Summer 2016

Characterization of insertion sequence IS605 in halanaerobium hydrogeniformans

Michael C. Sadler

Department:

Recommended Citation

Sadler, Michael C., "Characterization of insertion sequence IS605 in halanaerobium hydrogeniformans" (2016). *Masters Theses*. 7568.
https://scholarsmine.mst.edu/masters_theses/7568

CHARACTERIZATION OF INSERTION SEQUENCE IS605 IN *HALANAERОBIUM HYDROGENIFORMANS*

by

MICHAEL SADLER

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfilment of the Requirements for the Degree

MASTERS OF SCIENCE IN APPLIED AND ENVIRONMENTAL BIOLOGY

2016

Approved by

Ronald Frank, Advisor

Melanie Mormile

Dave Westenberg

ABSTRACT

Insertion sequences are the smallest prokaryotic transposable elements. These genes play a significant evolutionary role by promoting genome plasticity. Insertion sequences are highly diverse elements that have largely been uncharacterized. As such, the ability to accurately identify, annotate, and infer genomic impact of insertion sequences is lacking. The study of new insertion sequences contributes knowledge to their annotation and evolution. *Halanaerobium hydrogeniformans* is a unique organism with an abnormally high number of insertion sequences. A family of insertion sequences, IS200/605, showed several interesting distinctions from other elements in the genome, including severe open reading frame degradation, and was characterized in detail. This research uses bioinformatics tools to present an in depth characterization of a single insertion sequence family in *H. hydrogeniformans*. From these results insertion sequence activity can be inferred, including transposition capability, element interaction, and insertion sequence evolution.

ACKNOWLEDGMENTS

I would like to extend my greatest appreciation to Dr. Ronald Frank for his mentorship. He accepted me as his student when I had no background or experience in the tools or methodologies used in his lab. He spent a substantial amount of his time in guiding me through the research, the graduate program, and providing career advice. I could not have asked for a better advisor.

I would like to thank Dr. Melanie Mormile for serving on my committee and allowing me to spend time in her lab.

I would also like to thank Dr. David Westenberg for serving on my committee and providing essential early feedback on research.

TABLE OF CONTENTS

Page

APPENDICIES

LIST OF ILLUSTRATIONS

Page

LIST OF TABLES

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, which is responsible for the transposition activity, and they can carry non-transposition-essential genes known as accessory or passenger genes [1]. For some time after they were first described [2], transposable elements were thought to be junk DNA or selfish genes with little benefit to their hosts. It is now known that transposable elements play an important role in increasing genetic diversity by promoting 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 fossils (relics of transposable elements that have lost their ability to transpose) can 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 retrotransposon 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 terminal inverted repeats [6]. Many eukaryotic transposons are related to prokaryotic insertion sequences, and carry a variety of passenger genes [7].

1.2. INSERTION SEQUENCES

Insertion sequences are the smallest and simplest of prokaryotic transposable elements. Insertion sequences are highly diverse in structure and organization. Insertion sequences typically have an open reading frame (ORF), terminal inverted repeats, and direct repeats. Many insertion sequences also insert preferentially within their host genome. The differences between elements with regard to these features, as well as their catalytic mechanisms for transposition, are used to categorize insertion sequences into groups and families. There are 4 major groups and 29 distinct families. It is important to note that these are general insertion sequence characteristics, and do not apply to all insertion sequence families.

1.2.1. Organization. Insertion sequences are typically between 0.8 and 2.5 kb in size and carry a single open reading frame (ORF) required for transposition. Insertion sequence ORFs can be divided into structural domains that contribute to distinct functions. The N-terminal and C-terminal regions principally contain DNA-binding and catalytic domains, respectively. This orientation permits the coupling of synthesis and activity of the transposase [8], [9]. Further evidence of the purpose of this organization is that for a number of 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].

1.2.2. Terminal Inverted Repeats. With few exceptions, insertion sequences contain terminal inverted repeats (IR). These are generally imperfect IR of 10-40 bp in length near each terminus of the transposable element outside the ORF. Inverted repeats are short sequences that read the same 5' – 3' on each strand of DNA. The outermost base pairs are involved in strand cleavage and transfer during the transposition reaction, and the internal base pairs are recognized for transposase binding [11]. Additionally, endogenous insertion sequence promoters have been located in the terminal inverted repeat sequences upstream of the transposase gene which may provide a mechanism for auto regulation. Binding sites for host specific proteins have also been observed within or close to the IR that may also impact transposition activity or transposase expression [12].

1.2.3. Target Site Duplications. Another common feature to insertion sequences is a target site duplication that is generated on insertion. Staggered DNA cuts at the target site in the DNA backbone results in the duplication of the target DNA flanking the insertion sequence upon insertion. The target site duplication results in a direct repeat (DR). The size of the DR vary between families and elements, but typically range between 2-14 bp in length [13].

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 a 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 processes can result in a cascade of transposition events and numerous insertion sequences located proximally to one another [12]. The general structure of an insertion sequence is represented in Figure 1.1.

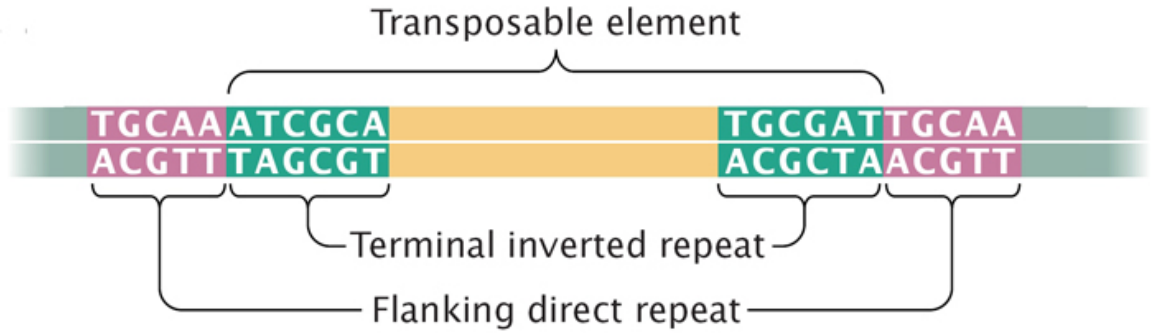


Figure 1.1. The general features of an insertion sequence containing inverted repeats and target site duplication (direct repeat) [14].

1.3. CATALYTIC CHEMISTRIES

Insertion sequences can be categorized into four groups based on their catalytic mechanisms for transposition. These groups are 1) the DDE, so called for the conserved catalytic DDE motif, 2) Y1, 3) S for their conserved tyrosine and serine residues at the catalytic site, and 4) Y2 group that shows similarity to proteins involved in rolling circle replication.

1.3.1. DDE. The majority of identified insertion sequences fall into the DDE family of transposases. Within this group there are dominant family members, including the IS3 and IS5 [1]. DDE family members feature a triad of negatively charged catalytic residues D (asp) D (asp) E (glu) that are highly conserved. The distance between each conserved residue is variable between families and highly conserved within families containing the DDE motif. Two transposition steps are common to all reactions catalyzed by DDE family members. The first is DNA cleavage through hydrolysis and the second is the attack of target site DNA by the free 3’OH on the element end. However, these family members transpose via a double stranded intermediate. Generating the 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 much less understood than their DDE counterparts. These 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, some IS607 family members also encode a second protein known as orfB or TnpB. This protein shows high sequence similarity to a protein encoded by members of the IS605 family. The TnpB 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 are members of a greater superfamily of endonucleases known as the HUH (H = Histidine, U = hydrophobic) endonuclease family.

This HUH superfamily acts exclusively on ssDNA, catalyzing DNA breakage and formation of a 5’ phosphotyrosine intermediate using the catalytic tyrosine residue and generating a free 3’OH at the cleavage site. Many HUH endonucleases 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].

Similar to the superfamily to which they belong, Y1 insertion sequences transpose via ssDNA intermediates and insert 3’ to a conserved, element-specific, penta- or tetranucleotide sequence. These Y1 family members also insert and excise preferentially from and into ssDNA [24], [25]. It is important to note that these transposable elements do not contain inverted repeats or generate target site duplications, common to the majority of identified insertion sequences.

1.3.4. Y2. Y2 insertion sequences, encoded by IS91, also fall within the HUH endonuclease superfamily. While Y1 transposases carry a single catalytic tyrosine, Y2 transposases carry two conserved tyrosine residues and appear to carry out transposition through a different mechanism. Y2 proteins also show similarities to proteins involved in rolling circle replication [1]. IS91 elements insert 3' to a conserved tetranucleotide sequence [26]. Relatively little is known about the transposition mechanisms of this family of insertion sequences compared to the more defined families.

See Figure 1.2 for the number of identified insertion sequences grouped by family and catalytic chemistry. See Figure 1.3 for the distribution of identified insertion sequences across prokaryotic genomes.

1.4. IS200/605

The IS200 and IS605 families of insertion sequences belong to the group of Y1 transposable elements briefly described in Section 1.3.3. The difference between these two families is that IS200 carries a single transposase gene (tnpA), while IS605 members encode a gene (tnpB) in addition to the tnpA The tnpB gene is dispensable for transposition [25].

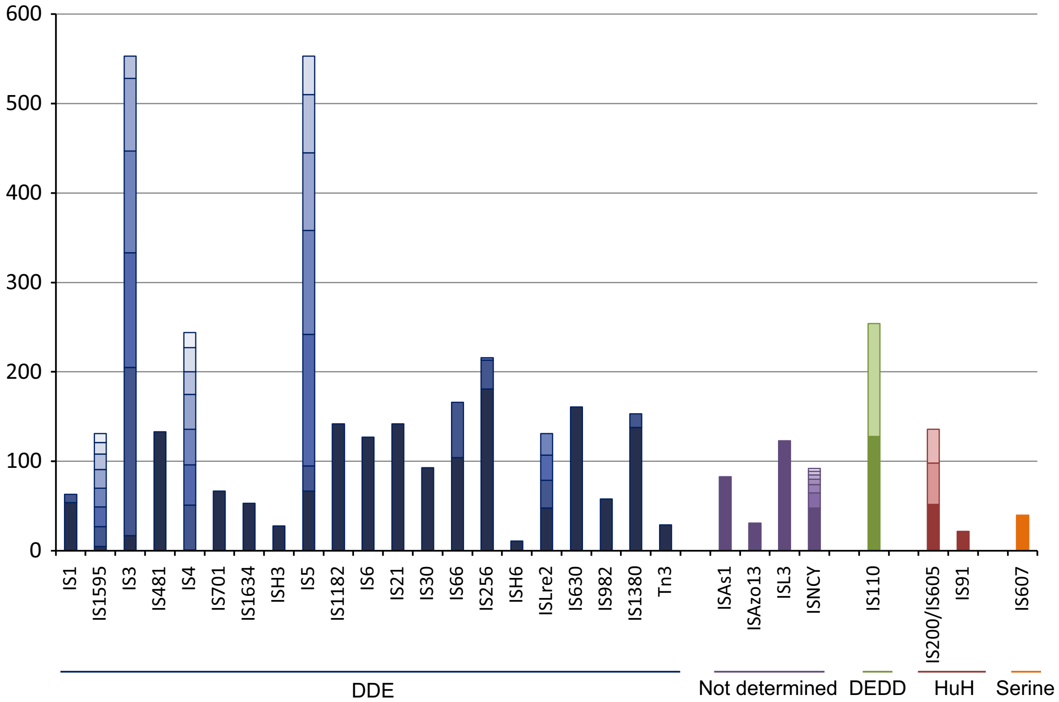


Figure 1.2. The number of identified insertion sequences grouped by catalytic chemistries and insertion sequence family [1]. DEDD represents a major subgroup of the DDE group. Shaded bars represent sub families

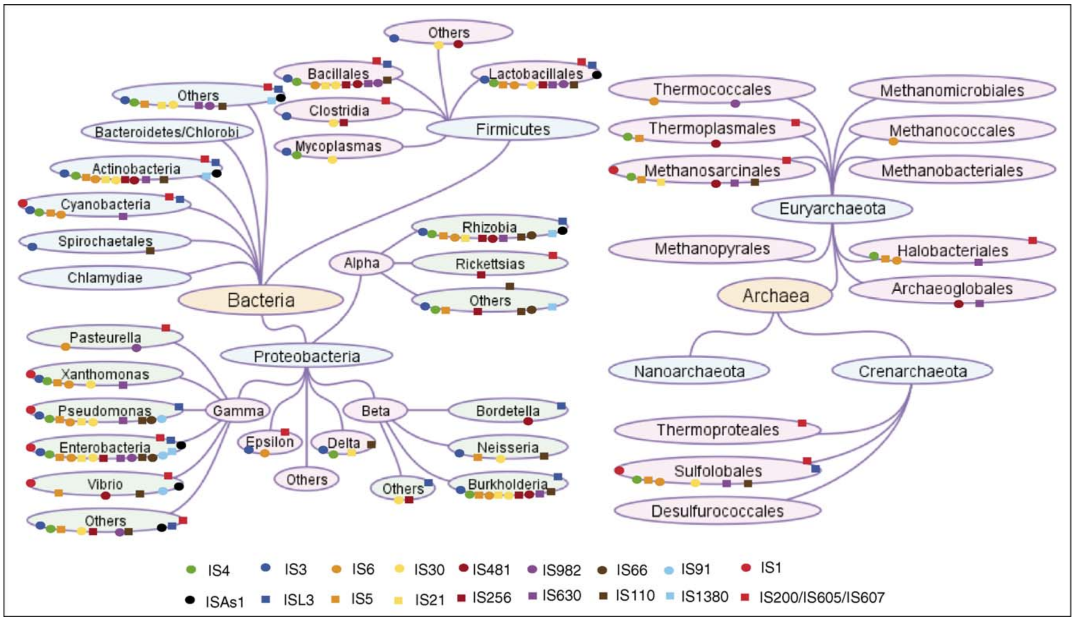


Figure 1.3. Distribution of identified insertion sequences families across prokaryotic genomes [5].

These elements do not contain terminal inverted repeats, nor do they generate target site duplications upon insertion. They contain secondary hairpin structure at both element ends that are necessary for transposition [24].

1.4.1. TnpA. The tnpA gene of the IS200 and IS605 families encodes the transposase. This protein contains the HUH motif and carries a single catalytic tyrosine. and inserts the element 3' to a specific tetra- or penta- nucleotide sequence [27].

The protein functions as an obligatory dimer [25]. For transposition to occur, each TnpA monomer binds an indispensable secondary hairpin structure present at each end of the element. In the well-characterized ISHp608 elements of the IS200/605 family, the sequence of these structures is the same at the left end (LE) and right end (RE) of the element [28], [24]. Transposition of the element occurs as circular ssDNA, and is strand specific [29]. The ability to differentiate between top and bottom strands in ISHp608 comes from a minor structural abnormality between the top and bottom strand [28]. Because this element transposes via a ssDNA intermediate, its transposition is coupled to the replication of host DNA, during which it has a preference for lagging strand template insertion [30]. These elements also have increased transposition rates with DNA repair mechanisms that produce large stretches of ssDNA [31].

The obligatory dimer of two TnpA proteins forms two functional conformations, a cis and a trans formation. In trans, the catalytic site is constructed of the HUH motif from one monomer, and the catalytic tyrosine from the other [32]. Conformation change from trans to cis results in strand breakage and the formation of 2 phosphotyrosine bonds, The reverse conformation change results in the insertion of the element into the target site [23], [33], [34].

Insertion into a new location starts with target recognition. Recognition is a result of DNA-DNA interaction of a tetranucleotide sequence 5’ to the LE hairpin structure and a target sequence. The target sequence is dependent on the sequence 5’ to the hairpin structure. When the active tetranucleotide sequence of the element was altered, new insertion sites were targeted [32]. Insertion occurs without target site duplication, and element excision precisely seals donor DNA. This transposition does not require host cell DNA repair factors [25].

1.4.2. TnpB. An ORF, known as orfB or tnpB is often encoded proximal to the Y1 tnpA of IS200. When together, they represent the IS605 family. For these family members, hairpins necessary for element transposition are found external to the two ORFs. OrfB is approximately 1200 nt in length and is dispensable for transposition [25]. OrfB is located in successive, divergent, or overlapping orientation with respect to tnpA [24]. Until recently the function of TnpB was largely unknown. There is now evidence to suggest that TnpB plays a role in transposition regulation of IS200 and IS605 elements. TnpB has been shown to decrease both excision and insertion of TnpA, decreasing excision approximately 200 times more efficiently than insertion [35]. The mechanism of how TnpB inhibits transposition is unknown but it is speculated that TnpB protein could competitively bind the hairpin structures at either end of the element, or bind the TnpA protein itself. TnpB directly impacts the activity of TnpA and does not act through host mediated factors [35].

The TnpB polypeptide typically contains 3 domains, an N-terminal HTH, a central domain, and a C-terminal zinc finger. TnpB is most variable in the N-terminal and most conserved in the C-terminal domain. The inhibitory action of TnpB on TnpA transposition is strictly dependent on the integrity of the zinc finger domain [35]. Zinc fingers perform a broad range of functions, primarily as interaction modules binding to a wide variety of nucleic acids, proteins, and other molecules [36].

The TnpB protein has been associated with members of the IS607 family. This family utilizes a different mechanism of transposition than the IS200/605 families. When associated with the transposases of IS607, tnpB is dispensable for transposition [20]. Additionally, homologues of tnpB are found in diverse eukaryotic transposable elements [37].

Several reported elements encode tnpB as the only ORF. This has resulted in the labeling of TnpB as a putative transposase gene (IS1341, IS809, and IS1136). However, the evidence for TnpB mediated transposition is absent and it is likely that these elements are non-autonomous derivatives of IS605 or IS607 families. See Figure 1.4 for a representative structure of a IS605 family member with divergent ORFs.

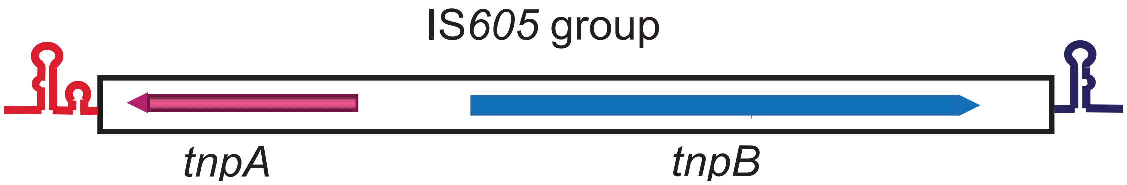


Figure 1.4. An illustration of an IS605 family member with divergent ORFs [1].

1.5. MINIATURE INVERTED REPEAT TRANSPOSABLE ELEMENTS

Miniature inverted repeat transposable elements (MITES) are partial copies of transposable elements that typically contain only the sequence or structures necessary for transposition. MITES can be impactful to host genomes as they can influence gene expression, alter mRNA stability, or influence transcription termination. In genomes without full length parent copies it can be extremely difficult to identify MITES, as they are often only present as short inverted repeat sequences [5]. MITES 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.

1.6. INSERTION SEQUENCE ANNOTATION

Insertion sequences, their nonautonomous derivatives, and MITES, represent a substantial portion of bacterial and archeal genomes. Insertion sequences are highly diverse with respect to their transposases and element ends. Because of this diversity accurate insertion sequence identification and annotation is difficult. Transposase genes of insertion sequences are often mislabeled as integrases, recombinases, pseudogenes, and hypothetical proteins [38]. The element ends containing inverted repeat and direct repeats, which are smaller and more diverse than the proteins themselves, are typically overlooked. It is even more rare that MITES are identified.

The development of high throughput sequencing has led to the generation of thousands of complete genomes and metagenomes. With the sheer quantity of insertion sequences and MITES, accurate identification and annotation requires more sophisticated methods than those currently available [39].

Several semi-automatic methods have been developed to aid in the identification of insertion sequences. Two of these are OASIS (Optimized Annotation System for Insertion Sequences) [40], and ISsaga (Insertion Sequence Semi-Automatic Genome Annotation) [38]. Enhanced methods to better visualize and organize these elements are being developed [41]. The underlying issue with these methods is that they rely on the quality of insertion sequence database libraries. While these methods expedite the identification of known insertion sequences, unique insertion sequences and MITES can be misidentified or completely overlooked. Even MITES of known and well characterized insertion sequences can be overlooked because they show such low similarity to the parent element, or a parent element may not be present in the genome being surveyed.

The inability to accurately identify insertion sequences and MITES leads to a severe bias towards characterized and complete copies of insertion sequences in surveys of insertion sequences across genomes [38]. Until more sophisticated methods are developed, or insertion sequence databases become more complete, a significant amount of manual curation is necessary when identifying and annotating insertion sequences and their MITES.

1.7. GENOMIC IMPACT OF TRANSPOSABLE ELEMENTS

Transposable elements were originally viewed as selfish DNA, serving little to no purpose to their host. However, it is now understood that transposable elements play a significant role in promoting genetic diversity, structure, and genomic plasticity [42].

1.7.1. Genomic Streamlining. Insertion sequences experience rapid expansion and loss within host genomes. This is accompanied by genomic rearrangement, and gene inactivation. With time, insertion sequences experience deletion that can be accompanied with deletion of host DNA, resulting in genome reduction. Insertion sequence degradation will lead to the development of non-functional, or non-autonomous elements, which are eventually cleared from the genome. The increase in transposable elements and reduction in genome size is most noted in new bacterial endosymbionts, and is only permissible with an increase in host dependence [5]. The relaxed selective pressure of new endosymbionts permits both the expansion of insertion sequences, and the ensuing genome reduction.

It has been observed that transposable element numbers increase in new bacterial endosymbionts compared to free living cells [43], 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 [44].

1.7.2. Insertional Mutation. Insertion sequences can effect genomes through direct impedance by inserting into and disrupting genes. Insertion sequence mediated disruption in a *Rickettsi* species resulted in non-pathogenicity by insertion into virulence genes [45], as well as a metronidazole resistant *H. pylori* by insertion within a gene necessary for pro-drug activation [46].

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 endogenous transcriptional promoters [12], and their insertion leads to changes in expression of flanking genes. Insertion sequences can also change expression by activating or inactivating promoter or repressor genes [47].

1.7.4. Genomic Rearrangement. Insertion sequences also impact genomes through a variety of chromosomal architecture changes. This activity stems from the multiple copies of elements with high sequence similarity. Recombination between two 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 [48], [49]. 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 [50]. Alternative transposition mechanisms can also result in inter-element sequence duplication [6].

