RECOVERY AND CHARACTERIZATION OF INSERTION SEQUENCE ELEMENTS FROM NATURAL ISOLATES OF THE HYPERTHERMOPHILIC ARCHAEON *SULFOLOBUS*

A thesis submitted to the

Division of Graduate Studies and Research
of the University of Cincinnati

in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in the Department of Biological Sciences
of the College of Arts and Sciences

2003

by

Zachary David Blount

B.S., Georgia Institute of Technology, 1999

Committee Chair: Dr. Dennis W. Grogan

ABSTRACT

Little study has been made of insertion sequences (IS) residing in *Sulfolobus* populations outside of Italy and Japan. Since 1999, the Grogan lab has isolated approximately 1700 *Sulfolobus* strains from Yellowstone National Park, Lassen National Park, New Zealand, the Kamchatka peninsula, and Italy. In this study, IS were recovered from these strains using a gene trap based upon 5-fluoroorotic acid selection for disruption of loci within the *pyr* operon. Seven distinct IS were recovered, two from both Kamchatka and North American strains. While two were novel, five were very similar to elements in the genomes of *S. solfataricus* and *S. tokodaii*. PCR screens specific for recovered IS showed some to be cosmopolitan, with evidence of their presence in all locations, while others were more restricted. These results indicate that widely separated *Sulfolobus* populations share a pool of closely related IS, though some elements may be active only in certain populations.

ACKNOWLEDGEMENTS

As with all achievements of any real significance or meaning, the work this thesis details owes much to many, and I would like to offer them my humble gratitude. I would like to thank my advisor and committee chair, Dr. Dennis Grogan for his help and advice, as well as the patience he showed toward my quirks and occasional delusions of grandeur (That grand synthesis of microbial molecular evolution will take a while longer than I had thought). I am grateful also that he allowed me the time to stumble around and get my bearings during that first year here, and that he supported my idea to remake Biology 552. I would also like to thank my committee members, Dr. Brian Kinkle and Dr. Ronald DeBry for not only their guidance and help, but also for the excellent instruction I received from them in the classes they taught.

I have had great luck in the people with whom I have shared the Grogan lab during my time here. I would like to thank graduate students Greg Bell and Mike Craig for showing me the ropes and helping me get settled in, as well as for their friendship. I have also appreciated the four undergraduates who have helped out in the lab. I would like to thank Josh Hansen for all the solutions he made, for pouring more plates than any one person should, and for braving the darkness of the Cincinnati subway with me not once, nor twice, but three times. Jason Holmes I thank for all the prep work and dishes he has done, as well as for being an engineering student…having him around was almost like being at Georgia Tech again. Lauren Hoffman I thank for providing amusement and for doing so many PCR reactions and gels. Amy Dill I thank just for being A-m-y. Please never change. I also thank Amy's little friend

Alfred the lab monkey for the inspiration his dedication has provided. All six I thank for making my time in the lab so fun and interesting.

I would also like to thank many of the friends and colleagues I have had the distinct honor and pleasure of knowing and learning from during my time here:

Rae Braudaway – my best friend, whose love, emotional support, and encouragement has helped me so much.

Heather Henry - without whom my first quarter would have been much harder, and who was always being willing to listen and offer an encouraging word.

Jerry Hinn – with whom I had many interesting discussion on subjects ranging from natural selection and departmental politics, to Buffy and Babylon 5.

Ami Terwilliger – who was a wonderful person to work with, a great, caring teacher, and a good friend. I wish her luck in medical school and beyond, and I hope her dreams remain magical.

Julie Stacey and Cathy Haywood – two people who never get enough appreciation for all the hard work they put into keeping the lab classes up and running. It has been an honor to call them friends and colleagues. I appreciate all the help they gave me.

Sabrina Mueller – who was a valuable colleague with whom I could discuss microbiology and off whom I could bounce new ideas. I hope to retain her counsel well into the future.

Joy Ingram – a person, aptly named, who has been one of only a few of my students to keep coming by to say hi. I will miss her.

My family, most especially Mom, Dad, Granny, and Paw Paw, has been a great source of strength for me. I thank them so very much for all their

encouragement, love, and emotional and financial assistance. I would also like to thank all my friends from home, most especially Lee Goodson, Tim Clemons, Matt and Karen Williams, Laurie Lawlor, Cherie Ropp, Ginny Coulter, and everyone else at the best independent bookstore in the South, for always being there and for being so understanding about how long it takes for me to reply to emails and letters.

My time in Cincinnati would not have been quite so pleasant had it not been for the wonderful system of parks with which this city and county are blessed. Everyone should be proud to pay the taxes that go to their upkeep and to fund the Cincinnati Parks Department and Board of Recreation as well as the Hamilton County Park District. I do not know how I would have remained sane had it not ben for all the beautiful places to walk and hike within minutes of my apartment.

I would also like to thank she of the clear, gray eyes for her inspiration.

Finally, I want to thank my college microbiology professors, especially Dr. Thomas Tornabene, for instilling in me such an abiding love of the microbial world, my high school AP biology teacher, Ms. Pugh, and, most of all, Mary Ann Watson, my second grade teacher. If it had not been for her, I do not know that I would have been here.

Εν οιδα οτι ογδεν οιδα.

-Socrates

I'm and explorer, okay? I like to find things out.

-Richard P. Feynman

Per usum meum…talis quails…

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Number Title Page

LIST OF ABBREVIATIONS

IS……Insertion Sequence(s)

IR……..…..Inverted Repeat

IRR……..…Right (Upstream) Inverted Repeat

IRL……...…Left (Downstream) Inverted Repeat

DR……...Direct Repeat

DRR……...…….Right (Upstream) Inverted Repeat

DRL……..…Left (Downstream) Inverted Repeat

ORF…….Open Reading Frame

ISC……..Insertion Sequence of Crenarchaeota

ISSt……..…Insertion Sequence of *Sulfolobus tokodaii*

DNA…….....Deoxyribonucleic Acid

RFLP……..….Restriction Fragment Length Polymorphism

5-FOA/FOA……..5-Fluoroorotic Acid

kDa……kilodalton

mL……...milliliter

µL……..…microliter

µg……..…microgram

ºC…….….degrees Celsius

pI……iso-electric point

Tn……transposon

M……...……molar concentration

DT……...…..Dextrose-Tryptone medium

DT.ura……Dextrose-Tryptone medium supplemented with uracil

DT.ura FOA…….Dextrose-Tryptone medium supplemented with uracil and FOA

TE……. Tris-Ethylenediamine tetra acetic acid buffer

EDTA……...Ethylenediamine tetra acetic acid

YNP99....Yellowstone National Park strain collection derived from year 1999 samples

YNP00....Yellowstone National Park strain collection derived from year 2000 samples

YNP01....Yellowstone National Park strain collection derived from year 2001 samples

L2…….……Lassen National Park strain collection derived from year 2000 samples

K2K……..…Kamchatka peninsula strain collection derived from year 2000 samples

NZ00……..…New Zealand strain collection derived from year 2000 samples

IT02……...……Naples, Italy strain collection derived from year 2002 samples

PCR…….Polymerase chain reaction

DMSO……...Di-methyl sulfoxide

dNTP…….Deoxynucleotide triphosphate

dH_2O……..Distilled water

A_{600}……Absorbance at 600 nm

aa…….amino acid

nt……...…nucleotide

bp……....base pair

G + C……Percentage of DNA sequence made up of guanine and cytosine

sacB……...….*Bacillus subtilis* gene encoding the enzyme levansucrase

lacZ……..…..*E. coli* gene encoding the enzyme β-galactosidase

lacS…….…….*Sulfolobus* gene encoding the enzyme β-galactosidase

nM……...nano Molar

kb…….kilobase pairs

UV……....ultraviolet light

TAE…….……..Buffer with Tris-HCl, acetic acid, and EDTA

TBE……..….Buffer with Tris-HCl, Boric acid, and EDTA

MITE……...Miniature Inverted repeat Transposable Element

Chapter 1
Insertion Sequences: A Review

I: Introduction

Insertion sequences (IS) are the most widely distributed autonomous transposable genetic elements, being present in all three domains of life. They are characterized by their small size, most falling into a range of only 700 bp to 2 kb, and their genetic simplicity. In general, IS are composed of open reading frame coding for a transposase that is flanked by terminal inverted repeats that define the ends of the element. The transposase recognizes these inverted repeats during its catalysis of the movements of the element from place to place within the host genome through the process of transposition. IS thus possess all the features required for transposition, minimizing their dependency of their genomic environment, and allowing them to insert into a wide variety of sites in the host genome with little or no sequence similarity. The ramifications of this property, together with their capacity to cross species boundaries, make them highly significant players in molecular evolution, while also making them very well suited to use as tools in biogeographic study.

II. IS Functional Structure

A parsimonious relationship between structure and function is very characteristic of IS, and this is reflected in their two major structural features. The terminal inverted repeats (IRs) range in size from ten to forty base pairs, and may be divided into two functional regions. The first is composed of the outermost two to three base pairs that are involved in the strand breakage and transfer reactions of

transposition. The second takes up the remainder of the IS and constitutes a transposase binding site that permits proper enzymatic positioning during transposition. Typically, the IR upstream of the central reading frame encoding the element's transposase also contains the ORF's promoter. As will be discussed, this permits control of transposase expression by transposase binding. (Mahillon & Chandler 1998, Chandler & Mahillon 2002, Craig 2002).

This stereotypical structural format is not universal. Three families of IS (IS91, IS110, and IS200/IS605) lack both IRs and DRs. Far more common is the possession of two to three ORFs in the central coding region. In some cases these multiple ORFs encode subunits of a multimeric transposase, while in others ribosomal frame shifting results in two ORFs being conjoined to encode a single protein (Chandler & Mahillon 2002). Occasionally other enzymes such as resolvases are encoded by a second or third ORF (Craig 2002).

III. Transposase Structure

Analysis of the few transposases that have been studied in depth has revealed most to have two functional domains. The N-terminal domain has been shown to have DNA-binding activity and displays a helix-turn-helix motif typical of proteins with this function (Haren et al. 1999, Chandler& Mahillon 2002). This portion of the transposase interacts with the second region of the IRs, thus permitting the enzyme to recognize the ends of the element. The C-terminal, or catalytic domain mediates the strand cleavage and rejoining reactions of transposition. In the vast majority of transposases, the catalytic domain possesses three conserved regions centered on the almost universal amino acid triad of DDE. These constitute the DD(35)E motif that

defines the phosphoryltransferase family of enzymes that includes RnaseH, RuvC, and many retroviral integrases (Haren et al. 1999, Chandler and Mallion 2002, Kulkowsk et al. 1992). The motif is integral to the active site of these enzymes, where it is responsible for the coordination of Mg^{+2} and other divalent metal ions that, in transposition, are involved in aiding nucleophilic attacking groups (Mahillon and Chandler 1998, Haren et al. 1999).

IV: Transposition

IS transposition is a multi-stage, DNA-mediated, site-specific recombination event that results in the transfer of an IS copy from one location in a host genome to any of a number of other non-homologous locations (Craig 2002). This process may be conservative, in which an IS is excised from its original site and transferred to the new one with no increase in copy number, or replicative, in which transfer takes place in conjunction with replication and copy number does increase. Conservative transposition begins with the assembly of the transposome, a highly organized nucleoprotein complex made up of the IS and a transposase multimer that permits the positioning at both terminal inverted repeats at least one N-terminal and one C-terminal transposase domain. In some instances, accessory proteins provided by the host cell are also a part of the assembly. This assembly phase is followed by donor DNA cleavage, in which the multimer's catalytic domains mediate nucleophilic attack by water molecules on the 3' ends of the IS, cleaving them from the host DNA, and leaving 3' hydroxyls. In most cases, these free hydroxyls then engage in a nucleophilic attack of the 5' linkages of the second strand, forming hairpin structures that are subsequently resolved to leave 5' phosphates, and resulting in complete

excision from the donor site (Craig 1997, Haren et al. 1999, Chandler & Mahillon 2002, Mizuuchi & Baker, 2002).

Unlike some large transposons and eukaryotic elements, it is at this point that most IS select their target site. This selection may be made directly through transposase interaction with the target, or may be influenced by host-specific accessory proteins, and the actual specificity of the process varies greatly from element to element. For most IS, the length of the target site seems more important and conserved than is its sequence. While some target preferences have been observed, it is quite rare that an IS has a single, specific preferred target sequence; instead consensus sequences are far more common. Other groups of IS prefer genomic regions marked by certain general sequence composition characteristics, such as high or low G+C composition. While none yet studied in detail seem to make entirely random selection, statistical analysis of a large number of sequenced insertion sites is often necessary to detect a pattern. In general, two factors seem to be dominant in impacting target selection for most elements. The first is the accessibility of a potential target, or the degree of close interaction that is possible between it and the transposase. This explains the observed bias of IS for insertion into bent DNA segments. More crucial to the survival of the element is the need for the element to largely avoid transposition into regions of the host chromosome that will negatively impact host viability. Due to the selective constraints stemming from this, it is common to find that many IS show preference for insertion into non-essential genes and the regions upstream of promoters where transcription will not be impeded (Craig 1997, Haren et al. 1999, Chander & Mahillon 2002).

Once a target site is selected, strand transfer takes place in which the transposase catalyzes nucleophilic attack on the target site by the element's terminal 3' hydroxyls. This incorporates the element into the new site in such a way that each end is initially bound to only one strand of its new location. In most IS this incorporation is staggered and leaves complementary single strand gaps at either end. This is eventually repaired in post-transfer processing by host enzymes. The result is duplication of the target sequence, and is the source of the direct repeats that flank typically flank IS in situ (Haren et al. 1999, Chandler & Mahillon 2002).

A major variation on the classical conservative mechanism may be referred to as circular intermediate transposition. In this a single 3' strand breakage initially takes place. The consequent 3' hydroxyl then attacks the 5' end of the same strand, cleaving it from the donor DNA. The result is circularization of one strand of the element. This is repeated for the second strand, and the element, now essentially a plasmid bound to a transposase multimer, then proceeds with strand transfer and target site integration in much the same manner as previously described. This mechanism was first studied in IS911, but is believed to be quite common in other elements and families (Chandler and Mahillon 2002, Rousseau et al. 2002, Craig 1997, Haren et al. 1999, Mizuuchi & Baker 2002).

Replicative transposition differs in that 5' cleavage of the element does not follow transposome assembly and 3' cleavage. Instead, the terminal 3' hydroxyls directly attack the target site, resulting in the element being simultaneously bound to both the donor and target sites. This structure, referred to as a Shapiro intermediate, is characterized by the presence of three-way junctions of double-stranded DNA on

either end. The host DNA polymerase recognizes these junctions as replication forks, replicates the element, and fills in the single stranded gaps from staggered integration. If the transposition is intermolecular, it causes cointegrate formation from the donor and recipient molecules, and if intramolecular, causes inversion of the intervening sequence. In the case of the former, resolution is possible by the action of an IS-encoded resolvase. In that of the latter, homologous recombination between the two IS copies can correct the inversion, though this occurs at an efficiency of only about fifty percent (Grindley 2002, Chandler and Mahillon 2002, Craig 1997).

High rates of transposition can greatly disrupt the genome of a host organism, negatively impacting its fitness, and thereby reducing IS fitness. IS are thus under selection for low levels of transposition, and they have evolved a wide range of mechanisms to regulate the process. In most IS, some control of transposase expression is maintained by the positioning of the promoter of its ORF in the transposase-binding domain of one of the inverted repeats. Thus, soon after translation starts, the N-terminal domain of transposase can block further transcription of the ORF. This mechanism is further refined in most elements of the IS1 and IS3 families. In these elements a low frequency translational frameshift is required to generate a full transposase protein. As a result, typically only the N-terminal domain of the transposase is produced, allowing transcription to be blocked while precluding catalytic activity (Escoubas et al. 1991). Regulation is also provided by the susceptibility of most transposases to host proteases, reducing their stability and thereby their activity. Transposome assembly is a final common target of regulatory mechanisms, as transposases tend to be inactive outside of a fully assembled

transposome. Such mechanisms typically take the form of requiring more complex protein-protein and protein-DNA interactions for proper assembly, thereby reducing its frequency of occurrence and precluding transposition (Craigie 1996, Craig 1997, Mahillon & Chandler, 1998, Chandler & Mahillon 2002, Mizuuchi & Baker 2002, Harren et al. 1999).

V: IS Families

Categorization of IS is difficult due to their small size and wide variety. This task is made more difficult by the great prevalence of isoforms, or IS copies with less than 10% DNA or 5% amino acid sequence divergence. Mahillon and Chandler (1998, 2002) developed the family system currently used to group elements. This system assigns IS to particular families based on ORF arrangement, transposase sequence similarity, inverted repeat similarity, and the generation of direct repeats of a given length. The system presently includes nineteen defined IS families of varying homogeneity, the basic characteristics of which are shown in table 1.1. There is also a large group of unclassified elements likely representing families for which an insufficient number of known members precludes definition. More than 800 IS have been officially assigned to families, with these assignments being catalogued at the IS-Finder Database (www-is.biotoul.fr/). Typically only one isoform from a given species is logged in the database. To date few of the many IS discovered in the course of genome sequencing have been included. Expansion of the current families and definition of new families is sure to increase in pace as these and other elements receive consideration.

Family	Members Identified*	Present in Bacteria ?	Present in Archaea ?	Length Range (bp)	IR	DR	ORF #	Mode of Transposition †
IS1	32	Y	Y	690 – 1170	16 – 46	8 – 12	1 – 2	Conservative or Replicative
IS3	91	Y	Y	1000 – 1610	5 – 43	3 – 5	2	Circular intermediate
IS4	50	Y	Y	1300 – 1960	7 – 22	9 – 12	1 – 3	Conservative
IS5	95	Y	Y	750 – 1640	4 – 40	2 – 9	1 – 2	Unknown
IS6	18	Y	Y	700 – 1650	14 – 37	0 – 8	1	Replicative
IS21	25	Y	?	1950 – 2630	12 – 50	4 – 8	2	Circular Intermediate?
IS30	24	Y	Y	1000 – 1260	7 – 29	0 – 3	1	Unknown
IS66	15	Y	?	2500 – 2720	9 – 24	8	3 – 5	Unknown
IS91	6	Y	?	1500 – 1850	0	0	1	Replicative
IS110	22	Y	Y	1200 – 2530	0 – 15	0 – 8	1	Circular Intermediate
IS200/ IS605	24	Y	Y	700 – 2000	0	0 – 8	1 – 3	Replicative?
IS256	41	Y	Y	1140 – 1500	14 – 42	0 – 9	1	Circular Intermediate
IS481	20	Y	?	950 – 1100	20 – 23	5 – 6	1	Unknown
IS630	22	Y	Y	950 – 1400	6 – 29	2	1 – 2	Conservative?
IS982	10	Y	Y	900 –1000	18 – 35	0	1 – 2	Unknown
IS1380	11	Y	?	1650 – 2070	6 –18	4	1	Unknown
ISAs1	9	Y	?	950 – 1550	5 – 25	8	1	Conservative?
ISL3	31	Y	Y	1190 – 2259	17 – 39	0 – 8	1 – 2	Unknown
Tn3	3	Y	?	>3000		5	1 – 2	Conservative

Table 1.1: Prokaryotic IS Families and Their Characterisitics (Adapted from Chandler & Mahillon 2002)

*Does not include IS discovered only from genome sequences

†Based on the best characterized IS in family

IS of the nineteen defined families are not limited to particular taxa. While there is variation in the apparent spread of family members among taxa, it is becoming clear that initial and current indications of provinciality in certain families is due to a lack of sufficient known examples. An example of this is the IS1 family. Originally thought to be limited to the enteric bacteria, specimens have been recently discovered in *Sulfolobus* genomes (Brugger et al. 2002). The degree of cosmopolitanism is very clear from the phylogenetic trees constructed for certain families such as IS5 and IS200/IS605. These do not show divergence of IS along the same evolutionary lines as their hosts, but rather evidence of extensive and common intertaxonomic transfers. IS that are commonly found in the same host, for instance, are often very highly diverged, while very closely related IS are routinely found in highly diverged hosts. This is true even at the domain level. All but six of the defined families include member elements discovered in Archaea. While there are certain bacterial and archaeal sub trees, there is no overall bifurcation of the trees into domain- specific IS. The pattern is instead largely the same as that seen at lower taxonomic levels. Such phylogenetic signals testify to cross-domain IS transfers, and indicate that IS-mediated horizontal gene transfer is likely to be of common occurrence (Brugger et al. 2002, Chandler & Mahillon 2002).

VI: IS in Evolution

IS have enormous potential to impact the evolution of their hosts through a variety mechanisms. Due to it being the means by which they were first discovered, the earliest noticed was the disruption of the genes into which IS transpose. Provided the gene is crucial enough, such disruption can kill the host. More subtly, the

transposition of an IS into regulatory and intergenic regions can impact the regulation of nearby genes. At least part of this effect is based on the partial promoter sequences that have been observed in the IRs of many elements. These allow the occasional generation of strong fortuitous promoters, up-regulating downstream genes. Alternately, disruption of an endogenous promoter can also lead to reduced gene expression (Mahillon 1999, Chandler and Mahillon 2002).

IS can also promote genome plasticity. The presence of multiple copies of an IS in a genome introduces large regions of high sequence similarity in areas where little or none previously existed. These islands of homology provide excellent substrates for the host's recombination machinery, and crossing over can take place between the copies more easily than would otherwise be the case at their sites of insertion. This leads to homologous intragenomic rearrangements that would not otherwise take place. This is also true of intramolecular replicative transposition, in which the resulting inversion is not always corrected. Similarly, recombination between IS copies on a single chromosome can also result in the excision of the intervening DNA, thus leading to gene deletion, while asymmetric recombination between them may lead to gene duplication. Further, if an IS that acts as a joint between regions subjected to rearrangement is deleted, this may also result in domain shuffling and new gene formation. All five situations can lead to the creation of new gene combinations that may generate new phenotypes. IS thus play a major role in the generation of variation at the molecular level upon which selective forces may act (Nevers & Saedler 1977, Haack & Roth 1995, Mahillon & Chandler 1998, Chandler & Mahillon 2002).

The autonomy of movement and unity of structure of IS also make them significant in horizontal gene transfer. IS can easily spread to other cell lines via plasmids or host chromosomal fragments containing them, something that is clearly common given phylogenetic evidence. More significant, however, is their capacity to similarly transfer segments of host DNA between cell lines via the formation of composite transposons. These structures result from the transposition of two copies of an IS into positions within a few hundred to a few thousand base pairs of each other. The transposase they encode can then interpret the inverted repeats marking the outer ends of the IS copies as defining a single element, and may consequently transpose the entire structure. This mobilizes the intervening segment of the host genome, allowing it to be moved to other parts of the genome or to resident plasmids. Such plasmids may then transfer to other, potentially highly diverged cells. In the new host, conservative transposition can integrate the transposon and its former host's DNA to the new chromosome. Alternately, replicative transposition can mediate integration of the entire plasmid (Chandler 1998). Transposon-mediated spread of antibiotic and heavy metal resistance genes are examples of this process (Campbell et al. 1994, Mahillon et al. 1999, Chandler and Mahillon 2002).

VII: IS as Molecular Markers

IS are not as readily adapted to use as tools in molecular biology as are transposons that carry genes conferring easily detected phenotypes. This is not to say that they are without their uses. Their self-contained structures, making them functional regardless of the genomic environment in which they find themselves, combined with their capacity to cross taxonomic boundaries make them highly

sensitive markers for the detection of gene flow. Further, the low frequency of transposition makes IS fairly stable in a given site of insertion. This property and the wide range of sites in a host genome open to their integration permits the pattern of IS insertion sites to be used for genomic fingerprinting. IS may thus be used to differentiate closely related strains, and elucidate the relationships between them through the tracking of gene flow and generation of high-resolution phylogenies (Mahillon & Chandler 1998, Lee et al.2001, Chandler & Mahillon 2002).

This utility has been extensively demonstrated in the study of bacterial pathogens. The first such was the use of the occurrence and intrachromosomal frequency of IS5 to study natural *Escherichia coli* isolates by Green *et al.* (1984). IS have since been used as the basis for the evolutionary and epidemiological analysis of *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, and species of *Salmonella*, *Bordetella*, *Staphylococcus*, and *Vibrio* (Arpin et al. 1996, Bik et al, 1996, Gordon et al, 1999, Kremer et al. 1999, Robinson et al. 1998, Sola et al. 2001, Stanley & Saunders 1996, van der Zee et al. 1997). Significantly, two phenotypically different *Mycobacterium* species, *ulcerans* and *marinum*, sharing 99.8% genetic identity, were demonstrated to be genetically isolated from each other on the basis of the presence of two high-copy-number IS in the former, but not the latter (Stinear et al 2000). In other studies, IS have been used to demonstrate the panmictic nature of global *Heliobacter pylori* populations (Salaun et al. 1998) and to delineate separate populations of the agricultural pathogen *Ralstonia solanacearum* (Lee et al. 2001). Further, the distribution of two IS in the genomes of strains of the periodontal pathogen

Actinobacillus actinomycetemcomitans has been demonstrated to be closely correlated with the population structure of the organism (Hayashida et al. 2000).

VIII: IS in Archaea

IS have been found to be common in Archaea, though considerable variation is observed between different groups and species in regard to the number and diversity of types. They have proven to be comparatively rare in the methanogens, as might be expected from their small genome sizes. *Methanobrevibacter smithii*, for instance, has been found to have only one IS, and *Methanococcus jannaschii* only a total of seven. Further, examination of *Methanobacterium thermoautotrophicum* has thus far shown evidence of no IS at all (Brugger et al. 2002). They have been found to be much more common amongst the thermophilic Euryarchaeota. *Pyrococcus furiosis*, for instance, was found to possess at least twenty-nine IS of only three types, with one having been found to have twenty-three copies present (Kawarabayasi et al. 1998). The complete genome sequence of *Thermoplasma volcanium*, despite its small size (1.5 Mb), showed the presence of twenty-four different IS types present in an aggregate copy number of twenty-seven. The IS200/IS605 family was particularly well represented. Finally, *Thermoplasma acidophilum* was found to host only four IS, each different (Brugger et al. 2002).

IS have been found to be even more common among the halophiles, as might be expected from their larger and more complex genomes. A large number of highly active IS are responsible for such a high rate of mutation in *Halobacterium salinarum* that cell division typically results in daughter cells with only an 80% likelihood of genetic identity (Charlebois & Doolittle 1989). The genome of *Halobacterium sp*.

NRC-1, a close relative and the first halophile to be fully sequenced, has been shown to contain eighty-two complete copies of IS belonging to both known and unknown families (Keller et al. 2000, Brugger et al. 2002). Interestingly, most of these reside in a megaplasmid that seems to be evolving into a second chromosome due the IS-mediated transfer of chromosomal genes to it (Ng et al. 1998).

IS in *Sulfolobus*

The largest number and diversity of IS yet found among the Archaea, indeed among prokaryotes, have been from species of the Crenarchaeotal genus *Sulfolobus* (See table 1.2). The largest and best studied genus of hyperthermophilic Archaea, *Sulfolobus* is made up of acidophilic obligate aerobes native to acidic, sulfur-containing hot springs, and are characterized by optimal growth at temperatures between 70 and 85 C and pH 2 to 4 (Grogan 1989). The ease with which they may be cultured has led to their popularity for the study of various aspects of hyperthermophile and Archaeal molecular biology.

