lentille virus, a Mimiviridae member infecting *Acanthamobea,* harbors Sputnik 2 – a virophage integrated in their genome (Desnues *et al.*, 2012). Transpovirons – another kind of mobile genetic elements, are ~7.4 kb in size with 6 to 8 protein coding genes. Free form of transpovirons were found to be replicating in high copy numbers in the giant virus factory and were shown to have genomic features facilitating integration in the genome of the viral host, and even virophage (Desnues *et al.*, 2012). In addition, 30 copies of miniature inverted-repeat transposable elements (MITEs), presumably proliferated by transposition, were found to colonize the genome of *Pandoravirus salinus*, (Sun *et al.*, 2015).

While gene transfer from diverse sources endows the giant viruses with variety of functional capabilities, gene transfer events from the giant viruses to their potential hosts have also been documented. Genes having high phylogenetic affinity to giant virus homologues have been found in a number of sequenced eukaryotes. NCLDV core genes were detected in 8 protists and a metazoan (*Hydra magnipapillata)*

genome (Filée, 2014). In *H. magnipapillata* genome, a large 400 kb region is putatively of viral origin, being acquired from a Mimiviridae member according to phylogenetic analysis (Filee, 2014). No Mimiviridae member infecting a metazoan has been reported yet, however, it is possible that the integrated Mimiviridae member infects a protist associated with *H. magnipapillata.* Giant virus-like genes have been found in several other protists, including *E. huxleyi (Read et al., 2013)* and *Chlorella (Blanc et al., 2010).* Some plant genomes, e.g., *Physcomitrella patens and Selaginella moellendorffii (Maumus et al., 2014)* also harbor giant virus like genes.

Virophages: The viral parasites of giant viruses

One of the intriguing aspects of some giant viruses is their association with 'virophage'. Virophages are small dsDNA viruses that depend on the NCLDV host for propagation. The first virophage of a NCLDV, Sputnik, was found to co-infect the amoebal host with *Acanthamoeba polyphaga* Mamavirus, a new strain of Mimivirus. It was shown that Sputnik multiplies within the virus factory – the cytoplasmic region dedicated to replication of Mimivirus (Claverie & Abergel, 2009), and can negatively affect the morphogenesis of the host NCLDV. Sputnik infection led to abortive forms and aberrant capsid assembly of the host virus (La Scola *et al.*, 2008). A number of other virophages were co-isolated with different strains of Mimiviruses, namely Sputnik-2, Sputnik-3 and Zamilon (Abrahão *et al.*, 2014). Similarly, Mavirus is a virophage associated with *Cafeteria roenbergensis* virus (Fischer *et al.*, 2010), while PgVV was found to be an integrated virophage in the genome of *Phaeocystis globosa* virus (PgV) (Santini *et al.*, 2013), a Mimivirus member infecting algae. So far, only members of the Mimiviridae family are found to have virophages associated with them. Virophage signatures have also been identified in metagenomics datasets from diverse aquatic environments including deep ocean to inland freshwater lakes, iced to hydrothermal lakes and human guts to animal associated habitats (Zhou *et al.*, 2013). Such widespread distribution of virophages naturally raises questions regarding their roles in giant virus evolution and ecological dynamics. Organic Lake virophage (OLV), a putative virophage assembled from the metagenome of Organic Lake, Antarctica, was suggested to stimulate secondary production through the microbial loop by reducing the cellular host mortality and increasing the frequency of algal bloom during summer (Yau *et al.*, 2011). In line with this observation, a mathematical modelling approach of independent infection and coinfection (with giant virus) of the cellular host by virophages suggested that presence of virophages might be beneficial to the host cellular organism. The model predicted reduced giant virus abundance and increased host abundance in the presence of virophages (Taylor *et al.*, 2014). Recently, Fischer & Hackl (2016) showed experimentally that Mavirus, the virophage associated with CroV, can integrate into the genome of the cellular host and can form infectious virions upon superinfection by CroV . The release of reactivated virophages appears to promote survival of host at the

population level, rather than at the infected cell level. Thus, the model based predictions and the latest experimental approach indicate that virophages might promote the survival of the cellular host populations, while having negative effect on the giant virus propagation.

The ecology of giant viruses: distribution, diversity and dynamics

As giant viruses can infect diverse hosts from protists to metazoans, it is not surprising that they have been isolated from a wide range of environments. Mimivirus was initially isolated from an amoeba co-culture present in the water of a cooling tower in Paris (Raoult *et al.*, 2004), while *Megavirus chilensis*, a Mimiviridae member infecting the same host was isolated from the coast of Chile (Arslan *et al.*, 2011). Later, a number of viruses infecting *Acanthamoeba* were detected from environments including contact lens fluid, decorative fountains, soil, hospital water, river and lake water, hypersaline water/soil and insect larvae (Pagnier *et al.*, 2013). Sequence signatures of giant viruses having similarity to Mimivirus and algae infecting Mimiviridae members were detected in the metagenomes generated from the saline soil of Kutch desert (India) (Kerepesi & Grolmusz, 2016). The isolation of *Pithovirus sybericum* and *Mollivirus sybericum*, from ~30,000 years old permafrost samples (Legendre *et al.*, 2014, Legendre *et al.*, 2015), indicated that some giant viruses can have remarkable persistence, even in extreme environments. This wide environmental distribution of *Acanthamoeba* infecting giant viruses reflects the ubiquity of their host organisms. We now have evidence that some amoeba infecting viruses might have a broader host range including humans. Giant viruses from both Mimi- and Marseilleviridae family were detected in respiratory and stool samples (Pagnier *et al.*, 2013). Mimivirus was found to successfully infect and replicate in mice (Khan *et al.*, 2007), human phagocytic cells and peripheral blood mononuclear cells *in vitro* (Ghigo *et al.*, 2008). Algal virus DNA sequences homologous to *Acanthocystis turfacea* Chlorella virus (ATCV-1) were detected in oropharyngeal virome of human subjects participating in a study that included measures of cognitive functioning (Yolken *et al.*, 2014). Interestingly, presence of ATCV-1 DNA sequences showed a significant statistical correlation with decreased cognitive measurements. Supporting this observation, inoculation of ATCV-1 in healthy mice led to decrease in several cognitive domains and altered expression of genes in hippocampus (Yolken *et al.*, 2014).

As will be discussed later in this chapter, diverse giant viruses with both algal and non-algal hosts inhabit aquatic ecosystems, whose diversity and ecological dynamics have been elucidated using established lab cultures, targeted amplicon based and metagenomics approaches. While metagenomics and targeted amplicon based techniques provide valuable insights on the diversity and distribution of giant viruses in diverse environments, development of new approaches and modification of the existing ones seem to be necessary for gauging the true abundance, diversity and host range of giant viruses. A recent meta-

analysis by Wilhelm *et al.* (2016) probed a number of metatranscriptomic libraries generated from several distinct ecological niches and revealed the existence of different giant viruses in these datasets (Figure 1.1). This approach has the potential to detect 'active' giant viruses that drive different ecological processes in various environmental contexts.

In 1979, Mayer and Taylor isolated a virus infecting marine alga *Micromonas pusilla*. Ultrastructure studies indicated that this was a giant virus with icosahedral capsid of > 100 nm diameter, although no molecular evidence was available at that time (Mayer & Taylor, 1979). In 1991, Cottrell & Suttle (1991) re-isolated viruses of *M. pusilla* from diverse coastal and oligotrophic marine locations and characterized them based on restriction fragment length polymorphism (RFLP) analysis. Since then marine virus research intensified, with increasing number of articles appearing in the scientific databases (Short, 2012). Our knowledge of the diversity of marine giant viruses largely comes from targeted amplicon based and whole community metagenome based studies. It is important to mention that, so far, the only marine giant viruses infecting known heterotrophic organisms are several members of the Mimivirus lineage, *Megavirus chilensis* and *Cafeteria roenbergensis* virus (CroV). Thus most research on giant virus diversity are focused on viruses that infect photosynthetic algae. A major leap forward in the study of the giant algal virus diversity was the recognition of conserved regions in the B family DNA polymerase core genes of the algal viruses (Chen & Suttle, 1995) which were classified in the Phycodnaviridae family based on their host range. Degenerate primers (AVS1 and AVS2) were designed from the conserved regions and were used to amplify sequences from several Phycodnaviridae isolates and also from natural sea water community, indicating that this primer set could be a useful tool to study algal viruses in different environments. PCR-DGGE analysis of fragments generated using these primers recovered viruses with high sequence similarity (>98%) from distant locations (Antarctica and British Columbia), demonstrating that closely related giant algal viruses can occur at geographically distant locations (Short & Suttle, 2002). A long term study using these primers on a single coastal site demonstrated that although algal virus diversity changes coincided with tide height, salinity or chl *a* concentration, it was not always coupled with the changes in the eukaryotic community, indicating succession among non-host eukaryotes (Steven & Curtis, 2003). Interestingly, diversity study using these primer sets from subtropical coastal water of Hawai'i led to detection of inteins in a number of putative Phycodnaviridae members in the *polB* genes (Culley *et al.*, 2008). However, ecological and evolutionary consequences of inteins in giant viruses remain unknown.



Figure 1.1: Maximum likelihood phylogenetic reconstruction of NCLDV major capsid protein sequences from environmental metatranscriptomic datasets generated from an alkaline soil sample (NCBI ID: SRP043976), the Amazon River and River Delta (SRP037995, SRP039544), the North Pacific Ocean (SRP052554), Station ALOHA in the tropical Pacific Ocean (CAM_SMPL_000824 at iMicrobe.us), and the North Sea (ERP004582). This tree demonstrates that active giant virus infections from diverse phylogenetic groups can be observed in many different environments. Adapted from Wilhelm *et al.* (2016).

While AVS primers provided novel insights on the algal giant viruses in aquatic systems, studies have revealed that these primers can't detect some of the Phycodnaviridae members like *Heterosigma akashiow* viruses and *Emiliania huxleyi* viruses and are biased towards amplifying *Micromonas pusilla* virus isolates (Short & Short, 2008, Clasen & Suttle, 2009), which might lead to skewed estimation of diversity and dynamics. Along with AVS primers targeting PolB, several other degenerate primers targeting *polB* or other conserved genes within the giant algal viruses have been developed and used to study the diversity of these viruses. '*mcp'* primers, designed from the Major capsid protein of the Phycodnaviridae members (Larsen *et al.*, 2008) revealed that a number of amplicons from cultured and environmental samples were phylogenetically closer to the Mimiviridae members rather than other Phycodnaviridae members. In retrospective, this was an interesting discovery – as discussed earlier, a number of algal viruses actually belong to the Mimiviridae clade (Monier *et al.*, 2008, Santini *et al.*, 2013). Neither AVS or '*mcp'* primers can amplify sequences from *Emiliania huxleyi* viruses (EhV). A primer set specifically targeting the major capsid proteins of EhV (Schroeder *et al.*, 2002) was used to assess their diversity across the North Atlantic, which greatly increased the known diversity and richness of EhVs (Rowe *et al.*, 2011).

Along with amplicon based targeted studies, a number of eye opening discoveries regarding the widespread diversity and phylogeny of giant viruses came from metagenomics. The first such study was carried out on the Global Ocean Sampling (GOS) dataset, revealing numerous 'Mimivirus-like' sequences. Phylogenetic analysis of these sequences suggested that a large proportion of these sequences probably originated from viruses infecting phytoplankton (Monier *et al.*, 2008). A subsequent study on the same dataset used DNA polymerase as a marker and estimated Mimivirus and related algal virus like sequences to be the second most abundant to bacteriophages. (Monier *et al.*, 2008). Interestingly, a survey of the Indian Ocean samples from the GOS expedition (Williamson *et al.*, 2012) revealed that proportion of various giant virus families largely varied between the 'viral fraction' and 'larger fraction' of the samples with a number of families (including Mimiviridae) being more represented in the larger fraction. This indicated that filtration approach for virome generation can cause biased estimation of giant viruses. Another large scale marine metagenomics (TARA oceans transect) study found that both proposed Megaviridae (extended Mimiviridae) and Phycodnaviridae members dominated the transect, with an estimated 10^4-10^5 NCLDV genomes/ml in the photic zone (Hingamp *et al.*, 2013). Together, these metagenomics studies placed marine NCLDVs in the spotlight - hinting to their roles in mortality of highly diverse eukaryotic phytoplanktons and global biogeochemical cycle. Recently, a virome library generated from coral *Montastraea cavernosa* was found to contain signatures of NCLDVs and ssRNA viruses (Correa *et al.*, 2013).The recognition of algal viruses in the Mimiviridae family also fueled

research on assessing the diversity of 'Mimivirus like' algal viruses. A primer set designed from MutS8, a gene found in a number of cultured algal Mimivirus representatives, revealed novel group of algal Mimiviruses in the marine environment (Wilson *et al.*, 2014).

A number of giant viruses have also been found to be associated with algal blooms – rapid growth of a particular alga that might cover a large area of the ocean, sometimes with a harmful effect on the co-occurring organisms and even human health (Anderson *et al.*, 2011). Giant viruses associated with algal blooms include *E. huxleyi* viruses (EhV) (Wilson *et al.*, 2002), *Phaeocystis globosa* virus (PgV) (Brussaard *et al.*, 2004), *Heterosigma akashiwo* virus (HaV) (Nagasaki & Yamaguchi, 1997) and *Aureococcus anophagefferens* virus (AaV) (Gastrich *et al.*, 2004). *Emiliania huxleyi* is the most abundant coccolithopore in the world's oceans, and can cause massive blooms visible from outer space. Given their ecological role in fixing a large amount of atmospheric carbon and impact on the planetary radiation budget, role of viruses in *E. huxleyi* blooms have been extensively studied. *In situ* measurement of a coccolithopore bloom in the western English Channel showed that the high reflectance area of the bloom characteristic of bloom collapse had a higher concentration of large virus like particles (LVLP) compared to the low reflectance area implying that viruses are likely the cause of the bloom demise (Wilson *et al.*, 2002). Two viruses showing lytic activity against *E. huxleyi* were also isolated from this bloom. Subsequent studies looked at the viral diversity and succession (Schroeder *et al.*, 2003, Martínez *et al.*, 2007), revealing genetically rich and stable EhV communities over years. A recent study tracked the complete lifecycle (from exponential growth to rapid demise) of an *E. huxleyi* bloom in the North Atlantic using both *in situ* biological measurements and multisatellite data addressing key questions regarding bloom turnover quantification and factors involved in its demise (Lehahn *et al.*, 2014). An estimated 24,000 tons of carbon was produced during the bloom, while high level of *EhV* activity (*in situ* concentration: 2.5×10^6 VLPs/ml) was detected during the demise phase. This is a striking example of viral turnover of photosynthetically fixed carbon, which can fuel marine food webs, short-circuit carbon transfer to higher trophic levels and can promote carbon export in the deep sea (Wilhelm & Suttle, 1999) Indeed, a natural question emerges from this observation regarding how EhVs can spread across the bloom to induce mortality across such a massive geographical space. Studies have shown that zooplanktons, especially copepods, can act as transmission vectors for EhVs in the bloom (Frada *et al.*, 2014). It was found that the copepod under study (*Calanus* sp.) can ingest the virus infected cells and can release the viruse back into the water in the fecal pellets. Thus they can transmit the virues between microscale and mesoscale (>1 km) patches, thereby mediating infection of healthy cells (Frada *et al.*, 2014). It was also demonstrated that EhVs can travel potentially longer distances through aerosols

(Sharoni *et al.*, 2015). The extensive body of research on the *E. huxleyi* blooms provides key information and set the stage for investigating the role of giant viruses in other eukaryotic algal bloom systems.

Phaeocystis globosa can form dense blooms in the temperate coastal regions of North Sea with a cell number reaching upto 10^8 cells/ml (Brussaard *et al.*, 2004). In both mesocosm and field studies, PgV abundance was found to be closely linked to that of their host and were found to be 30 to 100 fold higher than host abundance during the bloom maxima, indicating that they are important players in the collapse of the blooms (Brussaard *et al.*, 2005, Anne-Claire *et al.*, 2006). *Phaeocystis* can dominate as both single cells and colonial forms (Brussaard *et al.*, 2004). An ecosystem modelling study calibrated with mesocosm data found that fraction of infectious PgV increased over the course of the bloom – coinciding with the increase in single celled forms. This study also found prior to the initiation of the bloom, fraction of infectious PgV was very low (0.0005%), indicating that a large fraction of PgV standing stock lost infectivity as they persisted in the water column through autumn till spring (Ruardij *et al.*, 2005).

Brown tide blooms, caused by pelagophyte *Aureococcus anophagefferens*, are recurrent phenomenon along the east coast of USA (Gobler *et al.*, 2005) and have recently been observed in the coastal regions of China (Zhang *et al.*, 2012) and South Africa (Probyn *et al.*, 2001). These blooms are of a significant economic and health concern and their causes and effects have been investigated for decades (Gobler *et al.*, 2005). High abundance of *Aureococcus* (~1.0 X 10^5 cells/mL) can lead to high mortality of bay scallops and blue mussels, which could be due to a toxin-like activity. The toxin-like component seems to be chemical in nature (Gainey and Shumway, 1991); however, this component hasn't been identified or characterized. At the same time, brown tide blooms can destroy the eel grass habitats of the bivalves due to intense light attenuation (Gobler *et al.*, 2005). Fluctuation in nutrient dynamics is attributed to be the major driver of these blooms, anthropogenic activities likely playing a significant role (Gobler *et al.*, 2005). The genome of *Aureococcus* provided important insights on the genetic basis of its success over co-occurring phytoplankton during the brown tide blooms (Gobler *et a.l*, 2011). *Aureococcus* genome has more genes involved in organic carbon and nitrogen metabolism, providing an explanation for its success over other phytoplankton during elevated level of dissolved organic matter and reduced level of dissolved inorganic nitrogen (Gobler *et al.*, 2011). Suggested roles for viruses in controlling or modulating the severity of these blooms goes back to 1988, when Sieburth *et al.* (1988) observed the ultrastructure of *Aureococcus* collected from a bloom in Narragansett Bay. A number of cells were full with icosahedral virus particles, indicating that along with grazing, viruses might be one of the contributing factors to the senescence of the bloom. The first observation of virus mediated lysis of *Aureococcus* in culture was made by Milligan & Cosper (1994). They isolated a 'phage-like' virus, which was shown to lyse *Aureococcus*, suggesting that the same virus might play a role in brown tide bloom decline (Figure 1.2).

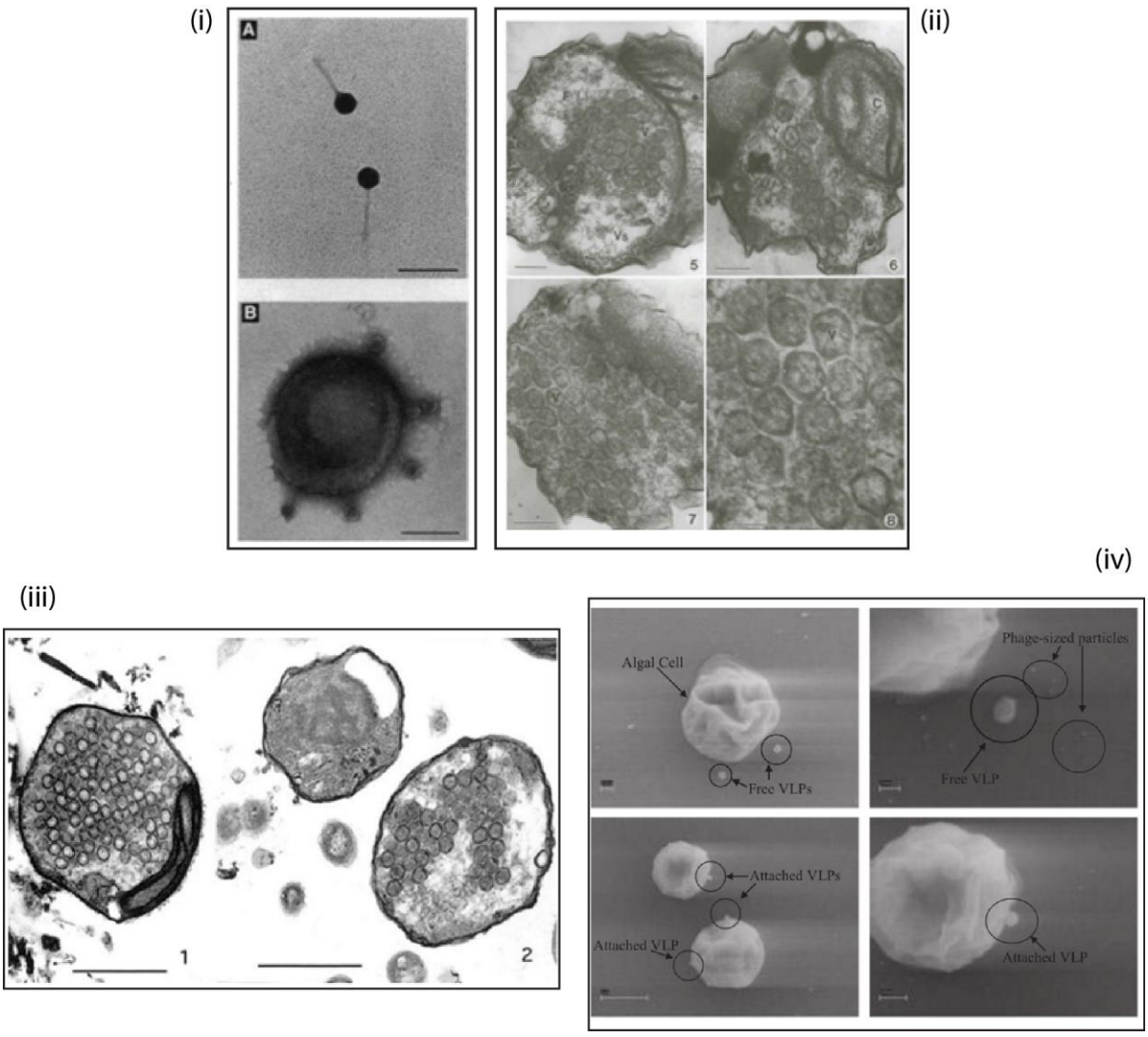


Figure 1.2: A number of ultrastructure analyses of *Aureococcus* and associated virus from different studies. Figure 2 (i) is adapted from Milligan & Cosper (1994), which shows detection of a phage like particle suggested to be responsible for mortality in *Aureococcus*. Panel (A) - Free phage like particle (scale bar = 100 nm), panel (B) - phage like particle attached to an *Aureococcus* cell (scale bar = 200 nm). Figure 2(ii), adapted from Gastrich *et al.* (1998), shows virus capsids inside the host cells at advanced stages of infection cycle. Scale bar = 0.25 µM. Notice the difference in the size of the particle compared to the early study by Milligan and Cosper (1994) (Figure 2(i)). Figure 2(iii) shows similarly large virus like particles inside the cells directly isolated from a 2002 bloom in New Jersey (Gastrich *et al.*, 2002)(Gastrich et al 2004) (scale bars = 1 µM). Figure 2(iv): Scanning electron micrograph of *Aureococcus* cultures exposed to viral lysates 30 min post-infection. Notice the large virus attached to the cell. (upper left: scale bar, 200 nm; upper right: scale bar, 1 µM; lower left: scale bar, 1 µM; lower right: scale bar, 200 nm). VLP - virus like particle.

The follow up study on the ultrastructure of the *Aureococcus* cells upon virus infection revealed some interesting details about the structure of the virus (Gastrich *et al.*, 1998). The capsid size of the newly formed virion particles seemed to be larger (~140-160 nm) compared to that of the 'phage-like' particle recorded previously (Figure 1.2). This is surprising, given that the cells were infected with viruses that were same as ones used in the 1998 study. A study by Gastrich *et al.* (2002) on the ultrastructure of virus infected cells directly from a bloom in New Jersey demonstrated similarly large icosahedral particles in the algal cells (Figure 1.2). This study also showed that as high as 37% of the algal cells can be infected during the late stage/peak of the bloom – indicating that viruses might be an important factor in the decline of the brown tide blooms (Probyn *et al.*, 2001). This observation strongly suggested that the virus infecting *Aureococcus* might not be a phage at all, but a virus much larger than that. However, the molecular characteristics of this virus were largely unknown, apart from its size and icosahedral shape. Subsequently, a large virus was isolated from the bloom water, and the nature of its interaction with *Aureococcus* was characterized in terms of host range and influence of abiotic factors like light and temperature (Gobler *et al.*, 2007). This particular virus exhibited strain specificity against *Aureococcus*, with several strains being resistant to its infection (Gobler *et al.*, 2007). Despite all the indications that a 'giant' virus infects *Aureocccus* rather than a phage from the ultrastructure analysis, the definitive evidence of this virus being the pathogen of *Aureococcus* was still lacking. It is possible that the virus lysate contains multiple isolates with distinct morphologies, including phages infecting cooccuring bacteria, thereby confounding the results. To resolve this a study was undertaken (Rowe *et al*., 2008) that strictly followed Koch's postulates to establish the cause of mortality in *Aureococcus* cells. They showed that the virus obtained from infection and lysis of *Aureococcus* has a giant icosahedral capsid and can attach to the *Aureococcus* cells (Figure 2). The virus was approximately 140 nm in diameter and was suggested to belong to 'Phycodnaviridae' clade of Nucleocytoplasmic Large DNA viruses (NCLDVs) based on its morphology and host. Together, the evidence suggest that the phage-like particle isolated by Milligan & Cosper (1994) might have originated from the bacteria in the *Aureococcus* culture.

An apparently common aspect of interaction of giant viruses with their hosts is the coexistence of different strains of hosts and viruses. A number of giant viruses were found to have different host specificity at the strain level. A study of the genotypic diversity of *E. huxleyi* and their viruses found a genetically rich, yet stable community of *E. huxleyi* and EhV in the Norwegian fjord over two separate years (Martínez *et al.*, 2007). Interestingly, the same two genotypes of EhV dominated the exponential and termination phase of both the blooms, even though initially at least four and six genotypes were coexisting during the 2003 and 2004 blooms, respectively. Host strain specificity was also reported for PgV, where some viral strains infected a broad range of host strains, while some were found to have a

more restricted host range (Brussaard *et al.*, 2005). Coexistence of such a diversity of hosts and viruses indicates that giant virus infection can influence the host clonal composition and abundance. *Heterosigma akashiwo* is a raphidophyte which causes ecosystem devastating algal blooms in the coastal waters of both the Northern and Southern Hemisphere. A number of clones of a giant dsDNA virus (HaV) have been isolated which are known to lyse different strains of *H. akashiow* (Nagasaki & Yamaguchi, 1998). A strain of *H. akashiwo* virus (HaV01) was inoculated into natural seawater containing *H. akashiwo* to assess its potential as a bloom control agent*;* however, little or no inhibition of growth was observed (Nagasaki *et al.*, 1999). This was attributed to the presence of both susceptible and resistant strains of *H. akashiwo*. Thus, from a practical point of view, intraspecies specificity of viruses might make them less attractive to control algal blooms.

The emergence of resistant strains is a likely mechanism of ensuring survival of the host at the population level upon viral attack. However, defense mechanisms at the individual cell level might also exist. One such example is the elegant 'Cheshire cat' strategy adopted by *E. huxleyi*. The diploid calcified *E. huxlei* cells are susceptible to EhV infection, however, the non-calcified haploid stages are 'invisible' to the viruses. Interestingly, exposing the diploid cells to EhV induces the haploid phase (Frada *et al.*, 2008). On the flipside of this coin, there are phaeoviruses, who can 'sneak' into the host genome during the gamete and/or spore formation (Stevens *et al.*, 2014), thus maintaining a 'provirus' like lifestyle. In opposition to 'acute' life strategy of majority of the giant algal viruses, phaeoviruses thus maintain a persistent existence within the host genome and have no noticeable negative impact on the host fitness. However, during the production of virus particles, the spore forming organs (zoidangium) of the host become markedy deformed (Stevens *et al.*, 2014). Phaeoviruses are the only known giant viruses to maintain a parsistent lifestyle to date.

Molecular aspects of giant virus – host interactions

One of the pivotal points denoting the ecological importance of viruses is their sheer abundance in nature. They are considered the most abundant biological entity, with an estimated ~10^{30} viruses dominating the planet's oceans (Suttle, 2007). While the sheer abundance of viruses leaves no doubt about their profound role in the global biogeochemical cycles, counting just virion particles can shift the focus away from the fact that free virus particles are an outcome of altered physiological profile of host cells, which eventually lead to production and assembly of viral components. The virocell concept was first introduced by Forterre (2011), which emphasized the fact that free virion particles are only one of the states that a virus may exist in, which is the extracellular infectious state. A virocell defines the autonomous state of vegetative replication of viruses inside the host cell, which is different from a 'ribocell' or a healthy cell

in terms of metabolic organization (Forterre, 2011). Thus, this concept inspires understanding the nanoscale underpinnings of a virocell, which can have biogeochemical implications at a large geographical scale. Indeed, the virocell is the melting pot for continuous host-virus evolutionary arms race and modulation of nutrient fluxes among microorganisms (Rosenwasser *et al.*, 2016). Despite the possibility that the interaction between giant viruses with their highly complex eukaryotic hosts result in a formidably large number of cellular processes to be affected, not much is known in this regard. Using high throughput techniques like transcriptomics and/or metabolomics, a few ecologically relevant host-virus systems have been studied to understand how viral infection can rewire the host metabolism during virus replication (Rosenwasser *et al.*, 2016). Among these, the *E. huxleyi* –EhV system is perhaps most extensively researched, given the profound biogeochemical implications of this system as discussed earlier. By combined transcriptomics and metabolomics experiment, it was shown that virus infection lead to increased transcription of host genes involved in fatty acid synthesis that facilitated the virus assembly (Pagarete *et al.*, 2011, Rosenwasser *et al.*, 2014, Rosenwasser *et al.*, 2016). Interestingly, sphingolipid biosynthesis genes encoded in the EhV genome were upregulated, while the host counterparts were downregulated. This indicated a putative host antiviral mechanism, which was compensated by the ability of the virus to direct synthesis of sphingolipids by the genes that it acquired from the host through HGT (Rosenwasser *et al.*, 2014). This is an example of how a particular host-virus pair coevolve, eventually shaping the ecological success of each other. Another notable aspect of *E. huxleyi* virocell is the remodeling of the antioxidant network during virus replication (Sheyn *et al.*, 2016), which will be discussed in detail in Chapter II. In the early phase of PBCV-1 infection of *Chlorella*, a number of host pathways were up and downregulated, having implications in both host defense and virus strategy to subvert such defenses (Rowe *et al.*, 2014). A number of studies have looked at the transcriptomic and metabolic rewiring of prokaryotic cells upon infection with phages. Phage infection of *Sulfitobacter* sp., a member of the *Roseobacter* lineage abundant in the marine system, led to increased metabolic activity in the cells compared to control, with 71% of the detected metabolites significantly elevated in the infected cells (Ankrah *et al.*, 2014). A recent study (Doron *et al.*, 2016) on a cyanophage having broad host range (Syn9 T4-like) infecting phylogenetically and ecologically distinct *Synechococcus* strains (WH7803, WH8102 and WH8109) showed that the virus adopts a similar temporal pattern of gene expression upon infecting these three strains. However, a large number of host response genes were clustered in the hypervariable genomic islands, indicating that these genomic islands might be central to host response and defense towards phage infection (Doron *et al.*, 2016). Howard-Varona *et al.* (2016) looked at a host-virus relationship of similar nature, where a single virus strain (podovirus φ38:1) can infect 13 out of 21 *Cellulophaga baltica* strains, a bacteria abundant in human gut (Howard-Varona *et al.*, 2016). Two of these host strains – NN016038 and no. 18, are nearly identical in terms of nucleotide

sequences (> 99.99%) of 93% of the genes. Yet, podovirus φ38:1 shows very different infection efficiency on these two strains, with high efficiency when infecting NN016038 (Howard-Varona *et al.*, 2016). The alternative host (no. 18) showing markedly low infection efficiency actually overexpressed DNA degradation genes and underexpressed translation genes, which eventually delayed and reduced phage DNA production. (Howard-Varona *et al.*, 2016). These studies reveal that even at strain level, host response varies markedly upon virus infection, giving insights into the mechanism of resistance and host-virus evolutionary arms race manifested in the physiological processes. Such findings might have important implications in the giant virus- eukaryotic host systems and might inspire research addressing similar questions in these systems.

Objectives

Research presented in this dissertation addresses a number of broad ecological and evolutionary questions pertaining to giant viruses using harmful brown tide algae *Aureococcus anophagefferens* and the giant virus infecting it (AaV*)* as an ecologically relevant model system. The possible role of AaV in the demise of the brown tide blooms was examined *in situ*, while the molecular basis of their interaction was also elucidated. With the insights obtained from this model system, we aimed to expand the knowledge on the evolutionary trajectories, ecological dynamics and molecular underpinnings of giant virus-host interactions.

In Chapter II, we aim to understand the evolutionary trajectory and phylogenetic affiliation of AaV in the context of other giant viruses through the lens of genomics. There are several reasons behind this approach. Firstly, the genome of a biological entity contains valuable information regarding its functional potential in an ecological context and signatures of its evolutionary history. Secondly, the functional landscape of a cellular organism or virus emerges from the genes. For this reason, the genome is a powerful reference to study the downstream processes that follows the central dogma. Finally, the genomic information can also be used to design molecular tools to probe the diversity and distribution of a biological entity. For the model system under study, the host genome has already been sequenced, and obtaining the genome sequence of the viral partner is a critical step to the address the research questions in the subsequent chapters of this dissertation.

In Chapter III, we studied the molecular basis of AaV- host interaction at gene transcription level throughout the course of infection. To achieve this we leveraged the available genome sequence data of both the host and the virus. As has been observed for other virocells, successful infection of the host by its virus is expected to produce an altered transcriptomic and metabolic response compared to an uninfected, healthy cell. One of our goals was to inspect the immediate response of the host upon virus infection, which might have implications in both host defense and viral subversion of such defenses. In addition, we were also interested in the dynamics of major cellular processes that might be modulated to facilitate virus production.

In Chapter IV and V, we address questions regarding the ecology of viruses that infect microeukaryotes, with an emphasis on giant viruses. The first question we asked is “how is the diversity and dynamics of AaV and associated algae infecting giant viruses modulated during Brown tide blooms?”. Answering this question is important to substantiate the possible role of AaV in the demise of the brown tide blooms.

The brown tide environment also provided a 'model ecosystem' to understand the ecological dynamics of giant algal viruses in the Mimiviridae group, which could further demonstrate their ecological relevance. We designed primers that specifically targeted the major capsid gene of algae infecting mimiviridae members (including AaV) to probe their dynamics throughout the peak and demise of a brown tide bloom. An important goal of this work was to provide new knowledge regarding the dynamics of this newly recognized phylogenetic group of viruses, which seem to be widely distributed in the world's oceans.

The seeming diversity and wide distribution of giant viruses in the world's oceans opens a new research frontier in marine virus ecology. New toolsets have to be developed to track their spatiotemporal dynamics, phylogeography, potential hosts and role in marine biogeochemistry.

In Chapter V, we explored metatranscriptomics as a potential tool to address pressing questions regarding marine viral dynamics – specifically the microeukaryote infecting viruses including the giants. Rather than capturing community DNA as is done with metagnomics, metatranscriptomics only captures RNA, but from both the cellular organisms and the actively infecting DNA viruses associated with this same community. In principle, this approach should also capture the sequences from RNA viruses – from both actively infecting and free virus particles. We analyzed time series metatranscriptomic datasets generated during the peak and demise of a brown tide bloom. We specifically asked the following questions,

1. Can we detect the active infection of AaV in community gene expression data?

2. Can we draw statistically meaningful correlations among eukaryotic viruses and their potential hosts using *Aureococcus* –AaV relationship as an *in situ* internal standard?

We later extended this approach to another time course metatranscriptomics dataset from Narragansett Bay to tease apart the temporal dynamics of the active microeukaryote virus community and also explored the possibility to draw potential host-virus relationships. We found metatranscriptomics to be a promising tool to tease apart spatiotemporal patterns in microeukaryote virus dynamics and host-virus association.

Together, this body of research provides key insights on the molecular processes underlying the giant algal virus – host interactions, gleans new knowledge regarding their ecosystem dynamics and tests molecular tools that will be valuable in future research on aquatic viruses infecting microeukaryotes.

List of References

Abrahão JS, Dornas FP, Silva LC, Almeida GM, Boratto PV, Colson P, La Scola B & Kroon EG (2014) *Acanthamoeba polyphaga* Mimivirus and other giant viruses: an open field to outstanding discoveries. *Virology journal* **11**:120.

Anderson DM, Cembella AD & Hallegraeff GM (2011) Progress in Understanding Harmful Algal Blooms: Paradigm Shifts and New Technologies for Research, Monitoring, and Management. *Annual Review of Marine Science* **4**: 143-176.

Ankrah NY, May AL, Middleton JL, Jones DR, Hadden MK, Gooding JR, LeCleir GR, Wilhelm SW, Campagna SR & Buchan A (2014) Phage infection of an environmentally relevant marine bacterium alters host metabolism and lysate composition. *The ISME journal* **8**: 1089-1100.

Anne-Claire B, Anna AMN, Marcel JWV & Corina PDB (2006) Virally induced mortality of *Phaeocystis globosa* during two spring blooms in temperate coastal waters. *Aquatic Microbial Ecology* **44**: 207-217.

Arslan D, Legendre M, Seltzer V, Abergel C & Claverie J-M (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci USA* **108**: 17486-17491.

Bamford DH (2003) Do viruses form lineages across different domains of life? *Research in Microbiology* **154**: 231-236.

Blanc G, Duncan G, Agarkova I, *et al.* (2010) The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *The Plant cell* **22**: 2943-2955.

Boyer M, Madoui M-A, Gimenez G, La Scola B & Raoult D (2010) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. *PLoS ONE* **5**: e15530.

Boyer M, Yutin N, Pagnier I, *et al.* (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA*. 106:21848-53.

Boyer M, Azza S, Barrassi L, *et al.* (2011) Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proc Natl Acad Sci USA* **108**: 10296-10301.

Brussaard CP, Short SM, Frederickson CM & Suttle CA (2004) Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton Phaeocystis globosa (Prymnesiophyceae). *Appl Environ Microbiol* **70**: 3700-3705.

Brussaard CPD, Kuipers B & Veldhuis MJW (2005) A mesocosm study of *Phaeocystis globosa* population dynamics: I. Regulatory role of viruses in bloom control. *Harmful Algae* **4**: 859-874.

Brussaard CPD, Short SM, Frederickson CM & Suttle CA (2004) Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton *Phaeocystis globosa* (Prymnesiophyceae). *Applied and Environmental Microbiology* **70**: 3700-3705.

Chen F & Suttle CA (1995) Amplification of DNA polymerase gene fragments from viruses infecting microalgae. *Appl Environ Microbiol* **61**: 1274-1278.

Clasen JL & Suttle CA (2009) Identification of Freshwater Phycodnaviridae and Their Potential Phytoplankton Hosts, Using DNA pol Sequence Fragments and a Genetic-Distance Analysis. *Applied and Environmental Microbiology* **75**: 991-997.

Claverie J-M & Abergel C (2009) Mimivirus and its Virophage. *Annual Review of Genetics* **43**: 49-66.
Claverie JM, Ogata H, Audic S, Abergel C, Suhre K & Fournier PE (2006) Mimivirus and the emerging concept of "giant" virus. *Virus Res* **117**: 133-144.

Colson P, Lamballerie X, Fournous G & Raoult D (2012) Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology* **55**:321-32.

Correa AMS, Welsh RM & Vega Thurber RL (2013) Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *The ISME journal* **7**: 13-27.

Cottrell MT & Suttle CA (1991) Wide-spread occurrence and clonal variation in viruses which cause lysis of a cosmopolitan, eukaryotic marine phytoplankter, *Micromonas pusilla*. *Marine Ecology Progress Series* **78**: 1-9.

Culley AI, Asuncion BF & Steward GF (2008) Detection of inteins among diverse DNA polymerase genes of uncultivated members of the Phycodnaviridae. *The ISME journal* **3**: 409-418.

DeAngelis PL, Jing W, Graves MV, Burbank DE & Van Etten JL (1997) Hyaluronan Synthase of Chlorella Virus PBCV-1. *Science* **278**: 1800-1803.

Desnues C, La Scola B, Yutin N, *et al.* (2012) Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci USA* **109**: 18078-18083.

Dodds JA & Cole A (1980) Microscopy and biology of *Uronema gigas*, a filamentous eucaryotic green alga, and its associated tailed virus-like particle. *Virology* **100**: 156-165.

Doron S, Fedida A, Hernandez-Prieto MA, *et al.* (2016) Transcriptome dynamics of a broad host-range cyanophage and its hosts. *The ISME journal* **10**: 1437-1455.

Filee J (2014) Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible part of the iceberg? *Virology* **466-467**: 53-59.

Fischer MG & Hackl T (2016) Genome Integration and Reactivation of the Virophage Mavirus In the Marine Protozoan Cafeteria roenbergensis. *bioRxiv*.

Fischer MG, Allen MJ, Wilson WH & Suttle CA (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA*. **107:** 19508-13.

Forterre P (1999) Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. *Molecular Microbiology* **33**: 457-465.

Forterre P (2006) Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA* **103**: 3669-3674.

Forterre P (2010) Defining life: the virus viewpoint. *Origins of life and evolution of the biosphere : the journal of the International Society for the Study of the Origin of Life* **40**: 151-160.

Forterre P (2011) Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *Comptes Rendus Chimie* **14**: 392-399.

Forterre P (2013) The virocell concept and environmental microbiology. *The ISME journal* **7**: 233-236.

Frada M, Probert I, Allen MJ, Wilson WH & de Vargas C (2008) The "Cheshire Cat" escape strategy of the coccolithophore Emiliania huxleyi in response to viral infection. *Proc Natl Acad Sci USA* **105**: 15944-15949.

Frada MJ, Schatz D, Farstey V, Ossolinski JE, Sabanay H, Ben-Dor S, Koren I & Vardi A (2014) Zooplankton may serve as transmission vectors for viruses infecting algal blooms in the ocean. *Current biology : CB* **24**: 2592-2597.

Gainey LF & Shumway SE (1991). The physiological effect of *Aureococcus anophagefferens* ("brown tide") on the lateral cilia of bivalve mollusks. *Biology Bulletin* **181**: 298-306.

Gastrich M, Anderson OR & Cosper E (2002) Viral-like particles (VLPS) in the alga, *Aureococcus anophagefferens* (Pelagophyceae), during 1999–2000 brown tide blooms in Little Egg Harbor, New Jersey. *Estuaries* **25**: 938-943.

Gastrich M, Leigh-Bell J, Gobler C, Roger Anderson O, Wilhelm S & Bryan M (2004) Viruses as potential regulators of regional brown tide blooms caused by the alga,Aureococcus anophagefferens. *Estuaries* **27**: 112-119.

Gastrich MD, Anderson OR, Benmayor SS & Cosper EM (1998) Ultrastructural analysis of viral infection in the brown-tide alga, *Aureococcus anophagefferens* (Pelagophyceae). *Phycologia* **37**: 300-306.

Ghigo E, Kartenbeck J, Lien P, Pelkmans L, Capo C, Mege JL & Raoult D (2008) Ameobal pathogen Mimivirus infects macrophages through phagocytosis. *PLoS Pathog* **4**: e1000087.

Gobler CJ, Lonsdale DJ & Boyer GL (2005) A review of the causes, effects, and potential management of harmful brown tide blooms caused by *Aureococcus anophagefferens* (Hargraves et Sieburth). *Estuaries* **28**: 726-749.

Gobler CJ, Anderson OR, Gastrich MD & Wilhelm SW (2007) Ecological aspects of viral infection and lysis in the harmful brown tide alga *Aureococcus anophagefferens*. *Aquatic Microbial Ecology* **47**: 25-36.

Granoff A, Came PE & Breeze DC (1966) Viruses and renal carcinoma of Rana pipiens. *Virology* **29**: 133-148.

Hingamp P, Grimsley N, Acinas SG, *et al.* (2013) Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *The ISME journal* **7**: 1678-1695.

Howard-Varona C, Roux S, Dore H, Solonenko NE, Holmfeldt K, Markillie LM, Orr G & Sullivan MB (2016) Regulation of infection efficiency in a globally abundant marine Bacteriodetes virus. *The ISME journal*. doi: 10.1038/ismej.2016.81

Iyer LM, Aravind L & Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *Journal of virology* **75**: 11720-11734.

Iyer LM, Balaji S, Koonin EV & Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research* **117**: 156-184.

Kawasaki T, Tanaka M, Fujie M, Usami S, Sakai K & Yamada T (2002) Chitin synthesis in Chlorovirus CVK2-infected chlorella cells. *Virology* **302**: 123-131.

Kerepesi C & Grolmusz V (2016) Giant viruses of the Kutch Desert. *Arch Virol* **161**: 721-724.

Khan M, La Scola B, Lepidi H & Raoult D (2007) Pneumonia in mice inoculated experimentally with *Acanthamoeba polyphaga* Mimivirus. *Microbial pathogenesis* **42**: 56-61.

Koonin EV & Yutin N (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* **53**: 284-292.

Koonin EV, Dolja VV & Krupovic M (2015) Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**: 2-25.

Krupovic M & Koonin EV (2015) Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Micro* **13**: 105-115.

Krupovic M, Bamford DH & Koonin EV (2014) Conservation of major and minor jelly-roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide viruses. *Biology direct* **9**: 1-7.

La Scola B, Desnues C, Pagnier I, *et al.* (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* **455**: 100-104.

Larsen JB, Larsen A, Bratbak G & Sandaa RA (2008) Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. *Appl Environ Microbiol* **74**: 3048-3057.

Lartigue A, Burlat B, Coutard B, Chaspoul F, Claverie J-M & Abergel C (2015) The Megavirus Chilensis Cu,Zn-Superoxide Dismutase: the First Viral Structure of a Typical Cellular Copper Chaperone-Independent Hyperstable Dimeric Enzyme. *Journal of virology* **89**: 824-832.

Legendre M, Bartoli J, Shmakova L, *et al.* (2014) Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci USA*. **111:**4274-4279.

Legendre M, Lartigue A, Bertaux L, *et al.* (2015) In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc Natl Acad Sci U S A* **112**: E5327-5335.

Lehahn Y, Koren I, Schatz D, *et al.* (2014) Decoupling physical from biological processes to assess the impact of viruses on a mesoscale algal bloom. *Current Biology* **24**: 2041-46.

Martínez JM, Schroeder DC, Larsen A, Bratbak G & Wilson WH (2007) Molecular Dynamics of Emiliania huxleyi and Cooccurring Viruses during Two Separate Mesocosm Studies. *Applied and Environmental Microbiology* **73**: 554-562.

Maumus F, Epert A, Nogué F & Blanc G (2014) Plant genomes enclose footprints of past infections by giant virus relatives. *Nat Commun* **5**:4268.

Mayer JA & Taylor FJR (1979) A virus which lyses the marine nanoflagellate Micromonas pusilla. *Nature* **281**: 299-301.

Milligan KL & Cosper EM (1994) Isolation of Virus Capable of Lysing the Brown Tide Microalga, Aureococcus anophagefferens. *Science* **266**: 805-807.

Monier A, Claverie J & Ogata H (2008) Taxonomic distribution of large DNA viruses in the sea. *Genome biology* **9**: R106.

Monier A, Larsen J, Sandaa R-A, Bratbak G, Claverie J-M & Ogata H (2008) Marine mimivirus relatives are probably large algal viruses. *Virology journal* **5**: 12.

Monier A, Pagarete A, de Vargas C, Allen MJ, Read B, Claverie JM & Ogata H (2009) Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome research* **19**: 1441-1449.

Nagasaki K & Yamaguchi M (1997) Isolation of a virus infectious to the harmful bloom causing microalga Heterosigma akashiwo (Raphidophyceae). *Aquatic Microbial Ecology* **13**: 135-140.

Nagasaki K & Yamaguchi M (1998) Intra-species host specificity of HaV (*Heterosigma akashiwo* virus) clones. *Aquatic Microbial Ecology* **14:** 109-112.

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

Pagarete A, Le Corguillé G, Tiwari B, Ogata H, de Vargas C, Wilson WH & Allen MJ (2011) Unveiling the transcriptional features associated with coccolithovirus infection of natural Emiliania huxleyi blooms. *FEMS Microbiology Ecology* **78**: 555-564.

Pagnier I, Reteno DGI, Saadi H, *et al.* (2013) A Decade of Improvements in Mimiviridae and Marseilleviridae Isolation from Amoeba. *Intervirology* **56**: 354-363.

Philippe N, Legendre M, Doutre G, *et al.* (2013) Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341**: 281-286.

Piacente F, Marin M, Molinaro A, *et al.* (2012) Giant DNA Virus Mimivirus Encodes Pathway for Biosynthesis of Unusual Sugar 4-Amino-4,6-dideoxy-d-glucose (Viosamine). *Journal of Biological Chemistry* **287**: 3009-3018.

Pickett-Heaps JD (1972) A possible virus infection in the green alga *Oedogonium. Journal of Phycology* **8**: 44-47.

Probyn T, Pitcher G, Pienaar R & Nuzzi R (2001) Brown tides and mariculture in Saldanha Bay, South Africa. *Marine pollution bulletin* **42**: 405-408.

Prudhomme S, Bonnaud B & Mallet F (2005) Endogenous retroviruses and animal reproduction. *Cytogenetic and Genome Research* **110**: 353-364.

Raoult D, Scola BL & Birtles R (2007) The Discovery and Characterization of Mimivirus, the Largest Known Virus and Putative Pneumonia Agent. *Clinical Infectious Diseases* **45**: 95-102.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M & Claverie J-M (2004) The 1.2-Megabase Genome Sequence of Mimivirus. *Science* **306**: 1344-1350.

Read BA, Kegel J, Klute MJ, *et al.* (2013) Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* **499**: 209-213.

Rosenwasser S, Ziv C, Creveld SG & Vardi A (2016) Virocell Metabolism: Metabolic Innovations During Host-Virus Interactions in the Ocean. *Trends Microbiol*. DOI: http://dx.doi.org/10.1016/j.tim.2016.06.006

Rosenwasser S, Mausz MA, Schatz D, *et al.* (2014) Rewiring Host Lipid Metabolism by Large Viruses Determines the Fate of *Emiliania huxleyi*, a Bloom-Forming Alga in the Ocean. *The Plant Cell Online* **26**: 2689-2707.

Rowe JM, Fabre M-F, Gobena D, Wilson WH & Wilhelm SW (2011) Application of the major capsid protein as a marker of the phylogenetic diversity of *Emiliania huxleyi* viruses. *FEMS Microbiology Ecology* **76**: 373-380.

Rowe JM, Dunlap JR, Gobler CJ, Anderson OR, Gastrich MD & Wilhelm SW (2008) Isolation of a non-phage-like lytic virus infecting *Aureococcus anophagefferens*. *Journal of Phycology* **44**: 71-76.

Rowe JM, Jeanniard A, Gurnon JR, Xia Y, Dunigan DD, Van Etten JL & Blanc G (2014) Global analysis of *Chlorella variabilis* NC64A mRNA profiles during the early Phase of *Paramecium bursaria* Chlorella virus-1 infection. *PLoS ONE* **9**: e90988.

Ruardij P, Veldhuis MJW & Brussaard CPD (2005) Modeling the bloom dynamics of the polymorphic phytoplankter Phaeocystis globosa: impact of grazers and viruses. *Harmful Algae* **4**: 941-963.

Santini S, Jeudy S, Bartoli J, *et al.* (2013) Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci USA*. **110:**10800-10805.

Schroeder CD, Oke J, Malin G & Wilson HW (2002) Coccolithovirus (Phycodnaviridae): Characterisation of a new large dsDNA algal virus that infects *Emiliana huxleyi*. *Arch Virol* **147**: 1685-1698.

Schroeder DC, Oke J, Hall M, Malin G & Wilson WH (2003) Virus succession observed during an *Emiliania huxleyi* bloom. *Appl Environ Microbiol* **69**: 2484-2490.

Sharoni S, Trainic M, Schatz D, Lehahn Y, Flores MJ, Bidle KD, Ben-Dor S, Rudich Y, Koren I & Vardi A (2015) Infection of phytoplankton by aerosolized marine viruses. *Proc Natl Acad Sci USA* **112**: 6643-6647.

Sheyn U, Rosenwasser S, Ben-Dor S, Porat Z & Vardi A (2016) Modulation of host ROS metabolism is essential for viral infection of a bloom-forming coccolithophore in the ocean. *The ISME journal* **10**: 1742-1754.

Short SM (2012) The ecology of viruses that infect eukaryotic algae. *Environ Microbiol* **14**: 2253-2271.

Short SM & Suttle CA (2002) Sequence Analysis of Marine Virus Communities Reveals that Groups of Related Algal Viruses Are Widely Distributed in Nature. *Applied and Environmental Microbiology* **68**: 1290-1296.

Short SM & Short CM (2008) Diversity of algal viruses in various North American freshwater environments. *Aquatic Microbial Ecology* **51**: 13-21.

Sieburth JM, Johnson PW & Hargraves PE (1988) Ultrastructure and ecology of *Aureococcus anophageferens* gen. et sp. nov. (Chrysophyceae): the dominant picoplankter during a bloom in Narragansett Bay, Rhode Island, Summer 1985. *Journal of Phycology* **24**: 416-425.

Steven MS & Curtis AS (2003) Temporal dynamics of natural communities of marine algal viruses and eukaryotes. *Aquatic Microbial Ecology* **32**: 107-119.

Stevens K, Weynberg K, Bellas C, Brown S, Brownlee C, Brown MT & Schroeder DC (2014) A Novel evolutionary strategy revealed in the Phaeoviruses. *PLoS ONE* **9**: e86040.

Sun C, Feschotte C, Wu Z & Mueller RL (2015) DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. *BMC Biology* **13**: 1-12.

Suttle CA (2007) Marine viruses - major players in the global ecosystem. *Nat Rev Micro* **5**: 801-812.

Takemura M (2001) Poxviruses and the origin of the eukaryotic nucleus. *J Mol Evol* **52**: 419-425.

Taylor BP, Cortez MH & Weitz JS (2014) The virus of my virus is my friend: Ecological effects of virophage with alternative modes of coinfection. *Journal of Theoretical Biology* **354**: 124-136.

Van Etten JL & Dunigan DD (2012) Chloroviruses: not your everyday plant virus. *Trends in Plant Science* **17**: 1-8.

Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M & Stough JMA (2016) Standing on the Shoulders of Giant Viruses: Five Lessons Learned about Large Viruses Infecting Small Eukaryotes and the Opportunities They Create. *PLoS Pathog* **12**: e1005752.

Wilhelm SW & Suttle CA (1999) Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**: 781-788.

Williams TA, Embley TM & Heinz E (2011) Informational Gene Phylogenies Do Not Support a Fourth Domain of Life for Nucleocytoplasmic Large DNA Viruses. *PLoS ONE* **6**: e21080.

Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW, Brami D, Thiagarajan M, McCrow JP, Tovchigrechko A, Yooseph S & Venter JC (2012) Metagenomic Exploration of Viruses throughout the Indian Ocean. *PLoS ONE* **7**: e42047.

Wilson WH, Gilg IC, Duarte A & Ogata H (2014) Development of DNA mismatch repair gene, MutS, as a diagnostic marker for detection and phylogenetic analysis of algal Megaviruses. *Virology* **466–467**: 123-128.

Wilson WH, Tarran GA, Schroeder D, Cox M, Oke J & Malin G (2002) Isolation of viruses responsible for the demise of an Emiliania huxleyi bloom in the English Channel. *Journal of the Marine Biological Association of the United Kingdom* **82**: 369-377.

Yau S, Lauro FM, DeMaere MZ, *et al.* (2011) Virophage control of antarctic algal host–virus dynamics. *Proc Natl Acad Sci USA* **108**: 6163-6168.

Yolken RH, Jones-Brando L, Dunigan DD, *et al.* (2014) Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc Natl Acad Sci USA*. **111:**16106-16111.

Yutin N & Koonin EV (2012) Proteorhodopsin genes in giant viruses. *Biology direct* **7**: 1-6.

Yutin N, Wolf YI & Koonin EV (2014) Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466–467**: 38-52.

Zhang Q-C, Qiu L-M, Yu R-C, Kong F-Z, Wang Y-F, Yan T, Gobler CJ & Zhou M-J (2012) Emergence of brown tides caused by *Aureococcus anophagefferens* Hargraves et Sieburth in China. *Harmful Algae* **19**: 117-124.

Zhou J, Zhang W, Yan S, Xiao J, Zhang Y, Li B, Pan Y & Wang Y (2013) Diversity of Virophages in Metagenomic Data Sets. *Journal of virology* **87**: 4225-4236.

CHAPTER II

Genome of Brown Tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution

Publication note

This chapter was previously published as a peer-reviewed article in *Virology*, authored by Mohammad Moniruzzaman, Gary R. LeCleir, Christopher M. Brown, Christopher J. Gobler, Kay D. Bidle, William H. Wilson and Steven W. Wilhelm.

My contribution to this work was lab experiments necessary for genome finishing, bioinformatics and phylogenetic analyses and writing the majority of the manuscript.

Citation info

"Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH & Wilhelm SW (2014) Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology*. **466-467:** 60-70."

Abstract

Aureococcus anophagefferens causes economically and ecologically destructive "brown tides" in the United States, China and South Africa. Here we report the 370,920 bp genomic sequence of AaV, a virus capable of infecting and lysing *A. anophagefferens*. AaV is a member of the nucleocytoplasmic large DNA virus (NCLDV) group, harboring 377 putative coding sequences and 8 tRNAs. Despite being an algal virus, AaV shows no phylogenetic affinity to the *Phycodnaviridae* family, to which most algae-infecting viruses belong. Core gene phylogenies, shared gene content and genome-wide similarities suggest AaV is the smallest member of the emerging clade "Megaviridae". The genomic architecture of AaV demonstrates that the ancestral virus had an even smaller genome, which expanded through gene duplication and assimilation of genes from diverse sources including the host itself – some of which probably modulate important host processes. AaV also harbors a number of genes exclusive to phycodnaviruses – reinforcing the hypothesis that *Phycodna-* and *Mimiviridae* share a common ancestor.

Introduction

The presence of viruses with large genomes in marine systems has stimulated questions concerning the origin and phylogenetic histories of these unique particles. Such viruses have been found to infect diverse hosts, including eukaryotic phytoplankton (Fitzgerald et al., 2007; Moreau et al., 2010; Santini et al., 2013; Wilson et al., 2005), non-photosynthetic protists like *Acanthamoeba* (includes Mimivirus (Raoult et al., 2004), Marseillevirus (Boyer et al., 2009) and most recently Pithovirus (Legendre et al., 2014)) and the zooplankton *Cafeteria roenbergensis* (Fischer et al., 2010). Genomic characterizations of these viruses have dramatically altered our perceptions of the breadth of functional potential for virus particles. Each of these viruses is characterized by a large genome size, spanning several hundred to thousands of kilobases, and an exceptionally diverse gene content that is atypical of most other viruses. Analyses of the genome architecture of these viral "leviathans" have revealed patterns of massive gene duplication (Suhre, 2005) and gene acquisition from diverse sources, including their putative hosts (Filee et al., 2008). With the discovery of large-genome viruses, established phylogenetic classifications of NCLDVs infecting unicellular eukaryotes NCLDVs have been challenged, with proposals to reclassify NCLDVs having the largest genomes in the 'Megaviridae' clade irrespective of their host range (Santini et al., 2013). Emerging information on the genomic sequence and architecture of these viruses has the potential to redefine our understanding of virus function and evolution, including kindling debate that these viruses may represent a new, fourth domain of life (Boyer et al., 2010; Williams et al., 2011).

Aureococcus anophagefferens, a unicellular microalga, is a pelagophyte which causes recurrent brown tide blooms in the coastal and estuarine waters of the eastern United States, South Africa (Gobler and Sunda, 2012) and China (Zhang et al., 2012). The economic and ecological effects of brown tides are significant; blooms cause severe light attenuation in the coastal waters, resulting in destruction of sea grass beds (Gobler and Sunda, 2012), an important nursery for marine life. Brown tides are also toxic to bivalves and have contributed towards the collapse of multiple shellfisheries (Gobler and Sunda, 2012). Early transmission electron micrographs of *A. anophagefferens* alluded to the importance of viruses in the ecology of this organism, as they revealed virus-like particles in natural populations (Sieburth et al., 1988). Subsequent studies demonstrated that viruses likely played an important role in the modulation of bloom events: Gastrich et al (Gastrich et al., 2004) found that up to 37.5% of the population of *A. anophagefferens* may be visibly infected during bloom peak (Gastrich et al., 2004), suggesting this virus may be present at total abundances as high as ~10^{20} particles in Great South Bay (NY).

Initial attempts at isolation of *A. anophagefferens*-specific viruses reported that phage-like tailed particles were present as lytic agents of cells in cultures (Garry et al., 1998; Milligan and Cosper, 1994). However, subsequent research identified and isolated a large, icosahedral virus with a diameter of ~140

nm that was morphologically consistent with the earlier observations from blooms of *A. anophagefferens* (Gastrich et al., 2002; Rowe et al., 2008) . Several aspects of viral infection dynamics of *A. anophagefferens* have already been investigated (Gastrich et al., 2004; Gobler et al., 2007), but a crucial step in understanding the molecular mechanisms of host-virus interactions and indeed the ecology of giant viruses is to decipher genomic information. The 56 Mbp genome of host, *A. anophagefferens*, has recently been described (Gobler et al., 2011). We now report the complete genome sequence of the Aureococcus anophagefferens virus (AaV) which we assembled from combined *Illumina*™ and *454*™ pyrosequencing data. We provide a comprehensive analysis of the gene content, genome architecture and phylogenetic position of AaV.

