12-7-2007

The Family Fuselloviridae : Diversity and Replication of a hyperthermic virus infecting the archaeon genus Sulfolobus

Adam Joseph Clore
Portland State University

THE FAMILY FUSELLOVIRIDAE: DIVERSITY AND REPLICATION OF A HYPERTHERMIC VIRUS INFECTING THE ARCHAEON GENUS *SULFOLOBUS*

by

ADAM JOSEPH CLORE

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

DOCTOR OF PHILOSOPHY
in
BIOLOGY

Portland State University
2008

DISSERTATION APPROVAL

The abstract and dissertation of Adam Joseph Clore for the Doctor of Philosophy in Biology were presented December 7, 2007, and accepted by the dissertation committee and the doctoral program.

COMMITTEE APPROVALS:

[redacted]

Kenneth Stedman, Chair

[redacted]

Michael Bartlett

[redacted]

Justin Courcelle

[redacted]

Niles Lehman

[redacted]

Dirk Iwata-Reuyl
Representative of the Office of Graduate Studies

DOCTORAL PROGRAM APPROVAL: [redacted]

Luis Ruedas, Director
Biology Ph.D. Program

ABSTRACT

An abstract of the dissertation of Adam Joseph Clore for the Doctor of Philosophy in Biology, presented December 7, 2007.

Title: The Family Fuselloviridae: Diversity and Replication of a Hyperthermic Virus Infecting the Archaeon Genus *Sulfolobus*

The virus family Fuselloviridae infects the hyperthermophilic and acidophilic Crenarchaeon genus *Sulfolobus* and has been isolated from terrestrial hotsprings worldwide. Two previously uncharacterized Fusselloviruses, SSV-I3 and SSV-L1, were isolated and sequenced and are compared to the five fully sequenced viruses presently in the public databases. Conserved promoters in all viruses and similar putative origins of replication suggest that these viruses use a transcriptional and genomic replication program similar to the relatively well-characterized SSV1. Pairwise comparisons of conserved genes in the seven virus genomes show that, like its host *Sulfolobus*, these viruses' genetic divergence correlates with geographic separation. Genome rearrangements, horizontal movement of genes between Fuselloviruses, other Crenarchaeal viruses, and other hosts are also discussed.

The development of a novel gene knockout system (LIPCR) for these viruses is presented with detailed methods. Use of this knockout system is demonstrated with two viral vectors that have fully and partially deleted integrase genes. The complete integrase deletion does not prevent virus replication but appears to prevent integration of the virus into the host genome and appears to decrease the relative fitness of the virus compared to a virus with a complete integrase gene. The partial integrase deletion removes the catalytic residues demonstrated to be necessary for enzymatic function but leaves the attachment site located within the integrase gene. Interestingly, this mutant appears to be still capable of integration in our lab host, *Sulfolobus solfataricus*. Possible reasons for this are discussed.

Acknowledgements

I would like to acknowledge the members of my committee: Dr. Michael Bartlett, Dr. Justin Courcelle, Dr. Niles Lehman, Dr. Dirk Iwata-Reuyl, and Dr. David Boone. I would like to thank my advisor, Dr. Kenneth Stedman, who trained me as his first PhD student, and managed to do it superbly without any practice.

I would also like to thank several members of the Stedman Lab and the fifth floor: Melissa DeYoung, James Laidler, and Michael Micorescu for their advice, support, and a constant open ear, as well as Melissa Kendall-Morris, Random Diessner, and Johanna Rigas.

A special thanks to the following: Jane Boone for infinite help with TAing, Susan Masta for all things phylogenetic. Chunfei Li for his help with electron microscopy, Dennis Prosen for advice and generously providing Phusion® polymerase, as well as Mark Nisenfeld and Leroy Laush, for maintaining, troubleshooting, and fixing just about everything.

I also gratefully acknowledge the National Science Foundation, Portland State University, and the Northwest Health Foundation. Without the financial support of these institutions none of this research would have been possible.

Most importantly, I would like to thank my wife, Cara, who constantly supported me in my graduate career through the long hours, low pay, and consistent underestimation of the time I would spend in the lab each day.

Table of Contents

List of Tables:

List of Figures:

Chapter 1: Introduction

Preface

Viruses are thought to be ubiquitous to all forms life, and it is estimated that there is nearly an order of magnitude more viruses than cellular organisms on Earth, totaling up to 10^{31} (91). Viruses are defined as obligate intracellular parasites, and as such they have no metabolism, and no means of independent replication. Therefore, they rely solely on cellular organisms for their reproduction (35). They take from their hosts nearly everything needed to replicate; often hijacking the cells' regulatory pathways do so (71). The means by which viruses take the resources they require to survive are in many instances very meager. HIV and most Lentivirus genomes contain around 9 kilobases of nucleic acids, yet within this small amount of genetic material and a small number of viral proteins they are able to effect efficient replication in mammals *via* reverse transcription, evasion of the host immune response, and production of all the necessary capsid proteins (71).

Examples of even simpler replication strategies are observed. The Hepatitis B virus contains a DNA genome just 3.2 kilobase pair in length that codes for 4 genes that allow for complete viral replication. Many other examples of extremely compact genomes in viruses exist (35). In nearly all viruses compact genomes are the rule, as are overlapping genes and a lack of intergenic regions compared to cellular organisms (35).

Because viruses depend on hosts for their replication, much of their gene regulation and control of biosynthetic pathways are similar, if not identical, to their hosts (71). This similarity, along with the relative simplicity of viruses, has made them ideal models to study the more complex workings of the cell. Early examples of work with phage T4 in *E. coli* led to the elucidation of the nature of genetic code, the discovery of messenger RNA, and have helped provide a great deal of understanding about how cells transcribe and translate the genome's information (50). The use of viruses as tools of molecular biology continues today, such as viral-based constructs that are being used to study gene therapy (29), and viral vectors are used to study protein interactions in cells (53).

The effect viruses have on their host is generally thought to be negative, and certainly to our individual experience this seems true. The effects viruses have at the population level may not be so (119). Viruses often target the most successful organisms where high population density provides ideal situations for efficient spread. This "killing the winner" is thought to even the odds of less competitive organisms and may help to control runaway populations that would otherwise drive the less fit into extinction (114). This would in turn keep diversity high and allow populations to respond more robustly to change. Viruses' horizontal movement of both their own genetic material and the occasional mis-packaged host nucleic acid allows for a horizontal spread of genes in a way that may otherwise never happen (22, 34, 39, 55).

Overall, viruses make a unique way of living, contribute to diversity and robustness of life, and have undoubtedly helped shaped life into what it is today. Viruses also provide us with unique models to study their more complex hosts. The aim of this research is to understand the site-specific integration in the virus SSV1 and the role it has on the replication of the virus, as well as the relationship and evolutionary history, this virus family shares. Understanding this viruses' replication, evolution and spread will give us a better understanding of not just the family Fuselloviridae but its archaeal host, *Sulfolobus*, and the ecology of thermal springs.

The Archaea

The Archaea are often referred to as the third domain of life, due to their late discovery with respect to Eukaryotes and Bacteria. The Archaea inhabit many of the most hostile habitats of the Earth. They are growing slowly in the cold depths of the ocean floor, in the dry lakes of Antarctica, in the hot acidic pools of thermal springs, and innumerable environments between these two extremes (21). Our initial perception of the Archaea was that they were confined to the most extreme niches of the earth and represented an ancient or "archaeal" type of life (28, 61). However environmental sequencing has shown the Archaea, particularly the mesophilic marine Crenarchaea, may represent one of the most numerous cell types in the world (51), and likely contribute a great deal to the fixation of carbon and the cycling of nutrients in the world's oceans (47). This suggests that our biosphere, and in consequence our lives, are heavily influenced by the Archaea.

When Carl Woese constructed the "universal" tree of life based on small subunit rDNA molecules he discovered that the Archaea is a clade of prokaryotes distinct from Bacteria (120). The group consists of a large number of diverse, single celled prokaryotic organisms that use a wide array of metabolism strategies and substrates in all environments known to support life (19). The Archaea are distinct from the Bacteria in many of their cellular functions and makeup, including DNA replication machinery, translation machinery, and the lipid content of their membranes (19).

When archaeal transcription initiation was first investigated, clear homologues to the DNA-dependant RNA polymerase II (125), TATA binding protein, and TFIIB (45) found in Eukaryotic transcription were seen in Archaea. This supports the hypothesis that Archaeal transcription is a simplified model of the Eukaryotic system (13). Archaeal transcription also has elements similar to bacterial transcription as well as completely unique elements (5). Examples of bacterial-like transcription mechanisms are the extensive use of bacterial-like repressors in transcriptional regulation and the commonality of polycistronic transcripts (12). Unique to Archaeal transcription are several transcriptional regulators that have no clear homologues (5).

Membranes of Archaea are distinctly different from that of Bacteria, being mainly composed of lipids with ether linkages to the glycerol heads, as opposed to the ester linkages found in most Bacteria. The pathways and enzymes used in the biosynthesis of these two types of lipids are, for the most part, unrelated as well (116).

The Domain Archaea and Sulfolobus' Place Within It

Within the Archaea there are several phyla, two of the most extensively represented and studied are the Euryarchaea and the Crenarchaea (19). Euryarchaea are composed of eight classes, which include the well-studied Methanogens and Halophiles. The Crenarchaea contains one class containing well-characterized hyperthermophiles, such as *Sulfolobus,* as well as a vast group of relatively poorly characterized mesophilic marine organisms (21).

Two other phyla, the Korarchaeota and the Nanoarchaeota, as well as a proposed phylum, the Ancient Archaeal Group, are present within the Archaea. The Koryarchaeota and the Ancient Archaeal Group are known only from environmental sequencing and are found in high temperature environments (10). The Nanoarchaeota is currently composed of one sequenced organism, *Nanoarchaeum equitans*, a small extracellular organism thought to be a parasite of the Crenarchaeon *Ignicoccus islandicus.* With a genome size of 490,885 base pairs *N. equitans* is the smallest archaeal genome sequenced to date (112).

Within the Crenarchaea and its single class, Thermoprotei, there are several orders including the order Sulfolobales. The Sulfolobales contains several genera, all of which are thermopiles with species that can be found in terrestrial sulfur springs (19). The genus *Sulfolobus* consists of many species that inhabit environments with temperatures above 70° C and pH below 3, and are commonly found in terrestrial sulfur springs (19). One of the most extensively studied species in this genus is *Sulfolobus solfataricus,* an aerobic hyperthermophilic and acidophilic organism capable of living facultatively as a chemolithoautotroph by the oxidation of H_2S and S^0 to H_2SO_4 or as chemoorganohetorotroph by the oxidation of a large variety of complex organic compounds (19). *S. solfataricus* strain P2 provides a good model organism because of its ability to grow as a heterotroph, and because it was one of the first Crenarchaea to have a complete genome sequence available, followed afterwards by another closely related species, *Sulfolobus tokodaii* (52, 100).

Table 1.1: Viruses of the Sulfolobales.

Family, genus	Genome*	Size	Shape	Integrase	Lysis
Fuselloviridae, fusellovirus (117)	ccc	14.8-17.4	spindle	Yes	No
Bicaudaviridae, bicaudavirus (82)	ccc	62.7	spindle w/tails	Yes	Yes
Ampullaviridae, ampullavirus (43)	ln	23.9	bottle	No	No
Guttaviridae, guttavirus (8)	ccc	~20	droplet	No	No
Rudiviridae, rudivirus (80)	ln	24.7-35.5	rod	No	No
Lipothrixviridae, α-lipothrixvirus (72)	ln	15.9	rod	No	No
Lipothrixviridae, β-lipothrixvirus (9)	ln	40.9	rod	No	No
Lipothrixviridae, γ-lipothrixvirus (15)	ln	21.1	rod	No	No
Lipothrixviridae, δ-lipothrixvirus (44)	ln	31.8	rod	No	No
Globuloviridae, globulovirus (42)	ln		drop	No	No
Unclassified viruses					
Sulfolobus tengchongensis spindle-shaped virus 1 (122)	ccc	75.3	spindle	No	No
Sulfolobus turreted icosahedral virus (89)	ln	17.7	ico**	No	No

*ccc = covalently closed circular DNA. ln = linear DNA. ** ico = icosahedral

Viruses of Sulfolobales

A surprising amount of morphological and genetic diversity is observed in the viruses infecting members of the kingdom *Crenarchaea,* mainly in the well-studied genus *Sulfolobus*. Currently described are seven families with ten genera as well as several unclassified viruses. All of these viruses have double stranded DNA genomes that lack RNA replication intermediates (81). All but one family, *Bicaudaviridae,* exits the cell by budding rather than cell lysis (82). Of the viruses with circular genomes, two are known to integrate themselves into the host via a bacterial-like integrase in the tyrosine recombinase family (Table 1.1).

The most common types of viruses and virus like particles found in enrichment cultures from solfataric hot springs are rod shaped (84, 88) and spindle shaped viruses (6). Unlike the rod shaped viruses commonly found in plants, these contain DNA genomes (82), and they show no similarity in coat protein genes to plant viruses. Based on these data the similarity to plant virus morphology is most likely the result of convergent evolution rather than a common ancestor.

SSV1 and the family Fuselloviridae

The family Fuselloviridae, commonly referred to as the SSV or Spindle Shaped Viruses, is the best characterized virus family to date within the Crenarchaea. The Sulfolobus Spindle-Shaped Virus 1 (SSV1), named SAV in the original manuscript, (68) is the type species and was isolated from *S. shibatae* strain B-12 that was isolated from a hot spring in Beppu, Japan. The SSV1 virus particle is approximately 60x90 nm with a 10 nm tail protruding from one end (68, 87) and is composed of at least three structural proteins and 15,465 base pairs of covalently closed circular, double stranded DNA (75). Like all other Fuselloviruses identified this species infects and replicates in *S.*

solfataricus and other closely related *Sulfolobus* strains but not in the more distant relative *Sulfolobus acidocaldarius* (95, 117).

The family Fuselloviridae is unique and confined to the Sulfolobales as hosts. Viruses with similar spindle shapes are commonly seen in hypersaline waters (30). However the particles seen in hypersaline waters are smaller than SSVs and seem to lack the small tail present on Fusellovirus capsids. Sequencing and characterization of two virus isolates from euryarchaeal *Haloarchaea* species from the Pink Lakes northwest of Victoria, Australia showed these viruses to be unique. Both isolates have ORFs and gene arrangements completely different from Fuselloviruses, linear genomes, and a virus capsid composed of unrelated structural proteins. Furthermore these viruses use different replication strategies, including the absence of integration (11). Based on these traits these viruses are named Salterprovirus-like particles and not placed in the family Fuselloviridae (30). Recently a spindle shaped virus-like particle, PAV1 was isolated from the euryarchaeal hyperthermophile *Pyrococcus abyssi* and its genome was sequenced. Again this virus was found to have no relation to Fuselloviruses (40).

Fuselloviruses, like other Sulfolobales viruses show few negative effects on the cell. They appear to exit the cells by budding and most hosts show no or little change in cellular growth rates upon infection in liquid media, and only a small amount of growth inhibition when grown on plates (37, 78, 81).

Despite extensive study, including crystallization of two proteins encoded by SSV1's 34 open reading frames, the functions of only four of these ORFs are known. These are the two viral coat proteins, VP1 and VP3, conserved in all Fuselloviruses, a DNA binding protein unique to SSV1 and also observed in the virion, VP2, and the integrase gene, discussed in more detail below (85-87, 117).

Putative functions have been suggested for several other genes. ORF B-251 is thought to be a possible copy number regulator based on weak similarity to the Bacterial DnaA protein, which functions in the regulation of gene expression and genome replication by unwinding local DNA in an ATP dependent manner (56). Two crystal structures exist of small proteins encoded by the SSV1 genome. F-93 forms a homodimer with a winged-helix motif similar to many DNA binding proteins (58). ORF D-63 forms a homodimer of monomers containing a two-helix motif. Based on the conserved surface residues seen in D-63, the SSV-K1 homologue F-61, and the SSV2 homologue D-57 Kraft et al. propose that this molecule may function as an adapter in the binding and assembling of macromolecules (57).

Based on the presence of turbid plaques seen when the virus infects lawns of *S. solfataricus,* lack of visible cell lysis in liquid culture or TEM, and by direct observation, the virus appears to bud from its host rather than exiting by cell lysis (95). Our observations show that virus production in the laboratory host *S. solfataricus* is higher than in its natural host *S. shibatae*. However this titer may be affected by other cell stresses such as oxidative damage and temperature change during growth in the lab, and a one base pair mismatch in the lab host's integration site, all of which could cause induction of the virus as described below.

UV induction of SSV1

SSV1 virus integrates into its host genome in a site-specific manner, and virus production can be induced by UV irradiation (87). Currently SSV1 is the only Fusellovirus shown to be effected by UV irradiation. Induction increases viral production in the natural host *S. shibatae* from a basal level of 10^2 to 10^3 plaque forming units/L in chronically infected cultures that were at an OD of 0.3 and in log phase of growth. Plaque forming units increase after a four hour eclipse to a maximum titer of 10^7 plaque-forming units/L 12 to 16 hours later when cultures are in late logarithmic growth phase and an OD_{600} of 0.6-0.7 (68). As most research with SSV1 is done in the well-characterized *S. solfataricus* rather than *S. shibatae,* rates of virus production during UV induction were measured as well. Results show similar but higher basal rate of viral production in chronically infected cultures and a higher titer of virus after UV induction, producing up to 10^9 plaque-forming units /L (95).

Evidence suggests stressors other than UV light also induce the virus. Mitomycin C effectively induces SSV1 (87), as does growth of Fusellovirus infected cultures on plates. When infected cultures are grown on plates a decreased growth rate compared to uninfected cells is observed, resulting in the presence of turbid halos caused by lower cell densities surrounding points of viral infection. Reduction in the growth rate of infected cells in liquid cultures is not observed except when viral production is induced (37). It is possible that the increased oxygen tension that cells grown on plates are exposed to is responsible for this difference in virus production (68).

Satellite Viruses

Two satellite viruses related to Fuselloviruses have been discovered and characterized in *Sulfolobus* strains. The first, pSSVx, was found together with SSV2 (7). This genetic element is double stranded circular DNA 5705 bp in size and contains 9 ORFs. pSSVx has a high degree of similarity in a cluster of ORFs and sequence features to pRN plasmids. The similarities include three genes conserved in pRN plasmids, a large ORF that is homologous to the family E type DNA polymerase identified in pRN1 (64), and found in other pRN plasmids, an ORF similar to the *copG* gene controlling plasmid copy number in pRN1 and pRN2 (65), and a putative plasmid regulation gene *plrA* (41). In addition pSSVx shares sequence similarity to putative single and double stranded origins of replication in pRN1 and pRN2 (7).

Based on the similarity of this element to the pRN family of non-conjugative Sulfolobus plasmids pSSVx is thought to replicate as a non-conjugative plasmid in the absence of a Fusellovirus. However Arnold et al. were not able to separate pSSVx from SSV2 in culture (7). The genetic element can only spread with a co-infection of a Fusellovirus. The virus strains SSV1 or SSV2 were both shown to be capable of allowing pSSVx to spread with the latter being a more efficient helper.

pSSVx also contains homologues of two genes conserved in all sequenced Fuselloviruses, a putative DnaA type copy number regulation gene, B-251 in SSV1, and a ORF of unknown function with weak similarities to several archaeal DNA gyrase-like genes, ORF A-153 in SSV1. The suggested function of the B-251 homologue as a copy number regulator was supported by a study of pSSVx transcription where it was shown the transcription of the B-251 homologue is proportional to pSSVx genome copy number in the cell (25). Based on the inability of the pRN plasmids to package into a capsid it was proposed that either or both of the ORFs homologous to the Fusellovirus ORFs plays a role in packaging of DNA into the viral capsid (7). However pSSVi (see below) contains no homologues to either of these genes and appears to package its DNA as well. Based on the similarities between pSSVx and the pRN family of plasmids and Fuselloviruses this genetic element is thought to be a plasmid/virus hybrid (7).

The second satellite virus was isolated very recently by others from our lab strain of *S. solfataricus* strain P2 and is called pSSVi. This element was discovered when SSV2 was transformed into *S. solfataricus*, producing extrachromosomal copies of both SSV2 and the smaller pSSVi (111). It is not known how this genetic element came to be in the Stedman lab strain, as other P2 strains from the DMSZ do not harbor pSSVi (111). pSSVi is a double stranded circular DNA 5740 bp in size and contains 8 ORFs. pSSVi shows less similarity to the pRN plasmids than pSSVx, having moderate homology only to the *copG* gene. The genetic element shares a homologous integrase gene with the Fuselloviridae, which seems to be most similar to the integrase in SSV1 and may be able to allow integration at SSV1 *att* sites (see Chapter 4). Like all integrases in the family Fuselloviridae this gene is partitioned upon integration by an internal attachment site. Based on the similarities to the pRN family of plasmids and Fuselloviruses this genetic element is also thought to be a plasmid/virus hybrid (111).

Both of these satellite viruses produce SSV-like particles upon co-infection with Fuselloviruses, however the particles of the satellite viruses are smaller, perhaps due to the smaller genome packaged (7, 111). Interestingly, pSSVi was observed to up-regulate viral production upon coinfection with SSV-I2, decreasing the rate of growth of infected cultures and producing higher virus titers (111). These two hybrids of virus and plasmid are an example of the movement of genes between *Sulfolobus* extrachromosomal elements and suggest that there is a close relationship between the viruses and plasmids in this genus.

Chapter 2: Sequences of the Novel Viruses SSV-L1 and SSV-I3 and Comparison to Other Sequenced Fuselloviruses

Abstract

This chapter describes the isolation, genome sequencing, and analysis of the Fusellovirus SSV-L1, from Lassen Volcanic National Park in the USA, and the analysis of the genome sequence of the Fusellovirus SSV-I3, isolated from Iceland, also sequenced in the Stedman lab. Both viruses are compared to each other and to five other published Fusellovirus genomes.

All 7 viruses have a similar genome composed of two distinct halves. One genome half contains conserved ORFs and core promoter sequences expressed late in the virus replication cycle, the other half lacks conservation of ORFs and core promoter sequences, and contains and a concentration of short repeat sequences. Unlike previously published data on Fuselloviruses, these comparisons show a clear correlation between virus sequence divergence and geographic distance separating the locations of virus isolation. Indirect evidence indicates that the origin of replication is in the non-conserved half of the genome.

Introduction

The family Fuselloviridae is the best-characterized virus family to date within the Crenarchaea, however little is known about how these viruses replicate, spread or their relationship to their host. The Sulfolobus Spindle-

Shaped Virus 1 (SSV1) is the type species and was isolated from *S. shibatae* strain B-12. *S. shibatae* was isolated from a hot spring in Beppu, Japan (68). The SSV1 virus particle is approximately 60x90 nm with a 10 nm long tail protruding from one end (68, 87). It is composed of at least three structural proteins and 15,465 base pairs of covalently closed circular, double stranded DNA (75). Like all other Fuselloviruses identified, this species infects and replicates in *S. solfataricus* and other closely related *Sulfolobus* strains but not in the more distantly related species such as *Sulfolobus acidocaldarius* (95, 117).

Four Fusellovirus genomes besides SSV1 are currently in the public database. SSV-I2[1] (106), and SSV-I4 (77), both isolated from Iceland in 1996 SSV-K1 isolated from the Kronotsky-Uzon Reserve in Kamchatka, Russia in 2000 (117), and SSV-RH1 isolated from the Norris Geyser Basin of Yellowstone National Park in the USA also in 2000 (117).

Previously Analyzed Fusellovirus

In 2004 the Fuselloviruses SSV1, SSV-I2, SSV-RH1 and SSV-K1 were compared to address the open reading frame (ORF) conservation and possible origins of this virus family (117). Eighteen ORFs, conserved in sequence and ORF order, were identified in one half of each genome of the four viruses. Based on the conserved 18 ORFs a common ancestor of these

[1] For updated nomenclature see Appendix 2

viruses was suggested (117). The other half of the genomes lack this conservation and primarily consist of ORFs that are not universally conserved in Fuselloviruses.

None of the ORFs in the non-conserved region have known function. Within the conserved half of the genome, the functions of only four ORFs are known. One of these is the integrase gene, and is the only gene to that appears to be orthologous to genes outside of the Fuselloviridae family (85-87, 117). Three other genes whose function is known code for proteins found in the virus capsid, the universally conserved viral coat proteins VP1 and VP3, and VP2, a DNA binding protein unique to SSV1 (83-85).

The Integrase Gene

The integrase protein is responsible for site-specific integration of the virus genome into tRNA genes in its host. Using Southern hybridization and PCR, integration of SSV1 was observed to integrate into a single arginyl tRNA gene in its hosts *S. shibatae* and *S. solfataricus* (95). Integration sites of SSV-RH and SSV-K1 in *S. solfataricus* were identified by PCR of integrated proviruses. SSV-RH1 integrates into one of five leucyl tRNA genes upon infection of *S. solfataricus* (117) while SSV-K1 integrates into at least three different tRNAs with similar sequences and a non-tRNA locus in *S. solfataricus* (117).

SSV1 Transcription

SSVI is the only Fusellovirus for which transcription has been studied. Promoters of these transcripts were the first archaeal promoters shown to contain the canonical TATA-box (87). Originally a total of 10 transcripts, named T1-T9 and a UV inducible T-ind transcript (Figure 2.1) were observed upon UV irradiation of latently infected *S. shibatae* cultures. These transcripts include all of the ORFs, with many of the transcripts being polycistronic and partially overlapping. Transcripts T1 and T2 start from the same promoter but terminate at different locations. Similarly, transcripts T4, T7, and T8 have the same promoter but terminate differently. The mechanisms of differential termination in Archaea are not known (87, 94).

Microarrays have recently been used to measure gene expression of SSV1 in latently infected *S. solfataricus* cultures after UV irradiation (37, 87). These microarray data supports Reiter's original observations and also identify a small monocistronic transcript named Tx that transcribes C124, thought to be the last ORF in T3 by Reiter *et al.* (see Figure 2.1) (37). Both studies measured transcription of latently infected cells after UV induction. The latter gives more detailed temporal data. Almost immediately after UV irradiation transcription of T-ind begins. Within four hours, 2 early transcripts, T5 and T6, are actively transcribed from promoters within 200 bases of the T-Ind transcript and extending away in both directions. By 8.5 hours post-induction all transcripts are up-regulated including transcripts that encode the coat protein genes and a putative copy regulator gene (37).

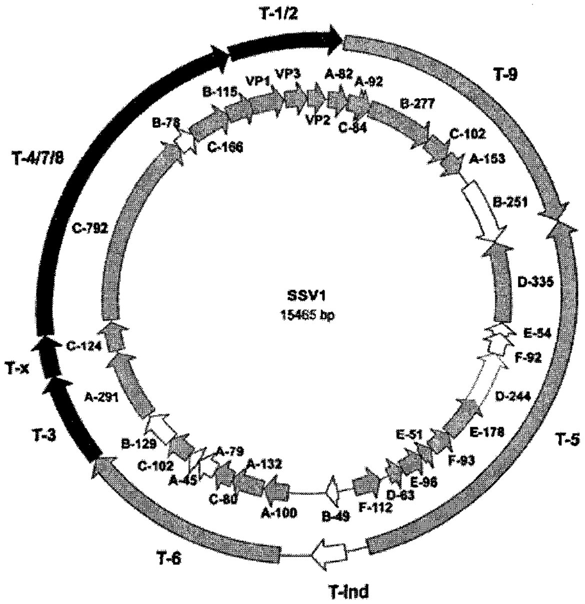

Figure 2.1: SSV1 UV induced transcripts and open reading frames. Outer circle shows transcripts as arrows. Lighter filled arrows represent early transcripts, darker filled arrows represent late transcripts. Grey filled arrows in the inner circle of ORFs represent constitutively expressed genes whose transcription was detected before UV irradiation. Adapted from (37) and (87)

Origin of Replication

Reiter, Frols and others have hypothesized that the origin of replication is located between the T5 and T6 transcripts based on the observations that both transcripts originate in this area and extend away in opposite directions, the presence of repeat sequences (87), and on unpublished data cited in (37).

However, as no direct evidence identifying an origin of replication in Fuselloviruses has been published, and no Fusellovirus sequence is similar to known origins of replication in Sulfolobus or its plasmids exists, the origin of replication in Fuselloviruses remains unknown.

Phylogeography in Fuselloviridae

In microbiology the Bass-Becking hypothesis that "everything is everywhere, the environment selects" is a commonly held concept for mesophilic microbes. This hypothesis appears to hold in many cases, such with the distribution of marine Crenarchea in the world's oceans (69), and the distribution of soil bacteria throughout the northern hemisphere (90). Hypothetically, great numbers and rapid growth of microbes, and possibly genetic recombination within these microbes allows for a saturation of dispersion resulting in nearly homogeneous diversity throughout the world (90).

Sulfolobus species appear to be unlike many mesophilic microbes in terms of the structure of their distribution. Multi-locus sequence typing of *Sulfolobus* strains isolated in thermal areas separated by distances ranging from single meters to thousands of kilometers show that sequence divergence is highly correlated to geographic distance (115). This difference is attributed, in part, to the barrier of "uninhabitable" area separating thermal springs from each other, restricting gene flow and allowing for genetic drift of isolated communities (115).

This observation of genetic differences in *Sulfolobus*, apparently due to geographic separation, is in contrast to PCR-amplified sequence data of Fuselloviruses from environmental samples from three hot springs in Yellowstone National Park (103). These results suggest that Fusellovirus populations in these springs change rapidly over space and time, and that individual hot springs contain comparable Fusellovirus sequence diversity to that seen between hot springs throughout the world. Therefore, there appears to be no correlation between sequence diversity and geographic separation within the Fuselloviruses(103). Why these viruses appear to lack similar pattern of distribution as the hosts is currently unknown.

Scope of Research

This chapter describes and analyzes the newly sequenced virus genomes from our lab, SSV-L1 and SSV-I3, and compares them to the five Fusellovirus genomes in the public database in terms of ORF content and conservation, promoter and origin of replication identification, and integration.

Finally, a phylogeographical comparison of these 7 Fusellovirus genomes from around the world contrasts the partial Fusellovirus genome data from Yellowstone National Park presented by Snyder in (103). This analysis requires a reassessment of the question of whether virus diversity changes with respect to geographic distance as is seen with the virus host.