The first IS recovered from *Sulfolobus* was discovered by Schleper et al. (1994) during experiments aimed at developing a transformation system for *Sulfolobus*

IS Type	IS Family	Species of Original Discovery	IS Length (bp)	IR Length (bp)	DR Length (bp)	ORFs
ISC774	IS6	*S. solfataricus*	782	19	0	2
ISSt796	IS1	*S. tokodaii*	796	21	8	1
ISSt847	IS6	*S. tokodaii*	847	32	?	1
ISC1041	IS30	*S. solfataricus*	1038	16	0	1
ISC1043	ISCL3	*S. solfataricus*	1043	14	2	2
ISC1048	ISC630/ Tc1	*S. solfataricus*	1048	23	2	1
ISC1058	IS5	*S. solfataricus*	1058	19	9	1
ISC1078	ISC630/ Tc1	*S. solfataricus*	1078	19	2	1
ISSt1145	NYC	*S. tokodaii*	1145	8	7	1
ISC1160	IS4	*S. solfataricus*	1160	12	6	3
ISC1173	IS1	*S. solfataricus*	1173	46	8	1
ISSt1173	IS1	*S. tokodaii*	1173	50	8	1
ISC1190	IS110	*S. solfataricus*	1190	0	8	2
ISC1212	IS5	*S. solfataricus*	1212	27	0 – 7	2
ISC1217	NYC	*S. solfataricus*	1148	13	6	1
ISC1225	IS4	*S. solfataricus*	1225	17	4 – 5	1
ISC1229	IS110	*S. solfataricus*	1229	0	0	1
ISC1234	IS5	*S. solfataricus*	1234	19	4	1
ISSt1234	IS110	*S. tokodaii*	1234	0	6	1

Figure 1.2: IS Recovered from *Sulfolobus* Species

IS Type	IS Family	Species of Original Discovery	IS Length (bp)	IR Length (bp)	DR Length (bp)	ORFs
ISC1250	IS256	*S. solfataricus*	1250	9	0	1
ISC1290	IS5	*S. solfataricus*	1290	40	0 – 9	1
ISC1316	IS200/ IS605	*S. solfataricus*	1316	0	0	1
ISSt1318	IS5	*S. tokodaii*	1318	26/27	5	1
ISC1332	IS256	*S. solfataricus*	1332	22	9	1
ISSt1350	IS4	*S. tokodaii*	1350	22	2	1
ISC1359	IS4	*S. solfataricus*	1359	25	4	1
ISC1395	ISC630/ Tc1	*S. solfataricus*	1395	60	0	2
ISC1439	IS4	*S. solfataricus*	1439	20	9	1
ISC1476	IS200/ IS605	*S. solfataricus*	1476	20	0	1
ISSt1484	IS110	*S. tokodaii*	1484	0	0	1
ISC1491	IS110	*S. solfataricus*	1491	0	0	2
ISSt1692	IS200/ 605	*S. tokodaii*	1692	0	0	2
ISSt1797	IS200/ IS605	*S. tokodaii*	1797	4	0	3
ISC1904	IS200/ IS605	*S. solfataricus*	1904	0	0	2
ISC1913	IS200/ IS605	*S. solfataricus*	1913	0	0	2
ISSt1922	IS200/ IS605	*S. tokodaii*	1922	0	0	2

Figure 1.2, continued: IS Recovered from *Sulfolobus* Species

solfataricus strain P1. They isolated three mutants with a stable Lac^- phenotype, PCR amplified, cloned, and sequenced the β-galactosidase gene of one of them. It was found to harbor a copy of an element now designated as ISC1217, a member of an undefined IS family. Subsequent analysis showed the remaining two mutants to have β-galactosidase genes similarly disrupted by the same IS. Southern analysis of the genome revealed it to harbor a total of eight copies of the element.

This was followed in 2000 by another unexpected discovery of IS by spontaneous mutation. To study the rate of mutation in various *Sulfolobus* species, Martusewitsche et al. (2000) isolated mutants defective in either of two pyrimidine biosynthesis pathway enzymes: orotate phosphoribosyl transferase or orotidine 5'monophosphate decarboxylase, which are encoded by the coordinately controlled *pyrE* and *pyrF* loci, respectively (Grogan & Gunsalus 1993, Jacobs & Grogan 1997). The mutation rate of *S. solfataricus* strains P1 and PH1 proved to be one to two orders of magnitude higher than in other species. The *pyrE* and *pyrF* loci of seven of the *S. solfataricus* mutants isolated were then screened by PCR, revealing that the dysfunction in all seven to be due to the transposition of IS into one of the two loci. Sequencing revealed four different IS to be responsible. One was identical to the previously characterized ISC1217, while three, ISC1359, ISC1058, and ISC1439 proved to be previously uncharacterized elements belonging to known IS families (ISC1359 and ISC1439 to IS4, and ISC1058 to IS5). These findings pointed to a high prevalence of active IS in *S. solfataricus* later confirmed by the genome sequence, and led Martusewitsch et al. to conclude that IS are responsible for the bulk of mutation events in this species.

While spontaneous mutation events have led to the discovery of many IS in *Sulfolobus*, most have been found in the course of sequencing efforts. The first such was the discovery of ISC1041 in *S. solfataricus* strain MT4 by Ammendola et al. (1998) when it happened to transpose into the glutamate dehydrogenase gene that the researchers had been attempting to clone. This was followed by the discovery of ISC1316 and ISC1332 in the sequence of the plasmid pNOB8 isolated from *Sulfolobus* strain NOB8-H2 (She et al. 1998), and that of ISC1913 during the sequencing of the pING family of plasmids from *S. islandicus* (She et al. 2000).

The greatest number of *Sulfolobus* IS thus found, however, have come from the complete genome sequences of *S. solfataricus* P2, and *S. tokodaii* (Kawarabayasi et al. 2001, She et al. 2001). Approximately ten percent of the *S. solfataricus* genome was found to be composed of an aggregate total of 201 complete copies of twenty-four distinct IS types, including examples of every *Sulfolobus* IS documented except ISC1041. *S. solfataricus* thus has the distinction of containing the largest number of IS found in any prokaryotic genome yet sequenced (Figure 1.1). A smaller number of IS, thirty-four total copies of twelve distinct types, were identified in the genome of *S. tokadaii*. Among the IS types discovered in the two species, eleven in each had close relatives in the other. A high incidence of IS is not apparently uniform across all *Sulfolobus* species. The partially completed genome sequence of *S. acidocaldarius*, a diverged member of the genus, has shown the presence of only a few complete IS copies (Ming et al. personal communication) and none have been found from spontaneous mutation events (Grogan et al. 2001).

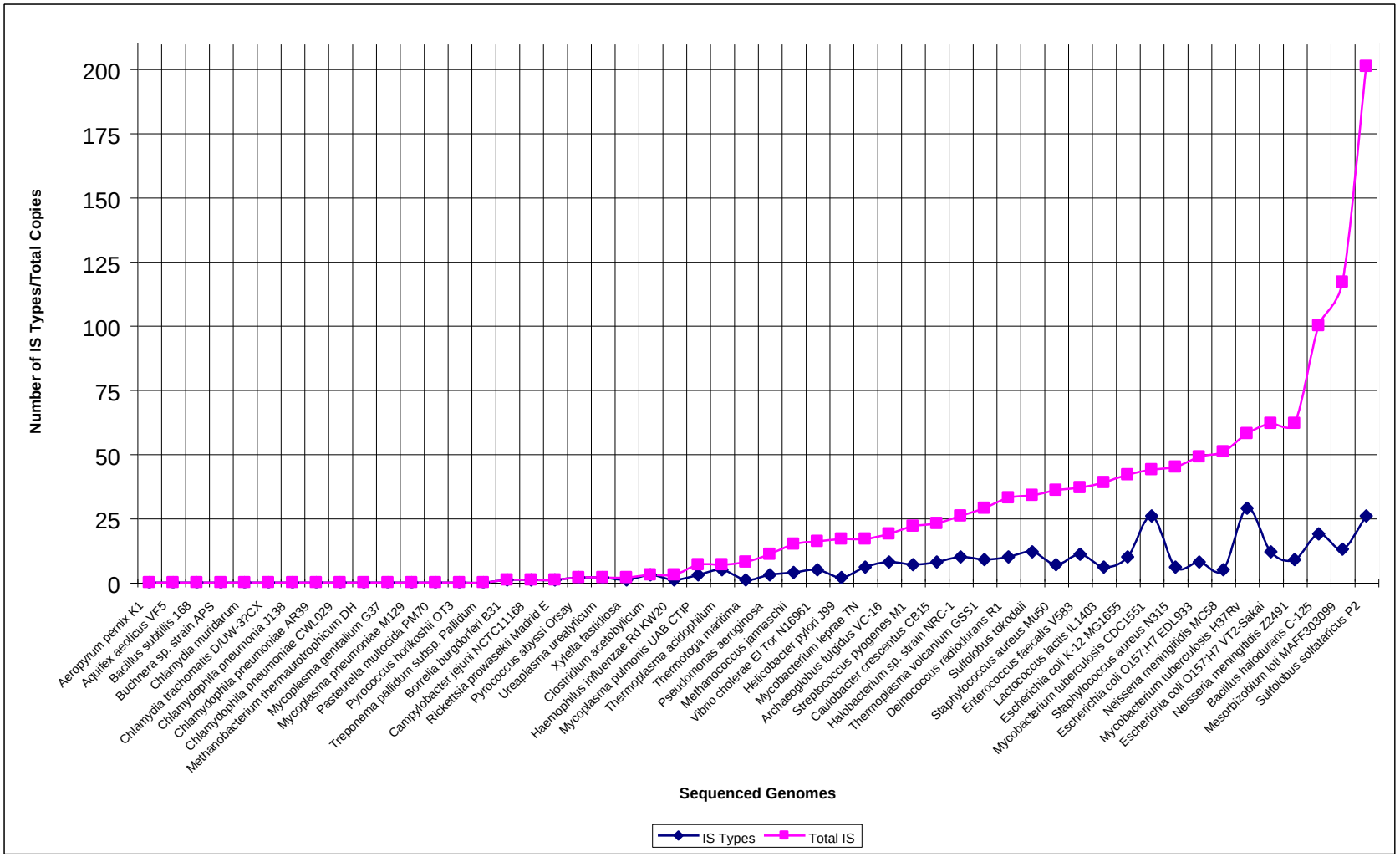


Figure 1.1: Number of IS and IS Types Discovered in Sequenced Prokaryote Genomes

The information concerning *Sulfolobus* IS that comes from the *S. solfataricus* and *S. tokodaii* is interesting and can not be discounted, but its leaves a number of important issues to be investigated. First, the genome sequences only give information about IS from two locations, namely Kyshu Island, Japan, and Naples, Italy. *Sulfolobus* populations are to be found in numerous locations worldwide. The types of IS to be found residing within these populations are not known. Second, its dominance of the populations of many sites in Iceland, North America, and Asia indicates *S. islandicus* to be the most cosmopolitan of known *Sulfolobus* species (Whitaker et al. 2001). As such, it stands to be of more interest in biogeographic studies, but little is known of the IS of this species. Finally, the large number of IS types in the two completely sequenced species that show close relation to elements in the other species is remarkable given the distances between the sites of their isolation. A specific question this leaves unanswered is that of whether or not similar close relationships are to be seen between IS found in other populations, indicating that a large number of well-separated population share a pool of closely related IS.

The project described in this thesis was intended to address these issues. It may be summarized by its four basic goals. The first was to use a gene trapping technique based upon the simple and reliable genetic selection modified from that used by Martusewitsch et al. (2000) to recover active IS from *Sulfolobus* strains derived from geographically separated hot spring populations. The second was to sequence and characterize the IS recovered. The third was to examine the geographic distributions of the IS recovered. The final goal was to provide a foundation for future studies of *Sulfolobus* genome dynamics, molecular genetics, population structure, and

biogeography by identifying active IS that might be useful as tools and genetic markers.

IX: References

1. Ammendola, S., Politi, L., & Scandurra, R. (1998). Cloning and sequencing of ISC1041 from the archaeon *Sulfolobus solfataricus* MT-4, a new member of the IS30 family of insertion elements. FEBS Letters, 428, 217 – 223.
2. Arpin, C., Lagrange, I., Gachie, J., Bebear, C. & Quentin, C. (1996). Epidemiological study of an outbreak of infection with *Staphylococcus aureus* resistant to lincosamides and streptogramin A in a French hospital. Journal of Medical Microbiology, 44, 303 – 310.
3. Bik, E, Gouw, R. & Mooi, F (1996). DNA fingerprinting of *Vibrio cholerae* strains with a novel insertion sequence element: a tool to identify epidemic strains. Journal of Clinical Microbiology, 34, 1453 – 1460.
4. Brugger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y., & Garrett, R. (2002). Mobile genetic elements in archaeal genomes. FEMS Microbiology Letters, 206, 131 – 141.
5. Campbell, A. (2002). Eubacterial Genomes, In Craig, N., Craigie, R., Gellert, M., & Lamhowitz., A. (Ed.). *Mobile DNA II* (pp. 1024 – 1039). Washington, D.C.: American Society for Microbiology Press.
6. Campbell, W., Cashore, A., Nakatsu, C., & Peel, M. (1994). Catabolic transposons. Biodegradation, 5, 323 – 342.

7. Chandler, M. (1998). Insertion Sequences and Transposons, In de Bruijin, F., Lupinski, J., & Weinstock, G. (Ed.). *Bacterial Genomics: Physical Structure and Analysis* (pp. 30 – 37). New York: Chapman & Hall.
8. Chandler, M. & Mahillon, J. (2002). Insertion Sequences Revisited, In Craig, N., Craigie, R., Gellert, M., & Lamhowitz, A. (Ed.). *Mobile DNA II* (pp. 305 – 366). Washington, D.C.: American Society for Microbiology Press.
9. Charlebois, R. & Doolittle, F. (1989). Transposable elements and genome structure in Halobacteria, In Berg, D. & Howe, M. (Ed.). *Mobile DNA* (pp. 297 – 307). Washington, D.C.: American Society for Microbiology Press.
10. Craig, N., (1997). Target site selection in transposition. Annual Review of Biochemistry, 66, 437 - 474.
11. Craig, N. (2002). Mobile DNA: An Introduction, In Craig, N., Craigie, R., Gellert, M., & Lamhowitz, A. (Ed.). *Mobile DNA II* (pp. 3 – 11). Washington, D.C.: ASM Press.
12. Craigie, R. (1996). Quality control in Mu DNA transposition. Cell, 85, 137 – 140.
13. Escoubas, J., Prere, M., Fayet, A., Salvignol, I., Gulas, O., Zerbib, O., Chandler, M. (1991. Translational control of transposition activity of the bacterial insertion sequence IS1. EMBO Journal, 10, 705 – 712.
14. Grindley, N. (2002). The movement of TN3-like elements: Tansposition and cointegrate resolution, In Craig, N., Craigie, R., Gellert, M., & Lamhowitz, A. (Ed.). *Mobile DNA II (*pp. 272 – 302). Washington, D.C.: ASM Press

15. Godon, S., Heym, B., Parkhill, J., Barrell, B. & Cole, S. (1999). New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. Microbiology, 145, 881 – 892.
16. Green, L., Miller, R., Dykhuizen, D., Hartl, D. (1984). Distribution of DNA insertion element IS5 in natural isolates of *E. coli*. Proceedings of the National Academy of Sciences (USA), 81, 881 – 892.
17. Grogan, D. (1989). Phenotypic characterization of the Archaebacterial genus *Sulfolobus*: Comparison of five wild-type strains. Journal of Bacteriology, 171, 6710 – 6719.
18. Grogan, D. & Gunsalus, R. (1993). *Sulfolobus acidocaldarius* synthesizes UMP via a standard de novo pathway: Results of a biochemical-Genetic Study. Jouirnal of Bacteriology, 175, 1500 – 1507.
19. Grogan, D., Carver, G., & Drake, J. (2001). Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon *Sulfolobus acidocaldarius*. Proceedings of the National Academy of Sciences (USA), 98, 7928 – 7933.
20. Haack, K. & Roth, J. (1995). Recombination between chromosomal IS200 elements supports frequent duplication formation in *Salmonella typhimurium*. Genetics, 141, 1245 – 1252.
21. Haren, L., Ton – Hoang, B., & Chandler, M. (1999). Integrating DNA: transposases and retroviral integrases. Annual Review of Microbiology, 53, 245 – 281.

22. Hayashida, H., Poulson, K., Takagi, O., & Kilian, M. (2000). Phylogenetic association of ISAa1 and IS150 – like insertion sequences in *Actinobacillus actinomycetemcomitans*. Microbiology, 146, 1977 – 1985.

23. Jacobs, K. & Grogan, D. (1997). Rates of spontaneous mutation in an archaeon from geothermal environments. Journal of Bacteriology, 179, 3298 – 3303.

24. Jenkins, R. & Reznikoff, W. (1997). Critical contacts between HIV – integrase and viral DNA identified by structure – based analysis and photo-crosslinking. EMBO Journal, 16, 6849 – 6859.

25. Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M. Ohfuku, Y., Funashashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., & Kikuchi, H. (1998). Complete sequence and gene organization of the genome of a hyperthermohilic archaebacterium, *Pyrococcus hoikoshii* OT3. DNA Research, 5, 55 – 76.

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

thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* Strain 7. DNA Research, 8, 123 – 140.

27. Kremer, K., van Soolingen, D., Frothingham, R., Haas, W., Hermans, P., Martin, C., Palittapongarnpim, P., Plikaytis, B., Riley, L., Yakrus, M., Muser, J., & van Embden, J. (1999). Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: Interlaboratory study of discriminatory power and reproducibility. Clinical Microbiology, 37, 3943 – 3950.

28. Kulkosky, J., Jones, K., Katz, R., Meck, J., & Skalka, A. (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retrobial/retrotransposon integrase and bacterial insertion sequence transposases. Molecular and Cellular Biology. 12, 2331 – 2338.

29. Lee, Y., Far, S., Chiu, L., & Hsia, K. (2001). Isolation of an insertion sequence from *Balslonia solanscearum* race I and the potential use for strain characterization and detection. Applied and Environmental Microbiology, 67, 3943 – 3950.

30. Mahillon, J. & Chandler, M. (1998). Insertion Sequences. Microbiology and Molecular Biology Reviews, 62, 725 – 774.

31. Mahillon, J., Leonard, C., & Chandler, M. (1999). IS Elements as Constituents of Bacterial Genomes. Research in Microbiology, 150, 675 – 687.

32. Martusewitsche, E., Sensen, C., & Schleper, C. (2000). High spontaneous mutation rate in the hyperthemophilic Archaeon *Sulfolobus solfataricus* is mediated by transposable elements. Journal of Bacteriology, 182, 2574 – 2581.
33. Ming, L. Brügger, K., & Garrett, R. personal communication.
34. Nevers, P/ & Sadler, H. (1977). Tansposable Genetic Elements as Agents of Gene Instability and Chromosomal Rearrangements. Nature, 268, 109 – 115.
35. Noll, K & Vargas, M. (1997). Recent advances in genetic analyses of hyperthermophilic Archaea and Bacteria. Archives of Microbiology, 168, 73 – 80.
36. Ohta, S., Tsuchida, K., Choi, S., Sekine, Y., Shiga, Y., Ohtsubo, E. (2002). Presence of a characteristic D-D-E motif in IS1 transposase. Journal of Bacteriology, 184, 6146 – 6154.
37. Robinson, A., Hollingshead, S., Musser, J., Parkinson, A., Briles, D., & Picard, R. (1998). The IS1167 insertion sequence is a phylogenetically informative marker among isolates of serotype 6B *Streptococcus pneumoniae*. Journal of Molecular Evolution, 47, 222 – 229.
38. Salaun, L., Audibert, C., Le Lay, G., Burucoa, C., Fouchere, J., & Picard, R. (1998). Panmictic structure of *Heliobacter pylori* demonstrated by the comparative study of six genetic markers. FEMS Microbiology Letters, 161, 231 – 139.
39. Schleper, C., Roder, R., Singer, T., & Zillig, W. (1994). An insertion element of the extremely thermophilic archaeon *Sulfolobus solfataricus* transposes into

the endogenous β-galactosidase gene. Molecular and General Genetics, 243, 91 – 96.

40. Shapiro, J. (1979). Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. Proceedings of the National Academy (USA), 76, 1933 – 1937.
41. She, Q., Phan, H. Garrett, R., Albers, S., Stedman, K, & Zillig, W. (1998). Genetic profile of pNOB8 from *Sulfolobus*: the first conjugative plasmid from an archaeon. Extremophiles 2, 417 – 425.
42. She, Q., Singh, R., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M., Chan-Weiher, C., Clausen, I., Curtis, B., Moors, A., Erauso, G., Fletcher, C., Gordon, P., Heikamp-de Jong, I., Jeffries, A., Kozera, K., Medina, N., Peng, X., Thi-Ngoc, H., Redder, P., Schenk, M., Theriault, C., Tolstrup, N., Charlebois, R., Doolittle, W., Duguet, M., Gaasterland, T., Garrett, R., Ragan, M., Sensen, C., & Van der Oost, J. (2001). Complete genome of the crenarchaeote *Sulfolobus solfataricus* P2. Proceedings of the National Academy of Sciences (USA), 98, 7835-7840.
43. Sola, C., Filliol, I., Legrand, E., Mokrousov, I., Rostogi, N. (2001). Mycobacterium tuberculosis phylogeny reconstructed based on combined numerical analysis with IS1081, IS6110, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. Journal of Molecular Evolution, 53, 680 – 689.

44. Stanley, J. & Saunders, N. (1996). DNA insertion sequences and the molecular epidemiology of *Salmonella* and *Mycobacterium*. Journal of Medical Microbiology, 45, 236 – 251.

45. Stedman, K., She, Q., Phan, H., Holz, I., Singh, H., Prangishvili, D., Garrett, R, & Zillig, W. (2000). The pING family of conjugative plasmids from the extremely thermophilic archaeon *Sulfolobus islandicus*: Insights into recombination and conjugation in Crenarchaeota. Journal of Bacteriology, 182, 7014 – 7020.

46. Stinear, T., Jenkin, G, Johnson, P., & Davies, J. (2000). Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinus* reveals evidence of recent divergence. Journal of Bacteriology, 182, 6322 – 6330.

47. Suzuki, T., Iwasaki, T., Uzawa, T., Hara, K, Nemoto, N, Kon, T., Toshiaki, U., Yamagishi, A., & Oshima, T. (2002). *Sulfolobus tokodaii* sp. nov. (f. *Sulfolobus* sp. strain 7), a new member of the genus *Sulfolobus* isolated from Beppu Hot Springs, Japan. Extremophiles, 6, 39 – 44.

48. Van der Zee, A., Mooi, R., van Embden, J., & Musser, J. (1997). Molecular evolution and host adaptation of Bordetella spp.: Phylogenetic analysis using multilocus enzyme electrophoresis and typing with three insertion sequences. Journal of Bacteriology, 179, 6609 – 6617.

49. Ng. W., Ciufo, S., Smith, T., Bumgarner, R., Baskin, D., Faust, J., Hall, B., Loretz, C., Seto, J., Slagel, J., Hood, L., & DasSarma, S. (1998). Snapshot of a large dynamic replicon in a halophilic archaeon: Megaplasmid or minichromosome? Genome Research, 11, 1131 – 1141.

50. Whitaker, R., Grogan, D., Taylor, J. (2001). The origin of a species? Biogeographical patterns of nucleotide variation in populations of *Sulfolobus*. Gordon Research Conference on Archaea: Ecology, Metabolism, and Molecular Biology, Andover, NH.

Chapter 2
Recovery of Active Insertion Sequences from *Sulfolobus* Strains

I: Introduction

Methods of IS Recovery

The advent of rapid DNA sequencing methods and consequent sequencing of the complete genomes of an increasingly long list of organisms has led to the discovery of a profusion of previously unknown insertion sequences. Complete genome sequencing has the benefit of uncovering all the IS present in a particular host, but it suffers from two major drawbacks as a tool with which to find and study IS. Most obvious is the enormous amount of time and money that must be invested in such an endeavor. Perhaps more important, however, is that genomic sequences yield only the sequence of the IS and their genomic locations. This information can be of limited value. One cannot necessarily determine from sequence alone whether or not a particular copy of an IS is active, much less the character of its activity, and intrachromosomal rearrangement can relocate an element exclusive of its endogenous transposase activity. The result is that genome sequences are deficient in operational information on the IS they uncover. These drawbacks still afford a role for older methods of recovering active IS.

Recovery of active IS from an organism of interest faces a number of difficulties arising from the nature of the elements. They are mobile, meaning that they are not necessarily localized in any particular genomic location. More importantly, unlike transposons that carry genes for antibiotic resistance and other

accessory functions, IS possess only those structural and genetic features required for transposition (Chandler & Mahillon 2002). This makes it impossible to directly select for them. IS recovery methods have thus relied on two characteristics typical of these elements that permit indirect selection and identification: the enlargement of the segments of DNA into which they transpose, and the effects, most notably inactivation, they can exert upon those genes that contain or lie close to their sites of insertion.

The most primitive method of IS recovery is based entirely upon the former characteristic. This involves the observation of the gross enlargement of a given restriction fragment during RFLP analysis. The low frequency of IS transposition makes this process tedious, not to mention prone to problems introduced by recombination events and gene duplications. These problems limit the utility of this method, and consequently few studies focused solely upon IS recovery have used it. It has, however, led to a number of incidental discoveries of IS, most notably in plants such as *Arabidopsis* and *Nicotiana* (Shepherd et al. 1982, Harberd et al. 1987, Schwarz – Sommer et al. 1987, Voytas & Ausubel 1988).

Far more useful has been the coupling of a search for IS-induced enlargement to a selection for a change in activity of a target gene or genes that can be caused by IS insertion. This is commonly called a gene or transposon trap. The selection enriches for mutants that may have an IS inserted into a defined and easily screened region of DNA. This region is then either amplified or isolated from a number of mutants and screened for enlargement indicative of an IS-mediated origin of the selected mutation (Gay et al. 1985). As most of the instances in which enlargement will not be observed

are screened out, this eliminates much of the tedium associated with the earlier RFLP method.

Types of Gene-Traps

Gene traps have proven to be not only an effective means of IS recovery, but also very flexible and easily adapted to different organisms. A wide variety have been developed so far, and may be divided into plasmid-borne forward selection, endogenous forward selection, plasmid-borne reverse selection, and endogenous reverse selection traps, depending on the direction of selection and location of the target reporter gene used.

The earliest traps used were plasmid-borne to take advantage of the development of the alkaline lysis technique of Birnboim and Doly (1979) that had made rapid plasmid isolation possible and reliable, thus permitting ease of recovery of the target for size screening. These were also generally based upon the selectable disruption of a target gene, making them forward in their direction of selection. Following selection of mutants, their resident plasmids were extracted, restriction digested, electrophoresed, and the band corresponding to the target gene examined for enlargement.

This was the pattern set by Gay et al., who reported the development of the first gene trap in 1985. They had constructed a plasmid vector bearing a *Bacillus subtilis sacB* gene, a kanamycin resistance gene, and a broad host range origin that permitted it to replicate in most Gram-negative bacteria. S*acB* gene encodes the enzyme levansucrase that converts sucrose to the fructose polymer levan, a substance that is toxic to Gram negatives. Gram-negative bacteria transformed by this plasmid

thus failed to grow on medium supplemented with sucrose and kanamycin unless *sacB* had been disrupted. These plasmids carried by the mutants were then examined for enlargement of *sacB*. *SacB* has since been used in a series of different vectors for the isolation of IS from a number of different organisms (Feng et al. 1997, Lee et al. 2001).

While *sacB* has proven popular, other genes have also been used as the bases of effective traps. *LacZ* and its blue/white selection system, for instance, has been successfully modified for this task from its use in ligation vectors (Cirillo et al. 1991, Waskar et al. 2000). Other systems have been based on an antibiotic resistance reporter gene under the control of a λcI857 repressor gene that acts as the target (Solenberg & Burgett 1989, Bartosik et al. 2003). These thus permit the ease of antibiotic selection without the need for replica plating.

Endogenous traps are based on a selectable target in the genome of the organism of interest. The genomic location removes the need for transformation and maintenance of a second selection pressure to prevent curing. However, for many years, the only way of detecting the enlargement of genomic target regions was to screen whole-genome digests by Southern blotting with target-specific probes. This made endogenous traps more difficult and time consuming to screen than plasmid-borne traps, and provides one reason for its less frequent usage. Indeed, for several years they were used primarily by researchers studying organisms not amiable to plasmid transformation. The best example of this is the trap designed around the nitrate reductase gene, the dysfunction of which yields resistance to the herbicide chlorate, used to discover a number of elements in *Nicotiana tabacum*, *Nicotiana*

plumbaginigolia, and *Arabidopsis thaliana* (Grandbastien et al. 1989, Tsay et al. 1993, Meyer et al. 1994). Since the advent of PCR, however, the screening of specific genomic segments has become both faster and easier, leading to a more common usage of endogenous genes such as *lacZ*. Despite this, plasmid-borne traps have continued in use because they can be used to screen organisms about which too little is known to permit the use of endogenous traps, as well as because they may be propagated in slow-growing organisms and then shuttled into fast-growing organisms for more rapid recovery (Solenberg et al. 1989, Waskar et al. 2000).

