

ANALYSIS OF ARCHAEOAL VIRUSES AND CHARACTERIZATION OF THEIR
COMMUNITIES IN YELLOWSTONE NATIONAL PARK

by

Benjamin Ian Bolduc

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, Montana

May 2014

©COPYRIGHT

by

Benjamin Ian Bolduc

2014

All Rights Reserved

ACKNOWLEDGEMENTS

I would like to thank the members of my committee: Dr. Mark Young, Dr. Brian Bothner, Dr. Valerie Copié, Dr. Matt Lavin and Dr. Sara Rushing. Their helpful advice and guidance throughout the years has helped me see new scientific perspectives and explore the often-multisided (and frustrating) results at the frontiers of environmental virology. I would like to specific thank my advisor, Dr. Mark Young, whose breadth of knowledge in my field is unparalleled – serving as an invaluable soundboard for all my virology questions.

I would like also like to thank Dr. Frank Roberto at Newmont Mining Corp., for sequencing my 96-well plates, providing guidance when experiments straight up failed, and giving me confidence that I could succeed. I would also like to thank Christine Hendrix, who worked at the Research Permit Office in Yellowstone National Park. Without her, I would never have been able to gain access to the hot springs from which all my research derives.

Finally, I would like to thank my family and friends, especially my aunts and grandparents who cared for me in my early years while my mother served in the military. To Jibryne “Gubby” Karter Jr., whose help during college and mentorship throughout my educational career gave me the opportunity to pursue whatever avenue of study I wanted. Lastly, I would like to thank my great-grandparents, Dominique and Joan Bolduc, who believed I would be “Dr.” long before I believed it. Their unwavering support (since before I can remember) gave me the courage to pursue my dreams, overcoming the greatest challenge of all: myself.

TABLE OF CONTENTS

1. INTRODUCTION	1
Archaea	1
<i>Crenarchaeota</i>	3
<i>Euryarchaeota</i>	4
<i>Thaumarchaeota</i>	4
<i>Nanoarchaeota</i>	5
<i>Korarchaeota</i>	6
Viruses and their Diversity	6
Archaeal Viruses	9
Spindle-Shaped Viruses	10
<i>Fuselloviridae</i>	11
<i>Bicaudaviridae</i>	12
Salterprovirus	14
Unclassified	15
Linear Viruses	16
<i>Lipothrixviridae</i>	17
<i>Rudiviridae</i>	18
Spherical Viruses	20
<i>Globulaviridae</i>	20
<i>Turriviridae</i>	21
<i>Halosphaerovirus</i>	22
Pleomorphic Viruses	23
Head-Tail Viruses	24
<i>Myoviridae/Siphoviridae</i>	25
<i>Podoviridae</i>	26
Other Unique Morphotypes	26
<i>Spiraviridae</i>	29
Viral Metagenomics	29
Sample Preparation	32
Sequencing	34
Bioinformatic Analysis	39
Research Directions	45
Literature Cited	52
2. IDENTIFICATION OF NOVEL POSITIVE-STRAND RNA VIRUSES BY METAGENOMIC ANALYSIS OF ARCHAEA-DOMINATED YELLOWSTONE HOT SPRINGS	78
Contribution of Authors and Co-Authors	78
Manuscript Information Page	79

TABLE OF CONTENTS - CONTINUED

Abstract	80
Introduction	81
Materials and Methods	84
YNP Hot Spring Sample Sites	84
Initial Screening of Viral Enriched Samples for RNA Genomes by 32P-labeling Experiments (RT-dependent Signal from RNA Viral Fraction)	84
Isolation of Cellular Fraction for Community Sequence Analysis	85
Isolation of Viral-Enriched Nucleic Acids	86
Isolation, Preparation of RNA-enriched Viral Fraction and Sequencing	86
Isolation, Preparation and Amplification of DNA-enriched Viral Fraction	87
Sequence Assembly and Analysis	87
Protein Structure Analysis	88
Phylogenetic Analysis of RdRps	88
Validation of Selected Assembled Contigs from RNA-enriched Viral Metagenomes by RT-PCR and DNA sequencing	89
Analysis of CRISPR in Cellular and Viral Metagenomes	90
Examining Viability of a Eukaryotic Virus in YNP Hot Springs	91
Comparing RNA Metagenomes to Other Environments	91
Results	91
Discussion	107
Acknowledgements	111
Literature Cited	113
 3. VIRAL COMMUNITY COMPOSITION IN YELLOWSTONE ACIDIC HOT SPRINGS ASSESSED BY NETWORK ANALYSIS	120
Contribution of Authors and Co-Authors	120
Manuscript Information Page	121
Abstract	121
Introduction	122
Materials and Methods	123
Sample Sites	123
Sample Collection	124
Processing Reads from Viral Metagenomes	125
Assembly of Viral Metagenomes	126
Analysis of Viral Alpha Diversity and Taxonomic Identification	126
Network Analysis of Viral Assemblies	127
Reassembly of Viral Contigs Based on Network Clusters	128

TABLE OF CONTENTS - CONTINUED

Minimum Sampling to Describe Viral Community Through Network Analysis.....	130
Results.....	130
Read Processing and Assembly	130
Taxonomic Analysis	137
Community Composition.....	137
Network Analysis.....	138
Discussion.....	146
Literature Cited	152
 4. CHARACTERIZATION OF VIRAL COMMUNITIES IN YELLOWSTONE HOT SPRINGS BY DEEP-SEQUENCING.....	 155
Contribution of Authors and Co-Authors	155
Manuscript Information Page	156
Abstract	157
Introduction.....	158
Materials and Methods.....	160
YNP Sample Site	160
Sample Collection.....	160
Nucleic Acid Extraction and Sequencing Preparation.....	161
Read Processing and Assembly	162
Rapid Identification of RdRps in Viral Metagenomes	163
Analysis of Viral Diversity	163
Network Analysis and Re-Assembly of Viral Populations.....	164
Results and Discussion	165
Sample Site, Isolation, and Extraction.....	165
Processing and Assembly of Reads	167
Viral Diversity	169
Network Analysis.....	169
Hallmark RNA Virus Genes	175
Conclusions.....	177
Literature Cited	180
 5. CONCLUDING REMARKS.....	 188
 APPENDICES	 197
APPENDIX A: Supporting Information for Chapter 2.....	198
APPENDIX B: Supporting Information for Chapter 3.....	205
APPENDIX C: Supporting Information for Chapter 4.....	210

TABLE OF CONTENTS - CONTINUED

LITERATURE CITED	218
------------------------	-----

LIST OF TABLES

Table	Page
1.1 List of Known Archaeal Viruses.....	46
2.1 Cellular, DNA viral and RNA viral assembly statistics	93
3.1 YNP Sampling Points and Characteristics.....	124
3.2. Sequencing and Assembly Statistics.....	129
3.3 DNA metavirome re-assemblies and CHAS-BLAST based on network analysis.....	132
3.4. Viral Community Structure Predicted from Assembly of Metagenomic Sequences.....	138
3.5 Network Analysis on Stepwise-Added Metagenomes.....	142
4.1 Sequencing and Assembly Statistics.....	168
4.2 GDD-containing sequences identified as RdRps.....	176

LIST OF FIGURES

Figure	Page
1.1 An overview of the three domains of life, including the recognized phyla of the <i>Archaea</i>	3
1.2 The <i>Fuselloviridae</i>	12
1.3 The <i>Bicaudaviridae</i>	14
1.4 The Salterprovirus, His1	15
1.5 The Unclassified Spindle-Shaped Viruses, PAV and TPV1	16
1.6 The <i>Lipothrixviridae</i>	18
1.7 The <i>Rudiviridae</i>	19
1.8 The <i>Globulaviridae</i>	21
1.9 The Proposed <i>Turriviridae</i> viruses	22
1.10 The Proposed <i>Halosphaerovirus</i>	23
1.11 A Panel of Pleomorphic Viruses.....	24
1.12 The Head-Tailed Viruses of the <i>Myoviridae</i> , <i>Siphoviridae</i> and <i>Podoviridae</i>	25
1.13 Archaeal Viruses with Exotic Morphologies	27
1.14 The <i>Spiraviridae</i> , ACV	29
2.1 Selected examples of screening hot springs for presence of RNA templates in the RNA virus enriched fractions.....	92
2.2 Hierarchical classification of contigs with MG-RAST.....	96
2.3 Detection and single stranded nature of RNA viral sequences within hot springs months after sampling for metagenomic analysis.....	98

LIST OF FIGURES – CONTINUED

Figure	Page
2.4 The RdRp of the putative archaeal viruses	100
2.5 The predicted capsid protein of the putative archaeal virus and its potential autoproteolytic activity	103
2.6 Phylogenetic tree of the RdRps.....	104
2.7 Analysis of cellular CRISPR direct repeat units and number of unique spacers matching archaeal species and RNA viral metagenome contigs.....	107
3.1 Population-based rank abundance curve.....	140
3.2 Metavirome network.....	141
3.3 Community temporal dynamics	146
4.1 Rarefaction curve for cDNA viromes and vDNA viromes.....	170
4.2 DNA metavirome network.....	171
4.3 RNA metagenome network.....	173
4.4 DNA metaviromes fraction contribution	174
4.5 RNA metaviromes fraction contribution	175

ABSTRACT

Viruses infecting the Archaea – the third domain of life – are the least understood of all viruses. Despite only 100 archaeal viruses being described, work on these viruses revealed a remarkable level of morphological and genetic diversity unmatched by their bacterial and eukaryotic counterparts, whose numbers range over 6000. Study of these archaeal viruses could gain insight into fundamental aspects of biology and reveal underlying evolutionary connections spanning the three domains of life, including the origin of life. In addition, we understand very little about their community structures in natural environments. To address these daunting tasks, a viral metagenomics approach was undertaken using next generation sequencing technologies. Despite this, only a fragmented view of the viral communities is possible in natural ecosystems. Therefore, this dissertation sought to apply a network-based approach in combination with viral metagenomics to not only describe natural viral communities, but to find and characterize the first RNA viruses out of acidic, high-temperature hot springs in Yellowstone National Park, USA. These hot springs harbor low complexity cellular communities dominated by several species of hyperthermophilic Archaea. The results of this dissertation show that this approach can identify distinct viral populations and provide insights into the viral community. Furthermore, the viral communities of these hot springs are relatively stable over the course of the sampling time period. In addition, a number of viral clusters – each representing a viral family at the taxonomic level – are likely previously uncharacterized DNA viruses infecting archaeal hosts. This approach demonstrates the utility of combining viral community sequencing with a network analysis to understand viral community structures in natural ecosystems. Additional analysis of these viral metagenomes led to the identification of novel RNA viral genome segments. Since no RNA virus infecting *Archaea* is known to exist, this dissertation also sought to more fully characterize these sequences. Genes for RNA-dependent RNA polymerases, a hallmark of positive-strand RNA viruses were identified, suggesting the existence of novel positive-strand RNA viruses likely replicating in hyperthermophilic archaeal hosts and are highly divergent from RNA viruses infecting eukaryotes and are even more distant from known bacterial RNA viruses.

INTRODUCTION

Archaea

The Archaea represent the third domain of life. These organisms were originally classified alongside bacteria based on cellular organization (i.e. lacking a nucleus or organelles) and not until the use of rRNA sequence led to the re-classification into the three kingdoms (265), separating bacteria into *Eubacteria* and *Archaeobacteria* and subsequently the three domains of *Archaea*, *Bacteria*, and *Eukarya* (266). Since then, a great deal of research has focused on multiple aspects of the *Archaea*, including the biochemistry, molecular biology, physiology, microbiology, and evolution. This wealth of information has revealed a number of broad similarities compared to the other domains. In terms of their cellular architecture, lacking a nucleus or membrane-bound organelles, *Archaea* resemble bacteria. The two domains also share many metabolic pathways (131) and similar genome organization (264). The *Archaea* also share a number of similarities with the *Eukarya*, including their DNA replication (28), transcription (113, 256), and translation machineries (5, 86). Not everything about *Archaea* is shared between the other domains; their ether-linked and isoprenoid-containing membranes (4, 36, 164) and their diverse energy-utilization pathways (245), particularly methanogenesis, are all unique to Archaea. Numerous reviews have been published on this subject (86, 87) and will not be reviewed here.

Since their classification, the *Archaea* were believed to be exclusively extremophiles – organisms surviving in the most extreme environments. At the time all

known *Archaea* fell into one of two groups; 1) the methanogens, consisting of extreme halophiles, sulfate-reducers and thermophiles, and 2) the “thermoacidophiles,” who were entirely thermophilic (263). These groups were named the *Euryarchaeota* and *Crenarchaeota*, respectively (266). The belief that *Archaea* were solely found in extreme conditions changed when a number of studies revealed *Archaea* to be widespread in the world’s oceans (65, 90). *Archaea* are now found to contribute up to 20% of the total biomass on Earth (67) and represent roughly 30% of the prokaryotic diversity in the ocean’s mesopelagic zone, and possibly the ocean’s single most abundant cell type (123). The *Archaea* have also been implicated in global biogeochemical cycling and may play a significant role in carbon and nitrogen cycling (88, 116). They are far more widespread than simply the ocean environments, inhabiting freshwater lakes, desert soils, Antarctic lakes, sewage, rumen, hydrothermal vents, hot springs, and salterns (reviewed extensively in (48, 66)). It is safe to say that members of the *Archaea* have thus far defined the boundaries for life on Earth, as they hold the record for both the highest growth temperatures and the highest salt concentrations (48).

As more archaeal genomes were isolated, the same rRNA sequencing leading Woese and colleagues to suggest the three-domain scheme also revealed several additional archaeal phyla. Recent discoveries have led to a number of proposed phyla, such as the *Korarchaeota*, isolated from a hot springs in Yellowstone National Park, USA (21), the ultra-small sized *Nanoarchaeota* (111), and the thermophilic ammonia-oxidizing *Thaumarchaeota* (46). Each of these phyla is clearly monophyletic within the

Archaea in terms of their lipid composition and rRNA sequencing. A figure presenting the recognized and proposed phyla is shown in Figure 1.1

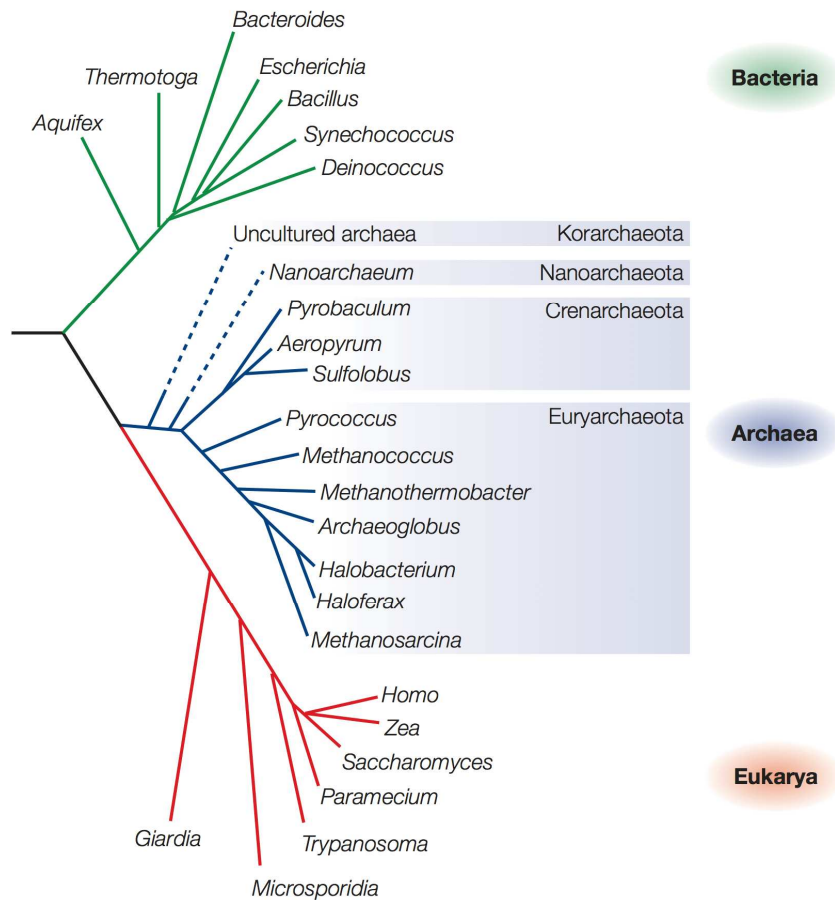


Figure 1.1. An overview of the three domains of life, including the recognized phyla of the *Archaea*. Figure taken from (5).

Crenarchaeota

The *Crenarchaeota* (from the Greek ‘*crenos*’ of origin) are a largely hyperthermophilic phylum historically known for possessing most of the record-breaking species (34, 230), including an archaeum isolated from a hydrothermal vent which actively reproduces at 121°C (124). They are found in a wide variety of hyperthermophilic environments, including hot springs, acid mine drainages, marine

hydrothermal vents and even near volcanic areas (48). They are of particular evolutionary interest due to their deep-rooted location in rDNA phylogenetic trees (86) and the possibility that the last common archaeal ancestor was also a thermophile (16).

Euryarchaeota

Alongside the *Crenarchaeota*, the *Euryarchaeota* were one of the two original phyla, then including the known methanogens and extreme halophiles (266). They are the most diverse of the archaeal phyla, including all known methanogens, many of the halophiles, some thermophilic and psychrophilic species, as well as mesophilic ones (5, 48). They are most noted for some members' methanogenic capabilities by reducing a variety of carbon compounds to methane (47) and their exceedingly high salt tolerances, often approaching saturation (37). The *Euryarchaeota* have been found in an exceedingly diverse array of environments, including the human body, varieties of soil types, lakes, hot springs, the deep subsurface and polar regions, salterns, and even the Dead Sea (48).

Thaumarchaeota

Once classified as mesophilic *Crenarchaea* (65, 90), the phylum *Thaumarchaeota* (from the Greek '*thaumas*' for wonder) was suggested when a more robust ribosomal protein analysis was performed on *Cenarchaeum symbiosum* (a marine archaeon inhabiting the tissue of a temperate water sponge (193)) and a well-represented selection of archaeal and bacterial sequences. This analysis showed mesophilic *Crenarchaeota* likely diverged from the *Archaea* before the speciation of *Euryarchaeota* and hyperthermophilic *Crenarchaeota* (46). Additional work by this group showed the

presence of a number of euryarchaeal-specific proteins and an absence of proteins specific to Crenarchaeota, further supporting the previous conclusions. This re-classification included all the currently known autotrophic ammonia-oxidizing *Archaea* (63, 88, 130, 165). This process is critical in nitrification, as the first and rate-limiting step, and is the only biological process that can convert reduced nitrogen to a biologically-accessible oxidized form (97), making members of the *Thaumarchaeota* major players in global nitrogen cycling (88).

Nanoarchaeota

The *Nanoarchaeota* (from the Greek ‘nano’ for dwarf) are a phylum of ultra small, obligate symbionts, originally discovered in a hydrothermal vent (111). Since their discovery, similar sequences have been found in other hydrothermal areas and the hot springs of YNP (109), with another strain recently isolated from another hot spring in YNP (182). Unable to survive without direct contact with its hosts, the cultured isolates entirely lack genes for the biosynthesis of amino acids, nucleotides, cofactors and lipids (254). The more recently cultured *Nanoarchaeota* does appear to possess the enzymes responsible for the oxidative TCA cycle and glycolysis, unlike the initial isolate (182). The *Nanoarchaeota* do appear to contain all the necessary components for replication, transcription, translation, and completion of the cell cycle with an unusually high percentage of its genome coding for proteins and stable RNA molecules (95%) (254). Its position as a phylum, and not as a fast-evolving euryarchaeal lineage, is a hotly debated topic in recent years (74, 145, 244). The addition of the second nanoarchaeon genome, whose features clearly contain affinities to the *Euryarchaeota*, strongly supports a deep-

branching clade of the *Archaea* (182). Additional sequences from the *Nanoarchaeota* will be necessary to resolve this issue.

Korarchaeota

The *Korarchaeota* (from the Greek ‘*koros/kore*’ for young man/woman) are a phylum argued to be one of the most ancient archaeal lineages diverging before the speciation of the *Crenarchaeota* and *Euryarchaeota* (21). Their genomes are a mosaic of crenarchaeal and euryarchaeal sequences, with ribosomal proteins and RNA polymerase subunits representing the former, and DNA replication, DNA repair, and cell division components representing the latter (82). The sequenced members of *Korarchaeota* also lack a number of biosynthetic pathways and are unable to use electron acceptors such as oxygen, nitrate, iron, or sulfur (112). Based on genome analysis, they are believed to use peptides as the principal carbon and energy source, with a strictly anaerobic, heterotrophic lifestyle – which might explain why complex enrichment cultures were able to support growth rather than pure cultures, allowing genomic characterization (82). Their members are believed to be widespread among hyperthermophilic environments, and have been found in both deep-sea hydrothermal vents (236) and terrestrial hot springs (15, 22, 199).

Viruses and their Diversity

Viruses are infectious biological agents infecting cellular life, with at least all cellular organisms studied to have a virus or virus-like element associated with them (246). They are considered the “most abundant biological entities on the planet” (42),

exceeding 10^{30} particles in the oceans and outnumbering *Archaea* and *Bacteria* combined by at least 15-fold (233). This staggering number becomes higher as viral abundance is often greater in soil (52). Viruses make up 5% of the biomass in the oceans, yet comprise 94% of the nucleic acid-containing particles (233). Due to their extraordinary numbers, viruses are responsible for lysing up to 20% of the ocean's prokaryotes each day (234) and are believed to play a major role in global carbon cycling through the viral shunt (259). This moves material (i.e. carbon) from living organisms (prokaryotes in the ocean) into dissolved organic carbon and particulate organic carbon available to other organisms for processes such as photosynthesis. In addition to global carbon cycling, viruses are believed to be major players in the nitrogen cycling (233, 259), underscoring their importance on a global level outside of human-based pathogenicity.

Viruses are an incredibly source of genetic diversity, arguably the largest untapped reservoir on Earth (233). Regardless of sampling location, the number of virus genotypes found is staggering. There were approximately 5000 different genotypes in near-shore marine environments, 1000 in human fecal samples, and 10000 to possibly 1 million in marine sediment samples (40, 41, 43). Overall, there are a predicted 100-million phage species with at least 2.5 billion phage-encoded open reading frames yet to be discovered (203). Drawing from these values, it is obvious the "that majority of viral diversity remains unknown" (42). Viruses are also known to serve as a means of horizontal gene transfer between organisms, potentially affecting microbial diversity and evolution (215).

Unlike all cellular organisms, viruses do not have a universal marker by which they can be classified and taxonomically organized. This is compounded by an incredibly array of diverse morphologies, genetic types, and modes of reproduction. Even if large groups of viruses shared genetically similar elements, many viruses have mutation rates so high that any similarities they once had to their brethren would be lost (71). Early taxonomies used the physical properties of the virion; based on the genome type, capsid symmetry, enveloped vs. non-enveloped capsids, and diameter of the virion (144). This led to the Baltimore classification system, which is based on the viral genome type and mode of replication (18). A more recent classification system, maintained by the ICTV, organizes viruses in a similar fashion to cellular taxonomy, through orders, families, genera and species (129). It is based on sequence comparison and argues for an organizational scheme where virus families likely evolved from a common ancestor.

In recent years a number of studies have revealed strikingly similar structural and assembly principles of virions (96, 197, 200), leading to the question of whether or not viruses formed lineages across different domains of life (19). This possibility offers higher-level organization and classification of viruses rooted in phylogeny and evolution, rather than sequence similarities limited to closely diverging species (i.e. on short evolutionary timescales) or genome-based, offering little information beyond the strandedness and nucleic acid type of a virus. With the discovery of the archaeal virus STIV (detailed below) and cryo-EM reconstruction, a structural similarity spanning all three domains of life was discovered (202). Structural conservation, while all sequence similarities were abolished, created the capability of classifying viruses

considered to share an ancestor and bring order to the viral universe (133). It is believed the viral “nonself” elements (opposed to the “self” used to determine the lineages) are the result of viral adaptation allowing the virus to change host systems or environments.

Archaeal Viruses

There are nearly 6300 prokaryotic viruses described morphologically; but, of these, only about 100 infect *Archaea* (1). They were first described in 1974 (240) in cultures infecting an extreme halophile, before the *Archaea* were classified alongside *Eukarya* and *Bacteria* in the three domains. The finding of another head-tail “phage” was unremarkable, as 96.3% of all prokaryotic viruses display this morphology (1). This changed in the early 1980s when Zillig and colleagues discovered filamentous and spindle-shaped viruses in Icelandic/Japanese hot springs with dsDNA genomes, unseen in bacterial viruses (120). Later expeditions in the mid 1990s found a wide assortment of other filamentous and spindle-shaped viruses from other hyperthermophilic environments (277, 278). Since then a remarkable amount of effort has been undertaken to not only find new archaeal viruses, but to understand their relationship to viruses from the other domains, between other archaeal viruses, and their relationships to their hosts.

Several factors contribute to the dramatic difference in viruses found from bacterial and archaeal domains. The first is the reliance on culturing of infected hosts to establish virus-host systems. Since it is estimated that less than 1% of microorganisms can be cultured (7, 80, 125), the vast majority of “hosts” and by extension their viruses, cannot be cultured. The second, albeit related to the first, is the difficulty of reproducing

many of the extreme conditions required for the *Archaea*. Viruses are known to infect 15 archaeal genera, compared to 179 of bacteria (1), likely illustrating the two challenges mentioned above. Of the archaeal viruses, they nearly exclusively infect hyperthermophiles or extreme halophiles.

Despite the relatively few numbers, archaeal viruses have been found to display an enormous diversity of virion morphologies (reviewed in (75, 187, 189)), including droplets, bottle-shaped, fusiform, bacilliform, linear, spherical, and icosahedral. It is noteworthy that the head-tail morphotype, seen in the earliest archaeal viruses, has only been seen in members of the Euryarchaeota, whereas most of the unusual morphologies occur within the Crenarchaeota (1). This is interesting considering head-tails are a minority of the observed morphologies in the hypersaline environment, and other unusual shapes predominate (224).

The following section organizes archaeal viruses by gross morphology and seeks to describe their genomic and structural properties, their relationship to other archaeal viruses and to viruses outside the domain *Archaea*. A comprehensive table of all known archaeal viruses is included in Table 1.1, at the end of this chapter.

Spindle-Shaped Viruses

Also known as fusiform viruses, these spindle-shaped viruses are common and exclusive to the *Archaea* dominating in environments where the *Archaea* are known to predominate (189, 195, 201, 278). Shaped like lemons, the vast majority of these viruses have circular, dsDNA genomes, and many have been shown to integrate into their host

genomes (117, 257). The only exception is virus His 1 infecting the Euryarchaeota, which has a linear dsDNA genome and does not appear to encode an integrase (25, 187).

Fuselloviridae. There are currently nine members of the *Fuselloviridae* family infecting two aerobic, hyperthermophilic species, *Sulfolobus* and *Acidianus*. The virions average 60 nm wide x 100 nm long, with short tail fibers that likely take part during virus attachment to its host (198, 229, 257).

The type virus of the *Fuselloviridae* is SSV1, first discovered in 1984 (then known as SAV 1) in Wolfram Zillig's lab (149). Its covalently-closed, circular genome is positively supercoiled (161) and 15.4-kb in length, roughly average for a member of the *Fuselloviridae*. Like many of this family, SSV1 can integrate into its host genome's tRNA gene, leaving a functionally intact gene upon integration (159). It does this through a site-specific recombinase, virally encoded, and is a member of the tyrosine recombinase family (257). Viral infection does not cause host lysis nor is the traditional lysogenic state established. Unlike many other archaeal viruses, the transcriptional patterns for SSV1 have been studied, with many of its transcripts constitutively expressed during its "carrier state" (89). This has led to the hypothesis that these viruses exist in this carrier state, balancing viral replication and cell division, releasing minimal amounts of infectious virions to the external environment, providing a means of protection against the harsh environment (190). SSV1 is a UV- and mytomycin C-inducible virus, temporarily slowing cell growth of the infected cultures while producing virions, but does not cause lysis (89, 217). This induction reveals early, late, and UV-inducible transcripts (89).

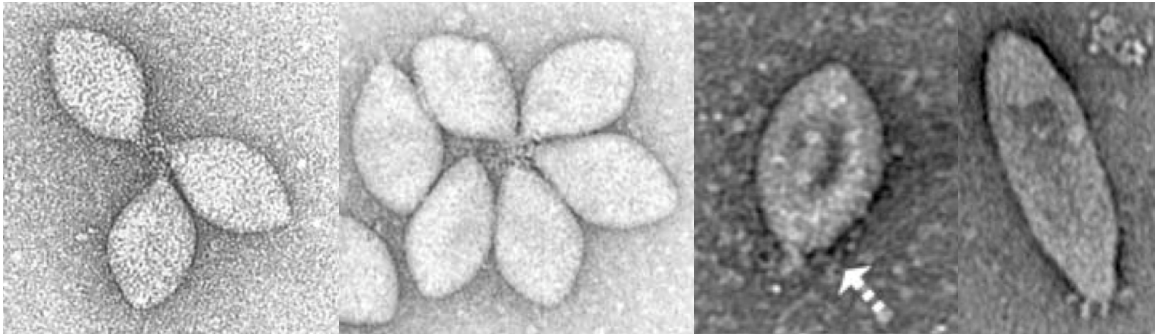


Figure 1.2. The *Fuselloviridae*. From left-to-right, SSV1, SSV2, SSV7 and ASV. Images for SSV1 and SSV2 taken from (229), SSV7 and ASV taken from (198).

Bicaudaviridae. The sole current member of this family is ATV, first discovered in 2006 infecting *Acidianus* (192). The most remarkable aspect of this virus is its extracellular tail development, occurring entirely outside of, and independent to its host. This development results in an 2X increase in volume and slightly diminished surface ratio, occurring at temperatures above 75°C (192). While several viruses do undergo structural changes upon binding to their host cell (20), none occur entirely independent of host. The mechanism for this morphological change is unknown, but analysis of its genome reveal a potential protein modified in the two-tailed versions of the virus that likely contributes to the change (192). It is believed this development is a survival strategy for environments with low cell density. ATV is also one of the few archaeal viruses that can lyse its host.

The 62.7-kb ATV genome is the second largest crenarchaeal virus genome, behind STSV1, another spindle-shaped virus. The genome contains very little similarity to other known archaeal viruses, but does contain several putative ORFs for which functions have been identified. Its genome contains four transposase genes, unique even compared to other archaeal viruses. ATV also encodes an integrase, and maintains a

functional gene upon integration (222). The integrated form is only detected in cells growing at 85°C, not at 75°C, suggesting induction under stressful conditions (192).

Two additional members have been proposed for inclusion into the *Bicaudaviridae*, STSV1 and the induced-provirus APSV1. STSV1 was originally isolated in a high temperature, acidic hot spring in China. Exhibiting a large 230 nm long x 107 nm wide spindle with a variable length tail (68 nm on average), STSV1 also contains the largest dsDNA genome (75.2-kb) among the crenarchaeal viruses (272). Similar to other members, its genome lacks similarity to other archaeal virus genomes. Unlike others, it is a nonlytic virus and does not appear to integrate into its host genome, despite encoding a tyrosine recombinase (272). The genome of STSV1 appears heavily modified, with a large degree of methylation as well as modified cytosine residues. These modifications may grant a degree of protection against enzymatic attack (272). The second proposed member, APSV1, was isolated in 2011 through induction of an integrated provirus revealed in the genome of *Aeropyrum pernix* (155). Upon analysis of the host chromosome, the authors noticed a recombinase from the tyrosine recombinase family and modified their culturing conditions to induce viral production. The APSV1 particle is a pleomorphic, spindle-shaped virion roughly 200 nm x 50 nm width with single or double-tailed forms that range from 20 to 50 nm to upwards of 700 nm (155). Like several of the *Fuselloviridae*, APSV1 has been seen to form rosette-like structures through interaction of their tail fibers. The APSV1 genome is a circular, double-stranded, and 38.0-kb in length. Being a provirus, an integrase gene was also found (155).

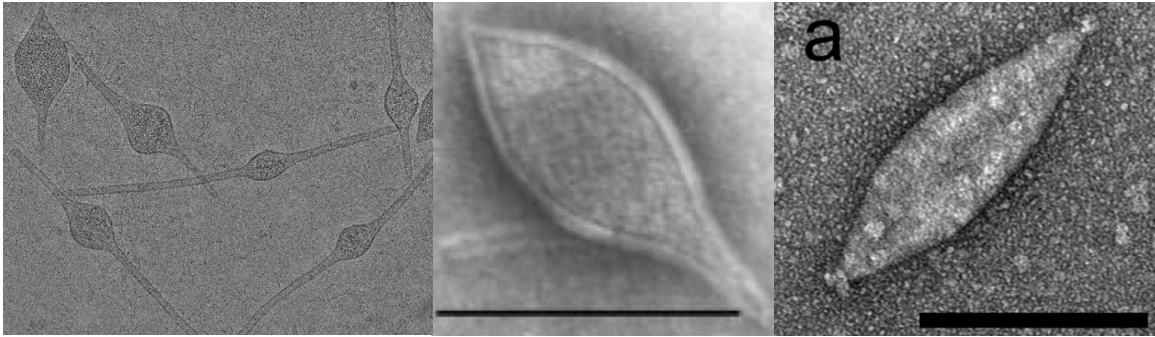


Figure 1.3. The *Bicaudaviridae*. From left-to-right, AT, STSV and APSV1. Image for ATV taken from (192), STSV from (272) and APSV1 from (155).

Salterprovirus. The only member of this genus, His1, infects the extremely halophilic *Euryarchaeota Haloarcula* (25). This genus originally contained His2, but recent analyses have shown His2 to be member of a group of pleomorphic viruses instead (177). His1 was also shown to be a non-lytic virus, opposed to previous reports of its potentially lytic nature (24). The virions are 44 nm wide by 74 nm long.

His1 is a linear, dsDNA, 14.4-kb genome, showing little similarity to other archaeal viruses, including those of the *Euryarchaeota*. The only significant similarity is the putative DNA polymerase of His1, showing homology to His2. Unlike other spindle-shaped viruses, His1 does not encode an integrase (25). The major structural protein, VP21 is a lipid-modified protein (178), unlike His2 containing a lipid bilayer (177).

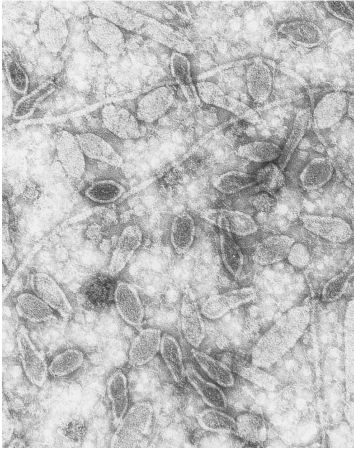


Figure 1.4. The Salterprovirus, His1. Image taken from (25)

Unclassified. To date, there are two spindle-shaped viruses that remain unclassified, PAV1 and TPV1, isolated from separate deep-sea hydrothermal vents. PAV1 was the first VLP isolated from a thermophilic euryarchaeote (93). A spindle-shape of 120 nm long x 60 nm wide with 15 nm tail fibers, PAV1 did not lyse its host or inhibit its growth. Initial work revealed a covalently-closed, dsDNA genome of approximately 18-kb that exists as an extrachromosomal element (seen in other viruses such as phage P4 (45)) in a carrier state, continuously producing virions (93). Attempts to induce viral product through UV irradiation, mitomycin C-treatment, or physiological stress were unsuccessful. Genome analysis of PAV1 revealed an 18.0-kb genome with nearly all the predicted genes on the same strand (92). The second unclassified, spindle-shaped virus is TPV1, the second virus isolated from a euryarchaeal thermophile (94). Its morphology is consistent with other spindle-shaped viruses, roughly 140 nm long x 80 nm wide with a 15 nm tail. The TPV1 genome is circular, dsDNA of approximately 21.5-kb. Like PAV1, it exists as an episomal element, with roughly 20 copies per host chromosome (94). As with PAV1, viruses are released continuously and do not cause

host cell lysis, but unlike PAV1, UV irradiation induces viral production. Although TPV1 encodes a putative tyrosine recombinase and is believed to integrate into its host (94), sites of integration remain to be seen.

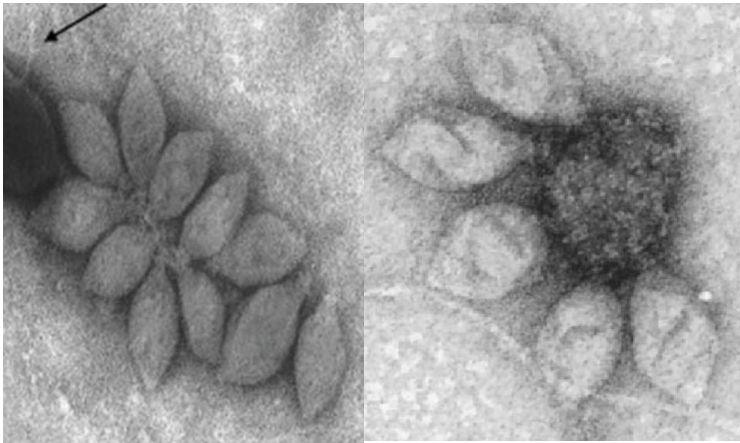


Figure 1.5. The Unclassified Spindle-Shaped Viruses, PAV and TPV1. Image for PAV taken from (93) and TPV1 from (94).

Linear Viruses

Linear virus particles are the most abundant morphology observed in hyperthermophilic environments (189, 195). Classified into two families, the flexible, filamentous *Lipothrixviridae* (13, 31, 103, 120) and the stiff, rod-like *Rudiviridae* (174, 188, 248), all members contain linear, dsDNA genomes and infect members of the *Crenarchaeota*. In addition, several studies have found a large number of shared genes such as glycosyl transferases and transcriptional regulators, as well as genome architecture, likely indicating a relatively recent ancestor diverged into the two families (191). Recent work also found members of each family to contain an identical four-helix bundle fold in its MCP, despite low sequence similarity (95). For these reasons the

families *Lipothrixviridae* and *Rudiviridae* have been classified into the order *Ligamenvirales* (189), owing to their common evolutionary history.

Lipothrixviridae. There are currently eleven members of the *Lipothrixviridae*, infecting *Sulfolobus*, *Acidianus*, and *Thermoproteus* (1). Except for TTV1, all members of the *Lipothrixviridae* are non-lytic, exist in stable carrier states with their host, and are not induced by UV irradiation or mytomycin C treatment (31, 104, 247, 278). Virions range dramatically in length and width, from 400 nm to nearly 2,000 nm in length, and from 24 nm to 40 nm in width. Their long, flexuous particles are composed of dsDNA bound to globular subunits in a helical fold, wrapped in a lipid shell derived from host lipids (13, 247, 278). Structural modeling of AFV3 virions shows a strong agreement between the length of the particle and its genome size (247). While structurally very similar, members of the *Lipothrixviridae* have been subdivided into four genera (*Alpha*-, *Beta*-, *Gamma*- and *Deltalipothrixvirus*) based on their unique terminal structures (190).

Genomes of the *Lipothrixviridae* range in size from 15.9-kb (TTV1) to 41.1-kb (AFV9). While few genes are conserved between the genera, such as the MCP, members of the *Betalipothrixvirus* contain a high degree of conservation and gene order. Both gene content and order are conserved for the ‘left halves’ of AFV3, 6-8, and SIFV while the right halves are variable in order for AFV7 and SIFV, suggesting a recombination event between genera sharing little sequence similarity (247). The most conserved genes within the *Lipothrixviridae* are among members of the *Betalipothrixvirus*, including transcriptional regulators, helicases, a methyltransferase, glycosyltransferase, and a nuclease (247). Also remarkable are the long ITRs present at the termini of each genome,

ranging from 11 bp in AFV1 to more than 800 bp in SIFV. In each AFV genome, these consist of terminal repeats repeated 2-3 times with constant spacing and a conserved sequence near the terminus, whereas in SIRV each end consists of 6-7 27-bp perfect direct repeats (174).

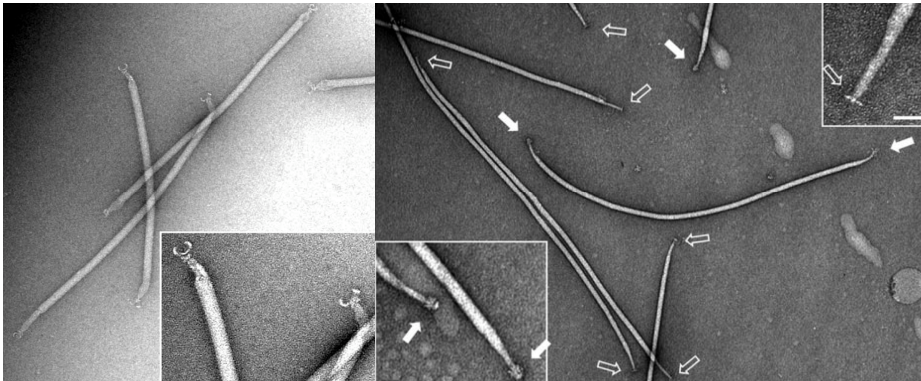


Figure 1.6. The *Lipothrixviridae*. AFV1 and a mixture of AFV6-8. Image for AFV1 taken from (31) and AFV6-8 from (247).

Rudiviridae. There are four members of the *Rudiviridae* family, infecting *Sulfolobus*, *Stygiolobus*, and *Acidianus*. As with the *Lipothrixviridae*, the *Rudiviridae* are non-lytic and are present in their hosts in a carrier state (188). They are distinguished from their sister family by possessing non-enveloped, stiff rods. Particles range in size from 610 nm to 900 nm in length and 25 nm in width. The length correlates to the size of the genome (174, 188). In a similar fashion to the *Lipothrixviridae*, virions are formed through the binding of a highly glycosylated protein to the dsDNA and the creation of a filament, which in turn creates a helical body (188). This overall architecture is shared by ssRNA viruses of the *Virgaviridae* and loosely with ssDNA viruses of the *Inoviridae*, but has not been previously observed in dsDNA viruses.

As members of the *Ligamenvirales*, *Rudiviridae* genomes contain linear, dsDNA ranging in size from 24.6-kb (ARV1) to 35.4-kb (SIRV2). At their ends are large ITRs, ranging from 1-kb (SRV) to 1.6-kb (SIRV2), with SRV, SIRV1, -2 having 4-5 copies of a direct repeat and ARV1 multiple degenerate copies of other diverse repeats (249). Conserved in all *Rudiviridae* is a 21-bp sequence located at the termini, thought to be a Holliday junction resolvase binding site (248). Also universally conserved are 17 predicted ORFs, including the major and minor capsid proteins, glycosyl transferases, a nuclease, Holliday junction resolvase, and transcriptional regulators (248, 249). Another 10 ORFs conserved in some, but not all genomes, while the unique sequences are generally located at the end of each genome (249). The genomes themselves are covalently closed, forming hairpin ends (188). Consistent with the presence of a Holliday junction resolvase and conserved terminal ends, SIRV1 replication produced head-to-head and tail-to-tail replicative intermediates (174), leading to the proposal that a self-priming mechanism is used during replication (187).

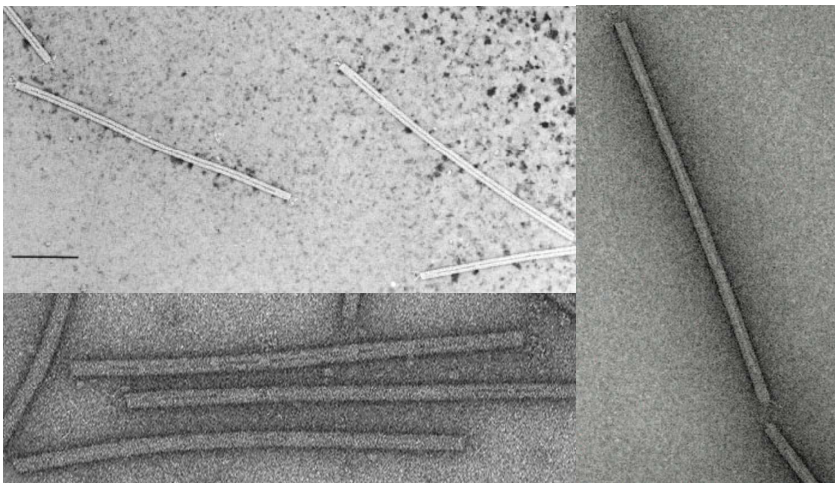


Figure 1.7. The *Rudiviridae*. Pictured top left is SIRV2, bottom left is ARV1, and right SRV. Images for SIRV2 taken from (188), ARV1 from (248) and SRV from (249).

Spherical Viruses

The six spherical viruses infect members of the *Crenarchaeota* (genus *Thermoproteus*, *Sulfolobus*) and the *Euryarchaeota* (genus *Haloarcula*) and are classified into families according to the presence/absence of a lipid envelope and structurally unique features.

Globulaviridae. There are currently two members of the *Globulaviridae*, PSV1 and TTSV1, both infecting anaerobic hyperthermophiles. PSV1, discovered in 2004 out of an enrichment culture of a terrestrial hot spring in YNP, USA, was the first proposed member of the *Globulaviridae*. Its spherical virion is approximately 100 nm in diameter, with spherical protrusions about 15 nm in diameter (101). TTSV1, also isolated from enrichments of a terrestrial hot springs, has a 70-nm diameter sphere lacking the protrusions seen in PSV (3). Morphologically similar, both virions consist of tightly packed nucleoproteins arranged in a superhelical conformation encased within an envelope of host-derived lipids (3, 101). Neither virus causes lysis in its host strain. Genomic analysis of the *Globulaviridae* mirror their shared morphologies - both possess linear, dsDNA genomes with nearly all ORFs located on one strand of the genome, exceptional for dsDNA viruses. Fifteen of the 38 TTSV1 ORFs were homologous with PSV, ranging from 20 – 49% identity while no ORFs matched to existing databases (3). In addition, the PSV genome encodes 190-bp ITRs at their termini, with covalently closed ends of their genome, not unlike the *Ligamenvirales* (101).

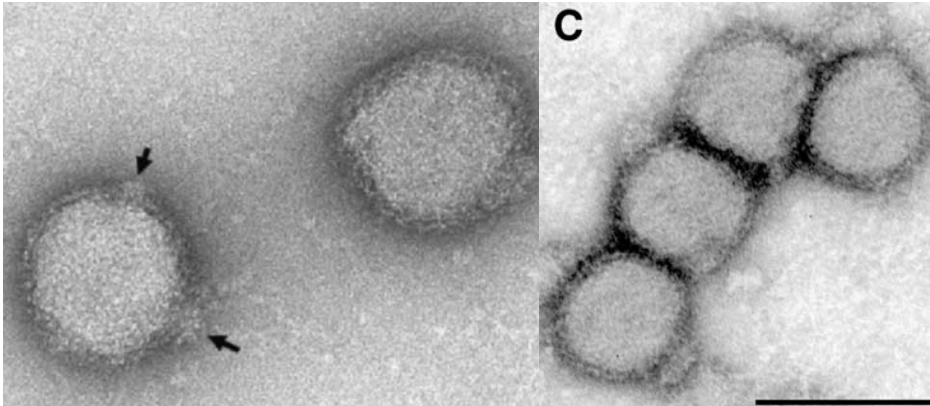


Figure 1.8. The *Globulaviridae*. Pictured left, PSV and right, TTSV1. Image for PSV taken from (101) and TTSV1 from (3).

Turriviridae. The two members of the proposed *Turriviridae*, STIV and STIV2, infect members of the thermoacidophilic *Sulfolobus*. STIV virions possess pseudo T=31 icosahedral symmetry with a diameter of 74 nm, an internal lipid membrane and exceptionally unique “turret-like” structures at each of the 12 5-fold vertices (202). Most remarkably, the double- β -barrel fold of the STIV MCP closely resembles that of eukaryal PBCV-1, adenovirus and bacteriophage PRD1, suggesting a common ancestry extending across all three domains of life (29, 127, 202). STIV2 is structurally quite similar to STIV1, with cryo-EM image reconstructions nearly indistinguishable between the two virions, except for the loss of petal-like decorations on the turrets of STIV2 (100, 126). Both *Turriviridae* genomes are circular, dsDNA of approximately 17.6-kb (STIV1) and 16.6-kb (STIV2). A large number of annotated ORFs are homologous between the two viruses, with some regions over 70% nucleotide identity (100). There is some disagreement over the presence and absence of the petals on the turret structure, as some propose that the decorated particles are a high-temperature form of the virus, while the undecorated result from morphological changes at low temperature (126).

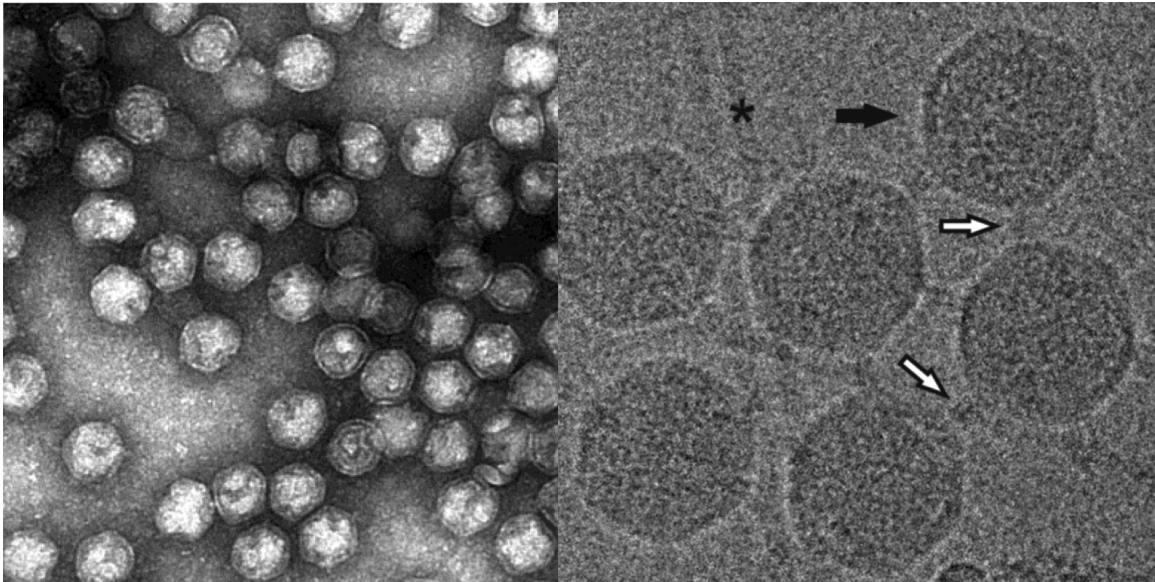


Figure 1.9. The proposed *Turriviridae* viruses. STIV (left) and STIV2 (right). Image for STIV taken from (64) and STIV2 from (100).

Halosphaerovirus. A very recently proposed group, the *halosphaeroviruses* include the remaining spherical viruses and are comprised of three extremely halophilic euryarchaeal viruses, SH1, PH1, and HHIV-2. SH1, the first archaeal spherical virus isolated (75) was the first described of this group (184). Initial characterization revealed strong similarities with STIV in virion architecture (as both contain icosahedral symmetry and an internal lipid layer) and structure of their MCPs, suggesting a common evolutionary heritage (119). This extends to all the halosphaeroviruses, as they all possess a spherical morphology, with 70-nm (SH1), 80-nm (HHIV-2), 51-nm (PH1) in diameter particles with an internal lipid layer (118, 184, 186). Unique to SH1 is the geometry of its capsid with $T = 28$ dextro, the first example of a complex capsid from single β -barrels (119). All members of this group contain linear, dsDNA genomes ranging from 28.0-kb to 30.8-kb with ITRs of 305-bp to 337-bp. Genome organization between the three members is highly similar, with a large number of homologous genes and overall

nucleotide similarities approaching 60% (118, 186). The nucleotide identities are so high that a detailed picture for the reasons of their genomic differences can be hypothesized (186). Recently an additional virus, SNJ1, has been suggested for inclusion into halosphaeroviruses. Originally identified as plasmid pHH205 of *Natrinema* J7-1 (150), SNJ1 is now believed to be not a member of the *Siphoviridae*, but instead halosphaerovirus (276). Like other members of this group, SNJ1 also possesses a spherical morphology, internal lipid layer and is believed to contain a PRD1-specific ATPase packaging motif conserved among the members of halosphaerovirus and *Turriviridae*. Unlike others in the group, SNJ1 has a circular, dsDNA genome, likely supporting a different replication mechanism (276).

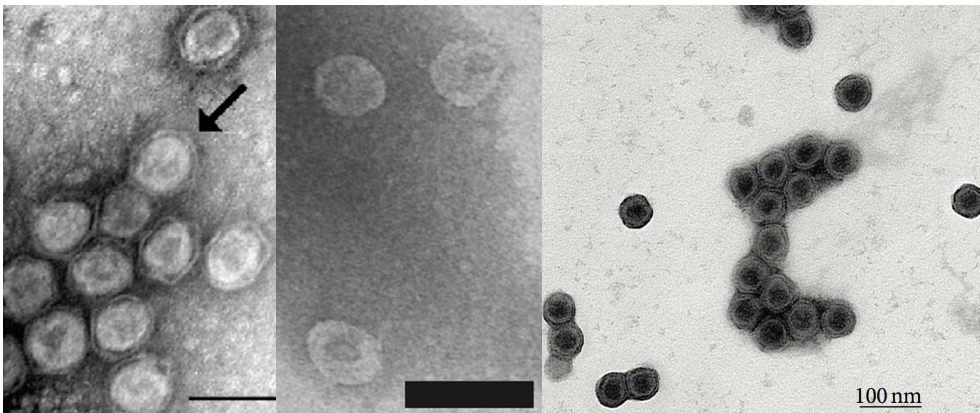


Figure 1.10. The proposed *Halosphaerovirus*. From left-to-right, SH1, HHIV2 and PH1. Image from SH1 from (184), HHIV-2 from (118) and PH1 from (186).

