

Elucidating Transposable Elements Survival  
Mechanisms: Invasion and Evasion by *Tc1/mariner*  
Superfamily DNA Transposons

-Analysis of plant *Tc1*-like elements horizontal transfers and rice  
*Mariner*-like elements autoregulation

By

Yuan Liu

A thesis submitted in conformity with the requirements for the degree of  
Master of Science

Department of Cell and Systems Biology

University of Toronto

©Copyright by Yuan Liu 2014

# Elucidating Transposable Elements Survival Mechanisms: Invasion and Evasion by *Tc1/mariner* Superfamily DNA Transposons

Yuan Liu

Master of Science

Department of Cell and Systems Biology  
University of Toronto

2014

## **Abstract**

Transposable elements are mobile DNA elements residing in almost all sequenced genomes. However, the kingdom Plantae has not previously been reported to harbor *Tc1*-like elements (TLEs). This study reports the characterization of several plant TLEs. It was proposed that the wide distribution of transposons may be partly due to horizontal transmissions, and some cases of highly similar TLEs from distantly related species were found in this study. Once a transposon invades a new genome, it may freely amplify. Being mutagenic in nature, they are under constant scrutiny of host inhibitory mechanisms that often keep transposition activity at a basal level. A potential evasive mechanism by a rice *Mariner*-like element is described here to auto-regulate activity via a subterminal transposase binding motif.

## **Acknowledgements**

I would like to express my thanks to my supervisory committee members, Dr. Guojun Yang, Dr. Tim Westwood, and Dr. David McMillen for their guidance and feedback during my graduate studies, as well as their patience with me.

I would also like to thank Dr. Yang for allowing me the opportunity to join his research team, along with the postdoctoral fellows Dr. Isam Fattash and Dr. Chiani Lee. I express my appreciation to Dr. Ichiro Inamoto, and the Espie Research Group for their assistance and permission to use their equipment.

Furthermore, Julia Pelz and Christy Simbeya, as well as many other fellow graduate students have helped made the graduate studies an experience of a life time. And last but not the least, I thank the support from outside of the University, from my parents, Charlie and Andrew Wang, Natalie Swieton, and my friends.

Funding was provided by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (Canada) (RGPIN 371565 to GY, RGPIN-2014-04709), Canadian Foundation for Innovation (CFI24456 and IOF-12 to GY), Ontario Research Foundation (ORF24456 to GY), and the University of Toronto.

## Acknowledgements of Published Material

Chapter II of this thesis contains published material regarding discovery and characterization of plant *Tc1*-like elements. The published article “*Tc1*-like transposable elements in plant genomes” can be found in <Mobile DNA>, 2014, 5:17 [1]. My Principle Investigator, Dr. Guojun Yang, was the other author of the article. Dr. Yang conceived the original study, and participated in the analysis, confirmation of my results, and drafting and revision of the manuscript. I wish to express my appreciation for his contribution to the published material.

For this thesis, the scope of the thesis study was expanded to report potential cases of horizontal transposon transfer events between plant species and some non-plant organisms. Some major additions were made to methods and results, including several plant elements for analysis and the construction of a phylogenetic tree to highlight plant *Tc1*-like elements distribution. Most of the bioinformatics analysis described here has been performed by me and the work done by Dr. Yang in the published material that is mentioned here including the analysis of *PpActin1* in the moss genome, analysis of conceptual *PpTc* transposase structures, creating Figure 5 and Figure 11, as well as confirmation of my results and editing of the texts.

# Table of contents

Acknowledgements .....	iii
Acknowledgements of Published Material.....	iv
Glossary .....	viii
List of Figures.....	ix
List of Tables .....	xi
Appendix Table List .....	xii
Chapter I: Introduction .....	1
I-1 Transposable elements.....	1
I-2 Classification .....	3
I-2.1 Retrotransposons.....	4
I-2.2 DNA transposons .....	5
I-3 Autonomy .....	8
I-4 <i>Tc1</i> /Mariner superfamily .....	10
I-5 Thesis objectives: elucidating TE survival mechanisms .....	10
Chapter II: Plant <i>Tc1</i> -like elements and their potential for horizontal transfer .....	12

II-1 Background.....	12
II-1.1 <i>Tc1</i> -like elements.....	12
II-1.2 Horizontal Transfer of Transposable Elements.....	13
II-2 Methods .....	16
II-2.1 Identification and retrieval of moss <i>Tc1</i> -like elements .....	16
II-2.2 Characterization of moss TLEs .....	16
II-2.3 Identification and analysis of TLEs in other genome databases	17
II-3 Results .....	18
II-3.1 <i>Tc1</i> -like elements in moss .....	18
II-3.2 Life history of moss TLEs .....	21
II-3.3 Expression of <i>PpTc1</i> in moss .....	22
II-3.4 <i>Tc1</i> -like elements in higher plant genomes .....	24
II-3.5 Cross-kingdom horizontal transfer of plant TLEs.....	29
II-3.6 <i>Tc1</i> -like MITEs in plant genomes .....	31
II-4 Discussion.....	32
II-4.1 Moss TLE elements and activity .....	32
II-4.2 Horizontal transposon transfer of plant TLEs .....	34

II-4.3 Life cycle of TE and HTT .....	36
Chapter III: Autoregulation of <i>Osmar</i> transposition .....	37
III-1 Background .....	37
III-1.1 Host-TE relationship .....	37
III-1.2 <i>Mariner</i> -like elements .....	40
III-2 Methods .....	42
III-2.1 Cloning .....	42
III-2.2 Yeast Excision Assay .....	44
III-2.3 Electrophoretic mobility shift assay (EMSA) .....	46
III-3 Results .....	48
III-3.1 Mutations in the 3' subterminal binding motif .....	48
III-3.2 Transposase binding affinity of 3' subterminal motifs .....	50
III-3.3 Effects of linker sequences between the 3' subterminal motifs .....	52
III-4 Discussion .....	57
Chapter IV: Concluding Remarks .....	60
References.....	61
Appendix.....	71

## Glossary

EMSA: Electrophoretic mobility shift assay

MLE: *Mariner*-like elements

ORF: Open reading frame

RTE: Relative transposition efficiency

TE: Transposable element

TIR: Terminal inverted repeat

TLE: *Tc1*-like elements

Tp: Transposase

TSD: Target site duplication

## List of Figures

Figure 1 Simplified illustration of a copy and paste retrotransposition. ....	4
Figure 2 Simplified illustration of a cut and paste DNA transposon. ....	6
Figure 3 Levels of autonomy.....	8
Figure 4 Summary of major structure componenets of moss TLEs, <i>PpTc1</i> and <i>PpTc2</i> . ....	18
Figure 5 Comparison of the putative transposases of <i>PpTc1</i> and <i>PpTc2</i> .....	20
Figure 6 Divergence chart for full-length copies of <i>PpTc1</i> and <i>PpTc2</i> . ....	21
Figure 7 Members with full length ORFs.....	23
Figure 8 Sequences of complete TIRs of plant TLEs compare to <i>Tc1</i> .....	25
Figure 9 Plant D34E motifs alignment.....	26
Figure 10 Phylogenetic relationship between plant <i>Tc1</i> -like elements.....	28
Figure 11 Horizontal transfers of plant TLEs. ....	30
Figure 12 Yeast excision assay.....	44
Figure 13 Oligo designs for EMSA experiments. ....	47
Figure 14 Comparison of excision frequency by <i>Osmar14</i> transposase in yeast between <i>Osmar14</i> NAS constructs.....	49

Figure 15 Comparison of excision frequency by <i>Osmar5</i> transposase in yeast between <i>Osmar5</i> NAS constructs.....	49
Figure 16 Gel image of EMSA experiment.....	51
Figure 17 Nucleotide sequences of the 3' end of different <i>Osmar</i> elements. .....	52
Figure 18 Comparison of excision frequency by <i>Osmar14</i> transposase in yeast between <i>Osmar14</i> NAS shift constructs. ....	54
Figure 19 Comparison of excision frequency by <i>Osmar5</i> transposase in yeast between <i>Osmar5</i> NAS and +5shift construct.....	55
Figure 20 Illustration of the effect of the 5bp shift between <i>Osmar5</i> and <i>Osmar14</i> .. ....	56

## List of Tables

Table 1 .....	43
---------------	----

## Appendix Table List

Table A1 Assembled transcripts of <i>PpTc1</i> from P0409 database .....	71
Table A2 <i>PpTc1</i> assembled transcripts that produce a conceptual full-length transposase bearing DD34E motifs. ....	73
Table A3 Full-length plant TLEs with intact TIRs .....	74
Table A4 <i>Tc1</i> -like transposases described in this study.....	75
Table A5 Oligo sequences for Osmar14 EMSA experiments.....	77

# **Chapter I: Introduction**

## **I-1 Transposable elements**

Genomes are dynamic. A major component of a dynamic genome is a plethora of mobile genomic elements collectively known as transposable elements (TEs). TEs are discrete pieces of DNA that are capable of mobilizing themselves or copies of themselves to a different location in the host genome. TEs may duplicated themselves in the process, resulting in increased copy numbers over time. Existence of TEs is nearly ubiquitous in all surveyed organisms with some minor exceptions of a few primal life forms [2-4].

These elements were part of our genomic ecosystem since the infancy of cellular life and successfully colonized most life on Earth. However, the importance and significance of their universality was not always immediately obvious to us. TEs were given many unfavourable names in the past at our different stages of understanding of their significance. They were called “jumping gene”, “selfish” DNA, or even “junk” DNA due to their lack of apparent function in a genome.

In the 1950s, at the beginning of her work with TEs, McClintock discovered Activator/Dissociator in maize [5, 6]. These elements were called “controlling elements” by McClintock. They contributed to the pigmentation in maize kernels, and they were capable of changing their locations between chromosomes. This discovery along with a similar demonstration of

bacterial mobile DNA implicated capability of chromosomal DNA to mobilize [7]. Mendel's genetics was widely accepted at the time and genes and genomes were thought to be static entities insusceptible to change. McClintock's discovery was thusly not well received for its implication of dynamic genomes and stochastic inheritable genetic material. This negative notion continued to be the attitude towards TEs.

The significance of TEs was later acknowledged by the discovery that *P* and *I* elements are, in fact, transposable elements [8, 9]. These fruit fly elements are responsible for the phenomenon known as hybrid dysgenesis [10-12], which involved genetic modifications that may lead to sterility. *P* elements were studied extensively in the early age of fly genetics, and they were even used as vectors to transform fruit flies [13]. These findings contradict the role of TEs as mere junk sequences that clutter genomes. The understanding that *P* elements were actually DNA transposons underscores the importance of TEs in scientific research and our complete understanding of genomic dynamics.

In 1983, Barbara McClintock was awarded a Nobel Prize in Physiology or Medicine for her discovery of mobile genetic elements in maize. This recognition marked the appreciation of the importance of transposable element in the scientific community, and the expansion of our knowledge on TEs over the decades.

With the advancement of technology and the advent of whole-genome sequencing, it is now apparent to us that TEs are not only important, but also abundant and diverse. In some eukaryotic genomes such as maize, about

85% of the genomic sequence is consisted of TEs [14]. Furthermore, it was found that nearly 55% of the human genome are TEs [15]. With the number of identified TEs accumulating rapidly, biologists classified them based on their transposition mechanism and structural properties.

## **I-2 Classification**

TEs vary greatly in number, size and structure. For example, *L1* elements in the human genome are thousands of base pairs long with two open reading frames that code for enzymes for transposition, while *Alu* elements are only hundreds of base pairs long with no coding sequences. To unify the naming and characterization system for TE discovery and research, universal classification systems for TEs were proposed [16, 17]. The classification systems focus on the characteristics that contribute to TEs mobilization, such as insertion sites, terminal sequences, TE sequence organization/structure, coding sequences and some distinctive sequences markers within the TEs. Subsequently, TEs were divided into classes, families, and other Linnaean hierarchies. Although the exact naming and classification system for TEs is a matter of debate, differences between major groups of TEs are well understood and a general classification system is widely used by the research community.

There are two classes of TEs that are distinguished by their methods of transposition and duplication. Class I TEs are retrotransposons that use a RNA intermediate and “copy-and-paste” transposition mechanism; and

Class II TEs are DNA transposons, and they utilize a DNA intermediate and “cut-and-paste” transposition system [4, 17-19].

### I-2.1 Retrotransposons

Retrotransposons use a “copy and paste” mechanism to mobilize (Figure 1). The genomic sequence of a retrotransposon is transcribed into a RNA intermediate, which is then reverse-transcribed into DNA sequence by a reverse transcriptase that is encoded by the internal sequences of retrotransposons. The new DNA copy can then be incorporated into the host

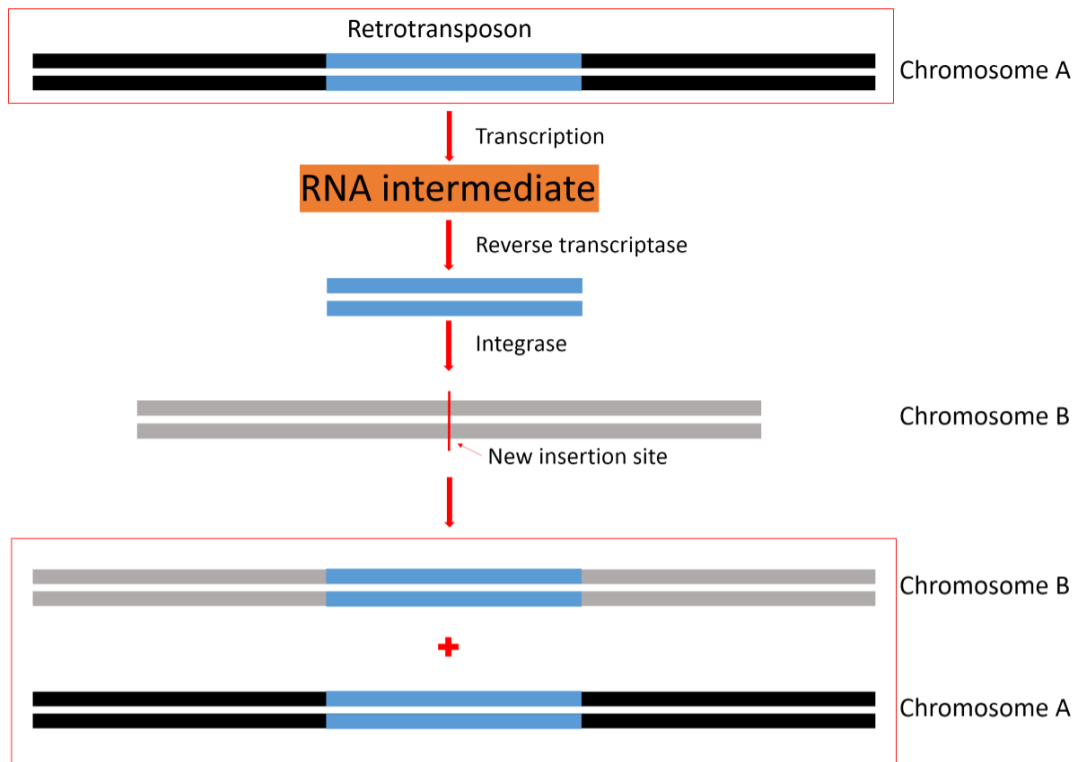


Figure 1 Simplified illustration of a copy and paste retrotransposition. The retrotransposon (blue) is transcribed into a RNA intermediate (orange) that is reverse transcribed and integrated into a new insertion site. The donor retrotransposon is retained, and an additional copy is produced in the process. Chromosomal DNA (long black and grey bars) arbitrarily represent two different loci.

genome by an integrase, which may also be encoded by the retrotransposon coding sequences. In the process, an additional copy of the original TE was created during each cycle of copy-and-paste resulting in an increase of TE copy numbers and expansion of the host genomic size. Members of retrotransposons include long-terminal repeat retrotransposons (LTRs), *DIRS*-like elements, *Penelope*-like elements, long interspersed elements (LINEs), and short interspersed elements (SINEs) [16, 20]. Notably, retroviruses are closely related to retrotransposons [21-23]. When a retrovirus loses its viral envelope and its ability to mobilize cellularly, it is reduced to what is called an endogenous retrovirus (ERV), which is regarded as a member of the retrotransposons. Retrotransposons are abundant, one extreme example is the *Alu* elements that is estimated to have 300,000 to over 1 million copies in the human genome occupying over 10% of the genomic content [24, 25]

### **I-2.2 DNA transposons**

The other major group of TEs are DNA transposons (Figure 2). The majority of DNA transposons adopted the “cut and paste” mechanism to carry out transposition. There are no RNA intermediates during the process and the TE genomic sequence is directly excised out of the chromosome and relocated into a different site. These elements encode a single gene called a transposase to carry out all transposition reactions. Transposases are specialized integrases and there are variations in the catalytic domain of

transposases from different families of DNA transposons [3]. Members of the DNA transposons include *Tc1/Mariner*, *hAT*, *Mutator*, *Merlin*, *Transib*, *P*, *PiggyBac*, *PIF-Harbinger*, *CACTA*, *Helitron* and *Maverick* [16, 18, 19].

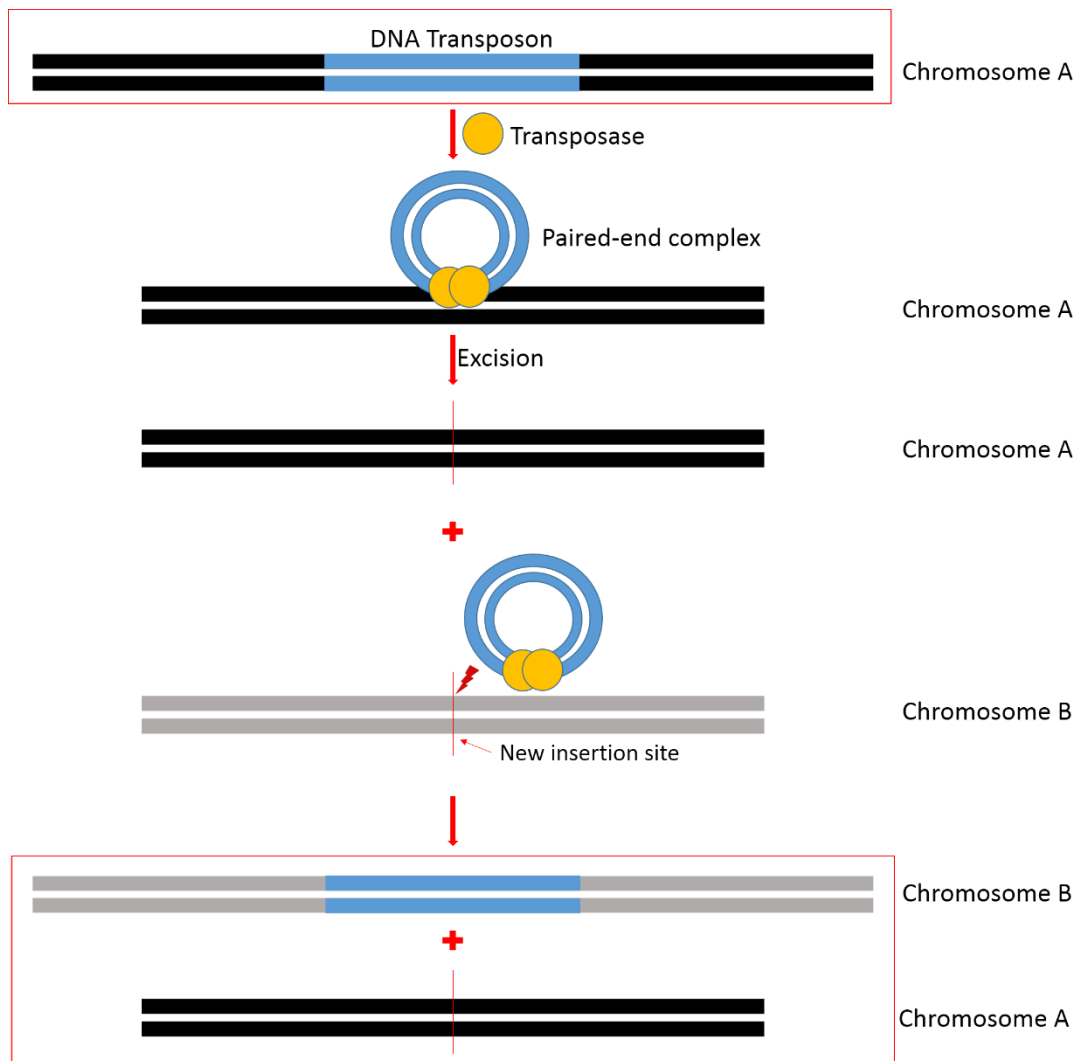


Figure 2 Simplified illustration of a cut and paste DNA transposon. The DNA transposon (blue) is cleaved out the donor chromosome and inserted into a new site by transposases (yellow solid circles). Transposases form synapses with the DNA to facilitate reactions (not shown). No duplicate of the original copy was produced in this cycle. Chromosomal DNA are shown as long black and grey bars, each representing a different locus.