1.8. INVASION – EXPANSION – EXTINCTION CYCLES

To a degree, insertion sequences provide a selective advantage to their host by increasing diversity and genomic plasticity [51]. However, insertion sequences are in general thought to be more damaging than beneficial to their host, and their persistence in genomes is questioned. It is hypothesized that insertion sequences undergo periodic invasion, expansion, and extinction cycles. These cycles are characterized by introduction to a new genome through horizontal gene transfer, expansion through replicative transposition, and extinction through unknown methods that eliminate, or degrade insertion sequences beyond recognition in a genome.

Insertion sequences have an extremely high, nearly identical, sequence similarity within genomes [52]. This unusually high sequence similarity is not due to evolutionary constraints, as insertion sequences between genomes show significant sequence divergence. Gene conversion, which is the homogenization of nearly identical sequences through recombination, has been proposed to be a mechanism for sequence conservation in obligate mutualistic endosymbionts [53]. However, evidence of gene conversion in insertion sequences of free living hosts is absent. Insertion sequences also show higher sequence conservation than gene duplicates, which would be subject to the same level of gene conversion. Additionally, successful transposition rates of insertion sequences are higher than substitution rates [54].

This leads to the hypothesis that insertion sequences are newly acquired to most genomes, and that they invade and rapidly expand within a genome. The patchy distribution of insertion sequences across genomes of highly related strains [55], [56], supports this hypothesis, and also shows that insertion sequences are not sustainable within a genome. Insertion sequences are selected against over time through down regulation of transposition, excision, and the preference for the majority to act in cis. The expansion of insertion sequences is permitted only because of the temporary benefits they might provide through genomic rearrangement and transfer of beneficial genes. As such, their persistence in the environment is dependent on horizontal gene transfer [54].

1.9. HALANAEROBIUM HYDROGENIFORMANS

Halanaerobium hydrogeniformans is an extremophile isolated from a haloalkaline lake in Washington State. This organism has gained attention due to its unique metabolic capabilities and potential for industrial applications. After the sequence of the genome was determined, 2463 genes were annotated. Among them were 72 transposase genes belonging to eight insertion sequence families [57]. This puts the bacterial genome at approximately 3% transposable elements, which is higher than in most bacterial genomes [5]. Because transposable elements are often misidentified, it was suspected that 72 transposable elements was a conservative estimate of the actual number encoded in the genome.

1.10. DATABASES AND BIOINFORMATIC TOOLS

1.10.1. NCBI. The National Center for Biotechnology Information (NCBI) was developed by the National Institutes of Health (NIH) after the need for computerized information processing in modern research was realized. NCBI's mission became "finding new approaches to deal with the volume and complexity of data in providing researchers with better access to analysis and computing tools to advance understanding of our genetic legacy and its role in health and disease." Data from the European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ) is shared with NCBI. NCBI is also host for numerous automated DNA and protein tools such as blastp, blastn, RefSeq, and ORF-finder. NCBI also provides access to DNA and protein sequences, mapping, structural data, and phylogenetic outputs [58].

1.10.2. 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 data and much of the software from EBI can be downloaded and installed locally, or run via online servers. The tools provided span DNA/RNA alignments, molecular structures, protein sequences, families, and motifs, taxonomy, and systems pathways [59].

1.10.3. PFAM. Pfam is a database containing protein families. Protein families are sets of proteins that share regions of high amino acid sequence similarity that are generated from multiple sequence alignments and hidden Markov models. These conserved regions can be used in the prediction of protein functionality when compared to known proteins [60].

1.10.4. Phylogeny.fr. Phylogeny.fr provides free web based phylogenetic analysis tools for the non-specialist. It permits automated and semi-automated phylogenetic relationships to be constructed between nucleotide or protein sequences using a multiple alignment process and can provide a newick output for various tree viewers [61].

1.10.5. ISfinder. ISfinder is an online public database providing general features (size, target sequence, family, inverted repeat sequences) for insertion sequences isolated from bacteria and archea. They rely on the scientific community to deposit sequences and information of characterized insertion sequences to enrich the database. ISfinder also provides a program ISbrowser that can be used to view identified and predicted insertion sequences in sequenced genomes [41], [62].

1.10.6. ISsaga. ISsaga is a tool of ISfinder that was developed to accurately identify and annotate insertion sequences with the use of a high-quality semi-automatic annotation system. This uses the ISfinder database to provide general prediction and annotation tools 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. Because the annotation accuracy of ISsaga is limited to the insertion sequence library of ISfinder, insertion sequences predicted by ISsaga have to be confirmed manually before being added to the ISfinder database [38], [63].

1.10.7. ExPASy. The Swiss Institute of Bioinformatics (SIB) has developed the Expert Protein Analysis System (ExPASy) web portal, offering access to numerous scientific resources, databases, and software tools. These tools are for areas of biology research including proteomics, genomics, phylogeny, structure, and more [64].

1.10.8. Sequence Alignment. Sequence alignments are made to determine the relatedness between two or more DNA or protein sequences. The services provided by the EBI offer programs for pairwise and multiple sequence alignment. Pairwise alignments are ideal for highlighting regions of similarity or dissimilarity that may confer a functional, structural, or evolutionary relationship between two sequences. These programs would include Needle, Stretcher, Water, Matcher, and LALIGN. The differences in these programs is that they utilize slightly different parameters to align sequences. Multiple sequence alignments are used to determine homology and evolutionary relatedness between sequences. These include Clustal Omega, an alignment program for three or more sequences [59].

1.10.9. Mfolds. DNA and RNA can contain secondary structure that is functional in a variety of biological processes. Mfolds and UNAFold are free web based programs developed to identify possible secondary structure and predict under what conditions they might form [65].

1.10.10. Argo. Argo is a Java based genome browser developed by The Broad Institute for viewing and annotation of whole genomes. It displays the sequence and annotation of DNA tracks. Files can be uploaded in SAM/BAM, FASTA, Genbank, GFF, BLAST, BED, WIG, and Genscan formats. 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 [66].

1.11. SUMMARY

This thesis presents a detailed characterization of an IS200/605 family members within *H. hydrogeniformans*. This family was selected for detailed characterization because of the unique characteristics of Y1 transposases. Six Y1 elements were originally annotated in the genome. After investigation this number rose to 23 elements and 1 MITE. Many of the 605 elements were misidentified by insertion sequence annotation software, and exhibit unique disruptions and fragmentation not typically observed in insertion sequences. The phylogeny of these elements in comparison to their structural differences suggests recombination between elements is occurring. These elements differ from reported IS200/605 family members in that their element ends are unique, and do not share common sequence between the right and left ends. This work is a detailed survey of an IS605 family of elements not reported elsewhere and provides a look at how insertion sequences might degrade within host genomes.

2. MATERIALS AND METHODS

2.1. INSERTION SEQUENCE IDENTIFICATION

The *Halanaerobium hydrogeniformans* genome sequence is recorded at the National Center for Biotechnology Information (NCBI), accession number CP002304.1. All genes annotated as insertion sequence, transposase, and integrase were used for a BLAST 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 open reading frame (ORF) from each different insertion sequence group was used for a BLAST search against the *H. hydrogeniformans* genome to identify partial insertion sequences that were annotated as pseudo or hypothetical genes. 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 originally annotated.

The elements belonging to the Y1 family were chosen for further investigation due to the numerous members present in the genome. This family was also chosen because of its distinct characteristics and the significant sequence dissimilarities between their replicates. Dissimilar replicates are inconsistent with reported high sequence similarity of insertion sequence between members of the same family within a bacterial genome [52].

The element families that were investigated in detail were given loci numbers for organization and further reference. Loci numbers were sorted 1-23 moving 5’-3’ from the origin of replication on the + strand.

2.1.1. Element Ends. The ends of a Y1 insertion sequence extend beyond the ORF. The element ends are defined as the nucleotide sequences of the element outside the ORF. These were identified by extracting 1000 nucleotides 5’ and 3’ of each ORF and aligning to identify the extent of homology between elements.

2.1.2. MITES. Miniature Inverted Repeat Transposable Elements (MITES) were identified by querying the genome with the element ends. Identified ends were matched with their corresponding ORFs. Element ends without corresponding ORFs were marked as potential MITES and examined further.

2.2. GENOME BROWSER

The Argo Genome Browser was used to visualize the genome of *Halanaerobium hydrogeniformans*. The genome was uploaded into Argo in Genbank format. Genes of interest were marked and categorized for further use. Visualization of gene positions allowed for a preliminary survey for insertion sequence position and proximity patterns. The genome browser was used to extract nucleotide and conceptual protein sequences for phylogenetic and alignment uses [66].

2.3. BLAST

Chosen sequences are aligned against a target database using a Basic Local Alignment Search Tool (BLAST). Databases can be queried with protein or nucleotide sequences.

For blastp, a conceptual protein sequence is used to query a protein database. This is used to identify potential gene products and conserved domains. Megablast is used to query a nucleotide sequence for closely related sequences for identification, working best if sequences show a 95% or higher similarity. Megablast was used to identify insertion sequence replicates within the genome. Discontiguous megablast is similar to megablast but allowing for greater mismatches and is intended for sequences with low similarity and cross-species comparisons. Discontiguous megablast was used to search for insertion sequence replicates that were misidentified or not annotated. 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. Blastn was used to search for MITES and element fragments against the genome. These BLAST tools are available free for use at the National Center for Biotechnology Information (NCBI). Algorithm parameters for BLAST searches used are in Table 2.1

2.4. ALIGNMENTS

Alignments were made between two or more protein sequences or two or more nucleotide sequences. Alignments are useful in comparing sequence similarity and structural differences. A number of alignment programs were used for pairwise and multiple sequence alignments. EMBOSS Needle and EMBOSS Stretcher utilize a Needleman-Wunsch algorithm to search for optimal global alignment between two sequences.

Table 2.1. Algorithm parameters for BLAST searches.

BLAST	**Blastp**	**Megablast**	**Discontinuous Megablast**	**Blastn**
Max Target Sequences	100	100	100	100
Expect Threshold	10	10	10	10
Word Size	6	28	11	11
Max Matches	0	0	0	0
Match/Mismatch	N/A	1, -2	2, -3	2, -3
Scoring Matrix	BLOSUM62	N/A	N/A	N/A
Gap Cost	Existence: 11 Extension: 1	Linear	Existence: 5 Extension: 2	Existence: 5 Extension: 2

Stretcher uses modifications that permit larger sequences to be globally aligned. LALIGN is a program for pairwise sequence alignment optimized for local alignment between two sequences [67]. Clustal Omega and Kalign are programs used to globally align multiple sequences [68]. All alignment programs are freely available for use from the European Bioinformatics Institute. Parameters and options used for alignment programs are found in Table 2.2 and Table 2.3.

Table 2.2. Alignment options for Clustal Omega.

Program	**Clustal Omega**
Dealign Input Sequences	NO
Clustering Guide Tree	YES
Clusteiring Iterations	YES
Combined Iterations	0
Tree Iterations	Default
HMM Iterations	Default

2.5. OPEN READING FRAME DISRUPTION

Insertion sequences can insert within genes disrupting the ORF. Automated identification of disrupted genes can be difficult. To identify if any of the Y1 insertion sequences inserted within a gene, 1000 nucleotides on either side of the insertion sequence (-1000/+1000) were extracted and spliced together. The 2000 nucleotide sequence frame was then searched with ORF Finder, a tool freely available for use from NCBI. Any ORF extending through position 0 (the middle of the extracted sequence) of the constructed ORF was conceptually translated and subjected to a blastp search against the NCBI database to identify potential protein products.

Table 2.3 Alignment parameters and options for pairwise alignment programs.

Program	Kalign	Needle	Stretcher	Water	Matcher	LALIGN
Gap Open	80	10	16	10	16	-12
Gap Extension	3	0.5	4	0.5	4	-4
Terminal Gap	3	NA		NA	NA	NA
Bonus Score	0	NA	NA	NA	NA	NA
Matrix	N/A	DNAfull	DNAfull	DNAfull	DNAfull	(+ 5) / (-4)
End Gap Penalty	NA	FALSE	NA	NA	NA	NA
End Gap Open	NA	10	NA	NA	NA	NA
End Gap Extension	NA	0.5	NA	NA	NA	NA
Alternatives Matrix	NA	NA	NA	NA	1	NA

2.6. PHYLOGENETIC ANALYSIS

Phylogenetic analysis was conducted with Phylogeney.fr. Extracted nucleotide sequences from insertion sequence ORF were input in FASTA format. Relationships of sequences were made using a MUSCLE sequence alignment without Gblock curation, and a maximum likelihood phylogenetic tree construction. Phylogenetic analysis was performed with the “one click” option for speed and alignment optimization [61], [69]. Mobile Elements in the genome showing significant deterioration were excluded from phylogenetic analysis, as the nucleotide sequences of these elements were too short to construct an accurate phylogenetic relationship.

2.7. SECONDARY STRUCTURE IDENTIFICATION

External to the ORF are conserved insertion sequence ends. In Y1 elements these ends contain hairpin structures necessary for transposition. Regions of the element ends showing potential for hairpin formation were identified by aligning the element left and right end nucleotide sequence with its respective reverse complement. The pairwise alignment program LALIGN was used to scan for regions with emphasis on local alignment. Regions showing significant alignment to their reverse complement were visually identified and subsequently examined with Mfolds, a DNA folding program, to view the potential physical structures. Mfolds DNA folding form was used under default conditions [65], [70].

3. RESULTS

3.1. INSERTION SEQUENCE IDENTIFICATION

ISsaga identified 16 insertion sequence families in *Halanaerobium hydrogenifomrans.* Initial observations reveal that these families are composed of few individual elements with varying levels in copy number.

Manual curation identified fewer families, with approximately the same number of total insertion sequences. Of note, ISsaga identified the presence of IS200/605, IS1341, and IS607 family members. In contrast, manual annotation resulted in the identification of one IS200 family member, and 22 IS605 family members. After detailed characterization, it was discovered that ISSaga misidentified these elements as there were no elements belonging to the IS1341 or IS607 families in *H. hydrogeniformans*. All misidentified elements showed high sequence similarities to the IS605 members. 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. Insertion sequence families IS1341, IS605, and IS607 are highlighted.

Detailed characterization of insertion sequences in *H. hydrogeniformans* was limited to the IS200 and IS605 family members. Each identified insertion sequence was given an independent locus number corresponding to its relative position to other detailed insertion sequences and the origin of replication. The elements are labeled locus 1-23 with increasing distance from the origin of replication. The locus numbers for each element, as well as some of the elements characteristics which are further discussed, are outlined in Table 3.2

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**

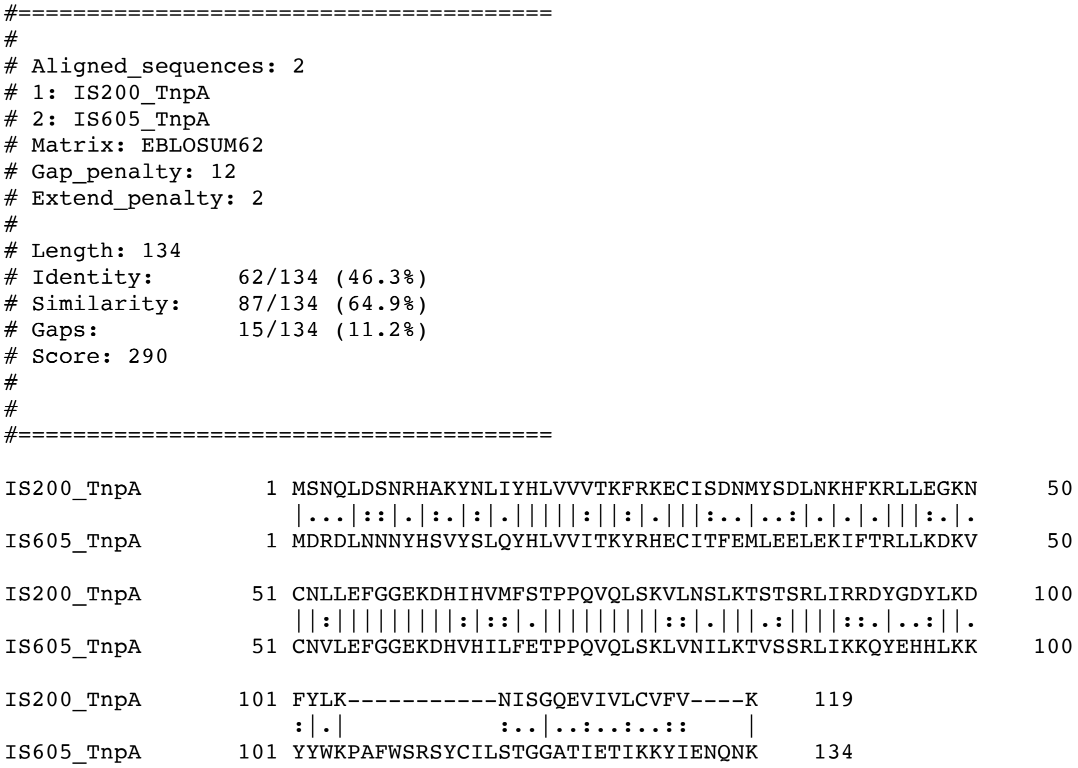
3.2. TnpA

There exist two different tnpA open reading frames (ORF). One belonging to an IS200 (locus 07), Accession number ADQ14068.1, in which it is the sole product of the insertion sequence.

Table 3.2. Characteristics of IS200 and IS605 elements in *H hydrogeniformans*.

Locus	**tnpA type**	**tnpB**	**LE**	**RE**	**tnpA halsa**	**tnpB halsa**	**Leading/ Lagging (tnpA)**
1	Type 2	1B	consensus	type 2	Halsa_0245	Halsa_0244	Lead
2	Type 5	2C	consensus	type 1	N/A	Halsa_0258	Lead
3	Type 5	2C	consensus	type 1	N/A	Halsa_0296	Lag
4	Type 5	2A	consensus	type 2	N/A	Halsa_0322	Lead
5	Type 5	1A	consensus	type 1	N/A	Halsa_0445	Lead
6	Type 5	2A	consensus	type 2	N/A	Halsa_0509	Lag
7	IS200_TnpA	N/A	unknown	unknown	Halsa_0613	N/A	Lag
8	Type 5	1A	consensus	type 1	N/A	Halsa_0624	Lag
MITE	NA	NA	Hairpin	type 2	NA	NA	Lag
9	Type 1	2A	consensus	type 1	Halsa_0741	Halsa_0742	Lag
10	Type 5	1A	consensus	type 1	N/A	Halsa_0809	Lag
11	Type 5	3	consensus	type 1	N/A	Halsa_0886	Lag
12	Type 5	3	consensus	type 1	N/A	Halsa_1064	Lag
13	Type 3	2A*	consensus	type 1	Halsa_1089	Halsa_1090	Lag
14	Type 5	MISC	consensus	type 2	N/A	Halsa_1216	Lead
15	Type 4	3*	consensus	MISC	Halsa_1228	Halsa_1227	Lead
16	Type 5	1A	consensus	type 1	N/A	Halsa_1236	Lag
17	Type 5	2B	consensus	type 1	N/A	Halsa_1482	Lag
18	Type 5	2B*	consensus	type 1	N/A	Halsa_1629	Lead
19	Type 5	2B	consensus	type 1	N/A	Halsa_1739	Lag
20	Type 2	2B	consensus	type 1	Halsa_2178	Halsa_2179	Lead
21	Type 5	3	consensus	type 2	N/A	Halsa_2207	Lead
22	Type 5	3	consensus	type 1	N/A	Halsa_2220	Lead
23	Type 5	3	consensus	type 1	N/A	Halsa_2306	Lag

The other tnpA belonging to the IS605 members, accession number WP_013405283.1, of which there are 22 complete, partial, or fragmented copies. Each TnpA protein contains a single Y1_Tnp superfamily domain. These will be referred to as the IS200 tnpA, and the IS605 tnpA. A protein alignment of each TnpA type is shown in Figure 3.1.



```
#=======================================
#
# Aligned_sequences: 2
# 1: IS200_TnpA
# 2: IS605_TnpA
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 134
# Identity:      62/134 (46.3%)
# Similarity:    87/134 (64.9%)
# Gaps:          15/134 (11.2%)
# Score: 290
#
#
#=======================================

IS200_TnpA         1 MSNQLDSNRHAKYNLIYHLVVVTKFRKECISDNMYSDLNKHFKRLLEGKN     50
                     |...|::|.|:.|:|.|||||:||:|.|||:..|..:|.|.|.|||:.|.
IS605_TnpA         1 MDRDLNNNYHSVYSLQYHLVVITKYRHECITFEMLEELEKIFTRLLKDKV     50

IS200_TnpA        51 CNLLEFGGEKDHIHVMFSTPPQVQLSKVLNSLKTSTSRLIRRDYGDYLKD    100
                     ||:|||||||||:|::|.|||||||||::|.|||.:||||::.|..:||.
IS605_TnpA        51 CNVLEFGGEKDHVHILFETPPQVQLSKLVNILKTVSSRLIKKQYEHHLKK    100

IS200_TnpA       101 FYLK-----------NISGQEVIVLCVFV----K    119
                     :|.|           :..|..:..:..::    |
IS605_TnpA       101 YYWKPAFWSRSYCILSTGGATIETIKKYIENQNK    134
```

Figure 3.1. IS200 TnpA and IS605 TnpA alignment.

3.2.1. IS200 tnpA. The IS200 tnpA ORF is 360nt long, consistent with other reported IS200 family members. Because this insertion sequence occurs without replicates, does not produce target site duplications, or contain inverted repeats, the element ends could not be identified. An attempt to identify secondary structures was made by aligning the nucleotide sequence on either end of the ORF with each other and each end with its reverse complement (Appendix A). However, regions showing significant alignment could not be identified above background levels. Additionally, it is unknown if any sequence showing alignment was part of the element ends.

3.2.2. IS605 tnpA. There exist 5 sub-types of the IS605 tnpA, as characterized by ORF structural differences, for a total of 22 individuals. Each subtype has a complete or partial divergent tnpB ORF. The 5 subtypes are described below and visualized in Figure 3.2 where blocks and triangles indicate 5' and 3' orientation. Full sequence alignments of all IS605 tnpA types are found in Appendix B.

Type 1 IS605 TnpA is a single replicate at locus 09 and is 405nt in length. This is the only 605 tnpA that could produce a functional protein as types 2-5 show significant degradation in the ORF.

