MANUAL
of
AQUATIC VIRAL ECOLOGY

MAVE Chapter 9, 2010, 82–91

The construction and analysis of marker gene libraries

Steven M. Short[1], Feng Chen[2], and Steven W. Wilhelm[3]*
[1]Department of Biology, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada
[2]Center of Marine Biotechnology, The University of Maryland Biotechnical Institute, Baltimore, MD 21202, USA
[3]Department of Microbiology, The University of Tennessee, Knoxville, TN 37922, USA

Abstract

Marker genes for viruses are typically amplified from aquatic samples to determine whether specific viruses are present in the sample, or to examine the diversity of a group of related viruses. In this chapter, we will provide an overview of common methods used to amplify, clone, sequence, and analyze virus marker genes, and will focus our discussion on viruses infecting algae, bacteria, and heterotrophic flagellates. Within this chapter, we endeavor to highlight critical aspects and components of these methods. To this end, instead of providing a detailed experimental protocol for each of the steps involved in examining virus marker gene libraries, we have provided a few key considerations, recommendations, and options for each step. We conclude this chapter with a brief discussion of research on a major capsid protein (*g20*) of cyanomyoviruses using this work as a case study for polymerase chain reaction primer design and development. By building on the experience of numerous labs, this chapter should not only be useful to the new virus ecologist, but also serve as a valuable resource to established research groups.

Introduction

Given the inherent difficulties associated with the isolation of purified viruses from aquatic environments, many researchers have chosen to explore the diversity and distribution of viruses using culture-independent molecular techniques. Due to their nature as obligate intracellular parasites, examination of viruses in the lab setting requires the concomitant maintenance and growth of the host organisms that they infect. Culture-independent approaches circumvent this constraint, allowing the researcher to characterize complex viral consortia directly. To achieve this however, one first requires sufficient knowledge of the genetic composition of the virus population in question. Within this chapter, we will describe how to interrogate the ecology of specific viruses in natural systems based on the limited amount of genetic information available from characterized viral isolates.

The characterization of viruses by these methods can briefly be described within a flow diagram that outlines the major steps in the construction and analysis of marker gene libraries (Fig. 1). Successful execution of this process however, requires careful application of appropriate controls and independent validations of individual steps. Within this chapter, we endeavor to highlight major components of these methods, discussing options and considerations in the specific step-by-step details. By building on the previous experience of numerous labs, this chapter should not only be useful to the new virus ecologist, but also serve as a valuable resource to established research groups.

Marker genes for viruses are typically amplified from aquatic samples for one of three purposes: 1) determining the presence of specific viruses, 2) determining the diversity of a group of related viruses, or 3) determining the abundance of a specific virus population based on the abundance of a marker gene. Within the context of this chapter, we will focus on the methods associated with 1 and 2 above, constraining our foci to viruses infecting algae, bacteria, and heterotrophic flagellates.

*Corresponding author: E-mail: steven.short@utoronto.ca

Acknowledgments

Publication costs for the Manual of Aquatic Viral Ecology were provided by the Gordon and Betty Moore Foundation. This document is based on work partially supported by the U.S. National Science Foundation (NSF) to the Scientific Committee for Oceanographic Research under Grant OCE-0608600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

The authors would like to thank colleagues and particularly previous students whom have helped work out the bugs of the described techniques. The authors also acknowledge the support of the Natural Science and Engineering Research Foundation Canada (SMS) and the National Science Foundation (NSF-OCE 0452409 SWW) for the support of their research programs that lead to the development of these ideas, and the Scientific Committee for Oceanographic Research for supporting working group 126 (marine virus ecology).

ISBN 978-0-9845591-0-7, DOI 10.4319/mave.2010.978-0-9845591-0-7.82

Suggested citation format: Short, S. M., F. Chen, and S. W. Wilhelm. 2010. The construction and analysis of marker gene libraries, p. 82–91. *In* S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.], Manual of Aquatic Viral Ecology. ASLO.

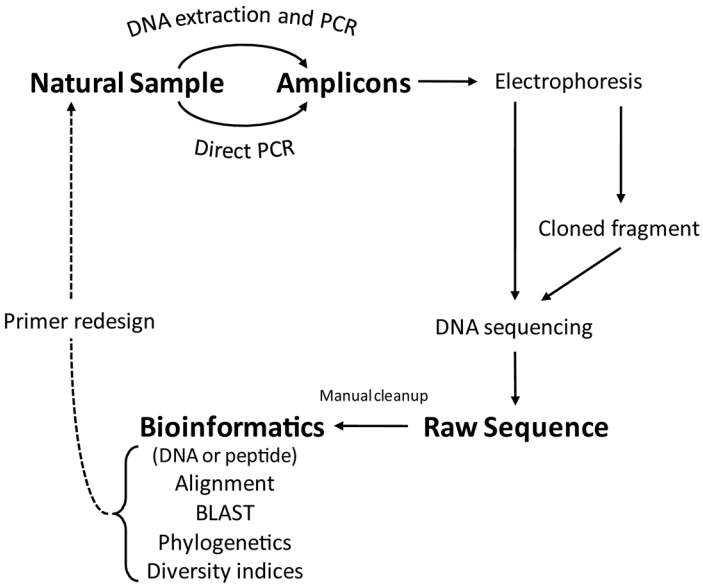

Fig. 1. Flowchart of steps from harvesting nucleic acids from virus samples to data analysis.

