The ISME Journal (2015) 9, 2162–2177

www.nature.com/ismej

ORIGINAL ARTICLE

Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis

Benjamin Bolduc[1,2], Jennifer F Wirth[1,2], Aurélien Mazurie[3] and Mark J Young[1,2]
[1]*Thermal Biology Institute, Montana State University, Bozeman, MT, USA;* [2]*Department of Plant Sciences and Plant Pathology and, Montana State University, Bozeman, MT, USA and* [3]*Bioinformatics Core Facility, Montana State University, Bozeman, MT, USA*

Understanding of viral assemblage structure in natural environments remains a daunting task. Total viral assemblage sequencing (for example, viral metagenomics) provides a tractable approach. However, even with the availability of next-generation sequencing technology it is usually only possible to obtain a fragmented view of viral assemblages in natural ecosystems. In this study, we applied a network-based approach in combination with viral metagenomics to investigate viral assemblage structure in the high temperature, acidic hot springs of Yellowstone National Park, USA. Our results show that this approach can identify distinct viral groups and provide insights into the viral assemblage structure. We identified 110 viral groups in the hot springs environment, with each viral group likely representing a viral family at the sub-family taxonomic level. Most of these viral groups are previously unknown DNA viruses likely infecting archaeal hosts. Overall, this study demonstrates the utility of combining viral assemblage sequencing approaches with network analysis to gain insights into viral assemblage structure in natural ecosystems.
The ISME Journal (2015) **9,** 2162–2177; doi:10.1038/ismej.2015.28; published online 30 June 2015

Introduction

Viral metagenomics has rapidly expanded our understanding of viral diversity. More than 40 viral metagenomics studies have been published in the past decade ranging from marine, freshwater, arctic, soil, human feces, gut and oral cavity environments (Rosario and Breitbart, 2011; Fancello *et al.,* 2012; Mokili *et al.,* 2012). Each study reveals not only new insights into the interplay between viruses and their hosts, but expands our understanding of the 'unknown virosphere.' A common trend emerging from these studies is the enormous viral diversity, both in terms of the number of virus types and their gene content. More than 10^{31} virus particles are estimated to exist in the oceans (Wilhelm and Suttle, 1999) with $>1.2\times10^{5}$ different genotypes (Angly *et al.,* 2006). Like cellular metagenomics analysis, the current challenge with viral metagenomics analysis is to move beyond the 'who-is-there' analysis to a more functional analysis of the role of the vast viral diversity in ecology and evolution in natural environments. As a step toward this long-term goal, there is an immediate need to improve our ability to define viral assemblage structure.

Correspondence: MJ Young, Thermal Biology Institute and Department of Plant Sciences and Plant Pathology, Montana State University, 119 PBB, Bozeman, MT 59717, USA. Email: myoung@montana.edu
Received 16 June 2014; revised 29 December 2014; accepted 12 January 2015; published online 30 June 2015

Despite the increasing number of environments examined, few studies have addressed the composition and interconnectivity of these viral assemblages as a whole. A variety of bioinformatic tools have been developed to analyze the structure and/or diversity of viral assemblages from metagenomics data sets (reviewed extensively in Fancello *et al.,* 2012; Mokili *et al.,* 2012). Although these tools provide taxonomic classification, functional assignment and estimates of virus community structure and diversity, they are frequently dependent on sequence comparisons against databases of known sequences or function, and do not account well for the large variation in sequences and functions typically associated with viral families. One promising direction is the use of protein-clustering techniques to organize the viral sequence space (Williamson *et al.,* 2008; Hurwitz and Sullivan, 2013; Solonenko *et al.,* 2013). These studies use protein clustering not only to aid in organizing the 'unknown' sequences constituting the majority of many metaviromes, but offers a means to measure viral assemblage diversity (Hurwitz *et al.,* 2012). Although this approach represents a significant advance forward, it is somewhat dependent on having assembled large contigs from metagenomic data sets and limited sequence variation within a virus population (Ojosnegros *et al.,* 2011). In natural environments, viruses may not exist as a single genotype, but instead as a cloud of highly related genotypes. When this natural sequence variation is

combined with the inherent sequencing errors of current sequencing technologies, complex viral assemblages producing large data sets and assembly algorithms not optimized for the variation, fragments of assembled genomes often result.

The acidic hot springs in Yellowstone National Park, USA (YNP) provide relatively simple, low-complexity environments that are dominated by a relatively few microbial species (Inskeep *et al.,* 2010). High temperatures (>80C), low pH (pH<3) conditions generally favor *Archaea*-dominated communities (Bolduc *et al.,* 2012). Bacteria and eukaryotes are few or in many cases, absent (Reysenbach *et al.,* 1994; Blank *et al.,* 2002; Kozubal *et al.,* 2012; Macur *et al.,* 2012; Jay *et al.,* 2013). These extreme environmental conditions favor not only the *Archaea* but also result in a relatively simplified microbial community structure. A number of viruses, exclusively archaeal, have been isolated out of these environments (Rice *et al.,* 2001; Mochizuki *et al.,* 2011; Pina *et al.,* 2011; Prangishvili, 2013). These conditions offer a tractable system to study virus–host relationships, as well as viral and host community structure and stability.

We have investigated the viral assemblage structure and stability of a YNP high temperature acidic hot spring using a network-based approach. We find that this approach allows us to define the viral assemblage structure and to determine that it is relatively stable over a 5-year sampling period.

Materials and methods

Sample sites

A YNP high temperature (72–93 °C), acidic (pH 2.0–4.5) hot spring was selected for this study based on previous work (Bolduc *et al.,* 2012). The Nymph Lake hot spring site 10 (NL10: 44.7536°N, 110.7237°W) was sampled at 10 different time points over a 5-year period (Table 1). NL10 is located in a relatively new thermal basin that has developed over the past 15 years (Lowenstern *et al.,* 2005).

Sample collection for pyrosequencing data sets

Hot spring samples for viral sequencing were collected by two different methods. The first method involved filtering 100 ml of hot spring water through two successive 0.8/0.2 μm filters (Pall Corporation, Port Washington, NY, USA) directly into four sterile 23 ml ultracentrifuge tubes, transported at ambient temperature back to the lab within 4 h and immediately centrifuged at 100 000 × *g* for 2 h. The supernatant was removed and the resulting virus-enriched pellet resuspended in a small volume (0.25–1.0 ml) of sterile water. In the second method, approximately 1.2-l of hot spring water was filtered as above. The filtrate was then spin concentrated (100 000 molecular weight cutoff) to a volume of 500 μl. Previous analysis had shown that free nucleic acids were not stable in these high temperature acidic environments, so prior treatment of samples with nucleases was not required to reduce this source of non-virion packaged DNA. Total nucleic acids from the viral pellet (method 1) and the concentrate (method 2) were extracted using DNA/RNA Viral Extraction Kit (Invitrogen by Life Technologies, Carlsbad, CA, USA) and eluted to a final volume of 60 μl. Extracted nucleic acids were RNase-treated with RNase ONE (Promega, Madison, WI, USA; 1U over 30 min) and re-extracted. These viral nucleic acids were amplified before sequencing by multiple displacement amplification (New England Biolabs, Ipswich, MA, USA) or whole-genome amplification (Sigma-Aldrich, St Louis, MO, USA). All viral samples from Sept 2008 to Sept 2009 were sequenced by the University of Illinois sequencing center using GS FLX 454 sequencing. The remaining samples were sequenced by the Broad Institute (Massachusetts Institute of Technology) using GS FLX 454 sequencing. Potential cellular DNA contamination was removed through a bioinformatic approach described below.

Sample collection for Illumina data sets

Approximately 60-l of hot spring water was collected and transported at ambient temperatures back to the

Table 1 Study site sampling points and characteristics

Date	*Metagenome*	*Amplification method*	*Temperature (°C)*	*pH*
September 2008	NL10 0809		90	4
January 2009	NL10 0901		92	4.3
March 2009	NL10 0903	MDA	84	5.1
April 2009	NL10 0904		92	4.5
August 2009	NL10 0908		90	4
September 2009	NL10 0909		92	4
October 2009	NL10 0910		83	3
February 2010	NL10 1002	WGA	87	3.1
June 2010	NL10 1006		89	4.5
December 2012	1.3 g cm^{-3}	Ovation	88	3
	1.4 g cm^{-3}			
	1.5 g cm^{-3}			

Abbreviations: MDA, multiple displacement amplification; WGA, whole-genome amplification.

lab and immediately filtered using a 0.4-μm filter (Millipore, Billerica, MA, USA) to remove cells and other debris. To concentrate the viruses, a previously described method for chemical flocculation of ocean viruses using $FeCl_3$ was modified (John *et al.,* 2010). Owing to the already low pH and high Fe^{3+} concentration, it was unnecessary to add additional $FeCl_3$. Instead, the filtered sample was pH adjusted to between pH 4.5 and 5.0 using NaOH, causing a visible orange-tint during precipitation. This flocculate was re-filtered onto fresh 0.4-μm filters and subsequently resuspended in a 5 mM citrate buffer (pH 3.0). This Fe-concentrated viral fraction was then centrifuged at $4000 \times g$ for 10 min to remove residual debris from the solution and the supernatant pelleted by centrifugation at $100\,000 \times g$ for 2 h. The pellet was resuspended in 1000-μl RNase/DNase-free H_2O (Zymo Research, Irvine, CA, USA) and applied as an overlay onto a preformed cesium chloride density step gradient with steps at 1.3, 1.4 and 1.5 g/ml. Based on previous research (Thurber *et al.,* 2009), it was believed that most viruses would fall within one of these interfaces. The gradient was spun at 15 000 r.p.m. in a Beckman MLS50 rotor (Beckman Coulter, Brea, CA, USA) for 2-h and separated in 500-μl fractions, taking each of the four interfaces. Following fractionation, each sample was dialyzed against citrate buffer in SLIDE-A-LYZER mini dialysis units (Thermo Fisher Scientific, Waltham, MA, USA).