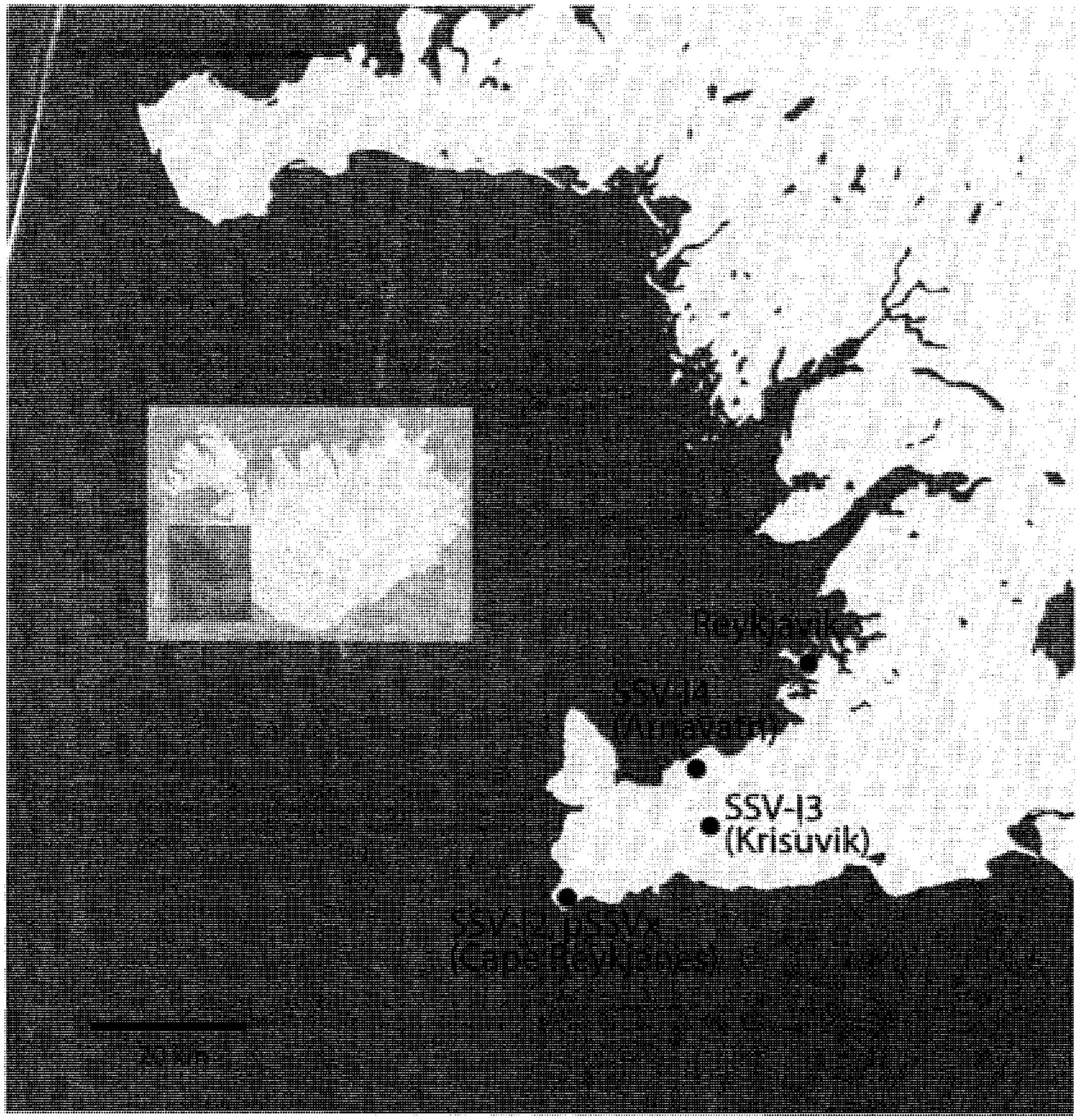

Figure 2.2: Location of isolation of Icelandic Fuselloviruses. Adopted from (6). Map modified from www.openstreetmap.org and is subject to the creative commons licensing agreement version 2.0 (www.creativecommons.org).

Results

The Virus SSV-I3

The virus SSV-I3 was originally isolated from a thermal spring in the Krisuvik solfatara in Southwestern Iceland in the summer of 1996 by W. Zillig, D. Prangishvili and I. Holz (Figure 2.1). The Krisuvik hot spring was above 90°C and below pH 4 (6). SSV-I3 was isolated in the same summer and approximately 20 kilometers from that of SSV-I2 and pSSVx, which were isolated from the Reykjanes thermal area. In even closer proximity is SSV-I4, isolated from Arnavatn approximately 7 km distant (see Figure 2.2) (6).

SSV-I3 has a genome composed of 15,230 base pairs of covalently closed circular DNA containing 32 ORFs. Among these are 15 ORFs observed in all Fusellovirus genomes. SSV-I3 is unique among all archaeal viruses in that it is the first fully sequenced Archaeal virus for which all of its ORFs are similar to previously known ORFs, most of which are most similar to SSV-I4's. Even excluding the highly similar to the ORFs in SSV-I4 all but two of SSV-I3 ORFs have homologues in the other Fuselloviruses (Table 2.2).

The Virus SSV-L1

The virus SSV-L1 was isolated from a thermal sulfur spring in the Devils Kitchen thermal area of Lassen Volcanic National Park in the summer of 2005. It has a genome composed of 14,461 base pairs of covalently closed circular DNA making it the smallest Fusellovirus genome sequenced to date. The

SSV-L1 genome contains 31 ORFs including 15 ORFs universally conserved in Fuselloviruses.

Conserved ORFs within the family Fuselloviradae

Previously, 18 conserved Fusellovirus ORFs have been described in SSV1, SSV-I2, SSV-K1 and SSV-RH1 (117). Conservation of 15 only of the 18 ORFs are observed with the addition of the 3 new viruses SSV-L1, SSV-I3 and SSV-I4 (Figure 2.3, Table 2.2). The 3 ORFs of the original 18 that are not completely conserved are homologues of the SSV1 ORFs A100, A79 and C80. All are encoded in the T6 transcript seperate from the 15 conserved ORFs (Figure 2.3).

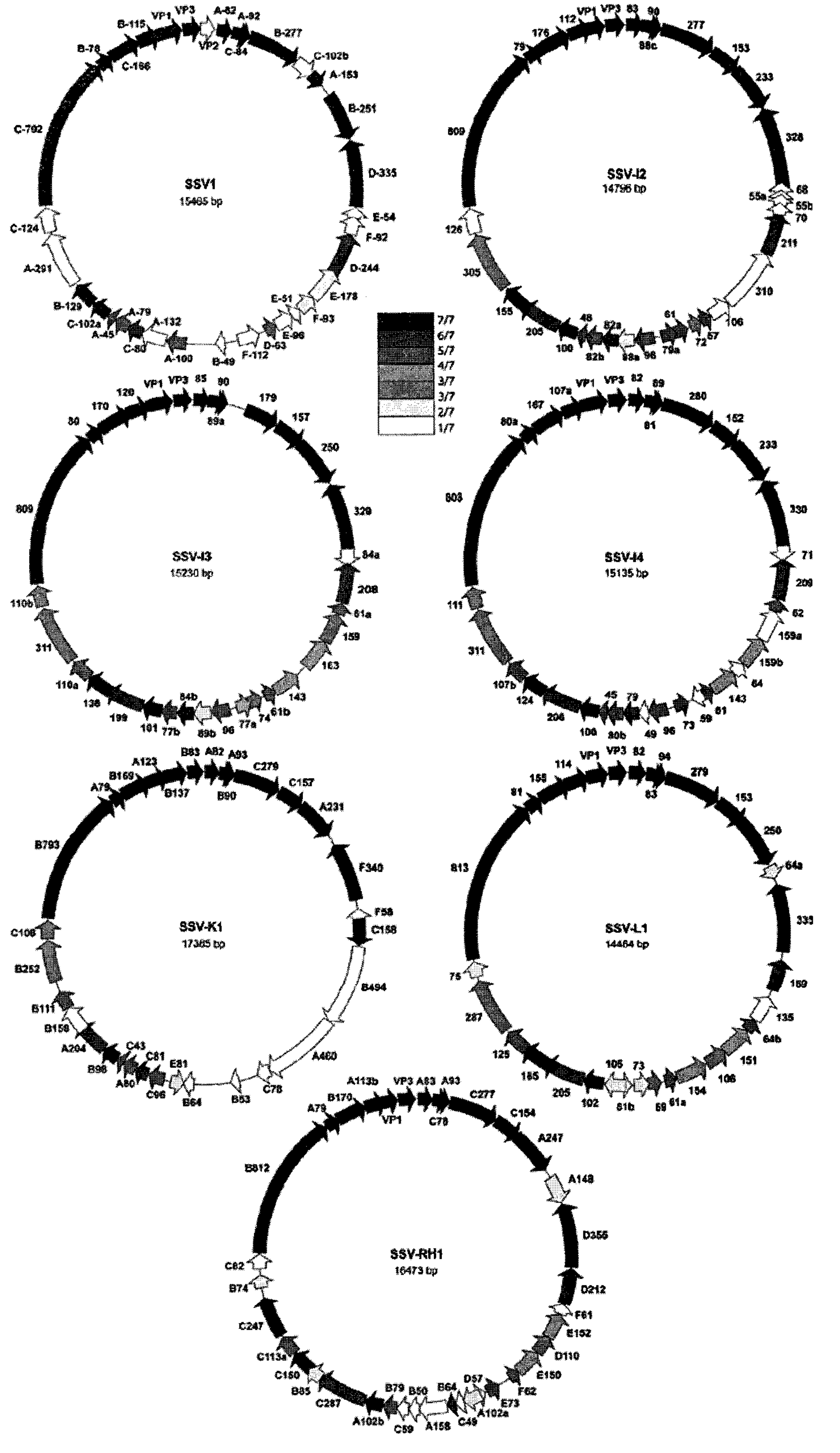
SSV1
15465 bp
B-115
VP1
VP3
A-82
A-92
VP2
C-84
B-277
C-102b
A-153
B-251
D-335
E-54
F-92
D-244
E-178
F-93
E-51
E-96
D-63
F-112
B-49
A-100
A-132
C-80
A-79
A-45
C-102a
B-129
A-291
C-124
C-792
B-78
C-166
SSV-I2
14796 bp
112
VP1
VP3
83
90
88c
277
153
233
328
68
55a
55b
70
211
310
106
57
72
61
79a
96
88a
82a
82b
48
100
205
155
305
126
809
79
176
7/7
6/7
5/7
4/7
3/7
3/7
2/7
1/7
SSV-I3
15230 bp
120
VP1
VP3
85
90
89a
179
157
250
329
84a
208
61a
159
163
143
61b
74
77a
96
89b
84b
77b
101
199
138
110a
311
110b
809
80
170
SSV-I4
15135 bp
107a
VP1
VP3
82
89
81
280
162
233
330
71
209
62
159a
159b
64
143
61
59
73
96
49
79
80b
45
106
206
124
107b
311
111
803
80a
167
SSV-K1
17385 bp
A123
B83
A82
A93
B137
B90
C279
C157
A231
F340
F58
C158
B494
A460
C78
B53
B64
E81
C81
C96
A80
C43
B98
A204
B156
B111
B252
C106
B793
A79
B169
SSV-L1
14464 bp
114
VP1
VP3
82
94
83
279
153
250
64a
339
169
135
64b
151
106
154
61a
59
73
61b
105
102
205
185
125
287
75
813
81
155
SSV-RH1
16473 bp
A113b
VP3
A83
A93
B170
VP1
C78
C277
C154
A247
A148
D355
D212
F61
E152
D110
E150
F62
E73
A102a
D57
B64
C49
A158
B50
C59
B79
A102b
C287
B85
C150
C113a
C247
B74
C62
B812
A79

Figure 2.3: Genome conservation in all Fuselloviruses (Previous page) Genome maps of all known Fusellovirus genomes aligned so that the end of the VP3 gene is at the top of each map. ORFs are labled as arrows. Arrows are filled with colors indicating conservation of ORFs . ORFs conserved in all seven genomes are Black, six genomes Blue, five genomes Purple, four genomes Green, three genomes Red, two genomes Yellow, and one genome White as described by the table in the Figure. ORFs are labled as annotated in the original publications (75, 106, 117) or herein. SSV-I4 sequence data from Genbank accession # EU030938.

Pair-wise Identity of Fusellovirus Genomes

SSV-I3 and SSV-I4 are isolates from hot springs that are geographically close to each other (7 kilometers) as well as being close to where SSV-I2 was isolated (21 and 24 kilometers respectively) (6) (Figure 2.1). To begin to assess the overall similarity of the Fuselloviruses to each other, the total nucleic acid sequence of all viruses was aligned and the pair-wise percent identity determined (Table 2.1). The two most similar sequences are SSV-I3 and SSV-I4 with 84% overall nucleotide identity. Similarities of SSV-I3 and SSV-I4 to SSV-I2 are also quite high, having 69 and 73% identity respectively. SSV-L1 and SSV-RH1 are 77% identical despite the geographic separation of 1000 km, much greater than that of the Icelandic viruses. All other viruses show between 51% and 61% identity to each other (Table 2.1).

Table 2.1: Pair-wise percent nucleotide identities of Fuselloviruses

	SSV1	SSV-I2	SSV-I3	SSV-I4	SSV-K1	SSV-L1	SSV-RH1
SSV1	100						
SSV-I2	52	100					
SSV-I3	52	69	100				
SSV-I4	51	73	84	100			
SSV-K1	52	55	58	58	100		
SSV-L1	51	61	61	61	55	100	
SSV-RH1	51	59	58	58	53	77	100

SSV-I3 and SSV-I4 ORFs

The majority of ORFs in SSV-I3 are most similar to ORFs in SSV-I4, including four SSV-I3 ORFs, 96, 110a, 311, and 110b, which share 100% nucleotide identity. The last 3 of these ORFs are adjacent in both virus genomes and are part of an area of over 1800 base pairs of contiguous identical sequence, that begins and ends in ORFs universally conserved in Fusellovirus genomes (Figure 2.4).

Contrasting with regions of 100% sequence identity in SSV-I3 and SSV-I4, less conserved sequence identity is seen in the universally conserved ORFs (Figure 2.4). The universally conserved SSV-I3 ORF 250 shares 86% amino acid identity with SSV1 ORF B251, followed by 75% identity to SSV-L1 ORF 250, and is only 48% and 47% identical to its geographically close relatives SSV-I2 ORF 233 and SSV-I4 ORF 233 respectively (Figure 2.5).

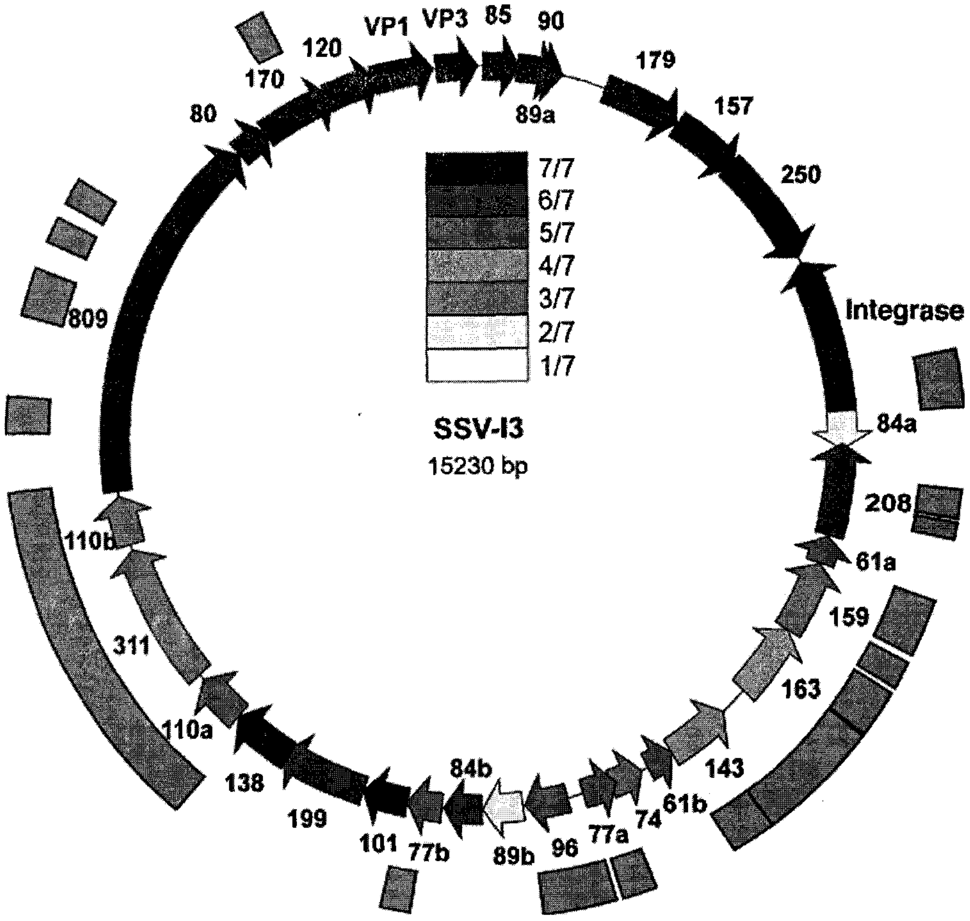

Figure 2.4: SSV-I3 genome sequence identities to the SSV-I4 genome. On the outside the red ring indicate identical nucleotide regions spanning more than 100 base pairs. Coloring of ORFs indicate conservation between the seven Fusellovirus genomes. Labels on inner circle are SSV-I3 ORF names.

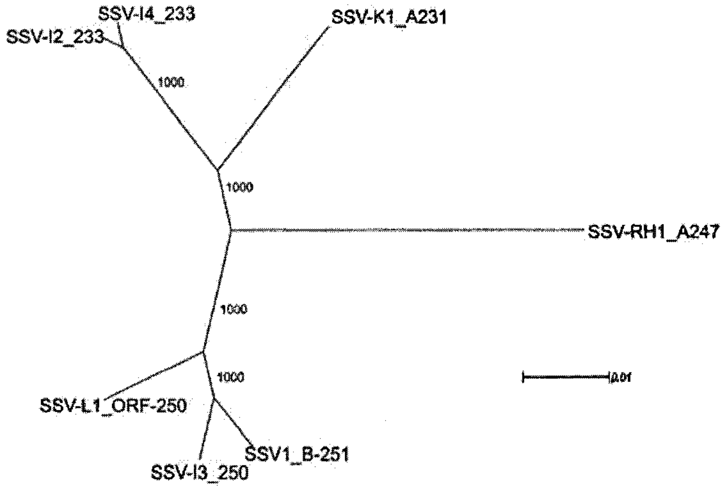

Figure 2.5: Neighbor-joining tree of universally conserved SSV-I3 ORF 250 and its homologues. Numbers represent bootstrap values out of 1000 replicates. Scale bar equals 0.01 substitutions per site.

SSV-L1 and SSV-RH1

The second most similar Fusellovirus genomes are those of SSV-L1 and SSV-RH1 (77% identity), the only 2 sequenced Fuselloviruses from North America. The greatest similarities between the SSV-L1 and SSV-RH1 genomes are located in the region lacking the universally conserved ORFs. Five of the 31 SSV-L1 ORFs are present only in SSV-L1 and SSV-RH1 (Figure 2.6). Of the 15 universally conserved ORFs in SSV-L1 only two share the highest identity to SSV-RH1 ORFs, the rest share higher identity to more distantly isolated Fuselloviruses including the SSV-I3 ORF 250 (Figure 2.5).

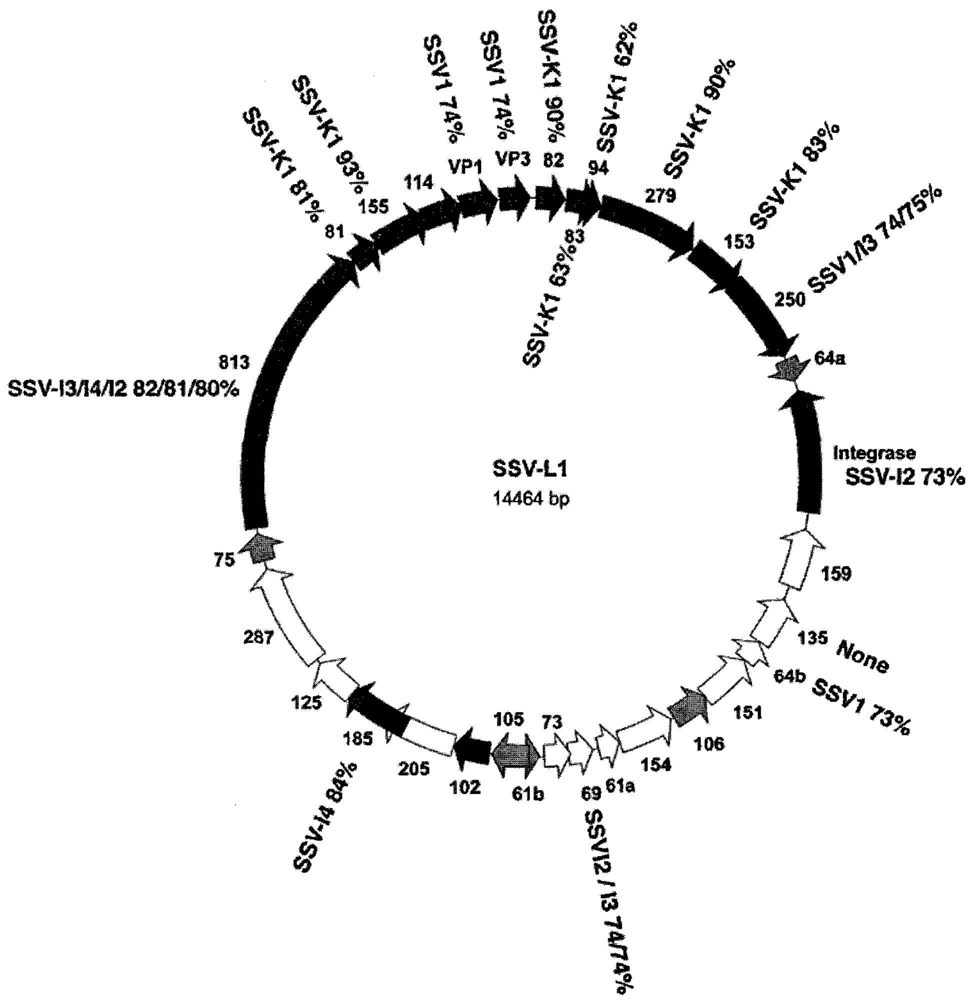

Figure 2.6: SSV-L1 genome highlighting universally conserved ORFs most similar to geographically distant viruses. ORFs conserved in all Fuselloviruses are filled in black. ORFs filled in grey are only present in SSV-L1 and SSV-RH1 genomes. ORFs filled in white are partially conserved. Virus names next to the SSV-L1 ORF labels indicate the virus genome with the most similar ORF followed by the percent amino acid identity. In the case of multiple similar ORFs, virus and percent identities are separated by slashes.

Biogeography of Fuselloviruses

Pair-wise nucleotide identity between Fusellovirus genomes (Table 2.1) supports change with respect to genetic distance but contradicts culture-independent PCR based studies (103).

Therefore, higher resolution phylogeographic techniques were applied to pair-wise comparisons between each of the 7 Fusellovirus genomes and the geographic distances separating the isolation locations. First, the sequence regions used by (103) were tested using the 251 bp sequence in the largest universally conserved ORF in SSV1, ORF C-792, and the coat protein genes. A linear regression of genetic distance as calculated by maximum likelihood using the Kimura model of evolution (54) plotted against the geographic distance separating the isolation locations confirmed the previous analysis (103), that no significant correlation was seen with either the coat protein sequences (data not shown) or the portion of the largest universally conserved ORF (Figure 2.7). In both cases p values calculated using the Mantel test with 999 replicates were greater than 0.05 and R^2 values were lower than 0.6.

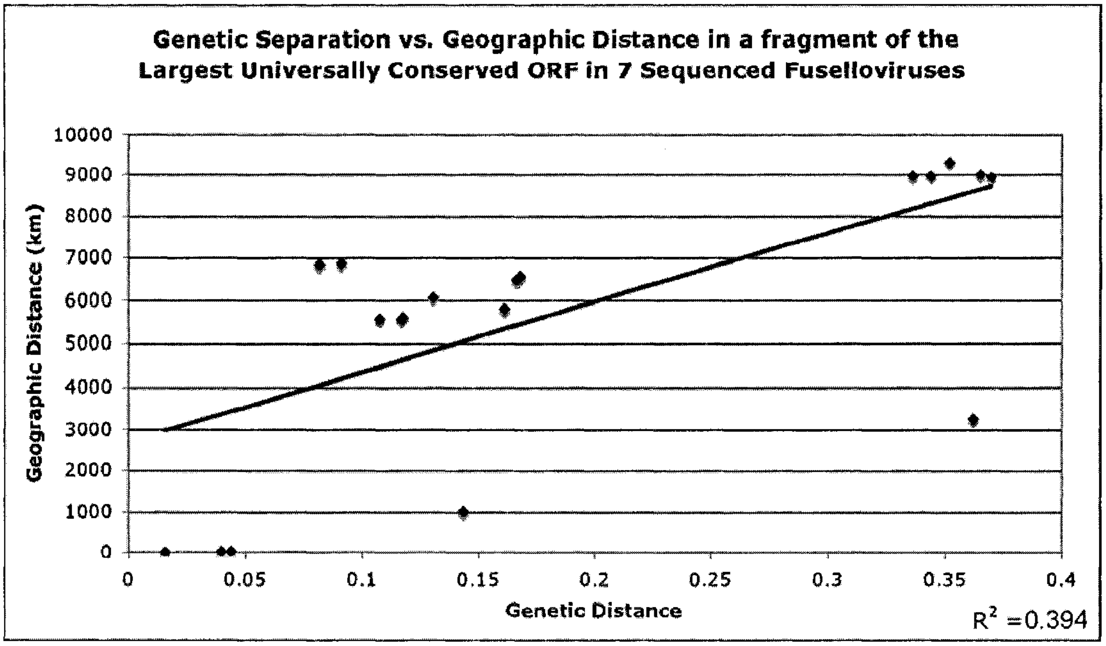

Figure 2.7: Correlation of geographic separation vs. genetic distance of the nucleotide sequence of the fragment of the largest universally conserved ORF used in (103). Geographic distance is a measurement of physical separation of the hotsprings from which the viruses were isolated. Genetic distance is the maximum likelihood value based on the Kimura model of evolution (54)

The 251 base pair sequence in the largest universally conserved ORF in SSV1, ORF C-792, and the coat protein genes make up a relatively small part of the virus genome. To determine if significant correlations can be observed using larger amounts of sequence data, linear regressions of the genetic distance, determined by amino acid similarities, of individual universally conserved ORFs verses geographic distance were calculated. The 15 universally conserved ORFs lack a statically significant correlation with the exception of the largest ORF, which does have a statically significant correlation (R^2 = 0.6916, p = 0.0040, Figure 2.8). Similar analysis of a concatenation of all 15 of the universally conserved ORF amino acid sequence also shows a good correlation between genetic distance and physical separation, ($R^2 = 0.748$, p = 0.0020, Figure 2.9). Similar correlations are also seen when comparing geographic distance to the genetic distance between degapped whole genome nucleic acid alignments of Fusellovirus genomes (R^2= 0.7244, p =0.0020, Figure 2.10).

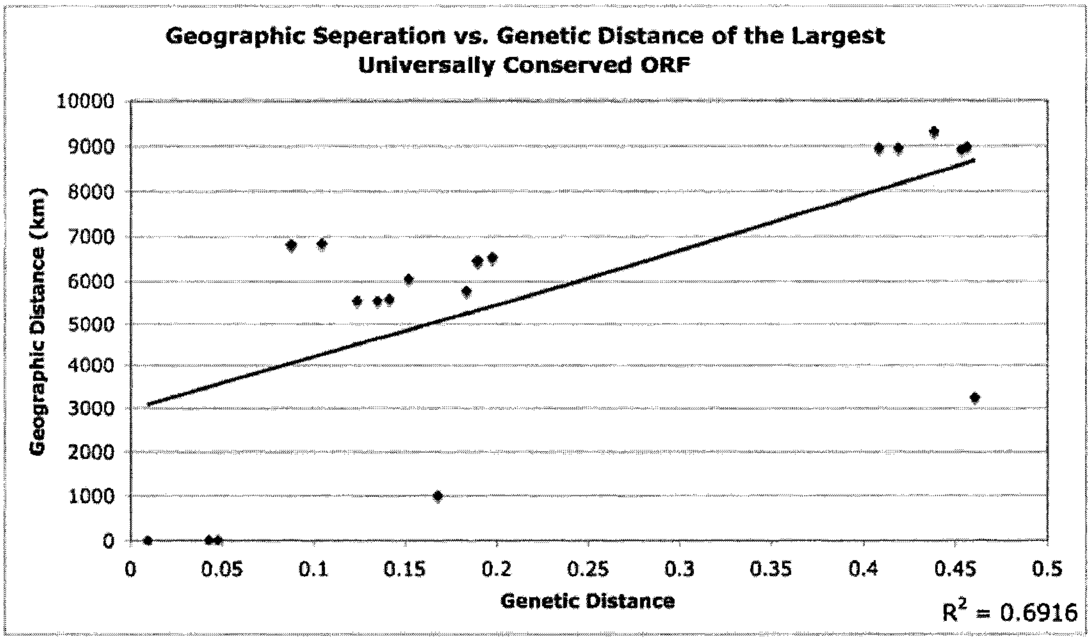

Figure 2.8: Correlation of genetic separation and geographic distance for the largest universally conserved Fusellovirus ORFs. Geographic distance is a measurement of physical separation of the hotsprings from which the viruses were isolated. Pair-wise genetic distances were calculated for amino acid sequences using maximum likelihood with the Jones-Tayer-Thorton model of evolution using Prodist in the Phylip package version 3.6.6 (33).

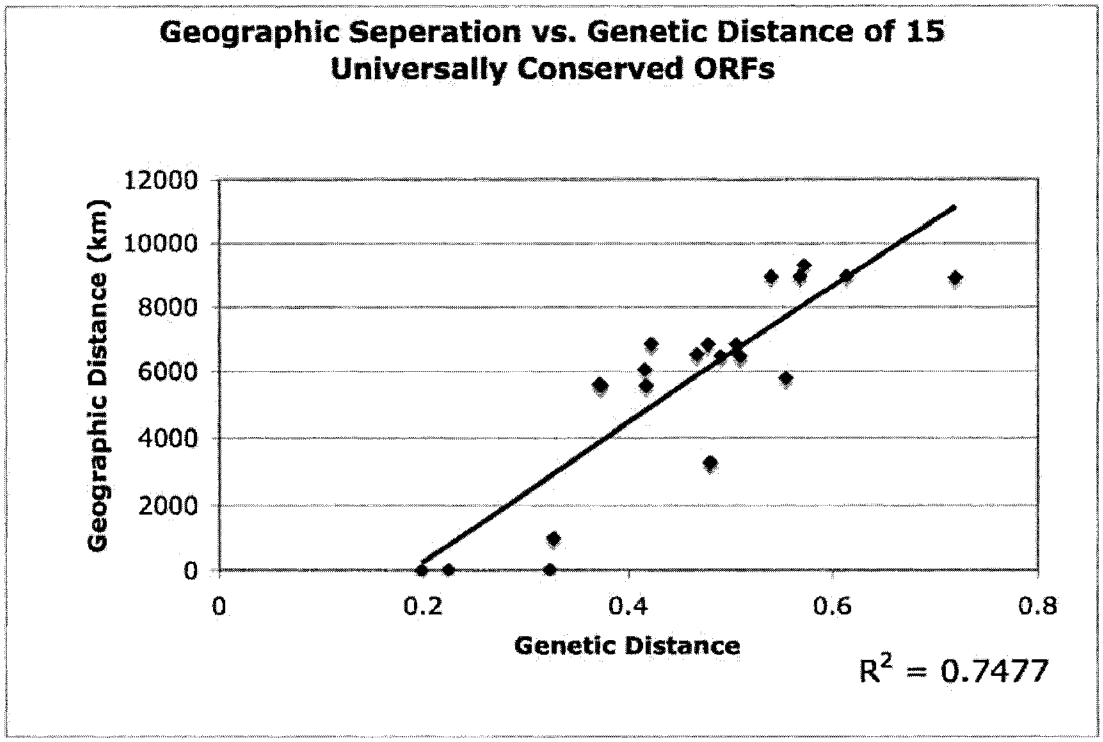

Figure 2.9: Correlation of genetic separation to geographic distance using concatenated universally conserved ORF products from Fuselloviruses. Data shown as in Figure 2.8.

