8-2019

Expanding the omics repertoire for model studies on a Chlorella-infecting giant virus

Samantha Coy
University of Tennessee, srose16@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Samantha Coy entitled "Expanding the omics repertoire for model studies on a Chlorella-infecting giant virus." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Microbiology.

Steven Wilhelm, Major Professor

We have read this dissertation and recommend its acceptance:

Alison Buchan, Shawn Campagna, Erik Zinser

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

EXPANDING THE OMICS REPERTOIRE FOR MODEL STUDIES ON A CHLORELLA-INFECTING GIANT VIRUS

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Samantha Rose Coy

August 2019

ACKNOWLEDGMENTS

"It takes a village…."

A few months ago, I made an acquaintance with a delightful and unusual family. They were a new husband and wife with a young daughter, but on top of that, the husband's brother was married to the wife's sister, and they all shared a house to save on finances. The sisters both worked, while the brothers took care of the kids at home and studied for their respective schooling. Amazed at how energized and joyful they were in the face of all these responsibilities, I lamely asked how difficult parenting was. The father shrugged with a smile and said, "It takes a village. We have a lot of people in our life who help out, and we couldn't do it without them." He paused and thought for a second, then continued, "Not only do they help out, but they are a part of how our daughter is raised. Everyone is a part of that in the way that we are this whole community of people coming together who get to do this together and that's pretty cool." This struck me, as I always thought of parenting as a bit more of a lone task. Reflecting on this as I am about to complete an arduous but rewarding PhD program, I am humbled at the realization that much like it takes a village to raise a child, it has also taken a village to get me here.

The first people I want to recognize are my parents, Art and Hollee. They are the ultimate role models for any situation life can throw at a person. They are selfless, hard-working, and the most faithful advocates a child can hope to have. My parents sacrificed next to everything so that my brother and I had opportunities to explore and excel in our chosen interests. They showed me I should not mistake myself for being too big to do a small job, or on the flip side, to feel too small to do a big job. I remember watching my mom hop on her bicycle to go work at the fast-food place down the road, load up the truck to go mow yards, and then come home to slave over a nutritious, home-cooked meal for our family. That said, you would be delighted to get to know her and find that she is one of the best critical thinkers I know (though she would never agree with me). Case in point: she designed and built both her and my brother's privacy fence, and anyone who sees them mistakes it for a professional build. My dad can be formidable in stature when you first meet him—I have told many people how he used to bench press over 400lbs-- but

he has the gentlest spirit and kindest attitude. How he loves my mother, my brother, and I is one of the most outstanding things about him. However, just like my mom, he is also incredibly hard working. While graduate school has its challenges and demands, my dad undoubtedly works longer hours and goes the extra mile more than anyone I have met—perhaps excluding my advisor, Dr. Wilhelm. The highest accolade I can hope for myself is to be recognized at even a fraction of the example my parents have set for being a good friend, worker, co-worker, neighbor, spouse, and eventually parent. I had the richest life because of them.

If you had asked me six years ago if I would be acknowledging my brother here, I probably would have laughed. My brother Ryan and I did not get along growing up, but somewhere along the way he met his amazing wife, Megan, became a rockstar software developer, and we started finding common ground again. The time Ryan and Megan took to invest in TC and I now amounts to many treasured memories and I am very thankful for the time out of the lab that I spent with them.

One of the reasons why my parents, Ryan, and Megan are such amazing people is the good ‘ole Missouri stock in each of them. People in Knoxville have playfully teased me about my pride for Missouri, but this pride mostly comes from a regard for the people there. One of the dearest families to me are the Thompsons: Kurt, Cheryl, Mia, and Ami. When I moved to Springfield, at the age of seven, I was too shy to talk to the other kids in the neighborhood and instead found my first friendships with the newlywed Kurt and Cheryl. I grew up across the street from them as they starting growing their family, and they soon started treating me as one of their own. I went on trips with them to Kansas to visit their relatives, proudly sat with Cheryl, Mia, and Ami at several high school football games that Kurt coached, and enjoyed the day-to-day life watching the girls say some of their first words, take their first steps, and on more than one occasion be mistaken as their teenage mother. Cheryl Thompson has one of the most genuine and loving hearts of anyone I have ever met. I told her my dreams to be a marine biologist someday, and when I was old enough, she arranged for me to stay with her sister Angela to do a summer internship in Key West, FL. This kindness that Cheryl showed me, and the risk that Angela

took in letting me into her home, to me represents the first domino to fall and lead me to where I am now. I am forever indebted to these people, love them more than most, and am completely inadequate in repaying them or acknowledging them as much as they deserve. Not being a part of their day-to-day life anymore is one of the hardest things about not living in Missouri.

Another person I would be remiss not to acknowledge is Dr. Teresa Carroll. I first knew her as only a mother of a girl I played basketball against in middle school. A few years passed and I ended up in Teresa's Zoology course as a freshman at Drury University, completely unsure of what to do with my life and somewhat hopeless for a satisfying future. She was the breath of fresh air I needed in my life that cultivated my interest in science, inspired my self-confidence, and set me up for graduate school. We bonded over our interest in science, she let me somewhat preemptively take a study abroad course with her to study tropical ecosystems, brought me in on her research, and lifted me up for all the success that came my way. Beyond the things she did for me, she was a role model in and of herself. Her journey in becoming a professor followed her earning not just one, but two PhD degrees—all the while raising three daughters and teaching courses at Drury. If I am ever feeling stretched thin, she is the person I reflect on and ask myself if I actually need to stretch the perception of what I can do. To this day, I still appreciate how lucky I am to have had her influence in my life.

There are many friends whom have been a much-needed outlet and refreshment during graduate school. Some have been in my life for only a short moment of time, while others have had a constant and enduring impact on me. Marissa Vigar and Regina Guazzo are two of the most treasured friendships I have formed that somehow have stood the test of time and space. That we have somehow managed to see each other at least once, if not two or three times a year, is a true feat despite spanning the country from coast-to-coast. Friends here in Knoxville have also had their influence; Danielle, my first roommate, who unknowingly made sure I persevered the first few years of graduate school. I often went to Jenny Only for advice on choices that came up in graduate school, and there is no one I would have trusted counsel from more. There is a big list of other

people that deserve to be told how special of an impact they have had on me, but for brevity's sake, I will simply name them here: Melissa and Charlie McClachlan, Melanie and Spencer Dixon, Bradley and Stephanie McMillan, Richard Kevorkian, Alex Emmons, Robbie Martin, Elizabeth McPherson, Ashley and Jordan Jenkins, Rudy Bonn, Kasey Rutherford, and Erica Freiert.

Each member of the Wilhelm lab is deserving of a big thank you. These people taught me almost everything I know, and were a joy to be around each day. The biggest thank you goes to Dr. Steven Wilhelm, my dissertation advisor, who cheerfully and gracefully guided me throughout the last six years of graduate school. I could not have had a more patient and encouraging advisor, and I always felt uplifted walking out of your office. A day does not go by that I feel deserving of the attention or investment that you bestowed on me. You are brilliant, and it is a privilege to have had even the smallest ounce of your time.

The last, and certainly not least, person on this list is my wonderful spouse and partner in life, TC Coy. Going through life with you has been the biggest adventure, and I am so thankful to have had you by my side. You kept me sane from day one, and your unenduring belief in me is all I need when I start to question myself. There is not a day that goes by that I do not acknowledge how lucky I am to have you in my life, and I am so grateful to you for your dedication to prioritizing my dreams. You are the rock in our family, and moving forward I want to help you reach your dreams as much as you have helped me. I love you, I don't deserve you, but I'll keep trying to be the person that does.

Despite these people being scattered all over the place, they are my village. Each of them has had such a special role in my life, and to complete this dissertation would not have been possible without their friendship, love, and support. They are a part of every success, have brought me through every failure, and keep me looking to the future. It is beyond luck that each of you have come into my life; I have thanked God continually for each of you and pray that I can sow back into your lives as much as you have mine. Thank you, thank you, thank you

ABSTRACT

Viruses are the most abundant biological entities in aquatic ecosystems. As top-down controls of plankton abundance and diversity, they are intrinsically linked to biogeochemical cycling, and by proxy, to global climate change. It is thus of great interest for researchers to understand the mechanics of viral infection and persistence among ecologically important phytoplankton assemblages. Viruses which infect eukaryotic algae are observed with diverse nucleic acid types, structures, and sizes, though most isolates to date bear large, dsDNA genomes comprised of genes normally only seen in cellular organisms. The *Chlorella* viruses are the model system for studying these entities, with many of the 'omics' approaches having been used to characterize the biology of this system. Here, we present data generated from epigenomic (*i.e.* DNA methylation) and metabolomic experiments of the prototype *Chlorella* virus, PBCV-1. In order to ask questions about virus DNA methylation, we first established a novel protocol for cryopreservation of PBCV-1 to control against epigenomic and genetic drift. This allowed for a baseline characterization of the DNA methylome profile in the prototype chlorovirus, PBCV-1, using PacBio's single-molecule, real-time (SMRT) sequencing software. The results of this study suggest the possibility of widespread epigenomic modifications, and that DNA methylation by viral restriction-modification associated enzymes is incomplete. Most instances of missing methylation marks are represented as hemimethylated palindromes, which are protected against the types of restriction enzymes encoded by these viruses and thus might represent an epigenomic regulatory function in the virus. Finally, we conducted a non-targeted metabolomics study of PBCV-1 infected *Chlorella* cells to make some of the first inferences of how viral infection alters the metabolic profile of this host system. Altogether, this work helps to distinguish the baseline epigenomic and metabolomic profiles of the *Chlorella*-PBCV-1 virus system for future comparison with more ecologically informative treatments (*i.e.* competition, sub-optimal light, nutrient limitation, etc.). This work will help to uncover general trends specific to algal-giant virus interactions that distinguish themselves from phage-bacteria systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : VIRUSES OF EUKARYOTIC ALGAE: DIVERSITY, METHODS FOR DETECTION, AND FUTURE DIRECTION

A version of this chapter was originally published in *Viruses*:

Samantha R. Coy, Eric R. Gann, Helena L. Pound, Steven M. Short, Steven W. Wilhelm. "Viruses of Eukaryotic Algae: Diversity, Methods for Detection, and Future Directions." *Viruses.* 2018;10(9): 487.

The conclusion of this article was revised here to reflect the dissertation focus on the on-going characterization of the model *Chlorella* virus system. S.R.C. and S.W.W. conceived the paper, and all listed authors contributed to the production of figures, text, and editing.

Abstract

The scope for ecological studies of eukaryotic algal viruses has greatly improved with the development of molecular and bioinformatic approaches that do not require algal cultures. Here, we review the history and perceived future opportunities for research on eukaryotic algal viruses. We begin with a summary of the 65 eukaryotic algal viruses that are presently in culture collections, with emphasis on shared evolutionary traits (e.g., conserved core genes) of each known viral type. We then describe how core genes have been used to enable molecular detection of viruses in the environment, ranging from PCR-based amplification to community scale "-omics" approaches. Special attention is given to recent studies that have employed network-analyses of "-omics" data to predict virus-host relationships, from which a general bioinformatics pipeline is described for this type of approach. Finally, we conclude with acknowledgement of how the field of aquatic virology is adapting to these advances, and highlight the need to properly characterize new virus-host systems that may be isolated using preliminary molecular surveys. Researchers can approach this work using lessons learned from the *Chlorella* virus system, which is not only the best characterized algal-virus system, but is also responsible for much of the foundation in the field of aquatic virology.

Introduction

Viruses infecting eukaryotic algae are extremely diverse. They have been reported with DNA or RNA genomes in various architectures (linear, circular, double-stranded, single-stranded, segmented) and sizes (4.4 to 638kb) (1). Some viruses accomplish infection with just a few viral genes at their disposal, while others maintain a gene arsenal nearly 100 times that size. Viruses infecting algae influence large ecological and biogeochemical processes. They direct the evolution of hosts through predator-prey selection and genetic exchange, consequently influencing algal fitness, population dynamics, and ultimately, microbial community structure. Infection can also alter the composition and distribution of organic matter in the environment (a process referred to as the aquatic ”viral shunt” (2)) and influence particle size-distribution, nutrient cycling, and biological system activity (e.g., respiration (3)). While algal viruses are important members in many aquatic environments, their contribution to these processes at the global scale primarily arises when they infect and lyse abundant bloom-forming algae. This includes harmful bloom formers and ecosystem scale specialists like coccolithophores that form blooms large enough to be observed from outer space (4).

It is a relatively recent realization that algal viruses are ecologically significant. In fact, the whole history of algal virus research has occurred primarily in just the last half century (Figure 1). While there have been sporadic observations of virus infection of algae cultures since the early 1970s (5, 6), the importance of algal viruses in natural systems was brought into the limelight by a series of observations of virus-like-particles associated with important bloom-forming algae (7-9). These findings inspired questions about the identity and evolutionary relationships within these virus-host systems. Such questions, however, required viruses to be isolated and genetically characterized.

One of the first algal-virus systems to achieve “model” status were the double-stranded DNA (dsDNA) viruses that infect the unicellular, ex-symbiotic, green alga *Chlorella* (10). The *Chlorella* virus-host model system remains the best characterized of all the algae-virus models, with genomes (11-15), transcriptomes (16, 17), and proteomes

(11) documented in the literature. Indeed, it was sequencing of the DNA polymerase B (*polB*) genes from *Chlorella* viruses PBCV-1 and NY-2A (18), and later from *Micromonas pusilla virus* SP1, that revealed a conserved amino acid sequence distinct from other known polB protein sequences. This observation enabled the development of degenerate PCR primers that selectively amplified these algal-virus *polB* genes (19, 20). The sequences of these PCR amplicons supported a unique monophyletic viral clade, now recognized as the family *Phycodnaviridae* of the Nucleocytoplasmic Large dsDNA Viruses (NCLDV). For a while the *Phycodnaviridae* was thought to be home to all of the large dsDNA algal viruses: perhaps even dominating the overall algal virus community. This perspective changed when sequencing of new isolates demonstrated that their "core" genes were more closely related to genes from the protist-infecting "giant viruses" of family *Mimiviridae* (21-23). In general, algal-infecting viruses are recognized as members of one of these two families, though future work may challenge the monophyletic nature of these groups. For example, clustering of the *Phycodnaviridae* is at times disrupted when homologs from other cellular or viral families are included in phylogenetic reconstructions (24-26).

There have also been increasing reports of single-stranded DNA (ssDNA) viruses, mostly infecting diatoms, RNA viruses (Table 1) (27, 28), and even parasites of these large algal viruses known as virophage (29, 30). The most informative reports on these systems have come from metagenomic and metatranscriptomic datasets that can detect the presence and activity of a wide range of DNA and RNA viruses. In turn, the known diversity of eukaryotic algal viruses has greatly expanded, at times even yielding putative full-length viral genome assemblies (31). Perhaps most promising is the possibility of predicting virus-host relationships in silico (31-33), whereas traditional methods have relied on virus isolation from a relative few cultivated algae. Shotgun -omics further create the opportunity to identify virus-host pairs from environmental data and place them in semi-quantitative ecological context. Indeed, these studies may even serve as preliminary assessments of the future cultivation requirements for isolating new virus-host systems. This burgeoning scientific frontier necessitates a review on the known diversity of

eukaryotic algal viruses, the molecular toolkit available for in situ studies on their ecology, and the direction aquatic virology is taking to adapt to these methodologies.

Diversity of cultured virus-host systems

The diversity of algal viruses mirrors that of their hosts, bearing in mind that the name “algae” does not denote a common evolutionary relationship. Indeed, algae have been observed in freshwater, marine, and terrestrial systems, in unicellular, colonial, or multicellular forms, and in disparate taxonomic lineages. Nevertheless, the diversity of algae can be depicted using an existing taxonomic framework that includes seven “supergroups” consisting of Excavata, Amoebozoa, Opisthokonta, Archaeplastida, the SAR group (Stramenophila, Alveolata, and Rhizaria), and a series of non-delineated, “cryptic” organisms collectively referred to as the *Incerta sedis* (34). Beyond this framework, the manner in which certain taxa are placed within eukaryotic phylogeny varies in the literature and is a subject of ongoing scientific debate. We adapted the schematic phylogeny presented by the TARA Oceans group (35) to illustrate the diversity of marine eukaryotic plankton, their relative abundance based on TARA Oceans 18S rDNA gene surveys, and lineage association with viruses that have been isolated and are maintained in lab cultures (Figure 2 and Table 1). This framework demonstrates that marine eukaryotic algae are known to occupy all but the Amoebozoa and Opisthokonta supergroups. Algae-infecting viruses have been isolated using hosts spanning almost all abundant planktonic lineages, though many are single systems or instances without genomic information to define viral phylogenetic placement (e.g., TampV). Although *Pyramimonadales* and *Raphidophyceae* were not abundant in the TARA Oceans 18S dataset, select species in these groups are known bloom-formers (36-39) making the available algal-virus system for these lineages ecologically informative. Viruses have also been isolated on important non-planktonic species, such as brown and red macroalgae (*Phaeophyceae* and *Rhodophyceae*). Abundant lineages without an algae-infecting virus include photosynthetic *Dictyochophyceae*, the Prasino Clade 7 group, the Chryso/Synuro group, and the Apicomplexans—though some of the highly represented lineages could be attributed to non-photosynthetic members. Establishing well characterized host-virus

systems in these lineages could be very useful for bloom-forming algae of these lineages. For example, it would be appealing to isolate a *Pseudochattonella* (*Dictyochophyceae*) infecting virus, as the host alga is responsible for fish kills. In 2016 *Pseudochattonella* was responsible for a massive fish kill in Peru amounting to an economic loss of ~$800 million dollars (40). In another interesting, albeit more complicated example, survival of the red-tide, bloom-forming ciliate *Mesodinium rubrum* depends on ingestion of photosynthetic cryptophytes to obtain necessary organelles (e.g., plastid, mitochondria, nucleus) (41). Viruses infecting cryptophyte prey may compete with this grazer, thus serving as an important control on the frequency and duration of red tides. Such broad trophic effects have been shown in studies on *Emiliania huxleyi,* where viral-infected cells are ingested by zooplankton at different rates than non-infected cells (42, 43).

Eukaryotic algal viruses in culture collections have been isolated from ~60 alga species (Table 1). Most of these are lytic, dsDNA viruses of the NCLDV group with a narrow, known host-range. The abundance of NCLDVs would imply that these are an ecologically relevant algal-virus type in the virus community, but whether or not these are the dominating type is unclear. This would certainly contrast with plant viromes which are dominated by RNA viruses. It is also possible that NCLDVs are more easily detected and isolated, thus explaining why only dsDNA viruses have been isolated from water samples that putatively contained other types of viruses. For example, electron micrographs of bloom-associated *Emiliania huxleyi* cells have been observed to simultaneously contain both small (50–60 nm) and large (185–200 nm) intracellular VLPs (44). Similar observations been made in *Pyramimonas orientalis* (45), but currently only one type of dsDNA virus has been isolated for this algae (46). It is possible that these viruses compete for algal infection, but they may also represent a case of virus-infecting virophage that are already known to co-occur with *Mimiviridae* (47, 48)*,* and perhaps even *Phycodnaviridae* (29, 32) viruses. Observations of co-occurring viruses are not limited to microscopy either; network analysis of metatranscriptomic data has linked the brown alga *Aureococcus anophagefferens* to its known dsDNA virus AaV as well as to uncharacterized ssDNA viruses (31), although the mechanism of this linkage (either direct, or via a co-occurring

microbial host of the virus) remains elusive. In short, algae may be infected by many types of viruses, potentially at the same time, and the numerically dominant virus type may not always represent that which is in the culture collection.

To date, there are four algal species that are known to be infected by diverse viruses comprised of different nucleic acid types. These include *Heterosigma akashiwo, Chaetoceros tenuissimus, Micromonas pusilla,* and *Heterocapsa circularisquama,* and in all cases the different virus types infect the same host strain (49). The coexistence of *Heterosigma akashiwo* viruses HaRNAV and HaDNAV is especially intriguing given these viruses exhibit opposite infection dynamics; the RNA virus has a high viral production rate, but a slower lytic cycle, whereas the DNA virus quickly replicates but produces fewer particles (50). It was hypothesized that coexistence could be maintained through variable host densities and viral decay rates, thus representing viruses that may have evolved as r- or k- strategists as has been proposed for *Heterocapsa* viruses (51), but is certainly not supported enough to be extrapolated as an explanation for all co-occurring viruses. Even virus isolates of the same nucleic acid type and species can exhibit considerable diversity. This can be extreme in some cases, where dsDNA viruses infecting the same algal host, which would be expected to cluster phylogenetically, are affiliated with NCLDV viral families *Mimiviridae* or *Phycodnaviridae* (e.g., *Phaeocystis globosa* Virus Groups I and Groups II (23, 52). It is possible that eukaryotic algae may commonly be infected by viruses of diverse replication strategies, and evolutionary histories, but the extent of this, as well as the factors that may allow this, needs more thorough investigation.

dsDNA viruses infecting eukaryotic algae

Most dsDNA viruses infecting algae are members of the NCLDV group, with the proposed exception of Tsv-N1 (53). Algal-NCLDV viruses have large genomes that encode hundreds of protein coding genes. Their evolutionary relationship has been inferred by core genes conserved across NCLDVs (54), placing them into either the family *Phycodnaviridae* or as extended members of the family *Mimiviridae*. Algal viruses of the latter group have recently been given the proposed distinction of Mesomimivirinae (55),

but for our purposes we will maintain the *Mimiviridae* description. The one exception to these two family assignments is HcDNAV, which shares closer similarity to the family *Asfarviridae* (56). To date, the NCLDV core gene compliment has been reduced to just a few genes (e.g., D5R packaging ATPase, D13L major capsid protein, and B family DNA polymerase), implying that the genetic diversity is huge among this group. Indeed, a genomic comparison among *Phycodnaviridae* members PBCV-1 (Chloroviruses), EsV-1 (Phaeoviruses), and EhV-86 (Coccolithoviruses) yielded only 14 conserved homologs from a pool of ~1000 genes (57). A more comprehensive look at these diverse genes can be found in genus-specific reviews of the *Phycodnaviridae* (14, 58-62).

It is anticipated that any single algal host can be permissive to many closely related virus variants, whereby phylogenetic comparisons of their core genes will reveal distinct clades (e.g., *Micromonas pusilla* and *Chlorella variabilis* viruses) with differences in latent phases, burst sizes, and genome size (14). In closely related viruses this is best resolved using concatenated alignments of marker protein sequences. At the same time, the origin of some of these genes is often attributed to gene transfer events. Many algal NCLDVs have acquired non-ancestral genes, but the majority of these appear to come from difference sources: Prasinoviruses acquire most of these from their host, Chlorovirus non-ancestral genes mostly derive from bacteria (63), and *Aureococcus anophagefferens* Virus (AaV) encodes a more even mixture of host, bacterial, archaeal, and viral genes (21). At the same time, it is worth noting that the origin of some genes could be difficult to ascertain if only a limited subset of viral (and host) homologs have been sequenced and annotated in public databases. Regardless, it has been suggested that viruses whose hosts are in closer association with bacteria tend to encode more putative non-ancestral genes, and that these genes cluster near the terminal ends of the viral genome (64). However, while the *Chlorella* algae is an endosymbiont of *Paramecium* that is certainly in close proximity to bacteria, the non-ancestral genes carried by the virus are evenly dispersed across its genome (14). In contrast, AaV displays terminal clusters of non-ancestral genes (21), but its host is a free-living photo/osmotroph. In either case, the biological implication of such high viral gene diversity, and how it is generated, is unclear.

It may help the virus acquire its specific needs for infection but has also been proposed to allow viruses to infect multiple hosts.

ssDNA viruses infecting eukaryotic algae

To date, the only ssDNA alga-infecting viruses that have been isolated are those which infect diatoms (Bacillariophyceae). In total, diatoms are a collective of an estimated 12,000–30,000 species, representing one of the most abundant phytoplankton groups in freshwater and marine environments (65). Most diatom-virus systems currently in culture are those infecting the cosmopolitan genus *Chaetoceros*. These isometric virus particles are ~35 nm in diameter and house circular, ssDNA genomes ranging from ~5.5–6.0 kb (66). The genomes generally encode four open reading frames consisting of an endonuclease (Rep), a major capsid protein, and two ORFs with unknown function. The capsid and replication initiating endonuclease are used in phylogenetic analyses. Three new members (whose genomes are ~4.5–4.7 kb) were recently reported from a de novo assembly of metagenomic reads from the mollusk *Amphibola crenata* and from sediment within an estuary in New Zealand (67). Phylogenetic analysis of the capsid proteins suggest this gene is a recent acquisition from ssRNA viruses, which is interesting, though not without precedent (68, 69). These metagenome assembled viruses have resulted in the taxonomic reclassification of diatom viruses into the family *Bacilladnaviridae* that includes cultured diatom viruses noted in Table 1 with asterisks (70). Many other ssDNA viruses are being detected in omics datasets (31), though resolving their specific host is an ongoing challenge.

RNA viruses infecting eukaryotic algae

Algae-infecting viruses with single (ss) and double-stranded (ds) RNA genomes have also been isolated and characterized, although most attention has been focused on the ssRNA isolates. Both virus groups encode an RNA-dependent RNA polymerase (RdRP), as well as proteases and helicases that can be used to infer distant evolutionary relationships. Most information on dsRNA algal viruses has been derived from the original

isolation papers describing the evolutionary relationships of the isolates. MpRV, a dsRNA virus of *Micromonas pusilla*, forms its own genus within the family *Reoviridae* (unassigned order) and has been proposed to be the ancestral line of the *Reoviridae* based on its placement between clades that demonstrate turreted or non-turreted virions (71). The other dsRNA virus isolate is *Chondrus crispus* virus (CcV), a toti-virus like entity. CcV represents an extraordinary case of a putative quasispecies virus that was accidentally discovered when a small band of dsRNA (~6 kb) was observed during host genomic preparation for sequencing (72). Similar dsRNA bands have been observed in extracts from all algal life phases, geographic locations, and in extracts from other red algae, though virus-like-particles and host lysis was not observed. The CcV system may represent either a latent or chronic (i.e., particle production below the limit of detection) viral infection that is ubiquitous among red algae, similar to known latent dsDNA viral infections of brown algae by Phaeoviruses (73). Since both *Chondrus crispus* and *Micromonas pusilla* are ecologically important algae, characterization of their relationship with these viruses is important and perhaps reflective of a need to search for more dsRNA viruses associated with algae.

ssRNA viruses have received considerably more attention since their hosts are common marine phytoplankton with some species capable of forming harmful blooms (38, 74, 75). Most of the alga-infecting ssRNA viruses are members of the order Picornavirales (Figure 3), with a few contradictions that are awaiting a taxonomic re-evaluation based on molecular data. The viruses infecting *Heterocapsa* and *Heterosigma* are the sole members of the families *Alvernaviridae* (unassigned order) and *Marnavirdiae* (order Picornavirales), respectively (67, 70), while the genus *Bacillarnavirus* (order Picornavirales) includes formal members *Chaetoceros socialis forma radians* RNA virus, *Chaetoceros tenuissimus* RNA virus 01, and *Rhizosolenia setigera* RNA virus 01. Other diatom viruses Csp03RNAV, AglaRNAV, and CtenRNAV type II are putative members of *Bacillarnavirus* based on phylogenetic relationships of replicase or structural proteins (1). The diatom viruses are generally thought to be highly species specific based on host-range experiments, with the exception of CtenRNAV type II which can infect four

Chaetoceros sp. in addition to *Chaetoceros tenuissimus* (66). These viruses and their hosts represent ecologically important systems that may reveal much on the persistence, co-existence, and competition of diatom viruses.

Culture independent approaches: expanding known diversity

PCR applications for estimating viral diversity and dynamics

Developing algal-virus model systems in the lab can inform much on the biology and ecology of algal viruses, but dependence on these systems is a limiting step. The ability to determine viral geographic distributions, population fluctuations, and diversity ultimately depends on analysis of environmental samples. Microscopic methods (76), flow cytometry (77-79), and infectivity assays (e.g., most probable number, plaque assay (10)) have been used to answer these questions, but these approaches lack taxonomic resolution and/or the relatively quick processing time that molecular techniques provide. To date, the principal molecular method for studying environmental algal viruses has been based on PCR amplification of conserved marker genes. Most of this work has focused on algal NCLDVs using polB (19) and the NCLDV major capsid protein (mcp) as gene targets (80): subsets of this community have been further examined using primers that specifically target the extended, algal *Mimiviridae* major capsid protein (AMmcp) (81). For reference, the potential amplification ranges of these primers are mapped against a phylogeny of sequenced virus isolates (Figure 4). There has been discussion on amplification bias of *polB* primers based on observations that environmental datasets tend to amplify prasinoviruses, even though these may be environmentally abundant viral types (51). The gene amplified by this primer set has also been suggested to be a poor marker for resolving within algal virus genera. For example, there are two distinct groups of *Phaeocystis globosa* infecting viruses, and these groups phylogenetically cluster into different families (1). Diversity may be better assessed using genome fluidity measurements of the pan-genome (82), but this would work better for describing viruses with full-genome sequences. Indeed, marker gene primer sets remain useful for elucidating environmental diversity of algal NCLDVs.

A recent clone library of PCR amplicons generated using the two mcp primer sets demonstrates a wide diversity of algal viruses isolated from marine and freshwater environments (81). This study also used PCR amplification to track the occurrence and dynamics of virus groups (defined by sequence clustering as operational taxonomic units, OTUs) over the course of a harmful brown-alga event. Biases aside, the approach used in that study has certainly expanded the known diversity of algal NCLDVs. It has also shown that cultured viral isolates are often distinct from environmental viruses, and that viruses are widely dispersed in the environment (80, 81, 83-86). Another recent group of primer sets was developed by Wilson et al. that amplifies a putative algal-*Mimiviridae* specific mismatch repair gene (*MutS*) (87). Novel groups of algal NCLDVs were detected in all of the samples tested, making this gene/primer set another potentially useful tool for studying virus diversity. RNA virus diversity has been assessed using primer sets targeting RNA dependent RNA polymerase (RdRP), a protein encoded by all RNA viruses (27, 88). This led to the discovery of a highly diverse super group of putative, marine, protist-infecting picorna-like viruses (88) that are consistently represented in metagenomic datasets (89). Moreover, alignments of conserved regions of RdRP form clades that are congruent with virion structure, host, and epidemiology (27).

While diversity can be addressed with degenerate primer PCR amplification, one of the major drawbacks of this approach is that it is generally not suitable for quantitative measurements (90). Indeed, degeneracies allow for biases in primer-binding and template amplification in mixed communities (91). Use of more specific primer sets and quantitative PCR approaches can avoid this issue (92, 93), but at the risk of not detecting closely related viruses. Even when using specific primer sets, recent duplications of marker genes can result in overestimation of viral abundances. One of the recent developments to overcome this is to spatially separate viruses and subject them to solid-phase, single-molecule PCR polony amplification (94). Family specific degenerate primers amplify diverse members without the issue of competitive amplification, then categorize and quantify the amplicons using probes for virus group specific genes. Of course, this method is also dependent on prior sequence knowledge on the virus types

of interest and has been validated only in cyanophage thus far, but it is certainly an appealing method for the study of eukaryotic alga infecting viruses. Another recently discovered application of PCR is its potential to link viruses and hosts. Microfluidics can be used to isolate infected single-cells that can then be subjected to simultaneous PCR detection of viral and host genes (95).

Using omics approaches to estimate virus diversity and dynamics

Because community scale genomics and transcriptomics are not dependent on target amplification, they are better suited for resolving viral diversity and can in some cases allow for the assembly of complete viral genomes. Though this is more readily accomplished in small RNA and DNA viruses (31, 67, 96), it has also been possible for some large dsDNA viruses and virophage (22, 29, 30). This potential is so valuable that a proposal was recently submitted to the International Committee on the Taxonomy of Viruses (ICTV) for the inclusion of metagenomic-assembled viruses into the official classification scheme (97). Not only was this approved, but it initiated a change in the primary approach ICTV uses for virus classification from phenotypic characterization based on viral isolates to molecular characterization based on viral DNA sequences. Since this time, metagenome assembled circular Rep-encoding single-stranded (CRESS) DNA viruses have been properly classified, including the *Bacilladnaviridae* (67), the putative vertebrate infecting *Smacoviridae* (98), and many more (70, 99). Some of the initial taxonomic classifications may also need to be reassessed in light of molecular methods, as classical taxonomy based on phenotype is not always congruent with phylogenetic clustering: The order Nidovirales may in fact belong to the Picornavirales.

While becoming more common, sequencing entire viral communities remains challenging and each experimental step must be considered in the context of existing biases and the project objectives. Virus particles have very low nucleic acid contents, necessitating amplification, concentration, or enrichment to obtain adequate sequencing depth. Simple approaches to do this involve concentration of environmental samples via filtration (100) or chemical flocculation (101). Virus enrichment can be done for specific

viral types with some quantitative applications. For example, dsDNA can be quantitatively amplified using fusion PCR primers, and adaptase will quantitatively amplify both ssDNA and dsDNA viruses (102). Rolling circle amplification can increase detection of circular viruses (67), and recombinant plant proteins that non-specifically bind dsRNA can select for dsRNA viruses (103). There are also methods to separate DNA and RNA viruses for separate analyses using hydroxyapatite-mediated techniques (104). One of the most appealing enrichment strategies recently used involves selection (*via* binding) of poly-A containing nucleic acid (i.e., mRNA) to focus on the active viral community (31). This is a useful signal to distinguish virus particles from active infection, as the former will not produce an mRNA signal, though this excludes some (+) ssRNA viruses that have polyadenylated genomes independent of infection (105). Though all of these methods are useful for improving detection, there are biases to be considered before making conclusions about viral abundances. These issues have been elucidated for sampling, extraction, and purification methods (106, 107), but these studies are not comprehensive.

The viral sequences generated from any sequencing approach are subjected to a general analytical workflow involving quality filtering, assembly, annotation, and diversity analyses. Many tools are available to perform this bioinformatic workflow (108), but few of these are designed to complete the full workflow. Moreover, careful understanding of the sequence databases searched in each workflow is necessary to know whether biases exist for particular virus types. GenBank and the nt/nr databases are preferred as these are continually updated and contain information for all virus types; however, their large size can slow processing considerably. To overcome this, creating custom workflows using marker genes of interest can speed up processing time while maintaining the ability to detect diverse virus types.

An example of a bioinformatics workflow using a custom marker gene database to interpret NGS sequences (i.e., Illumina™ paired-end sequencing) is shown in Figure 5. First, reads must be preprocessed to remove contaminating adapter sequences and trim low-quality reads. The next step involves assembly of reads into larger contigs, followed by contig annotation using a database of known sequences and a homology or alignment

search tool (BLAST, HMMER, Bowtie2, etc.). BLAST tools have commonly been used for this purpose in cellular organisms, and even in some virus studies (32), but may be less efficient for identifying novel virus homologs since they often have low pairwise sequence identities (109). An alternative to using sequence alignments are Hidden Markov models (MMS), which score hits to protein domains. These analyses can be done with the search tool HMMER to create a marker gene database (HMM-build) that can be queried against assembled contigs (110). Once viral contigs have been identified, the relevant gene hits can be extracted for post-processing (i.e., phylogenetic analysis). In many cases, especially when using small databases, it is useful to verify viral hits with a second similarity search of the extracted gene. Following verification, extracted viral hits can be placed unto an existing phylogenetic tree built with homologous reference sequences (e.g., pplacer (111)). Tree topology can be confirmed using a variety of other tree-building software (e.g., FastTree 2.1.7 (112), PhyML (113), RAxML (114), IQ-tree (115)) and methods (e.g., MrBayes for Bayesian tree-building (116)).

Information on virus abundance or activity can be inferred by mapping trimmed metagenomic or metatranscriptomic reads back to viral contigs normalized for between-sample comparisons (e.g., internal standards, library size, length, and reads per kilobase of transcript per million mapped reads [RPKM] values). However, there are some caveats to consider when examining environmental metatranscriptomes. Transcript abundance is not directly related to viral abundance for two reasons: First, biases are known to exist for highly transcriptionally active viruses, and second, single host organisms can support high viral loads. Moreover, virus metatranscriptomes can be contaminated with chimeras generated during assembly, remnant viral genes may be expressed from cells (117), and genomic duplications of marker genes could confound expression profiles. Some problems can be avoided with proper sampling and sequencing approaches mentioned previously, but others remain a significant obstacle for quantitative community analyses, though this has been resolved for bacteria-infecting viruses (102, 118). Until these confounding issues can be remedied and benchmarked for all viral types, they must be considered during the analysis of environmental data. A recent review by Nooij et al.

provides a comprehensive description of workflows that have been produced for viromic analyses, including specific applications, classification biases, and open-source availability (108).

Other downstream applications of omic assemblies

Another enticing application of community sequence data is the potential to deduce biological interactions using co-occurrence or network analyses. This is a relatively new approach that was developed for microbiome communities but has the potential to identify novel virus-host pairs (119). Two studies tracking the temporal dynamics of virus communities have been reported thus far (31, 32). From a metagenomics standpoint, these studies were striking because they generated putatively full-length Picornavirales and virophage genomes. Moreover, in the case of Moniruzzaman et al. 2017 (31) the viral genomes were generated from transcripts, indicating these virus genomes were actively expressed and were therefore produced from infected cells. Beyond these exciting findings, each study used network analyses to link potential virus-host pairs. Clusters created from sequencing data collected over the course of a brown-tide bloom (*Aureococcus anophagefferens*) linked the brown alga to its known virus, AaV, demonstrating the ability to extract known relationships with this approach. Several other clusters were generated from the same study, including smaller networks of single virus-host pairs and expected associations between *Prasinophyceae* and *Phycodnaviridae.* Roux et al. 2017 (32) focused on using networks to link virophage with giant NCLDV hosts and found strong specific associations with *Mimiviridae* and their extended alga-infecting members to drastically expand the diversity of known virophage hosts.

Altogether, predictions stemming from the studies noted above demonstrate how network analyses can generate testable hypotheses for future studies of algal virus-host interactions. By deducing sequences of virus-host pairs, one can attempt to confirm probable virus-host interactions. For example, a variation of fluorescent in-situ hybridization, deemed phageFISH, could be used to label virus and host genes in infected cells (120). Additionally, networks predicting viruses of cultured algae could be followed

up with virus tagging experiments(121). It might even be worthwhile to use more than one network building approach to look at ecosystem structures. Weiss et al. used real and mock in silico data to benchmark eight methods used for bacterial network analyses and found that some methods generate drastically different outputs (122). This is explained, in part, by differing strengths for detecting particular biological relationships (e.g., mutualism and commensalism) across different network approaches. It was also suggested that *p*-values of 0.001 should be used for high-precision network detection and rare OTUs should be removed prior to network construction.

Conclusions

The opportunities for algal virus ecologists are at an all-time high. Bioinformatic tools are becoming more accessible to a wide variety of scientists through the creation of publicly available genomic databases and graphic interfaces that mediate interactions with traditional command-line software (123). At the same time, researchers are increasing collaborations with one another by sharing methodologies in an interactive framework on *protocols.io* (e.g., Viral Ecology Research and Virtual Exchange network, or VERVE Net; https://www.protocols.io/groups/verve-net) and with cross-discipline collaborations fostered at research workshops funded by organizations like the *Gordon & Betty Moore Foundation* (GBMF) and the *Canadian Institute for Advanced Research* (CIFAR). The development of long-read sequencing methods, preemptively deemed “third-generation sequencing”, may address many of the issues with short-read assembly and viral quantification. DNA barcoding has been suggested as a cheap, reliable method to quickly track virus populations, and has recently been shown to recapitulate general viral community structures using sample volumes no bigger than a cup of water (124). New virus isolates can be discovered from sequencing of single aquatic viruses sorted by flow cytometry (125), as closely related, hyper diverse viruses are suggested to be difficult to assemble from metagenomes (126). Even better, isolation and sequencing of infected single-cells may allow for the identification of new virus-host systems. Network analyses of community sequence data predict ecological structures that may lead to the discovery

and isolation of several new algal-virus systems, bringing the scientific community “full-circle” to studying these systems in the lab. In light of that, it is immensely important to continue characterization of existing model giant virus systems in order to deduce conserved functions that might be novel or unique to these entitites. This objective is the cornerstone of the following dissertation work.

Algal-infecting NCLDVs include members of the families Phycodnaviridae and the more recently acknowledged Mimiviridae (Figure 1). The genetic diversity of isolates within these families is vast: a genomic comparison among just Phycodnaviridae members PBCV-1 (chloroviruses), EsV-1 (Phaeoviruses), and EhV-86 (Coccolithoviruses) yields only 14 conserved homologs from a pool of 1000 genes (57). This number is reasonable considering the associated hosts include freshwater, marine, unicellular, and multicellular algae bearing cell walls as different as anionic polysaccharides and calcareous plates. At the same time, a viral element that is conserved across this host diversity represents a likely important function of NCLDV viruses. These include ‘cell-like’ elements not normally not observed in viruses, including central components of protein-translation (e.g. tRNAs), inteins, and parts of DNA repair pathways. Another unique characteristic that has received less attention than these is an unusually high number of DNA-decorating methyltransferase (MTase) genes (Figure 2). These genes can be very concentrated in algal-NCLDV genomes, and yet, regression analyses indicate that genome size has little to do with their presence ($R=0.02$, $p=0.04$). This begs the question, what benefit do algal viruses gain from methylating their DNA? The first few chapters of this dissertation lay the groundwork for how we can ask questions about DNA methylation in algal-NCLDVs. We use the chloroviruses for these studies because viral strains have been shown to encode anywhere from 0 to 18 DNA methyltransferases, which can account for up to ~4.5% of a single virus’ protein coding potential (12, 13). Moreover, methylation patterns in the prototype chlorovirus, PBCV-1, may be understood against a rich research context including genomics, transcriptomics, proteomics, and cloning studies. Before establishing the ‘epigenomics’ of this system, however, it became apparent that we would need to ensure that PBCV-1 could maintain

a consistent methylation phenotype that we could call ‘wild type’. Historically, chlorovirus PBCV-1 has been maintained in the laboratory by serial propagation, thus subjecting it to genetic (or even epigenetic) drift. To control against this potential complication, we established a protocol for cryopreserving chlorovirus PBCV-1 which is described in the second chapter. By developing a successful method to cryopreserve virus PBCV-1, we became able to create a seed-stock system that allowed us to characterize a ‘base-line’ methylation pattern for chlorovirus PBCV-1 using *in silico* bioinformatic analyses and single-molecule real time (SMRT) sequencing data. This work is presented in the third chapter, and highlights potential novel functions of DNA methylation in algal giant viruses. In the final research chapter, we share the share data from the last remaining ‘omics’ study to be done in the PBCV-1 system: metabolomics. We describe the changing metabolic profile of the infected *Chlorella* cell, and highlight specific metabolites that account for most of the change over a six-hour infection cycle with PBCV-1. We conclude by identifying next steps to continue on the work presented in this dissertation.

