Virology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline

Melissa B. Duhaime [a], Matthew B. Sullivan [b,*]

[a] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA
[b] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

ARTICLE INFO

Keywords:
Ocean viruses
Metagenomics
viral concentration
viral purification
Next-generation sequencing
Environmental virology
Viral ecology

ABSTRACT

As new environments are studied, viruses consistently emerge as important and prominent players in natural and man-made ecosystems. However, much of what we know is built both upon the foundation of the culturable minority and using methods that are often insufficiently ground-truthed. Here, we review the modern culture-independent viral metagenomic sample-to-sequence pipeline and how next-generation sequencing techniques are drastically altering our ability to systematically and rigorously evaluate them. Together, a series of studies quantitatively evaluate existing and new methods that allow—even for ultra-low DNA samples—the generation of replicable, near-quantitative datasets that maximize inter-comparability and biological inference.

Introduction

The realization that ocean viruses are abundant (Bergh, 1989; Proctor and Fuhrman, 1990) and diverse (Angly et al., 2006) has fueled the rapidly growing field of "viral ecology" (Fuhrman, 1999; Wilhelm and Suttle, 1999; Wommack and Colwell, 2000; Weinbauer, 2004; Weinbauer and Rassoulzadegan, 2004; Suttle, 2005; Breitbart et al., 2007; Brussaard et al., 2008; Rohwer and Thurber, 2009). Broadly, viral ecology seeks to understand how the distribution of viruses and their genes impact a given host or ecosystem. Consistent with the universal tenets of ecology, this requires quantitative rigor as we attempt to track viral populations through space and time, quantify their impacts on measurable processes, and evaluate the underlying changes in genetic capacity of both virus and host. Because most (~85–99%) microbes remain resistant to routine cultivation techniques (Connon and Giovannoni, 2002; Rappe and Giovannoni, 2003), and not all viruses are easily cultured (Edwards and Rohwer, 2005), viral ecologists have relied heavily upon culture-independent techniques to understand the ecological role of viruses in nature. With rare exception, these culture-independent techniques are insufficiently validated for quantitative rigor due to a lack of isolates in culture collections, financial resources and tools needed to track viruses at the scales required to document naturally occurring viral diversity. Further notable, for a virology audience, is that the work to date is almost exclusively focused on double-stranded DNA viruses or phages.

Initially, culture-independent viral ecology utilized marker-based genetic diversity studies (Chen and Suttle, 1996; Fuller et al., 1998; Breitbart et al., 2004; Dorigo et al., 2004; Millard et al., 2004; Breitbart and Rohwer 2005; Zeidner et al., 2005; Sullivan et al., 2006; Sharon et al., 2007; Chenard and Suttle, 2008; Comeau and Krisch, 2008; Sullivan et al., 2008; Goldsmith et al., 2011) to survey environmental virus samples using marker genes ranging from major capsid proteins and photosynthesis genes to phosphate-related genes (e.g., *phoH*, though, notably, while originally thought to be phosphate stress related, this gene may have a different function, as summarized in Sullivan et al. (2010)). These datasets, i.e., counts of gene sequences belonging to particular phylogenetic lineages, provide an overview of gene diversity and how the presence or absence of particular lineages might change over space and time, but they are unlikely to provide quantitative data on lineage-specific abundances. This is because such surveys rely upon highly degenerate primer sets that are designed from limited sequence databases and require surprisingly permissive annealing temperatures (e.g., 35°C) to obtain products. In fact, the potential for biases (summarized in Table 1) may have, for at least some of these markers, led to limited ecological relevance, as they fail to correlate with any measured environmental parameter (e.g., T4 phage g20 in Sullivan et al. (2008)). Such marker-based efforts coupled to quantitative PCR (Matteson et al., 2011; Short et al., 2011; Hewson et al., 2012) are likely to improve a researcher's chances of being quantitative in a natural setting, particularly where extensive sequence data are available for well-contextualized primer design that allows non-degenerate primer

* Corresponding author.
E-mail address: mbsulli@email.arizona.edu (M.B. Sullivan).

0042-6822/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.virol.2012.09.036

Table 1
Known causes of amplification (PCR) errors and biases.

