Spring 2019

A genome wide survey of the insertion sequences in Halanaerobium hydrogeniformans, a haloalkaliphilic anaerobic bacterium

Kody Austin Bassett

Department:

A GENOME WIDE SURVEY OF THE INSERTION SEQUENCES IN *HALANAEROBIUM HYDROGENIFORMANS*, A HALOALKALIPHILIC ANAEROBIC BACTERIUM

By

KODY AUSTIN BASSETT

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN APPLIED AND ENVIRONMENTAL BIOLOGY

2019

Approved By:

Ronald Frank, Advisor

Melanie Mormile

Dave Westenberg

ABSTRACT

Insertion sequences (IS) are the smallest prokaryotic transposable elements. These mobile genes are highly diverse in size and structure, making them difficult to study. The transposition activity of insertion sequences has a significant role in evolution by enabling genomic plasticity via genome rearrangements. Insertion sequences are largely uncharacterized and thus methods to improve the ability to accurately identify, annotate, and infer genomic impact of insertion sequences is limited. A sequential set of methods using readily available genomic and bioinformatic tools was developed to accurately identify insertion sequences. This method was used to perform an entire genome survey of *Halanaerobium hydrogeniformans* to identify and characterize all insertion sequences. After characterization insertion sequence activity is inferred, including transposition capability, element interaction, and insertion sequence evolution.

ACKNOWLEDGMENTS

I would like to extend my greatest appreciation to Dr. Ronald Frank for being a mentor and advisor to me and for all the knowledge he has shared with me. He spent a substantial amount of his time in guiding me through the research and answering my questions. I could not have asked for a better advisor. I would like to thank Dr. Melanie Mormile for serving on my committee and providing career and research related advice. I would like to thank Dr. David Westenberg for serving on my committee and providing early feedback on research. I would also like to thank Brynn Shrom and Ivana Grimm for the research they performed on the project. A special thanks goes to Terry Wilson for being an amazing mentor to me. Her counseling as an advisor and positivity made it easier to complete this project and kept me motivated.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. TRANSPOSABLE ELEMENTS

Transposable elements are mobile DNA segments capable of excision and integration within their host genome. They carry a gene encoding a transposase (Tpase), which is responsible for the transposition activity. These elements can carry non-transposase genes known as accessory or passenger genes [1], such as antibiotic resistance genes. Transposable elements were discovered by Barbara McClintock when she was studying variation in maize kernel coloration [2]. At first, the science community believed transposable elements were junk DNA or selfish genes with little benefit to their hosts. It is now known that transposable elements are evolutionary agents that can increase genetic diversity through gene duplication, genomic rearrangements, and horizontal gene transfer [3]. Additionally, transposable elements have been shown to be the most abundant and ubiquitous genes in nature [4]. Transposable elements and their relics, that have lost the ability to transpose via mutations, represent a large portion of eukaryote genomes (80% in maize) but make up a relatively much smaller percentage of prokaryotic genomes [5].

Transposable elements are classified by structure and mechanisms of transposition and can be grouped into 2 classes. Class 1 transposable elements are composed of retrotransposons and retroposons, have similar structure to mRNA and retroviruses and are usually bound by long terminal repeats. This class of transposable element transpose via an RNA intermediate. Class 2 transposable elements are composed of insertion sequences and transposases, and transpose through DNA intermediates. They typically carry inverted

repeats at their terminal ends [6]. Also, many eukaryotic transposons are related to prokaryotic insertion sequences, and carry a variety of passenger genes [7].

1.2. INSERTION SEQUENCES

Insertion sequences are small, simple prokaryotic transposable elements that show high diversity in structure and organization. Insertion sequences typically have an open reading frame (ORF), terminal inverted repeats (IR), and produce a target site duplication (TSD) upon insertion. The high diversification in these features, as well as their catalytic mechanisms for transposition, are used to categorize insertion sequences into groups and families, of which there are 4 major groups and 29 distinct families. It is important to keep in mind that these are the basic characteristics of insertion sequences, and that not all insertion sequence families follow this basic outline.

1.2.1. Organization & Regulation. Insertion sequences are smaller compared to other transposable elements, typically between 0.8 and 2.5 kb in size, and carry a single open reading frame required for transposition. The N-terminal and C-terminal regions principally contain DNA-binding and catalytic domains, respectively. This orientation may permit the coupling of synthesis and activity of the transposase as amino acids are added to the C-terminus of the nascent polypeptide [8], [9]. Further evidence of the purpose of this organization is that for several insertion sequence families, DNA-binding domains isolated from the catalytic domains bind more readily than the whole protein. This suggests that the C terminal inhibits DNA binding to a degree through steric masking and provides an explanation for the preference of many transposases to act in cis (which is the preference for transposases to mobilize the gene from which is was encoded) [10].

Some insertion sequence families, IS3 and IS21, encode two ORFs. In both families the second ORF appears to play a role in transposition regulation using different methods. In the IS3 family, the two ORFs, orfA and orfB, overlap slightly. A slippery site results in the fusion of these two ORFs and the translation of a full-sized transposase. Translation of the transposase cannot occur without this fusion, and as a result might limit the frequency of IS3 transposition. In the IS21 family, the second ORF called IstB, encodes an ATPase helper gene that improves the efficiency of IS21 transposition.

1.2.2. Terminal Inverted Repeats. With a few exceptions, insertion sequences contain terminal IRs. These are short imperfectly-matched sequences that read the same 5' – 3' on each strand of DNA usually within 50 nucleotides of the ends of the element. The inverted repeats are involved in strand identification by the transposase enzyme. After the enzyme has bound to the DNA, the IRs are used in strand cleavage and transfer during the transposition reaction [11]. Additionally, the IRs may contain endogenous promoters or protein binding sites to allow regulation of gene expression by the element or the host, respectively [12].

1.2.3. Target Site Duplications. Most insertion sequences exhibit a target site duplication (TSD) that is generated on insertion. This target site duplication is often referred to as a direct repeat (DR). During transposition, staggered DNA cuts are made at the flanking ends of the target site. Upon insertion and DNA polymerase action, the short sequences flanking each end of the element are identical in the 5' – 3' direction. The size of the direct repeats varies between families and elements, but typically range between 2-14 bp in length [13]. However, there are some insertion sequence families that do not have

direct repeats. In some cases, recombination of the DR may have occurred between two insertion sequences resulting in a mismatch of the direct repeats flanking the target site.

1.2.4. Target Sequence. Some insertion sequences have a regional preference for insertion sites, inserting within an AT or GC rich area. Other elements require specific sequence ranging between 4-9 nt in length. Many Insertion sequences insert within or proximal to sequences that resemble their own terminal inverted repeats. These elements often transpose with an intermediate of an IR-IR junction (including members of IS30 and IS3 families). This process can result in a cascade of transposition events and numerous insertion sequences located in close proximity to one another [12]. The general structure of an insertion sequence is represented in Figure 1.1 [14].

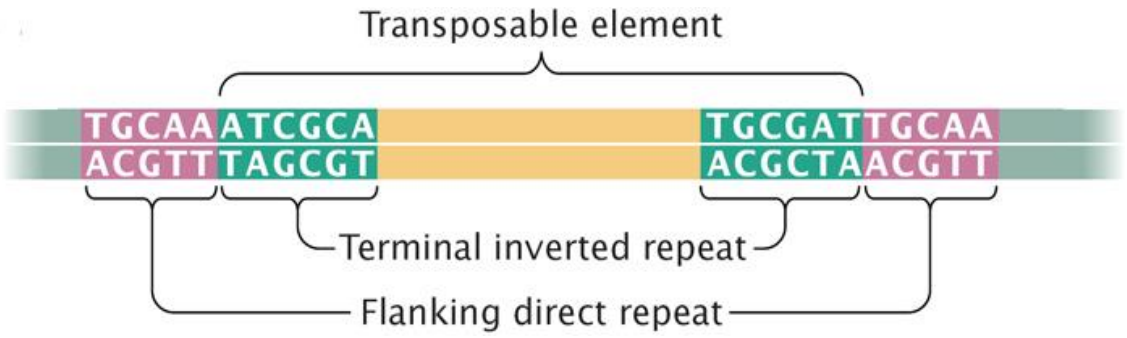

Figure 1.1. General Structure of an Insertion Sequence.

1.3. CATALYTIC MECHANISMS

There are four groups of insertion sequence families based on the catalytic mechanism of the transposase enzyme. These groups are 1) DDE, so called for the conserved catalytic DDE motif, 2) S, named for the serine residues at the catalytic site, 3) Y1, named for the single tyrosine as the catalyst and 4) Y2 group that shows similarity to proteins involved

in rolling circle replication. See Figure 1.2 for the number of identified insertion sequences grouped by family and catalytic chemistry.

1.3.1. DDE. The most common type of catalytic mechanism used by insertion sequences is DDE. This catalytic mechanism features a highly conserved triad of negatively charged catalytic residues D (asp) D (asp) E (glu). The distance between each residue varies between families but is highly conserved within families that use this mechanism for transposition. The DDE catalytic mechanism is separated into two steps. The first is DNA cleavage through hydrolysis and the second is the attack of the target site by the free 3'OH on the element end. However, DDE family members transpose via a double stranded intermediate. Generating this free dsDNA intermediate requires further processing of the second strand that is family specific [15]. The second strand is most commonly freed from its flanking DNA through the formation of a transient hairpin at the element end [16]. This DDE transposition mechanism has also been observed in host functions. For example, the RAG1/2 complex that catalyzes V(D)J recombination in developing lymphocytes is thought to have come from a domesticated transposase. RAG1 contains a highly conserved DDE motif [17].

1.3.2. Serine. Serine transposases are encoded by the IS607 family of insertion sequences and show some similarity to serine recombinases that catalyze inversion of DNA segments [18]. Although characterized groups of serine recombinases show an inversion of the typical DNA domain organization, Serine transposases show the typical domain organization with DNA binding and catalytic domains in the N-terminus and C-terminus respectively [19]. In addition to the transposase, IS607 family members also encode a second protein known as ORFB or TnpB, that shows high sequence similarity to a protein

encoded by members of the IS605 family. This second protein is not required for IS607 transposition [20]. IS607 elements in *E. coli* systems have been shown to insert with very low target sequence specificity, which is atypical for reactions catalyzed by serine recombinases [21].

1.3.3. Y1. The Y1 transposases which are among the smallest identified transposases (approximately 150aa in length), use a single catalytic tyrosine. These transposases act exclusively on single-stranded DNA [24]. The transposase catalyzes DNA breakage and the formation of a 5' phosphotyrosine intermediate using the catalytic tyrosine residue. This leaves a free 3'OH at the cleavage site. Like other HUH endonucleases, Y1 transposases recognize and bind DNA hairpin structures, cleaving ssDNA on either side of the stem or even within the hairpin structure itself [22]. These small hairpins can be identified and bound by the transposase through sequence specific recognition of the stem or loop, or through the recognition of structural irregularities in the stem [23]. As a result of this unique for transposition, the Y1 transposable elements do not contain inverted repeats and there is no target site duplication created upon transposition. The Y1 family members are also unique in that they insert and excise preferentially from and into ssDNA [24], [25].

1.3.4. Y2. Y2 insertion sequences, encoded by IS91, also fall within the HUH endonuclease superfamily. While Y1 transposases work via a single catalytic tyrosine, Y2 transposases use two tyrosine residues to carry out transposition with its unique mechanism, similar to proteins involved in rolling circle replication [1]. IS91 elements insert 3' to a conserved tetranucleotide sequence [26]. The mechanism of transposition is

not clear for the Y2 transposases since they are limited to a single IS family and research is limited.

1.4 TYPES OF TRANSPOSITION

After the transposase binds to one of the DNA strands at the terminal IRs, two different types of genome movement can occur based on the type of insertion sequence: conservative or replicative transposition [27].

1.4.1. Conservative Transposition. Conservative, cut-and-paste, transposition occurs when the insertion sequence is excised from the donor site with no formation of a plasmid and moved to the target site resulting in the movement of the insertion sequence from one location to another in the genome.

1.4.2. Replicative Transposition. Replicative, copy-and-paste, transposition occurs with the duplication of the IS upon fusion and formation of a cointegrate intermediate between two plasmids [27]. Upon resolution of this intermediate by recombination there are left two copies of the element in the genome, one at the original site and one copy at the target site.

1.5. INSERTION SEQUENCE FAMILIES

There are thousands of discovered insertion sequences and they all vary in length, ORF number and sizes, DRs, and terminal IRs. Most IS are grouped into major families based on similar characteristics and structure. It should be noted that the characteristics described below for each family are the most common characteristics and may not

represent every member. See Figure 1.2 [1] for the distribution of identified insertion sequences across prokaryotic genomes.

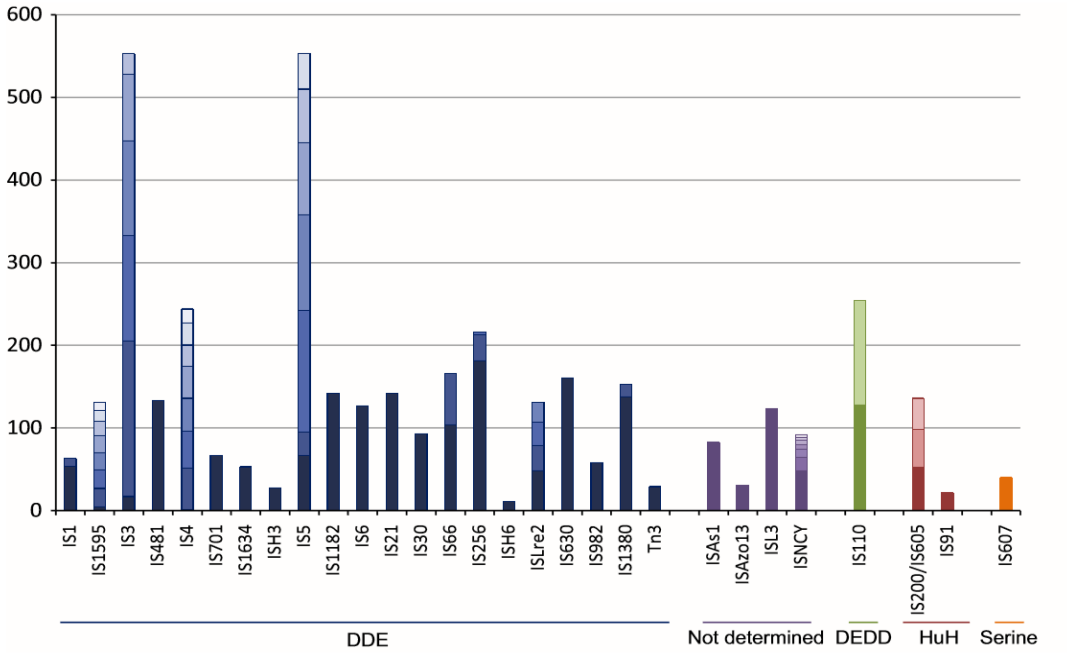

Figure 1.2. Insertion Sequence Diversity.

1.5.1. IS256. The IS256 family has average lengths ranging from 1,200 bp to 1,500 bp in which a single large ORF is encoded. However, there are members such as the closely related IST2 (*Thiobacillus ferrooxidans*) and IS6120 (*Mycobacterium smegmatis*) that appear to encode two open reading frames and may represent members of a distinct subgroup [28]. The DDE motif of this family usually contains a 112-residue spacer between the second D and E residues, along with a correctly placed K/R residue between them [28]. The inverted repeats range between 24 to 41 bp and the target site duplication is 8 bp except for a few members having a repeat of 9 bp [29].

1.5.2. IS30. The IS30 family is large with many similar elements, such as IS1086. IS30 elements have average nucleotide lengths of 1,221 bp that end with the dinucleotide sequence CA-3' [30]. Within the element resides a single ORF ranging from 293 to 383 codons with the stop codon close to the terminal end. IS30 transposase have a conserved DD(33)E motif (signifying the second aspartic acid is always the 33rd codon) as the catalytic mechanism [13]. The IS30 elements have been shown to utilize copy-and-paste transposition. The terminal IRs are highly homologous ranging from 20 to 30 bp. The direct repeats are highly variable within the family, but IS30 members commonly have repeats of 2 bp [31].

1.5.3. IS6. The IS6 family of elements was named after the directly repeated insertion sequences in transposon Tn6 [32]. IS6 elements are small compared to other families, having average nucleotide lengths of 789 bp (IS257) to 880 bp (IS6100) [13]. In one IS6 family member (IS26) the ORF is transcribed from a promoter located within the first 82 bp of the left end [33]. The IS6 family members are small in size, and most of the element makes up the ORF. The corresponding Tpase exhibits strong DDE catalytic chemistry. The terminal IRs are short related sequences ranging from 15 to 20 bp. IS6 family members commonly have directs repeats of 8 bp. Interestingly, studied IS6 members show exclusive use of cointegrates during transposition resulting in replicative transposition [13].

1.5.4. IS21. IS21 elements contain two ORFs and have short terminal IRs around 11 bp. IS21 family members have target site duplications of 4 bp or 8 bp, but most frequently have 5 bp repeats. IS21 elements exhibit two consecutive open reading frames. The first frame, called IstA, is closer to the putative promoter and longer than the second

frame, designated IstB. The arrangement of the two ORFs suggests that IS21 elements are polycistronic and IstA encodes a transposase while IstB encodes an ATPase used for regulation and target DNA recognition [34]. However, IstA has been shown to occur without IstB and still have reduced transposition activity. In IS21 elements the IstA is located in a -1 reading frame relative to the IstA, for example if IstA is in reading frame 2, IstB would be in reading frame 1. In IS21 elements IstA and IstB overlap by 1 nucleotide. The adenine of the stop codon, TGA, for IstA is the adenine of the ATG for IstB [13]. The IstA reading frame carries a DDE related motif but lacks the conserved K/R residue that most other DDE elements have, while the IstB reading frame has a conserved nucleoside triphosphate binding domain [35].

1.5.5. IS1182. The information about IS1182 family members is limited to general characteristics. IS1182 family members have average sizes ranging between 1,330 to 1,950 bp. Members do contain inverted repeats, but sizes vary, and have had direct repeats between sizes of 0 to 60. Elements contain a single ORF with a DDE catalytic motif [36].

1.5.6. IS3. One of the largest families, with over 80 members and found in many organisms the IS3 family has also been studied extensively. Although there is high variation within the family, there are several unique characteristics that all members share, including two ORFs. Members range in size between 1,200 and 1,550 bp, with few exceptions. The inverted terminal repeats are found to be between 20 and 40 bp. The majority of IS3 members terminate with 5’-TG-----CA-3’ and internally contain a GC rich region [13]. The two ORFs are partially overlapping, labeled orfA and orfB, in the relative translational reading phases 0 and -1, respectively. It has been demonstrated in at least three cases (IS3 subgroup IS150 [37], IS3 subgroup IS3[38], IS3 subgroup IS911[39]) that a

fusion protein, orfAB, is generated by programmed translational frame shifting. However, the products of orfA and orfB are also present as well. The frequency of this programmed frame shift varies between elements but is approximately 50% for IS150[37] and only 15% for IS911[39]. The predicted primary amino acid sequences from orfA exhibit a strong helix-turn-helix motif. While the orfB carries the catalytic DD(35)E domain and have a Tpase like structure. The IS3 family is highly divided containing several subgroups including IS3 ssgr IS150, IS3 ssgr IS3, IS3 ssgr IS911. These subgroups mostly differ in the alignments of the upstream orfA protein. Another correlation between subgroups is direct repeat size upon insertion, ranging between 3bp and 6bp.

1.5.7. ISNCY. The ISNCY family contains various elements that resemble insertion sequences but remain to be classified. These are elements whose nucleotide sequence is not known or limited and insufficient for family assignment or elements whose entire nucleotide sequence is known but show no significant relationship with more than one other element [13].

1.6. MINIATURE INVERTED REPEAT TRANSPOSABLE ELEMENTS

Miniature inverted repeat transposable elements (MITES) are partial copies of transposable elements. Partials are insertion sequences that are significantly smaller in size than the parent elements or isoforms, and do not usually contain an intact ORF. In genomes without full length parent copies it can be extremely difficult to identify MITES, as they often are present as short inverted repeat sequences [5]. Although they may not contain a transposase, MITES can still be transposed via a transposase if there are intact IRs. MITES are therefore impactful to host genomes as their insertion can influence gene expression,

alter mRNA stability, or influence transcription termination. MITES are important because they represent evidence of past insertion sequence activity and are important for understanding the evolution of insertion sequences within the host, and the impact of insertion sequences on the genome.

1.7. GENOMIC IMPACT OF INSERTION SEQUENCES

With the development of stronger bioinformatic tools allowing for further research, the importance of transposable elements has changed. Transposable elements were originally viewed as selfish DNA or a genomic parasite, serving little to no purpose to the host genome. It is now understood that transposable elements play a significant role in genomic diversity, structure, and genetic plasticity [40].

1.7.1. Genomic Streamlining. Genomic streamlining is an evolutionary theory that suggests there is a reproductive benefit to prokaryotes having a smaller genome size with less non-coding DNA and fewer non-essential genes. Insertion sequences experience rapid expansion and loss within host genomes via genomic rearrangement, and gene inactivation. With time, insertion sequences experience deletion along with flanking host DNA, resulting in genome reduction. These natural genomic occurrences will lead to the development of non-functional, or non-autonomous elements, which are eventually cleared from the genome. According to this theory, the relaxed selective pressure that tolerates both the expansion of insertion sequences, and the ensuing genome reduction is beneficial to the organism. It has been observed that insertion sequence numbers increase in new bacterial endosymbionts compared to free living cells [41], and that genomic reduction is correlated with insertion sequence expansion. This is evident in comparing three *Bordetella*

species *B. pertussis*, *B. parapertussis*, and *B. bronchiseptica*. The genome size of *B. bronchiseptica* is the largest of the three (5.34 Mb) and it harbors no insertion sequences, *B. parapertussis* has a reduced genome size (4.77 Mb) with over 100 insertion sequences, and B. pertussis with the smallest genome size (4.1 Mb) has over 260 identified insertion sequences. The phylogeny of the of the organisms suggested that *B. bronchiseptica* was the ancestral species of the three [42].