Type 2 IS605 tnpA has two replicates (loci 1 and 20). These ORFs align with the most 3' 234 nucleotides of type 1, and are the missing 171 nucleotides from the 5' end.

Type 3 IS605 tnpA exists as a single replicate at locus 13. Type 3 ORF is missing 171 nucleotides from the 5', 114 nucleotides from the 3' end, and aligns with the central 120 nucleotides of type 1.

Type 4 IS605 tnpA also occurs as a single replicate at locus 15, aligning with the most 3' 108 nucleotides of type 1.

Type 5 IS605 tnpA is the most commonly occurring with 17 individuals. It is also the most fragmented of the five types. Opposed to types 1-4, type 5 IS605 tnpA does not annotate as a pseudo or hypothetical gene by genomic annotation software or by ISsaga. This type is 122 nucleotides long, aligning with the most 5' 63 nucleotides and the most 3' 59 nucleotides of type 1.

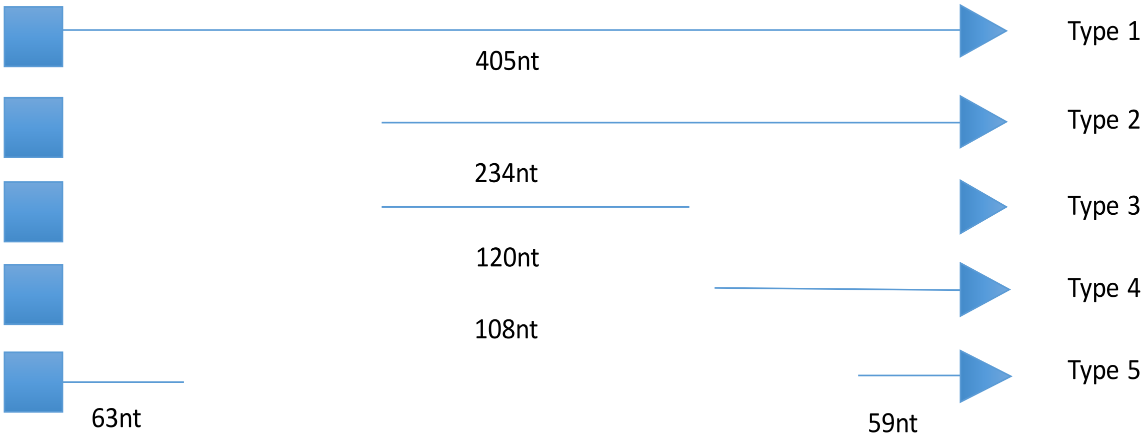


Figure 3.2. Relative IS605 tnpA sequence structures.

3.3. IS605 tnpB

There are 9 different 605 tnpB open reading frames present in the genome totaling 22 copies, each with a corresponding complete/partial/fragmented 605 tnpA (see table 3.2 for tnpA/tnpB pairings). These 9 different 605 tnpBs can be sorted into 3 primary groups and one miscellaneous group. These groups are described below and can be visualized in Figure 3.3 where blocks and triangles indicate 5' to 3' orientation. Full sequence alignments for all tnpB types are found in Appendix C

Type 1A tnpB has 4 replicates (loci 05, 08, 10, and 16). This ORF is 1254 nucleotides in length. This is not the most commonly occurring tnpB but it is the ORF most likely to produce a functional protein as types 2-3 and the miscellaneous group are sufficiently disrupted. This ORF encodes a protein containing three domains, a large ORFB_605 superfamily domain, a 605 central region, and a terminal Zn-ribbon binding domain. The element at locus 08 has inserted into and disrupted a sigma 54 interacting domain containing protein.

Type 1B tnpB is a single copy (locus 01) that aligns with type 1A ORFs. However, it contains a single nucleotide insertion at position 465 resulting in a frame shift and early translation termination.

Type2A tnpB has 3 replicates (loci 04, 06, and 09) and is 1382 nucleotides in length. These tnpB sequences align with type 1A ORFs with the exception of 2 additional 64 nucleotide inserts at position 433 and 1064. These inserts will be referred to as the left insert (LI) and right insert (RI) respectively.

Type 2A* tnpB is a single replicate (locus 13) and aligns with type 2A ORFs. It is classified as a type2A because it contains both LI and RI. It is denoted as a 2A* because it also is missing 173 nucleotides starting at nucleotide position 151.

Type 2B tnpB has 3 replicates (loci 17, 19, and 20) and has an ORF of 1318 nucleotides in length. This ORF aligns with type 2A tnpB but only contains the LI.

Type 2B* tnpB is a single replicate (locus 18) and has the same ORF and LI as type 2B ORFs. This element is denoted separately from type 2B because the ORF is disrupted by an insertion sequence 2.6kb in length. This sequence was identified manually and by ISsaga as a IS21 family member. Extraction of this element reveals that the remainder of the ORF aligns with other type 2B ORFs. Interestingly, this putative IS21 mobile element occurs in 3 replicates and is proximal to an IS605, IS256, IS200, and IS3.

Type 2C tnpB occurs in 2 replicates (loci 02, and 03) and is 1318 nucleotides in length. This tnpB aligns with type 2A ORFs with the exception that it contains only the RI.

Type 3 tnpB has 5 replicates (loci 11, 12, 21, 22, and 23) and is 724 nucleotides in length. This element aligns with type 2A ORFs with the exception that it contains a hybrid insert (HI) at position 433 and is missing the 463 nucleotides that exist between the LI and RI of type 2A. These inserts are further discussed in Section 3.4.

Type 3* tnpB is a single replicate (locus 15) and is classified as a type 3 tnpB because of its hybrid insert and absence of an interior sequence. This element is denoted separately from other type 3 ORFs as it is in a more progressed state of deterioration than the other type3 tnpBs. It totals 499 nucleotides in length, lacking a 173 nucleotide sequence at position 146, and a 52 nucleotide sequence at position 422.

A single miscellaneous (MISC) tnpB ORF (locus 14) exists in the genome and is 172 nucleotides in length. This MISC tnpB ORF contains only the most 5’ 102 nucleotides, and the most 3’ 70 nucleotides of type 1A ORFs. Due to the lack of internal sequence or inserts, this element cannot be confidently placed in any other group.

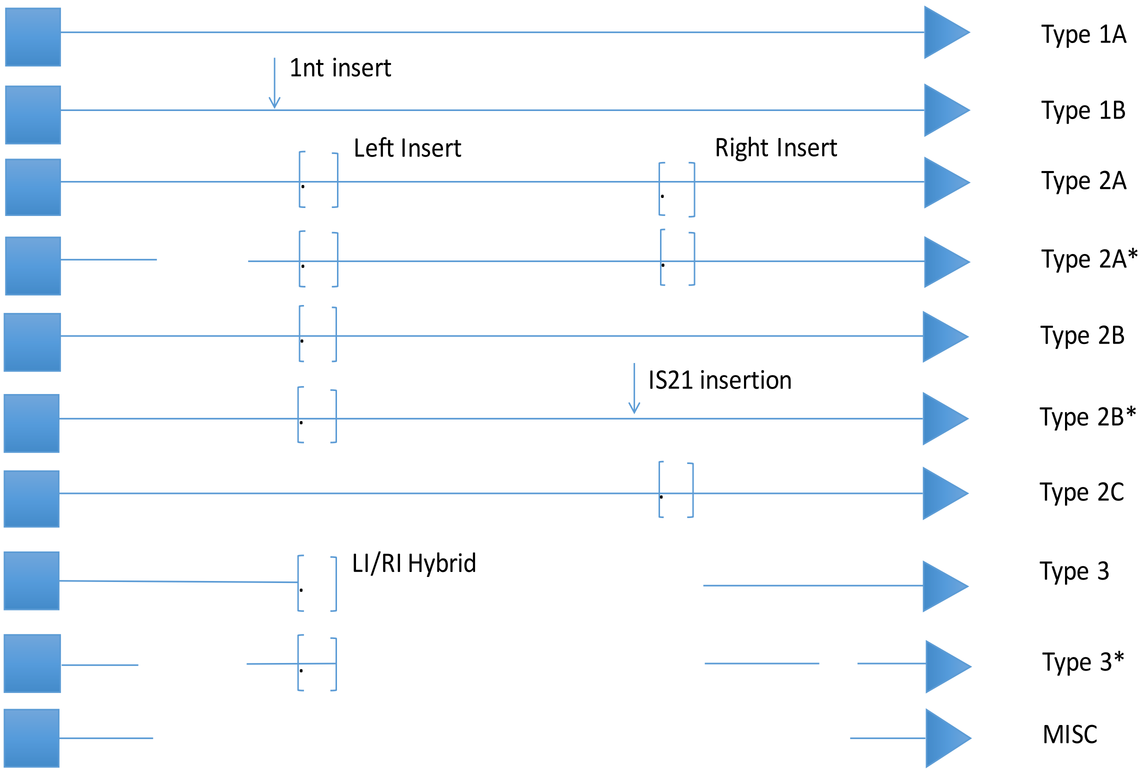


Figure 3.3. Relative IS605 tnpB sequence structures.

3.4. IS605 tnpA/tnpB INTER-ORF SPACE

The nucleotide sequence between the two divergent tnpA and tnpB ORFs is dependent on the IS605 tnpA ORF type present at each locus and varies on the tnpA end of the inter-ORF space. Figure 3.4 shows the nucleotide sequence alignments for the space between the ORFs. Each sequence is labeled with the IS605 tnpA ORF type it is present with. The inter-ORF sequence alignment for all loci is found in Appendix D.

```
Type_3      -----------CTCCATTTTTCCTTTTATAAGCAAACATATGTATGGTATAATTATAGTA      49
Type_4      ---------------ATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      45
Type_2      -----------CTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      49
Type_1      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Type_5      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
                           ****** ****** ***************** *************

Type_3      GAATGGAGGTGAAAAATCA 68
Type_4      GGATGGAGGTGAAAAATCA 64
Type_2      GGATGGAGGTGAAAAATCA 68
Type_1      GGATGGAGGTGAAAAGTCA 79
Type_5      GGATGGAGGTGAAAAGTCA 79
            * ************* ***
```

Figure 3.4. Inter-ORF sequence alignment.

3.5. ORF tnpB INSERTS

The inserts briefly discussed in Section 3.3 can be sorted into 3 groups using their location within the ORF and the most terminal 3 nucleotides on the 5' and 3' ends. The LI and RI are 64 nucleotides in length, while the HI is 67 nucleotides long. The structure of the three inserts are seen in Figure 3.5. All inserts share a common 61 nucleotide central region except where indicated. The LI however lacks a GCT sequence on its 3' end, and the RI insert lacks a TCA sequence on it's 5' end. The hybrid insert contains both the TCA and GCT sequences. This hybrid pattern persists internal to the insert ends between 4 mismatched nucleotides that are a total of 9 nucleotides apart. These inserts disrupt the IS605 tnpB ORF resulting in a non-functional protein. Insert sequence alignment for all inserts is found in Appendix E.

3.6. ORF tnpB PHYLOGENY

IS605 tnpB ORFs were used for the analysis because they contain a larger sequence for alignment. Only ORF tnpB types 1-3 are included in the phylogeny.

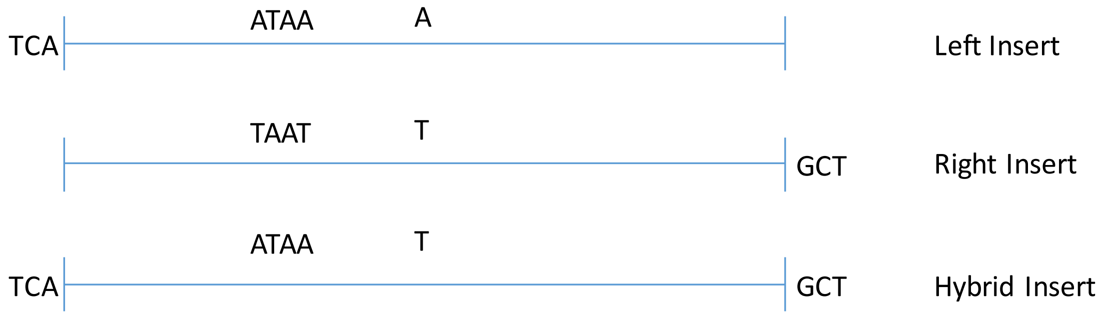


Figure 3.5 Relative insert sequence structure

Because of their more deteriorated state, types 2A*, 3* and MISC were excluded. The phylogenetic tree is located in Figure 3.6 and is labeled with the tnpB type and which locus it appears in (ex. T1A_05; Type 1A_Locus05). In the phylogenetic tree, we see that elements with structural similarities do not form a clade.

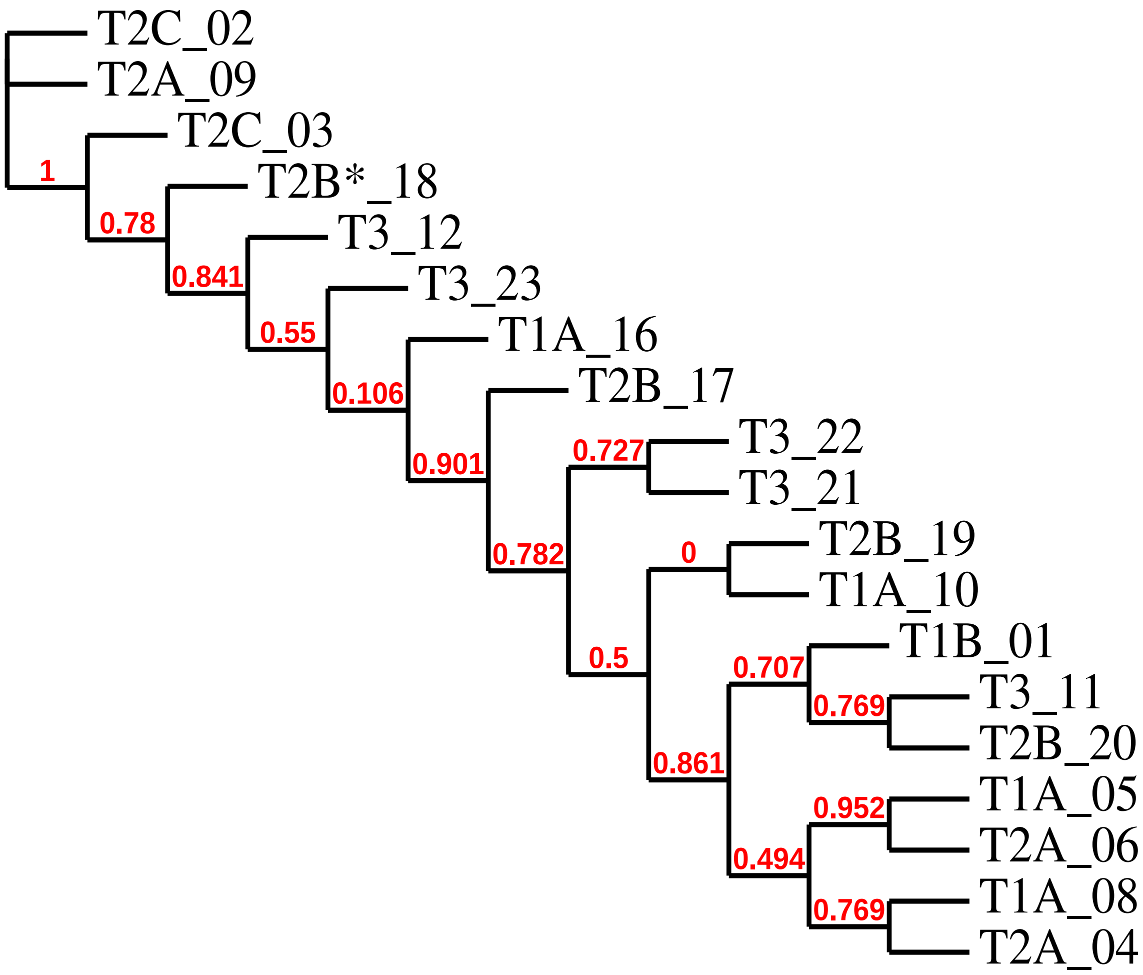


Figure 3.6. ORF tnpB phylogeny. Each tnpB is labeled with the type and locus number it appears in. Branches are labeled with branch support values. Branch length is ignored.

3.7. ELEMENT ENDS

The left end (LE) of the element, as defined as the sequence downstream of the IS605 tnpA ORF, is composed of a 60 nucleotide sequence for all but one of the 22 elements. The LE of locus 13 is missing 11 nucleotides on its tnpA end. The LE of the element begins with a TTTAT sequence (tnpB encoding strand) which is consistent with IS200/605 family members. The left end sequence alignment for all loci is found in Appendix F.

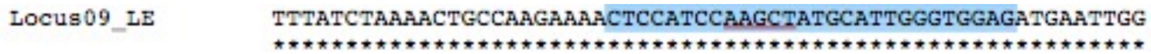
The right end (RE) of the element, as defined as the sequence downstream of the 605 tnpB ORF, can be sorted into two groups and one miscellaneous based on the presence of a 28 nucleotide insert. Type 1, the consensus RE present for all elements unless otherwise stated, extends 132 nucleotides past the 3' end of the tnpB ORF. Type 2 is present at five loci (loci 01, 04, 06, 14, and 21). Type 2 RE contains a 28 nucleotide insert at position 99, and has a total length of 160 nucleotides. This 28 nucleotide insert does not show significant sequence similarity to the IS605 tnpB ORF inserts described in Section 3.4. The miscellaneous RE (locus 15) extends only 23 nucleotides past the 3' end of its respective tnpB ORF. Unlike the locus 18 tnpB disruption where the remainder of the element can be clearly identified beyond the putative IS21 family member, the remainder of the RE for locus 15 cannot be located. The right end sequence alignment for all loci is found in Appendix G.

3.8. HAIRPIN STRUCTURES

Both LE and RE sequences of the IS605 elements contain a hairpin structure required by IS200/605 family members for transposition.

3.8.1. Left End Structure. The LE has only one possible hairpin structure. It is composed of a 10 base pair stem, and 8 nucleotide loop starting 23 nucleotides form the 5' end of element (tnpB encoding strand). The LE sequence alignment highlighting the structure is seen in Figure 3.7. Figure 3.8 shows the structure of the LE hairpin. The LE reverse complement alignments for identification of potential LE structure is found in Appendix H.

3.8.2. Right End Structures. The RE has 3 potential structures. Structures 1, 2, and 3, begin 52, 78, and 91 nucleotides from the 3' end of the tnpB ORF respectively. These structures form an imperfect stem with 8 out of 10, 9 out of 11, and 11 out of 13 base pairs with a 5, 7, and 8 nucleotide loop respectively. The 28 nucleotide insert present in the RE of 5 elements is inserted within structures 2 and 3, but not structure 1. The RE structures 1, 2, and 3 are in Figure 3.9, 3.10, and 3.11 respectively. The RE alignment of sequences highlighting these structures is shown in Figure 3.12. The RE reverse complement alignment for the identification of potential RE structures is found in Appendix I.



```
Locus09_LE          TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG
                    ************************************************************
```

Figure 3.7. Highlighted sequence of left end structure.

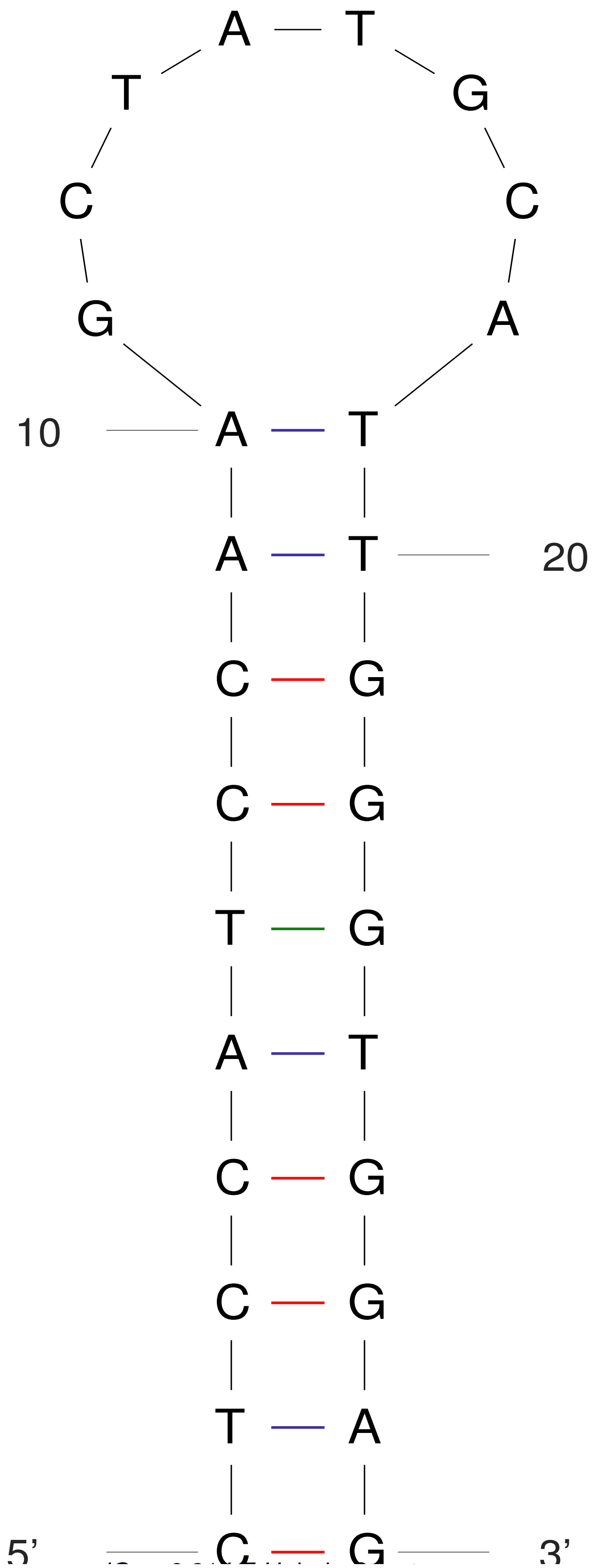


Figure 3.8. Left end hairpin structure.