Error/Bias	Experimental solutions	References
1. Stochastic events (amplification differences early in PCR, polymerase error, primer misannealing)	Limit PCR cycles, mix replicate PCRs	(Higuchi et al., 1993; Wagner et al., 1994; Kanagawa 2003)
2. Differential template amplification at each round	Limit PCR cycles	(Suzuki and Giovannoni 1996)
3. Heteroduplex formation	Reconditioning PCR	(Speksnijder et al., 2001; Thompson et al., 2002)
4. Incompletely extended primer	Limit PCR cycles	(Judo et al., 1998)
5. Template switching during DNA synthesis leading to chimeric amplicons	Limit PCR cycles	(Odelberg et al., 1995; Patel et al., 1996)
6. Differential primer binding among degenerate primers (G/C > A/T) leading to skewed template amplification	Primer design	(Polz and Cavanaugh 1998)
7. Cycler ramp speed; if ramping is too steep (e.g., 6 C/s) there is strong bias against high % G+C regions	2.2 C/s found to be the optimal rate	(Aird et al., 2011)
8. Polymerase choice	Be cognizant of the influence, e.g., in one study, AccuPrime Taq HiFi decreased biases against high %G+C seen with Phusion HF; adding betaine may help	(Aird et al., 2011)
9. Incomplete denaturation	Elongate the initial denaturation step towards 3 m and the cycle denaturation towards 8s	(Aird et al., 2011)
10. Primer annealing temperature	Least % G+C biased when lowered from 72 to 65 C	(Aird et al., 2011)

sets to be used. To date, however, it has been financially and practically impossible to *quantitatively* evaluate these markers' PCR conditions across diverse target templates. New sequencing technologies and isolate collections change this.

As environmentally-relevant cultures became available, genomics partnered with experiments, modeling and metagenomics began to more comprehensively map out phage–host interactions in the environment, with a prominent example being virus-encoded photosynthesis genes (Mann et al., 2003). Cyanobacterial viruses (cyanophage) encode photosynthesis genes (Mann et al., 2003; Lindell et al., 2004; Millard et al., 2004; Sharon et al., 2009) that are expressed during infection (Lindell et al., 2005; Clokie et al., 2006; Dammeyer et al., 2008), aid phage fitness (Bragg and Chisholm, 2008; Hellweger, 2009), and impact photosystem evolution globally (Zeidner et al., 2005; Sullivan et al., 2006). Cyanophage genomics has highlighted other "auxiliary metabolic genes" (AMGs, Breitbart et al., 2007), including AMGs involved in scavenging commonly limiting nutrients like phosphate and nitrogen (Sullivan et al., 2005, 2010; Weigele et al., 2007; Millard et al., 2009). Metagenomic studies bring in the ability to document the ubiquity of these observations in the surface oceans (Dammeyer et al., 2008; Dinsdale et al., 2008; Williamson et al., 2008), while also revealing other AMGs that viruses may use to influence microbial metabolism (Sharon et al., 2011).

Indeed, metagenomics represents the best current means of documenting the taxonomic composition and genetic potential of uncultured virus communities. However, analytical tools are surprisingly lacking and viral metagenomic datasets are challenging to obtain. Tools are now emerging such as metaVir (Roux et al., 2011) and VIROME (Wommack et al., 2011), but for the most part, two datasets have, particularly in the oceans, been routinely utilized to provide ecological context for new genetic findings in viral ecology (e.g., 363 and 334 google scholar citations on 20th September 2012 for Angly et al. (2006), Dinsdale et al. (2008)). These datasets offer qualitative information that can powerfully help one evaluate ubiquity of a new gene or viral type in surface ocean viral communities. However, they are now known to suffer from a number of issues that render them non-quantitative with respect to taxon or gene abundances (artifacts summarized below).

Recent critique of the broad scientific endeavor warns of the increasing threat of the 'creeping cracks of bias' (Sarewitz 2012), as "science's internal controls on bias [are] failing, and bias and error [are] trending in the same direction towards the pervasive over-selection and over-reporting of false positive results." As viral ecologists, this warns us to take heed as we forge new territory applying emerging technologies to tackle age-old problems and theories. Comfortingly, modern sequencing and computational capabilities coupled with newly developed informatics provide the opportunity to bring rigor into what are likely to become foundational sequence-based methods in viral ecology. Here, we review a series of recent papers that provide a roadmap towards a nearly quantitative metagenomic sample-to-sequence toolkit for studying environmental virus communities.

From sample to metagenome, knowledgeably

The viral ecology metagenomic sample-to-sequence pipeline (overview in Fig. 1) is experimentally challenging at each step, as a sample progresses from concentration and purification of viral particles to amplification of the resulting DNA for sequencing preparation. However, enough knowledge is amassing that this pipeline is primed to advance to 'routine' use in environmental virology studies.