Materials and procedures

Viral gene markers—Viruses are probably the most diverse biological entity in the biosphere. Despite the fact that no universal gene marker (like the 16S and 18S ribosomal ribonucleic acid genes from prokaryotes and eukaryotes, respectively) is available for all viruses, many studies have demonstrated that certain genes are conserved among certain groups of viruses that infect closely related hosts. By designing oligonucleotide primers that hybridize to conserved regions of these marker genes, many researchers have used PCR to amplify virus marker genes from environmental samples to investigate the genetic diversity of specific groups of viruses in variety of aquatic environments (*see* Table 1). Currently, viral capsid related genes and virus-encoded deoxyribonucleic acid (DNA)/ribonucleic acid (RNA) polymerase gene are the most widely used genetic markers for aquatic viruses, and various PCR primer sets have been designed to target these genetic markers (Table 1). Studies of these virus marker genes have demonstrated that viruses in the marine environments are much more diverse than might be expected based on the limited numbers of cultivated viruses. With the recent rapid increase in the number of microbial genes and genomes available in public sequence databases, many viral signature genes (e.g., genes involved in photosynthesis or DNA replication) have been identified. By taking advantage of the plethora of information now available in sequence databases (e.g., NCBI's GenBank database at http://www.ncbi.nlm.nih.gov/Genbank/), polymerase chain reaction (PCR)-based methods can provide a rapid, sensitive, and economical approach to explore the diversity of viral genes or viral groups in nature and address important questions about the distribution, diversity, and even activity of virus in aquatic ecosystems.

Sample collection and preparation—The history and details of proper sample collection and processing before PCR amplification of virus genes are numerous. In some cases, virus markers can readily be amplified directly from unaltered whole water samples. In other cases, preconcentration of virus particles may be required; this process is thoroughly explained in another chapter (Wommack et al. 2010, this volume). For qualitative purposes, PCR amplification is often most successful from concentrated virus communities. However, the variety of steps involved in either ultrafiltration or ultracentrifugation increases the potential for particle loss, which can complicate quantitative analyses. Ultimately, the ambient abundance of viruses and the sensitivity of the particular assay will dictate the approach taken in preparing samples for analysis.

Similarly, a debate continues as to whether nucleic acids need to be extracted from virus samples prior to PCR amplification, or whether viral genetic material can be directly amplified. Many of the early studies on virus diversity in aquatic systems employed virus concentrates (*see* Wommack et al. 2010, this volume) as starting material. More recently, researchers have directly amplified marker elements from unextracted virus-bearing samples (Short and Short 2008; Wilhelm and Matteson 2008). Moreover, when comparing PCR amplification of unextracted virus concentrates and polyethylene glycol (PEG) precipitated virus concentrates to extracted viral DNA, the extracted DNA often produced poor PCR amplification yields (Chen et al. unpubl. results). While the approach of using unextracted virus DNA is often quite successful and requires only slight changes to the PCR protocol, its efficacy may depend on the capsid/membrane composition of the virus in question. Nonetheless, a simple freeze/heat treatment consisting of 3 repetitions of freezing virus samples until solid followed heating to 95°C for 2 min has been used to generate PCR-amplifiable virus DNA from a variety of aquatic samples (Chen et al. 1996; Short and Short 2008; Short and Suttle 2002).

Primer design—In targeting a specific population or group of microorganisms in aquatic environments using PCR-based methods, primer design is often the most critical and challenging step. Thankfully, because PCR is a well established technique, many excellent volumes have been written on the optimization and application of PCR, and most include some discussion of the critical considerations for primer design (e.g., Altshuler 2006; Atlas 1993; Innis et al. 1990; Mcpherson and Moller 2006), and some focus entirely on primer design (Yuryev 2007). In addition, freely available software can be found on the Internet that can aid in primer design. For example, the program OligoAnalyzer 3.1 is available at the Integrated DNA Technologies Web site (http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/). This particular software allows the user to enter oligonucleotide (primer) sequences and provides general analytical information such as predicted melting temperatures for each primer, as well as more complicated but useful information such as the primer's potential for hairpin formation, self-dimer formation, and hetero-dimer formation. This software also allows the user to

Table 1. Conserved genes and primers used in the construction and analysis of marker gene libraries

