CHAPTER EIGHT

Preparation of Metagenomic Libraries from Naturally Occurring Marine Viruses

Sergei A. Solonenko[*], **Matthew B. Sullivan**[*,†,1]
[*]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, USA
[†]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, USA
[1]Corresponding author: e-mail address: mbsulli@email.arizona.edu

Contents

Abstract

Microbes are now well recognized as major drivers of the biogeochemical cycling that fuels the Earth, and their viruses (phages) are known to be abundant and important in microbial mortality, horizontal gene transfer, and modulating microbial metabolic output. Investigation of environmental phages has been frustrated by an inability to culture the vast majority of naturally occurring diversity coupled with the lack of robust, quantitative, culture-independent methods for studying this uncultured majority. However, for double-stranded DNA phages, a quantitative viral metagenomic sample-to-sequence workflow now exists. Here, we review these advances with special emphasis on the technical details of preparing DNA sequencing libraries for metagenomic sequencing from environmentally relevant low-input DNA samples. Library preparation steps broadly involve manipulating the sample DNA by fragmentation, end repair and adaptor ligation, size fractionation, and amplification. One critical area of future research and development is parallel advances for alternate nucleic acid types such

Methods in Enzymology, Volume 531
ISSN 0076-6879
http://dx.doi.org/10.1016/B978-0-12-407863-5.00008-3

as single-stranded DNA and RNA viruses that are also abundant in nature. Combinations of recent advances in fragmentation (e.g., acoustic shearing and tagmentation), ligation reactions (adaptor-to-template ratio reference table availability), size fractionation (non-gel-sizing), and amplification (linear amplification for deep sequencing and linker amplification protocols) enhance our ability to generate quantitatively representative metagenomic datasets from low-input DNA samples. Such datasets are already providing new insights into the role of viruses in marine systems and will continue to do so as new environments are explored and synergies and paradigms emerge from large-scale comparative analyses.

1. ON THE IMPORTANCE OF ENVIRONMENTAL VIRUSES AND VIRAL METAGENOMICS

Viruses infect all forms of life from the smallest microbes to the largest plants and animals. The outcomes of these infections can range from no discernible impact (some chronic or lysogenic infections) to death (lytic infections), but together viruses likely have profound impacts across all ecosystems on Earth as they number over $\sim 10^{31}$ planet-wide—approximately 10 times more viruses than prokaryotes (Wommack & Colwell, 2000). Particularly, well studied are marine bacterial viruses (phages) (Suttle, 2007), which kill ~20–40% of bacteria per day (Suttle, 2005; Weinbauer, 2004), move 10^{29} genes per day (Paul, 1999), and exist as a prophages within the genomes of about half the microbes at any given time (Paul, 2008). This implicates marine viruses in altering global biogeochemical cycling (the "viral shunt" keeps substrates from higher trophic levels, Fuhrman, 1999; Wilhelm & Suttle, 1999), structuring microbial communities (with most theory focused on "kill the winner," Thingstad, 2000; Weinbauer & Rassoulzadegan, 2004), and moving genes from one host to another, possibly driving microbial niche differentiation (e.g., Sullivan et al., 2006).

One phage–host system—cyanobacterial viruses (cyanophages) that infect abundant, marine *Prochlorococcus* and *Synechococcus* (Sullivan, Waterbury, & Chisholm, 2003)—has been relatively well studied due to its ecological importance and amenability to culturing. In fact, cyanophages harbor core "host" photosynthesis genes that are expressed during infection (Clokie, Shan, Bailey, Jia, & Krisch, 2006; Dammeyer, Bagby, Sullivan, Chisholm, & Frankenberg-Dinkel, 2008; Lindell, Jaffe, Johnson, Church, & Chisholm, 2005; Thompson et al., 2011), can recombine with

host copies to alter the evolutionary trajectory of their host's photosystems (Ignacio-Espinoza & Sullivan, 2012; Lindell et al., 2004; Sullivan et al., 2006), and are modeled to improve phage fitness by boosting photosynthesis during infection (Bragg & Chisholm, 2008; Hellweger, 2009). This "photosynthetic phage" paradigm demonstrates that an infected cell is intimately controlled by its viral predator and calls for deeper investigation to document other coevolutionary paradigms in representative model systems from the diversity of viruses and hosts in nature.

Problematically, however, the bulk of microbial hosts and their viruses have not yet been cultivated. In fact, 85% of 1100 genome-sequenced phages derive from only 3 of the 45 known bacterial phyla (Holmfeldt et al., 2013), and these statistics are worse for archaeal and eukaryotic hosts. This is changing as new marine phage–host systems emerge (Holmfeldt et al., 2013; Zhao et al., 2013). However, the disparity between known potential hosts and those in culture led environmental virologists to culture-independent methods (e.g., metagenomics) to survey natural viral communities. Environmental viral metagenomes preceded those of their microbial hosts by 2 years with the development of the linker-amplified shotgun library method (Breitbart et al., 2002; Schoenfeld et al., 2008; Tyson et al., 2004; Venter et al., 2004) and even inspired Norman Anderson (Viral Defense Foundation) and N. Leigh Anderson (Plasma Proteome Institute) to propose sequencing, cataloging, and tracking viruses in human blood to treat human disease (Anderson, Gerin, & Anderson, 2003). Such efforts have not yet been realized, but in the environmental sciences, application of viral metagenomics has indeed led to a number of important discoveries (Breitbart, 2012).

Environmental viral metagenomic studies over the past decade have revealed how little we know—the bulk of viral metagenomes are (Cesar Ignacio-Espinoza, Solonenko & Sullivan, 2013) or completely new to science (reviewed in Hurwitz & Sullivan, 2013)—but new biology has emerged including evidence for recombination between ssDNA and ssRNA viruses (Rosario, Duffy, & Breitbart, 2012), delineation of compositional differences between freshwater and marine viral communities (Roux, Krupovic, Poulet, Debroas, & Enault, 2012), and the discovery of novel and diverse auxiliary metabolic genes found in viral metagenomes (Sharon et al., 2011). More recent work expands the above "photosynthetic virus" paradigm from photosynthesis genes in cyanophages to diverse host metabolic genes in a majority of phages (Hurwitz, Hallam, & Sullivan, in review-a; Hurwitz & Sullivan, in review-b). This, in combination with decades-old coliphage studies, suggests that the metabolic output of an

uninfected cell drastically differs from that of a metabolically reprogrammed virus-infected cell. While few quantitative data are available, ocean virus–microbe interactions clearly impact the global carbon cycle, often dictating whether carbon in any individual microbial cell is sequestered to the deep ocean or released to the atmosphere through respiration of viral lysates (Fuhrman, 1999).