An alternative to forward selection traps, reverse selection traps operate on the principle that IS can potentially greatly increase the expression of gene with a weak promoter downstream of its insertion site due either to the creation of a fortuitous promoter, or by the element's possession of outwardly-directed promoters (Prentki et al. 1986, Saedler et al. 1974, Ciampie et al., 1982.). They require a target region for IS insertion, and a downstream, selectable reporter gene that has been rendered silent by upstream promoter disruption. They possess an advantage over forward selection traps in that it is very unlikely that a simple point or frameshift mutation will trigger expression of the reporter, thus greatly reducing the background of non-IS-induced mutation. Plasmid-borne forward selection traps have proven useful and efficient, though they are not common at this point, likely due in part to the greater complexity of the vectors they require (Szeverenyi et al. 1996). On the other hand, while an endogenous reverse trap is conceivable, the manipulation that would be required of the target and reporter of the organism under study would limit its application to only well

characterized organisms, and I have been unable to uncover any report of one being used.

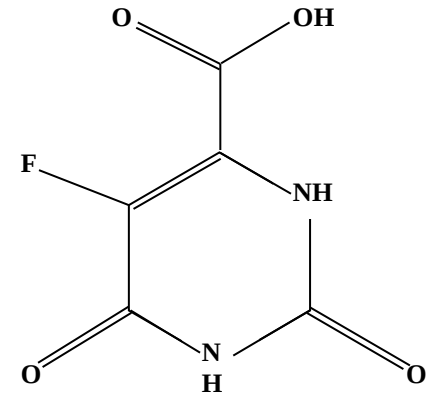


Figure 2.1: 5-Fluoroorotic Acid (5-FOA)

IS Trapping in *Sulfolobus*

The recovery of IS from *Sulfolobus* species is greatly facilitated by a reliable and simple endogenous gene trap based on the selection of pyrimidine auxotrophic mutants with 5-fluoroorotic acid (5-FOA) (Figure 2.1). This selection has long been used to study the various aspects of *Sulfolobus* mutation and DNA repair. Its use as an IS trap in *Sulfolobus* species was first reported by Martusewitsch et al. (2000), and is based on the characteristics of the well described *Sulfolobus* pyrimidine biosynthesis pathway.

The pyrimidine biosynthesis pathway of *Sulfolobus* and its relation to the *pyr* operon that controls it is depicted in Figure 2.2 (Grogan & Gunsalus 1993). Selective conditions are created when one plates *Sulfolobus* cells on a medium supplemented with 5-FOA and uracil. In wild type cells, the 5-FOA is incorporated into the pathway in place of orotate. The result of this is the production of non-functional pyrimidines that impair the function of the nucleic acids into which they may be incorporated, leading to cell lethality. The enzymes that catalyze the biochemical steps that succeed the level of 5-FOA incorporation, orotate phosphoribosyl transferase and orotidine monophosphate decarboxylase, are encoded by the *pyrE* and *pyrF* genes, respectively. In mutant cells in which either of these two genes, or their common promoter are

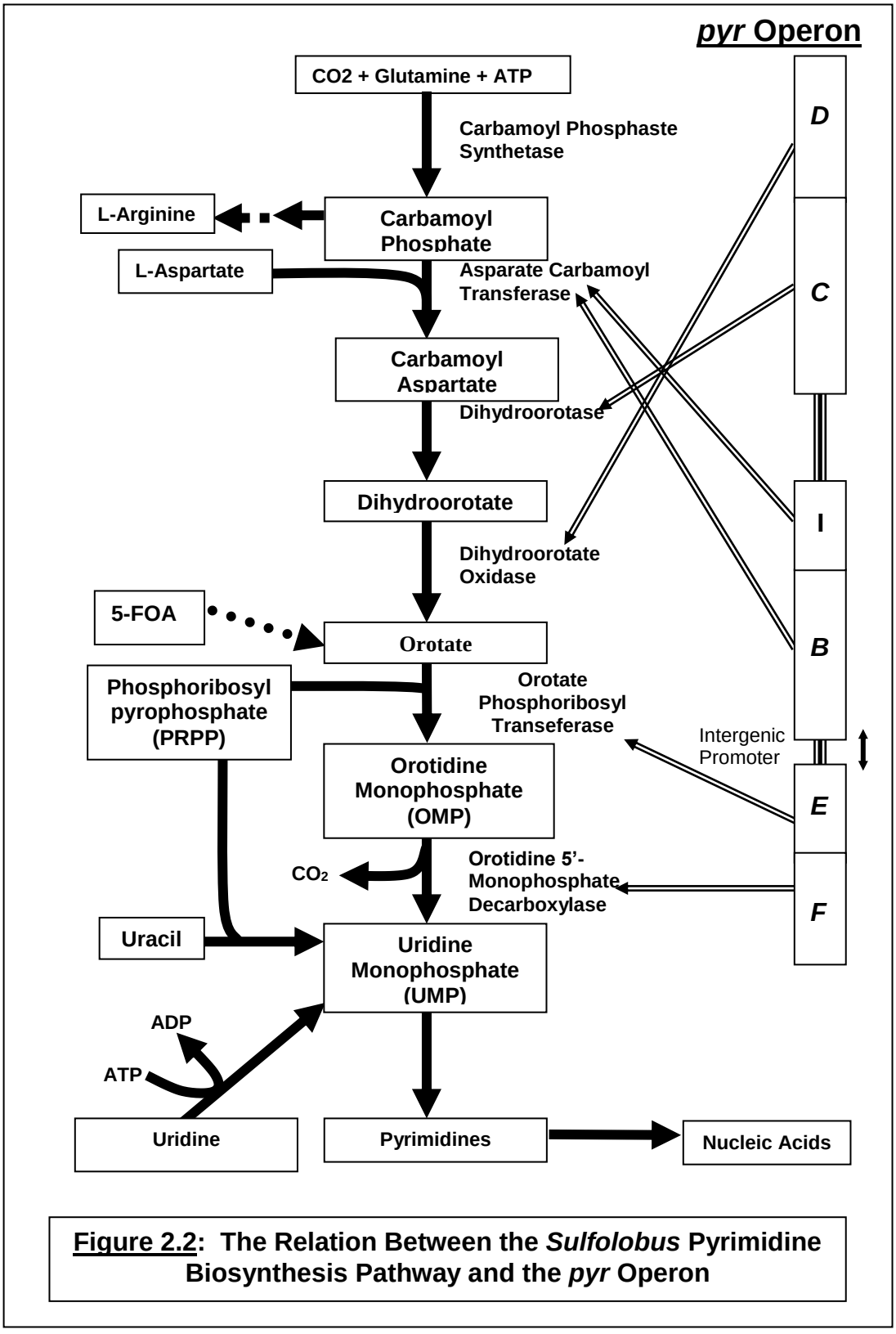


Figure 2.2: The Relation Between the *Sulfolobus* Pyrimidine Biosynthesis Pathway and the *pyr* Operon

disrupted, 5-FOA incorporation is blocked. This prevents the production of lethally dysfunctional pyrimidines, and, as 5-FOA is not toxic in its unincorporated state, the mutant cells remain viable. The uracil in the medium is then converted to uridine monophosphate via a salvage pathway, thus permitting continued pyrimidine synthesis and cell growth. Therefore, only cells with mutations in this region will initiate colony formation on the supplemented medium. PCR may then be used to screen the mutants for enlargement of the target region.

Project Overview

Since 1999, the Grogan lab has built a collection of about 1700 *Sulfolobus* strains derived from populations residing in the hot springs of a number of widely separated geographic regions. This collection, summarized in table 2.1, is one of the largest yet gathered of *Sulfolobus*. It is also potentially one of the most diverse, due not only to the large number of strains, but also to the direct plating methodology followed in the isolation of them.

This is in contrast to most such collections that represent strains isolated from enrichment cultures that would presumably have favored the retention of those strains in the sample that were either fastest growing or numerically superior. The bulk of these strains represent the unrecognized and cosmopolitan species *S. islandicus* (Whitaker 2001), while those from New Zealand represent at least two groups of unknown species related to *S. acidocaldarius* (Blount, unpublished results) and members of the genus *Sulphurisphaera* (Bell 2001), and those of Naples Italy were found to be *S. solfataricus* (Whitaker, personal communication).

Region	Year Sampled	Set Designation	Strains Collected	Dominant Species in Set
Yellowstone National Park, Wyoming, USA	1999	YNP99	351	*S. islandicus*
	2000	YNP00	850	
	2001	YNP01	46	
New Zealand	2000	NZ00	160	Unknown relative of *S. acidocaldarius*
Kamchatka Peninsula, Russia	2000	K2K	206	*S. islandicus*
Lassen National Park, California, USA	2000	L2K	40	*S. islandicus*
Naples, Italy	2002	It02	62	*S. solfataricus*

Table 2.1: Summary of the Grogan Lab *Sulfolobus* Natural Isolate Collection

The initial object of this project was to use the *pyrE/F* gene trap of Martusewitsch et al. (2000) to recover IS from a number of the strains in this collection for sequencing and other analyses. Recovery entailed the screening of chosen strains using three protocols of varying levels of screening thoroughness. In the first and most cursory (Protocol I), a frozen cell suspension of the strain of interest was used to inoculate a 3 mL broth culture of permissive DT medium. A sample of the mature culture was then plated on selective medium supplemented with 5-FOA and uracil. Five resulting mutant colonies were chosen and used to inoculate selective broth cultures from which genomic DNA was extracted and PCR probed for target site enlargement. Protocol I was used only for the first series of exploratory screenings. Protocol II was designed to eliminate the likelihood of screening sibling mutants and thus increase likelihood of recovering IS. In it, a strain interest was revived from frozen storage by streaking for isolation on a DT medium plate. Five of the wild type colonies that grew on this plate were chosen and used to inoculate five independent

DT broth cultures. Samples of these were then spread on selective DT.ura FOA medium plates. A single mutant colony was then chosen from each selective plate for target site screening. Protocol III was a more thorough variant of Protocol II in which five mutants were screened from each selective plate. These protocols are depicted in Figure 2.3.

Strains were chosen for gene trapping for a number of reasons. Initially, strains were chosen for their membership in one or both of two sets previously singled out from those in the YNP99, YNP00, NZ00, K2K, and L2K collections by examination of the EcoRI restriction profile of their genomes. The first set was made up on forty-five strains determined on the basis of these profiles to represent the greatest diversity amongst those examined (Bell, 2001). The second was made up of forty-four strains observed to have high intensity bands in their restriction profiles, possibly indicating the presence of plasmids. Later strains were chosen on the basis of cursory examination of genomic restriction profiles to increase examination of strains representing within-set diversity. An attempt was also made to screen all isolates belonging to the small collections of L2K, YNP01, and IT02. In a final series of thorough screenings using protocol three designed to recover previously undiscovered IS, strains were selected on the basis of the findings from a series of PCR probings using primers specific for IS recovered from earlier screens. A final total of 270 strains were screened by one or more of the three protocols.

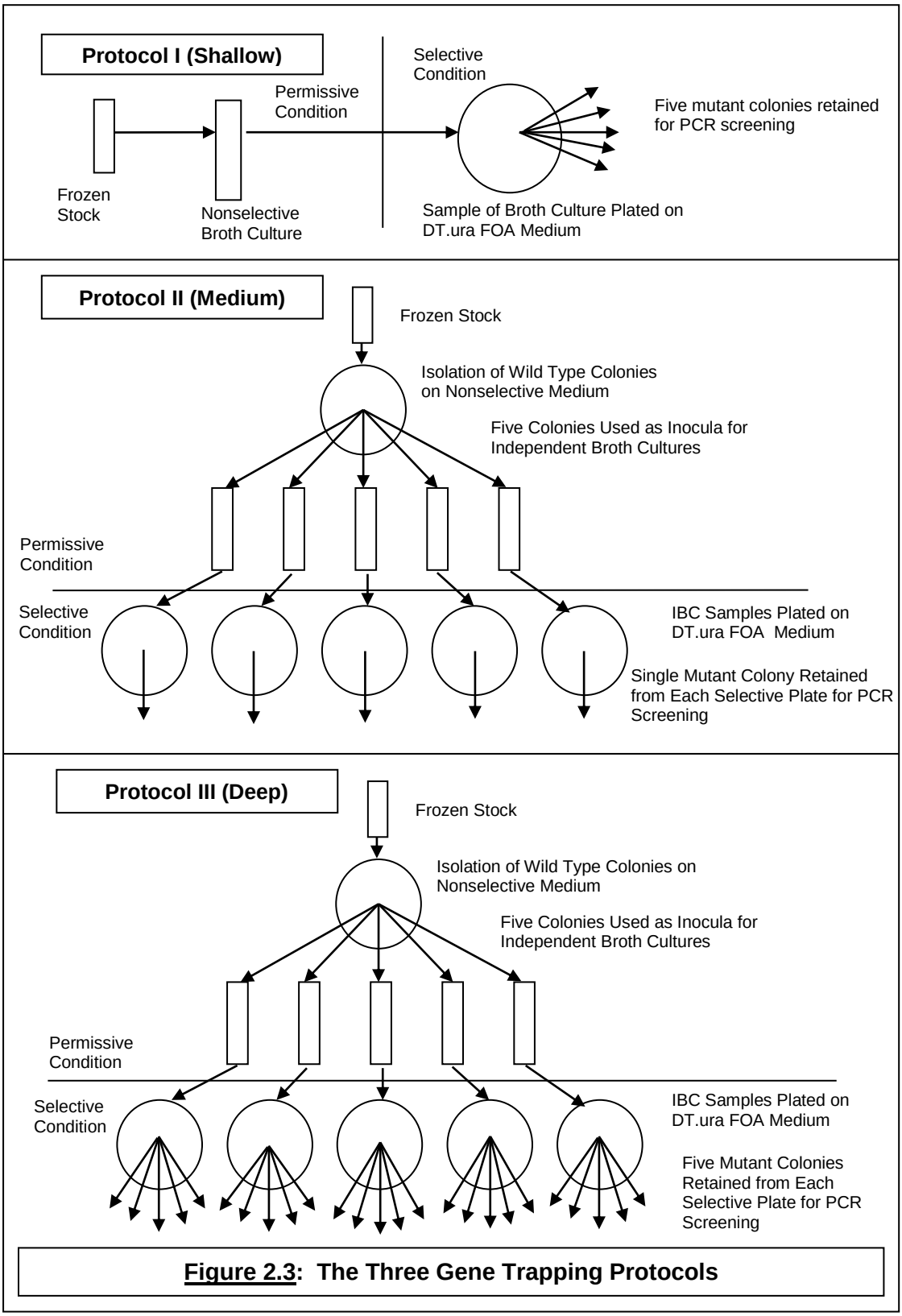


Figure 2.3: The Three Gene Trapping Protocols

II. Materials and Methods

Media

The non-selective growth medium DT contained Dextrin10 at a 0.2% concentration as a carbon source, and Bacto Tryptone at a 0.1% concentration as a nitrogen source. The medium also contained the following components per liter: 3 g K_2SO_4, 0.5 g NaH_2PO_4, 0.3 g $MgSO_4$-$7H_2O$, 0.1 g $CaCl_2$-$2H_2O$, 350 μL 50% H_2SO_4, and 0.02 mL of a concentrated mineral solution of 5% (w/v) $FeCl_3$-$6H_2O$, 0.5% $CuCl_2$-$2H_2O$, 0.5% $CoCl_2$-$6H_2O$, 0.5% $MnCl_2$-$4H_2O$, and 0.5% $ZnCl_2$ in a 1 M solution of HCl.

The selective growth medium DT.ura FOA was identical in composition to DT medium, but supplemented with 20 μg/mL of uracil and either 50, 100, or 150 μg/mL of 5-FOA to facilitate the growth of pyrimidine auxotrophs.

For plating, media were solidified with the addition of 6.5 g/L of Gelrite© brand gellam gum.

Revival of Stored Isolates for IS Screens

Strains chosen for use in gene trapping were revived from cell suspensions frozen in 12% (w/v) DMSO. In Protocol I, inocula were transferred from frozen stock to 3 mL of DT broth medium in screw cap culture tubes using sterile wooden inoculating sticks. Inoculated tubes were tightly capped and incubated at approximately 75º C while tilted at a 25º angle until they displayed an A_{600} of 0.1 or higher. In Protocols II and III, inocula were streaked for isolation on DT plates according to a standard t-streak pattern using sterile wooden inoculation sticks. These plates were inverted, sealed in plastic bags, placed in coffee cans, and incubated at

approximately 75º C for five to fourteen days, depending on the speed of colony formation.

Independent Broth Cultures

Following incubation of the revival plates for type two and three protocols, five colonies were chosen from each and used as inocula for 3 mL DT broth cultures. These were incubated at approximately 75º C while tilted at a 25º angle until they displayed an A_{600} of greater than 0.100.

Isolation of Pyrimidine Auxotrophs for IS Screening

Mutants displaying a phenotype for pyrimidine auxotrophy were isolated by spreading 0.1 to 0.8 mL of the mature revival or independent broth cultures on DT.ura FOA plates. Once dry, plates were inverted, sealed in plastic bags, placed in coffee cans, and incubated at approximately 75º C for seven to fourteen days, depending on the speed of colony formation.

Upon removal from incubation, selective plates were examined for colony growth. Those bearing mature, fully formed colonies (i.e. diameters of 1 – 2 mm), were retained, while those displaying only immature colony development were placed back in incubation. Mutant colonies were then chosen for PCR screening of the target region of the gene trap from those plates retained as fully incubated. In this process, precedence was given to larger colonies on the assumption that this would eliminate the possibility of screening mutants with leaky phenotypes that would be unlikely to be the result of IS insertion into the target region. Once chosen, sterile wooden inoculating sticks were used to transfer inocula from colonies to 3 mL DT.ura FOA broths. In Protocol I, five colonies were chosen from each plate, in Protocol II only

one was chosen from each, and in Protocol III, five colonies were chosen from each plate. In Protocol III, inocula from chosen mutant colonies were also streaked on a separate DT.ura FOA plate to act as a reserve (This was also done with some of the mutants examined from protocol two trapping runs). Mutant broth cultures were then incubated until they reached an A_{600} of 0.15 to 0.4.

DNA Extraction

DNA was extracted from mutant cultures using a procedure derived from that of Pitcher et al. (1989). The cultures were pelleted by centrifugation for seven minutes at 6000 rpm. The supernatant was decanted, and the cells resuspended in 1 mL of distilled water and transferred to 2 mL Eppendorf tubes. The cells were again pelleted and the supernatant decanted. The cells were resuspended in 120 μL of a pH 7 Tris and EDTA (TE) buffer prior to being lysed by the addition of 270 μL of a solution of N-lauryl sarcosine and guanidinium thiocyanate. To this was added 200 μL of 7.5 M ammonium acetate. Seven hundred microliters of a solution of chloroform and *iso*-amyl alcohol was then added to extract the proteins into an organic phase. Following a brief agitation and incubation, this mixture was centrifuged and the aqueous phase carefully drawn off and transferred to fresh Eppendorf tubes. DNA was precipitated by the addition of 300 mL of iso-propyl alcohol, and then pelleted by centrifugation. The pellet was washed in 70% ethanol plus ammonium acetate, repelleted, and dissolved in 40 uL of TE buffer.

PCR Screening of Mutants

Polymerase chain reaction was used to screen extracted mutant DNA for the presence of IS within the target region. All Protocol I mutants and approximately half

the mutants examined under Protocol II were limited to examination of the 585 bp *pyrE* locus. Later screens examined a 735 bp segment containing the *pyrE* gene, the intergenic, promoter-containing region, and the end of the *pyrB* gene immediately upstream of its start codon. The *pyrF* gene was excluded from these screens because *pyrF* mutations, IS-mediated or otherwise, had previously been found to be comparatively rare (Martusewitsch et al. 2000, Grogan et al. 2001). Amplifications were carried out in 20 μL reaction mixtures containing 200 μM dNTPs, 10 mM Tris-Hcl (pH 9.0), 50 mM KCl, 0.1% Triton X-100, 0.5 units of Taq DNA polymerase (Promega or New England Biolabs), and approximately 50 ng of mutant genomic DNA. Primers were added to a final concentration of 2.5 pM. For examination of the *pyrE* locus only, the primers SsopyrE2for (5'-GAAGATCTCTACGTATGAATTTCGC-3') and SsopyrE1rev (5'-CGGGATCCATTGCTAATATTACTCTAC-3') were used. For examination of *pyrE* and the intergenic promoter region, primers SsoINTER1for (5'-CGAATATTCTAAAGTAGTCATCTCTGG-3') and SsopyrE1rev were used. All primers were designed from *S. solfataricus* P2 sequence (She et al. 2001).

Mutants derived from NZ00 strains were screened using the primer pair of SAPYRE-S (5'-TTTCATATGGATTTCGTGAAAGCTCTAC-3') and SAPYRE-A (5'-TTTGGATCCCTAGCTTTTTCCAATATTTTTCAC-3') that were designed from *S. acidocaldarius* sequence to amplify the *pyrE* locus. Unfortunately, though some success was noted, amplifications were inconsistent, and the New Zealand strains were not included with later gene trapping runs.

All PCR reactions were carried out in a MJ Research, Inc. PTC-100 programmable thermocycler under the following cycling conditions: 2 minutes at 94˚ C followed by 25 cycles of denaturation (1 minute at 94˚ C), annealing (1 minute at 57˚ C), and extension (1 minute at 70˚ C). Five μL of the reaction products were then loaded and run on 1% agarose mini-gels with either 1X TAE or 0.5X TBE + EtBr buffer. The gels run with 1X TAE were stained with ethidium promide and destained with dH_2O. All gels were visualized with UV, and photos taken using a Nucleotech digital camera system. Genomic DNA extracts yielding no amplification products were agitated and used as template in a second round of amplifications. Successful amplification was observed in approximately 50% to 60% of such second attempts. In the interest of time, third attempts at amplification were not made.

Mutant DNA yielding amplification products enlarged by more than 100 bp were retained for further analysis.

Localization of Insertions

A series of PCR amplifications under conditions identical to those described above were carried out with each enlarged locus using a set of nested primers designed from *S. solfataricus* (She et al. 2001) sequence with the aid of the web-based Primer 3 tool (Rozen, S., & Skaletsky, H. 2000) to amplify specific portions of the screened region (See figure 2.4 and table 2.2). The object of this was to localize the putative IS within the screened region so as to facilitate rapid sequencing.

Storage of Mutants with Trapped IS

Those mutants found to have IS within the screened portion of the target region were, provided storage streaks had been made, inoculated into 3 mL DT.ura FOA broth cultures. These were incubated until they displayed an A_{600} of greater than 0.300. At this point, the cells were pelleted by centrifugation and the supernatant decanted. The pellets were resuspended in 500 μL of 12% w/v DMSO, transferred by pipette to 2 mL plastic screw cap freezer tubes and then placed in storage in a –70˚ C freezer.

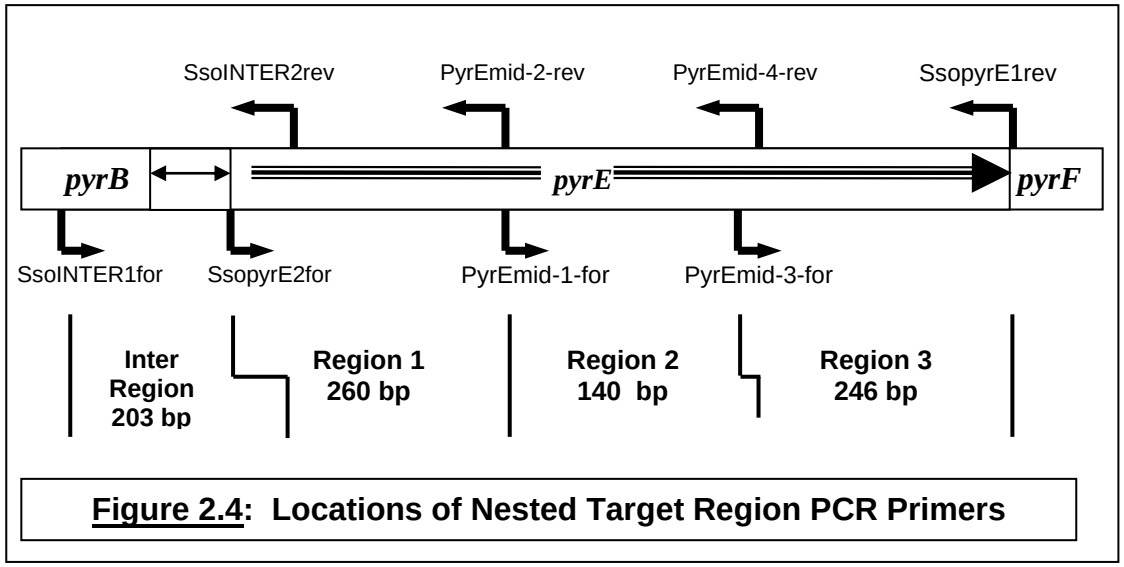


Figure 2.4: Locations of Nested Target Region PCR Primers

Name	Sequence (5' – 3')	Tm	Annealing Positions (5'-3')
SsolNTER1for	cgaatattctaaagtagtcatctctgg	74	1 → 27
SsolNTER2rev	actaaccttacctgatgttaaaacg	68	180 → 203
SsopyrE2for	gaagatctctacgtatgaatttcgc	72	114 → 138
PyrEmid-1-for	gcttgtaaccttaaagagcctatgg	72	349 → 374
PyrEmid-2-rev	ccataggctctttaaggttacaagc	72	349 → 374
PyrEmid-3-for	tccatatgagaaagcaacattgg	62	489 → 511
PyrEmid-4-rev	tgcgtctgaaactttacctcc	62	517 → 538
SsopyrE1rev	cgggatccattgctaatattactctag	76	712 → 735

Table 2.2.: **Nested Primer Set Information**

III. Results and Discussion

Isolation of Mutants

To select for *pyrE* and *pyrF* mutants, 1 x 10^7 – 1 x 10^8 cells were plated on DT.ura FOA medium. Though the concentration of uracil in this medium was held at 20 μg/mL throughout the study, the 5-FOA concentration varied. In the early, Protocol I gene trapping runs, this medium contained 5-FOA at the concentration of 50 μg/mL used by Martusewitsch et al. in their study. After the Protocol I runs netted fewer IS than expected, the concentration was increased to 100 μg/mL to reduce colony formation by cells possessing leaky phenotypes and consequently the background mutants among those screened for IS. The concentration was later increased to a final concentration of 150 μg/mL to provide even greater stringency. These increases in selection stringency were not observed to have much discernable effect, though no focused investigation was made into this.

The apparent frequency of mutation to resistance to 5-FOA was generally observed to be between 10^{-6} to 10^{-7} for most strains from the regions sampled outside of Italy, though instances of apparent rates as high as 10^{-5} and as low as 10^{-8} were not uncommon (data not shown). This is in line with previous observations. The observed apparent mutation rates for the IT02 strains were generally higher than those noted for other strains, as had been expected from the results described for previously studied strains of *S. solfataricus* (Martusewitsch et al. 2000), the dominant species in this set.

Examination of Mutants for Trapped IS

PCR was used to screen the target region of the gene trap of chosen mutants for gross enlargement (>100 bp) indicative of IS insertion. As expected from the work of Bell (2001), *S. solfataricus* primers were found to be quite effective in amplification of loci from *S. islandicus* mutants. Amplification of loci of mutants derived from New Zealand isolates related to *S. acidocaldarius* and *Sulfurisphaera* was considerably more problematic. While there was some success using the *S. acidocaldarius* primers, it was found to be very inconsistent and unreliable, and the NZ00 strains were consequently excluded from later runs.