Samples were extracted using ZR viral RNA/DNA kit (Zymo Research) according to the manufacturer's instructions. Extracted nucleic acids were RNase-treated with RNase ONE (Promega; 1U over 30 min) and re-extracted. Extracted viral DNA was then amplified using the Ovation Ultralow library system (NuGEN Technologies, San Carlos, CA, USA) according to the manufacturer's instructions. The amplified nucleic acids were sequenced by the University of Illinois sequencing center using the Illumina MiSeq v3 system (Illumina, San Diego, CA, USA) with paired-end reads (2 x 300 nt).

Processing and assembly of 454 reads from viral metagenomes

Briefly outlined, adapter and primer sequences were trimmed from individual sequencing reads using Tagcleaner (http://tagcleaner.sourceforge.net/) and subsequently filtered for quality using a python script that used a sliding quality window with average read quality of 25 across 50 bp (all custom python scripts are available on github.com/). To aid assembly, highly duplicated sequencing reads were identified using CD-HIT-454 (Niu *et al.,* 2010) at 99% identity and reduced to the single largest, representative read. Preliminary analysis also revealed a proportion of sequence reads from the viral sets that overlapped with sequence reads from a corresponding cellular fraction. As it was not evident if these overlapping reads represented viral sequences present in the cellular fraction or contaminating cellular DNA in the viral fraction, it was decided to remove these reads from the data set before assembly. These reads were removed by aligning sequences using Newbler gsMapper version 2.7 (Roche 454 Life Sciences, Branford, CT, USA) at 70% identity over 50 bp.

Remaining reads from each time point were assembled using the 454 gsAssembler software program version 2.7 (Roche 454 Life Sciences) with default parameters, except for the minimum identity between sequences (98%) and the minimum overlap (50 bp; sequencing and assembly statistics are presented in Supplementary Table S1). These stringent assembly conditions were used to prevent misassembly between reads of different viruses (Breitbart *et al.*, 2002). An additional assembly was generated that cross-assembled the filtered reads (briefly described in Bolduc *et al.,* 2012) of the entire DNA data set (hereafter referred to as cross-assemblies). This cross-assembly was used for analysis by the Virome pipeline (described below).

Processing and assembly of Illumina reads from viral metagenomes

Reads were pre-binned according to their sequencing barcodes. Standard MiSeq V3 adapter sequences were trimmed from individual sequencing reads using Trimmomatic V0.32 (http://www.usadellab.org/cms/?page=trimmomatic) and quality checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Primers used in the Ovation Ultralow library system amplification were identified as over-represented k-mers. To remove these sequences, Jellyfish (Marcais and Kingsford, 2011) was used to identify all over-represented k-mers between the range of 16 and 32 and then a python script compared all identified k-mers common with the viral metaviromes. The common k-mers were assembled using Minimo from the amos package (version 1.5; http://sourceforge.net/ projects/amos/ and Treangen *et al.,* 2002) at 100% nucleotide identity and an overlap length of 50% and the adapter sequences were trimmed with Trimmomatic. Following trimming, DNA sequences were assembled using IDBA-UD (Peng *et al.,* 2011,2012) using a k-mer range of 28–124 and default parameters.

Analysis of viral alpha diversity and taxonomic identification

Viral reads were analyzed using a combination of GAAS (version 0.17; http://sourceforge.net/projects/gaas; Angly *et al.,* 2009), Circonspect (version 0.2.6; http://sourceforge.net/projects/circonspect/) and PHACCS (version 1.1.3; http://sourceforget.net/projects/phaccs/; Angly *et al.,* 2005). Community structure for individual, mixed (that is, averaged), and cross-assembled samples were modeled using

PHACCS based on the contig spectra and average genome sizes using available models. Cross-assembly generated contigs over 300 bp, which were uploaded to the VIROME web server and analyzed as previously described (Wommack *et al.,* 2012).

Analysis of viral diversity of Illumina data sets
Rarefaction curves were constructed by an in-house version analogous to Metavir (Roux *et al.,* 2011). Briefly, 250 000 paired-end, trimmed reads are randomly selected from each metagenome and rarified in 12 500 sequence increments, for a total of 20 rarified sets. These sets are clustered with Uclust (Edgar, 2010) at 97%, 90% and 75% nucleotide identities. The resulting clusters and input sequences are plotted using the matplotlib library for python (Hunter, 2007).

Network analysis: optimization
To determine the optimum parameters for use in the final network, the number of viral groups and data represented were evaluated as a function of both high-scoring pair (HSP) length and minimum e-value cutoff, as these two parameters were the most influential in changing the network topology. For this optimization, the HSP-lengths were varied between 50 and 500 in 25-bp increments and e-values varied from e^{-10} to e^{-50}. In addition, the minimum number of contig members in each viral assembly group was evaluated using 50 and 100 contig thresholds. Overall, these optimization runs evaluated 162 total variations (Supplementary Figure S1).

Network analysis
Contigs from each viral assemblage were compared through an 'all-verses-all' comparison using BLASTN with default parameters (except that the max target sequences were set to 10 000 and an e-value cutoff minimum of 10^{-5}). HSPs were filtered to remove those below a 50-bp minimum alignment length, 75% nucleotide identity and e-value of 10^{-10}. Filtered HSPs were subsequently converted to a directed graph, with contigs as graph nodes, their HSP connecting them (the query-target relationship) as graph edges and their e-value stored as the edge property, *weight*. The weight serves as a measure of strength connecting two contigs to each other. The community detection algorithm (the Louvain method; Blondel *et al.,* 2008) was then applied to the graph. This method utilizes both edge weight and edge numbers connecting contigs to maximize cluster modularity within the network. The effect is contigs separated into partitions, more highly related within a partition than between them. In this work, we refer to these viral partitions as viral *groups*. Viral groups containing <50 contigs were removed from downstream analyses because of their relatively minor impact on the results and disproportional computational requirements. Graph networks were visualized using Gephi (Bastian *et al.,* 2009) using the OpenOrd layout (developed from the force-directed Fruchterman-Reingold; Fruchterman and Reingold, 1991) and optimized using Gephi's Force Atlas 2 plugin. Both of these force-directed graph-drawing algorithms assign forces to both nodes (contigs) and their connecting edges (HSPs). Nodes are repulsed from each other as if electrically charged, while their edges are springs serving as an attractive force. The strength of the spring is correlated to the edge weight. Network statistics and topological properties, such as modularity, clustering coefficient and degree were calculated using Gephi. In the case of modularity, the Louvain method calculates its own modularity internally, which is not accessible by the user. The contributions of each viral assemblage's contigs and reads to each group were calculated during the network analysis. These contributions provide a snapshot of the viral populations over time. As controls, all viruses, including the 67 archaeal viruses and the Illumina loading control (bacteriophage ϕX174) in the NCBI RefSeq database (including the Illumina loading control, bacteriophage ϕX174) were seeded as full-length nucleotide sequences alongside the initial BLAST and the network analysis repeated. In addition to the use of full-length sequences, virus reference sequences were fragmented into 1000-bp segments with 250-bp overlaps between segments. These fragmented genomes were compared against the established network as an alternative approach to using full-length sequences. A list of the archaeal viruses used in the seeding is in Supplementary Table S2.

Network analysis controls: creation of synthetic networks consisting of both random and artificial metagenomes
To establish the fundamental biological nature of the viral groups, two artificial networks were constructed. In the first network, random sequence contigs were generated from the initial contigs using python's random module, which uses the Mersenne Twister algorithm (Matsumoto and Nishimura, 1998) as its pseudorandom number generator. These random contig sequences had similar characteristics (contig length, abundance and known error rate of the sequencing technology used) as the experimental viral metagenomic data sets. The randomized contigs were then subjected to the identical network analysis as described above, including the additional virus-based 'seeding'.

The second artificial network was constructed from a known set of viruses from the *Podoviridae* family. Briefly, 228 *Podoviridae* were downloaded from NCBI by using the NCBI's taxonomy browser, selecting the *Podoviridae* taxon. The GI numbers associated with each virus were then used to identify

their full taxonomic lineage. Lineage information was used to select 85 viral species from the list; a minimum of five members from each sub-family was randomly selected after which members from the remaining species were randomly selected. To these viruses, three archaeal viruses were added; *Sulfolobus* turreted icosahedral virus 1 (STIV-1), *Aeropyrum* coil-shaped virus (ACV) and *Acidianus* two-tailed virus (ATV). The 88 viral genomes were then fragmented using Grinder (Angly *et al.,* 2012) with the following parameters to match the NL10 viral metagenomes: total reads, 175 000; read distribution, 500 normal 100; mutation distribution, uniform; homopolymer distribution, Balzer; abundance model, powerlaw; quality levels, 30 10, mutational distribution, 1.45; chimeric percentage, 5. These parameters were found through an exhaustive approach, varying each parameter and evaluating all permutations (264 combinations; Supplementary Figure S1) through assembly with gsAssembler and selecting for the inferred read error, quality-passing reads and contig statistics (total contigs, large contigs, average total contig length and average large contig length) best matching the NL10 viral assemblages. Following assembly with gsAssembler (as above), the resulting contigs were subjected to the above network analysis and analyzed under the same conditions.