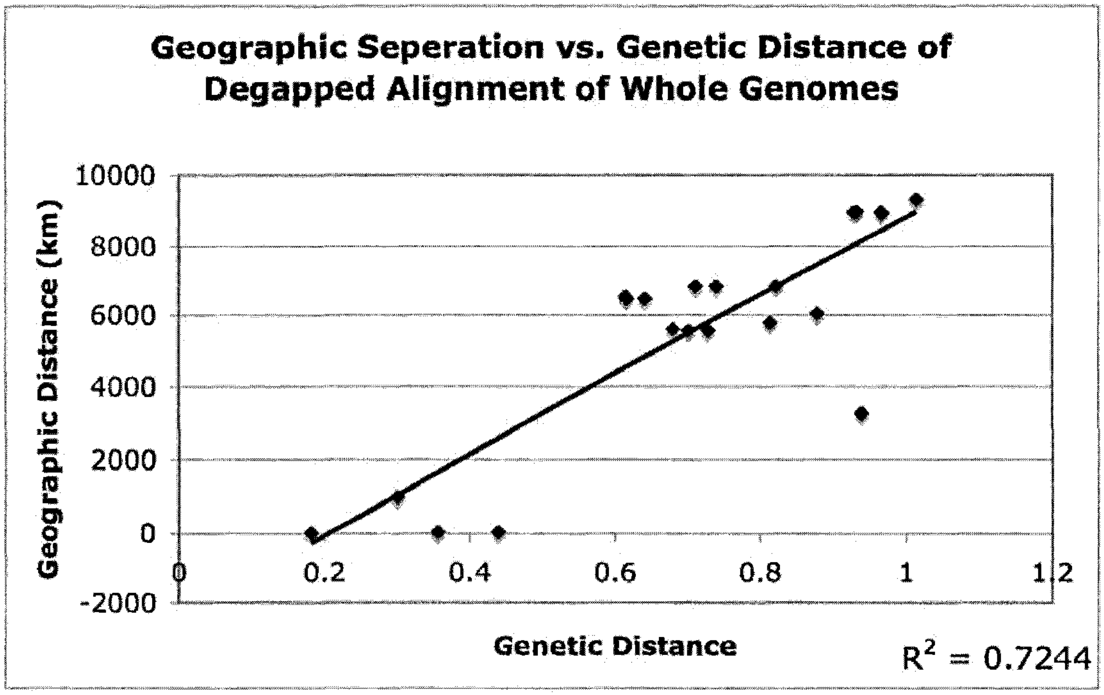

Figure 2.10: Correlation of genetic separation to geographic distance using de-gapped nucleic acid alignment of Fuselloviruses genomes. Data shown as in Figure 2.6.

Conservation of Promoter Regions

Transcription has been mapped in SSV1 and the location and putative BRE and TATA boxes for all 11 known transcripts have been identified during replication of latent viruses induced by UV-irradiation (37, 87). Due to the similar genomic structure in other Fuselloviruses similar transcription patterns are expected but have not been tested (106). To analyze the similarity of the promoters of all Fuselloviruses, the nucleotide sequences of the areas 200 bases upstream of the first ORF in each putative transcript in all Fuselloviruses were aligned to the demonstrated SSV1 transcript start sites and putative TATA and BRE sequences (37, 87, 41). For most putative transcripts the promoter regions are well conserved in all viruses (Figure 2.11). Exceptions are the putative T5 promoter and the T-ind promoter, which show no conservation between any other Fuselloviruses and the SSV1 promoters.

Possible sequencing error in SSV-K1

What appears to be the start of the transcript in SSV-K1 homologous to the T3 transcript in SSV-1 starts 57 codons before the annotated start codon for SSV-K1 ORF 252 in a possible -2 frameshift. This suggests a sequencing error, a defective gene, or a transcriptional or translational slippage event. The 57 codons that precede ORF 252 in SSV-K1 are similar to the N-termini of the longer homologues in other viruses (which range from 287 to 311 amino acids in length) corroborating the promoter data and suggesting a sequencing error.

Location of Tx Transcript in SSV-RH1.

Two similar ORFs are located in the Tx transcript homologue area in the SSV-RH1 genome, ORF B74 and ORF C82. These genes may be the result of a gene duplication event, so the possible promoter area upstream of each gene was aligned to the other viruses' putative Tx promoters. While both genes have a possible promoter, the one preceding the first ORF, ORF B74, has more similarity to the Tx promoters of other viruses (Figure 2.11).

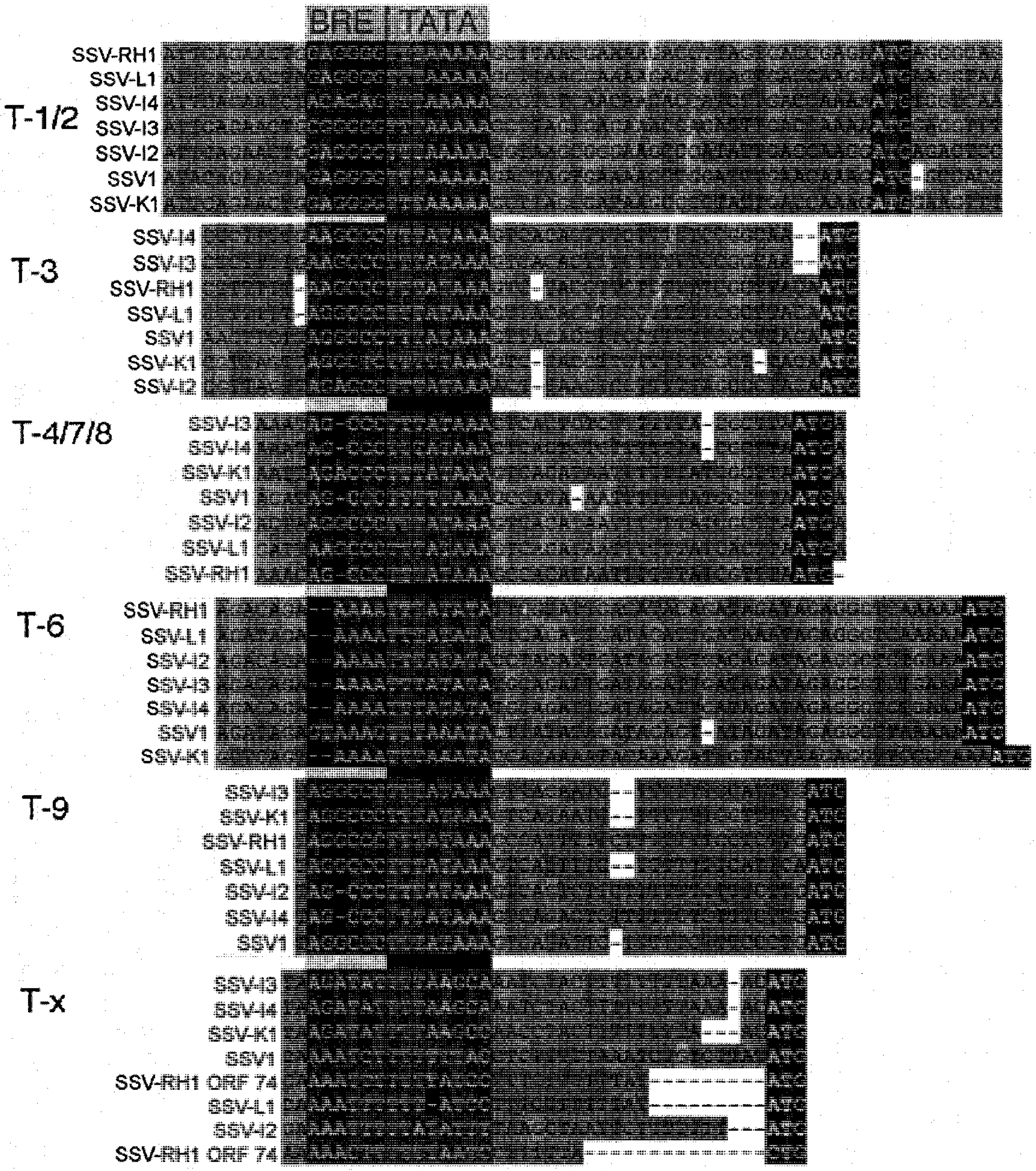

Figure 2.11: Alignment of conserved Fusellovirus putative promoter regions. Shaded with white letters areas indicate putative TFB recognition elements (BRE), TATA boxes (TATA) and the start codons of the first ORF in the transcript. SSV-RH1 ORF 74 and SSV-RH1 ORF 82 are two genes in transcript Tx thought to be the result of a gene duplication event, both were putative promoters were aligned separately.

Putative Promoters in the T5 Transcript

The T5 transcript in SSV1 is in the opposite orientation with respect to all other transcripts and contains the largest number of non-conserved ORFs, 2 of which have been shown to be nonessential (105). The similar regions in other Fusellovirus genomes also show this opposite orientation and a lack of conserved ORFs in this region (Figure 2.3). The first ORF in the SSV1 T5 transcript, ORF F-112, shows no similarity to other Fusellovirus ORFs. The putative promoter region also has a non-canonical and non-conserved promoter, having the -40 to -20 sequence of A**GTAAGAC***TTAAATA*CTAAT (37) with putative BRE and TATA box in bold and italic respectively.

The SSV1 T5 promoter region is not similar to sequences found in any other Fuselloviruses. To locate promoters in other Fuselloviruses in regions analogous to the SSV1 T5 transcript, sequences were aligned that were 200 bases upstream and 20 bases into all Fusellovirus ORFs in the same orientation as the SSV1 T5 transcript. No sequences were universally conserved in Fuselloviruses, however there is a conserved region containing a possible TATA box upstream of SSV-I2 ORF 61, SSV-I3 ORF 61, SSV-I4 ORF 61, SSV-RH1 ORF-F61, and SSV-L1 ORF 62 (Figure 2.12 A).

In addition, a possible promoter was seen upstream of the integrase gene in SSV-I2, SSV-I3, SSV-I4, and SSV-K1 (Figure 2.12 B). This putative promoter region is coding regions of ORFs that precede the integrase gene.

These predicted ORFs are non-conserved and in different orientation in SSV-I3 and SSV-I4 (Figure 2.3). This putative promoter suggests a possible monocistronic transcript of the integrase gene in these viruses. No sequence similarity is observed in SSV1, SSV-RH1 or SSV-L1.

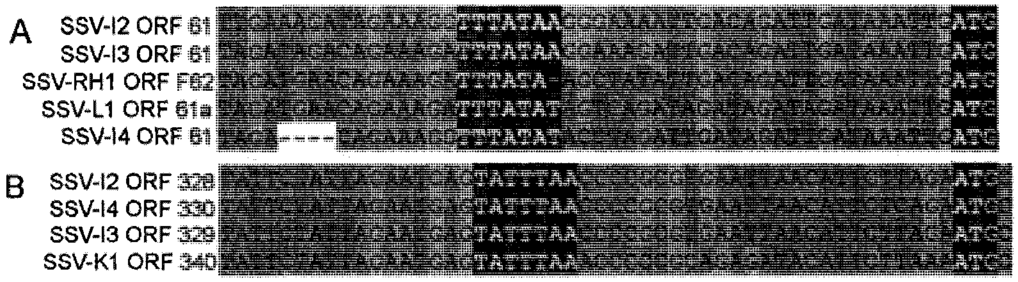

Figure 2.12: Possible previously unannotated Fusellovirus promoters in the T5 transcript area **(A)** and upstream of the integrase gene **(B)**. Start codon and possible TATA box darkened.

Table 2.2 (Following two pages) Genes and ORFs in the 7 Fusellovirus genomes as annotated above (Figure 2.3). Similar genes are grouped in rows, darker shading represent conservation within more of the genomes. TMH indicates predicted trans membrane helices (See methods for prediction). *Indicates ORFs likely disrupted in the SSV1 genome (105).These genomes failed to produce virus when transformed into *S. solfataricus*, suggesting that they are essential.

SSV1	SSV-I2	SSV-I3	SSV-I4	SSV-K1	SSV-RH1	SSV-L1	TMH
VP1	88bVP1	VP1	VP1	B137VP1	VP1	VP1	Y
*A153	153	157	152	C157	C154	153	N
*B78	79	80	80a	A79	A79	81	N
A82	83	83	82	B83	A83	82	Y
A92	90	90	89	B90	A93	94	Y
*B115	112	120	107a	A123	A113b	114	N
*B129	155	138	124	B158	C150	185	N
*B251	233	250	233	A231	A247	250	N
*B277	276	179	280	C279	C277	279	N
C102a	100	101	100	B98	A102b	102	Y
*C166	176	170	167	B169	B170	155	N
*C792	809	809	808	B793	B812	813	N
C84	88c	89a	81	A82	C78	83	N
*D335	328	329	330	F340	D335	335	N
VP3	VP3	VP3	VP3	VP3	VP3	VP3	Y
C80	82a	84b	79	C81	B64		N
	205	199	206	A204	C287	205	N
D244	211	208	209		D212	159	N
A100	96	96	96	C96			N
		110a	107b	B111	C113a	125	Y
	61	61b	61		F62	61a	N
A45	48	77b	45	C43			N
A79	82b		80b	A80	B79		N
D63	57	61a	62			64b	N
	79a	77a	73		E73	73	N
		163	159b		E152	151	N
		143	143		E150	154	N
	305				C247	287	N
	72	74				69	N
		311	311	B252			N
		110b	111	C108			Y
E178					A148		Y
F93				E81			N
	88a	89b					N
		159	159a				N
		84a	71				N
					D57	61b	Y
					A102a	105	N

SSV1	SSV-I2	SSV-I3	SSV-I4	SSV-K1	SSV-RH1	SSV-L1	TMH
					B85	64a	N
					B74	75	Y
					D110	106	N
A132							N
A291							N
B49							N
C102b							N
C124							Y
E51							N
E54							N
E96							N
F112							N
F92							Y
VP2							N
	68						N
	55a						N
	70						N
	55b						N
	310						N
	106						N
	126						Y
			64				N
			59				N
			49				N
				F58			Y
				C158			N
				B494			N
				A460			N
				C78			N
				B53			N
				B64			N
					F61		N
					C49		N
					B50		N
					A158		N
					C59		N
					C82		Y
						135	N

Putative Origins of Replication

As yet the precise location of origin of replication (*oriV*) of the Fuselloviruses remains unknown. To help determine the *oriV* location, analysis of the GC skew and purine skew was performed with all genomes. Figure 2.13 shows CG skew and purine skew in the 7 aligned Fusellovirus genomes. All genomes were manually aligned so that the first ORF of the putative transcript homologous to the T5 transcript in SSV1 is located at the five-kilobase pairmark on the graph. In both CG skew and purine skew analysis there is a sharp drop in skew in all genomes near the putative origin of replication in SSV1, with the exception of SSV-K1.

Integration in Fuselloviruses

Integrase attA sites have been demonstrated for SSV1 (95), SSV-I2, SSV-K1, and SSV-RH1 (117). The putative integrase gene in SSV-L1 shares a sequence of 49 identical base pairs with the 3' end of the *S. solfataricus* tRNA 30 (glycine, CCC anticodon), and tRNA 8 (glycine CCG anticodon) 47 base pairs of this sequence are identical to the SSV-I2 *attP* site, differing only in the first two nucleotides (106) (Figure 2.15). Compared to the SSV-I2 *attP* site, the putative SSV-L1 *attP* site is less similar to its geographically closest neighbor SSV-RH1, as well as the virus that shares the most similar integrase protein sequence, SSV-K1 (Figure 2.14, Figure 2.15).

Not surprisingly due to their overall similarity, SSV-I3 and SSV-I4 have the same putative integration site. Interestingly, this site is very similar to *attP* of SSV-K1 (Figure 2.16). SSV-K1 has been shown to integrate into three tRNA genes in *S. solfataricus*. tRNA 40 and tRNAs 26 and 32 each of the latter with 4 nucleotide differences relative to the virus genome (117). In SSV-I3 the situation is a near perfect reversal, showing a 52 base pair match to tRNA 32 with one nucleotide difference from tRNA 26 and four differences from tRNA 40.

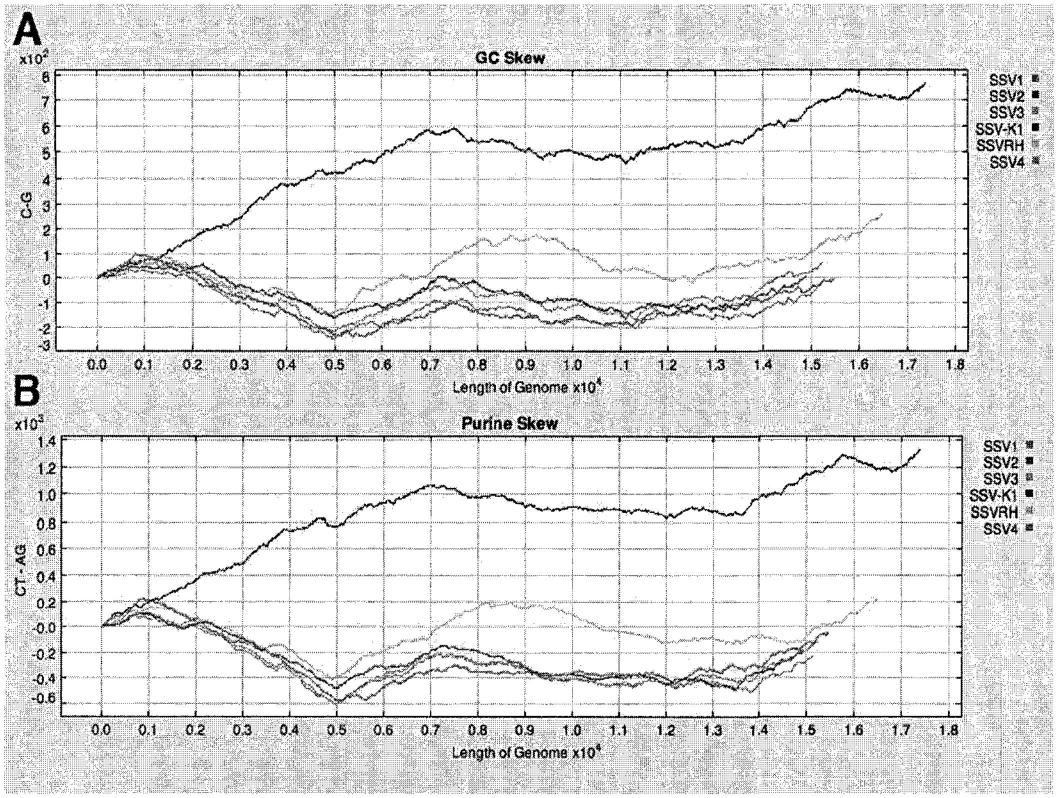

Figure 2.13: Putative origin identification. (A) GC skews of Fusellovirus genomes (B) Purine skew of the same data. In both cases a window of 10 base pairs was used. All genomes were aligned so that the first ORF orientated in the same direction of the T5 transcript in SSV1 is located at the five-kilobase pairmark on the graph. (See Figure 2.3)

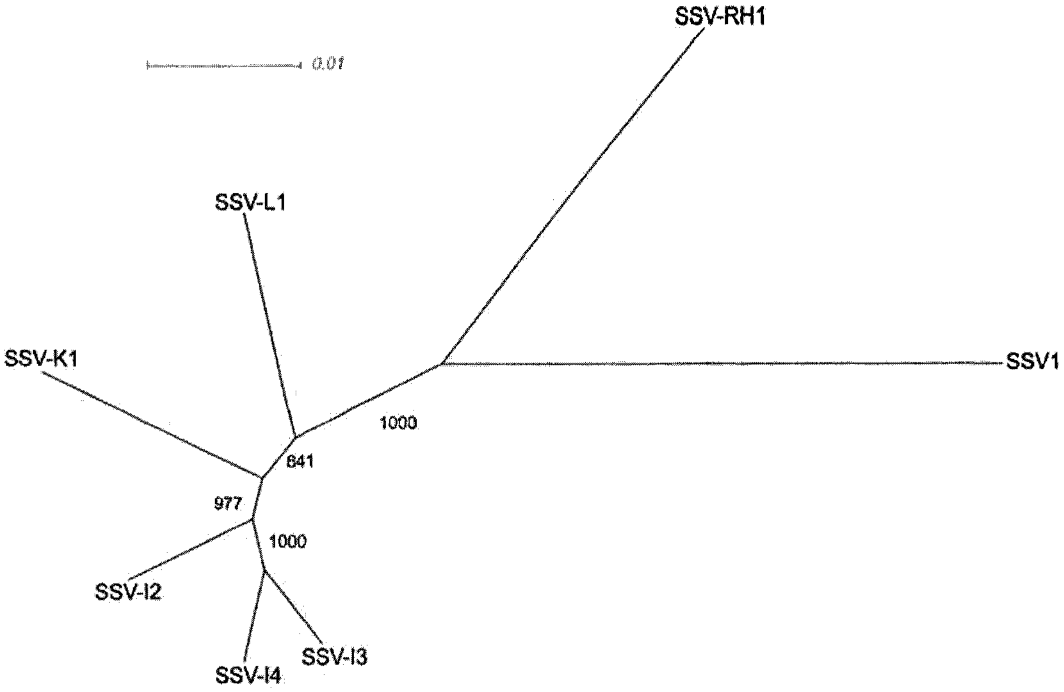

Figure 2.14: Fusellovirus integrase proteins. Neighbor-joining tree of the amino acid sequences of viral integrase genes (108). Bootstrap support based on 1000 replicas is labeled. Scale bar equals 0.01 substitutions per site.

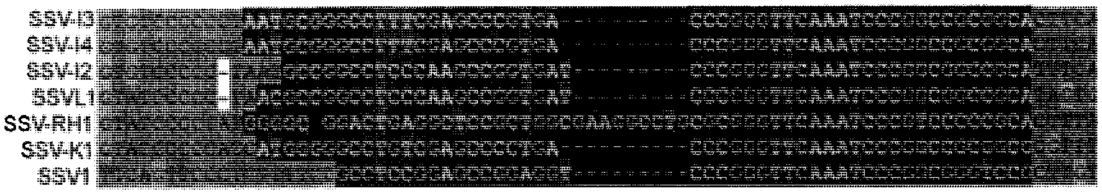

Figure 2.15: Putative Fusellovirus *attP* sites. The integrase gene of each virus was aligned with all known tRNA and snoRNA genes in the *S. solfataricus* genome and then to each other. Areas of exact match to *S. solfataricus* tRNA genes are shaded with white letters.

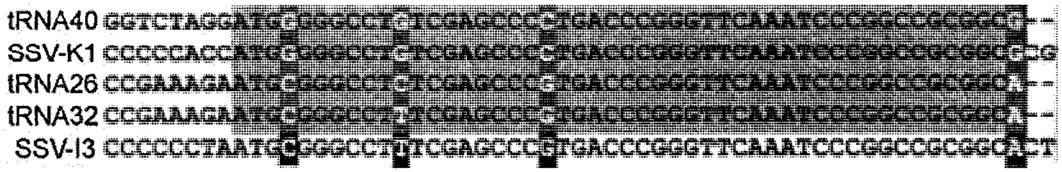
tRNA40 GGTCTAGGATGCGGGCCTCTCGAGCCCCTGACCCGGGTTCAAATCCCGGCCGCGGCG--
SSV-K1 CCCCCACCATGCGGGCCTCTCGAGCCCCTGACCCGGGTTCAAATCCCGGCCGCGGCCG
tRNA26 CCGAAAGAATGCGGGCCTCTCGAGCCCGTGACCCGGGTTCAAATCCCGGCCGCGGCA--
tRNA32 CCGAAAGAATGCGGGCCTTTCGAGCCCGTGACCCGGGTTCAAATCCCGGCCGCGGCA--
SSV-I3 CCCCCCTAATGCGGGCCTTTCGAGCCCGTGACCCGGGTTCAAATCCCGGCCGCGGCACT

Figure 2.16: Predicted *attP* site of SSV-I3. Known *attP* sites of SSV-K1 are included with its experimentally demonstrated *attA* sites. The attachment sites are shaded. Shading is inverted where sequences mismatch.

Conservation of Fusellovirus Structural Genes

In SSV1 the viral structural proteins VP1 and VP3 are the major components of the virus capsid (86). When the VP1 gene sequence and protein sequence were compared in SSV1 it was found that VP1 is post-translationally truncated at its N-terminal end, and that both VP1 and VP3 are very similar, particularly in their hydrophobic C-termini (86). VP1 and VP3 genes are well conserved in the other viruses (Figure 2.17). The highly conserved hydrophobic areas in VP1 and VP3 are of similar length to membrane spanning alpha helices and contain charged amino acid residues at the outward ends of the helices indicating the area between the membrane spanning domains faces outwards with respect to the cell based on modeling predictions (104).

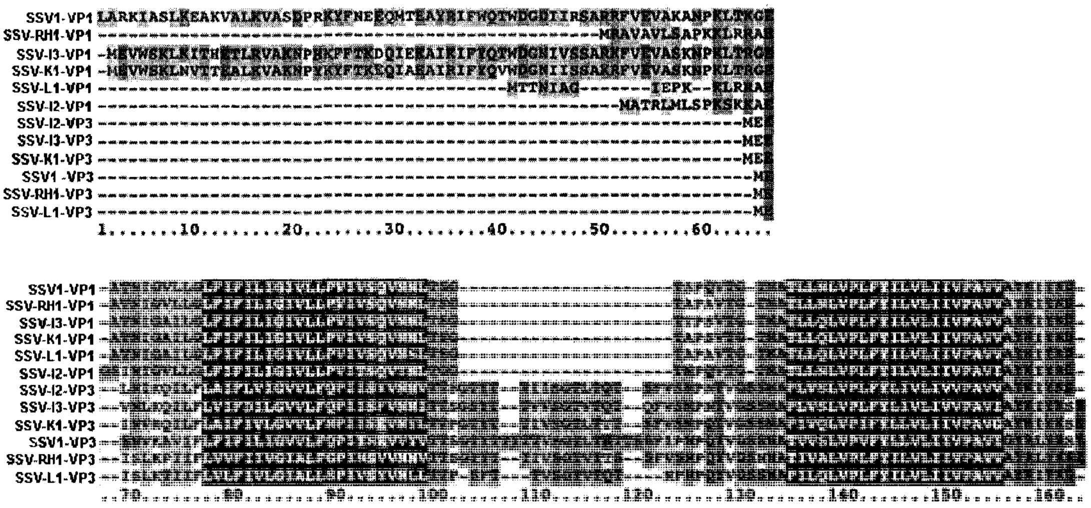

Figure 2.17: Alignments of VP1 and VP3 protein sequences from all Fuselloviruses. Last residue in the top row (Glutamic acid) is the post-translational cleavage site in the SSV1 VP1 protein. Predicted trans-membrane domains are darkened.

Repeated Sequence Analysis

To search for further evidence of recombination, tandem and inverted repeat sequences (26) were located and mapped in all Fusellovirus genomes (Figure 2.18). Repeats cluster near the T-Ind, T5 and T6 promoters in SSV1 and similar areas in other Fuselloviruses.

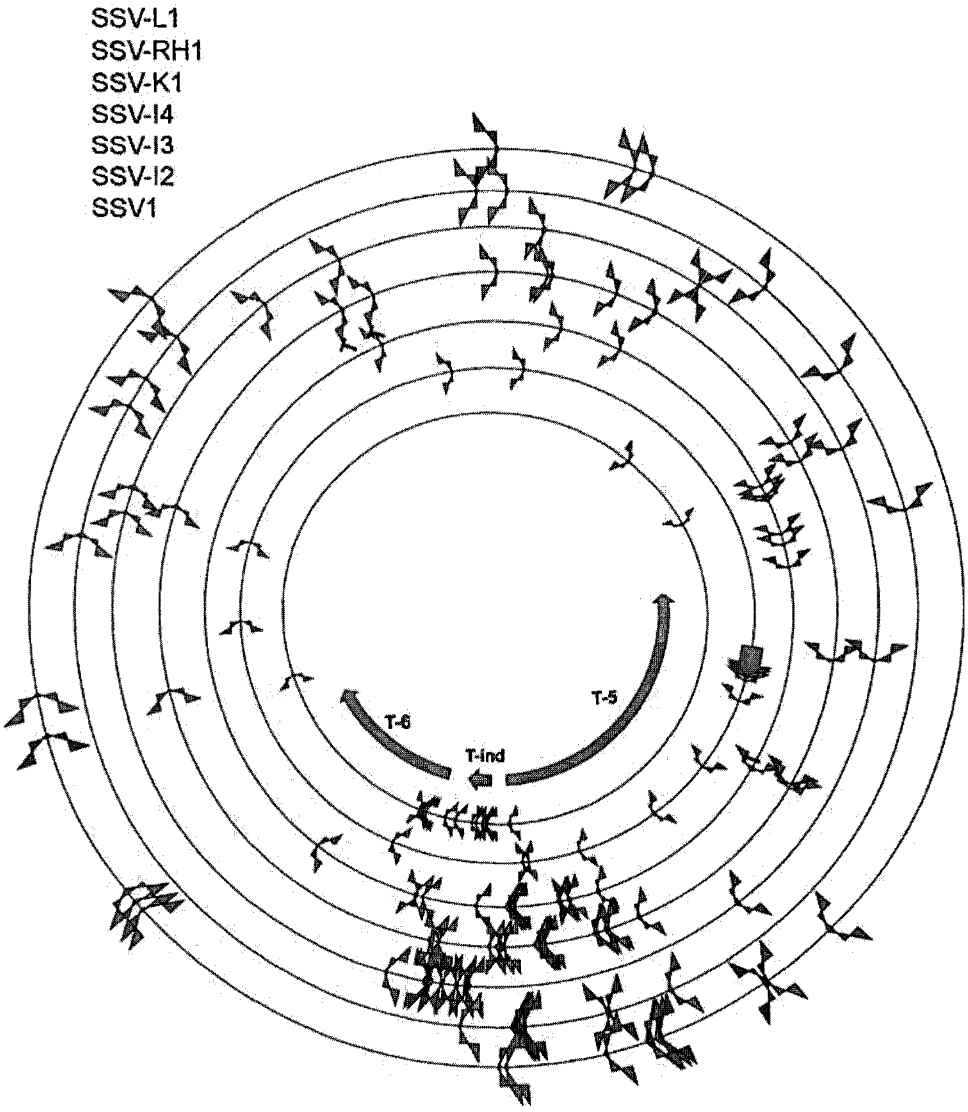

Figure 2.18: Repeat elements in Fusellovirus genomes. Circles represent genomes ordered from outside in are SSV-L1, SSV-RH1, SSV-K1, SSV-I4, SSV-I3, SSV-I2, and SSV1. Red elements are tandem repeats, blue elements are inverted repeats. All repeat elements annotated contain a minimum length of 12 bases (14). All genomes are aligned so that the first ORF of the putative transcript homologous to T5 transcript in SSV1 is at the bottom of the map. The T5, T6, and T-Ind transcripts in SSV1 are indicated as red arrows inside the SSV1 genome.

