

Sequencing and characterization of virus genomes

William H. Wilson^{1*} and Declan Schroeder^{2*}

¹Bigelow Laboratory for Ocean Sciences, 180 McKown Point Road, POB 475, West Boothbay Harbor, Maine, 04575, USA

²Marine Biological Association of the UK, Citadel Hill, Plymouth, PL1 2PB, UK

Abstract

By unraveling the genetic code of viruses, genome sequencing offers a new era for aquatic virus ecology giving access to ecological function of viruses on an unprecedented scale. Although this chapter starts with the suggestion to that virus genome sequencing should be conducted professionally if financially feasible, we essentially try and guide the reader through some of the procedures that will direct a novice through a genome sequencing project. Arguably, the most important recommendation is to start with as high purity virus nucleic acid as possible. We use the adage, junk in equals junk out. Once sequence information is obtained, there is plenty of free, user-friendly software available to help build, annotate, and then compare sequence data. Acquiring metadata is another important aspect that is not often considered when embarking on a genome project. A new initiative by the Genomic Standards Consortium has introduced Minimum Information about a Genome Sequence (MIGS) that allows standardization of the way the data are collected to make it useful for downstream post-genomic analyses. Most viruses sequenced to date have produced surprises, and there is more to come from the other 10³¹ viruses still to be sequenced. This chapter focuses on sequencing purified virus isolates rather than virus metagenomes.

Introduction

By way of a preface, this is not a detailed list of step-by-step methods on how to sequence a virus genome. Sequencing projects, particularly for large DNA viruses (100 kb–1200 kb), are significant undertakings for any lab-based project, and the magnitude of such an onerous task is often underestimated. Here, a sequencing project is defined as all the steps from

*Corresponding author: E-mail: ¹wwilson@bigelow.org and ²dsch@mba.ac.uk

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.

Research discussed here was funded partly from the Natural Environment Research Council of the UK (NERC) Environmental Genomics thematic program (NE/A509332/1 and NE/D001455/1) and partly from the US National Science Foundation (EF0723730). DCS is funded by NERC and through the NERC core strategic research programme Oceans2025 (R8-H12-52). We thank Mike Allen for providing Table 1.

ISBN 978-0-9845591-0-7, DOI 10.4319/mave.2010.978-0-9845591-0-7.134

Suggested citation format: Wilson, W. H., and D. Schroeder. 2010. Sequencing and characterization of virus genomes, p. 134–144. In S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.], Manual of Aquatic Viral Ecology. ASLO.

obtaining a clonal virus isolate through to generation of a completely annotated virus genome. The first question to ask is whether it is financially feasible to get your virus of choice sequenced by a professional facility, i.e., large-scale, high throughput sequencing and bioinformatics facilities. It takes several skill sets and some dedicated expensive equipment to do the job efficiently, a facility not often afforded by a standard aquatic virology laboratory. Exceptions to this are smaller-scale sequencing projects (1 kb–50 kb) such as the small RNA-virus genomes (Lang et al. 2004; Shirai et al. 2008) or some of the smaller, straightforward (e.g., no extensive repeat regions) DNA viruses (Rohwer et al. 2000). With an increase in new technologies such as 454 (www.454.com), high throughput mass sequencing is becoming more accessible. However, it is still expensive to sequence a single virus genome unless you work together with other researchers to get several viruses on a single run or can negotiate with a facility to use up spare capacity on a high throughput run. The first swathe of sequence data are only a starting point, and although you can get 99% of the genome sequence very quickly, i.e., within 1–2 weeks (Lander and Waterman 1988), finishing the genome can take up to 95% of the time and budget of an entire sequencing project. In this chapter, we will discuss some of the options when considering the practicality of finishing a virus genome. Clearly, there will be projects on a tight budget that will want to attempt to glean basic sequence information to help develop hypotheses on their viruses. For this, we will provide some basic protocols

with Web links and citations to similar projects. This chapter will also explore some of the thinking behind why sequencing a genome should be considered, we will provide an assessment of some of the current techniques and explore how sequence data can be verified once a basic genome annotation has been conducted. Guidelines for collecting metadata associated with the genomes will also be provided as metadata is increasingly important as more viral genomes are sequenced and comparative analyses become feasible. Verification of functional assignment (annotation) is essentially how characterization is defined in this chapter—a much more significant process than the grunt of obtaining the sequence data.

Materials and procedures

Before embarking down the road of virus genome sequencing and annotation, a number of key validation checks should be made:

- Make every effort to start with a clonal and axenic isolate. *Why clonal?* Mutation rates in viruses are known to be very high, especially when compared with its cellular hosts, and because the progeny of a single infection event may lead to the production of many variants of the original, it is, therefore, imperative not to complicate the sequence analysis even further by starting off with a mixture of similar genotypes. This can be done by performing either dilution to extinction (Nagasaki and Bratbak 2010, this volume) or plaque assay (Schroeder et al. 2002) experiments. *Why axenic?* Bacterial sequence contamination is a major problem when sequencing novel genomes. Many sequencing programs have come unstuck because of this very issue. Consequently, give yourself every opportunity of generating sequencing information of your virus by starting off with a well-defined clean system.
- Aim to get a large starting quantity of virus. Sequencing protocols are very wasteful, so it is of utmost importance that large quantities of virus, and thus genomic material, be produced (a minimum of 100 ng, though ideally aim for up to 10 µg). Ideally, this needs to be done in one single event from the same starting virus inoculum (i.e., from a clonal virus preparation such as a plaque resuspension). This is to avoid amplifying variants of the original generated by successive and continuous re-inoculation.
- Determine an efficient virus concentration protocol (Lawrence and Steward 2010, this volume). Before proceeding to the next step of extracting virus nucleic acids, it is best to test whether you have any carry over of cellular genomes. This can be easily achieved by using universal ribosomal DNA primer sets. If these PCRs produce positive results, treat your virus concentrate with commercially available nucleases. Since virus nucleic acids are still protected by their capsid, this treatment will have little or no effect on its genome. However, do remember to inactivate the enzymes before proceeding to the nucleic acid extraction phase.

- Nucleic acid quality assessment. Test the quality of the nucleic acids generated using either PCR or restriction digestion. This will help determine if your nucleic acid is suitable for downstream sequencing.