Pleomorphic Viruses

Originally consisting of only a single species, HRPV1 (179), the pleomorphic viruses have recently added 5 members and reclassified His2 (formerly Salterprovirus), now being tentatively placed into the pleolipoviruses group (177, 221). Members of this

group possess pleomorphic virions consisting of host-derived lipids and two major structural proteins facing opposite sides of their envelope (177). Genomic characterization has revealed a high degree of synteny with a collinear cluster of conserved genes. All the pleomorphic viruses encode homologous major structural proteins as well as those involved in viral assembly (221). The most remarkable aspect of the pleolipoviruses is the nature of their genomes and the resulting differences in their replication strategy, despite possessing regions of extraordinarily high similarity. HRPV-1, HRPV-2, HRPV-6 are ssDNA genomes with identified rep proteins, HGPV-1 and HRPV-3 are dsDNA genomes, and His2 is a linear dsDNA likely using protein-primed replication (221).

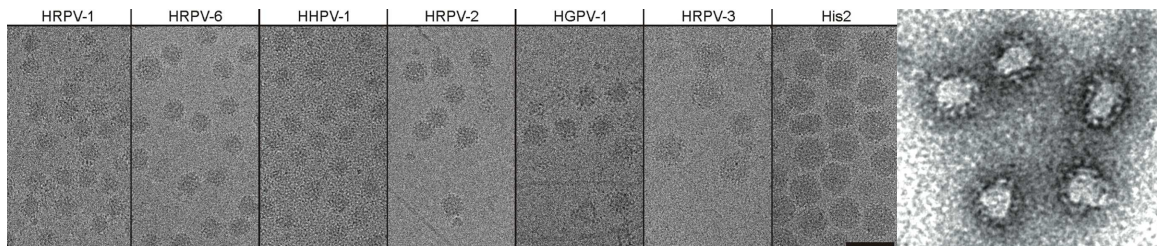


Figure 1.11. A panel of pleomorphic viruses. Pleomorphic viruses and a close up of HRPV1. Panel of images taken from (177) and HRPV1 from (206).

Head-Tail Viruses

Archaeal head-tail viruses are found exclusively among the *Euryarchaeota*, infecting extreme halophiles or anaerobic methanogens (181). Their morphologies are characteristic of bacterial phage, with non-enveloped virions containing icosahedral heads and various length tails. Surprisingly, the head-tail morphology is rarely observed in hypersaline environments, instead of spindle, star and linear shapes despite

representing the majority of known haloviruses (75, 185, 216). It is believed this misrepresentation results from biased approaches for isolation, due to the common method of screening through production of plaques on cell lawns (224). More than 43 head-tail viruses have been reported, with many of them not studied beyond basic description. Regardless, all of them have been assigned to the bacteriophage families *Myoviridae* and *Siphoviridae* with a single member of the *Podoviridae* (1, 14, 181). Due to the relatively large number of euryarchaeal viruses classified into the *Myoviridae* and *Siphoviridae* and the high degree of similarity that exists between them at both genomic and proteomic levels, they will be discussed together.

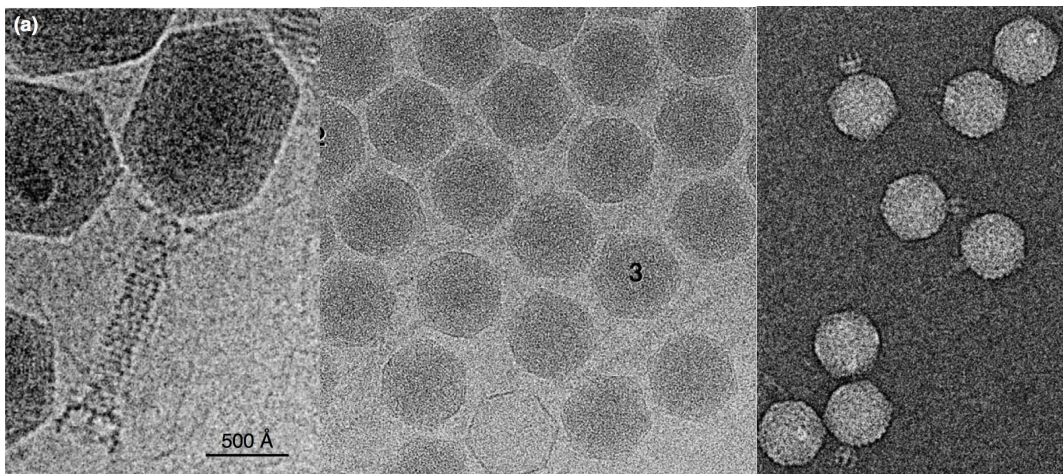


Figure 1.12. The head-tailed viruses of the *Myoviridae*, *Siphoviridae* and *Podoviridae*. T4 (left) image from (211), T5 (center) from (79) and P22 (right) from (49).

Myoviridae/Siphoviridae. *Myoviridae* are non-enveloped head-tail viruses with a “collar” separating the head and contractile tail. The *Siphoviridae* are characterized by a non-enveloped virion with non-contractile tails. More than 30 euryarchaeal viruses have been reported out of the *myoviridae* while about 10 have been reported for the *siphoviridae* (14). Of those that have been subsequently characterized, they show

incredible similarities to those of the bacteriophages. Though distant, a large number of putative genes are homologous to bacteriophages, including genes related to DNA synthesis, modification, replication and virus assembly (220, 237, 238). Both morphological and genetic similarities suggest genetic exchange between *Archaea* and *Bacteria* in hypersaline environments (185).

Podoviridae. The *podoviridae* are non-enveloped, head-tail viruses with short, non-contractile tails. The sole archaeal virus belonging to this formerly bacterial-only group is HSTV-1, a lytic virus whose circular, dsDNA genome contains several putative genes homologous to DNA replication, nucleotide metabolism, and structural proteins of tailed phages (14, 180). Analysis of the MCP reveals the canonical HK97-like fold, found in phage HK97 (258) as well other phage and herpesviruses (17, 79), supporting an evolutionary lineage of head-tail viruses evolving before the split of the three domains (180).

Other Unique Morphologies

Requiring novel families due to their truly exceptional viral morphology, these viral architectures are found exclusively in the *Crenarchaeota*. They are the bottle-shaped *Ampullaviridae* with ABV (102), the bacilliform *Clavaviridae* with APBV1 (156), and the droplet-shaped *Guttaviridae* consisting of SNDV and APOV1 (11, 155). Viruses from these three families possess circular (except ABV), dsDNA genomes and do not lyse their hosts.

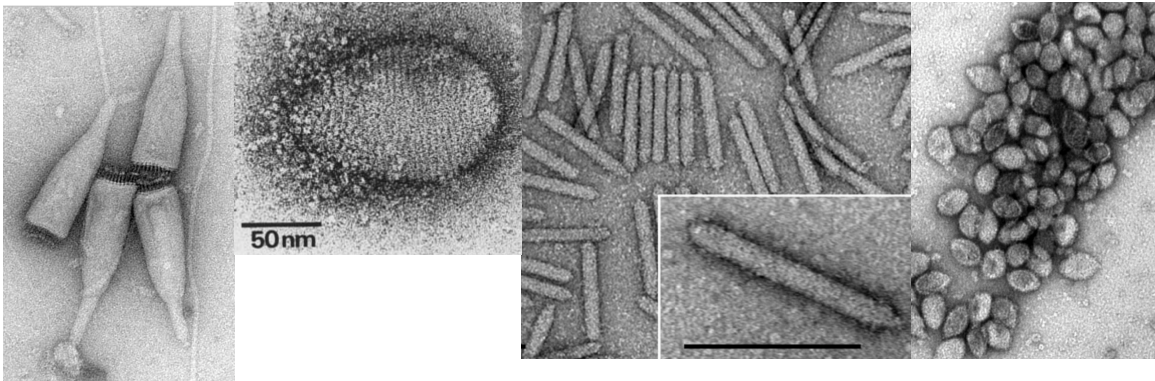


Figure 1.13. Archaeal viruses with exotic morphologies. Left-to-right, ABV, SNDV, APBV1, APBV1 close-up (inset) and APOV1. Images for ABV taken from (102), SNDV from (12), APBV1 from (156) and APOV1 from (155).

ABV virions appear bottle-shaped, with an overall length of 230-nm and a 75-nm width at its widest (the “bottom” of the bottle), tapering to a point of approximately 4-nm. The virion structure contains no icosahedral, spherical, or helical symmetry. The bottom is covered with a number of thin filaments that readily bind to other virions but do not appear to be involved in adsorption (102). The narrow ends can form rosette-like structures and readily attach to membrane vesicles of their host cells. An envelope covers a funnel-shaped core formed by “a torroidally supercoiled nucleoprotein filament” (102, 189). It is hypothesized that changes in the protein core could be involved in injecting or transporting the genome into the host cell. Genome analysis of ABV revealed long ITRs with 3 of 57 ORFs assignable to a function. Of these 3 ORFs, one was to a DNA polymerase family entirely consisting of linear dsDNA genomes with IRFs and covalent-linked terminal proteins (173). In agreement with the DNA polymerase family, the genome also revealed a putative RNA molecule essential for packaging linear dsDNA in the phi29 capsid (50, 51).

APBV1 is a rigid bacilliform, roughly 140-nm x 20-nm with one pointy end and one rounded end, lacking any distinctive surface features or inner structures (156). The genome is 5278-bp, the smallest prokaryotic dsDNA genome known and rivals the smallest eukaryal dsDNA viruses, such as SV40 (5243-bp, (85)) of the *polyomaviridae*. APBV1 contains 14 ORFs located on the same strand with none of them matching sequences in extant databases (156). It does integrate into its host, but exists as a pseudo-stable carrier state.

The *Guttaviridae*'s SNDV was first observed from a *Sulfolobus* isolate from New Zealand in 1996 out of Wolfram Zillig's lab. Its droplet-like morphology was densely covered by thin, long fibers with a helical structure on its surface (278). Further work revealed virion sizes ranging from 110-nm to 185-nm in length and 95-nm to 70-nm in width. The beehive-like ribbed pattern on its surface, covered with the long tail fibers are suspected to preferentially bind to breaks or folds in the S-layer of its host (11). Its 15.5-kb, heavily methylated circular genome does not integrate into its host, but persists as a pseudo-stable carrier state. This state depends on the balance between host and virus replication, and if disturbed, the host can become cured of the virus (11). The second virus of the *Guttaviridae*, APOV1, was found through induction of a provirus region in *Aeropyrum pernix*. APOV1 is a regular oval-shaped virion 80-nm by 60-nm with no observed surface structures or tail. It has a 13.7-kb genome containing 21 ORFs, 5 of which showed sequence similarities to other sequences. Three of these were to putative genes in hyperthermophiles, with the remaining two to *Fuselloviridae* integrase and DnaA-like proteins (155).

Spiraviridae. The recently proposed *Spiraviridae* family consists of one member, ACV, isolated from *A. pernix* (154). Its virion is non-enveloped and cylindrical, formed by a coiling fiber of a single nucleoprotein. This morphology has never been seen before in any virus. Equally exceptional is the nature of its genome, a single-stranded, circular genome of 24.9-kb (154). This is the first single-stranded genome isolated from the crenarchaeota, the second of the *Archaea*, and is twice as large as the previously largest ssDNA genome, Pf4 (255). The architecture of the single nucleoprotein allows for multiple levels of helical organization, with two nucleoprotein strands containing ssDNA that are in turn intertwined into a helical formation. This allows for “a unique architectural solution” for a ssDNA genome that apparently does not impose a constraint on genome size (154).

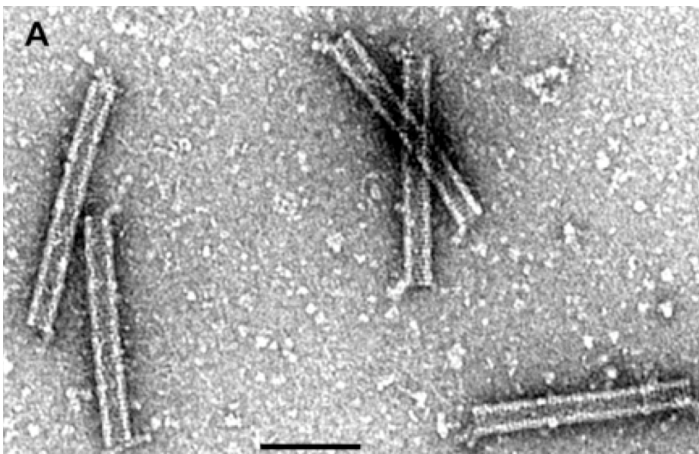


Figure 1.14. The *Spiraviridae*, ACV. Image taken from (154).

Viral Metagenomics

Environmental virology (and virology in general) has long been dependent on culturing potential hosts in order to study viruses and their virus-host interactions.

Cultivation-dependent methods, however, suffer from a serious limitation – not all organisms can be successfully cultured in laboratories. Culture-dependent methods are limited by a number of factors that essentially boil down to culturing conditions – the inability to perfectly mimic all environmental conditions appropriate for every microorganism in a laboratory. This bias that 99 to 99.9% has been estimated from a variety of diverse habitats of microorganisms are unable to be cultured (7, 59, 241). Metagenomics, a term first used in 1998 (99), was developed to overcome the limitations of culture-dependent approaches by directly sequencing microbial communities from their environments. It was later applied to viral communities, bypassing the need for culturing the host of a virus and directly sequencing viral genomes (43). It has been estimated that viruses outnumber cells by 10-100 (38, 78, 233); and is believed that, coupled with the inability to culture their hosts, the viral world is not only extraordinarily rich (in genotypes), but also essentially unknown.

In recent years, many groups have turned to culture-independent viral metagenomic approaches to investigate viral diversity in diverse natural environments. Since the advent of viral metagenomics, hundreds of papers have been published from differing environments, from marine to freshwater to soil and air (43, 107, 212, 226, 242, 260). These studies have revealed a number of exciting discoveries, but also a similar trend – between 50-70% of all sequences do not have a match to any sequence of viral or cellular origins (reviewed in (210)). In one study, nearly 99% of all phage sequences did not match to any extant database (68). Despite the recent development of cheaper, next-generation sequencing, (152, 219), these numbers appear to be holding constant. Of the

matching viral sequences in environmental samples, the sequence identities are low, often below 50% identity (210).

One of the main challenges in viral metagenomics is addressing the vast number of sequences without significant sequence similarities. These likely represent completely novel viruses, underscoring our lack of understanding in the viral sequence space (210). Analysis of these unknown sequences is often perplexing due to differences between researchers on the types of sequence search algorithms used, databases searched against, the sample type, and methods of processing the raw data (157). Depending on any one of these factors, the actual number of “unknown” sequences can vary greatly. Several other issues complicate the analysis of viral metagenomes, such as the presence of GTAs (81, 137, 227) and other “contaminating” cellular sequences. These gene transfer agents can appear as tailed bacteriophages, but transfer random fragments of host DNA. Further complicating the GTA issue are the findings that many viruses carry metabolic genes (AMGs) and can even transfer an entire metabolic pathway (44, 158). In an analysis of ocean viral communities, viral-encoded genes for energy and carbohydrate metabolism, Fe-S cluster synthesis and photosystem were found (44). Recent advances in preparatory methods, as well as the advent of next-generation sequencing, alleviate, but do not entirely eliminate these issues (described below). In general, there are three phases to any viral metagenomic study – sample preparation, sequencing, and post-sequencing analysis (commonly known as bioinformatics) (157).

Sample Preparation

While extensive work has been done to further each phase, the most critical step is sample generation. Each sample type poses its own unique challenges for which separate protocols are often empirically determined; they all serve the same function – isolation of viral nucleic acid from its native environment. The earliest viral metagenomes (marine) required large volume concentration steps, such as tangential flow filtration (TFF) and subsequent purification of viral particles through the use of cesium density gradients (43). Despite their large numbers, these concentration and purification steps are often necessary to ensure that sufficient quantities and purity of genetic material is available for sequencing. Unfortunately this method suffers from being unable to isolate large eukaryotic viruses, ssRNA viruses, those sensitive to CsCl, or any viruses too large or too small to pass through the filtration steps (43, 78, 269). This basic method was developed due to the “contamination” potential of the isolation. Inherent to any concentration or filtration steps is the possibility of non-target nucleic acid being isolated (and thus amplified) alongside viral. At ~2.5 Mb, the average microbial genome (225) is 50 times larger than the average 50-kb DNA viral genome (231, 232, 268). In environments where viruses outnumber cells (discussed above) 10:1 or 100:1, any numerical advantage the viral material would have is quickly overwhelmed by the larger microbial genomes. Free DNA in these environments (170, 171) further reduces the target to noise ratio. For these reasons most sample preparations incorporate DNase- and RNase-treatment steps to remove free nucleic acid. Following removal, purification of viruses from cells using cesium density gradients is often employed (ref), or an additional

filtration step to microorganisms larger than the average virus size (269). More recent sample preparation methods employ the use of FeCl_3 -precipitation that captures nearly 100% of SYBR-stainable viruses, is more quantitative than TFF and a substantially greater virus-to-cell ratio (122). One study examined the reproducibility and quantitative nature of this method and found it was not only quantitative and reproducible, but that it “maximized inter-comparability and biological inference” (73). Following isolation and purification, viral nucleic acids can be extracted through a number of different methods, such as chloroform and phenol-chloroform, formamide, cetyltrimethylammonium bromide (CTAB) extractions, and silica-based columns. Selection of the extraction method is highly dependent on the sample type and the research questions with the ultimate goal of generating viral nucleic acid of high purity and quantity. Despite the concentration of several orders of magnitude, many viral samples (especially those of marine) must subsequently be amplified to generate sufficient sequencing material. Traditional approaches used linker-amplified shotgun libraries (LASLs) (43, 239) and random amplified shotgun libraries (RASLs) (205) to meet the sequencing requirements from as little as 1 – 100 ng of input material and produce a relatively unbiased, representative sample containing fragments proportional to their concentration in the original sample (205). Later on, phi29-based amplification strategies were employed to further increase the amount of sequenceable material and allow for even lower starting input (9). These cloning-dependent methods, however, suffer from being both expensive and time-consuming (239). In the case of phi29, they are liable to systematic biases towards circular and ssDNA (128), the formation of chimeras (139) as well as stochastic

biases that skew taxonomic representation (274, 275). With the advent of massively parallel sequencing, dramatically increasing throughput and reducing costs per base (described below), a shift in sample preparation toward cloning-independent approaches (98) has occurred. Instead of preparing nucleic acid for cloning, DNA is (potentially) amplified and fragmented to a size amenable to the type of sequencing employed. This dramatically changed the field of environmental virology by enabling research to focus on amplification and relying on the sequencing platform to generate massive amounts of data rather than sacrifice amplification improvements to the bottleneck inherent in cloning coupled with Sanger sequencing. To this end, several new amplification strategies have been developed such as linker amplification (106, 108) and Nextera. Recent improvements to the linker strategy has allowed for sub-nanogram quantities of DNA to amplify into 1 – 5 µg of nearly quantitative, sequence-ready material (72).

Sequencing

The second step in any viral metagenomes pipeline is the selection of the sequencing platform. Prior to the last decade, nearly all sequencing was performed through capillary-based application of the Sanger biochemistry (214, 235) – the so-called “first-generation” technology. The first viral marine (43), gut (41), marine sediment (40), and viral RNA metagenomes (57, 58) all utilized this platform. The later onset of studying RNA viruses is attributed to methodological challenges in amplifying viral RNA out of environmental samples and through the use of degenerate primers (57) as well as random amplification (56). Through improvement over the decades, the Sanger-based technology can generate single read lengths of over 1,000-bp and raw-base accuracies

approaching 99.999% (223). After the completion of the Human Genome Project, it became clear that the cost and time required from Sanger biochemistry was insufficient for sequencing larger numbers of human genomes. It was not until the high-throughput and high parallelization of pyrosequencing, a principle used for genotyping since the mid 1990s (207, 209), that the cost and time required dramatically decreased and the era of “next-generation sequencing” (NGS) began. A number of extensive reviews on the history, chemistry and advancement of many NGS technologies have been published elsewhere (152, 223) and will not be reiterated here, except for the two most widespread in use and relevant to the advancement of viral metagenomics. These two NGS technologies are 454 pyrosequencing and Illumina technology (also known as Solexa). Both utilize the principle of sequencing by synthesis, a process in which a polymerase or ligase extends a large number of DNA strands in parallel (91). Nucleotides can be provided one at a time and read at the end of a repeating cycle, or labeled with an identifying tag so they can be added all at once with both 454 and Illumina using the former method. While their basic principles have existed for decades, it was not until technological improvements in surface chemistry, optics, computation, microscopy and engineering, among others (223), allowed for a rapid advancement into a post-NGS world. In many aspects, advances in NGS technologies are proceeding so quickly that reviews on the subject are outdated by the time they are published.

Early adopters of 454 technology were interested in sequencing and diagnosis of novel viruses in human disease (32, 162, 168) and plants (132). Continuing the trend, recent use of the technology has been applied as a diagnostics tool in virology, enabling

testing on an individualized scale and detection of ultra-low-abundance of antiviral drug-resistance as well as a means of studying genome variability, discovery of new pathogens, and studying viral communities in healthy and diseased individuals (23, 196). Least surprising is the driving force behind this rapid conversion – human clinical significance – with 17 of the top 20 sequenced viruses causing human disease (196). 454 pyrosequencing begins with library construction, where target DNA molecules are fragmented to a desired size and linkers containing universal priming sites are ligated to each end. These adapter-ligated sequences are then subjected to emulsion PCR (70), a process in which each molecule of DNA is separated into a single strand, bound to the surface of a single 28- μ m bead and clonally amplified using PCR. Each bead (often numbering in the millions) is then bound to PicoTiterPlate wells (142) and then sequenced via the pyrosequencing method (208, 209), where cyclic rounds of specific nucleotides are added and nucleotide incorporation generates a signal of light, whose position and intensity is recorded via CDD. Current 454 pyrosequencing on the GS FLX+ platform can generate up to 1,000-bp reads and 700 Mb in less than a day with a consensus accuracy of 99.997%.

454 technology, in addition to being the first commercialized product on the market, has become the *de facto* sequencing technology after its emergence in 2007 with longer read lengths and lower error rates versus other NGS technologies. A number of novel viruses have been discovered through its application, including those associated with colony collapse disorder and potential reservoirs for animal-to-human transmissions, as well as proventricular dilatation disease (PDD), a disease characterized by ataxia and

seizures (55, 110, 140, 147). Additionally, the use of 454 pyrosequencing has allowed researchers to directly detect human pathogens without *a priori* genetic information (162). Viruses have not been only discovered through 454 pyrosequencing but fully assembled as well, without the need for host culturing (35, 69).

Illumina (still commonly referred to its original name, Solexa) technology was developed internally (30) and is based on reversible dye-terminators and engineered polymerases (83, 243). The objective was to create a sequencing platform of very high throughput and low cost, capable of delivering “routine whole human genome sequencing” (30). As with 454 pyrosequencing, the first step in the sequencing process is target fragmentation and subsequent ligation of adapter molecules with universal priming sites attached to each end. These DNA molecules are then attached to a solid substrate via a flexible linker complementary to the priming sites and amplified using bridge PCR (2, 83). This generates amplicons identical to the initial template that are both tethered to, and immobilized to a single physical location on the solid substrate. Following bridge PCR, each “location” consists of ~1,000 clonal amplicons, with several million total clusters. Afterward, the amplicons are linearized, sequencing primers are hybridized to the universal priming site and the sequencing commences. During the sequencing process, four nucleotides are chemically modified to become “reversible terminators” containing a chemically-cleavable fluorescent tag and modified in such a way that only one nucleotide can be incorporated during each cycle. These nucleotides are added to the linearized amplicons, single-base extension occurs, the entire physical surface is captured in four channels, and the nucleotides are cleaved of their fluorescent tag and the 3’

hydroxyl position is unblocked to allow for additional extension. This cycle then continues with the next round of nucleotides and proceeds until the desired number of cycles is attained. As improvements in technology become available, the read length (originally 36-bp single reads) has dramatically improved (currently 300-bp, paired-end reads).

The single most important impact is one of cost. At less than \$0.07/Mb, Illumina sequencing is almost 40,000X cheaper than Sanger (\$2600/Gb) and 150X cheaper than 454 (\$10) (143). Improvement in the near future will cause the price to drop to less than \$0.01/Mb, allowing for the sequencing of an entire human genome for less than \$1000 (115). With throughput approaching 1 Tb, Illumina technology allows for not only discovery of novel viruses, but the ability to discriminate within-host variants of highly diverse pathogens (26), rhinovirus evolution during human infection (54), and RNA virus replication fidelity (76). The ultra-deep coverage provided by Illumina sequencing can even allow the detection of ultra-low abundance viruses (~100 copies) from heavily host-contaminated clinical samples (146) and can be used to detect rare, drug-resistant HIV mutations (141).

The main drawback of Illumina sequencing is downstream analysis (detailed below). Ultra-deep coverage, extraordinarily large number of reads and read quality (219, 223) in addition to large computational resources, are often required to process such vast sums of data. 454 pyrosequencing has longer read lengths, but suffers from homopolymers and being the most expensive (per Mb) NGS technology. Nevertheless, improvements in NGS sequencing technology are accelerating faster than Moore's Law.

Bioinformatic Analysis

The final step in any viral metagenomics project is post-sequence analysis; a process in which sequence data collected from the prior stage is assembled and analyzed. The ultimate objective of any project is full characterization of the community – essentially, *who is there* and *what are they doing*. Specifically, these questions refer to the composition of community members, the contribution of each member towards the total community (also referred to as a “population”) and intra- and inter-species heterogeneity of their genes/genomes (219). In terms of viral metagenomics, this translates to assembly of viral genomes, viral community composition, viral biogeography and virus-host interactions. To answer these questions, sequence data is often, but not always assembled. Before the advent of NGS, these fundamental questions about the nature of the community and its members were limited in scope, owing to the limitations of the sequencing technology and did not require assembly. The earliest viral metagenomics projects were pre-NGS with sequences derived entirely from Sanger sequencing (see above) (39, 41, 43, 57, 84). Bioinformatic analyses of the sequences were often read-based, using BLAST (6) or other homology search tools against previously identified public and private databases to establish taxonomic and functional identification. While groundbreaking and offering researchers their first glimpses into environmental viral communities, the results from these works were essentially classification into known versus unknown sequences. Sequences either matched to one of the three major phage families – *podo-*, *myo-* or *sipho-viridae* or were classified as unknown (40, 43). During this “first generation” of viral metagenomes it was found that

>50% of viral sequences were unable to be classified (78) and only the most abundant viruses were sampled (261), exposing a clear gap in the field's understanding of viruses and their communities. Around this time, the field was beginning to see the vast levels of diversity within the viral world – with more than 1 million different genotypes within 1-kg of marine sediment (40) and possibly up to 100 million globally (203). In response, tools such as the Phage Proteomic Tree (204) and PHACCS (8) were developed. The Phage Proteomic Tree is a first-of-its-kind, genome-based taxonomy for phage. PHACCS (Phage Communities from Contig Spectrum) builds models of potential community structures until they match experimental input, offering the first such measure of viral community structure – including their richness, evenness, and overall diversity. It soon became clear that the current era's sequencing technology could not keep pace with other advances in the field. Despite the enormous cost of “deep” Sanger-based, metagenomic studies, it was revealed that only the most abundant viruses were sampled, that significant evolution of viruses can occur rapidly, and that viruses play a substantial role in metabolic processes (261). At this point few tools were available to understand/analyze any unknown sequences and all diversity measurements pointed to continually increasing signs of under-sampling wherever a viral community was analyzed. The arrival of NGS, with its massive throughput and low cost, provided an opportunity to analyze the viral community without cloning biases, with minimal sample, and the ability to sequence deep enough to “reach past” the most abundant viruses. Unfortunately, as with any technological advancement and its ability to overcome previous obstacles, is not without its own set of shortcomings. In the case of NGS it is sequence assembly. The single

greatest challenge to any sequencing project is the ability to correctly put together (i.e. “assemble”) fragments from numerous diverse and similar species. This challenge is confounded by uneven community composition and abundances (273) with some regions of many genomes receiving poor or uneven coverage, with some receiving none at all (219). Coverage variation can be due to chance or varying target abundance or biases in amplification and sequencing, none of which are easily resolved (153). Sanger sequencing with its few, but longer, reads is relatively easily assembled. However, for millions (even billions) of reads ranging from 36 bp to 700-800 bp an entirely different approach is required. A number of complex assembly algorithms, requiring computationally expensive, high-performance computing platforms, have been developed to overcome this problem (153, 219, 223). While there is no “silver bullet” algorithm, nearly all are based on graph structures. The three major graph types are Overlap-Layout-Consensus (OLC), which rely on overlap graphs, de Bruijn Graphs (DBG), based on K-mers, and greedy graphs, using either or both of the previous (153). While these assembler algorithms are described in detail elsewhere (143, 152, 219, 223), a brief description of each will be mentioned here. Sanger-data and 454-data typically use OLC algorithms and are generally optimized for larger genomes, such as bacteria and Archaea. There are a number of programs invoking the OLC algorithm, but the most notable are 454 Life Technologies’ Newbler (148), Celera (160), CAP/PCAP and Arachne (27). OLC assemblers work in three phases: find all overlaps in an all-verses-all pairwise comparison, construct a read layout with all overlaps, and finally a multiple sequence alignment based on the read layout (153). The most expensive step of this process is the

first stage, where all possible overlaps must be found. Due to this fact, most programs institute a “seed” and extension heuristic, where very short fragments from each sequence are compared against each other. Comparing large numbers of short sequences is less memory intensive than comparing whole reads and is the most frequently applied efficiency. Each of the above OLC-based programs employs different strategies within each of the three phases, generating a range of pros and cons for researchers to select depending on their data and particular research objectives. Greedy-based algorithms, such as those employed by SSAKE (253) and VCAKE (121), simply extend any initial contig or read by adding the next highest-scoring overlap against any other contigs or reads (153). The graph built through these programs is dramatically simplified due to only examining the highest-scoring overlap. While efficient, these assemblers suffer most heavily from “short-sightedness” – by selecting the highest-scoring match at each step-wise extension, they may not generate the best overall sequence. For example, they may extend one contig while sacrificing other contigs which could have grown larger (153). The de Bruijn graph assemblers (k-mer based specifically, below) are the most widely used for NGS sequencing, owing to their avoidance of computationally intensive all-verses-all overlaps, compressing redundant sequences (reducing memory costs) and possessing the ability to deal with the vast amounts of data generated from Illumina sequencing. On the other hand, these graphs store the actual sequence, potentially consuming all available memory (153). K-mer graphs are a subset of de Bruijn graphs, where reads are *in silico* fragmented into shorter fragments (called k-mers) with their suffix and prefixes representing nodes on a graph and edges connecting them are k-mers

with that suffix and prefix (53). “Solving” the graph is simply finding a path that traverses each edge exactly once, also known as the Eulerian cycle. Unfortunately repeats, palindromes, and sequencing error complicate the graph transversal, introducing multiple potential paths (153).

Despite the significant advances in assembly algorithms, most are not optimized for small, highly heterogeneous sequences such as those from a viral population, nor are most designed to handle metagenomic sequences. As recently as 2011, there were *no de novo* metagenome-specific assemblers available until Meta-IDBA was released (176). Until this release, single-genome assemblers were used in metagenomic studies (270, 271). By 2012, additional assemblers, such as Meta-Velvet (163) and IDBA-UD (175) – both extensions of their original programming – as well as MAP (136) and Genovo (138), were developed. Still, wide ranges of target abundance and population heterogeneity can introduce unresolvable graph structures in metagenomic sequences (219).

Following NGS assembly of viral metagenomes, few analysis tools are available to annotate and characterize the resulting assembled sequences (known as contigs). As viral sequences lack a single gene shared among all viruses (i.e. the 16S rRNA gene), molecular phylogenetic analysis is impossible (183). Similarity analyses such as BLAST often fail to find homologous relationships (and thus annotation) with around 90% of sequences from viral metagenomics projects lacking similarity to any extant sequence databases (114). There are, however, two annotation pipelines developed recently for analysis of viral metagenomes; Viral Informatics Resource for Metagenome Exploration (VIROME) (267) and Metavir (213). Metavir is an interactive web server that

characterizes a metavirome using a suite of four major tools. It uses the GAAS tool (10), which compares viral contigs against viral sequences in NCBI's Refseq database (194) for taxonomic classification. It then uses marker genes – viral sequences associated with specific viral families – and aligns them against the input sequences. This alignment is then used to generate phylogenetic trees. Finally, rarefaction curves are generated for the viral metagenomes through the clustering program Uclust (77). Sequences are rarefied by subsampling different numbers of inputs sequences and clustering, the resulting number of clusters can be plotted against the number of sequences sampled, allowing for a plot of the rarefaction curves (213). Unfortunately, the major limitation of Metavir lies in its dependence on homology searches against previously characterized viral genomes present in the RefSeq database and its reliance on conserved marker genes. While unknown sequences cannot be classified, they can be incorporated into the rarefaction analysis. VIROME, developed shortly after Metavir, is an analysis pipeline comparing viral sequences against UniRef and five annotated protein databases, as well as predicted protein sequences from environmental databases (267). This allows for taxonomic and/or environmental characterization and functional annotation for every predicted ORF in the input viral metagenome. Additionally, environmental terms and metadata associated with the various protein libraries also serves to further group sequences relationships into features such as nucleic acid types, ecosystems (forest, wetland, ocean, coastal, etc.) and physio-chemical (acidic, alkaline, saline, low-high temp, etc.), among others. Finally, VIROME pre-filters sequences based on sequence quality, performs a tRNA scan, and identifies rRNA sequences – all known to increase computational overhead and interfere

with downstream sequence analysis (267). At present, only Metavir and VIROME offer a user-friendly, web-based viral metagenomics workflow designed to perform comparative analysis that does not require the researcher to have a dedicated bioinformatics infrastructure in place. The only limitation of VIROME is not one of analysis, but one of CPU time and cost. A full plate of 454 sequencing (see above explanation) is approximately 1 million reads (500-bp average read length) and generates 500 Mb. With the shift towards Illumina sequencing, the average output can range from 50 to 1000 Gb increasing the minimum CPU time to 100,000 CPU hours and \$50,000, far more than the VIROME infrastructure was designed to handle.

Research Directions

Over the past decade, advancements in sequencing technology and improvements in methodologies allow researchers to study the unseen and essentially uncharacterized virosphere. These culture- and sequence-independent approaches enable the study of both low and highly abundant viruses, entirely novel viruses, and uncovered the incredible diversity of viruses within several ecosystems. Next-generation sequencing has not only allowed for the complete *de novo* re-assembly of viruses from metagenomic samples, as a by-product in many cases, but also probed both deeply and broadly into the place of viral communities within the virus-host dynamics of a system. For all these significant advances, little work has been done in studying the archaeal viruses. Despite groundbreaking work revealing exotic particle morphologies and genomes with no homology to extant databases, only 40 archaeal viruses were known at the beginning of

this work – compared to the more than 5500 viruses infecting *Bacteria* and *Eukarya*. More incredible is the absence of an archaeal RNA virus. Initial work searching for RNA viruses in environmental samples from a metagenomics perspective found a high diversity of picorna-like viruses; upwards of 98% of all the sequences belonging to positive-sense ssRNA viruses. Other work focusing on RNA viruses found similar results, but none were able to identify an archaeal host.

The hot springs of Yellowstone National Park (YNP) thus provided a unique opportunity to study an archaeal rich environment and search for the first archaeal RNA viruses. These hot springs had previously shown to be dominated by the *Archaea*. In addition, preliminary work in virus-host isolation and biogeochemical characterization of many springs in YNP had laid the fundamental framework to allow an exhaustive RNA virus-centric investigation.

Table 1.1 List of Known Archaeal Viruses. Archaeal viruses are colored by phylum they infect. *Crenarchaeota* (blue), *Euryarchaeota* (red) and *Thaumarchaeota* (green)

Morphology	Family/genus	Virus Name	Abbreviation	Host	Genome Size (kb)	Genome Type	C/ L	Ref
Spindle	Fuselloviridae	Sulfolobus spindle-shaped virus 1	SSV1	<i>Sulfolobus</i>	15.4	dsDNA	C	(169)
		Sulfolobus spindle-shaped virus 2	SSV2	<i>Sulfolobus</i>	14.7	dsDNA	C	(229)
		Sulfolobus spindle-shaped virus 3	SSV3	<i>Sulfolobus</i>	15	dsDNA	C	(228)
		Sulfolobus spindle-shaped virus 4	SSV4	<i>Sulfolobus</i>	15.1	dsDNA	C	(33)
		Sulfolobus spindle-shaped virus 5	SSV5	<i>Sulfolobus</i>	15.3	dsDNA	C	(198)
		Sulfolobus spindle-shaped virus 6	SSV6	<i>Sulfolobus</i>	15.6	dsDNA	C	
		Sulfolobus spindle-shaped virus 7	SSV7	<i>Sulfolobus</i>	17.6	dsDNA	C	
		Sulfolobus spindle-shaped virus 8	SSV8	<i>Sulfolobus</i>	16.4	dsDNA	C	(257)
		Sulfolobus spindle-shaped virus 9	SSV9	<i>Sulfolobus</i>	17.3	dsDNA	C	
	Bicaudaviridae	Acidianus spindle-shaped virus 1	ASV1	<i>Acidianus</i>	24.1	dsDNA	C	(198)
		Acidianus two-tailed virus	ATV	<i>Acidianus</i>	62.7	dsDNA	C	(105)
	Salterprovirus	His virus 1	His 1	<i>Haloarcula</i>	14.4	dsDNA	L	(25)
	"Spiraviridae"	Aeropyrum pernix coil-shaped virus	ACV	<i>Aeropyrum</i>	24.9	ssDNA	C	(154)
	Unclassified	Pyrococcus abyssi virus 1	PAV1	<i>Pyrococcus</i>	18	dsDNA	C	(93)
		Thermococcus prieurii virus 1	TPV1	<i>Thermococcus</i>	21.5	dsDNA	C	(94)
Aeropyrum pernix spindle-shaped virus 1		APSV1	<i>Aeropyrum</i>	38	dsDNA	C	(155)	
Sulfolobus tengchongensis spindle-shaped virus 1		STSV1	<i>Sulfolobus</i>	75.2	dsDNA	C	(272)	
Bottle	Ampullaviridae	Acidianus bottle-shaped virus	ABV	<i>Acidianus</i>	23.8	dsDNA	L	(102)
Bacilliiform	Clavaridae	Aeropyrum pernix bacilliiform virus 1	APBV1	<i>Aeropyrum</i>	5.2	dsDNA	C	(156)
	Guttaviridae	Sulfolobus neozealandicus droplet-shaped virus	SNDV	<i>Sulfolobus</i>	20	dsDNA	C	(11)
Aeropyrum pernix ovoid virus 1		APOV1	<i>Aeropyrum</i>	13.7	dsDNA	C	(155)	

Table 1.1 – Continued

Linear	Lipothrixviridae	Desulfurolobus ambivalens filamentous virus	DAFV	<i>Acidianus</i>	56	dsDNA	L	(277)
		Acidianus filamentous virus 1	AFV1	<i>Acidianus</i>	21	dsDNA	L	(31)
		Acidianus filamentous virus 2	AFV2	<i>Acidianus</i>	31.7	dsDNA	L	(103)
		Acidianus filamentous virus 3	AFV3	<i>Acidianus</i>	40.4	dsDNA	L	
		Acidianus filamentous virus 6	AFV6	<i>Acidianus</i>	39.5	dsDNA	L	(247)
		Acidianus filamentous virus 7	AFV7	<i>Acidianus</i>	36.8	dsDNA	L	
		Acidianus filamentous virus 8	AFV8	<i>Acidianus</i>	38.1	dsDNA	L	
		Acidianus filamentous virus 9	AFV9	<i>Acidianus</i>	41.1	dsDNA	L	(33)
		Sulfolobus islandicus filamentous virus 1	SIFV	<i>Sulfolobus</i>	40.8	dsDNA	L	(13)
		Thermoproteus tenax virus 1	TTV1	<i>Thermoproteus</i>	15.9	dsDNA	L	
		Thermoproteus tenax virus 2	TTV2	<i>Thermoproteus</i>	16	dsDNA	L	(120)
		Thermoproteus tenax virus 3	TTV3	<i>Thermoproteus</i>	27	dsDNA	L	
		Thermoproteus tenax virus 4	TTV4	<i>Thermoproteus</i>	17	dsDNA	L	(278)
		Sulfolobus islandicus rod-shaped virus 1	SIRV1	<i>Sulfolobus</i>	32.3	dsDNA	L	(277)
	Rudiviridae	Sulfolobus islandicus rod-shaped virus 2	SIRV2	<i>Sulfolobus</i>	35.4	dsDNA	L	(174)
		Stygiolobus rod-shaped virus	SRV	<i>Stygiolobus4</i>	28	dsDNA	L	(249)
		Acidianus rod-shaped virus 1	ARV1	<i>Acidianus</i>	24.6	dsDNA	L	(248)
	Globulaviridae	Pyrobaculum spherical virus 1	PSV1	<i>Pyrobaculum</i>	28.3	dsDNA	L	(101)
		Thermoproteus texax spherical virus 1	TTSV1	<i>Thermoproteus</i>	21.6	dsDNA	L	(3)
	"Turriviridae"	Sulfolobus turreted icosahedral virus	STIV	<i>Sulfolobus</i>	17.6	dsDNA	C	(202)
		Sulfolobus turreted icosahedral virus 2	STIV2	<i>Sulfolobus</i>	16.6	dsDNA	C	(100)
Spherical	"Halosphaerovirus"			<i>Haloarcula</i> <i>Haloferax</i>				(184)
		Spherical halovirus 1	SH1	<i>Halorubrum</i>	30.9	dsDNA	L	
		PH1	PH1	<i>Haloarcula</i>	28	dsDNA	L	(186)
		Haloarcula hispanica icosahedral virus 2	HHIV-2	<i>Haloarcula</i>	30.5	dsDNA	L	(14)
		SNJ1	SNJ1	<i>Natrinema</i>	16.3	dsDNA	C	

Table 1.1 – Continued

Pleomorphic	Pleolipoviruses	His virus 2	His 2	<i>Haloarcula</i>	16	dsDNA	L	(24)
		Haloarcula hispanica pleomorphic virus 1	HHPV-1	<i>Haloarcula</i>	8	dsDNA	C	(179)
		Halorubrum pleomorphic virus 1	HRPV-1	<i>Halorubrum</i>	7	ssDNA	C	(206)
		Halorubrum pleomorphic virus 2	HRPV-2	<i>Halorubrum</i>	10.6	ssDNA	C	
		Halorubrum pleomorphic virus 3	HRPV-3	<i>Halorubrum</i>	8.7	dsDNA	C int	(14)
		Halogeometricum pleomorphic virus 1	HGPV-1	<i>Halogeometricum</i>	9.7	dsDNA	C int	
		Halorubrum pleomorphic virus 6	HRPV-6	<i>Halorubrum</i>	8.5	ssDNA	C	(177)
		HF1	HF1	<i>Haloferax</i>	77.7	dsDNA	L	(166)
		HF2	HF2	<i>Halorubrum</i>	75.9	dsDNA	L	
		phiH	phiH	<i>Halobacterium</i>	59	dsDNA	L	(218)
Head-tail	Myoviridae	phiCh1	phiCh1	<i>Natrialba</i>	58.5	dsDNA	L	(262)
		Halorubrum sodomense head-tail virus 2	HSTV-2	<i>Halorubrum</i>	68.2	nd		(14)
		Halorubrum head-tail virus 1	HRTV-1	<i>Halorubrum</i>		nd		(135)
		Halorubrum head-tail virus 2	HRTV-2	<i>Halorubrum</i>		nd		
		Halorubrum head-tail virus 3	HRTV-3	<i>Halorubrum</i>		nd		
		Halorubrum head-tail virus 5	HRTV-5	<i>Halorubrum</i>		nd		
		Halorubrum head-tail virus 6	HRTV-6	<i>Halorubrum</i>		nd		
		Halorubrum head-tail virus 7	HRTV-7	<i>Halorubrum</i>		nd		
		Haloarcula japonica head-tail virus 1	HJTV-1	<i>Haloarcula</i>		nd		
		Haloarcula japonica head-tail virus 2	HJTV-2	<i>Haloarcula</i>		nd		(14)
		Halorubrum head-tail virus 8	HRTV-8	<i>Halorubrum</i>		nd		
		Halogramum head-tail virus 1	HGTV-1	<i>Halogramum</i>		nd		
		Haloarcula head-tail virus 1	HATV-1	<i>Haloarcula</i>		nd		
		Halorubrum sodomense head-tail virus 3	HSTV-3	<i>Halorubrum</i>		nd		
		Halorubrum head-tail virus 9	HRTV-9	<i>Halorubrum</i>		nd		
		Halorubrum head-tail virus 10	HRTV-10	<i>Halorubrum</i>		nd		
		Haloarcula head-tail virus 2	HATV-2	<i>Haloarcula</i>		nd		

Table 1.1 – Continued

Head-Tail	Myoviridae	Halophilic head-tail virus 1	HTV-1	<i>E200-7</i>	nd		(14)
		Halorubrum head-tail virus 11	HRTV-11	<i>Halorubrum</i>	nd		
		Halorubrum head-tail virus 12	HRTV-12	<i>Halorubrum</i>	nd		
		Halobacterium halobium virus 1	Hh1	Halobacterium	37.6	dsDNA	L (172)
		Halobacterium halobium virus 3	Hh3	Halobacterium	29.6	dsDNA	L (172)
		Halophage Ja. 1	Ja.1	Halobacterium	230	dsDNA	L (252)
		Halobacterium salinarium virus 1	Hs1	Halobacterium		nd	(240)
		Halobacterial phage phiN	phiN	Halobacterium	56		L (250)
		Halophage S45	S45	Halobacterium		dsDNA	L (61)
		Bacteriophage S5100	S5100	Halobacterium		nd	(60)
	Siphoviridae	Halobacterium phage B10	B10	Halobacterium		nd	(251)
		psiM1	psiM1	<i>Methanobacterium</i>	30.4	dsDNA	L (151)
		BJ1	BJ1	<i>Halorubrum</i>	43	dsDNA	L (167)
		Haloarcula vallismortis head-tail virus 1	HVTV-1	<i>Haloarcula</i>	101.3	nd	(14)
		Halorubrum head-tail virus 4	HRTV-4	<i>Halorubrum</i>		nd	(14)
		Haloarcula hispanica head-tail virus 1	HHTV-1	<i>Haloarcula</i>		nd	(135)
		Haloarcula californiae head-tail virus 1	HCTV-1	<i>Haloarcula</i>		nd	(135)
		Halorcula californiae head-tail virus 2	HCTV-2	<i>Haloarcula</i>		nd	(14)
		Halorcula californiae head-tail virus 5	HCTV-5	<i>Haloarcula</i>		nd	
		Haloarcula vallismortis head-tail virus 2	HVTV-2	<i>Haloarcula</i>		nd	(14)
	Podoviridae provirus	Haloarcula hispanica head-tail virus 2	HHTV-2	<i>Haloarcula</i>		nd	
		Haloarcula sinaiensis head-tail virus 1	HSTV-1	<i>Haloarcula</i>	32.2	dsDNA	C
		Nvi-pro1	Nvi-pro1	<i>Nitrosophaera</i>	24	-	(134)

Table 1.1 – Continued

Unclassified	psiM2	psiM2	<i>Methanobacterium</i>	26	dsDNA	L
	phiF1	phiF1	<i>Methanobacterium</i>	85	dsDNA	L
	phiF3	phiF3	<i>Methanobacterium</i>	36	dsDNA	C
	PG	PG	<i>Methanobacterium</i>	50	dsDNA	
	PSM1	PSM1	<i>Methanobacterium</i>	35	dsDNA	
	VTA	VTA	<i>Methanobacterium</i>	nd		
	S50.2	S50.2	Halobacterium	nd		
	S4100	S4100	Halobacterium	nd		(251)
	S41	S41	Halobacterium	nd		
	S50.2 Vm	S50.2 Vm	Halobacterium	nd		
	S41 Vm	S41 Vm	Halobacterium	nd		(62)

Literature Cited

1. **Ackermann, H. W., and D. Prangishvili.** 2012. Prokaryote viruses studied by electron microscopy, p. 1843-1849, *Arch Virol*, vol. 157.
2. **Adessi, C., G. Matton, G. Ayala, G. Turcatti, J.-J. Mermoud, P. Mayer, and E. Kawashima.** 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms, p. e87-e87, *Nucleic Acids Res*, vol. 28. Oxford Univ Press.
3. **Ahn, D.-G., S.-I. Kim, J.-K. Rhee, K. P. Kim, J.-G. Pan, and J.-W. Oh.** 2006. TTSV1, a new virus-like particle isolated from the hyperthermophilic crenarchaeote *Thermoproteus tenax*., p. 280-290, *Virology*, vol. 351.
4. **Albers, S.-V., Z. Szabó, and A. J. M. Driessen.** 2006. Protein secretion in the Archaea: multiple paths towards a unique cell surface, p. 537-547, *Nature Publishing Group*, vol. 4.
5. **Allers, T., and M. Mevarech.** 2005. Archaeal genetics — the third way, p. 58-73, *Nature Reviews Genetics*, vol. 6.
6. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**:3389-3402.
7. **Amann, R.** 2000. Who is out there? Microbial Aspects of Biodiversity, p. 1-8, *Syst. Appl. Microbiol.*, vol. 23. Urban & Fischer Verlag.
8. **Angly, F., B. Rodriguez-Brito, D. Bangor, P. McNairnie, M. Breitbart, P. Salamon, B. Felts, J. Nulton, J. Mahaffy, and F. Rohwer.** 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information., p. 41, *BMC Bioinformatics*, vol. 6.
9. **Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. *Plos Biol* **4**:2121-2131.
10. **Angly, F. E., D. Willner, A. Prieto-Davó, R. A. Edwards, R. Schmieder, R. Vega-Thurber, D. A. Antonopoulos, K. Barott, M. T. Cottrell, C. Desnues, E. A. Dinsdale, M. Furlan, M. Haynes, M. R. Henn, Y. Hu, D. L. Kirchman, T. McDole, J. D. McPherson, F. Meyer, R. M. Miller, E. Mundt, R. K. Naviaux, B. Rodriguez-Mueller, R. Stevens, L. Wegley, L. Zhang, B. Zhu, and F. Rohwer.** 2009. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes, p. e1000593, *PLoS Comput Biol*, vol. 5.

11. **Arnold, H. P., U. Ziese, and W. Zillig.** 2000. SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon *Sulfolobus*. *Virology* **272**:409-416.
12. **Arnold, H. P., U. Ziese, and W. Zillig.** 2000. SNDV, a Novel Virus of the Extremely Thermophilic and Acidophilic Archaeon *Sulfolobus*, p. 409-416, *Virology*, vol. 272.
13. **Arnold, H. P., W. Zillig, U. Ziese, I. Holz, M. Crosby, T. Utterback, J. F. Weidmann, J. K. Kristjanson, H. P. Klenk, K. E. Nelson, and C. M. Fraser.** 2000. A Novel Lipothrixvirus, SIFV, of the Extremely Thermophilic Crenarchaeon *Sulfolobus*, p. 252-266, *Virology*, vol. 267.
14. **Atanasova, N. S., E. Roine, A. Oren, D. H. Bamford, and H. M. Oksanen.** 2012. Global network of specific virus-host interactions in hypersaline environments., p. 426-440, *Environmental Microbiology*, vol. 14.
15. **Auchtung, T. A., C. D. Takacs-Vesbach, and C. M. Cavanaugh.** 2006. 16S rRNA Phylogenetic Investigation of the Candidate Division "Korarchaeota" p. 5077-5082, *Appl Environ Microb*, vol. 72.
16. **Auguet, J.-C., A. Barberán, and E. O. Casamayor.** 2009. Global ecological patterns in uncultured Archaea, p. 182-190, *The ISME Journal*, vol. 4. Nature Publishing Group.
17. **Baker, M. L., W. Jiang, F. J. Rixon, and W. Chiu.** 2005. Common Ancestry of Herpesviruses and Tailed DNA Bacteriophages, p. 14967-14970, *J. Virol.*, vol. 79.
18. **Baltimore, D.** 1971. Expression of animal virus genomes., p. 235-241, *Bacteriol Rev*, vol. 35.
19. **Bamford, D. H.** 2003. Do viruses form lineages across different domains of life?, p. 231-236, *Research in Microbiology*, vol. 154.
20. **Bamford, D. H., J. Caldentey, and J. K. H. Bamford.** 1995. Bacteriophage Prd1: A Broad Host Range Dsdna Tectiviruses With an Internal Membrane, p. 281-319. *In* F. A. M. Karl Maramorosch and J. S. Aaron (ed.), *Advances in Virus Research*, vol. Volume 45. Academic Press.
21. **Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace.** 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences, p. 9188-9193, *Proceedings of the National Academy of Sciences*, vol. 93. National Acad Sciences.