Materials and Methods

AaV production and purification

The original AaV virus was isolated in 2002 and has been maintained in culture since (see Rowe et al. 2008 for a description of the morphology and infectious potential of the particle). AaV was added at a ratio of 8-12 viruses/cell to a 650 mL culture of *Aureococcus anophagefferens* CCMP1984 at 18°C, 14:10 light/dark cycle. The culture was monitored for virus and cell numbers. During lysis and at each step of purification, 40µL of lysate was fixed with glutaraldehyde (0.5% final) for 15 min at 4°C, then stored at -80°C until staining and counting. Lysate was filtered through GF/F filters (Whatman) to remove debris and the majority of bacteria. Viral enrichment was performed using precipitation with PEG8000 (Lawrence and Steward, 2010). PEG8000 was added to filtered lysate (8 g per 100 mL of lysate) and completely dissolved by gentle mixing. The PEG/lysate solution was left overnight at 4°C, followed by centrifugation for 35 min at 4°C, 10,000xg. Supernatants were carefully decanted, leaving 3-4 mL of supernatant in each bottle, residual liquid used to thoroughly rinse the bottles and the contents pooled. Approximately 10 mL of this concentrated virus solution was further concentrated to a volume of 1.5mL using a 30kDa cutoff Centricon filter (Millipore). Viruses were purified on an Optiprep™ (Iodixanol) step gradient. Four steps, 25%, 30%, 35% and 40% were prepared by diluting the 60% Optiprep stock with MilliQ H_2O. From each concentration, 2.63 mL was bottom loaded in a 12 mL ultracentrifuge tube, the lightest added first and the heaviest last (Lawrence and Steward, 2010). The 1.5 mL sample was then loaded on the top of the gradient. An identical gradient was prepared as a balance. The gradient was centrifuged in an SW41 rotor (Beckman Coulter) for 14 hours and 45 minutes at 39,000 rpm. Starting from the top of the gradient and working down, 14 fractions of 0.6-1.0 mL were collected. The density of each fraction was determined. A 5 µl sample of each fraction was diluted into 995 µl 0.22 µm filtered media. Each diluted sample (40 µl) was fixed with 0.8 µl glutaraldehyde for 15 minutes and

viruses were enumerated by flow cytometry after staining with SYBR gold (Brussaard, 2004). Bacterial concentrations were determined simultaneously with virus counts. Fractions with the highest concentrations of virus and lowest concentrations of bacteria were pooled for extraction of the viral genomic DNA.

AaV DNA extraction

The extraction protocol closely followed that for the *Emiliania huxleyi* virus (Schroeder et al., 2002). Briefly, the sample was treated with 5mg/mL proteinase K in a lysis buffer consisting of 20 mM EDTA (pH 8.0) and 0.5% SDS at 65°C, heated for approximately 1 hour to break up the capsid. Ten percent of the sample volume of phenol was then added and the DNA was extracted with chloroform: isoamyl alcohol (CIA). Additional chloroform: isoamyl alcohol (24:1) and precipitation steps were inserted to reduce spectral interference from iodixanol, allowing DNA quantification by a NanoDrop™ 1000 (Thermo Scientific).

Genome assembly

The AaV genome was sequenced at an extremely high depth using Illumina™ technology and was further complemented by 454™ sequencing. 128,117,014 Illumina™ paired-end reads of 100 bp length and 115,372 454™ single reads (avg. length of 272bp) were used for assembling the genome of AaV. After removing the Illumina™ and 454™ specific adapters and trimming these reads based on quality scores (limit 0.04), a hybrid *de novo* assembly was performed on these reads in *CLC Genomics Workbench* 5.0 (www.clcbio.com) (paired distance range 120-400, K-mer size 64) resulting in 185,000 contigs with a 200 bp contig size cut-off. However, when a 2,000bp size cut-off was imposed, number of contigs was reduced to 136. The largest contig obtained was 121.7 Kbp in length. Preliminary homology search of this 2,000bp subset against a local NCBI nr database (Benson et al., 2005) using BLASTX algorithm (Gish and States, 1993) revealed that several of these contigs originated from host genome, mitochondria and chloroplast sequences, indicating contamination from these sources during DNA extraction step and were not studied further. To identify the contigs of viral origin, we performed tblastx analysis (E-value < $1e^{-5}$) of these contigs against all the NCLDV virus genomes available as of November 21, 2012. Seven large contigs (A: 121,756 bp , B: 70,494 bp, C: 59,347 bp, D: 32,251 bp, E: 31,234 bp, F: 28,705 bp and G: 25,370 bp) were found to be putatively of viral origin based on the observation that they had highest sequence similarity to large dsDNA viruses with best hits to different sequenced NCLDVs. Sequence statistics analysis revealed that six of these seven putative contigs had very similar GC content ranging from 27.7% (contig 'D') to 29.4% (contig 'F'). Among all the 136 contigs analyzed, these were

the only ones having GC contents below 30% whereas contig 'G' had a slightly higher GC content of 31.4%.

Assuming that gaps between these contigs of viral origin resulted from repetitive sequences, we performed an all-vs-all nucleotide BLAST of these seven contigs in search of shared sequence similarities at the ends. This analysis revealed that contig pair (G,F); (F,A); (A,C); (C,D); (D,E); (E,B); (E,C) and (B,F) indeed shared sequence similarity at their ends. We hypothesized that these contig pairs represent contiguous sequences, and designed forward and reverse primers for the corresponding pairs. The PCR products spanning the gaps were purified and cloned using TOPO® TA cloning® kit in One Shot® TOPO10 chemically competent *E. coli* cells (Invitrogen™, CA) and sequenced using Sanger method. When the gaps were large, a second round of primers were designed from the ends of the sequences obtained from first round of reactions. The generated sequences were used to close the gaps between these contigs to obtain a continuous sequence of 370,920bp, which represents the genome of AaV.

Contig 'G' (25,370bp) represents the 5` terminal region of the genome, which consists of 20 DUF285 domain-containing CDSs in tandem (interspersed by three other CDSs). This contig also has a slightly higher GC% (31.4%) compared to the other viral contigs. As discussed in the main text, this region is also highly repetitious. To verify that contig 'G' was not an artifact of misassembled reads, we designed seven different sets of primers spanning seven non-overlapping regions of Contig 'G'. Upon PCR amplification, each of these primer pairs generated amplicons of the expected size, which confirmed that Contig 'G' is not a product of misassembled reads due to its highly repetitious nature, but an authentic part of the AaV genome. As part of our ongoing genome analysis, several other genes (including the DNA polymerase, major capsid, putative ion channel and mechanosensitive channel) have already been verified by cloning and Sanger sequencing.

The genome sequence data of AaV has been deposited in the GenBank database (accession no. KJ645900).

Genome Annotation

Putative coding sequences (CDS) were predicted using Prodigal web server (Hyatt et al., 2010). We defined a CDS having a minimum length of 50 consecutive codons bordered by a start and a stop codon. Homologous genes for the determined CDSs were detected by carrying out BLASTp analysis (Altschul et al., 1997) of the CDSs against the NCBI non-redundant protein database (nr) (Benson et al., 2005) with an E-value cutoff of 1e-05 to avoid false positive matches. Protein domains were detected using NCBI Conserved Domain Database (CDD) (Marchler-Bauer et al., 2013), pfam (Punta et al., 2012)

and Interpro (Quevillon et al., 2005) servers. Functional annotation resulted from integrating BLASTp results with the results obtained from these databases. tRNAs were predicted in the tRNAscanSE (Schattner et al., 2005) server using the general tRNA model.

Phylogenetic analysis

Putative homologs of the query proteins were identified by separate BLASTp (Altschul et al., 1997) searches against the viruses, eukaryotes, bacteria and other taxonomic subgroups defined in the GenBank nr database (Benson et al., 2005). For each of the query proteins, a representative set of homologues were selected. When available, homologs from the host algae were included in the phylogenetic reconstructions. Multiple sequence alignments were performed in MEGA 5.0 software (Tamura et al., 2011) using MUSCLE algorithm (Edgar, 2004) followed by manual refinement. Evolutionary models having the highest likelihood for each set of alignments were determined using Prottest 3.2.1 (Darriba et al., 2011). Maximum Likelihood phylogenetic reconstruction was performed for each set of alignments in TREEFINDER (Jobb et al., 2004). The Expected-Likelihood Weights (ELW) of 1,000 local rearrangements were used as confidence values for the nodes.

A number of CDSs had no homologs outside a particular domain of life or outside NCLDVs (NCLDV specific hypothetical CDSs). Phylogenetic origins of these CDSs were assigned directly to the respective domains of life they had best match to (E-value < 1e-05).

Other analyses

BLASTn comparison of the whole genomes was performed and illustrated using Blast Ring Image Generator (BRIG) (Alikhan et al., 2011) with an E-value cutoff of 1e-05. For Nucleocytoplasmic Virus Orthologous Group (NCVOG) analysis, a BLASTp (Altschul et al., 1997) search of AaV CDSs was carried out against a database containing all NCLDV proteins belonging to different NCVOG categories as constructed by Yutin et al (Yutin et al., 2009). (Downloaded from ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG). Hits with E-values lower than 1e-05 were assigned to their representative NCVOGs. Conserved motifs upstream of the PARCELs were analyzed using the MEME suite (Bailey and Elkan, 1994). Whole genome dot plot was constructed in YASS (Yet Another Similarity Searcher) server (Noé and Kucherov, 2005).

Results and Discussion

General genome features

Aureococcus anophagefferens Virus (AaV) has a linear double-stranded DNA genome with a size of 370,920 bp (Appendix Fig. 2.5). Applying a conservative annotation process, we identified 377 putative coding sequences in the genome of AaV, with a coding density of 88.3%. Such a high coding density is typical of large dsDNA viruses. The genome is A+T rich with a G+C content of 28.7%, in stark contrast to the host, which has a very high G+C content (69.5%) (Gobler et al., 2011). AT richness of the genome is also reflected in the codon usage of AaV; ~25% of the codons contain only A or T, whereas ~41% of the codons contain at least two A or T (excluding the stop codons). Putative coding sequence (CDS) length ranges from 52 to 2076 (with an average of ~290) amino acids.

Among the putative coding regions, 53% (200 of the 377) have significant (E-value < 1e-05) sequence similarity to proteins in the NCBI *nr* database, with 67 best matches to nucleocytoplasmic large DNA viruses (NCLDVs), 56 with eukaryotes, 72 with bacteria/bacteriophage and 5 best matches to archaea (Fig. 2.1a). Twenty eight of the CDSs having best match to NCLDVs have no homologs in the three domains of life (NCLDV specific hypothetical CDSs). Six AaV CDSs were most similar to sequences from the host (in terms of best Blastp hit) *Aureococcus anophagefferens*, indicating the possibility of horizontal gene transfer (HGT) between the host and the virus (Appendix Table 2.2). Among the 177 CDSs having no matches in the NCBI nr database, 34 had significant hits to NCBI environmental database (env_nr). Eighty seven (23%) of the AaV CDSs could be assigned within the Cluster of Orthologous Group categories (COG) (Appendix Fig. 2.5, Appendix Fig. 2.3), providing insight into their potential function.

A gene complement typical of NCLDVs

AaV harbors 9 of the 10 genes that are present in most members of all the NCLDV lineages defined by Yutin et al (Yutin et al., 2009) (Appendix Table 2.1), suggesting that AaV belongs to this group. Most NCLDVs characterized to date harbor genes involved in the basic biochemistry of "living" (*i.e.*, cellular) organisms: replication, transcription and translation. At least 18 genes within AaV can be categorized in the NCVOG category 'Transcription and RNA processing' (Appendix Fig. 2.5, Appendix Fig. 2.6). AaV has two copies of RNA polymerase II large subunit (Rpb1) (AaV_242, AaV_320), possibly as a result of paralogous expansion (Appendix Table 2.3), and two non-paralogous copies of RNA polymerase II small subunits (Rpb2) (AaV_222, AaV_370). Such a phenomenon has also been reported in *Phaeocystis globosa* virus 16T (Santini et al.,

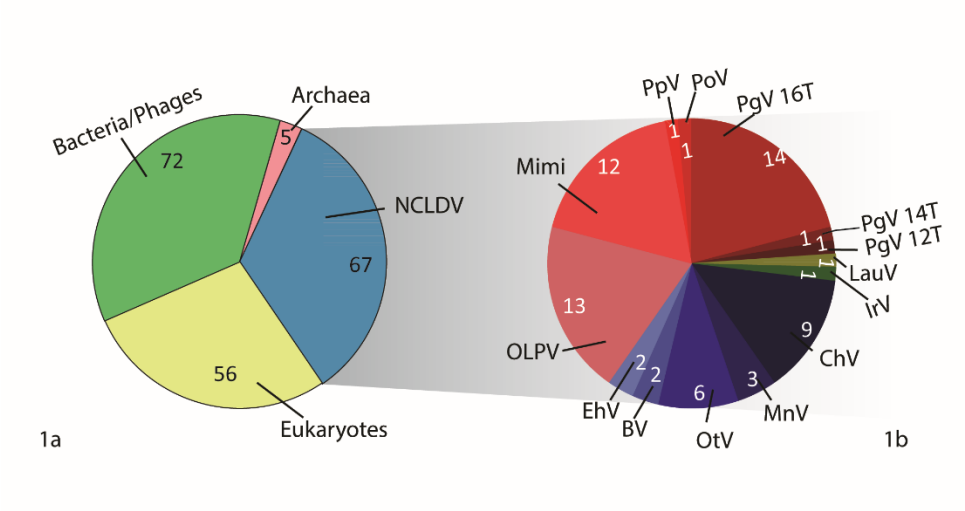


Figure 2.1: Best BLASTp hits of AaV proteome against a) NCLDVs and three domains of life. b) Different NCLDV viruses. IrV- Iridoviridae, PpV- *Phaeocystis pouchetii* virus, PoV- *Pyramimonas orientalis* virus, PgV16T- *Phaeocystis globosa virus* 16T, PgV 14T- *P. globosa virus* 14T, PgV 12T- *P. globosa virus* 12T, ChV- *Chlorella* viruses, MnV- *Micromonas* viruses, OtV- *Ostreococcus tauri* viruses, BV- *Bathycoccus* sp. virus, EhV- *Emiliania huxleyi* viruses, OLPV-Organic Lake Phycodnaviruses, Mimi-Mimiviridae, LauV- Lausannevirus. Hits to the Megaviridae clade are presented as shades of red whereas phycodnavirus hits are in shades of blue. Irido- and Lausannevirus hits are in shades of green.

2013), which shows gene duplication of Rpb2 subunits. AaV also harbors six other eukaryotic RNA polymerase subunits. Proteins involved in transcription initiation and elongation, namely Transcription initiation factor TFIIB (AaV_203), TATA-box binding protein (AaV_117) and a transcription elongation factor TFIIS (AaV_381), are present in the AaV genome. The acetylation state of histones in the chromatin is an important regulator of transcription in eukaryotic organisms (Marmorstein and Roth, 2001). ELP3-histone acetyl transferase genes are present in the *Cafeteria roenbergensis* virus (Fischer et al., 2010) and the *Phaeocystis globosa* virus genome. AaV also harbors this gene (AaV_368), and phylogenetic analysis suggests a eukaryotic origin for this gene (Appendix Fig. 2.7a). The maintenance of this gene in viruses infecting hosts with diverse lifestyles indicate that it was possibly present in the common ancestor of these viruses and may have an important function in modulating the transcriptional state of the host or virus maturation.

Although members of the NCLDV family are mostly independent of the host for replication and transcription, they typically depend on host protein synthesis machinery (Koonin and Yutin, 2010). However, genes coding for tRNAs and proteins involved in translation have been found in large viruses (Raoult et al., 2004; Wilson et al., 2005). Consistent with this, we identified eight tRNA genes in AaV (Appendix Table 2.4). Among these, five are present as a cluster starting at position 322,252 bp. Elongation factor 5A (AaV_110) and eIF 1α (AaV_118), genes involved in translation elongation and initiation, respectively, also occur in AaV. Translation elongation factor 5A is unique to AaV among large viruses (Appendix Fig. 2.7b).

While elevated photosynthetically active radiation has been shown to accelerate the virus-mediated lysis of the host (Gobler et al., 2007), higher intensity UV radiation can introduce DNA damage, including pyrimidine dimers (Weinbauer et al., 1997). Photolyase genes have been found in a number of large viruses (e.g., Fischer et al., 2010) and their role in repairing pyrimidine dimers in a NCLDV has been demonstrated (Srinivasan et al., 2001). A class II photolyase (AaV_082) in AaV likely plays a similar role. A lambda-type exonuclease (AaV_159), a Holliday junction resolvase (AaV_201) and a dUTPase gene (AaV_318) are all present in AaV and relevant given the high A+T content in its genome. A MutS7 gene, putatively involved in mismatch repair, has been found in all members of the *Mimiviridae* family (Ogata et al., 2011), and is also present in AaV. The MutT gene (nudix hydrolase) is involved in preventing the mis-incorporation of dGMP and thus transversion mutations (Akiyama et al., 1989) and there are two copies in the AaV genome (AaV_234, AaV_173).

Enzymes involved in ubiquitination are found in all NCLDV lineages . AaV encodes five E3 ubiquitin ligases, one E2 ubiquitin ligase, one POZ domain protein (part of the SCF-E3 complex) and also

a Ulp1 family thiol protease, a deubiquitination protein.This arsenal of proteins likely contributes toward its ability to overcome the host's defense against viral infection by interfering with Ub signaling (Iyer et al., 2006).

Unique CDSs derived from the host and other sources

It has been established that host-derived genes in cyanophage and large DNA viruses play key roles in resource acquisition by infected hosts and facilitate viral synthesis and host lysis (Hill, 2006; Monier et al., 2009; Wilson et al., 2005). For a number of AaV CDSs, no homologs could be identified in other NCLDVs (Appendix Table 2.5). While Blastp identified only six genes with best matches to host proteins, our phylogenetic reconstructions imply that at least 13 genes were possibly acquired from the host. (Appendix Table 2.5, Appendix Fig. 2.7). Among these, glucuronyl hydrolase (AaV_078) (Appendix Fig. 2.7c) and pectate lyases (AaV_003, 375, 038) (Appendix Fig. 2.7d) are both present in the host alga (Gobler et al., 2011). Brown tides are associated with severe light attenuation, which minimizes light available for photosynthesis (Gobler et al., 2011; Gobler and Sunda, 2012). As such, glucuronyl hydrolase and pectate lyase genes in AaV may permit *A. anophagefferens* to derive energy from abundant sources of organic carbon during blooms, enhancing the ability of the infected cell to generate energy when light levels are reduced and as the chloroplast is degraded in the late stages of infection (Gastrich et al., 1998). Another putative host-derived gene is an intramembrane rhomboid family serine protease (AaV_077) (Appendix Fig. 2.7e). Given that *A. anophagefferens* can use complex sources of organic nitrogen, this protease may enhance the ability of infected *A. anophagefferens* to access nitrogen (Gobler et al., 2011; Gobler and Sunda, 2012). Collectively, the presence of genes associated with the degradation of organic compounds by *A. anophagefferens* in AaV suggests they were obtained from the host and retained due to enhanced fitness they provide.

AaV harbors a calpain family thiol protease (AaV_045) not reported in any other NCLDVs. (Appendix Fig. 2.7f). Calpains have diverse functions, including cell cycle regulation and caspase-independent apoptosis (Nemova et al., 2010). Modulation of host cell apoptosis by NCLDVs, especially among the *Phycodnaviridae* (reviewed in Bidle and Vardi, 2011), is an emerging feature of host-virus interactions and appears to be a shared trait amongst diverse algal viruses (Bidle and Vardi, 2011). Studies have demonstrated a pivotal role for viral induction of the host programmed cell death machinery in the propagation of *Emiliania huxleyi* viruses (EhV) (Bidle et al., 2007; Vardi et al., 2009), with this trait being conserved among a mixed population of both host and EhV genotypes (Vardi et al., 2012). In other systems (e.g., Hepatitis C) activation of calpains have been shown to inhibit the host's extrinsic apoptotic signaling pathway, which is necessary for successful infection. (Simonin et al., 2009).

Consequently, the possession of such a protease in the AaV genome might allow for avoidance of the host's virus exclusion strategy.

Among NCLDVs, ion channel proteins have only been reported in the *Phycodnaviridae*. Potassium channel proteins in *Chlorella* viruses (Thiel et al., 2011) are well-studied and critical in the infection process (Greiner et al., 2009, Neupartl et al., 2007). We located a putative potassium ion channel protein (AaV_153) with a length of 157aa as well as a putative small conductance mechanosensitive ion channel protein (AaV_043). Small conductance mechanosensitive channels are implicated in counteracting the osmotic pressure inside the cells (Wilson et al., 2013). The presence of two ion channel proteins, which operate under two distinct stimuli (ionic and mechanical) in AaV raise the possibility that they might have important (and different) roles during different stages of infection.

AaV has two prenyl transferases, commonly involved in lipid metabolism. One of these, undecaprenyl pyrophosphate synthase, was putatively derived from the host (AaV_255) (Appendix Fig. 2.7g). The second protein, UbiA prenyltransferase (AaV_373) (Appendix Fig. 2.7h), is a key enzyme in the biosynthesis of Ubiquinone (Szkopinska, 2000), a critical molecule in the respiratory electron transport chain. The presence of UbiA along with an AIM24 domain protein (AaV_144) (Appendix Fig. 2.7i), that has a role in mitochondrial biogenesis (Hess et al., 2009), imparts the potential for AaV to further modulate the host's energy generating processes.

Carbohydrates can mediate interactions in diverse virus-host systems. The genome of AaV harbors a carbohydrate sulfotransferase (AaV_102) (Appendix Fig. 2.7j), a gene involved in producing sulfated carbohydrates. Heparan sulfate, a sulfated carbohydrate, is known to be a surface receptor for a number of viruses including Vaccinia and Herpes Simplex (Zhu et al., 2011). In the case of *Emiliania huxleyi*, EhV-86 encodes C-type lectin-containing protein that associates with purified lipid rafts from 2 hours post-infected host cells, arguing that EhV infection occurs at the interface between virus proteins and host lipid-raft sugar-lipid moieties (Rose et al., 2014). No studies have been conducted on the molecular mechanisms of AaV-host interactions yet, so whether sulfated carbohydrates have any role in such interaction remains an open question. The *Aureococcus* genome is highly enriched in sulfatase genes that encode proteins that degrade sulfonated polysaccharides (29) which may assist in discouraging the attachment of AaV to its cell surface.

Finally, AaV encodes phaeophorbide *a* oxygenase (PaoA; AaV_372) (Appendix Fig. 2.7k), which is a key enzyme in chlorophyll catabolism (Pruzinska et al., 2003). PaoA was also present in Organic Lake Phycodnavirus 2 (Yau et al., 2011), a virus assembled using metagenomic data from a hypersaline lake in Antarctica. Occurrence of this gene in two viruses from distinct geographic locations

suggests that it was probably present in the common ancestor and might play role in modulating the host cellular processes. However, the possibility of independent acquisition of the gene through HGT cannot be discounted, either.

Putative role(s) for repetitive DNA elements in AaV

Repetitive DNA elements occur frequently in large DNA virus genomes. For example, *Cafeteria roenbergensis* virus (CroV) and Mimivirus have FNIP repeats (Pfam: PF05725) (Fischer et al., 2010; Raoult et al., 2004) while three distinct families of repeats with no homology within available databases were reported in EhV-86 (Allen et al., 2006). Approximately 11.3% of the genome of AaV is comprised of a lysine-enriched domain of unknown function (DUF285, Pfam: 03382), which is distantly related to leucine rich repeats. DUF285 domain regions are sequestered in putative coding sequences, resulting in a large paralogous family of 50 ORFs (Appendix Table 2.3, Appendix. Fig. 2.5), which is 13.25% of the total gene content. These CDSs range from 101 amino acids to 708 amino acids in length and all of them contain $\geq$ 1 copy of either partial or complete DUF285 domain defined in the Pfam database. Phylogenetic analysis reveals that this sequence probably originated in a bacterium (Appendix Fig. 2.7l). Repeats characterized by DUF285 domains occur sporadically in unicellular microbes, especially in the obligate endosymbiotic class Mollicutes, and also in unicellular photosynthetic eukaryotes (Roske et al., 2010). ORFs containing these domains were termed PARCELs (Palindromic Amphipathic Repeat Coding Elements), characterized by repeating elements displaying dyad symmetry and variable hydrophilic and conserved hydrophobic regions (Roske et al., 2010). These ORFs are also found as part of some bacterial mobile elements and plasmids. It has been suggested that PARCELs have spread in diverse bacterial and eukaryotic lineages through HGT and intra-genomic shuffling (Roske et al., 2010). The sequence characteristics of PARCELs endow them with potential roles in gene expansion and recombination (Roske et al., 2010).

Twenty of these PARCELs are present as tandem repeats at the 5‘extremity of AaV genome (genome location **A**: 1,908 bp – 22,906bp, interrupted by three other CDSs), while another cluster of 9 genes are present near the other end of the genome (genome location **B**: 330,172bp – 339,020bp) (Fig. 2.2). Both clusters are on the positive strand while the rest of the PARCELs are evenly distributed on the negative strand (Appendix Fig. 2.8). Additionally, two distinct conserved domains (which we denote as Motif_A and Motif_B) have been found to be consistently present at the upstream regions of the positive and the negative strand PARCELs, respectively (Appendix Fig. 2.9). In the genome of host *A. anophagefferens* we have identified occurrences of the DUF285 motif at 90 distinct loci distributed across 32 scaffolds. The presence of PARCELs in both the host and virus genomes is intriguing. Since

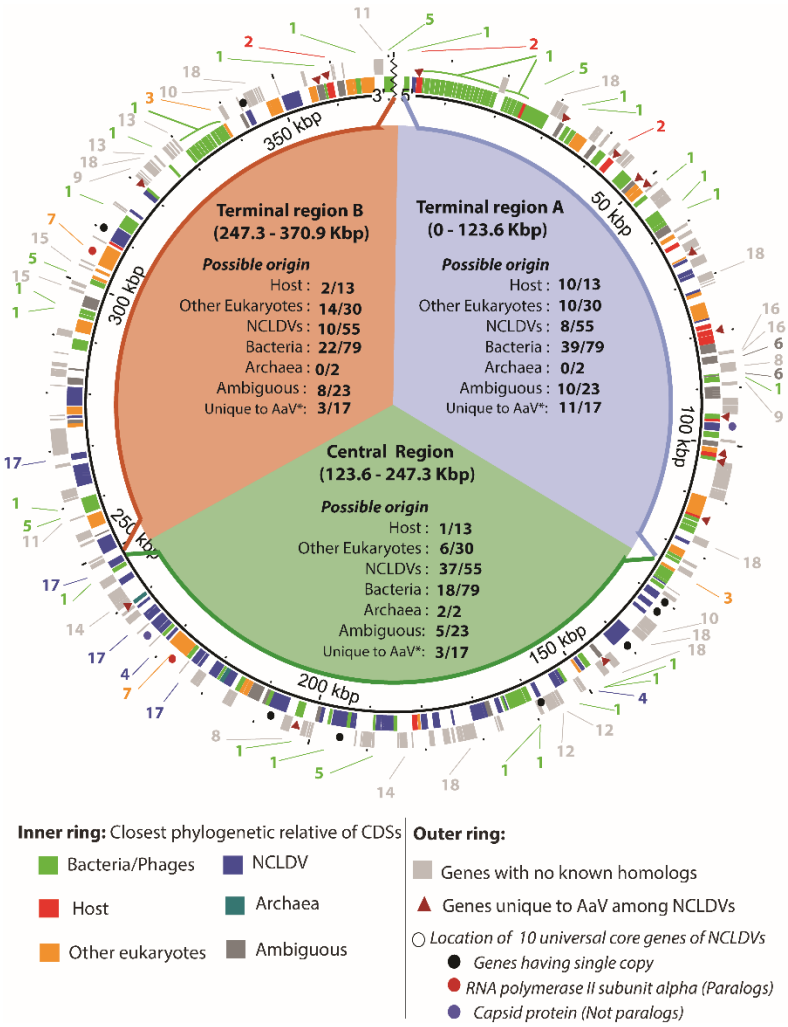


Figure 2.2: Locations of the genes based on possible phylogenetic origins. To demonstrate the biased distribution of genes derived from different sources, the genome was divided into three equal sections to represent the central core and two terminal regions. Each of the sections are 123.6 Kbp in length. The location of the paralogous genes are shown using the numbers that represent each of the paralog groups (Appendix Table 2.4). Genes with putative origin in host (red), other eukaryotes (orange), bacteria (green), archaea (teal) are shown. Genes for which the origin could not be inferred are depicted in gray (ambiguous origin). Locations of genes unique to AaV are marked with dark red triangles whereas universal NCLDV core genes are denoted by circles. The pie chart inside the genome map represents the three equal sections of the genome. Number of genes of different origins (Host, Other Eukaryotes, NCLDVs, Bacteria, Archaea and Ambiguous origin) located in each of the regions are presented inside the respective sections of the pie chart. For example, the central region harbors 32 of the 49 genes putatively derived from an NCLDV ancestor.* *Genes 'unique to AaV' refer to the genes that are only found in AaV, to date, among the NCLDVs.*

phylogenetic analysis suggests the possible origin of host PARCELs in bacteria (Appendix Fig. 2.7l), the most parsimonious scenario is the mobilization of this sequence to AaV from the host (upon uptake from bacteria) and its subsequent intra-genomic duplication. Alternatively, it is possible that this sequence is in flux between the host and the virus, playing a role in host-virus coevolution. Lineage-specific gene expansion contributed to the genome growth of other NCLDVs substantially (Iyer et al., 2006), and a similar mechanism is probably in effect in the genome of AaV. Duplicated genes can also go through neo-functionalization, a process where the daughter copy assumes a new function distinct from the mother gene (Liu et al., 2011). In accordance with this mechanism, we have found one of the PARCELs (AaV_220) containing a U-box domain fused to it, which probably functions as an E3 ubiquitin ligase (Ohi et al., 2003). The presence of conserved motifs at the upstream of the PARCELs and their orientation in antiparallel directions (Appendix Fig. 2.9) inside the genome suggests that they are subject to intra-genomic mobilization. PARCELs at the extremities of the genome (AaV_001 & AaV_382) are present as inverted repeats (Appendix Fig. 2.8), which potentially mediate circularization of AaV genome as has been found in Mimivirus and some other NCLDVs (Raoult et al., 2004).

The phylogenetic position and evolutionary history of AaV

DNA polymerase gene-based phylogeny clustered AaV in the *Mimiviridae* family (Fig. 2.3). The only other algal virus that clusters in *Mimiviridae* family is *P. globosa* 16T. Two viruses from metagenomes generated in a hypersaline Antarctic lake (Organic Lake Phycodnavirus 1 and 2) (Yau et al., 2011), having no known hosts, also cluster in the *Mimiviridae*. Based on common marker genes and large genome sizes, it has been hypothesized that these viruses have a common ancestor, and a new group called 'Megaviridae' has been proposed that further extends the *Mimiviridae* family and is independent of hosts (Santini et al., 2013). Sixty-seven AaV proteins have their highest sequence similarity to large DNA viruses, among which, 65% are to Megaviridae family (Fig. 2.1b). Recruitment of whole genomes of NCLDVs to AaV demonstrate more coverage from the Megaviridae members relative to other NCLDV viruses (Fig. 2.4), supporting our suggestion that AaV is more similar to Megaviridae than to phycodnaviruses. Yutin et al. (Yutin et al., 2013) grouped the proteins of all the Megaviridae members, generating the 'Mimivirus cluster of orthologous groups (MimiCOGs)' (Yutin et al., 2013). Fifty two MimiCOG family genes are commonly shared among this group and, despite having the smallest genome, AaV possesses 46 of these 'core' genes (Appendix Table 2.1).

Figure 2.3: Maximum likelihood phylogenetic trees of (A) B family DNA polymerase (AaV_141), (B) Major capsid protein (AaV_096) and (C) A32-like virion packaging ATPase (AaV_165) with other NCLDV members. The capsid homologs from the Poxviridae family are highly divergent and were not included in the major capsid protein phylogenetic analysis. The trees were constructed from amino acid alignments of the respective proteins. The Expected-Likelihood Weights (ELW) of 1,000 local rearrangements were used as confidence values for the nodes. The abbreviations are as follows: AaV (*Aureococcus anophagefferens* virus), CroV (*Cafeteria roenbergensis* virus), OLPV 1 (Organic Lake Phycodnavirus 1), CeV (*Chrysochromulina ericina* virus), PGV 16T (*Phaeocystis globosa* virus 16T), APMV (*Acanthamoeba polyphaga* mimivirus), *M. chilensis (Megavirus chilensis), M. monve (Moumouvirus monve),* McV (*Mollascum contagiosum* virus), CpV (Canarypox virus), SpV (Swinepox virus), LsdV (Lumpy skin disease virus), YmtV (Yaba monkey tumor virus), AmeV 'L' (*Amsacta moorei* entomopoxvirus L), ASFV (African swine fever virus), ESV 1 (*Ectocarpus siliculosus* virus 1), FsV (*Feldmannia* species virus), ATCV 1 (*Acanthocystis turfacea* Chlorella virus 1), OsV5 (*Ostreococcus* virus OsV5), *P. salinus (Pandoravirus salinus),* EhV-86 (*Emiliania huxleyi* virus 86), Hav 01 (*Heterosigma akashiwo* virus 01), LdV 1 (Lymphocystis disease virus 1), IsknV (Infectious spleen and kidney necrosis virus), Frv 3 (Frog virus 3), SgIV (Singapore grouper iridovirus), IRV 6 (Invertebrate iridescent virus 6), HvaV 3e (*Heliothis virescens* ascovirus 3e), *P. sibericum (Pithovirus sibericum),* LauV (Lausannevirus), C8V (Cannes 8 virus).

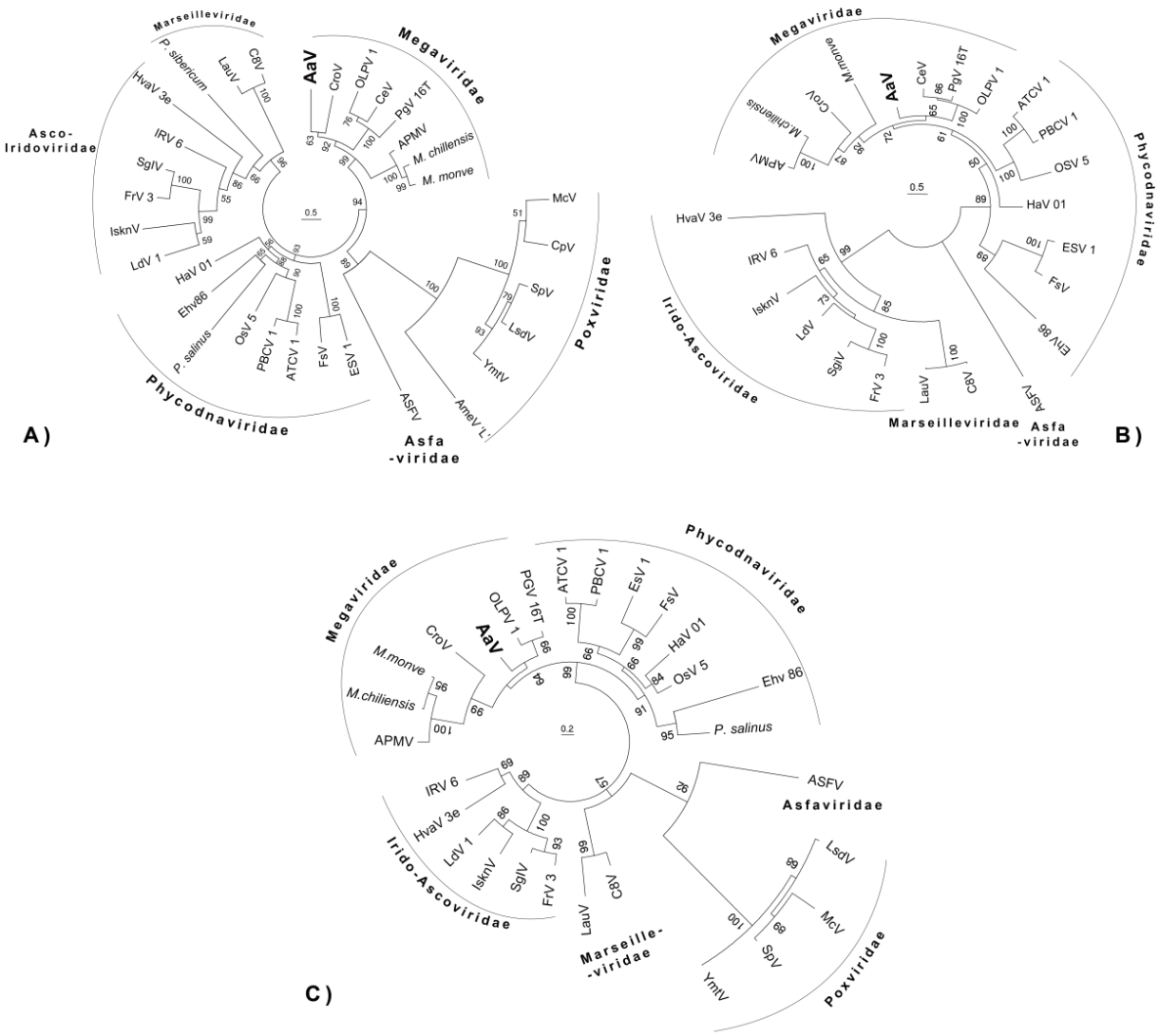


Figure 2.3. Continued.

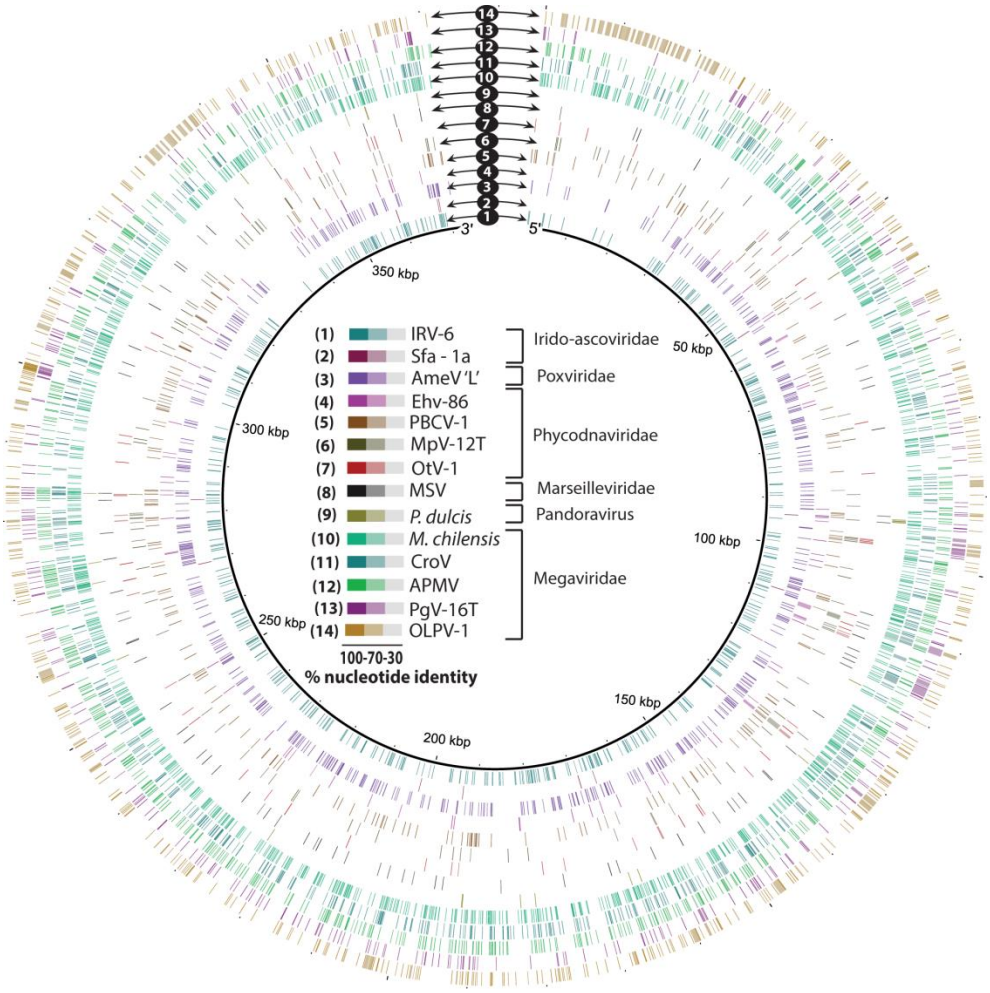


Figure 2.4: BLASTn hits of whole genomes of NCLDV members recruited against the genome of AaV. The pairwise alignments (E-value < 1-e05) that each of the compared genomes produced with that of AaV were mapped on the AaV genome. The density of colors on each ring represents the identity percentage of nucleotides shared by a particular virus with AaV. This figure illustrates the fact that AaV genome shares more regions of sequence similarity to the Megaviridae family members than other NCLDV families.

Apart from their large genome size, a key feature of the members of the Megaviridae group is the presence of both asparagine synthetase and MutS7 genes. Based on our analyses, it is evident that AaV belongs to this proposed clade; however, it lacks the asparagine synthetase gene, raising questions concerning the universality of this gene within the Megaviridae. Other traits consistent with other Megaviridae to date (the presence of a "virus factory" within infected cells as well as the presence of virophage) are also not evident in our observations to date: while these latter traits require deeper investigation (*e.g.,* TEM observations of infected cells in an effort to define the presence of virus factories), the data we have gathered to date highlights the difficulty of defining 'core' genes or traits in any system (Kislyuk et al., 2011).

Based on synapomorphies, it has been suggested that Phycodna- and Mimiviruses originated from a common ancestor (Iyer et al., 2006). Indeed, in contrast to our conclusion that AaV belongs within the Megaviridae, 22 of the AaV genes had highest sequence similarity with phycodnaviruses, and in a number of cases to genes exclusively present in phycodnaviruses. Among these are a phosphate starvation-induced protein (AaV_210) (Appendix Fig. 2.7m), a RNA polymerase sigma factor 70 (AaV_076, putatively host derived), two copies SCF ubiquitin ligase (AaV_357, AaV_123), a zinc finger domain protein (AaV_380) and two paralogs of laminin G domain-containing protein (AaV_024, AaV_386). A number of hypothetical genes exclusive to phycodnaviruses were also found in the AaV genome (Appendix Table 2.6). An intriguing finding was the presence of 4 paralogous copies of a phycodnavirus-specific hypothetical gene (Paralog group 17; Appendix Table 2.3) in the NCVOG cluster 1343, hereinafter denoted as 'AaV and phycodnavirus-specific highly similar genetic element' (AP_HGE). This gene is also found in several strains of *Paramecium bursaria* - *Chlorella* virus in multiple copies (Yutin et al., 2013). In AaV, these elements share very high sequence homology, with average pairwise similarity of 90.5% and 80% at nucleotide and amino acid level, respectively. Furthermore, 90% of the 143 nucleotides immediately upstream and 82% of the 50 nucleotide positions immediately downstream of these ORFs are also fully conserved (Appendix Fig. 2.10). Although AP_HGEs do not match known mobile genetic elements, the possibility that the conserved sequence signatures may contribute to the mobility of these ORFs cannot be ruled out.

Based on the similarity between AaV CDSs and proteins from other sources (Fig. 2.1a), we hypothesized that AaV has acquired genes from diverse sources. To reconstruct the evolutionary history of the AaV genes, we carried out a comprehensive maximum likelihood phylogenetic analyses for genes having homologs in diverse domains of life. According to the phylogenetic analyses, 78 genes possibly originated from bacteria, including the 50 genes in paralog group 1 (DUF285 domain containing proteins)

(Appendix Table 2.3). Fifty-five genes showed highest phylogenetic affinity to NCLDVs. Eight genes were found to have closest similarity to the corresponding host protein (Appendix Fig. 2.7b, c, e and g, Appendix Table 2.2), indicating a relatively recent horizontal gene transfer. Thirty one genes appear to be acquired from eukaryotes other than the host. Another five proteins clustered with bacteria and the host (as the only eukaryote), indicating a history of gene transfer among bacteria, *Aureococcus* and AaV. Finally two genes possibly originated from archaea as suggested by our phylogenetic analysis.

The Genomic architecture of AaV

An initial observation, that most genes having phylogenetic affinity to the host proteins are located at the terminal regions of the genome, prompted us to further investigate the spatial distribution of the CDSs in the AaV genome. We partitioned the genome map into three sections of near-equal length (123.6 Kbp starting from 5` end (Terminal region A), 123.6 Kbp – 247.3 Kbp (Central region) and 247.3 Kbp – 370 Kbp (Terminal region B); Fig. 2.2). Terminal region A contains 10 of the 13 genes having phylogenetic affinity to the host, with three in tandem (AaV_076, 077 & 078), suggesting the concurrent acquisition of these genes from the host. Twenty-five of the 31 genes having origin in other eukaryotes are also situated within the two terminal regions. Interestingly, of the 17 genes that are unique to AaV among the NCLDVs (Appendix Table 2.5), 14 are found in these two terminal sections. Genes unique to AaV are unlikely to have been vertically inherited from the ancestral virus and were possibly acquired through HGT. Sixty of the 78 genes putatively derived from bacteria are distributed in the terminal regions, most of which are paralogous copies of DUF285 domain-containing ORFs. One of the most interesting observations was the presence of seven universal NCLDV specific 'core' genes (Yutin et al., 2009) in the central region along with a capsid protein (AaV_247). This region also harbors 20 of the 28 NCLDV specific hypothetical CDSs that are found in AaV (Appendix Table 2.6). In total, the central region accommodates 37 of the 55 AaV genes putatively having origin in NCLDVs. Taken together, these observations suggest that the ancestral version of AaV was probably a much smaller NCLDV, and that this virus has expanded its genome by accruing genes at the terminal regions from diverse sources including the host itself.

It has been proposed that the Mimivirus genome evolved to its "giant" size by accumulating genes from host and other eukaryotic organisms at the terminal regions of the genome (Filee et al., 2008) and acquisition of bacterial genes by lineages of NCLDVs has also been demonstrated (Filee et al., 2007). Being a member of the Megaviridae clade, AaV shows a similar pattern of gene acquisition. In contrast to Mimivirus, however, AaV is unlikely to have intimate contact with bacteria inside the host, since *Aureococcus anophagefferens* is not known to be phagotrophic (Gobler and Sunda, 2012). *A.*

anophagefferens can degrade foreign organic matter to derive energy and thus likely comes into frequent contact with foreign nucleic acids (Gobler et al., 2011), although one would anticipate that long stretches (especially as intact ORFs) would almost always be disrupted. Whether free DNA from the environment is assimilated by the host and made available to the virus during replication is an open question. Should intact non-host genes be present within infected *A. anophagefferens*, mechanisms like strand invasion might contribute to the assimilation of these foreign genes during replication of AaV inside the host (Filee et al., 2007).

Conclusions

We have presented the complete genome sequence of AaV, a virus that infects a marine pelagophyte that causes harmful brown tides. AaV is a large dsDNA virus, yet the smallest described member of an emerging clade, the Megaviridae, which harbors nucleocytoplasmic large DNA viruses with diverse hosts including both non-photosynthetic protists and photosynthetic algae. Despite having a smaller genome size, AaV shares a large number of core genes with other members of this growing group, which points to a common evolutionary history of AaV and some of the largest and most complex viruses. This observation suggests genome size is not a definitive criterion for the proposed 'Megaviridae' family. Analysis of the genome architecture suggests that the ancestral virus of AaV was probably much smaller in terms of genome size, and likely followed an evolutionary pathway involving massive gene accumulation and gene duplication from the host as well as other organisms. The genome of AaV harbors putative functional and hypothetical genes from both *Phycodna*- and *Mimiviridae* clades and thereby enhances our understanding of the evolutionary history of these two diverged families which may have a common ancestor (Iyer et al., 2006). Moreover, AaV has several genes novel to the NCLDV group with possible roles in regulation of host cell processes. Several genes likely acquired from *A. anophagefferens* possibly allow AaV to facilitate the acquisition of resources by its host during viral infection.

Most of the Megaviridae family members isolated so far infect one single host, *Acanthamoeba polyphaga*, a non-photosynthetic phagotroph, although many researchers do not consider the amoeba to be the "native" host. AaV joins this group as one of two Megaviridae members having known photosynthetic hosts, and in the current case it brings a substantial ecological history (Gastrich et al., 2002; Gastrich et al., 2004; Gobler et al., 2007; Gobler and Sunda, 2012; Rowe et al., 2008). Algal viruses of the *Phycodnaviridae* family have been extensively studied regarding their seasonal dynamics (Martínez et al., 2007), diversity (reviewed in Short, 2012) and capability of modulating host cellular processes (Vardi et al., 2012) – however it is evident that algae-infecting NCLDVs exist across at least two distinct phylogenetic clades, and we anticipate new Megaviridae members with photosynthetic hosts

will be described in near future. Because of the socioeconomic and environmental impact of brown tides, both *A. anophagefferens* and AaV have been studied extensively from physiological and ecological perspectives. Now, with the availability of genome sequences for both AaV and its host (Gobler et al., 2011), it is possible to develop a biomolecular experimental model system for teasing apart not only the dynamics of Megaviridae family, as well as to begin to experimentally gain insight into the genomic and phylogenetic evolution of this group.

Acknowledgements

The authors thank a decade of students and collaborators for their continued support of work with the *Aureococcus anophagefferens* system. Isolation and purification efforts of AaVs and genomic DNA was supported by the National Science Foundation (OCE-1061883 to KDB) and an Institute of Marine and Coastal Sciences postdoctoral fellowship to CMB. Sequencing, assembly and annotation were also supported by the National Science Foundation (EF-0949120 to SWW and EF-0949162 to WHW).

List of References

Akiyama, M., Maki, H., Sekiguchi, M., Horiuchi, T., 1989. A specific role of MutT protein: to prevent dG.dA mispairing in DNA replication. Proc. Natl. Acad. Sci. U.S.A. 86, 3949-3952.

Alikhan, N.-F., Petty, N., Ben Zakour, N., Beatson, S., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12, 402.

Allen, M.J., Schroeder, D.C., Wilson, W.H., 2006. Preliminary characterisation of repeat families in the genome of EhV-86, a giant algal virus that infects the marine microalga *Emiliania huxleyi*. Arch. Virol. 151, 525-535.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28-36.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2005. GenBank. Nucleic Acids Res. 33, D34-38.

Bidle, K.D., Haramaty, L., Barcelos e Ramos, J., Falkowski, P., 2007. Viral activation and recruitment of metacaspases in the unicellular coccolithophore, *Emiliania huxleyi*. Proc. Natl. Acad. Sci. U.S.A. 104, 6049-6054.

Bidle, K.D., Vardi, A., 2011. A chemical arms race at sea mediates algal host–virus interactions. Curr. Opin. Microbiol. 14, 449-457.

Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B., Raoult, D., 2010. Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. PLoS ONE 5, e15530.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M.G., Suzan-Monti, M., La Scola, B., Koonin, E.V., Raoult, D., 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc. Natl. Acad. Sci. USA. 107(45), 19508-19513.

Brussaard, C.P., 2004. Optimization of procedures for counting viruses by flow cytometry. Appl. Environ. Microbiol. 70, 1506-1513.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 27(8), 1164-1165.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792-1797.

Filee, J., Pouget, N., Chandler, M., 2008. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. BMC Evol. Biol. 8, 320.

Filee, J., Siguier, P., Chandler, M., 2007. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet. 23, 10-15.

Fischer, M.G., Allen, M.J., Wilson, W.H., Suttle, C.A., 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc. Natl. Acad. Sci. USA. 107(45), 19508-19513.

Fitzgerald, L.A., Graves, M.V., Li, X., Hartigan, J., Pfitzner, A.J., Hoffart, E., Van Etten, J.L., 2007. Sequence and annotation of the 288-kb ATCV-1 virus that infects an endosymbiotic *Chlorella* strain of the heliozoon *Acanthocystis turfacea*. Virology 362, 350-361.

Garry, R.T., Hearing, P., Cosper, E.M., 1998. Characterization of a lytic virus infectious to the bloom-forming microalga *Aureococcus anophagefferens* (Pelagophyceae). J. Phycol. 34, 616-621.

Gastrich, M., Anderson, O.R., Cosper, E., 2002. Viral-like particles (VLPS) in the alga, *Aureococcus anophagefferens* (Pelagophyceae), during 1999–2000 Brown tide blooms in Little Egg Harbor, New Jersey. Estuaries 25, 938-943.

Gastrich, M., Leigh-Bell, J., Gobler, C., Roger Anderson, O., Wilhelm, S., Bryan, M., 2004. Viruses as potential regulators of regional brown tide blooms caused by the alga, *Aureococcus anophagefferens*. Estuaries 27, 112-119.

Gastrich, M.D., Anderson, O.R., Benmayor, S.S., Cosper, E.M., 1998. Ultrastructural analysis of viral infection in the brown-tide alga, *Aureococcus anophagefferens* (Pelagophyceae). Phycologia 37, 300-306.

Gish, W., States, D.J., 1993. Identification of protein coding regions by database similarity search. Nat. Genet. 3, 266-272.

Gobler, C.J., Anderson, O.R., Gastrich, M.D., Wilhelm, S.W., 2007. Ecological aspects of viral infection and lysis in the harmful brown tide alga *Aureococcus anophagefferens*. Aquat. Microb. Ecol. 47, 25-36.

Gobler, C.J., Berry, D.L., Dyhrman, S.T., Wilhelm, S.W., Salamov, A., Lobanov, A.V., Zhang, Y., Collier, J.L., Wurch, L.L., Kustka, A.B., Dill, B.D., Shah, M., VerBerkmoes, N.C., Kuo, A., Terry, A., Pangilinan, J., Lindquist, E.A., Lucas, S., Paulsen, I.T., Hattenrath-Lehmann, T.K., Talmage, S.C., Walker, E.A., Koch, F., Burson, A.M., Marcoval, M.A., Tang, Y.-Z., LeCleir, G.R., Coyne, K.J., Berg, G.M., Bertrand, E.M., Saito, M.A., Gladyshev, V.N., Grigoriev, I.V., 2011. Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. Proc. Natl. Acad. Sci. U.S.A. 08(11), 4352-7.

Gobler, C.J., Sunda, W.G., 2012. Ecosystem disruptive algal blooms of the brown tide species, *Aureococcus anophagefferens* and *Aureoumbra lagunensis*. Harmful Algae 14, 36-45.

Greiner, T., Frohns, F., Kang, M., Van Etten, J.L., Kasmann, A., Moroni, A., Hertel, B., Thiel, G., 2009. *Chlorella* viruses prevent multiple infections by depolarizing the host membrane. J. Gen. Virol. 90, 2033-2039.

Hess, D.C., Myers, C.L., Huttenhower, C., Hibbs, M.A., Hayes, A.P., Paw, J., Clore, J.J., Mendoza, R.M., Luis, B.S., Nislow, C., Giaever, G., Costanzo, M., Troyanskaya, O.G., Caudy, A.A., 2009. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. PLoS Genet. 5, e1000407.

Hill, E., 2006. The cyanophage molecular mixing bowl of photosynthesis genes. PLoS Biol. 4, e264.

Hyatt, D., Chen, G.-L., LoCascio, P., Land, M., Larimer, F., Hauser, L., 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.

Iyer, L.M., Balaji, S., Koonin, E.V., Aravind, L., 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 117, 156-184.

Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.

Kislyuk, A., Haegeman, B., Bergman, N., Weitz, J., 2011. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 12, 32.

Koonin, E.V., Yutin, N., 2010. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53, 284-292.

Lawrence, J.,E., Steward G.,F., 2010 Purification of viruses by centrifugation. Wilhelm, S.,W., Weinbauer, M.,G., Suttle, C.,A., (Eds.), Manual of Aquatic Viral Ecology (MAVE), Association for the Sciences of Limnology and Oceanography , Texas. pp 166-181.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., Claverie, J.-M., 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a Pandoravirus morphology. Proc. Natl. Acad. Sci. U.S.A. publication ahead of print 10.1073/pnas.1320670111.

Liu, S.-L., Baute, G.J., Adams, K.L., 2011. Organ and cell type–specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. Genome Biol. Evol. 3, 1419-1436.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Lu, S., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Bryant, S.H., 2013. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 41, D348-352.

Marmorstein, R., Roth, S.Y., 2001. Histone acetyltransferases: function, structure, and catalysis. Curr. Opin. Genet. Dev. 11, 155-161.

Martínez, J.M., Schroeder, D.C., Larsen, A., Bratbak, G., Wilson, W.H., 2007. Molecular dynamics of *Emiliania huxleyi* and cooccurring viruses during two separate mesocosm studies. Appl. Environ. Microbiol. 73, 554-562.

Milligan, K.L., Cosper, E.M., 1994. Isolation of virus capable of lysing the brown tide microalga, *Aureococcus anophagefferens*. Science 266, 805-807.

Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.M., Ogata, H., 2009. Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome res. 19, 1441-1449.

Moreau, H., Piganeau, G., Desdevises, Y., Cooke, R., Derelle, E., Grimsley, N., 2010. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J. Virol. 84, 12555-12563.

Nemova, N.N., Lysenko, L.A., Kantserova, N.P., 2010. Proteases of the calpain family: Structure and functions. Russ. J. Dev. Biol. 41, 318-325.

Neupartl M, *et al.* (2008) Chlorella viruses evoke a rapid release of K+ from host cells during the early phase of infection. *Virology* 372(2):340-348.

Noé, L., Kucherov, G., 2005. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. 33, W540-W543.

Ogata, H., Ray, J., Toyoda, K., Sandaa, R.A., Nagasaki, K., Bratbak, G., Claverie, J.M., 2011. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. ISME J. 5, 1143-1151.

Ohi, M.D., Vander Kooi, C.W., Rosenberg, J.A., Chazin, W.J., Gould, K.L., 2003. Structural insights into the U-box, a domain associated with multi-ubiquitination. Nat. Struct. Mol. Biol. 10, 250-255.

Pruzinska, A., Tanner, G., Anders, I., Roca, M., Hortensteiner, S., 2003. Chlorophyll breakdown: pheophorbide a oxygenase is a Rieske-type iron-sulfur protein, encoded by the accelerated cell death 1 gene. Proc. Natl. Acad. Sci. U.S.A. 100, 15259-15264.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R.D., 2012. The Pfam protein families database. Nucleic Acids Res. 40, D290-D301.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116-W120.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.-M., 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306, 1344-1350.

Rose, S.L., Fulton, J.M., Brown, C.M., Natale, F., Van Mooy, B.A.S., Bidle, K.D., 2014. Isolation and characterization of lipid rafts in *Emiliania huxleyi*: a role for membrane microdomains in host–virus interactions. Environ. Microbiol. 16(4), 1150-1166.

Roske, K., Foecking, M., Yooseph, S., Glass, J., Calcutt, M., Wise, K., 2010. A versatile palindromic amphipathic repeat coding sequence horizontally distributed among diverse bacterial and eukaryotic microbes. BMC Genomics 11, 430.s

Rowe, J.M., Dunlap, J.R., Gobler, C.J., Anderson, O.R., Gastrich, M.D., Wilhelm, S.W., 2008. Isolation of a non-phage-like lytic virus infecting *Aureococcus anophagefferens*. J. Phycol. 44, 71-76.

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K.E., Noordeloos, A.A.M., Brussaard, C.P.D., Claverie, J.-M., 2013. Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc. Natl. Acad. Sci. U.S.A. 110(26), 10800.

Schattner, P., Brooks, A.N., Lowe, T.M., 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33, W686-W689.

Schroeder, D.C., Oke, J., Malin, G., Wilson, W.H., 2002. Coccolithovirus (Phycodnaviridae): characterisation of a new large dsDNA algal virus that infects *Emiliana huxleyi*. Arch. Virol. 147, 1685-1698.

Short, S.M., 2012. The ecology of viruses that infect eukaryotic algae. Environ. Microbiol. 14, 2253-2271.

Sieburth, J.M., Johnson, P.W., Hargraves, P.E., 1988. Ultrastructure and ecology of *Aureococcus anophageferens* gen. et sp. nov. (Chrysophyceae): the dominant picoplankter during a bloom in Narragansett Bay, Rhode Island, summer 19851. J. Phycol. 24, 416-425.

Simonin, Y., Disson, O., Lerat, H., Antoine, E., Biname, F., Rosenberg, A.R., Desagher, S., Lassus, P., Bioulac-Sage, P., Hibner, U., 2009. Calpain activation by Hepatitis C virus proteins inhibits the extrinsic apoptotic signaling pathway. Hepatology 50, 1370-1379.

Srinivasan, V., Schnitzlein, W.M., Tripathy, D.N., 2001. Fowlpox virus encodes a novel DNA repair enzyme, CPD-photolyase, that restores infectivity of UV light-damaged virus. J. Virol. 75, 1681-1688.

Suhre, K., 2005. Gene and genome duplication in *Acanthamoeba polyphaga* Mimivirus. J. Virol. 79, 14095-14101.

Szkopinska, A., 2000. Ubiquinone. Biosynthesis of quinone ring and its isoprenoid side chain. Intracellular localization. Acta Biochim. Pol. 47, 469-480.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731-2739.

