environmental
microbiology

Environmental Microbiology (2012) **14**(9), 2526–2537 doi:10.1111/j.1462-2920.2012.02791.x

Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method

Melissa B. Duhaime,[1,†] Li Deng[1,2], Bonnie T. Poulos[1] and Matthew B. Sullivan[1,*]

[1] *Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA.*

[2] *Helmholtz Zentrum München-German Research Center for Environmental Health, Institute of Groundwater Ecology, Neuherberg, Germany.*

Summary

Metagenomics generates and tests hypotheses about dynamics and mechanistic drivers in wild populations, yet commonly suffers from insufficient (< 1 ng) starting genomic material for sequencing. Current solutions for amplifying sufficient DNA for metagenomics analyses include linear amplification for deep sequencing (LADS), which requires more DNA than is normally available, linker-amplified shotgun libraries (LASLs), which is prohibitively low throughput, and whole-genome amplification, which is significantly biased and thus non-quantitative. Here, we adapt the LASL approach to next generation sequencing by offering an alternate polymerase for challenging samples, developing a more efficient sizing step, integrating a 'reconditioning PCR' step to increase yield and minimize late-cycle PCR artefacts, and empirically documenting the quantitative capability of the optimized method with both laboratory isolate and wild community viral DNA. Our optimized linker amplification method requires as little as 1 pg of DNA and is the most precise and accurate available, with G + C content amplification biases less than 1.5-fold, even for complex samples as diverse as a wild virus community. While optimized here for 454 sequencing, this linker amplification method can be used to prepare metagenomics libraries for sequencing with next-generation platforms, including Illumina and Ion Torrent, the first of which we tested and present data for here.

Introduction

Microbial processes drive much of the biogeochemistry that fuels the planet (Falkowski *et al.*, 2008), and viruses meddle with these microbial processes at the level of the single cell hosts they infect, resulting in modulation of local- and global-scale biogeochemical processes. This has been best demonstrated in the ocean cyanobacteria and their viruses (cyanophages), whose genomes contain metabolically and environmentally significant genes, including genes for photosynthesis (Mann *et al.*, 2003; Lindell *et al.*, 2004; Millard *et al.*, 2004; Sullivan *et al.*, 2006), phosphate stress response (Sullivan *et al.*, 2005), nitrogen stress response (Sullivan *et al.*, 2010), and nucleotide scavenging (Sullivan *et al.*, 2005). In model systems, these core photosynthesis genes are expressed (Clokie *et al.*, 2006; Lindell *et al.*, 2007) and translated into proteins (Lindell *et al.*, 2005) during infection, and are predicted to boost phage fitness (Bragg and Chisholm, 2008; Hellweger, 2009). Further, cyanophages shuffle genes in niche-defining host genomic islands (Coleman *et al.*, 2006; Kettler *et al.*, 2007; Rodriguez-Valera *et al.*, 2009), resulting in viral-driven changes of the host cell surface (Avrani *et al.*, 2011).

Yet, in most environments, there are few such model systems available, and ultimately a community-scale context is needed to understand the extent to which model system findings are a reliable proxy for wild populations. Researchers commonly turn to whole community sequencing, metagenomics (Handelsman *et al.*, 1998), to probe viral and microbial diversity, protein function and population genomics. Many of these studies are hindered by limited biomass, a consequence of targeted genomics [e.g. stable-isotope probing (Neufeld *et al.*, 2007), cell sorting (Woyke *et al.*, 2009)], low cell density microbial communities (Biddle *et al.*, 2008) or, as in virus studies, small target genome sizes. For example, a typical 20 l ocean virus sample yields on the order of 1 pg to 1 ng

Received 30 November, 2011; revised 5 April, 2012; accepted 4 May, 2012. *For correspondence. E-mail mbsulli@email.arizona.edu; Tel. (+1) 520 626 9100; Fax (+1) 520 621 9903. †Present address: Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA.

DNA, while 454 pyrosequencing and Illumina require 1–5 μg for standard library prep, with slightly less DNA necessary for recent methodological advances, such as linear amplified deep sequencing (LADS), which requires 3–40 ng DNA (Hoeijmakers *et al*., 2011) and Nextera, which requires > 50 ng (Marine *et al*., 2011). To date, viral researchers have relied on linker amplification shotgun libraries (LASLs; Breitbart *et al*., 2002; Vega Thurber, 2009) or whole-genome amplification methods [e.g. multiple displacement amplification (MDA); Angly *et al*., 2006; Dinsdale *et al*., 2008] to generate sufficient material. However, the former suffers from cloning biases and does not scale for next-generation sequencing and the latter suffers from stochastic amplification biases, which render the resulting metagenomes non-quantitative (Abulencia *et al*., 2006; Zhang *et al*., 2006; Arriola *et al*., 2007) and can skew a community's taxonomic profile (Yilmaz *et al*., 2010), rendering cross-sample comparison meaningless.

Two recent developments set the stage for progress, particularly in environmental viral genomics. First, a new precipitation method improves aquatic viral concentration efficiencies from < 25% (typical of tangential flow filtration) to nearly 100% (John *et al*., 2011). Second, the Broad Institute recently modified LASL protocols (Breitbart *et al*., 2002) for 454 pyrosequencing (Henn *et al*., 2010). Briefly, in this linker amplification (LA) modification for next-generation platforms, DNA is sheared, blunt-end repaired and linker-ligated, then gel-sized to a narrow size range before PCR amplification to generate greater quantities of the target DNA. Genome sequencing of viral isolates suggested that these features minimize inherent PCR biases (Henn *et al*., 2010), generally thought to be due to heterogeneous fragment lengths and variable primer site annealing.

Here, we further improve upon the LA method through assessment of *sensitivity* – by answering 'how low can we go?' with respect to starting DNA concentrations, *efficiency* – by identifying and optimizing steps where sample loss occurs, *accuracy* – by empirically quantifying sequence biases introduced, and *applicability* – by successful application of the method to multiple next-generation sequencing platforms.

Results and discussion

LA method optimizations

Briefly (Fig. 1), extracted DNA is sheared to 400–800 bp using an ultrasonic technique (Covaris). The sheared DNA is end-repaired to facilitate ligation of oligonucleotide linkers. Linker-ligated DNA is then size fractionated (400–800 bp) to target the properly ligated DNA. A small-scale PCR titration is performed to determine the optimal cycle

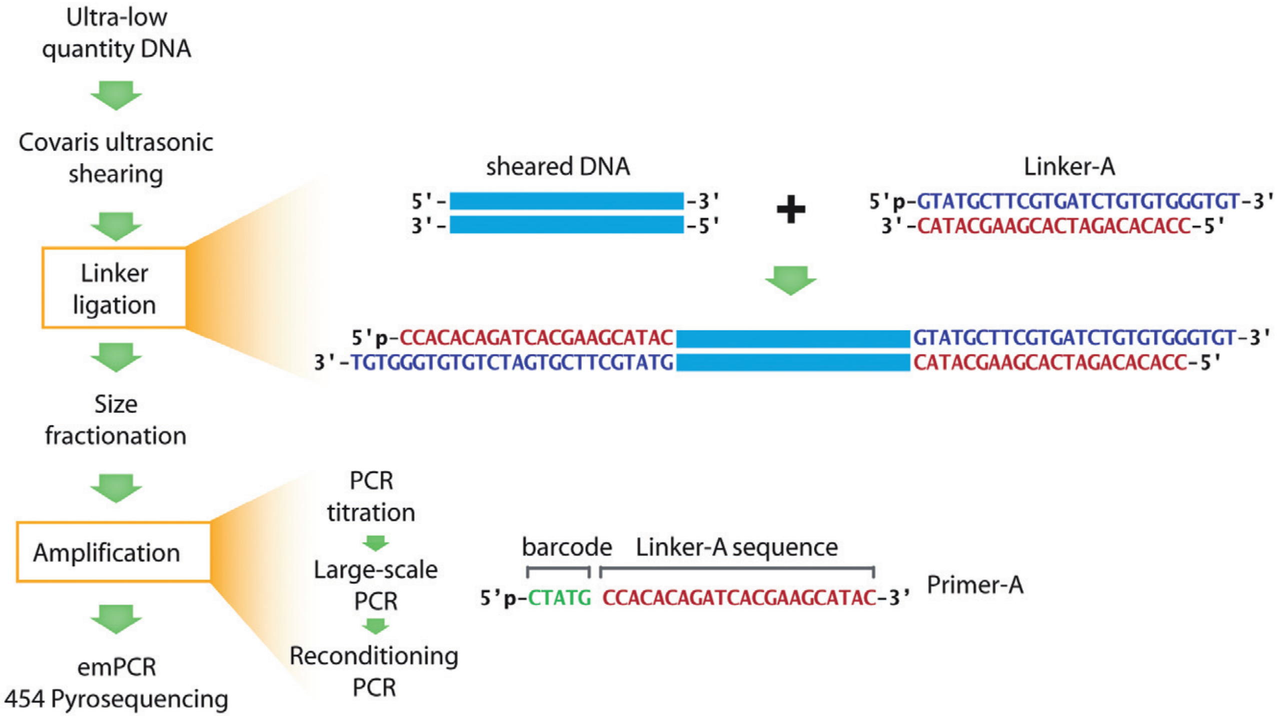

Ultra-low quantity DNA
Covaris ultrasonic shearing
Linker ligation
Size fractionation
Amplification
emPCR
454 Pyrosequencing
sheared DNA
5'-
-3'
3'-
-5'
Linker-A
5'p-GTATGCTTCGTGATCTGTGTGGGTGT-3'
3'-CATACGAAGCACTAGACACACC-5'
5'p-CCACACAGATCACGAAGCATAC
GTATGCTTCGTGATCTGTGTGGGTGT-3'
3'-TGTGGGTGTGTCTAGTGCTTCGTATG
CATACGAAGCACTAGACACACC-5'
PCR titration
Large-scale PCR
Reconditioning PCR
barcode
Linker-A sequence
5'p-CTATG CCACACAGATCACGAAGCATAC-3'
Primer-A

Fig. 1. Linker amplification (LA) method schema. This study assesses an optimized LA method, with particular focus on providing new bar-codes in the linker ligation step to facilitate pooling of samples, as well as quantitative evaluation of the impact of amplification on resulting isolate and community DNA genomic sequencing.

Table 1. Summary of treatments studied in linker amplification sequence analysis.

