ARCHAEOAL HOST VIRUS INTERACTIONS

by

Jennifer Fulton Wirth

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Microbiology

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 2011
APPROVAL

of a dissertation submitted by

Jennifer Fulton Wirth

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citation, bibliographic style, and consistency and is ready for submission to The Graduate School.

Mark Young

Approved for the Department of Microbiology

Mark Jutila

Approved for The Graduate School

Dr. Carl A. Fox
STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with “fair use” as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted “the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part.”

Jennifer Fulton Wirth

July 2011
ACKNOWLEDGEMENTS

I would like to thank all of the members of my committee: Mark Young, Trevor Douglas, Mike Franklin, Mensur Dlakic and Martin Teintze. A special thank you to Mark, who often frustrated me, but also supported and encouraged me, both professionally and personally, through this process.

Thank you to all the current and past members of the lab: Jamie Snyder, Sue Brumfield, Mary Bateson, Peter Suci, Ben Bolduc, Becky Hochstein, Debbie Willits, Alice Ortmann, Josh Spuhler, Eric Gillitzer, Blake Wiedenheft, Danielle Bouchard, Lars Liepold, Hulda Thorisdottir. I owe Debbie Willits a special thank you for teaching me everything I needed to know during my first few years, and for being a good friend. Thank you also to Jamie Snyder for helping me with so much in the last two years, for being a great professional role model and a good friend.

I would like to thank Matt Lavin for hours of discussions about our data, and helping us understand what it all means. I would also like to thank Frank Roberto for more hours of discussions about our data and performing so much sequencing for me.

Finally, I would like to thank my family. Thanks to my parents for supporting me in so many ways, and for instilling me with a love of learning. Thanks to my husband, Jon, for tolerating my mania when I’m stressed and helping me see the lighter side of everything. Even though he has no idea what it all means, thanks to Finn for making it all worthwhile.
# TABLE OF CONTENTS

1. INTRODUCTION AND BACKGROUND .............................................................................. 1

The Domain Archaea ........................................................................................................ 1

*Sulfolobus solfataricus* as a Model Crenarchaeote .................................................. 6

Archaeal Viruses ............................................................................................................. 8

*Sulolobus* Turreted Icosahedral Virus (STIV) ............................................................ 12

STIV Genome .................................................................................................................. 14

STIV Structure ............................................................................................................... 15

STIV Transcriptional Analysis ....................................................................................... 18

Genetic Systems for Viruses of *Sulfolbus* .................................................................. 20

STIV Genetic System ...................................................................................................... 21

Virus-Host Interactions ................................................................................................. 21

CRISPR/Cas: A Newly Identified Prokaryotic Immune System ................................. 23

History of the Discovery of the CRISPR/Cas System ................................................. 23

Model for Function of CRISPR/Cas System in Cultured Isolates ............................... 27

Origins and Evolution of CRISPR/Cas Systems ......................................................... 32

CRISPR/Cas Function in Natural Populations .............................................................. 33

Objectives of this Research ........................................................................................ 34

2. DEVELOPMENT OF AN INFECTIOUS CLONE OF *SULFOLOBUS*
TURRETED ICOSAHEDRAL VIRUS (STIV) AS A TOOL FOR EXAMINING STIV GENETICS ................................................................. 36

Introduction .................................................................................................................. 36

Materials and Methods ................................................................................................. 39

Construction of an STIV Full-Length Clone ................................................................. 39

Removal of *Nco*I Site in *E. coli* Vector .................................................................. 42

Construction of STIV Gene Disruption Mutants ......................................................... 42

Preparation of Cloned Viral DNA for Transformation into *S. solfataricus* ............... 46

Transfection of *S. solfataricus* with Viral DNA ......................................................... 46

Detection of Infectious Progeny Virus Following Transfection with Cloned Viral DNA .......................................................................................................................... 47

Results ........................................................................................................................... 49

Construction of the Full-Length STIV/TA Clone and Demonstration of Infectivity ................................................................................................................................. 49

STIV Infectious Clone Allows for Targeted Gene Alterations .................................... 52

Discussion ..................................................................................................................... 55
3. VIRAL METAGENOMICS AT ROAD SIDE SPRINGS ..................................................... 60

Introduction ............................................................................................................... 60
Materials and Methods .......................................................................................... 63
Site Selection ........................................................................................................... 63
Sample Preparation and Virus Purification ........................................................... 64
Cell and Virus Enumeration by Epifluorescence Microscopy ................................. 66
TEM Analysis of Samples ....................................................................................... 67
Nucleic Acid Extraction ......................................................................................... 67
16S rDNA Sequencing ............................................................................................ 69
Differential Nucleic Acid Digestion ....................................................................... 69
Viral Metagenomic Sequencing and Assembly ...................................................... 69
Confirmation of RSSV contig Assembly by PCR Amplification, Cloning, and Sequencing ........................................................................................................... 70
Colony Hybridization ............................................................................................. 71
Mapping of Nymph Lake Metagenomic Reads to RSSV Contig .............................. 71
Open Reading Frame Prediction and Annotation .................................................... 72
Coat Protein Amplification, Cloning and Overexpression ..................................... 73
Results ..................................................................................................................... 75
Transmission Electron Microscopy ......................................................................... 75
16S rDNA Sequencing ........................................................................................... 76
Viral Metagenomics and RSSV Contig Assembly .................................................. 76
Confirmation of RSSV Contig Assembly by PCR Amplification, Cloning and Sequencing ........................................................................................................... 78
Colony Hybridization to Define RSSV Ends .......................................................... 80
Mapping Nymph Lake Metagenomic Reads to RSSV Contig ............................... 80
RSSV Contig Annotation ......................................................................................... 81
Coat Protein Overexpression .................................................................................. 83
Detection of RSS Viral Sequence in Other Yellowstone Hot Springs ................. 84
Discussion .............................................................................................................. 84

4. CRISPR/CAS ADAPTIVE IMMUNE FUNCTION IN A NATURAL ENVIRONMENT .................................................................................................................. 88

Introduction ........................................................................................................... 88
Materials and Methods .......................................................................................... 92
Site Selection and Chemical Analysis ..................................................................... 92
TABLE OF CONTENTS - CONTINUED

Host Sample Collection and Preparation .......................................................... 92
DNA Sequencing ............................................................................................... 93
Viral Sample Collection and Preparation ........................................................ 94
CRISPR Analysis .............................................................................................. 95
Cell and Virus Enumeration ............................................................................ 96
Results ................................................................................................................ 96
The NL Hot Springs has a Relatively Stable Water Chemistry that Supports a Low Biomass Microbial Community ......................................................... 96
The NL Hot Springs Supports a Low Complexity Archaeal Cellular and Viral Community .......................................................................................... 98
Temporal Dynamics of a CRISPR Locus .......................................................... 100
Spacer Presence Correlates to the Subsequent Presence of Matching Viral Sequence ....................................................................................... 106
Discussion ......................................................................................................... 111

5. CONCLUDING REMARKS AND FUTURE DIRECTIONS................................. 117
Development of an STIV Genetic System ........................................................ 118
Establishing Culture Independent Detection (Metagenomics) of Archaeal Viruses .......................................................................................... 122
Examining the CRISPR/Cas System in Natural Environments....................... 124

REFERENCES CITED ...................................................................................... 128
APPENDICES ..................................................................................................... 139

APPENDIX A: Development of a Genetically Tractable Strain of S. solfataricus 2-2-12 .................................................................................. 140
APPENDIX B: Tn5 Transposon Mutagenesis of STIV ....................................... 148
APPENDIX C: High-Throughput Mutagenesis Using Recombineering in STIV ....................................................................................... 158
APPENDIX D: STIV A197 Complementation ..................................................... 167
LIST OF TABLES

Table | Page
--- | ---
1.1 Viruses of the Crenarchaeotes | 10
2.1 Primers used for the Construction and Sequencing the Full-Length STIV Infectious Clone and Various Mutant Clones | 41
2.2 Infection and Transfections of a Variety of *Sulfolobus* spp. with STIV | 51
2.3 Mutant Viruses Constructed from the pFLSTIV Clone | 52
3.1 RSS Primers | 74
3.2 Annotation of RSS Viral Contig | 83
4.1 Cellular Metagenome Sequencing and Assembly Statistics | 94
4.2 Viral Metagenome Sequencing and Assembly Statistics | 95
4.3 16S Genes (>500 bps) Identified in Three Cellular Metagenomes | 100
4.4 BLASTx Matches to Viral Metagenomic Contigs | 101
4.5 Spacer Matches into Each of the Viral Metagenomes Allowing Zero Mismatches (100% Identity) | 106
4.6 Spacer Matches into Each of the Viral Metagenomes Allowing 5 Mismatches (83% Identity) | 107
A.1 *PyrEF* Primers | 147
B.1 TN5ΔRK6yori Primers | 157
C.1 Recombineering Primers | 166
D.1 A197 Complementation Primers | 175
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Phylogenetic Tree of SSU rDNA of Archaea</td>
<td>4</td>
</tr>
<tr>
<td>1.2</td>
<td>Phylogenetic Tree Showing the Placement of the Proposed Phylum, Thaumarchaeota, in the Archaeal Tree of Life</td>
<td>6</td>
</tr>
<tr>
<td>1.3</td>
<td>Electron Micrographs of <em>S. Solfatarius</em></td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>Lysis of <em>S. solfataricus</em> by STIV</td>
<td>13</td>
</tr>
<tr>
<td>1.5</td>
<td>Genome Map of STIV</td>
<td>15</td>
</tr>
<tr>
<td>1.6</td>
<td>Structure of STIV</td>
<td>17</td>
</tr>
<tr>
<td>1.7</td>
<td>Comparison of Capsid Architecture of STIV with a Eukaryal Virus and a Bacteriophage</td>
<td>18</td>
</tr>
<tr>
<td>1.8</td>
<td>Transcriptional Timing of STIV ORFs</td>
<td>19</td>
</tr>
<tr>
<td>1.9</td>
<td>CRISPR Locus and <em>cas</em> Genes from <em>Streptococcus thermophilus</em></td>
<td>23</td>
</tr>
<tr>
<td>1.10</td>
<td>Stages of CRISPR/Cas Function</td>
<td>27</td>
</tr>
<tr>
<td>1.11</td>
<td>CRISPR/Cas Interference Stage</td>
<td>29</td>
</tr>
<tr>
<td>2.1</td>
<td>Schematic of the Construction of the Full-Length STIV Clone</td>
<td>40</td>
</tr>
<tr>
<td>2.2</td>
<td>Full-Length Clone of STIV</td>
<td>49</td>
</tr>
<tr>
<td>2.3</td>
<td>Analysis of pFLSTIV Transformation into <em>S. solfataricus</em> 2-2-12</td>
<td>50</td>
</tr>
<tr>
<td>2.4</td>
<td>Analysis of ΔB116, A197 D151E Transformations into <em>S. solfataricus</em> 2-2-12</td>
<td>53</td>
</tr>
<tr>
<td>2.5</td>
<td>Analysis of the Phenotype of ΔB116 STIV</td>
<td>54</td>
</tr>
<tr>
<td>2.6</td>
<td>TEMs of Particles from A197 D151E STIV Transformed Cultures</td>
<td>55</td>
</tr>
<tr>
<td>3.1</td>
<td>Road Side Springs</td>
<td>63</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.2</td>
<td>Schematic of Sampling Method</td>
<td>64</td>
</tr>
<tr>
<td>3.3</td>
<td>100kDa Molecular Weight Cutoff Tangential Flow Filtration System</td>
<td>65</td>
</tr>
<tr>
<td>3.4</td>
<td>Transmission Electron Micrographs of Concentrated Sample from Road Side Springs</td>
<td>75</td>
</tr>
<tr>
<td>3.5</td>
<td>Analysis of Extracted Nucleic Acid from Road Side Spring</td>
<td>76</td>
</tr>
<tr>
<td>3.6</td>
<td>Sequence Coverage Map of the Putative Road Side Spring Viral Contig</td>
<td>78</td>
</tr>
<tr>
<td>3.7</td>
<td>PCR of Various Regions of the RSSV Contig</td>
<td>79</td>
</tr>
<tr>
<td>3.8</td>
<td>Colony Hybridization Membrane with Colonies Positive for RSS Sequence</td>
<td>80</td>
</tr>
<tr>
<td>3.9</td>
<td>Nymph Lake Reads Mapped to RSSV Contig</td>
<td>81</td>
</tr>
<tr>
<td>3.10</td>
<td>Open Reading Frame Map of RSSV Contig</td>
<td>83</td>
</tr>
<tr>
<td>4.1</td>
<td>Water Chemistry from 6/23/08 to 10/14/09</td>
<td>97</td>
</tr>
<tr>
<td>4.2</td>
<td>Analysis of Aug. 2007 Metagenome</td>
<td>98</td>
</tr>
<tr>
<td>4.3</td>
<td>Variation of Spacer Sequence Content Within the Monitored CRISPR Locus Over 25 Months of Sampling</td>
<td>104</td>
</tr>
<tr>
<td>4.4</td>
<td>Inverse Simpson Index of Diversity of Spacers Over Time</td>
<td>105</td>
</tr>
<tr>
<td>4.5</td>
<td>Increasing Spacer Matches into Viral Metagenomes Over Time</td>
<td>108</td>
</tr>
<tr>
<td>4.6</td>
<td>Decreasing Spacer Matches of Cellular Metagenomes Over Time</td>
<td>109</td>
</tr>
<tr>
<td>4.7</td>
<td>Individual Spacers Matched to Viral Metagenome Reads</td>
<td>110</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.8</td>
<td>Proposed Model of the Function of CRISPR/Cas in the Context of a Nutrient Limiting Environment and Frequency Dependent Selection</td>
<td>112</td>
</tr>
<tr>
<td>A.1</td>
<td>Analysis of 5-FOA Selected Colonies by PCR and Restriction Digest of pyrEF Gene Cassette</td>
<td>145</td>
</tr>
<tr>
<td>B.1</td>
<td>PCR Screen of Tn5 transposon Insertion into Native STIV</td>
<td>153</td>
</tr>
<tr>
<td>B.2</td>
<td>Construction of the TN5ΔRK6yori Transposon</td>
<td>155</td>
</tr>
<tr>
<td>D.1</td>
<td>ΔA197STIV Complemented with pSESD/A197</td>
<td>171</td>
</tr>
<tr>
<td>D.2</td>
<td>Dot Blots of qPCR Positive ΔA197STIV and pSESD/A197 Cultures</td>
<td>172</td>
</tr>
<tr>
<td>D.3</td>
<td>PCR Spanning SacII Mutation Shows A197 is WT After Replication in <em>S. solfataricus</em></td>
<td>173</td>
</tr>
</tbody>
</table>
ABSTRACT

Viruses are the most abundant biological entity on earth, and virus-host interactions are one of the most important factors shaping microbial populations (Suttle, 2007b). The study of both the cellular and viral members of the domain Archaea is a relatively new field. Thus, the viruses (and their cellular hosts) of Archaea are poorly understood as compared to viruses of Bacteria and Eukarya. This work has sought to expand our understanding of archaeal viruses by two general approaches. The first is by developing and implementing the use of a genetic system for a crenarchaeal virus, *Sulfolobus* turreted icosahedral virus (STIV), isolated from a hot (82°C) acidic (pH 2.2) pool in Yellowstone National Park, USA. The second approach has been to look at viral communities and their interactions with their cellular hosts in natural environments.

We have developed a genetic tool, an infectious clone for STIV, which has allowed for genetic analysis of this virus. A number of viral genes have been knocked out, and their functions investigated using this tool. We have determined that at least three viral genes, A197, B345 and C381, are required for viral replication, while one gene, B116, is not essential. Work continues investigating function for other STIV genes as well as specific interactions with its host, *Sulfolobus solfataricus*.

We have performed total community sequencing (metagenomics) for both the cellular and viral populations of several hot springs in Yellowstone National Park. We have been able to assemble a near full-length putative novel viral genome from one of these sites. We have also performed an in depth analysis of the function of a newly described bacterial and archaeal adaptive immune system (CRISPR/Cas) in a natural environment. This study has provided insights into the function of this immune system in a complex nutrient limited environment, which would not have been observed by studying cultured isolates in a laboratory.
Viruses have recently been shown to be the most abundant biological entity on the planet (Suttle, 2007b). Virus-host interactions are a major driver of cellular evolution and shaper of microbial ecology on earth. The domain Archaea, as a newly discovered and described branch of the tree of life, and the viruses that infect them, are poorly understood as compared to Bacteria and Eukarya.

This work describes recent advances in understanding these virus-host interactions. We have made significant advances in our understanding of the specific life cycle of the Sulfolobus turretted icosahedral virus (STIV) and its interaction with its host Sulfolobus solfatarius through the construction and implementation of new genetic tools for this virus. We have also expanded our understanding of virus-host interactions in a natural environment through both metagenomic sequencing of virus and host components of hot springs in Yellowstone National Park, and by following the dynamics of the newly described prokaryotic immune system, CRISPR/Cas.

The Domain Archaea

All life on Earth is divided into three major domains: Eukarya, Bacteria and Archaea. Prior to the advent of DNA sequencing, all single celled anucleated cells were grouped together as prokaryotes. With the introduction of molecular techniques,
particularly 16s rDNA analysis, the degree of relatedness of two organisms could be more accurately measured. This led to the division of the prokaryotes into two separate domains: Bacteria and Archaea (Woese et al., 1990). Archaea are an interesting blend of biochemical machinery thought previously to belong exclusively to either Bacteria or Eukarya. Interestingly, in the area of information processing, Archaea are more similar to Eukarya than they are to Bacteria. In particular, the DNA replication genes (helicases, replicases and primases) are closely related to eukaryal homologs and not to bacteria (Gribaldo and Brochier-Armanet, 2006). Some ribosomal proteins are also shared between Archaea and Eukarya while none are shared with Bacteria. Other information processing systems, such as transcription, translation and RNA metabolism are also more eukaryal in nature. However, Archaea are similar to Bacteria in other ways, such as the use of polycistronic transcripts, the organization of their chromosomes with rapid and bidirectional replication (Grabowski and Kelman, 2003) and the use of Shine-Delgarno sequences for translation initiation (Gribaldo and Brochier-Armanet, 2006). The use of multiple origins of replication (Robinson and Bell, 2005) and chromatin proteins in transcriptional regulation both unknown in Bacteria, suggesting a level of complexity somewhere in between Bacteria and Eukarya.

While Archaea have intriguing homology with both Bacteria and Eukarya, they also have some distinctive features that are uniquely archaeal. Most notable is the phospholipid organization in their cell walls. Archaea have an isoprenoid ether and a glycerol-1-phosphate backbone, while both Eukarya and Bacteria have ester linkages
and glycerol-3-phosphate backbones (Gribaldo and Brochier-Armanet, 2006). Methanogenesis and survival at extreme high temperatures (>72°C) are also both unique to Archaea (Gribaldo and Brochier-Armanet, 2006).

On the rDNA-based universal tree of life, Archaea are shown branching more closely with Eukarya. In other words, Archaea and Eukarya may share a common ancestor, while the link with Bacteria is through the Last Universal Common Ancestor (LUCA) (Woese et al., 1990). However, the mosaic nature of archaean genomes (i.e. some similarity to Bacteria, some similarity to Eukarya and some unique features) has led to some controversy. Some researchers hypothesize that the last common ancestor was a protoeukaryote, some say it was more prokaryal while others propose that Archaea emerged from a genome fusion between a protoeukaryote and a protoprokaryote (Grabowski and Kelman, 2003; Woese et al., 1991). Regardless of whether Archaea are placed nearer to Bacteria or to Eukarya there remain some inconsistencies. Lateral gene transfer, gene duplications and gene loss may account for some of these inconsistencies (Grabowski and Kelman, 2003), but make resolution of the placement of Archaea on the tree of life more difficult.

The domain Archaea is further subdivided into two major phyla, Crenarchaeae and Euryarchaeae, with at least two other proposed phyla, Nanoarchaeae (Woese et al., 1991; Garrity, 2001) and Korarchaeota (Fig. 1.1). Crenarchaeotes tend to be thermophiles that either oxidize or reduce sulfur for energy. For the cultured organisms
there is only one class within the Crenarchaeota, *Thermoprotei*, which contains four orders: *Thermoproteales*, *Desulfurococcales*, *Caldisphaerales*, and *Sulfolobales*.

---

Figure 1.1 Phylogenetic tree of SSU rDNA of Archaea. (Gribaldo and Brochier-Armanet, 2006)

The Euryarchaeotes comprise a more varied group of methanogens and halophiles (Pace, 2009) as well as thermophiles and are made up of eight classes across a broad spectrum of environments: *Thermococcales*, *Methanococci*, *Archaeoglobi*, *Thermoplasmata*, *Methanopyri*, *Methanobacteria*, *Halobacteria*, and *Methanomicrobia*. 
Nanoarchaea currently comprises only one sequenced isolate, the extremely small obligate symbiont of *Igniococcus hospitalis*, *Nanoarchaeum equitans*. The Korarchaeota phylum has primarily been identified and defined by uncultured members, with only one organism, *Candidatus Korarchaeum cryptofilum*, maintained in an enrichment culture (Elkins et al., 2008). Both of these phyla likely represent deeply branching lineages within the Archaea.

The thermophilic Crenarchaeotes are relatively easy to grow under laboratory conditions. As a result, for many years our knowledge of Archaea came primarily through these cultured isolates. However, the recent move to look at populations *in situ* via metagenomic sequencing, rather than selecting for only the species that are easy to grow in culture, has dramatically broadened our understanding of the diversity within this domain.

Thaumarchaea are a recently proposed phylum that form a deeply branching cluster of environmental sequences comprised of all the archaeal ammonia oxidizers and other environmental sequences of organisms with unknown metabolisms (Pester et al.). These organisms were originally classified as mesophilic Crenarchaeotes, but analysis of 53 ribosomal proteins common to Archaea and Eukarya demonstrated that they branched off before the separation of Crenarchaeota and Euryarchaeota (Fig. 1.2).
Figure. 1.2 Phylogenetic tree showing the placement of the proposed phylum, Thaumarchaeota, in the archaeal tree of life (Pester et al.).

While Archaea are known to inhabit extreme environments, such as high temperature, low pH, high salt, etc, we also now know that they thrive in almost any environment. Archaea are found in soils, in the human digestive system, in the ocean, and in freshwater. They co-exist with bacteria and eukaryotes in almost every environment. However, the extreme environments, such as many Yellowstone thermal features, tend to be dominated by archaeal species.

**Sulfolobus solfataricus** as a Model Crenarchaeote

*Sulfolobus* spp. are hyperthermophilic acidophilic crenarchaeotes that were first identified by Thomas Brock working in Yellowstone National Park in the early 1970s. There are a number of sequenced *Sulfolobus* species, including *S. acidocaldarius* (Chen et al., 2005), *S. tokodaii* (Kawarabayasi et al., 2001), seven isolates of *S. islandicus* (two from Lassen National Park, two from Yellowstone National Park and three from Kamchatka, Russia) (Reno et al., 2009), and two isolates of *S. solfataricus* (strain P2 from Italy and strain 98/2 (She et al., 2001). Despite some significant difficulties (detailed
below), *Sulfolobus solfataricus* has become a model organism for the crenarchaeal phylum.

*Sulfolobus solfataricus* strain P2 (figure 1.3) is an aerobic heterotroph that oxidizes sulfur and grows optimally at 78°C and pH 2-4. It is relatively easy to grow for a hyperthermophilic acidophile and thus has become one of the best studied members of the Crenarachaeal phylum. Its genome has been sequenced, and numerous gene expression studies have been performed (Chong et al., 2007; Cooper et al., 2009; Frols et al., 2009; Maaty et al., 2009; Ortmann, 2008; She et al., 2001; Simon et al., 2009; Snijders et al., 2006). However, it is not an organism for which there is a tractable genetic system, likely because of the large number of insertion sequences (IS elements) present in the genome. Most genetic studies have been performed in a related strain *S. solfataricus* strain 98/2. Strain 98/2 has also recently been sequenced and has proven to be remarkably closely related to the P2 strain.

![Figure 1.3. Electron micrographs of *S. solfataricus*. A. Scanning electron micrograph. B. Transmission electron micrograph (Sue Brumfield).](image)
*Sulfolobus solfataricus* has a relatively large genome for the Crenarchaeotes, at approximately 3 megabases. Its genome is quite plastic with approximately 200 insertion (IS) elements that vary in size from 700 – 2000 bps (Brugger, 2007). There are also numerous fragments of formerly functional IS elements. Together, these IS elements account for nearly 10% of the entire genome (Brugger, 2007).

*Sulfolobus* spp. are ubiquitous around the world. They have been detected and isolated from springs in Italy, Iceland, Yellowstone National Park, Kamchatka, New Zealand and Japan (Zillig et al., 1998). Members of this genus are likely found in any environment within their growth conditions.

**Archaeal Viruses**

It has only recently been recognized that there is a high abundance of viruses on this planet. There are estimates of $10^{31}$ viruses globally (Suttle, 2007a), and in most aquatic environments, the numbers of viruses exceed the number of prokaryotic cells by one order of magnitude (Wommack and Colwell, 2000). This is an enormous reservoir of genetic material and biological diversity that is relatively untapped and little understood. Our limited knowledge is particularly evident in the study of viruses of the domain Archaea. There are only approximately 50 known archaeal viruses as compared to more than 5000 viruses infecting members of Eukarya and Bacteria. This is likely not a reflection of actual low numbers, but rather because we have only recently begun to
look for viruses of Archaea. The diversity uncovered in the relatively new area of study of archaeal viruses has been astounding.

The field of archaeal virology was pioneered by Wolfram Zillig and his colleagues, working primarily with *Sulfolobus* isolates. His laboratory collected samples from hot (75-95°C) acidic (pH 1-5) pools all over the globe. These samples were grown in enrichment cultures, and the DNA from total cultures examined. Any samples with unusual restriction fragment patterns, suggesting the presence of a high copy extrachromosomal element, were then spotted onto lawns of potential viral hosts to screen for plaque formation. This simple and straightforward method yielded over 30 novel viruses and 25 novel plasmids of various *Sulfolobus* spp. (Zillig et al., 1998). This screening method obviously relies on culturable hosts and production of a plaque phenotype. A true understanding of crenarchaeal viral diversity is not possible with such methods, but this was, and continues to be, a viable method of isolating novel viruses and studying them in depth.

The majority of crenarchaeal viruses have little homology to anything in the public databases except other crenarchaeal viruses. The morphological diversity of these viruses also exceeds that of the other domains. While the vast majority of bacteriophage (96%, 4950 of ~5100 total viruses) described have a head and tail morphology (Ackermann, 2001), this morphotype is rare for crenarchaeal viruses (Prangishvili and Garrett, 2004). Discovery of new viruses infecting hyperthermophilic crenarchaeotes has led to the creation of a number of new viral families. The
filamentous viruses, TTV1, TTV2, TTV3, TTV4, AFV and SIFV (Table 1.1), make up the family *Lipothrixviridae*. The rod shaped viruses, SIRV1 and SIRV2, make up the *Rudiviridae* family. The spindle-shaped viruses, or SSVs, make up the *Fusseloviridae*.

Several new families have also recently been added to account for this novel diversity. The family *Guttaviridae* has been proposed for the droplet-shaped virus, SNDV, *Ampullaviridae* for the bottle-shaped virus, *Bicaudaviridae* for the two-tailed virus, *Globulaviridae* for the spherical viruses (TTSV1 and PSV) and the *Sulfolobus* turreted icosahedral viruses (STIV1 and STIV2) will likely comprise a new viral family as well (Prangishvili and Garrett, 2004).