Thiel, G., Baumeister, D., Schroeder, I., Kast, S.M., Van Etten, J.L., Moroni, A., 2011. Minimal art: Or why small viral K+ channels are good tools for understanding basic structure and function relations. Biochim. Biophys. Acta . 1808, 580-588.

Vardi, A., Haramaty, L., Van Mooy, B.A., Fredricks, H.F., Kimmance, S.A., Larsen, A., Bidle, K.D., 2012. Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. Proc. Natl. Acad. Sci. U.S.A. 109, 19327-19332.

Vardi, A., Van Mooy, B.A.S., Fredricks, H.F., Popendorf, K.J., Ossolinski, J.E., Haramaty, L., Bidle, K.D., 2009. Viral glycosphingolipids induce lytic infection and cell death in marine phytoplankton. Science 326, 861-865.

Weinbauer, M.G., Wilhelm, S.W., Suttle, C.A., Garza, D.R., 1997. Photoreactivation compensates for UV damage and restores infectivity to natural marine virus communities. Appl. Environ. Microbiol. 63, 2200-2205.

Williams, T.A., Embley, T.M., Heinz, E., 2011. Informational gene phylogenies do not support a fourth domain of life for Nucleocytoplasmic Large DNA viruses. PLoS ONE 6, e21080.

Wilson, M.E., Maksaev, G., Haswell, E.S., 2013. MSCS-like mechanosensitive channels in plants and microbes. Biochemistry 52, 5708-5722.

Wilson, W.H., Schroeder, D.C., Allen, M.J., Holden, M.T.G., Parkhill, J., Barrell, B.G., Churcher, C., Hamlin, N., Mungall, K., Norbertczak, H., Quail, M.A., Price, C., Rabbinowitsch, E., Walker, D., Craigon, M., Roy, D., Ghazal, P., 2005. Complete genome sequence and lytic phase transcription profile of a coccolithovirus. Science 309, 1090-1092.

Yau, S., Lauro, F.M., DeMaere, M.Z., Brown, M.V., Thomas, T., Raftery, M.J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J.M., Gibson, J.A., Cavicchioli, R., 2011. Virophage control of Antarctic algal host–virus dynamics. Proc. Natl. Acad. Sci. U.S.A. 108, 6163-6168.

Yutin, N., Colson, P., Raoult, D., Koonin, E.V., 2013. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virology J. 10, 106.

Yutin, N., Wolf, Y., Raoult, D., Koonin, E., 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. Virology J. 6, 223.

Zhang, Q.-C., Qiu, L.-M., Yu, R.-C., Kong, F.-Z., Wang, Y.-F., Yan, T., Gobler, C.J., Zhou, M.-J., 2012. Emergence of brown tides caused by *Aureococcus anophagefferens* Hargraves et Sieburth in China. Harmful Algae 19, 117-124.

Zhu, W., Li, J., Liang, G., 2011. How does cellular heparan sulfate function in viral pathogenicity? Biomed. Environ. Sci. 24, 81-87.

Chapter II Appendix

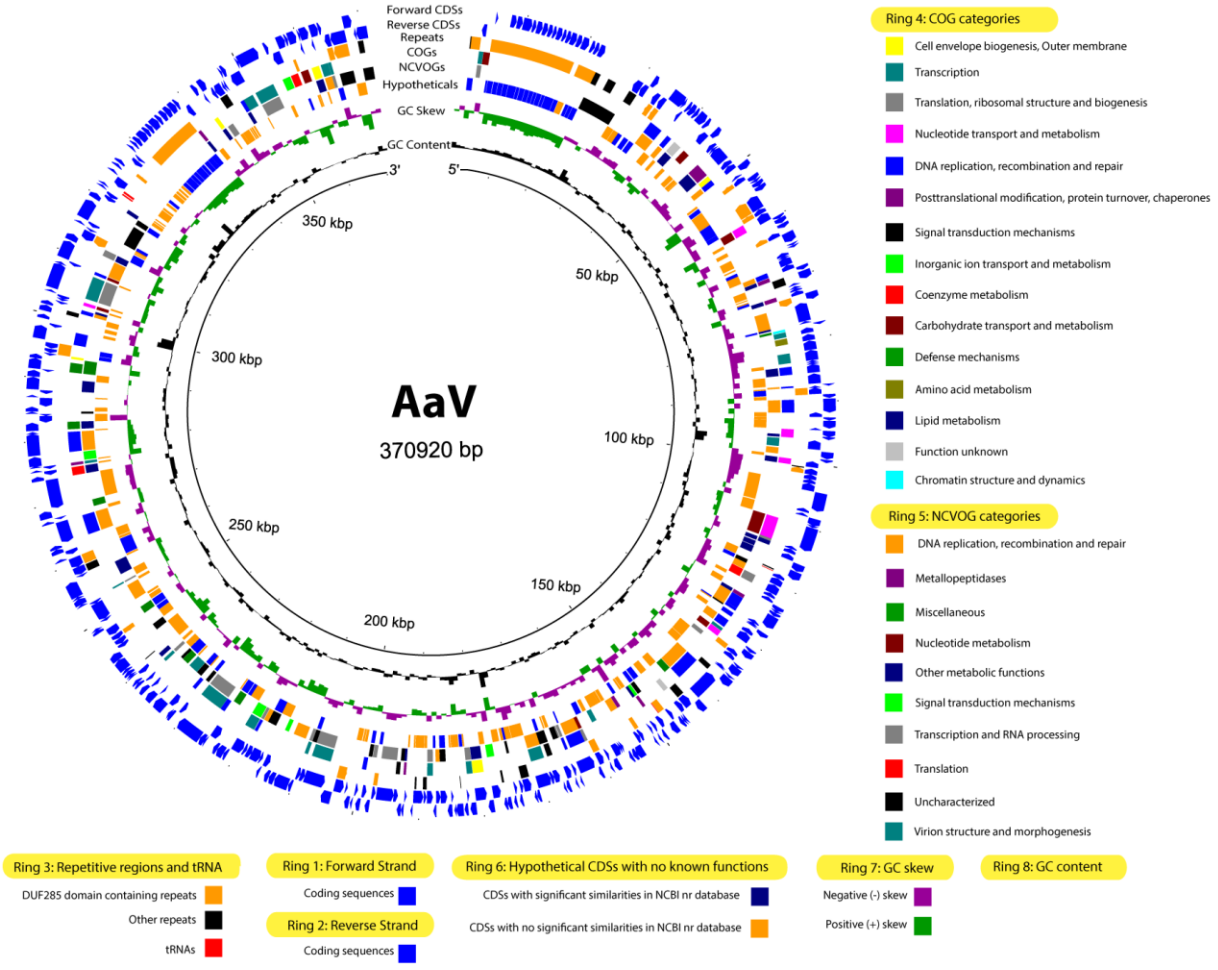


Figure 2.5: Genome map of AaV depicting forward and reverse strand genes, COG and NCVOG categories, repetitive regions, GC content, GC skew and hypothetical genes with no known functions.

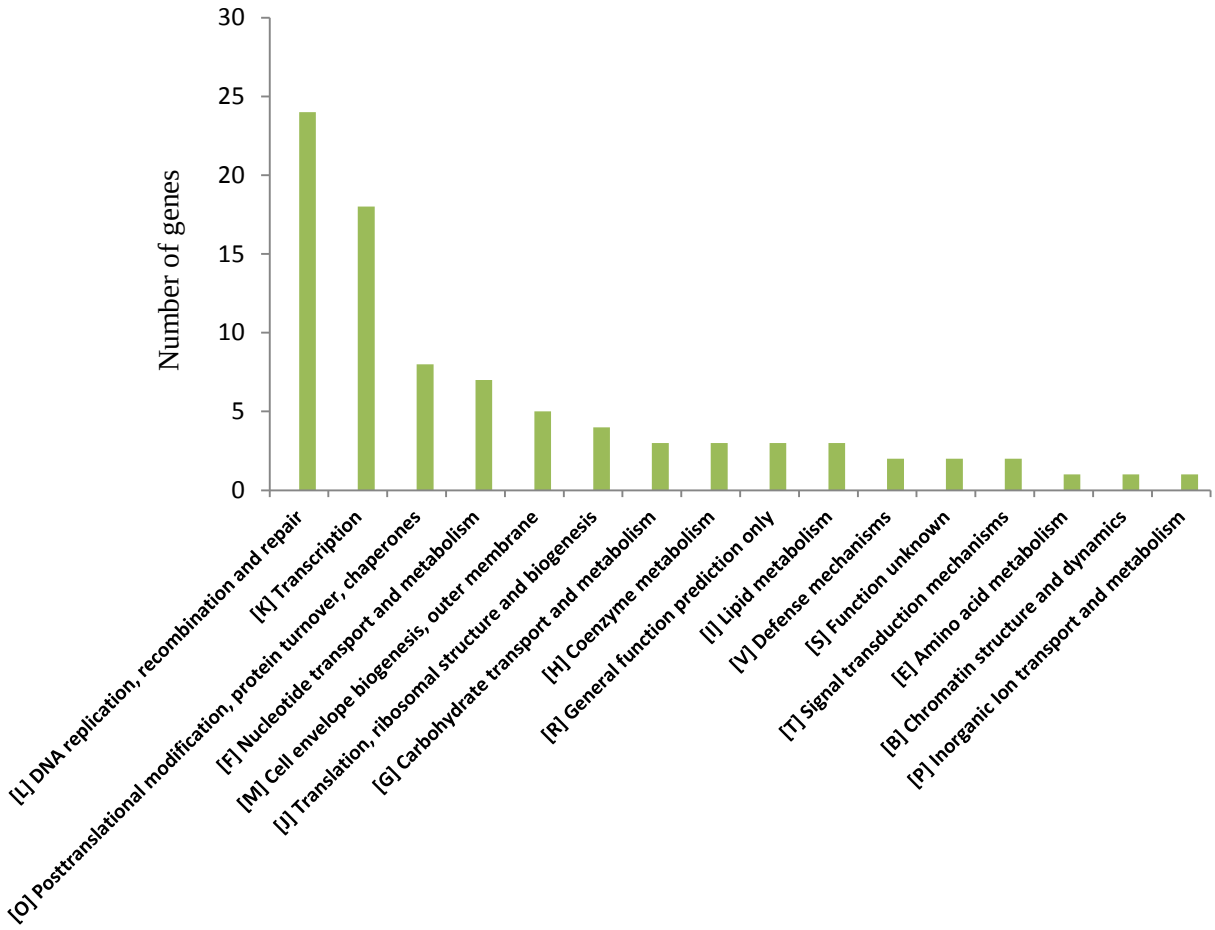


Figure 2.6: Number of AaV genes classified in different cluster of orthologous groups (COG) categories.

Figure 2.7: Maximum likelihood phylogenetic trees of -

a) ELP3-Histone acetyl transferase (AaV_368)

b) Translation elongation factor 5a (AaV_110)

c) Glucuronyl hydrolase (AaV_078)

d) Putative pectate lyases (AaV_003, 038 and 375)

e) Rhomboid family serine protease (AaV_077)

f) Calpain family thiol protease (AaV_045)

g) Undecaprenyl pyrophosphate synthase (AaV_077)

h) UbiA prenyltransferase (AaV_373)

i) AIM24 domain containing protein (AaV_144)

j) Carbohydrate sulfotransferase (AaV_102)

k) Phaeophorbide a oxygenase (AaV_372)

l) DUF285 domain containing CDSs.

m) PhoH family protein (AaV_210)

Edge supports at the nodes are calculated as 1000 iterations of expected likelihood weights (LR-ELW). Sequences from different sources are depicted as: AaV – Red, eukaryotes – blue, bacteria – green, archaea - purple, other viruses – black. Details about phylogenetic tree construction are available in the 'SI Materials and Methods' section.

Figure 2.7(a): ELP3-Histone acetyl transferase (AaV_368)

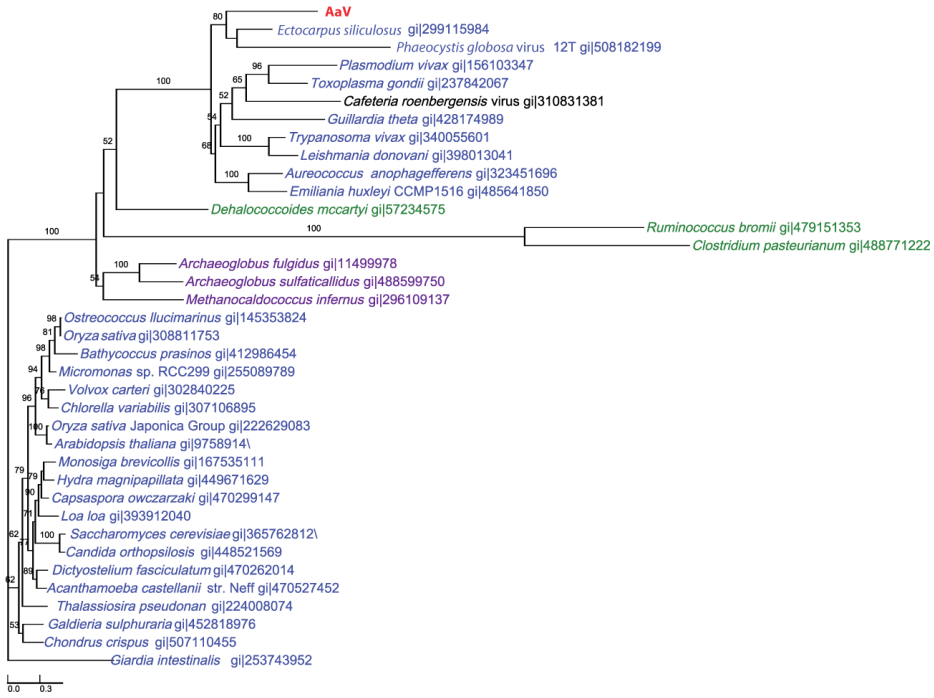


Figure 2.7(b): Translation elongation factor 5a (AaV_110)

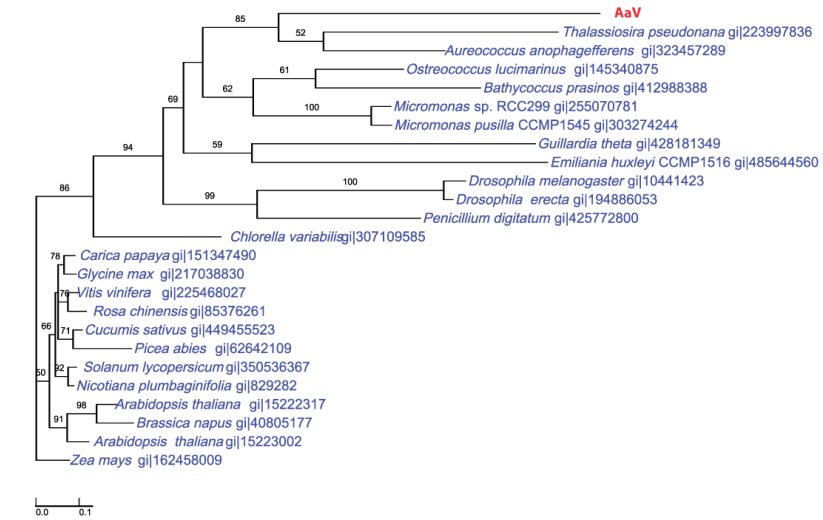


Figure 2.7 Continued.Figure 2.7(c): Glucuronyl hydrolase (AaV_078)

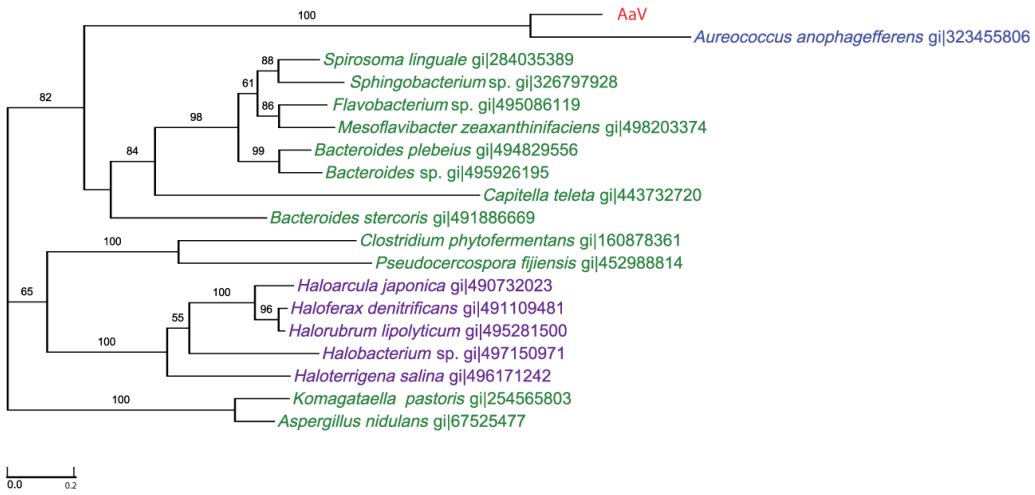


Figure 2.7(d): Putative pectate lyases (AaV_003, 038 and 375)

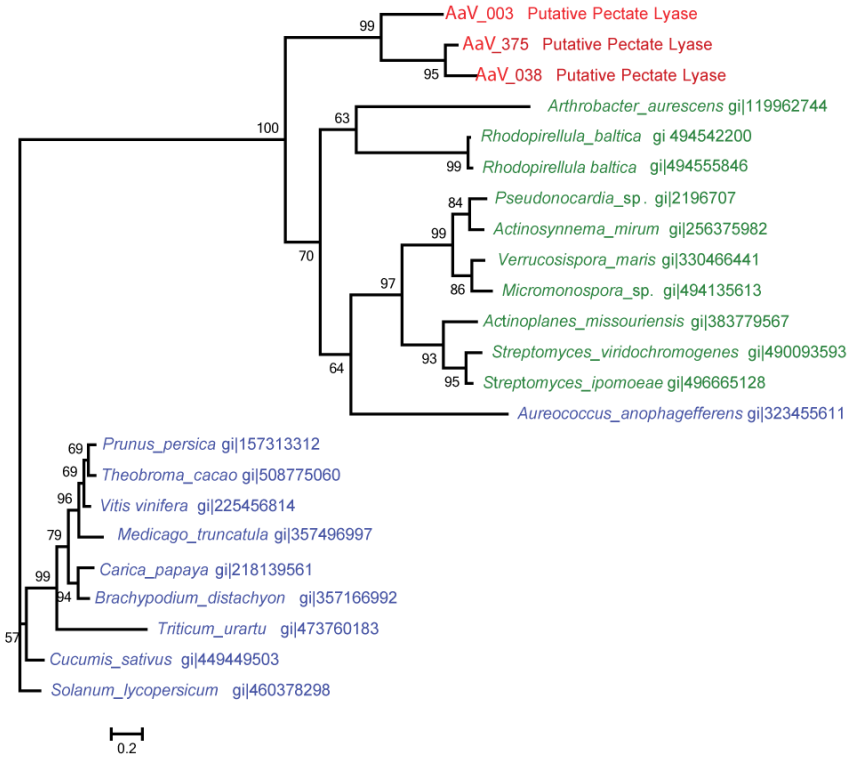


Figure 2.7 Continued.

Figure 2.7(e): Rhomboid family serine protease (AaV_077)

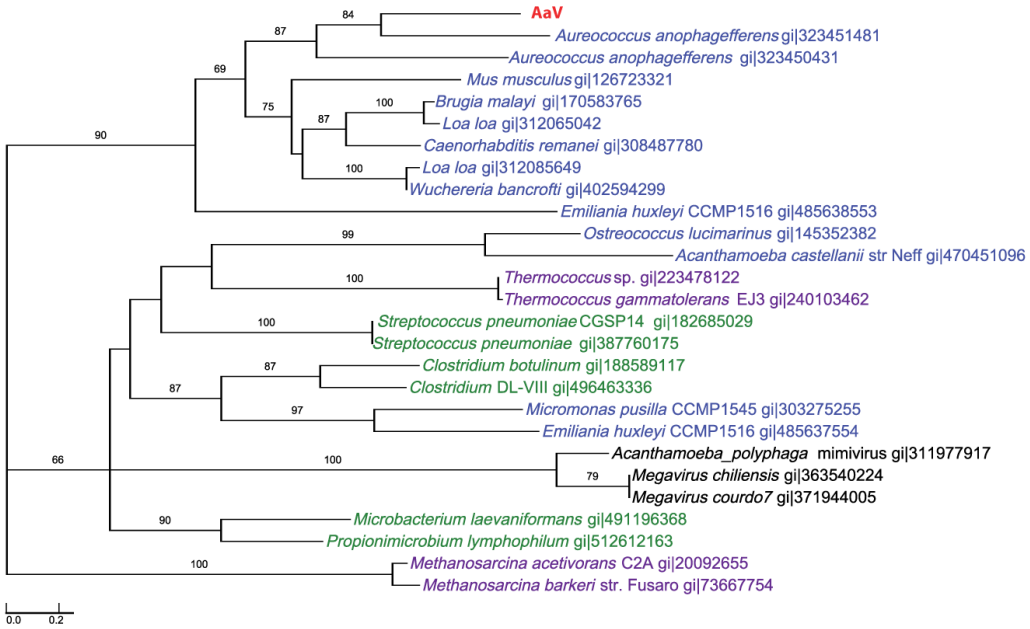


Figure 2.7(f): Calpain family thiol protease (AaV_045)

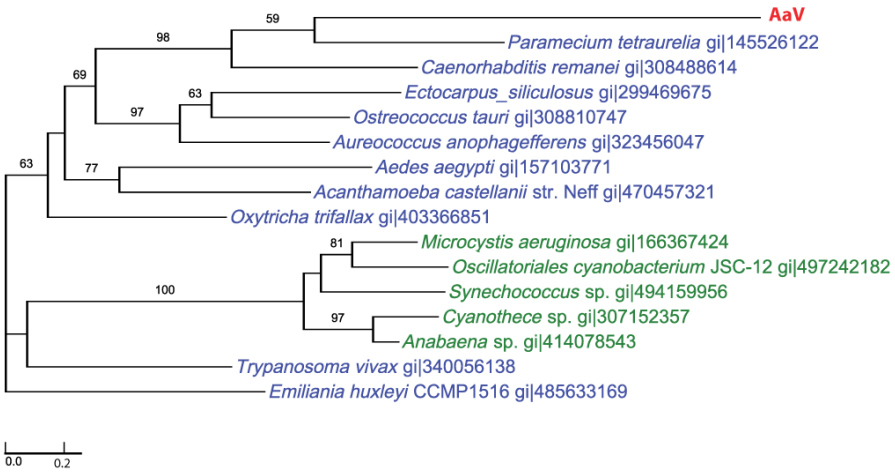


Figure 2.7 Continued.

Figure 2.7(g): Undecaprenyl pyrophosphate synthase (AaV_077)

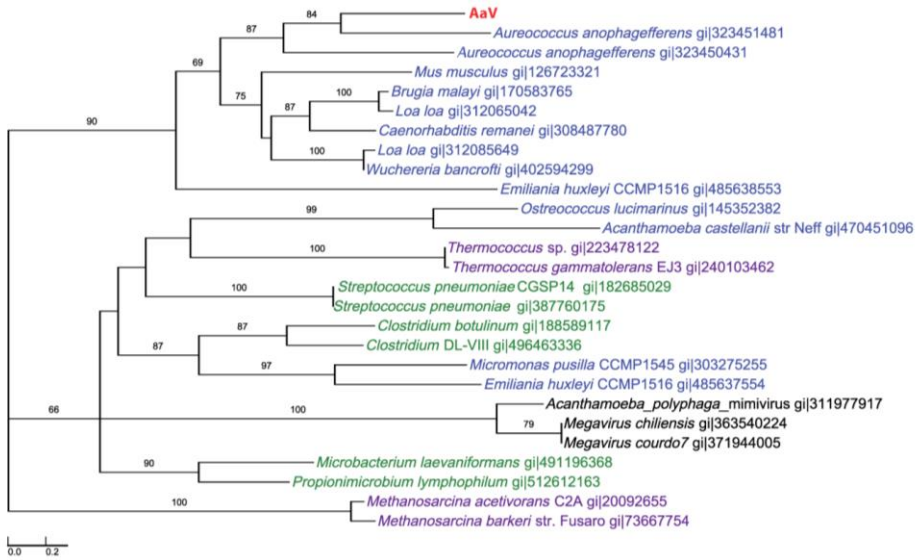


Figure 2.7(h): UbiA prenyltransferase (AaV_373)

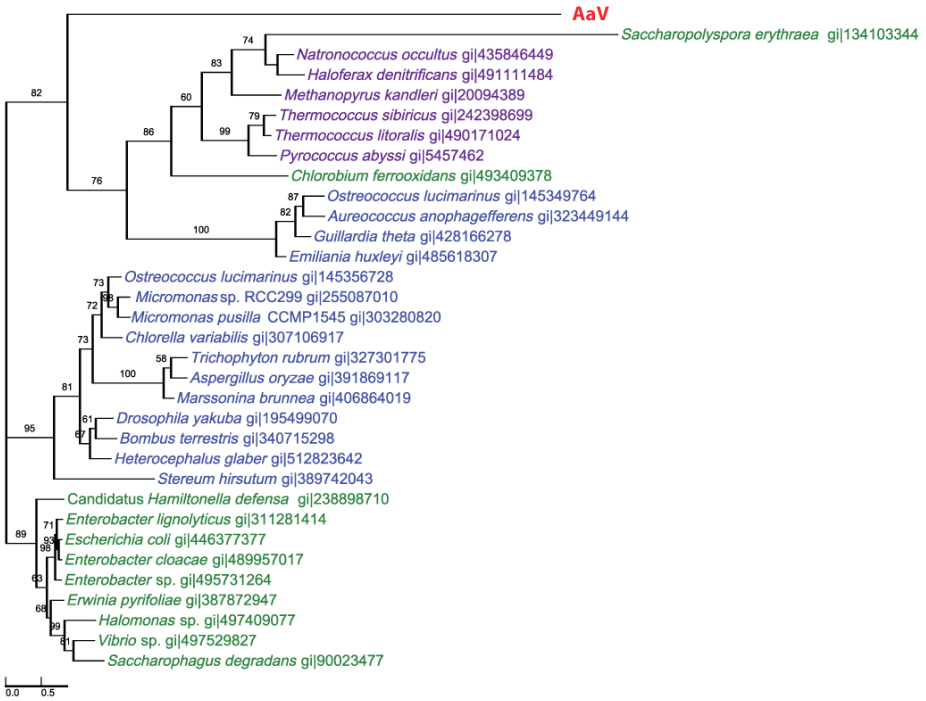


Figure 2.7 Continued.

Figure 2.7(i): AIM24 domain containing protein (AaV_144)

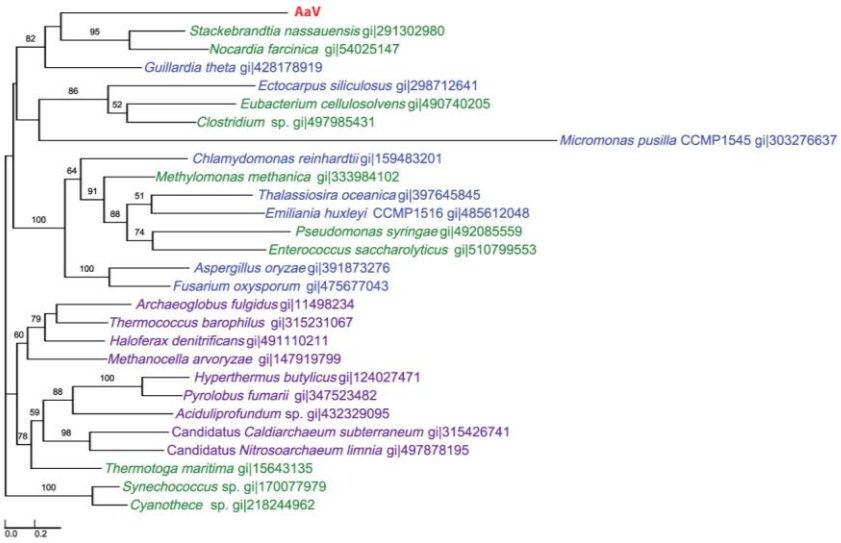


Figure 2.7(j): Carbohydrate sulfotransferase (AaV_102)

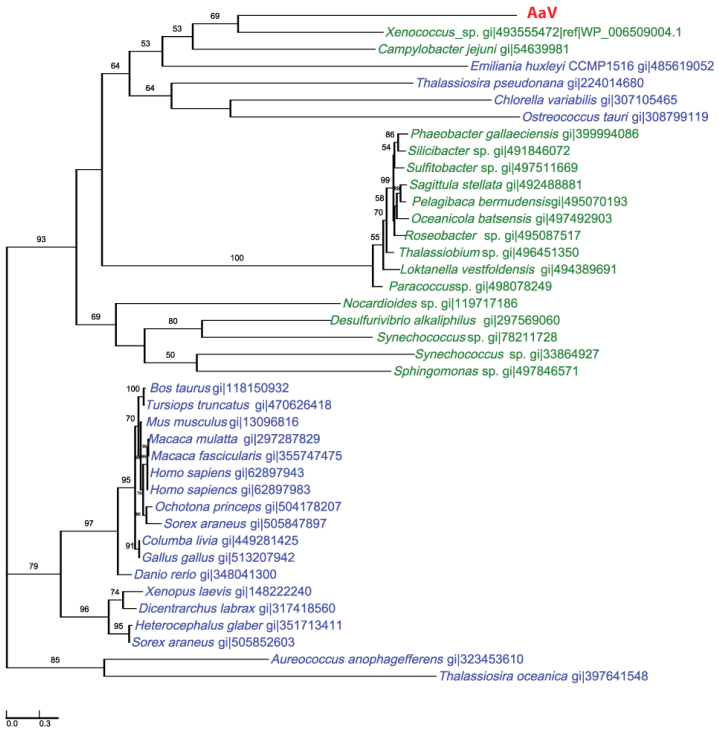


Figure 2.7 Continued.

Figure 2.7(k): Phaeophorbide a oxygenase (AaV_372)

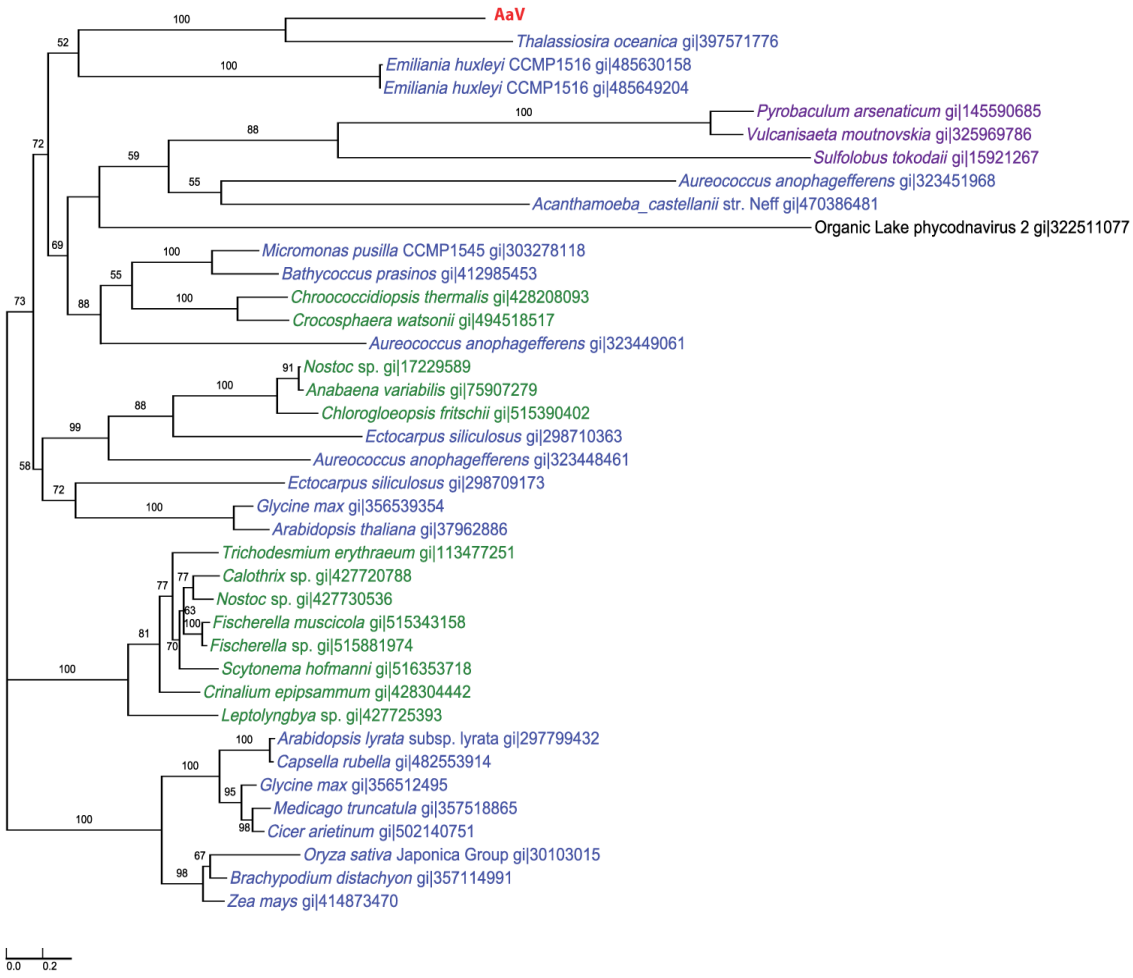


Figure 2.7 Continued.

Figure 2.7(l): DUF285 domain containing CDSs.

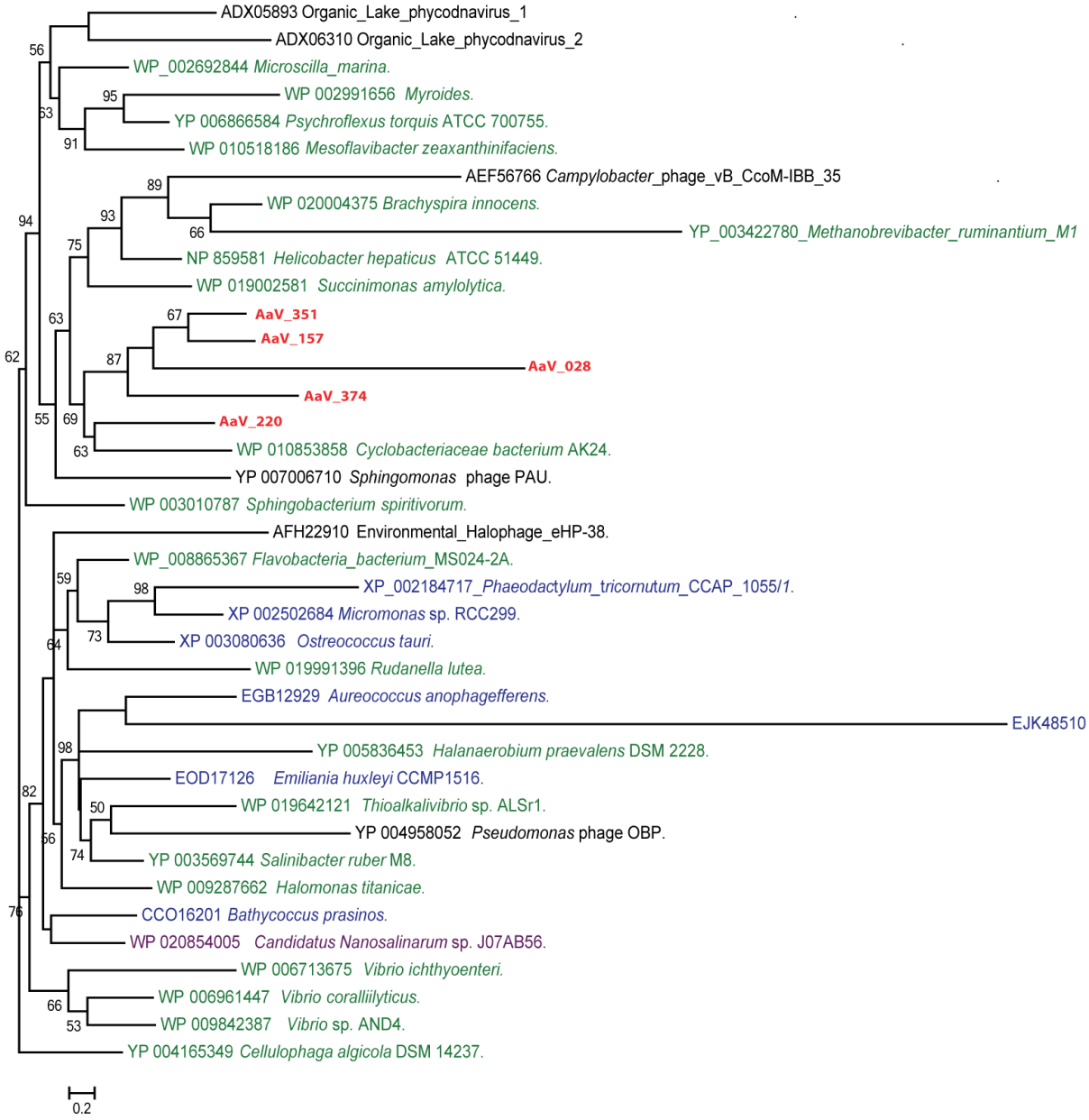


Figure 2.7 Continued.

Figure 2.7(m): PhoH family protein (AaV_210)

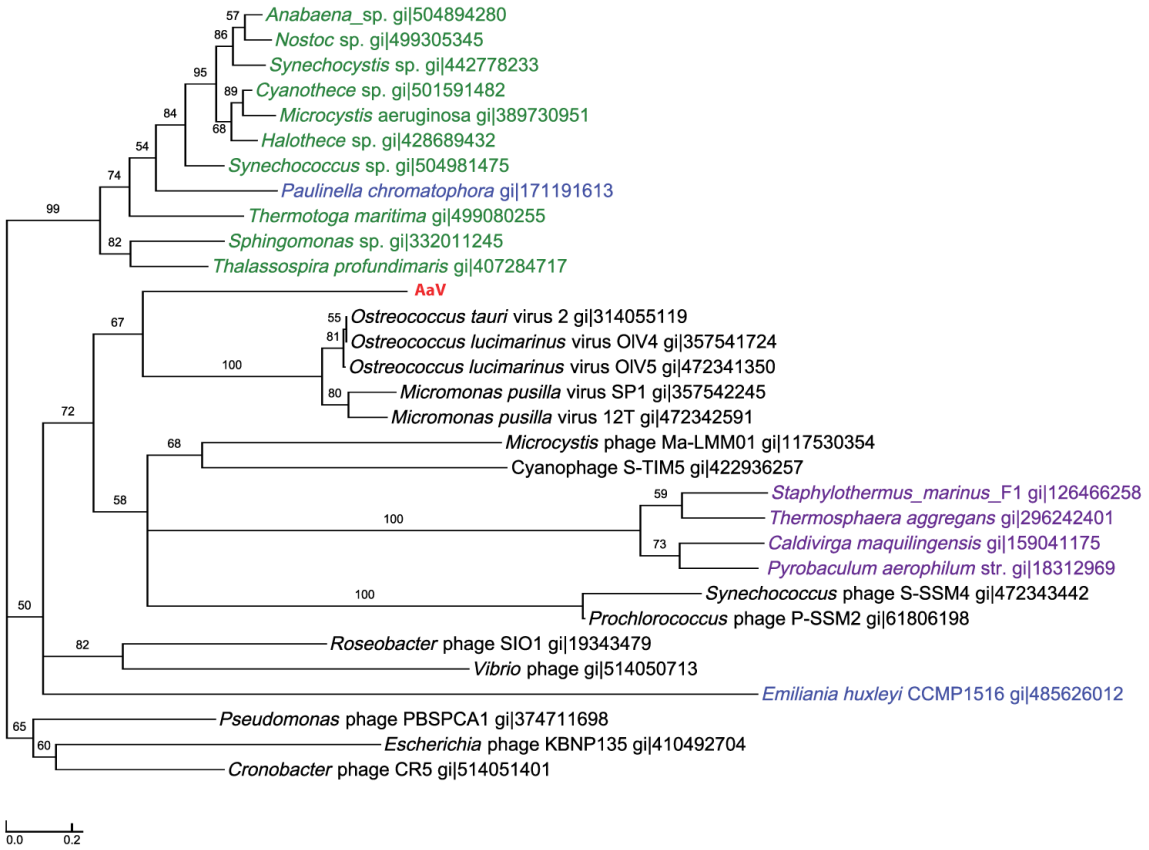


Figure 2.7 Continued.

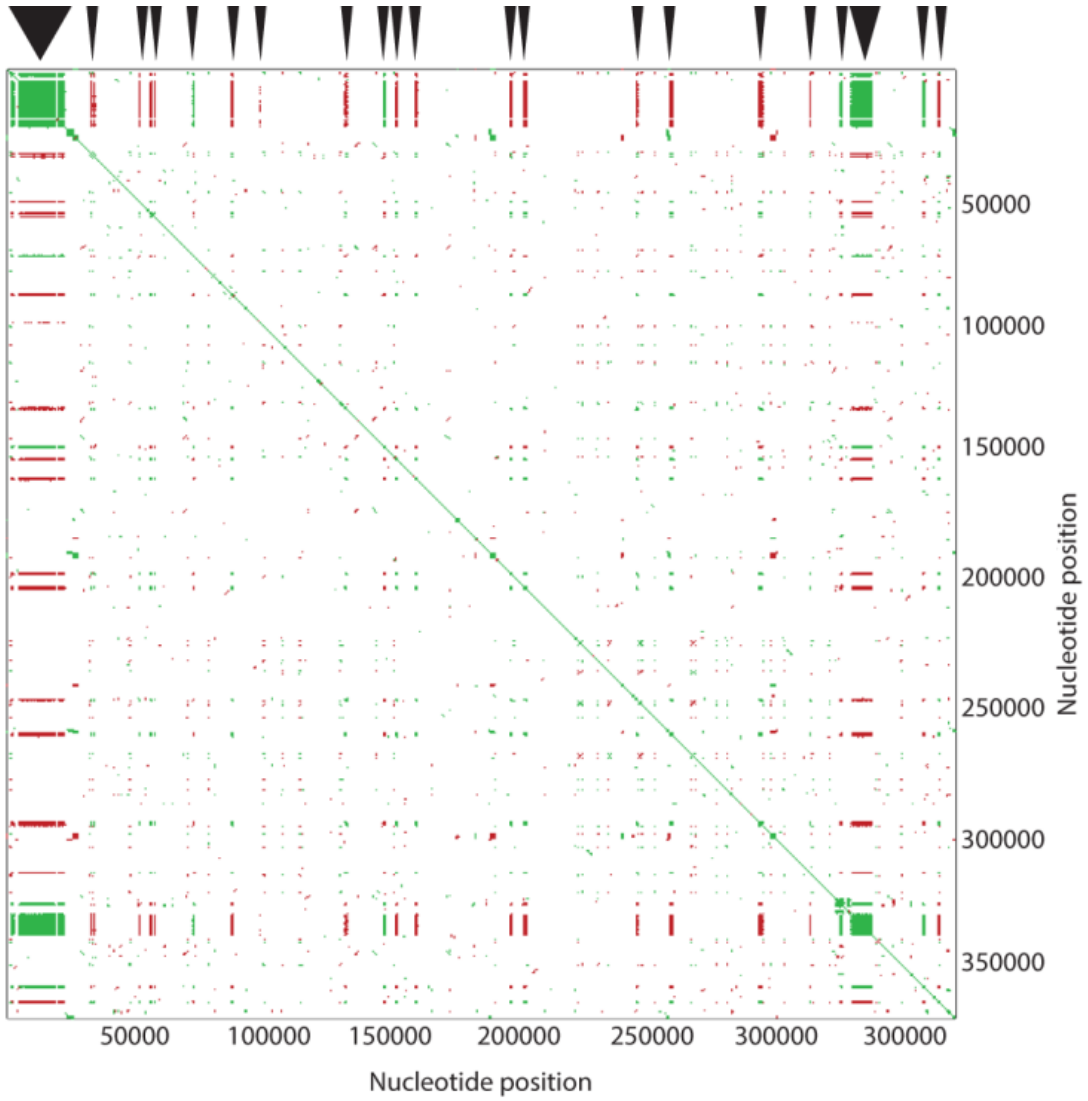


Figure 2.8: Whole genome dot plot of AaV. The DUF285 domain containing repetitive regions are marked with black arrowheads. Forward repeats are in green and inverted repeats are in red.



Figure 2.9(a)



Figure 2.9(b)

Figure 2.9: MEME sequence conservation logo of intergenic motifs present at the upstream regions of the DUF_ORFs. S5a) Intergenic motifs present at the upstream of the DUF_ORFs on positive strand. S5b) Intergenic motifs at the upstream region of the DUF_ORFs on negative strand.

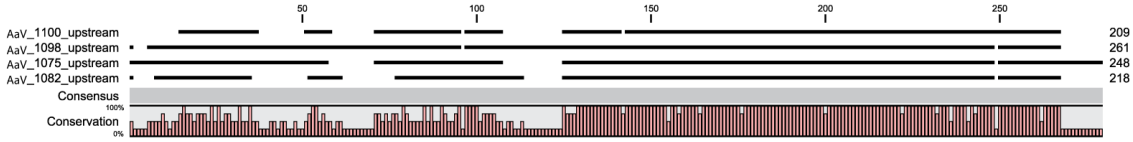


Figure 2.10(a)

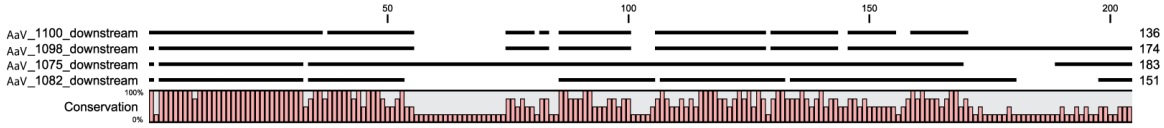


Figure 2.10(b)

Figure 2.10: Sequence conservation at the immediate upstream and downstream regions of ‘AaV and Phycodnavirus specific highly conserved genetic elements (AP_HGEs) clustered in paralog group 17. S5a) Sequence alignment of immediate upstream regions of the ORFs in paralog group 17. S5b) Sequence alignment of immediate downstream regions of the ORFs in paralog group 17.

Table 2.1: Mimivirus cluster of orthologous groups (MimiCOGs) present in AaV. The universal NCLDV core genes present in AaV are listed below.

mimiCOGs	**AaV gene identifier**	**Annotation**
CLS10031	AaV_134	Putative transcription factor
CLS10039	AaV_247, AaV_096	Capsid protein (2 copies)
CLS10089	AaV_193	DNA directed RNA polymerase II subunit rpb3
CLS10259	AaV_171	DNA mismatch repair ATPase (MutS)
CLS10201	AaV_084	Topoisomerase Type IA
CLS10230	AaV_293	DNA topoisomerase type IIA
CLS10090	AaV_298	DNA directed RNA polymerase subunit rpb9/M
CLS10250	AaV_224	RNA polymerase subunit RPB5
CLS10261	AaV_174	DNA directed RNA polymerase K subunit/rpb6
CLS10249	AaV_131	Putative DNA Directed RNA pol II subunit E`/ RPB 7
CLS10024	AaV_290	ERV/ALR sulphydryl oxidase
CLS10086	AaV_383	Fucosylgalactoside 3-alpha-galactosyltransferase
CLS10030	AaV_201	Putative Holliday junction resolvase
CLS10056	AaV_158	Putative metal dependent hydrolase
CLS10088	AaV_234, AaV_173	Nucleoside diphosphate hydrolase (MutT)
CLS10052	AaV_065	Proliferating cell nuclear antigen (PCNA)
CLS10035	AaV_208	Protein Disulfide Isomerase
CLS10216	AaV_271	RNA polymerase subunit Rpb10
CLS10047	AaV_034	Replication factor C subunit 2
CLS10049	AaV_034	Replication factor C subunit 2
CLS10258	AaV_125	Putative ribonuclease H1
CLS10130	AaV_109	Ribonucleoside di phosphate reductase alpha subunit
CLS10028	AaV_117	Putative TATA-box binding family protein.
CLS10057	AaV_381	Transcription Elongation factor TFIIS

Table 2.1 Continued.

mimiCOGs	AaV gene identifier	Annotation
CLS10055	AaV_203	Transcription factor TF IIB
CLS10011	AaV_074	Ubiquitin conjugating enzyme E2
CLS10214	AaV_066	SUMO-1 specific cysteine protease
CLS10066	AaV_130	VV_A18 like Helicase
CLS10218	AaV_159	Putative lambda-type exonuclease
CLS10212	AaV_200	Hypothetical protein found in other NCLDVs
CLS10222	AaV_214	Hypothetical protein
CLS10233	AaV_186	Hypothetical protein
CLS10236	AaV_175	Conserved hypothetical protein
CLS10043	AaV_250	Hypothetical protein
CLS10046	AaV_328	Hypothetical protein
CLS10070	AaV_116	Hypothetical protein (similar to other NCLDVs)
CLS10081	AaV_179	Hypothetical protein similar to NCLDVs
		Universal NCLDV core genes
CLS10071	AaV_361	Putative VLTF-3 like Transcription Factor
CLS10262	AaV_226, AaV_323	Putative D5 Primase/Helicase
CLS10015	AaV_180, AaV_269	Putative ATP-dependent RNA helicase
CLS10104	AaV_141	B family DNA polymerase
CLS10076	AaV_242, AaV_320	DNA directed RNA polymerase II largest subunit
CLS10053	AaV_002, AaV_222, AaV_370	RNA polymerase II second largest subunit
CLS10219	AaV_211, AaV_213	mRNA capping Enzyme
CLS10252	AaV_132	Ribonucleotide reductase small subunit
CLS10068	AaV_165	Putative A32 Virion packaging ATPase

Table 2.2: CDS name, location, length, functional annotation and probable phylogenetic origin of putative AaV coding sequences. The method of inferring the probable phylogenetic origins of the CDSs is detailed in the 'Supporting Materials and Methods' section.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_020	19144..20211	356	Hypothetical protein			
AaV_148	146630..147202	191	Putative Oxogluterate/Iron dependent dioxygenase			
AaV_141	137793..141878	1362	B family DNA polymerase	0417 (DNA replication, recombination & repair)	0038 (DNA replication, recombination & repair)	NCDLV
AaV_274	255031..257448	806	Beta-1,4 Galactosyltransferase		0044 (Other metabolic functions)	Eukaryotes
AaV_318	302875..303384	170	Putative deoxyuridine 5'_triphosphate nucleotidohydrolase (dUTPase)	0756 (Nucleotide transport and metabolism)	1068 (Nucleotide metabolism)	Eukaryotes
AaV_322	308412..308873	154	Cytosine-C5 specific DNA methyl transferase	0270 (DNA replication, recombination & repair)	1066 (Nucleotide metabolism)	Host/Bacteria [1]
AaV_222	205303..208587	1095	RNA polymerase beta subunit	0085 (Transcription)	0271 (Transcription and RNA processing)	NCLDV
AaV_171	161291..163936	882	DNA mismatch repair ATPase (MutS)	0249 (DNA replication, recombination & repair)	0105 (DNA replication, recombination & repair)	Bacteria
AaV_088	89626..91770	715	Putative DNA topoisomerase IA	0550 (DNA replication, recombination & repair)	0036 (DNA replication, recombination & repair)	Ambiguous
AaV_002	3383..4105	241	RNA polymerase II second largest subunit	0085 (Transcription)	0271 (Transcription and RNA processing)	NCLDV
AaV_142	141990..143108	373	DnaJ/Heat shock protein 40		0046 (Miscellaneous)	
AaV_001	1905..2891	329	DUF285 domain containing protein			Bacteria
AaV_010	11074..11631	186	DUF285 domain containing protein			Bacteria
AaV_011	11745..12623	293	DUF285 domain containing protein			Bacteria
AaV_012	12716..13486	257	DUF285 domain containing protein			Bacteria
AaV_013	13577..14275	233	DUF285 domain containing protein			Bacteria
AaV_014	14368..14907	180	DUF285 domain containing protein			Bacteria
AaV_015	14999..15829	277	DUF285 domain containing protein			Bacteria
AaV_016	15936..16475	180	DUF285 domain containing protein			Bacteria
AaV_017	16568..17224	219	DUF285 domain containing protein			Bacteria
AaV_018	17257..18195	313	DUF285 domain containing protein			Bacteria
AaV_019	18288..19043	252	DUF285 domain containing protein			Bacteria
AaV_021	20592..21182	197	DUF285 domain containing protein			Bacteria
AaV_022	21273..21896	208	DUF285 domain containing protein			Bacteria
AaV_023	21986..22906	307	DUF285 domain containing protein			Bacteria

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_343	326048..326563	172	DUF285 domain containing protein			Bacteria
AaV_348	330172..330837	222	DUF285 domain containing protein			Bacteria
AaV_349	330972..332114	381	DUF285 domain containing protein			Bacteria
AaV_350	332196..332816	207	DUF285 domain containing protein			Bacteria
AaV_351	332923..333537	205	DUF285 domain containing protein			Bacteria
AaV_352	333630..334799	390	DUF285 domain containing protein			Bacteria
AaV_353	334892..335470	193	DUF285 domain containing protein			Bacteria
AaV_354	335709..336482	258	DUF285 domain containing protein			Bacteria
AaV_355	336560..337399	280	DUF285 domain containing protein			Bacteria
AaV_356	337455..338417	321	DUF285 domain containing protein			Bacteria
AaV_374	358160..358588	143	DUF285 domain containing protein			Bacteria
AaV_004	5311..5613	101	DUF285 domain containing protein			Bacteria
AaV_005	5704..6642	313	DUF285 domain containing protein			Bacteria
AaV_006	6759..7586	276	DUF285 domain containing protein			Bacteria
AaV_007	7677..8525	283	DUF285 domain containing protein			Bacteria
AaV_008	8619..9473	285	DUF285 domain containing protein			Bacteria
AaV_009	9567..10982	472	DUF285 domain containing protein			Bacteria
AaV_212	197013..7567	185	DUF285 domain containing protein			Bacteria
AaV_220	202174..203622	483	Putative U-box E3 Ubiquitin Ligase			Bacteria
AaV_264	246081..246857	259	DUF285 domain containing protein			Bacteria
AaV_277	259229..260821	531	DUF285 domain containing protein			Bacteria
AaV_309	294069..4698	210	DUF285 domain containing protein			Bacteria
AaV_310	294791..295633	281	DUF285 domain containing protein			Bacteria

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_325	312989..314404	472	DUF285 domain containing protein			Bacteria
AaV_028	32900..33796	299	DUF285 domain containing protein			Bacteria
AaV_029	34078..34986	303	DUF285 domain containing protein			Bacteria
AaV_382	363982..364887	302	DUF285 domain containing protein			Bacteria
AaV_047	51851..52633	261	DUF285 domain containing protein			Bacteria
AaV_052	55463..57676	738	DUF285 domain containing protein			Bacteria
AaV_053	57735..58514	260	DUF285 domain containing protein			Bacteria
AaV_123	124100..124657	186	S-phase-kinase-associated protein-1 (Skp1)	5201 (Posttranslational modification, protein turnover, chaperones)	1299 (Other metabolic functions)	Eukaryote [2][8]
AaV_195	182353..182919	189	Hypothetical protein			*Aureococcus*
AaV_101	101690..102451	254	Putative Methyltransferase			*Aureococcus* [3]
AaV_175	165262..166164	301	Conserved hypothetical protein		1216 (Uncharacterized)	NCLDV
AaV_179	168119..168661	181	Hypothetical protein		1137 (Uncharacterized)	NCLDV
AaV_186	174276..174677	134	Hypothetical protein		1129 (Uncharacterized)	NCLDV
AaV_151	148486..149202	239	Hypothetical protein		1278 (Uncharacterized)	NCLDV
AaV_045	50257..51312	352	Calpain family thiol protease			Eukaryote [4]
AaV_104	104297..107164	956	Hypothetical protein			
AaV_107	108381..109499	373	Hypothetical protein			
AaV_108	109524..109736	71	Hypothetical protein			
AaV_114	117031..118317	429	Putative DegV domain (partial) containing protein			
AaV_121	123080..123763	228	Predicted coiled-coil domain-containing protein			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_122	123797..124021	75	Hypothetical protein			
AaV_131	130830..131572	162	Putative RNA pol II subunit E`	1095 (Transcription)		
AaV_135	134016..134372	119	Hypothetical protein			
AaV_136	134406..135632	409	Putative glycoprotein B domain containing protein (partial)			
AaV_139	136031..136324	98	Hypothetical protein			
AaV_140	136427..137716	430	Hypothetical protein			
AaV_145	145420..5713	98	Hypothetical protein			
AaV_146	145729.. 6028	100	Hypothetical protein			
AaV_147	146049..6585	179	Hypothetical protein			
AaV_154	150361..150672	104	Hypothetical protein			
AaV_156	151137..151451	105	Hypothetical protein			
AaV_160	154815..155429	205	Hypothetical protein			
AaV_161	155545..156018	158	Hypothetical protein			
AaV_167	159137..159310	58	Hypothetical protein			
AaV_172	164176..164424	83	Hypothetical protein			NCLDV
AaV_174	164986..165231	82	DNA directed RNA polymerase K subunit/rpb6		0522 (Transcription and RNA processing)	
AaV_176	166194..166748	185	Hypothetical protein			
AaV_177	166936..167169	78	Hypothetical protein			
AaV_181	170980..171525	182	Hypothetical protein			
AaV_182	171552..171923	124	Hypothetical protein			
AaV_185	173791..174270	160	Putative transmembrane domain containing protein			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_187	174735..175178	148	Hypothetical protein			
AaV_188	175566..176231	222	Putative membrane protein			
AaV_189	176240..176728	163	Putative valine-glycine-serine rich repeat containing protein			
AaV_191	178304..179185	294	Hypothetical protien with zinc finger domain (partial).			
AaV_196	183084..184208	375	Hypothetical protein			
AaV_197	184258..184440	61	Hypothetical protein			
AaV_202	187303..187746	148	Hypothetical protein			NCLDV
AaV_206	191007..191261	85	Hypothetical protein			
AaV_209	192802..193404	201	Putative thioredoxin like fold containing protein		0629 (Uncharacterized)	NCLDV
AaV_213	197704..198378	225	Putative m-RNA capping enzyme (partial)		1117 (Transcription and RNA processing)	
AaV_217	200378..200692	105	Hypothetical protein			
AaV_218	200860..201582	241	Putative Serine/threonine-protein kinase (partial).			
AaV_221	203792..205306	505	Hypothetical protein			
AaV_225	210024..210560	179	Hypothetical protein			
AaV_227	212937..213791	285	Replication factor C subunit		0071 (Miscellaneous)	Eukaryote
AaV_228	213859..214551	231	C3HC4 domain containing E3 ubiquitin ligase		0330 (Signal transduction regulation)	Eukaryote
AaV_230	215469..215882	138	Putative Zinc/RING finger domain containing protein			
AaV_236	221152..221664	171	Hypothetical protein			
AaV_237	221708..222805	366	Putative calcium binding site containing protein			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_238	222846..223472	209	Hypothetical protein			NCLDV
AaV_240	225025..225234	70	Hypothetical protein			
AaV_245	231970..232170	67	Hypothetical protein			
AaV_251	237069..237815	249	Hypothetical protein			
AaV_254	238832..239308	159	Hypothetical protein			
AaV_256	240065..240277	71	Putative DNA repair protein (partial)			
AaV_257	240298..240570	91	Hypothetical protein			
AaV_258	240600..241157	186	Hypothetical protein			
AaV_259	241206..241637	144	Hypothetical protein			
AaV_260	241670..243403	578	Hypothetical protein			
AaV_266	248513..249073	187	Hypothetical protein			
AaV_267	249117..249308	64	Hypothetical protein			
AaV_272	253700..254653	318	Hypothetical protein			
AaV_273	254676..255014	113	Hypothetical protein			
AaV_278	260956..261342	129	Hypothetical protein			
AaV_280	262232..262993	254	Hypothetical protein			
AaV_283	267011..267430	140	Hypothetical protein			
AaV_285	267923..269080	386	Hypothetical protein		1343 (Miscellaneous)	NCLDV
AaV_289	275278..275532	85	Hypothetical protein			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_291	276016..276195	60	Hypothetical protein			
AaV_294	281257..281448	64	Hypothetical protein			
AaV_299	284698..285405	236	Hypothetical protein			
AaV_301	286982..287176	65	Hypothetical protein			
AaV_304	290100..291083	328	Hypothetical protein			
AaV_313	299211..299843	211	Hypothetical protein			
AaV_314	300376..301308	311	Hypothetical protein			Bacteria
AaV_319	303468..303761	98	Hypothetical protein			
AaV_321	308022..308285	88	Hypothetical protein			
AaV_327	314945..315265	107	Hypothetical protein			
AaV_027	31567..32358	264	Hypothetical protein			Ambiguous
AaV_329	315876..318812	979	Putative DNA polymerase III subunit alpha (partial)		0420 (Uncharacterized)	
AaV_330	319354..320202	283	Hypothetical protein		1012 (Uncharacterized)	NCLDV
AaV_333	321538..321732	65	Hypothetical protein			
AaV_334	322006..322194	63	Hypothetical protein			
AaV_341	324353..324706	118	Hypothetical protein			
AaV_342	325392..325799	136	Hypothetical protein			
AaV_346	328908..329252	115	Putative methyl malonate semi aldehyde dehydrogenase family protein (partial)			
AaV_347	329593..330012	140	Hypothetical protein			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_366	346929..347225	99	Hypothetical protein			
AaV_369	349969..350133	55	Hypothetical protein			
AaV_371	354719..355279	187	Hypothetical protein			
AaV_377	360343..360543	67	Hypothetical protein			
AaV_384	367466..368788	441	Putative repeat containinig protein			
AaV_031	37417..37650	78	Hypothetical protein			
AaV_033	38163..38678	172	Hypothetical protein			
AaV_035	39630..40175	182	Putative 3-octaprenyl-4-hydroxybenzoate carboxy-lyase/Phytase (partial)			
AaV_037	41888..42493	202	Putative DUF2722 domain containing protein (partial)			
AaV_039	43759..44658	300	HMG box domain containing protein			
AaV_048	52775..53518	248	Hypothetical protein			
AaV_049	53545..53913	123	Hypothetical protein			
AaV_050	53964..54668	235	Putative DUF2457 domain containing protein (partial)			
AaV_051	54677..55405	243	Hypothetical protein.			
AaV_055	60074..60565	164	Possible chromosome segregation ATPase (Partial)			
AaV_056	60577..61338	254	Putative glycosyl transferase family protein (partial)			Eukaryote
AaV_057	61475..62128	218	Hypothetical protein.			
AaV_059	63059..64540	494	Hypothetical protein.			
AaV_060	64895..65119	75	Hypothetical protein.			
AaV_067	69400..70629	410	Putative ribosomal protein (partial)			
AaV_087	88926..89519	198	Hypothetical protein.			
AaV_090	92530..93009	160	Hypothetical protein.			
AaV_091	93064..94224	387	Hypothetical protein.			
AaV_098	99094..99567	158	Putative uncharacterized protein (DUF3478 domain containing)			
AaV_102	102469..103098	210	Putative Carbohydrate Sulfotransferase			Bacteria

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_103	103117..104265	383	Hypothetical protein.			
AaV_105	107166..107714	183	Hypothetical protein.			
AaV_106	107937..108344	136	Hypothetical protein.			
AaV_116	118429..119058	210	Hypothetical protein		0158 (Uncharacterized)	NCLDV
AaV_117	119114..120205	364	Putative TATA-box binding family protein.		0313 (DNA replication, recombination & repair)	
AaV_119	121727..122290	188	Putative phage structural protein like protein			Bacteria
AaV_120	122439..122993	185	Putative phage structural protein			Bacteria
AaV_126	127225..127446	74	Hypothetical protein		0842 (Uncharacterized)	NCLDV
AaV_127	127462..127626	55	Hypothetical protein.			
AaV_129	128389..129087	233	Hypothetical protein.			
AaV_134	133230..133919	230	Putative VLTF2 like transcription factor		1164 (Transcription and RNA processing)	NCLDV
AaV_137	135629..135787	53	Hypothetical protein.			
AaV_138	135832..136002	57	Hypothetical protein.			
AaV_143	143105..144397	431	Hypothetical protein.			
AaV_155	150669..151025	119	Hypothetical protein.			
AaV_162	156071..156364	98	Hypothetical protein.			
AaV_163	156381..156659	93	Hypothetical protein.			
AaV_164	156665..157759	365	Putative adenylate kinase domain protein (partial)			
AaV_166	158670..159092	141	Putative SecD domain protein (partial).			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_183	171934..173172	413	Hypothetical protein.			
AaV_184	173271..173750	160	Putative GTPase domain containing protein (partial).			
AaV_190	176938..178203	422	Putative Serine threonine protein kinase haspin		0285 (Signal transduction regulation).	NCLDV
AaV_194	182038..182274	79	Hypothetical protein.			
AaV_198	184545..184799	85	Hypothetical protein.			
AaV_200	185534..186550	339	Hypothetical protein		1423 (Uncharacterized)	NCLDV
AaV_204	188832..189269	146	Hypothetical protein.			
AaV_207	192023..192304	94	Hypothetical protein.			
AaV_214	198480..199118	213	Hypothetical protein		0645 (Uncharacterized)	NCLDV
AaV_216	199847..200278	144	Hypothetical protein.			
AaV_219	201574..201786	71	Hypothetical protein.			
AaV_223	208620..209321	234	Hypothetical protein		1024 (Uncharacterized) (exclusive to iridoviruses, this virus also have this)	Bacteria
AaV_229	214591..215430	280	Putative metallopeptidase like protein			
AaV_231	215893..216300	136	Hypothetical protein.			
AaV_246	232167..233117	317	Hypothetical protein		0632 (Uncharacterized)	NCLDV
AaV_248	234898..235167	90	Hypothetical protein.			
AaV_250	236557..237018	154	Hypothetical protein		0628 (Uncharacterized)	NCLDV
AaV_252	237807..238262	152	Hypothetical protein.			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_262	244564..245160	199	Hypothetical protein.			
AaV_263	245193..246002	270	Hypothetical protein.			
AaV_265	247052..248308	419	Hypothetical protein		1343 (Miscellaneous)	NCLDV
AaV_268	249555..249911	119	Hypothetical protein.			
AaV_270	252750..253178	143	Hypothetical protein.			
AaV_275	257563..257847	95	Hypothetical protein.			
AaV_276	257971..259206	412	Hypothetical protein			Bacteria
AaV_279	261481..262206	242	Hypothetical protein.			
AaV_284	267419..267574	52	Hypothetical protein.			
AaV_286	269194..273804	1537	Hypothetical protein.			
AaV_288	274558..275208	217	Putative SAM dependent methyltransferase	2227 (Coenzyme metabolism)	1191 (Other metabolic functions)	Ambiguous
AaV_296	282932..283957	342	Hypothetical protein.			
AaV_297	283986..284255	90	Hypothetical protein.			
AaV_300	285487..286824	446	Putative HNH endonuclease			
AaV_302	287500..288033	178	Class I DNA binding protein			
AaV_305	291080..291283	68	Hypothetical protein.			
AaV_025	29246..30481	412	Hypothetical protein.			
AaV_308	293627..293983	119	Putative bZIP transcription factor	1792 (Cell envelope biogenesis, outer membrane)		
AaV_312	298184..298663	160	Hypothetical protein.			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_317	302587.. 2871	95	Hypothetical protein.			
AaV_026	30573..31532	320	Hypothetical protein.			
AaV_326	314610..4840	77	Hypothetical protein.			
AaV_328	315300..315833	178	Hypothetical protein		1083 (Uncharacterized)	NCLDV
AaV_332	320815..1393	193	Hypothetical protein.			
AaV_340	322911..4200	430	Hypothetical protein.			
AaV_344	327382..328068	229	Hypothetical protein.			
AaV_345	328236..328646	137	Hypothetical protein.			
AaV_358	339301..340125	275	Hypothetical protein.			
AaV_360	342738..342983	82	Hypothetical protein.			
AaV_362	344117..344638	174	Putative SAP-domain protein			
AaV_363	344700..345923	408	Hypothetical protein.			
AaV_364	346121..346342	74	Hypothetical protein.			
AaV_365	346456..346752	99	Hypothetical protein.			
AaV_379	361876..362145	90	Putative membrane protein.			
AaV_385	368783..369082	100	Hypothetical protein.			
AaV_032	37640..38116	159	Hypothetical protein.			
AaV_040	44865..45500	212	Hypothetical protein.			
AaV_041	45573..45782	70	Hypothetical protein.			
AaV_046	51299..51655	119	Putative seryl-tRNA synthetase partial			
AaV_062	65628..66191	188	Hypothetical protein			NCLDV
AaV_063	66212..66391	60	Hypothetical protein.			
AaV_068	70852..71013	54	Hypothetical protein.			
AaV_069	71345..71524	60	Hypothetical protein.			