Once the initial validation has been done, you can proceed with the assurance that you have done everything possible to mitigate sources of producing junk sequence data. The next few steps entail the extraction and manipulation of nucleic acids, depending on the format available to you:

- Nucleic acid preparation. The choice of nucleic acid extraction will depend on the type of virus genome (RNA or DNA, ss or ds), the quantity and quality produced by the method, and the budget available to you. There are a number of commercially available kits (e.g., Qiagen) that will perform a perfectly adequate job. Alternatively, the universally tried and tested phenol-chloroform method for nucleic acid extraction (Lawrence and Steward 2010, this volume) normally delivers good results.
- Nucleic acid random fragmentation. Depending upon the size of the genome and sequencing strategy used, fragmentation may be necessary. For larger dsDNA viruses (>100 kb), the nucleic acids will need to be fragmented into smaller clonable sizes (e.g., 1–4 kb) for shotgun cloning and Sanger-based sequencing. This can be done enzymatically by controlled DNaseI treatment (Rohwer et al. 2000) or physically by sonication (Wilson et al. 2005).

DNaseI treatment—Generation of DNA fragments by DNase digestion involves digesting DNA for a range of times, then picking the time that gives optimal-sized DNA fragments (typically 1000–4000 bp). In a 50 µL reaction volume, resuspend 8 µg DNA in 50 mM Tris-HCl (pH 7.6), 10 mM MnCl₂, 100 µg mL⁻¹ bovine serum albumin, and 0.01 SU mL⁻¹ DNase I. Remove 5 µL aliquots (adding to 45 µL TE buffer, pH7.6) 0, 0.5, 1, 2, 5, 10, 15, and 30 min after addition of the digestion mixture and immediately transfer to a tube containing 25 µL Tris-buffered (pH 7.0) phenol (typically the shorter incubations, up to 2 min, give optimally sized fragments). After a phenol:chloroform (1:1) and two chloroform extractions, precipitate the fragmented DNA, wash with 70% ethanol, and dry. Resuspend fragmented DNA in 23 µL of Blunt-ending Mix (100 µM dNTPs, 1 × T4 DNA Pol Buffer) and heat at 65°C for 30 min to resuspend DNA and inactivate any DNase I that was carried over. After cooling to room temperature, add 2.5 U Klenow fragment and 5 U T4 DNA polymerase then incubate the reaction at 37°C for 1 h. The fragmented and blunt-ended virus DNA can be run on a 1% agarose gel prior to excising fragments in the 1000–4000 bp range using a standard gel extraction procedures before downstream cloning (NB do not excise fragments smaller than 1000 kb, as downstream cloning will preferentially clone the smaller fragments).

Sonication—Generation of DNA fragments by sonication is performed by placing a microcentrifuge tube containing the buffered DNA sample into an ice-water bath in a cup-horn sonicator. Sonication is conducted for a varying number of

10-s bursts using maximum output and continuous power. Exact conditions for sonication should be empirically determined for a given DNA sample before a preparative sonication is performed. Typically, 100 µg DNA in TE buffer is split into 10 aliquots of 35 µL; 5 are subjected to sonication for increasing numbers of 10 s bursts. Aliquots from each time point are run on an agarose gel to determine optimal-sized DNA fragments (1–4 kb). Once optimal sonication conditions are determined, the remaining 5 aliquots (approximately 8 µg) are sonicated according to those predetermined conditions. DNA can be blunt-ended and size-selected as above prior to downstream cloning.

- Cloning. If newer “next generation sequencing” (see below) options are chosen, any nucleic acid fragmentation or cloning will be conducted by the sequencing facility. Fragmented DNA can be cloned into a wide range of commercially available cloning vectors (e.g., www.promega.com/vectors/cloning_vectors.htm) and/or cloning kits that are available from a wide range of molecular reagents companies (e.g., Promega, Invitrogen, New England Biolabs) that make cloning almost fool-proof. However, remember that cloning procedures work best with high purity insert DNA. It is worth checking the quality of DNA inserts with A260/280 ratios prior to cloning with specific spectrophotometry devices such as NanoDrop (www.nanodrop.com). Smaller DNA or RNA genomes can be cloned whole in specifically designed bacterial artificial chromosome (BAC)- or Yeast Artificial Chromosome (YAC)-based cloning and sequencing vectors, EPICENTRE Biotechnologies is one company that provides a range of options for this (www.epibio.com). Clones can then be sequenced directly, usually using Sanger-based technology (for basic explanation and animation, see <http://www.dnai.org/text/mediashowcase/index2.html?id=552>).
- Sequencing. For laboratory-based ‘in-house’ projects, clone libraries (or PCR products) are typically run with Sanger-based technology. A search of Web sites reveals numerous tips and protocols for improving reads and reducing the cost of sequencing (typically by diluting sequencing enzyme) (e.g., www.nucleics.com). Note the first tip on this site is “Use clean DNA” (junk in = junk out!). If the sequencing project can afford to send DNA directly to a sequencing facility, there are numerous options currently available, from Sanger-based sequencing services to the new generation high throughput services:
- Sanger-based sequencing. This technology utilizes DNA polymerase and chain terminating fluorescently bases to create four series of labeled DNA fragments. Sequencing platforms (e.g., ABI & Beckman) capable of resolving these fragments can accurately and efficiently resolve on average ~700–800 bp. Therefore, M13 *E. coli* based vectors are routinely used to create ~1 kb size clone libraries. The amount of sequence generated is dependent on the size of the virus genome. The general rule of thumb is to generate sequence data of at least 8-fold coverage of your genome. This approach will provide up to 99% coverage of your genome, and then finishing approaches will be required to obtain a full genome (see below).
- 454 Sequencing Technology (Roche’s Genome Sequencer FLX system; www.454.com). Pyrosequencing is based on a method developed by Ronaghi et al. (1996) and uses enzyme-mediated luminescence during DNA synthesis. The 454 sample preparation first relies on fragmentation of the genomic DNA followed by binding each fragment to a microscopic bead. Pyrosequencing is then used to determine the sequence of each bead-associated DNA fragment. This technology can produce sequence reads of around 200–400 bp, providing around 80–120 Mb per run. The plate platform design of 454 allows the utility of splitting and/or dividing plates, a cost-saving measure that does not appear to be available for other technologies. 454 has been successfully used for finishing or genome assembly purposes (see below).
- Illumina’s Solexa technology (www.illumina.com). The second commercially available next-generation technology also fragments genomic DNA, which is then ligated to a glass surface where bridge amplification creates multiple clusters of identical sequences (Bentley et al. 2008). These then go on to be sequenced by a synthesis step using fluorescently labeled terminators with imaging following each successive base addition. The read lengths are much shorter than those from 454’s instrument (~35 bp), therefore this technology is not appropriate for de novo sequencing. However, it comes into its own if you have a genome to align the sequence against as each run (flowcell with 8 channels) produces many times more reads than 454.
- SOLiD technology (ABI) (http://www3.appliedbiosystems.com/AB_Home/applicationtechnologies/SOLIDSystemSequencing/index.htm). This most recent addition to the next-generation sequencing platforms amplifies fragmented genomic DNA by emulsion PCR on beads, followed by cyclic array (polony) sequencing (Shendure et al. 2005). To date, this technology has yet to be thoroughly exploited by sequencing enthusiasts.
- Assembly. Assembly is a crucial component of the annotation process. The major sequencing facilities (e.g., Sanger Institute, Genoscope, DOE Joint Genome Institute) are continually developing and optimizing their own assembly programs. Many assembly programs are however freely available for download, e.g., Phred/Phrap (<http://www.phrap.org/>), Staden (<http://staden.sourceforge.net/overview.html>), Celera Assembler (<http://apps.sourceforge.net/mediawiki/wgs-assembler>), AMOS (<http://amos.sourceforge.net/>), and ARACNE (<http://www.broad.mit.edu/science/programs/genome-biology/computational-rd/computational-research-and-development>), all being suited for virus genome assembly. Many commercial packages such as DNASTAR