Pool	Treatment	Input DNA (ng)	PCR cycles	Barcode (5′–3′)	Linker	Reads (post-QC)
***Pseudoalteromonas* phage H105/1: clonal virus lysate**						
1	cyc15A	10	15	CGACA	CCA CAC AGA TCA CGA AGC ATA C	4 306
	cyc15B			CATAG		1 626
	cyc15C			ATGTA		7 582
	cyc18A	1	18	CGTGT		8 072
	cyc18B			ACGTG		11 889
	cyc18C			TGAGT		12 739
	cyc20A	0.1	20	CTCTA		8 729
	cyc20B			ACTCT		3
	cyc20C			TGCTG		5 186
	cyc25A	0.01	25	CTATG		8 722
	cyc25B			AGCAT		8 006
	cyc25C			TCGCA		6 091
	cyc30A	0.001	30	CTGAG		7 250
	cyc30B			ATCAG		8 201
	cyc30C			TCATA		10 992
2	cyc15rA	10	15 + 3[a]	CGACA	CCA CAC AGA TCA CGA AGC ATA C	1 287
	cyc15rB			CATAG		2 088
	cyc15rC			ATGTA		1 710
	cyc18rA	1	18 + 3	CGTGT		1 183
	cyc18rB			ACGTG		436
	cyc18rC			TGAGT		900
	cyc20rA	0.1	20 + 3	CTCTA		4 431
	cyc20rB			ACTCT		4
	cyc20rC			TGCTG		1 194
	cyc25rA	0.01	25 + 3	CTATG		2 768
	cyc25rB			AGCAT		1 157
	cyc25rC			TCGCA		1 527
	cyc30rA	0.001	30 + 3	CTGAG		1 775
	cyc30rB			ATCAG		5 195
	cyc30rC			TCATA		1 252
	unamp	n/a	No amp	None	ACG AGT GCG TAT ATC GCG AGT CAT	30 279
Biosphere2 Ocean: environmental virus sample						
1	B2cyc15A	10	15	CGACA	CCA CAC AGA TCA CGA AGC ATA C	222 421
	B2cyc25A	0.1	25	CAGAT		212 093
	B2cyc15rA	10	15 + 3	ACGTG		119 144
	B2cyc25rA	0.1	25 + 3	TACGA		111 680
2	B2cyc15B	10	15	CGACA	CCA CAC AGA TCA CGA AGC ATA C	261 245
	B2cyc25B	0.1	25	CAGAT		340 488
3	B2cyc15C	10	15	CGACA	CCA CAC AGA TCA CGA AGC ATA C	246 313
	B2cyc25C	0.1	25	CAGAT		310 311
4	unamp A	n/a	No amp	None	ACG AGT GCG TAT ATC GCG AGT CAT	132 639
5	unamp B	n/a	No amp	None	None	160 879

Triplicates are differentiated as A, B and C; reconditioned samples are identified with an 'r'. When text in a row is blank, refer to the previously listed text; for example, input DNA for each of cyc15A, cyc15B, and cyc15C is 10 ng.

a. Three additional cycles represent the reconditioning PCR.

n/a, not applicable; No amp, no amplification.

number (lowest number of cycles resulting in a high molecular weight product) for large-scale PCR. A three-cycle reconditioning step is performed to reduce heteroduplexes, increase product yield, and enrich for high molecular weight DNA (Thompson *et al*., 2002) that can be sequenced with next generation sequencers. Here, each step of this LA method was assessed and optimized (Table 1; Table S1), empirically determining the effects of DNA concentration, PCR cycle number, and reconditioning PCR on the resulting datasets generated from each a clonal virus isolate and an environmental viral community (Table 1).

First, we designed a new set of 5 bp barcodes to label DNA samples uniquely during amplification and allow pooling of multiple samples on the sequencing plate (Fig. 1, Table 1).

Second, we identified an alternative high-fidelity polymerase (LA TaKaRa HS) to complement that previously used (*Pfu* Turbo HotStart; Henn *et al*., 2010). Both enzymes yield product from starting DNA concentrations as low as 100 fg in only 30–35 PCR cycles (Table S2). However, differences emerged. Based on sensitivity, TaKaRa outperformed *Pfu* for microbial 16S samples and an isolate genome dilution series, while the opposite held

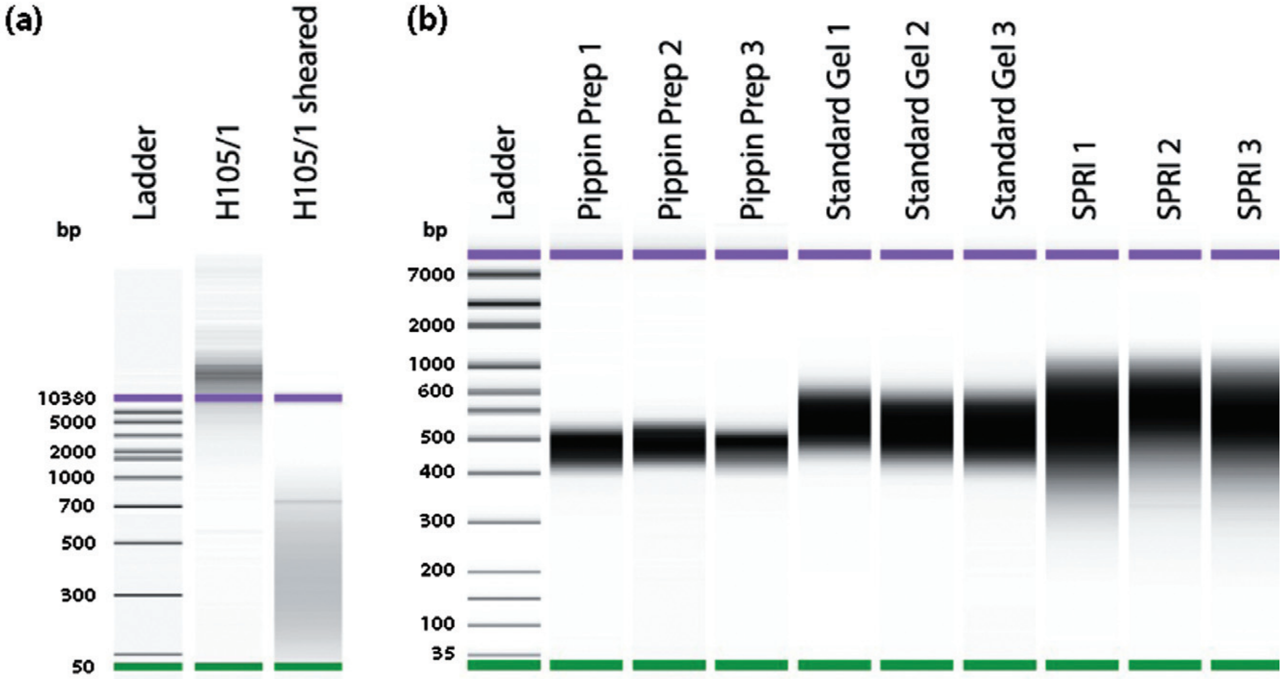


Fig. 2. A. Comparison of sheared H105/1 genomic DNA versus unsheared. DNA was run on Agilent chip (DNA 7500 ladder).
B. Comparison of size-fractionation methods. Size-fractionation of sheared DNA targeting the 400–600 bp range. DNA was run on Agilent chip (high-sensitivity ladder).

true for DNA extracted from a varied collection of ocean virus concentrates (Table S2). However, the sensitivity of *Pfu* came at a cost, as this enzyme amplified a 'no template control' at 30 and 35 cycles, while TaKaRa did not (Table S2). Further, in select samples, TaKaRa yielded more product and with a broader size range than *Pfu* (Fig. S1). Finally, while *Pfu* was more sensitive, amplifying some samples that could not be amplified by TaKaRa (Table S2), the LA TaKaRa HS enzyme enriched for sequences that were of extremely low abundance, 'rares', in the original sample (addressed below in sequence analysis, as well as in companion paper, Hurwitz *et al*., 2012). In summary, we found that enzyme choice depended on sample type and study question. If one enzyme is unable to amplify a challenging sample, success with the other is likely. Furthermore, the use of LA TaKaRa HS polymerase may be of particular interest to studies targeting components of the rare biosphere (*see Results and discussion: Assessment of systematic biases due to amplification*).

Third, we sought to determine the most efficient and precise post-shearing size selection protocol. Efficiency and precision are necessary as environmental samples often yield ultra-low DNA concentrations, yet next-generation sequencing libraries require significant amounts of DNA, e.g. 1–5 μg, of precise sizes, e.g. 400–800 bp for 454 pyrosequencing (Roche Diagnostics GmbH, 2009a) and 300–450 bp for Illumina (2009; Illumina Sample Preparation Guide).

Of the three sizing fractionation methods tested for target recovery efficiency (fraction recovered DNA in target 400–600 bp size range), throughput (ease of applicability to numerous samples simultaneously), and risk of cross-sample contamination, Pippin Prep, an automated optical electrophoretic system that does not require gel extraction, was the most efficient and reproducible (94–96% of input DNA, $n = 3$; Table S3), with the tightest, most specific sizing (Fig. 2B). Pippin Prep and Solid Phase Reversible Immobilization (SPRI), a method based on size-specific capture of DNA by AmPure beads, were equally as high-throughput with low risk of cross-contamination. Yet, SPRI was the least efficient, recovering 46–50% of the targeted size fraction post-shearing ($n = 3$), as it can not be used to bound the upper size range (leaving DNA > 600 bp) due to a sub-optimal ratio of PEG:DNA, the mechanism used to remove DNA fragments (Hawkins *et al*., 1994). However, to size fractionate DNA for Illumina libraries (targeting 300–450 bp), the double-SPRI (dSPRI) method has been shown to be a viable method (Rodrigue *et al*., 2010), as it caps both high and low size ranges. Of the three methods tested, standard gel extraction had moderate target recovery efficiency (64–74%). However, this efficiency varies greatly with researcher proficiency and, further, the size selection takes orders of magnitude more time and risks cross-sample contamination. Based on this comparative analysis, we recommend the Pippin Prep automated electrophoretic system to prepare samples for 454 pyrosequencing libraries.

Finally, we quantified the full process post-shearing for seven phage lysates and environmental virus samples and found that the concentration, blunt-end repair, linker ligation and size fractionation (gel-based) was 5.7% ± 2.3% efficient, with respect to total DNA recovery (Table S4; note this calculation is based on gel size fractionation – it is likely that efficiency is higher with the newly assessed Pippin Prep, the most efficient sizing method tested; Table S3). We have found the typical increase in DNA yield due to LA to range from 10- to 1200-fold (Table S5).

Empirical evaluation of amplification biases

After optimizing the LA method, we sought to determine quantitatively the influence of starting DNA concentration and amplification parameters, such as cycle number and reconditioning (three extra rounds of amplification to reduce heteroduplex formation; Thompson *et al*., 2002), on sequence data from the clonal virus isolate, H105/1, and environmental virus community DNA from the Biosphere2 ocean (B2O; Oracle, AZ).