In the course of Protocol I and about half of Protocol II gene trapping runs, only the *pyrE* locus was examined in this manner. This locus was focused upon because primers specific for this locus in both *S. solfataricus* and *S. acidocaldarius* were in hand, and both Martusewitsch et al. and Grogan et al. (2001) had shown that mutation of the *pyrF* region was comparatively rare. The portion of the target screened was

later extended to include the intergenic promoter region in order to increase then number of IS recovered. About half the IS recovered by Martusewitsch et al. had transposed into this area. This broadening of the portion of the target screened greatly increased frequency of IS capture.

A summary of all gene-trapping results is given in table 2.3. A total of 270 *Sulfolobus* strains isolated from populations residing in the hot springs of five well-separated geographic regions were used in gene trapping experiments. The target genes of over 1700 mutants were screened for enlargement. Of these, 237 derived from 102 separate strains were discovered to display evidence of IS insertion into the target. The Italian strains displayed the highest apparent frequency of insertion-mediated mutation, as was consistent with the results of Martusewitsch and colleagues. Considerable variation was observed in apparent frequency of insertion-mediated mutation among the sets of *S. islandicus*, though they were overall much lower than for the Italian set. This indicates that insertion is not the dominant mode of mutation in this species.

Sample Set	Strains Examined	Mutants Screened	Enlarged Loci Found	Strains Yielding IS	IS Insertions Per Region of Trap*				
					Intergenic Region	Region 1	Region 2	Region 3	Not Known
YNP99	46	298	38	10	21	4	2	3	8
YNP00	33	216	14	11	5	8	2	2	0
YNP01	46	199	3	3	0	0	1	2	0
NZ00	10	56	0	0	0	0	0	0	0
K2K	53	469	58	19	29	9	7	4	9
L2K	30	328	56	21	12	19	9	12	4
IT02	52	167	68	38	31	11	8	7	17
Totals	270	1733	237	102	98	51	29	30	38

Table 2.3: **Summary of Gene Trapping Results**
*Intergenic region not consistently screened by all protocols.

Gene Trapping Protocols

Three different protocols of varying thoroughness were used over the course of gene trapping runs. Protocol I was the most superficial, and involved the plating of a sample of a broth culture directly revived from frozen stock on selective medium. A number of mutant colonies, five being most common, were chosen from those that arose for PCR screening of their target loci. While this protocol permitted rapid examination of mutants from many different strains, it was hampered by the likelihood of examining multiple sibling mutants arising from the same mutation event. This was no doubt one cause of the comparatively small number of IS recovered using it (Table 2.4). Exacerbating this, very few mutants examined under protocol one had their intergenic regions screened.

Sample Set	Strains Examined	Mutants Screened	Enlarged Loci Found	Strains Yielding IS	IS Insertions Per Region of Trap*			
					Intergenic Region	Region 1	Region 2	Region 3
YNP99	20	80	5	4	0	0	2	3
YNP00	19	74	5	4	0	2	1	2
YNP01	Set not screened using this protocol							
NZ00	7	35	0	0	0	0	0	0
K2K	18	93	1	1	0	1	0	0
L2K	14	99	6	6	5	1	0	0
IT02	Set not screened using this protocol							
Totals	78	381	17	15	5	4	3	5

Table 2.4: Results of Gene Trapping with Protocol I
*Intergenic region not consistently screened by this protocol.

Protocol II was somewhat more penetrating. A strain of interest was revived by streaking frozen stock material on non-selective DT medium, and independent broth cultures were inoculated from five of the resulting colonies. Samples of each independent broth culture were then plated on selective medium. A single mutant

colony arising on each selective plate was then chosen and its target locus PCR screened. The benefit of this practice was to eliminate the possibility of screening sibling mutants, thus increasing the actual number of mutations examined. While the Protocol II gene trapping runs did appear to yield the recovery of a greater number of enlarged loci than Protocol I (Table 2.5), this seems to be due more to addition of the intergenic region to the portion of the target screened than it does to the method of choosing mutants. However, it is certain that a greater number of independent insertion events were discovered than was the case with Protocol I.

Sample Set	Strains Examined	Mutants Screened	Enlarged Loci Found	Strains Yielding IS	IS Insertions Per Region of Trap*				
					Intergenic Region	Region 1	Region 2	Region 3	Not Known
YNP99	26	117	12	6	6	2	0	0	4
YNP00	10	41	6	5	5	1	0	0	0
YNP01	46	199	3	3	0	0	1	2	0
NZ00	5	21	0	0	0	0	0	0	0
K2K	26	123	11	10	1	4	1	0	5
L2K	17	80	24	12	1	10	4	8	1
IT02	52	167	74	38	31	11	8	7	17
Totals	182	748	124	74	44	28	14	17	27

Table 2.5: **Results of Gene Trapping with Protocol II**
*Intergenic region not consistently screened by this protocol.

Protocol III combined the two other protocols to provide greater thoroughness. It followed protocol II in the isolation of mutants from independent broth cultures, but five mutants were chosen for PCR screening from each IBC. This had the benefit of ensuring the screening of mutants certain to have arisen from at least five separate mutation events while increasing the chances of effectively sampling the mutation events that occurred within the independent broth cultures. Excluding the effect of screening the intergenic region, this approximately doubled the apparent frequency of

IS recovery, but this is likely skewed by the screening of multiple instances of the same insertion event. It is also notable that the bulk of the enlarged loci discovered by this method were derived from mutants of comparatively few isolates (Table 2.6). With these reservations made clear, it must be pointed out that the Protocol III screens led to the recovery of examples of all the unique IS types ultimately found in the entire course of trapping efforts. It may be recommended that, provided Protocols I and II are used for initial examination of environmental strains to increase the diversity of sampling, Protocol III is the best means of rapidly recovering a high diversity of IS from a small number of strains found to be of particular interest.

Sample Set	Strains Examined	Mutants Screened	Enlarged Loci Found	Strains Yielding IS	IS Insertions Per Region of Trap				
					Intergenic Region	Region 1	Region 2	Region 3	Not Known
YNP99	9	101	21	2	15	2	0	0	4
YNP00	6	101	8	3	0	5	1	2	0
YNP01	Set not screened using this protocol								
NZ00	Set not screened using this protocol								
K2K	13	253	46	9	28	4	6	4	4
L2K	6	149	26	4	6	8	5	4	3
IT02	Set not screened using this protocol								
Totals	34	604	101	18	49	19	12	10	11

Table 2.6: **Results of Gene Trapping with Protocol III**

Localization of IS within Target Region

A second series of PCR reactions were carried out with enlarged loci using a set of nested primers designed from *S. solfataricus* sequence (She et a. 2001) to specifically amplify four different regions of the target, and thus localize the insertion into one of them. The INTER region was made up of 180 bp composed of the first part of the intergenic promoter-containing area and the last portion of the *pyrB* gene. Region one was composed of the first 245 bp of the *pyrE* gene. Region two was

comprised of 162 bp in the center of *pyrE*. Finally, region three was made up of the final 218 bp of *pyrE*, including the 14 bp overlap of it with the *pyrF* sequence.

Localization was originally done to facilitate complete sequencing of insertions, but it soon became clear that it could also be used to determine the regional preference of insertion displayed by the elements recovered. In respect to the latter use, the pooled results for all gene trapping runs, as well as the results of Protocol I and II runs should be examined with caution, as screening of the intergenic region was not routine in all cases. This naturally skews the localization findings, though it is possible to compare the results for the three regions within *pyrE*. As the intergenic region was routinely screened for all Protocol III trapping, as well as the Protocol II trapping runs with the Italian strains, their results maybe used for comparison between the three *pyrE* regions and the intergenic region. Considering this, it is very clear that there is a pronounced tendency toward insertion in the intergenic region, with it being the site of approximately half of insertions into the screened target. There is a similarly large proportion of insertions into region one of *pyrE* compared to regions two and three, both of which had similar, lower levels of insertion. Martusewitch et al. had found that three of the seven insertions they studied were into the intergenic region, a finding consistent with my own, but they had also showed an equal proportion of insertions into the three *pyrE* regions. This may, however, simply be due to sampling error. The preference for the intergenic region fits with previous observations that many transposases show pronounced tendencies to guide transposition into intergenic regions where gene disruption is less likely (Craig 1997, Haren et al. 1999, Chandler

& Mahillon 2002). The mechanism behind this is unknown, though it is likely that some signal common to coding regions of a chromosome is the basis.

Insertions

Among the enlarged loci discovered in mutants from Yellowstone, Lassen, and Kamchatka isolates, even different categories of insertions could be discerned on the basis of size. The first were those of only about 100 bp. These were comparatively few in number, and were largely ignored as their small size indicated they were more likely to be non-autonomous mobile elements such as MITEs than true IS. The remaining six categories divided into insertions of approximately 0.7 kb, 0.8 kb, 1.1 kb, 1.2 kb, 1.3 kb, and 2 kb. Each category was eventually found to correspond to a single type of IS, with the exception of the 1.1 kb insertion category that was composed of two closely related but still distinct IS of 1057 and 1058 bp in length. Five of these were closely related to IS discovered in the genome of *S. solfataricus* and/or *S. tokodaii* (She et al. 2001, Kawarabayasi et al. 2001), and two were wholly novel. The 1.2 kb insertions, found in mutants from both the K2K and L2K sets, were the most common, representing approximately half of all insertions recovered, and all but one of the 56 from the Lassen set. This indicates them to either be present in high copy number amongst the strains that bear them, or simply highly active. The 1.1 kb insertions were the next most common, being responsible for about half the enlargements discovered in K2K, YNP99, and YNP00 mutants. One of the two types of IS in this category was only recovered from mutants of one YNP99 strain. The remaining enlargements among the YNP99 and YNP00 sets were due to the 1.3 kb and 0.8 kb insertions. The latter were also responsible for all the enlargements noted

amongst the YNP01 mutants screened. Neither was observed outside of the Yellowstone collections. Finally, the 0.7 and 2 kb insertions were much more rare, the latter only being found in one instance of an L2K mutant, and the former only among mutants of two K2K strains. These IS will be discussed in greater detail in Chapter III.

Due to the dominance of the Italian set by *S. solfataricus* and how well the IS of this species have been characterized, comparatively little effort was put into being as complete in the examination of the enlarged loci discovered in IT02 mutants. In general, enlargements were due to insertions in the range of 1.1 to 1.5 kb, and the few that were sequenced proved to be identical to elements previously characterized. It is unclear as to how many distinct IS types were responsible for these insertions, though it is likely that many of the twenty within or close to this size range that were discovered in the *S. solfataricus* genome would be represented among them (She et al. 2001, Brugger et al. 2002).

IV. Conclusions

This has been the first major search for active IS from natural strains of *Sulfolobus* using a gene trap methodology. It further constitutes the first examination of *S. islandicus* strains for IS. As was expected, it was found that a gene trap based upon 5-fluoroorotic acid selection for mutation in the pyrE and pyrF loci, as well as their common promoter, described by Martusewitsch et al. (2000) was useful for the recovery of presumably active IS from such natural strains. Using this trap, six categories of insertion based on length and representing seven distinct IS types were recovered. Five of the IS types were found to have close relatives in the two

completely sequenced Sulfolobus genomes, while two were found to be novel elements not previously described, though relics of close relatives of both were found in the genome sequences. These results indicate that active IS are common constituent features of the genomes of natural strains of *Sulfolobus*, and establishes that *S. islandicus* strains possess such active IS.

The FOA selections and PCR screens of mutant loci were focused solely upon the recovery of IS, and records pertinent to the determination of mutation rates were largely neglected as a result. It is consequently difficult to make any conclusions regarding comparative mutation rates between strains of the different geographic regions, nor the impact of IS upon them. However, it is still possible to note that the low apparent frequency of IS-mediated mutation in the *S. islandicus* strains from Yellowstone National Park, Lassen National Park, and the Kamchatka peninsula relative that observed for the *S. solfataricus* strains from Naples, Italy indicates that IS are not as dominant a source of molecular variation in *S. islandicus* as they are in *S. solfataricus*. While there were major variations in the frequency of IS recovery among the different *S. islandicus* strain collections, meaningful extrapolation about their respective populations can not be made, as it is impossible to know how well these populations are represented by the strains examined.

Considering the large number of strains and mutants examined in the course of this work, it is odd that so few distinct IS were recovered. There are a number of possible explanations for this. As was noted earlier, the results indicate that IS are not as significant sources of mutation in *S. islandicus* as they are in *S. solfataricus*, and it is possible and likely that the number of recovered IS was limited by the former

possessing fewer resident IS types than the latter. However, this might also be an artifact of the gene trap methodology. Gene traps are limited in the extent of the target. While this certainly bears upon the benefits of the methodology, it can also be a drawback because it limits the IS that can be recovered to those that can transpose into the target. It is unlikely that this represents a barrier to the recovery of IS that lack stringent target sequence requirements. However, it may well exclude the possibility of recovering those elements with such requirements if the target used does not meet them. Conversely, those IS with sequence requirements met by the target locus will be over represented among the IS recovered. Due to this, it would be prudent to follow up this study with another that uses a different gene trap such as the *Sulfolobus lacS* gene that has already proven amenable to such use (Schleper et al. 1994) to see if the same IS are recovered.

The use of a different target gene would permit the addressing another issue. It was noted that reliable recovery of the same IS type from mutants of a given strain was possible over multiple independent trapping runs. In the few cases in which such multiple independent insertions of the same IS into the target of the same strain were sequenced, the elements were found to be identical. This could be considered evidence that the recovery of a given IS was perhaps largely dictated by its position relative the target. This would be consistent with findings that target accessibility is a dominant factor in the insertion of bacterial IS into a given location (Craig 1997, Haren et al. 1999, Chandler & Mahillon 2002). Presumably a different gene trap would be accessible to a different IS or number of IS.

V. References

1. Baltz, R., Hahn, D., McHenney, M, & Solenberg, P. Transposition of TN5096 and related transposons in *Streptomyces* species. Gene, 115, 61 – 65.

2. Bartosik, D., Sochacka, M., & Baj, J. (2003). Identification and characterization of transposable elements of *Paracoccus pantrophus*. Journal of Bacteriology, 185, 3753 – 3763.

3. Beck, P., Dingermann, T., & Winckler, T. (2002). Transfer RNA gene-targeted retrotransposition of *Dictyostelium* TRE5-A into a Chromosomal UMP Synthase Gene Trap. Journal of Molecular Biology, 318, 273 – 285.

4. Bell, G. (2001). Genetic diversity of natural *Sulfolobus* populations. Masters Thesis, University of Cincinnati.

5. Blount, Z. (2002). Unpublished finding.

6. Brugger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y., & Garrett, R. (2002). Mobile elements in archaeal genomes. FEMS Microbiology Letters, 206, 131 – 141.

7. Bureau, T., & Wessler, S. (1992). Tourist: A large family of small inverted repeat elements frequently associate with maize genes. The Plant Cell, 4, 1283 – 1294.

8. Chandler, M. & Mahillon, J. (2002). Insertion Sequences Revisited, In Craig, N., Craigie, R., Gellert, M., & Lamhowitz, A. (Ed.). *Mobile DNA II* (pp. 305 – 366). Washington, D.C.: American Society for Microbiology Press.

9. Ciampi, M, Schmid, M., & Roth, J. (1982). Transposon TN10 provides a promoter for transcription of adjacent sequences. Proceedings of the National Academy of Sciences (USA), 79, 5016 – 5020.

10. Cirillo, J., Barletta, R., Bloom, B., & Jacobs, W. Jr. (1991). A novel transposon trap for *Mycobacteria*: isolation and characterization of IS1096. Journal of Bacteriology, 173, 7772 – 7780.

11. Daboussi, M., Langin, T., & Brygoo, Y. (1992). Fot1, a new family of fungal transposable elements. Molecular and General Genetics, 232, 12 – 16.

12. Feng, X., Ou, L, & Ogram, A. (1997). Cloning and sequence analysis of a novel insertion element from plasmids harbored by the carbofuran-degrading bacterium, *Sphingomonas sp. CFPO6*. Plasmid, 37, 169 – 179.

13. Gay, P., Coq, W., Steinmetz, M., Berkelman, T., & Kado, C. (1985). Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. Journal of Bacteriology, 164, 918 – 921.

14. Granbastien, M., Spielmann, & Caboche, M. (1989). Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature, 337, 376 – 380.

15. Grogan, D., Carver, G., & Drake, J. (2001). Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermophilic archaeon *Sulfolobus acidocaldarius*. Proceedings of the National Academy of Sciences (USA), 98, 7928 – 7933.

16. Kallastu, A., Horak, R., & Kivisaar, M. (1998). Identification and characterization of IS1411, a new insertion sequence which causes transcriptional activation of the phenol degradation genes in *Pseudomonas putida*. Journal of Bacteriology, 180, 5306 – 5312.

17. Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Otsuka, R., Nakazawa, H., Takamiya, M., Kato, Y., Yoshizawa, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Masuda, S., Yanagii, M., Nishimura, M., Yamagishi, A., Oshima, T., & Kikuchi, H. (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* Strain 7. DNA Research, 8, 123 – 140.

18. Lee, Y., Fan, S., Chiu, L, & Hsia, K. (2001). Isolation of an insertion sequence from *Ralstonia solanacearum* Race 1 and its potential use for strain characterization and detection. Applied and Environmental Microbiology, 67, 3943 – 3950.

19. Marberd, N., Flavell, R., & Thompson, R. (1987). Identification of a transposon-like insertion in a Glu-1 allele of wheat. Molecular and General Genetics, 209, 326 – 332.

20. Martusewitsch, E., Sensen, C., & Schleper, C. (2000). High spontaneous mutation rate in the hyperthermophilic archaeon *Sulfolobus solfataricus* is

mediated by transposable elements. Journal of Bacteriology, 182, 2574 – 2581.

21. McClintock, B. (1948). The significance of responses of the genome to challenge. Science, 226, 792 – 801.

22. Meyer, C., Pouteau, S., Rouze, P., & Caboche, M. (1994). Isolation and molecular characterization of dTnp1, a mobile and defective transposable element of *Nicotiana plumbaginifolia*. Molecular and General Genetics, 242, 194 – 200.

23. Pitcher, D., Saunders, N., & Owen, R. (1989). Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. Letters in Applied Microbiology, 8, 151 – 156.

24. Prentki, P., Teter, B., Chandler, M., & Galas, D. (1986). Functional promoters created by the insertion of transposable element IS1. Journal of Molecular Biology, 191, 383 – 393.

25. Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S. & Misener, S. (eds.). Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.
Web Tool At: www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi
(Code available at www-genome.wi.mit.edu/genome_software/other/primer3.html.)

26. Saedler, H., Reif, H., Hu, S., & Davidson, N. (1974). IS2, a genetic element for turn-off and turn-on gene activity in *E. coli*. Molecular and General Genetics, 132, 265 – 289.

27. Schleper, C., Roder, R., Singer, T., & Zillig W. (1994). An insertion element of the extremely thermophilic archaeon *Sulfolobus solfataricus* transposes into the endogenous β-galactosidase gene. Molecular and General Genetics, 243, 91 – 96.

28. Schwarz-Sommer, Z., Leclercq, L., Gobel, E., & Saedler, H. (1987). Cin4, an insert altering the structure of the A1 gene in *Zea mays*, exhibits properties of nonviral retrotransposons. The EMBO Journal, 6, 3873 – 3880.

29. Shapiro, J. (1969). Mutations caused by the insertion of genetic material into the galactose operon of *Escherichia coli*. Journal of Molecular Biology, 40, 93 – 105.

30. She, Q., Singh, R., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M., Chan-Weiher, C., Clausen, I., Curtis, B., Moors, A., Erauso, G., Fletcher, C., Gordon, P., Heikamp-de Jong, I., Jeffries, A., Kozera, K., Medina, N., Peng, X., Thi-Ngoc, H., Redder, P., Schenk, M., Theriault, C., Tolstrup, N., Charlebois, R., Doolittle, W., Duguet, M., Gaasterland, T., Garrett, R., Ragan, M., Sensen, C., & Van der Oost, J. (2001). Complete genome of the crenarchaeote *Sulfolobus solfataricus* P2. Proceedings of the National Academy of Sciences (USA), 98, 7835-7840.

31. Shepherd, N., Schwarz-Sommer, Z., Wienand, U., Sommer, H., Deumlin, B., Peterson, P., & Saedler, H. (1982). Cloning of a genomic fragment carrying the insertion element Cin 1 of *Zea mays*. Molecular and General Genetics, 188, 266 – 271.

32. Solenberg, P. & Bergett, S. (1989). Method for selection of transposable DNA and characterization of a new insertion sequence, IS493, from *Streptomyces lividans*. Journal of Bacteriology, 171, 4807 – 4813.

33. Szeverenyi, I., Hodel, A., Arber, W., & Olasz, F. (1996). Vector for IS element entrapment and functional characterization based on turning on expression of distal promoterless genes. Gene, 174, 103 – 110.

34. Tsay, Y., Frank, M., Page, T., Dean, C., & Crawford, N. (1993). Identification of a Mobile Endogenous Transposon in *Arabidopsis thaliana*. Science, 260, 342 – 344.

35. Voytas, D. & Ausubel, F. (1988). A copia-like transposable element family in Arabidopsis thaliana. Nature, 336, 242 – 244.

36. Waskar, M., Kumar, D., Kumar, A., & Srivastava, R. (2000). Isolation of a novel insertion sequence from *Mycobacterium fotuitum* using a trap vector based on inactivation of a *lacZ* reporter gene. Genetics and Molecular Biology, 146, 1157 – 1162.

37. Whitaker, R.. Personal communication.

Chapter 3
Molecular Analysis of Recovered IS

I. Introduction

There are a number of benefits to the discovery of IS by gene trapping versus genome sequence analysis that are relevant to the characterization of the elements recovered. The preferential trapping of active sequences greatly increases the level of confidence that characteristics identified in an IS recovered in this manner are not artifacts introduced by drift, as is the case with inactive copies no longer under selective restraint. A second benefit is that, in order to function properly as a gene trap, the sequence of the target must be known. It is thus possible to precisely determine the joint between the IS and the native DNA into which it has inserted, and to therefore identify and characterize the terminal inverted repeats of the element. The same is true of the directly repeated sequences marking the sites of insertion.

Knowledge of the sequence of the target also permits rapid extraction of IS sequence for assembly and analysis. Much can be gleaned from the DNA sequence of an IS, inclusive of sequence composition analysis, open reading frame (ORF) identification, and putative encoded protein analysis. The Basic Local Alignment Search Tool (BLAST) search program set provided by the National Center for Bioinformatics can then be used to identify local sequence similarities between the IS DNA and putative protein sequences and those stored in databases such as NIH's GenBank (Altschul et al. 1990, Altschul et al. 1997). In this manner, related sequences can be identified for use in phylogenetic reconstruction to examine the evolutionary relationships of the IS. Comparison of phylogeny of examples of IS to

the geographic distribution of the hosts from which they are recovered can then be used to gain perspective on gene flow between different populations, while comparison of IS and host phylogenies can shed light on evidence of horizontal gene exchange. At the same time, identification of relatives can permit the placement of an element in an IS family. This is important for further putative characterization of recovered IS, as family members often share operational characteristics.

Possession of IS sequences can also be used to construct PCR primers to specifically amplify close relatives from strain chromosomal DNA. With these primers, large numbers of strains can be screened for the presence of recovered IS. This has a number of applications as a means of both identifying strains that should be subjected to gene trap screening and giving some indication of geographic distribution of screened IS. The capacity for IS-specific amplification can also contribute to the construction of Southern blotting probes that are useful in studying the variation in number and site of insertion of the IS in the genomes of strains of interest. Such studies can be of great value, as is discussed in Chapter I.

II. Materials and Methods

Sample Preparation for Sequencing

Enlarged mutant loci chosen for sequencing were amplified by polymerase chain reaction in 50 to 60 μL reaction mixtures containing 200 μM dNTPs, 10 mM Tris-HCl (pH 9.0), 50 mM KCl, 0.1% Triton X-100, 1 unit of Taq DNA polymerase (Promega or New England Biolabs), and approximately 100 ng of mutant genomic DNA. Primers were added to a final concentration of 2.5 pM. In most cases, the entire target, including the intergenic region, was amplified using the primers

SsoINTER1for (5'-CGAATATTCTAAAGTAGTCATCTCTGG-3') and SsopyrE1rev (5'-CGGGATCCATTGCTAATATTACTCTAC-3'). Reactions were carried out in a MJ Research, Inc. PTC-100 programmable thermocycler under the following cycle conditions: 2 minutes at 94 C followed by 25 cycles of denaturation (1 minute at 94 C), annealing (1 minute at 57 C), and extension (1 minute at 70 C).

To examine for successful amplification, one μL of the reaction products were loaded and run on 1% agarose mini-gels with either 1X TAE or 0.5X TBE + EtBr buffer. The gels run with 1X TAE were stained with ethidium bromide and destained with dH_2O. All gels were visualized with UV, and photos taken using a Nucleotech digital camera system. Those product mixtures displaying successful amplification were purified by use of a Millipore Microcon® YM-100 centrifuge filter device with a nominal molecular weight limit of 100 kDa according to the manufacturer's directions, with all centrifugations carried out in a Spectrafuge 14M bench top centrifuge. Purified products were finally suspended in 25 to 30 μL of dH2O, and 0.5 to 1 mL of each run on a 1% agarose minigel as described above. This permitted both verification of purity by examination for evidence of contaminants and product quantification by comparison to known concentrations of a molecular weight marker (Promega λ DNA cut with HindIII and EcoRI). Products were considered to be ready for sequencing provided they were found to be sufficiently pure and of a concentration of greater than (0.44 ng/μL)(product length in bp/100). Products with concentrations greater than twice this were diluted as appropriate. Products were submitted for sequencing in 12 μL total volume samples with the appropriate primer added to each at a final concentration of 583.3 nM.

Sequencing

Sequencing was carried out by the Cincinnati Children's Hospital Medical Center Genomics Facility. A chain termination sequencing reaction was performed using the ABI PRISM BigDye Terminator Cycle Sequencing Kit. The sequencing thermocycle used had the following thermal profile: 95º C for five minutes, followed by thirty cycles of 95º C for 0.5 minutes, 56º C for 0.5 minutes, and 60º C for 4 minutes. The reaction products were then cleaned, separated by capillary electrophoresis, and analyzed by an ABI PRISM 3700 DNA Analyzer. Text sequences were examined and edited for miscalls by visual inspection of the electropherograms printouts for each reaction. Sequencing results from this facility typically produced approximately 700 bp of useful sequence, requiring at least one forward and one reverse sequencing run to retrieve the full sequence of any single IS.

Sequence Assembly

Sequence analysis began with the identification of the joints between the IS responsible for the enlarged target and the target itself and the assembly of the forward and reverse sequences into a single sequence. This was accomplished by performing a pair wise alignment of the sequence obtained with the forward primer and that of the wild type target obtained from the *S. solfataricus* genome (She et al. 2001) using the "BLAST 2 Sequences" tool (Tatusova & Madden 1999) available from the National Center for Bioinformatics BLAST (Basic Local Alignment Search Tool) web page (http://www.ncbi.nlm.nih.gov/BLAST/). The reverse complement of the sequence obtained with the reverse primer was then generated using the reverse complement tool of the web-based Sequence Manipulation Suite (Stothard 2000), and a pair wise

alignment performed between it and the wild type target. The portion of this containing the IS was then aligned with the forward sequence to determine overlap, and the two joined to assemble the full sequence of the IS. Alignments of the terminal 20 – 100 base pairs of either end were used to identify the inverted repeats of the element, while direct repeats were identified by visual inspection of the joints between the ends of the IS and the target.