The challenge to developing a quantitative understanding of viral roles in ecosystems has been the lack of optimized tools to study viruses in a quantitative manner. For viral community sequence space, however, there is now an optimized, quantitative ocean viral metagenomic sample-to-sequence workflow (Fig. 8.1) that has been thoroughly evaluated using replicated metagenomic analyses to understand impacts of choices made in viral particle concentration and purification, nucleic acid amplification, and sequencing library preparation and platform choice (Duhaime, Deng, Poulos, & Sullivan, 2012; Duhaime & Sullivan, 2012; Hurwitz, Deng, Poulos, & Sullivan, 2013; John et al., 2011; Solonenko et al., 2013). This new quantitative data type has facilitated exciting discoveries, including uncovering the most abundant viruses in the oceans (Zhao et al., 2013) and advancing informatic solutions to organize unknown viral sequence space (*sensu* Yooseph 2007 protein clusters) (Hurwitz et al., 2013). This organization is tremendously powerful for viromic studies, as it helped reveal that the core Pacific Ocean virome (POV) is made of only 180 proteins, its pan-genome is relatively well sampled (~422k proteins), and the bulk of these proteins—even those core to all samples—are functionally

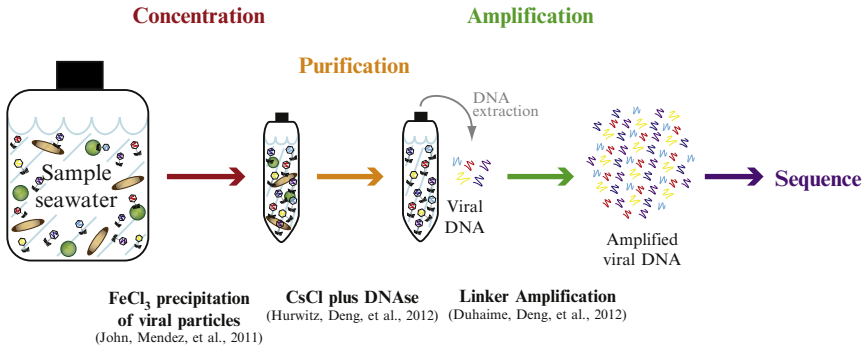

Figure 8.1 Overview of the environmental viral metagenomic sample-to-sequence workflow. The four basic steps in the creation of viral metagenomic data are illustrated, including references for suggested protocols for sequencing dsDNA viruses from marine samples. *Reprinted with permission from Duhaime and Sullivan (2012).*

unknown, but abundant, and presumably driving viral effects on ecosystem function. Further, the POV dataset has revealed that viral metabolic reprogramming extends far beyond cyanophage manipulation of photosynthesis, as it appears that Pacific Ocean viruses manipulate all of central microbial metabolism during infection, which profoundly alters our perception of viral roles in global carbon cycling. Specifically, Pacific Ocean virus gene content suggests that viral communities manipulate all starvation-related central metabolic pathways during infection in ways that could define viral niche space across hosts and the water column. Finally, protein clusters are powerful ecological inference tools. Specifically, they can (i) serve as a universal metric for comparing community viral diversity—something currently problematic due to reliance upon quantification derived from assembly output not yet tuned for metagenomic datasets—and (ii) offer a basis on which one can apply OTU-based ecological theory, independent of known function, using new and expanding community tools (e.g., QIIME).

Clearly, viral metagenomics will lead to myriad discoveries and, with careful optimization of the sample-to-sequence workflows, to help develop a more comprehensive understanding of the roles viruses play in the function of Earth's ecosystems.

2. THE DNA VIRAL METAGENOMIC SAMPLE-TO-SEQUENCE PIPELINE

Prior to constructing sequencing libraries, one needs to obtain a viral community concentrate and nucleic acids. This sample-to-sequence workflow (Fig. 8.1) is relatively well established now for double-stranded DNA (dsDNA) viruses and involves prefiltration to remove cellular material, concentration and purification of viral particles, and DNA extraction. While choice of prefilter is dependent upon environmental microbial concentrations and types, as well as the research questions being investigated, the remaining steps are now relatively well constrained (exceptions in the following paragraph) as follows. Viral particles are concentrated by $FeCl_3$ precipitation (John et al., 2011), with choice of purification (DNAse alone, DNAse + cesium chloride density gradient ultracentrifugation, or DNAse + sucrose density gradient ultracentrifugation) (Hurwitz et al., 2013), and the resulting limiting DNA (usually less than a few tens of nanograms) available for linker amplification techniques yielding metagenomes

that are ±1.5-fold biased by %G+C content (e.g., Duhaime et al., 2012), which sharply contrasts up to ±10,000-fold biases of phi29-based whole-genome amplification methods (Yilmaz, Allgaier, & Hugenholtz, 2010; Zhang et al., 2006), although this value for phi29-based amplification may be an overestimate, since the measurements were done under the challenging conditions of single-cell amplification.

Based upon SYBR Gold particle counts, the current sample-to-sequence workflow captures the vast majority of detectable viral particles. However, there remain issues and opportunities for research and development, particularly for studies needing to document less common phage types. These include the following: (i) very large viruses are problematic because the prefilters are either too small (0.2 μm) or else coselect many microbes (e.g., 0.8 or 0.45 μm), (ii) lipid-containing viruses may require tweaks to concentration and purification protocols, (iii) the current methods are optimized for dsDNA viruses. On this latter point, it is possible that RNA viruses are missed because RNA is not commonly extracted from viral concentrates, and ssDNA viruses are missed because we cannot detect these well by staining (Holmfeldt, Odic, Sullivan, Middelboe, & Riemann, 2012) and density gradients often select against them (Thurber, Haynes, Breitbart, Wegley, & Rohwer, 2009). Notably, however, some studies have enriched for ssDNA viruses using one of the inherent systematic biases of the phi29 whole-genome amplification enzyme (Kim & Bae, 2011; Kim et al., 2008).