Although it should be noted that *Helitron* and *Maverick* do not utilize the cut-and-paste transposition system.

The majority of DNA transposons share these three key features required for cut-and-paste transposition: terminal inverted repeats, target site and transposase [26]. Terminal inverted repeats (TIRs) are the terminal sequences of a transposon, flanking the internal coding sequences. TIRs are recognition sites for transposases to bind, two *MosI* transposases bind to the TIRs and form a paired-end complex to facilitate transposition reactions [27-29]. The transposases catalyze the excision reaction at the termini of the TIRs and cleave the double-stranded DNA transposon out of the donor site. The transposases subsequently form a synapse with the specific target sequence of a new site, and catalyze the integration reaction into chromosomal sequence [29, 30].

Cut-and-paste DNA transposons do not generate additional copies during each cycle of transposition. However, they can still achieve duplication through host DNA replication and repair systems [26]. A DNA transposon can transpose during host replication processes from a replicated location on the chromosome to an unreplicated site; alternatively, the TE can be restored via gap repair by homologous recombination [26]. In general, DNA transposons are much more modest in copy numbers when compared to retrotransposon. Due to a less efficient method for duplication, only 1.6% of the human genome are DNA transposons while over 40% of the genome are retrotransposons [24, 25]. There are, however, many cases of DNA transposons that excel in a genome to achieve many copies, such as small

fragmented elements called miniature inverted repeats (MITEs) in some plant genomes [31].

### I-3 Autonomy

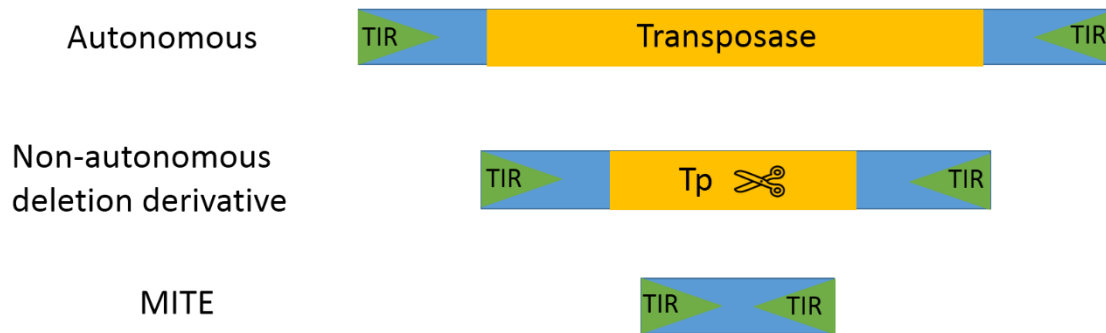


Figure 3 Levels of autonomy. Simplified illustration of the structural differences between autonomous, non-autonomous, and MITE elements. DNA transposons were used in this illustration. Element is blue; Transposase coding sequence is yellow; TIRs are green; Scissors indicate a deletion derivative of the transposase gene.

Most elements observed today are relics of past activities and they are presently immobilized. Full-length autonomous elements can rely on self-encoded specialized enzymes to transpose and sustain copy numbers in the host genome. However, most elements that can be found today have lost their capability to encode for an active protein (Figure 3). These elements must utilize the enzymes encoded by the autonomous copies to carry out transposition reactions, and we call them non-autonomous elements [26, 32].

In some cases, a non-autonomous element maintains the full length and has high sequence similarity to the autonomous copies, but has accumulated many point mutations and indel mutations. The coding sequences of these

elements have become defective and translate non-functional proteins or may not even encode for a long polypeptide at all. Non-autonomous elements also include deletion-derivatives of autonomous elements. In these cases, they only maintain high sequence similarities with autonomous copies in the flanking regions (terminal sequences of TE), and the coding sequences are truncated or deleted. They are shorter in length compared to the autonomous elements and full-length non-autonomous elements, and result in variable sizes of non-autonomous versions. In the most extreme scenario, the elements have lost virtually all internal coding sequences, and only terminal flanking sequences that are crucial for transposition reaction remain. These elements are called miniature inverted-repeats transposable elements (MITEs).

Non-autonomous elements may still maintain activity through utilizing enzymes encoded by autonomous elements, since only the terminal sequences are required for recruiting dedicated enzymes to carry out transposition reactions. MITEs are especially successful at utilizing transposases of related autonomous elements to amplify to high copy numbers, making them the ultimate non-autonomous transposons [33-35]. MITEs were originally discovered in plants, and they were classified into two families [36]. Based on sequence similarities, *Tourist* MITEs are related to the *PIF/Harbinger* superfamily, and *Stowaway* MITEs belong with the *Tc1/Mariner* superfamily [37].

## **I-4 *Tc1*/Mariner superfamily**

One of the largest group of DNA transposons is the *Tc1*/*Mariner* superfamily, it is nearly ubiquitous in eukaryotic genomes [38]. The *Tc1*/*mariner* superfamily was first identified in nematode and insect genomes [39]. The superfamily was named after *Tc1* in *Caenorhabditis elegans* [40], and *Mariner* in *Drosophila mauritiana* [41]. This superfamily is characterized by two TIRs of typically 12-28 nt flanked by dinucleotide target site duplications (TSDs) of “TA”. The transposases of this superfamily contain a triad catalytic motif consisting of either two aspartic acid residues and a glutamate residue (DDE) in *Tc1*-like elements (TLEs); or three aspartic acids (DDD) in *Mariner*-like elements (MLEs). These two types of domains are collectively known as the DDE/D motif. The pocket formed by the triad residues contains the metal ions required in the DNA cleavage reaction during transposition [29]. Based on the number of residues between the second and third catalytic residues of the DDE/D motif, *Tc1*/*mariner* transposases can be characterized as DD34E, DD34D, DD31-33D, DD35E, DD37D, DD37E, or DD39D, and each type of transposase domains designates a sub-group of the *Tc1*/*mariner* superfamily [42-44].

## **I-5 Thesis objectives: elucidating TE survival mechanisms**

The lifecycle of TEs is initiated when a founding copy enjoys a stage of proliferation when it is first derived or transmitted. This family of TEs then becomes inactivated through accumulated mutation or host inhibitory

mechanism until activities of all copies in the family are lost in the host and its offspring. Survival of a lineage of a TE may be achieved by horizontal transfer into an uncolonized host or adoption of mechanisms that minimize its own activity to avoid deleterious effects to the host and thus evade the scrutiny of host defense mechanisms [25, 45, 46].

This thesis aims to elucidate the properties of TEs that allow their ubiquitous and resilient presence in eukaryotic genomes. This study, consisting of two projects, will explore the survival mechanisms of TEs: 1) through horizontal transfers for TEs to invade into other genomes, and 2) through autoregulation to evade host inhibitory mechanisms and persist in a genome.

Focusing on the *Tc1/mariner* superfamily, the thesis projects will utilize bioinformatical tools to screen available genomic database and characterize elements to find new potential active elements to expand our understanding of TE mobility, and experimentally analyze a subterminal binding motif for an active transposase to study the molecular mechanisms of autoregulation.

# Chapter II: Plant *Tc1*-like elements and their potential for horizontal transfer

## II-1 Background

### II-1.1 *Tc1*-like elements

*Tc1* was first characterized in *Caenorhabditis elegans* genome. It has a DD34E catalytic domain in its transposase, and has distinctive terminal sequences of 5' CAGT...GTCA-3', flanked by TSD of dinucleotide "TA". Subsequent reports found that *Tc1*-like elements are widespread, and especially abundant in teleost fish [47]. Some of elements of this group were demonstrated to be active, including *Tc1* and *Tc3* that naturally exist in *C.elegans* [40, 48, 49], *Minos* from *Drosophila hydei* [50], and *Impala* from *Fusarium oxysporum* [51, 52]. Notably, *Sleeping Beauty* was an artificially constructed element derived from dormant fish TLEs [53], and was designed as a genetic tool for gene delivery in mammalian genomes. Subsequent reports after the discovery of *Tc1* show that TLEs are widespread in eukaryotes [19]. However, although its sister group, MLEs, are widespread in plant species, TLEs have not yet been reported in a plant genome.

In this study, TLEs in plant genomes are reported. Two prevalent elements in the moss genome are characterized and analyzed, leading to the discovery and characterization of patchily distributed TLEs throughout the plant kingdom. Moreover, these elements were analyzed as potential candidates for episodes of horizontal transposon transfer.

## II-1.2 Horizontal Transfer of Transposable Elements

TEs are vertically transmitted to the genome of the host offspring. TEs persist symbiotically within the same lineage and continue to proliferate and evolve. Given its nature to mobilize and insert within the host genome, it was postulated that transposable elements are also capable of horizontally spreading into other genomes [45, 46, 54-56]. Many convincing cases were reported since the 1990s on the issue of horizontal transfer of transposable elements, notably the *P* elements in fruit flies [57]. Characteristic signs that suggest a horizontal transposon transfer (HTT) has occurred are: 1) display of patchy distribution; 2) highly similar sequences that is unexpected for divergent host species; and 3) discrepancy between TE and host lineages. With the arrival of the genomic era and extensive sequencing projects, screening for highly similar sequences between divergent species and tracing patchy distribution has become an straightforward task for bioinformaticians. Many cases of HTT were subsequently proposed based on genomic data, especially in fruit flies [45]. These pieces of evidence support the capability of TEs to not only vertically transmit, but to horizontally invade other genomes as well.

The exact mechanism for TEs to transfer horizontally is still unknown, despite their logical capability demonstrated by transposition. Some of the proposed routes of transfer include feeding, pathogens and parasites [45]. Through these channels, vectors such as bacteria and viruses are the logical culprit for HTT, given their tendency to invade eukaryotic cells and exchange genetic material. While not frequently observed, some examples

demonstrate the capability of vectors to intake host TEs. A retrotransposon was found to integrate into a poxvirus from a snake host [58], and insect TEs were identified in a baculovirus [59]. Another notable candidate is the *Wolbachia* genus of bacteria. They are endocelelluar parasites and are capable of exchanging genetic material with the host [60-62].

Additionally, a highly probable piece of evidence is the HTT of a *Mariner*-like element between a parasitoid wasp and its lepidopteran hosts [63]. This transfer may have involved a virus that is symbiotic to the wasp, and it produces viral pockets that are injected along with the eggs into the host. Any TE that is packaged into the viral particle is then exposed to the host cell. Some TE-derived fragments of DNA have been found in these kinds of viruses and support their role in the HTT between the parasitoid wasp and its hosts.

HTT may result in significant consequences in molecular evolution. Once a TE horizontally transmit into a distant host, it is likely unhindered by host defensive mechanisms that otherwise inhibit native transposons [45]. The TEs may freely amplify in the new host genome until novel inhibitory mechanisms are developed. TE proliferation contributes directly to the genome size and re-arrangement of chromosomal material [25, 64-67]. Moreover, TEs can be domesticated by a host genome and contribute to a novel functional genetic sequence [68-71]. It is important to note that TEs also often capture host genes and transpose captured genes as its internal sequence. Although it has not been observed previously, a HTT event that carries a piece of functional genetic material may result in an interesting

route of horizontal gene transfer between eukaryotic species. These theories highlight the importance for increased knowledge in HTT mechanisms. It is crucial to fully comprehend the capability of TEs to facilitate exchange of genetic material between organisms to construct a complete evolutionary history of TEs and perhaps all genomes.

HTT has been reported for both DNA transposons and retrotransposons. However, the most distant transfers have been exclusively DNA transposons, namely the *Tc1/mariner* and *hAT* superfamilies [45]. Indeed, *Tc1/mariner* and *hAT* superfamilies of DNA transposons are capable of incorporation into a highly divergent host when introduced by experiments [27, 39, 57, 72-77]. Proposed events of HTT of these superfamilies can cross species of different phyla and even kingdoms.

Much of the mechanisms underlying eukaryotic HTT is still mysterious. The most extensively studied cases of HTT focused on fruit flies, which heavily dominated our knowledge of eukaryotic HTT [45]. It is important to further our scope and understanding beyond a small group of organisms, as HTT is an important aspect that may play a major role in the observed diversity and abundance of eukaryotic TEs. In this study, several pieces of evidence have been found to support potential episodes of horizontal transfer of DNA transposons of the *Mariner* family between plants, bacteria, fungi and insects. These discoveries shed light into the process of horizontal genomic invasion by TEs.

## II-2 Methods

### II-2.1 Identification and retrieval of moss *Tc1*-like elements

To identify transposons that belong to the *Tc1*-subsuperfamily, the *Tc1* transposase peptide sequence was used as the query sequence to search against GenBank databases of *P. patens* genome with the default parameters. Each returned sequence was retrieved and inspected for TIRs with flanking TSD that end with 5'TA/CAGT...ACTG/TA-3', as well as internal coding sequences that bear a DD34E domain (both TSD and TIRs are defining properties of TLEs). Complete elements were searched against its host genome to obtain all of the members in its family. Nucleotide sequences of full-length TLE copies were retrieved with MITE Analysis Kit (MAK) function MEMBER (<http://labs.csb.utoronto.ca/yang/MAK/>) [78]. Members of each family were retrieved with MAK with zero tolerance for end sequence mismatches, as the terminal sequences are conserved for TLEs.

### II-2.2 Characterization of moss TLEs

Alignments of all retrieved members in each *P. patens* TLE (abbrev. *PpTc*) family were completed with CLUSTAL to generate a consensus sequence (<http://www.ebi.ac.uk/>). The consensus sequence was used as the input for the DIVERGENCE function of MAK. The returned divergence values are calculated percentage of sequence differences based on pairwise alignments of each member to its consensus sequence. Divergence value of each member of the family were organized into groups with a unit of one half

percentage, and the number of members in each group was plotted on a graph.

HTH motifs were predicted with pfam domain database (<http://pfam.xfam.org>) and by NCBI webservice. Phyre2 was used to construct a conceptual protein structure (<http://www.sbg.bio.ic.ac.uk/phyre2/>). The phylogenetic tree was generated with Genebee webservice with 1000 bootstrap value ([http://www.genebee.msu.su/services/phtree\\_reduced.html](http://www.genebee.msu.su/services/phtree_reduced.html)).

Moss TLEs *PpTc1* and *PpTc2* consensus sequences were used to search against the assembled transcripts database Pp0409 on the moss genome browser (<http://www.cosmoss.org/>) [79]. Each hit was inspected for an ORF that may translate conceptually into a potential transposase sequence with APE program (<http://biologylabs.utah.edu/jorgensen/wayned/ape/>). The loci of transcripts were cross-referenced to the nucleotide BLAST hits to avoid redundancy. The sequences were also used to search for moss small RNA databases at [cosmoss.org](http://www.cosmoss.org).

### **II-2.3 Identification and analysis of TLEs in other genome databases**

The peptide sequences of the putative transposases of *PpTc1* and *PpTc2* were used to search against other plant and non-plant genomes in WGS and nr/nt databases at the NCBI webservice. Hits and their flanking sequences were collected and inspected to identify putative transposase or TIR sequences.

Plant genomes were also screened for MITEs sequences with program MITE Digger (<http://labs.csb.utoronto.ca/yang/MITEDigger/>) [80]. Each returned MITE family was inspected for 5' CAGT...ACTG-3' terminal sequences that were flanked by dinucleotides TA. Plant genomes that contained full-length TLEs were also screened with MAK function TOPDOWN for any related MITE families.

## II-3 Results

### II-3.1 *Tc1*-like elements in moss

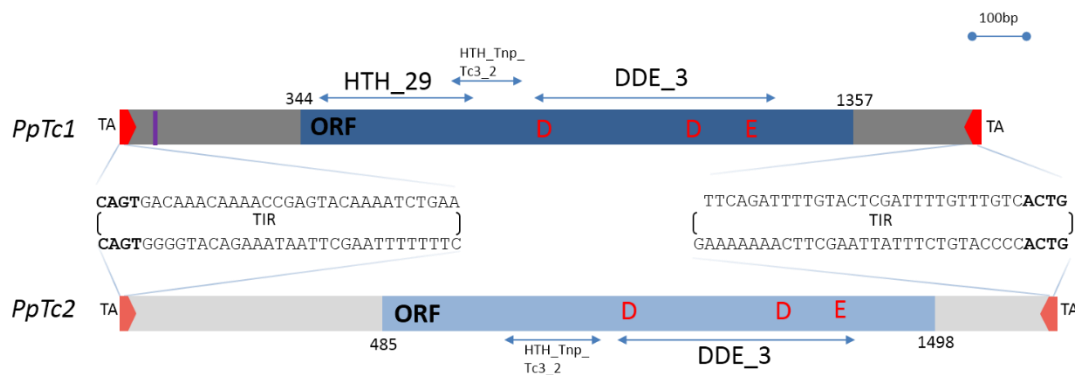


Figure 4 Summary of major structure components of moss TLEs, *PpTc1* and *PpTc2*. Each element has two TIRs (pentagons), an ORF (blue bars) that codes for a theoretical transposase with DD34E catalytic domain (red letters), and flanking dinucleotide TSD “TA”. Purple bar: hypothetical promoter in *PpTc1*. All structural components are scaled representations.

It was previously found that *Mariner*-like elements, the sister group of TLEs, are widespread in plant genomes. In contrast, despite the abundance of TLEs in other kingdoms, there has been no previous identification of TLEs in the kingdom Plantae. To investigate whether plant genomes contain

TLEs, moss genome sequence databases were screened because mosses are among the first terrestrial plants and they are evolutionarily ancient.

When the sequence of *Tc1* transposase was used as the query sequence for BLAST search against the moss (*Physcomitrella patens*) genome database [81], 118 significant hits were returned with low e values ( $<10^{-8}$ ). Closer inspection of the returned sequences revealed that many of these elements have complete terminal inverted repeats (TIRs) with terminal 5'-CAGT ... ACTG-3' sequences flanked by TSDs of dinucleotides TA, and contain open reading frames (ORFs) for a theoretical transposase bearing a DD34E motif (Figure 4). These are characteristics of TLEs and therefore the elements were classified as TLEs and designated *PpTcs*. The *PpTc* elements can be classified into two distinct groups, *PpTc1* and *PpTc2*.

The full-length *PpTc1* elements are 1584 bp long with TIRs of 33 bp. It has an ORF of 338 aa with two helix-turn-helix domains and a catalytic DD34E domain. A total of 85 copies are retrieved from the *P. patens* genome sequence database. Among them, 75 are full length copies that bear intact ends with average sequence identity of 96.3%, and 52 of which are highly similar copies with >98% sequence identity, although there are no identical copies. Nine copies carry intact full-length ORFs of 338 aa.

The full-length *PpTc2* elements are 1709 bp long, with TIRs of 33 bp. A total of 22 copies of *PpTc2* were retrieved from the genome database. The 20 full-length copies have an average sequence identity of 96.6%, of which eight copies contain full-length ORFs of 338aa.