### Table 1.1. Viruses of the Crenarchaeotes

<table>
<thead>
<tr>
<th>Common Name</th>
<th>Family</th>
<th>Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfolobus spindle-shaped virus (SSV 1,2,RH,K,5,6,7)</td>
<td><em>Fusselviridae</em></td>
<td>Various <em>Sulfolobus</em> spp.</td>
</tr>
<tr>
<td>Sulfolobus turreted icosahedral virus (STIV 1,2)</td>
<td>Proposed new family</td>
<td><em>Sulfolobus sofataricus</em></td>
</tr>
<tr>
<td>Sulfolobus islandicus rod-shaped virus (SIRV 1,2)</td>
<td><em>Rudiviridae</em></td>
<td><em>Sulfolobus islandicus</em></td>
</tr>
<tr>
<td>Thermoproteus tenax virus (TTV 1,2,3,4)</td>
<td><em>Lipothrixviridae</em></td>
<td><em>Thermoproteus tenax</em></td>
</tr>
<tr>
<td>Acidianus filamentous virus (AFV)</td>
<td><em>Lipothrixviridae</em></td>
<td><em>Acidianus</em> spp.</td>
</tr>
<tr>
<td>Sulfolobus islandicus filamentous virus (SIFV)</td>
<td><em>Lipothrixviridae</em></td>
<td><em>Sulfolobus islandicus</em></td>
</tr>
<tr>
<td>Sulfolobus neozelandicus droplet-shaped virus (SNDV)</td>
<td><em>Guttaviridae</em></td>
<td><em>Sulfolobus neozelandicus</em></td>
</tr>
<tr>
<td>Sulfolobus tengchongensis spindle-shaped virus (STSV1)</td>
<td><em>Fusselviridae</em></td>
<td><em>Sulfolobus tengchongensis</em></td>
</tr>
<tr>
<td>Acidianus bottle-shaped virus (ABV)</td>
<td><em>Ampullaviridae</em></td>
<td><em>Acidianus</em> spp.</td>
</tr>
<tr>
<td>Acidianus two-tailed virus (ATV)</td>
<td><em>Bicaudaviridae</em></td>
<td><em>Acidianus</em> spp.</td>
</tr>
<tr>
<td>Thermoproteus tenax spherical virus (TTSV1)</td>
<td><em>Globuloviridae</em></td>
<td><em>Thermoproteus tenax</em></td>
</tr>
<tr>
<td>Pyrobaculum spherical virus (PSV)</td>
<td><em>Globuloviridae</em></td>
<td><em>Pyrobaculum</em> spp.</td>
</tr>
</tbody>
</table>
All crenarchaeal viruses described have dsDNA genomes, both linear and circular. The *Fuselloviridae* have been shown to exist both extrachromosomally as a plasmid and to integrate into their hosts' genomes in a specific manner at a tRNA gene, using a virally encoded integrase. Integration of the SSV viral genome is not required for completion of the viral replication cycle. Interestingly, the process results in a split integrase gene. The integrated viral genome can only be excised if there is an extrachromosomal copy of the genome present with an intact integrase gene (Prangishvili and Garrett, 2004). No other family of crenarchaeal virus is known to integrate into its host chromosome(s).

The majority of crenarchaeal viruses are maintained stably in their host, with only STIV (Brumfield et al., 2009) TTV1 (Prangishvili and Garrett, 2004), and SIRV2 being lytic (Prangishvili and Quax, 2011; Quax et al., 2011). Even UV or mytomyocin C induction of integrated SSVs does not result in lysis (Prangishvili and Garrett, 2004). Like the study of the host cellular organisms, archaeal virology has been limited by what hosts can be cultured in the laboratory. However, with more and more metagenomes becoming publicly available, we are seeing a rapid expansion of our understanding of archaeal viruses as well as environments in which they are detected. Because of their uniqueness, undoubtedly many archaeal viruses are still not being identified, and the field stands only to expand exponentially as more sequences become available and are correctly identified.
*Sulfolobus Turreted Icosahedral Virus (STIV)*

*Sulfolobus* turreted icosahedral virus (STIV) has been the subject of detailed structural, genetic, transcriptomic, proteomic and biochemical studies. This virus arguably serves as the model archaeal virus. As a result, more is known about STIV than most other viruses infecting members of the domain Archaea. Like most viruses isolated from crenarchaeal hosts, STIV has little in common with viruses of eukaryotic and bacterial hosts and should be considered the founding member of a new virus family. Despite this lack of obvious homology to other viruses, STIV has a components of gene content, replication strategy, and particle structure reminiscent of viruses of Eukarya and Bacteria, suggesting an evolutionary relationship between viruses from all domains of life (Khayat, 2005; Rice, 2004).

STIV was originally isolated from an acidic (pH 2.2) high temperature hot spring (82°C) in the Rabbit Creek thermal area of Yellowstone National Park (YNP) (Rice et al., 2004). The virus was isolated from an enrichment culture of an unsequenced *Sulfolobus* isolate, YNPRC179. This enrichment culture was screened by TEM for the presence of virus-like particles which were subsequently purified and characterized (Rice, 2004).

STIV has a limited host rage. The virus has been shown to infect the Italian isolates *S. solfataricus* strains P1, P2 and a highly susceptible strain derived from *S. solfataricus* P2, termed strain 2-2-12. Infections of other species, including *S. islandicus*, *S. acidocaldarius*, and *S. tokodaii*, have been unproductive (Ortmann, 2008; Rice, 2004; Chapter 2, table 2.2). The full host range of the virus is not known.
STIV is the first lytic virus infecting a member of the genus *Sulfolobus*. Infection of *Sulfolobus solfataricus* P2 strain 2-2-12 produces plaques and high virus numbers 36 hours post infection in liquid cultures (figure 1.4.a). Greater than 90% of cells viewed by electron microscopy show evidence of virus infection (figure 1.4.b) with densely packed cells (more than 50 viruses visible within a cell) observed (Brumfield et al., 2009). TEM analysis of a time course of the infection clearly demonstrates lysis of infected cells. Lysis appears to be an active process directed by virus gene expression, with alteration of the host S-layer preceding cell membrane disruption (figure 1.4.c,d.). TEM and SEM

Figure 1.4. Lysis of *S.solfataricus* by STIV. a. Lysed *S. solfataricus* 2-2-12 at t=48 post infection with STIV. b. STIV plaques on a lawn of *S. solfataricus* 2-2-12. c. SEM of STIV infected *S. solfataricus* 2-2-12 with pyrimids on the cell surface. d. TEM of STIV infected *S. solfataricus* 2-2-12 with pyrimids on the cell surface (Brumfield et al., 2009).
analyses of these cells show pyramid-like structures on the cell surface just prior to lysis. Recent work has shown that the viral gene, C92, plays a role in the formation of these structures (Snyder et al., 2011) and may be acting in concert with the cellular ESCRT (Endosomal Sorting Complex Required for Transport – a cellular system involved in trafficking proteins) (Samson, 2008) system to form the pyramids and traffic nascent viral particles to the cell surface (Snyder and Young). The exact mechanism by which this is accomplished, the specific host and virus genes that are involved and the biochemical details of the lysis process have yet to be elucidated.

**STIV Genome:** The virus encapsidates a circular 17,663 bp dsDNA genome that replicates episomally with no evidence of integration into the host genome (Thorrisdottir, 2006) (Figure 1.5). The current annotation of the genome includes 38 open reading frames (ORFs) (Maaty, 2006; Rice, 2004). Overall, there is little similarity to ORFs coded in the STIV genome with other genes in the databases. Bioinformatics approaches found similarity for three STIV ORFs to proteins in the public databases (C557, B116, C92) (Rice, 2004). One further ORF, B164, was predicted by homology modeling to be an ATPase (Maaty, 2006). The other 34 ORFs have no homology to proteins in the public databases. The lack of homology at the primary amino acid level to other known proteins suggests that function must be determined through further genetic, biochemical and or structural analysis. Crystal structures exist for STIV non-structural proteins, which have provided insights into their function. A197 is a putative glycosyl-transferase (Larson, 2006), and the virus capsid has been demonstrated to be
glycosylated (Maaty, 2006), however it has not been shown that A197 is responsible. B116 is a DNA binding protein (Larson, 2007b), and F93 is a winged-helix DNA binding protein (Larson, 2007a), suggesting that both may be involved in controlling viral and/or host gene expression.

**STIV Structure:** The virion is ~74 nm in diameter with an outer protein shell built on a pseudo $T=31$ lattice (figure 1.6.a). This is the first virus with this symmetry to be described. At each of the twelve five-fold axes there are turret-like appendages projecting from the virion surface from which the virus gets its name. These turrets project 13 nm from the virus particle surface and are approximately 24 nm in diameter at their broadest point (Rice et al., 2004). The function of these turrets is unknown, but

![Figure 1.5. Genome Map of STIV. Asterisks denote proteins shown to be associated with the particle. Light gray are cytoplasmic protein, dark gray are membrane proteins and black are proteins found in both the cytoplasm and the membrane (adapted from Maaty, 2011).](image)
one can speculate that they may be involved in receptor-mediated cell surface attachment, injection of DNA into the cell, and/or packaging of DNA into the capsid. Mass spectrometry has identified eleven proteins in the virus particle, nine of which are viral encoded (C557, C381, B345, A223, B164, B130, B109, A78, A55) and two host-derived (SSO7D and SSO0881) (Maaty, 2006). ORF B345 was identified by MALDI mass spectrometry as the major capsid protein component of the virion. The major capsid protein is glycosylated, but the nature of the carbohydrates and the chemical linkage to the coat protein subunit are unknown. Four of these ORFs are thought to code for proteins involved with the turret structure formation, including an ATPase (ORF B164). The functions of the remaining four viral coded virion proteins are unknown. The two host proteins found within the virion are a small DNA binding proteins (Sso7D) likely involved with packaging the viral genome and a host protein (SSo 0881) with a potential vacuolar sorting VPS24 domain that is found in both Saccharomyces cerevisiae and mammals and is conserved across the Sulfolobus family. The virion also contains a host derived internal lipid layer surrounding the nucleocapsid genomic core (figure 1.6.b). This internal membrane is comprised of a subset of host lipids that are highly enriched for acidic lipids (Khayat, 2005). The major capsid protein forming the exterior capsid shell is thought to anchor to the internal lipid through insertion of its C-terminus into the lipid bilayer. The mechanism by which this membrane is acquired and its function is unknown, but one can speculate that the function of this lipid envelope is to help
protect the viral genome from acid hydrolysis and or to help stabilize the virion to its high temperature environment.

![Cryo-EM reconstruction of STIV showing turrets (orange) at the five-fold axes of symmetry (Rice, 2004).](image1.png)

![Chimera cut-away image of STIV showing the internal lipid membrane (yellow) (Maaty, 2006).](image2.png)

**Figure 1.6.** Structure of STIV. a. Cryo-EM reconstruction of STIV showing turrets (orange) at the five-fold axes of symmetry (Rice, 2004). B. Chimera cut-away image of STIV showing the internal lipid membrane (yellow) (Maaty, 2006).

Comparative structural analysis has revealed interesting insights into virus evolution (Maaty, 2006). The cryo-EM reconstruction of the viron shows an overall virion architecture that is similar to viruses of the other two domains of life (figure 1.7). The crystal structure of the major capsid protein of STIV also reflects a common overall structure that is shared with viruses that infect eukaryotic host (adenovirus) and bacterial hosts (PRD1). This is the first example of a virus morphology found in all three domains of life. Structure predictions for several other proteins from STIV further support this relationship among viruses from all domains of life. C381 and A223 both share predicted structural similarity as determined by fold recognition algorithms with
the P5 protein of PRD1. The virion architecture, along with the tertiary structure of the capsid protein itself, suggests a virus lineage that spans all domains of life, despite the inability to detect this homology at the nucleotide or amino acid sequence level.

Figure 1.7. Comparison of the capsid architecture of STIV with a eukaryal virus and a bacteriophage.

**STIV Transcriptional Analysis:** Microarray experiments following an infected culture and comparing it to an uninfected culture have provided insights into the timing of events during viral infection as well as the host’s response to the virus. In a near-synchronous, single replication cycle, there was evidence for limited temporal regulation of viral gene expression. Transcription of virus genes was first detected at 8 hours post infection (hpi), with a peak at 24 hpi. Lysis of cells was first detected at 32 hpi. There were essentially three sets of genes that appeared to be co-regulated. A number of genes were first detected at 8 hpi: C92 and B121 as well as two gene clusters, B116, C118, A53, and A66 and the region from C57 to A106 along with A106 and A137, suggesting that these two regions might be transcribed as operons. The majority of the genes, including all of the structural proteins, showed peak transcription
at 16 hpi. A small set of genes was not transcribed until late (F75, F93 and D66 at 24 hpi). Two genes, A61 and B124, were not detected (figure 1.8). One gene, C67, was transcribed in coordination with the viral structural genes, but it has not been identified as part of the virion.

![Figure 1.8. Transcriptional timing of STIV ORFs. A Genome map of the virus showing differentially regulated genes by color-coding (Ortmann, 2008).]

During the infection, 177 host genes were determined to be differentially expressed, with 124 genes up regulated and 53 genes down regulated. The up regulated genes were dominated by genes associated with DNA replication and repair and those of unknown function, while the down regulated genes, mostly detected at 32 hpi, were associated with energy production and metabolism. The observed patterns of transcription suggest that up regulated genes are likely used by the virus to reprogram the cell for virus replication, while the down regulated genes reflect the eminent lysis of the cells. The majority of the host genes (94%), however, showed no differential regulation between infected and uninfected cultures, suggesting that the virus life cycle is tailored to avoid a host stress response. This is further corroborated by the absence
of typical heat shock or UV-induction stress response type genes. And in fact, some genes that are up-regulated under other stresses are found to be down-regulated in response to STIV infection.

**Genetic Systems for Viruses of Sulfolobus**

The discovery of these novel viruses and plasmids led to the first development of shuttle vectors for *Sulfolobus*, allowing for the beginning of genetic analysis of this genus. Both viruses and plasmids were altered to include the *lacS* gene for screening and an *E. coli* vector for propagation in that host. Later work introduced selectable markers and ultimately allowed for recombination based knockouts of endogenous cellular genes (Jonuscheit, 2003).

*Sulfolobus* spindle-shaped virus (SSV) was the first crenarchaeal virus for which a shuttle vector was constructed. An *E. coli* vector, pUC18, was randomly cloned into SSV1 viral genome at an *Alul* site. This vector was engineered with a *pyrEF* gene cassette, allowing for uracil selection, and a *lacS* reporter gene. This construct can be propagated in both *E. coli* and *Sulfolobus* spp. and has robust selection and screening in both cell types. Because the vector produces virus particles, it is self-spreading through the culture, overcoming low transformation efficiency. However, one limitation in this system is that the SSV virus irreversibly integrates into the host genome. This system also relies upon transformation of the vector into an uracil auxotrophic *Sulfolobus* strain that lacks an endogenous *pyrEF* gene. There are other vectors for *Sulfolobus* spp. that
utilize the antibiotic hygromycin (Cannio 2001), however they have not been as widely used.

**STIV Genetic System:** While STIV is one of the better understood viruses infecting a crenarchaeal host, much remains to be examined. The ability to understand the replication cycle of this virus, the details of its interactions with its host, and the function of its viral gene products would be greatly facilitated by the development of a genetic system for STIV. There is currently only one other crenarchaeal virus genetic system, for *Sulfolobus* spindle-shaped virus (SSV) (Stedman, 1999). In order to expand our knowledge of STIV's life cycle as well as contribute to the genetic tools available for the study of *Sulfolobus* spp., we undertook the construction of an infectious genetic clone of STIV. With this genetic clone in hand, we then were able to disrupt and knock out several open reading frames (ORFs) in the virus genome (described in chapter2).

**Virus-Host Interactions**

Viral predation is likely a major factor, if not the most important factor, in shaping cellular (prey) populations. The manner in which predator (virus) and prey (microbe) interact has been described as the “Kill-the-Winner” theory. This theory predicts repetitive cycling of predator and prey populations. As one organism becomes the most dominant it will be rapidly reduced by predation from viruses. This allows one viral type to briefly become dominant, but then decline as its host declines (Rodriguez-Brito et al., 2010). The dominant cellular population is replaced by another population,
and the cycle continues. This constant and rapid evolution of the most abundant microbial strain and its predators is part of an arms race termed the “Red Queen” dynamic (Gandon, 2008). Regions of the host genome responsible for viral defense are often seen to undergo rapid changes. For example, cell surface receptors and the Clusters of Regularly Interspaced Palindromic Repeats (CRISPRs; described in detail below) are both areas that see high levels of genetic mutation in order to escape viral predation, while genes involved in entry and the viral target of a CRISPR spacer are areas within a viral genome that mutate rapidly in order to evade host defenses (Gandon, 2008). The result is constantly escalating infectivity and resistance.

Looking into complex natural environments, this mechanism is not clear. Analysis of infectivity and resistance of a soil host-parasite system showed that bacteria were more resistant to contemporary phage than to past or future phage infections. In this series of experiments, researchers infected stocks of *Pseudomonas flourescens* SBW25 with an associated phage, SBW25Φ2, in the presence or absence of the naturally occurring soil community. In the arms race dynamic, one would expect that the host would be most resistant to a previous phage and least resistant to a future phage. Instead, they found that the host was most resistant to the contemporary phage. This suggests that the continual escalation of resistance and infectivity is not occurring, but rather the operation of fluctuating selection dynamics, where a parasite specializes on a particular host, and the fitness of each fluctuates through time. In this scenario, co-evolution can continue indefinitely and maintain genetic diversity within the
populations, rather than relying on continual introduction of novel mutations and the occasional purging of diversity as occurs with the arms race dynamic (Gomez, 2011). How the CRISPR/Cas immune system might impact viral dynamics within a population characterized by fluctuating selection dynamics has not previously been explored.

**CRISPR/Cas: A Newly Identified Prokaryotic Immune System**

The CRISPR/Cas (Clusters of Regularly Interspaced Palindromic Repeats and the CRISPR associated genes) is a newly described adaptive immune system found in Bacteria and Archaea. The CRISPR loci (figure 1.9) are encoded on the cellular chromosome and are usually physically linked to a set of conserved cas genes. The CRISPR array maintains memory of viral encounters that the cell has seen which allows for a rapid response upon subsequent infection. The cas genes encode the protein components that mediate an interference mechanism.

![Figure 1.9](image)

*Figure 1.9 CRISPR locus and Cas genes from* *Streptococcus thermophilus* *strain DGCC7710* *(Deveau, 2010).*

**History of the Discovery of the CRISPR/Cas System:** CRISPRs were first identified in 1987 in *E.coli* (Ishino et al., 1987). While sequencing the *iap* gene, researchers noted an unusual structure downstream from the gene. They observed five sets of 29 bp identical repeats separated by 32 bp “spacer” sequences. The function was unknown,
but they speculated that the dyad symmetry of the repeat units might function to stabilize mRNAs upon transcription of the region.

With the advent of higher throughput methods of sequencing, ever-increasing numbers of microbial genomes were being sequenced. The analysis of these genomes has shown that these repeat/spacer loci are ubiquitous in Bacteria and Archaea. Nearly half of all bacterial species (537 of 1193 total) and more than 83% of all archaeal species (80 of 96 total) sequenced to date contain these repetitive arrays (CRISPRdb, 2011).

The CRISPR loci can vary greatly in size; from as few as two to 375 repeat/spacer units have been observed. They are flanked by a conserved leader sequence of up to 500 base pairs that tends to be AT rich and contains the promoter. This is followed by a series of identical repeats of 21-48 bps separated by unique spacers of 26-72 bps. There are no open reading frames encoded within the locus (Deveau, 2010). Some organisms have repeats with palindromic structures (thus the name), while others lack this structure. A single organism can have anywhere from one to 20 or more of these CRISPR arrays in its genome (CRISPRdb).

It was also recognized in analyzing large numbers of microbial genomes that there were a set of genes that were almost always associated with the CRISPR arrays (Jansen et al., 2002). These were termed the CRISPR associated genes (cas genes). These genes are often, although not always, found as a cassette flanking the array. Moreover, species that lack CRISPRs, also lack cas genes. The CRISPR/Cas system was originally predicted to be involved in either DNA repair or chromosome partitioning
because of the predicted nucleic acid binding and modifying functions of these genes (Jansen, 2002).

In 2005, three independent groups noted that the CRISPR spacers were not, in fact, “junk” sequences. The majority of all spacers had no homology to anything in the non-redundant databases. However, the small number of spacers that did match something in the databases almost universally matched viruses, plasmids or other mobile elements, and they were almost always exact matches. This led to the idea that CRISPRs might be functioning as an adaptive immune system (Mojica, 2005; Pourcel, 2005; Haft, 2005).

The Cas protein machinery consists of three different types of genes: the core genes (cas1-6), subtype specific genes (cse, csa, cst, csm, csy, csh, csn and csd) and repeat associated mysterious protein (as many as 90 RAMPs) modules (cmr1-6) (Haft, 2005). Any given CRISPR-containing organism will encode between 4 and 10 cas genes: several “core” genes, usually linked to the CRISPR locus, and one or more of the eight subtype sets along with a RAMP module (Makarova, 2006).

The core cas genes consist of cas1-6 and are beginning to have their functions defined. Cas1 and cas2 are the mostly highly conserved and have come to function as markers for the presence of the CRISPR/Cas system in a species. Cas1 is a DNA endonuclease (Wiedenheft, 2009) with RNA and DNA binding activity (Han, 2009). The structure of Cas1 shows a novel fold with two distinct domains, an N-terminal domain made up of ten β-strands and two α-helices, and a C-terminal domain made up of ten α-
helices and contains a conserved divalent metal ion-binding site (Wiedenheft, 2009). Cas2 is a small protein with an N-terminal β-strand followed by a polar amino acid and has been shown to have endoribonuclease activity (Beloglazova, 2008). These two proteins, along with Cas4 are thought to be involved in the acquisition and integration of new spacers. Cas4 appears to be a RecB-like protein with cysteine residues near the C-terminus that may be involved in DNA binding (Haft, 2005; Jansen 2002). Cas3 has an HD domain that acts as a metal-dependent nuclease on ds-oligonucleotides and a DEAD/H box helicase domain (Haft, 2005; Jansen 2002). This protein is thought to be involved in interference of an invading nucleic acid (Haft, 2005; Jansen, 2002). Cas5 is a protein of unknown function with a conserved N-terminal domain, but a divergent C-terminus across different organisms (Haft, 2005). Cas6 has RNase activity and is involved in processing the CRISPR transcript into short RNAs (Carte, 2008; Haft, 2005).

In 2007, it was shown that CRISPR/Cas system does function as an adaptive immune system under laboratory conditions. This was first demonstrated when Streptococcus thermophilus was infected by the lytic phage 2972; those members of the cellular population that survived had acquired a new CRISPR spacer specific to that phage. Deletion of phage specific spacers in a resistant strain made that strain susceptible once again. They also demonstrated that this response was enhanced by incorporation of multiple spacers to the same phage, but was abolished by a single mismatch within the spacer (Barrangou, 2007).
Model for Function of CRISPR/Cas System in Cultured Isolates: The CRISPR/Cas system is thought to function in two distinct stages (figure 1.10a): adaptation and interference. The first stage, adaptation, is the integration of a new spacer into the host.

Figure 1.10. Stages of CRISPR/Cas function (Deveau, 2010).
encoded CRISPR array the first time that the host encounters a virus or other mobile element. The CRISPR locus is constitutively transcribed, and the mRNA cleaved into functional units, crRNAs between 68-75 nts in length. The interference stage is directed by the crRNAs against subsequent invasion by a previously encountered mobile element (figure 1.10b).

During adaptation, spacers have been shown to be incorporated in a polar manner at the leader end of the array (Lillestol, 2009; Barrangou, 2007) (figure 1.10a.). This was first predicted by analyzing two related Sulfolobus strains, S. Solfataricus strains P1 and P2. Comparison of their CRISPRs showed conservation at the downstream end, but high levels of variability at the leader end. It was subsequently demonstrated in S. thermophilus that new phage specific spacers were incorporated at the leader end of the array (Barrangou, 2007). This suggests that the array itself may provide a record of previous infections, with the more recent encounters encoded at the upstream end and older infections at the downstream end. This agrees with the observation that the most downstream repeats are often degenerate, suggesting that the distal end of the array is no longer functional.

The Interference stage can be further broken into six steps (figure 1.11). The CRISPR array is constitutively transcribed from the leader end (figure 1.11.1). While the exact function of the leader is unknown, it has been suggested that it is involved in acquisition of new spacers, and in E. coli, it has been shown that the leader containsthe promoter (Pul, 2010). The CRISPR transcript is processed into short crRNAs the size of
one repeat/spacer unit. In *E. coli*, a set of five Cas proteins (CasABCD and E; alternatively, Cse1-4 and Cas5e), termed the CASCADE (CRISPR Associated Complex for Figure 1.11 CRISPR/Cas Interference stage (Deveau, 2010).

Anti-viral Defense) complex, has been shown to be involved with the crRNA processing. Specifically, CasE is required for processing, and further *in vitro* activity assays showed
that CasE is a metal-independent endoribonuclease that specifically cleaves precursor-crRNAs into mature crRNAs. Post-processing, the short crRNAs remain bound to the CASCADE complex.

Cloning of the crRNAs bound to the CASCADE complex showed far more products from the leader end of the array than from the downstream end. This suggests that transcription is terminated prematurely, providing further evidence that the downstream end represents older viruses, from which the host no longer needs protection.

The crRNAs function along with Cas proteins to target invading nucleic acids, likely both DNA and RNA depending on the system. In *Streptococcus epidermidus*, researchers showed the system targeted DNA by inserting a self-splicing intron into the CRISPR target sequence within the nickase gene on a conjugative plasmid. The mRNA from this gene would reconstitute the CRISPR target by splicing, while the DNA would not contain the target sequence. The intron resulted in disrupting the interference mechanisms, showing that the DNA was the target of the CRISPR spacer (Maraffini, 2008). However, in *Pyrococcus furiosus* it was shown that a complex of RAMP proteins bound to a crRNA could recognize and cleave complementary RNA (Hale, 2009). In *E.coli* and *Pyrococcus furiosus*, it has been shown that the CASCADE complex along with specific crRNAs and Cas3 are all required to direct interference against an invading nucleic acid.
A similar CASCADE like complex has recently been described from *S. solfataricus* and has been termed aCASCADE (archaeal CASCADE). The aCASCADE complex purified from *S. solfataricus* contains Csa2 (Cas7/CasC) and Cas5a (CasD/Cas5e) and bound crRNAs. When expressed in *E. coli*, the Csa2-Cas5a complex co-purified with Cas6, Csa5 and possibly Csa4 (all either at lower abundances or with only weak interactions). This complex bound to crRNAs also recognized single stranded “target” DNA *in vitro*. All *E. coli* CASCADE components except CasA and CasB were identified in *S. solfataricus*, and the system appears to function similarly as in *E. coli* (Lintner, 2011).

Given the presence of a crRNA target sequence with the host encoded CRISPR locus itself, self-recognition becomes an important factor in avoiding auto-immunity. The existence of a short motif in *S. thermophilus*, termed a proto-spacer associate motif, or PAM, just outside of the proto-spacer region of the target sequence, and not present in the host CRISPR locus, may allow the host cell to discriminate between self and non-self sequence (Horvath, 2009). In *S. epidermidis* there is no such motif. However, it has been shown in this organism that pairing between nucleotides in the repeat region distinguishes the host CRISPR from a target foreign sequence (Maraffini, 2010).

Recently the CRISPR/Cas system has been shown to function in the archaeal model organism *S. solfataricus* (Manica, 2011). Researchers engineered ORF406 from the *Sulfolobus* plasmid pNOB8 into the *E.coli/S.solfataricus* shuttle vector, pMJ03 that is based on the virus SSV1 and produces virus particles and plaques. *S. solfataricus* P2 contains a spacer match (with 7 mismatches) to ORF406. They constructed a number of
versions of this plasmid, the WT version with seven mismatches, an exact match between spacer and ORF406, and two deletions in the middle of the protospacer, one of 1bp and one of 3bps. They demonstrated significant impairment of transformation efficiency with the complementary spacer; however, they still demonstrated reduction of transformation with seven mismatches as well as with both deletion mutants. This clearly demonstrates an immune function for the CRISPR/Cas system in *S. solfataricus*; however, in contrast to other systems, interference was still effective with up to three mismatches and detectable at seven mismatches (Manica, 2011).

**Origins and Evolution of CRISPR/Cas Systems:** While the CRISPR/Cas system is a heritable adaptive immune system, it also appears to propagate by horizontal gene transfer (HGT). The phylogenetic tree of CRISPR systems does not agree with the established phylogeny of Bacteria and Archaea (Karginov, 2010). Furthermore, closely related organisms often have divergent spacer content, while the converse is also true (Held, 2009; Godde, 2006; Haft, 2005). Comparison of Sargasso Sea sequences and a closely related *Shewanella oneidensis* strain shows that there is a high degree of conservation among the genes flanking the CRISPR region between these two organisms. However, *S. oneidensis* does not contain a CRISPR, suggesting that the Sargasso Sea organism has acquired a CRISPR system by HGT (Godde, 2006). CRISPRs have been identified on plasmids (Haft, 2005; Godde, 2006; Manica, 2011) providing a potential mechanism of exchange.
**CRISPR/Cas Function in Natural Populations:** The operation and function of the CRISPR/cas system in natural environments have not been extensively analyzed. Most work describing the function and the dynamics of the CRISPR/cas system has been performed in clonal bacterial populations in the laboratory. The mechanism of function in archaeal systems is less well defined, and function within a natural population has not been directly shown. An examination of natural archaeal and bacterial populations associated with acid mine drainage by metagenomic approaches has demonstrated that there is a high amount of diversity within the CRISPR arrays in non-clonal natural populations (Tyson, 2008). Hyperdiversity within this region in natural populations is thought to play a role in allowing the population as a whole to respond to phage predation, while any given individual within the population has a very limited set of spacers, and thus is only protected from a small subset of potential viral infections.