Discussion

Definition of the Fusellovirus Core.

The comparison of seven Fusellovirus genomes isolated from around the world allows further refinement of number of universally conserved Fusellovirus ORFs from 18, based on the comparison of SSV1, SSV-I2, SSV-K1, and SSV-RH1 (117), to 15. From sequence conservation, promoter conservation and ORF synteny, these 15 "core ORFs" appear to be necessary for successful virus replication. Gene disruption data in SSV1 (105) supports the necessity of these core ORFs in viral replication, as genomes with a putatively disrupted ORF in any 1 of 9 core genes tested failed to replicate when transformed into *S. solfataricus* (putatively disrupted ORFs are noted with an asterisk in Table 2.2). One of these putatively disrupted genes was the integrase gene. In Chapter 4 it is shown that a complete deletion of the integrase gene does not stop replication, but most likely lacks the ability to compete with or spread as efficiently as integrase containing viruses.

The 15 universally conserved ORFs lowers the previously apparent minimum number of Fusellovirus core genes from 18 that were based on the comparison of SSV1, SSV-I2, SSV-K1, and SSV-RH1 (117). The 3 ORFs of the original 18 that are not completely conserved are homologues of the SSV1 ORFs A100, A79 and C80. All are encoded in the T6 transcript away from the 15 conserved ORFs and near genes that are non-conserved (Figure 2.3), and all absent only in North American strains (Table 2.2). The A100 homologue is absent in SSV-L1 and SSV-RH1 and A79 and C80 are absent only in SSV-L1. The addition of more genomes to the Fusellovirus database may further reduce the number of universally conserved genes below 15. Analysis of many more genomes will be needed to truly assess what makes up a Fusellovirus genome core.

Remodeling in the Non-core Region of Fusellovirus Genomes

Four lines of evidence suggest genomic regions outside of the area containing the core ORFs appears to undergo more remodeling than that of the core ORF region itself.

First, the concentration of inverted and tandem repeat sequences shown in Figure 2.18 is in the region of the genome with the least conserved ORFs (Figure 2.3). Repeat elements such as these are correlated with increased genomic rearrangements in microbes (3, 4) suggesting they may be targets for recombination, however at this time a mechanism for recombination is not known.

Second, the variation in Fusellovirus genome size, ranging from 17.3 kilobase pairs, to 14.4 kilobase pairs, is mainly in the non-core region. Within the seven viral genomes the combined sequence length of the 15 universally conserved ORFs varies by only 207 nucleotides or 2.4%. The remaining 2.7 kilobase pair size variance is made up of differences in the variable genes, which occupy between 5.9 and 8.8 kilobase pair of sequence, a variance of 33%. Intergenic regions make up less than 5 % of the genome most of this variance in the ORFs.

Third, the nearly identical sequence, including ORFs and intergenic regions in SSV-I3 and SSV-I4, suggest a recent horizontal transfer may have taken place between these viruses or their ancestors. Five ORFs unique to the non-core region of the SSV-L1 and SSV-RH1 genomes may also be a result of recent recombination between these two isolates from the western United States or their ancestors (Figure 2.6).

Lastly, homologues of Fusellovirus non-core ORFs are found in extrachromosomal elements outside the family Fuselloviridae (Table 2.3). While two core ORFs are also found in non-Fusellovirus extrachromosomal elements, it is remarkable that these less conserved ORFs are also shared.

Table 2.3: Extrachromosomal elements with similar ORFs to those found in Fusellovirus genomes. E values denote the best BLAST score between the Fusellovirus ORF, marked with an asterisk, and the virus or plasmid ORF in the leftmost column. Core Fusellovirus ORFs are white, non-core Fusellovirus ORFs are shaded grey.

Related virus/plasmid	SSV1	SSV-I2	SSV-I3	SSV-I4	SSV-K1	SSV-L1	SSV-RH1	e value
ATV ORF55	B-251	233	250*	233	A231	250	A247	8 E-12
STSV1 ORF69	B-251	233	250*	233	A231	250	A247	5 E-06
pSSVx ORF288	B-251*	233	250	233	A231	250	A247	1 E-12
ATV ORF07	F-112							0.33
AFV1 ORF59a	A-45	48	77*	45	C43			0.005
SIRV1 ORF15	C-102							8 E-05
SIRV2 ORF22	C-102							1 E-04
SIRV1 gp132		88a*	89b					2 E-11
SIRV2 gp08		88a	89b*					4 E-06
AFV1 ORF223		205*	199	206	A204	205	C287	7 E-03
pSSVi ORF336	D335							4 E-136
pSSVi ORF735					A460			6 E-22
pSSVx ORF154	A153	153	157	152	C157	153	C154*	3 E-37
AFV1 ORF157							A-158	4 E-34

*ATV- *Acidianus* Two Tailed Virus (82), STSV-*Sulfolobus tengchongensis* spindle-shaped virus (122), AFV-*Acidianus* Filamentous Virus (15) SIRV *Sulfolobus islandicus* Rod Shaped Virus (78). pSSVi and pSSVx are virus/plasmid hybrids (7, 111)

The most obvious example of gene movement into Fuselloviruses is in SSV-K1. In the area analogous to the SSV1 T5 transcript, a section of the SSV-K1 genome is inverted relative to the other viruses (Figure 2.3). Included in this inversion are four predicted ORFs. The SSV-K1 ORF B-494, is similar (higest e value 1^{-34}) to several helicase and DEAD box containing proteins found in several Archaea including *Sulfolobus* species (63). ORF A-460 shows high similarity (e value $6x10^{-22}$) to a protein of unknown function in *S. solfataricus* and similarity (e value $2x10^{-2}$) to pSSVi ORF 735, which is thought to be a primase, polymerase and helicase (111). The ORF immediately downstream of A-460 also shows similarity to the C-terminal section of the pSSVi ORF-735 (e value 3×10^{-11}) and may be part of an original full-length gene, sequencing error, or gene fusion in pSSVi. Function of these genes in SSV-K1 is unknown, however the similarity to genes found in the virus/plasmid hybrid pSSVi suggests that these genes may have been acquired from a genetic element similar to pSSVi or visa-versa.

Transcription in Fuselloviruses is Conserved

Apart from SSV1 the only other analysis of Fusellovirus promoters was done for SSV-I2 (106). That study noted that all of the TATA boxes in putative promoters were conserved between SSV1 and SSV-I2 with the exception of T-ind (106). This study analyzed all core promoter elements within all seven genomes and found two previously undetected putative promoters. This analysis confirmed that all viruses share similar promoters with the exception of the SSV1 T5 and T-Ind promoters. Therefore, overall transcriptional regulation is likely conserved in Fuselloviruses.

T-ind Transcript

Two of the characterized promoters in SSV1, those of T-ind and T-5, are not conserved in any of the other viruses. ORFs in this region of Fusellovirus genomes are also not conserved (Figure 2.3). The T-ind transcript is apparently responsible for the strong UV irradiation-mediated induction of

SSV1 (86). This induction has only observed in the SSV1 virus (106). Unlike most archaeal transcripts, mapping of the T-ind transcript indicates a promoter region devoid of the canonical TATA box. The lack of homology of the T-Ind promoter or transcript homology suggests that SSV1 may be unique in its ability to be induced by UV irradiation.

T5 transcript reassignment

The T5 promoter sequence of SSV1 is similar the canonical archaeal promoter sequences, but no similar sequences are seen in similarly oriented transcription units in other Fuselloviruses (Figure 2.3). Expression of these ORFs has not yet been shown. Moreover, most ORFs in this study and in other studies were annotated by finding a start codon followed by at least 50 uninterrupted codons, a rather liberal annotation which may be incorrect (75, 106, 117).

A putative promoter two to three ORFs downstream of the first ORF in the putative T5-like transcript was identified in the SSV-I2, SSV-I3, SSV-I4, SSV-RH1 and SSV-L1 (Figure 2.12 A). The sequence of this putative promoter is similar to canonical promoters suggesting this may be the start of the T5 transcript analogue, and that the ORFs upstream of this promoter may not be expressed. Were this to be the case an intergenic region would be present near the location of the putative *oriV* and no overlap of the oppositely oriented T5 and T6 transcript analogues would be present (Figure 2.3). This arrangement would more closely match that seen in SSV1. Northern blots or similar analyses are needed to determine which ORFs are truly expressed in these Fuselloviruses.

An Integrase Transcript?

A second putative promoter was identified upstream of the integrase gene in SSV-I2, SSV-I3, SSV-I4 and SSV-K1 (Figure 2.12). No transcripts have been reported in this location in SSV1. Nevertheless, it was noted by Frols et al. (37) that more integrase gene mRNA was observed than mRNA from other ORFs in the T5 transcript under non-inducing conditions despite being the last gene in the polycistronic transcript (37). This suggests that the integrase gene may have its own, non-conserved, promoter in some Fuselloviruses.

Fusellovirus Replication Origins:

As yet the precise location of the viral origin of replication (*oriV*) of the Fuselloviruses remains unknown. In SSV1 indications that the origin is located near the T-ind promoter are based on a number of indirect observations. First is the opposite orientation of the reading frames extending away in both directions from this location and a number of inverted repeat sequences (75). Secondly, preliminary results from the Bell and Schleper labs mentioned in (37) used 2D DNA gels to map the *oriV* to this region. Lastly, the I-Ind transcript is strongly upregulated upon virus induction (37). In addition to this 1700 bases containing the putative origin of replication were used in the construction of a shuttle vector, named pEXSS, reportedly capable of replication in *Sulfolobus* (23, 25). The functionality of this plasmid is uncertain however, as a number of labs are unable to replicate the published data (K. Stedman, personal communication). Furthermore, recently published gel shift data suggests an unknown protein found in *S. shibatae* (the original host of SSV1) called STRIP (SSV1T5/T6 region interacting protein), binds to sequences found only in SSV1 near the T5 and T6 promoter start site. This binding does not affect the *in vitro* transcription of the T5 or T6 promoters and is hypothesized to be involved with DNA replication (83).

Both GC skew and Purine skew analysis of the SSV1 genome support the data above that the *oriV* is located between the T5 and T6 transcripts. It also suggests that *oriVs* are conserved in location in all analyzed genomes except possibly the SSV-K1 virus (Figure 2.13). No detectable sequence similarity was found in alignments within 700 base pairs of the minimum GC skew indicating that the sequence of the putative *oriVs* are not conserved in these viruses (data not shown).

GC skew is a common method of locating putative origins of replication in prokaryotes and in their plasmids and viruses, including those found in the Archaea (122). The decrease in guanine in the leading strand, calculated by the formula of (G-C)/(G+C) and plotted over a given window of bases is used to calculate the skew. Purine skew is also an indication of replication origins (36). Care must be taken with the interpretation of GC and Purine skew data however, as both can be inaccurate indicators of replication origins when genomes have recent rearrangements (36). This may be the case with SSV-K1, which contains a region with little similarity, in both the orientation and sequence of ORFs, relative to other Fuselloviruses. ORFs in this region are similar to those found in the pRN plasmid family, indicating that this may be a recent insertion. Nevertheless, after removal of the inverted sequence in SSV-K1 and reanalysis of the GC and Purine skew, a characteristic dip seen in the other six viruses was undetectable (data not shown). Whether SSV-K1 has a unique *oriV* location, or is similar to that of the other viruses and masked by recent insertions, is unknown at this time.

Phylogeography in Fuselloviridae

The significant correlation between genetic difference and geographic distance between the seven Fusellovirus genomes indicates the dispersal of these viruses and/or their hosts are limited, most likely by the barrier of inhospitable areas separating the hot springs. The most statistically significant correlations to geographic distance are observed with the complete de-gapped nucleotide sequence (Figure 2.10), and the amino acid sequence of the concatenated universally conserved ORFs (Figure 2.9). A weaker but statistically significant correlation of genetic difference to geographic distance is observed the largest conserved ORF (Figure 2.8). Correlations between genetic separation and geographic distance between amino acid sequence of the smaller universally conserved ORFs or the nucleotide sequence used by Snyder et al (103) to geographic distance are not significant, indicating that a large portions of the virus need to be analyzed to observe a correlation to genetic distance.

The weak signal showing the correlation of genetic difference to geographic distance in Fuselloviruses may be explained by the mechanisms that limit genetic drift (66). Selection pressure on genes and sequence elements, such as promoters, and replication origins, prevents drift (66). Assuming correct annotation, nearly all of the Fusellovirus genomes are made up of genes and sequence elements, therefore it is reasonable to assume much of the genome is constrained from drift. Limited movement of viruses between hotsprings could also reduce genetic drift by allowing an exchange of genetic information.

Previous work on *Sulfolobus* strains isolated from thermal areas separated by distances spanning several meters to thousands of kilometers, show a correlation of genetic difference to geographic distance (115) Similar to that observed here. This correlation was also attributed to the barrier of uninhabitable area separating thermal springs from each other, restricting gene flow and allowing for genetic drift of isolated communities.

In contrast, culture independent environmental data collected from three hot springs in Yellowstone National Park suggest that virus populations change rapidly over time and space, that individual hot springs carry a great deal of diversity, and that no correlation is seen in viruses with respect to genetic and geographic distance (103).

The SSV data used in that study were based on amplification and sequencing of part of the putative Fusellovirus coat protein genes and an approximately 250 base pair section of the largest universally conserved open reading frame (103). Phylogenetic trees of the coat protein sequence place the published fully sequenced Fusellovirus genomes in a separate monophyletic clade from the environmental samples with the exception of SSV-K1, which groups with environmental sequences closest to the other fully sequenced virus clades (103). Phylogenetic trees of the sequence of the portion of the largest ORF placed fully sequenced Fusellovirus genomes interspersed within many other sequences but on longer branches than nearly all environmental sequences (103). Rarefaction curves from these data show that the actual diversity of the hot springs were not represented by the hundreds of samples used in this study (103). In both cases the sequences used for these analyses were amplified from environmentally isolated DNA using a 0.2-micron filter retentate. As Fusellovirus particles are generally around 60x90 nm in size it is likely many viruses were not retained in the filter. In addition, because these sequences are culture independent, some or most may be from proviruses integrated into host genomes rather than actively replicating viruses. Integrated provirus would have no selection to keep their genome intact; therefore they would quickly drift, as is seen in the inactive virus fragment in *S. shibatae* (85). Therefore the reported virus diversity in Yellowstone hot springs (103) may overestimate the actual diversity present by the inclusion non-functional proviruses.

Even if all of the environmental sequences analyzed by Snyder (103) were from active viruses the region analyzed is likely too small to observe a correlation of genetic difference to geographic distance. This is demonstrated by using the same sequence area in the seven fully sequenced viruses to make pair-wise comparisons, which correlate poorly to distance (Figure 2.7).

Finally, the genetic separation of the Fuselloviruses may be apparent only over large distances. The short distance separating the Yellowstone hotsprings may not provide an effective enough barrier to dispersion to allow drift, as hypothesized by Snyder et al. (103).

To summarize, different Fusellovirus genomes show a direct correlation between genetic separation and geographic distance over large scales using regions of the genome spanning multiple ORFs. This trend may not be apparent with viruses separated by small geographic differences, and may be blurred by the inclusion of non-replicating virus sequence, or by looking at small areas of the genome. To truly understand the movement of these viruses between hot springs and their genetic separation from one another many more virus genomes will need to be sequenced and compared.

Diversity in Islandic viruses.

A comparison of the genome similarity of Fuselloviruses shows that total nucleotide identity ranges from 51% when comparing the distantly separated viruses SSV1 and SSV-L1 to 86% when comparing the nucleotide identities of SSV-I3 and SSV-I4 (Table 2.1). Within the latter pair, a stretch of over 1800 base pairs of identical sequence and several other large stretches of identical sequence (Figure 2.4) are evidence for 4 evolutionary scenarios. The first scenario is that these viruses only recently diverged from each other, the second that there is a very high degree of selection at the nucleic acid level of these sequences, third that this sequence has been recently laterally transferred, and fourth, a combination of the above 3 scenarios. The first seems most likely based on the observation that several other regions of the genome share hundreds of base pairs of identical sequence, the geographic closeness of the isolation (7km), and the overall sequence identity (Figure 2.4). It is puzzling however, that the region of the genome containing the 15 conserved genes does not share a similar conservation, and suggests that recombination or large regions of the genome may have taken place between SSV-I3 and SSV-I4 rather than a recent divergence from a common ancestor. Sequencing of many whole genomes or a metagenome sequencing of hotsprings at different timepoints could possibly resolve this issue. It should be noted that as SSV-I3 and SSV-I4 were sequenced in different labs at different times, (this work and (77) respectively) the probability of contamination is extremely low.

Sequence differences between SSV-I3 and SSV-I4 are much less than the differences between these viruses and SSV-I2, from a similar but slightly more distant hot spring (Figure 2.2 and Table 2.1). The 1800 base pair identical sequence shared by SSV-I3 and SSV-I4 is not shared with SSV-I2 (data not shown). This suggests that the SSV-I2 lineage diverged from that of SSV-I3 and SSV-I4 earlier that SSV-I3 and SSV-I4 diverged from each other despite being separated by similar geographic distances. As SSV-I2 was isolated in a co-culture with the virus/plasmid hybrid pSSVx it is also possible the relationship of this extra chromosomal element has had an impact on the evolution of the virus (77).

VP Genes

The genes for the two universally conserved SSV1 structural proteins, VP1 and VP3, are very similar, both between viruses and between the two genes (Figure 2.17). The similarity of the genes led to the proposal that they may have originated from a gene duplication event (86). However, the VP1 genes are dissimilar in their N-termini. In SSV1 it was demonstrated that VP1 is post-translationally cleaved (86) and suggested that the post-translational cleavage of the VP1 protein occurs in the SSV-I2 virus (106), SSV-RH1 and SSV-K1 (117). This study suggests that this is common for all SSVs (Figure 2.17). The protease responsible for this cleavage has yet to be characterized, it is tempting to speculate that it is encoded by one of the 15 completely conserved ORFs.

Immuno-electron microscopy observations indicate that SSV1 VP proteins aggregate at the membrane before virion formation (124). Membrane spanning domains predicted here suggest that the proteins may be embedded in the membrane prior to virus assembly. These hydrophobic alpha helices may also aid in holding the virion particle together once the virus is assembled.

Surprisingly, the DNA binding protein encoded by the VP2 is found in the purified capsid of SSV1 (86). Its gene is not present in any other sequenced Fusellovirus including those analyzed in this study (106, 117), and is one of the only non-conserved genes in the region of universally conserved ORFs (Figure 2.3). Furthermore, amplification of environmental DNA using conserved primers flanking the VP1 and VP3 genes have produced very few amplifications indicating the presence of a VP2-like gene (103). This suggests that SSV1 is an anomaly with respect to the VP2 gene, and poses an as yet unanswered question about the necessity and role of the VP2 protein in virus replication.

Integrase Genes

tRNA genes are common integration sites for prokaryotic viruses. This is probably due to the ubiquity of the gene, its slow rate of change, and its slightly palindromic structure (118). This palindrome allows homologous subunits of the integrase protein, each having similar binding affinities, to bind to each side of the attachment site (26). Cleavage assays done with the SSV1 integrase show that the cleavage site is located at the 5' end of the viral attachment site (*attP)* (96), and have shown the integrase protein alone to be sufficient for *in vitro* recombination (70). Along with an off-center cleavage site, no clear palindromes are seen in the SSV1-integrase *attP* sites or that of the other Fuselloviruses (Figure 2.15). Palindromic *attP* sites with cleavage sites located near the center of the palindrome are common features of bacterial *attP* sites (118).

Based on alignments of the *attP* sites, all of which share almost identical 3, sequences (Figure 2.15), it seems that *attP* are evolutionarily constrained (Figure 2.15).

The geographically and genetically distantly related SSV-L1 and SSV-I2 viruses appear to share the same integration site in *S. solfataricus*, the tRNA

30 gene, despite having differences in the integrase gene itself (Figure 2.14). The SSV-K1 *attP* site is in the opposite orientation in the integrase gene relative to all other known Fuselloviral *attP* sites, yet the sequence is very similar to the SSV-I3 and SSV-I4 *attP* site (Figure 2.15). SSV-K1 is thought to be more promiscuous than the other tested SSVs in its ability to integrate into multiple tRNA genes with up to four mismatches within its 49 base pair *attP* site (134), however, a thorough study of all Fusellovirus integration sites has not been undertaken. Presumably, inversion of the SSV-K1 *attP* site allows the virus to integrate into the genome in the opposite orientation. How the inversion happened is also not known, and puzzling as the *attP* sites are the only sites of homology between the 5' region of the SSV-K1 integrase and the same regions in SSV-I3 and SSV-I4 integrase. This suggests that the *attP* sites of theses viruses are under very different evolutionary constraints that the gene they reside in.

Other Conserved Genes

While the majority of the universally conserved ORFs in Fuselloviruses have unknown function, several have clues to possible function. ORFs A-82, A-92, C-102a and their homologues all have predicted membrane spanning domains (Table 2.2). Genes of similar size with similar membrane spanning domains are found in small multidrug resistance proteins (76), and viral and membrane fusion proteins (101).

Summary

These analyses suggest that the Fuselloviruses likely have conserved transcription cycles, especially in the later part of transcription when the most conserved parts of the genomes are expressed. It is also likely that the Fuselloviruses have a conserved location of the origin of replication but not in the origin binding proteins. About half of each of the viral genomes contain non-universally conserved ORFs. The function of these ORFs are remain a mystery, however the large number of repeat elements and evidence of gene movement found in the non-conserved region suggests that this area is frequently recombined. This is supported by the observation that ORFs in this area are very similar to ORFs in one or two other Fuselloviruses and other viruses and plasmids (Table 2.2).

In pair-wise comparisons of the conserved genes a positive correlation between the amount of genetic change and the geographic distance between the locations of virus isolation is seen, indicating that these viruses are limited in their spread and that this spread seems to be a gradual “island hopping” rather than a rapid and diffuse dispersal. This observation may be particularly useful in modeling the spread of viruses to distant hot springs as well as determining the age and history of unstudied hot springs. These results contrast with biogeography studies of viruses of Sulfolobus in Yellowstone

National Park by Snyder (103) and agree with the results of the studies of the biogeography of Sulfolobus strains around the world by (115).

Methods

Isolation of SSV-I3

The SSV-I3 virus was isolated from the Krisuvik solfatara in Southeastern Iceland in the summer of 1996 (6). The hot spring was above 90°C and below pH 4. Water and sediment samples were enriched for Sulfolobales by adding a standard *Sulfolobus* media similar to that used previously (20) and screened for virus production as described previously (6). Viral genomes were extracted as in (106). The viral genomic DNA was cut into 4 fragments using the restriction enzyme EcoRI, inserted into a pUC based vector, and replicated in *E. coli* using standard methods (93). Sequencing was done using an Applied Biosystems Big Dye Terminator sequencing kit and sequenced on an ABI 3700 sequencer using the manufacturer's protocols.

Isolation of SSV-L1

The SSVL virus was isolated from a thermal spring in The Devils Kitchen area of Lassen Volcanic National Park in California. The hot spring had a pH 1.8 and a temperature of 94°C. Water and sediment samples were enriched and the virus was isolated as described above. Viruses were detected as described previously (Stedman et al., 2003). Three fragments of the viral genome were amplified from conserved sites in the virus genome. Forward and reverse primers were made using the sequence and complement sequence of the three primers listed in table 2.4. PCR amplifications used LIPCR as described in Chapter 3. Temperature profiles for the amplification were as follows: an initial denaturation at 98°C for 3 minutes, all subsequent denaturations at 98°C for 15 seconds, annealing at 63°C for 15 seconds, and extension at 72 °C for 6 minutes. Reactions were cycled 30 times with a final 6-minute extension at 72°C. These amplicons covered the entire virus sequence and were cloned using the TOPO zero blunt kit (Invitrogen) then transformed into *E. coli* chemically competent DH5α cells using standard methods (93).

Table 2.4 SSVL Amplification Primers

Name	Sequence
Univ7 F	ATT CAG ATT CTG WAT WCA GAA
Univ8 R	TCS CCT AAC GCA CTC ATC
Univ3 F	CAA TCG CCA TAG GCT ACG G

Sequence data was assembled using Sequencher (Gene Codes Corporation) and the Lasergene package (DNA Star). Annotation was done using Artemis release 9 (92). ORFs were identified manually as sequences with a start and stop codon that were a minimum of 50 codons in length. For comparison, the genome was compared to the protein database and to the known Fusellovirus genes using tBLASTn (2). The program FgenesV, by

Softberry Inc. (Mount Kisco, NY.), was also used to validate these predictions. With the exception of minor differences these methods gave the same results.

Trans Membrane Predictions

Trans membrane predictions were tested on all ORFs using TMHMM program in Biology Workbench (http://workbench.sdsc.edu/), which uses a trained hidden Markov model to predict the location and orientation of trans membrane alpha helices (104).

Biogeography

All protein and DNA alignments were made with Clustal X (108) and edited manually with Seaview (38). To produce the concatenated ORF alignment, individual protein sequences were aligned with Clustal X with default settings (108) and edited manually with Seaview (38). These alignments were concatenated manually in Word (Microsoft Corporation). A distance matrix was made from this concatenated alignment using Protdist and DNAdist respectively in the Phylip program Ver. 3.66, (33). The Jones-Taylor-Thorton and Kimura model of evolution were used respectively. All other settings were left to default. Whole genome alignments were produced using all Fusellovirus sequences starting from the stop codon of the VP3 gene. These sequences were aligned Clustal X with default settings, edited manually with Seaview (38) For degapped nucleotide alignments the whole genome alignments were degapped with Clustal X. Measuring geographic distances was done using Google Earth version 2.0. Graphing was performed and R squared values were calculated using Excel (Microsoft Corporation). P values were obtained using the Mantel test with the TFPA program developed by Mark Miller (http://bioweb.usu.edu/mpmbio/index.htm) and used 999 replicates.

AttP *site Prediction*

AttP site predictions were made by first aligning all known tRNA and snoRNA genes in the *S. solfataricus* genome. Integrase genes were then aligned to the RNA alignments using Clustal X (108) and edited manually with Seaview (38).

Other Bioinformatic Tools

Repeat sequences were identified using bl2seq (107) by decreasing the word size to 7, the gap x dropoff to 20 and increasing the reward for a match to 9. Tandem repeats were identified using the program Tandem Repeats with expect set to 30 (14). Only repeats with e values above 1x10^{13} were included.

ORF comparisons for the genome maps (Figure 2.10) and gene table (Table 2.2) were made by comparing all Fuselloviruse ORF protein sequences to the Genbank non-redundant database using pBLAST (2). All settings were left to default. To compare the two virus genomes not in the database (SSV-L1 and SSV-I3) to each other bl2seq (107) was used with default settings. Any Fusellovirus hits from the Blast searches were considered to be matches, however the lowest Fusellovirus match had an e value of $2x10^{-2}$.

GC skew and Pyrimidine excess were mapped using GraphDNA (Viral Bioinformatics Resource Center, University of Victoria). To align putative *ori* sites, sequence alignments were edited manually in MS Word.

Individual genes' Neighbor-Joining trees were constructed using Clustal X and viewed with Dendrograph (46).1000 bootstraps were used.

All plasmid maps were drawn with the online program Savvy version 0.1 by Malay K Basu (http://www.bioinformatics.org/savvy/) and edited in Adobe Illustrator.

Chapter 3: LIPCR: A Technique to Create Site-specific Knockouts and Insertions in Fusellovirus Genomes

Abstract

Long Inverse PCR (LIPCR) was developed as a technique to insert and remove genetic material from double stranded circular DNA viruses. This method can create deletion mutants by amplifying and circularizing part of a genome. Products of LIPCR can be also be ligated to an insert to create insertion or replacement mutants. As the amplification area can be controlled by the design of primers this method is site specific.

The utility of LIPCR is demonstrated by the creation of two novel Fusellovirus shuttle vectors capable of replication in *E. coli* and *S. solfataricus.* An insertion mutant was created by amplifying 15.5 kilobase pair of the 18.5 kilobase pair genome of the original SSV1 Fusellovirus shuttle vector and replacing the 3 kilobase pair section with different DNA. A deletion mutant was created by amplifying a 17.5 kilobase pair fragment of the insertion mutant genome and circularizing the amplicon by ligation. Both constructs are shown to be capable of replication in both *E. coli* and *S. solfataricus.*

The following is expanded from a technical note written for BioRad (appendix A).

Introduction

The SSV1 virus is the type species of the Fusellovirus family, a widespread but poorly understood virus family capable of replication in *Sulfolobus solfataricus* (95). The 15.5 kilobase pair , double stranded circular DNA genome of SSV1 contains 34 ORFs, only four of which encode proteins of known function (75).

To begin to understand the function and necessity of these ORFs, mutagenesis has been used previously in one published study in SSV1 (105). In this study insertion mutants were created by the insertion of the pBluescript plasmid into restriction sites in the complete SSV1 genome. Using this method 2 ORFs, ORF E178 and E51, were shown to be tolerate the insertion of pBluescript and replicate in *S. solfataricus.* The plasmid pKMSD48, containing pBluescript inserted into ORF E178, was further shown to replicate stably as a plasmid in *E. coli* and to replicate, spread, and produce particles that are indistinguishable from the wild type virus in *S. solfataricus* (105).

To further understand the function and necessity of the remaining ORFs in the Fuselloviruses, a method to create site-specific deletions and insertions that was not dependent on restriction sites was needed. Here I describe long inverse PCR (LIPCR) in the SSV1 genome. Insertion mutagenesis is demonstrated by replacement of the pBluescript plasmid in pKMSD48 with a plasmid conferring a different antibiotic resistance. Deletion mutagenesis is demonstrated by the complete deletion of a gene in the shuttle vector genome. A graphical overview of this method is shown in Figure 3.1.

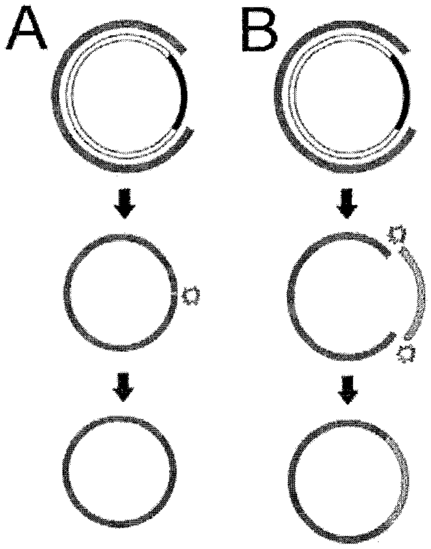

Figure 3.1: LIPCR overview. (A) Deletion: Amplification of the entire template (inner circle) except for the region to be deleted (in black) The grey LIPCR amplicon is ligated (*) in the second step to form the circular genome lacking the deleted area. (B) Insertion: Part or all of the circular template (inner circle) is amplified (grey outer circle). This amplicon is ligated in the second step to an insertion sequence (hatchmarked) to form the final circular genome.

Results

Long PCR Optimization

Several different DNA polymerases ranging in processivity and fidelity were tested for their ability to amplify the entire 15.5 kilobase pair SSV1 virus genome using pKMSD48 as template using M13 primers, which have binding sites in the pBluescript portion of the shuttle vector. DNA polymerases tested were *Taq,* Vent, Deep Vent, and Phusion (all purchased from New England Biolabs) and Pfu (purchased from Stratagene). For each DNA polymerase the manufacturers' guidelines were followed for the use of buffers and dNTP concentrations. Amplification conditions for each DNA polymerase were optimized using annealing temperatures spanning 50-72°C, Mg concentrations spanning 1-5 mM, and template concentrations spanning 30 femtomolar to 30 picomolar. Phusion polymerase was the only polymerase found to produce an amplicon greater than 10 kilobase pairs in my hands (Figure 3.2, Panel A, Lane 1, and Panel B, Lane 1).