Reassembly of viral contigs based on network clusters
Owing to a high network-clustering coefficient and the number of viral group-exclusive viral reference matches, it was believed each viral group could be reduced to a set of 'representative' pan-genomes. Contigs within each viral group were re-assembled using the Cap3 assembler (Huang and Madan, 1999) with a 50-bp minimum contig overlap length and 98% minimum overlap identity (identical to the initial assembly conditions). Assembly characteristics (number of new contigs, average lengths and so on) were analyzed using a custom python script, which also handled running the Cap3 assembler program.

Network comparison of NL10 to another high temperature, acidic YNP hot spring
BLASTN was used to compare contigs generated from a previous viral metagenomic study from another high temperature, acidic hot spring in the Crater Hills region of YNP (CH041) and the viral groups present in the NL10 viral assemblages. Viral contigs from CH041 were compared with each of NL10 viral group's contigs with an e-value cutoff of 10^{-10} and minimum identity of 70%. Matches to any contig were added to its viral group count total.

Analysis of minimum sampling to describe viral assemblages through network analysis
Contigs from each metagenome were added in stepwise, time-forward manner, adding contigs to the each previous sampling point (beginning with only NL10 0901 contigs) and performing the network analysis before the addition of the next time point. This continued until all sampling points were used. The same process was repeated but in a reversed 'direction', starting with NL10 1006 and moving backward in time. This resulted in 18 additional networks that were parsed as above. Network statistics and topological properties for each network were calculated using Gephi (Table 2).

Table 2 Network analysis on stepwise-added metagenomes

# *Metagenomes*	1[a]	2[b]	3[c]	4[d]	5[e]	6[f]	7[g]	8[h]	9[i]
Contigs	211	2202	6148	9963	14 741	19 977	24 527	28 134	30 132
Edges	942	10 012	36 665	68 197	121 704	186 584	237 728	277 340	304 585
Average degree	8.929	9.094	11.927	13.690	16.512	18.680	19.385	19.716	20.217
Modularity	0.682	0.923	0.946	0.957	0.956	0.956	0.956	0.959	0.959
Viral groups	8	28	55	68	81	84	92	103	105
Avg. clustering coefficient	0.688	0.720	0.750	0.765	0.770	0.768	0.766	0.766	0.766
# *Metagenomes*	1[j]	2[k]	3[l]	4[m]	5[n]	6[o]	7[p]	8[q]	9[r]
Contigs		509	3583	9120	13 995	17 776	22 889	26 618	30 132
Edges		1344	12 711	39 773	78 338	115 037	178 631	238 101	304 585
Average degree		5.281	7.095	8.722	11.195	12.943	15608	17.89	20.217
Modularity		0.900	0.962	0.969	0.962	0.965	0.962	0.961	0.958
Viral groups		19	51	86	94	106	108	105	105
Avg. clustering coefficient		0.650	0.712	0.730	0.742	0.751	0.758	0.764	0.766

[a]0809. [b]0809-0901. [c]0809-0903. [d]0809-0904. [e]0809-0908. [f]0809-0909. [g]0809-0910. [h]0809-1002. [i]0809-1006. [j]Did not have a major population. [k]1006-1002. [l]1006-0910. [m]1006-0909. [n]1006-0908. [o]1006-0904. [p]1006-0903. [q]1006-0901. [r]1006-0809.
The 9 454 metagenomes were added one-by-one successively to the previous metagenome (1, 1+2, 1+2+3, etc.) and analyzed using the network-based analysis. The number of contigs present in the network is given under 'Contigs' with the number of connections between all contigs as 'Edges.' Average degree refers to the average number of connections incoming/outgoing from each contig and denotes relative connectivity between contigs. Modularity is a measure of separation and distinction of viral groups. The greater the value (approaches 1) the more connections tend to exist between members within a viral group than between members of differing viral groups. This also serves as a measure of the overall organizational level of the network, with values >0.3 indicating community structure. The average clustering coefficient measures how well contigs tend to associate with each other.

Scripts and data availability
Scripts used in the network analysis are available at Github (https://github.com/bolduc/blast2network). The nine 454 and three Illumina viral metagenomes have been deposited in the NCBI database under Bioproject PRJNA273640.

Results

Analysis of the viral content of single YNP hot spring was evaluated over a 5-year time period. A total of 10 viral-enriched samples were collected and analyzed by community sequencing technology over the course of a 5-year period. Nine viral metagenomes were generated during a 2-year period, which were subjected to DNA sequencing using 454 technology, whereas the 10th sample (taken 3 years after the last 454 sample) was performed using virus samples further enriched on CsCl gradients and sequenced using Illumina technology (Table 1). Although the pH and temperature of the site is relatively stable, there is a seasonal variation in the pH of the hot spring based on the amount of surface water mixing with water coming up from the deep subsurface. The influx of more neutral water likely affects the pH and geochemical composition of the site, with larger variations possible throughout the year.

Read processing and assemblies
In order to assemble high confidence viral assemblies, low-quality reads, cellular carry-over reads and highly duplicated reads were removed and subsequently assembled with high stringency parameters. Initial assembly of these reads using the 454 gsAssembler software program (version 2.7) was unable to assemble several of the viral metagenomes until the 'large and complex genome' option was enabled. This option is generally evoked during processing of large prokaryote or eukaryotic genomes and skips the assembly of high-copy (>500) repeats and the last level of detangling. The detangling phase is responsible for resolving complex graph structures created when contig ends overlap between multiple contigs, often the result of repeats. Resulting assemblies resulted in a large number of small contigs with an average contig length less than the average read length. Aligning these contigs against viral sequences known to exist in the NL10 site revealed complete genome coverage with a wide range of sequencing coverage. Such uneven sequencing coverage often prevents optimal assembly (Li and Godzik, 2006) and we suspected these uneven regions resulted in premature breaking of the contigs. To reduce the hyper-deep coverage regions, highly duplicated reads were removed at 99% identity with CD-HIT-454, reducing the number of reads by 14% (on average) for the viral metagenomes (Supplementary Table S1). Viral reads were subsequently filtered against potential cellular carry-over sequences further reducing the number of reads for the viral metagenomes by an additional 1%. This resulted in a reduction from ~1.61 million reads (620 Mbp) to ~1.44 million reads (521 Mbp) for the viral metagenomes. Similar processing of Illumina reads removed ~9% of reads, reducing the total raw reads from ~6.08 million (3.6 Gb) to 5.53 million (2.0 Gb).

Our assembly strategy improved the number and length of assembled contigs of the nine 454 data sets. The initial assembly of sequences from NL10 hot springs resulted in a large number of viral contigs under stringent conditions. Separate assemblies of the nine 454 viral metagenomes generated 50 735 total contigs (average contig length of 565 bp), assembling 81.8% of the reads, with 16.1% singletons (the remaining classified as repeats, chimeric or too short). The cross-assembly of all nine metagenomes reads produced fewer overall contigs (27 608), maintained the average contig length of 562 bp and increased the number of reads assembling into contigs (86.6% with 10.2% singletons) as compared with the individual assemblies. Assembly of the Illumina data sets generated 118 261 contigs, assembling 19% of the reads. Cross-assembling the three Illumina data sets resulted in a 65.1% overall reduction in contigs (41 332), revealing a large number of duplicated sequences between samples. To further support the concept that viral groups (described below) are representative of common and/or related viral types, each group was re-assembled using similar assembly parameters (Supplementary Table S3). In nearly all cases, assemblies of viral groups resulted in the creation of fewer and larger contigs. For example, viral group 29 contains 391 contigs, representing 0.8% of all contigs and 0.26% of the total read-bp. Reassembly reduced the number of contigs by 53.2% (183 contigs) and increased the average contig length by 42.9%. It is suspected these reductions are due to the presence of the same sequence at multiple time points, as well as overlap between sequences that are not sequenced at all time points because of insufficient sequencing depth at any individual time point, isolation or amplification biases or stochastic effects during sequencing. Overall, viral group re-assemblies resulted in a 54% decrease in contig numbers, a 28% increase in average contig size and a 30% increase in the largest contig length.