## II-3.2 Life history of moss TLEs

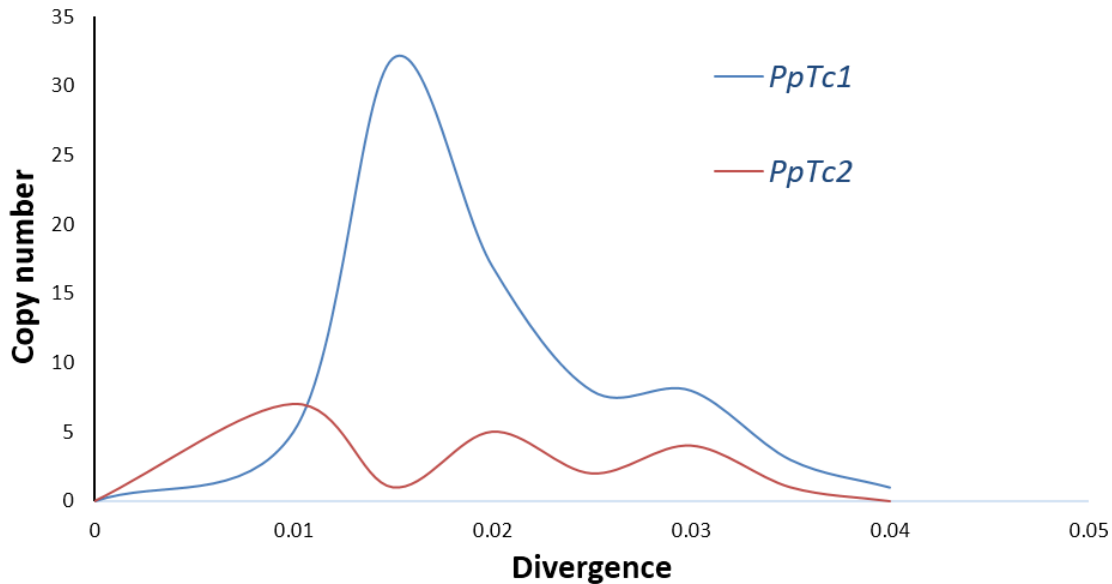


Figure 6 Divergence chart for full-length copies of *PpTc1* and *PpTc2*. Consensus was created from CLUSTAL alignment of full-length copies. Average Divergence: *PpTc1* -  $2.18 \pm 0.08\%$ ; *PpTc2* -  $2.17 \pm 0.20\%$ .

The age of a TE can be estimated by the rate of accumulation of mutations relative to other members of the family, acting as a molecular clock that measures divergence from the founding copy in the genome. The founding copy can be approximated as the consensus sequence of the family, and the subsequent duplicated copies may be similarly divergent when they were produced at a similar age. Using this theory, the age of the moss TLEs can be estimated and the life cycle of the elements can be observed by the clustering of similarly divergent copies. The sequence divergence score was calculated for each copy and placed into groups of similar values to simulate an age group. The number of copies within each group was counted and plotted against the range of divergence values (Figure 6). This graph would theoretically denote the life history of the *PpTcs* within the moss genome

based on the number of members in each age group. The *PpTc1* family has an average divergence value of  $2.18 \pm 0.08\%$  with a significant peak at 1.5% divergence, suggesting that a recent burst of amplification events of this family occurred about 1.5 *Mys* ago according to a rate of 1% sequence divergence per million years [82]. The *PpTc2* family has an average sequence divergence value of  $2.17 \pm 0.20\%$  with the most recent peak at about 1%, suggesting that *PpTc2*, similar to *PpTc1*, recently amplified about one *My* ago. The amplification events in both elements have decreased since then, reduced to a mild population at present, as indicated by the lack of identical copies.

Although *PpTc1* and *PpTc2* bear identical extreme terminal sequences 5'-CAGT...ACTG-3', their internal sequences do not share significant similarities and even the coding sequences in the full-length ORFs of the two elements share low nucleic sequence similarities. When the putative peptide sequences of the two transposases were aligned, they only share 26% (89/338) sequence identity with 47% positive (161/338) (Figure 5). These observations suggest that either the two elements shared a very distant common ancestor or they individually invaded the moss genome. However, the very similar intra-family sequence divergence levels of the two families suggest that they invaded and amplified in the moss genome at a similar evolutionary age.

### **II-3.3 Expression of *PpTc1* in moss**

The high intra-family sequence similarity in *PpTc1* and *PpTc2* and the presence of multiple copies of elements that contain intact transposase

coding sequences indicate that these moss TLEs are potentially active. The expression of transposase is required for transposition activity; therefore, it is important to determine whether *PpTc1* and *PpTc2* are actively expressed. Extensively sequenced transcriptome of *P.patens* is available. Expressed sequence tags (ESTs) obtained from protonemal tissue were assembled into transcript database Pp0409, which contains 47,557 entries and cover 98% of the genomic sequences ([www.cosmoss.org](http://www.cosmoss.org)).

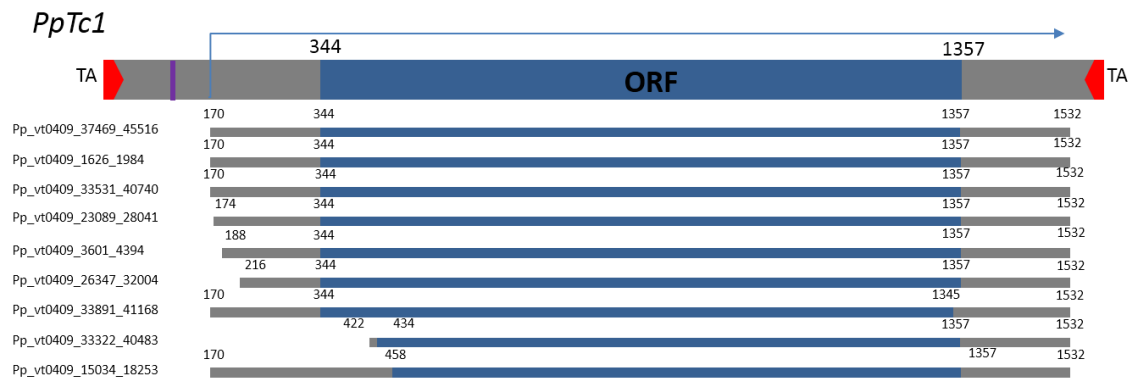


Figure 7 Members with full length ORFs. Summary of assembled transcripts that encode for a full-length (>300aa) intact transposase bearing the DD34E motif. Gene model of *PpTc-1* constructed based on representative copy from Fig. 1. TSD is “TA”; Red pentagons: TIRs; Blue bars: ORFs; Purple: predicted promoter. All representations are scaled with labeled ORF start and end positions.

For *PpTc1*, there were 68 assembled transcripts from the Pp0409 database, each containing the DNA sequence of the ORF (Appendix Table A1). To compare with a native gene of moss plants, expression of *PpTc1* is higher than that of PpAct1 (Moss actin gene), which had 17 transcripts. Each of these transcripts was inspected and found to correspond to different copies of *PpTc1* element. Nine of the *PpTc1* transcripts can be conceptually translated into a full-length intact transposase (Figure 7, Appendix Table

A2). Each of these transcripts bearing intact ORFs is derived from a specific copy of the nine genomic copies of *PpTc1* bearing intact transposase coding sequences, suggesting that these elements are actively transcribed to yield mature mRNA. There were no transcripts that match the ORF sequences of *PpTc2*.

Due to the abundant numbers of genomic copies for *PpTc* elements, the host may degrade TE mRNAs into siRNAs by Dicer proteins and inhibit TE activity through small-RNA silencing (See section II-1.1). Searches for *PpTc* related small-RNA were conducted against the moss databases and there were no returned hits. This suggests that the *PpTc1* mRNAs are not targets of small-RNA silencing and they may be translated into functional transposase proteins, which in theory are capable of carrying out active transposition of all *PpTc1* copies. However, the lack of identical copies in the genome databases indicates that there are no current activities.

Therefore, moss elements must be under other means of repressive pressure from the host, such as silencing by over-production of transposase [83]. On the other hand, *PpTc2* elements have no mature transcripts in the genome databases and they may be inhibited pre-transcriptionally, for example, by DNA-methylation silencing.

### **II-3.4 *Tc1*-like elements in higher plant genomes**

Moss is a lower plant form. The presence of TLEs in moss may suggest that a lineage of TLE has branched into the plant kingdom. To determine whether TLEs have proliferated throughout plant genomes, the predicted transposase sequences of *PpTc1* and *PpTc2* were used as query sequences to

search against plant genomic sequence databases. Segments of *Tc1*-like transposase coding sequences were identified in 10 plant genomes. In each genome, a conserved region including at least the second (aspartic acid) and the third (glutamic acid) residue of the DD34E catalytic motif could be retrieved.

In the rice (*Oryza sativa*) genome, a single copy of full-length (1557bp) TLE (*OsTc1*) with TIRs of 29 bp encodes an intact ORF of 375 aa. The ends of *OsTc1* are mutated into 5'-CACA...TGTG-3' (Figure 8). Complete DD34E motifs are also present in the TLEs from wheat (*Triticum urartu*), birch (*Betula nana*), pear (*Pyrus x bretschneideri*) and black cottonwood (*Populus trichocarpa*). Although only one end of these elements can be retrieved because the genome sequence contigs end at the other end of the DD34E motif. Most of the terminal sequences carry the extreme terminal sequences (5'-CAGT...ACTG'-3) flanked by dinucleotides TA like *Tc1*



Figure 8 Sequences of complete TIRs of plant TLEs compare to *Tc1*. Asterisks indicate elements that only have one sequence end (CAGT) available in the genome contig, and 28bp from these sequence termini are shown here. Degrees of background shading indicate level of conservation among these elements, black = highly conserved, gray = somewhat conserved. Elements and host genomes are summarized in Appendix Table A3 and A4.

```

Tc1      GFVVFQQDNDFKHTSL----HVRSWFQRRHVHLLDWPFSQSPDLNPIEHLWEEEL
Impala  --IFMHDNASVETAR----IVKALLEELGVDLMTWPPYSPDLNPIENLWALM
PpTc1   KVVFQHDNDPKHTAK----SVQFWLSSQPFQLLRWPAQSPDLNPIEMHFWALL
PpTc2   QLILME DGAPVHRS S----LP LQWRRAHGIEKLFWPANSPDLNPIENVVMV
BnTc1   NLYFQQ DNAPHTSA----KSARFMKQNGLKMLLWPANSPDLNPIEHIWHL
BnTc2   DVIVLE DNAPCHSSK----ATCAARQNLGITSLKHPNSPDNLNPIENLWDQV
BnTc3   DALVVE DGASCHWSK----QTNKGREKLIHVNLNHPQSPDLNPIENVVCLQL
BnTc4   DILVVE DGAPCHTCK----LAKEARSKLGIPLSLIHPSPSPDLNPIENVWQLL
BrTc1   DFVLMHDNARCHTAR----VSRQFLREKELRTMDWPALSPDLNPIEHLWDEL
BrTc2   --VMQDNAPAHACE----NTMEEMRERSIIPIDWPFPNSPDLNPIEAVWDM
CsTc1   NSVVVM DNAPFKRA----DIQELLEQQGHKILWLPAYSPDLNPIEHMMAWV
HvTc1   GAVIVM DNVSFHKKRQ----DTQAAIQKAGFILEYLPYSPDLNPIEHKWAQA
HvTc2   KIHIL DNSGYHCSQ----RVKDAALEKAIIVLHLYLPPYSPDLNPIERLWKVM
HvTc3   NSVMVM DNASFHKKRQ----DIQDAIKDAGFILEYLPVYSPDLNPIEKKWAHA
HvTc4   ---QHDLAPAHSAK----TTGKWFTHDGHITVLLNWPANSPDLNPIENVLWDIV
LsTc2   KIHIL DNSGYHCSQ----RVKDAALEKAIIVLHLYLPPYSPDLNPIERLWKL
LsTc1   GAVIVR DNVSFHKKRQ----DIQAAIQKAGFILEYLPYSPDLNPIEHKWAQA
OsTc1   LFQLMH DNARPH TAR----VVRQT LAAANINVLPPWPAQSPDLNPIEHAWDM
PtTc1   GWVFQHDNDPKHTAK----ATKEWLKHKH IKVM EWPSQSPDLNPIE-----
PxbTc1 DFVLMHDNARCHTAR----VSRQFLREKELRTMDWPALSPDLNPIEHLWDEL
PxbTc2 NAMCMHDNARAH TAQ----VVDEYLHDVGIHKMEWPARS PDLNPIEHAWDER
TuTc1   NWIFQQDNDFKHTSI----LVRNWLAEANGVAVMQWPSQSPDLNPIEHLWAEV
AgTLE   HYIFQH DNDSKHTSR----TVKCYLANQDVQVLPWPALSPDLNPIENLWSTL
AltLE   ATIFPQ DNDPKHTSR----KAKKCLQDLDMRVLQWPPQSPDLNPIEHLWDVL
CatLE   GWVFQHDNDPKHTAK----ATKEWLKHKH FKVL EWPSQSPDLNPIENLWREL
CctLE   RFTFQQ DNDPKHTAK----ITKEWLHNSVTVLEWPSQSPDLNPIEHLWRDL
DvtLE   RYKLYQ DNDPKHTSF----LCRTWLLYNCSKVIDT PAQSPDLNPIENLWAF
HmtLE   EVIFQQ DNDPKHTAK----IVKNWLASQDFQVLEWPAQSPDLNPIENLWSQL
MstLE   GVSVMV DNASFHRSR----ILTPMFAHHDME LMFLPAYSPDFNPIETLEAVV
RctLE   DFIFQQ ELAPAH TAK----SINTWFNDHGITVLDWPANSPDLNPIENMGGIV
RdtLE   QVIFQH DNDPKHTSK----LVKEYLKDQSYNILEWPAQSPDLNPIENMWSLL
TvtLE   NVIFQQ DNDPKHTSK----KAKRCFKENKIQVLEWPAQSPDLNPIEHLWGLV
WeTLE   DIIFQQ DGDPKHTAK----IVKEWIGKQHFQLMEWPAQSPDLNPIENLWSIV
VctLE   ---MDNAS IIRAA----PVCEVVTAKGLEAIFT PPYSPDFNPIENVNAFTKV
Mos1    RVIFLHDNASEHTAR----AVRDTLET LNWEVLP HAAYS PDLPASDYHLFAS
Soymar1 TIFIQQ DNARTEINPDPEFVQAA TQDGFDIRLMCQFPNS PDFNVL D LGFFSA
Osmar5  TIWIQQ DNARTEILPIDDAQGVAVAAQSGLD IRLVNQFPNS PDMMCLDLGFFAS

```

Figure 9 Plant D34E motifs alignment. Alignment shows D34E region from all plant *Tc1*-like peptide segments, with some non-plant TLEs and plant MLEs of various DDD motifs. Sequences are shaded to highlight the conserved blocks near the catalytic residues. Degrees of background shading indicate level of conservation among these elements, black = highly conserved, gray = somewhat conserved. Elements and host genomes are summarized in Appendix Table A4.

(Figure 8). Shorter ORFs (72 aa to 184 aa) that contain at least the second (D) and the third (E) residues of the triad motif can be found in the genomes of barley (*Hordeum vulgare*), Chinese cabbage (*Brassica rapa*), hemp (*Cannabis sativa*), and lettuce (*Lactuca sativa*), as well as birch and pear.

To illustrate the transposase sequence similarities between plant TLEs and previously reported TLEs in animals and fungi, the regions containing the second (D) and the third (E) residue of the DD34E motifs on plant TLEs were aligned together with representative animal TLEs and plant MLEs (Figure 9). The residues surrounding the catalytic triad motif showed strong

sequence conservation. The third residues of the triad catalytic motif and residues surrounding them are distinct between TLEs and MLEs.

To understand the evolutionary relationship between the plant TLEs, an unrooted phylogenetic tree was constructed for plant TLEs and some non-plant TLEs from a wide range of species (Figure 10). The non-plant TLEs include well-characterized elements such as *Tc1*, *Tc3* and *Impala*. *Soymar1* and *Osmar* are plant MLEs and they were included as an outgroup.

Furthermore, non-characterized elements that showed significant sequence homology with the plant TLEs were also included. The transposases of plant TLEs are well diverged, even elements from the same genome may belong to different clades, e.g. *PpTc1* and *PpTc2* are grouped into distant clades.

Grouping of plant and non-plant TLEs further support that plant TLEs are from different lineages of TLEs. Generally, TLEs from both plant and non-plant genomes show two major clades each related to *Tc1* and *Tc3*.

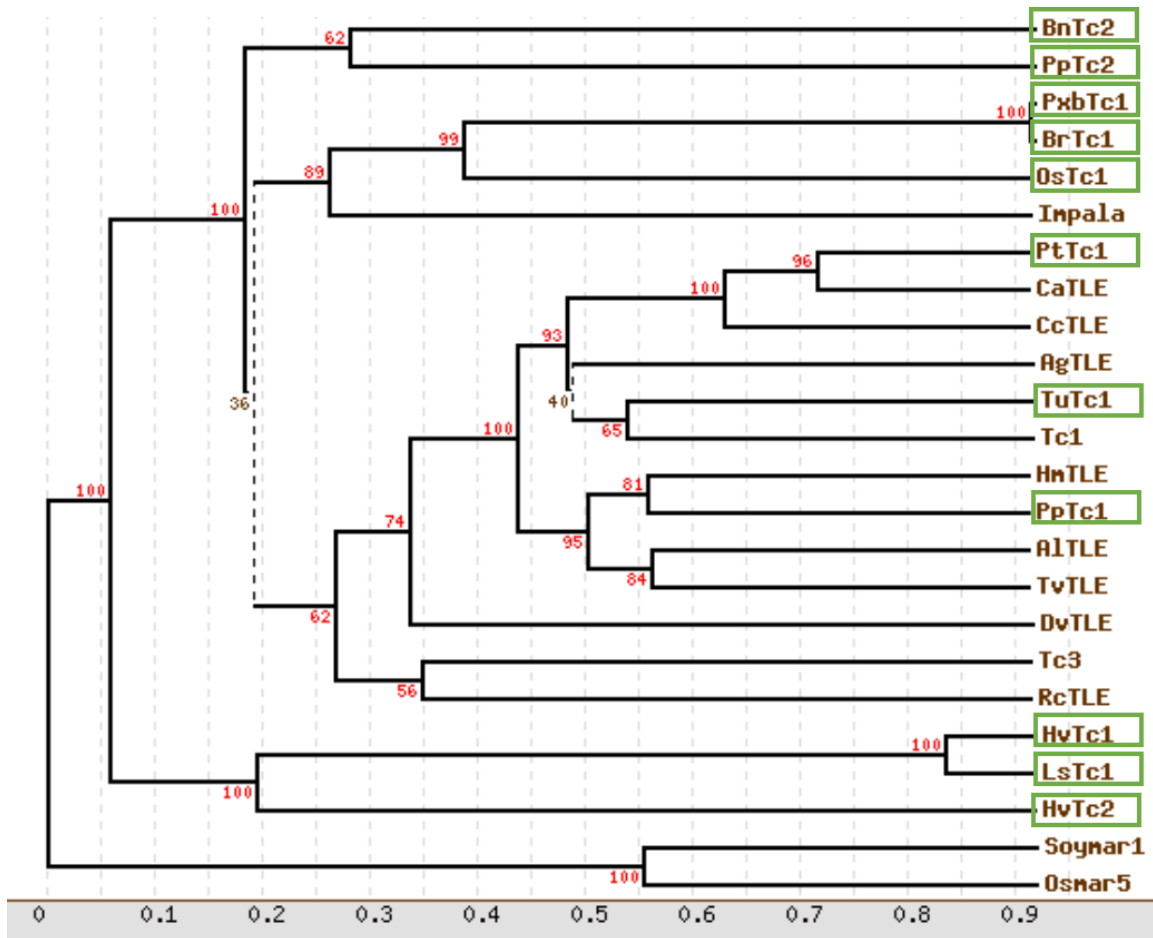


Figure 10 Phylogenetic relationship between plant *TcI*-like elements. The tree includes other representatives of non-plant TLEs and plant MLEs (green boxes). The tree was constructed based on the alignment of the entire DDE/D motif peptide sequences. The tree was generated with Neighbour-Joining algorithm with bootstrap values (1000 reps) indicated on each node. Elements and host genomes are summarized in Appendix Table A4.