Concentration: At 10^6–10^8 viruses per ml, viruses are abundant with respect to other biological entities; yet, many lab-based assays (e.g., metagenomics, proteomics, cultivation of less abundant viruses) require concentration to obtain sufficient material. Tangential Flow Filtration (TFF) has served as the gold standard for over two decades (Suttle et al., 1991; Wommack et al., 2010), in spite of costly set-up requirements, as well as minimally repeatable and inefficient concentration success (Fuhrman et al., 2005; Colombet et al., 2007; Wommack et al., 2010; John et al., 2011). Recently, a new chemistry-based concentration method—$FeCl_3$-precipitation—has emerged that captures nearly all SYBR-stainable viral particles, is easy to implement, and requires a relatively inexpensive set of filters and chemicals (John et al., 2011). For these reasons, $FeCl_3$-precipitation is fastly becoming the standard viral concentration technique on global oceanographic research campaigns (e.g., Tara Oceans, Malaspina, LineP), as well as studies in bioreactors (e.g., EBPR sludge) and freshwater lakes (e.g., the Great Lakes in USA). New publications are forthcoming that will show the applicability of this technique to a variety of aquatic environments.

In a study designed to quantitatively evaluate biases of TFF and $FeCl_3$-precipitation concentration methods, DNA was extracted from 1080 L of viruses purified from the viral-fraction (particles < 0.2 μm) of a large-scale ocean analog, the Biosphere2 Ocean, and used to generate triplicate metagenomes from each concentration treatment. Analyses of these data showed that TFF

Please cite this article as: Duhaime, M.B., Sullivan, M.B., Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology (2012), http://dx.doi.org/10.1016/j.virol.2012.09.036

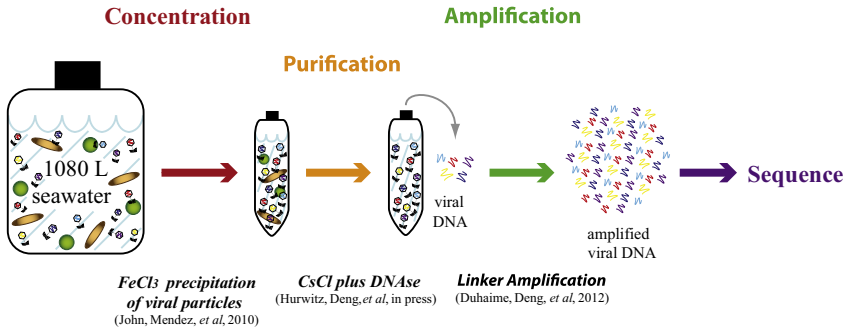

Fig. 1

concentrated viral metagenomes were prone to trace bacterial DNA contamination, with the viral-to-bacteria ratio of taxonomically assigned reads from the optimal $FeCl_3$ protocol were >2, while that of the TFF treatments was <1 (Hurwitz et al., in press). A finer taxonomic evaluation revealed that TFF concentrated viral metagenomes contained significantly fewer abundant viral types (*Podoviridae* and *Phycodnaviridae*) and more variable *Myoviridae* signals, as compared to their $FeCl_3$-precipitated counterparts (Hurwitz et al., in press). However, these taxonomic evaluations must be interpreted cautiously, as they are based upon the less than one-third of the data with database 'hits'. Alternative methods, such as 'protein clustering'—grouping predicted proteins by sequence similarity to use as discreet units, regardless of known function (Yooseph et al., 2007, 2008)—and 'shared k-mer'—relating sequences based on shared DNA words of 'k' length—analyses, use more or all of the data, respectively. These more comprehensive methods suggest the concentration efficiencies of $FeCl_3$ and TFF are likely broadly comparable. Notably, attention to TFF pore sizes (100 kDa was used, but smaller 30 kDa or 50 kDa pore sizes may have minimized the loss of *Podoviridae*), and choice of pre-filter (0.2 μm pre-filter was used, which may undersample the larger *Phycodnaviridae*, also observed in (Wommack et al. (2010) critically depends upon the target virus group.