Effect of starting DNA, cycle number and reconditioning on read depth. In order to discern which treatments have an effect on read depth, the deviation of amplified read depths from unamplified was compared (all coloured lines of Fig. 3A). There was *no* significant difference between treatments of different starting quantities of DNA and numbers of PCR cycling ($P = 0.13–0.82$, pairwise two-tailed *t*-tests; Table S6). This is in contrast to the popularly used MDA, with which limiting template DNA concentration (1 ng) imposes dramatic representational biases on resulting sequence data (Wu *et al*., 2006). At first pass, there was a significant difference between reconditioned and non-reconditioned treatments ($P = 1.6E\text{-}05$). However, this was due to the unintentional reduced sequencing effort of the reconditioned samples (fewer reads causing even some regions of the genome to approach, though never reach, zero coverage; Table 1, Fig. 2A) and subsequent scaling of individual datasets by sequencing effort, which artificially inflates the magnitude of reconditioned read depths. When low coverage areas (< sevenfold) are masked from the genome and the same test performed, there is *no* significant difference between reconditioned and non-reconditioned samples ($P = 0.33$). Aiming for average genome coverage of 15× should help to minimize these low coverage areas and avoid scaling issues that can interfere with cross-dataset comparisons when coverage is disparate. Based on this absence of discernible biases imposed by reconditioning, this step is highly recommended, as it has been shown to minimize heteroduplex formation during amplification (Thompson *et al*., 2002) and will ultimately result in a threefold increase in product yield, an important consideration in low DNA samples.

Assessment of systematic biases due to amplification. On the whole, there was notable difference between the amplified and unamplified samples (Fig. 3A), which we hypothesized to result from systematic biases introduced during amplification (e.g. %G + C). Indeed, by normalizing the relative frequency of %G + C-binned reads to those observed in the unamplified treatment, we found the LA method to under-represent regions of the H105/1 genome with < 40% G + C and over-represent regions above, by 0.5- and 1.5-fold respectively (Fig. 3B). Extending this assessment to the B2O metagenome, we found a similar onefold over- and under-representation of reads, relative to that seen in unamplified treatments (Fig. 4). Notably, the %G + C of the B2O metagenome (12–84%) spans a much wider range than H105/1 (31–55%), and is a range characteristic of most sequenced dsDNA virus metagenomes (Fig. S2). At the %G + C extremes seen in the B2O community assemblage, both high and low %G + C reads are under-represented in amplified treatments, a common phenomenon inherent to PCR amplification (Hoeijmakers *et al*., 2011). Regardless, these systematic differences resulting in 0.5- to 1.5-fold biases are a marked improvement over the stochastic representation biases of whole genome amplification, which can lead to 100s-fold (Woyke *et al*., 2009) to 10 000s-fold changes (Zhang *et al*., 2006) and have been shown to render MDA-generated metagenomes non-quantitative (Yilmaz *et al*., 2010). Our optimized LA results are comparable to observations of the new LADS (Hoeijmakers *et al*., 2011) and amplification-free (Kozarewa *et al*., 2009) methods, though these methods require significantly more DNA, 3–40 ng and 100s of ng of input DNA, respectively, which is often unattainable from environmental samples.

Our experimental design allowed us to evaluate variability in replicate metagenome preparations to assess how PCR cycling and reconditioning impacted resultant datasets. We anticipated that increased cycle number might exacerbate the %G + C biases. However, the lack of such a trend indicates that replicate LA datasets prepared with varying amounts of starting DNA (1 pg to 10 ng) and using different cycling conditions are quantitatively comparable (Figs 3 and 4), with few significant differences between treatments (Table S7a). As with the read depth comparison, there was a higher incidence of significant differences between non-reconditioned and reconditioned samples, especially at higher cycle numbers (Table S7b). Indeed, a slight positive trend existed between reconditioned samples and cycle number, which did not exist for the non-reconditioned samples (Fig. S3). Thus, the additional 3-cycle reconditioning step resulted in a 10-fold increase in DNA – and DNA of higher molecular weight (Fig. S1) – at the cost of only slight %G + C bias at higher cycle numbers (Fig. 3B). Importantly, all biases imposed across all treatments are still never more than 0.5- to 1.5-fold (Figs 3 and 4).

As a common goal of ecology is cross-community comparisons of diversity, we also examined how LA impacted diversity profiles, represented as rarefaction curves of protein clusters with at least 20 members (Fig. 5A). Regardless of cycle number or reconditioning, the rarefaction curves of amplified treatments were nearly identical, while those from the unamplified treatments were less diverse (Fig. 5). Quantification of 'singleton' reads (defined at 90% identity), indicate that rare reads from the original community DNA are enriched for in the amplified treatments (over 62% percent of the original reads; Fig. 5B), with more than 10% fewer singletons in the unamplified treatments.

We hypothesize that this enrichment of rare reads is a feature of the LA TaKaRa polymerase enzyme used during PCR. A related study has shown the enrichment of 'rares' does *not* occur in similarly prepared samples where the *Pfu* Turbo master mix was used, with the resultant

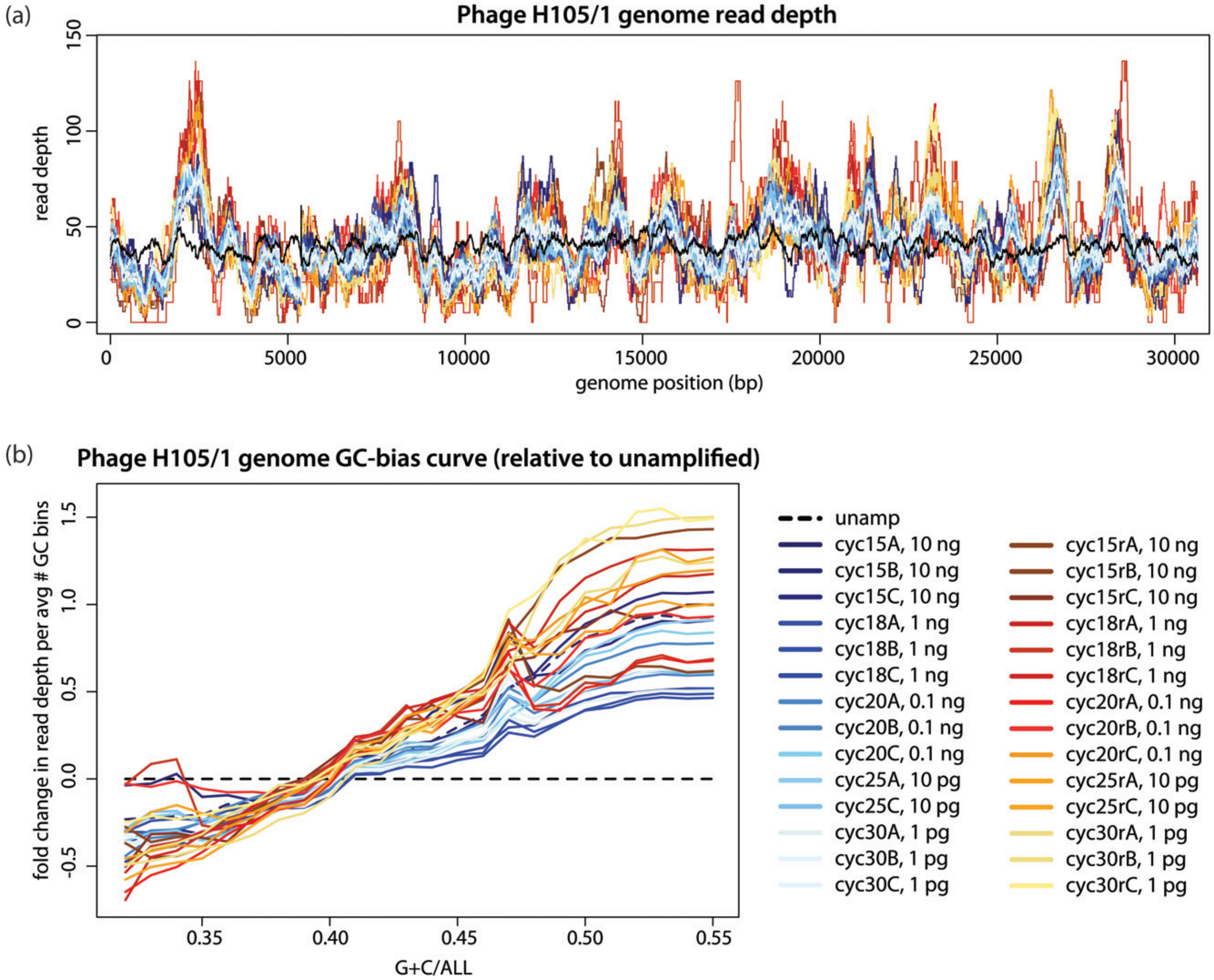


Fig. 3. Quantitative evaluation of resulting isolate genome sequencing data.
A. Read depth of treatments, as mapped to the Phage H105/1 reference genome: 15–30 PCR cycles, with (red) and without (blue) reconditioning, and unamplified (black) genomic DNA. Counts are scaled by the total number of nucleotides per treatment (Table 1) and multiplied by a factor (1 222 442, the average number of nucleotides in all treatments), to scale to a relatable 'read depth' value.
B. H105/1 genome 'GC-bias curve' representing the relationship between %G + C and read depth, as calculated in a 500 bp sliding window across the genome. Colour scale shared between (A) and (B).

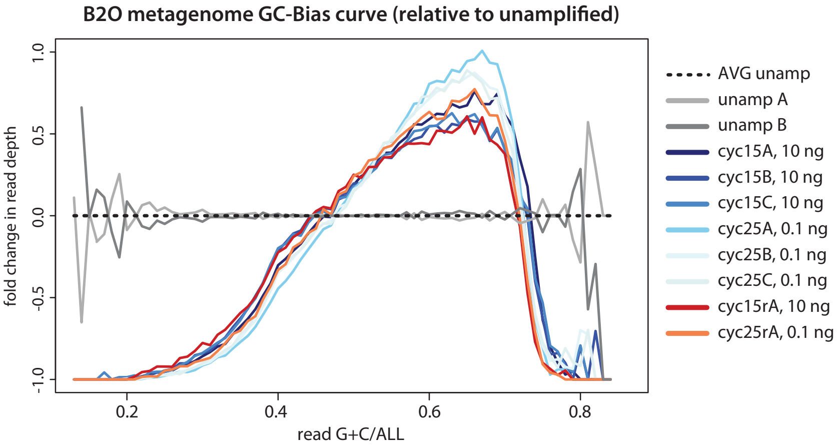


Fig. 4. Biosphere2 ocean metagenome 'GC-bias curve' representing the %G + C of each read per treatment, relative to the average %G + C of the unamplified treatments.

amplified datasets thus appearing more diverse (data published in companion paper; Hurwitz *et al*., 2012). This trend is supported by Wu and colleagues, who found pyrotag datasets prepared with a *Pfu* polymerase to be less diverse than the same sample amplified with a TaKaRa polymerase, which they attribute, quite generally, to possible differences in the enzymes' processivity or proof-reading (Wu *et al*., 2010). These observations may be related to the fact that during PCR, hydrolytic deamination of dCTP to dUTP in already amplified products can inhibit further amplification of these fragments by *Pfu*, as it is an archaeal family B polymerase known to stall at dU-containing DNA (Lasken *et al*., 1996). However, the *Pfu* Turbo master mix contains a dUTPase enhancement