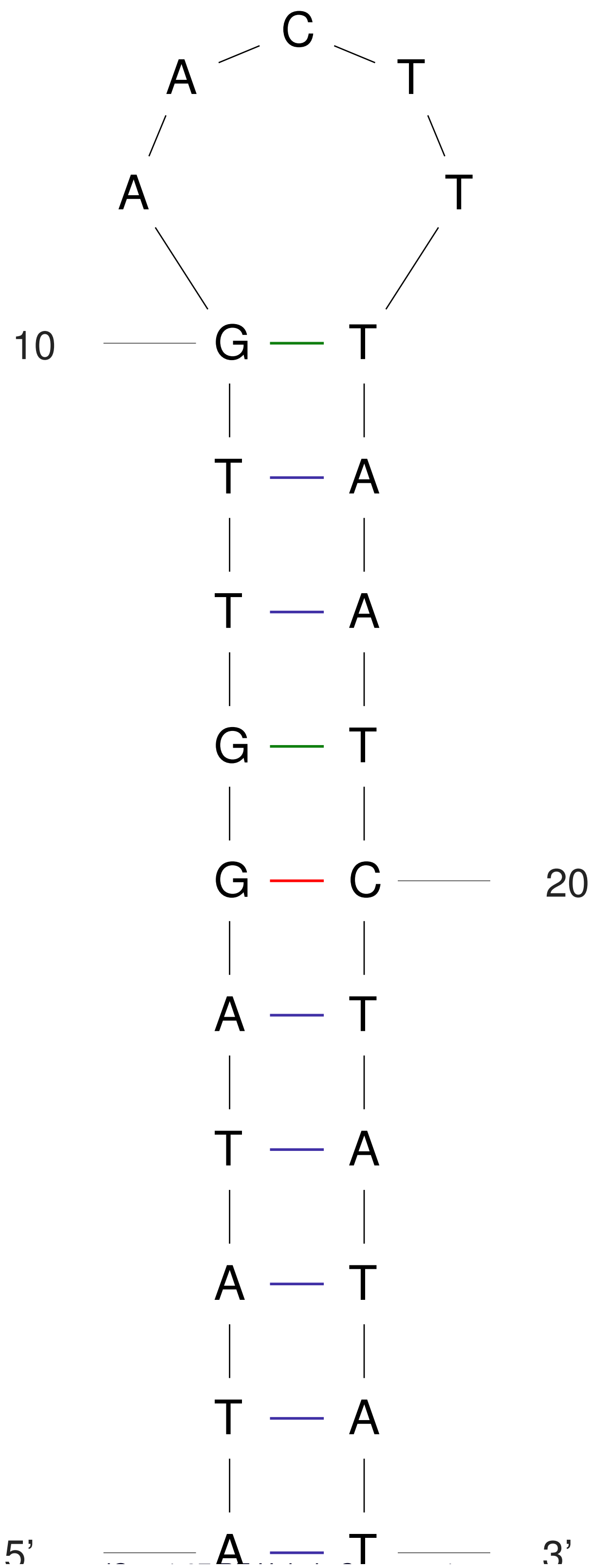


Figure 3.9. Right end hairpin structure 1.

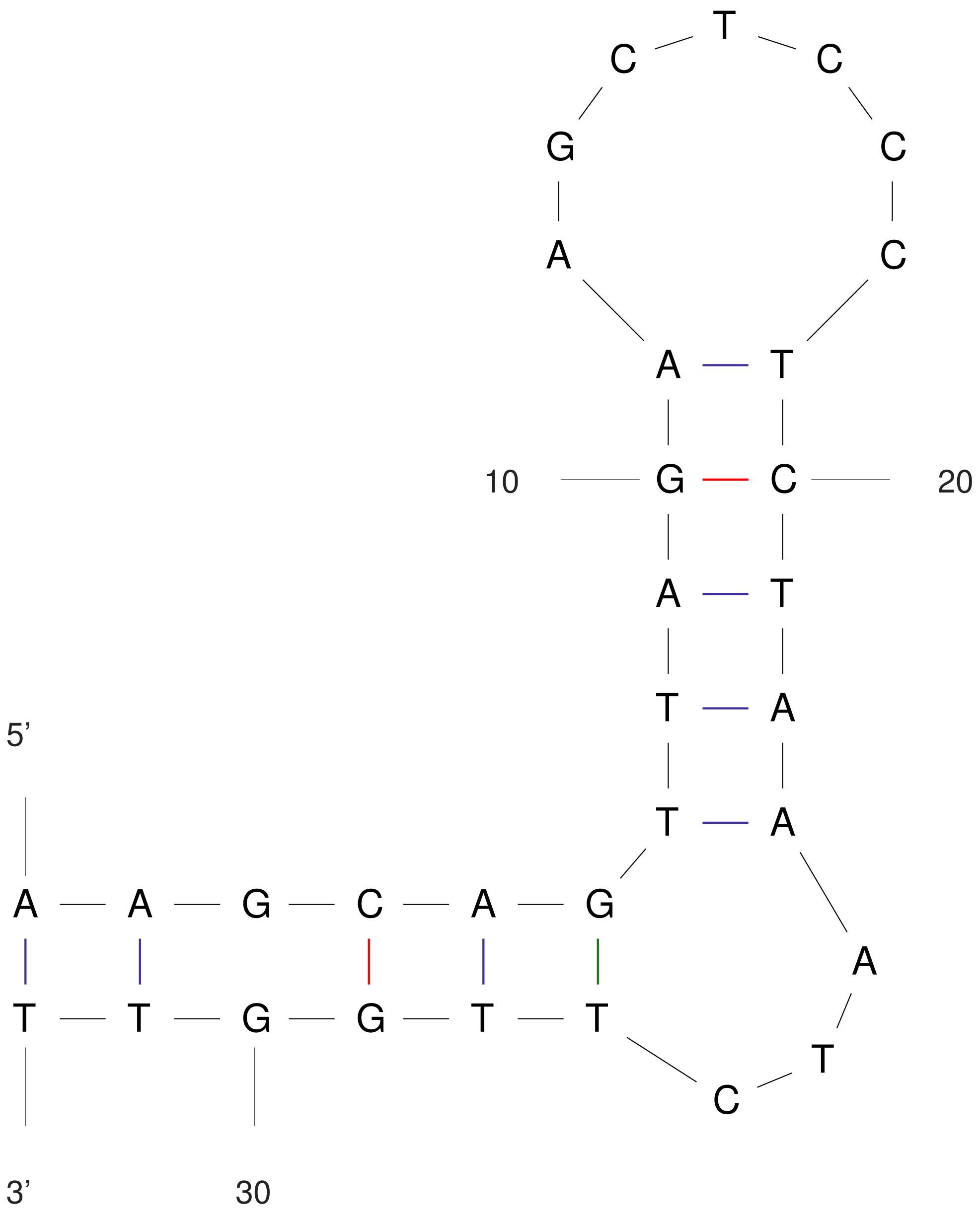


Figure 3.10. Right end hairpin structure 2.

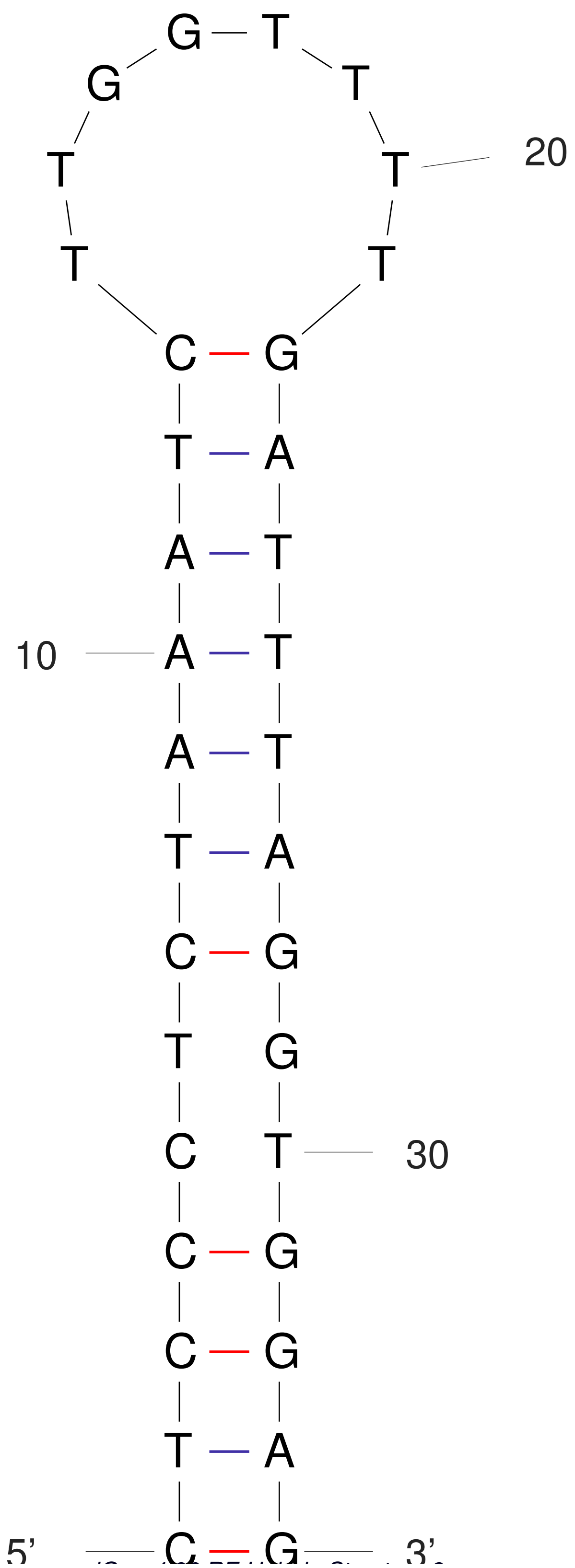
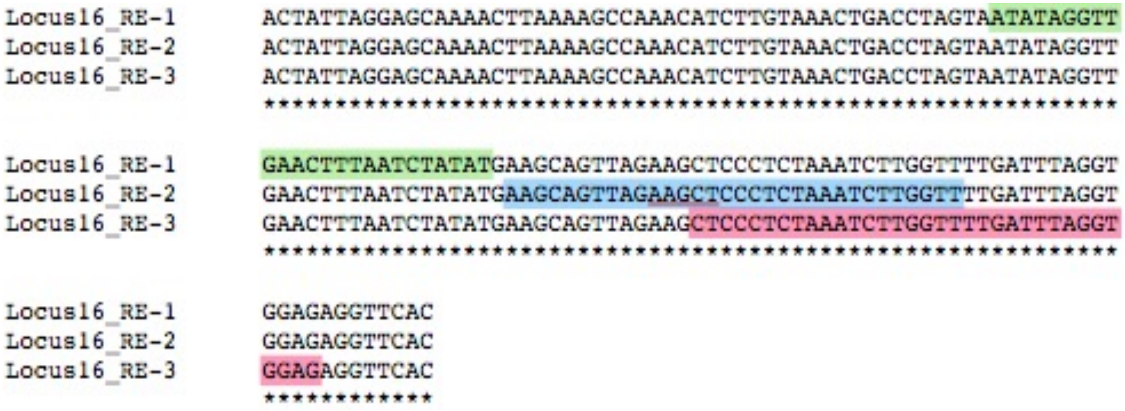


Figure 3.11. Right end hairpin structure 3.



```
Locus16_RE-1      ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT
Locus16_RE-2      ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT
Locus16_RE-3      ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT
                  ************************************************************

Locus16_RE-1      GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGGT
Locus16_RE-2      GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGGT
Locus16_RE-3      GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGGT
                  ************************************************************

Locus16_RE-1      GGAGAGGTTCAC
Locus16_RE-2      GGAGAGGTTCAC
Locus16_RE-3      GGAGAGGTTCAC
                  ************
```

Figure 3.12. Highlighted sequence of right end structures.

3.8.3. ORF tnpB Insert Structure.

The 605 tnpB ORF inserts also contain a secondary hairpin structure. This structure is an imperfect stem with 7 out of 9 bp and a 5 nucleotide loop. The structure is shown in Figure 3.13. The insert sequence highlighting the structure is shown in Figure 3.14. The reverse complement alignments used to identify potential tnpB insert structure is found in Appendix J.

3.9. MINIATURE INVERTED REPEAT TRANSPOSABLE ELEMENTS

One IS605 MITE was identified within the genome. This MITE is approximately 271 nucleotides in length beginning at nucleotide positon 843,418 in the genome and is closely located to locus 09. This MITE contains the last 58 nucleotides of tnpB ORF, no sequence of the 605 tnpA ORF, 28 nucleotides of the LE, and the entire 160 nucleotides of group 2 RE. The 28 nucleotides aligning with the LE contain the predicted secondary hairpin structure of the LE.

The LE, tnpB portion, and RE of the MITE are aligned with representatives in Figures 3.15, 3.16, and 3.17.

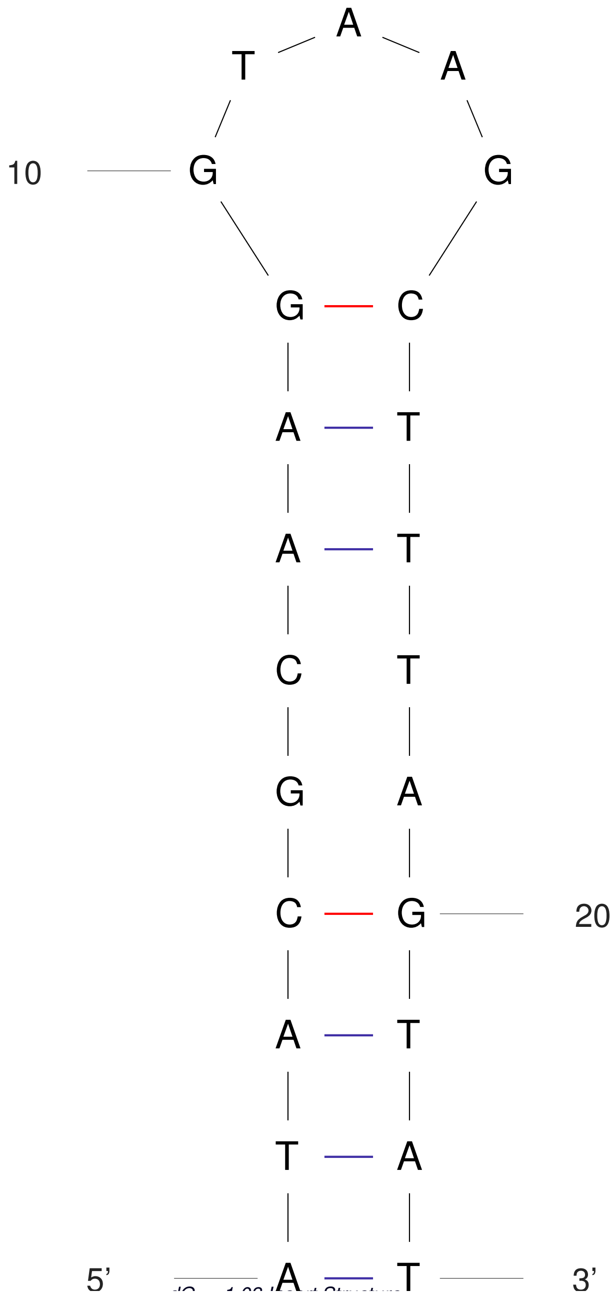
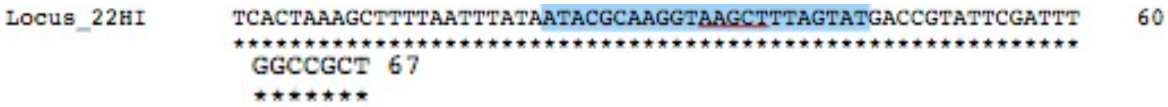


Figure 3.13. ORF tnpB insert hairpin structure.



```
Locus_22HI      TCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    60
                ************************************************************
                 GGCCGCT 67
                 *******
```

Figure 3.14. Highlighted sequence of tnpB ORF insert structure.

```
MITE_LE            1 -----------------------CTCCATCCAAGCTATGCATTGGGTGGA     27
                                            |||||||||||||||||||||||||||
Locus09_LE         1 TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGA     50

MITE_LE           28 G---------     28
                     |
Locus09_LE        51 GATGAATTGG     60
```

Figure 3.15. MITE LE alignment.

```
Locus16_tnpB    1151 ATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGACCTATCAAA   1200
                                                                  |||||
MITE_tnpB          1 ---------------------------------------------TCAAA      5

Locus16_tnpB    1201 GAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATC   1250
                     ||||.||||||||||||||||||||.||||||||||||||||||||||||
MITE_tnpB          6 GAGGTGAGAGATAATGGATTTGTGGCCAATCCTTCAAGATTAAGGGTATC     55

Locus16_tnpB    1251 CTAA   1254
                     ||||
MITE_tnpB         56 CTAA     59
```

Figure 3.16. MITE tnpB alignment.

```
MITE_RE            1 ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGT     50
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
Locus01_RE         1 ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGT     50

MITE_RE           51 AATATAGGTTGAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCCGA    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
Locus01_RE        51 AATATAGGTTGAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCCGA    100

MITE_RE          101 CGCGAAGCTAGTATCTCTTTTGATGCAAATCTTGGTTTTGATTTAGGTGG    150
                     |||||||||||.||||||||||||||||||||||||||||||||||||||
Locus01_RE       101 CGCGAAGCTAGCATCTCTTTTGATGCAAATCTTGGTTTTGATTTAGGTGG    150

MITE_RE          151 AGAGGTTCAC    160
                     ||||||||||
Locus01_RE       151 AGAGGTTCAC    160
```

Figure 3.17. MITE RE alignment.

4. DISCUSSION

4.1. INSERTION SEQUENCE IDENTIFICATION

Y1 elements in *Halanaerobium hydrogeniformans* were chosen for detailed characterization because of their progressed stages of decay, their misidentification by the semi-automatic insertion sequence annotation program, and because Y1 elements do not have a strong preference for cis transposition (which is the preference for a transposase to act on the element that it was transcribed from). Many of the IS605 elements were identified as solo tnpB (IS1341) elements because partial IS605 tnpA sequences were not detected. Additionally, the most closely related tnpB in the ISfinder library (what ISsaga relies on for annotation), was a tnpB of an IS607, a serine transposase. This explains why many of the IS605's encoding only the tnpB were identified as a IS607 and suggests that for insertion sequences deposited in the ISfinder library, this IS605 is the closest relative to the IS607's or that other IS605 tnpB's have also been misidentified.

The misidentification of many of the IS605 elements in *H. hydrogeniformans* by ISsaga 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.

4.2. ORF tnpB PHYLOGENY

A phylogenetic tree between the major tnpB ORFs was constructed to determine the order of transposition events of the IS605 elements. The tnpB ORF was chosen because it is the most consistent sequence between the 22 IS605 elements. ORFs of tnpB types 2A*, 3*, and MISC were not included in the phylogenetic analysis due to their further degraded state. To eliminate the effects of the inserts and the missing inter-insert sequence on the phylogeny, the inserts were manually removed and phylogeny was inferred without G-blocks curation.

It was hypothesized that elements sharing structural similarities (LI/RI/HI) would form a clade on the phylogenetic tree, and that it could be inferred when deletion and insertion events took place. If these elements were replicating without recombination, tnpB's with similar structure (LI/RI/HI) should form a clade. For example, all type 2A's would clade together, all type 2B's would clade together, and all type 2C's would clade together. Figure 3.6 however shows that tnpB ORFs sharing structural similarities do not form a clade. This strongly indicates that recombination between tnpB ORFs is occurring.

This evidence of insertion sequence recombination is contrary to past research. Insertion sequences were screened for evidence of recombination by searching for break points and for pairs of insertion sequence fragments showing more similarity to one another. The research concluded that there was no evidence of recombination or gene conversion [54]. It should be noted however, that the research excluded IS200/605 elements from the survey and was limited complete and annotated insertion sequences. Whereas the results presented here contain elements that were originally annotated as pseudogenes but later manually identified as tnpB disrupted ORFs.

4.3. INVASION – EXPANSION – EXTINCTION CYCLES

Insertion sequences have a high sequence similarity within genomes [52]. There is a lack of evidence for recombination and gene conversion between elements [54]. Insertion sequences have a patchy distribution among genomes [56]. These observations have led to the generally accepted hypothesis that insertion sequences undergo invasion, expansion, and extinction cycles in their free living hosts.

Contrary to the DDE family of insertion sequences that show strong preference for cis action [9], Y1 elements do not. 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. Thus, Y1 elements with degraded ORFs encoding nonfunctional TnpA protein can still replicate, so long as they maintain the secondary structures necessary for transposition and there is at least one functional transposase encoded somewhere in the genome. This lack of cis preference has allowed for the observation of IS605 elements in various stages of degradation.

The present IS605 elements may be in the extinction phase of the insertion sequence cycle. Because they are not immediately selected against, elements with deletions and disruptions can accumulate in the genome and be observed. Other families of insertion sequences may degrade in similar ways within their host genome, and their degraded states have not been observed because, unlike Y1 elements, their disrupted copies are removed from the genome.

Furthermore, recombination between insertion sequences may help explain the rapid extinction of elements in a genome. While ORF disruptions or fatal mutations may accumulate in one element, they can spread throughout replicates in the genome via recombination. This would reduce the number of elements with an intact transposase gene.

4.4. TYPE 5 tnpA

It is worth noting that the most commonly occurring 605 tnpA is type 5 (17 of the 22 elements). The abundance of IS605 elements containing a type 5 tnpA may be a result of increased rates of transposition relative to the other IS605 tnpA types. Presented here are two possibilities for an increased rate of transposition for elements containing a type 5 605 tnpA. Either size reduction increases transposition frequency, or the missing tnpA nucleotide sequence could have a regulatory function as well as encode a TnpA protein.

IS605 exclusively excises from, and preferentially inserts into ssDNA. This preference leads to a bias towards lagging strand template insertion when transposition is coupled with host replication [30]. As element size increases, the probability that both ends of the element exist as ssDNA decreases. Alternatively, as Okazaki fragment size increases, so does the probability that the element ends exist as ssDNA in the lagging strand template.

Thus, as element size decreases there is an increase in genome replication associated transposition events [30]. The 282 nucleotide size reduction of an element with a type 5 tnpA may increase the frequency of transposition by increasing the time spent in a ssDNA state during replication. It would be expected however that the size reduction of type 3 tnpB (530nt) would also increase the rate of transposition. The discrepancy in copy number of these elements, (6 type 3 tnpB vs 17 type 5 tnpA), does not support this. However, the accuracy of the phylogenetic tree is diminished by recombination events between elements, and it cannot be inferred which of the elements existed in the genome first or which has a higher relative replication rate.

