Fall 2018

Role of Miniature Inverted-Repeat Transposable Elements in Genome Evolution of Potamopyrgus antipodarum

Emily Lyon

ROLE OF MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS IN GENOME EVOLUTION OF POTAMOPYRGUS ANTIPODARUM

by

Emily Lyon

A thesis submitted in partial fulfillment of the requirements
for graduation with Honors in the Biology

_____

Maurine Neiman
Thesis Mentor

Fall 2018

All requirements for graduation with Honors in the
Biology have been completed.

_____

Lori Adams
Biology Honors Advisor

ROLE OF MINIATURE INVERTED-REPEAT TRANSPOSABLE ELEMENTS IN GENOME EVOLUTION OF POTAMOPYRGUS ANTIPODARUM

by

Emily Ann Lyon

A thesis submitted in partial fulfillment of the requirements for graduation with Honors in the Department of Biology

Maurine Neiman
Thesis Mentor

Fall 2018

All requirements for graduation with Honors in the Department of Biology have been completed.

Lori Adams, PhD.
Biology Honors Advisor

ABSTRACT

Transposable elements are genomic parasites that move within the genome and can cause gene and genome evolution. Transposable elements make up significant portions of many eukaryotic genomes but have been little studied in animals. This research study focuses on characterizing and identifying a type of transposable element, called miniature inverted-repeat transposable elements (MITEs) within the *Potamopyrgus antipodarum* genome. MITEs are particularly small transposable elements which can occur in thousands of copies within a genome. This research is conducted using the genome of *P. antipodarum*, a species of mud snail that is native to New Zealand. This genome is currently being annotated by the Neiman Lab. My research focuses primarily on identifying and characterizing the MITEs that are present within this genome. Whenever assembling a new genome, the transposable element content of that genome should be assessed, which can only be done after these elements are identified and characterized. I identified the likely superfamilies and key characteristics of these MITEs. In my research I also assessed the genomic locations of these MITEs, determining if they are inserted in exons, introns, or intergenic regions. I discuss the proportions of MITEs inserted in these genomic locations and the implications of these insertions. Finally, these genomic insertions are assessed both based on MITE families and on all MITE sequences.

ACKOWLEDGEMENTS

None of this project would have been possible without the support of Kyle McElroy, a graduate student in the Neiman Lab. He was not only instrumental in conceptualizing this project, but also in every step since. He provided constant guidance in using the various programs and bioinformatic processes necessary for this research project. Kyle was also influential in determining the direction of the project throughout. Both Kyle McElroy and Maurine Neiman were also crucial in the communication of this research. I have presented and written about this research several times, each with the indispensable aid of Maurine and Kyle. These two have been such inspiring mentors throughout my research project.

I also want to thank all members of the Neiman Lab. The collective work of these individuals is what created the research that my project was built upon. The reference genome assembly of *P. antipodarum* was particularly instrumental for my research. Maurine Neiman, John Logsdon, Jeffrey Boore, Laura Bankers, Peter Fields, Joe Jalinsky, Katelyn Larkin, Kyle McElroy, Joel Sharbrough, Cindy Toll, and Peter Wilton all worked on creating this reference genome. In particular, Joel Sharbrough assisted greatly with this project, not only by creating the script I used to determine MITE sequence locations within the genome, but also by assisting with conceptualizing certain aspects of this project.

I also want to acknowledge the Biology Honors Program at the University of Iowa. This program is responsible for my initial participation in research, and for the eventual authorship of this paper. Lori Adams was essential in the writing process and the eventual final product of this paper. Her continued interest in my work inspires me to want to share my research with others.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Genomic changes drive evolution. Transposable elements (TEs) are a major source of these genomic changes. TEs are genetic sequences of varying lengths that exist and sometimes move within the genome of a host organism. These TEs use the host's resources to move and replicate within the genome. TEs can make up a substantial fraction (>50%) of a genome, especially in multicellular eukaryotes, and can alter gene function and phenotypes, induce chromosome rearrangements, provide raw materials for new genes, and even cause speciation (Feschotte & Pritham, 2007). TE dynamics can also produce small genomic changes that lead to major diversification of genome architecture (Feschotte & Pritham, 2007).

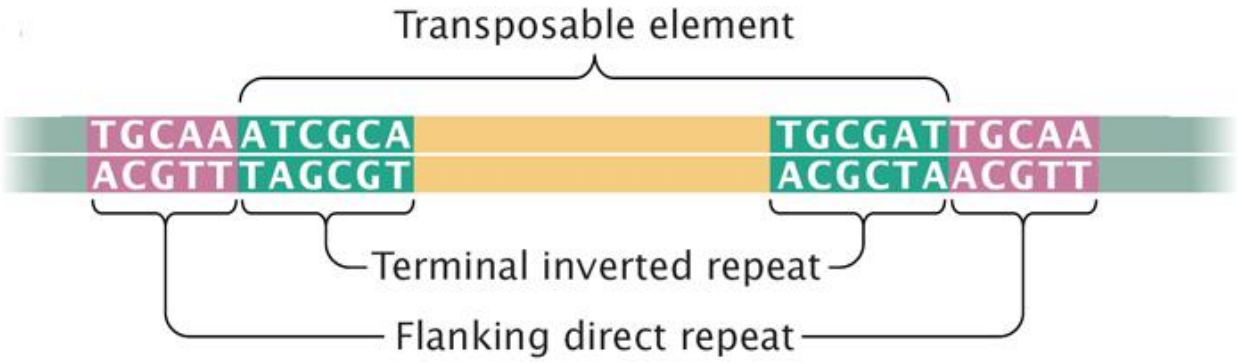
There is immense variation in the types of TEs. TEs are usually either DNA transposons (which move by a cut and paste mechanism) or retrotransposons (which move by a copy and paste mechanism). Within these categories a TE can be autonomous or non-autonomous. Autonomous elements can move themselves, while the mobility of non-autonomous TEs is reliant on similar autonomous elements. My honors thesis focus is on a subclass of TEs called miniature inverted repeat transposable elements (MITEs). MITEs are a nonautonomous (movement reliant on other TEs) subclass of DNA transposons, themselves a distinct class of TEs (Robillard et al, 2016). MITEs and DNA transposons are both found in most eukaryotic genomes. While much is known about the MITEs and TEs present in some organisms (e.g., maize), there is little known about TEs in the vast majority of eukaryotes, even though TEs are present in all but a few taxa. Because TEs can have huge impacts on genome evolution and contribute to speciation, they can also have serious deleterious consequences including sterility and lethality, broad characterization of TE activity across eukaryotes is important (Feschotte & Pritham, 2007). Here I focus on characterizing the abundance and dynamics of MITES in mollusks, an ancient, diverse, and biologically and economically important phylum.