(<http://www.dnastar.com/products/seqmanpro.php>) and DNA BASER (<http://www.dnabaser.com/>) arguably produce more user friendly software, much more suited for the part-time assembler.

- **Finishing.** To finish or not to finish, that is the question! The next phase, the finishing strategies are arguably the most onerous and time-consuming tasks. Recent developments, i.e., next generation sequencing, are being touted as the savior of many aspiring genome assembler. As these technologies are for those with substantial research budgets, closing physical gaps are mainly achieved through cumbersome phases of PCR amplifications on genomic DNA with primer pairs positioned on independent contig ends. However, if there is a hole in the sequence, which may be due to a physical barrier introduced by the fragmentation/cloning or some troublesome spot for sequencing, it may be best to focus on identifying the arrangement of the contigs with respect to one another and then “filling in” the holes with PCR amplifications across gaps and subsequent sequencing of those products. Other strategies include sequencing libraries with much larger inserts using cosmids or fosmids (around 40 kb inserts). These vectors provide greater cloning flexibility and construct stability, however, if larger gaps need filling, BACs and YACs can be used:
- **Bacterial Artificial Chromosome (BAC).** BAC vectors can accommodate up to 300-kb size DNA fragments. These large fragments are usually generated by restriction endonuclease digestion, separated on a pulse field gel (Sandaa et al. 2010, this volume), excised and gel purified. Electroporation is the transformation method of choice when creating BAC libraries.
- **Yeast Artificial Chromosome (YAC):** YAC vectors can accommodate similar size ranges as BAC vectors; however, yeast cells can offer some advantages over cloning in bacterial cells. DNA fragments containing repeat sequences are difficult to propagate in bacterial cells because prokaryotes do not have such extensive DNA elements in their genomes. Since yeasts are eukaryotes, they tolerate such sequences better. This is an important point as large dsDNA viruses contain extensive repetitive repeats (Schroeder et al. 2009; Wilson et al. 2005), a problem not always easy to resolve (Delaroque et al. 2003).
- **Metadata and MIGS.** All authors of genome sequence data must consider the corresponding metadata associated with the organism or virus that is being sequenced. This requires standardization of the way the data are collected to make it useful for downstream post-genomic analyses e.g., comparative genomics. To address the issues surrounding development of better metadata descriptions of genomic investigations (including whole genome sequencing and metagenomics), the Genomic Standards Consortium (GSC) (http://gensc.org/gc_wiki/index.php/Main_Page) was recently formed. GSC introduced the Minimum Informa-

tion about a Genome Sequence (MIGS) specification (Field et al. 2008), an ongoing process that has the intent of promoting community participation in its development and discussing the resources that will be required to develop improved mechanisms of metadata capture and exchange. It is worth checking the GSC Web portal (URL above) for continuous updates in MIGS implementation. Example metadata required for MIGS compliancy include (but is not limited to) environmental parameters (e.g., location of isolation, physicochemical parameters, type of habitat), pathogenicity parameters; propagation conditions, treatment during collection, nucleic acid extraction procedures, and sequencing procedures (e.g., see example MIGS compliant report in Figure 1).

An excellent example of the application of MIGS and use of metadata for comparative analysis and annotation of all publicly available genomes (including viruses) can be found at the Integrated Microbial Genomes (IMG) Web portal (<http://img.jgi.doe.gov/cgi-bin/w/main.cgi>) (Markowitz et al. 2007). It is a user-friendly interface allowing navigation of genome data along its three key dimensions (genes, genomes, and functions), and groups together the main comparative analysis tools.

- **Annotation.** This is often the most intimidating aspect of the process—making sense of all that sequence! Arguably the main reason for the anxiety is the enormity of the task at hand. To alleviate this unwelcome uncertainty of “will I be able to recognize any of the genes?” or “what if I miss genes?” in my new uncharted virome, the ever-growing field of bioinformatics has come to our rescue—or has it? Unfortunately, bioinformatics has created its own share of mayhem and confusion, i.e., what software package is best? As pointed out at the start of this chapter, “leave it to the professionals,” i.e., get someone onboard who has some experience in this area. That said, if you go back to the basics, the task will not be as intimidating as first thought. A good start is a book called *Bioinformatics for Dummies* (Claverie and Notredame 2003), which as its cover states is “a painless and thorough introduction to the field.”