The nucleic acid extraction step is particularly challenging for microbial samples and thought to be one of the largest sources of bias in microbial metagenomes (Morgan, Darling, & Eisen, 2010). However, this step is unlikely to be problematic for environmental viruses because microbes have incredible diversity in cell membranes resulting in highly variable accessibility of their DNA. In contrast, viruses use a relatively simple method for protecting their DNA—protein capsids—which lends itself to nearly universally effective DNA extraction protocols. Protocols to date have largely focused on extracting DNA from viral concentrates, but there are also methods available for studying RNA and ssDNA metagenomes (Culley, Lang, & Suttle, 2006; Filiatrault et al., 2010; Roux et al., 2012). In fact, recent work suggests that RNA viruses may represent half of the viruses in the oceans (Steward et al., 2013), and methods exist to simultaneously separate ssDNA, dsDNA, and RNA from the same viral sample (Andrews-Pfannkoch, Fadrosh, Thorpe, & Williamson, 2010). Clearly, viruses with other nucleic acid types are promising targets for exploration in the environment. However, we focus here on DNA viruses since the

sample-to-sequence pipeline is now well understood. Specifically, this chapter focuses on DNA library construction from natural viruses for metagenomic sequencing, including optimizations necessary for obtaining high-quality data from limiting DNA input amounts that are common to such samples.

3. THE LIBRARY PREPARATION PROCESS

Over the last decade, many variations in library preparation have emerged. However, the overall process is relatively constrained to manipulating the sample DNA by fragmentation, end repair and adaptor ligation, size fractionation, and amplification (Fig. 8.2).

3.1. Fragmentation

Obtaining the desired size of genomic DNA for sequencing library preparation requires fragmenting the DNA using a variety of options (summarized in Table 8.1 and detailed below). The overall goals of these methods are identical—to create fragments of the desired size while minimizing loss through efficient DNA recovery and narrowing the resulting fragment length distribution—but each method has strengths and weaknesses.

Traditional DNA fragmentation for genome sequencing projects was done using hydrodynamic shearing, nebulization, or enzymatic digestion, but these approaches have significant limitations for application to metagenomics. Nebulization mechanically breaks long DNA strands by forcing a nucleic acid solution through a narrow opening with varied air pressure. The advantages of nebulization are (i) random breakage with a relatively small fragment size range and (ii) no need for expensive equipment beyond pressurized air, while the disadvantages are (i) low throughput as only one sample can be fragmented per nebulizer and (ii) loss of up to 50% of total DNA which necessitates several micrograms of input DNA as starting material (Quail, 2010; Quail et al., 2008). Another mechanical shearing method, hydrodynamic shearing, uses the shear forces generated when repeatedly streaming a DNA sample through a narrow opening to generate large (>2 kb) and relatively tightly sized fragments, a great advantage for mate-pair protocols. As in nebulization, some material is lost, and a high sample minimum of several micrograms of DNA is required (see HydroShear Technical Brochure, 2009). Alternative to mechanical shearing, traditional protocols have used enzymatic digestion by endonucleases either with specific and known cleavage sites

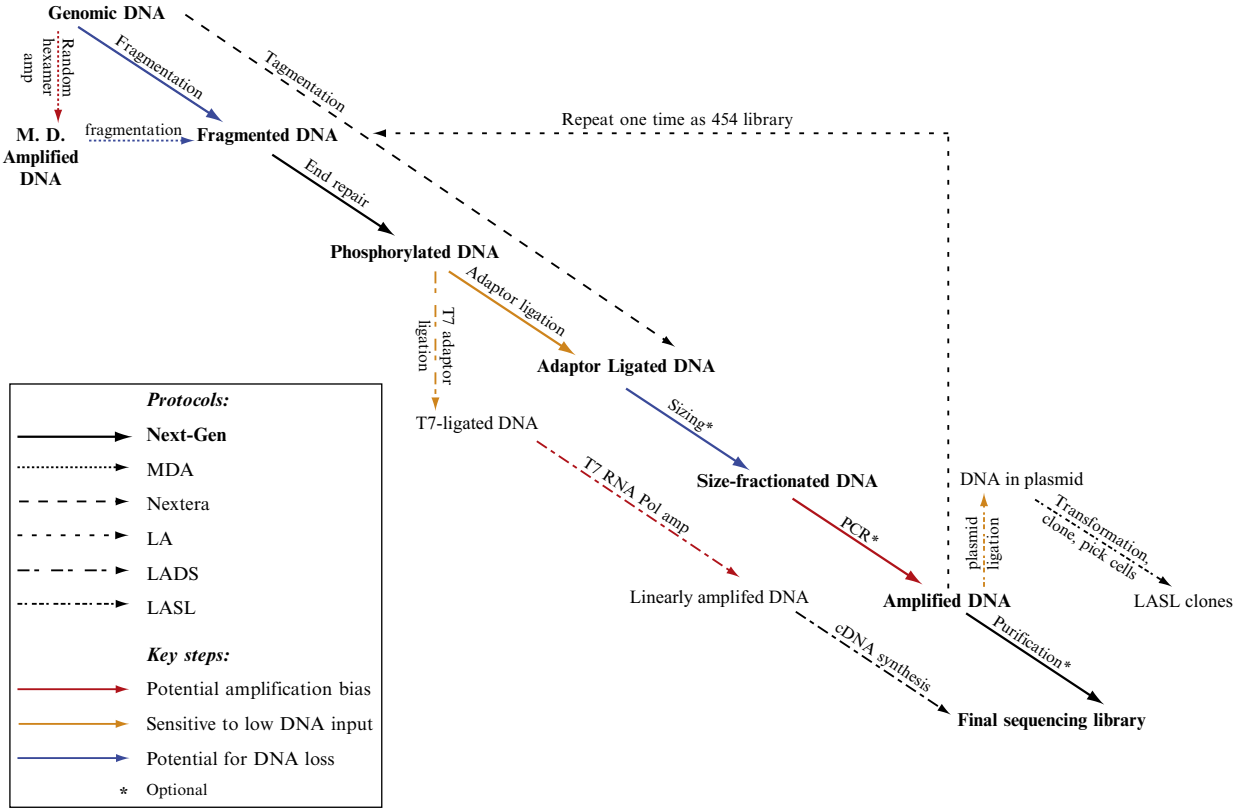

Figure 8.2 Schematic of common steps in next-generation sequencing library preparation. The methods represented include multiple displacement amplification (MDA, a phi29 whole-genome amplification method), an amplification of raw genomic DNA; linear amplification for deep sequencing (LADS), an alternative to PCR amplification for amplifying library DNA; linker-amplified shotgun library (LASL), a clone library protocol which shares many steps with sequencing library preparation; as well as several next-generation library construction methods, some of which may not require sizing, PCR, or purification (see Table 8.1). This figure highlights several steps in the procedure that are associated with issues that may impact the success or quality of the constructed library, in particular, amplification bias, ligation conditions, and choice of fragmentation method.