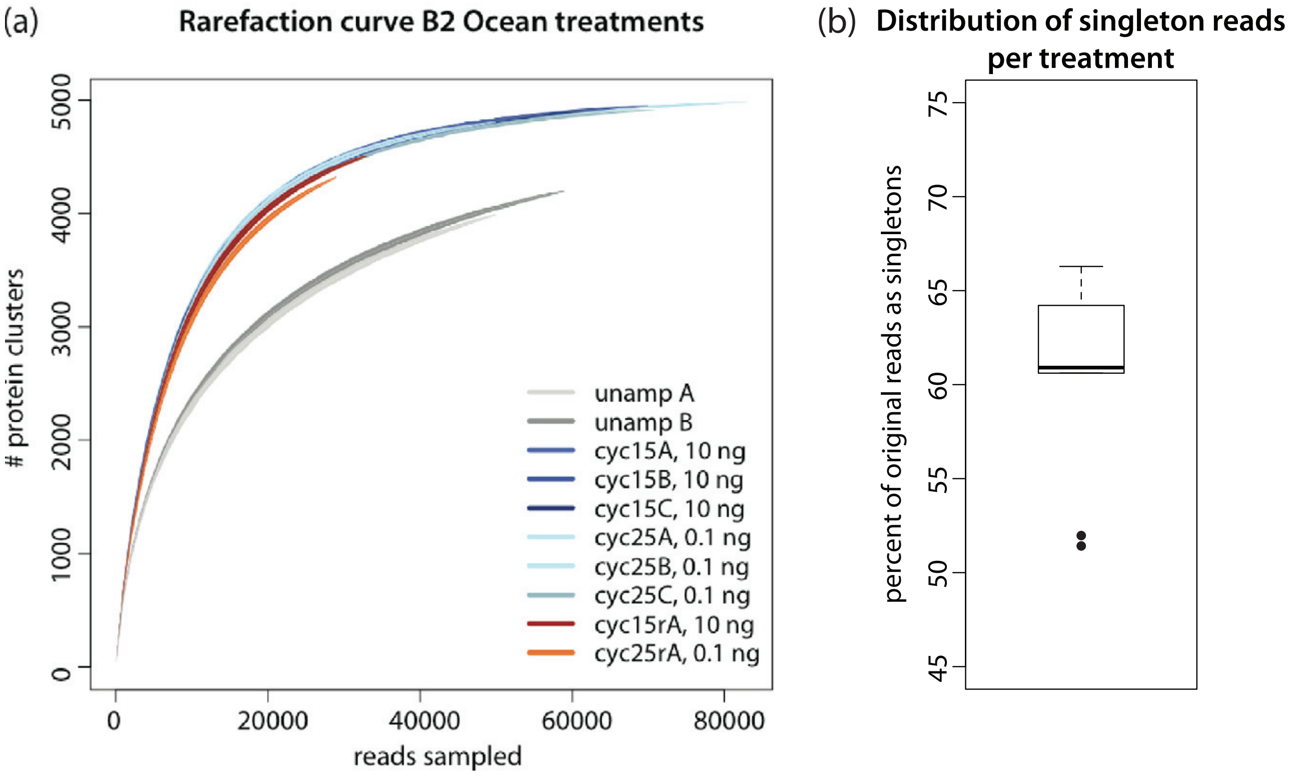


Fig. 5. Protein cluster diversity in the Biosphere 2 ocean viral community.
A. Rarefaction curve representing the relative sequence diversity of each treatment, as measured by protein clusters (with > 20 sequence members) derived from all amplified and unamplified treatments. Higher diversity in the amplified treatments is likely due to a preferential amplification of the rare biosphere by the LA TaKaRa-HS DNA polymerase used in PCR.
B. Boxplot representing the range of percent original reads as singletons in each treatment. The two outliers at 51% are the unamplified treatments.

factor (Hogrefe *et al*., 2002) that effectively deaminates dUTP back to dCTP, allowing continued amplification of already amplified fragments and thus presumably maintaining the original relative abundances of DNA fragments in the sample. Thus, one alternate explanation may be that if the TaKaRa enzyme is at all hindered by the presence of dUTP in already amplified templates, the enzyme will preferentially amplify the low abundance, yet to be amplified rare reads, as we observe. These rare reads amplified to detection by TaKaRa do not appear to be artefacts of amplification, as they have sequence similarity to established protein clusters at the same frequency as the abundant reads in the same dataset (Hurwitz *et al*., 2012). Whether this enrichment of rare reads is beneficial or detrimental depends heavily on the downstream application of the sequenced dataset.

Application to other sequencing platforms. While the protocol and analysis presented here are developed for 454 pyrosequencing, with minor adjustments this LA protocol is appropriate for samples intended for other sequencing platforms. For instance, shearing conditions can be altered to size templates for shorter read-length sequencing technologies (e.g. 200 bp for Ion Torrent's current 316 chip). Further, template mixtures can be used to sequence with Illumina technologies to circumvent issues with the non-random linker sequence on linker-amplified templates, which causes erroneous base-calling by the Illumina software. To this end, we recently pooled linker amplified DNA from 20 freshwater cyanophage isolates at a ratio of 1:1 with known phi29 DNA template (our template mixture) and successfully generated high quality, full-length Illumina sequence data (Fig. S4; L. Deng and M.B. Sullivan, unpubl. data). Alternate approaches to prep libraries for Illumina sequencing include (i) adding a 3′-U during PCR amplification to the LA linker-containing DNA, such that the linker can be cleaved at the U with, e.g. the USER Enzyme (New England Biolabs; Beverly, MA, USA), followed by an S1 nuclease clean-up of remnant ssDNA, resulting in linker-free DNA for Illumina library prep (described by Rodrigue *et al*., 2009); or (ii) performing the LA method with Illumina linkers. The benefit of the former being that the amplification behaviour of linkers in this study has been rigorously tested and assessed here.

With ever-declining sequencing costs and sequence analysis tools becoming more readily available, metagenomics has become a standard tool for investigating wild communities. As such, it is essential to optimize and popularize robust methods to preserve a quantitative metagenomic signal. The optimized LA method presented here enables a several orders of magnitude increase in DNA yield that results in minimally biased (0.5- to 1.5-fold as compared with unamplified) metagenomes appropriate for comparative analyses from samples with limiting amounts (< 1 ng) of DNA.

Experimental procedures

The development and optimization of the LA method described below and (detailed in Table S1) are derived from an accumulation of knowledge gained from preparing over 50 varied samples (Tables S2–5; Figs S1–5). The analyses of sequence biases introduced by the method are based on a single phage isolate, *Pseudoaltermonas* phage H105/1 (herein H105/1; Duhaime *et al*., 2011), and an environmental sample from the Biosphere2 ocean (herein B2O; Oracle, AZ, USA).

Linker amplification

Isolation of DNA. To isolate genomic DNA, H105/1 was grown on *Pseudoalteromonas* sp. H105, the lysate was polyethylene glycol/NaCl precipitated and cesium chloride (CsCl) purified, as previously described (Sambrook and Russell, 2001; Duhaime *et al*., 2011). Genomic DNA was extracted using Wizard Prep Resin (Promega; Madison, WI, USA) and mini-columns (Promega) as described by Henn and colleagues (2010). For environmental virus community DNA, 1080 l seawater was 0.2 μm filtered, and the virus-containing filtrate iron chloride precipitated and concentrated per John and colleagues (2011). The resuspended viral concentrate was treated with DNase I (100 units ml^{-1}) for 2 h at room temperature, then DNase activity inactivated with 100 mM EDTA/EGTA. This virus preparation was purified on a CsCl gradient (recovering 7.1 × 1011 viruses ml-1 from the $\rho = 1.4–1.52$ CsCl fraction) (Sambrook and Russell, 2001) and DNA was extracted from 0.6 ml using Wizard Prep Resin (as above). Lysate and environmental DNA was ligated to Adaptor-A (Fig. 1), diluted, and amplified at various PCR cycle numbers (Table 1) per the LA protocol described below.

Preparation of sheared, Linker-A Ligated DNA. Covaris Adaptive Focused Acoustics (AFA) was used to shear DNA to 400–800 bp, with the following parameters: 130 μl of DNA in Tris EDTA (TE) buffer (up to 5 μg total DNA), duty cycle of 5%, intensity of 3200 cycles per burst, at 6–8°C for 62 or 120 s, for the Covaris E210 and S2 models respectively. DNA fragment sizes were determined before and after shearing on a DNA 7500 or High Sensitivity chip in the Agilent Bioanalyzer 2100 (Agilent Technologies; Santa Clara, CA, USA). Following shearing, low yield (< 500 ng) samples were concentrated two- to threefold (final volume 35–50 μl) with Amicon Ultra-0.5100 kDa filter units (Millipore; Billerica, MA, USA), according to manufacturer's directions.

Hemi-phosphorylated Linker-A (Fig. 1) was prepared by annealing a synthesized forward linker (5′-phosphorylated-GTA TGC TTC GTG ATC TGT GTG GGT GT-3′; 1.14 mM in TE) to the reverse linker (5′-CCA CAC AGA TCA CGA AGC ATA C-3′; 1.14 mM in TE; *not* phosphorylated in order to promote unidirectional ligation) in a 50 mM NaCl buffer. Equal volumes of forward and reverse linkers were heated to 100°C, slowly cooled on the bench to room temperature, and placed on ice for 5 min. Resultant Linker-A was diluted to 10 μM in TE and stored at −20°C for subsequent use.

Blunt-end repair of sheared DNA, ligation to Linker-A and required cleanup reactions (Fig. 1) were performed as described (Henn *et al*., 2010) using the End-It DNA End-Repair kit (Epicentre Biotechnologies; Madison, WI, USA), Fast-Link DNA Ligation kit (Epicentre Biotechnologies) and Min-Elute Reaction Clean-up kit (Qiagen; Valencia, CA, USA) respectively. Sheared, linker-ligated DNA was then size-fractionated.

H105/1 genomic DNA, 10 μg in 130 μl (77.4 ng μl^{-1}) was sheared to 400–600 bp (Covaris) to test three sizing methods in triplicate: gel electrophoresis, SPRI and a digital optical electrophoretic system, Pippin Prep (Sage Science; Beverly, MA, USA). Each method was tested with 13.7 μl of sheared DNA at a concentration of 14.08 ng μl^{-1} (183 ng total DNA). For gel extraction, sheared DNA was run on a 1.5% agarose gel stained with ethidium bromide for 90 min at 80 V. Gel fragments in the 400–600 bp size range were excised, purified to remove agarose (Qiagen MinElute Gell Extraction kit), and eluted in 20 μl EB buffer, as described (Henn *et al*., 2010). SPRI with Agencourt AMPure XP beads (Beckman Coulter; Danvers, MA, USA) was used to remove the less than 400 bp fragments (detailed in *Results and discussion: LA method optimizations*). To specifically retain the desired greater than 400 bp fragments a bead to DNA ratio of 55:100 was used, as determined per Roche (Roche Diagnostics GmbH, 2009a), such that the final reaction contained 13 μl sheared DNA, 7 μl TE buffer, and 11 μl Ampure beads. DNA was eluted in 20 μl EB buffer. Pippin Prep samples were run in pre-cast 2% agarose gel cassettes, pre-stained with ethidium bromide (Sage Science), set to recover the 400–600 bp fragments. DNA was recovered in 39–42 μl final volumes (Table S3). Following each method, DNA accurately retained in the target 400–600 bp range was quantified to determine target recovery efficiency using an Agilent Bioanalyzer 2100 (Agilent Technologies; Santa Clara, CA, USA).