The CRISPR/Cas system is just one of many known anti-viral defense systems operating in Bacteria and Archaea. Other known virus defense mechanisms in Bacteria and Archaea include the Restriction Modification System (Bickle, 1993), Abortive Resistance (Chopin, 2005) and altering cell surface receptors (Forde, 1999). While work in culture has shown that addition of a single new spacer into a functioning CRISPR array is sufficient to provide protection against virus infection, it is suspected that in natural populations multiple mechanisms are playing a role in protecting cells from predation. These added layers of complexity make it much more difficult to determine the exact function of the CRISPR/cas system in a complex natural environment.
Objectives of this Research

The three primary objectives of this research were to gain a better understanding of the STIV life cycle and its interaction with its host *S. solfataricus*, to expand our knowledge of archaeal viral diversity through culture-independent sequencing approaches, and to probe the function of the newly described CRISPR/Cas immune system function in a natural environment.

**Objective #1:** Develop a genetic system for STIV In order to better elucidate the life cycle and interactions with its host, *S. solfataricus*. Specifically, we constructed an infectious clone of the virus and used it to probe gene function by gene disruptions, site-directed mutagenesis, knockouts and gene complementation experiments.

**Objective #2:** Demonstrate the feasibility of metagenomics sequencing in a high temperature acidic environment and avoid the biases inherent in culturing hosts and their viruses and to expand our understanding of viruses that we might not be able to access using traditional culturing methods. Specifically we sequenced the total viral fraction of Road Side Spring thermal feature in Yellowstone National Park (YNP) and assembled and analyzed a near full-length novel viral genome. This was the first viral metagenomic sequencing effort undertaken in a hot acidic environment.

**Objective #3:** Better understand the function of the newly described CRISPR/Cas immune system in natural environments as compared to what has been observed in
laboratory cultures. We used a paired host-virus metagenome sequencing approach combined with analysis of a PCR amplified clone library over the course of two years of sampling at a Nymph Lake hot spring (YNP). These data sets were compared to a time course of viral metagenomic samples spanning the same time period. Our goal was to better define how this immune system (CRISPR/Cas) functions in a complex natural environment and determine if its function could be correlated to the viral population present in that environment.
CHAPTER 2

DEVELOPMENT OF AN INFECTIOUS CLONE OF *SULFOLOBUS* TURRETED ICOSAHEDRAL VIRUS (STIV) AS A TOOL FOR EXAMINING STIV GENETICS

Introduction

There is only a limited understanding of the viruses that infect cellular hosts belonging to the Archaea. To date, only a limited number of archaeal viruses have been isolated. There are a total of 49 crenarchaeal and euryarchaeal viral genomes sequenced (Ackermann and Kropinski, 2007; Happonen et al., 2010; Prangishvili et al., 2006; Redder et al., 2009; Snyder et al., 2003). Few of the open reading frames (ORFs) in any of these viruses have significant similarity to known genes in the public databases. Infectious clones have been developed only for members of the *Fuselloviridae* (Stedman et al., 1999; Jonascheit et al., 2003; Berkner and Lipps, 2008) and the spherical euryarchaeal viruses (Porter and Dyall-Smith, 2008).

*Sulfolobus solfataricus* has emerged as a model organism for the phylum Crenarchaeota. It is an aerobic heterotrophic thermoacidophile (≈80°C and pH 2-4 optimum) that can be easily cultured in the laboratory. Members of the *Sulfolobales* have been detected in high temperature (72-85°C) acidic (pH 2.0-4.0) hot springs, fumaroles, and mud pots throughout the world (Brock et al., 1972). The genome of *S. solfataricus* strain P2 has been sequenced (She et al., 2001) and a genetic system has been developed for a close relative of this strain (Berkner and Lipps, 2008; Contursi et
al., 2003; Schelert et al., 2004; Wagner et al., 2009). Because of the ease of culturing this genus, 15 viruses have been identified that infect *Sulfolobus* spp. (Arnold et al., 2000a; Arnold et al., 2000b; Prangishvili et al., 1999; Rice et al., 2004; Schleper et al., 1992; Stedman et al., 2003; Wiedenheft et al., 2004; Xiang et al., 2005).

*Sulfolobus* turreted icosahedral virus (STIV) was originally isolated from a high temperature (82°C) acidic (pH 2.2) hot spring in the Rabbit Creek thermal area of Yellowstone National Park, USA. The virus was isolated from an enrichment culture of an uncharacterized *Sulfolobus* species (Rice et al., 2004). The purified virus infects the Italian isolate of *Sulfolobus solfataricus* strain P2. The virus particle is a 74 nm icosahedron (T=31) with turret-like projections extending from its five-fold axes of symmetry (Rice et al., 2004; Khayat et al., 2005; Khayat et al., 2010; Chi-yu Fu et al., 2010). The virus encapsidates a 17663 base pair double-stranded circular DNA genome encoding 38 open reading frames. Few of the ORFs have any similarity to known genes in the public databases and thus are difficult to annotate without structural or functional studies (Khayat et al., 2005; Larson et al., 2007a; Larson et al., 2007b; Larson et al., 2006; Maaty et al., 2006; Rice et al., 2004; Chi-yu, 2010; Khayat, 2010). The replication cycle of STIV is not known, but comparative genomic and structural analysis has provided insights into several STIV gene functions. Proteomic analysis of purified virions has demonstrated that the virion is comprised of nine viral proteins (C557, C381, B345, A223, B164, B130, B109, A78, A55) and two host proteins (SSO7D and SSO0881) (Maaty et al., 2006). B345 has been determined to be the major capsid protein
(MCP) (Rice et al., 2004; Khayat et al., 2005), and its structural analysis has provided evolutionary insights into the relationship between icosahedral viruses from all three domains of life (Khayat et al., 2005; Maaty et al., 2006). B164 has been shown by homology modeling to likely be a viral ATPase (Maaty et al., 2006). The X-ray crystal structures of three additional STIV proteins (F93, B116 and A197) have been determined. F93 is a winged-helix DNA-binding protein that is a putative transcriptional regulator (Larson et al., 2007a). B116 is a novel DNA binding protein that has been shown to bind DNA in a non-specific manner (Larson et al., 2007b). A197 is a putative glycosyl-transferase (Larson et al., 2006).

While STIV is one of the better understood viruses infecting a crenarchaeal host and has emerged as a model crenarchaeal virus system, much remains to be examined. The ability to understand the replication cycle of this virus, the details of its interactions with its host, and the function of its gene products would be greatly facilitated by the development of an infectious clone that allows for the directed manipulation of the STIV genome. There is currently only one other crenarchaeal virus for which there is an infectious clone, *Sulfolobus* spindle-shaped virus (SSV) (Stedman et al., 1999). This chapter describes the development of an infectious STIV clone and its utility for probing viral gene function.
Materials and Methods

Construction of an STIV Full-length Clone

A 21 kb full-length STIV clone (17.7kb viral genome plus 3.9 kb *E.coli* plasmid vector) was constructed by sequential addition of PCR products to the pCRII-Topo-TA *E.coli* cloning vector (Invitrogen). Briefly outlined, native viral DNA isolated from purified STIV virions was used as the template for all PCR reactions. The 17.7 kb genome of STIV was PCR amplified using iProof high fidelity DNA polymerase (BioRad) in three separate overlapping segments (7436bp, 4488bp and 5849bp segments) using oligonucleotide primers that either spanned unique restriction endonuclease sites or, in the case of 1F and 1R, created a single base change that engineered a new unique BamHI site (table 2.1).

“A” overhangs were added to the blunt iProof generated PCR products by the following reaction: 2 μl PCR product, 2 μL 5x GoTaq buffer, 1 U GoTaq, 1 μL 20 mM dATPs in a 10 μl reaction at 72°C for 10 minutes. Each PCR product was cloned into the pCRII-Topo-TA plasmid (Invitrogen) per manufacturer’s protocol.

The full-length genome of STIV was constructed within the pCRII-Topo-TA plasmid by sequential addition of the three viral PCR products (figure 2.1). This was accomplished by first cloning the 4488 bp STIV genome segment from position 14586-1410 (through STIV position 0) into the pCRII-Topo-TA plasmid to produce an 8461 bp clone. This 8461 bp clone was used as the destination plasmid for the addition of all subsequent STIV PCR products. Because the cloning strategy involved using the unique
An Ncol site in STIV to insert subsequent pieces into this destination vector, and the pCRII-Topo-TA vector also contains a Ncol site, we removed the vector Ncol site (described below). This restriction site fell within the Kan^R gene, and filling the Ncol site inactivated the Kan^R gene. Therefore the destination vector had Amp resistance but was Kan sensitive.

![Diagram](image)

Figure 2.1. Schematic of the construction of the full-length STIV clone.

The Kan sensitive 8461 bp clone was digested with Ncol and BstXI (Promega) to remove a 51 bp fragment from the plasmid polylinker region to the Ncol site in the STIV genome sequence. A 11408 bp STIV clone, encompassing the STIV genome from positions 7188-14623, was digested with BstXI and Ncol and ligated into BstXI and Ncol linearized 8461 bp clone using T4 DNA ligase (Promega) producing a 15802 bp.
Table 2.1. STIV Primers.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Restriction site</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIV1F</td>
<td>cggtgagaagatcaaaagaggtggatccagagaaggcg</td>
<td>engineered</td>
</tr>
<tr>
<td>STIV 1R</td>
<td>cgccctcttgatcaccctcttcgtcttcaccgc</td>
<td>engineered</td>
</tr>
<tr>
<td>STIV 2F</td>
<td>gataagagagagaatccatgggtgttatcccccttccc</td>
<td>Ncol</td>
</tr>
<tr>
<td>STIV 2R</td>
<td>gggagggggtacaacatgattacaattcttctctatc</td>
<td>Ncol</td>
</tr>
<tr>
<td>STIV 3F</td>
<td>gctaaaacttcatctcaggtgaattaggaatacg</td>
<td>AvrII</td>
</tr>
<tr>
<td>STIV 3R</td>
<td>cgtatctcactacatctgtagattgagatgttttaggc</td>
<td>AvrII</td>
</tr>
<tr>
<td>A197 D151NF</td>
<td>gcattggatagccgaacagatcttccccccctcactc</td>
<td>engineered</td>
</tr>
<tr>
<td>A197 D151NR</td>
<td>gcattggatagccgaacagatcttccccccctcactc</td>
<td>engineered</td>
</tr>
<tr>
<td>A197 D151HF</td>
<td>gcattggatagccgaacagatcttccccccctcactc</td>
<td>engineered</td>
</tr>
<tr>
<td>A197 D151HR</td>
<td>gcattggatagccgaacagatcttccccccctcactc</td>
<td>engineered</td>
</tr>
<tr>
<td>A197 D151EF</td>
<td>gcattggatagccgaacagatcttccccccctcactc</td>
<td>engineered</td>
</tr>
<tr>
<td>A197 D151ER</td>
<td>gcattggatagccgaacagatcttccccccctcactc</td>
<td>engineered</td>
</tr>
<tr>
<td>A197 Xmal F</td>
<td>cttcttagttagcttttttcag</td>
<td>XmalI</td>
</tr>
<tr>
<td>A197 AvrIl R</td>
<td>ctaattcacttctactgtgtagat</td>
<td>AvrII</td>
</tr>
<tr>
<td>A197 KO AgeI R</td>
<td>gcatacgggtgctctctctactgtcact</td>
<td>engineered</td>
</tr>
<tr>
<td>A197 KO AgeI F</td>
<td>gcctagtgggtgtctactgtcact</td>
<td>engineered</td>
</tr>
</tbody>
</table>
| A197 AgeI     | caatactcagggtggatgatatgtgatgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt
intermediate clone. In the final assembly step, a 9821 bp STIV clone encompassing positions 1376-7224 was digested with NotI (Promega) and AvrII (New England Biolabs) and ligated into NotI and AvrII digested 15802 bp intermediate clone. The final construct was a 21,582 bp clone (17697 bp of STIV genome with complementary terminal repeated sequences and 3885 bp of plasmid vector) and designated pFLSTIV. *E. coli* (DH10B XL2 Blue, Stratagene) harboring pFLSTIV were grown and plasmid DNA purified by standard protocols (Sambrook et al., 1989). The final construct was confirmed by restriction endonuclease digest mapping and by selected DNA sequencing.

**Removal of Ncol Site in *E.coli* Vector**

Approximately 100 ng of the 8461 bp clone was digested with Ncol in a 20 μL reaction. The reaction was stopped at t=0, 5 s, 10 s, 30 s, 1 m, and 5 m by transfer to preheated (55°C) Proteinase K (40 μg, Ambion) in 50 μL TE (pH 8). The digests were visualized on a gel and those with the appearance of a linear molecule (t=5 s and t=10 s) were cleaned-up with a PCR clean-up kit (Qiagen), followed by end repair with Klenow fragment (Promega), ligation with T4 DNA polymerase (Promega), and transformation into *E.coli* (XL2 Blue cells, Stratagene). Colonies were grown on LB Amp and were patched onto LB Kan plates. Colonies that were Amp\(^R\) and Kan\(^S\) were analyzed.

**Construction of STIV Gene Disruption Mutants**

Different strategies were employed to construct both gene disruptions and gene modifications within the STIV genome. Frame-shift mutations were introduced into STIV
genes B116, A197, and C381; deletion mutations were created in A197 and B345; and amino acid additions and substitution mutations were introduced into B345 and A197.

The frame-shift mutations in B116, A197, and the single amino acid insert in B345 were performed by selected restriction endonuclease digestion of pFLSTIV, followed by end repair with Klenow fragment (Promega) to introduce a frame shift or insert within each targeted ORF. To create a B116 frame shift mutation, pFLSTIV was digested with PacI. After the end repair reaction, two base pairs were added, creating a frame shift at position 85, introducing a stop codon at position 88 and altering the intervening codons. The presence of the frame shift was confirmed in the progeny viral particles by DNA sequencing. To create an A197 frame shift mutation, pFLSTIV was digested with SacII and end repaired. This resulted in the insertion of two nucleotides, creating a frame shift at residue 179 and the introduction of a stop codon at position 183. An amino acid insertion into B345 was created by digestion of pFLSTIV with Bsu361 followed by end repair, inserting three nucleotides, resulting in the insertion of a single Serine residue at position 223. Each end repair reaction was subsequently ligated and transformed into E. coli. Colonies were screened by PCR using primers flanking the region of interest and spanning the filled restriction endonuclease site (table 2.1). Constructs were digested with the appropriate restriction endonuclease to confirm the loss of the targeted restriction endonuclease site and were confirmed by DNA sequencing with primers A197 SacII screen F and R, B116 PacI screen F and R, or B345 5'F and B345 mid R (table 2.1).
Point mutations were introduced into A197 and C381 by site directed mutagenesis using the QuickChange kit (Stratagene). For A197, a 3842 bp fragment was amplified from the WT clone using iProof polymerase (BioRad) and cloned into the pSMART HCKan vector (Lucigen). This fragment encompassed the A197 gene and two unique restriction sites, XmaI and SacII (Promega). Three point mutation in the putative active site of A197 were engineered: D151E, D151N, and D151H. The XmaI and SacII sites were subsequently used to excise the mutated A197 gene from the subclone and ligate it into the full-length clone. All mutations were confirmed by sequencing with the A197 R seq primer (table 2.1). For C381, a 4007 bp fragment was amplified from the WT clone using iProof polymerase and cloned into the pCRII-Topo-TA vector (Invitrogen). This fragment included the C381 gene and the unique restriction sites, SacII (Promega) and PacI (New England Biolabs). One mutant was the result of a point mutation that changed Lysine 158 into a stop codon, and the other mutation was a single base pair deletion at base pair 105, which resulted in a frame shift. The two unique restriction sites were used to ligate the mutated piece back into the full-length viral genome. All mutations were verified by DNA sequencing with primers C381 331F and C381 602R or C381 331R and C381 602R (table 2.1).

The deletion mutant of B345 was created by deleting the C-terminal 22 amino acids, which had been hypothesized to interact with the internal lipid layer (Maaty et al., 2006). A B345 ‘sub-clone’ was created by amplifying the B345 and C557 genes from the STIV genome. This 4488 bp fragment was cloned into the pCRII-Topo-TA vector
following manufacturer’s guidelines. In order to delete 66 bps from the C-terminal end of B345, inverted primers were designed that would engineer a unique AgeI site in the place of the 22 amino acids. After amplification, the product was digested with AgeI (Promega) and ligated, recircularizing the plasmid. The mutant was screened for the presence of the AgeI site and by DNA sequenced with primers B345-AgeI seq -F and R (table 2.1). The B345 C-terminal deletion mutant was ligated back into the STIV full-length genome using the unique restriction sites Bsu36I and BsiWI (New England Biolabs).

A deletion mutant of A197 was constructed by removing 569 nucleotides of the gene, leaving 5 nucleotides at the 5’ end and 15 nucleotides at the 3’ end. An A197 ‘sub-clone’ was constructed by amplifying the region of STIV encompassing the A197 gene and spanning Xmal and AvrII unique restriction sites using the A197 Xmal F and A197 AvrII R primers. The resulting product was cloned into the pCRII-Topo-TA vector to produce a 9085 bp clone. The deletion of the A197 gene was then constructed by inverted PCR using primers A197 KO AgeI F and A197 KO AgeI R, which added an AgeI site at the ends of the amplicon. The PCR product was then digested with AgeI and ligated together, resulting in a circular plasmid with 569 bps of the A197 gene deleted. This product was then transformed into XL2 Blue E. coli cells (Stratagene) and correct colonies were screened for the presence of the AgeI site and further confirmed by DNA sequencing with primers A197 AgeI screen F and R (table 2.1). The A197 deletion
mutant was ligated back into the STIV full-length genome using the unique restriction sites Xmal and AvrII.

Preparation of Cloned Viral DNA for Transformation into S. solfataricus

Purified wild type (WT) pFLSTIV, gene disruption and point mutation clones were digested with BamHI (Promega) restriction endonuclease, followed by ligation with T4 DNA ligase to remove the pCRII-Topo-TA vector. This resulted in circular STIV genomes similar to the native genome with the exception of a single base change, which introduces a unique BamHI site into a non-coding region. To examine the effects of maintaining small amounts of heterologous DNA sequences on infectivity, one WT construct was also prepared, leaving a 106 bp fragment of the pCRII-Topo-TA vector in the STIV chromosome. This was constructed by digestion with SacI and Apal (Promega), and the ends were repaired with Klenow fragment and ligated.

Transfection of S. solfataricus with Viral DNA

Transformation of S. solfataricus strain 2-2-12 (Ortmann et al., 2008) was performed using a protocol adapted from Elferink (Elferink et al., 1996), Schleper (Schleper et al., 1992) and Albers (Albers et al., 2005). S. solfataricus cultures were grown overnight to an OD$_{650}$ ~ 0.2. Cells were cooled on ice and collected by centrifugation at 5000 x g for 5 minutes. The cell pellet was washed three times with cold 20 mM sucrose, each time reducing the total volume (50 mL, 25 mL, 1 mL). The final cell pellet was resuspended in sucrose at ~10$^{10}$ cells ml$^{-1}$ (500 mL of OD$_{650}$ ~ 0.2
culture reduced to 500 µl). Approximately 500 ng of DNA was used for each transformation. Electroporations were performed using the BioRad Gene Pulser Xcell at 1500 V, 25 µF, and 400 ohms for a time constant of 9 msec (1 mm gap cuvette).

Electroporated cells were mixed with 1 mL cold ddH2O and incubated on ice for 10 min, transferred to 80°C for 10 min and finally into 25 mL prewarmed medium 182 pH 2.5 (DSMZ), and the cultures were incubated at 80°C. Control transfections included native STIV DNA isolated directly from purified virus and mock transfections. In addition to S. solfataricus, transfections were performed in S. tokodaii (DSMZ), S. islandicus (R. Whitaker), S. acidocaldarius (ATCC), and Sulfolobus sp strain 2025 (Worthington et al., 2003) and were assayed for STIV production as described below. The full-length pFLSTIV clone, the clone with the cloning vector removed and the STIV sequence recircularizes, a linear version of the clone with the cloning vector removed and a clone containing a truncated (106 bps) cloning vector were all tested for ability to replicate.

Detection of Infectious Progeny Virus Following Transfection with Cloned Viral DNA

Transfected S. solfataricus cultures were monitored for up to 72 hours post transfection by absorbance at OD_{650} for cell growth and by quantitative PCR (qPCR) for production of viral genomes using PCR primers to the major capsid protein (Ortmann et al., 2008). Cultures that were positive by qPCR were screened by transmission electron microscopy (TEM, Leo 912AB) for presence of virus particles. Western blot analysis with polyclonal antibodies to the STIV major capsid protein (MCP ORF B345) was used to
confirm the presence of virus. Western blot analysis using polyclonal antibodies to STIV B116 and C381 were used for confirming the genetic disruptions of the B116 and C381 gene products.

The presence of infectious virus resulting from transfection with the cloned STIV DNA was assayed by both passage of virus to new cell cultures and by plaque assays. Transfected cell cultures were centrifuged at 5,000 x g for 5 minutes, the supernatant was filtered through 0.8/0.2 µm filters (Pall Acrodisc 25 mm PF Syringe Filter), and was concentrated ~10 fold by filtration through a 100 kDa MWCO filter (Amicon Millipore). The concentrated supernatant was used as a source of inoculum. Infections using cell-free concentrated transfection supernatants were performed by addition of ~ 3x10⁹ VLPs to early log (OD₆₅₀ ~0.2) S. solfataricus cells in medium 182 at pH 2.5 and incubation at 80°C. Cell cultures were subsequently analyzed by qPCR and TEM for presence of virus genomes and virus particles as described above. Plaque assays were performed as previously described (Brumfield et al., 2009). Briefly, dilutions of cell free supernatants of cultures transfected with cloned viral DNA were incubated with 500 µl early log S. solfataricus cells at OD₆₅₀ ~0.2 for 15 minutes at 80°C. Medium 182 with 0.2% w/v Phytagel (Sigma) at pH 2.5 and 80°C was added to the culture and immediately spread onto Medium 182 (1% w/v Phytagel) pH 2.5 plates and incubated at 80°C. Plaques were visually scored and tested for virus production by PCR for the MCP.
Results

Construction of the Full-length STIV Clone and Demonstration of Infectivity

A full-length STIV clone was successfully constructed by PCR amplification of three overlapping segments of the native STIV genome and insertion into the *E. coli* pCRII-Topo-TA plasmid (figure 2.2a). An *EcoRI* digest of the full-length clone produced the expected banding pattern (figure 2.2b), and DNA sequencing of selected regions of the cloned viral DNA confirmed that the cloned DNA was identical to the native STIV genome. The only exception was the introduction of a new *BamHI* restriction endonuclease site to facilitate the removal of the pCRII-Topo-TA vector sequence prior to transfection. The creation of this *BamHI* site did not result in any altered protein coding capacity within the viral genome.

![Figure 2.2.](image)

Figure 2.2. Full length clone of STIV. a. Genome map of full-length STIV infectious clone with STIV ORFs shown in black and the pCRII-Topo-TA cloning vector shown in gray. b. *EcoRI* digest of full-length STIV infectious clone.
Transfection of the full-length clone into *S. solfataricus* resulted in the production of both progeny viral particles (figure 2.3a) and genomes (figure 2.3b).

Purification of virus from the primary transformed cultures and passage of this virus into early log *S. solfataricus* cells demonstrated that the cloned virus was infectious and produced plaques similar to native virus. Transfection of DNA without prior removal of the *E. coli* vector was not successful. However, the presence of 106 bp of vector sequence in the non-coding region STIV genome still maintained infectivity. This

![Figure 2.3. Analysis of pFLSTIV transformation into *S. solfataricus* 2-2-12. a. TEMs of viral particles from primary transformed cultures. B. Averages and standard deviation of time course comparison of STIV transfection (clone) vs. a natural infection cycle (native virus particles) as measured by qPCR. C. Western blot with αB345 (major capsid protein) antibody: 1. purified virus, 2. infected culture, 3. transformed culture.](image-url)
indicates that the presence of the entire 3.9 kb of pCRII-Topo-TA vector sequence prevents viral replication. Likewise, transfection with linear BamHI digested STIV genome was non-infectious. Transformation of the circular cloned viral DNA resulted in an infection cycle that was delayed by 20 hrs as compared to native STIV infection, but it eventually reached the same magnitude as infection with native virus, as estimated by qPCR (figure 2.3b). Subsequent infection with virus particles obtained from a transfected culture produced a similar timing as native virus. In an attempt to assess the host range of STIV, various hosts were tested for infection with native virus and transfection with the cloned STIV DNA. The pattern of successful transfection closely followed the pattern of host competency for infection with native STIV (table 2.2).

A number of *Sulfolobus* species were tested for the ability to replicate STIV (purified virus or cloned viral DNA). Only *S. solfataricus* supported viral replication.

Table 2.2. Infection and Transfection of a variety of *Sulfolobus* spp. with STIV.

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Infection Results (Genomes/mL at MOI~1.0)</th>
<th>Transfection Results (Genomes/mL per 1µg DNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. solfataricus</em> P2</td>
<td>$1.3 \times 10^7 \pm 8.2 \times 10^6$</td>
<td>$2.1 \times 10^8 \pm 2.8 \times 10^8$</td>
</tr>
<tr>
<td><em>S. solfataricus</em> 2-2-12 (susceptible strain)</td>
<td>$5.7 \times 10^{11} \pm 1.45 \times 10^{11}$</td>
<td>$1.5 \times 10^{11} \pm 8.5 \times 10^{10}$</td>
</tr>
<tr>
<td><em>S. solfataricus</em> P2 #5 (resistant strain)</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td><em>S. solfataricus</em> P1</td>
<td>+, low level of infection (Rice et al., 2003)</td>
<td>Unknown</td>
</tr>
<tr>
<td><em>S. tokadaii</em></td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td><em>S. islandicus</em> (strains Y.G.57.14 and Y.N.15.51)</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td><em>Sulfolobus</em> strain PBL2025</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td><em>S. acidocaldarius</em></td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>
STIV Infectious Clone Allows for Targeted Gene Alterations

Selected STIV gene disruptions and point mutations were performed to demonstrate the ability of the STIV full-length clone to serve as the basis for a viral genetic system. STIV ORFs A197, B116, B345, and C381 were chosen for gene modification experiments. These ORFs were chosen based on some prior knowledge of gene function from structural or proteomics studies. Disruption of A197, B345, and C381 were lethal to virus replication, with no new viral genomes detected by qPCR (table 2.3). In contrast, transfection with cloned viral DNA harboring a B116 frame shift

Table 2.3. Mutant viruses constructed from the pFLSTIV clone. All constructs were transformed in S. solfataricus 2-2-12 host cells and screened for phenotype by qPCR, Western blot, and/or cell plaque assay.

<table>
<thead>
<tr>
<th>STIV ORF</th>
<th>Mutation</th>
<th>Method of construction</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>B116</td>
<td>Frame shift at position 85 and stop codon at position 88</td>
<td>Endonuclease digestion (PacI)</td>
<td>Delayed replication cycle, produces virions, small plaque phenotype</td>
</tr>
<tr>
<td>A197</td>
<td>Frame shift at position 179 and stop codon at 183.</td>
<td>Endonuclease digestion (SacII)</td>
<td>Lethal</td>
</tr>
<tr>
<td>A197</td>
<td>D151E</td>
<td>Site Directed Mutagenesis</td>
<td>Delayed replication cycle, produces particles</td>
</tr>
<tr>
<td>A197</td>
<td>D151N, D151H</td>
<td>Site Directed Mutagenesis</td>
<td>Lethal</td>
</tr>
<tr>
<td>A197</td>
<td>Deletion of 569 nts L158-&gt;Stop</td>
<td>Inverse PCR</td>
<td>Lethal</td>
</tr>
<tr>
<td>C381</td>
<td>Insertion of serine residue at position 223</td>
<td>Endonuclease digestion (Bsu361)</td>
<td>Lethal</td>
</tr>
<tr>
<td>B345</td>
<td>Deletion of c-term 22 a.a.</td>
<td>Inverse PCR</td>
<td>Lethal</td>
</tr>
</tbody>
</table>
mutation produced infectious virus. This cell culture was positive by qPCR for the MCP gene (figure 2.4a) and produced virus particles (figure 2.4b). The virus isolated from the supernatant of this culture was infectious when introduced into fresh *S. solfataricus* cells, and maintained the ΔB116 mutation (figure 2.4c). A comparison between the wild type and B116 frame-shift virus infections showed that the B116 frame-shift produced a delayed infection cycle of 24 hours as compared to wild type, but eventually reached similar levels (figure 2.4a).