Taxonomic group	Gene	Gene product	Primer sequences	References
Cyanophage	*psbA*	photosynthetic gene D1	**58-VDIDGIREP-66:** GTN GAY ATH GAY GGN ATH MGN GAR CC **331-MHERNAHNFP-340:** GGR AAR TTR TGN GCR TTN CKY TCR TGC AT	(Zeidner et al. 2003)
Cyanophage	*psbA & psbD*	photosynthetic gene D1 & D2	**psbAF:** GTN GAY ATH GAY GGN ATH MGN GAR CC **psbAR:** GGR AAR TTR TGN GC	(Millard et al. 2004)
Cyanophage & Cyanobacteria	psbA	Photosynethetic gene D1	**psbA-93F:** TAY CCN ATY TGG GAA GC **psbA-341R:** GTR TTR AAG GGD GAR CT	(Wang and Chen 2008)
Myoviridae	g20	viral capsid assembly protein	**CPS1:** GTA GWA TTT TCT ACA TTG AYG TTG G **CPS2:** GGT ARC CAG AAA TCY TCM AGC AT	(Fuller et al. 1998)
Myoviridae	g20	viral capsid assembly protein	**CPS3:** TGG TAY GTY GAT GGM AGA **CPS4:** CAT WTC WTC CCA HTC TTC **CPS8:** AAA TAY TTD CCA ACA WAT GGA	(Zhong et al. 2002)
Myoviridae	g23	major capsid protein	**MZIA1bis:** GAT ATT TGI GGI GTT CAG CCI ATG A **MZIA6:** CGC GGT TGA TTT CCA GCA TGA TTT C	(Filee et al. 2005)
Myoviridae (cyanophage that infect *Anabaena* & *Nostoc*)	MCP	major capsid protein	**AN15 MCPF5:** GTT CCT GGC ACA CCT GAA GCG **AN15 MCPR5:** CTT ACC ATC GCT TGT GTC GGC ATC	(Baker et al. 2006)
Nodularia-specific cyanophage (myovirus & siphovirus)	g23	major capsid protein	**CAP1:** ATT TGY GGY GTT CAG CCK ATG A C**AP2:** AAC RAY TTC RCG GTT GAT TTC CA	(Jenkins and Hayes 2006)
T7-like Podophage	T7-DNA polymerase	DNA polymerase	**T7DPol230F:** ARG ARM RIA AYG GIT **T7DPol510R:** GTR TGD ATR TCI CC **HECTORPol29F:** GCA AGC AAC TTT ACT GTG G **HECTORPol711R:** CGA GAG ATA CAC CAA CGA A **PARISPol25F:** ATA CTA CAC GCT ACT CTG G **PARISPol701R:** GAG TGG CAA GAG GAG TTA T	(Breitbart et al. 2004)
Cyanophage of *Microcystis*			**Sheath RTF:** ACA TCA GCG TTC GTT TCG G **Sheath RTR:** CAA TCT GGT TAG GTA GGT CG	(Takashima et al. 2007)
EhV	MCP	major capsid protein	**MCP-F2:** TTC GCG CTC GAG TCG ATC[1] **MCP-R2:** GAC CTT TAG GCC AGG GAG	(Schroeder et al. 2003)
Phycodnaviridae	MCP	major capsid protein	**MCP Fwd:** GGY GGY CAR CGY ATT GA **MCP Rev:** TGI ARY TGY TCR AYI AGG TA	(Larsen et al. 2008)
Phycodnaviridae	*Pol*	DNA polymerase	**AVS1:** GARGGIGCIACIGTIYTIGAYGC **AVS2:** GCIGCRTAICKYTTYTTISWRTA **POL:** SWRTCIGTRTCICCRTA	(Chen and Suttle 1995)
HaV			**PKN61A:** GAT CTG ACT CATG ACC CAA CG **PKN61B:** CCA CCA TCA GAA TCA TCA CCC AGT **PKN62A:** AGA TGA AGA CGA TGA TGA CGA CGA T **PKN62B:** GTG AAG ATG AAA AGG AAA GCA AGG A **PAGB01A:** GTC AAG AAC AAT CGA ACC GTA AT **PAGB01B:** TTT ATT CTG ATT TCT GTC CCG	(Nagasaki et al. 2001)
T7-like Podophage	T7-DNA polymerase	DNA polymerase	**T7DPol230F:** ARG ARM RIA AYG GIT **T7DPol510R:** GTR TGD ATR TCI CC **HECTORPol29F:** GCA AGC AAC TTT ACT GTG G **HECTORPol711R:** CGA GAG ATA CAC CAA CGA A **PARISPol25F:** ATA CTA CAC GCT ACT CTG G **PARISPol701R:** GAG TGG CAA GAG GAG TTA T	(Culley and Steward 2007)

1. The 40 bp GC-clamp sequence associated with this primer was omitted for brevity.
2. The codes for mixed bases are: R = A, G; Y = C, T; M = A, C; K = G, T; S = C, G; W = A, T; H = A, C, T; B = C, G, T; V = A, C, G; D = A, G, T; n = A, G, C, T; I = inosine.

directly compare their primer sequences to sequences archived in the GenBank database.

As a general guideline, several criteria should be considered when designing PCR primers for the analysis of aquatic viruses:

1) the target gene should be evolutionarily conserved among the viruses of interest;
2) at least one region with a minimum of 6 consecutive amino acids (or >16 nucleotides) that is conserved only among the target organisms can be identified in multiple sequence alignments;
3) when multiple regions are available for primers, regions with the least degeneracy should be considered;
4) at sites of 4-fold degeneracy where G, A, T, and C should all be considered, the practical degeneracy of the primer can be reduced by using an inosine residue;
5) the desired size of PCR products may vary for different applications (e.g., shorter PCR amplicons ranging from 150–400 bp are ideal for DGGE applications and quantitative PCR (qPCR), whereas longer amplicons ranging from 500–800 bp are desirable for the phylogenetic analyses of clone libraries;
6) more than one set of primers should be designed and tested when multiple target regions are available;
7) because the design of specific PCR primers relies on the number of known target sequences, it is important to include as many related sequences as possible when creating sequence alignments for primer design;
8) PCR primers should be modified (redesigned) as more sequences belonging to the target organisms become available.

In some cases, PCR primers (e.g., primers that target the *g20* gene of cyanomyoviruses) were originally designed based on a limited number of gene sequences. This can result in poorly constrained sequence information since the specificity of primers was not well defined in the first place. Although it can easily be argued that poorly constrained sequence data are more valuable than no data at all, it is nonetheless important to use as much sequence data from representative groups of viruses when designing and redesigning PCR primers (Fig. 1). For example, using newly available sequence data, *g20* primers specific for cyanomyoviruses have been modified, and much higher PCR specificity has been achieved (Chen et al. unpubl. data; more details are described below; Fig. 2).

PCR amplification—PCR is a widely used in vitro technique that generates millions or even billions of copies of specific gene fragments. There are numerous general and field-specific procedural references for PCR, and almost all of the major scientific vendors distribute PCR reagents and equipment. Therefore, this section will only provide a simple guide to help neophyte molecular biologists get started; obviously there are far too many options that could be considered for a particular PCR application to discuss them here. Whenever possible, the procedure outlined in published literature describing the use of a particular set of primers should be followed. However, researchers should not be surprised when they need to troubleshoot previously described conditions for a particular reaction. In our experience, different Taq DNA polymerases, thermal cyclers and reagents, and even different workers, can have a dramatic influence on PCR results.