### II-3.5 Cross-kingdom horizontal transfer of plant TLEs

On the phylogenetic tree constructed from plant TLE transposases, some elements in different species such as *OsTc1* from rice and *PxbTc1* from pear were highly similar, unexpected from vertical transmission from a common ancestor considering the divergence of the two species. There was high incongruence of the phylogenetic relationship between the elements and the host species. It suggests that these elements may have not been derived from a common ancestor and vertically inherited by the hosts, instead horizontal transfer events may be the underlying mechanism. To investigate whether any of these elements may have been involved in horizontal transfer, the plant TLEs were used as query sequences to search against all of the currently available genomic sequence databases, in an attempt to find highly similar sequences in other genomes.

Five of the plant TLEs have significant matches in the databases of other genomes with overall DNA sequence similarity >90% (Figure 11). For the rice TLE *OsTc1* element, a nearly identical copy (99% sequence identity) exists in the genome of the corn earworm *Helicoverpa zea*. In fact, the insect element was previously characterized as *H<sub>z</sub>Tc1* [84]. Both elements bear complete TIRs at both ends and are flanked by the dinucleotides TA. Both elements encode an intact open reading frame (375 aa). There are only 16 mismatches between the two elements, although there is no significant sequence similarity in the flanking sequences. The flanking sequences of *OsTc1* are of apparent rice origin and share similarity to the rice sequence contigs AAAA02048058, AAAA02044200, AAAA02043435, and

AAAA02049671. The flanking sequence of *H<sub>z</sub>Tc1* contains a P450 gene that is of apparent insect origin.

For the *Brassica* TLE *BrTc1*, highly similar sequences were found in the genomes of pear (*Pyrus x bretschneideri*) and cabbage moth (*Plutella*

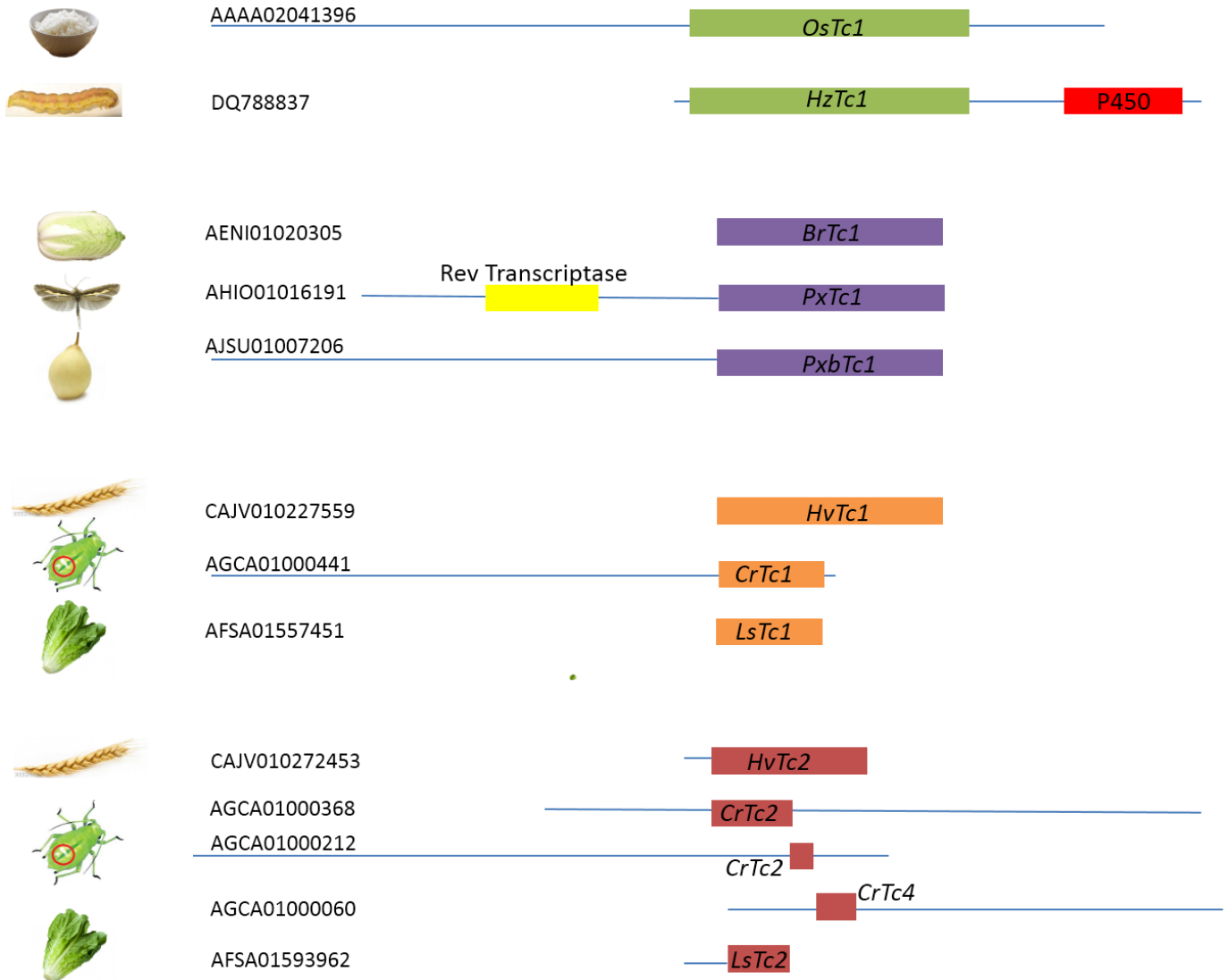


Figure 11 Horizontal transfers of plant TLEs. Thin blue lines: flanking sequences in the contig of each accession with no significant sequence similarity; Red: P450, an insect gene; Yellow: reverse transcriptase; Red circle on aphid abdomen: endosymbiont, *Candidatus Regiella insecticola*.

*xylostella*) [85, 86]. While these three elements are identical in DNA sequence, there is no detectable sequence homology in the flanking regions. The flanking sequences of *PxTc1* carry a reverse transcriptase gene that shares ~70% sequence identity with another reverse transcriptase gene in the *Brassica* genome. The element in cabbage moth has 1125 full-length nearly-identical copies in the genome database. Two different TLEs in the barley (*Hordeum vulgare*) genome have segments of highly similar matches (>97% sequence identity) in other genomes [87], both involve the lettuce (*Lactuca sativa*) and an endosymbiont bacterium (*Candidatus Regiella insecticola*) in pea aphid species (*Acyrtosiphon pisum*) [88, 89]. The presence of these highly similar sequences of TLEs in different genome databases suggest that the patchy distribution of TLEs in plant genomes may have resulted from horizontal transfers from other organisms.

### **II-3.6 *Tc1*-like MITEs in plant genomes**

Miniature inverted-repeat transposable elements (MITEs) are the most reduced form of non-autonomous elements with no internal sequences of their full-length relatives. They usually achieve much higher copy numbers than their autonomous counterparts, by using the transposases coded by the autonomous elements. To determine whether TLE-related MITEs exist in plant genomes, currently available plant genome databases were screened for hypothetical *Tc1*-related MITEs with MITE Digger. Two such MITE families were identified, one in Chinese cabbage (*Brassica rapa*) and the other in potato (*Solanum tuberosum*). Both MITEs bear the conserved terminal nucleotides of 5'-CAGT...ACTG-3' flanked by TA (Figure 8). The

MITE in potato, designated as St-TLM, has 9 copies. ST-TLM is 150bp in length with TIRs of 21bp. The overall sequence similarities of St-TLM copies is 70.0%, indicating high sequence divergence between members. The MITE in *B. rapa*, here designated as Br-TLM, is 198 bp long and it has 147 copies bearing TIRs of 24bp. The overall sequence similarity of Br-TLM copies is 85.6%. The average divergence of each member to the Br-TLM consensus sequence is  $7.479 \pm 0.3197\%$ . The lower copy number and sequence similarities of the potato MITE suggests that this is an older family and has probably lost some of its members from the genome. Generally, the relatively high sequence divergence of the two MITE families suggest that they have been inactive in transposition, probably due to loss of the TLEs that had provided them transposases from the genome.

## **II-4 Discussion**

### **II-4.1 Moss TLE elements and activity**

The identification of TLEs in plant genomes expands our knowledge on the distribution and diversity of *Tc1/mariner* elements. Elements belonging to the *Mariner*-subgroup have been found previously to be widespread in plant genomes [43]. TLEs, however, have not been reported in plants. In fact, *PpTc1* and *PpTc2* are the first *Tc1/mariner* elements described in moss. They not only expand the range of distribution of TLEs into plants, but also provide information for the development of TE-based tools for gene discovery in moss in the future.

As shown by the sequence data collected from each family, *PpTc* elements have undergone a recent wave of proliferation with many highly similar copies (Figure 6). The lack of identical copies in the moss genome also suggests that their transposition activities are currently debilitated. Although most copies of *PpTc* elements have lost the capability of encoding a functional transposase due to mutations that interrupt the transposase coding sequences, both families have several members bearing full-length intact transposase-coding genes. *PpTc1* elements are actively expressed in moss with many detected transcripts but no small-RNAs that target its coding sequence. These results indicate that *PpTc1* activity is inhibited post-transcriptionally, although there may still be functional transposases that can assist in some undetected activity. The absence of transcripts from *PpTc2* may indicate a high level of pre-transcription repression. Availability of intact transposase coding sequence also allows the *PpTc* elements to carry out transposition activities when the inhibitory mechanism is relieved, e.g. during stress. Alternatively, the activities of these elements may be restricted to certain tissues or specific temporal stages during the life cycle of the plant. It is important to recall that the transcript data were collected from protonemal tissue. Further investigation on the repression of the transposition activities of both families will facilitate our understanding of the interaction between TEs and their host genomes, which will be the focus in Chapter III.

## **II-4.2 Horizontal transposon transfer of plant TLEs**

HTT is an important aspect of TE propagation, since it opens new channels for TEs to bypass evolutionary barriers to invade other species that may be vacant of similar TEs. HTT events often associate with patchy distribution, incongruence between host and TEs evolutionary relationships and unexpected sequences in divergent organisms [56, 90]. The phylogenetic relationship between the plant TLEs was a puzzling result at first sight and it corresponded with these criteria. Close inspection and searches of non-plant sequences revealed highly probable episodes of HTT that crossed kingdoms of life. These discoveries raised interesting propositions for the role of HTT in transmittance of DNA transposons into the kingdom Plantae and perhaps genomes of other forms of life.

HTT is rare, although also most likely under-reported. The vast amount of genome data hosts many uncharacterized TEs and some of the most extensively studied genomes for HTT, such as the fruit flies, dominate the breadth of our knowledge on the matter of HTT. From previously reported cases of HTT, DNA transposons are primarily responsible for cross-phyla HTT and beyond. In fact, all known cases of long-distance horizontal transfer events are DNA transposons. This observation is validated by the ease of introducing DNA transposons to new genomes in experimental scenarios and their prowess as gene-delivery tools.

The exact mechanism of HTT is still unknown. The most probable route of transmission is through a vector that is likely to exchange DNA material with the host. The HTT between pea aphid endosymbiotic bacteria and

cabbage and wheat may be a candidate for this hypothesis. The bacterium, *Candidatus Regiella insecticola*, lives in the hemocoel of aphids. This species of bacterium is exposed to both aphid cells and aphid ingested plant matter, which provides an opportunity for HTT between the bacteria and multiple plant sources. It should also be noted that *Candidatus Regiella insecticola* in pea aphids is believed to provide a beneficial trait for protection against parasitoids, and a previous demonstration of HTT has been made between a parasitoid and insects through a viral vector [89].

Free-DNA in body fluids may also facilitate HTT events through feeding, which is likely the passage responsible for the HTT involving aphid endosymbiont. All of the proposed HTTs in this study include a herbivore with plants that are often direct food sources. Insects that feed on multiple food sources may transmit their TEs into different plants. The reported horizontal transfer of a *Mutator*-like element between rice and millet may be mediated by an insect that feed on both species [91]. Similarly, the transfers of TLEs found in lettuce and barley may have been mediated by the pea aphid via its symbiont *Candidatus Regiella insecticola*, and similarly between pear and cabbage via moth. It should be noted that diamond back moth is not known to feed on pear trees.

Not all the TLEs described in this study are candidates for HTT. It is likely that vertically inherited TLEs are widespread in plant genomes as well, given the breadth of the plant species that host TLEs. The dual-route transmission of TEs, however, highlights the importance for our

understanding of HTT, which may skew our knowledge of evolutionary history of TEs of all categories.

### **II-4.3 Life cycle of TE and HTT**

HTT creates a route for TEs to quickly bypass evolutionary processes and spread into other genomes. The new hosts lack inhibitory mechanisms that target these newcomers and these TEs may initiate unhindered proliferation. The increased mutational load from TE amplification may be detrimental to the host if it is left unchecked. If the TEs were continued to multiply, inhibitory mechanisms (See III-1.1) then may target these new TEs in the genome and result in decreased activity. Inactivity and accumulated mutation cause loss of intact coding sequences and the TEs become evolutionary relics that are commonly found today. This life cycle of TE underscores the source of accumulation of the inactive TEs, i.e. “junk” DNAs in genomes.

However, HTTs allow TEs to escape unfavourable genomic environments, where epigenetic silencing causes loss of transposition activity. Once a TE invades a new genome where it is unrecognized by the inhibitory mechanisms, the founding copy initiates a new life cycle, and the persistence in this new lineage would rely on adaptive traits that all TEs to evade inhibitory mechanisms and survive with basal levels of activities. A proposed mechanism for evasion will be discussed next in Chapter III.

## Chapter III: Autoregulation of *Osmar* transposition

### III-1 Background

#### III-1.1 Host-TE relationship

Although the presence of TEs in genomes can be best described as parasitic in nature, there is, however, a benefit for an organism to harbour TEs in its genome. Mobilization of genomic sequences assists with host species genetic diversity and long-term fitness [2, 25]. In a shorter time-frame, it was also observed that TEs are often activated during stress events [92-94], which constituted McClintock's original theory that TE activity may contribute to genetic diversity and fitness through accelerated molecular evolution [95]. *Ty5* integrase, for example, can be modified through phosphorylation to lose its specificity for heterochromatin during stress in *S. cerevisiae* resulting in non-specific integration [94]. TEs are also a source for new genetic sequences that can be incorporated into the host genome. TE regulatory elements may be "domesticated" by the host and perform functions for host genes [68, 96]; sometimes even the TE element themselves can be domesticated [68-71, 97]. These behaviours additionally add to the potency of TEs as a source of additional host genetic diversity.

While there are some characterized examples of beneficial aspects of TEs, TEs are mutagenic in nature and may impose detrimental effects to the host. In theory, a successful TE must achieve maximal number of copies in a genome without killing the host by introducing exceedingly high volume of mutations; in turn, the host must ensure that TEs remain only basally active

and non-threatening, and therefore, a delicate balance is established in genomes allowing both TEs and the host to survive through evolutionary ages symbiotically [25, 98-100].

Many TEs have evolved to target specific regions in the host genome, where their activities would cause minimal damage to the host [25]. For example, *L1* elements with mutated endonucleases (endonucleases are required for endogenous integration) have been observed to target telomere regions [101]. Some *Penelope*-like elements also lack an active endonuclease and integrate specifically into telomeres [102]. In addition, *Ty5* in *S. cerevisiae* targets gene-poor heterochromatin. The *Ty5* integrase has a targeting domain in the C-terminus that binds to a structure in the heterochromatin [103, 104]. This specificity, interestingly, is thought to be a mimic of *Esc1p*, a native protein that possesses the same targeting sequence [105]. Similarly, *L1* and *Penelope* elements integrate into telomeres with a mechanism similar to that of telomerase. These observations additionally highlight the close relationship between TEs and host.

Some TEs target genic regions of the host, but adopt mechanisms to avoid causing damage to host genomes. For example, *Tn7* in *E. coli* encodes a DNA-binding protein that has high specificity and ensures insertion into a safe site for the host [106]. Similarly, *P* element in *D. melanogaster* targets 500 bp upstream of transcription start site (TSS) of genes to avoid direct insertion into important open reading frames [107]. *Ty1* and *Ty3* in *S. cerevisiae* target upstream of genes that are transcribed by RNA polymerase

III [108, 109]. These mechanisms allow TEs to transpose near genes without interrupting transcription of host genes.

Despite the properties of some TEs to avoid harm to the host organism, other unregulated transposition events may cause deleterious effect in the host genome if they are left unchecked, and eukaryotes have evolved mechanisms that limit or inhibit TE activities [26, 110-112]. These defense mechanisms inactivate TEs and result in a majority of observed TEs to be dormant copies. Some of the most significant epigenetic mechanisms adopted by eukaryotes are DNA methylation [113-116], RNA-mediated silencing [117-119], and histone modifications [120].

DNA methylation is an addition of 5-methyl group to a cytosine, which represses transcription of TEs. The majority of DNA methylation events are attributed to repression of retrotransposons, and inability of *Arabidopsis thaliana* to carry out cytosine methylation result in rapid retrotransposon proliferation [121]. Another defense mechanism is the small-RNA inhibition, which includes two mechanisms to carry out post-transcriptional degradation [25, 122]. Small interfering RNAs (siRNAs) are generated from double stranded RNAs (dsRNAs) that are derived from TE sequences. siRNAs are short (20~24 bp) sequences processed by Dicer proteins, and they target TE mRNA strands that are complimentary to the siRNAs. The RNA-induced silencing complex (RISC) formed by the siRNA and an Argonaute protein recognizes mRNAs of target TEs and cause their degradation. Another small-RNA inhibition mechanism, called PIWI-interacting RNAs (piRNAs), uses the PIWI proteins and piRNAs generated

from TE RNA sequences to repetitively cleave TE mRNA [123]. Small-RNA inhibition may control both DNA transposons and retrotransposons, as demonstrated with *Tc1* elements in *C. elegans* [124], *P* elements in *D. melanogaster* [125], and *L1* elements [126].

Thus, TEs and their hosts have developed an intricate relationship. The detrimental mutational effects of TEs cause the host to adopt defense mechanisms that effectively inhibit TE activities. TEs appear to have evolved to minimize damage to the host and to target specific “safe” sites. Interestingly, internal sequences of TEs may play a role in the TE-host dynamic. In this study, a rice *Mariner*-like element was analyzed for self-regulation to limit its own activity and evade inhibitory mechanisms.

### **III-1.2 *Mariner*-like elements**

The other major sub-group of the *Tc1/mariner* superfamily of DNA transposons is the *Mariner*-like elements (MLEs). MLEs carry DDD catalytic domains, with the same target site of dinucleotide “TA” as its sister group, the TLEs [39]. Originally only well-characterized in insects and fungi, MLEs were subsequently found to be widespread in both animal and plant genomes [43].

*Osmars* are a group of MLEs found in rice genome *Oryza sativa*. There were 34 identified *Osmars* and about 22000 *Stowaway* MITEs that associate with them [127]. It was demonstrated that 6 of the *Osmars* encode transposases that carry variable functionality in a yeast excision system, and *Osmar5* and *Osmar14* transposases can achieve significantly high excision

activity [34, 74]. In particular, *Osmar14* transposase can excise a MITE, *Ost35*, with significantly high efficiency. Although DNA transposons are widespread and abundant, few have been demonstrated to be actively transposing.