Full-length amplification of the SSV1 genome (without smaller nonspecific bands) using Phusion polymerase was found to require more precise temperature profiles than standard, short fragment PCR. This was observed when using different Applied Biosystems Geneamp 2700 thermocyclers, where it was necessary to optimize annealing temperatures separately in each machine, presumably due to minor variations between thermocyclers. Using an MJ research Dyad thermocycler produced more consistent full-length amplification in different heating blocks, which did not require individual optimization. For each primer set, template concentration was optimized using five-fold dilutions of template, spanning approximately 30 femtomolar to 30 picomolar, to find concentrations where full-length amplification occurred without smaller nonspecific products. Optimization was needed for each DNA preparation and it was often found that full-length amplicons were only produced within a template concentration range less than an order of magnitude (data not shown). dNTPs were found to allow successful reactions only at concentrations of 200 μM (data not shown). Reaction volumes also affected the efficiency of the reaction, with the most reproducible results seen in small (20μl) reactions (data not shown).

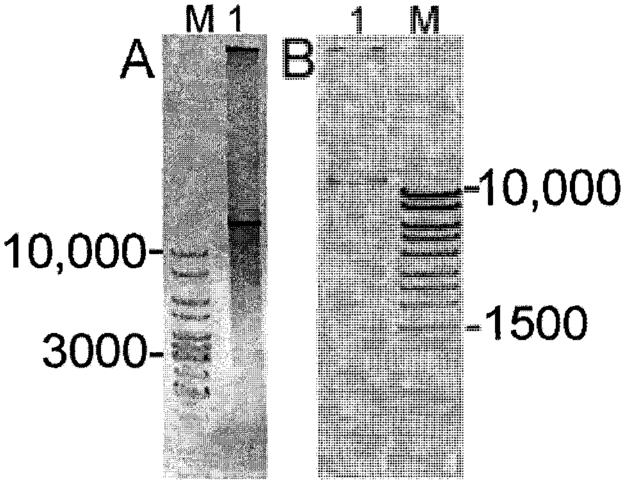

Figure 3.2: LIPCR generates large products **(A)** LIPCR of 15.5 kilobase pair SSV1 genome from pKMSD48 using M13 Forward and Reverse primers **(B)** Amplification of the 17.5 kilobase pair pAJC96 using Del Forward and Reverse primers. Markers (M) are Fermentas Generuler 1 kb ladder in panel A and Massruler 1kb ladder in panel B. Sizes of representative bands are indicated in base pairs.

Gene Replacement with LIPCR

The full length SSV1 genome was amplified by LIPCR using M13 forward and reverse primers. The pKMSD48 shuttle vector purified from *E. coli* was used as template (Table 3.1). Annealing temperature was optimized using an MJ Dyad Gradient thermocycler using 20 μl reactions with 2°C gradations in temperature from 54-72°C. A 62°C annealing temperature was optimal as it produced a single LIPCR product with a length in excess of 10 kilobase pair when analyzed by agarose gel electrophoresis, (Figure 3.2 Panel A, Lane 1). In most cases the optimal temperature for the primer sequences was 7-10 degrees higher than those calculated by the manufacturer (Integrated DNA Technology).

After gel purification, the LIPCR product was ligated into a pCR TOPO Blunt II plasmid conferring Kanamycin resistance. 4 μl of the 6 μl reaction was transformed into the chemically competent StAble 3 strain of *E. coli* (Invitrogen). 36 hours after plating the entire transformation mixture approximately 75 colonies were observed on LB plates containing 35 μg per ml Kanamycin, a transformation efficiency of approximately 10^4 colonies/μg. Plasmid preparations from 3 ml cultures grown from 5 of the smallest colonies, isolated using alkaline lysis, and digested with EcoRI, were visualized by agarose gel electrophoresis. One of these five colonies contained a plasmid appearing to contain the predicted fragment sizes of 5166, 3952, 3498, 3193, 2439 and 2143 base pairs shown in Figure 3.3 Lane 3. This plasmid was named pAJC97, and was used as template for the LIPCR mediated gene removal. The other four preparations contained bacterial plasmids that lacked the insertion of a full-length LIPCR product.

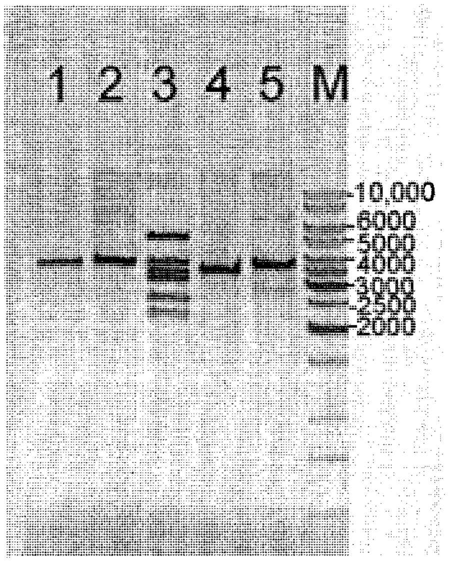

Figure 3.3: Confirmation of LIPCR insertion: EcoRI digest of five plasmids generated by the ligation of the SSV1 LIPCR product to a TOPO vector (Invitrogen). Expected bands are 5166, 3952, 3498, 3193, 2439 and 2143 base pair in size. Lane 3 contains the expected bands, the plasmid from this culture is named pAJC97. Other lanes lack a full length PCR product. The marker is Fermentas 1kb Generuler. Sizes in base pairs of representative bands are marked.

Gene Deletion with LIPCR

Deletion mutagenesis was performed by LIPCR using pAJC97 as template and using primers constructed so that the 3' ends of the oligonucleotides faced away from the area to be deleted, the viral integrase gene (Del Forward and Del Reverse, Table 3.1). This allows amplification of the entire shuttle vector except for the area to be deleted. NarI and SexAI restriction sites were added to the primers to allow directional cloning in future applications. Annealing temperature was optimized in an MJ Dyad Gradient thermocycler using 20 μl reactions with 2° C gradations in temperature ranging from 54-72°C. A 65°C annealing temperature produced a single product larger than 10 kilobase pairs shown in Figure 3.2 Panel B.

Terminal phosphates were added to the LIPCR product and the product was circularized. The product was then transformed into StAble 3 cells as the pAJC97 construct. The transformation efficiency was approximately 10^4 colonies/μg. After 36 hours growth the smallest colonies on the plate were picked, grown in LB with 35 μg / ml kanamycin, from which plasmids were purified and digested with EcoRI. Of the 8 colonies picked 3 appeared to have the expected sized fragments when separated by agarose gel electrophoresis. One of these was chosen to be pAJC96. As in the pAJC97 screening the remaining plasmid preps contained plasmids smaller than the full-length shuttle vector and were assumed to be partial-length amplifications (data not shown).

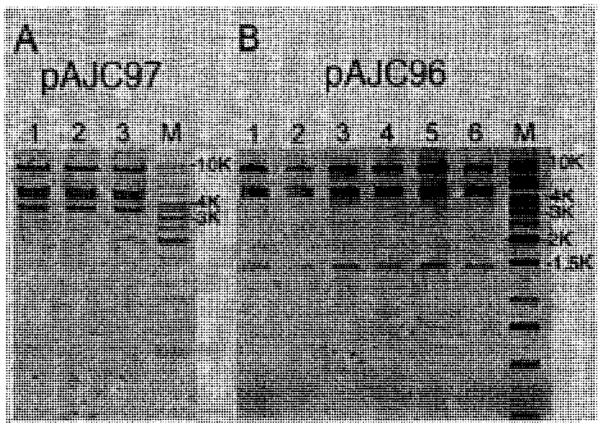

Figure 3.4: The LIPCR-created shuttle vectors are capable of replication in *E. coli* and *S. solfataricus*. Three and 6 independent clones of (A) pAJC97, and (B) pAJC96 respectively, were isolated from *E. coli* transformed with double stranded circular DNA from pAJC96 and pAJC97 transformed *S. solfataricus*. Plasmids were digested with HindIII and appear to contain the predicted fragment sizes of 7807, 4632, 4200, and 3415 for pAJC97, and 7807, 4632, 4200, and 1434 for pAJC96. The rightmost lane in each gel (M) contains a Fermentas 1 kilobase pair Generuler ladder. Sizes of representative bands are marked in base pairs.

Confirmation of Accurate Replication of the LIPCR Products

To determine if deleterious changes had been introduced by LIPCR, the LIPCR-produced shuttle vectors pAJC97 and pAJC96 were transformed separately into *S. solfataricus* strain P2. Both produced halos of growth inhibition when spotted on lawns of uninfected *S. solfataricus,* suggesting the ability of pAJC96 and pAJC97 replicate and spread as virus. Details of the replication and spread of pAJC97 and pAJC96 are described in Chapter 4.

To confirm that pAJC96 and pAJC97 could replicate in *S. solfataricus*, double stranded circular DNA was isolated from *S. solfataricus* cultures grown from freezer stocks of the original strains transformed with pAJC97 and pAJC96. Each plasmid DNA was transformed into *E. coli* StAble 3 cells. Plasmid DNA was isolated from six randomly chosen colonies from the transformation of pAJC97 and three randomly chosen colonies of pAJC96, all of which produced the expected bands when digested with EcoRI and separated by agarose gel electrophoresis (Figure 3.4).

Further support for LIPCR replication fidelity was generated by sequencing of 3 kilobases of the second LIPCR product, pAJC96. Ligation junctions had the expected sequences, and the other sequence data showed no changes from published SSV1 or bacterial plasmid sequences with the exception of the changes added by Del Forward and Del Reverse primers.

Discussion

The successful construction of the shuttle vectors pAJC97 and pAJC96 demonstrate that LIPCR is an effective way to insert and remove specific DNA sequences in Fusellovirus genomes. In theory, this method can be used with any plasmid or virus with a circular genome with appropriate primers and optimization. The fidelity of replication with Phusion polymerase appears to be high based on three kilobases of error free sequence in the product of two rounds of LIPCR, and by the ability of the DNA constructs produced by the LIPCR method to replicate in both hosts. Complete sequencing of an LIPCR product will be necessary to fully assess the accuracy and fidelity of this method.

Several drawbacks do exist with this procedure. As LIPCR takes 6-8 hours per experiment to run and requires optimization with each new set of primers and each new template used, a considerable amount of time must be spent in optimization. Cloning and transformation with large PCR products is inefficient. Using the StAble 3 cells, the efficiency is 3 to 4 orders of magnitude lower that of supercoiled pUC18 plasmid (data not shown). This can pose a serious problem if smaller nonspecific products are produced during LIPCR as these smaller products transform with a higher relative efficiency. Without gel purification, transformation of ligated LIPCR products that showed a single high molecular weight band when separated by agarose gel electrophoresis produced orders of magnitude more clones, nearly all of which contained plasmids with less than full-length amplifications (data not shown).

The amount of template required for amplification of the full-length product is also quite high and the amplification products yields are relatively low in comparison to traditional PCR. The 3 and 6 picomols of template used in the pAJC96 and pAJC97 amplifications is enough to be visible as very faint bands when entire 20 μl reactions are separated on agarose gels and stained with ethidium bromide. This amount of template is orders of magnitude greater than amounts needed to amplify small (250 or 1100 base pairs) amplicons with either Phusion or Taq polymerase (data not shown). The large amount of template needed in LIPCR may produce a high percentage of transformations containing the original template if the amplicon is co-purified by agarose gel electrophoresis.

Possible solutions to this problem would be to purify the LIPCR template from a strain of *E. coli* capable of Dam methylation, amplify the template with LIPCR, and prior to gel purification digest the template with DpnI, a restriction enzyme that cleaves only Dcm methylated DNA thus eliminating (or greatly reducing) the amount of template used for transformation. A similar method is used in kits such as Stratagene's Quickchange kit to remove templates after mutagenic PCR. Another possibility would be to amplify the product in two halves, gel purify each half, ligate them together to form the final product, and screen for clones that contain both halves and have been ligated together in the proper orientation using techniques such as restriction fragment length polymorphism. Finally, as demonstrated in the construction of pAJC96, the antibiotic marker can be changed by inserting a different bacterial plasmid into the shuttle vector.

While Phusion polymerase was the only polymerase found to successfully produce a full-length amplicon, other polymerases were not as extensively tested and may work as well or better with further optimization. With the introduction of new highly possessive polymerases and polymerase cocktails it is likely that equally effective, or possibly more effective polymerase may be available for the LIPCR procedure described here.

Materials and Methods

Virus Shuttle Vector Construction and Isolation

The pKMSD48 shuttle vector, a fusion between the bacterial plasmid pBluescript SK+ and the SSV1 virus genome, was provided by Kenneth Stedman (105). Shuttle vector genomes were purified from *E. coli* using alkaline lysis essentially as described in (17). Briefly, 3 ml of cells were resuspended after centrifugation at 14,000 x g for 2 minutes in 100 μl of a solution of 50 mM Glucose, 25 mM Tris/HCl pH 8.0, and 10 mM EDTA. 200 μl of 0.2 M NaOH and 1% (w/v) SDS was added and the tube mixed by inversion. Finally 150 μl 3 M potassium acetate solution was added and mixed well. The mixture was centrifuged at maximum speed in a tabletop centrifuge for 20 minutes after which the supernatant was extracted three times with 1 volume (450 μl) of phenol/chloroform/isoamyl alcohol (25:24:1). The DNA was precipitated by the addition of 0.8 volumes (360 μl) of isopropanol. The precipitate was washed twice with 1 ml of 70% ethanol, dried, and dissolved in 50 μl sterile water.

Primer Design

Standard M13 forward (-20) and M13 reverse (-27) primers were used unchanged for the construction of the pAJC97 insertion plasmid. Primers for the pAJC96 deletion amplification (Del Forward and Del Reverse) were designed to remove the complete integrase gene from the virus and allow the directional cloning of different replacement genes (Table 3.1). Primers were designed so that their 5' end flanked the gene to be removed. The length of the primer was extended in the 3' direction for approximately 25 bases, and stopped when a GC clamp of at least one base was present and when predicted annealing temperatures were between 55-60˚C. See Table 3.1 for sequences.

Primer sequences were checked for hairpins and other secondary structure using Mfold http://www.bioinfo.rpi.edu/applications/mfold/old/dna/ with Na^+ concentrations set at 50mM and Mg^{2+} concentrations set at 0 mM. Primers' 3' ends were changed if the secondary structure had a predicted melting temperature (Tm) of fewer than 15°C below that of the duplexed or hairpinned primers. The primer sequence was changed to insert restriction endonuclease cleavage sites for directional cloning as shown in Table 3.1. Final Tm predictions were calculated with Hyther, a program that predicts of nucleic acid hybridization thermodynamics taking into account mispairing http://ozone2.chem.wayne.edu/Hyther/hytherm1main.html. The Tm was adjusted by increasing or decreasing the length of the 5' end of the primer to allow the predicted Tms of both primers to be within 3°C of each other and between 55 and 60°C.

Table 3.1: Primers used in PCR. Bold letters show restriction sites. Italics show start codon of removed gene.

Name	Sequence	Tm
M13 Forward	5'-GTAAAA CGACGGCCAGT-3'	53.0*
M13 Reverse	5'-CAGGAAACAGCTATGAC-3'	47.3*
Del Forward	5'-CGTCTTATCTTTCGT*CAT*TTC**ACCTGGT**ACTATTATGG-3'	58.3*
Del Reverse	5'-GGGGTCTGACA**GGCGCC**GTATCACTATC-3'	55.4*

*All Tms calculated as described above.

Temperature Optimization and PCR Conditions

Amplification of the SSV1 viral genome using pKMSD48 template purified from *E. coli* used M13 Forward and Reverse primers (Table 3.1). All LIPCR amplifications for cloning were carried out in an MJ Research Dyad thermocycler. Temperature calculations were estimated by the thermocycler and all reactions were carried out in 20 μL volumes.

The conditions used for LIPCR amplification followed MJ Research's recommendations (Dennis Prosen, personal communication). This included an initial denaturation at 98°C for 3 minutes, and subsequent denaturations at 98°C for 15 seconds. Annealing was 15 seconds long. Optimal annealing temperatures for LIPCR products used to create pAJC97 and pAJC96 were 62 and 65 °C respectively. Extension was 8 minutes long at 72 °C. Reactions were cycled 30 times with a final 8-minute extension at 72°C. Materials for LIPCR are described in Table 3.2.

Table 3.2: LIPCR materials

Reagent	pAJC97 LIPCR	pAJC96 LIPCR
Buffer	1X HF Buffer	1X HF Buffer
dNTPs	0.2mM/base	0.2mM/base
Template	3pM	6pM
Forward primer	M13 F, 250nM	Del F, 250nM
Reverse primer	M13 R, 250nM	Del R, 250nM
Polymerase	0.02 U/μl, Phusion	0.02 U/μl Phusion

Ligation

The PCR products of both LIPCRs were gel purified by cutting the full-length band from a 0.8% low-melt agarose gel stained with ethidium bromide. The agarose was digested with Beta Agarase (New England Biolabs) following the manufacturer's instructions and resuspended in sterile water. The final amounts of products were roughly quantified by ethidium bromide fluorescence relative to known DNA standards. For the LIPCR product that would become pAJC97, 8 ng of DNA was cloned into Invitrogen's TOPO Blunt II ® vector using the Zero Blunt PCR Cloning Kit following manufacturer's instructions.

Circularization of the LIPCR amplicon that would become pAJC96 was done using 500 ng of gel purified LIPCR product that was added to a reaction mixture of 1X T4 Ligase buffer containing 1mM ATP, 2 μl of PEG_{4000} (Sigma), and 10 units of Polynucleotide Kinase (New England Biolabs) in a total volume of 40 μl. The reaction was incubated at 30°C for 30 minutes then denatured at 65°C for 20 minutes. After denaturation 20 units of T4 Ligase (New England Biolabs) were added and incubated at 16 degrees for 4 hours.

Transformation into Sulfolobus

Shuttle vectors purified from *E. coli* were transformed by electroporation into the host *Sulfolobus solfataricus* as described previously (106). Briefly, fresh 50 ml cultures of *S. solfataricus* were grown to an OD600 of 0.2 in a standard Sulfolobus media similar to that used previously (20). Cells were washed 5 times with one volume of ice-cold sterile 20mM sucrose and resuspended in 400 μl 20 mM sucrose. Transformation was done using electroporation at 15 kV/cm, 400 ohms and 25 μF with 1 μg of DNA purified from *E. coli* as described above. After electroporation cells were immediately placed in 1 ml of 80°C growth medium and diluted after 1 hour into 30 ml of growth medium at 80°C.

Detection of Virus Production in S. solfataricus

Virus production in *S. solfataricus* from amplified viral genomes was detected in two ways. First viral halos of growth inhibition were observed by spotting *S. solfataricus* that was transformed with the shuttle vector onto lawns of non-infected *S. solfataricus* (105). Secondly, purification of shuttle vector genomes from infected Sulfolobus strains and restriction endonuclease digestion showed the presence of reproducing virus. Agarose gel electrophoresis was done using standard molecular methods (91).

Sequencing

Sequencing reactions used Big Dye Terminator readymix solutions version 3.1(Applied Biosystems) according to the manufacturer's instructions. Each reaction used 4 *μ*l of BDT, 1.6 pmol of primer, and 400 ng of template. Sequencing was performed by the KECK genomics center at Portland State University and at the Oregon Health and Science University's sequencing core. The template for sequencing reactions was pAJC96 DNA purified from *E. coli.*

Restriction fragment size was calculated with pDraw32 (AcaClone software).

Chapter 4: SSV1 Viral Integrase is not Essential

This chapter is based on the following publication:

Adam J. Clore and Kenneth M. Stedman "SSV1 integrase is not essential"

Virology 2007 Apr 25; 361(1):103-11. Epub 2006 Dec 18.

Abstract

All known Fusellovirus genomes contain an integrase gene. The SSV1 integrase gene product has been demonstrated to be sufficient to recombine viral and host attachment sites in vitro. The gene conservation suggests that integration is a necessary function for virus replication. To test this hypothesis, three Sulfolobus-*E. coli* shuttle vectors were constructed using the LIPCR method described in chapter 3. These shuttle vectors contain the SSV1 genome with either a complete integrase gene, a partial integrase gene lacking catalytic residues, or no integrase gene. The ability of all of the shuttle vectors to replicate and spread in *Sulfolobus solfataricus* is demonstrated. The vector lacking the entire integrase gene does not integrate into the SSV1 *attA* site, while both vectors containing either the partial or complete integrase gene appear to integrate. Competition assays suggest that vectors lacking the integrase gene may not replicate or infect as well as those containing it.

Introduction

The integrase gene is a unifying feature present in all Fusellovirus genomes (see chapter 2) and is the only gene showing distinct homology to genes found outside of this virus family (96). The integrase gene product belongs to the highly conserved tyrosine recombinase protein family, a family that has homologous proteins found in viruses that infect all domains of life and in their eukaryotic and prokaryotic hosts (32).

Members of the tyrosine recombinase protein family span a large range of functions, such as topoisomerases, resolvases, restriction endonucleases, and regulators of gene expression (73). The hallmark trait of tyrosine recombinases is a tetrad of conserved, non-consecutive amino acid residues Arg-His-Arg-Tyr present in the catalytic domain. The basic Arg-His-Arg residues are involved in coordinating the DNA so that the scissile phosphate of the DNA backbone is aligned with the active site tyrosine, the fourth conserved residue (32). Monomers of integrase have one active site per peptide, therefore four monomers are needed to completely recombine two strands of DNA (109). Recombination of the strands was demonstrated to occur by the formation and resolution of a Holliday junction when intermediates of Cre recombinase and Lambda integrase were crystallized (18, 24).

Two subtypes of integrases have been observed in the Crenarchaea and their extrachromosomal elements. One subtype is found in plasmids and some viruses that have a viral attachment site (*attP)* outside of the integrase gene, and another found in the Fuselloviruses and related extrachromosomal elements that contain an *attP* site located within the integrase gene (97).

Fusellovirus Integrase Sequence Similarities

The Fusellovirus integrase genes share weak sequence similarity to the well-characterized phage integrases and eukaryotic recombinases (73). These similarities cluster in the catalytic domain located in the C-terminal end of the protein and group near the active site in three conserved boxes. There is approximately 40% identity at the amino acid level within these boxes compromising approximately 80 residues between the SSV1 integrase and its most similar characterized integrases, those belonging to the Xer D family of recombinases (59)(Figure 4.1).

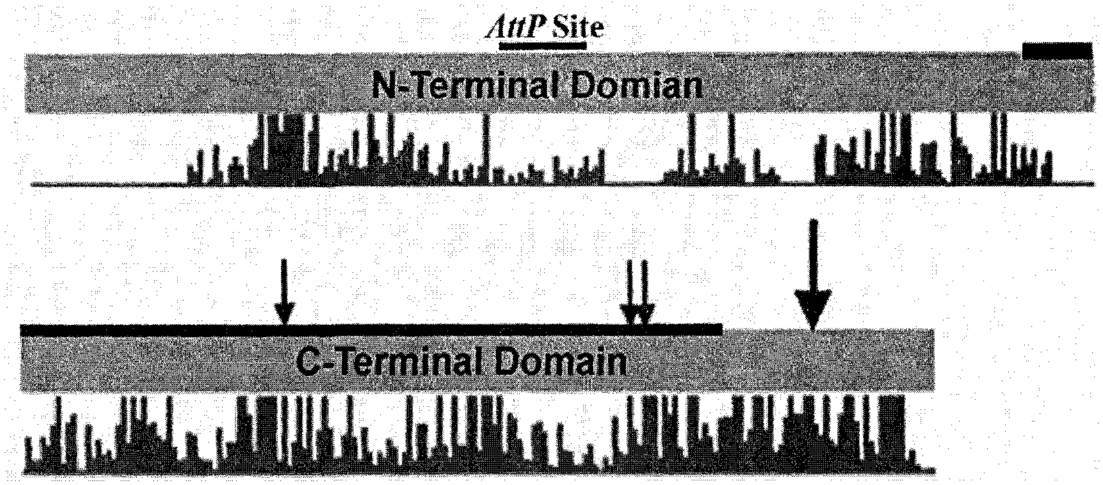

Figure 4.1: Conservation of Fusellovirus integrase protein domains. The location of the attachment site is labeled *attP*, with a thin black overline. The positions of the predicted catalytic tetrad are indicated with arrows, with a large arrow for the active site tyrosine. The histogram shows weighted homology of aligned amino acids of integrases from all 5 published Fuselloviruses based on a BLOSUM matrix. The thick black line indicates the portion of the integrase gene deleted in the δInt shuttle vector.

SSV-1 Viral Integration

The Sulfolobus Spindle-Shaped Virus 1 (SSV1) is the type species of the family Fuselloviridae and was isolated from *Sulfolobus shibatae* strain B-12. *S. shibatae* was isolated from a hot spring in Beppu, Japan. The SSV1 virus particle is composed of at least three structural proteins and 15,465 base pairs of covalently closed circular, double stranded DNA (75). Like all other Fuselloviruses identified, this species infects and replicates in *S. solfataricus* and other closely related *Sulfolobus* strains but not in more distant relatives such as *Sulfolobus acidocaldarius* (95, 117).

SSV1 viral integration occurs specifically in the 3' end of the host arginine CCG tRNA gene (tRNA 30 in the *S. solfataricus* P2 genome). Integration is in the same tRNA in both *S. shibatae* (123) and *S. solfataricus* (95). In both cases the homologous host and viral attachment sites (*attA* and *attP* respectively) extend to the 3' end of the tRNA gene, so the integration does not result in a change in the sequence of the tRNA. *In vitro* studies show the integrase is capable of transferring the phosphodiester bonds between short double-stranded DNA fragments containing *attA* and *attP* sequences in the absence of host or viral accessory proteins (70) (96).

The tyrosine recombinase tetrad of amino acid residues is present in the SSV integrases, however, as with some other divergent integrases, the histidine residue of this tetrad has been replaced with a lysine in all sequenced sequenced SSV integrase genes (73). Based on alignments and secondary structure predictions, several amino acid residues, including the conserved tetrad, are thought to be homologous to those found all other tyrosine recombinase family integrases. Changes of any of these residues in the SSV1 integrase limit or abolish the ability to recombine *attP* and *attA* containing segments of linear double stranded DNA *in vitro* (59).

SSV attP Site Location in Integrase Gene

The location of the Fusellovirus *attP* site within the 5' half of the integrase gene causes a gene disruption upon integration. The internal *attP* sites in Fusellovirus integrases are unique with the exception of integrase genes found in some Myxophage (49, 67). The Myxophage Mx8 and Mx9 integrase genes are partitioned upon integration but are distinct from Fusellovirus integrases in that the *attP* site is located approximately 30 codons from the 3' end of the gene. These last 30 amino acids of the Myxophage integrase protein contains none of the conserved residues needed for catalysis, leading to minimal change after the disruption. It is speculated that the truncation may be involved in regulating prophage excision (110). In contrast, the SSV1 *attP* site is located within 66 codons of the 5' end of the gene, upstream of all sequences encoding amino acids shown to be critical for enzymatic activity. How the virus solves this problem of excision, or whether the SSV1 provirus is ever excised, is unknown.

While the functions of the SSV1 integrase protein have been studied *in vitro*, its role *in vivo* is not clear. It is not known if integration is a critical part of virus replication. It is not known if excision of the provirus occurs. Finally, any advantage to the virus of integration is not known. To begin to investigate these questions integrase gene deletions in SSV1 were created using a recently developed long inverse PCR (LIPCR) technique. The resulting shuttle vectors were tested in *S. solfataricus* (Chapter 3).

LIPCR

Long Inverse PCR (LIPCR) was developed as a technique to insert and remove genetic material from double stranded circular DNA viruses. This method can create deletion mutants by amplifying and circularizing part of a genome. LIPCR products can be also be ligated to an insert to create insertion or replacement mutants. This method is site specific the amplification since the region amplified can be controlled by primer design. Details of this method are described in Chapter 3.

Results

Creation of Shuttle Vectors with LIPCR

To investigate the necessity of the viral integrase gene and parts thereof in replication, three *E. coli/S. solfataricus* shuttle vectors were constructed using LIPCR (see Table 4.1 for a list of strains and plasmids). All shuttle vectors are based on the pKMSD48 vector described in (105), and referred to hereafter as the pBluescript vector. The pBluescript vector consists of a pBluescript plasmid inserted into a Sau3A1 site near the 3' end of gene E-178 in the T5 transcript of SSV1. It was shown to replicate stably as a plasmid in *E. coli* and to replicate and spread as a virus in *S. solfataricus* (105). The pBluescript vector was used as template to amplify the SSV1 genome from the M13 primer binding sites in the pBluescript plasmid. This amplicon was ligated into a TOPO PCR Blunt II plasmid to create a vector named pAJC97, referred to hereafter as the wild-type vector. The wild-type vector was used to generate an otherwise identical shuttle vector lacking the entire integrase gene (pAJC96) referred to hereafter as the ΔInt vector (see Figure 4.2).

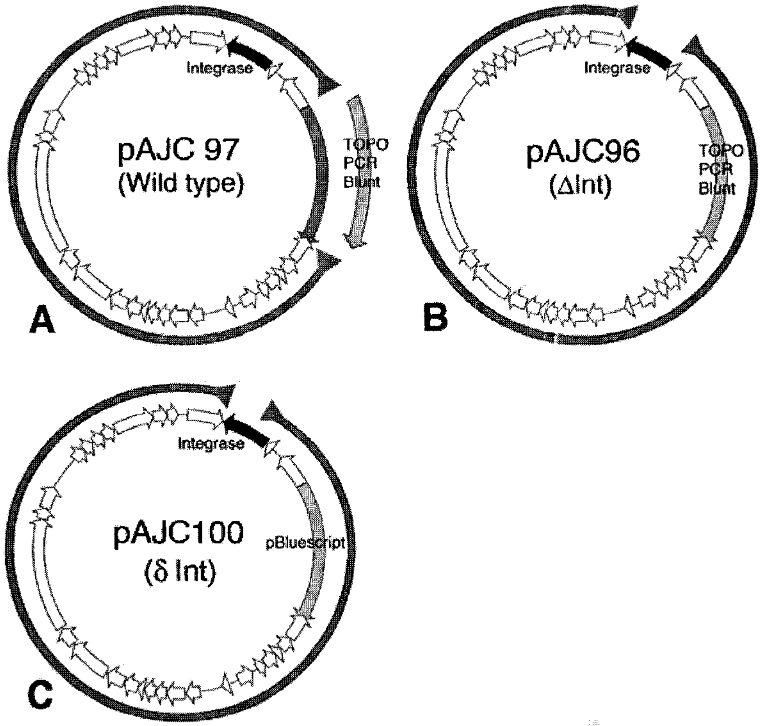

Figure 4.2: Long inverse PCR (LIPCR) amplification strategy to create integrase-lacking shuttle vectors. Grey circles represent amplicons. **(A)** Integrase containing shuttle vector pAJC97 (wild-type vector). **(B)** Integrase lacking shuttle vector pAJC96 (ΔInt). **(C)** The partial integrase containing shuttle vector pAJC100 (δInt).