1.7.2. Insertional Mutation. Easily observed and most likely has the greatest effect on the organism, insertional mutations are the direct insertion into genes causing disruption and mutation. Insertion sequence mediated disruption in a Rickettsi species resulted in non-pathogenicity by insertion into virulence genes [43], and a metronidazole resistant H. pylorus by insertion within genes necessary for pro-drug activation [44].

1.7.3. Gene Expression. Although over 80% of genes in prokaryotic genomes encode proteins, not all insertion events cause a direct disruption. Insertion into intergenic regions can still impact the host genome. Some mobile elements carry transcriptional promoters [12], and their insertion leads to changes in expression of flanking genes. Insertion sequences can also change expression by activating or inactivating repressor genes though endogenous promoters [45].

1.7.4. Genome Rearrangement. Insertion sequences also impact genomes through a variety of chromosomal architecture changes, stemming from the multiple copies of elements with high sequence similarity. Recombination between 2 IR of a single insertion sequence can result in an inversion. Direct inversion of elements carrying endogenous promoters has been shown to increase pathogenicity through phase variation in a number

of organisms [46], [47]. Recombination can also occur between elements resulting in the inversion of the entire sequence between the elements, or in the deletion of sequence between the elements and the formation of a single hybrid element [48]. Alternative transposition mechanisms can also result in intergenic sequence duplication [6].

1.8. HALANAEROBIUM HYDROGENIFORMANS

Our preliminary analyses indicate that *Halanaerobium hydrogeniformans* has many insertion sequences that may or may not be active. *H. hydrogeniformans* is a haloalkaliphilic anaerobe isolated from Washington State's Soap Lake. Among its many metabolic properties and in addition to the ability to produce hydrogen, this organism has the enzymatic ability to convert glycerol into 1,3-propanediol, an industrially useful monomer, with a 60.3% conversion rate [49]. The genome was first sequenced to get insights into the metabolic and adaptive properties that enabled *H. hydrogeniformans* to form hydrogen from cellulosic material. However, after examining the genome, it was found that insertion sequences were scattered across the genome. This extremophile seemed to have a very large number of annotated transposases compared to the average number in *E. coli*. [49] Also, the *H. hydrogeniformans* genome is approximately 3% transposable elements, higher than in most bacterial genomes [5]. The first step in characterizing the insertion sequences, and to provide answers to questions concerning this microorganism, was to perform a genome wide survey to identify all insertion sequences in *H. hydrogeniformans* using *in silico* methods. A preliminary screen of the genome has estimated a total of 127 isoforms and partials belonging to 10 different insertion sequence families, created from several transposition events occurring throughout the existence of this extremophile.

1.9. DATABASES AND BIOINFORMATIC TOOLS

1.9.1. ISfinder & ISsaga. ISfinder is an online public database providing general features (size, TSD, IS family, IRs, ORF) for insertion sequences isolated from bacteria and archaea. They rely on the scientific community to submit sequences and information of characterized insertion sequences to enrich the database. ISfinder also provides a browser that can be used to view identified and predicted insertion sequences in sequenced genomes [36], [50]. ISsaga is a tool of ISfinder that was developed for researches to accurately identify and annotate insertion sequences with the use of a high quality semi-automatic annotation system. ISsaga uses the ISfinder database of submitted insertion sequences to provide general prediction for potential insertion sequences in a genome. It provides genomic context of individual insertion sequences, visual display of genomic positions, and a small array of tools to find element ends, target site duplications, and inverted repeats. The annotation accuracy of ISsaga is limited to the ISfinder library; therefore, insertion sequences predicted by ISsaga are confirmed manually before being added to the ISfinder database [51], [52].

1.9.2. Argo. Argo is a genome browser developed by The Broad Institute, used for viewing and annotation of whole genomes. It displays the sequence and annotation of DNA tracks. Argo supports many file formats including FASTA and Genbank. This program is useful in determining relative position to other genes, as well as extracting DNA and protein sequences for further phylogenetic or structural analysis [53].

1.9.3. NCBI. The National Center for Biotechnology (NCBI) was developed by the National Institute of Health (NIH) after the need for computerized information processing

in modern research was realized. Data from the European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ) is shared with NCBI. NCBI has become an essential tool for most biologists and provides access to DNA and protein sequences, mapping, structural data, and phylogenetic outputs [54]. NCBI is also the host to several automated DNA and protein tools such as BLAST, RefSeq, and ORF-finder.

1.9.4. EBI. The European Bioinformatics Institute (EBI) is part of the EMBL and provides the most up-to-date and comprehensive range of basic research and computational biology tools for researchers in academia and industry. The tools provided span DNA/RNA alignments, molecular structures, protein sequences, families, and motifs, taxonomy, and systems pathways [55]. Many of the tools consistently used were part of the EMBOSS analysis tool pack created by the EBI. EMBOSS Sixpack and Transeq are sequence translation tools that were consistently used during ORF identification.

1.9.5. Phylogeny fr. Phylogeny.fr is a free web based phylogenetic analysis tool for non-specialists. It provides several options to create a customized workflow that caters to the user. An automated or semi-automated phylogenetic relationship can be constructed between nucleotide or protein sequences using a multiple alignment process and provides a newick output for various tree viewers [56].

1.10. SUMMARY

This thesis presents a genome wide survey of all the insertion sequences within *H. hydrogeniformans*. After investigation 15 unique insertion sequence were identified and characterized. This full genome survey expands upon earlier genomic studies done in this organism along with providing a framework for future insertion sequence research.

2. METHODS

2.1. GENOME BROWSER

The Argo genome browser tool was used to visualize the genome of *H. hydrogeniformans*. The *H. hydrogeniformans* genome was imported from the National Center for Biotechnology Information (NCBI), accession number CP002304.1. The genome, in Genbank format, was uploaded into Argo. Insertion sequences were marked and categorized by family. Argo was used for visualization of element positions, strand orientation, and gene proximity, transposase or otherwise. The primary use of the genome browser was for nucleotide extraction and insertion sequence identification [53].

2.2. BLAST

Chosen sequences were aligned against a target database using a Basic Local Alignment Search Tool (BLAST). These BLAST tools are available free for use at NCBI. Databases can be queried with protein or nucleotide sequences using the various types of BLAST that NCBI provides.

2.2.1. BLASTn. A BLASTn search uses a nucleotide sequence as the query and searches a nucleotide database. BLASTn is slower than megablast and discontiguous megablast but allows a word-size of seven bases. This permits the comparison of short sequences with low similarity. Megablast and discontinuous megablast have preset parameters for use with more similar sequences.

2.2.1.1. Megablast. Megablast was used to query a nucleotide sequence for closely related individuals within the *H. hydrogeniformans* genome, working best if sequences show a 95% or higher similarity. Megablast was used once a reference insertion sequence had been chosen to identify IS replicates within the genome.

2.2.1.2. Discontinuous megablast. Discontiguous megablast allows for greater mismatches and is intended for sequences with low similarity or cross-species comparisons. Discontiguous megablast was sparsely used to search for IS replicates that were not annotated.

2.2.1.3. Tblastn. A tblastn searches translated nucleotide database using a protein query sequence. This type of BLAST is commonly used to identify other insertion sequences once a reference protein has been identified. Although nucleotide sequences can change regularly due to mutations, it is common that the protein sequences remain the same because of codon overlap. This is a mechanism that prevents the loss of functional proteins due to mutation.

2.2.2. BLASTp. A conceptual protein sequence is used to query a protein database. This was used when the reference had been selected and the open reading frame (ORF) identified. The identified ORF was used to search for isoforms or partials that were not identified in the preliminary survey.

2.2.3. Other Types of BLAST. Some special cases, such as solos or partials require additional searches. Blastx can be used to search protein databases with a translated nucleotide query. Finally, a tblastx searches a translated nucleotide database using a translated nucleotide query.

2.3. INSERTION SEQUENCE IDENTIFICATION

An extensive survey of the insertion sequences within the *H. hydrogeniformans* genome was performed to identify insertion sequences. All elements identified during the preliminary screening, excluding the IS200/605 family, were further investigated. Four characteristics were collected from the genome wide survey that were used for insertion sequence identification: location and size of the insertion sequence in the genome, size and structure of inverted repeats (IR)s, size and structure of direct repeats (DR)s or target size duplication (TSD), and structure of the open reading frame (ORF). All identified elements were given loci numbers for organization along with unique family names, from ISHaHy2 (IS3 subgroup IS150) to ISHaHy15 (IS1182), upon there addition to the ISFinder database.

2.3.1. BLAST of the Genome to the ISfinder Database. To survey the genome for insertion sequences, a set of programs and procedures were used simultaneously to predict the location of insertion sequences in the genome. All genes annotated as insertion sequence, transposase, and integrase were used in a blastp search against genbank to determine potential products. The results were used as a query against the ISfinder library to confirm insertion sequence identity. After confirmation, a representative ORF from each different insertion sequence group was used for a blastn search against the *H. hydrogeniformans* genome to identify partial insertion sequences that were annotated as pseudogenes or hypothetical proteins. Insertion sequences in the genome were then identified with ISsaga to compare the identity results from manual and semi-automatic library-based methods. ISsaga scans for insertion sequences in annotated genomes by comparing potential sequences against the ISfinder database. It then performs a blastn for replicons within the genome to identify partial elements or potential mobile elements not

in the ISfinder library. These predictions, manual and semi-automatic, were organized into a table containing all the theoretical insertion sequences in the genome. This table also included: an approximated ORF location in the genome, IS family, and the highest percentage similarity sequence for each insertion sequence found in the genome.

2.3.2. Element End Identification. The ends of the insertion sequences were unknown and needed to be identified. A reference sequence was chosen for every family using the list organized from 2.3.1. The reference sequence is an element representative of most members in a family and was chosen based on the predicted ORF size and strand orientation. Element ends are typically within a few hundred nucleotides of the start and stop of any protein coding sequences contained within the element. A sequence including 800 nucleotides flanking the predicted ORF on both sides was extracted from the genome using Argo. This sequence was used as the query in a BLASTn of the entire *H. hydrogeniformans* genome. The BLAST report includes alignments to similar sequences in the genome that potentially belong to the same insertion sequence family as the reference sequence. These other insertion sequences were identified as isoforms or partials in subsequent steps, based on size and similarity to the reference sequence.

2.3.3. Inverted Repeats Identification. Inverted repeats are generally located within 50 nucleotides of each end of the element (Figure 2.1). They play a significant role in the recognition of the element by the transposase and are required for transposition activity. Using the reference sequence, the first 50 and last 50 bases were extracted and placed into a separate document. The last 50 bases were converted to the revert complement using an online reverse complement tool [57].

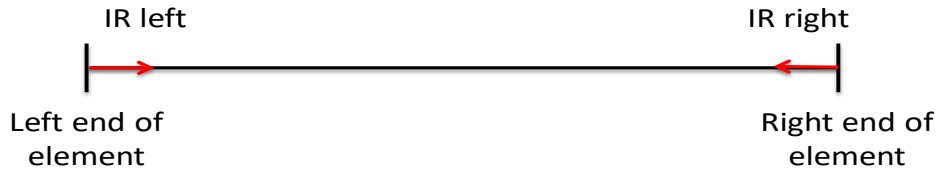

Figure 2.1. Model of Inverted Repeats (IR)s.

This converted sequence was then aligned with the first 50 bases that were not converted and matches were highlighted (Figure 2.2). Inverted repeats are imperfect matches and the best fraction of matches was chosen as the inverted repeats (e.g. in Figure 2.2 the best IR score for this element is 19/28, even though there is a total of 28/50 matches, or a score of .56. This score is smaller than 19/28, which is a score of .68).

```
19/28
AGTAGTGTACAATAAGTTGTGTAAATAGATTTTTTTCAATAAAAAAAGAG
GATAACGTCAAGAATCTTGTGTAAATAAATCAGTATCGTTCTTTAAATAG
```

Figure 2.2. The Imperfect Inverted Repeats of ISHahy7.

2.3.4. Direct Repeats Identification. Direct repeats are perfect matches located immediately adjacent to both ends of the element (Figure 2.3). The direct repeats are a result of a duplication of the insertion site during the process of transposition and are also known as target site duplications (TSD). The direct repeats vary in size based on the insertion sequence family. Together, the inverted repeats and direct repeats were used to verify that the correct ends were chosen previously.

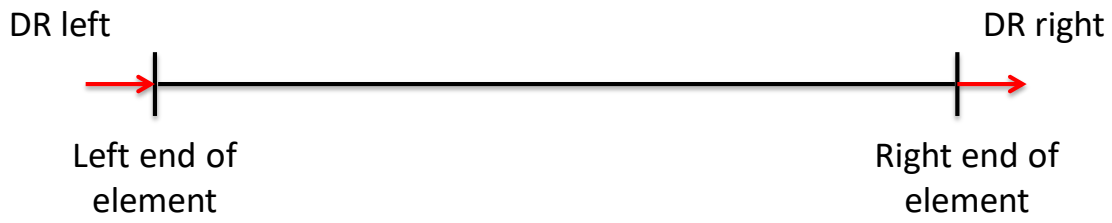

Figure 2.3. Model of a Target Site Duplication.

2.3.5. Open Reading Frame Identification. Insertion sequences harbor genes that encode the transposase proteins that copy or move the element. Some families of elements carry additional genes as well. After the element ends have been determined and verified by inverted and direct repeat identification, a search for ORFs is necessary. Although there are many ORF identification programs with varying degrees of sophistication, EMBOSS Transeq and Sixpack [55] were used for their simplicity. Transeq and Sixpack were used to translate the nucleotide reference element into three frames depending on strand orientation. After translation, long uninterrupted amino acid sequences were highlighted in the EMBOSS Transeq results. The first methionine in the uninterrupted amino acid sequence was assigned as the start of the ORF, and the first stop codon, symbolized by an asterisk was assigned as the stop of the ORF. EMBOSS Sixpack was used to align the translated ORF to the DNA sequence of the element and the location of the first and last nucleotides of the ORF were recorded.

2.4. ORF DISRUPTION

Insertion sequences can insert within genes disrupting the coding region. Automated identification of disrupted genes is difficult. To identify if any of the identified

insertion sequences inserted within a gene, the insertion sequence plus 1000 nucleotides on either side were extracted in Argo. The insertion sequence was then deleted along with one of the DRs and the extended regions were spliced together. The 2000 nucleotide sequence was then translated using the same methods from 2.3.5 to look for theoretical ORFs. These ORFs were subjected to a BLASTp search against the NCBI database to identify potential protein products.

2.5. SYNONYMOUS SUBSTITUTION RATES

The insertion sequences that are found in the *H. hydrogeniformans* genome are most likely the result of multiple invasions at different times. A synonymous substitution rate analysis was performed on families with high variation between isoforms to determine if relative time of invasion events could be estimated. A multiple sequence alignment of the amino acids was performed using Clustal omega [55]. The amino acid alignment was then converted to a codon alignment using Pal2Nal, a program that converts protein alignments into codon alignments [58]. SNAP (Synonymous Non-synonomous Analysis Program) was then used to calculate the synonymous and non-synonymous substitution rates [59].

2.6. PHYLOGENETIC ANALYSIS

To make evolutionary comparisons between different sequences belonging to the same family a phylogenetic analysis was performed using Phylogeny.fr. Though various online tools exist for phylogenetic analysis this one provided a customized “a la carte” option. The nucleotide sequence of the reference insertion sequences was extracted in FASTA format and uploaded into Phylogeney.fr to generate a tree [56], [60]. The

alignment program MUSCLE was chosen to perform the multiple sequence alignment with a Gblocks alignment curation. The phylogeny was then constructed using maximum likelihood and the Drawtree program.

2.7. GENOMIC MAP OF IS

A genomic map of the insertion sequences relative to all genes was developed using a genome viewer called CiVI, or Circular Visualization for Microbial Genomes [61]. The table of insertion sequences used to develop the genome can be found in appendix section.

2.8. LEADING/LAGGING STRAND BIAS

A comparison was done to determine if any insertion sequence families appeared to favor transposition to the leading or lagging strand. At the start of replication, in plasmid DNA, two DNA polymerase bind to the origin on alternate strands (DNA polymerase reacts 5' – 3') and begin working in opposite directions from the origin. Knowing this, we can theoretically assign each IS as being a leading strand template or lagging strand template.

3. RESULTS

3.1. INSERTION SEQUENCE IDENTIFICATION

ISsaga identified 16 insertion sequence families within the *H. hydrogeniformans* genome. Table 3.1 presents the number of unique insertion sequences per family and the total number of elements belonging to that family as identified by ISsaga. Highlighted are all insertion sequence families that were further studied in this body of work. Detailed characterization of insertion sequences in *H. hydrogeniformans* was done on all families excluding members of the IS200 and IS605 families. Of the families studied in detail, manual curation identified 15 unique insertion sequences with a total of 72 loci. Of these loci, 60 were identified as isoforms and 12 were identified as partials. Each insertion sequence was given an independent locus number corresponding to its relative position to other isoforms and the origin of replication. The locus numbers for each element, unique family name, as well as some of the elements characteristics which are further discussed, are outlined in the table found in the appendix section. ISsaga identified 7 loci that upon manual annotation were concluded to be false positives. Two IS91 elements were deemed false-positives. The IS91 elements were both solo accessory genes that ISSaga annotated as integrase, qualities of an IS91 false-positive [36]. IS21 helper genes are common false-positives because they are ATPase-like genes. It should be noted that not all characteristics were identifiable for partial insertion sequences due to variation in sizes.

Table 3.1. Insertion Sequences in *H. hydrogeniformans* as identified by ISsaga.

Family	**Unique IS**	**Total IS**
IS200_IS605_ssgr_IS1341	1	5
IS3_ssgr_IS407	1	3
IS3_ssgr_IS3	4	4
IS6	2	7
IS607	2	15
ISNCY_ssgr_IS1202	1	4
IS256	4	14
ISNCY	1	2
IS30	3	12
IS3_ssgr_IS150	3	16
IS200_IS605	2	4
IS1182	2	2
IS21	2	3
IS3_ssgr_IS51	1	8
IS3	1	8
IS110	1	1
Total	**31**	**108**

3.2. IS3 FAMILY MEMBERS

The genome of *H. hydrogeniformans* contains 29 loci that harbor either functional, defective, or fossil remnants of the IS3 family of bacterial insertion sequences (Figure 3.1). The elements have been divided into five groups (ISHahy2, 3, 4, 5, and a solo partial element). ISHahy4 and ISHahy5 are most closely related to different IS3 elements in *Halanaerobium praevalens*, ISHahy2 is most like an IS3 element in *Halobacteroides halobius*, ISHahy3 is most similar to an IS3 element in *Acetohalobium arabaticum*, and the solo partial element has highest similarity to sequences in *Bacillus cereus.* The transposase of typical IS3 elements is encoded by two overlapping out-of-frame open reading frames (ORF)s. and a programmed translational frameshift at a slippery site upstream of the stop codon in orfA. The slippery site is followed by potential hairpin sequence that is characteristic of many programmed -1 ribosomal frameshift (PRF) mechanisms. An example of this potential hairpin structure is shown in Figure 3.1 below.

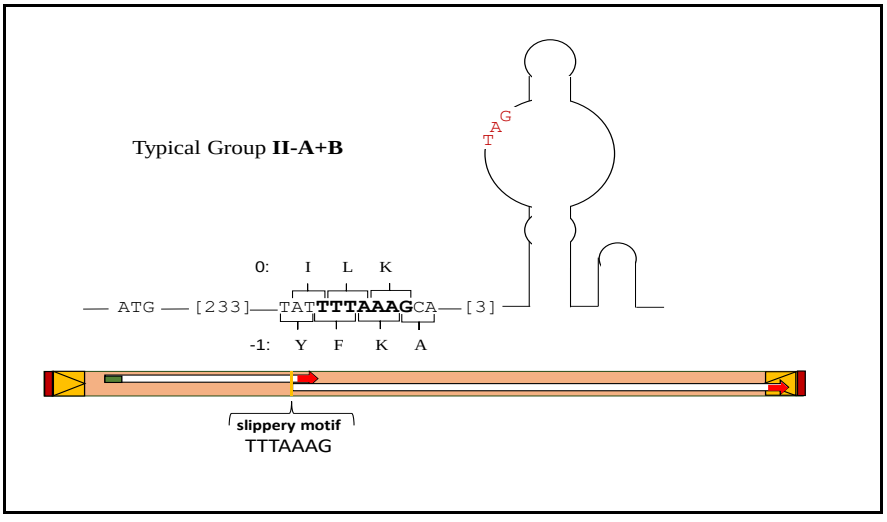

Figure 3.1. Structure of IS3 Programmed Frameshift and Hairpin Structure.

3.2.1. ISHahy2. Eleven copies of ISHahy2 (loci numbered clockwise from the origin) are found in the *H. hydrogeniformans* genome, and all are 91% or more similar at the nucleotide level encoding transposases 98% or more similar at the protein level. The structure of the reference element for this group (locus 8) is 1212 nt in length (Figure 3.2). Most of the loci had a target size duplication (TSD) of 3 nt. Locus 4 exhibits a 4 nt TSD and two, loci 5 and 11, do not exhibit a TSD. Locus 1, being truncated at the right end cannot show a TSD. The element starts and ends with 5'TG-CA3' and the inverted repeats show strong symmetry (29 of 41 nt). An open reading frame ORFA begins 68 nucleotides from the left end and 21 nucleotides from the IRL (inverted repeat-left). A heptameric slippery site beginning at codon 85 of orfA allows a PRF extending into orfB that encodes a full-length 375 amino acid orfAB transposase. The stop codon of orfB is 17 nucleotides from the right end of the element and internal to IRR (inverted repeat-right). Eight of the eleven copies of this element are identical in nucleotide sequence, loci 3, 4, 5, 6, 7, 8, 10, and 11 (locus 3 has a single nucleotide substitution and locus 7 has a single nucleotide insertion). Two copies, loci 2 and 9, are 99% similar to each other but only 94% similar to the reference copy group at the nucleotide level. The remaining copy, locus 1 is 92% similar to the reference copy group and 91% similar to loci 2 and 9 at the nucleotide level largely due to the fact that this copy is missing 32 nucleotides from its right end. Two elements, locus 7 of the reference copy group and locus 2 not of the reference copy group have a single nucleotide insertion at precisely the same location in orfA immediately following the heptameric slippery site. This insertion brings orfA and orfB into the same frame and potentially encodes a full-length transposase without the need for the PRF. All eleven elements have identical IRL sequences. The eight loci identical to the reference

copy have IRR sequences that show the greatest symmetry to IRL (29 of 41 nt), the two loci 2 and 9 (27 of 41), and the single element locus 1 is missing most of the IRR.