Polymerase evaluation, barcode design. Two high fidelity polymerases, *Pfu* Turbo Hotstart (Stratagene; La Jolla, CA, USA) and TaKaRa LA Taq Hotstart (Takara Bio; Shiga, Japan), both with 3′ to 5′ exonuclease (‘proof-reading’) activity and an antibody quencher for hotstart capability, were assessed for efficiency and sensitivity. Both enzyme reactions used 1–2 μl Linker-A ligated and size-fractionated DNA with 0.5 μl (5 pmol) of the 10 μM PCR phos-A primer. For the *Pfu* Turbo Hotstart system, 12.5 μl *Pfu* Turbo Hotstart 2× Master Mix (0.1 U *Pfu* Turbo μl^{-1}) was used, while the TaKaRa LA HS system required 2.5 μl 10× PCR buffer, 4.0 μl of 2.5 mM dNTP mix (10 nmol each), and 0.25 μl TaKaRa LS Taq HS (5 U LA TaKaRa μl^{-1}). Both reactions were brought to 25 μl with nuclease-free water. For all reactions prepared for the sequence analysis (H105/1 phage genome and B2O environmental DNA), the TaKaRa LA Taq Hotstart system was used.

A series of 5 bp barcodes was added to the ‘phos-A PCR primer’ (5′-p-CCACACAGATCACGAAGCATAC-3′) (Henn *et al*., 2010), such that a sample-specific, unique barcode would be added at the 5′ end of each DNA fragment during amplification (Fig. 1, barcodes listed in Table 1). The barcodes were designed such that (i) no consecutive duplicate nucleotides exist, (ii) barcodes differ by at least 2 nucleotides, (iii) no C appears at the 3′ end, as Linker-A has a 5′ C (Fig. 1), and (iv) no G appears at the 5′ end, as the 454 emPCR primers have a 3′ G (forward: 5′-CGTATCGCCTCCCT CGCGCCATCAG-3′; reverse: 5′-CTATGCGCCTTGCCAG CCCGCTCAG-3′) (Roche Diagnostics GmbH, 2009b).

Small-scale PCR titration. To determine the minimum number of PCR cycles needed for amplification of each sample, a small-scale PCR titration was performed with varying cycle numbers [95°C for 2 min, (95°C for 30 s, 60°C for 60 s, 72°C for 90 s) × 15, 18, 20, 25, or 30, 72°C for 10 min]. Generally, three cycle numbers were tested according to the amount of DNA available after ligation and sizing, as a log-linear relationship exists between input DNA and cycle number needed for amplification (Fig. S5). Generally, 1–10 ng DNA samples were run for 15–20 cycles, 0.1–1 ng for 18–25 cycles, 10–100 pg for 22–30 cycles, and less than 10 pg for 25–35 cycles. If DNA was not amplified by 35 cycles, more DNA was added to the PCR reaction (up to 10% of the PCR reaction volume) or the sample was concentrated using AMPure XP beads (80 μl beads to 100 μl DNA) and PCR titration attempted once more.

PCR products were analysed on 1.5% agarose gels with 0.5 ng μl^{-1} ethidium bromide, run in 1× Tris-acetate-EDTA (TAE) buffer at 90 V for 30 min, using 5 μl PCR product mixed with 1 μl 6× Blue/Orange Loading Dye (Promega). Quick Load 100 bp DNA Ladder (New England Biolabs; Beverly, MA, USA), 250 bp DNA Ladder (Invitrogen; Carlsbad, CA, USA), or 1 kB Plus DNA Ladder (Invitrogen) was used for fragment size determination.

Large-scale amplification and ‘reconditioning PCR’. Sample DNA was amplified (per PCR protocol above) using the number of cycles determined by the small-scale titration. Depending on quantity of starting DNA, six to ten reactions were performed in this step, resulting in up to 250 μl PCR product, to ensure sufficient DNA for sequencing (1–5 μg per standard 454 pyrosequencing or Illumina library).

To both increase yield and minimize heteroduplex formation, a ‘reconditioning PCR’ step was added, *sensu* Thompson and colleagues (2002). To recondition, the amplified DNA is diluted 10-fold in a fresh PCR reaction mix (200 μl reactions with 2.5 μl TaKaRa LA HS and 20 μl of small-scale PCR product as template, all other reagents in same proportions as in small-scale titration) and amplified for three cycles [95°C for 2 min, (95°C for 30 s, 60°C for 60 s, 72°C for 90 s) × 3, 72°C for 10 min], effectively increasing the primer-to-template ratio by replenishing the reaction with new primer. All PCR products previously generated were reconditioned, resulting in up to 2.5 ml of final reconditioned product, which was then concentrated to 250 μl using Amicon Ultra-0.5100 kDa centrifugal columns, purified with the MinElute PCR Purification (Qiagen) kit, and DNA eluted off the mini-columns with 25–40 μl TE buffer warmed to 80°C. Note that in order to determine the effect of this treatment on resultant sequence data, parallel treatments of phage H105/1 genome and B2O environmental DNA were *not* reconditioned (Table 1). Finally, amplified samples were quantified (Quant-iT Pico Green for dsDNA; Invitrogen) and, where sample pooling was necessary, samples with unique barcodes were mixed in equimolar amounts for sequencing library preparation. Libraries were sequenced using GS FLX Titanium 454 pyrosequencing (no paired-ends) or the Illumina HiSeq (paired-end reads, one channel with raw output of 19.9 GB).

Sequence analysis

Phage genome analysis. Genomic reads were mapped per treatment to the complete H105/1 genome (RefSeq NC_015293) using gsMapper (v 2.5.3). Read depths were divided by the total number of nucleotides per treatment to normalize for sequencing effort. %G + C and average read depths were calculated over a 500 bp sliding window to generate 'GC bias' plots.

Environmental virus metagenome. Biosphere2 ocean metagenome reads were subjected to quality control to remove reads that (i) contained an ambiguous base, (ii) were at least two standard deviations from the mean sequence length per plate, (iii) were at least two standard deviations from the mean quality score per plate, or (iv) were identified by cd-hit454 (default parameters) as emulsion-PCR replicates. %G + C was calculated per read. For the rarefaction analysis, reads from all treatments were assembled using Velvet (Zerbino and Birney, 2008) with automatic coverage estimation and a *k*-mer length of 21. Genes were predicted on all raw reads and contigs greater than 100 bp using Prodigal (metagenomic gene finding, remainder of parameters as default) (Hyatt *et al*., 2010). Protein clusters with 60% within-cluster identity were built using cd-hit (Li and Godzik, 2006) with word length of 4. Original reads were non-redundantly mapped to protein cluster representative sequences using blastx (Altschul *et al*., 1990) with 60% identity cut-off. Rarefaction curves were generated by sampling protein clusters of > 20 members without replacement, increasing sampling effort in 5000-read steps (100 replicates) using R scripts. To compare the number of 'rares', reads were clustered per treatment using cdhit-est (Li and Godzik, 2006) at 90% within-cluster identity, considering both strands, and with a word size of 8. The number of singletons per treatment was defined as the number of reads in single-member clusters.

Statistical tests. Tests comparing the relative read depths and %G + C bias between unamplified and amplified treatments were performed using a series of paired two-tailed Student's *t*-tests. To quantify the magnitude of read depth and %G + C bias, the area between the unamplified and amplified curves was calculated by trapezoid integration. Pairwise tests were performed to compare (i) all reconditioned versus all non-reconditioned samples ($n = 14$ pairs), (ii) the reconditioned and non-reconditioned per cycle number ($n = 3$ pairs), (iii) all treatments of a cycle number against each of the remaining cycle number treatments, reconditioned and non-reconditioned combined ($n = 6$ pairs), and (iv) 'c' repeated, testing reconditioned and non-reconditioned separately ($n = 3$ pairs).

The LA protocol is available at http://eebweb.arizona.edu/faculty/mbsulli/protocols.htm. All sequence data is available on the CAMERA web portal, tracking number CAM_P_0000912.

Acknowledgements

We thank N. Strange-Thomann, M. Osburne and V. Rich for discussion during methodological optimization, V. Denef for discussions during analysis, members of the Tucson Marine Phage Lab for comments on manuscript drafts. Biosphere2 Ocean samples were collected by A. Gregory, N. Solonenko and C. Decker. Phage isolate sequencing was performed by Emory University GRA Genomic Center; environmental samples were sequenced by the University of Arizona Genetics Core Facility. We further acknowledge the UITS Research Computing Group and the Arizona Research Laboratories Biotech Computing for high-performance computing access and support. The Joint Genome Institute Community Sequencing Program is acknowledged for providing the freshwater cyanophage pooling experiment on Illumina HiSeq (CSP grant awarded to L.D.) and discussions regarding possible modifications for Illumina library prep (personal communication with C. Hoover); the remainder of the work was funded by a Gordon and Betty Moore Foundation grant to M.B.S.

References

Abulencia, C.B., Wyborski, D.L., Garcia, J.A., Podar, M., Chen, W., and Chang, S.H. (2006) Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Appl Environ Microbiol* **72:** 3291–3301.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., and Carlson, C. (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4:** e368.

Arriola, E., Lambros, M.B., Jones, C., Dexter, T., Mackay, A., and Tan, D.S. (2007) Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. *Lab Invest* **87:** 75–83.

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011) Genomic island variability facilitates *Prochlorococcus*-virus coexistence. *Nature* **474:** 604–608.

Biddle, J.F., Fitz-Gibbon, S., Schuster, S.C., Brenchley, J.E., and House, C.H. (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci USA* **105:** 10583–10588.

Bragg, J.G., and Chisholm, S.W. (2008) Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS ONE* **3:** e3550.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., and Mead, D. (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99:** 14250–14255.

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

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., and Delong, E.F. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311:** 1768–1770.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., and Brulc, J.M. (2008) Functional metagenomic profiling of nine biomes. *Nature* **452:** 629–632.

Duhaime, M.B., Wichels, A., Waldmann, J., Teeling, H., and Glöckner, F.O. (2011) Ecogenomics and genome landscapes of marine *Pseudoalteromonas* phage H105/1. *ISME J* **5:** 107–121.

Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* **320:** 1034–1039.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5:** R245–R249.

Hawkins, T.L., O'Connor-Morin, T., Roy, A., and Santillan, C. (1994) DNA purification and isolation using a solid-phase. *Nucleic Acids Res* **22:** 4543–4544.