A reduced element size increasing transposition frequency is an unlikely reason for the disproportional number of type 5 tnpA. This explanation relies on genome replication associated transposition and a preference for lagging strand template insertion. As seen in Table 3.2, there is no skew for or against insertion into the lagging strand template (10 of 22 tnpAs on leading strand)

The TnpB protein serves as a potential IS605 transposition regulatory protein and has been shown to inhibit IS605 excision and insertion. It is hypothesized that TnpB protein inhibits transposition by binding the terminal DNA hairpin structures or the TnpA protein itself. TnpB mediated transposition inhibition is dependent on the terminal Zn finger domain [35]. However, it has not been established what this domain interacts with. It is possible that the TnpB protein binds ssDNA of the IS605 tnpA ORF sequence, inhibiting TnpA binding or dimerization and preventing transposition. If the region of binding were missing

(Figure 3.2, type 5 tnpA) TnpB could not inhibit transposition and elements without this sequence would have an increased rate of transposition.

Alternatively, the disproportional number of type 5 tnpAs may be a relic of early formation after insertion sequence acquisition, and selective pressure against functional TnpA proteins. Without an accurate phylogenetic tree, it cannot be determined when this type of tnpA formed.

4.5. ORF tnpB INSERT

Left and right inserts (LI & RI) contain a common core 58 nucleotides and are distinguishable by their most 5' and 3' three nucleotides. All LI contain a TCA as the most 5' three nucleotides, while all RI contain a GCT as the most 3' three nucleotides. The hybrid insert (HI) is 67 nucleotides in length and contains both TCA and GCT trinucleotide sequences at the 5' and 3' ends of the insert as seen in Figure 3.5. This pattern indicates that a recombination event has occurred between a LI and a RI to form a hybrid insert.

This same hybrid pattern persists internal to the insert ends. The LI contains an ATAA and a A at nucleotide positions 20 and 33 respectively, while the RI contain a TAAT and T at these positions. The hybrid insert contains the ATAA and T at positions 20 and 33 indicating LI towards the 5' end and RI towards the 3' end. This suggests that the initiating endonuclease for recombination between these inserts has a higher affinity for the sequence between positions 20 and 33 of the insert.

These hybrid inserts (HI) are the product of recombination from a LI and a RI of either the same or different elements, (e.g. a single T2A that contains both a LI and RI, or a T2B and a T2C that contain a LI and a RI respectively). If the recombination event were to take place between a LI and RI of different elements, the results would be one element containing a hybrid insert with LI and RI characteristics at the 5’ and 3’ ends respectively, excluding the ORF regions between inserts (type 3 tnpB, Figure 3.3), and another element containing a HI with the LI and RI characteristics at the 3’ and 5’ ends respectively, with the sequence between the inserts being duplicated.

If the recombination took place between a LI and RI of the same element, only one product capable of transposition could be formed, that is a hybrid insert with LI and RI characteristics at the 5’ and 3’ ends, (type 3 tnpB, Figure 3.3).

No inserts were observed in the genome showing LI or RI characteristics at the 3’ and 5’ ends, nor were inter insert sequence duplications identified. It is hypothesized that all type 3 elements containing a HI are a result of recombination between a LI and a RI of a type 2A tnpB.

The LI and RI show high sequence similarity, indicating that that they originated from the same source. The differentiating three nucleotide sequence at either end suggests an imprecise excision of the insert before insertion into the IS605 tnpB ORF. The 64 nucleotide sequence of the insert, or any part of it, is not found in the genome outside a tnpB ORF.

The independent insertion of all the LI and RI to the same relative location within the tnpB ORF is unlikely. Their reoccurrence in tnpB ORF is thus likely a result of two insertion events and the replication of those elements. As such it is also hypothesized that the presence of these inserts in the tnpB ORF does not impede transposition of the IS605 elements.

4.6. ELEMENT ENDS AND STRUCTURES

Element ends of IS200/605 family members contain hairpin structures indispensable for transposition. In characterized IS200/605 elements, left end (LE) and right end (RE) structures are the same for each element [32], [28].

The LE sequence for all IS605 elements in *H. hydrogeniformans* is highly conserved and stretches 60nt downstream of the tnpA ORF. The LE has the potential to form a single hairpin structure (Figure 3.7) but shows no sequence homology to the RE.

The RE of the IS605 elements is 132 nucleotides in length and has the potential to form 3 different hairpin structures (Figure 3.9, 3.10, and 3.11). Highlighted sequences of the RE structures (Figure 3.12) show that structures 2 and 3 have significant overlap, making them mutually exclusive. Structures 1 and 3 are separated by 14 nucleotides, so it may be possible to form both structures simultaneously. The base pairing in the stems of structures 1 and 2 however, are separated by a single nucleotide. It is not clear if structures 1 and 2 are exclusive or competitive, as a single nucleotide space may permit both structures to co-exist.

There has previously been speculation that the terminal hairpins structures of IS200/605 elements serve as a transcriptional terminator as well as prevents ribosome binding. It has since been established that they play a mechanistic role in transposition [24]. Potentially competing and mutually exclusive structures may further serve a regulatory role by preventing the mechanistic hairpin structure from being bound by a TnpA monomer.

Competitive structures have been reported before, although in these instances it was clear which structures were mechanistic as only a single common structure was observed between the LE and the RE [24].

Of characterized IS200/605 elements, it is unknown whether the TnpA binds the terminal hairpins through structure recognition or DNA sequence recognition in the stem or loop of the structures [23]. This is the first known report to describe a characterized IS605 element that does not contain the same secondary structure at both the LE and RE. This difference in LE and RE structure, while maintaining transposable capability of the element, suggests that the hairpin is recognized from structure alone. However, there could be a short conserved sequence in both LE and RE structures recognized by TnpA.

Underlined in the highlighted LE structure sequence (Figure 3.8) and the highlighted RE structure 2 sequence (Figure 3.12) is a common AAGCT. This pentanucletoide sequence is presented in the hairpin loop in both structures. The sequence and location in the hairpin is the strongest similarity between any of the potential structures. This suggests that RE structure 2 is mechanistic, implying that RE structure 1 and 3 are potentially regulatory, and that TnpA recognizes a pentanucleotide sequence AAGCT in the loop of the hairpin structure.

At five loci, a 28 nucleotide long sequence has inserted into structure 2 and 3. This insert occurs immediately after nucleotide 21 of structure 2 and nucleotide 8 of structure 3 disrupting both structures. Because elements containing this RE insert have replicated (loci 01, 04, 06, 14, and 21), it is not completely preventing replication.

This supports the notion that RE structure 1 is the mechanistic structure. However, as the insert occurs toward the end of the structure, the AAGCT pentanucletoide sequence in the loop of structure 2 could still be presented. It is possible that the insert only reduces the affinity of TnpA for the RE structure 2.

Elements surveyed by Ronning [28] were shown to excise in a strand specific manner dependent on a secondary loop containing a T in the stem of the structure. RE structure 2 (Figure 3.10) contains a 3 nucleotide secondary loop containing a T. This secondary loop however is not present in the LE structure of elements described here.

Of note is the secondary structure of the tnpB ORF insert (Figure 3.13). In the reported strand, this hairpin contains an AAGCT pentanucleotide sequence in the loop. Additionally, an AAGGT sequence can be found in a more similar position compared to the pentanucleotide sequence in the LE structure and the RE structure 2. The complement strand hairpin also contains an AAGCT sequence in a similar position compared to the two structures. The implications of this observation are unknown at this time.

4.7. MITES

Bacterial MITES are typically difficult to identify as they are short elements and the parental transposable elements are often no longer present in the genome. A single IS605-related MITE was located in *H. hydrogeniformans*. This element is 247 nucleotides in length. It contains 28 nucleotides of the LE (Figure 3.15), the last 59 nucleotides of the IS605 tnpB ORF (Figure 3.16), and the entire 160 nucleotides of a RE containing the 28 nucleotide insert (Figure3.17). These sequences occur in succession, without gaps. The 28 nucleotides of the LE contain the entire LE hairpin structure. The LE hairpin structure and an intact RE make it likely that this element is transposable.

4.8. CONCLUSIONS

Although only a single element contains an intact IS605 tnpA, all IS605 elements reported here contain intact hairpin structures and are likely capable of transposition by a TnpA acting in trans. Dissimilar RE and LE structure sequences suggest that hairpin recognition may be independent of hairpin sequence, although a conserved pentanucleotide sequence present in the hairpin loop is suggestive of a sequence specific recognition. Unique to our findings, the inserts in the tnpB ORF provide structural differences that can be used to infer recombination between insertion sequences. Because these Y1 elements do not rely on the integrity of their ORF for transposition, their detailed survey in a single genome provides a snapshot of how insertion sequences degrade during invasion-expansion-extinction cycles.

4.9. FUTURE DIRECTIONS

The results presented here explore interesting insertion sequence activity within *Halanaerobium hydrogeniformans.* However, they only provide a snapshot of activity. While there is evidence indicating element recombination, direct evidence for insertion sequence recombination is absent. Similarly, it is hypothesized that all the IS605 elements discussed are transposable due to their intact secondary structures. However, direct evidence of transposition is still needed. The LE and RE of the element do not share a common secondary structure sequence. It is unknown what commonality between the structures is essential for TnpA recognition. Future directions for research would address these issues.

Amplification and sequencing of the IS605 insertion sequences from a new sample of *H. hydrogeniformans* from the environment and comparison of the tnpB ORFs and flanking sequences could elucidate transposition and recombination hypotheses. TnpA and TnpB binding assays with elements containing mutated hairpin structures could help determine the functional sequences of the hairpins and the mechanism by which TnpB inhibits transposition

Immediate future research would include the comparison of the IS605 elements in *H. hydrogeniformans* to those in *Halanaerobium saccharolyticum*. *H. saccharolyticum* is the closest relative to *H. hydrogeniformans* and the sequence of its genome is currently being determined. Partial genome sequences from the project are available. Initial observations from the partial sequences show that *H.*

saccharolyticum contains a highly similar TnpA and TnpB to those described here (approximately 90% nucleotide similarity) in its genome in fragmented and partial copies. Interestingly, *Halanaerobium praevalens*, which is closely related to *H. hydrogeniformans* and *H. saccharolyticum*, does not contain any of the IS605 insertion sequences described in this work. Comparison of the elements and syntenic regions between *H. hydrogeniformans*, *H. saccharolyticum,* and *H. praevalens* would help us understand the origins of these elements, their continued activity in genomes, and the manner in which they decompose.

APPENDIX A.

IS200 ELEMENT ENDS AND REVERSE COMPLEMENTS

```
#=======================================
#
# Aligned_sequences: 2
# 1: IS200_RE
# 2: IS200_LE
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 76
# Identity:      37/76 (48.7%)
# Similarity:    37/76 (48.7%)
# Gaps:          31/76 (40.8%)
# Score: 80.5
#
#
#=======================================

IS200_RE          22 TATTAGTTTACAACTAAGTT--ATA------TCATAATATAAATATACTA     63
                     |||||.|||         ||  .||      ||        |||.|.|.|
IS200_LE           1 TATTATTTT---------TTGCTTACCCGGGTC--------AATTTTCAA     33

IS200_RE          64 TAATATAACAATAGAGTCAATAATTT     89
                     |    |||.||.||.||||  |||||
IS200_LE          34 T----TAATAAAAGGGTCA--AATTT     53
```

```
>>IS200_LE_RC                                      (53 nt)
 Waterman-Eggert score: 46;  68.0 bits; E(1) <  9.7e-18
55.3% identity (55.3% similar) in 47 nt overlap (53-7:2-48)

         50        40        30        20        10
IS200_ AAATTTGACCCTTTTATTAATTGAAAATTGACCCGGGTAAGCAAAAA
       :: ::   :: :::  : : :  ::: :   :: ::  :   :::::
IS200_ AATTTGACCCTTTTATTAATTGAAAATTGACCCGGGTAAGCAAAAAA
              10        20        30        40

>--
 Waterman-Eggert score: 46;  12.4 bits; E(1) <  0.41
90.9% identity (90.9% similar) in 11 nt overlap (53-43:24-34)

         50
IS200_ AAATTTGACCC
       ::: :::::::
IS200_ AAAATTGACCC
            30
```

```
>--
 Waterman-Eggert score: 31;  8.6 bits; E(1) <  1
87.5% identity (87.5% similar) in 8 nt overlap (11-4:46-53)

       10
IS200_ AAAAAATA
       ::: ::::
IS200_ AAATAATA
          50

>--
 Waterman-Eggert score: 30;  8.3 bits; E(1) <  1
100.0% identity (100.0% similar) in 6 nt overlap (35-30:26-31)


IS200_ AATTGA
       ::::::
IS200_ AATTGA
          30

>>IS200_LE_RC                                             (53 nt)
 Waterman-Eggert score: 50;  13.4 bits; E(1) <  0.23
65.0% identity (65.0% similar) in 40 nt overlap (1-39:15-53)

               10         20        30
IS200_ TATTATTTTTTGCTT-ACCCGGGTCAATTTTCAATTAATA
       ::::: ::     :: :::::::: ::     :: :::::
IS200_ TATTAATTGAAAATTGACCCGGGT-AAGCAAAAAATAATA
           20        30         40        50


>>IS200_RE_RC                                             (101 nt)
 Waterman-Eggert score: 78;  73.9 bits; E(1) <  5.6e-19
63.3% identity (63.3% similar) in 60 nt overlap (98-40:23-78)

               90        80        70        60        50
IS200_ CTTTATT-TGAAATTATTGACTCTATTGTTATATTATAGTATATTTATATTATGATATAA
       :: :::: : : ::::: :    :::  :::::::::  ::::  :   :: : :  :::
IS200_ CTCTATTGTTATATTATAG----TATATTTATATTATGATATAACTTAGTTGTAAACTAA
             30        40            50        60        70

>--
 Waterman-Eggert score: 66;  11.2 bits; E(1) <  0.99
66.7% identity (66.7% similar) in 63 nt overlap (92-34:19-79)

        90        80          70         60         50        40
IS200_ TTGAAATTATTGACTCTATTGT--TATATT-ATAGTAT-ATTTATATTATGATATAACTT
       ::::   :::::  : :::: :  :::::: ::: ::: :: ::  ::: : : :::  :
IS200_ TTGACTCTATTGT-TATATTATAGTATATTTATATTATGATATAACTTA-GTTGTAAACT
```

```
>>IS200_RE_RC                                             (101 nt)
 Waterman-Eggert score: 116;  18.3 bits; E(1) <  0.03
69.0% identity (69.0% similar) in 58 nt overlap (15-72:7-63)

           20        30        40        50        60        70
IS200_ TATAAAATATTAGTTTACAACTAAGTTATATCATAATATAAATATACTATAATATAAC
       :::   : :::: :: ::      ::::::: ::: ::::  :::: ::: :::::::
IS200_ TATTTGAAATTA-TTGACTCTATTGTTATATTATAGTATATTTATATTATGATATAAC
         10         20        30        40        50        60

>--
 Waterman-Eggert score: 92;  14.9 bits; E(1) <  0.28
62.1% identity (62.1% similar) in 58 nt overlap (1-58:44-101)

               10        20        30        40        50
IS200_ TATATATCCTCTCCTATAAAATATTAGTTTACAACTAAGTTATATCATAATATAAATA
       ::: :::  : :  :::::  :: : ::  :: : ::  :::::  : :  ::: :::
IS200_ TATTTATATTATGATATAACTTAGTTGTAAACTAATATTTTATAGGAGAGGATATATA
            50        60        70        80        90       100

>--
 Waterman-Eggert score: 90;  14.6 bits; E(1) <  0.33
58.1% identity (58.1% similar) in 93 nt overlap (8-99:3-94)

        10        20         30        40        50        60
IS200_ CCTCTCCTATAAAATATT-AGTTTACAACTAAGTTATATCATAATATAAATATACTATAA
       ::: :  :  ::: :::: : : ::    ::  :::::  ::: :   : :::  :::::
IS200_ CCTTTATTTGAAATTATTGACTCTATTGTTATATTATAGTATATTTATATTATGATATAA
             10        20        30        40        50        60
```

APPENDIX B.

IS605 tnpA SEQUENCE ALIGNMENTS

```
CLUSTAL O(1.2.1) multiple sequence alignment


Locus13      ------------------------------------------------------------      0
Locus01      ------------------------------------------------------------      0
Locus20      ------------------------------------------------------------      0
Locus15      ------------------------------------------------------------      0
Locus09      ATGGATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATTTAGTT      60

Locus13      ------------------------------------------------------------      0
Locus01      ------------------------------------------------------------      0
Locus20      ------------------------------------------------------------      0
Locus15      ------------------------------------------------------------      0
Locus09      GTAATTACAAAATACAGACATGAATGTATTACTTTTGAAATGCTTGAAGAATTAGAAAAA      120

Locus13      ---------------------------------------------------GGAGAAAAA      9
Locus01      ---------------------------------------------------GGAGAAAAA      9
Locus20      ---------------------------------------------------GGAGAAAAA      9
Locus15      ------------------------------------------------------------      0
Locus09      ATATTCACCAGATTACTCAAGGACAAAGTTTGTAATGTTCTAGAGTTTGGAGGAGAAAAA      180

Locus13      GATCATGTGCATATCCTCTTTGAAAATCCACCTCAGGTACAATTATCTAAGTTAGTTAAT      69
Locus01      GATCATGTGCATATCCTCTTTGAAACTCCATCTCAGGTACAATTATCTAAGTTAGTTAAT      69
Locus20      GATCATGTGCATATCCTCTTTGAAACTCCACCTCAGGTACAATTATCTAAGTTAGTTAAT      69
Locus15      ------------------------------------------------------------      0
Locus09      GATCATGTGCATATCCTCTTTGAAACTCCACCTCAGGTACAATTATCTAAGTTAGTTAAT      240

Locus13      ATATTAAAAACTGTATCTTCAAGACTTATCAAAAAGCAATATGAACACCAT---------      120
Locus01      ATATTAAAAACTGTATCTTCAAGACTTATCAAAAAGCAATATGAACACCATCTGAAAAAA      129
Locus20      ATATTAAAAACAGTATCTTCAAGACTTATCAAAAAGCAATATGAACACCATCTGAAAAAA      129
Locus15      ---------------------------------------------------------AAA      3
Locus09      ATATTAAAAACAGTATCTTCAAGACTTATCAAAAAGCAATATGAACACCATCTGAAAAAA      300

Locus13      ------------------------------------------------------------      120
Locus01      TATTATTGGAAACCTGCTTTTTGGTCTAGAAGCTACTGCATTTTGTCTACTGGTGGTGCT      189
Locus20      TATTATTGGAAACCTGCTTTTTGGTCTAGAAGCTACTGCATTTTGTCTACTGGTGGTGCT      189
Locus15      TATTATTGGAAACCTGCTTTTTGGTCTAGAAGCTACTGCATTTTGTCTACTGGTGGTGCT      63
Locus09      TATTATTGGAAACCTGCTTTTTGGTCTAGAAGCTACTGCATTTTGTCTACTGGTGGTGCT      360

Locus13      ---------------------------------------------      120
Locus01      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG      234
Locus20      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG      234
Locus15      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG      108
Locus09      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG      405
```

```
CLUSTAL O(1.2.1) multiple sequence alignment


Locus_09      ATGGATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATTTAGTT      60
Locus_06      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATTTAGTT      60
Locus_04      ATGGATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_03      ATGGATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_19      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_10      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_12      ATGAGTAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_02      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_11      ATGAATAGAGACTTAAATAACAATTATTATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_18      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_22      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_17      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_16      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_08      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_05      ATGAATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_23      ATGGATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_21      ATGGATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
Locus_14      ATGGATAGAGACTTAAATAACAATTATCATTCTGTTTATAGTCTACAATATCATT--ATT      58
              ***  ********************** ***************************   **

Locus_09      GTAATTACAAAATACAGACATGAATGTATTACTTTTGAAATGCTTGAAGAATTAGAAAAA      120
Locus_06      ------------------------------------------------------------      60
Locus_04      ------------------------------------------------------------      58
Locus_03      ------------------------------------------------------------      58
Locus_19      ------------------------------------------------------------      58
Locus_10      ------------------------------------------------------------      58
Locus_12      ------------------------------------------------------------      58
Locus_02      ------------------------------------------------------------      58
Locus_11      ------------------------------------------------------------      58
Locus_18      ------------------------------------------------------------      58
Locus_22      ------------------------------------------------------------      58
Locus_17      ------------------------------------------------------------      58
Locus_16      ------------------------------------------------------------      58
Locus_08      ------------------------------------------------------------      58
Locus_05      ------------------------------------------------------------      58
Locus_23      ------------------------------------------------------------      58
Locus_21      ------------------------------------------------------------      58
Locus_14      ------------------------------------------------------------      58


Locus_09      ATATTCACCAGATTACTCAAGGACAAAGTTTGTAATGTTCTAGAGTTTGGAGGAGAAAAA      180
Locus_06      ------------------------------------------------------------      60
Locus_04      ------------------------------------------------------------      58
Locus_03      ------------------------------------------------------------      58
Locus_19      ------------------------------------------------------------      58
Locus_10      ------------------------------------------------------------      58
Locus_12      ------------------------------------------------------------      58
Locus_02      ------------------------------------------------------------      58
Locus_11      ------------------------------------------------------------      58
Locus_18      ------------------------------------------------------------      58
Locus_22      ------------------------------------------------------------      58
Locus_17      ------------------------------------------------------------      58
Locus_16      ------------------------------------------------------------      58
```

```
Locus_08        ------------------------------------------------------------      58
Locus_05        ------------------------------------------------------------      58
Locus_23        ------------------------------------------------------------      58
Locus_21        ------------------------------------------------------------      58
Locus_14        ------------------------------------------------------------      58

Locus_09        GATCATGTGCATATCCTCTTTGAAACTCCACCTCAGGTACAATTATCTAAGTTAGTTAAT      240
Locus_06        ------------------------------------------------------------      60
Locus_04        ------------------------------------------------------------      58
Locus_03        ------------------------------------------------------------      58
Locus_19        ------------------------------------------------------------      58
Locus_10        ------------------------------------------------------------      58
Locus_12        ------------------------------------------------------------      58
Locus_02        ------------------------------------------------------------      58
Locus_11        ------------------------------------------------------------      58
Locus_18        ------------------------------------------------------------      58
Locus_22        ------------------------------------------------------------      58
Locus_17        ------------------------------------------------------------      58
Locus_16        ------------------------------------------------------------      58
Locus_08        ------------------------------------------------------------      58
Locus_05        ------------------------------------------------------------      58
Locus_23        ------------------------------------------------------------      58
Locus_21        ------------------------------------------------------------      58
Locus_14        ------------------------------------------------------------      58

Locus_09        ATATTAAAAACAGTATCTTCAAGACTTATCAAAAAGCAATATGAACACCATCTGAAAAAA      300
Locus_06        ------------------------------------------------------------      60
Locus_04        ------------------------------------------------------------      58
Locus_03        ------------------------------------------------------------      58
Locus_19        ------------------------------------------------------------      58
Locus_10        ------------------------------------------------------------      58
Locus_12        ------------------------------------------------------------      58
Locus_02        ------------------------------------------------------------      58
Locus_11        ------------------------------------------------------------      58
Locus_18        ------------------------------------------------------------      58
Locus_22        ------------------------------------------------------------      58
Locus_17        ------------------------------------------------------------      58
Locus_16        ------------------------------------------------------------      58
Locus_08        ------------------------------------------------------------      58
Locus_05        ------------------------------------------------------------      58
Locus_23        ------------------------------------------------------------      58
Locus_21        ------------------------------------------------------------      58
Locus_14        ------------------------------------------------------------      58

Locus_09        TATTATTGGAAACCTGCTTTTTGGTCTAGAAGCTACTGCATTTTGTCTACTGGTGGTGCT      360
Locus_06        ------------------------------------------GTATCTTCTGGTGGTGCT      78
Locus_04        ------------------------------------------GTATCTTCTGGTGATGCT      76
Locus_03        ------------------------------------------GTATCTTCTGGTGATGCT      76
Locus_19        ------------------------------------------GTATCTTCTGGTGATGCT      76
Locus_10        ------------------------------------------GTATCTTCTGGTGATGCT      76
Locus_12        ------------------------------------------GTATCTTCTGGTGATGCT      76
Locus_02        ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_11        ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_18        ------------------------------------------GTATCTTCTGGTGGTGCT      76
```

```
Locus_22      ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_17      ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_16      ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_08      ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_05      ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_23      ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_21      ------------------------------------------GTATCTTCTGGTGGTGCT      76
Locus_14      ------------------------------------------GTATCTTCTGGTGGTGCT      76
                                                         * *** ****** ****

Locus_09      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     405
Locus_06      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     123
Locus_04      GCTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_03      GCTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_19      GCTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_10      GCTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_12      GCTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_02      ACTATTGAGACAATTAAAAAGTATATTGAAAATAAGAATAAATAG     121
Locus_11      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_18      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_22      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_17      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_16      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_08      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_05      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_23      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_21      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
Locus_14      ACTATTGAGACAATTAAAAAGTATATTGAAAATCAGAATAAATAG     121
               ******************************** ***********
```