References

1. Short SM, Staniewski MA, Chaban YV, Long AM, Wang D. Diversity of viruses infecting eukaryotic algae. In: P. H, Abedon ST, editors. Viruses of microorganisms. Poole, UK: Caister Academic Press; 2018. p. 211-44.

2. Wilhelm SW, Suttle CA. Viruses and nutrient cycles in the sea - viruses play critical roles in the structure and function of aquatic food webs. BioScience. 1999;49(10):781-8.

3. Fuhrman JA. Marine viruses and their biogeochemical and ecological effects. Nature. 1999;399(6736):541-8.

4. Holligan PM, Viollier M, Harbour DS, Camus P, Champagne-Philippe M. Satellite and ship studies of coccolithophore production along a continental-shelf edge. Nature. 1983;304(5924):339-42.

5. Wilhelm SW, Bird JT, Bonifer KS, Calfee BC, Chen T, Coy SR, et al. A student's guide to giant viruses infecting small eukaryotes: from *Acanthamoeba* to *Zooxanthellae*. Viruses. 2017;9(3).

6. Van Etten JL, Lane LC, Meints RH. Viruses and virus-like particles of eukaroytic algae. Microbiological Reviews. 1991;55(4):586-620.

7. Bratbak G, Egge JK, Heldal M. Viral mortality of the marine alga *Emiliania-huxleyi* (*Haptophyceae*) and termination of algal blooms. Marine Ecology Progress Series. 1993;93(1-2):39-48.

8. Nagasaki K, Ando M, Itakura S, Imai I, Ishida Y. Viral mortality in the final stage of *Heterosigma akashiwo* (*Raphidophyceae*) red tide. Journal of Plankton Research. 1994;16(11):1595-9.

9. Gastrich MD, Anderson OR, Benmayor SS, Cosper EM. Ultrastructural analysis of viral infection in the brown-tide alga, *Aureococcus anophagefferens* (*Pelagophyceae*). Phycologia. 1998;37(4):300-6.

10. Van Etten JL, Burbank DE, Kuczmarski D, Meints RH. Virus-infection of culturable *Chlorella*-like algae and development of a plaque assay. Science. 1983;219(4587):994-6.

11. Dunigan DD, Cerny RL, Bauman AT, Roach JC, Lane LC, Agarkova IV, et al. *Paramecium bursaria Chlorella Virus* 1 proteome reveals novel architectural and regulatory features of a giant virus. Journal of Virology. 2012;86(16):8821-34.

12. Fitzgerald LA, Graves MV, Li X, Feldblyum T, Nierman WC, Van Etten JL. Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect *Chlorella* NC64A. Virology. 2007;358(2):472-84.

13. Fitzgerald LA, Graves MV, Li X, Feldblyum T, Hartigan J, Van Etten JL. Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect *Chlorella* Pbi. Virology. 2007;358(2):459-71.

14. Jeanniard A, Dunigan DD, Gurnon JR, Agarkova IV, Kang M, Vitek J, et al. Towards defining the chloroviruses: a genomic journey through a genus of large DNA viruses. BMC Genomics. 2013;14.

15. Quispe CF, Esmael A, Sonderman O, McQuinn M, Agarkova I, Battan M, et al. Characterization of a new chlorovirus type with permissive and non-permissive features on phylogenetically related algal strains. Virology. 2017;500:103-13.

16. Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova IV, et al. Microarray analysis of *Paramecium bursaria Chlorella Virus* 1 transcription. Journal of Virology. 2010;84(1):532-42.

17. Blanc G, Mozar M, Agarkova IV, Gurnon JR, Yanai-Balser G, Rowe JM, et al. Deep RNA sequencing reveals hidden features and dynamics of early gene transcription in *Paramecium bursaria Chlorella Virus* 1. PLOS One. 2014;9(3):10.

18. Grabherr R, Strasser P, Vanetten JL. The DNA-polymerase gene from *Chlorella* viruses PBCV-1 and NY-2A contains an intron with nuclear splicing sequences. Virology. 1992;188(2):721-31.

19. Chen F, Suttle CA. Amplification of DNA-polymerase gene fragments from viruses infecting microalgae. Applied and Environmental Microbiology. 1995;61(4):1274-8.

20. Chen F, Suttle CA, Short SM. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. Applied and Environmental Microbiology. 1996;62(8):2869-74.

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

22. Zhang WJ, Zhou JL, Liu TG, Yu YX, Pan YJ, Yan SL, et al. Four novel algal virus genomes discovered from Yellowstone Lake metagenomes. Scientific Reports. 2015;5.

23. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, et al. Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(26):10800-5.

24. Yutin N, Koonin EV. Hidden evolutionary complexity of nucleo-cytoplasmic large DNA viruses of eukaryotes. Virology Journal. 2012;9.

25. Yutin N, Koonin EV. Pandoraviruses are highly derived phycodnaviruses. Biology Direct. 2013;8.

26. Maruyama F, Ueki S. Evolution and phylogeny of large DNA viruses, *Mimiviridae* and *Phycodnaviridae* including newly characterized *Heterosigma akashiwo* virus. Frontiers in Microbiology. 2016;7.

27. Culley AI, Lang AS, Suttle CA. High diversity of unknown picorna-like viruses in the sea. Nature. 2003;424(6952):1054-7.

28. Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. Are we missing half of the viruses in the ocean? ISME Journal. 2013;7(3):672-9.

29. Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ, et al. Virophage control of antarctic algal host-virus dynamics. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(15):6163-8.

30. Gong CW, Zhang WJ, Zhou XW, Wang HM, Sun GW, Xiao JZ, et al. Novel virophages discovered in a freshwater lake in China. Frontiers in Microbiology. 2016;7.

31. Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Gobler CJ, Wilhelm SW. Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. Nature Communications. 2017;8.

32. Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. Nature Communications. 2017;8.

33. Allen LZ, McCrow JP, Ininbergs K, Dupont CL, Badger JH, Hoffman JM, et al. The Baltic Sea virome: diversity and transcriptional activity of DNA and RNA viruses. Msystems. 2017;2(1).

34. Burki F. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harbor Perspectives in Biology. 2014;6(5).

35. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348(6237).

36. Kataoka T, Yamaguchi H, Sato M, Watanabe T, Taniuchi Y, Kuwata A, et al. Seasonal and geographical distribution of near-surface small photosynthetic eukaryotes in the western North Pacific determined by pyrosequencing of 18S rDNA. FEMS Microbiology Ecology. 2017;93(2).

37. Gowing MM. Large viruses and infected microeukaryotes in Ross Sea summer pack ice habitats. Marine Biology. 2003;142(5):1029-40.

38. Honjo T. Overview on bloom dynamics and physiological ecology of *Heterosigma akashiwo*. Smayda TJ, Shimizu Y, editors. 1993. 33-41 p.

39. Karosiene J, Kasperoviciene J, Koreiviene J, Savadova K, Vitonyte I. Factors promoting persistence of the bloom-forming *Gonyostomum semen* in temperate lakes. Limnologica. 2016;60:51-8.

40. Leon-Munoz J, Urbina MA, Garreaud R, Iriarte JL. Hydroclimatic conditions trigger record harmful algal bloom in western Patagonia (summer 2016). Scientific Reports. 2018;8:10.

41. Gustafson DE, Stoecker DK, Johnson MD, Van Heukelem WF, Sneider K. Cryptophyte algae are robbed of their organelles by the marine ciliate *Mesodinium rubrum*. Nature. 2000;405(6790):1049-52.

42. Vermont AI, Martinez JM, Waller JD, Gilg IC, Leavitt AH, Floge SA, et al. Virus infection of *Emiliania huxleyi* deters grazing by the copepod *Acartia tonsa*. Journal of Plankton Research. 2016;38(5):1194-205.

43. Evans C, Wilson WH. Preferential grazing of *Oxyrrhis marina* on virus-infected *Emiliania huxleyi*. Limnology and Oceanography. 2008;53(5):2035-U12.

44. Brussaard CPD, Kempers RS, Kop AJ, Riegman R, Heldal M. Virus-like particles in a summer bloom of *Emiliania huxleyi* in the North Sea. Aquatic Microbial Ecology. 1996;10(2):105-13.

45. Moestrup HT, H. A. An ultrastructural study of the flagellate *Pyramimonas orientalis* with particular emphasis on golgi apparatus activity and the flagellar apparatus. Protoplasma. 1974;81:247-69.

46. Sandaa RA, Heldal M, Castberg T, Thyrhaug R, Bratbak G. Isolation and characterization of two viruses with large genome size infecting *Chrysochromulina ericina* (*Prymnesiophyceae*) and *Pyramimonas orientalis* (*Prasinophyceae*). Virology. 2001;290(2):272-80.

47. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, et al. The virophage as a unique parasite of the giant mimivirus. Nature. 2008;455(7209):100-U65.

48. Fischer MG, Suttle CA. A virophage at the origin of large DNA transposons. Science. 2011;332(6026):231-4.

49. Mojica KDA, Brussaard CPD. Factors affecting virus dynamics and microbial host-virus interactions in marine environments. FEMS Microbiology Ecology. 2014;89(3):495-515.

50. Lawrence JE, Brussaard CPD, Suttle CA. Virus-specific responses of *Heterosigma akashiwo* to infection. Applied and Environmental Microbiology. 2006;72(12):7829-34.

51. Short SM. The ecology of viruses that infect eukaryotic algae. Environmental Microbiology. 2012;14(9):2253-71.

52. Brussaard CPD, Short SM, Frederickson CM, Suttle CA. Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton *Phaeocystis globosa* (*Prymnesiophyceae*). Applied and Environmental Microbiology. 2004;70(6):3700-5.

53. Pagarete A, Grebert T, Stepanova O, Sandaa RA, Bratbak G. Tsv-N1: A novel DNA algal virus that infects *Tetraselmis striata*. Viruses-Basel. 2015;7(7):3937-53.

54. Iyer LM, Aravind L, Koonin EV. Common origin of four diverse families of large eukaryotic DNA viruses. Journal of Virology. 2001;75(23):11720-34.

55. Gallot-Lavallee L, Blanc G, Claverie JM. Comparative genomics of *Chrysochromulina Ericina* Virus and other microalga-infecting large DNA viruses highlights their intricate evolutionary relationship with the established *Mimiviridae* family. Journal of Virology. 2017;91(14).

56. Ogata H, Toyoda K, Tomaru Y, Nakayama N, Shirai Y, Claverie JM, et al. Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. Virology Journal. 2009;6.

57. Allen MJ, Schroeder DC, Holden MTG, Wilson WH. Evolutionary history of the *Coccolithoviridae*. Molecular Biology and Evolution. 2006;23(1):86-92.

58. Derelle E, Monier A, Cooke R, Worden AZ, Grimsley NH, Moreau H. Diversity of viruses infecting the green microalga *Ostreococcus lucimarinus*. Journal of Virology. 2015;89(11):5812-21.

59. Finke JF, Winget DM, Chan AM, Suttle CA. Variation in the genetic repertoire of viruses infecting *Micromonas pusilla* reflects horizontal gene transfer and links to their environmental distribution. Viruses-Basel. 2017;9(5).

60. Dunigan DD, Fitzgerald LA, Van Etten JL. Phycodnaviruses: a peek at genetic diversity. Virus Research. 2006;117(1):119-32.

61. Nissimov JI, Pagarete A, Ma F, Cody S, Dunigan DD, Kimmance SA, et al. Coccolithoviruses: A review of cross-kingdom genomic thievery and metabolic thuggery. Viruses-Basel. 2017;9(3).

62. Clerissi C, Grimsley N, Ogata H, Hingamp P, Poulain J, Desdevises Y. Unveiling of the diversity of prasinoviruses (*Phycodnaviridae*) in marine samples by using high-throughput sequencing analyses of PCR-amplified DNA polymerase and major capsid protein genes. Applied and Environmental Microbiology. 2014;80(10):3150-60.

63. Filee J. Genomic comparison of closely related giant viruses supports an accordion-like model of evolution. Frontiers in Microbiology. 2015;6.

64. Filee J, Pouget N, Chandler M. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic Large DNA viruses. BMC Evolutionary Biology. 2008;8(320).

65. Malviya S, Scalco E, Audic S, Vincenta F, Veluchamy A, Poulain J, et al. Insights into global diatom distribution and diversity in the world's ocean. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(11):E1516-E25.

66. Kimura K, Tomarua Y. Discovery of Two Novel Viruses Expands the Diversity of single-stranded DNA and single-stranded RNA viruses infecting a cosmopolitan marine diatom. Applied and Environmental Microbiology. 2015;81(3):1120-31.

67. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology. 2017;504:114-21.

68. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biology Direct. 2012;7.

69. Krupovic M, Koonin EV. Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses. Scientific Reports. 2014;4.

70. King AMQ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, et al. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). Archives of virology. 2018.

71. Attoui H, Jaafar FM, Belhouchet M, de Micco P, de Lamballerie X, Brussaard CPD. *Micromonas pusilla* reovirus: a new member of the family *Reoviridae* assigned to a novel proposed genus (*Mimoreovirus*). Journal of General Virology. 2006;87:1375-83.

72. Rousvoal S, Bouyer B, Lopez-Cristoffanini C, Boyen C, Collen J. Mutant swarms of a totivirus-like entities are present in the red macroalga *Chondrus crispus* and have been partially transferred to the nuclear genome. Journal of Phycology. 2016;52(4):493-504.

73. Delaroque N, Maier I, Knippers R, Muller DG. Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (*Phaeophyceae*). Journal of General Virology. 1999;80:1367-70.

74. Nagai K, Matsuyama Y, Uchida T, Yamaguchi M, Ishimura M, Nishimura A, et al. Toxicity and LD(50) levels of the red tide dinoflagellate *Heterocapsa circularisquama* on juvenile pearl oysters. Aquaculture. 1996;144(1-3):149-54.

75. Nagasaki K. Dinoflagellates, diatoms, and their viruses. Journal of Microbiology. 2008;46(3):235-43.

76. Noble RT, Fuhrman JA. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquatic Microbial Ecology. 1998;14(2):113-8.

77. Marie D, Brussaard CPD, Thyrhaug R, Bratbak G, Vaulot D. Enumeration of marine viruses in culture and natural samples by flow cytometry. Applied and Environmental Microbiology. 1999;65(1):45-52.

78. Brussaard CPD, Marie D, Bratbak G. Flow cytometric detection of viruses. Journal of Virological Methods. 2000;85(1-2):175-82.

79. Brussaard CPD. Optimization of procedures for counting viruses by flow cytometry. Applied and Environmental Microbiology. 2004;70(3):1506-13.

80. Larsen JB, Larsen A, Bratbak G, Sandaa RA. Phylogenetic analysis of members of the *Phycodnaviridae* virus family, using amplified fragments of the major capsid protein gene. Applied and Environmental Microbiology. 2008;74(10):3048-57.

81. Moniruzzaman M, Gann ER, LeCleir GR, Kang Y, Gobler CJ, Wilhelm SW. Diversity and dynamics of algal *Megaviridae* members during a harmful brown tide caused by the pelagophyte, *Aureococcus* anophagefferens. FEMS Microbiology Ecology. 2016;92(5).

82. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12:10.

83. Short SM, Suttle CA. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. Applied and Environmental Microbiology. 2002;68(3):1290-6.

84. Short SM, Short CM. Diversity of algal viruses in various North American freshwater environments. Aquatic Microbial Ecology. 2008;51(1):13-21.

85. Clasen JL, Suttle CA. Identification of freshwater *Phycodnaviridae* and their potential phytoplankton hosts, using DNA pol sequence fragments and a genetic-distance analysis. Applied and Environmental Microbiology. 2009;75(4):991-7.

86. Rowe JM, Fabre MF, Gobena D, Wilson WH, Wilhelm SW. Application of the major capsid protein as a marker of the phylogenetic diversity of *Emiliania huxleyi* viruses. FEMS Microbiology Ecology. 2011;76(2):373-80.

87. Wilson WH, Gilg IC, Duarte A, Ogata H. Development of DNA mismatch repair gene, MutS, as a diagnostic marker for detection and phylogenetic analysis of algal megaviruses. Virology. 2014;466:123-8.

88. Culley AI, Steward GF. New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. Applied and Environmental Microbiology. 2007;73(18):5937-44.

89. Culley A. New insight into the RNA aquatic virosphere via viromics. Virus Research. 2018;244:84-9.

90. Sullivan MB. Viromes, not gene markers, for studying double-stranded DNA virus communities. Journal of Virology. 2015;89(5):2459-61.

91. Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. Applied and Environmental Microbiology. 1998;64(10):3724-30.

92. Short SM, Short CM. Quantitative PCR reveals transient and persistent algal viruses in Lake Ontario, Canada. Environmental Microbiology. 2009;11(10):2639-48.

93. Short CM, Rusanova O, Short SM. Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada. ISME Journal. 2011;5(5):810-21.

94. Baran N, Goldin S, Maidanik I, Lindell D. Quantification of diverse virus populations in the environment using the polony method. Nature Microbiology. 2018;3(1).

95. Tadmor AD, Ottesen EA, Leadbetter JR, Phillips R. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. Science. 2011;333(6038):58-62.

96. Culley AI, Mueller JA, Belcaid M, Wood-Charlson EM, Poisson G, Steward GF. The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. Mbio. 2014;5(3).

97. Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy in the age of metagenomics. Nature Reviews Microbiology. 2017;15(3):161-8.

98. Varsani A, Krupovic M. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. Archives of Virology. 2018;163(7):2005-15.

99. Adams MJ, Lefkowitz EJ, King AMQ, Harrach B, Harrison RL, Knowles NJ, et al. Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2017). Archives of Virology. 2017;162(8):2505-38.

100. Wommack KE, Hill RT, Colwell RR. A simple method for the concentration of viruses from natural water samples. Journal of Microbiological Methods. 1995;22(1):57-67.

101. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environmental Microbiology Reports. 2011;3(2):195-202.

102. Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenk SM, Goldsmith DB, et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. Peerj. 2016;4.

103. Kobayashi K, Tomita R, Sakamoto M. Recombinant plant dsRNA-binding protein as an effective tool for the isolation of viral replicative form dsRNA and universal detection of RNA viruses. Journal of General Plant Pathology. 2009;75:87-91.

104. Andrews-Pfannkoch C, Fadrosh DW, Thorpe J, Williamson SJ. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. Applied and Environmental Microbiology. 2010;76(15):5039-45.

105. Shatkin AJ. Animal RNA viruses - genome structure and function. Annual Review of Biochemistry. 1974;43:643-65.

106. Steward GF, Culley A. Extraction and purification of nucleic acids from viruses. Manual of Aquatic Viral Ecology 2010. p. 154-65.

107. Hurwitz BL, Deng L, Poulos BT, Sullivan MB. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. Environmental Microbiology. 2013;15(5):1428-40.

108. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG. Overview of virus metagenomic classification methods and their biological applications. Frontiers in Microbiology. 2018;9:21.

109. Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL. Profile hidden markov models for the detection of viruses within metagenomic sequence data. Plos One. 2014;9(8).

110. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Research. 2013;41(12).

111. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. Bmc Bioinformatics. 2010;11.

112. Price MN, Dehal PS, Arkin AP. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution. 2009;26(7):1641-50.

113. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010;59(3):307-21.

114. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.

115. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution. 2015;32(1):268-74.

116. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17(8):754-5.

117. Gallot-Lavallee L, Blanc G. A glimpse of Nucleo-Cytoplasmic Large DNA Virus biodiversity through the eukaryotic genomics window. Viruses-Basel. 2017;9(1):14.

118. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. Peerj. 2017;5.

119. Sullivan MB, Weitz JS, Wilhelm S. Viral ecology comes of age. Environmental Microbiology Reports. 2017;9(1):33-5.

120. Allers E, Moraru C, Duhaime MB, Beneze E, Solonenko N, Barrero-Canosa J, et al. Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. Environmental Microbiology. 2013;15(8):2306-18.

121. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, et al. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. Nature. 2014;513(7517):242-+.

122. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME Journal. 2016;10(7):1669-81.

123. Bolduc B, Youens-Clark K, Roux S, Hurwitz BL, Sullivan MB. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. ISME Journal. 2017;11(1):7-14.

124. Flaviani F, Schroeder DC, Balestreri C, Schroeder JL, Moore K, Paszkiewicz K, et al. A pelagic microbiome (viruses to protists) from a small cup of seawater. Viruses-Basel. 2017;9(3).

125. Wilson WH, Gilg IC, Moniruzzaman M, Field EK, Koren S, LeCleir GR, et al. Genomic exploration of individual giant ocean viruses. ISME Journal. 2017;11(8):1736-45.

126. Martinez-Hernandez F, Fornas O, Gomez ML, Bolduc B, de la Cruz Pena MJ, Martinez JM, et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nature Communications. 2017;8.

127. Dodds JA, Cole A. Microscopy and biology of *Uronema gigas*, a filamentous eukaryotic green-alga, and its associated tailed virus-like particle. Virology. 1980;100(1):156-65.

128. Schvarcz CR, Steward GF. A giant virus infecting green algae encodes key fermentation genes. Virology. 2018;518:423-33.

129. Weynberg KD, Allen MJ, Gilg IC, Scanlan DJ, Wilson WH. Genome sequence of *Ostreococcus tauri* Virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. Journal of Virology. 2011;85(9):4520-9.

130. Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, Grimsley N. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. Journal of Virology. 2010;84(24):12555-63.

131. Martinez JM, Boere A, Gilg L, van Lent JWM, Witte HJ, van Bleijswijk JDL, et al. New lipid envelope-containing dsDNA virus isolates infecting *Micromonas pusilla* reveal a separate phylogenetic group. Aquatic Microbial Ecology. 2015;74(1):17-28.

132. Brussaard CPD, Noordeloos AAM, Sandaa RA, Heldal M, Bratbak G. Discovery of a dsRNA virus infecting the marine photosynthetic protist *Micromonas pusilla*. Virology. 2004;319(2):280-91.

133. Maat DS, Biggs T, Evans C, van Bleijswijk JDL, van der Wel NN, Dutilh BE, et al. Characterization and temperature dependence of arctic *Micromonas polaris* viruses. Viruses-Basel. 2017;9(6).

134. Tomaru Y, Katanozaka N, Nishida K, Shirai Y, Tarutani K, Yamaguchi M, et al. Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga *Heterocapsa circularisquama*. Aquatic Microbial Ecology. 2004;34(3):207-18.

135. Kim J, Kim CH, Takano Y, Jang IK, Kim SW, Choi TJ. Isolation and physiological characterization of a new algicidal virus infecting the harmful dinoflagellate *Heterocapsa pygmaea*. Plant Pathology Journal. 2012;28(4):433-8.

136. Onji M, Nakano S, Suzuki S. Virus-like particles suppress growth of the red-tide-forming marine dinoflagellate *Gymnodinium mikimotoi*. Marine Biotechnology. 2003;5(5):435-42.

137. Bettarel Y, Kan J, Wang K, Williamson KE, Cooney S, Ribblett S, et al. Isolation and preliminary characterisation of a small nuclear inclusion virus infecting the diatom *Chaetoceros cf. gracilis*. Aquatic Microbial Ecology. 2005;40(2):103-14.

138. Nagasaki K, Tomaru Y, Takao Y, Nishida K, Shirai Y, Suzuki H, et al. Previously unknown virus infects marine diatom. Applied and Environmental Microbiology. 2005;71(7):3528-35.

139. Tomaru Y, Toyoda K, Suzuki H, Nagumo T, Kimura K, Takao Y. New single-stranded DNA virus with a unique genomic structure that infects marine diatom *Chaetoceros setoensis*. Scientific Reports. 2013;3.

140. Tomaru Y, Takao Y, Suzuki H, Nagumo T, Nagasaki K. Isolation and Characterization of a single-stranded RNA virus infecting the bloom-forming diatom *Chaetoceros socialis*. Applied and Environmental Microbiology. 2009;75(8):2375-81.

141. Tomaru Y, Takao Y, Suzuki H, Nagumo T, Koike K, Nagasaki K. Isolation and characterization of a single-stranded DNA virus infecting *Chaetoceros lorenzianus Grunow*. Applied and Environmental Microbiology. 2011;77(15):5285-93.

142. Tomaru Y, Shirai Y, Toyoda K, Nagasaki K. Isolation and characterisation of a single-stranded DNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus*. Aquatic Microbial Ecology. 2011;64(2):175-84.

143. Shirai Y, Tomaru Y, Takao Y, Suzuki H, Nagumo T, Nagasaki K. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus Meunier*. Applied and Environmental Microbiology. 2008;74(13):4022-7.

144. Kimura K, Tomaru Y. Isolation and characterization of a single-stranded DNA virus infecting the marine diatom *Chaetoceros sp Strain SS628-11* isolated from western Japan. Plos One. 2013;8(12).

145. Toyoda K, Kimura K, Hata N, Nakayama N, Nagasaki K, Tomaru Y. Isolation and characterization of a single-stranded DNA virus infecting the marine planktonic diatom *Chaetoceros sp (strain TG07-C28)*. Plankton & Benthos Research. 2012;7(1):20-8.

146. Tomaru Y, Shirai Y, Suzuki H, Nagumo T, Nagasaki K. Isolation and characterization of a new single-stranded DNA virus infecting the cosmopolitan marine diatom *Chaetoceros dehilis*. Aquatic Microbial Ecology. 2008;50(2):103-12.

147. Tomaru Y, Toyoda K, Kimura K, Takao Y, Sakurada K, Nakayama N, et al. Isolation and characterization of a single-stranded RNA virus that infects the marine planktonic diatom *Chaetoceros sp (SS08-C03)*. Phycological Research. 2013;61(1):27-36.

148. Eissler Y, Wang K, Chen F, Wommack KE, Coats DW. Ultrastructural characterization of the lytic cycle of an intracellular virus infecting the diatom *Chaetoceros cf wighamii* (*Bacillariophyceae*) from Chesapeake Bay, USA. Journal of Phycology. 2009;45(4):787-97.

149. Tomaru Y, Toyoda K, Kimura K, Hata N, Yoshida M, Nagasaki K. First evidence for the existence of pennate diatom viruses. ISME Journal. 2012;6(7):1445-8.

150. Nagasaki K, Tomaru Y, Katanozaka N, Shirai Y, Nishida K, Itakura S, et al. Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. Applied and Environmental Microbiology. 2004;70(2):704-11.

151. Kim J, Kim CH, Youn SH, Choi TJ. Isolation and physiological characterization of a novel algicidal virus infecting the marine diatom *Skeletonema costatum*. Plant Pathology Journal. 2015;31(2):186-91.

152. Kim J, Yoon SH, Choi TJ. Isolation and physiological characterization of a novel virus infecting *Stephanopyxis palmeriana* (*Bacillariophyta*). Algae. 2015;30(2):81-7.

153. Kapp M, Knippers R, Mueller DG. New members of a group of DNA viruses infecting brown algae. Phycological Research. 1997;45(2):85-90.

154. Henry EC, Meints RH. A persistent virus-infection in *Feldmannia* (*Phaeophyceae*). Journal of Phycology. 1992;28(4):517-26.

155. Maier I, Wolf S, Delaroque N, Muller DG, Kawai H. A DNA virus infecting the marine brown alga *Pilayella littoralis* (*Ectocarpales, Phaeophyceae*) in culture. European Journal of Phycology. 1998;33(3):213-20.

156. Nagasaki K, Yamaguchi M. Isolation of a virus infectious to the harmful bloom causing microalga *Heterosigma akashiwo* (*Raphidophyceae*). Aquatic Microbial Ecology. 1997;13(2):135-40.

157. Tai V, Lawrence JE, Lang AS, Chan AM, Culley AI, Suttle CA. Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (*Raphidophyceae*). Journal of Phycology. 2003;39(2):343-52.

158. Lawrence JE, Chan AM, Suttle CA. A novel virus (HaNIV) causes lysis of the toxic bloom-forming alga *Heterosigma akashiwo* (*Raphidophyceae*). Journal of Phycology. 2001;37(2):216-22.

159. Castberg T, Thyrhaug R, Larsen A, Sandaa RA, Heldal M, Van Etten JL, et al. Isolation and characterization of a virus that infects *Emiliania huxleyi* (*Haptophyta*). Journal of Phycology. 2002;38(4):767-74.

160. Baudoux AC, Brussaard CPD. Characterization of different viruses infecting the marine harmful algal bloom species *Phaeocystis globosa*. Virology. 2005;341(1):80-90.

161. Wilson WH, Schroeder DC, Ho J, Canty M. Phylogenetic analysis of PgV-102P, a new virus from the English Channel that infects *Phaeocystis globosa*. Journal of the Marine Biological Association of the United Kingdom. 2006;86(3):485-90.

162. Jacobsen A, Bratbak G, Heldal M. Isolation and characterization of a virus infecting *Phaeocystis pouchetii* (*Prymnesiophyceae*). Journal of Phycology. 1996;32(6):923-7.

163. Suttle CA, Chan AM. Viruses infecting the marine prymnesiophyte *Chrysochromulina spp*. - isolation, preliminary characterization, and natural-abundance. Marine Ecology Progress Series. 1995;118(1-3):275-82.

164. Mirza SF, Staniewski MA, Short CM, Long AM, Chaban YV, Short SM. Isolation and characterization of a virus infecting the freshwater algae *Chrysochromulina parva*. Virology. 2015;486:105-15.

165. Johannessen TV, Bratbak G, Larsen A, Ogata H, Egge ES, Edvardsen B, et al. Characterisation of three novel giant viruses reveals huge diversity among viruses infecting *Prymnesiales* (*Haptophyta*). Virology. 2015;476:180-8.

166. Wagstaff BA, Vladu IC, Barclay JE, Schroeder DC, Malin G, Field RA. Isolation and characterization of a double stranded DNA megavirus infecting the toxin-producing haptophyte *Prymnesium parvum*. Viruses-Basel. 2017;9(3).

167. Nagasaki K, Kim J-J, Tomaru Y, Takao Y, Nagai S. Isolation and characterization of a novel virus infecting *Teleaulax amphioxeia* (*Cryptophyceae*). Plankton & Benthos Research. 2009;4(3):122-4.

168. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 2016;33:1870-4.

169. Lefever S, Pattyn F, Hellemans J, Vandesompele J. Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. Clinical Chemistry. 2013;59(10):1470-80.

170. Gulvik CA, Effler TC, Wilhelm SW, Buchan A. De-MetaST-BLAST: A tool for the validation of degenerate primer sets and data mining of publicly available metagenomes. Plos One. 2012;7(11).

Appendix

Figure 1.1 Timeline of eukaryotic algal virus research

Colored bars represent the annual citations and publications generated from a Web of Science Citation Report using the field tag TS = (algal virus) for all databases. The search was conducted on 8 May 2018 at 11:00 a.m. Citation Report results were visualized as heatmaps using custom R scripts. Electron micrograph image (127) and electrophoretic gel (19) reprinted by permission. Network analysis (31) reprinted under authority of Creative Commons.

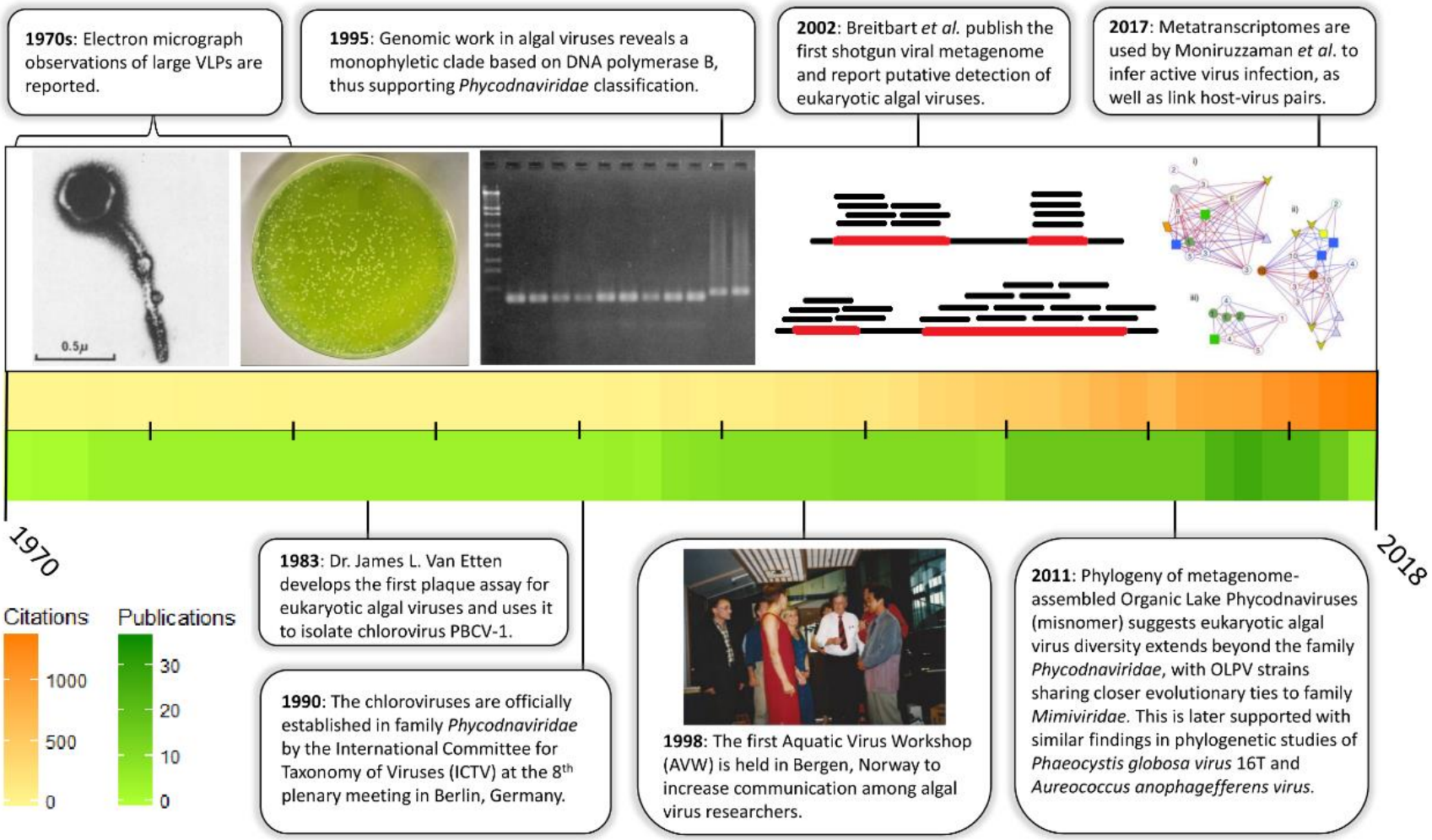

1970s: Electron micrograph observations of large VLPs are reported.
1995: Genomic work in algal viruses reveals a monophyletic clade based on DNA polymerase B, thus supporting *Phycodnaviridae* classification.
2002: Breitbart *et al.* publish the first shotgun viral metagenome and report putative detection of eukaryotic algal viruses.
2017: Metatranscriptomes are used by Moniruzzaman *et al.* to infer active virus infection, as well as link host-virus pairs.
0.5µ
1970
2018
Citations
1000
500
0
Publications
30
20
10
0
1983: Dr. James L. Van Etten develops the first plaque assay for eukaryotic algal viruses and uses it to isolate chlorovirus PBCV-1.
1990: The chloroviruses are officially established in family *Phycodnaviridae* by the International Committee for Taxonomy of Viruses (ICTV) at the 8th plenary meeting in Berlin, Germany.
1998: The first Aquatic Virus Workshop (AVW) is held in Bergen, Norway to increase communication among algal virus researchers.
2011: Phylogeny of metagenome-assembled Organic Lake Phycodnaviruses (misnomer) suggests eukaryotic algal virus diversity extends beyond the family *Phycodnaviridae*, with OLPV strains sharing closer evolutionary ties to family *Mimiviridae*. This is later supported with similar findings in phylogenetic studies of *Phaeocystis globosa virus* 16T and *Aureococcus anophagefferens virus*.

Figure 1.2 Known virus interactions with eukaryotic algal lineages

(a) Schematic phylogeny adapted from de Vargas et al. demonstrating known virus-interactions with eukaryotic alga lineages. The phylogeny was originally constructed on recognized eukaryotic plankton lineages that were detected in TARA Oceans datasets, which included hits to all aquatic algal containing lineages. We collapsed the original tree to highlight these lineages in the context of their current phylogenetic placement. Green lines denote lineages with photosynthetic algal representatives, whereas the text color indicates whether all or only some representatives are phototrophic-green or black text, respectively; (b) Yellow boxes denote the top ten most abundant, planktonic, phototroph-associated lineages based on 18S rDNA surveyed in the TARA Oceans study. Asterisks denote lineages that were artificially grouped for simplicity, and their full descriptions can be found at http://taraoceans.sb-roscoff.fr/EukDiv/; (c) Red boxes denote algal-lineages that have an isolated algae-infecting virus in culture collection, though these are not all marine systems. The virus isolates are listed in Table 1.