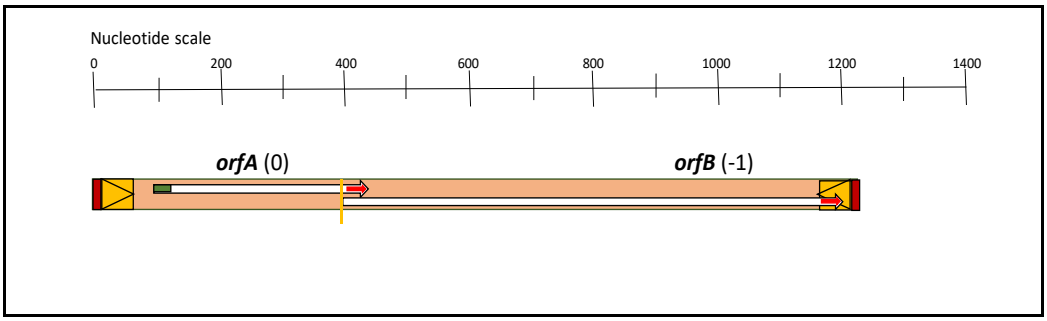

Figure 3.2. ISHahy2 Structure.

3.2.2. ISHahy3. Three copies of ISHahy3 are found in the *H. hydrogeniformans* genome. Two are full length elements and one is a partial element. The structure of the reference element for this group (locus 1) is 1360 nt in length. The element starts and ends with 5'TG-CA3' and the inverted repeats show weak symmetry (9 of 16 nt). However, if a single nucleotide gap is allowed in IRR the matches increase to 18/27. An open reading frame ORFA begins 73 nucleotides from the left end. A heptameric slippery site beginning at codon 90 of orfA allows a PRF extending into orfB that encodes a full-length 404 amino acid orfAB transposase. The stop codon of orfB is 74 nucleotides from the right end of the element and lies entirely outside the IRR (Figure 3.3). The reference element has a TSD of 4 nt and locus 2 does not have TSD. The second full length ISHahy3 element is 98% identical to the reference element it contains many nucleotide substitutions, as well as a single nucleotide deletion in orfA such that a programmed -1 ribosomal frameshift would not coincide with orfB but with the third reading frame that harbors numerous stop codons.

It is unlikely that this element encodes a full-length transposase. The third copy of ISHahy3 is a partial element missing the first 807 nucleotides. The remaining 552 nucleotides of the right end contain a single nucleotide deletion and two nucleotide substitutions. The site of the truncation does not appear to harbor any other insertion sequences.

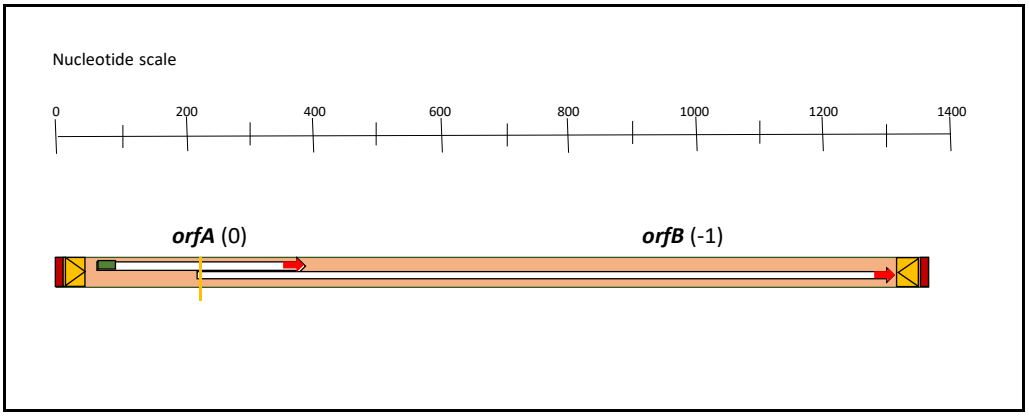

Figure 3.3. ISHahy3 Structure.

3.2.3. ISHahy4. Ten copies of ISHahy4 are found in the *H. hydrogeniformans* genome, all 99.5% similar at the nucleotide level. The structure of the reference element for this group (locus 2) is 1411 nt in length. None of the elements conform to the 5'TG-CA3' ends common to the IS3 family. All ten elements begin with 5'CT and end with CG3'. They have weaker inverted repeats than ISHahy2, 11 of 16 nucleotides. An open reading frame orfA begins 73 nucleotides from the left end, 41 nucleotides from IRL. A heptameric slippery site beginning at codon 100 of orfA allows a PRF extending into ORFB that encodes a full-length 422 amino acid orfAB transposase. The stop codon of orfB is 130 nucleotides from the right end of the element and 93 nucleotides from IRR (Figure 3.4)**.** Half of the loci, including the reference, have a TSD of 4 nt. Two loci, 2 and 4, have a TSD of 2. Loci 7 and 8 have a TSD of 0 nt. Locus 10 has a TSD of 5 nt. Five of

the ten loci, 2, 4, 6, 8, and 10 are identical, and locus 9 has only a single nucleotide substitution. Locus 7 has a single nucleotide insertion at position 714 (orfB) shifting the reading frame to +1 and disrupting the putative orfAB transposase 210 amino acids short of the carboxy terminus. Locus 5 has a single nucleotide insertion and deletion at positions 504 and 590 respectively, disrupting 29 amino acids of the putative orfAB transposase. Locus 1 has several single nucleotide insertions and deletions distributed throughout orfB, and locus 3 has several similar indels, but is the only element among the ten that has an indel (deletion) in orfA. All ten elements are identical outside the ORFs and locus 7 has a perfect 75 nucleotide tandem duplication of the region between the left end and the orfA 5'-ATG. Additionally, two small fragments of ISHahy4 exist in the genome, one (116 nt) 93% identical to nucleotides 2-118 of the reference copy and the other (75nt) 100% identical to nucleotides 1-75 of the reference copy.

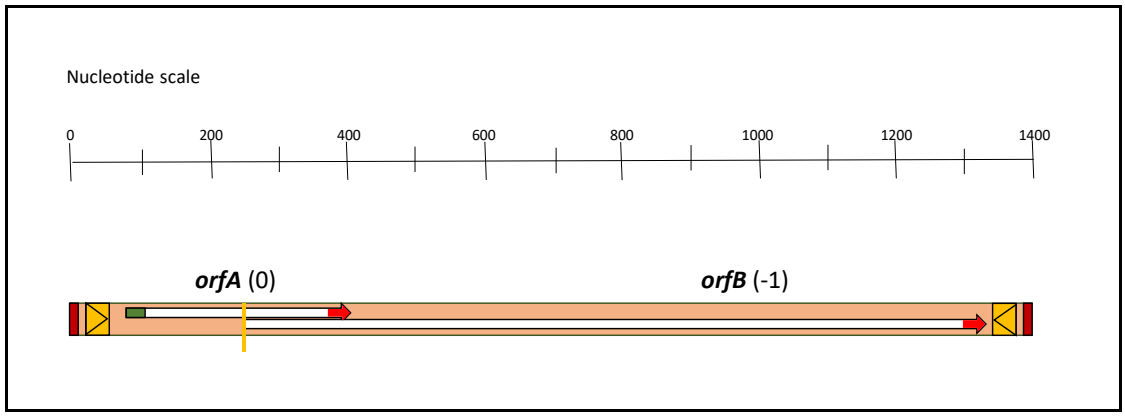

Figure 3.4. ISHahy4 Structure.

3.2.4. ISHahy5. Two copies of ISHahy5 are found in the *H. hydrogeniformans* genome. Only one is full length and the other is a partial element. The structure of the full-length reference element for this group (locus 1) is 1227 nt in length. The element starts

and ends with 5'TG-CA3' and the inverted repeats show weak symmetry (13 of 31 nt). ORFA begins 38 nucleotides from the left end and 6 nucleotides from the IRL. A heptameric slippery site beginning at codon 88 of orfA allows a PRF extending into orfB that encodes a full-length 390 amino acid orfAB transposase. Interestingly, another potential heptameric slippery site begins at codon 96 still within the orfA coding region. The orfA and orfB overlap is 14 codons. The stop codon of orfB is 18 nucleotides from the right end and internal to the IRR (Figure 3.5). Neither of the loci have target site duplications (TSD) in this family. The partial ISHahy5 element is 617 nucleotides in length but is comprised of three non-contiguous portions of the element. The first 159 nucleotides align with positions 36 to 194 of the element starting just after the IRL and containing the start codon of orfA. The second region, 176 nucleotides, corresponds to positions 458 to 635 and contain an internal segment of orfB. The remaining 282 nucleotides align with positions 946 to 1227 containing the orfB stop as well as the IRR.

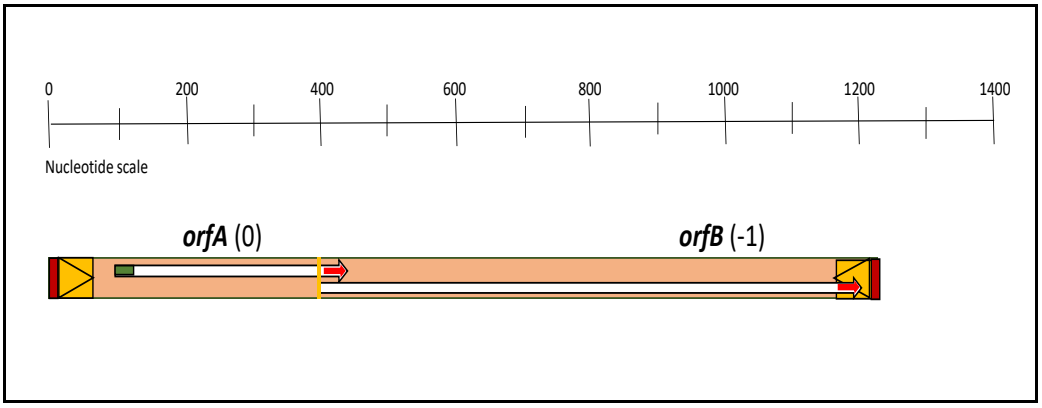

Figure 3.5. ISHahy5 Structure.

3.2.5. Solo Partial Element. A single copy of a partial (724 nucleotides) and fragmented element belonging to the IS3 family was identified by its similarity to ISBce13

in *Bacillus cereus*. This remnant shows 71% positives to 66 amino acids of ORFA, and 68% positives to 164 amino acids of the carboxyl end of orfB in *B. cereus*. Inverted repeats and direct repeats are missing as well as the overlapping region of orfA and orfB.

3.3. ISHAHY6

The genome of *H. hydrogeniformans* contains seven loci that harbor either functional, defective, or fossil remnants of the IS6 family of bacterial insertion sequences. The elements belong to a single ISHahy6 group and are 99% similar at the nucleotide level and greater than 98% similar at the protein level if indels are excluded. ISHahy6 is most closely related to IS6 like elements found in *Halanaerobium congolense* with a 92% similarity at the protein level.

Seven copies of ISHahy6 are found in the *H. hydrogeniformans* genome. Six are full length elements and one is a partial element. The structure of the reference element for this group (locus 7) is 1714 nt in length. The element starts and ends with 5'TG-CA3' and the inverted repeats have a weak symmetry of 14 of 19 nt. An open reading frame begins 160 nucleotides from the left end. A stop codon is 194 nt from the right end of the element and lies outside of the IRR. This ORF encodes a single full-length 453 amino acid IS6 transposase (Figure 3.6). All loci have TSD of 8 nt. The partial does not contain an intact TSD. Two of the loci, loci 3 and 6, are identical to the reference element. Locus 1 is 99% identical at the nucleotide level, but as a result of four indels within the ORF there is premature stop. The ORF of locus 4 is also truncated due to an inserted base at nucleotide 536 of the element. A single deletion at nucleotide 846 of locus 5 causes a frameshift and premature stop as well. The partial ISHahy6 element (locus 2) is comprised of 2 non-

contiguous segments that total 479 nt in length. The first segment is 359 nt in length and corresponds to positions 152 to 511 of ISHahy6. The second segment makes up the other 120 nt and corresponds to positions 1 to 121

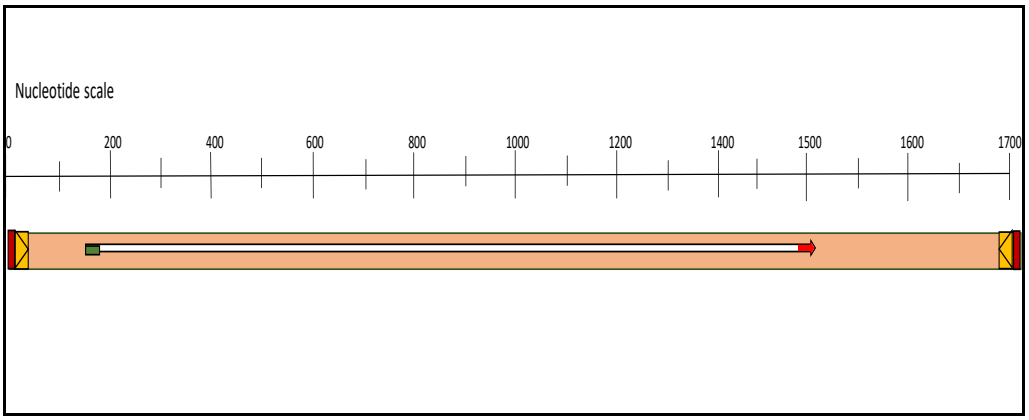

Figure 3.6. ISHahy6 Structure.

3.4. ISHAHY7

The genome of *H. hydrogeniformans* contains 12 loci that harbor either functional, defective, or fossil remnants of the IS256 family of bacterial insertion sequences. The elements belong to a single ISHahy7 group and are greater than 98% similar at the protein level if indels are excluded. ISHahy7 is most closely related (88% positives at protein level) to IS256 like elements found in *Halanaerobium salsuginis.*

Twelve copies of ISHahy7 are found in the *H. hydrogeniformans* genome. Eight loci are full length elements and four are partial elements. The structure of the reference element for this group (locus 3) is 1327 nt in length. The element starts and ends with 5’AG-TC3’ and the inverted repeats have a symmetry of 19 of 28 nt. An open reading frame begins 118 nucleotides from the left end. A stop codon is 28 nt from the right end of the element and lies right next to the IRR. This ORF encodes a single full-length 393 amino

acid IS256 transposase (Figure 3.7). All isoforms have the TSD of 8 nt. Loci 4,5,6,7,10, and 11 are identical to the reference element. Locus 1 is 98% identical at the protein level but is riddled with nucleotide substitutions along with a single deletion at nucleotide 1291 resulting in an ORF extension of 12 nucleotides. The nucleotide substitutions disrupt the inverted repeats resulting in a weaker symmetry of 18 of 30. The first partial IS, locus 2, is a very short sequence of 117 nt and corresponds to positions 1 to 116 of the reference element. The second partial, locus 8, is comprised of 2 non-contiguous segments that total 627 nt in length. The first segment is 121 nt in length and corresponds to positions 2 to 122 of ISHahy7. The second segment makes up the other 506 nt and corresponds to positions 247 to 752 which contains the start of the ORF. The third partial, locus 9, is a single short segment that is 290 nt that correspond to the end of the element from positions 1039 to 1328. The fourth partial, locus 10, is a formed from 3 contiguous segments that total 880 nt in length and make up the first 880 nt of the element.

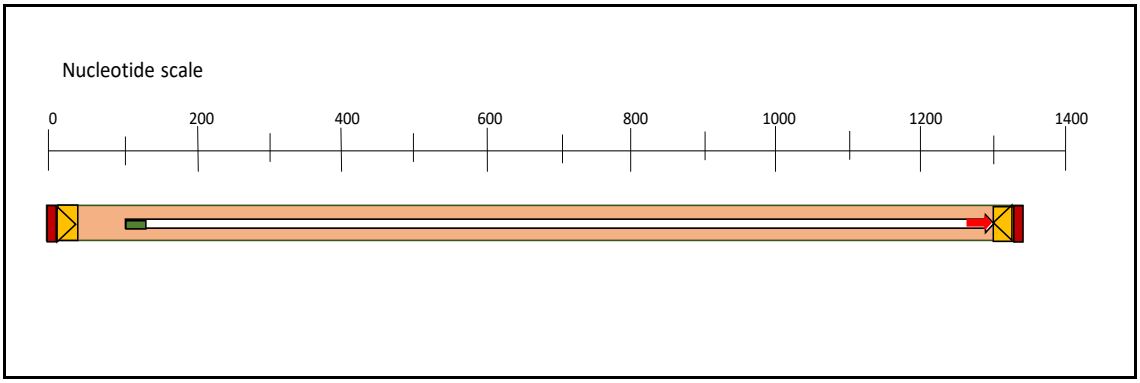

Figure 3.7. ISHahy7 Structure.

3.5. IS30 FAMILY MEMBERS

The genome of *H. hydrogeniformans* contains 12 loci that harbor either functional or defective forms of the IS30 family of bacterial insertion sequences. The elements have been divided into four groups (ISHahy 8, 9, 10, and 11). ISHahy8 is most closely related (90% positives at protein level) to IS30 like elements found in *Halanaerobium congolense.* ISHahy9 is most closely related (95% positives at protein level) to IS30 like elements found in *Halanaerobium saccharolyticum.* ISHahy10 and ISHahy11 are equally related (77% positives at the protein level) to IS30 like elements found in Clostridiales and are 96% similar at the protein level to each other.

3.5.1. ISHahy8. Seven copies of ISHahy8 are found in the *H. hydrogeniformans* genome. All loci are full length elements. The structure of the reference element for this group (locus 1) is 1246 nt in length. The element starts and ends with 5'TT-CC3' and the inverted repeats have a symmetry of 19 of 26 nt. An open reading frame begins 109 nucleotides from the left end. A stop codon is 68 nt from the right end of the element and is outside of the IRR. This ORF encodes a single full-length 356 amino acid IS30 transposase (Figure 3.8). Locus 5 has a TSD of 2 nt while all other loci have no TSD. Locus 2 is 99% identical at the protein level resulting in an isoform. As a result of the two indels located at nucleotide 38 and 267 the ORF is disrupted and truncated resulting in inactivity of this isoform.

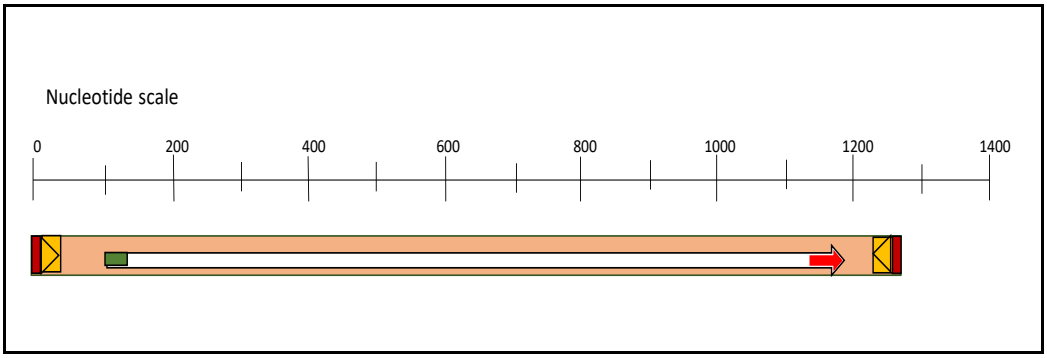

Figure 3.8. ISHahy8 Structure.

3.5.2. ISHahy9. Three copies of ISHahy9 are found in the *H. hydrogeniformans* genome. All loci are full length elements. The structure of the reference element for this group (locus 1) is 1217 nt in length. The element starts and ends with 5'GC-AG3' and the inverted repeats have a symmetry of 19 of 27 nt. An open reading frame begins 122 nucleotides from the left end. A stop codon is 47 nt from the right end of the element and touches the IRR. This ORF encodes a single full-length 349 amino acid IS30 transposase (Figure 3.9). Loci 1 and 3 have a TSD size of 13, locus 2 has a TSD size of 9. Locus 2 is identical to the reference. Locus 3 is 99% identical with a single nucleotide insertion at position 50 of the element.

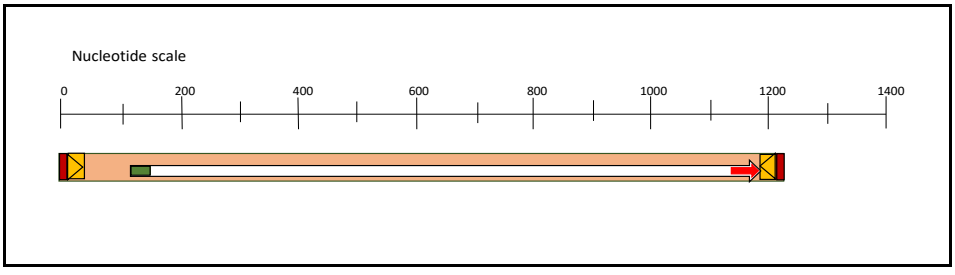

Figure 3.9. ISHahy9 Structure.

3.5.3. ISHahy10. A single copy of ISHahy10 is found in the *H. hydrogeniformans* genome. The structure of the single locus 1220 nt in length. The element starts and ends with 5'GT-AG3' and the inverted repeats have stronger symmetry of 21 of 26 nt. An open reading frame begins 102 nucleotides from the left end. A stop codon is 75 nt from the right end of the element and is outside the IRR. This ORF encodes a single full-length 347 amino acid IS30 transposase (Figure 3.10). The single locus has a TSD of 7 nt.