The crystal structures of some transposases have provided insight of details of the transposition reaction that facilitates mobility of DNA transposons. An animal *Mariner*-like element *Mos1* from *Drosophila mauritiana* was the homologous model transposase for *Osmar* transposases (<http://www.sbg.bio.ic.ac.uk/phyre2/>). The *Mos1* transposase protein contains a bipartite helix-turn-helix DNA-binding domain and an aspartic acid triad for its catalytic center [29, 128], with a general tertiary structure similar to the computationally constructed *Osmar* transposase structure. To accomplish excision, two *Mos1* transposases bind to TIRs of the DNA elements as two monomers or a dimer and form a paired-end complex between two TIRs to catalyze strand cleavage [29, 128]. Similarly, *Osmars* have two binding sites in each terminus to allow transposase recognition and attachment [127]. Interestingly, a duplicate of the 3' terminal binding site can be found upstream in the 3' sub-terminal region in many *Osmars*. It was shown previously in *Osmar5* that this subterminal duplicated site also binds to transposase, thereby providing a third binding site for transposases [127]. Mutations within the 3' subterminal binding site resulted in reduced binding ability by the transposase to this motif in *Osmar5*, similar to the terminal binding sites [127]. This indicates a crucial binding function for the *Osmar5* 3' subterminal motif. However, the *Osmar14* 3' subterminal motif was found to have a repressive effect on transposition [34]. Mutations near the

*Osmar14* subterminal motifs have variable degrees of alleviation effect and result in increased transposition. Additionally, *Ost35* lacks the subterminal motif of *Osmar14*, but shares similar terminal sequences. It is able to amplify dramatically by *Osmar14* transposase. These findings suggest that *Osmar14* transposase is highly functional, but the subterminal repressive motif within the *Osmar14* DNA elements prevents its proliferation.

This study attempts to elucidate the molecular mechanism of autoregulation by the *Osmar* subterminal motif. Transposition assays and DNA-protein mobility shift assays were performed with mutational constructs of *Osmar14* and *Osmar5*.

## **III-2 Methods**

### **III-2.1 Cloning**

*Osmar14*NAS and *Osmar5*NAS were non-autonomous shortened versions of *Osmar14* and *Osmar5* elements. Different *Osmar5*NAS and *Osmar14*NAS constructs with mutations in the subterminal region were designed and cloned into plasmid vectors for downstream experiments. Subterminal regions of *Osmar14*NAS and *Osmar5*NAS were mutated into constructs in Table 1. Synthesized mutagenic primers from Integrated DNA Technologies, Inc. were used for polymerase chain reaction (PCR) based mutagenesis. Generated mutant constructs were digested with *HpaI* and ligated into *HpaI* restriction site within the *ade2* reporter gene of plasmid

<b>Table 1</b>	
<b>Construct</b>	<b>Mutation description</b>
<b>OSM5-N20</b>	Sub-terminal box1 mutated into SmaI digestion site
<b>OSM5-5'subTIR</b>	3'sub-terminal TIR introduced to 5'
<b>OSM5-N20-5'subTIR</b>	Sub-terminal box1 mutated into SmaI digestion site, 3'sub-terminal TIR introduced to 5'
<b>OSM14-5'subTIR</b>	3'sub-terminal TIR introduced to 5'
<b>OSM14-N20</b>	Sub-terminal box1 mutated into SmaI digestion site
<b>OSM14-N20-5'subTIR</b>	Sub-terminal box1 mutated into SmaI digestion site, 3'sub-terminal TIR introduced to 5'
<b>OSM5-3'subboxes-shift+5</b>	Sub-terminal boxes shifted 5 bp towards 5'
<b>OSM5-3'subboxes-shift-5</b>	Sub-terminal boxes shifted 5 bp towards 3'
<b>OSM14-3'subboxes-shift+10</b>	Sub-terminal boxes shifted 10 bp towards 5'
<b>OSM14-3'subboxes-shift-10</b>	Sub-terminal boxes shifted 10 bp towards 3'
<b>OSM14-3'subboxes-shift+5</b>	Sub-terminal boxes shifted 5 bp towards 5'
<b>OSM14-3'subboxes-shift-5</b>	Sub-terminal boxes shifted 5 bp towards 3'
<b>OSM14-3'subboxes-reverse*</b>	The two sub-terminal boxes reversed in position and direction
<b>OSM14-3'subT-Tlinker*</b>	Deletion of 1 Thymine in the 3'subT linker

pWL89A, resulting in pOsm14- and pOsm5- series of mutant plasmids. The pOsm14- and pOsm5- plasmids were used to transform *E. coli* and transformants were selected on Luria Broth plates containing antibiotic carbenicillin (1% Tryptone, 0.5% yeast extract, 1% sodium chloride, 100ug/ml carbenicillin, 1.5% agar). Plasmid DNA was extracted from selected colonies and correct plasmid sequences were confirmed via

restriction digest and Sanger sequencing. Sequencing was performed at the TCAG sequencing facility.

### III-2.2 Yeast Excision Assay

To test the effect of the mutations on transposition activity, a yeast-based excision assay system was used to test excision efficiency of transposase with different constructs (Figure 12). In addition to the antibiotic resistance gene and the *Osmar* sequences, pOsm14- and pOsm5- also carry the *ura3*

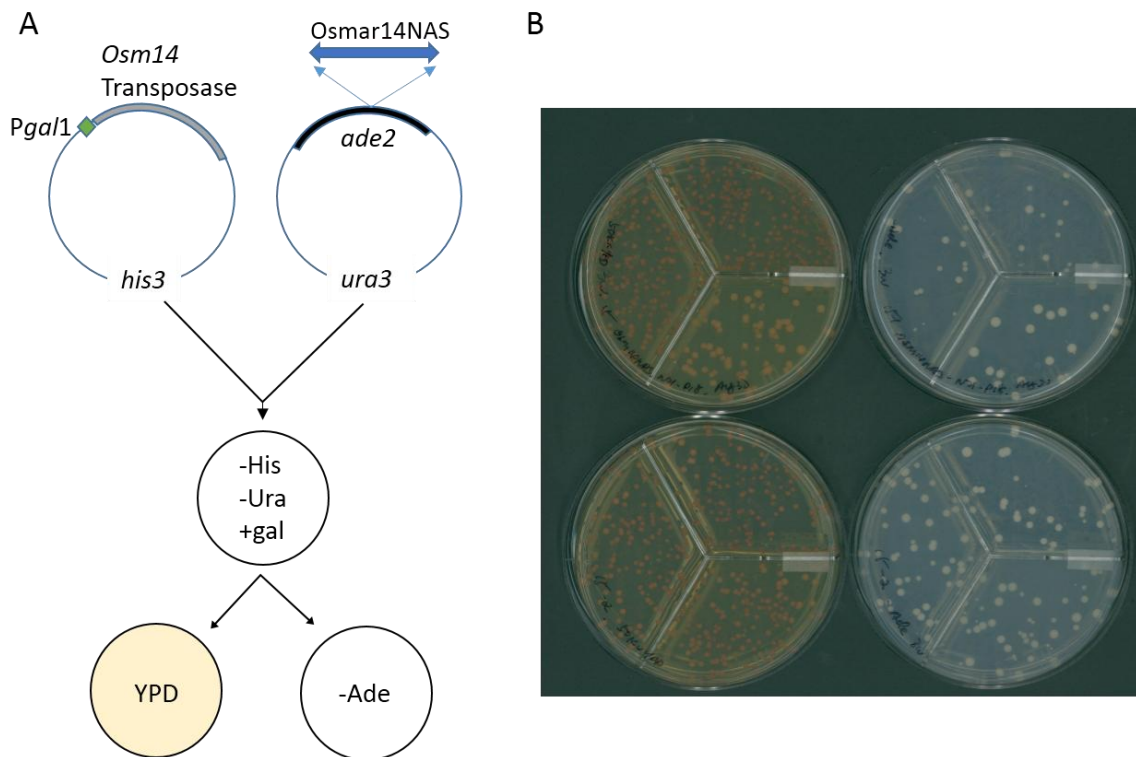


Figure 12 Yeast excision assay A) An example of yeast excision assay summary using *Osmar14* transposase-element combination. Blue circles: plasmids with embedded transposase coding sequence, grey; and *ade2*, yellow; Selection genes: *his3* and *ura3*; Black circles represent growth plates with key selection and induction nutrient. When the *Osmar14NAS* is excised out, *ade2* gene is restored and the cells can be selected on adenine drop-out plates B) Examples of growth plates: yellow/left plates, YPD media; white/right plates, -Ade media (see Methods)

gene. Plasmids pOsm14tp and pOsm5tp bear the *his3* gene, *Osmar* transposase genes and *gal1* promoter upstream of the transposase coding sequence. When a successful excision event occurs in a cell, the *ade2* gene is restored and allows the growth of a colony on the adenine-dropout plates. As a negative control, plasmid PRS413 carrying *his3* but lacking *Pgal1* and transposase coding sequences was used in place of pOsm14tp and pOsm5tp plasmids.

Cell cultures of haploid yeast strain DG2523 were prepared with 100mM lithium acetate to become competent. Competent yeast cells were double transformed with a pOsm14- or pOsm5- plasmid and a corresponding transposase plasmid via lithium acetate/heat shock method. The transformation reaction uses 50uL competent cells, 5.8uL of 5mg/mL denatured salmon sperm DNA, 400uL of PEG buffer (100mM LiAc, 1x TE buffer pH 8.0, 40% PEG3350) and heat shocked at 42C for 45 minutes. Transformed cells were grown on agar plates with 1.7% yeast nitrogen base, 5% ammonium sulfate, 2% galactose, 1% raffinose, complete supplement mixture lacking histidine and uracil (CSM-HU), and 1.5% agar. Plates were incubated for 3 days at 30°C until colonies appeared, and then were placed in room temperature for three weeks for TE excisions to occur. Colonies were picked after three weeks and each colony was resuspended in water and grown on agar plates with 1.7% yeast nitrogen base, 5% ammonium sulfate, 2% glucose, complete supplement mixture lacking adenine (CMS-ADE), and 1.5% agar. Each selected colony was also diluted by a factor of  $10^5$  and plated onto agar plates containing YPD (2% bacteriological peptone, 1% yeast extract, 2% D-glucose, 1.5% agar).

Excision frequency were calculated from number of *ade2*-revertant colonies on the adenine dropout plates relative to the number of colonies on the YPD plates. Several batches of yeast excision assays have been performed with significant variation in raw excision frequency values which are attributable to growth conditions. Nonetheless, the results showed reliably reproducible qualitative trend which is reported in detail here with a single patch of excision assay.

### **III-2.3 Electrophoretic mobility shift assay (EMSA)**

The *Osmar14* transposase coding sequence was cloned into plasmid pMal-c2x (New England Biolab) between *BamHI* and *HindIII* restriction sites. The Osm14 transposase was fused with a maltose-binding protein tag. The plasmid was used to transform *E. coli* cells BL21\*DE3, and transformants were selected on LB-Agar-carbenicillin plates. Transformed cells were grown overnight in 50mL of 2XYT growth media (1.6% Tryptone, 1% yeast extract, 0.5% NaCl) at 37°C with 100mg/L carbenicillin. The overnight culture was used to inoculate 1L 2XYT media with 0.5% glucose. The 1L culture was grown for 2 hours at 37°C and then induced with 0.5mM IPTG. Induced culture was further grown at 30°C for 4 hours. Cells were harvested and lysed with a Constant Systems Cell Disrupter (Constant Systems Ltd) at 30 psi in 20mL of phosphate buffered saline (PBS: 0.8% NaCl, 0.02% KCl, 10mM Na<sub>2</sub>HPO<sub>4</sub>, and 1.8mM KH<sub>2</sub>PO<sub>4</sub>) with 1X yeast protease inhibitor cocktail. The fusion protein Osm14Tp-MBP was bound with amylose beads resin (New England Biolab) by shaking at 4°C for 4 hours. The resins were washed with cold 1x PBS. Purified protein was eluted with

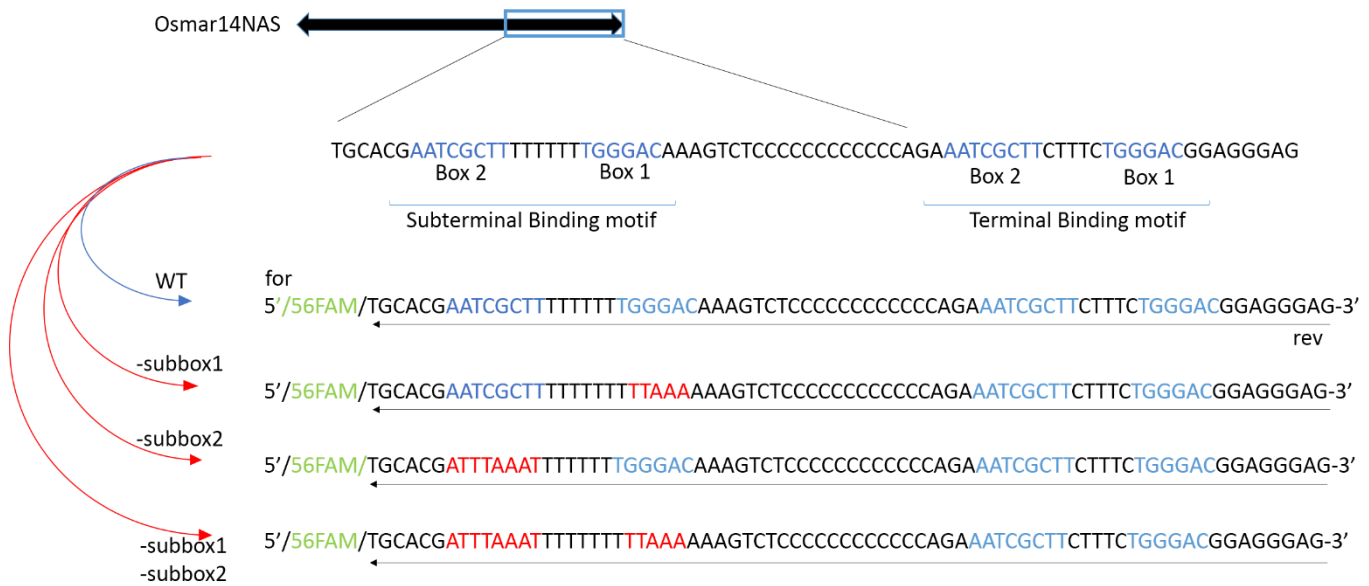


Figure 13 Oligo designs for EMSA experiments. The subterminal repeat boxes were mutated by substituting the transposase binding core into A and T nucleotides. Forward oligos were annealed with complementary reverse oligos. On the 5' end of forward oligos is a 6FAM fluorescent tag used for visualization. Red, mutated sequences; blue, wild type repeat boxes.

200 $\mu$ L of 1M amylose. The protein was quantified with Bradford method (Biorad).

Double-stranded DNA fragments were produced by annealing synthesized DNA oligos in TE buffer, and each “Forward” oligo carries a 5'-6FAM fluorescence tag (IDT) (Figure 13, Appendix Table A5). In each reaction, 1.5 $\mu$ g of Osm14Tp-MBP, or MBP control, was incubated with 100ng of annealed DNA fragments in buffer containing 15nM Tris (pH 7.5), 1mM DTT, 0.3mg/ml BSA, 33 $\mu$ g/ml salmon-sperm DNA, 10% glycerol, 0.1% NP-40, and 0.1mM EDTA at room temperature overnight. The incubation samples were separated on a 6% polyacrylamide gel with chilled 0.5 x TBE running buffer (45mM Tris-borate, 1mM EDTA). Gels were visualized under LT-9900 Illumatool Bright Light System (Lighttools Research), with

excitation source set at 488 $\pm$ 10nm and gels viewed through a 530nm YFP filter. 6FAM absorbs at 495nm and emits at 520nm.

### **III-3 Results**

#### **III-3.1 Mutations in the 3' subterminal binding motif**

*Osmar14NAS* has been previously shown to carry a 3' subterminal motif that has significant repressive effects on the transposition efficiency of the entire elements [34]. Molecular mechanisms that underlie this effect were not well understood. However, it was clear that the repressive effect was well pronounced, and mutations of the motif could result in elevated transposition efficiency.

The 3' subterminal motif has two repeat boxes (Figure 13). These are direct repeats of the 3' terminal sequences. *Osmar14NAS* and *Osmar5NAS* were both mutated at a site previously designated N20 (Mutation N20), which mutates the subterminal motif repeat box 1. Relative to the wild type *Osmar14NAS* (Relative transposition efficiency to control, RTE =  $1\pm 0.63$ ), *Osmar14NAS-N20* showed a significant increase in excision frequency (RTE =  $2.94\pm 0.95$ ) (Figure 14). *Osmar5NAS-N20*, on the other hand, showed decreased excision frequency (RTE =  $0.62\pm 0.26$ ) (Figure 15). These results suggest that the repressive effect of the subterminal motifs is only found in *Osmar14*. Mutations in the same motif in *Osmar5* actually have the opposite effect.

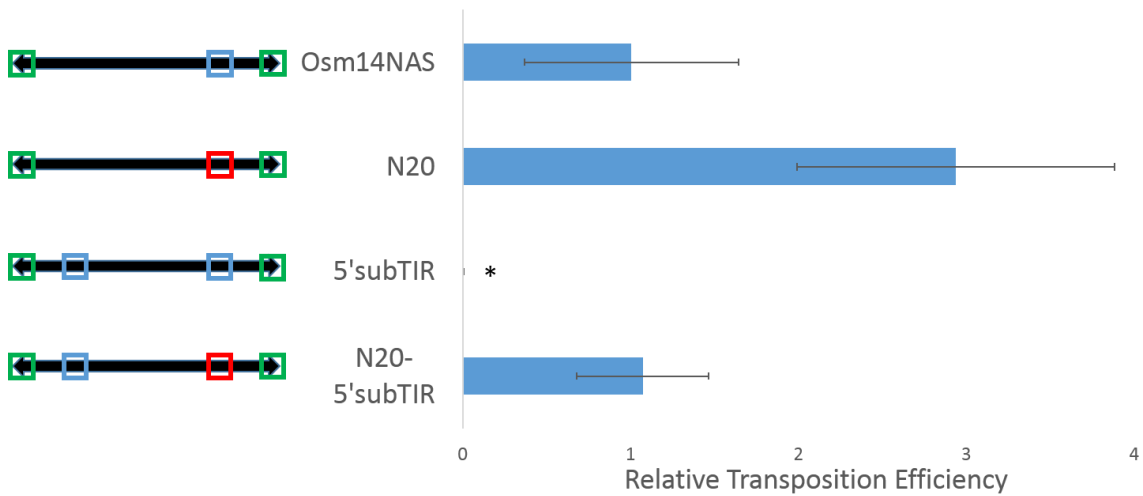


Figure 14 Comparison of excision frequency by *Osmar14* transposase in yeast between *Osmar14*NAS constructs. Left, illustration of the mutations done to the subterminal region: blue box, wild type sequences; red boxes, altered sequences. \* indicate  $p < 0.1$  in a student's t-test.

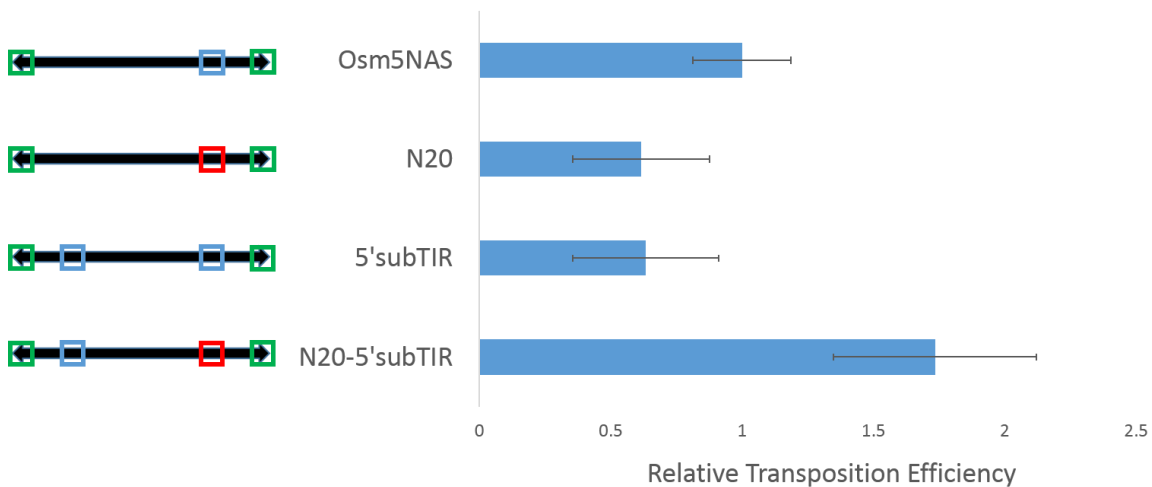


Figure 15 Comparison of excision frequency by *Osmar5* transposase in yeast between *Osmar5*NAS constructs. Left, illustration of the mutations done to the subterminal region: blue box, wild type sequences; red boxes, altered sequences.