![Graph](image1)

**Figure 2.4.** Analysis of ΔB116, A197 D151E transformations into *S. solfataricus* 2-2-12. a. Averages and standard deviations for ΔB116, A197 D151E, and WT STIV infections. b. TEM of particle from ΔB116 transformed culture. c. PCR spanning filled *PacI* site and digestion of with *PacI*: 1. PCR of WT transformed culture, 2. *PacI* digest of WT, 3. PCR of ΔB116 transformed culture, 4. *PacI* digest of ΔB116.
Western blot analysis with B116 antibodies did not detect any B116 protein in the B116 frame-shift transformed cells (figure 2.5a). Plaque assay analysis showed that the virus still lysed its host cells but manifested as a small plaque phenotype. The average plaque size produced by the B116 frame-shift mutant was reduced by ~90% as compared to wild type virus (figure 2.5 b, c and d). The single conserved amino acid substitution within A197 was also positive by qPCR (for the MCP gene (figure 2.5a) and

Figure 2.5. Analysis of the phenotype of ΔB116 STIV. A. Western blot of ΔB116 and WT infected cells: Lanes 1-3 purified B116 protein loading controls of 1 μg, 100 ng, 10 ng, respectively. 4. WT t=24, 5. WT t=48, 6. ΔB116 t=24, 7. ΔB116 t=48, 8. Uninfected S. solfataricus strain 2-2-12. B. ΔB116 small plaque phenotype (arrows). C. WT STIV plaques (arrows). D. PCR from plaques of ΔB116 virus: 1,3. plaque, 2,4. lawn.
produced virus particles (figure 2.6). The remaining mutant viruses were never detected by qPCR in transformed cultures indicating that the mutations were lethal and the targeted ORF is essential to STIV replication.

Figure 2.6. TEMs of particles from A197 D151E STIV transformed cultures.

**Discussion**

Construction of an infectious clone of STIV provides a valuable resource for the genetic analysis of this virus and for probing its interaction with its *Sulfolobus* host. The observation that the cloned viral DNA is infectious demonstrates that the viral or host proteins associated with the virion are not strictly required for infection. The demonstration that the cloned viral DNA can be used as a template for targeted gene disruption and point mutations of STIV ORFs should greatly facilitate detailed genetic and biochemical analysis of this virus. The full-length STIV cloned genome also provides the foundation for exploring the use of STIV as a vector for gene delivery and as a heterologous protein expression system for *Sulfolobus*. This is only the second crenarchaeal virus for which an infectious clone has been produced.
STIV was originally isolated from an uncharacterized *Sulfolobus* sp. from Yellowstone National Park. It was subsequently demonstrated to infect the Italian isolate *S. solfataricus* P2 (Rice et al., 2004), as well as *S. solfataricus* strain 2-2-12 (a highly susceptible strain derived from *S. solfataricus* P2) (Ortmann et al., 2008) and *S. solfataricus* P1 (a second Italian isolate; unpublished). STIV has not successfully replicated in any other *Sulfolobus* species or strains tested to date, including several Yellowstone isolates (table 2.2). Our attempts to use transfection with the cloned STIV DNA failed to produce viral infections except in *S. solfataricus* isolates. The inability to overcome the host range of STIV by transfection of cloned DNA is in contrast to examples from other archaeal viruses. His1 and His2 are viruses that infect Haloarchaeal hosts (euryarchaeotes). His2 only infects one host, *Haloarcula hispanica*, while His1 infects *H. hispanica* and one uncharacterized *Halorubrum* species. However, transfection was successful in six of 12 other hosts tested spanning four genera (Porter and Dyall-Smith, 2008; Porter et al., 2005; Porter et al., 2008). Of particular interest is the inability to transfect or infect isolates from Yellowstone National Park, which presumably would be more closely related to the isolate from which STIV was obtained than the lab strains *S. solfataricus* P2 and P1 originally isolated from Italy. The inability to transfect any of these isolates suggests that there is an inherent biochemical inability to support viral replication rather than simply a lack of a cell-surface receptor, corroborating previous microarray results in which no obvious membrane proteins
were differentially regulated when comparing *S. solfataricus* strain 2-2-12 to resistant strains of *S. solfataricus* P2 (Ortmann et al., 2008).

The inability to observe infection when the transfected DNA either contains large amounts of foreign sequences (e.g. the plasmid vector) or is a linear molecule is not surprising. For many viruses, there is only a limited ability to accommodate extra DNA within the virion. The requirement of removing the pCRII-Topo-TA vector appears to be a size limitation rather than the inability to tolerate an insertion into the particular site that was chosen. A construct was made in which a small fragment (106 bp) of the pCRII-Topo-TA vector was left in the STIV genome. This construct successfully replicated in *S. solfataricus*, producing viable virus particles. The requirement for a circular genome indicated that viral DNA replication either relies upon a circular DNA replication mechanism or a circular template for transcription.

The possible phenotypic outcomes of targeted viral gene modifications are either no change from the wild type phenotype, lethality or an altered phenotype. Our B345, B116 and A197 gene disruption and point mutations experiments were guided by structural information. A197 is a putative glycosyl-transferase, B345 is the major capsid protein, and B116 is a putative DNA binding protein. Creating frame shift mutations, which results in introducing downstream premature stop codons within the A197, C381, and introducing a serine residue into B345 were non-infectious, indicating that these genes are essential. The point at which the defect interferes with viral life cycle is as yet undefined. It is not surprising that eliminating glycosyl transferase
function would eliminate infectivity. The MCP of STIV is glycosylated (Maaty et al., 2006), and it is suspected that the A197 gene product may be responsible. Eliminating MCP glycosylation may prevent assembly and/or release of virus from cells. It is more surprising that a single serine residue addition to the MCP eliminates infectivity. Based on the crystal structure and cryo-EM image reconstruction of STIV, the added serine residues lies within a short surface exposed loop region of the MCP and would seem unlikely to significantly disrupt MCP function. The fact that this mutation is lethal suggests that this region may be interacting with a critical ligand required for infection.

The crystal structure of the B116 protein indicates that it is a highly conserved DNA binding protein that is found in a wide diversity of crenarchaeal viruses (Larson et al., 2007b). The ability to knock it out and maintain a viable virus demonstrates that it is a non-essential gene. The altered phenotype of small plaques and delayed infection cycle indicate that it is a crippled virus as compared to wild type. While the structure and sequence comparison to proteins of other crenarchaeal viruses suggests that B116 is a DNA binding protein (Larson et al., 2007b), its DNA target has not yet been demonstrated. Gel-shift experiments with fragmented STIV genomic DNA have not yielded any binding sites. No change in viral transcription patterns, other than timing, was observed for this mutant. Together this suggests that B116 is not a regulator of STIV transcription. It may be involved with modulating host transcription, but its function remains unknown.
The construction of a full-length infectious clone for STIV provides a valuable genetic tool for investigation of the STIV replication cycle and its interactions with its host. With both viral and host genetic tools now in hand, it should be possible to gain a greater understanding into the archaeal lifestyle and in particular adaptations required for life at high temperatures.
CHAPTER 3

VIRAL METAGENOMICS AT ROAD SIDE SPRING

Introduction

The detection and isolation of viruses directly from high temperature (>80°C) acidic (pH<4) hot springs, fumaroles and soils has long been challenging. These extreme environments tend to have low biomass (typically <10^6 cells/ml) and low free viral abundance (<10^5/ml) as compared to eutrophic freshwater lakes (Bettarel et al., 2004) or near shore marine environments (Suttle, 2005). Establishing laboratory cultures from these environments poses challenges due to the extreme and poorly defined growth conditions that must be mimicked in the laboratory. Likewise, culture-independent approaches (i.e. PCR amplification) for looking at viruses in the natural environment are problematic because of our rudimentary knowledge of viral diversity in these environments and the lack of universally conserved signatures that can be used for virus detection and isolation. Metagenomic sequencing has recently become the standard for gaining a foothold into environments about which we know little such as Yellowstone National Park hot springs. When this project was first undertaken it was not known whether viral metagenomics in a low abundance environment would be feasible.

Traditionally, most viruses have been initially isolated from the environment by first culturing their hosts in the laboratory and subsequently isolating the viruses from
the established cultures. The obvious advantage of this approach is that once one has isolated the virus, one has also identified its host and has established a means by which to maintain virus production. The disadvantage of culture-dependent approaches is that they are likely to severely underestimate the total diversity of viruses present in that environment due to limitations in culturing host organisms. Culturing hosts as a method for identifying and characterizing viruses not only limits us to viruses of hosts we are currently able to culture, but also introduces culturing bias in the population structures of the viruses themselves (Snyder et al., 2004). Forgoing enrichment cultures in favor of direct detection of viruses and subsequent isolation of genetic material from environmental samples provides us with the potential to uncover a greater understanding of the diversity of viruses in these environments. It also provides the foundation for more detailed examination of virus-host interactions within these unusual natural communities.

The dominant viral hosts present in high temperature acidic environments are members of the Archaea (Valentine, 2007). Our understanding of Archaea in these environments is just beginning to be explored. Compared to Bacteria and Eukarya, only a relatively small number of viruses have been identified and characterized from the Archaea. Of the over 5000 known viruses, only 49 infect archaeal hosts. Little more than a decade ago there were only four characterized viruses of Archaea (Zillig et al., 1994). Today there are approximately 25 viruses that have been identified and isolated from high temperature terrestrial environments, and these represent only four host
genera: *Sulfolobus, Acidianus, Pyrobaculum* and *Thermoproteus* (Prangishvili and Garrett, 2005).

Despite the inherent difficulties of culturing archaeal hosts from high temperature acidic environments, it has provided the majority of viruses that have been isolated to date. A number of viruses have been isolated from *Sulfolobus* spp., partly because of their relative ease of culturing and their ubiquitous presence in high temperature acidic environments. Consequently, the majority of Archaeal viruses discovered to date infect *Sulfolobus* spp.

Culture-independent approaches to virus isolation from high temperature environments are based on taking advantage of the relative abundance and the unique size and density of virus particles found free of their hosts. Our general approach was to filter a relatively large volume of environmental sample (ranging from liters to hundreds of liters), simultaneously concentrating the cells and the viruses from the sample. In the second step, the virus was separated from the cells. In the final step the virus was further purified and concentrated.

We undertook a viral metagenomics sequencing project at Road Side Spring in the Norris Geyser Basin of Yellowstone National Park beginning in the summer of 2005, both to expand our understanding of viral communities present in hot acidic environments and as a proof of concept that metagenomic sequencing is a tractable method for looking at viruses in these environments. Our goals were two-fold. First was to show that enough genetic material could be obtained to perform metagenomic
sequencing in an environment with relatively low biomass. The second goal was to show that in such a low-complexity environment it is possible to assemble large viral contigs, in contrast to more complex environments like the ocean (Rodriguez-Brito et al., 2010).

**Materials and Methods**

**Site Selection**

Road Side Springs (RSS N44 45.214 W110 43.515) in Yellowstone National Park was chosen for this study because of the following criteria: it is hot (72°C), acidic (pH 4.2), large volume (~3 x 5 m) and relatively clear (little suspended sediment) (figure 3.1).

![Figure 3.1 Road Side Spring.](image)
The pool is also located close to the road, making it an easy location for setting up large volume filtration systems requiring pumps and a generator.

**Sample Preparation and Virus Purification**

Approximately 400 L of Road Side Spring water (from ~10 cm below the surface at the middle of the pool) was concentrated to 4 L by tangential flow filtration (TFF) through a 100,000 Dalton molecular weight cutoff (MWCO) filter on site at RSS. Figure 3.2 presents a schematic of the tangential flow process. Figure 3.3 shows a tangential

Figure 3.2 Schematic of sampling method. Sample is pumped to a glass wool filter to remove sediments and a 100 kDa molecular weight cutoff filter to concentrate the sample in a sampling bottle, while the filtrate is directed back to the pool.
flow filter. Briefly, the hot spring water was initially pumped (using a Masterflex I/P double headed long shaft peristaltic pump) through a glass wool filter (constructed from 2 ft of 2 in diameter PVC pipe with hose fittings on each end and densely packed with glass wool) to remove as much sediment as possible. It was then directed through a 100 kDa MWCO tangential flow filter (Amersham Biosciences UFP-100-C-9A) applying backpressure in order to force the water through the filter tangentially. Pressure was adjusted to achieve an equal flow rate between the retentate and the filtrate (approximately 1L/minute). The filtrate (water and particles smaller than 100 kDa) was directed back into RSS, while the retentate was directed into a 10 L sterile carboy. The

Figure 3.3 100 kDa molecular weight cutoff tangential flow filtration system.
retentate was further concentrated by using a second TFF with both the intake and retentate lines in the carboy and sufficient backpressure to direct approximately half the volume to filtrate (returned to RSS).

The final 4 L volume of retentate was returned to the laboratory at ambient temperature. This 4 L sample was then centrifuged (6000 x g for 20 min) to remove sediments and cells. The resulting supernatant was filtered through a double 0.45 µm filter (Pall FP-450 0.45 µm 47 mm filter #66480) to remove any remaining cells and sediments. The sample was further concentrated 10 fold (to approximately 200 mL), using the 10,000 Da MWCO tangential flow filter unit (Amersham Biosciences UFP-10-C4A) in a manner similar to that described for the in field 100,000 MWCO tangential flow unit.

The retentate was then loaded into four 26 mL ultracentrifuge tubes and spun at 100,000 x g for 2 hours (Beckman Ultracentrifuge model L8-M with T-30 rotor) to concentrate all virus particles. Most of the supernatant was then removed and all four pellets resuspended in a total of approximately 200 µL hot spring water (supernatant).

**Cell and Virus Enumeration by Epifluorescence Microscopy**

Direct cells and virus like particle (VLPs) counts of RSS samples were determined by Epifluorescence Microscopy (Ortmann and Suttle, 2009). For cell counts total hot spring water at various dilutions was applied to a 0.02 µm glass filter (Whatman Anodisc 25 filters) by vacuum filtration. For viral counts the sample was first filtered through 0.8/0.2 µm filter (Pall) and then applied to the 0.02 µm glass filter by vacuum filtration. The filter
was placed in ~30 µl 10X Sybr Gold (Invitrogen) stain for 15 minutes in the dark and then air dried. Filters were mounted on glass slides with 15 µl 50% glycerol and 50% 1X PBS and coverslipped with additional 20 µl glycerol/PBS. Counts were performed on a Nikon Eclipse E600 flourescent microscope. Total cells or viruses in at least 10 fields of view were counted and divided by the total number of fields counted. Cells/mL (VLPs/mL) was calculated as follows:

\[
\text{Cells per ml} = \frac{\text{Avg. # of cells counted} \times (\text{area of filter mm}^2/\text{area of field of view mm}^2) \times \text{volume filtered ml}^{-1}}{}.
\]

**TEM Analysis of Samples**

Concentrated samples were screened by transmission electron microscopy (TEM, Leo 912AB) for presence of virus particles.

**Nucleic Acid Extraction**

Nucleic acid was extracted from the total sample pellet and from the viral pellet for both PCR-based 16S rDNA and viral nucleic acid analysis. For total chromosomol DNA extraction, approximately 50 ml of RSS environmental water sample was centrifuged at 6000 x g for 10 minutes, the supernatant removed and the cell pellet resuspended in 250 µl TEN (10 mM Tris-HCL, 1 mM EDTA, 150 mM NaCl, pH 8.0). Then 250 µl TENST (TEN with 0.12% w/v Triton X-100, 1.6% w/v N-lauryl sarcosine) was added and the sample was incubated for 30 minutes at room temperature prior to extraction by phenol chloroform (see below).
In order to ensure that samples contacting virus particles were free of any exogenous cellular or free nucleic acid, the samples were nuclease treated prior to nucleic acid extraction from the virions. DNase RQ1 (5 units) and 10x DNase buffer (Promega) were added to the sample and incubated at 37°C for 30 minutes. The samples were subsequently treated with RNase A (1mg/mL) and incubated for 5 minutes. The DNase and RNase were then inactivated or eliminated by the Proteinase K during the nucleic acid extraction.

For viral nucleic acid extraction from virions, the sample was brought up to a total volume of 200 µl with 0.1% total concentration of SDS (2 µl of 10% SDS) 400 µg proteinase K (Ambion) and the remainder of the volume TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). The sample was incubated at 37°C ramping to 55°C over 30 minutes and then phenol extracted. The phenol-chloroform extraction was performed similarly for both cellular and viral samples. A volume of hot phenol (55°C) equal to the sample volume was added, vortexed well and centrifuged at 14000 rpm for five minutes. The aqueous layer was removed, and the process was repeated with phenol, followed by a phenol:CH₃Cl:IAA (25:24:1) extraction. The DNA was precipitated with ethanol (2x volume), sodium acetate (0.10 x volume, 3M at pH 5.2) and collected by centrifugation. The DNA pellet was washed with 70% ethanol and was resuspended in a final volume of 50 µl TE (pH 8.0).
16S rDNA Sequencing

16S rDNA PCR was performed on both the cellular fraction and on the viral fraction. Primers specific for both Archaea and Bacteria were used (348Fa and U1391R or 61 bacF and U1391R) to amplify the cellular 16S rDNA. For the cellular 16S rDNA, the PCR products were cloned, and one 96-well plate was sequenced at Idaho National Laboratories. For the viral 16S rDNA amplification, universal 16S primers were used (U515F and U1381R).

Differential Nucleic Acid Digestion

Total virus nucleic acid was digested with DNase and RNase in separate reactions to determine the nature of the nucleic acid. 10 μL of sample was digested with either DNase RQ1 (1 unit) and 10x DNase buffer (Promega) in at 20 μL reaction at 37°C for 30 minutes or with RNase A (1mg/mL) and incubated for 5 minutes. Digests were visualized on an ethidium bromide gel.

Viral Metagenomic Sequencing and Assembly

The viral nucleic acid was provided to Lucigen Corp. (Madison, WI) for amplification and library construction (nano-library). E. coli harboring the clones was grown overnight and mini prepped (Millipore Montage Plasmid Mini Prep 96 kit). Sanger sequencing was performed in 96 well format, using T7 promoter and terminator primers and the BigDye Terminator v3.1 sequencing kit (Applied Biosystems). Sequencing plates were sent to Idaho National Laboratory (Idaho Falls, ID) and were
determined on an Applied Biosystems 3730 DNA Analyzer. A total of 1152 reads were obtained.

Sequences were cleaned up (trimmed of vector and corrected for ambiguous base calls) and assembled using Sequencher 4.5. Assemblies were performed at 90% identity with 20 bp overlap.

**Confirmation of Genome Assembly by PCR Amplification, Cloning, and Sequencing**

PCR amplification using a combination of primer sets (Table 3.1) and the iProof (BioRad) high fidelity polymerase was performed to amplify long fragments of the viral contig (table 3.1) as well as to attempt to amplify off the ends of the 18167 bp contig to determine if the viral genome was circular. Cycling parameters were as follows: initial denaturation of 98°C for 30 s; 40 cycles of 98°C for 10 s; 48-61°C (depending on primer set) for 10 s; 72°C for 1-5 m (depending on primer set) followed by a final extension of 72°C for 10 m.

PCR using the Taq (Promega) polymerase was used to amplify seven short regions of the viral contig for confirmation of the assembly and to span regions of low coverage. Cycling parameters were as follows: initial denaturation of 95°C for 5 m; 40 cycles of 95°C for 30 s; 48-55°C (depending on primer set) for 30 s; 72°C for 1 m followed by a final extension of 72°C for 10 m.
PCR products of the approximate correct length were cloned into the pCRII-Topo-TA vector (Invitrogen) per manufacture’s protocols. Clones were sequenced using the T7 promoter and M13 reverse primers at Idaho National Laboratory.

**Colony Hybridization**

In order to screen the RSS library in *E. coli* for clones containing sequence near either end of the RSSV contig, Dig-labeled probes were designed near the ends of the RSSV contig (RSSV_5’F and RSSV_5’R or RSSV_3’F and RSSV_3’R primer sets). The Dig-labeled probe was constructed using the DIG High Prime DNA labeling kit (Roche). *E. coli* harboring clones from the RSS library were plated on LBKan plates and grown overnight at 37°C. Colonies were blotted onto Nylon membrane (Millipore Immobilon NY+) and air-dried at room temperature for 30 minutes. The membrane was UV crosslinked, washed, hybridized with the Dig-labeled probe, and detected using the Dig High Prime DNA labeling kit. Fluorescent spots were matched to candidate colonies which were then patched onto a new LBKan plate, grown overnight, picked into liquid media, grown overnight, plasmid prepped and sequenced (as above).

**Mapping Nymph Lake Viral Metagenomic Reads to RSSV Contig**

The largest RSS contig (18167 bps) was subjected to BLAST analysis (BLASTn) against all metagenomic data sets from Nymph Lake and Crater Hills (described in chapter 4). All reads that matched with an e-value below 10e-3 were then mapped to the RSSV contig using GS Mapper (Newbler GS Assembler v2.3) at 90% identity.
Open Reading Frame Prediction and Annotation

The 18167 bp RSS contig was submitted to the AlterORF pipeline (Pedroso et al., 2008) for ORF-finding and annotation. ORFs were identified by looking for gaps of at least 45 amino acids between stop codons in all reading frames. These ORFs were translated and then blasted against the following databases: CDD, Conserved Domain Database (RPS-BLAST), and all its sub-databases (CD, Conserved Domains; COG, Cluster of Orthologous Groups; KOG; Cluster of Orthologous Groups (Eukaryotic sequences); PFAM, Protein Families; PRK, Protein Clusters; SMART, Simple Modular Architecture Research Tool; TIGRfam, TIGR protein Families); PDB, Protein Data Bank (BLASTp); Uniprot/Swissprot, Universal Protein Resource, Swiss-Prot section (manually curated proteins)(BLASTp); UniRef90 (BLASTp); Uniprot clustered proteins (90% identity)(BLASTp); NR: Non-Redundant protein database by NCBI (BLASTp).

The “best” single match for each ORF was then chosen by the following criteria: Matches with less than 20% of hit coverage were discarded, “informative” descriptions were given priority over “uninformative” (i.e. hypothetical) descriptions, e-values were used to determine precedence between two informative or two uninformative matches. The NR, PDB and KOG databases were given lower priority compared to the other databases; thus matches from these sources were only chosen if the other databases did not return significant results.
Coat Protein Amplification, Cloning and Over Expression

The coat protein was amplified from the 9kb long viral clone using Taq polymerase and RSS_cpf and either RSS_cpr or RSS_cpr_his. Cycling parameters were as follows: initial denaturation of 95°C for five minutes, 35 cycles of 95°C for 30 s, 46°C for 30 s, 72°C for 1 m followed by a final extension of 72°C for 10 m. It was amplified either from the start codon through the stop codon for expression of a WT protein or through the last codon before the stop for the addition of a His tag. In both cases an Ndel site was added to the 5’ end and a Xhol site was added to the 3’ end for the WT protein and a BamHI site for the His tagged protein. Amplicons were digested with Xhol and Ndel (New England Biolabs) or BamHI and ligated into the pET30a vector (also digested with the appropriate enzymes). Clones were transformed into XL2 Blue cells (Stratagene), grown overnight in LB Kan and sequenced using the T7 promoter and terminator primers.

Correct clones without a His tag were transformed into BL21 cells and grown overnight in 5 mL LB Kan. Cultures were diluted into 100 mL at ~12 hours and grown to an OD ~0.4 (approximately 3 hrs) and induced with 1mM IPTG. Samples (1 mL) were taken at t=0, 1, 2, 3 and 4 hrs post induction. Cells were pelleted, frozen overnight at -20°C, thawed and resuspended in 50 μL SDS buffer, boiled for 5 minutes and analyzed on an 12.5 % Homogenous SDS PAGE gel (GE Healthcare).

Clones with a His tag were transformed and induced as described above. At t=4 hours cells were pelleted, resuspended in lysis buffer (50mM NaH2PO4, 300mM NaCl,
Table 3.1  RSS Primers

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence</th>
<th>Amplicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSS 7k 5’F</td>
<td>GCCTGAACCTGTATACCAGCCG</td>
<td>Long PCR of RSSV</td>
</tr>
<tr>
<td>RSS 7k 3’R</td>
<td>CGACCCATTCTCTCCTACATGGC</td>
<td>Long PCR of RSSV</td>
</tr>
<tr>
<td>RSS 7k 5’R</td>
<td>CGGCTGTATACAAAGTTCAGGC</td>
<td>Inverted PCR</td>
</tr>
<tr>
<td>RSS 7k 3’F</td>
<td>GCCATGTAGGAATAAGGTTCG</td>
<td>Inverted PCR</td>
</tr>
<tr>
<td>RSS 9k 5’F</td>
<td>GCCGTTATTGTAGTAGAATACCGCC</td>
<td>Long PCR of RSSV</td>
</tr>
<tr>
<td>RSS 9k 3’R</td>
<td>CGGGGCCACCTACGGCCATTTACG</td>
<td>Long PCR of RSSV</td>
</tr>
<tr>
<td>RSS 9k 5’R</td>
<td>GCCATAGGAATACGTCGAGGC</td>
<td>Inverted PCR</td>
</tr>
<tr>
<td>RSSV 17K 5’R</td>
<td>CGGCGGCTATATCATTGCGGTAG</td>
<td>Near full length RSSV</td>
</tr>
<tr>
<td>RSSV 17K 5’R #2</td>
<td>GTCAGGACATCAGAGTACGTTACG</td>
<td>Near full length RSSV</td>
</tr>
<tr>
<td>RSSV 17K 3’F</td>
<td>GGACACTTGAGTGTTGCTCAGG</td>
<td>Inverted PCR</td>
</tr>
<tr>
<td>RSSV 17K 3’F #2</td>
<td>CCGTGGTTTTTACGGGACAC</td>
<td>Inverted PCR</td>
</tr>
<tr>
<td>RSS_cpf (Ndel)</td>
<td>CATATGATAATTCCACTGGCATTGGGG</td>
<td>Coat protein</td>
</tr>
<tr>
<td>RSS_cpr (XhoI)</td>
<td>CTCGGAGCCACGAACAGCAAG</td>
<td>Coat protein</td>
</tr>
<tr>
<td>RSS_cpr_his (BamHI)</td>
<td>CGATGGAGTCATCCGACAGACAGC</td>
<td>Coat protein</td>
</tr>
<tr>
<td>16S U515F</td>
<td>GTGCCAGCMGCCGCGGTAA</td>
<td>Universal 16S</td>
</tr>
<tr>
<td>16S U1391 R</td>
<td>GACGGGCGGCTGTRCA</td>
<td>Universal 16S</td>
</tr>
<tr>
<td>16S 348 Fa</td>
<td>TCCAGGCCCTACCCGG</td>
<td>Archaeal 16S</td>
</tr>
<tr>
<td>16S 61 bacF</td>
<td>CTGTTAACACATGCAAG</td>
<td>Bacterial 16S</td>
</tr>
<tr>
<td>T7 promoter</td>
<td>TAATACGACTCACTATAGG</td>
<td>Sequencing primer</td>
</tr>
<tr>
<td>T7 terminator</td>
<td>GCTAGTTATCTGCAGCGG</td>
<td>Sequencing primer</td>
</tr>
<tr>
<td>M13R</td>
<td>CAGGAACACAGCTATGACC</td>
<td>Sequencing primer</td>
</tr>
<tr>
<td>RSSV_12F</td>
<td>CGTCGACGACATCAGCAGTACG</td>
<td>119 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_1311R</td>
<td>GGATCTGATCAGGCTGGTGTG</td>
<td>119 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_1454F</td>
<td>GACATGGCAGAATCTCTGTGAATGC</td>
<td>152 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_1606R</td>
<td>CCTTGCCCTATGGCCTGCAATCC</td>
<td>152 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_9953F</td>
<td>CCAAGGGCTTGCCAGGTTT</td>
<td>214 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_10167R</td>
<td>CCGTGAGGCTCATAATCAACAGACG</td>
<td>214 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_11189F</td>
<td>CCGCCTGACATCTTGATACCAG</td>
<td>292 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_11481R</td>
<td>CTCATGGATGCTTCTTTCGTC</td>
<td>292 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_12076F</td>
<td>CCCACGATATCGGTGGAATTGC</td>
<td>189 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_12265R</td>
<td>CCTGCAAGAATAGCCATATAAACGCCC</td>
<td>189 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_13525F</td>
<td>GGCAACACCCACCTCTTACG</td>
<td>221 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_17537R</td>
<td>GGCAAGAATATGGGCTCAGC</td>
<td>221 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_17770F</td>
<td>CCTGACCAGCGGCCATGTC</td>
<td>223 bp RSSV contig confirmation</td>
</tr>
<tr>
<td>RSSV_18993R</td>
<td>GGACACGGATATAATCGGCGG</td>
<td>223 bp RSSV contig confirmation</td>
</tr>
</tbody>
</table>
10 mM Imidazole, 1 mg/mL lysozyme), and sonicated 6x for 10 s on ice. The protein was purified using the Qiagen Ni-NTA resin according to the manufacturer’s protocol. The supernatant, cell pellet, all washes and elutions were analyzed by western blot using an anti-6x His antibody (Qiagen).