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_165	157836..158666	277	Putative A32 Virion packaging ATPase		0249 (Virion structure and morphogenesis)	NCLDV
AaV_170	160747..161298	184	Putative Deoxynucleoside kinase	0480 (Translation, ribosomal structure and biogenesis)	1067 (Nucleoside/Nucleotide_kinase,Nucleotide metabolism)	Bacteria
AaV_311	296033..298075	681	Superfamily I Helicase	1112 (DNA replication, recombination & repair)		Ambiguous
AaV_233	217851..218639	263	Replication factor C small subunit 2		0001 (DNA replication, recombination & repair)	Archaea
AaV_306	291345..292946	534	Putative ABC-transporter family protein	1132 (Defense mechanisms)	0002 (Miscellaneous)	Eukaryote
AaV_113	115796..116926	377	Putative adenine specific DNA methyltransferase		0234 (Other metabolic functions	Bacteria
AaV_144	144543..145367	275	Putative AIM24 domain containing protein	2013 (Function unknown)		Ambiguous
AaV_180	168698..170956	753	Putative VV D6R-type Helicase	1061 (Transcription / DNA replication, recombination, & repair)	0031 (DNA replication, recombination & repair)	NCLDV
AaV_034	38715..39557	281	Replication factor C subunit 2	2256 (DNA replication, recombination & repair)	1351 (DNA replication, recombination & repair)	Ambiguous
AaV_066	68577..69206	210	SUMO-1 specific cysteine protease	5160 (Posttranslational modification, protein turnover, chaperones)	0246 (Other metabolic functions)	NCLDV
AaV_331	320186..320782	199	CDP-alcohol phosphatidyltransferase	0558 (Lipid metabolism)		Bacteria
AaV_082	84029..85357	443	Putative DNA photolyase class II	0415 (DNA replication, recombination & repair)	1004 (Other metabolic functions)	Ambiguous
AaV_093	94629..95558	310	Putative Cytosine Deaminase	0402 (Nucleotide transport and metabolism / General function prediction only)		Bacteria
AaV_226	210628..212871	748	Putative D5 Primase/Helicase		0023 (DNA replication, recombination & repair)	Ambiguous
AaV_094	95563..96015	151	Putative deoxycytidylate deaminase	2131 (Nucleotide transport and metabolism)	1064 (Other metabolic functions)	*Aureococcus*
AaV_030	35117..37414	766	Putative Superfamily II RNA helicase	4581 (DNA replication, recombination & repair)	0030 (DNA replication, recombination & repair)	
AaV_316	302038..302403	122	Putative CMP/dCMP deaminase		1064 (Other metabolic functions)	

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_070	71580..72029	150	Hypothetical protein			NCLDV
AaV_074	77080..77388	103	Hypothetical protein		0329 (Other metabolic functions)	NCLDV
AaV_080	83441..83692	84	Hypothetical protein.			
AaV_081	83732..84004	91	Hypothetical protein.			
AaV_083	85539..87119	527	Hypothetical protein.			
AaV_085	87554..87763	70	Putative DNA photolyase (partial)			
AaV_089	91752..92459	236	Hypothetical protein.			
AaV_092	94246..94542	99	Hypothetical protein.			
AaV_249	235356..236498	381	Hypothetical protein		1343 (Miscellaneous)	NCLDV
AaV_298	284330..284653	108	DNA directed RNA polymerase subunit rpb9/M		0521 (Uncharacterized)	Ambiguous
AaV_239	223635..224891	419	Hypothetical protein		1343 (Miscellaneous)	NCLDV
AaV_243	230431..231069	213	Hypothetical protein		1278 (Uncharacterized)	NCLDV
AaV_232	216327..217844	506	Hypothetical protein		1131 (Uncharacterized)	NCLDV
AaV_096	96238..97650	471	Putative Capsid protein		0022 (Virion structure and morphogenesis)	NCLDV
AaV_124	124721..126679	653	NAD dependent DNA ligase	0272 (DNA replication, recombination & repair)	0035 (DNA replication, recombination & repair)	Bacteria
AaV_071	72063..72725	221	Cyclophilin type peptydyl-prolyl cis-trans isomerase	0652 (Posttranslational modification, protein turnover, chaperones)	0711 (Other metabolic functions)	Eukaryote
AaV_210	193409..194077	223	Phosphate starvation-inducible protein PhoH	1702 (Signal transduction mechanisms)		Bacteria
AaV_099	99594..100517	308	Putative Polynucleotide-kinase-3 phosphatase	4088 (Nucleotide transport and metabolism)	0243 (Other metabolic functions)	Ambiguous
AaV_064	66416..67648	411	Hypothetical protein.			
AaV_065	67665..68477	271	Proliferating cell nuclear antigen (PCNA)/DNA polymerase sliding clamp	0592 (DNA replication, recombination & repair)	0241 (DNA replication, recombination & repair)	NCLDV
AaV_199	184821..185537	239	Oxogluterate/Iron dependent dioxygenas			Bacteria
AaV_036	40319..41701	461	Oxogluterate/Iron dependent dioxygenase	3128 (Function unknown)	1166 (Other metabolic functions)	Bacteria

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_193	181068..182021	318	DNA directed RNA polymerase II subunit rpb3	0202 (Transcription)	0635 (RNA_polymerase,Transcription and RNA processing)	Eukaryotes
AaV_224	209350..209919	190	RNA polymerase subunit RPB5	2012 (Transcription)	0273 (Transcription and RNA processing)	Ambiguous
AaV_128	127678..128382	235	DNA (cytosine-5) methyltransferase	0863 (DNA replication, recombination & repair)		Bacteria
AaV_295	281695..282939	415	Putative N6 Adenine specific DNA methyltransferase	0286 (Defense mechanisms)	0234 (Other metabolic functions)	Ambiguous
AaV_315	301421..302026	202	Putative DNA polymerase epsilon subunit	0847 (DNA replication, recombination & repair)	0047 (DNA replication, recombination & repair)	Eukaryotic
AaV_044	49279..50178	300	DNA polymerase X family protein	1796 (DNA replication, recombination & repair)	004 (DNA replication, recombination & repair)	Eukaryote
AaV_215	199152..199841	230	Crossover endonuclease Mus81	1948 (DNA replication, recombination & repair)		Ambiguous
AaV_084	87334..87597	88	Topoisomerase Type IA	0550 (DNA replication, recombination & repair)	0033 (DNA replication, recombination & repair)	Bacteria
AaV_293	277850..281224	1125	DNA topoisomerase type IIA	0187 (DNA replication, recombination & repair)	0037 (DNA replication, recombination & repair)	NCLDV
AaV_307	292958.. 3623	222	DnaJ/Heat shock protein 40		0046 (Miscellaneous)	Eukaryotes
AaV_100	100780..101676	299	Family 14 Glycosyltransferase			Eukaryote
AaV_157	152372..152920	183	DUF285 domain containing protein			Bacteria
AaV_149	147261..148250	330	Putative DUF285 domain containing protein			Bacteria
AaV_169	159671..160642	324	Putative DUF285 domain containing protein			Bacteria
AaV_168	159671..160642	324	DUF285 domain containing protein			Bacteria
AaV_290	275529..275972	148	ERV/ALR sulphydryl oxidase	5054 (Posttranslational modification, protein turnover, chaperones)	0052 (Virion structure and morphogenesis)	NCLDV
AaV_110	113516..113947	144	Eukaryotic Translation Elongation factor 5A	0231 (Translation, ribosomal structure and biogenesis)		*Aureococcus* [2]
AaV_078	80252..81715	488	Family 88 Glycosyl Hydrolase (Unsaturated Glucuronyl hydrolase)	4225 (General function prediction only)		*Aureococcus*

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_378	360585..361874	430	Glycosyl Transferase family 25	3306 (Cell envelope biogenesis, outer membrane)	1198 (Other metabolic functions)	Ambiguous
AaV_192	179182..180990	603	Putative glycosyltransferase	0438 (Cell envelope biogenesis, outer membrane)	0067 (Other metabolic functions)	NCLDV
AaV_359	341969..342628	220	family 25 Glycosyltransferase	3306(Cell envelope biogenesis, outer membrane)	0068 (Other metabolic functions)	Ambiguous
AaV_235	219361..221037	559	Putative Helicase/ E3 Ubiquitin Ligase	0553 (Transcription/DNA replication,recombination,repair)	0330 (Signal transduction regulation)	NCLDV
AaV_269	249951..252746	932	Putative ATP-dependent RNA helicase		0031 (DNA replication, recombination & repair)	NCLDV
AaV_073	74109..77027	973	Putative DEADDEAh box helicase		0032 (DNA replication, recombination & repair)	Eukaryote
AaV_368	347831..349624	598	Putative Histone acetyl transferase	1243 (Transcription/Chromatin structure and dynamics)		Eukaryote
AaV_075	77645..78022	126	Class I DNA binding Protein	5648 (Chromatin structure and dynamics)	0071 (Miscellaneous)	*Aureococcus* [5]
AaV_097	97924..98979	352	Putative HNH endonuclease		0072 (DNA replication, recombination & repair)	Bacteria
AaV_201	186590..187279	230	Putative Holliday junction resolvase		0278 (DNA replication, recombination & repair)	NCLDV
AaV_282	263555..266908	1118	Putative DNA repairing ATPase	0419 (DNA replication, recombination & repair)	0308 (DNA replication, recombination & repair)	NCLDV
AaV_153	149853..150323	157	Putative Ion channel domain containing protein		1344 (Other metabolic functions)	
AaV_024	23004..29231-3	2076	Putative Concanavalin A-like lectin/glucanase superfamily protein		0107 (Uncharacterized)	Bacteria
AaV_386	369148..370920	591	Putative Concanavalin A-like lectin/glucanase superfamily protein		0108 (Uncharacterized)	Bacteria
AaV_111	113987..114646	220	Class 3 Lipase	3675 (Lipid metabolism)	0225 (Other metabolic functions).	Bacteria
AaV_042	45835..48150	772	Putative Lon protease	0466 (Posttranslational modification, protein turnover, chaperones)	0228 (Other metabolic functions)	Bacteria
AaV_247	233190..234893	568	Putative Capsid protein 2		0022 (Virion structure and morphogenesis)	NCLDV

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_043	48433..49218	262	Small conductance mechanosensetive channel	0668 (Cell envelope biogenesis, outer membrane)		Ambiguous
AaV_281	263017..263565	183	Putative membrane protein			NCLDV
AaV_158	152969..153559	197	Putative metal dependent hydrolase		1120 (Metallopeptidase)	NCLDV
AaV_324	311932..312966	345	Putative type II DNA methyltransferase	2263 (Translation, ribosomal structure and biogenesis)	0234 (Other metabolic functions)	
AaV_287	273849..274556	236	S-adenosyl L-methionine dependent Methyltransferase	2226 (Coenzyme metabolism)	1191 (Other metabolic functions)	Eukaryote
AaV_367	347283..347885	201	Putative RNA methylase	0742 (DNA repilcation, recombination & repair)	0564 (Other metabolic functions)	Eukaryotes
AaV_211	194094..196826	911	mRNA capping Enzyme		1117 (Transcription and RNA processing)	NCLDV
AaV_173	164421..164867	149	Nucleoside diphosphate hydrolase (MutT)	0494 (DNA replication, recombination, & repair / General function prediction only)	0236 (Transcription and RNA processing)	Bacteria
AaV_234	218742..219233	164	Nucleoside diphosphate hydrolase (MutT)	0494 (DNA replication, recombination, & repair / General function prediction only)	0236 (Transcription and RNA processing)	Bacteria
AaV_375	358656..359930	425	Putative pectate lyase	3866 (Carbohydrate transport and metabolism)		Host/Bacteria
AaV_003	4177..5256	364	Putate pectate lyase	3866 (Carbohydrate transport and metabolism)		Host/Bacteria
AaV_038	42485..43642	386	Putative Pectate Lyase	3866 (Carbohydrate metabolism)		Host/Bacteria
AaV_261	243436..244581	382	Putative HD superfamily phosphohydrolase	1078 (General function prediction only)	0603 (Uncharacterized)	NCLDV
AaV_303	288135..290054	640	Oxogluterate/Iron dependent dioxygenase		1166 (Other metabolic functions)	Bacteria
AaV_208	192373..192768	132	Protein Disulfide isomerase	0526 (Posttranslational modification, protein turnover, chaperones)	0317 (Other metabolic functions)	NCLDV
AaV_383	364951..367395	815	fucosylgalactoside 3-alpha-galactosyltransferase		0059 (Uncharacterized)	Eukaryotes
AaV_077	79189..80175	329	Rhomboid family Serine protease	0705 (Amino acid metabolism)		*Aureococcus*
AaV_125	126731..127180	150	Putative ribonuclease H1	0328 (DNA replication, recombination & repair)	1352(DNA replication, recombination & repair)	Bacteria

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_079	81762..83384	541	(Ribonuclease R)	0557 (Transcription)		Ambiguous
AaV_372	355315..356661	449	Putative Phaeophorbide a Oxygenase	2146 (Inorganic ion transport and metabolism)		Eukaryotes
AaV_152	149359..149850	164	C3H2C3-type E3 Ubiquitin ligase/RING-H2 finger protein	5194 (Posttranslational modification, protein turnover, chaperones)	0330 (Signal transduction and regulation)	Eukaryotes
AaV_292	276204.. 7649	482	C3HC4 type E3 Ubiquitin ligase		0330 (Signal transduction)	Eukaryotes
AaV_253	238304..238804	167	RNA polymerase Rpb5, C-terminal domain	2012 (Transcription)	0273 (Transcription and RNA processing)	NCLDV
AaV_076	78130..78954	275	RNA polymerase sigma factor 70	0568 (Transcription)		*Aureococcus* [6]
AaV_241	225244..226005	254	Ribonuclease H	0258 (DNA replication, recombination & repair)		Bacteria
AaV_357	338505..339020	172	SCF ubiquitin ligase	5201 (Posttranslational modification, protein turnover, chaperones)	1299 (Other metabolic functions)	Eukaryotes [2][8]
AaV_058	62257..63033	259	Putative Tetratricopeptide domain containing protein			Host/Bacteria [7]
AaV_381	363515..363985	157	Transcription Elongation factor TFIIS/DNA directed RNA polymerase subunit M	1594 (Transcription)	0272 (Transcription and RNA processing)	Eukaryotes
AaV_203	187801..188745	315	Transcription factor TF IIB	1405 (Transcription)	1127 (Transcription and RNA processing)	NCLDV
AaV_205	189357..190922	522	Hypothetical protein			
AaV_112	114866..115678	271	Type II DNA modification methyl transferase		0234 (Other metabolic functions)	Bacteria
AaV_373	356847..358094	416	UbiA Prenyltransferase	0382 (Coenzyme metabolism)		Ambiguous
AaV_132	131399..132376	326	Ribonucleotide reductase small subunit	0208 (Nucleotide transport and metabolism)	0276 (Nucleotide metabolism)	NCLDV
AaV_255	239312..240016	235	Undecaprenyl pyrophosphate synthase	0020 (Lipid metabolism)		Archaea
AaV_159	153635..154858	408	Putative lambda-type exonuclease		1192 (DNA replication, recombination & repair)	NCLDV
AaV_061	65210..65515	102	Putative WWE domain containing protein			Eukaryote

Table 2.2 Continued.

Gene locus tag	Position in genome (Start..End)	Length (aa)	Annotation	COG category	NCVOG category	Phylogenetic affinity
AaV_376	359986..360270	95	Putative zf-DHHC domain containing protein			
AaV_380	362217..363461	415	Putative Zinc finger domain containing protein	5189 (Transcription / Cell division and chromosome partitioning)	0072 (DNA replication, recombination & repair)	Eukaryote
AaV_086	87814..88797	328	DUF285 domain containing protein			Bacteria
AaV_244	231420..232019	200	Ribonulease HII	0164 (DNA replication, recombination & repair)		Bacteria
AaV_109	109833..113513	1227	Ribonucleoside di phosphate reductase alpha subunit	0209 (Nucleotide transport and metabolism)	1353 (Nucleotide metabolism)	Eukaryote
AaV_242	226063..230052	1330	DNA directed RNA polymerase II largest subunit	0086 (Transcription)	0274 (Transcription and RNA processing)	Eukaryote
AaV_320	304100..307966	1289	DNA directed RNA polymerase largest subunit	0086 (Transcription)	0274 (Transcription and RNA processing)	Eukaryote
AaV_370	350213..353542	1110	DNA directed RNA polymerase second largest subunit	0085 (Transcription)	0271 (Transcription and RNA processing)	NCLDV
AaV_271	253423..253656	78	RNA polymerase subunit Rpb10	1644 (Transcription)	1368 (Transcription and RNA processing)	Ambiguous
AaV_178	167195..168094	300	Serine threonine protein phosphatase	0639 (Signal transduction mechanisms)	0995 (Uncharacterized). Renized by deltablast	Ambiguous
AaV_054	58603..59985	461	Thymidylate synthase	0207 (Nucleotide transport and metabolism)	1136(Nucleotide metabolism)	Ambiguous [8]
AaV_118	120292..121665	458	Translation elongation factor EF-1 alpha	5256 (Translation, ribosomal structure and biogenesis)	0064 (Translation)	Eukaryote
AaV_361	343105..344124	340	Putative VLTF-3 like Transcription Factor		0262 (Transcription and RNA processing)	NCLDV
AaV_130	129177..130508	444	VV_A18 like Helicase	1061 (Transcription/DNA replication,recombination,repair)	0076 (DNA replication, recombination & repair)	NCLDV
AaV_323	309009..311711	901	D5-ATPase-Helicase	3378 (General function prediction only)	0023 (DNA replication, recombination & repair)	NCLDV
AaV_133	133148..132465	684	DUF285 domain containing protein			Bacteria
AaV_072	72876-73979	1104	DUF285 domain containing protein			Bacteria

1. *Aureococcus* mitochondria is among the closest relatives in phylogenetic reconstruction.
2. *Aureococcus* is a close relative along with *Thalassiosira. A cluster of Aav, T. pseudonana and Aureococcus* in the phylogenetic reconstruction.
3. Other picoeukaryotes are monophylatic with this gene.
4. *Aureococcus* seems to be of close in terms of phylogenetic affinity with other picoeukaryotes.
5. Has paraphily with *Micromonas* and other picoeukaryotes.
6. FIrst hit in Blast is to bacteria, but phylogeny reveals origin in the host.
7. *Thalassiosira* is close in terms of phylogenetic affinity.
8. *Aureococcus* is the best hit in Blastp.

Table 2.3: Paralogous genes in AaV clustered in groups.

Paralog group	CDSs	Annotation	Copies
1	AaV_001, 004 - 19, 020- 023, 343, 348 – 356, 374, 212, 264, 277, 309, 30, 325, 028, 029, 382, 047, 052, 053, 157, 149, 169, 168, 086, 133, 072	DUF285 domain containing proteins	50
2	AaV_003, AaV_375, AaV_038	Putative pectate lyase	3
3	AaV_357, AaV_123	SCF ubiquitin ligase	2
4	AaV_243, AaV_151	Hypothetical proteins similar to NCLDVs	2
5	AaV_386, AaV_024, AaV_276, AaV_314, AaV_204	Laminin G domain proteins and derivatives	5
6	AaV_082, AaV_085	Putative CPD photolyases	2
7	AaV_242, AaV_320	DNA directed RNA polymerase II largest subunit	2
8	AaV_221, AaV_083	Hypothetical proteins (Orfans)	2
9	AaV_087, AaV_332	Hypothetical proteins (Orfans)	2
11	AaV_275, AaV_385	Hypothetical proteins (Orfans)	2
12	AaV_160, AaV_161	Hypothetical proteins (Orfans)	2
13	AaV_341, AaV_347	Hypothetical proteins (Orfans)	2
14	AaV_196, AaV_258	Hypothetical proteins (Orfans)	2
15	AaV_312, AaV_317	Hypothetical proteins (Orfans)	2
16	AaV_080, AaV_081	Hypothetical proteins (Orfans)	2
17	AaV_285, AaV_249, AaV_239, AaV_265	Hypothetical proteins similar to NCLDVs	4
18	AaV_114, AaV_140, AaV_067,AaV_143, AaV_183, AaV_025, AaV_340, AaV_363	Hypothetical proteins (Orfans)	8
		Total number of CDSs in paralog groups	**94**

Table 2.4: tRNAs in the AaV genome.

tRNA #	Locus tag	Start	End	Type	Anti-codon
1	AaV_335	322252	322333	Leu	TAA
2	AaV_336	322372	322445	Ile	TAT
3	AaV_337	322625	322707	Ser	TGA
4	AaV_338	322727	322812	Tyr	GTA
5	AaV_339	322836	322906	Gln	TTG
6	AaV_150	148390	148317	Arg	TCT
7	AaV_115	118395	118324	Thr	AGT
8	AaV_095	96105	96031	Thr	TGT

Table 2.5: Genes uniquely present in AaV among the nucleocytoplasmic large DNA viruses (NCLDVs)

Gene ID	**Putative function**	**Probable phylogenetic origin**
AaV_045	Calpain family thiol protease	Eukaryote
AaV_102	Putative Carbohydrate sulfotransferase	Bacteria
AaV_331	CDP-alcohol phosphotidyl transferase	Bacteria
AaV_078	Unsaturated Glucuronyl hydrolase	*Aureococcus* (Host)
AaV_003, AaV_375, AaV_038(Paralogs)	Putative pectate lyase	Bacteria/*Aureococcus**
AaV_144	AIM24 domain containing protein	Ambiguous
AaV_093	Putative cytosine deaminase	Bacteria
AaV_077	Rhomboid family serine protease	*Aureococcus* (Host)
AaV_110	Eukaryotic translation initiation factor 5A	*Aureococcus* (Host)
AaV_373	UbiA prenyl transferase	Ambiguous
AaV_255	Undecaprenyl pyrophosphate synthase	Archaea
AaV_061	WWE domain containing protein	Eukaryote
AaV_027	Hypothetical protein	Ambiguous
AaV_220	Putative DUF285 domain containing E3 ubiquitin ligase	Bacteria
AaV_043	Small conductance mechanosensitive channel	Ambiguous

Table 2.6: Putative AaV CDSs having homologs only in the NCLDVs (NCLDV specific ORFans). Genes present inside the central region (123.6 Kbp – 247.3 Kbp) are denoted in the final column.

Gene ID	Start	End	NCLDV families with homologs	Gene inside the central region?
			Megaviridae only	
AaV_070	71580	72029		No
AaV_074	77080	77388		No
AaV_172	164176	164424		Yes
AaV_209	192802	193404		Yes
AaV_232	216327	217844		Yes
AaV_250	236557	237018		Yes
			Phycodnaviridae only	
AaV_062	65628	66191		No
AaV_126	127225	127446		Yes
AaV_134	133230	133919		Yes
AaV_238	222846	223472		Yes
AaV_239	223635	224891		Yes
AaV_243	230431	231069		Yes
AaV_246	232167	233117		Yes
AaV_249	235356	236498		Yes
AaV_265	247052	248308		Yes
AaV_281	263017	263565		No
AaV_285	267923	269080		No
AaV_330	319354	320202		No
			Both Phycodnaviridae and Megaviridae	
AaV_116	118429	119058		No
AaV_151	148486	149202		Yes
AaV_158	152969	153559		Yes
AaV_175	165262	166164		Yes
AaV_179	168119	168661		Yes
AaV_186	174276	174677		Yes
AaV_201	186590	187279		Yes
AaV_202	187303	187746		Yes
AaV_214	198480	199118		Yes
AaV_328	315300	315833		No

CHAPTER III

The transcriptomic landscape of *Aureococcus* virocell

Publication note:

This chapter is a version of an article in preparation for submission to *mBio* by Mohammad Moniruzzaman, Eric R. Gann, Stephen P. Dearth, Shawn R. Campagna and Steven W. Wilhelm.

My contribution to this work was experimental design, sample collection and processing, transcriptomic data analysis and most of the writing of the manuscript.

Abstract

While *Aureococcus anophagefferens* virus (AaV) plays a critical role in regulating the fate of Brown tide blooms, the molecular aspects of the infection process remain unknown. We studied the transcriptomic response of the *Aureococcus anophagefferens CCMP1984* cultures that were challenged with AaV over an entire infection cycle. A massive transcriptional reprogramming of the host was evident as early as 5 minutes post-infection, with modulation of specific processes likely related to both host defense mechanism and viral takeover of the cell. Infected *Aureococcus* showed a relative suppression of host-cell transcripts associated with photosynthesis, cytoskeleton formation, fatty acid and carbohydrate biosynthesis. In contrast, host cell processes related to protein synthesis, polyamine biosynthesis, cellular respiration, transcription and RNA processing were overrepresented compared to the healthy cultures at different stages of the infection cycle. Interestingly, a large number of selenoproteins were overexpressed during infection. Overexpression of the redox active selenoproteins also indicated that viral replication and assembly were possibly under progress in a highly oxidative environment. Out of 384 AaV genes, we detected expression of 381. These genes demonstrated a clear temporal-expression pattern, with increasing relative expression observed for majority of the genes over time. We also detected a putative early promoter motif for AaV, which was highly similar to the early promoter elements of two other Mimiviridae members, indicating some degree of evolutionary conservation of this promoter element within this clade. This study provides first series of insights into the *Aureococcus* 'virocell', and paves the way to generate compelling hypothesis regarding metabolic and regulatory processes which play critical roles in virus propagation.

Introduction

In the formative paper on 'viral shunt' (1999), Wilhelm and Suttle pointed to the critical role of viruses in the microbial loop. Viruses lyse and directly release cellular organic and inorganic nutrients back to the environment, where some become available for microbial growth, while othjer organic material may be exported to the deep ocean (Wilhelm & Suttle, 1999). With an estimated 10^{31} virus particles in the sea (Angly *et al.*, 2005), the geographical scale and impact of these processes are enormous – viral activity can turn over an estimate of 150 gigatons of carbon per year (Suttle, 2007). This role of viruses in the global biogeochemical cycle raise the inevitable question – "How do they do it?" Indeed, the answer lies within the molecular details of the infection process. It is historically thought that viruses encode the minimal amount of genomic information necessary in order to instruct host cells to produce new viruses. Using almost entirely the host machineries, hundreds of virus particles can be produced from one host cell. For example, Hepatitis B virus encodes only four overlapping genes in a 3.2 kb genome (Liang, 2009), whereas a large number of *Picornavirales* members, which are widespread in the ocean, only code for one or two proteins (Lang *et al.*, 2009).

This paradigm was challenged by discovery of 'giant' eukaryotic viruses – viruses that rival even bacterial cells in terms of their physical size and genomic content (Raoult *et al.*, 2004, Moniruzzaman *et al.*, 2014, Wilhelm *et al.*, 2016). It was eventually revealed that a major portion of the genomic content of these giant viruses was acquired from the eukaryotic hosts and other sources through horizontal gene transfer (HGT), some of which are passed vertically through the course of viral evolution (Filee *et al.*, 2007, Koonin & Yutin, 2010). Together, this genomic content can potentially make these large viruses largely autonomous of the host cell, and can even empower them to control individual processes in the complex eukaryotic cells to produce virus specific macromolecules (Claverie & Abergel, 2010).

Despite our knowledge concerning the genomic potential of these viruses, the molecular underpinnings of conversion from a healthy host cell into a 'virocell' (Forterre, 2013) is largely unknown. Of particular interest are the eukaryotic algae and their giant viruses, owing to the role of phytoplanktons in global carbon cycling (Li, 1994, Worden *et al.*, 2004) and the formation of rapid and sometimes devastating algal blooms (Anderson *et al.*, 2011). Critical insights have been obtained regarding the modulation of cellular processes of *Emiliania huxleyi* – the most abundant coccolithopore alga in the world's ocean – upon infection by its virus, EhV (Vardi *et al.*, 2012). This includes the production of viral glycosphingolipids (vGSLs) through a sophisticated virus-mediated regulation of host lipid biosynthesis and modulation of host redox state (Vardi *et al.*, 2009, Rosenwasser *et al.*, 2016). In parallel, transcriptomic analyses of *Chlorella variabilis* infected by a giant virus, PBCV-1, has helped elucidate

host-processes during the early phase of infection (Rowe *et al.*, 2014). However, giant viruses infecting eukaryotic algae are enormously diverse, infecting a potentially wide range of photosynthetic hosts (*e.g.*, Moniruzzaman *et al.* (2016)). And while they do share a few core proteins, there are a large number of genes that vary from one virus to another. As a consequence, significant differences in the molecular basis of interactions can be expected between different eukaryotic host-virus pairs.

In this study, we employed transcriptomics to resolve the molecular response of *Aureococcus anophagefferens* to infection by AaV. Prior research has resolved the genomic complement of both AaV and *Aureococcus* (Gobler *et al.*, 2011, Moniruzzaman *et al.*, 2014). However, no information is available on the progressive changes in the molecular processes of *Aureococcus* cells upon infection, which might provide critical information on the metabolic pathways and cellular components that can impact virus production. Additionally, the possible roles of the large number of genes that AaV acquired from host, other organisms and its NCLDV ancestor remain to be elucidated. Our experimental design examined different stages of the AaV infection cycle to capture the dynamic nature of host cellular response and viral transcriptional landscape. This study provides first insights into the molecular interaction between an environmentally disruptive harmful alga and a giant virus.

Materials and Methods

Experimental setup

Aureococcus anophagefferens culture was maintained in modified L-1 medium (Guillard, 2003) at an irradiation level of 100 *microeinsteins* m^{-2} s^{-1} at a 14:10 (h) light-dark cycle. Prior to the experiment, *Aureococcus* cultures were grown to a mid-log phase concentration of ~9 x 10^5 cells/ml. Five biological replicates (2.0 L) of *Aureococcus* cells were started within two hours of the onset of the light cycle. The cultures were inoculated with AaV at a multiplicity of infection (MOI) of ~18. The MOI was determined by counting DNA containing virus particles through fluorescence microscopy, since a plaque assay is not available for estimating the infectious AaV titer. For each of the biological replicates, controls inoculated with the same volume of heat-killed viral lysate was also established. The heat killed lysate was generated by microwaving for 5 minutes. Samples were collected at 5 min, 30 min, 1 h, 6 h, 12 h and 21 h after inoculation. For RNA extraction, 250 ml samples were filtered through 0.8-µM pore-size ATTP filters (EMD Millipore, Darmstadt, Germany) and immediately flash frozen in liquid nitrogen and stored at -80° C until further processing. Unfiltered samples (for cell enumeration) and samples passed through 0.45-µM polyvinylidene fluoride syringe filters (Merck, Darmstadt, Germany) for free virus count were preserved in 0.5% glutaraldehyde at -80° C from each sample at each time point.

Cell and free virus densities

Aureococcus cells were enumerated using a GUAVA-HT6 flow cytometer (EMD Millipore, Darmstadt, Germany) gated on the red chlorophyll fluorescence. Free virus particle densities from each time point was determined following Ortmann and Suttle, 2009. Samples were thawed at room temperature and diluted 100 times using L-1 medium prior counting. The diluted samples were collected on 25-mm diameter Whatman Anodisc (Sigma-Aldrich, St. Louis, MO, USA) inorganic membrane filters having a nominal pore-size of 0.02 µM. The filters were allowed to air-dry for 15 mins following incubation with 15 µL of 4,000X diluted Syber Green (Lonza, Rockland, ME, USA). The filters were then fixed using an anti-fade solution (50:50 PBS/glycerol and 0.1% p-Phenylenediamine) (Noble & Fuhrman, 1998). Slides were observed through a Leica DM5500 B microscope at 1000X magnification with a L5 filter cube (excitation filter: 480/40, suppression filter: BP 527/30) (Leica Microsystems CMS GmbH, Hesse, Germany). For each sample, 20 random fields (1 µM X 1 µM) were enumerated and averaged. The following formula was used to estimate the VLPs/ml in each sample:

$$VLPs/ml = V_f * \frac{A_a}{A_g * V_f} * D$$

Where, V_f = average virus count/field, A_a = total filterable area of Anodisc (excluding the O-ring), A_g = Area of eyepiece grid, V_f= volume filtered (mL), D = dilution factor.

RNA extraction and sequencing

Three of the 5 biological replicate experiments were used for RNA extraction and analysis for this study. The other two samples were preserved for metabolome extraction, results of which will be described elsewhere. RNA was extracted with MO BIO Powerwater RNA isolation kit (MO BIO Laboratories, Carlsbad, CA, USA) following a 2 min bead beating step using Lysing Matrix E 2 mL tubes (MP Biomedicals, Santa Ana, CA, USA). The manufacturer's protocol was followed with slight modification. Specifically, the DNAse treatment step was performed twice to ensure sufficient purity of the RNA. RNA was quantified using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific, Waltham, MA, USA) and RNA integrity was checked with an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). Extracted RNA was sequenced in the HudsonAlpha Genomic Services Lab (Huntsville, AL, USA). The RNA samples were poly-A selected to enrich for mRNA. Samples were sequenced using a Illumina® NextSeq® sequencer at a target depth of ~25 million reads per sample and 76 bp read length. Standard protocols by Illumina® were followed for library preparation, poly dT bead selection and sequencing.

Bioinformatics analysis

Sequencing reads were trimmed in CLC Genomics Workbench 9.0 (Qiagen, Hilden, Germany). Reads with a quality score cut-off of 0.3 or with ambiguous bases ('N's) were discarded. Reads passing quality control were mapped to the *Aureococcus* genome sequence (NCBI. Accession no ACJI00000000) with stringent mapping criteria (95% similarity, 70% length matching). Differential expression of genes in the virus-treated samples compared to the controls was determined at each time point using edgeR program implemented in the CLC Genomics Workbench 9.0. P-values were adjusted for False Discovery Rate (FDR) using Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995). Reads were also mapped to the AaV genome using the same parameters. The number of reads mapped to each AaV gene was rarefied by library size. Values from biological replicates at each time point were averaged prior to hierarchical clustering. Data was visualized using a heatmap of AaV gene expression in PRIMER 7.0 statistical analysis program (Clarke & Gorley, 2015).

The enrichment of genes within the framework of Gene Ontology (GO) terms (positive or negative fold changes) was determined using BiNGO (Maere et al., 2005). Only genes showing an absolute fold change of ≥1.5 were used for this analysis. The distribution of reads within KEGG pathways was determined using a z-test as implemented in 'GAGE' R package (Luo et al., 2009). This analysis employed input from all the genes, irrespective of fold-change level or statistical significance, and looked for coordinated expression changes in a particular pathway. The resulting P-values for both the analyses were adjusted for false discovery rate (FDR) using Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995). We considered a FDR p-value cut-off of 0.1 for the GO analysis and 0.15 for the pathway enrichment analysis as significant

Results

Cell growth and virus infection dynamics

Cultures inoculated with heat-killed lysates displayed the growth pattern of a healthy *Aureococcus* culture, reaching a cell density of approximately ~1.1 x10^6 cells/ml by 24 h (Figure 3.1). In contrast, the virus infected cultures didn't show any significant increase in cell density over the course of infection, indicating that a proportion of cells were infected during the first cycle of virus propagation (Figure 3.1). Consistent with previous studies, free-virus titer increased around 24 h after infection and steadily increased up to ~3.5 x 10^7 VLPs/mL by 30 h post infection (Figure 3.1). Complete lysis of the culture usually takes 48-72 hours during routine virus production in lab, encompassing 2-3 infection cycles (Rowe *et al.*, 2008).

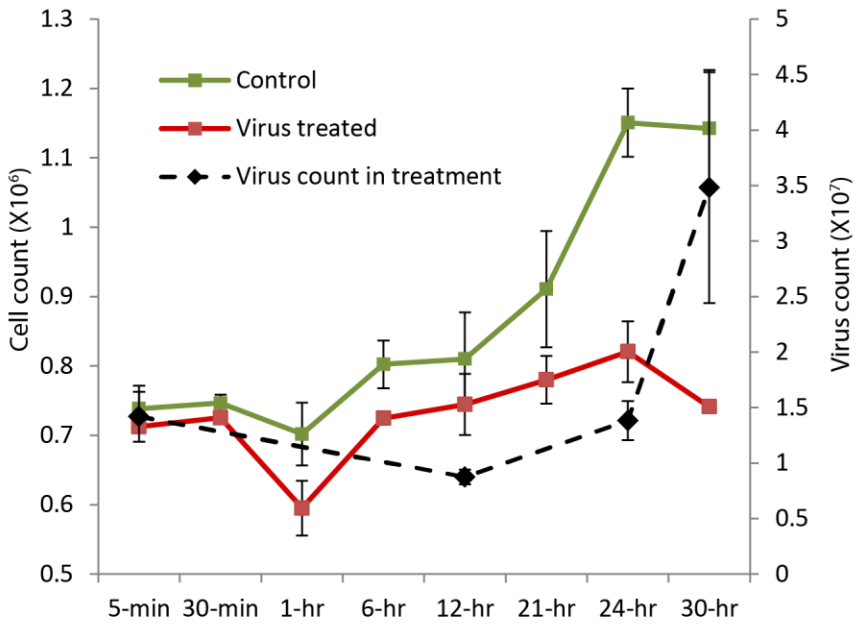


Figure 3.1: *Aureococcus* and AaV count over the course of infection. Cell counts are average of three biological replicates, while virus counts are average of two biological replicates. Green – cell count in control cultures, red – cell counts in infected cultures, black – virus counts.

RNA-seq output and read mapping statistics

After quality trimming, between ~18.2 and 29.4 million reads were obtained from the 36 samples. In the control samples, > 80% of the reads could be mapped to the host nuclear, chloroplast and mitochondrial genomes (Figure 3.2). In the virus-treated samples, the proportion of virus transcripts steadily increased over time (Figure 3.2, 3.3). As a result, the fraction of reads mapped to the host genome slightly decreased in the virus-treated samples at the later time points (Figure 3.2). About ~20% of the reads from all the samples could not be aligned to the host or viral genomes, which likely originated as a result of contamination during library preparation and sequencing.

Gene expression dynamics of AaV

Transcripts from 116 viral genes were present in the infected culture as early as 5 min post-infection (Figure 3.4, 3.5). While only ~0.007% reads could be mapped to viral genome from sequence libraries at 5 min, ~15% of the reads originated from viral transcripts by the 21 h time point (Figure 3.3). To resolve temporal patterns of virus gene expression, we performed a hierarchical clustering using the average number rarefied reads per library that mapped to viral genes over the time course. Clear temporal patterns in gene expression were observed, with some genes expressed either immediately or within one hour of infection, while reads from other genes appeared late into the infection (Figure 3.5). The temporal clustering provided an opportunity to resolve promoter sequences associated with early and late gene expression. We grouped the genes first expressed within 5 minutes to 6 hours as 'early to intermediate' class, and the genes expressed within 12-21 hours as 'late' class. MEME (Bailey *et al.*, 2009) was used in the discriminative mode to detect motifs enriched in the 'early to intermediate' set of genes compared to the 'late' class. A motif with the general pattern "[AT][AT][AT][TA]AAAAATGAT[ATG][AG][AC]AAA[AT]" was found in the first class of genes with an E-value of 2.1e-151 (Figure 3.6). This motif encompasses the octamer "AAAAATGA". When we searched for the AaV specific octamer motif, we found 47.5% (127 genes) of the 'early to intermediate' class genes contain this motif with exact match on their upstream, while only 22% (24 genes) of the 'genes in the 'late' category harbored it in the upstream regions. This evidence strongly indicates that the motif detected by MEME likely harbors the early promoter in AaV. A search for late promoter motif in the second set of sequence resulted in a highly degenerate motif with a large E-value (3.53e+011) without any match to the previously reported late promoter motifs of giant viruses (Figure 3.7).

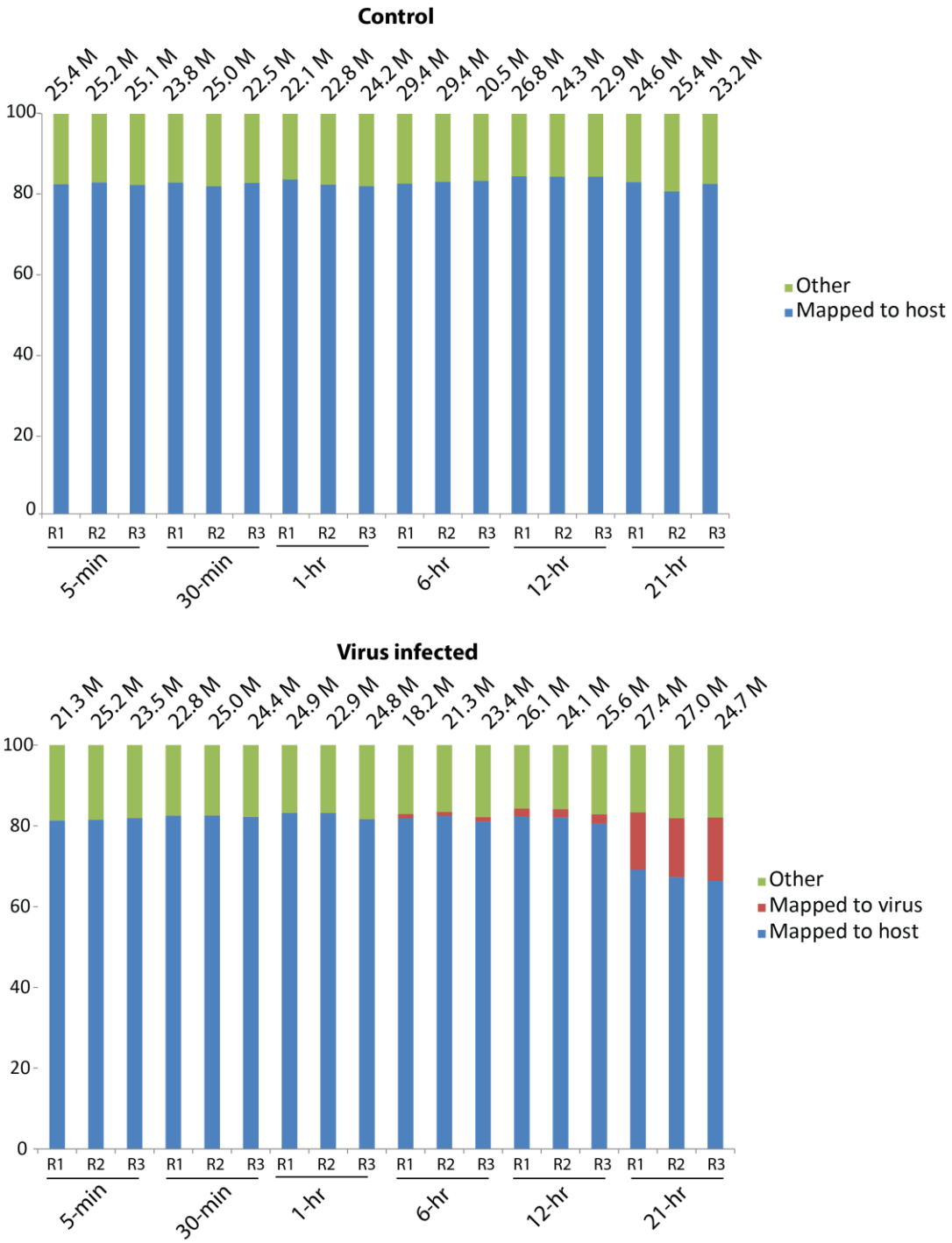


Figure 3.2: Percentage of reads mapped from each sequence library on host and virus genomes. "Other" indicates the fraction of reads that could not be mapped at a 95% sequence identity level. Total number of reads (in millions) for each library are mentioned on the respective bar graphs. Biological replicates from each time point are indicated as R1, R2 and R3.

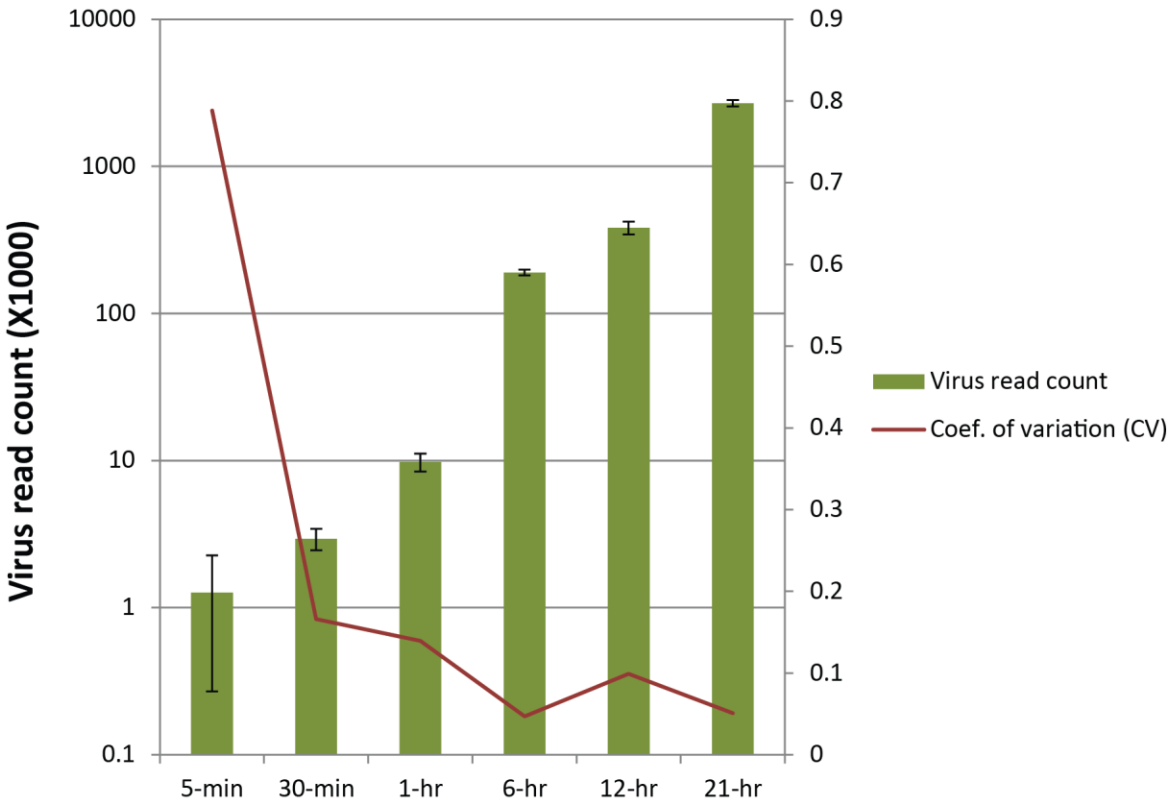


Figure 3.3: Average number of reads mapped from the virus treated samples to AaV genome over time. The read counts were rarefied by library size prior averaging. The coefficient of variation for each time point is plotted on the secondary axis.

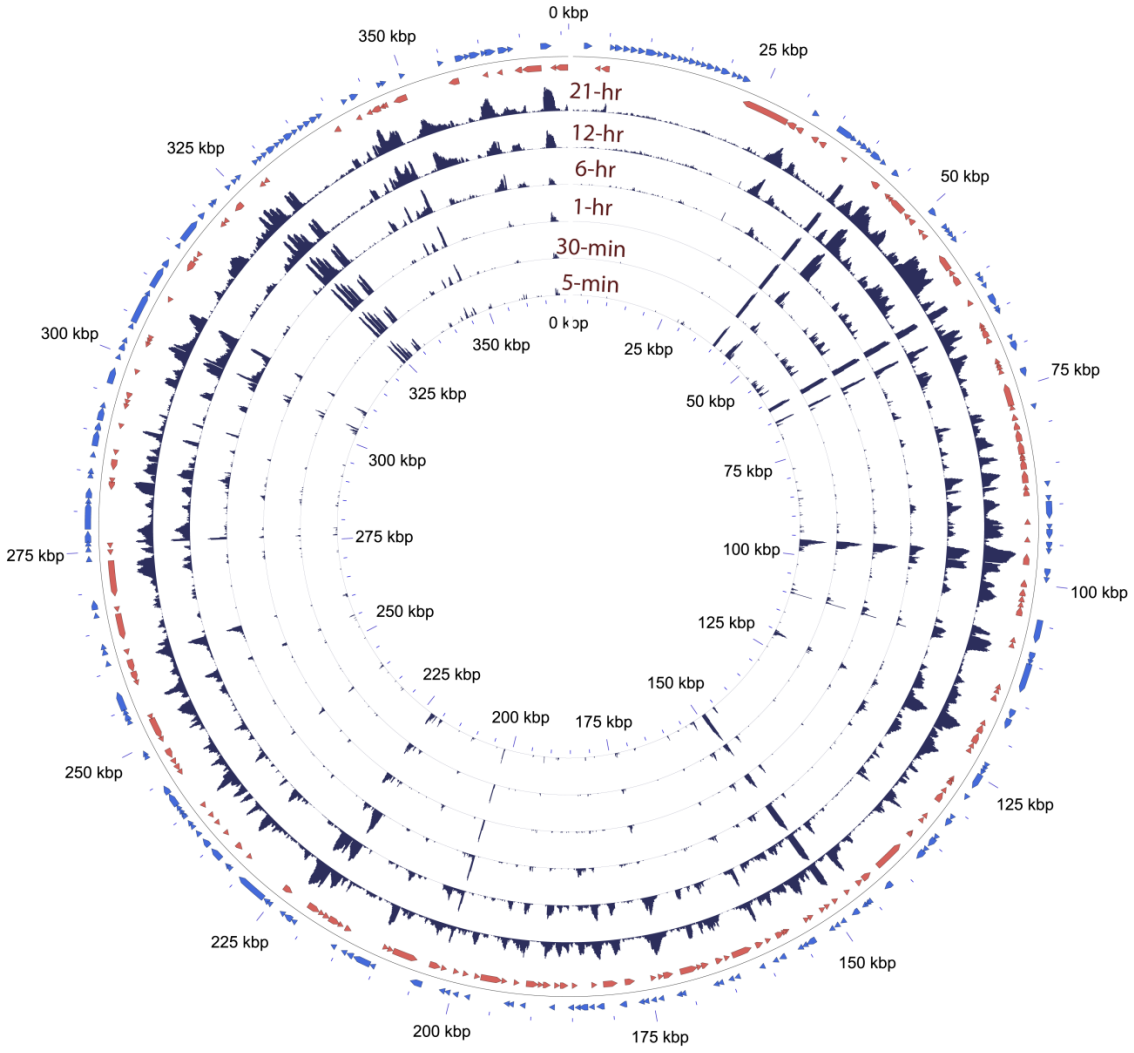


Figure 3.4: Trend in expression of individual AaV genes over time. The read mapping from each timepoint was converted to coverage graphs with a 100bp sliding window. Three replicates from each time points were averaged after rarefaction by library size. The two outermost rings represent the forward (blue) and reverse (red) coding sequences.

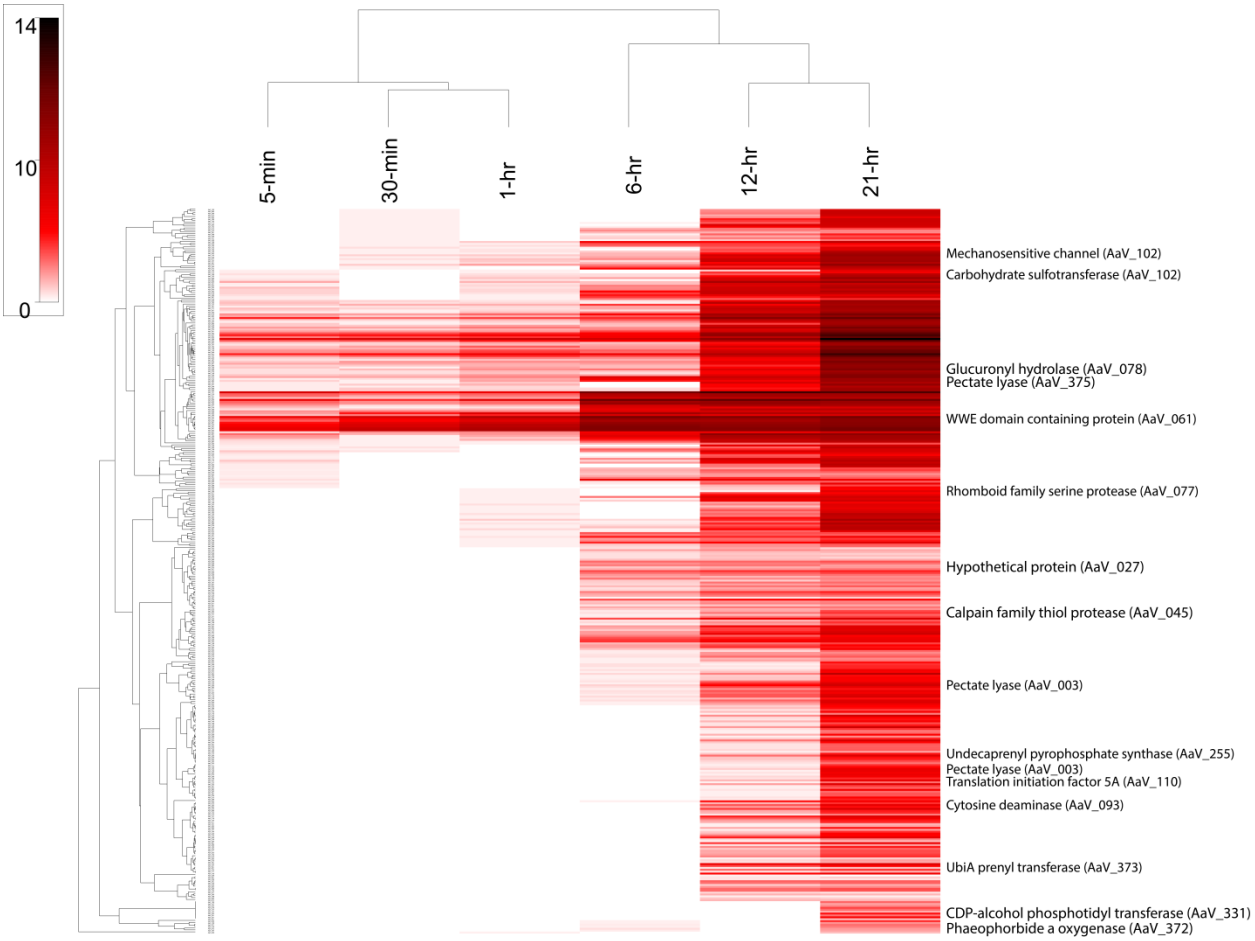


Figure 3.5: Heatmap showing temporal pattern in AaV gene expression. The read mapping data over individual genes was hierarchically clustered after 4th root transformation. Expression pattern of genes unique to AaV (acquired by HGT) are marked on the right panel.