Molecular Characterization of Recovered IS

Once recovered, IS sequences were characterized using a battery of web based tools and services. Reverse complements were constructed and G+C percentages were determined using the appropriate tools of the Sequence Manipulation Suite (Stothard 2000). Putative open reading frames were identified with either the ORF finder program bundled with the NEBcutter restriction profile tool (Vincze & Roberts 2003) or that provided by NCBI (Tatusova & Tatusov 2003). It is worth nothing that the former proved by far the greater utility and ease of use of the two. Putative translations of the ORFs so discovered were BLAST-searched against the NCBI protein database to identify nearest relations (Altschul et al. 1997). This also permitted identification of conserved domains in the primary sequence and presumptive IS family assignment. Putative protein molecular weights and theoretical isoelectric points were determined using the Expert Protein Analysis System (ExPASy) "Compute pI/MW tool". When no relatives of near isoformic degree were discovered, BLAST searches of nucleotide sequence were also done (Altschul et al. 1997) to detect possibly fragmented copies not noticed in previous sequence deposits to GenBank. Promoters were detected using the Neural Network Promoter Prediction

tool of the Berkley Drosophila Genome Project (Reese 2001) and confirmed by visual inspection.

IS-Specific PCR Screening

Internal PCR primers were designed using the Primer 3 web-based tool (Rozen & Skaletsky 2000) to specifically amplify each unique IS type recovered (See Table 3.1). A series of PCR reactions were then performed using these primers to probe the chromosomal DNA of a number of strains from each of the five sampled regions to detect the presence of these IS and related isoforms. Amplifications were carried out in 15 μL reaction mixtures containing 200 μM dNTPs, 10 mM Tris-Hcl (pH 9.0), 50 mM KCl, 0.1% Triton X-100, 0.425 units of Taq DNA polymerase (Promega or New England Biolabs), and approximately 50 ng of mutant genomic DNA. Primers were added to a final concentration of 2.5 pM. Reactions were carried out in a MJ Research, Inc. PTC-100 programmable thermocycler under the following cycling

IS	Primer	Primer Sequence	Position Relative Standard Orientation
ISC735	ISC735for	5'-GGGTCGACGGTGTCCAACATTTTCATTCACC-3'	710 - 733
	ISC735rev	5'-GGGGATCCTATAAAACATTAGAAGGGCGGG-3'	28 - 50
ISC796	ISC796for	5'-GGGTCGACTCGTCGTAAGCTACAAATTCTGG-3'	766 – 789
	ISC796rev	5'-GGGGATCCGAGTATTACTTATGGGTAGGAAGC-3'	37 – 61
ISC1057/ ISC1058b	ISC1057for	5'-CCGTCGACGCTTTGTTGATCTCTTCAATTTATTTC-3'	1030 – 1057
	ISC1057rev	5'-CCGGATCCGCTTTGTGGGCTACTTCAAAATTATAC-3'	2 – 28
ISC1205	ISC1205for	5'-CCGTCGACAGATTTTCATCAAACCCAGCAC-3'	1163 – 1176
	ISC1205rev	5'-CCGGATCCTCGGTAATACAAAGCTCACAG-3'	19 – 42
ISC1288	ISC1288for	5'-CCGTCGACGAGTGTCCCAAGTGCAAATAA-3'	1262 – 1285
	ISC1288rev	5'-CCGGATCCGTGTCCCAAGCGCAAATAAT-3'	1 – 24
ISC1926	ISC1926for	5'-GGGTCGACAAGGGCTGAATCCTTTCTCAC-3'	1 – 21
	ISC1926rev	5'-GGGGATCCTCATTCATGATCGAGTGTAGGG-3'	1868 – 1859
	ISC1926midfor	5'-GAAGAACCTAAGGATGCCACAC-3'	537 – 559

Table 3.1: IS-Specific Primers Used as PCR Probes

conditions: 2 minutes at 94 C followed by 29 cycles of denaturation (1 minute at 94 C), annealing (1 minute at 57 C), and extension (1 minute at 70 C). Five μL of the reaction products were then loaded and run on 1% agarose mini-gels with either 1X TAE or 0.5X TBE + EtBr buffer. The gels run with 1X TAE were stained with ethidium promide and destained with dH_2O. All gels were visualized with UV, and photos taken using a Nucleotech digital camera system.

Phylogenetic Analysis

The amino acid sequences of proteins returned by BLASTP searches using the predicted primary sequences of putative recovered IS open readings frames, as well as those of other IS found to be in the same family as listed at the IS-Finder web site, were retained for phylogenetic reconstruction. These were aligned using ClustalX version 1.8 for Windows 95 (Thompson et al. 1997). Alignment was carried out using the Gonnet matrix using the parameters of 10.0 open gap, 0.10 pair wise-gap, and 0.2 multiple-gap extension penalties. Hydrophilic and residue-specific penalties were also enabled.

The PAUP version 3.11 for Macintosh software package was used to reconstruct phylogenies from the aligned amino acid sequences. Reconstruction was performed using a maximum parsimony analysis using a heuristic search with random stepwise addition of taxa through 200 replications, and branchswapping using the tree-bisection reconnection (TBR) option. The trees generated were then subjected to bootstrap analysis with 1000 replicates using the same heuristic parameters for the original reconstruction, save for the reduction to random stepwise addition to 10

replicates per search. Trees corresponding to the 50% bootstrap consensus rule were retained.

III. Results

DNA Sequences of Recovered IS

Putative IS-bearing fragments were selected for sequencing on a number of bases. Initially, when the number of enlarged loci recovered was low, at least two representatives of each size class discovered in the different sample sets screened were sequenced. Later, when a greater number of enlarged loci had been identified, and sequence data had permitted analysis of IS examples for restriction sites, insertions were subjected to restriction digestion to putatively identify them prior to sequencing. Later in the course of the project, an attempt was made to sequence as many examples of certain individual IS of independent derivation as possible, provided, of course, that multiple examples were available, so as to facilitate phylogenic inquiry. An exception of the above regimen was made with the enlargements recovered from the IT02 sample set. A representative number of putative IS sequenced from among these, and all were found to be identical to elements already reported to have been discovered in the *S. solfataricus* genome. As the strains composing the IT02 set were determined to be identical to *S. solfataricus* P2 (Whitaker 2003), this finding together with the large number of enlargements found led to the cessation of sequencing work with them. Sequencing efforts led to the identification of seven distinct IS types among the elements recovered: ISC735, ISC796, ISC1057, ISC1058b, ISC1205, ISC1288, and ISC1926 (See table 3.2). ISC (Insertion Sequence Crenarchaeota) designations were

	ISC735	ISC796	ISC1057	ISC1058b
Length (bp)	735	796	1057	1058
%G+C	41%	43%	41%	39%
IR Length (bp)	18	21	8 – 9	8
DR Length (bp)	8	8	8	8
ORFs	1	1	1	1
Family	IS6	IS1	IS5	IS5
Closest Documented Relative, Source Organism (%Identity/%Similarity; aa Measured Over)	ISt847, *S. tokodaii* (36/55, 211)	ISt796, *S. tokodaii* (88/94, 244)	ISC1058, *S. solfataricus* (YNP 83/90, 299) (K2K 85/89, 299)	ISC1058, *S. solfataricus* P2 (72/82, 299)
Special Features	COG3316 Conserved Domain	IS1 Family Conserved Domains COG3677 and InsB	Transposase 11 DDE Conserved Domain	Variant of ISC1057
Sets Recovered From	K2K	YNP99, YNP00, YNP01	YNP99, YNP00, K2K	YNP99

	ISC1205	ISC1288	ISC1926
Length	1204 – 1211	1279 – 1288	1926
G+C	45 – 46%	43%	41%
IR Length (bp)	17 – 20	34	0
DR Length (bp)	4 – 7	5	0
ORFs	1	1	2
Family	Undefined	IS5	IS200/IS605
Closest Documented Relative, Source Organism (%Identity/%Similarity; aa Measured Over)	ISC1217, *S. solfataricus* P2 (32/48, 320)	ISC1290, *S. solfataricus* P2 (92/93, 307)	ISC1913, *S. solfataricus* P2 (ORF I: 90/95, 213) (ORF II: 83/89, 376)
Special Features	Transposase 29 Conserved Domain		Resolvase Conserved Domain COG0675.1 Conserved Domain
Sets Recovered From	L2K, K2K	YNP99, YNP00	L2K

Table 3.2: Characteristics of Recovered Insertion Sequences

given on the basis of the length in base pairs of the first example of the type to be fully sequenced.

Some trouble was encountered during sequencing attempts. It was rapidly discovered that sequencing using Intergenic Region-specific primers was prone to a high frequency of nucleotide miscalls, dirty sequence, and sequencing failures. Attempts were made to correct these problems with new primers. However, neither longer nor shorter primers based on the old ones, nor entirely new primers designed to bind to slightly different positions than the old ones achieved a noticeable improvement in sequence quality or reliability. It is assumed that possible secondary structure characteristics in the Intergenic Region that are expressed at low annealing temperatures may interfere with sequencing. Similarly, problems were encountered in sequencing using the regional primers pyrEmid2rev and pyrEmid1for, though alternate primers with higher levels of success were ultimately discovered (not shown.). These problems led to the preferential sequencing of IS that had inserted into Region Three of *pyrE*. Significant difficulties were also encountered in the sequencing of ISC1058b and ISC1288 that did not seem to be related to the region of their insertion. Aside from some secondary structural characteristics not readily apparent, it is unclear as to what the cause of this may be for ISC1058b. However, the long terminal inverted repeats, 33 bp, of the element are likely at fault in the problems in sequencing ISC1288. At the relatively permissive annealing temperature of the sequencing reaction used, it is quite likely that the IRs of an element could anneal to

each other, generating a hairpin structure, and thus preventing the sequencing of the IS.

In the course of gene trapping, perhaps a half dozen instances were noted of target enlargements less than that caused by the insertion of an IS. Two of these were sequenced to determine if these were due to the insertion of non-autonomous mobile genetic elements. In the first case, the enlargement was due to duplication of approximately 100 bp of the intergenic region. In the second case, the enlargement was due to the insertion of a segment of DNA that showed no significant similarity to any known sequence, but also did not show the presence of characteristics such as terminal inverted repeats typical of non-autonomous elements.

Profiles of Recovered IS

ISC735

The smallest IS yet recovered from any *Sulfolobus* species, ISC735 bears the further distinction of being the only IS discovered in the course of this project to display no significant similarity to any catalogued nucleotide sequences. In the course of IS-specific PCR screening it showed the lowest frequency of positive amplifications of any element, with only eleven strains from Kamchatka displaying its presence. This indicates that it is also likely to possess the most restricted geographic range of any of the recovered elements as well. This is odd considering it was found to display a high apparent level of activity in the one strain from which it was recovered.

A putative ORF extends over approximately 88% of the element. The stop codon lies partly within the right inverted repeat, and the start codon (ATG) is preceded by putative TATA box (33 bp) and ribosomal binding site elements (18 bp),

though the intervening distance is greater than is typical. The ORF encodes a putative protein of 214 amino acids that was predicted to have a molecular weight of 24.5 kDa and a theoretical pI of 10.33. BLAST searches with the amino acid sequence revealed its closest documented relative to be the 233 aa transposase encoded by ISSt847, an IS present in *S. tokodaii* in at least five full or partial copies. The BLAST score between the two was rather low, however, displaying an identity of only 36% and a similarity of only 55% over 211 amino acids. The only other *Sulfolobus* IS to which it displayed a significant degree (29% identity, 51% similarity over 134 aa) of similarity is ISC774, an element discovered in the *S. solfataricus* P2 genome. It displayed similar levels of relationship to a number of elements discovered in strains of *Halobacterium*, *Aquifex*, *Methanococcus*, and *Pyrococcus*. Interestingly, a significant relationship was also detected with a putative protein detected in the *Arabidopsis* genome. These relationships are shown the phylogram shown in figure 3.1 based on predicted transposase sequences.

The conserved domain protein COG3316 was detected within the internal segment stretching from aa41 to aa208. This domain is characteristic of transposases related to that of IS240-A from *Archaeoglobus fulgidus*. This element belongs to the IS6 family, and ISC735 was given its family designation on this basis. Members of this family are generally quite small, with the bacterial members typically ranging from 789 to 800 bp. All members of this family that have been studied operationally so far have been shown to transpose via a cointegrate intermediate and are likely replicative. As in the case of ISC735, no member encodes a resolvase, and are thus presumably dependent upon those supplied by the host or other elements present in the

same genome (Chandler & Mahillon 2002). It is possible that the apparent high activity of ISC735 is related more to an accumulation of copies in the host chromosome. Though no particular target site specificity has been noted for any member of IS6 (Chandler & Mahillon 2002), there are indications that this is not be the case for ISC735. All enlargements attributed to it in this study involved its insertion into the intergenic region, and the two sequenced examples, both representing independent insertions, displayed transposition into the same position six nucleotides from the start of the *pyrE* gene.

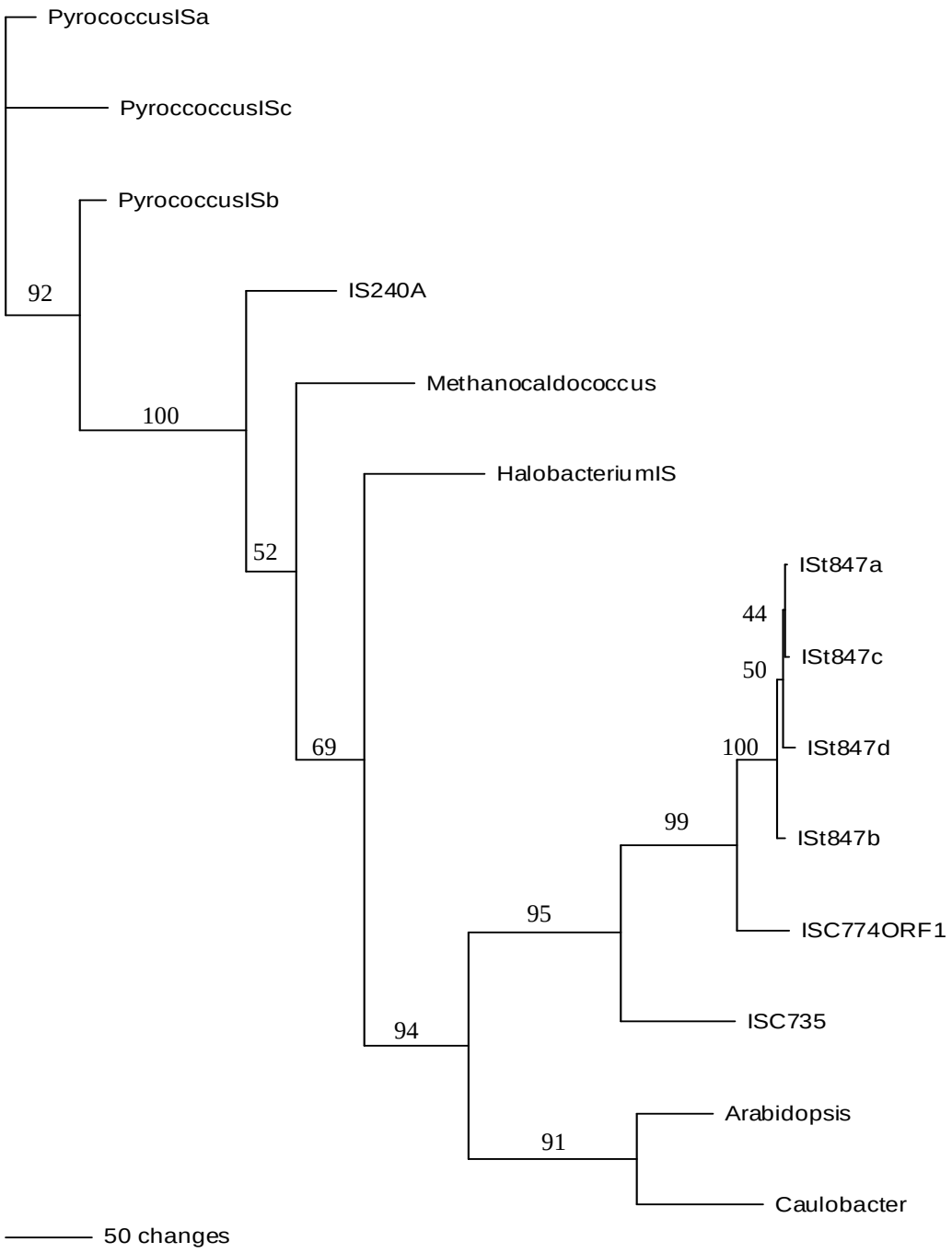


Figure 3.1: Phylogram with Bootstrap Values of the Relationship Between ISC735 and Other Members of the IS6 Family Based on Predicted Transposase Sequences

ISC796

ISC796 is bounded by 21 bp imperfect inverted repeats, produces 8 bp direct repeats, and carries a single putative open reading frame of 734 bp that spans approximately 92% of the element. The ORF begins with a typical ATG start codon, and ends on an ochre stop codon. A putative ribosomal binding site is located 8 bp upstream of the start codon, and a possible Archaeal promoter of moderate strength is present further upstream within the latter half of the left inverted repeat. Interestingly, a strong bacterial promoter is present as well, indicating that the element is capable of subsisting within a bacterial host. A BLAST search with the ISC796 nucleotide sequence revealed a high level (84% over 794 nt) of sequence similarity to ISSt796, an IS present in the *S. tokodaii* genome in five copies. The search also identified six stretches of 76 to 266 nucleotides in the *S. solfataricus* P2 genome that displayed 85% to 93% identity to segments of ISC796, indicating a fragmentation of previously active copies.

Analysis of the putative ORF predicted the encoded protein to be 244 amino acids in length with a molecular weight of 28.7 kDa and a theoretical pI of 9.9. As expected from the nucleotide results, a high BLAST score was shown between the protein's primary sequence and that of ISSt796, with an identity of 84% and a similarity of 94% over the full sequence length. Lower BLAST scores between it and the transposases of ISC1173 (38% identity, 52% similarity over 230 aa) and ISSt1173 (34% identity, 52% similarity over 223 aa).

The presence of five conserved protein domains in two separate regions was also predicted. Three of these, Transposase 12, COG3677, and IS1 pfam03811 were predicted to occupy a region from aa9 to aa97 with alignments of 70.5%, 30.7%, and 86.3% respectively. The Transposase 12 domain is characteristic of the transposases of ISL3 family elements and displayed an alignment score of only 30.7%. COG3677 and IS1 pfam03811/InsA both correspond to the DNA binding domain of IS1 family transposases and displayed an alignment with the predicted protein sequence of 70.5% and 92%, respectively. The region of the predicted protein stretching from aa113 to aa231 showed alignment scores of 86.3% and 95.9%, respectively, to the Transposase 27 and COG1662 domains, both of which correspond to the InsB domain of IS1 transposases. An IS1 DDE motif, as defined by Ohta et al. (2002), was also detected in this same region that was identical to that demonstrated for ISSt796. This was expected, as InsB is the catalytic domain. These results led to the assignment of ISC796 to the IS1 family. The relationship between ISC796 and other selected members of the IS1 family are shown in the transposase sequence-based phylogram shown in figure 3.2.

In the course of IS-specific PCR screening, ISC796 showed the highest frequency of amplification of any recovered IS. Amplification of greater than 70% was observed in the screening of all sets save for that of NZ00 strains, for which the observed frequency was approximately 31% (See figure 3.6). This indicates that ISC796 is quite cosmopolitan and a common resident of *Sulfolobus*. This was born out from preliminary DNA-DNA hybridization results using ISC796-specific probes that showed one to six copies of a related element to be present in the genomes of six

Lassen strains. In light of this, it is interesting to that ISC796 was not commonly recovered. Only six strains, all from Yellowstone, yielded a mutant found to have resulted from its transposition into the target. No instances of multiple insertion events were observed in the same strain, indicating relatively low levels of transposition. As IS1 family elements display a low frequency of transposition ($\sim 10^{-7}$) (Chandler & Mahillon 2002), this observation is in line with ISC796's family assignment.

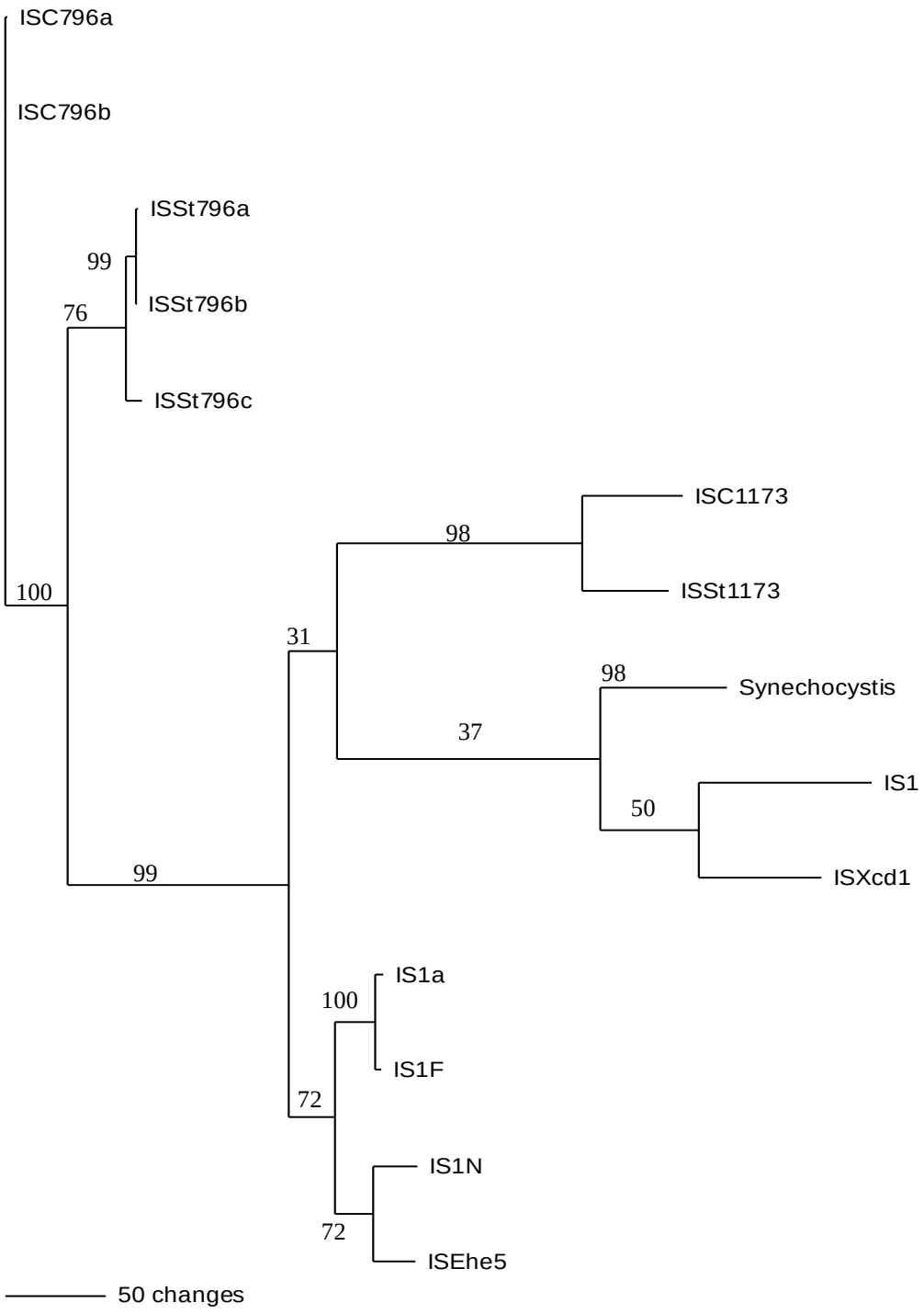


Figure 3.2: Phylogram with Bootstrap Values of the IS1 Family Based on Transposase Sequences

In all documented bacterial IS1 family elements, the InsA DNA-binding and the InsB catalytic domains are encoded by two ORFs. A programmed frameshift mediated by an A_6C sequence in the region overlapped by the two permits the expression of the complete, two- domain transposase. In most circumstances, however, the *insA* gene is the only one transcribed, and the resulting InsA-containing protein binds to the terminal inverted repeats. As the promoter that jointly controls both genes is located in the IRL sequence, this effectively blocks transcription (Ohta et al. 2002). This mode of transcriptional regulation is not possible in the Archaeal IS1 elements so far documented, as all have been found to possess only one ORF. However, the positioning of a strong to moderately strong promoter within an inverted repeat does indicate that transposition regulation, at least in ISC796, depends heavily on its occlusion by the DNA binding domain of the expressed transposase. As this means that RNA polymerase will be physically blocked from access to the promoter, control of transcription will be much higher than with those elements dependent upon weak promoters for regulation. This likely explains the observed low frequency of recovery.

ISC1057 and ISC1058b

Approximately half of the IS recovered from Kamchatka and Yellowstone strains were found to be elements related to ISC1058, an IS present in multiple copies in the *S. solfataricus* genome. Three distinct elements could be discerned among these. Two were 1057 bp elements found in Yellowstone and Kamchatka isolates, respectively. These were found to display nucleotide identities to ISC1058 of 82% and 86%, respectively. The most pronounced difference noted between the 1057 bp

elements and ISC1058 was in the possession by the former of inverted repeats of only 8 to 9 bp, compared to the 19 bp of the latter. Due to a high level of nucleotide sequence identity (93%) between them, it seemed unnecessary to consider them entirely separate elements, and were given the joint designations of ISC1057 variants Yellowstone and Kamchatka, respectively.

Both ISC1057 variants produce direct repeats of 8 bp, possess inverted repeats of 8 to 9 bp, and display G+C contents of 41%. They were also found to both possess a putative ORF extending over 85% of their length and encoding a potential protein of 299 aa that terminate with an opal stop codon. Putative ribosomal binding sites, and strong to moderate promoters, both bacterial and Archaeal, were also detected upstream of their start codons.

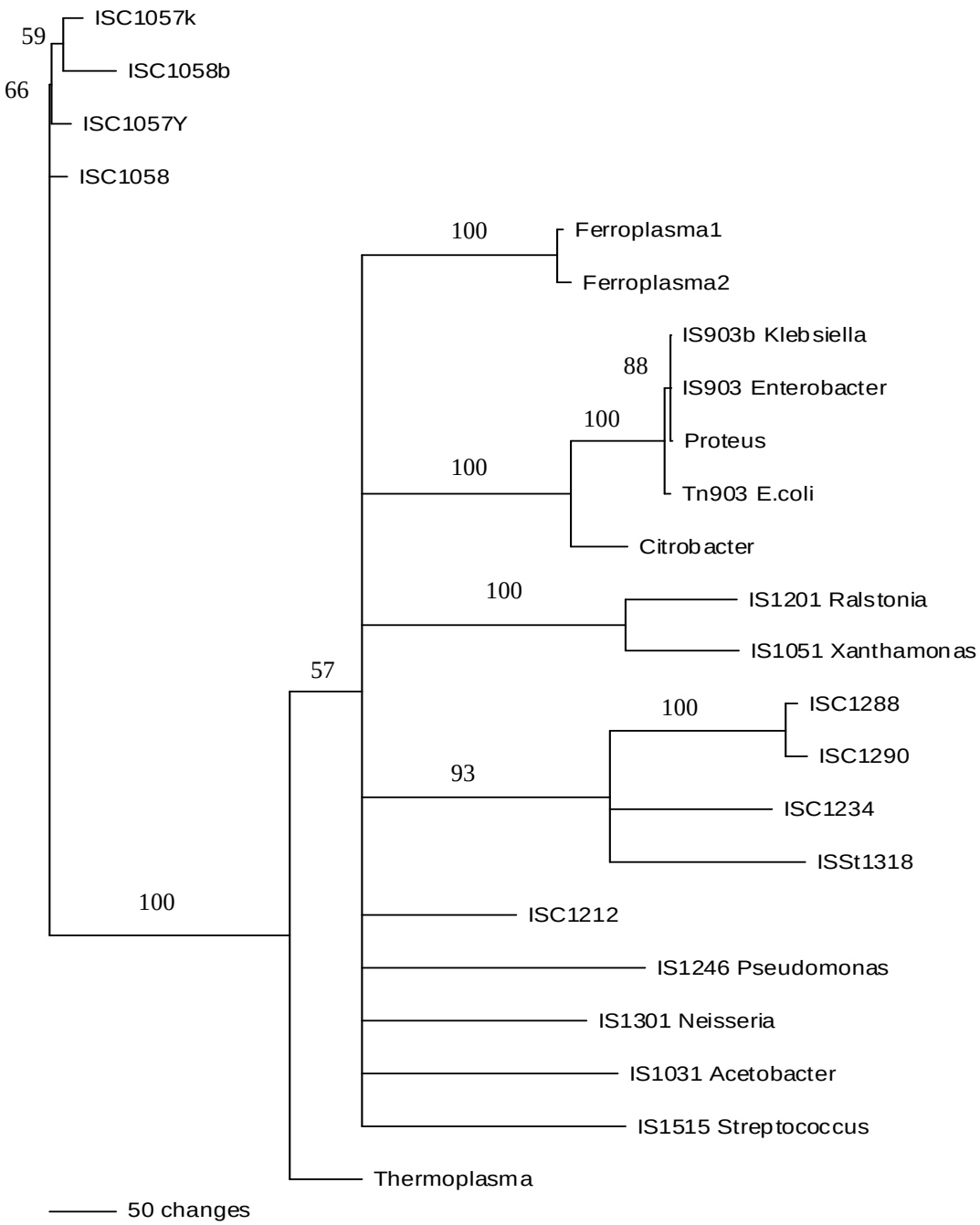
The potentially encoded proteins were found to possess molecular weights of 34.97 kDa for the Yellowstone variant, and 34.88 kDa for the Kamchatka variant, and isoelectric points of 9.86 and 9.79, respectively. In BLAST searches the primary sequence of both predicted proteins displayed high similarity scores to the transposase of ISC1058, as was expected. The Yellowstone variant showed an 83% identity and 90% similarity to the ISC1058 protein over 299 aa, and the Kamchatka an 85% identity and 89% similarity over the same stretch. Pair wise alignment showed the two to display 88% identity and 91% similarity to each other. Both displayed low-level BLAST scores (Identities between 24% and 32% and similarities between 41% and 48% over 230 to 280 aa) to a large number of transposases associated with IS discovered in a organisms ranging from *E. coli* and *Citrobacter* to *Ferroplasma* and *Shewanella*. A putative DDE conserved domain characteristic of a

transposase/integrase superfamily inclusive of the IS4 and IS5 families and bacteriophage λ was identified in both predicted protein sequences. In the Yellowstone sequence, this was between aa87 and aa282, producing an alignment of 99.6%, and was between aa92 and aa282 in the Kamchatka sequence for an alignment of 93.1%. The two elements were assigned to the large, heterogeneous IS5 family on the basis of both the putative conserved domain and the previous assignment of ISC1058 to this same family.