APPENDIX C

IS605 tnpB SEQUENCE ALIGNMENTS

Type 1A/B

```
CLUSTAL O(1.2.1) multiple sequence alignment


T1B_01          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT          60
T1A_16          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT          60
T1A_10          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT          60
T1A_08          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT          60
T1A_05          ATGCGATTATCATTTAAATTCAACCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT          60
                *********************** ************************************

T1B_01          GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT          120
T1A_16          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT          120
T1A_10          GAATTAGCCTGGCATATTAGTAAACTATATAATACAGTCAATTATGAGGTTAAAAACAAT          120
T1A_08          GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT          120
T1A_05          GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT          120
                ***************     **** ********* ********** ** ******* ***

T1B_01          AAAGATGTAAAAGTTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT          180
T1A_16          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT          180
T1A_10          AAAGATATAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT          180
T1A_08          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT          180
T1A_05          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT          180
                ****** ****** **********************************************

T1B_01          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA          240
T1A_16          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGTTCAGGACTGGAAA          240
T1A_10          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA          240
T1A_08          GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA          240
T1A_05          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA          240
                ************************** ******************* *************

T1B_01          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA          300
T1A_16          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA          300
T1A_10          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA          300
T1A_08          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA          300
T1A_05          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA          300
                ***************************************************** ******

T1B_01          GGACCACCTAATTTTAAACATATGAACAGCAATCCCTGTGAAATAATTTTTACCAATTTA          360
T1A_16          GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA          360
T1A_10          GGATCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA          360
T1A_08          GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA          360
T1A_05          GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA          360
                **  ************************* ******************************

T1B_01          GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T1A_16          GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA          420
T1A_10          GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T1A_08          GCTGTTAGGATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T1A_05          GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
                ******** **** *************************************** ******

T1B_01          TATAATGTGAAAAGCTCTTAATTTGAGCTGCCTGAAAGCAGTTCAAAGCATTATAGATTT          480
T1A_16          TATAATGTGAAGGCTCTTAATTTTGAGCTGC-CTGAAGCAGTTCAAAGCATTATAGATTT          479
T1A_10          TATAATGTGAAAGCTCTTAATTTTGAGCTGC-CTGAAGCAGTTCAAAGCATTATAGATTT          479
T1A_08          TATAATGTGAAGGCTCTTAATTTTGAGCTGC-CTGAAGCAGTTCAAAGCATTATAGATTT          479
```

```
T1A_05      TATAATGTGAAAAGCCTTAATTTTGAGCTGC-CTGAAGCAGTTCAAAGCATTATAGATTT      479
            ***********      * * **********    *************************

T1B_01      AGATGCTGTCCAGCAGATAAAGATAAAGCAGGACCGTATTTCTAAAAGATGGTATCTACT      540
T1A_16      AGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCTCTT      539
T1A_10      AGATGCTGGCCAGCAGATAAAGATTAAGCAGGACCGCATTTCTAAAAGATGGTATCTACT      539
T1A_08      AGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCTACT      539
T1A_05      AGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCTACT      539
            ******** *************** *********** ********************  *

T1B_01      AATCATCTATAAAACCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAGTTGA      600
T1A_16      AATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGTAGA      599
T1A_10      AATCATCTATAAGGTCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAATAGA      599
T1A_08      AATCATCTATAAGACCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAGTTGA      599
T1A_05      AATCATCTATAAGACCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAGTAGA      599
            *** ***** **     * **   ******** *** **  ************** * **

T1B_01      TTTAGGCCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATATTAT      660
T1A_16      TCTAGGTCTTGATAATTTGGCTACTTTAATATTTAAAAACAATTCTGATTGTTATATTAT      659
T1A_10      TCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAACTCTGAGTGTTATATTAT      659
T1A_08      TCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATATTAT      659
T1A_05      TCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGAGTGTTATATTAT      659
            * **** ********************** ************ ***** ***********

T1B_01      CAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACTACA      720
T1A_16      CAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACTACA      719
T1A_10      CAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACTACA      719
T1A_08      CAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACTACA      719
T1A_05      CAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACTACA      719
            ************************************************************

T1B_01      AAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAAATA      780
T1A_16      AAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAAATA      779
T1A_10      AAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAAATA      779
T1A_08      AAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAAATA      779
T1A_05      AAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAAATA      779
            ************************************************************

T1B_01      TCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAAAAT      840
T1A_16      TCTGAGATTAAAGAGAAAAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAAAAT      839
T1A_10      TCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGCTGCAAAAT      839
T1A_08      TCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATATCCATAAAGCTAGCTGCAAAAT      839
T1A_05      TCTGAGATTAAAGAGAAGAAATTATATTAGCAATTATCTCCATAAAGCTAGTTGCAAAAT      839
            ***************** ************  ***** ************* ********

T1B_01      AGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAATAT      900
T1A_16      AGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAATAT      899
T1A_10      AGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAATAT      899
T1A_08      AGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAATAT      899
T1A_05      AGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATCGGAGATATAAAAAATAT      899
            ****************************************** *****************

T1B_01      TAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCAATCCAGAGATTAAAAAAATT      960
T1A_16      TAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAAATT      959
T1A_10      TAAAAAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAAATT      959
T1A_08      TAAACAATGCAGCAATCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAAATT      959
T1A_05      TAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAAATT      959
```

```
            **** ********** *********************** ********************

T1B_01      AATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTATAC        1020
T1A_16      AATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGGAATTGATGAAAGCTATAC        1019
T1A_10      AATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTATAC        1019
T1A_08      AATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTATAC        1019
T1A_05      AATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTATAC        1019
            ***************************************** ******************

T1B_01      TTCTGGGTGTAGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATAAATCCAG        1080
T1A_16      TTCTGGATGTAGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATAAATCCAG        1079
T1A_10      TTCTGGGTGTAGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATAAATCCAG        1079
T1A_08      TTCTGGATGTAGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATAAATCCAG        1079
T1A_05      TTCTGGATGTAGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATAAATCCAG        1079
            ****** *****************************************************

T1B_01      AAGAATTGCTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTGATCAGAA        1140
T1A_16      AAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTGATCAGAA        1139
T1A_10      AAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGTCTATTAATTAATGCTGATCAGAA        1139
T1A_08      AAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTGATCAGAA        1139
T1A_05      AAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTGATCAGAA        1139
            ******* **************************** ***********************

T1B_01      TGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGACCTATCAA        1200
T1A_16      TGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGACCTATCAA        1199
T1A_10      TGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGACCTATCAA        1199
T1A_08      TGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGACCTATCAA        1199
T1A_05      TGGTAGCTTTAATATACTTCGTAAATATCATAAGGATAAATGTATTCTCAGACCCATCAA        1199
            ****** ******************** ***** ******************** *****

T1B_01      AGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCTAA     1255
T1A_16      AGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCTAA     1254
T1A_10      AGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCTAA     1254
T1A_08      AGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCTAA     1254
T1A_05      AGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCTAA     1254
            *******************************************************
```

Type 2A

```
CLUSTAL O(1.2.1) multiple sequence alignment


T1A_10      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T1A_05      ATGCGATTATCATTTAAATTCAACCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T1A_16      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T1A_08      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2A_09      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2A_06      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2A_04      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
            *********************** ************************************

T1A_10      GAATTAGCCTGGCATATTAGTAAACTATATAATACAGTCAATTATGAGGTTAAAAACAAT        120
T1A_05      GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT        120
T1A_16      GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T1A_08      GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT        120
T2A_09      GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2A_06      GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2A_04      GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT        120
            ***************:  : **** ********* ********** **.******* ***

T1A_10      AAAGATATAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T1A_05      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T1A_16      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T1A_08      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2A_09      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2A_06      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2A_04      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
            ******.*****************************************************

T1A_10      GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
T1A_05      GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
T1A_16      GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGTTCAGGACTGGAAA        240
T1A_08      GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
T2A_09      GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
T2A_06      GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTAAGGACTGGAAA        240
T2A_04      GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
            **************************.******************* *.***********

T1A_10      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T1A_05      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T1A_16      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA        300
T1A_08      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2A_09      AGTTTTTTTTATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2A_06      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2A_04      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
            *********:******************************************* ******

T1A_10      GGATCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA        360
T1A_05      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA        360
T1A_16      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA        360
T1A_08      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA        360
T2A_09      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA        360
T2A_06      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA        360
T2A_04      GGATCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA        360
            **.*********************************************************
```

```
T1A_10          GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T1A_05          GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T1A_16          GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA          420
T1A_08          GCTGTTAGGATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T2A_09          GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T2A_06          GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
T2A_04          GCTGTTAGGATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA          420
                ********.****.***************************************.******

T1A_10          TATAATGTGAAAG-----------------------------------------------          433
T1A_05          TATAATGTGAAAA-----------------------------------------------          433
T1A_16          TATAATGTGAAGG-----------------------------------------------          433
T1A_08          TATAATGTGAAGG-----------------------------------------------          433
T2A_09          TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA          480
T2A_06          TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA          480
T2A_04          TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA          480
                ***********..

T1A_10          -----------------CTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA          476
T1A_05          -----------------GCCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA          476
T1A_16          -----------------CTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA          476
T1A_08          -----------------CTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA          476
T2A_09          CCGTATTCGATTTGGCCCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA          540
T2A_06          CCGTATTCGATTTGGCCTCCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA          540
T2A_04          CCGTATTCGATTTGGCCCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA          540
                                   *****************************************

T1A_10          TTTAGATGCTGGCCAGCAGATAAAGATTAAGCAGGACCGCATTTCTAAAAGATGGTATCT          536
T1A_05          TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT          536
T1A_16          TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT          536
T1A_08          TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT          536
T2A_09          TTTAGATGCTGTCCAGCAGATAAAGATAAAGCAAGATCATATCTCTAAAAAATGGTATCT          600
T2A_06          TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT          600
T2A_04          TTTAGATGCTTTACAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT          600
                **********  .**************:*****.** *. ** *******.*********

T1A_10          ACTAATCATCTATAAGGTCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAAT          596
T1A_05          ACTAATCATCTATAAGACCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAGT          596
T1A_16          CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT          596
T1A_08          ACTAATCATCTATAAGACCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAGT          596
T2A_09          CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT          660
T2A_06          CTTAATTATCTATAAGACCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAAT          660
T2A_04          ACTAATCGTCTATAAGAGCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAAT          660
                . **** .**** **..  .*.**.. ********.*** **. **************.*

T1A_10          AGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAACTCTGAGTGTTATAT          656
T1A_05          AGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGAGTGTTATAT          656
T1A_16          AGATCTAGGTCTTGATAATTTGGCTACTTTAATATTTAAAAACAATTCTGATTGTTATAT          656
T1A_08          TGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATAT          656
T2A_09          TGATTTAGGCCTTGATAACTTAGCTGTACTAACATTTAAAGATAATTCTGATTGTTATAT          720
T2A_06          AGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTTTGATTGTTATAT          720
T2A_04          AGATCTAGGTCTTGATAATTTGGCTACTTTAATATTTAAAAACAATTCTGATTGTTATAT          720
                :*** **** ******** **.***. : *** *******.* ** * *** ********

T1A_10          TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTAATAAAGAAATTGCCAGACT          716
T1A_05          TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTAATAAAGAAATTGCCAGACT          716
```

```
T1A_16      TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT     716
T1A_08      TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT     716
T2A_09      TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT     780
T2A_06      TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT     780
T2A_04      TATCAATGGTAAAACTATTAAATCTAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT     780
            ************************ ***********************************

T1A_10      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA     776
T1A_05      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA     776
T1A_16      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA     776
T1A_08      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA     776
T2A_09      ACAAAGCATTAGAATTAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA     840
T2A_06      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA     840
T2A_04      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA     840
            *************** ********************************************

T1A_10      ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGCTGCAA     836
T1A_05      ATATCTGAGATTAAAGAGAAGAAATTATATTAGCAATTATCTCCATAAAGCTAGTTGCAA     836
T1A_16      ATATCTGAGATTAAAGAGAAAAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA     836
T1A_08      ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATATCCATAAAGCTAGCTGCAA     836
T2A_09      ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA     900
T2A_06      ATATCTGAGATTAAAGAGAAGAAATTATATTAGCGATTATCTCCATAAAGCTAGTTGCAA     900
T2A_04      ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA     900
            ********************.************..*****.************* *****

T1A_10      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA     896
T1A_05      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATCGGAGATATAAAAAA     896
T1A_16      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA     896
T1A_08      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA     896
T2A_09      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA     960
T2A_06      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATCAGAGATATAAAAAA     960
T2A_04      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA     960
            ********************************************* .*************

T1A_10      TATTAAAAAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA     956
T1A_05      TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA     956
T1A_16      TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA     956
T1A_08      TATTAAACAATGCAGCAATCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA     956
T2A_09      TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA     1020
T2A_06      TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCAATCCAGAGATTAAAAAA     1020
T2A_04      TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA     1020
            *******.********** ***********************.*****************

T1A_10      ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA     1016
T1A_05      ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA     1016
T1A_16      ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGGAATTGATGAAAGCTA     1016
T1A_08      ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA     1016
T2A_09      ATTAATTGAATACAAAGTTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA     1080
T2A_06      ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA     1080
T2A_04      ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA     1080
            ***************** **************************.***************

T1A_10      TACTTCTGGGTGTA----------------------------------------------     1030
T1A_05      TACTTCTGGATGTA----------------------------------------------     1030
T1A_16      TACTTCTGGATGTA----------------------------------------------     1030
T1A_08      TACTTCTGGATGTA----------------------------------------------     1030
T2A_09      TACTTCCGGATGTACTAAAGCTTTTAATTTTAATTACGCAAGGAAAGCTTTAGTATGACC     1140
```

```
T2A_06      TACTTCTGGGTGTACTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACC        1140
T2A_04      TACTTCTGGATGTACTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACC        1140
            ****** **.****

T1A_10      ------------------GTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA        1072
T1A_05      ------------------GTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA        1072
T1A_16      ------------------GTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA        1072
T1A_08      ------------------GTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA        1072
T2A_09      GTATTCGATTTGGCCGCTCTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA        1200
T2A_06      GTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA        1200
T2A_04      GTATTCGATTTGGCCGCTGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA        1200
                               *****************************************

T1A_10      AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGTCTATTAATTAATGCTG        1132
T1A_05      AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG        1132
T1A_16      AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG        1132
T1A_08      AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG        1132
T2A_09      AATCCAGAAGAATTACCAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG        1260
T2A_06      AATCCAGAAGAATTACCAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG        1260
T2A_04      AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGCGGGCCTATTAATTAATGCTG        1260
            **************** **********************.*** ****************

T1A_10      ATCAGAATGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC        1192
T1A_05      ATCAGAATGGTAGCTTTAATATACTTCGTAAATATCATAAGGATAAATGTATTCTCAGAC        1192
T1A_16      ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC        1192
T1A_08      ATCAGAATGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC        1192
T2A_09      ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC        1320
T2A_06      ATCAGAATGGTAGCTTTAATATACTTCGTAAATATCATAAGGATAAATGTATTCTCAGAC        1320
T2A_04      ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC        1320
            ************* ******************** ***** *******************

T1A_10      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT        1252
T1A_05      CCATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT        1252
T1A_16      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT        1252
T1A_08      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT        1252
T2A_09      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT        1380
T2A_06      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT        1380
T2A_04      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT        1380
            * **********************************************************

T1A_10      AA  1254
T1A_05      AA  1254
T1A_16      AA  1254
T1A_08      AA  1254
T2A_09      AA  1382
T2A_06      AA  1382
T2A_04      AA  1382
            **
```

Type 2A*

```
CLUSTAL O(1.2.1) multiple sequence alignment


T2A_09          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT      60
T2A*_13         ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT      60
                ************************************************************

T2A_09          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT      120
T2A*_13         GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT      120
                ********************************* **************************

T2A_09          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT      180
T2A*_13         AAAGATGTAAAAGCTGTCTATACTGAA---------------------------------      147
                ***************************

T2A_09          GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA      240
T2A*_13         ------------------------------------------------------------      147


T2A_09          AGTTTTTTTTATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA      300
T2A*_13         ------------------------------------------------------------      147


T2A_09          GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA      360
T2A*_13         --------------------TATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA      187
                                    ****************************************

T2A_09          GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA      420
T2A*_13         GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA      247
                *************.***************************************.******

T2A_09          TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA      480
T2A*_13         TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA      307
                ************************************************************

T2A_09          CCGTATTCGATTTGGCCCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA      540
T2A*_13         CCGTATTCGATTTGGCCTCCTTAATTTTGAGCTGCTTGAAGCAGTTCAAAGCATTATAGA      367
                *****************  **************** ************************

T2A_09          TTTAGATGCTGTCCAGCAGATAAAGATAAAGCAAGATCATATCTCTAAAAAATGGTATCT      600
T2A*_13         TTTAGATGCTGTCCAACAGATAAAGATAAAGCAAGATCATATCTCTAAAAGATGGTATCT      427
                ***************.**********************************.*********

T2A_09          CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT      660
T2A*_13         ACTAATCATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT      487
                . **** *****************************************************

T2A_09          TGATTTAGGCCTTGATAACTTAGCTGTACTAACATTTAAAGATAATTCTGATTGTTATAT      720
T2A*_13         AGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGAGTGTTATAT      547
                :*** **** ******** **.***. : ***********.* ******** ********

T2A_09          TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT      780
T2A*_13         TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT      607
                ************************************************************

T2A_09          ACAAAGCATTAGAATTAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA      840
```

```
T2A*_13      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       667
             *************** ********************************************

T2A_09       ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA       900
T2A*_13      ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATATCCATAAAGCTAGTTGCAA       727
             ****************************************.*******************

T2A_09       AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA       960
T2A*_13      AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATCGGAGACATAAAAAA       787
             ********************************************* ***** ********

T2A_09       TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       1020
T2A*_13      TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       847
             ************************************************************

T2A_09       ATTAATTGAATACAAAGTTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1080
T2A*_13      ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       907
             ***************** ******************************************

T2A_09       TACTTCCGGATGTACTAAAGCTTTTAATTTTAATTACGCAAGGAAAGCTTTAGTATGACC       1140
T2A*_13      TACTTCCGGATGTACTAAAGCTTTTAATTTTAGTTACGCAAGGTAAGCTTTAGTATGACC       967
             ********************************.**********:****************

T2A_09       GTATTCGATTTGGCCGCTCTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1200
T2A*_13      GTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGCAATTATGATA       1027
             ******************.****************************** ** *******

T2A_09       AATCCAGAAGAATTACCAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1260
T2A*_13      AATCCAGAAGAATTACTAGGGGTCTCTTTAAAACTAACGAGGGCTTATTAATTAATGCTG       1087
             **************** **.************************ ***************

T2A_09       ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1320
T2A*_13      ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1147
             ************************************************************

T2A_09       CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1380
T2A*_13      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1207
             ************************************************************

T2A_09       AA 1382
T2A*_13      AA 1209
             **
```