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., Stange-Thomann, N., Osburne, M.S., Berlin, A.M., and Kelly, L. (2010) Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE* **5:** e9083.

Hoeijmakers, W.A., Bártfai, R., Françoijs, K.J., and Stunnenberg, H.G. (2011) Linear amplification for deep sequencing. *Nat Protoc* **6:** 1026–1036.

Hogrefe, H.H., Hansen, C.J., Scott, B.R., and Nielson, K.B. (2002) Archaeal dUTPase enhances PCR amplifications with archaeal DNA polymerases by preventing dUTP incorporation. *Proc Natl Acad Sci USA* **99:** 596–601.

Hurwitz, B.H., Deng, L., Poulos, B.T., and Sullivan, M.B. (2012) Evaluation of methods to concentrate and purify wild ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol*, in press.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11:** 119.

Illumina (2009) Paired-End Sequencing Sample Preparation Guide. Catalog # PE-930-1001 Part # 1005063 Rev. B.

John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K., and Kern, S. (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* **3:** 195–202.

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., and Rodrigue, S. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:** e231.

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

Lasken, R.S., Schuster, D.M., and Rashtchian, A. (1996) Archaebacterial DNA polymerases tightly bind uracil-containing DNA. *J Biol Chem* **271:** 17692–17696.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658–1659.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* **101:** 11013–11018.

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

Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., and Rector, T. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449:** 83–86.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003) Bacterial photosynthesis genes in a virus. *Nature* **424:** 741.

Marine, R., Polson, S.W., Ravel, J., Hatfull, G., Russell, D., and Sullivan, M., *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:** 8071–8079.

Millard, A., Clokie, M.R., Shub, D.A., and Mann, N.H. (2004) Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* **101:** 11007–11012.

Neufeld, J.D., Vohra, J., Dumont, M.G., Lueders, T., Manefield, M., and Friedrich, M.W. (2007) DNA stable-isotope probing. *Nat Protoc* **2:** 860–866.

Roche Diagnostics GmbH (2009a) GS FLX Titanium General Library Preparation. Method Manual USM-0048.B.

Roche Diagnostics GmbH (2009b) Amplicon Fusion Primer Design Guidelines for GS FLX Titanium Series Lib-A Chemistry. Technical Bulletin TCB No. 013-2009: 1–3.

Rodrigue, S., Malmstrom, R.R., Berlin, A.M., Birren, B.W., Henn, M.R., and Chisholm, S.W. (2009) Whole genome amplification and *de novo* assembly of single bacterial cells. *PLoS ONE* **4:** e6864.

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., and Alm, E.J. (2010) Unlocking short read sequencing for metagenomics. *PLoS ONE* **5:** e11840.

Rodriguez-Valera, F., Martin-Cuadrado, A.B., Rodriguez-Brito, B., Pasić, L., Thingstad, T.F., and Rohwer, F. (2009) Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* **7:** 828–836.

Sambrook, J., and Russell, D.W. (2001) *Molecular Cloning: A Laboratory Manual*, 3rd edn. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press.

Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005) Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3:** e144.

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

Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., and Weigele, P.R. (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* **12:** 3035–3056.

Thompson, J.R., Marcelino, L.A., and Polz, M.F. (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR. *Nucleic Acids Res* **30:** 2083–2088.

Vega Thurber, R. (2009) Current insights into phage biodiversity and biogeography. *Curr Opin Microbiol* **12:** 582–587.

Woyke, T., Xie, G., Copeland, A., González, J.M., Han, C., and Kiss, H. (2009) Assembling the marine metagenome, one cell at a time. *PLoS ONE* **4:** e5299.

Wu, J.Y., Jiang, X.T., Jiang, Y.X., Lu, S.Y., Zou, F., and Zhou, H.W. (2010) Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiol* **10:** 255.

Wu, L., Liu, X., Schadt, C.W., and Zhou, J. (2006) Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. *Appl Environ Microbiol* **72:** 4931–4941.

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

Zerbino, D.R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., and Chisholm, S.W. (2006) Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24:** 680–686.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Comparison of DNA polymerase efficiency and effect of reconditioning PCR on two phage lysates, TUSD #20 and #23. LA-TaKaRa yielded more product and with a broader size range than PFU Turbo HotStart. Original PCR product is diluted 10× for input in reconditioning reaction, and thus results in a 10-fold increase in product. Also, note the enrichment for high molecular weight product following reconditioning.

Fig. S2. %G + C range of virus communities, as seen using various amplification (linker amplification, phi29 multiple displacement amplification, and linker amplified shotgun libraries) and sequencing platforms (454 pyrosequencing, Sanger sequencing). (A) in-house sequence data, (B) and (C) are identifiable by name and available on the CAMERA web portal, including all metadata and publication references.

Fig. S3. The magnitude of difference in %G + C bias between unamplified and amplified treatments was assessed by the magnitude of difference between the integrated area under their %G + C-bias plot curves (Fig. 3). As cycle number increased (and quantity of starting material decreased), there was a slight trend in the deviation of amplified from unamplified treatments.

Fig. S4. Quality score profile of Illumina reads from 20 pooled freshwater cyanophage genomes (L. Deng and M.B. Sullivan, unpubl. data) generated from a linker amplified library. Boxplot generated by Fastx Toolbox.

Fig. S5. Log-linear relationship between PCR cycle numbers and starting DNA, as it is diluted to extinction. Test was performed with dilution of a single phage genome, H105/1.

Table S1. Detailed break-down of linker amplification method, including time and cost estimates from sample to sequence for a typical 20 l aquatic virus sample.

Table S2. Comparison of DNA polymerase sensitivity and specificity.

Table S3. Comparison of sheared DNA size-fractionation techniques.

Table S4. Evaluation of linker amplification efficiency.

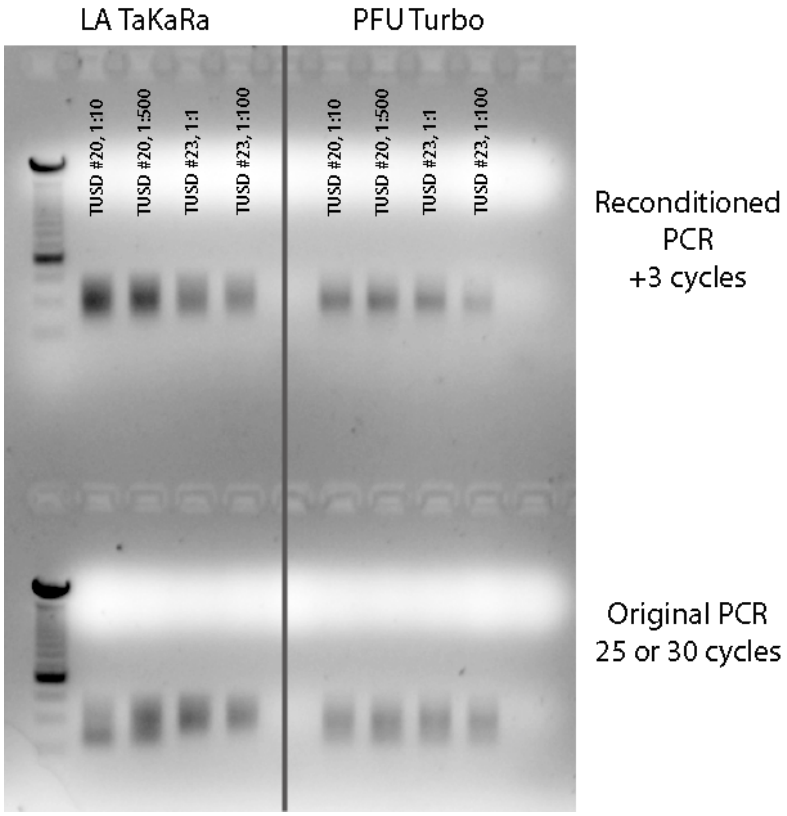
Table S5. Increase in DNA as a result of linker amplification, from both environmental and phage lysate DNA preps.

Table S6. Two-tailed paired Student's *t*-tests comparing the integrated areas between unamplified read depth curve and curve of each amplified treatment.

Table S7. Two-tailed paired Student's *t*-tests comparing the integrated areas between unamplified %G + C bias curve of unamplified and each amplified treatment for H105/1.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Supplementary Figure 1. Comparison of DNA polymerase efficiency and effect of reconditioning PCR on two phage lysates, TUSD #20 and #23. LA-TaKaRa yielded more product and with a broader size range than PFU Turbo HotStart. Original PCR product is diluted 10X for input in reconditioning reaction, and thus results in a 10-fold increase in product. Also, note the enrichment for high molecular weight product following reconditioning.



Supplementary Table 5. Increase in DNA as a result of Linker Amplification, from both environmental and phage lysate DNA preps. Starting DNA concentration reflects that after the shearing and linker ligation steps.

	Sample	**DNA ng/ul**	**ng DNA (in 20 µl PCR)**[1]	**ng DNA after PCR**[2]	**fold-increase**
Environmental	SIO-3 NT	0.34	6.8	88.53	13
	SIO-4 NT	0.39	7.8	165.43	21
	SIO-3 CsCl	0.25	5	51.74	10
	SIO-4 CsCl	0.21	4.2	46.15	11
	SIO-3 sucrose	0.19	3.8	1146.53	302
	SIO-4 sucrose	0.25	5	123.93	25
Clonal lysate	TUSD #20	0.05	1	1197.17	1197
	TUSD #21	0.22	4.4	1161.31	264
	TUSD #22	0.18	3.6	1361.94	378
	TUSD #23	0.1	2	1287.34	644
	TUSD #24	0.23	4.6	1218.65	265
	Phage PSS2	0.05	1	1235.82	1236

[1] PCR reactions run with **PFU-Turbo Hot Start** for 25 cycles

[2] PCR product pooled, purified.

Supplementary Table 6. Two-tailed paired Student's t-tests comparing the integrated areas between unamplified read depth curve and curve of each amplified treatment. Probabilities are reported. To test for the effect of cycle number on degree of deviation from unamplified read depth, a test was performed for each combination of cycle numbers; none were significant, with all p > 0.05.

All treatments

cycles	*15*	*18*	*20*	*25*	*30*
15					
18	*0.819*				
20	*n.a.*	*n.a.*			
25	*0.212*	*0.233*	*n.a.*		
30	*0.103*	*0.337*	*n.a.*	*0.611*	

Non-reconditioned

cycles	*15*	*18*	*20*	*25*	*30*
15					
18	*0.225*				
20	*n.a.*	*n.a.*			
25	*0.239*	*0.188*	*n.a.*		
30	*0.168*	*0.304*	*n.a.*	*0.446*	

Reconditioned

cycles	*15*	*18*	*20*	*25*	*30*
15					
18	*0.282*				
20	*n.a.*	*n.a.*			
25	*0.718*	*0.131*	*n.a.*		
30	*0.556*	*0.136*	*n.a.*	*0.926*	

Supplementary Table 7. Two-tailed paired Student's t-tests comparing the integrated areas between unamplified %G+C bias curve of unamplified and each amplified treatment for H105/1. Significant (p <0.05) values are yellow; highly significant (p <0.001) values are orange. "n.a." indicates not enough data to test, as the 20-cycle treatment was dropped from the analysis due to failed sequencing.