As novices, you will require a software package that can:

- **Identify open reading frames (ORFs).** This is classically defined as a string of sequence starting with a start (AUG) and ending with a stop codon (UAG, UAA, or UGA). As some organisms do not exclusively recognize AUG as a start codon, you might want your software to allow you to look at ORFs between two stop codons.
- **View relevant features in all six possible frames.** This allows you to quickly see where the larger ORFs are located on the genome.
- **Upon selecting an ORF, provide a predicted amino acid sequence.** This is an important feature as it gives you the flexibility to search (BLAST) Web-based genomic databases (NCBI, EBI, etc.) for DNA or protein homologs (though you do not need a putative amino acid sequence

-> genome catalogue	
gcat identifier	000026_GCAT
genome report title	Emiliana huxleyi Virus 86
publication	PMID:12209309
->NCBI Genome Projects Database	
taxonomic group	Virus
taxid	181082
accession	AJ890364
relevance	Giant algal virus
ncbi organism name	Emiliana huxleyi virus 86
ncbi status	complete
migs	
->Investigation	
->->Study	
Nucleic Acid ("organism")	
subspecific genetic lineage	none reported
number of replicons	1
source material identifiers	
collection	Plymouth Virus Collection
source material identifier	EhV-86
specific host (taxid)	280463
host specificity and range (taxid)	2903
Health/disease status of specific host at time of collection	Crashed; Virus isolated from a crashed Emiliana huxleyi bloom
normally pathogenic or not	lytic
phenotype	
propagation	lytic
encoded traits	phosphate permease

encoded traits	RNA polymerase
encoded traits	DNA polymerase
encoded traits	Sphingolipid biosynthesis
environment	
geographical location	
latitude	50° 13.79' N
longitude	4° 9.59' W
depth or altitude	15 metres
time of sample collection	1999-07-27
habitat type	marine
Assay	
sample processing	
Isolation and Growth conditions	PMID: 12209309
Nucleic acid preparation	
nucleic acid extraction method	Phenol/Chloroform
dna amplification method	Not applied
sequencing method	dideoxysequencing
data processing	
assembly	
assembly method	phrap
estimated error rate	1 error within 10,000 bases
method of calculation	estimation from phred quality scores

Fig. 1. Example MIGS-compliant report for EhV-86. Adapted from supplementary table supplied with Field et al. (2008). Modified screenshots taken from the Genome Catalogue (<http://gensc.sf.net>). Read-only information is imported from the NCBI Genome Projects Database and the Genomes Online Database (<http://www.genomesonline.org>) to place each record into context. Values are only given for fields in MIGS marked as "minimal" (M, or mandatory) although more information may be available online. Figure reproduced by permission of M. Allen.

to conduct homology searches, NCBI will allow you to do all permutations).

- Give you these aforementioned features and can perform all these functions in real time, i.e., easily accessible (point and click), integrated software with automated tools and user friendly output formats.

We have personal experience with two software packages, namely Artemis (www.sanger.ac.uk/Software/Artemis/) and DNASTAR (www.dnastar.com) (Schroeder et al. 2009; Wilson et al. 2005). Both are more than suited for custom annotation of viromes, as are many other packages in the market place, so the choice is ultimately up to the user. An alternate strategy for ORF identification is to use automated software such as AMIGene (www.genoscope.cns.fr/agc/tools/amiga/form.php), GLIMMER (www.cbcu.edu/software/glimmer/), and GeneMark (<http://exon.biology.gatech.edu/>)—all of which have been used to annotate the latest giant virus, *Feldmannia* sp. virus 158 (Schroeder et al. 2009).

An important resource to BLAST against is the increasing volume of virus metagenomics data that is now available. Without doubt, metagenomics has revolutionized the study of microbiology and has revealed an incredible amount of

genetic diversity, particularly in the marine environment, and in particular among virus communities. The various methodologies of metagenomics have been discussed and reviewed extensively (Allen and Wilson 2008; Delwart 2007; Hall 2007; Handelsman 2004; Kunin et al. 2008; Riesenfeld et al. 2004; Streit and Schmitz 2004; Tringe and Rubin 2005; Wommack et al. 2008). A useful starting point for metagenomics data can be found at the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) Web portal (<http://camera.calit2.net/>). CAMERA is making raw environmental sequence data accessible along with associated metadata, pre-computed search results, and high-performance computational resources.

- Verification of functional assignment (annotation)
 - I) From ORFs to coding sequences (CDS) with putative function:
- Phylogeny. ORFs that have a BLAST homolog is likely to be coding for a gene of a putative function, i.e., coding sequences (CDS). CDS homologs identified by the various BLAST outputs should only be considered as indicators of possible function. An important first step in verify puta-

tive gene identities is by performing a phylogenetic analysis. This sort of analysis adds weight to the BLAST analysis by identifying its closest neighbor and whether it groups with rigorously, well-characterized, and peer reviewed homologs. Many phylogenetic software are freely available for use (e.g., PHYLIP—<http://evolution.genetics.washington.edu/phylip.html>). The National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/) provides a phylogenetic analysis for BLAST hit search outputs at a click of a button. Alternatively, specific Web pages have dedicated links for identifying virus CDSs based on phylogenetic profiling (www.igs.cnrs-mrs.fr/phydbac/Mimi/indexvirus.html).

II) From CDS to Gene:

- Reverse transcriptase PCR (RT-PCR): The ability to detect the mRNA of a CDS is an important validation step, i.e., proof that the CDS is being expressed, of it performing a particular function. Many commercial kits such as those provided by Quantace (Sensimix™—www.quantace.com/country.asp) and Qiagen (OneStep RT-PCR—[http://www1.qiagen.com/Products/Pcr/QiagenReverse Transcriptases/OneStepRtPcr.aspx](http://www1.qiagen.com/Products/Pcr/QiagenReverse%20Transcriptases/OneStepRtPcr.aspx)) allow for rapid detection of CDS target. Other uses of this technology are to link levels of expression with time point of infection or infectious phenotype, i.e., quantitative PCR or real time RT-PCR). Quantitative PCR utilizes the five or so cycles where DNA molecules are synthesized logarithmically from scarcely detectable to the log-linear phase. Therefore, the analysis is characterized by the threshold cycle (Ct)—the sooner the threshold is reached the higher the starting number of target molecules. Using standard curve samples (DNA for DNA molecules and RNA for RNA molecules) with known concentrations, it is possible to determine the copy number of the molecule in question.
- Microarrays. RT-PCR looks at the expression of one gene at a time. Microarrays can carry thousands of gene-specific probes to detect multiple targets in a single sample (Allen and Wilson 2006; Allen et al. 2010, this volume). Sample cDNA is hybridized to a platform (e.g., a microscope slide) containing spots of DNA (60-70-mer oligonucleotide probes, each diagnostic for a target of interest). In transcriptional microarrays, positive hybridization indicates up-regulation of a gene, hence confirming transcription of a CDS. As an example, microarrays were employed in the annotation of the EhV-86 genome (Wilson et al. 2005). Functional information from preliminary expression results can be used to determine correct reading frames for disputed CDSs. In addition, it can be used to help to identify new and unannotated CDSs. The primary use of the microarray is to assign virus transcripts into kinetic classes with the distinct aim of helping to determine the function of coordinately expressed genes with no database homologues (Allen et al. 2006a). An opportunistic use of the microarray is to use it as a tool for

genome diversity analysis (Allen et al. 2007). Very simply, the array can be used as a hybridization tool to determine presence or absence (or highly divergent) of genes in genomes of related coccolithoviruses. Rather than focusing on a single gene, the microarray will allow the formation of a diversity index based on whole genomes without the need to sequence these genomes. This can help to reveal core coccolithovirus genes and identify variable and absent genes between coccolithovirus genomes (Allen et al. 2006c).