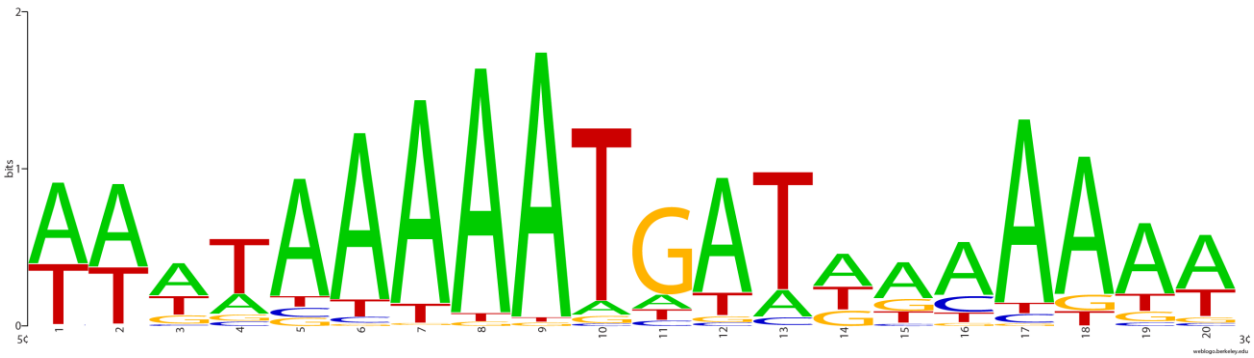


Figure 3.6: Sequence logo of putative early promoter motif of AaV.

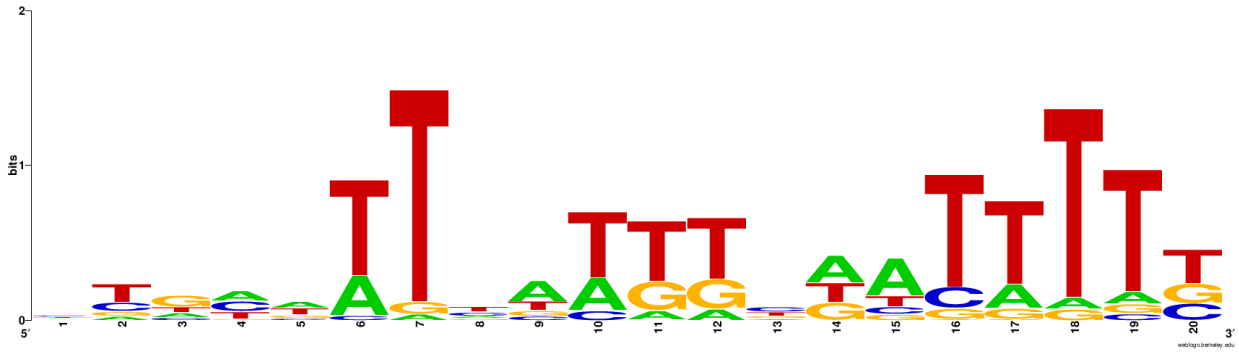


Figure 3.7: Sequence logo of the motif overrepresented in the ‘late’ class of genes compared to the ‘early to intermediate’ gene class in AaV.

Genome-wide read mapping revealed that the expression of different genes had a large spatiotemporal variation (Figure 3.4). Only three of the annotated genes from the AaV genome (AaV_004 & AaV_115 – hypothetical proteins and AaV_336 – leucyl tRNA) were not detected. The two terminal DUF285 domain rich regions showed lower variation in expression values compared to other viral genes (Figure 3.4). While relative expression of most of the viral genes varied over several orders of magnitude during the time course, expression of DUF285 regions stayed within one order of magnitude. As a striking contrast, expression of major capsid protein was found to be dramatically high at 21 hour – encompassing more than 50% of the virus specific reads and ~6% of the entire libraries at that time point.

There are 137 genes from AaV which have NCVOG (Yutin *et al.*, 2009) and/or COG (Tatusov *et al.*, 2000) assignments, giving insights into their potential function (Appendix, Table 3.2). Based on the cluster analysis, we found 7 of 9 annotated viral-methyltransferases to be expressed within the 'early to intermediate' timeframe – anywhere from 5 min to 6 h. Three of 4 genes with NCVOG category 'virion structure and morphogenesis' (AaV_165, 247 and 290) were expressed late – consistent with previous observations that genes involved in virus structural components are usually expressed late during infection (Fischer *et al.*, 2010, Legendre *et al.*, 2010). The exception was Major capsid protein (AaV_096), the major structural component of the virus, which was found to be expressed immediately (5 min) after infection. Three ubiquitin ligases (AaV_228, 235 and 298) and two proteases (AaV_042, 066) were also found to be expressed 5 min post-infection, alongside a number of putative transcription factors (Appendix, Table 3.2). As discussed in chapter II, AaV has a number of genes unique among NCLDVs, which are putatively acquired by horizontal gene transfer. Among these, three carbohydrate metabolism genes (carbohydrate sulfotransferase: AaV_102, glucuronyl hydrolase: AaV_078 and pectate lyase:AaV_375) were expressed immediately after infection (Figure 3.5). However, the majority of HGT acquired genes were found to be expressed only at or after 6 h (Figure 3.5). Most of the genes with COG or NCVOG classifications did not follow a specific temporal pattern, and were distributed in both 'early to intermediate' or 'late' class.

Global transcriptional remodeling of the virus infected host

Viral infection induced a dramatic and rapid reprogramming of the host cell, which was reflected in the number of *Aureococcus* genes that were differentially expressed compared to the uninfected culture (Figure 3.8). Even at 5-min post infection, we observed 13.36% of the 11,570 genes of *Aureococcus* were differentially expressed, with 412 genes having fold changes of > 1.5 and 588 genes with a fold change < -1.5 (FDR $p<0.05$) (Figure 3.8). With exception of the 1 h time point, the number of genes over or underrepresented compared to control showed a tendency to increase over time, with the highest number of genes observed to be differentially expressed occurring at the 12 h time point (42.86%).

The number of differentially expressed host genes was dramatically reduced at 1-h compared to other time points: only 82 genes were found to be differentially expressed (Figure 3.8). All of these differentially expressed genes showed negative fold-change. Since the general trend of increasing number of differentially expressed genes over time did not apply to 1 h time point, it presented an 'anomaly' that invited further investigation. There were ~9,800 virus reads on average in the three virus-treated biological replicates from time-point 1 hour (Figure 3.3), which fits well within the trend of increasing viral reads over time. This diminished the possibility of sample mislabeling or other potential sources of human errors. An nMDS analysis coupled with a hierarchical clustering using Bray-Curtis similarity showed that control and infected samples from 1-hr time-point clustered together and showed > 97.5% similarity between the replicates (Figure 3.9), whereas control and treatment suggest that reduced instances of changes in expression in the1-h samples had a biological basis.

A large number of GO categories and KEGG pathways were differentially represented across time points, except for at 1 h (Figure 3.10). Notable GO categories that were downregulated in response to virus infection included actin, microtubule motor activity, photosynthesis, processes related to fatty acid biosynthesis, carbohydrate biosynthesis and transmembrane transport. Overrepresented GO categories included mitochondrial membrane, transcription, regulation of oxidoreductase, a wide array of methyltransferase activity and protein synthesis related processes (Figure 3.10). Among the overrepresented KEGG pathways were ribosome, glutathione and sulfur metabolism, nucleotide excision repair and beta-alanine metabolism. Underrepresented KEGG pathways included endocytosis, glycosphingolipid biosynthesis, fatty acid biosynthesis, galactose metabolism, autophagy regulation and phagosome (Figure 3.10). Both GO and pathway analyses found carbohydrate and lipid biosynthesis were underrepresented, while protein synthesis was detected to be overrepresented by both approaches. Even though only a few genes were differentially expressed at 1 hour, pathway analysis detected ribosome, butanoate and sulfur metabolism to be overexpressed at this time point (Figure 3.10).

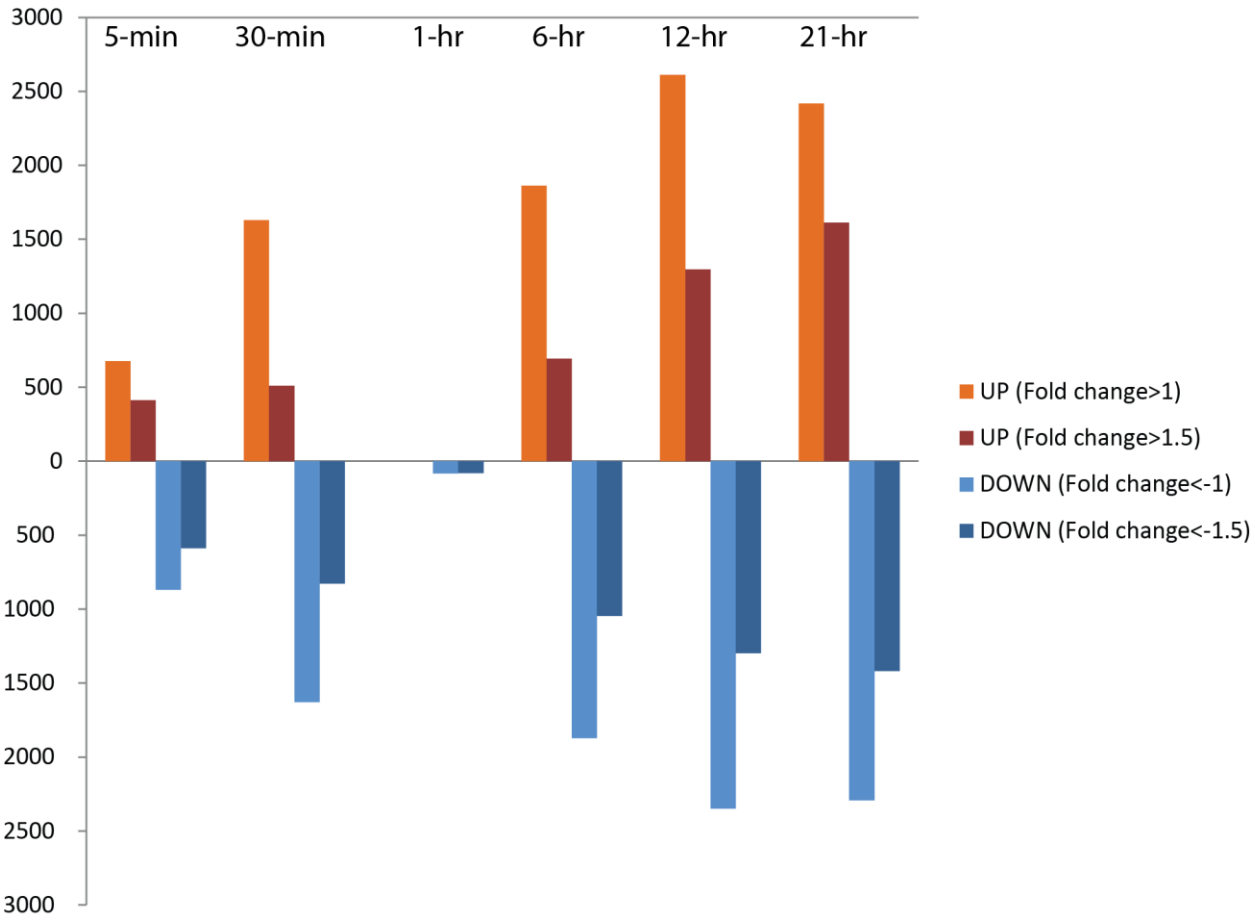


Figure 3.8: Total number of genes up- and downregulated in the host across different time points compared to the healthy cultures. Number of upregulated genes are on positive Y-axis, while number of downregulated genes are represented on the negative Y-axis.

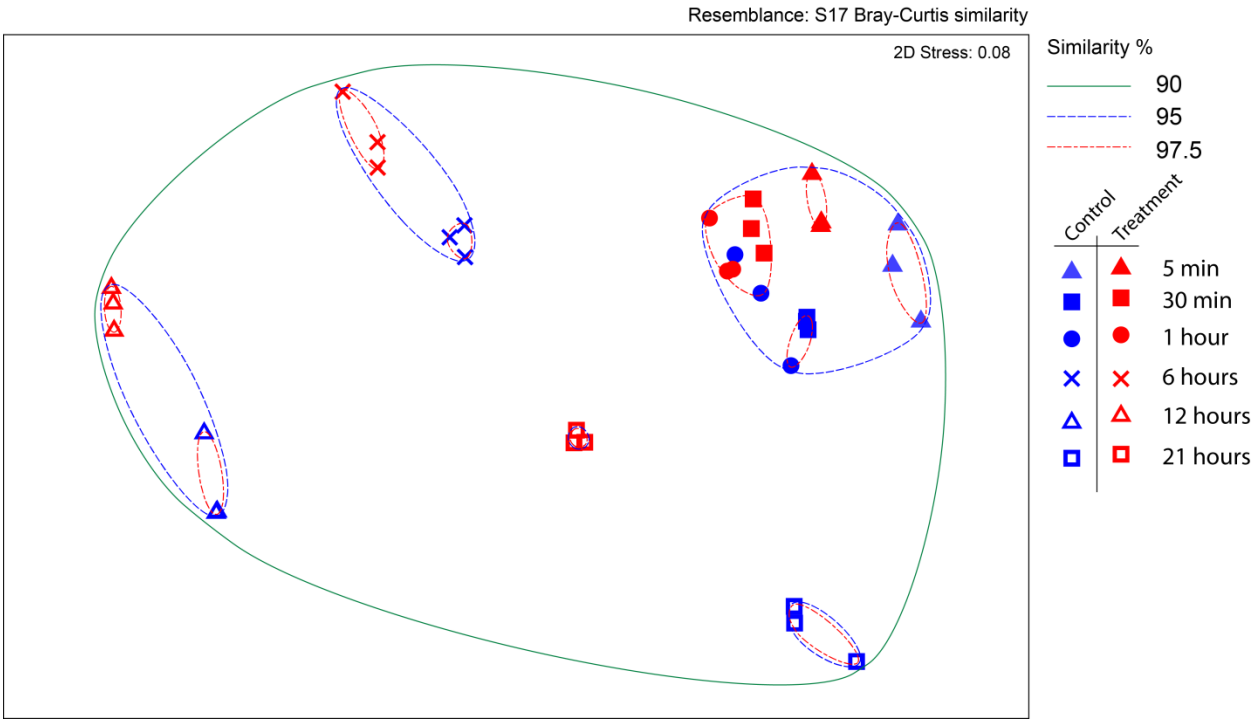


Figure 3.9: nMDS plot of the TMM normalized gene expression data from all replicate samples. The circles drawn around the samples denote different level of similarity (90, 95 and 97.5%) among the samples obtained from hierarchical clustering of the samples using Bray-Curtis dissimilarity metric.

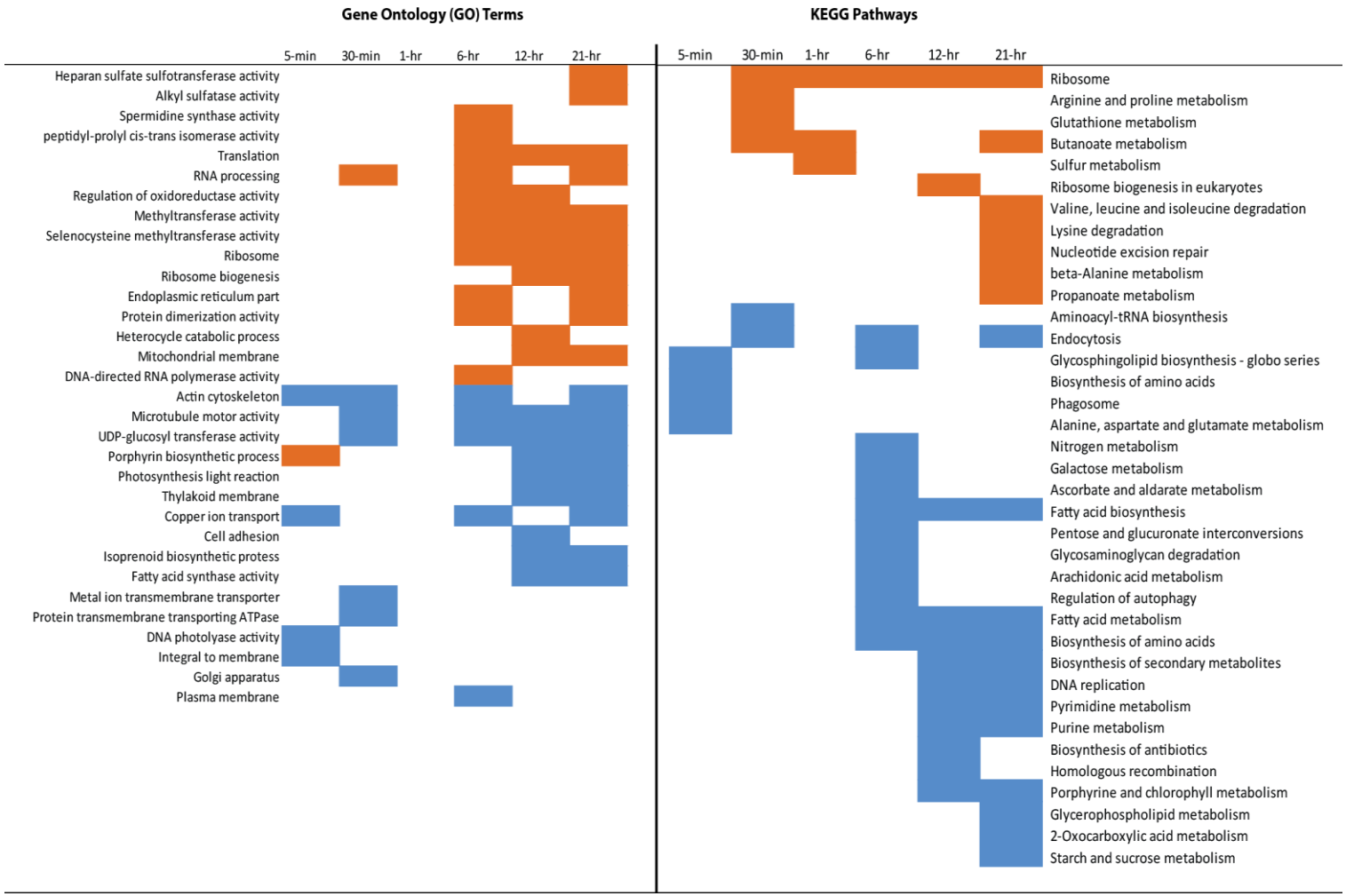


Figure 3.10: Notable GO categories and KEGG pathways over- or underrepresented in the virus treated samples over different time points. The overrepresented terms/pathways are shown in orange, while the underrepresented ones are as blue rectangles.

Downregulation of photosynthesis related processes upon virus infection

We observed significant expression reduction in ~20% of the 62 light harvesting complex (LHC) genes in *Aureococcus* as early as 5 min after infection, indicating that the light harvesting capacity of the infected cells decreased compared to the healthy culture (Figure 3.11). The number of LHC genes showing negative fold-change compared to control increased through the infection time course – reaching >60% by 30 min and 6 h. By 12 and 21 h, almost all the 62 LHC genes showed significant negative expression, with the exception of 6 genes at 12 h and 2 genes at 21 h that were proportionally increased (Figure 3.11). Chloroplast genes encoding proteins for photosystem I and II during the infection (*psaA, psaB, psaL, psbA, psbC and psbD*) were downregulated by 5 min post-infection (Table 3.1). However, some of the components of the photosystem II, namely *psbA, psbD, psbB, psbD and psbK* were found to be overrepresented in the later stages of infection (12 and 21 hours) compared to controls. Taken together, the data indicate that photosynthetic capacity of the virus infected cells decreased during the early and intermediate stage of infection –throughout the light cycle.

While LHC proteins and photosystem structural proteins were generally underexpressed, we found porphyrin biosynthesis related genes to be overrepresented immediately after infection (Figure 3.10). Although a number of steps in the chlorophyll biosynthesis pathways were overrepresented in the virus infected culture leading to increased porphyrin derivatives, the step leading to chlorophyll synthesis from chlorophyllide-a had no significant change compared to control. This indicated that porphyrin derivatives likely accumulated early in the infected cell, leading to photooxidative damage of different cellular components, including chloroplasts (Reinbothe *et al.*, 1996). In contrast, biosynthesis of porphyrin was suppressed in the infected cells during the late phase of infection (Figure 3.10). Additionally, DNA photolyase activity was found to be downregulated 5 min post-infection (Figure 3.10) – a process critical in repairing UV-mediated formation of pyrimidine dimers in DNA (Thiagarajan *et al.*, 2011).

Figure 3.11: The expression pattern of all the light harvesting complex (LHC) protein encoding genes in virus treated *Aureococcus* cultures compared to the control. Fold changes are Log2 converted. Positive fold changes are in red, while negative fold changes are shown in blue. Non-significant (FDR p>0.1) fold changes are shown in gray.



Figure 3.11. Continued.

Table 3.1: Over and underexpressed genes encoding components of photosystem I and II across the time course compared to uninfected control. Fold changes (FC) and corresponding FDR p values are shown. **Significant (FDR p<0.1) positive (orange) and negative (green) fold changes are marked.

	5-min		**30-min**		**1-hr**		**6-hr**		**12-hr**		**21-hr**	
Gene	*FC*	*FDR p*	*FC*	*FDR p*	*FC*	*FDR p*	*FC*	*FDR p*	*FC*	*FDR p*	*FC*	*FDR p*
psaA	**-1.84**	0.00	**-1.58**	0.00	**-1.54**	0.05	**-1.89**	0.00	**-1.43**	0.00	**-1.74**	0.00
psaB	**-1.74**	0.00	**-1.30**	0.05	-1.47	0.65	**-1.28**	0.11	1.03	1.00	-1.20	0.59
psaC	-1.06	1.00	-1.91	0.30	-1.80	1.00	1.25	0.74	1.62	0.36	**-2.21**	0.06
psaD	-1.41	0.88	-1.18	1.00	-1.83	0.86	1.16	0.96	1.28	0.51	-1.20	0.97
psaJ	1.64	1.00	-2.63	0.74	-6.73	1.00	-1.73	0.96	-1.82	1.00	2.31	0.85
psaL	**-2.34**	0.00	**-1.50**	0.09	-1.38	1.00	**-1.82**	0.00	-1.14	0.41	-1.06	0.99
psaM	1.00	1.00	-2.81	0.91	-6.72	1.00	1.07	1.00	-1.82	1.00	1.00	1.00
psbA	**-1.31**	0.09	**-1.46**	0.00	-1.56	0.17	-1.07	0.79	1.57	0.00	1.51	0.04
psbB	1.11	0.88	1.68	0.00	-1.02	1.00	1.52	0.00	1.37	0.01	-1.04	0.90
psbC	**-1.61**	0.00	**-1.26**	0.08	-1.65	0.11	-1.06	0.93	1.07	0.83	1.08	0.93
psbD	**-1.85**	0.00	**-1.60**	0.00	-1.56	0.36	**-1.58**	0.00	1.24	0.06	1.29	0.06
psbE	-1.37	0.88	-1.11	1.00	-1.21	1.00	-1.40	0.44	1.13	1.00	-1.16	1.00
psbF	-1.33	1.00	-2.63	0.75	-1.14	1.00	-1.50	0.90	2.88	0.22	-2.42	0.70
psbH	-1.57	0.96	**-3.14**	0.02	-1.78	1.00	-1.23	1.00	-1.07	1.00	-1.55	0.45
psbI	1.25	0.46	1.73	0.09	-1.37	1.00	**-1.36**	0.10	-1.23	0.40	-1.14	0.77
psbJ	-1.31	1.00	1.23	1.00	-2.08	1.00	-1.11	1.00	2.13	0.70	2.31	0.85
psbK	1.16	1.00	-1.17	1.00	-1.56	1.00	-1.04	1.00	1.03	1.00	4.47	0.07
psbL	-1.38	1.00	-1.16	1.00	-3.05	1.00	-1.58	0.83	2.93	0.34	-3.39	0.61
psbN	-3.76	1.00	-1.97	1.00	1.00	1.00	1.06	1.00	-6.52	0.74	3.57	1.00
psbT	-1.35	0.92	-1.29	0.95	-1.13	1.00	1.07	1.00	1.16	1.00	1.44	0.83
psbV	-1.47	0.65	-1.38	0.60	-1.52	1.00	**-1.99**	0.05	1.40	0.41	1.68	0.26
psbX	-1.05	1.00	1.62	0.79	3.50	1.00	-1.31	0.91	3.70	0.39	-1.06	1.00

Changes in selenoproteome expression

Aureococcus has 59 predicted selenoproteins – the highest reported amongst all eukaryotes (Gobler *et al.*, 2011). Out of these, 35 showed significant (FDR $p<0.1$) fold-changes at least at one time point. While a high number of selenoproteins were found to be negatively differentially expressed immediately post-infection, at 12 and 21 h a large number of selenoproteins showed increased expression relative to controls (Figure 3.12). In accordance with this large scale overrepresentation of selenoproteome, selenocysteine (Sec) methyltransferase activity was found to be overrepresented in the infected culture during the last three time points (Figure 3.10). Sec-methyltransferase plays an important role in removing the toxic effect of higher levels of selenium in plants (Neuhierl & Bock, 1996). Additionally, O-phosphoseryl-tRNA(Sec) selenium transferase, a gene involved in producing selenocysteinyl-tRNA from L-seryl-tRNA(Sec) was also found to be overexpressed (Figure 3.13). Cystathionine beta-lyase and Sec lyase, two genes involved in conversion of Sec into methionine and alanine, also showed increased expression compared to control (Figure 3.13). Although no known selenium transporter has been characterized in *Aureococcus,* it is known that opportunistic transport of selenium using phosphate transporters might be a common mechanism in plants, fungi and algae (Lazard *et al.*, 2010). *Aureococcus* has six annotated phosphate transporters. Among these, either 3or 4 of the transporters showed significant positive differential expression at 12 h and 21 h, respectively, during infection (Figure 3.14). Five of the overexpressed selenoprotein genes were methionine sulfoxide reductases (Figure 3.12), genes involved in reversing the oxidation of methionine by reactive oxygen species (ROS), thereby repairing the oxidative damage in proteins (Moskovitz, 2005). Three copies of glutathione peroxidases were also overexpressed compared to control, which are crucial in reducing H_2O_2 or organic hydroperoxides, thereby minimizing oxidative damage to cellular components (Ursini *et al.*, 1995). Another selenoprotein, dehydroascorbate reductase, which is involved in recycling of ascorbate, was overexpressed. Sel U and Sel H, two selenoproteins involved in redox functions and Sep15, a transcript whose product is involved in protein folding in endoplasmic reticulum (Labunskyy *et al.*, 2009) were overexpressed in the late stage of virus infection (Figure 3.12). Interestingly, recombinant Sel H has demonstrated significant glutathione peroxidase activity in human breast cancer cell line (LLC1) (Novoselov *et al.*, 2007). A number of redox active proteins not incorporating selenium were also overexpressed during the last two time points. This includes a copy of Cu-Zn superoxide dismutase (Aurandraft_59136), which dismutates superoxide anion (O_2^-), leading to production of H_2O_2.



Figure 3.12: The expression pattern of 35 selenoproteins in *Aureococcus* showing significant fold change (FDR p<0.1) in the virus infected culture at least at one time point. The significant fold changes are marked with asterisk.

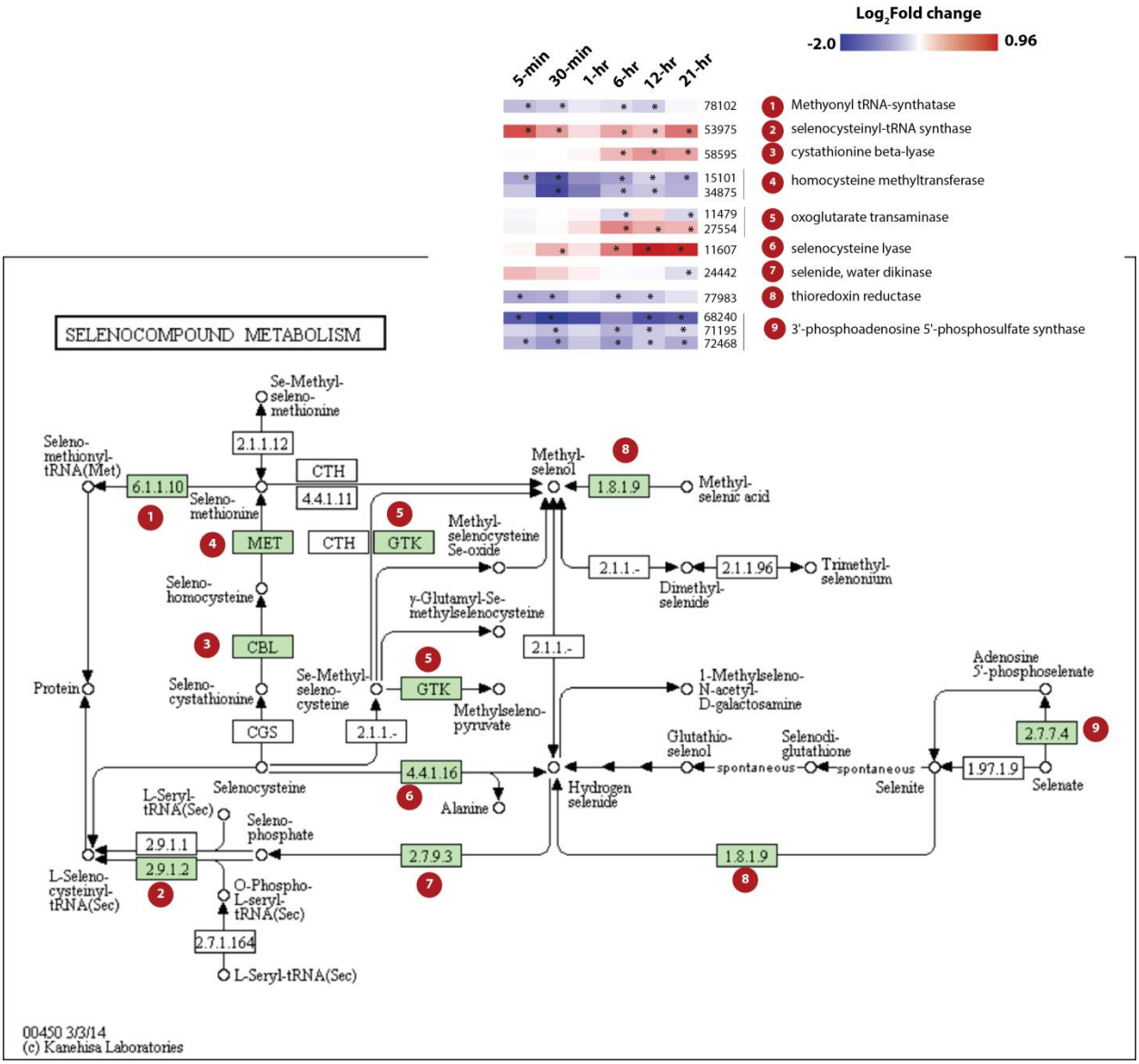


Figure 3.13: KEGG pathway "Selenocompound metabolism" with *Aureococcus* gene mapped on it. Enzyme coding genes having homologs in *Aureococcus* genome are denoted in green. The expression pattern of these genes (Log2 fold change) compared to the control are presented in the heatmap. Significant (FDR $p<0.1$) fold changes are marked with asterisk.

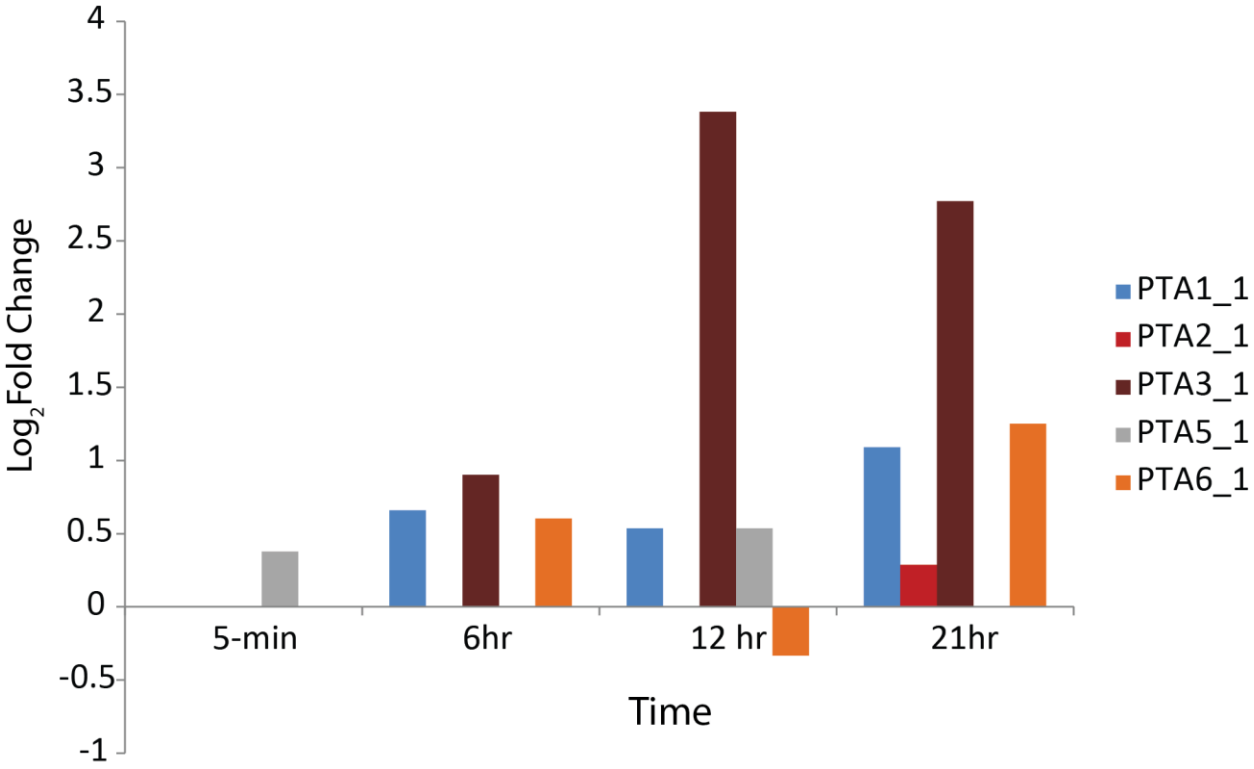


Figure 3.14: Up- or downregulation of six annotated phosphate transporters in *Aureococcus* over six time points.

Increased polyamine biosynthesis

Data exploration using GO analysis revealed an overrepresentation of spermidine synthase activity at 6 h post infection. Ornithine decarboxylase converts L-ornithine to putrescine, which is further converted to spermidine by spermidine synthase. *Aureococcus* homologs of ornithine decarboxylase and spermidine synthase were found to be up-regulated in the virus-infected culture from 6 h onwards (Figure 3.15). Additionally, a homolog of N-carbamoylputrescine amydase (Aurandraft_59241) was also found to be upregulated during this time. This suggests that cellular spermidine and putrescine pools increased during the intermediate and late phase of infection.

Discussion

Transcriptional landscape of AaV

This study provides initial insight into the gene expression dynamics of an algal virus in the Mimiviridae (proposed 'Megaviridae') clade (Santini *et al.*, 2013). AaV has a 21-30 h infection cycle, with free virus production observable by 21 hours post-infection, which steadily increases over time (Brown & Bidle, 2014). As might be anticipated a high coefficient of variation existed within replicates for the samples from the early time points , indicative of the rapid temporal changes in gene expression that occur at the beginning of the infection process (Figure 3.3). Viral reads appeared within the host cell as early as 5 min – indicating a rapid modulation of host cellular processes directed towards transcribing the viral mRNAs (Figure 3.4). Fast transcription of viral genes was also observed in Chlorella virus PBCV-1 transcriptome, where viral transcripts were detected 7 min post-infection (Blanc *et al.*, 2014). It has been demonstrated that for Mimivirus (Legendre et al., 2010), Cafeteria roenbergensis virus (Fischer *et al.*, 2010) and PBCV-1 (Blanc *et al.*, 2014), almost all the genes are detected during the course of infection. AaV is not an exception as only three genes did not have detectable expression by end of the infection cycle. However, it is possible that expression of these genes could not be detected owing to lack of sequencing depth. An interesting observation was the low expression level of the terminal DUF285 domain containing genes compared to other viral genes across all the time points. The regulatory mechanism behind the low expression dynamics of these genes is unknown – however, the role of methylation or binding of specific factors suppressing the expression of these genes can't be ruled out.

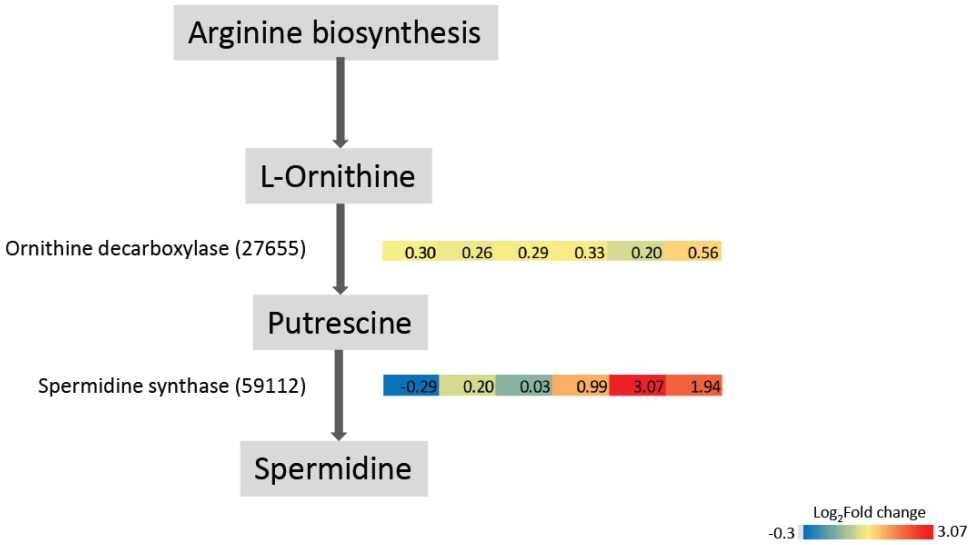


Figure 3.15: Spermidine biosynthesis pathway in *Aureococcus*. The expression changes (Log_2 fold change) compared to control are presented as heatmaps for each of the genes at right side of the pathway. The heatmaps represent fold changes over progressive sampling time from left to right.

We investigated the association of any particular promoter sequences with the temporal pattern of gene expression in AaV. The putative AaV early promoter includes an octamer ' 'AAAAATGA' (Figure 3.6), which is similar to the reported Mimivirus and Cafeteria roenbergensis virus early promoter motif "AAAATTGA" (Fischer *et al.*, 2010, Legendre *et al.*, 2010), with only one mismatch (at the fifth position). In contrast, AaV early promoter motif was not similar to that of algal virus PBCV-1, which is a member of Phycodnaviridae. This could indicate some degree of evolutionary conservation of the early promoter motif across the Mimiviridae clade. Analyzing the genome sequences of other Mimiviridae members might provide further support to this observation.

Giant viruses contain a diverse array of genes and even functional proteins (e.g. (e.g., Fischer *et al.*, 2014)) within their capsids, which they use to transform their host's cellular environment. In case of AaV, the immediately expressed genes included both proteases and ubiquitin ligases, which possibly participate in degrading the host proteins. Several transcription factors and RNA polymerase subunits were also expressed at the same time, which likely allowed transcription of the virus genes independent of the host apparatus (Appendix, Table 3.2). Interestingly, two of these HGT acquired genes are involved in carbohydrate metabolism (one a paralog of pectate lyase, and the second an unsaturated glucuronyl hydrolase) were expressed at 5 min. It is known that unsaturated glucuronyl hydrolases remove the terminal unsaturated sugar from the oligosaccharide products released by polysaccharide lyases (Jongkees & Withers, 2011). Healthy *Aureococcus* cells are surrounded by a fibrous glycocalyx, which is absent from the virus infected cells (Gastrich *et al.*, 1998): it is thus compelling to speculate that the role of viral polysaccharide lyase and glucuronyl hydrolase during infection is to make the cell membrane accessible for virus attachment. A large number of virus and host proteins are packaged into the capsid of CroV (Fischer *et al.*, 2014) and Mimivirus (Renesto *et al.*, 2006). Future proteomic analyses of AaV will be necessary to find if any of the proteins could be involved in degrading the extracellular matrix of the cell and other infection associated processes.

The *Aureococcus* 'virocell' – transcriptional remodeling upon virus infection

A surprising outcome of this study was the rapid transcriptional response of the host after virus infection, indicating several possibilities (Figure 3.8). In one way, a part of the differentially expressed gene pool might reflect host defense response to virus attack. In other way, they might include genes that are rapidly manipulated by the virus to transform the cellular environment in favor of virus propagation. This rapid change at transcription level perhaps represents how the transformation of a healthy cell into a 'virocell' is initiated. 'Virocell' is an interesting concept which emphasizes the fact that a microorganism infected by virus operate very differently from a healthy co-existing cell at a molecular level (Forterre,

2011). Our study provides a peak into the processes relevant to the *Aureococcus* virocell. Very little information concerning the molecular events associated with AaV infection of *Aureococcus* is available. For this reason, we first explored the cellular processes affected in the host using Gene Ontology (GO) (Harris et al., 2004) and KEGG pathway (Kanehisa & Goto, 2000) enrichment analysis. Individual genes are involved in multiple pathways, which is why applying FDR corrected p-value might be too stringent for such pathway analysis, missing important changes. Additionally, GO analysis using a fold-change cut-off will also reduce power to detect particular processes which were differentially regulated across all or several time-points at a reduced, yet significant expression level than the imposed cut-off. Nevertheless, these analyses provided the first look into the molecular machineries and metabolism of *Aureococcus* affected in response to virus infection. It also helped us to detect interesting processes relevant to virus infection and carry out detailed investigation on them across all the time points.

Aureococcus has a large number of nucleus-encoded light harvesting complex (LHC) proteins, which augment the photosynthetic reaction center in collecting light energy, compared to the competing phytoplankton (Gobler *et al.*, 2011). It is known that lower light level can delay the virus-mediated lysis of *Aureococcus*, and photosynthetic efficiency is not significantly different between infected and non-infected cultures at least within 24 hours of infection (Gobler *et al.*, 2007). However, the molecular basis of how virus infection can influence the host photosynthetic capacity is largely unknown. Immediately after infection, transcript levels of photosynthesis related processes went down relative to control, with increasing number of LHC proteins being downregulated as infection progressed (Figure 3.11). AaV propagation was found to be adversely affected at low light – with cultures incubated in low light (~3 µmol quanta m^{-2} s^{-1}) taking more than 7 days to be reduced to $<10^4$ cells/mL compared to high light (~110 µmol quanta m^{-2} s^{-1}) incubated culture, which took 3 days to be reduced to similar concentration (Gobler *et al.*, 2007). Downregulation of photosynthesis was also observed in *Chlorella* upon infection with PBCV-1 (Seaton *et al.*, 1995) and *Heterosigma akashiwo* infected by both RNA and DNA viruses (Philippe *et al.*, 2003). Photosynthesis was found to be downregulated in a wide range of plants in response to pathogen invasion (e.g., virus and bacteria) (Bilgin *et al.*, 2010) and was suggested to be an adaptive response to biotic attack. It is important to mention that downregulation of gene expression doesn't necessarily mean immediate loss of function – specifically, the proteins involved in light reaction might have a long functional lifetime (Bilgin *et al.*, 2010). Thereby, the actual effect of immediate downregulation of photosynthesis gene expression on viral attack remains to be elucidated. It was proposed that slow turnover of many photosynthesis related proteins allows the host to redirect resources for immediate defense mechanisms without dramatically reducing its photosynthetic capacity (Bilgin *et al.*, 2010). High light requirement of the virus and *Aureococcus's* capacity to grow in a low light

environment might itself act as a natural defense mechanism at the community level, where delayed virus production might eventually lead to fewer host-virus contacts and subsequent infection.

Porphyrin derivatives are precursors of chlorophyll biosynthesis in photosynthetic organisms. We observed an upregulation of porphyrin biosynthesis (Figure 3.10) with no significant change in expression of the gene involved in formation of chlorophyll as end-products from these chlorophyll precursors. In photosynthetic organisms, different chlorophyll precursors are formed as part of its biosynthetic pathway. However, accumulation of such precursors, especially protoporphyrin IX, can lead to photosensitivity (Inagaki *et al.*, 2015). Unbound chlorophyll and other porphyrin precursors can absorb light energy and produce oxygen radicals inside the cell. It was also demonstrated that various porphyrin derivatives might have broad antiviral activity –however, the activity was mostly extracellular. For example, an alkylated porphyrin (chlorophillide) was found to cause damage to the hepatitis B-virus capsid (Guo *et al.*, 2011), leading to loss of virion DNA. Thus, increased porphyrin level might increase the oxidative stress, making the cellular environment hostile for the invading viral components (Mock *et al.*, 1998). This observation is in contrast with EhV infected *E. huxleyi* cells, where no oxidative burst could be detected immediately after infection (Evans *et al.*, 2006). In the infected *Aureococcus* cultures, we also noticed downregulation of DNA photolyase activity (Figure 3.10). Together, porphyrin upregulation and concomitant downregulation of DNA photolyase might work as one of the first lines of host defense – an oxidative intracellular environment with a suppression of DNA repair activity. Recently, photolyase was reported to be part of CroV proteome (Fischer *et al.*, 2014). It was also found that majority of the packaged proteins, including photolyase, were late proteins (Fischer *et al.*, 2014). Interestingly, a photolyase is also encoded into AaV genome, which was found to be expressed late (12 h post-infection). It is possible that proteins involved in subversion of host defense are also packaged in AaV.

The majority of the selenoproteins characterized to date have redox-active functions; however, they can also have a wide range of biological roles (Labunskyy *et al.*, 2014). Some viruses can encode Sec containing proteins, with bioinformatics evidence provided for several mammalian viruses (Taylor *et al.*, 1997): indeed a Sec-containing glutathione peroxidase experimentally characterized in HIV-1 (Zhao *et al.*, 2000). Given the unrivaled compendium of *Aureococcus* selenoproteins within the eukaryotic domain, we were interested in how the expressions of these proteins would be modulated by AaV infection and their possible role in virus propagation. The results provide insights on at least two key aspects that are altered in the *Aureococcus* virocell compared to a healthy cell – the cellular redox status and Sec requirement. *Aureococcus* shows a sigmoidal increase in growth rate in response to increasing selenium concentration in culture, while anthropogenic enrichment of selenium in the coastal water was suggested to enhance its growth, especially during the peak of the blooms (Gobler et al, 2013). Our study indicates

that a number of upregulated selenoproteins are possibly involved in viral protein synthesis and preventing oxidation of these viral proteins, especially during the late phase of infection (Figure 3.12). It has been demonstrated that Sec-containing methionine sulfoxide reductases (MSR) are efficient, showing 10-50 fold higher enzymatic activity compared to the Cys-contatining MSRs (Kim *et al.*, 2006). Additionally, selenoproteins deemed crucial in regulating the redox state of the cells (e.g., glutathione peroxidase and dehydroascorbate reductase) were also overrepresented during infection. The cellular pro-/antioxidant balance is a highly complex process, involving a cascade of enzymatic activity and interconnected pathways. As was aptly put by Schwarz (1996), "it is difficult to distinguish between association and causation as well as between primary and secondary effects of a given virus on ROS mediated cellular injury." In EhV, redox regulation of cellular environment was found to be crucial for viral replication. Observations from transcriptome data coupled to targeted experiments revealed an elevated production of Glutathione (GSH) along with H_2O_2 accumulation at the end of the infection - which was suggested to play a role in apoptosis and viral release (Sheyn *et al.*, 2016). While in AaV, we observed upregulation of a superoxide dismutase homolog, suggesting accumulation of H_20_2 (no catalase is annotated in *Aureococcus* genome), no further inference can be made on its role without targeted experiments. However, upregulation of a large number of selenoproteins involved in protein damage repair, folding and other redox functions can indeed be interpreted as a response to increase in cellular oxidative stress under which viral protein synthesis and assembly were likely progressing. Additionally, increase in Sec biosynthesis, its conversion to other amino acids (Figure 3.13) and sec-methyltransferase activity point to the possibility of increased requirement of selenium for the infected cells. Data on effect of selenium deficiency on virus infection is scarce. A study linked selenium deficiency to genomic changes in coxackievirus B3, suggesting such changes could result from the lack of antioxidant effect exerted by selenium (Beck *et al.*, 1995). Investigating the effect of selenium deficiency on AaV and other algal virus replication might reveal interesting findings in this regard.

The role and fate of actin cytoskeleton upon invasion by diverse viruses has been extensively researched (Cudmore *et al.*, 1997). Viral subversion of actin cytoskeleton can be implicated in diverse processes - a few of which are viral surfing of cell surface, internalization, nuclear localization and even disseminasion from the host cell (Spear & Wu, 2014). However, the role of actin cytoskeleton in giant virus replication is largely unknown. It was demonstrated that actin suppressing drugs had no negative effect on Chlorella virus replication even at a concentration that is inhibitory to host cell growth (Nietfeldt *et al.*, 1992). Mimivirus and poxviruses produce distinct viral factories in the cytoplasm and the viral genome do not enter the host nucleus (Mutsafi *et al.*, 2010). In contrast, the large DNA genome of phycodnavirus (including Chlorella viruses) (Yamada *et al.*, 2006), iridoviruses (Williams *et al.*,

2005) and asfarviruses (Eulálio *et al.*, 2007) are trafficked into the host nucleus after entry into the cytoplasm. However, the dense meshwork generated by actin cytoskeleton imposes a significant barrier to this transport process (Dauty & Verkman, 2005). In the ultrastructure of AaV infected *Aureococcus* cells, we could not locate any feature that can be likened to 'virus factories' (although no molecular evidence is available yet) (Gastrich *et al.*, 1998, Gastrich *et al.*, 2002). It is possible that suppression of actin cytoskeleton can facilitate the trafficking of AaV genome into the nucleus and might also explain why Chlorella virus replication is not affected by cytoskeleton disruptive drugs. A question that emerges from these observations is whether cytoskeleton suppression is a common feature of giant viruses whose replication involves trafficking of the genome into host nucleus.

Cellular polyamines can also have imporant roles during virus infection, which can range anywhere from facilitating the virus replication to mediating virus resistance (Walters, 2003). The possible role of polyamines in propagation of several viruses (Raina *et al.*, 1981, Baumann *et al.*, 2007) has been previously noted. In some viruses, polyamines are packaged into the capsid where they can function as cations to neutralize the negative charges of the phosphate groups of nucleic acid (Gibson & Roizman, 1971). Interestingly, *Chlorella* virus PBCV-1, NY-2A and MT325 encode a complete polyamine biosynthetic pathway in their genomes, enabling them to synthesize homospermidine and putrescine (Baumann *et al.*, 2007). While no polyamine biosynthesis genes could be located on AaV genome, upregulation of host polyamine biosynthesis alongside adoption of polyamine sysnthesis genes by *Chlorella* viruses might indicate a widespread role of polyamines in giant virus life cycle.

Conclusion

Host virus interactions at nanoscale eventually shape ecosystem processes at geographical scales. Resolving the molecular aspects of ecologically relevant host-virus interactions is critical to understand the role of viruses in the biogeochemical processes. In this study, we gleaned insights into the transcriptomic response of a harmful alga upon infection by a giant virus. The ultimate fate of a cell going through lytic infection is to produce progeny viruses, which is accomplished through a completely different transcriptomic and metabolic trajectory relative to a healthy cell. The most likely outcome of this massive transcriptional response is a reprogrammed metabolic profile - specific metabolites might regulate viral replication and might be incorporated in the virion particles. The altered virocell metabolism might even influence large scale ecological processes; for example, differential uptake or release of specific compounds by virocells might alter the nutrient dynamics, thereby affecting the coexisting microbial communities. This study will help generate new hypotheses regarding individual

metabolic or regulatory processes that might have important biogeochemical consequences, and perhaps more importantly place the "virocell" into a better ecological context.

Funding

This project was funded by a Penley Fellowship (2015) awarded to Mohammad Moniruzzaman and a Kenneth and Blair Mossman Endowed Professorship to Steven W. Wilhelm.

Acknowledgement

The authors thank Mark McDonald for valuable help in setting up the experiment and sample collection.

List of References

Anderson DM, Cembella AD & Hallegraeff GM (2011) Progress in Understanding Harmful Algal Blooms: Paradigm Shifts and New Technologies for Research, Monitoring, and Management. *Annual Review of Marine Science* **4**: 143-176.

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J & Rohwer F (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**: 1-9.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW & Noble WS (2009) MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**: W202-W208.

Baumann S, Sander A, Gurnon JR, Yanai-Balser GM, Van Etten JL & Piotrowski M (2007) Chlorella viruses contain genes encoding a complete polyamine biosynthetic pathway. *Virology* **360**: 209-217.

Baumann S, Sander A, Gurnon JR, Yanai-Balser GM, Van Etten JL & Piotrowski M (2007) Chlorella viruses contain genes encoding a complete polyamine biosynthetic pathway. *Virology* **360**: 209-217.

Beck MA, Shi Q, Morris VC & Levander OA (1995) Rapid genomic evolution of a non-virulent Coxsackievirus B3 in selenium-deficient mice results in selection of identical virulent isolates. *Nat Med* **1**: 433-436.

Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**: 289-300.

Bilgin DD, Zavala JA, Zhu JIN, Clough SJ, Ort DR & DeLucia EH (2010) Biotic stress globally downregulates photosynthesis genes. *Plant, Cell & Environment* **33**: 1597-1613.

Blanc G, Mozar M, Agarkova IV, Gurnon JR, Yanai-Balser G, Rowe JM, Xia Y, Riethoven J-J, Dunigan DD & Van Etten JL (2014) Deep RNA Sequencing Reveals Hidden Features and Dynamics of Early Gene Transcription in *Paramecium bursaria* Chlorella Virus 1. *PLoS ONE* **9**: e90989.

Brown CM & Bidle KD (2014) Attenuation of virus production at high multiplicities of infection in Aureococcus anophagefferens. *Virology*. **466-467:** 71-81

Claverie JM & Abergel C (2010) Mimivirus: the emerging paradox of quasi-autonomous viruses. *Trends in genetics* **26**: 431-437.

Clarke, KR, Gorley, RN, 2015. PRIMER v7: User Manual/Tutorial. PRIMER-E, Plymouth, 296pp.

Cudmore S, Reckmann I & Way M (1997) Viral manipulations of the actin cytoskeleton. *Trends Microbiol* **5**: 142-148.

Dauty E & Verkman AS (2005) Actin Cytoskeleton as the Principal Determinant of Size-dependent DNA Mobility in Cytoplasm: A new barrier for non-viral gene delivery. *Journal of Biological Chemistry* **280**: 7823-7828.

Eulálio A, Nunes-Correia I, Salas J, Salas ML, Simões S & Pedroso de Lima MC (2007) African swine fever virus p37 structural protein is localized in nuclear foci containing the viral DNA at early post-infection times. *Virus Research* **130**: 18-27.

Evans C, Malin G, Mills GP & Wilson WH (2006) Viral infection of emiliania huxleyi (prymnesiophyceae) leads to elevated production of reactive oxygen specie. *Journal of Phycology* **42**: 1040-1047.

Filee J, Siguier P & Chandler M (2007) I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends in genetics* **23**: 10-15.

Fischer MG, Allen MJ, Wilson WH & Suttle CA (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA*. **107:**19508-19513

Fischer MG, Kelly I, Foster LJ & Suttle CA (2014) The virion of Cafeteria roenbergensis virus (CroV) contains a complex suite of proteins for transcription and DNA repair. *Virology* **466-467**: 82-94.

Forterre P (2011) Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *Comptes Rendus Chimie* **14**: 392-399.

Forterre P (2013) The virocell concept and environmental microbiology. *The ISME journal* **7**: 233-236.

Gastrich MD, Anderson OR & Cosper EM (2002) Viral-like particles (VLPS) in the alga,Aureococcus anophagefferens (pelagophyceae), during 1999–2000 brown tide blooms in Little Egg Harbor, New Jersey. *Estuaries* **25**: 938-943.

Gastrich MD, Anderson OR, Benmayor SS & Cosper EM (1998) Ultrastructural analysis of viral infection in the brown-tide alga, Aureococcus anophagefferens (Pelagophyceae). *Phycologia* **37**: 300-306.

Gibson W & Roizman B (1971) Compartmentalization of Spermine and Spermidine in the Herpes Simplex Virion. *Proc Natl Acad Sci USA* **68**: 2818-2821.

Gobler CJ, Anderson OR, Gastrich MD & Wilhelm SW (2007) Ecological aspects of viral infection and lysis in the harmful brown tide alga *Aureococcus anophagefferens*. *Aquat Microb Ecol* **47**: 25-36.

Gobler CJ, Berry DL, Dyhrman ST, *et al.* (2011) Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *Proc Natl Acad Sci USA.* **108:**4352-4357

Guillard RRL. Culture methods. In: Hallegraeff GM, Anderson DM, Cembella AD, editors. Manual on Harmful Marine Microalgae. 2nd Edition, UNESCO Publishing (2003).

Guo H, Pan X, Mao R, *et al.* (2011) Alkylated Porphyrins Have Broad Antiviral Activity against Hepadnaviruses, Flaviviruses, Filoviruses, and Arenaviruses. *Antimicrobial Agents and Chemotherapy* **55**: 478-486.

Harris MA, Clark J, Ireland A, *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: D258-261.

Inagaki N, Kinoshita K, Kagawa T, Tanaka A, Ueno O, Shimada H & Takano M (2015) Phytochrome B Mediates the Regulation of Chlorophyll Biosynthesis through Transcriptional Regulation of *ChlH* and *GUN4* in Rice Seedlings. *PLoS ONE* **10**: e0135408.

Jongkees SA & Withers SG (2011) Glycoside cleavage by a new mechanism in unsaturated glucuronyl hydrolases. *Journal of the American Chemical Society* **133**: 19334-19337.

Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27-30.

Kim H-Y, Fomenko DE, Yoon Y-E & Gladyshev VN (2006) Catalytic Advantages Provided by Selenocysteine in Methionine-S-Sulfoxide Reductases. *Biochemistry* **45**: 13697-13704.

Koonin EV & Yutin N (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* **53**: 284-292.

Labunskyy VM, Hatfield DL & Gladyshev VN (2014) Selenoproteins: molecular pathways and physiological roles. *Physiological reviews* **94**: 739-777.

Labunskyy VM, Yoo MH, Hatfield DL & Gladyshev VN (2009) Sep15, a thioredoxin-like selenoprotein, is involved in the unfolded protein response and differentially regulated by adaptive and acute ER stresses. *Biochemistry* **48**: 8458-8465.

Lang AS, Rise ML, Culley AI & Steward GF (2009) RNA viruses in the sea. *FEMS microbiology reviews* **33**: 295-323.

Lazard M, Blanquet S, Fisicaro P, Labarraque G & Plateau P (2010) Uptake of Selenite by Saccharomyces cerevisiae Involves the High and Low Affinity Orthophosphate Transporters. *The Journal of Biological Chemistry* **285**: 32029-32037.

Legendre M, Audic S, Poirot O, *et al.* (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome research* **20**: 664-674.

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

Liang TJ (2009) Hepatitis B: the virus and disease. *Hepatology (Baltimore, Md)* **49**: S13-21.

Luo W, Friedman MS, Shedden K, Hankenson KD & Woolf PJ (2009) GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**: 1-17.

Maere S, Heymans K & Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**: 3448-3449.

Mock H-P, Keetman U, Kruse E, Rank B & Grimm B (1998) Defense Responses to Tetrapyrrole-Induced Oxidative Stress in Transgenic Plants with Reduced Uroporphyrinogen Decarboxylase or Coproporphyrinogen Oxidase Activity. *Plant physiology* **116**: 107-116.

Moniruzzaman M, Gann ER, LeCleir GR, Kang Y, Gobler CJ & Wilhelm SW (2016) Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, Aureococcus anophagefferens. *FEMS Microbiology Ecology*. **92:**fiw058

Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH & Wilhelm SW (2014) Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology*. **466-467:** 60-70

Moskovitz J (2005) Methionine sulfoxide reductases: ubiquitous enzymes involved in antioxidant defense, protein regulation, and prevention of aging-associated diseases. *Biochimica et biophysica acta* **1703**: 213-219.

Mutsafi Y, Zauberman N, Sabanay I & Minsky A (2010) Vaccinia-like cytoplasmic replication of the giant Mimivirus. *Proc Natl Acad Sci USA* **107**: 5978-5982.