One of the most important considerations for PCR is lab hygiene. Because of its sensitivity, PCR reactions can easily be contaminated with amplifiable DNA. It is much easier to take proactive measures to prevent contamination that to have to track down the source of contamination after it has been detected. All reagents should be dispensed into small portions or working stocks before their use. This practice has the double benefit of preventing the loss of large stocks of reagents in the event that they become contaminated, and it also minimizes the number of freeze-thaw cycles that a reagent endures. The use of aerosol barrier tips for automatic pipettors, frequent sanitization of lab benches, and dedicated lab spaces or sterile hoods for setting up PCRs are also highly recommended. Although lab coats are generally recommended as essential personal protective equipment, they must be washed frequently if workers are to wear them when setting up PCR reactions; a dirty sleeve can be a major reservoir for contaminating nucleic acids! As a final comment, although it may seem obvious, it cannot be stressed enough that positive and negative controls must be included in every single PCR experiment.

PCR reactions are set up via the creation of a master mix that includes all reagents except the template nucleic acid. Generally, it is wise to prepare a slightly larger volume master mix that is absolutely necessary because the wasted reagents represent a trivial expense, and minor pipettor inaccuracies can lead to a short fall when dispensing the master mix into individual reaction tubes. The following reagents and concentrations are typical for many PCR reactions:

- PCR buffers are usually supplied at a 10× or 2× concentration with the polymerase enzyme. The buffers components are somewhat variable and are optimized by the manufacturer for use with a particular thermally-stable DNA polymerase enzyme.
- $MgCl_2$ is usually supplied in a 50 or 25 mM stock. The working concentration can vary between 1.5 to 4.0 mM depending on the primer sequences. For any particular PCR protocol, the optimal working concentration should be empirically determined as it can have a dramatic effect on the yield of PCR products and the stringency of the reaction.
- dNTPs can be purchased individually, or in mixtures of all four nucleotides. Generally, dNTPs are mixed and stored as stock solutions with each dNTP at a concentration of 10 mM, or 40 mM total for all dNTPs. For most PCR protocols, final concentrations of 0.2 mM of each dNTP is sufficient and provides ample product yield without negatively affecting the PCR specificity or fidelity.
- oligonucleotide primers can be ordered as lyophilized

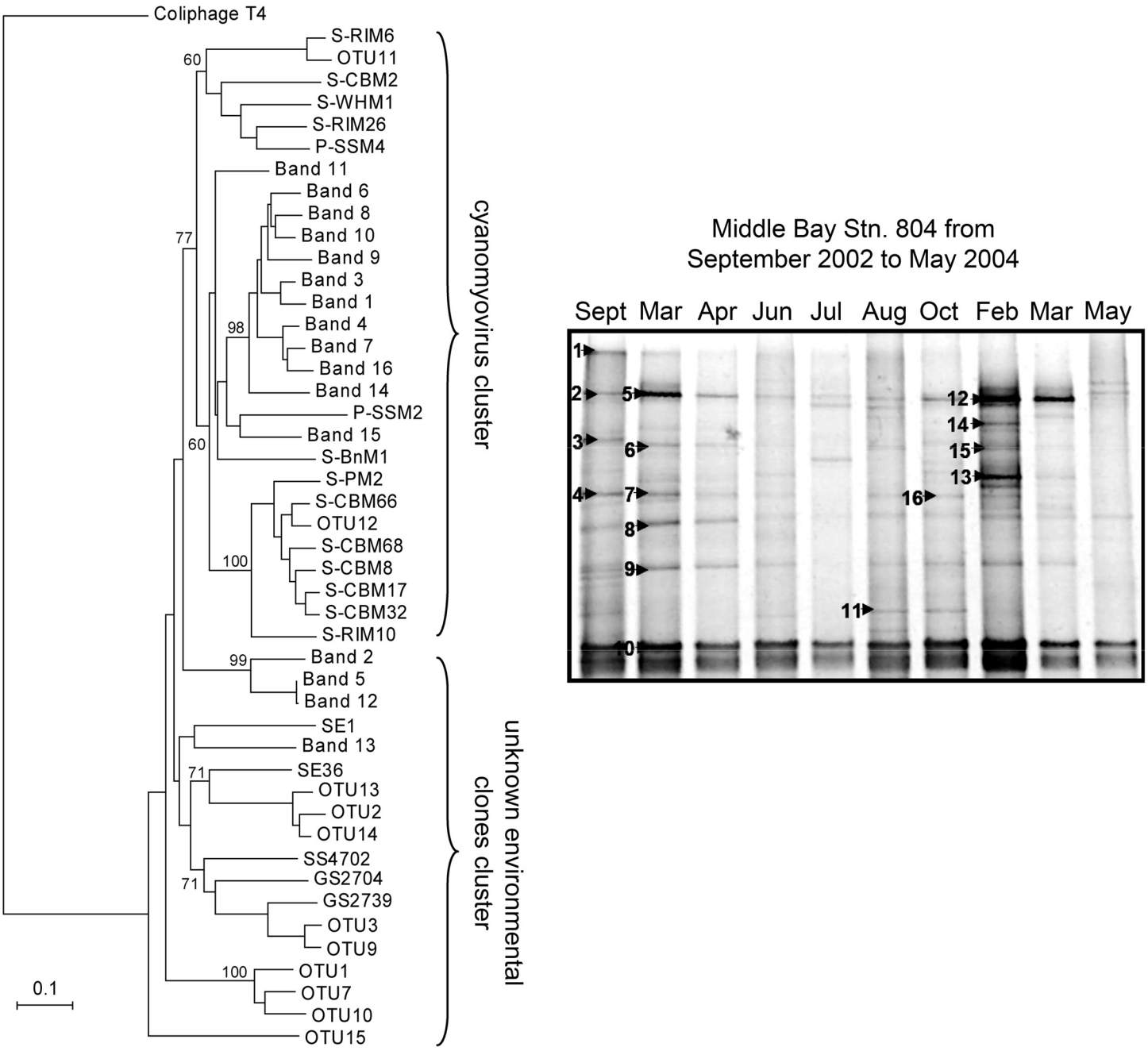

Fig. 2. Left: Phylogenetic analysis of cyanomyovirus g20 gene sequences (ca. 390 bp) from excised DGGE bands. The bootstrap values (>50) were shown on the major nodes. Right: DGGE profile of PCR-amplified g20 gene fragments at Sta. 804 in the Chesapeake Bay from September 2002 to May 2004.