Very little is known about how molluskan genomes evolve. The extent to which the insights from the organisms in which MTES have been well characterized (insects, worms, plants) can apply to mollusks is limited because these other organisms have not shared evolutionary histories with mollusks for hundreds of millions of years.

Because TE insertion and excision can cause mutations that are at least mildly deleterious, natural selection generally acts to remove TEs from a population. TEs might impose fitness penalties through insertion within a gene, ectopic recombination between copies of a TE, and TE insertion effects on gene expression (Barron et al, 2014). DNA transposons are especially likely to affect gene expression because this type of TE tends to insert relatively close to genes and often affect gene regulation (Feschotte & Pritham, 2007)

All DNA transposons - including MITES - have key features by which they can be identified and classified into families of TEs that share recent common descent (Figure 1). Some examples of these features include the terminal inverted repeats (TIRs) and flanking direct terminal repeats (DTRs). The transposon protein targets a specific sequence during genome insertion, which becomes the DTR (Wicker et al, 2007). MITEs include the key features of TIRs and TSDs, but do not contain the transposase gene. MITEs are formed when deletions occur in DNA transposons but share the same defining features of DNA transposons. Individual MITE elements are usually only 100-600 bp in length. They retain the TIRs and TSDs of the autonomous element, allowing the transposase of the MITE's parental element to move the MITE. MITES move in a cut-and-paste manner, which means that element movement does not directly generate a new copy. MITEs can have particularly striking impacts on the genome because of their proliferative nature, creating large-scale rearrangements and deletions, and altering gene expression (Yang et al, 2013).

Indeed, MITES replicate very rapidly, usually have high sequence homogeneity, and occur in high copy number (up to thousands of copies per genome) compared to other types of TEs. MITEs often experience amplification "bursts", defined as rapid accumulation in a genome, which allows MITEs to drive genomic variation and changes in gene expression (Feschotte & Pritham, 2007). The specific conditions that contribute to differential activity in MITE families remains poorly understood. There is a paradox where MITEs can be present in thousands of copies within a genome and not cause the activity of the autonomous parental element of that MITE to be downregulated. It would be expected that many MITEs copies using the transposase of the autonomous parental element would inhibit the parental



https://www.nature.com/scitable/content/many-transposable-elements-have-common-characteristics-29563

Figure 1: **Diagram of Transposon Insertion** Key features present within MITEs and all DNA transposons: TIRs and TSDs. TIRs are inverted complements of each other, while DTRs are direct copies of each other present on the outside of the TIRs. Both TIRs and TSDs are key MITE characteristics that must be present. The yellow portion in this figure represents the area of sequence homology present in MITEs between the two TIRs

element from being able to move within the genome because its transposase was always occupied. If MITEs become abundant within the genome it would also be expected that the high number of MITEs would activate host silencing mechanisms that would also silence the autonomous copy. Despite these predictions, Feschotte and Pritham (2007) found that genomes with high numbers of TEs also tended to have MITE copy numbers. This creates what is known as the MITE paradox. The reason for this paradox is still unknown and is an area of further study. Answering this question has to start with characterizing the MITEs present within a genome.

Another feature of MITEs that warrant further study is why MITEs can occur in these high copy numbers. MITEs can have thousands of copies within a genome, and all these copies of a single MITE sequence make up what is known as a MITE family. The sequences of a MITE family all match the same consensus sequence. MITE families are all moved by the same consensus sequence. It is a MITE family that has thousands of copies within the genome, and each of these copies individually is a MITE sequence. Signatures of recent amplification bursts include relatively high MITE copy number compared to other MITE families within the genome and relatively high sequence identity across MITE sequences (Feschotte & Pritham, 2007).

The mechanisms responsible for MITE replication and the selective forces that operate against replication influence MITE placement within a genome, whether and how MITES affect gene expression, and whether MITEs are actively replicating. It is not clear if actively moving MITEs tend to be located near genes (Wessler, 2010) or distributed evenly throughout the genome (Lu et al, 2012). Both Wessler et al. (2010) and Lu et al. (2012) found that those MITEs near a gene influence that gene. The relatively small size of MITEs means that these elements might impose a relatively low selective burden, possibly explaining their high copy number (Barron, 2014). Active MITEs do not contain any open reading frames for genes but may become integrated into a genome by providing raw genetic material for evolution.

MITEs are a unique and important type of TE in the *Potamopyrgus antipodarum* genome, the system in which I will be doing my research. *P. antipodarum* are a species of mud snail that is native to New Zealand. They have been used primarily as a natural system for studying the maintenance of sex

because they have both sexual and asexual lineages coexisting in New Zealand lakes. *P. antipodarum* is also a highly invasive organism, and its genome has also been instrumental in studying polyploidy, invasiveness, coevolution, and much more. One of the reasons classifying MITEs within *P. antipodarum's* genome is important is because the Neiman lab and others are currently creating a reference genome for P. antipodarum. When studying a new genome, it is important to catalogue what is in it in order to assess what unique features the genome has that may warrant further study.

The *P. antipodarum* genome is a particularly good system for studying MITEs because TEs in general compose at least 25% of the *P. antipodarum* genome, with MITEs alone making up 7.8% of the genome. This is a notably high MITE content relative to the MITE abundance in the majority of other eukaryotic genomes, though it is hard to determine this with complete certainty because many researchers do not search for MITEs specifically. The first step before any further analysis of MITEs can be done in the *P. antipodarum* genome is identification and classification of these elements within this genome. I used MITE Hunter software to catalog and identify MITE families within the *P. antipodarum* genome. Next, I conducted phylogenetic analysis to determine possible autonomous elements and genomic positions for MITEs, assign MITEs to possible superfamilies, and determine how recently MITEs may have undergone amplification bursts. MITEs that inserted into the genome or experienced amplification bursts relatively long ago will have acquired more mutations (i.e. tend to have larger pairwise nucleotide distances between copies) relative to more recently inserted MITEs. Accordingly, by calculating pairwise nucleotide distances for the copies of a given MITE family, I will provide information critical to a better understanding of the evolutionary history of that family.