Type 2B/2B*/2C

```
CLUSTAL O(1.2.1) multiple sequence alignment


T2C_03          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2C_02          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2A_09          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2B_17          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2B*_18         ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2A_06          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2A_04          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2B_20          ATGCGATTATCATTTAAATTCAACCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
T2B_19          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT        60
                *********************** ************************************

T2C_03          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2C_02          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2A_09          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2B_17          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2B*_18         GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2A_06          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2A_04          GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT        120
T2B_20          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
T2B_19          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT        120
                ********************************** *************************

T2C_03          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2C_02          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2A_09          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2B_17          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2B*_18         AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2A_06          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2A_04          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2B_20          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
T2B_19          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT        180
                ************************************************************

T2C_03          GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTAAGGACTGGAAA        240
T2C_02          GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTAAGGACTGGAAA        240
T2A_09          GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
T2B_17          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTAAGGACTGGAAA        240
T2B*_18         GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTAAGGACTGGAAA        240
T2A_06          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTAAGGACTGGAAA        240
T2A_04          GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
T2B_20          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA        240
T2B_19          GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCTGTTAGCTCAGGACTGGAAA        240
                ************************** ************* ******* ***********

T2C_03          AGTTTTTTTTATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA        300
T2C_02          AGTTTTTTTTATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2A_09          AGTTTTTTTTATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2B_17          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAAGGTCAGCCA        300
T2B*_18         AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2A_06          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2A_04          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA        300
T2B_20          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA        300
T2B_19          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA        300
```

```
             ********* **************************************** ** ******

T2C_03       GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
T2C_02       GGGTCACCTAATTTTAAACATATGAACAGTAATTCCTGTGAAATAATTTTTACCAATTTA       360
T2A_09       GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
T2B_17       GGGTCACCGAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATCTG       360
T2B*_18      GAGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
T2A_06       GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
T2A_04       GGATCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
T2B_20       GGATCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
T2B_19       GGATCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
             *  ***** ************************ *********************** *

T2C_03       GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2C_02       GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2A_09       GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2B_17       GCTGTTAGAATTAAAGATAATAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2B*_18      GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2A_06       GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2A_04       GCTGTTAGGATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2B_20       GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
T2B_19       GCTGTTAGAATTAGAGATAATAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA       420
             ******** **** ****** ***************************************

T2C_03       TATAATGTGAAG------------------------------------------------       432
T2C_02       TATAATGTGAAG------------------------------------------------       432
T2A_09       TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA       480
T2B_17       TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA       480
T2B*_18      TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA       480
T2A_06       TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA       480
T2A_04       TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA       480
T2B_20       TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA       480
T2B_19       TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA       480
             ************

T2C_03       ----------------GCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       476
T2C_02       ----------------GCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       476
T2A_09       CCGTATTCGATTTGGCCCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       540
T2B_17       CCGTATTCGATTTGGCCTTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       540
T2B*_18      CCGTATTCGATTTGGCC-TCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       539
T2A_06       CCGTATTCGATTTGGCCTCCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       540
T2A_04       CCGTATTCGATTTGGCCCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       540
T2B_20       CCGTATTCGATTTGGCCTCCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       540
T2B_19       CCGTATTCGATTTGGCCCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA       540
                                ***************************************** 

T2C_03       TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT       536
T2C_02       TTTAGATGCTGTCCAGCAGATAAAGATAAAGCAAGATCATATCTCTAAAAAATGGTATCT       536
T2A_09       TTTAGATGCTGTCCAGCAGATAAAGATAAAGCAAGATCATATCTCTAAAAAATGGTATCT       600
T2B_17       TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT       600
T2B*_18      TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT       599
T2A_06       TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT       600
T2A_04       TTTAGATGCTTTACAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT       600
T2B_20       TTTAGATGCTTTACAGCAGATAAAGGTTAAGCAGGACCGTATTTCTAAAAGATGGTATCT       600
T2B_19       TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGCATTTCTAAAAGATGGTATCT       600
             ********** * ************ * ***** ** *  ** ******* *********
```

```
T2C_03       CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT       596
T2C_02       CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT       596
T2A_09       CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT       660
T2B_17       ACTAATCATCTACAAAGTTAAAGAGGCAAAAGAAAATAAGAAATCTAACATAATGGCAGT       660
T2B*_18      CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT       659
T2A_06       CTTAATTATCTATAAGACCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAAT       660
T2A_04       ACTAATCGTCTATAAGAGCGAGGAAATAAAAGAAAATAATAACCCTAACATAATGGCAAT       660
T2B_20       ACTAATTATCTATACGACCGAGGAAATAAAAGGAAATAATAACCCTAACATAATGGCAGT       660
T2B_19       ACTAATCATCTATAAGGTCGAGGAAATAAAAGAAAATAATAACTCTAACATAATGGCAGT       660
               ****  **** *      * **   ***** ** *** **  ************** *

T2C_03       TGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATAT       656
T2C_02       TGATTTAGGCCTTGATAACTTAGCTGTACTAACATTTAAAGATAATTCTGATTGTTATAT       656
T2A_09       TGATTTAGGCCTTGATAACTTAGCTGTACTAACATTTAAAGATAATTCTGATTGTTATAT       720
T2B_17       AGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATAT       720
T2B*_18      TGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATAT       719
T2A_06       AGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTTTGATTGTTATAT       720
T2A_04       AGATCTAGGTCTTGATAATTTGGCTACTTTAATATTTAAAAACAATTCTGATTGTTATAT       720
T2B_20       TGATCTAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATAT       720
T2B_19       AGATATAGGTCTTGATAATTTGGCTACTTTAACATTTAAAAACAATTCTGATTGTTATAT       720
              *** **** ******** ** ***    *** ******* * **** ************

T2C_03       TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       716
T2C_02       TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       716
T2A_09       TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       780
T2B_17       TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       780
T2B*_18      TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       779
T2A_06       TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       780
T2A_04       TATCAATGGTAAAACTATTAAATCTAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       780
T2B_20       TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       780
T2B_19       TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT       780
             ************************ ***********************************

T2C_03       ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       776
T2C_02       ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       776
T2A_09       ACAAAGCATTAGAATTAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       840
T2B_17       ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       840
T2B*_18      ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       839
T2A_06       ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       840
T2A_04       ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       840
T2B_20       ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       840
T2B_19       ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA       840
             *************** ********************************************

T2C_03       ATATCTGAGATTAAAGAGAAAAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA       836
T2C_02       ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGCTGCAA       836
T2A_09       ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA       900
T2B_17       ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGCTGCAA       900
T2B*_18      ATATCTGAGATTAAAGAGAAAAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA       899
T2A_06       ATATCTGAGATTAAAGAGAAGAAATTATATTAGCGATTATCTCCATAAAGCTAGTTGCAA       900
T2A_04       ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA       900
T2B_20       ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGCTGCAA       900
T2B_19       ATATCTGAGATTAAAGAGAAAAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA       900
             ******************** ************ ******************** *****

T2C_03       AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATCGGAGATATAAAAAA       896
T2C_02       AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA       896
```

```
T2A_09        AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA       960
T2B_17        AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA       960
T2B*_18       AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA       959
T2A_06        AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATCAGAGATATAAAAAA       960
T2A_04        AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA       960
T2B_20        AATAGTTAATTTAGCAGTTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA       960
T2B_19        AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATCGGAGATATAAAAAA       960
              ******* ******** ****************************  *************

T2C_03        TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATTCAGAGATTAAAAAA       956
T2C_02        TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       956
T2A_09        TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       1020
T2B_17        TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       1020
T2B*_18       TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCA-GAGATTAAAAAA       1018
T2A_06        TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCAATCCAGAGATTAAAAAA       1020
T2A_04        TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       1020
T2B_20        TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       1020
T2B_19        TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA       1020
              ****************** *********************** **   ************

T2C_03        ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1016
T2C_02        ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1016
T2A_09        ATTAATTGAATACAAAGTTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1080
T2B_17        ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGGAATTGATGAAAGCTA       1080
T2B*_18       ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1078
T2A_06        ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1080
T2A_04        ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1080
T2B_20        ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1080
T2B_19        ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA       1080
              ***************** ************************** ***************

T2C_03        TACTTCTGGGTGTACTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACC       1076
T2C_02        TACTTCTGGATGTACTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACC       1076
T2A_09        TACTTCCGGATGTACTAAAGCTTTTAATTTTAATTACGCAAGGAAAGCTTTAGTATGACC       1140
T2B_17        TACTTCTGGATGTA----------------------------------------------       1094
T2B*_18       TACTTCTGGGTGTA----------------------------------------------       1092
T2A_06        TACTTCTGGGTGTACTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACC       1140
T2A_04        TACTTCTGGATGTACTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACC       1140
T2B_20        TACTTCTGGGTGT-----------------------------------------------       1093
T2B_19        TACTTCTGGATGT-----------------------------------------------       1093
              ****** ** ***

T2C_03        GTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1136
T2C_02        GTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1136
T2A_09        GTATTCGATTTGGCCGCTCTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1200
T2B_17        ------------------GTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1136
T2B*_18       ------------------GTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1134
T2A_06        GTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1200
T2A_04        GTATTCGATTTGGCCGCTGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1200
T2B_20        -----------------AGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1136
T2B_19        -----------------AGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA       1136
                                 *****************************************

T2C_03        AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGCGGGCCTATTAATTAATGCTG       1196
T2C_02        AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1196
T2A_09        AATCCAGAAGAATTACCAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1260
T2B_17        AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1196
```

```
T2B*_18         AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1194
T2A_06          AATCCAGAAGAATTACCAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1260
T2A_04          AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGCGGGCCTATTAATTAATGCTG       1260
T2B_20          AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1196
T2B_19          AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG       1196
                **************** ********************** ********************

T2C_03          ATCAGAATGGTAGTTTTAATATACTTCGTAAATATCATAACGATAAATGTATTCTCAGAC       1256
T2C_02          ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1256
T2A_09          ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1320
T2B_17          ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1256
T2B*_18         ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1254
T2A_06          ATCAGAATGGTAGCTTTAATATACTTCGTAAATATCATAAGGATAAATGTATTCTCAGAC       1320
T2A_04          ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1320
T2B_20          ATCAGAATGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1256
T2B_19          ATCAGAATGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC       1256
                ************* ******************** ***** *******************

T2C_03          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1316
T2C_02          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1316
T2A_09          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1380
T2B_17          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1316
T2B*_18         CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1314
T2A_06          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1380
T2A_04          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1380
T2B_20          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1316
T2B_19          CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT       1316
                ************************************************************

T2C_03          AA 1318
T2C_02          AA 1318
T2A_09          AA 1382
T2B_17          AA 1318
T2B*_18         AA 1316
T2A_06          AA 1382
T2A_04          AA 1382
T2B_20          AA 1318
T2B_19          AA 1318
                **
```

Type 3

```
CLUSTAL multiple sequence alignment by Kalign (2.0)


T1A_16      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT
T2A_09      ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT
T3_23       ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT
T3_22       ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT
T3_21       ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT
T3_12       ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT
T3_11       ATGCGATTATCATTTAAATTCAACCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT


T1A_16      GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT
T2A_09      GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT
T3_23       GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT
T3_22       GAATTAGCCTAGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT
T3_21       GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT
T3_12       GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTTAATTATCAGGTTAAAAACAAT
T3_11       GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT


T1A_16      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT
T2A_09      AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT
T3_23       AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT
T3_22       AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT
T3_21       AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT
T3_12       AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT
T3_11       AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT


T1A_16      GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGTTCAGGACTGGAAA
T2A_09      GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA
T3_23       GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA
T3_22       GACTACCTTCACTCCCATAACAGACAACAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA
T3_21       GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA
T3_12       GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCATTTAGCTCAGGACTGGAAA
T3_11       GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA


T1A_16      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA
```

```
T2A_09     AGTTTTTTTTATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA
T3_23      AGTTTTTTTAATTCACTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA
T3_22      AGTTTTTTTAATTCACTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA
T3_21      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA
T3_12      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGGCAGCCA
T3_11      AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA


T1A_16     GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA
T2A_09     GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA
T3_23      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA
T3_22      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA
T3_21      GGATCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATCTG
T3_12      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA
T3_11      GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA


T1A_16     GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA
T2A_09     GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA
T3_23      GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA
T3_22      GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA
T3_21      GCTATTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA
T3_12      GCTGTTAGAATTAAAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAATCTAAA
T3_11      GCTGTTAGAATTAGAGATAATAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA


T1A_16     TATAATGTGAAGG-----------------------------------------------
T2A_09     TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA
T3_23      TATAATGTGAAAAGCACTAAAGCTTTTAATTTATAA------------------------
T3_22      TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAA------------------------
T3_21      TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAA------------------------
T3_12      TATAATGTGAAAGTCACTAAAGCTTTTAATTTATAA------------------------
T3_11      TATAATGTGAAAGTCACTAAAGCTTTTAATTTATAA------------------------


T1A_16     -----------------CTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA
T2A_09     CCGTATTCGATTTGGCCCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGA
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------


T1A_16     TTTAGATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCT
T2A_09     TTTAGATGCTGTCCAGCAGATAAAGATAAAGCAAGATCATATCTCTAAAAAATGGTATCT
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------


T1A_16     CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT
T2A_09     CTTAATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGT
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------
```

```
T1A_16     AGATCTAGGTCTTGATAATTTGGCTACTTTAATATTTAAAAACAATTCTGATTGTTATAT
T2A_09     TGATTTAGGCCTTGATAACTTAGCTGTACTAACATTTAAAGATAATTCTGATTGTTATAT
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------

T1A_16     TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT
T2A_09     TATCAATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACT
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------

T1A_16     ACAAAGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA
T2A_09     ACAAAGCATTAGAATTAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAA
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------

T1A_16     ATATCTGAGATTAAAGAGAAAAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA
T2A_09     ATATCTGAGATTAAAGAGAAGAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAA
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------

T1A_16     AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA
T2A_09     AATAGTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAA
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------

T1A_16     TATTAAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA
T2A_09     TATTAAACAATGCAGCAAGCTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAA
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
T3_11      ------------------------------------------------------------

T1A_16     ATTAATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGGAATTGATGAAAGCTA
T2A_09     ATTAATTGAATACAAAGTTAAACTAAAAGGTATCAAAGTTGTTGAAATTGATGAAAGCTA
T3_23      ------------------------------------------------------------
T3_22      ------------------------------------------------------------
T3_21      ------------------------------------------------------------
T3_12      ------------------------------------------------------------
```

```
T3_11       ------------------------------------------------------------


T1A_16      TACTTCTGGATGTAGT--------------------------------------------
T2A_09      TACTTCCGGATGTACTAAAGCTTTTAATTTTAATTACGCAAGGAAAGCTTTAGTATGACC
T3_23       ----------------------------------TACGCAAGGTAAGCTTTAGTATGACC
T3_22       ----------------------------------TACGCAAGGTAAGCTTTAGTATGACC
T3_21       ----------------------------------TACGCAAGGAAAGCTTTAGTATGACC
T3_12       ----------------------------------TACGCAAGGAAAGCTTTAGTATGACC
T3_11       ----------------------------------TACGCAAGGAAAGCTTTAGTATGACC


T1A_16      --------------------TCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA
T2A_09      GTATTCGATTTGGCCGCTCTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA
T3_23       GTATTCGATTTGGCCGCTGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA
T3_22       GTATTCGATTTGGCCGCTGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA
T3_21       GTATTCGATTTGGCCGCTGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA
T3_12       GTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA
T3_11       GTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATA


T1A_16      AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG
T2A_09      AATCCAGAAGAATTACCAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG
T3_23       AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG
T3_22       AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG
T3_21       AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG
T3_12       AATCCAGAAGAATTACTAGAGGTCTCTTTAAAAATAACGAGGGCCTATTAATTAATGCTG
T3_11       AATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTG


T1A_16      ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC
T2A_09      ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC
T3_23       ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC
T3_22       ATCAGAATGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC
T3_21       ATCAGAATGGTAGCTTTAATATACTTCGTAAATATCATAACGATAAATGTATTCTCAGAC
T3_12       ATCAGAATGGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC
T3_11       ATCAGAATGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGAC


T1A_16      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT
T2A_09      CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT
T3_23       CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT
T3_22       CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT
T3_21       CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT
T3_12       CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT
T3_11       CTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCT


T1A_16      AA
T2A_09      AA
T3_23       AA
T3_22       AA
T3_21       AA
T3_12       AA
T3_11       AA
```

Type 3*

```
CLUSTAL O(1.2.1) multiple sequence alignment


T3*_15          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT      60
T3_11           ATGCGATTATCATTTAAATTCAACCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT      60
                *********************** ************************************

T3*_15          GAATTAGCCTGGCATATTAGTAAACTATATAATACAGTCAATTATCAGATTAAAAATAAT      120
T3_11           GAATTAGCCTGGCATTGCTCTAAATTATATAATATAGTCAATTATCAGATTAAAAATAAT      120
                ***************:  : **** ********* *************************

T3*_15          AAAGATGTAAAAGCTGTCTATACTGA----------------------------------      146
T3_11           AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT      180
                **************************

T3*_15          ------------------------------------------------------------      146
T3_11           GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGCTCAGGACTGGAAA      240
                

T3*_15          ------------------------------------------------------------      146
T3_11           AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA      300
                

T3*_15          -------------------ATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA      187
T3_11           GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA      360
                                   *****************************************

T3*_15          GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA      247
T3_11           GCTGTTAGAATTAGAGATAATAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA      420
                ******************** ***************************************

T3*_15          TATAATGTGAAGGTCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGA      307
T3_11           TATAATGTGAAAGTCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGA      480
                ***********.*********************************:**************

T3*_15          CTGTATTCGATTTGGCCGCTCTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTTTGA      367
T3_11           CCGTATTCGATTTGGCCGCTATTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGA      540
                * ******************.***********************************:***

T3*_15          TAAATCCAGAAGAATTTCCAGAGGTCTCTTTAAAACTAAGGAGGGCCTATTAATT-----      422
T3_11           TAAATCCAGAAGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGC      600
                ****************:* ******************** ***************

T3*_15          -----------------------------------------------AATGTATTCTCAG      435
T3_11           TGATCAGAATGGTAGCTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAG      660
                                                               *************

T3*_15          ATCTATCAAAGAGGCGAGAGATAATGGGTTTGTGGCCAATCCTTCAAGATTAAGGGTACC      495
T3_11           ACCTATCAAAGAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATC      720
                * *************************.*******.********************** *

T3*_15          TAAA      499
T3_11           CTAA      724
                 :**
```

Type MISC

```
CLUSTAL O(1.2.1) multiple sequence alignment


T1A_16          ATGCGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT       60
MISC_14         ATACGATTATCATTTAAATTCAAGCCTAAATTAAGCCATAAGCAATTAGTAATAATTAAT       60
                ** *********************************************************

T1A_16          GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTATCAGATTAAAAATAAT       120
MISC_14         GAATTAGCCTGGCATTGCTCTAAATTATATAATACAGTCAATTCTCAGACCTATCAAAGA       120
                ******************************************* *****   *  * *

T1A_16          AAAGATGTAAAAGCTGTCTATACTGAATTAGAAACTAGATATAAAAATAACTGGCATAAT       180
MISC_14         GGCGAGAGAT--------------------------------------------------       130
                   **   *

T1A_16          GACTACCTTCACTCCCATAACAGACAGCAGGCATTAAAGCAGTTAGTTCAGGACTGGAAA       240
MISC_14         ------------------------------------------------------------       130


T1A_16          AGTTTTTTTAATTCTCTCAAAGATTATAAAAAGAATCCTCAAAAATATAAGGGTCAGCCA       300
MISC_14         ------------------------------------------------------------       130


T1A_16          GGGTCACCTAATTTTAAACATATGAACAGTAATCCCTGTGAAATAATTTTTACCAATTTA       360
MISC_14         ------------------------------------------------------------       130


T1A_16          GCTGTTAGAATTAGAGATAACAAATTACTCTTATCCTTATCTAAAAAGATACAGTCTAAA       420
MISC_14         ------------------------------------------------------------       130


T1A_16          TATAATGTGAAGGCTCTTAATTTTGAGCTGCCTGAAGCAGTTCAAAGCATTATAGATTTA       480
MISC_14         ------------------------------------------------------------       130


T1A_16          GATGCTGTCCAGCAGATAAAGATTAAGCAGGACCGTATTTCTAAAAGATGGTATCTCTTA       540
MISC_14         ------------------------------------------------------------       130


T1A_16          ATTATCTACAAAGTTAAAGAGGCAAAAGAAAGTAAGAAATCTAACATAATGGCAGTAGAT       600
MISC_14         ------------------------------------------------------------       130


T1A_16          CTAGGTCTTGATAATTTGGCTACTTTAATATTTAAAAACAATTCTGATTGTTATATTATC       660
MISC_14         ------------------------------------------------------------       130


T1A_16          AATGGTAAAACTATTAAATCCAAAAATTCTTATTTTAATAAAGAAATTGCCAGACTACAA       720
MISC_14         ------------------------------------------------------------       130


T1A_16          AGCATTAGAATGAGGCAGTTAGCTACCAGTAAAATTAGAGATACTAAACGAATAAAATAT       780
MISC_14         ------------------------------------------------------------       130


T1A_16          CTGAGATTAAAGAGAAAAAATTATATTAGAGATTATCTCCATAAAGCTAGTTGCAAAATA       840
```

```
MISC_14     ------------------------------------------------------------      130

T1A_16      GTTGATTTAGCAATTGAAAATCAAGTAGAAACTATTGTAATTGGAGATATAAAAAATATT      900
MISC_14     ------------------------------------------------------------      130

T1A_16      AAACAATGCAGCAAACTTAAATCTTTTGTCCAAATACCGATCCAGAGATTAAAAAAATTA      960
MISC_14     ------------------------------------------------------------      130

T1A_16      ATTGAATACAAAGCTAAACTAAAAGGTATCAAAGTTGTTGGAATTGATGAAAGCTATACT      1020
MISC_14     ------------------------------------------------------------      130

T1A_16      TCTGGATGTAGTTCAGTAGATCTGGAAAAAATAAATAAAAGTAACTATGATAAATCCAGA      1080
MISC_14     ------------------------------------------------------------      130

T1A_16      AGAATTACTAGAGGTCTCTTTAAAACTAACGAGGGCCTATTAATTAATGCTGATCAGAAT      1140
MISC_14     ------------------------------------------------------------      130

T1A_16      GGTAGTTTTAATATACTTCGTAAATACCATAACGATAAATGTATTCTCAGACCTATCAAA      1200
MISC_14     ------------------------------------------------------------      130

T1A_16      GAGGCGAGAGATAATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCTAA     1254
MISC_14     ------------AATGGATTTGTGGACAATCCTTCAAGATTAAGGGTATCCTAA     172
                        ******************************************
```

APPENDIX D.

tnpA/tnpB INTER-ORF SEQUENCE ALIGNMENT

```
CLUSTAL O(1.2.1) multiple sequence alignment


Locus_13      -----------CTCCATTTTTCCTTTTATAAGCAAACATATGTATGGTATAATTATAGTA      49
Locus_03      -AAAAATCAAACCTCCATTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      59
Locus_02      AAAGATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_18      AAAAATCAAACCTCCATGTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_16      AAAAATCAAACCTCCATGTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_08      AAAAATCAAACCTCCATGTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_09      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_05      AAAGATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_10      AAAAATAAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_11      AAAGATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_12      AAAGATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_14      AAAGATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_17      AAAGATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_19      AAAGATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_20      -----------CTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      49
Locus_15      ---------------ATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      45
Locus_01      -----------CTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      49
Locus_23      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_22      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_21      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_06      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
Locus_04      AAAAATCAAACCTCCATTTTTTCTTTTACAAGCAAACATATGTATGATATAATTATAGTA      60
                                *** ****** ***************** *************

Locus_13      GAATGGAGGTGAAAAATCA      68
Locus_03      GGATGGAGGTGAAAAGTCA      78
Locus_02      GGATGGAGGTGAAAAATTA      79
Locus_18      GGATGGAGGTGAAAAATCA      79
Locus_16      GGATGGAGGTGAAAAATCA      79
Locus_08      GGATGGAGGTGAAAAATCA      79
Locus_09      GGATGGAGGTGAAAAGTCA      79
Locus_05      GGATGGAGGTGAAAAATCA      79
Locus_10      GGATGGAGGTGAAAAATCA      79
Locus_11      GGATGGAGGTGAAAAATCA      79
Locus_12      GGATGGAGGTGAAAAATCA      79
Locus_14      GGATGGAGGTGAAAAATCA      79
Locus_17      GGATGGAGGTGAAAAATCA      79
Locus_19      GGATGGAGGTGAAAAATCA      79
Locus_20      GGATGGAGGTGAAAAATCA      68
Locus_15      GGATGGAGGTGAAAAATCA      64
Locus_01      GGATGGAGGTGAAAAATCA      68
Locus_23      GGATGGAGGTGAAAAATCA      79
Locus_22      GGATGGAGGTGAAAAATCA      79
Locus_21      GGATGGAGGTGAAAAATCA      79
Locus_06      GGATGGAGGTGAAAAATCA      79
Locus_04      GGATGGAGGTGAAAAATCA      79
              * ************* * *
```

APPENDIX E.