Taxonomic analysis
To reduce the computational cost associated with analyzing the nine 454 generated metaviromes, only the cross-assembled metagenome was taxonomically classified using the VIROME pipeline. Briefly, VIROME is an annotation system developed to characterize viral sequences. Sequences are first filtered for quality, the presence of ribosomal RNA and false duplicate sequences and other known contaminants such as vector sequences. Sequences are subsequently analyzed for the presence of

transfer RNA sequences and compared against UniRef, and MetaGenomes Online. Sequences not assigned to these collections have their open reading frames (ORFs) predicted and compared against five annotated protein databases (ACLAME, COG, GO, KEGG and SEED). VIROME then annotates each sequence based on best-matches against each database (Wommack *et al.,* 2012). Analysis of contigs ≥ 300 bp revealed more than half were unable to be assigned. Those sequences that were assigned were almost exclusively archaeal viruses (97.8% of all ORFs). The remaining 2.2% were to temperate phage of the order *Caudovirales*. Archaeal virus homologs matched to families known to infect hyperthermophilic *Archaea*; spread between the *Rudiviridae* (32%), including *Sulfolobus islandicus* rod-shaped virus 1 and 2 (SIRV-1, 2), *Acidianus* rod-shaped virus (ARV) and the *Lipothrixviridae* (31%), which include nearly all the *Acidianus* filamentous virus strains (AFV) and *Sulfolobus islandicus* filamentous virus (SIFV). The remaining matches were spread evenly between the *Fuselloviridae, Bicaudaviridae* and the proposed *Turriviridae*, including STIV-1 and 2, and hyperthermophilic archaeal viruses 1 and 2. It is unclear if the taxonomic assignment of the phage ORFs found in the viral fractions is due to the presence of phage through contamination or actual presence of phage within the hot springs or viral ORFs with their closest known paralogs found in cellular genomes. However, these results support previous analysis that archaeal viruses and their archaeal hosts dominate high temperature acidic hot springs.

Community composition

The community composition was modeled using PHACCS, including viral genotype richness, most abundant genotype, evenness and diversity (Supplementary Table S4). The PHACCS-analyzed metaviromes were significantly lower diversity and genotype compared with freshwater and marine metaviromes (Tamaki *et al.,* 2011; McDaniel *et al.,* 2013; Tseng *et al.,* 2013) and slightly lower than those of other springs in YNP (Schoenfeld *et al.,* 2008). The number of estimated genotypes varied between 192 and 1342 although the different estimators separate the samples into a low richness and moderate–low richness with ranges of 192 to 494, and 1145 to 1342, respectively. This is lower than the number of genotypes seen in marine (53 000; McDaniel *et al.,* 2013) and freshwater (352–22 000; Tseng *et al.,* 2013) but on par with a study in a wastewater treatment plant (423–560; Tamaki *et al.,* 2011) and lower than a neutral, boiling spring in YNP (1310–1440; Schoenfeld *et al.,* 2008). The most abundant genotypes were inversely related to the number of genotypes, ranging from 3% to 7% for the low richness and 2% to 4% for the moderate–low richness NL10 groups, in agreement with the community evenness, which was stable between 0.87 and 0.94. This is similar to the YNP hot spring and significantly lower than the wastewater treatment plant and freshwater marine studies. Cross-contig spectra from PHACCS showed 129 genotypes common to all time points with high evenness (0.94) and low Shannon-Wiener diversity of 4.5. GAAS (Angly *et al.,* 2009) was also used to estimate the viral genome complexity. GAAS, which estimates the average size of genomes in assemblages, was unable to find enough similarities when compared against the NCBI reference viral genomes with a percent similarity above 50% and a length of 20% of the viral read. In order to estimate the average genome size, viral reads were recruited against the NCBI viral genomes using gsMapper at 90%, 80% and 70% identity across a 50-bp overlap. A weighted average was calculated based on the number of reads aligning to a reference genome and the size of the genome relative to other reference genomes also containing matches.

Rarefaction curves were used to assess species richness for the Illumina data sets. An in-house script mirroring the methodology of Metavir (Roux *et al.,* 2011) was applied to each of the metaviromes. In order to balance the computational demands and time of sampling all reads, a subsampling of 250 000 reads from each data set was selected. The rarefaction analysis showed curves approaching a plateau in the Illumina sample (Supplementary Figure S2), suggesting a more complete sampling of the viral sequence space in the Illumina data and support low genotype richness. In support of this analysis, electronmicrographs of each step gradient fraction showed a correlation between the morphological diversity seen by transmission electron microscope analysis and the estimated richness (data not shown).

Network analysis

Owing to the temporal nature of the samples, it was hypothesized that a network-based approach would reveal patterns of viral assemblage composition and dynamics. To this end, an all-versus-all BLAST was performed on all contigs (>100 bp) from the viral metagenomes. Roughly, 70% of all contigs (35 215 out of 50 735) were found to have a significant match of 75% identity across a 200-bp minimum length to another metagenome contig. To estimate the number of distinct viral groups within the large BLAST-based viral network, the Louvain method was applied. This method has been used in modeling disease outbreaks and identifying and containing computer viruses within social (Lee and Cunningham, 2013) and peer-to-peer networks (Ruehrup *et al.,* 2013). A total of 947 viral groups (that is, an individual cluster of sequences defined by their division between other clusters based on maximizing their modularity, where stronger and more frequent connections within a cluster exist than between members of differing clusters) containing ≥ 2 contig members were detected (for the purposes of this analysis, remaining singleton

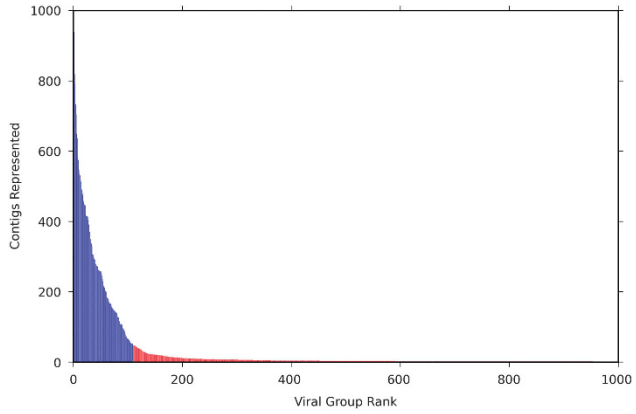

Figure 1 Viral group membership rank abundance curve. Viral groups are ranked according to the number of contigs present within the group. Those with >50 contigs members are included in the NL10 network are blue, with those <50, are red. For brevity, singletons (those without a match during the BLAST analysis or those not included in a viral group by the partitioning algorithm) are excluded.

contigs (15 520) were not considered). The viral group membership-based rank abundance curve is shown in Figure 1. The most dominant population for the viral networks represents 3.2% of the contigs in the network and 0.2% of its reads (a read-based rank abundance is shown in Supplementary Figure S3). This agrees well with the numbers estimated by PHACCS, considering each viral group can represent multiple species simultaneously (discussed below). The network shows a large number of low population groups, with 726 viral groups in the network containing <10 members (comprising only 4.8% of all contigs and <0.01% of reads) and 113 viral groups between 10 and 49 members (4.5% of all contigs), suggesting insufficient sequencing. For the purposes of this study, a viral network was defined has having ≥50 contig members. Even with a minimum membership of ≥50 contigs, >60% of the contigs and 94% of unique reads-bp are represented in the network (Supplementary Figures S4 and S5), strongly suggesting the largest populations represent the dominant viral assemblage's members. For this work and subsequent visualizations, we defined the major viral populations as those groups within the viral assemblages containing ≥50 contigs. Under this threshold, 110 viral groups comprise the Nymph Lake viral assemblage (Figure 2).

Network visualization

Figure 2 visualizes the NL10 viral network and provides a non-metric view of how viral groups are organized. Points represent contigs and connecting lines (edges) are colored by their source contig. Clusters of sequences represent viral groups; each viral group is uniquely, but randomly colored (110 different color variants) to better distinguish groups. Edge length between contigs is inversely proportional to their weight (defined as the -log(e-value)) with strongly connected contigs closer to each other. Except where viral groups are connected (by edges between their members), viral group positions are random and have no BLAST-level relationship to surrounding groups. Long edges connecting distant groups (charged points) result from balancing edge weights (springs) and minimizing interactions between other groups, which can result in underestimation of the true relationship strength between distally connected viral groups.

Owing to the network-based nature of this approach, topographical properties can be calculated to more precisely define the network and to describe the biological organization within and between viral groups. As a whole, members within each viral group are more strongly related to each other than they are to members of other viral groups within the network. A network metric measuring this property is known as *modularity*, which measures the separation strength of individual groups within the larger community (Newman and Girvan, 2004). Networks with higher modularity values have more densely connected vertices and are separated from other dense groups through fewer, sparser connections. Generally, in real-world applications, network modularites >0.3 are considered structured and rarely exceed 0.7 (Newman, 2006). The modularity index was exceedingly high (0.96) for the NL10 viral network. A second metric, known as the clustering coefficient, measures how well connected nodes are with their neighbors. The clustering coefficient was also very high (0.764) for the viral network. These two values together provide overwhelmingly that well-supported and highly connected groups defining the viral assemblages. Despite the densely connected viral groups, there were only 17 connected components in the network, suggesting that few viral groups are entirely isolated from each other.

Sampling effect on network topology

In order to better understand the minimum sampling required of the viral assemblages to define the network with high confidence, network analyses were performed by sequentially adding metagenomes and their network topologies evaluated with the same criteria as above. Using the clustering coefficient, modularity and the number of populations as a measure of 'completeness' compared with the final network, five metagenomes were required to achieve sufficient sampling (Table 2).