3.5.4. ISHahy11. A single copy of ISHahy11 is found in the *H. hydrogeniformans* genome. The structure of the single locus 1220 nt in length. The element starts and ends with 5'GT-AG3' and the inverted repeats have stronger symmetry of 20 of 26 nt. An open reading frame begins 104 nucleotides from the left end. A stop codon is 73 nt from the right end of the element and is outside the IRR. This ORF encodes a single full-length 347 amino acid IS30 transposase (Figure 3.10). The single locus has a TSD of 7 nt.

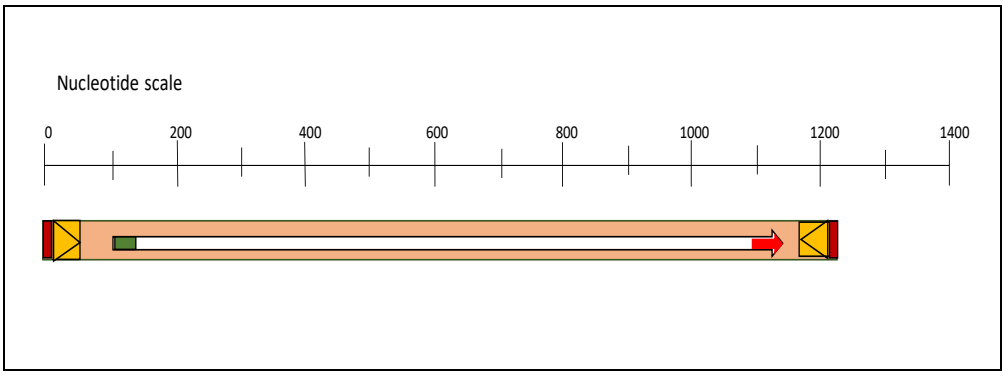

Figure 3.10. ISHahy10 and ISHahy11 Structure.

3.6. ISHAHY12

The genome of *H. hydrogeniformans* contains 4 loci that harbor either functional, defective, or fossil remnants of the IS21 family of bacterial insertion sequences. The

elements belong to a single group labeled ISHahy12. IS21 elements contain two consecutive ORFs called IstA and IstB. Note, that IstB is labeled with an h for helper gene in the appendix table. IstA is most closely related (98% positives at protein level) to IS21 like elements found in *Halanaerobium kushneri.* While IstB is most closely related (99% positives at protein level) to ATPase genes found in *Halanaerobium congolense.*

Four copies of ISHahy12 are found in the *H. hydrogeniformans* genome. Three of the identified loci are full length elements and one is an uncharacterized partial. The structure of the reference element for this group (locus 2) is 2517 nt in length. The element starts and ends with 5'TG-CA3' and the inverted repeats have a symmetry of 17 of 26 nt. The IstA open reading frame begins 351 nucleotides from the left end. The IstA stop codon is 5'TAA3' and ends at nucleotide 1688 nt. This ORF encodes a single full-length 445 amino acid IS21 transposase. The IstB ORF begins at nucleotide 1688 and ends at 2443 outside of the IRR. IstB encodes an ATPase helper gene in the -1 frame of IstA that helps regulate IS21 transposition (Figure 3.11).

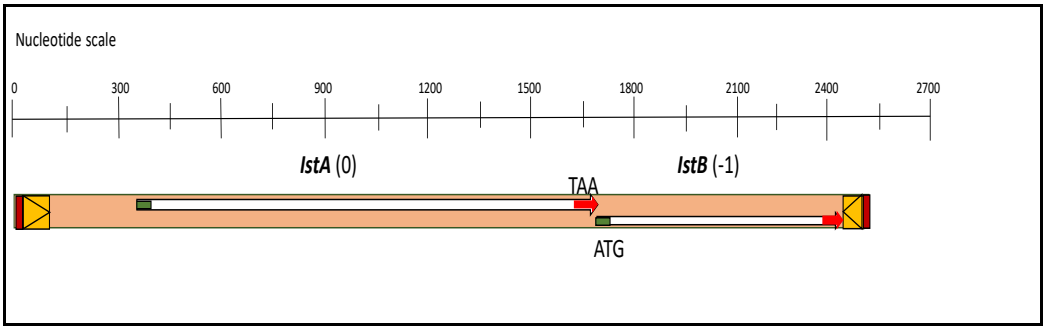

Figure 3.11. ISHahy12 Structure Including IstA and IstB Overlap Site.

All loci have a TSD of 6 nt. Locus 3 is identical to the reference element; however, it has been interrupted by ISHahy7 locus 6 such that the first 103 bases of ISHahy12 locus

3 are upstream of the remaining element (Figure 3.12). Locus 4 contains 19 indels that result in a truncated IstA and IstB ORF. An insertion at position 910 of the element results in a -1 frameshift and premature stop at position 917 resulting in a shortened polypeptide of 188 aa. The IstB ORF begins at position 1689, near identical to the reference, but a series of 16 deletions starting at position 2123 results in an early stop and truncated ORF of 148 aa. The ISHahy12 partial, locus 1, was uncharacterized because of two separate insertion events caused by ISHahy4 locus 2 and the uncharacterized IS110 element.

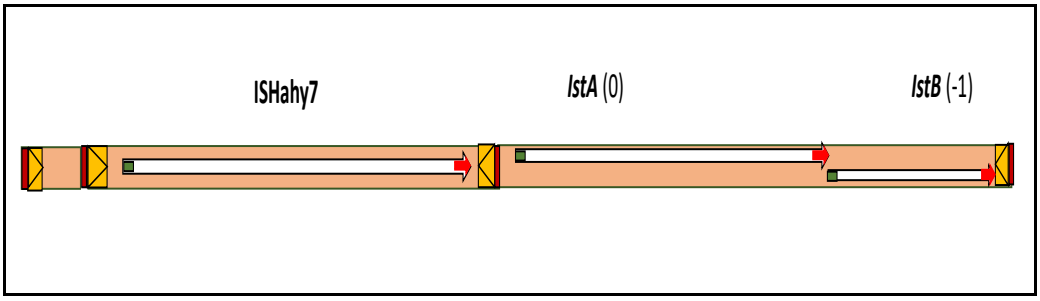

Figure 3.12. ISHahy12 Locus 3 Structure After ISHahy7 Interruption.

3.7. ISNCY FAMILY MEMBERS

The genome of *H. hydrogeniformans* contains 6 loci that harbor either functional or defective transposase belonging to the ISNCY family of bacterial insertion sequences. The elements have been divided into two groups (ISHahy 13 and 14). ISHahy13 is most closely related (76% positives at protein level) to ISNCY like elements found in *Candidatus Desulfosporosinus infrequens*. ISHahy14 is most closely related (95% positives at protein level) to an unidentified transposase found in *Halanaerobium kushneria*

3.7.1. ISHahy13. Four copies of ISHahy13 are found in the *H. hydrogeniformans* genome. All loci are full length elements. The structure of the reference element for this

group (locus 4) is 1497 nt in length. The element starts and ends with 5'TG-CA3' which is common in this family. The inverted repeats have a symmetry of 19 of 24 nt. An open reading frame begins 54 nucleotides from the left end. A stop codon is 101 nt from the right end of the element and is outside of the IRR. This ORF encodes a single full-length 447 amino acid ISNCY transposase (Figure 3.13). All loci have a target size duplication of 17. None of the ISHahy13 elements are identical to the reference. Locus 1 is 99% identical to the reference but two indels (insertions) have disrupted the ORF via a frameshift. The indels are located at positions 748 and 798. If these indels are removed the resulting polypeptide is identical to the reference. Locus 2 is 99% similar to the reference sequence as well with two nucleotide substitutions and three indels (deletions) at positions 229, 237, and 1086 and does not result in a frameshift. Locus 4 is 98% positive at the protein level which is a result of 23 nucleotide substitutions spread evenly across the element.

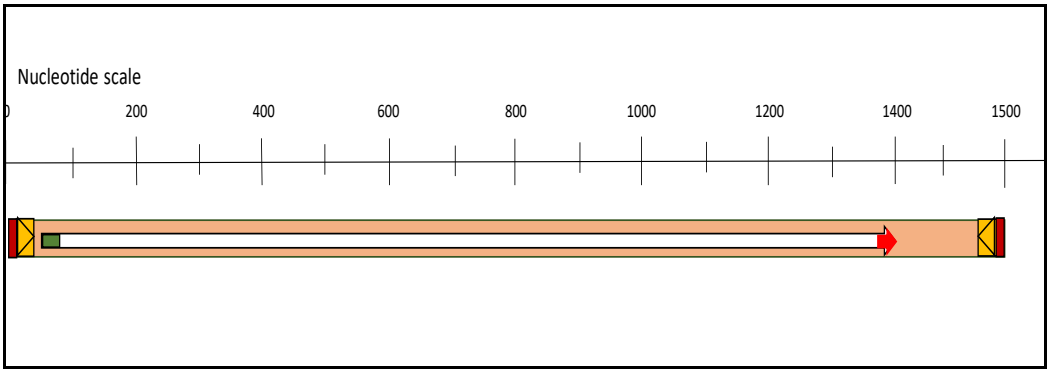

Figure 3.13. ISHahy13 Structure.

3.7.2. ISHahy14. Two full size copies of ISHahy14 are found in the *H. hydrogeniformans* genome. The structure of the reference element for this group (locus 1) is 1739 nt in length. The element starts and ends with 5'TA-TA3' and the inverted repeats have a weaker symmetry of 11 of 16 nt. An open reading frame begins 136 nucleotides

from the left end. A stop codon is 168 nt from the right end of the element. This ORF encodes a single full-length 478 amino acid ISNCY transposase (Figure 3.14). Locus 1 has a TSD of 12 and locus 2 has TSD of 14. Locus 2 is 100% identical at the protein level.

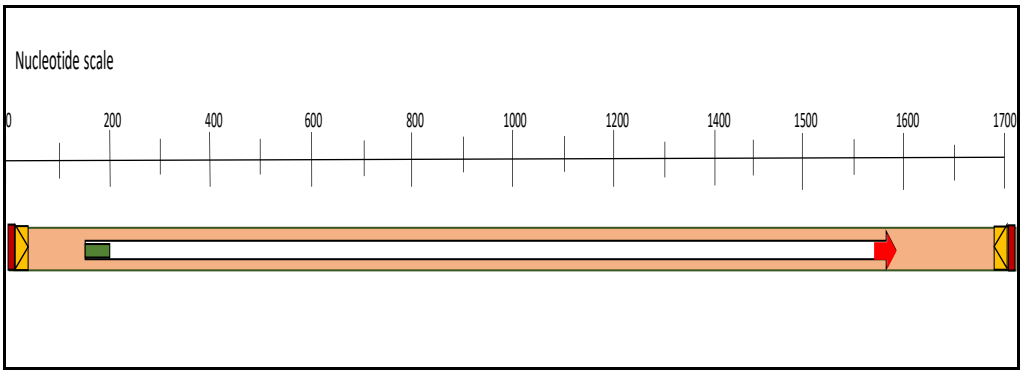

Figure 3.14. ISHahy14 Structure.

3.8. ISHAHY15

The genome of *H. hydrogeniformans* contains 2 loci that harbor a functional or defective transposase belonging to the IS1182 family of bacterial insertion sequences. The elements belong to a single group labeled ISHahy15. ISHahy15 is most closely related (84% positives at protein level) to IS1182 like elements found in *Halanaerobium salsuginis*.

Two full size copies of ISHahy15 are found in the *H. hydrogeniformans* genome. The structure of the reference element for this group (locus 1) is 1920 nt in length. The element starts and ends with 5'GG-CC3' and the inverted repeats have a strong symmetry of 26 of 27 nt. An open reading frame begins 166 nucleotides from the left end. A stop codon is 210 nt from the right end of the element. This open reading frame (ORF) encodes a single full-length 514 amino acid IS1182 transposase (Figure 3.15). Neither loci have a

target site duplication (TSD). Locus 2 contains a single indel resulting in a frameshift and truncated ORF.

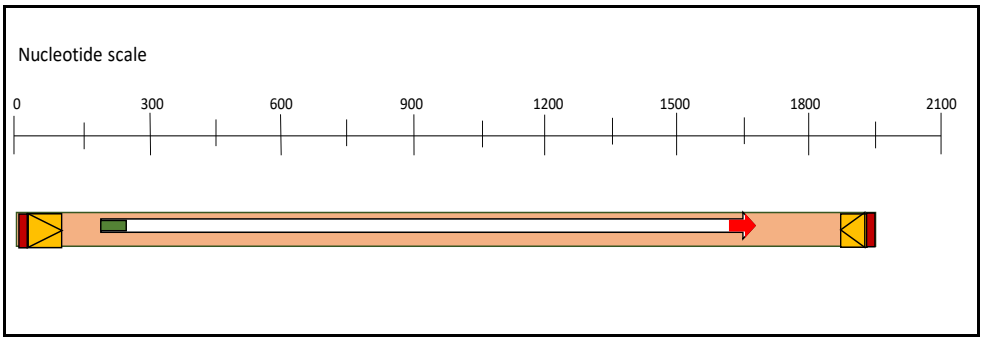

Figure 3.15. ISHahy15 Structure.

4. DISCUSSION

4.1. INSERTION SEQUENCE IDENTIFICATION

All transposable elements in *Halanaerobium hydrogeniformans*, excluding members of the IS200/605 family, were characterized in detail to complete a genome wide survey of the insertion sequences found in this organism. The IS200/605 family members were previously studied by a former Missouri S&T graduate student, Mike Sadler. Among the remaining elements a total of 99 elements belonging to 15 unique insertion sequences were identified. These 15 unique insertion sequences represent nine different insertion sequence families (Table 4.1).

All the insertion sequences discovered in this project belong to the DDE group of insertion sequences, utilizing the conserved triad of D(asp)D(asp)E(glu) to transpose (Figure 4.1). The distance between each residue is highly variable between each family, but within families it is highly conserved.

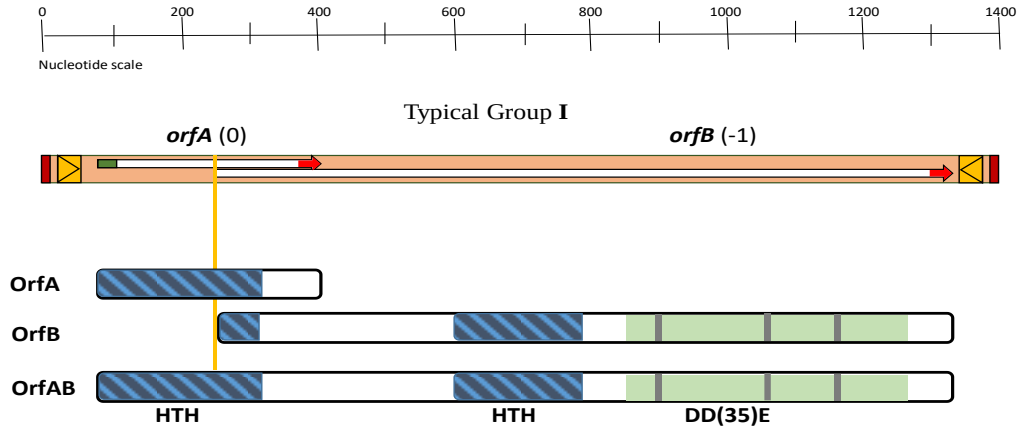

Figure 4.1. Catalytic DDE Structure of an IS3 Family Member.

Each identified element was mapped on the genome relative to other insertion sequences, and other genes (Figure 4.2). It does not appear that the insertion sequences have a strand bias; however, several families have preferential insertion near relative elements of the same family, such as the IS3 and IS30 families. These families transpose via an inverted repeat – inverted repeat junction (IR-IR) intermediate and insert within or nearby sequences that resemble their own terminal inverted repeats. It does appear that this type of preferential insertion exists in the *H. hydrogeniformans* genome.

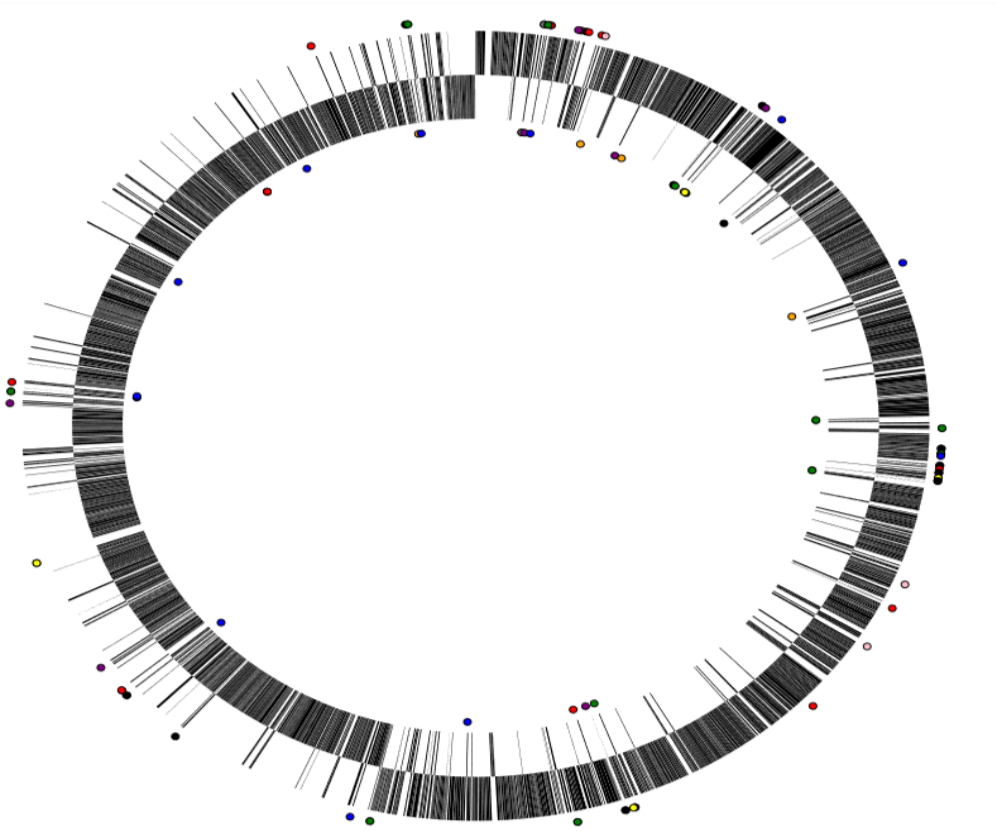

Figure 4.2. Genome Map of Insertion Sequences. Insertion sequences are shown as the different colored dots above (positive strand) and below (negative strand) genes located across the genome.

The misidentification of several IS21 and IS91 elements in *H. hydrogeniformans* by the ISsaga program highlights the need for more developed automated insertion sequence annotation programs, and the importance of manual curation for the identification of insertion sequences. It also indicates the limits of library-based annotation software. IS21 helper genes are common false-positives because they are ATPase-like genes. The misidentified IS21 elements were deemed false-positives after checking that they were helper genes with no nearby IS21 transposase and that they only aligned with an ATP DNA binding domain. Similarly, the IS91 elements were also deemed false-positives. The IS91 elements were both solo accessory genes that ISSaga annotated as integrase, qualities of an IS91 false-positive [36]. A single IS110 element was identified by ISsaga as being present in the genome. We were unable to describe this element because there are no other IS110 elements in this organism to compare to and the IRs and TSD could not identified. Searches to find similar IS110 elements in another organism were inconclusive as well. This solo IS110 is severed by the interactions of ISHahy4 locus 2 and ISHahy12 locus 1 in the genome.

4.1.1. ISHahy2. ISHahy2 belongs to the IS3 family of insertion sequences**.** Eleven copies of ISHahy2 were identified in the genome. This insertion sequence is similar to other identified IS3 elements [13]. ISHahy2 is 1212 nt long, within the reported range of 1200 to 1550 nt. ISHahy2, like other IS3 family members [37,38,39], terminates in 5’-TG-CA-3’ and contains two separate overlapping open reading frames (ORFs) that form a single orfAB by programmed ribosomal frameshift (PRF). Another characteristic of IS3 family members is a large IR match score, which ISHahy2 had (29/41) as well. ISHahy2

contains mostly full-size loci: some that harbor functional genes and some that have become defective.

Table 4.1. Identified ISHahy Elements and Their Corresponding Families.

IS Name	IS Family	Number of Loci
ISHahy2	IS3	11
ISHahy3	IS3	3
ISHahy4	IS3	10
ISHahy5	IS3	2
ISHahy6	IS6	7
ISHahy7	IS256	12
ISHahy8	IS30	7
ISHahy9	IS30	3
ISHahy10	IS30	1
ISHahy11	IS30	1
ISHahy12	IS21	4
ISHahy13	ISNCY	4
ISHahy14	ISNCY	2
ISHahy15	IS1182	2

The size of the ISHahy2 family and similarity between isoforms is characteristic of an expansion in the invasion-expansion-extinction cycle. After a new transposase invades a foreign genome it will begin to expand throughout the genome by transposition. Mutations and other insertion/deletion (indel) events will begin to occur, disturbing and disrupting the ORF and render the element defective. As mutations and indel events continue, the element can deteriorate into a partial element or fossil of a prior insertion event. DDE family transposases show strong preference for cis action [10], meaning the ORF must be intact for an element to transpose. Strong cis action increases selective pressure against elements with disrupted, or otherwise non-functional protein encoding ORFs because these elements have a reduced ability to replicate. Most of the elements in

this family are identical or near identical and have mostly intact ORFs. Therefore, it appears the ISHahy2 elements are in the expansion part of the cycle.