METHODS & RESULTS

Identification and characterization of MITEs

I used MITE Hunter software (Han & Wessler, 2010) to identify possible MITEs in the *de novo* assembled genome of *P. antipodarum*. I then used a several-step process to either confirm possible MITEs as true MITEs or reclassify the putative MITE as simple random repeats or as another category of TE. First, I used the Kimura two-parameter model (as implemented in RepeatMasker (Smit et al, 2015) to calculate the pairwise nucleotide divergence values within each putative MITE family. These divergence values served as an indicator of how recently these MITEs were active within the genome.

In order to assess copies of a MITE family I used the MEMBER function of a toolkit named MITE analysis kit (MAK) (Yang & Hall, 2003). MAK uses several levels of analysis concerning MITEs in an automated process to retrieve sequences of MITE families and to determine the positions of these sequences. MAK has several functions that can be used to investigate different aspects of MITE families, such as possible autonomous parental elements as well as copy positions within the genome (Yang & Hall, 2003). I used MAK to extract at least 20 of the best blast hits for each potential MITE family. I selected the best blast hits by choosing the sequences with the highest bit scores, meaning they have the highest alignment score from the blast alignment and are the most confident matches to the consensus sequence of the MITE family. I then align these 20 copies in MEGA software (Kumar et al, 2016) using the Clustal W function (Figure 2). The alignment of these MITE family copies allows me to identify the TIRs and ensure that there was a consistent TSD for each MITE family. I used these sequence alignments to identify and characterize the TSDs and TIRs, calculate the length, and estimate the sequence homology of each of these possible MITEs.

The MEMBER function of MAK also provided the copy number of each potential MITE in the genome, which helped to either verify or discredit a sequence as a MITE. The primary criteria I used to verify a sequence as a MITE were (in order of importance) presence of TIRs, greater than 90% sequence homology among copies, consistent TSD sequence or length of sequence, at least 10 copies of the MITE

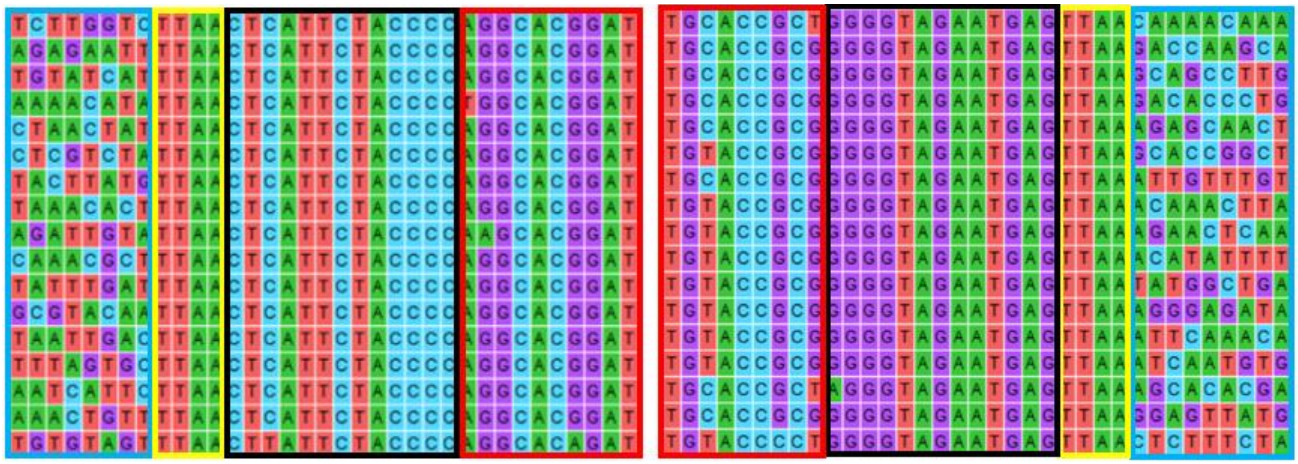


Figure 2: MITEs have characteristic TIRs, TSDs, and sequence homology between the TIRs This is example of 17 copies of MITE family 12_50900 aligned in MEGA. Different elements of MITE copies are outlined in colors as follows: black: TIRs; yellow: TSDs; red: sequence homology between the TIRs; and blue: nonhomologous sequences surrounding each MITE. White space between two alignments represents omitted sequence of continued sequence homology between two TIRs

within the genome, and lack of transposase gene (proving it is a nonautonomous element). I used the approaches described above to verify 21 MITE sequences in the *P. antipodarum* draft genome assembly (Table 1). Feschotte and Pritham (2007) lists DNA transposon superfamilies along with characteristics of that family, including their conserved TSDs. I used these TSDs to determine what superfamily the MITEs I manually identified in *P. antipodarum's* genome belong to.

Detecting related autonomous elements

Autonomous elements are responsible for moving MITEs around within the genome and allowing them to replicate. The implications are that detection of intact autonomous parental elements of a MITE within the genome suggests that MITE may still be active. With this logic in mind, I executed a case study with the goal of identifying the parental autonomous element of one MITE (11_11380) in the *P. antipodarum* genome. First, I used the LONG function, as implemented within MAK (Yang & Hall, 2003), to extract possible autonomous parental elements of MITE family 11_11380. I then blasted the output from LONG against NCBI's conserved domain database (Marchler-Bauer, 2004), which allowed me to determine whether any of the sequences in LONG's output contained the transposase gene. This CDD search did identify a sequence that encoded the transposase gene, making it likely that this sequence from the LONG output for the 11_11380 MITE family was a parental autonomous DNA transposon.