Network viral groups: what they represent

To evaluate how well the network clusters (populations predicted known virus families) all viruses in NCBI's database were 'seeded' in a parallel network analysis, including 67 archaeal virus genomes, both as full-length genome sequences and as fragmented

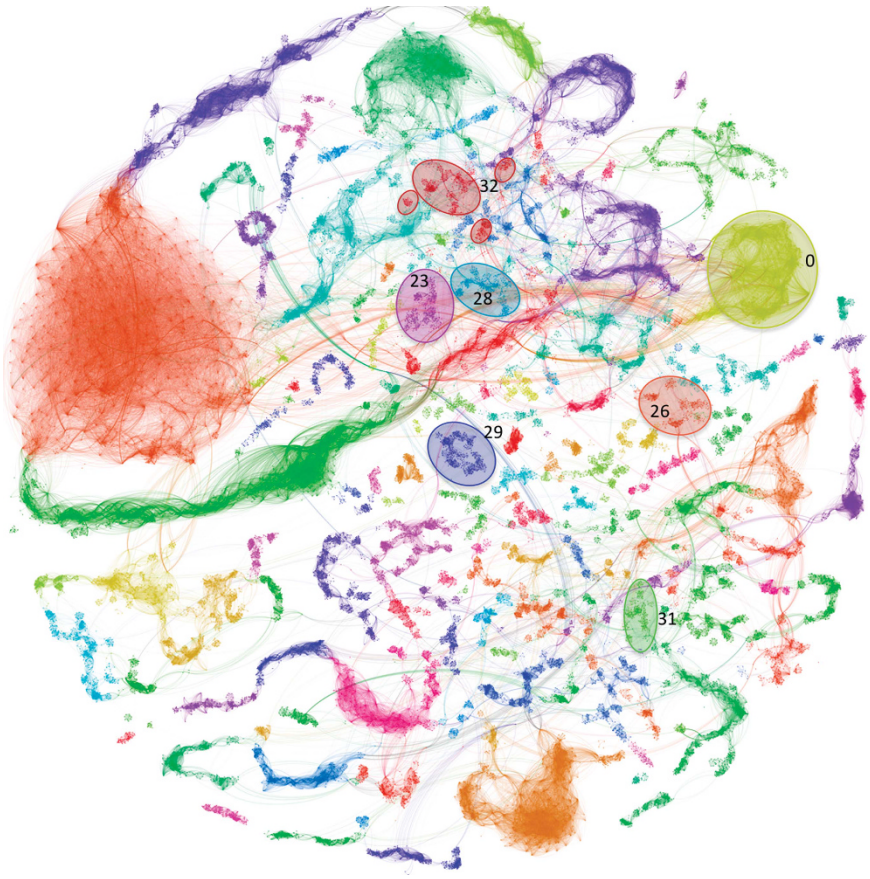

Figure 2 Metavirome network. A visualization of the all-verses-all BLAST network. Viral groups are assigned a unique color and were visualized in Gephi. Separation of the viral groups was accomplished using Gephi's Force Atlas 2 plugin, a force-directed layout algorithm. Ovals highlight viral groups whose members matched against seeded viruses in a parallel network analysis. Despite the large number of seeded viruses (1812), only 26 viruses were assigned to viral groups. All 26 viruses were crenarchaeal viruses. Edge connections are lines connecting members within a viral group and between members of differing viral groups. Care should be taken to not associate distance between viral groups as indicative of sequence similarity. To enhance clarity and distinctions between viral groups, only those containing 50 contigs or more are included in the figure. Viral groups numbering: group 0 = SIRV-1,2; group 23 = ASV-1, SSV-1-2, 4-9; group 26 = ATV; group 28 = AFV-1; group 29 = STIV-1,2; group 31 = AFV2-3, 6-9, SIFV; group 32 = STST-1,2 and ARSV-1, SRV.

genomes (Supplementary Table S2). These archaeal viruses comprise members originally isolated from high temperature environments (37 of the 67 viruses) and viruses not expected to be present in hot spring environments (30 mostly halophile viruses from mesophilic environments), such as those infecting members of the Euryarchaeota (for a review of archaeal viruses, see Prangishvili, 2013; Dellas *et al.,* 2014). The vast majority of known viral genomes did not map to any of the network clusters (1786/1812). As expected, this included all viruses infecting Bacteria and Eukarya or archaeal haloviruses (30 viruses) not expected to be present in previously in acidic, thermophilic environments. Only 26 of the 1812 total viruses, all of which were exclusively archaeal viruses commonly found in thermophilic environments, representing 6 families were found associated with 7 network-defined viral assemblages, representing 6.6% of all contigs (Figure 2). Of these 26 viruses, all of the *Fuselloviridae* (9/9) and nearly all *Lipothrixviridae* (7/9) were found associated with single viral groups (AFV-1 was found in its own group). Of the four characterized *Rudiviridae*, SIRV-1 and SIRV-2 (Prangishvili *et al.,* 1999) were found associated together within their own distinct groups (ARV-1 and SRV are described below). Spherical viruses STIV-1 and STIV-2 (from the proposed 'Turrividae'; Rice *et al.,* 2004; Happonen *et al.,* 2010) were also found tightly centered within a single group as well. The largest viral group in the seeded network was the SIRV-1 and 2 group, comprising 3.2% of the network contigs (0.15% of reads) and 1.9% of all contigs. The only viral group containing members from differing families involved the *Rudiviridae* viruses ARV-1 and SRV, and the unclassified fusiform viruses STSV-1 and STSV-2 (Xiang *et al.,* 2005; Erdmann *et al.,* 2013). A number of other archaeal viruses were found exclusively in single clusters, including ATV; 1.4% network contigs (Prangishvili *et al.,* 2006b) and AFV-1 (1.3% network contigs; Bettstetter *et al.,* 2003). A number of thermophilic viruses were not detected. These include *Acidianus* bottle-shaped virus (ABV), *Aeropyrum pernix* bacilliform virus 1 (APBV-1), coil-shaped virus (ACV) and spindle-shaped virus 1 (APSV-1), *Thermoproteus*

tenax spherical virus 1 (TTSV-1), *Pyrococcus abyssi* virus 1 (PAV-1), *Hyperthermophilic archaeal* virus 1 (HAV-1), *Thermococcus prieurii* virus 1 (TPV-1) and *Pyrobaculum* spherical virus (PSV). Further supporting these viruses' absence from the network, all metagenome reads were aligned against each of the viral references, finding only PSV with reads aligning (3) at any significant (e-value cutoff of 10^{-5}) level. The thermophilic archaeal viruses not associated with network clusters are more commonly associated with more neutral thermophilic environments, providing a possible explanation for their absence in this analysis.

A parallel analysis using fragmented versions of the viral reference sequences found 1194 total matches against the viral groups, with 95% (1133-/1194) of these fragments associated with the same viral groups as their full-length reference counterparts (Supplementary Table S5). The remaining 61 fragments were found associated with well-connected, adjacent viral groups. This suggests these fragments to be more distally related to the 'core' genome of the viral group (detailed below). Overall, the association of the 26 seeded known archaeal viruses isolated from thermophilic acidic environments defined network cluster supports the conclusion the network clusters defined viral assemblages represents a taxonomic unit at or below the family level.

Network viral groups—synthetic networks

To establish the biological significance of the viral groups in the network, two artificially generated viral metagenomes were subjected to the same network analysis. In the first metagenome, randomized versions of the contigs from the original viral assemblages were created. Approximately the same GC% and length, these contigs did not generate sufficient overlap during the BLAST to form a viral group.

As one of the fundamental assumptions in the network analysis is that viral groups correspond to a sub-family to species-level organization, an artificial metagenome using family members of the *Podoviridae* and three archaeal viruses (ATV, STIV and ACV) was designed to test this hypothesis. Nine synthetic metagenomes (each approximately 175 000 reads) was generated using Grinder and 88 *Podoviridae* and archaeal viruses as read-references, and assembled using gsAssembler under the same conditions as the original assemblies. The 19 968 contigs were network analyzed as the NL10 viral network. Nearly 95% of the contigs (18 910) had a BLAST alignment (1058 singletons) and organized into 64 viral groups ranging in size between 87 and 813 contigs (Supplementary Figure S6). Eighty-seven of the 88 (98.9%) were identified in the network analysis, with 52 of the reference sequences (59%) associated with a single viral group. During this analysis it was noted that (a) although reference sequences were associated with single groups, multiple references would share that group and b) the remaining ~40% of reference sequences were spread amongst multiple viral groups, seemingly contrary to the NL10 analysis. For example, Enterobacteria phage N4 and IME11 are each associated with the same two viral groups, whereas Pseudomonas phage LUZ24, phiIBB-PAA2 and vB PaeP p2-10 Or1 are all associated with the same three viral groups. This suggests that closely related viruses associated with multiple viral groups were maintaining their association. It was also noted that the Enterobacteria phage were both members of the N4-like genus, whereas the Pseudomonas phage were all members of the Luz24-like genus. When examined at the sub-family level, only 5 of the 64 viral groups contained members from multiple references (Supplementary Figure S7), supporting the initial hypothesis that the network analysis grouped viruses corresponding at the sub-family level. The size of the viral reference sequence also corresponded to the number of viral groups. The largest *Podoviridae* genome, Cellulophaga phage phi4:1 was associated with five viral groups and is approximately 146-kb. Mapping the 1076 contigs from these five viral groups against their original sequence revealed 100% of the contigs matched at 99% pairwise identity, although several regions across the genome were of low coverage (~1). Reassembly within these five viral groups resulted in 314 contigs and an average contig reduction (within each group) of 70%. Impressively, the largest contig sizes within each group increased from an average of 990 bp to 6.65 kb. Further assembly of the contigs between viral groups revealed a single, 146-kb genome spanning 99% of the original sequence at 99.4% identity. Reassembly between all contigs of the five viral groups generated an identical sequence. Furthermore, these five viral groups represent two connected components, where contigs can be connected to any other contig within the component through in-between connections (Supplementary Figure S8). These two components correspond to the two sections of the phi4:1 genome separated by a low coverage region. On average, reassembly of the *Podoviridae* network viral groups reduced the total contigs by 80% and increased the maximum contig size by 143%.