Purification: Once a viral concentrate is obtained, the researcher must consider the nature of their research questions to make decisions about how to purify the viral particles away from co-concentrated materials. For example, viral ecologists commonly want to be able to link observations made in sequenced metagenomes back to the uncultured viruses, yet to do so requires confidence in being able to purify viral particles away from contaminating environmental or microbial DNA. To date, concentrated viral DNA preparations have been screened by amplifying with universal 16S microbial rDNA PCR primer sets to evaluate the presence of any contaminating microbial DNA prior to further purification, or the resulting metagenomes are examined by looking at genetic markers to estimate bacterial genome equivalents (McDaniel et al., 2008; Schoenfeld et al., 2008; Steward and Preston 2011). To this end, three purification procedures have become commonplace in viral ecology—DNAse alone, sucrose plus DNAse, and CsCl plus DNAse—and the decision about which method to use has largely been based on anecdotal observations.

A new study uses triplicate metagenomes made from the above 1080 L viral concentrate for each of these three purification methods to evaluate the impacts of purification (Hurwitz et al., in press). These analyses suggest that purification procedures resulted in metagenomes that were statistically indistinguishable at taxonomic levels of family, genus and species (taxonomy could only be assigned to ~30% of reads), suggesting that choice of purification method had much less impact than that of concentration method or polymerase used in amplification. Both 'protein clustering' and 'shared k-mer' analyses (again, both use more or all of the reads, respectively) suggested that any two samples within a treatment tended to share ~80% of the reads between them, differing predominantly at the level of 'rare' (k-mer=1) reads. Comfortingly, this suggests that purification method choice only minimally impacts the resulting viral metagenomic sequence data. Further, while the most repeatable (e.g., inferred virus-to-bacteria read ratio) metagenomes were those that were purified using DNAse plus CsCl, DNAse alone was comparable for two out of three replicates. This latter finding is promising as viral ecologists seek to expand beyond dsDNA viruses to those with buoyant densities outside the density range normally collected from CsCl purification gradients (e.g., ssDNA phages; Thurber et al., 2009; Holmfeldt et al., 2012).

Amplification and library construction: Though often an order of magnitude more abundant than the microbes they infect, the typical viral genome size is 1–2 orders smaller than their hosts'. As such, viral DNA needed for metagenomic sequencing library preparation is often limiting and must be amplified. Most virus metagenome sequencing projects to date have turned to either linker amplification shotgun libraries (LASLs) or whole genome amplification methods, e.g., multiple displacement amplification (MDA). These methods suffer from being prohibitively low-throughput, in the case of LASLs, or, in the case of MDA, are prone to systematic biases particularly relevant for viruses (e.g., selection for single-stranded and circular DNA templates; Kim et al., 2008; Kim and Bae, 2011), as well as stochastic biases (e.g., 100s–10,000 s-fold biases in coverage; Zhang et al., 2006; Woyke et al., 2009), which skew the taxonomic representation of a community in non-repeatable ways (Yilmaz et al., 2010).

New methods, such as linker amplification for deep sequencing (LADS; Hoeijmakers et al., 2011) and Nextera, have offered more quantitative amplification options for samples with limiting DNA (but note the documented Nextera bias against low % G+C content; Marine et al., 2011). However, these methods still may suffer from PCR amplification issues (see Table 1) or require too much DNA. Linker amplification of tightly sized products has long been considered relatively robust to amplification issues (Rohwer et al., 2001), as it is designed to avoid most of these known PCR biases. LADS requires 3–40 ng (Hoeijmakers et al., 2011) and Nextera >50 ng (Marine et al., 2011) of DNA, whereas a typical 20 L open ocean viral-fraction metagenomic sample might yield much less than this (~1 pg to 1 ng DNA). Clearly, further study is still needed.

Thus, by further optimizing and rigorously assessing Henn et al., 2010 existing high-throughput linker amplification (LA) technique, Duhaime et al. (2012) provide an option for such 'ultra-low' DNA samples. Specifically, using the above 1080 L seawater sample, replicate 454-sequenced viral genomes and metagenomes were generated and analyzed. Together, these data suggest LA to be highly replicable with minimal systematic biases, i.e., <1.5-fold

Please cite this article as: Duhaime, M.B., Sullivan, M.B., Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology (2012), http://dx.doi.org/10.1016/j.virol.2012.09.036

biases due to % G+C content (Duhaime et al., 2012). Thus, while not as high-throughput as some library prep methods (e.g., Nextera), the optimized LA method is now documented to provide precise, nearly-quantitative next-generation sequencing-ready DNA (1–5 μg) from sub-nanogram DNA amounts.