Results

Transmission Electron Microscopy

Transmission electron microscopy (TEM) analysis of the concentrated viral sample from RSS revealed a large number of viral like particles of different morphologies

Figure 3.4 Transmission electron micrographs of concentrated sample from Road Side Springs.
A small (~30-40 nm) pleomorphic spherical VLP was the predominant morphology observed (fig. 3.4d, black arrow). A phenol chloroform extraction produced stainable amounts of viral nucleic acid as imaged on an ethidium bromide stained gel.

Differential nuclease treatment indicated that the visible nucleic acid was primarily DNA (figure 3.5 lanes 1 and 2). An EcoRI digest produced a visible banding pattern (figure 3.5 lane 3). 16S rDNA PCR indicated that there was no detectable DNA of cellular origin in the viral fraction (figure 3.5 lanes 4, 5, and 6).

Figure 3.5 Analysis of extracted nucleic acid from RSS. Lanes: 1. DNase treated, 2. RNase treated, 3. EcoRI digest, 4. 16SrDNA PCR16S rDNA amplification of extracted viral nucleic acid, 5. 16S rDNA amplification of nucleic acid extracted from cell pellet from RSS, 6. 16S rDNA no template control.

16S rDNA Sequencing

One plate (96 clones) of 16S rDNA PCR products from the cellular pellet was sequenced: half a plate (48 clones) using archaeal specific primers and half a plate (48
clones) using bacterial primers. The top archaeal matches were: *Sulfolobus solfataricus* 16S rRNA gene complete sequence (95% identity, 5 clones), *Caldisphaera lagunensis* gene for 16S rRNA (92% identity, 4 clones) and uncultured *Staphylothermus* sp. clone SK213 16S ribosomal RNA gene (92% identity, 2 clones). The top bacterial matches were: uncultured *Aquificales* 16S rRNA gene, partial sequence (99% identity, 15 clones), uncultured *Hydrogenobaculum* sp. clone SK262 16S rRNA, partial sequence (99% identity, 7 clones), uncultured *Hydrogenothermus* sp. clone OPPB005 16S ribosomal RNA gene, partial sequence (98% identity, 2 clones).

**Viral Metagenomic Sequence and RSSV Assembly**

1152 reads were sequenced (average read length of 640 bps) with approximately 20% of reads either failing or of poor quality and were not included in the assembly. A total of 26 contigs (assemblies of sequencing reads into contiguous sequence) larger than 1000 bp and containing more than two reads were assembled. The largest contig assembled consisted of 151 clones and was 17491 bps and was termed putative RSS viral contig (RSSV) (figure 3.6). This contig was subjected to extensive analysis to determine if it is a complete, or near complete, viral genome.

The contig has a GC content of 57%. The read coverage was relatively even across the entire length of the contig, with average coverage of 5X. Areas of low coverage (<3x, blue checks) were confirmed by follow up PCR amplification, cloning and sequencing (see below). Four additional plates of sequence and colony hybridization experiments directed towards detecting sequence at the ends of the existing contig (see
below) extended the contig to 18167 kb. Because this sequence was detected by PCR in the nearby Nymph Lake hot spring #10, sequencing reads from eight viral metagenomes from that site were mapped to the RSSV contig. The Nymph Lake reads extended the contig to 19324 bps.

**Confirmation of Assembled RSSV Contig by PCR Amplification, Cloning and Sequencing**

Primers were designed to various regions of the RSSV contig both to confirm the contig assembly and to produce large clones of the putative viral genome. It was possible to amplify and clone large regions of the RSSV contig, one of approximately 7kb and the other of approximately 9kb (figure 3.7a). Both of these clones produced the expected banding pattern in an *EcoRI* digest. Seven sets of primers were designed to

![Figure 3.6 Sequence coverage map of the putative RSS viral contig. Areas of low coverage shown in blue checks or white, stop codons shown as red slash marks. Green arrows are forward sequences, red arrows are reverse sequences.](image-url)
span low-coverage regions of the viral contig and to verify the contig assembly (table 3.1). Six of these primer sets produced the expected size amplicon (figure 3.7b). The primer set RSSV_12F and RSSV_1311 was designed to amplify beginning 12 base pairs from the 5’ end of the contig. This primer set failed to produce a product.

Primers were also designed to amplify off the ends of the 18167 bp contig in order to determine if the viral genome is circular. Inverted PCR (primers directed outward from the ends of the contig) would only produce an amplicon if the genome is circular. Various extension times (1-5 m) and annealing temperatures (53-61°C) were tested. Multiple bands were always amplified. Cloning and sequencing of these products resulted in sequences that did not align to the existing 18167 bp contig, indicating that the RSSV is not likely circular.

Figure 3.7  PCR of various regions of the RSSV contig.  a.  1. Negative control, 2. 7kb RSSV clone, 3. 9kb RSSV clone; b. PCR using seven different primer sets spanning the RSSV contig: 1. RSSV_12F and RSSV_1311R, 2. RSSV_9953F and RSSV_10167R, 3. RSSV_1454F and RSSV_1606R, 4. RSSV_11189F and RSSV_11483R, 5. RSSV_12076F and RSSV_12265R, 6. RSSV_17352F and RSSV_17537R, 7. RSSV_18770F and RSSV_18993R, 8. RSSV_18770F and RSSV_18993R RSS viral sample 4/18/05, 9. RSSV_18770F and RSSV_18993R No template control. All templates were RSS cellular sample from 08/02/07, unless otherwise indicated.
Colony Hybridization to Define the RSSV Ends

We screened plates of *E. coli* colonies harboring the RSS library by colony hybridization with probes designed to the ends of the RSSV contig. A total of eight plates (approximately 400 colonies) were analyzed, four plates hybridized with a probe to the 5’ end of the contig and four plates with a probe to the 3’ end. From these plates, only 11 colonies from the 5’ probe plates and 8 colonies from the 3’ probe plates were positive (figure 3.8). All 5’ sequences added internally to the existing RSSV contig, while the 3’ sequences extended the contig by nearly 1000 bps.

Figure 3.8 Colony hybridization membrane with colonies positive for RSS sequence.

Mapping Nymph Lake Viral Metagenomic Reads to RSSV Contig

A total of 9812 metagenomic reads from all Nymph Lake viral metagenomes (described in chapter 4) mapped to the RSSV contig at 90%. Approximately 600 bps
were added to 5’ end of the contig and 550 bps to the 3’ end. The entire contig has an average of 40X coverage, but the extended ends drop below 10X coverage (figure 3.9).

Figure 3.9 Nymph Lake reads mapped to RSSV contig. Green indicates coverage >40X, red is coverage <10X.

**RSSV Contig Annotation**

A total of 50 ORFs larger than 45 amino acids were identified in all six reading frames (figure 3.10). Nucleotide BLAST (BLASTn) analysis returned no matches to anything in the non-redundant database. Likewise there were no matches to the COG, KEGG or Conserved Domain Databases. Protein based BLAST of translated sequences (BLASTp) produced annotations for fifteen ORFs (table 3.2). All of these matched to Pyrobaculum Spherical Virus (PSV) (Haring, 2004) and/or *Thermoproteus tenax* spherical virus 1 (TTSV1) (Ahn, 2006), most annotated only as hypothetical proteins. One ORF matched the PSV Coat Protein (VP2) and TTSV1 gp-20, as well as to a cellular regulatory protein from *Burkholderia dolosa* (table 3.2).

There was some synteny evident in comparing the RSS genome to PSV and TTSV1 (fig. 3.10). Specifically there were three sets of genes of three to four genes that were syntenous between RSSV and TTSV1 (Pink, green and blue boxes in fig 3.10). TTSV1 ORFs gp4, gp5, and gp6 (identities between 50-67% and e-values from 6.0e-5 to 2.0e-69), gp13, gp14, gp15 and gp16 (identities between 40-51% and e-values from 4.0e-10 to
6.0e-25) and gp20, gp 21 and gp22 (identities between 30-45% and e-values from 9.0e-14 to 1.0e-44) exhibited synteny with sets of ORFs within RSSV. The rest of the genome did not exhibit synteny. There were only two genes that were syntenous between RSSV and PSV (green boxes in fig. 3.10): PSV ORFs gp25 (identity 41%, e-value 2.0e-24) and gp26 (identity 60%, e-value 2.0e-18).

Figure 3.10  Open reading frame map of RSSV contig compared to TTSV1 and PSV. Black lines are stop codons. All gaps larger than 45 codons are considered putative ORFs. Boxes are ORFs with some annotation assigned. Syntenous genes are indicated with common colors and lines between genomes. Putative coat protein is circled.
Table 3.2  Annotation of RSS viral contig: BLASTp against the NR database

<table>
<thead>
<tr>
<th>Coord.</th>
<th>Query Value</th>
<th>Identity</th>
<th>Similarity</th>
<th>Query Coverage</th>
<th>Hit Coverage</th>
<th>Hit Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-1618</td>
<td>2.00E-06</td>
<td>29.86</td>
<td>52.08</td>
<td>25.97</td>
<td>24.44</td>
<td>hypothetical protein PyrSV_gp02</td>
</tr>
<tr>
<td>2317-3447</td>
<td>1.00E-04</td>
<td>41.61</td>
<td>59.44</td>
<td>73.74</td>
<td>89.14</td>
<td>hypothetical protein PyrSV_gp03</td>
</tr>
<tr>
<td>3573-4097</td>
<td>5.00E-04</td>
<td>40.74</td>
<td>62.96</td>
<td>30.86</td>
<td>62.07</td>
<td>hypothetical protein PyrSV_gp08</td>
</tr>
<tr>
<td>4087-3325</td>
<td>5.00E-14</td>
<td>28.02</td>
<td>49.57</td>
<td>90.87</td>
<td>98.24</td>
<td>hypothetical protein PyrSV_gp15</td>
</tr>
<tr>
<td>comp.6291-5003</td>
<td>1.00E-17</td>
<td>31.23</td>
<td>49.80</td>
<td>94.68</td>
<td>97.62</td>
<td>hypothetical protein PyrSV_gp16</td>
</tr>
<tr>
<td>12098-15156</td>
<td>7.00E-22</td>
<td>31.10</td>
<td>51.51</td>
<td>80.74</td>
<td>98.98</td>
<td>hypothetical protein PyrSV_gp17</td>
</tr>
<tr>
<td>5766-6671</td>
<td>3.00E-13</td>
<td>28.91</td>
<td>48.98</td>
<td>86.76</td>
<td>89.87</td>
<td>hypothetical protein PyrSV_gp21</td>
</tr>
<tr>
<td>7573-8058</td>
<td>2.00E-24</td>
<td>41.21</td>
<td>57.58</td>
<td>99.38</td>
<td>100.00</td>
<td>hypothetical protein PyrSV_gp25</td>
</tr>
<tr>
<td>8062-8358</td>
<td>2.00E-18</td>
<td>60.81</td>
<td>77.03</td>
<td>70.71</td>
<td>77.08</td>
<td>hypothetical protein PyrSV_gp26</td>
</tr>
<tr>
<td>11490-12251</td>
<td>8.00E-32</td>
<td>44.97</td>
<td>67.20</td>
<td>72.05</td>
<td>74.80</td>
<td>virus coat protein VP2 [Pyrobaculum spherical virus]</td>
</tr>
</tbody>
</table>

Coat Protein Over Expression

Over expression of the viral coat protein was not detected either by direct visualization on a protein gel of either WT, nickel column affinity purification for His tagged proteins or by western blot using anti-His antibodies to the heterologous His tag.
Detection of RSS Viral Sequence in Other Yellowstone Hot Springs

Primers designed to the RSS coat protein (table 3.1) were used to look for the presence of the viral sequence in five sites in the Crater Hills region, and one other site in the Nymph Lake area of Yellowstone National Park. The sequence was detected in three sites, CHANN041 (84°C, pH 1.5), CH#3 (88°C, pH 2.0) and CH#4 (75°C, pH 1.5), in the Crater Hills region. It was also detected in Nymph Lake hot spring #10 (86 – 92°C, pH3) and was detected in all of our viral metagenomes from this site (chapter 4).

The viral sequence was detected in the Road Side Spring site at numerous sampling time points including 3-18-05, 4-16-07, 8-2-07, 5-29-09 and 5-11-11.

Discussion

High temperature acidic hot springs in Yellowstone National Park tend to be dominated by members of the domain Archaea and are generally sites with low abundance of free viral particles. Road Side Spring is approximately 10°C cooler and one order of magnitude less acidic than sites that have been shown to be dominated by Archaeal organisms. Detection of bacterial species in this site is in accordance with these environmental parameters.

Concentration of environmental samples approximately $10^6$ fold allowed for TEM based visualization of 5-7 virus morphologies. The abundance of free virus particles in the concentrated sample allowed us to extract detectable quantities of nucleic acid. Despite high numbers of VLPs detected by epiflourescence and a diversity of
morphologies detected by TEM, the banding pattern seen in the EcoRI digest of total viral nucleic acid, indicates that the viral fraction was dominated by a single genotype. This was corroborated by the ability to assemble such a large contig (19324 bps) from metagenomic data.

It is likely that we have sequenced and assembled a near full length novel virus. The 19324 bp RSS contig is within the range of a full-length genome for an archaeal virus. Subsequent sequencing efforts, even those directed towards the ends of the contig, primarily added sequence to the interior of the contig, rather than extending the ends, suggesting that we have the majority of this genome. The ability to PCR amplify both long fragments and six short fragments, all of the expected size, gives us confidence that the assembly is accurate.

The inability to over express the coat protein may be a hurdle that could be overcome with different expression systems. Expression using the pET vector in the Rosetta cell line (which carries a plasmid to express rare tRNAs) as well as using the Gateway (Invitrogen) system was also attempted. This gateway clone could be used in yeast or other expression systems in the future. Researchers studying PSV have also been unable to over express the PSV coat protein (personal communication, M. White).

The RSS viral sequence was detected across multiple time points spanning several years, and at numerous sites throughout Yellowstone with vastly different physical properties. This suggests that this virus is ubiquitous within high temperature acidic hot springs within Yellowstone National Park.
Challenges in annotating viral genomes from the domain Archaea, are not uncommon, as there are few viral sequences from the domain Archaea in the public databases, and those that are available tend to match best to their close relatives. However, the annotation of the coat protein gene on this contig by identity with the coat protein from PSV and a structural protein from TTSV1 gives us high confidence that this is indeed a viral genome. The high number of matching ORFs between these three viruses and synteny evident for several sets of ORFs further confirms that this is a viral genome.

PSV was isolated from an enrichment culture from Obsidian Pool in Yellowstone National Park (85°C and pH 6) while TTSV1 was isolated from an enrichment culture from a hot spring in Indonesia (grown at 88°C and pH 6). Both viruses are spherical shaped viruses. TTSV is approximately 70 nm in diameter with a 21 Kb linear dsDNA genome (Ahn, 2006), while PSV is approximately 100 nm in diameter with a 28 Kb linear dsDNA genome (Haring, 2004). While many archaeal viruses have few or poor quality matches to sequences in the public databases, both TTSV1 and PSV are unusual in that neither of them matches to anything other than each other. The RSSV contig has 15 putative ORFs with matches to one or both of these viruses with amino acid identities between 30-60%. The archaeal viral family *Globuloviridae* has been proposed for PSV and TTSV1 (Ahn, 2006; Haring, 2004), and likely the putative RSS virus would be the third member of this family.
This study demonstrated the feasibility of performing metagenomic sequencing on the viral fraction of a high temperature acidic aqueous environment. More importantly, it showed us that total community sequencing in low complexity environments can yield high coverage of the dominant community members, even with a small sequencing project. Recent improvements in sequencing technology and assembly of metagenomic data sets have streamlined this process and are allowing us to perform much more sophisticated analyses (chapter 4).

The next step in this project is to identify the viral morphotype and host organism. Efforts in this area are being made to develop a fluorescence activated viral sorting (viral FACS) protocol of direct environmental samples. A viral FACS assay using nucleic acid probes to the viral sequence should allow for identification of host organism (infected cells) as well as the viral morphotype associated with this genome.
CHAPTER 4

CRISPR/CAS ADAPTIVE IMMUNE FUNCTION IN A NATURAL ENVIRONMENT

Introduction

In natural environments, viruses play a critical role in defining the microbial community (Beltran Rodriguez-Brito, 2010; Furhman, 1999; Kunin et al., 2008; Rodriguez-Brito et al., 2010; Wommack and Colwell, 2000). However, we have only a limited understanding of the temporal dynamics of virus-cellular interactions and their effects on defining the microbial community structure in natural environments. Two models have been proposed to describe the coevolution of virus (parasite)-host dynamics (Gandon et al., 2008). The ‘arms race dynamics’ (ARD) holds that both host and virus continually accumulate adaptive mutations. The fluctuating selection dynamics model (FSD) proposes that virus and host frequencies oscillate over time because of negative frequency-dependent selection. While under controlled laboratory conditions the dynamics of host-virus interactions have been extensively studied (Barrangou et al., 2007; Brouns et al., 2008; Prangishvili et al., 2001), there are few studies that have examined host-virus interactions in natural environments (Breitbart et al., 2004; Chen et al., 2009; Desnues et al., 2008; Rodriguez-Brito et al., 2010). This is due in part to the enormous complexity of both host and viral communities present in most natural environments. In contrast, extreme environments often have much less
complex microbial community structures (Chivian et al., 2008; Inskeep et al., 2010; Schoenfeld et al., 2008; Tyson et al., 2004).

A characteristic function of an adaptive cellular immune system is to maintain, at least temporarily, a memory of an invading pathogen so that the host can respond to future assaults by that same element. The recently described adaptive immune system termed the CRISPR/Cas system (Clusters of Regularly Interspaced Palindromic Repeats and CRISPR associated genes) consists of 23-50bp repeat structures separated by short 30-40bp spacer sequences (Grissa et al., 2007a; Grissa et al., 2007b). CRISPR spacer sequences are derived from mobile DNA elements such as viruses, plasmids, and transposable elements (Mojica et al., 2005) and protect the host organism against such elements. Short crRNAs transcribed from the CRISPR array function as a nucleic acid mediated interference system. The function of the CRISPR/Cas system in Archaea is just beginning to be elucidated (Lillestol et al., 2009; Manica et al., 2011; Shah and Garrett, 2010; Shah et al., 2009); however, it is likely to operate in a similar fashion as it does in Bacteria. It has not been directly demonstrated that the CRISPR/Cas systems function in natural environments.

The spacer sequence content within CRISPR loci likely reflects a cell’s recent past interactions with viruses and other mobile genetic elements. CRISPR loci vary greatly in length, containing from as few as two repeat/spacer units to hundreds of repeat/spacer units. The number of different CRISPR loci within a genome can also vary from one to at least 23 (CRISPRdb (Grissa et al., 2007b)). Often, but not always, cas genes are found
adjacent to CRISPR loci. Acquisition of new CRISPR spacer sequences is thought to occur upon exposure of the cell to a virus which results in the polar addition of new spacer sequence proximal to the leader of the of the CRISPR locus, with degeneration and loss of older or downstream spacer sequences. Under laboratory conditions, rapid evolution of CRISPR loci has been observed (Barrangou et al., 2007; Held and Whitaker, 2009; Kunin et al., 2008), and the newly acquired spacer provides specific protection against the represented phage (Barrangou et al., 2007).

In environmental samples, analysis of closely related but geographically isolated organisms has shown conservation of repeats and flanking regions with no conservation of the CRISPR spacer sequences (Kunin et al., 2008). The hypervariability of CRISPR spacer content is thought to provide a population level resistance to invading viruses (Tyson and Banfield, 2008).

In contrast to this generally held view, adaptive immune function of CRISPR/Cas has recently been questioned by several groups. Babu et al. (2011) have shown that while the system may have immune function it also appears to be involved in DNA repair. Touchon et al. (2011) and Touchon and Rocha (2010), in comparing hundreds of strains of *Escherichia* and *Salmonella*, have demonstrated that the rate of change over hundreds of thousands of generations is not suggestive of an effective adaptive immune system.

The CRISPR/Cas system likely plays a role in defining the viruses that can successfully infect potential host species. While the host may rapidly acquire a new
virus derived spacer sequence(s) within its CRISPR loci to provide protection against infection, a subset within the viral population can evade inactivation due to viral mutations which are mis-matched to the CRISPR spacer sequence(s) (Deveau et al., 2008). The ability to evade the CRISPR/Cas defense system imparts fitness to that viral genotype, potentially allowing it to become the new dominant member of the viral population. Whether the coevolution of the cellular CRISPR/Cas-virus dynamics follows more of an ARD or FDS model is currently unknown.

High temperature acidic environments are typically dominated by a few archaeal host species and their viruses (Hall et al., 2008; Inskeep et al., 2010; Tyson et al., 2004). The Nymph Lake hot spring #10 (referred herein as the NL hot spring) is a high temperature acidic hot spring (86 – 92°C, pH3) located in Yellowstone National Park (YNP). Initial 16S rDNA analysis suggested that the NL hot spring supports a low complexity archaeal cellular community, and we also suspect that it supports a relatively low complexity archaeal viral community. The NL hot spring provides a small island-like environment in which the total microbial community can be accessed using cellular and viral metagenomic approaches. We therefore have used the NL hot spring as a natural environment to examine the temporal dynamics of the cellular CRISPR/Cas system in response to the viral community.
Site Selection and Chemical Analysis

Nymph Lake #10 (N44 45.214 W110 43.432) was chosen as the site for this study because of its high temperature (86 – 92°C) and low pH (3). The site is located north of Nymph Lake in a geologically new thermal area that has been relatively dynamic over the past decade (Lowenstern et al., 2005).

Water chemistry was analyzed using a water chemistry kit from Energy labs in Billings, Montana. Samples were collected, preserved according to their protocols and shipped to their facilities for analysis.

Host Sample Collection and Preparation

Host samples were collected at the following time points: Aug. 2007, Jan., April, June, Aug. and Sept. 2008, Jan., March, April, June, July and Aug. 2009. Metagenomic sequencing was performed for the Aug. 2007, Aug. 2008, and Aug. 2009 samples. PCR amplified CRISPR clone libraries were prepared for all time points.

Host samples were collected in 500 mL sterile Nalgene bottles and transported at ambient temperature to the lab. Samples were filtered onto three 0.45μm filters (Pall) per 500 mL of sample frozen at -80°C.

DNA was extracted from the filters by incubating the filters in lysozyme (1 mg/ml) at 37°C for 1 hr followed by Proteinase K (0.2 mg/ml) and SDS (1%) at 55°C for 1 hr. DNA was extracted from the lysate using phenol:chloroform:isoamyl alcohol
(25:24:1) followed by ethanol precipitation (0.1X volume NH₄OAC and 2X volume ethanol). DNA concentration varied from 200ng/μL to undetectable levels. Dilutions of the DNA were made for PCR amplification.

**DNA Sequencing**

The DNA from the August 2007 sample was sent to the Joint Genome Institute (Walnut Creek, CA) where 454 Titanium sequencing was performed along with paired end read Sanger sequencing of a small insert library.

DNA from the August 2008 and August 2009 samples were amplified using the Whole Genome Amplification kit (Sigma) and sent to the DNA Sequencing Facility of the Roy J. Carver Biotechnology Center at the University of Illinois (Urbana-Champaign, IL) for 454 Titanium sequencing. A total of 370034 reads were obtained for the August 2008 sample and 408256 reads for the August 2009 sample. Sequencing statistics are presented in table 4.1.

PCR was performed using Taq enzyme (Promega) and various dilutions of extracted DNA. Primers used were NLcF (CGCGTAAATGCACGTGTAAA) and 2732R (CTAATCCCTTCTATCCGCG). Cycling conditions were as follows: initial denaturation of 95°C for five minutes; 35 cycles of 95°C for 30 s; 55°C for 30 s; 72°C for 1 m, followed by a final extension of 72°C for 10 m.

PCR products were cloned into pCRII-Topo-TA vector (Invitrogen). Colonies were grown and plasmid isolated using the Millipore Montage Plasmid Mini Prep 96 kit. Sequencing reactions were performed in 96 well format using Topo F primer
(GCGAATTGGGCCCTCTA). Sequences were determined on an Applied Biosystems 3730 DNA Analyzer at Idaho National Laboratory (Idaho Falls, ID).

Table 4.1  Cellular metagenome sequencing and assembly statistics

<table>
<thead>
<tr>
<th>Site, Type, Date</th>
<th>Sequencing Center</th>
<th>Number of Reads</th>
<th>Number of Bases</th>
<th>Number Aligned Reads</th>
<th>Number Aligned Bases</th>
<th>Inferred Read Error</th>
<th>Reads &gt; 301 coverage</th>
<th>Number Assembled</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL2007-08 cellular</td>
<td>JGI</td>
<td>579976.00</td>
<td>244 MB</td>
<td>550854 (94.98%)</td>
<td>233112020 (95.36%)</td>
<td>0.01</td>
<td>0.01</td>
<td>1400.00</td>
</tr>
<tr>
<td>NL2008-08 cellular</td>
<td>IL</td>
<td>370034.00</td>
<td>143 MB</td>
<td>267838 (72.38%)</td>
<td>104187883 (72.83%)</td>
<td>0.02</td>
<td>0.02</td>
<td>802.00</td>
</tr>
<tr>
<td>NL2009-08 cellular</td>
<td>IL</td>
<td>408256.00</td>
<td>134 MB</td>
<td>229339 (56.18%)</td>
<td>78959242 (59.04%)</td>
<td>0.02</td>
<td>0.02</td>
<td>800.00</td>
</tr>
</tbody>
</table>

Viral Sample Collection and Preparation

Virus-enriched samples were collected at the following time points: Jan., Aug. and Sept. 2008, Jan., March, April, Aug. and Sept. 2009. Samples for viral metagenomic sequencing and PCR analysis were collected by filtering 100mL of hot spring water through two 0.8/0.2 µm filters (Pall) directly into four sterile 23 mL ultracentrifuge tubes. These were transported at ambient temperature back to the lab within four hours and immediately spun at 100,000 x G for two hours. The supernatant was
removed and the samples resuspended in a small volume (0.25 – 1.0 mL) of sterile water.

The sample from Jan. 2008 was sent to JGI for amplification (using NEB Phi29 polymerase for 16h at 30°C) and 454-pyrosequencing. Samples from Aug. and Sept. 2008, and Jan., March, April, Aug. and Sept. 2009 samples were amplified by Whole Genome Amplification per manufacturer’s protocols (Sigma) and then sent to the Illinois sequencing center for 454 sequencing. Assemblies were performed using Newbler GS Assembler v2.3. Sequencing statistics are presented in table 4.2.

Table 4.2. Viral metagenome sequencing and assembly statistics

<table>
<thead>
<tr>
<th>Site, Type, Date</th>
<th>2008-01</th>
<th>2008-08</th>
<th>2008-09</th>
<th>2009-01</th>
<th>2009-03</th>
<th>2009-04</th>
<th>2009-09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing Center</td>
<td>JGI</td>
<td>IL</td>
<td>IL</td>
<td>IL</td>
<td>IL</td>
<td>IL</td>
<td>IL</td>
</tr>
<tr>
<td>Number of Reads</td>
<td>239369</td>
<td>52242</td>
<td>198855</td>
<td>166621</td>
<td>211161</td>
<td>203754</td>
<td>164818</td>
</tr>
<tr>
<td>Number of Bases</td>
<td>89.6 MB</td>
<td>18.3 MB</td>
<td>78.1 MB</td>
<td>66.2 MB</td>
<td>82.4 MB</td>
<td>79.6 MB</td>
<td>63.9 MB</td>
</tr>
<tr>
<td>Number Assembled</td>
<td>88488</td>
<td>8203</td>
<td>60173</td>
<td>41435</td>
<td>52212</td>
<td>55451</td>
<td>46109</td>
</tr>
<tr>
<td>Large Contigs Number (&gt;1000 bps)</td>
<td>103</td>
<td>588</td>
<td>519</td>
<td>509</td>
<td>657</td>
<td>464</td>
<td>540</td>
</tr>
<tr>
<td>Large Contig Bases</td>
<td>0.3 MB</td>
<td>1.6 MB</td>
<td>0.9 MB</td>
<td>0.9 MB</td>
<td>1.1 MB</td>
<td>0.8 MB</td>
<td>0.9 MB</td>
</tr>
<tr>
<td>Average Large Contig Size</td>
<td>2533</td>
<td>2690</td>
<td>1735</td>
<td>1689</td>
<td>1722</td>
<td>1814</td>
<td>1613</td>
</tr>
<tr>
<td>Largest Contig Size</td>
<td>23.2 KB</td>
<td>24.9 KB</td>
<td>11.1 KB</td>
<td>8.5 KB</td>
<td>8.5 KB</td>
<td>9.0 KB</td>
<td>7.7 KB</td>
</tr>
<tr>
<td>All Contigs Number</td>
<td>1216</td>
<td>1296</td>
<td>4480</td>
<td>5579</td>
<td>7337</td>
<td>5565</td>
<td>6770</td>
</tr>
<tr>
<td>All Contigs Bases</td>
<td>36.9 MB</td>
<td>15.7 MB</td>
<td>24.87</td>
<td>24.73</td>
<td>27.21</td>
<td>27.98</td>
<td>24.03</td>
</tr>
</tbody>
</table>

CRISPR Analysis

In order to identify CRISPR-containing reads, direct repeat sequences were subjected to BLAST analysis against the August 2007 cellular metagenome to identify sequence reads that contained possible spacers. These reads were then submitted to
the CRISPRfinder website (http://crispr.u-psud.fr/Server/CRISPRfinder.php) to identify spacers.