The three archaeal viruses seeded alongside the *Podoviridae* were each associated with single viral groups in the network analysis and shared no other contigs from other reference sequences. The number of contigs associated with each virus ranged from 118 (ACV) to 412 (STIV) and roughly corresponds to their relative read-based abundance from Grinder. As above, reassembly of the viral groups resulted in a dramatic reduction in contigs (82%) and increased largest contig size (148%). Alignments of the contigs against their references revealed complete coverage of the genomes with all contigs aligning. As with the *Podoviridae*, each alignment revealed low coverage regions likely preventing full-length assemblies.

Comparing the NL10 viral groups against geochemically similar sites

To assess if these similar geochemical environments (acidic, high temperature) would yield differing community compositions, contigs from another hot spring in YNP, CH041, were compared with contigs of the viral groups present in the NL10 viral network. Of the 12 922 contigs in the CH041 viral assemblages, 4259 (32.9%) had strong matches (e-value 10^{-30}, HSP length 50 bp, HSP identity 75%) to 100 of the 110 viral groups (92%). It is clear that many of the viral groups in NL10 are shared in other environments of similar geochemistry. Although there is a small positive correlation ($R^2 = 0.42$) between the size of the viral groups size and the number of CH041 matches, a number of outliers exist.

Temporal stability of viral groups

We have examined the stability of the viral populations using the viral groups within the network as a proxy for presence (Figure 3). Unlike traditional measures of diversity and rank abundance plots, this network-based approach can temporally discriminate within and between viral groups. Twenty-seven of the 29 largest viral groups (based on contig membership) contain members (contigs) from each sampling time point (Figure 3). Typically, around 10% of the members of a given viral group were detected at any time point, though varying from near zero to 50%. We hypothesize lower sequencing depth is responsible for those time points where near zero members were detected. A global analysis of all viral groups in the network revealed that 76 of the 110 viral groups (69%) were present at every sampling point. This increases to 94 viral groups if only one time point is excluded (86%) and 109 viral groups (>99%) consist of members spanning at least half of the sampling points. Overall, these results indicate that each viral group persists in the NL10 environment and the majority of viruses are present at similar levels during a 2-year sampling period of the 454 data sets. It also suggests that at any given sampling point, the sequencing depth was insufficient to detect a majority of the viral contig members. However, by analyzing a time series, the insufficient sequencing depth is overcome through sequence accumulation and contig coalescence of the viral groups.

In order to evaluate if sufficient sequencing depth at a single time point could re-create the network and provide long-term temporal dynamics, the network analysis was repeated using Illumina sequencing technology out of the same environment, 3 years later. Improvements in virus isolation and nucleic acid extraction and amplification were also incorporated. When applied to the Illumina-based contigs, 75.3% (89 088 of 118 259) were retained within the network after filtering out viral groups containing <50 members. This resulted in 184 well-defined viral groups and represents approximately 76% of all Illumina contigs.

Seeding of the Illumina-based network with known viruses as described above revealed 26 archaeal viruses associated within seven viral groups. Twenty-five of these 26 were also found in the 454 based, time course network. As expected, all 26 viruses infected members of the Crenarchaeota, as seen in the network analysis out of the same spring. The largest group (5042 members) represents 4.3% of all contigs (5.7% of the network) and includes all of the recognized *Fuselloviridae*. Not included in this viral group is *Aeropyrum* pernix spindle-shaped virus (APSV-1), not found previously. The seventh largest group (1455 members; 1.2%) matches to the proposed *Turriviridae*, STIV-1 and STIV-2. One recently described virus, initially found through *in silico* identification of proviral sequences (Mochizuki *et al.,* 2011), is APSV-1. Although *A. pernix* has not been seen in this environment, its presence (499 members; 0.4%) indicates that an *A. pernix*-like organism may be a low abundance member of the cellular population or that a potential new host for the APSV-1-like exists in the springs. All the *Lipothrixviridae* sequences were found within a single group, including *Acidianus'* AFV-2-3, 6-9 and SIFV. The two *Rudiviridae* viruses found previously were also found within a single group, SIRV-1 and 2 (777 members, 0.66%). It is extraordinary that alignments using the major capsid proteins of these linear double-stranded DNA viruses group AFV-3, 6-9 and SIFV into a distinct phylogenetic group (Prangishvili and Krupovic, 2012), which is in excellent agreement with the separation in the viral network.

Discussion

The overall objective of this study was to determine the viral assemblage structure and stability in a Yellowstone hot spring by applying a network approach to a time series of viral metagenomic data. We found that the NL10 hot springs contained 110 well-supported viral groups that were persistent over a 5-year time period (Figure 3). The known archaeal viruses represent only 6.3% of the viral groups. Only 7 out of 110 viral groups had members matching to 26 known viruses from acidic hot spring environments. This suggests that the remaining 103 viral clusters represent novel archaeal viruses and provides a rationale for a directed search for these unknown viruses.

A network-based approach was designed and applied to nine viral metagenomic data sets created using 454 based sequencing technology over a 2-year sampling period and one viral metagenomic data set created using Illumina sequencing technology 3 years later. The network approach provides an additional tool for viral assemblage analysis by using BLAST-level relationships and applying a partitioning algorithm to organize viral sequence space. BLAST HSPs connect contigs and the Louvain

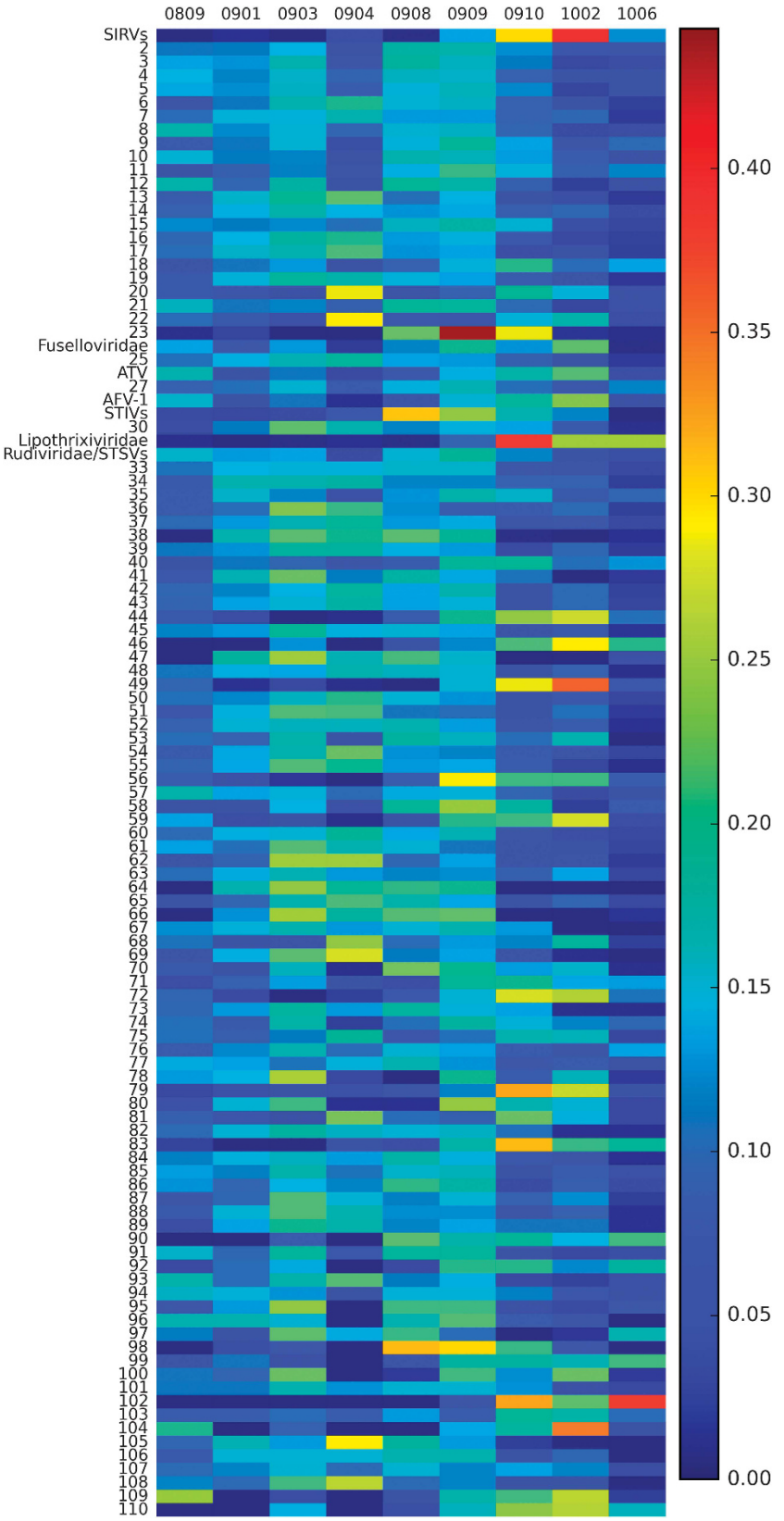

Figure 3 Temporal dynamics of NL10 viral groups. Columns represent time points and rows, viral groups. The percentage of contigs contributed to the viral group from each time point is indicated by the color bar (shown to the right). Viral groups are ordered by their overall rank (number of contigs they contain), with identification of viruses associated with that viral group indicated. All 110 viral groups in the network are included. A read-based version of this figure is included in Supplementary Material.