III) From Gene to Protein:

- Proteomics. Analysis and characterization of the complete set of proteins (proteome) of a virus is one way to determine if structural genes are eventually translated into proteins (Allen et al. 2008; Clokie et al. 2008). Currently, the method is not commonly used for aquatic viruses, however it is a promising tool which combines 2D-gel electrophoresis followed by quantitative or semiquantitative mass spectrometry-based analysis of virus proteomes. Analysis of interactions between virus and host proteomes is leading to the new field of interactomics, an emerging area that uses genomic and proteomic tools to determine the full set of interactions between viruses and their hosts (Viswanathan and Fruh 2007).

Assessment and discussion

The fundamental basis of life is written in nucleic acid. For viruses, this code is written in a variety of forms, be it single- or double-stranded, RNA or DNA. To date, genomic analysis of viruses has had the biggest impact on the study of virus diversity. Regardless of the nature of the genome, the pinnacle of assessing any biological entity's genetic diversity is to sequence its entire genome, then crucially, compare it to other genomes to assess the magnitude of the changes. The best (and most comprehensive) assessment of virus diversity would be to sequence the entire genome of every virus on the planet. This, of course, is beyond the realms of possibility and plausibility. At the time of writing, the viral genomes page on NCBI (www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239) contained links to 3235 reference sequences for 2178 virus genomes and 41 reference sequences for viroids. The vast majority of these viruses are medically or agriculturally related, essentially reflecting the levels of funding for these important areas of research. Only a fraction (1%–2%) of the genomes are from aquatic viruses (Table 1). However, it is clear from this small but diverse collection of viruses that their hosts are equally diverse and include bacteria, archaea, algae, amoeba, invertebrates, and vertebrates, with virus genomes ranging in size from a few thousand bases to over a million bases.

These fully sequenced viruses represent only a minute fraction of the estimated 10^{31} viruses in aquatic environments, yet they have revealed a plethora of novelty and have altered our view of viruses as simple 'bags of genes'. It is common to identify genes involved in core virus functions such as RNA polymerase, DNA polymerase, and structural proteins, yet it

Table 1. Selection of viruses of aquatic origin listed at NCBI that have completed genomes (as of February 2009). NB. The list is not exhaustive, the focus is on viruses that infect primary producers, and the microbial components of aquatic ecosystems. It is worth noting that descriptors for search terms are limited, and it is currently not possible to pull out virus genomes that are of aquatic (either marine or freshwater) origin. As increasing numbers of virus genomes are added to databases, it is important that appropriate meta-data, such as proposed by MIGS (Field et al. 2008), is made available to allow researchers to refine searches for specific groups of genomes. Although the aim of this table is to point the reader in the direction of aquatic virus reference genomes and a virus genome Web portal, the sequencing pipeline for each type of virus (i.e., RNA or DNA) is essentially the same once you have access to an adequate concentration of good quality nucleic acid. Acquiring enough high purity nucleic acid is arguably the biggest challenge of any aquatic virus sequencing project.

Virus	Family	Accession*	Size	Reference
DNA viruses				
Acanthamoeba polyphaga Mimivirus	Mimiviridae	NC_006450	1,181,404	(Raoult et al. 2004)
Emiliana huxleyi virus 86	Phycodnaviridae	NC_007346	407,339	(Wilson et al. 2005)
Paramecium bursaria Chlorella virus NY2A	Phycodnaviridae	NC_009898	368,683	(Fitzgerald et al. 2007b)
Paramecium bursaria Chlorella virus AR158	Phycodnaviridae	NC_009899	344,691	(Fitzgerald et al. 2007b)
Ectocarpus siliculosus virus 1	Phycodnaviridae	NC_002687	335,593	(Delaroque et al. 2001)
Paramecium bursaria Chlorella virus 1	Phycodnaviridae	NC_000852	330,743	(Li et al. 1997)
Paramecium bursaria Chlorella virus FR483	Phycodnaviridae	NC_008603	321,240	(Fitzgerald et al. 2007a)
Paramecium bursaria Chlorella virus MT325	Phycodnaviridae	DQ491001	314,335	(Fitzgerald et al. 2007a)
Shrimp white spot syndrome virus	Nimaviridae	NC_003225	305,107	(Yang et al. 2001)
Cyprinid herpesvirus 3	Herpesviridae	NC_009127	295,146	(Aoki et al. 2007)
Acanthocystis turfacea Chlorella virus 1	Phycodnaviridae	NC_008724	288,047	(Fitzgerald et al. 2007c)
Prochlorococcus phage P-SSM2	Myoviridae	NC_006883	252,401	(Sullivan et al. 2005)
Synechococcus phage S-PM2	Myoviridae	NC_006820	196,280	(Mann et al. 2005)
Crocodilepox virus	Poxviridae	NC_008030	190,054	(Afonso et al. 2006)
Ostreococcus virus OsV5	Phycodnaviridae	NC_010191	185,373	(Derelle et al. 2008)
Synechococcus phage syn9	Myoviridae	NC_008296	177,300	(Weigle et al. 2007)
Prochlorococcus phage P-SSM4	Myoviridae	NC_006884	178,249	(Sullivan et al. 2005)
Microcystis phage Ma-LMM01	Myoviridae	NC_008562	162,109	(Yoshida et al. 2008)
Feldmannia species virus	Phycodnaviridae	NC_011183	154,641	(Schroeder et al. 2009)
Thermus phage P23-45	Siphoviridae	NC_009803	84,201	(Minakhin et al. 2008)
Synechococcus phage P60	Podoviridae	NC_003390	47,872	(Chen and Lu 2002)
Cyanophage Syn5	Podoviridae	NC_009531	46,214	(Pope et al. 2007)
Vibriophage VpV262	Podoviridae	NC_003907	46,012	(Hardies et al. 2003)
Prochlorococcus phage P-SSP7	Podoviridae	NC_006882	44,970	(Sullivan et al. 2005)
Archaeal BJ1 virus	Siphoviridae	NC_008695	42,271	(Pagaling et al. 2007)
Phormidium phage Pf-WMP4	Podoviridae	NC_008367	40,938	(Liu et al. 2007)
Roseobacter phage SIO1	Podoviridae	NC_002519	39,898	(Rohwer et al. 2000)
Pseudoalteromonas phage PM2	Corticoviridae	NC_000867	10,079	(Mannisto et al. 1999)
Penaeus merguensis densovirus	Parvoviridae	NC_007218	6,321	(Sukhumsirichart et al. 2006)
RNA Viruses				
Beluga Whale Coronavirus SW1	Coronaviridae	NC_010646	31,686	(Mihindukulasuriya et al. 2008)
Micromonas pusilla reovirus	Reoviridae	NC_008171 – NC_008181	25,563	(Attoui et al. 2006)
			segmented	
Dolphin morbillivirus	Paramyxoviridae	NC_005283	15,702	(Rima et al. 2005)
Salmon pancreas disease virus	Togaviridae	NC_003930	11,919	(Weston et al. 2002)
Chaetoceros tenuissimus RNA virus	Unclassified	AB375474	9,431	(Shirai et al. 2008)
Marine RNA virus JP-A	Seawater sample†	NC_009757	9,236	(Culley et al. 2007)
Marine RNA virus JP-B	Seawater sample†	NC_009758	8,926	(Culley et al. 2007)
Heterosigma akashiwo RNA virus SOG263	Marnaviridae	NC_005281	8,587	(Lang et al. 2004)
Seal picornavirus type 1	Picornaviridae	NC_009891	6,718	(Kapoor et al. 2008)
Chaetoceros salsugineum Nuc Incl virus	Unclassified	NC_007193	6,000	(Nagasaki et al. 2005b)
Marine RNA virus SOG	Seawater sample†	NC_009756	4,449	(Culley et al. 2007)
Heterocapsa circularisquama RNA virus	Unclassified	NC_007518	4,375	(Nagasaki et al. 2005a)