Neuhierl B & Bock A (1996) On the mechanism of selenium tolerance in selenium-accumulating plants. Purification and characterization of a specific selenocysteine methyltransferase from cultured cells of Astragalus bisculatus. *European journal of biochemistry / FEBS* **239**: 235-238.

Nietfeldt JW, Lee K & Van Etten JL (1992) Chlorella virus PBCV-1 replication is not affected by cytoskeletal disruptors. *Intervirology* **33**: 116-120.

Noble RT & Fuhrman JA (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquatic Microbial Ecology* **14**: 113-118.

Novoselov SV, Kryukov GV, Xu XM, Carlson BA, Hatfield DL & Gladyshev VN (2007) Selenoprotein H is a nucleolar thioredoxin-like protein with a unique expression pattern. *J Biol Chem* **282**: 11960-11968.

Ortmann AC & Suttle CA (2009) Determination of virus abundance by epifluorescence microscopy. *Methods in molecular biology (Clifton, NJ)* **501**: 87-95.

Philippe J, Janice EL, Curtis AS & Paul JH (2003) Effects of viral infection on photosynthetic processes in the bloom-forming alga Heterosigma akashiwo. *Aquatic Microbial Ecology* **31**: 9-17.

Raina A, Tuomi K & Mantyjarvi R (1981) Roles of polyamines in the replication of animal viruses. *Medical biology* **59**: 428-432.

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M & Claverie J-M (2004) The 1.2-Megabase Genome Sequence of Mimivirus. *Science* **306**: 1344-1350.

Reinbothe S, Reinbothe C, Apel K & Lebedev N (1996) Evolution of Chlorophyll Biosynthesis—The Challenge to Survive Photooxidation. *Cell* **86**: 703-705.

Renesto P, Abergel C, Decloquement P, Moinier D, Azza S, Ogata H, Fourquet P, Gorvel J-P & Claverie J-M (2006) Mimivirus Giant Particles Incorporate a Large Fraction of Anonymous and Unique Gene Products. *Journal of virology* **80**: 11678-11685.

Rosenwasser S, Ziv C, Creveld SG & Vardi A (2016) Virocell Metabolism: Metabolic Innovations During Host-Virus Interactions in the Ocean. *Trends Microbiol*. (in press)

Rowe JM, Dunlap JR, Gobler CJ, Anderson OR, Gastrich MD & Wilhelm SW (2008) Isolation of a non-phage-like lytic virus infecting aureococcus anophagefferen. *Journal of Phycology* **44**: 71-76.

Rowe JM, Jeanniard A, Gurnon JR, Xia Y, Dunigan DD, Van Etten JL & Blanc G (2014) Global analysis of *Chlorella Variabilis* NC64A mRNA profiles during the early phase of *Paramecium bursaria* Chlorella Virus-1 Infection. *PLoS ONE* **9**: e90988.

Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M & Abergel C (2013) Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci* **110**: 10800-10805.

Schwarz KB (1996) Oxidative stress during viral infection: A review. *Free Radical Biology and Medicine* **21**: 641-649.

Seaton G, Lee K & Rohozinski J (1995) Photosynthetic Shutdown in Chlorella NC64A Associated with the Infection Cycle of Paramecium bursaria Chlorella Virus-1. *Plant physiology* **108**: 1431-1438.

Sheyn U, Rosenwasser S, Ben-Dor S, Porat Z & Vardi A (2016) Modulation of host ROS metabolism is essential for viral infection of a bloom-forming coccolithophore in the ocean. *The ISME journal* **10**: 1742-1754.

Spear M & Wu Y (2014) Viral exploitation of actin: force-generation and scaffolding functions in viral infection. *Virologica Sinica* **29**: 139-147.

Suttle CA (2007) Marine viruses [mdash] major players in the global ecosystem. *Nat Rev Micro* **5**: 801-812.

Tatusov R, Galperin M, Natale D & Koonin E (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33-36.

Taylor EW, Nadimpalli RG & Ramanathan CS (1997) Genomic structures of viral agents in relation to the biosynthesis of selenoproteins. *Biological trace element research* **56**: 63-91.

Thiagarajan V, Byrdin M, Eker APM, Müller P & Brettel K (2011) Kinetics of cyclobutane thymine dimer splitting by DNA photolyase directly monitored in the UV. *Proc Natl Acad Sci USA* **108**: 9402-9407.

Ursini F, Maiorino M, Brigelius-Flohe R, Aumann KD, Roveri A, Schomburg D & Flohe L (1995) Diversity of glutathione peroxidases. *Methods in enzymology* **252**: 38-53.

Vardi A, Van Mooy BAS, Fredricks HF, Popendorf KJ, Ossolinski JE, Haramaty L & Bidle KD (2009) Viral Glycosphingolipids Induce Lytic Infection and Cell Death in Marine Phytoplankton. *Science* **326**: 861-865.

Vardi A, Haramaty L, Van Mooy BA, Fredricks HF, Kimmance SA, Larsen A & Bidle KD (2012) Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proc Natl Acad Sci U S A* **109**: 19327-19332.

Walters DR (2003) Polyamines and plant disease. *Phytochemistry* **64**: 97-107.

Wilhelm SW & Suttle CA (1999) Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**: 781-788.

Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M & Stough JMA (2016) Standing on the Shoulders of Giant Viruses: Five Lessons Learned about Large Viruses Infecting Small Eukaryotes and the Opportunities They Create. *PLoS Pathog* **12**: e1005752.

Williams T, Barbosa-Solomieu V & Chinchar VG (2005) A Decade of Advances in Iridovirus Research. *Advances in Virus Research.* **65:** 173-248.

Worden AZ, Nolan JK & Palenik B (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography* **49**: 168-179.

Yamada T, Onimatsu H & Van Etten JL (2006) Chlorella Viruses. *Advances in Virus Research.* **66:** 293-336.

Yutin N, Wolf Y, Raoult D & Koonin E (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virology journal* **6**: 223.

Zhao L, Cox AG, Ruzicka JA, Bhat AA, Zhang W & Taylor EW (2000) Molecular modeling and in vitro activity of an HIV-1-encoded glutathione peroxidase. *Proc Natl Acad Sci USA* **97**: 6356-6361.

Chapter III Appendix

Table 3.2: AaV genes with COG/NCVOG assignments and time point when their expressions were first observed. Genes with no COG/NCVOG assignments are not shown.

Gene name	Annotation	COG category	NCVOG category	Time of first expression
AaV_320	DNA directed RNA polymerase largest subunit	0086 (Transcription)	0274 (Transcription and RNA processing)	5 min
AaV_193	DNA directed RNA polymerase II subunit rpb3	0202 (Transcription)	0635 (RNA_polymerase,Transcription and RNA processing)	5 min
AaV_124	NAD dependent DNA ligase	0272 (DNA replication, recombination & repair)	0035 (DNA replication, recombination & repair)	5 min
AaV_042	Putative Lon protease	0466 (Posttranslational modification, protein turnover, chaperones)	0228 (Other metabolic functions)	5 min
AaV_084	Topoisomerase Type IA	0550 (DNA replication, recombination & repair)	0033 (DNA replication, recombination & repair)	5 min
AaV_088	Putative DNA topoisomerase IA	0550 (DNA replication, recombination & repair)	0036 (DNA replication, recombination & repair)	5 min
AaV_235	Putative Helicase/ E3 Ubiquitin Ligase	0553 (Transcription/DNA replication,recombination,repair)	0330 (Signal transduction regulation)	5 min
AaV_318	Putative deoxyuridine 5'_triphosphate nucleotidohydrolase (dUTPase)	0756 (Nucleotide transport and metabolism)	1068 (Nucleotide metabolism)	5 min
AaV_306	Putative ABC-transporter family protein	1132 (Defense mechanisms)	0002 (Miscellaneous)	5 min
AaV_203	Transcription factor TF IIB	1405 (Transcription)	1127 (Transcription and RNA processing)	5 min
AaV_381	Transcription Elongation factor TFIIS/DNA directed RNA polymerase subunit M	1594 (Transcription)	0272 (Transcription and RNA processing)	5 min
AaV_308	Putative bZIP transcription factor	1792 (Cell envelope biogenesis, outer membrane)	-	5 min
AaV_375	Putative pectate lyase	3866 (Carbohydrate transport and metabolism)	-	5 min
AaV_078	Family 88 Glycosyl Hydrolase (Unsaturated Glucuronyl hydrolase)	4225 (General function prediction only)	-	5 min
AaV_066	SUMO-1 specific cysteine protease	5160 (Posttranslational modification, protein turnover, chaperones)	0246 (Other metabolic functions)	5 min
AaV_380	Putative Zinc finger domain containing protein	5189 (Transcription / Cell division and chromosome partitioning)	0072 (DNA replication, recombination & repair)	5 min
AaV_118	Translation elongation factor EF-1 alpha	5256 (Translation, ribosomal structure and biogenesis)	0064 (Translation)	5 min

Table 3.2 Continued.

Gene name	Annotation	COG category	NCVOG category	Time of first expression
AaV_075	Class I DNA binding Protein	5648 (Chromatin structure and dynamics)	0071 (Miscellaneous)	5 min
AaV_274	Beta-1,4 Galactosyltransferase	-	0044 (Other metabolic functions)	5 min
AaV_228	C3HC4 domain containing E3 ubiquitin ligase	-	0330 (Signal transduction regulation)	5 min
AaV_292	C3HC4 type E3 Ubiquitin ligase	-	0330 (Signal transduction regulation)	5 min
AaV_151	Hypothetical protein	-	1278 (Uncharacterized)	5 min
AaV_246	Hypothetical protein	-	0632 (Uncharacterized)	5 min
AaV_179	Hypothetical protein	-	1137 (Uncharacterized)	5 min
AaV_211	mRNA capping Enzyme	-	1117 (Transcription and RNA processing)	5 min
AaV_269	Putative ATP-dependent RNA helicase	-	0031 (DNA replication, recombination & repair)	5 min
AaV_096	Putative Capsid protein	-	0022 (Virion structure and morphogenesis)	5 min
AaV_316	Putative CMP/dCMP deaminase	-	1064 (Other metabolic functions)	5 min
AaV_024	Putative Concanavalin A-like lectin/glucanase superfamily protein	-	0107 (Uncharacterized)	5 min
AaV_329	Putative DNA polymerase III subunit alpha (partial)	-	0420 (Uncharacterized)	5 min
AaV_097	Putative HNH endonuclease	-	0072 (DNA replication, recombination & repair)	5 min
AaV_159	Putative lambda-type exonuclease	-	1192 (DNA replication, recombination & repair)	5 min
AaV_117	Putative TATA-box binding family protein.	-	0313 (DNA replication, recombination & repair)	5 min
AaV_361	Putative VLTF-3 like Transcription Factor	-	0262 (Transcription and RNA processing)	5 min
AaV_112	Type II DNA modification methyl transferase	-	0234 (Other metabolic functions)	5 min
AaV_141	B family DNA polymerase	0417 (DNA replication, recombination & repair)	0038 (DNA replication, recombination & repair)	30 min
AaV_192	Putative glycosyltransferase	0438 (Cell envelope biogenesis, outer membrane)	0067 (Other metabolic functions)	30 min
AaV_043	Small conductance mechanosensetive channel	0668 (Cell envelope biogenesis, outer membrane)	-	30 min
AaV_368	Putative Histone acetyl transferase	1243 (Transcription/Chromatin structure and dynamics)	-	30 min
AaV_288	Putative SAM dependent methyltransferase	2227 (Coenzyme metabolism)	1191 (Other metabolic functions)	30 min

Table 3.2 Continued.

Gene name	Annotation	COG category	NCVOG category	Time of first expression
AaV_175	Conserved hypothetical protein	-	1216 (Uncharacterized)	30 min
AaV_142	DnaJ/Heat shock protein 40	-	0046 (Miscellaneous)	30 min
AaV_214	Hypothetical protein	-	0645 (Uncharacterized)	30 min
AaV_249	Hypothetical protein	-	1343 (Miscellaneous)	30 min
AaV_153	Putative Ion channel domain containing protein	-	1344 (Other metabolic functions)	30 min
AaV_134	Putative VLTF2 like transcription factor	-	1164 (Transcription and RNA processing)	30 min
AaV_293	DNA topoisomerase type IIA	0187 (DNA replication, recombination & repair)	0037 (DNA replication, recombination & repair)	1 hr
AaV_054	Thymidylate synthase	0207 (Nucleotide transport and metabolism)	1136(Nucleotide metabolism)	1 hr
AaV_109	Ribonucleoside di phosphate reductase alpha subunit	0209 (Nucleotide transport and metabolism)	1353 (Nucleotide metabolism)	1 hr
AaV_208	Protein Disulfide isomerase	0526 (Posttranslational modification, protein turnover, chaperones)	0317 (Other metabolic functions)	1 hr
AaV_079	Putative Exoribonuclease (Ribonuclease R)	0557 (Transcription)	-	1 hr
AaV_077	Rhomboid family Serine protease	0705 (Amino acid metabolism)	-	1 hr
AaV_324	Putative type II DNA methyltransferase	2263 (Translation, ribosomal structure and biogenesis)	0234 (Other metabolic functions)	1 hr
AaV_357	SCF ubiquitin ligase	5201 (Posttranslational modification, protein turnover, chaperones)	1299 (Other metabolic functions)	1 hr
AaV_232	Hypothetical protein	-	1131 (Uncharacterized)	1 hr
AaV_200	Hypothetical protein	-	1423 (Uncharacterized)	1 hr
AaV_113	Putative adenine specific DNA methyltransferase	-	0234 (Other metabolic functions	1 hr
AaV_226	Putative D5 Primase/Helicase	-	0023 (DNA replication, recombination & repair)	1 hr
AaV_201	Putative Holliday junction resolvase	-	0278 (DNA replication, recombination & repair)	1 hr
AaV_242	DNA directed RNA polymerase II largest subunit	0086 (Transcription)	0274 (Transcription and RNA processing)	6 hr
AaV_244	Ribonulease HII	0164 (DNA replication, recombination & repair)	-	6 hr
AaV_171	DNA mismatch repair ATPase (MutS)	0249 (DNA replication, recombination & repair)	0105 (DNA replication, recombination & repair)	6 hr
AaV_322	Cytosine-C5 specific DNA methyl transferase	0270 (DNA replication, recombination & repair)	1066 (Nucleotide metabolism)	6 hr

Table 3.2 Continued.

Gene name	Annotation	COG category	NCVOG category	Time of first expression
AaV_295	Putative N6 Adenine specific DNA methyltransferase	0286 (Defense mechanisms)	0234 (Other metabolic functions)	6 hr
AaV_282	Putative DNA repairing ATPase	0419 (DNA replication, recombination & repair)	0308 (DNA replication, recombination & repair)	6 hr
AaV_234	Nucleoside diphosphate hydrolase (MutT)	0494 (DNA replication, recombination, & repair / General function prediction only)	0236 (Transcription and RNA processing)	6 hr
AaV_173	Nucleoside diphosphate hydrolase (MutT)	0494 (DNA replication, recombination, & repair / General function prediction only)	0236 (Transcription and RNA processing)	6 hr
AaV_076	RNA polymerase sigma factor 70	0568 (Transcription)	-	6 hr
AaV_065	Proliferating cell nuclear antigen	0592 (DNA replication, recombination & repair)	0241 (DNA replication, recombination & repair)	6 hr
AaV_071	Cyclophilin type peptydyl-prolyl cis-trans isomerase	0652 (Posttranslational modification, protein turnover, chaperones)	0711 (Other metabolic functions)	6 hr
AaV_128	DNA (cytosine-5) methyltransferase	0863 (DNA replication, recombination & repair)	-	6 hr
AaV_271	RNA polymerase subunit Rpb10	1644 (Transcription)	1368 (Transcription and RNA processing)	6 hr
AaV_044	DNA polymerase X family protein	1796 (DNA replication, recombination & repair)	004 (DNA replication, recombination & repair)	6 hr
AaV_094	Putative deoxycytidylate deaminase	2131 (Nucleotide transport and metabolism)	1064 (Other metabolic functions)	6 hr
AaV_372	Putative Phaeophorbide a Oxygenase	2146 (Inorganic ion transport and metabolism / General function prediction only)		6 hr
AaV_378	Glycosyl Transferase family 25	3306 (Cell envelope biogenesis, outer membrane)	1198 (Other metabolic functions)	6 hr
AaV_323	D5-ATPase-Helicase	3378 (General function prediction only)	0023 (DNA replication, recombination & repair)	6 hr
AaV_038	Putative Pectate Lyase	3866 (Carbohydrate transport and metabolism)	-	6 hr
AaV_099	Putative Polynucleotide-kinase-3 phosphatase	4088 (Nucleotide transport and metabolism)	0243 (Other metabolic functions)	6 hr
AaV_030	Putative Superfamily II RNA helicase	4581 (DNA replication, recombination & repair)	0030 (DNA replication, recombination & repair)	6 hr
AaV_174	DNA directed RNA polymerase K subunit/rpb6	-	0522 (Transcription and RNA processing)	6 hr
AaV_383	fucosylgalactoside 3-alpha-galactosyltransferase	-	0059 (Uncharacterized)	6 hr
AaV_126	Hypothetical protein	-	0842 (Uncharacterized)	6 hr
AaV_223	Hypothetical protein	-	1024 (Uncharacterized) (exclusive to iridoviruses, this virus also have this)	6 hr
AaV_186	Hypothetical protein	-	1129 (Uncharacterized)	6 hr
AaV_303	Oxogluterate/Iron dependent dioxygenase	-	1166 (Other metabolic functions)	6 hr

Table 3.2 Continued.

Gene name	Annotation	COG category	NCVOG category	Time of first expression
AaV_073	Putative DEADDEAh box helicase	-	0032 (DNA replication, recombination & repair)	6 hr
AaV_209	Putative thioredoxin like fold containing protein	-	0629 (Uncharacterized)	6 hr
AaV_255	Undecaprenyl pyrophosphate synthase	0020 (Lipid metabolism)	-	12 hr
AaV_222	RNA polymerase beta subunit	0085 (Transcription)	0271 (Transcription and RNA processing)	12 hr
AaV_132	Ribonucleotide reductase small subunit	0208 (Nucleotide transport and metabolism)	0276 (Nucleotide metabolism)	12 hr
AaV_110	Eukaryotic Translation Elongation factor 5A	0231 (Translation, ribosomal structure and biogenesis)	-	12 hr
AaV_241	Ribonuclease H	0258 (DNA replication, recombination & repair)	-	12 hr
AaV_125	Putative ribonuclease H1	0328 (DNA replication, recombination & repair)	1352(DNA replication, recombination & repair)	12 hr
AaV_373	UbiA Prenyltransferase	0382 (Coenzyme metabolism)	-	12 hr
AaV_093	Putative Cytosine Deaminase	0402 (Nucleotide transport and metabolism / General function prediction only)	-	12 hr
AaV_082	Putative DNA photolyase class II	0415 (DNA replication, recombination & repair)	1004 (Other metabolic functions)	12 hr
AaV_170	Putative Deoxynucleoside kinase	0480 (Translation, ribosomal structure and biogenesis)	1067 (Nucleoside/Nucleotide_kinase,Nucleotide metabolism)	12 hr
AaV_178	Serine threonine protein phosphatase	0639 (Signal transduction mechanisms)	0995 (Uncharacterized). Renized by deltablast	12 hr
AaV_367	Putative RNA methylase	0742 (DNA repilcation, recombination & repair)	0564 (Other metabolic functions)	12 hr
AaV_315	Putative DNA polymerase epsilon subunit	0847 (DNA replication, recombination & repair)	0047 (DNA replication, recombination & repair)	12 hr
AaV_180	Putative VV D6R-type Helicase	1061 (Transcription / DNA replication, recombination, & repair)	0031 (DNA replication, recombination & repair)	12 hr
AaV_130	VV_A18 like Helicase	1061 (Transcription/DNA replication,recombination,repair)	0076 (DNA replication, recombination & repair)	12 hr
AaV_261	Putative HD superfamily phosphohydrolase	1078 (General function prediction only)	0603 (Uncharacterized)	12 hr
AaV_311	Superfamily I Helicase	1112 (DNA replication, recombination & repair)	-	12 hr
AaV_210	Phosphate starvation-inducible protein PhoH	1702 (Signal transduction mechanisms)	-	12 hr
AaV_215	Crossover junction endonuclease Mus81	1948 (DNA replication, recombination & repair)	-	12 hr
AaV_224	RNA polymerase subunit RPB5	2012 (Transcription)	0273 (Transcription and RNA processing)	12 hr

Table 3.2 Continued.

Gene name	Annotation	COG category	NCVOG category	Time of first expression
AaV_287	S-adenosyl L-methionine dependent Methyltransferase	2226 (Coenzyme metabolism)	1191 (Other metabolic functions)	12 hr
AaV_034	Replication factor C subunit 2	2256 (DNA replication, recombination & repair)	1351 (DNA replication, recombination & repair)	12 hr
AaV_036	Oxogluterate/Iron dependent dioxygenase	3128 (Function unknown)	1166 (Other metabolic functions)	12 hr
AaV_359	family 25 Glycosyltransferase	3306(Cell envelope biogenesis, outer membrane)	0068 (Other metabolic functions)	12 hr
AaV_111	Class 3 Lipase	3675 (Lipid metabolism)	0225 (Other metabolic functions).	12 hr
AaV_003	Putate pectate lyase	3866 (Carbohydrate transport and metabolism)	-	12 hr
AaV_290	ERV/ALR sulphydryl oxidase	5054 (Posttranslational modification, protein turnover, chaperones)	0052 (Virion structure and morphogenesis)	12 hr
AaV_152	C3H2C3-type E3 Ubiquitin ligase/RING-H2 finger protein	5194 (Posttranslational modification, protein turnover, chaperones)	0330 (Signal transduction and regulation)	12 hr
AaV_298	DNA directed RNA polymerase subunit rpb9/M	-	0521 (Uncharacterized)	12 hr
AaV_307	DnaJ/Heat shock protein 40	-	0046 (Miscellaneous)	12 hr
AaV_250	Hypothetical protein	-	0628 (Uncharacterized)	12 hr
AaV_328	Hypothetical protein	-	1083 (Uncharacterized)	12 hr
AaV_243	Hypothetical protein	-	1278 (Uncharacterized)	12 hr
AaV_239	Hypothetical protein	-	1343 (Miscellaneous)	12 hr
AaV_116	Hypothetical protein	-	0158 (Uncharacterized)	12 hr
AaV_074	Hypothetical protein	-	0329 (Other metabolic functions)	12 hr
AaV_330	Hypothetical protein	-	1012 (Uncharacterized)	12 hr
AaV_265	Hypothetical protein	-	1343 (Miscellaneous)	12 hr
AaV_285	Hypothetical protein	-	1343 (Miscellaneous)	12 hr
AaV_247	Putative Capsid protein 2	-	0022 (Virion structure and morphogenesis)	12 hr
AaV_386	Putative Concanavalin A-like lectin/glucanase superfamily protein	-	0108 (Uncharacterized)	12 hr
AaV_158	Putative metal dependent hydrolase	-	1120 (Metallopeptidase)	12 hr
AaV_190	Putative Serine threonine protein kinase haspin	-	0285 (Signal transduction regulation).	12 hr
AaV_233	Replication factor C small subunit 2	-	0001 (DNA replication, recombination & repair)	12 hr

Table 3.2 Continued.

Gene name	Annotation	COG category	NCVOG category	Time of first expression
AaV_227	Replication factor C subunit	-	0071 (Miscellaneous)	12 hr
AaV_002	RNA polymerase II second largest subunit	0085 (Transcription)	0271 (Transcription and RNA processing)	21 hr
AaV_253	RNA polymerase Rpb5, C-terminal domain	2012 (Transcription)	0273 (Transcription and RNA processing)	21 hr
AaV_123	S-phase-kinase-associated protein-1 (Skp1)	5201 (Posttranslational modification, protein turnover, chaperones)	1299 (Other metabolic functions)	21 hr
AaV_165	Putative A32 Virion packaging ATPase	-	0249 (Virion structure and morphogenesis)	21 hr

CHAPTER IV

Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, *Aureococcus anophagefferens*

Publication note:

This chapter was published as a peer-reviewed article in *FEMS Microbiology Ecology* with the authors Mohammad Moniruzzaman, Eric R. Gann, Gary R. LeCleir, Yoonja Kang, Christopher J. Gobler, and Steven W. Wilhelm.

My contribution to this work was experimental design, processing the samples received from the collaborators, wet lab works (shared with Eric R. Gann), bioinformatics analyses, statistical analyses and majority of the writing.

Citation:

Moniruzzaman M, Gann ER, LeCleir GR, Kang Y, Gobler CJ & Wilhelm SW (2016) Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, *Aureococcus anophagefferens*. *FEMS Microbiology Ecology*. **92:** fiw058.

Abstract

Many giant dsDNA algal viruses share a common ancestor with Mimivirus – one of the largest viruses, in terms of genetic content. Together, these viruses form the proposed 'Megaviridae' clade of Nucleocytoplasmic Large DNA viruses (NCLDV). To gauge Megaviridae diversity in nature we designed degenerate primers targeting the Major Capsid Protein (MCP) genes of algae-infecting viruses within this group and probed the clade's diversity during the course of a brown tide bloom caused by the harmful pelagophyte, *Aureococcus anophagefferens*. We amplified target sequences in water samples from two distinct locations (Weesuck Creek and Quantuck Bay, NY) covering twelve weeks concurrent with the proliferation and demise of a bloom. In total, 475 amplicons clustered into 145 Operational Taxonomic Units (OTUs) at 97% identity. One OTU contained 19 sequences with ≥ 97 % identity to AaV, a member of the Megaviridae clade that infects *A. anophagefferens*, suggesting AaV was present during the bloom. Unifrac analysis showed clear temporal patterns in the algal Megaviridae dynamics, with a shift in the virus community structure that corresponded to the *Aureococcus* bloom decline in both the locations. Our data provide insights regarding the environmental relevance of algal Megaviridae members and raise important questions regarding their phylodynamics across different environmental gradients.

Introduction

Nucleocytoplasmic Large DNA viruses (NCLDV) have garnered increased attention from both ecologists and evolutionary biologists in the recent years owing to their unprecedented genomic complexity, phylogenetic history, and widespread roles in the global biogeochemical cycles. They are ubiquitous in the planet's ecosystem, infecting eukaryotic hosts ranging from metazoans to unicellular algae in terrestrial, marine and freshwater systems (Koonin & Yutin, 2001). Of current interest are the NCLDVs which infect eukaryotic algae in marine and freshwater ecosystems. Although numerically less abundant than cyanobacteria, photosynthetic picoeukaryotes contribute significantly to carbon fixation due to their relatively larger size (Li, 1994). For example, at one Pacific Ocean coastal site, picoeukaryotes (cells ≤ 2 µM) were found to contribute 76% of the net carbon production compared to their cyanobacterial counterparts (Worden *et al.*, 2004). In line with this observation, a recent study demonstrated that eukaryotes in the size range from 1.8 µM to 2.8 µM can contribute up to 44% of the CO_2 fixation in the subtropical and tropical northeast Atlantic Ocean (Jardillier *et al.*, 2010). NCLDVs infecting unicellular marine and freshwater eukaryotes are thus likely an important component of the aquatic ecosystem as part of the 'viral shunt' (Wilhelm & Suttle, 1999). Indeed, as a special case of viral shunt, giant viruses have been found to induce episodic mass-scale mortality of algal blooms: the virus-induced demise of large scale blooms of the coccolithopore *Emiliania huxleyi* is a remarkable example of such a phenomenon (Wilson *et al.*, 2002, Lehahn *et al.*, 2014).

Phylogenetic analyses have revealed that NCLDVs share a common ancestor and a set of core genes despite their diverse genetic repertoire and host range (Iyer *et al.*, 2006). According to International Committee on Taxonomy of Viruses (ICTV) classification, giant viruses infecting algae are classified as belonging to the 'Phycodnaviridae' family encompassing six distinct genera (Wilson, 2012). The majority of the algal viruses with available complete genomes or partial genetic information form a monophyletic cluster congruent to this classification (Wilson, 2012). However, recent studies have shed new light and thus demanded re-evaluation of this classification system (Santini *et al.*, 2013). Mimiviridae is a clade consisting of NCLDVs infecting non-photosynthetic protists – *Acanthamoeba* and *Cafeteria roenbergensis*. Interestingly, analysis of the Global Ocean sampling (GOS) metagenomics dataset coupled with available DNA polymerase gene sequences of three algal viruses revealed that numerous Mimivirus-like sequences probably originated from algal viruses in the ocean (Monier *et al.*, 2008). This unconventional relationship was further clarified by phylogenetic analyses of two recently reported giant algal viruses – PgV 16T, a virus infecting prymnesiophite *Phaeocystis globosa* (Santini *et al.*, 2013) and AaV, a virus infecting the pelagophyte *Aureococcus anophagefferens*, (Moniruzzaman *et al.*, 2014). These studies demonstrated that despite infecting algae, AaV and PgV 16T form a

monophyletic cluster within the Mimiviridae. Several other giant viruses (for which limited genetic information is available) infecting algal hosts in the class Prymnesiophyceae (*Chrysochromulina* virus) and Parsinophyceae (*Pyramimonas orientalis* virus) also cluster in this group. This extended Mimiviridae clade was proposed to be named 'Megaviridae' based on large genome size and gene conservation patterns of the members (Santini *et al.*, 2013). Phylogenetic studies have pointed out that the algal and non-algal viruses in the Megaviridae form two distinct sister-clades in phylogenies based on several core genes (Yutin *et al.*, 2013, Moniruzzaman *et al.*, 2014). Inspite of the conflict between molecular phylogeny and taxonomical assignments, sufficient conservation of the core genes of the algal viruses across both Phycodnaviridae and Megaviridae clades have allowed scientists to design degenerate primers that include members from both these clades (Chen & Suttle, 1995, Larsen *et al.*, 2008). Such primers, based on the Major capsid and the DNA polymerase gene, have provided valuable insights regarding virus dynamics in both local and global geographical contexts (Short, 2012, Adriaenssens & Cowan, 2014).

Since the first metagenomic studies on algal viruses in the *Megaviridae* clade, subsequent efforts have pointed to the global diversity, distribution and possible quantitative importance of algal *Megaviridae* members in the world's oceans. Specifically, in terms of core gene abundance, Megaviridae were found to be the second most abundant group right after the bacteriophages in the GOS database (Ogata *et al.*, 2011), while Megaviridae and Phycodnaviridae were found to dominate the Tara Oceans Metagenome datasets (Hingamp *et al.*, 2013). Indeed, the ecological role of the known members of algal Megaviridae family are well recognized – AaV is implicated in the collapse of harmful brown tide blooms along the US East Coast (Gastrich *et al.*, 2004) while PgV is associated with the mortality of *Phaeocystis globosa*, a bloom forming algae in the North Atlantic and the North Sea coasts (Brussaard *et al.*, 2004).

Although these studies provide a first indication of the global ecological significance of algal Megaviridae, data on virus diversity and community structure within ecological contexts are scarce. Using degenerate primers targeting a region of the Major Capsid Protein (MCP) gene, we have focused in on the algal Megaviridae members, and studied their diversity during a harmful brown tide bloom. Ecological impacts of these brown tides have included mass mortality of bivalve shellfish and seagrasses (Gobler and Sunda, 2012). *A. anophagefferens* virus (AaV), an algal Megaviridae member, is a potential agent that contributes to the brown tide collapse (Gastrich *et al.*, 2004), so this bloom environment provided an excellent 'experimental system' to understand the dynamics of AaV and co-occurring algal Megaviruses. Our study provides insights on the community structure of algal Megaviruses during different phases of the bloom and also elucidates the possible contribution of these viruses in the bloom dynamics.

Materials and Methods

Sample collection and processing

During a brown tide event in 2013, samples were collected from two geographically close and connected locations, Quantuck Bay (40°48'11.95"N, 72°37'13.11"W) and Weesuck Creek in Shinnecock Bay (40°50'16.34"N, 72°34'11.75"W) along the southern shore of Long Island, New York. Quantuck Bay and Shinnecock Bay have experienced regular brown tide blooms during the past three decades (Gobler & Sunda, 2012). These locations were sampled roughly once a week from mid-May to mid-August (Figure 4.1) as part of a regularly occurring brown tide survey. For the current study, samples collected on June 6, 2013 onwards were processed. Whole water samples (30 mL) were immediately centrifuged using a Sorvall Lynx 4000 centrifuge (Thermo Scientific) equipped with a fixed angle rotor (Model no. F14-14x50cy, K-factor 798) at 29,000Xg for 30 mins (pre-chilled at 4°C) and the pellet was re-suspended in 1.5 ml final volume using L1-Se medium and frozen at -80°C until further processing.

Aureococcus, Heterotrophic bacteria and Cyanobacterial enumeration

The methods used to enumerate picoeukaryotes and bacteria have been described previously (Koch *et al.*, 2013). The field samples were collected in acid-washed 20 L carboys and processed within an hour of collection. *Aureococcus* counts were made on whole water preserved using a filter-sterilized glutaraldehyde solution (1% v/v final concentration) and stored at 4° C. Enzyme linked immunosorbent assays (ELISA) using a monoclonal antibody specific for *Aureococcus anophagefferens* (Stauffer *et al.*, 2008) were used to count *Aureococcus* cells via their fluorescein isothiocyanate (FITC) fluorescence using a flow cytometer (FACScan; Becton, Dickinson and Company). Samples were stained with SYBR Green I (Life Technologies) for estimating the abundance of heterotrophic bacteria, while picocyanobacteria and picoeukaryote densities were quantified based on pigment fluorescence. Phycocyanin-containing cyanobacteria (PC cyanobacteria) and phycoerythrin-containing cyanobacteria (PE cyanobacteria) were distinguished based on their chlorophyll *a* and phycoerythrin content using the software Cyflogic (CyFlo Ltd., Finland). Based on the prior fluorescence characteristics of cyanobacteria present during brown tides, such groups are likely to be *Synechococcus* sp. (Gobler *et al.*, 2011). PC cyanobacteria, due to their lack of phycoerythrin, can be detected in lower left corner of the chlorophyll a vs. phycoerythrin window and phycoerythrin-enriched cyanobacteria should be found in the center of the window.

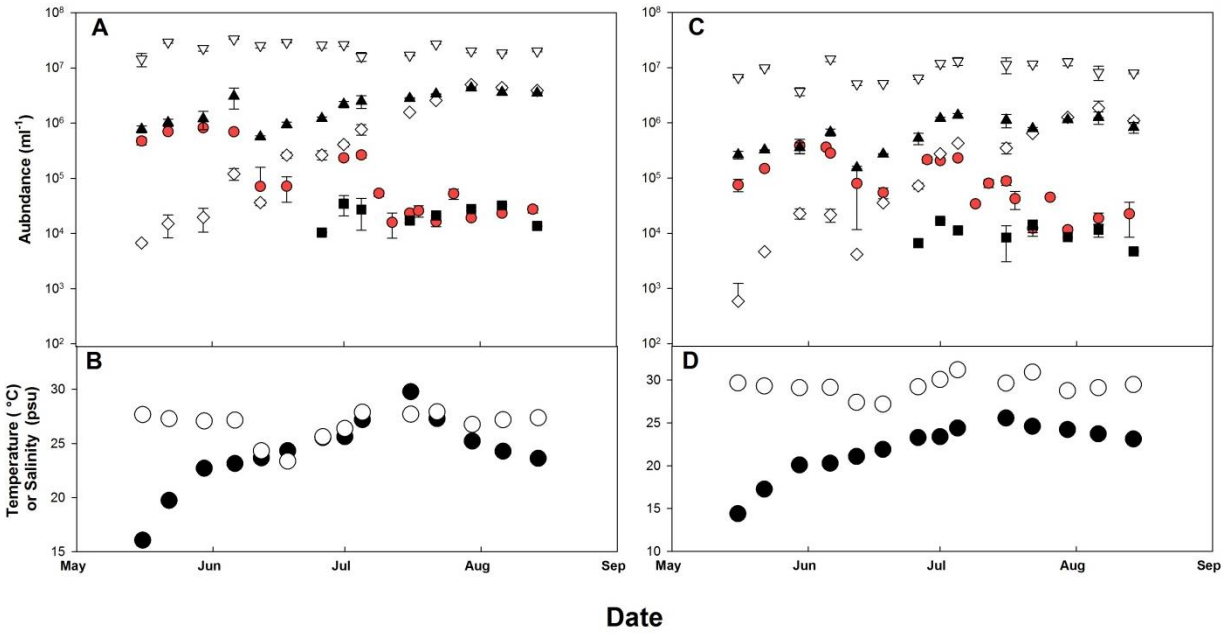


Figure 4.1: Dynamics of *Aureococcus* (●), other picoeukaryotes (▲), phycocyanin-rich cyanobacteria (◊) and phycoerythrin-rich cyanobacteria (■) during the sampling period in Quantuck Bay (**A**) and Weesuck Creek (**B**). Samples starting from June 6 were processed for this study. Temperature (●) and salinity (○) estimates during the sampling period for Quantuck Bay (**C**) and Weesuck Creek (**D**) are also given.

Primer designing and sequence data generation

Full length major capsid nucleotide sequences from both Megaviridae and Phycodnaviridae members were downloaded from the NCBI nr database. These sequences and the AaV MCP were aligned using MEGA 6.0. To facilitate the inspection of the alignment for conserved regions, protein alignments were also generated in the same program. Conserved regions specific to the algal Megaviridae were detected and degenerate primers designed (AM-mcp-F: 5`-TGGMTSCAMATYTGGWR-3', AM-mcp-R: 5` GCRTCNGTRTARTTRAA-3`) (Appendix Figure 4.5). Due to its highly divergent nature, *Pyramimonas orientalis* virus (PoV) could not be included in the scope of the designed primers. The specificity of this primer set was verified using a previously collected water sample from Quantuck Bay (January, 2012) and also an AaV lysate generated from lab cultures of *A. anophagefferens* CCMP 1984.

PCR was carried out for each sample in a 25 µL reaction volume containing 0.5 µL of Forward (2 µM), 0.5 µL Reverse (2 µM) primers, 1 µL concentrated sample,12.5 µL EconoTaq Plus Green master mix (Lucigen) and 10.5 µL sterile deionized water. A temperature cycle consisting of 95° C for 10 minutes (initial denaturation), 95° C for 30 seconds (denaturation), 52.7° C for 45 seconds (annealing), 72° C for 1 min (elongation) for a total of 37 cycles and a final extension at 72° C for 10 minutes was performed. The PCR products were separated by agarose gel electrophoresis. In none of our samples were DNA bands outside the range of 500-700 nucleotides, which was the expected size range of the MCP amplicons. DNA was excised and purified using a QIAquick gel extraction kit (Qiagen). Purified PCR products were cloned in chemically competent *E. coli* cells using TOPO TA cloning kit according to manufacturer's protocol (Invitrogen). The transformed cells were screened using the 'blue-white' selection procedure - white colonies were randomly picked from LB + kanamycin + X-gal plates after overnight incubation. 24 colonies from each of the samples were grown in 96 well plates and sent for Sanger sequencing in the Clemson University Genomics Institute. For three samples from Quantuck Bay, additional clones were sequenced by the University of Tennessee Molecular Biology Resource Facility. After removal of partial and low quality sequences, 475 sequences were curated for further analyses. Number of sequences curated for each weekly sample varied from 13 to 24 amplicons for Weesuck Creek and 17 to 32 amplicons for Quantuck Bay.. The sequences were submitted in the NCBI database under the accession number 'KT445288 - KT445765'.

OTU clustering, phylogeny and community diversity analysis

Following primer trimming, the 475 amplicons were clustered into 145 Operational Taxonomic Units (OTU) using UPARSE (Edgar, 2013) with a 97% OTU radius (greedy clustering algorithm). For phylogenetic analysis, MCP amplicons from previous studies (Larsen *et al.*, 2008, López-Bueno *et al.*,

2009, Rozon & Short, 2013, Zhong & Jacquet, 2014) were downloaded from NCBI and were clustered into OTUs following the same procedure. As part of our ongoing research on freshwater Megaviridae members, we included 19 algal Megaviridae OTUs from the Tennessee River, amplified using both AM-mcp (this study) and *mcp* (Larsen et al., 2008) primers (NCBI accession numbers: KT445288 – KT445315).

Before phylogenetic reconstruction, we translated the OTUs into amino acid sequences using FrameBot (Wang *et al.*, 2013) with a reference sequence list of MCPs from the Megaviridae and Phycodnaviridae. Translated OTUs and reference sequences from both Megaviridae and Phycodnaviridae members were aligned in MEGA 6.0 software (Tamura *et al.*, 2013) using the MUSCLE algorithm (Edgar, 2004). A maximum likelihood phylogenetic tree was constructed in Treefinder (Jobb *et al.*, 2004) with LG+F model with invariant sites. Expected likelihood ratio (1000 iterations) was used as the confidence value for the nodes. The tree was annotated using the iTOL interface (Letunic & Bork, 2011). Unweighted Unifrac (Lozupone & Knight, 2005) analysis was carried out on the amplicon sequences from Quantuck Bay (QB) and Weesuck Creek (WC) in the FastUnifrac web interface (Hamady *et al.*, 2009).

Results

Host cell dynamics

Aureococcus abundance varied widely during the course of the sampling period, which encompassed the peak and gradual decline of a brown tide bloom (Figure 4.1). For the purposes of this study, we consider cells densities exceeding 10^5 cells mL^{-1} as a 'bloom', as this species had been shown to cause harm or mortality in multiple species of bivalves and eelgrass at this level (e.g., Talmage and Gobler, 2012, Cosper et al., 1987). *Aureococcus* counts peaked towards the 'bloom' level at a concentration of 8.2×10^5 cells mL^{-1} and 3.9×10^5 cells mL^{-1} at the end of May in Quantuck Bay and Weesuck Creek, respectively. A second peak was observed at the end of June following a brief decline from the earlier peak. After this, *Aureococcus* cells started to decrease and reached a concentration of 2.8×10^4 (QB) and 2.3×10^4 cells mL^{-1} (WC) on the last sampling date (August 13, 2013) (Figure 4.1). During the last six time points corresponding to the sampling for viral analysis, *Aureococcus* cells remained at a lower abundance of $\sim 10^4$ cells mL^{-1} in both the locations. The phycocyanin-rich cyanobacterial concentrations started at 6.75×10^3 cells mL^{-1} in Quantuck Bay and 5.9×10^2 cells mL^{-1} in Weesuck Creek in May and steadily increased to a level of 1.9×10^6 (WC) and 3.9×10^6 cells mL^{-1} (QB) at the end of the sampling period (Figure 4.1). Picoeukaryotes other than *Aureococcus* persisted at high densities throughout the bloom with an average concentration of 2.2×10^6 cells mL^{-1} (QB) and 7.3×10^5 cells mL^{-1} (WC). In both locations, the concentration of picoeukaryotes steadily increased even during the demise of the brown tide

(Figure 4.1). Heterotrophic bacterial abundance was also high throughout the bloom and varied only marginally during the sampling period: between 1.4 x 10^7 and 3.3 x 10^7 cells mL^{-1} in Quantuck Bay and between 6.6 x 10^6 and 1.4 x 10^7 cells mL^{-1} in Weesuck Creek (data not shown).

Phylogenetic diversity of the MCP Operational Taxonomic Units

All the 475 amplicons from 24 sampling points were clustered into 145 OTUs at 97% OTU radius (Figure 4.3 and Appendix Figure 4.4). The most abundant OTU in our dataset had 87% amino acid similarity to the AaV MCP and was represented by 27 amplicons from both the locations (Figure 4.2). AaV was the third most abundant OTU in the dataset with 19 amplicons. Seventy one of the OTUs were singletons (Appendix Figure 4.6). Rarefaction curves constructed for the sequences from both the locations did not reach the plateau, which indicates that our sampling depth was not enough to capture the entire diversity of the algal Megaviridae members within the scope of the primer sets used (data not shown).

We performed a phylogenetic analysis of the OTUs from this study and OTUs from other studies amplified using the Larsen et al *mcp* primers. As expected, all the OTUs amplified in this study clustered in the algal Megaviridae sister clade, providing confirmation on the specificity of the primer sets (Figure 4.3). The *mcp* primers were designed from conserved regions of viruses in both Phycodnaviridae and algal Megaviridae. Consequently, sequences amplified by these primer sets clustered in both clades (Figure 4.3). A cluster consisting of AaV and PoV, two known cultured representatives, harbored a number of these sequences, while another cluster was found to harbor the Organic Lake Phycodnavirus (OLPV) (assembled from metagenomes) and sequences amplified by *mcp* primers. Interestingly, one of the sub-clades was found to solely harbor sequences amplified from freshwater studies without any cultured representatives and included one Tennessee River OTU (Moniruzzaman and Wilhelm, unpublished data) (Figure 4.3). Five other OTUs from the Tennessee River formed a monophyletic cluster with several brown tide OTUs. While a number of the brown tide OTUs clustered with cultured representatives, a majority of the sequences formed new clusters, indicating the significant unexplored diversity of algal Megaviridae in this system.

Algal Megaviridae community dynamics

We used unweighted Unifrac (Lozupone & Knight, 2005) to inspect any spatio-temporal patterns in the algal Megaviridae community structure. A Unifrac significance test on the samples pooled over the locations revealed no significant difference (P= 0.46) between these locations. However, the P-test (Martin, 2002), based on parsimony changes, indicated otherwise (P <0.001). Unifrac significance is

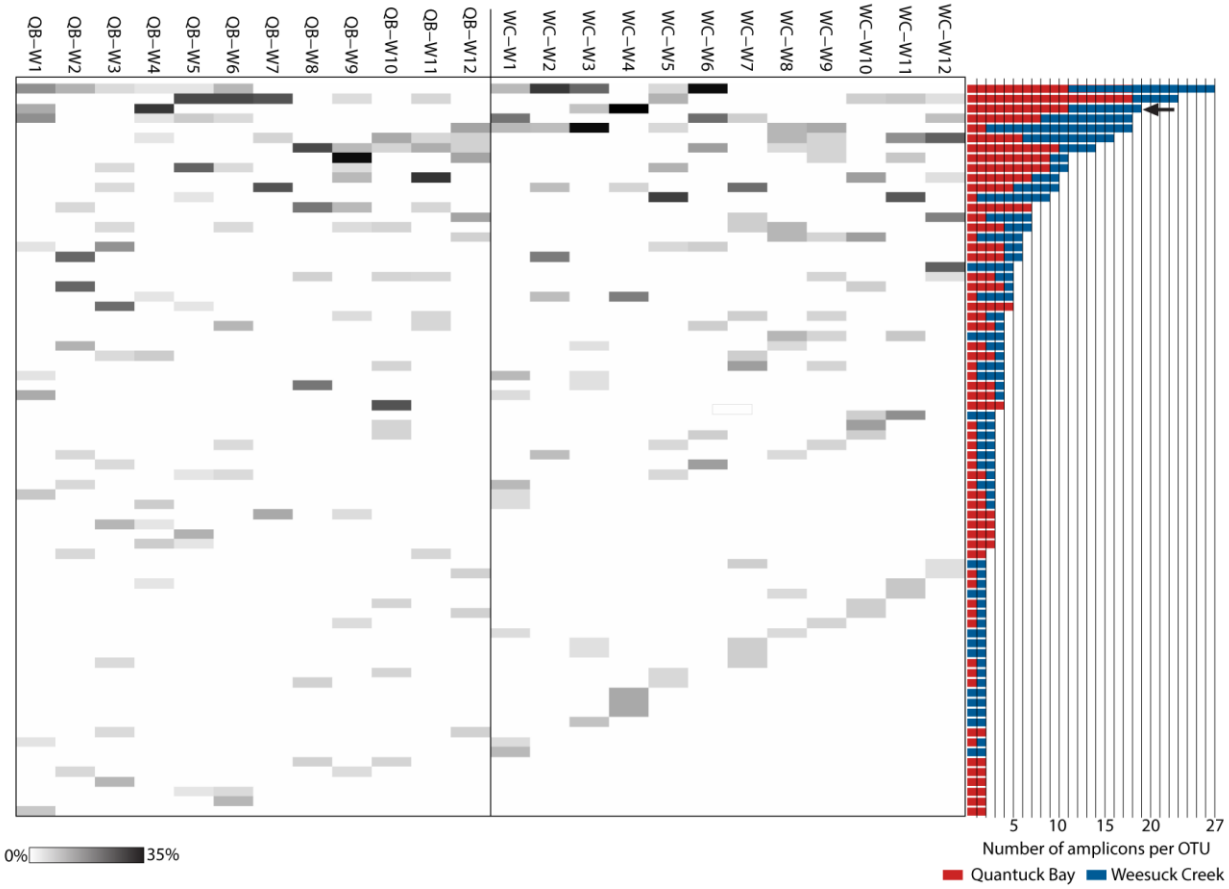


Figure 4.2: Heatmap of the distribution of amplicons representing the brown tide OTUs across different weekly samples in Quantuck Bay and Weesuck Creek (normalized to the library size). The right panel shows total number of amplicons per OTU in each of the locations. The OTUs are sorted in a descending order starting with the most abundant one (27 amplicons) on the first row of the heatmap. The OTU representing AaV is in the third row of the heatmap and indicated with an arrow.

Figure 4.3: Maximum likelihood phylogenetic analysis of MCP Operational Taxonomic Units (OTU) from this study and from previous studies using the *mcp* primers (Larsen et al, 2008). Reference sequences from Phycodnaviridae and Megaviridae are indicated in bold letters. Part of the outermost circle marked in light green indicates the Phycodnaviridae clade, while red and blue indicates algal and non-algal Megaviridae sister-clades, respectively. The innermost circle is color-coded according to the studies where they were generated. Nodes having confidence value ≥50% is marked with solid gray circles. Abbreviations: CroV – *Cafeteria roenbergensis* virus, Mou – *Acanthamoeba polyphaga* Moumouvirus, Mimi – Mimivirus, M.goul – *Moumouvirus goulette,* M.chil – *Megavirus chiliensis*, M.cour – *Megavirus courdo,* AaV – *Aureococcus anophagefferens* virus, PoV – *Pyramimonas orientalis* virus, OLPV – Organic Lake Phycodnavirus, CeV – *Chrysochromulina ericina* virus, PgV – *Phaeocystis globosa* virus, PpV – *Phaeocystis pouchetii* virus, PBCV – *Paramecium bursaria Chlorella* virus, ATCV – *Acanthocystis turfacea Chlorella* virus, BpV – *Bathycoccus parsinos* virus, MpV – *Micromonas pusilla* virus, OsV – *Ostreococcus* sp. virus, OlV – *Ostreococcus lucimarinus* virus.

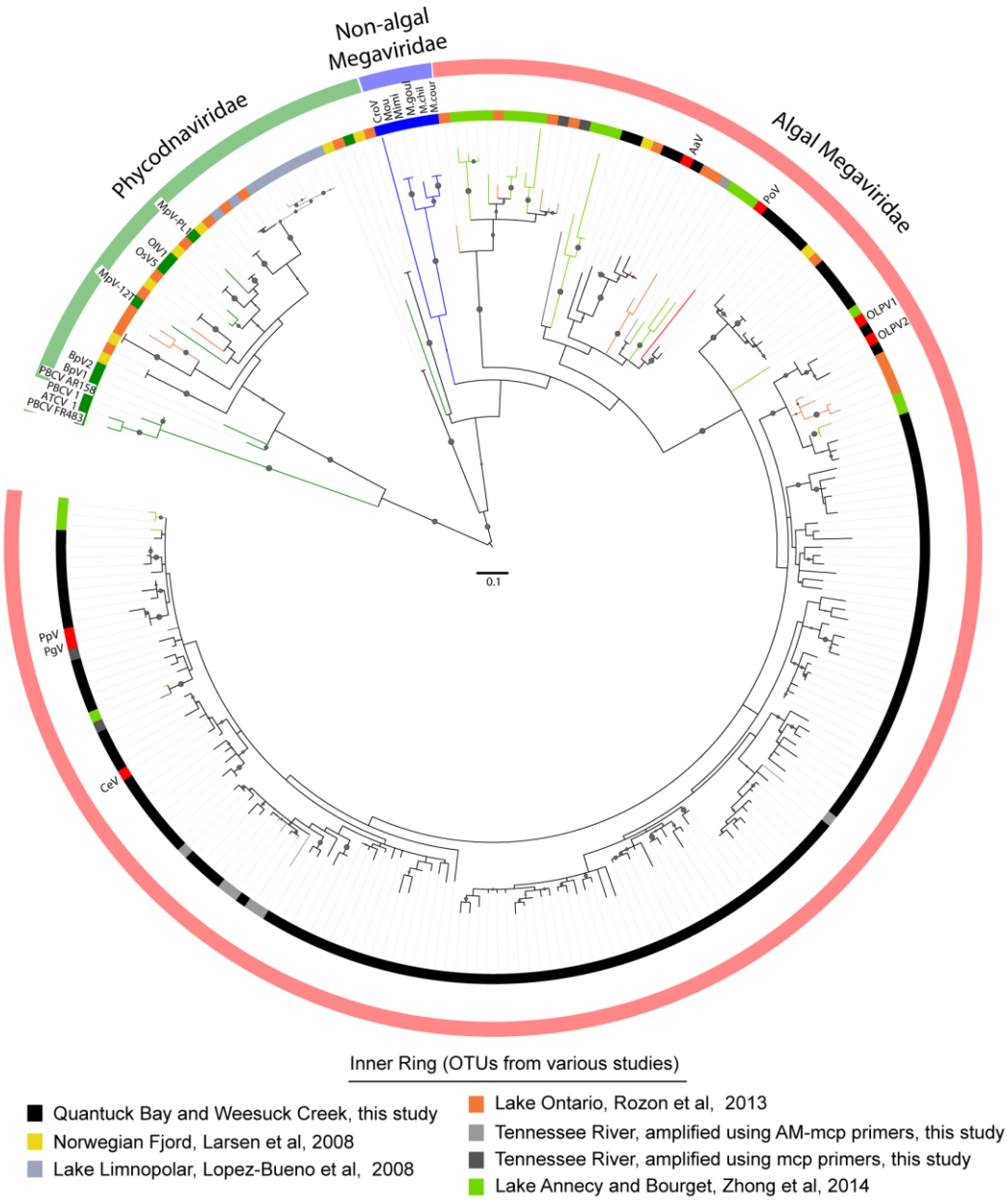


Figure 4.3. Continued.

based upon the phylogenetic branch lengths unique to each of the compared environments (Lozupone & Knight, 2005), while P-test is based upon the minimum number of changes required in the tree topology to explain the differences between the environments compared to the trees where the environments are randomly assigned (Martin, 2002).

Principal coordinates analysis of the pairwise Unifrac distances between the samples from both Quantuck Bay and Weesuck Creek showed clear separation of the earlier weeks (Week 1 to 6 for QB and Week 1 to 7 for WC) from the later weeks of the sampling (Figure 4.4). This separation was found along the Principal component 1, which explained 28.86% of the variation in this dataset. This observation was further supported by a UPGMA clustering of the samples pooled over locations, which showed two distinct clusters comprising of Week 1 - 6 samples and Week 7 – 12 (data not shown) As mentioned previously, the earlier six weeks of sampling included periods of high abundance of *Aureococcus* (~10^5 cells mL^{-1}), while cell concentration reached and remained at a lower abundance (10^4 cells mL^{-1}) over the last six sampling points. Thus, the shift we observed in the Megaviridae community structure corresponded to the major phase shift of the brown tide event. In our dataset, amplicons representing the most abundant OTU (OTU-1, 27 copies) and AaV (OTU-3, 19 copies) were retrieved exclusively from the bloom phase, while a number of OTUs were found only from the weeks representing the decline phase (e.g., OTU-8 and 10, Figure 4.2). Similar bias towards either of the phases was found for several other abundant amplicons. To test whether there was any association between the OTUs and the phases of the bloom, we pooled the 11 most abundant OTUs ($\geq$ 10 amplicons) over the locations and categorized them according to bloom and decline phases. A Pearson's chi square test rejected the null hypothesis of no association (P-value < 0.001, df = 10, simulated P-value with 2000 replicates) between these OTUs and the bloom phases.

Discussion

In light of the recent information regarding the potential ecological significance of algal Megaviruses, we explored the diversity of this intriguing phylogenetic group in both marine and freshwater ecosystems. Through marker gene analyses, we first aimed to resolve the potential diversity of AaV and other algae infecting members of Megaviridae during a brown tide cause by *A. anophagefferens* in order to develop a better quantitative, ecological understanding of these viruses. Several studies have indicated that these giant viruses may play a role in collapsing algal blooms. As a first step towards characterizing their ecology, we needed to identify suitable *in situ* conditions to study their dynamics, and brown tide events provided this opportunity.

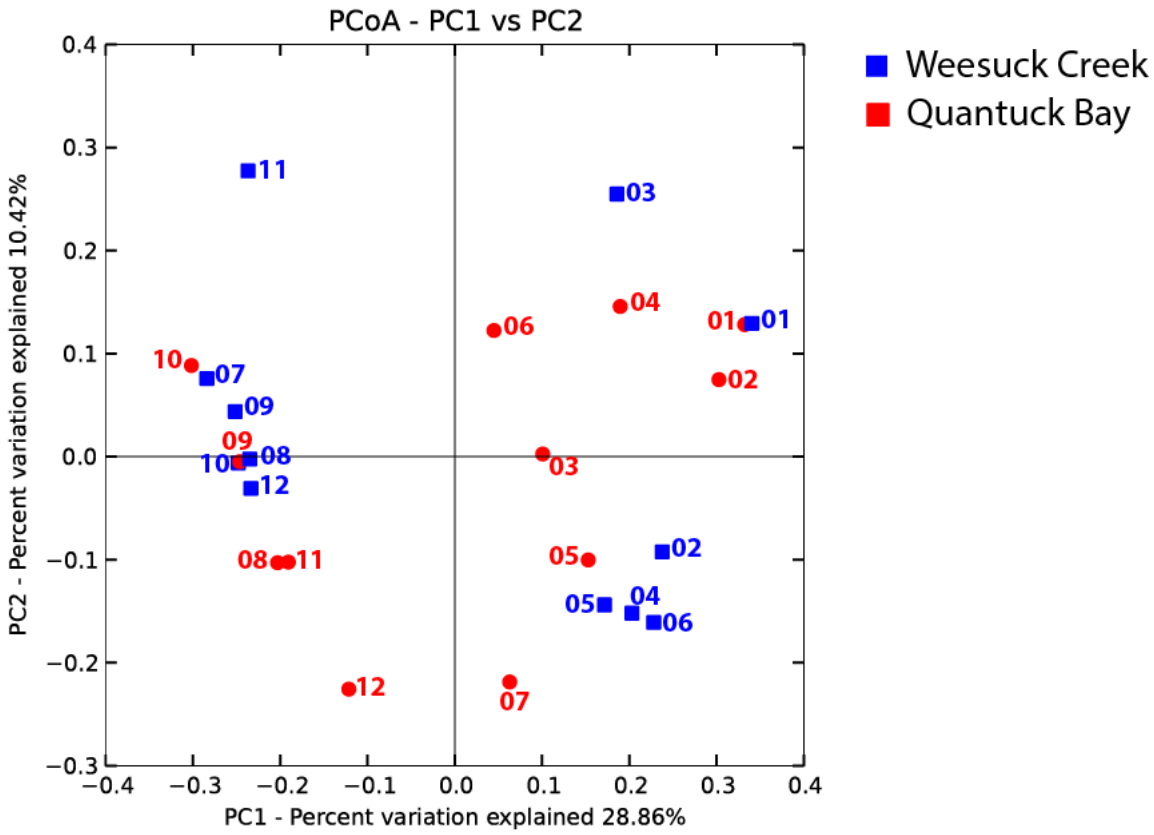


Figure 4.4: Principal coordinate analysis of the pairwise Unifrac distances between the weekly samples. Each sample is color coded (Weesuck Creek: Blue rectangle, Quantuck Bay: Red circle) along with the number of week it was collected.

In spite of a limited availability of sequence information (five cultured isolates and two candidates from metagenomics studies), our primer sets successfully amplified a broad spectrum of putative algal Megaviridae diversity with an apparent high specificity during the peak and demise of a brown tide. While this primer set was fairly specific for the environments tested, data from different environmental contexts will be necessary to further validate its specificity. It is important to mention that the reverse primer motif that we used was found to be conserved in the algal Megaviridae members, but highly diverged in the sequenced Phycodnaviridae or non-algal Megaviridae family MCP genes, which might be a factor contributing to the specificity of this primer set to algal Megaviridae members (Appendix Figure 4.5).

A phylogenetic comparison of these sequences with studies based on primers encompassing viruses from both Phycodnaviridae (ICTV specified) and Megaviridae provided several insights. Sequences from other studies using *mcp* (Larsen *et al.*, 2008) primers fell within both Phycodnaviridae and Megaviridae clade (Figure 4.3). This reflects the approach taken in the primer design where members from both Phycodna- and Megaviridae were included. It is important to note that the *mcp* (Larsen et al., 2008) primer set was designed from the regions that are conserved in the major capsid genes of both Phycodnaviridae and algal Megaviridae members. The original study was based on the established classification scheme of algal viruses, where giant viruses infecting algae are thought to form a monophyletic cluster. However, our study is based on several recent observations supporting the idea that algal viruses are probably not monophyletic - some of the cultured algal viruses and representatives in the metagenomes cluster with giant Mimivirus and *Cafeteria roenbergensis* virus (Monier et al., 2008, Santini et al., 2013, Moniruzzaman et al., 2014). These observations clearly indicate that our understanding of the algal virus phylogeny is evolving, which will also impact the algal virus classification scheme and how we will explore the diversity of these viruses in future.