Table 8.1 A summary of several common library prep protocols available for Illumina, 454, and Ion Torrent sequencing systems

	Input DNA	Fragmentation method	DNA ends treatment	Ligation method	Adaptor type	Sizing method	Amplification	Sequencing	References
Illumina TruSeq	1 μg	Acoustic shear	End Repair & A-Tailing	T/A overhang	Y-adaptors	Gel extraction	Adaptor specific	Illumina	TruSeq Sample Prep Guide
454 GS FLX+	1 μg	Nebulization	End Repair	Blunt ended	Dual dsDNA or Y-adaptors	Bead	None	454	GS FLX+ Library Prep Manual
Ion Torrent	100 ng or 1 μg	Acoustic or enzymatic shear	End Repair	Blunt ended	Dual dsDNA	Gel extraction	Adaptor specific	Ion PGM	Ion Torrent Library Prep Manual
Multiple displacement amplification	1–100 ng	Endonuclease	LC Dependent	LC dependent	LC dependent	LC dependent	Random Hexamer	LC dependent	Yilmaz et al. (2010)
Linker-amplified library construction	>10 pg	Acoustic shear	End Repair	Blunt ended	Dual dsDNA	Bead or gel extraction	Adaptor specific	454	Duhaime et al. (2012)
Linker-amplified for deep sequencing	3–40 ng	Nebulization	End Repair & A-Tailing	T/A overhang	Identical dsDNA	Gel extraction	Transcription	Illumina	Hoeijmakers et al. (2011)

Continued

Table 8.1 A summary of several common library prep protocols available for Illumina, 454, and Ion Torrent sequencing systems—cont'd

	Input DNA	Fragmentation method	DNA ends treatment	Ligation method	Adaptor type	Sizing method	Amplification	Sequencing	References
Linker-amplified shotgun library	1 μg	HydroShear	End Repair	Blunt ended	Identical dsDNA	None	Adaptor specific	Sanger	Breitbart et al. (2002)
Nextera XT	1 ng	Simultaneous fragmentation and tagging			Dual dsDNA	Bead	Adaptor specific (limited cycle)	Illumina	Nextera XT Sample Prep Guide

DNA amounts refer to the recommended starting DNA necessary for the protocol (unsheared viral dsDNA). Four fragmentation options are represented across these protocols, but most are intercompatible except for the transposase, where fragmentation and adaptor attachment happen in one step. Adaptor types are Y-adaptor, which includes two separate adaptors that share a region of homology and form a Y structure during ligation, dual adaptors, two different adaptors ligated on either end of a genomic DNA fragment, and identical adaptors, where the same adaptor is ligated on both ends of the genomic DNA fragment. Some methods of attaching dual adaptors generate many adaptor combinations, requiring a purification/enrichment step to obtain properly ligated library fragments (ones with different adaptors on each end). MDA is done before fragmentation and is thus compatible with many different types of downstream sequencing preparation, with the affected steps marked as library construction (LC) dependent.

for controlled genomic DNA fragmentation or with more permissive cleavage sites for nonspecific shearing of DNA. Advantages of enzymatic digestion include (i) no need for equipment investment, (ii) random digestion (for nonspecific enzymes), (iii) marginally tunable sizing by adjusting the restriction reaction conditions, while the disadvantages are (i) nonrandom fragmentation (for specific cut-site restriction endonucleases), (ii) poor control for generating large fragments (e.g., NEB Fragmentase kit), and (iii) lower reproducibility (Adey et al., 2010; Linnarsson, 2010).

In contrast, newer library preparation protocols fragment DNA using acoustic shearing or tagmentation (Nextera kit, Illumina TruSeq kit, Duhaime et al., 2012). To generate fragmented DNA, acoustic shearing simply uses cavitation to randomly break up the DNA (Quail, 2010), while tagmentation combines fragmentation with adaptor attachment in one transposition reaction (Adey et al., 2010). These two methods pervade modern library protocols due to several desirable features. First, both can produce fragments with narrow size distributions that are optimal for short-read sequencing (e.g., 150–300 bp, Henn et al., 2010), which is not efficiently done with nebulization or enzymatic digestion (Quail et al., 2008). Notably, downstream sizing may not be needed for acoustic shearing but is required for tagmentation to remove small fragments where size distributions extend as low as 40 bp (Nextera XT manual; Adey et al., 2010). Second, acoustic shearing and tagmentation are high-efficiency methods: acoustic shearing because it incurs virtually no sample loss because it is performed in closed tubes, and tagmentation because it reduces sample manipulation. Third, acoustic shearing, in particular, has reduced chance of contamination because the entire process is done in a closed tube. Finally, both methods can be scaled for high-throughput work. For example, acoustic shearing can already be done in 96-well plate format and has recently been utilized in microfluidic applications (Tseng, Lomonosov, Furlong, & Merten, 2012), with development heading toward automated microfluidic μl-scale sequencing library preparation (Vyawahare, Griffiths, & Merten, 2010). The disadvantages of these methods are that acoustic shearing requires expensive equipment or fee-for-service access, while tagmentation leads to slight %G+C biases in genomes (Marine et al., 2011) and metagenomes (Solonenko et al., 2013), presumably due to insertion biases inherent to the transposase (Adey et al., 2010).

3.2. Insert size choices

Many library preparation options should be tuned to accommodate the type of sequence data best suited to the research question being addressed. For example, metagenomic sequencing has predominantly relied on data derived from a single sequencing read per DNA fragment. However, two sequencing reads per DNA fragment (paired-end sequencing) can be obtained by attachment of different sequencing adaptors to DNA fragment ends to allow directional sequencing off each end. This strategy can be used to provide longer "reads" for small-insert libraries where the two sequencing reads overlap each other. For large-insert libraries, such paired-end data can drastically increase metagenomic assembly contig sizes (e.g., Rodrigue et al., 2010). Several assembly algorithms use paired-end information for genome scaffolding, with Allpaths-lg (Gnerre et al., 2011), the most popular, and options in Velvet (Zerbino, McEwen, Margulies, & Birney, 2009), Abyss (Simpson et al., 2009), and SOAP-denovo (Luo et al., 2012) also available. Notably, these algorithms were designed for single genome assembly and have problems handling large differences in coverage (>100) present in metagenomic data, in which high coverage contigs may be mistaken for repeat regions or lead to misassembly due to heterogeneity, while low-coverage contigs may become overly fragmented due to low read overlap (Peng, Leung, Yiu, & Chin, 2012). Two recently published methods, IDBA-UD (Peng et al., 2012) and MetaVelvet (Namiki, Hachiya, Tanaka, & Sakakibara, 2012), address the above issues and are capable of analyzing metagenomic paired-end data, but either method has yet to be used on viral metagenomic data.