*Most data obtained from NCBI Genome www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239

†Complete genome sequence obtained from a metagenomic analysis of RNA extracted from seawater.

has becoming increasingly obvious that most viruses also harbor the ability to alter and manipulate host metabolism in highly specific ways to maximize the chance of successful infection. For example, genes involved in sphingolipid production, photosynthesis, and carbon metabolism have all been identified on virus genomes in recent years (Clokic and Mann 2006; Lindell et al. 2005; Lindell et al. 2004; Mann et al. 2005; Mann et al. 2003; Sandaa et al. 2008; Wilson et al. 2005; Yoshida et al. 2008). Complete genome sequencing projects are becoming commonplace, yet these projects are still restricted to relatively small numbers of isolates making comparative genomics of viruses still a young science. Preliminary comparative genomic analyses are making promising progress in helping to determine evolutionary and taxonomic status of certain groups of viruses (Allen et al. 2006c; Hendrix et al. 1999; Iyer et al. 2006; Rohwer and Edwards 2002). Full genome comparisons of aquatic viruses have been completed recently, particularly with algal virus and cyanophage genomes (Allen et al. 2006b; Delaroque et al. 2003; Fitzgerald et al. 2007a; Fitzgerald et al. 2007b; Fitzgerald et al. 2007c; Sullivan et al. 2005), this can help determine the selective advantage of containing certain genes, especially if a gene is missing in a closely related virus isolate. This allows downstream hypothesis testing with virus host systems to help determine the true function of observed differences between genomes.

Despite falling costs and the obvious scientific advantages of having complete coverage, it is often difficult for researchers to justify sequencing closely related viruses (Allen et al. 2006b; Delaroque et al. 2003). However, subtle differences at the genetic level can have profound effects on the infection success of closely related virus isolates. This diversity often remains hidden within a genome and is not immediately obvious until full genomic sequences become available.

A quick analysis of the size spectrum of genomes reveals a significant gap between the 407kb EhV-86 and the 1,181kb Mimivirus (Table 1). One explanation is simply lack of isolated representatives in this size range and here lies a methodological conundrum. Viruses in this genome size range are likely to be too large to pass through a 0.2 μm filter. Standard procedure includes passing a water sample through a 0.2 μm filter and looking for viruses in the filtrate. However, most large viruses will not be filterable through such a pore size, so even at this early sampling stage, researchers often introduce a size bias in their sampling strategy. Numerous Mimivirus sequence homologs have been identified in the Venter Sargasso Sea environmental database (Ghedini and Claverie 2005; Monier et al. 2008a; Monier et al. 2008b). Indeed, these authors suggest that their data are indicative of high concentrations of Mimivirus 'relatives' in the ocean. Only a concerted sampling effort to specifically isolate giant viruses in these environments will identify new giant viruses that will fill the gap at the top end of the genome size range.

Certainly the biggest dilemma when trying to choose new

viruses to sequence is trying to determine where to start and how to justify some viruses over others? Options for justification could be global importance of the host (clear implications for biogeochemical cycling and gene transfer processes); whether the host is sequenced (useful for downstream post-genomic analysis of virus-host interactions); extraordinary size of virus (help explain why the virus is so large); other virus relatives already sequenced (useful for comparative genomics and evolutionary determination); unusual host niche (does the genetic signature provide clues of a selective advantage in that niche) or exploitation opportunities (viruses from e.g., extreme environments). Whichever isolates are chosen, all will help to answer specific hypotheses as well as give novel information to help generate new questions and hypotheses for future projects.

Future advances will involve using tools such as microarrays, proteomics, and interactomics to help determine functionality of unknown genes. Sequence information should be considered as a starting point for asking questions and developing hypotheses about the role of viruses. It is an exciting new era for virus ecology and when used in combination with more traditional approaches, virus genomics will give us access to their ecological function on an unprecedented scale.

Comments and recommendations

Start with as high purity nucleic acid preparation as you can manage. Do not rush it, proceed with care. If your budget can stand it, get as much professional help as you can, particularly with bioinformatics. However, if your budget is limited, there is plenty of user friendly software now available to help build, annotate, and then compare sequence data. But do not give up. Most viruses sequenced to date have produced surprises in their genomes, and there are more to come from the other 99.9999% (or 10^{31}) viruses yet to be sequenced.