Next, I used MEGA to align this longer putatively autonomous element sequence to the MITE and determine whether the MITE and autonomous element sequences share TSDs, TIRs, and some sequence homology. A MITE needs to have the same TIRs and TSD as the autonomous element for the autonomous element to be able to move the MITE sequence within the genome. Sequence homology, especially at the ends of the MITE and autonomous element, supports that a deletion in the autonomous element created the MITE. This alignment also helped me determine whether this MITE likely resulted from a deletion in this autonomous element. I did find one sequence in the LONG output for this MITE

Table 1: MITEs that were verified manually

MITE Name	TIR Length	TSD	MITE Length	Copy Number	Superfamily
1_57582	13	TTAA	638	277	*PiggyBac*
1_6220	4	3(varies)	984	772	*CACTA*
10_4549	42	4 (varies)	785	46	*Banshee*
11_11380	12	TTAA	310	700	*PiggyBac*
12_50900	13	TTAA	508	886	*PiggyBac*
14_56701	13	6 (varies)	417	292	*Maverick*
15_21317	42	TA	223	39	Tc1/*mariner*
16_61215	7	9 (varies)	71	135	*Merlin*
2_59093	13	6 (varies)	405	697	*Maverick*
3_25813	13	TTAA	203	784	*PiggyBac*
3_39804	7	TA	450	15	Tc1/*mariner*
3_4722	14	6 (varies)	436	683	*Maverick*
3_52544	10	5 (varies)	443	83	*Maverick*
4_41337	14	TTAA	243	189	*PiggyBac*
4_58523	17	TTAA	330	34	*PiggyBac*
4_5858	5	6 (varies)	374	371	*Maverick*
4_7439	13	TTAA	592	58	*PiggyBac*
5_44591	18	TTAA	703	779	*PiggyBac*
6_56181	17	4 (varies)	507	55	*Banshee*
6_7449	24	3 (varies)	164	915	*CACTA*
9_24080	16	TTAA	680	765	*PiggyBac*

that contained a gene for transposase, had the same TSDs and TIRs as the MITE, shared internal sequence homology with the MITE, and even blasted to a known *PiggyBac* transposase (Figure 3). The shared TSDs of the two elements were TTAA, which is a known TSD of the *PiggyBac* family, which further supports that this longer sequence is the autonomous parental element of the identified MITE. This means that MITE family 11_11380 could still be moving within the *P. antipodarum* genome because its transposase (in the *PiggyBac* parental element) is still present in the genome.

Determining the genomic locations of MITEs

An additional 45 putative MITEs from the MITE-HUNTER output were verified by Kyle McElroy using the program PASTEClassifier for addition automated classification (Hoede et al, 2014). PASTEClassifier, with the output from MITE Hunter, provides the TIRs of the possible MITE sequences. Only those MITEs with TIRs that were less than 80% of the entire sequence were added to my data set. This limitation of 80% was set to conservatively rule out non-TE inverted repeats. These possible MITEs were added to the data set to obtain a larger sample size for tests of MITE impact on genes, for a total of 66 different MITEs. I then used RepeatMasker with the consensus sequences of these MITEs to annotate the locations for all copies of these 66 MTIE families in the draft *P. antipodarum* genome assembly.

I used tePositions_v2.py, an in-house custom python script written by Joel Sharbrough, with the RepeatMasker output to sort these MITE sequences into insertions into exons, introns, or intergenic regions based on the gene models in the current annotation of the draft *P. antipodarum* genome assembly. This program also provided the coordinates of the MITE relative to important genomic landmarks (e.g., 5'UTR, introns, exons). In order to focus only on putative insertions that were likely to represent true insertions and rule out false positive hits, I confined

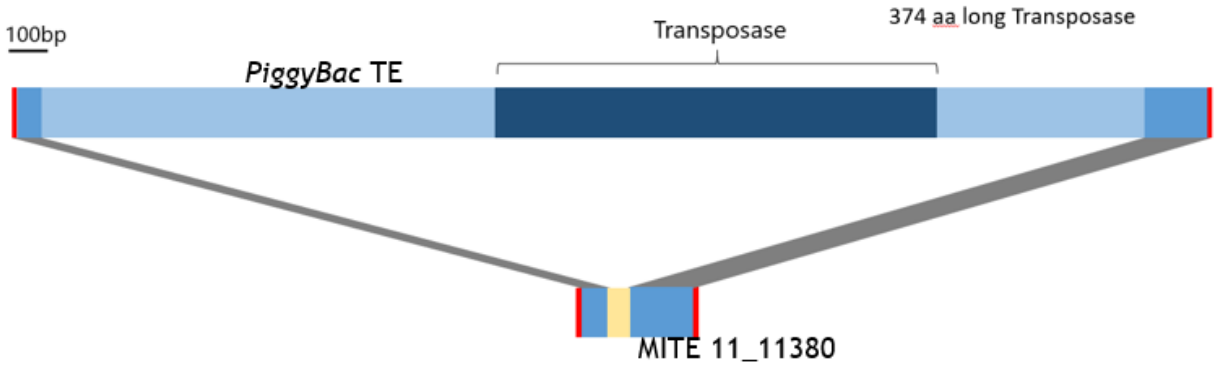


Figure 3: Schematic of autonomous *PiggyBac* element that likely controls MITE family 11_11380 The top block represents autonomous PiggyBac TE. The bottom block represents the consensus sequence of MITE family 11_11380. Color coding of schematic is as follows: lightest blue: portions of *PiggyBac* TE that did not align to the MITE family consensus sequence; intermediate shade of blue: sequences present in both autonomous *PiggyBac* TE and MITE family 11_11380; dark blue: transposase gene; red: identical TSDs for the two sequences; and light-yellow: short sequence in MITE family 11_11380 that did not align to the *PiggyBac* TE.

my analysis to sequences that comprised at least half of the length of the consensus sequence for the focal MITE.

All of the MITEs I analyzed were inserted in both introns and intergenic regions (Table 2). The number of introns a MITE family was inserted into ranged from 2 to 2,258 MITE sequences retained in introns per MITE family. Each MITE family was also present in intergenic regions, with a range of 22 to 12,764 MITE sequences inserted into intergenic regions of the genome per MITE family. All MITE families had more sequence inserted in intergenic regions than introns. Only 9 distinct MITEs were inserted into an exon, with 1 to 6 insertions in exons for each of these nine MITEs.

The MITE families with the highest number of exonic insertions had 6 sequences inserted into exons. For the 22 genes with MITE sequences inserted into their exons, I used Blast2GO (Götz et al, 2008) and blasted the genes using blastp (Marchler-Bauer et al, 2004) to determine the function of the gene (Table 3). Determining the function of these gens is particularly important because they have a MITE insertion in the coding sequence of the gene. Accordingly, the function of these genes is relatively likely to be affected by the MITE.