PCR amplified sequence reads were manually edited (trimmed and corrected for ambiguous base calls) and then submitted to the CRISPR finder website using the CRISPRtionary tool to compare repeat and spacers between reads. Zero mismatches were allowed for defining unique spacers. A presence/absence binary file was generated by the CRISPRtionary tool.

The CRISPR spacers from the cellular metagenome and the PCR were aligned to the viral metagenome reads and contigs using the FASTA program (Pearson and Lipman, 1988). Alignments were analyzed at 100% nt identity and at 83% nt identity and no more than 5 bp mismatches. Matches to the viral metagenomes were normalized to the size of the metagenome for comparison between data sets.

Cell and Virus Enumeration

Cells and viruses were counted by epiflourescence microscopy as previously described (Ortmann and Suttle, 2009).

Results

The NL Hot Springs has a Relatively Stable Water Chemistry that Supports a Low Biomass Microbial Community

The physical properties and general water chemistry of the NL hot springs remained relatively constant over the monitoring period (figure 4.1). Chemical analysis
of the hot spring water samples at five points over the course of nine months indicates that this environment is a sulfate-chloride hot springs found within YNP (Inskeep et al., 2010; Snyder et al., 2007). The major physical change observed was seasonal fluctuations in the volume of surface water in the hot spring.

Fluorescent-based total cell counts of NL hot spring environmental samples indicate that it has a low cell density (6 x 10^6 cells/ml) and low density of virus like particles (VLP) (10^5 VLPs/ml). While in many other types of environmental samples the VLP to cell ratio is typically ~10, it is known that this ratio can vary greatly between different types of environmental samples (Suttle, 2007; Wommack and Colwell, 2000).

![Figure 4.1. Water chemistry from 6/23/08 – 10/14/09. Not detected: Antimony, Barium, Copper, Lead, Selenium, and Mercury.](image)

We suspect that the low cell and virus densities and VLP to cell ratio reflects a nutrient limited environment limiting cellular replication and minimizing virus exposure to the extreme environment free from its intracellular environment.
The NL Hot Springs Supports a Low Complexity Archaeal Cellular and Virus Community

Total community DNA sequencing of the NL hot spring cellular fraction was performed at three time points each one year apart. A total of 263 MB of DNA sequence was generated with 14.2MB representing a unique sequence (an average of 18X base coverage) indicating significant DNA sequence coverage of at least the dominant cellular genetic components present in the hot spring (table 4.1). There was 11-30% overlap between sequencing reads among the three time points, indicating that common organisms were present during the two-year sampling period.

Analysis of the August 2007 cellular metagenome indicated an environment dominated by a single archaeal organism. A GC plot of individual sequence reads showed one predominate peak (figure 4.2a) and principal component analysis (PCA) of oligonucleotide frequency distributions (figure 4.2b) further indicated a single dominant

Figure 4.2. Analysis of Aug. 2007 metagenome. A. GC plot of all reads from the NL August 2007 cellular metagenome. B. PCA analysis of oligonucleotide frequency of the NL August 2007 cellular metagenome.
cellular community member (provided by Doug Rusch, JVCI). A total of seventeen 16S genes larger than 500 bps were found in all three of the cellular metagenomes. These 16S genes showed similarity to thermophilic sulfur-dependent organisms in the *Sulfolobales* family or uncultured archaeal clones (Table 4.3). This concurs with the analysis of clone libraries generated using a targeted PCR-directed amplification of the 16S rDNA gene. The initial 16S rDNA sequencing also indicated a low complexity environment.

Total viral community DNA sequencing was performed on eight environmental samples over the two-year period. Analysis of these data sets also demonstrated a low complexity of likely novel archaeal viral genotypes. A large portion (15-36% of reads) of each of the viral metagenomes assembled into contigs (Table 4.2), many contigs having sequence coverage depth in excess of 100X. The largest contigs in these metagenomes were in the range of full-length viral genomes, with the largest at 24 Kb, further indicating a lack of complexity within this environment. Four randomly chosen large contigs were confirmed by PCR-based amplification and DNA sequencing. BLASTx searches of the non-redundant NCBI database with all viral metagenome contigs produced few matches to known viruses with the exception of isolated open reading frames with homology to proteins from a few thermophilic archaeal viruses, notably SSVs, STIV, and ASV (Table 4.4). There was little apparent cross contamination of cellular sequences in the environmental virus fraction. Only 1.5-7.5% of the sequences reads in
the viral fraction showed significant homology (>98% identity over a 50 bp region) to sequence reads derived from the paired cellular fraction prepared at the same time.

Table 4.3. 16S genes (>500 bps) identified in all three cellular metagenomes.

<table>
<thead>
<tr>
<th>Contig</th>
<th>Contig length</th>
<th>Number of reads</th>
<th>Top BLASTn hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLO078c-contig03130</td>
<td>13409</td>
<td>4376</td>
<td>Uncultured archaean clone HS3wa_S2_16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig00060</td>
<td>1404</td>
<td>13</td>
<td>Uncultured Pandoraea sp. clone CHINA6 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig00065</td>
<td>1428</td>
<td>261</td>
<td>Hydrogenobaculum Hydrogenobaculum sp. NOR3L3B partial 16S rRNA gene, strain NOR3L3B</td>
</tr>
<tr>
<td>NLO078c-contig00083</td>
<td>742</td>
<td>26</td>
<td>Uncultured archaean clone SK428 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig00091</td>
<td>843</td>
<td>3</td>
<td>Uncultured Geothermobacterium sp. clone BV129 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig00109</td>
<td>1377</td>
<td>175</td>
<td>Uncultured archaean clone KOZ125 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig00152</td>
<td>1579</td>
<td>26</td>
<td>Uncultured nanoarchaeote 16S rRNA gene, clone OP-9</td>
</tr>
<tr>
<td>NLO078c-contig00164</td>
<td>711</td>
<td>2</td>
<td>Uncultured Rhabdolumis clone GN24DB2 18S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig00840</td>
<td>1150</td>
<td>312</td>
<td>Uncultured Aquilales bacterium clone YNP_SSsp_B60 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig01538</td>
<td>1434</td>
<td>177</td>
<td>Uncultured Thermocinetis sp. clone YNP_SBC_BP2A_B3 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig01715</td>
<td>1378</td>
<td>24</td>
<td>Uncultured archaean gene for 16S rRNA, partial sequence, clone: H028S9A9S</td>
</tr>
<tr>
<td>NLO078c-contig01752</td>
<td>837</td>
<td>17</td>
<td>Uncultured archaean clone Tn1-a26 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO078c-contig01781</td>
<td>1247</td>
<td>331</td>
<td>Uncultured Vulcaniaeta sp. clone SK409 16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO080c-contig00571</td>
<td>518</td>
<td>7</td>
<td>Uncultured nanoarchaeote 16S rRNA gene, clone OP-9</td>
</tr>
<tr>
<td>NLO080c-contig03639</td>
<td>1520</td>
<td>71</td>
<td>Uncultured archaean clone HS3wa_S2_16S ribosomal RNA gene, partial sequence</td>
</tr>
<tr>
<td>NLO080c-contig03472</td>
<td>966</td>
<td>17</td>
<td>S. tokodai str.7 DNA, complete genome</td>
</tr>
<tr>
<td>NLO090c-contig13344</td>
<td>2613</td>
<td>89</td>
<td>Uncultured nanoarchaeote 16S rRNA gene, clone OP-9</td>
</tr>
</tbody>
</table>

Temporal Dynamics of a CRISPR Locus

Three CRISPR loci were identified in the initial cellular metagenome (August 2007). All three shared a common direct repeat structure that belongs to the Sulfolobales DRIII family (Lillestol et al., 2009). These three CRISPR loci contained a total of only 26 unique CRISPR spacer sequences. Two of these CRISPR loci were located at the end of existing contigs (contig 821 and contig 1277) and may be truncated in the middle of a CRISPR locus, so that the exact length of these CRISPR
loci cannot be determined. The third locus was located in the middle of a large (37 Kb) contig (contig 2635) and is a complete CRISPR locus with only 5 repeat/spacer units.

Table 4.4. BLASTx matches to viral metagenome contigs.

<table>
<thead>
<tr>
<th>Query</th>
<th>Alignment Title</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLv0808contig00002</td>
<td>hypothetical protein LKI_03710 [Leuconostoc kimchii IMSNU11.154]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0808contig00002</td>
<td>dehydrogenase (flavoprotein) [Leuconostoc citreum KM20]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0808contig00002</td>
<td>FAD(NAD)-dependent oxidoreductase [Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0808contig00035</td>
<td>ABC transporter ATP-binding and permease protein [Leuconostoc kimchii IMSNU11.154]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0808contig00035</td>
<td>peptide ABC transporter ATPase [Leuconostoc citreum KM20]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0809contig00944</td>
<td>ORF A460 [Sulfobolus virus Kamchatka 1]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0809contig01648</td>
<td>transposase, IS605 OrfB family [Sulfobolus islandicus M.16.4]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0809contig01776</td>
<td>ribonuclease-diphosphate reductase, adenosylcobalamin-dependent [Sulfobolus islandicus L.S.2.15]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0809contig01776</td>
<td>ribonuclease-diphosphate reductase large subunit [Sulfobolus tokodaii str. 7]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0809contig02003</td>
<td>ORF B494 [Sulfobolus virus Kamchatka 1]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0901contig00538</td>
<td>hypothetical protein pRNA_06 [Sulfobolus islandicus]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0901contig00538</td>
<td>hypothetical protein ST0314 [Sulfobolus tokodaii str. 7]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0901contig00538</td>
<td>RnpA protein [Sulfobolus islandicus]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0901contig00538</td>
<td>phosphatidyltransferase, P4 family [Sulfobolus islandicus L.S.2.15]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0903contig00024</td>
<td>hypothetical protein ST0954 [Sulfobolus tokodaii str. 7]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0903contig00024</td>
<td>transposase, IS605 OrfB family [Sulfobolus islandicus M.16.27]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0903contig00024</td>
<td>hypothetical protein pHeV14_21 [Sulfobolus islandicus]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0903contig00024</td>
<td>hypothetical protein ST2552 [Sulfobolus tokodaii str. 7]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0903contig01098</td>
<td>bifunctional DNA primase/polymerase [Sulfobolus islandicus M.14.25]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0903contig01469</td>
<td>ORF B494 [Sulfobolus virus Kamchatka 1]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0904contig03367</td>
<td>transposase, IS605 OrfB family [Sulfobolus islandicus L.S.2.15]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0904contig03367</td>
<td>putative transposase [Sulfobolus sp. L00 11]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0904contig01588</td>
<td>hypothetical protein M1425_0506 [Sulfobolus islandicus M.14.25]</td>
<td>1.17E-17</td>
</tr>
<tr>
<td>NLv0904contig03316</td>
<td>coat protein [Sulfobolus turreted icosahedral virus]</td>
<td>2.62E-17</td>
</tr>
<tr>
<td>NLv0904contig03316</td>
<td>hypothetical protein STIV2_A345 [Sulfobolus turreted icosahedral virus 2]</td>
<td>5.47E-165</td>
</tr>
<tr>
<td>NLv0908contig01462</td>
<td>transposase, IS605 OrfB family [Sulfobolus islandicus L.S.2.15]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0908contig01462</td>
<td>putative transposase [Sulfobolus sp. L00 11]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0908contig04313</td>
<td>putative end-flament protein [Acidobacterium end-flament]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0908contig04313</td>
<td>putative end-flament protein [Sulfobolus sp. end-flament]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0908contig04495</td>
<td>ribonuclease-diphosphate reductase (red) [Sulfobolus solitaricus 982]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0909contig00488</td>
<td>putative end-flament protein [Sulfobolus sp. end-flament]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0909contig00488</td>
<td>hypothetical protein Saci_1002 [Sulfobolus acidocaldarius DSM 639]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0909contig02168</td>
<td>ORF B494 [Sulfobolus virus Kamchatka 1]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0909contig02168</td>
<td>ORF A460 [Sulfobolus virus Kamchatka 1]</td>
<td>0</td>
</tr>
<tr>
<td>NLv0909contig02775</td>
<td>hypothetical protein M164_0757 [Sulfobolus islandicus M.16.4]</td>
<td>0</td>
</tr>
</tbody>
</table>

There were no cas genes found in the neighborhood (within 10Kb) of any of these loci. However, there were cas genes found within this metagenome. Cas4 family, Csa2 family and Cmr1 and CRISPR locus-related DNA binding proteins were identified.
The small size of the CRISPR locus and the lack of neighboring cas genes suggested that this might be an inactive locus. However, during the two-year time span we were always able to detect the locus, and there were no changes in the sequence of direct repeat units or the leader sequences, indicating that this locus is under selection pressure to maintain its function. Furthermore, RT-PCR indicated that this locus is being transcribed, suggesting that the locus is indeed active (data not shown). Bioinformatic analysis of contig 2635 did not provide direct evidence that this locus was plasmid based or part of a transposable element.

The temporal dynamics of the CRISPR locus originally identified on contig 2635 were examined over a two-year time course. PCR amplification and cloning of the contig 2635 CRISPR locus over eleven independent time points over the course of two years demonstrated both periods of relative stability and maintenance of spacer composition as well as periods of rapid spacer diversification (figure 4.3). The number of clones sequenced at each time point varied, from 23-66 clones. A total of 510 clones were sequenced across the two-year time period. The CRISPR locus varied in total spacer/repeat units throughout this time, with as few as zero and as many as 13 spacers detected, with the vast majority having 5 spacer/repeat units per loci. There were a total of 243 unique spacers detected at this locus. Average spacer length was 36 nts and ranged from 20-42 nts, with 98% longer than 34 nts. There was little variation in length (24bp) or sequence content of the direct repeats. Two classes of CRISPR spacer sequences were detected (figure 4.3). The first class, and
the majority of CRISPR spacers, appeared and disappeared rapidly, with 133 CRISPR spacers (55% of total unique CRISPR spacers detected) detected only at a single time point. The second class represented clusters of CRISPR spacers that were detected as a linked set of spacers over multiple time points. This second class consisted of 15 spacers (7% of total CRISPR spacer sequences) making up three sets. The most dominant CRISPR spacer sequence set consists of the same five linked spacer sequences and was detected at all but two time points: June 29, 2009 and August 2009. The second set consists of five linked spacers that were first detected in June 2008 and were found periodically throughout the time course. The third set consists of five spacers that were first detected at low levels beginning in September 2008 and were found continually at low levels with the exception of April 2009, when it was not detected. On June 29, 2009 it became the dominant set detected.

Gradual change does not appear to be the main mechanism of change in the monitored CRISPR locus. Sets of CRISPR spacer sequences appeared and disappeared in blocks, although this was not universal. There were 33 cases (120 clones) where we observed the addition of a single or multiple CRISPR spacers to the locus that at a previous time point did not contain that spacer sequence. There were also five examples (20 clones) of CRISPR spacers that were added or deleted internally in the locus, and there was one example (2 clones) of spacer sequence “stuttering,” or duplication. However, the majority of all change that was observed was the complete turnover of full sets of CRISPR spacer sequence groups (321 clones).
Figure 4.3. Variation of spacer sequence content within the monitored CRISPR locus over 25 months of sampling. Unique spacer sequences (243) are represented across the horizontal axis. The detection of a particular spacer sequence within the same sampling time or between sampling points is represented on the vertical axis. Examples of groups of CRISPR spacer sequences that are linked and are detected as blocks (black arrows) or as individual spacers (gray arrow) are indicated. Spacer-less CRISPR sequences are indicated as black boxes.

A new class of CRISPR loci, we termed “spacer-less,” was detected which accounted for 10% (49/510 clones) of all PCR generated sequence reads across the entire time course and one read within the metagenomic data set. They consisted of the CRISPR leader sequence, followed by single complete direct repeat, followed by the downstream sequence. No CRISPR spacer sequences were present. At three time
points, these were the dominant clone types present: 16 clones (42% of clones sequenced at this time point) in August 2008, 12 clones (42% of clones sequenced at this time point) in September 2008 and 12 clones (37% of clones sequenced at this time point) in January 2009. At the same time, previously dominant CRISPR spacer sequence sets become subdominant within the clones sequenced at that particular time (figure 4.3).

An inverse Simpson diversity index showed an oscillating and overall increasing CRISPR spacer sequence diversity with time (figure 4.4). The diversity was stable and low for the first three sample points spanning ten months. The subsequent fourteen months were characterized by oscillations with a periodicity of five to seven months.

![Figure 4.4. Inverse Simpsons index of diversity of spacers across time (solid line). Number of spacer-less reads across time (dotted line).](image)
Moreover, the oscillations appear to be negatively correlated to the appearance of spacer-less CRISPR reads, suggesting a potential involvement of these read types in the reduction of diversity.

Spacer Presence Correlated to the Subsequent Presence of Matching Viral Sequence

Total viral enriched community DNA sequencing was performed on environmental samples collected in Jan. Aug. and Sept. 2008, and Jan., Mar., Apr., Aug. and Sept. 2009. CRISPR spacer sequences from the PCR library were aligned to the viral metagenome reads and contigs using the FASTA program. Alignments with zero mismatches (table 4.5) and 1-5 bp mismatches (table 4.6) were analyzed. There were a total of 342 exact matches and 662 1-5bp mismatch hits. Forty out of 243 (17%) total unique CRISPR spacer sequences accounted for all of the exact matches. This was in

Table 4.5. Spacer matches into each of the viral metagenomes allowing zero mismatches (100% identity). Normalized to the size of the metagenome. Shaded boxes indicate paired viral and cellular metagenomes.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Metagenome MB -&gt;</td>
<td>0.7 1.9 2.4 2.7 3.5 2.9 3.4</td>
<td>3.9</td>
<td>0.7</td>
<td>1.9</td>
<td>2.4</td>
<td>2.7</td>
<td>3.5</td>
<td>2.9</td>
<td>3.4</td>
</tr>
<tr>
<td>26</td>
<td>NL0708c-0bp</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>NL0804p-0bp</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>50</td>
<td>NL0806p-0bp</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>55</td>
<td>NL0808p-0bp</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>NL0809p-0bp</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>27</td>
<td>NL0901p-0bp</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>NL0903p-0bp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>NL0904p-0bp</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>NL0906p-0bp</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>NL0908p-0bp</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 4.6. Spacer matches into each of the viral metagenomes allowing 5 mismatches (~83% identity). Normalized to the size of the metagenome. Shaded boxes indicate paired viral and cellular metagenomes.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Metagenome M8 -&gt; 0.7</td>
<td>1.9</td>
<td>2.4</td>
<td>2.7</td>
<td>3.5</td>
<td>2.9</td>
<td>3.4</td>
<td>3.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>NL0708c</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>NL0804p</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>NL0806p</td>
<td>4</td>
<td>2</td>
<td>12</td>
<td>11</td>
<td>14</td>
<td>10</td>
<td>15</td>
<td>29</td>
</tr>
<tr>
<td>55</td>
<td>NL0808p</td>
<td>2</td>
<td>3</td>
<td>15</td>
<td>8</td>
<td>12</td>
<td>9</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>NL0809p</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>27</td>
<td>NL0901p</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>21</td>
<td>NL0903p</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>NL0904p</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>12</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>48</td>
<td>NL0906p</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>NL0908p</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

contrast to the NL CRISPR spacer content of the viral metagenomes obtained from a similar acidic high temperature YNP spring 30km away where <1% of the NL CRISPR spacer sequences matched the viral population present at this second hot spring (data not shown). This demonstrates that there is a correlation between the cellular CRISPR spacer sequence content and the viral population sequence content present within the same environment.

Unexpectedly, the appearance of CRISPR spacer sequences correlated to a maintenance or increase in the number of matches in the viral metagenomes at subsequent time points (figure 4.5). When we relaxed the match stringency to allow for up to 5 bp mismatches there was a weak trend of increasing matches over time. This suggests that viruses are not rapidly mutating their proto-spacers to escape CRISPR immunity. While this trend was observed for the group of the CRISPR spacer sequences
as a whole detected at a particular time, it was not as evident when following individual CRISPR spacer sequences over time. Not all matches to individual spacers increased over time. However, within a set of linked CRISPR spacers where more than one spacer matched into the viral metagenomes, in every case one of the matching spacers increased in viral matches over time.

Figure 4.5. Increasing spacer matches into viral metagenomes over time. Data is normalized to the size of the metagenome. Solid diamonds are exact matches and solid line is linear regression for exact matches \((R^2=0.84)\). Solid squares are \(\leq 5\) bp mismatches and dotted line is linear regression for \(\leq 5\) bp mismatches \((R^2=0.34)\).

The position of the matching CRISPR spacers within an array with regards to the leader end was analyzed. Twenty-one out of forty matching spacers were either at the 5’ end or one spacer downstream from the 5’ end, while only one matching spacer was positioned at the 3’ end. This concurs with the existing view of CRISPR array evolution, that more recent viral encounters are represented by the upstream spacers.
CRISPR spacer sequences from the PCR library were also aligned to the cellular metagenome reads and contigs using the FASTA program, and total Spacer-containing Non-CRISPR reads (SNCs) were analyzed (figure 4.6). The number of SNC matches decreased over time for both exact matches (36% reduction) and the 1-5 bp mismatches (80% reduction). This suggests that intracellular virus population is being modulated by the existence of a matching CRISPR spacer. Alternatively, SNCs could be declining in abundance over time because cellular genomes containing them are being selected against.

![Decreasing spacer matches to cellular metagenomes (non-CRISPR matches) over time. Data is normalized to the size of the metagenome. Solid diamonds are exact matches and solid line is linear regression for exact matches (R2=0.87). Solid squares are ≤5 bp mismatches and dotted line is linear regression for ≤5 bp mismatches (R2=0.86).](image)

The total number of viral reads that each CRISPR spacer sequence aligned to over time was analyzed as an indication of abundance of the viral genotypes being targeted by CRISPR spacers (figure 4.7). There was a significant peak spanning a five-
month period (encompassing time points, NLv0808, NLv0809, and NLv0901). Fourteen of 40 CRISPR spacer sequences mapped to this peak. There was no correlation either as to when a CRISPR spacer sequence was first detected in the host derived PCR library or whether it was represented in this peak. Spacers that had been detected months before August 2008, as well as spacers that were detected months after January 2009, were found to peak at that time point. There is also a second smaller peak observed 12 months later, suggesting the possibility of oscillating seasonal variation. Six of the

![Figure 4.7](image-url)

Figure 4.7. Individual spacers matched to viral metagenome reads, normalized to the size of the viral metagenome.

fourteen spacers that mapped to the first peak also mapped to this second peak. An additional seven spacers only peaked at the second time point. The remainder of the 40 spacers had few exact matches and showed no apparent pattern over time.
Discussion

The overall findings of this study demonstrate that the cellular CRISPR may be episodically advantageous but prone to inactivation by excision of the spacers in a natural hot spring environment. By following a single CRISPR locus over the course of two years, we observed both periods of rapid change on the order of less than a month and periods of little change for periods up to seven months in the dominant CRISPR spacer sequence content. However, change at this CRISPR locus was characterized by substitution of entire blocks of CRISPR spacer sequences rather than gradual polar addition of individual CRISPR spacer sequences as has been observed by others in natural environments (Andersson and Banfield, 2008) or under laboratory conditions (Barrangou et al., 2007; Manica et al., 2011).

The detection of blocks of CRISPR spacer sequences that are present early in the sampling period, diminish for a period and then reappear again as the dominate CRISPR spacer sequence block (e.g. CRISPR spacer blocks #2 and #3), suggests that CRISPR spacer sequence blocks can rapidly move within the population. Movement and spread within the host population could be a function of being localized on a mobile genetic element (e.g. plasmids or transposable elements), selected from the sub-dominant CRISPR type existing within the resident population, or movement (immigration) from an external source (e.g. another hot springs). However, analysis of contig 2635 did not provide direct evidence that this locus was plasmid based or part of a transposable element.
The observation of a significant number of spacer-less CRISPRs introduces the possibility of a new mechanism for change within this locus that results in the complete “sweep” of all spacer sequences (figure 4.8). The spacer-less CRISPRs may be a cellular driven mechanism by which the CRISPR locus undergoes homologous recombination between repeats to remove all spacers. This locus would then be ready for rapid acquisition of a complete block of spacers from another locus within the genome or from other members in the population. The mechanism by which this transfer could

![Figure 4.8. Proposed model of the function of CRISPR/Cas in the context of a nutrient limiting environment and frequency dependent selection.](image-url)
occur is not known. Alternatively the spacer-less CRISPRs might be the result of an active viral counter-response in order to evade the host encoded immune response, or in the absence of positive selection to maintain a locus (i.e. viral infection pressure), the spacers are eliminated.

Surprisingly, there appears to be a bias for viruses to be present in the environment that match the CRISPR spacer sequences after the initial detection of the spacer in the cellular population (Tables 4.5 and 4.6). This is not the expectation of an immunity system whereby the host retains memory of the infectious agent. One would expect that if the CRISPR/Cas system were 100% effective as an antiviral defense system, that the presence of spacer sequences would preclude detection of viral genomic sequences identical to the spacer sequences present at the same time or at subsequent times. This is not the case. We observe multiple examples of identical matches between the CRISPR spacer sequences and viral genome sequences in paired samples. In fact, there is maintenance or even an increase in exact matches of spacer sequences to the viral population after the time point the CRISPR spacer sequence is first detected.

We outline a general model that integrates aspects of the conventional Frequency Dependent Selection (FDS) model within the ecological context of the NL hot spring environment (Fig. 4.8). We suspect that the NL hot springs represents a nutrient limited environment for both cellular and viral replication, that cellular replication and virus production rates are low, and that virus infection seldom leads to the rapid killing
of infected hosts. There is both low cell (~$10^{-6}$) and virus particle (~$10^{-5}$) density reducing the probability of interaction between the two. Thus, most cells are likely to be infected by non-lytic viruses, many of which can integrate into the host genome. We also speculate that polar addition of a new CRISPR spacer is rare and is primarily used during the initial response to a new virus within the population. In contrast, movement of preexisting CRISPR spacer blocks via recombination with a plasmid encoded locus provides a mechanism for rapid spread of resistance within the population.

With these assumptions operating, we interpret our results in the following way. When a virus infects a cell that encodes a matching spacer to that virus it can be targeted for inactivation, but that inactivation is not complete. The virus can still replicate in these cells at reduced levels. The virus may also be able to replicate at higher levels within a small subpopulation of cells lacking matching spacer sequences. This virus may also integrate into the host genome. Integration would result in a selection pressure for the host to inactivate its CRISPR/Cas system in order to avoid targeting its own genome. Excision of the CRISPR array and generation of spacer-less CRISPRs may be a mechanism to achieve this.

As a result of creating a spacer-less CRISPR, virus replication would increase in the cell. Cells containing a spacer-less CRISPR locus also become more susceptible to infection by other viruses. Infection by another virus leads to a low level chronic infection of the cell. Movement of a block of spacer sequences from within the population via recombination at the spacer-less CRISPR loci to these chronically infected
cells reduces the virus load. This would result in an indefinite oscillation between spacer-containing and spacer-less CRISPRs, while sometimes driving virus numbers down, but often resulting in a tolerance situation where the virus can be stably maintained at low levels in the population.

At times when the replication pressure from a particular virus type increases, there is a counter acting cellular pressure to distribute the CRISPR spacer block containing the corresponding matching spacer sequence within the population. Over time there would be an increase in the likelihood of the CRISPR spacer sequence population to match the viral community because it provides a mechanism to maintain virus production in a limiting resource environment.