The pBluescript vector was used as template to construct a third vector lacking most of the 3' portion of the integrase gene thereby removing sequence encoding the conserved catalytic residues Arg-Lys-Arg and causing a frameshift in the sequence coding for the active site tyrosine. The N-terminal domain including the *attP* site remains intact. This partial deletion mutant was named pAJC100 and is referred to hereafter as the δInt vector. Figure 4.1 illustrates the shuttle vector constructions, strain descriptions are listed in Table 4.1. Details of the LIPCR amplification are described in methods.

Integrase Lacking and Containing Vectors Replicate in S. solfataricus

The ΔInt, δInt, and wild-type shuttle vectors were transformed individually into *S. solfataricus* strain P2 by electroporation. The transformed *S. solfataricus* cells were grown in liquid culture for 7 days, each day 2 *μ*l of each culture was placed on lawns of uninfected *S. solfataricus* to assay for halos of growth inhibition, an indication of viral infection (106). Transformed cultures did not produce halos on lawns made daily for the first three days after transformation. However all cultures spotted 96 hours post-transformation and produced halos of growth inhibition (Figure 4.3, Panel A). (Note that in the printed image the faint halo produced by δInt is not visible but was clearly distinguishable by eye).

Virus could be propagated in uninfected *S. solfataricus* by inoculating fresh media from the edge of a halo, or by using the cell-free supernatant of infected cultures (data not shown). These infected cultures were capable of producing halos and viral DNA, indicating that the presence of this halo is the result of an infective agent (data not shown).

To confirm that the halos seen on plates were caused by the presence of the correct virus vector, PCR was used to amplify fragments of the viral genomes surrounding the integrase gene from total DNA isolated from the transformed cultures. Primers used for this amplification were Int F and Int R (Table 4.2). In all cases these amplifications produced the predicted sized fragments of 1953 base pairs for the wild type virus and vector, 1467 base pairs for the δInt vector, and 951 base pairs for the ΔInt vector, indicating the presence of the correct virus in each culture (Figure 4.3, Panel B).

Finally, the particles from infected cultures containing the shuttle vector with the largest deletion, ΔInt, were indistinguishable from the wild type virus in both size and shape when observed with the transmission electron microscope (Figure 4.3, Panel C). Images of wild type and δInt particles were also observed with TEM (data not shown).

Table 4.1: Strains and Plasmids

Strain	Description	Citation
S. solfataricus P2	Wild type	DSM1617
S. shibatae B12	Wild type	DSM5389
Stbl3 E. coli	F– mcrB mrr hsdS20(rB–, mB–) recA13 supE44 ara-14 galK2 lacY1 proA2 rpsL20(StrR) xyl-5 λ– leu mtl-1	Invitrogen
Plasmid/virus		
SSV1	Wild type virus, Fusellovirus type strain	(75)
pKMSD48 (pBluescript vector)	pBluescript II SK+ in SSV1 (Sau3AI selection)	(105)
pAJC97 (wild type vector)	SSV1 portion of pKMSD48, cloned into TOPO PCR Blunt II	This Work
pAJC96 (ΔInt vector)	pAJC97 with integrase gene removed	This Work
pAJC100 (δInt vector)	pKMSD48 with C-term of integrase removed	This Work

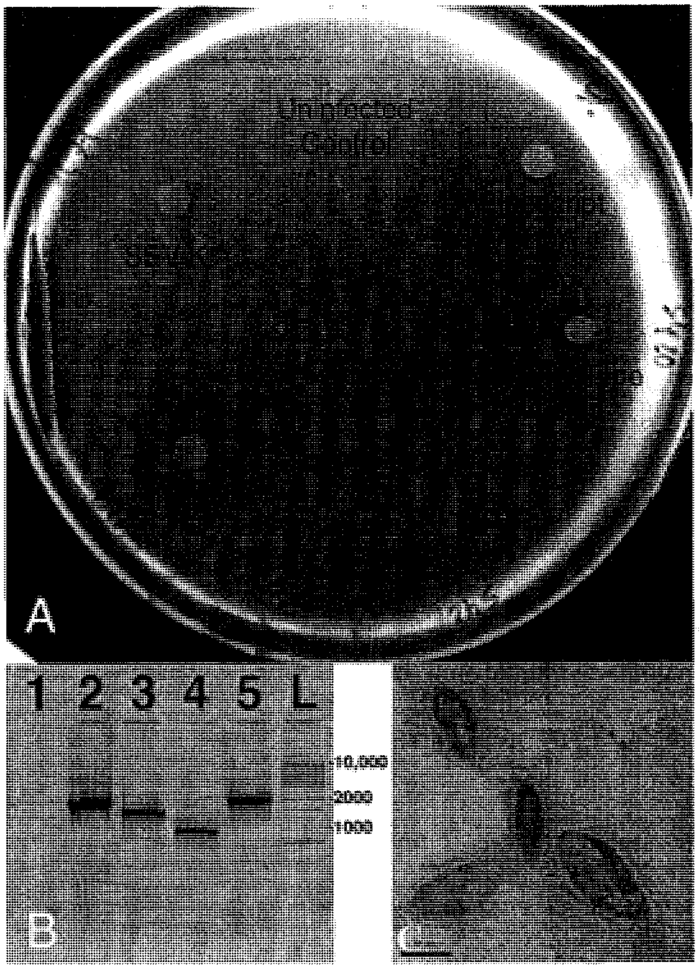

Figure 4.3: The integrase gene in SSV1 is not essential for virus replication. **(A)** Halos of growth inhibition produced by transformed cultures on an uninfected *S. solfataricus* lawn. The virus used for transformation is labeled on the plate below each spot (SSV-K1 is a positive control). **(B)** PCR amplification of total DNAs isolated from *S. solfataricus* transformants using primers Int F and Int R flanking the integrase gene. Lane 1: uninfected *Sulfolobus* culture Lane 2: wild-type vector transformation. Lane 3: δInt vector transformation. Lane 4; ΔInt vector transformation: Lane 5: wild-type SSV1 virus transformation. L GeneRuler 1 kb DNA ladder (Fermentas). **(C)** Negatively stained transmission electron micrograph of culture supernatant from *S. solfataricus* transformed with ΔInt vector. Bar represents 50 nm.

ΔInt does not Integrate at the attA *site; δInt and the Wild-type Vector Appear to Integrate at the* attA *site*

While removal of the integrase gene does not prevent viral replication, it is expected to abolish the ability of the virus to integrate. To test if the integrase lacking shuttle vectors integrate, a PCR based assay was developed. The assay uses primers that flank the SSV1 *attA* site, the argininyl (tRNA 30) and consist of a host-specific primer, Provirus 1, and virus-specific primer, Int F, (Table 4.2). This method is illustrated schematically in Figure 4.4.

Three integration assays were performed. The first test used total DNA extracts from transformed cultures as template (the same DNA used as template for the PCR shown in Figure 4.3, Panel B). No PCR product was observed using the DNA isolated from uninfected cells or cells transformed with the ΔInt vector, indicating that there was no integrated virus. A product of the predicted size of 2241 base pairs was seen using the DNA isolated from cells infected with the wild-type vector, the wild-type virus. Surprisingly, a similar sized PCR product was observed with DNA isolated from cells transformed with the δInt vector, indicating an integrated virus (Figure 4.5).

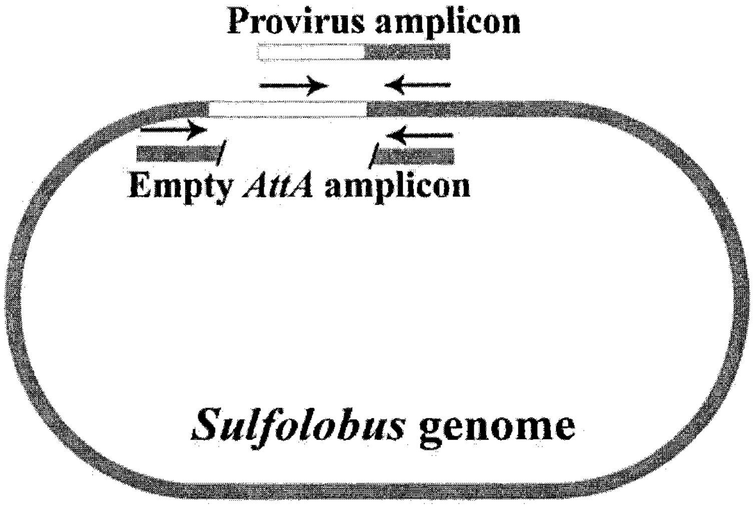

Figure 4.4: PCR assay for proviral integration. Provirus (white) integrated into host chromosome (grey) and the location of the primers used (arrows). Primers above the circle allow amplification only in the presence of a provirus (primers Int F and Provirus), primers below the circle represent primers that allow amplification only in the absence of a provirus (primers AttA and Provirus) see Table 4.2 for primer data.

To confirm these results, the *attA* integration assay was repeated with all cultures regrown from freezer stocks. The assay was also repeated with alternate primers (Provirus2 and Int F2, Table 4.2). Repeated assays show the same results as with the originally transformed DNA (data not shown).

Southern hybridization assays were used to test for integration in areas other than the wild-type *attA* site but were inconclusive due to poor resolution (data not shown).

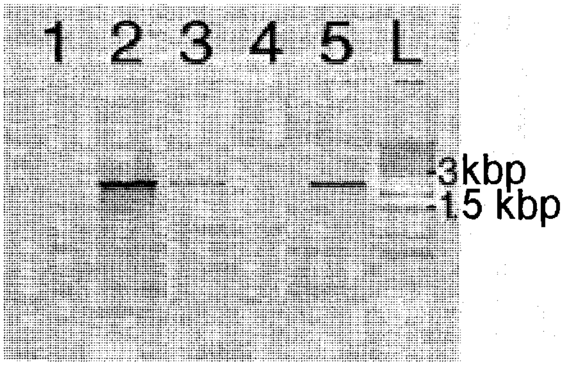

Figure 4.5: PCR assay for proviral integration of Fusellovirus shuttle vectors into the *S. solfataricus attA* site. PCR test for integration in total DNA prepared from transformed cultures using primers Provirus 1 and Int F (amplicon 2241 basepairs). Lane 1: uninfected *S. solfataricus* culture Lane 2: wild-type vector transformation. Lane 3: δInt vector transformation. Lane 4: ΔInt vector transformation Lane 5: wild-type SSV1 virus transformation. L: GeneRuler 1 kb DNA ladder (Fermentas). Representative bands are marked in kilobase pairs.

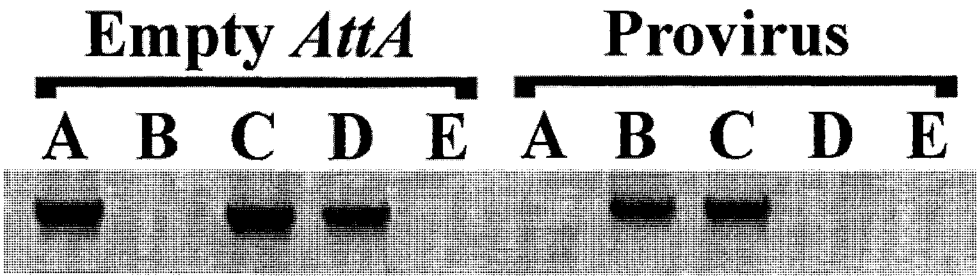

Figure 4.6: Integration is persistent. **Empty *AttA*:** PCR products using primers Provirus and AttA using *S. solfataricus* total DNA extracts as templates. A positive amplification indicates an empty *attA* sites, e.g. the absence of provirus. **Provirus:** PCR products using primers "IntF" and "Provirus" using *S. solfataricus* total DNA extracts as templates. A positive amplification indicates attachment sites containing provirus. Primer data in Table 4.2. **A**, uninfected culture; **B**, SSV1-infected culture grown from a single cell; **C**, culture transformed with wild-type vector; **D**, culture transformed with ΔInt vector; **E**, no template.

Fusellovirus Integration May be Irreversible

Fusellovirus integrase genes are partitioned upon integration since the *attP* is located within the integrase gene. How or if the virus excises its genome after integration into the host genome is unknown. To determine if the viruses studied here excise their genomes from their host after integration, transformed cultures were assayed for empty *attA* sites using primers (Int F" and Provirus Table 4.2), illustrated in Figure 4.4. Cells were assayed for *attA* sites containing virus as described above. In a culture of *S. solfataricus* infected with a wild type virus and grown from a single infected cell, cells lacking proviruses were undetectable using PCR. Cultures transformed with the same vector but not grown from a single infected cell showed a mixture of provirus containing cells and cells lacking a provirus (Figure 4.7 Lanes C).

SSV1 Constructs Containing the Integrase Gene Out-compete Those Without

The ubiquitous presence of the integrase gene in Fusellovirus genomes suggests that it confers a selective advantage; however the reduction in genome size by 7 percent due to the deletion of the integrase gene could also be advantageous. In order to compare the effect of the integrase gene on the ability of the virus to compete, cultures of the wild-type vector and ΔInt vectors were each grown from freezer stocks and were mixed at three different ratios of wild type to ΔInt infected cultures, 1:1, 1:10, and 10:1. PCR with the Int F and Int R primers were used to qualitatively detect each vector in co-culture.

Because these primers amplify the area spanning the integrase gene, different sized amplicons are produced from each virus.

Within 24 hours of co-culture, there was less of the smaller amplicon produced from ΔInt virus template than the larger amplicon produced by the wild type vector template. Within 96 hours, approximately 12 host generations, the ΔInt vector was nearly undetectable by PCR in all three of the co-culture (Figure 4.7).

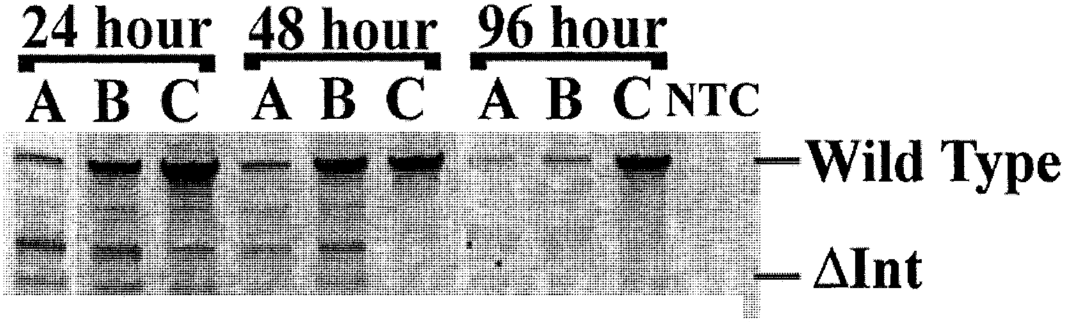

Figure 4.7: Competition between viruses with and without an integrase gene. PCR using primers Int R and Int F flanking the integrase gene. Templates were total DNA from co-cultures at indicated times. Bands corresponding to the size predicted of 952 base pairs for ΔInt vector and 1963 base pairs for the wild-type vector are labeled. Bands between the two predicted amplicon sizes are nonspecific amplification associated with the ΔInt vector. Lanes labeled A are PCR products from templates isolated from co-cultures contained starting ratios of 1 parts *S. solfataricus* transformed with wild-type vector to 10 parts transformed with ΔInt vector. Lanes labeled B are DNA from a co-culture with a ratio of 1:1; lanes labeled C are DNA a co-culture with a ratio of 10:1. The far right lane labeled NTC is a no-template PCR control.

Discussion

Integration is not Required for Replication

Fusellovirus shuttle vectors that lack the entire integrase gene are able to maintain infections in cultures by persisting as an episomal plasmid. They are capable of replication in *S. solfataricus* strain P2 , as demonstrated by: a) the production of virus particles after transformation of ΔInt vector genomes, b) the ability to amplify virus DNA from cultures days after transformation and after regrowth from cryogenic storage, c) the ability of cultures infected with ΔInt vector genomes to produce halos of growth inhibition on lawns of uninfected cells, halos that themselves contain infectious particles.

The ability of the ΔInt vector and δInt vector transformed cultures to form halos of growth inhibition remains after extended growth. Cultures infected with ΔInt vector and δInt vector were capable of producing halos on lawns of uninfected cells after of 24 days of growth (data not shown). During this time the liquid media was replaced every 6-7 days, making it highly unlikely that a non-replicating virus or any other material introduced during the transformation would be responsible for the halo production.

While all shuttle vectors created in this study replicate in *S. solfataricus*, the ΔInt vector seems to be incapable of integrating into the *attP* site of SSV1, based on the repeated lack of amplification using the two sets of primers shown in Figure 4.5. Repeated Southern hybridizations were attempted to show a more conclusive result than a lack of a PCR product but were unsuccessful. Based on the site-specificity of most tyrosine recombinase type integrases (73), and the ability of the SSV1 integrase to only recombine its specific *att* sites *in vitro* (96), it is not unexpected to find that a virus deficient in both an *attA* site and an integrase cannot integrate. Other tests for integration, such as detecting radioactively labeled Fusellovirus genomes integrated into non-labeled cells, could help to clarify this result.

SSV1 virus' lack of a requirement for integration is unlike the eukaryotic retroviruses and bacteriophage Mu that absolutely require integration for replication (26). It may also be unlike phage Lambda which, like SSV1, uses a tyrosine recombinase, but without integration cannot be stably maintained in the absence of constant induction (62). However many parameters of SSV1 induction are not known. For instance, fluctuations in halo size and intercellular DNA levels appear to occur for unknown reasons, it cannot be ruled out that normal laboratory culture of Fuselloviruses induces these them to some extent.

Integration of the δInt Vector

One of the most surprising findings of this work is that the δint vector, which contains the 3' end of the integrase gene but none of the sequence encoding the putative catalytic residues, shows signs of integration (Figure 4.5). However, in the integration assay the consistently fainter PCR bands produced by the δInt vector transformed culture suggest that this integration may be less efficient than viruses containing the full-length integrase (Figure 4.5). Experiments to measure the relative amounts of integrated versus episomal genomes were not directly performed however.

How this integration takes place is truly puzzling. The *S. solfataricus* P2 genome contains fragments of integrase genes similar to Fusellovirus integrases partitioned at putative *attP* sites that are most likely the result of past infections (100). However, there are no signs of active virus infection in this strain, and nothing to suggest that these disrupted integrase gene fragments are expressed or functional. Furthermore, these gene fragments appear to be integrated into sites other than the SSV1 *attA* site. Assuming that the integrases that made these insertions are site-specific like the homologues used to identify them, they should lack the specificity to recognize the SSV1 *attP* site and integrate the δInt vector. Moreover, none of the integrase genes in *S. solfataricus* strain P2 genome are complete (100).

Sulfolobus does have a homologous recombination system which has been demonstrated to recombine foreign DNA into its chromosome, however efficient recombination requires nearly 1000 bases of homologous sequence (121). The attachment sites in SSV1 and its host shares only 44 base pairs that contains a single mismatch, and some sequence similarity in the flanking regions. It is possible however, that the constantly present extra-chromosomal form of the δInt vector is able to recombine inefficiently in spite of this small region of homology. This seems unlikely, but the minimum sequence length needed for recombination in *S. solfataricus* is not precisely known. It could be measured by transforming ΔInt based shuttle vectors containing different sized regions of homology to a portion of a marker gene, such as the beta-galactosidase encoding gene (27).

Another possibility is that the integrase containing plasmid pSSVi, which was recently identified from a substrain of *S. solfataricus* strain P2 (111), may have assisted in the integration of the δInt construct. This plasmid was recently observed in the same strain of *S. solfataricus* strain P2 used in this study after the transformation of SSV-I2 genomic DNA into this strain by another group. The 5.5 kilobase pair pSSVi plasmid contains an integrase gene that was shown to integrate only into tRNA 31 coding for the arginine with the GCG anticodon and not the SSV1 *attA* site in the tRNA 30 coding for arginine with the anticodon CCG. The sequence of these tRNAs is very similar however, having only two mismatches relative to the 44 base pair SSV1 *attP* site. As the SSV1 integrase is capable of tolerating a single mismatch between its *attP* site and the laboratory host *S. solfataricus attA* site, it is possible that the pSSVi integrase could, integrate viruses containing SSV1 *attP* into the SSV1 *attA* site (117). The integrase mutants constructed in this study would provide a good model to test this hypothesis, as the *attP* sequence could be inserted into the ΔInt vector and either be co-transformed with pSSVi into uninfected cells or transformed into cells containing pSSVi.

It is not known when or how the strain of *S. solfataricus* strain P2, used in all experiments described in this dissertation, acquired pSSVi. This strain was brought from Wolfram Zillig's collection to the Stedman lab (K. Stedman, personal communication). No signs of the pSSVi infection described in (111) were observed in the strain in our lab, such as the 5.5 kilobase pair band of the plasmid DNA, the presence of smaller satellite virus particles seen in the co-infection of pSSVi and SSV1 or SSV-I2, or the decrease in growth rate of cells and increase in total virus titer when co-infected (111). As the manuscript describing pSSVi was not published until after the experiments described were completed, the presence of pSSVi in the cultures used in these experiments was not rigorously tested for and therefore cannot be excluded as a contributing factor to the results obtained.

Integration is Permanent

Figure 4.7 shows that cultures grown from a single wild-type virus-infected colony have no detectable empty *attA* sites while transformed cultures consistently contain empty *attA* sites. These data suggest that integration of the virus is permanent. Alternatively, the same results would be obtained if virus excision takes places along with rapid reintegration. The presence of empty *attA* sites in the cultures not passaged from a single colony suggests that virus spread may be slow, or that a subpopulation of cells remains immune to viral infection or viral integration. The latter seems most likely as incompletely infected cultures have been previously observed in other SSV-based infections even after continued growth of these cultures (48).

Integration and its Relationship to Fitness

While replication of the virus is possible without integration into the SSV1 *attA* site, the effect of a missing integrase gene on fitness is less clear. Competition experiments between cultures transformed with the ΔInt vector and the wild-type vector suggest that viruses containing the integrase gene cause the production of more viral DNA than those lacking the integrase gene in co-cultures (Figure 4.7).

Attempts at single colony isolation of strains infected with the ΔInt vector support the hypothesis that the that the ΔInt vector is less fit than the wild type virus in it's ability to spread in infected cultures. When single colonies from the wild-type vector cultures were grown on lawns of uninfected *S. solfataricus,* halos were produced by 6 out of 112 colonies. No halos were observed from spots of the ΔInt vector isolates despite the screening of over 200 colonies (data not shown).

Very little is known about the mechanisms and rate limiting steps of SSV1 virus replication and spread, making the reason or reasons for the relative decrease of the ΔInt vector DNA in co-culture unclear. Possibilities include a more efficient re-infection of transiently cured cells by the wild-type vector, a faster relative growth rate of cells infected by the wild-type vector compared to those infected by ΔInt, or displacement of the ΔInt vector by the wild-type vector. None of these have been tested.

There are several caveats to this competition experiment. As the PCR used in this assay amplifies across the integrase gene and *attP* site using template extracted from cellular sources, only inter-cellular circular viral DNA is detected, therefore the information that can be extrapolated from this experiment is restricted to the relative amount of the extrachromosomal ΔInt vector DNA present in the cell in relation to the relative amount of extrachromosomal wild type vector DNA present in the cell. It is possible that the metric of intracellular extrachromosomal viral DNA is not a true indicator of the amount of virus genome, which is the sum of extrachromosomal DNA, integrated provirus, and extracellular virion particles. Other caveats include the inability of the assay to detect the growth rate of wild-type or ΔInt infected cells, which may or may not be replicating at different rates and influencing virus production, and the possibility of a pSSVi contamination confounding the results.

To ultimately understand the advantage conferred by the integrase gene and integration, a quantification of intracellular virus DNA production and particle production, is needed. Using Southern hybridizations to track total intracellular viral DNA levels combined with an assay for virus particle production, such as a halo forming assay (95) or a virus particle count with fluorescent staining (113), would help to clarify whether these competition data are a true measure of virus production.

In considering possible reasons why integrase-lacking viruses may be less productive than integrase-containing viruses there are some possibilities that are unlikely. Transcriptional expression of provirus genes from host promoters is unlikely to occur as all of the proviral genes are transcribed divergently from the central region of the provirus (87). The closest host promoter, that of the tRNA gene, a transcript is approximately 6.5 kilobase pairs away from an ORF oriented in the proper direction. No research has been published on maximum lengths of transcripts in *Sulfolobus* however, so this cannot be completely ruled out. It is also not likely that the removal of the integrase gene disrupts the function of other SSV1 genes, as the integrase gene is located at the end of the T5 polycistronic transcript (87). Additionally, the insertion of an *E. coli* plasmid in the integrase containing transcript does not seem to prevent virus integration (shown in pAJC97 and (105)).

Conclusion

The study of SSV1 integration in its hyperthermophilic archaeal host *S. solfataricus* provides a unique contrast to many of the well-studied integrating viruses such as phages Lambda and P1. This study shows that removal of the integrase gene from the SSV1 virus does not stop the virus from replicating and infecting new cells. The main change observed in replication of the ΔInt virus is the inability to integrate into the SSV1 *attP* site. Therefore viral integration appears to be an optional step in the replication of SSV1 and probably all Fuselloviruses. It is doubtful that an integrase deficient Fusellovirus would be competitive with other Fuselloviruses containing integrases in the environment based on the gene's complete conservation in all sequenced genomes and based on the results of the competition assay.

It also appears that a deletion of the integrase catalytic domain, a deletion that removes several residues known to be required for recombination in vitro (96), does not stop proviral integration at the SSV1 *attA* site, although the reason for this phenomenon is not known. To our knowledge, this is the first directed functional mutagenesis study of any archaeal virus. It will be interesting to see if Fuselloviruses with further modified integrase genes also compete well with the wild-type or if Fuselloviruses that lack an integrase gene can be found in the environment.

Materials and methods

Isolation of DNA

Isolation of total DNA from Sulfolobus used methods described previously (105). Briefly, 15 ml of late logarithmically growing cells (OD600 nm = 0.7) were centrifuged for 5 min at 2000 g. The cell pellet was resuspended in 500 μl of TEN (10 mM Tris/HCl, 10 mM EDTA, 150 mM NaCl, pH 8.0). A total of 500 μl of TENST (TEN plus 0.12% Triton X-100 and 1.6% N-lauryl sarcosine) was added, and the mixture was incubated for 30 min on ice. A total of 1 ml of a mixture of phenol/chloroform/isoamyl alcohol (25:24:1) was added and mixed by vortexing, and the phases were separated by centrifugation for 20 min in a microcentrifuge at maximum rpm. The aqueous phase was phenol/chloroform-extracted two more times, and RNAse was added. The DNA was precipitated by the addition of 0.8 volumes of isopropanol. The precipitate was washed twice with 70% ethanol, dried, and dissolved in 30 μl sterile water.

Isolation of viral (extrachromosomal) DNA from *S. solfataricus* used the alkaline lysis technique essentially as described in (17). 100 ml of cells were resuspended after pelleting in 100 μl of a solution of 50 mM Glucose, 25 mM Tris/HCl pH 8.0, and 10 mM EDTA. 200 μl of a solution of 0.2M NaOH and 1% (w/v) SDS was added. Finally 150 μl of a 3M potassium acetate solution (60 ml 5M potassium acetate, 11.5 ml Acetic Acid, and 28.5 ml of sterile water) was added and mixed well. The mixture was centrifuged at maximum speed in a tabletop centrifuge for 20 minutes after which the supernatant was extracted three times with 1 volume (450 μl) of phenol/chloroform/isoamyl alcohol (25:24:1). The DNA was precipitated by the addition of 0.8 (360 μl) volumes of isopropanol. The precipitate was washed twice with 1 ml of 70% ethanol, dried, and dissolved in 50 μl sterile water.

Isolation of viral (extrachromosomal) DNA from *E. coli* used the alkaline lysis essentially as described in (17). Briefly, 3 ml of cells were resuspended after centrifugation at 14,000 x g for 2 minutes in 100 μl of a solution of 50 mM Glucose, 25 mM Tris/HCl pH 8.0, and 10 mM EDTA. 200 μl of 0.2 M NaOH and 1% (w/v) SDS was added and mixed by inversion. Finally 150 μl 3 M potassium acetate solution was added and mixed well. The mixture was centrifuged at maximum speed in a tabletop microcentrifuge for 20 minutes after which the supernatant was extracted three times with 1 volume (450 μl) of phenol/chloroform/isoamyl alcohol (25:24:1). The DNA was precipitated by the addition of 0.8 volumes (360 μl) of isopropanol. The precipitate was washed twice with 1 ml of 70% ethanol, dried, and dissolved in 50 μl sterile water.

Primer Design for LIPCR Construction

All oligonucleotides used are listed in Table 4.2. Standard M13 forward (-20) and M13 reverse (-27) primers were used unchanged for the construction of the wild type vector. Primers for the ΔInt vector (ΔInt Forward and ΔInt Reverse) were designed to remove the complete integrase gene beginning at the start codon, base 1968, and continuing to the end of the stop codon, base 961 of the SSV1 genome (GenBank accession number X07234). Primers for the δInt vector (δInt Forward and δInt Reverse) were designed to remove the C-terminal portion of the integrase gene integrase gene from base 1462 to base 1051 of the SSV1 genome. Primers were designed so that their 5' end flanked the area to be removed. The length of the primer was extended in the 3' direction until a GC clamp of at least one base was present and predicted annealing temperatures were between 55-60°C.

Primer sequences were checked for hairpins and other secondary structure using Mfold http://www.bioinfo.rpi.edu/applications/mfold/old/dna/ with Na^+ concentrations set at 50mM and Mg^{2+} concentrations set at 0 mM.

Full-length Vector Construction Using LIPCR

All plasmids and strains used are listed in Table 4.1. The pBlusecript vector (105) consisting of pBluescript II SK+ in the SSV1 virus was used as template for the LIPCR to create the wild type vector and the δInt vector using primers described above. Amplification used Phusion high fidelity DNA polymerase in an MJ Research Dyad thermocycler. Temperature calculations were estimated by the thermocycler and all reactions were carried out in 20 μL volumes. Materials used in each PCR are described in Table 4.3. LIPCR conditions were as follows: an initial denaturation at 98°C for 3 minutes, and subsequent denaturations at 98°C for 15 seconds. Annealing temperatures were 62°C for the ΔInt vector and 65°C for the wild type vector and the δInt vector. All annealing times were 15 seconds long. Extension was for 8 minutes at 72 °C. Reactions were cycled 30 times with a final 8-minute extension at 72°C.