22. **Barns, S. M., R. E. Fundyga, M. W. Jeffries, and N. R. Pace.** 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment., p. 1609-1613, *Proceedings of the National Academy of Sciences*, vol. 91.
23. **Barzon, L., E. Lavezzo, V. Militello, S. Toppo, and G. Palù.** 2011. Applications of Next-Generation Sequencing Technologies to Diagnostic Virology, p. 7861-7884, *IJMS*, vol. 12.
24. **Bath, C., T. Cukalac, K. Porter, and M. L. Dyll-Smith.** 2006. His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus, p. 228-239, *Virology*, vol. 350.
25. **Bath, C., and M. L. Dyll-Smith.** 1998. His1, an Archaeal Virus of the Fuselloviridae Family That Infects *Haloarcula hispanica*.
26. **Batty, E. M., T. H. N. Wong, A. Trebes, K. Argoud, M. Attar, D. Buck, C. L. C. Ip, T. Golubchik, M. Cule, R. Bowden, C. Manganis, P. Klenerman, E. Barnes, A. S. Walker, D. H. Wyllie, D. J. Wilson, K. E. Dingle, T. E. A. Peto, D. W. Crook, and P. Piazza.** 2013. A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples, p. e66129, *PLoS ONE*, vol. 8.
27. **Batzoglou, S.** 2002. ARACHNE: A Whole-Genome Shotgun Assembler, p. 177-189, *Genome Research*, vol. 12.
28. **Bell, S. D., and S. P. Jackson.** 1998. Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features, p. 222-228, *Trends in Microbiology*, vol. 6. Elsevier.
29. **Benson, S. D., J. K. H. Bamford, D. H. Bamford, and R. M. Burnett.** 2004. Does common architecture reveal a viral lineage spanning all three domains of life?, p. 673-685, *Molecular Cell*, vol. 16.
30. **Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H.**

- Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschield, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, et al.** 2008. Accurate whole human genome sequencing using reversible terminator chemistry, p. 53-59, *Nature*, vol. 456.
31. **Bettstetter, M., X. Peng, R. A. Garrett, and D. Prangishvili.** 2003. AFV1, a novel virus infecting hyperthermophilic archaea of the genus *acidianus*, p. 68-79, *Virology*, vol. 315.
 32. **Bishop-Lilly, K. A., M. J. Turell, K. M. Willner, A. Butani, N. M. E. Nolan, S. M. Lentz, A. Akmal, A. Mateczun, T. N. Brahmbhatt, S. Sozhamannan, C. A. Whitehouse, and T. D. Read.** 2010. Arbovirus Detection in Insect Vectors by Rapid, High-Throughput Pyrosequencing, p. e878, *PLoS Negl Trop Dis*, vol. 4.
 33. **Bize, A., X. Peng, M. Prokofeva, K. MacLellan, S. Lucas, P. Forterre, R. A. Garrett, E. A. Bonch-Osmolovskaya, and D. Prangishvili.** 2008. Viruses in acidic geothermal environments of the Kamchatka Peninsula, p. 358-366, *Research in Microbiology*, vol. 159.
 34. **Blöchl, E., R. Rachel, S. Burggraf, D. Hafenbradl, H. W. Jannasch, and K. O. Stetter.** 1997. *Pyrolobus fumarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C., p. 14-21, *Extremophiles*, vol. 1.
 35. **Blomstrom, A. L., F. Widen, A. S. Hammer, S. Belak, and M. Berg.** 2010. Detection of a Novel Astrovirus in Brain Tissue of Mink Suffering from Shaking Mink Syndrome by Use of Viral Metagenomics, p. 4392-4396, *J Clin Microbiol*, vol. 48.
 36. **Boucher, Y., M. Kamekura, and W. F. Doolittle.** 2004. Origins and evolution of isoprenoid lipid biosynthesis in archaea, p. 515-527, *Mol Microbiol*, vol. 52.
 37. **Bowers, K. J., and J. Wiegel.** 2011. Temperature and pH optima of extremely halophilic archaea: a mini-review, p. 119-128, *Extremophiles*, vol. 15. Springer.
 38. **Breitbart, M.** 2012. Marine Viruses: Truth or Dare, p. 425-448, *Annu. Rev. Marine. Sci.*, vol. 4.
 39. **Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2004. Diversity and population structure of a near-shore marine-

sediment viral community, p. 565-574, *Proceedings of the Royal Society B: Biological Sciences*, vol. 271.

40. **Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2004. Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**:565-574.
41. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2003. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**:6220-6223.
42. **Breitbart, M., and F. Rohwer.** 2005. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**:278-284.
43. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**:14250-14255.
44. **Breitbart, M., L. R. Thompson, C. A. Suttle, and M. B. Sullivan.** 2007. Exploring the vast diversity of marine viruses, p. 135, *OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY SOCIETY-*, vol. 20. The Oceanographic Society; 1999.
45. **Briani, F., G. Dehò, F. Forti, and D. Ghisotti.** 2001. The Plasmid Status of Satellite Bacteriophage P4. *Plasmid* **45**:1-17.
46. **Brochier-Armanet, C., B. Boussau, S. Gribaldo, and P. Forterre.** 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology* **6**:245-252.
47. **Cavicchioli, R.** 2010. Archaea — timeline of the third domain, p. 51-61, *Nature Publishing Group*, vol. 9. Nature Publishing Group.
48. **Chaban, B., S. Y. M. Ng, and K. F. Jarrell.** 2006. Archaeal habitats — from the extreme to the ordinary, p. 73-116, *Can. J. Microbiol.*, vol. 52.
49. **Chang, J., P. Weigele, J. King, W. Chiu, and W. Jiang.** 2006. Cryo-EM Asymmetric Reconstruction of Bacteriophage P22 Reveals Organization of its DNA Packaging and Infecting Machinery, p. 1073-1082, *Structure*, vol. 14.
50. **Chen, C., and P. Guo.** 1997. Magnesium-induced conformational change of packaging RNA for procapsid recognition and binding during phage phi29 DNA encapsidation., p. 495-500, *J. Virol.*, vol. 71.

51. **Chen, C., and P. Guo.** 1997. Sequential action of six virus-encoded DNA-packaging RNAs during phage phi29 genomic DNA translocation., p. 3864-3871, *J. Virol.*, vol. 71.
52. **Comeau, A. M., G. F. Hatfull, H. M. Krisch, D. Lindell, N. H. Mann, and D. Prangishvili.** 2008. Exploring the prokaryotic virosphere. *Research in Microbiology* **159**:306-313.
53. **Compeau, P. E. C., P. A. Pevzner, and G. Tesler.** 2011. How to apply de Bruijn graphs to genome assembly, p. 987-991, *Nature Biotechnology*, vol. 29. Nature Publishing Group.
54. **Cordey, S., T. Junier, D. Gerlach, F. Gobbini, L. Farinelli, E. M. Zdobnov, B. Winther, C. Tapparel, and L. Kaiser.** 2010. Rhinovirus Genome Evolution during Experimental Human Infection, p. e10588, *PLoS ONE*, vol. 5.
55. **Cox-Foster, D. L., S. Conlan, E. C. Holmes, G. Palacios, J. D. Evans, N. A. Moran, P.-L. Quan, T. Briese, M. Hornig, D. M. Geiser, V. Martinson, D. vanEngelsdorp, A. L. Kalkstein, A. Drysdale, J. Hui, J. Zhai, L. Cui, S. K. Hutchison, J. F. Simons, M. Egholm, J. S. Pettis, and W. I. Lipkin.** 2007. A metagenomic survey of microbes in honey bee colony collapse disorder, p. 283-287, *Science*, vol. 318.
56. **Culley, A. I.** 2006. Metagenomic Analysis of Coastal RNA Virus Communities, p. 1795-1798, *Science*, vol. 312.
57. **Culley, A. I., A. S. Lang, and C. A. Suttle.** 2003. High diversity of unknown picorna-like viruses in the sea. *Nature* **424**:1054-1057.
58. **Culley, A. I., A. S. Lang, and C. A. Suttle.** 2006. Metagenomic analysis of coastal RNA virus communities. *Science* **312**:1795-1798.
59. **Daniel, R.** 2004. The soil metagenome – a rich resource for the discovery of novel natural products, p. 199-204, *Current Opinion in Biotechnology*, vol. 15.
60. **Daniels, L. L., and A. C. Wais.** 1990. Ecophysiology of Bacteriophage-S5100 Infecting *Halobacterium-Cutirubrum*, p. 3605-3608, *Appl Environ Microb*, vol. 56.
61. **Daniels, L. L., and A. C. Wais.** 1984. Restriction and modification of Halophage S45 in *Halobacterium*, p. 133-136, *Current Microbiology*, vol. 10. Springer.
62. **Daniels, L. L., and A. C. Wais.** 1998. Virulence in phage populations infecting *Halobacterium cutirubrum*, p. 129-134, *FEMS Microbiology Ecology*, vol. 25. Wiley Online Library.

63. **de la Torre, J. R., C. B. Walker, A. E. Ingalls, M. Könneke, and D. A. Stahl.** 2008. Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol, p. 810-818, *Environmental Microbiology*, vol. 10.
64. **Dellas, N., C. M. Lawrence, and M. J. Young.** 2013. A Survey of Protein Structures from Archaeal Viruses, *Life*.
65. **DeLong, E. F.** 1992. Archaea in coastal marine environments., p. 5685-5689, *Proceedings of the National Academy of Sciences*, vol. 89.
66. **DeLong, E. F.** 1998. Everything in moderation: archaea as 'non-extremophiles', p. 649-654, *Curr. Opin. Genet. Dev.*, vol. 8.
67. **DeLong, E. F., and N. R. Pace.** 2001. Environmental diversity of bacteria and archaea, p. 470-478, *Systematic Biology*, vol. 50. Oxford University Press.
68. **Desnues, C., B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, M. Furlan, L. Wegley, B. Chau, Y. Ruan, D. Hall, F. E. Angly, R. A. Edwards, L. Li, R. V. Thurber, R. P. Reid, J. Siefert, V. Souza, D. L. Valentine, B. K. Swan, M. Breitbart, and F. Rohwer.** 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites, p. 340-343, *Nature*, vol. 452.
69. **Donaldson, E. F., A. N. Haskew, J. E. Gates, J. Huynh, C. J. Moore, and M. B. Frieman.** 2010. Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat., p. 13004-13018, *J. Virol.*, vol. 84.
70. **Dressman, D., H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein.** 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations, p. 8817-8822, *Proceedings of the National Academy of Sciences*, vol. 100. National Acad Sciences.
71. **Duffy, S., L. A. Shackelton, and E. C. Holmes.** 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**:267-276.
72. **Duhaime, M. B., L. Deng, B. T. Poulos, and M. B. Sullivan.** 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method, p. 2526-2537, *Environmental Microbiology*, vol. 14.
73. **Duhaime, M. B., and M. B. Sullivan.** 2012. Ocean viruses Rigorously evaluating the metagenomic sample-to-sequence pipeline, p. 1-5, *Virology*. Elsevier.

74. **Dutilh, B. E., B. Snel, T. J. G. Ettema, and M. A. Huynen.** 2008. Signature Genes as a Phylogenomic Tool, p. 1659-1667, *Molecular Biology and Evolution*, vol. 25.
75. **Dyall-Smith, M., S.-L. Tang, and C. Bath.** 2003. Haloarchaeal viruses: how diverse are they?, p. 309-313, *Research in Microbiology*, vol. 154.
76. **Eckerle, L. D., M. M. Becker, R. A. Halpin, K. Li, E. Venter, X. Lu, S. Scherbakova, R. L. Graham, R. S. Baric, T. B. Stockwell, D. J. Spiro, and M. R. Denison.** 2010. Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing, p. e1000896, *PLoS Pathog*, vol. 6.
77. **Edgar, R. C.** 2010. Search and clustering orders of magnitude faster than BLAST, p. 2460-2461, *Bioinformatics*, vol. 26.
78. **Edwards, R. A., and F. Rohwer.** 2005. Viral metagenomics, p. 504-510, *Nature Reviews Microbiology*, vol. 3. Nature Publishing Group.
79. **Effantin, G., P. Boulanger, E. Neumann, L. Letellier, and J. F. Conway.** 2006. Bacteriophage T5 Structure Reveals Similarities with HK97 and T4 Suggesting Evolutionary Relationships, p. 993-1002, *J Mol Biol*, vol. 361.
80. **Eilers, H., J. Pernthaler, F. O. Glockner, and R. Amann.** 2000. Culturability and In Situ Abundance of Pelagic Bacteria from the North Sea, p. 3044-3051, *Appl Environ Microb*, vol. 66.
81. **Eiserling, F., A. Pushkin, M. Gingery, and G. Bertani.** 1999. Bacteriophage-like particles associated with the gene transfer agent of *Methanococcus voltae* PS., p. 3305-3308, *J. Gen. Virol.*, vol. 80 (Pt 12).
82. **Elkins, J. G., M. Podar, D. E. Graham, K. S. Makarova, Y. Wolf, L. Randau, B. P. Hedlund, C. Brochier-Armanet, V. Kunin, I. Anderson, A. Lapidus, E. Goltsman, K. Barry, E. V. Koonin, P. Hugenholtz, N. Kyrpides, G. Wanner, P. Richardson, M. Keller, and K. O. Stetter.** 2008. A korarchaeal genome reveals insights into the evolution of the Archaea., p. 8102-8107, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105.
83. **Fedurco, M.** 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies, p. e22-e22, *Nucleic Acids Res*, vol. 34.
84. **Fierer, N., M. Breitbart, J. Nulton, P. Salamon, C. Lozupone, R. Jones, M. Robeson, R. A. Edwards, B. Felts, S. Rayhawk, R. Knight, F. Rohwer, and R. B. Jackson.** 2007. Metagenomic and small-subunit rRNA analyses reveal the

- genetic diversity of bacteria, archaea, fungi, and viruses in soil., p. 7059-7066, *Appl Environ Microb*, vol. 73.
85. **Fiers, W., R. Contreras, G. Haegemann, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert, and M. Ysebaert.** 1978. Complete nucleotide sequence of SV40 DNA., p. 113-120, *Nature*, vol. 273.
 86. **Forterre, P., C. Brochier, and H. Philippe.** 2002. Evolution of the Archaea, p. 409-422, *Theoretical Population Biology*, vol. 61.
 87. **Forterre, P., S. Gribaldo, and C. Brochier-Armanet.** 2007. Natural History of the Archaeal Domain, p. 17-28, *Archaea*. Blackwell Publishing Ltd.
 88. **Francis, C. A., K. J. Roberts, J. M. Beman, A. E. Santoro, and B. B. Oakley.** 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean., p. 14683-14688, *Proceedings of the National Academy of Sciences*, vol. 102.
 89. **Fröls, S., P. Gordon, M. A. Panlilio, and C. Schleper.** 2007. Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays, *Virology*.
 90. **Fuhrman, J. A.** 1992. Novel major archaeobacterial group from marine plankton. *Nature* **356**:148-149.
 91. **Fuller, C. W., L. R. Middendorf, S. A. Benner, G. M. Church, T. Harris, X. Huang, S. B. Jovanovich, J. R. Nelson, J. A. Schloss, D. C. Schwartz, and D. V. Vezenov.** 2009. The challenges of sequencing by synthesis, p. 1013-1023, *Nature Biotechnology*, vol. 27. Nature Publishing Group.
 92. **Geslin, C., M. Gaillard, D. Flament, K. Rouault, M. Le Romancer, D. Prieur, and G. Erauso.** 2007. Analysis of the First Genome of a Hyperthermophilic Marine Virus-Like Particle, PAV1, Isolated from *Pyrococcus abyssi*, p. 4510-4519, *Journal of Bacteriology*, vol. 189.
 93. **Geslin, C., M. Le Romancer, G. Erauso, M. Gaillard, G. Perrot, and D. Prieur.** 2003. PAV1, the First Virus-Like Particle Isolated from a Hyperthermophilic Euryarchaeote, "Pyrococcus abyssi", p. 3888-3894, *Journal of Bacteriology*, vol. 185.
 94. **Gorlas, A., E. V. Koonin, N. Bienvenu, D. Prieur, and C. Geslin.** 2011. TPV1, the first virus isolated from the hyperthermophilic genus *Thermococcus*, p. 503-516, *Environmental Microbiology*, vol. 14.
 95. **Goulet, A., S. Blangy, P. Redder, D. Prangishvili, C. Felisberto-Rodrigues, P. Forterre, V. Campanacci, and C. Cambillau.** 2009. *Acidianus filamentous*

virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage, p. 21155-21160, *Proceedings of the National Academy of Sciences*, vol. 106. National Acad Sciences.

96. **Grimes, J. M., J. N. Burroughs, P. Gouet, J. M. Diprose, R. Malby, S. Zientara, P. P. C. Mertens, and D. I. Stuart.** 1998. The atomic structure of the bluetongue virus core. *Nature* **395**:470-478.
97. **Gruber, N., and J. N. Galloway.** 2008. An Earth-system perspective of the global nitrogen cycle, p. 293-296, *Nature*, vol. 451.
98. **Hall, N.** 2007. Advanced sequencing technologies and their wider impact in microbiology, p. 1518-1525, *J. Exp. Biol.*, vol. 210.
99. **Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman.** 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, p. R245-R249, *Chemistry & Biology*, vol. 5. Elsevier.
100. **Happonen, L. J., P. Redder, X. Peng, L. J. Reigstad, D. Prangishvili, and S. J. Butcher.** 2010. Familial Relationships in Hyperthermo- and Acidophilic Archaeal Viruses, p. 4747-4754, *J. Virol.*, vol. 84.
101. **Häring, M., X. Peng, K. Brugger, R. Rachel, K. O. Stetter, R. A. Garrett, and D. Prangishvili.** 2004. Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the *Globuloviridae*, p. 233-242, *Virology*, vol. 323.
102. **Häring, M., R. Rachel, X. Peng, R. A. Garrett, and D. Prangishvili.** 2005. Viral diversity in hot springs of Pozzuoli, Italy, and characterization of a unique archaeal virus, *Acidianus* bottle-shaped virus, from a new family, the *Ampullaviridae*.
103. **Haring, M., G. Vestergaard, K. Brugger, R. Rachel, R. A. Garrett, and D. Prangishvili.** 2005. Structure and Genome Organization of AFV2, a Novel Archaeal Lipothrixvirus with Unusual Terminal and Core Structures, p. 3855-3858, *Journal of Bacteriology*, vol. 187.
104. **Häring, M., G. Vestergaard, K. Brugger, R. Rachel, R. A. Garrett, and D. Prangishvili.** 2005. Structure and genome organization of AFV2, a novel archaeal lipothrixvirus with unusual terminal and core structures, p. 3855-3858, *Journal of Bacteriology*, vol. 187. Am Soc Microbiol.
105. **Häring, M., G. Vestergaard, R. Rachel, L. Chen, R. A. Garrett, and D. Prangishvili.** 2005. Virology: Independent virus development outside a host, p. 1101-1102, *Nature*, vol. 436.

106. **Henn, M. R., M. B. Sullivan, N. Stange-Thomann, M. S. Osburne, A. M. Berlin, L. Kelly, C. Yandava, C. Kodira, Q. Zeng, M. Weiland, T. Sparrow, S. Saif, G. Giannoukos, S. K. Young, C. Nusbaum, B. W. Birren, and S. W. Chisholm.** 2010. Analysis of High-Throughput Sequencing and Annotation Strategies for Phage Genomes, p. e9083, PLoS ONE, vol. 5.
107. **Hewson, I., J. G. Barbosa, J. M. Brown, R. P. Donelan, J. B. Eaglesham, E. M. Eggleston, and B. A. LaBarre.** 2012. Temporal Dynamics and Decay of Putatively Allochthonous and Autochthonous Viral Genotypes in Contrasting Freshwater Lakes, p. 6583-6591, Appl Environ Microb, vol. 78.
108. **Hoeijmakers, W. A. M., R. Bartfai, K.-J. Francoijs, and H. G. Stunnenberg.** 2011. Linear amplification for deep sequencing. Nat. Protocols **6**:1026-1036.
109. **Hohn, M. J., B. P. Hedlund, and H. Huber.** 2002. Detection of 16S rDNA sequences representing the novel phylum "Nanoarchaeota"; indication for a wide distribution in high temperature biotopes., p. 551-554, Syst. Appl. Microbiol., vol. 25.
110. **Honkavuori, K. S., H. L. Shivaprasad, B. L. Williams, P.-L. Quan, M. Hornig, C. Street, G. Palacios, S. K. Hutchison, M. Franca, M. Egholm, T. Briese, and W. I. Lipkin.** 2008. Novel Borna Virus in Psittacine Birds with Proventricular Dilatation Disease, p. 1883-1886, Emerging Infectious Diseases, vol. 14.
111. **Huber, H., M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter.** 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont, p. 63-67, Nature, vol. 417.
112. **Huber, R., H. Huber, and K. O. Stetter.** 2000. Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties., p. 615-623, FEMS Microbiology Reviews, vol. 24.
113. **Huet, J., R. Schnabel, A. Sentenac, and W. Zillig.** 1983. Archaeobacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type., p. 1291-1294, EMBO J., vol. 2.
114. **Hurwitz, B. L., and M. B. Sullivan.** 2013. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology, p. e57355, PLoS ONE, vol. 8.
115. **Illumina** 2014, posting date. HiSeq X Ten Sequencing System. Illumina. [Online.]
116. **Ingalls, A. E., S. R. Shah, R. L. Hansman, L. I. Aluwihare, G. M. Santos, E. R. M. Druffel, and A. Pearson.** 2006. Quantifying archaeal community

- autotrophy in the mesopelagic ocean using natural radiocarbon., p. 6442-6447, Proceedings of the National Academy of Sciences, vol. 103.
117. **Iverson, E., and K. Stedman.** 2012. A genetic study of SSV1, the prototypical fusellovirus., p. 200, Front Microbiol, vol. 3.
 118. **Jaakkola, S. T., R. K. Penttinen, S. T. Vilen, M. Jalasvuori, G. Ronnholm, J. K. H. Bamford, D. H. Bamford, and H. M. Oksanen.** 2012. Closely Related Archaeal Haloarcula hispanica Icosahedral Viruses HHIV-2 and SH1 Have Nonhomologous Genes Encoding Host Recognition Functions, p. 4734-4742, J. Virol., vol. 86.
 119. **Jäälinoja, H. T., E. Roine, P. Laurinmäki, H. M. Kivelä, D. H. Bamford, and S. J. Butcher.** 2008. Structure and host-cell interaction of SH1, a membrane-containing, halophilic euryarchaeal virus, p. 8008-8013, Proceedings of the National Academy of Sciences, vol. 105. National Acad Sciences.
 120. **Janekovic, D., S. Wunderl, I. Holz, W. Zillig, A. Gierl, and H. Neumann.** 1983. TTV1, TTV2 and TTV3, a family of viruses of the extremely thermophilic, anaerobic, sulfur reducing archaeobacterium Thermoproteus tenax, p. 39-45, Molecular and General Genetics MGG, vol. 192. Springer.
 121. **Jeck, W. R., J. A. Reinhardt, D. A. Baltrus, M. T. Hickenbotham, V. Magrini, E. R. Mardis, J. L. Dangl, and C. D. Jones.** 2007. Extending assembly of short DNA sequences to handle error, p. 2942-2944, Bioinformatics, vol. 23.
 122. **John, S. G., C. B. Mendez, L. Deng, B. Poulos, A. K. M. Kauffman, S. Kern, J. Brum, M. F. Polz, E. A. Boyle, and M. B. Sullivan.** 2010. A simple and efficient method for concentration of ocean viruses by chemical flocculation, p. 195-202, Environmental Microbiology Reports, vol. 3.
 123. **Karner, M. B., E. F. DeLong, and D. M. Karl.** 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature **409**:507-510.
 124. **Kashefi, K., and D. R. Lovley.** 2003. Extending the upper temperature limit for life, p. 934-934, Science, vol. 301. American Association for the Advancement of Science.
 125. **Kellenberger, E.** 2001. Exploring the unknown. The silent revolution of microbiology., p. 5-7, EMBO Rep., vol. 2.
 126. **Khayat, R., C. y. Fu, A. C. Ortmann, M. J. Young, and J. E. Johnson.** 2010. The Architecture and Chemical Stability of the Archaeal Sulfolobus Turreted Icosahedral Virus, p. 9575-9583, J. Virol., vol. 84.

127. **Khayat, R., L. Tang, E. T. Larson, C. M. Lawrence, M. Young, and J. E. Johnson.** 2005. Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses, p. 18944-18949, *Proceedings of the National Academy of Sciences*, vol. 102. National Acad Sciences.
128. **Kim, K. H., and J. W. Bae.** 2011. Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses, p. 7663-7668, *Appl Environ Microb*, vol. 77.
129. **King, A. M., M. J. Adams, E. J. Lefkowitz, and E. B. Carstens.** 2011. Virus taxonomy: IXth report of the International Committee on Taxonomy of Viruses, vol. 9. Access Online via Elsevier.
130. **Könneke, M., A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl.** 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon, p. 543-546, *Nature*, vol. 437.
131. **Koonin, E. V., A. R. Mushegian, M. Y. Galperin, and D. R. Walker.** 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol* 25:619-637.
132. **Kreuze, J. F., A. Perez, M. Untiveros, D. Quispe, S. Fuentes, I. Barker, and R. Simon.** 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses, p. 1-7, *Virology*, vol. 388. Elsevier Inc.
133. **Krupovic, M., and D. H. Bamford.** 2010. Order to the Viral Universe, p. 12476-12479, *J. Virol.*, vol. 84.
134. **Krupovič, M., A. Spang, S. Gribaldo, P. Forterre, and C. Schleper.** 2011. A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea, p. 82-88, *Biochem. Soc. Trans*, vol. 39.
135. **Kukkaro, P., and D. H. Bamford.** 2009. Virus-host interactions in environments with a wide range of ionic strengths, p. 71-77, *Environmental Microbiology Reports*, vol. 1.
136. **Lai, B., R. Ding, Y. Li, L. Duan, and H. Zhu.** 2012. A de novo metagenomic assembly program for shotgun DNA reads, p. 1455-1462, *Bioinformatics*, vol. 28.
137. **Lang, A. S., and J. T. Beatty.** 2007. Importance of widespread gene transfer agent genes in alpha-proteobacteria., p. 54-62, *Trends in Microbiology*, vol. 15.
138. **Laserson, J., V. Jovic, and D. Koller.** 2011. Genovo: De Novo Assembly for Metagenomes, p. 429-443, *Journal of Computational Biology*, vol. 18.

139. **Lasken, R. S., and T. B. Stockwell.** 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction, p. 19, *BMC Biotechnol*, vol. 7.
140. **Lauck, M., D. Hyeroba, A. Tumukunde, G. Weny, S. M. Lank, C. A. Chapman, D. H. O'Connor, T. C. Friedrich, and T. L. Goldberg.** 2011. Novel, Divergent Simian Hemorrhagic Fever Viruses in a Wild Ugandan Red Colobus Monkey Discovered Using Direct Pyrosequencing, p. e19056, *PLoS ONE*, vol. 6.
141. **Le, T., J. Chiarella, B. B. Simen, B. Hanczaruk, M. Egholm, M. L. Landry, K. Dieckhaus, M. I. Rosen, and M. J. Kozal.** 2009. Low-Abundance HIV Drug-Resistant Viral Variants in Treatment-Experienced Persons Correlate with Historical Antiretroviral Use, p. e6079, *PLoS ONE*, vol. 4.
142. **Leamon, J. H., W. L. Lee, K. R. Tartaro, J. R. Lanza, G. J. Sarkis, A. D. deWinter, J. Berka, and K. L. Lohman.** 2003. A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions, p. 3769-3777, *Electrophoresis*, vol. 24.
143. **Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law.** 2012. Comparison of Next-Generation Sequencing Systems, p. 1-11, *Journal of Biomedicine and Biotechnology*, vol. 2012.
144. **Lwoff, A., and P. Tournier.** 1966. The classification of viruses. *Annual Reviews in Microbiology* 20:45-74.
145. **Makarova, K. S., A. V. Sorokin, P. S. Novichkov, Y. I. Wolf, and E. V. Koonin.** 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea, p. 33, *Biology Direct*, vol. 2.
146. **Malboeuf, C. M., X. Yang, P. Charlebois, J. Qu, A. M. Berlin, M. Casali, K. N. Pesko, C. L. Boutwell, J. P. DeVincenzo, G. D. Ebel, T. M. Allen, M. C. Zody, M. R. Henn, and J. Z. Levin.** 2013. Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification, *Nucleic Acids Res*, vol. 41.
147. **Maori, E., S. Lavi, R. Mozes-Koch, Y. Gantman, Y. Peretz, O. Edelbaum, E. Tanne, and I. Sela.** 2007. Isolation and characterization of Israeli acute paralysis virus, a dicistrovirus affecting honeybees in Israel: evidence for diversity due to intra- and inter-species recombination, p. 3428-3438, *J Gen Virol*, vol. 88.
148. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B.**

- Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors, *Nature*.
149. **Martin, A., S. Yeats, D. Janekovic, W. D. Reiter, W. Aicher, and W. Zillig.** 1984. Sav-1, a Temperate Uv-Inducible DNA Virus-Like Particle from the Archaeobacterium *Sulfolobus-Acidocaldarius* Isolate B-12. *Embo Journal* **3**:2165-2168.
 150. **Mei, Y., J. Chen, D. Sun, D. Chen, Y. Yang, P. Shen, and X. Chen.** 2007. Induction and preliminary characterization of a novel halophage SNJ1 from lysogenic *Natrinema* sp. F5. *Canadian journal of microbiology* **53**:1106.
 151. **Meile, L., U. Jenal, D. Studer, M. Jordan, and T. Leisinger.** 1989. Characterization of Psi-M1, a Virulent Phage of *Methanobacterium-Thermoautotrophicum* Marburg, p. 105-110, *Archives of Microbiology*, vol. 152.
 152. **Metzker, M. L.** 2009. Sequencing technologies — the next generation, p. 31-46, *Nature Reviews Genetics*, vol. 11. Nature Publishing Group.
 153. **Miller, J. R., S. Koren, and G. Sutton.** 2010. Assembly algorithms for next-generation sequencing data, p. 315-327, *Genomics*, vol. 95. Elsevier Inc.
 154. **Mochizuki, T., M. Krupovič, G. Pehau-Arnaudet, Y. Sako, P. Forterre, and D. Prangishvili.** 2012. Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome., p. 13386-13391, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109.
 155. **Mochizuki, T., Y. Sako, and D. Prangishvili.** 2011. Provirus Induction in Hyperthermophilic Archaea: Characterization of *Aeropyrum pernix* Spindle-Shaped Virus 1 and *Aeropyrum pernix* Ovoid Virus 1, p. 5412-5419, *Journal of Bacteriology*, vol. 193.
 156. **Mochizuki, T., T. Yoshida, R. Tanaka, P. Forterre, Y. Sako, and D. Prangishvili.** 2010. Diversity of viruses of the hyperthermophilic archaeal genus *Aeropyrum*, and isolation of the *Aeropyrum pernix* bacilliform virus 1, APBV1, the first representative of the family Clavaviridae, p. 347-354, *Virology*, vol. 402. Elsevier Inc.
 157. **Mokili, J. L., F. Rohwer, and B. E. Dutilh.** 2012. Metagenomics and future perspectives in virus discovery, p. 63-77, *Current Opinion in Virology*, vol. 2. Elsevier B.V.

158. **Monier, A., A. Pagarete, C. de Vargas, M. J. Allen, B. Read, J. M. Claverie, and H. Ogata.** 2009. Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus, p. 1441-1449, *Genome Research*, vol. 19.
159. **Muskhelishvili, G., P. Palm, and W. Zillig.** 1993. SSV1-encoded site-specific recombination system in *Sulfolobus shibatae*, p. 334-342, *Molecular and General Genetics MGG*, vol. 237. Springer.
160. **Myers, E. W.** 2000. A Whole-Genome Assembly of *Drosophila*, p. 2196-2204, *Science*, vol. 287.
161. **Nadal, M., G. Mirambeau, P. Forterre, W.-D. Reiter, and M. Duguet.** 1986. Positively supercoiled DNA in a virus-like particle of an archaebacterium. *Nature* **321**:256-258.
162. **Nakamura, S., C.-S. Yang, N. Sakon, M. Ueda, T. Tougan, A. Yamashita, N. Goto, K. Takahashi, T. Yasunaga, K. Ikuta, T. Mizutani, Y. Okamoto, M. Tagami, R. Morita, N. Maeda, J. Kawai, Y. Hayashizaki, Y. Nagai, T. Horii, T. Iida, and T. Nakaya.** 2009. Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach, p. e4219, *PLoS ONE*, vol. 4.
163. **Namiki, T., T. Hachiya, H. Tanaka, and Y. Sakakibara.** 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads, p. e155-e155, *Nucleic Acids Res*, vol. 40.
164. **Ng, S. Y. M., B. Zolghadr, A. J. M. Driessen, S. V. Albers, and K. F. Jarrell.** 2008. Cell Surface Structures of Archaea, p. 6039-6047, *Journal of Bacteriology*, vol. 190.
165. **Nicol, G. W., and C. Schleper.** 2006. Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle?, p. 207-212, *Trends in Microbiology*, vol. 14.
166. **Nuttall, S. D., and M. L. Dyll-Smith.** 1993. HF1 and HF2: novel bacteriophages of halophilic archaea, p. 678-684, *Virology*, vol. 197. Elsevier.
167. **Pagaling, E., R. D. Haigh, W. D. Grant, D. A. Cowan, B. E. Jones, Y. Ma, A. Ventosa, and S. Heaphy.** 2007. Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China., p. 410, *BMC Genomics*, vol. 8.
168. **Palacios, G., J. Druce, L. Du, T. Tran, C. Birch, T. Briesse, S. Conlan, P.-L. Quan, J. Hui, J. Marshall, J. F. Simons, M. Egholm, C. D. Paddock, W.-J. Shieh, C. S. Goldsmith, S. R. Zaki, M. Catton, and W. I. Lipkin.** 2008. A New

- Arenavirus in a Cluster of Fatal Transplant-Associated Diseases, p. 991-998, *N Engl J Med*, vol. 358.
169. **Palm, P., C. Schleper, B. Grampp, S. Yeats, P. McWilliam, W. D. Reiter, and W. Zillig.** 1991. Complete nucleotide sequence of the virus SSV1 of the archaeobacterium *Sulfolobus shibatae*., p. 242-250, *Virology*, vol. 185.
 170. **Paul, J. H., W. H. Jeffrey, and M. F. DeFlaun.** 1987. Dynamics of extracellular DNA in the marine environment., p. 170-179, *Appl Environ Microb*, vol. 53.
 171. **Paul, J. H., S. C. Jiang, and J. B. Rose.** 1991. Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration., p. 2197-2204, *Appl Environ Microb*, vol. 57.
 172. **Pauling, C.** 1982. Bacteriophages of *Halobacterium halobium*: isolation from fermented fish sauce and primary characterization, p. 916-921, *Canadian Journal of Microbiology*, vol. 28. NRC Research Press.
 173. **Peng, X., T. Basta, M. Häring, R. A. Garrett, and D. Prangishvili.** 2007. Genome of the Acidianus bottle-shaped virus and insights into the replication and packaging mechanisms, p. 237-243, *Virology*, vol. 364.
 174. **Peng, X., H. Blum, Q. She, S. Mallok, K. Brugger, R. A. Garrett, W. Zillig, and D. Prangishvili.** 2001. Sequences and Replication of Genomes of the Archaeal Rudiviruses SIRV1 and SIRV2: Relationships to the Archaeal Lipothrixvirus SIFV and Some Eukaryal Viruses, p. 226-234, *Virology*, vol. 291.
 175. **Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin.** 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, p. 1420-1428, *Bioinformatics*, vol. 28.
 176. **Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin.** 2011. Meta-IDBA: a de Novo assembler for metagenomic data, p. i94-i101, *Bioinformatics*, vol. 27.
 177. **Pietila, M. K., N. S. Atanasova, V. Manole, L. Liljeroos, S. J. Butcher, H. M. Oksanen, and D. H. Bamford.** 2012. Virion Architecture Unifies Globally Distributed Pleolipoviruses Infecting Halophilic Archaea, p. 5067-5079, *J. Virol.*, vol. 86.
 178. **Pietilä, M. K., N. S. Atanasova, H. M. Oksanen, and D. H. Bamford.** 2012. Modified coat protein forms the flexible spindle-shaped virion of haloarchaeal virus His1., *Environmental Microbiology*.
 179. **Pietilä, M. K., S. Laurinavičius, J. Sund, E. Roine, and D. H. Bamford.** 2010. The single-stranded DNA genome of novel archaeal virus halorubrum

pleomorphic virus 1 is enclosed in the envelope decorated with glycoprotein spikes., p. 788-798, *J. Virol.*, vol. 84.

180. **Pietilä, M. K., P. Laurinmäki, D. A. Russell, C.-C. Ko, D. Jacobs-Sera, R. W. Hendrix, D. H. Bamford, and S. J. Butcher.** 2013. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story., *Proc. Natl. Acad. Sci. U.S.A.*
181. **Pina, M., A. Bize, P. Forterre, and D. Prangishvili.** 2011. The archeoviruses, p. 1035-1054, *FEMS Microbiology Reviews*, vol. 35.
182. **Podar, M., K. S. Makarova, D. E. Graham, Y. I. Wolf, E. V. Koonin, and A.-L. Reysenbach.** 2013. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park, p. 1-1, *Biology Direct*, vol. 8. *Biology Direct*.
183. **Polson, S. W., S. W. Wilhelm, and K. E. Wommack.** 2010. Unraveling the viral tapestry (from inside the capsid out), p. 1-4, *The ISME Journal*. Nature Publishing Group.
184. **Porter, K., P. Kukkaro, J. K. H. Bamford, C. Bath, H. M. Kivelä, M. L. Dyall-Smith, and D. H. Bamford.** 2005. SH1: A novel, spherical halovirus isolated from an Australian hypersaline lake, p. 22-33, *Virology*, vol. 335.
185. **Porter, K., B. E. Russ, and M. L. Dyall-Smith.** 2007. Virus–host interactions in salt lakes, p. 418-424, *Current Opinion in Microbiology*, vol. 10.
186. **Porter, K., S.-L. Tang, C.-P. Chen, P.-W. Chiang, M.-J. Hong, and M. Dyall-Smith.** 2013. PH1: An Archaeovirus of *Haloarcula hispanica* Related to SH1 and HHIV-2, p. 1-17, *Archaea*, vol. 2013.
187. **Prangishvili, D.** 2003. Evolutionary insights from studies on viruses of hyperthermophilic archaea, p. 289-294, *Research in Microbiology*, vol. 154.
188. **Prangishvili, D., H. P. Arnold, D. Gotz, U. Ziese, I. Holz, J. K. Kristjansson, and W. Zillig.** 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.
189. **Prangishvili, D., P. Forterre, and R. A. Garrett.** 2006. Viruses of the Archaea: a unifying view, p. 837-848, *Nature Publishing Group*, vol. 4.
190. **Prangishvili, D., and R. A. Garrett.** 2005. Viruses of hyperthermophilic Crenarchaea. *Trends Microbiol* **13**:535-542.

191. **Prangishvili, D., R. A. Garrett, and E. V. Koonin.** 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res* **117**:52-67.
192. **Prangishvili, D., G. Vestergaard, M. Häring, R. Aramayo, T. Basta, R. Rachel, and R. A. Garrett.** 2006. Structural and Genomic Properties of the Hyperthermophilic Archaeal Virus ATV with an Extracellular Stage of the Reproductive Cycle, p. 1203-1216, *J Mol Biol*, vol. 359.
193. **Preston, C. M., K. Y. Wu, T. F. Molinski, and E. F. DeLong.** 1996. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov., p. 6241-6246, *Proceedings of the National Academy of Sciences*, vol. 93. National Acad Sciences.
194. **Pruitt, K. D., T. Tatusova, and D. R. Maglott.** 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, p. D61-D65, *Nucleic Acids Res*, vol. 35.
195. **Rachel, R., M. Bettstetter, B. P. Hedlund, M. H x000E4 ring, A. Kessler, K. O. Stetter, and D. Prangishvili.** 2002. Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments, p. 2419-2429, *Arch Virol*, vol. 147.
196. **Radford, A. D., D. Chapman, L. Dixon, J. Chantrey, A. C. Darby, and N. Hall.** 2012. Application of next-generation sequencing technologies in virology, p. 1853-1868, *J Gen Virol*, vol. 93.
197. **Ravantti, J. J., A. Gaidelyte, D. H. Bamford, and J. K. H. Bamford.** 2003. Comparative analysis of bacterial viruses Bam35, infecting a gram-positive host, and PRD1, infecting gram-negative hosts, demonstrates a viral lineage, p. 401-414, *Virology*, vol. 313.
198. **Redder, P., X. Peng, K. Brugger, S. A. Shah, F. Roesch, B. Greve, Q. She, C. Schleper, P. Forterre, R. A. Garrett, and D. Prangishvili.** 2009. Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism, p. 2849-2862, *Environmental Microbiology*, vol. 11.
199. **Reigstad, L. J., S. L. Jorgensen, and C. Schleper.** 2009. Diversity and abundance of Korarchaeota in terrestrial hot springs of Iceland and Kamchatka, p. 346-356, *The ISME Journal*, vol. 4. Nature Publishing Group.
200. **Reinisch, K. M., M. L. Nibert, and S. C. Harrison.** 2000. Structure of the reovirus core at 3.6 Å resolution., p. 960-967, *Nature*, vol. 404.

201. **Rice, G., K. Stedman, J. Snyder, B. Wiedenheft, D. Willits, S. Brumfield, T. McDermott, and M. J. Young.** 2001. Viruses from extreme thermal environments. *Proc Natl Acad Sci U S A* **98**:13341-13345.
202. **Rice, G., L. Tang, K. Stedman, F. Roberto, J. Spuhler, E. Gillitzer, J. E. Johnson, T. Douglas, and M. Young.** 2004. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life, p. 7716-7720, *Proceedings of the National Academy of Sciences*, vol. 101. National Acad Sciences.
203. **Rohwer, F.** 2003. Global phage diversity, p. 141, *Cell*, vol. 113. Elsevier.
204. **Rohwer, F., and R. Edwards.** 2002. The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage, p. 4529-4535, *Journal of Bacteriology*, vol. 184.
205. **Rohwer, F., V. Seguritan, D. H. Choi, A. M. Segall, and F. Azam.** 2001. Production of shotgun libraries using random amplification., p. 108-112- 114-106- 118, *Biotech.*, vol. 31.
206. **Roine, E., M. K. Pietila, L. Paulin, N. Kalkkinen, and D. H. Bamford.** 2009. An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. *Mol Microbiol* **72**:307-319.
207. **Ronaghi, M.** 2003. Pyrosequencing for SNP genotyping. *Methods in molecular biology (Clifton, NJ)* **212**:189.
208. **Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén.** 1996. Real-time DNA sequencing using detection of pyrophosphate release., p. 84-89, *Anal. Biochem.*, vol. 242.
209. **Ronaghi, M., M. Uhlén, and P. Nyren.** 1998. A sequencing method based on real-time pyrophosphate. *Science* **281**:363-365.
210. **Rosario, K., and M. Breitbart.** 2011. Exploring the viral world through metagenomics., p. 289-297, *Current Opinion in Virology*, vol. 1.
211. **Rossmann, M. G., V. V. Mesyanzhinov, F. Arisaka, and P. G. Leiman.** 2004. The bacteriophage T4 DNA injection machine, p. 171-180, *Curr Opin Struc Biol*, vol. 14.
212. **Roux, S., F. Enault, A. Robin, V. Ravet, S. Personnic, S. Theil, J. Colombet, T. Sime-Ngando, and D. Debroas.** 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics, p. e33641, *PLoS ONE*, vol. 7.

213. **Roux, S., M. Faubladier, A. Mahul, N. Paulhe, A. Bernard, D. Debroas, and F. Enault.** 2011. Metavir: a web server dedicated to virome analysis, p. 3074-3075, *Bioinformatics*, vol. 27.
214. **Sanger, F., G. Air, B. Barrell, N. Brown, A. Coulson, C. Fiddes, C. Hutchison, P. Slocombe, and M. Smith.** 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**:687.
215. **Sano, E., S. Carlson, L. Wegley, and F. Rohwer.** 2004. Movement of viruses between biomes, p. 5842-5846, *Appl Environ Microb*, vol. 70. Am Soc Microbiol.
216. **Santos, F., A. Meyerdierks, A. Peña, R. Rosselló-Mora, R. Amann, and J. Antón.** 2007. Metagenomic approach to the study of halophages: the environmental halophage 1. *Environmental Microbiology* **9**:1711-1723.
217. **Schleper, C., K. Kubo, and W. Zillig.** 1992. The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of transfection with viral DNA., p. 7645-7649, *Proceedings of the National Academy of Sciences*, vol. 89. National Acad Sciences.
218. **Schnabel, H., W. Zillig, M. Pfäffle, R. Schnabel, H. Michel, and H. Delius.** 1982. Halobacterium halobium phage øH., p. 87-92, *EMBO J.*, vol. 1.
219. **Scholz, M. B., C.-C. Lo, and P. S. G. Chain.** 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis., p. 9-15, *Current Opinion in Biotechnology*, vol. 23.
220. **Senčilo, A., D. Jacobs-Sera, D. A. Russell, C.-C. Ko, C. A. Bowman, N. S. Atanasova, E. Osterlund, H. M. Oksanen, D. H. Bamford, G. F. Hatfull, E. Roine, and R. W. Hendrix.** 2013. Snapshot of haloarchaeal tailed virus genomes., *RNA Biol*, vol. 10.
221. **Sencilo, A., L. Paulin, S. Kellner, M. Helm, and E. Roine.** 2012. Related haloarchaeal pleomorphic viruses contain different genome types, p. 5523-5534, *Nucleic Acids Res*, vol. 40.
222. **She, Q., K. Brügger, and L. Chen.** 2002. Archaeal integrative genetic elements and their impact on genome evolution. *Research in Microbiology* **153**:325-332.
223. **Shendure, J., and H. Ji.** 2008. Next-generation DNA sequencing, p. 1135-1145, *Nature Biotechnology*, vol. 26.
224. **Sime-Ngando, T., S. Lucas, A. Robin, K. P. Tucker, J. Colombet, Y. Bettarel, E. Desmond, S. Gribaldo, P. Forterre, M. Breitbart, and D. Prangishvili.** 2011. Diversity of virus–host systems in hypersaline Lake Retba, Senegal. *Environmental Microbiology* **13**:1956-1972.

225. **Simon, M., and F. Azam.** 1989. Protein content and protein synthesis rates of planktonic marine bacteria., p. 201-213, *Marine ecology progress series*. Oldendorf, vol. 51.
226. **Snyder, J. C., B. Wiedenheft, M. Lavin, F. F. Roberto, J. Spuhler, A. C. Ortmann, T. Douglas, and M. Young.** 2007. Virus movement maintains local virus population diversity., p. 19102-19107, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104.
227. **Stanton, T. B.** 2007. Prophage-like gene transfer agents—Novel mechanisms of gene exchange for *Methanococcus*, *Desulfovibrio*, *Brachyspira*, and *Rhodobacter* species, p. 43-49, *Anaerobe*, vol. 13.
228. **Stedman, K. M., A. Clore, and Y. Combet-Blanc.** 2006. Presented at the SYMPOSIA-SOCIETY FOR GENERAL MICROBIOLOGY.
229. **Stedman, K. M., Q. She, H. Phan, H. P. Arnold, I. Holz, R. A. Garrett, and W. Zillig.** 2003. Relationships between fuselloviruses infecting the extremely thermophilic archaeon *Sulfolobus*: SSV1 and SSV2. *Research in Microbiology* **154**:295-302.
230. **Stetter, K. O.** 1996. Hyperthermophilic procaryotes, p. 149-158, *FEMS Microbiology Reviews*, vol. 18. Wiley Online Library.
231. **Steward, G. F., and F. Azam.** 2000. Analysis of marine viral assemblages, p. 159-165, *Microbial Biosystems* (Bell, C., Brylinsky, M. and Johnson-Green, P., Eds.).
232. **Steward, G. F., J. L. Montiel, and F. Azam.** 2000. Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments, p. 1697-1706, *Limnology and Oceanography*, vol. 45.
233. **Suttle, C. A.** 2007. Marine viruses — major players in the global ecosystem, p. 801-812, *Nature Publishing Group*, vol. 5.
234. **Suttle, C. A.** 1994. The significance of viruses to mortality in aquatic microbial communities. *Microbial Ecology* **28**:237-243.
235. **Swerdlow, H., and R. Gesteland.** 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res* **18**:1415-1419.
236. **Takai, K., and K. Horikoshi.** 1999. Genetic diversity of archaea in deep-sea hydrothermal vent environments., p. 1285-1297, *Genetics*, vol. 152.

237. **Tang, S.-L., S. Nuttall, K. Ngui, C. Fisher, P. Lopez, and M. Dyll-Smith.** 2002. HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome., p. 283-296, *Mol Microbiol*, vol. 44.
238. **Tang, S. L., S. Nuttall, and M. Dyll-Smith.** 2004. Haloviruses HF1 and HF2: Evidence for a Recent and Large Recombination Event, p. 2810-2817, *Journal of Bacteriology*, vol. 186.
239. **Thurber, R. V., M. Haynes, M. Breitbart, L. Wegley, and F. Rohwer.** 2009. Laboratory procedures to generate viral metagenomes, p. 470-483, *Nat Protoc*, vol. 4.
240. **Torsvik, T., and I. Dundas.** 1974. Bacteriophage of *Halobacterium salinarum*. *Nature* **248**:680.
241. **Torsvik, V., J. Goksøyr, and F. L. Daae.** 1990. High diversity in DNA of soil bacteria., p. 782-787, *Appl Environ Microb*, vol. 56. Am Soc Microbiol.
242. **Tseng, C.-H., P.-W. Chiang, F.-K. Shiah, Y.-L. Chen, J.-R. Liou, T.-C. Hsu, S. Maheswararajah, I. Saeed, S. Halgamuge, and S.-L. Tang.** 2013. Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances, p. 1-13, *The ISME Journal*. Nature Publishing Group.
243. **Turcatti, G., A. Romieu, M. Fedurco, and A. P. Tairi.** 2007. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis, p. e25-e25, *Nucleic Acids Res*, vol. 36.
244. **Urbonavičius, J., S. Auxilien, H. Walbott, K. Trachana, B. Golinelli-Pimpanau, C. Brochier-Armanet, and H. Grosjean.** 2007. Acquisition of a bacterial RumA-type tRNA(uracil-54, C5)-methyltransferase by Archaea through an ancient horizontal gene transfer, p. 323-335, *Mol Microbiol*, vol. 67.
245. **Valentine, D. L.** 2007. Adaptations to energy stress dictate the ecology and evolution of the Archaea, p. 316-323, *Nature Publishing Group*, vol. 5.
246. **Van Regenmortel, M., and C. Fauquet.** 2000. Virus taxonomy: classification and nomenclature of viruses: seventh report of the International Committee on Taxonomy of Viruses.
247. **Vestergaard, G., R. Aramayo, T. Basta, M. Häring, X. Peng, K. Brugger, L. Chen, R. Rachel, N. Boisset, R. A. Garrett, and D. Prangishvili.** 2008. Structure of the acidianus filamentous virus 3 and comparative genomics of related archaeal lipothrixviruses., p. 371-381, *J. Virol.*, vol. 82.

248. **Vestergaard, G., M. Häring, X. Peng, R. Rachel, R. A. Garrett, and D. Prangishvili.** 2005. A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus *Acidianus*, p. 83-92, *Virology*, vol. 336.
249. **Vestergaard, G., S. A. Shah, A. Bize, W. Reitberger, M. Reuter, H. Phan, A. Briegel, R. Rachel, R. A. Garrett, and D. Prangishvili.** 2008. Stygiolobus Rod-Shaped Virus and the Interplay of Crenarchaeal Rudiviruses with the CRISPR Antiviral System, p. 6837-6845, *Journal of Bacteriology*, vol. 190.
250. **Vogelsang-Wenke, H., and D. Oesterhelt.** 1988. Isolation of a halobacterial phage with a fully cytosine-methylated genome, p. 407-414, *Molecular and General Genetics MGG*, vol. 211. Springer.
251. **Wais, A. C., and L. L. Daniels.** 1985. Populations of bacteriophage infecting *Halobacterium* in a transient brine pool, p. 323-326, *FEMS Microbiology Letters*, vol. 31. Elsevier.
252. **Wais, A. C., M. Kon, R. E. MacDonald, and B. D. Stollar.** 1975. Salt-dependent bacteriophage infecting *Halobacterium cutirubrum* and *H. halobium*., p. 314-315, *Nature*, vol. 256.
253. **Warren, R. L., G. G. Sutton, S. J. M. Jones, and R. A. Holt.** 2007. Assembling millions of short DNA sequences using SSAKE, p. 500-501, *Bioinformatics*, vol. 23.
254. **Waters, E., M. J. Hohn, I. Ahel, D. E. Graham, M. D. Adams, M. Barnstead, K. Y. Beeson, L. Bibbs, R. Bolanos, M. Keller, K. Kretz, X. Lin, E. Mathur, J. Ni, M. Podar, T. Richardson, G. G. Sutton, M. Simon, D. Soll, K. O. Stetter, J. M. Short, and M. Noordewier.** 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism., p. 12984-12988, *Proceedings of the National Academy of Sciences*, vol. 100.
255. **Webb, J. S., M. Lau, and S. Kjelleberg.** 2004. Bacteriophage and Phenotypic Variation in *Pseudomonas aeruginosa* Biofilm Development, p. 8066-8073, *Journal of Bacteriology*, vol. 186.
256. **Werner, F.** 2007. Structure and function of archaeal RNA polymerases, p. 1395-1404, *Mol Microbiol*, vol. 65.
257. **Wiedenheft, B., K. Stedman, F. Roberto, D. Willits, A.-K. Gleske, L. Zoeller, J. Snyder, T. Douglas, and M. Young.** 2004. Comparative genomic analysis of hyperthermophilic archaeal *Fuselloviridae* viruses, p. 1954-1961, *J. Virol.*, vol. 78. *Am Soc Microbiol*.