To test the robustness of the repressive effect of the 3' subterminal motifs and whether its placement in 3' is a significant property, the 3' subterminal motifs were introduced into the mirrored locus on the 5' subterminal region

(Mutation: *Osmar14*NAS-5'subTIR, RTE =  $0.0039 \pm 0.0039$ ,  $p < 0.1$ ) (Figure 14). This mutation obliterated transposition activity in *Osmar14*NAS, which supports the notion that the 3' sub-terminal region of *Osmar14* has a repressive effect. When abolishment of the 3' subterminal motifs (mutation N20) was combined with the 5' sub-terminal addition of the motifs, the transposition activity was restored to wild type level (Mutation N20-5'subTIR, RTE =  $1.07 \pm 0.39$ ). This suggests that the repressive effect of the 3' subterminal motifs is additive when the motifs are present on both corresponding subterminal loci, and either locus may have the same repressive effect (Figure 14). Therefore, in *Osmar14*, abolishing the 3' subterminal motif increases activity, and addition of a 5' mirrored subterminal region renders the element immobile.

Contrary to what was seen in *Osmar14*, *Osmar5*NAS mutants of the same nature showed decreased transposition efficiency regardless of addition or abolishment of the sub-terminal repeats (mutation *Osmar5*-5'subTIR, RTE =  $0.63 \pm 0.28$ ) (Figure 15). Switching the 3' subterminal motifs into the 5' locus resulted in an increase in activity (Mutation: *Osmar5*-5'subTIR-N20, RTE =  $1.73 \pm 0.39$ ).

### **III-3.2 Transposase binding affinity of 3' subterminal motifs**

Transposase binding affinity in the subterminal motifs has been previously demonstrated in *Osmar5* [127]. *Osmar14* constructs in this study were designed to specifically target the sequences of the subterminal repeat boxes and the effect of these boxes on transposase binding to the elements. DNA

oligos of the 75 nucleotides on the 3' terminus were used for transposase binding. Oligos contain both terminal and subterminal binding sites.

EMSA gels revealed the presence of a single bound transposase at the terminal motif and unbound DNA in all mutant constructs (Figure 16). *Osmar14*NAS wild type showed a heavier upper band indicating double transposase binding of both terminal and subterminal motifs simultaneously. When box 1 was mutated and replaced with *SwaI* restriction recognition sequence (ATTTAAAT), transposase binding affinity was not reduced in the subterminal motif. This mutation is very similar to the N20 mutation in

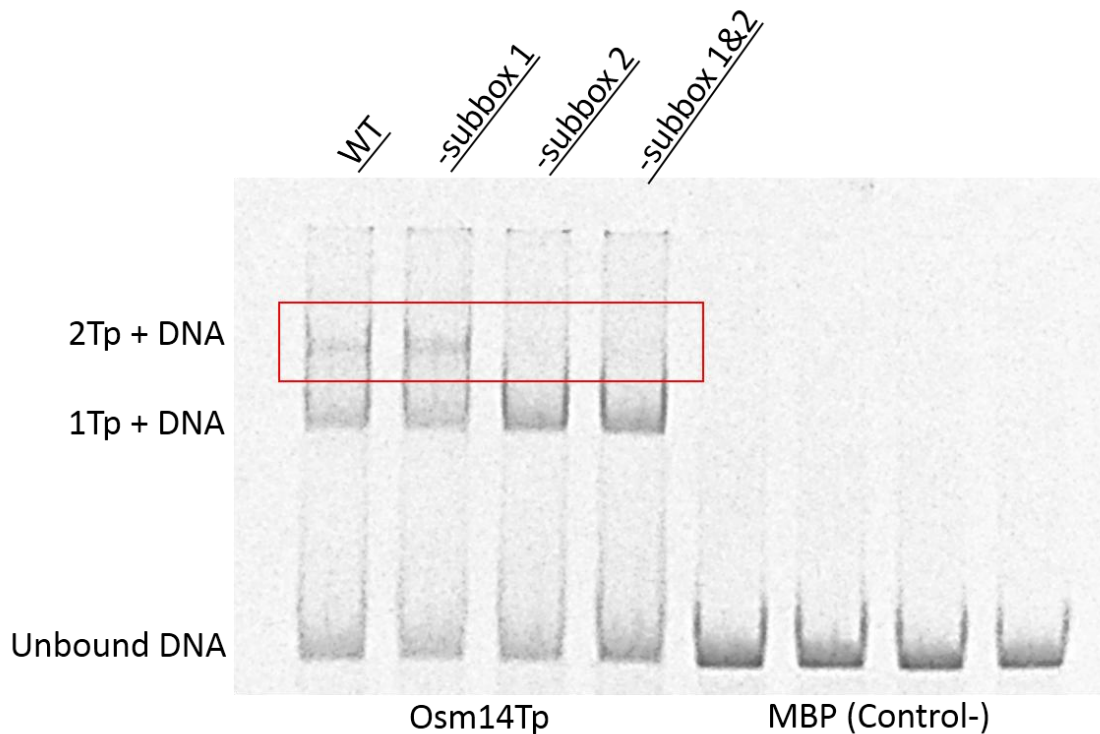


Figure 16 Gel image of EMSA experiment. The EMSA used *Osmar14* transposase (Osm14Tp) and annealed DNA oligos of 75 nucleotides from 3' *Osmar14*NAS bearing different mutations. Maltose binding protein (MBP) experiments with the same oligos showed no binding.

the same box sequences, suggesting this locus is mainly responsible for tuning transposition efficiency, but not the binding of transposases. Box 2 mutation on the other hand, resulted in disappearance of the upper band. When both repeat boxes of the subterminal motif were destroyed, upper band also disappeared. Box 1, therefore, is highly related to mediating excision suppression, whereas Box 2 is crucial for transposase binding.

### III-3.3 Effects of linker sequences between the 3' subterminal motifs

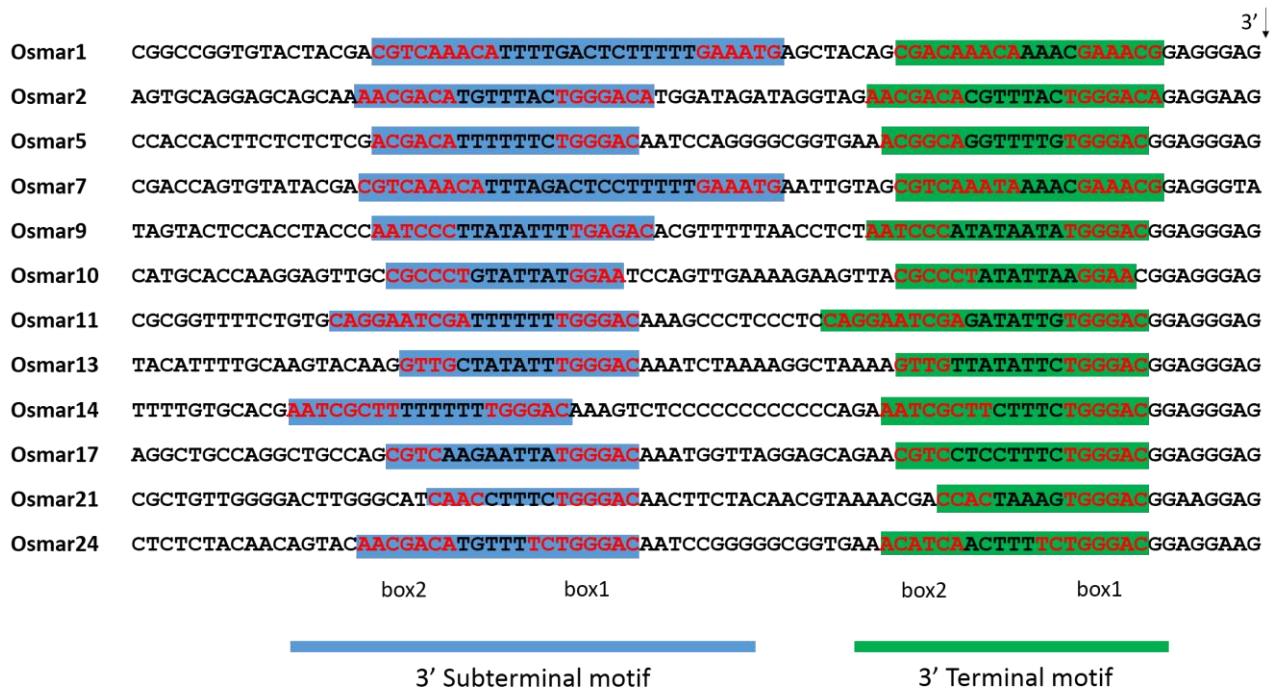


Figure 17 Nucleotide sequences of the 3' end of different *Osmar* elements. Red letters, repeat boxes that are found in both 3' subterminal and terminal motif. Blue/Green shading, the theoretical transposase binding cores. Blue/Green bars, legend for motifs, and approximate motif location at terminal and subterminal region.

It was previously demonstrated that *Osmar5* transposase is much more efficient than *Osmar14* in yeast excision assays to transpose their native full-length elements [34]. Yet, *Osmar14* transposase is capable of high activities as shown by mutated non-autonomous *Osmar14* variants and related MITEs that can achieve high excision frequency. The difference in transposase activity may lie within the 3' subterminal region of the *Osmar* elements, which is shown here to have different effects in *Osmar14* and *Osmar5* excision frequency in the yeast assays.

Many *Osmars* have a 3' sub-terminal region that carries two boxes of direct repeats of the terminal sequences (Figure 17), which have been demonstrated to bind transposase in *Osmar5* and *Osmar14*. The positions of the repeats boxes, however, vary between *Osmar* species. The differences in their positions can be attributed mainly to the linker lengths between the 3' terminal and the sub-terminal motifs. This linker sequence length is 5 bp longer in *Osmar14* than *Osmar5*. To investigate the significance of this difference between linker sizes, several mutant constructs were designed to shift the position of the 3' sub-terminal motif in *Osmar5*NAS and *Osmar14*NAS relative to the terminal motif.

The OSM14-3'subboxes-5Shift construct shifts the *Osmar14*NAS sub-terminal motif (binding core including both repeat boxes and the linker sequence inbetween) 5 nucleotides downstream towards the 3' end, effectively **shortening** the distance between the terminal and the sub-terminal motifs by 5 nucleotides. This shift makes the new linker length between the *Osmar14*NAS terminal and subterminal binding motifs the

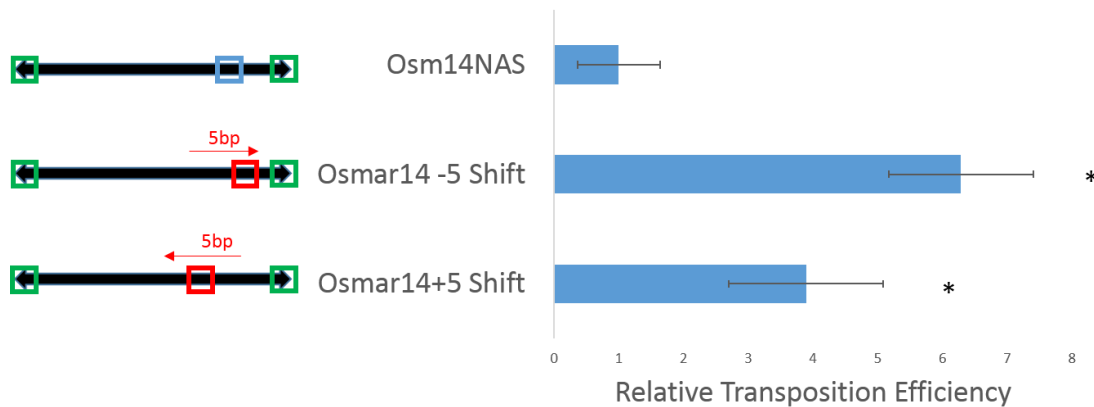


Figure 18 Comparison of excision frequency by *Osmar14* transposase in yeast between *Osmar14NAS* shift constructs. The subterminal transposase binding core (both boxes and linker sequence) is shifted. Left, illustration of the shifts done to the subterminal region: blue box, wild type sequences; red boxes, altered sequences; green boxes, TIRs. \*indicate  $p < 0.1$  in a student t-test.

same as the linker in *Osmar5NAS*. The -5 in *Osmar14NAS* linker length resulted in significant increase in transposition efficiency (RTE =  $6.29 \pm 1.12$ ) (Figure 18).

Another construct shifting *Osmar5NAS* sub-terminal binding motif core upstream (Mutant: *Osmar5NAS+5Shift*) **increases** the distance between *Osmar5NAS* 3' terminal and sub-terminal motifs by 5 nucleotides. This makes the new linker length in *Osmar5NAS* the same as the linker of *Osmar14NAS*, and cause a **decrease** in transposition efficiency (RTE =  $0.14 \pm 0.09$ ) (Figure 19). The *Osmar14 +5Shift* (RTE =  $3.89 \pm 1.20$ ) mutation relocates the subterminal motif into the relative position of *Osmar5*, and **increases** the excision efficiency. Therefore, the relative position of the 3' sub-terminal motif to the terminal motif, i.e. the linker length, is crucial to its effect on transposition activity.

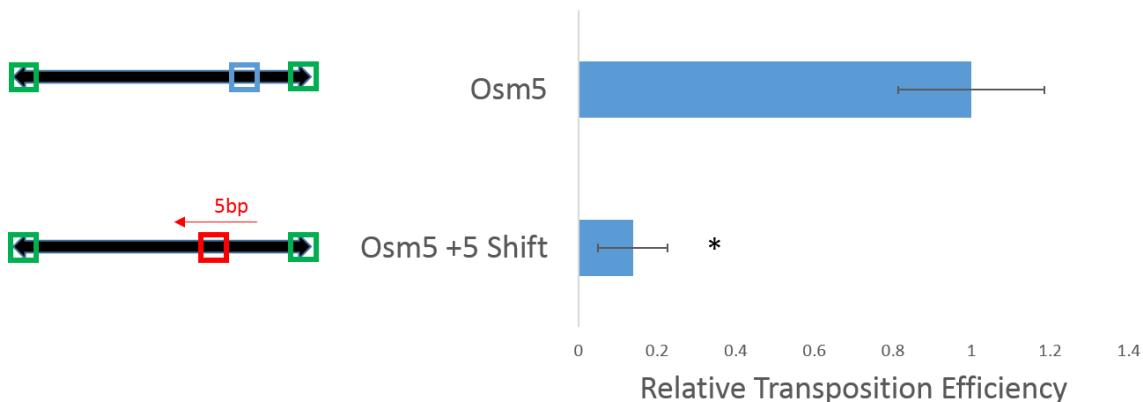


Figure 19 Comparison of excision frequency by *Osmar5* transposase in yeast between *Osmar5*NAS and +5shift construct. The subterminal transposase binding core (both boxes and linker sequence) is shifted 5bp towards 5'. Left, illustration of the shift done to the subterminal region: blue box, wild type sequences; red box, altered sequences; green boxes, TIRs. \*indicate  $p < 0.1$  in a student t-test.

*Osmar14*+10Shift and *Osmar14*-10Shift showed slight increases in excision efficiency that were much less pronounced than the +5/-5 shift mutations. These findings are the first data to reveal a cyclical relationship between linker length and excision efficiency. A subsequent experiment in the Yang Lab (unpublished work performed by Dr. C. Lee) using *Osmar14*NAS shift constructs with +1/-1 linker length increments further supported this cyclical trend with peaks at +5 and -5 shifts, and troughs at the -10, 0 and +10 shifts. The linker length between the terminal and subterminal motifs is 41 bp, which suggests that the cyclical effect of the linker length corresponds with the rotational phase between the terminal and subterminal motifs on a DNA double-helix (1 rotation ~10.5 bp) (Figure 20). When the motifs are off-

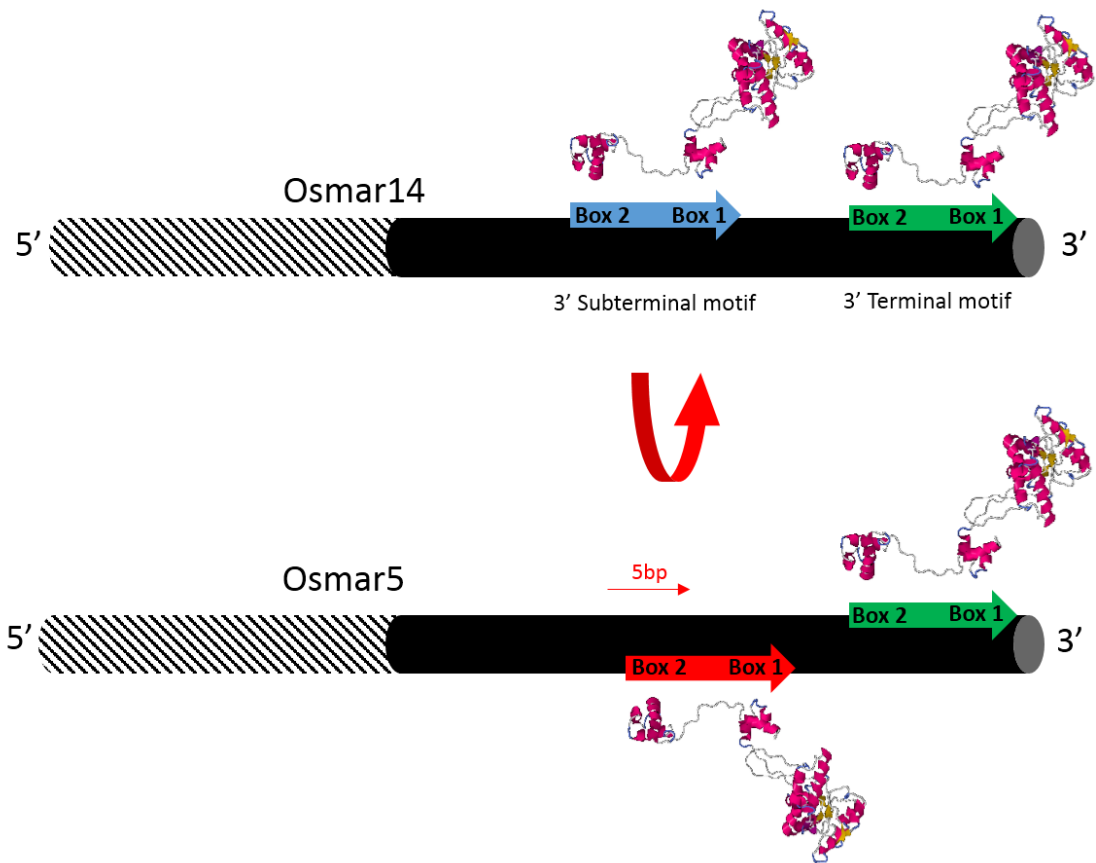


Figure 20 Illustration of the effect of the 5bp shift between *Osmar5* and *Osmar14*. Black bar, the 3' sequences of *Osmars*; the dashed region illustrates the 5' end. 5bp shifts of the subterminal motif changes the rotational phase relative to the 3' terminal motif. Top, in phase, *Osmar14*; Bottom, out of phase, *Osmar5* relative arrangement of the binding motifs.

phase (+5 and -5 shifts), the excision frequency is dramatically increased, and when the motifs are in-phase (0, -10, +10 shifts) the excision frequency drops to the minima.