As mentioned earlier, another common occurrence is the IR-IR junction intermediate that forms in IS30 and IS3 family members. This makes them more likely to insert near other family members, and more so if they have the same or related IR. This can also occur between IS30 and IS3 family members as can be seen by the insertion of ISHahy11 (IS30 family) into ISHahy2 locus 1 (IS3 family). This insertion event separated the last 32 nt of ISHahy2 locus 1 from the rest of the element. The 32 nt were not found beyond the ISHahy11 because of a different IS3 insertion event. However, if the IRL (inverted repeat left) and IRR (inverted repeat right) were intact, possible transposition could occur moving the entire IS3 and IS30 complex. The ISHahy2 elements have various target site duplications (TSD). Two elements have a TSD of 0, one element had a TSD of 4, and the remaining loci had TSD of 3.

4.1.2. ISHahy3. ISHahy3 belongs to the IS3 family of insertion sequences. Three copies of ISHahy3 were identified in the genome, two of them full size and the third a partial. ISHahy2 is 1360 nt in length, terminating in 5'-TG-CA-3', and harboring two separate ORFs that form the single orfAB. However, ISHahy3 members have a low IR match score of 9/16. If a single nucleotide gap is allowed the IR score increases to 18/27. An indel may have occurred resulting in a disrupted IR which could weaken or prevent transposase binding. The second locus of ISHahy3 has many nucleotide substitutions and a single deletion preventing the normal programmed -1 ribosomal frameshift from occurring. This likely means that a transposase is not encoded by this element. The small size of the ISHahy3 family and fragmentation of two of the three loci is representative of

the extinction phase in the invasion-expansion-extinction cycle. A single ISHahy3 element had a TSD of 4, the other loci had a TSD of 0. The partial did not contain a TSD.

4.1.3. ISHahy4. ISHahy4 belongs to the IS3 family of insertion sequences. Ten copies of ISHahy4 were identified in the genome, eight of them full size and two that are partials. ISHahy4 is larger in size compared to other IS3 elements found in *H. Hydrogeniformans*. ISHahy4 is 1411 nt in size and terminates in 5'-CT-CG-3'. IShahy4 elements also have weaker IR match scores compared to ISHahy2. Five of the ten loci are identical copies, while locus 9 only has a single nucleotide substitution. These loci likely encode a full-size transposase. The remaining loci are plagued by varying degrees of insertions and deletions resulting in frameshifts and disrupted ORFs. However, all ten elements are identical outside the ORFs and locus 7 has a perfect 75 nucleotide tandem duplication of the region between the left end and the orfA ATG. Like the ISHahy2, the large size of the ISHahy4 family is representative of a family exhibiting expansion, although the degree of fragmentation and presence of partials may mean this family is finished expanding and beginning the extinction phase due to selection pressures. There is high variation in the TSDs of the ISHahy4 elements. Two loci have a TSD of 0, two loci have a TSD of 2, 5 loci have a TSD of 4, and a single locus has a TSD of 5.

4.1.4. ISHahy5. ISHahy5 belongs to the IS3 family of insertion sequences. Two copies of ISHahy5 were identified in the genome, one full size and the other a partial. ISHahy5 is 1227 nt in length, terminating in 5'-TG-CA-3', and containing two separate ORFs that form the single orfAB. This is like other IS3 elements described by Mahillon and Chandler [13]. However, ISHahy5 members had one of the lowest IR match scores of 13/31. The second locus of ISHahy5 is a partial comprised of three non-contiguous portions

of the element. This partial contains portions of both orfA and orfB and is likely a fossil of a prior insertion event. The deteriorated partial indicates that this family may be undergoing extinction. Neither of the IShahy5 elements contain a TSD.

4.1.5. ISHahy6. IShahy6 belongs to the IS6 family of insertion sequences. Seven copies of ISHahy6 were identified in the genome, six elements are full size and the other is a partial. ISHahy6 is 1714 nt in length, which is twice the size (789 nt to 880 nt) of other identified IS6-like elements [33]. The ISHahy6 IRs were like other IS6 elements with a match score of 14/19 and the ISHahy6 TSD size was the same as reported in the literature with a TSD of 8 nt [13]. As a result of indels and substitutions, three of the elements are truncated and result in a premature stop. The partial copy of ISHahy6 is comprised of two non-contiguous portions of the element. ISHahy6 appears to have been in the genome for a significant amount of time, as evidenced by seven copies, However, this family now appears to be undergoing slow extinction as evidenced by the deteriorating isoforms and the non-contiguous partial.

4.1.6. ISHahy7. ISHahy7 belongs to the IS256 family of insertion sequences. Twelve copies of ISHahy7 are found in the *H. hydrogeniformans* genome. Eight of the loci are full length elements and four are partial elements. The length of this family (1327 nt) is similar to IS256 elements described by Mahillon and Chandler [13] and the IR have a somewhat lower match score of 19 nt compared to 24 to 41 nt. Some ISHahy7 members have a TSD of 8 nt which matches well to other IS256 elements reported elsewhere [29]. All isoforms, excluding locus 1, are identical. Locus 1 is full size, but as a result of substitutions and a single deletion the ORF is extended, most likely disabling the transposase. Of the four partials, two are very short in length and do not contain any portion

of the ORF. The other two partials are made up of contiguous or non-contiguous segments that contain at least part of the ORF. It is most likely that this family has begun to enter the extinction cycle as illustrated by the varying degrees of partials and isoforms.

4.1.7. ISHahy8. ISHahy8 belongs to the IS30 family of insertion sequences. Seven copies of ISHahy8 are found in the *H. hydrogeniformans* genome. All seven loci are full length elements. The length of this family (1327 nt) is larger than other IS30 family members. Also, ISHahy8 does not end in the dinucleotide sequence CA-3' that is common in IS30 elements in other genomes [30]. The IRs are similar to those described in the literature. Locus 5 had a TSD of 2 nt which is seen in the literature; however, all other loci do not have a TSD. This could be a result of recombination that has occurred between ISHahy8 members resulting in a mismatch of the TSD. A single copy of ISHahy8 (locus 2) has two indels that result in a frameshift and defective ORF. The lack of partials and relative similarity between isoforms likely means this family is not undergoing extinction yet but is probably in the expansion phase based on the number of copies. It should be noted that this family most commonly uses the copy and paste mechanism of transposition.

4.1.8. ISHahy9. ISHahy9 belongs to the IS30 family of insertion sequences. Three full size copies of ISHahy9 are found in the *H. hydrogeniformans* genome. The length of this family (1217 nt) is the same as previously identified IS30 family members in other genomes [30]. Like ISHahy8, ISHahy9 also does not end in the dinucleotide sequence CA-3' that is common in other IS30 family members. The IR score for ISHahy9 is the same as ISHahy8. However, the ISHahy9 target site duplications are much larger at 13 and 9. ISHahy9 locus 3 contains a single insertion, resulting in a frameshift. However, this frameshift does not cause a premature stop and a full size ORF is still produced, although

it is unclear whether this ORF encodes a transposase or not. The ISHahy9 is likely a relatively recent invasion event and is in the process of expanding.

4.1.9. ISHahy10 and ISHahy11. ISHahy10 and ISHahy11 belong to the IS30 family of insertion sequences. There is a single copy each of ISHahy10 and ISHahy11 in the genome. ISHahy10 and ISHahy11 are only 96% similar at the protein level and therefore below the threshold of 98% similarity deemed by ISSaga as necessary to be considered isoform elements. Both elements are equal in length and the same size as IS30-like elements described elsewhere [13]. Like other IS30 members found in *H. hydrogeniformans*, neither ISHahy10 or ISHahy11 terminate in the CA-3' that is common in IS30 elements found outside of *H. hydrogeniformans* [30]. ISHahy10 and ISHahy11 differ slightly in IRs with ISHahy10 having a higher score of 21/26 compared to ISHahy11 which has an IR score of 20/26. This difference is caused by a single nucleotide substitution (thymine to a guanine) in ISHahy11 at position 2. A single insertion and deletion within the ORF results in two separate elements, even though they remain the same size and encode the same size transposase. It is unknown which element inserted first in a new invasion event. However, it is most likely that after a single transposition event the isoform diverged via mutations causing the isoform to become its own unique element. To discern the order of events, it must be determined whether the ISHahy10 and ISHahy11 ORFs encode viable transposases.

4.1.10. ISHahy12. ISHahy12 belongs to the IS21 family of insertion sequences and contains two ORFs (IstA and IstB). The ORFs overlap such that the terminating adenine of the IstA stop codon, TAA or TGA, is the starting adenine for the ATG start codon of IstB in the -1 frame. IstA encodes the transposase, while IstB is an ATPase like helper gene

that is not required for transposition. The *H. hydrogeniformans* genome contains four copies of ISHahy12, three of which are full length, and one is an uncharacterized partial. The ISHahy12 elements had much larger IRs and are much longer than any of the other IS found in *H. hydrogeniformans* at 2517 nt. Interestingly, three of the four copies of ISHahy12 have been affected by insertions or mutation events. Locus 3 has been interrupted by ISHahy7 locus 6 such that the first 103 bases are upstream of the remaining element. However, both ORFs and IRs are intact and identical to the reference, such that the transposase could be encoded and have a binding recognition site. This could result in the transposition of both the ISHahy12 and ISHahy7 found within. Locus 4 contains many indels and results in two truncated ORFs such that neither IstA or IstB encodes a viable product.

The partial element, locus 1, could not be described fully due to two separate insertion events by the unidentified IS110 element and ISHahy4 locus 2. We are unsure in what order the transposition events occurred, but the ISHahy4 element was identified as an uninterrupted isoform suggesting it was likely the last transposition event at this location. A possible order of the transposition events is that the ISHahy12 inserted at this locus, either as a new invasion or as an expansion. Then the IS110 inserted within the ISHahy12, and finally the ISHahy4 locus 2 inserted into the IS110. There are no sequences upstream or downstream of the ISHahy4 locus 2 that resemble other characterized ISHahy12 elements, nor were any significant IRs or TSDs identified. This means that the ISHahy4 locus 2 did not directly insert into the ISHahy12 locus 1. Instead the IS110 inserted first and the split ISHahy12 was further dissected via insertion events and other mutations making it unrecognizable.

4.1.11. ISHahy13 and ISHahy14. ISHahy13 and ISHahy14 belong to the non-classified group of insertion sequences designated ISNCY. This group includes insertion sequences whose nucleotide sequence is not well-known and insufficient for family assignment or elements whose entire nucleotide sequence is known but show no significant relationship with more than one other element. There were four full length ISHahy13 elements and two full length ISHahy14 elements identified in the *H. hydrogeniformans* genome. There is a considerable difference between the nucleotide length of the two elements, ISHahy13 is 1497 nt long while ISHahy14 is larger containing 1739 nt. The IRs and TSD are also significantly different between the two elements. There does not appear to be any significant similarities between the ISHahy13 and ISHahy14 elements found in *H. hydrogeniformans*. These two element groups were very likely the result of separate invasion events. ISHahy13 contains more loci as well as a greater variation in ORF sequence, which could be evidence that ISHahy13 has been in the genome longer. ISHahy13 loci 2 and 3 contain many insertion and deletions that result in ORF fragmentation. Loci four is an exact copy of Locus one. There are only two copies of ISHahy14 found in the genome, both of which encode a full-size transposase. This could be the result of a more recent invasion

4.1.12. ISHahy15. Two copies of ISHahy15 were found in the *H. Hydrogeniformans* genome. ISHahy15 has the second largest nucleotide sequence of 1920 nt, which falls within the nucleotide range of other identified IS1182 members (1330 to 1950 nt). The ISHahy15 elements had the strongest IR symmetry of all ISHahy elements with a score of 26/27. A single indel in locus 2 results in a frameshift and disrupted. ORF.

Since there are only two copies of ISHahy15 in the genome, it appears that this transposase invaded recently compared to other larger families or is highly regulated.

4.2. ORF DISRUPTION

Although there are preferential insertion sites and target sequences that control transposition, most transposition events result in a random insertion across the genome. These random insertions could insert within a functional gene rendering it inactive. Most ISHahy elements were checked for ORF disruption via excising the insertion sequence from the genome and then searching for an intact ORF. No significant ORF disruptions were identified, other than the two insertion events discussed in previous sections. This is not surprising since the random insertion within essential genes would likely reduce fitness; therefore, preventing the insertion event from being transferred to subsequent generations. Some insertion sequences and genomes, such as the IS3 and IS30 family members, contain regulatory features that prevent insertion sequences from destroying the host genome.

4.3. SYNONYMOUS SUBSTITUTION RATES

To identify the order of invasion of the IS30 family of elements (ISHahy8, 9, 10, 11), a synonymous substitution rate analysis was performed. Our hypothesis was that the earlier the invasion event the more synonymous substitutions would have to accumulate in the transposase. Synonymous substitutions are neutral to fitness because the resulting amino acid (aa) does not change. Results were inconclusive as the IS30 family members were too similar to be able to measure differences in the substitution rates.

4.4. PHYLOGENETIC ANALYSIS

Phylogenies were constructed for the IS3 family elements and the IS30 family elements to analyze the order of transposition events. Plotted in the phylogeny are the reference sequence of each unique ISHahy and any of the sequences that differ from the reference. The closest ancestor to each reference sequence was included in the phylogeny as well.

4.4.1. IS30 Family Phylogeny. The IS30 family phylogeny indicates that these elements resulted from three separate invasion events. The three IS30 family elements separate into three clades (Figure 4.3). The ISHahy8 elements are closely related to IS30-like elements found in *Selenihalanaerobacter shriftii*. The ISHahy9 elements are sister clades with the ISHahy10 and ISHahy11 elements. The ISHahy9 elements are closely related to IS30-like elements found in *Orenia marismortui*. The ISHahy10 and ISHahy11 elements are closely related to IS30-like elements found in *Clostridiales* species. It should be noted that like the IS3 phylogeny, the IS30 elements are all highly related to IS30-like elements found in *H. congolense*, but these were comparisons were excluded from this phylogeny to highlight differences between the elements.

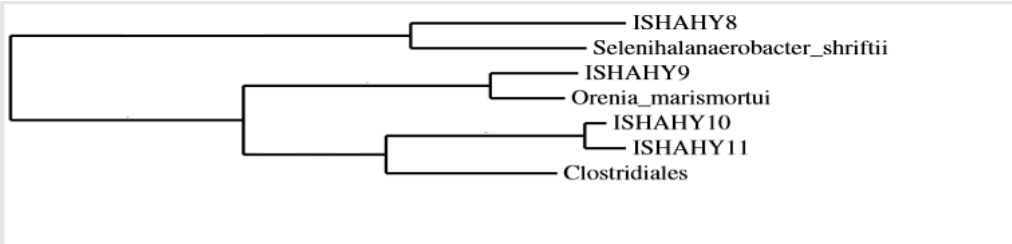

Figure 4.3. IS30 Phylogeny.

4.4.2. IS3 Family Phylogeny. From the phylogeny of the IS3 family it appears that each unique ISHahy element inserted via its own invasion event, as can be seen by the separation into five clades (Figure 4.3). The ISHahy 2 elements are sister clades with the ISHahy5 and both are closely related to different insertion sequences found in *Halanaerobium congolense*. Similarly, ISHahy3 and ISHahy4 are sister clades and similar to different elements in *H. congolense* as well. This close relationship to *H. congolense* is seen across many of the ISHahy elements and may be indicative that *H. hydrogeniformans* and *H. congolense* diverged from one another relatively recently in their evolution history. Since so many insertion sequences are closely related *to H. congolense* we included the next closest relative in the phylogeny. It's this result that suggests that the 5 element groups originated as separate events. The solo partial is most closely related to an insertion sequence found in *Bacillus cereus*.

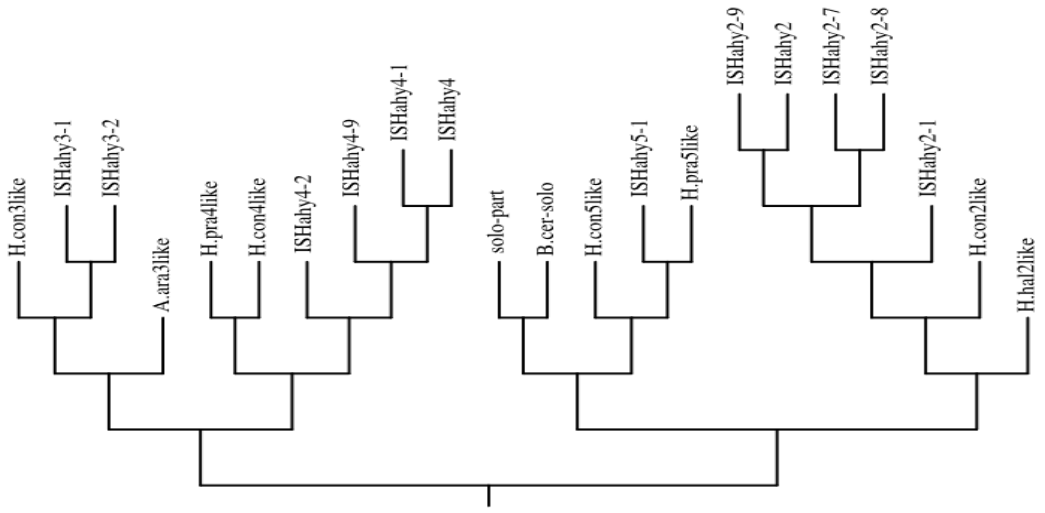

Figure 4.4. IS3 Phylogeny.

4.5. LEADING/LAGGING STRAND BIAS

Each insertion sequence was analyzed as to whether it resides on the leading or lagging strand to determine if there was any strand bias. The leading strand refers to the continuous replication that occurs in the 5' -> '3 direction via DNA polymerase. The lagging strand refers to the discontinuous replication of Okazaki fragments in the 3' -> 5' direction as indicated in Figure 4.5. From the origin, assuming that the first nucleotide is near the origin, two forks will separate and replicate in opposite directions meeting at the middle of the genome, at approximately 1.3 million base pairs. In the clockwise direction the insertion sequences found on the positive strand will be templates for the lagging strand, while insertion sequences on the minus strand will be templates for the leading strand. In the counterclockwise direction, insertion sequences on the positive strand will be the leading strand template and insertion sequences on the minus strand will be lagging strand templates. ISHahy4 shows partial strand bias with 60% on the lagging strand template. Both ISHahy5 loci are also on the lagging strand. The IS30 family appears to show strand bias with 10 of the 12 loci appearing on the lagging strand. However, it is difficult conclude if there is significant strand bias in this organism or within families because the sample size is too small to be statistically accurate.

4.6. CONCLUSION

An extensive study of insertion sequences found in the *H. hydrogeniformans* genome revealed a total of 15 unique ISHahy elements scattered among 72 loci around the genome. Many of these elements are full size and could produce an active transposase; however, many defective and partial elements were also identified. This study provided

significant contributions to the world-wide database of insertion sequences. These are the first elements to be described in a *Halanaerobium* species and submitted to ISfinder. All future elements with 98% similarity will have the ISHahy designation.

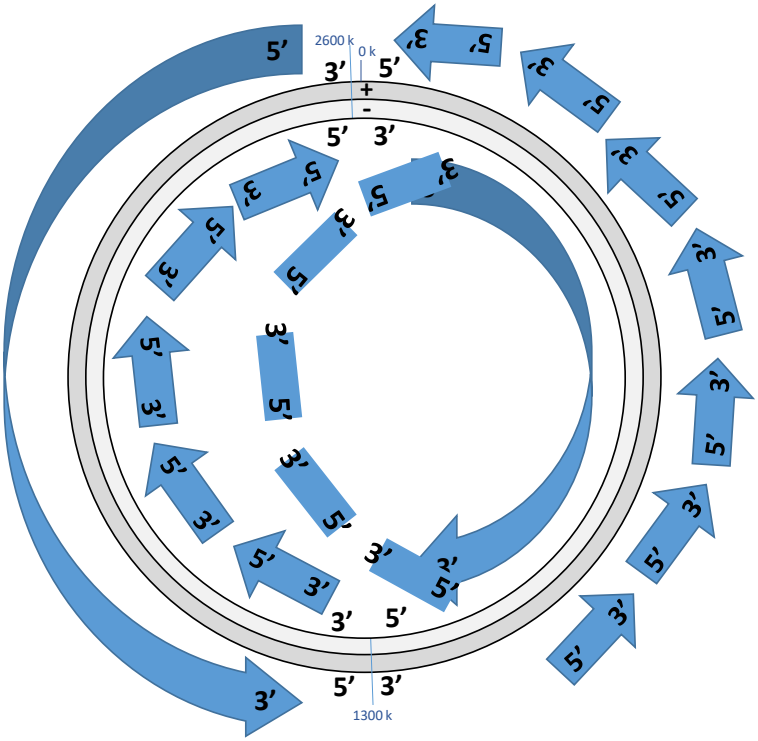

Figure 4.5. How elements were assigned to the leading or lagging strand. The leading strand is shown by the continuous arrow going in the 5' – 3' direction. The lagging strand is shown as Okazaki fragments.

4.7. FUTURE DIRECTIONS

The results presented here explore interesting insertion sequence activity within *H. hydrogeniformans*. However, they only provide a snapshot of the activity at the moment the isolate was taken. While there is evidence indicating potential element transposition,

direct experimental evidence for insertion sequence transpsition is absent. Future directions for research could address these issues. One approach requires the growth of the original isolate over many generations followed by digestion, amplification and sequencing of the ISHahy elements. If *H. hydrogeniformans* has active transposition, using this technique would provide evidence as to which insertion sequences have transposed and roughly how many generations it took for this event to occur. A second approach would be to collect and sequence other *H. hydrogeniforman* isolates from the same environment and compare them to the ISHahy identified in this isolate using PCR analysis.

As shown in the phylogenies, other *Halanaerobium* species contain very similar insertion sequences to the ISHahy. Future research could include a comparison of all ISHahy elements to those in the closely related *H. congolense* or *H. saccaralyticum* including genome location. However, this will have to wait until these other *Halanaerobium* species have fully sequenced genomes. Initial observations from the contigs show that *H. congolense* and other *Halanaerobium* species contain highly similar ORFs to those described here (approximately 90% nucleotide similarity). Comparison of the ISHahy element to other species would help us understand the origins of these elements, their continued activity in genomes, and the manner in which they degenerate.