Table 4.2: Primers

Primer	Sequence	Description
Int F	5'-ATGGTAAGGAACATGAAGATGAAGAAGAG-3'	Amplifies area surrounding integrase gene
Int R	5'-TAGAATACAAGGTGGACAAATGAGTCCTTC-3'	
δInt F	5'-AGATATTAGATCTTTTATTCAAGGGCGTAAACCG-3'	Amplifies δInt vector
δInt R	5'-CGCTATCAGCTCTGCAAAGAGTCGGTAAGCCT-3'	
M13 R	5'-CAG GAA ACA GCT ATG AC-3'	Amplifies wild type vector
M13 F	5'-GTA AAA CGA CGG CCA GT-3'	
ΔInt F	5'-CGTCTTATCTTTCGTCATTTCACCTGGTACTATTATGG-3'	Amplifies ΔInt virus
ΔInt R	5'-GGGGTCTGACAGGCGCCGTATCACTATC-3'	
Provirus	5'-AACGTTACCGGAGATGTTGC-3'	Amplifies in presence of Provirus
Int F	5'-ATGGTAAGGAACATGAAGATGAAGAAGAG-3'	
Provirus 2	5'-TTGCACAGACTGCTGGAATC-3'	
Int F2	5'-GTTTACGCCCTTGAATAAAAGATCTAATATCTA-3'	
AttA	5'-GACATAATTATACGTGAAAGAAAAGGGCG-3'	Amplifies Empty *attA* site with "Provirus"
Provirus	5'-AACGTTACCGGAGATGTTGC-3'	

Table 4.3: LIPCR Reagents

Reagent	Wild type amplification	ΔInt amplification	δInt amplification
Buffer	1X HF Buffer	1X HF Buffer	1X HF Buffer
dNTPs	0.2mM/base	0.2mM/base	.2mM/base
Template	3pM pBluescript vector	6pM wild type vector	4pM pBluescript vector
Forward primer	M13 F, 250nM	ΔInt F, 250nM	δInt F, 250nM
Reverse primer	M13 R, 250nM	ΔInt R, 250nM	δInt R, 250nM
Phusion Polymerase	0.02 U/μl,	0.02 U/μl	0.01U/μl

Viral Transformation and Detection in S For the construction of the wild-type vector, the LIPCR amplified DNA was ligated into the TOPO Blunt II vector using the Zero Blunt TOPO PCR cloning kit (Invitrogen) following the manufacturer's protocols. For the construction of the δInt vector and the ΔInt vector, the LIPCR amplified DNA was phosphorylated with T4 kinase (Fermentas), and circularized with T4 ligase (Fermentas) following manufacturer's protocols. All products were transformed into StAble 3 chemically competent cells (Invitrogen) following the manufacturer's protocols. Extrachromosomal DNA isolated from transformed clones was screened by restriction endonuclease digestion with HindIII or EcoRI (Fermentas). Plasmids containing the appropriate restriction fragments were sequenced across the ligation junctions to ensure that correct ligation took place

Viral Transformation and Detection in Sulfolobus

The wild type, ΔInt, and δInt vector genomes isolated from E. coli were electroporated into *S. solfataricus* strain P2 as previously described (95) and assays for viral activity using the spot on lawn technique were done as previously described (105). Briefly, lawns of *S. solfataricus* strain P2 were spread on Gelrite plates in a 0.2% Gelrite soft-layer with a standard Sulfolobus media similar to that used previously (20). 2 μl of transformed P2 culture were spotted on the lawns which were then incubated for 48 to 72 h until a consistent lawn was present and halos were observed in the positive controls. All plates were made in triplicate every 48 h post-transformation. Preparations of viral DNA from infected strains were also analyzed by restriction endonuclease digestion, PCR amplification using viral specific primers Int R and Int F (see Table 4.2), and transmission electron microscopy of negatively stained virus particles (Figure 4.3).

Conditions for the amplification of viral DNA with primers Int F and Int R for diagnostic purposes were as follows: after a 5-min denaturation at 94 °C, reaction mixtures were subjected to 35 amplification cycles of denaturation (94 °C for 15 seconds), annealing (51 °C for 15 seconds), extension (72 °C for 1 min and 45 seconds), and a final extension (72 °C for 4 minutes). Conditions for the following PCRs were the same with the exception of the following changes: First and second tests for virus integration PCR (primers Int F and Provirus) and (Int F2 and Provirus2) used an annealing temperature of 50°C and an extension time of 2 minutes. PCR provirus test, (primers attA and Provirus) used an annealing temperature of 48°C

TEM Imaging

TEM preparations used the supernatant of infected cultures that were centrifuged for 15 min at 2000g. 2 μl of the supernatant was absorbed onto a carbon/formvar grid and negatively stained by floating the grids on a solution of 2% uranyl acetate for 15 seconds followed by wicking off the excess uranyl acetate with a paper towel and drying for 10 minutes at room temperature. The prepared grids were viewed on a JEOL 2000 TEM.

Cell culture

Sulfolobus strains were grown in liquid culture at pH 3.2 with moderate shaking at 80° in long-necked Erlenmeyer flasks. The liquid medium used was similar to that of (20) and contained 0.1% yeast extract (Sigma) and 0.2% sucrose (Fermentas) as carbon source. Solid media were made by adding of Gelrite (Sigma) to the medium at a final concentration of 0.6% w/v. Soft layers for overlays were made by the addition of Gelrite to 0.2%. For long term growth infected strains of *S. solfataricus* were grown for 24 days in the media described above at 80 °C with shaking. Every week the cultures were diluted 1:200 with fresh media. Spot plates and PCR using Int R and Int F as described above were used to determine the presence or absence of virus at the end of the test.

Competition Assays

Cultures infected with the wild type vector and the ΔInt vector were grown from freezer stocks in the media described above to an OD_{600} nm of 0.7, then mixed and added to fresh media and grown as described above. Ratios of 1:1, 1:10, and 10:1 were created by adding 1 ml and 10 ml or 5 ml and 5 ml of the respective cultures to 40 ml of fresh media and incubated at 80 °C with shaking. 5 ml samples were removed daily for DNA extraction and the volume was replaced with fresh media to keep the cultures actively growing. Spot on lawn tests were done to determine viral activity and PCR with Int F and Int R primers (Table 4.2) was used to confirm which virus strain was present.

Sequencing

Sequencing reactions used Big Dye Terminator readymix solutions version 3.1(Applied Biosystems) according to the manufacturer's instructions. Each reaction used 4 μl of BDT, 1.6 pmol of primer, and 400 ng of template. Sequencing was performed by the KECK genomics center at Portland State University and at the Oregon Health and Science University's sequencing core. The template for sequencing reactions was pAJC96 DNA purified from *E. coli.*

Restriction fragment size was calculated with pDraw32 (AcaClone software).

Chapter 5: Summary and Discussion

In conclusion, the data presented here give a more precise picture about the replication of the Fuselloviruses in their archaeon hosts. Sequencing and annotation of two new Fusellovirus genomes and the comparison of these viruses to the four previously studied viruses and a newly available virus genome allowed several new insights into the similarities and differences of this virus family.

While the viruses lack a common sequence or pattern in the area thought to be the origin of replication, GC skew, and purine skew indirectly support DNA replication originating from the area that is near the T5 and T6 transcript start sites. Future research will be needed to demonstrate this biochemically. Promoter regions are conserved in the 7 Fuselloviruses based on the similarity of the promoter regions of nine of the eleven SSV1 promoters. This suggests a conserved mechanism for temporal expression of the basic virus genes, especially highly conserved late genes. The two non-conserved promoters regulate the T5 transcript of largely non-conserved ORFs and the T-ind transcript found only in SSV1, the only Fusellovirus to date shown to be dramatically up-regulated by UV irradiation (117). This suggests that SSV1 is somewhat of an outlier in the Fuselloviruses with respect to its sensitivity to ultraviolet irradiation and possibly its regulation.

With respect to ORF conservation, the Fusellovirus genome seems to be partitioned in two regions, one conserved and the other not. Within this non-conserved region are the T5, T6, and T3 transcripts, both of which contain ORFs not present in all 7 viruses. Non-conserved ORFs are particularly common near the beginning of the T5 and T6 transcripts.

The transcript with the least conservation, T5 in SSV1, shows little similarity in promoter sequence, ORF pattern, and coding strand orientation to the similar putative transcripts in other Fuselloviruses. Clusters of repeat sequences, and in the case of SSV-K1 what appears to be a recent gene insertion, suggest that this area of the genome is a hotspot for recombination. The idea of a recombination hotspot is also supported by several ORFs with high homology to ORFs in other virus families, suggesting a recent movement of genes between virus families. Why it is expressed early in the transcription cycle of SSV1 remains a mystery, as does the reason for this non-conserved portion of the genome to exist at all.

Based on the geographic change seen in the pair-wise comparison of each of the seven viruses conserved ORFs there is clearly a change with respect to genetic distance in Fuselloviruses, indicating that their spread is limited. This may be due the barrier of inhospitable environments that separates individual hot springs, and may also be compounded by population bottlenecks that occur as these ephemeral hot springs change. This suggests that the spread of the viruses occurs through a slow and gradual process of island hopping. This is opposed to a large scale and rapid spread from something like a large volcanic eruption spreading particles throughout the world in a single event. This type of spread would be instantaneous showing no correlation between geographic and genetic distance.

Interestingly, SSV-L1 seems to be somewhat of an outlier with respect to biogeography. It has conserved ORFs that when viewed individually seem to be more similar to distantly related viruses. The concatenation of all SSV-L1s ORFs however still shows a positive correlation between genetic and geographic distance suggesting that these individual changes average out to show an overall correlation. This correlation is the weakest of all of the viruses however, and removal of the SSV-L1 data from the pair-wise comparison improves the r-squared value of the other data (data not shown). This suggests that the SSV-L1 virus may have experienced a slightly different history than the other viruses.

To aid in the study of unknown ORFs in the Fusellovirus family a new method for inducing deletions into the circular genomes of the SSVs (LIPCR) was developed, which can be used not just with the integrase genes but also any other part of the viral genome. This method allowed the creation and testing of two new SSV1 virus mutants. Testing of these mutants showed that a virus completely lacking the integrase can replicate but not integrate at least not into the wild-type *attA* site, indicating integration in these viruses is not essential for infection. Direct competition under laboratory conditions shows that viruses without the integrase gene appear to be at a disadvantage compared to the wild type. A partial integrase deletion mutant appears to integrate apparently without a functional integrase.

The Benefits of Lysogeny

One of the most intriguing questions involving virus integration, and in a more general sense lysogeny, is what benefit lysogeny confers to the virus compared to a purely virulent replication strategy. The widespread distribution of integrating viruses, the conservation of the mechanism of using tyrosine recombinases, and the parallel evolution of other integration mechanisms such as those used by serine recombinases and non site-specific integrases to accomplish integration suggest its benefit. One hypothesis as to why, put forth by Echols, is that integration decreases the need for maintenance by the virus as the host continually replicates the integrated genome as it multiplies (31). This may be why the SSV viruses do not seem to excise their genomes once integrated.

A second hypothesis supporting lysogeny put forth by Levin et al. is that under low cell densities virulent viruses would have a disadvantage in leaving the host in the when the chances of finding new hosts are poor (60). This may be particularly relevant to the high temperature acidic environments where *Sulfolobus* is found. While typical cell densities range from 10^6 to 10^8 cells/ml in these springs (102), the amount of free virus particles observed using culture-independent techniques are much lower than all other examined environments (74). Fuselloviruses themselves are not tolerant of the high temperature conditions in which there hosts thrive and lose infectivity within minutes when stored at high temperature acidic conditions (S. Morris, R. Diessner, S. Lee and K. Stedman, manuscript in preparation). Together these observations suggest that while cell densities in these springs may be normal the ability of viruses to move from one cell to the next may be poor, and that vertical transmission of the virus in these cases may be particularly beneficial.

A third theory is that lysogens may aid in the fitness of the host, by the result of added genes and/or the impartment of immunity to superinfection. Horizontal gene transfer can result in the addition of genes beneficial to the host, such as the presence of pathogenicity islands, the prevalence of virally encoded virulence factors (1), the addition of metabolic genes (55), and the evolution of DNA replication enzymes (34). Within *Sulfolobus* evidence of host and viral gene transfer has been observed suggesting a long and complex history between the viruses and their hosts (39, 81, 85). Remnants of horizontal gene transfer mediated by integrases can be seen in the defective proviruses integrated in the genomes of Sulfolobus and other extreme thermophiles (81, 85, 98), the evidence of N- and C-terminal integrase fragments separated by a pRN-like plasmid element in the genome of *S. solfataricus* (79), and a large variety of plasmids and other insertion elements also present in *Sulfolobus* species (99). Whether any of these are beneficial to Sulfolobus is unclear.

Integrase Genes in Archaea

All of the viruses sequenced contain an integrase gene that is a member of the tyrosine recombinase family of proteins. Interestingly this gene is found at the end of the least conserved transcript, SSV1 T5, in the Fuselloviruses, and is not dramatically affected by the insertion of several kilobase pairs of foreign DNA upstream in its transcript, as is seen in the creation of the SSV1 based shuttle vectors that are completely capable of integration (105). Some promoter-like sequences are found directly upstream of the integrase gene in the Icelandic SSVs and SSV-K1, suggesting, along with microarray data in SSV1, that the integrase gene may have its own promoter as well as the polycistronic T5 promoter known to transcribe it.

Several attempts, most notably (73) have been made to create alignments of the tyrosine recombinase family of proteins as a way of determining the relatedness of this group. The proteins differ greatly in function and relatedness so alignments are generally made of just the conserved boxes found near the catalytic residues (73). Phylogenetic analysis of an amino acid alignment of all annotated putative and actual integrase catalytic domains in the Archaea and their viruses generates a tree similar to that of the 16S rDNA tree (Figure 5.1), however the internal nodes of this tree resolve poorly, with bootstrap values below 50%, most likely due to sequence saturation. The branches, circled in the figure resolve well and show clear relatedness between integrases in similar genera and in the viruses within them.

Fusellovirus integrases form a monophyletic clade containing some putative integrases annotated in *Sulfolobus* species. It seems likely that the integrases within the genomes of *Sulfolobus* are proviral integrases based on the commonality of integrated viruses found in the two sequenced *Sulfolobus* species (52, 100). The Fusellovirus integrase clade does not include integrases found in plasmids of Sulfolobus with the exception of pSSVi. These proteins do not partition the integrase gene upon insertion and are not thought to be directly related to the Fusellovirus integrases (97).

The recent advancements in metagenomic sequencing will undoubtedly increase the number of integrase genes in the database. With this additional information it may be possible to fill in some of the gaps in the phylogenetic tree that currently cannot be resolved. This could lead us to a better understanding of how this common viral gene made its way into (or out of) the family Fuselloviridae.

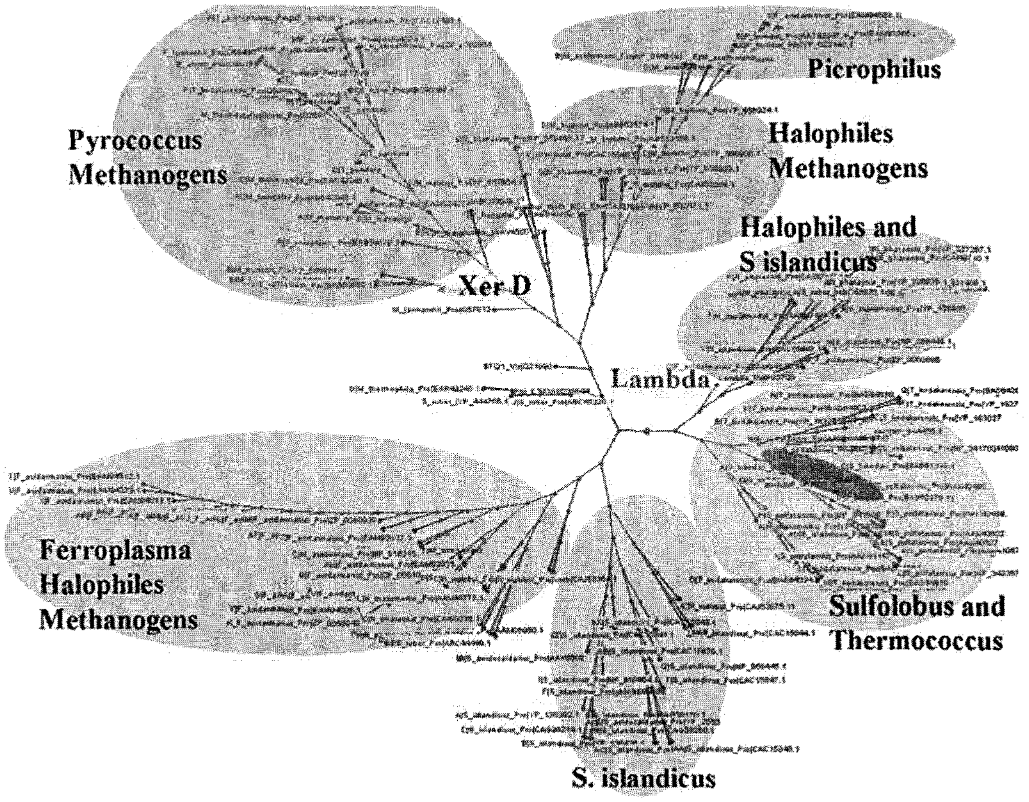

Figure 5.1: An unrooted tree of integrases in the Archaea. Tree constructed using the neighbor-joining algorithm in Clustal X (108) and viewed with Hypertree (16). Red Dots on nodes indicate bootstrap values above 70%, blue dots below 70%.

References

1. **Abedon, S. T., and L. , J.T.** 2005. Why bacteriophage encode exotoxins and other virulence factors. Evolutionary Bioinformatics Online **1:**97-110.

2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-10.

3. **Andersson, S. G., C. Alsmark, B. Canback, W. Davids, C. Frank, O. Karlberg, L. Klasson, B. Antoine-Legault, A. Mira, and I. Tamas.** 2002. Comparative genomics of microbial pathogens and symbionts. Bioinformatics **18 Suppl 2:**S17.

4. **Andersson, S. G., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. Alsmark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, H. H. Winkler, and C. G. Kurland.** 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature **396:**133-40.

5. **Andrade, M. A., C. Ouzounis, C. Sander, J. Tamames, and A. Valencia.** 1999. Functional classes in the three domains of life. J Mol Evol **49:**551-7.

6. **Arnold, H. P.** 1998. Isolation and characterization of novel viruses in the crenarchaeon genus *Sulfolobus*. Ludwig-Maximilians-Universität, München.

7. **Arnold, H. P., Q. She, H. Phan, K. Stedman, D. Prangishvili, I. Holz, J. K. Kristjansson, R. Garrett, and W. Zillig.** 1999. The genetic element pSSVx of the extremely thermophilic crenarchaeon Sulfolobus is a hybrid between a plasmid and a virus. Mol Microbiol **34:**217-26.

8. **Arnold, H. P., U. Ziese, and W. Zillig.** 2000. SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon Sulfolobus. Virology **272:**409-16.

9. **Arnold, H. P., W. Zillig, U. Ziese, I. Holz, M. Crosby, T. Utterback, J. F. Weidmann, J. K. Kristjanson, H. P. Klenk, K. E. Nelson, and C. M. Fraser.** 2000. A novel lipothrixvirus, SIFV, of the extremely thermophilic crenarchaeon Sulfolobus. Virology **267:**252-66.

10. **Auchtung, T. A., C. D. Takacs-Vesbach, and C. M. Cavanaugh.** 2006. 16S rRNA phylogenetic investigation of the candidate division "Korarchaeota". Appl Environ Microbiol **72:**5077-82.

11. **Bath, C., T. Cukalac, K. Porter, and M. L. Dyall-Smith.** 2006. His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. Virology **350:**228-39.

12. **Bell, S. D.** 2005. Archaeal transcriptional regulation--variation on a bacterial theme? Trends Microbiol **13:**262-5.

13. **Bell, S. D., and S. P. Jackson.** 2001. Mechanism and regulation of transcription in archaea. Curr Opin Microbiol **4:**208-13.

14. **Benson, G.** 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27:**573-80.

15. **Bettstetter, M., X. Peng, R. A. Garrett, and D. Prangishvili.** 2003. AFV1, a novel virus infecting hyperthermophilic archaea of the genus acidianus. Virology **315:**68-79.

16. **Bingham, J., and S. Sudarsanam.** 2000. Visualizing large hierarchical clusters in hyperbolic space. Bioinformatics **16:**660-1.

17. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res **7:**1513-23.

18. **Biswas, T., H. Aihara, M. Radman-Livaja, D. Filman, A. Landy, and T. Ellenberger.** 2005. A structural basis for allosteric control of DNA recombination by lambda integrase. Nature **435:**1059-66.

19. **Boone, D. R., R. W. Castenholz, and G. M. Garrity.** 2001. Bergey's manual of systematic bacteriology, 2nd ed. Springer, New York.

20. **Brock, T. D., K. M. Brock, R. T. Belly, and R. L. Weiss.** 1972. Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. Archiv fur Mikrobiologie **84:**54-68.

21. **Brock, T. D., and T. D. Brock.** 1994. Biology of microorganisms, 7th ed. Prentice Hall, Englewood Cliffs, N.J.

22. **Brussow, H., C. Canchaya, and W. D. Hardt.** 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev **68:**560-602.

23. **Cannio, R., P. Contursi, M. Rossi, and S. Bartolucci.** 1998. An autonomously replicating transforming vector for Sulfolobus solfataricus. J Bacteriol **180:**3237-40.

24. **Chen, Y., U. Narendra, L. E. Iype, M. M. Cox, and P. A. Rice.** 2000. Crystal structure of a Flp recombinase-Holliday junction complex: assembly of an active oligomer by helix swapping. Mol Cell **6:**885-97.

25. **Contursi, P., R. Cannio, S. Prato, Q. She, M. Rossi, and S. Bartolucci.** 2007. Transcriptional analysis of the genetic element pSSVx: differential and temporal regulation of gene expression reveals correlation between transcription and replication. J Bacteriol **189:**6339-50.

26. **Craig, N. L.** 2002. Mobile DNA II. ASM Press, Washington, D.C.

27. **Cubellis, M. V., C. Rozzo, P. Montecucchi, and M. Rossi.** 1990. Isolation and sequencing of a new beta-galactosidase-encoding archaebacterial gene. Gene **94:**89-94.

28. **Di Giulio, M.** 2007. The tree of life might be rooted in the branch leading to Nanoarchaeota. Gene **401:**108-13.

29. **Douglas, J. T.** 2007. Adenoviral vectors for gene therapy. Mol Biotechnol **36:**71-80.

30. **Dyall-Smith, M., S. L. Tang, and C. Bath.** 2003. Haloarchaeal viruses: how diverse are they? Res Microbiol **154:**309-13.

31. **Echols, H.** 1972. Developmental pathways for the temperate phage: lysis vs lysogeny. Annu Rev Genet **6:**157-90.

32. **Esposito, D., and J. J. Scocca.** 1997. The integrase family of tyrosine recombinases: evolution of a conserved active site domain. Nucleic Acids Res **25:**3605-14.

33. **Felsenstein, J.** 2005. PHYLIP (Phylogeny Inference Package) 3.6 ed. Distributed by the author.

34. **Filee, J., P. Forterre, and J. Laurent.** 2003. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. Res Microbiol **154:**237-43.

35. **Flint, S. J.** 2000. Principles of virology : molecular biology, pathogenesis, and control. ASM Press, Washington, D.C.

36. **Freeman, J. M., T. N. Plasterer, T. F. Smith, and S. C. Mohr.** 1998. Patterns of Genome Organization in Bacteria. Science **279:**1827a-.

37. **Frols, S., P. M. Gordon, M. A. Panlilio, C. Schleper, and C. W. Sensen.** 2007. Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays. Virology **365:**48-59.

38. **Galtier, N., M. Gouy, and C. Gautier.** 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci **12:**543-8.

39. **Garcia-Vallve, S., A. Romeu, and J. Palau.** 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res **10:**1719-25.

40. **Geslin, C., M. Le Romancer, G. Erauso, M. Gaillard, G. Perrot, and D. Prieur.** 2003. PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, "Pyrococcus abyssi". J Bacteriol **185:**3888-94.

41. **Greve, B., S. Jensen, K. Brugger, W. Zillig, and R. A. Garrett.** 2004. Genomic comparison of archaeal conjugative plasmids from Sulfolobus. Archaea **1:**231-9.

42. **Haring, M., X. Peng, K. Brugger, R. Rachel, K. O. Stetter, R. A. Garrett, and D. Prangishvili.** 2004. Morphology and genome

organization of the virus PSV of the hyperthermophilic archaeal genera Pyrobaculum and Thermoproteus: a novel virus family, the Globuloviridae. Virology **323:**233-42.

43. **Haring, M., R. Rachel, X. Peng, R. A. Garrett, and D. Prangishvili.** 2005. Viral diversity in hot springs of Pozzuoli, Italy, and characterization of a unique archaeal virus, Acidianus bottle-shaped virus, from a new family, the Ampullaviridae. J Virol **79:**9904-11.

44. **Haring, M., G. Vestergaard, K. Brugger, R. Rachel, R. A. Garrett, and D. Prangishvili.** 2005. Structure and genome organization of AFV2, a novel archaeal lipothrixvirus with unusual terminal and core structures. J Bacteriol **187:**3855-8.

45. **Hausner, W., J. Wettach, C. Hethke, and M. Thomm.** 1996. Two transcription factors related with the eucaryal transcription factors TATA-binding protein and transcription factor IIB direct promoter recognition by an archaeal RNA polymerase. J Biol Chem **271:**30144-8.

46. **Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp.** 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics **8:**460.

47. **Ingalls, A. E., S. R. Shah, R. L. Hansman, L. I. Aluwihare, G. M. Santos, E. R. Druffel, and A. Pearson.** 2006. Quantifying archaeal

community autotrophy in the mesopelagic ocean using natural radiocarbon. Proc Natl Acad Sci U S A **103:**6442-7.

48. **Jonuscheit, M., E. Martusewitsch, K. M. Stedman, and C. Schleper.** 2003. A reporter gene system for the hyperthermophilic archaeon Sulfolobus solfataricus based on a selectable and integrative shuttle vector. Mol Microbiol **48:**1241-52.

49. **Julien, B.** 2003. Characterization of the integrase gene and attachment site for the Myxococcus xanthus bacteriophage Mx9. J Bacteriol **185:**6325-30.

50. **Karam, J. D., and J. W. Drake.** 1994. Molecular biology of bacteriophage T4. American Society for Microbiology, Washington, DC.

51. **Karner, M. B., E. F. DeLong, and D. M. Karl.** 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature **409:**507-10.

52. **Kawarabayasi, Y., Y. Hino, H. Horikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, H. Kosugi, A. Hosoyama, S. Fukui, Y. Nagai, K. Nishijima, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Kato, T. Yoshizawa, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, S. Masuda, M. Yanagii, M. Nishimura, A. Yamagishi, T. Oshima, and H. Kikuchi.** 2001. Complete genome

sequence of an aerobic thermoacidophilic crenarchaeon, Sulfolobus tokodaii strain7. DNA Res **8:**123-40.

53. **Kay, B. K., J. Winter, and J. McCafferty.** 1996. Phage display of peptides and proteins : a laboratory manual. Academic Press, San Diego.

54. **Kimura, M., and T. Ota.** 1972. On the stochastic model for estimation of mutational distance between homologous proteins. J Mol Evol **2:**87-90.

55. **Klein, M., M. Friedrich, A. J. Roger, P. Hugenholtz, S. Fishbain, H. Abicht, L. L. Blackall, D. A. Stahl, and M. Wagner.** 2001. Multiple lateral transfers of dissimilatory sulfite reductase genes between major lineages of sulfate-reducing prokaryotes. J Bacteriol **183:**6028-35.

56. **Koonin, E. V.** 1992. Archaebacterial virus SSV1 encodes a putative DnaA-like protein. Nucleic Acids Res **20:**1143.

57. **Kraft, P., D. Kummel, A. Oeckinghaus, G. H. Gauss, B. Wiedenheft, M. Young, and C. M. Lawrence.** 2004. Structure of D-63 from sulfolobus spindle-shaped virus 1: surface properties of the dimeric four-helix bundle suggest an adaptor protein function. J Virol **78:**7438-42.

58. **Kraft, P., A. Oeckinghaus, D. Kummel, G. H. Gauss, J. Gilmore, B. Wiedenheft, M. Young, and C. M. Lawrence.** 2004. Crystal structure

of F-93 from Sulfolobus spindle-shaped virus 1, a winged-helix DNA binding protein. J Virol **78:**11544-50.

59. **Letzelter, C., M. Duguet, and M. C. Serre.** 2004. Mutational analysis of the archaeal tyrosine recombinase SSV1 integrase suggests a mechanism of DNA cleavage in trans. J Biol Chem **279:**28936-44.

60. **Levin, B. R., F. M. Stewart, and L. Chao.** 1977. Resource-Limited Growth, Competition, and Predation: A Model and Experimental Studies with Bacteria and Bacteriophage. American Naturalist **111:**3-25.

61. **Lewalter, K., and V. Muller.** 2006. Bioenergetics of archaea: ancient energy conserving mechanisms developed in the early history of life. Biochim Biophys Acta **1757:**437-45.

62. **Lieb, M.** 1953. The establishment of lysogenicity in Escherichia coli. J Bacteriol **65:**642-51.

63. **Lim, J., T. Thomas, and R. Cavicchioli.** 2000. Low temperature regulated DEAD-box RNA helicase from the Antarctic archaeon, Methanococcoides burtonii. J Mol Biol **297:**553-67.

64. **Lipps, G., S. Rother, C. Hart, and G. Krauss.** 2003. A novel type of replicative enzyme harbouring ATPase, primase and DNA polymerase activity. Embo J **22:**2516-25.

65. **Lipps, G., M. Stegert, and G. Krauss.** 2001. Thermostable and site-specific DNA binding of the gene product ORF56 from the Sulfolobus

islandicus plasmid pRN1, a putative archael plasmid copy control protein. Nucleic Acids Res **29:**904-13.

66. **MacArthur, R. H., and E. O. Wilson.** 2001. The theory of island biogeography. Princeton University Press, Princeton.

67. **Magrini, V., M. L. Storms, and P. Youderian.** 1999. Site-specific recombination of temperate Myxococcus xanthus phage Mx8: regulation of integrase activity by reversible, covalent modification. J Bacteriol **181:**4062-70.

68. **Martin, A., S. Yeats, D. Janekovic, W.-D. Reiter, W. Aicher, and W. Zillig.** 1984. SAV1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldaricus isolate B12. EMBO J **3:**2165-2168.

69. **Massana, R., E. F. DeLong, and C. Pedros-Alio.** 2000. A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. Appl Environ Microbiol **66:**1777-87.

70. **Muskhelishvili, G.** 1994. The archaeal SSV integrase promotes intermolecular excisive recombination in-vitro. Syst. Appl. Microbiol **16:**506-508.

71. **Myers, G.** 1998. Viral regulatory structures and their degeneracy. Addison-Wesley, Reading, Mass.

72. **Neumann, H., V. Schwass, C. Eckerskorn, and W. Zillig.** 1989. Identification and characterization of the genes encoding three structural proteins of the Thermoproteus tenax virus TTV1. Molecular and General Genetics MGG **217:**105-110.

73. **Nunes-Duby, S. E., H. J. Kwon, R. S. Tirumalai, T. Ellenberger, and A. Landy.** 1998. Similarities and differences among 105 members of the Int family of site-specific recombinases. Nucleic Acids Res **26:**391-406.

74. **Ortmann, A. C., B. Wiedenheft, T. Douglas, and M. Young.** 2006. Hot crenarchaeal viruses reveal deep evolutionary connections. Nat Rev Microbiol **4:**520-8.

75. **Palm, P., C. Schleper, B. Grampp, S. Yeats, P. McWilliam, W. D. Reiter, and W. Zillig.** 1991. Complete nucleotide sequence of the virus SSV1 of the archaebacterium Sulfolobus shibatae. Virology **185:**242-50.