Though most extensively demonstrated for 454 sequencing, the LA method can also be used on other next-generation platforms, such as Illumina and Ion Torrent (Duhaime et al., 2012). For Ion Torrent, one need alter only shearing conditions to generate appropriately sized templates, as the barcodes do not pose a problem for this technology. For Illumina, the adaptations are more substantive, though commonly worked around, as there is a need to overcome problems with Illumina base-calling software due to the non-random nature of the barcodes present on LA (or any multiplexed) DNA. Strategies around this include (i) pooling with a known template, (ii) introducing a mechanism to cleave these non-random motifs for linker-free DNA (sensu Rodrigue et al., 2009), or (iii) introducing a string of four degenerate bases (i.e., "NNNN," sensu Bartram et al., 2011) between the Illumina sequencing primer and the LA primer (e.g., "Primer-A"; Duhaime et al., 2012), which targets the LA linker sequence and introduces a barcode when sample pooling is desired. As has been used for Illumina amplicon sequencing (Bartram et al., 2011; Caporaso et al., 2011), a possible modification to LA for optimal success with pooled samples may be the inclusion of a third sequencing primer, the "index sequencing primer", to ensure efficient indexing and minimal barcode loss. Finally, efforts by the DOE Joint Genome Institute to identify library construction inefficiencies are resulting in successful Illumina sequencing from DNA amounts approaching those common in environmental virology. Specifically, current amplification protocols successfully amplify less than 1 ng DNA in only five cycles, while unamplified libraries can now be made from as little as 25 ng of DNA (Chia-Lin Wei, personal communication).

Given the limiting nature of environmental sampling, improvements in amplification and library construction efficiency offer opportunity for substantive gain. We posit that the most promising areas for improvement are systematic evaluation of high-fidelity polymerases, as well as reducing DNA sizing losses and linker ligation inefficiencies in library construction.

Conclusions

Viruses appear critical to community dynamics and ecosystem function in any environment, yet remain the most understudied and mysterious component of microbial communities due to sampling, experimental and informatic challenges. Our tools are now approaching the quantitative rigor and throughput needed to map their myriad forms and functions–at least for double-stranded DNA viruses. Current studies aimed at evaluating inter-comparability of metagenomic datasets across myriad sequencing platforms and library preparation techniques, as well as those enabling access to single-stranded DNA and all RNA viruses (e.g., Andrews-Pfannkoch et al., 2010), will fill other critical voids in this toolkit. Looking forward, these efforts to develop quantitative rigor, along with emerging game-changing methods (e.g., Allen et al., 2011; Tadmor et al., 2011; Allers et al., submitted for publication; Deng et al., in preparation) and the required body of theory to interpret new data scales and types (e.g., Flores et al., 2011) will undoubtedly transform our ability to "see" in viral ecology. These advances portend a time when viral ecology will advance from a descriptive to a predictive science for the most abundant and likely most diverse biological entities on Earth.

Protocol availability

The most current versions of protocols for the $FeCl_3$-precipitation, the purification and LA methods discussed above are at http://eebweb.arizona.edu/Faculty/mbsulli/protocols.htm, maintained and complete with suggestions and updates from the scientific community.

Acknowledgments

Past and present Tucson marine phage lab members; Cindi Hoover, Rex Malmstrom, Susannah Tringe, Chia-Lin Wei and Tanja Woyke at the DOE Joint Genome Institute; and Maureen Coleman, Jed Fuhrman, Seth John, David Mead, Forest Rohwer, and Grieg Steward are thanked for years of discussion on evaluating and developing robust viral ecology methods. MBS acknowledges the Gordon and Betty Moore Foundation for funding, and two anonymous reviewers for their constructive suggestions and comments.