APPENDIX A.

INSERTION SEQUENCE TABLE

IS Name	Locus	family	orf_L	orf_R	orf_bp	aas	ori (+/-)	IS_L	IS_R	length	DR	IR	IS length(Partials)	lead/lag template
ISHahy 2	1p	IS3 ssgr IS150	64692	64964	273	90	+	-	-	1182	-	-	64608(1)-65788(1181)	lag
ISHahy 2	1a-p	IS3 ssgr IS150	65006	65806	801	266	+	-	-	1182	-	-	64608(1)-65788(1181)	lag
ISHahy 2	2	IS3 ssgr IS150	260425	259298	1128	375	-	260493	257381	1213	3	27/41	-	lead
ISHahy 2	3	IS3	644987	644109	879	292	-	645303	644092	1212	3	29/41	-	lead
ISHahy 2	3a	IS3 ssgr IS150	645235	644948	288	95	-	645303	644092	1212	3	29/41	-	lead
ISHahy 2	4	IS3	655977	656855	879	292	+	655661	656872	1212	4	29/41	-	lag
ISHahy 2	4a	IS3 ssgr IS150	655729	656016	288	95	+	655661	656872	1212	4	29/41	-	lag
ISHahy 2	5	IS3	682116	682994	879	292	+	681800	683011	1212	0	29/41	-	lag
ISHahy 2	5a	IS3 ssgr IS150	681868	682155	288	95	+	681800	683011	1212	0	29/41	-	lag
ISHahy 2	6	IS3	715540	714662	879	292	-	715856	714645	1212	3	29/41	-	lead
ISHahy 2	6a	IS3 ssgr IS150	715788	715501	288	95	-	715856	714645	1212	3	29/41	-	lead
ISHahy 2	7	IS3	1158485	1157358	1128	375	-	1158553	1157341	1213	3	29/41	-	lead
ISHahy 2	8	IS3	1215113	1215946	834	277	+	1214752	1215963	1212	3	29/41	-	lag
ISHahy 2	8a	IS3 ssgr IS150	1214820	1215107	288	95	+	1214752	1215963	1212	3	29/41	-	lag
ISHahy 2	9	IS3 ssgr IS150	1401964	1402842	879	292	+	1401648	1402859	1212	3	27/41	-	lead
ISHahy 2	9a	IS3 ssgr IS150	1401716	1402003	288	95	+	1401648	1402859	1212	3	27/41	-	lead
ISHahy 2	10	IS3	1995557	1996435	879	292	+	1995241	1996452	1212	3	29/41	-	lead
ISHahy 2	10a	IS3 ssgr IS150	1995309	1995596	288	95	+	1995241	1996452	1212	3	29/41	-	lead
ISHahy 2	11	IS3	2552844	2553722	879	292	+	2552528	2553739	1212	0	29/41	-	lead
ISHahy 2	11a	IS3 ssgr IS150	2552596	2552883	288	95	+	2552528	2553739	1212	0	29/41	-	lead
ISHahy 3	1	IS3 ssgr IS150	56281	55334	948	315	-	56619	55260	1360	4	9/16	-	lead
ISHahy 3	1a	IS3 ssgr IS150	56547	56260	288	95	-	56619	55260	1360	4	9/17	-	lead
ISHahy 3	2	IS3 ssgr IS150	61440	61739	300	99	+	61368	62726	1359	0	9/16	-	lag
ISHahy 3	2a	IS3 ssgr IS150	61576	62652	1077	358	+	61368	62726	1359	0	9/17	-	lag
ISHahy 3	3-p	IS3 ssgr IS3	-	-	-	-	+	-	-	-	-	-	67055()-67606()	lag
ISHahy 4	1	IS3 ssgr IS407	257963	256992	972	323	-	258332	256922	1411	4	11/16	-	lead
ISHahy 4	1a	IS3 ssgr IS407	258259	257948	312	103	-	258332	256922	1411	4	11/16	-	lead
ISHahy 4	2	IS3 ssgr IS51	275443	275754	312	103	+	275370	276680	1411	2	11/16	-	lag
ISHahy 4	2a	IS3 ssgr IS150	275739	276710	972	323	+	275370	276680	1411	2	11/16	-	lag
ISHahy 4	3	IS3 ssgr IS407	339801	338830	972	323	-	340170	338760	1411	4	11/16	-	lead
ISHahy 4	3	IS3 ssgr IS51	676624	676935	312	103	+	676551	677961	1411	2	11/16	-	lag
ISHahy 4	4a	IS3 ssgr IS150	676920	677891	972	323	+	676551	677961	1411	2	11/16	-	lag
ISHahy 4	5	IS3 ssgr IS51	694059	694370	312	103	+	693986	695376	1411	4	11/16	-	lag
ISHahy 4	5a	IS3 ssgr IS150	694355	695326	972	323	+	693986	695376	1411	4	11/16	-	lag
ISHahy 4	6	IS3 ssgr IS51	701588	701899	312	103	+	701515	702925	1411	4	11/16	-	lag
ISHahy 4	6a	IS3 ssgr IS150	701884	702855	972	323	+	701515	702925	1411	4	11/16	-	lag
ISHahy 4	7	IS3 ssgr IS51	710047	710358	312	103	+	709974	711383	1410	0	11/16	-	lag
ISHahy 4	7a	IS3 ssgr IS407	710342	711313	972	323	+	709974	711383	1410	0	11/16	-	lag
ISHahy 4	8	IS3 ssgr IS51	1170463	1170774	312	103	+	1170390	1171800	1411	0	11/16	-	lag
ISHahy 4	8a	IS3 ssgr IS150	1170759	1171730	972	323	+	1170390	1171800	1411	0	11/16	-	lag
ISHahy 4	9	IS3 ssgr IS51	1658139	1658450	312	103	+	1658066	1659476	1411	4	11/16	-	lead
ISHahy 4	9a	IS3 ssgr IS150	1658435	1659406	972	323	+	1658066	1659476	1411	4	11/16	-	lead
ISHahy 4	10	IS3 ssgr IS51	2550619	2550930	312	103	+	2550546	2551956	1411	5	11/16	-	lead
ISHahy 4	10a	IS3 ssgr IS150	2550915	2551886	972	323	+	2550546	2551956	1411	5	11/16	-	lead

ISHahy 6	1	IS6	589825	559567	259	85	-	59985	58272	1714	8	14/19	-	lead
ISHahy 6	2-p	IS6	-	-	-	-	+	-	-	479	-	-	92530-92889 & 93159-93279	lag
ISHahy 6	3	IS6	175162	173805	1358	453	-	175322	173609	1714	8	14/19	-	lead
ISHahy 6	4	IS6	279460	280215	756	251	+	279299	281012	1713	8	14/19	-	lag
ISHahy 6	5	IS6	1169224	1168184	1041	346	-	1169706	1167993	1714	8	14/19	-	lead
ISHahy 6	6	IS6	1694610	1695967	1358	453	+	1694450	1696162	1713	8	14/19	-	lead
ISHahy 6	7	IS6	1983253	1984610	1358	453	+	1983093	1984806	1714	8	14/19	-	lead
ISHahy 7	1	IS256	95285	96478	1194	397	+	95169	96494	1326	8	18/30	-	lag
ISHahy 7	2-p	IS256	-	-	-	-	+	-	-	117	-	-	102108(1)-102225(116)	lag
ISHahy 7	3	IS256	114190	115371	1182	393	+	114074	115400	1327	8	19/28	-	lag
ISHahy 7	4	IS256	847231	848412	1182	393	+	847115	848441	1327	8	19/28	-	lag
ISHahy 7	5	IS256	970297	971478	1182	393	+	970181	971507	1327	8	19/28	-	lag
ISHahy 7	6	IS256	1161504	1162685	1182	393	+	1161388	1162714	1327	8	19/28	-	lag
ISHahy 7	7	IS256	1185478	1184297	1182	393	-	1185594	1184268	1327	8	19/28	-	lead
ISHahy 7	8-p	IS256	-	-	-	-	+	-	-	619	-	-	1664295-1664913	lead
ISHahy 7	9-p	IS256	-	-	-	-	+	-	-	290	-	-	1664928(1039)-1665217(1328)	lead
ISHahy 7	10	IS256	2004998	2006179	1182	393	+	2004882	2006208	1327	8	19/28	-	lead
ISHahy 7	11-p1	IS256	-	-	-	-	-	-	-	880	-	-	2339230-2338351	lead
ISHahy 7	11-p2	IS256	-	-	-	-	-	-	-		-	-		lag
ISHahy 7	11-p3	IS256	-	-	-	-	-	-	-		-	-		lag
ISHahy 7	12	IS256	2463277	2464458	1182	393	+	2463161	2464487	1327	8	19/28	-	lead
ISHahy 8	1	IS30	297922	298992	1071	356	+	297814	299059	1246	0	19/26	-	lag
ISHahy 8	2	IS30	482017	482512	486	161	+	481336	482579	1244	0	19/26	-	lag
ISHahy 8	3	IS30	1657699	1656629	1071	356	-	1657807	1656562	1246	0	19/26	-	lag
ISHahy 8	4	IS30	2001588	2000518	1071	356	-	2001696	2000451	1246	0	19/26	-	lag
ISHahy 8	5	IS30	2170908	2169838	1071	356	-	2171016	2169771	1246	2	19/26	-	lag
ISHahy 8	6	IS30	2397298	2396228	1071	356	-	2397406	2396161	1246	0	19/26	-	lag
ISHahy 8	7	IS30	2546348	2545278	1071	356	-	2546456	2545211	1246	0	19/26	-	lag
ISHahy 9	1	IS30	684267	685316	1050	349	+	684146	685362	1217	13	19/27	-	lag
ISHahy 9	2	IS30	1316375	1315326	1050	349	-	1316496	1315280	1217	9	19/27	-	lag
ISHahy 9	3	IS30	1999384	1998335	1050	349	-	1999506	1998289	1218	13	19/27	-	lag
ISHahy 10	1	IS30	1420021	1421064	1044	347	+	1419920	1421135	1220	7	21/26	-	lead
ISHahy 11	1	IS30	66905	65862	1044	347	-	67006	65791	1220	0	20/26	-	lead
ISHahy12	1-p	IS21	275366	274314	1053	350	-	-	-	-	-	-	-	lead
ISHahy12	1H-p	IS21	274188	273433	756	251	-	-	-	-	-	-	-	lead
ISHahy12	2	IS21	707201	708538	1338	445	+	706851	709367	2517	6	17/26	-	lag
ISHahy12	2H	IS21	708538	709293	756	251	+	706851	709367	2517	6	17/16	-	lag
ISHahy12	3	IS21	1162962	1164299	1388	445	+	1161277	1161379	2517	6	17/26	-	lag
ISHahy12	3H	IS21	1164299	1165054	756	251	+	1162715	1165128	2517	6	17/26	-	lag
ISHahy12	4	IS21	1816707	1817273	567	188	+	1816357	1818858	2502	6	17/26	-	lead
ISHahy12	4H	IS21	1818045	1818491	447	148	+	1816357	1818858	2502	6	17/26	-	lead
ISHahy 13	1	ISNCY ssgr IS1202	130153	128808	1293	429	-	130206	128708	1499	17	19/24	-	lead
ISHahy 13	2	ISNCY ssgr IS1202	184124	182808	1317	438	-	184201	182708	1494	17	19/24	-	lead
ISHahy 13	3	ISNCY ssgr IS1202	496022	494679	1344	447	-	496075	494579	1497	17	19/24	-	lead
ISHahy 13	4	ISNCY ssgr IS1202	2543048	2541705	1344	447	-	2543101	2541605	1497	17	19/24	-	lag
ISHahy 14	1	ISNCY	99732	101168	1437	478	+	99596	101334	1739	12	11/16	-	lag
ISHahy 14	2	ISNCY	1597559	1598995	1437	478	+	1597423	1599161	1739	14	11/16	-	lead
ISHahy 15	1	IS1182	820471	822015	1545	514	+	820306	822225	1920	0	26/27	-	lag
ISHahy 15	2	IS1182	829402	830946	1545	514	+	892237	831155	1919	0	26/27	-	lag
Solo not identifiable	-	IS110	277258	276773	486	161	-	-	-	-	-	-	-	lead
Partial not identifiable	-	IS1182	117461	118687	1227	408	+	-	-	-	-	-	-	lag
Solo Partial	-	IS3 ssgr IS3	63952	64548	597	198	+	-	-	-	-	-	-	lag
False Positive	-	IS21	77	1498	1422	473	+	-	-	-	-	-	-	lag
False Positive	-	IS21	274	3	272	90	-	-	-	-	-	-	-	lead
False Positive	-	IS21	2577433	2578455	1023	340	+	-	-	-	-	-	-	lead
False Positive	-	IS91	79694	79029	666	221	-	-	-	-	-	-	-	lead
False Positive	-	IS91	1384811	1383876	936	311	-	-	-	-	-	-	-	lag

APPENDIX B.

REFERENCE ISHAHY NUCLEOTIDE SEQUENCES

ISHahy2

TGGATCGGGTATAATTAATTAAACAGTTTTTTCTCGGACATGATATAATAAGATAATAA
CAAACTAAAATGGAGGAAAATATCATGCCTACAAAATATCCTGAAGAAATCAAAAGAAA
AGTTGTTGCTCTGGCCAATAATGGTAAAAATCAAACTGAAATACTCAATGAATATGGAA
TGGCAAGGTCCACACTTCATAAATGGATAAAAGACTATAATAACTCAGGTTCATTCAGC
GCTAAAGATAATAGATCTGATAAAGAAAAAGAATTAATTAAATTACAAAAAGAAAACAA
GCAGTTAAAAATGGAGAATGATATTTTAAAGCAAGCGGCGCTGATAATGGGACGAAAGT
AGCAGTTATTAAGGCAAACAGGGATAAATACAGTATTAGCGCCATGTGCAGAGCACTCA
ATATATCAAGGGGTATGATCTATTATACCCCTAAAGAAAAACAGGTTGATGTTGAACTA
GAAAGCGAAGTGATTTCCATTTACAAAGCAAGTAGAAATCACTATGGAACCAGAAAAAT
CAAAAGAGAATTAGCTAAAAAAGGTTATCAGGTGTCCAAGCGAAGAATAGGTAAAATAA
TGAAAAAATATAATCTAGTTTCTACTTATACTAAAAAACAATACAAAGTTCATTCTCCA
AGCTGTAATGAAGATAAAATTGCCAATATTGTAAACAGAGAATTTAACAAAGAAGAAGC
TCTAGACGTTGTTGTCAGTGATTTAACCTATGTTAATGTAAAAGGAAAATGGAACTATG
TCTGTCTGATCATAGATCTCTTCAACCGTGAATTTGTTGGTTATGCAGCAGGTAAAAAG
AAAAATGCCGAATTAGTAACTGAGGCTTTTAAAAGTATTAAAAGACCACTAAATCAAAT
TAATATTTTACATACTGATAGAGGTAATGAATTCAAAAATAAAGCTATCGATGATATTT
TAGTCAGGTTTGATATTGAGCGTTCTTTAAGCAATAAAGGATGCCCATATGATAATGCA
GTGGCAGAAGCAGCCTTTAAAGTAGTTAAGACTGAATTTGCTTATGACAGAATATTTAA
CAGCTTTGAAGAGCTGGAATATGAGCTATTTGACTATGTTAACTGGTATAATAACCACA
GAATCCACGGGTCGTTAGATTACCTAACACCTGTTGAATATAGATATTTAATGTTCGAT
AAAAAATTGTTGTAAAAAAGTATTGACAATCCA

ISHahy3

TGTGACCCCCCCTAAATATATTGGACAGATACCAATTTAACATCTTATAATAATTACAG
GGGGTAAAAGCAAATGCCTAAAAGTTATGATCCTGAAGAAAAAATGGAAATTGTTCTTC
GAGCTATCAAAGGTGAAAAAATATCTGATCTTGCTGATGAATATAGTGTCAGCCGTAAT
TCCATCTACCTATGGAAAAAAGAATTCTTAAACGGAGGTATGAGTAAACTCAGTGGTGA
ATCTGCTACTGAACAGGAAGCTAAACTTAAGAAAAAAGACGAACAAATTAAAGAGATGG
AAAAAATTATTGGCCAACAGAAAGTCCAGATGGAAATCCTTAAAAAAAAGCCCTGGCAG
AACTAAATACTGTCGATAAAATCAAGATAATTAATCAGCTTAAATCAGAATATACTGTA
GCAGAACTATGCCGAACTTTTGATATAGCCAGGAGCACCTATTATTACCGTCAGAACAA
TGATCAAAAAGAGAAAGATACTAATACTGAAGATCAACTTTACACTGGTCATCCAGCTT
ATGATAAAGATGGTAATTTAGTTCCAGAAAAAGAGGTTGTTAAATTGGTTAAAAAGTAC
TGTGATGAGTCTCCTCATCTGGGCTATAGAATGGTAACTGATTACTTGAATTATACAGA
AAATCTAAAAGTTAATCATAAGCGTATTTACCGTATAATGAAAGTCTTAGATCTACTAC
AGGATAAGAGAGTTCCAAAACCTAAAAATTATCAATTAAAACAAAAACATGAACTCACT
GGACCTGACCAGCTCTGGGAGATGGATATGGTCCAGATGTACATAGATAACAGTGGCCA
GTGGGTATATATGTTTGATATTATTGATGTATTTACCAGAGAGATTGTAGGTCATCATG
AAGGATTAAGATGCCGCACAAAAGAAGCTCTTAAAGCCCTTGAAAAGGCAATAGAAAAT
CGGAATACTGATAATCTAATATTAAGAACTGATAATGGTACCCAGTTTAGAAGTCGAGA
GTTTCAGACCAGGATTAGAGAGCTTGATATATCTCATGAAAGAACAATGGTAAACACAC
CAGAAGAAAACGCCCATATTGAGAGCTTTCACGGCACATTAAAAAGAGCTGAAGTATAT
CAAAAACATTACCGCAGCATAACCCACTGCAGGAATTCTATTGCTAAATTTGTTGATAA
ATACAACAATAGGAGACCACATTCATCTGTTGGTAAAATACCACCAGCAGTTTATCATA
AAAATGTTCTTAACAACTTAGTTTCAGGAATTAGATTTGCTGCATAAGCAGTTAATTTA
TTGAAAAAACTATGTTTAACAAATCATATTAAAGTTTGTCCAAGATTAGAGGGGCTAGC
TCA

ISHahy4

CTATATAGCCCCCTAACAGAATTGGAGAAAAAGTTTGAAAGTGATATACTCTAATTACA
AGGGGGAAACACCAATGGCAAACAGAAAATATTCCGATGAAACTAAAGAACAAATTGTA
AAAGAATGTCGCGAAATAGGTAACACAGCTCTTGTAGCAAGACGACATAATATTTCTAA
GCATACTGTTTACAGCTGGGTCAAAAAAGCTAAAGAAACAGGATCAGTTAGATCTCTTC
CTAAAGATGAAAAAAAGCAAATGAAAGAGATAGAAAATAGATTAAGTAAAATGAGCGAT
GAAAATGATAAGCTCAAAAAAATTGTAGCAGAAAAAGAATTAGAATTAGCGATTTTAAG
GGAGTTGAGAGATAAAGTAAACCCCCGATAGCCCTCAAAGTTCAGATTGCATCAAAGTG
GATAAATAAAGGGTATAAAATCTCTATTGTTTTAGACTTTGTTGGGCTTAATTCTTCCA
CTTACTACAGTAATATAAATAGAAAAACTGAGAGTGAAAGTACTAATAGCAGCAATTCC
AATAATCCTCAAGGAAGACCTGTCCCTGGGTATTCTCTAACTGAATCAGGTGAAAAAAT
ATCTGATGAACAGATTAAAGAATGGCTCTTAGAACTGGTTGCAGGAGATGGCTTCCCTT
ATGGTTACAGGAAACTTACAGTCTGTTTAAAAGAAGACTATAACTTGAAAATAAATAAG
AAAAAAGTATACAGGTTATGCAAAGAACTGGATATATTAAGATCGCAAAGAAAAATCAA
AAAATTTAGACCTAAAAAGATTGCAAAACAGGAAGAAATTACAGAACCAAATCAACTCT
GGCAGATGGATTTAAAATACGGCTACATAAATGGAACAGATCAGTTCTTTTTCCAGATG
TCAGTAATTGATGTCTTTGATAAGACTGTTATAGATTATCACCTGGGACTAAGCTGTAA
AGCTAAAGATACCTGCAGGGTATTAAAGGCTGCTTTAAATAAAAGAAAGCTGTATAAAG
GCATGAATTTGCCTAAAATTAGAACAGATAATGGACCACAATTTGTCTCTAAATTATTT
GGAGACACCTGTGAAAAACTGGGGGTAGAGCATCAGAGAATTCCAGTTAGAACACCTAA
TATGAATGCTCATATAGAATCATTTCATTCGGTTTTAGAAAAAGATTGTTATTCAATTA
ATGAATTCAGTAGTTTTATTGACGCCTATAAAAAAGTCAGTGAGTATATGAATTATTAT
AACAACAGATACCGTCATGGCAGTCTTAATGATATGCCTCCAGCAAAATTTTATAAACT
GGCTAAAGCAGAAAAAATAGTTGCTGAACCAGTACTCGCCTAAATCAAAAAATGAGAAA
CTAAGAATGAGCGAGCTTAACAAACATATTTTCCATATTTAGGGGGTTGAACCG