258. **Wikoff, W. R., L. Liljas, R. L. Duda, H. Tsuruta, R. W. Hendrix, and J. E. Johnson.** 2000. Topologically linked protein rings in the bacteriophage HK97 capsid., p. 2129-2133, *Science*, vol. 289.
259. **Wilhelm, S. W., and C. A. Suttle.** 1999. *Viruses and Nutrient Cycles in the Sea*. American Institute of Biological Sciences Circulation, AIBS, 1313 Dolley Madison Blvd., Suite 402, McLean, VA 22101. USA
260. **Williamson, K. E., K. E. Wommack, and M. Radosevich.** 2003. Sampling natural viral communities from soil for culture-independent analyses, p. 6628-6633, *Appl Environ Microb*, vol. 69. Am Soc Microbiol.
261. **Williamson, S. J., D. B. Rusch, S. Yooseph, A. L. Halpern, K. B. Heidelberg, J. I. Glass, C. Andrews-Pfannkoch, D. Fadrosh, C. S. Miller, G. Sutton, M. Frazier, and J. C. Venter.** 2008. The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples, p. e1456, *PLoS ONE*, vol. 3.
262. **Witte, A., U. Baranyi, R. Klein, M. Sulzner, C. Luo, G. Wanner, D. H. Krüger, and W. Lubitz.** 1997. Characterization of *Natronobacterium magadii* phage phi Ch1, a unique archaeal phage containing DNA and RNA., p. 603-616, *Mol Microbiol*, vol. 23.
263. **Woese, C. R.** 1987. Bacterial evolution., p. 221-271, *Microbiol. Rev.*, vol. 51.
264. **Woese, C. R.** 1994. There must be a prokaryote somewhere: microbiology's search for itself., p. 1-9, *Microbiol. Rev.*, vol. 58.
265. **Woese, C. R., and G. E. Fox.** 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences* **74**:5088-5090.
266. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* **87**:4576-4579.
267. **Wommack, K. E., J. Bhavsar, S. W. Polson, J. Chen, M. Dumas, S. Srinivasiah, M. Furman, S. Jamindar, and D. J. Nasko.** 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences., p. 427-439, *Stand. Genomic Sci.*, vol. 6.
268. **Wommack, K. E., and R. R. Colwell.** 2000. Virioplankton: Viruses in Aquatic Ecosystems, p. 69-114, *Microbiology and Molecular Biology Reviews*, vol. 64.
269. **Wommack, K. E., T. Sime-Ngando, D. M. Winget, S. Jamindar, and R. R. Helton.** 2010. Filtration-based methods for the collection of viral concentrates

from large water samples, p. 110-117, Manual of Aquatic Viral Ecology (MAVE).

270. **Wooley, J. C., A. Godzik, and I. Friedberg.** 2010. A primer on metagenomics, p. e1000667, PLoS Comput Biol, vol. 6. Public Library of Science.
271. **Wooley, J. C., and Y. Ye.** 2010. Metagenomics: facts and artifacts, and computational challenges, p. 71-81, Journal of computer science and technology, vol. 25. Springer.
272. **Xiang, X., L. Chen, X. Huang, Y. Luo, Q. She, and L. Huang.** 2005. *Sulfolobus tengchongensis* Spindle-Shaped Virus STSV1: Virus-Host Interactions and Genomic Features, p. 8677-8686, J. Virol., vol. 79.
273. **Xie, G., P. S. G. Chain, C. C. Lo, K. L. Liu, J. Gans, J. Merritt, and F. Qi.** 2010. Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing, p. 391-405, Molecular Oral Microbiology, vol. 25.
274. **Yilmaz, S., M. Allgaier, and P. Hugenholtz.** 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Meth 7:943-944.
275. **Zhang, K., A. C. Martiny, N. B. Reppas, K. W. Barry, J. Malek, S. W. Chisholm, and G. M. Church.** 2006. Sequencing genomes from single cells by polymerase cloning, p. 680-686, Nature Biotechnology, vol. 24.
276. **Zhang, Z., Y. Liu, S. Wang, D. Yang, Y. Cheng, J. Hu, J. Chen, Y. Mei, P. Shen, D. H. Bamford, and X. Chen.** 2012. Temperate membrane-containing halophilic archaeal virus SNJ1 has a circular dsDNA genome identical to that of plasmid pHH205, p. 1-9, Virology. Elsevier.
277. **Zillig, W., A. Kletzin, C. Schleper, I. Holz, D. Janekovic, J. Hain, M. Lanzendörfer, and J. K. Kristjansson.** 1994. Screening for Sulfolobales, their Plasmids and their Viruses in Icelandic Solfataras, p. 609-628, Syst. Appl. Microbiol., vol. 16. Gustav Fischer Verlag, Stuttgart · Jena · New York.
278. **Zillig, W., D. Prangishvilli, C. Schleper, M. Elferink, I. Holz, S. Albers, D. Janekovic, and D. Götz.** 1996. Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea., p. 225-236, FEMS Microbiology Reviews, vol. 18.

CHAPTER TWO

IDENTIFICATION OF NOVEL POSITIVE-STRAND RNA VIRUSES BY
METAGENOME ANALYSIS OF ARCHAEA-DOMINATED
YELLOWSTONE HOT SPRINGS

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Benjamin I. Bolduc

Contributions: Conceived the study, collected and analyzed output data, and wrote the manuscript.

Co-Author: Daniel P. Shaughnessy

Contributions: Collected and analyzed output data.

Co-Author: Yuri I. Wolf

Contributions: Collected and analyzed output data.

Co-Author: Eugene Koonin

Contributions: Collected and analyzed output data, discussed the results and implications, provided valuable insight and edited the manuscript during the final stages.

Co-Author: Francisco F. Roberto

Contributions: Provided valuable insight and edited the manuscript during the final stages.

Co-Author: Mark J. Young

Contributions: Obtained funding, assisted in conceiving the study, discussed the results and implications, provided valuable insight and edited the manuscript at all stages

Manuscript Information Page

Authors: Benjamin I. Bolduc, Daniel P. Shaughnessy, Yuri I. Wolf, Eugene Koonin, Francisco F. Roberto, Mark J. Young

Journal of Virology

Status of Manuscript:

- ☐ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-review journal
- ☐ Accepted by a peer-reviewed journal
- ☒ Published in a peer-reviewed journal

Publisher: American Society for Microbiology

Issue in which Manuscript Appears Here: Volume 86, Issue 10, 5562-5573 (2012)

CHAPTER TWO

IDENTIFICATION OF NOVEL POSITIVE-STRAND RNA VIRUSES BY
METAGENOME ANALYSIS OF ARCHAEA-DOMINATED
YELLOWSTONE HOT SPRINGS

Published: *Journal of Virology*, **86**, 10, 5562-73

Benjamin Bolduc^{1,4}, Daniel P. Shaughnessy^{1,3}, Yuri I Wolf⁵, Eugene Koonin⁵, Francisco F. Roberto⁶, and Mark Young^{1,2,3}

Thermal Biology Institute,¹ Depts. of Microbiology,² Plant Sciences and Plant Pathology,³ Chemistry and Biochemistry,⁴ Montana State University, Bozeman, MT 59717 USA, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA⁵, Idaho National Laboratory, Idaho Falls, ID 83415, USA⁶

Abstract

There are no known RNA viruses that infect Archaea. Filling this gap in our knowledge of viruses will enhance our understanding of the relationships between RNA viruses from the three domains of cellular life, and in particular could shed light on the origin of the enormous diversity of RNA viruses infecting eukaryotes. We describe here identification of novel RNA viral genome segments from high temperature acidic hot springs in Yellowstone National Park, USA. These hot springs harbor low complexity cellular communities dominated by several species of hyperthermophilic Archaea. A viral metagenomics approach was taken to assemble segments of these RNA virus genomes from viral populations directly isolated from hot spring samples. Analysis of these RNA metagenomes demonstrated unique gene content that is not generally related to known RNA viruses of Bacteria and Eukarya. However, genes for RNA-dependent RNA polymerase (RdRp), a hallmark of positive-strand RNA viruses, were identified in two contigs. One of these contigs is approximately 5,600 nucleotides in length and encodes a polyprotein that also contains a region homologous to the capsid protein of nodaviruses, tetraviruses and birnaviruses. Phylogenetic analyses of the RdRps encoded in these contigs indicate that the putative archaeal viruses form a unique group distinct from the RdRps of RNA viruses of Eukarya and Bacteria. Collectively, our findings suggest the existence of novel positive-strand RNA viruses that likely replicate in hyperthermophilic archaeal hosts and are highly divergent from RNA viruses that infect eukaryotes and even

more distant from known bacterial RNA viruses. These positive-strand RNA viruses might be direct ancestors of RNA viruses of eukaryotes.

Introduction

In contrast to viruses that infect Bacteria and Eukarya, little is known about the viruses that infect Archaea. Fewer than 50 archaeal viruses have been discovered compared to more than 5,500 viruses known to infect the hosts from the other two domains of life (2). Of the few archaeal viruses that have been characterized, mainly those infecting members of the phylum Crenarchaeota, many have revealed both unusual gene content (55) and unique virion morphologies (42, 53, 54, 66). To date, all archaeal viruses possess dsDNA genomes with the exception of one ssDNA virus infecting Archaea of the genus *Halorubrum* (60). No RNA viruses infecting Archaea have been described. Eukaryotic RNA viruses are a dominant viral form: they are numerous, diverse, and widespread, found to infect animals, plants, and many unicellular forms (17, 28, 38, 40, 71). Only two narrow groups of bacterial RNA viruses are known, and these do not show a close relationship with the viruses infecting eukaryotes (38, 40). Thus, the origin of the RNA viruses of eukaryotes remains an enigma. In this context, the search for RNA viruses infecting Archaea appears to be of special interest because the discovery of such viruses would have the potential to shed new light on the origin of viruses of eukaryotes.

One factor limiting the discovery of archaeal viruses has been the reliance on culture-dependent approaches for virus isolation. Many Archaea inhabit extreme environments, such as high temperature and high salinity conditions, making culture-

dependent approaches for virus discovery difficult. Over the past decade, direct sequencing of viral communities from environmental samples (viral metagenomics) has been used to more fully understand viral diversity in natural environments, including marine (13), sedimentary (10), human and animal feces (11, 14, 33), gut (23) and freshwater (59) environments. Viral metagenomics overcomes many of the limitations of traditional culture-based approaches for detection of new viruses. In general, viral metagenomics studies have revealed enormous diversity and abundance of viruses. For example, more than 5,000 different viral genotypes were identified in 200 liters of sea water (12). Although most viral metagenomics studies have focused primarily on dsDNA viruses (5, 73), recent advances in methodology have enabled the implementation of RNA viral metagenomics as an approach for investigating RNA viral communities. Studies in marine environments have revealed a diverse community of RNA viruses broadly distributed throughout multiple viral taxa (16, 17). In these metagenomic studies, RNA viruses infecting Bacteria or Archaea have not been identified, supporting the view that the major hosts for RNA viruses in the marine environment are unicellular eukaryotes (74). Metagenomic analyses of RNA viruses in fresh water have shown that the majority of sequences did not show significant similarity to known viruses in public databases, following the trend of marine environments. Those sequences that did match, were broadly distributed to nearly 30 families of viruses infecting a variety of eukaryotes but did not include the few groups of identified RNA viruses infecting Bacteria or putative RNA viruses of Archaea (19). The absence of evidence of RNA viruses infecting Archaea in both marine and freshwater environments (5, 13, 17, 19) appears surprising in

light of the indications that Archaea comprise up to 40% of the planet's microbial biomass and up to one-third of the ocean's prokaryotes (34).

The recently described Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) adaptive immunity system can be used to link viral genome sequences to cellular hosts. The CRISPR system appears to be an active defense mechanism against invading nucleic acids, including viruses, through a genetic interference pathway (7, 29, 44, 45, 69). The CRISPR system is present in ~40% and ~90% of sequenced bacterial and archaeal genomes, respectively (41). The CRISPR loci is composed of a leader sequence that directs transcription of a non-coding RNA that spans multiple copies of an ~25 nt direct repeat region (DR) that are separated by ~35 nt spacer sequences which can form long arrays. Each spacer sequence within a CRISPR array is unique. Immunity requires a sequence match between the invading nucleic acid and the spacer sequences that lie between DRs. The sequence content of the DRs can often be used assign a CRISPR type at the cellular family level (22). It has been shown that the CRISPR loci provide a record of a cell's interaction with viruses (45, 65). Accordingly, the CRISPR DR and spacer content present in an environmental sample can be used to connect the viruses present in that environment with bacterial or archaeal hosts.

The acidic hot springs in Yellowstone National Park USA (YNP) offer an opportunity to search for archaeal RNA viruses in an environment where Archaea predominate. Based on previous rRNA gene sequence analysis, these environments are inhabited by microbial communities of low complexity that are dominated by fewer than 10 archaeal species (30). In these low pH (pH<4), high temperature (>80°C) springs,

bacteria and eukaryotes are scarce or in many cases absent (9, 30, 57). Using traditional culture-dependent approaches, many dsDNA archaeal viruses have been isolated from these high temperature acidic hot springs in YNP (58) and other thermal hot springs worldwide (6, 48, 52). We used a viral metagenomic approach to directly search for archaeal RNA viruses in several acidic hot springs found in YNP. We report here the detection of putative archaeal RNA virus genomes.

Materials and Methods

YNP Hot Spring Sample Sites

Twenty eight YNP high temperature, acidic hot springs were initially screened for the presence of RNA in the viral enriched fraction (Table S1) between October and November 2008. In depth sampling for metagenomic analysis was performed on 3 of these sites; NL10, NL17 and NL18. A total of 7 paired DNA viral and RNA viral samples were collected. Nymph Lake hot springs NL 10 was collected in October 2009, February 2010 and June 2010. Nymph Lake hot springs NL17 and NL18 were collected in October 2009 and June 2010. Nymph Lake hot springs NL10 was also sampled in August 2008 and 2009 for total cellular DNA.

Initial Screening of Viral Enriched Samples for RNA Genomes by ³²P-labeling Experiments (RT-dependent Signal from RNA Viral Fraction)

A viral enriched fraction was created from each sample by filtration of 26 mL of hot spring water through two successive 0.8/0.2 µm Acrodisc 25 mm PF filters (Millipore) to remove cells, collection of the virus particles in the flow through,

centrifugation at 30,000 X g for 2 hrs and resuspension of the virus pellet into 200 μ L sterile water. Total viral nucleic acids were extracted from the virus enriched fraction using the mirVana miRNA isolation kit (Ambion). Following nucleic acid extraction, DNA was removed by DNase I (Ambion) treatment. Aliquots of RNA-enriched extracted viral nucleic acids were treated (controls) or not treated with RNase ONE (Promega) prior to cDNA synthesis. First-strand cDNA synthesis reactions were carried out using 1 mM random hexamers in the presence of 10 μ Ci of α -³²P dCTP (MP BioMedicals). Reactions were incubated at 25 C for 10 min, at 50 C for 90 min then terminated at 85 C for 5 min followed by RNaseH treatment. TCA precipitated nucleic acids were collected on glass filter fibers (Whatman G/F) and scintillation counts of precipitated and non-precipitated material were collected.

Isolation of Cellular Fraction for
Community Sequence Analysis (Fig. S1).

Cells were isolated from ~500-mL of hot spring water by filtration onto 0.45 μ m Pall filters (Millipore) and stored at -20 C. Total DNA was isolated from cells using phenol:chloroform:isoamyl alcohol 25:24:1 extraction of the filters followed by ethanol precipitation. DNA was amplified using GenomePlex Whole Genome Amplification (Sigma). Amplification products were purified using QIAquick PCR Purification kit (Qiagen). Approximately 10 μ g of cellular DNA was provided to the University of Illinois Champaign-Urbana Sequencing Center (Urbana, IL) and subjected to 454 Titanium pyrosequencing (Roche).

Isolation of Viral-enriched Nucleic Acids

~1.2 L of hot spring water was filtered through two successive Pall 0.8/0.2 μm filters (Millipore). The filtrate flow through was concentrated by Ultra centrifugal filters (Millipore) using a 100k MWCO to a volume of 500 μL . Total viral nucleic acids were extracted from 200 μL of each sample with the Purelink Viral RNA/DNA Mini Kit (Invitrogen) and eluted to a final volume of 60 μL .

Isolation, Preparation of RNA-enriched Viral Fraction and Sequencing

An improved method to isolate nucleic acids from the viral-enriched fraction was developed after initial RT-PCR screening. To obtain the RNA viral-enriched fractions, 30 μL of total extracted nucleic acid was treated with 3 U RNase-free DNase I (Ambion) at 37 C for 20 min, followed by addition of EtOH to 37%. This mixture was applied to the RNA/DNA Mini Kit columns (Purelink), eluted, and re-extracted following the manufacturer's protocols. The eluted nucleic acids were amplified with the Transplex Whole Transcriptome Amplification kit (Sigma). Reactions were purified using QIAquick PCR purification kits (Qiagen) followed by passage through Amicon Ultra-0.5 filter columns (100 kDa MWCO). Approximately 100 ng of purified, amplified viral cDNA products were provided to the Broad Institute (Cambridge, MA) for sequencing using the Titanium chemistry on the 454 FLX pyrosequencer (Roche).

Isolation, Preparation and Amplification of DNA-enriched Viral Fraction

To obtain the DNA-enriched viral fraction, 10 μ L of viral enriched total nucleic acid were amplified with the GenomePlex Whole Genome Amplification kit (Sigma). Reactions were purified as described above for the RNA-enriched viral fractions. Approximately 5 μ g of purified, amplified viral DNA was provided to the Broad Institute (Cambridge, MA) for sequencing as with the RNA-enriched viral fractions.

Sequence Assembly and Analysis

DNA sequences were trimmed of both primer and tag sequences. To aid in assembly, highly represented duplicated sequences were identified and removed using CD-HIT-454 (50) with a 40 bp overlap and 98% sequence identity. All 7 viral RNA datasets were then cross-assembled using gsAssembler version 2.5 (Roche) at 98% sequence identity with a 50 bp overlap. Cellular and DNA viral metagenomic data sets were assembled using the same procedures. The assembled RNA viral contigs (greater than 100 bp) were compared against the Nymph Lake viral DNA metagenomes as well as the GreenGenes rRNA dataset (18). Sequences matching either DNA viral reads or ribosomal sequences were excluded from future analysis. These filtered RNA viral contigs were subsequently analyzed against NCBI-nr and nt databases using BLASTN, BLASTX, TBLASTX, BLASTP, and PSI-BLAST for iterative search of protein sequence databases (3). The RPSTBLASTN program from the NCBI BLAST package was used to compare contigs against the conserved domain database. The resulting metagenomes were also analyzed using the MG-RAST metagenomics analysis server

(49) to assign taxonomies. Contigs showing significant matches (E-value 1×10^{-3}) to RdRps and structural proteins were subjected to a more rigorous examination using HHpred (68) against the pdb70 database (November 2010 release). Sequence assembly and analysis are outlined in Fig. S2.

Protein Structure Analysis

Secondary structure predictions for RdRp and CP sequences were performed using the HHpred software. Sequence to structure threading was performed using the Phyre server (35) and HHpred; MODELLER (21) was used to build the 3D model; visualization was rendered using Pymol (64). Multiple alignment of 3D structures of capsid proteins was extracted from the DALI database (27).

Phylogenetic Analysis of RdRps

Seed alignments of viral RdRp (cd01699) and group II intron reverse transcriptases (cd01651) were downloaded from the NCBI CDD database (47). These alignments were used to generate position-specific scoring matrices (PSSM) that served as queries for PSI-BLAST iterative search (4) against positive strand ssRNA virus sequences in the NCBI RefSeq database, producing 988 matches. The detected RdRp sequences were clustered into 29 groups using the NCBI BLASTCLUST program (<http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html>), roughly corresponding to various taxa of ssRNA viruses. Additionally, a cluster of 31 non-redundant sequences of Group II intron reverse transcriptases and a cluster with two RdRp sequences from Contig00002 and Contig00228 were added to the data set.

Multiple sequence alignments within clusters were produced using the MUSCLE program (20). Cluster alignments were progressively aligned using the HAlign program (67), at each step the highest-scoring pair of alignments replaced the two “parent” alignments; alignment positions containing less than 33% of non-gap characters were removed. For the purpose of phylogenetic analysis alignment positions containing less than 50% of non-gap characters were further eliminated from the final alignment. The phylogenetic tree of the full dataset (1021 sequences) was reconstructed using the FastTree program with default parameters (56). The optimal sequence evolution model for the dataset was selected using the ProtTest program (1). A representative dataset of 104 sequences was selected using the full tree as a guidance; phylogenetic tree was reconstructed using the RAxML program (70) with PROTGAMMALGF evolutionary model, as indicated by the ProtTest analysis. Bootstrap analysis with 100 replications and Shimodaira-Hasegawa log-likelihood test for alternative tree topologies derived from constraint trees was performed using RAxML.

Validation of Selected Assembled Contigs from RNA-enriched Viral Metagenomes by RT-PCR and DNA Sequencing

Up to two years after the original samples were collected for the production of the viral and cellular metagenomes, additional NL10, NL17 and NL18 cellular and viral fraction samples were collected in June 2011 and September 2011 as described above to determine if selected RNA viral genomes were still present in these hot springs. Viral nucleic acids were extracted using the ZR Viral RNA Kit (Zymo Research) in combination with the on-column DNase digestion, as recommended by the manufacturer.

Forward and reverse PCR primers designed from selected metagenomic contigs (Table S2) were used with SuperScript III One-Step RT-PCR with Platinum Taq (Invitrogen). To test for RT dependency, the RT was excluded from the procedure. To test for strand-specificity, only one primer was added during the cDNA synthesis stage and the 2nd primer added after the 94 C heat-kill of the reverse transcriptase and activation of Platinum Taq. PCR products were cloned into the pCR2.1 vector using the TopoTA cloning kit (Invitrogen). Sequencing of the products was performed using Sanger sequencing with BigDye terminator v3.1 on ABI 3100.

Analysis of CRISPR in Cellular and Viral Metagenomes

Reads from the cellular metagenomes were analyzed for CRISPR related sequences with the CRISPR Recognition Tool (CRT) (8) using the program's default parameters. Reads identified as containing CRISPRs using CRT were then analyzed using CRISPRFinder (25). All reads identified as CRISPRs with CRT were also identified as CRISPRs with CRISPRFinder. CRISPR-containing reads were parsed for direct repeats (DRs) and spacers generated from CRT. The resulting CRISPR spacers were compared against the viral RNA and DNA metagenomes using BLASTN with an ungapped alignment and 100% nucleotide identity across the entire length of the spacer. Reads with spacers matching to RNA contigs were further analyzed by extraction of their associated DRs and identification of their closest reference microbial genome.

Examining Viability of an Eukaryotic Virus in YNP Hot Springs

To test for the potential of possible external contamination of eukaryotic viruses into the examined hot springs, 10 ng (4800 genomes) of purified Cowpea chlorotic mottle virus (CCMV), a highly stable ssRNA plant virus, was mixed with 50-mL of hot spring water in a Falcon tube and placed in the hot springs. Aliquots of 50 μ L were taken from the samples at varying time intervals and analyzed with qRT-PCR using SSoFast EvaGreen Supermix (Life Sciences) on a RotorGene-Q system (Qiagen).

Comparing RNA Metagenomes to Other Environments

To assess whether similar RNA viral genomes were present in YNP high temperature circa neutral hot spring environments that are dominated by bacteria, viral RNA metagenomes were compared against transcriptome libraries of Mushroom and Octopus Springs YNP (36, 43) using BLASTN.

Results

An initial survey of 28 YNP hot springs was performed to determine which hot springs contained a detectable RNA signal in the isolated RNA viral fraction (Table S1). The surveyed sites were selected based on being high temperature ($>80^{\circ}$ C) and low pH (<4.0) springs likely to be dominated by Archaea. This initial screen of viral fractions was based on a RNA-dependent reverse-transcriptase (RT) assay where 32 P-dCTP incorporation into first strand cDNA was determined in RNase-treated controls and non-treated viral-enriched samples, both of which had previously been DNase-treated. While

24 sites showed little difference between incorporation of ^{32}P - dCTP of RNase treated and non-treated samples, 3 hot springs showed a reproducible and statistically significant difference (Figure 2.1). Resampling of these hot springs 12 months later resulted in nearly identical detection of the RNA signal in the RNA viral fraction suggesting that RNA viral community was being maintained in these hot springs. The three sites (NL10, NL17, and NL18) with the highest RT-dependent signal were selected for generating metagenomic libraries.

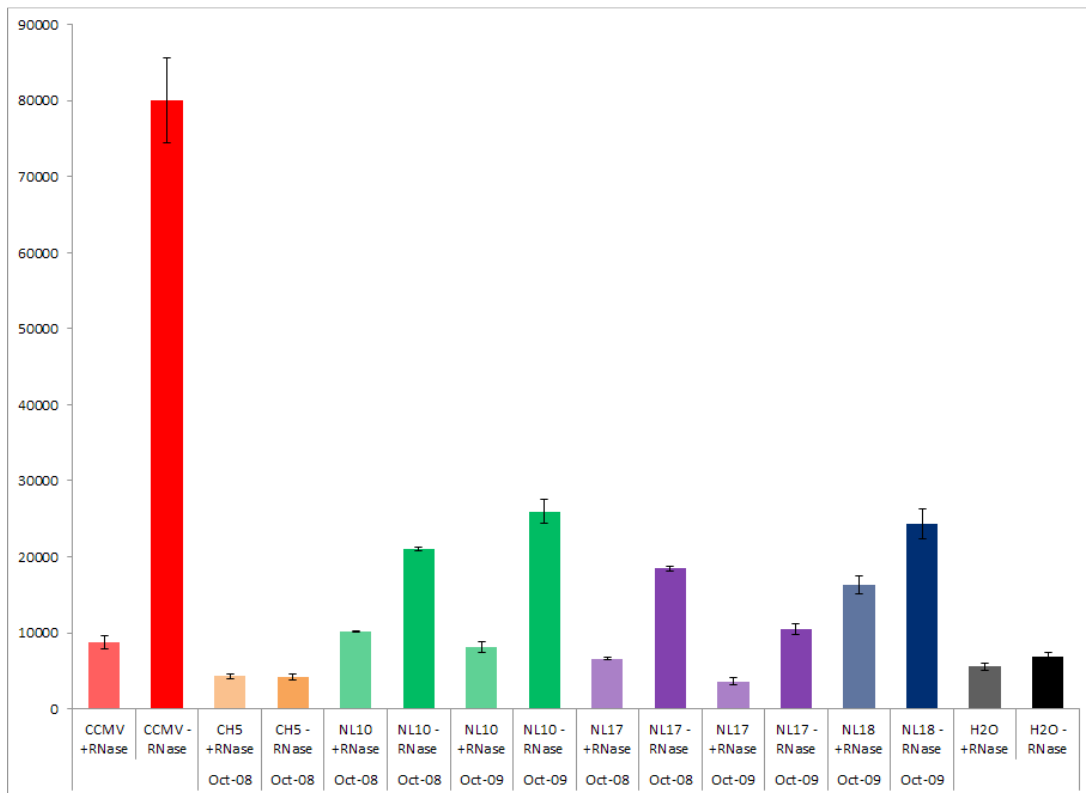


Figure 2.1. Selected examples of screening hot springs for presence of RNA templates in the RNA virus enriched fractions. ^{32}P dCTP incorporation into RT-dependent first strand cDNA synthesis is indicated. For each hot spring sample the control of prior RNase treatment of the sample is indicated (+RNase). Resampling of selected hot spring 12 months later demonstrated the maintenance of RNA signal in the viral fraction. Positive controls of a known +ssRNA virus (Cowpea chlorotic mottle virus (CCMV)) and water negative controls are indicated.

A stringent approach was taken to assemble the RNA viral and DNA viral metagenomes. Removal of highly duplicated sequence reads at 98% similarity with CD-HIT-454 resulted in a 41.6% reduction in reads across the viral metagenomes (Table 2.1). These duplicated sequences are likely indicators of both the amplification bias and the high sequencing depth of the relatively low complexity viral communities found in these hot springs. Removal of these highly represented sequencing reads significantly improved the overall assembly resulting in much higher confidence contigs.

Table 2.1. Cellular, DNA viral and RNA viral assembly statistics

	DNA Metagenomes	RNA Metagenomes		Cellular Metagenomes
Hot Spring Site	NL10, NL17, NL18	NL10, NL17, NL18		NL10
Date of Samples	Oct 2009, Feb & June 2010	Oct 2009, Feb & June 2010		Aug 2008 & 2009
		Pre-DNA Filtration	Post-DNA filtration	
Total Reads	1075407	356886	90227	632479
Total Bases (MB)	34.0	4.2	1.1	229
Reads Assembled	72.90%	34.1%	34.1%	64.6%
Largest Contig	21541	5846	5846	23333
Large Contigs (> 1000 bp)	3411	519	79	4820
Total Contigs (> 100 bp)	27746	9696	2636	22649
Average Contig Length	573.9	434.3	409.5	716.8
Average Coverage	37.8	36.8	34.2	22.7

The initial analysis of the viral RNA metagenomes indicated that they contained a high proportion of sequences found in the paired viral DNA metagenomes. This finding suggests that the initial treatment of the isolated viral nucleic acid with DNase was insufficient to remove 100% of the viral DNA sequences likely due to its vast excess compared to RNA in the viral fraction. To address this bias towards DNA viral sequences, RNA contigs were compared against the DNA viral metagenomes using BLASTN. Matches of RNA contigs (E-value 1×10^{-30}) against reads from the DNA metagenomes were excluded from further analysis. More than 72% of the contigs in the initial RNA metagenomic data sets were removed as a result of matching sequences in the viral DNA datasets (7,060 of the initial 9,696 contigs). This screening procedure, while removing the majority of contigs, provided confidence that the remaining contigs assembled from the RNA viral metagenomes were derived from RNA viruses present in that hot spring and not derived from viral DNA sequences.

The assembly of sequences from the Nymph Lake hot spring sites resulted in a large number of contigs under stringent assembly conditions (98% identity within at least 50 bp overlap) (Table 2.1). The 14 viral samples (7 RNA enriched viral samples and 7 DNA enriched viral samples) generated ~2.8 million reads and 807 Mb of sequence. After de-duplication, there were 877k reads and 590 Mb of unique sequence. Assembled viral DNA metagenomes generated 27,746 total contigs, with an average large contig size (>1000 bp) of 1,898 bp. The assembled viral RNA metagenomes generated 9,696 total contigs, with an average large contig length of 1,361 bp. Nearly 73% of the sequence reads in the DNA viral metagenomes were either fully or partially assembled into contigs

whereas 36.8% of sequence reads in the RNA viral metagenomes assembled into contigs. The assembled viral DNA metagenomic contigs had an average length of 574 bp with an average read depth of 38-fold. The final assembled viral RNA metagenomes had an average contig length of 410 bp with an average read depth of 34-fold base coverage.

Cellular DNA metagenomes, comprising 632k sequencing reads which represented 229 Mb of sequence were assembled under the same high stringency parameters as the viral datasets (Table 2.1). The NL0808 and NL0908 cellular assemblies generated 22,649 contigs with an average contig length of 716 bp and an average read depth of 36-fold base coverage. Sixty five percent of the reads assembled into contigs.

The analysis of the assembled cellular metagenomes revealed that 81.0% of the contigs to have a significant match against the server's protein databases (Figure 2.2A). The majority of the sequences (57%) matched to Archaea, primarily to the Crenarchaeota (86.9%) and to a lesser degree, the Euryarchaeota (11.6%). Matches to Nanoarchaeota, Korarchaeota and Thaumarchaeota were also detected. A smaller portion of the cellular contigs matched to Bacteria (20.6%), mostly to delta- and gamma-proteobacteria and firmicutes (clostridia and bacilli). However, overall these matches to the Bacteria were weak compared to the matches to the Archaea. This could reflect shared genes between Archaea and Bacteria and/or a statistical bias towards bacterial sequences due to their overrepresentation as compared to archaeal sequences in the public databases. In support of this possibility, further analysis of the cellular metagenomes using the ribosomal databases available on the MG-RAST server (Ribosomal Database Project, SILVA rRNA Database Project and Greengenes) revealed 4 matches against the Crenarchaeota

(*Thermoprotei*), one match to the *Nanoarchaea* and a single distal match to *Sulfurihydrogenibium yellowstonense*, a bacterium of the phylum Aquificales originally isolated from a YNP hot spring. Overall, these results confirm that the YNP hot springs analyzed here are dominated by Archaea.

Examination of the putative viral RNA contigs showed that the majority of sequences (56%) did not align to known sequences (Figure 2.2C). A small fraction of the sequences matched known viruses (3%). The remaining sequences matched sequences from Archaea (20%), Bacteria (10%), and Eukaryota (11%).

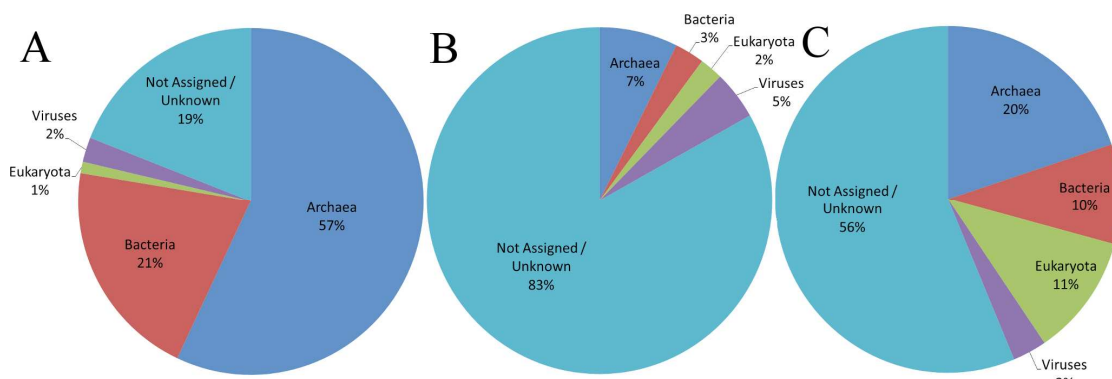


Figure 2.2. Hierarchical classification of contigs with MG-RAST. Classification of (A) cellular, (B) viral DNA and (C) viral RNA sequences based on the M5NR database in MG-RAST. Sequences with insufficient significance or those that did not match were classified as “Not Assigned / Unknown.”

Analysis of the RNA viral contigs using NCBI’s protein and nucleotide non-redundant databases, the conserved domain database (CDD) as well as Uniprot/Swissprot (31) detected several sequences with similarity to RNA-dependent RNA polymerases (RdRp), a ss (+) RNA virus “hallmark gene” (39). The largest contig in the RNA viral dataset (5.6 kb) as well as smaller contigs with similarity to RdRps were confirmed to be

present and persistent in the hot springs over time. Resampling of the NL 10 and NL 18 RNA viral fraction and extraction of the total RNA out of the cellular fraction 18 months after the last sampling for the metagenomic analysis showed the continued presence of the RNA genome segments. An RT-PCR assay with primers designed from these RNA contigs was performed on samples taken from the same hot springs over an 18 month period (Figure 2.3). These assays demonstrated that sequences were single-stranded because amplifiable product was generated only when cDNA synthesis used a specific, directional primer. The longest RT-dependent Contig00002 is 5662 nt in length, composed of 258 reads and contains a single, large ORF that encodes a putative viral polyprotein encompassing an RdRp and a putative capsid protein as detailed below (the sequence of Contig00002 was deposited in GenBank with the Accession Number). Omission of the RT, template or pre-treatment of samples with RNase prior to the RT-PCR assay eliminated the signal, indicating that the source of the signal was an RNA template in the viral fraction (Figure 2.3). This analysis supported the original contig assemblies and demonstrated that these putative RNA viruses are stable members of the viral community within the explored hot springs.

Further search for RdRps, as well as viral structural (capsid) proteins, was performed using a combination of BLAST, MG-RAST and HHpred. Two contigs were identified as containing significant similarity to known RdRps. A similar search examining the DNA enriched viral metagenome dataset revealed no similarity to RdRps but did show many hits to capsid proteins, mainly those of archaeal DNA viruses, as expected.

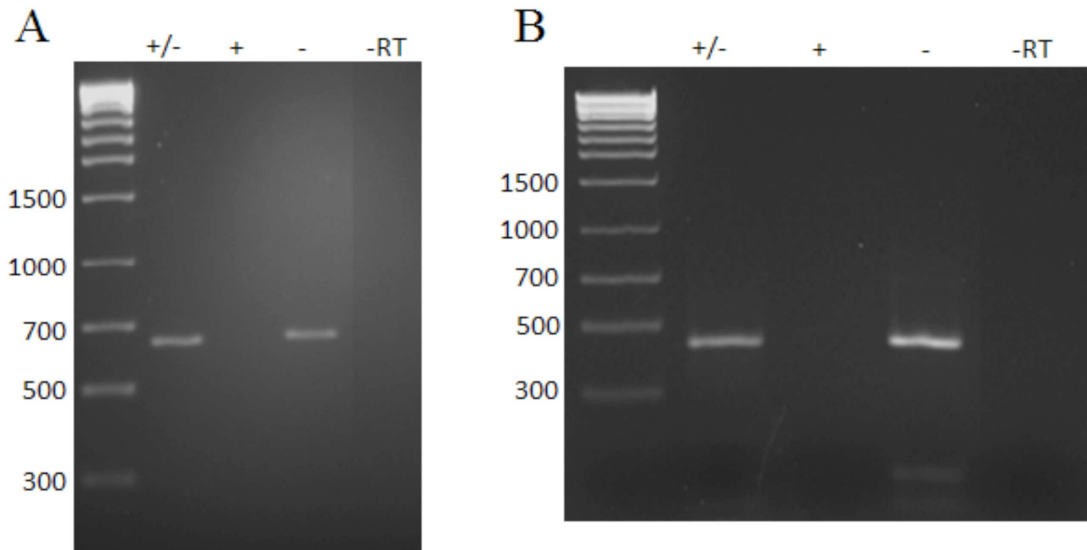


Figure 2.3. Detection and single stranded nature of RNA viral sequences within hot springs months after sampling for metagenomic analysis. Total nucleic acids were extracted from either the viral fraction or total cellular fraction of NL 10 (B) and NL 18 (A) eighteen months after the original sampling for metagenomic analysis, DNase-treated and subjected to RT-PCR. The detection and strand-specific nature of Contig00009 (A) and Contig00002 (B) were determined using contig specific primer sets (Table S2). Either “forward (+),” “reverse (-)” or both (+/-) primers were added to the initial first-strand cDNA synthesis mix and incubated as described in the Materials & Methods. After first strand synthesis, the complete primer pair was added (if necessary) during for the PCR stage. Controls of minus reverse-transcriptase (-RT) and MW marker (bp) are indicated.

Contig00002 contains a single long ORF potentially coding for a 198 kDa (1809 amino acids) polyprotein that spans almost the entire length of the contig, suggesting that a complete or nearly complete RNA viral genome was assembled. In the middle part of this polyprotein (approximately between amino acids 800-1020), a putative RdRp domain was identified using several search methods. A search of the Conserved Domain Database at the NCBI using the RPS-BLAST program detected a highly significant similarity (E-value of approximately 6×10^{-11}) to the RdRp domain profile (no significant similarity to any other domain was detected for this or other parts of the polyprotein). A

BLASTP search of the non-redundant protein database at the NCBI yielded statistically significant similarity (E-value 2×10^{-4} , 23% identity in an alignment of 341 amino acid residues) to the RdRp of Pariacato nodavirus and limited, not significant (E-value of approximately 0.1) similarity to the RdRps of several other RNA viruses including tombusviruses and pestiviruses. The second iteration of the PSI-BLAST search resulted in highly significant similarity to the RdRps of a variety of positive-strand RNA viruses of eukaryotes. This putative RdRp sequence contains all three highly conserved motifs, A (D-x(4,5)-D) and B (S/T)G-x(3)-T-x(4)-N(S/T), which are involved in nucleotide selection as well as metal binding (24) and motif C which contains the highly conserved GDD motif and is an essential part of the Mg^{2+} binding site (32, 37) (Figure 2.4A). Modeling of the inferred protein RdRp onto the X-ray structure of the (+) ss Norwalk virus RdRp structure showed that the putative archaeal virus RdRp contained the principal structural elements of the Palm domain of the RdRps of eukaryotic positive-strand RNA viruses (Figure 2.4B). Matches to RdRps were also detected for Contig00228 (comprising 10 reads) which encompassed three overlapping short ORFs each of which showed approximately 70% amino acid sequence identity to the predicted RdRp of Contig00002 (Figure 2.4A). Thus, this contig appears to encode an RdRp of a putative archaeal virus that is related to but clearly distinct from the putative virus encoded by Contig00002.

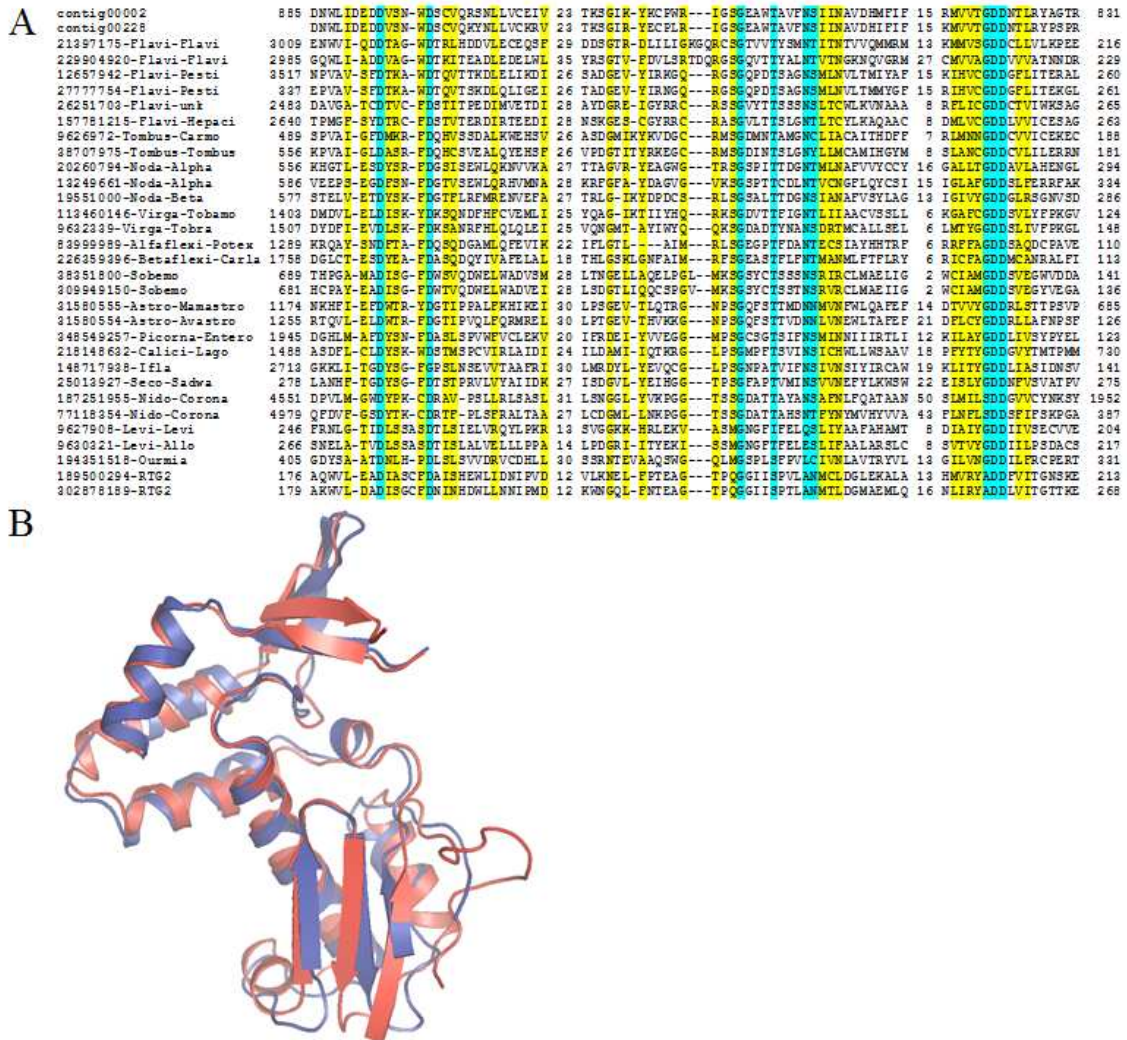


Figure 2.4. The RdRp of the putative archaeal viruses. (A) Multiple sequence alignment of the putative archaeal virus RdRps with homologs from positive-strand RNA viruses of eukaryotes and bacteria, and reverse transcriptases from bacterial Group 2 introns. The (nearly) universally conserved amino acid residues in the three signature motifs of the RdRps are highlighted in cyan and partially conserved residues are highlighted in yellow. The top two sequences are from the putative archaeal viruses identified in this work. The rest of the sequences include representatives of the 29 clusters of viral RdRps identified as described under Methods. The two sequences at the bottom (RTG2) are reverse transcriptases. Each sequence is denoted by the GenBank identifier (GI number) and the name of the virus group (typically, family) and subgroup (typically genus). The numbers denote the lengths (number of amino acids) in less well conserved regions between the conserved motifs and the distances between the ends of the respective proteins and the aligned segments. (B) Structural model of the central core region of the putative archaeal virus RdRp from Contig00002 using the calicivirus RdRp as a template (PDB 3bso (75)).

A comparison of the Contig00002 polyprotein sequence to databases of protein family profiles using the HHPred program supported the finding of highly significant similarity to viral RdRps and additionally led to the detection of sequences that were similar to capsid proteins of positive-strand eukaryotic RNA viruses within the nodavirus family. This sequence of highest similarity to capsid proteins is approximately 90 amino acids in length and is located C-terminal to the RdRp, spanning amino acid residues 1562-1652 of the polyprotein. The alignment between this portion of the polyprotein and the capsid proteins of eukaryotic viruses demonstrated a level of similarity that was not statistically significant although the alignments included approximately 30% identical amino acid residues. Nevertheless, a more detailed, direct comparison of the Contig00002 polyprotein sequence to the nodavirus capsid protein sequences as well as the related capsid protein sequences of tetraviruses and nodaviruses allowed us to extend the alignment and identify the main structural elements of the capsid proteins (Figure 2.5). The nodavirus capsid protein adopts an 8-strand jelly roll fold that is distantly related to the structures of the capsid proteins of other small icosahedral positive-strand RNA viruses. In addition, unlike other well-characterized viral capsid proteins, nodaviruses possess an autoproteolytic activity that is involved in processing of the capsid protein precursor (the α protein) into the mature large (β) and small (γ) capsid proteins (61, 63). The nodavirus capsid protein is the founding member of the A6 peptidase superfamily that employs an unusual catalytic mechanism involving an interaction between the active aspartate of the capsid protein precursor and a conserved asparagine that is proximal to the scissile peptide bond in the C-terminal part of the protein, at the boundary between β

and γ capsid proteins (76). These residues, which are essential for autocatalytic cleavage of the nodavirus capsid protein precursor, are also conserved in the sequences of capsid proteins of tetraviruses. Counterparts to both the catalytic aspartate and the cleavage site asparagine were detected in the Contig00002 polyprotein, and regions surrounding these two conserved amino acids were found to have the highest levels of sequence similarity (Figure 2.5). Exhaustive analysis of the parts of the Contig00002 polyprotein outside the RdRp and capsid protein regions failed to detect similarity to any other known proteins or domains. Collectively, the observation of a large polyprotein (a common feature of eukaryotic positive-strand RNA viruses) containing putative RdRp and capsid proteins in addition to a possible autoproteolytic activity suggest that this contig corresponds to a near full-length genome of a previously unknown positive-strand RNA virus.

Phylogenetic analysis of the identified RdRp sequences showed that they formed a lineage distinct from known RdRps of Eukaryotic and Bacterial RNA viruses (Figure 2.6). The putative RdRps encoded by Contig00002 and Contig00228 did not show specific affinity with any of the three superfamilies of eukaryote positive-strand RNA viruses (picorna-like, alpha-like, and flavi-like), or the only known bacterial lineage (leviviruses), and statistical tests showed that association with any of these major lineages or additional distinct lineages such as nodaviruses could not be convincingly ruled out (Supplemental Material). These results are compatible with a 'central' position of the new RdRp in the tree and hence with its possible ancestral status.

Figure 2.5. The predicted capsid protein of the putative archaeal virus and its potential autoproteolytic activity. Multiple alignment of the putative archaeal virus capsid protein with the capsid protein sequences of nodaviruses, tetraviruses and birnaviruses. Sequences are identified either by PDB ID or by GenBank GI numbers. Secondary structure is shown for the three available crystal structures denoted by the corresponding PDB codes (l: loop; H: helix; E: beta-strand). Alignment columns with homogeneity of 0.4 or greater (see Methods) are highlighted in yellow. The catalytic Asp is highlighted in cyan; the cleavage site Asn is highlighted in green.

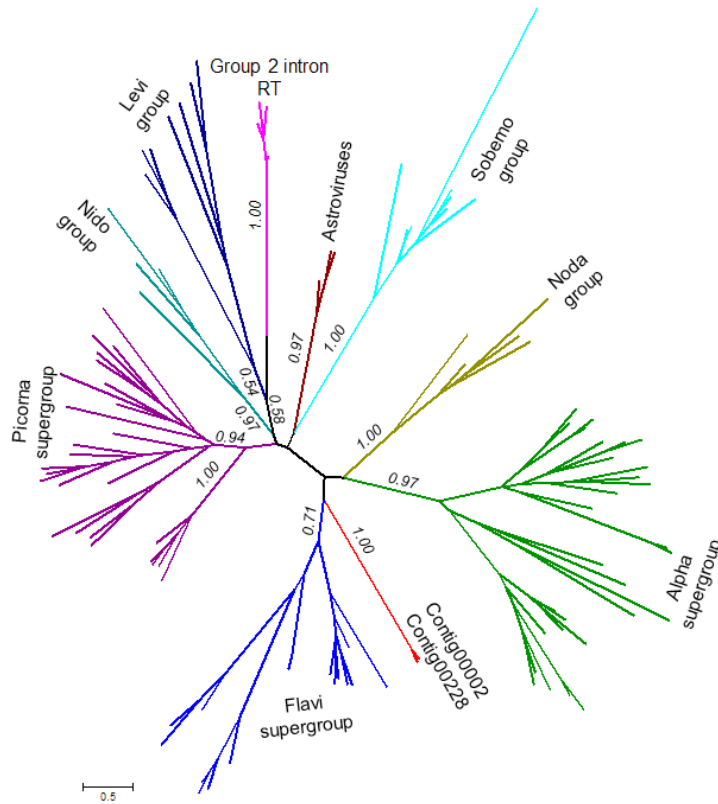


Figure 2.6. Phylogenetic tree of the RdRps. The unrooted phylogenetic tree was generated as described under Methods. Bootstrap support values greater than 0.5 are indicated for deep internal branches. Large groups of positive-strand RNA viruses from eukaryotes and bacteria (levi group) and the bacterial retro-transcribing elements (RT) are indicated and shown by unique colors.

To investigate the possibility that the detected RNA virus genomes were an introduced contaminant of the hot springs from external sources, we tested the ability of a RNA plant virus known for its high stability to be maintained within the hot spring environment. Samples of the CCMV were mixed with hot spring water in 50 ml Falcon tubes and placed back into the hot springs. Aliquots were taken at regular intervals for up to 30 min. and quantified using a qRT-PCR assay which can detect as few as 10 viral genome copies. Within 1 min of sampling, there was no detectable amount of the CCMV RNA (data not shown). This experiment indicates that it is unlikely that an externally

introduced RNA virus survives long enough to be detected, especially in the form of long RNA segments, under the high temperature acidic conditions of the hot springs from which the putative archaeal virus sequences were obtained.

To further probe the possibility that the detected RNA viral genomes replicate in archaeal hosts, these sequences were compared to the sequences produced by environmental transcriptomic analysis of two high temperature YNP hot springs at neutral to alkaline pHs (Mushroom and Octopus Springs) known to be dominated by thermophilic bacteria. These hot springs harbor a relatively high complexity thermophilic bacterial community that consists of chloroflexi, cyanobacteria, acidobacteria and chlorobi (36). A BLASTN analysis found that the great majority of the contigs (97% of the total RNA viral contigs) did not have statistically significant matches to the metatranscriptome of these neutral to alkaline YNP hot springs. Three percent of the RNA viral contigs did show a significant match. However, the average length of the matching contigs was 342 bp, much shorter than the overall average contig length. This analysis indicates that there is little overlap between the RNA viral community present in the Archaea-dominated acidic hot springs and the bacteria- dominated neutral alkaline hot springs, and provides further evidence that the detected putative viral RNA genomes are associated with archaeal hosts.

Finally, we analyzed the CRISPR direct repeats (DR) and spacer content present in our cellular metagenomic datasets in an attempt to link the putative RNA viral genomes directly to a specific host type and to investigate potential host immunity to RNA viruses. CRISPR DRs and spacers were extracted from the cellular metagenomic

data sets. The spacer sequences (10349) were compared with the assembled viral RNA and DNA metagenomes. Forty six spacers (0.44%), associated with 4 types of DRs, were identical to RNA sequences within the viral metagenomic dataset (Figure 2.7). A similar comparison of the spacers to the DNA viral metagenome revealed 628 (6.07%) identical spacers (data not shown). The majority of matching spacer sequences of the RNA metagenome (44/46) were related to DRs of the archaeal species *S. islandicus* and *S. acidocaldarius*. These were the only significant matches found in the CRISPRfinder direct repeat database. These findings suggests that the RNA viral genomes replicate in an archaeal host belonging to the Sulfolobales, a cellular type commonly found in NL 10 and acidic hot springs worldwide, and elicit CRISPR-mediated immune response. However, we cannot categorically rule out the unlikely possibility that the CRISPR spacer matches to the RNA viral metagenome are in fact to DNA viral sequences that are still present in our viral RNA metagenomic dataset.