Central to this model is the movement within the population by acquisition of spacer blocks via recombination at the spacer-less CRISPRs. Within the cellular population there is maintained a broad distribution of diverse CRISPR spacer sequences, which represent the population’s past collective memory of the viral populations that have replicated within individual cells. This model holds that spacers will be preferentially acquired as existing spacer blocks via recombination within the population rather than created de novo. An advantage of this model is that it reduces the need to ‘invent’ new spacer sequences, which is an apparently slow process via polar addition at the leader end of a CRISPR locus. Oscillation of host and viral genotypes by slow (low pressure) negative selection frequency pressure occurs to maintain a state of dynamic equilibrium within overall host-virus populations.
In many aspects the appearance of the spacer-less CRISPR loci may be related to the anti-CRISPR system recently proposed for *Escherichia* and *Salmonella* (Touchon et al., 2011; Touchon and Rocha 2010). In these systems, CRISPR loci lacking *cas* genes prevent invasion by genetic elements containing functional CRISPRs/cas located on mobile genetic elements. That is because the CRISPR loci lacking *cas* genes contain spacers with sequences targeting the *cas* genes on the invading mobile genetic elements. By analogy, in our system the system is directed at invading viruses. We have no evidence that the spacers within a block are targeting any sequences other than viruses.

Under our proposed model, the host achieves keeping the virus at low level at a minimum cost of maintaining the CRISPR system. Frequency dependent selection explains the observed pattern at the CRISPR loci. In a nutrient-limited environment with low host population densities, spacer-containing CRISPR loci could be at an advantage because infection leads to chronic infections. Viral genome integration leads to a selective advantage because it favors spacer-less CRISPR loci. Frequency dependent selection would cycle this pattern of an increase and decrease in the frequency of spacer-containing CRISPR loci in the cellular genome.
The primary objective of the research described in this thesis was to significantly contribute to a better understanding of the interactions between archaeal viruses and their host organisms. The specific objectives were to:

1. To develop a genetic system for STIV and probe gene function.
2. To develop a culture-independent approach (metagenomics) for identifying novel viruses and to determine if this approach could be used to access archaeal viruses not accessible using traditional culturing methods.
3. To define how the CRISPR/Cas prokaryotic immune system functions in a natural environment and determine if its function could be correlated to the viral population present in that environment.

Each of these objectives was successfully accomplished.

Virus-host interactions are a major driver of microbial diversity in every environment. Archaeal viruses and hosts are poorly understood, compared to viruses and hosts of Eukarya or Bacteria. This work has sought to increase our understanding of this unusual group of viruses on a number of fronts: in depth study of the life cycle of a cultured isolate (STIV), expansion of our understanding of the total diversity of archaeal viruses by looking beyond cultured isolates, and understanding the specific function of the newly described CRISPR/Cas immune system in a natural environment.
Development of an STIV Genetic System

The research reported in this thesis has described significant advances in our ability to probe the function of STIV with the development of a STIV infectious clone. The development of an STIV genetic system allows one to knock out STIV genes by design and to examine in detail the consequence of targeted gene disruption on STIV replication and the host response to viral infection. Using a targeted gene disruption approach, a number of essential STIV genes have been identified, including the major coat protein (B345), a glycosyl-transferase (A197) and a structural protein thought to be a part of the turret appendage (C381). A non-essential gene, a DNA-binding protein (B116), has also been identified. The development of an infectious clone and associated targeted gene disruption approaches should provide a valuable tool for continued probing of STIV function and host-virus interactions.

In addition to the work described here, the STIV infectious clone/genetic system has been the basis for additional targeted genetic analysis of STIV functions. STIV C92 has recently been knocked out, and it’s role in STIV egress is currently being investigated (Snyder, 2011 and current work). Gene disruptions of STIV F93 (a DNA binding protein), B164 (a putative ATPase), A510 (a putative integrase) and A223 (structural protein thought to be part of the turrets) are currently being constructed. Recent work also includes affinity tagging STIV proteins with 6 x His and FLAG tags in order to perform pull-down assays, to look at interaction partners with viral proteins and to visualize sub-cell localization of STIV proteins by immunolabeling.
A major hurdle to overcome in order to better understand STIV-host interactions is to develop a \textit{Sulfolobus} strain susceptible to STIV that is also genetically tractable. Thus far, the virus has only been shown to replicate in \textit{S. solfataricus} strains P1, \textit{S. solfataricus} P2 and the \textit{S. solfataricus} P2 derivative, \textit{S. solfataricus} 2-2-12, all of which cannot be readily genetically modified, likely because of the instability of their genomes due to large numbers of mobile insertion sequences (Martusewitsch \textit{et al.}, 2000; Rolfsmeier and Blum, 1995). All cellular knockouts and over expression experiments have been performed in either a \textit{S. solfataricus} P1 derivative, strain PH1-16 (Martusewitsch \textit{et al.}, 2000), or a \textit{Sulfolobus} strain 98/2 derivatives (Rolfsmeier and Blum, 1995; Schelert \textit{et al.}, 2004). Thus we cannot simultaneously knockout cellular genes and infect with the virus. This significantly impairs our ability to probe specific interactions between the virus and cellular proteins.

One set of host genes, the ESCRT system genes, is hypothesized to be involved in viral assembly, transport and egress (Snyder and Young, 2011). The ESCRT related protein, VPS4 has been inactivated in \textit{S. solfataricus} PH1-16 by expression of a trans-dominant mutant (Samson \textit{et al.}, 2008). The viral encoded gene, C92, which is hypothesized to interact with the ESCRT system, has also been over expressed in this strain. Over expression of C92 alone in \textit{S. solfataricus} PH1-16 produced pyramid structures similar to those seen during viral infection and associated with viral lysis (Snyder \textit{et al.}, 2011). A genetically tractable and virally susceptible host would allow us to look at the course of a virus infection in a VPS4 knockout host. One might expect that
the virus would replicate and assemble, but not be trafficked to the cell membrane or lyse the cell.

Preliminary work reported in this thesis has focused on development of a *S. solfataricus* host that can both support STIV infection and be genetically modified (Appendix 1). We have undertaken the development a uracil auxotroph of the STIV susceptible host, *S. solfataricus* 2-2-12. Uracil auxotrophy would allow us to utilize a *pyrEF* gene cassette as a selectable marker, both in over expression vectors and in the STIV infectious clone. To this end we have grown *S. solfataricus* 2-2-12 on solid media containing 5-flouro-uracil and uracil supplementation. 5-flouro-uracil is toxic if incorporated into the uracil synthesis pathway and so selects for knockouts or disruptions of the cellular *pyrEF* gene cassette. The high number of insertion elements in the *S. solfataricus* genome means that most auxotrophs found in our screen have an IS insertion in one of these two genes. Unfortunately, the instability of this genome works both in our favor to select for the auxotroph, but also results in frequent reversions back to wild type. Our approach relied on UV induction to increase the mobilization of IS elements in the *S. solfataricus* genome. However, our results indicate that we may have generated more mutations (either IS elements or generalized DNA damage) than could be tolerated. Other potential approaches to this problem are described in Appendix 1.

A second approach, which has not yet been undertaken, is to select for a strain of the *S. solfataricus* PH1-16 that supports STIV replication. This would involve plating
on solid media and testing colonies for STIV production. This approach was successful in developing *S. solfataricus* 2-2-12 from strain P2 (Ortmann, 2008). However with that strain, we began with a cell population that had a minor component that was highly susceptible to STIV, and several rounds of colony purification resulted in section of a highly susceptible strain. In the case of *S. solfataricus* PH1-16, there is no indication that any cells are being infected, and thus this approach is less likely to be successful.

A complete library of gene knockouts for every STIV gene is an ultimate goal. To date, the STIV knockouts performed have been guided by structural, transcriptomic, proteomic or bioinformatic information. This has allowed us to perform targeted mutations of genes of interest. A full catalog of which genes are essential and which are non-essential would require either a brute force approach using the current tools or development of a high-throughput method to look for any viable knockouts. Attempts to develop transposon and recombineering based high throughput methods to generate mutant viruses were both unsuccessful (Appendices 2 and 3). In both cases, only WT viruses or significantly truncated deletion mutants were detected.

Development of a true shuttle vector, which can replicate in both *Sulfolobus* and *E. coli* would be very useful. This would allow us to recover genomes that replicate in *Sulfolobus* and amplify them by transformation into *E.coli*. It would also streamline the current knockout procedure by eliminating the requirement of removing the *E.coli* vector prior to transformation into *Sulfolobus*. This is currently both a time consuming
and challenging step, particularly when DNA yields for such a large construct and transformation efficiency into *Sulfolobus* are both critical issues.

Identification of non-essential genes would allow us to construct a smaller replicon that would be easier to manipulate and propagate in *E.coli*. This would also enable us to insert other genes, such as selectable markers and reporter genes as well as multi-cloning sites for heterologous over expression, without exceeding the packaging limit of the virion. The most widely used current over expression vectors for *Sulfolobus* are based on the virus SSV1. While this tool has been useful in over expression studies, the SSV1 background results in the integration of the plasmid into the host chromosome and the inability to clear the culture of the virus once it is transfected. An STIV based over expression vector would potentially still maintain the self-spreading nature of using a viral vector that overcomes low transformation efficiency in cultures. This virus has not been detected to integrate into the Sulfolobus chromosome, which would eliminate unintended consequences of viral integration.

Establishing Culture-Independent Detection (Metagenomics) of Archaeal Viruses

The field of metagenomics has made significant advances since we first undertook a small viral metagenomics project at Road Side Springs in Yellowstone National Park. Our ability to successfully assemble a previously unknown near full-length archaeal viral-like genome directly from environmental samples, using only a very modest sequencing effort, provided confidence that we could extend a metagenomics
approach to probe cellular and viral communities in other hot springs. This has proven to be quite useful for probing, both over space and time, virus and host population dynamics.

Since this project began, advances have been made in a number of our methods. Improvements in sampling procedures have allowed us to sample much smaller volumes of water than described in Chapter 3 of this work. The procedure now involves spin filtering a total of 1 L hot spring water through a 100 kDa molecular weight cut-off filter instead of filtering hundreds of liters through multiple filtration devices (Chapter 3). Several new generations of amplification technology have emerged since this project began. We now perform the amplification in our own hands using commercially available kits with little evidence of non-specific amplification (Chapter 4). Finally, the improvement in tools for metagenomic sequencing (454 and 454 Titanium, Chapter 4) and assembly (Newbler GS Assembler v2.3) now allow for almost overnight sequencing and assembly of huge data sets. The challenge remains in probing these data sets to answer specific questions.

We have generated time courses of metagenomic data sets of viral and cellular samples spanning several years from two sites: Nymph Lake hot springs #10 (described in chapter 4) and Crater Hills hot springs CHANN041 (“Alice Springs”) (Ortmann, in preparation), along with several RNA viral metagenomic data sets from two other sites: Nymph Lake #17 and #18 (Bolduc, in review). Our lab is currently analyzing a total of 42 independent metagenomes from all of these sites, including viral DNA, viral RNA and
total cellular metagenomes. The depth of this sequencing project is unparalleled to date in any environment. The time course allows us to look at change through time in the viral and cellular populations. In particular, it allows us to address questions such as whether the CRISPR/Cas immune system is having an impact on the viral population at that site (Chapter 4).

The field of metagenomics has now matured beyond the point where labs simply sequence everything in any given environment to “see who’s there.” Now, we are beginning to be able to probe complex interactions involving all members of a population. The amount of data generated in even a relatively simple system such as a hot acidic pool in Yellowstone is enormous, and thus, experimental design and asking specific questions become critical. Integration of metagenomic sequencing with metabolic pathway modeling, single cell genomics and microbial community modeling is allowing us to better understand how complex natural environments function. This work utilized the cellular and viral metagenomes from the Nymph Lake #10 site to probe the function of the CRISPR immune system in this particular environment and has provided us with insights that that would not be apparent by studying this system in a clonal cultured population.

**Examining the CRISPR/Cas System in Natural Environments**

The discovery and description of the CRISPR/Cas immune system in Archaea and Bacteria is one of the more important breakthroughs in microbiology in recent years. It
is also an example of a significant advance in our understanding of biology that has come out of the relatively new field of bioinformatics. The recognition that spacer sequences in the CRISPR arrays were derived from viruses and other extrachromosomal elements opened a new field of study. As this field matures, it is becoming apparent that the CRISPR/Cas immune system picture is not as straightforward as it initially appeared. A number of studies have brought into question the idea that CRISPRs universally function as an immune system (Babu et al., 2011; Touchon et al., 2011).

There are many open questions in the field of CRISPR/Cas immunity research. How cells recognize invading DNA, the mechanism by which new spacers are incorporated, how (and if) the CRISPR system selects specific proto-spacers for incorporation are but a few of the intriguing questions in this area. The exact function of the numerous cas genes is also an area of work still in progress.

The role of viruses in these interactions (evading and potentially inactivating CRISPR/Cas) is an emerging area of study. Clearly, viruses are mutating non-essential sequences to introduce mismatches into proto-spacer (Andersson and Banfield, 2008; Deveau et al., 2008; Heidelberg et al., 2009) and proto-spacer associated motif (PAM) (Semenova et al., 2009) regions. However, another possibility, and one which this particular study points to, is that viruses encode an active anti-CRISPR system that is either shutting down or crippling the cellular CRISPR immunity. Viruses of all domains of life are quite adept at overcoming whatever cellular defenses the host evolves.
Undoubtedly, there are mechanisms by which viruses are evading the CRISPR/Cas system.

Even more intriguing is the possibility that some viruses are actually benefiting from the association with a cellular CRISPR system. Our work shows that some viruses are maintained with exact spacer matches, and that they even obtain some positive advantage when hosts in their environment contain exact spacers to them. This is suggestive of an eukaryotic tolerance type mechanism and may be a mechanism for viruses and cells in complex natural environments with limited resources to tolerate each other rather than expend large amounts of energy in the constant escalation of an arms race.

Another point of CRISPR function that has been called into question by both the results reported in this thesis and others (Touchon and Rocha, 2010) is that the polar addition of spacers at the leader end upon exposure to a novel virus is a common mechanism for change at a CRISPR locus. While this is clearly one mechanism of adaptation, our results demonstrate that this is not the only mechanism by which a cell responds to a changing viral population. Rather, we propose that the cellular population draws from its collective library of existing CRISPR spacer sequences either via selection of members within the population carrying those sequences or by recombination. Exploring how this occurs and whether spacer-less CRISPRs play a role in this process are some of the future directions for the work described in Chapter 4 of this thesis.
This work has significantly contributed to our understanding of Archaeal virology. The development of a genetic system for STIV has opened new areas of research not previously possible for this virus. Exciting discoveries are being made in understanding the life cycle of STIV and its link to viruses from other domains of life (Snyder and Young, 2011). Our pilot metagenomics study at Road Side Springs in Yellowstone paved the way for larger more in-depth viral metagenomics projects throughout Yellowstone. And finally, our characterization of the function of the CRISPR/Cas immune system in a natural environment over time has led to a new model for the function of this system in nutrient limiting environments, which is in contrast to the function described to date in the literature (based on laboratory studies).
REFERENCES CITED
REFERENCES CITED


Grissa I, Vergnaud G, Pourcel C (2007b). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*** **8**.


APPENDIX A

DEVELOPMENT OF A GENETICALLY TRACTABLE STRAIN OF S. SOLFATARICUS STRAIN 2-2-12
Introduction

*Sulfolobus solfataricus* strain 2-2-12, a derivative of *S. so* strain P2 that is highly susceptible to STIV infection, is not a strain for which directed genetic manipulation can be easily performed. The genomes of both *S. so* strain P2 and Sso strain 2-2-12 are highly plastic with numerous mobile genetic elements (insertion sequence, IS). Furthermore, there is no readily available selectable marker available for use in the Sso 2-2-12 strain. There are several expression vectors available for use in other *Sulfolobus* strains that contain a *pyrEF* gene cassette (Jonuscheit, 2003; Martusewitsch *et al.*, 2000). In order to utilize this gene cassette for selection the strain into which the plasmid is transformed must have the *pyrEF* genes inactivated.

The *pyrEF* cassette functions in the uracil biosynthesis pathway. *PyrE* codes for an orotate phospho-ribosyltransferase (*ura5* or *ura10* homologs) while *pyrF* codes for an orotidine-5’-phosphate decarboxylase (*ura3* homolog). Knocking out either of these genes produces a uracil auxotroph that cannot synthesize uracil. 5-FOA is converted to 5-flourouracil by the *pyrEF* gene products. Incorporation of 5-flourouracil into RNA is toxic to the cells. However, if uracil is supplemented to a Δ*pyrEF* strain, the conversion step to the toxic compound is bypassed and cells grow normally (Boeke *et al.*, 1987). Selection of Δ*pyrEF* knockouts can be performed by growth on 5-FOA plates (Vonstein *et al.*, 1995). We constructed a Δ*pyrEF* uracil auxotroph of *S. solfataricus* strain 2-2-12 by selection for uracil auxotrophy.
Methods

We attempted two approaches to obtain a uracil auxotrophic S. solfataricus 2-2-12 strain. The first was adapted from Martusewitsch et al. (Martusewitsch et al., 2000). Cultures were grown to late log (OD$_{650}$ ~0.9) in liquid media (Brock’s 0.4% sucrose, 0.1% tryptone, pH 3.5) and then plated on uracil (10 μg/mL) supplemented 5-FOA (50 μg/mL). The resultant colonies were then screened on Brock’s media plates (pH 3.5), no uracil/5-FOA Brock’s plates, and Brock’s tryptone plates and grown at 75°C for 7 days.

In the second method S. solfataricus 2-2-12 cells were grown to an OD$_{650}$~0.2 in Brock’s media and then UV induced. Cells (2 mL culture) were spread evenly in a thin layer across the Petri plate by swirling the plate and then exposed to UV using a UV stratalinker (50mJ). These cells were then inoculated into 25 mL prewarmed Brock’s media (as above) and incubated at 80°C for 1, 4 or 24 hours. At t=1, 4, and 24 hrs post UV exposure 1 mL and 500 μL of growing cultures were plated onto uracil (10 μg/mL) supplemented 5-FOA (50 μg/mL) Brock’s media plates and grown at 75°C for 7 days. Controls included plating the UV induced cultures on no uracil 5-FOA plates and Brock’s/tryptone plates (no 5-FOA) as well as plating non-UV induced cultures on uracil supplemented 5-FOA plates.

All colonies were picked into liquid cultures of Brock’s/tryptone (0.1%) media and Brock’s no uracil media. Cultures (with uracil supplementation) were grown to an OD$_{650}$ of 0.2 and glycerol stocked at -80°C.
Glycerol stocked colonies were inoculated into 25 mL Brocks media/tryptone, grown to an OD650 ~0.2 and then transformed with pSESD (transformation protocol as described in chapter2). Cells were recovered in Brocks/tryptone for one hour, 24 hrs or 48 hrs (in separate experiments) and then placed on ice for 20 minutes, spun at 6000 x g for 5 minutes and washed with cold no tryptone Brock’s media three times in order to remove the tryptone for passage into uracil-free media. Cells were mock transformed as a control. The original recovery cultures (in Brocks/tryptone media) were also maintained to monitor growth in the absence of selection.

PCR was performed using primers flanking the pyrEF cassette (pyrEFR and pyrEFL) and the iProof polymerase (BioRad). Cycling parameters were as follows: initial denaturation of 98°C for 30 s; 35 cycles of 98°C for 10 s; 57°C for 10 s; 72°C for 30 s followed by a final extension of 72°C for 10 m.

PCR amplicons were cloned into pCRII-Topo-TA vector (Invitrogen) per manufacturer’s protocol with the adaptation of adding A overhangs to the blunt amplicon in the following reaction: 2 μL PCR product, 2 μL 5x GoTaq Buffer, 1 U GoTaq (Promega), 1 μL 20 mM dATPs in a 10 μL reaction at 72°C for 10 m.

Sequencing was performed using the TopoF and TopoR primers designed just upstream and downstream from the multicloning site in the pCRII-Topo-TA vector. Sequences were sent to Nevada Genomics (Reno,NV) for determination.
Results

Growth was not observed on any plates from the first round of selection where cultures were grown to late log prior to plating on 5-FOA plates. Control cultures plated with no selection produced colonies with one week.

Approximately six days after UV induction, colonies were visible on the t=4 and 24 hr recovery plates and on no selection plates. There was no growth on the t=1 hr recovery or the uninduced plates.

Colonies from the t=4 and 24 hr recovery plates grew in liquid culture with uracil supplementation. No growth was visible in the cultures lacking uracil supplementation indicating that the pyrEF cassette was non-functional.

Colony PCR on five colonies produced larger than expected bands for the WT pyrEF gene cassette suggesting that an IS had inserted within this region. Colony #4 had a slightly larger band than colonies 1,2,3 or 5 (figure A.1a). An EcoRI digest of clones (in pCRII-Topo-TA Vector) of PCR products from colonies #2,3 and 4 showed different banding patterns suggesting at least two different insertion events having occurred (figure A.1b).

Sequencing of colony #3 showed a Sso insertion sequence IS1058 in the pyrE gene (SSO0615). While colony #4 had the Sulfolobus IS4 family insertion sequence in the pyrF gene (SSO616). Colonies #3 and #4 were used for transformation with the pSESD shuttle vector (Snyder et al., 2011) which carries a pyrEF cassette. Growth of the colonies out of glycerol stocks was much slower than for controls. Glycerol stocked S.
solfataricus 2-2-12 grew to an OD$_{650}$ ~0.2 within 48 hrs of inoculation. Colonies $S$. $\Delta$pyr$E$F #3 and #4 took up to one week to reach an OD$_{650}$ of 0.2.

Figure A.1. Analysis of F-FOA selected colonies by PCR and restriction digest of pyr$E$F gene cassette. a. PCR with pyr$E$FL and pyr$E$FR primers flanking the pyr$E$F gene cassette. Lanes 1-5. Colonies from 5-FOA selection plates. 6. WT $S$. solfatarius 2-2-12. 7. Negative. b. EcoRI digest of colonies 2,3, and 4.

Very little growth was observed in pSESD transformed $S$. $\Delta$pyr$E$F #3 and #4 cultures in Brock’s tryptone media. After 4 days the cultures reached OD$_{650}$ ~0.15 and were just beginning to be visibly turbid. However, further incubation did not result in an increase in OD$_{650}$ and between days 6-8 the cultures turned brown indicating the cultures had died. Three independent attempts to transform these isolates with pSESD varying the time of recovery prior to washing and passaging into no tryptone media yielded no growth.
Transformed cultures were also plated onto uracil free Brocks media plates at t=24 hrs post transformation. Plates were incubated at 75°C for up to two weeks (until the plates dried out). No colonies were observed on any of these plates.

Discussion

The slow growth of stocked cultures before and after transformation with pSESD suggests that the selected Sso strains were crippled. UV induction is random mutagenesis approach that often leads to large deletions. Our PCR and sequencing only focused on mutations in the pyrEF gene cassette and would not have detected other mutations.

Attempts to grow uninduced *S. solfataricus* 2-2-12 on 5-FOA never produced auxotrophic mutants. No colonies ever grew on the 5-FOA plates. This suggests that the concentration of 5-FOA was too toxic to allow for enough growth for selection to occur, despite the success of others in generating pyrEF auxotrophic mutants in *S. solfataricus* P1. The concentrations of 5-FOA used by Martusewitsch et al. and in our experiments were approximately twice what was determined for *S. acidocaldarius* (Grogan, 1991). Growing uninduced *S. so2-2-12* on dilutions of the concentration of 5-FOA may be a next step in generating specific mutations only in the pyrEF cassette.

Martusewitsch et al. screened a total of 16 isolates of *S. solfataricus* P1 and Sso PH1 that grew in the presence 5-FOA with uracil supplementation. Only two of these grew in the absence of uracil. This suggests that even in the absence of a stress like UV
induction, obtaining stable knockouts of this gene cassette without other unintended mutations may be a rare event requiring a much larger screening effort.

Table A.1. PyrEF Primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyrEFL</td>
<td>GCGGTACCACTTGCGAATAATGCTGC</td>
</tr>
<tr>
<td>PyrEFR</td>
<td>GCAGGTACCTCGTGTAGATTTCCTCCC</td>
</tr>
<tr>
<td>Topo F seq</td>
<td>CGAGCTCGGATCCACTAGTAA</td>
</tr>
<tr>
<td>Topo R seq</td>
<td>CGGAATTGGGCCCTCTA</td>
</tr>
</tbody>
</table>
APPENDIX B

TN5 TRANSPOSON MUTAGENESIS OF STIV
Introduction

In an effort to develop a high throughput method of disrupting each of the ORFs in the STIV genome, we attempted to develop a transposon mutagenesis approach. The advantages to such a method are that transposition events are theoretically random and thus will generate a library of mutants, which can be transformed as a pool into S. so. This allows replication in the native host to select for any viable mutants. Screening then focuses on characterizing the phenotypes of any replicons.

To do this we used the EZ TN5 Transposomics Kit from Epicentre (Madison, WI). The kit provides a TN5 based transposon with an RK6γ origin of replication and a Kan^R gene flanked by 19 bp Mosaic End (ME) sequences on either end. The EZ TN5 transposase recognizes the ME sequence and inserts randomly into any target sequence present in the reaction. The transposase generates a 9 bp staggered cut and inserts the transposon by a transesterification reaction. This product is transformed into E. coli and the E. coli DNA repair system fills in the single stranded gap, resulting in a 9bp direct repeat on either side of the transposon. In the absence of the transposase, the transposon is stable and will not “jump” to a new site. The transposon itself codes for a RK6γ ori and a Kan^R gene, is 2001 bp in length and contains one BamHI site.

We utilized several different target DNAs: pFLSTIV digested with BamHI to remove the pCRII-Topo-TA-vector, native STIV DNA isolated directly from the virus, and the full length STIV clone (which contains a pUC ori, a functional Amp^R gene, a deactivated Kan^R gene and is 21582 bp with two BamHI sites flanking the STIV
Because the transposon would not be removed prior to transformation into \textit{S. solfataricus}, and we suspect that we might exceed the a genome packaging limit for the STIV virion, we also constructed a shorter version of the transposon containing only a Kan$^R$ gene and the ME sequences. The RK6γ ori is not required for replication in \textit{E.coli} when this construct is inserted into the pFLSTIV clone because the pCRII-Topo-TA vector carries a pUC ori.

\section*{Methods}

\textbf{Transposition into STIV}

Native STIV DNA, a digested (with \textit{BamHI}) and religated pFLSTIV (chapter 2) and the intact pFLSTIV clone were used as the target sequence for transposition. Transposition was performed according to manufacturer’s protocol using equimolar ratios of target sequence to transposon as well as 10:1 ratio of target to transposon.

In the case of transposition into native STIV, the DNA was isolated by phenol chloroform extraction and ethanol precipitation (Sambrook \textit{et al.}, 1989). The DNA was microdizalized into ddH$_2$O by spotting 20µL onto a Type VS filter (0.025um) (Millipore) and floating it on ddH$_2$O for approximately 20 minutes. This was used for transposition with the WT TN5 transposon. Control reactions included STIV DNA only and TN5 only. All were plated on Kan LB plates.

In the case of transposition into the pFLSTIV clone with the pCRII-Topo-TA vector removed, the WT TN5 transposon was inserted into the recircularized WT clone (ligated
as described in Chapter 2). This product was then digested with \textit{EagI} (10 \mu L transposition reaction digested in a 20 \mu L reaction), which cuts three times in the pCRII-Topo-TA vector, but not within either the STIV or the TN5 sequence. The digested DNA was then microdialized (as above). Transformations were plated on Amp, Kan and Amp/Kan LB plates.

For transposition into the intact pFLSTIV we used our TN5\textDelta RK6\textgamma ori transposon (described below). Transposition of TN5\textDelta RK6\textgamma ori with no target DNA was performed as a control. Transformations were plated on Amp, Kan and Amp/Kan LB plates.

Transformations into \textit{E. coli} were performed by electroporation (2.5 kV, 25 \mu F, 200 ohms in a 2 mm gap cuvette for a time constant \~5 ms) into Transformax \textit{pir+} ECD100D \textit{E.coli} (Epicentre) and plated onto the appropriate antibiotic LB plates.

\textbf{PCR Screening of TN5 Clones}

PCR using TN5 5’ seq R and various STIV primers (table B.1) was performed to determine whether clones contained STIV sequence. PCR cycling conditions were as follows: initial denaturation of 95°C for five minutes; 35 cycles of 95°C for 30 s; 60°C for 30 s; 72°C for 4 m followed by a final extension of 72°C for 10 m.