ISHahy5

TGAGTTAAGATATATAATAAACTCAGGAGGTGCCCAAATGGAAAACGAAGAAGTTACA
CTGAAGAATTCAAAAGAGATGCTGTTGAACTCAGTCTCAACTCTGATAAATCTGTTAAA
GAAATTGCTGAAGATCTCGGCATTAATTATGGTAATCTTAATCGCTGGCGTAGAGAATA
TAGAAACAAAGGTAAACATGCTTTTCCTGGTAATGGAAAACAGAATTTAACACCTGAAC
AAAAGAAAATAAAAGAAATTGAAGATGAACTTAGAGAAACCAAATTAGAACGTGATATA
TTAAAAAAAGCAGTAGGCATCTTTTCGAAAAAACCGAAGTAATCTATGGTTTTATCCGG
GACCACAGAGATCAATTCCCTGTGACGAAAATGTGCCAGGTATTAGCAGTTTCCCGGTC
AGGTTTCTATGATTGGTTAGATAGAGAACCCAGTCAAAGAGAAATCGAAAATAAGAAAC
TAAAACTAGAAATAGCTAAAATATACTGGCAGCATAGCGGCCGCTATGGAAGCCCTAGA
ATACATCGTCAGCTGATAAAAGAAGGTCATAACTGTAATATAAAAAGAGTTGTGAGACT
TATGAAAGTAATGGGTCTTAGGGCAATTCAGAAAAAGAAATTCAAAAAGACAACTGATT
CAAATCACAACCTACCTTTAAAAGAAAATTTGCTTAATAGAAATTTTGATGTATCTAAG
CCTAACAAGGTTTGGGCTTCAGATATTACATATATACCTACAGCTGAAGGCTGGCTTTA
CTTAGCTGTAGTAATAGATCTTTATTCCCGCAAAATTGTTGGTTGGTCAATTAATAAAA
GAATGACCAGACAGCTTGTCATAGACGCTCTTAACATGAGGATAAAAAACAGAAATCCT
AAAAAAGGTTTAATATTTCATTCTGATAGAGGCAGCCAGTATGCAAGCCATGATTTTCA
GAAAGAACTATGGAAAAATGGTATAAGATCCTCAATGAGTCGCAAAGGAGACTGCTGGG
ATAATGCAGTAGTAGAAAGTTTCTTCTCCACATTAAAAACAGAATTAATTTATCAAGAT
AATTATAAAACTAGACAGCAAGCAAGACAGGAAATTTTCCAATATATTGCTGTTTATTA
CAACAGAATCAGAATGCATTCTACATTAGATTATAAAAGTCCAGAAGACTACGAAAACG
AGAGAAAACCATCTAAACTATGTGTCTAATTTAATGGGGAAACCTCA

ISHahy6

TGTTACGTCAAATCTAATAGTTTTTTCAAAAATATTAATAGTTATCATAGTTGTCAGGC
TAGGAATCTCCATAAAAAGAAGGAATCCCTCCCCCTGAGCTAGTAAATAGTAATTGAGA
AAGTCACAACCTACTAGACCCAGAAAGGAAGGGATGTAAGTTATGTTTTCTCGAAAAAT
ACCTAATTTCCTGCTTAAACCTATCATGAATTTTGTTCTTTTGACCTATCTTTTTTTCG
CCAGACTTTTCCAGGTTGATTTTGAACCCAATAATAAACTCGATGATAGTTATACTAAA
TTTAACAGTTTTGATGACCCACAACCTATTATAGAATATAATCAGGTTGACTATCGCGA
TATTCTAGCTGAAGCTGAAAAAAATGGTGAAACTATTCAGCCTGTTAATCGCAATAAAC
CTCTAACTGTAAAGGTTGAGGAGTGCTCTCAATGTGGAGCTCCCAAAAAATATCTCTAC
AGTTTCGGTCATGATCCAGATGGATACCAAAAGTTCCAGTGTAAACTCTGTAAACATCA
GTGGGCTCCTAAGAAACCTAAAAAACCTAAGAACCACCCTACATATCGCTGTCCTTTCT
GCAACTATGCTCTTGCTAAAGAGAAAAAGCGCAAGCATTTTACCAAGTATAAATGCCGC
AATGATGACTGTTCTAAATGGAAAAATGAACATAAGCGTTATCGCTACAGAGCCTATAA
TTTTGATATCAATAAGCTTGAACTTTCCAGACCTGATAAAGAACCTGTTAACCTTGATC
ATTCTCACTATGGAAACTTTACCATCTCTAAGGCTATTGACTTTTATGTCAGTCTTGGT
CTTTCTTTAAGACAGGCTGAAAGAGCTCTTAAACTTGCCTATGGTGTTTCACCTTCAGC
TCAAACCATCCAAAATTGGACTGTCTCTTTAGCTTACAGACTTGCTCCTAAAATCAATG
AGCTTGATCTGCCTTTATCTGGTATTGTTGCTGTTGATGAAACATATATTAAAATAAAA
GGCAGCTGGCATTATCTTTTTACTGCCATAGATGGTGAAAATGGCTGTGTTATCGCTCG
CATCTTTCTAAAAATCGCGACGCTAAAGGTGCAATAACCATTTTAAAAAGAATAATCGA
CCAATATCAAGACCAAAAATTTGTTCTAGTTACTGATATGGCTCCAATTTACAGAGCTG
CTGTCCATGCTGCTAAAGTGTTTTTGAAAACCAATATTGACCACAAACAACTTAAGGGG
CTATTTGCTAACGATGATAACTCAGATGAAATTTACAGACCTTATAAAAACATCATTGA
AAGGTTTTTCGGTACTTATAAAGCCCATTACAAACGCCATAAAAGCTTTAGCTCTTTTG
ATGGTGCACTTACTCATATAACTCTTTATCAGCTCTATTTTAATTATTTAAAACCTCAT
AGTTCCTTTAATGATAAACCACCACTGATAGTGGAAGGAGCAAGAGGTCAGCCTATCGA
ATCCTGGGCCCAACTTATCAGATGGATCACCAAAACTGATAGGTAATTAAATATCTTTT
AGAAAAATCATGTTTTATCTCTTTAGATATTTATAATCACTAAATACCATTTTCTTACT
TTATTTACAACTTTAAACTTTTGACATTCTGATCATTTTCATTATAATCGTAATTAGTG
GTGCTCTATTACTATTTGCTTACATTTTATGGTAACTATCATATTTTATTTGACGTTAT
CA

ISHahy7

AGTAGTGTACAATAAGTTGTGTAAATAGATTTTTTTCAATAAAAAAAGAGGAATCCTTC
TTTGAATGGTTGAAATATTTATTAGCGAAATAAACTCAACCCAAAGGAGGACTCCTCATG
AACAGTATACCCGAAAAAAACAGTGATGGTCAAGTTGAATTTCACGATTTAATCATTG
AACTCATCAAAAATTTTCTCGAGAATTTCCTCAAGGCTGAATTAACTGAATTTCTAAAC
TATGAAAAACATGAATACTCAGGTAGAAACTCTGGCAATAGTCGTAATGGATCTTATCT
TCGTGATTTCTTAACTCAGTTTGGCAGTATTAAAGGCTTAAATGTTCCTAGAGATAGAA
TGGTGAATTCCAAACTGAACTATTCCAACCATATAAACGCTATGATAACTGGCTTGAAG
AAGCTATAATAAACATGTATGCTAATGGCCTTTCCACCCGCTATGTAGCTGATTGGATA
GAGCAGATGTATGGACAAAAATATAGCCCTACTACTATTAGTAATCTAACTAATGTTGC
TCTTGAAGAGGTTAAAAAGTGGAAAGAAAGACCACTTCAAAAACGGTACAGCGTTATTT
TTATTGATGGCATGAGCATAAAAGTCAGACGAGATACTGTTGCAAATGAATCTGTATAT
ATTATCATTGGTATCAATGAAGACGGCTATCGTGAAATACTTGATTTCTACATTGGTGC
AACTGAATCTGCTGCTTTATGGGAAGAAGTACTAAGTAATTTAAAAGAACGCGGAGTCC
AGGAAGTCCTACTAGGTGTTATAGATGGACTCCCAGGACTTAAAGATTCTTTTCTAAAG
TATTCCCTAAAGCGGATGTGCAGCGTTGTATAGTTCATAAAGTGCGTAATACAATAGTC
AAAGTTAGAAAAAAAGATACTGATGAAATCGTTAAAGATTTAAAAAAGATCTATAGATC
TCCCAGCAGAGAGTTTGCAGAAAAAGCTTTAGAAGAATTTGATTTTAAATGGAGTAAAA
TCTATCCTAAAGTTACTCAAAGCTGGTACGTAGATAAAGATGAACTATTAACATTTTAT
AAATATCCAGAAAGCATACATAAAGCCATATACACAACAAACTGGATTGAAAGAGCCAA
TAAAGAAATCAAAAAAAGATTAAAGCCTATGAATAGTTTGCCTAATGTACAAGCAGCTG
AAAAAATAATTTATTTAAAGATTATTGAGTACAACTCAAAATGGTCTGATAGAAAGATG
AGAGGATTTTTAGCTGCAAAAGATCAGCTCCATCAACTATTTAAAGAACGATACTGATT
TATTTACACAAGATTCTTGACGTTATC

ISHahy8

TTTGAATTGTATAATTAAATGCACGCGGTGATTAGATAAAAAAATAGACACATACCTGT
ACCTTTTGTATAATGTTAATAGCGACAAAAACATTAGGAGGTACAGAATATGTGTCAAA
GTAATTGTAACACAAAACGAAGCAAAGGAAAACACCTGAATTTTGATGATCGCAAATTA
ATTAAACATTTATATAATGTCCAAGAAAAAAATTATACTGAAATAGGCGAAGAATTAAA
CTGCCATAGAACAACAATCAGTCGAGAAATAAAAAAAGGTGAGGTAGAACTTGATAATG
GTGATGGAACAACCAGAAAAGAATATGTACCAGAGATAGCACAGAAAGTATATGAATTT
AATAATTCGAATAAAGGCCCTGATTTAAAAATCGATAAAAATAAAGAGTTAGCAGAATT
TATAGAAGAAAAAATCAAGGAATTAAGATCTCCTGCAGCAGTGGCAAAAGACATTGAAG
AATCAGATAAATTTGAAATTCAATTACACTGGAAAACAATCTATAATTACATAGATAAA
GGTGTTTTGAATATAGACAGAAGCGATCTCCCACATGGTAATTATAAAACTGGGAAAGA
CAGACCAAAAGAAAGTGAAAGCACTACGAGGCGTAAGGAAGGTCGTACAATTAGAGATA
GGCCTGAAGGAGCTGATACCAGGGAGGAATTTGGGCATTGGGAGATGGACTTAGTAGAA
GGGTTAAAACAAAAAGATGAACCTGCCTTACTAGTGCTGACAGAAAGACAAACTAGGCA
GGAAATCATAGAAAAAATACCAAACAAAAAAGCAAAGTCAGTGGTAAAAGGACTTGATC
GTATAGAAAGACGATTTGGGGTTGTTAATTTTAGGGAAACATTTAAATCGATTACTACA
GACAACGGTTCAGAATTTGCAGATTATGAAGGTATAGAACAATCATATACAGGTAGCAG
TATACCTAGAACTAGTTTGTATTATTGTGATGCATACTGTAGCTGGCAAAGGGGATCAA
ATGAAGTAGCTAATAAGTTTATCAGAAGATTTTTACCTAAAGGCACAAGTTTTAAAGGT
ATTAAGAGAAAACTGATAAAAAAGATTCAGGATTTTATAAATACTTATCCTAGAAAAAT
GTTTAACTATGAAAACTCAGATAAATTATTTAAAGAGAAATTAAGCTTTAGTGTCTAGA
AGTAATGAAGGTACAGGTGATTTTAAGAAAGATTAAAAAAGCGTGCATTGCATGTTGCA
ATTCACC

ISHahy9

GCTTAATTGTAAAATGAACTTGAACTAATTAAAAAAAACTAATAAAAAATGCCCATAGT
GGCACAAACTGTAATATAATTGAATTACTACAAACAATAAACATTACAGGAGGGTGCAC
ACTATGGACTACTTATATGATACACCAAATTCTCGAAAAAATAAACACCTTAATGCTTA
TGATCGCGGTCAAATTGCTTTATTACATTCAGAAGGAATGTCACCTTATGCAATTGGTA
AACGCTTAGGTAGAGCTTCAAATACAATTAGAAACGAGTTGAAACGTGGTACAGTTTCT
CAAATAAAGGCCAATAAAAAGGTTGATATTTATTTCCCTGATGTTGGTCAAAGAGTTTA
TGAAAATAATCGTAAAAATTGCGGACCTAAGTTTAAACTCCTAGAGTGCGAAGATTTCA
TAGAACATGTTTTAGATAAATTTTACAACTCAGATCATTCAATTGATTCTATTTGTGGA
TCAGCTCAGGTGCATAATAAATTTCCAAATTCAAAGATGGTTTGCACCAAAACTCTCTA
TAATTACATAGATGCTGGACTACTTAAGATTAAAAATATTGATCTACCATTGAAATTAA
AGCGTTCTACAAAACCAAAACGTATTAAGCAGAATAAAAAGAAACTGGGTACCAGCATT
GATGAACGCCCTGAAAGCGTTAATGATCGCAGCGAATTTGGCCACTGGGAAATTGATAC
TATTATTGGTAAAAAGACTAAAGATGAAGCAGCACTACTTACTATGACAGAACGTACAA
CTCGCTCACAAATTATTCGCAAAATTGATGACAAAACATCTTGTTCTGTTCAGGAAGCT
ATGACAAAGCTAATCAAAGAAACTGGAGATCTTTTTTCTACAGTCTTTAAAAGCATTAC
CAGTGACAATGGTTCTGAATTCTCAGAGCTAGCCAGTGTAGAAGAAATAGTTGGCACTA
AAATTTATTATACTCATCCCTATTCAGCCTGGGAAAGAGGAACAAATGAACGTCACAAT
GGCTTGATAAGAAGGTTTATACCTAAAGGTAGAAGTATAAATGAATTTTCAATTGAAGC
TATTGCTAGAGTTCAAAATTGGTGTAATACTTTACCTAGAAAAATATTAGGATATTTAA
CTCCTGATGAAGCTTTTGAAGACCAACTAAAACTAATTCTATACAATTAAAATTATTAC
AGTTTGAAACTAGTTCAATTTAACATTGCAATTCAAG

ISHahy10

GTTTCATTGTAAAATTAAATGCAACAGACAAAAAAATTAGAACCATAGCTAAAAACCTA
GTATGATATAAGTGACCAAACAAAATCATATGGAGGTTTAAGCTATGGCTCATACTAAA
GATAATACCACAACTCAAAGAACTTTTAAACATCTTTCTTCTTATGAAAGAGGTAAAAT
TGCTGCTCTTCTGCAAGAAGGATACTCTCAAAGAAAGATTGCTGAAAAGCTAGGCAGAA
ATCACAGCACTATTAATCGTGAAATTAAAAGAGGTACTACAACTCAACTTAACTATGAT
CTGTCTACTTATGAACAGTATTTCCCTGAAACTGGTCAAGCTGTCTTTGAGAAAAATCG
TTCTCACTGCGGTAATAAATCTAAATTGCTTAAAGTGGAATCTTTTCTACAACATGCTG
AAAAAATGATCTTGAAAAATGATTGGTCTCCAGATGTTGTTGTTGGACACGCTCTAAAA
AATAATGAATTTACTAAAGATGAAATGGTATCCACTAAAACTCTCTATAATTATATAGA
TCAAAATTTATTAGATGTTAAAAATATTGATCTTCAACTTAAAGTTCGCAGAAAACAAA
GAGATCCTAATAAAAGAAAGCATAAAAGACTTCAGGGTAAAAGTATTGAAGAAAGACCA
GAAACAGTTGATGATCGAAAAGAATTCGGCCACTGGGAAATTGATACCGTCAGAGGTAC
TAGAACTAAAGATAATGTCCTTTTAACTATTACTGAAAGATCCACTCGTCAGCATTTAA
TTAGACTCTTAGAAGATAAGAGCTCTGCTGCTGTAGATCAAGCAATTAAGAAACTTAAA
GTTCAATTTTCTAATGTTTTTAAAAAGGTATTTAAAACCATTACAGCTGATAATGGAAC
TGAATTTGCTAATTTATATAATCATGATGTTGATGTTTACTATGCTCATCCTTATTCAG
CCTGGGAAAGAGGCACTAATGAAAGACATAATGGTCTCATTAGAAGATTCATTCCTAAA
GGAGAACAAATAAGCAACTATACAGAAAAACAGATTCGGAGAATACAAAATTGGTGTAA
TAATTATCCAAGAAAATTATTGGATTATTCTACTCCAAATGAATTATTCCAAAAAGAAC
TTCAAGCTATTATTAACCCTGTTTAAACAATTTAAAATTCAAAATACTATATTATACTA
GTCTTTTCTATAGGTGTTGCATTTAATATTGCAATTTAAG

ISHahy11

GGTTCATTGTAAAATTAAATGCAACAGACAAAAAAATTAGAACCATAGCTAAAAACCTA
GTATGATATAAGTGACCAAACAAAATCATATGGAGGTTTAAGCTATGGCTCATACTAAA
GATAATACCACAACTCAAAGAACTTTTAAACATCTTTCTTCTTACGAAAGAGGCAAAAT
TGCTGCTTTTTTGCAAGAAGGATATTCTCAAAGAAAGATTGCTGAAAAATTAGGAAGAT
CTCCCAGTACTATTAATCGTGAAATTAAAAAAGGCACCACAACTCAACTAAACTATGAT
CTGTCTACCTATGAGCAGTATTTCCCTGAAACTGGTCAGGCTGTTTATGAGAAAAATCG
TTCTCACTGCGGTAATAAACATAAATTACTTAAAGTAGAAACATTTCTAAATTATGCAG
AAAAGATGATTTTAGAAAATAGTTGGTCTCCAGATGTTGTTGTTGGCCATGCTCTGAAA
AATAATAAATTTACTAAAGATAAAATGGTATCTGCTAAAACTTTGTATAATTATATAGA
CCAAAATCTACTAGATGTTAAAAATATTGATCTTCAGCTTAAAGTTCGTAGAAAAAAAA
GAGTTCCTAATAAAAGAAAGCATAAAAGACTTAAGGGTAAAAGTATTGAAGAAAGACCA
GAAACAGTTGATGACCGAAAAGAATTTGGCCACTGGGAAATTGATACTGTTAGAGGCAC
TAGAGCCAAAGACAATGTTCTTTTGACAATTACCGAAAGAACTACTCGTCAACATTTGA
TTAGAGTTTTAGAAGATAAGAGCTCTGCTGCTGTAGATCAAGCAATTAAAAAGCTAAAA
GTTCAATATTCTAATGTTTTTAATAAGCTATTTAAAACAATTACGGCTGATAATGGAAC
TGAATTTACTAATTTACATAATCACGATATTAATGTTTACTATGCTCATCCCTATTCAG
CTTGGGAAAGAGGCACTAATGAAAGACACAATGGTCTCATTAGACGATTCATTCCTAAA
GGAGAACAAATAAGTAAATATACAGAAAAACAAATTCAAAGAATACAAAATTGGTGTAA
TAATTATCCAAGAAAATTATTGAATTATTTTACTCCAAATGAATTATTCCAAAAAGAAC
TTCAATCTATTATCAGCGATGTATAAATAGTTTTAAATATAATATACTTCGTTATACTT
GGTTTTTCTATTAGTGTTGCATTTAATATTGCAATTTAAG

ISHahy12

TGTATATGATAATCTAAAAGTGACCCAGGATAATAATTTAAAATTGACCCGTTTAAATA
GATAATATAAACTGTCCATAATAAAATAAAAATTATGGGCAGGTGAAAAAGGTGATAAA
ATTGAATGAAAAAGCCGATATTCTAATTAAACATTTTGTCGAGGGTCAATCAATTAGAA
AAATAGCTAGAGAGTCAGAATTTTCAAGAAATACTATCCGCAAATATATCCGAGATCAT
GAAGAAAAACTTCAGAATCTGGATAAAGCAAAATCCAGAGATGAAATATTAACCTTAAT
TGAATCTCTAGTAGAAACTCCTGAGTATGACAGCTCTAACCGTAATAAATATAAAATGA
ATGAAGCTATCAAAAAAGATATTGATTACTGCATTAAAGAAAACAAAAAACGGCGAGCA
ACAGGTATGCGTAAGCTGCAGCGTAAAAAGATTGATATTCATGAATATCTGCTCGGTAA
AGGTCATGATATTAGTTACTCAACAGTCTGCAACTACATCAGAAACACCTATGAAAATG
CAGCCAAAGAAGCATATATAAAACAGCTCTATCATGAAGGAGAATCCCTTCAGTTCGAC
TATGGTGAAGCCAATTTAGAAATTGCAGGTAAAAACAAAACCCTTAATATGGCTCTATT
CACAACTGGTTTTGGTTTTCATCACTATGGTAAATTTTATGCTAATAAAAAGATGACTT
CATTTTTAGATGCTCATGTTAAAGCATTTAAGAACTTCAATGGAGTATATCATGAAGTA
GTTTATGATAATTTAAAACAGGCTGTAAAAAGATTCGTTGGTCCTTCTGAAAAAGAAGC
TACTGAGGATCTAGTTAAGCTTTCCCTCTACTATGGCTTTAAATACCGCTTTTGTAACG
CATATCGCGGGAACGAAAAAGGCAAAGTTGAGAAAGGCGTTGATTTCGTCAGGCGGCGG
ATCTTCTCCCAGAAAACAAGTTTTGAATCCATAGAAGAAGCAAATGAGTATCTTGCTGC
AGGACTAAAAAGGCTTAACTCAATGAAAAAAGCTGAATTAGATAACAGATCACCATATG
AGCTCTGGCAGCAGGAAATACCATATCTACTGCCTTTAAAACCTCCATATGAGGCCTGC
AGAGAAAAAGAATGTAGAGTCAATAAATACAGCTTTGTAACCTTTGAGCAGAATAAATA
TTCTGTACCTGATCATCTTGTTGGAAGATTTGTTAACAGCAGAATTTACTCAGATTATA
TAAAAATATACTACAACGATGAGCTGGTAGCAAAGCATCAGAGAATATATGAAGTTCAT
AAATGCAGTATTGAAATAGAACATTATCTTCTAACTCTAAAAAGAAAGCCTGGTGCATT
GAAAAAATCCCTAGCTTTTCATCAAATGGCACCAGACCTCAAAGAGATATATAAATTAT
ACTACGATAATTATACCAGAAAAACTAAAGAATTCATAGAGCTTTTAGAACTGGTAAGT
GAGAAAGGTCTGGATGAAATTAAGGGTGCTATAGAAACTCTTTCTAAAAACAAAAAGTC
AATGGTAACAACTGCTAATATCAAAGTGTTAGTTCAAAAGGAGAATACTCCTGATTTTA
CAATTTCCAGTGAAATAAATGAGAATTCTCTGGAACTATTAAAAAAATGGGATACTATG
TTTAGTCTCCAAAATCATCAGGAGGTGATCCAGTAATGAATAAAGAAATAATACTGGAA
ACTAAAAATTACATCAAAGAACTAAAGATTGCTGGTATTAGAAACGACTTAAATCAAAA
ACTGGCTGAAGCTTATCGCAACAATACACCTTACGAAGAACTATTAAGAGATCTATTTC
GAGATGCATATGATATGCGTAAAGAAAATGGCCGTAAAAATAGGATTAAAAATGCTAAA
TTCCCATACAAAAAATATTTGGATACCCTAATTGTAGACTATCTACCTGAGAGTGCTCA
AGCTAAGTTTAAAGAGTTAAAGTCGCTTAAATTTATTGAAGAAAATAGAAACATAATTT
TCTCAGGAAATCCTGGGACTGGAAAAAGTCATTTGTCTATAGGCCTTGGTATCAAGGCC
TGTAACGAGGGCTATAAAGTCTTCTTTGCCACAGTACCACAGTTTATCAACCAATTGAA
AGAAACCCAGAGCGAAAGAAGATTAACTAATTTTGAATCTAAATTCCAAAAATATGACT
TAATAATTTTAGATGAATTGGGCTACATCTCTTTTGACCAGAAAGGAAGCGAACTTTTA
TTCTCCTTTCTTTCACTAAGAGCAGAACGCAAATCTACTATCATAACTTCCAATCTCTC
ATTTGAGAGGTGGAATGAGGTGTTTAATGATCCAGCATTAACTGCAGCTATGGTAGATA
GATTGACCCACAAATCATATGTAATTAATATGAATGGTAATTCTTACCGAATGAAAGAA
ACTGAAGAGTGGCTGGGTCAATAAAATTAATATTATTTTTTGCTTACCCGGGTCAATTT
TCAATTAATAAAAGGGTCAAATTTCACTTGACAAAAACA