For a given gene I asked first whether it had a MITE inserted in one of its introns. I then assessed how many MITE sequences were present in a given gene's introns and how many MITE families are present in the introns of that gene. Each MITE family has distinct TSDs, TIRs, and aligns to a separate consensus sequence than all other MITE families. For introns one of my assessments was purely focused on MITE families and seeing how many distinct MITE families are present in the introns of each gene. This is the output that is represented in Figure 4. This is very different from counting the MITE sequences that are present in the introns of a gene.

Table 2: Genomic characterization of MITE insertion per MITE family

MITE families with no exon insertions left blank for that column.

MITE Name	Intron	Intergenic	Exon
1_16768	15	1917	
1_57582	149	475	
1_6220	328	12764	
10_23747	104	1371	
10_33869	421	617	
10_4549	43	179	2
10_5133	18	80	
11_10048	49	200	
11_11380	1060	1645	1
11_48230	9	246	
11_4947	920	1666	
11_49549	773	1453	
12_11930	100	498	
12_32852	47	251	
12_36597	25	569	
12_39013	248	867	
12_50900	606	1664	
12_56801	429	2815	
13_2981	104	583	
13_45149	24	117	

13_45651	19	1198	
14_56701	369	933	1
15_21317	192	1671	
15_21629	56	278	
15_54427	79	516	
15_6096	514	1098	
15_7273	145	2624	
16_2345	125	250	
16_43288	667	756	
16_45436	180	1171	
16_56600	493	818	
16_57507	38	450	
16_61215	1772	4500	6
16_8351	92	939	
2_23242	193	3030	
2_5617	11	22	
2_59093	1009	1458	2
3_22345	157	2805	
3_44702	65	714	
3_25813	1852	3805	
3_39804	20	94	
3_9890	455	674	
3_4722	725	1127	1

3_52544	51	276	5
4_23504	1286	4647	
4_24956	24	390	
4_25574	314	411	
4_41337	136	253	2
4_47140	101	194	
4_52037	339	1571	
4_58523	108	448	
4_5858	299	410	
4_62157	1303	3452	
4_7439	46	164	
4_9858	93	407	
5_44591	157	1112	
5_54873	2	99	
5_8363	115	280	
5_9100	882	5999	
6_51448	2258	5079	
6_56181	117	2227	
6_61336	20	86	
6_7449	911	4984	2
8_27886	267	545	
9_15823	684	811	
9_24080	338	1224	

Table 3: Predicted function of genes with exonic MITE insertion

MITE Name	Gene	Gene Function
10_4549	maker-1505-snap-gene-0.12	uncharacterized
	maker-183-snap-gene-5.17	Acyl-CoA dehydrogenase
11_11380	maker-725-snap-gene-0.18	Isoleucine—tRNA ligase, mitochondrial like
14_56701	snap_masked-390-processed-gene-2.4	ATP-dependent RNA helicase
16_61215	maker-272-snap-gene-7.17	Uncharacterized
	maker-204-snap-gene-8.26	Glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1-like
	maker-785-snap-gene-0.39	Multidrug resistance-associated protein 1-like
	snap_masked-242-processed-gene-13.14	Uncharacterized
	maker-301-snap-gene-12.6	Phosducine-like protein 3
	snap_masked-479-processed-gene-4.11	Chymotrypsinogen B
2_59093	maker-508-snap-gene-0.24	Leucine-rich repeat-containing protein
	maker-56-snap-gene-26.16	Origin recognition complex subunit 2
3_4722	maker-360-snap-gene-0.9	Ankyrin repeat protein
3_52544	maker-387-snap-gene-1.14	Mannose-1-phosphate guanyltransferase
	maker-439-snap-gene-0.15	Cation-dependent mannose-6-phosphate receptor-like
	maker-76-snap-gene-6.23	Cell death regulator Aven-like

	maker-715-snap-gene-1.26	Inosine-uridine preferring nucleoside hydrolase-like
	maker-139-snap-gene-2.19	Pre-mRNA splicing factor SPF27
4_41337	maker-246-snap-gene-1.8	IQ domain-containing protein D
	maker-634-snap-gene-0.11	Thioredoxin reductase mitochondrial
6_7449	snap_masked-582-processed-gene-1.12	Complement C1q-like protein
	maker-43-snap-gene-14.15	Structural maintenance of chromosomes protein 4-like