### III-4 Discussion

Transposases carry out transposition reactions and mobilize DNA elements. *Osmar* transposases can be computationally modeling based on architecture of *Mos1* transposase. The predicted tertiary structure contains a bipartite DNA-binding helix-turn-helix domain (Figure 20). This bipartite structure is consistent with the dual DNA binding boxes in the *Osmar* 3' sub-terminal motif and both terminal motifs. Analysis of the helix-turn-helix motifs of the *Osmar5* transposase suggested that HTH1 was crucial for the specific binding to the element terminal transposase binding sites while HTH2 enhanced the overall binding with the Box 1 and Box 2 [127]. It should also be noted that the HTH2 motif in *Mos1* transposase was likely responsible for dimerization [29].

The 3' subterminal box 2 in *Osmar14NAS* has much higher transposase binding affinity than subterminal box 1, although subterminal box1 has a much stronger repressive effect. This is consistent with the notion that HTH1 would facilitate specific binding of transposase to box 2, while HTH2 positions the catalytic domain towards to the element terminus via interaction with box 1.

As shown by EMSA, binding to the sub-terminal and the terminal motifs are not exclusive (Figure 16), allowing three transposases to simultaneously bind to the full-length *Osmar* DNA element. In *Osmar14*, 3' terminal motif is 41 bp away from the sub-terminal motif and the DNA double helix is ~10.5 base pairs per turn. The two binding motifs on the 3' *Osmar14* are

nearly perfectly in-phase and the **sub**terminally-bound transposase may sterically clash with the terminally-bound transposase. This clash may prevent the formation of the pair-end complex between the 5' and 3' terminal transposases (Figure 20), effectively repress transposition. In addition, the repressive effect is additive when the repeat boxes are introduced to the 5' sub-terminal, as shown by construct 5'subTIR which obliterated transposition.

In contrast to *Osmar14*, *Osmar5* has the two 3' binding motifs 36 bps apart, nearly perfectly out-of-phase. When transposases bind simultaneously onto the two binding sites on the 3' of *Osmar5*, the transposase are placed on the opposite faces of the double helix. This arrangement results in higher *Osmar5* transposition activity than wild type *Osmar14*. The sub-terminally bound transposase is exposed to the 3' sequence end, and may carry out cleavage. Alternatively, it may attract the 5' transposase and assist in the formation of the paired-end synapse. Therefore, the sub-terminal motif can either repress or enhance transposition efficiency depending on its rotational relationship to the 3' terminal motif. This may have significant evolutionary impact on the TE family.

TEs play a major role in reshaping the host genomic landscapes.

Hyperactivity of TEs introduces mutations and may be harmful for hosts.

Therefore, hosts evolved and adopted inhibitory mechanisms to suppress activity of its TE populations [25, 99]. TEs themselves may have evolved to autoregulate their activities to escaped silencing. The sub-terminal binding site on *Osmar* sequences may be the result of selection imposed by the host

defense mechanisms, whereas the relative position of the sub-terminal motif to the terminal motif tunes the transposition activity of *Osmars*. This limits the rate of transposition, and allows a basal level of activity by the *Osmars* to sustain in the host genome.

## Chapter IV: Concluding Remarks

TEs are a major dynamic force for genomic evolution. They are diverse and abundant in eukaryotic genomes. One aspect of TEs is their potential to horizontally transmit to other genomes. HTT is supported by ample amount of evidence and it may play a major role in shaping the evolutionary history [45]. Once a TE escapes its native host and invades an uncolonized genome, it may initiate a new life cycle. The founding copy amplifies in the new genome, and it may continue to grow until mutations accumulate and host defense systems start to inactivate TE activity. TEs are subject to inhibitory mechanisms because unchecked accumulation of TEs are likely to cause deleterious effects to the host [25]. TEs have evolved to minimize harm to the host through targeting specific genomic regions [25], or in the case of *Osmars* to autoregulate activity through molecular arrangement of secondary competitive transposase binding motifs. The vast abundance of TEs is a testament to their persistence to survive and coexist with host genomes. Studying molecular mechanisms that allow transposable elements to persist will help us paint a clearer picture of genomic evolution.

## References

1. Liu, Y. and G. Yang, *Tc1-like transposable elements in plant genomes*. Mobile DNA, 2014. 5(1): p. 17.
2. Feschotte, C. and E.J. Pritham, *DNA transposons and the evolution of eukaryotic genomes*. Annu Rev Genet, 2007. 41: p. 331-68.
3. Yuan, Y. and S. Wessler, *The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies*. Proc Natl Acad Sci U S A, 2011. 108: p. 7884 - 7889.
4. Wicker, T., et al., *A unified classification system for eukaryotic transposable elements*. Nat Rev Genet, 2007. 8: p. 973 - 982.
5. McClintock, B., *Controlling elements and the gene*. Cold Spring Harb Symp Quant Biol, 1956. 21: p. 197-216.
6. Mc, C.B., *The origin and behavior of mutable loci in maize*. Proc Natl Acad Sci U S A, 1950. 36(6): p. 344-55.
7. Shapiro, J.A., *Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli*. J Mol Biol, 1969. 40(1): p. 93-105.
8. Finnegan, D.J., *Transposable elements*. Curr Opin Genet Dev, 1992. 2(6): p. 861-7.
9. Biemont, C., *A brief history of the status of transposable elements: from junk DNA to major players in evolution*. Genetics, 2010. 186(4): p. 1085-93.
10. Picard, G., et al., *Non-mendelian female sterility and hybrid dysgenesis in Drosophila melanogaster*. Genet Res, 1978. 32(3): p. 275-87.
11. Rubin, G.M., M.G. Kidwell, and P.M. Bingham, *The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations*. Cell, 1982. 29(3): p. 987-94.
12. Engels, W.R., *The P Family of Transposable Elements in Drosophila*. Annual Review of Genetics, 1983. 17(1): p. 315-344.
13. Rubin, G.M. and A.C. Spradling, *Genetic transformation of Drosophila with transposable element vectors*. Science, 1982. 218(4570): p. 348-53.
14. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics*. Science, 2009. 326(5956): p. 1112-5.
15. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution*. Nat Rev Genet, 2009. 10(10): p. 691-703.

16. Wicker, T., et al., *A unified classification system for eukaryotic transposable elements*. Nat Rev Genet, 2007. 8(12): p. 973-82.
17. Kapitonov, V.V. and J. Jurka, *A universal classification of eukaryotic transposable elements implemented in Repbase*. Nat Rev Genet, 2008. 9(5): p. 411-2; author reply 414.
18. Jurka, J., et al., *Repetitive sequences in complex genomes: structure and evolution*. Annu Rev Genomics Hum Genet, 2007. 8: p. 241-59.
19. Yuan, Y.W. and S.R. Wessler, *The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies*. Proc Natl Acad Sci U S A, 2011. 108(19): p. 7884–7889.
20. Capy, P., *Classification and nomenclature of retrotransposable elements*. Cytogenet Genome Res, 2005. 110(1-4): p. 457-61.
21. Varmus, H., *Retroviruses*. Science, 1988. 240(4858): p. 1427-35.
22. Wright, D.A. and D.F. Voytas, *Potential retroviruses in plants: Tat1 is related to a group of Arabidopsis thaliana Ty3/gypsy retrotransposons that encode envelope-like proteins*. Genetics, 1998. 149(2): p. 703-15.
23. Song, S.U., et al., *An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus*. Genes Dev, 1994. 8(17): p. 2046-57.
24. Kazazian, H.H., Jr. and J.V. Moran, *The impact of L1 retrotransposons on the human genome*. Nat Genet, 1998. 19(1): p. 19-24.
25. Levin, H.L. and J.V. Moran, *Dynamic interactions between transposable elements and their hosts*. Nat Rev Genet, 2011. 12(9): p. 615-27.
26. Craig, N.L., *Mobile DNA II*. 2002, Washington, D.C.: ASM Press. xviii, 1204 p., 32 p. of plates.
27. Ding, S., et al., *Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice*. Cell, 2005. 122(3): p. 473-83.
28. Medhora, M., K. Maruyama, and D.L. Hartl, *Molecular and functional analysis of the mariner mutator element Mos1 in Drosophila*. Genetics, 1991. 128(2): p. 311-8.
29. Richardson, J., et al., *Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote*. Cell, 2009. 138: p. 1096 - 1108.
30. Hickman, A.B., et al., *Molecular architecture of a eukaryotic DNA transposase*. Nat Struct Mol Biol, 2005. 12(8): p. 715-21.

31. Feschotte, C., X. Zhang, and S.R. Wessler, *Miniature inverted-repeat transposable elements (MITEs) and their relationship with established DNA transposons*. *Mobile DNA II*, 2002: p. 1147-1158.
32. Kazazian, H.H., Jr., *Mobile elements: drivers of genome evolution*. *Science*, 2004. 303(5664): p. 1626-32.
33. Dufresne, M., et al., *Transposition of a fungal miniature inverted-repeat transposable element through the action of a Tc1-like transposase*. *Genetics*, 2007. 175(1): p. 441-52.
34. Yang, G., et al., *Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE*. *Science*, 2009. 325(5946): p. 1391-4.
35. Fattash, I., et al., *Miniature Inverted-repeat Transposable Elements (MITEs): discovery, distribution and activity*. *Genome*, 2013. 56: p. 475 - 486.
36. Bureau, T.E. and S.R. Wessler, *Tourist: a large family of small inverted repeat elements frequently associated with maize genes*. *Plant Cell*, 1992. 4(10): p. 1283-94.
37. Bureau, T.E. and S.R. Wessler, *Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants*. *Plant Cell*, 1994. 6(6): p. 907-16.
38. Feschotte, C. and E. Pritham, *DNA transposons and the evolution of eukaryotic genomes*. *Annu Rev Genet*, 2007. 41: p. 331 - 368.
39. Plasterk, R., Z. Izsvak, and Z. Ivics, *Resident aliens: the Tc1/mariner superfamily of transposable elements*. *Trends Genet*, 1999. 15: p. 326 - 332.
40. Eide, D. and P. Anderson, *Transposition of Tc1 in the nematode *Caenorhabditis elegans**. *Proc Natl Acad Sci U S A*, 1985. 82(6): p. 1756-60.
41. Jacobson, J., M. Medhora, and D. Hartl, *Molecular structure of a somatically unstable transposable element in *Drosophila**. *Proc Natl Acad Sci U S A*, 1986. 83: p. 8684 - 8688.
42. Doak, T., et al., *A proposed superfamily of transposase genes - transposon-like elements in ciliated protozoa and a common D35e motif*. *Proc Natl Acad Sci U S A*, 1994. 91: p. 942 - 946.
43. Feschotte, C. and S. Wessler, *Mariner-like transposases are widespread and diverse in flowering plants*. *Proc Natl Acad Sci U S A*, 2002. 99: p. 280 - 285.

44. Shao, H. and Z. Tu, *Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons*. Genetics, 2001. 159(3): p. 1103-15.
45. Schaack, S., C. Gilbert, and C. Feschotte, *Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution*. Trends Ecol Evol, 2010. 25(9): p. 537-46.
46. Gilbert, C., et al., *A role for host-parasite interactions in the horizontal transfer of transposons across phyla*. Nature, 2010. 464(7293): p. 1347-50.
47. Radice, A.D., et al., *Widespread occurrence of the Tc1 transposon family: Tc1-like transposons from teleost fish*. Mol Gen Genet, 1994. 244(6): p. 606-12.
48. Collins, J., E. Forbes, and P. Anderson, *The Tc3 family of transposable genetic elements in Caenorhabditis elegans*. Genetics, 1989. 121(1): p. 47-55.
49. Ruan, K. and S. Emmons, *Precise and imprecise somatic excision of the transposon Tc1 in the nematode C-elegans*. Nucleic Acids Res, 1987. 15: p. 6875 - 6881.
50. Pavlopoulos, A., et al., *The DNA transposon Minos as a tool for transgenesis and functional genomic analysis in vertebrates and invertebrates*. Genome Biol, 2007. Suppl 1: p. S2.
51. Carr, P., et al., *The transposon impala is activated by low temperatures: use of a controlled transposition system to identify genes critical for viability of Aspergillus fumigatus*. Eukaryot Cell, 2010. 9: p. 438 - 448.
52. Hua-Van, A., T. Langin, and M. Daboussi, *Aberrant transposition of a Tc1-mariner element, impala, in the fungus Fusarium oxysporum*. Mol Genet Genomics, 2002. 267: p. 79 - 87.
53. Ivics, Z., et al., *Molecular reconstruction of Sleeping beauty, a Tc1-like transposon from fish, and its transposition in human cells*. Cell, 1997. 91: p. 501 - 510.
54. Daniels, S.B., et al., *Evidence for horizontal transmission of the P transposable element between Drosophila species*. Genetics, 1990. 124(2): p. 339-55.
55. Houck, M.A., et al., *Possible horizontal transfer of Drosophila genes by the mite Proctolaelaps regalis*. Science, 1991. 253(5024): p. 1125-8.

56. Silva, J.C., E.L. Loreto, and J.B. Clark, *Factors that affect the horizontal transfer of transposable elements*. *Curr Issues Mol Biol*, 2004. 6(1): p. 57-71.
57. Rubin, E.J., et al., *In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria*. *Proc Natl Acad Sci U S A*, 1999. 96(4): p. 1645-50.
58. Piskurek, O. and N. Okada, *Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals*. *Proc Natl Acad Sci U S A*, 2007. 104(29): p. 12046-51.
59. Fraser, M.J., G.E. Smith, and M.D. Summers, *Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses*. *J Virol*, 1983. 47(2): p. 287-300.
60. Dunning Hotopp, J.C., et al., *Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes*. *Science*, 2007. 317(5845): p. 1753-6.
61. Raychoudhury, R., et al., *Modes of acquisition of Wolbachia: horizontal transfer, hybrid introgression, and codivergence in the Nasonia species complex*. *Evolution*, 2009. 63(1): p. 165-83.
62. Klasson, L., et al., *Horizontal gene transfer between Wolbachia and the mosquito Aedes aegypti*. *BMC Genomics*, 2009. 10: p. 33.
63. Yoshiyama, M., et al., *Possible horizontal transfer of a transposable element from host to parasitoid*. *Mol Biol Evol*, 2001. 18(10): p. 1952-8.
64. Biemont, C. and C. Vieira, *Genetics: junk DNA as an evolutionary force*. *Nature*, 2006. 443(7111): p. 521-4.
65. Bourque, G., *Transposable elements in gene regulation and in the evolution of vertebrate genomes*. *Curr Opin Genet Dev*, 2009. 19(6): p. 607-12.
66. Brookfield, J.F., *The ecology of the genome - mobile DNA elements and their hosts*. *Nat Rev Genet*, 2005. 6(2): p. 128-36.
67. Chenais, B., et al., *The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments*. *Gene*, 2012. 509(1): p. 7-15.
68. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements*. *Genome Res*, 2008. 18(11): p. 1752-62.

69. Majumdar, S., A. Singh, and D.C. Rio, *The human THAP9 gene encodes an active P-element DNA transposase*. *Science*, 2013. 339(6118): p. 446-8.
70. Volff, J.N., *Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes*. *Bioessays*, 2006. 28(9): p. 913-22.
71. Miller, W.J., et al., *Molecular domestication--more than a sporadic episode in evolution*. *Genetica*, 1999. 107(1-3): p. 197-207.
72. Kodama, K., S. Takagi, and A. Koga, *The Toll element of the medaka fish, a member of the hAT transposable element family, jumps in Caenorhabditis elegans*. *Heredity (Edinb)*, 2008. 101(3): p. 222-7.
73. Weil, C.F. and R. Kunze, *Transposition of maize Ac/Ds transposable elements in the yeast Saccharomyces cerevisiae*. *Nat Genet*, 2000. 26(2): p. 187-90.
74. Yang, G., C.F. Weil, and S.R. Wessler, *A rice Tc1/mariner-like element transposes in yeast*. *Plant Cell*, 2006. 18(10): p. 2469-78.
75. Emelyanov, A., et al., *Trans-kingdom transposition of the maize dissociation element*. *Genetics*, 2006. 174(3): p. 1095-104.
76. Evertts, A.G., et al., *The hermes transposon of Musca domestica is an efficient tool for the mutagenesis of Schizosaccharomyces pombe*. *Genetics*, 2007. 177(4): p. 2519-23.
77. Zhang, J.K., et al., *In vivo transposon mutagenesis of the methanogenic archaeon Methanosarcina acetivorans C2A using a modified version of the insect mariner-family transposable element Himar1*. *Proc Natl Acad Sci U S A*, 2000. 97(17): p. 9665-70.
78. Yang, G., *MAK, a computational tool kit for automated MITE analysis*. *Nucleic Acids Research*, 2003. 31(13): p. 3659-3665.
79. Lang, D., et al., *Representation and high-quality annotation of the Physcomitrella patens transcriptome demonstrates a high proportion of proteins involved in metabolism in mosses*. *Plant Biol (Stuttg)*, 2005. 7: p. 238 - 250.
80. Yang, G., *MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements*. *BMC Bioinformatics*, 2013. 14: p. 186.
81. Rensing, S., et al., *The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants*. *Science*, 2008. 319: p. 64 - 69.

82. Ochman, H. and A.C. Wilson, *Evolution in bacteria: evidence for a universal substitution rate in cellular genomes*. J Mol Evol, 1987. 26(1-2): p. 74-86.
83. Lohe, A. and D. Hartl, *Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation*. Mol Biol Evol, 1996. 13: p. 549 - 555.
84. Chen, S. and X. Li, *Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome P450 genes*. BMC Evol Biol, 2007. 7: p. 46.
85. Wu, J., et al., *The genome of the pear (Pyrus bretschneideri Rehd.)*. Genome Res, 2013. 23(2): p. 396-408.
86. You, M., et al., *A heterozygous moth genome provides insights into herbivory and detoxification*. Nat Genet, 2013. 45(2): p. 220-5.
87. International Barley Genome Sequencing, C., et al., *A physical, genetic and functional sequence assembly of the barley genome*. Nature, 2012. 491(7426): p. 711-6.
88. Truco, M.J., et al., *An Ultra High-Density, Transcript-Based, Genetic Map of Lettuce*. G3 (Bethesda), 2013.
89. Hansen, A.K., C. Vorburger, and N.A. Moran, *Genomic basis of endosymbiont-conferred protection against an insect parasitoid*. Genome Res, 2012. 22(1): p. 106-14.
90. Loreto, E.L., C.M. Carareto, and P. Capy, *Revisiting horizontal transfer of transposable elements in Drosophila*. Heredity (Edinb), 2008. 100(6): p. 545-54.
91. Diao, X., M. Freeling, and D. Lisch, *Horizontal transfer of a plant transposon*. PLoS Biol, 2006. 4(1): p. e5.
92. Chen, D., et al., *Global transcriptional responses of fission yeast to environmental stress*. Mol Biol Cell, 2003. 14(1): p. 214-29.
93. Grandbastien, M.A., et al., *Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae*. Cytogenet Genome Res, 2005. 110(1-4): p. 229-41.
94. Dai, J., et al., *Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin*. Mol Cell, 2007. 27(2): p. 289-99.
95. McClintock, B., *The significance of responses of the genome to challenge*. Science, 1984. 226(4676): p. 792-801.

96. Batut, P., et al., *High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression*. Genome Res, 2013. 23(1): p. 169-80.
97. Cordaux, R., et al., *Birth of a chimeric primate gene by capture of the transposase gene from a mobile element*. Proc Natl Acad Sci U S A, 2006. 103(21): p. 8101-6.
98. Le Rouzic, A. and P. Capy, *The first steps of transposable elements invasion: parasitic strategy vs. genetic drift*. Genetics, 2005. 169(2): p. 1033-43.
99. Venner, S., C. Feschotte, and C. Biemont, *Dynamics of transposable elements: towards a community ecology of the genome*. Trends Genet, 2009. 25(7): p. 317-23.
100. Goodier, J.L. and H.H. Kazazian, Jr., *Retrotransposons revisited: the restraint and rehabilitation of parasites*. Cell, 2008. 135(1): p. 23-35.
101. Morrish, T.A., et al., *Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres*. Nature, 2007. 446(7132): p. 208-12.
102. Gladyshev, E.A. and I.R. Arkhipova, *Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes*. Proc Natl Acad Sci U S A, 2007. 104(22): p. 9352-7.
103. Gai, X. and D.F. Voytas, *A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin*. Mol Cell, 1998. 1(7): p. 1051-5.
104. Xie, W., et al., *Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p*. Mol Cell Biol, 2001. 21(19): p. 6606-14.
105. Brady, T.L., et al., *Retrotransposon target site selection by imitation of a cellular protein*. Mol Cell Biol, 2008. 28(4): p. 1230-9.
106. Kuduvalli, P.N., J.E. Rao, and N.L. Craig, *Target DNA structure plays a critical role in Tn7 transposition*. EMBO J, 2001. 20(4): p. 924-32.
107. Bellen, H.J., et al., *The Drosophila gene disruption project: progress using transposons with distinctive site specificities*. Genetics, 2011. 188(3): p. 731-43.
108. Bushman, F.D., *Targeting survival: integration site selection by retroviruses and LTR-retrotransposons*. Cell, 2003. 115(2): p. 135-8.