ISHahy13

TGTGATTATCTTGTGATAATGTCATAGTGAAATTATCTTGATAAAATGTCACTATGAGA
GGAGAAAAATATTATCTTATGTCCCAAACACAATTGAATAGGTATCATGTCATTTCTAT
GGTCATTGATAAAAGTATGACCAACTCAGAAGCTGCACAGGCTTTAGACCTTAGTGTTC
GCCAAATAATTCGTTTGAAGAAAGGTGTGAAAAAAGAAGGTCCATCATTTTTAGTTCAT
AAAAATAAAGGTCGTAAACCTAGTCATGCTTTTTCTGAGGATTTCGTCAGTAAAATTGT
TGCCCTGAAAAAATCAGACTCGTATCAAAGTGCTAACTTCATGCATTTTCAAGAATTAC
TAGAAGAACATGAAAACATCACCATTAGTTACAACGCACTTCACACTATTTTAAGCAGC
AATGGTGTCAAAAGTCCCAAAAAGCATCGCAAAAAGAAAGTTCATCACCGTCGTAAACG
CAAAGCAAAAGAAGGCCAATTAATCCAAATTGATGCCAGTCCACACGACTGGTTTGGTA
CTGGAGAGAAATATGACATTCATGGTGCCATAGATGATTCTACCGGTAAAATAATGGGC
CTTTACATGACTAAAAACGAGTGCCGACAGGGATATTTTGAAACTTTAAGATCTGTGGT
CTTAAATTATGGTATACCCGTCAGTATTTATACCGATAGACATACTATCTTCCGTTCAC
CTAAGGCAGACAAACTTACAATTGAAGAACAACTTGCTGGTAAAACTGTTAAAGATACT
CAGTTTGGAAGAGCCATGAAAGAACTTGGAATTACTATGATCCCTGCACGCTCACCGCA
AGCTAAAGGCCGTATTGAAAGATTATGGGATACTCTCCAAAGCAGACTGCCTGTTGAAT
TTAAGATAGCAGGTATATCCACTATCGAGCAGGCTAATGAGTTTTTTGCTGACTACTTA
ATCAAATTTAATGAAAAATTTGCTGTAGCACCTGAAGACACCATTTGCGCTTTTAGAAC
TCTTCCTGAAGAAATTTGTATAGATAATATACTTTGCGTTAAGGAAGAAAGAATACTTG
ATAATGGACTAACTTTTTCTTTTTATAATCAGCGCTTCAAGGTCGTTACTAATAGTATG
TCAGTCTATCCAAAATCGAAAATCAAAGTTCTTCTTAGTCCTAAATTCGGTGTGAAAGC
ACAATATGGTGAAAAAATTTATGATGTTATCCATTGGACTGAACATAAAAAATCAAAAG
TTAAAAAGAGTAAAACTAATTCTAATAACAAATACAGACCAGACGATGGGCATTATTAT
AAATATGGACATCAAGCCTGGCCAAAAGTAACTTTTGAAGACAGTGATCACGATATACT
GAAAATGCTGGAAAAAATATTTCTCACTCAAATTGCCTGATATGTTTGGTTTAAACCCC
AAACATATCTCCCCTTTGGGGGTGACATTTTCTCAAGATAAAATTGGTATAAAAAAGTG
ACATTTTCAAAAGTTATTGACA

ISHahy14

TAGTCGTTTGTATAATGTCTAAGGCCCGAAATAACAGGGCTCATGACATAAATTTTTGG
TGACTTCCTCCTTAAAATATGGTATAATATAGTTAGAAATACTAAAAATACCACAAATC
AAAAGGAGGAAATCATAATGTCTTTTAATTCTAAAAAGCAATTGTCTTTCGGTGATCTT
TATGAACAGGCCAAAGATTGGGCTCAAAATGATAAACCTCAATTCCTTGAAATGCTCGA
CCAATATCTTGATTTATCTGAATTTATTCCTGACGGTTTCTACAATGCTTACTACAAAT
ATTTTGGTAGAAATAGAATTTACAGACTGGAATCAATGCTTTCTGCATTTATACTGCAA
AAAATACTGGGTATTCCAACTCTTGTCCTTCTAATTAATATCTTAACTTTAAGCAGTGA
TTTAAAAGAATTCTGTGGTTTTAACTCTGTACCTGATATCTCTCAGTTTTCTCGTTTTA
AGACAAAATTTGAAGATCATTTAGAAGACTTTTTCTATCACCTTGTTGATGTTACTGAA
CCTCTCTGCAGAAAAATTGATCCTTTAATGGCTGATCTTTTTATCTATGATACAACTGG
TTTTGAACCCTATGTGATTGAAAACAATCCTAAATACATAAATAATATTATTCGCAGAC
TTAAAAATATCTACAAAAATGATAAAAATGTTAATATTTATGGTTTTGCGTATCAGTCT
ATGCCTTCTTCTGCTCAAGTTAATAATGAGATAAAACAACTCTATATTAACGGCCATTT
TTGTTATGTCTACAAAGCTGGTATTGTCACTAATGGTCTTGGAATTATCCGTCACATTT
CTTTTTTTGATGATCAATTTAAAGATAATCACCCTGAAATACCTATTGAGAAAAAGACT
GATTCACCTGATGAAGATAAATCTATTGGTGATTCAACTTCTTTAAAACCAGTTTTAGA
AGATTATTTTAAACTTCACAGTAATCATCAATATTCAACTTTTATCGCTGATTCAGCCT
TTGACAGTTATCAAACTTATCCTTTTTTTACTGAAAGATTTTGGTTTTGATAAAGCAGTC
ATACCACTTAATTTTAGAAACACTAAATCAAGTTTACCACAGCCAGAATACAATGAAAA
CGGCTGGCCTTTATGTCCTAAAGACTCTTCCTTACCAATGAAACCTAATGGTTGGTGTC
GCGGTAAAAACCGCAGTCCCAGATTTAAGTTTGTCTGTCCTAAAATTAATCATAAGGGT
GGTAAAAGAAACTGCTACTGTGAGAACCCCTGTACTGACTCCAGTTACAGTAGAGTTAT
TTATACCTATCCTGACCAGGATTTAAGAACTTACCCTGGTATCATCAGAGATACAGATG
ATTGGATAAACCTCTACAGAAAACGAGGAGTAGTTGAACAGACAATTAACTATTTTAAA
GATGCCATGGTTACTGGTAATTTAAAGACTCAGAACCTAAAAAGTGTTAAATCAGATGT
TTTTCTAGCCGGAATCACTCAACTTTTAACATTAATTCTAGCTGATAAAATGGATAAAC
CAGAAAATATTAGATCTTTAAGATCACTAATTGCTTAGAATTCAAATCTTTTAATATCT
TTTTTAAAACCTGTATTCATAGGTTTATTTTGTCATGTCTAATTTTATTAGAACATTTT
CTTATTCAATTTTTAAATTCTTAAAATTGAATTCAACCTACTATTACAGTTTTTCACTA
TTTTCTTTGCCTATTTTGCAAATACCTA

ISHahy15

ATGAATAATCTTTTAGATCAGAGTATAAAACTCAATTTATTGGATACTGAAACAATAGC
TATTGATGCTTCAAAACTGGAAGCATATGAGCGAGCTAAACCAAGGTCAAAGATAGATA
AGGAAAATAATTTTACCCCTGACTGGGGAACCAAATTTGATTCCCATAAAAATCAGATA
ACCTGGTATGGCTGGAAGATACATGCTGCAGTTGAAACTAAATCTGAAATACCAATAGC
TTTAACTTTAACTCCAGCTAATCATGCAGATAAAACTCAGGCAATACCATTAGTTGAAA
AGGTTAGTGAGTTTTTAGCTAAAAGAGATTTAATTAAACCTAAATATTGGACTATGGAT
TCAGGCTATGACTACACTGATATCTATGAGTATATTCTCTTTGAGCAGGAATCTCAGGC
CATTACCCCTATTAACAAACGTAATGCTAAACAACCACCAGCTGGATTTTATGATTTTA
AAGGAACACCAGTCTGCAGCGGTGGTTACAAAATGTATTACTGGGGTCATTACAAAGGA
GTCAACAAATTCAGATGTCCACATGTATGTGGTAAAGTTGACTGTATTCACGGTACTAA
GTGGTGTTCTGACAAAGATTATGGTCGAGTAACAAAAACAAGACCAAAAGATAATCCAA
GGTACATATCAACTCCACATAGAGATTCTAGAACCTGGAAAAAGATCTACAACAGAAGA
ACTAGTGTTGAAAGAACTTTTTCTAGATTGAAAGAGCATCTGAATTTAGCCAACTTAAC
TGTAATGGGAGCCAAAAAAGTTAAAACTCATTTACTGTTAAGCTCAATCAGTTTGATAG
CCGCAAGAATAGCAGCAGAGAAAATCAAGCTGCAAAATCAGGTTTTAGCTGCTTAA

BIBLIOGRAPHY

1. Siguier P, Gourbeyre E, Chandler M: Bacterial insertion sequences: their genomic impact and diversity. FEMS microbiology reviews 2014, 38(5):865-891.

2. McClintock B: Induction of Instability at Selected Loci in Maize. Genetics 1953, 38(6):579-599.

3. Cerveau N, Leclercq S, Bouchon D, Cordaux R: Evolutionary Dynamics and Genomic Impact of Prokaryote Transposable Elements. In: Evolutionary Biology - Concepts, Biodiversity, Macroevolution and Genome Evolution. Springer-Verlag Berlin Heidelberg; 2011: 291-312.

4. Aziz RK, Breitbart M, Edwards RA: Transposases are the most abundant, most ubiquitous genes in nature. Nucleic acids research 2010, 38(13):4207-4217.

5. Siguier P, Filee J, Chandler M: Insertion sequences in prokaryotic genomes. Current opinion in microbiology 2006, 9(5):526-531.

6. Gray YH: It takes two transposons to tango: transposable-elementmediated chromosomal rearrangements. Trends in Genetics 2000, 16(10):461-468.

7. Hickman AB, Chandler M, Dyda F: Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. Crit Rev Biochem Mol Biol 2010, 45(1):50-69.

8. Haren L, Ton-Hoang B, Chandler M: Integrating DNA: Transposases and Retroviral Integrases. Annu Rev Microbiol 1999, 53:245-281.

9. Chandler M, Mahillon J: Insertion Sequences Revisted. In: Mobile DNA II. Edited by Craig NL, Craigie R, Gellert M, Lambowitz A. Washington DC: American Society for Microbiology Press; 2002: 305-366.

10. Duval-Valentin G, Chandler M: Cotranslational control of DNA transposition: a window of opportunity. Molecular cell 2011, 44(6):989996.

11. Ichikawa H: Two domains in the terminal inverted-repeat sequence of transposon Tn3. Gene 1990, 86(1):11-17.

12. Nagy Z, Chandler M: Regulation of transposition in bacteria. Res Microbiol 2004, 155(5):387-398.

13. Mahillon J, Chandler M: Insertion Sequences. Microbiology and molecular biology reviews: MMBR 1998, 62(3):725-774.

14. Many Transposable Elements Have Common Characteristics [http://www.nature.com/scitable/content/many-transposable-elements-havecommon-characteristics-29563]

15. Turlan C, Chandler M: Playing second fiddle: second-strand procesing and liberation of transposable elements from donor DNA. Trends in microbiology 2000, 8(6):268-274.

16. Curcio MJ, Derbyshire KM: The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol 2003, 4(11):865-877.

17. Schatz DG: Antigen receptor genes and the evolution of a recombinase. Semin Immunol 2004, 16(4):245-256.

18. Smith MCM, Thorpe HM: Diversity in the Serine recombinases. Molecular Microbiology 2002, 44(2):299-307.

19. Boocock MR, Rice PA: A proposed mechanism for IS607-family serine transposases. Mob DNA 2013, 4(1):24.

20. Kersulyte Dangeruta, Asish K. Mukhopadhyay, Mutsinori Shirai, Teruko Nakazawa, Berg DE: Funcitonal organizaiton adn insertion specificity of IS607, a chimeric delement of helicobacter pylori. Journal of bacteriology 2000, 182(19):5300-5308.

21. Grindley NDF, Whiteson KL, Rice PA: Mechanisms of site-specific recombination. Annu Rev Biochem 2006, 75:567-605.

22. Bikard D, Loot C, Baharoglu Z, Mazel D: Folded DNA in action: hairpin formation and biological functions in prokaryotes. Microbiology and molecular biology reviews : MMBR 2010, 74(4):570-588.

23. Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B: Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nature reviews Microbiology 2013, 11(8):525-538.

24. Ton-Hoang B, Guynet C, Ronning DR, Cointin-Marty B, Dyda F, Chandler M: Transposition of IShp608, member of an unusual family of bacterial insertion sequences. The EMBO Journal 2005, 24(18):33253338.

25. Kersulyte D, Velapatino B, Dailide G, Mukhopadhyay AK, Ito Y, Cahuayme L, Parkinson AJ, Gilman RH, Berg DE: Transposable Element ISHp608 of Helicobacter pylori: Nonrandom Geographic Distribution, Functional Organization, and Insertion Specificity. Journal of bacteriology 2002, 184(4):992-1002.

26. Garcillan-Barcia M, And Cruz, F.: Distribution of IS91 family insertion sequences in bacterial genomes: evolutionary implicaitons. FEMS Microbiology Ecology 2002, 42:303-313.

27. Griffiths, Anthony J.F.; Gelbart, William M.; Miller, Jeffrey H.; Lewontin, Richard C.: Introduction to Genetic Analysis. 7th ed. New York: W H Freeman & Co; c1999.

28. Guilhot, C., B. Gicquel, J. Davies, and C. Martin. 1992. Isolation and analysis of IS6120, a new insertion sequence from Mycobacterium smegmatis. Mol. Microbiol. 6:107–113.

29. Guedon, G., F. Bourgoin, M. Pebay, Y. Roussel, C. Colmin, J. M. Simonet, and B. Decaris. 1995. Characterization and distribution of two insertion sequences, IS1191 and iso-IS981, in Streptococcus thermophilus: does intergeneric transfer of insertion sequences occur in lactic acid bacteria cocultures? Mol. Microbiol. 16:69–78.

30. Dalrymple, B., P. Caspers, and W. Arber. 1984. Nucleotide sequence of the prokaryotic mobile genetic element IS30. EMBO J. 3:2145–2149.

31. Caspers, P., B. Dalrymple, S. Iida, and W. Arber. 1984. IS30, a new insertion sequence of Escherichia coli K12. Mol. Gen. Genet. 196:68–73.

32. Berg, D. E., J. Davies, B. Allet, and J. D. Rochaix. 1975. Transposition of R factor genes to bacteriophage lambda. Proc. Natl. Acad. Sci. USA 72:3628– 3632.

33. Mollet, B., S. Iida, and W. Arber. 1985. Gene organization and target specificity of the prokaryotic mobile genetic element IS26. Mol. Gen. Genet. 201:198–203.

34. Arias-Palomo, E. and J. M. Berger (2015). "An Atypical AAA+ ATPase Assembly Controls Efficient Transposition through DNA Remodeling and Transposase Recruitment." Cell 162(4): 860-871.

35. Solinas, F., A. M. Marconi, M. Ruzzi, and E. Zennaro. 1995. Characterization and sequence of a novel insertion sequence, IS1162, from Pseudomonas fluorescens. Gene 155:77–82.

36. IS FINDER [https://www-is.biotoul.fr/]

37. Vogele, K., E. Schwartz, C. Welz, E. Schiltz, and B. Rak. 1991. High-level ribosomal frameshifting directs the synthesis of IS150 gene products. Nucleic Acids Res. 19:4377–4385.

38. Sekine, Y., N. Eisaki, and E. Ohtsubo. 1994. Translational control in production of transposase and in transposition of insertion sequence IS3. J. Mol. Biol. 235:1406–1420.

39. Polard, P., M. F. Pr`ere, M. Chandler, and O. Fayet. 1991. Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS911. J. Mol. Biol. 222:465–477.

40. Mira AK, Lisa. and Siv GE Anderson: Microbial genome evoluiton: sources of variability. Current opinion in microbiology 2002, 5(5):506512.

41. Moran NA, Plague GR: Genomic changes following host restriction in bacteria. Curr Opin Genet Dev 2004, 14(6):627-633.

42. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL et al: Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat Genet 2003, 35(1):32-40.

43. Simser JA, Rahman MS, Dreher-Lesnick SM, Azad AF: A novel and naturally occurring transposon, ISRpe1 in the Rickettsia peacockii genome disrupting the rickA gene involved in actin-based motility. Mol Microbiol 2005, 58(1):71-79.

44. Debets-Ossenkopp YJ, Pot RGJ, Westerloo DJ, Goodwin A, Vandenbroucke-Grauls CMJE, Berg DE, Hoffman PS, Kusters JG: Insertion of a Mini-IS605 and Deletion of Adjacent Sequences in the Nitroreductase (rdxA) Gene Cause Metronidazole Resistance in Helicobacter pylori NCTC11637. Antimicrobial Agents and Chemotherapy 1999, 43(11):2657-2662.

45. Barker CS, Pruss BM, Matsumura P: Increased motility of Escherichia coli by insertion sequence element integration into the regulatory region of the flhD operon. Journal of bacteriology 2004, 186(22):75297537.

46. Ziebuhr W, Krimmer V, Rachid S, Lo¨ßner I, Go¨tz F, Hacker J: A novel mechanism of phase variation of virulence in Staphylococcus

47. Epidermidis: evidence for control of the polysaccharide intracellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. Molecular Microbiology 1999, 32(2):345-356.

48. Van der Woude MW, Baumler AJ: Phase and antigenic variation in bacteria. Clin Microbiol Rev 2004, 17(3):581-611, table of contents.

49. Mormile MR: Going from microbial ecology to genome data and back: studies on a haloalkaliphilic bacterium isolated from Soap Lake, Washington State. Frontiers in microbiology 2014, 5:628.

50. Kichenaradja P, Siguier P, Perochon J, Chandler M: ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. Nucleic acids research 2010, 38(Database issue):D62-68

51. Varani AM, Siguier P, Gourbeyre E, Charneau V, Chandler M: ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. Genome biology 2011, 12(3):R30.

52. ISsaga – IS Semi-automatic Genomic Annotation [http://issaga.biotoul.fr/ISsaga2/about.php]

53. Argo Genome Browser [http://www.broadinstitute.org/annotation/argo]

54. NCBI: Our Mission [http://www.ncbi.nlm.nih.gov/home/about/mission.shtml]

55. The European Bioinformatics Institute [http://www.ebi.ac.uk/]

56. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M et al: Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic acids research 2008, 36(Web Server issue):W465-469.

57. Reverse Complement [https://www.bioinformatics.org/sms/rev_comp.html]

58. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006;34(Web Server issue):W609-12.

59. Korber B. (2000). HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers.

60. Phylogeny.fr Robust Phylogenetic Analysis For The Non-Specialist [http://www.phylogeny.fr/]

61. *CiVi: circular genome visualization with unique features to analyze sequence elements* Lex Overmars; Sacha A. F. T. van Hijum; Roland J Siezen; Christof Francke Bioinformatics 2015; doi: 10.1093/bioinformatics/btv249

VITA

Kody Bassett was born in Saint Robert, Missouri. He attended Waynesville High School in Waynesville, Missouri. In 2017 he graduated with a Bachelors of Science in Biological Sciences from Missouri University of Science and Technology located in Rolla, Missouri. After finishing his undergraduate degree, Kody continued his academic career by pursuing a Masters degree in Applied and Environmental Biology in 2017. He received his Masters degree in Applied and Environmental Biology from Missouri University of Science and Technology in May 2019.