Currently, paired-end sequencing libraries are limited to small (<800 bp) insert sizes due to limitations in bridge amplification clustering (Illumina Paired End Sample Prep Guide, Rev. E., February 2011) and emPCR (GS FLX Ti General Preparation Method Manual, April 2009). One way to overcome this hard limit is by mate pairing (similar to long-range paired-end or paired-end tag libraries), whereby longer DNA fragments are circularized by ligating the two ends together and then fragmented down to <800 bp size compatible with paired-end library construction and sequencing. Current mate-pair library size limits are 40 kbp, with mate-pair creation efficiency decreasing and size distribution variation increasing as insert sizes increase (Asan et al., 2012). As well, a major reason why mate-pair libraries are not standard in environmental metagenomic surveys is the requirement for prohibitively large amounts of starting DNA, for

example, 50 μg for a 35-kb mate-pair library (Asan et al., 2012). Successful use of mate-pair data yields a new level of organization to metagenomic data (Iverson et al., 2012). While such quantities are currently impossible for viral metagenomic studies, there is potential for creative amplification-based solutions which could augment environmental DNA to the point where environmental virologists may also benefit from mate-pair data.

3.3. End repair and adaptor ligation: A key step in low-input DNA library construction

Fragmentation commonly results in ssDNA ends which require repair to prepare for dsDNA adaptor ligation (Table 8.1). In fact, end repair is part of every protocol except tagmentation, where the transposition reaction leaves no damage to DNA ends and includes addition of adaptors. Some protocols, such as Illumina and LADS, utilize A-tailing to create an overhang to which T-tailed adapter sequences are ligated so as to leverage improved efficiency over blunt-end ligation and prevent concatenation of template DNA (Bratbak, Wilson, & Heldal, 1996). However, because A-tailing adds another step to the procedure in which DNA may be lost (i.e., DNA binding to tubes, Ellison, English, Burns, & Keer, 2006), many protocols utilize blunt-end ligation for adding adaptor sequence to the fragments (Table 8.1).

The indispensable step in library preparation is the addition of adaptors to the genomic DNA fragments, which eventually act as a primer site during the sequencing reaction. Most protocols achieve this using ligation, with the exception of tagmentation, where the transposition reaction attaches the adaptors (Table 8.1). Adaptor sequences vary by sequencing technology and application (overview in Fig. 8.3). Adaptors can contain just the sequencing primer site, commonly also with a barcode incorporated to identify pooled libraries sequenced together on one run. Custom barcodes are easy to develop for the 454 and Ion Torrent systems (examples available at http://www.eebweb.arizona.edu/faculty/mbsulli/protocols/TMPL_LAs.pdf), but more complicated for Illumina sequencing where barcoding of the first several sequenced bases disrupts the identification of clustered reads on the sequencing plate (Rohland & Reich, 2012). Particular library methods can have variations in the attached sequences, including the T7 promoter for transcription in the LADS protocol, the mosaic end sequence that is necessary for transposition in the Nextera tagmentation protocol, and sequences specific for amplifying library fragments (e.g., P5 and P7 sequences in the TruSeq Illumina protocol and LADS, or the A-linker in Linker Amplification). A Y-adaptor instead of dual dsDNA adaptors has also

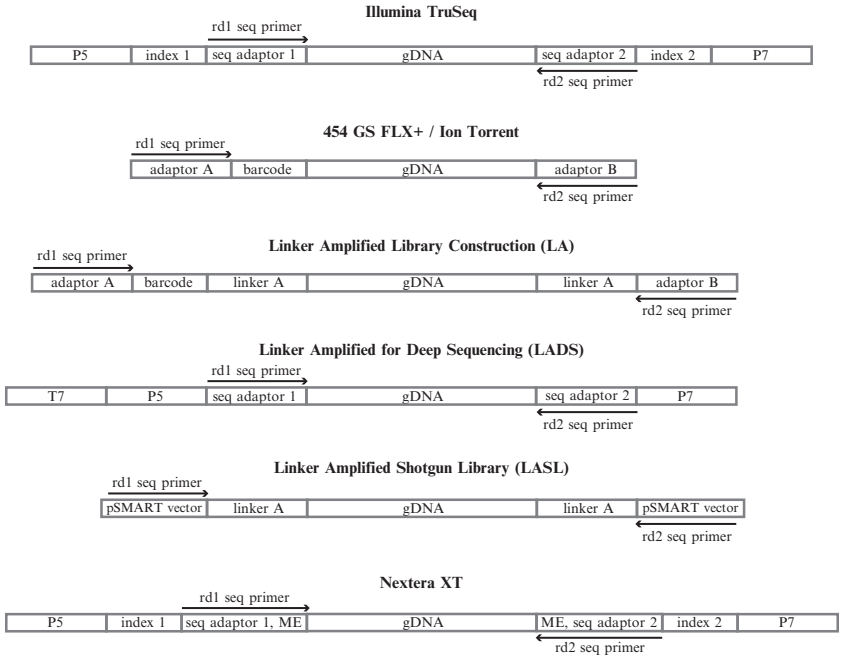

Figure 8.3 Overview schematic of adaptor sequences involved in commonly used library preparation technologies. This figure represents an overview of finished library fragments generated using each library preparation method discussed in this review. Particular focus is placed on (1) the presence of index or barcode regions that allow a library to be pooled with other libraries for efficient sequencing, (2) the location of sequencing primers illustrated as arrows here, indicating the parts of each fragment that will appear in the final sequencing output (and potentially require trimming), and (3) auxiliary sequences such as T7, P5, and P7 that are important for library amplification. ME stands for mosaic end, a subsection of the Nextera sequencing adaptor that allows transposition to occur.

been used to prevent the loss of library DNA due to the attachment of incorrect combinations of adaptors (see 454 General vs. 454 Rapid prep kits, also Zheng et al., 2010).

The success of adaptor ligation is critical to the generation of a robust sequencing library, particularly for low-input DNA samples where optimizing the adaptor-to-template (calculated as free DNA ends) ratio is critical (Solonenko et al., 2013). For particularly low-input DNA libraries, adaptors can also be used to amplify DNA prior to sequencing preparations as shown for LADS (Hoeijmakers, Bartfai, Francoijs, & Stunnenberg, 2011) and LA (Duhaime et al., 2012) in Fig. 8.3. Notably, LADS and LA amplifications are much preferable to amplifications using random hexamer primers and

phi29, which result in nonquantitative and nonreproducible library composition (Yilmaz et al., 2010) where quantities can vary as much as 10,000-fold from starting concentrations (Zhang et al., 2006).