Our primer sets revealed a high degree of diversity within the algal Megaviridae clade. This supports a metagenomic study that showed high degree of diversity within extended Mimiviridae family (Monier et al, 2008) in the world's oceans. Indeed, genetic information for only a few of the algal Megaviridae members is available, and this phylogenetic study reveals a number of clades with no cultural representatives. Previously, phylogenetic studies using primers specific to algal virus MCP revealed freshwater and marine specific clades (Zhong & Jacquet, 2014), while similar demarcation was found among freshwater, estuarine and marine DNA polymerase sequences (Gimenes *et al.*, 2012). Our meta-analysis indeed revealed a clade consisting of freshwater MCP sequences within the algal Megaviridae (Figure 4.3), however, the rest of the freshwater sequences were found to cluster in clades dominated by the sequences obtained from Quantuck Bay and Weesuck Creek. We note, however, that

both of these sampling locations were estuarine and within 1 - 2 km of freshwater tributaries, providing a potential eco-geographical explanation for this clustering of sequences obtained from the Tennessee river and brown tide samples. Extensive sampling of algal Megaviridae MCP across different environmental gradients will be necessary to get better insights on the phylodynamics of these viruses.

So far, AaV is the only virus known to infect and lyse *Aureococcus* in culture. Our data indicates that viral genotypes with ≥98% sequence similarity to AaV are present and are possibly a recurrent component of brown tides. Although viruses infecting *Aureococcus* other than AaV are not known, such a phenomenon might not be surprising. Distinct viral genotypes within the Megaviridae family from both marine and freshwater systems can infect the same host, *Acanthamoeba castellanii* (Raoult *et al.*, 2004, Arslan *et al.*, 2011). As a more relevant scenario, two recently reported members of the algal Megaviridae (HeV RF02 and PkV RF01) show cross-species infectivity (Johannessen *et al.*, 2015). If such cross-specificity is ubiquitous in the world's oceans, that will result in complex dynamics of both host and virus diversity (Short, 2012) and impact host-virus coevolution. One important consequence will be the possibility of viruses with broad-host range to have variable control over the population dynamics of different hosts. We found association of several OTUs with the peak and decline of the bloom event, suggesting that beyond AaV, multiple virus-host interactions shaped the bloom trajectory. The most abundant OTU in this study had 87% amino acid similarity to AaV MCP. Both AaV and this amplicon showed association with the 'bloom' phase of the sampling period. Although these provide tantalizing evidence for cross-specific infectivity involving *Aureococcus* or phylogenetically close algal species, we must note that our observation only provides a hypothesis that must be rigorously tested. Ultimately, such association can only be established in the context of large datasets generated on broader spatial and temporal scales with prior knowledge on interacting pairs of hosts and viruses.

We used Unifrac (Lozupone & Knight, 2005), a widely used phylogeny based distance metric to inspect the patterns of viral diversity and identify potential forces shaping algal Megaviridae community in the studied systems. Unifrac significance tests and phylogenetic P-tests (Martin, 2002) produced conflicting results regarding the similarity between Quantuck Bay and Weesuck Creek algal Megaviridae community over the sampling period. Unifrac significance is based upon the phylogenetic branch lengths unique to each of the compared environments whereas P-test is based upon the minimum number of changes required in the tree topology to explain the differences between the environments compared to the trees where the environments are randomly assigned. The conflicting results in our study indicate that although environment specific clusters were formed, they did not contribute enough unique branch length fractions to have a significant difference in the phylogenetic history. However, it is important to note that these bays are interconnected and had similar densities and dynamics of *Aureococcus* (Gobler *et al.*,

2011), other phototrophs and heterotrophs during the study period. Therefore, this apparent conflict could be an artifact of the proximity of sampling locales and low sampling depth in our study. Due to the high number of samples compared (24) and limited sequences per sample, we opted for unweighted Unifrac rather than the weighted version. Brown tide blooms are highly dynamic in nature and are initiated as a response to a multitude of bottom-up and top-down factors (Gobler *et al.*, 2005) of which several are already known. It has been demonstrated that these blooms emerge within a summer pico-algal assemblage that developed after the collapse of a spring diatom bloom (Sieracki *et al.*, 2004). Due to its unusual ability to utilize dissolved organic matter (DOM) (Lomas *et al.*, 2001), *Aureococcus* can outcompete co-occurring picoalgae when dissolved inorganic nitrogen (DIN) is low (Keller & Rice, 1989, Gobler *et al.*, 2002). In our analysis, temperature was negatively correlated with *Aureococcus*, while picoeukaryotic algal (other than *Aureococcus*) abundances seemed to be positively correlated (Supplementary Table 1). This observation can partially be explained by the fact that the apparent maximum temperature for brown tide bloom maintenance is 25° C (Nuzzi & Waters, 1989). Due to low sampling frequency, however, this analysis could not be extended to elucidate the relationship between viral dynamics and the abiotic factors. However, the fact that temperature had a contrasting relationship with the potential host pools of the Megaviridae (*Aureococcus* and other picoeukaryotes) indicates that such relationships are likely to be complex.

Little information is available regarding the co-occurring algal and viral diversity in the brown tide affected environments. It has been previously shown that *Aureococcus* can dominate yet still co-exist with several other known phytoplankton including *Ostreococcus*, *Thalassiosira*, *Phaeodactylum* and the cyanobacterium, *Synechococcus* (Gobler *et al.*, 2011). A study of estuaries along the mid-Atlantic coast which are historically prone to brown tides (including Quantuck Bay and Shinnecock Bay) found that protist diversity was high throughout the study and that dramatic fluctuations occurred within a very short temporal scale, regardless of study site. (Vigil *et al.*, 2009). Surprisingly, protistan diversity was high even during *Aureococcus* bloom events (Vigil *et al.*, 2009). The annually recurring community was found to be relatively stable at the same sampling sites, indicating the presence of a highly dynamic yet a resilient community over the long term. Interestingly, an *Ostreococcus*-like population bloomed in summer 2001 in West Neck Bay which has history of brown tides (O'Kelly *et al.*, 2003). These bloom forming cells were found to harbor giant icosahedral virus particles. The high abundance of picoeukaryotes in our dataset throughout the sampling period indicates that the *Aureococcus* blooms were not monospecific and other phytoplankters coexisted. Along with *Aureococcus*, this picoeukaryote pool likely harbored hosts of the algal Megaviridae identified during this study.

While the demise of the *Aureococcus* bloom witnessed during this study was sharp, we note post-peak cell densities remained at 10^4 cell/ml, a level commonly present in NY estuaries when this species is not in bloom (Gobler et al., 2011). The existence of a background density of 10^4 cells mL^{-1} may be related to the persistence of a viral resistant sub-population of *Aureococcus*. Prior work with laboratory propagated, *Aureococcus*-specific viruses demonstrated that the addition of these viruses to field populations of *Aureococcus* was capable of significantly reducing cell densities by 70% in 72 hrs (Gobler et al. 2007). The persistence of nearly 10^5 cells mL^{-1} following the experiment was interpreted as the persistence of a viral resistant sub-population. During the present study, the decline in *Aureococcus* cell densities represented a loss of >95% of the cells present at the peak of the bloom. Such decline is consistent in magnitude with prior studies that described an increase in *Aureococcus* cells infected with viruses from <1% to nearly 40% at the bloom peak (Gastrich et al 2004) suggesting the current bloom may have experienced similar rates of viral infection. However, other factors including grazing (Gobler et al., 2005) and reduction in the DOM pool (Sieracki et al., 2004) might also contribute to this decline.

Several studies revealed dramatic shifts in both algal and viral populations following viral lysis of the dominant phytoplankton (Castberg *et al.*, 2001, Jardillier *et al.*, 2005). A mesocosm study demonstrated that *E. huxleyi* bloom collapse followed changes in both the abundance and composition of the viral, bacterial, and algal communities (Castberg *et al.*, 2001). Similarly, we observed a shift in the viral community which coincided with the extended period of low abundance of *Aureococcus* (~10^4 cells mL^{-1}) after the peak. Interestingly, the non-brown tide picoeukaryote densities steadily increased over the low abundance period suggesting that a succession in both the phytoplankton and algal Megaviridae community followed the bloom decline. The possible role of AaV in this decline can be explained by the 'killing the winner' hypothesis (Thingstad, 2000), where viral lysis of the most abundant algae releases new nutrients and opens a niche that other algal species compete for. However, the contribution of additional biotic and abiotic factors in this shift cannot be discounted. Additionally, the role of algal Megaviridae-host interactions other than AaV in the bloom decline remains an open question.

In the last few decades, we have come to appreciate the critical role of aquatic viruses in maintaining the diversity and dynamics of the abundant primary producers in the ocean (Wilhelm and Suttle, 1999). Due to their well-recognized contribution to the global biogeochemical cycle, there is an increasing consensus regarding the need to incorporate viruses in the global climate models (Danovaro *et al.*, 2011). In this study, we have extended our current understanding of viral diversity and dynamics to a newly recognized phylogenetic group of giant algal viruses. Given the massive and episodic mortality events these viruses appear to cause (Wilson *et al.*, 2002, Gastrich *et al.*, 2004), the ability to track these viruses and their effects is of paramount interest to establish appropriate ecological models. Through *in*

situ identification of AaV using molecular markers, we have reinforced its role in the mortality of a harmful algal bloom and have established a foundation for further studies. Moreover, the highly diverse algal Megaviridae community was shown to respond to environmental changes, revealing their contribution to the eukaryotic phytoplankton dynamics along with the relatively well-studied Phycodnaviridae clade. Their intriguing phylogenetic affinity to viruses infecting hosts with fundamentally different life cycles, possible quantitative importance, and role in the evolution of the enormously diverse eukaryotic phytoplankton open up a new and challenging frontier of research questions.

Funding

This research was support by the *National Science Foundation* (EF-0949120, OCE-1061352 to SWW) and the *Kenneth & Blaire Mossman Endowed Professorship* to the University of Tennessee.

Acknowledgements

The authors thank a decade of students and collaborators for their continued support of work with the *Aureococcus anophagefferens* system. We also thank two anonymous reviewers for comments that greatly improved the manuscript.

List of References

Adriaenssens EM, Cowan DA. Using signature genes as tools to assess environmental viral ecology and diversity. *Appl Environ Microbiol* 2014; **80**: 4470-4480.

Arslan D, Legendre M, Seltzer V, et al. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci USA* 2011; **108**: 17486-17491.

Brussaard CP, Short SM, Frederickson CM, et al. Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton *Phaeocystis globosa* (Prymnesiophyceae). *Appl Environ Microbiol* 2004; **70**: 3700-3705.

Buskey EJ, Liu HB, Collumb C, et al. The decline and recovery of a persistent Texas brown tide algal bloom in the Laguna Madre (Texas, USA). *Estuaries* 2001; **24**: 337-346.

Castberg T, Larsen A, Sandaa RA, et al. Microbial population dynamics and diversity during a bloom of the marine coccolithophorid *Emiliania huxleyi* (Haptophyta). *Mar Ecol Prog Ser* 2001; **221**: 39-46.

Chen F, Suttle CA. Amplification of DNA polymerase gene fragments from viruses infecting microalgae. *Appl Environ Microbiol* 1995; **61**: 1274-1278.

Cosper EM, Dennison WC, Carpenter EJ, et al. Recurrent and persistent brown tide blooms perturb coastal marine ecosystem. *Estuaries* 1987; **10:** 284-290.

Danovaro R, Corinaldesi C, Dell'anno A, et al. Marine viruses and global climate change. *FEMS Microbiol Rev* 2011; **35**: 993-1034.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**: 1792-1797.

Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Meth* 2013; **10**: 996-998.

Gastrich M, Leigh-Bell J, Gobler C, et al. Viruses as potential regulators of regional brown tide blooms caused by the alga, *Aureococcus anophagefferens*. *Estuaries* 2004; **27**: 112-119.

Gimenes MV, Zanotto PM, Suttle CA, et al. Phylodynamics and movement of Phycodnaviruses among aquatic environments. *ISME J* 2012; **6**: 237-247.

Gobler CJ, Sunda WG. Ecosystem disruptive algal blooms of the brown tide species, *Aureococcus anophagefferens* and *Aureoumbra lagunensis*. *Harmful Algae* 2012; **14**: 36-45.

Gobler CJ, Renaghan MJ, Buck NJ. Impacts of nutrients and grazing mortality on the abundance of *Aureococcus anophagefferens* during a New York brown tide bloom. *Limnol and Oceanogr* 2002 **47**: 129-141.

Gobler CJ, Lonsdale DJ, Boyer GL. A review of the causes, effects, and potential management of harmful brown tide blooms caused by *Aureococcus anophagefferens* (Hargraves et Sieburth). *Estuaries* 2005; **28**: 726-749.

Gobler CJ, Berry DL, Dyhrman ST, et al. Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *Proc Natl Acad Sci USA* 2011; **108:** 4352-4357.

Gobler CJ, Anderson OR, Gastrich MD, et al. Ecological aspects of viral infection and lysis in the harmful brown tide alga *Aureococcus anophagefferens*. *Aquat Microbiol Ecol* 2007; **47**: 25-36

.

Hamady M, Lozupone C, Knight R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 2009; **4**: 17-27.

Hingamp P, Grimsley N, Acinas SG, et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* 2013; **7**: 1678-1695.

Iyer LM, Balaji S, Koonin EV, et al. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* 2006, **117**: 156-184.

Jardillier L, Zubkov MV, Pearman J, et al. Significant CO_2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME J* 2010; **4**: 1180-1192.

Jardillier L, Bettarel Y, Richardot M, et al. Effects of viruses and predators on prokaryotic community composition. *Microb Ecol* 2005; **50**: 557-569.

Jobb G, von Haeseler A, Strimmer K. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* 2004; **4**: 18.

Johannessen TV, Bratbak G, Larsen A, et al. Characterisation of three novel giant viruses reveals huge diversity among viruses infecting Prymnesiales (Haptophyta). *Virology* 2015; **476**: 180-188.

Keller AA, Rice RL. Effects of nutrient enrichment on natural populations of the brown tide phytoplankton *Aureococcus anophagefferens* (Chrysophyceae). *J Phycol* 1989; **25**: 636-646.

Koch F, Sañudo-Wilhelmy SA, Fisher NS, et al. Effect of vitamins B1 and B12 on bloom dynamics of the harmful brown tide alga, *Aureococcus anophagefferens* (Pelagophyceae). *Limnol Oceanogr* 2013; **58**: 1761-1774.

Koonin EV, Yutin N. Nucleo-cytoplasmic Large DNA Viruses (NCLDV) of Eukaryotes. In: *eLS.* John Wiley & Sons Ltd, Chichester, 2012. http://www.els.net DOI: 10.1002/9780470015902.a0023268

Larsen JB, Larsen A, Bratbak G, et al. Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. *Appl Environ Microbiol* 2008; **74**: 3048-3057.

Lehahn Y, Koren I, Schatz D, et al. Decoupling physical from biological processes to assess the impact of viruses on a mesoscale algal bloom. *Curr Biol* 2014; **24:** 2041-2046.

Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 2011; **39**: W475-W478.

Li WKW. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnol Oceanogr* 1994; **39**: 169-175.

Lomas MW, Glibert PM, Clougherty DA, et al. Elevated organic nutrient ratios associated with brown tide algal blooms of *Aureococcus anophagefferens* (Pelagophyceae). *J Plankton Res* 2001; **23**: 1339-1344.

López-Bueno A, Tamames J, Velázquez D, et al. High diversity of the viral community from an Antarctic lake. *Science* 2009; **326**: 858-861.

Lozupone C, Knight R.UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005; **71**: 8228-8235.

Martin AP. Phylogenetic Approaches for describing and comparing the diversity of microbial communities. *Appl Enviorn Microbiol* 2002; **68**: 3673-3682.

Monier A, Larsen J, Sandaa R-A, et al. Marine mimivirus relatives are probably large algal viruses. *Virology J* 2008; **5**: 12.

Moniruzzaman M, LeCleir GR, Brown CM, et al. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology* 2014; **466-467:** 60-70.

Nuzzi R, Waters RM, et al. The Spatial and Temporal Distribution of "Brown Tide" in Eastern Long Island. In: Cosper EM, Bricelj VM, Carpenter, EJ (eds). *Novel Phytoplankton Blooms*. Springer-Verlag, 1989. 118-137.

O'Kelly CJ, Sieracki ME, Thier EC, et al. A transient bloom of *Ostreococcus* (Chlorophyta, Prasinophyceae) in West Neck Bay, Long Island, New York. *J Phycol* 2003; **39**: 850-854.

Ogata H, Monier A, Claverie JM. Distribution of Giant Viruses in Marine Environments. In: Ceccaldi HJ, Dekeyser I, Girault M, Stora G (eds). *Global Change: Mankind-Marine Environment Interactions.* Springer, Netherlands, 2011. 157-162.

Raoult D, Audic S, Robert C, et al. The 1.2-Megabase Genome Sequence of Mimivirus. *Science* 2004; **306**: 1344-1350.

Rozon RM, Short SM. Complex seasonality observed amongst diverse phytoplankton viruses in the Bay of Quinte, an embayment of Lake Ontario. *Freshw Biol* 2013; **58**: 2648-2663.

Santini S, Jeudy S, Bartoli J, et al. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci USA* 2013; **110**:10800-10805.

Short SM. The ecology of viruses that infect eukaryotic algae. *Environ Microbiol* 2012; **14**: 2253-2271.

Sieracki ME, Gobler CJ, Cucci TL, et al. Pico- and nanoplankton dynamics during bloom initiation of *Aureococcus* in a Long Island, NY bay. *Harmful Algae* 2004; **3**: 459-470.

Stauffer BA, Schaffner RA, Wazniak C, et al. Immunofluorescence flow cytometry technique for enumeration of the brown tide alga, *Aureococcus anophagefferens*. *Appl Environ Microbiol* 2008; **74**: 6931-6940.

Talmage SC and Gobler CJ. Effects of CO_2 and harmful alga *Aureococcus anophagefferens* on growth and survival of oyster and scallop larvae. *Mar Ecol Prog Ser* 2012; **464**:121-134.

Tamura K, Stecher G, Peterson D, et al. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* 2013; **30**: 2725-2729.

Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* 200; **45**: 1320-1328.

Vigil P, Countway PD, Rose J. Rapid shifts in dominant taxa among microbial eukaryotes in estuarine ecosystems. *Aquat Microbiol Ecol* 2009; **54**: 83-100.

Wang Q, Quensen JF, Fish JA, et al. Ecological patterns of nifH genes in four terrestrial climatic zones explored with targeted metagenomics using FrameBot, a new informatics tool. *mBio* 2013; **4**: e00592-13.

Wilhelm SW, Suttle CA. Viruses and nutrient cycles in the sea. *BioScience* 1999; **49**: 781-788.

Wilson WH, Tarran GA, Schroeder D, et al. Isolation of viruses responsible for the demise of an Emiliania huxleyi bloom in the English Channel. *J Mar Biol Assoc U.K.* 2002; **82**: 369-377.

Wilson WH, Van Etten JL, Schroeder DC, et al. Family: Phycodnaviridae. In: Andrew MQ, King EL, Michael JA, Eric BC (eds.).*Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier, London, 2012. 248-262.

Worden AZ, Nolan JK, Palenik B. Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnol Oceanogr* 2004; **49**: 168-179.

Yutin N, Colson P, Raoult D, et al. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virology J* 2013; **10**: 106.

Zhong X, Jacquet S. Contrasting diversity of phycodnavirus signature genes in two large and deep western European lakes. *Env Microbiol* 2014; **16**: 759-773.

Chapter IV Appendix

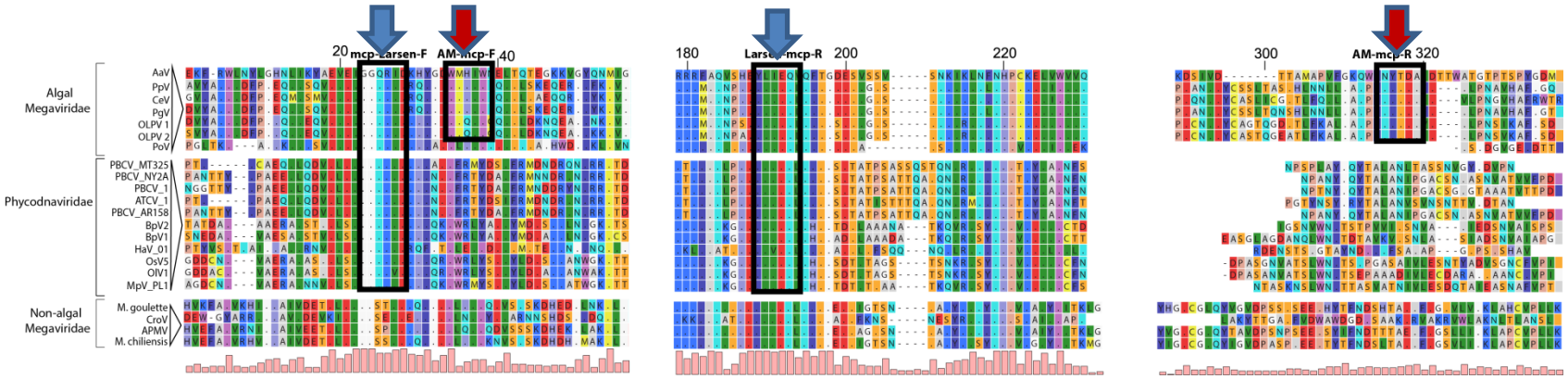


Figure 4.5: Alignment of Major Capsid Protein (MCP) sequences from the Megaviridae and Phycodnaviridae clade used to design the degenerate primers specific to algal Megaviridae. Region Marked in red arrows indicate the primers designed in this study (AM-mcp), while regions indicated in blue are the primers designed by Larsen et al, 2008. AaV: *Aureococcus anophagefferens* virus, PpV: *Phaeocystis pouchetii* virus, CeV: *Chrysochromulina ericina* virus, PgV: *Phaeocystis globosa* virus, OLPV: Organic Lake Phycodnavirus, PoV: *Pyramimonas orientalis* virus, PBCV: *Paramecium bursaria Chlorella* virus, ATCV: *Acanthocystis turfacea Chlorella* virus, BpV: *Bathycoccus parsinos* virus, HaV: *Heterosigma akashiwo* virus, OsV: Ostreococcus virus, OlV: Ostreococcus lucimarinus virus, MpV: Micromonas pusilla virus, M.goulette: *Moumouvirus goulette*, CroV: *Cafeteria roenbergensis* virus, APMV: *Acanthamoeba polyphaga* Mimivirus, M.chiliensis: *Megavirus chiliensis*.

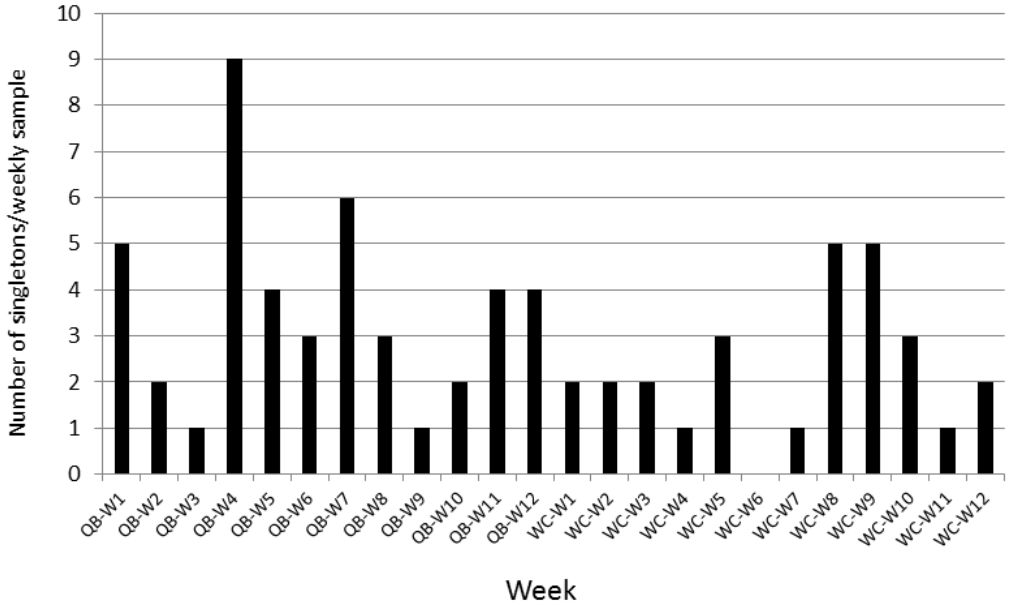


Figure 4.6: Distribution of singleton OTUs in each of the weekly samples from both the sampling locations. QB – Quantuck Bay, WC – Weesuck Creek.

Table 4.1: Spearman's correlation co-efficient (ρ) indicating significant relationship between cell counts vs. environmental variables. (PC_Cyano: Phycocyanin rich cyanobacteria.)

Variable	**by Variable**	**Spearman ρ**	**P-value**
PC_Cyano	*Aureococcus anophagefferens*	-0.6087	0.0016
Picoeukaryotes	Heterotrophs	0.6322	0.0003
Picoeukaryotes	PC_Cyano	0.7980	<.0001
Temperature	*Aureococcus anophagefferens*	-0.4548	0.0256
Temperature	Heterotrophs	0.3870	0.0419
Temperature	PC_Cyano	0.7619	<.0001
Temperature	Picoeukaryotes	0.6798	<.0001
Salinity	Heterotrophs	-0.6525	0.0002

CHAPTER V

Virus-host infection dynamics for marine microeukaryotes resolved from metatranscriptomics

Publication note:

This chapter is the draft of a manuscript that will be submitted to *Proceedings of National Academy of Sciences* authored by Mohammad Moniruzzaman, Louie Wurch, Harriet Alexander, Christopher G. Gobler, Sonya Dyhrman and Steven W Wilhelm.

My contribution to this article was data gathering from the collaborators, all the bioinformatics and statistical analyses and writing majority of the manuscript.

Abstract

Metatranscriptomics has emerged as an important tool in microbial ecology that can resolve the functional landscape of both prokaryotes and eukaryotes in a community. In this study, we evaluated the potential of metatranscriptomics to probe virus dynamics and virus-host relationships in marine systems. RNA-seq data from polyadenylation-selected samples were examined from microbial communities in time-courses in two highly productive environments: a brown tide bloom in Quantuck Bay, NY, and a diatom dominated community in Narragansett Bay, NY. Active virus infections by a diversity of giant viruses (NCLDVs) infecting algal and non-algal hosts were found in both sites. Ongoing infections of the brown tide causative agent (*Aureococcus anophagefferens*) by a known *Mimiviridae* (AaV) were observed during both the peak and decline of a brown tide bloom. Remarkably, bloom decline was also accompanied by increased activity for viruses other than AaV, including (+)ssRNA viruses in an unassigned group of the Picornavirales. In Narragansett Bay, better temporal resolution revealed active NCLDVs with both 'boom-and-bust' as well as 'steady-state infection'-like ecologies. From the data we assembled 18 complete or near-complete genomes belonging to the Picornavirales order – greatly expanding the current genome compendium of Picornavirales. Finally, statistical co-occurrence examinations of the dsDNA, ssRNA and dsRNA virus markers with picoeukaryotic diversity within the RNA-seq data revealed a wealth of potential virus-host relationships (a putative "*who infects whom*") that include known as well as novel interactions. These findings validate lab-observed relationships as well as define new potential interactions between viruses and hosts in marine surface waters and establish a mechanism to begin to address active virus infection at the community level.

Introduction

Viruses that infect marine microbes are an integral component of aquatic ecosystems with a diversity spectrum spanning the entire Baltimore classification scheme (Breitbart, 2012). The potential association of viruses with global-scale biogeochemistry, bloom termination events and the constraints they place on microbial community diversity has driven scientific interest in virus ecology (Weitz & Wilhelm, 2012, Brum *et al.*, 2015). Amongst these, giant dsDNA viruses belonging to the Nucleocytoplasmic Large DNA Virus (NCLDV) group infect protists with diverse lifestyles (Short, 2012) and are thought to be abundant in the world's oceans (Ogata *et al.*, 2011). Individually some of these viruses have been shown to be potential drivers of algal bloom collapse, with *Emiliania huxleyi* virus (EhV) and *Aureococcus anophagefferens* virus (AaV) being prominent examples (Gastrich *et al.*, 2004, Lehahn *et al.*, 2014). However, despite their possible importance, only a few ecologically-relevant host-NCLDV systems are established in culture and little is known about their ecology. As a specific example, viruses infecting obligate heterotrophs in the candidate Mimiviridae family are often isolated using *Acanthamoeba* as a 'tool', while the true ecologically relevant host remains unknown (Arslan *et al.*, 2011). Indeed, CroV is the only heterotrophic picoeukaryote infecting virus in this group for which the environmental host is known (Fischer *et al.*, 2010).

The term 'giant virus' is used as a catch-all to describe viruses that are incredibly diverse in size - some of them falling within the size class of small bacteria (Moniruzzaman *et al.*, 2016, Wilhelm *et al.*, 2016). To this end, culture independent approaches for their study are challenging: "viromes" – large metagenomic datasets enriched with viral sequences, are usually generated on samples filtered through ≤ 0.22 or 0.45 µM pore-size filters to exclude bacteria and small eukaryotes (Brum *et al.*, 2015). This approach however largely excludes giant viruses that range from 100 nM to ~1.5 µM (Abergel *et al.*, 2015). Additionally, the large genomes of these viruses often contain genes acquired from putative hosts, bacteria and other sources (Filee *et al.*, 2008), forwarding the question 'What can be considered truly viral?' in metagenomics analyses. Thus there is a need for new toolsets to complement (and perhaps steer) culture-based isolation approaches that can overcome these collective aforementioned issues.

The marine virosphere is not limited to large dsDNA viruses alone, as viruses containing all nucleic acid types (ss- and dsRNA as well as ssDNA) have been described that infect some form of marine picoeukaryotes (Lang *et al.*, 2009, Labonte & Suttle, 2013). It has been suggested that RNA viruses comprise a major fraction of the marine viroplankton and may infect ecologically important organisms ranging from diatoms and dinoflagellates to fish (Lang *et al.*, 2009). Yet methods for virus enumeration have missed these as nucleic acid stains (for enumeration) and DNA-based metagenomics approaches

miss RNA viruses. (Steward *et al.*, 2013). Indeed, little is known about RNA virus diversity and distribution – the first RNA virus infecting a marine protist was only discovered in 2003 (Tai *et al.*, 2003), and the high diversity of ssDNA viruses was only recently described (Labonte & Suttle, 2013). Collectively these observations illustrate that we have barely scratched the surface of the marine virosphere in terms of host-range and phylodynamics with respect to viruses infecting eukaryotes.

We examined metatranscriptomes of marine picoeukaryotes from two highly productive marine sites on the east coast of USA – Quantuck Bay and Narragansett Bay. Quantuck Bay experiences recurring Brown Tide blooms caused by the pelagophyte *Aureococcus anophagefferens* (Gobler *et al.*, 2011) which are shaped by a giant virus (AaV) (Gastrich *et al.*, 1998, Gastrich *et al.*, 2004, Gobler *et al.*, 2004). In contrast, Narragansett Bay is a highly productive system with seasonal diatom blooms, but little information about the eukaryotic virus community in this system is available. Within the context of these microbial communities we set out to extract and study signature genes specific to the double-stranded DNA giant viruses that exist as evidence of ongoing and active infection. Subsequently we applied this approach within these metatranscriptomes for viruses infecting eukaryotes with different nucleic acid genomes (ssDNA, ssRNA and dsRNA) as well as to the host community itself, resulting in emergent relationships of putative virus-host pairings. Our observations demonstrate the diversity, dynamics and genomic landscape of a large cross-section of eukaryotic plankton-infecting viruses in these systems as well as provide insight into the active plankton members that may be serving as potential hosts. We demonstrate that such information can be a powerful tool to link viruses to their potential hosts *in situ* in the context of the interactive network among eukaryotic microorganisms and viruses.

Materials and Methods

Experimental design

Quantuck Bay: The detail on sample collection and processing from the Quantuck Bay study site is described in Wurch et al (2016). Briefly, samples were collected from a bloom in Quantuck Bay (Latitude = 40.806395; Longitude = -72.621002), NY that occurred from late May to early July, 2011 covering the initiation, peak and demise of the bloom . *Aureococcus* cell were counted from glutaraldehyde (1% final v/v) fixed whole water samples using an enzyme-linked immunosorbent assay (ELISA) with a monoclonal antibody as described previously (Koch *et al.*, 2013). Planktonic chlorophyll *a* was collected and measured fluorometrically (Welschmeyer, 1994) on 0.2 and 5 µm filters. *In situ* samples from June 22nd (3rd sampling point) was also used to carry out nutrient amendment experiments. Briefly, bottles

were filled with natural sea water from the bloom and were amended with 25 µM ammonium only (+N), 4 µM phosphate only (+P), and 25 µM ammonium and 4 µM phosphate (+N&P) in triplicate. Three additional bottles with no nutrient addition were used as control. The samples were then incubated for 24 hours in a floating chamber at 0.5 m in eastern Shinnecock Bay at the Stony Brook - Southampton Marine Science Center under one layer of neutral density cover to mimic the light and temperature levels of Quantuck Bay. Samples for *Aureococcus* cell density measurement and total RNA extraction were collected at T=0 and T=24 hours.

Narragansett Bay: The detail sampling procedure is described in Alexander *et al* (2015). Briefly, samples were collected from a long term sampling site in Narragansett Bay (41°34′12″ N, 71°23′24″ W) during 2012 on May 16 (NB-S1), May 21 (NB-S2), May 30 (NB-S3), June 4 (NB-S4) and June 8 (NB-S5). Sample collection and processing was completed within 0830 and 0900 local time to reduce the influence of diel signals. 2.0 L of water from each sample was filtered on 5.0-µm pore size PC filters using a peristaltic pump. The filters were snap frozen in liquid nitrogen and stored at -80^0 C until RNA extraction. Water collected along with NB-S3 was also used for nutrient amendment experiments. For this, triplicate 2.5L bottles were filled with water pre-filtered through a 200-µm mesh and amended with specific nutrients to create +N, +P, -N, -P treatments alongside an ambient control. The +N and +P treatments were designed to eliminate nitrogen and phosphate stress signals, whereas the –N and –P treatments were supplemented with everything except the nutrient in question. N and P amendment concentrations were ~10-fold the seasonal average N and P concentrations measured at the station II in the surface waters of Narragansett Bay. The +P and +N amendment contained 3 µM phosphate and 10 µM nitrate, respectively. The –P amendment contained 10 µM nitrate, 68 µM silica, 4.6 µM iron and f/5 vitamins. The –N treatment was amended with 3 µM phosphate, 68 µM silica, 4.6 µM iron and f/5 vitamins. The f/5 media ratios (Guillard, 1975) were followed for silica, iron and vitamin amendments. Bottles were incubated for 48 hours in a flow-through incubator at ambient temperature and photosynthetically active radiation. After the end of the incubation, treated and control samples were filtered and stored for RNA extraction in the same manner for the *in situ* samples.

RNA extraction and sequencing

For the Quantuck Bay study site, approximately 25 ml of natural seawater from each of the *in situ* and nutrient amendment samples were pre-filtered through 5 µm polycarbonate (PC) filters and collected on 0.2 µm PC filters. The samples were flash frozen immediately after filtration and transferred to -80^0 C. Prior RNA extraction, CTAB buffer (Teknova, CA, USA) amended by polyvinylpyrrolidone (1%

mass/vol) was added to each of the samples. UltraClean® Plant RNA Isolation Kit (MO BIO Laboratories, CA, USA) was used to extract the RNA according to manufacturer's protocol which was quantified spectrophotometrically. RNA samples were sequenced in the Columbia Sequencing Center (NY, USA) using Illumina™ HiSeq™ platform with poly-A enrichment at a depth of 50 million 100bp single end reads. Two more replicate samples were sequenced from June 22 (QB-S3) at a depth of 100 million reads (Table 5.1).

For Narragansett Bay, replicate filters from each treatment and *in situ* samples were pooled, representing 6 L of water for each sample. RNA was extracted using RNeasy Mini Kit (Qiagen, Hilden, Germany) according to a modified yeast RNA extraction protocol. Briefly, lysis buffer and RNA-clean zircon beads were added to the filter. Samples were then vortexed for 1 min, placed on ice for 30 s, and then vortexed again for 1 min. The resulting RNA was eluted in water and possible DNA contamination was removed using a TURBO DNA-free Kit (Thermo Fisher Scientific, MA, USA). RNA from each triplicate was pooled by sample or treatment. >1,000 ng RNA from each sample then went through a poly-A selection using oligo-dT beads followed by library preparation with TruSeq RNA Prep Kit (Illumina, CA, USA). The samples were sequenced with an Illumina™ HiSeq2000™ at the Columbia University Genome Center to produce ~60 million; 100bp paired-end reads per sample (Table 5.1).

Read assembly and screening for virus and eukaryote specific contigs

Sequence reads from both locations were quality trimmed (stringent trimming (quality score ≤0.03), No 'N's allowed, 70bp size cutoff) in CLC genomics workbench 8.0 (Qiagen, Hilden, Germany). Separate assemblies were performed on all the site specific *in situ* samples together, which resulted in 2,455,926 contigs for Quantuck Bay and 9,525,233 contigs for Narragansett Bay at a 100bp size cut-off.

For selecting contigs specific to Major Capsid proteins of NCLDV, a HMM profile was created after aligning the MCP sequences from complete giant virus genomes and several reported MCP genes available in NCBI. The HMM profile was queried against the translated contig libraries to select the putative MCP candidate contigs using HMMER (Eddy, 2011). For selecting eukaryotic RPB1 contigs, HMM profile specific to domain "RPB1-C-term (NCBI CDD ID: cd02584)" and "RPB1-N-term (NCBI CDD ID: cd02733)" was used to query the contig libraries. All the MCP and RPB1 candidate contigs detected in this manner were queried against NCBI Refseq database and only contigs with first BLASTx hits (E-value cut-off $\leq 10^{-3}$) to MCP and RPB1 were kept for further analysis.

Table 5.1: RNA–seq library sizes after quality trimming. The Quantuck Bay samples were sequenced as single reads, while paired-end sequencing was carried out for the Narragansett Bay samples (denoted by X2).

	Sample	Number of reads
Quantuck Bay	June 14, 2011	43,755,618
	June 16, 2011	40,228,517
	June 22, 2011	214,681,748
	Nutrient amendment-control	41,156,535
	+Nitrogen	39,539,666
	+Phosphorus	36,797,290
	+Nitrogen and +Phosphorus	37,442,936
Narragansett Bay	16-05-2012	88,207,422 X2
	21-05-2012	53,911,567 X2
	30-05-2012	68,734,805 X2
	4-06-2012	45,344,970 X2
	8-06-2012	55,025,398 X2
	Nutrient amendment - control	38,713,917 X2
	-Phosphorus	45,764,698 X2
	+Phosphorus	37,528,376 X2
	-Nitrogen	36,671,310 X2
	+Nitrogen	49,217,942 X2

To detect contigs originating from viruses other than NCLDVs, we combined the proteins derived from all the viruses having algal, fungal and protozan hosts available on NCBI database. This protein database was queried against the contig libraries using tBLASTn with an E-value cut-off of $\leq 10^{-3}$. All the candidate contigs screened by this procedure were then queried against NCBI Refseq database using BLASTx. Only contigs having topmost hits to different viruses were kept for further analysis. All these contigs had best hits to diverse eukaryotic viruses, which is probably due to the fact that the samples were poly-A selected prior to sequencing. These virus contigs were binned into distinct viral groups according to their best BLASTx hits. Percentage of reads recruited to individual viral groups was calculated for determining proportional abundance of different viral groups over different time points. For detailed phylogenetic analysis of ssRNA and ssDNA viruses, subset of these contigs harboring RdRP (pfam id: PF05183) and viral replicase (pfam: PF03090) motif were selected using HMM profile specific to these motifs.

Genomic and phylogenetic analysis

Reference sequences for MCP (giant viruses), RdRP (RNA viruses), viral replicase (ssDNA virus) and RPB1 (eukaryotes) were downloaded from NCBI Refseq database. A number of RPB1 sequences representing several eukaryotic groups were also collected from MMETSP (Keeling *et al.*, 2014) peptide collections, for which no representative genomes are available in NCBI yet. The reference sequences were aligned in MEGA 6.0 (Tamura *et al.*, 2013). Maximum likelihood phylogenetic trees were constructed in PhyML (Guindon *et al.*, 2010) with LG model, gamma shape parameter and frequency type estimated from the data. aLRT SH-like statistic was calculated for branch support. The eukaryotic classification scheme by Adl et al (2005) was followed.

Selected contigs were translated to amino acid sequences and were placed on the reference trees in a maximum likelihood framework using pplacer (Matsen *et al.*, 2010). The placement files were converted to trees with pendant edges showing the best placement of the contigs using 'guppy tool of pplacer. The placement trees were visualized and annotated using iTOL interface (Letunic & Bork, 2016).

ORFs were predicted on the complete or near-complete picornavirales genomes using CLC genomic workbench 8.0 (www.clcbio.com). The genome annotation with predicted features was assisted with pfam (Punta *et al.*, 2012) and Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2013) search.

Statistical analysis

Quality trimmed reads were mapped to the selected viral and eukaryotic contigs from individual read libraries with high stringency (97% identify, 70% length fraction matching) in CLC genomics workbench 8.0. The read mapping values were normalized by library size and length and expressed as rarefied counts

per kilobase (RCK). For calculating the relative proportion of individual viral phylogenetic groups, number of reads mapped to the contigs from a specific viral group was summed and normalized by library size.

RCK values of viral and eukaryotic contigs upto 225 base pairs were converted into matrices separately for Quantuck Bay and Narragansett Bay datasets, which included mapping statistics from both *in situ* and nutrient amendment libraries. Group averaged hierarchical clustering was performed on these matrices using Pearson's correlation coefficient in PRIMER 7.0. SIMPROF test (Clarke *et al.*, 2008) was applied on the clusters with 5% significance level and 1000 permutations to identify statistically distinct clusters. Selected clusters were visualized and annotated in Cytoscape 3.0 (Cline *et al.*, 2007).

Results and discussion

Sample collection and data processing

Quantuck Bay (NY, USA) experienced a massive brown tide bloom from late May to early July in 2011 (Wurch et al, 2016). Samples collected on June 14 and June 16 represented the peak of the bloom, with no significant difference in *Aureococcus* cell concentrations between those dates (Figure 5.1). A third sample was collected on June 22 which represented the initial phase of bloom decline. Additionally, mesocoms (including a control and three nutrient amendments: nitrate (+N), phosphate (+P), nitrate and phosphate (+N&P) were carried out using water collected on June 22. The samples were pre-filtered through a 5-µM filter, followed by collection on a 0.2-µM nominal pore-size filter, limiting them to primarily bacteria and small eukaryotes. The second set of data was generated from a long-term monitoring site in Narragansett Bay (Alexander *et al.*, 2015). Five samples were sequenced during May and June of 2012 (NB-S1 to NB-S5) (Table 5.1). Samples were collected on filters with 5 µM nominal pore-size, thereby excluding any eukaryotic cells < 5 µM as well as most of the free bacteria and virus particles. Nutrient amendment experiments (control, +N, +P, -N, -P, +N&+P) were also performed using water collected on May 30 (NB-S3). Details about how these amendments were achieved are provided in materials and methods section. Samples were sequenced as previously described (Alexander *et al.*, 2015) (Wurch et al, 2016) after selection for poly-adenylated RNA sequences: this approach collected information from the microeukaryotic community as well as active infections (by dsDNA and ssDNA viruses) of this community. This approach also collected materials from cell-associated (either ongoing infections or potentially attached particles) for viruses with ssRNA and dsRNA genomes.

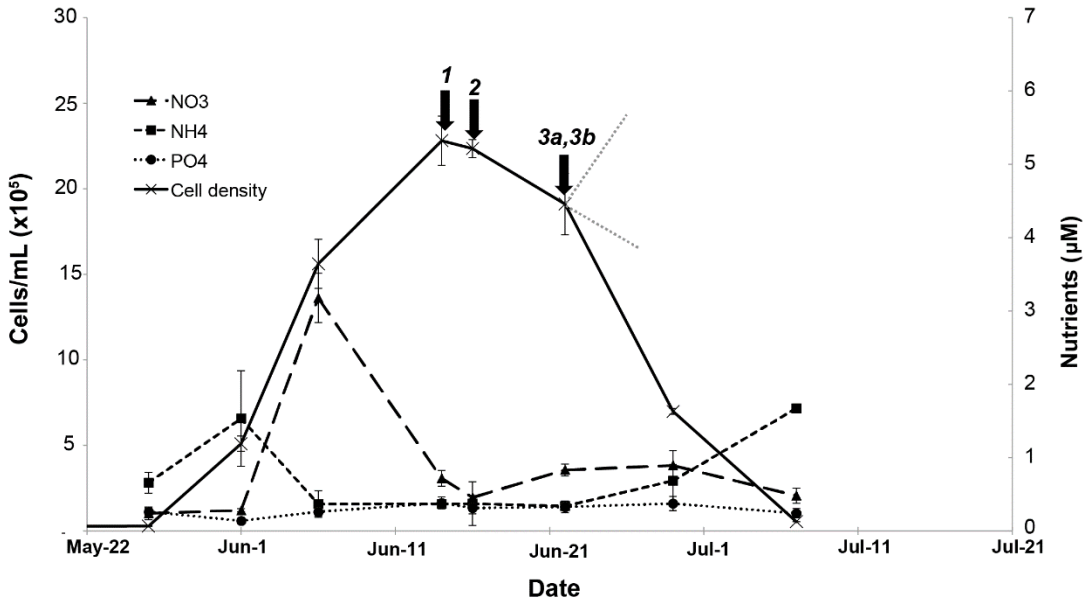


Figure 5.1: *Aureococcus* cell counts and nutrient dynamics during a 2011 brown tide bloom in Quantuck Bay, NY. The arrows represent the dates of sample collection. Reproduced from Wurch et al. (2016).

In total, ~500 million reads across 10 samples were generated from Narragansett Bay, and ~450 million reads from 7 samples were generated from Quantuck Bay study sites. Sequence reads from a particular study site were assembled together *de novo* to obtain a library of representative community transcripts (see materials and methods). Individual libraries were subsequently mapped to the virus and eukaryote specific contigs detected as described in the materials and methods section. Raw read counts were normalized by library size and contig length– giving a metric that we define as Rarefied Counts per Kilobase (RCK). This value was used to compare the expression level of contigs between different samples.

Temporal dynamics of active giant virus infections

To explore NCLDVs in metatranscriptomic datasets, we screened contig libraries for ten conserved core genes of NCLDVs (including major capsid protein) (Yutin *et al.*, 2009) and normalized count of reads assignable to these contigs from each of the libraries. At both sites, a higher number of contigs originating from NCLDV-specific Major Capsid Protein (MCP) can be identified compared to other potential candidates (Figure 5.2); indeed the abundance of reads mapped to MCP contigs is higher than the sum of specific reads to all other phylogenetic candidate gene contigs (Figure 5.2) for all samples except QB-S3 that confirms previous efforts suggesting MCP is a suitable marker for NCLDVs in molecular datasets (Moniruzzaman et al. 2016). A previous transcriptome study of Mimivirus demonstrated that MCP gene is expressed immediately after infection and is one of the top 20 most expressed genes (Legendre *et al.*, 2010). Interestingly, no other core genes made this list in Mimivirus. Additionally, the most enriched virus-specific biological function was viral capsid after 24 hours of viral infection of *Emiliania huxleyi* (Rosenwasser *et al.*, 2014). Thus, the higher abundance of MCP specific fragments in our datasets is consistent with these observations. Although only distant homologs of MCP are present in *Poxviridae* (Yutin *et al.*, 2009) and there are no homologs in recently discovered Pandora- and Pithoviruses (Abergel *et al.*, 2015), ubiquity of this gene in all other NCLDV families makes it a suitable candidate for phylogenetic probing of metatrancriptomics data.

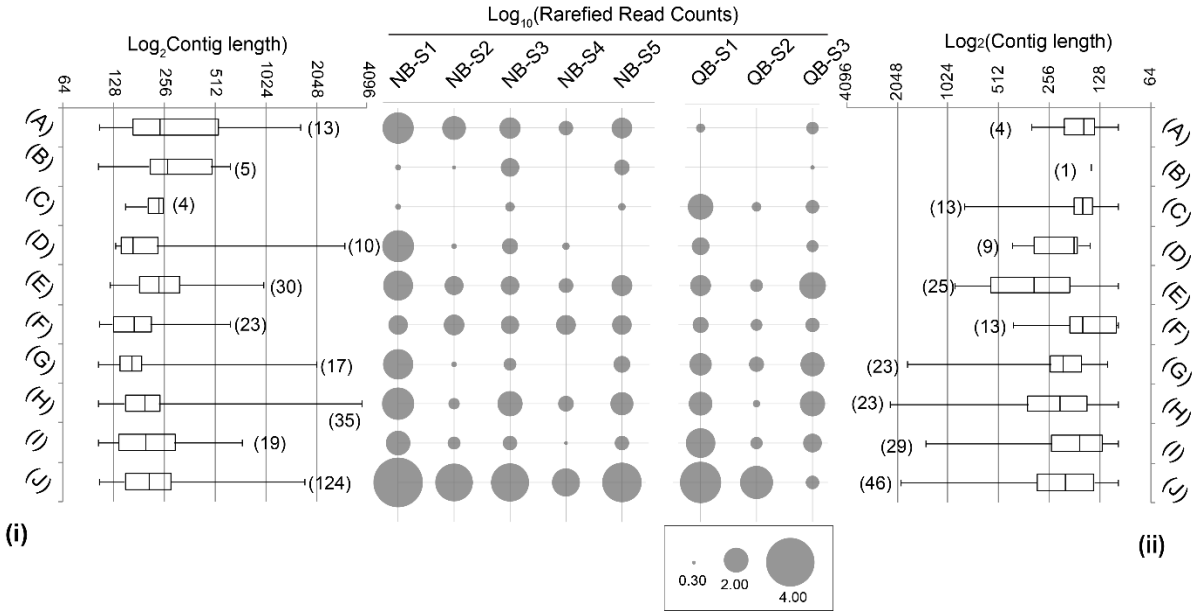


Figure 5.2: Abundance of 9 NCLDV core genes including major capsid protein (MCP) in terms of normalized read counts and number of contigs recovered (up to 100bp length) from Quantuck Bay (right panel) and Narragansett Bay (left panel). The box and whisker plots represent the range of the contig lengths with number of contigs recovered for each gene in brackets. The filled circles represent the rarefied abundances of each contig in each sample. No contigs could be detected from myristolyated envelope protein, a core NCLDV gene. Core genes are indicated on the X-axes as follows: A) A32 virion packaging ATPase, B) VLFT3 like transcription factor, C) Superfamily II helicase II, D) mRNA capping enzyme, E) D5 helicase/primase, F) Ribonucleotide reductase small subunit, G) RNA polymerase large subunit, H) RNA polymerase small subunit, I) B-family DNA polymerase, J) Major capsid protein.

The data provides unprecedented insight into the diversity of active NCLDV infections. Several observations are common to both study sites, irrespective of sampling time. Firstly, based on the phylogenetic placement of the contigs, NCLDV members from different families are consistently present, with highest number of contigs placed within the *Mimiviridae* family, followed by *Phycodnaviridae*, from both study sites (Figure 5.3 (A &B)). A large number of contigs have strong phylogenetic affinities to AaV as well as other algae-infecting members of the *Mimiviridae* clade. Very limited information on the ecological dynamics of the Mimiviridae members infecting putative algal hosts is available (Moniruzzaman *et al.*, 2016), with almost no ecological information on the *Mimiviridae* members infecting non-algal hosts. Our data indicate that these viruses are an important component of the marine virosphere and are as active as the better-studied *Phycodnaviridae* group. Interestingly, three contigs placed in the *Asfarviridae* clade recruit ~7% (BT-S1) to ~21% (BT-S2) of the MCP reads in Quantuck Bay samples (Figure 5.3A). The only known *Asfarviridae*-like NCLDV isolated from ocean infects a dinoflagellate (Ogata *et al.*, 2009). However, recently NCLDVs related to *Asfarviridae* have also been reported to infect *Vermamoeba*, a free-living amoebae (Reteno *et al.*, 2015). In the Quantuck Bay datasets, we detect active infection of AaV during the bloom - with high stringency (read similarity ≥ 97%), 1368 and 604 reads can be mapped across the AaV genome from QB-S1 and S2 respectively after library size normalization, while 236 reads are mapped to the QB-S3 bloom decline sample (Figure 5.4). Across the entire genome, 15 AaV genes have more than 10 reads, although two thirds these have no similarity to genes with known functions (ORFans). Highly expressed ORFans have also been recorded for Mimivirus: 17 of the top 20 most highly expressed genes were hypothetical genes (Legendre *et al.*, 2010). These observations indicate that such genes are important in propagation of AaV and other NCLDVs, despite our lack of knowledge addressing their function. However, some known genes such as AaV's MCP is also among the highly expressed functional genes, with 121 total reads mapped to this gene across the three *in situ* samples from Quantuck Bay. This observation is consistent with the expression pattern of MCP in the cultured based transcriptomics experiment described in Chapter III. Both total reads mapped across the AaV genome (Figure 5.4) and specifically to MCP gene (Figure 5.3A) progressively decline throughout the sampling period, with least number of reads mapped from S3. This might indicate that AaV activity was present, but reduced during the bloom decline stage - an observation that is supported by a recent study where AaV amplicons were only detected during the peak of the bloom (Moniruzzaman *et al.*, 2016).

Figure 5.3 (A & B): Phylogenetic placement of major capsid protein contigs from Quantuck Bay (prefix: Q MC-) and Narragansett Bay (prefix N MC-) on a reference tree of NCLDVs with icosahedral capsids. Node support (aLRT-SH statistic)>50% are shown as dark circles. Contigs upto 200bp are shown, with their expression level (RCK values) in individual samples as a heatmap on the outer rings. The contig that represents the MCP of AaV is marked with a black arrow. The reference sequences are shown in bold italic typeface. Abbreviations: MsV-Marseillevirs, LauV: Lausannevirus, Ws Irido: Weisenia iridescent virus, SG Irido: Singapore Grouper iridescent virus, He Asco: Heliothis virescens ascovirus, AsfV: Asfarvirus, EhV86: Emiliania huxleyi virus 86, HaV01: Heterosigma akashiwo virus 01, PBCV1: Paramacium bursaria Chlorella virus 1, ATCV 1: Acanthocystis turfacea Chlorella virus 1, BpV1: Bathycoccus parsinos virus 1, MpV12T: Micromonas pusilla virus 12T, OlV1: Ostreococcus lucimarinus virus 1, AaV: Aureococcus anophagefferens virus, CeV: Chrysochromulina ericina virus, PpV: Phaeocystis pouchetii virus, PgV: Phaeocystis globosa virus, PoV: Pyramimonas orientalis virus, Mega: Megavirus chilensis, Moumou: Moumouvirus goulette, Mimi: Mimivirus, CroV: Cafeteria roenbergensis virus.

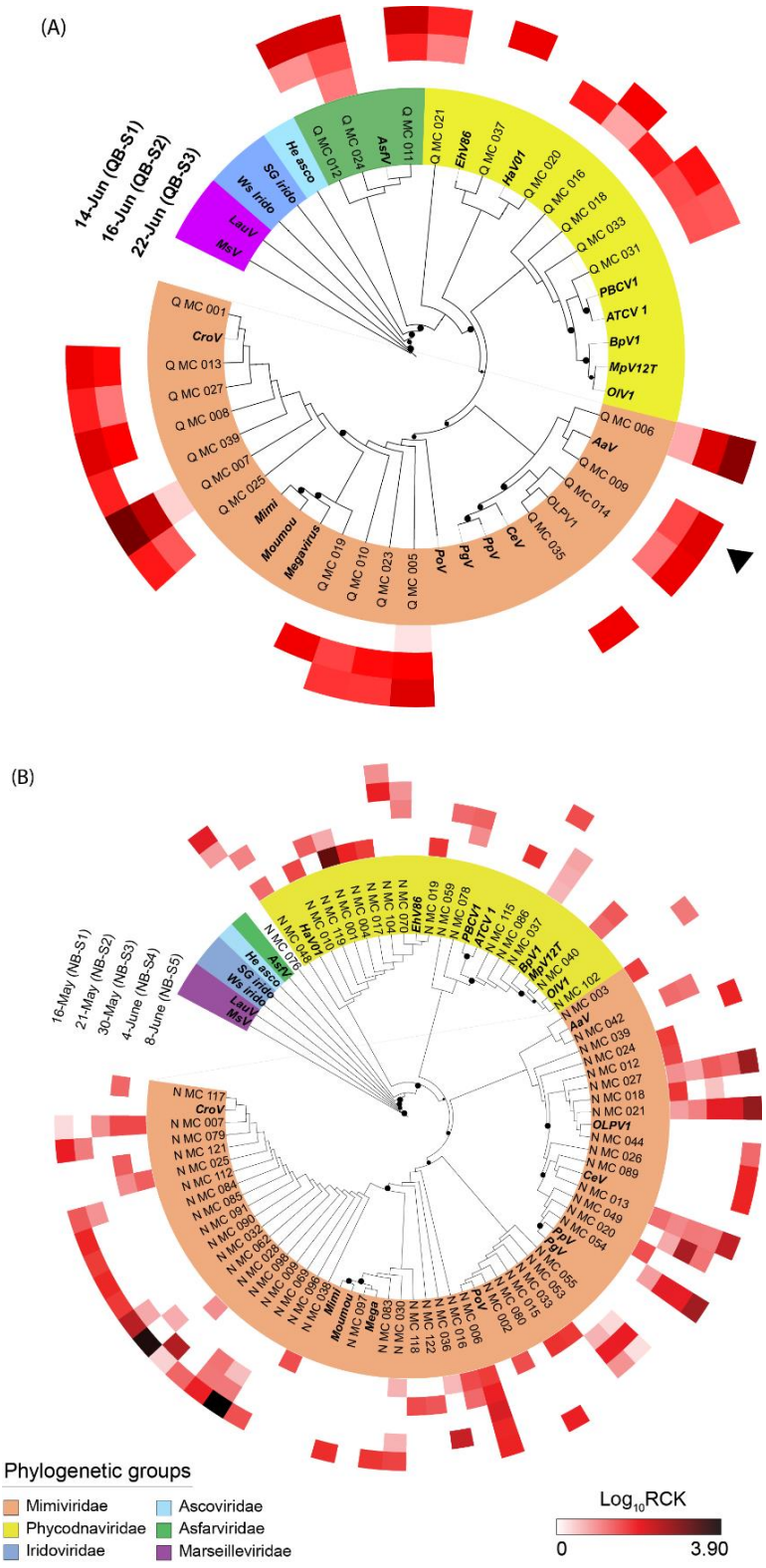


Figure 5.3. Continued.

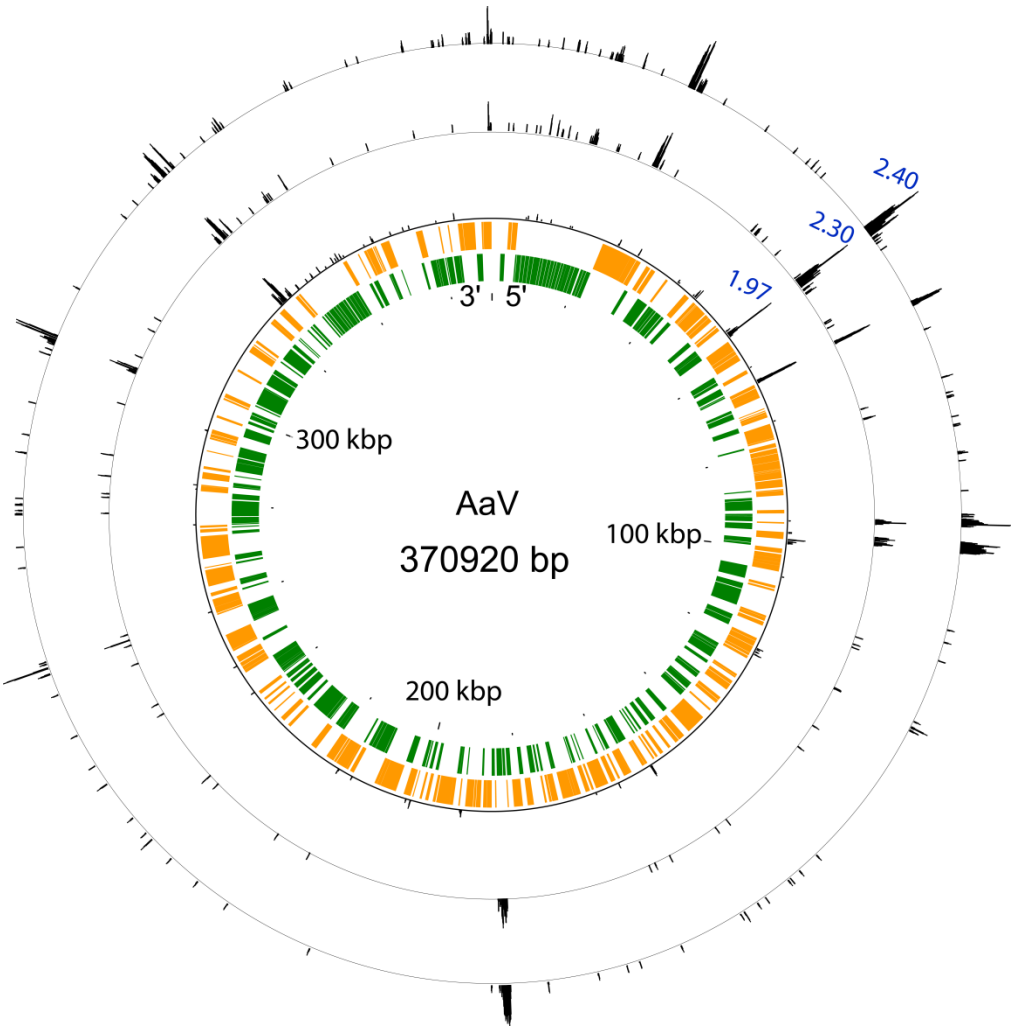


Figure 5.4: Normalized read mapping statistics to AaV genome from the three Quantuck Bay samples. From outermost to innermost ring: Log converted number of reads mapped from QB-S1 (June 14), QB-S2 (June 16) and QB-S3 (June 22), respectively. Forward CDSs are in green while reverse CDSs are shown in orange. The value in blue on each ring represents the highest number of reads (log converted) mapped from the particular sample.