109. Lesage, P. and A.L. Todeschini, *Happy together: the life and times of Ty retrotransposons and their hosts*. Cytogenet Genome Res, 2005. 110(1-4): p. 70-90.
110. Le Rouzic, A., S. Dupas, and P. Capy, *Genome ecosystem and transposable elements species*. Gene, 2007. 390(1-2): p. 214-20.
111. Slotkin, R.K. and R. Martienssen, *Transposable elements and the epigenetic regulation of the genome*. Nat Rev Genet, 2007. 8(4): p. 272-85.
112. Hollister, J.D. and B.S. Gaut, *Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression*. Genome Res, 2009. 19(8): p. 1419-28.
113. Alves, G., A. Tatro, and T. Fanning, *Differential methylation of human LINE-1 retrotransposons in malignant cells*. Gene, 1996. 176(1-2): p. 39-44.
114. Kuramochi-Miyagawa, S., et al., *DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes*. Genes Dev, 2008. 22(7): p. 908-17.
115. Selker, E.U., et al., *The methylated component of the Neurospora crassa genome*. Nature, 2003. 422(6934): p. 893-7.
116. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites*. Trends Genet, 1997. 13(8): p. 335-40.
117. Malone, C.D. and G.J. Hannon, *Small RNAs as guardians of the genome*. Cell, 2009. 136(4): p. 656-68.
118. van Rij, R.P. and E. Berezikov, *Small RNAs and the control of transposons and viruses in Drosophila*. Trends Microbiol, 2009. 17(4): p. 163-71.
119. Ghildiyal, M. and P.D. Zamore, *Small silencing RNAs: an expanding universe*. Nat Rev Genet, 2009. 10(2): p. 94-108.
120. Huda, A., L. Marino-Ramirez, and I.K. Jordan, *Epigenetic histone modifications of human transposable elements: genome defense versus exaptation*. Mob DNA, 2010. 1(1): p. 2.
121. Tsukahara, S., et al., *Bursts of retrotransposition reproduced in Arabidopsis*. Nature, 2009. 461(7262): p. 423-6.
122. Czech, B. and G.J. Hannon, *Small RNA sorting: matchmaking for Argonautes*. Nat Rev Genet, 2011. 12(1): p. 19-31.

123. Aravin, A.A., G.J. Hannon, and J. Brennecke, *The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race*. Science, 2007. 318(5851): p. 761-4.
124. Sijen, T. and R.H. Plasterk, *Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi*. Nature, 2003. 426(6964): p. 310-4.
125. Brennecke, J., et al., *An epigenetic role for maternally inherited piRNAs in transposon silencing*. Science, 2008. 322(5906): p. 1387-92.
126. Yang, N. and H.H. Kazazian, Jr., *L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells*. Nat Struct Mol Biol, 2006. 13(9): p. 763-71.
127. Feschotte, C., et al., *DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs*. Nucleic Acids Res, 2005. 33(7): p. 2153-65.
128. Richardson, J.M., et al., *Mechanism of Mos1 transposition: insights from structural analysis*. EMBO J, 2006. 25(6): p. 1324-34.

## Appendix

**Table A1 Assembled transcripts of *PpTc1* from P0409 database**

Assembled Transcript name, location and EST	Bit Score
Pp_vt0409_16265_19780 cluster 22612 subcluster 16265 assembled from 38 ESTs	1359
Pp_vt0409_2338_2851 cluster 15845 subcluster 2338 assembled from 41 ESTs	1341
Pp_vt0409_26739_32467 cluster 17084 subcluster 26739 assembled from 31 ESTs	1338
Pp_vt0409_22484_27308 cluster 1901 subcluster 22484 assembled from 41 ESTs	1249
Pp_vt0409_21190_25764 cluster 29325 subcluster 21190 assembled from 41 ESTs	1245
Pp_vt0409_16198_19697 cluster 18030 subcluster 16198 assembled from 41 ESTs	1245
Pp_vt0409_2135_2608 cluster 21995 subcluster 2135 assembled from 38 ESTs	1245
Pp_vt0409_33891_41168 cluster 24259 subcluster 33891 assembled from 41 ESTs	1241
Pp_vt0409_23089_28041 cluster 31877 subcluster 23089 assembled from 32 ESTs	1241
Pp_vt0409_21080_25616 cluster 25885 subcluster 21080 assembled from 41 ESTs	1241
Pp_vt0409_39081_47530 cluster 32726 subcluster 39081 assembled from 40 ESTs	1237
Pp_vt0409_28788_34955 cluster 30767 subcluster 28788 assembled from 40 ESTs	1237
Pp_vt0409_15034_18253 cluster 20439 subcluster 15034 assembled from 39 ESTs	1237
Pp_vt0409_9611_11695 cluster 31496 subcluster 9611 assembled from 40 ESTs	1237
Pp_vt0409_37469_45516 cluster 18478 subcluster 37469 assembled from 41 ESTs	1233
Pp_vt0409_35827_43506 cluster 15175 subcluster 35827 assembled from 39 ESTs	1233
Pp_vt0409_35135_42663 cluster 36286 subcluster 35135 assembled from 41 ESTs	1233
Pp_vt0409_23051_27987 cluster 31844 subcluster 23051 assembled from 9 ESTs	1233
Pp_vt0409_22179_26957 cluster 4371 subcluster 22179 assembled from 40 ESTs	1233
Pp_vt0409_1848_2267 cluster 27757 subcluster 1848 assembled from 41 ESTs	1233
Pp_vt0409_1285_1593 cluster 15715 subcluster 1285 assembled from 40 ESTs	1233
Pp_vt0409_7424_9044 cluster 29069 subcluster 7424 assembled from 39 ESTs	1229
Pp_vt0409_32884_39946 cluster 9817 subcluster 32884 assembled from 43 ESTs	1227
Pp_vt0409_26369_32030 cluster 20169 subcluster 26369 assembled from 37 ESTs	1225
Pp_vt0409_19520_23734 cluster 17132 subcluster 19520 assembled from 41 ESTs	1225
Pp_vt0409_1626_1984 cluster 27110 subcluster 1626 assembled from 39 ESTs	1225
Pp_vt0409_33531_40740 cluster 26180 subcluster 33531 assembled from 38 ESTs	1221
Pp_vt0409_21894_26610 cluster 19939 subcluster 21894 assembled from 41 ESTs	1221
Pp_vt0409_11044_13402 cluster 1585 subcluster 11044 assembled from 39 ESTs	1221
Pp_vt0409_20200_24553 cluster 11512 subcluster 20200 assembled from 40 ESTs	1220
Pp_vt0409_34556_42002 cluster 7898 subcluster 34556 assembled from 38 ESTs	1217
Pp_vt0409_26550_32238 cluster 27008 subcluster 26550 assembled from 41 ESTs	1217
Pp_vt0409_14579_17704 cluster 13297 subcluster 14579 assembled from 39 ESTs	1217
Pp_vt0409_19255_23405 cluster 10743 subcluster 19255 assembled from 39 ESTs	1213

Pp_vt0409_16561_20145 cluster 4038 subcluster 16561 assembled from 38 ESTs	1213
Pp_vt0409_1472_1815 cluster 22049 subcluster 1472 assembled from 39 ESTs	1213
Pp_vt0409_20074_24398 cluster 19526 subcluster 20074 assembled from 38 ESTs	1211
Pp_vt0409_37104_45081 cluster 30907 subcluster 37104 assembled from 39 ESTs	1210
Pp_vt0409_1113_1380 cluster 7162 subcluster 1113 assembled from 39 ESTs	1209
Pp_vt0409_38597_46914 cluster 24136 subcluster 38597 assembled from 40 ESTs	1205
Pp_vt0409_9993_12147 cluster 14198 subcluster 9993 assembled from 35 ESTs	1205
Pp_vt0409_23995_29128 cluster 33874 subcluster 23995 assembled from 36 ESTs	1202
Pp_vt0409_34390_41799 cluster 7501 subcluster 34390 assembled from 38 ESTs	1201
Pp_vt0409_17556_21346 cluster 31700 subcluster 17556 assembled from 11 ESTs	1201
Pp_vt0409_10895_13223 cluster 7353 subcluster 10895 assembled from 39 ESTs	1201
Pp_vt0409_6004_7319 cluster 19776 subcluster 6004 assembled from 38 ESTs	1201
Pp_vt0409_1750_2133 cluster 28426 subcluster 1750 assembled from 33 ESTs	1197
Pp_vt0409_678_832 cluster 28649 subcluster 678 assembled from 32 ESTs	1197
Pp_vt0409_31034_37709 cluster 22321 subcluster 31034 assembled from 36 ESTs	1193
Pp_vt0409_3720_4540 cluster 18902 subcluster 3720 assembled from 30 ESTs	1193
Pp_vt0409_26347_32004 cluster 20147 subcluster 26347 assembled from 5 ESTs	1191
Pp_vt0409_9341_11375 cluster 4647 subcluster 9341 assembled from 38 ESTs	1190
Pp_vt0409_15674_19023 cluster 27182 subcluster 15674 assembled from 38 ESTs	1181
Pp_vt0409_11072_13431 cluster 1612 subcluster 11072 assembled from 32 ESTs	1181
Pp_vt0409_26950_32707 cluster 3648 subcluster 26950 assembled from 39 ESTs	1178
Pp_vt0409_3912_4766 cluster 15894 subcluster 3912 assembled from 34 ESTs	1177
Pp_vt0409_31035_37710 cluster 22322 subcluster 31035 assembled from 23 ESTs	1173
Pp_vt0409_25191_30587 cluster 2622 subcluster 25191 assembled from 30 ESTs	1173
Pp_vt0409_33491_40693 cluster 19861 subcluster 33491 assembled from 19 ESTs	1171
Pp_vt0409_1802_2193 cluster 27714 subcluster 1802 assembled from 31 ESTs	1169
Pp_vt0409_4888_5970 cluster 17544 subcluster 4888 assembled from 17 ESTs	1168
Pp_vt0409_32857_39915 cluster 9794 subcluster 32857 assembled from 31 ESTs	1158
Pp_vt0409_12656_15338 cluster 2562 subcluster 12656 assembled from 31 ESTs	1157
Pp_vt0409_36169_43920 cluster 34975 subcluster 36169 assembled from 29 ESTs	1153
Pp_vt0409_22676_27540 cluster 13891 subcluster 22676 assembled from 21 ESTs	1153
Pp_vt0409_3601_4394 cluster 35237 subcluster 3601 assembled from 22 ESTs	1149
Pp_vt0409_4237_5175 cluster 10268 subcluster 4237 assembled from 6 ESTs	1145
Pp_vt0409_6281_7637 cluster 17300 subcluster 6281 assembled from 24 ESTs	1115
Pp_vt0409_4252_5191 cluster 10280 subcluster 4252 assembled from 17 ESTs	1111
Pp_vt0409_26543_32231 cluster 27001 subcluster 26543 assembled from 29 ESTs	1089
Pp_vt0409_2382_2904 cluster 3505 subcluster 2382 assembled from 2 ESTs	1012
Pp_vt0409_29499_35849 cluster 10612 subcluster 29499 assembled from 2 ESTs	971
Pp_vt0409_33322_40483 cluster 2975 subcluster 33322 assembled from 2 ESTs	961

**Table A2 *PpTc1* assembled transcripts that produce a conceptual full-length transposase bearing DD34E motifs.**

<b>Accession</b>	<b>#EST evidence</b>	<b>Scaffold</b>	<b>Start</b>	<b>end</b>	<b>length</b>
Pp_vt0409_37469_45516	41	89	326953	328314	1362
Pp_vt0409_1626_1984	39	107	912421	913782	1362
Pp_vt0409_33531_40740	38	67	526593	527954	1362
Pp_vt0409_23089_28041	32	34	407831	409188	1358
Pp_vt0409_3601_4394	22	121	304597	305940	1344
Pp_vt0409_26347_32004	5	4	361588 6	361720 1	1316
Pp_vt0409_33891_41168	41	69	294398	295759	1362
Pp_vt0409_33322_40483	2	65	181156 4	181267 3	1110
Pp_vt0409_15034_18253	39	226	98064	99425	1362

**Table A3 Full-length plant TLEs with intact TIRs**

<b>Element</b>	<b>Organism</b>	<b>Accession</b>	<b>Start</b>	<b>End</b>	<b>#full-length</b>	<b>TIR</b>
<i>PpTc1</i>	<i>Physcomitrella patens</i>	ABEU01007491	6844	8431	75	CAGTGACAAACA AAACCGAGTACA AAATCTGAA
<i>PpTc2</i>	<i>Physcomitrella patens</i>	ABEU01006878	161594	163306	20	CAGTGGGGTACA GAAATAATTCGA ATTTTTTTC
<i>OsTc1</i>	<i>Oryza sativa Indica</i>	AAAA02041396	2423	3983	1	CACACATCAAAA GTGTCTAGGGATC AA
<i>Br-TLM</i>	<i>Brassica rapa</i>	AENI01000564	34190	34387	147	CAGTGAAACCTCT ATAAATTAATA
<i>St-TLM</i>	<i>Solanum tuberosum</i>	AEWC01004763	8450	8599	9	CAGTCATACCTCT CTATAACA

**Table A4 *Tc1*-like transposases described in this study.**

Element	Organism	Accession	ORF start	ORF end	Complete DD34E Triad?
<b>Plant</b>					
<i>PpTc1</i>	<i>Physcomitrella patens</i>	ABEU01007491	7186	8199	Y
<i>PpTc2</i>	<i>Physcomitrella patens</i>	ABEU01006878	162826	161813	Y
<i>OsTc1</i>	<i>Oryza sativa Indica</i>	AAAA02041396	3821	2697	Y
<i>BnTc1</i>	<i>Betula nana</i>	CAOK01056615	1484	1978	N
<i>BnTc2</i>	<i>Betula nana</i>	CAOK01550459	168	1214	Y
<i>BnTc3</i>	<i>Betula nana</i>	CAOK01014729	14272	14472	N
<i>BnTc4</i>	<i>Betula nana</i>	CAOK01486111	2	244	N
<i>BrTc1</i>	<i>Brassica rapa</i>	AENI01020305	162	572	N
<i>BrTc2</i>	<i>Brassica rapa</i>	AENI01036930	17	328	N
<i>CsTc1</i>	<i>Cannabis sativa</i>	AGQN01308320	302	517	N
<i>HvTc1</i>	<i>Hordium vulgare</i>	CAJV010227559	1	1684	Y
<i>HvTc2</i>	<i>Hordium vulgare</i>	CAJV010272453	49	555	Y
<i>HvTc3</i>	<i>Hordium vulgare</i>	CAJV012609061	1716	2114	N
<i>HvTc4</i>	<i>Hordium vulgare</i>	CAJV011622646	1	222	N
<i>LsTc1</i>	<i>Lactuca sativa</i>	AFSA01593962	2	394	N
<i>LsTc2</i>	<i>Lactuca sativa</i>	AFSA01593962	87	485	N
<i>PtTc1</i>	<i>Populus Trichocarpa</i>	AARH01030986	1	714	Y
<i>PxbTc1</i>	<i>Pyrus x bretschneideri</i>	AJSU01007483	3055	3606	Y
<i>PxbTc2</i>	<i>Pyrus x bretschneideri</i>	AJSU01007483	3055	3606	N

<i>TuTc1</i>	<i>Triticum urartu</i>	AOTI010070343	376	1368	Y
<b>Non-plant</b>					
<i>AgTLE</i>	<i>Anopheles gambiae</i>	AAD03793	1	260	Y
<i>AlTLE</i>	<i>Albugo laibachii</i>	CCA18549	1	255	Y
<i>CaTLE</i>	<i>Crotalus adamanteus</i>	AFJ51821	1	325	Y
<i>CcTLE</i>	<i>Cyprinus carpio</i>	AET85182	1	332	Y
<i>DvTLE</i>	<i>Drosophila virilis</i>	AAA88882	1	348	Y
<i>HmTLE</i>	<i>Hydra magnipapillata</i>	ABRM01017102	12567	13541	Y
<i>Impala</i>	<i>Fusarium oxysporum</i>	AGBI01000009.1	69358	70380	Y
<i>MsTLE</i>	<i>Micromonas sp. RCC299</i>	XP_002506953	1	368	Y
<i>RcTLE</i>	<i>Rana catesbeiana</i>	ACO51862	1	334	Y
<i>RdTLE</i>	<i>Rhizopus delemar</i>	AACW02000171	5110	5613	N
<i>Tc1</i>	<i>Caenorhabditis elegans</i>	P03934	1	273	Y
<i>TvTLE</i>	<i>Trametes versicolor</i>	AEJI01000976	281	1297	Y
<i>VcTLE</i>	<i>Volvox carteri f. nagariensis</i>	ACJH01008053	1817	1945	N
<i>WeTLE</i>	<i>Wolbachia endosymbiont</i>	AAGB01000014	2003	4914	Y

**Table A5 Oligo sequences for Osmar14 EMSA experiments**

Oligos:	Sequence
For-6FAM-Osm14NAS-3'ter75bp Mut-subT-Box2	5'/56FAM/TGCACGATTTAAATTTTTTTTGGACAAAGTCTCCCCCCCCCCCCCAGAAATCGCTTCTTTCTGGGACGGAGGGAG-3'
Rev-Osm14NAS-3'ter75bp Mut-subT-Box2	5'CTCCCTCCGTCCCAGAAAGAAGCGATTTCTGGGGGGGGGGGGGAGACTTTGTCCCAAAAAAAAAATTTAAATCGTGCA-3'
For-6FAM-Osm14NAS-3'ter75bp Mut-subT-Box1	5'/56FAM/TGCACGAATCGCTTTTTTTTTTTTAAAAAAGTCTCCCCCCCCCCCCCAGAAATCGCTTCTTTCTGGGACGGAGGGAG-3'
Rev-Osm14NAS-3'ter75bp Mut-subT-Box1	5'CTCCCTCCGTCCCAGAAAGAAGCGATTTCTGGGGGGGGGGGGGAGACTTTTTTTAAAAAAAAAAAAGCGATTCGTGCA-3'
For-6FAM-Osm14NAS-3'ter75bp WT	5'/56FAM/TGCACGAATCGCTTTTTTTTTTTGGACAAAGTCTCCCCCCCCCCCCCAGAAATCGCTTCTTTCTGGGACGGAGGGAG-3'
Rev-Osm14NAS-3'ter75bp WT	5'CTCCCTCCGTCCCAGAAAGAAGCGATTTCTGGGGGGGGGGGGGAGACTTTGTCCCAAAAAAAAAAAGCGATTCGTGCA-3'
For-6FAM-Osm14NAS-3'ter75bp Mut-subT-Box1+2	5'/56FAM/TGCACGATTTAAATTTTTTTTTTTTAAAAAAGTCTCCCCCCCCCCCCCAGAAATCGCTTCTTTCTGGGACGGAGGGAG-3'
Rev-Osm14NAS-3'ter75bp Mut-subT-Box1+2	5'CTCCCTCCGTCCCAGAAAGAAGCGATTTCTGGGGGGGGGGGGGAGACTTTTTTTAAAAAAAAAAATTTAAATCGTGCA-3'