IS605 tnpB ORF INSERT SEQUENCE ALIGNMENT

```
CLUSTAL O(1.2.1) multiple sequence alignment


Locus_15HI      TCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGACTGTATTCGATTT    60
Locus_22HI      TCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_18LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_13LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_17LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_09LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_19LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_20LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_04LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_06LI      TCACTAAAGCTTTTAATTTATAATACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_21HI      TCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_11HI      TCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_12HI      TCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_23HI      GCACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    60
Locus_02RI      ---CTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    57
Locus_04RI      ---CTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    57
Locus_06RI      ---CTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    57
Locus_13RI      ---CTAAAGCTTTTAATTTTAGTTACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    57
Locus_09RI      ---CTAAAGCTTTTAATTTTAATTACGCAAGGAAAGCTTTAGTATGACCGTATTCGATTT    57
Locus_03RI      ---CTAAAGCTTTTAATTTTAATTACGCAAGGTAAGCTTTAGTATGACCGTATTCGATTT    57
                   ****************::.:*********:*************** ***********

Locus_15HI      GGCCGCT 67
Locus_22HI      GGCCGCT 67
Locus_18LI      GGCC--- 64
Locus_13LI      GGCC--- 64
Locus_17LI      GGCC--- 64
Locus_09LI      GGCC--- 64
Locus_19LI      GGCC--- 64
Locus_20LI      GGCC--- 64
Locus_04LI      GGCC--- 64
Locus_06LI      GGCC--- 64
Locus_21HI      GGCCGCT 67
Locus_11HI      GGCCGCT 67
Locus_12HI      GGCCGCT 67
Locus_23HI      GGCCGCT 67
Locus_02RI      GGCCGCT 64
Locus_04RI      GGCCGCT 64
Locus_06RI      GGCCGCT 64
Locus_13RI      GGCCGCT 64
Locus_09RI      GGCCGCT 64
Locus_03RI      GGCCGCT 64
                ****
```

APPENDIX F.

IS605 LEFT END SEQUENCE ALIGNMENT

```
CLUSTAL O(1.2.1) multiple sequence alignment


Locus_04      TTTATATAAAATTGCCAAGAAAACTCCATCCAAGCTATGCATTGGCTGGAGATGAATTGG      60
Locus_05      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTAGGTGGAGATGAATTGG      60
Locus_21      TTTATTTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTAGGTGGAGATGAATTGG      60
Locus_03      TTTATCTGAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_01      TTTATCTAAAACTGCCAAGAAAACCCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_15      TTTATCTAAATATGCCAAGAAAACTCCATCCAAGCTATGCATTGTGTGGAGATAAATTGG      60
Locus_11      TTTATCTAAAACTGCCAAGAAAACTCCTTCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_12      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATAAATTGG      60
Locus_14      TTTATCTAAAACTGCCAAGAAAACTCCCTCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_16      TTTATCTAAAATTGCCAAGAAAACTCCATCCAAGCTATGCATTGGCTGGAGATGAATTGG      60
Locus_17      TTTATCTAAAACTGCTAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATAAATTGG      60
Locus_18      TTTATCTAAAATTGCCAAGAAAACTCCATCCAAGCTATGCATTGGCTGGAGATGAATTGG      60
Locus_23      TTTATCTAAAATTGCCAAGAAAACTCCATCCAAGCTATGCATTGGCTGGAGATGAATTGG      60
Locus_02      TTTATCTAAAATTGCCAAGAAAACTCCATCCAAGCTATGCATTGGCTGGAGATGAATTGG      60
Locus_22      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGCTGGAGATGAATTGG      60
Locus_06      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_08      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_10      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_19      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_09      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_20      TTTATCTAAAACTGCCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG      60
Locus_13      TTTATCTAAAATTGTCAAGAAAACTCCATTCAAGCTATGCATTGGGGGA-----------      49
              ***** *.**: **  ******** ** * *************.   *.
```

APPENDIX G.

IS605 RIGHT END SEQUENCE ALIGNMENT

```
CLUSTAL O(1.2.1) multiple sequence alignment


Locus_01        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_21        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_14        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_04        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_06        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_02        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_12        GCTATTAGGAGCAAAATTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_13        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_17        GCTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_19        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_08        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_23        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_20        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_11        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_09        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_03        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_16        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_10        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_05        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_18        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
Locus_22        ACTATTAGGAGCAAAACTTAAAAGCCAAACATCTTGTAAACTGACCTAGTAATATAGGTT      60
                 *************** *******************************************

Locus_01        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCCGACGCGAAGCTAGCATCTCTTT      120
Locus_21        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCCGACGCGAAGCTAGTATCTCTTT      120
Locus_14        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCCGACGCGAAGCTAGTATCTCTTT      120
Locus_04        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCCGACGCGAAGCTAGTATCTCTTT      120
Locus_06        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCCGACGCGAAGCTAGTATCTCTTT      120
Locus_02        GAACTTTAATCTATATGAAGCAGTTAAAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_12        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_13        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGATTTTGATTTAGG-      119
Locus_17        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_19        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCTAAATCTTGGTTTTGATTTAGA-      119
Locus_08        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCTAAATCTTGGTTTTGATTTAGA-      119
Locus_23        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_20        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_11        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_09        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_03        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_16        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_10        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_05        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_18        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAATCTTGGTTTTGATTTAGG-      119
Locus_22        GAACTTTAATCTATATGAAGCAGTTAGAAGCTCCATCTAAATCTTGGTTTTGATTTAGG-      119
                ************************** ******* **  *  *     *   ** *

Locus_01        TGATGCAAATCTTGGTTTTGATTTAGGTGGAGAGGTTCAC  160
Locus_21        TGATGCAAATCTTGGTTTTGATTTAGGTGGAGAGGTTCAC  160
Locus_14        TGATGCAAATCTTGGTTTTGATTTAGGTGGAGAGGTTCAC  160
Locus_04        TGATGCAAATCTTGGTTTTGATTTAGGTGGAGAGGTTCAC  160
Locus_06        TGATGCAAATCTTGGTTTTGATTTAGATGGAGAGGTTCAC  160
Locus_02        ---------------------------TGGAGAGGTTCAC  132
Locus_12        ---------------------------TGGAGAGGTTCAC  132
```

```
Locus_13      ---------------------------TGGAGAGGTTCAC  132
Locus_17      ---------------------------TGGAGAGGTTCAC  132
Locus_19      ---------------------------TGGAGAGGTTCAC  132
Locus_08      ---------------------------TGGAGAGGTTCAC  132
Locus_23      ---------------------------TGGAGAGGTTCAC  132
Locus_20      ---------------------------TGGAGAGGTTCAC  132
Locus_11      ---------------------------TGGAGAGGTTCAC  132
Locus_09      ---------------------------TGGAGAGGTTCAC  132
Locus_03      ---------------------------TGGAGAGGTTCAC  132
Locus_16      ---------------------------TGGAGAGGTTCAC  132
Locus_10      ---------------------------TGGAGAGGTTCAC  132
Locus_05      ---------------------------TGGAGAGGTTCAC  132
Locus_18      ---------------------------TGGAGAGGTTCAC  132
Locus_22      ---------------------------TGGAGAGGTTCAC  132
                                         *************
```

APPENDIX H.

IS605 LEFT END REVERSE COMPLEMENT ALIGNMENT

```
>>>Locus_09LE, 60 nt vs lalign-I20160419-034638-0135-52459438-pg.bsequence library

>>Locus_09LE_RC                                           (60 nt)
 Waterman-Eggert score: 36;  88.8 bits; E(1) <  6.5e-24
88.9% identity (88.9% similar) in 9 nt overlap (60-52:16-24)

      60
Locus_ CCAATTCAT
       ::::: :::
Locus_ CCAATGCAT
          20

>--
 Waterman-Eggert score: 33;  10.9 bits; E(1) <  0.84
72.2% identity (72.2% similar) in 18 nt overlap (36-20:36-53)

            30
Locus_ AGCTTGGATGG-AGTTTT
       :: ::   ::: ::::::
Locus_ AGTTTTCTTGGCAGTTTT
          40        50

>>Locus_09LE_RC                                           (60 nt)
 Waterman-Eggert score: 86;  26.4 bits; E(1) <  4.1e-05
65.2% identity (65.2% similar) in 46 nt overlap (15-60:1-46)


            20        30        40        50        60
 Locus_ CCAAGAAAACTCCATCCAAGCTATGCATTGGGTGGAGATGAATTGG
        ::::   : ::::: ::::   ::   :::: ::::: :   ::::
 Locus_ CCAATTCATCTCCACCCAATGCATAGCTTGGATGGAGTTTTCTTGG
                10        20        30        40

 >--
  Waterman-Eggert score: 56;  17.6 bits; E(1) <  0.017
 72.7% identity (72.7% similar) in 22 nt overlap (29-50:11-32)

        30        40        50
 Locus_ TCCAAGCTATGCATTGGGTGGA
        ::::  : :::::: :  ::::
 Locus_ TCCACCCAATGCATAGCTTGGA
                20        30

 >--
  Waterman-Eggert score: 34;  11.2 bits; E(1) <  0.78
 69.2% identity (69.2% similar) in 26 nt overlap (23-47:14-38)

              30        40
 Locus_ ACTCCATCCA-AGCTATGCATTGGGT
        :: : :: :: :::: :: :: : ::
 Locus_ ACCCAATGCATAGCT-TGGATGGAGT
             20         30
```

APPENDIX I.

IS605 RIGHT END REVERSE COMPLEMENT ALIGNMENT

```
 Waterman-Eggert score: 39;  11.0 bits; E(1) <  1
73.3% identity (73.3% similar) in 15 nt overlap (32-18:114-128)

        30        20
Locus_ ATGTTTGGCTTTTAA
       : :::: :::  :::
Locus_ AAGTTTTGCTCCTAA
           120

>>Locus_16RE_RC                                          (132 nt)
 Waterman-Eggert score: 82;  20.4 bits; E(1) <  0.013
64.4% identity (64.4% similar) in 59 nt overlap (63-120:13-70)

             70        80        90       100        110       120
Locus_ ACTTTAATCTATATGAAGCAGTTAGAAGCTCCCTCTAAAT-CTTGGTTTTGATTTAGGT
       :: : :::: : :  ::: : ::::: :   : ::::: : :::  : : :::: : ::
Locus_ ACCTAAATCAAAACCAAG-ATTTAGAGGGAGCTTCTAACTGCTTCATATAGATTAAAGT
             20        30         40        50        60        70

>--
 Waterman-Eggert score: 82;  20.4 bits; E(1) <  0.013
68.4% identity (68.4% similar) in 38 nt overlap (41-78:46-83)

               50        60        70
Locus_ CTGACCTAGTAATATAGGTTGAACTTTAATCTATATGA
       :: ::    : :::::: :: :: :: :: :::::: :
Locus_ CTAACTGCTTCATATAGATTAAAGTTCAACCTATATTA
          50        60        70        80

>--
 Waterman-Eggert score: 62;  16.0 bits; E(1) <  0.23
64.7% identity (64.7% similar) in 34 nt overlap (91-124:9-42)

              100       110       120
Locus_ CTCCCTCTAAATCTTGGTTTTGATTTAGGTGGAG
       ::::  :::::::        :::::::  ::::
Locus_ CTCCACCTAAATCAAAACCAAGATTTAGAGGGAG
       10        20        30        40
```

```
>>Locus_16RE_RC                                           (132 nt)
 Waterman-Eggert score: 42;  146.7 bits; E(1) <  1.2e-40
66.7% identity (66.7% similar) in 21 nt overlap (129-109:48-68)

    130       120       110
Locus_ AACCTCTCCACCTAAATCAAA
       :::  :: ::  :: :: :::
Locus_ AACTGCTTCATATAGATTAAA
        50        60

>--
 Waterman-Eggert score: 39;  11.0 bits; E(1) <  1
62.5% identity (62.5% similar) in 24 nt overlap (117-94:65-88)

            110       100
Locus_ TAAATCAAAACCAAGATTTAGAGG
       ::::    :::: : :::   :::
Locus_ TAAAGTTCAACCTATATTACTAGG
           70        80
```

APPENDIX J.

IS605 tnpB ORF INSERT REVERSE COMPLEMENT ALIGNMENT

```
>>Locus_22HI_RC                                           (67 nt)
 Waterman-Eggert score: 35;  83.1 bits; E(1) <  4.2e-22
100.0% identity (100.0% similar) in 7 nt overlap (40-34:55-61)

      40
Locus_ AAAGCTT
       :::::::
Locus_ AAAGCTT
           60

>>Locus_22HI_RC                                           (67 nt)
 Waterman-Eggert score: 70;  19.0 bits; E(1) <  0.0087
63.4% identity (63.4% similar) in 41 nt overlap (3-43:25-65)

             10        20        30        40
Locus_ ACTAAAGCTTTTAATTTATAATACGCAAGGTAAGCTTTAGT
       ::::::::::    :   :: ::   :    ::::::::::
Locus_ ACTAAAGCTTACCTTGCGTATTATAAATTAAAAGCTTTAGT
           30        40        50        60
```

```
66.7% identity (66.7% similar) in 27 nt overlap (21-47:21-47)

               30        40
Locus_ TAATACGCAAGGTAAGCTTTAGTATGA
       : ::::  ::: : : :::  :::: :
Locus_ TCATACTAAAGCTTACCTTGCGTATTA
               30        40

>--
 Waterman-Eggert score: 40;  11.7 bits; E(1) <  0.74
100.0% identity (100.0% similar) in 8 nt overlap (6-13:55-62)

          10
Locus_ AAAGCTTT
       ::::::::
Locus_ AAAGCTTT
           60

>--
 Waterman-Eggert score: 38;  11.2 bits; E(1) <  0.85
76.9% identity (76.9% similar) in 13 nt overlap (11-23:45-57)

               20
Locus_ TTTTAATTTATAA
       :: ::: ::: ::
Locus_ TTATAAATTAAAA
           50

>--
 Waterman-Eggert score: 34;  10.3 bits; E(1) <  0.97
62.9% identity (62.9% similar) in 35 nt overlap (22-54:14-46)

                30        40        50
Locus_ AATACG--CAAGGTAAGCTTTAGTATGACCGTATT
       ::::::  ::   :::   :::   ::  ::::::
Locus_ AATACGGTCATACTAAAGCTTACCTTG--CGTATT
            20        30        40

>--
 Waterman-Eggert score: 30;  9.3 bits; E(1) <  1
66.7% identity (66.7% similar) in 15 nt overlap (3-17:51-65)
```

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-element-mediated 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 tetroviral itegrases**. *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):989-996.

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-have-common-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):3325-3338.

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. Kersulyte D, Akopyants NS, Clifton SW, Roe BA, Berg DE: **Novel sequence organization and insertion specificity of IS605 and IS606: chimaeric transposable elements of *Helicobacter pylori***. *Gene* 1998, **223**(1):175-186.

28. Ronning DR, Guynet C, Ton-Hoang B, Perez ZN, Ghirlando R, Chandler M, Dyda F: **Active site sharing and subterminal hairpin recognition in a new class of DNA transposases**. *Molecular cell* 2005, **20**(1):143-154.

29. Guynet C, Hickman AB, Barabas O, Dyda F, Chandler M, Ton-Hoang B: **In vitro reconstitution of a single-stranded transposition mechanism of IS608**. *Molecular cell* 2008, **29**(3):302-312.

30. Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, Sommer S, Chandler M: **Single-stranded DNA transposition is coupled to host replication**. *Cell* 2010, **142**(3):398-408.

31. Pasternak C, Ton-Hoang B, Coste G, Bailone A, Chandler M, Sommer S: **Irradiation-induced Deinococcus radiodurans genome fragmentation triggers transposition of a single resident insertion sequence**. *PLoS genetics* 2010, **6**(1):e1000799.

32. Barabas O, Ronning DR, Guynet C, Hickman AB, Ton-Hoang B, Chandler M, Dyda F: **Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection**. *Cell* 2008, **132**(2):208-220.

33. Montano SP, Rice PA: **Moving DNA around: DNA transposition and retroviral integration**. *Current opinion in structural biology* 2011, **21**(3):370-378.

34. He S, Guynet C, Siguier P, Hickman AB, Dyda F, Chandler M, Ton-Hoang B: **IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer**. *Nucleic acids research* 2013, **41**(5):3302-3313.

35. Pasternak C, Dulermo R, Ton-Hoang B, Debuchy R, Siguier P, Coste G, Chandler M, Sommer S: **ISDra2 transposition in Deinococcus radiodurans is downregulated by TnpB**. *Mol Microbiol* 2013, **88**(2):443-455.

36. Krishna SS: **Structural classification of zinc fingers: SURVEY AND SUMMARY**. *Nucleic acids research* 2003, **31**(2):532-550.

37. Bao W, Jurka J: **Homologues of bacterial TnpB_*IS605* are widespread in diverse eukaryotic transposable elements**. *Mobile DNA* 2013, **4**(12).

38. 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.

39. Kamoun C, Payen T, Hua-Van A, Filee J: **Improving prokaryotic transposable elements identificaiton using a combination of *de novo* and profile HMM methods**. *BMC Genomics* 2013, **14**(1).

40. Robinson DG, Lee MC, Marx CJ: **OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences**. *Nucleic acids research* 2012, **40**(22):e174.

41. 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.

42. Mira AK, Lisa. and Siv GE Anderson: **Microbial genome evoluiton: sources of variability**. *Current opinion in microbiology* 2002, **5**(5):506-512.

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

44. 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.

45. 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.

46. 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.

47. 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):7529-7537.

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

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

50. Ling A, Cordaux R: **Insertion sequence inversions mediated by ectopic recombination between terminal inverted repeats**. *PLoS One* 2010, **5**(12):e15654.

51. Cerveau N, Leclercq S, Leroy E, Bouchon D, Cordaux R: **Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from Wolbachia endosymbionts**. *Genome biology and evolution* 2011, **3**:1175-1186.

52. Wagner A, Lewis C, Bichsel M: **A survey of bacterial insertion sequences using IScan**. *Nucleic acids research* 2007, **35**(16):5284-5293.

53. Cordaux R: **Gene conversion maintains nonfunctional transposable elements in an obligate mutualistic endosymbiont**. *Molecular biology and evolution* 2009, **26**(8):1679-1682.

54. Wagner A: **Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes**. *Molecular biology and evolution* 2006, **23**(4):723-733.

55. Sawyer SA, Dykhuizen DE, DuBose RF, Green L, Mutangadura-Mhlanga T, Wolczyk DF, Hartl DL: **Distribution and abundance of insertion sequences among natural isolates of Escherichia coli**. *Genetics* 1987, **115**(1).

56. Qiu N, He J, Wang Y, Cheng G, Li M, Sun M, Yu Z: **Prevalence and diversity of insertion sequences in the genome of Bacillus thuringiensis YBT-1520 and comparison with other Bacillus cereus group members**. *FEMS Microbiol Lett* 2010, **310**(1):9-16.

57. 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.

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

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

60. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A *et al*: **The Pfam protein families database: towards a more sustainable future**. *Nucleic acids research* 2016, **44**(D1):D279-285.

61. 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.

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

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

64. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E *et al*: **ExPASy: SIB bioinformatics resource portal**. *Nucleic acids research* 2012, **40**(Web Server issue):W597-603.

65. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic acids research* 2003, **31**(13):3406-3415.

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

67. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R: **The EMBL-EBI bioinformatics web and programmatic tools framework**. *Nucleic acids research* 2015, **43**(W1):W580-584.

68. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J *et al*: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega**. *Mol Syst Biol* 2011, **7**:539.

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

70. **DNA folding form** [http://unafold.rna.albany.edu/?q=mfold/DNA-Folding-Form]

VITA

Mike Sadler was born in Ogden, Utah. He attended Davis High School in Kaysville, Utah. In 2012 he graduated with a Bachelors of Science in Microbiology from Weber State University in Ogden, Utah. After a brief time in the work force, Mike returned to academia to pursue a Masters degree in Applied and Environmental Biology in 2014. He earned the degree from Missouri University of Science and Technology in July 2016.