method partitions contigs into clusters of high relatedness. Both significance and number of relationships between contigs influence the probability a contig will be placed into one viral cluster (referred to as a viral group) over another. This is due to the fact that the Louvain method seeks to maximize modularity by maximizing the number of connections within a viral group and minimizing the connections between groups. These relationships can represent shared gene content (frequently the case in horizontal gene transfer between viruses), overlapping regions on the ends of contigs that were unable to be assembled or highly related fragments of a viral genome sequenced at multiple time points. In addition, sequence data can be annotated with time series information in order to track the contribution of specific sampling points to the viral group. Read recruitment from each of the metagenomes onto the viral groups show a correlation between the number of members in the group and the number of reads aligning to the group. As the length of a contig generally corresponds to the number of reads assembling to it, larger contigs that overlap other large contigs likely have a greater number of reads represented in the community and could serve as a proxy for relative abundance.

The validity of the network approach was supported by three lines of evidence. First, seeding the viral network with known archaeal, bacterial and eukaryotic viruses (a total of 1812 viral genomes) resulted in only 26 archaeal viruses previously identified from acidic hot spring environments mapped to viral clusters generated from the NL10 viral metagenomic data sets. Furthermore, fragmentation of these 26 archaeal virus genomes mapped to the same network viral clusters as the non-fragmented genomes. Examination of the network clusters to which the 26 archaeal viruses mapped made biological sense. For example, the highly related archaeal viruses STIV and STIV-2 mapped to the same single network cluster. Similarly, the nine related *Fuselloviridae* viruses all mapped to the same single network cluster. Second, an identical network analysis was performed on a synthetic viral metagenomic data set derived from 85 *Podoviridae* genomes and 3 archaeal viruses having properties similar to the NL10 viral metagenome data sets. Analysis of the resulting viral assemblies also made biological sense. Closely related *Podoviridae* members formed a single network cluster, whereas distantly related *Podoviridae* members formed their own individual network clusters. Finally, network analysis of synthetic metagenomic data sets of randomized sequences, but having the overall properties of the NL10 viral metagenomic data sets, failed to generate any network clusters.

We believe the individual network viral groups represent the sequence cloud of highly related viruses, at least at the sub-family taxonomic level. Of the subset of viral groups to which known archaeal viruses are members, most match to virus families (that is, *Lipothrixviridae*, *Fuselloviridae,* 'Turriviridae') or single viruses (that is, AFV-1, ATV). It is tempting to speculate that those groups with few members represent virus types/families with smaller pan viral genome sizes as compared with those with larger contig membership. Although is not possible to eliminate the possibility that contaminating cellular or other non-viral sequences are present in the network clusters, we believe that they are relatively minor. This is because we produced the viral metagenomes from samples greatly enriched in virions and because we were able to extensively filter the viral metagenomic data sets before assembly and network analysis against known cellular genome sequences in the public databases and cellular metagenomes produced from the same environmental samples.

The success of this network approach to create biologically meaningful groupings of viral assemblages is likely due to the inherent imperfect nature of metagenomic data sets. Most assembly programs are dependent on having high quality, accurate, even deep sequence coverage created from nearly homogenous template sources. This is unlikely the case for environmental viral sources that comprise complex mixtures of different virus types, each of which likely exists as heterogeneous mixtures of sequences. The result of traditional assembly of such mixtures is usually only fragments (contigs). Network analysis allows grouping of these fragments based on sequences relationships less stringent than common assembly algorithms. The network analysis described here provides a complementary approach to protein-clustering analysis to organize viral groupings using metagenomic data sets. However, it has the advantage of not being dependent on having large (>10 kb) contigs for the analysis.

The 110 identified viral groups likely represent the upper bound for number of major viral groups present in this hot spring environment. Although somewhat lower than the PHACCS richness estimates, the overall number of populations and their size give only a relative abundance of their represented members and genome completeness, and thus care must be taken to not equate network rank abundance with species abundance. This is complicated by the fact that PHACCS estimates richness in terms of genotypes at 98% identity and the network analysis does not make such assumptions. Taxonomic classification of the cross-assembled temporal metavirome revealed roughly half of ORFs were unable to be classified, following the trend of what other groups have reported (Prangishvili *et al.,* 2006a; Yoshida *et al.,* 2013). Of the classifiable ORFs, most matched to previously characterized archaeal viruses (~97%), providing confidence that our viral metagenomic data sets were accessing the archaeal viral assemblages thought to dominate these types of hot spring environments. ORFs lacking homology suggests an unexplored large community of archaeal viruses that remains to be explored.

The viral groups provide a roadmap to explore these unknown viruses. PCR primers can be designed against members in viral groups lacking identification in order to quantify uncharacterized viruses in environmental samples and to aid in their identification and isolation.

Although viral groups can be used to identify viral taxonomic groups, incorporation of time series information expands the analytical power of the network analysis. The heat map in Figure 3 shows the contribution of contigs from each time point for each viral group in the network. Nearly 70% of viral groups contain contigs from all sampling points, suggesting that many of the viruses present in the hot spring are persistent through time. This result is the same when a read-based version of the viral groups is examined (Supplementary Figure S9). If viral groups represent virus species or sub-families, then this organizational scheme can be combined with a read-based proxy for abundance to estimate persistence and relative stability of viruses within the environment. Read-based abundance methods suffer the drawback of being unable to represent an entire genome, even when combined with clustered sequence data associated with a particular viral species. Abundances of reads corresponding to one region of a viral genome may differ from another leading to an uneven representation of that viral type. Similarly, reads from multiple segments of a genome may not be known to be from the same virus, inflating the number of virus species. The network's viral groupings provide a pan-genome platform, specific for closely related viruses that avoid these limitations.

The variations in contig and read-based contributions of the viral groups can be influenced by a number of factors, including the amplification methods and sequencing technologies used to generate the metagenomes. As the smaller metagenomes are more likely to contribute fewer contigs and reads to a viral group, their contribution may be underestimated. Although this is true for many of the viral groups, notable exceptions exist, such as viral groups 3–6, 8, 10, 12 and the *Rudiviridae*/STSV group. More reads and contigs are contributed for the second smallest assembly, NL10 0809. The same can be said for NL10 1006, with group 92, 99 and 102 having their largest contributions. These exceptions can result from biases in amplification and sequencing or represent changes in the viral assemblage. We believe this is likely the latter and that methodological biases have a diminished influence on the network analysis.

Analysis of the same site 3 years after the 454 metaviromes using Illumina sequencing addresses the abundance and persistent of viruses in the hot spring. Twenty-five of the 26 known archaeal viruses are present in the same six viral groups between both networks. The composition of the associated members is also the same between the networks, strongly suggesting that the same viruses (or highly related relatives) are maintained in the hot spring on a long-term basis. This is significant because of the length of time passed (3 years) and the radically different methodologies ($FeCl_3$ precipitation, ovation amplification, Illumina sequencing) used to generate the sequence data. Most striking is the overall similarity between the two networks, considering the metagenomes created using 454 based sequencing technology were subjected to simple physical filtration, whereas the Illumina samples were further purified on cesium chloride gradients. As only 19% of the Illumina data assembled, compared with 82% of the 454 data, and 2.3x more contigs were generated (than the 454 based metagenomic data sets combined), it is reasonable to assume the Illumina data are more fragmented. Despite this, over 75% of the contigs were present in the Illumina-based network. With a cutoff of 50 members, 184 viral groups exist. However, we believe the more fragmented data set contributes to a larger number of smaller viral groups that would be merged with larger groups if more overlaps existed. Considering such a small percentage of reads assembled and the vast majority of contigs were included in the network, we believe that as more contigs assemble and are included, the overall number of viral groups will collapse toward 110 viral groups. Continuing efforts in assembling Illumina-based viral metagenomes is an area of active research and will provide powerful supporting evidence toward this hypothesis.

Our network approach is distinct from previous viral network analysis (Emerson *et al.,* 2012). The objective of this work was not to assemble full-length genomes—although it can organize viral sequence space and aid in assembly, as seen in the reassembly of full-length genomes in the *Podoviridae* network. In contrast, our goal was to gain a better understanding of viral assemblages' structures and stabilities. The high stringency assemblies ensure that each contig corresponds to a single viral genotype rather than a consensus sequence. This allows for much greater discriminatory power when analyzing the connections between contigs within and between viral groups underlying the network structure. Depending on the homogeneity of each viral genome's sequence space, the Louvain method can detect those subtle differences lost when composite genome fragments are utilized.