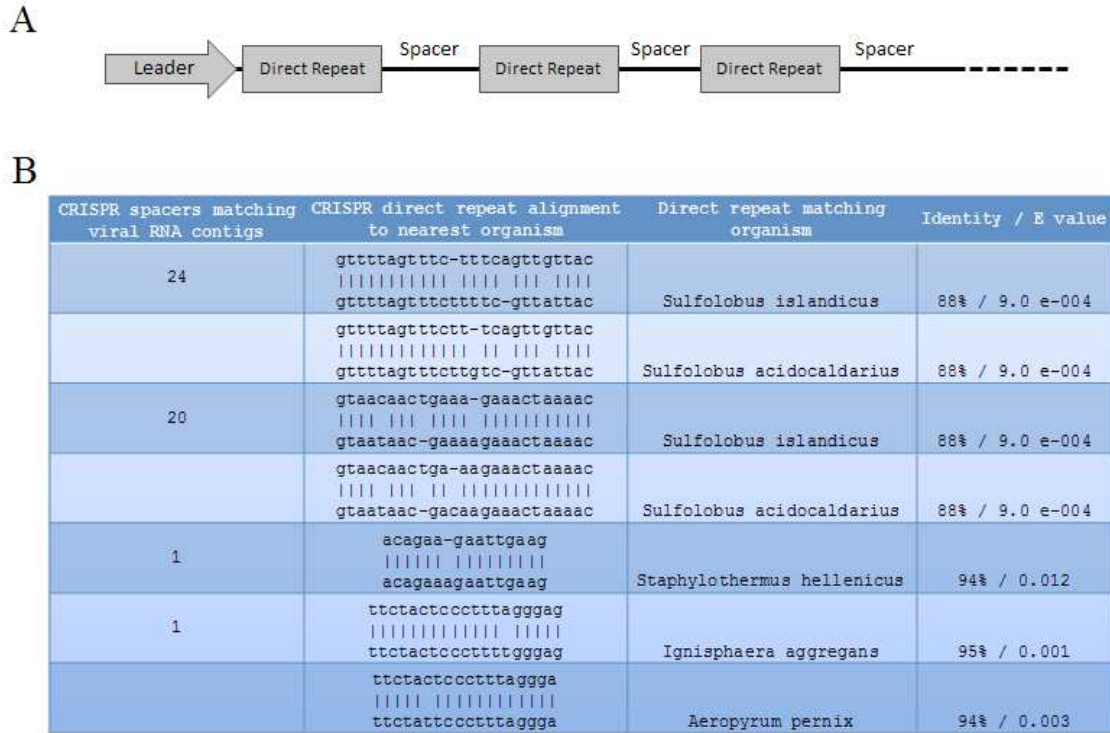


Figure 2.7. Analysis of cellular CRISPR direct repeat units and number of unique spacers matching archaeal species and RNA viral metagenome contigs. (A) Schematic representation of the CRISPR loci indicating the direct repeat structure interspaced with spacer regions. (B) The number of CRISPR spacer sequences with 100% match to RNA viral contigs and the alignment of the direct repeat structures (top sequence) with the closest sequenced organism (bottom sequence). Percentage of identity and E values are indicated.

Discussion

To our knowledge, this work is the first attempt on viral RNA community analysis in extreme environments. The results demonstrate the presence of putative RNA viral genomes in high temperature acidic hot springs dominated by Archaea. Several lines of experimental and comparative genomic evidence support the conclusion that the assembled genome segments belong to RNA viral genomes. In particular, two contigs encoded protein sequences with a specific, significant similarity to the RdRp, a hallmark

protein of positive-strand RNA viruses. One of these, the 5.6 kb Contig00002, might comprise a (nearly) complete viral genome with signature features of positive-strand RNA viruses of eukaryotes. Specifically, this contig encodes a polyprotein that is similar in size to the polyproteins of diverse eukaryote positive-strand RNA viruses that contain both the RdRp and capsid protein domains. Polyproteins of positive-strand RNA viruses of eukaryotes, in particular viruses of the picorna-like superfamily, are processed by viral proteases to yield individual functional proteins. A distinct protease domain was not detected in the virus polyprotein encoded in Contig00002. However, the catalytic site and the cleavage site of the nodavirus capsid autoprotease were conserved in the region of the putative archaeal virus polyprotein that is similar to the viral capsid proteins (Figure 2.5). In nodaviruses, the capsid protein precursor is encoded in a distinct segment of genomic RNA such that the protease is involved only in the autocatalytic maturation of the capsid proteins (15, 62, 76). In the novel virus described here, this protease might perform more versatile roles by catalyzing additional cleavage events and releasing the RdRp, the capsid protein and possibly other mature viral proteins that remain to be identified.

The results of this study do not unequivocally demonstrate the existence of archaeal RNA viruses. It cannot be ruled out that the detected RNA genome segments derive from viruses replicating in bacterial hosts present in these hot springs. However, several lines of independent experimental and comparative genomic evidence suggest that the identified RNA genome segments originate likely from viruses replicating in archaeal hosts. First, analysis of both the overall composition of the cellular metagenomes by MG-RAST and the 16S rDNA content shows that the microbiota of the

sampled hot springs consists primarily of Archaea. Second, the putative RNA viral sequences were recovered from multiple hot springs at multiple time points. Third, experimental exposure of a plant RNA virus that is known for high stability to the acidic hot springs environment shows that the virus is rapidly degraded under these conditions. Fourth, when we compared the transcriptome of a YNP neutral alkaline hot springs that are known to be dominated by bacteria rather than Archaea, there was no overlap with the RNA viral sequences detected in the YNP acidic hot springs. Taken together, this experimental evidence suggests that any RNA virus sequence that is consistently recovered from the hot springs replicates in archaeal cells.

This evidence is complemented by the sequence analysis results which clearly indicate that the novel virus detected in this study 1) is a positive-strand RNA virus related to positive-strand RNA viruses of eukaryotes, and 2) it does not belong to any known virus family. This conclusion is supported both by the topology of the phylogenetic tree of the RdRps (Figure 2.6) and by the unique arrangement of protein domains in the polyprotein. It should be further noted that the only known family of bacterial positive-strand RNA viruses, the Leviviridae, is a group of limited diversity that does not show evolutionary affinity to any particular groups of viruses infecting eukaryotes and might not even share a common origin with eukaryotic viruses (40). Thus, should the novel virus described here derive from bacteria, its genome organization would be completely unprecedented among genetic elements far isolated from bacterial hosts.

Finally, when we compared the transcriptome of YNP neutral alkaline hot springs that are known to be dominated by bacterial and not archaeal communities, there was no overlap with the RNA viral sequences detected in the YNP acidic hot springs. This supports the notion that the RNA viral communities present in archaeal dominated hot springs is distinct from the viral communities present in bacterial dominated hot spring environments.

Interestingly, we observed multiple matches between archaeal CRISPR spacers and the RNA metagenomes isolated from the hot springs. The interpretation of this observation is complicated by the fact that none of these matches were to the contigs containing sequences homologous to RNA viruses of eukaryotes. Nevertheless, the identification of these spacers might indicate that archaeal RNA viruses elicit CRISPR-mediated immunity, a possibility that is of particular interest given that at least some archaeal CRISPR systems have been shown to target RNA, in contrast to bacterial CRISPRs that appear to only target DNA (26, 72).

The definitive demonstration of the existence of archaeal RNA viruses awaits isolation of viral particles capable of infecting archaeal hosts and producing infectious progeny. However, assuming that the preliminary indications presented here are born out, the implications for the origin and evolution of positive-strand RNA viruses of eukaryotes will be fundamental. At present, the prokaryotic ancestry of eukaryotic RNA viruses remains unclear. The leviviruses, the only known group of positive-strand RNA viruses of prokaryotes (bacteria), do not seem to be direct ancestors of the viruses of eukaryotes, and the closest homolog of the latter in prokaryotes appears to be the RT of

bacterial retro-transcribing elements (40). In contrast, the putative RNA virus of Archaea identified here could be related to the direct ancestors of eukaryotic viruses as indicated by the homology of the RdRps and the capsid proteins. Moreover, it is notable that the strongest similarity for both of these proteins was detected with the respective proteins of nodaviruses, a family of viruses of eukaryotes that is only distantly related to other positive-strand RNA viruses and is among the simplest known groups of viruses with respect to genome architecture and the protein repertoire (40). The nodavirus family seems to be a good candidate for the ancestral group of eukaryotic RNA viruses that might have evolved directly from the archaeal progenitors. Notably, the putative direct ancestral relationship between RNA viruses of Archaea and eukaryotes mimics similar relationships that have been recently described for some of the key functional systems of the eukaryotic cells such as the ubiquitin network (51) or the membrane remodeling and cell division apparatus (46). From the methodological standpoint, this study demonstrates that viral metagenomic approaches have the potential to yield information that is essential to ultimately isolate and characterize novel RNA viruses and their cellular hosts.

Acknowledgements

This work was supported by National Science Foundation grant numbers DEB-0936178 and EF-080220 and National Aeronautics and Space Administration grant number NNA-08CN85A. YIW and EVK are supported by the Department of Health and Human Services intramural program (NIH, National Library of Medicine). We thank

Mary Bateson, Jamie Snyder, Martin Lawrence, Nikki Dellas, and Brian Bother for critical reading of this manuscript.

Literature Cited

1. **Abascal, F., R. Zardoya, and D. Posada.** 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**:2104-2105.
2. **Ackermann, H. W.** 2007. 5500 Phages examined in the electron microscope. *Arch Virol* **152**:227-243.
3. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic Local Alignment Search Tool. *J Mol Biol* **215**:403-410.
4. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**:3389-3402.
5. **Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. *Plos Biol* **4**:2121-2131.
6. **Arnold, H. P., U. Ziese, and W. Zillig.** 2000. SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon *Sulfolobus*. *Virology* **272**:409-416.
7. **Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath.** 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**:1709-1712.
8. **Bland, C., T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, and P. Hugenholtz.** 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**:209.
9. **Blank, C. E., S. L. Cady, and N. R. Pace.** 2002. Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. *Appl Environ Microbiol* **68**:5123-5135.
10. **Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2004. Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**:565-574.
11. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2003. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**:6220-6223.

12. **Breitbart, M., and F. Rohwer.** 2005. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**:278-284.
13. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**:14250-14255.
14. **Cann, A. J., S. E. Fandrich, and S. Heaphy.** 2005. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**:151-156.
15. **Cheng, R. H., V. S. Reddy, N. H. Olson, A. J. Fisher, T. S. Baker, and J. E. Johnson.** 1994. Functional implications of quasi-equivalence in a T = 3 icosahedral animal virus established by cryo-electron microscopy and X-ray crystallography. *Structure* **2**:271-282.
16. **Culley, A. I., A. S. Lang, and C. A. Suttle.** 2003. High diversity of unknown picorna-like viruses in the sea. *Nature* **424**:1054-1057.
17. **Culley, A. I., A. S. Lang, and C. A. Suttle.** 2006. Metagenomic analysis of coastal RNA virus communities. *Science* **312**:1795-1798.
18. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* **72**:5069-5072.
19. **Djikeng, A., R. Kuzmickas, N. G. Anderson, and D. J. Spiro.** 2009. Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**:e7264.
20. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792-1797.
21. **Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali.** 2006. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* **Chapter 5**:Unit 5 6.
22. **Garrett, R. A., S. A. Shah, G. Vestergaard, L. Deng, S. Gudbergdottir, C. S. Kenchappa, S. Erdmann, and Q. She.** 2011. CRISPR-based immune systems of the Sulfolobales: complexity and diversity. *Biochem Soc Trans* **39**:51-57.
23. **Gill, S. R., M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson.** 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312**:1355-1359.

24. **Gohara, D. W., S. Crotty, J. J. Arnold, J. D. Yoder, R. Andino, and C. E. Cameron.** 2000. Poliovirus RNA-dependent RNA polymerase (3D(pol)) - Structural, biochemical, and biological analysis of conserved structural motifs A and B. *J Biol Chem* **275**:25523-25532.
25. **Grissa, I., G. Vergnaud, and C. Pourcel.** 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**:W52-57.
26. **Hale, C. R., P. Zhao, S. Olson, M. O. Duff, B. R. Graveley, L. Wells, R. M. Terns, and M. P. Terns.** 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**:945-956.
27. **Holm, L., and P. Rosenstrom.** 2010. Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**:W545-549.
28. **Holmes, E. C.** 2009. RNA virus genomics: a world of possibilities. *J Clin Invest* **119**:2488-2495.
29. **Horvath, P., D. A. Romero, A. C. Coute-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, and R. Barrangou.** 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**:1401-1412.
30. **Inskip, W. P., D. B. Rusch, Z. J. Jay, M. J. Herrgard, M. A. Kozubal, T. H. Richardson, R. E. Macur, N. Hamamura, R. Jennings, B. W. Fouke, A. L. Reysenbach, F. Roberto, M. Young, A. Schwartz, E. S. Boyd, J. H. Badger, E. J. Mathur, A. C. Ortmann, M. Bateson, G. Geesey, and M. Frazier.** 2010. Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5**:e9773.
31. **Jain, E., A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B. E. Suzek, M. J. Martin, P. McGarvey, and E. Gasteiger.** 2009. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**.
32. **Kamer, G., and P. Argos.** 1984. Primary Structural Comparison of Rna-Dependent Polymerases from Plant, Animal and Bacterial-Viruses. *Nucleic Acids Res* **12**:7269-7282.
33. **Kapoor, A., J. Victoria, P. Simmonds, C. Wang, R. W. Shafer, R. Nims, O. Nielsen, and E. Delwart.** 2008. A highly divergent picornavirus in a marine mammal. *J Virol* **82**:311-320.
34. **Karner, M. B., E. F. DeLong, and D. M. Karl.** 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**:507-510.

35. **Kelley, L. A., and M. J. Sternberg.** 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4**:363-371.
36. **Klatt, C. G., J. M. Wood, D. B. Rusch, M. M. Bateson, N. Hamamura, J. F. Heidelberg, A. R. Grossman, D. Bhaya, F. M. Cohan, M. Kuhl, D. A. Bryant, and D. M. Ward.** 2011. Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J* **5**:1262-1278.
37. **Koonin, E. V., V. P. Boyko, and V. V. Dolja.** 1991. Small Cysteine-Rich Proteins of Different Groups of Plant Rna Viruses Are Related to Different Families of Nucleic Acid-Binding Proteins. *Virology* **181**:395-398.
38. **Koonin, E. V., and V. V. Dolja.** 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit Rev Biochem Mol Biol* **28**:375-430.
39. **Koonin, E. V., T. G. Senkevich, and V. V. Dolja.** 2006. The ancient Virus World and evolution of cells. *Biol Direct* **1**:29.
40. **Koonin, E. V., Y. I. Wolf, K. Nagasaki, and V. V. Dolja.** 2008. The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat Rev Microbiol* **6**:925-939.
41. **Kunin, V., R. Sorek, and P. Hugenholtz.** 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**:R61.
42. **Lawrence, C. M., S. Menon, B. J. Eilers, B. Bothner, R. Khayat, T. Douglas, and M. J. Young.** 2009. Structural and Functional Studies of Archaeal Viruses. *J Biol Chem* **284**:12599-12603.
43. **Liu, Z., C. G. Klatt, J. M. Wood, D. B. Rusch, M. Ludwig, N. Wittekindt, L. P. Tomsho, S. C. Schuster, D. M. Ward, and D. A. Bryant.** 2011. Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *ISME J* **5**:1279-1290.
44. **Makarova, K. S., N. V. Grishin, S. A. Shabalina, Y. I. Wolf, and E. V. Koonin.** 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**:7.
45. **Makarova, K. S., D. H. Haft, R. Barrangou, S. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost, and E. V. Koonin.** 2011. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**:467-477.

46. **Makarova, K. S., N. Yutin, S. D. Bell, and E. V. Koonin.** 2010. Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat Rev Microbiol* **8**:731-741.
47. **Marchler-Bauer, A., S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant.** 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**:D225-229.
48. **Martin, A., S. Yeats, D. Janekovic, W. D. Reiter, W. Aicher, and W. Zillig.** 1984. Sav-1, a Temperate Uv-Inducible DNA Virus-Like Particle from the Archaeobacterium *Sulfolobus-Acidocaldarius* Isolate B-12. *Embo Journal* **3**:2165-2168.
49. **Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards.** 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**.
50. **Niu, B., L. Fu, S. Sun, and W. Li.** 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11**:187.
51. **Nunoura, T., Y. Takaki, J. Kakuta, S. Nishi, J. Sugahara, H. Kazama, G. J. Chee, M. Hattori, A. Kanai, H. Atomi, K. Takai, and H. Takami.** 2011. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39**:3204-3223.
52. **Prangishvili, D., H. P. Arnold, D. Gotz, U. Ziese, I. Holz, J. K. Kristjansson, and W. Zillig.** 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.
53. **Prangishvili, D., and R. A. Garrett.** 2004. Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. *Biochem Soc Trans* **32**:204-208.
54. **Prangishvili, D., and R. A. Garrett.** 2005. Viruses of hyperthermophilic Crenarchaea. *Trends Microbiol* **13**:535-542.

55. **Prangishvili, D., R. A. Garrett, and E. V. Koonin.** 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res* **117**:52-67.
56. **Price, M. N., P. S. Dehal, and A. P. Arkin.** 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**:1641-1650.
57. **Reysenbach, A. L., G. S. Wickham, and N. R. Pace.** 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **60**:2113-2119.
58. **Rice, G., K. Stedman, J. Snyder, B. Wiedenheft, D. Willits, S. Brumfield, T. McDermott, and M. J. Young.** 2001. Viruses from extreme thermal environments. *Proc Natl Acad Sci U S A* **98**:13341-13345.
59. **Rodriguez-Brito, B., L. L. Li, L. Wegley, M. Furlan, F. Angly, M. Breitbart, J. Buchanan, C. Desnues, E. Dinsdale, R. Edwards, B. Felts, M. Haynes, H. Liu, D. Lipson, J. Mahaffy, A. B. Martin-Cuadrado, A. Mira, J. Nulton, L. Pasic, S. Rayhawk, J. Rodriguez-Mueller, F. Rodriguez-Valera, P. Salamon, S. Srinagesh, T. F. Thingstad, T. Tran, R. V. Thurber, D. Willner, M. Youle, and F. Rohwer.** 2010. Viral and microbial community dynamics in four aquatic environments. *Isme Journal* **4**:739-751.
60. **Roine, E., M. K. Pietila, L. Paulin, N. Kalkkinen, and D. H. Bamford.** 2009. An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. *Mol Microbiol* **72**:307-319.
61. **Schneemann, A., T. M. Gallagher, and R. R. Rueckert.** 1994. Reconstitution of Flock House provirions: a model system for studying structure and assembly. *J Virol* **68**:4547-4556.
62. **Schneemann, A., and D. Marshall.** 1998. Specific encapsidation of nodavirus RNAs is mediated through the C terminus of capsid precursor protein alpha. *J Virol* **72**:8738-8746.
63. **Schneemann, A., W. Zhong, T. M. Gallagher, and R. R. Rueckert.** 1992. Maturation cleavage required for infectivity of a nodavirus. *J Virol* **66**:6728-6734.
64. **Schrodinger, LLC.** 2010. The PyMOL Molecular Graphics System, Version 1.3r1.
65. **Snyder, J. C., M. M. Bateson, M. Lavin, and M. J. Young.** 2010. Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. *Appl Environ Microbiol* **76**:7251-7258.

66. **Snyder, J. C., K. Stedman, G. Rice, B. Wiedenheft, J. Spuhler, and M. J. Young.** 2003. Viruses of hyperthermophilic Archaea. *Res Microbiol* **154**:474-482.
67. **Soding, J.** 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**:951-960.
68. **Soding, J., A. Biegert, and A. N. Lupas.** 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**:W244-W248.
69. **Sorek, R., V. Kunin, and P. Hugenholtz.** 2008. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**:181-186.
70. **Stamatakis, A., P. Hoover, and J. Rougemont.** 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* **57**:758-771.
71. **Suttle, C. A.** 2005. Viruses in the sea. *Nature* **437**:356-361.
72. **Terns, M. P., and R. M. Terns.** 2011. CRISPR-based adaptive immune systems. *Curr Opin Microbiol* **14**:321-327.
73. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**:66-74.
74. **Weinbauer, M. G.** 2004. Ecology of prokaryotic viruses. *Fems Microbiology Reviews* **28**:127-181.
75. **Zamyatkin, D. F., F. Parra, J. M. Alonso, D. A. Harki, B. R. Peterson, P. Grochulski, and K. K. Ng.** 2008. Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. *J Biol Chem* **283**:7705-7712.
76. **Zlotnick, A., V. S. Reddy, R. Dasgupta, A. Schneemann, W. J. Ray, Jr., R. R. Rueckert, and J. E. Johnson.** 1994. Capsid assembly in a family of animal viruses primes an autoproteolytic maturation that depends on a single aspartic acid residue. *J Biol Chem* **269**:13680-13684.

CHAPTER THREE

VIRAL COMMUNITY COMPOSITION IN YELLOWSTONE
ACIDIC HOT SPRINGS ASSESSED BY
NETWORK ANALYSIS

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Benjamin I. Bolduc

Contributions: Conceived the study, collected and analyzed output data, and wrote the manuscript.

Co-Author: Jennifer Wirth

Contributions: Collected data and edited the manuscript during the final stages.

Co-Author: Aurélien Mazurie

Contributions: Provided valuable insight and edited the manuscript during the final stages.

Co-Author: Mark J. Young

Contributions: Obtained funding, discussed the results and implications, provided valuable insight and edited the manuscript at all stages.

Manuscript Information Page

Authors: Benjamin I. Bolduc, Jennifer Wirth, Aurélien Mazurie , Mark J. Young

ISME Journal

Status of Manuscript:

- ☒ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-review journal
- ☐ Accepted by a peer-reviewed journal
- ☐ Published in a peer-reviewed journal

CHAPTER THREE

VIRAL COMMUNITY COMPOSITION IN YELLOWSTONE

ACIDIC HOT SPRINGS ASSESSED BY

NETWORK ANALYSIS

Prepared

Benjamin Bolduc^{1,4}, Jennifer Wirth^{1,3}, Aurélien Mazurie, and Mark Young^{1,2,3}

Thermal Biology Institute,¹ Depts. of Microbiology,² Plant Sciences and Plant Pathology,³ Chemistry and Biochemistry,⁴ Montana State University, Bozeman, MT 59717 USA

Abstract

Our understanding of viral community structure in natural environments remains a daunting task. Total viral community sequencing (e.g. viral metagenomics) provides a tractable approach. However, even with the availability of next-generation sequencing technology it is only possible to obtain a fragmented view of viral communities in natural ecosystems. In this study we applied a network-based approach in combination with viral metagenomics to investigate viral community structure in the high temperature, acidic hot springs of Yellowstone National Park, USA. Our results show that this approach can identify distinct viral populations and provide insights into the viral community stability using metagenomic sequence data. We identified 82 viral clusters in the hot springs environment, with each network cluster likely representing a viral family at the taxonomic level. Most of these viral clusters are previously unknown DNA viruses infecting archaeal hosts. Overall, this study demonstrates the utility of combining viral community sequencing approaches with network analysis to gain insights into viral community structure in natural ecosystems.

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 (7, 11, 26, 28). 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 (25) with more than 1.2×10^5 different genotypes (2). Like with 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 the ecology and evolution in natural environments. As a step towards this long-term goal, there is an immediate need to improve our ability to define viral communities structures.

Despite the increasing number of environments examined, few studies have addressed the composition and interconnectivity of these viral communities as a whole. A variety of bioinformatic tools have been developed to analyze the structure and/or diversity of viral communities from metagenomics datasets (reviewed extensively in (10, 17)), including PHAACS (1) and Metavir (23). While these tools provide taxonomic classification, functional assignment and estimates of virus community structure and diversity, they typically only provide a snapshot of a specific time and place, are dependent on sequence comparisons to sequences of known function, and do not account

well for the large variation in sequences and functions typically associated with viral families.

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 (12). High temperatures ($>80^{\circ}\text{C}$), low pH ($\text{pH}<3$) conditions generally favor *Archaea*-dominated communities (4, 6, 22). These conditions offer a tractable system to study virus-host relationships as well as viral and host community structure and community structure stability.

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

Materials and Methods

Sample Sites

A YNP high-temperature ($72\text{--}93^{\circ}\text{C}$), acidic ($\text{pH } 2.0\text{--}4.5$) hot spring was selected for this study based on previous work (6). The Nymph Lake hot spring site 10 (NL10: 44.7536°N , 110.7237°W) was sampled at 9 different time points over a 21 -month period (Table 3.1). Nymph Lake 10 is located in a relatively new thermal basin that has developed over the past 15 years (15).

Table 3.1. YNP Sampling Points and Characteristics

Date	Metagenome	Amplification Method	Temperature	pH
Sep-08	NL10 0809	MDA	90	4
Jan-09	NL10 0901		92	4.3
Mar-09	NL10 0903		84	5.1
Apr-09	NL10 0904		92	4.5
Aug-09	NL10 0908		90	4
Sep-09	NL10 0909		92	4
Oct-09	NL10 0910	WGA	83	3
Feb-10	NL10 1002		87	3.1
Jun-10	NL10 1006		89	4.5

Sample Collection

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 four hours and immediately centrifuged at 100,000 $\times$ g for 2 hours. 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 liters of hot spring water was filtered as above. The filtrate was then concentrated with Ultra centrifugal filters using a 100,000 molecular weight cutoff to a volume of 500 μL . 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) and

eluted to a final volume of 60 μ L. Extracted viral nucleic acids were amplified prior to sequencing by Multiple Displacement Amplification (MDA; New England Biolabs, Ipswich, MA, USA) or Whole Genome Amplification (WGA; Sigma-Aldrich, St Louis, MO, USA). All viral samples from Aug-08 to Sept-09 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.

Processing 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 custom python script that used a sliding quality window with average read quality of 25 across 50 bp. Preliminary analysis revealed a wide range of sequencing coverage within the metagenomes. Such uneven sequencing coverage ranges often prevents optimal assembly (14). To aid assembly, highly duplicated sequencing reads were identified using CD-HIT-454 (20) 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. Since 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 dataset prior to assembly. These reads were removed by aligning sequences using Newbler gsMapper version 2.7 at 70% identity over 50 bp.

Assembly of Viral Metagenomes

Remaining sequencing reads from each time point were assembled using the 454 gsAssembler software program version 2.7 with default parameters, except for the minimum identity between sequences (98%) and the minimum overlap (50 bp; sequencing and assembly statistics are presented in Table 3.2). These stringent assembly conditions were used to prevent mis-assembly between reads of different viruses (8). An additional assembly was generated that cross-assembled the filtered reads (briefly described in (6) of the entire DNA dataset (hereafter referred to as cross assemblies). This cross assembly was used for analysis by the Virome pipeline (described below).

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>) (3), Circonspect (version 0.2.6; <http://sourceforge.net/projects/circonspect/>) and PHACCS (version 1.1.3; <http://sourceforaget.net/projects/phaccs/>) (1). GAAS, which estimates the average size of genomes in communities, was unable to find enough similarities when compared against the NCBI reference viral genomes with a percent similarity above 50 percent 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. Community structure for individual, mixed (i.e.

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 were uploaded to the VIROME web server and analyzed as previously described (27).

Network Analysis of Viral Assemblies

Contigs from each metagenome set were compared through an “all-vs.-all” BLAST comparison using MegaBLAST with the default parameters (except: max target sequences = 1000, e-value cutoff 10^{-3}). The BLAST file was parsed using a custom python script that 1) evaluated all hit-scoring pairs (HSPs) using the following criteria: a minimum query alignment length of 50 bp and a minimum identity of 70%, 2) created a directed graph in *.graphml format, 3) applied the Louvain method to detect distinct viral populations (i.e. clusters of related contigs) within a large community network (5), and 4) applied a user-defined cutoff for the minimum number of members (i.e. contigs) in a viral cluster/population. For all datasets, the viral clusters’ membership cutoff was selected based on maximizing the percentage of data used while minimizing more rare, inactive populations (Supplemental Figure 1). For a more detailed explanation of how the viral network is created, see additional methods in the supplemental material. In order to allow better visualization, a viral population was defined as a network cluster containing ≥ 100 contigs. Networks were visualized within Gephi (16) using the OpenOrd layout and further optimized using the Force Atlas 2 plug-in. Network statistics and topological properties of each network were calculated using Gephi. As controls, genomes of 67

known archaeal DNA viruses were seeded into the viral metagenomes to test if they associated with individual network clusters.

The stability of the viral network was performed using a custom python script that analyzed each member of every community, referenced which metagenome the member was derived from and summed the contribution from each time point per community.

Reassembly of Viral Contigs Based on Network Clusters

Due to the high clustering coefficient of the network and the number of viral reference sequence matches exclusive to each viral cluster, it was believed that each cluster could be reduced to a group of “representative” pan-genomes. Contigs within each community (defined through the network analysis, above) were re-assembled using the Minimo assembler (version 1.6, <http://sourceforge.net/projects/amos/>), based on the AMOS infrastructure (v. 3.1.0) with a 35-bp minimum contig overlap length and 98% minimum overlap identity. Resulting assemblies were parsed using a python script to determine the total number of new contigs, their average length, as well as largest and smallest contigs.

Table 3.2. Sequencing and Assembly Statistics

Metagenome (Site Date)	# Reads	Mbp	Read Length	Post-Reduction Reads	Reduction	Reads Assembled (%)	# Contigs	Contig Length*	# Lg Contigs (>1 kb)	Lg Contig Length*
NL10 0809	201,434	68.44	409	174461	13.4%	92.3%	4235	600	650	1779
NL10 0901	168,300	58.58	414	149119	11.4%	90.2%	4916	569	685	1758
NL10 0903	212,474	73.61	408	191571	9.8%	89.0%	6705	548	848	1766
NL10 0904	205,811	71.11	406	184185	10.5%	90.1%	5507	516	601	1848
NL10 0908	166,526	52.56	402	137707	17.3%	87.2%	5908	549	757	1698
NL10 0909	204,818	69.94	401	182757	10.8%	88.3%	7447	541	924	1716
NL10 0910	123,820	32.34	347	98413	20.5%	82.1%	6086	528	685	1681
NL10 1002	154,473	39.07	316	132787	14.0%	68.0%	5757	526	647	1603
NL10 1006	176,955	40.75	318	146042	17.5%	49.0%	4174	707	837	1857
NL10 Meta- Assembly	1,614,611	506.39	369	1397042	13.9%	86.6%	27608	562	3515	1929

Minimum Sampling to Describe Viral Community Through Network Analysis

Contigs from each metagenome were added in stepwise, time-forward fashion, adding contigs to the each previous sampling point (beginning with only NL10 0901 contigs) and performing a network analysis (described above) before the addition of the next time point. This continued until all sampling points were used, and then the same process was repeated but reversed “direction”, starting with NL10 1006 and moving backwards in time. This resulted in 18 additional networks that were parsed as above while network statistics and topological properties of each network were calculated using Gephi.

Results

Read Processing and Assembly

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 (Supplemental Figure 1). Removal of highly duplicated reads at 98% identity with CD-HIT-454 reduced the number of reads by 14% (on average) for the viral metagenomes (Table 2). 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 (506.4 Mbp) to ~1.40 million reads (430.5 Mbp) for the viral metagenomes.

Sequential assembly strategies improved the number and length of assembled contigs. The initial assembly of sequences from NL10 hot springs resulted in a large

number of viral contigs under stringent conditions. Assembly of the 9 individual viral metagenomes generated 50 735 total contigs (average large contig length of 1.74-kb), assembling 81.8% of the reads, with 16.1% singletons (the remaining classified as repeats, chimeric or too short). A cross-assembly of each individual metagenome's reads produced fewer overall contigs (27 608), increased the average large contig length (1.9-kb) and increased the number of reads assembling into contigs (86.6% with 10.2% singletons) compared to its individual counterparts. To further support the concept that network clusters (described below) are representative of common and/or related viral types, each cluster defined by our network analysis was re-assembled using similar assembly parameters (Table 3). In nearly all cases, the re-assembly resulted in the creation of fewer, larger contigs. For example, cluster 43 contains 504 contigs, representing 1.01% of the total community abundance. Re-assembly reduced the number of contigs by 53.2% (236 contigs) and increased the average contig length by 27%. 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 due to insufficient sequencing depth at individual time points, isolation or amplification biases or stochastic effects during sequencing. Overall, these cluster re-assemblies resulted in a 55% decrease in the number of contigs, a 27% increase in average contig size and a 23% increase in the largest contig size.

Table 3.3. DNA metavirome re-assemblies and CHAS-BLAST based on network analysis

Name	Pre-Assembly			Post-Assembly			Contig Reduction	Average Contig Increase	CHAS BLAST
	Contigs	Max	Average	Contigs	Max	Average			
Population 1	1811	4894	519	917	5023	621	49.4%	20%	X
Population 2	1542	4912	521	657	4913	669	57.4%	28%	X
Population 3	1291	3731	428	660	3731	530	48.9%	24%	X
Population 4	1127	7555	804	396	11637	1092	64.9%	36%	X
Population 5	1059	8265	617	573	8285	709	45.9%	15%	X
Population 6	1038	4563	506	520	5558	613	49.9%	21%	X
Population 7	1035	8571	663	363	14334	845	64.9%	27%	X
Population 8	1012	4564	425	352	4599	547	65.2%	29%	X
Population 9	972	4168	490	441	4168	576	54.6%	18%	X
Population 10	964	11111	845	390	20353	1006	59.5%	19%	X
Population 11	919	5135	514	366	5135	642	60.2%	25%	X
Population 12	911	3146	471	421	3146	599	53.8%	27%	X
Population 13	910	7671	577	521	10325	691	42.7%	20%	X
Population 14	896	4051	507	299	4052	716	66.6%	41%	X
Population 15	858	5050	548	305	5050	714	64.5%	30%	X

Population 16	834	3950	492	326	4061	690	60.9%	40%	X
Population 17	813	3855	448	335	4025	559	58.8%	25%	X
Population 18	810	4134	540	313	4475	722	61.4%	34%	X
Population 19	799	11240	644	296	11245	864	63.0%	34%	X
Population 20	762	2958	450	342	3659	595	55.1%	32%	X
Population 21	758	3299	458	440	3595	535	42.0%	17%	X
Population 22	755	3386	428	616	5044	475	18.4%	11%	X
Population 23	736	5680	504	195	6826	708	73.5%	40%	X
Population 24	715	4733	465	342	5372	571	52.2%	23%	X
Population 25	713	3768	407	557	3769	458	21.9%	13%	X
Population 26	713	2312	342	328	2319	448	54.0%	31%	X
Population 27	701	4089	513	258	4092	621	63.2%	21%	X
Population 28	696	4642	541	256	4642	710	63.2%	31%	X
Population 29	693	6179	468	232	6197	525	66.5%	12%	X
Population 30	688	7683	437	463	7683	517	32.7%	18%	X
Population 31	644	5640	487	342	6133	588	46.9%	21%	X
Population 32	595	6190	430	181	6649	544	69.6%	27%	X
Population 33	586	4205	623	189	4419	802	67.7%	29%	X
Population 34	585	2043	361	222	2043	515	62.1%	43%	X
Population 35	575	12346	649	263	12950	635	54.3%	-2%	X

Population 36	537	4895	674	167	4893	916	68.9%	36%	X
Population 37	536	8820	872	164	8829	1034	69.4%	19%	X
Population 38	530	5483	534	186	5483	702	64.9%	31%	X
Population 39	529	4438	401	178	5888	523	66.4%	30%	X
Population 40	527	3817	500	313	4250	563	40.6%	13%	X
Population 41	527	2710	553	177	5219	729	66.4%	32%	X
Population 42	513	3430	446	330	3513	514	35.7%	15%	X
Population 43	504	2838	491	236	3312	626	53.2%	27%	X
Population 44	484	6571	376	185	6748	533	61.8%	42%	X
Population 45	478	1719	309	194	1719	434	59.4%	40%	X
Population 46	445	1960	423	326	3161	481	26.7%	14%	X
Population 47	418	3230	508	251	4454	638	40.0%	26%	X
Population 48	414	2852	436	256	3169	542	38.2%	24%	X
Population 49	398	2461	488	183	2530	583	54.0%	19%	X
Population 50	388	2020	408	154	2045	551	60.3%	35%	X
Population 51	378	7206	478	114	7205	556	69.8%	16%	X
Population 52	346	6320	703	130	6320	883	62.4%	26%	X
Population 53	337	3583	383	131	4743	521	61.1%	36%	X
Population 54	335	3313	403	124	3314	564	63.0%	40%	X
Population 55	319	4178	635	148	6029	825	53.6%	30%	X

Population 56	319	4295	664	182	6432	710	42.9%	7%	X
Population 57	314	2265	429	139	2265	556	55.7%	30%	X
Population 58	313	3901	549	120	3901	604	61.7%	10%	X
Population 59	295	8598	798	98	9889	950	66.8%	19%	X
Population 60	288	8981	947	84	23069	1312	70.8%	39%	X
Population 61	269	9508	593	74	9508	660	72.5%	11%	X
Population 62	246	2304	390	118	2446	507	52.0%	30%	X
Population 63	234	2115	527	55	3255	746	76.5%	42%	X
Population 64	229	14464	1024	81	16177	1017	64.6%	-1%	X
Population 65	225	6119	521	177	6119	596	21.3%	14%	X
Population 66	213	9437	582	146	9624	664	31.5%	14%	
Population 67	201	3124	608	79	4264	756	60.7%	24%	X
Population 68	197	4150	615	57	5951	799	71.1%	30%	X
Population 69	178	11316	1241	64	13823	1520	64.0%	22%	X
Population 70	174	2495	470	121	2978	548	30.5%	17%	X
Population 71	168	3319	662	74	4317	912	56.0%	38%	X
Population 72	145	2235	367	56	2235	471	61.4%	28%	X
Population 73	143	1774	497	116	2634	545	18.9%	10%	X
Population 74	143	4060	717	66	4659	946	53.8%	32%	X
Population 75	133	3110	550	80	3764	612	39.8%	11%	X

Population 76	130	2111	439	89	2334	523	31.5%	19%	
Population 77	129	4652	777	46	5809	1329	64.3%	71%	X
Population 78	124	5120	846	42	7008	1340	66.1%	58%	X
Population 79	119	1710	472	93	1710	529	21.8%	12%	
Population 80	111	1450	333	93	1465	355	16.2%	7%	X
Population 81	108	6958	1331	17	21111	2971	84.3%	123%	
Population 82	107	6212	1320	23	11443	2443	78.5%	85%	X

Taxonomic Analysis

To reduce the computational cost associated with analyzing 9 metaviromes, only the cross-assembled metagenomes were taxonomically classified using the VIROME pipeline. Analysis of contigs > 300 bp revealed more than half of the cross-assembled metaviromes could not be assigned. Assignable viral sequences with hits to viruses were nearly exclusively archaeal, with 97.8% of ORFs matching to archaeal viruses. The remaining 2.2% were to temperate phage of the order *Caudovirales*. Archaeal virus homologs matched to families known to infect the hyperthermophilic members of the Archaea; spread between the *Rudiviridae* (32%) which includes *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 homologs were spread evenly between the *Fuselloviridae*, *Bicaudaviridae* and a number of unclassified viruses including *Sulfolobus* turreted icosahedral virus 1 and 2 (STIV-1, 2) and hyperthermophilic archaeal virus 1 and 2. It is not clear if the taxonomic assignment of the phage ORFs found in the viral fractions is a consequence of cellular nucleic acids present in the viral fraction or truly 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 (Table 3.4). The

metaviromes were relatively low diversity and genotype. The number of estimated genotypes varied between 192 to 1 342, although the different estimators separates the samples into a low richness and moderate-low richness with ranges of 192 to 494, and 1145 to 1 342 respectively. The most abundant genotypes were inversely related to the number of genotypes, ranging from 3-7% for the low richness and 2-4% for the moderate-low richness NL10 groups. The community evenness was stable, between 0.87 and 0.94. Cross-contig spectra from PHACCS showed 129 genotypes common to all time points and high evenness (0.94) with a Shannon-Wiener index showing a low diversity of 4.5.

Table 3.4. Viral Community Structure Predicted from Assembly of Metagenomic Sequences

Metagenome	Scenario	Error (fit)	Richness (Genotypes)	Evenness	Most abundant genotype (%)	Shannon- Wiener index
NL10 0809 DNA	lognormal	3935.9	192	0.88	7.31%	4.62 nats
NL10 0901 DNA	lognormal	2680.7	234	0.89	6.13%	4.85 nats
NL10 0903 DNA	lognormal	747.03	334	0.89	4.96%	5.19 nats
NL10 0904 DNA	lognormal	908.76	226	0.91	4.94%	4.94 nats
NL10 0908 DNA	lognormal	861.64	376	0.89	4.91%	5.27 nats
NL10 0909 DNA	power	1678.6	267	0.93	5.68%	5.17 nats
NL10 0910 DNA	power	180.57	494	0.94	3.43%	5.86 nats
NL10 1002 DNA	power	794.13	1145	0.91	4.02%	6.41 nats
NL10 1006 DNA	power	26.554	1342	0.95	2.15%	6.82 nats
NL DNA Cross	lognormal	2393.7	129	0.94	4.52%	4.59 nats

Network Analysis

Due to the temporal nature of the samples, it was hypothesized that a network-based approach would reveal patterns of viral community composition and dynamics. To

this end, MegaBLAST was performed across the assembled metagenomes in an all-vs.-all comparison. Roughly 96% of all contigs (48 894 out of 50 735) were found to have a significant match of 70% identity across a 50 bp minimum length (E-value cutoff of 10^{-15} of the largest contig in each pairwise relationship) to another contig in the metagenome. To estimate the number of distinct populations 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 (13) and peer-to-peer networks (24). A total of 384 viral populations (i.e. an individual cluster of sequences distinctly separated from other clusters) containing ≥ 2 contig members were detected (for the purposes of this analysis, remaining single contigs were not considered). A population-based rank abundance curve is shown in Figure 3.1. The most dominant population for the viral networks represents 3.82% of the total community (excluding singletons). This agrees well with the numbers estimated by PHACCS, considering each network cluster can represent multiple species simultaneously (below). The network shows a large number of low-population clusters, with 244 populations in the viral network containing fewer than 10 members (1 985 including singletons), suggesting insufficient sequencing. The remaining clusters with ≥ 10 members in the viral network consist of 140 populations (98.1% of non-singleton contigs). Even with a minimum membership of ≥ 100 contigs, more than 94.4% of the viral contigs are represented in the community, strongly suggesting the largest populations represent the dominant viral community members. For this work we define the major viral populations as those clusters within the community containing ≥ 100 contig members. In order to reduce

network complexity and aid in visualization, only those passing these criteria are included (Figure 3.2). Under this threshold, the viral community is comprised of 82 members.

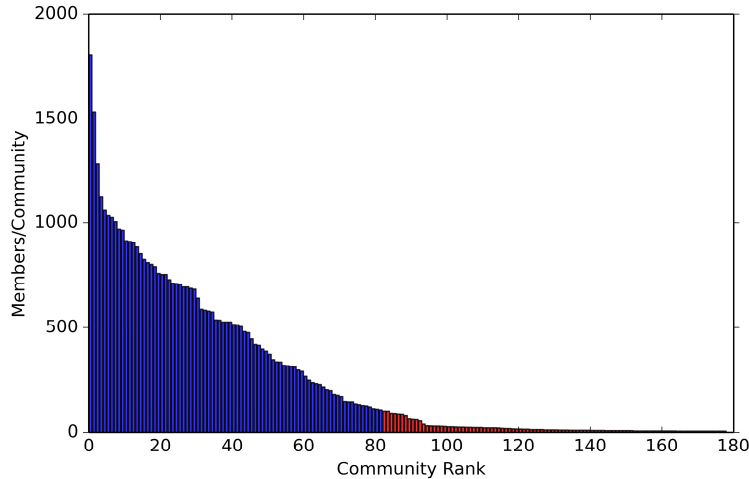


Figure 3.1. Population-based rank abundance curve. Populations are ranked according to the number of contigs (i.e. “membership”) they contain. Those with fewer than 100 members are red, with fewer than 2 members excluded for brevity.

This network approach allows for topographical properties to be calculated in order to describe the relationships between and within the populations. As a whole, the separation of populations within the viral community network is more similar within their populations than between the populations. One such metric that measures this property is modularity, which measures the separation strength of individual groups within the larger community (19). 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 modularities greater than 0.3 are considered structured and rarely exceed 0.7 (18). A second metric, known as the clustering coefficient, measures how well connected nodes are with their neighbors.

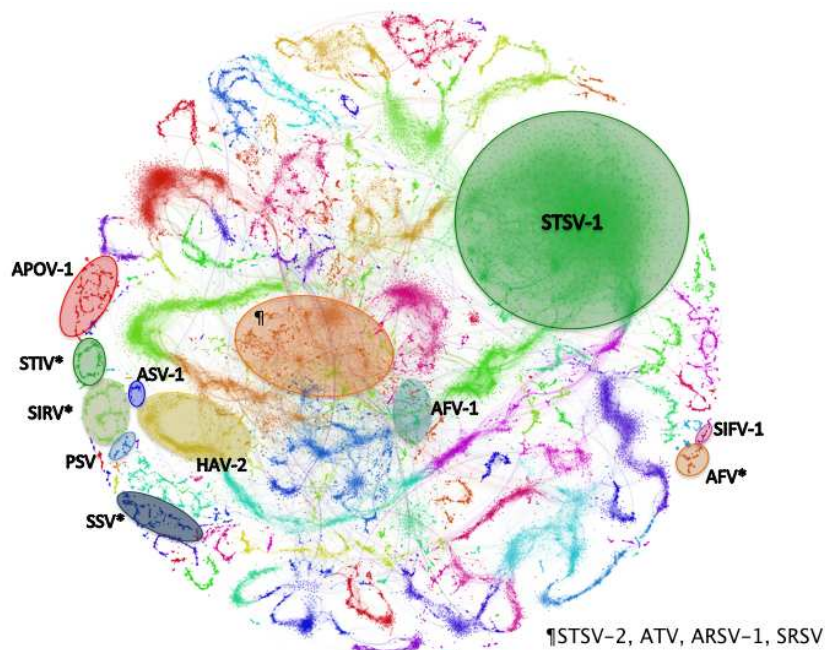


Fig 3.2. Metavirome network. Distinctions between viral populations are created by applying the Louvain method to contig relationships established through an all-vs-all BLAST. Each population is randomly assigned a unique color. A parallel network analysis was performed with 67 known archaeal viruses, allowing for identification of 12 communities correspond to viruses known to typically inhabit high temperature acidic hot springs virus. Archaeal viruses not shown do not have matches to the network. For brevity, only communities containing more than 50 members are included. STIV* = STIV-1,2. SIRV* = SIRV-1,2. AFV* = AFV-2,3,6-9. SSV* = SSV-1,2,4-9.

The clustering coefficient was 0.749 and modularity index was 0.941 for the viral network, indicating well-supported and highly connected clusters defining the viral community.

In order to better understand the minimum sampling required of the viral community to define the network with high confidence, network analyses were performed on sequentially added 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 to the final network, 5 metagenomes were required to achieve sufficient sampling (Table 3.5).

Table 3.5. Network Analysis on Stepwise-Added Metagenomes

# Metagenomes	1 ¹	2 ²	3 ³	4 ⁴	5 ⁵	6 ⁶	7 ⁷	8 ⁸	9 ⁹
Contigs	273	5420	12049	17582	23846	30968	37085	42074	44714
Edges	2454	38759	121736	216932	365756	550281	691156	803160	879459
Average Degree	11.79	14.302	20.207	24.677	30.677	35.539	37.274	38.178	39.337
Modularity	0.433	0.917	0.932	0.941	0.938	0.939	0.938	0.941	0.941
Populations	11	36	55	55	63	65	76	77	82
Avg. Clustering Coefficient	0.693	0.73	0.742	0.752	0.752	0.75	0.749	0.749	0.749
# Metagenomes	1 ¹⁰	2 ¹¹	3 ¹²	4 ¹³	5 ¹⁴	6 ¹⁵	7 ¹⁶	8 ¹⁷	9 ¹⁸
Contigs		1094	8084	16796	23194	28602	35381	40383	44479
Edges		4781	47624	127497	232836	337955	518282	688037	863162
Average Degree		8.74	11.782	15.182	20.077	23.632	29.297	34.076	38.812
Modularity		0.877	0.951	0.953	0.946	0.949	0.946	0.945	0.942
Populations		22	52	76	74	80	84	86	83
Avg. Clustering Coefficient		0.681	0.717	0.728	0.735	0.741	0.745	0.748	0.749

So as to evaluate how well the network clusters/populations predicted known virus families, 67 archaeal virus genomes were “seeded” in a parallel network analysis. 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. Twenty-nine archaeal viruses, representing 14 families, and 5 unclassified species were found deeply rooted within 13 network clusters/populations, representing 20% of the contigs (Figure 3.2). Of these 29 viruses, all of the *Fuselloviridae* (9 viruses) and nearly all *Lipothrixviridae* (7/8) were contained within their own clusters (AFV-1 was found in its own cluster). Of the four characterized *Rudiviridae*, SIRV-1 and SIRV-2 were found associated together within their own distinct populations (ARV-1 and SRV-1 are the other two *Rudiviridae* described below). Spherical viruses STIV-1 and STIV-2 (from the proposed “Turriviridae”) are also found tightly centered within a single network population as well. The largest population in the seeded network was a mixed *Acidianus* two-tailed virus (ATV), *Acidianus* rod-shaped virus 1 (ARV-1), STSV-2 and *Stygiolobus* rod-shaped virus (SRV), comprising 3.8% of the overall network composition. A number of other archaeal viruses were found exclusively in single clusters, including *Pyrobaculum* spherical virus (PSV; 0.6%), Hyperthermophilic Archaeal virus 2 (HAV2; 1.8%), STSV-1 (2.7%), *Sulfolobus* islandicus filamentous virus (SIFV; 0.21%), *Acidianus* filamentous virus 1 (AFV-1; 1.9%) and the *Guttaviridae* virus *Aeropyrum pernix* ovoid virus 1 (APOV-1, 2.4%). The grouping of recognized viral species within the same network cluster supports the

conclusion that a network cluster represents a taxonomic unit at the viral family level or higher. The remaining 38 seeded viruses were found to not share any sequence similarity to any contig present in a network cluster. As expected, these include all of the haloviruses (30 viruses) not seen previously in hot springs environments. Surprisingly, eight viruses were not detected, even though these viruses have been previously isolated from hot spring environments. 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) and *Thermococcus prieurii* virus 1 (TPV-1). In order to further support the absence of these viruses from the network analysis, every individual metagenome sequencing read was aligned against each of the viral references, finding no reads aligning at any significant (e-value cutoff of 10^{-3}) level.

To see if a similar geochemical environment or location could yield differing community compositions (or compare the viral populations from two geochemically similar conditions), contigs from another high temperature, acidic hot spring in YNP were compared to the major populations present in the NL10 viral community. Seventy-eight of the 82 major populations (95%) in NL10 were found to contained strong (average e-value 10^{-12}) matches to contigs present in CHAS. With an average identity of 80.9% over a length of 386 bp, it is clear that the largest populations in NL10 are present in other environments of similar geochemistry. While there is a small positive correlation ($R^2 = 0.229$) between the population size and the number of CHAS matches, a number of

outliers exist. The SIRV-1,2 group (population 14) is the second largest matching in CHAS. Population 3 is the largest matching group in CHAS, with as many matches as the next 6 largest matching groups combined.

We have examined the viral community stability using the defined viral network clusters (Figure 3.3). Unlike traditional measures of diversity and static rank abundance plots, the network-based approach used here has the advantage of being able to discriminate between time points within and between viral populations. It was found that the largest 9 viral clusters (largest defined as the total number of contig members) and the “STIV”-cluster all contain members from each sampling time point (Figure 3). However, at no time point were a majority (>50%) of the contigs members of a cluster detected. Typically, around 10% of the contig members of a given network cluster were detected at a given time point, but this value varies from near zero to 30%. We argue that in those sampling time points where near zero contig members were detected, that these samples typically had lower sequencing depth. A more global analysis revealed that 58 of the major viral populations (representing 79.3% of the network) were present at every sampling point. Seventy-eight of the major clusters, accounting for 93.0% of the data, consists of members spanning at least half of the sampling points. Overall, these results indicate that each individual network cluster (viral population) is relatively stable and present at similar levels during the two-year sampling period. It also suggests that at any given sampling point, the sequencing depth was not sufficient to detect a majority of the viral contig members within a given network cluster. However, by analyzing a time

series, sufficient sequencing depth at each individual time was accomplished for defining stable network clusters that define the viral community structure.