References

Aird, D., Ross, M.G., et al., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12, R18.
Allen, L.Z., Ishoey, T., et al., 2011. Single virus genomics: a new tool for virus discovery. PLoS One 6 (3), e17722.
Allers, E., Moraru, C., et al., Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses, Proc. Natl. Acad. Sci. USA, Submitted for publication.
Andrews-Pfannkoch, C., Fadrosh, D.W., et al., 2010. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. Appl. Environ. Microbiol. 76 (15), 5039–5045.
Angly, F.E., Felts, B., et al., 2006. The marine viromes of four oceanic regions. PLoS Biol. 4 (11), e368.
Bartram, A.K., Lynch, M.D., et al., 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. Appl. Environ. Microbiol. 77 (11), 3846–3852.
Bergh, O., 1989. High abundance of viruses found in aquatic environments. Nature 340, 467–468.
Bragg, J.G., Chisholm, S.W., 2008. Modelling the fitness consequences of a cyanophage-encoded photosynthesis gene. PLoS One 3, e3550.
Breitbart, M., Miyake, J.H., et al., 2004. Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol. Lett. 236 (2), 249–256.
Breitbart, M., Rohwer, F., 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13 (6), 278–284.
Breitbart, M., Thompson, L.R., et al., 2007. Exploring the vast diversity of marine viruses. Oceanography 20, 353–362.
Brussaard, C.P., et al., 2008. Global scale processes with a nano-scale drive: the role of marine viruses. ISME J. 2 (6), 575–578.
Caporaso, J.G., Lauber, C.L., et al., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. USA 108 (1), 4516–4522.
Chen, F., Suttle, C.A., 1996. Evolutionary relationships among large double-stranded DNA viruses that infect microalgae and other organisms as inferred from DNA polymerase genes. Virology 219 (1), 170–178.
Chenard, C., Suttle, C.A., 2008. Phylogenetic diversity of sequences of cyanophage photosynthetic gene psbA in marine and freshwaters. Appl. Environ. Microbiol. 74 (17), 5317–5324.
Clokie, M.R.J., Shan, J., et al., 2006. Transcription of a photosynthetic T4-type phage during infection of a marine cyanobacterium. Environ. Microbiol. 8, 827–835.
Colombet, J., Robin, A., et al., 2007. Virioplankton pegylation: use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. J. Microbiol. Methods 71 (3), 212–219.
Comeau, A.M., Krisch, H.M., 2008. The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. Mol. Biol. Evol. 25 (7), 1321–1332.
Connon, S.A., Giovannoni, S.J., 2002. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl. Environ. Microbiol. 68 (8), 3878–3885.
Dammeyer, T., Bagby, S.C., et al., 2008. Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. Curr. Biol. 18 (6), 442–448.
Deng, L., A. Gregory, et al., Contrasting strategies of viruses that infect photo- and hetero- trophic bacteria revealed by viral-tagging, in preparation.
Dinsdale, E.A., Edwards, R.A., et al., 2008. Functional metagenomic profiling of nine biomes. Nature 452 (7187), 629–632.
Dorigo, U., Jacquet, S., et al., 2004. Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, lake Bourget. Appl. Environ. Microbiol. 70, 1017–1022.

Please cite this article as: Duhaime, M.B., Sullivan, M.B., Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology (2012), http://dx.doi.org/10.1016/j.virol.2012.09.036