3.4. Sizing and other options

From gel sizing to beads and chip-based systems, there are many options available for controlling the size of library fragments (summarized in Table 8.1). Gel sizing has traditionally been used for DNA sizing, but it is problematic for low-input DNA samples as there may be too little DNA to visualize it on a gel and the protocols suffer from inefficient yields (~50%) and intersample contamination (Duhaime et al., 2012). Sizing is, however, critical for targeting a small range of fragment sizes so as to improve final library quality (Linnarsson, 2010; Quail et al., 2008). An overabundance of small DNA fragments may alter the stoichiometry of adaptor ligation reactions or overpopulate the library during PCR amplification steps. Tightly sized input DNA is also particularly valuable for downstream analyses (e.g., scaffolding for genome assembly) that rely upon paired-end or mate-pair information (Simpson et al., 2009). Acoustic shearing can even produce a fragment distribution that is narrow enough that sizing can be skipped (Solonenko et al., 2013). The Pippin Prep is a more accurate method of gel sizing, and while it requires more investment in equipment, this method is recommended for low-DNA viral metagenomic protocols (Duhaime et al., 2012). The LabChip XT system is another automated sizing method with greater accuracy compared to gel sizing, but currently this has a higher price point. By far, the most cost-effective and high-throughput sizing method uses carboxylic acid coated beads (SPRI, Ampure XP, or My One) to capture different sizes of DNA (Borgstrom, Lundin, & Lundeberg, 2011; Rohland & Reich, 2012). Lastly, columns commonly used to remove extra nucleotides, primers, or adaptors and adaptor dimers may also function as a sizing step, as small DNA fragments are removed (>100 bp for QiaQuick PCR Cleanup Kit).

3.5. Amplification protocols for enrichment, quantity, and signal detection

Once DNA has been processed as above, there remains only the need to amplify the resultant DNA molecules before sequencing. Amplification serves several purposes in metagenomic sequencing library protocols. First, limited amplification cycles (10 or fewer for Illumina TruSeq prep) enrich

the DNA pool for molecules containing correctly ligated adaptors. Second, for low-input DNA samples, amplification can be used to augment sample DNA so as to have enough material to survive library preparation loss steps. Amplification is also used to improve signal detection when a pool of synchronized sequenced reads is required (e.g., 454, Illumina, Ion Torrent). Commonly, this is a separate, final step in library preparation before sequencing—an amplification to create the ~1000 copies that are read by the sequencer. Notably, these PCRs are done with each template isolated in some manner: 454 and Ion Torrent utilize emPCR on a primer-covered bead, while Illumina uses bridge amplification to generate localized "clusters" on a primer-covered sequencing plate (Metzker, 2010). Third, the amplification step is of critical importance and associated choices should not be made lightly. This is because whole-genome amplification methods lead to nonquantitative metagenomes (Yilmaz et al., 2010), while PCR-based amplification is prone to several biases including stochasticity of amplification, heteroduplex formation, chimeric amplicons, and %G+C bias due to the polymerase, high-temperature amplification conditions, and differential priming (reviewed in Duhaime & Sullivan, 2012). However, for PCR-based amplification methods, conditions can be optimized to yield less biased products (Adey et al., 2010), including adjustment of cycling conditions and addition of stabilizing compounds (Schwientek, Szczepanowski, Ruckert, Stoye, & Puhler, 2011), linear amplification (LADS, Hoeijmakers et al., 2011) to lower cross-amplicon competition for primers (Shaw, 2002), and leaving out the amplification step entirely when DNA amounts are not limiting (>1 μg (Kozarewa et al., 2009) for Illumina, standard 454 protocol). Because emPCR and bridge amplification physically isolate amplicons from each other, the signal amplification reactions are a minimal source of bias, with artificial duplicates being the largest issue and observed only for emPCR-based technologies (454 and Ion Torrent; Gomez-Alvarez, Teal, & Schmidt, 2009). Notably, single-molecule sequencing developments may improve these technologies further (Wanunu, 2012).

3.6. Library quantification

The final step of any library preparation procedure is quantification of the library before loading the library for sequencing by emPCR for 454 or Ion Torrent and bridge amplification for Illumina. Correct quantification prevents the library DNA from being overloaded, which can lead to mixed signals, or underloaded, which underutilizes sequencing capacity. Library

concentration information also gives the user the opportunity to strategically pool several libraries when sequencing depth requires less than one run or lane. Several methods are available for this procedure including qPCR, and titration-free qPCR (Zheng et al., 2010), but typically this step is done by the sequencing center and is not a choice for the user to make.

3.7. Sequencing reaction and technologies

Ultimately, each sequencing technology differs not only in preparation (reviewed here) but also in type of sequencing data generated (reviewed in Glenn, 2011; Kircher & Kelso, 2010; Metzker, 2010). Briefly, two important features are the cost efficiency of sequencing data and the read length. Illumina sequencing is the current leader in cost with tens of millions of reads per run, with high potential to overwhelm downstream bioinformatic processing pipelines (Chiang, Clapham, Qi, Sale, & Coates, 2011). 454 GS FLX produces the longest reads available in a next-generation system, an important characteristic for assembly, as well as routine metagenomic analysis (Wommack, Bhavsar, & Ravel, 2008). Beyond these predominantly genome-centered reviews, our own previous work used replicated metagenomics to evaluate the impact of sequencing platforms on the resulting viral metagenomes and showed that the choice of sequencing technology may be less of an influence on the content of metagenomic data than choices made during library preparation (Solonenko et al., 2013).

4. CONCLUSIONS

As new library preparation methods are developed, viral metagenomics continues to become less expensive and more reproducible, as well as more accessible to an expanding diversity of viral types. While the viral metagenomic sample-to-sequence workflow is relatively well established now for dsDNA viruses, there is a need for parallel research and development toward quantitative metagenomic processing steps for accessing ssDNA and RNA viruses in the environment. Mindful of this, it is clear that modern sequencing capacity now empowers metagenomics to adopt experimental designs involving technical replicates (Knight et al., 2012) and that such designs have proven critical for understanding impacts of library preparation methods and sequencing platforms on the resulting viral metagenomes (Solonenko et al., 2013). Implied in these goals is the use of efficient, replicable methods for generating viral metagenomes, an important part of metagenomic experimental design. Making informed

choices at key steps in metagenomics library preparation, such as fragmentation, ligation, and amplification, may reduce the chances of unexpected failure of library preparation or bias in metagenomic sequencing data. As such, refined metagenomic datasets coupled with myriad emerging viral ecology tools that allow access to single viral genomes, link wild viruses to their hosts, and evaluate viral community morphology (Allen et al., 2011, 2013; Brum, Schenck, & Sullivan, 2013; Deng et al., 2013; Tadmor, Ottesen, Leadbetter, & Phillips, 2011) are transforming the landscape of questions that researchers can ask. Together, these advances beckon a new era for the field where we can finally develop a mechanistic understanding of the principles governing variations in natural virus and microbial communities, one virus and one host at a time.