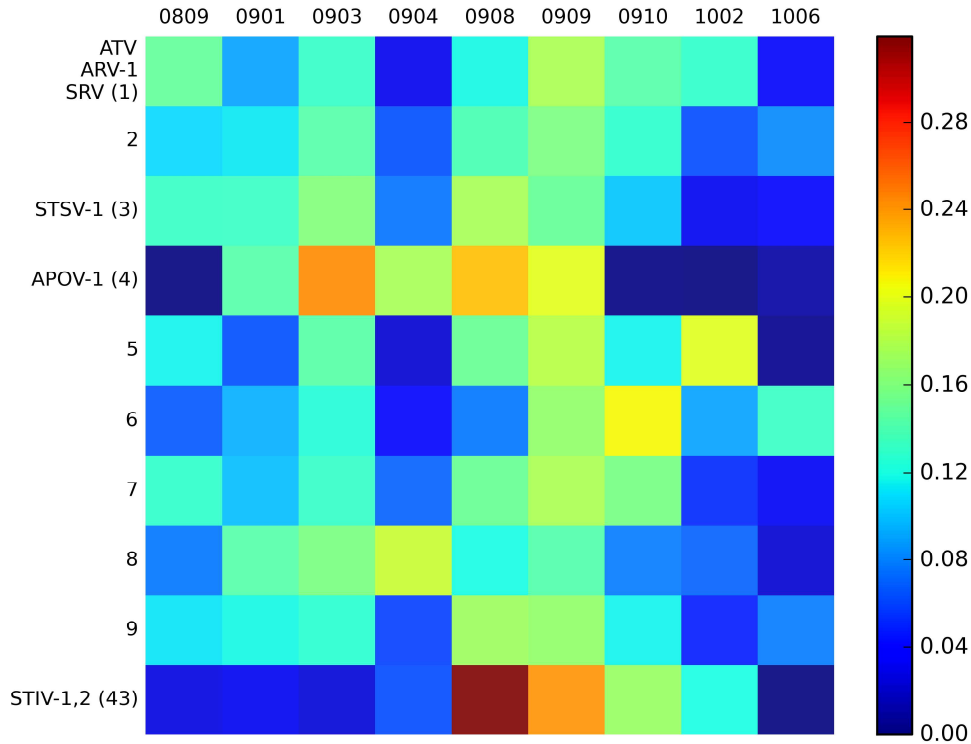


Figure 3.3. Community temporal dynamics. Each row in the heat map represents a particular viral population (i.e. network cluster) with identification where possible with each column representing a time point. Each point represents the fraction of contigs present relative to the total contigs of each population. Change between time points is identified through the difference of the fractions. The color bar (shown to the right) is shown to the right, with larger fractions (red) and smaller (blue). The largest nine viral populations and the STIV-associated cluster are included.

Discussion

The overall objective of this study was to determine the viral community structure and stability by combining temporal viral metagenomic datasets with a community network approach in selected YNP hot springs. Taxonomic classification of the cross-assembled metavirome revealed roughly half of ORFs were unable to be classified,

following the trend of what other groups have reported (21, 29). The lack of homology suggests a yet-unexplored large community of archaeal viruses remaining elusive. Classifiable ORFs matched almost entirely to previously characterized archaeal viruses (~97%).

In an effort towards a more complete understanding of the viral community structure and its stability, a network-based approach was designed and applied simultaneously to 9 metaviromes. This approach expands the current extent of viral community analysis by creating a network of connections between contigs assembled from individual metaviromes, representing differing sampling dates. Connections between contigs represent shared genes, overlapping edges from incomplete viral genomes or similar contigs present at multiple sampling times. Seeding the viral network with known archaeal viruses supports the validity of the network analysis by tightly grouping highly related viruses and distinguishing more distant ones. It indicates that the known archaeal viruses represent only 16% of the populations (13/82 major clusters had members matching known viruses) and 20% of all the metagenomic data. This points to the possibility that the remaining network clusters represent viral populations of entirely novel viral families. In addition, the network approach is not dependent on sequences present in public databases, providing a *de novo* approach to identification of related and unclassified sequences. The re-assembly of viral clusters with subsequent contig reduction and length increase offers an additional advantage of introducing this type of network approach in a viral metagenomics assembly pipeline. PCR primers can be designed against members in network clusters lacking known viral families in an in order

to quantify viruses in environmental samples and to aid in their identification and isolation.

Through sequential addition of metagenomes, it was revealed that roughly 5 of the 9 time points were required to achieve similar topological properties as to the “final” network. This threshold indicates insufficient sequencing depth at various sampling points and further supports a temporal study to overcome these limitations.

We believe the network clusters represent pan-genomes of highly related viruses. Of the subset of network clusters to which known archaeal viruses are members, some of the network clusters match to single viruses (i.e. APOV-1, STSV) or highly related viruses (i.e. STIV-1, 2), while others possess matches multiple viruses belonging to entire virus families, such as the *Fuselloviridae* and *Lipothrixviridae*. It is tempting to speculate that those network clusters with few members represent virus types/families with smaller pan viral genome size as compared to those with large contig membership. Additional deep sequencing of these communities will be required to determine if this is the case.

We believe the 82 viral populations represent the upper bound for number of major viral groups present in this hot spring environment. While 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 weight *network* rank abundance with *species* abundance. This is complicated by the fact PHACCS estimates richness in terms of genotypes at 98% identity and the network analysis does not make such assumptions. Due to the underlying nucleotide BLAST that creates inter-contig connections, the more highly related

sequences are at the nucleotide level, the more likely they are to be detected by the Louvain algorithm to be within the same cluster/population. Shared genes also influence these connections, and thus their associations, between contigs and their clusters. This is due to the fact the Louvain method seeks to maximize modularity by maximizing the number of weighted connections within a cluster and minimizes connections between clusters. Heavily shared genes (such as viral glycosyltransferases) can influence the network by tightening the connections between otherwise unrelated genomes/populations. Read recruitment from each of the metagenomes onto the network clusters shows a correlation between the size of the community and the number of reads present (data not shown). Since the length of a contig generally increases with the increases in reads assembling to it, larger contigs that overlap other large contigs likely have a greater number of reads represented in the community as well and could serve as a proxy for relative abundance. The population size can signify a single genome with numerous contigs 'tiled' across the genome or groups of highly variable genes present at each time point. In the reference-seeded network, read recruitment and re-assemblies suggest it is the former, rather than the latter. Despite this drawback, many of the viral populations show trends across time that likely correspond to abundances within the environment.

This work provides a more comprehensive picture of the underlying stability of the overall viral community, moving beyond static numbers. In general we observed a stable viral community structure through the course of this study. However, it is important to point out that the combination of possible insufficient sequencing depth and

possible amplification bias makes it difficult to observe dynamic fluctuation with network clusters/viral populations. We speculate that the underlying viral population is kept in balance by the supporting capacity of the host community.

Mapping contigs generated from another YNP hot spring, with similar temperature and pH revealed an astonishingly high number of major populations to be present in both environments. Seventy-eight of the 82 major NL10 populations were found having a large number of CHAS contigs mapped against them. The remarkable similarities seen through the viral populations is likely a consequence of their geochemical similarities with difference between the two communities pointing to underlying abiotic and biotic factors, such as a changing host community or different elemental concentrations.

Our network approach is distinct from previous viral network analysis (9). The objective of this work was not to assemble full-length genomes (though it is a byproduct of this process) or to measure abundance through read recruitment. In contrast, our goal was to gain a better understanding of viral community structure and stability. The high stringency assemblies ensure that each contig likely only represents one viral genotype, rather than a consensus sequence for a given virus type. This allows for much greater discriminatory power when analyzing the connections governing the underlying network structure. Depending on the homogeneity of each viral genome's sequence space, the Louvain method can detect those subtle differences that are lost when composite genome fragments are utilized (i.e. through <98% assembly identities).

While this network approach has the potential to greatly expand our understanding of viral community structure, it is not without limitations. As with many studies, sampling bias can affect the members of the community being sampled, sequences being amplified or even the completeness of the sequences. All of these factors affect the number of reads assembling into contigs, which in turn expands the sequence space and allows for more connections within the network. If an insufficient amount of sequence space has been covered between two regions on the same genome, no connection will be created and the network analysis will fail to include them within the same population.

Literature Cited

1. **Angly, F., B. Rodriguez-Brito, D. Bangor, P. McNairnie, M. Breitbart, P. Salamon, B. Felts, J. Nulton, J. Mahaffy, and F. Rohwer.** 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information., p. 41, BMC Bioinformatics, vol. 6.
2. **Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. *Plos Biol* **4**:2121-2131.
3. **Angly, F. E., D. Willner, A. Prieto-Davó, R. A. Edwards, R. Schmieder, R. Vega-Thurber, D. A. Antonopoulos, K. Barott, M. T. Cottrell, C. Desnues, E. A. Dinsdale, M. Furlan, M. Haynes, M. R. Henn, Y. Hu, D. L. Kirchman, T. McDole, J. D. McPherson, F. Meyer, R. M. Miller, E. Mundt, R. K. Naviaux, B. Rodriguez-Mueller, R. Stevens, L. Wegley, L. Zhang, B. Zhu, and F. Rohwer.** 2009. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes, p. e1000593, PLoS Comput Biol, vol. 5.
4. **Blank, C. E., S. L. Cady, and N. R. Pace.** 2002. Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. *Appl Environ Microbiol* **68**:5123-5135.
5. **Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre.** 2008. Fast unfolding of communities in large networks, p. P10008, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008. IOP Publishing.
6. **Bolduc, B., D. P. Shaughnessy, Y. I. Wolf, E. V. Koonin, F. F. Roberto, and M. Young.** 2012. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J Virol* **86**:5562-5573.
7. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2003. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**:6220-6223.
8. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**:14250-14255.
9. **Emerson, J. B., B. C. Thomas, K. Andrade, E. E. Allen, K. B. Heidelberg, and J. F. Banfield.** 2012. Dynamic Viral Populations in Hypersaline Systems as Revealed by Metagenomic Assembly, p. 6309-6320, *Appl Environ Microb*, vol. 78.

10. **Fancello, L., D. Raoult, and C. Desnues.** 2012. Computational tools for viral metagenomics and their application in clinical research, p. 1-13, *Virology*. Elsevier.
11. **Gill, S. R., M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson.** 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312**:1355-1359.
12. **Inskeep, W. P., D. B. Rusch, Z. J. Jay, M. J. Herrgard, M. A. Kozubal, T. H. Richardson, R. E. Macur, N. Hamamura, R. Jennings, B. W. Fouke, A. L. Reysenbach, F. Roberto, M. Young, A. Schwartz, E. S. Boyd, J. H. Badger, E. J. Mathur, A. C. Ortmann, M. Bateson, G. Geesey, and M. Frazier.** 2010. Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5**:e9773.
13. **Lee, C., and P. Cunningham.** 2013. Benchmarking community detection methods on social media data, arXiv preprint arXiv:1302.0739.
14. **Li, W., and A. Godzik.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658-1659.
15. **Lowenstern, J. B., U. o. Utah, and G. Survey.** 2005. Steam Explosions, Earthquakes, and Volcanic Eruptions: What's in Yellowstone's Future? U.S. Geological Survey.
16. **Mathieu Bastian, S. H. M. J.** 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks, p. 1-2.
17. **Mokili, J. L., F. Rohwer, and B. E. Dutilh.** 2012. Metagenomics and future perspectives in virus discovery, p. 63-77, *Current Opinion in Virology*, vol. 2. Elsevier B.V.
18. **Newman, M.** 2006. Modularity and community structure in networks, p. 8577-8582, *Proceedings of the National Academy of Sciences*, vol. 103.
19. **Newman, M. E. J., and M. Girvan.** 2004. Finding and evaluating community structure in networks., p. 026113, *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 69.
20. **Niu, B., L. Fu, S. Sun, and W. Li.** 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11**:187.
21. **Prangishvili, D., P. Forterre, and R. A. Garrett.** 2006. Viruses of the Archaea: a unifying view, p. 837-848, *Nature Publishing Group*, vol. 4.

22. **Reysenbach, A. L., G. S. Wickham, and N. R. Pace.** 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **60**:2113-2119.
23. **Roux, S., M. Faubladier, A. Mahul, N. Paulhe, A. Bernard, D. Debroas, and F. Enault.** 2011. Metavir: a web server dedicated to virome analysis, p. 3074-3075, *Bioinformatics*, vol. 27.
24. **Ruehrup, S., P. Urbano, and A. Berger.** 2013. Botnet detection revisited: theory and practice of finding malicious P2P networks via Internet connection graphs, ... *IEEE Conference on*.
25. **Wilhelm, S. W., and C. A. Suttle.** 1999. Viruses and Nutrient Cycles in the Sea. American Institute of Biological Sciences Circulation, AIBS, 1313 Dolley Madison Blvd., Suite 402, McLean, VA 22101. USA
26. **Willner, D., M. Furlan, R. Schmieder, J. A. Grasis, D. T. Pride, D. A. Relman, F. E. Angly, T. McDole, J. Mariella, Ray P, and F. Rohwer.** 2011. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity, p. 4547-4553, *Proceedings of the National Academy of Sciences*, vol. 108. National Acad Sciences.
27. **Wommack, K. E., J. Bhavsar, S. W. Polson, J. Chen, M. Dumas, S. Srinivasiah, M. Furman, S. Jamindar, and D. J. Nasko.** 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences., p. 427-439, *Stand. Genomic Sci.*, vol. 6.
28. **Yau, S., F. M. Lauro, M. Z. DeMaere, M. V. Brown, T. Thomas, M. J. Raftery, C. Andrews-Pfannkoch, M. Lewis, J. M. Hoffman, J. A. Gibson, and R. Cavicchioli.** 2011. Virophage control of antarctic algal host-virus dynamics., p. 6163-6168, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108.
29. **Yoshida, M., Y. Takaki, M. Eitoku, T. Nunoura, and K. Takai.** 2013. Metagenomic Analysis of Viral Communities in (Hado)Pelagic Sediments, p. e57271, *PLoS ONE*, vol. 8.

CHAPTER FOUR

CHARACTERIZATION OF VIRAL COMMUNITIES IN
YELLOWSTONE HOT SPRINGS
BY DEEP-SEQUENCING

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Benjamin I. Bolduc

Contributions: Conceived the study, collected and analyzed output data, and wrote the manuscript and edited the manuscript during all stages.

Co-Author: Mark J. Young

Contributions: Obtained funding, discussed the results and implications, provided valuable insight and edited the manuscript during the final stages.

Manuscript Information Page

Authors: Benjamin I. Bolduc, Mark J. Young

Status of Manuscript:

- ☒ Prepared for submission to a peer-reviewed journal
- ☐ Officially submitted to a peer-review journal
- ☐ Accepted by a peer-reviewed journal
- ☐ Published in a peer-reviewed journal

CHAPTER FOUR

CHARACTERIZATION OF DNA AND VIRAL COMMUNITIES
IN YELLOWSTONE HOT SPRINGS
BY DEEP SEQUENCING

Prepared

Benjamin Bolduc^{1,4} , and Mark Young^{1,2,3}

Thermal Biology Institute,¹ Depts. of Microbiology,² Plant Sciences and Plant Pathology,³ Chemistry and Biochemistry,⁴ Montana State University, Bozeman, MT 59717 USA

Abstract

Understanding viral community structure remains a challenging task in environmental virology. Advancements in sequencing have met this challenge head-on, but still provides an inadequate picture of the community. Despite these recent advancements, there are no ultra-deep sequencing projects of viral communities targeting extreme environments dominated by the *Archaea*. We report here the first such study investigating viral community structure in the high temperature, acidic hot springs of Yellowstone National Park, USA. These hot springs harbor low complexity cellular communities dominated by several species of hyperthermophilic *Archaea*. We applied a novel networks-based approach with viral metagenomics to DNA and RNA viruses using Illumina-based sequencing. This approach identified 77 viral groups in the DNA viral community and 24 in the RNA viral community. We expand on previous findings and present a near-complete picture of both DNA and RNA viral communities. Analysis of the RNA viral metagenomes led to the identification of novel RNA viral genome segments. In addition, several RNA-dependent RNA polymerases, hallmark genes for single-stranded, positive-sense RNA viruses, were identified in the RNA virus community. These sequences were later confirmed to be RNase-sensitive and present throughout the course of study.

Introduction

Developments in sequencing technologies and new methodologies have seen the discovery of more archaeal viruses in the past decade than in the previous 50 years combined (2, 18, 19, 36, 37, 45-48). These discoveries advance beyond the field of archaeal virology and into virology as a whole. Archaeal viruses are noted most often for their exceptional morphologies, such as the turreted, T=31 icosahedral STIV (58), the bottle-shaped ABV (42) and the beehive appearing SNDV (1). Beyond extraordinary morphologies, archaeal viruses include the largest single-stranded DNA viral genome (35), the smallest known double-stranded DNA genome infecting a prokaryote (37) and the only virus with an extracellular development stage outside its host (51). Another exceptional feature is that many archaeal viruses, especially those of the *Crenarchaeota* exist stably with their host (known as a “carrier state”) – existing in low copy numbers and avoiding host integration (49). Research on the euryarchaeal HHPV-1 has revealed close evolutionary relationships between ssDNA and dsDNA viruses (59) and work with the crenarchaeal virus STIV-1 has shown a similar organization of capsid proteins between viruses infecting each domain of life, suggesting an evolutionary connection predating the three-domain split (58). More recently work on SIRV2 and STIV has shown the use of a unique egress mechanism, in which virus-associated pyramids (VAPs) are formed before cell lysis and consist of a single protein (14, 52, 63). Clearly, discoveries made through the study of archaeal viruses reveal the underlying host-virus relationships and viral evolution.

Despite these discoveries, there are no ultra-deep sequencing projects of metaviromes targeting extreme environments dominated by the *Archaea*. Sequence- and

culture-dependent approaches are of particular interest in studying not only archaeal viruses, but viral communities as well. Overcoming the reliance on primers targeting conserved regions allows for capture of more diverse sequences and no longer restricts virus discovery. Sequence-independent approaches can, in theory, amplify any sequence from a wide variety of samples, such as feces (10, 11, 68), vaginal tract (26) and human gut (54, 65). The advent of 454 pyrosequencing (29) and NGS ushered in a new era for viral metagenomics, with one study assembling 608 putative ssDNA viral genomes alone, effectively doubling the number of ssDNA viruses in extant databases (28). The combination of ever increasing read length and high-throughput has enable researchers to discover new viruses (4, 7, 15), and capture a glimpse of viral diversity in the oceans (21, 28), freshwater (60), sediments (73), the human gut (reviewed in (55)), and oral cavity (69). At the time, the short read length of Illumina reads prohibited characterization of communities (70), especially in a field where read length matters (70). Recent advancements in Illumina technology have dramatically increased read lengths, currently capable of delivering 50 million 300-nt paired-end reads, equivalent to 15 Gb. However, nearly all deep-sequencing projects have focused on viruses associated with human disease. Illumina sequencing has been used to explore zoonotic viruses (71), understand intra-host population diversity (3, 39, 41) and reveal aspects of quasispecies diversity and evolution seen through replication fidelity (16). It has also been used to detect novel and previously unseen viruses (13, 62), but rarely viral communities. In one outstanding case, Illumina sequencing provided nearly full length genomes from as little as 100 copies of an RNA virus (32).

The hot springs in Yellowstone National Park (YNP) that are low pH (pH < 4), high temperature (> 80°C) typically have low complexity microbial communities dominated by few species (23). Bacteria and eukaryotes are few or in many cases, absent (6, 22, 24, 31, 56). These extreme environmental conditions favor not only the *Achaea* but also a relatively simplified community. A number of viruses (exclusively archaeal) have been isolated out of these environments (36, 47, 48, 57).

In this work, we analyzed paired DNA and RNA viral samples out of an extreme environment in YNP using Illumina-based sequencing. To our knowledge, this is the first use of Illumina to explore the viral community present in an archaeal-dominated environment. Previous work in this hot spring revealed the presence of positive-strand RNA viruses, whose potentially host was from the *Archaea* (9). Here, we expand upon those findings and present a near-complete picture of both DNA and RNA viral communities.

Materials and Methods

YNP Sample Site

A high-temperature (72-93°C), acidic (pH 2.0 – 4.5) hot spring was selected for this study based on previous work (9). The Nymph Lake hot spring site 10 (NL10: 44.7536°N, 110.7237°W) was sampled in (January 2014). This site is located in a relatively new thermal basin that has developed in the past 15 years (30).

Sample Collection

Roughly 60-L of hot springs water was collected and transported at ambient temperatures back to the lab and immediately filtered using a 0.4-µm filter (Millipore) 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 (25). Due to the already low pH and high Fe concentration, it was unnecessary to add additional FeCl_3 . Instead, the filtered sample was pH adjusted to between pH 4.5 – 5.0 using NaOH, causing a visible orange-tint during precipitation. This flocculate was then re-filtered onto a fresh 0.4- μm filter and completely resuspended in a citrate buffer. This Fe-concentrated viral fraction was then centrifuged at 4k g for 10 min to remove any residual debris from the solution and the resulting supernatant pelleted by high-speed centrifugation at 37,000 rpm for 24 hr. The pellet was resuspended in 1000- μL RNase/DNase-free H_2O (Zymo Research) and applied as an overlay onto a preformed cesium density gradient with steps at 1.2 g/mL, 1.3 g/mL, 1.4 g/mL and 1.5 g/mL. Based on previous research (66), it was believed that most viruses would fall within one of these interfaces. This gradient was spun at 15,000 rpm in a Beckman MLS50 rotor for 2 hr 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 (Fisher Scientific).

Nucleic Acid Extraction and Sequencing Preparation

Samples were extracted using ZR viral RNA/DNA kit (Zymo Research) according to the manufacturer's instructions. Extracted nucleic acids were split into 2 halves of approximately equal volume with the "RNA viral samples" being DNase-treated with DNase I (Zymo Research; 2.5 U, 60 min) and re-extracted and the "DNA viral samples" being RNase-treated with RNase ONE (Promega; 1U over 30 min) and re-extracted. Previous work on nuclease-digestible material out of this site (9) had shown

insufficient DNase-digestion in the RNA viral fraction, and as such, the RNA viral samples were DNase-treated twice more, for a total of three rounds of DNase digestion.

The extracted viral RNA was amplified using the Ovation RNA-Seq System V2 (NuGEN) according to the manufacturer's instructions. Amplified cDNA (from the RNA viral) and the extracted viral DNA were then amplified using the Ovation Ultralow library system (NuGEN) according to the manufacturer's instructions. In the case of the cDNA, no shearing was performed since the majority of amplified products were between 150 and 550 bp. The amplified nucleic acids were sequenced by the University of Illinois sequencing center using the Illumina MiSeq v3 system with paired-end reads (2 x 300 nt).

Read Processing and Assembly

Reads received from the Illinois sequencing center were already pre-binned according to their sequencing barcodes. Briefly outlined, 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/>). During the quality check step, a large number of over represented k-mers were identified. To remove these sequences, Jellyfish (33) was used to identify all over-represented k-mers between the range of 16 – 32 and then a python script then was written to compare all identified k-mers common to either all RNA or DNA viral metaviromes. The common k-mers were then assembled (also handled by the python script) using Minimo from the amos package (version 1.5; <http://sourceforge.net/projects/amos/> and (67)) at 100% nucleotide identity and an overlap length of 50%. The assembled sequences matched to PCR and sequencing

primers used during the Illumina sequencing process. An additional sequence could not be identified but we suspect it is the Ovation RNA Seq V2 primer used during cDNA synthesis. This adapter sequence, as well as all the sequencing, amplification and adapter sequences were trimmed with Trimmomatic V0.32. Following trimming, RNA sequences were assembled using VICUNA (72) using the Ovation RNA-Seq V2 primer and Illumina sequences for read trimming (sequence assembly statistics shown in Table 1). DNA sequences were assembled using IDBA-UD (43, 44) using a k-mer range of 28 – 124 and default parameters.

Rapid Identification of RdRPs in Viral Metagenomes

In order to rapidly identify RNA-dependent RNA polymerase (RdRP) sequences, often considered a hallmark gene for positive-stranded single-stranded RNA viruses (27), a pre-filtered and pipelined version of HHpred was used (64). Briefly, assembled sequences are translated in all six reading frames. The ORFs are then pre-filtered if they contain the canonical GDD box (also known as “motif C”) and then subjected to analysis using the HMM-HMM comparison program HHblits (53), invoking the prediction of secondary-structure, a matrix of 0.25, and then first searching against the nr database filtered to 20% nucleotide identity, and subsequently against the SCOP and PDB databases.

Analysis of Viral Diversity

Rarefaction curves were constructed by an in-house version analogous to Metavir (61). 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 (17) at 97%, 90% and 75% nucleotide identities. The resulting clusters and input sequences are plotted using the matplotlib library (20) for python.

Network Analysis and Re-Assembly of Viral Populations

Viral metagenomes were analyzed using a novel network analysis technique (Bolduc et al, *manuscript in prep.*). Briefly, contigs from each DNA and RNA metavirome were compared to each other via BLASTn with default parameters, except e-value cutoff of 10^{-5} . The resulting file was parsed; HSPs were evaluated for a minimum alignment length of 50 bp, 75% nucleotide identity and e-value of 10^{-10} . The surviving contigs and their relationships (HSPs) were converted to a directed graph, with contigs as graph nodes and their HSP information (which show the pairwise relationship between a target and query sequence) stored in the graph's edges properties. One of these properties, the edge *weight*, is the measure of strength connecting two contigs to each other. Higher values are associated with higher strength and "closeness" while lower values will have contigs more distal to each other. A community detection algorithm (the Louvain method (8)) was applied to the graph. This method utilizes the edge's weight (the log of the e-value) as well as the number of edges connecting contigs/nodes in the network to maximize the modularity of resulting partitions – essentially grouping more related sequences. Clusters containing fewer than 100 contigs/nodes were removed from downstream analyses. Graph networks were visualized using Gephi (34) using the OpenOrd layout (developed from the force-directed Fruchterman-Reingold) and optimized using Gephi's Force Atlas 2 plugin. Both of these force-directed graph-

drawing algorithms assign forces to both nodes and their connecting edges. Nodes are repulsed from each other as if electrically charged, while their edges are springs which serve as an attractive force. The strength of the spring is correlated to the edge weight. Network statistics and topological properties were calculated using Gephi. The contributions of each metavirome “fraction” to each cluster were also calculated through the use of in-house scripts. As controls, all publically available virus genomes (including the Illumina loading control, bacteriophage ϕ X174) were seeded into the initial BLAST and the network analysis was repeated identically as above.

Due to the highly diverse nature of viral genomes, even the latest metagenome assemblers cannot fully assemble every variant genome, reconstruct segments of viral genomes with dramatically different sequencing depths or reconstruct individual members in viral families with highly conserved regions. To partially overcome this limitation, each network cluster (defined by the community detection algorithm) was re-assembled using the Minimo assembler (version 1.6), based on the AMOS infrastructure (v. 3.1.0 (67)) with a 35-bp minimum overlap and 98% overlap identity.

Results and Discussion

Sample Site, Isolation and Extraction

A single site in the Nymph Lake region was selected based on previous work characterizing the cellular and viral communities. This previous work, as well as work by others has shown a low pH, high temperature combination favors the dominance of *Archaea*. We sought to exploit this aspect and selected a sampling site where preliminary investigations showed the near exclusivity of *Archaea* and a viral community dominated

by their viruses. One aspect lacking in the methodology of prior work was the simple filtration step to separate cells, cellular debris and other particulate matter from the viral component. Removal of such material is critical for downstream analyses, as cells outnumber viruses in these environments ten-to-one and each cells contain 1000x more genetic material. A 1% carryover results in 100x more cellular material to be sequenced over viral. Following large volume isolation of hot spring water and filtering to remove the cells and other particulates, viruses were concentrated using a chemical flocculation commonly employed in wastewater treatment plants to remove viruses. This method has recently been extended to marine environments and possesses a high recovery rate (> 95%), can concentrate given its adsorption onto filters, and can be scaled to accommodate nearly any feasible sample volume. After low-speed centrifugation to remove non-resuspended material and other environmental debris, a high-speed centrifugation step further concentrated viruses. Finally, a CsCl gradient separated the viral concentrate into distinct densities of 1.2, 1.3, 1.4 and 1.5 g/mL – the range of buoyant densities encompassing most viruses (66). Although not quantitative, these densities provided a guide for possible virus classification and grouping. Following isolation and extraction of total nucleic acids, RNA samples were DNase treated twice to prevent carryover of cellular and viral DNA seen previously (9) and DNA samples were treated with RNase.

Due to the extraordinarily low (< 100 fg) amount of RNA in these samples following extraction and nuclease treatments, the Ovation RNA-Seq V2 system was selected as a sequence-independent approach to amplify the low input material. It has been noted that other similar sequence-independent, RNA amplification kits are more

sensitive to rRNA amplification and do not produce as many reads aligning to their target genome (32). After cDNA synthesis and amplification, roughly 5 ng was produced per sample. The cDNA and DNA (~10 ng/sample) were then amplified using the Ovation Ultra Low library system, generating >1000 ng/DNA sample and ~ 200 ng/cDNA sample. Unfortunately, the amount of input material isolated from vDNA2 (1.2 g/mL) was less than 100 ng total and was not sequenced alongside its companion fractions.

Amplified viral cDNA and vDNA was sequenced using the MiSeq platform with the latest V3 chemistry due to the longer read lengths (300 nt) and the higher throughput (>15 million paired reads). A total of 11,513,837 paired-end reads, totaling ~7 Gb of sequence were generated (Table 4.1).

Processing and Assembly of Reads

Reads from each fraction were quality trimmed against the standard Illumina primers and quality checked using the FastQC program. During this phase it was noted that a large number of cDNA reads (> 50-90%) contained over-represented k-mers. These k-mers were almost exclusively located towards the ends of reads, suggesting the presence of the Ovation RNA-Seq amplification primers. Assembly of the k-mers generated a single sequence, which was trimmed from the cDNA reads (in addition to the standard Illumina primers). On average, trimming removed 14% of the reads (Table 4.1). Following trimming, reads were assembled initially using MIRA (5, 12), IDBA-UD (43), VICUNA (72), MetaVelvet (38) using their default parameters (data not shown). The assemblies were evaluated for overall contig lengths, N50 contig sizes, number of contigs and the percent of reads assembling into contigs. Based on these criteria, VICUNA was selected to perform the RNA viral assemblies and IDBA-UD for the DNA viral

assemblies. We suspect that the differences in assembler performance are related to the sequencing depth, overall coverage and diversity found between the viral samples. It should be noted that VICUNA was originally developed to assemble highly heterogeneous RNA viruses out of complex (i.e. heavily host-contaminated) samples. Not restricting an assembler to perform equally for all viromes is an advantage in cases where fundamental differences in viral populations and the community in which it exists create dynamically different graphs.

Table 4.1 Sequencing and Assembly Statistics

Sample Name	Density (g/mL)	Raw Reads	Trimming			Assembly*	
			Paired	Unpaired	Removed	Contigs	Assembled (%)
vDNA3	1.3	2,015,924	1,774,502	235,468	5,954	43,481	23.2%
vDNA4	1.4	1,617,630	1,496,718	116,114	4,798	29,522	16.9%
vDNA5	1.5	2,445,089	2,265,053	173,495	6,541	45,258	16.9%
vDNAx	N/A	6,078,643	5,536,273	525,077	17,293	87,563	16.2%
cDNA1	1.2	1,189,667	1,036,607	133,342	19,718	11,485	41.5%
cDNA2	1.3	1,279,779	1,069,619	176,575	33,585	13,284	53.8%
cDNA3	1.4	1,776,551	1,452,308	281,305	42,938	18,360	61.1%
cDNA4	1.5	1,189,197	898,623	236,815	53,759	11,952	55.2%
cDNAx	N/A	5,435,194	4,457,157	828,037	150,000	41,332	21.3%

*Assemblies of vDNA viromes are performed using IDBA-UD. Assemblies of cDNA viromes are performed using VICUNA. “x” assemblies represent combined fractions.

An assembly of the metaviromes revealed differences in the number of contigs generated and the percentage of reads assembling (Table 4.1). On average, 19% of vDNA metaviromes assembled, compared to 53% of the cDNA metaviromes. This difference could be attributed to the difference in assembler chosen or, more likely, a function of viral diversity in the hot springs. Fewer assembling reads corresponds to higher levels of sequence divergence, and by extension, are more diverse.

Viral Diversity

Rarefaction curves are often used to assess species richness and whether or not more sampling would lead to the discovery of additional species. An in-house script mirroring the methodology of Metavir was applied to each of the metaviromes. The subsampling of 250,000 reads from each dataset was chosen to balance the analysis' completeness (of sampling all reads) vs. computation time. It is clear from the slope of the curves that cDNA metaviromes are more sampled (and arguably) less diverse than their vDNA counterparts (Figure 4.1). Sampling each fraction individually shows a difference in sample diversity between cDNA3 and other cDNA steps. It is interesting to note that electronmicrographs show a correlation between the morphological diversity seen and the estimated richness – with cDNA3 showing a greater range of morphologies than any of the other fractions (data not shown). Such a clear difference is not seen in the vDNA viromes. While the addition of each sequence in the cDNA increases the number of clusters by less than 5%, the addition of each sequence in the vDNA viromes increases clusters by nearly 40%.

Network Analysis

A network based analysis of the vDNA and cDNA metaviromes was performed as described previously (Bolduc et al, *manuscript in prep*). This network analysis creates a graph of nodes and connections based on relationships between sequences. In the case of BLASTs, the query and target sequences are nodes in the network, while the HSPs themselves serve as the connections (i.e. edges) between them. The HSP attributes are transferred to the edges – with e-value, nucleotide identities or any other HSP-stored information can be used as a measure of the 'strength' between two sequences. The

Louvain method was then used to “partition” the network into distinct groupings, called clusters, which represent the optimized modularity for each partition. This in effect creates a network where clusters have high modularity – meaning highly connected contigs within clusters and fewer, sparser connections between contigs of other clusters. Due to the fact the connections are based on nucleotide similarities, the resulting clusters share a high degree of similarity with each other, but much less so against other clusters.

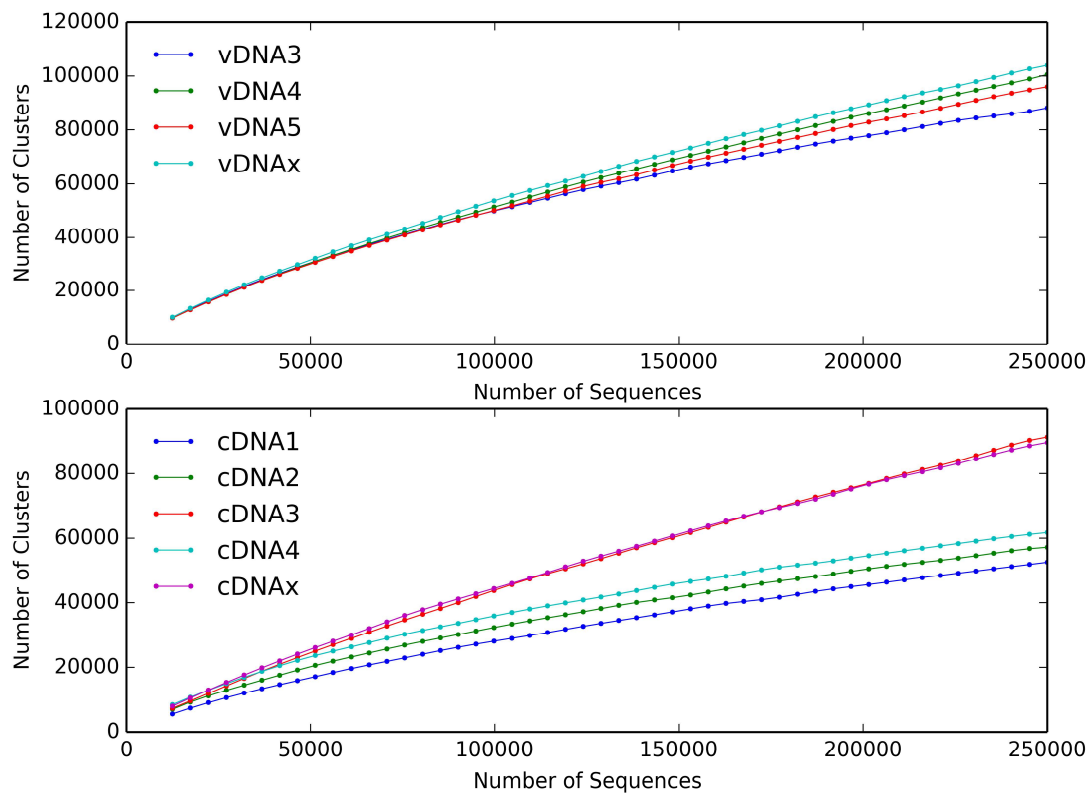


Figure 4.1. Rarefaction curve for vDNA viromes (top) and cDNA viromes (below).

When the network analysis was applied to the DNA metaviromes, 84.5% (55 844) of the 66 051 contigs possessed a match to another contig. Of these, 90.0% (50 254) were retained within the network after filtering out viral groups (identified by the Louvain method; Supplemental Figure 2) containing fewer than 100 members (Supplemental

Figure 1). This resulted in 77 well-defined viral groups and represents approximately 76% of all viral DNA contigs (Figure 4.2). As described previously (Bolduc et al, *manuscript in prep*), these viral groups likely correspond to viruses at the family taxonomic level.

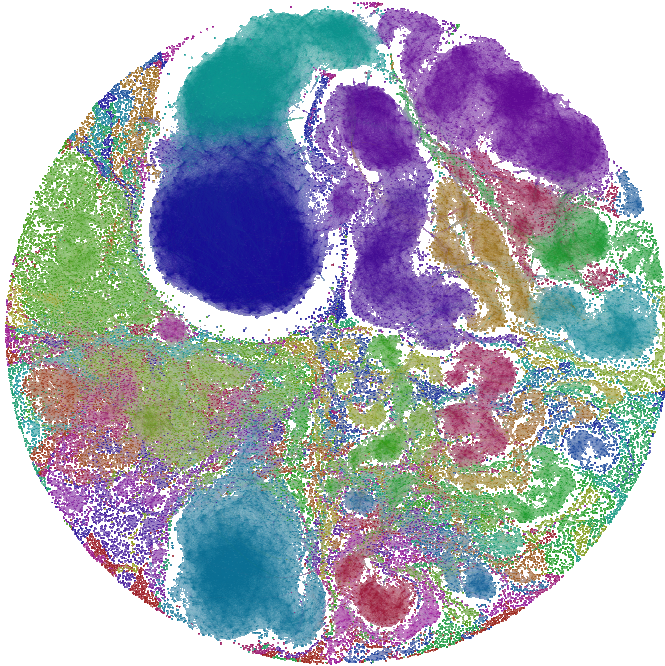


Figure 4.2. DNA metavirome network. Each cluster is given a unique color, with relationships between sequences denoted as an edge. Colored by its parent sequence.

Seeding of the DNA virus network with all publically available viruses revealed 24 viruses associated with 7 groups. All 24 viruses infected members of the Crenarchaeota – all known to exist in these environments, and all seen in the previous network analysis out of the same spring. The largest group (4 216 members) represents 8.39% of the network and includes all of the *Fuselloviridae* – ASV-1, SSV-1, 2, 4-9 – seen in this environment. The second largest group (2 123 members; 4.22%) was associated exclusively with *Acidianus* two-tailed virus (ATV). The third through six

largest groups have no viral match, nor do they match with sequences in the publically available nucleotide and protein databases. It is likely these represent entirely novel virus families. The seventh largest group (1 004 members; 2.00%) matches to the proposed *Turriviridae*, *Sulfolobus* turreted icosahedral virus 1 (STIV-1) and STIV-2. One recently described virus, found initially through *in silico* identification of proviral sequences (36), is *Aeropyrum* pernix ovoid virus (APOV-1). Although *A. pernix* has not been seen in this environment, its presence (655 members; 1.30%) indicates that *A. pernix* might be a low abundance member of the cellular population or a potential new host for APOV-1 exists in the springs. The remaining viruses found seeded within viral groups were from the proposed order Ligamenvirales, a higher-order organization of the *Rudiviridae* and *Lipothrixviridae* (50). Nearly all the filamentous viruses from *Acidianus*, AFV-3, 6-9 and *Sulfolobus islandicus* filamentous virus (SIFV) were found within one group, alongside the rudiviruses *Stygiolobus* rod-shaped virus (SRV) and *Sulfolobus islandicus* rod-shaped virus 1 and 2 (SIRV-1, 2). This group (639 members; 1.27%) was found distinct from, but closely associated with two other groups representing AFV-1 (622 members; 1.24%) and AFV-2 (381 members; 0.76%). It is extraordinary that alignments using the major capsid proteins (MCPs) of these linear dsDNA viruses group AFV-3, 6-9 and SIFV into a distinct phylogenetic group, separate from AFV-1 (50) – in excellent agreement with the separation in the viral network. The close association of SIRV-1 and 2 and SRV with SIFV is also seen in the MCP analysis, and it is known that several sets of genes overlap between the viruses, despite being in different families (40).

When applied to the RNA viromes, only 42.28% (23 287) of contigs were found to have a connection with another contig. Only 30.7% of these (7 160; Supplemental

Figure 4) were viral groups having 100 or more members (23 total groups; Supplemental Figure 2) and represents 13.0% of all contigs. Unlike the vDNA metaviromes, the RNA network does not match to previously characterized viruses. It is likely that these 23 viral groups represent entirely unknown RNA viruses existing in these hyperthermophilic environments.

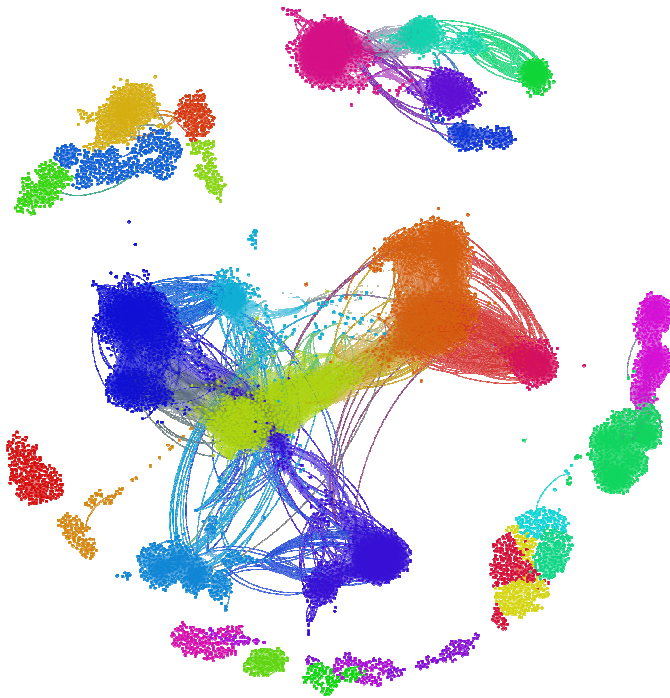


Figure 4.3. RNA metavirome network. Each cluster is given a unique color, with relationships between sequences denoted as an edge. Colored by its parent sequence.

The network also revealed a stark contrast in the distribution of contigs within viral groups between the DNA and RNA viromes. Essentially, despite the clear separation of the viral fractions by a cesium density gradient, viral groups within the DNA metavirome network were relatively well-distributed – meaning the contribution from each fraction was relatively equally distributed within a group (Figure 4.4). In contrast, the distribution for the RNA metavirome network clearly indicated distinct

contributions from each density (Figure 4.5). In some cases over 50% of the contigs from a viral groups are derived from a single fraction. It is evident that the vast majority of major RNA viral groups are separated into a cDNA1/2 or cDNA3/4 dichotomy. This agrees well with the assembly data, as combining reads from all four fractions leads to a *decrease* in the number of reads assembled. Increasing the sample diversity introduces more unresolvable points for graph reduction of de Bruijn graphs during sequence assembly. The number of contigs also goes down (when compared to the sum total of each individual fraction) not due to larger, fewer contigs, but fewer reads are going to the [potentially] same contig.

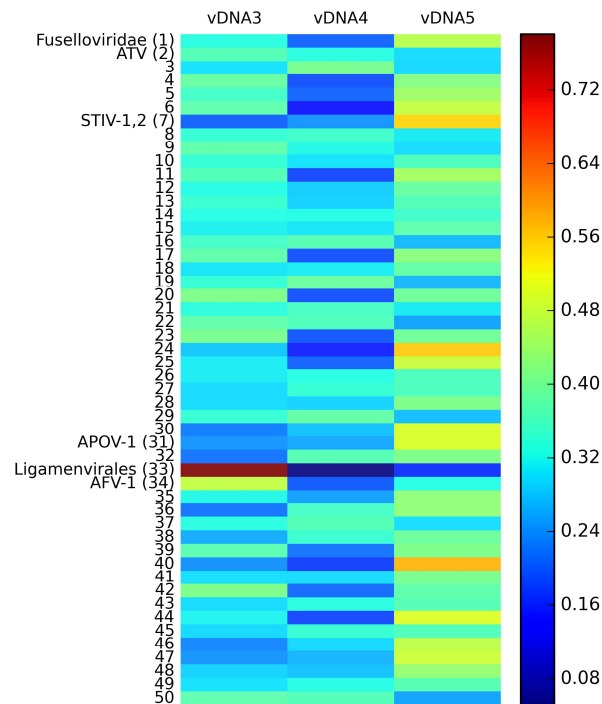


Figure 4.4. DNA metaviromes fraction contribution. Gradient fraction information of each contig was extracted from their viral groups. The percentage each fraction contributes to the overall viral group is indicated according to the figure legend (right), with larger contributes more red and fewer contributes more blue. For brevity, only the 50 largest viral groups are presented.

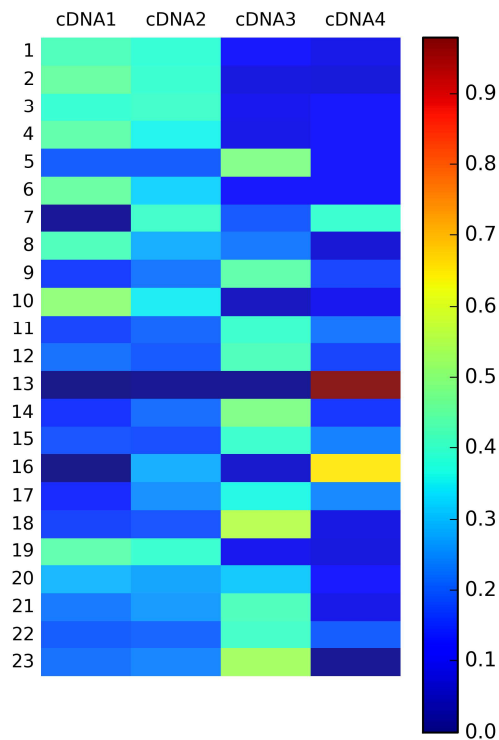


Figure 4.5. RNA metaviromes fraction contribution. Gradient fraction information of each contig was extracted from their viral groups. The percentage each fraction contributes to the overall viral group is indicated according to the figure legend (right), with larger contributes more red and fewer contributes more blue. Only viral groups containing 100 or more members are included.

Hallmark RNA Virus Genes

Previous work searching for RNA viruses out of these extreme environments (9) revealed a group of RdRp-containing sequences highly distal to known RdRps of bacterial and eukaryotic viruses – yet clearly related – despite lacking all but the universally conserved motifs at the primary sequence level. The result was sequence homology based on predicted secondary and tertiary structure, a computationally challenging effort. As an alternative to searching all ORFs against public databases, a directed approach was taken to more efficiently find strong candidates for RNA viruses. The highly conserved GDD motif of RdRps (27) was selected as a “bait” sequence due to

its importance in the replication cycle of positive, single-stranded RNA viruses. Seven sequences were identified during this process, three of which (dg-5807-cDNA3, dg-785-cDNA2 and dg-2474-cDNA2; Table 4.2) were given a probability by HHblits of being an RNA-dependent RNA polymerase > 99.99%. While these three sequences were highly similar at the protein level, they did not assemble together, suggesting 2-3 highly related RNA viruses. Comparison of these sequences with output from IDBA-UD, MIRA, and Meta-Velvet assemblers found nearly identical versions of the sequences, further supporting a clear distinction between the sequences (data not shown). The remaining four sequences were found closely associated in the same cluster during the network analysis (above) with estimated RdRp probabilities by HHblits ~45%. These hits were the only ones identified by HHblits, with no other significant alignments. Additional searches against NCBI's public protein and nucleotide databases revealed no additional matches to these sequences. Primers designed against the three sequences with a high probability showed at least two (dg-2474 and dg-785) present 3 months after the initial sampling period.

Table 4.2. GDD-containing sequences identified as RdRps

Contig	Sample	Length	Probability
dg-5807	cDNA3	260	100
dg-785	cDNA2	396	100
dg-2474	cDNA2	282	100
dg-9725	cDNA4	218	46.8
dg-7187	cDNA3	253	46.3
dg-284	cDNA3	653	43.4
dg-2335	cDNA2	285	42.7

Conclusions

The overall objective of this research was to determine the viral community structure through use of a novel network approach by in-depth sequencing of a single time point from a high-temperature, acidic hot spring in YNP. Previous work out of this environment showed a relatively stable viral community with clearly defined viral groups likely representing viral families at the taxonomic level. This work sought to delve deeper into the network approach by asking the question: *Is deep sequencing of a single time point sufficient to regenerate the viral groups of a time course-based network analysis?* Analysis of the DNA metaviromes revealed only a small portion (19.2%) matching to known viruses, and of these viruses, all were exclusively from the Crenarchaeota. Furthermore, each viral group either matched to a specific virus, or a group of viruses from the same family. In the case of the *Rudiviridae-Lipothrixviridae* group, substantial evidence supports a phylogenetic relationship grouping the two families into a larger viral order. Despite lacking taxonomic information for the remainder of the viral groups, the network shows distinct separation between groups, suggesting these groups represent additional novel viruses whose sequences are not yet in public databases. Fractionation of viruses during processing did not appear to separate the contributions of sequences towards a viral group for the DNA metaviromes, yet did so for the RNA metaviromes. This is partially explained by the large taxonomic space represented by many of the viral groups and the resulting large variation in their specific densities. Network analysis of the RNA metaviromes revealed even less taxonomic characterization, with no known viruses matching to sequences within the viral groups. Analysis of the contigs led to the identification of many GDD-containing sequences, several of which were further

analyzed and identified as RNA-dependent RNA polymerases, a hallmark of positive-sense, single-stranded RNA viruses. Further work showed the continued presence of these sequences out of the original environment and a density-dependent nature. Future work will endeavor to generate full-length sequence from these RdRp-containing contigs and seek to determine their hosts.

The viral networks presented in this work present an alternative and parallel view compared to those in Chapter 3. The number of viral groups previously reported using 454 pyrosequencing over nine time points revealed 82 viral groups. In this work, utilizing a different extraction methodology, sequencing technology and analysis pipeline – 77 viral groups were identified. With few exceptions, nearly all of the major viral groups associated with a specific group of viruses are the same between the two studies. Most surprising is the clear separation within viral groups between studies, with groups containing specific members (i.e. virus seeds) being maintained. In some cases, such as the separation of ATV and ARV-1 from their combined viral group into separate groups, deeper sequencing at a specific time point provided the higher resolution necessary for such demarcation. The presence of the same major groups between studies suggests that the NL10 viral populations are stable and persist over long time periods (5 years). However, the radical changes in many of their rank abundances between studies also suggest that these viral populations are changing temporally. It is tempting to speculate that such variation was qualitatively captured during the analysis of the temporal-based analyses, and was more accurately defined through deep sequencing at a specific time point. Given the distinct viral groups presented in this work, future studies would aim to

use specific probes designed against sequences of the viral groups, quantitatively measuring the actual abundances of their viruses.

References Cited

1. **Arnold, H. P., U. Ziese, and W. Zillig.** 2000. SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon *Sulfolobus*. *Virology* **272**:409-416.
2. **Atanasova, N. S., E. Roine, A. Oren, D. H. Bamford, and H. M. Oksanen.** 2012. Global network of specific virus-host interactions in hypersaline environments., p. 426-440, *Environmental Microbiology*, vol. 14.
3. **Batty, E. M., T. H. N. Wong, A. Trebes, K. Argoud, M. Attar, D. Buck, C. L. C. Ip, T. Golubchik, M. Cule, R. Bowden, C. Manganis, P. Klenerman, E. Barnes, A. S. Walker, D. H. Wyllie, D. J. Wilson, K. E. Dingle, T. E. A. Peto, D. W. Crook, and P. Piazza.** 2013. A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples, p. e66129, *PLoS ONE*, vol. 8.
4. **Bishop-Lilly, K. A., M. J. Turell, K. M. Willner, A. Butani, N. M. E. Nolan, S. M. Lentz, A. Akmal, A. Mateczun, T. N. Brahmbhatt, S. Sozhamannan, C. A. Whitehouse, and T. D. Read.** 2010. Arbovirus Detection in Insect Vectors by Rapid, High-Throughput Pyrosequencing, p. e878, *PLoS Negl Trop Dis*, vol. 4.
5. **Biswas, A., D. Ranjan, and M. Zubair.** 2011. Parallelization of MIRA Whole Genome and EST Sequence Assembler, Workshop on Parallel Algorithms and Software for Analysis of Massive Graphs (ParGraph).
6. **Blank, C. E., S. L. Cady, and N. R. Pace.** 2002. Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. *Appl Environ Microbiol* **68**:5123-5135.
7. **Blomstrom, A. L., F. Widen, A. S. Hammer, S. Belak, and M. Berg.** 2010. Detection of a Novel Astrovirus in Brain Tissue of Mink Suffering from Shaking Mink Syndrome by Use of Viral Metagenomics, p. 4392-4396, *J Clin Microbiol*, vol. 48.
8. **Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre.** 2008. Fast unfolding of communities in large networks, p. P10008, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008. IOP Publishing.
9. **Bolduc, B., D. P. Shaughnessy, Y. I. Wolf, E. V. Koonin, F. F. Roberto, and M. Young.** 2012. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J Virol* **86**:5562-5573.
10. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2003. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**:6220-6223.