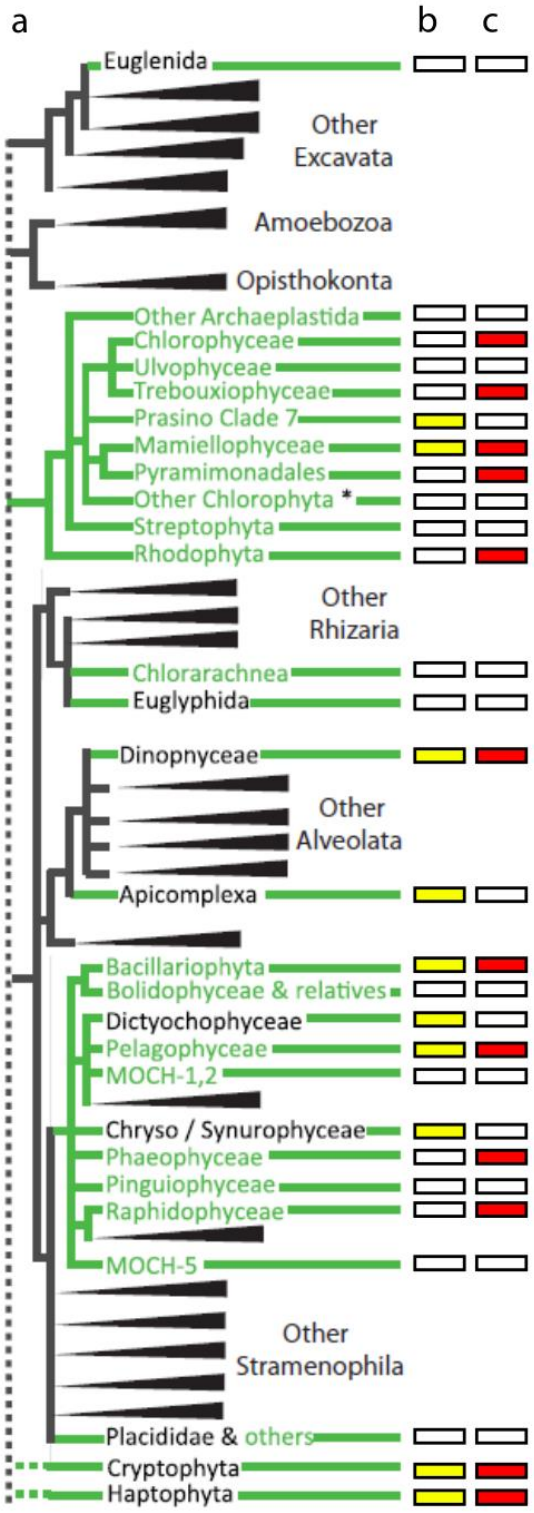

a
b
c
Euglenida
Other Excavata
Amoebozoa
Opisthokonta
Other Archaeplastida
Chlorophyceae
Ulvophyceae
Trebouxiophyceae
Prasino Clade 7
Mamiellophyceae
Pyramimonadales
Other Chlorophyta *
Streptophyta
Rhodophyta
Other Rhizaria
Chlorarachnea
Euglyphida
Dinopnyceae
Other Alveolata
Apicomplexa
Bacillariophyta
Bolidophyceae & relatives
Dictyochophyceae
Pelagophyceae
MOCH-1,2
Chryso / Synurophyceae
Phaeophyceae
Pinguiophyceae
Raphidophyceae
MOCH-5
Other Stramenophila
Placididae & others
Cryptophyta
Haptophyta

Table 1.1 Algal viruses currently in culture collection

Host Algae	Type	Size (kbp or knt)	Code	References
Chlorophyceaea				
Tetraselmis spp.	dsDNA	668	TetV	Schvarcz et al. 2018 (128)
Tetraselmis striata	dsDNA	31	Tsv-N1	Pagarete et al. 2015 (53)
Trebouxiophyceae				
Chlorella variabilis NC64A	dsDNA	287–369	PBCV-1	Jeanniard et al. 2013 (14)
Chlorella variabilis Syngen 2-3	dsDNA	327	OSy-NE5	Quispe et al. 2017 (15)
Chlorella heliozoae SAG 3.83	dsDNA	288–327	ATCV-1	Jeanniard et al. 2013 (14)
Micratinium conductrix Pbi	dsDNA	302–329	CVM	Jeanniard et al. 2013 (14)
Mamiellophyceae				
Ostreococcus lucimarinus	dsDNA	182–196	OlV1	Derelle et al. 2015 (58)
Ostreococcus tauri	dsDNA	184–192	OtV5	Weynberg et al. 2011 (129)
Ostreococcus mediterraneus	dsDNA	193	OmV1	Derelle et al. 2015 (58)
Bathycoccus sp. RCC1105	dsDNA	187–198	BpV	Moreau et al. 2010 (130)
Micromonas pusilla CCMP1545	dsDNA	186–195	MpV-02T	Martinez Martinez et al. 2015 (131)
Micromonas pusilla LAC38	dsDNA	173–205	MpV1	Finke et al. 2017 (59)
Micromonas pusilla LAC38	dsRNA	25.5	MpRV	Brussaard et al. 2004 (132)
Micromonas polaris	dsDNA	191–205	MpoV	Maat et al. 2017 (133)
Pyramimonadales				
Pyramimonas orientalis	dsDNA	560	PoV	Sandaa et al. 2001 (46)
Rhodophyta				
Chondrus crispus	dsRNA	6	CcV	Rousvoal et al. 2016 (72)
Dinophyceaea				
Heterocapsa circularisquama	dsDNA	356	HcDNAV	Ogata et al. 2009 (56)
Heterocapsa circularisquama	ssRNA	4.4	HcRNAV	Tomaru et al. 2004 (134)
Heterocapsa pygmea	dsDNA	ND	HpygDNAV	Kim et al. 2012 (135)
Gymnodinium mikimotoi	ND	ND	GM6/GM7	Onji et al. 2003 (136)
Bacillariophyta				
Chaetoceros cf. *gracilise*	ND	ND	CspNIV	Bettarel et al. 2005 (137)
Chaetoceros salsugineum	ssDNA	6	CsalDNAV*	Nagasaki et al. 2005 (138)
Chaetoceros setoensis	ssDNA	5.8	CsetDNAV*	Tomaru et al. 2013 (139)

Table 1.1 (continued)

Host Algae	Type	Size (kbp or knt)	Code	References
Chaetoceros socialis f. *radians*	ssRNA	9.4	CsfrRNAV	Tomaru et al. 2009b (140)
Chaetoceros lorenzianus	ssDNA	5.8	ClorDNAV*	Tomaru et al. 2011 (141)
Chaetoceros tenuissimus	ssDNA	5.6	CtenDNAV-I*	Tomaru et al. 2011 (142)
Chaetoceros tenuissimus	ssDNA	5.6	CtenDNAV-II*	Kimura and Tomaru 2015 (66)
Chaetoceros tenuissimus	ssRNA	9.4	CtenRNAV	Shirai et al. 2008 (143)
Chaetoceros tenuissimus, *Chaetoceros* spp.	ssRNA	9.6	CtenRNAV-II	Kimura and Tomaru 2015 (66)
Chaetoceros spp. SS628-11	ssDNA	5.5	Csp07DNAV*	Kimura et al. 2013 (144)
Chaetoceros spp. TG07-C28	ssDNA	ND	Csp05DNAV	Toyoda et al. 2012 (145)
Chaetoceros debilis	ssDNA	ND	CdebDNAV	Tomaru et al. 2008 (146)
Chaetoceros sp. SS08-C03	ssRNA	9.4	Csp03RNAV	Tomaru et al. 2013 (147)
Chaetoceros cf. *wighamii*	ssDNA	7-8	CwNIV	Eissler et al. 2009 (148)
Asterionellopsis glacialis	ssRNA	9.5	AglaRNAV	Tomaru et al. 2012 (149)
Thalassionema nitzschioides	ssDNA	5.5	TnitDNAV	Tomaru et al. 2012 (149)
Rhizosolenia setigera	ssRNA	11.2	RsetRNAV	Nagasaki et al. 2004 (150)
Skeletonema costatum	ND	ND	ScosV	Kim et al. 2015 (151)
Stephanopyxis palmeriana	ND	ND	SpalV	Kim et al. 2015 (152)
Pelagophyceae				
Aureococcus anophagefferens	dsDNA	370	AaV	Moniruzzaman et al. 2014 (21)
Phaeophyceae				
Ectocarpus fasciculatus	dsDNA	340	EfasV	Kapp et al. 1997 (153)
Ectocarpus siliculosus	dsDNA	320	EsV	Kapp et al. 1997 (153)
Feldmannia irregularis	dsDNA	180	FirrV	Kapp et al. 1997 (153)
Feldmannia simplex	dsDNA	220	FlexV	Kapp et al. 1997 (153)
Feldmannia species	dsDNA	170	FsV	Henry and Meints 1992 (154)
Hincksia hinckiae	dsDNA	240	HincV	Kapp et al. 1997 (153)
Myriotrichia clavaeformis	dsDNA	320	MclaV	Kapp et al. 1997 (153)
Pilayella littoralis	dsDNA	280	PlitV	Maier et al. 1998 (155)

Table 1.1 (continued)

Host Algae	Type	Size (kbp or knt)	Code	References
Raphidophyceae				
Heterosigma akashiwo	dsDNA	ND	HaV	Nagasaki et al. 1997 (156)
Heterosigma akashiwo	dsDNA	180	O1s1	Lawrence et al. 2006 (50)
Heterosigma akashiwo	ssRNA	9.1	HaRNAV	Tai et al. 2003 (157)
Heterosigma akashiwo	ND	ND	HaNIV	Lawrence et al. 2001 (158)
Haptophyta				
Emiliania huxleyi	dsDNA	415	EhV	Castberg et al. 2002 (159)
Phaeocystis globosa	dsDNA	466	PgV-16T (Group I)	Baudoux et al. 2005 (160)
Phaeocystis globosa	dsDNA	177	PgV-03T (Group II)	Baudoux et al. 2005 (160)
Phaeocystis globosa	dsDNA	176	PgV-102P	Wilson et al. 2006 (161)
Phaeocystis pouchetii	dsDNA	485	PpV	Jacobsen et al. 1996 (162)
Chrysochromulina brevifilum, Chrysochromulina strobilus	dsDNA	ND	CbV	Suttle and Chan 1995 (163)
Chrysochromulina ericina	dsDNA	510	CeV	Sandaa et al. 2001 (46)
Chrysochromulina parva	dsDNA	485	CpV	Mirza et al. 2015 (164)
Haptolina ericina, Prymnesium kappa	dsDNA	530	HeV-RF02	Johannessen et al. 2015 (165)
Prymnesium kappa, Haptolina ericina	dsDNA	ND	PkV-RF01	Johannessen et al. 2015 (165)
Prymnesium kappa	dsDNA	507	PkV-RF02	Johannessen et al. 2015 (165)
Prymnesium parvum	dsDNA	ND	PpDNAV	Wagstaff et al. 2017 (166)
Cryptophyta				
Teleaulax amphioxeia	ND	ND	TampV	Nagasaki et al. 2009 (167)

Table 1. Summary of all reported eukaryotic algal viruses that have been isolated. A range of genome sizes (kbp or knt) represents multiple virus strains associated with the same host species, and in this case, only the type virus is reported under the code column. Asterisks denote original names for some of the diatom ssDNA viruses, which have since been renamed and placed into genera of the family *Bacilladnaviridae* (*Chaetoceros setoensis* DNA virus = *Diatodnavirus*; *Chaetoceros salsugineum* DNA virus 1 = *Chaetoceros protobacilladnavirus* 1; *Chaetoceros* sp. DNA virus 7 = *Chaetoceros protobacilladnavirus* 2; *Chaetoceros lorenzianus* DNA virus = *Chaetoceros protobacilladnavirus* 3; *Chaetoceros tenuissimus* DNA viruses type I and II = *Chaetoceros protobacilladnavirus* 4). ND = Not detected or reported.

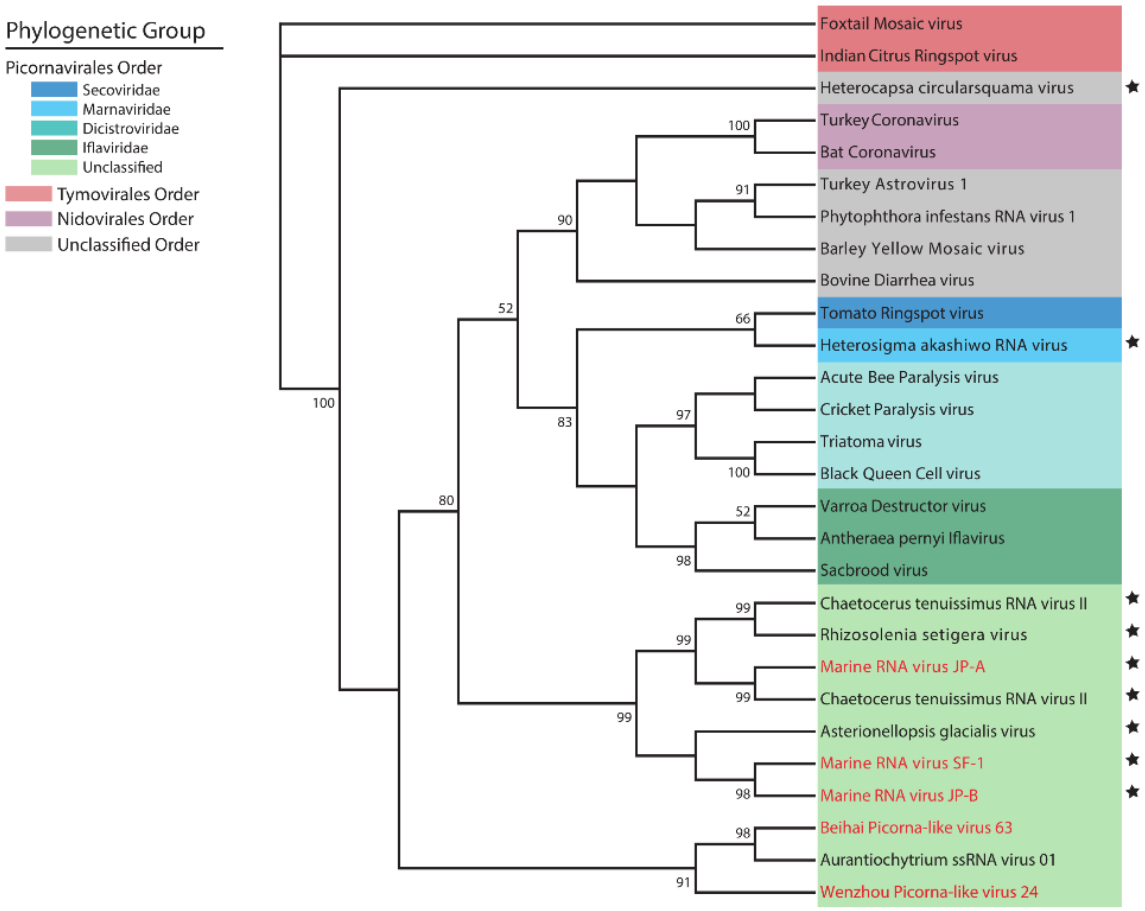


Figure 1.3 Diversity of single-stranded RNA viruses

Diversity is depicted based on phylogeny of RNA-dependent RNA polymerase (Rdrp NCBI CDD:01699) reference sequences downloaded from NCBI RefSeq database. Sequences were aligned and trimmed in Mega7 (168) and an unrooted maximum likelihood phylogeny was created using PhyML 3.0 with LG model (113). Empirical equilibrium frequencies were used with aLRT SH-like statistics for branch support. Phylogenetic groups are color coded with algal viruses denoted by a star. Viral isolates from metagenomic assemblies are in red text.

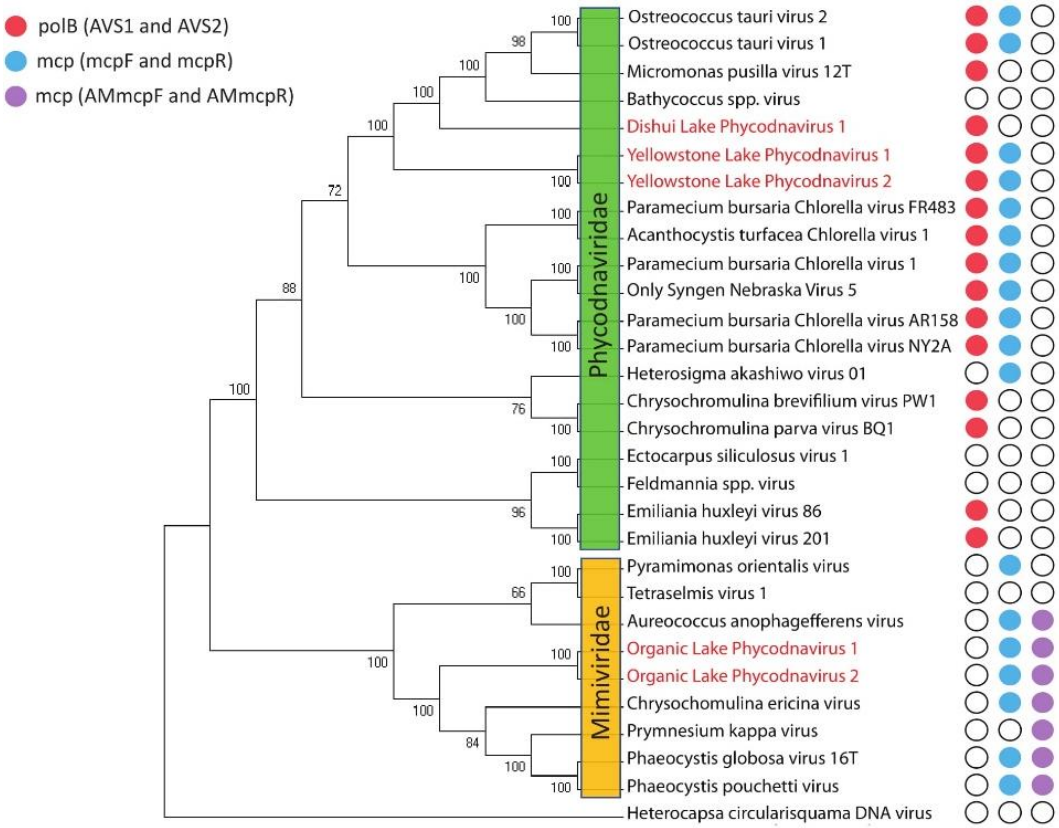


Figure 1.4 Diversity of dsDNA viruses and specificity of published primers

Phylogenetic tree depicting the evolutionary relationships of algal NCLDVs based on amino acid alignment (ClustalW) of the core gene, DNA polymerase B. The tree was built using the maximum likelihood method based on the JTT matrix-based model with 200 iterations in MEGA7 (168). Viruses belong either to the family *Phycodnaviridae* or are recognized “extended members” of the family *Mimiviridae*. The recently discovered dinoflagellate infecting virus, *Heterocapsa circularisquama* DNA virus, was used to root the tree and shows little similarity to other algal NCLDVs despite being a large DNA virus. Viruses in red text denote metagenome assembled viral genomes, meaning their association with an alga host is putative. Colored dots to the right indicate the viruses can be putatively PCR amplified by the respective PCR primer set based on ≥90% match between each primer and its respective target binding site. This equates to ≤2 primer mismatches, which has been shown to be capable of producing a PCR reaction, albeit at lower efficiency (for RT-qPCR) (169). The same study shows that three or more mismatches in the same primer completely inhibit a PCR reaction, and is an observation that aligns with failed PCR reactions reported for *Ectocarpus siliculosus* virus 1 and *Feldmannia spp*. virus (19). PCR amplification predictions were done using motif searches in CLC Genomics and the software De-MetaST-BLAST (170).

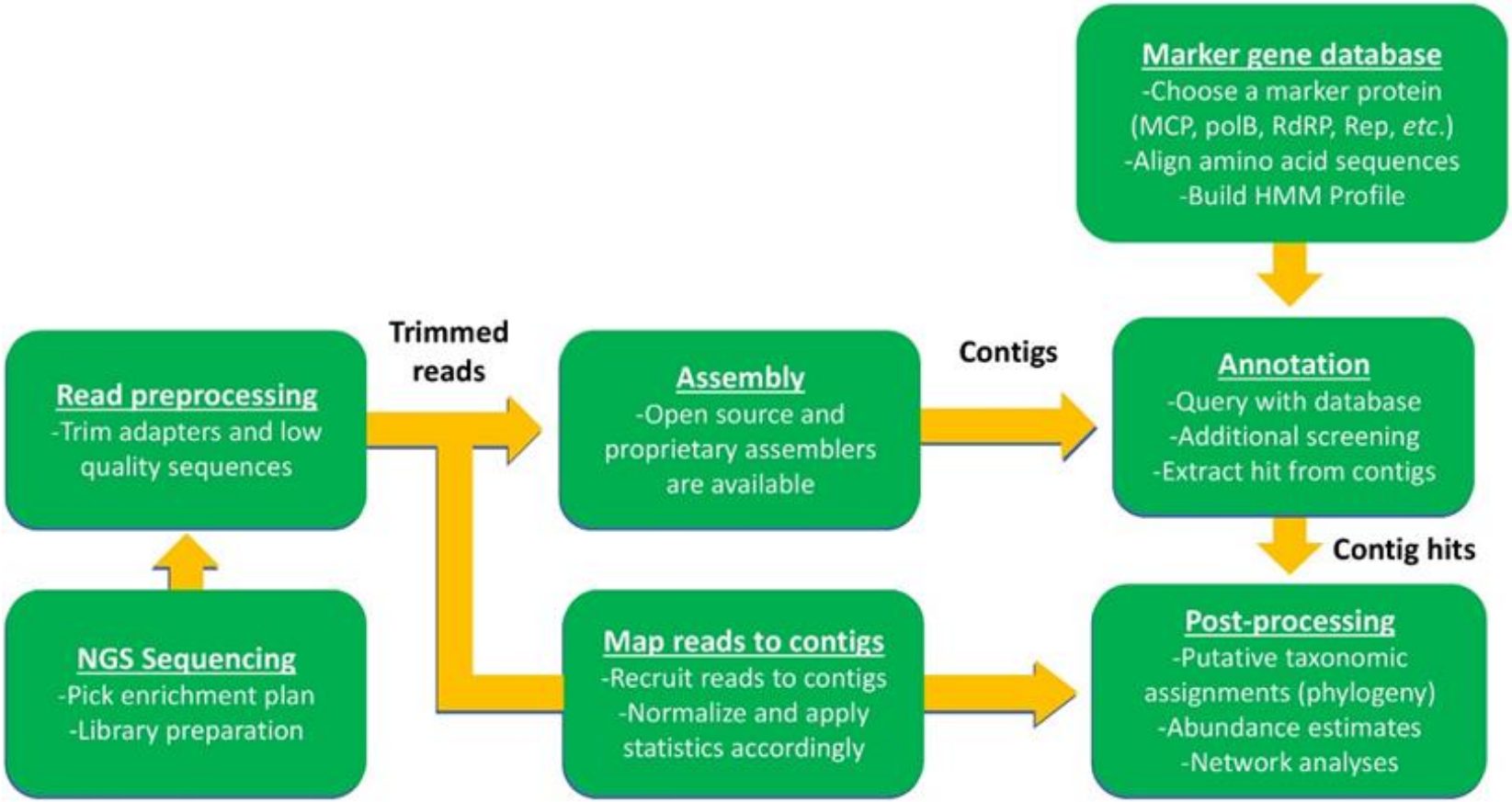


Figure 1.5 General bioinformatic pipeline using marker gene probing

This framework follows that used by Moniruzzaman et al. 2017 (31), where viral activity was assessed using marker gene detection from environmental mRNA. Though this framework was modeled off the cited study, it is flexible enough to incorporate both metagenomic and metatranscriptomic applications.

CHAPTER 2 : CRYOPRESERVATION OF PBCV-1 DURING AN ACTIVE INFECTION CYCLE OF ITS HOST

A version of this chapter was originally published in *PLoS One*:

Samantha R. Coy, Alyssa N. Alsante, James L. Van Etten, Steven W. Wilhelm. “Cryopreservation of Paramecium bursaria Chlorella Virus-1 during an active infection cycle of its host.” *PLoS One.* 2018;14(3): e0211755.

S.R.C. and S.W.W. conceived the paper, and all listed authors contributed to the production of figures, text, and editing.

Abstract

Best practices in laboratory culture management often include cryopreservation of microbiota, but this can be challenging with some virus particles. By preserving viral isolates researchers can mitigate genetic drift and laboratory-induced selection, thereby maintaining genetically consistent strains between experiments. To this end, we developed a method to cryopreserve the model, green-alga infecting virus, *Paramecium bursaria Chlorella virus 1* (PBCV-1). We explored cryotolerance of the infectivity of this virus particle, whereby freezing without cryoprotectants was found to maintain the highest infectivity (~2.5%). We then assessed the cryopreservation potential of PBCV-1 during an active infection cycle in its *Chlorella variabilis* NC64A host, and found that virus survivorship was highest (69.5 ± 16.5 %) when the infected host is cryopreserved during mid-late stages of infection (*i.e.,* coinciding with virion assembly). The most optimal condition for cryopreservation was observed at 240 minutes post-infection. Overall, utilizing the cell as a vehicle for viral cryopreservation resulted in 24.9 – 30.1 fold increases in PBCV-1 survival based on 95% confidence intervals of frozen virus particles and virus cryopreserved at 240 minutes post-infection. Given that cryoprotectants are often naturally produced by psychrophilic organisms, we suspect that cryopreservation of infected hosts may be a reliable mechanism for virus persistence in non-growth permitting circumstances in the environment, such as ancient permafrosts.

Introduction

Viruses are abundant components of all biological systems and they likely infect every lineage of eukaryotic algae. Their impact is most readily noticed following infection and lysis of abundant bloom forming algae (1-3), though lytic activity of all algal viruses contributes to significant biomass recycling *via* the 'viral shunt' (4). To date, 65 eukaryotic algal viruses have been isolated and developed as laboratory strains (5, 6). Most of these are maintained through serial propagation on their respective hosts. Though this has been effective for culturing many strains over the last few decades (7, 8), each passage allows for genetic mutations that can accumulate in a population (9), leading to a deviation from a standard 'wild-type.' Moreover, it is imperative to control evolution following the development of genetically tractable algal hosts (10) and (ultimately) virus systems. Although seed-stock systems can be developed without cryopreservation, many systems are not amenable to this either because the virus particles are degraded during purification efforts or lose their infectivity during storage. Moreover, it can take time to achieve axenic status with new virus isolates, thus making contaminating bacterial activity a significant source of degradation. Thus, a protocol for successful virus cryobiological preservation that is applicable to a wide variety of algae-virus systems would offer an opportunity to universally improve virus management and distribution in the laboratory.

Cryopreservation is not a new concept in biological sciences. For most protocols, it involves controlled cooling of biota to sub-freezing temperatures to achieve biological cessation while preserving viability. This most often manifests as slow-cooling at a rate of 1° C / min in the presence of osmoprotectant(s) (*e.g.,* dimethylsulfoxide (DMSO), glycerol) for long-term storage at -130° C or below (11). Too slow a cooling rate can result in higher intracellular concentration of osmoprotectants, resulting in toxicity, whereas too fast a cooling rate allows the formation of intracellular ice crystals which can rupture cell membranes (12). The thawing process is typically quick, as microbial death is commonly associated with slow thaw rates. Though cryopreservation is a standard method for

maintaining cellular organisms, it has rarely been utilized for the preservation of algal viruses.

One eukaryotic algal virus cryopreservation protocol is in existence. It was developed for HaV, a dsDNA virus that infects the red tide forming dinoflagellate *Heterosigma akashiwo* (13). Researchers investigated a combination of cryoprotectants and storage temperatures with the highest recovery (8.3% of infectious virus) employing flash freezing of HaV particles suspended in 20% DMSO. This protocol has been adapted for a handful of other algal viruses with viable recovery ranging from < 1% to 27% (14-16). The typical low recovery in these procedures is likely due to physiological differences between viruses and cells including differences in permeability, osmolarity tolerance, and toxicity to osmoprotectants. It is also clear that these protocols deviate from the standard method which controls the cooling rate; to our knowledge this has not been tested as a matter of improving virus particle survival. Owing to these complications, we decided to take a new approach by investigating cryopreservation recovery and stability of actively infecting, cell-associated algal viruses.

Chloroviruses are large (> 300 kb) dsDNA viruses in the family *Phycodnaviridae* (17). They are members of the proposed order the Megavirales (18), also known as “giant” viruses, and remain the best characterized algal-virus system to date. Isolated in the early 1980’s (7), the prototype chlorovirus *Paramecium bursaria Chlorella virus 1* (PBCV-1) has been maintained through serial propagation on its host, *Chlorella variabilis* NC64A. PBCV-1 is inactivated by freezing, though other closely related virus strains, including other chloroviruses, persist through freeze/thaw events (19, 20). As a great deal of research has centered on PBCV-1, including genomics (21), transcriptomics (22, 23), and proteomics (21), it is important to develop a successful cryopreservation protocol for this strain that may serve as a model for preserving algal viruses. There are several reports of cryopreservation techniques for eukaryotic algae (24-28) which might be adapted for the preservation of actively replicating chloroviruses.

Here, we tested the cryo-potential of chlorovirus PBCV-1 using a protocol that yielded consistent recovery (~50% viable cells) of four strains of algae over 15 years: *Chlorella vulgaris* C-27, *Chlorella vulgaris* M-207A7, *Nannochloropsis oculate* ST-4, and *Tetraselmis tetrathlele* T-501 (29). Owing to the close relationship between *C. vulgaris* and *C. variabilis,* as well as the consistent results across unique algae, we elected to determine if these results could be recapitulated in PBCV-1. To test this, we attempted cryopreservation of both the virus particle as well as the virus replicating in its host.

Materials and methods

Virus particle cryopreservation

Chlorella variabilis NC64A was infected with PBCV-1 during mid-logarithmic growth at standard culturing conditions (25°C; continuous light exposure at 30µEin/m2/s) using Modified Bold's Basal Medium (30). Following complete lysis, the viral lysate was pre-filtered through a sterile, 0.45 µm polycarbonate syringe filter and titered by plaque assay (31, 32) for initial infectivity assessments. Cryoprotectant choice was guided by Nakanishi et al. (29), in which a combination of 5% DMSO (v/v), 5% ethylene glycol (v/v), and 5% proline (w/v) was found to consistently produce the highest algal recoveries. Stock solutions of each cryoprotectant were made at a concentration of 30% with sterilized Milli-Q water and combined in a 1:1:1 ratio to yield a final concentration of 10% for each compound. For virus particle cryopreservation, 1 mL of PBCV-1 particles (7.82x 108 plaque forming units (PFUs) per ml) was added to 1 mL of ice-chilled cryoprotectant solution contained in a 2-mL cryovial. The cryovials were incubated on ice for 45 min, then transferred to a freeze-rate controlled container (Mr. Frosty, Thermo Fisher Scientific Inc., USA) filled with isopropanol for overnight incubation at -80° C. The next morning, cryovials were transferred to a -150° C freezer. At the designated recovery times, vials were removed from the freezer and set in a 40° C water bath. After thawing, the samples were serially diluted ten-fold in 50 mM Tris-HCl (pH = 7.8) and virus infectivity was determined by plaque assay (31). Virus viability was calculated as a percentage by

comparison to the initial virus particle stock titer before cryopreservation. Long-term experiments assessed the stability of virus infectivity in particles stored at -150° C.

Infected Chlorella cryopreservation

Chlorovirus PBCV-1 was propagated as described above and titered to obtain infectious PFUs/ml. This virus particle stock was used to infect late-logarithmically growing *C. variabilis* NC64A at an M.O.I. of 5, at which point infected cultures were returned to standard incubation conditions. At 1, 10, 30, 60, 120, 180, 240, 300, and 360 min post-infection (PI), 1 mL aliquots of infected cells were mixed with 1 mL of ice-chilled cryoprotectants [final concentration: 5% DMSO (v/v), 5% ethylene glycol (v/v), and 5% proline (w/v)] in duplicates. The mixture was incubated on ice for 45 min, then transferred to a freeze-rate controlled container (Mr. Frosty, Thermo Fisher Scientific Inc., USA has a -1C/min cooling rate) filled with isopropanol for overnight incubation at -80° C. The next morning, cryovials were immediately transferred to a -150° C freezer. At the designated recovery times, vials were removed from the freezer and placed in a 40° C water bath. After thawing, the infected cells were pelleted in a Sorvall Legend RT Benchtop Centrifuge at 3,700 rpm (~3,000 rcf) for 10 min: (free virus requires higher speeds for pelleting). Cell pellets were re-suspended in 2 mL of 0.01M HEPES solution (pH = 6.5). Suspensions were immediately diluted and plaque assayed, plating late-infection treatments first. Viability was determined as a percentage of the pre-frozen cellular concentration (3.57×10^6 cells/mL), as only surviving infected cells would be capable of producing plaques. Long-term experiments were conducted in the same manner, though only time points 10, 180, and 240 min PI were collected and assayed. The complete step-by-step method can be found at protocols.io (33).

Results

Following the cryopreservation procedures of other algal virus researchers (13-16), we investigated the cryo-potential of the PBCV-1 particle. Cryoprotectant alone treatments elicited a lethal effect: ~87% of the infectious virus particles were inactivated

in the presence of these chemicals following 24 hr exposure at 4° C. Given this effect, we decided to freeze PBCV-1 particles at -150° C without any cryoprotectants. This resulted in ~2.5% recovery of the infectious virus population, which was stable for storage periods of up to one year (Fig 1). Seeing room for improvement, we tested the cryo-potential of PBCV-1 in an infected, cell-associated state.

The PBCV-1 replication cycle requires about 6-8 h to release nascent virus particles (34). Post-infection sampling times for cryopreservation (10, 30, 60, 120, 180, 240, 300, 360 min PI) followed similar sampling strategies used in PBCV-1 transcription studies (22, 23). Specifically, these time points were collected across distinct physiological phases in the PBCV-1 lifecycle and thus represent likely unique conditions for cryopreservation. Following 24-h storage of cryopreserved, infected cells, we found that late stages of infection were more conducive to virus survival than early stages (Fig 2). Thus, we followed cryo-stability for one year in one early (10 min PI) and two late infection stages (180 and 240 min PI) (Fig 3). Small day-to-day fluctuations in virus titers were common, but were typically consistent among treatments, suggesting human error. Despite these fluctuations, the virus particle stock control, 180-min, and 240-min PI treatment yielded an acceptable relative standard deviation (RSD) for these plate counts (35) across all recovery assessments, indicating cryo-stability (Table 1). Cryo-stability was not observed in the 10 min PI samples (Table 1). In comparison to virus particle cryopreservation, the cell-associated method yielded significant improvement in survivorship for the optimal 240-minute treatment (24.9 – 30.1 fold increases).

Discussion

The current maintenance strategy for chloroviruses involves serial propagation on the alga host followed by lysate particle storage at 4°C. Chloroviruses are relatively stable under these conditions, though even PBCV-1 is known to degrade after several years of storage. In any case, many algae-virus systems are less amenable to long-term storage at 4°C. For example, new algae-virus systems are not always quickly made axenic, and are thus susceptible to degradation from contaminating bacteria. On the other hand,

viruses propagated on axenic hosts can still degrade. For reasons unknown, chloroviruses are more stable in lysates (bacterial-free) than in particle stocks purified by sucrose density gradients (36), but they always eventually lose their infectivity. Serial propagation of viruses is therefore often required. Even if this is done infrequently, it can still promote genetic drift and result in deviation from wild-type status. This is concerning for all virus types, though RNA viruses, which have the fastest mutation rates, would be most susceptible (9, 37). Beyond considering spontaneous, replication-associated errors, chloroviruses encode putative enzymes involved in genomic rearrangements. For example, GIY-YIG mobile endonucleases and an IS607 transposon may be involved in insertions/deletions and/or gene loss/duplications observed in genomic comparisons of chloroviruses (38, 39). Thus, maintenance of wild-type strains is important for consistency between experiments. Virology labs could follow the microbial culture collection strategy, which typically uses a cryo-banking/seed-stock system for the dissemination of microbial specimens. The purpose of the seed-stock system is to minimize serial propagation of microbiota. The American Type Culture Collection (ATCC) suggests that consumers transfer their cultures no more than five-times after propagation from the thawed culture collection stock. Though a seemingly strict standard, it is not difficult to imagine the consequences of violating this. For example, the United States Pharmacopeia and National Formulary requires test organisms to be maintained this way for routine antibiotic efficacy screens, and non-compliance can undermine therapeutic treatment (35). Although there is no direct clinical link to maintaining algal viruses this way, the logic is consistent with any research requirements. The cryopreservation protocol described here can help researchers better set up these cryo-banking/seed stock systems.

Standard cryopreservation techniques are not designed for the unique structure and physiology of virus particles. Indeed, cryoprotectants are classified by their permeability across cell membranes, which often coincides with their molecular weight (24). Smaller compounds, such as ethylene glycol and DMSO, are considered penetrating cryoprotectants, while larger compounds (*e.g.* amino acids; L-proline) are typically non-penetrating. That said, the exclusion size threshold has not been established for most

viruses so it is not clear which, if any of these compounds penetrate the viral capsid. It is generally thought that virus capsids are permeable to water and ions, though the latter diffuses much slower; this mechanism has been used to osmotically rupture capsids (40, 41), including PBCV-1 (42). The final cryoprotectant solution used for PBCV-1 particle cryopreservation has an estimated osmolarity of ~150 mOsmoles/L, which is comparable to the storage buffer used for this virus. In light of this, we propose that the lethal effect the cryoprotectants have on the PBCV-1 particle is not the result of osmotic stress, and that inactivation instead occurred by toxicity of cryoprotectants or oxidative stress. This would be consistent with viruses not being metabolically active and therefore unable to repair damage caused by this treatment. It is also consistent with the observation that Mimivirus, a giant virus relative which also contains an internal lipid membrane, is said to be inactivated by lipophilic compounds such as DMSO (43). That said, DMSO is often used as a stabilizer for freezing of enveloped virus particles (44). This discrepancy may be due to unique properties between external and internal membranes, or even system differences between animal and plant viruses, which imparts resistance in some cases over others. Regardless, the mechanism of inactivation may be better ascertained by looking at survivorship of virion particles via epifluorescent microscopy, flow cytometry (45-47), or using bioassays to quantify oxidative stress.

Although the algal cell is in a sub-optimal physiological state during infection, it is apparently robust enough to survive and maintain an active infection during cryopreservation. That said, fewer infectious virus were recovered when the cell was cryopreserved during early infection stages. This might be explained by differences in adsorption rates and synchronicity of infection, resulting in fewer infected cells at the start of the experiment. Most, if not all cells are infected at the later stages of infection (3-4 hr PI). Regardless of any differences in synchronicity, the algal cell will be completely arrested during cryopreservation, and will only continue the infection cycle after thawing. Internal, mature viruses that have not yet lysed their host cell might still be inactivated by cryoprotectants, thus reducing viral burst size, but our experiments did not account for this. We also did not account for inefficiencies in infection rates; though we infected at

M.O.I. values based on infectious particle counts, it is possible that all the cells were not infected. Had we plated the infected cell population prior to cryoprotection we could have corrected for this in our results. In any case, accounting for infection inefficiency can only improve PBCV-1 survivorship and the success of our method.

The general classification of cryoprotectants based on membrane permeability is consistent in the infected cell treatment. Although the *C. variabilis* NC64A genome encodes a secondary active transporter for the uptake of proline, radio-labeled solute uptake experiments revealed that PBCV-1 infection abolishes its activity (48). With that in mind, the tonicity of the cryoprotectant mixture would equate to ~90 mOsmoles/L, as only DMSO and ethylene glycol are penetrating, and many of the components in the MBBM media would be spent by late-logarithmic growth. This concentration is comparable to buffers routinely used in our lab for handling *C. variabilis* (40 mOsmoles/L), so there is little concern of osmotic stress. The chances of osmotic stress were also low considering the consistent success associated with this cryopreservation formula across eukaryotic algae, including two *Chlorella* spp. (29). Our results are likely applicable to any algal virus whose host can be cryopreserved. That said, we expect that researchers may still have to adjust their cryoprotectant mixture to account for system differences related to osmolarity tolerance and cryoprotectant toxicity. There has also been research indicating that axenicity impacts cryopreservation survival in microalgae. In this light, it is possible that the bacterial community produces secondary metabolites which promote survival (49). In another scenario, organisms with psychrophilic tendencies might be adapted to freeze situations and cryoprotectant additives may not be necessary.

The goal of this study was to develop a long-term cryopreservation method for chlorovirus PBCV-1, but there are also interesting ecological implications of this research. Recent metagenomic and isolation efforts indicate that giant viruses of microeukaryotes (e.g., *Phycodnaviridae* and *Mimiviridae)* are widely distributed in nature (50, 51), but it is not well understood how these viruses persist in the environment. Freezing events represent a potential mechanism of inactivation for some algal viruses, though chlorovirus ATCV-1 is stable during these conditions (19). In two other studies, a closely related giant

virus of the family *Mimiviridae* (52), as well as a second giant virus in the family *Molliviridae* (53), were revived from 30,000 year old permafrost. Both of these viruses were revived using *Acanthamoeba spp.,* one of the main hosts for many giant viruses. That said, there have been questions about whether *Acanathamoeba* and other protists used for laboratory viral propagation are the natural or primary hosts of these ancient viruses (54). Although these viruses might be able to withstand freezing temperatures on their own, the results of this study suggest that a natural host might serve as a better vehicle for surviving freezing. Indeed, many microbes produce natural cryoprotectants (*e.g.* L-proline, trehalose, betatine, etc.) or encode machinery to transport these osmoprotectants into the cell. Following this thought process, it is possible that environments containing frozen, infected cells might contain naturally cryopreserved algal-virus systems. These systems may be deciphered following advances in single-cell sorting and sequencing techniques. Indeed, a similar approach has been successfully utilized to identify and sequence single virus genomes in the ocean (55). Though this latter study sorted virus particles, flow-cytometry sorting of viral infected cells may be achieved using fluorescent probes specific for viral marker genes (*e.g.,* major capsid protein) or dyes to detect viral-induced host phenotypes (*e.g.,* membrane blebbing). As a proof of concept, viral genetic sequences recovered from Siberian permafrost could be used to probe for still frozen viral-infected host cells, thereby testing the natural host range of these viruses.

To our knowledge, this is the first report of successful cryopreservation of a eukaryotic algal virus during its infection cycle. We expect that respective cellular hosts will provide more suitable physiological conditions for cryopreservation and storage of algal viruses that infect eukaryotic algae. We also recommend that laboratories working with algal viruses establish cryopreserved seed-stock systems to better preserve wild-type controls for future experimentation, especially in lieu of future modification of these viral systems.

References

1. Bratbak G, Egge JK, Heldal M. Viral mortality of the marine alga *Emiliania-huxleyi* (*Haptophyceae*) and termination of algal blooms. Marine Ecology Progress Series. 1993;93(1-2):39-48.

2. Rowe JM, Dunlap JR, Gobler CJ, Anderson OR, Gastrich MD, Wilhelm SW. Isolation of a non-phage-like lytic virus infecting *Aureococcus anophagefferens*. Journal of Phycology. 2008;44(1):71-6.

3. Nagasaki K, Tomaru Y, Nakanishi K, Hata N, Katanozaka N, Yamaguchi M. Dynamics of *Heterocapsa circularisquama* (*Dinophyceae*) and its viruses in Ago Bay, Japan. Aquatic Microbial Ecology. 2004;34(3):219-26.

4. Wilhelm SW, Suttle CA. Viruses and nutrient cycles in the sea - viruses play critical roles in the structure and function of aquatic food webs. BioScience. 1999;49(10):781-8.

5. Coy SR, Gann ER, Pound HL, Short SM, Wilhelm SW. Viruses of eukaryotic algae: diversity, methods for detection, and future directions. Viruses. 2018;10(9).

6. Short SM, Staniewski MA, Chaban YV, Long AM, Wang D. Diversity of viruses infecting eukaryotic algae. In: P. H, Abedon ST, editors. Viruses of microorganisms. Poole, UK: Caister Academic Press; 2018. p. 211-44.

7. Van Etten JL, Burbank DE, Xia Y, Meints RH. Growth-cycle of a virus, PBCV-1, that infects *Chlorella*-like algae. Virology. 1983;126(1):117-25.

8. Castberg T, Thyrhaug R, Larsen A, Sandaa RA, Heldal M, Van Etten JL, et al. Isolation and characterization of a virus that infects *Emiliania huxleyi* (Haptophyta). Journal of Phycology. 2002;38(4):767-74.

9. Peck KM, Lauring AS. Complexities of viral mutation rates. Journal of Virology. 2018;92(14):8.

10. Waller RF, Cleves PA, Rubio-Brotons M, Woods A, Bender SJ, Edgcomb V, et al. Strength in numbers: collaborative science for new experimental model systems. PloS Biology. 2018;16(7):10.

11. Mazur P. Freezing of living cells - mechanisms and implications. American Journal of Physiology. 1984;247(3):C125-C42.

12. Mazur P, Leibo SP, Chu EHY. A two-factor hypothesis of freezing injury - evidence from chineses-hamster tissue-culture cells. Experimental Cell Research. 1972;71(2):345-55.

13. Nagasaki K, Yamaguchi M. Cryopreservation of a virus (HaV) infecting a harmful bloom causing microalga, Heterosigma akashiwo (Raphidophyceae). Fisheries Science. 1999;65(2):319-20.