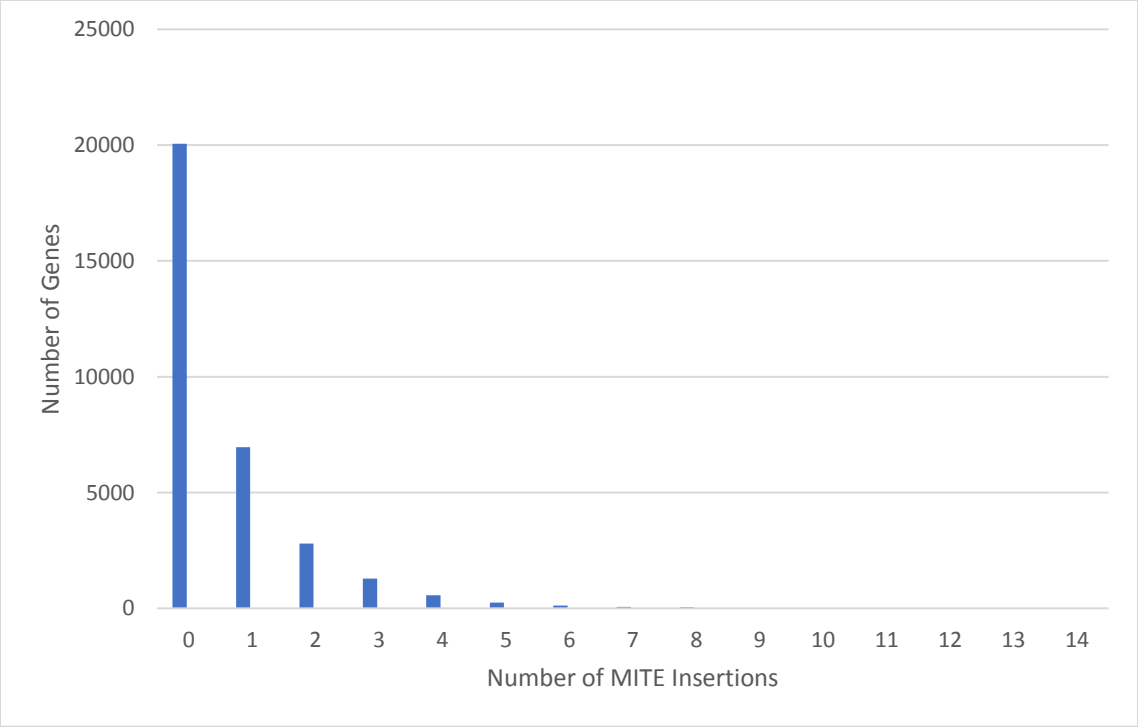


Figure 2: **MITE families per gene** Number of distinct MITE families' insertions into each gene's introns, counting each MITE family only once
The Y axis represents the number of genes with a certain number of MITE families inserted in its introns. The X axis represents the range of observed MITE families inserted into intron of gene.

Each MITE family can have up to thousands of copies within the genome, each copy having sequence homology to the MITE family it belongs to. In my assessment of MITE sequences within the introns of genes I simply counted how many MITE sequences in total are in the introns of a particular genes. This means that even if a MITE family had multiple copies within the same gene, I would count all these copies, along with the copies of any other MITE family that occurs in this gene's introns. The number of occurrences of MITE copies of all MITE families within a gene's introns is what was measured in the output of Figure 5.

Out of all the 32,148 predicted genes in *P. antipodarum's* genome, 62.3% of the genes (20,052 genes) did not have a MITE sequence from any family inserted into any introns. 21.6% of the genes (6,949 genes) only had one MITE family inserted into an intron. The gene that had the most MITE family insertions, gene maker-6-snap-gene-20.17-mRNA-1, had 14 different MITE families inserted in its introns (Figure 4). When looking at individual MITE sequences, instead of MITE families, 20.1% of the genes have only one MITE sequence inserted in its introns. The gene with the most MITE sequences inserted in its introns had 21 MITE sequence insertions (Figure 5). The gene with the most insertions for both MITE sequences and MITE families is maker-6-snap-gene-20.17-mRNA-1. This gene is single copy, ruling out the likelihood that duplicated genes may be more tolerant of MITE insertions. The protein produced from this gene blasts to a E3 ubiquitin-protein ligase HUWE 1-like isoform.

Having a MITE inserted into the intron of a gene can still affects gene expression or function. When MITEs become integrated and retained in the genome they are more likely to take on novel function as they mutate and become less likely to move. MITE families that have a relatively high pairwise nucleotide divergence values (calculated using RMasker) are likely to represent MITEs that inserted into the genome longer ago than MITEs with relatively low

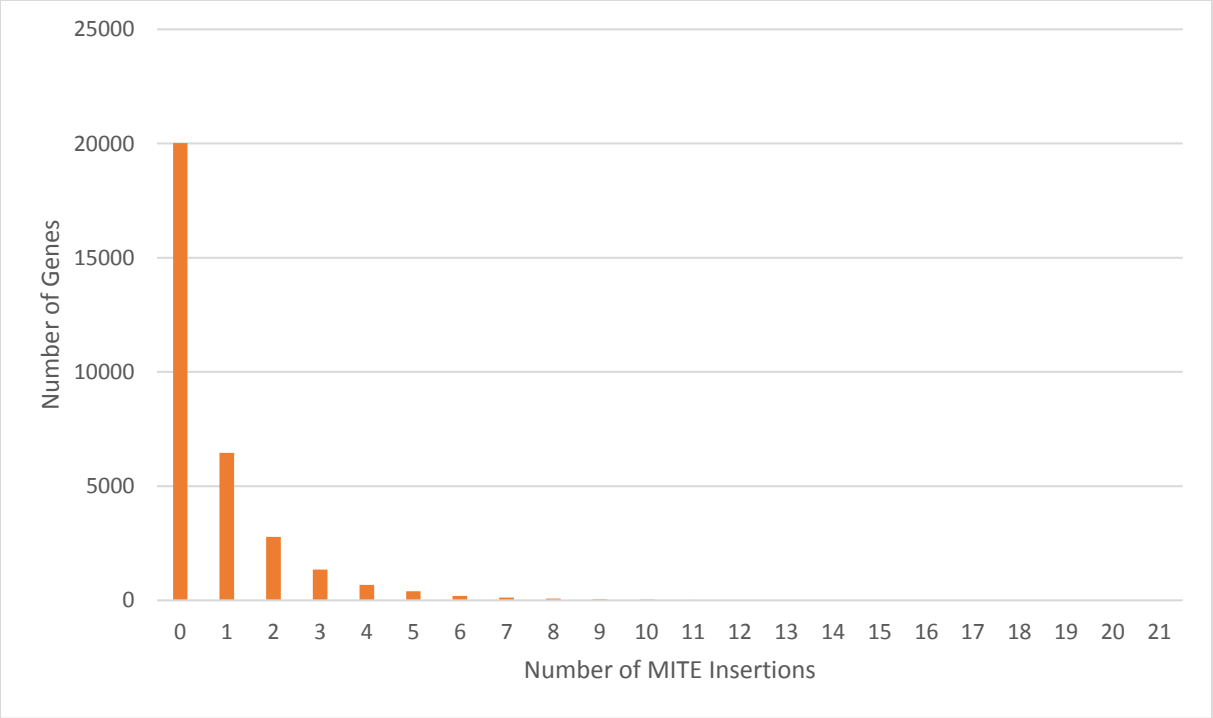


Figure 3: All MITE Sequence Insertions Number of MITE sequences inserted into each gene's introns, counting each MITE sequence even if a family has multiple copies within the genome The X axis represents the range of observed MITE sequences inserted into intron of a gene. The Y axis represents how many genes had a certain number of MITE sequences inserted in its introns.

pairwise divergence values. I predicted that as older MITEs became degraded from being stationary in the genome, they would be more likely to become integrated into the genome and take on function. As MITEs became integrated or took on some function they would have been more likely to be inserted in introns, because their presence in introns would be less likely to disrupt gene function than the presence of actively moving MITEs. Younger MITEs are more likely to be actively moving within the genome and have lower pairwise divergence values because they just recently inserted into the genome. Therefore, I would have predicted younger MITEs would be more likely to be retained in intergenic regions (Barron et al, 2014) where they would have less influence on gene function. Therefore, my prediction would have been a positive correlation between X and Y. I used a Pearson's correlation to determine whether this prediction was met but did not detect a significant correlation ($R = -0.2717$, $p=0.027742$) (Figure 6). This p value is not significant at 0.01 confidence level, showing that there is not a statistically significant correlation to this data. This analysis outcome implies that MITEs are just as likely to be inserted into introns (vs intergenic regions) regardless of time since the original insertion.