ACKNOWLEDGMENTS

We thank Christine Schirmer for assistance with figures and tables and technical discussions as well as Jennifer Brum and Natalie Solonenko for comments on the chapter. Funding was provided by the Gordon and Betty Moore Foundation to M. B. S. and an NSF IGERT Comparative Genomics Training Grant to S. A. S.

REFERENCES

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., & Turner, E. H. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, *11*(12), R119.

Allen, L. Z., Ishoey, T., Novotny, M. A., McLean, J. S., Lasken, R. S., & Williamson, S. J. (2011). Single virus genomics: A new tool for virus discovery. *PLoS One*, *6*(3), e17722.

Allers, E., Moraru, C., Duhaime, M., Beneze, E., Solonenko, N., Barerro-Canosa, J., et al. (2013). Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environmental Microbiology*, *15*(8), 2306–2318.

Anderson, N. G., Gerin, J. L., & Anderson, N. L. (2003). Global screening for human viral pathogens. *Emerging Infectious Diseases*, *9*(7), 768–774.

Andrews-Pfannkoch, C., Fadrosh, D. W., Thorpe, J., & Williamson, S. J. (2010). Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Applied and Environmental Microbiology*, *76*(15), 5039–5045.

Asan, Geng, C., Chen, Y., Wu, K., Cai, Q., Wang, Y., et al. (2012). Paired-end sequencing of long-range DNA fragments for de novo assembly of large, complex mammalian genomes by direct intra-molecule ligation. *PLoS One*, *7*(9), e46211.

Borgstrom, E., Lundin, S., & Lundeberg, J. (2011). Large scale library generation for high throughput sequencing. *PLoS One*, *6*(4), e19119.

Bragg, J. G., & Chisholm, S. W. (2008). Modelling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS One*, *3*, e3550.

Bratbak, G., Wilson, W., & Heldal, M. (1996). Viral control of *Emiliania huxleyi* blooms? *Journal of Marine Systems*, *9*(1), 75–81.

Breitbart, M. (2012). Marine viruses: Truth or dare. *Annual Review of Marine Science*, *4*, 425–448.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(22), 14250–14255.

Brum, J. R., Schenck, R. O., & Sullivan, M. B. (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *The ISME Journal* advance online publication, 2 May 2013. http://dx.doi.org/10.1038/ismej.2013.67.

Cesar Ignacio-Espinoza, J., Solonenko, S. A., & Sullivan, M. B. (2013). The global virome: Not as big as we thought? *Current Opinion in Virology*.

Chiang, G. T., Clapham, P., Qi, G., Sale, K., & Coates, G. (2011). Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute. *BMC Bioinformatics*, *12*, 361.

Clokie, M. R. J., Shan, J., Bailey, S., Jia, Y., & Krisch, H. M. (2006). Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environmental Microbiology*, *8*, 827–835.

Culley, A. I., Lang, A. S., & Suttle, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science*, *312*(5781), 1795–1798.

Dammeyer, T., Bagby, S. C., Sullivan, M. B., Chisholm, S. W., & Frankenberg-Dinkel, N. (2008). Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Current Biology*, *18*(6), 442–448.

Deng, L., Gregory, A., Yilmaz, S., Poulos, B. T., Hugenholtz, P., & Sullivan, M. B. (2012). Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging. *mBio*, *3*(6), e00373-12. http://dx.doi.org/10.1128/mBio.00373-12.

Duhaime, M., Deng, L., Poulos, B., & Sullivan, M. B. (2012). Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: A rigorous assessment and optimization of the linker amplification method. *Environmental Microbiology*, *14*, 2526–2537.

Duhaime, M., & Sullivan, M. B. (2012). Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology*, *434*, 181–186.

Ellison, S. L., English, C. A., Burns, M. J., & Keer, J. T. (2006). Routes to improving the reliability of low level DNA analysis using real-time PCR. *BMC Biotechnology*, *6*, 33.

Filiatrault, M. J., Stodghill, P. V., Bronstein, P. A., Moll, S., Lindeberg, M., Grills, G., et al. (2010). Transcriptome analysis of Pseudomonas syringae identifies new genes, noncoding RNAs, and antisense activity. *Journal of Bacteriology*, *192*(9), 2359–2372.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, *399*, 541–548.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759–769.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1513–1518.

Gomez-Alvarez, V., Teal, T. K., & Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *The ISME Journal*, *3*(11), 1314–1317.

Hellweger, F. L. (2009). Carrying photosynthesis genes increases ecological fitness of cyanophage *in silico*. *Environmental Microbiology*, *11*, 1386–1394.

Henn, M. R., Sullivan, M. B., Stange-Thomann, N., Osburne, M. S., Berlin, A. M., Kelly, L., et al. (2010). Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS One*, *5*(2), e9083.

Hoeijmakers, W. A., Bartfai, R., Francoijs, K. J., & Stunnenberg, H. G. (2011). Linear amplification for deep sequencing. *Nature Protocols*, *6*(7), 1026–1036.

Holmfeldt, K., Odic, D., Sullivan, M. B., Middelboe, M., & Riemann, L. (2012). Cultivated single-stranded DNA phages that infect marine Bacteroidetes prove difficult to detect with DNA-binding stains. *Applied and Environmental Microbiology*, *78*(3), 892–894.

Holmfeldt, Karin, et al. (2013). Twelve previously unknown phage genera are ubiquitous in global oceans. *Proceedings of the National Academy of Sciences*, *110*(31), 12798–12803.

Hurwitz, B. H., Deng, L., Poulos, B., & Sullivan, M. B. (2013). Evaluation of methods to concentrate and purify wild ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology*, *15*(5), 1428–1440. http://dx.doi.org/10.1111/j.1462-2920.2012.02836.x.

Hurwitz, B. H., & Sullivan, M. B. (2013). The Pacific Ocean Virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One*, *8*, e57355.