14. Kim J, Kim CH, Youn SH, Choi TJ. Isolation and physiological characterization of a novel algicidal virus infecting the marine diatom *Skeletonema costatum*. Plant Pathology Journal. 2015;31(2):186-91.

15. Kim J, Yoon SH, Choi TJ. Isolation and physiological characterization of a novel virus infecting *Stephanopyxis palmeriana* (*Bacillariophyta*). Algae. 2015;30(2):81-7.

16. Kim J, Kim CH, Takano Y, Jang IK, Kim SW, Choi TJ. Isolation and physiological characterization of a new algicidal virus infecting the harmful dinoflagellate *Heterocapsa pygmaea*. Plant Pathology Journal. 2012;28(4):433-8.

17. Jeanniard A, Dunigan DD, Gurnon JR, Agarkova IV, Kang M, Vitek J, et al. Towards defining the chloroviruses: a genomic journey through a genus of large DNA viruses. BMC Genomics. 2013;14.

18. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, et al. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Archives of Virology. 2013;158(12):2517-21.

19. Long AM, Short SM. Seasonal determinations of algal virus decay rates reveal overwintering in a temperate freshwater pond. ISME Journal. 2016;10(7):1602-12.

20. Bubeck JA, Pfitzner AJP. Isolation and characterization of a new type of chlorovirus that infects an endosymblotic *Chlorella* strain of the heliozoon *Acanthocystis turfacea*. Journal of General Virology. 2005;86:2871-7.

21. Dunigan DD, Cerny RL, Bauman AT, Roach JC, Lane LC, Agarkova IV, et al. Paramecium bursaria chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. Journal of Virology. 2012;86(16):8821-34.

22. Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova IV, et al. Microarray analysis of paramecium bursaria chlorella birus 1 transcription. Journal of Virology. 2010;84(1):532-42.

23. Blanc G, Mozar M, Agarkova IV, Gurnon JR, Yanai-Balser G, Rowe JM, et al. Deep RNA sequencing reveals hidden features and dynamics of early gene transcription in paramecium bursaria chlorella virus 1. PloS One. 2014;9(3):10.

24. Hubalek Z. Protectants used in the cryopreservation of microorganisms. Cryobiology. 2003;46(3):205-29.

25. Benson EE. Cryopreservation of phytodiversity: A critical appraisal of theory practice. Critical Reviews in Plant Sciences. 2008;27(3):141-219.

26. Day JG, Watanabe MM, Morris GJ, Fleck RA, McLellan MR. Long-term viability of preserved eukaryotic algae. Journal of Applied Phycology. 1997;9(2):121-7.

27. Taylor R, Fletcher RL. Cryopreservation of eukaryotic algae - a review of methodologies. Journal of Applied Phycology. 1998;10(5):481-501.

28. Rhodes L, Smith J, Tervit R, Roberts R, Adamson J, Adams S, et al. Cryopreservation of economically valuable marine micro-algae in the classes *Bacillariophyceae, Chlorophyceae, Cyanophyceae, Dinophyceae, Haptophyceae, Prasinophyceae, and Rhodophyceae*. Cryobiology. 2006;52(1):152-6.

29. Nakanishi K, Deuchi K, Kuwano K. Cryopreservation of four valuable strains of microalgae, including viability and characteristics during 15 years of cryostorage. Journal of Applied Phycology. 2012;24(6):1381-5.

30. Dunigan DD, Agarkova I. Formulation of MBBM (modified Bold's Basal medium). protocolsio. 2016: doi 10.17504/protocols.io.etwbepe.

31. Van Etten JL, Burbank DE, Kuczmarski D, Meints RH. Virus-infection of culturable *Chlorella*-like algae and development of a plaque assay. Science. 1983;219(4587):994-6.

32. Dunigan DD, Agarkova I. PBCV-1 virus plaque assay. protocolsio. 2016: doi 10.17504/protocols.io.estbeen.

33. Coy SR, Alsante A, Wilhelm SW. Long term cryopreservation of chloroviruses by infection of *Chlorella*. protocolsio. 2018: doi 10.17504/protocols.io.wa2fage.

34. Dunigan DD, Fitzgerald LA, Van Etten JL. Phycodnaviruses: a peek at genetic diversity. Virus Research. 2006;117(1):119-32.

35. Convention. USP. U.S. Pharmacopeia National Formulary 2018: UPS41-NF36: Nielsen bookata; 2018.

36. Agarkova I, Hertel B, Zhang XZ, Lane L, Tchourbanov A, Dunigan DD, et al. Dynamic attachment of Chlorovirus PBCV-1 to Chlorella variabilis. Virology. 2014;466:95-102.

37. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral nutation rates. Journal of Virology. 2010;84(19):9733-48.

38. Filee J, Pouget N, Chandler M. Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. BMC Evolutionary Biology. 2008;8(320).

39. Filee J, Siguier P, Chandler M. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends in Genetics. 2007;23(1):10-5.

40. Cordova A, Deserno M, Gelbart WM, Ben-Shaul A. Osmotic shock and the strength of viral capsids. Biophysical Journal. 2003;85(1):70-4.

41. Roos WH, Ivanovska IL, Evilevitch A, Wuite GJL. Viral capsids: mechanical characteristics, genome packaging and delivery mechanisms. Cellular and Molecular Life Sciences. 2007;64(12):1484-97.

42. Wulfmeyer T, Polzer C, Hiepler G, Hamacher K, Shoeman R, Dunigan DD, et al. Structural organization of DNA in *Chlorella* viruses. PLoS One. 2012;7(2).

43. Claverie JM, Abergel C. Virus Taxonomy Ninth Report of the International Committee on Taxonomy of Viruses: Elsevier Inc.; 2012. p. 223-8.

44. Wallis C, Melnick JL. Stabilization of enveloped viruses by dimethyl sulfoxide. Journal of Virology. 1968;2(9):953-4.

45. Noble RT, Fuhrman JA. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquatic Microbial Ecology. 1998;14(2):113-8.

46. Brussaard CPD, Marie D, Bratbak G. Flow cytometric detection of viruses. Journal of Virological Methods. 2000;85(1-2):175-82.

47. Brussaard CPD. Optimization of procedures for counting viruses by flow cytometry. Applied and Environmental Microbiology. 2004;70(3):1506-13.

48. Agarkova I, Dunigan D, Gurnon J, Greiner T, Barres J, Thiel G, et al. Chlorovirus-mediated membrane depolarization of chlorella alters secondary active transport of solutes. Journal of Virology. 2008;82(24):12181-90.

49. Amaral R, Pereira JC, Pais A, Santos LMA. Is axenicity crucial to cryopreserve rnicroalgae? Cryobiology. 2013;67(3):312-20.

50. Wilhelm SW, Coy SR, Gann ER, Moniruzzaman M, Stough JMA. Standing on the shoulders of giant viruses: five lessons learned about large viruses infecting small eukaryotes and the opportunities they create. PLoS Pathogens. 2016;12(8).

51. Kerepesi C, Grolmusz V. The "giant virus finder" discovers an abundance of giant viruses in the Antarctic dry valleys. Archives of Virology. 2017;162(6):1671-6.

52. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(11):4274-9.

53. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, et al. In-depth study of *Mollivirus sibericum*, a new 30,000-y-old giant virus infecting *Acanthamoeba*. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(38):E5327-E35.

54. Wilhelm SW, Bird JT, Bonifer KS, Calfee BC, Chen T, Coy SR, et al. A student's guide to giant viruses infecting small eukaryotes: from *Acanthamoeba* to *Zooxanthellae*. Viruses. 2017;9(3).

55. Wilson WH, Gilg IC, Moniruzzaman M, Field EK, Koren S, LeCleir GR, et al. Genomic exploration of individual giant ocean viruses. ISME Journal. 2017;11(8):1736-45.

Appendix

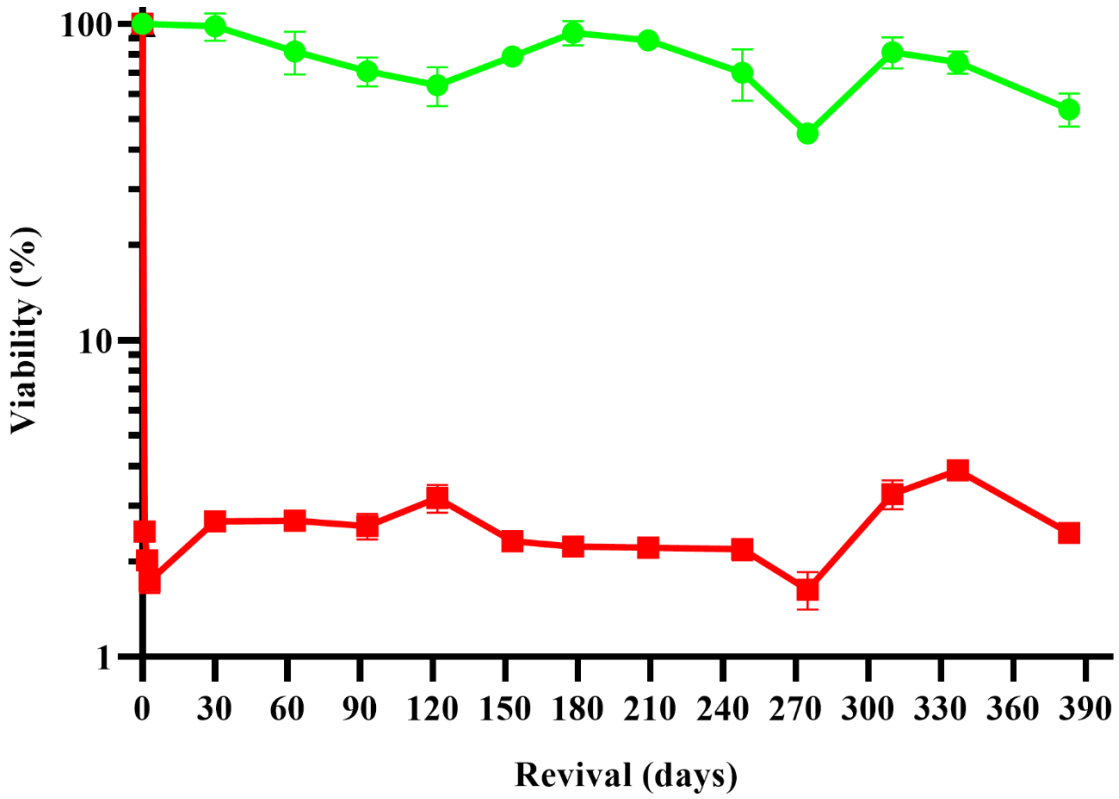


Figure 2.1 Cryo-stability of the PBCV-1 particle

Viability of chlorovirus PBCV-1 was determined by plaque assaying viruses that had been stored as particles either at 4°C or -150°C. Green circles represent virus particles stored at 4° C, while red squares denote virus particles stored at -150° C. Error bars are represented as the standard deviation of biological and technical replicates.

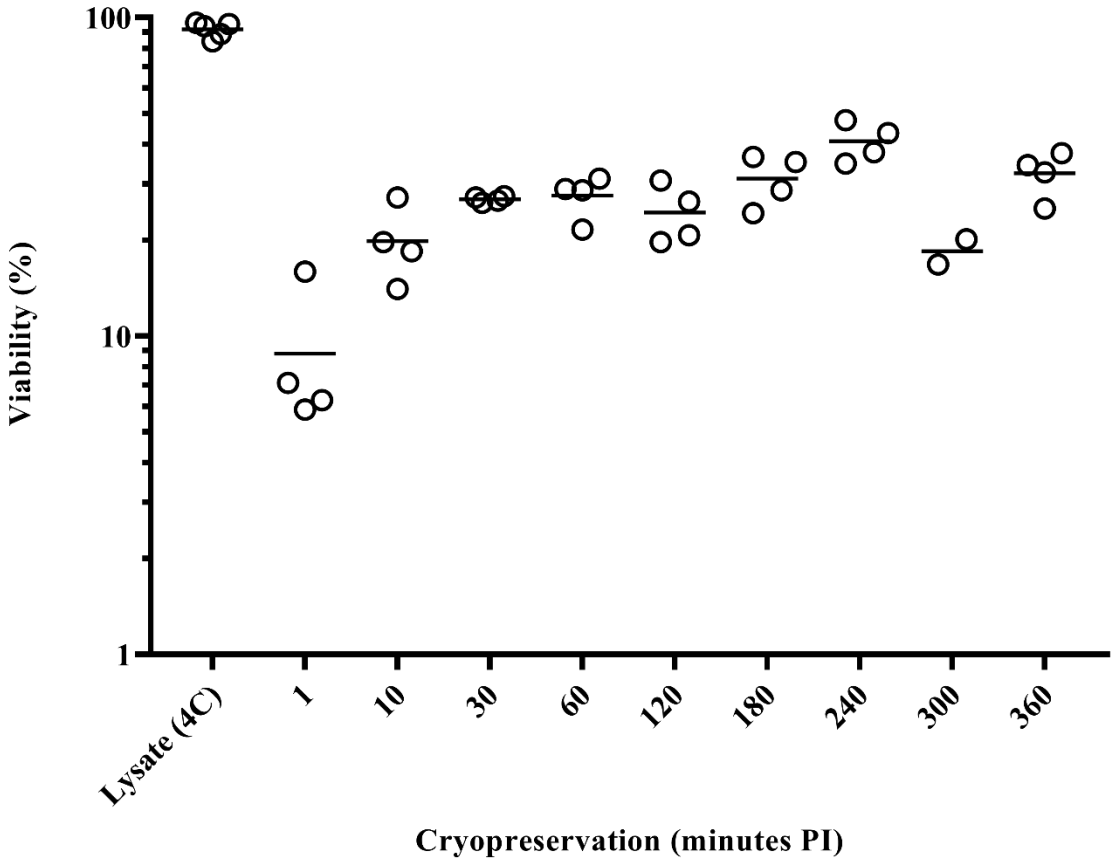


Figure 2.2 Recovery of infectious PBCV-1 frozen at different times post infection

Viability of chlorovirus PBCV-1 was assayed by monitoring plaque formation of cell-associated viruses that were collected at different times during an active infection cycle of the NC64A host. Open circles denote replicate plaque titers, with the average represented by the solid line.

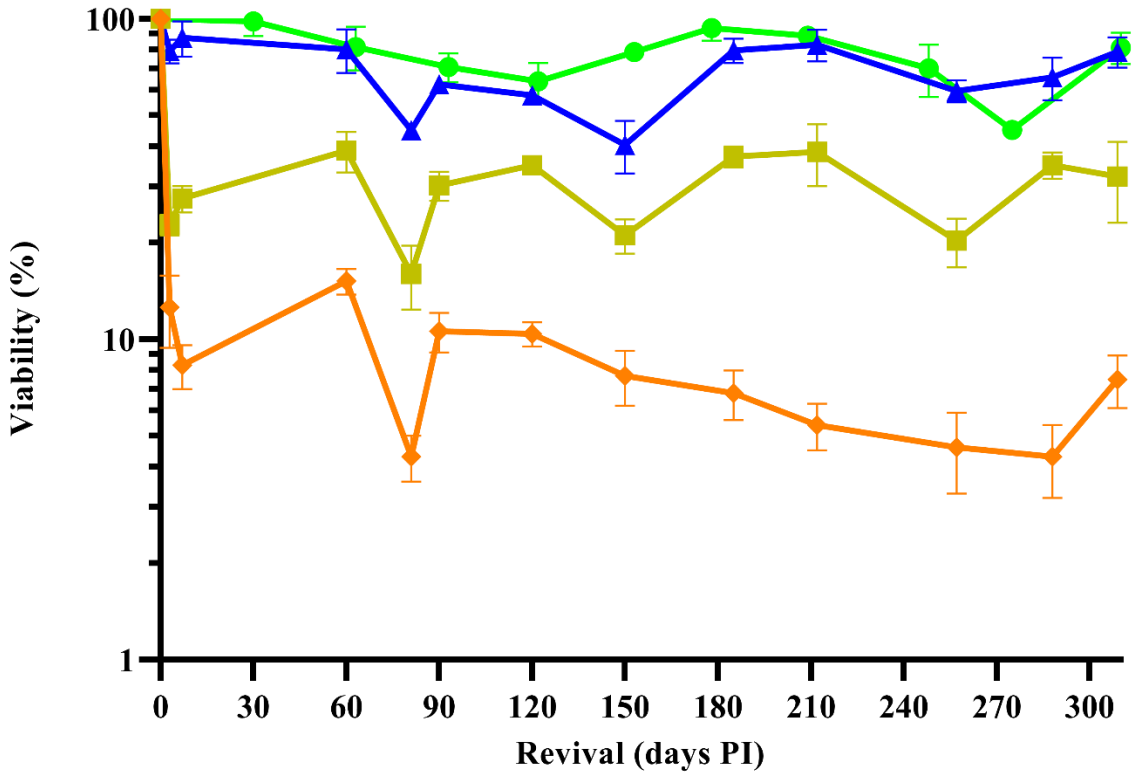


Figure 2.3 Long-term cryo-stability of PBCV-1 frozen during host infection

Infectious chlorovirus PBCV-1 was monitored by plaque assay in virus particle stocks stored at 4° C (green circles) and in cryopreserved, PBCV-1-infected host cultures. Blue triangles, yellow squares, and orange diamonds represent virus viability following storage of infected cells cryopreserved after 240, 180, and 10 minutes PI. Error bars represent the standard deviation among biological and technical replicates.

Table 2.1 Statistical assessment of PBCV-1 infectivity across one year of storage

Treatment	N	Average	SD	RSD	95%CI
Virus Particle Stock (4° C)	67	75.1	16.9	22.5*	71.1 - 79.2
Virus Particle Stock (-150° C)	124	2.53	0.61	24.0*	2.42 – 2.64
Cell-associated virus 10 minutes PI (-150° C, +CPA)	79	7.56	3.38	44.7	6.81 – 8.31
Cell-associated virus 180 minutes PI (-150° C, +CPA)	82	31.9	10.9	34.2*	29.5 – 34.3
Cell-associated virus 240 minutes PI (-150° C, +CPA)	82	69.5	16.5	23.8*	65.9 – 73.0

+CPA, cryoprotectants present as described in materials and methods section. Asterisks (*) denote an acceptable RSD (*i.e.*, Coefficient of Variation) for plaque assays based on a 35% threshold used in bacterial plating standards set from chapter 1223 by the U.S. Pharmacopeia and National Formulary.

CHAPTER 3 : SMRT SEQUENCING OF PBCV-1 VIRIONS REVEALS DYNAMIC METHYLATION PATTERNS IN ADENINES TARGETED BY RESTRICITON MODIFICATION SYSTEMS

This chapter is in prep for manuscript submission. :

Samantha R. Coy, Eric R. Gann, Spiridon E. Papoulis, Michael E. Holder, Nadim J. Ajami, Joseph F. Petrosino, Erik R. Zinser, James L. Van Etten, Steve W. Wilhelm. “SMRT sequencing of *Paramecium bursaria Chlorella Virus-1* reveals dynamic methylation patterns in adenines targeted by restriction modification systems.”

S.R.C. performed the experiment, wrote the R code to organize the data, analyzed data, and wrote the manuscript. E.R.G wrote the python code used to conduct some of the *in silico* analyses for motif enrichment. S.E.P. wrote and documented the code to identify methyltransferases in publicly available viral genomes, and produced the data used to generate Figure 3.3. M.E.H., N.J.A., and J.F.P. performed SMRT sequencing and initial *de novo* genome assembly. J.L.V. provided research materials and direction for the study. S.W.W. guided conception of the paper and provided funding for research materials. All listed authors contributed to the production of figures, text, and editing.

Abstract

Chloroviruses (family Phycodnaviridae) infect eukaryotic, freshwater, unicellular green algae. A unique feature of these viruses is an abundance of DNA methyltransferases (MTases), with strains dedicating anywhere between 0 - 4.5% of their protein coding potential to these genes. This diversity highlights just one of the long-standing values of the chlorovirus model system, where group-wide epigenomic characterization might begin to elucidate the function(s) of DNA methylation in large dsDNA viruses. We characterized DNA modifications in the prototype chlorovirus, PBCV-1, using single molecule real time (SMRT) sequencing (aka PacBio). This was contrasted against total available sites predicted in silico based on DNA sequence alone. The SMRT-software detected N6-methyl-adenine (m6A) at GATC and CATG recognition sites, which are known methylation motifs associated with enzymes M.CviAI and M.CviAII, respectively. At the same time, PacBio analyses indicated that 10.9% of the PBCV-1 genome is associated with large interpulse duration ratio (ipdRatio) values, the primary

metric for DNA modification identification. This represents 20.6x more sites than all available target adenines in CATG and GATC motifs, and contrasts against analyses in Escherichia coli wherein all sites with similar ipdRatio values can be accounted for by known motifs. Cross comparisons of methylation status between biological replicates for each target tetramer indicate ~81% of sites are stably methylated, and ~2% are stably unmethylated. The remaining 17% of sites are stochastically methylated between the biological replicates. When these are paired together with their palindromic reverse compliments, we show that palindromes exist in completely non-methylated states, fully methylated states, and hemi-methylated states. Given these sites are targeted by not just methyltransferases, but by restriction endonucleases that are encoded by PBCV-1 as virus-originating restriction modification (RM) systems, there is a strong selective pressure to modify all target sites. The finding that most instances of non-methylation are associated with hemi-methylation is congruent with observations that hemi-methylated palindromes are resistant to cleavage by restriction endonucleases recognizing these sequences. However, that some hemi-methylated sites are consistently not methylated might represent a unique biological function for PBCV-1. This study serves as a baseline for future investigation into the epigenomics of chloroviruses and their giant virus relatives.

Introduction

Viruses infecting eukaryotic algae play a critical role in aquatic ecosystems. Their lytic activity results in redistribution of organic matter from the particulate pool into a dissolved to particulate continuum that can be assimilated by the remaining microbial community, thus providing ecosystem stability and driving biodiversity (1-5). In the oceans, this activity is modeled to account for the daily, virus-driven cycling of up to a quarter of the total organic carbon in the surface oceans (6).

Although viruses are the most abundant entities in aquatic systems, only a relatively small number of eukaryotic algal viruses have been isolated and are maintained in laboratories (7, 8). This group represents diverse nucleic acid types, architectures, and

sizes, but the majority of isolates have large, double-stranded DNA genomes. A phylogenetic comparison of these large viruses places them into the proposed monophyletic order, the *Megavirales* (9), also more commonly known as ‘giant’ viruses. Giant viruses have long interested researchers for not only their size, often overlapping with that of cellular organisms (10), but their unique genomic content. Indeed, these entities encode genes not normally observed in a virus, including central components of protein translation and DNA repair. Of interest to this study is an unusually high number of DNA methyltransferases. For instance, viruses that infect freshwater chlorella-like green algae encode up to 18 distinct DNA methyltransferases, representing up to ~4.5% of their total protein coding potential (11). Many of these enzymes share only distant homology to methyltransferases encoded by cellular organisms, and thus represent enzymes uniquely adapted to these viruses.

DNA methyltransferases catalyze the transfer of a reactive methyl group from the common cellular metabolite, S-adenosyl-methionine, to form either methylated cytosine (m4C, m5C) or methylated adenine (m6A). Most known enzymes have evolved to modify cytosines or adenines in specific nucleotide sequences that are recognized by the enzyme. These often range from two to six base-pairs, though there are examples of more complicated and promiscuous recognition sequences (12). In any case, the direct consequence of DNA methylation is a change in the primary and secondary structure of the DNA (13), which can influence a variety of DNA recognition and protein binding interactions.

DNA methylation is known to regulate a diverse number of physiological processes. The most well studied examples include its role as a silencing molecule in gene expression, and as a component of bacterial restriction modification systems. However, a multitude of novel functions that are directed by DNA methylation have been recently identified (14-16), suggesting this is a major regulatory molecule for a variety of DNA-protein activities. Given this modification is found in many algal-infecting giant viruses, we suspect that it is used to confer functions not previously known to promote

virus fitness. Before this question can be investigated, however, it is necessary to quantify the distribution and stability of DNA methylation in some of the model virus systems.

The chloroviruses are the model system for studying giant, algal-infecting viruses. Isolated over 35 years ago, this system has expanded to include several hundred virus strains acquired across the globe (17). Genomic comparisons of these isolates unveiled a great diversity of DNA methyltransferases with strains encoding up to 18 distinct enzymes. The number of encoded enzymes aligns well with the amount of methylated DNA measured in each respective viral genome, ranging from 0.12% to 47.5% and 0% to 37% of cytosines and adenines, respectively (18, 19). The prototype chlorovirus strain, PBCV-1, has been subjected to genomic, transcriptomic, and proteomic studies (20-22), thus providing a rich biological context from which to assess its 'epigenomic' state. Previous studies have determined that 1.86% of cytosines (m5C) and 1.45% of adenines (m6A) are methylated in the PBCV-1 genome (18, 19), and that these modifications arise from the activity of up to five putative methyltransferases (Table 1). Target sequences of only the m6A decorating methyltransferases have been deduced, and a full inventory of their genomic occurrence suggests that only a fraction of these sites require m6A modification to yield an m6A pool size of 1.45%. This perceived incomplete methylation of all target sites represents a potential regulatory function for the virus, and thus warrants a higher resolution analysis of DNA methylation in PBCV-1.

In this study, we characterized methylation pattern and stability by *in silico* bioinformatic analysis of potential methylation sites occurring in the PBCV-1 genome followed with site specific measurements using single-molecule real time (SMRT) sequencing (aka PacBio). This technique allowed us to determine the methylation status of each target site, as well as its stability across space (*i.e.* separate virus populations) and time (*i.e.* multiple generations). During sequencing, we also identified signatures of DNA modification other than methylation and predict that these modifications bear important consequences for viral infectivity. This study establishes a baseline for future investigation into DNA modifications in *Chlorella* viruses and serves as a model for initiating these studies in other algal-infecting giant viruses.

Materials and methods

Distribution of methyltransferases encoded by viruses

To identify methyltransferases in publicly available viral genomes, we chose a strategy that uses both BLAST 2.7.1+ and HMMER 3.1b2 to generate alignments to a reference database derived from experimentally characterized 'Gold Standard' methyltransferases found in New England Biolabs' REBASE (12). To find protein profiles that identify functional motifs of methyltransferases, we used hmmscan with gathering cutoffs to collect Pfams (from release 31) (23) represented in our Gold Standard database including: PF05869.11 (Dam), PF00145.17 (DNA_methylase), PF07669.11 (Eco57I), PF13651.6 (EcoRI_methylase), PF12161.8 (HsdM_N), PF02086.15 (MethyltransfD12), PF02384.16 (N6_Mtase), PF01555.18 (N6_N4_Mtase), and PF12564.8 (TypeIII_RM_meth). Gold Standard methyltransferases that did not contain one of the Pfams used in this study were queried in BLAST searches to identify putative methyltransferases in viral proteomes and can be found in 'BLASTexceptions.fasta'. A viral protein was considered a methyltransferase if a protein profile aligned with hmmsearch exceeded gathering cutoffs or the viral protein aligned to a methyltransferase sequence via BLAST with the query protein being at least 75% of the alignment length and the evalue was <1E-5.

Viral assembly metadata from RefSeq, GenBank, and NCBI taxonomy were downloaded on June 26th, 2018 and stored in 'Viral_assemblydat.tsv' and 'nodes.dmp'. If a virus was included in both RefSeq and GenBank, the RefSeq assembly was preferentially used as a query. Because viral assemblies deposited in GenBank are inconsistently annotated with coding sequences, we chose to translate all frames in these viral genomes to avoid differences in annotation approaches. Methyltransferase pfams and characterized methyltransferase sequences lacking pfams (outlined above) were queried with HMMER and BLAST, respectively, to annotate putative methyltransferases found in the translated frames of each viral genome. To map viruses to the hosts they infect, mappings were downloaded from the Virus-Host DB

(https://www.genome.jp/virushostdb/). Jupyter notebooks and associated source code used for methyltransferase annotation can be found at www.github.com/SEpapoulis/MTannotation.

PBCV-1 in-silico analysis of DNA modification predictions

Functional motif sequences known to be targeted by PBCV-1 methyltransferases (Table 1) are individually considered as potential methylation sites *in silico*. That said, it is useful to determine the distribution of these sites to make hypotheses about how they might be selected for or against in certain genomic contexts. Frequency of motif sites was assessed using two methods. SeqIndyEn.py was used to evaluate the enrichment of PBCV-1 methyltransferase motifs. Assuming that each nucleotide has an equal probability of occurring, we would expect a four base-pair sequence to occur once in a 256 base-pair window.We calculated an enrichment score based on the number of observed motifs in a window subtracted from the expected number (n-2). Next, we used SeqIndyDep.py to locate genomic regions that are depleted in the motif sites. The expected number of motif sites in these gaps was calculated by counting the length of sequence between two neighbouring motifs, dividing by 256, and multiplying the quotient by 2 to account for both motifs. Altogether, this methodology allowed us to identify fold enriched sites as >1 and fold depleted sites as <1. Recognizing that this approach does not account for amino acid codon dependencies, we also did a parallel enrichment/depletion analysis using the open source, on-line software DistAMo [27]. This scores motif frequency on the basis of codon redundancy, wherein a motif sequence that could be substituted to yield the same codon scores as enriched, and a sequence that can be substituted with the motif of interest to yield the same codon scores as depleted in the original sequence. We also investigated the occurrence of methylation motifs inside genes (motif_CDS.py; DistAMo). All scripts described here are available on GitHub: www.github.com/wilhelmlab/pacbio-scripts/.

Preparation of viral DNA for SMRT sequencing

Three batch cultures of *Chlorella variabilis* (NC64A) were grown in Modified Bold's Basal medium at 25° C under continuous light (30 µmol photons m^{-2} s^{-1}) and gentle shaking (150 rpm) (24). Virus PBCV-1 was added to *C. variabilis* NC64A during mid-logarithmic growth ($\mu = 1.22 \pm 0.06$) at an M.O.I of ~5 (Figure 1). These cultures were incubated at normal culturing conditions described above until the host population had visibly completely lysed, which occurred after three days.

Cell lysate was extracted in triplicate 30 ml samples using a phenol-chloroform method that selectively extracts viral DNA (25). Viral DNA quality and quantity were assessed spectrophotometrically with a NanoDrop 1000 (Thermo Fisher Scientific, Inc., DE, USA). DNA was visualized using agarose gel electrophoresis to confirm the presence of high molecular weight DNA (>40kb), and subjected to a PCR screen for 16S rRNA gene targets to monitor cellular contamination (Figure 2A-B). Extracted DNA was stored at -20°C prior to sequencing.

SMRT sequencing preparation, assembly, and analysis

For each culture, 10 µg of high-quality (absorbance >1.8 for 260/230 and 260/280 spectrophotometric ratios), high-molecular weight DNA was used for SMRT sequencing on the Pacific Biosciences RSII platform. Each sample was loaded into a single-molecule-real-time (SMRT) cell that was prepped with the DNA Template Prep Kit 3.0 and DNA Polymerase Binding Kit P6 v2 (using the P6-C4 chemistry). Viral genomes were initially *de novo* assembled and polished using HGAP and Quiver for comparison to the PBCV-1 reference sequence on NCBI (Accession: NC_000852). Concomitantly, viral reads were aligned to the PBCV-1 reference genome to make epigenomic comparisons across biological replicates. Reference genome recruitment was done using the Pacbio SMRT analysis platform (protocol version = 2.3.0, method=RS Resequencing). Instead of using the Pacbio automated pipeline, which includes DNA modification detection and motif analysis, we manually executed modification and motif detection tools from

command-line to incorporate analysis options not offered in the GUI version. We used ipdSummary.py and motifMaker.sh to identify modifications and motifs, respectively. The command-line analysis produced four files for each sample that were collectively used for downstream analyses: Modifications.gff, Modifications.csv, Motifs.gff, and Motif_Summary.csv. We also repeated this process with different specified read mapping coverages in order to test for coverage effects, as the modification detection software is considered to be skewed towards false positive detection as sensitivity increases (*i.e.* high coverage). However, as we were primarily interested in m6A modification and wanted to decrease the chance of false positive detection, we subsampled our reads with a desired 30X coverage as only 25X is reported sufficient for m6A identification at all genomic positions. All commands used on the command-line are described in the following text.

Commands used for modification and motif detection

Paths to directories specifying files have been simplified for brevity. A full description of the arguments included here can be found from the help manual included in the command-line tool. We admit that some usage might need to be altered as the software has gone through updates since our analyses were completed. The four output files that are generated from this (in their final form) are highlighted in red. This process was reiterated to generate files at coverage (X) of 30, 95, 170, 255, and full coverage. It is also important to note that we maintained use of the default minScore threshold for identifying motifs based on a modificationQV score, though adjustment is recommended to prevent false positive motif identification. Typically this minScore threshold can be chosen based on a break in the modificationQV values of nucleotides as a function of per-strand coverage; modified bases have a higher modificationQV value. However, choosing this threshold is slightly arbitrary in that some nucleotide sites cannot be confidently assigned (Supplemental Figures 7,8). Indeed, at the maximum coverage there is not a clear break to set a minScore threshold (Supplemental Figures 9,10).

```
$ ipdSummary.py /PBCV1-1C_AlignedReads.cmp.h5
--reference /PBCV_1.fasta
--gff /PBCV1-1C_Modifications.gff
--csv /PBCV1-1C_Modifications.csv
--identify m6A,m4C,m5C_TET
--methylFraction
--maxCoverage X

$ motifMaker.sh find
--fasta /PBCV_1.fasta
--gff /PBCV1-1C_Modifications.gff
--minScore 30.0
--output /PBCV1-1C_Motif_Summary.csv

$ motifMaker.sh reprocess
--fasta /PBCV_1.fasta
--gff /PBCV1-1C_Modifications.gff
--motifs /PBCV1-1C_Motif_summary.csv
--output /PBCV1-1C_Motifs.gff
```

DNA methylation stability analysis

A series of R functions were created to enable bioinformatic comparisons of methylation status between biological replicates. Using the function all_motifs.R the four output files generated during command-line modification and motif analysis were combined into one master file. This master file was cross-referenced with those generated from the biological replicates to determine stability of methylated bases. Specifically, we used Pacbio's methylFrac output to do this comparison, which reports how many reads aligning to a particular site are methylated with 95% confidence intervals. This statistic has been validated by another research group who recapitulated the methylFrac output with their own molecule-aggregated, single nucleotide analysis in Python (26). Thus, we

cross-referenced site-specific methylFrac values between biological references to determine site-specific stability. Sites were grouped by methylation status (stably methylated; stably non-methylated; and stochastically methylated) and analyzed for commonalities in gene function and transcriptional timing, where applicable. We also did these analyses in the context of the palindrome, wherein a CATG or GATC site on the forward strand is complimented by the same sequence on the reverse strand. Thus, palindromes were defined and grouped as methylated on both strands, hemimethylated, or lacking methylation on both strands. Attempts to validate PacBio methylation patterns were done using restriction enzyme digests that target GATC tetramers in methylated or non-methylated forms.

Results

Viral methyltransferases in the NCBI database

In order to determine whether methyltransferase enrichment is common in "giant", algal-infecting viruses, we queried the NCBI database for these enzymes (Figure 3, Table 2). This demonstrated that all of the top twenty eukaryotic virus hits were comprised of algal-infecting viruses. Since these are in the 'giant' class, we performed a correlation to determine genome size as a driver of methyltransferase enrichment, and found that this poorly explained eukaryotic virus methylation (R^2= 0.16784). Thus, it is plausible that function(s) associated with viral DNA methylation are instead conserved across similar host-virus systems (*i.e.,* algae-infecting). This justifies further study of DNA methylation in *Chlorella* virus PBCV-1 as a model system for understanding this DNA modification in similar host-virus systems.

Distribution of m6A-targeted nucleotide sequences in the PBCV-1 genome

The methyltransferase target motifs CATG and GATC occur a total of 3,498 times in the PBCV-1 genome. As palindromes, this represents 1,749 distinct genomic locations that can be methylated. A context-independent prediction based on the number of nucleotide combinations that can occur in an oligomer with a defined length (4^n) yields an

expected frequency of once every 256 bp. On average, GATC sites occur once every 388 bps, with an index of dispersion of 477, while CATG sites occur once every 364 bps with an index dispersion of 490. This contrasts with *E. coli* DNA sequences, which encode GATC sites once every 243 bp on average (27, 28). Efforts to identify genomic regions enriched or depleted in these motif sites, again based on a context-independent window size of 256 bp, yields an association with many ORFS (Table 3; Table 4).

A brief glance at the top 10 enriched or depleted regions reveals some trends. First, it appears there are more protein coding genes (coined 'major' ORFs for this system (20)) associated with motif enrichment than non-protein coding genes ('minor' ORFs (20)). While it appears that early transcriptional genes are preferentially enriched in one or both of these motifs, correlational analysis between all ORFs (size normalized), their expression status, and the amount of methylation revealed no relationship (R=0.008 for GATC;0.004 for CATG). On the other hand, there are several regions of the genome that are depleted in both tetramers. Almost all of the top ten most depleted regions are larger than 2kb, including the polycistronic region encoding eleven tRNAs.

Although sequence-independent approaches have been widely used to estimate motif frequency, it is important to consider that motif presence may experience strong selection for protein coding requirements. To assess motif frequency as a function of codon redundancy, we used the open-source, online software DistAMo (29). A broad analysis of all genes annotated with significant Z-scores, the metric for enrichment or depletion, indicated several things. First, although Z-scores are not quantitatively comparable to the frequency scores calculated in the sequence independent approach, many sites were confirmed as enriched or depleted between both approaches (Figure 4A-C). For example, the sequence independent approach identified two gene-associated regions with 4.5x the amount of methylation motifs; these regions were confirmed as enriched in the amino acid redundancy analysis. However, DistAMo revealed that the region impacting gene *A219/222/226R* was enriched in only CATG sites (Figure 4A), whereas the sequence independent approach indicated that both motifs occurred more frequently than expected. On the other hand, the region impacting gene *A656L* was

confirmed to be enriched in only GATC sites as expected (Figure 4B). Second, in contrast to our previous findings, it is the minor, non-coding ORFs that are more commonly enriched in methylation motifs (Figure 4; Table 5). Following this, the few depleted genes identified by DistAMo are all major, protein-coding ORFs. This includes *A351L, A402R, A422R, A486L, A607R, and A625R*. Half of these (*A402R, A486L, A607R*) encode hypothetical proteins with no known homologs, but the remaining genes all encode orthologs for genome integration. *A351L* and *A422R* encode domains used by homing endonucleases for DNA binding, whereas *A625R* is a putative transposase ortholog. Last, since DistAMo can compute motif frequency at any defined window size, we observed enrichment/depletion at scales greater than genes (4kb – 40kb). In some cases, CATG and GATC sequence frequency counteracts one another (Figure 4A-B), while in other regions there is an enrichment or depletion of both tetramers (Figure 4C). The largest depleted region, at the middle of the genome (~165kb), contains the largest non-protein coding region where 11 viral tRNAs are encoded as a polycistron. These are not annotated in the DistAMo graphs since they are not annotated as genes on NCBI, though they are known to lie between annotated genes *A326L* and *a331L*. It is only at the larger window size analysis that this region registers with an overall depletion in either methylation target.

SMRT sequencing of the PBCV-1 genome

Past sequencing efforts indicate that the PBCV-1 genome has a length of 330,611 bp and a linear architecture with terminal inverted repeats that fold together to form hairpin loops (20). *De novo* assemblies generated in this study were 362,016 ± 3,579 bp with extra length discrepancies represented by additional inversions at each termini (Figure 5). A closer inspection of these sequences indicated the extra length represented sequence and/or assembler artifacts, an issue previously encountered during Pacbio sequencing of another chlorovirus (30). This was supported with BLAST homology searches of PBCV-1 ORFs annotated in the reference genome against the *de novo* assemblies, in which genes located in the terminal ~15kb of each end were duplicated.

The corrected viral genome length yielded the expected genome size of ~331kb, thus validating a genome recruitment approach for methylome analyses.

Genome recruitments accumulated 1466 ± 149 read coverage per nucleotide site per strand across the three biological replicates, with decreasing coverage occurring at the terminal ends (Figure 6). PacBio modification detection software requires only 25-fold coverage to detect m6A (31). m5C is reported as detectable at a higher coverage of 250-fold, but is optimally identified following TET1 modification (32)—for this reason PacBio software only supports algorithm detection of m5C that has been TET1 modified. As our viral genomes were not prepared this way, we were unable to identify m5C sites in this study. Due to this limitation, and the fact that high coverage is associated with an increased rate of false positive discovery (33), we decided to randomly sub-sample from the read pool to obtain a maximum recruitment coverage of 30 fold to focus on the known m6A motifs. Using these settings, PacBio software identified a large number of nucleotides associated with modification, as indicated by a high inter pulse duration ratio (ipdRatio), the primary metric for modification detection (Figure 7).

72,170 nucleotide sites are marked with an ipdRatio > 2, which accounts for 10.9% of the PBCV-1 genome (including both forward and reverse strands). This represents ~20.6x more sites than all of the available adenines associated with GATC and CATG tetramers. While some of these events can hypothetically be accounted for by secondary peaks known to form in close proximity to methylated sites, there is typically only one secondary peak associated with each m6A site (34). Indeed, it is this unique signature that assists with PacBio software identification of m6A. Accounting for this background noise, as well as the 1.86% of cytosines purportedly methylated from past studies, this still yields ~62,500 nucleotide sites unaccounted for with an unusually high ipdRatio. Some of these peaks occur in unexpected regions, including the 3,617 bp sequence that does not contain either expected tetramer as determined from *in silico* analysis (associated with genes *A121* and *A122/123R*). A different visualization of this data in the context of motifs detected (Table 6) demonstrates that many cytosines, guanines, and thymines have high ipdRatio values, but only adenines separate as clear populations of

motif associated modification, and non-modified bases (Figure 8). Longer degenerate cytosine and guanine decorated motifs are detected by PacBio in each virus replicate, but at low frequency. Moreover, these are replaced with other low-frequency degenerate sequences at higher coverage analyses (Table 7). Because these long degenerate strings are likely false positives (as PacBio analyses are skewed towards this (35)) our analyses henceforward focus on the methylation status of the 3,498 adenines targeted by PBCV-1 methyltransferases in CATG and GATC contexts.