Duhaime, M., Deng, L., et al., 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. Environ. Microbiol. 14 (9), 2526–2537.
Edwards, R.A., Rohwer, F., 2005. Viral metagenomics. Nat. Rev. Microbiol. 3 (6), 504–510.
Flores, C.O., Meyer, J.R., et al., 2011. Statistical structure of host-phage interactions. Proc. Natl. Acad. Sci. USA 108, e288.
Fuhrman, J.A., 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399, 541–548.
Fuhrman, J.A., Liang, X., et al., 2005. Rapid detection of enteroviruses in small volumes of natural waters by real-time quantitative reverse transcriptase PCR. Appl. Environ. Microbiol. 71 (8), 4523–4530.
Fuller, N.J., Wilson, W.H., et al., 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 (6), 2051–2060.
Goldsmith, D.B., Crosti, G., et al., 2011. Development of phoH as a novel signature gene for assessing marine phage diversity. Appl. Environ. Microbiol. 77 (21), 7730–7739.
Hellweger, F.L., 2009. Carrying photosynthesis genes increases ecological fitness of cyanophage *in silico*. Environ. Microbiol. 11, 1386–1394.
Henn, M.R., Sullivan, M.B., et al., 2010. Analysis of high-throughput sequencing and annotation strategies for phage genomes. PLoS One 5 (2), e9083.
Hewson, I., Barbosa, J.G., et al., 2012. Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes. Appl. Environ. Microbiol. 78 (18), 6583–6591.
Higuchi, R., Fockler, C., et al., 1993. Kinetic PCR analysis—real-time monitoring of DNA amplification reactions. Bio-Technology 11 (9), 1026–1030.
Hoeijmakers, W.A., Bartfai, R., et al., 2011. Linear amplification for deep sequencing. Nat. Protoc. 6 (7), 1026–1036.
Holmfeldt, K., Odic, D., et al., 2012. Cultivated single-stranded DNA phages that infect marine bacteroidetes prove difficult to detect with DNA-binding stains. Appl. Environ. Microbiol. 78 (3), 892–894.
Hurwitz, B.H., Deng, L., et al. Evaluation of methods to concentrate and purify wild ocean virus communities through comparative, replicated metagenomics. Environ. Microbiol, http://dx.doi.org/10.1111/j.1462-2920.2012.02836.x, in press.
John, S.G., Mendez, C.B., et al., 2011. A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ. Microbiol. Rep. 3 (2), 195–202.
Judo, M.S.B., Wedel, A.B., et al., 1998. Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res. 26 (7), 1819–1825.
Kanagawa, T., 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). J. Biosci. Bioeng. 96 (4), 317–323.
Kim, K.H., Bae, J.W., 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl. Environ. Microbiol. 77 (21), 7663–7668.
Kim, K.H., Chang, H.W., et al., 2008. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl. Environ. Microbiol. 74 (19), 5975–5985.
Lindell, D., Jaffe, J.D., et al., 2005. Photosynthesis genes in marine viruses yield proteins during host infection. Nature 438 (7064), 86–89.
Lindell, D., Sullivan, M.B., et al., 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc. Natl. Acad. Sci. USA 101 (30), 11013–11018.
Mann, N.H., Cook, A., et al., 2003. Bacterial photosynthesis genes in a virus. Nature 424, 741.
Marine, R., Polson, S.W., et al., 2011. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl. Environ. Microbiol. 77 (22), 8071–8079.
Matteson, A.R., Loar, S.N., et al., 2011. Molecular enumeration of an ecologically important cyanophage in a Laurentian great lake. Appl. Environ. Microbiol. 77 (19), 6772–6779.
McDaniel, L., Breitbart, M., et al., 2008. Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. PLoS One 3, e3263.
Millard, A., Clokie, M.R., et al., 2004. Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. Proc. Natl. Acad. Sci. USA 101 (30), 11007–11012.
Millard, A.D., Zwirglmaier, K., et al., 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: Implications for mechanisms of cyanophage evolution. Environ. Microbiol. 11, 2370–2387.
Odelberg, S.J., Weiss, R.B., et al., 1995. Template-switching during DNA-synthesis by thermus-aquaticus DNA-polymerase-I. Nucleic Acids Res. 23 (11), 2049–2057.
Patel, R., Lin, C., et al., 1996. Formation of chimeric DNA primer extension products by template switching onto an annealed downstream oligonucleotide. Proc. Nat. Acad. Sci. USA 93 (7), 2969–2974.
Polz, M.F., Cavanaugh, C.M., 1998. Bias in template-to-product ratios in multitemplate PCR. Appl. Environ. Microbiol. 64 (10), 3724–3730.
Proctor, L.M., Fuhrman, J.A., 1990. Viral mortality of marine bacteria and cyanobacteria. Nature 343, 60–62.
Rappe, M.S., Giovannoni, S.J., 2003. The uncultured microbial majority. Ann. Rev. Microbiol. 57, 369–394.
Rodrigue, S., Malmstrom, R.R., et al., 2009. Whole genome amplification and de novo assembly of single bacterial cells. PLoS One 4 (9), e6864.
Rohwer, F., Seguritan, V., et al., 2001. Production of shotgun libraries using random amplification. Biotechniques 31 (1), 108–118.
Rohwer, F., Thurber, R.V., 2009. Viruses manipulate the marine environment. Nature 459, 207–212.