References

- Afonso, C. L., and others. 2006. Genome of crocodilepox virus. *J. Virol.* 80:4978-4991.
- Allen, M. J., and W. H. Wilson. 2006. The coccolithovirus microarray: an array of uses. *Briefings in Functional Genomics and Proteomics* 5:273-279.
- Allen, M. J., T. Forster, D. C. Schroeder, M. Hall, D. Roy, P. Ghazal, and W. H. Wilson. 2006a. Locus-specific gene expression pattern suggests a unique propagation strategy for a giant algal virus. *J. Virol.* 80:7699-7705.
- , D. C. Schroeder, A. Donkin, K. J. Crawford, and W. H. Wilson. 2006b. Genome comparison of two Coccolithoviruses. *Virol. J.* 3:15.
- , ———, M. T. G. Holden, and W. H. Wilson. 2006c. Evolutionary history of the Coccolithoviridae. *Mol. Biol. Evol.* 23:86-92.
- , J. Martinez-Martinez, D. C. Schroeder, P. J. Somerfield, and W. H. Wilson. 2007. Use of microarrays to assess viral diversity: from genotype to phenotype. *Environ. Microbiol.* 9:971-982.

- , J. A. Howard, K. S. Lilley, and W. H. Wilson. 2008. Proteomic analysis of the EhV-86 virion. *Proteome Sci.* 6:11.
- , and W. H. Wilson. 2008. Aquatic virus diversity accessed through omic techniques: A route map to function. *Curr. Opin. Microbiol.* 11:226-232.
- , B. Tiwari, M. E. Futschik, and D. Lindell. 2010. Construction of microarrays and their application to virus analysis, p. 34-56. *In* S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.], *Manual of Aquatic Viral Ecology*. ASLO.
- Aoki, T., and others. 2007. Genome sequences of three koi herpesvirus isolates representing the expanding distribution of an emerging disease threatening koi and common carp worldwide. *J. Virol.* 81:5058-5065.
- Attoui, H., F. M. Jaafar, M. Belhouchet, P. De Micco, X. De Lamballerie, and C. P. D. Brussaard. 2006. *Micromonas pusilla* reovirus: a new member of the family Reoviridae assigned to a novel proposed genus (*Mimoreovirus*). *J. Gen. Virol.* 87:1375-1383.
- Bentley, D. R., and others. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53-59.
- Chen, F., and J. R. Lu. 2002. Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl. Environ. Microbiol.* 68:2589-2594.
- Claverie, J. M., and C. Notredame. 2003. *Bioinformatics for dummies*. Wiley.
- Clokic, M. R. J., and N. H. Mann. 2006. Marine cyanophages and light. *Environ. Microbiol.* 8:2074-2082.
- , and others. 2008. A proteomic approach to the identification of the major virion structural proteins of the marine cyanomyovirus S-PM2. *Microbiology-Sgm* 154:1775-1782.
- Culley, A. I., A. S. Lang, and C. A. Suttle. 2007. The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. *Virology J.* 4:69.
- Delaroque, N., D. G. Muller, G. Bothe, T. Pohl, R. Knippers, and W. Boland. 2001. The complete DNA sequence of the *Ectocarpus siliculosus* virus EsV-1 genome. *Virology* 287:112-132.
- , W. Boland, D. G. Muller, and R. Knippers. 2003. Comparisons of two large phaeoviral genomes and evolutionary implications. *J. Mol. Evol.* 57:613-622.
- Delwart, E. L. 2007. Viral metagenomics. *Rev. Med. Virol.* 17:115-131.
- Derelle, E., and others. 2008. Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. *PLoS ONE* 3:e2250.
- Field, D., and others. 2008. The minimum information about a genome sequence (MIGS) specification. *Nat. Biotech.* 26:541-547.
- Fitzgerald, L. A., M. V. Graves, X. Li, T. Feldblyum, J. Hartigan, and J. L. Van Etten. 2007a. Sequence and annotation of the 314-kb MT325 and the 321-kb FR483 viruses that infect *Chlorella Pbi*. *Virology* 358:459-471.
- , ——, ——, ——, W. C. Nierman, and J. L. Van Etten. 2007b. Sequence and annotation of the 369-kb NY-2A and the 345-kb AR158 viruses that infect *Chlorella NC64A*. *Virology* 358:472-484.
- , ——, ——, J. Hartigan, A. J. P. Pfitzner, E. Hoffart, and J. L. Van Etten. 2007c. Sequence and annotation of the 288-kb ATCV-1 virus that infects an endosymbiotic *Chlorella* strain of the heliozoon *Acanthocystis turfacea*. *Virology* 362:350-361.
- Ghedini, E., and J. M. Claverie. 2005. Mimivirus relatives in the Sargasso Sea. *Virol J.* 2:62.
- Hall, N. 2007. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 210:1518-1525.
- Handelsman, J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68:669-685.
- Hardies, S. C., A. M. Comeau, P. Serwer, and C. A. Suttle. 2003. The complete sequence of marine bacteriophage VpV262 infecting *Vibrio parahaemolyticus* indicates that an ancestral component of a T7 viral supergroup is widespread in the marine environment. *Virology* 310:359-371.
- Hendrix, R. W., M. C. M. Smith, R. N. Burns, M. E. Ford, and G. F. Hatfull. 1999. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. U.S.A.* 96:2192-2197.
- Iyer, L. M., S. Balaji, E. V. Koonin, and L. Aravind. 2006. Evolutionary genomics of nucleocytoplasmic large DNA viruses. *Virus Res.* 117:156-184.
- Kapoor, A., and others. 2008. A highly divergent picornavirus in a marine mammal. *J. Virol.* 82:311-320.
- Kunin, V., A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. 2008. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* 72:557-578.
- Lander, E. S., and M. S. Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231-239.
- Lang, A. S., A. I. Culley, and C. A. Suttle. 2004. Genome sequence and characterization of a virus (HaRNAV) related to picorna-like viruses that infects the marine toxic bloom-forming alga *Heterosigma akashiwo*. *Virology* 320:206-217.
- Lawrence, J. E., and G. F. Steward. 2010. Purification of viruses by centrifugation, p. 166-181. *In* S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.], *Manual of Aquatic Viral Ecology*. ASLO.
- Li, Y., Z. Lu, L. Sun, S. Ropp, G. F. Kutish, D. L. Rock, and J. L. Van Etten. 1997. Analysis of 74 kb of DNA located at the right end of the 330-kb *Chlorella* virus PBCV-1 genome. *Virology* 237:360-377.
- Lindell, D., M. B. Sullivan, Z. I. Johnson, A. C. Tolonen, F. Rohwer, and S. W. Chisholm. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U.S.A.* 101:11013-11018.
- , J. D. Jaffe, Z. I. Johnson, G. M. Church, and S. W. Chisholm. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438:86-89.