The motifs identified in the motifMaker.sh program are identified based on their association with a Phred quality score, known as the ModificationQV, that exceeds a defined threshold. This threshold is given a default value of 30, which corresponds to a p-value of 0.001. It is recommended that users observe the distribution of data for this metric in order to adjust the threshold value appropriately—especially in high coverage datasets. In observing ModificationQV scores, we found that only adenine modification status could be better resolved in comparison to the ipdRatio graphs (Figure 9), though some target adenosines cluster with the non-modified adenines in either case (Figure 8; Figure 10). Using the default QV threshold, which was appropriate for our study, Pacbio software detected 96.4 ± 0.3% and 84.1 ± 0.2% of CATG and GATC sites, respectively, as methylated (Table 7). This accounts for 1.59 ± 0.004 % of the total adenine pool, which is close to historic measurements of 1.45% (19) and represents closer to ~3200 as opposed to ~2900 methylated targets. It also means that only about 300 target sites are not methylated, though it is not clear if these sites are consistent between each replicate or whether they are random occurrences. To determine this, we analyzed site-specific methylation status across biological replicates to infer modification stability.

Using annotations directly made by PacBio software, we parsed sites identified with m6A modification. Stability of these sites was determined by comparison of a methylFrac value, which represents the percent of reads aligning to that site which are identified with m6A. Mean and standard deviation calculation of this value between the three replicates enables a direct binning strategy for one of three modification characteristics: stably non-methylated sites (average methylFrac ~ 0), variably

methylated sites (methylFrac ~ 1.0 in one or two of the three sequenced replicates, resulting in an average value of ~0.33 or ~0.66), and stably methylated sites (average methylFrac ~1.0). By comparing methylFrac values averaged for each GATC and CATG site (treating targets on each strand as independent), it becomes clear that each characteristic type is represented (Figure 11). An estimate on the size of each population yields 2,825 tetramers (80.7%) that are almost always methylated, 457 tetramers (13.1%) that are methylated in two replicates, 143 tetramers (4.1%) that are methylated in only one replicate, and 73 tetramers (2.1%) that are almost never methylated. Thus, over 17% of target sites are variably methylated while the remaining 83% maintain a stable modification status.

After determining methylation stability of each GATC and CATG site, we analyzed these within their palindromic context. Palindromic reverse complimentary sequences targeted by DNA methyltransferases can exist in one of three states: methylation occurring on both strands, methylation missing on both strands, or methylation occurring on only one strand (*i.e.* hemi-methylation). By mapping methylFrac status of adenines on both the forward and reverse strand of each palindrome, it is clear that all three types of palindromes are represented in the PBCV-1 genome, though GATC sites are more often associated with non-methylation (Figure 12). Among all 1,749 palindromes, 1,083 sites (61.9%) demonstrated methylation on average in >75% of reads for both strands on a palindrome. Alternatively, 542 sites (31%) were defined as hemi-methylated with only one strand containing methylation in >75% of reads. The remaining 124 sites (7.1%) were defined as more stochastic with methylation occurring on <75% of reads on either strand.

Only three palindromes were annotated as stably not methylated on both strands. This includes two loci located within neighboring genes encoding minor capsid proteins (CATG - *A383R*; GATC - *A384dL*) and one within a hypothetical protein (GATC - *A432R*). The genes impacted by these sites are all expressed late in the PBCV-1 infection cycle, their protein products are packaged/part of the PBCV-1 virion (20, 21), and there are no known homologs in the NCBI database outside of close viral relatives. All sites had a ModificationQV value <30, and all but one had an ipdRatio value <2 (1.64 ± 0.78).

Confirmation of modification vacancies at these sites was attempted with digestion of the PBCV-1 genome using commercial restriction endonucleases DpnI, DpnII, and Sau3AI, which cut methylated GATC sites, non-methylated GATC sites, and GATC sites independent of methylation status, respectively (Figure 13). Although DpnII treatment yields a smear distinct from control DNA not treated with an enzyme, an expected ~24,000 bp band representing the DNA between the two GATC non-methylated palindromes was absent. This indicates that some factor is inhibiting enzymatic digestion at the three locations marked as non-methylated.

The next example of stable non-methylation occurs in hemi-methylated palindromes. Although we defined this by a threshold of <75% of reads on one strand, and >75% of reads on the other, high-confidence hemi-methylated sites can be distinguished as one strand lacking methylation in all reads (methylFrac=0). These account for 32 GATC palindromes, and 12 CATG palindromes (Figure 12). While ipdRatio values are on average under 2 for the non-methylated strand in CATG palindromes (1.69 ± 0.44), these values far exceed this threshold in the case non-methylated strands in GATC palindromes (3.75 ± 1.34). And while a handful of these GATC sites are annotated as modified in some way other than methylation, most bear no modification annotation. There is no obvious clustering of either type of stable hemi-methylated palindrome (GATC sites occur on average every 6,830 bp with an index dispersion of 9,574 bp; CATG sites are dispersed on average every 23,629 bp with an index dispersion of 39,102 bp). All that said, there is no apparent pattern perceived here related to how certain palindrome types are dispersed according to gene ontology, gene location (N terminus to C terminus), or transcriptional status of the gene.

Although the overall analysis of palindromes indicates that each type of methylation is present, functional implications may be better understood in the context of specific genomic units. There are several ways to do this, ranging from regulatory regions to whole genes. We chose to map palindrome methylation for enriched regions identified in sequence independent *in silico* analyses (Figure 14A) and capsid proteins encoded by PBCV-1 (Figure 14B). In most cases methylation is stable, but there are a few completely

non-methylated palindromes (*A383R*) or stable hemi-methylated palindromes (*A430L, A622L, A011L*). In general, though, there is typically a mix in the types of patterns observed per gene. These findings present opportunities to explore the consequence of this patterning on viral activity and fitness.

Discussion

Many tools have been developed over the last few decades to allow molecular insight into host-virus interactions (8). Most notable are improvements in the 'omics' techniques, enabling increasingly higher resolution studies ranging from single organisms to whole communities. Indeed, *Chlorella* virus PBCV-1 has been subjected to genomics, transcriptomics, and proteomics, with the findings of this study establishing some of the first epigenomic observations in this system.

What can be concluded about CATG and GATC motif distribution in the PBCV-1 genome is that it is not random. Moreover, using two approaches to assess motif frequency, while not exhaustive, helps delineate the selective forces acting on motif patterning. Analyses based on a sequence independent approach found that the top ten enriched regions mostly associate with major, protein-coding genes. Since 92.8% of the PBCV-1 genome is occupied by protein coding genes, this trend is somewhat expected. It is also congruent with methylation motif enrichment in prokaryotes whose genomes share a similar coding density of ~90% (36). On the other hand, when we accounted for local codon redundancy most protein coding genes lose their enrichment status to be dominated by minor, non-protein coding genes. This seems more logical considering minor genes are non-protein coding and would thus not be under this selective pressure . However, this is complicated by the fact that most minor genes overlap with protein coding genes, albeit as smaller units. It is also important to note that certain parts of the protein might be more flexible to amino acid substitutions than others, including regions not significantly impacting protein structure or domain function. Since DistAMo assumes amino acid sequence is not flexible, it is difficult to conclude whether selection is greater on the protein sequence as opposed to any other factor (*i.e.,* RNA secondary structure).

In any case, genes/regions marked as enriched or depleted across both analyses provide higher confidence candidates that are truly enriched or depleted. For example, genes *A219/222/226R* and *A656L* might be good candidates for investigation into flexibility of the codon/amino acid sequence, as well as how variable methylation at these sites impacts the virus. In our PacBio studies, the motif-enriched region impacting gene *A219/222/226* maintains nearly complete methylation in all but a few sites. However, determining the effects of this is dependent on making PBCV-1 genetically tractable, which has not yet been accomplished.

Another useful application of the *in silico* analysis is that it identified many regions depleted in motifs targeted for methylation. The few genes marked as depleted by DistAMo represent all major, protein-coding ORFs, with half bearing functions associated with genome integration. *A351L* and *A422R* encode domains used by homing endonucleases for DNA binding, whereas *A625R* is a transposase ortholog. That these proteins are all involved in genome integration processes might represent a negative selective pressure to side-effects of palindromes and/or methylation that can occur on these sites. In rice, hypermethylation occurs in transposable elements following whole genome duplication, a marker for angiosperm evolution (37). Concomitantly, this modification inhibits their transposition to stabilize the integrity of the chromosome and decrease nearby gene expression. Some of the larger chloroviruses encode many putative transposases (some with internal resolvases) and homing endonucleases (11), which are thought to be involved in genomic rearrangements and gene duplications (17). These larger viruses are more heavily methylated (17, 18), which might also function to stabilize chlorovirus genomes at a certain size threshold. It would be interesting to see if hypermethylation occurs more frequently in transposases of these larger viruses. Another striking case of motif depletion in PBCV-1, which was identified in both *in silico* analysis approaches, was that the middle of the genome, where the viral tRNA polycistron is encoded, comprises one of the largest regions lacking either target motif. Motif depletion in tRNA, and even rRNA genes, has been observed in bacterial genomes, too; these genes exhibit the lowest frequency of GATC motifs in *E. coli* (28). Hypotheses

related to this depletion in bacterial genes have suggested that selection against the palindrome occurs in these regions due to its increased ability to form secondary structures that might interfere with constitutive expression of these genes (28). Though not related to methylation, this is a factor worth considering. Another similarity between PBCV-1 and bacteria is how motifs are dispersed. In *E. coli,* GATC motifs are never separated by more than 2kb. This has been hypothesized to promote mismatch repair efficiency*,* presumably because this function is less efficient when the GATC methyl-director is separated by greater distances (28). Since PBCV-1 spacing is similar to *E. coli* (Table 2)*,* and the *Chlorella variabilis* host genome encodes a MutS homolog (XP_005846525), there might be a methyl-directed function associated with these two components. Although PBCV-1 does not encode its own MutS protein, several giant virus relatives encode their own MutS homologs, suggesting it is important for virus fitness (38). Finally, another observation from our sequence independent approach showed that CATG and GATC motifs counteract one another as enriched or depleted in certain regions. Though it is not clear if this is meaningful for PBCV-1, it is interesting that GATC enrichment and depletion in *Escherichia coli* marks the genome replication of origin and termination, respectively (29). Thus, methylation might have similar implications for PBCV-1 replication.

Results from the Pacbio software suggests PBCV-1 exhibits high, but not complete, methylation of CATG and GATC tetramers. Total adenine methylation accounts for 1.59 ± 0.004 % of all adenines, which is close to historical HPLC measurements of 1.45% (18, 19). We identified three putative, completely non-methylated palindromes, but failed to confirm these with restriction digestion. Providing these are indeed true negatives, there are some intriguing biological consequences at stake. Both of the adenine decorating methyltransferases in PBCV-1 are associated with RM systems. These RM systems have been shown to digest and recycle the host genome for viral DNA replication, while protecting the viral genome against self-digestion (39). Thus, there are potential deleterious consequences for PBCV-1 if a target site is not methylated. Indeed, this selective pressure is apparently strong enough to dictate complete methylation of RM

targeted motifs in >100 bacterial chromosomes analyzed with Pacbio sequencing (40). Complete methylation would seem especially necessary for PBCV-1 since the viral restriction endonucleases are packaged in the virus particle (39). Failure to digest the viral genome during *in vitro* restriction digestion indicates that the three palindromes identified by Pacbio as fully non-methylated are able to resist restriction by one of several means. First, it is possible that these sites are false negatives for m6A. This seems unlikely, given kinetic fingerprints of these sites do not generally exhibit a high ipdRatio value on either strand that is indicative of m6A. Second, it is possible that some other type of modification is present on or at least in the vicinity of these palindromes, which allows evasion of endonuclease recognition. If this were true, one would expect that this would interfere with polymerase kinetics to yield a high ipdRatio. That said, it is clear that some modifications do not elicit strong effects and are consequently poorly detected, if at all. Native 5mC does not protrude into the major groove of DNA like m6A does, and thus elicits a subtle impact on polymerase kinetics. Indeed, this is why TET1 modification is used to improve the signal of 5mC (32). Thus, there is no evidence to refute alternative modifications, though there is also not enough evidence to conclude that other modifications are present. In fact, it is possible that these are false negatives deriving from poor *in silico* predictions of ipdRatio values for the non-modified control sequence. However, the original description of the *in silico* control demonstrated that it is effective at identifying all true positives at coverages similar to what we used here (31, 33). Thus, future investigation is needed to confirm the modification status of these three palindromes.

In considering ipdRatio values as the purported primary metric for modification (34, 40), it is worthwhile to reiterate that several high ipdRatio events occur in the PBCV-1 genome that are not associated with motifs known to be targeted for methylation. To check if this is common in other systems, we analyzed ipdRatio distributions in Pacbio data for *Escherichia coli* K-12 (MG1655). This bacterium encodes three active methyltransferases that recognize GATC, CCWGG, and AACNNNNNNGTGC/GCACNNNNNNGTT contexts. Pacbio data indicated that >99% of

these sites were marked as methylated, which accounts for roughly 63,522 target sites. Total number of sites exceeding an ipdRatio >2 yielded 154,230 sites, accounting for 1.7% of the bacterial genome (both forward and reverse strands). This represents an enrichment of only ~2.4x in comparison to the ~20.5x enrichment observed in PBCV-1. One might propose that the ~65,000 nucleotides indicated as modified (not with methylation or a repeated, detectable motif) is an artifact of Pacbio's high error rate, which is ~11-15% on average (41). However, since the mapped reads were randomly sampled to 30-fold coverage, and the errors are random, this error rate reduces to <1%, which is far below the number of loci with high ipdRatio values.

Barring a poor *in silico* prediction of non-modified polymerase kinetics, it is surmisable that some other modification is responsible for the high number of peaks observed in PBCV-1. Indeed, this has been hypothesized to explain the thousands of kinetic variation events that were observed in *E. coli* (33). These would be random and not associated with adenines according to the distribution of ipdRatio and ModificationQV scores. One possible explanation is that oxidative stress induced lesions are occurring; these randomly occur in at least 20 different DNA base configurations (42), and have been shown to elicit kinetic effects across multiple neighboring sites (43). However, an algorithm to test this does not exist in the Pacbio software yet. In any case, we favor this idea as oxidative stress is a considerable challenge for virus replication, which is presumably why so many giant viruses, including PBCV-1 (44), encode machinery to mitigate this stress (10). Stress induced DNA modifications might become rampant in lytic virus progeny, thus representing a potential cause of some progeny being non-infectious. Indeed, chlorovirus PBCV-1 has a burst size of ~1000 progeny, yet, only ~30% of this population is capable of forming plaques (18). It is unlikely that cellular organisms, for which most modification analyses have been based on, would have as many oxidative stress induced DNA modifications as viruses because these organisms can repair these lesions as long as they are alive. Viruses, on the other hand, might encounter an environment less conducive to DNA repair followed by genome packaging in a metabolically inactive virion. Moreover, our extraction protocol does not selectively

acquire nucleic acid from only infectious virus. Still, it is also possible that some type of modification not yet characterized, which has nothing to do with oxidative stress, could be responsible. Recent computational studies have identified a variety of DNA modification systems (45), and other Pacbio work has confirmed the presence of novel modifications including phosphorothionation (46). Giant viruses, whose proteins are dominated by those with unknown functions might also be capable of this. We also maintain this possibility in light of the fact that past HPLC-MS measurements of PBCV-1 nucleosides determined percent of methylated nucleotides based on relative pools as opposed to absolute quantitative measurements (19). This approach would mean that peaks associated with uniquely modified nucleotides, perhaps with drastically different retention times, would not be considered in nucleotide pool estimations.

A more common observation from this study is that hemimethylation can be common (~31% of palindromes), though only a few of these sites are stable (2.51%). The status of these hemimethylated GATC sites is questionable given their 'non-methylated' strand often exhibits a large ipdRatio value. Despite that, it is not biologically impossible that hemi-methylation is in the PBCV-1 genome; this is known to occur in many cellular organisms including human cell lines (47, 48) and bacteria (49). Moreover, the endonuclease targeting CATG sequences, M.CviAI, has been shown to be sensitive to hemi-methylation (50). It is thus reasonable that the GATC targeting endonuclease is also unable to cleave hemimethylated targets, or at the very least at a much slower rate. Additionally the viral replication time is much longer (a few hours) than bacteria with stable methylation patterns (less than 20 minutes) (49). The PBCV-1 genome is also much smaller with fewer target sites than bacteria, making it even more possible that all targets can be completely methylated.. Thus, some factor might block complete methylation, such as a protein that competes for this binding site. Altogether, these observations suggest hemimethylation is biologically permissible, and invites future investigation into whether this is a unique characteristic of these types of enzymes that relates to specific viral activity.

Finally, there are a few caveats to point out about our approach. First, we analyzed stability using the methylFrac value computed by Pacbio software. Though this has been validated with external analyses (26), an inherent weakness is that the methylFrac analysis does not consider molecule specific effects. Namely, the 30 reads that align to a given site can derive from any number of molecules, each of which should in reality be considered as sub-populations or perhaps 'quasi-species'. Other tools exist to analyze this data for population-level epigenomic variants (26), but our data could not be assessed this way as we used the *in silico* control data as opposed to a whole genome amplified control. Second, we sub-sampled our reads to a coverage of 30-fold, which is reportedly appropriate for methylation detection at all genomic positions according to Pacbio. That said, the 95% confidence variables for the methylFrac value were at times quite large, which might be impacted by reads deriving from molecules with diverse methylation profiles. This effect cannot be accounted for until the smalr package created by Beaulaurier *et al.* (26) is updated to analyze read effects with an *in silico* control.

DNA methylation has been identified or inferred in many types of giant viruses using restriction mapping (18), cloning (19), and genomics (11, 51-54). The enzymes responsible for these modifications are at times paired with a cognate restriction endonuclease, thus forming a viral restriction modification system, as is the case for PBCV-1 (51, 53). In chloroviruses, however, the majority (~75%) of methyltransferases are not paired with a restriction endonuclease and are instead annotated as 'orphan' methyltransferases (55). Orphans have been identified as regulators of cellular and viral activities (15, 56), though a function has not yet been described in giant viruses. These enzymes are not likely used for chlorovirus restriction evasion. This is evident because there is no need to protect DNA from eukaryotic hosts, which encode no native restriction endonucleases, or other RM-carrying chloroviruses as these entities prevent co-infection via membrane depolarization (57). This observation, combined with the fact that chloroviruses and their giant virus relatives encode some of the highest numbers of methyltransferases among viruses, indicates a novel and seemingly biologically important use of methyltransferases for viral fitness. This is supported by the observation that

phylogeny of some of these enzymes reflects a long evolutionary history within viruses, instead of a recent acquisition from cellular organisms by horizontal gene transfer. Complementing this concept, is that other enzymes encoded by chloroviruses have seemed to adapt to side-effects of genomic methylation. For example, topoisomerase II from PBCV-1 processes DNA at a rate of 30-50x faster than human topoisomerases (58). This rate is believed to stem from an adaptation to higher incidences of DNA methylation in chloroviruses (59), as methylation is known to slow topoisomerase processing much like it does with DNA polymerase (59). It would be interesting to see if different variations of palindrome methylation impact topoisomerase activity. It is also possible that these markers assist with genome condensation for viral DNA packaging (60). In another example, hemimethylated palindromes have been shown to control promotor activation, in some cases allowing gene expression only transiently following DNA replication (16). Though we did not see an obvious correlation between transcriptional profile and methylation here, it is possible that the methylation profile during an active viral infection might better explain transcriptional orchestration of PBCV-1. In any case, information provided here establishes a useful framework for investigating DNA methylation in chlorovirus PBCV-1, as well as initiating these studies in other systems with more ‘orphan’ methyltransferases.

References

1. Anesio AM, Bellas CM. Are low temperature habitats hot spots of microbial evolution driven by viruses? Trends in Microbiology. 2011;19(2):52-7.

2. Weinbauer MG, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? Environmental Microbiology. 2004;6(1):1-11.

3. Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC. Drivers of bacterial beta-diversity depend on spatial scale. Proc Natl Acad Sci USA. 2011;108(19):7850-4.

4. Thingstad TF, Lignell R. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat Microb Ecol. 1997;13(1):19-27.

5. Weitz JS, Stock CA, Wilhelm SW, Bourouiba L, Coleman ML, Buchan A, et al. A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. ISME J. 2015;9(6):1352-64.

6. Wilhelm SW, Suttle CA. Viruses and nutrient cycles in the sea - viruses play critical roles in the structure and function of aquatic food webs. BioScience. 1999;49(10):781-8.

7. Short SM, Staniewski MA, Chaban YV, Long AM, Wang D. Diversity of viruses infecting eukaryotic algae. In: P. H, Abedon ST, editors. Viruses of microorganisms. Poole, UK: Caister Academic Press; 2018. p. 211-44.

8. Coy SR, Gann ER, Pound HL, Short SM, Wilhelm SW. Viruses of eukaryotic algae: diversity, methods for detection, and future directions. Viruses. 2018;10(9).

9. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, Bideshi DK, et al. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol. 2013;158(12):2517-21.

10. Wilhelm SW, Bird JT, Bonifer KS, Calfee BC, Chen T, Coy SR, et al. A student's guide to giant viruses infecting small eukaryotes: from *Acanthamoeba* to *Zooxanthellae*. Viruses. 2017;9(3).

11. Fitzgerald LA, Graves MV, Li X, Feldblyum T, Nierman WC, Van Etten JL. Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect Chlorella NC64A. Virology. 2007;358(2):472-84.

12. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Research. 2015;43(D1):D298-D9.

13. Nathan D, Crothers DM. Bending and flexibility of methylated and unmethylated EcoRI DNA. Journal of Molecular Biology. 2002;316(1):7-17.

14. Wion D, Casadesus J. N-6-methyl-adenine: an epigenetic signal for DNA-protein interactions. Nature Reviews Microbiology. 2006;4(3):183-92.

15. Murphy J, Mahony J, Ainsworth S, Nauta A, van Sinderen D. Bacteriophage Orphan DNA Methyltransferases: Insights from Their Bacterial Origin, Function, and Occurrence. Appl Environ Microbiol. 2013;79(24):7547-55.

16. Casadesus J, Low D. Epigenetic gene regulation in the bacterial world. Microbiology and Molecular Biology Reviews. 2006;70(3):830-+.

17. Jeanniard A, Dunigan DD, Gurnon JR, Agarkova IV, Kang M, Vitek J, et al. Towards defining the chloroviruses: a genomic journey through a genus of large DNA viruses. BMC Genomics. 2013;14.

18. Van Etten JL, Lane LC, Meints RH. Viruses and virus-like particles of eukaroytic algae. Microbiological Reviews. 1991;55(4):586-620.

19. Van Etten JL, Schuster AM, Girton L, Burbank DE, Swinton D, Hattman S. DNA methylation of viruses infecting a eukaryotic Chlorella-like green alga. Nucleic Acids Research. 1985;13(10):3471-8.

20. Dunigan DD, Cerny RL, Bauman AT, Roach JC, Lane LC, Agarkova IV, et al. Paramecium bursaria chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. J Virol. 2012;86(16):8821-34.

21. Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova IV, et al. Microarray analysis of paramecium bursaria chlorella birus 1 transcription. J Virol. 2010;84(1):532-42.

22. Blanc G, Mozar M, Agarkova IV, Gurnon JR, Yanai-Balser G, Rowe JM, et al. Deep RNA sequencing reveals hidden features and dynamics of early gene transcription in paramecium bursaria chlorella virus 1. PloS One. 2014;9(3):10.

23. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Research. 2019;47(D1):D427-D32.

24. Dunigan D, Agarkova I. Formulation of MBBM (Modified Bold's Basal Medium) 2016 [

25. Dunigan D, Agarkova I. Viral DNA Miniprep Procedure. 2016.

26. Beaulaurier J, Zhang X-S, Zhu S, Sebra R, Rosenbluh C, Deikus G, et al. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nature Communications. 2015;6:7438.

27. Barras F, Marinus MG. Arrangement of dam methylation sites (gatc) in the Escherichia-coli chromosome. Nucleic Acids Research. 1988;16(20):9821-38.

28. Marinus MG, Lobner-Olesen A. DNA Methylation. EcoSal Plus. 2014;6(1).

29. Sobetzko P, Jelonek L, Strickert M, Han WX, Goesmann A, Waldminghaus T. DistAMo: A Web-Based Tool to Characterize DNA-Motif Distribution on Bacterial Chromosomes. Frontiers in Microbiology. 2016;7.

30. Quispe CF, Esmael A, Sonderman O, McQuinn M, Agarkova I, Battan M, et al. Characterization of a new chlorovirus type with permissive and non-permissive features on phylogenetically related algal strains. Virology. 2017;500:103-13.

31. Rakkhumkaew N, Shibatani S, Kawasaki T, Fujie M, Yamada T. Hyaluronan synthesis in cultured tobacco cells (BY-2) expressing a chlorovirus enzyme: Cytological studies. Biotechnology and Bioengineering. 2013;110(4):1174-9.

32. Clark TA, Lu XY, Luong K, Dai Q, Boitano M, Turner SW, et al. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. Bmc Biology. 2013;11.

33. Feng ZX, Fang G, Korlach J, Clark T, Luong K, Zhang XG, et al. Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic. PLoS Comput Biol. 2013;9(3):10.

34. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods. 2010;7(6):461-U72.

35. Biosciences P. Base Modification: From Sequencing Data to a High Confidence Motif List GitHub2014 [Available from: https://github.com/PacificBiosciences/Bioinformatics-Training.wiki.git.

36. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics. 2015;15(2):141-61.

37. Zhang J, Liu Y, Xia EH, Yao QY, Liu XD, Gao LZ. Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(50):E7022-E9.

38. Ogata H, Ray J, Toyoda K, Sandaa RA, Nagasaki K, Bratbak G, et al. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. ISME J. 2011;5(7):1143-51.

39. Agarkova IV, Dunigan DD, Van Etten JL. Virion-associated restriction endonucleases of chloroviruses. J Virol. 2006;80(16):8114-23.

40. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, et al. The Epigenomic Landscape of Prokaryotes. Plos Genetics. 2016;12(2).

41. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics & Bioinformatics. 2015;13(5):278-89.

42. Cooke MS, Evans MD, Dizdaroglu M, Lunec J. Oxidative DNA damage: mechanisms, mutation, and disease. Faseb Journal. 2003;17(10):1195-214.

43. Clark TA, Spittle KE, Turner SW, Korlach J. Direct detection and sequencing of damaged DNA bases. Genome integrity. 2011;2:10-.

44. Kang M, Duncan GA, Kuszynski C, Oyler G, Zheng JY, Becker DF, et al. Chlorovirus PBCV-1 Encodes an Active Copper-Zinc Superoxide Dismutase. Journal of Virology. 2014;88(21):12541-50.

45. Iyer LM, Zhang DP, Burroughs AM, Aravind L. Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. Nucleic Acids Research. 2013;41(16):7635-55.

46. Ahlgren NA, Chen Y, Needham DM, Parada AE, Sachdeva R, Trinh V, et al. Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral

infection, phosphorothioation DNA modification and multiple restriction systems. Environmental Microbiology. 2017;19(6):2434-52.

47. Xu CH, Corces VG. Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. Science. 2018;359(6380):1166-9.

48. Ehrlich M, Lacey M. DNA Hypomethylation and Hemimethylation in Cancer. In: Karpf AR, editor. Epigenetic Alterations in Oncogenesis. Advances in Experimental Medicine and Biology. 7542013. p. 31-56.

49. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nature Biotechnology. 2012;30(12):1232-+.

50. Luo GZ, Wang F, Weng XC, Chen K, Hao ZY, Yu M, et al. Characterization of eukaryotic DNA N-6-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. Nature Communications. 2016;7.

51. Stough JMA, Yutin N, Chaban YV, Moniruzzaman M, Gann ER, Pound HL, et al. Genome and Environmental Activity of a Chrysochromulina parva Virus and Its Virophages. Frontiers in Microbiology. 2019;10.

52. Schroeder DC, Park Y, Yoon HM, Lee YS, Kang W, Meints RH, et al. Genomic analysis of the smallest giant virus - Feldmannia sp virus 158. Virology. 2009;384(1):223-32.

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

54. Schvarcz CR, Steward GF. A giant virus infecting green algae encodes key fermentation genes. Virology. 2018;518:423-33.

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

56. Sternberg N, Coulby J. Cleavage of the Bacteriophage-P1 packaging site is regulated by adenine methylation. Proc Natl Acad Sci USA. 1990;87(20):8070-4.

57. Greiner T, Frohns F, Kang M, Van Etten JL, Kasmann A, Moroni A, et al. Chlorella viruses prevent multiple infections by depolarizing the host membrane. J Gen Virol. 2009;90:2033-9.

58. Fortune JM, Lavrukhin OV, Gurnon JR, Van Etten JL, Lloyd RS, Osheroff N. Topoisomerase II from Chlorella virus PBCV-1 has an exceptionally high DNA cleavage activity. Journal of Biological Chemistry. 2001;276(26):24401-8.

59. Dickey JS, Van Etten JL, Osheroff N. DNA methylation impacts the cleavage activity of chlorella virus topoisomerase II. Biochemistry. 2005;44(46):15378-86.

60. Wulfmeyer T, Polzer C, Hiepler G, Hamacher K, Shoeman R, Dunigan DD, et al. Structural organization of DNA in *Chlorella* viruses. PLoS One. 2012;7(2).

61. Xia YN, Vanetten JL. DNA methyltransferase induced by pbcv-1 infection of a chlorella-like green-alga. Molecular and Cellular Biology. 1986;6(5):1440-5.

62. Xia YN, Burbank DE, Uher L, Rabussay D, Vanetten JL. Restriction endonuclease activity induced by pbcv-1 virus infection of a chlorella-like green alga. Molecular and Cellular Biology. 1986;6(5):1430-9.

63. Zhang YP, Nelson M, Nietfeldt JW, Burbank DE, Vanetten JL. Characterization of chlorella virus pbcv-1 CviAII restriction and modification system. Nucleic Acids Research. 1992;20(20):5351-6.

64. Zhang Y, Nelson M, Vanetten JL. A single amino-acid change restores DNA cytosine methyltransferase activity in a cloned chlorella virus pseudogene. Nucleic Acids Research. 1992;20(7):1637-42.

Appendix

Table 3.1 Characteristics of methyltransferases encoded by PBCV-1

Name	Gene	Mod	Motif	Txc[a]	Notes
M.CviAI	*A581R*	m6A	GATC	E	Part of viral RM system (61, 62)
M.CviAII	*A251R*	m6A	CATG	E	Part of viral RM system (63)
M.CviAIV	*A530R*	m5C	RGCB	L	Pseudogene, non-functional (64)
M.CviAIIIP	*A517L*	m5C	-	E	Putative G + C rich sequence (64)
M.CviAV	*A683L*	m5C	-	EL	Putative non-functional gene

[a]Transcriptional status reprinted from supplemental tables in Dunigan *et al.* 2012 (20). E=Early; L=Late; EL= Early-Late. Mod=Modification Type

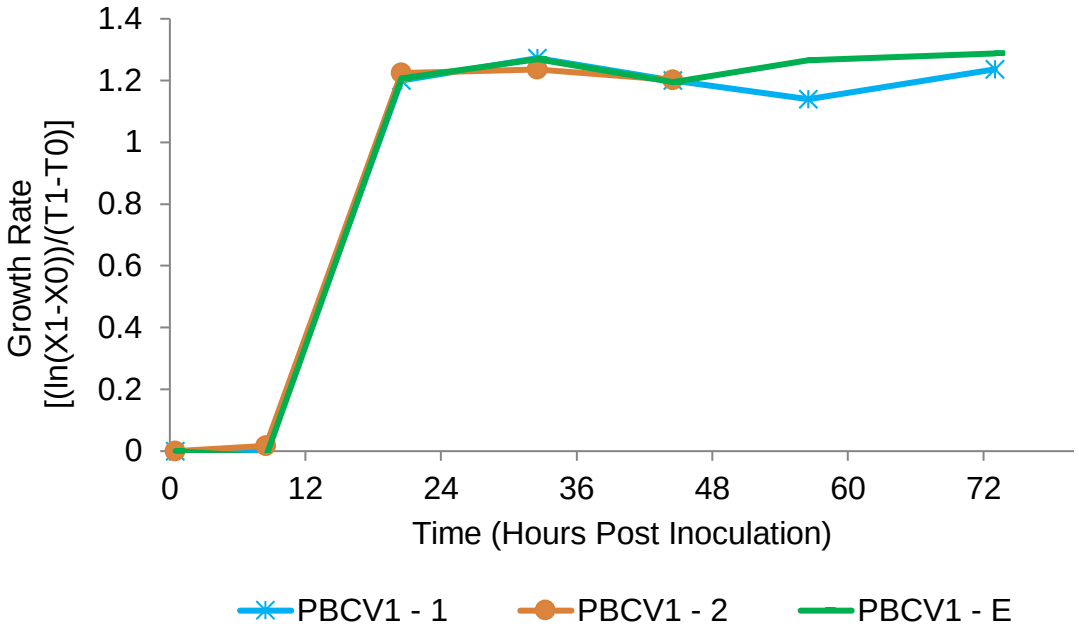


Figure 3.1 Growth dynamics of C. variabilis cultures prior to infection at 72 h

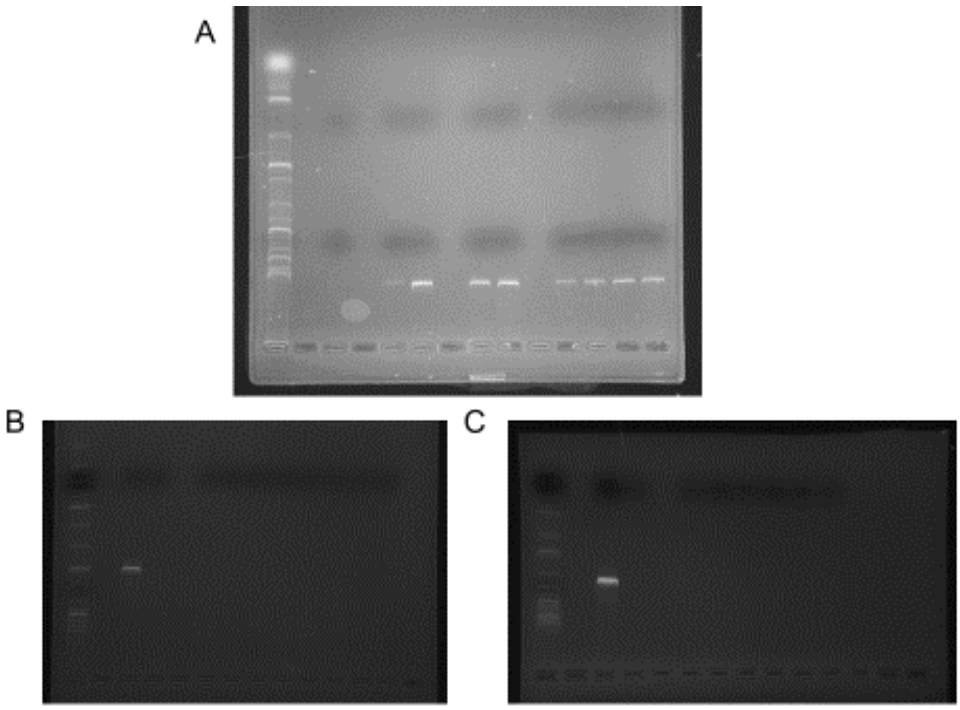


Figure 3.2 Agarose gel electrophoresis visualization of PBCV-1 genomic DNA

From left to right, 1) 1kb extension ladder (NEB); 2) NA; 3) Loading Buffer Control; 4) NA; 5/6) PBCV-1 Rep1(1C); 7) NA; 8/9) PBCV1-Rep2(2A); 10) NA; 11-14) PBCV-1 Rep3 (E1). 70-80ng of DNA were loaded into each well and stained with Midori Direct for UV visualization. Agarose gel electrophoresis results of 16SrDNA amplification using the universal 27F and 1522R primers. From left to right, wells represent 1) 1kb Plus DNA Ladder; 2) NA; 3) E. coli DNA 4) Non-Template Control; 5) NA; Lanes 6 thru 10) PBCV1-Rep1; Lanes 11 thru 13) PBCV1-Rep2; Second Gel, from left to right 1) 1kb Plus DNA Ladder; 2) NA; 3) E. coli DNA; 4) Non-Template Control; 5) NA; Lanes 6 thru 7) PBCV-1 Rep2 DNA; Lanes 8 thru 11) PBCV-1 Rep3 DNA.

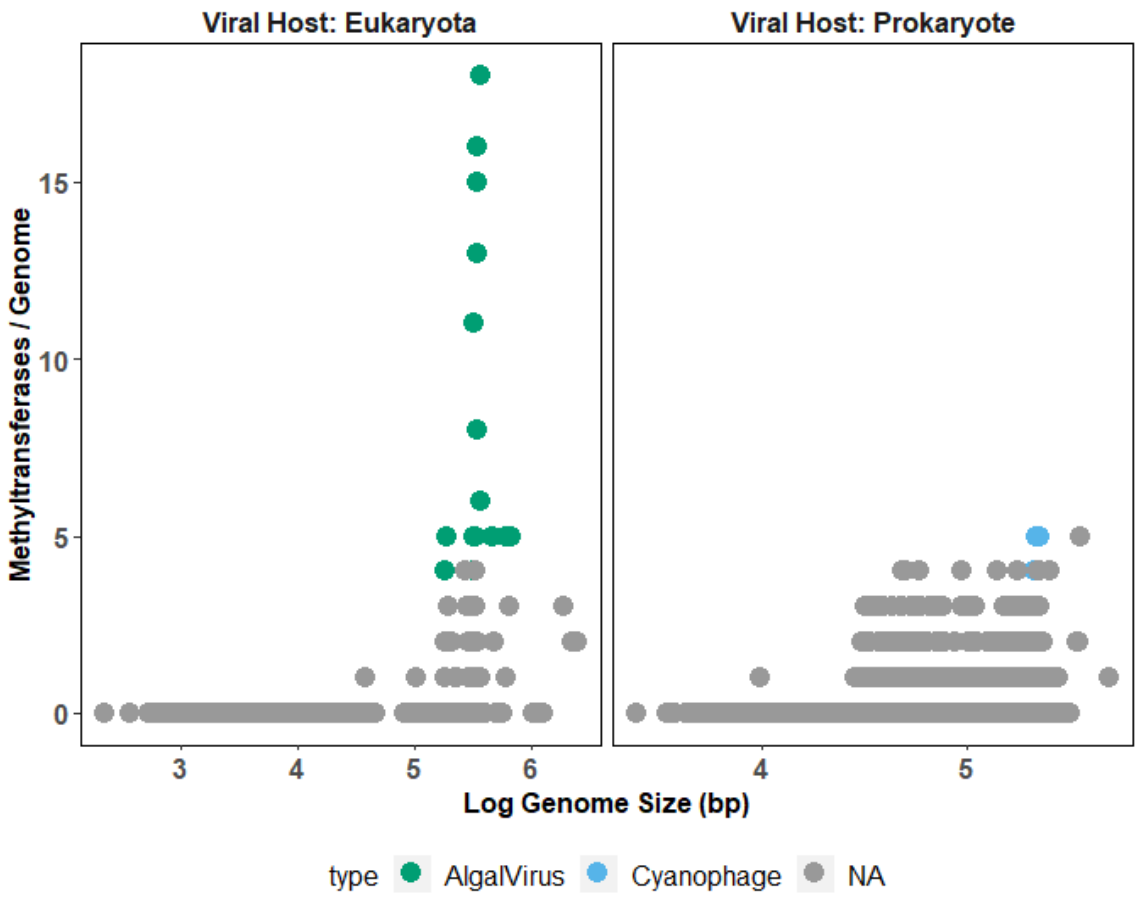


Figure 3.3 Summary of methyltransferases encoded in public viral genomes

The top twenty viral genomes encoding the most methyltransferases were color-coded depending on whether they associate with a photosynthetic host. Some of these are not visible because they overlap with other viruses, though a list of these are provided in Table S1. All other hits, independent of methyltransferase status, were color coded grey and are defined as non-applicable (NA). n=10,708 sequences.