(a) Probability that differences between cycle numbers and starting DNA amounts is due to chance.

(b) Probability that differences between reconditioned and non-reconditioned samples is due to chance. All reconditioned samples are compared to all non-reconditioned ("all treatments"); reconditioning is also tested within replicates of each cycle number.

(a) **All data: %G+C range 0.31-0.55**

All treatments (n = 6 for all)

cycles	*15*	*18*	*20*	*25*	*30*
15					
18	*0.279*				
20	*n.a.*	*n.a.*			
25	*0.641*	*0.151*	*n.a.*		
30	*0.954*	*0.037*	*n.a.*	*0.444*	

Non-reconditioned (n = 3 for all)

cycles	*15*	*18*	*20*	*25*	*30*
15					
18	*0.001*				
20	*n.a.*	*n.a.*			
25	*0.211*	*0.102*	*n.a.*	*n.a.*	
30	*0.101*	*0.175*	*n.a.*	*0.460*	

Reconditioned (n = 3 for all)

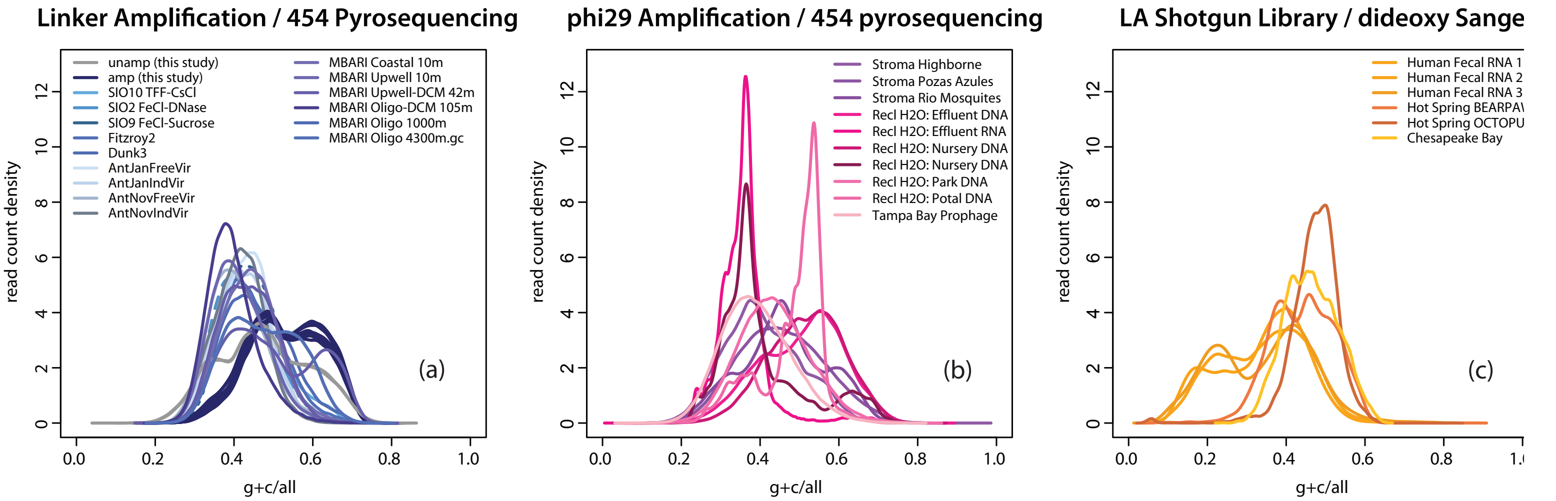
cycles	*15*	*18*	*20*	*25*	*30*
15					
18	*0.986*				
20	*n.a.*	*n.a.*			
25	*0.726*	*0.706*	*n.a.*		
30	*0.312*	*0.204*	*n.a.*	*0.021*	

(b) **All data: %G+C range 0.31-0.55**

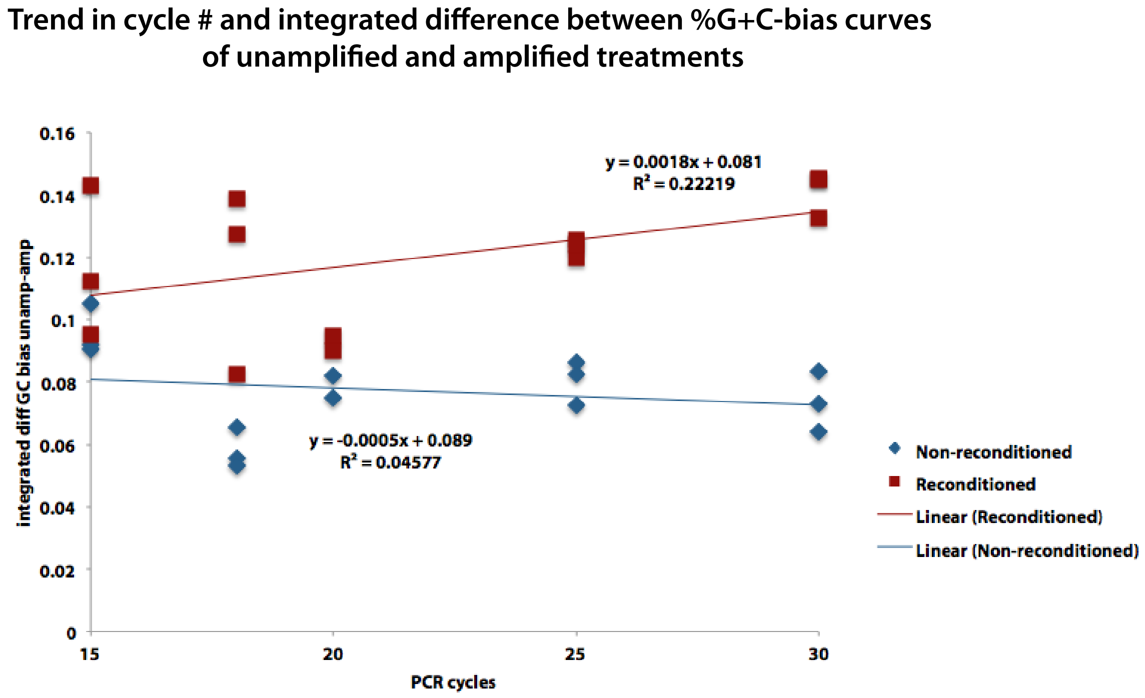
Non-recond. vs. Reconditioned

all treatments (n = 14)	4.2E-05
15 cycles, 10 ng (n = 3)	*0.314*
18 cycles, 1 ng (n = 3)	*0.057*
20 cycles, 0.1 ng (n = 3)	*n.a.*
25 cycles, 10 pg (n = 3)	*0.010*
30 cycles, 1 ng (n = 3)	*0.019*

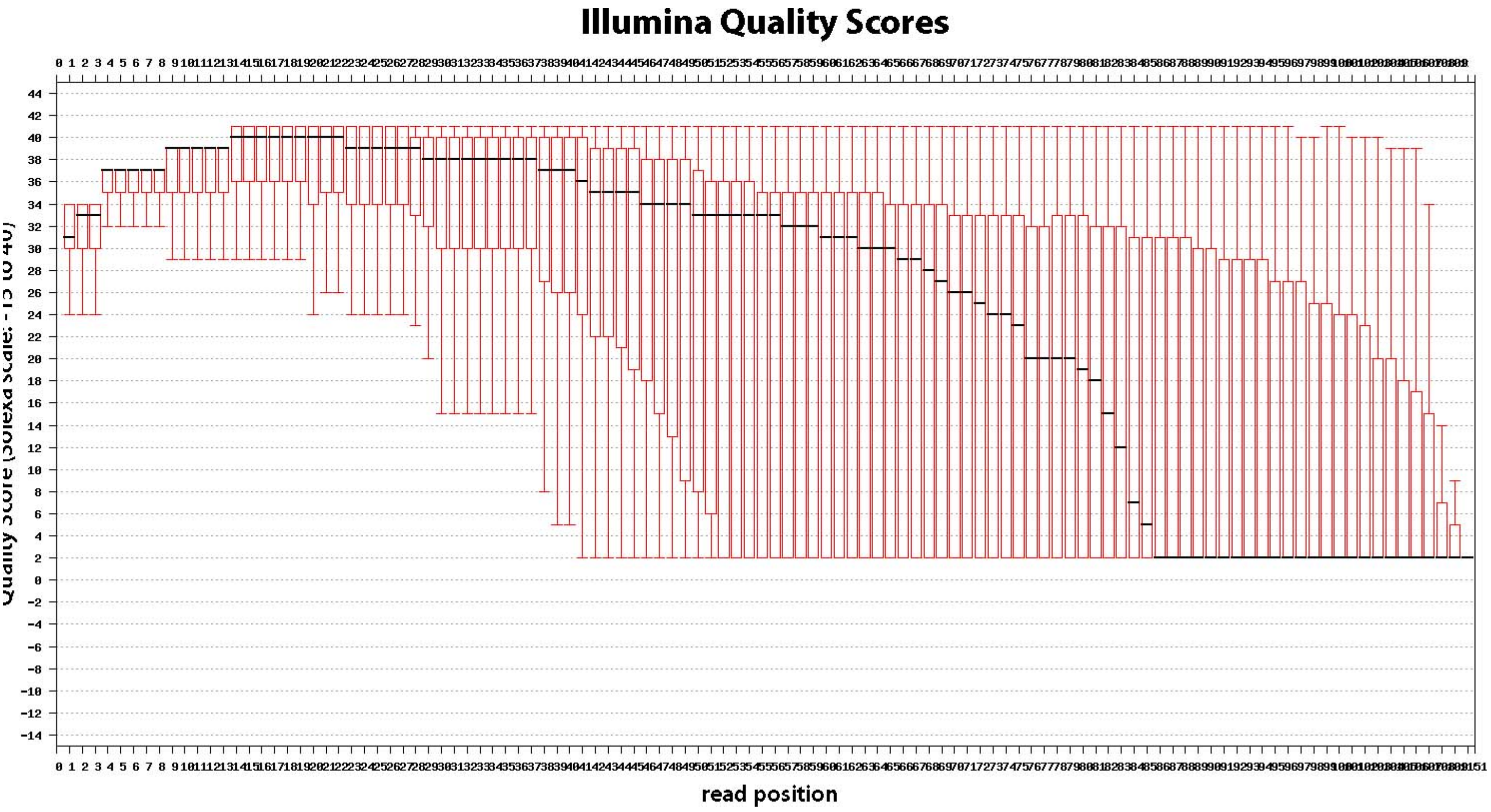
Supplementary Figure 2. %G+C range of virus communities, as seen using various amplification (linker amplification, phi29 multiple displacement amplification, and linker amplified shotgun libraries) and sequencing platforms (454 pyrosequencing, Sanger sequencing). (a) in-house sequence data, (b) and (c) are identifiable by name and available on the CAMERA web portal, including all metadata and publication references.