stocks and can be reuspended in sterile, pure water or TE buffer (10 mM Tris, 0.1 mM EDTA, pH 7.5) for long-term storage at a concentration of 100 µM. Generally, aliquots of working stocks are made at 10 µM and the final concentration of each primer in a PCR reaction can range from 0.1 to 1.0 µM (i.e., a total of 10 to 100 pmol of primer in a final reaction volume of 50 µL) depending on the primer. Generally, PCR with degenerate primers require slightly higher primer concentrations, but the optimal primer concentration should be determined empirically.

- thermally stable DNA polymerases are the key ingredient in PCR as these enzymes withstand the extreme temperature fluctuations of thermal cycling. For many years, *Taq* DNA polymerase was the standard enzyme using for PCR. However, many vendors now produce various enzymes or enzyme mixtures that are optimized

for long PCR (amplification of fragments >10 Kb), or high fidelity amplification. Additionally, most manufacturers now produce reasonably priced hot-start enzymes that are not active until after the initial denaturation step. These hot-start enzymes are very useful as they prevent amplification artifacts produced by nonspecific primer annealing during the initial ramping up to the denaturation temperature.

- H_2O is added in sufficient volume to bring the total volume up to that desired for each reaction (the total volume for individual reactions is typically 25 or 50 µL depending on the desired yield). Although it is often overlooked as a potential source of amplification difficulties, H_2O quality is critically important. When possible, certified nuclease-free water should be used, but good results can be obtained with pure water that has been ultrafiltered and is ion free (i.e., 18.3 MΩ-cm resistivity).

Cloning and sequencing—By design, PCR methods for amplifying nucleic acids from aquatic viruses use universal primers that target related but different gene sequences. Because Sanger (dideoxy-based) sequencing reactions are confounded when more than one template is present, gene fragments from natural populations must be separated before sequencing reactions can be conducted. The most common approach to separate individual amplified gene fragments is to clone the PCR products into a plasmid vector, transform bacterial cells with the recombinant plasmids, and purify plasmids from individual isolated bacterial colonies; generally, each colony will contain only one type of recombinant plasmid. Purified plasmids can then be used as templates for sequencing reactions. Other methods like denaturing gradient gel electrophoresis (DGGE) can also be used to separate individual gene fragments from complex mixtures of PCR products. For these methods, individual bands that theoretically represent only a single DNA fragment are excised from the gel and are re-amplified with second round of PCR. After these second round PCR products are purified, they can then be used as templates for sequencing reactions.

Cloning PCR products has become relatively routine, and many manufacturers produce kits that can be used. Although the cost of the kits may exceed the cost of reagents prepared in-house, the time savings and efficacy of the kits far exceeds the relatively minor increased cost of cloning. The same statement can be made for most kit-based molecular methods, and therefore we have included below, lists of some common kits that can be used for many of the steps involved in the creation of marker gene libraries. The list of kits that we have provided is not meant to indicate any preferences or be all inclusive. Rather, the lists that follow are included to simply suggest a few reliable sources for these kits; many other manufacturers produce similar kits that may be equally cost effective and efficient. Most cloning or DNA purification kits include detailed instructions and trouble-shooting guides, and generally the manufacturer's recommendations and protocols should be followed. The competent cells used for bacterial plasmid transformation are often included in cloning kits, or they can be purchased separately. Although competent cells prepared by individual labs are considerably less expensive than commercially prepared cells, the effort to produce them may not be worth the cost savings unless they will be used routinely. Most general molecular biology manuals provide a protocol for the preparation of competent cells (Ausubel et al. 2002; Sambrook et al. 1989). Two types of kits are available for cloning PCR products. Some are based on a TA-cloning method that takes advantage of the single deoxadenosine overhang left by *Taq* DNA polymerase and other non-proofreading polymerase enzymes, while others are designed to clone blunt-ended PCR products. In either case, the number of colonies that contain recombinant plasmids with the desired PCR fragment can be greatly enhanced by loading all of the PCR reaction in an agarose gel, excising the fragment of the appropriate size, and purifying the fragment using a commercial gel extraction kit. In our experience, this step greatly reduces the possibility of ligating primer-dimers or other PCR artifacts into the plasmid vector, thereby enhancing the recovery of clones containing the gene fragment of interest.

Like PCR product cloning, DNA sequencing has become routine despite the high cost of the instruments used for automated sequence analysis. Generally, because high throughput or multi-user sequencing facilities offer sequencing services at significantly reduced cost compared with sequencing within individual labs, they have become the most common option for nucleotide sequencing. Many academic institutions and private companies provide sequencing services at a reasonable cost, and a brief web search should reveal many options for sequencing services. Sequencing reagents are produced by several manufacturers and vary depending on the automated sequencing instrument used. Most, if not all, sequencing facilities will recommend specific reagent kits and protocols for their users. The most important consideration for obtaining good sequencing results is the purity of the sequencing template as a poor quality template DNA is the most common cause for failed sequencing reactions. Therefore, no matter if sequencing templates are purified plasmids or PCR products, we highly recommend the use of commercial DNA purification kits because of their ease of use and the consistent DNA purity that they provide.