The third ISC1058 relative was only recovered from one YNP99 strain, in which it was responsible for at least three separate insertion events. As it was found to be 1058 bp in length, it has been given the designation ISC1058b to differentiate it from the *S. solfataricus* element, to which it displays 80% nucleotide identity. It displays a length and genetic organization (A single ORF occupying 85% of the element downstream of promoter and rbs elements.) similar to those of the ISC1057 isoforms, and, considering that it displays high sequence identities to both (88% to the Kamchatka variant, and 93% to the Yellowstone variant), it is questionable as to whether or not it should be considered a distinct element. However, the 299 amino acid protein predicted to be encoded by its open reading frame shows considerably lower similarity levels to both (74% identity, 81% identity to the Kamchatka variant, and 77% identity, 83% similarity to the Yellowstone variant, both over 299 aa.), was considered to justify the decision to do so. It also displayed a markedly lower level of similarity to the ISC1058 transposase (72% identity and 81% similarity over 299 aa) than did the ISC1057 predicted proteins. Further, the NCBI protein BLAST program did not identify any putative conserved domains (Despite the visually-confirmed

presence of a putative DDE motif.), unlike the two ISC1057 isoforms. The inferred phylogenetic relationship between the ISC1057 regional variants, ISC1058b, ISC1058, and other members of the IS5 family are shown in the phylogram in figure 3.3 based on predicted transposase sequences.

The primers used in ISC1057-specific PCR screens were designed from the sequence of the Yellowstone variant. As both ISC1057 variants and ISC1058b possess nearly identicalend sequences, these primers were capable of amplifying all three. In the course of screening, amplification was observed from strains of all geographic regions except Italy (figure 3.6). The frequency of amplification was highest (~83%) amongst the Yellowstone strains. Only three New Zealand strains were found positive, but roughly half of the strains from Lassen and Kamchatka did. Considering this, it is surprising that it was never recovered from any Lassen strain. It is odd that no amplification was observed amongst the Italian strains, considering that the strains recovered from the Naples hot springs were found to be identical to *S. solfataricus* P2, the genome sequence of which showed a high copy number (14) of ISC1058 (She et al. 2001, Brugger et al. 2002). It is possible that the differences noted between terminal sequences of the elements recovered from the *S. islandicus* strains and ISC1058 was sufficient to interfere in the annealing of the primers and thus prevent successful amplification. Though it was not attempted, it is possible that this problem could be overcome by lowering the annealing temperature of the PCR thermocycle.



<u>Figure 3.3</u>: Phylogram with Bootstrap Values of the Relationship Between Recovered and Previously Identified Members of the IS5 Family Based on Predicted Transposase Sequences

ISC1205

ISC1205 is perhaps the most remarkable IS discovered in the course of this project. Among the *S. islandicus* strains examined, it was recovered more often than all other IS types combined. Indeed, it was responsible for all but one of the IS-mediated mutations observed amongst the Lassen strains examined, and approximately half of those in the Kamchatka strains. In addition, it ranks as among the most novel IS recovered. BLAST searches with the nucleotide sequence of the type sequence recovered from L2K strain 24, revealed significant scores to only five segments of 40 to 268 bp in the *S. tokodaii* genome with identities of 84% to 93%, and three segments of 84 to 348 bp spread over some 3 kb of the *S. solfataricus* genome that displayed identities of 81% to 85%. This is taken as evidence of the fragmentation of at least one copy of an isoform in each organism at some point in the past.

A number of examples ISC1205 isoforms were sequenced from among those recovered from both Lassen and Kamchatka strains. The analysis of these, summarized in table 3.3, revealed a great degree of consistency in characteristic features amidst minor variation. They display some minor differences in element and IR length (1204 to 1211 bp and 17 to 20 bp, respectively). There is also little discernable pattern in either the length or sequence of the direct repeats, though some sequence requirement may be assumed, as two sequenced examples from Lassen (24a and 24b in table 3.3) and at least two partially sequenced examples from Kamchatka and Lassen (not shown) had independently inserted into the same location. If target site selection were totally unguided, this would not be expected. All, however, were

Strain of Origin	Length (bp)	G+C	DR Sequence	IR (Sequence)	Predicted ORF Data			
					Position (bp)	Protein Length	pI	MW (kDa)
L2K 24a (Type)	1205	46%	TATTA	TAGCAGTTGTTCTACTTT TAGCAGTTGTCCTACTTT	89 - 1189	366	9.5	41.9
L2K 24b	1205	46%	TATTA	TAGCAGTTGTTCTACTTT TAGCAGTTGTCCTACTTT	89 – 1189	366	9.5	41.9
L2K 28	1205	45%	CTAGATG	TAGCAGTTGTTCTACTTT TAGCAGTTGTTCTACTTT	89 – 1189	366	9.3	41.9
L2K 15	1205	45%	TAGATGA	TAGCAGTTGTTCTACTTT TAGCAGTTGTTCTACTTT	89 – 1189	366	9.5	41.9
L2K 22	1211	45%	?	TAGCAGTTGTTCTACTTT TAG-AGTTGTTCTACTTT	90 – 1184	364	9.9	41.5
L2K 19	1204	45%	ATTATT	-AGCAGTTGTTCTACTTT TAGCAGTTGTTCTACTTT	88 –1188	366	9.3	41.9
L2K 14	1205	45%	ACAAAG	TAGCAGTTGTTCTACTTT TAGCAGTTGTTCTACTTT	89 – 1189	366	9.3	41.9
K2K 12-8	1205	46%	ATCCTAA	TAGCAGTTGTTCTACTTT TAGCAGTTGTTCTACTTT	89 - 1189	366	9.2	41.8
K2K 12-1	1206	46%	TAAT	---AGCAGTTGTTCTACTTT AGCAGCAGTTGTTCTACTTT	88 – 1188	366	9.2	41.7

Table 3.3 : Data from Multiple Sequenced Examples of ISC1205 Isoforms

found to possess a single putative open reading frame spanning approximately 90% to 91% of their length, with the positioning almost precisely the same in each. All showed the presence of a potential rbs 9 bp upstream of the start codon (CTG), as well as a strong bacterial promoter, though only a weak to moderate Archaeal promoter. All the ORFs terminate in an ochre stop codon partially contained within the right inverted repeat.

Translation of the putative ORFs identified a predicted protein of predominantly 366 amino acids in length. In all cases, the theoretical isoelectric points for these were between 9.2 and 9.9, and the molecular weights were predicted to be between 41.5 and 41.9 kDa. All of these predicted proteins were found to display evidence of a Transposase 29 conserved domain. This is a recently created conserved domain that is based around the transposase sequence of ISC1217. Protein

BLAST searches revealed a marked dearth of closely related proteins in the database. A hypothetical 131 aa protein, Sso0132, the code for which is within the largest nucleotide stretch in the *S. solfataricus* genome to which the DNA sequences of the ISC1205 isoforms showed simililarity, produced the highest returned score (75% identity, 86% similarity over 131 aa), further bolstering the supposition of a fragmented isoform in this organism. Considerably lower scores were returned for the transposases of ISC1217 (32% identity, 48% similarity over 370 aa) and its isoformic relative ISSt1145 (32% identity, 53% similarity over 279 aa), as was expected from the conserved domain prediction. A number of other putative transposases and hypothetic proteins from organisms including *Ferroplasma*, *Bradyrhizobium*, and *Desulfitobacterium* produced scores of between 28% identity and 52% similarity over 100 to 171 amino acids. These results were consistent for all predicted ISC1205 protein sequences.

ISC1205 could not be placed in a recognized and defined IS family. This is also the case for ISC1217 from *S. solfataricus*, and it is likely that a new family will be defined at some point in the future that will be based around ISC1217 and ISC1205. The similarity between these *Sulfolobus* IS and putative bacterial transposases, low though they are, indicates that this family is not restricted by host domain. No putative DDE domain was detected in the alignment of the transposase sequences, indicating that the family does not belong to the DDE superfamily, and is thus of independent derivation.

Pairwise alignments of the sequenced examples revealed remarkable conservation in both nucleotide and amino acid sequences. These results are

summarized in table 3.4. The two examples recovered and sequenced from Lassen strain 24 proved to be identical in the course of this analysis, as did those found in strains 15 and 19. These two groups are thus listed together. In no comparison was a greater than 5% difference at the nucleotide level or 11% at the level of amino acid identity observed between isoforms. Interestingly, there was little variation in the differences observed between examples from the same set and those of difference sets. Indeed, the greatest differences were between the Lassen examples from strains 24a/24b and 22. These patterns are reflected in the phylogenetic tree generated from the transposase sequences of the isoforms and those elements in the database to which they showed similarity (Firgure 3.4).

		Lassen Isoforms					Kamchatka Isoforms	
		24a/24b	**28**	**22**	**15/19**	**14**	**12-8**	**12-1**
Lassen Isoforms	**24a/ 24b**	-	96%	95%	95%	96%	96%	96%
	28	89%/91%	-	99%	99%	99%	97%	96%
	22	88%/89%	93%/93%	-	99%	99%	96%	95%
	15/ 19	89%/91%	93%/93%	93%/93%	-	99%	97%	96%
	14	89%/91%	93%/93%	93%/93%	93%/93%	-	97%	96%
Kamchatka Isoforms	**12-8**	89%/90%	89%/90%	89%/90%	90%/91%	90%/91%	-	96%
	12-1	89%/90%	90%/91%	89%/90%	90%/91%	90%/91%	90%/91%	-

Table 3.4: Nucleotide Identities (Top) and Protein Identities/Similarities (Bottom) Between Distinct ISC1205 Isoforms from Recovered from Kamchatka and Lassen Strains

ISC1205 specific PCR screens revealed the element to be apparently quite widespread. Positive amplification was noted for strains from all sampled regions except Italy. Interestingly, unlike other elements, it also showed very high frequency of positive amplification among the New Zealand strains probed. It was detected in all Lassen strains probed, also, a result that indicates a prevalence that

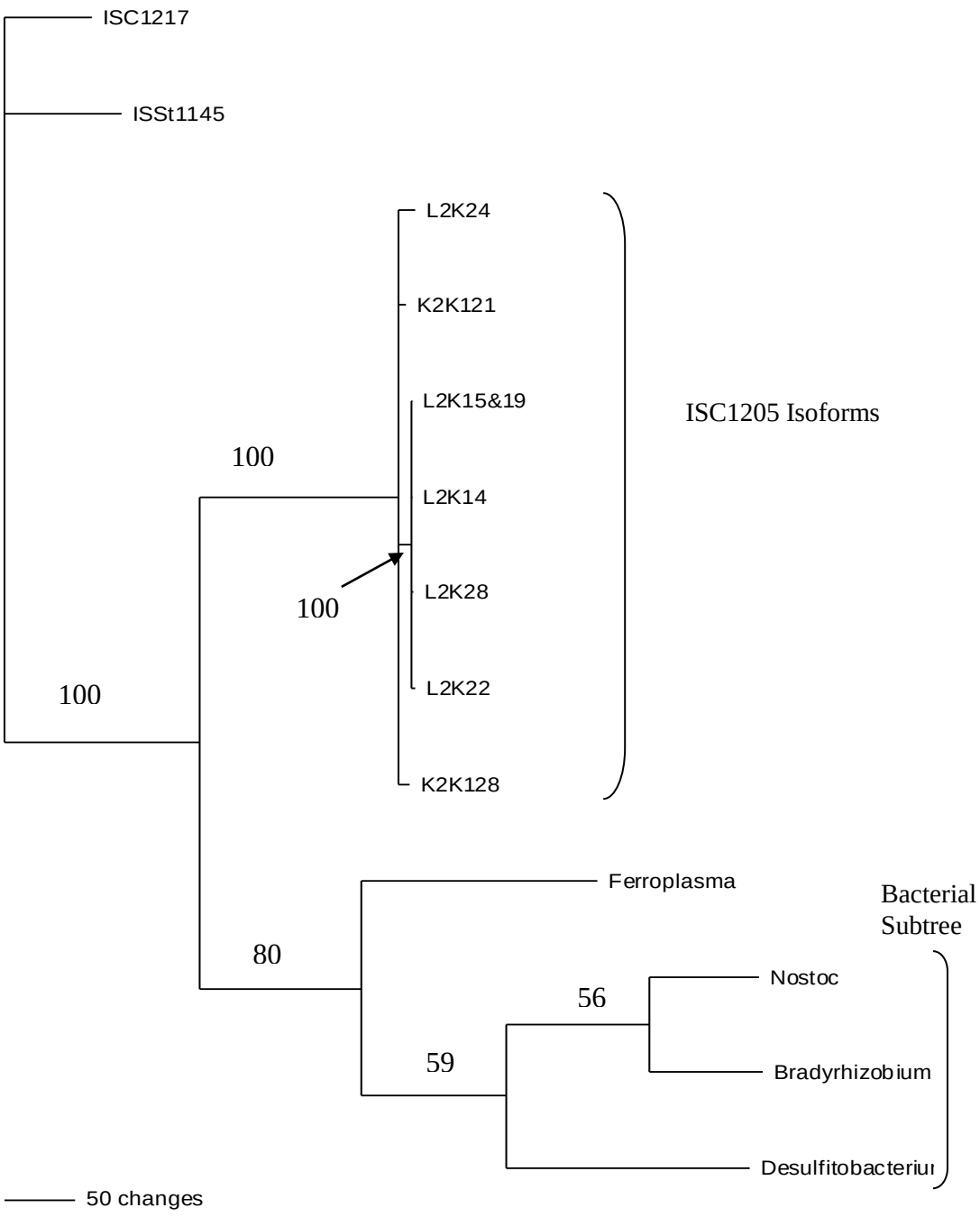


Figure 3.4: Phylogram with Bootstrap Values Based on Transposase Amino Acid Sequences of ISC1205 Isoforms and Related Elements

explains the high levels of recovery noted for it in this set. Unexpected, however, were the low frequencies of amplification noted among the Kamchatka strains. The high frequency of recovery of the element from other strains is at odds with this result. Of course, an isoform was later trapped in one of the strains showing negative amplification, and it is likely that the actual proportion of strains carrying the element are higher than is indicated by the screening results. The lack of amplification from the Italian isolates is odd considering the fragmented copy that was found in the *S. solfataricus* genome. It is possible, though, that fragmentation has occurred to all copies that once resided in the population to such a degree that detection by PCR probing is no longer possible.

ISC1205 displays the highest G+C value of any IS type recovered in this project. The 45% to 46% value observed is considerably higher than both the 41% generally observed for the other elements (table 3.2) and the 38% for *Sulfolobus* itself. This is consistent with the 50% G+C value for the ISC1217 nucleotide sequence. It is unclear as to what, if any, significance this might have. However, this difference in base pair composition, along with the possession of a strong bacterial promoter may indicate that ISC1205 and ISC1217 have only recently crossed domains from bacterial hosts of different composition.

It is also possible that the odd patterns of similarity observed between ISC1205 isoforms compared to their geographic distribution may be due to their being carried between populations by bacterial hosts. It is known that spore forming bacilli are among the bacterial populations of hot springs inhabited by *Sulfolobus*. It is

conceivable that windblown spores of cells hosting isoforms are involved in the transfer of these IS between regions on a sporadic basis. It would be interesting to examine specimens from the bacterial populations residing in the same hot springs sampled for *Sulfolobus* isolation for the presence of similar IS to test this hypothesis.

ISC1288

ISC1288 was commonly recovered from YNP99 and YNP00 strains during gene trapping. Despite this, only one example was successfully completely sequenced due to the problems encountered during sequencing, likely because of the long, 34 bp inverted repeats featured by the element. Direct repeats of 5 bp were observed, with at least two partially sequenced and independently derived examples transposing into the same location in Region 2 of *pyrE*. A single putative open reading frame spanning the element from an ATG start codon beginning at nt144 to an ochre stop codon ending at bp 185. At only 73%, this is the lowest proportion of putative coding to noncoding sequence of any type of element recovered. A putative rbs was detected 9 bp upstream of the start codon, as well as a bacterial promoter of moderate strength, though no Archaeal element of any detectable strength was noted.

The ORF was predicted to encode a protein of 313 aa. This protein was found to possess a theoretical pI of 9.98, and a molecular weight of 36.9 kDa. No conserved domains were detected. When used in a protein BLAST search, the primary sequence was found to display a high similarity score (92% identity, 93% similarity over 307 aa) with the transposase of ISC1290, an element found in the *S. solfataricus* genome. Lower scores were observed for its relationship with ISC1234 (37% identity, 55%

similarity over 269 aa) and ISSt1319 (29% identity, 48% similarity over 283 aa). Aside from these *Sulfolobus* sequences, a single other hit was noted to a 248 aa hypothetical protein in the genome of *Kluyveromyces lactis (*26% identity, 44% similarity over 118 aa). As all of the *Sulfolobus* proteins have been previously identified as belonging to the IS5 family of elements, ISC1288 was similarly assigned (Figure 3.3). A nucleotide BLAST search showed an extraordinarily high 96% identity with ISC1290. Coupled with the high amino acid similarity scores, it is likely prudent to regard ISC1288 as an isoform of ISC1290, rather than a distinctly different IS.

PCR screening with primers designed to amplify ISC1288 led to its identification of residence in strains in Lassen and Italy, as well as Yellowstone. Considering the relation to ISC1290, it was unsurprising that amplification was observed in all but two of the Italian strains probed. However, in spite of its common recovery, it was noted to be present in only 35% of the Yellowstone strains tested. No amplification at all was noted in the probing of New Zealand and Kamchatka strains, and only a low frequency of amplification was noted for the Lassen strains screened. It thus seems to be limited to North America and Asia.

ISC1926

The largest IS recovered in this project, and, indeed, the largest recorded to be found in *Sulfolobus,* is ISC1926. It was also the most rarely recovered IS. Only a single Lassen mutant was observed to have experienced transposition of this element into the target region. Aside from this, it is unique in its respect to a complete lack of inverted repeats. Direct repeats of its insertion site were also not observed. These

characteristics were the same as those for ISC1913, an element present in two complete and two partial copies in the *S. solfataricus* genome, and ISSt1924, an element present in the *S. tokodaii* genome. ISC1926 is closely related to these previously documented IS, displaying 90% and 78% identity to them, respectively, at the nucleotide level. Also like these elements, ISC1926 is predicted to possess two open reading frames. The first of these begins with an alternative ATC start codon at nt42 and extends to an amber stop codon ending at nt707. The second open reading frame is far larger, beginning with a standard ATG start codon at nt785, and stretching to nt1918, where it terminates in another amber stop codon. Together, these open reading frames compose 93% of the element, the highest proportion of coding sequence among the recovered IS. Potential ribosomal binding sites precede these putative ORFs, but no properly positioned promoter of any significant strength could be detected upstream of either. This may explain in part the low apparent rate of transposition and recovery of the element during gene trapping.

ORF I was predicted to encoded a protein of 221 amino acids. This predicted protein was found to have a theoretical isoelectric point of 9.27, and a molecular weight of 25.3 kDa. It is also clearly a resolvase that indicates a high likelihood that ISC1926 utilizes a replicative mode of transpostion. Three conserved protein domains were detected within the primary sequence of this protein. The first, COG2452.1 showed 100% alignment from aa12 to aa210, and is listed as an active DNA integrase/resolvase domain. The second, showing 92.8% alignment between aa67 and aa197, was the pfam00239 Resolvase N-terminal Domain. Finally, an AlphA Transcriptional Regulator Domain was detected between aa10 and aa67, and showing

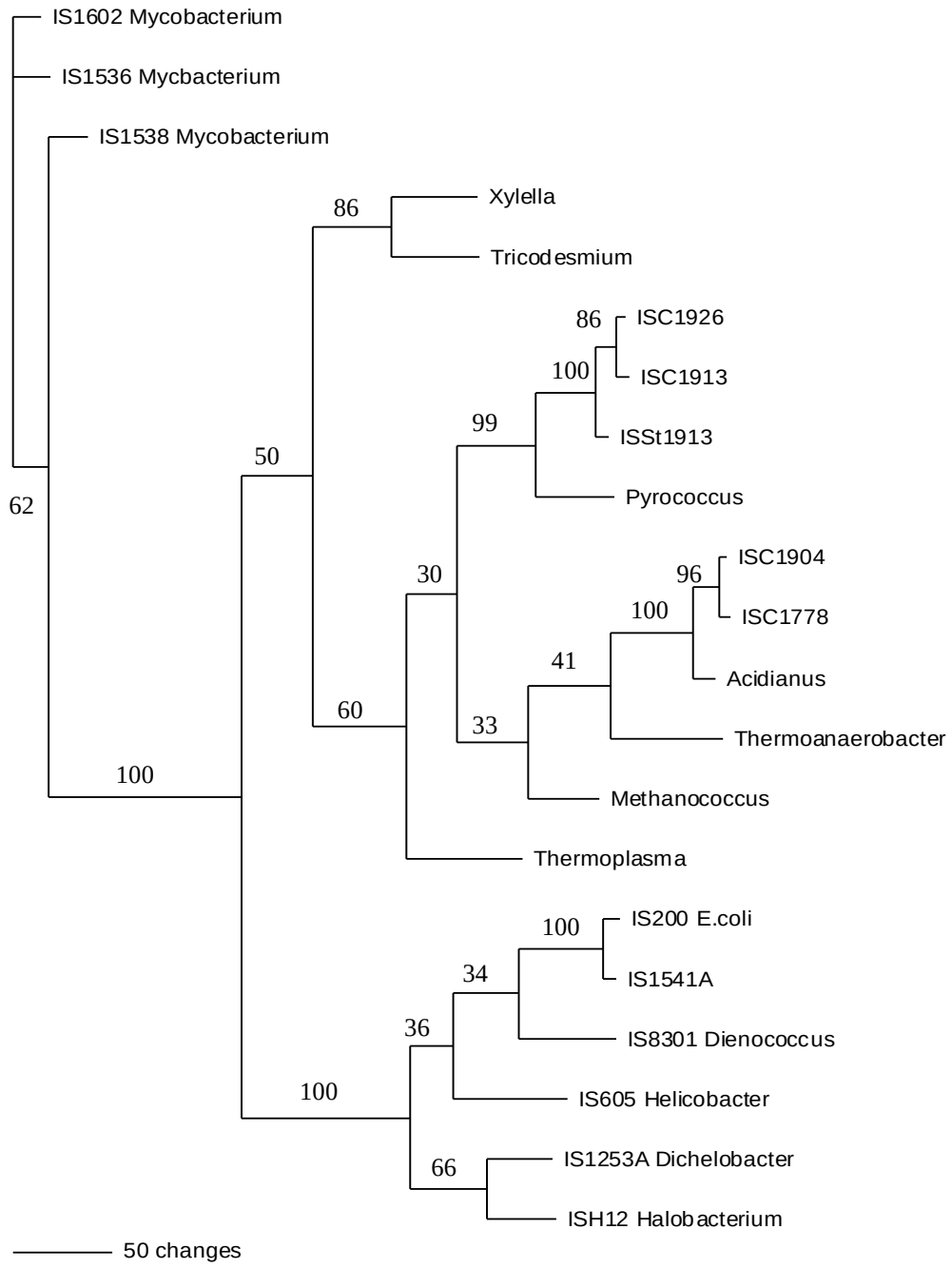
84.3% alignment. High BLAST scores were returned for a search with the primary sequence. The highest, as could be expected, were to the first ORFs of ISC1913 and ISSt1924 (94% identity, 97% similarity and 90% identity, 95% similarity, respectively with each over 213 aa). Lower scores (42% to 63% identity and 61% to 78% similarity) were observed over 190 to 214 aa comparisons between other proteins identified in *Pyrococcus*, *Methanococcus*, *Acidianus*, and *Thermoanaerobacter*, as well as that encoded by ORF I of ISC1904.

The second open reading frame appears to encode transposase of 378 amino acids. The predicted primary sequence was found to indicate a theoretical pI of 9.98, and a molecular weight of 43.9 kDa. It, too, detected a putative conserved domain, COF0675, that displayed 80.8% alignment from aa19 to aa344. This domain characterizes a number of IS discovered in both bacteria and Archaea that belong to the IS605/IS200 family of elements. This predicted protein showed a considerably lower level of conservation. Far lower BLAST scores were returned from a search with its primary sequence. As before, the highest returned scores were for ISC1913 and ISSt1924, this time to the second ORF of each (83% identity, 89% similarity and 76% identity, 86% similarity, respectively, over 376 aa). As before, lower scores were returned from comparison to a number of other proteins. These included the transposases of ISC1904 and its relative in *S. tokodaii* (24% identity, 40% similarity and 29% identity, 46% similarity, respectively over 299 aa), and a number of hypothetical proteins in *Pyrococcus*, *Methanococcus*, and *Ferroplasma* (23% to 41% identity, 39% to 58% similarity over 194 to 381 amino acids.).

On the basis of the BLAST searches as well as the relationship to ISC1913 and ISSt1924, the ISC1926 was assigned to the IS605/IS200 family. This family is made up primarily of composite elements that seem to have been generated by the recombination of two elements. Indeed, IS605, originally recovered from *Helicobacter pylori*, possess two open reading frames, one of which corresponds to a potentially autonomous copy of IS200, originally identified in Salmonella typhimurium, and the other to a transposase related to that of IS1341 (Chandler & Mahillon 2002). Not much is known of the members of this family, though the common presence of a resolvase, as seen in ISC1926 and its relatives in *Sulfolobus*, strongly indicates replicative transposition. The relationship between ISC1926 and a number of other elements, including those found in *Sulfolobus*, are depicted in the phylogram constructed on the basis of putative resolvase sequences that is shown in figure 3.5.