11. **Cann, A. J., S. E. Fandrich, and S. Heaphy.** 2005. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**:151-156.
12. **Chevreur, B.** 2005. MIRA: an automated genome and EST assembler, Ruprecht-Karls University, Heidelberg, Germany.
13. **Coetzee, B., M.-J. Freeborough, H. J. Maree, J.-M. Celton, D. J. G. Rees, and J. T. Burger.** 2010. Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard, p. 157-163, *Virology*, vol. 400. Elsevier Inc.
14. **Daum, B., T. E. F. Quax, M. Sachse, D. J. Mills, J. Reimann, O. Yildiz, S. Hader, C. Saveanu, P. Forterre, S. V. Albers, W. Kuhlbrandt, and D. Prangishvili.** 2014. Self-assembly of the general membrane-remodeling protein PVAP into sevenfold virus-associated pyramids, *Proc. Natl. Acad. Sci. U.S.A.*
15. **Desnues, C., B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, M. Furlan, L. Wegley, B. Chau, Y. Ruan, D. Hall, F. E. Angly, R. A. Edwards, L. Li, R. V. Thurber, R. P. Reid, J. Siefert, V. Souza, D. L. Valentine, B. K. Swan, M. Breitbart, and F. Rohwer.** 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites, p. 340-343, *Nature*, vol. 452.
16. **Eckerle, L. D., M. M. Becker, R. A. Halpin, K. Li, E. Venter, X. Lu, S. Scherbakova, R. L. Graham, R. S. Baric, T. B. Stockwell, D. J. Spiro, and M. R. Denison.** 2010. Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing, p. e1000896, *PLoS Pathog*, vol. 6.
17. **Edgar, R. C.** 2010. Search and clustering orders of magnitude faster than BLAST, p. 2460-2461, *Bioinformatics*, vol. 26.
18. **Erdmann, S., B. Chen, X. Huang, L. Deng, C. Liu, S. A. Shah, S. Moine Bauer, C. L. Sobrino, H. Wang, Y. Wei, Q. She, R. A. Garrett, L. Huang, and L. Lin.** 2013. A novel single-tailed fusiform *Sulfolobus* virus STSV2 infecting model *Sulfolobus* species, *Extremophiles*.
19. **Gorlas, A., E. V. Koonin, N. Bienvenu, D. Prieur, and C. Geslin.** 2011. TPV1, the first virus isolated from the hyperthermophilic genus *Thermococcus*, p. 503-516, *Environmental Microbiology*, vol. 14.
20. **Hunter, J. D.** 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**:0090-0095.
21. **Hurwitz, B. L., and M. B. Sullivan.** 2013. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology, p. e57355, *PLoS ONE*, vol. 8.

22. **Inskeep, W. P.** 2012. Microbial iron cycling in acidic geothermal springs of Yellowstone National Park: integrating molecular surveys, geochemical processes, and isolation of novel Fe-active microorganisms, p. 1-16.
23. **Inskeep, W. P., D. B. Rusch, Z. J. Jay, M. J. Herrgard, M. A. Kozubal, T. H. Richardson, R. E. Macur, N. Hamamura, R. Jennings, B. W. Fouke, A. L. Reysenbach, F. Roberto, M. Young, A. Schwartz, E. S. Boyd, J. H. Badger, E. J. Mathur, A. C. Ortmann, M. Bateson, G. Geesey, and M. Frazier.** 2010. Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* 5:e9773.
24. **Jay, Z. J., D. B. Rusch, S. G. Tringe, C. Bailey, R. M. Jennings, and W. P. Inskeep.** 2013. Predominant Acidilobus-Like Populations from Geothermal Environments in Yellowstone National Park Exhibit Similar Metabolic Potential in Different Hypoxic Microbial Communities, p. 294-305, *Appl Environ Microb*, vol. 80.
25. **John, S. G., C. B. Mendez, L. Deng, B. Poulos, A. K. M. Kauffman, S. Kern, J. Brum, M. F. Polz, E. A. Boyle, and M. B. Sullivan.** 2010. A simple and efficient method for concentration of ocean viruses by chemical flocculation, p. 195-202, *Environmental Microbiology Reports*, vol. 3.
26. **Kim, T. K., S. M. Thomas, M. Ho, S. Sharma, C. I. Reich, J. A. Frank, K. M. Yeater, D. R. Biggs, N. Nakamura, R. Stumpf, S. R. Leigh, R. I. Tapping, S. R. Blanke, J. M. Slauch, H. R. Gaskins, J. S. Weisbaum, G. J. Olsen, L. L. Hoyer, and B. A. Wilson.** 2009. Heterogeneity of Vaginal Microbial Communities within Individuals, p. 1181-1189, *J Clin Microbiol*, vol. 47.
27. **Koonin, E. V., and V. V. Dolja.** 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit Rev Biochem Mol Biol* 28:375-430.
28. **Labonté, J. M., and C. A. Suttle.** 2013. Previously unknown and highly divergent ssDNA viruses populate the oceans., *The ISME Journal*.
29. **Leamon, J. H., W. L. Lee, K. R. Tartaro, J. R. Lanza, G. J. Sarkis, A. D. deWinter, J. Berka, and K. L. Lohman.** 2003. A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions, p. 3769-3777, *Electrophoresis*, vol. 24.
30. **Lowenstern, J. B., U. o. Utah, and G. Survey.** 2005. Steam Explosions, Earthquakes, and Volcanic Eruptions: What's in Yellowstone's Future? U.S. Geological Survey.
31. **Macur, R. E., Z. J. Jay, W. P. Taylor, M. A. Kozubal, B. D. Kocar, and W. P. Inskeep.** 2012. Microbial community structure and sulfur biogeochemistry in

mildly-acidic sulfidic geothermal springs in Yellowstone National Park, p. 86-99, *Geobiology*, vol. 11.

32. **Malboeuf, C. M., X. Yang, P. Charlebois, J. Qu, A. M. Berlin, M. Casali, K. N. Pesko, C. L. Boutwell, J. P. DeVincenzo, G. D. Ebel, T. M. Allen, M. C. Zody, M. R. Henn, and J. Z. Levin.** 2013. Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification, *Nucleic Acids Res*, vol. 41.
33. **Marcais, G., and C. Kingsford.** 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, p. 764-770, *Bioinformatics*, vol. 27.
34. **Mathieu Bastian, S. H. M. J.** 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks, p. 1-2.
35. **Mochizuki, T., M. Krupovič, G. Pehau-Arnaudet, Y. Sako, P. Forterre, and D. Prangishvili.** 2012. Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome., p. 13386-13391, *Proc. Natl. Acad. Sci. U.S.A.*, vol. 109.
36. **Mochizuki, T., Y. Sako, and D. Prangishvili.** 2011. Provirus Induction in Hyperthermophilic Archaea: Characterization of *Aeropyrum pernix* Spindle-Shaped Virus 1 and *Aeropyrum pernix* Ovoid Virus 1, p. 5412-5419, *Journal of Bacteriology*, vol. 193.
37. **Mochizuki, T., T. Yoshida, R. Tanaka, P. Forterre, Y. Sako, and D. Prangishvili.** 2010. Diversity of viruses of the hyperthermophilic archaeal genus *Aeropyrum*, and isolation of the *Aeropyrum pernix* bacilliform virus 1, APBV1, the first representative of the family Clavaviridae, p. 347-354, *Virology*, vol. 402. Elsevier Inc.
38. **Namiki, T., T. Hachiya, H. Tanaka, and Y. Sakakibara.** 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads, p. e155-e155, *Nucleic Acids Res*, vol. 40.
39. **Ninomiya, M., Y. Ueno, R. Funayama, T. Nagashima, Y. Nishida, Y. Kondo, J. Inoue, E. Kakazu, O. Kimura, K. Nakayama, and T. Shimosegawa.** 2012. Use of Illumina Deep Sequencing Technology To Differentiate Hepatitis C Virus Variants, p. 857-866, *J Clin Microbiol*, vol. 50.
40. **Oke, M., M. Kerou, H. Liu, X. Peng, R. A. Garrett, D. Prangishvili, J. H. Naismith, and M. F. White.** 2010. A Dimeric Rep Protein Initiates Replication of a Linear Archaeal Virus Genome: Implications for the Rep Mechanism and Viral Replication, p. 925-931, *J. Virol.*, vol. 85.
41. **Out, A. A., I. J. H. M. van Minderhout, J. J. Goeman, Y. Ariyurek, S. Ossowski, K. Schneeberger, D. Weigel, M. van Galen, P. E. M. Taschner, C.**

- M. J. Tops, M. H. Breuning, G.-J. B. van Ommen, J. T. den Dunnen, P. Devilee, and F. J. Hes.** 2009. Deep sequencing to reveal new variants in pooled DNA samples. *Human Mutation* **30**:1703-1712.
42. **Peng, X., T. Basta, M. Häring, R. A. Garrett, and D. Prangishvili.** 2007. Genome of the Acidianus bottle-shaped virus and insights into the replication and packaging mechanisms, p. 237-243, *Virology*, vol. 364.
 43. **Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin.** 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, p. 1420-1428, *Bioinformatics*, vol. 28.
 44. **Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin.** 2011. Meta-IDBA: a de Novo assembler for metagenomic data, p. i94-i101, *Bioinformatics*, vol. 27.
 45. **Pietila, M. K., N. S. Atanasova, V. Manole, L. Liljeroos, S. J. Butcher, H. M. Oksanen, and D. H. Bamford.** 2012. Virion Architecture Unifies Globally Distributed Pleolipoviruses Infecting Halophilic Archaea, p. 5067-5079, *J. Virol.*, vol. 86.
 46. **Pietilä, M. K., P. Laurinmäki, D. A. Russell, C.-C. Ko, D. Jacobs-Sera, R. W. Hendrix, D. H. Bamford, and S. J. Butcher.** 2013. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story., *Proc. Natl. Acad. Sci. U.S.A.*
 47. **Pina, M., A. Bize, P. Forterre, and D. Prangishvili.** 2011. The archeoviruses, p. 1035-1054, *FEMS Microbiology Reviews*, vol. 35.
 48. **Prangishvili, D.** 2013. The Wonderful World of Archaeal Viruses, p. 565-585, *Annu. Rev. Microbiol.*, vol. 67.
 49. **Prangishvili, D., R. A. Garrett, and E. V. Koonin.** 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res* **117**:52-67.
 50. **Prangishvili, D., and M. Krupovič.** 2012. A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales", p. 791-795, *Arch Virol*, vol. 157.
 51. **Prangishvili, D., G. Vestergaard, M. Häring, R. Aramayo, T. Basta, R. Rachel, and R. A. Garrett.** 2006. Structural and Genomic Properties of the Hyperthermophilic Archaeal Virus ATV with an Extracellular Stage of the Reproductive Cycle, p. 1203-1216, *J Mol Biol*, vol. 359.
 52. **Quax, T. E. F., S. Lucas, J. Reimann, G. Pehau-Arnaudet, M.-C. Prevost, P. Forterre, S.-V. Albers, and D. Prangishvili.** 2011. Simple and elegant design of

a virion egress structure in Archaea., p. 3354-3359, Proc. Natl. Acad. Sci. U.S.A., vol. 108.

53. **Remmert, M., A. Biegert, A. Hauser, and J. Söding.** 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, *Nature Methods*. Nature Publishing Group.
54. **Reyes, A., M. Haynes, N. Hanson, F. E. Angly, A. C. Heath, F. Rohwer, and J. I. Gordon.** 2010. Viruses in the faecal microbiota of monozygotic twins and their mothers., p. 334-338, *Nature*, vol. 466.
55. **Reyes, A., N. P. Semenkovich, K. Whiteson, F. Rohwer, and J. I. Gordon.** 2012. Going viral: next-generation sequencing applied to phage populations in the human gut, p. 607-617, *Nature Publishing Group*, vol. 10. Nature Publishing Group.
56. **Reysenbach, A. L., G. S. Wickham, and N. R. Pace.** 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **60**:2113-2119.
57. **Rice, G., K. Stedman, J. Snyder, B. Wiedenheft, D. Willits, S. Brumfield, T. McDermott, and M. J. Young.** 2001. Viruses from extreme thermal environments. *Proc Natl Acad Sci U S A* **98**:13341-13345.
58. **Rice, G., L. Tang, K. Stedman, F. Roberto, J. Spuhler, E. Gillitzer, J. E. Johnson, T. Douglas, and M. Young.** 2004. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life, p. 7716-7720, *Proceedings of the National Academy of Sciences*, vol. 101. National Acad Sciences.
59. **Roine, E., P. Kukkaro, L. Paulin, S. Laurinavicius, A. Domanska, P. Somerharju, and D. H. Bamford.** 2010. New, Closely Related Haloarchaeal Viral Elements with Different Nucleic Acid Types, p. 3682-3689, *J. Virol.*, vol. 84.
60. **Roux, S., F. Enault, A. Robin, V. Ravet, S. Personnic, S. Theil, J. Colombet, T. Sime-Ngando, and D. Debroas.** 2012. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics, p. e33641, *PLoS ONE*, vol. 7.
61. **Roux, S., M. Faubladier, A. Mahul, N. Paulhe, A. Bernard, D. Debroas, and F. Enault.** 2011. Metavir: a web server dedicated to virome analysis, p. 3074-3075, *Bioinformatics*, vol. 27.
62. **Runckel, C., M. L. Flenniken, J. C. Engel, J. G. Ruby, D. Ganem, R. Andino, and J. L. DeRisi.** 2011. Temporal Analysis of the Honey Bee Microbiome

Reveals Four Novel Viruses and Seasonal Prevalence of Known Viruses, Nosema, and Crithidia, p. e20656, PLoS ONE, vol. 6.

63. **Snyder, J. C., S. K. Brumfield, K. M. Kerchner, T. E. F. Quax, D. Prangishvili, and M. J. Young.** 2013. Insights into a viral lytic pathway from an archaeal virus-host system., p. 2186-2192, J. Virol., vol. 87.
64. **Soding, J., A. Biegert, and A. N. Lupas.** 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res **33**:W244-W248.
65. **Stern, A., E. Mick, I. Tirosh, O. Sagy, and R. Sorek.** 2012. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome, p. 1985-1994, Genome Research, vol. 22.
66. **Thurber, R. V., M. Haynes, M. Breitbart, L. Wegley, and F. Rohwer.** 2009. Laboratory procedures to generate viral metagenomes, p. 470-483, Nat Protoc, vol. 4.
67. **Treangen, T. J., D. D. Sommer, F. E. Angly, S. Koren, and M. Pop.** 2002. Next Generation Sequence Assembly with AMOS. John Wiley & Sons, Inc.
68. **van den Brand, J. M. A., M. van Leeuwen, C. M. Schapendonk, J. H. Simon, B. L. Haagmans, A. D. M. E. Osterhaus, and S. L. Smits.** 2012. Metagenomic Analysis of the Viral Flora of Pine Marten and European Badger Feces, p. 2360-2365, J. Virol., vol. 86.
69. **Willner, D., M. Furlan, R. Schmieder, J. A. Grasis, D. T. Pride, D. A. Relman, F. E. Angly, T. McDole, J. Mariella, Ray P, and F. Rohwer.** 2011. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity, p. 4547-4553, Proceedings of the National Academy of Sciences, vol. 108. National Acad Sciences.
70. **Wommack, K. E., J. Bhavsar, and J. Ravel.** 2008. Metagenomics: Read Length Matters, p. 1453-1463, Appl Environ Microb, vol. 74.
71. **Wu, Z., X. Ren, L. Yang, Y. Hu, J. Yang, G. He, J. Zhang, J. Dong, L. Sun, J. Du, L. Liu, Y. Xue, J. Wang, F. Yang, S. Zhang, and Q. Jin.** 2012. Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces., p. 10999-11012, J. Virol., vol. 86.
72. **Yang, X., P. Charlebois, S. Gnerre, M. G. Coole, N. J. Lennon, J. Z. Levin, J. Qu, E. M. Ryan, M. C. Zody, and M. R. Henn.** 2012. De novo Assembly of highly diverse viral populations, p. 475, BMC Genomics, vol. 13. BioMed Central Ltd.

73. **Yoshida, M., Y. Takaki, M. Eitoku, T. Nunoura, and K. Takai.** 2013. Metagenomic Analysis of Viral Communities in (Hado)Pelagic Sediments, p. e57271, PLoS ONE, vol. 8.

CONCLUDING REMARKS

Viruses play an important role on Earth, affecting all cellular life forms. As the most abundant biological entities on the planet, they outnumber cells by at least 15-fold. Their lysis of prokaryotes in the ocean is believed to play a major role in the global carbon cycle, with evidence supporting their affect of nitrogen cycling as well. Viruses are a major source of genetic diversity and can serve as a means of gene transfer between organisms, affecting microbial diversity. Viruses are also believed to be a major driving source of microbial evolution. Yet despite their wide-reaching impact, the viruses infecting the *Archaea* are the least studied. Despite more than 6300 prokaryotic described, only ~100 of those are from the *Archaea*. None of the archaeal viruses is RNA, and only two of them are single-stranded DNA viruses. One of the major reasons for this difference lies in the difficulty of culturing members of the *Archaea*, with many requiring challenging/difficult culturing conditions. Another aspect is the fact that much of the biology and biochemistry of *Archaea* is simply unknown. It comes as no surprise that nearly all archaeal viruses described originate from a successfully cultured host. As the most common method for screening for new viruses, host members are cultured using various enrichment medias and, if culture establishment is successful, the media is screen for virus-like particles (VLPs). This culture-dependence methodology is the single-greatest methodological limitation to studying archaeal viruses. In essence, if it cannot be cultured, no new viruses can be found. This is a particularly troubling problem, especially in cases where EM studies reveal a large assortment of VLPs directly observed in

environmental samples that are not typically observed in enrichment cultures. Archaeal viruses, despite being described, are far from the same level of characterization as many of their bacterial or eukaryotic counterparts. This discrepancy in characterization is due to many factors, such as the extreme environmental conditions required by their hosts and the challenge of developing successful host-virus model systems. To overcome the limitation of culturing, culture-independent methods were developed that allowed the study of viruses regardless of host, ushering in the era of viral metagenomics. Within the first few years, it became clear our understanding of viruses was more lacking than previously thought. The vast diversity of viruses in the oceans, their impact on global nutrient cycling, the role they play as drivers of microbial evolution – all were found through viral metagenomics. The impact of new sequencing technologies such as pyrosequencing and Illumina only expanded and accelerated viral metagenomics. Viral metagenomes increased in scale, the amount of data generated, the depth and coverage of their targeted viral genomes. A veritable shift in the bottleneck of virus discovery and understanding changed from the ability in sequencing viruses to bioinformatic analysis. In spite of these discoveries, viral metagenomics has not been applied to archaeal viruses in extreme environments with the same vigor as their oceanic, freshwater or soil counterparts.

This body of work sought to apply viral metagenomics to the archaeal-dominated, hot springs of Yellowstone National Park (YNP). These low pH, high temperature environments favor the *Archaea* and coincidentally provide a simple, tractable system to study archaeal viruses and their hosts. Conversely, the combination of extreme environmental conditions and simple microbial background yields a low biomass system,

presenting a methodological challenge. In essence, this work sought to describe the archaeal virus community in terms of enhanced dynamic resolution and provide a more accurate description than previously available and search for the first known archaeal RNA virus. Chapter 2 describes the first viral metagenomes (“viromes”) generated from a high temperature, acidic hot spring. Hot spring site selection for metagenomic analysis was directed by a novel reverse-transcriptase incorporation assay that determined the amount of RNA in the viral fraction. Sites with high levels of RNA were selected for subsequent metagenomic analysis, alongside paired DNA metaviromes and cellular metagenomes. The cellular metagenomes confirmed the springs were archaeal-dominated and possessed a simple community structure, following a pattern seen in previous studies. Both DNA and RNA metaviromes were found to be essentially unknown, with less than half of either matching to sequences in extant databases. Examination of the RNA metavirome revealed one near full-length RNA virus sequence, coding for a single polyprotein, complete with a capsid protein and the hallmark of all positive-sense, single-stranded RNA viruses – an RNA-dependent RNA polymerase (RdRp). Closer examination of the RdRp showed conservation of the major motifs, despite having only an 11% identity to its closest homolog. Phylogenetic reconstructions placed the RdRp in a distinct lineage separate from both bacterial and eukaryotic viruses. Also found during analysis of this sequence was an autoproteolytic site containing the same conserved residues as those present in nodaviruses and tectiviruses. This site is essential for proteolytic cleavage of the peptide to form the capsid protein precursor. It is remarkable to note that these two viral families also contain a long polyprotein in their genome structure. Also found during the RNA virome analysis was an additional, shorter RdRp-

containing sequence that was similar, but clearly distinct from the near full-length viral sequence.

Describing viral communities is currently one of the most challenging aspects of viral metagenomics. A large number of fragmented genomes resulting from viral assemblies combined with uneven sequencing depth, biases in amplification and sequencing and biases in sampling all contribute to obfuscating community structures. The near absence of sequence similarity of archaeal viruses to extant databases makes identification of the fragmented genomes impossible in many cases. This complicates phylogenetic analysis and, in many cases, removes such genomes from other downstream analyses, further reducing the detail in describing the viral community. Chapter 3 sought to tackle this problem by applying a network-based analysis to describe the viral communities initially outlined in Chapter 2. This approach utilized an all-verses-all BLAST comparing all contigs against themselves – with contigs as nodes and their edges annotated as/by the high scoring pair information. A community detection algorithm that identified clusters of sequences using the information of the edges then analyzed the network. This was able to clearly separate eighty-two viral populations from the larger community, enabling for the first time a measure on the number of particular viral groups present in an environment (more below). This also allowed for a detailed temporal investigation of the DNA viral community revealing patterns of stability of most viral populations and change for the smaller viral groups. Due to the temporal nature of the sampling, it was possible to visualize the number of sequences represented at any time point for any given viral group. Chapter 3 also took advantage of new developments in metavirome analysis pipelines, revealing the vast majority (> 97%) of classifiable

sequences as matching archaeal viruses. This mirrored the composition of the cellular metagenome, supporting the relationship and interaction between viral and cellular communities. Most remarkable was the “seeding” of all known viruses against the viral network, showing that only archaeal viruses matched to specific viral clusters – with the matching viruses corresponding to those identified by the metavirome analysis pipelines. This led to the hypothesis that despite only a small fraction of the viral clusters having a matching virus, the other clusters were entirely unknown archaeal virus families not yet described. A visual representation of the network was the first such attempt in the literature to organize viral populations and move beyond the over-simplified community characterizations using abstract numbers. Finally, the network and its viral groups also showed that they could be re-assembled into larger sequences, and in some cases, full-length versions of their reference counterparts. This suggested that the overall sequencing depth at any particular time point was insufficient to fully sequence any one viral genome, but in association with the network analysis and subsequent clustering of viral groups was sufficient to overcome this limitation and allow full genome regeneration.

With the foreknowledge that previous work revealed inadequate sequencing depth only compensated by a time series and novel network analysis – it was hypothesized greater sequencing at a specific time point would reveal a similar viral community structure. In chapter 4, an analysis of these viral communities applied advancements in sequencing technologies and sampling methods, generating 150x more data than results presented in chapters 2 and 3. To improve virus concentration, a chemical flocculation method used to remove viruses during wastewater treatment was exploited. To enhance purity, cesium-purified viral fractions instead of filter purification was used in isolating

viruses. The first of the next-generation sequencing (NGS) platforms, 454 pyrosequencing, used in chapters 2 and 3, was replaced with the most recent version of the Illumina sequencing technology. This upgrade dramatically increased sequencing depth and represents the first time Illumina's use in a metavirome context. The previous network analysis is again used to characterize both DNA and RNA metaviromes, finding a dramatically similar number of viral communities as previously seen in chapter 3. In addition, seeding the network with all known viruses revealed a similar profile as before – with only archaeal viruses matching, with each virus family matching to a single group. Similarly, re-assembly of the viral groups also re-assembles full-length viral genomes. Taken together, the results of chapter 4 serve to confirm the hypothesis that viral community is relatively stable in the extreme environments of Yellowstone and that sufficiently deep sequencing of a single time point is adequate to re-create the same network as seen with the temporal sampling. A more detailed analysis of the RNA metavirome revealed additional RdRp-containing sequences, suggesting the RNA virus community was also stable in this environment, yet at an incredibly low abundance compared to the viral DNA community. These RdRps were clearly separate from those described in chapter 2, suggesting that the community may fluctuate and change more dramatically than its DNA counterpart.

Collectively, the results in this dissertation provide new insights into the characterization of archaeal virus communities and provide the first evidence of archaeal RNA viruses. Analysis of these communities indicates the presence of nearly 82 distinct viral groups in the DNA metaviromes and 23 in the RNA metaviromes. Many of these groups are entirely uncharacterized and present new avenues for research. The network-

analysis based approach offers the field of virology a new way to examine viral communities, by not only grouping virus populations but provide the ability to study temporal and spatial dynamics. The findings of RNA virus hallmark genes in an archaeal-dominated hot spring sampled over the course of years by multiple different approaches is circumstantial evidence for the presence of archaeal RNA viruses. The work provided here is the first serious attempt at searching for archaeal RNA viruses.

Future research into the hosts of these RNA viruses would provide the required evidence supporting them as the first described archaeal RNA viruses. A virus-based, fluorescent in-situ hybridization (FISH) could link segments of viral genomes against host-specific (i.e. identifying) genes. Hosts identified through this method could also be sorted through the use of fluorescence-activated cell sorting (FACS) – a specialized form of flow cytometry. If these RNA viruses indeed have archaeal hosts, these findings would have considerable impact on the origin and evolution of RNA viruses, and by extension, viruses as a whole. Additional future research into the community structure would involve design of virus group-specific probes for quantitative PCR (qPCR) and tracking of each virus group at various timescales over the course of days, weeks, months, or longer. These studies would allow, for the first time, detailed quantitative measurements of viral abundances of viruses neither cultured nor previously described. The advantage of the network analysis and clustering of viral sequences into specific family-level groupings enables further experimental avenues, such as qPCR design, of entirely novel viruses. Although not described at detail in this dissertation, clustered regularly interspaced short palindromic repeats (CRISPs) are an acquired immune system found in prokaryotes that contain sequences derived from virus challenges, essentially “a history

of past viral infections.” These sequences, which can be identified from the cellular metagenomes, can be linked to specific host genera. By comparing these CRISPR-derived sequences against the viral sequences found in the network, one can link a viral group to specific host genera. Future work would seek to exploit this association, and provide an additional means of linking virus to host.

The network approach described herein offers the virology community a new way to analyze metaviromes and their communities. In addition to organizing viruses – both known and unknown – into groups, providing a means to track temporal and spatial dynamics, and allow for re-assembly into full-length viral genomes, the networks lays the groundwork for more sophisticated types of analysis. The emergence of NGS has resulted in an unprecedented amount of data whose exponential growth shows no signs of slowing down. Dealing with such vast amounts of data presents a daunting task for the field – a challenge that can be overcome through the use of previously established methodologies originating in another field, such as the sophisticated network partitioning algorithm used in this dissertation. Networks (and the field of graph theory) – such as the ones used herein – can easily scale to the levels of sequencing now seen with current technologies. By virtue of their nature, these networks can also store more than just BLAST information to connect one sequence to another; they can be annotated with several layers of additional information. Each layer can be organized further to reveal higher levels of order that may show global patterns of community dynamics or allow for other analytical network algorithms to be applied. Neither case is currently possible using traditional metagenomic analysis pipelines. The fundamental problems of environmental virology, the vast “unknown” sequences and the computational “nightmare” of assembling

fragments of a highly diverse viral community can be partially addressed using the above methods. The unprecedented amount of viral diversity and the complex relationship with their hosts further complicates a complete elucidation of a viral community in the natural environment. Higher resolution of viral communities, i.e. a more detailed understanding of the community, can be obtained with deeper sequencing and pairing cellular metagenomes. The major factors limiting the resolution of the networks presented in this dissertation was the sequencing depth and computational power available at the time of analysis. Both factors will be quickly overcome given the quickening pace of advancements in the field. The quiet revolution in the field of environmental virology, beginning with the application of NGS to viral metagenomics, will only be accelerated through the application of third-generation NGS and network-based approaches, such as the ones presented here, and the development of more powerful annotation pipelines. In closing, future work should seek to marry sequence-similarity (nucleotide and protein) networks and sequence-annotations, and combine it with abundance-based measurements to gain an unparalleled insight into dynamics of viral communities and understanding of its members. The entirety of this dissertation sought to lay the groundwork and take the first steps into this next generation of environmental virology, by enhancing virus isolation protocols, providing a new way to analyze viral metagenomic data, and showing a glimpse into the world of RNA viruses in extreme environments, presenting experimental evidence of how archaeal RNA viruses may be direct ancestors of their eukaryotic counterparts.

APPENDICES

APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER 2

Chapter 2:

Published: Journal of Virology, **2012**, 86, 5562-5573, doi:10.1128/JVI.07196-11

Further Supporting materials are available free of charge via the Internet at <http://jvi.asm.org/content/86/10/5562/suppl/DC1> . This material contains additional methods and results sections as well as figures and tables.

Table S1. Location of hot springs surveyed in Yellowstone National Park, USA.

Sample Name	Lat	Long	Date	pH	Temperature (°C)
Amp-21	N44 48.06666	W110 43.72332	Aug-09	2.0	76
Amp-22	N44 47.82834	W110 43.46664	Aug-09	2.0	90
Mon-1	N44 41.09832	W110 45.24498	Aug-09	1.5	86
Mon-2	N44 41.09502	W110 45.22998	Aug-09	1.5	75
Mon-3	N44 42.755	W110 45.22998	Aug-09	1.5	70
Mon-4	N44 42.74833	W110 45.23334	Aug-09	1.5	87
Mon-5	N44 41.79498	W110 45.22998	Aug-09	1.5	79
Mon-6	N44 41.07666	W110 45.23502	Aug-09	1.5	79
Mon-7	N44 41.055	W110 45.22002	Aug-09	1.5	69
NL-10	N44 45.214	W110 43.432	Aug-09	4.5	93
NL-17	N44 45.11832	W110 43.70832	Aug-09	2.0	81
NL-18	N44 45.12666	W110 43.71336	Aug-09	2.0	83
NL-19	N44 45.12168	W110 43.71834	Aug-09	2.0	78
NL-20	N44 45.12834	W110 43.72164	Aug-09	3.0	85
SL-1	N44 21.45167	W110 47.61333	Aug-09	6.0	83
SL-2	N44 21.38	W110 47.77666	Aug-09	3.0	89
SL-3	N44 21.225	W110 47.84	Aug-09	2.5	83
SL-4	N44 21.01	W110 47.35016	Aug-09	5.0	80

SL-5	N44 21.1	W110 47.99833	Aug-09	4.0	81
SL-6	N44 21.1	W110 47.98	Aug-09	5.0	85
Syl-8	N44 41.99502	W110 45.91998	Aug-09	3.5	77
Syl-9	N44 41.95836	W110 46.03836	Aug-09	5.0	82
Syl-10	N44 41.96832	W110 46.07832	Aug-09	2.0	86
Syl-11	N44 43.63666	W110 46.09002	Aug-09	2.0	87
Syl-12	N44 41.95836	W110 46.11666	Aug-09	1.5	80
Syl-13	N44 41.93334	W110 46.10832	Aug-09	5.0	81
Syl-14	N44 42	W110 46.24002	Aug-09	2.0	73
Syl-15	N44 42.03498	W110 45.96666	Aug-09	4.5	85

TABLE S2. PCR primers used to confirm selected contigs from the RNA virus metagenomes.

Name	Sequence	Start	End	Length
Contig00002 2192F	TTTAGCGGTCACGCCGGCTG	2192	2211	20
Contig00002 2756R	CGCTCCAGCTCACGTCGTGG	2737	2756	20
Contig00009 702F	ATTGTGAGCTACGCGCGGGG	702	721	20
Contig00009 1337R	GGCGAGGAAACCCGTGTGCT	1318	1337	20

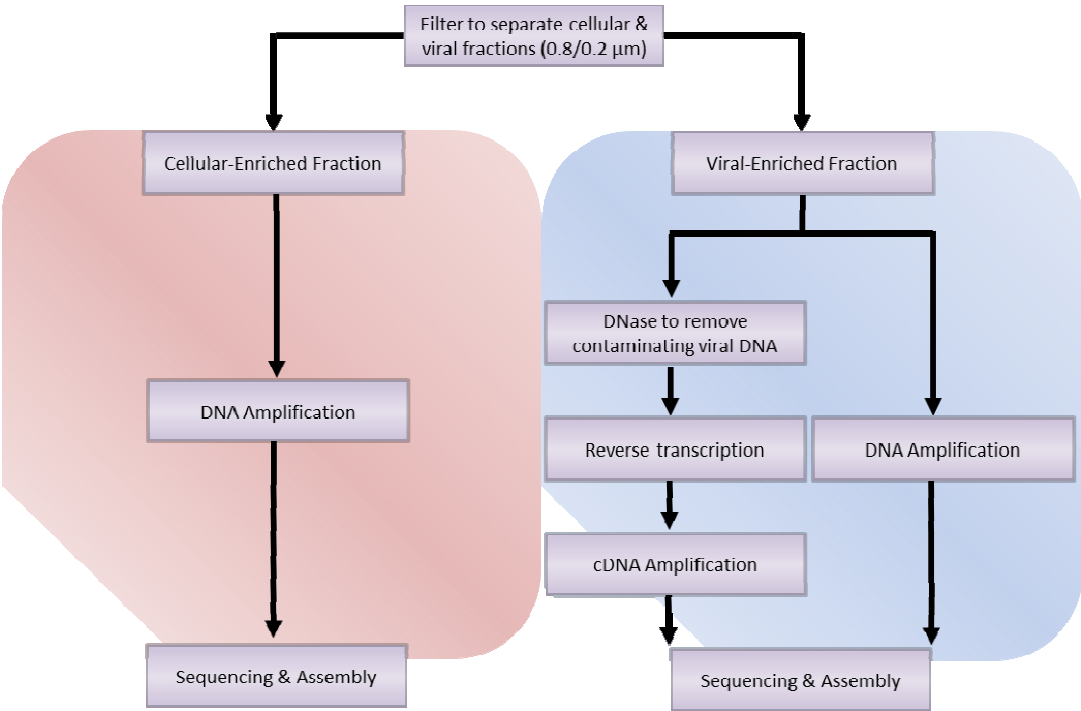


FIG.

S1. Overview of the experimental flow to produce viral RNA enriched, viral DNA enriched, and cellular enriched metagenomes.

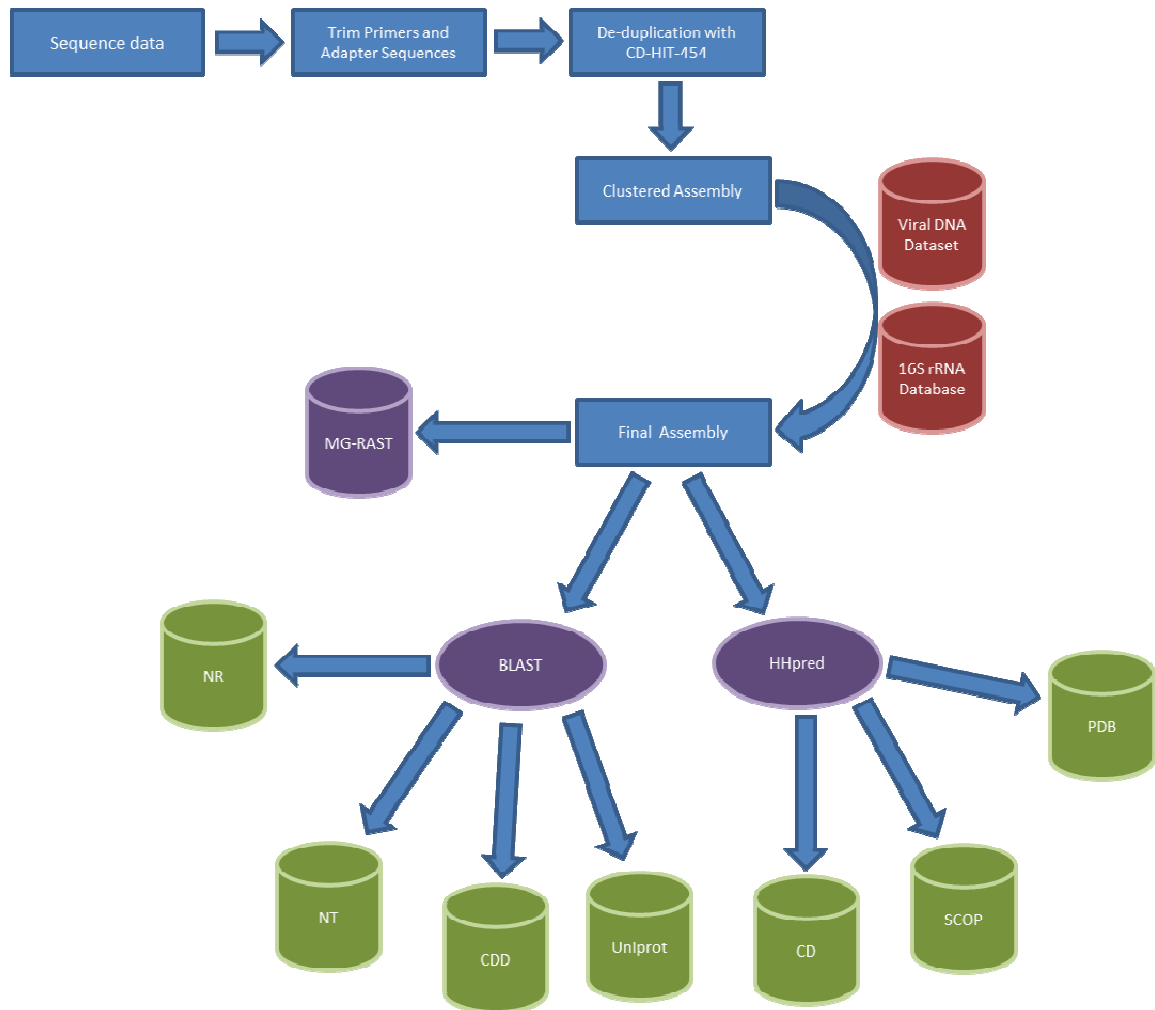


FIG. S2. Metagenomic viral RNA assembly outline. Raw sequence data was processed through a series of quality checks and redundant sequences are removed. Assembled contigs were compared against the paired viral DNA metagenomes and 16S rRNA databases. Matching contigs were removed from further analysis. Contigs are compared against a number of databases using a NCBI BLAST, MG-RAST and selected contigs against HHpred.

Statistical tests on constrained tree topologies with different positions of the putative archaeal virus RdRp. Likelihood and p-values for constrained trees (see Methods).

Tree #	LL	Δ LL	p
Unconstrained	-67716.4	n/a	>0.05
1	-67733.8	-17.5 $\square$ 16.2	>0.05
2	-67717.8	-1.5 $\square$ 13.9	>0.05
3	-67732.1	-15.7 $\square$ 18.6	>0.05
4	-67742.6	-26.3 $\square$ 20.4	>0.05
5	-67731.2	-14.8 $\square$ 19.3	>0.05
6	-67728.9	-12.5 $\square$ 14.0	>0.05
7	-67735.7	-19.3 $\square$ 20.5	>0.05
8	-67739.2	-22.8 $\square$ 20.0	>0.05
9	-67731.3	-14.9 $\square$ 18.7	>0.05
10	-67718.9	-2.6 $\square$ 11.5	>0.05

LL – Log Likelihood

Δ LL – Log Likelihood difference compared to the unconstrained tree p – p-value

Tree topologies in Newick format

APPENDIX B

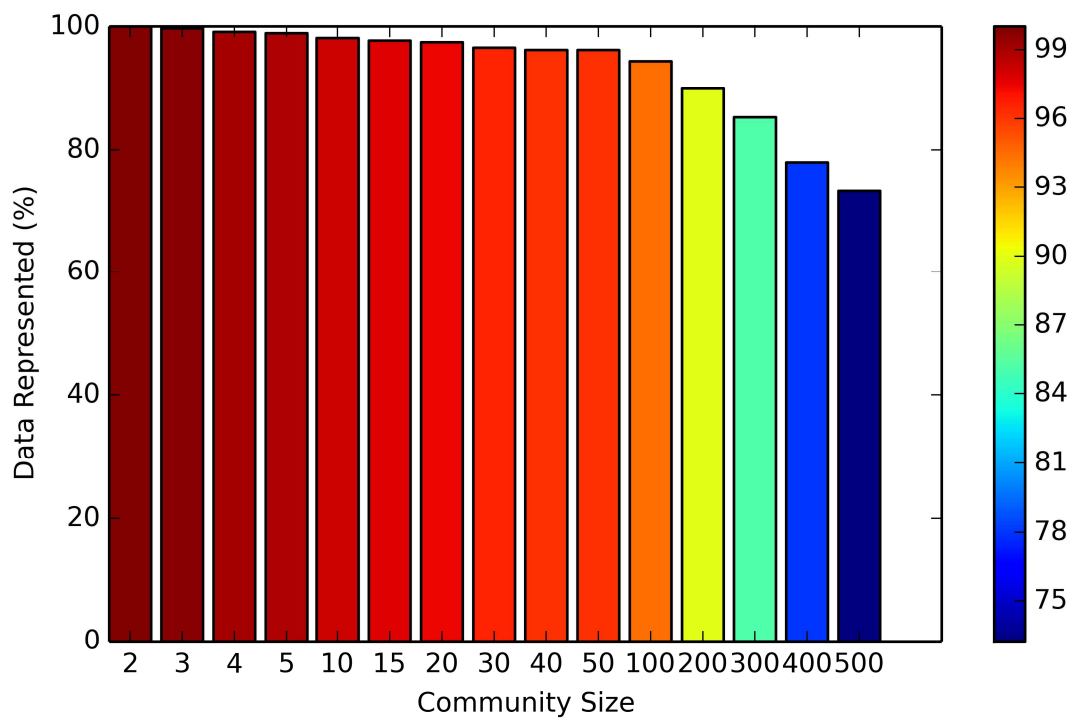
SUPPORTING INFORMATION FOR CHAPTER 3

Chapter 3:

Prepared

This contains 1 figure and provides an extensive description of materials and methods.

Supplemental Figure 1. Membership Cutoff Values Based on Percentage of Data Used



Supplemental Methods

Network construction. The viral network is initially constructed from an all-verses-all BLAST of assembled viral contigs. The resulting BLAST file contains information of all queries (initial or “originating” contigs) and their matches (final or “target” contigs) with the HSP (high-scoring pair) serving as the relationship between the query and target. A script was developed that first evaluated each HSP to pass user-defined criteria of e-value, length of the HSP and the overall identity of the match – as a filter to remove superfluous and statistically insignificant connections. It is possible to extend the criteria to any number of conditions and is only limited by the information stored within the BLAST file. Other conditions could also be added from third-party or other external analysis programs, such as filtering contigs that only have a match to another database. After filtering, the contigs and their HSPs are converted into a network graph. The contigs serve as nodes while the HSP information is stored within the edges, connecting the source node (original or query contig) to the target node (target or final contig). It is possible to store all HSP information within each edge’s “properties,” which can be subsequently queried by network algorithms or other analysis tools. Network algorithms in clustering and partitioning frequently utilize the most important property, the edge *weight*. For the network in this study, the edge weight is the log of the e-value, unless in cases where the e-value = 0, the weight is converted to 300. The e-value was chosen because it is the only way to compare connection strengths between contigs in a statistically meaningful manner. All filtered HSPs and their query and target contigs were converted to a network graph, meaning any number of connections could be created between any two contigs – given that it contains an HSP between them. Following

network creation, the Louvain method is applied to the network. This method is one of the most widely used community detection methods available and scales to 100 million nodes and billions of connections. This greedy optimization method seeks to optimize the modularity of partitions (or groups, clusters) within the network. Modularity measures the strength of divisions between resulting clusters – with high values associated for clusters with many connections within the cluster, but sparse between clusters. During this two-phase optimization, edge's weights are taken into account, and thus highly significant matches connecting two contigs are given greater prominence. It is therefore possible for multiple contigs to be connected through the pairwise comparisons of BLAST and the Louvain method to place the contigs in separate “communities” (i.e. clusters or groups) due to the strengths those contigs may have with other, more “related” sequences. Given sufficient sequencing and coverage of target viral genomes, this method is capable of differentiating individual members of a viral family, despite the fact multiple shared genes exist between the members. On a more global level, this method can successfully partition an entire viral community into its constituent members (this is elaborated within the main manuscript).

Subsequent analyses (describe within the main manuscript) are then applied after partitioning/clustering. Visualization is unnecessary for these analyses, but is performed to allow the researcher to “see” the community structure. To visualize, the force-directed layout algorithm, OpenOrd, implemented in the Gephi visualization program, is then applied to the weighted network. (OpenOrd is based on Fruchterman Rheingold, a classic force-vector) This algorithm assists in distinguishing clusters by using the edge weights and optimally placing contigs close to each other having strong weights, and further from

each other having weak weights. After this layout-partitioning algorithm is finished, another force-directed layout algorithm, Force Atlas 2 is used. This essentially tightens the clusters and imposes a gravity-based limit to the overall spread of the network. This is necessary due to the large number of sequences within the network and the repulsive effect between contigs. It also serves to “fine-tune” the clusters previously established during OpenOrd.

APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER 4

Chapter 4:

Prepared

This contains 4 figures.

Figure 1. Viral groups of the DNA metaviromes, defined by the network analysis, and their rank abundance. Those smaller than 100 members are highlighted in red.

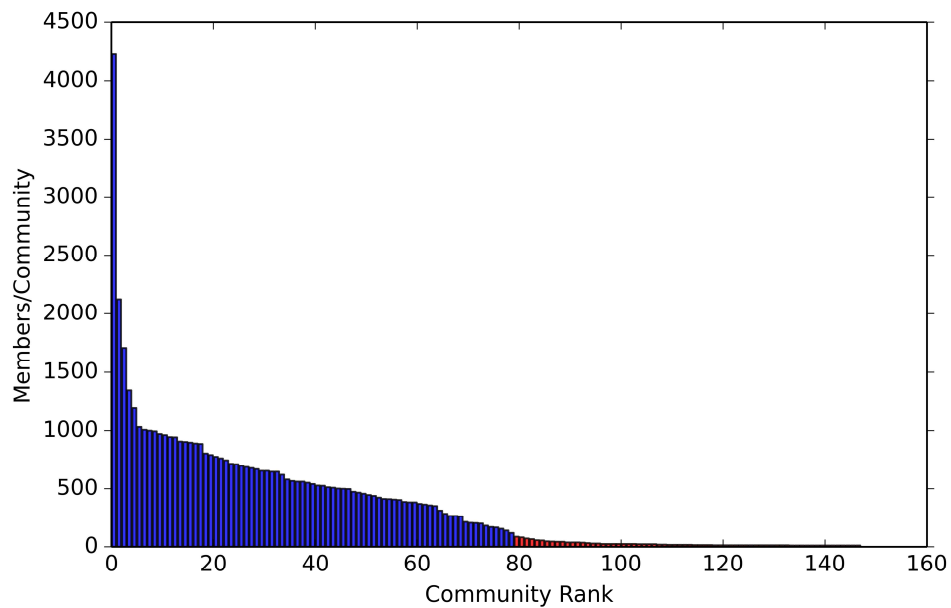


Figure 2. Viral groups of the RNA metaviromes, defined by the network analysis and their rank abundance. Those smaller than 100 members are highlighted in red.

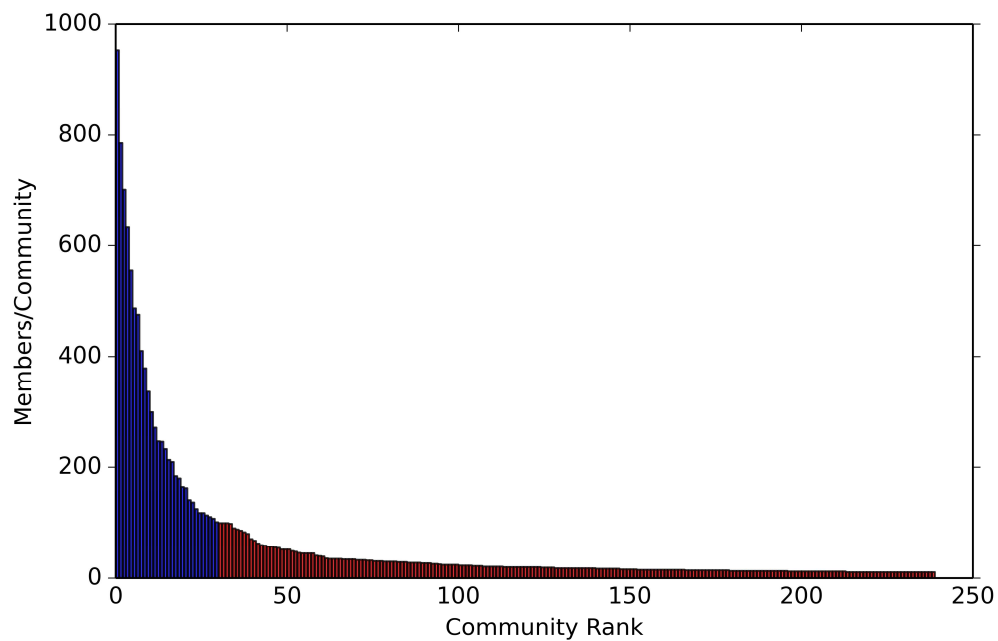


Figure 3. The percentage of data represented as a function of community size in the DNA metaviromes.

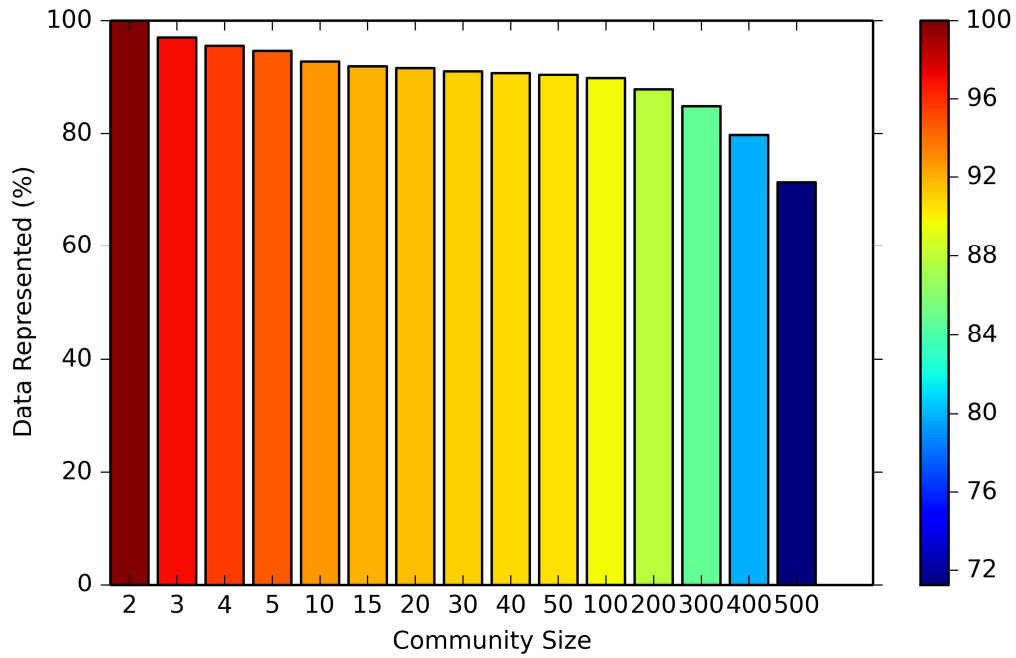
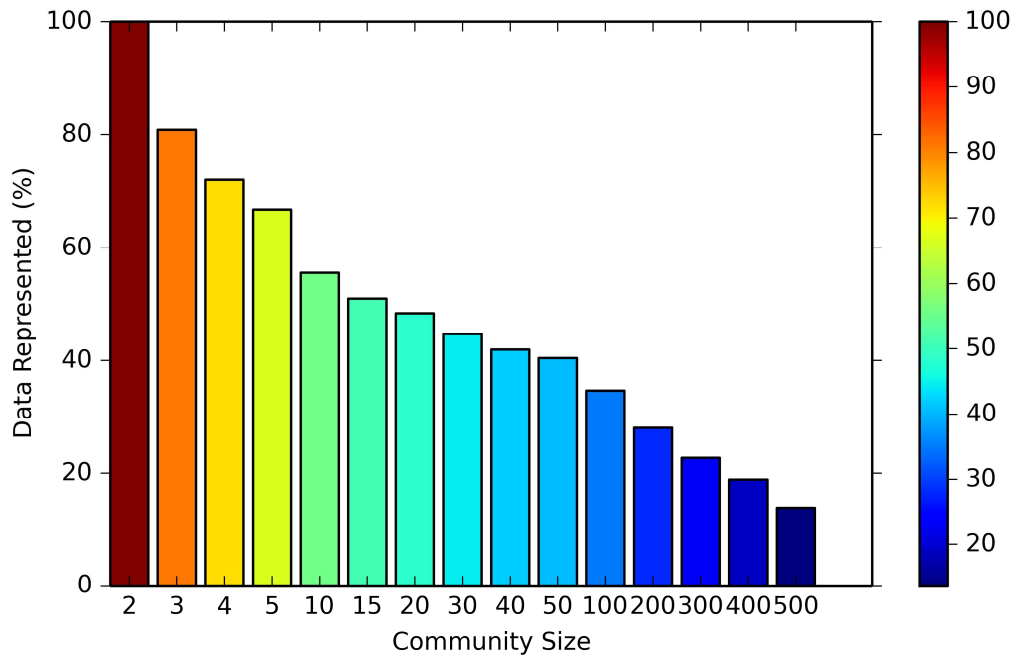


Figure 4. The percentage of data represented as a function of community size in the RNA metaviromes.



LITERATURE CITED

Abascal F, Zardoya R, Posada D (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: 2104-2105.

Ackermann HW (2007). 5500 Phages examined in the electron microscope. *Arch Virol* **152**: 227-243.

Ackermann HW, Prangishvili D (2012). Prokaryote viruses studied by electron microscopy. *Arch Virol*. pp 1843-1849.

Adessi C, Matton G, Ayala G, Turcatti G, Mermod J-J, Mayer P *et al* (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research*. Oxford Univ Press. pp e87-e87.

Ahn D-G, Kim S-I, Rhee J-K, Kim KP, Pan J-G, Oh J-W (2006). TTSV1, a new virus-like particle isolated from the hyperthermophilic crenarchaeote *Thermoproteus tenax*. *Virology*. pp 280-290.