Roux, S., Faubladier, M., et al., 2011. Metavir: a web server dedicated to virome analysis. Bioinformatics 27 (21), 3074–3075.
Sarewitz, D., 2012. Beware the creeping cracks of bias. Nature 485 (7397), 149.
Schoenfeld, T., Patterson, M., et al., 2008. Assembly of viral metagenomes from yellowstone hot springs. Appl. Environ. Microbiol. 74 (13), 4164–4174.
Sharon, I., Alperovitch, A., et al., 2009. Photosystem I gene cassettes are present in marine virus genomes. Nature 461 (7261), 258–262.
Sharon, I., Battchikova, N., et al., 2011. Comparative metagenomics of microbial traits within oceanic viral communities. ISME J. 5 (7), 1178–1190.
Sharon, I., Tzahor, S., et al., 2007. Viral photosynthetic reaction center genes and transcripts in the marine environment. ISME J. 1 (6), 492–501.
Short, C.M., Rusanova, O., et al., 2011. Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada. ISME J. 5 (5), 810–821.
Speksnijder, A., Kowalchuk, G.A., et al., 2001. Microvariation artifacts introduced by PCR and cloning of closely related 16S rRNA gene sequences. Appl. Environ. Microbiol. 67 (1), 469–472.
Steward, G.F., Preston, C.M., 2011. Analysis of a viral metagenomic library from 200 m depth in Monterey Bay, California constructed by direct shotgun cloning. Virol. J. 8, 287.
Sullivan, M.B., Coleman, M., et al., 2005. Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol. 3 (5), e144.
Sullivan, M.B., Coleman, M.L., et al., 2008. Portal protein diversity and phage ecology. Environ. Microbiol. 10, 2810–2823.
Sullivan, M.B., Huang, K.H., et al., 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared to T4-like myoviruses from diverse hosts and environments. Environ. Microbiol. 12, 3035–3056.
Sullivan, M.B., Lindell, D., et al., 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. 4, e234.
Suttle, C.A., 2005. Viruses in the sea. Nature 437 (7057), 356–361.
Suttle, C.A., Chan, A.M., et al., 1991. Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. Appl. Environ. Microbiol. 57 (3), 721–726.
Suzuki, M.T., Giovannoni, S.J., 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. 62 (2), 625–630.
Tadmor, A.D., Ottesen, E.A., et al., 2011. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. Science 333 (6038), 58–62.
Thompson, J.R., Marcelino, L.A., et al., 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by reconditioning PCR. Nucleic Acids Res. 30 (9), 2083–2088.
Thurber, R.V., Haynes, M., et al., 2009. Laboratory procedures to generate viral metagenomes. Nat. Protoc. 4 (4), 470–483.
Wagner, A., Blackstone, N., et al., 1994. Surveys of gene families using polymerase chain reaction—PCR selection and PCR drift. Syst. Biol. 43 (2), 250–261.
Weigele, P.R., Pope, W.H., et al., 2007. Genomic and structural analysis of Syn9, a cyanophage infecting marine Prochlorococcus and Synechococcus. Environ. Microbiol. 9 (7), 1675–1695.
Weinbauer, M.G., 2004. Ecology of prokaryotic viruses. FEMS Microbiol. Rev. 28 (2), 127–181.
Weinbauer, M.G., Rassoulzadegan, F., 2004. Are viruses driving microbial diversification and diversity? Environ. Microbiol. 6 (1), 1–11.
Wilhelm, S.W., Suttle, C.A., 1999. Viruses and nutrient cycles in the sea. Bioscience 49 (10), 781–788.
Williamson, S.J., Rusch, D.B., et al., 2008. The Sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3 (1), e1456.
Wommack, K.E., Colwell, R.R., 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. 64, 69–114.
Wommack, K.E., Polson, S.W., et al., 2011. VIROME: a standard operating procedure for classification of viral metagenome sequences. Stand. Genomic Sci. 4, 427–439.
Wommack, K.E., Sime-Ngando, T., et al. (2010). MAVE: Filtration-based methods for the collection of viral concentrates from large water samples. Manual of Aquatic Viral Ecology. In: Wilhelm, S.W., Weinbauer, M.C., Suttle, A., Proceedings of ASLO, pp. 110–117.
Woyke, T., Xie, G., et al., 2009. Assembling the marine metagenome, one cell at a time. PLoS One 4 (4), e5299.
Yilmaz, S., Allgaier, M., et al., 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat. Methods 7 (12), 943–944.
Yooseph, S., Li, W.Z., et al., 2008. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. BMC Bioinformatics 9, 182.
Yooseph, S., Sutton, G., et al., 2007. The Sorcerer II global ocean sampling expedition: expanding the universe of protein families. PLoS Biol. 5 (3), e16.
Zeidner, G., Bielawski, J.P., et al., 2005. Potential photosynthesis gene recombination between Prochlorococcus and Synechococcus via viral intermediates. Environ. Microbiol. 7 (10), 1505–1513.
Zhang, K., Martiny, A.C., et al., 2006. Sequencing genomes from single cells by polymerase cloning. Nat. Biotechnol. 24 (6), 680–686.

Please cite this article as: Duhaime, M.B., Sullivan, M.B., Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. Virology (2012), http://dx.doi.org/10.1016/j.virol.2012.09.036