ISC1926-specific PCR screening showed a lower level of positive amplification among the strains of the five sampled region than any other element aside from ISC735, indicating it to have a quite limited geographic distribution. Amplification was only observed among the strains of Lassen National Park, and Italy. The screening of the Italian strains was something of a surprise, as only one of the 38 strains screened showed amplification, and ISC1913 is known to be present in the population from the *S. solfataricus* P2 genome sequence. Due to this, the screen was repeated, with the same results observed. As will be discussed below, problems were encountered in the amplification of a small, 100 bp fragment using the forward and reverse primers for the entire element, leading to the need for screening with a forward

primer annealing to its midpoint. Pair wise alignment of the nucleotide sequences of ISC1926, ISC1913, and ISSt1924 did not show variation in this area sufficient to preclude successful annealing of this primer to any of them.



<u>Figure 3.5</u>: Phylogram with Bootstrap Values of Representative Members of the IS605/IS200 Family Based on Putative Resolvase Amino Acids Sequences

IS-Specific PCR Screens

Six sets of primers were designed to specifically amplify recovered IS and their close relatives. Due to their sequence similarity, the same set of primers was capable of amplifying both ISC1057 and ISC1058b. To examine the geographic distribution of recovered IS, these primers were used to screen the chromosomal DNA of twenty to eighty-nine strains isolated from each region sampled. For the purposes of these screens, all Yellowstone strains examined were drawn from the YNP00 sample set. The results of these screens are summarized in figure 3.6.

A strain was considered to have been shown positive for the presence of an IS or its isoformic relatives only if a product band was observed of a size approximately that of the IS probed for. In general, any observed products were of the expected size, though there were two exceptions noted in the course of the screens. When probing the L2K strains for the presence of ISC1205, four strains displayed the presence of a single band either significantly larger or significantly smaller than that expected. The strong signals these variant bands represented, together with the fact that several other strains displayed similar variant bands in addition to bands of expected size led me to believe that these represented copies of ISC1205 altered by recombination events. Though there was no evidence that these bands represented active copies, the four strains in question were still counted as positive.

The other case was that of the probings for the presence of ISC1926. The initial PCR screens for this element were performed using the forward and reverse primers for the amplification of the entire element. Strangely, virtually every reaction run using these primers resulted in the appearance of a small band of approximately 100 bp, regardless

of whether or not the expected 1900 bp band was noted. Though it was speculated that this could represent the amplification of small non-autonomous mobile elements that utilize the terminal sequences of ISC1926, it is questionable that such elements could be sufficiently active in strains lacking the full element and its transposase to avoid being lost from host genomes. A second series of screens using the reverse primer used previously and a forward primer designed to amplify from the middle of the sequence (This primer was on hand anyway, as it was necessary to sequence the middle portion of the exceptionally long element) was then carried out. This eliminated the small band problem, and allowed for a more accurate assessment of the geographic distribution of the IS.

Positive amplification from at least one strain from each sampled region using a given set of IS-specific primers can be taken as definite evidence of the presence of relatives of that IS in the region. However, the meaning of a lack of observation of positive amplifications from any of the screened strains of a region is not so clear. One major problem with interpreting such results is that it is impossible to know how well all the strains in the collection represent the populations from which they are derived, much less those selected for the IS-specific screens. It is certainly possible for relatives of the IS screened for to be present in the populations sampled, but not in the strains actually examined.

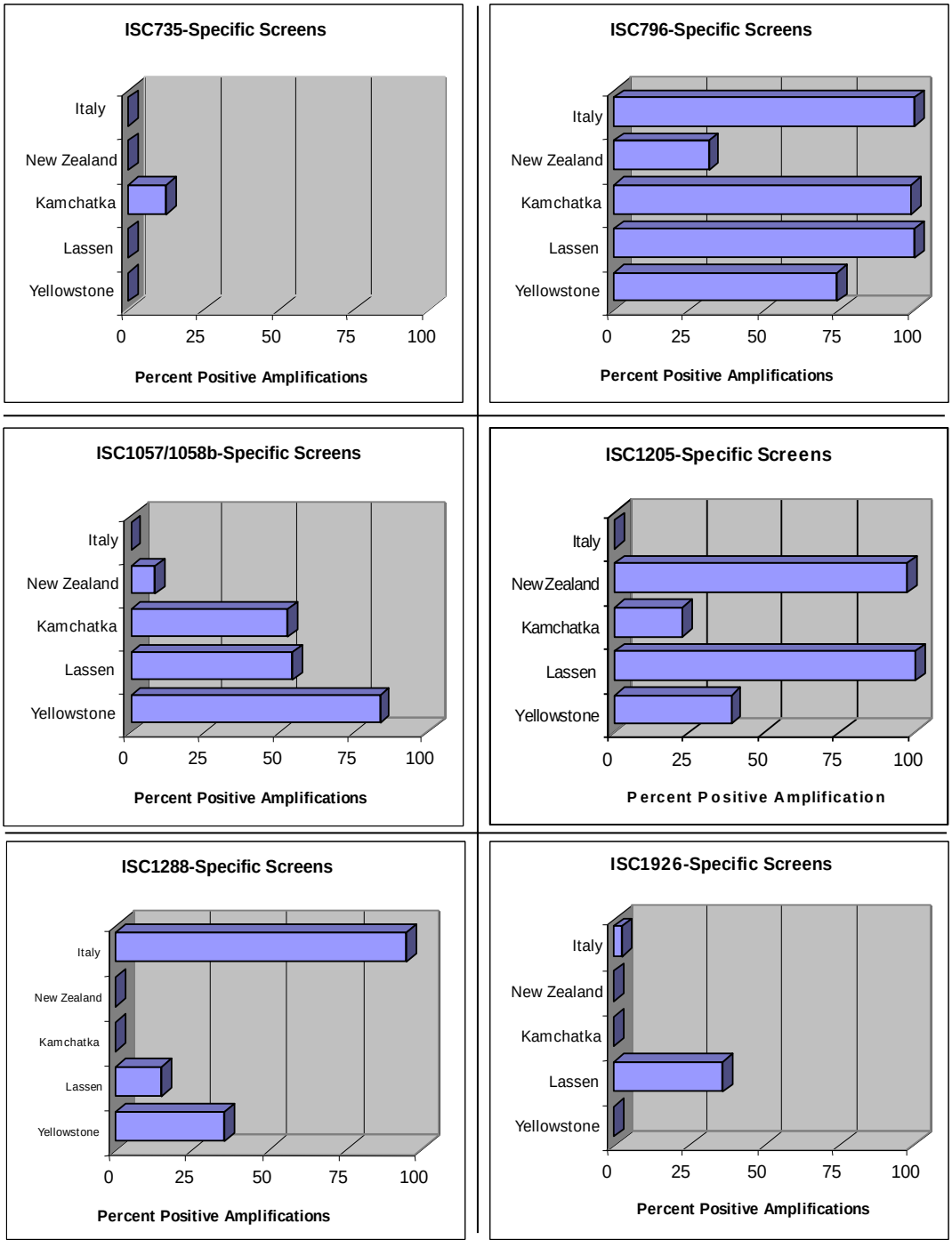


Figure 3.6: Frequency of IS-Specific PCR Screens Showing Positive Amplification from Strains of Five Sampled Geographic Regions

IV. Concluding Remarks

This project has led to the recovery of seven distinct types of insertion sequences from *Sulfolobus* strains collected from Wyoming, northern California, and the Kamchatka peninsula. Of these, four, ISC1057, ISC1058b, ISC1288, and ISC1926, showed high levels of similarity to elements previously identified in the genome of *S. solfataricus* strain P2. Indeed, ISC1288 is best regarded as an isoform of ISC1290. Two of the seven, ISC1926 and ISC796, showed high similarity to elements identified in the genomes of *S. tokodaii*. Fragments of the latter were also found in the *S. solfataricus* genome. Of the remaining two elements, ISC1205 showed only slight similarity at the amino acid level to elements identified in the two *Sulfolobus* genomes, but evidence was found in both of fragmented copies of closer relatives. Finally, ISC735 showed very little similarity to any element previously reported on an amino acid level, and none at the nucleotide level, though a possible relic of a relative was discovered in the *S. solfataricus* genome.

Great variation was displayed by these elements in terms of their geographic distribution as indicated by PCR-screening of strains derived from populations in the five regions sampled. ISC796 appeared ubiquitous, with evidence of its presence in all five regions, with ISC1057/ISC1058b and ISC1206 being indicated to be present in four of the five locations. ISC1288 showed itself to be of somewhat more restricted distribution, with amplification noted for three sets. ISC1926 seemed much more restricted in its range, being shown to be present only in the Lassen and Italian populations. Finally, ISC735 seemed to be almost provincial, with evidence of its

presence only being demonstrated in Kamchatka strains, though the relic in *S. solfataricus* indicates that it is not restricted to this region.

These findings indicate that many, widely separated *Sulfolobus* populations share a pool of closely related IS, with some elements being endemic to or restricted to only a few locations. This is consistent with the finding that almost all of the IS in *S. tokodaii* have close relatives in *S. solfataricus*. It would indicate that the IS do not respect species boundaries within the *Sulfolobus* genus, as the shared IS relatives show greater similarities than do their hosts. This may be extended to other thermophiles given the number of recovered IS with relatives in other thermophile genomes. This indicates at least some degree of horizontal gene transfer between the lineages, and, potentially, between populations. However, high-resolution phylogenies of *S. islandicus* strains from Iceland, Kamchatka, Lassen, and Yellowstone (Figure 3.7) show that there is also clear separation between them, thus indicating that any such transfer is not frequent enough to prevent the divergence of the populations (Whitaker 2001). For the most part, the phylogenies generated for those instances in which multiple examples of an IS type have been isolated from different regions support this separation, though the evidence provided by the ISC1205 tree is difficult to interpret. More examples of the different IS types will need to be sequenced and analyzed before much can be said with certainty using their respective phylogenies. However, to return to the IS-specific PCR screens, it is interesting to note that there was a distinguishable relationship between the geographic location of a population and the IS it shared with other populations. Specifically, the closer the populations spatially, the greater the number of IS types they apparently shared. Yellowstone and Lassen,

for instance, showed the presence all the same IS, save for ISC1926 that was shown to be present in the latter, but not the former. The populations of these two locations, then, shared three of the same IS with the Kamchatka and New Zealand populations. The Kamchatka and New Zealand populations also showed the presence of the

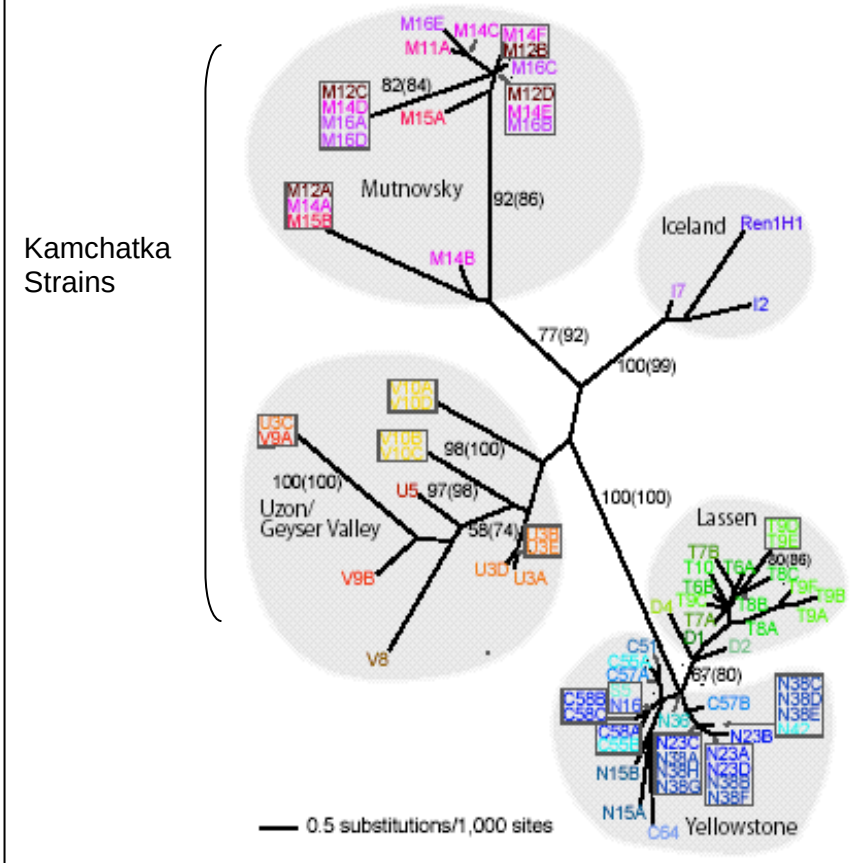


Figure 3.7: High Resolution Phylogeny of *S. islandicus* Strains from Geographically-Separated Populations

same IS, save for ISC735 that was only shown in Kamchatka. The results observed for the Italian strains are more difficult to fit into this pattern, as the PCR data can be supplemented with genomic data. The PCR screens showed the presence of only ISC796, ISC1288, and ISC1926. Looking at the genome of *S. solfataricus*, an ISC796 relative is present in fragments, while multiple copies of close relatives of both ISC1288 and ISC1926 are also present. However, a relative of ISC1057, ISC1058, is also present, as are fragmented copies of ISC1205 and ISC735. What this likely indicates is that the closer two populations are, the more frequent will be periodic events that permit the trading of IS from one to the other, and

thus providing for greater similarity in the portion of the IS "pool" that they share, while still permitting the possible maintenance of IS more or less peculiar to each. However, the frequency of such events does not proceed to zero, as the sharing of IS types between even widely separated populations indicates. Still, the frequency will decline, and this may impact the degree to which an IS is maintained in a population beyond than to which it may be considered endemic. The lower the frequency of such transfers between two populations, the more likely it is that they will begin to diverge in the constituency of their IS populations.

The above has important ramifications for gaining a better understanding of the pool of *Sulfolobus* IS. While most of the recovered IS did show close relationships to previously identified elements, it is of significance that two that did not. This is

especially evident in regard to ISC1205, which was found to be quite cosmopolitan and active. This indicates that at least one IS that likely plays a major role in the molecular evolution of a number of *Sulfolobus* populations was missed by genome sequencing, emphasizing the need for the use of gene trapping methods in screening for these elements, thus permitting examination of more and more diverse strains. Neither by gene trapping nor by PCR screening were all seven IS found in the strains from any one single region. It is thus clear that concentration of study on any one population or region will lead to the missing of potentially important and interesting elements that might prove to be of great utility in answering questions of biogeography and molecular genetics asked of *Sulfolobus* and its populations.

V. Recommendations for Future Study

There are a number of avenues of future study that should be undertaken to extend the findings of this project. Certainly gene trapping runs to recover IS should continue to be performed on new strains brought into the collection, especially when they are being derived from populations in previously unsampled locations. These should be done not only with the trap used by this project, but also with others in order to maximize recovery of diverse elements. Work should also be done to overcome the difficulties experienced in examining the gene trap targets of the New Zealand strains. To test for the possibility of a shared pool of active IS, thermophilic bacteria from the same hot springs should also be isolated and examined in this manner. This could contribute to evaluating the hypothesis that horizontal gene transfer is a pervasive phenomenon among hypthermophiles (Aravind et al. 1998). New sets of isolates should also be PCR screened for the presence of previously identified elements, no

only to further determine the distribution of the IS, but also to identify specific isolates for use in gene trapping in order to possibly trap isoforms for biophylogeographic study. Further, primers should be designed for specific amplification of IS identified in the sequenced genomes with which to screen current and new sets for their presence in order to better understand the extent to which IS are shared between *Sulfolobus* populations, and the effect of geographic separation upon this.

A major area of future focus should also be the obtaining of information on the intragenomic positions and copy numbers of characterized IS using DNA-DNA bybridization techniques. This will not only provide a means of strain typing, but also of generating high-resolution phylogenies that would be of great value in study of gene flow between populations. The value of such information to the future use of *Sulfolobus* as a model in biogeographic study should be clear. It would also provide a basis for investigation into the operational characteristics of *Sulfolobus* IS. Work such as this is greatly needed due to the lack of substantial information regarding transposition in thermophiles.

VI. References

1. Altschul, S., Gish, W., Miller, W., Meyers, E., & Lipman D. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215, 403 – 410.

2. Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25, 3389 – 3402.

   (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi)

3. Aravind, L., Tatusov, R., Wold, Y., Walker, D., & Koonin, E. (1998). Evidence for massic gene exchange between Archaeal and bacterial hyperthermophiles. Trends in Genetics, 14, 442 – 444.

4. Chandler, M. & Mahillon, J. (2002). Insertion Sequences Revisited, In Craig, N., Craigie, R., Gellert, M., & Lamhowitz, A. (Ed.). *Mobile DNA II* (pp. 305 – 366). Washington, D.C.: American Society for Microbiology Press.

5. Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., & Bairoch A. (2003).

   ExPASy: the proteomics server for in-depth protein knowledge and analysis Nucleic Acids Research, 31, 3784 - 3788. (http://us.expasy.org/)

6. Martusewitsch, E., Sensen, C., & Schleper, C. (2000). High spontaneous mutation rate in the hyperthermophilic archaeon *Sulfolobus solfataricus* is mediated by transposable elements. Journal of Bacteriology, 182, 2574 – 2581.

7. Ohta, S. Tsuchida, K., Choi, S., Sekine, Y., Shiga, Y., & Ohtsubo, E. (2002). Presence of a characteristics D-D-E motif in IS1 transposase. Journal of Bacteriology, 184, 6146 – 6154.

8. Reese, M.G. (2001). Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Computational Chemistry, 26, 21 - 56.

   Web tool at: http://www.fruitfly.org/seq_tools/promoter.html

9. Rozen, S., & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S. & Misener, S. (eds.). Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.
   Web Tool At: www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi

10. She, Q., Singh, R., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M., Chan-Weiher, C., Clausen, I., Curtis, B., Moors, A., Erauso, G., Fletcher, C., Gordon, P., Heikamp-de Jong, I., Jeffries, A., Kozera, K., Medina, N., Peng, X., Thi-Ngoc, H., Redder, P., Schenk, M., Theriault, C., Tolstrup, N., Charlebois, R., Doolittle, W., Duguet, M., Gaasterland, T., Garrett, R., Ragan, M., Sensen, C., & Van der Oost, J. (2001). Complete genome of the crenarchaeote *Sulfolobus solfataricus* P2. Proceedings of the National Academy of Sciences (USA), 98, 7835-7840.

11. Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. BioTechniques 28: 1102-1104. (http://www.ualberta.ca/~stothard/javascript/)

12. Tatusova, T. & Madden, T. (1999). Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. FEMS Microbiology Letters, 174, 247 – 250. (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html)

13. Tatusova, T. & Tatusov, R. (2003). ORF Finder. National Center for Bioinformatics. (http://www.ncbi.nih.gov/gorf/gorf.html)

14. Vincze, T., Posfai, J., & Roberts, R. (2003). NEBcutter: a program to cleave DNA with restriction enzymes. Nucleic Acids Research, 31, 3688 – 3691. (http://tools.neb.com/NEBcutter)

15. Whitaker, R. Personal communication.

Appendix

IS Sequences

ISC735

Nucleotide Sequence (Predicted ORF in bold):

```
   1 TAGGGTGTCC AACATTTTTT GTTTTTTGTA TAAAACATTA GAAGGGCGGG TGAGGTTGTA
  61 TTATTGATAG GGAATGATTA GTAAGACTAA GTTAAGGCCC CTCCCCGCCC TAGTCCTAAG
 121 TTCTGTTTTT TCAGTTTTAA ACAATTTTCC ACGCTTTCCC GGTAGTGAAA GTAATGCTAG
 181 AACTTTGTCA GCTTATTACT TAGGTCTTTC TTTGAGGAGG GTTGAAGACC TGTTTCACGT
 241 CCCTAAAAAT ACTGTCCAGT ATTATTGGAG GAAATTAGCC AATTACTTTT CTCCCCCCTC
 301 GTGTAGCGGT CATTATGCAG TTGACGAGAC CAAGATTAGA GTGGTTAACG GTTTACTTTA
 361 CTGGTTGTGG GTCGTGAGGG ATTTAAACAC GGGGAAAGTG ATTGCAGTTA GGCTTTCAAA
 421 AACTAGGAGC GGATTAGATG TAATACTACT ATTTAAAGGC AAAAGGATTA GGCTGGAGAA
 481 GACCATTAAC ATTCTCCCCG ACGGTGGGCC TTGGTATAAT AAACTTTCAA CACTCGGTGT
 541 TAAACACGAG CACGTGACTT TTGGCAAGAA AAACCTAGTT GAACAAGTGG TCAGAAGTTT
 601 AAAGTTAAGA CTTGCAAATA TGGATAAGCA CTTTCCGCCC AACGCCAGTA AGGGTTCAAT
 661 AATTAGGTGG GGTAAGGCGT TTTTCACCTT ATTCAACCTT TTTCAACAAG GTGAATGAAA
 721 ATGTTGGACA CCCTC
```

ORF Predicted Amino Acid Sequence:

```
MISKTKLRPL PALVLSSVFS VLNNFPRFPG SESNARTLSA YYLGLSLRRV
EDLFHVPKNT VQYYWRKLAN YFSPPSCSGH YAVDETKIRV VNGLLYWLWV
VRDLNTGKVI AVRLSKTRSG LDVILLFKGK RIRLEKTINI LPDGGPWYNK
LSTLGVKHEH VTFGKKNLVE QVVRSLKLRL ANMDKHFPPN ASKGSIIRWG
KAFFTLFNLF QQGE
```

ISC796

Origin: YNP01 90'-18

Nucleotide Sequence (Predicted ORF in bold):

```
   1 GGTAGTGTCG TCGTTAAGTT ATTTATATTT GTATAACGAG TATTACTTAT GGGTAGGAAG
  61 CCTGTATTTA GGCAAGACGT TTCTTGTCCC TCTTGTGGTA GTCATCATGT TGTTAAGTGT
 121 GGTAGGCCTT TGGGTAGGCA GAAGTTTTTG TGTAGGGATT GTGGTAAGTA CTTCTTGGGT
 181 GATGCTAGTT ATCATCATCA TTCTAGGAAG TTGAGGGAGG AGGCTTTGAG AATGTATGCT
 241 AATGGTATGA GTATGAGGGC TATTTCTAGG GTGCTTAACG TACCTCTTGG TACTGTTTTC
 301 ACTTGGATTA AGCGTTATGG TAGGAAAAAG CATGAGAAGT TGGTTGAGTT GTGGGGTAGG
 361 GCTAAGGAGC TGGTCAAGGG TAAGGTTGTT GCTAAGGTTG TTGATGAGAT GTGGACTTAC
 421 TTGTACAAGA ATGCTAGGGC TTTTTACAAG TGGGTTTTCA CTTGTTACGT GTCCACGAAG
 481 CTGGGAGTTT ACCTCATTTA CTCTGTGGGG GATAGGGATG AGAGTACTTT CCTTGAGGTC
 541 AAAAAGTATT TGCCTGACGA GGGTAGATGG GTGAGCGATG ATTATAACTT GTACTTCTGG
 601 TTGAAAGACC ACACGGTTGT CTCGCCAGTT AACCCGAACG AGTCCTTTCA TTCCTCATTA
 661 AGGGATAGGC TAATTAGATT CAAGAGAGCA ACGAAGGCAG TAAATAGGAG CATTCGCACC
 721 ATGATGTACT CCATAGCCCT AGTCTTATGG GAGAGAAGGT TAATCCCAGA ATTTGTAGCT
 781 TAACGACGAC ACTATC
```

ORF Predicted Amino Acid Sequence:

```
MGRKPVFRQD VSCPSCGSHH VVKCGRPLGR QKFLCRDCGK YFLGDASYHH
HSRKLREEAL RMYANGMSMR AISRVLNVPL GTVFTWIKRY GRKKHEKLVE
LWGRAKELVK GKVVAKVVDE MWTYLYKNAR AFYKWVFTCY VSTKLGVYLI
YSVGDRDEST FLEVKKYLPD EGRWVSDDYN LYFWLKDHTV VSPVNPNESF
HSSLRDRLIR FKRATKAVNR SIRTMMYSIA LVLWERRLIP EFVA
```

ISC1057, Yellowstone Variant

Origin: YNP00 58-82

Nucleotide Sequence (Predicted ORF in bold):

```
   1 TGCTTTGTGG GCTACTTCAA AATTATACAA AAATTTATAA ATAATTATGC CGGAGTATTA
  61 CCCATGGGAA AGAGTAAGTA CAAGAGGGAT TGGCACAAGT ACGACGAGAA CGTTATAACG
 121 AGATATACCC TAATGTTCCC CTTCTACGTC TTCGAACACT GGTGGGATTT ACTAGCAGAG
 181 GAGAATAGGC ATGCCAAGAA AACCTACAAG GCACCAAAGG AGTTCAACGA ATTCCTAGCA
 241 TTCCTCCACA TCTTCCTACC TTATAGGGCC ATAGAAGGAG TATTGAGAGC ACTAGAGAGA
 301 CTGAAAATCA TCCCAACAAG CCTAGACTAC TCAACAATAT GGGAAAGAGT AAGAAACATG
 361 AACATAAAAT TCCCAGAGGC AAATGACCAA CTTGAAGTAA TAGCAGACGC AACGGGAATA
 421 AGCACAAACA AGGGAGGACA ATACATTATA GCAAAATGGG GAAAAACCAA GGACTCAAAA
 481 TTCCTCAAGA TCGAAATAGT AATGGATAAG GACCAATTCA ACGTAATAAA CGCTGAAGTA
 541 ACCAGCAACG AGGTTCAGAC TGCAGTTAAG ACGGTTAAGG ATTTACAAGA TAAGGGAAAG
 601 AAGGTCAAGA AGTTTTATGG AGATAAAGCT TATGATGCTA ATGAGGTTTA CAAGACTGGG
 661 GTTGAGGTTG TTGTCCCACC TAGGAAGAAC GCTTCTACTA GGCGTGGCCA TCCTGCTAGG
 721 AGAAAGGCTG TAAGGGAGTT CAAGAGGTTG GGTTATAATC GTTGGAGGGA GGAGAAGGGT
 781 TATGGTGTTA GGTGGAGGAT TGAGTCCTTA TTCTCTGCTG TGAAGCGTAC TTTTGGGGAA
 841 TCTGTTAGGG CTACAAGTTT TTTAGGACAA GTGGTTGAGG CTAAGCTCAA GTTCTGGGCT
 901 TACGCATGGA TGGTCCACTT GGCTAATTCT GTAGTCGGTA GAGCTCCGGG TATTAGGGTG
 961 TGAGCTTGCG AATAACGTTG AAATAAATAT TAATTACTGA AAAATTCTCA GTGTATCATA
1021 TCATGGTTAT GAAATAAATT GAAGAGATCA ACAAAGC
```

ORF Predicted Amino Acid Sequence:

```
MGKSKYKRDW HKYDENVITR YTLMFPFYVF EHWWDLLAEE NRHAKKTYKA
PKEFNEFLAF LHIFLPYRAI EGVLRALERL KIIPTSLDYS TIWERVRNMN
IKFPEANDQL EVIADATGIS TNKGGQYIIA KWGKTKDSKF LKIEIVMDKD
QFNVINAEVT SNEVQTAVKT VKDLQDKGKK VKKFYGDKAY DANEVYKTGV
EVVVPPRKNA STRRGHPARR KAVREFKRLG YNRWREEKGY GVRWRIESLF
SAVKRTFGES VRATSFLGQV VEAKLKFWAY AWMVHLANSV VGRAPGIRV
```