Common UA- or TA-based PCR cloning kits:

- Fermentas InsTAclone™ PCR Cloning Kit (http://www.fermentas.com/)
- Invitrogen TOPO TA Cloning® Kit (http://www.invitrogen.com/)
- Promega pGEM-T and pGEM-T Easy Vector Systems (http://www.promega.com/)
- Stratagene StrataClone™ PCR Cloning Kit (http://www.stratagene.com/)

Common blunt-end PCR cloning kits:

- Clontech In-Fusion™ PCR Cloning Kits (http://www.clontech.com/). *Note*: although this kit does not require deoxyadenosine ("A") overhangs on PCR fragments to be cloned; blunt-end polishing is also not required.

- Fermentas CloneJET™ PCR Cloning Kit (http://www.fermentas.com/)
- Invitrogen Zero Blunt® TOPO® PCR Cloning Kit (http://www.invitrogen.com/)
- Stratagene StrataClone™ Blunt PCR Cloning Kit (http://www.stratagene.com/)

Common gel extraction kits:

- Fermentas DNA gel extraction kit (http://www.fermentas.com/)
- Invitrogen PureLink™ Gel Extraction Kit (http://www.invitrogen.com/)
- Promega Wizard®" DNA Clean up system (http://www.promega.com/)
- Qiagen QIAquick Gel Extraction kit (http://www.qiagen.com/)
- Stratagene StrataPrep® DNA Gel Extraction Kit (http://www.stratagene.com/)

Common plasmid miniprep kits:

- Fermentas GeneJET™ Plasmid Miniprep Kit (http://www.fermentas.com/)
- Invitrogen ChargeSwitch® NoSpin Plasmid Micro Kit (http://www.invitrogen.com/)
- Promega Wizard® Plus Minipreps DNA purification system (http://www.promega.com/)
- Qiagen QIAprep Spin Miniprep Kit (http://www1.qiagen.com/)
- Stratagene StrataPrep® Plasmid Miniprep Kit (http://www.stratagene.com/)

Common PCR cleanup kits:

- Applied Biosystems DNAclear™ kit (http://www.appliedbiosystems.com/)
- Fermentas DNA gel extraction kit (http://www.fermentas.com/)
- Invitrogen ChargeSwitch® PCR Clean-Up Kit (http://www.invitrogen.com/)
- Promega Wizard® DNA Clean up system (http://www.promega.com/)
- Qiagen QIAquick PCR purification kit (http://www1.qiagen.com/)
- Stratagene StrataPrep® PCR Purification Kit (http://www.stratagene.com/)

Bioinformatic analysis—Once sequences have been obtained from a marker gene clone library, the steps involved in sequence analysis include 1) sequence editing, 2) sequence alignment, 3) phylogenetic inference, 4) drawing phylograms, and 5) calculating diversity indices (Fig. 1). Although the analysis of clone library sequences can seem daunting to the uninitiated, references such as Hall's book *Phylogenetic Trees Made Easy* (2008) offer excellent advice and background information that will walk beginners through the essential elements of sequence analysis; more in-depth discussions of phylogenetic inference can be found in advanced texts (Felsenstein 2004; Graur and Li 2000; Hillis et al. 1996).

By its very nature, bioinformatic analysis is computationally intensive and is conducted using a variety of software. In recent years, computer software and hardware has changed dramatically, and most of these changes have resulted in easy to use and widely available bioinformatic software. For example, Macintosh computers now use an Intel chip that allows them to use the Windows operating system, and there are Windows emulators available for both Linux and Unix operating systems. Therefore, to ensure that this discussion is useful to the broadest possible audience, we have focused on the use of Windows-based software that is freely available on the World Wide Web (most of the software listed in the following paragraphs is also available in versions compatible with Unix or Macintosh operating systems). For the sake of brevity, we will not discuss the parameters that must be considered when analyzing genetic libraries. Instead, we will simply point readers to the excellent texts mentioned in the preceding paragraph, and provide a brief list of some of the available free software, noting their major functions and the Web site from which they can be downloaded:

- BioEdit (Hall 1999). This software can be used for sequence editing and much more. It is available free of charge at http://www.mbio.ncsu.edu/BioEdit/BioEdit.html. This software package can be used to view the chromatograms produced by several different types of automated sequencers, and it can also be used to analyze the physical properties of nucleic acid or amino acid sequences. Further, it can be used to translate DNA sequences, search sequences for defined motifs, conduct BLAST searches locally or to the GenBank database, align sequences using ClustalW, and it produces publication quality prints of sequence alignments. This is an extremely useful program that has far too many functions to list here.
- ClustalX (Thompson et al. 1997). This software is the most widely used sequence alignment software available. It can be used generate pairwise and multiple alignments of nucleotide and amino acid sequences, and a variety of parameters such as gap penalties and the substitution matrix can be set by the user. It is downloadable for free from http://bips.u-strasbg.fr/fr/Documentation/ClustalX/.
- Mega 4 (Tamura et al. 2007). This software can be used to align nucleic acid or amino acid sequences, estimate evolutionary distances using a variety of models, build phylogenetic tress via neighbor joining or maximum parsimony methods, and test phylogenetic tree reliability via interior branch tests or bootstrap analysis. In addition, Mega 4 has extensive tree viewing, manipulation, and editing tools that can be used to create publication quality trees in a variety of file formats. This software is free and can be downloaded from http://www.megasoftware.net/.
- MrBayes (Ronquist and Huelsenbeck 2003). This software is used for Bayesian phylogenetic inference. Bayesian inference of phylogeny has become very popular among molecular systematists and is based on the posterior probability distribution of trees using a Markov chain Monte

Carlo simulation technique that approximates these posterior probabilities. Although this software is operated through command lines and is not as easy user friendly as other graphical interface programs, excellent documentation is provided with the software, and Hall (2008) provides a good tutorial to help beginning users get started. MrBayes is available for free download from http://mrbayes.csit.fsu.edu/.