Ignacio-Espinoza, J. C., & Sullivan, M. B. (2012). Phylogenomics of T4 cyanophages: Lateral gene transfer in the "core" and origins of host genes. *Environmental Microbiology*, *14*, 2113–2126.

Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L., & Armbrust, E. V. (2012). Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota. *Science*, *335*(6068), 587–590.

John, S. G., Mendez, C. B., Deng, L., Poulos, B., Kauffman, A. K. M., Kern, S., et al. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports*, *3*(2), 195–202.

Kim, K. H., & Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Applied and Environmental Microbiology*, *77*(21), 7663–7668.

Kim, K. H., Chang, H. W., Nam, Y. D., Roh, S. W., Kim, M. S., Sung, Y., et al. (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and Environmental Microbiology*, *74*(19), 5975–5985.

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing—Concepts and limitations. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524–536.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., et al. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nature Biotechnology*, *30*(6), 513–520.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, *6*(4), 291–295.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., & Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, *438*(7064), 86–89.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., & Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(30), 11013–11018.

Linnarsson, S. (2010). Recent advances in DNA sequencing methods—General principles of sample preparation. *Experimental Cell Research*, *316*(8), 1339–1343.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18.

Marine, R., Polson, S. W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., et al. (2011). Evaluation of a transposase protocol for rapid generation of shotgun high-throughput

sequencing libraries from nanogram quantities of DNA. *Applied and Environmental Microbiology*, *77*(22), 8071–8079.
Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews. Genetics*, *11*(1), 31–46.
Morgan, J. L., Darling, A. E., & Eisen, J. A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One*, *5*(4), e10209.
Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155.
Paul, J. H. (1999). Microbial gene transfer: An ecological perspective. *Journal of Molecular Microbiology and Biotechnology*, *1*, 45–50.
Paul, J. H. (2008). Prophages in marine bacteria: Dangerous molecular time bombs or the key to survival in the seas? *The ISME Journal*, *2*(6), 579–589.
Peng, Y., Leung, H. C., Yiu, S. M., & Chin, F. Y. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428.
Quail, M. A. (2010). *DNA: Mechanical breakage*. *Encyclopedia of Life Sciences*. Chichester: John Wiley & Sons, Ltd.
Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., et al. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, *5*(12), 1005–1010.
Rodrigue, S., Materna, A. C., Timberlake, S. C., Blackburn, M. C., Malmstrom, R. R., Alm, E. J., et al. (2010). Unlocking short read sequencing for metagenomics. *PLoS One*, *5*(7), e11840.
Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939–946.
Rosario, K., Duffy, S., & Breitbart, M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: Insights gained from metagenomics. *Archives of Virology*, *157*(10), 1851–1871.
Roux, S., Krupovic, M., Poulet, A., Debroas, D., & Enault, F. (2012). Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One*, *7*, e40418.
Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M., & Mead, D. (2008). Assembly of viral metagenomes from Yellowstone hot springs. *Applied and Environmental Microbiology*, *74*(13), 4164–4174.
Schwientek, P., Szczepanowski, R., Ruckert, C., Stoye, J., & Puhler, A. (2011). Sequencing of high G+C microbial genomes using the ultrafast pyrosequencing technology. *Journal of Biotechnology*, *155*(1), 68–77.
Sharon, I., Battchikova, N., Aro, E. M., Giglione, C., Meinnel, T., Glaser, F., et al. (2011). Comparative metagenomics of microbial traits within oceanic viral communities. *The ISME Journal*, *5*(7), 1178–1190.
Shaw, C. A. (2002). Theoretical considerations of amplification strategies. *Neurochemical Research*, *27*, 1123–1131.
Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123.
Solonenko, S., Ignacio-Espinoza, J. C., Alberti, A., Cruaud, C., Hallam, S. J., Konstantinidis, K. T., et al. (2013). Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics*, *14*(1), 320.
Steward, G. F., Culley, A. I., Mueller, J. A., Wood-Charlson, E. M., Belcaid, M., & Poisson, G. (2013). Are we missing half of the viruses in the ocean? *The ISME Journal*, *7*(3), 672–679.

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., & Chisholm, S. W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology*, *4*, e234.

Sullivan, M. B., Waterbury, J. B., & Chisholm, S. W. (2003). Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature*, *424*, 1047–1051.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361.

Suttle, C. A. (2007). Marine viruses—Major players in the global ecosystem. *Nature Reviews. Microbiology*, *5*, 801–812.

Tadmor, A. D., Ottesen, E. A., Leadbetter, J. R., & Phillips, R. (2011). Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science*, *333*(6038), 58–62.

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems. *Limnology and Oceanography*, *45*(6), 1320–1328.

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., et al. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), E757–E764.

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature Protocols*, *4*(4), 470–483.

Tseng, Q., Lomonosov, A. M., Furlong, E. E., & Merten, C. A. (2012). Fragmentation of DNA in a sub-microliter microfluidic sonication device. *Lab on a Chip*, *12*(22), 4677–4682.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, *304*(5667), 66–74.

Vyawahare, S., Griffiths, A. D., & Merten, C. A. (2010). Miniaturization and parallelization of biological and chemical assays in microfluidic devices. *Chemistry & Biology*, *17*(10), 1052–1065.

Wanunu, M. (2012). Nanopores: A journey towards DNA sequencing. *Physics of Life Reviews*, *9*(2), 125–158.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, *28*(2), 127–181.

Weinbauer, M. G., & Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? *Environmental Microbiology*, *6*(1), 1–11.

Wilhelm, S. W., & Suttle, C. A. (1999). Viruses and nutrient cycles in the sea. *Bioscience*, *49*(10), 781–788.

Wommack, K. E., Bhavsar, J., & Ravel, J. (2008). Metagenomics: Read length matters. *Applied and Environmental Microbiology*, *74*(5), 1453–1463.

Wommack, K. E., & Colwell, R. R. (2000). Virioplankton: Viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews*, *64*, 69–114.

Yilmaz, S., Allgaier, M., & Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods*, 7(12), 943–944.

Zerbino, D. R., McEwen, G. K., Margulies, E. H., & Birney, E. (2009). Pebble and rock band: Heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One*, *4*(12), e8407.

Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W., et al. (2006). Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology*, *24*(6), 680–686.

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., et al. (2013). Abundant SAR11 viruses in the ocean. *Nature*, *494*(7437), 357–360.
Zheng, Z., Advani, A., Melefors, O., Glavas, S., Nordstrom, H., Ye, W., et al. (2010). Titration-free massively parallel pyrosequencing using trace amounts of starting material. *Nucleic Acids Research*, *38*(13), e137.