ISC1057, Kamchatka Variant

Origin: Kamchatka 3-19

Nucleotide Sequence (Predicted ORF in bold):

```
   1 TGCTTTGTGG GCTACTTCAA AATTATACAA AAATTTATAA CTAATTATAT CGAGTATTAC
  61 CCATGGGAAA GAGTAAGTAC AAGAGGGATT GGAGCAAATA CGATGAGAAC GTTATAACTA
 121 GATATACCCT AATGTTCCCC TTCTATGTCT TCGAACACTG GTGGGATTTA CCAGCAAAGG
 181 AGAATAGGAA CGCCAAGAAA ACCTACAAGG CACCGAAAGA GTTCAACGAA TTCCTAGCAT
 241 TCCTCCACAT CTTCCTACCC TATAGGGCAA TAGAAGGAGT ATTAAGAGCA CTAGAAAGAC
 301 TGAAAATCTT CCCAACAAGC CTCGATTATT CAACAATATG GGAAAGAGTA AGAAACATGA
 361 ACATAACCTT CCCGGAGGCA AGTGATGAAC TTGAAGTAAT AGCAGACGCA ACGGGAATAA
 421 GCACAAACAA GGGAGGACAA TACATCATAG CAAAATGGGG TAAAACTAGA GACTCAAAAT
 481 TCCTCAAGAT CGAAATAGTA ATGGACAAGG ACGAATTCAA CGTAATAAAC GCTGAAGTAA
 541 CTAGCAACGA GGTTGAGACT GCAGTTAAGA CGGTTAAGGA TTTACAAGAT AAGGGTAAGA
 601 AGGTCAAGAA GTTTTATGGG GATAAGGCTT ATGATGCCAA TGAGGTTTAC AAGACCGGGG
 661 TTGAGGTTGT TGTCCCACCT AGGAAGAACG CTTCTACTAA ACGTGGTCAT CCTGCTAGGA
 721 GGAAGGTTGT GAGGGAGTTC AAGAAACTTG GCTATAATCG TTGGAGGGAG GAGAGGGGTT
 781 ATGGTGTTAG GTGGAGGGTT GAGTCCTTGT TCTCTGCTGT GAAGCGTACT TTTGGGGAGT
 841 CTGTTAGGGC TACAAGTTTT TTAGGGCAAG TGGTTGAGGC TAAGCTCAAG TTCTGGGCTT
 901 ATGCATGGAT GGTCCACTTG GCTAATTCTG TAGTCGGTAG GGCTCCGGGT ATTAGGGTGT
 961 GAGCTTGAGA ATAACGTTGA AATAAATATT AATTACTGAA AAATTCTCAA TATATTATCT
1021 CATACTTATG AAATAAATTG AAGAGATCAT ACAAAGC
```

ORF Predicted Amino Acid Sequence:

```
MGKSKYKRDW SKYDENVITR YTLMFPFYVF EHWWDLPAKE NRNAKKTYKA
PKEFNEFLAF LHIFLPYRAI EGVLRALERL KIFPTSLDYS TIWERVRNMN
ITFPEASDEL EVIADATGIS TNKGGQYIIA KWGKTRDSKF LKIEIVMDKD
EFNVINAEVT SNEVETAVKT VKDLQDKGKK VKKFYGDKAY DANEVYKTGV
EVVVPPRKNA STKRGHPARR KVVREFKKLG YNRWREERGY GVRWRVESLF
SAVKRTFGES VRATSFLGQV VEAKLKFWAY AWMVHLANSV VGRAPGIRV
```

ISC1058b

Origin: YNP99 9-16

Nucleotide Sequence (Putative ORF in bold):

```
   1 TGCTTTGTGG GCTACTTAAA AATTATACAA AAATTTATAA ATAATTATAC CGAGTATTAC
  61 CCATGGAAAA GAGTAAGTAC AAGAGGGATT GGAACAAGTA CGACAAGAAC GTCATAACGA
 121 GATATACCCT AATGTTCCCC TTCTACGTCT TCAAACACTG GTGGGATTTA CTAGCAAAGA
 181 AAAATAAGCA TGCCAAAAAA ACCTACAAGG CACCAAAGGA GTTCAACGAA TTCCTAGCAT
 241 TCCTCCACAT CTTCCTACCT TATAGGGCCA TAGAAGGAGT ATTGAGAGCA CTAAAAAGAC
 301 TGAAAATCAT CCCAACAAGC CTAGACTACT CAACATTATG GGAAAGAGTA AAAAACATGA
 361 TCATAAAATT CCCAGAGGCA AATGACCAAC TTGAAGTAAT AGCAGACGCA ACGGGAATAA
 421 CCACAAACAG GGGAGGACAA TACATTATAC CAAAATGGGG TAAAACCAAG GACTCAAAAT
 481 TCCTCAAGAT TGAAATAGTA ATGGATAAGG ACCCATTCGA CGTAATAAAC GGTGAAGTAA
 541 CCAGCAACGA GGTTGAGTCT GCAGTTAAGT CAGTTAAGGA TTTTCAAGAT AAGGGAAAGA
 601 AGGTCAAGAA GTTTTATGGA GATAAAGCTT ATTATGCTAA TGAGGTTTAC AAGACTGGGG
 661 TTGAGGTTGT TGTCCCCCCT AGGAAGAACG TTTTTACTAG GCGTGGCCAT CCTGCTAGGA
 721 GAAAGGCTGT AAAGGAGTTC AAGAGGGTGG GTTATAATCG TTGGAAGGAG GAAAAGGGTT
 781 ATGGTGTTAG GTGGAGGAAT GAGTCCTTAT TTTTTTCTGT TAAGCGTACT TTTGGGGAAT
 841 TTGTTAGGGC TTCAAGTTTT TTAGGACAAG TGGTTAAGGC TAAGCTTAAG TTTTGGGTTT
 901 ATGCATGGAT GGGCCCCTTG GATAATTTTG TAGTCGGTAG AGCTCCGGGT ATTAGGGTGT
 961 GAAGCTTTCG AATACCGTTG AAATTAATTT TAATTACTGA AAAATTTTCA ATGTATCATA
1021 TCATGCTTAT GAAATTAATT GAAGAGATCA ACAAAGCA
```

ORF Predicted Amino Acid Sequence:

```
MEKSKYKRDW NKYDKNVITR YTLMFPFYVF KHWWDLLAKK NKHAKKTYKA
PKEFNEFLAF LHIFLPYRAI EGVLRALKRL KIIPTSLDYS TLWERVKNMI
IKFPEANDQL EVIADATGIT TNRGGQYIIP KWGKTKDSKF LKIEIVMDKD
PFDVINGEVT SNEVESAVKS VKDFQDKGKK VKKFYGDKAY YANEVYKTGV
EVVVPPRKNV FTRRGHPARR KAVKEFKRVG YNRWKEEKGY GVRWRNESLF
FSVKRTFGEF VRASSFLGQV VKAKLKFWVY AWMGPLDNFV VGRAPGIRV
```

ISC1205, Lassen Variant 1 (Type)

Origin: Lassen Strain 24

Nucleotide Sequence (Putative ORF in bold):

```
   1 TAGCAGTTGT TCTACTTTTC CTCGGTAATA CAAAGCTCAC AGTACTACCG ATAAAGTTTT
  61 TAGATTTACG TTTAGTTGAT TAAAAACTGT GAAGTCCGAG AAAGAGCTGG ACATTGCAAG
 121 GACCGAGTTC ATCAAGTCCT TCAACTACCT GATTGGAACA CTGAGGATGA ACGGATTGAG
 181 AAGAAAGGTC GCAGTAGGCT TAGCACTCAT GACCTTGATA GGAGGAAGGG CCAGCATTAG
 241 AAACGCATCC ATCACGTTCA AGCTAAACTA CGCCAACTTG TTAAAAACAC TGGAGAACCT
 301 AGAGAACACA TGGAGAGACT ACCTTGAAGC GTTAAGCAAA GTGATAATTG GACCGGTAGT
 361 AGTGATAATT GACGACACCT TCGACCACAA ACTCTATTCC AGAGTAGAGG GCATAGCAAG
 421 CAAGTACGGG AACTACTTCG CGTGGTGCTC CACGCACAAG AGATTCGAGC CCGGCATACA
 481 AGTCCTCACA ATAGCCTTAT ACGACTTAGC CATGGGAAAG AGCTATTTGA TAGGAGCTTT
 541 CCCATACGCC ACGAGAAAGA TGTGGGAAAG CGGGATGGTA AGTGAGTTCA AGACCAAGAT
 601 CGAGATGGCT GCGGAAATTA TCGAGATCCT CAAAGAGCGG TTCCATGTGG CGAGGGTAGT
 661 GTTTGACTCC TGGTATTGGT CGGAGAAGCT GGTGAAGGGT AGTGTAGTTT CGGAATTGAA
 721 GTCTAATAGG AGGCTCATCA GGGTTAGGCC TTTGGAGGGA GGGAAGACGT TGGGGGTGGA
 781 GGGGCACCCC CATGTCGGAG ATCTCCCTCC AGGGTCTTAC TTAGCTGAGT TGACCCTAGG
 841 AGATCAAGTT ATTACTATAA AGTTGTTAAT ACTGGTATAT AAAGATAACA GGCTCAATTT
 901 GTACACTACT GACCTTAACT TGAGCGATGA GGAGATAGAG GCAACTTGGA AGATTAGGTG
 961 GGAGATAGAG AAGTTACACA AGGATATTAA GGCTCTGGGT ATGCAAGATT CCTCTTTCCT
1021 CAAGAGGAAG AGGCTTCAAG GTTATTTACT CCTCTTCGTG ATGGTGGTCA ACGCGGCCAG
1081 AGACTTGGTC ACGTCCCTTA ACTTGAAGAG CGTGGAGGAA CTTCGCCGGT TCGTTGAAAT
1141 ACGTTTAGGA GGTGCTCTGG GTTTGATGAA AATCTTTAAG CTGCGTTAAA GTAGGACAAC
1201 TGCTA
```

ORF Predicted Amino Acid Sequence:

```
VKSEKELDIA RTEFIKSFNY LIGTLRMNGL RRKVAVGLAL MTLIGGRASI
RNASITFKLN YANLLKTLEN LENTWRDYLE ALSKVIIGPV VVIIDDTFDH
KLYSRVEGIA SKYGNYFAWC STHKRFEPGI QVLTIALYDL AMGKSYLIGA
FPYATRKMWE SGMVSEFKTK IEMAAEIIEI LKERFHVARV VFDSWYWSEK
LVKGSVVSEL KSNRRLIRVR PLEGGKTLGV EGHPHVGDLP PGSYLAELTL
GDQVITIKLL ILVYKDNRLN LYTTDLNLSD EEIEATWKIR WEIEKLHKDI
KALGMQDSSF LKRKRLQGYL LLFVMVVNAA RDLVTSLNLK SVEELRRFVE
IRLGGALGLM KIFKLR
```

ISC1205, Lassen Variant 2

Origin: Lassen Strain 28

Nucleotide Sequence (Putative ORF in bold):

```
   1 TAGCAGTTGT TCTACTTTTC CTCGGTAATA CAAAGTTCAC TGTATTATCG ATAAAGTTTT
  61 TAGATTTACG TTTAGTTGAT TAAAAACTGT GAAGTCCGAG AAAGAGCTGG ACATTGCAAG
 121 GACCGAGTTC ATTAAGTCCT TCAATTTCGT AGTCGGGACA CTGAGGATGA ACGGATTGAG
 181 CAGAAAGGTC GCAGTAGGCT TAGCACTCAT GACCTTGATA GGAGGAAGGG CCAGCATTAG
 241 AAACGCATCC ATCACGTTCA AGCTAAACTA CGCCAACTTG TTAAAAACAC TGGAGAACCT
 301 AGAGAACACA TGGAGTGACT ACCTCGAAGC GTTAAGCAAA GTGATAGTTG GACCGGTAGT
 361 AGTGATAATT GACGACACCT TCGACCACAA ACTCTATTCC AGAGTAGAGG GCATAGCAAG
 421 CAAGTACGGG AACTACTTCG CGTGGTGCTC CACACACAAG AGATTCGAGC CCGGCATACA
 481 AGTCCTCACA ATAGCCTTAT ACGACTTAGC CATGGGAAAG AGCTATCTGA TAGGAGCTTT
 541 CCCATACGCC ACGAGAAAGA TGTGGGAAAG CGGGATGGTA AGTGAGTTCA AGACCAAGAT
 601 CGAGATGGCT GCGGAAATTA TCGAGGTTCT CAAAGAGCGG TTCCATGTAG TGAGGGTAGT
 661 GTTTGACTCC TGGTATTGGT CGGAGAAGCT TGTGAGGGAT AGTGTAGTTT CTGAGTTGAA
 721 GTCCAACAGG AGGCTTCTAA GGGTTAGGCC TTTGGAGGGA GAGAAGACGT TGGGGGTGGA
 781 GGGGCACCCC CATGTCGGAG ATCTCCCTCC AGGGTCTTAC TTAGCTGAGC TGACCCTAGG
 841 AGACCAAGTT ATTACTATAA AGTTGTTAAT ACTGGTATAT AAAGATAACA GGCTCAATTT
 901 GTACACTACT GACCTTAACT TGAGCGATGA GGAGATAGAG GCAACTTGGA AGATTAGGTG
 961 GGAGATAGAG AAGTTTCACA AGGATATTAA GGCTCTGGGT ATGCAAGATT CCTCTTTCCT
1021 CAAGAGGAAG AGGCTTCAAG GTTATCTGCT CCTCTTCGTG ATGGTGGTTA ACACGGTCAG
1081 AGATTTGATC AGCTCCCTTA ACTTGAAGAG CGTGGAGGAA CTTCTCCGGT TCGTTGAAAT
1141 ACGTTTAGGA GGTGCTCTGG GTTTGATGAA AATCTTTAAG CTGCGTTAAA GTAGAACAAC
1201 TGCTA
```

ORF Predicted Amino Acid Sequence:

```
VKSEKELDIA RTEFIKSFNF VVGTLRMNGL SRKVAVGLAL MTLIGGRASI
RNASITFKLN YANLLKTLEN LENTWSDYLE ALSKVIVGPV VVIIDDTFDH
KLYSRVEGIA SKYGNYFAWC STHKRFEPGI QVLTIALYDL AMGKSYLIGA
FPYATRKMWE SGMVSEFKTK IEMAAEIIEV LKERFHVVRV VFDSWYWSEK
LVRDSVVSEL KSNRRLLRVR PLEGEKTLGV EGHPHVGDLP PGSYLAELTL
GDQVITIKLL ILVYKDNRLN LYTTDLNLSD EEIEATWKIR WEIEKFPKDI
KALGMQDSSF LKRKRLQGYL LLFVMVVNTV RDLISSLNLK SVEELLRFVE
IRLGGALGLM KIFKLR
```

ISC1205, Kamchatka Variant

Origin: Kamchatka 12-1

Nucleotide Sequence (Putative ORF in blold):

```
   1 AGCAGTTGTT CTACTTTTCC TCGGTAATAC AAAATTCCCA GTACTACCGA TAAAGTTTTT
  61 AGATTTACGT TTAGTTGATT AAAAACTGTG AACCCCGAGA AAGAGCTGGA CATTGCAAGG
 121 ACCGAGTTCA TTAAGTCCTT CAACCACCTG ATTGGAACAC TGAGGATGAA CGGGTTAAGC
 181 AGAAAGGTCG CAGTAGGCTT AGCACTCATG ACCTTGATAG GAGGAAGGGC CAGCATTAGA
 241 AACGCATCCA TCACGTTCAA GCTAAACTAC GCCAACTTGT TAAAAACACT GGAGAACTTA
 301 GAGAACACAT GGAGAGACTA CCTTGAAGCG TTAAGCAAAG TGATAATTGG ACCGGTAGTA
 361 GTGATAATTG ACGACACCTT CGACCACAAA CTCTATTCCA GAGTAGAGGG CATAGCAAGC
 421 AGATACGGGA ATTACTTCGC tGGGTGCTCC ACGCACAAGA GATTCGAGCC CGGCATACAA
 481 GTCCTCACAA TAGCCTTATA CGACTTAGCC ATGGGAAAGA GCTATCTGAT AGGAGCTTTC
 541 CCATACGCCA CGAGAAAGAT GTGGGAAAGC GGGATGGTAA GTGAGTTCAA GACCAAGATC
 601 GAGATGGCTG CGGAAATTAT CGAGGTTCTC AAAGAGCGGT TCCATGTGGC GAGGGTAGTG
 661 TTTGACTCCT GGTATTGGTC GGAGAAGCTG GTGAAGGGTA GTGTAGTTTC GGAATTGAAG
 721 TCCAACAGGA GGCTCATCAG AGTTAGGCCT TTGGAGGGAG AGAAGACGTT GGGGGTGGAG
 781 GGGCACCCCC ATGTCGGAGA TCTCCCTCCA GGGTCTTACT TGGCTGAGTT GACCCTAGGA
 841 GACCAAGTTA TTACTATAAA GTTGTTAATA CTGGTATATA AAGATAACAG GCTCAATTTG
 901 TACACTACTG ACCTTAACTT GAGCGATGAG GAGATAGAGG CAACTTGGAA GATTAGGTGG
 961 GAGATAGAGA AGTTTCACAA GGATATTAAG GCTCTGGGTA TGCAAGATTC CTCTTTCCTC
1021 AAGAGGAAGA GGCTTCAAGG TTATTTACTC CTCTTCGTGA TGGTGGTTAA CACGGTCAGA
1081 GATTTGATCA GCTCCCTTAA CTTGAAGAGC GTGGAGGAAA TTCTCCGGTT CGTTGAAATA
1141 CGTTTAGGAG GTGCTCTGGG TTTGATGAAA ATCTTTAAGC TGTGTTAAAG TAGAACAACT
1201 GCTGCT
```

ORF Predicted Amino Acid Sequence:

```
VNPEKELDIA RTEFIKSFNH LIGTLRMNGL SRKVAVGLAL MTLIGGRASI
RNASITFKLN YANLLKTLEN LENTWRDYLE ALSKVIIGPV VVIIDDTFDH
KLYSRVEGIA SRYGNYFAGC STHKRFEPGI QVLTIALYDL AMGKSYLIGA
FPYATRKMWE SGMVSEFKTK IEMAAEIIEV LKERFHVARV VFDSWYWSEK
LVKGSVVSEL KSNRRLIRVR PLEGEKTLGV EGHPHVGDLP PGSYLAELTL
GDQVITIKLL ILVYKDNRLN LYTTDLNLSD EEIEATWKIR WEIEKFHKDI
KALGMQDSSF LKRKRLQGYL LLFVMVVNTV RDLISSLNLK SVEEILRFVE
IRLGGALGLM KIFKLC
```

ISC1288

Origin: YNP99 9-23

Nucleotide Sequence (Putative ORF in bold):

```
   1 GAGAGGGTCC CAGGGGCAAA TAATTGAATA AATTTATGTA AAGGTTTATA AATATTATTT
  61 CAAAAGATAT CGTGAATTAC TGAAAAATTT TGAATTTTCG CTTGAAGCAA AAAGTCGCGT
 121 AATACTAAAA GGGAACCGGA ACGATGTACG TACCGATTAA CATGAAGCAT TGGATAAAAT
 181 ACTACAACGG CCCGGCCCCT TATGAACATT TATCAGCAGG GCATAAAACA CTTGGAAAAA
 241 TGAACTACAT GCTGATCACG TCTGAAGTGC TGCCGCGTTT TTCTGACCTC TTAATATTGA
 301 GAGCGTTAAT AGTTATGATT GGGGGGACGT GCTCCTACAG GGACGGGCGG AGCTACTACA
 361 AGTCAGACGT GGGGGTAAGG TGGTTCCTAG GCGAGTACAA GTCTAAGTCG GAGATCCATA
 421 GGAGGGCAAA GAATTTTAGG GGAGAGGTAA AGACTTTGTT CAAGGAGTCC CCCAAGGAGT
 481 TGGAGGGGAA GATGAGTAAA CTTGCTGACT ACTTACCTAG CAGGGCGTTA TACGGAAAGG
 541 TTGAAAAGCT GGGGATCGTG GATTCCTTCC TAATCGAGGT ACCCTTCGGG AAGAGGAACA
 601 AGGAAACATT GAAAAAGAAG TTTGAGCTAC ACCTAAGGCA GAGGAAGTAC AGGGAGGCGG
 661 CTAACACGCT CTTCTTTTAC ATTAAGTGCA AAGTGAGGAG GAGGTTCAAG GGAGAGTTTA
 721 CAAAGAAGAG GGACAGGAGT TACTTCGGCT TCAAGGTCTT CAACCTCATG TCGCCAACAA
 781 TGATAGTTCA CGAGATTCAA GTGGAGCTGG CCAATTTTCC GGACAATAAG GGGGGCTTCT
 841 CTCGCAGCGG TTATAAGGTA GTGGATAGGG GCTTCGTGGG GAAGTCCTCG ACCTGGTTGA
 901 TAGGTTTCTC TAGTTTCAGG AGGTATGTGG AGTTCTTTGG GATCTTCTTG AGGAGGTATT
 961 GGAGGCCTTA CGCTACTGAA AAGGGTATGG TCGAGCTCTT TGTCTACGTT ATCGCGTTGA
1021 TTTACAACTC CTACATCTAC ACTTCTGTGT TATCGCGTGT TCCGGAGAGT CAACTCGCCC
1081 ACTAACTTGT ACCGCGAGAG TTGATCAAGG TAGTGTGGAC TATGTTGGAA ATTCTGTTTT
1141 TCTCTATACT TGATTATTTT CCATTACAAT ATAAGCTTAA TCTCTCGTTA TACTGAAAGA
1201 TATAATTATA TTTTATTCTT AATTAATTAT AAAATTGTTT TTCCTATCAT ACTATATTTA
1261 TTCAATTATT TGCGCTTGGG ACACTCTC
```

ORF Predicted Amino Acid Sequence:

```
MYVPINMKHW IKYYNGPAPY EHLSAGHKTL GKMNYMLITS EVLPRFSDLL
ILRALIVMIG GTCSYRDGRS YYKSDVGVRW FLGEYKSKSE IHRRAKNFRG
EVKTLFKESP KELEGKMSKL ADYLPSRALY GKVEKLGIVD SFLIEVPFGK
RNKETLKKKF ELHLRQRKYR EAANTLFFYI KCKVRRRFKG EFTKKRDRSY
FGFKVFNLMS PTMIVHEIQV ELANFPDNKG GFSRSGYKVV DRGFVGKSST
WLIGFSSFRR YVEFFGIFLR RYWRPYATEK GMVELFVYVI ALIYNSYIYT
SVLSRVPESQ LAH
```

ISC1926

Origin: Lassen 11

Nucleotide Sequence (Putative ORFs in bold):

```
   1 AAGGGCTGAA TCCTTTCTCA CGTTAATAGA ATCTTTTTA TATTCATTTA TTATCAACTA
  61 ATTTTGTGGA GAGACTACTG AGGCCTAAGG AGGCTTGCCA ACTACTCAGC ATTTCATACT
 121 CAACTCTCCT ACGGTGGATT AGAGAAGGGA AAATAAGGGT GGTAACGACT GAAGGAGGGA
 181 AGTACAGAAT ACCTTACAGC GAAATTAAGA AGTACTTAGA GAAGAGGGAG GAAATAAGGG
 241 CAGTAATTTA CGCAAGAGTT TCATCATCAG ATCAAAAAGA AGATTTGGAG AGACAAATAA
 301 ACTACCTAAC AAATTACGCA ACAGCAAAGG GTTACAAGGT AGTTGAGGTG TTGAAAGATA
 361 TAGCTAGCGG GTTAAACACG CAAAGGAAAG GATTGCTGAA GCTCTTCAAA CTTGTTGAGG
 421 GGAGGAGTGT TGACGTCGTA TTAATAACAT ACAAAGACAG ACTAACGCGT TTTGGATTTG
 481 AGTACATTGA AGAGCTCTTC TCAACCATGG GAGTTAAGAT TGAAGTAGTT TTCGGAGAAG
 541 AACCTAAGGA TGCCACACAA GAACTTGTGG AAGATTTGAT TTCCATTATT ACATCATTCG
 601 CTGGTAAAAT TTACGGTATG AGGAGTCATA AGAAGACAGT CCTAGTTCAA GGTGTAAAAA
 661 AGTTGATAGG TGAGTTAAGT GGAGAGGACG ATAAAGTTAA GGGTTAGGGT TGACTATATT
 721 ACATACTCAG CACTTAAGGA AGTTGAGGGG GAGTACAGAG AGGTTCTAGA GGACGCAATA
 781 AATTATGGGC GTGTCAAACA AAACTACCTC CTTCACTAGA ATTAAAGCTG GAGTTTACAA
 841 GACTGAGAGG GAAAAGCACA AGGACTTACC ATCCCATTAC ATCTACACCG CTTGTGAAGA
 901 TGCAAGCGAG AGGTTGGACA GCTTTGAGAA GTTGAAGAAG AGAGGTAGGA GTTACACTGA
 961 GAAACCTTCA GTGAGGAAGG TCACTGTGCA TCTAGATGAT CATCTGTGGA AGTTCAGTCT
1021 CGATAAGATC TCAATTTCCA CAATGCAAGG TAGGGTTTTC ATTTCACCAA CCTTCCCTAA
1081 GATCTTCTGG AGATATTATA ACACGGAGTG GAGGATTGCG AGTGAAGCCA GGTTTAAGTT
1141 GTTGAAGGGA AATGTTGTAG AGTTCTTCAT AGTTTTTAAG AGGGACGAGC CTAAACCTTA
1201 TGAACCTAAG GGTTTCATCC CCGTCGACCT TAACGAGGAT TCGGTCTCTG TATTAGTTGA
1261 TGGAAAACCG ATGCTTTTAG AGACTAACAC TAAGAGGATT ACTCTGGGCT ATGAGTATAG
1321 GAGGAAGGCA ATAACAACTC GTAGGTCAGC TGAGGATAGA GAAGTGAAGA GGAAGTTAAA
1381 GAGGCTGAGG GAGAGGGATA AGAAAGTAGT CATTAGGAGG AAGTTGGCTA AGCTGATCGT
1441 TAAAGAGGCT TTTGAAAGTA TGAGTGCAAT TGTCTTAGAG GCCTTGCCAA GGAGACCTCC
1501 AGAGCATATG ATAAAGGACG TGAAAGACTC TCAGCTTAGG TTGAGGATTT ATAGATCGGC
1561 ATTTTCCTCA ATGAAGAATG CTATTATAGA GAAGGCTAAG GAGTTTAGAG TCCCCGTAGT
1621 CTTAGTTAAT CCCTCATATA CTTCTTCAAC TTGTCCAATC CACGGGGCGA AGATCGTTTA
1681 CCAACCCGAT GGGGGCGATG CCCCAAGGGT TGGTGTTTGT GAGAAGGGGA AGGAAAAGTG
1741 GCATAGGGAT GTGGTTGCCC TCTATAATTT GAGGAAAAGG GCTGGAGATG TGAGCCCCGT
1801 GCCGTTGGGC TCGAAGGAGT CCCATGACCC ACCTACCGTT AAGTTAGGCA GGTGGTTGAG
1861 GGCTAAGTCC CTACACTCGA TCATGAATGA ACATAAAATG ATTGAAATGA AAGTGTAGGG
1921 ACAAAC
```

ORF I Predicted Amino Acid Sequence (Putative Resolvase):

IHLLSTNFVE RLLRPKEACQ LLSISYSTLL RWIREGKIRV VTTEGGKYRI
PYSEIKKYLE KREEIRAVIY ARVSSSDQKE DLERQINYLT NYATAKGYKV
VEVLKDIASG LNTQRKGLLK LFKLVEGRSV DVVLITYKDR LTRFGFEYIE
ELFSTMGVKI EVVFGEEPKD ATQELVEDLI SIITSFAGKI YGMRSHKKTV
LVQGVKKLIG ELSGEDDKVK G

ORF II Predicted Amino Acid Sequence (Putative Transposase):

MGVSNKTTS FTRIKAGVYK TEREKHKDLP SHYIYTACED ASERLDSFEK
LKKRGRSYTE KPSVRKVTVH LDDHLWKFSL DKISISTMQG RVFISPTFPK
IFWRYYNTEW RIASEARFKL LKGNVVEFFI VFKRDEPKPY EPKGFIPVDL
NEDSVSVLVD GKPMLLETNT KRITLGYEYR RKAITTRRSA EDREVKRKLK
RLRERDKKVV IRRKLAKLIV KEAFESMSAI VLEALPRRPP EHMIKDVKDS
QLRLRIYRSA FSSMKNAIIE KAKEFRVPVV LVNPSYTSST CPIHGAKIVY
QPDGGDAPRV GVCEKGKEKW HRDVVALYNL RKRAGDVSPV PLGSKESHDP
PTVKLGRWLR AKSLHSIMNE HKMIEMKV