With five *in situ* samples over a period of approx 4 weeks, data from Narragansett Bay allowed us to observe the dynamics of the NCLDVs across a broader temporal scale. Some members from Phycodna- and *Mimiviridae* clades show continuous activity over a prolonged period, while 'boom-bust' like relationships (Short, 2012) are possibly present for some other members (Figure 5.3B). For example, N_MC_006 and N_MC_021 have RCK values within a magnitude between samples over all the time points, an observation supporting presence of both the hosts and their viruses over a long time. While this scenario is consistent with a 'slow-and-steady' infection dynamics, it can also be explained by persistent infections of the plankton – where ongoing virus production doesn't necessarily lead to host (or at least total community) mortality (Floge, 2014). The expression values of other contigs, however, reflect a 'boom-and-bust' like scenario – with number of mapped reads varying over several orders of magnitude between time points. One striking example of such phenomenon is contig N_MC_069 in the non-algal *Mimiviridae* family, whose RCK value dropped by two orders of magnitude on May 21 and May 30 compared to May 16 (Figure 5.3 B).

Diversity and dynamics of viruses infecting microeukaryotes beyond the NCLDVs

As part of our analyses, 579 and 599 contigs from Quantuck Bay and Narragansett Bay, were respectively assigned to viruses that are not classified as NCLDVs Our analyses suggest the majority of these contigs were from (+)ssRNA viruses. At both sites, the main contributors of the (+)ssRNA virus diversity were from a yet unassigned group of marine RNA viruses in the Picornavirales order (Culley *et al.*, 2003), contributing 62% of the total Quantuck Bay and 74% of the Narragansett Bay non-NCLDV virus contigs (Figure 5.5A, B). Viruses with known hosts in this group infect a number of diatoms (e.g., *Chaetoceros* sp., *Asterionellopsis glacialis* and *Rhizosolenia setigera)* (Lang *et al.*, 2009), and a marine fungoid protist (*Aurantiotrychium*) (Takao *et al.*, 2006). The closest phylogenetic relative of this group is *Marnaviridae*, which currently have only one member – HaRNAV, that infects marine raphidophyte *Heterosigma akashiwo* (Tai *et al.*, 2003). The second major group of (+)ssRNA viruses belonged to *Dicistroviridae* family, with 90 and 36 contigs from Quantuck Bay and Narragansett Bay, respectively (Figure 5.5A, B). Interestingly, the only dsRNA viruses detected in both locations were similar to viruses in the *Totiviridae*, *Partitiviridae* and *Hypoviridae* family – all of which are known viruses of Fungi (Wilson, 2012). This established the unique possibility that these viruses may be infecting fungi that are parasitic on algae, as have been proposed recently for samples collected in the Laurentian Great Lakes (Edgar *et al.*, 2016). While some ssDNA virus contigs from Quantuck Bay cluster with the *Nanoviridae* family, others from both locations donot form any definitive cluster with known circular DNA viruses, thus potentially representing novel circular DNA virus groups in the ocean (Figure 5.6). No (–) ssRNA virus contigs were detected.

Figure 5.5 (A &B): Phylogenetic placement of (+)ssRNA virus contigs harboring RdRP motifs from Quantuck Bay (prefix: Q_sR_OV_) and Narragansett Bay (prefix: N_OV_) on reference trees. Node support (aLRT-SH statistic)>50% are shown as dark circles. Contigs upto 225bp are shown, with their expression level (RCK values) in individual samples as a heatmap on the outer rings. The reference sequences are shown in bold italic typeface.

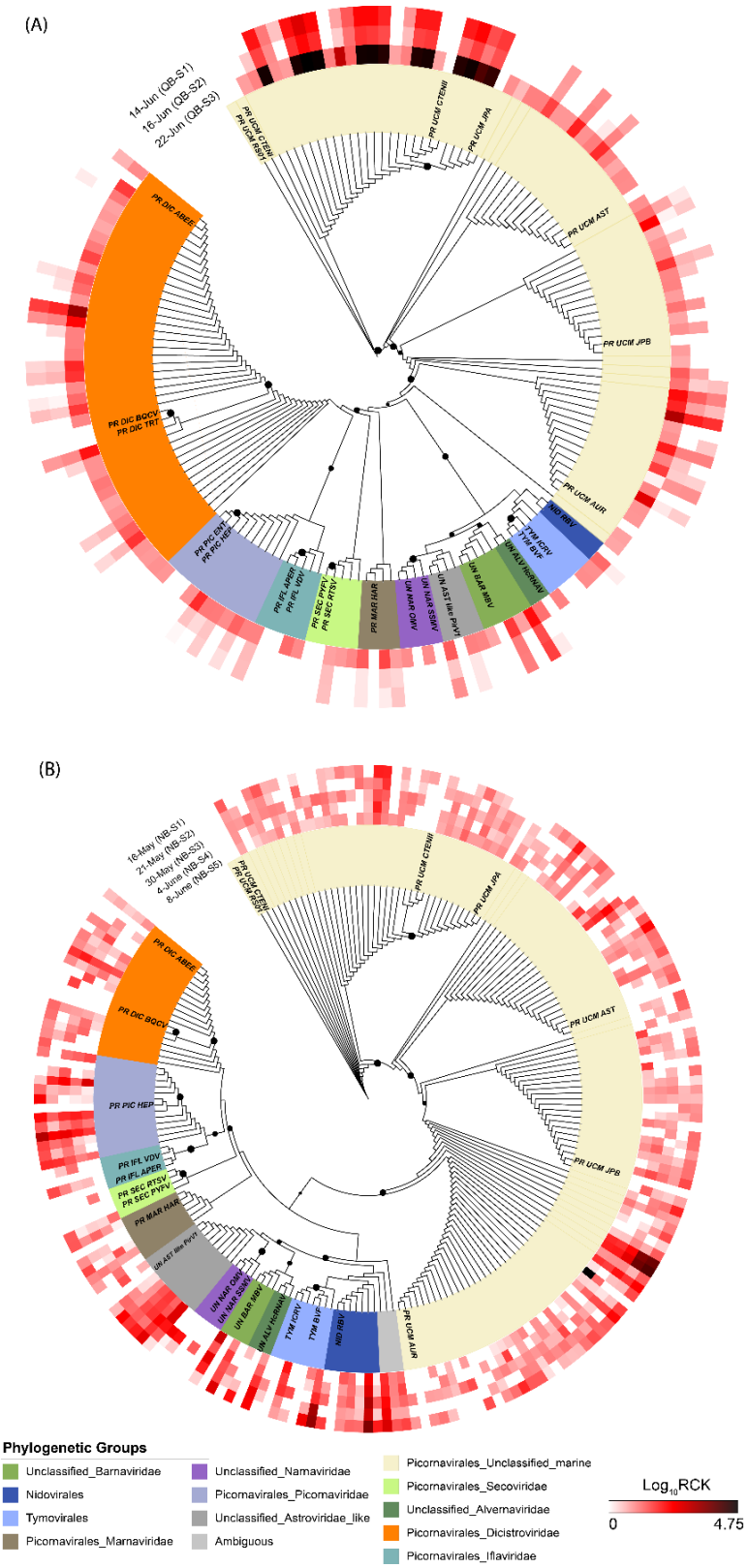


Figure 5.5 (A &B). Continued.

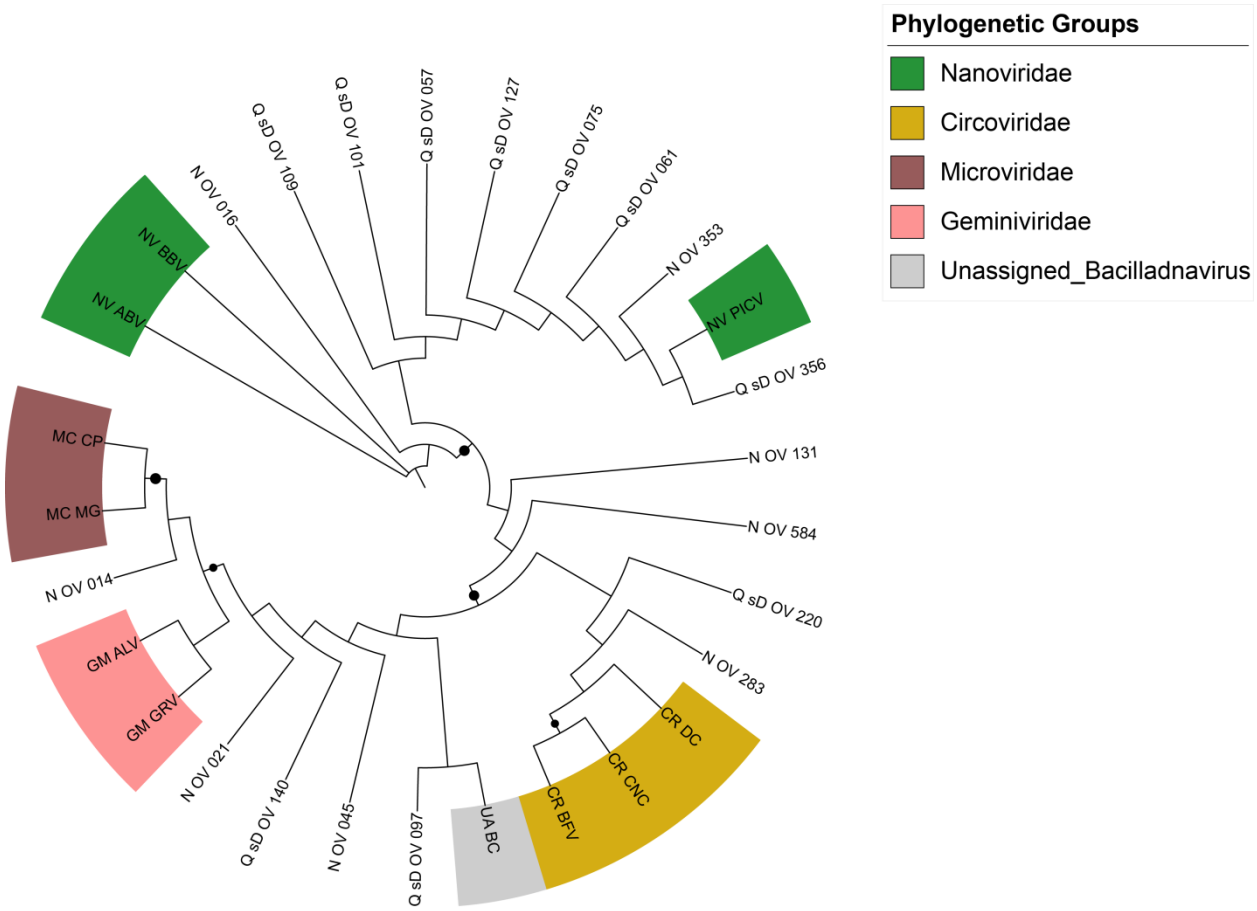


Figure 5.6: Phylogenetic placement of ssDNA virus contigs (prefix: N_OV_ and Q_sR_OV_) harboring viral replicase motifs from Quantuck Bay and Narragansett Bay on a reference tree. Node support (aLRT-SH statistic)>50% are shown as dark circles. Contigs upto 225bp are shown. The reference sequences are shown in bold italic typeface.

To assess how the activity of individual virus groups changed over time, we measured the proportion of reads that mapped to different virus groups from individual libraries. In the Narragansett Bay, the majority of the reads originate from unclassified marine Picornavirales, *Dicistroviridae*, *Secoviridae* and *Picornaviridae* families across all the time points (Figure 5.7). The unclassified marine Picornavirales group recruits anywhere from ~68% (NB-S1) to ~98% (NB-S5) of the reads specific to viruses other than NCLDVs. In Quantuck Bay, reads coming from both unclassified marine Picornavirales and ssDNA viruses dominate across the first two time points (Figure 5.8). However, a remarkable shift in the proportional abundance of viral reads is observed on the third day (QB-S3), when the unclassified marine Picornavirales became dominant, representing 93% of the total viral fragments other than NCLDVs (Figure 5.8). Overall, 2.4% of the entire QB-S3 library (~4.3 million fragments) can be mapped to these viral contigs, compared to 0.043% and 0.027% of reads from QB-S1 and QB-S2, respectively. This indicate a remarkable increase in activity of this group, concordant with the onset of Brown tide bloom decline. In the nutrient amendment samples (based on the water collected on the same date as QB-S3) these virus transcripts also increase by an order of magnitude relative to QB-S1 and QB-S2, which further validates the observed increase in viral activity in QB-S3. Phylogenetic analysis confirms these ssRNA virus contigs to be part of the unclassified marine picornavirales group (Figure 5.5A). The reason for an increase in the activity of these viruses concomitant with brown tide bloom decline is not known. *Aureococcus* blooms are not mono-specific – they include diatoms, dinoflagellates and high densities (~10^4 cells/ml) of heterotrophic protists alongside *Aureococcus*, even during the bloom peak (Sieracki *et al.*, 2004). Additionally, overall picoplankton abundance can be maintained at high densities, even during the course of bloom decline, likely reflecting the succession in picoeukaryote community (Moniruzzaman *et al.*, 2016). These observations lead to the hypothesis that virus infection in other eukaryotic plankton benefitting from the *Aureococcus* bloom decline resulted in the observed virus activity. Together, these observations suggest a much broader ecological role for viruses during phytoplankton bloom decline, a phenomenon to be investigated.

Studies examining RNA virus dynamics in the ocean are limited and major questions regarding their phylodynamics, host range and relative abundance remain unanswered. Our study confirms the findings of a number of studies revealing diverse Picornavirales phylogenetically distinct from the established families to be dominant members in several marine study sites (Steward *et al.*, 2013, Miranda *et al.*, 2016) demonstrating that this unassigned virus group is a major component of the marine virioplankton.

Marnaviridae	0.707	0.353	0.325	2.225	0.028
Other Picornavirales	0.234	0.099	0.171	0.304	0.004
Narnaviridae	0.904	0.105	0.330	0.969	0.045
Alphatetaviridae	0.881	0.277	0.495	2.521	0.090
Virgaviridae	0.032	0.127	0.413	1.971	0.146
Potyviridae	0.000	0.006	1.992	0.000	0.000
PirV-1 like	3.948	1.490	1.821	1.076	0.511
Fusariviridae	0.000	0.000	0.000	0.000	0.001
Alvarnaviridae	0.042	0.006	0.001	0.008	0.000
Totiviridae	0.017	0.108	0.003	0.074	0.008
Partitiviridae	0.002	0.095	1.016	0.074	0.017
ssDNA virus	2.350	3.963	1.484	5.206	0.300
DNA-RNA hybrid virus like	0.000	0.000	0.002	0.000	0.002
Bacillaridnavirus	0.065	0.115	0.431	0.895	0.048

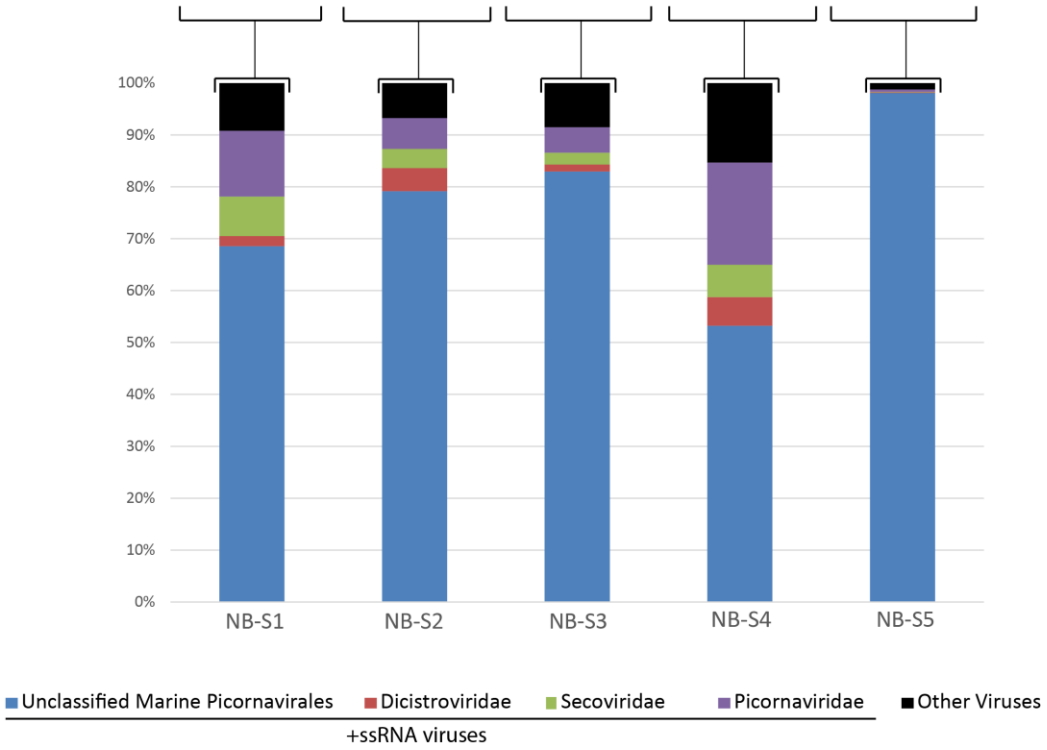


Figure 5.7: Proportional abundance of reads (in percentage) assigned to different viral groups in the Narragansett Bay study site over the five *in situ* samples. The 'other viruses' portion (in black) is further broken down into viral groups in the tables which represent a minor contribution towards the diversity.

Marnaviridae	0.075	0.176	0.025
Secoviridae	0.155	0.203	0.072
Picornaviridae	0.107	0.065	0.012
Other Picornavirales	0.767	1.007	2.824
Potyviridae	0.895	1.091	0.057
PirV-1 like	0.330	0.545	0.030
Fusariviridae	0.000	0.000	0.005
Alvarnaviridae	0.005	0.000	0.004
Tymovirales	0.000	0.000	0.000
Nodaviridae	0.000	0.000	0.000
Fisavirus like	0.032	0.083	0.016
Halastavi arva RNA virus like	0.037	0.203	0.042
Totiviridae	0.000	0.000	0.001
Partitiviridae	0.021	0.028	0.000
Hypoviridae	0.037	0.120	0.003
Unclassified Mycovirus	0.053	0.037	0.016

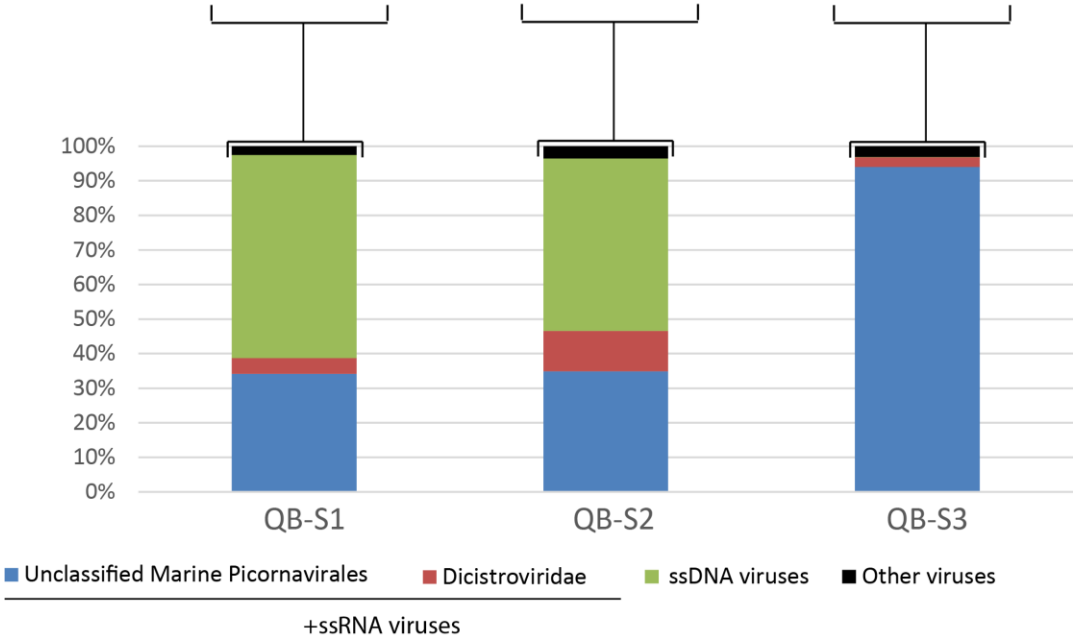


Figure 5.8: Proportional abundance of reads (in percentage) assigned to different viral groups in the Quantuck Bay study site over the three *in situ* samples. The 'other viruses' portion (in black) is further broken down into viral groups in the tables which represent a minor contribution towards the diversity.

Additionally, separate studies have shown ssDNA viruses, including *Circoviridae* members, in distinct oceanic regions (Rosario *et al.*, 2009, Labonte & Suttle, 2013). Owing to their small size, detection and quantification of both RNA viruses and ssDNA viruses pose significant technical challenges (Miranda *et al.*, 2016). Our results clearly point out the power of metatranscriptomic approach in this regard by allowing simultaneous analysis of the dynamics of DNA and RNA viruses. RNA viruses have a diverse size range, with Picornavirales members in a size range of ~25-30 nm (Wilson, 2012). Moreover, the filtration method applied to our samples (Material and Methods) allowed detection of ongoing virus infection (for DNA and RNA viruses) and cell-surface associated RNA viruses. Our poly-A selected metatranscriptomes enriched for viral sequences that are polyadenylated revealing active transcripts from DNA viruses (inside the host). It is important to mention that some (+)ssRNA viruses have poly-A tailed genomes (e.g., Picorna- and Togaviruses) even outside the host (Shatkin, 1974). Therefore, owing to their nature, the (+) ssRNA viral diversity captured by this approach might reflect both actively replicating and freely existing viruses.

Recovery of 18 novel complete or near-complete Picornavirales genomes

A survey of the available non-segmented Picornavirales genomes in the NCBI revealed that members have an average genome size of ~8,470 bp, with the maximum being 10,985 bp (NC_021566, *Iflaviridae*) and minimum being 6,580 bp (NC_01297, *Picornaviridae*). Nine of the assembled virus like contigs from each study site are >7,000 bp in length and have best hits to different Picornavirales members. Phylogeny and feature analysis of existing (+)ssRNA virus genomes suggest that these contigs are complete or near-complete Picornavirales genomes (Figure 5.9). A RNA dependent RNA polymerase gene based phylogeny places 12 of these genomes with the 'unclassified marine Picornavirales' group and 6 others within the *Dicistroviridae* family (Figure 5.9). Sixteen of these genomes are dicistronic – they harbor two ORFs coding for structural and non-structural proteins, while 2 are monocistronic (Figure 5.9). While all the genomes have the structural and non-structrual protein motifs characteristic of Picornavirales ORFs, a glycosyl transferase domain is found in N_OV_137 (Figure 5.9).

To our knowledge, the presence of glycosyltransferase domains has only been reported in members of the *Endornaviridae* family dsRNA viruses (Song *et al.*, 2013). The majority of the viruses in this group have dicistronic genomes, however, we did recover two monocistronic genomes (Q_sR_OV_022, Q_sR_OV_041), revealing differences in genome architecture among the members of this group (Figure 4). Remarkably, one of the viruses (N_OV_001) has a reverse orientation of the genes with the first ORF encoding for the structural protein, which is unusual for dicistronic viruses in the Picornavirales order with only one documented case for a *Dicistroviridae* member (Greninger & DeRisi, 2015).

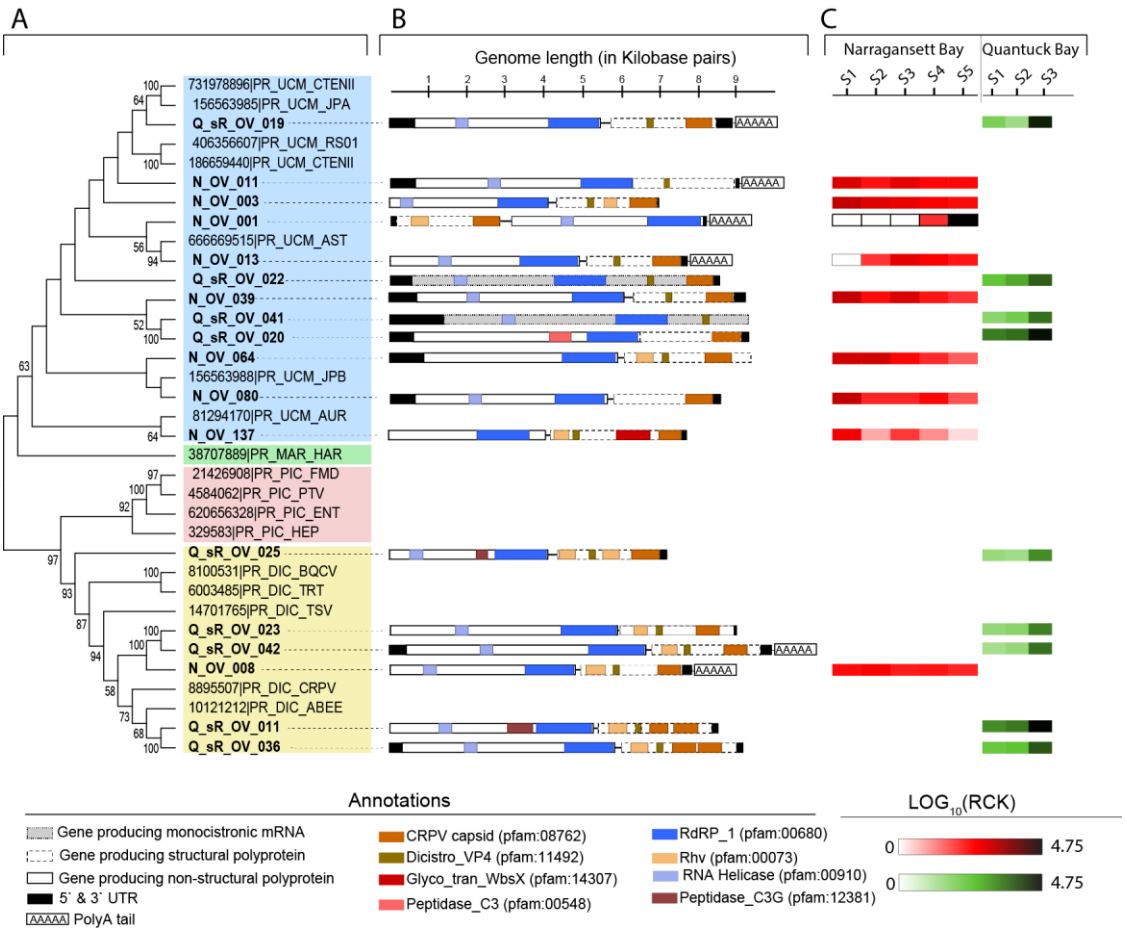


Figure 5.9: Complete or near-complete Picornavirales genomes recovered from both Quantuck Bay and Narragansett Bay study sites. Panel (A) shows the phylogenetic classification of these contigs in a topology-only maximum likelihood tree, with contigs from Quantuck Bay having prefix 'Q_sR_OV_' an contigs from Narragansett Bay having prefix 'N_OV_'. Panel (B) shows the genome architecture of these contigs with protein domains and putative CDSs. Panel (C) shows the activity of these viruses in terms of RCK values (Log_{10} converted) in across different *in situ* samples.

All the (+)ssRNA virus genome assemblies from Quantuck Bay samples had higher RCK values during bloom decline (QB-S3) compared to the earlier time points corresponding to the bloom peak (Figure 4). N_OV_001, a candidate virus from Narragansett Bay, was not present in the first three sampling points; however, it recruited 83% of the total viral fragments from the last sampling point (~0.55% of the entire library) (Figure 5.9). During the sampling period, Narragansett Bay was experiencing the spring diatom blooms. Concordant with this observation, up to 40% of the reads from the metatranscriptome libraries could be mapped to *Thalassiosira*, *Skeletonema* and other diatoms (Alexander *et al.*, 2015). The closest phylogenetic relative of N_OV_001 is a virus infecting diatom *Asterionellopsis glacialis* (Figure 5.9).

Who infects whom? - Resolving virus-host relationships using metatranscriptomics

The data presented an opportunity to generate hypotheses concerning ecologically meaningful relationships among these diverse protists and their pathogens, with the established *AaV-Aureococcus* association acting as a *de facto* internal standard. Transcripts from DNA viruses must originate within the host cells, and thereby for a particular host-virus pair, a significant and strong positive correlation can be expected for gene expression among viruses and their hosts. Building on this idea, host gene expression of at least a subset of the host's genome is a prerequisite to observe gene expression of a virus specific to that host, as evidenced by transcriptomic landscape of host-virus dynamics in culture (Legendre *et al.*, 2010, Rowe *et al.*, 2014) and induced blooms in mesocosms (Pagarete *et al.*, 2011). We inspected statistical co-occurrences among the contigs containing virus and eukaryote-specific marker genes based on their expressions. Since poly-A selected metatranscriptomes are largely depleted of ribosomal RNA marker genes, we opted for functional genes suitable for phylogenetic analysis. Expression of MCP (dsDNA viruses), RdRP (RNA viruses) and viral replicase (ssDNA viruses) were compared to the functional eukaryotic marker gene RNA polymerase II large subunit (RPB1). Hierarchical clustering of a Pearson's correlation matrix followed by a SIMPROF analysis (Clarke *et al.*, 2008) was used to detect statistically distinct clusters which contained both viral and eukaryotic marker genes that had been classified into phylogenetic groups by placing the putative host contigs on a RPB1 reference tree (Figure 5.10).

Our overall analysis produced a number of statistically distinct clusters harboring both viral and eukaryotic contigs (Figure 5.11). Interestingly, one cluster (Figure 5.11, A(ii)) harbored both AaV and *Aureococcus*, demonstrating that ecologically relevant relationships between viruses and their hosts can be discerned using transcriptome sequence data. Close inspection revealed interesting relationships among the coexisting eukaryotic and viral components. Cluster A(ii), while containing both *Aureococcus* and AaV, also harbored another Mimiviridiae member, several ssDNA and (+)ssRNA virus contigs along

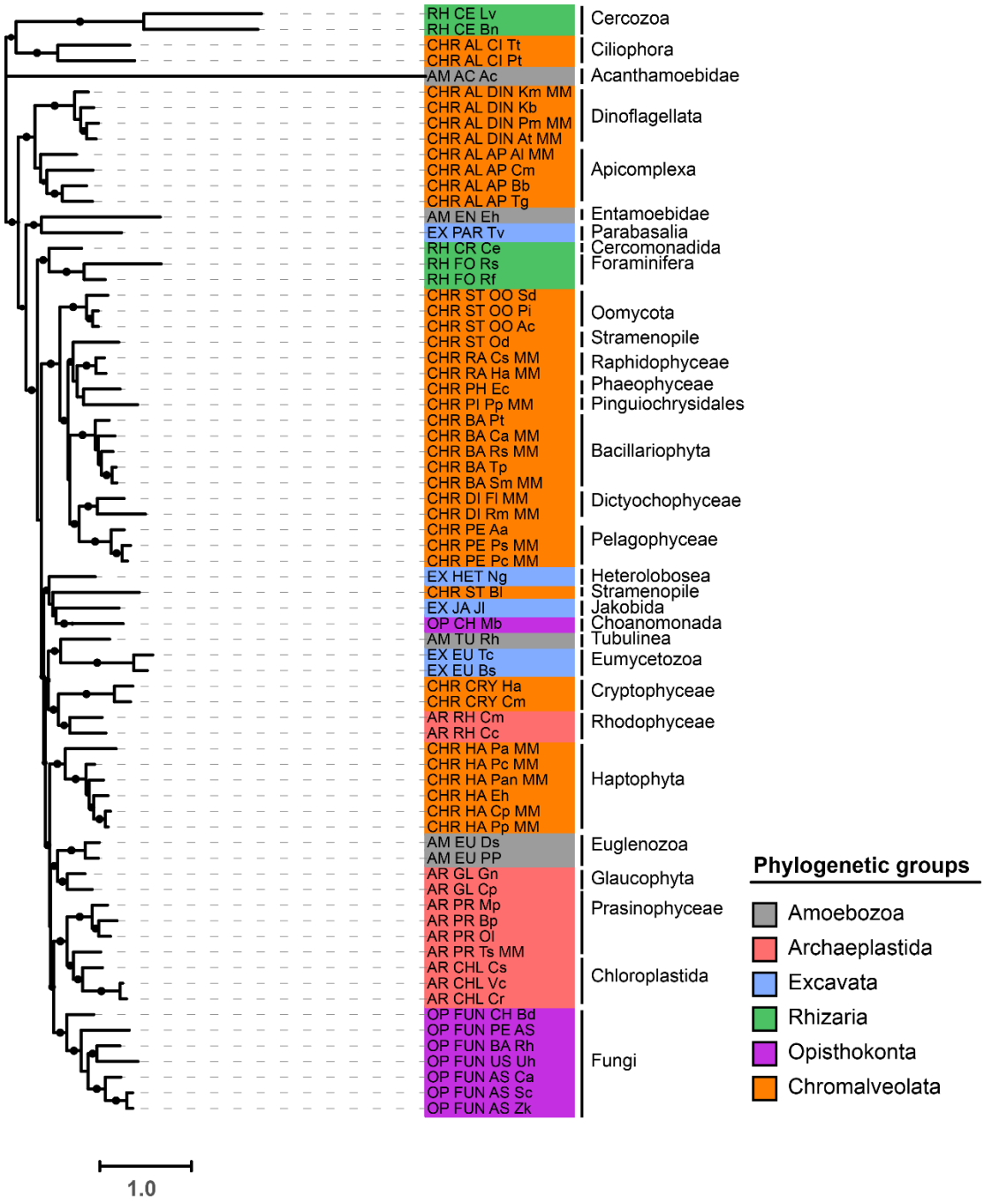


Figure 5.10: Maximum likelihood phylogenetic tree of eukaryotic RPB1 (RNA polymerase II large subunit) used as reference to classify the RPB1 contigs assisted by pplacer.

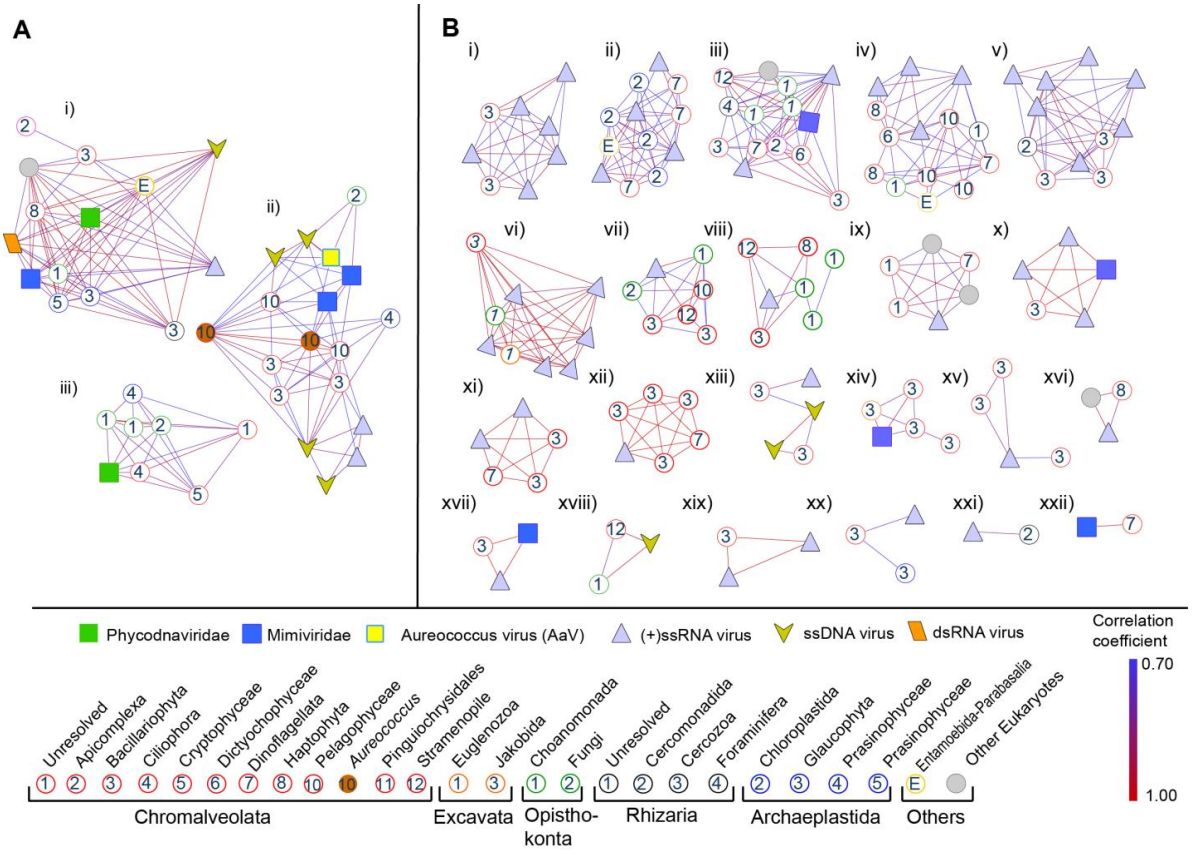


Figure 5.11: Representative SIMPROF clusters containing both viral and eukaryotic members from Quantuck Bay (Panel A) and Narragansett Bay (Panel B). Contigs are shown as nodes and the correlations as the connecting edges. Phylogenetic classification of the contigs are shown in the bottom panel. *Aureococcus* (dark brown circles) and AaV (bright yellow square) are in cluster A(ii).

with eukaryotes belonging to prasinophyceae and pelagophyceae (Figure 5.11). The possibility of *Aureococcus* being infected by more than on virus in nature cannot be discounted (and is perhaps likely). Moreover, the potential for AaV to infect closely related Pelagophytes remains a possibility (although this has not been seen in lab studies) (Gobler *et al.*, 2007). One cluster, A(i), which contains both a Phycodna- and a *Mimiviridae* member also includes a RPB1 contig phylogenetically placed in Cercozoa group (Figure 5.11). Interestingly, although no cercozoan host-NCLDV pairs currently exist in culture, a recent study showed integration of NCLDV genes in the genome of a cercozoan *Bigelowella natans* (Blanc *et al.*, 2015). This integrated NCLDV with the *B. natans* genome potentially belongs to *Phycodnaviridae*, as revealed by phylogenetic analysis of the MCP gene.

Similar clusters of phylogenetically distinct eukaryotes and viruses were also found in Narragansett Bay. For example, cluster B(iii) contained a *Mimiviridae* and several ssRNA virus contigs connected to a number of choanomonada, stramenopile, diatom and dinoflagellate members (Figure 5.11). The majority of the eukaryotic contigs belonged to diatoms followed by dinoflagellates in the Narragansett Bay sample, which reflects the protist composition of this site that experiences seasonal diatom blooms (Alexander *et al.*, 2015). A large number of contigs having phylogenetic affinity to choanomonada were found in both Quantuck bay and Narragansett Bay locations, and found in several of the representative SIMPROF clusters (Figure 5). While larger networks of viruses and eukaryotes existed, clusters with fewer members are important since they revealed more specific relationships. For example, cluster B(xiv) contained one *Mimiviridae*, one jakobida (heterotrophic flagellate) and several diatom contigs (Figure 5.11). To date the obligate heterotrophs known to be infected by *Mimiviridae* members are *Cafeteria roenbergensis* (Fischer et al., 2010), *Acanthamoeba* (Abrahão et al., 2014) and *Vermaamoeba* spp (Reteno *et al.*, 2015). Additionally; Cluster B(xviii) harbors a ssDNA virus, a stramenopile and a choanomonada member, while cluster B(xxii) revealed a one-to-one relationship between a *Mimiviridae* and a dinoflagellate (Figure 5.11). Only one dinoflagellate – *Heterocapsa circularisquama* , has been shownto be infected by a NCLDV(Kenji *et al.*, 2001). Cluster B(x) and B(xvii) consists of *Mimiviridae,* diatoms and ssRNA viruses. No diatom is known to be infected by a NCLDV yet, although a large number of ssRNA virus contigs in our study are phylogenetically closest to diatom infecting RNA viruses in the unclassified marine picornavirales group (Figure 5.5 (A& B)). A number of clusters (e.g., B(xii)) were enriched with both ssRNA virus and diatom contigs. However, relationship among the ssRNA viruses and the eukaryotes need to be interpreted with caution, since these contigs might originate both from free virus particles and/or within the hosts.

Several clusters contained fungal contigs along with other eukaryotes – pointing to the possibility of a parasitic relationship with phytoplanktons and other protists as discussed previously. The AaV-*Aureococcus* cluster (A(ii)) harbored a fungus contig and a *Barnaviridae* member – a family of virus with fungi as the only known hosts (Figure 5.11) (Wilson, 2012). Several other clusters, e.g., A(iii) and B(vii) also contained fungal contigs. While such observations are certainly not definitive, they might point to the existence of parasitic relationships resulting in complicated ecological interactions involving protists, fungi and fungal viruses in marine ecosystems.

Samples over longer time course will be necessary to draw more statistically robust relationships,further narrowing down the potential interacting partners. One limitation of reference independent assembly of high throughput data is fragmented contigs originating from same transcript – which is illustrated by two *Aureococcus* specific RPB1 contigs in cluster A(ii) that originated from a single coding sequence. Deeper sequencing and development of bioinformatics methods will be necessary to produce longer contigs. These limitations notwithstanding, our analysis provides a 'proof-of-principle' for inferring the complex relationships among diverse eukaryotic protists and their viruses using metatrsnscriptomics data.

Conclusion

In this study we demonstrate that metatranscriptomics can provide a comprehensive view of the marine virosphere by detecting key viral players active in an ecological context. Also, by simultaneously detecting viruses with distinct nucleic acids and strandedness, metatranscriptomics can largely overcome technical limitations in studying different virus groups owing to their size range and nucleic acid types. In the last two decades we have learned a lot about the diversity and dynamics of the phages in the ocean, but the eukaryotic virosphere remains elusive, with little known about who is infecting whom in the environment. Establishing ecologically relevant host-virus pairs in culture is an important way to understand this problem, which remains a formidable, yet necessary task for the huge diversity of microeukaryotes and their viruses. As demonstrated in our study, analyzing the vast wealth of information captured by metatranscriptomics in a statistical framework can be a right step towards answering this vital question.

Acknowledgements

The authors thank Gary LeCleir and Eric Gann for assistance in data collection and analyses.

Funding

This work was funded by grants from the National Science Foundation, the Gordon & Betty Moore Foundation (SWW) and the Kenneth & Blaire Mossman Endowment to the University of Tennessee (SWW).

List of References

Abergel C, Legendre M & Claverie J-M (2015) The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS microbiology reviews*. **39: 779-96.**

Abrahão JS, Dornas FP, Silva LC, Almeida GM, Boratto PV, Colson P, La Scola B & Kroon EG (2014) *Acanthamoeba polyphaga* Mimivirus and other giant viruses: an open field to outstanding discoveries. *Virology Journal* **11**:120.

Adl SM, Simpson AG, Farmer MA, *et al.* (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *The Journal of eukaryotic microbiology* **52**: 399-451.

Alexander H, Jenkins BD, Rynearson TA & Dyhrman ST (2015) Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proc Natl Acad Sci USA* **112**: E2182-E2190.

Arslan D, Legendre M, Seltzer V, Abergel C & Claverie J-M (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci USA* **108**: 17486-17491.

Blanc G, Gallot-Lavallee L & Maumus F (2015) Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proc Natl Acad Sci U S A* **112**: E5318-5326.

Breitbart M (2012) Marine viruses: truth or dare. *Ann Rev Mar Sci* **4**: 425-448.

Brum JR, Ignacio-Espinoza JC, Roux S, *et al.* (2015) Patterns and ecological drivers of ocean viral communities. *Science* **348**.

Clarke KR, Somerfield PJ & Gorley RN (2008) Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. *Journal of Experimental Marine Biology and Ecology* **366**: 56-69.

Cline MS, Smoot M, Cerami E, *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature protocols* **2**: 2366-2382.

Culley AI, Lang AS & Suttle CA (2003) High diversity of unknown picorna-like viruses in the sea. *Nature* **424**: 1054-1057.

Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS computational biology* **7**: e1002195.

Edgar RE, Morris PF, Rozmarynowycz MJ, D'Souza NA, Moniruzzaman M, Bourbonniere RA, Bullerjahn GS, Phuntmart V, Wilhelm SW & McKay RML (2016) Adaptations to photoautotrophy associated with seasonal ice cover in a large lake revealed by metatranscriptome analysis of a diatom bloom. *J Gt Lakes Res* **in press**. http://dx.doi.org/10.1016/j.jglr.2016.07.025

Filee J, Pouget N & Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC evolutionary biology* **8**: 320.

Fischer MG, Allen MJ, Wilson WH & Suttle CA (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci U S A* **107**: 19508-19513.

Floge, Sheri A., "Virus Infections of Eukaryotic Marine Microbes" (2014). *Electronic Theses and Dissertations.* Paper 2206. http://digitalcommons.library.umaine.edu/etd/2206

Gastrich M, Leigh-Bell J, Gobler C, Roger Anderson O, Wilhelm S & Bryan M (2004) Viruses as potential regulators of regional brown tide blooms caused by the alga,Aureococcus anophagefferens. *Estuaries* **27**: 112-119.

Gastrich MD, Anderson OR, Benmayor SS & Cosper EM (1998) Ultrastructural analysis of viral infection in the brown-tide alga, *Aureococcus anophagefferens* (Pelagophyceae). *Phycologia* **37**: 300-306.

Gastrich MD, Leigh-Bell JA, Gobler C, Anderson OR & Wilhelm SW (2004) Viruses as potential regulators of regional brown tide blooms caused by the alga, *Aureococcus anophagefferens*: a comparison of bloom years 1999-2000 and 2002. *Estuaries* **27**: 112-119.

Gobler CJ, Anderson OR, Gastrich MD & Wilhelm SW (2007) Ecological aspects of viral infection and lysis in the harmful brown tide alga *Aureococcus anophagefferens*. *Aquat Microb Ecol* **47**: 25-36.

Gobler CJ, Deonarine S, Leigh-Bell J, Gastrich MD, Anderson OR & Wilhelm SW (2004) Ecology of phytoplankton communities dominated by *Aureococcus anophagefferens*: The role of viruses, nutrients, and microzooplankton grazing. *Harmful Algae* **3**: 471-483.

Gobler CJ, Berry DL, Dyhrman ST, *et al.* (2011) Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *Proc Natl Acad Sci USA.* **108:**4352-4357.

Greninger AL & DeRisi JL (2015) Draft genome sequence of Laverivirus UC1, a Dicistrovirus-like RNA virus featuring an unusual genome organization. *Genome Announcements* **3**: e00656-00615.

Guillard RRL (1975) in Culture of Marine Invertebrate Animals, eds Smith WL, Chanley MH (Springer, Boston).

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W & Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307-321.

Keeling PJ, Burki F, Wilcox HM, *et al.* (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* **12**: e1001889.

Kenji T, Keizo N, Shigeru I & Mineo Y (2001) Isolation of a virus infecting the novel shellfish-killing dinoflagellate Heterocapsa circularisquama. *Aquatic Microbial Ecology* **23**: 103-111.

Koch F, Sañudo-Wilhelmy SA, Fisher NS & Gobler CJ (2013) Effect of vitamins B1 and B12 on bloom dynamics of the harmful brown tide alga, *Aureococcus anophagefferens* (Pelagophyceae). *Limnology and Oceanography* **58**: 1761-1774.

Labonte JM & Suttle CA (2013) Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME journal* **7**: 2169-2177.

Lang AS, Rise ML, Culley AI & Steward GF (2009) RNA viruses in the sea. *FEMS microbiology reviews* **33**: 295-323.

Legendre M, Audic S, Poirot O, *et al.* (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome research* **20**: 664-674.

Lehahn Y, Koren I, Schatz D, *et al.* (2014) Decoupling physical from biological processes to assess the impact of viruses on a mesoscale algal bloom. *Current Biology* **24**.

Letunic I & Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*. **44:**W242-W245.

Marchler-Bauer A, Zheng C, Chitsaz F, *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41**: D348-352.

Matsen FA, Kodner RB & Armbrust VE (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**: 1-16.

Miranda JA, Culley AI, Schvarcz CR & Steward GF (2016) RNA viruses as major contributors to Antarctic virioplankton. *Environ Microbiol*. DOI: 10.1111/1462-2920.13291

Moniruzzaman M, Gann ER, LeCleir GR, Kang Y, Gobler CJ & Wilhelm SW (2016) Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, *Aureococcus anophagefferens*. *FEMS Microbiology Ecology*. **92:** fiw058.

Ogata H, Monier A & Claverie J-M (2011) Distribution of Giant Viruses in Marine Environments. *Global Change: Mankind-Marine Environment Interactions,*(Ceccaldi H-J, Dekeyser I, Girault M & Stora G, eds.), pp. 157-162. Springer Netherlands.

Ogata H, Toyoda K, Tomaru Y, Nakayama N, Shirai Y, Claverie J-M & Nagasaki K (2009) Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virology journal* **6**: 1-8.

Pagarete A, Le Corguillé G, Tiwari B, Ogata H, de Vargas C, Wilson WH & Allen MJ (2011) Unveiling the transcriptional features associated with coccolithovirus infection of natural Emiliania huxleyi blooms. *FEMS Microbiology Ecology* **78**: 555-564.

Punta M, Coggill PC, Eberhardt RY, *et al.* (2012) The Pfam protein families database. *Nucleic Acids Research* **40**: D290-D301.

Reteno DG, Benamar S, Bou Khalil J, Andreani J, Armstrong N, Klose T, Rossmann M, Colson P, Raoult D & La Scola B (2015) Faustovirus, an Asfarvirus-related new lineage of giant viruses infecting amoebae. *Journal of virology*. **89**: 6585-6594.

Rosario K, Duffy S & Breitbart M (2009) Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of General Virology* **90**: 2418-2424.

Rosenwasser S, Mausz MA, Schatz D, *et al.* (2014) Rewiring Host Lipid Metabolism by Large Viruses Determines the Fate of Emiliania huxleyi, a Bloom-Forming Alga in the Ocean. *The Plant Cell Online* **26**: 2689-2707.

Rowe JM, Jeanniard A, Gurnon JR, Xia Y, Dunigan DD, Van Etten JL & Blanc G (2014) Global analysis of *Chlorella variabilis* NC64A mRNA profiles during the early phase of *Paramecium bursaria* Chlorella Virus-1 infection. *PLoS ONE* **9**: e90988.

Shatkin AJ (1974) Animal RNA Viruses: Genome Structure and Function. *Annual Review of Biochemistry* **43**: 643-665.
Short SM (2012) The ecology of viruses that infect eukaryotic algae. *Environmental Microbiology* **14**: 2253-2271.

Sieracki ME, Gobler CJ, Cucci TL, Thier EC, Gilg IC & Keller MD (2004) Pico- and nanoplankton dynamics during bloom initiation of *Aureococcus* in a Long Island, NY bay. *Harmful Algae* **3**: 459-470.

Song D, Cho WK, Park S-H, Jo Y & Kim K-H (2013) Evolution of and horizontal gene transfer in the Endornavirus genus. *PLoS ONE* **8**: e64270.

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M & Poisson G (2013) Are we missing half of the viruses in the ocean? *The ISME journal* **7**: 672-679.

Tai V, Lawrence JE, Lang AS, Chan AM, Culley AI & Suttle CA (2003) Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (Raphidophyceae). *Journal of Phycology* **39**: 343-352.

Takao Y, Mise K, Nagasaki K, Okuno T & Honda D (2006) Complete nucleotide sequence and genome organization of a single-stranded RNA virus infecting the marine fungoid protist Schizochytrium sp. *The Journal of general virology* **87**: 723-733.

Tamura K, Stecher G, Peterson D, Filipski A & Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**: 2725-2729.

Weitz JS & Wilhelm SW (2012) Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol Rep* **4**: 17.

Welschmeyer NA (1994) Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments. *Limnology and Oceanography* **39**: 1985-1992.

Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M & Stough JMA (2016) Standing on the shoulders of giant viruses: 5 lessons learned about large viruses infecting small eukaryotes and the opportunities they create. *PLoS Pathogens* **In press**. http://dx.doi.org/10.1371/journal.ppat.1005752

Wilson WH, Van Etten, J.L., Schroeder, D.C., Nagasaki, K., Brussaard, C., Bratbak, G., Suttle, C. (2012) Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. (Andrew M.Q. King EL, Michael J. Adams , Eric B. Carsten, ed.) pp. 1338. Elsevier London, UK.

Wurch LL, Alexander H, Frischkorn K, Haley ST, Gobler CJ and Dyhrman ST (2016) Phosphorus availability controls the development of a marine harmful algal bloom caused by the pelagophyte, *Aureococcus anophagefferens.* **In review**.

Yutin N, Wolf Y, Raoult D & Koonin E (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virology journal* **6**: 223.

CHAPTER VI

Conclusion

Conclusion

While viruses within distinct phylogenetic categories and every nucleic acid types can infect and lyse eukaryotic phytoplankton, phylogenetically related giant dsDNA viruses have been found to be associated with algal blooms (Short, 2012). Ecological phenomena like algal blooms are of significant biogeochemical and/or economic importance. At the same time, giant viruses have caught the attention of evolutionary biologists due to their apparently complex evolutionary trajectory which is intertwined with that of their eukaryotic hosts (Iyer *et al.*, 2006, Moreira & Brochier-Armanet, 2008, Boyer *et al.*, 2009, Maumus *et al.*, 2014). Association of a giant virus with the harmful bloom alga *Aureococcus anophagefferens* prompted investigation into the role of this virus in modulating the dynamics of these blooms (Gastrich *et al.*, 2004, Gobler *et al.*, 2007). However, molecular tools to track the dynamics of this virus *in situ* were not available, limiting such investigation to lab settings. In addition, almost no information is available on the fine grain detail of the cellular events associated with AaV infection. Research presented in this dissertation commenced with the recognition that recurrent brown tide blooms, although a nuisance, can be an excellent 'model ecosystem' to study the viral dynamics associated with algal blooms *in situ.* We realized that developing molecular tools will not only allow us to understand AaV-*Aureococcus* relationship better, but also will broaden our understanding on eukaryotic virus-host dynamics as a whole. The findings obtained from this collection of studies can be summarized as follows –

1. AaV is a Nucleocytoplasmic Large DNA Virus (NCLDV), part of a virus group which infect diverse eukaryotes and share a common viral ancestor. Phylogenetically, AaV doesn't belong to the Phycodnaviridae family of alga infecting NCLDVs. Rather, it is the smallest 'giant' in the Mimiviridae (proposed Megaviridae) clade, which harbors some of the largest viruses ever discovered. The genome of AaV encodes footprints of intimate coevolution with its host – a large number of genes were possibly acquired from *Aureococcus* or other eukaryotes and bacteria.

2. Upon infection, AaV induces a rapid and massive reprogramming in *Aureococcus,* as evidenced by the altered cellular transcriptomic profile. A large number of cellular pathways are up- or downregulated compared to healthy cultures, implying viral modulation of the metabolic processes converting a healthy cell into a 'virocell', which is geared towards virus production. In addition, almost all the genes of AaV were expressed over the course of infection. Viral gene expression was temporally coordinated, with different genes possibly associated with different stages of the infection process.

3. Algal viruses in the Mimiviridae clade (including AaV) associated with brown tide blooms are highly diverse and dynamic – responding to changes in bloom condition. A possible explanation could be that these algal viruses were responding to the succession in the microeukaryotic community at different stages of the bloom. Algal Mimiviridae members are possibly an important component of the aquatic virosphere, and the primers developed in this study can be used to study them in diverse ecosystems.

4. Using metatranscriptomics, we determined AaV to be an active component of the brown tide blooms and detected diversity and dynamics of eukaryotic viruses with different nucleic acid types from both Quantuck Bay and Narragansett Bay study sites. Notably, we showed that time course metatranscriptomics can be used to reveal statistical co-occurrences of viruses and their putative hosts, and can be a powerful tool to predict such relationships. In fact, we argue that metatrancriptomics is potentially superior to metagenomics in drawing such correlations, since it focuses on the active members, largely excluding the relationships that might include transient or inactive viral community.

Studying the eukaryotic virus community *in situ* and at the level of individual host-virus pairs remain a formidable yet necessary task to fully understand the contribution of eukaryotic planktons to the marine biogeochemistry. This dissertation presents a number of studies that took both culture dependent and ecosystem level approaches to investigate giant virus-host interaction from nanoscale to ecosystem scale. Using both contemporary high throughput 'omics' approach and more 'traditional' techniques, we were able to address both finer details and big picture questions regarding giant virus ecology. With the necessary genomic information finally available, "*Aureococcus* – AaV" pair joins the small number of marine algal virus –host model systems. This opens up avenue for hypotheses driven research on this system, results of which will have broader implications for the field. A number of questions that will catch the attention of the researchers in near future can be envisioned. The molecular aspects of different 'virocells' will be investigated in search for markers of infection and commonality/differences in the viral strategies for virocell conversion across diverse hosts. Accomplishing this goal will also necessitate isolation of new host-virus pairs. Giant viruses forming distinct lineages within the same phylogenetic group (Mimivirus and Megavirus chilensis) can infect the same host (Arslan *et al.*, 2011), giving rise to the question, "how prevalent is such promiscuous relationship in natural giant virus communities?" Researching this question will provide important insights on the mechanism and extent to which giant viruses can drive the eukaryotic dynamics in the aquatic systems. However, marine eukaryotic community is enormously diverse (de Vargas *et al.*, 2015), and teasing apart the host-virus relationships

within this complex community will be challenging. Techniques will have to be developed for such predictions in a 'high throughput' manner, so that predictive models can be built incorporating these relationships. Metagenomics and targeted amplicon based approaches are already being used for predicting host-virus networks. However, as we have demonstrated the 'proof of principle' in this thesis, time series metatranscriptomics can be a suitable way to address this question along with gauging the diversity and dynamics of the active host and viral community. Already underway, the coevolution of giant viruses and their hosts will be more extensively researched, especially the function of specific viral genes in modulating host processes. Mechanism and preference in gain and loss of genes by giant viruses might necessitate 'long term evolution' experiments in lab – an approach already applied to Mimivirus (Boyer *et al.*, 2011).

This is an exciting era for aquatic virus research with scopes for significant contribution towards understanding the Earth's biosphere in a changing climate. Research presented in this dissertation was not only limited to a particular host-virus system, but also attempted to resolve long standing questions on marine virus ecology. We believe this body of work has the potential to propel the field forward with new knowledge and research questions that were generated.

List of References

Arslan D, Legendre M, Seltzer V, Abergel C & Claverie J-M (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci USA* **108**: 17486-17491.

Boyer M, Yutin N, Pagnier I, *et al.* (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA*. **106:** 21848-53.

Boyer M, Azza S, Barrassi L, *et al.* (2011) Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proc Natl Acad Sci USA* **108**: 10296-10301.

de Vargas C, Audic S, Henry N, *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**.

Gastrich M, Leigh-Bell J, Gobler C, Roger Anderson O, Wilhelm S & Bryan M (2004) Viruses as potential regulators of regional brown tide blooms caused by the alga, *Aureococcus anophagefferens*. *Estuaries* **27**: 112-119.

Gobler CJ, Anderson OR, Gastrich MD & Wilhelm SW (2007) Ecological aspects of viral infection and lysis in the harmful brown tide alga Aureococcus anophagefferens. *Aquatic Microbial Ecology* **47**: 25-36.

Iyer LM, Balaji S, Koonin EV & Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* **117**: 156-184.

Maumus F, Epert A, Nogué F & Blanc G (2014) Plant genomes enclose footprints of past infections by giant virus relatives. *Nat Commun* **5**:4268.

Moreira D & Brochier-Armanet C (2008) Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC evolutionary biology* **8**: 12.

Short SM (2012) The ecology of viruses that infect eukaryotic algae. *Environ Microbiol* **14**: 2253-2271.

Vita

Mohammad “Monir” Moniruzzaman was born in Jhenaidaha district of Bangladesh, a small country in South Asia. He spent his childhood in Narayanganj, a port city of Bangladesh, where his father was employed in the Adamjee Jute Mills - the largest jute processing industry of the world before it was shut down in 2002. He completed his undergrad and MS in Microbiology from the University of Dhaka. He joined the PhD program in UTK in 2011.

Apart from being fascinated by the secrets of the natural world, he is also passionate about photography. His favorite subjects are the people around him.