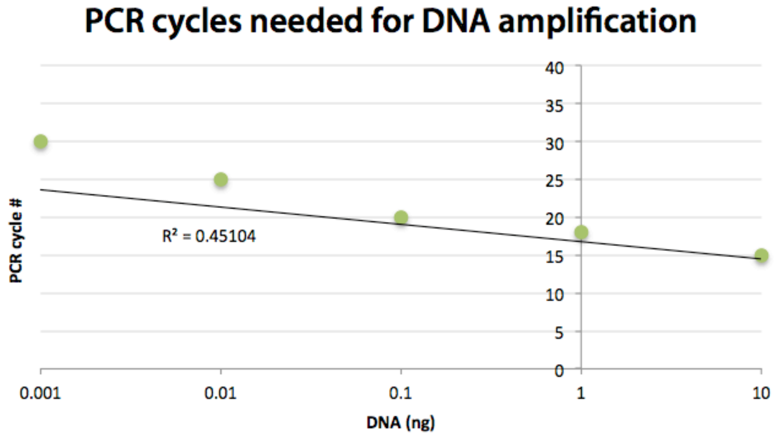
Supplementary Figure 3. The magnitude of difference in %G+C bias between unamplified and amplified treatments was assessed by the magnitude of difference between the integrated area under their %G+C-bias plot curves (Figure 3). As cycle number increased (and quantity of starting material decreased), there was a slight trend in the deviation of amplified from unamplified treatments.



Supplementary Figure 4. Quality score profile of Illumina reads from 20 pooled freshwater cyanophage genomes (Deng and Sullivan, unpublished) generated from a linker amplified library. Boxplot generated by Fastx Toolbox.



Supplementary Figure 5. Log-linear relationship between PCR cycle numbers and starting DNA, as it is diluted to extinction. Test was performed with dilution of a single phage genome, H105/1.



Supplementary Table 1. Detailed break-down of linker amplification method, including time and cost estimates from sample-to-sequence for a typical 20 L aquatic virus sample.

1.	**Ligation**	*time/sample*	*$/sample*	*kit used*
a	Ultrasonic shearing of DNA to 400-800 bp, Covaris	1 week*	$20**	*time depends on turn-around of facility, actually process approx. 5 min; **price includes Agilent Bioanalyzer fragment sizing
b	Concentrate Sheared DNA	15-20 min	$3.40	Amicon Ultra-0.5ml 30K (UFC503096)
c	End repair DNA	1 hr	$5	Epicentre Biotechnologies
d	Clean-up reaction	10 min	$2.30	Qiagen MinElute or QiaQuick (#28104)
e	Ligate Fwd and Rev linker to DNA	2 hr	$2.40	Fast-Link Ligation Kit, Epicentre (LK0750)
f	Clean-up reaction	10 min	$2.30	Qiagen MinElute or QiaQuick
f-i	Size Fractionation: SPRI Beads	20 min	$0.85	Agencourt AMPure XP; Beckman Coulter A63880
	or			
f-ii	Size Fractionation: gel-sizing	3 hr	$1.40	Seakem GTG Agarose
	DNA recovery: gel extraction	0.5 hr	$2.20	QiaQuick MinElute gel extraction kit
	or			
f-iii	Size Fractionation: Pippin Prep	5 min	$11.25	Sage Science cassette kit
2.	**Amplification**	*time/sample*	*$/sample*	*kit used*
a	Small-scale PCR titration: determine optimal cycle #	1-2hr	$26.30	LA TaKaRa HS polymerase
b	Large-scale PCR	1-2hr		
c	Reconditioning PCR	0.5 hr		
d	PCR clean-up	10 min	$2.50	Qiagen MinElute or QiaQuick
e	Pico Green Quantification	0.5hr	$1.00	
TOTAL		**1-2 days work (1 week wait*)**	**$67.50**	

Supplementary Table 2. Comparison of DNA polymerase sensitivity and specificity. PFU Turbo HotStart out-performed LA-TaKaRa in 13 virus samples (blue), while the opposite held for one microbial sample (orange). However, the sensitivity comes at a cost, as PFU 'no template control' amplified at 30 and 35 cycles (red), while TaKaRa did not. All samples had been subjected to the LA protocol through ligation to Linker A and amplification (Supp. Table 1). ✚, amplification confirmed by strong band in post-PCR gel; −, no detectable amplification; ±, faint band indicating small degree of amplification; n.d., no data. ++microbial samples were sheared by nebulization, rather than Covaris; **below detection limit of PicoGreen assay, which in practice is 1-5 pg.

			20 cycles		*25 cycles*		*30 cycles*		*35 cycles*	
	Sample	**Starting DNA (pg/µl)**	**PFU**	**TAK**	**PFU**	**TAK**	**PFU**	**TAK**	**PFU**	**TAK**
Trial 1	Tara #23 DCM 1:100	229	✚	✚	✚	✚	✚	✚	✚	✚
	Tara #30 DCM 1:100	1495	✚	✚	✚	✚	✚	✚	✚	✚
	Fitzroy viral 1:100	1.9	✚	✚	✚	✚	✚	✚	✚	✚
	Dunk viral 1:100	4.9	✚	✚	✚	✚	✚	✚	✚	✚
	Sept/09 B2O micro (16S) 1:100++	17.4	✚	✚	✚	✚	✚	✚	✚	✚
	Aug/09 P26 10m	< 1 **	✚	✚	✚	✚	✚	✚	✚	✚
	Jun/09 P26 10m	< 1	✚	✚	✚	✚	✚	✚	✚	✚
	Jun/09 P4 10m 1:10	855	✚	✚	✚	✚	✚	✚	✚	✚
	Jun/08 P26 10m	6.3	−	−	±	±	✚	✚	✚	✚
	Feb/09 P26 1000m	6	−	−	✚	±	✚	✚	✚	✚
	Jun/09 P4 500m	2.2	−	−	✚	±	✚	✚	✚	✚
	Jun/09 P12 2000m	< 1	−	−	−	−	✚	−	✚	✚
	Dec/09 Line67 Open DCM	872	✚	✚	✚	✚	✚	✚	✚	✚
	Apr/09 SIO Rep 3-NT	175	±	±	✚	✚	✚	✚	✚	✚
	No Template Control	n.a.	−	−	−	−	✚	−	✚	−
Trial 2	TUSD #20 1:30	< 1	−	−	✚	−	✚	✚	✚	✚
	TUSD #20 1:300	< 0.1	−	−	−	−	✚	±	✚	✚
	TUSD #23 1:30	80.4	−	−	±	−	✚	✚	✚	✚

	TUSD #23 1:300	8.4	−	−	−	−	+	−	+	+
	Jun/08 P26-10m	< 1	−	−	+	−	+	+	+	+
	Feb/09 P26 2000m	3.7	−	−	+	−	+	+	+	+
	Aug/08 P26 10m	< 1	−	−	+	−	+	+	+	+
	Aug/08 P26 1000m	< 1	−	−	−	−	+	±	+	+
	Feb/09 P26 10m	3.2	−	−	+	±	+	+	+	+
	Jun/08 P26 2000m	< 1	−	−	−	−	+	−	+	+
	No Template Control	n.a.	−	−	−	−	−	−	−	−
Trial 3	Sept/09 B2 micro (16S)	28	−	±	n.d.	+	−	+	−	+
	Jun/10 SIO micro 1:100 (16S)	638	−	−	±	±	+	+	+	+
	Jun/10 SIO micro 1:1000 (16S)	1.9	−	−	−	−	±	±	+	+
	Dec/09 Fitzroy micro 1:10	2900	+	+	+	+	+	+	+	+
	Dec/09 Dunk micro 1:200	1220	+	+	+	+	+	+	+	+
	No Template Control	n.a.	n.d.	n.d.	−	−	n.d.	n.d.	−	−
Trial 4	TUSD phage mix (n = 2)	100000	n.d.	n.d.	+	+	+	+	+	+
	TUSD phage mix (n = 2)	10000	n.d.	n.d.	+	+	+	+	+	+
	TUSD phage mix (n = 2)	3300	n.d.	n.d.	±	+	+	+	+	+
	TUSD phage mix (n = 2)	1000	n.d.	n.d.	−	±	+	+	+	+
	TUSD phage mix (n = 2)	200	n.d.	n.d.	−	−	±	+	+	+
	No Template Control	n.a.	n.d.	n.d.	−	−	−	−	−	−

Supplementary Table 3. Comparison of sheared DNA size-fractionation techniques. All data are post-size selection and determined by the Agilent Bioanalyzer. The three tests were performed with 13.7 µl of DNA (183 ng DNA total) from the same pool of sheared starting DNA (Fig. 2a)

Sample	Volume recovered in target 400-600 bp range (µl)	[DNA] recovered in target 400-600 bp range (ng/µl)	DNA recovered in target 400-600 bp range (ng)	% recovery of sheared starting DNA (183 ng start)	*TARGET RECOVERY EFFICIENCY* (% DNA recovered in 400-600 bp)	Actual size range recovered (*Fig. 2b*)
Pippin Prep A	42	1.27	53.34	29	94	*400-600 bp*
Pippin Prep B	39	1.24	48.36	26	95	
Pippin Prep C	39	1.37	53.43	29	96	
Standard Gel 1A	20	3.95	79	43	64	*400-750 bp*
Standard Gel 2A	20	4.4	88	48	74	
Standard Gel 3A	20	4.39	87.8	48	72	
Ampure Bead 1	20	5.91	118.2	65	49	*400-950 bp*
Ampure Bead 2	20	5.05	101	55	46	
Ampure Bead 3	20	5.34	106.8	58	50	

Supplementary Table 4. Evaluation of Linker Amplification efficiency. Percent DNA recovery (post-shearing) of the concentration, blunt-end repair/reaction clean-up, Linker-A ligation/reaction clean-up, and size fractionation (gel)/recovery of DNA from gel. This is the template DNA used in PCR amplification.

	Sample ID	**ng DNA sheared**	**ng DNA recovered**	**% recovery**
Environmental samples	SIO-3 NT	151.7	13.42	8.8
	SIO-4 NT	148.5	15.78	10.6
	SIO-3 CsCl	150.4	9.8	6.5
	SIO-4 CsCl	129.6	8.09	6.2
	SIO-3 sucrose	148.8	7.7	5.2
	SIO-4 sucrose	149.9		3.5
	TUSD #20a	149.3	5.32	3.6
Clonal phage lysates	TUSD #20b	49.8	2.11	4.2
	TUSD #21	151.2	7.64	5.1
	TUSD #22	149	6.27	4.2
	TUSD #23a	148.5	6.3	4.2
	TUSD #23b	54	1.86	3.4
	TUSD #24	149.6	8.22	5.5
	PSS-2	42.4	3.87	9.1
	Avg. ± s.d.	54	1.86	**5.7 ± 2.3**