Overall, this work provides a more comprehensive picture of the underlying virus community structure and stability that should be generally applicable to diverse natural environments. For example, mapping contigs generated from a second YNP hot spring (CH041), with similar temperature and pH, revealed a high number of major populations to be present in both environments. Nearly 92% of the viral groups present in NL10 were found in CH041, indicating common viral populations in hot spring environments sharing common geochemical features. It will be interesting to apply a similar network analysis approach to examine more complex viral assemblages from ocean to human microbiome communities.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Susan Brumfield, Matthew Lavin and Jamie Snyder for their critical reading of the manuscript. This research was supported by NSF-DEB-1342876.

References

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P *et al.* (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinform* **6**: 41.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *Plos Biol* **4**: 2121–2131.

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R *et al.* (2009). The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. *PLoS Comput Biol* **5**: e1000593.

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res* **40**: e94.

Bastian M, Heymann S, Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. San Jose, CA, USA.

Bettstetter M, Peng X, Garrett RA, Prangishvili D. (2003). AFV1, a novel virus infecting hyperthermophilic archaea of the genus acidianus. *Virology* **315**: 68–79.

Blank CE, Cady SL, Pace NR. (2002). Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. *Appl Environ Microbiol* **68**: 5123–5135.

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. (2008). Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* **2008**: P10008.

Bolduc B, Shaughnessy DP, Wolf YI, Koonin EV, Roberto FF, Young M. (2012). Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J Virol* **86**: 5562–5573.

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

Dellas N, Snyder JC, Bolduc B, Young MJ. (2014). Archaeal viruses: diversity, replication, and structure. *Ann Rev Virol* **1**: 399–426.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF. (2012). Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl Environ Microbiol* **78**: 6309–6320.

Erdmann S, Chen B, Huang X, Deng L, Liu C, Shah SA *et al.* (2013). A novel single-tailed fusiform Sulfolobus virus STSV2 infecting model Sulfolobus species. *Extremophiles* **18**: 51–60.

Fancello L, Raoult D, Desnues C. (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology* **434**: 162–174.

Fruchterman TMJ, Reingold EM. (1991). Graph drawing by force-directed placement. *Software Practice Exp* **21**: 1129–1164.

Happonen LJ, Redder P, Peng X, Reigstad LJ, Prangishvili D, Butcher SJ. (2010). Familial relationships in hyperthermo- and acidophilic archaeal viruses. *J Virol* **84**: 4747–4754.

Huang X, Madan A. (1999). CAP3: a DNA sequence assembly program. *Genome Res* **9**: 868–877.

Hunter JD. (2007). Matplotlib: a 2D graphics environment. *Comput Sci Engineer* **9**: 0090–0095.

Hurwitz BL, Deng L, Poulos BT, Sullivan MB. (2012). Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol*, 1428–1440.

Hurwitz BL, Sullivan MB. (2013). The Pacific ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8**: e57355.

Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA, Richardson TH *et al.* (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5**: e9773.

Jay ZJ, Rusch DB, Tringe SG, Bailey C, Jennings RM, Inskeep WP. (2013). Predominant acidilobus-like populations from geothermal environments in Yellowstone National Park exhibit similar metabolic potential in different hypoxic microbial communities. *Appl Environ Microbiol* **80**: 294–305.

John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S *et al.* (2010). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Reports* **3**: 195–202.

Kozubal MA, Macur RE, Jay ZJ, Beam JP, Malfatti SA, Tringe SG *et al.* (2012). Microbial iron cycling in acidic geothermal springs of Yellowstone National Park: integrating molecular surveys, geochemical processes, and isolation of novel Fe-active microorganisms. *Front Microbiol* **3**: 109.

Lee C, Cunningham P. (2013). Benchmarking community detection methods on social media data. *arXiv preprint arXiv*: 1302.0739.

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

Lowenstern JB, Christiansen RL, Smith RB, Morgan LA, Heasler H. (2005). *Steam Explosions, Earthquakes, and Volcanic Eruptions – What's in Yellowstone's Future?* U.S. Geological Survey Fact Sheet 2005–3024.

Macur RE, Jay ZJ, Taylor WP, Kozubal MA, Kocar BD, Inskeep WP. (2012). Microbial community structure and sulfur biogeochemistry in mildly-acidic sulfidic geothermal springs in Yellowstone National Park. *Geobiology* **11**: 86–99.

Marcais G, Kingsford C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Matsumoto M, Nishimura T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans Model Comp Simul (TOMACS)* **8**: 3–30.

McDaniel LD, Rosario K, Breitbart M, Paul JH. (2013). Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environ Microbiol* **16**: 570–585.

Mochizuki T, Sako Y, Prangishvili D. (2011). Provirus Induction in Hyperthermophilic Archaea: characterization of Aeropyrum pernix spindle-shaped virus 1 and Aeropyrum pernix ovoid virus 1. *J Bacteriol* **193**: 5412–5419.

Mokili JL, Rohwer F, Dutilh BE. (2012). Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* **2**: 63–77.

Newman M. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci USA* **103**: 8577–8582.

Newman ME, Girvan M. (2004). Finding and evaluating community structure in networks. *Physical Rev E* **69**: 026113.

Niu B, Fu L, Sun S, Li W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11**: 187.

Ojosnegros S, Perales C, Mas A, Domingo E. (2011). Quasispecies as a matter of fact: viruses and beyond. *Virus Res* **162**: 203–215.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2011). Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics* **27**: i94–i101.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Pina M, Bize A, Forterre P, Prangishvili D. (2011). The archeoviruses. *FEMS Microbiol Rev* **35**: 1035–1054.

Prangishvili D, Arnold HP, Gotz D, Ziese U, Holz I, Kristjansson JK *et al.* (1999). A novel virus family, the Rudiviridae: structure, virus-host interactions and genome variability of the Sulfolobus viruses SIRV1 and SIRV2. *Genetics* **152**: 1387–1396.

Prangishvili D, Forterre P, Garrett RA. (2006a). Viruses of the Archaea: A Unifying View. *Nat Rev Microbiol* **4**: 837–848.

Prangishvili D, Vestergaard G, Häring M, Aramayo R, Basta T, Rachel R *et al.* (2006b). Structural and genomic properties of the Hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle. *J Mol Biol* **359**: 1203–1216.

Prangishvili D, Krupovic M. (2012). A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales". *Arch Virol* **157**: 791–795.

Prangishvili D. (2013). The wonderful world of archaeal viruses. *Ann Rev Microbiol* **67**: 565–585.

Reysenbach AL, Wickham GS, Pace NR. (1994). Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **60**: 2113–2119.

Rice G, Stedman K, Snyder J, Wiedenheft B, Willits D, Brumfield S *et al.* (2001). Viruses from extreme thermal environments. *Proc Natl Acad Sci USA* **98**: 13341–13345.

Rice G, Tang L, Stedman K, Roberto F, Spuhler J, Gillitzer E *et al.* (2004). The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc Natl Acad Sci USA* **101**: 7716–7720.

Rosario K, Breitbart M. (2011). Exploring the viral world through metagenomics. *Curr Opin Virol* **1**: 289–297.

Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D *et al.* (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**: 3074–3075.

Ruehrup S, Urbano P, Berger A, D'Alconzo A. (2013). Botnet detection revisited: theory and practice of finding malicious P2P networks via Internet connection graphs. *Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on*, Turin, Italy, 435–440.

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D. (2008). Assembly of viral metagenomes from Yellowstone hot springs. *Appl Environ Microbiol* **74**: 4164–4174.

Solonenko SA, Ignacio-Espinoza JC, Alberti A, Cruaud C, Hallam S, Konstantinidis K *et al.* (2013). Sequencing platform and library preparation choices impact viral metagenomes. *BMC Genomics* **14**: 320.

Tamaki H, Zhang R, Angly FE, Nakamura S, Hong P-Y, Yasunaga T *et al.* (2011). Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol* **14**: 441–452.

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. (2009). Laboratory procedures to generate viral metagenomes. *Nat Protocols* **4**: 470–483.

Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M. Next generation sequence assembly with AMOS.*Curr Prot Bioinform* 2002; **Chapter 11**: Unit 11.8.

Tseng C-H, Chiang P-W, Shiah F-K, Chen Y-L, Liou J-R, Hsu T-C *et al.* (2013). Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J* **7**:1–13.

Wilhelm SW, Suttle CA. (1999). Viruses and nutrient cycles in the sea viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**: 781–788.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI *et al.* (2008). The Sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One* **3**: e1456.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S *et al.* (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* **6**: 427–439.

Xiang X, Chen L, Huang X, Luo Y, She Q, Huang L. (2005). Sulfolobus tengchongensis spindle-shaped virus STSV1: virus-host interactions and genomic features. *J Virol* **79**: 8677–8686.

Yoshida M, Takaki Y, Eitoku M, Nunoura T, Takai K. (2013). Metagenomic analysis of viral communities in (Hado)Pelagic sediments. *PLoS One* **8**: e57271.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)