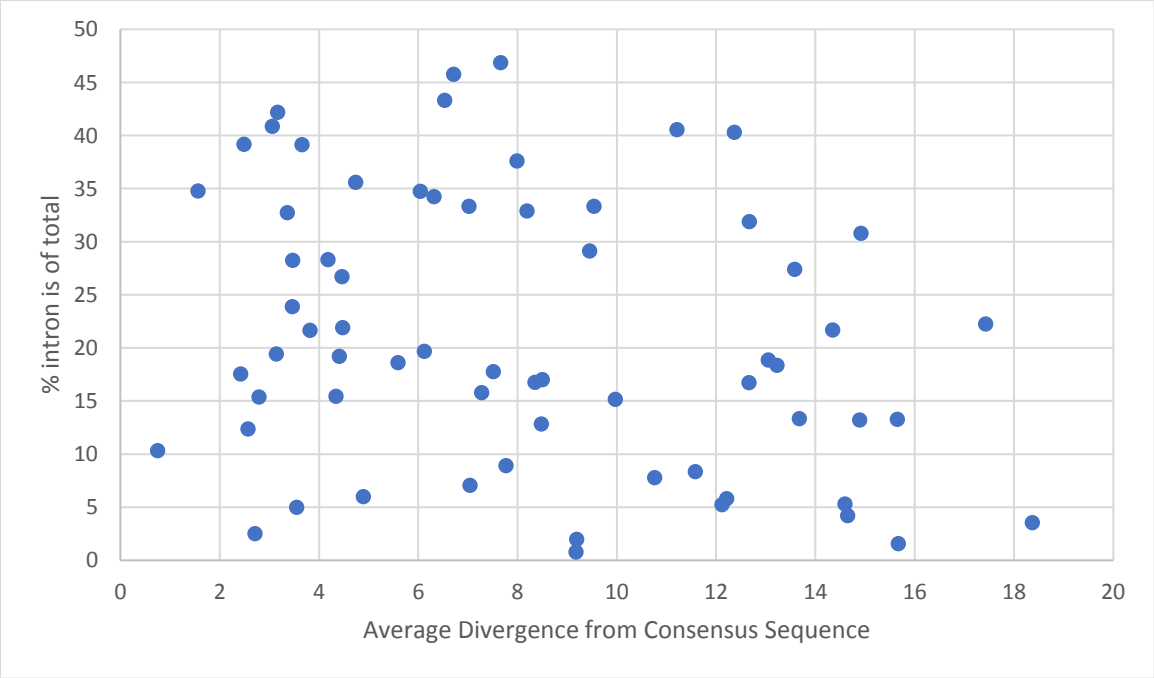


Figure 4: Age of MITE family does not influence likelihood of retention into introns Younger families of MITEs on left of graph with older MITEs on right of graph
The X axis represents the average Kimura 2-parameter divergence of the MITE sequences from the consensus sequence of that MITE family. The Y axis represents the percent of all MITE sequence insertions that are in introns for the same MITE family.

DISCUSSION

I identified 1 autonomous element of a MITE family. An autonomous element typically undergoes an internal deletion, which gives right to a MITE. Shortly after this, MITEs often go through amplification bursts and accumulate copies quickly within the genome. Despite this, the autonomous element within the genome remain at low copy numbers, and now the transposase of this autonomous element is responsible for moving the all the copies of this MITE within the genome. It is a paradox that these MITEs can accumulate to such high copy numbers within the genome, and occupy the transposase of their parental autonomous element, and still the parental autonomous element can persist in the genome with MITEs so successfully parasitizing its transposon activity. By finding this autonomous *PiggyBac* element within the genome and assessing the intergenomic relationships between this PiggyBac element and MITE 11_11380, the *P. antipodarum* genome can start to be used to assess the MITE paradox.

Characterizing the MITEs within the *P. antipodarum* genome is also an important first step in order to do more analysis later, in this case more focused on TE mediated genome evolution of the *P. antipodarum* genome. I manually identified 21 MITE families, most of which belonged to the *PiggyBac* or *Maverick* superfamilies. This result implies that these superfamilies may have recently undergone an amplification burst within the *P. antipodarum* genome. I also determined the genomic locations of these MITE families and found that while most (79.62%) MITE insertions are in intergenic regions, 20.36% of all MITE sequences are present in the introns of genes. These results suggest that MITEs in the *P. antipodarum* genome may affect gene expression by being inserted in introns. Lu et al. (2012) found that when MITEs are inserted near genes they reduce the expression of those genes. However, Guo et al (2017) found that nearby MITE insertions had no effect on genes. The effects of intronic MITE insertions on gene

expression is one area that warrants further study in the *P. antipodarum* genome. that MITEs may not only affect genome evolution but also gene sequence, expression, and evolution. Only a very small fraction - 0.02%- of all MITE insertions were in the exons of genes. My blast searches did not reveal any clear trend regarding gene identity for the genes harboring exonic MITE insertions. Of all the 32,148 predicted genes in *P. antipodarum's* genome, 37.74% of these genes (12,131 genes) had MITE sequences inserted into their introns. Of these 12,132 that had intronic MITE sequence insertions, 53.18% of these genes (N = 6,949) only had 1 MITE sequence insertion. On the other hand, there were 12,096 genes that had MITE families inserted into their introns. Of these genes, 57.45% (N=6,949 genes) only had 1 MITE family inserted in its introns. The results of these two groups are similar, though one gene (maker-6-snap-gene-20.17-mRNA-1) had 14 MITE families inserted in its introns while it had 21 MITE sequences in its introns (demonstrating that at least some of these insertions are copies of the same MITE family with the same sequence). There are several caveats to these results, however. The genome assembly of *P. antipodarum* is fragmented, so the view of intronic and intergenic MITE insertions may be inaccurate. MITE sequences that are predicted to be intronic may be intergenic if a predicted gene is actually two genes, and MITE sequences that are predicted to be intergenic may actually be intronic if a gene gets split into two different contigs in the genome assembly. Predicted genes may also not be actual genes, though it does appear that many of the predicted genes with MITEs inserted in them have blast results. However, gene models have not be carefully inspected for completion, which also may undermine the results of this research.