- Liu, X. Y., M. Shi, S. L. Kong, Y. Gao, and C. C. An. 2007. Cyanophage Pf-WMP4, a T7-like phage infecting the freshwater cyanobacterium *Phormidium foveolarum*: Complete genome sequence and DNA translocation. *Virology* 366:28-39.
- Mann, N. H., A. Cook, A. Millard, S. Bailey, and M. Clokie. 2003. Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature* 424:741.
- , and others. 2005. The genome of S-PM2, a “photosynthetic” T4-type bacteriophage that infects marine *Synechococcus* strains. *J. Bacteriol.* 187:3188-3200.
- Mannisto, R. H., H. M. Kivela, L. Paulin, D. H. Bamford, and J. K. H. Bamford. 1999. The complete genome sequence of PM2, the first lipid-containing bacterial virus to be isolated. *Virology* 262:355-363.
- Markowitz, V. M., and others. 2008. The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.* 36(Database issue):D528-D533 [doi:10.1093/nar/gkm846].
- Mihindukulasuriya, K. A., G. Wu, J. S. Leger, R. W. Nordhausen, and D. Wang. 2008. Identification of a novel coronavirus from a beluga whale by using a panviral microarray. *J. Virol.* 82:5084-5088.
- Minakhin, L., and others. 2008. Genome comparison and proteomic characterization of *Thermus thermophilus* bacteriophages P23-45 and P74-26: Siphoviruses with triplex-forming sequences and the longest known tails. *J. Mol. Biol.* 378:468-480.
- Monier, A., J. M. Claverie, and H. Ogata. 2008a. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol.* 9:R106 [doi:10.1186/gb-2008-9-7-r106].
- , J. Larsen, R.-A. Sandaa, G. Bratbak, J.-M. Claverie, and H. Ogata. 2008b. Marine mimivirus relatives are probably large algal viruses. *Virol. J.* 5:12.
- Nagasaki, K., Y. Shirai, Y. Takao, H. Mizumoto, K. Nishida, and Y. Tomaru. 2005a. Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. *Appl. Environ. Microbiol.* 71:8888-8894.
- , Y. Tomaru, Y. Takao, K. Nishida, Y. Shirai, H. Suzuki, and T. Nagumo. 2005b. Previously unknown virus infects marine diatom. *Appl. Environ. Microbiol.* 71:3528-3535.
- , and G. Bratbak. 2010. Isolation of viruses infecting photosynthetic and nonphotosynthetic protists, p. 92-101. *In* S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.], *Manual of Aquatic Viral Ecology*. ASLO.
- Pagalang, E., and others. 2007. Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. *BMC Genomics* 8:13.
- Pope, W. H., and others. 2007. Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: A “horned” bacteriophage of marine *Synechococcus*. *J. Mol. Biol.* 368:966-981.
- Raoult, D., and others. 2004. The 1.2-megabase genome sequence of mimivirus. *Science* 306:1344-1350.
- Riesenfeld, C. S., P. D. Schloss, and J. Handelsman. 2004. Metagenomics: genomic analysis of microbial communities. *Ann. Rev. Genet.* 38:525-552.
- Rima, B. K., A. M. J. Collin, and J. A. P. Earle. 2005. Completion of the sequence of a cetacean morbillivirus and comparative analysis of the complete genome sequences of four morbilliviruses. *Virus Genes* 30:113-119.
- Rohwer, F., A. Segall, G. Steward, V. Seguritan, M. Breitbart, F. Wolven, and F. Azam. 2000. The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnol. Oceanogr.* 45:408-418.
- , and R. Edwards. 2002. The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol.* 184:4529-4535.
- Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242:84-89.
- Sandaa, R. A., M. Clokie, and N. H. Mann. 2008. Photosynthetic genes in viral populations with a large genomic size range from Norwegian coastal waters. *FEMS Microbiol. Ecol.* 63:2-11.
- Sandaa, R.-A., S. M. Short, and D. C. Schroeder. 2010. Fingerprinting aquatic virus communities, p. 9-18. *In* S. W. Wilhelm, M. G. Weinbauer, and C. A. Suttle [eds.], *Manual of Aquatic Viral Ecology*. ASLO.
- Schroeder, D. C., J. Oke, G. Malin, and W. H. Wilson. 2002. Coccolithovirus (*Phycodnaviridae*): Characterisation of a new large dsDNA algal virus that infects *Emiliania huxleyi*. *Arch. Virol.* 147:1685-1698.
- , and others. 2009. Genomic analysis of the smallest giant virus - *Feldmannia* sp. virus 158. *Virology* 384:223-232.
- Shendure, J., and others. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728-1732.
- Shirai, Y., Y. Tomaru, Y. Takao, H. Suzuki, T. Nagumo, and K. Nagasaki. 2008. Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus* Meunier. *Appl. Environ. Microbiol.* 74:4022-4027.
- Streit, W. R., and R. A. Schmitz. 2004. Metagenomics—the key to the uncultured microbes. *Curr. Opin. Microbiol.* 7:492-498.
- Sukhumsirichart, W., P. Attasart, V. Boonsaeng, and S. Panyim. 2006. Complete nucleotide sequence and genomic organization of hepatopancreatic parvovirus (HPV) of *Penaeus monodon*. *Virology* 346:266-277.
- Sullivan, M. B., M. L. Coleman, P. Weigele, F. Rohwer, and S. W. Chisholm. 2005. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biol.* 3:790-806.
- Tringe, S. G., and E. M. Rubin. 2005. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* 6:805-814.
- Viswanathan, K., and K. Fruh. 2007. Viral proteomics: global

- evaluation of viruses and their interaction with the host. *Expert Rev. Proteomics* 4:815-829.
- Weigele, P. R., and others. 2007. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ. Microbiol.* 9:1675-1695.
- Weston, J., and others. 2002. Comparison of two aquatic alphaviruses, salmon pancreas disease virus and sleeping disease virus, by using genome sequence analysis, monoclonal reactivity, and cross-infection. *J. Virol.* 76:6155-6163.
- Wilson, W. H., and others. 2005. Complete genome sequence and lytic phase transcription profile of a coccolithovirus. *Science* 309:1090-1092.
- Wommack, K. E., J. Bhavsar, and J. Ravel. 2008. Metagenomics: Read length matters. *Appl. Environ. Microbiol.* 74:1453-1463.
- Yang, F., J. He, X. H. Lin, Q. Li, D. Pan, X. B. Zhang, and X. Xu. 2001. Complete genome sequence of the shrimp white spot bacilliform virus. *J. Virol.* 75:11811-11820.
- Yoshida, T., and others. 2008. Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. *J. Bacteriol.* 190:1762-1772.