76. **Paulsen, I. T., R. A. Skurray, R. Tam, M. H. Saier, Jr., R. J. Turner, J. H. Weiner, E. B. Goldberg, and L. L. Grinius.** 1996. The SMR family: a novel family of multidrug efflux proteins involved with the efflux of lipophilic drugs. Mol Microbiol **19:**1167-75.

77. **Peng, X.** 2008. Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of

Sulfolobus and the implications for plasmid survival. Microbiology **154:**383-91.

78. **Peng, X., H. Blum, Q. She, S. Mallok, K. Brugger, R. A. Garrett, W. Zillig, and D. Prangishvili.** 2001. Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. Virology **291:**226-34.

79. **Peng, X., I. Holz, W. Zillig, R. A. Garrett, and Q. She.** 2000. Evolution of the family of pRN plasmids and their integrase-mediated insertion into the chromosome of the crenarchaeon Sulfolobus solfataricus. J Mol Biol **303:**449-54.

80. **Prangishvili, D., H. P. Arnold, D. Gotz, U. Ziese, I. Holz, J. K. Kristjansson, and W. Zillig.** 1999. A novel virus family, the Rudiviridae: Structure, virus-host interactions and genome variability of the sulfolobus viruses SIRV1 and SIRV2. Genetics **152:**1387-96.

81. **Prangishvili, D., P. Forterre, and R. A. Garrett.** 2006. Viruses of the Archaea: a unifying view. Nat Rev Microbiol **4:**837-48.

82. **Prangishvili, D., G. Vestergaard, M. Haring, R. Aramayo, T. Basta, R. Rachel, and R. A. Garrett.** 2006. Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle. J Mol Biol **359:**1203-16.

83. **Qureshi, S. A.** 2007. Protein-DNA interactions at the Sulfolobus spindle-shaped virus-1 (SSV1) T5 and T6 gene promoters. Can J Microbiol **53:**1076-83.

84. **Rachel, R., M. Bettstetter, B. P. Hedlund, M. Haring, A. Kessler, K. O. Stetter, and D. Prangishvili.** 2002. Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. Arch Virol **147:**2419-29.

85. **Reiter, W.-D., and P. Palm.** 1990. Identification and characterization of a defective SSV1 genome integrated into a tRNA gene in the archaebacterium <i>Sulfolobus</i> sp. B12. Molecular Genetics and Genomics (Historical Archive) **221:**65-71.

86. **Reiter, W.-D., P. Palm, A. Henschen, F. Lottspeich, W. Zillig, and B. Grampp.** 1987. Identification and characterization of the genes encoding three structural proteins of the Sulfolobus virus-like particle SSV1. Molecular and General Genetics MGG **206:**144-153.

87. **Reiter, W.-D., P. Palm, S. Yeats, and W. Zillig.** 1987. Gene expression in archaebacteria: Physical mapping of constitutive and UV-inducible transcripts from the <i>Sulfolobus</i> virus-like particle SSV1. Molecular Genetics and Genomics (Historical Archive) **209:**270-275.

88. **Rice, G., K. Stedman, J. Snyder, B. Wiedenheft, D. Willits, S. Brumfield, T. McDermott, and M. J. Young.** 2001. Viruses from extreme thermal environments. Proc Natl Acad Sci U S A **98:**13341-5.

89. **Rice, G., L. Tang, K. Stedman, F. Roberto, J. Spuhler, E. Gillitzer, J. E. Johnson, T. Douglas, and M. Young.** 2004. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. Proc Natl Acad Sci U S A **101:**7716-20.

90. **Roberts, M. S. C., Frederick M.** 1995. Recombination and Migration Rates in Natural Populations of Bacillus subtilis and Bacillus mojavensis. Evolution **49:**1081-1094.

91. **Rohwer, F.** 2003. Global phage diversity. Cell **113:**141.

92. **Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944-5.

93. **Sambrook, J., and D. W. Russell.** 2001. Molecular cloning : a laboratory manual, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

94. **Santangelo, T. J., and J. N. Reeve.** 2006. Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. J Mol Biol **355:**196-210.

95. **Schleper, C., K. Kubo, and W. Zillig.** 1992. The Particle SSV1 from the Extremely Thermophilic Archaeon Sulfolobus is a Virus: Demonstration of Infectivity and of Transfection with Viral DNA. PNAS **89:**7645-7649.

96. **Serre, M. C., C. Letzelter, J. R. Garel, and M. Duguet.** 2002. Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase. J Biol Chem **277:**16758-67.

97. **She, Q., K. Brugger, and L. Chen.** 2002. Archaeal integrative genetic elements and their impact on genome evolution. Res Microbiol **153:**325-32.

98. **She, Q., X. Peng, W. Zillig, and R. A. Garrett.** 2001. Gene capture in archaeal chromosomes. Nature **409:**478.

99. **She, Q., B. Shen, and L. Chen.** 2004. Archaeal integrases and mechanisms of gene capture. Biochem Soc Trans **32:**222-6.

100. **She, Q., R. K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M. J. Awayez, C. C. Chan-Weiher, I. G. Clausen, B. A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P. M. Gordon, I. Heikamp-de Jong, A. C. Jeffries, C. J. Kozera, N. Medina, X. Peng, H. P. Thi-Ngoc, P. Redder, M. E. Schenk, C. Theriault, N. Tolstrup, R. L. Charlebois, W. F. Doolittle, M. Duguet, T. Gaasterland, R. A. Garrett, M. A. Ragan, C. W. Sensen, and J. Van der Oost.** 2001. The complete

genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci U S A **98:**7835-40.

101. **Shmulevitz, M., and R. Duncan.** 2000. A new class of fusion-associated small transmembrane (FAST) proteins encoded by the non-enveloped fusogenic reoviruses. Embo J **19:**902-12.

102. **Siering, P. L., J. M. Clarke, and M. S. Wilson.** 2006. Geochemical and biological diversity of acidic, hot springs in Lassen Volcanic National Park. Geomicrobiology Journal **23:**129-141.

103. **Snyder, J. C.** 2005. VIRUS DYNAMICS, ARCHAEAL POPULATIONS, AND WATER CHEMISTRY OFTHREE ACIDIC HOT SPRINGS IN YELLOWSTONE NATIONAL PARK. MONTANA STATE UNIVERSITY, Bozeman.

104. **Sonnhammer, E. L., G. von Heijne, and A. Krogh.** 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol **6:**175-82.

105. **Stedman, K., C. Schleper, E. Rumpf, and W. Zillig.** 1999. Genetic requirements for the function of the archaeal virus SSV1 in *Sulfolobus solfataricus*: construction and testing of viral shuttle vectors. Genetics **152:**1397-1405.

106. **Stedman, K. M., Q. She, H. Phan, H. P. Arnold, I. Holz, R. A. Garrett, and W. Zillig.** 2003. Relationships between fuselloviruses infecting the

extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2. Res Microbiol **154:**295-302.

107. **Tatusova, T. A., and T. L. Madden.** 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett **174:**247-50.

108. **Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins.** 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25:**4876-82.

109. **Tirumalai, R. S., E. Healey, and A. Landy.** 1997. The catalytic domain of lambda site-specific recombinase. Proc Natl Acad Sci U S A **94:**6104-9.

110. **Tojo, N., and T. Komano.** 2003. The IntP C-terminal segment is not required for excision of bacteriophage Mx8 from the Myxococcus xanthus chromosome. J Bacteriol **185:**2187-93.

111. **Wang, Y., Z. Duan, H. Zhu, X. Guo, Z. Wang, J. Zhou, Q. She, and L. Huang.** 2007. A novel Sulfolobus non-conjugative extrachromosomal genetic element capable of integration into the host genome and spreading in the presence of a fusellovirus. Virology **363:**124-33.

112. **Waters, E., M. J. Hohn, I. Ahel, D. E. Graham, M. D. Adams, M. Barnstead, K. Y. Beeson, L. Bibbs, R. Bolanos, M. Keller, K. Kretz,**

X. Lin, E. Mathur, J. Ni, M. Podar, T. Richardson, G. G. Sutton, M. Simon, D. Soll, K. O. Stetter, J. M. Short, and M. Noordewier. 2003. The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A **100:**12984-8.

113. **Wegley, L., P. Mosier-Boss, S. Lieberman, J. Andrews, A. Graff-Baker, and F. Rohwer.** 2006. Rapid estimation of microbial numbers in water using bulk fluorescence. Environ Microbiol **8:**1775-82.

114. **Weinbauer, M. G., and F. Rassoulzadegan.** 2004. Are viruses driving microbial diversification and diversity? Environ Microbiol **6:**1-11.

115. **Whitaker, R. J., D. W. Grogan, and J. W. Taylor.** 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science **301:**976-8.

116. **White, D.** 2007. The physiology and biochemistry of prokaryotes, 3rd ed. Oxford University Press, New York.

117. **Wiedenheft, B., K. Stedman, F. Roberto, D. Willits, A. K. Gleske, L. Zoeller, J. Snyder, T. Douglas, and M. Young.** 2004. Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses. J Virol **78:**1954-61.

118. **Williams, K. P.** 2003. Traffic at the tmRNA gene. J Bacteriol **185:**1059-70.

119. **Wills, C.** 1996. Yellow fever, black goddess : the coevolution of people and plagues. Addison-Wesley Pub., Reading, MA.

120. **Woese, C. R., and G. E. Fox.** 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A **74:**5088-90.

121. **Worthington, P., V. Hoang, F. Perez-Pomares, and P. Blum.** 2003. Targeted disruption of the alpha-amylase gene in the hyperthermophilic archaeon Sulfolobus solfataricus. J Bacteriol **185:**482-8.

122. **Xiang, X., L. Chen, X. Huang, Y. Luo, Q. She, and L. Huang.** 2005. Sulfolobus tengchongensis spindle-shaped virus STSV1: virus-host interactions and genomic features. J Virol **79:**8677-86.

123. **Yeats, S., P. McWilliams, and W. Zillig.** 1982. A plasmid in the archaebacterium Sulfolobus acidocaldarius. EMBO J **1:**1035-1038.

124. **Zillig, W., H. P. Arnold, I. Holz, D. Prangishvili, A. Schweier, K. Stedman, Q. She, H. Phan, R. Garrett, and J. K. Kristjansson.** 1998. Genetic elements in the extremely thermophilic archaeon Sulfolobus. Extremophiles **2:**131-40.

125. **Zillig, W., K. O. Stetter, and D. Janekovic.** 1979. DNA-dependent RNA polymerase from the archaebacterium Sulfolobus acidocaldarius. Eur J Biochem **96:**597-604.

Appendix A: Long Inverse PCR Using iProof™ Polymerase

tech note 5337

Adam Clore and Kenneth Stedman, Biology Department and Center for Life in Extreme Environments, Portland State University, PO Box 751, Portland, OR 97207 USA

Introduction

Viruses have long been used as model systems to probe fundamental questions in molecular biology. The use of viruses to this end dates back to the 1930s, when the study of the T4 bacteriophage led to, among other things, the elucidation of the function of messenger RNA and the deciphering of the genetic code (Mathews et al. 1983). Using viruses as models for molecular study remains important today as we strive to understand new systems, tackle emerging diseases, and develop new tactics to fight pathogens.

Our laboratory's research focuses on the SSV1 virus. This UV-inducible virus was isolated from Sulfolobus shibatae, an acidic hyperthermophilic archaeon that lives in acidic sulfur springs with pH near 3 and temperature of around 80°C (Grogan et al. 1990). The 15.5 kilobase pair double-stranded circular DNA genome of the SSV1 virus contains several short repeated sequences and 34 open reading frames (ORFs), of which only four have known functions (Palm et al. 1991). The remaining ORFs show no similarity to any genes in public databases.

To investigate the function of the uncharacterized ORFs in SSV1, our laboratory has developed a method of long inverse PCR to quickly and effectively produce site-directed mutants. Inverse PCR was first described by Howard Ochman and colleagues (1988) and was designed to amplify regions of unsequenced DNA that flank regions of known sequence. In this technique, the DNA is first digested with a restriction enzyme and the fragment containing the known sequence and flanking regions is ligated to form a circle. Next, using primers oriented outward from the area of known sequence, the rest of the fragment (i.e., the flanking regions) is amplified and can then be sequenced.

Unlike the method described by Ochman et al., the procedure we use amplifies the entire viral genome or slightly less (up to 20 kb). After amplification, the linear amplicon can be ligated together to produce a deletion mutant, the amplicon can be ligated to an insert to produce replacement mutants, or the entire genome can be amplified and ligated using primers containing mismatches to produce site-directed mutants. Transformation with these mutants produces a higher percentage of positive clones than transposon mutagenesis and other methods. This technique should be useful for rapidly producing site-directed mutations in other viruses with relatively large circular genomes, in plasmids, and in episomal DNA where other methods used to induce mutations prove ineffective.

In this report, we use iProof polymerase to amplify the entire 15.5 kilobase pair genome of the SSV1 virus from a shuttle vector consisting of the viral genome and an inserted bacterial plasmid. Further, we amplified the entire viral genome and replaced the original bacterial plasmid with one conferring resistance to a different antibiotic. Finally, this product was amplified from another site in the viral genome to remove a specific gene. As a result of the high fidelity of the iProof polymerase, both of these constructs show no detectable mutations and their ability to to infect and reproduce in their host is similar to that of the wild-type virus.

Methods

Shuttle Vector Construction

A fusion between the bacterial plasmid pBluescript SK+ and the SSV1 virus was constructed as previously described by Stedman et al. (1999). Briefly, the 2,961 bp bacterial plasmid was inserted into a neutral site in the viral genome and was found to replicate similarly to the wild-type virus. Packaging this extra DNA seems to pose no problem for the virus since replication, stability, insertion, and virion structure all are comparable to wild type (Stedman et al. 1999). Shuttle vector genomes were purified from E. coli using alkaline lysis as described in Stedman et al. (1999).

Amplification Primer Design

To amplify the entire SSV1 genome from the original shuttle vector, standard M13 forward (-20) and M13 reverse (-27) primers were used with their sequences unchanged (Table 1). Primers for the second amplification (Del right and Del left), which used product from the first PCR as template, were designed to remove the complete gene from the virus and to allow the directional cloning of different genes. To this end, primers were designed so that their 5' ends flanked the ORF to be removed, overlapping the start codon. The length of the primer was extended in the 3' direction for approximately 25 bases, and stopped when a GC clamp of at least one base was present (Table 1).

Table 1. Primers used in PCR. Bold letters show restriction sites, green letters indicate mispaired bases. Italics indicate the start codon of the removed gene.

Name	Sequence	Tm,°C
M13 F (–20)	GTAAAACGACGGCCAGT	53.0
M13 R (–27)	CAGGAAACAGCTATGAC	47.3
Del right	CGTCTTATCTTTCGTCATTTCACCTGGTACTATTATGG	58.3
Del left	GGGGTCTGACAGGCGCCGTATCACTATC	55.4

Primer sequences were checked for hairpins and other secondary structures using mfold software (http://www.bioinfo.rpi.edu/applications/mfold/old/dna/; Zuker 2003), with Na+ concentrations set at 50 mM and Mg2+ concentrations set at 0 mM. Primers were redesigned with different sequences if the Tm of the hairpin structure was within 15°C of the predicted Tm of the duplexed primer/template pair. Bases were modified to allow the insertion of restriction endonuclease cleavage sites for directional cloning (Table 1). Final Tm predictions were calculated with Hyther software (http://ozone2.chem.wayne.edu/Hyther/hytherm1main.html), which predicts nucleic acid hybridization thermodynamics, taking into account mispairing. The Tm was adjusted by increasing or decreasing the 5' end of the primer to allow the predicted Tm values of forward and reverse primers to be within 3°C of each other and between 55 and 60°C.

PCR Conditions

Amplification was carried out using the primers listed in Table 1 and the PCR reagents listed in Table 2, in a DNA Engine Dyad(r) thermal cycler equipped with a gradient block. Temperature calculations were estimated by the instrument and all reactions were carried out in 20 *μ*l volumes.

Table 2. PCR parameters.

Reagent	M13 amplification	Del amplification
Buffer	1x HF buffer	1x HF buffer
dNTPs	0.2 mM/base	0.2 mM/base
Template	3 pM	6 pM
Forward primer	M13 F, 250 nM	Del right, 250 nM
Reverse primer	M13 R, 250 nM	Del left, 250 nM
Polymerase	0.02 U/μl, iProof	0.02 U/μl iProof

Optimization of specific annealing temperatures is critical for decreasing nonspecific product production, especially with templates that contain repetitive elements such as the SSV1 genome. Therefore, temperature optimization was carried out for both primer sets. Figure 1 shows a temperature optimization with the M13 primers, starting at 3°C below the calculated annealing temperature of 53°C for standard PCR and increasing to above the optimal Tm. In most cases, the optimal temperature was 7-10°C higher than the calculated annealing temperature for standard PCR conditions.

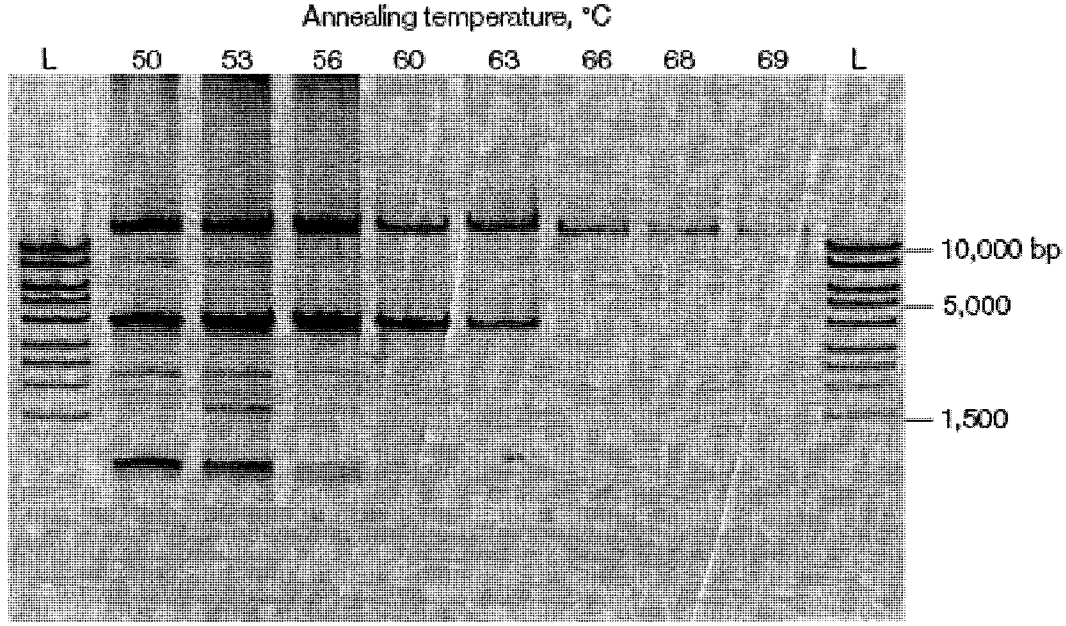

Fig. 1. Annealing temperature optimization. Long inverse PCR of the SSV1 shuttle vector was performed using varying annealing temperatures (indicated at top of gel). PCR products were run on an agarose gel to determine which was the lowest annealing temperature to prevent nonspecific amplification, in this case, 66°C. Lane L, MassRuler, high range DNA ladder (Fermentas).
The conditions used for amplification of each construct followed the manufacturer's guidelines. This included an initial denaturation step at 98°C for 3 min, followed by 30 cycles of 15 sec denaturation at 98°C, 15 sec annealing with the temperature optimized as described above (66°C for M13, 64°C for Del), and an 8 min extension at 72°C. A final 8 min extension was done at 72°C after the 30 cycles.

Ligation

Because iProof polymerase generates blunt-end DNA fragments during amplification, the M13 amplicon was cloned into the pCR Blunt II-TOPO vector (Invitrogen) following the manufacturer's instructions. This kit is designed to accept inserts that lack a 5' phosphate; therefore, no modification to the amplicon was necessary. The PCR product was gel-purified and the gel containing the band was digested with _-agarase I (New England Biolabs). The product was then quantitated relative to a standard by gel fluorescence, and 8 ng was added to one TOPO reaction kit as directed by the manufacturer.

Ligation of the deletion construct was carried out in a similar manner. The PCR product was circularized by adding 5' phosphates to the amplicon and ligating the blunt ends produced by iProof polymerase to each other. Gel-purified PCR product (500 ng) was added to a reaction of 1x T4 ligase buffer containing 1 mM ATP (New England Biolabs), 2 μl PEG4000 (Sigma) and 10 U polynucleotide kinase (New England Biolabs) in a total volume of 40 μl. The reaction was incubated at 37°C for 1 hr, after which 20 U T4 ligase were added and incubated at 16°C for 4 hr.

Transformation Into E. coli

For the M13 construct, the entire ligation was transformed by heat shock into the StAble 3 strain of E. coli cells (Invitrogen). Transformation typically gave low yields (104 colonies/μg transformed). After 48 hr of growth, the smallest colonies were selected from the plates and grown in LB broth. Plasmids were purified from 5 ml liquid cultures by alkaline lysis. Preparations were screened for full-length constructs by restriction fragment length polymorphism (RFLP) analysis.

For the Del amplicon, a 10 μl aliquot of the reaction (125 ng) was transformed into chemically competent StAble 3 cells and plated as described above.

Transformation Into Sulfolobus

Shuttle vectors purified from E. coli were transformed by electroporation into the host S. solfataricus as described previously (Stedman et al. 2003).

Viral Production

Viral production of the amplified viral genomes was detected in three ways. First, transformed S. solfataricus was spotted onto lawns of uninfected S. solfataricus and the cultures were examined for the presence of viral plaques. Second, RFLP analysis of purified shuttle vector genomes from infected strains was used to detect the presence of reproducing virus. Finally, PCR was used to amplify the area surrounding the removed gene and the PCR products were run on a gel to ascertain that bands of the correct size were produced in the different mutants. PCR conditions were as follows: initial denaturation at 95°C for 5 min, subsequent denaturation at 95°C for 15 sec, annealing at 52°C for 15 sec, and extension at 72°C for 1.5 min. After 30 cycles, a final extension at 72°C for 5 min was used.

Results and Discussion

Both the M13 amplification of the 15.5 kilobase pair viral genome and the subsequent 18.5 kilobase pair Del construct amplified from the PCR product of the M13 amplification yielded functional viruses upon transformation into the laboratory host, S. solfataricus.

RFLP analysis (Figure 2) showed that, of the first five colonies screened, one contained the correct insert. This method required substantially less screening of colonies than other methods, such as transposon mutagenesis and partial restriction digestion and ligation (data not shown).

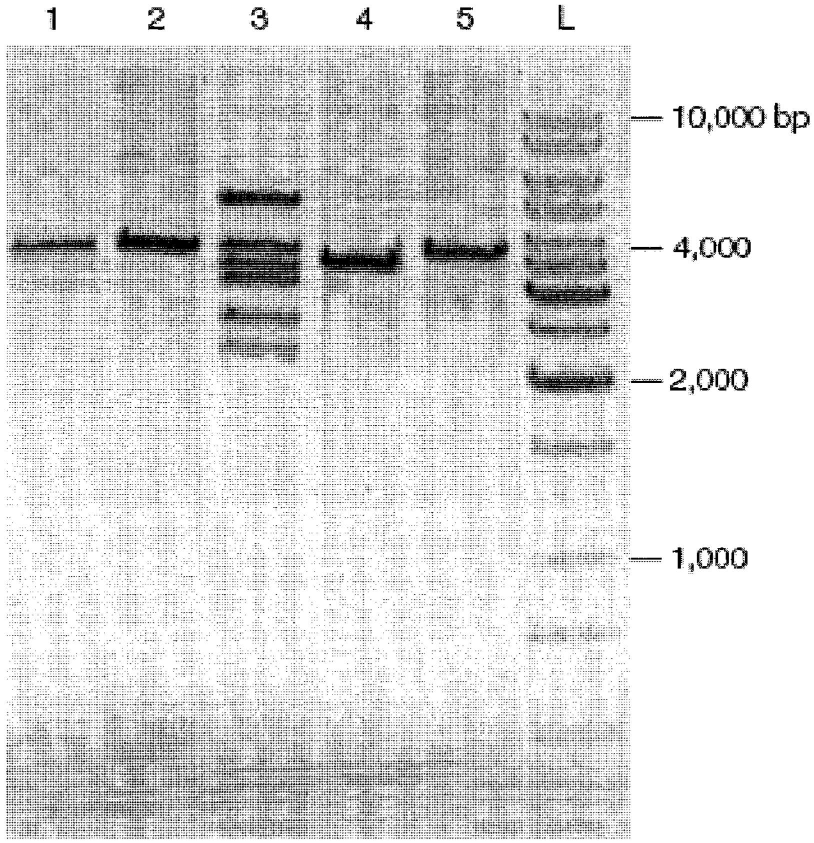

Fig. 2. RFLP screening of clones. DNA purified from E. coli transformed with the M13 PCR product was treated with EcoRI to determine whether the full-length clone was present. Of five transformations tested, only one, in lane 3, contained the 19.2 kilobase pair full-length clone. Lane L, 1 kilobase pair GeneRuler DNA ladder (Fermentas).

We used three different methods to verify that virus was being produced by the amplified viral genomes in S. solfataricus. First, viral plaques were seen after S. solfataricus that was transformed with the shuttle vector was spotted onto lawns of uninfected S. solfataricus (data not shown). Second, purification of shuttle vector genomes from infected strains and RFLP analysis showed the presence of reproducing virus (data not shown). Finally, PCR of the area surrounding the removed gene showed bands of correct size in the different mutants (Figure 3).

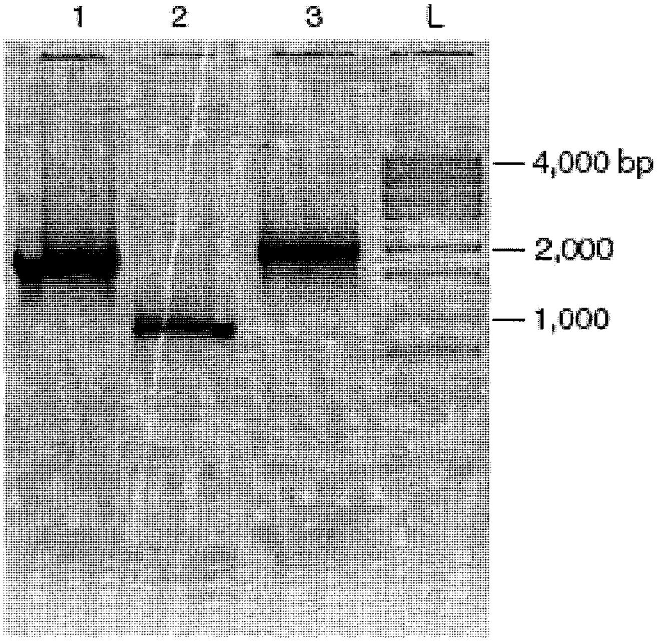

Fig. 3. Amplification of sequences surrounding the removed gene. Sequences were amplified by PCR and analyzed in a gel. Lane 1, M13 amplicon; lane 2, Del amplicon; lane 3 wild-type virus; lane L, 1 kb ladder (Fermentas).

Proper amplification of the template required stricter adherence to specific reaction parameters, including template concentration, dNTP concentration, and reaction volume, than traditional PCR (data not shown). Template concentration had to be optimized for each primer set as well as for each DNA extract. It was often found that amplification was successful over only a narrow concentration range (less than an order of magnitude). The presence of varying amounts of contaminating proteins, sheared DNA, or both in individual preparations and the sensitivity of the reaction may have led to varying amounts of template required for each preparation and the necessity for individual optimization with each extract. Successful reactions occurred only at dNTP concentrations of 200 μM. Reaction volumes also affected the efficiency of the reaction, with the best results seen in small (20 μl) reactions.

We compared other high-fidelity polymerases in the same procedure, but none was effective at amplifying the template without producing smaller, nonspecific bands (data not shown). In addition to this, iProof polymerase had the fastest extension time of any of the high-fidelity polymerases, allowing completion of reactions in 8 hr as opposed to over 20 hr for other polymerases.

In summary, this method represents a rapid and efficient method for amplifying and mutating large plasmids and circular viral genomes.

References

Grogan D et al., Isolate B12, which harbours a virus-like element, represents a new species of the archaebacterial genus Sulfolobus, Sulfolobus shibatae, sp. nov, Arch Microbiol 154, 594-599 (1990)

Mathews CK et al., pp 1-7 in Bacteriophage T4, American Society for Microbiology, Washington DC (1983)

Ochman H et al., Genetic applications of an inverse polymerase chain reaction, Genetics 120, 621-623 (1988)

Palm P et al., Complete nucleotide sequence of the virus SSV1 of the archaebacterium Sulfolobus shibatae, Virology 185, 242-250 (1991)

Stedman K et al., Genetic requirements for the function of the archaeal virus SSV1 in Sulfolobus solfataricus: construction and testing of viral shuttle vectors, Genetics 152, 1397-1405 (1999)

Stedman KM et al., Relationships between fuselloviruses infecting the extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2, Res Microbiol 154, 295-302 (2003)

Zuker M, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res 31, 3406-3415, (2003)

This work was partially supported by a NSF-Microbial Observatories Grant: MCB-0132156.

This tech note was current as of the date of writing (2005) and not necessarily the date this version (Rev A, 2005) was published.

Hyther is a trademark of Wayne State University. pBluescript is a trademark of Stratagene. pCR and TOPO are trademarks of Invitrogen Corp.

Practice of the patented polymerase chain reaction (PCR) process requires a license. The DNA Engine Dyad thermal cycler is an Authorized Thermal Cycler and may be used with PCR licenses available from Applied Biosystems. Its use with Authorized Reagents also provides a limited PCR license in accordance with the label rights accompanying such reagents. Some applications may also require licenses from other third parties.

Appendix B: Proposed Changes to Virus nomenclature

The current body of knowledge of Fuselloviruses was compiled at different times by many people. Subsequently, the naming of viruses and their predicted genes are not uniform, causing unnecessary confusion. To help correct this problem the following Fusellovirus nomenclature standards are proposed and will be submitted to the International Committee on the Taxonomy of Viruses. For consistency and to ease interpretation these standards are used throughout this dissertation.

Viruses

All viruses other than the type species SSV1 will be named starting with "Sulfolobus Spindle-Shaped" followed by a 1 to 3 letter abbreviation of the isolation location, followed by a number sequential to the order in which the viruses from that location were entered into the public domain. These changes will affect the following viruses changing their name as dfollows: SSV2 to SSV-I2, SSV-RH to SSV-RH1, and SSV4 to SSV-I4. The virus SSV-K1 will remain unchanged as it fits the proposed nomenclature. SSV-I1 will be skipped to avoid confusion with previous publications by changing names as little as possible.

Sequence Numbering and Naming

Open reading frames will be named by the number of amino acids encoded in the ORF, as has been done for all published SSV genomes to date. Since origins of replication have not been unambiguously determined, nucleotide sequence numbering will begin with the first base after the stop codon of the universally conserved VP3 structural protein and will proceed clockwise with the coding strand. This is the numbering used in the SSV-RH1 and SSV-K1 genome annotations (117). ORFs encoding the same number of amino acid residues will be differentiated by a lower case letter following the ORF number, starting alphabetically from the first ORF encountered by moving clockwise from the nucleotide sequence start. Due the large number of name changes required to bring the previously annotated genomes into compliance it is suggested that this only be used for future annotation. For consistency these rules will be followed in this dissertation.