These limitations of the genome assembly may also have affected my search for parental autonomous elements. While I was able to find a *PiggBac* element in the genome that was a likely parental element of MITE family 11_11380, this was the only parental autonomous

element I was able to identify. A better assembled genome may have helped with this search, but there are other limitations to searches for autonomous parental elements of MITEs. DNA transposons can also cause the formation of heterochromatin, which may make finding autonomous parental elements even harder (Feschotte & Pritham, 2007).

The presence of autonomous parental elements in the genome would lead to the continued proliferation of MITEs. If MITEs are still moving within the genome, it would be logical that they would more closely resemble the consensus sequence (meaning a lower divergence value). However, the divergence values of MITEs did not seem to correlate with the likelihood that a MITE sequence would be retained in an intron (vs an intergenic region). Those MITEs that are inserted into introns are more likely to have effects on gene expression and evolution. Therefore, it would seem more likely that MITEs that had been inserted into the genome for longer (and become more integrated into the genome) would be in introns. It would be detrimental to the organism to have gene expression levels changing with the insertions and excisions of MITEs that are still active in the genome and not mutated to the point where they can no longer move within the genome. My research did not support this prediction however, because of the lack of significant correlation between divergence values and retention of MITE families in introns.

CONCLUSION

P. antipodarum is an ideal model for the study of MITEs. The ongoing annotation of *P. antipodarum* genome has shown that at least a quarter of the genome is derived from TEs. DNA transposons make up over half of the total amount of TEs in this genome, with MITEs composing over half of DNA transposon content. Therefore, MITEs make up a sizable portion of *P. antipodarum* genome, having at least 21 distinct MITE families in it. A third of *P. antipodarum's* predicted genes have at least one MITE insertion, with 20.38% of all MITE sequence insertions being in genes. All of this may support that MITEs can influence gene expression. When MITEs insert into gene introns they can have regulatory influences on that gene's expression by helping to determine intron length.

Though I was able to find MITEs within the *P. antipodarum* genome, characterizing autonomous elements within the genome was more difficult: I was only able to find and verify one such element. *P. antipodarum* has a large quantity of TEs, and a sizable number of MITEs. Feschotte and Pritham (2007) found that genomes with high numbers of TEs also tended to have large quantities of nonautonomous elements. This is a kind of paradox because it would be expected that the activity of nonautonomous elements would be detrimental to their autonomous parental elements. This is expected because MITEs occurring in high copy numbers would occupy the transposase activity of that autonomous element and possibly trigger host repression mechanisms of both the autonomous and nonautonomous elements. Therefore, it would be expected that increased MITE copy number would repress the activity of the autonomous parental element. However, the findings of my research supported the conclusions of Feschotte and Pritham (2007) that genomes with large numbers of TEs also had high quantities of nonautonomous elements. My characterizations of MITEs within the P. antipodarum genome

may help to explain this MITE paradox. One of the key features that contribute to this MITE paradox is that MITEs occur in the genome in such high copy numbers.

These high copy numbers are one of the reasons why MITEs are so interesting and warrant further study. A MITE's tendency to accumulate in high numbers not only produce the MITE paradox, but also means MITEs can insert into the introns of many different genes and have many chances to influence gene expression in an organism. Changes in expression of multiple genes can eventually lead to large scale genome evolution due to MITE activity. Identifying and characterizing the MITEs present within a genome is the first step to assessing the mechanisms, maintenance and effects of these high MITE copy numbers.

LITERATURE CITED

Barron, M., Fiston-Lavier, A., Petrov, D., & Gonzalez, J. (2014). Population genomics of transposable elements in Drosophila. *The Annual Review of Genetics,48*, 561-581.

Feschotte, C., & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, *41*(1), 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448

Guo, C., Spinelli, M., Ye, C., Li, Q. Q., & Liang, C. (2017). Genome-Wide Comparative Analysis of Miniature Inverted Repeat Transposable Elements in 19 Arabidopsis thaliana Ecotype Accessions. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-02855-1

Han, Y., & Wessler, S. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*. doi:10.1093/nar/gkq862

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., & Quesneville, H. (2014). PASTEC: An automatic transposable element classification tool. *PLoS ONE*, *9*(5). https://doi.org/10.1371/journal.pone.0091929

Kumar S, Stecher G, and Tamura K (2016) Molecular Biology and Evolution 33:1870-1874

Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W., & Kuang, H. (2012). Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in oryza sativa. *Molecular Biology and Evolution*, *29*(3), 1005–1017. https://doi.org/10.1093/molbev/msr282

Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue): W327-31.

Robillard, É., Le Rouzic, A., Zhang, Z., Capy, P., & Hua-Van, A. (2016). Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proceedings of the National Academy of Sciences*, *113*(51), 14763–14768. https://doi.org/10.1073/pnas.1524143113

S. Götz et al. "High-throughput function annotation and data mining with the Blast2GO suite", Nucleic Acids Research, Vol. 36, June, 2008, pp. 3420-3435.

Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015 <http://www.repeatmasker.org>.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., … Schulman, A. H. (2007, December). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2165

Yang, G., & Hall, T. C. (2003). MAK, a computational tool kit for automated MITE analysis. *Nucleic Acids Research*,*31*(13), 3659-3665. doi:10.1093/nar/gkg531

Yang, G., Fattash, I., Lee, C. N., Liu, K., & Cavinder, B. (2013). Birth of three stowaway-like MITE families via microhomology-mediated miniaturization of a Tc1/Mariner element in the yellow fever mosquito. *Genome Biology and Evolution*, *5*(10), 1937–1948. https://doi.org/10.1093/gbe/evt146