Albers S-V, Szabó Z, Driessen AJM (2006). Protein secretion in the Archaea: multiple paths towards a unique cell surface. *Nature Publishing Group*. pp 537-547.

Allers T, Mevarech M (2005). Archaeal genetics — the third way. *Nature Reviews Genetics*. pp 58-73.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic Local Alignment Search Tool. *J Mol Biol* **215**: 403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Amann R (2000). Who is out there? Microbial Aspects of Biodiversity. *Syst. Appl. Microbiol*. Urban & Fischer Verlag. pp 1-8.

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 Bioinformatics*. p 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*. p e1000593.

Arnold HP, Ziese U, Zillig W (2000a). SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon *Sulfolobus*. *Virology* **272**: 409-416.

- Arnold HP, Ziese U, Zillig W (2000b). SNDV, a Novel Virus of the Extremely Thermophilic and Acidophilic Archaeon *Sulfolobus*. *Virology*. pp 409-416.
- Arnold HP, Zillig W, Ziese U, Holz I, Crosby M, Utterback T *et al* (2000c). A Novel Lipothrixvirus, SIFV, of the Extremely Thermophilic Crenarchaeon *Sulfolobus*. *Virology*. pp 252-266.
- Atanasova NS, Roine E, Oren A, Bamford DH, Oksanen HM (2012). Global network of specific virus-host interactions in hypersaline environments. *Environmental Microbiology*. pp 426-440.
- Auchtung TA, Takacs-Vesbach CD, Cavanaugh CM (2006). 16S rRNA Phylogenetic Investigation of the Candidate Division "Korarchaeota". *Applied and Environmental Microbiology*. pp 5077-5082.
- Auguet J-C, Barberán A, Casamayor EO (2009). Global ecological patterns in uncultured Archaea. *The ISME Journal*. Nature Publishing Group. pp 182-190.
- Baker ML, Jiang W, Rixon FJ, Chiu W (2005). Common Ancestry of Herpesviruses and Tailed DNA Bacteriophages. *J. Virol.* pp 14967-14970.
- Baltimore D (1971). Expression of animal virus genomes. *Bacteriol Rev.* pp 235-241.
- Bamford DH, Caldentey J, Bamford JKH (1995). Bacteriophage Prd1: A Broad Host Range Dsdna Tectivirus With an Internal Membrane. In: Karl Maramorosch FAM, Aaron JS (eds). *Advances in Virus Research*. Academic Press. pp 281-319.
- Bamford DH (2003). Do viruses form lineages across different domains of life? *Research in Microbiology*. pp 231-236.
- Barns SM, Fundyga RE, Jeffries MW, Pace NR (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proceedings of the National Academy of Sciences*. pp 1609-1613.
- Barns SM, Delwiche CF, Palmer JD, Pace NR (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 9188-9193.
- Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S *et al* (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709-1712.
- Barzon L, Lavezzo E, Militello V, Toppo S, Palù G (2011). Applications of Next-Generation Sequencing Technologies to Diagnostic Virology. *IJMS*. pp 7861-7884.

Bath C, Dyall-Smith ML (1998). His1, an Archaeal Virus of the Fuselloviridae Family That Infects *Haloarcula hispanica*.

Bath C, Cukalac T, Porter K, Dyall-Smith ML (2006). His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. *Virology*. pp 228-239.

Batty EM, Wong THN, Trebes A, Argoud K, Attar M, Buck D *et al* (2013). A Modified RNA-Seq Approach for Whole Genome Sequencing of RNA Viruses from Faecal and Blood Samples. *PLoS ONE*. p e66129.

Batzoglou S (2002). ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research*. pp 177-189.

Bell SD, Jackson SP (1998). Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends in Microbiology*. Elsevier. pp 222-228.

Benson SD, Bamford JKH, Bamford DH, Burnett RM (2004). Does common architecture reveal a viral lineage spanning all three domains of life? *Molecular Cell*. pp 673-685.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG *et al* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. pp 53-59.

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

Bishop-Lilly KA, Turell MJ, Willner KM, Butani A, Nolan NME, Lentz SM *et al* (2010). Arbovirus Detection in Insect Vectors by Rapid, High-Throughput Pyrosequencing. *PLoS Negl Trop Dis*. p e878.

Biswas A, Ranjan D, Zubair M (2011). Parallelization of MIRA Whole Genome and EST Sequence Assembler. *Workshop on Parallel Algorithms and Software for Analysis of Massive Graphs (ParGraph)*.

Bize A, Peng X, Prokofeva M, MacLellan K, Lucas S, Forterre P *et al* (2008). Viruses in acidic geothermal environments of the Kamchatka Peninsula. *Research in Microbiology*. pp 358-366.

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC *et al* (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**: 209.

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.

- Blöchl E, Rachel R, Burggraf S, Hafenbradl D, Jannasch HW, Stetter KO (1997). *Pyrolobus fumarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. *Extremophiles*. pp 14-21.
- Blomstrom AL, Widen F, Hammer AS, Belak S, Berg M (2010). Detection of a Novel Astrovirus in Brain Tissue of Mink Suffering from Shaking Mink Syndrome by Use of Viral Metagenomics. *Journal of Clinical Microbiology*. pp 4392-4396.
- Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. IOP Publishing. p 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.
- Boucher Y, Kamekura M, Doolittle WF (2004). Origins and evolution of isoprenoid lipid biosynthesis in archaea. *Molecular Microbiology*. pp 515-527.
- Bowers KJ, Wiegel J (2011). Temperature and pH optima of extremely halophilic archaea: a mini-review. *Extremophiles*. Springer. pp 119-128.
- 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 U S A* **99**: 14250-14255.
- Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P *et al* (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220-6223.
- Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P *et al* (2004a). Diversity and population structure of a near-shore marine-sediment viral community. *Proceedings of the Royal Society B: Biological Sciences*. pp 565-574.
- Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P *et al* (2004b). Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**: 565-574.
- Breitbart M, Rohwer F (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**: 278-284.
- Breitbart M, Thompson LR, Suttle CA, Sullivan MB (2007). Exploring the vast diversity of marine viruses. *OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY SOCIETY*-. The Oceanographic Society; 1999. p 135.

Breitbart M (2012). Marine Viruses: Truth or Dare. *Annu. Rev. Marine. Sci.* pp 425-448.

Briani F, Dehò G, Forti F, Ghisotti D (2001). The Plasmid Status of Satellite Bacteriophage P4. *Plasmid* **45**: 1-17.

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008). Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology* **6**: 245-252.

Cann AJ, Fandrich SE, Heaphy S (2005). Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes* **30**: 151-156.

Cavicchioli R (2010). Archaea — timeline of the third domain. *Nature Publishing Group*. Nature Publishing Group. pp 51-61.

Chaban B, Ng SYM, Jarrell KF (2006). Archaeal habitats — from the extreme to the ordinary. *Can. J. Microbiol.* pp 73-116.

Chang J, Weigele P, King J, Chiu W, Jiang W (2006). Cryo-EM Asymmetric Reconstruction of Bacteriophage P22 Reveals Organization of its DNA Packaging and Infecting Machinery. *Structure*. pp 1073-1082.

Chen C, Guo P (1997a). Sequential action of six virus-encoded DNA-packaging RNAs during phage phi29 genomic DNA translocation. *J. Virol.* pp 3864-3871.

Chen C, Guo P (1997b). Magnesium-induced conformational change of packaging RNA for procapsid recognition and binding during phage phi29 DNA encapsidation. *J. Virol.* pp 495-500.

Cheng RH, Reddy VS, Olson NH, Fisher AJ, Baker TS, Johnson JE (1994). Functional implications of quasi-equivalence in a T = 3 icosahedral animal virus established by cryo-electron microscopy and X-ray crystallography. *Structure* **2**: 271-282.

Chevreur B (2005). MIRA: an automated genome and EST assembler. *Ruprecht-Karls University, Heidelberg, Germany*.

Coetzee B, Freeborough M-J, Maree HJ, Celton J-M, Rees DJG, Burger JT (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. *Virology*. Elsevier Inc. pp 157-163.

Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH, Prangishvili D (2008). Exploring the prokaryotic virosphere. *Research in Microbiology* **159**: 306-313.

Compeau PEC, Pevzner PA, Tesler G (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*. Nature Publishing Group. pp 987-991.

Cordey S, Junier T, Gerlach D, Gobbini F, Farinelli L, Zdobnov EM *et al* (2010). Rhinovirus Genome Evolution during Experimental Human Infection. *PLoS ONE*. p e10588.

Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA *et al* (2007). A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*. pp 283-287.

Culley AI, Lang AS, Suttle CA (2003). High diversity of unknown picorna-like viruses in the sea. *Nature* **424**: 1054-1057.

Culley AI (2006). Metagenomic Analysis of Coastal RNA Virus Communities. *Science*. pp 1795-1798.

Culley AI, Lang AS, Suttle CA (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**: 1795-1798.

Daniel R (2004). The soil metagenome – a rich resource for the discovery of novel natural products. *Current Opinion in Biotechnology*. pp 199-204.

Daniels LL, Wais AC (1984). Restriction and modification of Halophage S45 in *Halobacterium*. *Current Microbiology*. Springer. pp 133-136.

Daniels LL, Wais AC (1990). Ecophysiology of Bacteriophage-S5100 Infecting *Halobacterium-Cutirubrum*. *Applied and Environmental Microbiology*. pp 3605-3608.

Daniels LL, Wais AC (1998). Virulence in phage populations infecting *Halobacterium cutirubrum*. *FEMS Microbiology Ecology*. Wiley Online Library. pp 129-134.

Daum B, Quax TEF, Sachse M, Mills DJ, Reimann J, Yildiz O *et al* (2014). Self-assembly of the general membrane-remodeling protein PVAP into sevenfold virus-associated pyramids. *Proc. Natl. Acad. Sci. U.S.A.*

de la Torre JR, Walker CB, Ingalls AE, Könneke M, Stahl DA (2008). Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environmental Microbiology*. pp 810-818.

Dellas N, Lawrence CM, Young MJ (2013). A Survey of Protein Structures from Archaeal Viruses. *Life*.

DeLong EF (1992). Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences*. pp 5685-5689.

DeLong EF (1998). Everything in moderation: archaea as ‘non-extremophiles’. *Curr. Opin. Genet. Dev.* pp 649-654.

- DeLong EF, Pace NR (2001). Environmental diversity of bacteria and archaea. *Systematic Biology*. Oxford University Press. pp 470-478.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* **72**: 5069-5072.
- Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M *et al* (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*. pp 340-343.
- Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ (2009). Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**: e7264.
- Donaldson EF, Haskew AN, Gates JE, Huynh J, Moore CJ, Frieman MB (2010). Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. *J. Virol.* pp 13004-13018.
- Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 8817-8822.
- Duffy S, Shackelton LA, Holmes EC (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**: 267-276.
- Duhaime MB, Deng L, Poulos BT, Sullivan MB (2012). Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environmental Microbiology*. pp 2526-2537.
- Duhaime MB, Sullivan MB (2012). Ocean viruses Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology*. Elsevier. pp 1-5.
- Dutilh BE, Snel B, Ettema TJG, Huynen MA (2008). Signature Genes as a Phylogenomic Tool. *Molecular Biology and Evolution*. pp 1659-1667.
- Dyall-Smith M, Tang S-L, Bath C (2003). Haloarchaeal viruses: how diverse are they? *Research in Microbiology*. pp 309-313.
- Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X *et al* (2010). Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing. *PLoS Pathog*. p e1000896.

- Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.
- Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. pp 2460-2461.
- Edwards RA, Rohwer F (2005). Viral metagenomics. *Nature Reviews Microbiology*. Nature Publishing Group. pp 504-510.
- Effantin G, Boulanger P, Neumann E, Letellier L, Conway JF (2006). Bacteriophage T5 Structure Reveals Similarities with HK97 and T4 Suggesting Evolutionary Relationships. *Journal of Molecular Biology*. pp 993-1002.
- Eilers H, Pernthaler J, Glockner FO, Amann R (2000). Culturability and In Situ Abundance of Pelagic Bacteria from the North Sea. *Applied and Environmental Microbiology*. pp 3044-3051.
- Eiserling F, Pushkin A, Gingery M, Bertani G (1999). Bacteriophage-like particles associated with the gene transfer agent of methanococcus voltae PS. *J. Gen. Virol.* pp 3305-3308.
- Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, Randau L *et al* (2008). A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc. Natl. Acad. Sci. U.S.A.* pp 8102-8107.
- Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF (2012). Dynamic Viral Populations in Hypersaline Systems as Revealed by Metagenomic Assembly. *Applied and Environmental Microbiology*. pp 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*.
- Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY *et al* (2006). Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* **Chapter 5**: Unit 5 6.
- Fancello L, Raoult D, Desnues C (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology*. Elsevier. pp 1-13.
- Fedurco M (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research*. pp e22-e22.
- Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R *et al* (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria,

archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology*. pp 7059-7066.

Fiers W, Contreras R, Haegemann G, Rogiers R, Van de Voorde A, Van Heuverswyn H *et al* (1978). Complete nucleotide sequence of SV40 DNA. *Nature*. pp 113-120.

Forterre P, Brochier C, Philippe H (2002). Evolution of the Archaea. *Theoretical Population Biology*. pp 409-422.

Forterre P, Gribaldo S, Brochier-Armanet C (2007). Natural History of the Archaeal Domain. *Archaea*. Blackwell Publishing Ltd. pp 17-28.

Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proceedings of the National Academy of Sciences*. pp 14683-14688.

Fröls S, Gordon P, Panlilio MA, Schleper C (2007). Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays. *Virology*.

Fuhrman JA (1992). Novel major archaeobacterial group from marine plankton. *Nature* **356**: 148-149.

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X *et al* (2009). The challenges of sequencing by synthesis. *Nature Biotechnology*. Nature Publishing Group. pp 1013-1023.

Garrett RA, Shah SA, Vestergaard G, Deng L, Gudbergdottir S, Kenchappa CS *et al* (2011). CRISPR-based immune systems of the Sulfolobales: complexity and diversity. *Biochem Soc Trans* **39**: 51-57.

Geslin C, Le Romancer M, Erauso G, Gaillard M, Perrot G, Prieur D (2003). PAV1, the First Virus-Like Particle Isolated from a Hyperthermophilic Euryarchaeote, "Pyrococcus abyssi". *Journal of Bacteriology*. pp 3888-3894.

Geslin C, Gaillard M, Flament D, Rouault K, Le Romancer M, Prieur D *et al* (2007). Analysis of the First Genome of a Hyperthermophilic Marine Virus-Like Particle, PAV1, Isolated from *Pyrococcus abyssi*. *Journal of Bacteriology*. pp 4510-4519.

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS *et al* (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**: 1355-1359.

Gohara DW, Crotty S, Arnold JJ, Yoder JD, Andino R, Cameron CE (2000). Poliovirus RNA-dependent RNA polymerase (3D(pol)) - Structural, biochemical, and biological analysis of conserved structural motifs A and B. *J Biol Chem* **275**: 25523-25532.

Gorlas A, Koonin EV, Bienvenu N, Prieur D, Geslin C (2011). TPV1, the first virus isolated from the hyperthermophilic genus *Thermococcus*. *Environmental Microbiology*. pp 503-516.

Goulet A, Blangy S, Redder P, Prangishvili D, Felisberto-Rodrigues C, Forterre P *et al* (2009). Acidianus filamentous virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 21155-21160.

Grimes JM, Burroughs JN, Gouet P, Diprose JM, Malby R, Zientara S *et al* (1998). The atomic structure of the bluetongue virus core. *Nature* **395**: 470-478.

Grissa I, Vergnaud G, Pourcel C (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52-57.

Gruber N, Galloway JN (2008). An Earth-system perspective of the global nitrogen cycle. *Nature*. pp 293-296.

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L *et al* (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**: 945-956.

Hall N (2007). Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* pp 1518-1525.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*. Elsevier. pp R245-R249.

Happonen LJ, Redder P, Peng X, Reigstad LJ, Prangishvili D, Butcher SJ (2010). Familial Relationships in Hyperthermo- and Acidophilic Archaeal Viruses. *J. Virol.* pp 4747-4754.

Haring M, Vestergaard G, Brugger K, Rachel R, Garrett RA, Prangishvili D (2005). Structure and Genome Organization of AFV2, a Novel Archaeal Lipothrixvirus with Unusual Terminal and Core Structures. *Journal of Bacteriology*. pp 3855-3858.

Häring M, Peng X, Brugger K, Rachel R, Stetter KO, Garrett RA *et al* (2004). Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the Globuloviridae. *Virology*. pp 233-242.

Häring M, Rachel R, Peng X, Garrett RA, Prangishvili D (2005a). Viral diversity in hot springs of Pozzuoli, Italy, and characterization of a unique archaeal virus, *Acidianus* bottle-shaped virus, from a new family, the Ampullaviridae.

Häring M, Vestergaard G, Brugger K, Rachel R, Garrett RA, Prangishvili D (2005b). Structure and genome organization of AFV2, a novel archaeal lipothrixvirus with unusual terminal and core structures. *Journal of Bacteriology*. Am Soc Microbiol. pp 3855-3858.

Häring M, Vestergaard G, Rachel R, Chen L, Garrett RA, Prangishvili D (2005c). Virology: Independent virus development outside a host. *Nature*. pp 1101-1102.

Henn MR, Sullivan MB, Stange-Thomann N, Osburne MS, Berlin AM, Kelly L *et al* (2010). Analysis of High-Throughput Sequencing and Annotation Strategies for Phage Genomes. *PLoS ONE*. p e9083.

Hewson I, Barbosa JG, Brown JM, Donelan RP, Eaglesham JB, Eggleston EM *et al* (2012). Temporal Dynamics and Decay of Putatively Allochthonous and Autochthonous Viral Genotypes in Contrasting Freshwater Lakes. *Applied and Environmental Microbiology*. pp 6583-6591.

Hoeijmakers WAM, Bartfai R, Francoijs K-J, Stunnenberg HG (2011). Linear amplification for deep sequencing. *Nat Protocols* **6**: 1026-1036.

Hohn MJ, Hedlund BP, Huber H (2002). Detection of 16S rDNA sequences representing the novel phylum "Nanoarchaeota": indication for a wide distribution in high temperature biotopes. *Syst. Appl. Microbiol*. pp 551-554.

Holm L, Rosenstrom P (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**: W545-549.

Holmes EC (2009). RNA virus genomics: a world of possibilities. *J Clin Invest* **119**: 2488-2495.

Honkavuori KS, Shivaprasad HL, Williams BL, Quan P-L, Hornig M, Street C *et al* (2008). Novel Borna Virus in Psittacine Birds with Proventricular Dilatation Disease. *Emerging Infectious Diseases*. pp 1883-1886.

Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S *et al* (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**: 1401-1412.

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*. pp 63-67.

Huber R, Huber H, Stetter KO (2000). Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. *FEMS Microbiology Reviews*. pp 615-623.

Huet J, Schnabel R, Sentenac A, Zillig W (1983). Archaeobacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. *EMBO J*. pp 1291-1294.

Hunter JD (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9**: 0090-0095.

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*. p e57355.

Illumina (2014). HiSeq X Ten Sequencing System. Illumina: San Diego, CA.

Ingalls AE, Shah SR, Hansman RL, Aluwihare LI, Santos GM, Druffel ERM *et al* (2006). Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proceedings of the National Academy of Sciences*. pp 6442-6447.

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.

Inskeep WP (2012). Microbial iron cycling in acidic geothermal springs of Yellowstone National Park: integrating molecular surveys, geochemical processes, and isolation of novel Fe-active microorganisms. pp 1-16.

Iverson E, Stedman K (2012). A genetic study of SSV1, the prototypical fusellovirus. *Front Microbiol*. p 200.

Jaakkola ST, Penttinen RK, Vilen ST, Jalasvuori M, Ronnholm G, Bamford JKH *et al* (2012). Closely Related Archaeal Haloarcula hispanica Icosahedral Viruses HHIV-2 and SH1 Have Nonhomologous Genes Encoding Host Recognition Functions. *J. Virol*. pp 4734-4742.

Jäälinoja HT, Roine E, Laurinmäki P, Kivelä HM, Bamford DH, Butcher SJ (2008). Structure and host-cell interaction of SH1, a membrane-containing, halophilic euryarchaeal virus. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 8008-8013.

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE *et al* (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* **10**.

Janekovic D, Wunderl S, Holz I, Zillig W, Gierl A, Neumann H (1983). TTV1, TTV2 and TTV3, a family of viruses of the extremely thermophilic, anaerobic, sulfur reducing archaeobacterium *Thermoproteus tenax*. *Molecular and General Genetics MGG*. Springer. pp 39-45.

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. *Applied and Environmental Microbiology*. pp 294-305.

Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER *et al* (2007). Extending assembly of short DNA sequences to handle error. *Bioinformatics*. pp 2942-2944.

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. *Environmental Microbiology Reports*. pp 195-202.

Kamer G, Argos P (1984). Primary Structural Comparison of Rna-Dependent Polymerases from Plant, Animal and Bacterial-Viruses. *Nucleic Acids Res* **12**: 7269-7282.

Kapoor A, Victoria J, Simmonds P, Wang C, Shafer RW, Nims R *et al* (2008). A highly divergent picornavirus in a marine mammal. *J Virol* **82**: 311-320.

Karner MB, DeLong EF, Karl DM (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507-510.

Kashefi K, Lovley DR (2003). Extending the upper temperature limit for life. *Science*. American Association for the Advancement of Science. pp 934-934.

Kellenberger E (2001). Exploring the unknown. The silent revolution of microbiology. *EMBO Rep*. pp 5-7.

Kelley LA, Sternberg MJ (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4**: 363-371.

Khayat R, Tang L, Larson ET, Lawrence CM, Young M, Johnson JE (2005). Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 18944-18949.

Khayat R, Fu Cy, Ortmann AC, Young MJ, Johnson JE (2010). The Architecture and Chemical Stability of the Archaeal *Sulfolobus* Turreted Icosahedral Virus. *J. Virol*. pp 9575-9583.

Kim KH, Bae JW (2011). Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Applied and Environmental Microbiology*. pp 7663-7668.

Kim TK, Thomas SM, Ho M, Sharma S, Reich CI, Frank JA *et al* (2009). Heterogeneity of Vaginal Microbial Communities within Individuals. *Journal of Clinical Microbiology*. pp 1181-1189.

King AM, Adams MJ, Lefkowitz EJ, Carstens EB (2011). *Virus taxonomy: IXth report of the International Committee on Taxonomy of Viruses*, vol. 9. Access Online via Elsevier.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Hamamura N, Heidelberg JF *et al* (2011). Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J* **5**: 1262-1278.

Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*. pp 543-546.

Koonin EV, Boyko VP, Dolja VV (1991). Small Cysteine-Rich Proteins of Different Groups of Plant Rna Viruses Are Related to Different Families of Nucleic Acid-Binding Proteins. *Virology* **181**: 395-398.

Koonin EV, Dolja VV (1993). Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit Rev Biochem Mol Biol* **28**: 375-430.

Koonin EV, Mushegian AR, Galperin MY, Walker DR (1997). Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol* **25**: 619-637.

Koonin EV, Senkevich TG, Dolja VV (2006). The ancient Virus World and evolution of cells. *Biol Direct* **1**: 29.

Koonin EV, Wolf YI, Nagasaki K, Dolja VV (2008). The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat Rev Microbiol* **6**: 925-939.

Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I *et al* (2009). Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology*. Elsevier Inc. pp 1-7.

Krupovic M, Bamford DH (2010). Order to the Viral Universe. *J. Virol.* pp 12476-12479.

Krupovič M, Spang A, Gribaldo S, Forterre P, Schleper C (2011). A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. *Biochem. Soc. Trans.* pp 82-88.

Kukkaro P, Bamford DH (2009). Virus-host interactions in environments with a wide range of ionic strengths. *Environmental Microbiology Reports*. pp 71-77.

Kunin V, Sorek R, Hugenholtz P (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**: R61.

Labonté JM, Suttle CA (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME Journal*.

Lai B, Ding R, Li Y, Duan L, Zhu H (2012). A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics*. pp 1455-1462.

Lang AS, Beatty JT (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends in Microbiology*. pp 54-62.

Laserson J, Jojic V, Koller D (2011). Genovo: De Novo Assembly for Metagenomes. *Journal of Computational Biology*. pp 429-443.

Lasken RS, Stockwell TB (2007). Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol*. p 19.

Lauck M, Hyeroba D, Tumukunde A, Weny G, Lank SM, Chapman CA *et al* (2011). Novel, Divergent Simian Hemorrhagic Fever Viruses in a Wild Ugandan Red Colobus Monkey Discovered Using Direct Pyrosequencing. *PLoS ONE*. p e19056.

Lawrence CM, Menon S, Eilers BJ, Bothner B, Khayat R, Douglas T *et al* (2009). Structural and Functional Studies of Archaeal Viruses. *J Biol Chem* **284**: 12599-12603.

Le T, Chiarella J, Simen BB, Hanczaruk B, Egholm M, Landry ML *et al* (2009). Low-Abundance HIV Drug-Resistant Viral Variants in Treatment-Experienced Persons Correlate with Historical Antiretroviral Use. *PLoS ONE*. p e6079.

Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, deWinter AD *et al* (2003). A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*. pp 3769-3777.

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.

Liu L, Li Y, Li S, Hu N, He Y, Pong R *et al* (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*. pp 1-11.

Liu Z, Klatt CG, Wood JM, Rusch DB, Ludwig M, Wittekindt N *et al* (2011). Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *ISME J* **5**: 1279-1290.

Lowenstern JB, Utah Uo, Survey G (2005). *Steam Explosions, Earthquakes, and Volcanic Eruptions: What's in Yellowstone's Future?* U.S. Geological Survey.

- Lwoff A, Tournier P (1966). The classification of viruses. *Annual Reviews in Microbiology* **20**: 45-74.
- 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*. pp 86-99.
- Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**: 7.
- Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biology Direct*. p 33.
- Makarova KS, Yutin N, Bell SD, Koonin EV (2010). Evolution of diverse cell division and vesicle formation systems in Archaea. *Nat Rev Microbiol* **8**: 731-741.
- Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P *et al* (2011). Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**: 467-477.
- Malboeuf CM, Yang X, Charlebois P, Qu J, Berlin AM, Casali M *et al* (2013). Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic Acids Research*.
- Maori E, Lavi S, Mozes-Koch R, Gantman Y, Peretz Y, Edelbaum O *et al* (2007). Isolation and characterization of Israeli acute paralysis virus, a dicistrovirus affecting honeybees in Israel: evidence for diversity due to intra- and inter-species recombination. *Journal of General Virology*. pp 3428-3438.
- Marcais G, Kingsford C (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. pp 764-770.
- Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C *et al* (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**: D225-229.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA *et al* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*.
- Martin A, Yeats S, Janekovic D, Reiter WD, Aicher W, Zillig W (1984). Sav-1, a Temperate Uv-Inducible DNA Virus-Like Particle from the Archaeobacterium *Sulfolobus Acidocaldarius* Isolate B-12. *Embo Journal* **3**: 2165-2168.

- Mathieu Bastian SHMJ (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. pp 1-2.
- Mei Y, Chen J, Sun D, Chen D, Yang Y, Shen P *et al* (2007). Induction and preliminary characterization of a novel halophage SNJ1 from lysogenic *Natrinema* sp. F5. *Canadian journal of microbiology* **53**: 1106.
- Meile L, Jenal U, Studer D, Jordan M, Leisinger T (1989). Characterization of Psi-M1, a Virulent Phage of *Methanobacterium-Thermoautotrophicum* Marburg. *Archives of Microbiology*. pp 105-110.
- Metzker ML (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics*. Nature Publishing Group. pp 31-46.
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al* (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**.
- Miller JR, Koren S, Sutton G (2010). Assembly algorithms for next-generation sequencing data. *Genomics*. Elsevier Inc. pp 315-327.
- Mochizuki T, Yoshida T, Tanaka R, Forterre P, Sako Y, Prangishvili D (2010). Diversity of viruses of the hyperthermophilic archaeal genus *Aeropyrum*, and isolation of the *Aeropyrum pernix* bacilliform virus 1, APBV1, the first representative of the family Clavaviridae. *Virology*. Elsevier Inc. pp 347-354.
- 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. *Journal of Bacteriology*. pp 5412-5419.
- Mochizuki T, Krupović M, Pehau-Arnaudet G, Sako Y, Forterre P, Prangishvili D (2012). Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proc. Natl. Acad. Sci. U.S.A.* pp 13386-13391.
- Mokili JL, Rohwer F, Dutilh BE (2012). Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*. Elsevier B.V. pp 63-77.
- Monier A, Pagarete A, de Vargas C, Allen MJ, Read B, Claverie JM *et al* (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Research*. pp 1441-1449.
- Muskhelishvili G, Palm P, Zillig W (1993). SSV1-encoded site-specific recombination system in *Sulfolobus shibatae*. *Molecular and General Genetics MGG*. Springer. pp 334-342.
- Myers EW (2000). A Whole-Genome Assembly of *Drosophila*. *Science*. pp 2196-2204.

- Nadal M, Mirambeau G, Forterre P, Reiter W-D, Duguet M (1986). Positively supercoiled DNA in a virus-like particle of an archaebacterium. *Nature* **321**: 256-258.
- Nakamura S, Yang C-S, Sakon N, Ueda M, Tougan T, Yamashita A *et al* (2009). Direct Metagenomic Detection of Viral Pathogens in Nasal and Fecal Specimens Using an Unbiased High-Throughput Sequencing Approach. *PLoS ONE*. p e4219.
- Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*. pp e155-e155.
- Newman M (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*. pp 8577-8582.
- Newman MEJ, Girvan M (2004). Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. p 026113.
- Ng SYM, Zolghadr B, Driessen AJM, Albers SV, Jarrell KF (2008). Cell Surface Structures of Archaea. *Journal of Bacteriology*. pp 6039-6047.
- Nicol GW, Schleper C (2006). Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle? *Trends in Microbiology*. pp 207-212.
- Ninomiya M, Ueno Y, Funayama R, Nagashima T, Nishida Y, Kondo Y *et al* (2012). Use of Illumina Deep Sequencing Technology To Differentiate Hepatitis C Virus Variants. *Journal of Clinical Microbiology*. pp 857-866.
- Niu B, Fu L, Sun S, Li W (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11**: 187.
- Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H *et al* (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39**: 3204-3223.
- Nuttall SD, Dyll-Smith ML (1993). HF1 and HF2: novel bacteriophages of halophilic archaea. *Virology*. Elsevier. pp 678-684.
- Oke M, Kerou M, Liu H, Peng X, Garrett RA, Prangishvili D *et al* (2010). A Dimeric Rep Protein Initiates Replication of a Linear Archaeal Virus Genome: Implications for the Rep Mechanism and Viral Replication. *J. Virol*. pp 925-931.
- Out AA, van Minderhout IJHM, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K *et al* (2009). Deep sequencing to reveal new variants in pooled DNA samples. *Human Mutation* **30**: 1703-1712.

Pagaling E, Haigh RD, Grant WD, Cowan DA, Jones BE, Ma Y *et al* (2007). Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. *BMC Genomics*. p 410.

Palacios G, Druce J, Du L, Tran T, Birch C, Briese T *et al* (2008). A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. *N Engl J Med*. pp 991-998.

Palm P, Schleper C, Grampp B, Yeats S, McWilliam P, Reiter WD *et al* (1991). Complete nucleotide sequence of the virus SSV1 of the archaeobacterium *Sulfolobus shibatae*. *Virology*. pp 242-250.

Paul JH, Jeffrey WH, DeFlaun MF (1987). Dynamics of extracellular DNA in the marine environment. *Applied and Environmental Microbiology*. pp 170-179.

Paul JH, Jiang SC, Rose JB (1991). Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. *Applied and Environmental Microbiology*. pp 2197-2204.

Pauling C (1982). Bacteriophages of *Halobacterium halobium*: isolation from fermented fish sauce and primary characterization. *Canadian Journal of Microbiology*. NRC Research Press. pp 916-921.

Peng X, Blum H, She Q, Mallok S, Brugger K, Garrett RA *et al* (2001). Sequences and Replication of Genomes of the Archaeal Rudiviruses SIRV1 and SIRV2: Relationships to the Archaeal Lipothrixvirus SIFV and Some Eukaryal Viruses. *Virology*. pp 226-234.

Peng X, Basta T, Häring M, Garrett RA, Prangishvili D (2007). Genome of the *Acidianus* bottle-shaped virus and insights into the replication and packaging mechanisms. *Virology*. pp 237-243.

Peng Y, Leung HCM, Yiu SM, Chin FYL (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*. pp 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*. pp 1420-1428.

Pietila MK, Atanasova NS, Manole V, Liljeroos L, Butcher SJ, Oksanen HM *et al* (2012). Virion Architecture Unifies Globally Distributed Pleolipoviruses Infecting Halophilic Archaea. *J. Virol*. pp 5067-5079.

Pietilä MK, Laurinavičius S, Sund J, Roine E, Bamford DH (2010). The single-stranded DNA genome of novel archaeal virus halorubrum pleomorphic virus 1 is enclosed in the envelope decorated with glycoprotein spikes. *J. Virol*. pp 788-798.

Pietilä MK, Atanasova NS, Oksanen HM, Bamford DH (2012). Modified coat protein forms the flexible spindle-shaped virion of haloarchaeal virus His1. *Environmental Microbiology*.

Pietilä MK, Laurinmäki P, Russell DA, Ko C-C, Jacobs-Sera D, Hendrix RW *et al* (2013). Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proc. Natl. Acad. Sci. U.S.A.*

Pina M, Bize A, Forterre P, Prangishvili D (2011). The archeoviruses. *FEMS Microbiology Reviews*. pp 1035-1054.

Podar M, Makarova KS, Graham DE, Wolf YI, Koonin EV, Reysenbach A-L (2013). Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biology Direct*. Biology Direct. pp 1-1.

Polson SW, Wilhelm SW, Wommack KE (2010). Unraveling the viral tapestry (from inside the capsid out). *The ISME Journal*. Nature Publishing Group. pp 1-4.

Porter K, Kukkaro P, Bamford JKH, Bath C, Kivelä HM, Dyall-Smith ML *et al* (2005). SH1: A novel, spherical halovirus isolated from an Australian hypersaline lake. *Virology*. pp 22-33.

Porter K, Russ BE, Dyall-Smith ML (2007). Virus–host interactions in salt lakes. *Current Opinion in Microbiology*. pp 418-424.

Porter K, Tang S-L, Chen C-P, Chiang P-W, Hong M-J, Dyall-Smith M (2013). PH1: An Archaeovirus of *Haloarcula hispanica* Related to SH1 and HHIV-2. *Archaea*. pp 1-17.

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 (2003). Evolutionary insights from studies on viruses of hyperthermophilic archaea. *Research in Microbiology*. pp 289-294.

Prangishvili D, Garrett RA (2004). Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. *Biochem Soc Trans* **32**: 204-208.

Prangishvili D, Garrett RA (2005). Viruses of hyperthermophilic Crenarchaea. *Trends Microbiol* **13**: 535-542.

Prangishvili D, Forterre P, Garrett RA (2006a). Viruses of the Archaea: a unifying view. *Nature Publishing Group*. pp 837-848.

Prangishvili D, Garrett RA, Koonin EV (2006b). Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. *Virus Res* **117**: 52-67.

Prangishvili D, Vestergaard G, Häring M, Aramayo R, Basta T, Rachel R *et al* (2006c). Structural and Genomic Properties of the Hyperthermophilic Archaeal Virus ATV with an Extracellular Stage of the Reproductive Cycle. *Journal of Molecular Biology*. pp 1203-1216.

Prangishvili D, Krupovič M (2012). A new proposed taxon for double-stranded DNA viruses, the order “Ligamenvirales”. *Arch Virol*. pp 791-795.

Prangishvili D (2013). The Wonderful World of Archaeal Viruses. *Annu. Rev. Microbiol*. pp 565-585.

Preston CM, Wu KY, Molinski TF, DeLong EF (1996). A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 6241-6246.

Price MN, Dehal PS, Arkin AP (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641-1650.

Pruitt KD, Tatusova T, Maglott DR (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*. pp D61-D65.

Quax TEF, Lucas S, Reimann J, Pehau-Arnaudet G, Prevost M-C, Forterre P *et al* (2011). Simple and elegant design of a virion egress structure in Archaea. *Proc. Natl. Acad. Sci. U.S.A.* pp 3354-3359.

Rachel R, Bettstetter M, Hedlund BP, H x000E4 ring M, Kessler A, Stetter KO *et al* (2002). Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. *Arch Virol*. pp 2419-2429.

Radford AD, Chapman D, Dixon L, Chantrey J, Darby AC, Hall N (2012). Application of next-generation sequencing technologies in virology. *Journal of General Virology*. pp 1853-1868.

Ravanti JJ, Gaidelyte A, Bamford DH, Bamford JKH (2003). Comparative analysis of bacterial viruses Bam35, infecting a gram-positive host, and PRD1, infecting gram-negative hosts, demonstrates a viral lineage. *Virology*. pp 401-414.

Redder P, Peng X, Brugger K, Shah SA, Roesch F, Greve B *et al* (2009). Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interval viral recombination mechanism. *Environmental Microbiology*. pp 2849-2862.

Reigstad LJ, Jorgensen SL, Schleper C (2009). Diversity and abundance of Korarchaeota in terrestrial hot springs of Iceland and Kamchatka. *The ISME Journal*. Nature Publishing Group. pp 346-356.

Reinisch KM, Nibert ML, Harrison SC (2000). Structure of the reovirus core at 3.6 Å resolution. *Nature*. pp 960-967.

Remmert M, Biegert A, Hauser A, Söding J (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*. Nature Publishing Group.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F *et al* (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*. pp 334-338.

Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature Publishing Group*. Nature Publishing Group. pp 607-617.

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 U S A* **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. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 7716-7720.

Rodriguez-Brito B, Li LL, Wegley L, Furlan M, Angly F, Breitbart M *et al* (2010). Viral and microbial community dynamics in four aquatic environments. *Isme Journal* **4**: 739-751.

Rohwer F, Seguritan V, Choi DH, Segall AM, Azam F (2001). Production of shotgun libraries using random amplification. *Biotech*. pp 108-112- 114-106- 118.

Rohwer F, Edwards R (2002). The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. *Journal of Bacteriology*. pp 4529-4535.

Rohwer F (2003). Global phage diversity. *Cell*. Elsevier. p 141.

Roine E, Pietila MK, Paulin L, Kalkkinen N, Bamford DH (2009). An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. *Mol Microbiol* **72**: 307-319.

Roine E, Kukkaro P, Paulin L, Laurinavicius S, Domanska A, Somerharju P *et al* (2010). New, Closely Related Haloarchaeal Viral Elements with Different Nucleic Acid Types. *J. Virol.* pp 3682-3689.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* pp 84-89.

Ronaghi M, Uhlén M, Nyren P (1998). A sequencing method based on real-time pyrophosphate. *Science* **281**: 363-365.

Ronaghi M (2003). Pyrosequencing for SNP genotyping. *Methods in molecular biology (Clifton, NJ)* **212**: 189.

Rosario K, Breitbart M (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology.* pp 289-297.

Rossmann MG, Mesyanzhinov VV, Arisaka F, Leiman PG (2004). The bacteriophage T4 DNA injection machine. *Current Opinion in Structural Biology.* pp 171-180.

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

Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S *et al* (2012). Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS ONE.* p e33641.

Ruehrup S, Urbano P, Berger A (2013). Botnet detection revisited: theory and practice of finding malicious P2P networks via Internet connection graphs. ... *IEEE Conference on.*

Runckel C, Flenniken ML, Engel JC, Ruby JG, Ganem D, Andino R *et al* (2011). Temporal Analysis of the Honey Bee Microbiome Reveals Four Novel Viruses and Seasonal Prevalence of Known Viruses, Nosema, and Crithidia. *PLoS ONE.* p e20656.

Sanger F, Air G, Barrell B, Brown N, Coulson A, Fiddes C *et al* (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**: 687.

Sano E, Carlson S, Wegley L, Rohwer F (2004). Movement of viruses between biomes. *Applied and Environmental Microbiology.* Am Soc Microbiol. pp 5842-5846.

Santos F, Meyerdierks A, Peña A, Rosselló-Mora R, Amann R, Antón J (2007). Metagenomic approach to the study of halophages: the environmental halophage 1. *Environmental Microbiology* **9**: 1711-1723.

Schleper C, Kubo K, Zillig W (1992). The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of

transfection with viral DNA. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 7645-7649.

Schnabel H, Zillig W, Pfäffle M, Schnabel R, Michel H, Delius H (1982). Halobacterium halobium phage øH. *EMBO J*. pp 87-92.

Schneemann A, Zhong W, Gallagher TM, Rueckert RR (1992). Maturation cleavage required for infectivity of a nodavirus. *J Virol* **66**: 6728-6734.

Schneemann A, Gallagher TM, Rueckert RR (1994). Reconstitution of Flock House provirions: a model system for studying structure and assembly. *J Virol* **68**: 4547-4556.

Schneemann A, Marshall D (1998). Specific encapsidation of nodavirus RNAs is mediated through the C terminus of capsid precursor protein alpha. *J Virol* **72**: 8738-8746.

Scholz MB, Lo C-C, Chain PSG (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology*. pp 9-15.

Schrodinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Sencilo A, Paulin L, Kellner S, Helm M, Roine E (2012). Related haloarchaeal pleomorphic viruses contain different genome types. *Nucleic Acids Research*. pp 5523-5534.

Senčilo A, Jacobs-Sera D, Russell DA, Ko C-C, Bowman CA, Atanasova NS *et al* (2013). Snapshot of haloarchaeal tailed virus genomes. *RNA Biol*.

She Q, Brügger K, Chen L (2002). Archaeal integrative genetic elements and their impact on genome evolution. *Research in Microbiology* **153**: 325-332.

Shendure J, Ji H (2008). Next-generation DNA sequencing. *Nature Biotechnology*. pp 1135-1145.

Sime-Ngando T, Lucas S, Robin A, Tucker KP, Colombet J, Bettarel Y *et al* (2011). Diversity of virus–host systems in hypersaline Lake Retba, Senegal. *Environmental Microbiology* **13**: 1956-1972.

Simon M, Azam F (1989). Protein content and protein synthesis rates of planktonic marine bacteria. *Marine ecology progress series*. Oldendorf. pp 201-213.

Snyder JC, Stedman K, Rice G, Wiedenheft B, Spuhler J, Young MJ (2003). Viruses of hyperthermophilic Archaea. *Res Microbiol* **154**: 474-482.

Snyder JC, Wiedenheft B, Lavin M, Roberto FF, Spuhler J, Ortmann AC *et al* (2007). Virus movement maintains local virus population diversity. *Proc. Natl. Acad. Sci. U.S.A.* pp 19102-19107.

Snyder JC, Bateson MM, Lavin M, Young MJ (2010). Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. *Appl Environ Microbiol* **76**: 7251-7258.

Snyder JC, Brumfield SK, Kerchner KM, Quax TEF, Prangishvili D, Young MJ (2013). Insights into a viral lytic pathway from an archaeal virus-host system. *J. Virol.* pp 2186-2192.

Soding J (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951-960.

Soding J, Biegert A, Lupas AN (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**: W244-W248.

Sorek R, Kunin V, Hugenholtz P (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**: 181-186.

Stamatakis A, Hoover P, Rougemont J (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* **57**: 758-771.

Stanton TB (2007). Prophage-like gene transfer agents—Novel mechanisms of gene exchange for *Methanococcus*, *Desulfovibrio*, *Brachyspira*, and *Rhodobacter* species. *Anaerobe*. pp 43-49.

Stedman KM, She Q, Phan H, Arnold HP, Holz I, Garrett RA *et al* (2003). Relationships between fuselloviruses infecting the extremely thermophilic archaeon *Sulfolobus*: SSV1 and SSV2. *Research in Microbiology* **154**: 295-302.

: Biogeographical diversity of archaeal viruses. *SYMPOSIA-SOCIETY FOR GENERAL MICROBIOLOGY*.

Stern A, Mick E, Tirosh I, Sagy O, Sorek R (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Research*. pp 1985-1994.

Stetter KO (1996). Hyperthermophilic procaryotes. *FEMS Microbiology Reviews*. Wiley Online Library. pp 149-158.

Steward GF, Azam F (2000). Analysis of marine viral assemblages. *Microbial Biosystems* (Bell, C., Brylinsky, M. and Johnson-Green, P., Eds.). pp 159-165.

Steward GF, Montiel JL, Azam F (2000). Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnology and Oceanography*. pp 1697-1706.

Suttle CA (1994). The significance of viruses to mortality in aquatic microbial communities. *Microbial Ecology* **28**: 237-243.

Suttle CA (2005). Viruses in the sea. *Nature* **437**: 356-361.

Suttle CA (2007). Marine viruses — major players in the global ecosystem. *Nature Publishing Group*. pp 801-812.

Swerdlow H, Gesteland R (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res* **18**: 1415-1419.

Takai K, Horikoshi K (1999). Genetic diversity of archaea in deep-sea hydrothermal vent environments. *Genetics*. pp 1285-1297.

Tang S-L, Nuttall S, Ngui K, Fisher C, Lopez P, Dyll-Smith M (2002). HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome. *Molecular Microbiology*. pp 283-296.

Tang SL, Nuttall S, Dyll-Smith M (2004). Haloviruses HF1 and HF2: Evidence for a Recent and Large Recombination Event. *Journal of Bacteriology*. pp 2810-2817.

Terns MP, Terns RM (2011). CRISPR-based adaptive immune systems. *Curr Opin Microbiol* **14**: 321-327.

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009). Laboratory procedures to generate viral metagenomes. *Nat Protoc*. pp 470-483.

Torsvik T, Dundas I (1974). Bacteriophage of *Halobacterium salinarium*. *Nature* **248**: 680.

Torsvik V, Goksøyr J, Daae FL (1990). High diversity in DNA of soil bacteria. *Applied and Environmental Microbiology*. Am Soc Microbiol. pp 782-787.

Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M (2002). Next Generation Sequence Assembly with AMOS. John Wiley & Sons, Inc.

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. *The ISME Journal*. Nature Publishing Group. pp 1-13.

Turcatti G, Romieu A, Fedurco M, Tairi AP (2007). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Research*. pp e25-e25.

Urbonavičius J, Auxilien S, Walbott H, Trachana K, Golinelli-Pimpaneau B, Brochier-Armanet C *et al* (2007). Acquisition of a bacterial RumA-type tRNA(uracil-54, C5)-methyltransferase by Archaea through an ancient horizontal gene transfer. *Molecular Microbiology*. pp 323-335.

Valentine DL (2007). Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nature Publishing Group*. pp 316-323.

van den Brand JMA, van Leeuwen M, Schapendonk CM, Simon JH, Haagmans BL, Osterhaus ADME *et al* (2012). Metagenomic Analysis of the Viral Flora of Pine Marten and European Badger Feces. *J. Virol*. pp 2360-2365.

Van Regenmortel M, Fauquet C (2000). Virus taxonomy: classification and nomenclature of viruses: seventh report of the International Committee on Taxonomy of Viruses.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Vestergaard G, Häring M, Peng X, Rachel R, Garrett RA, Prangishvili D (2005). A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus Acidianus. *Virology*. pp 83-92.

Vestergaard G, Aramayo R, Basta T, Häring M, Peng X, Brugger K *et al* (2008a). Structure of the acidianus filamentous virus 3 and comparative genomics of related archaeal lipothrixviruses. *J. Virol*. pp 371-381.

Vestergaard G, Shah SA, Bize A, Reitberger W, Reuter M, Phan H *et al* (2008b). Stygiolobus Rod-Shaped Virus and the Interplay of Crenarchaeal Rudiviruses with the CRISPR Antiviral System. *Journal of Bacteriology*. pp 6837-6845.

Vogelsang-Wenke H, Oesterhelt D (1988). Isolation of a halobacterial phage with a fully cytosine-methylated genome. *Molecular and General Genetics MGG*. Springer. pp 407-414.

Wais AC, Kon M, MacDonald RE, Stollar BD (1975). Salt-dependent bacteriophage infecting *Halobacterium cutirubrum* and *H. halobium*. *Nature*. pp 314-315.

Wais AC, Daniels LL (1985). Populations of bacteriophage infecting *Halobacterium* in a transient brine pool. *FEMS Microbiology Letters*. Elsevier. pp 323-326.

Warren RL, Sutton GG, Jones SJM, Holt RA (2007). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*. pp 500-501.

- Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, Barnstead M *et al* (2003). The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. *Proceedings of the National Academy of Sciences*. pp 12984-12988.
- Webb JS, Lau M, Kjelleberg S (2004). Bacteriophage and Phenotypic Variation in Pseudomonas aeruginosa Biofilm Development. *Journal of Bacteriology*. pp 8066-8073.
- Weinbauer MG (2004). Ecology of prokaryotic viruses. *Fems Microbiology Reviews* **28**: 127-181.
- Werner F (2007). Structure and function of archaeal RNA polymerases. *Molecular Microbiology*. pp 1395-1404.
- Wiedenheft B, Stedman K, Roberto F, Willits D, Gleske A-K, Zoeller L *et al* (2004). Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses. *J. Virol. Am Soc Microbiol*. pp 1954-1961.
- Wikoff WR, Liljas L, Duda RL, Tsuruta H, Hendrix RW, Johnson JE (2000). Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*. pp 2129-2133.
- Wilhelm SW, Suttle CA (1999). Viruses and Nutrient Cycles in the Sea. American Institute of Biological Sciences Circulation, AIBS, 1313 Dolley Madison Blvd., Suite 402, McLean, VA 22101. USA
- Williamson KE, Wommack KE, Radosevich M (2003). Sampling natural viral communities from soil for culture-independent analyses. *Applied and Environmental Microbiology*. Am Soc Microbiol. pp 6628-6633.
- 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*. p e1456.
- Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, Relman DA *et al* (2011). Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proceedings of the National Academy of Sciences*. National Acad Sciences. pp 4547-4553.
- Witte A, Baranyi U, Klein R, Sulzner M, Luo C, Wanner G *et al* (1997). Characterization of Natronobacterium magadii phage phi Ch1, a unique archaeal phage containing DNA and RNA. *Molecular Microbiology*. pp 603-616.
- Woese CR, Fox GE (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences* **74**: 5088-5090.
- Woese CR (1987). Bacterial evolution. *Microbiol. Rev.* pp 221-271.

Woese CR, Kandler O, Wheelis ML (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* **87**: 4576-4579.

Woese CR (1994). There must be a prokaryote somewhere: microbiology's search for itself. *Microbiol. Rev.* pp 1-9.

Wommack KE, Colwell RR (2000). Virioplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews.* pp 69-114.

Wommack KE, Bhavsar J, Ravel J (2008). Metagenomics: Read Length Matters. *Applied and Environmental Microbiology.* pp 1453-1463.

Wommack KE, Sime-Ngando T, Winget DM, Jamindar S, Helton RR (2010). Filtration-based methods for the collection of viral concentrates from large water samples. *Manual of Aquatic Viral Ecology (MAVE).* pp 110-117.

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.* pp 427-439.

Wooley JC, Godzik A, Friedberg I (2010). A primer on metagenomics. *PLoS Comput Biol.* Public Library of Science. p e1000667.

Wooley JC, Ye Y (2010). Metagenomics: facts and artifacts, and computational challenges. *Journal of computer science and technology.* Springer. pp 71-81.

Wu Z, Ren X, Yang L, Hu Y, Yang J, He G *et al* (2012). Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. *J. Virol.* pp 10999-11012.

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.* pp 8677-8686.

Xie G, Chain PSG, Lo CC, Liu KL, Gans J, Merritt J *et al* (2010). Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Molecular Oral Microbiology.* pp 391-405.

Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ *et al* (2012). De novo Assembly of highly diverse viral populations. *BMC Genomics.* BioMed Central Ltd. p 475.

- Yau S, Lauro FM, DeMaere MZ, Brown MV, Thomas T, Raftery MJ *et al* (2011). Virophage control of antarctic algal host-virus dynamics. *Proc. Natl. Acad. Sci. U.S.A.* pp 6163-6168.
- Yilmaz S, Allgaier M, Hugenholtz P (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Meth* **7**: 943-944.
- Yoshida M, Takaki Y, Eitoku M, Nunoura T, Takai K (2013). Metagenomic Analysis of Viral Communities in (Hado)Pelagic Sediments. *PLoS ONE*. p e57271.
- Young M, Bolduc B, Shaughnessy DP, Roberto FF, Wolf YI, Koonin EV (2013). Reply to "codon usage frequency of RNA virus genomes from high-temperature acidic-environment metagenomes". *J. Virol.* pp 1920-1921.
- Zamyatkin DF, Parra F, Alonso JM, Harki DA, Peterson BR, Grochulski P *et al* (2008). Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. *J Biol Chem* **283**: 7705-7712.
- Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW *et al* (2006). Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology*. pp 680-686.
- Zhang Z, Liu Y, Wang S, Yang D, Cheng Y, Hu J *et al* (2012). Temperate membrane-containing halophilic archaeal virus SNJ1 has a circular dsDNA genome identical to that of plasmid pHH205. *Virology*. Elsevier. pp 1-9.
- Zillig W, Kletzin A, Schleper C, Holz I, Janekovic D, Hain J *et al* (1994). Screening for Sulfolobales, their Plasmids and their Viruses in Icelandic Solfataras. *Syst. Appl. Microbiol.* Gustav Fischer Verlag, Stuttgart · Jena · New York pp 609-628.
- Zillig W, Prangishvilli D, Schleper C, Elferink M, Holz I, Albers S *et al* (1996). Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. *FEMS Microbiology Reviews*. pp 225-236.
- Zlotnick A, Reddy VS, Dasgupta R, Schneemann A, Ray WJ, Jr., Rueckert RR *et al* (1994). Capsid assembly in a family of animal viruses primes an autoproteolytic maturation that depends on a single aspartic acid residue. *J Biol Chem* **269**: 13680-13684.