- Phylogeny.fr: robust phylogenetic analysis for the non-specialist (Dereeper et al. 2008). This free web service incorporates several alignment and phylogenetic tools into a user friendly website that can be used to reconstruct and analyze phylogenetic relationships between molecular sequences in a single-step or, for more experienced users, an "A la carte" menu can be used to tailor various aspects of the phylogenetic workflow. This site also includes extensive documentation. The site can be accessed at http://www.phylogeny.fr/.
- EstimateS: Statistical estimation of species richness and shared species from samples. Version 8.0.0, R. K. Colwell. 2006. This software can be used to calculate a variety of biodiversity functions, estimators, and indexes based on a range of biological data. For example, EstimateS can be used to compute rarefaction and species accumulation curves, as well as a variety of different species richness estimators for data from marker gene libraries. EstimateS is a free software application that can be downloaded from http://viceroy.eeb.uconn.edu/estimates. Excellent supporting documentation for the software is also available at the same Web site.
- Rarefaction Calculator (http://www2.biology.ualberta.ca/jbrzusto/rarefact.php), Analytic Rarefaction (http://www.uga.edu/strata/software/index.html), and DOTUR (http://schloss.micro.umass.edu/software/) (Schloss and Handelsman 2005) are other free software applications that can be used to estimate rarefaction curves for data from marker gene libraries. We have included them because of their simplicity and ease of use.

Assessment

As mentioned above, the genetic diversity of cyanomyoviruses in various aquatic environments has been investigated extensively. However, a large proportion of environmental *g20* sequences do not appear to be from myoviruses that infect *Synechococcus* and *Prochlorococcus* since they cluster outside clades containing sequences from laboratory isolates (Marston and Sallee 2003; Short and Suttle 2005; Wang and Chen 2008; Wilhelm et al. 2006; Zhong et al. 2002). For example, among 207 clones retrieved from diverse marine environments, about 80% did not cluster with known cyanomyoviruses (Zhong et al. 2002). More than 60% of DGGE band sequences recovered from both marine and freshwater environments were outside the cyanomyovirus cluster (Short and Suttle 2005). This problem occurs because the PCR primers were designed based on the limited cyanomyovirus *g20* gene sequences. More specific PCR amplification can be achieved when more gene sequences become available. By redesigning the *g20* gene primers based on the newly available cyanomyovirus *g20* sequences, a high proportion of environmental clones fell into the Cyanomyovirus cluster (Fig. 2, left panel). The modified *g20* primer set SMP-1F and SMP-2R (Wang and Chen unpubl. data) were designed based on nearly 30 cyanomyovirus *g20* sequences, and could be useful for specifically monitoring the population dynamics of cyanomyoviruses in the natural environment. With modified *g20* primers, 75% of DGGE band sequences fell within the Cyanomyovirus cluster. The seasonal shift on cyanomyovirus populations in the Chesapeake Bay can be seen from the DGGE analysis of *g20* amplicon (Fig. 2, right panel).

The *g20* gene study is just one of many examples showing the difficulty or limitation of using molecular tools to explore the diversity of microbes in nature. Many steps related to PCR amplification (i.e., *Taq* enzymes, number of PCR cycles, etc.) could also cause the biased results. Therefore, it is important to optimize the PCR conditions before a large quantity of samples are analyzed. Finally, this study is also limited by the availability of sequences in publicly available databases. While many *g20* amplicons fall outside the clusters associated with known cyanophage isolates, the highest identity remains that of cyanophage *g20* genes. As such, the investigator must ultimately understand that the interpretation of molecular data from culture independent studies is at the mercy of the available data in molecular repositories. While this will no doubt improve over time, in the case of some understudied virus groups, data reanalysis in subsequent years may result in different interpretations.

Discussion

The application of molecular tools to questions concerning the ecology of viruses is a rapidly changing area. Already in the last several years, advances in DNA sequencing technologies have exponentially expanded the available database of genetic information from viruses (Zeidner et al. 2003). Given the rate of advancement in both the theory and technology associated with this area of research, it is perhaps most important to caution researchers to be sure that they have fully examined the most recent literature prior to establishing a new program of research. Ultimately though, different laboratories use different tools, and researchers are encouraged to adapt their own available tool sets and materials when addressing questions of marine virus diversity.

With respect to choices regarding the use of established primer sets, it is important that investigators carefully follow recommended protocols when adapting techniques developed in another lab (and as such that these protocols are well documented for publication). Sometimes even slight changes in instrumentation (e.g., the type of thermal cycler) or basic sources of reagents (e.g., similar polymerases from different

vendors) can markedly influence the success of a molecular biological exercise. As with so many other biological systems, much of the molecular biology of aquatic viruses comes down to proper validation, optimization, and the use of both positive and negative controls to get the best possible data.

Comments and recommendation

The molecular examination of viruses in aquatic communities is just one of the many areas of virus ecology where researchers are making tremendous and rapid strides forward. As PCR-based molecular techniques have improved our qualitative understanding of microbial diversity, quantitative molecular approaches for studying virus communities, although in their infancy, will allow us to better understand processes associated with either the entire virus community or specific virus populations. While many challenges remain in the adaptation of lab techniques (e.g., quantitative PCR) to field studies, these challenges and others associated with PCR-based approaches will undoubtedly be solved in the near future. As such, perhaps the most important recommendation to both the neophyte and the experienced researcher is to complete a thorough examination of the peer-reviewed literature prior to taking on any project. While we provide what we feel are sound recommendations in the current review, the trajectory of this field of research is steep and demands that students of this field need to be up-to-date on the most recent advances.