Table 3.2 Top twenty viruses with the most methyltransferase genes

Host	Accession	Virus Type*	Viral Species/Strain	MTase
Prokaryote	GCA_002593925	Cyanophage	*Synechococcus phage ACG-2014f*	5
Prokaryote	GCA_002593945	Cyanophage	*Synechococcus phage ACG-2014f*	5
Prokaryote	GCA_002594045	Cyanophage	*Synechococcus phage ACG-2014f*	5
Prokaryote	GCA_002594185	Cyanophage	*Synechococcus phage ACG-2014f*	5
Prokaryote	GCA_002594565	Cyanophage	*Synechococcus phage ACG-2014f*	5
Prokaryote	GCF_000898015	Bacteriophage	*Cronobacter phage vB_CsaM_GAP32*	5
Prokaryote	GCA_002593785	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002593805	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002593845	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002593885	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002593985	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594025	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594065	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594085	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594105	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594165	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594385	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594405	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594485	Cyanophage	*Synechococcus phage ACG-2014f*	4
Prokaryote	GCA_002594505	Cyanophage	*Synechococcus phage ACG-2014f*	4
Eukaryote	GCF_000873685	NCLDV	*Paramecium bursaria Chlorella virus NY2A*	18
Eukaryote	GCF_000871245	NCLDV	*Paramecium bursaria Chlorella virus AR158*	16
Eukaryote	JX997170	NCLDV	*Paramecium bursaria Chlorella virus IL-5-2s1*	15
Eukaryote	JX997182	NCLDV	*Paramecium bursaria Chlorella virus NY-2B*	15
Eukaryote	JX997172	NCLDV	*Paramecium bursaria Chlorella virus MA-1D*	15

*NCLDV = Nucleocytoplasmic Large dsDNA Viruses

Table 3.2 (continued)

Host	Accession	Virus Type	Viral Species/Strain	MTase
Eukaryote	JX997183	NCLDV	*Paramecium bursaria Chlorella virus NYs1*	13
Eukaryote	JX997160	NCLDV	*Paramecium bursaria Chlorella virus CVB-1*	11
Eukaryote	HQ704802	NCLDV	*Organic Lake Phycodnavirus 1*	8
Eukaryote	GCF_000922335	NCLDV	*Aureococcus anophagefferens virus*	6
Eukaryote	GCF_000847045	NCLDV	*Paramecium bursaria Chlorella virus 1*	5
Eukaryote	GCF_000889395	NCLDV	*Cafeteria roenbergensis virus BV-PW1*	5
Eukaryote	GCF_000905435	NCLDV	*Ostreococcus lucimarinus virus OlV5*	5
Eukaryote	GCF_000907415	NCLDV	*Phaeocystis globosa virus 16T*	5
Eukaryote	JX997163	NCLDV	*Paramecium bursaria Chlorella virus CVM-1*	5
Eukaryote	KY322437	NCLDV	*Tetraselmis virus 1*	5
Eukaryote	JX997176	NCLDV	*Paramecium bursaria Chlorella virus NE-JV-1*	5
Eukaryote	GCF_000887855	NCLDV	*Ostreococcus tauri virus 2*	4
Eukaryote	GCF_001887825	NCLDV	*Only Syngen Nebraska Virus 5*	4
Eukaryote	JX997159	NCLDV	*Paramecium bursaria Chlorella virus CVA-1*	4
Eukaryote	JX997154	NCLDV	*Paramecium bursaria Chlorella virus AP110A*	4

*NCLDV = Nucleocytoplasmic Large dsDNA Viruses

Table 3.3 Top ten genomic regions enriched in motifs in a 256 bp window

Location	Genes Impacted	MC	GATC	CATG	Txc	Annotations
111716-111972	*A219/222/226R*	4.5	4	5	Early	Glycosyltransferase [4.0E-6]
315461-315717	*A656L*	4.5	9	0	Early	Collagen Triple Repeat (20 copies) [9.1E-11]
244295-244551	*A505L a509R, a508R*	4	6	2	Early	Hypothetical protien
265551-265807	*A552R, a553L*	4	2	6	Early	Transcription Factor TFIID [3.1E-7]
2793-3049	*A005R*	3.5	3	4	Early	Ankyrin repeat [6.3e-11]
108294-108550	*A214L*	3.5	4	3	Early	Hypothetical protein
167096-167352	*A330R, a331L*	3.5	5	2	Early-Late	Ankyrin repeat [1.3e-07]
286981-287237	*A598L, a599R*	3.5	1	6	Early-Late	Histidine Decarboxylase [3.0E-53]
289453-289709	*A604L*	3.5	1	6	Early	Hypothetical protein
22131-22387	*A035L*	3	1	5	Late	Hypothetical protein

Gene names denoted with an upper-case ‘A’ are defined as major ORFs that have been detected in transcripts and/or proteomes, whereas minor ORFS have not been detected in those studies and are denoted with a lower-case ‘a’ (20, 21). Motif concentration, denoted as MC, represents fold enrichment or depletion of motifs based on a window size of 256 base pairs (see Materials and Methods). GATC and CATG columns list the number of each motif observed in the window. Txc denotes the stage at which transcripts for major ORFs are detected. Annotations are listed for only the major ORFs and tRNAs; only one is listed per gene, and those given with an e-value represent the highest confidence annotation based on COG, Pfam, or KEGG hits.

Table 3.4 Top ten genomic regions depleted in motifs in a 256 bp window

Location	Genes Impacted	MC	Txc	Annotations
62092-65709	*A121R*	-28	Early-Late	Hypothetical protein
	A122/123R		Early	Autotransporter adhesion [1.0E-12] (glycoprotein)
10109-12557	*A014R*	-19	Late	Hypothetical protein
	A018L		Late	Glycoprotein repeat [1.2E-11]
	a016L, a017L			
126568-128978	*A251R*	-18	Early	M.CviAII (CATG) Methyltransferase
	A252R		Early	R.CviAII (CATG) Restriction Endonuclease
	A253R		Early	Hypothetical protein
	A254R		Late	Hypothetical protein
	a253aR, a252bL, a251bL, a252aL, a251aL			
180346-182661	*A363R, A368L*	-18	Early	Hypothetical proteins
	A366L		Early-Late	Hypothetical protein
	a367R, a365L			
299832-302126	*A623aL*	-17	n/a	Hypothetical protein
	A623L		Early	AN1-like Zinc finger [1.7E-12]
	A624R		Late	Predicted membrane protein [3.4E-26]
	A625R		Late	Transposase IS605 OrfB Family [2.0E-20]
	A627R		Late	Hypothetical protein
	a626L, a626aR			
297089-299232	*A619L, A620L, A621L*	-16	Late	Hypothetical proteins
	A622L		Late	Capsid Protein
	a621bL, a621aR, a620aR			

Table 3.4 (continued)

Location	Genes Impacted	MC	Txc	Annotations
195248-197302	*A401R, A403R*	-16	Early-Late	Hypothetical proteins
	A402R, A404R, A405R		Late	Hypothetical proteins
	A404aL		n/a	Hypothetical protein
253391-255424	*A532L*	-15	Late	Hypothetical protein
	A532aL, A534R		n/a	Hypothetical proteins
	A533R, A535L, A536L		Early-Late	Hypothetical proteins
163902-165909	*A328L*	-15	n/a	Hypothetical protein
	A329R		Late	Hypothetical protein
	a329aL		n/a	9/10 Putatively functional tRNAs
	Lys-3, Tyr-1, Ile-1, Leu-1, Lys-2, Asn-2, Arg-1, Lys-1, Asn-1, Pseudo-tRNA-1			
39482-41467	*A075L*	-15	Early-Late	Exostosin Family [5.5E-9]
	A075cR		n/a	Hypothetical protein
	A075bl		n/a	Hypothetical protein
	A076L		n/a	Hypothetical protein
	A077L		Early	Hypothetical protein
	A078R		Early	N-carbamoylputrescine amidohydrolase

Gene names denoted with an upper-case ‘A’ are defined as major ORFs that have been detected in transcripts and/or proteomes, whereas minor ORFS have not been detected are denoted with a lower-case ‘a’ (20, 21). Motif concentration, denoted as MC, represents fold enrichment or depletion of motifs based on a window size of 256 base pairs (see Materials and Methods). GATC and CATG columns list the number of each motif observed in the window. Txc denotes the stage at which transcripts for major ORFs are detected. Annotations are listed for only the major ORFs and tRNAs; only one is listed per gene, and those given with an e-value represent the highest confidence annotation based on COG, Pfam, or KEGG hits.

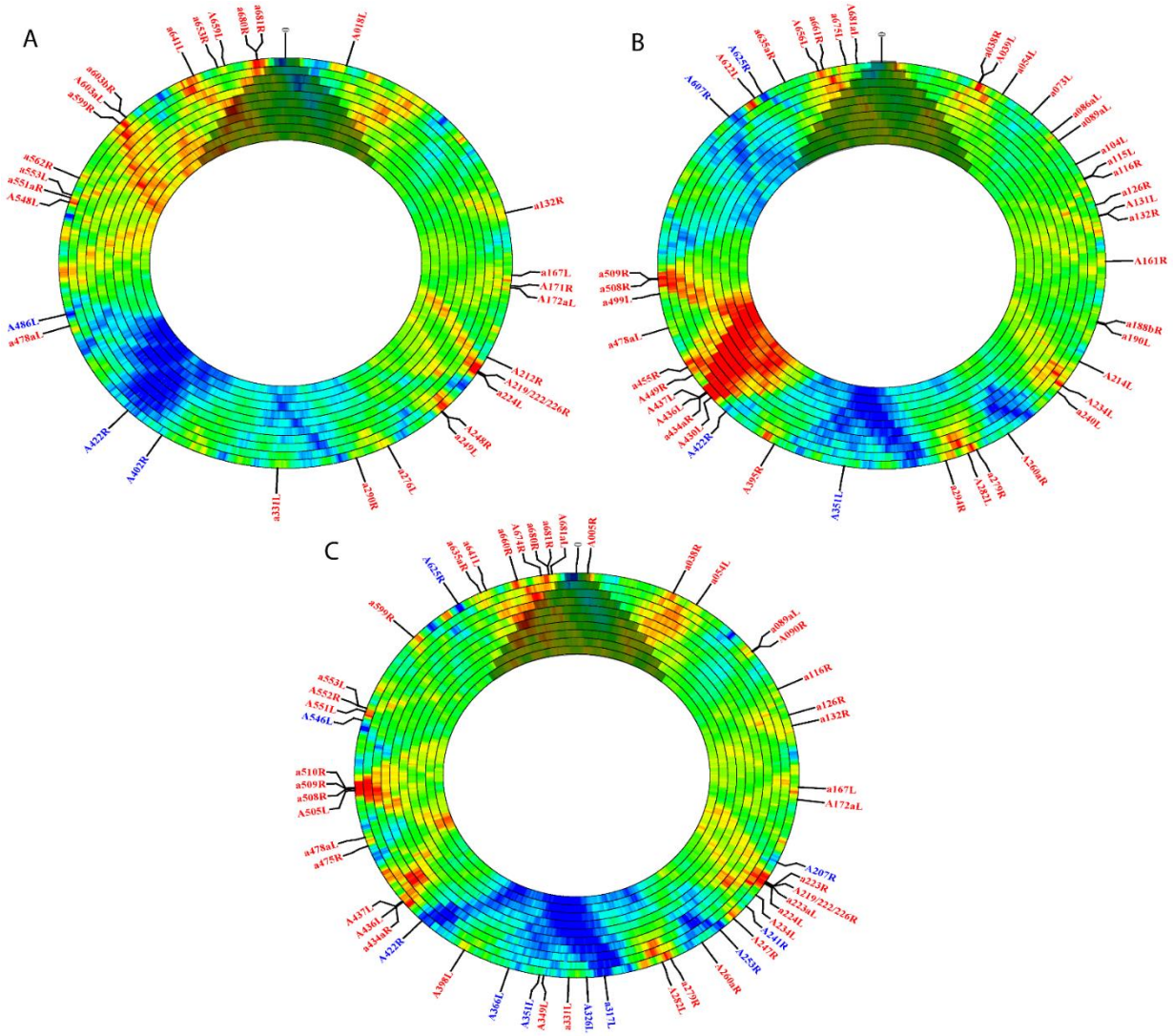


Figure 3.4 PBCV-1 motif frequency as a function of coding sequence

Frequency was determined in the context of local codon flexibility at different genomic increments. The PBCV-1 genome was analyzed for motif frequency of CATG (A), GATC (B), and both CATG and GATC (C) tetramers. Panels show increasing window sizes, ranging from 4 to 40kb moving from the outer to inner ring. Each ring increases incrementally by 4kb, but the step size is always 4kb. Z-score is represented with color; red indicates more enriched scores and blue indicates more scores. Green represents expected frequency with no depletion or enrichment. Since the DistAMo software was created for circular chromosomes, there is artificial data coded across the termini for PBCV-1. This region has been shaded over to discourage faulty interpretation. Genes with significant Z-scores (greater than the absolute value of 2) are annotated by color and are included in Supplemental Table S2.

Table 3.5 PBCV-1 ORFs enriched or depleted in GATC or CATG motifs

Gene Name	motif	zScore	Start	End	Accession
a478aL	GATC	4.2	231306	231812	NP_048835.2
a126R	GATC	3.87	66620	66820	NP_048474.1
a508R	GATC	3.75	244334	244567	NP_048864.1
a661R	GATC	3.65	316543	316851	NP_049017.1
A437L	GATC	3.64	212519	212830	NP_048794.2
a509R	GATC	3.57	244423	244728	NP_048865.1
a279R	GATC	3.54	142146	142364	NP_048633.1
a434aR	GATC	3.45	212284	212472	YP_004678953
a116R	GATC	3.4	59398	59643	NP_048464.1
a038R	GATC	3.35	23584	23823	NP_048386.1
a190L	GATC	3.19	97525	97743	NP_048537.1
A436L	GATC	3.04	212299	212490	NP_048793.2
a499L	GATC	3.04	240483	240716	NP_048855.1
A622L	GATC	3	298138	299700	NP_048978.1
a086aL	GATC	2.93	45101	45229	YP_004678889
a294R	GATC	2.88	149926	150150	NP_048648.1
A234L	GATC	2.85	115777	116103	NP_048582.1
A214L	GATC	2.67	108265	108672	NP_048561.1
a089aL	GATC	2.67	47826	47972	YP_004678892
a132R	GATC	2.6	69533	69805	NP_048480.1
a635aR	GATC	2.59	307064	307258	YP_004678992
a054L	GATC	2.58	29800	30123	NP_048402.1
a073L	GATC	2.55	38417	38626	NP_048421.1
A430L	GATC	2.47	210155	211468	NP_048787.1
a240L	GATC	2.46	117770	117967	NP_048588.1
A260aR	GATC	2.41	133700	133897	NP_048614.3
a675L	GATC	2.4	321966	322334	NP_049031.1
A282L	GATC	2.4	143630	145339	NP_048636.1
A656L	GATC	2.22	315127	315849	NP_049012.2
a115L	GATC	2.18	59265	59495	NP_048463.1
A449R	GATC	2.18	217799	218380	NP_048806.1
A039L	GATC	2.15	23623	24078	NP_048387.1
A161R	GATC	2.14	81345	81716	NP_048509.1

Table 3.5 (continued)

Gene Name	motif	zScore	Start	End	Accession
A681aL	GATC	2.11	324693	324869	YP_004678999
a104L	GATC	2.1	55054	55344	NP_048452.1
A395R	GATC	2.09	191505	191753	NP_048752.1
A131L	GATC	2.09	69359	69769	NP_048479.1
a188bR	GATC	2.08	97258	97398	YP_004678909
a455R	GATC	2.03	220218	220661	NP_048812.1
A607R	GATC	-2.04	290633	291808	NP_048963.2
A351L	GATC	-2.2	173636	174712	NP_048708.1
A422R	GATC	-2.21	205267	206259	NP_048779.2
A625R	GATC	-2.39	300424	301722	NP_048981.2
a553L	CATG	4.15	265624	265839	NP_048909.1
A172aL	CATG	3.38	88944	89111	YP_004678906
a132R	CATG	3.14	69533	69805	NP_048480.1
a167L	CATG	3.09	85677	85880	NP_048515.1
A603aL	CATG	2.84	289390	289575	YP_004678983
a478aL	CATG	2.8	231306	231812	NP_048835.2
a641L	CATG	2.8	308469	308726	NP_048997.1
a224L	CATG	2.79	112197	112463	NP_048572.1
a551aR	CATG	2.75	264946	265074	YP_004678972
a680R	CATG	2.68	323837	324100	NP_049036.1
A219/222/226R	CATG	2.56	110893	112926	NP_048569.4
a331L	CATG	2.49	167096	167299	NP_048687.1
a249L	CATG	2.42	125191	125499	NP_048598.1
A212R	CATG	2.31	107615	107782	NP_048559.2
a276L	CATG	2.31	140462	140746	NP_048630.1
a290R	CATG	2.27	148366	148773	NP_048644.1
a603bR	CATG	2.24	289445	289591	YP_004678984
a562R	CATG	2.17	270374	270571	NP_048918.1
a681R	CATG	2.14	323857	324081	NP_049037.1
A018L	CATG	2.14	12367	16374	NP_048366.1
a653R	CATG	2.08	314450	314647	NP_049009.1
A402R	CATG	-2.01	195325	196008	NP_048759.1
A486L	CATG	-2.2	234401	234859	NP_048842.1
A422R	CATG	-2.3	205267	206259	NP_048779.2

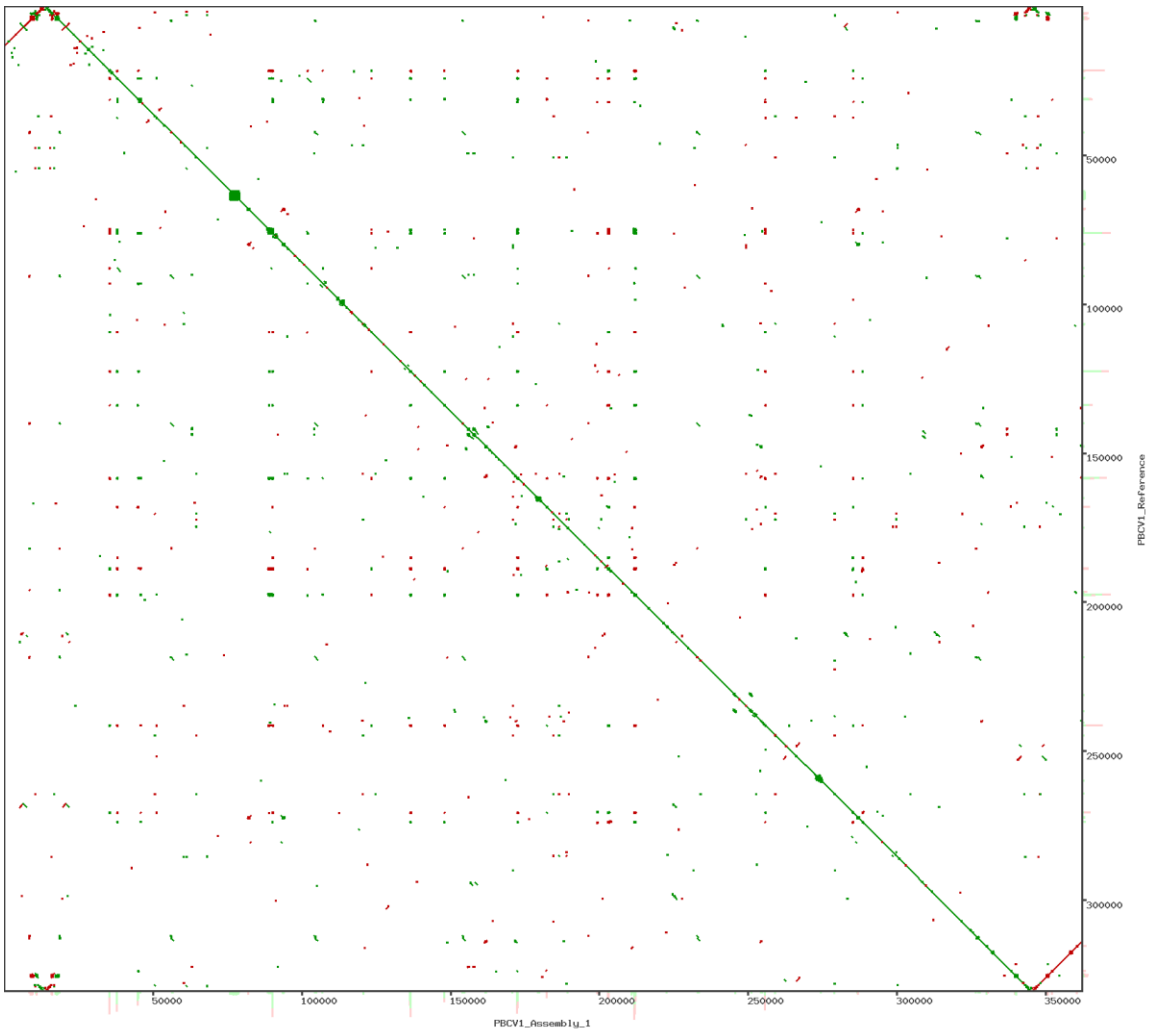


Figure 3.5 Dot plot alignments between de novo and reference sequences

RefSeq (y-axis) and a representative *de novo* assembled Pacbio sequence (x-axis).

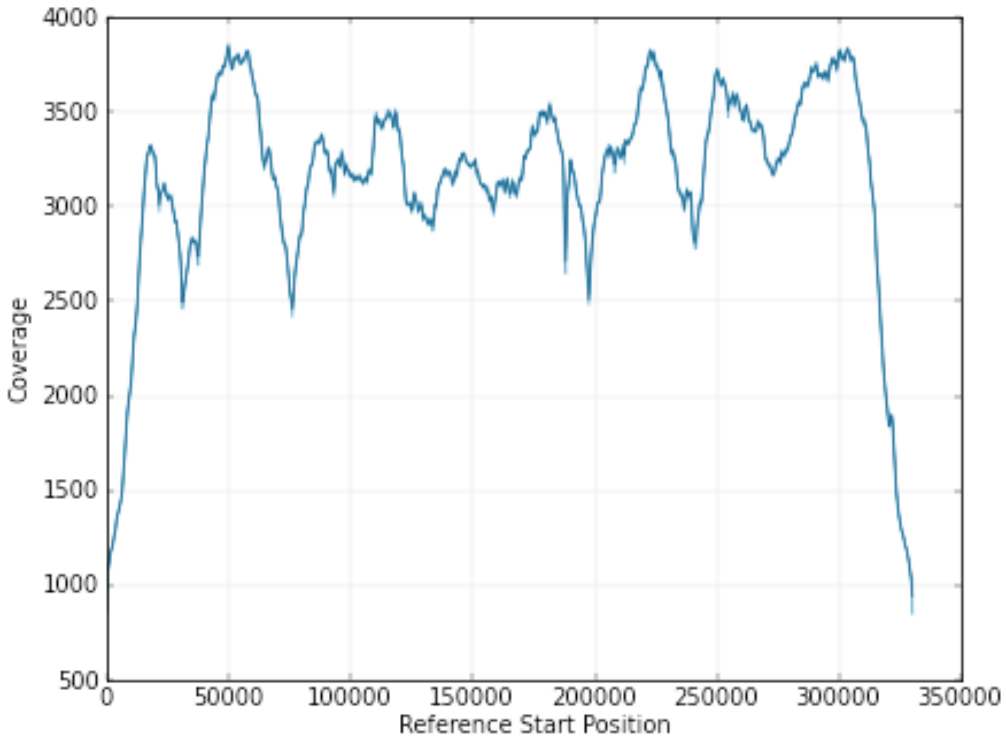


Figure 3.6 Representative read coverage for Pacbio generated reads

Reads were mapped against the PBCV-1 reference genome (x-axis).

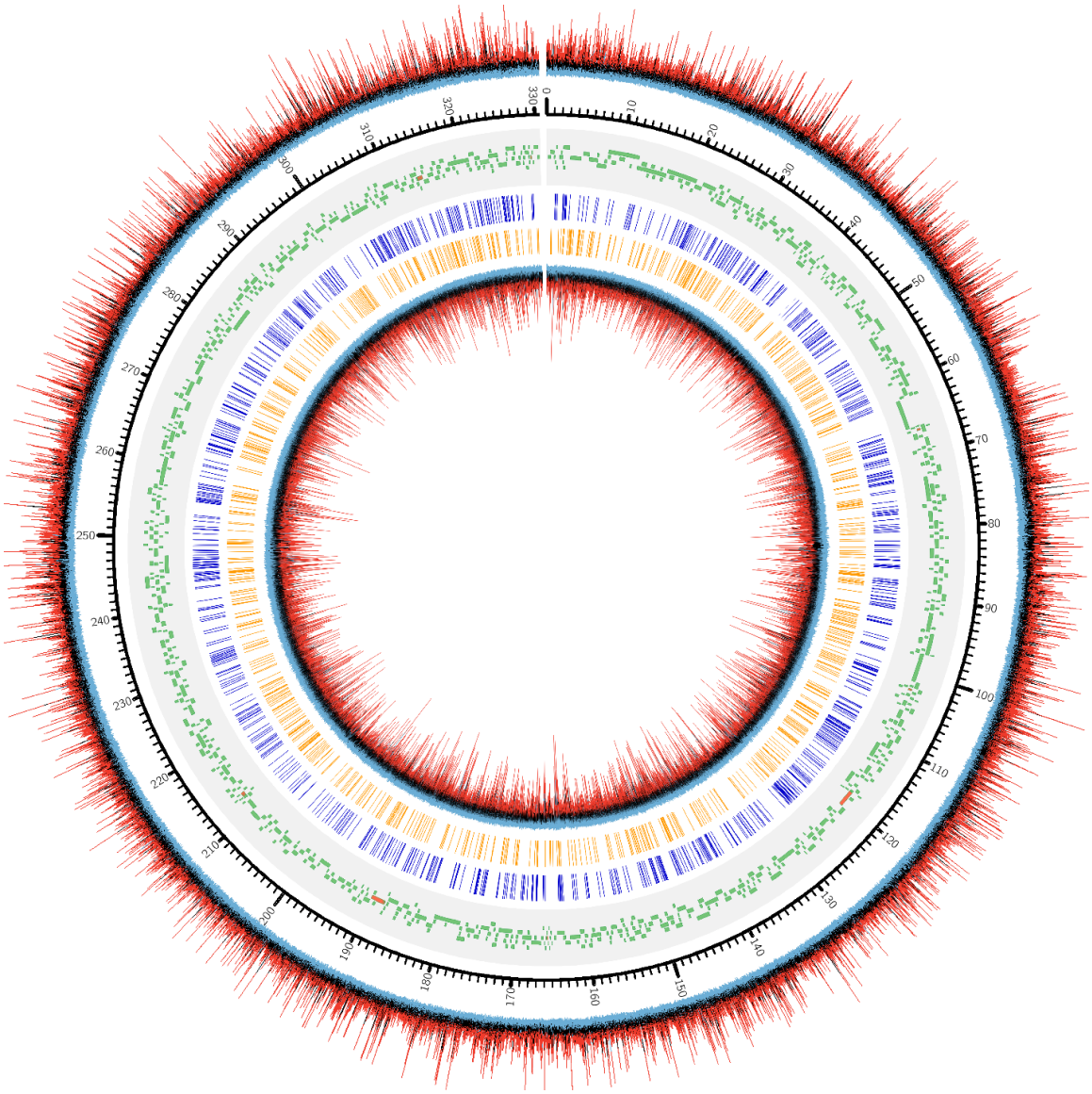


Figure 3.7 PacBio data for chlorovirus PBCV-1

Starting from the outermost ring and going inward: 1) ipdRatio for each nucleotide on the forward strand, 2) genomic positions for PBCV-1 (kbp), 3) PBCV-1 potential protein coding sequences and tRNA encoding sequences (Dunigan et al., 2012), 4) CATG sites, 5) GATC sites, and 6) ipdRatio for each nucleotide on the reverse strand. The color coding of the ipdRatio is artificial to denote peak values, with a red peak denoting a value >2. Black peaks indicate nucleotide kinetics similar to a non-modified base (ipdRatio ~1), and blue peaks indicate a value <0.5. This plot was made in CIRCOS (Krzywinski et al., 2009) with the intent of displaying ipdRatio peak height occurrence and diversity. Replicates exhibited CIRCOS plots with nearly imperceptible differences, which is why only one representative is shown here.

Table 3.6 Characteristics of PBCV-1 motifs defined as modified by Pacbio

Replicate	Coverage	Motif	Modification	Fraction	modQV (μ)	IpdRatio (μ)
PBCV1-R1	30	CATG	m6A	0.9672586	52.807804	5.4444346
	30	GATC	m6A	0.8443396	50.525837	4.466902
	30	GNNNNVNH	modified_base	0.052710593	35.31804	3.0786736
	30	CNNNNRNH	m5C	0.019287998	37.418095	2.5839715
	255	CATG	m6A	0.98612654	326.93134	5.298204
	255	GATC	m6A	0.870283	297.26627	4.273591
	255	DTNRRDDDG	modified_base	0.17467625	46.228996	1.6845644
	255	TNNNDNNH	modified_base	0.11900781	45.164227	1.6761321
	255	TNNNCRVH	modified_base	0.08545584	42.325905	1.6427859
	1509.6927	CATG	m6A	0.9916759	1318.0654	5.2829576
	1499.319	GATC	m6A	0.9817217	988.5249	4.248782
	1460.0975	AGDVAAAAW	m6A	0.4939759	259.4878	1.4247562
	1488.6095	TNNNNNNH	modified_base	0.16766186	71.82713	1.5380374
	1486.9062	TNNNDNDG	modified_base	0.09227239	66.023094	1.5094546
PBCV1-R2	30	CATG	m6A	0.9672586	52.807804	5.4444346
	30	GATC	m6A	0.8443396	50.525837	4.466902
	30	GNNNNVNH	modified_base	0.052710593	35.31804	3.0786736
	30	CNNNNRNH	m5C	0.019287998	37.418095	2.5839715
	255	MNNGANGCAGYA	m6A	1	141.16667	1.7216667
	255	CATG	m6A	0.9889012	328.59653	5.322953
	255	GATC	m6A	0.8832547	300.22498	4.2772593
	255	TNNNDNNH	modified_base	0.11322659	44.60365	1.674846
	255	TNVRDDDG	modified_base	0.11033353	43.268555	1.6505132
	255	TNNNCRVH	modified_base	0.07934619	42.278	1.6454105
	1653.5283	CATG	m6A	0.9916759	1410.751	5.3039694
	1643.3647	GATC	m6A	0.9829009	1047.8032	4.2681665
	1575.7595	AGDVAAAAW	m6A	0.4759036	273.10126	1.4377215
	1625.5973	TNNNNNNH	modified_base	0.16515067	71.32479	1.5295854
	1616.6515	TNRVNNDG	modified_base	0.16131958	67.71468	1.509661

Table 3.6 (continued)

Replicate	Coverage	Motif	Modification	Fraction	modQV (μ)	IpdRatio (μ)
PBCV1-R3	30	GNATWATNGCA	modified_base	1	38.6	2.722
	30	CATG	m6A	0.963929	52.84226	5.422868
	30	GATC	m6A	0.8402123	50.331226	4.3777704
	30	GNNNNVNH	modified_base	0.05321101	35.40935	3.1023834
	30	GNVVNTBH	modified_base	0.04541603	34.830223	2.9693081
	30	CNNNNRNH	m5C	0.021161688	37.25434	2.5690968
	255	TNAGAGTTNKNNNNNNG	m6A	1	76.6	1.464
	255	DNTNNGCATAANT	modified_base	1	48.6	1.7650001
	255	CATG	m6A	0.98612654	324.51773	5.2549305
	255	WNNNNNGANGCAGCA	m6A	0.9166667	139.36363	1.6345454
	255	WGAGGCNNNTNYA	m6A	0.875	89.28571	1.4628571
	255	ANNKNTNTNNGCNTNNTT	modified_base	0.875	46.57143	1.6557142
	255	GATC	m6A	0.8649764	298.743	4.2358856
	255	TNNNAGTTNGNANTNNNT	m6A	0.85714287	69.5	1.6666666
	255	HNNNNNAGGCMNTTG	m6A	0.85714287	90.333336	1.6550001
	255	TNACGANAANTNNNNNA	m6A	0.8333333	97.4	1.5059999
	255	TNNNTTGANNNAGNNNTG	m6A	0.8333333	86.6	1.7739999
	255	ANANNNAGNGNGNNAYT	m6A	0.75	78.5	1.7916666
	255	AGAGAAWAA	m6A	0.75	103	1.6749998
	255	ANANTNANAGANNANNY	m6A	0.72727275	61.375	1.5675
	255	YNNAGGNWAAANT	m6A	0.6666667	64	1.62
	255	TNNNNNNASYTASTA	m6A	0.6666667	79.4	1.8400002
	255	ANNNNTNAGNAAAAA	m6A	0.57894737	78.63636	1.4945455
	255	GNANNNNHANNTGGCA	m6A	0.5714286	75.5	1.58875
	255	AGNAAATTTT	m6A	0.5	82.28571	1.8385714
	255	ADKYAGYANY	m6A	0.41666666	127.825	2.0889997
	255	TNRADRRG	modified_base	0.27981222	45.78859	1.6870131
	255	TNNNDNNH	modified_base	0.11198833	44.66702	1.6770811
	255	TNNNCRVH	modified_base	0.08045703	42.001972	1.6423571
	254.94911	TVNNNDDG	modified_base	0.061388757	42.930527	1.6568396

Table 3.6 (continued)

Replicate	Coverage	Motif	Modification	Fraction	modQV (μ)	IpdRatio (μ)
PBCV1-R3	1469.5	GNNANNTNGCANTNNCA	m6A	1	343.83334	1.5016667
	1340.2858	ANNAGANNNAGCAA	m6A	1	268.7143	1.5757143
	1397	AATGANGAANNNT	m6A	1	344	1.4133333
	1342.5695	CATG	m6A	0.9900111	1203.7657	5.2254577
	1338.9729	GATC	m6A	0.9811321	919.5919	4.2007504
	1264.5454	AGNVAAAAWH	m6A	0.48125	234.22078	1.4315586
	1340.3623	TNNNNNNH	modified_base	0.16286087	69.2657	1.5390226
	1336.2245	TNRVNNDG	modified_base	0.15905227	64.91661	1.5147673
	1318.8641	TGYNNNNG	modified_base	0.13299957	64.89689	1.5278397
	1242.2812	AGKNNNNH	m6A	0.054108746	144.39024	1.4278698

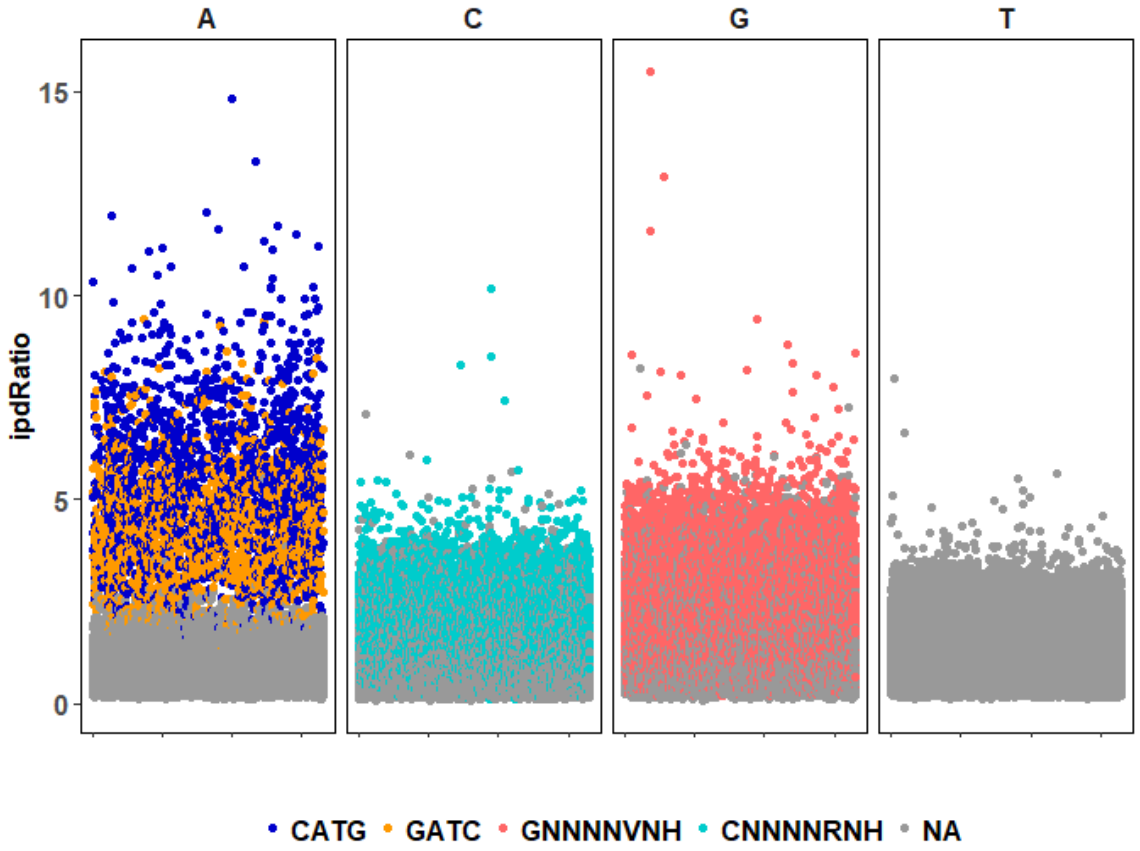


Figure 3.8 ipdRatio score for nucleotides in one replicate of the PBCV-1 genome

Dot color denotes association with a motif detected by motifMaker.sh. There are nearly imperceptible differences between replicates, which is why only one is shown here.

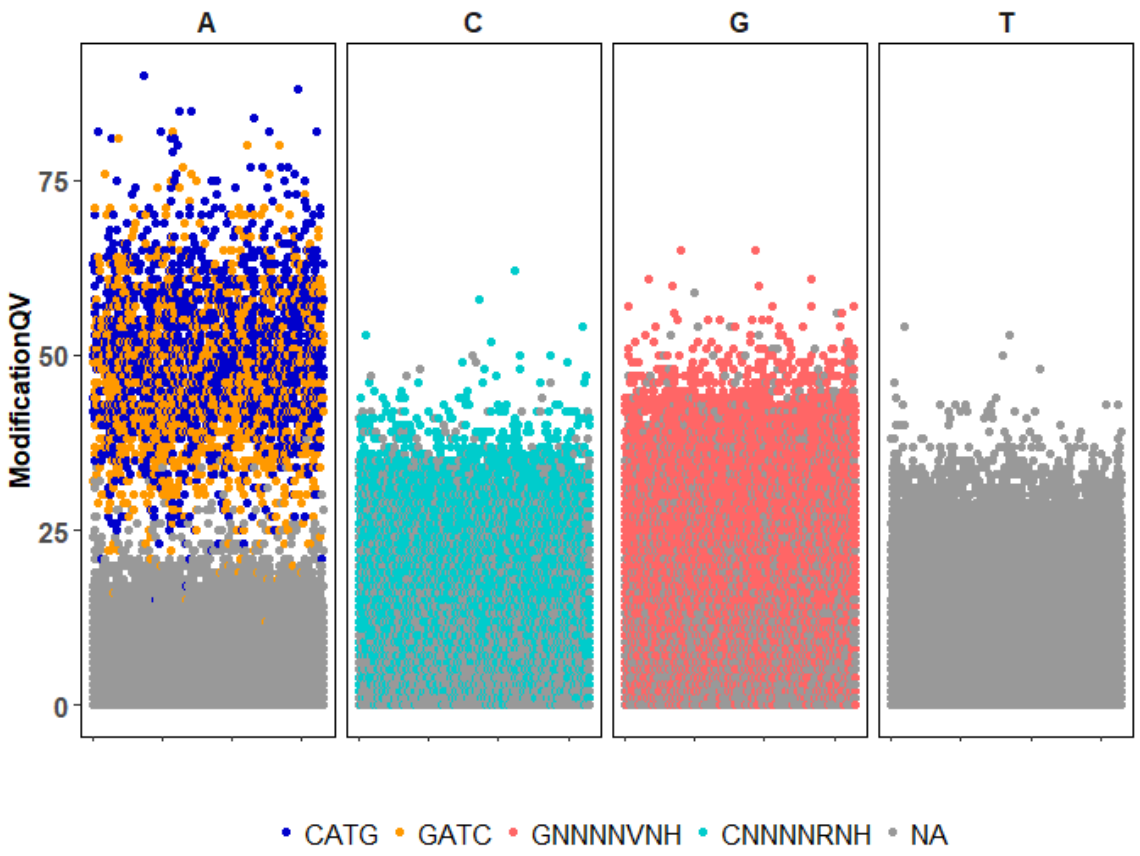


Figure 3.9 Average modification QV scores for all PBCV-1 nucleotides

Dot color denotes association with a motif detected by motifMaker.sh. Motifs not detected in all three replicates are not shown.