\textbf{Construction of Truncated TN5\textDelta RK6\textgamma ori Transposon}

Primers were designed to amplify the Kan\textsuperscript{r} gene and add the 19 bp ME sequences to the ends of the amplicon (table B.1). PCR was performed with the iProof (BioRad) polymerase with the following cycling conditions: initial denaturation of 98°C
for 30 s; 35 cycles of 98°C for 10 s; 57°C for 10 s; 72°C for 30 s followed by a final extension of 72°C for 10 m. The PCR product was excised from the ethidium bromide gel, cloned into pCRII-Topo-TA vector and sequenced. This clone served as the template for reamplification of the truncated TN5 construct. The PCR reaction was then treated with DpnI to remove the clone template DNA and was cleaned up with the Qiaquick PCR clean-up kit (Qiagen) and used for transposition reactions.

**Transformation in S. solfataricus**

The transposed pFLSTIV was prepared for transformation into *S. solfataricus* by digestion with *Bam*HI to remove the *E. coli* vector, as described in chapter 2. The transposon would remain in the STIV genome unless the transposition event occurred in the *E.coli* vector.

**Results**

**Transposition into Native STIV**

Transposition into Native STIV DNA resulted in more colonies on the negative control plates (TN5 only) than on the STIV TN5 plates, indicating that TN5 transposition into itself is a significant background issue in the absence of sufficient or appropriate quality target DNA. In order to check for the presence of STIV sequence in clones obtained from this reaction, PCR using the TN5 5’ seq R primer and three different primers from STIV (designed to different regions within the genome: STIV_MCP_3’R, STIV_B116F and STIV_1865R) in separate PCR reactions on eight colonies was
performed. Each colony was positive for STIV sequence with at least one of the primers, however, several clones produced multiple bands, suggesting that the transposon had inserted more than once into the STIV genome (figure B.1). Significant deletions of regions of the STIV genome were observed. *EcoRI* digests of all of these clones found no full-length STIV clones.

Figure B.1  PCR Screen of TN5 transposon insertion into Native STIV. Lanes 1-8 primer set TN5 5’ seq R and STIV_MCP_3’R, lanes 9-16 primer set TN5 5’ seq R and STIV_B116F, lanes 17-24 primer set TN5 5’ seq R and STIV_1865R. 25. WT STIV with STIV_1865R and STIV_886F, 26. ddH2O with TN5 5’ seq R and STIV_1865R. STIV TN5 clones are in the same order for each primer set.

**Transposition into pFLSTIV with pCRII-Topo-TA Vector Removed**

Transposition into the pre-digested pFLSTIV was plated on Amp, Kan and Amp/Kan plates. Colonies grew on the Kan plates and approximately 10 colonies grew on the Amp plates, but there was no growth on the Amp/Kan plates. This indicated that the Topo vector had been efficiently removed from the STIV sequence, and that the *EagI*
digest eliminated any cases of transposition into the Topo vector. Therefore any Kan\textsuperscript{R} colonies should have been STIV sequence with the TN5 insert. The small amount of Amp background may have been religated Topo vector with no TN5 inserts. However, an EcoRI digest screen of 48 colonies found no full-length STIV clones. This again suggested that transposition events resulted in STIV genome deletions or rearrangements.

**Construction of Truncated TN5ΔRK6γori Transposon**

PCR of the TN5 transposon to amplify the Kan\textsuperscript{R} gene and add ME sequences to the ends gave somewhat ambiguous results with multiple bands present on the gel (figure B.2a). However, cutting out the larger bright band just above the 1018 marker, cloning and sequencing it showed that we had a construct with the 19 bp ME sequences flanking the Kan\textsuperscript{R} gene with approximately 150 bps of intervening sequence (figure B.2b). This clone was then used as PCR template with the same primers to obtain more concentrated stocks of TN5ΔRK6γori construct.

**Transposition of TN5ΔRK6γori into pFLSTIV**

The truncated TN5ΔRK6γori construct (figure B.2b) was transposed into pFLSTIV and plated on Kan and Amp. A control reaction of TN5ΔRK6γori only was also performed. There were no colonies on plates for the control reaction, indicating that the TN5ΔRK6γori does indeed lack the ori and cannot replicate itself as was the case with the WT TN5 transposon provided by the manufacturer. The TN5ΔRK6γori plus STIV
plated on two Kan plates produced a total of eleven colonies. The same reaction plated on Amp produced >100 colonies, but not a lawn. The pFLSTIV construct transformed into *E.coli* with no transposon produced confluent colonies. Viral DNA isolated from 6 of the eleven colonies appeared large enough to potentially be full-length STIV clones. Sequencing these clones using primers designed to sequence across the transposon/STIV junction (TN5_5’_seq_R and TN5_3’_seq_F) found only one case of transposition into STIV with no deletion of the flanking STIV sequence.

Figure B.2. Construction of the TN5ΔRK6yori transposon. A. Gel of PCR amplified TN5ΔRK6yori. B. Schematic of TN5ΔRK6yori construct.

**Transformation into *S.solfolobus* 2-2-12**

The single example of what appeared to be pFLSTIV with a transposon in native STIV sequence (into ORF A61) was transformed into *S.solfataricus* 2-2-12 (as described in Chapter 2), along with WT pFLSTIV as a positive control. QPCR screening of
transformed cultures with primers to the STIV major capsid protein (STIV 5’F and STIV mid R) gave no indication of viral replication.

Discussion

Transposition into native STIV or recircularized pFLSTIV with the Topo vector removed were our first choices for transposon-based mutagenesis approaches because both of these methods should have produced a large pool of mutants that could be transformed into *S. solfataricus*. We would then screen for mutant viral genotypes by looking for altered plaque phenotypes. However, in practice both of these methods were not viable. For native STIV, we hypothesized that the quantity or quality of the target DNA may not have been appropriate. DNA extraction from native virus has always been somewhat problematic, with some evidence that the viral genome is tightly bound to histone-like proteins that are not readily removed, and generally results in low yields of viral DNA. Multiple bands observed in a number of our clones by PCR screen, as well as a significant background of TN5 only colonies both indicates that the target DNA may not have been of sufficient quantity. More than one transposition event per STIV genome may account for the high frequency (100% of clones we screened) of truncated clones, as recombination between two transposons may have cut out the intervening STIV sequence.

The ligated DNA in our circularized pFLSTIV did not appear to be an appropriate transposition target. The reasons for this were not investigated, however, we
speculated that the relaxed circle DNA or the presence of the ligation junction may have affected our results.

The TN5ΔRK6γori was developed in order to reduce the size of the transposon so as to be less likely to exceed a potential STIV DNA packaging limit. The removal of the BamHI site in the transposon also allowed for the removal of the Topo vector after the transposition event. We were able to isolate one apparently full length STIV clone using this strategy. This particular clone was not viable when transformed into S. solfataricus. However this positive transposition events indicates that the general approach, with further optimization, can likely be successfully developed as a random mutagenesis approach.

Further efforts to develop any one of these three strategies may ultimately yield a library of STIV gene disruptions. A better method for native STIV DNA extraction seems the most likely method to generate a pool of knockouts to screen. Overall, TN5 transposon based mutagenesis was not a high throughput straightforward method in our hands to generate a large number of STIV mutants rapidly.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>TN5 no ori F</td>
<td>CTGTCTTTTATACACATCTCACAGGGTGCTCTAAAACTCTCTG</td>
</tr>
<tr>
<td>TN5 no linker R</td>
<td>CTGTCTTTTATACACATCTCGTGAGGTGAGGGATGGTGTATTTG</td>
</tr>
<tr>
<td>TN5_5’_seq_R</td>
<td>CACCTGATTGCCCGACATTATC</td>
</tr>
<tr>
<td>TN5_3’_seq_F</td>
<td>GCAGAGCATTACGCTGACTTTG</td>
</tr>
<tr>
<td>STIV_MCP_3’R</td>
<td>CCAGATGCAGGTACAGATGTT</td>
</tr>
<tr>
<td>STIV_B116F</td>
<td>GCCAAGGGTCTTTGTTACACATTCATG</td>
</tr>
<tr>
<td>STIV_1865R</td>
<td>GCACTCTTTACGTCTAATTTATCTTCGCC</td>
</tr>
</tbody>
</table>
APPENDIX C

HIGH THROUGHPUT MUTAGENESIS USING RECOMBINEERING IN STIV
In addition to random transposition as a method for rapidly generating mutant STIV viral genomes (Appendix 2), we attempted to use the recombineering method. Recombineering is an in vivo directed recombination method that uses mutagenic oligos or PCR amplicons and the *E. coli* phage λ recombination proteins. The strength of this method is that it overcomes difficulties in mutagenizing large DNA molecules which cannot easily be PCR amplified.

The λ phage encodes a set of recombination proteins called “Red” proteins which require approximately 50 bps of homology to direct recombination. The Red system in *E. coli* consists of three proteins, Exo, Gam and Beta. Gam functions to inhibit the *E. coli* RecBCD and SbcCD nuclease which normally degrade linear DNA. The Exo protein is a dsDNA exonuclease which binds dsDNA ends and degrades one strand in a 5’-3’ direction, resulting in the 3’ ssDNA overhang which is required for recombination. The Beta protein is a ssDNA-binding protein which promotes annealing of complementary DNA. Beta can bind ssDNA longer than 35 nucleotides. Recombineering with dsDNA (i.e. PCR amplicons) requires the function of all three of these proteins, while recombineering with a ssDNA oligo only requires the Beta protein. The inappropriate expression of Exo and Beta can lead to unwanted rearrangements, while prolonged expression of Gam can lead to plasmid instability (Sawitzke, 2007).

*E. coli* strains harboring defective λ prophage with the lysis, structural and DNA replication genes deleted has been developed for recombineering. In this work we used
a strain modified from *E. coli* DH10B (λCI857(cro-bioA)-Tet<sup>+</sup>), SW102. The Red genes are expressed under the control of λ phage *pL* operon and a temperature sensitive repressor, CI857 (<37°C there is no expression). However, researchers report rare subpopulations of spontaneously induced cells (Sawitzke, 2007). Cells harboring the target DNA and mutagenic oligos (or PCR product) are induced by briefly raising the temperature to 42°C, at which point the CI857 protein is denatured, the *pL* operon is transcribed, and recombination can take place. Cells are then cooled to <37°C, shutting off transcription of the Red proteins and ideally avoiding off-target recombination (Sawitzke, 2007).

**Methods**

We utilized a Recombineering kit provided by the Biological Resources Branch of the National Cancer Institute designed for use with BAC clones. Our methods were adapted from protocols provided with this kit.

**Preparation of Cells for Transformation**

SW102 cells were streaked onto a LB plate (no antibiotic). A colony was inoculated into a 2 mL overnight culture of LB or LB/Amp (for cells pre-transformed with pFLSTIV, see below). At OD<sub>650</sub>~0.4 500 μL was transferred into 25 mL LB or LB/Amp (for cells previously transformed with pFLSTIV, see below) in a 50 mL Erlenmeyer flask. This was grown at 32°C for ~3 hours to an OD<sub>650</sub>~0.6. Cells were chilled on ice for ~2 m and
pelleted (6000 x g for 5 m). The supernatant was removed, cells were washed twice in ice-cold ddH₂O and resuspended in ~50 μL ddH₂O.

**Transformation of pFLSTIV and Oligos into SW102 Cells**

The pFLSTIV clone and ssDNA oligos were transformed into SW102 cells by electroporation (2.5kV, 200 ohms, 25μF, 2 mm gap cuvette, time constant ~ 5 ms). Cells were recovered in NZY media for 1 hr, plated on LB/Amp and grown overnight at 32°C.

We performed these transformations two different ways. First, pFLSTIV was transformed into SW102 cells, cultures were recovered per standard protocol except all growth conditions were at 32°C. Cells were plated onto LB/Amp plates, grown overnight at 32°C and then colonies were inoculated into LB/Amp and prepared as described above. Prepared cells harboring the pFLSTIV were then induced (described below) and transformed with ssDNA oligos. As an alternative method we induced the cells and then co-transformed pFLSTIV and the ssDNA oligos. In both cases we transformed ~5 ng pFLSTIV and 200 ng dsDNA oligos.

Cells were recovered for 1 hour in NZY media (no antibiotic) and then plated onto appropriate plates (LB/Amp for Xmal and Pfol recombineering and LB/Kan for Ncol recombineering).

**STIV Recombineering Targets**

We designed oligos to target three different unique restriction sites within the STIV genome. In each case the oligo was approximately 100 bps in length with the
mutagenized bps near the middle of the oligo. One oligo targeted the \textit{N}col site that had been filled in a Kan$^R$ gene of the pCRII-Topo-TA vector during construction of the pFLSTIV clone (Chapter 2). Altering this region to reconstitute the \textit{N}col site back to the WT pCRII-Topo-TA sequence provided us with selection as this reactivated the Kan$^R$ gene on the plasmid. The other two oligos spanned the unique \textit{Pfo} site in C557 and the \textit{XmaI} site in A510. As a control we also performed a mock recombineering experiment in the absence of a ssDNA oligo.

\textbf{Heat Shock Induction}

Once cells reached an OD$_{650}$~0.6 10 mL were transferred to a 50mL Erlenmeyer flask and incubated at 42°C for 15 minutes. They were then immediately placed on ice and swirled for 2-5 minutes and then prepared for transformation with either a ssDNA oligo alone (for transformation into cells already harboring pFLSTIV) or both a ssDNA oligo and pFLSTIV. A control culture was retained at 32°C throughout and then prepared for transformation.

\textbf{Screening Recombineered Clones}

We screened the recombineering experiments by colony PCR with primers flanking the restriction sites that were either removed or introduced. The product was digested with the appropriate restriction endonuclease (5\mu L PCR product, 1 \mu L 10X Buffer, 1 U enzyme in a 10 \mu L reaction at 37°C for 1 h) and visualized on an ethidium bromide gel.
Significant background issues in the experiments with no selection led to a number of efforts to reduce non-recombined background. Because our recombineering oligos functioned to abolish existing restriction sites in STIV we used these restriction enzymes after the recombination experiment to remove any WT background. In order to do this, cells were recovered post-transformation in 1 mL LB (no antibiotic) for 1 h at 32°C. Then 1 mL LB with 100 μg/mL Amp was added for a final concentration of Amp of 50 μg/mL. This was grown overnight at 32°C. Plasmid DNA was isolated from this culture using the PureYield Plasmid MiniPrep kit (Promega). Approximately 50 ng of the purified plasmid was digested with the appropriate enzyme in a 10 μL reaction for 1 h at 37°C. This was then transformed into XL2 Blue cells (Stratagene) per manufacturer’s protocol and plated on LB/Amp plates. As a control, WT pFLSTIV was digested with each of the above restriction endonucleases.

We also performed double digests of the plasmid with either XmaI or PfoI and SacI. There is a unique SacI site within the STIV genome. Therefore, if the desired recombination event took place a full-length STIV genome linear molecule would result but WT STIV genomes would be cut twice.

Results

Recombination using the oligo to restore the NcoI site in the vector KanR gene was successful. Recombined colonies grew on both Amp and Kan plates, however,
colony numbers were very low on the Kan plates (5 colonies) as compared to Amp (>100), indicating that this was a rare event.

The recombination experiments without selection did not result in the isolation of the correct clones. There was evidence that recombination was occurring, however, not as intended. We observed small plasmids in our DNA prep of recombineered colonies. All digests of PCR amplicons spanning targeted restriction sites indicated that the products were all WT.

We were not successful in sufficiently reducing the background by digestion of the WT plasmid after recombination. We hypothesized that a linear plasmid may have been able to recircularize and was still replicating. However, our results indicated that it was more likely that we were simply not cutting all of the DNA. Incomplete digestion leaving even a small amount of intact supercoiled plasmid would preferentially transform and make finding rare recombination events unlikely. Dilutions of the plasmid DNA reduced total numbers of colonies, but did not change the ratio of WT to recombined colonies.

Discussion

Sawitzke, et. al. describe the frequency for recombination events using a ssDNA oligo to engineer a point mutation as $10^5$ per $10^8$ viable cells. The amount of time cultures are allowed to recover prior to plating is also critical. If cultures are plated early, the number of colonies to be screened will be reduced, but the parental copies
may also still be present in the colony requiring further purification, and in our case, potentially confounding the results. We relied on a PCR screen to find colonies lacking a particular restriction site. In every case we observed digestion of PCR products. If the parental copies were present at much higher concentrations than the recombined copies, our PCR screen may never detect the recombined copies.

If a longer outgrowth period is used, each colony will be relatively pure, but the frequency of recombined colonies will be reduced approximately 4-16 fold. This would theoretically only produce $6.3 \times 10^3 - 2.5 \times 10^4$ recombination events per $10^8$ viable cells, requiring a larger screening effort than we undertook.

Oligos for recombineering can be designed to either the leading strand or the lagging strand, however, recombineering efficiency can be nearly 30 fold higher with oligos to the lagging strand. We did not take this into account in our design. Attention to this detail might improve efficiency.

The *E. coli* Mismatch Repair system (MMR) targets 1-3 bp mismatches between strands of DNA and repairs them. This system can interfere with recombineering strategies engineering short mutations. For this reason we tried longer regions of mismatch with our second round of recombineering. However, we did not detect any recombineered *Xma*I or *Pfo*I sites in these experiments.

The small size of plasmids observed in digests of recombineered clones was confounding. One explanation may be that pFLSTIV is not stable when manipulated. The Topo vector is a high copy number vector with a pUC ori and was not designed for
such a large insert. Similar truncations were observed in our TN5 transposon based mutagenesis experiments (Appendix B). Both lines of experimentation might benefit from cloning STIV into a low copy number plasmid designed for large inserts.

Table C.1. Recombineering Primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Xma</em> recomb</td>
<td>GAGTCTATTTTGTAGGAAAAATAATGCCCTGGGATGTAGTATTTGAGTTATCAATTCCAAAATGGC</td>
</tr>
<tr>
<td><em>Pfo</em> recomb</td>
<td>GACGGGGACTTTTGGAGCATCGCATTCGAGGAGAACATATGGGGATATAGTAGGATATAGGCATAATT</td>
</tr>
<tr>
<td><em>Xma</em> 5bp change</td>
<td>GAAAAATTTTTAAACTCTTAGTTGAGTATTTTGAGGAAAAATATGCCCTGGGATGTAGTATTTTGGAG</td>
</tr>
<tr>
<td><em>Pfo</em> 5bp change</td>
<td>CTGCAACCATCATACCCGATGACGGGACTTTTGGAGCATCCGAGAATATGGGGAGTAAGATTAGTAT</td>
</tr>
<tr>
<td><em>Nco</em> recomb</td>
<td>CGCCAGGCTAGGCGAGCGATCGCTGCTGGAGGAGGAGGACGCATTGGGATGCGATGCTGCTGGAGGAAATG</td>
</tr>
<tr>
<td></td>
<td>ATCCGGAATATGCTGCTGGAGGAGGACGCATTGGGATGCGATGCTGCTGGAGGAGGACGCATTGGGATG</td>
</tr>
</tbody>
</table>
APPENDIX D

A197 COMPLEMENTATION
Introduction

Classic genetics approaches to demonstrating gene function involve either disrupting or knocking out genes followed by complementing the function with a heterologous expressed protein product. We attempted to do this for the A197 open reading frame in STIV. The A197 ORF had been genetically altered in pFLSTIV in several ways (Chapter 2, Wirth et al., 2011). We deleted the ORF almost in its entirety, leaving only 20 bps of the gene (KO_A197STIV), we disrupted the ORF with a digestion and fill reaction that created a frame-shift and inserted a premature stop codon, and we made three point mutations (D151E, D151N, and D151H) in what is thought to be the active site of the enzyme (Larson, 2006).

Methods

Construction of the A197 Gene Disruption and Deletion Mutants

The construction of the A197 mutant viral constructs is described in Chapter 2.

Construction of the A197 Expression Vector

The A197 gene was amplified from WT STIV DNA using A197_Ndel_F and A197_SalI_R adding an Ndel site upstream from the start of the gene and a SalI site to the downstream end of the gene. The Sulfolobus expression vector, pSeSD (provided by Qunxin She, University of Copenhagen) was also digested with Ndel and SalI. The two products were then ligated producing a shuttle vector with A197 cloned downstream.
from the *araS* promoter and with a 6X his tag on the C-terminus (figure D.1).

Sequencing with C92seqF and C92seqR (which were designed to the pSESD vector) confirmed that the A197 gene was correct and in-frame with the 6X His tag.

**A197 Complementation in *S.solfataricus* 2-2-12**

Complementation of both ΔA197STIV (gene disruption) and KO_A197STIV (gene deletion) were performed by co-transformation of 300 ng of the mutant virus genome (prepared as previously described) along with 1µg of pSESD/A197. Transformations were performed as described for the pFLSTIV transformations (Chapter 2). Protein expression was induced at 24hrs post transformation with 1mL of 10% Arabinose (for a final concentration of 0.4%). One mL of culture was pulled every 24 hours for qPCR and western beginning with t=0 at transformation. Cultures were followed for 72 hours post transformation.

**Dot Blots for Detection of Mutant Virus and A197 Over Expression**

Dot blots were performed by spotting 1µL of sample onto a Nylon membrane (Millipore Immobilon NY+), blocking with 5% powdered milk in TBS/Tween (0.1%), incubating with the appropriate antibody in TBS/Tween (0.1%), and detecting with the Opti-4CN Colorimetric detection kit (BioRad) using an anti-rabbit secondary antibody linked to a Horse-radish peroxidase enzyme. The α-B345 (major capsid protein) antibody was used to detect the presence of virus in transformed cultures. The α-A197 polyclonal antibody was used to detect the over expression of A197.
QPCR for Detection of Viral Genomes in Culture

QPCR using the STIV_5’F and STIV_midR primers (Chapter 2) designed to the B345 ORF was used to detect viral genomes in transformed cultures.

PCR for Detection of pSESD in Transformed Cultures

PCR using the C92seqF and C92seqR primers was performed to determine if the pSESD plasmid was being maintained in culture over time.

Results

In our first complementation experiment the gene disruption mutant of ΔA197STIV was co-transformed with pSESD/A197. The pSESD plasmid alone, ΔA197STIV alone and no DNA were transformed as controls. These cultures were not induced for over expression of A197. PCR to detect the pSESD plasmid showed that it was maintained over the course of the experiment. We did not develop a qPCR assay for this primer set and so we did not quantitate the levels of plasmid in the cultures. However, visualizing the PCR products on an ethidium bromide gel did not give any indication of decreasing concentration over time (figure D.1a).

QPCR of cultures transformed with both ΔA197STIV and pSESD/A197 indicated that the viral genome was replicating. Two of the triplicate transformations of ΔA197STIV and pSESD/A197 produced viral genomes at levels comparable to WT virus or WT pFLSTIV (Chapter 2), the third culture showed no evidence of viral replication.
The no complementation control of ΔA197STIV was negative by qPCR (as in Chapter 2), as was PSESD alone (figure D.1b).

Figure D.1. ΔA197STIV complemented with pSESD/A197. A. PCR for the presence of pSESD in transformed cultures: lanes 1-3, t=0 ΔA197STIV and pSESD/A197 replicates, 4. t=0 pSESD, 5-7 t=24 ΔA197STIV and pSESD/A197, 8. t=24 pSESD only, 9-11. t=48 ΔA197STIV and pSESD/A197, 12. t=48 pSESD only, 13. No template control. B. QPCR results of ΔA197STIV and pSESD/A197. Red, green and dark blue lines are ΔA197STIV and pSESD/A197 replicates 1,2 and 3, yellow line is mock transformation, light blue is pSESD only.

A dot blot to detect the major capsid protein of STIV indicated that viral particles were being produced (figure D.2a) in the cultures that were positive by qPCR. A dot blot to detect the A197 protein indicated that the protein was present in the cultures that were qPCR and B345 positive (figure D.2b).

PCR spanning the A197 mutation site in STIV using primers outside of the A197 ORF (STIV_SacII_screen_F & STIV_SacII_screen_R) on the ΔA197STIV and pSESD/A197 transformed cultures showed that the ΔA197STIV construct had a WT A197 gene (figure
Figure D.2. Dot blots of qPCR positive ∆A197STIV and pSESD/A197 cultures. a. Detection of B345 (major capsid protein). 1-2 t= 24 ∆A197STIV and pSESD/A197; 3. t=24 ∆A197STIV; 4-5 t=48 ∆A197STIV and pSESD/A197; 6. t=48 ∆A197STIV; 7-8. t=72 ∆A197STIV and pSESD/A197; 9. t=72 ∆A197STIV; 10. Purified STIV; 11. Purified B345 protein; 12. t=48 Mock transformed culture. b. Detection of a197. 1-2 t=24 ∆A197STIV and pSESD/A197; 3. t=24 ∆A197STIV; 4. t=24 pSESD 5-6. t=48 ∆A197STIV and pSESD/A197; 7. t=48 ∆A197STIV; 8. t=48 pSESD; 9-10. t=72 ∆A197STIV and pSESD/A197; 11. t=72 ∆A197STIV; 12. t=72 pSESD; 13. Purified WT STIV; 14. t=48 Mock transformed culture.

D.3a). PCR to double check the input material indicated that prior to transformation the ∆A197STIV construct had a disrupted A197 gene (figure D.3b).
In order to test complementation of the lethal ΔA197 phenotype a true deletion mutant of the viral gene was constructed (KO_A197STIV). This was to remove any possibility of recombination of a disrupted, but otherwise intact, A197 open reading frame with the WT A197 gene on the over expression vector. Transformation of KO_A197STIV into S. solfataricus (as described in Chapter 2) did not result in production of viral genomes or particles. Transformation of KO_A197STIV and pSESD/A197, as described above also did not produce detectable levels of viral replication.

**Discussion**

The co-transformation of a WT A197 gene and a disrupted gene apparently allowed for recombination in vivo to reconstitute the WT viral genome. The results from our complementation experiment with the gene disrupted ΔA197STIV and pSESD/A197
suggested that A197 expression was not being directed from an over expression vector. We did not have purified A197 to use as a standard on the dot blot and therefore did not test for sensitivity of the αA197 antibody. However, the levels detected in WT infected cultures were similar to what was seen in our complemented cultures.

To overcome recombination and demonstrate complementation we performed the near complete deletion of A197 in pFLSTIV (KO_A197STIV). Co-transformation of this construct with pSESD was performed. PCR over the time course only detected A197 on the pSESD plasmid, no product was produced using the STIV_SaclI_screen_F & STIV_SaclI_screen_R primers, indicating no replication of the KO_A197STIV, because the screening primers should have produced a 345bp product spanning the deletion site of A197.

The reasons for the failure of the complementation of KO_A197STIV with pSESD/A197 have not been fully investigated. It is possible that the deletion of A197 produced polar effects in the viral genome, potentially disrupting other essential genes. Our construct left 20 nts from the A197 ORF in the viral genome. Future efforts could include construction of a deletion mutant that ensures neighboring ORFs are not impacted.

These complementation experiments were performed in S. solfataricus strain 2-2-12. This strain is not a uracil auxotroph and therefore selection for the pSESD plasmid was not performed. It is possible that complementation would not have been possible in this strain as successful viral replication may have represented sufficient negative
selection pressure for the cells to clear the pSESD plasmid in the absence of positive selection.

Table D.1. A197 Complementation Primers

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A197_AgeI_screen_F</td>
<td>CAAATTCTAGGGTGTTGGAATAATG</td>
</tr>
<tr>
<td>A197_AgeI_screen_R</td>
<td>CGTTTATAAGCAATCCTAAAAACG</td>
</tr>
<tr>
<td>A197_Knockout_AgeI_F</td>
<td>GCCCACCAGGCTAGAATAAAAGTAATACTCATG</td>
</tr>
<tr>
<td>A197_Knockout_AgeI_R</td>
<td>GCATACCGTGCTTCTATATAGCTCAC</td>
</tr>
<tr>
<td>A197_NdeI_F</td>
<td>GATCTCATATGAGAAACGCTTTTTTTTTTATACTTCG</td>
</tr>
<tr>
<td>A197_SalI_R</td>
<td>GCAATATTTTCTCGACCTTTTTTTCTAGCTACGTGATC</td>
</tr>
<tr>
<td>STIV_SacII_screen_F</td>
<td>GGGTATGATGTGTGTTTGTGTC</td>
</tr>
<tr>
<td>STIV_SacII_screen_R</td>
<td>CTATCAGCTTGTCTCCCG</td>
</tr>
<tr>
<td>STIV_XmaI_F</td>
<td>CTCTTAGTTAGAGTCTATTTTGTGAGG</td>
</tr>
<tr>
<td>STIV_AvrII_R</td>
<td>CCTAATTCCACTAGGTAGTAG</td>
</tr>
<tr>
<td>C92_Seq_F</td>
<td>ACCCGAGAATGAGGTGAGAAGC</td>
</tr>
<tr>
<td>C92_Seq_R</td>
<td>CAGTAGGTACCCGGGAAAAA</td>
</tr>
</tbody>
</table>