References

Altshuler, M. L. 2006. PCR troubleshooting: The essential guide. Caister Academic Press.

Atlas, R. M. 1993. Detecting gene sequences using the polymerase chain reaction, p. 267-270. *In* P. F. Kemp, B. F. Sherr, E. B. Sherr, and J. J. Cole [eds.], Handbook of methods in aquatic microbial ecology. Lewis Publishers.

Ausubel, F. M., and others [eds.]. 2002. Short protocols in molecular biology: a compendium of methods from current protocols in molecular biology, 5th ed. Wiley.

Baker, A. C., V. J. Goddard, J. Davy, D. C. Schroeder, D. G. Adams, and W. H. Wilson. 2006. Identification of a diagnostic marker to detect freshwater cyanophages of filamentous cyanobacteria. Appl. Environ. Microbiol. 72:5713-5719.

Breitbart, M., J. H. Miyake, and F. Rohwer. 2004. Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol. Lett. 236:249-256.

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

———, ———, and S. M. Short. 1996. Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. Appl. Environ. Microbiol. 62:2869-2874.

Culley, A. I., and G. F. Steward. 2007. New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. Appl. Environ. Microbiol. 73:5937-5944.

Dereeper, A., and others. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36:W465-W469.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates.

Filee, J., F. Tetart, C. A. Suttle, and H. M. Krisch. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. Proc. Natl. Acad. Sci. USA 102:12471-12476.

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

Graur, D., and W. Li. 2000. Fundamentals of molecular evolution, 2nd ed. Sinauer Associates.

Hall, B. G. 2008. Phylogenetic trees made easy: a how-to manual, 3rd ed. Sinauer Associates.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

Hillis, D. M., C. Moritz, and B. K. Mable [eds.]. 1996. Molecular systematics, 2nd ed. Sinauer Associates.

Innis, M. A., D. H. Gefland, J. J. Sninsky, and T. J. White. 1990. PCR protocols: A guide to methods and applications. Academic Press.

Jenkins, C. A., and P. K. Hayes. 2006. Diversity of cyanophages infecting the heterocystous filamentous cyanobacterium Nodularia isolated from the brackish Baltic Sea. J. Mar. Biol. Assoc. U.K. 86:529-536.

Larsen, J. B., A. Larsen, G. Bratbak, and R. A. Sandaa. 2008. Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. Appl. Environ. Microbiol. 74:3048-3057.

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

McPherson, M. J., and S. G. Møller. 2006. PCR, 2nd ed. Taylor & Francis.

Millard, A., M. R. J. Cloki.e., D. A. Shub, and N. H. Mann. 2004. Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. Proc. Natl. Acad. Sci. USA 101:11007-11012.

Nagasaki, K., K. Tarutani, M. Hamaguchi, and M. Yamaguchi. 2001. Preliminary analysis of *Heterosigma akashiwo* virus DNA. Microbes Environ. 16:147-154.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press.

Schloss, P. D., and J. Handelsman. 2005. Introducing DOTUR, a computer program for defining operational taxonomic

units and estimating species richness. Appl. Environ. Microbiol. 71:1501-1506.

Schroeder, D. C., and C. A. Suttle. 2002. Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. Appl. Environ. Microbiol. 68:1290-1296.

———, J. Oke, M. Hall, G. Malin, and W. H. Wilson. 2003. Virus succession observed during an *Emiliania huxleyi* bloom. Appl. Environ. Microbiol. 69:2484-2490.

Short, C. M., and C. A. Suttle. 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl. Environ. Microbiol. 71:480-486.

Short, S. M., and C. M. Short. 2008. Diversity of algal viruses in various North American freshwater environments. Aquat. Microb. Ecol. 51:13-21.

Takashima, Y., and others. 2007. Development and application of quantitative detection of cyanophages phylogenetically related to cyanophage Ma-LMM01 infecting *Microcystis aeruginosa* in fresh water. Microbes Environ. 22:207-213.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24:1596-1599.

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

Wang, K., and F. Chen. 2008. Prevalence of highly host-specific cyanophages in the estuarine environment. Environ. Microbiol. 10:300-312.

Wilhelm, S. W., M. J. Carberry, M. L. Eldridge, L. Poorvin, M. A. Saxton, and M. A. Doblin. 2006. Marine and freshwater cyanophages in a Laurentian Great Lake: Evidence from infectivity assays and molecular analyses of g20 genes. Appl. Environ. Microbiol. 72:4957-4963.

———, and A. R. Matteson. 2008. Freshwater and marine virioplankton: a brief overview of commonalities and differences. Freshw. Biol. 53:1076-1089.

Wommack, K. E., T. Sime-Ngando, D. M. Winget, S. Jamindar, and R. R. Helton. 2010. Filtration-based methods for the collection of viral concentrates from large water samples, p. 110-117. *In* S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.], Manual of Aquatic Viral Ecology. ASLO.

Yuryev, A. [ed.]. 2007. PCR primer design. Humana Press.

Zeidner, G., and others. 2003. Molecular diversity among marine picophytoplankton as revealed by psbA analyses. Environ. Microbiol. 5:212-216.

Zhong, Y., F. Chen, S. W. Wilhelm, L. Poorvin, and R. E. Hodson. 2002. Phylogenetic diversity of marine cyanophage isolates and natural virus communities as revealed by sequences of viral capsid assembly protein gene g20. Appl. Environ. Microbiol. 68:1576-1584.