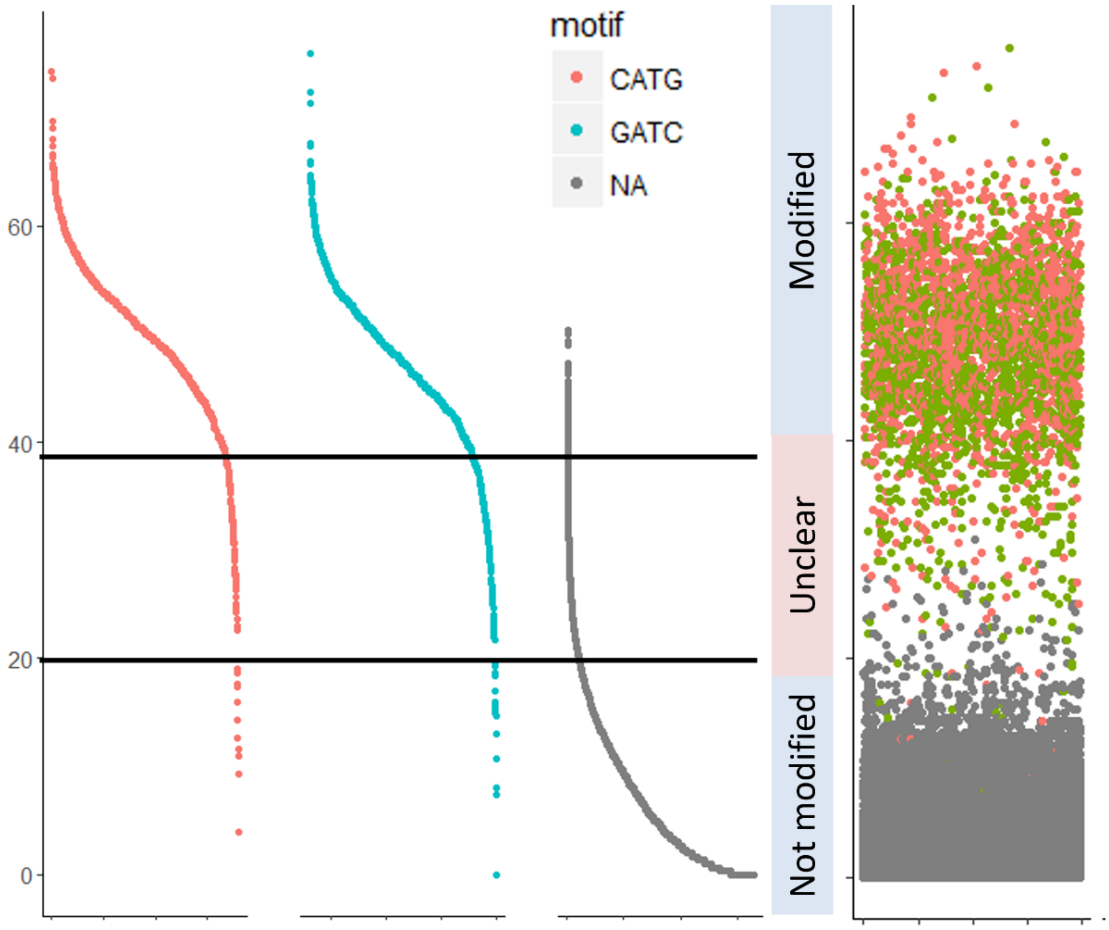


Figure 3.10 Re-visualization of PBCV-1 adenine data as a rank ordered distribution

Boundaries have been overlaid to demonstrate how one could confidently identify modified and non-modified sites, yet, there is an unclear region wherein the modification status is uncertain. Thus, we used the default Modification QV value of 30 as a threshold for deciding modification status.

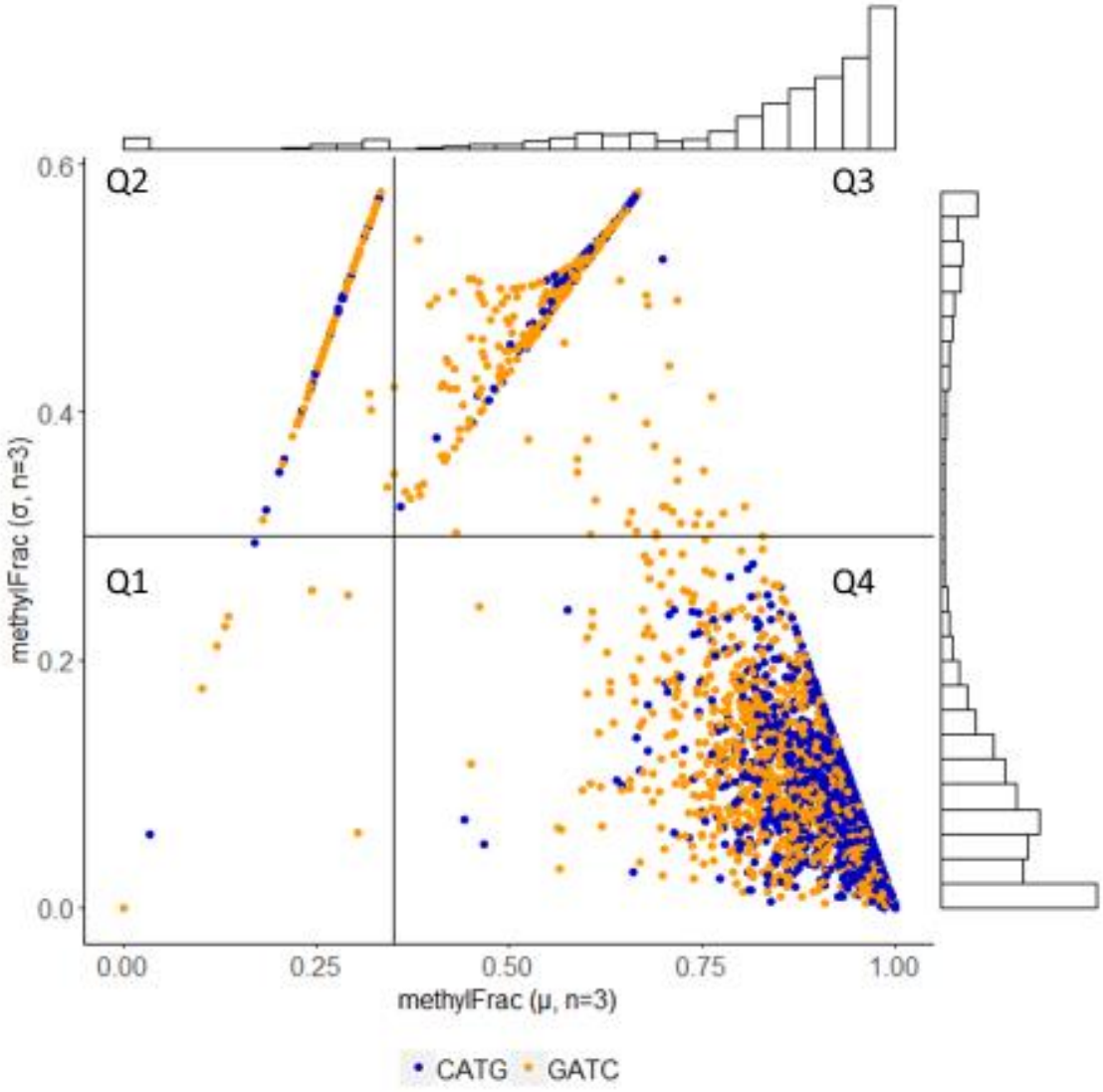


Figure 3.11 Methylation stability of CATG and GATC tetramers in PBCV-1

Each dot represents the average MethylFrac value computed for three biological replicates. An average value approaching one indicates stable methylation, whereas closer to zero indicates stable non-methylation. Standard deviation is used to define methylation variation between the three biological replicates. Histogram plots provide an estimate of how many events occur with the given coordinates. Q1 represents 73 events (21.9% CATG, 57% GATC); Q2 represents 457 events (23.1% CATG, 76.9% GATC); Q3 represents 143 events (24.3% CATG, 75.7% GATC); Q4 represents 2,825 events (58.1% CATG, 41.9% GATC).

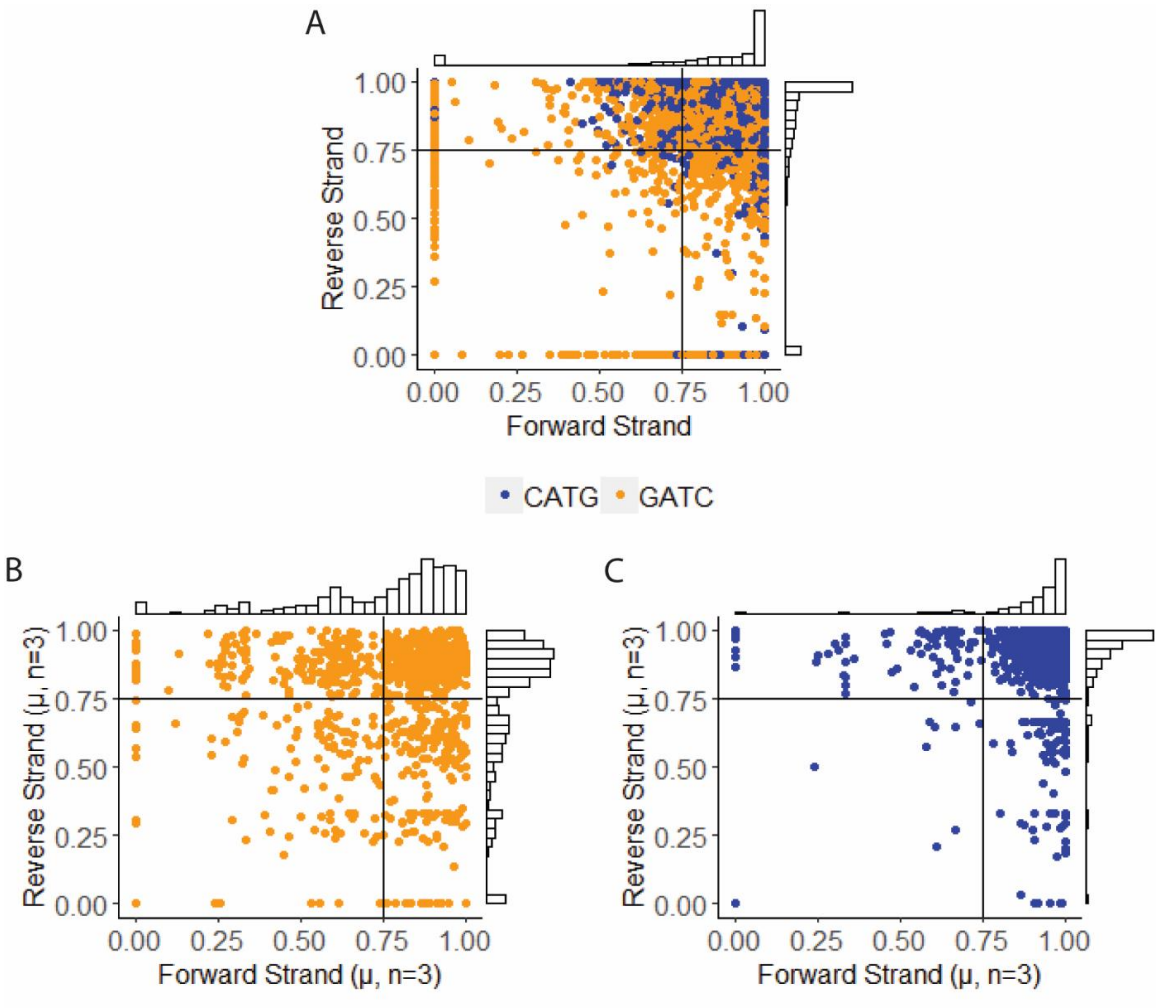


Figure 3.12 Methylation stability of CATG and GATC

Stability in palindromes across three sequenced replicates of PBCV-1 (A). The averaged methylFrac for the forward and reverse strand of a single palindrome are plotted for the separated motifs (B-C). Histogram plots provide an estimate of how many events occur with the given coordinates, with lines marking thresholds for defining complete methylation, hemimethylation, and stochastic methylation.

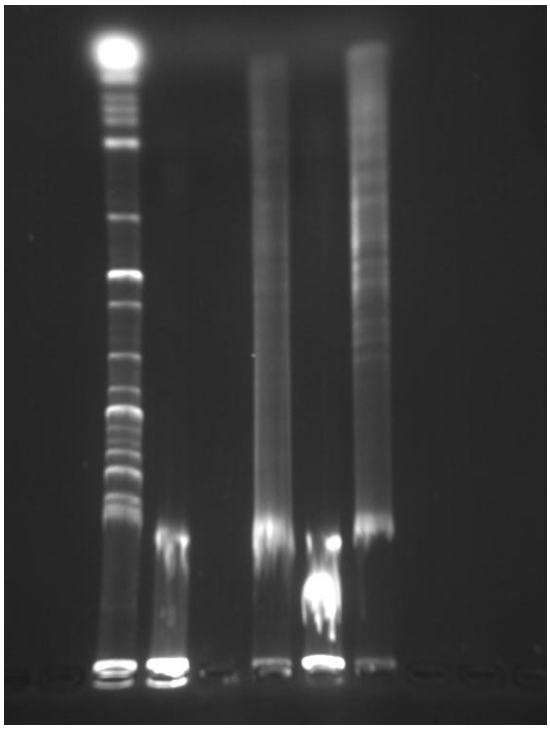


Figure 3.13 Restriction digestion analysis of chlorovirus PBCV-1 genomic DNA

Wells from left to right denote 1) 40kb Extension Ladder; 2) PBCV-1 genomic DNA; 3) Loading Buffer; 4) PBCV-1 DNA + DpnI; 5) PBCV-1 DNA + DpnII; 6) PBCV-1 DNA + Sau3AI. Electroporation was carried out on a 0.7% agarose gel.

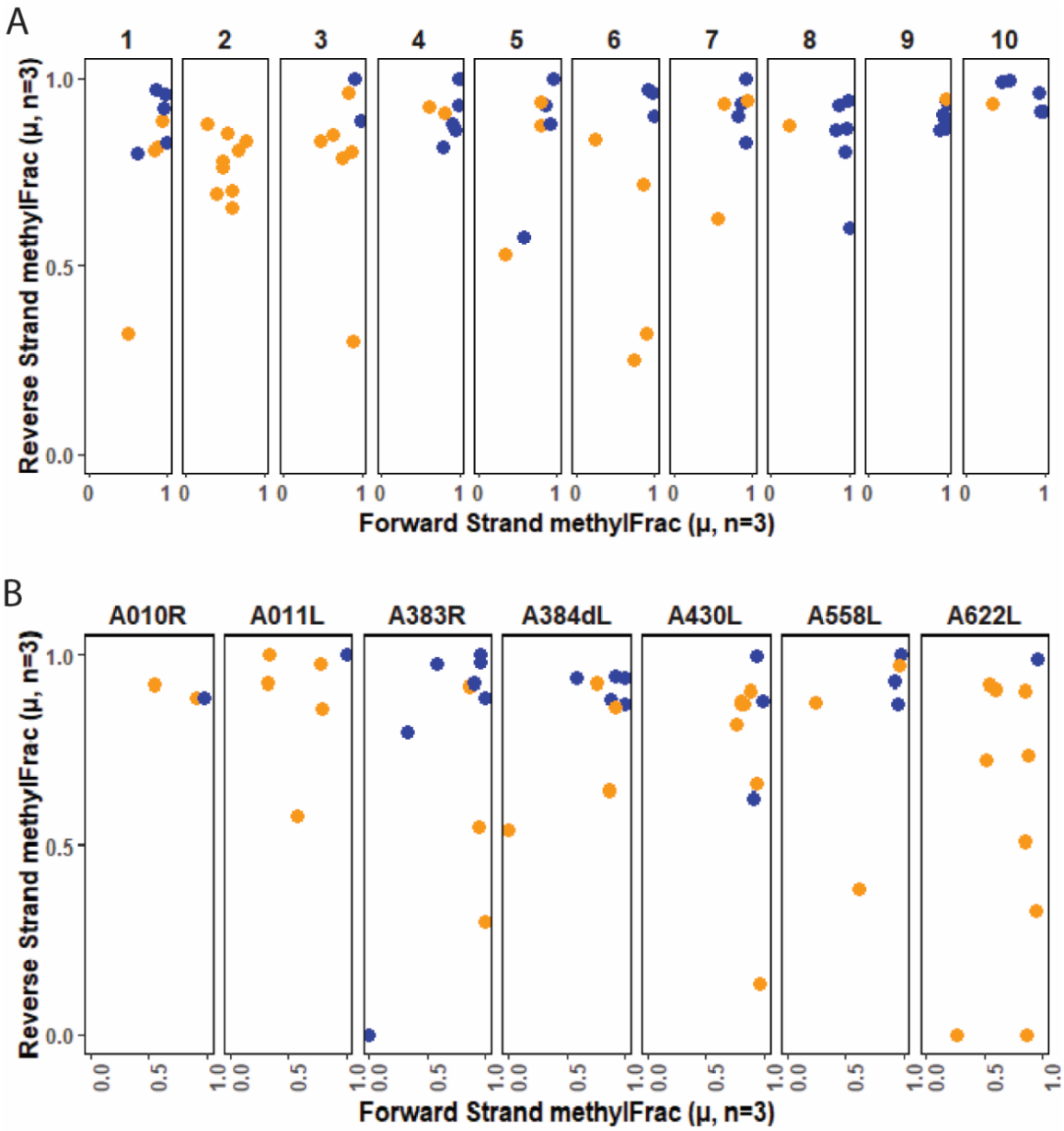


Figure 3.14 Methylation stability of palindromes grouped in genomic contexts

Palindromes are graphed for the top ten 256 bp regions previously identified in in silico analyses (Table 2) as enriched in one or both methyltransferases (A). Palindrome methylation characteristics can also be graphed by functional protein groups, with capsid proteins shown here (B). The major capsid protein is encoded by *A430L.*

CHAPTER 4 : INFECTION OF CHLORELLA VARIABILIS WITH PBCV-1 INDUCES OXIDATIVE STRESS AND CHANGES IN BIOCHEMICALLY RELATED, INTRACELLULAR AMINO ACICDS

This chapter is in prep for manuscript submission.

Samantha R. Coy, Katarina Jones, Brennan J. Hughes, Hector F. Castro-Gonzalez, Shawn Campagna, James L. Van Etten, Steve W. Wilhelm. “Infection of *Chorella variabilis* with PBCV-1 induces oxidative stress and changes in intracellular amino acids linked by biochemistry.”

S.R.C. performed the experiment, extracted metabolites, analyzed data, and wrote the manuscript. K.J. processed the samples on the UPLC and performed preliminary analyses and quantification of the isotope-labeled metabolites. B.J.H. enumerated virus particles. H.F.G. and S.C. helped to guide the metabolomics approach. J.L.V. provided research materials and direction for the study. S.W.W. guided conception of the paper and provided funding for research materials. All listed authors contributed to the production of figures, text, and editing.

Abstract

Viruses have a significant influence on carbon and nutrient cycles in aquatic ecosystems. However, this impact has typically only been investigated as a consequence of lytic infection, and interest is growing in the metabolism of virus infected cells, coined ‘virocells’. Here, we investigate changes in the metabolic profile across the course of *Paramecium bursaria chlorella virus 1* (PBCV-1) infection in the green alga, *Chlorella variabilis* NC64A. In total, 102 small (<1000 Dalton), water-soluble metabolites were detected and monitored for metabolic fold changes across the infection cycle. Over a quarter of these (28.4%) demonstrated significant fold changes (>1.5 fold change; p-value <0.05) within seven minutes of infection, which increased to 68.6% of detected metabolites by the end of infection. Since viruses are primarily made of proteins and nucleic acid, we expected virocells to deplete these host resources for their replication. Depleted and enriched amino acids group by biochemical similarities, indicating dynamic physiologic and/or structural requirements. Specifically, aromatic and basic amino acids

are collectively enriched, whereas amidic, hydroxylic, and sulfur-containing amino acids are collectively depleted in infected cultures. Detected nucleotides are almost immediately enriched, and often continue to be throughout the course of infection. We also utilized stable-isotope-labeled standards to quantify glutamine, glutamate, and alpha-ketoglutarate, and compare these as ratios to detect nitrogen limitation, TCA cycle enrichment, and oxidative stress. Altogether, these studies serve as a baseline for understanding viral-mediated metabolic changes in the chlorovirus virocell, which serve as a model system for their giant, algal-infecting virus relatives.

Introduction

Viruses are more abundant than any other biological member of aquatic ecosystems (1). Their abundance is inextricably linked to a significant ecological footprint including alteration of biogeochemical fluxes, directed evolution of cellular hosts, and maintenance of ecosystem scale biodiversity. Much attention has been given to understanding these processes as a consequence of infection, but they typically occur in the aftermath of infection, specifically following lysis (2). The virus infected cell, coined the 'virocell', is a distinct metabolic state that is driven by the virus to support its replication (3-5). Indeed, because viruses are the biggest reservoir for genetic diversity (1), their infection can introduce virus-encoded auxillary metabolic genes (vAMGs) that provide the virocell with novel metabolic abilities (6).

Understanding the virocell's metabolism is a growing interest within aquatic biology, though most focus has been on ecologically relevant prokaryotes like heterotrophic bacteria (7). Recently, however, attention has turned to the effect of viruses on eukaryotic algae (8), as these organisms can rapidly proliferate to form blooms big enough to be seen from outer space (9). These coccolithophore infecting viruses are members of a unique phylogenetic group known as the Nucleocytoplasmic large dsDNA viruses (NCLDVs; *i.e.,* 'giant viruses'), whose members infect different types of freshwater and marine algae (10). Giant virus genomes encode hundred to thousands of genes, the majority of which are hypothetical proteins bearing little to no homology with functionally

characterized genes. Thus, the potential for vAMGs is high in these systems, and there is a need to investigate other giant virus-algae systems to quantify these effects.

The best studied model system of the giant, algae-infecting viruses are the chloroviruses. These large (genomes >300 kbp; particle sizes ~190 µm diameter) entities infect unicellular, ex-symbiotic, green, chlorella-like algae, with most characterization studies done on the prototype virus, *Paramecium bursaria Chlorella Virus 1* (PBCV-1). PBCV-1 encodes 416 predicted protein-encoding sequences, the majority of which are in the unknown-function category (11). Its infection cycle is completed in 6-8 hours (12), with distinct transcriptional phases divided by the onset of DNA replication (13, 14). Indeed, genes transcribed 'early' in the infection cycle express transcripts before replication (<60min PI), whereas genes transcribed 'late' in the infection cycle are expressed after DNA replication (>60 min PI). DNA replication putatively signals late transcription, as inhibition with aphidicolin can prevent the transcription of most 'late' expressed genes (13, 14). Unique physiological events have also been characterized to occur throughout the infection cycle, which likely bear important consequences on virocell metabolism.

Addition of chlorovirus PBCV-1 to *Chlorella variabilis* cultures demonstrates rapid viral adsorption to its host, followed by viral-mediated digestion of the host cell wall, and injection of its virion contents (15). Concomitant upon fusion, structual potassium ion channels in the virion membrane function to export host intracellular potassium ions, thus disrupting the host cell's electrochemical gradient to promote membrane depolarization (16, 17). While this has been shown to prevent secondary infection of competing viruses (18), a side effect of this is decreased activity of secondary active transporters (19). Consequently, the virocell might become nutritionally dependent on recycling of intracellular resources as opposed to extracellular uptake. This complicates the transition from a GC rich host genome (67.2%) (20) that must be recycled to produce up to a thousand PBCV-1 particles whose genomes are AT rich (60%) (11, 12), a discrepancy common among algal viruses (21). Another discrepancy is the elemental composition between hosts and viruses, the latter of which are comprised of mostly proteins and

nucleic acid, which make viruses arguably more nitrogen and phosphorous rich than cellular organisms (22). Altogether, there are many insights that can be gained from a metabolomic investigation of the *Chlorella variabilis* virocell.

In this study, we characterized the intracellular metabolic profile of *Chlorella variabilis* during the PBCV-1 infection cycle. This was accomplished using ultra performance liquid chromatography coupled with mass spectrometry for determination of small (<1000 Daltons), water-soluble metabolite concentrations in a virocell relative to a non-infected host. During this analysis, we used stable-isotopically-labeled internal standards to quantify glutamine, glutamate, and alpha-ketoglutarate, which are associated with central metabolic processes putatively necessary for viral infection. Glutamate to glutamine ratios are also known markers for monitoring nitrogen stress in bacteria (23). This study establishes a baseline for investigation into *Chlorella* virocells, which may bear important consequences for virus-host interactions in the environment as well as the symbiosis that occurs between *Chlorella* and its *Paramecium* host.

Materials and methods

Growth conditions and treatments

Three two liter cultures of *Chlorella variabilis* strain NC64A were grown at 25ºC with magnetic stirring on a plate (~150 rpm) in Modified Bold's Basal Medium (MBBM) (24). After reaching a culture density of 5.0 x 10^6 cells/mL (*i.e.* late exponential growth), the cultures were combined into a sterile Nalgene bottle, and promptly dispensed as 450mL aliquots into 12 sterile 1 L flasks. These flasks were returned to the incubator to acclimate for one hour with shaking (~150 rpm) instead of stirring to accommodate a larger number of flasks. At this point, flasks were retrieved for sample collection and treatment. First, a 10mL metabolite sample was collected unto a 0.22 µm polycarbonate filter, flash frozen, and transferred to a -80ºC freezer for storage. For nucleic acid extraction, 30mL were pelleted at 10,000 x g for 1 min at 4ºC. The cell pellet was flash frozen and transferred to the -80ºC freezer until nucleic acid extraction. For cell and virus-like-particle (VLP) counts,

a one milliliter sample was fixed at a final concentration of 0.5% glutaraldehyde for 15 min on ice, in the dark. This was flash frozen and stored at -80ºC until analysis by flow cytometry. This process was done for each flask to complete T0 sample collection. At this point, flasks 7 thru 12 were infected with virus PBCV-1 at an MOI of 4. At 7 min post-infection, all flasks were sampled again for metabolites, nucleic acid, cell, and VLP measurements, then returned to the incubator. Sample collection was repeated at 30, 60, 120, 240, and 360 min.

Metabolite extraction

An extraction solvent was prepared containing a 40:40:20 mixture of HPLC grade methanol, acetonitrile, and water with 0.1 M formic acid (25). Half of this stock included an internal standard cocktail comprised of heavy labelled L-glutamic acid (^{15}N, 98%, #CLM-2411-0.01), L-glutamine (1-^{13}C, 99%, #CLM-3612-PK), and alpha-ketoglutaric acid (^{13}C5, 99%, #CLM-2411-PK), yielding a final concentration of 12.5 µM for each standard. All steps of the extraction process were performed in a 4° C cold room with minimal light unless otherwise noted.

Filters were removed from the storage freezer, immediately unfolded, and placed cell side down in sterile plastic petri dishes containing 800 µl of the cocktail-amended solvent. 750 µl of non-amended solvent was then pipetted over the top facing side of the filter, and the petri dishes were placed in a -20ºC freezer for a twenty min extraction. The filters were then flipped over to expose the cell side, and the filter was rinsed using a pipette and the excess solvent contained in the dish. This was repeated a total of twenty times, at which point excess solvent was transferred to a microfuge tube and set aside. 500 µl of fresh, non-amended extraction solvent was added back to the filter contained in the petri dish, which was used to rinse the filter five times. Using tweezers, the filter was then blotted against the petri dish held at an angle to remove excess liquid from the filter. Excess solvent was then pipetted ten times over parts of the petri dish containing imprinted cells (*i.e.,* stuck). All solvent contained in the dish was then transferred to the microfuge tube containing the initial extraction, and the contents were centrifuged for 5

min at 13.3 xG. The supernatant was transferred to a fresh microfuge tube, while the pellet was resuspended in 200 µl of fresh, non-amended extraction solvent for a secondary, twenty minute extraction at 4°C. After this incubation, the pellet suspension was centrifuged for five min at 13.3 xG, and the supernatant was combined with the previously collected extract. Samples were extracted in batches of three, and held on ice in the dark for no longer than 6 hours. Batches of 24 samples were then dried using liquid nitrogen and stored until all samples were collected. All samples for this experiment were collected over a two day period, at which point the dried samples were resuspended in 300 µl and transferred to a UPLC vial. Samples were randomized for UPLC-MS analysis, and all samples were run in a single batch to minimize instrumental drift effects.

Metabolite UPLC-MS Analyses

Extracted metabolites were dried under nitrogen and resuspended in HPLC grade water (Fischer Scientific, Hampton, NH, USA) prior to mass analysis. Using an established untargeted metabolomics method (26) metabolites were analyzed with ultra-performance liquid chromatography coupled to high resolution mass spectrometry (UPLC-HRMS) (Thermo Scientific, San Jose, CA, USA). The water-soluble metabolites were separated using a Synergi Hydro RP column (100mm x 2.1mm, 2.6 µm, 100 Å) and an UltiMate 3000 pump (Thermo Fischer). All solvents used were HPLC grade (Fischer Scientific). An Exactive Plus Orbitrap MS (Thermo Fischer) was utilized for the full scan mass analysis. Using an open source software package, Metabolomic Analysis and Visualization Engine (MAVEN) (27), which is based on XCMS (28), metabolites were identified by chromatographic retention time and exact mass. Area under the curve was integrated and further statistical analyses performed. Selected metabolites (alpha-ketoglutarate, glutamate, and glutamine) were quantified using an isotopically labeled internal standard (Cambridge Isotope Laboratories, Inc., Tewksbury, MA, USA). Heatmaps and dotplots were generated from metabolite fold changes in the R language (29). Partial Least Squares-Discriminant Analysis (PLS-DA) (30) was performed to test for metabolite differences between groups. Multivariate analyses were conducted in PRIMER 7.0 using default settings unless otherwise noted.

Results

The Chlorella-PBCV1 infection cycle

The PBCV-1 infection cycle is known to last 6-8 h, at which point the cells lyse and release new viruses. We observed a steady decrease in chlorophyll-a fluorescence in infected cultures over this time course, a decrease in cell numbers, and a concomitant increase in viruses at the end of infection (Figure 1). In comparison, the non-infected cultures displayed a gradual increase in fluorescence and cell concentration that is congruent with normal growth. Despite some fluctuation occurring in the infected cultures, the overall infection dynamics reflect a typical infection profile of PBCV-1.

Changes in the metabolic profile due to infection

A total of 102 small, water soluble metabolites were identified in our samples and normalized by cell concentrations. An Analysis of Similarity (ANOSIM) calculation for the metabolic profile between treatments yields an R value of 0.512 (p-value <0.001). To visually quantify similarity, we used non-metric multidimensional scaling (nMDS) and clustering for samples grouped by treatment and time. This identified treatment-specific trends in the metabolic profile, wherein only the infected cultures show a change in measured metabolites that correlates with time (Figure 2). The non-infected control maintains 90% similarity in the metabolic profile across the six hour experiment, with the exception of four time points that all derive from the same flask (and a few outliers). Infected cultures, on the other hand, overlap with the controls at the beginning of infection, but become more distinct and cluster by similar times after 30 minutes post-infection. Nonetheless, all samples excluding the seeming outliers maintain a high similarity in the metabolic profile (>80%). A two-way SIMPER analysis between treatment and time identified seven metabolites that contribute to 70% of distance-based dissimilarity, though malate dominated the contribution (Table 1). Malate and succinate are central components of metabolism with involvement in major metabolic pathways (*i.e.* TCA cycle).

An alternative quantitative assessment of metabolite alterations using fold change thresholds (>1.5) and tests of significance (p-value <0.05) reveals rapid shifts in the metabolic profile. In the non-infected control, metabolite peak areas compared to T0 measurements show slightly decreased fold changes with less than a quarter of these bearing significance at the final time point (Figure 3A). Notable exceptions with more consistently different metabolite deviances include alpha-ketoglutamate, uridine, 2-isopropylmalate, and trehalose/sucrose. Virus-infected cultures, on the other hand, demonstrate pronounced fold differences compared to T0 measurements, with over a quarter of the metabolites exhibiting significant differences within seven min of infection (Figure 3C). Comparing metabolites across the treatments for each time point can amplify or weaken the signal (Figure 3 B), but the trends appear to reflect those generally observed in the viral treatment across time (Figure 3C; Figure 4). The specific dynamics among metabolites previously identified to drive distance-based similarity metrics (Figure 2) are also unveiled with this method. Malate, which was shown to primarily influence metabolic profile differences, does not change in the non-infected cultures but significantly increases within min of viral infection. The same general pattern occurs with succinate, another central metabolite in the TCA cycle, and all other drivers of viral-related metabolic change except for leucine and/or isoleucine. Indeed, amino acids as a group can be categorized across a spectrum from significantly enriched to significantly depleted, though there is rarely enough metabolites represented in this dataset to analyze all the intermediates of a particular amino acid synthesis or catabolic pathway. Despite this, it is clear that aromatic amino acids (phenylalanine, tyrosine, and tryptophan) and basic amino acids (histidine, arginine, and to a lesser extent lysine) exhibit enrichment patterns while amidic, sulfur-containing, and hydroxylic amino acids exhibit general depletion. The common acidic, cytosolic amino acids, aspartate and glutamate, exhibit little change across the infection profile. Nucleotides and their precursors/derivatives are generally depleted across time in host cultures, though the fold change is rarely large or significant. Indeed, uridine is the only nucleotide metabolite of consequence in non-infected cultures. Virocell intracellular nucleotide levels distinguish themselves, as they are nearly all significantly enriched, including their nucleoside RNA counterparts. Synthesis shows

preference for pyrimidine intermediates thymidine and uridine, as well as an enrichment in available intracellular adenosine. These molecules are enriched by seven min post-infection, and increase in availability over the course of infection. Stochasticity is observed in other pathway-related metabolites. However, it should be noted that glutathione, a metabolite associated with redox status of cells (31), is clearly enriched in virocells and drives metabolic differences.

To summarize these findings, 28.4% of the detected metabolites exhibit a significant fold change (>[1.5]), p-value <0.05) in virocells compared to controls within seven min of infection (Figure 4B). This percentage generally increases over the course of infection so that ~70% of detected metabolites are significantly different by the end. Significant metabolites are more often enriched than depleted, though both types of changes are observed during early stages of infection (prior to DNA synthesis).

Differences in metabolites targeted for quantification

Central metabolites glutamate, glutamine, and alpha-ketoglutarate were quantified using heavy-labeled internal standards. The ratios of these metabolites can help to indicate the fate of glutamate, an acidic, cytosolic amino acid that has many metabolic fates (Figure 5). First, it is an amino acid that serves as a basic building block for other amino acids *via* transamination, leading to the formation of alpha-ketoglutarate. However, glutamate can also form alpha-ketoglutarate in a separate reaction involving glutamate dehydrogenase. In either case, this end product can be cycled back into glutamate, or it can enter the TCA cycle to make energy. One of the amino acids formed from glutamate is glutamine, an amidic amino acid named because its R group contains an amide. The ratio of glutamate to glutamine is used to monitor nitrogen limitation, within an increased ratio towards glutamate indicating growth under N limitation (23). Finally, glutamate can also combine with cysteine and glycine to form a tripeptide, glutathione, that has well characterized anti-oxidizing properties. Glutathione can again be converted back into glutamate via 5-oxoprolinase and gamma-glutamyl transferase. Altogether, that we can look at rates between precursors and derivatives of glutamate might indicate something

about N limitation, TCA cycle activity, and/or redox conditions. That said, glutamate to glutamine ratios did not significantly change over the experiment in either control (Figure 6A). There was a slight decrease in the ratio of glutamate to alpha-ketoglutarate, though the fold change is barely significant (Figure 6B). Cell concentrations of three labeled metabolites indeed indicate that only glutamate decreases, while alpha-ketoglutarate and glutamine levels are consistent across infection (Figure 6C). Although we did not use a heavy-labeled standard for glutathione, peak area can be used to determine ratio of metabolites within a sample, with changes in that ratio being comparable regardless of normalization. Glutamate to glutathione ratios were detected to decrease within 30 min of infection, and continued to drop by nearly two orders of magnitude by the end of infection (Figure 7A). We thus decided to compare peak area ratios between reduced glutathione (GSH) and its oxidized form, glutathione disulfide (GSSG), to determine redox cycling patterns (Figure 7B). This indicated a ratio increase towards the oxidized form starting at 30 min post-infection, with greater significance at one hour (Figure 7C).

Discussion

Our approach targets small (<1000 Daltons), water soluble metabolites, and the number of metabolites detected is representative of other studies using this methodology (7). Based on this data, host culture metabolism maintains 90% similarity across the course of the experiment, with the exception of a few outliers. The four control datapoints that cluster together and away from the rest of the non-infected data all derive from the same flask, including a T0, T7, T240, and T360 sample. It is thus likely that sampling and/or processing account for this deviance as opposed to a biological reason. Virocell metabolic profiles also maintain high similarity (>80%) over the course of infection, though there is clear separation across the length of the PBCV-1 infection cycle. Because of the low number of metabolites detected (considering there are likely orders of magnitude more in the cell), dissimilarity is explained by only a handful of metabolites (~10). We focus on those contributing to 70% of deviance, which yielded a mixture of seven metabolites that are involved in diverse metabolic pathways. Most of distance-based

differences were explained by malate, a major component of the TCA cycle. Other intermediates in the TCA cycle were detected in this study, including succinate, alpha-ketoglutarate, and fumarate, and all showed a general increase in their intracellular presence. This might indicate that the virocell is generating more energy than that found in a normal host culture to accelerate viral replication. Other metabolites identified as contributors of dissimilarity can be grouped together by oxidative stress management (glutathione, trehalose/sucrose), but others do not (glycodeoxycholate and isoleucine/leucine). Not much is known about glycodeoxycholate function in algae, though in mammals it is regarded as a bile salt that solubilizes fats for absorption and is itself absorbed. Since both *Chlorella* and PBCV-1 have lipids, it is possible that the virus is using this to recycle host lipids for its virion structure. Isoleucine/leucine are aliphatic amino acids, and although SIMPER analysis revealed they were driving differences, fold-change comparisons indicate this was not as large as other metabolites.

Distance-based similarity is not comparable to fold changes of significance, wherein nearly 70% of detected metabolites surpass the deviance threshold (>1.5 fold; p-value <0.05) compared to non-infected controls. Group-based analyses do however begin to elucidate trends in viral-driven metabolism. First, it is clear that changes in amino acids are explained often by having similar biochemistries. Aromatic and basic amino acids are specifically enriched, whereas sulfur-containing, amidic, and hydroxylic amino acids are depleted. Whether this is the consequence of effects on shared synthesis pathways, degradation kinetics, or localized effects of infection is not clear. For example, aromatic enrichment could be explained by the fact that these amino acids all derive from the shikimate synthesis pathway. However, transcriptional analyses in PBCV-1 infected *Chlorella* hosts reveal that the shikimate pathway, which is responsible for aromatic amino acid synthesis, is completely shut down within one hour of infection (32). It was originally hypothesized that this was a host-defense strategy to reduce amino acid availability to the virus, but our results indicate that the aromatics are more enriched than they were in non-infected cultures. Thus, it is more likely that enrichment is the result of as build-up of thee metabolites present at the start of infection or degradation of proteins. Intracellular

partitioning of amino acids has been demonstrated in microbes and plants, wherein aromatic and basic amino acid have been shown to accumulate in vacuoles whereas acidic amino acids exist almost exclusively in the cytosol (33). If this occurs in *Chlorella,* it is possible that plastidular transport might be compromised in virocells, making the stock of aromatics that are stored in these vesicles look like enrichment when compared to the non-infected culture (Figure 3B). On the other hand, investigation of just the infected culture aromatic content across time reveals there is some enrichment, indicating degradation is going on too (Figure 3C). In any case, the shutdown of the shikimate pathway and potential stagnation of aromatic synthesis represents a problem for the cell. Indeed, plants direct 20-30% of photosynthetically fixed carbon to the production of phenylalanine and its derivatives, which constitute up to 45% of plant biomass, impact growth, development, and reproduction, and are precursors for defense compounds (*i.e.* salicylic acid, tannins, and flavonoids) (36). Thus, some of these might be useful to the virocell and require recycling of aromatics from other sources. Finally, there is also research demonstrating that aromatic amino acids yield protective effects against abiotic stresses. Tyrosine, tryptophan, and to a lesser extent phenylalanine, all absorb ultraviolet light. An accumulation of these amino acids might reduce UV penetration and damage to the nucleic acid. Aromatics are also often enriched in transmembrane proteins putatively because they are enriched in regions containing the highest lipid density to prevent lethal, chain-reactive oxidation of lipids. Altogether, there are many exciting mechanisms that might be driven by aromatic amino acid enrichment in virocells. Depleted amino acids, on the other hand, might be a higher requirement for virus protein production and/or nutritional requirements.

Another trend in virocell metabolism is the increase in intracellular nucleotides and their derivatives. It is expected that these would all generally decrease, because the PBCV-1 virus is known to utilize both viral-encoded general nucleases and restriction-modification (RM) systems to selectively digest and recycle the host genome (37). While this supports viral replication, there is nevertheless a putatively significant challenge presented for the virus to achieve production of its progeny. Specifically, the PBCV-1

genome must produce up to a thousand copies of a 331-kbp, AT rich (60%) genome from a 46.2-Mbp host genome that is GC rich (67.1%) (21). Although the host genome is much larger, estimates of the number of adenines and thymines required to make 1000 viral progeny exceed host genome availability by 12.6x. Although chloroplast genomes and mitochondrial genomes might alleviate this (38), only the chloroplast has been shown to be digested (39) and its genome does not provide enough nucleotides to make up the difference (124-kbp, 33.9% GC). That said, it is interesting that our data shows that intracellular adenosine and thymidine, the nucleoside precursors of adenines and thymines, are significantly enriched before the start of replication and throughout the rest of the infection cycle. Synthesis of thymine might be achieved via the cycling of other nucleotides, such as cytosine, uracil, and its intermediates. Many of these intermediates are upregulated, though cytosine and uracil itself were not detected in our dataset (though they are obviously present). Another nucleotide that may contribute to this, though it was not detected here, is 5-methylcytosine. The *Chlorella variabilis* genome is methylated with this nucleotide in CNG and CNN sequences (32), and this nucleotide can be converted into thymine via a one-step deaminase reaction. Adenine might also be produced from purine metabolism of guanine and its derivatives, though complete pathway analyses are also limited. Pathway-specific standards might be used to better define the cycling of purines and pyrimidines to promote replication of GC deviant viruses.

Viruses are primarily composed of nucleic acids and proteins, implying that the stoichiometry of virions is distinct from hosts and might accelerate rapid limitation of nitrogen and amino acids during infection (22). This is especially important considering virions have less stoichiometric plasticity, meaning host cellular stoichiometry has a direct impact on virus production. This has been observed in *Chlorella* and PBCV-1; the virus has a calculated 17:5:1 requirement of C/N/P and viral production will change upon nutrient availability instead of altering this elemental ratio. Indeed, viral progeny numbers decrease under phosphorous-limiting conditions(40) and when the host is grown on nitrogen species that are more difficult to assimilate (*i.e.* nitrates and nitrites) (41). That said, we expected to observe nitrogen limitation over the course of infection but did not.

This is likely due to our media choice; modified Bolds Basal Medium (24) is a nutrient rich medium containing 3mM KNO_3 and a supplement of bactopeptone. Past research has shown that the bactopeptone is a more bioavailable source of nitrogen to *Chlorella* (42), but we expected it to be depleted by time of infection. We expected this because the cultures were in mid-log phase in batch cultures. On top of that, virus-induced host membrane depolarization decreases the efficiency of secondary active transporters, including those involved in amino acid transport (19), making exogenous nutrient sources less available. For a study more interested on the effects of nitrogen and carbon availability, it would be useful to redo this experiment while either monitoring exogenous nitrogen concentrations or using a truly nitrogen-limited medium. We chose to use standard culturing medium because all other omics characterization studies of PBCV-1 have used this medium.

While glutamate : glutamine ratios did not indicate nitrogen limitation, we did see a decrease in glutamate indicating it was being utilized or not replished as quickly by the virocell. This was complimented by an increase in glutathione, another derivative of glutamate metabolism. Moreover, the ratio of glutathione to its oxidized form, glutathione disulfide, increased over the course of infection by up to two orders of magnitude, reaching a peak-area ratio of around 1. This was puzzling at first, as oxidative stress studies in human oncogenics indicate that the ratio should decrease with increased oxidative stress (43). However, the enzyme glutathione reductase, which regenerates free glutathione from its oxidized form, is continually produced in the presence of oxidative stress. Thus, it is possible that this is a continued, heightened stress in the virocell that is never mitigated and thus requires glutathione to be continuously recycled. Heightened protein activity has been observed in other chlorovirus encoded proteins, including topoisomerase, which processes DNA at a much faster rate (44). In any case, there is possibly a unique relationship between redox conditions of the virocell and the ability of the virus to complete replication. Indeed, pre-treatment of *Emiliania huxleyi* cultures with glutathione prevents viral replication (45). In terms of a virocell metabolic marker, it would be interesting to compare this glutathione redox state ratios in different nutrient conditions

to determine how other stresses might cause toxic reactive oxygen species, and whether the signals can be stress-delineated.

Chlorovirus PBCV-1 clearly drives the metabolic activities of its host during infection. These studies establish metabolic changes within seven min of infection, which is congruent with detection of viral transcripts (13). That said, there were some issues with flow cytometry cell counts of the infected treatment due to instrument defects discovered after processing. To circumnavigate this, we are in the process of extracting nucleic acid from cell pellets collected and frozen during the experiment to perform qPCR counts of both the host and virus. We expect this to remove some of the noise displayed in this data, and to perhaps delineate more consistent trends across the infection cycle. In any case, this data will be a useful contribution to understanding algae host-virus interactions, especially within this particular system.

References

1. Suttle CA. Marine viruses - major players in the global ecosystem. Nature Reviews Microbiology. 2007;5(10):801-12.

2. Wilhelm SW, Suttle CA. Viruses and Nutrient Cycles in the Sea - Viruses play critical roles in the structure and function of aquatic food webs. Bioscience. 1999;49(10):781-8.

3. Forterre P. The virocell concept and environmental microbiology. Isme Journal. 2013;7(2):233-6.

4. Forterre P. Manipulation of cellular syntheses and the nature of viruses: The virocell concept. Comptes Rendus Chimie. 2011;14(4):392-9.

5. Rosenwasser S, Ziv C, Van Creveld SG, Vardi A. Virocell Metabolism: Metabolic Innovations During Host-Virus Interactions in the Ocean. Trends in Microbiology. 2016;24(10):821-32.

6. Breitbart M. Marine Viruses: Truth or Dare. In: Carlson CA, Giovannoni SJ, editors. Annual Review of Marine Science, Vol 4. Annual Review of Marine Science. 42012. p. 425-48.

7. Ankrah NYD, May AL, Middleton JL, Jones DR, Hadden MK, Gooding JR, et al. Phage infection of an environmentally relevant marine bacterium alters host metabolism and lysate composition. Isme Journal. 2014;8(5):1089-100.

8. Schleyer G, Shahaf N, Ziv C, Dong YH, Meoded RA, Helfrich EJN, et al. In plaque-mass spectrometry imaging of a bloom-forming alga during viral infection reveals a metabolic shift towards odd-chain fatty acid lipids. Nature Microbiology. 2019;4(3):527-38.

9. Holligan PM, Viollier M, Harbour DS, Camus P, Champagnephilippe M. Satellite and ship studies of coccolithophore production along a continental-shelf edge. Nature. 1983;304(5924):339-42.

10. Coy SR, Gann ER, Pound HL, Short SM, Wilhelm SW. Viruses of Eukaryotic Algae: Diversity, Methods for Detection, and Future Directions. Viruses-Basel. 2018;10(9).

11. Dunigan DD, Cerny RL, Bauman AT, Roach JC, Lane LC, Agarkova IV, et al. Paramecium bursaria Chlorella Virus 1 Proteome Reveals Novel Architectural and Regulatory Features of a Giant Virus. Journal of Virology. 2012;86(16):8821-34.

12. Van Etten JL, Lane LC, Meints RH. Viruses and virus-like particles of eukaryotic algae. Microbiological Reviews. 1991;55(4):586-620.

13. Blanc G, Mozar M, Agarkova IV, Gurnon JR, Yanai-Balser G, Rowe JM, et al. Deep RNA Sequencing Reveals Hidden Features and Dynamics of Early Gene Transcription in Paramecium bursaria Chlorella Virus 1. Plos One. 2014;9(3).

14. Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova IV, et al. Microarray Analysis of Paramecium bursaria Chlorella Virus 1 Transcription. Journal of Virology. 2010;84(1):532-42.

15. Milrot E, Shimoni E, Dadosh T, Rechav K, Unger T, Van Etten JL, et al. Structural studies demonstrating a bacteriophage-like replication cycle of the eukaryote-infecting Paramecium bursaria chlorella virus-1. Plos Pathogens. 2017;13(8).

16. Frohns F, Kasmann A, Kramer D, Schafer B, Mehmel M, Kang M, et al. Potassium ion channels of chlorella viruses cause rapid depolarization of host cells during infection. Journal of Virology. 2006;80(5):2437-44.

17. Neupartl M, Meyer C, Woll I, Frohns F, Kang M, Van Etten JL, et al. Chlorella viruses evoke a rapid release of K+ from host cells during the early phase of infection. Virology. 2008;372(2):340-8.

18. Greiner T, Frohns F, Kang M, Van Etten JL, Kasmann A, Moroni A, et al. Chlorella viruses prevent multiple infections by depolarizing the host membrane. Journal of General Virology. 2009;90:2033-9.

19. Agarkova I, Dunigan D, Gurnon J, Greiner T, Barres J, Thiel G, et al. Chlorovirus-Mediated Membrane Depolarization of Chlorella Alters Secondary Active Transport of Solutes. Journal of Virology. 2008;82(24):12181-90.

20. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, et al. The Chlorella variabilis NC64A Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex. Plant Cell. 2010;22(9):2943-55.

21. Wilhelm SW, Bird JT, Bonifer KS, Calfee BC, Chen T, Coy SR, et al. A Student's Guide to Giant Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. Viruses-Basel. 2017;9(3).

22. Jover LF, Effler TC, Buchan A, Wilhelm SW, Weitz JS. The elemental composition of virus particles: implications for marine biogeochemical cycles. Nature Reviews Microbiology. 2014;12(7):519-28.

23. Flynn KJ, Dickson DMJ, Alamoudi OA. The ratio of glutamine : glutamate in microalgae - a biomarker for n-status suitable for use at natural cell densities. Journal of Plankton Research. 1989;11(1):165-70.

24. Dunigan D, Agarkova I. Formulation of MBBM (Modified Bold's Basal Medium) 2016 [

25. Rabinowitz JD, Kimball E. Acidic acetonitrile for cellular metabolome extraction from Escherichia coli. Analytical Chemistry. 2007;79(16):6167-73.

26. Lu WY, Clasquin MF, Melamud E, Amador-Noguez D, Caudy AA, Rabinowitz JD. Metabolomic Analysis via Reversed-Phase Ion-Pairing Liquid Chromatography Coupled to a Stand Alone Orbitrap Mass Spectrometer. Analytical Chemistry. 2010;82(8):3212-21.

27. Melamud E, Vastag L, Rabinowitz JD. Metabolomic Analysis and Visualization Engine for LC-MS Data. Analytical Chemistry. 2010;82(23):9818-26.

28. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. Analytical Chemistry. 2006;78(3):779-87.

29. Team RC. R: A language and enviroment for statistical computing. In: Computing RFfS, editor. Vienna, Austria2018.

30. Sanchez G. DiscriMiner: Tools of the Trade for Discriminat Analysis. 2013.

31. Buchanan BB, Balmer Y. Redox regulation: A broadening horizon. Annual Review of Plant Biology. 2005;56:187-220.

32. Rowe JM, Jeanniard A, Gurnon JR, Xia Y, Dunigan DD, Van Etten JL, et al. Global Analysis of Chlorella variabilis NC64A mRNA Profiles during the Early Phase of Paramecium bursaria Chlorella Virus-1 Infection. Plos One. 2014;9(3).

33. Sekito T, Fujiki Y, Ohsumi Y, Kakinuma Y. Novel families of vacuolar amino acid transporters. Iubmb Life. 2008;60(8):519-25.

34. Seaton GGR, Lee K, Rohozinski J. Photosynthetic shutdown in chlorella nc64a associated with the infection cycle of paramecium-bursaria chlorella virus 1. Plant Physiology. 1995;108(4):1431-8.

35. Scheer H. Chlorophyll breakdown in aquatic ecosystems. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(43):17311-2.

36. Maeda H, Dudareva N. The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants. In: Merchant SS, editor. Annual Review of Plant Biology, Vol 63. Annual Review of Plant Biology. 632012. p. 73-105.

37. Agarkova IV, Dunigan DD, Van Etten JL. Virion-associated restriction endonucleases of chloroviruses. Journal of Virology. 2006;80(16):8114-23.

38. Orsini M, Costelli C, Malavasi V, Cusano R, Concas A, Angius A, et al. Complete sequence and characterization of mitochondrial and chloroplast genome of Chlorella variabilis NC64A. Mitochondrial DNA Part A. 2016;27(5):3128-30.

39. Van Etten JL, Burbank DE, Joshi J, Meints RH. DNA synthesis in a Chlorella-like algal folloiwng infection with the virus PBCV-1. Virology. 1984;134(2):443-9.

40. Clasen JL, Elser JJ. The effect of host Chlorella NC64A carbon : phosphorus ratio on the production of Paramecium bursaria Chlorella Virus-1. Freshwater Biology. 2007;52(1):112-22.

41. Cheng YS, Labavitch J, VanderGheynst JS. Organic and Inorganic Nitrogen Impact Chlorella variabilis Productivity and Host Quality for Viral Production and Cell Lysis. Applied Biochemistry and Biotechnology. 2015;176(2):467-79.

42. Kamako S, Hoshina R, Ueno S, Imamura N. Establishment of axenic endosymbiotic strains of Japanese Paramecium bursaria and the utilization of carbohydrate and nitrogen compounds by the isolated algae. European Journal of Protistology. 2005;41(3):193-202.

43. Zitka O, Skalickova S, Gumulec J, Masarik M, Adam V, Hubalek J, et al. Redox status expressed as GSH:GSSG ratio as a marker for oxidative stress in paediatric tumour patients. Oncology Letters. 2012;4(6):1247-53.

44. Fortune JM, Lavrukhin OV, Gurnon JR, Van Etten JL, Lloyd RS, Osheroff N. Topoisomerase II from Chlorella virus PBCV-1 has an exceptionally high DNA cleavage activity. Journal of Biological Chemistry. 2001;276(26):24401-8.

45. Sheyn U, Rosenwasser S, Ben-Dor S, Porat Z, Vardi A. Modulation of host ROS metabolism is essential for viral infection of a bloom-forming coccolithophore in the ocean. Isme Journal. 2016;10(7):1742-54.

Appendix

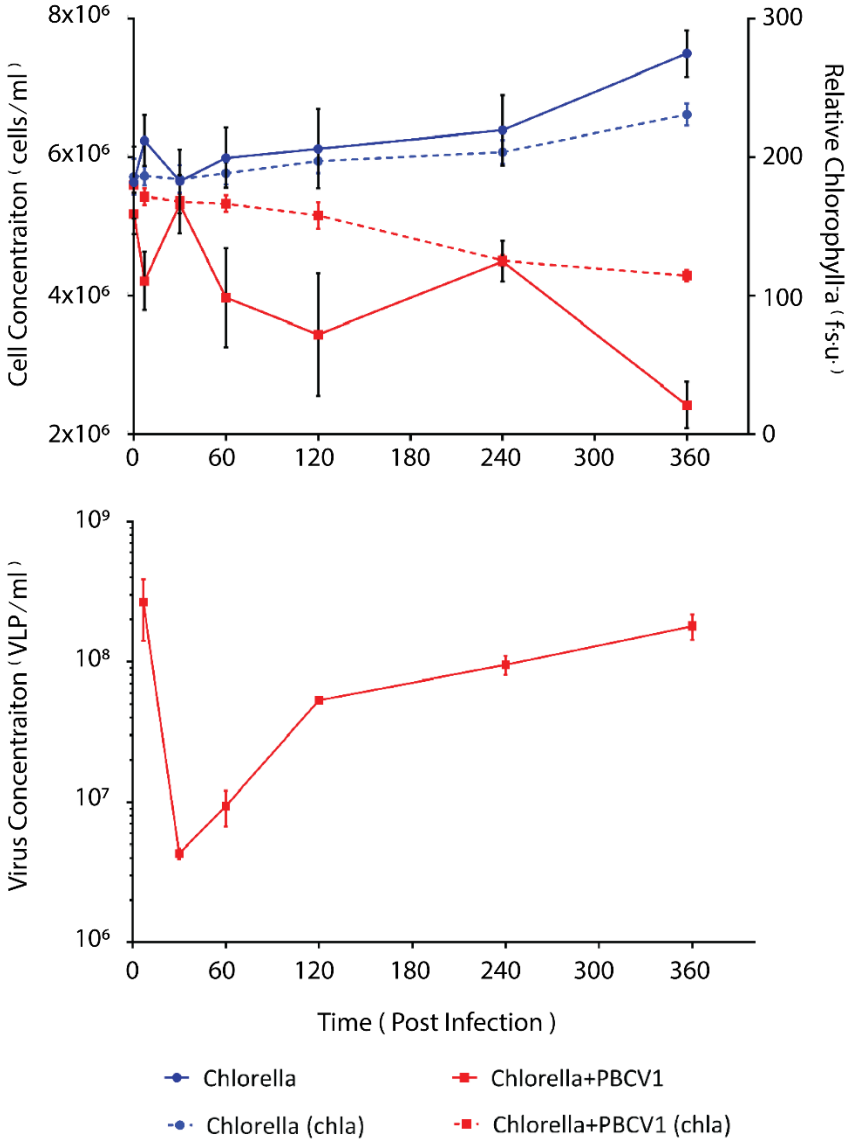


Figure 4.1 Growth dynamics of Chlorella and Chlorella virocells

Culture cell counts (solid lines) and chlorophyll fluorescence (dotted lines) across the experiment for non-infected (blue circles) and infected (red squares) treatments (A). Free virus like particles were also counted in the infected cultures over the six-hour experiment.

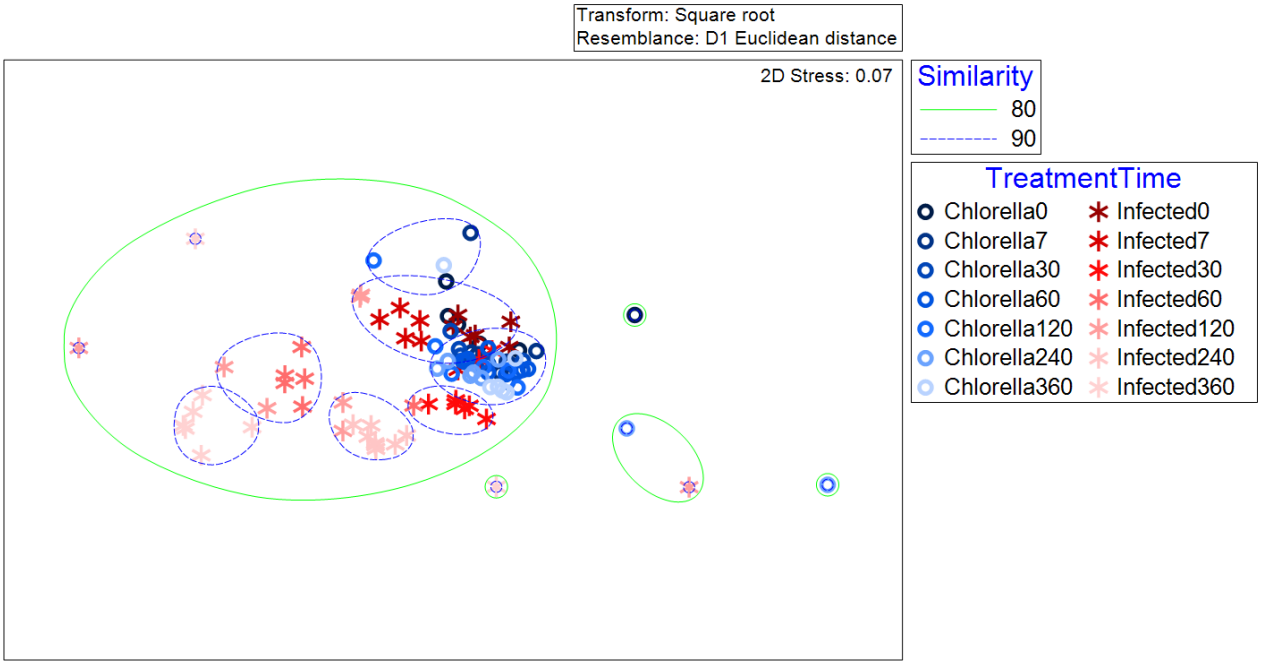


Figure 4.2 Spatial similarity of samples by treatment and time

Distances were calculated using non-metric, multi-dimensional scaling (nMDS) combined with clustering analyses (solid green and blue dotted circles). The samples were square root transformed, and mapped using a Euclidean distance resemblance. Non-infected treatments are defined in blue as Chlorella, while infected cultures are defined in red. The number linked to the treatment indicates the time post-infection that the samples were collected at.

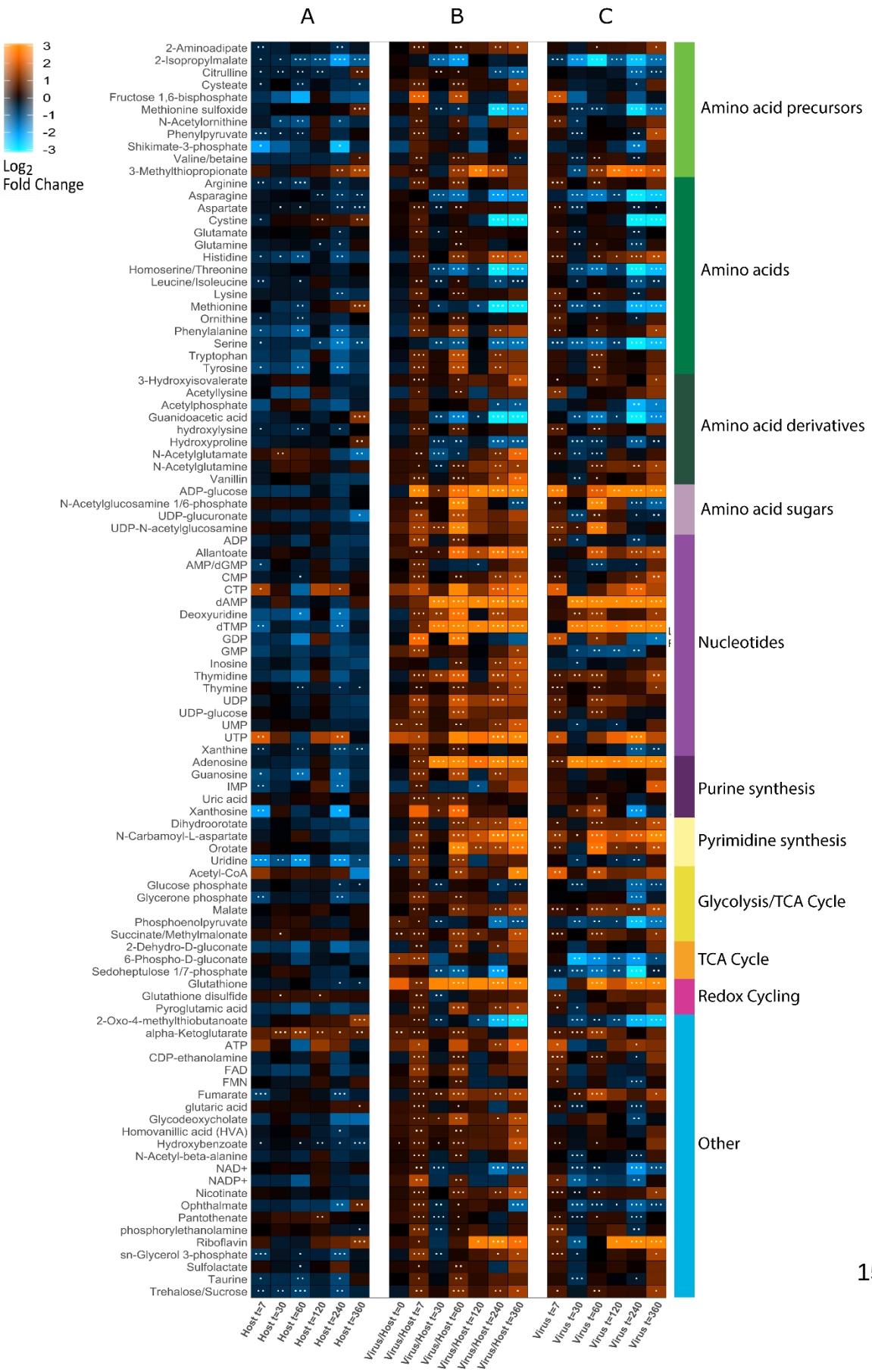
Table 4.1 Metabolites driving metabolic profile dissimilarity between treatments

Metabolite	Host μ	Infected μ	Sq Dist μ	Sq Dist σ	%CV	Contrib%
Malate	15.7	20.9	52.1	0.77	1.48	46.21
Pyroglutamic Acid	4.77	6.41	8.21	0.63	7.67	7.28
Trehalose/Sucrose	4.41	5.86	5.37	0.6	11.2	4.76
Leucine/Isoleucine	7.14	6.61	4.58	0.57	12.4	4.06
Succinate/Methylmalonate	4.55	6.12	4.27	0.92	21.5	3.78
Glutathione	0.298	1.46	3	0.84	28	2.66
Glycodeoxycholate	2.52	3.55	2.73	0.63	23.1	2.42

Host and Infected mean (μ) peak-area values for metabolites are shown, with mean square distance (μ), standard deviation (σ), and coefficient of variation (%CV) shown to describe variability. Contrib% indicates the percent contributing to the distances calculated between samples and thus represent drivers of unique virocell metabolism. Only the top 70% of contributing metabolites are shown here.

Figure 4.3 Detected metabolites and their fold change

Changes in the non-infected control were observed across time compared to T0 measurements (A), as well as in the infected culture across time compared to T0 measurements (C). Metabolite fold change in the infected culture compared to non-infected culture was also done at each time point (B). Bars to the right group metabolites by functional KEGG pathway-related classifications, whereas metabolites grouped into more than one pathway were often placed in the ‘other’ group. Significance is marked with asterisks.


A
B
C
3
2
1
0
-1
-2
-3
Log2 Fold Change
2-Aminoadipate
2-Isopropylmalate
Citrulline
Cysteate
Fructose 1,6-bisphosphate
Methionine sulfoxide
N-Acetylornithine
Phenylpyruvate
Shikimate-3-phosphate
Valine/betaine
3-Methylthiopropionate
Arginine
Asparagine
Aspartate
Cystine
Glutamate
Glutamine
Histidine
Homoserine/Threonine
Leucine/Isoleucine
Lysine
Methionine
Ornithine
Phenylalanine
Serine
Tryptophan
Tyrosine
3-Hydroxyisovalerate
Acetyllysine
Acetylphosphate
Guanidoacetic acid
hydroxylysine
Hydroxyproline
N-Acetylglutamate
N-Acetylglutamine
Vanillin
ADP-glucose
N-Acetylglucosamine 1/6-phosphate
UDP-glucuronate
UDP-N-acetylglucosamine
ADP
Allantoate
AMP/dGMP
CMP
CTP
dAMP
Deoxyuridine
dTMP
GDP
GMP
Inosine
Thymidine
Thymine
UDP
UDP-glucose
UMP
UTP
Xanthine
Adenosine
Guanosine
IMP
Uric acid
Xanthosine
Dihydroorotate
N-Carbamoyl-L-aspartate
Orotate
Uridine
Acetyl-CoA
Glucose phosphate
Glycerone phosphate
Malate
Phosphoenolpyruvate
Succinate/Methylmalonate
2-Dehydro-D-gluconate
6-Phospho-D-gluconate
Sedoheptulose 1/7-phosphate
Glutathione
Glutathione disulfide
Pyroglutamic acid
2-Oxo-4-methylthiobutanoate
alpha-Ketoglutarate
ATP
CDP-ethanolamine
FAD
FMN
Fumarate
glutaric acid
Glycodeoxycholate
Homovanillic acid (HVA)
Hydroxybenzoate
N-Acetyl-beta-alanine
NAD+
NADP+
Nicotinate
Ophthalmate
Pantothenate
phosphorylethanolamine
Riboflavin
sn-Glycerol 3-phosphate
Sulfolactate
Taurine
Trehalose/Sucrose
Host t=7
Host t=30
Host t=60
Host t=120
Host t=240
Host t=360
Virus/Host t=0
Virus/Host t=7
Virus/Host t=30
Virus/Host t=60
Virus/Host t=120
Virus/Host t=240
Virus/Host t=360
Virus t=7
Virus t=30
Virus t=60
Virus t=120
Virus t=240
Virus t=360
Amino acid precursors
Amino acids
Amino acid derivatives
Amino acid sugars
Nucleotides
Purine synthesis
Pyrimidine synthesis
Glycolysis/TCA Cycle
TCA Cycle
Redox Cycling
Other

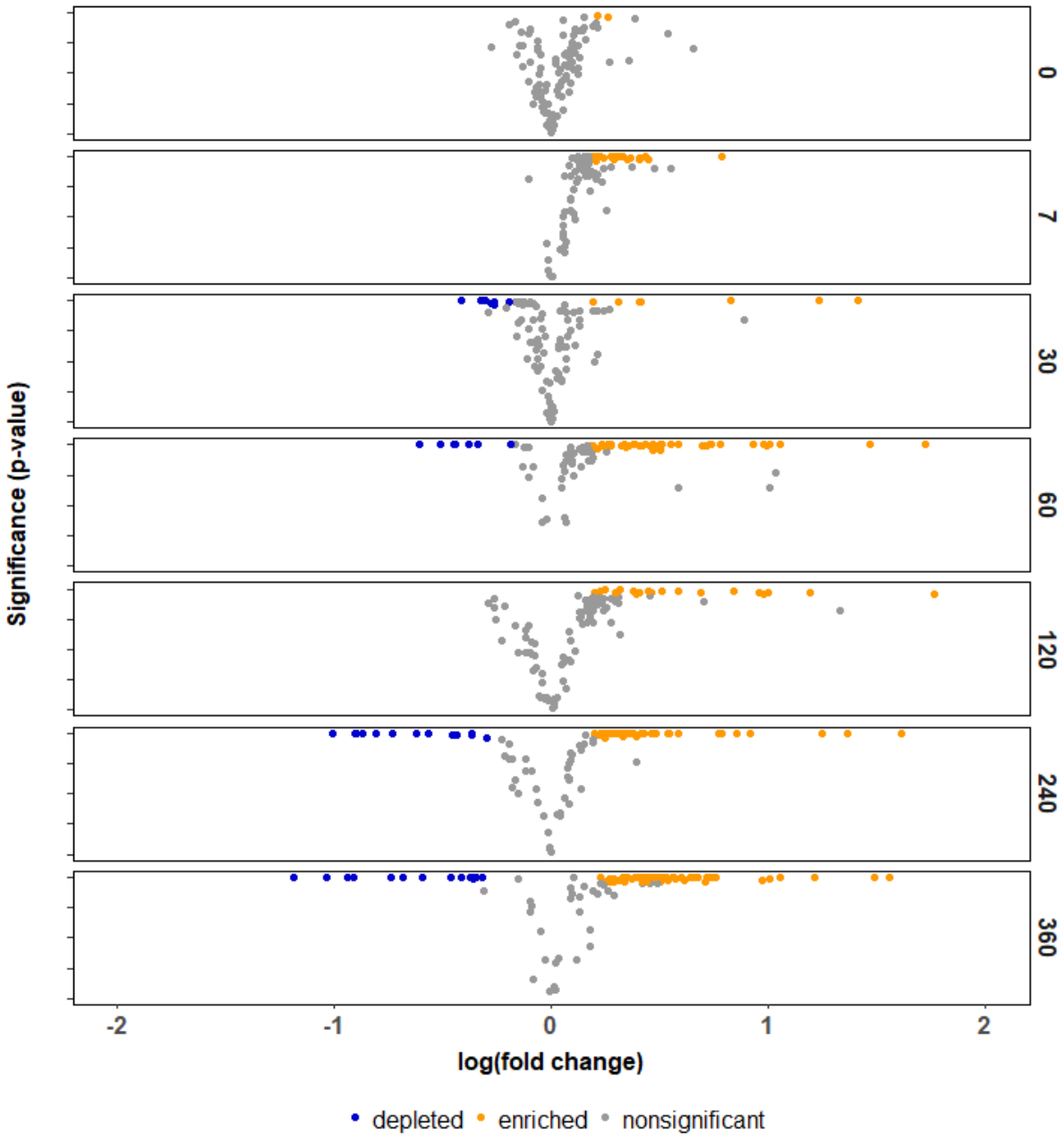


Figure 4.4 Metabolic profile shifts in Chlorella virocells

Metabolites marked as enriched (orange) or depleted (blue) in infected cultures compared to non-infected cultures, faceted by time post-infection. The threshold for defining this used a fold-change of 1.5 (before log transformation) and a p-value significance of <0.05.

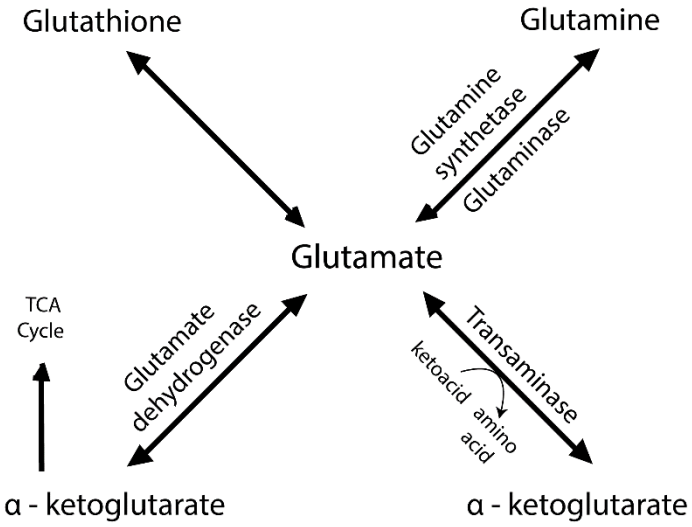


Figure 4.5 Metabolic fates of glutamate

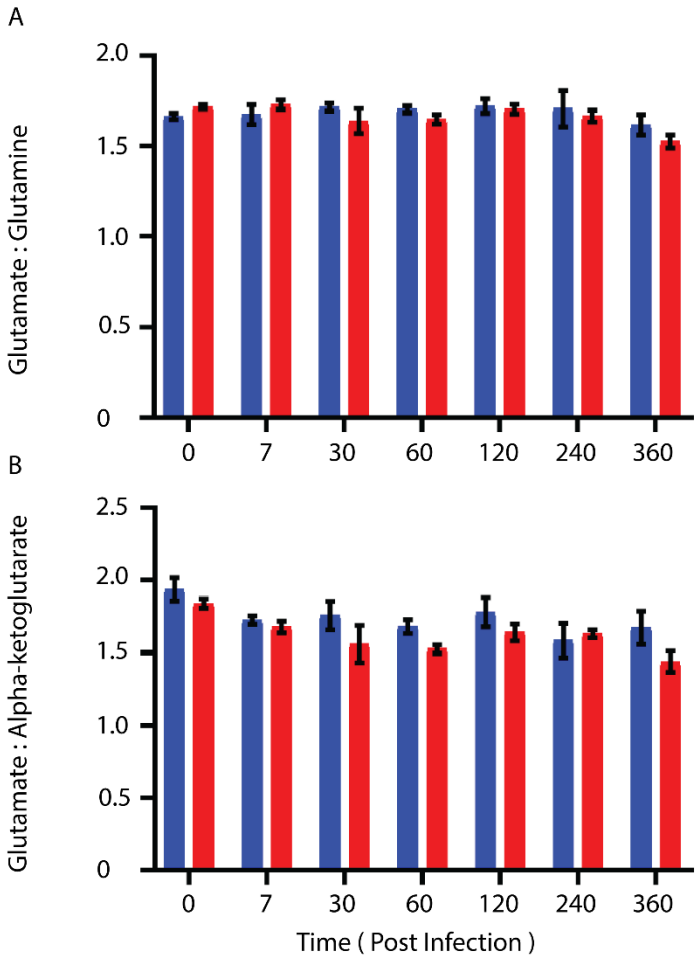


Figure 4.6 Molar ratio of stable-isotope labeled central metabolites

The blue bars represent non-infected cultures and the red bars represent infected cultures.

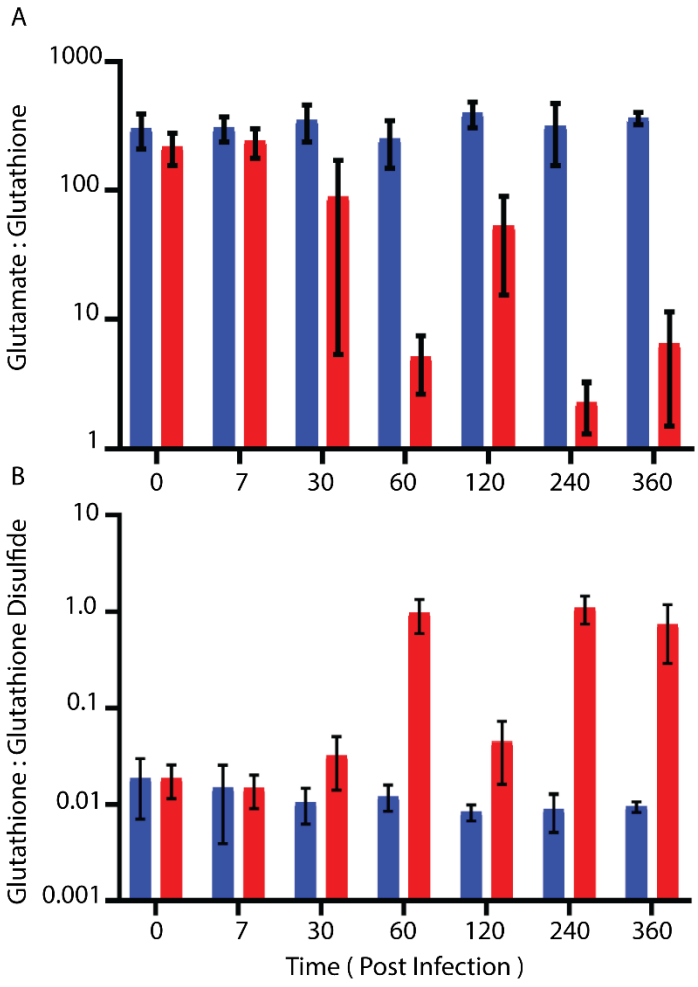


Figure 4.7 Peak area of metabolites involved in redox cycling

The blue bars represent non-infected cultures and the red bars represent infected cultures.

CHAPTER 5 : CONCLUSIONS

Viruses are important members of aquatic food webs that contribute significantly to host mortality, leading to cascading effects on nutrient cycling, host evolution, and community biodiversity. That said, most of the focus on viruses has been specifically geared towards bacteriophage. Viruses of eukaryotic phototrophs are gaining recognition as important ecological members as their hosts can proliferate to form massive blooms, destroy pristine habitats, and serve as important biogeochemically important taxa (*i.e.* diatoms and dinoflagellates). Understanding the interactions of viruses is important for predicting their ecology, though there is much unknown considering these viruses are i) generally not amenable to virus genetic manipulation, ii) over half of their genomes, which contains hundreds of genes, are hypothetical proteins with no known function, and iii) there are only ~65 viruses in culture compared to algal host species that are estimated to be on the order of hundreds of thousands. Thus, developing model systems to gain insight into the life cycle and success of these viruses is a useful starting point. For this dissertation, we performed model characterization studies in *Paramecium bursaria chlorella virus 1* (PBCV-1) and its green, unicellular host, *Chlorella variabilis* (NC64A). We developed a cryopreservation protocol to properly maintain viral strains, which has previously been a challenge to research labs and culture distributors alike. Additionally, we performed the first baseline assessments of both the 'epigenome' and metabolome of the virion and virocell, respectively.

Before we could make any observations about the epigenome or metabolome, we had to make sure that the viral genome could be controlled against both genetic and epigenetic alterations. This has traditionally been accomplished using cryopreservation, though studies have focused on cellular organisms as opposed to viruses. We show that the host can be used as a vehicle for virus cryopreservation, so that the virus is not exposed to putatively toxic effects of cryoprotectants and or freeze stress. Although cryopreservation is a common tool in bacteria and other cell culturing labs, we do recommend that someone check for freeze derived alterations in the epigenetic profile. This could be done by making batch measurements of the nucleosides before and after cryopreservation to determine if m5C and m6A levels are consistent.

Once we had a cryopreservation method in store, we decided to freeze down PBCV-1 as a source stock that we could call ‘wild type’. To better define this ‘wild type’ we decided to perform methylation and metabolomic analyzed using standard culturing conditions for comparison to other lab group studies. Our PacBio work revealed that the virus had dynamic methylation patterns, which is intriguing given these adenines are also targeted for restriction in their non-methylated form. We also found an enrichment of nucleotides bearing a modified signal, more so than can be accounted for by the known methyltransferases encoded by PBCV-1. Whether these are true modifications, artifacts from interpreting viral sequence modifications with non-homologous cellular sequence data, or some other reason is unknown. There is a benefit in monitoring genome modification changes across generations of virus, which might unveil something about virus burst size or activity. We also analyzed the metabolic profile and found unique, amino acid changes were shared among biochemically related groups. For future direction, there is a special opportunity to determine whether the disruption of the shikimate pathway, responsible for aromatic amino acid synthesis and the increased intracellular content, is a viral or host mediated function to promote or arrest infection, respectively. There is also opportunity to combine this data with that generated by other groups (*i.e.* transcriptomes) interested in the GC discrepancy between viruses and hosts. Finally, recent studies have demonstrated that oxidative stress is a double edged sword in virocells wherein some ROS is needed to promote viral infection, yet, too much can be lethal to the cell. Our study showed a clear enrichment of glutathione, a marker for redox cycling and oxidative stress. There have also been investigations into sucrose/trehaolse and other intermediates in the TCA cycle as antioxidants or mitigators of osmotic stress. Defining more specific mitigators of oxidative stress for this system would be an important step towards understanding how oxidative stress might damage or inactivate viruses during replication.

VITA

Samantha Rose Coy was born in Kansas City, MO (USA). She attended Kickapoo High School where she graduated in 2009. After graduation, Samantha began her undergraduate academic career at Drury University where she pursued a Bachelor of Arts, majoring in both Environmental Science and Biology. Living in the Midwest with a desire to do marine work, Samantha learned to find ways to intersect her hobbies with what eventually developed into academic and research interests. She was an avid SCUBA diver, and she joined a regional cave diving team to increase her skills as well as gain access to limited dive sites. Her unique hobbies and excitement in her university courses gained the attention of one of her professors, Dr. Teresa Carroll, whom she began a research relationship with. Samantha connected the cave diving team with Dr. Carroll to give her access to new research sites, as the caves in Missouri are deep (100-300 feet) underwater cave systems feeding surface springs. Samantha took a study abroad course with Dr. Carroll to Roatan Honduras where she made baseline quantitative estimates of the incidence and distribution of coral disease and corallivore activity. During her studies, she learned about the unique microbiology that drives both the health and disease of the coral animal, and decided she wanted to continue learning about this work. She was selected to do a summer internship in Key West, Florida where she worked as an assistant educator at Reef Relief, a not for profit local organization dedicated to preserving local coral reef ecosystems in the Florida Keys. The next summer, Samantha was selected to do a research internship at Duke University's Marine Lab in Beaufort, NC. While there, she had her first true microbiology research experience where she worked on developing a molecular diagnostic tool to determine *in situ,* strain specific activity of *Prochlorococcus* clades. This was such a positive experience that she was invited to join the lab in the field the following winter for a 28 day research cruise in the Pacific Ocean. University of Tennessee professors Dr. Wilhelm and Dr Zinser happened to be sharing ship time with the Duke group at this time, and it was through this serendipitous meeting that Samantha ended up applying to and attending graduate school at the University of Tennessee. She enrolled in the Microbiology doctoral program in the Fall of 2013, joined

the lab of Dr. Steven Wilhelm, and received her Ph.D. in August 2019. Next, she will return to her interests in coral reef microbiology through a post-doctoral position based at Rice University in Houston, TX. She has also come ‘full circle’ to collaborating with her undergraduate research advisor, Dr. Teresa Carroll, as a need for microbiology based skills has arisen to address some of her work. Samantha plans to pursue a career in academia or a government-run research laboratory.