DD34E DNA Transposable Elements of Mosquitoes: Whole-genome survey, evolution, and transposition

Monique R. Coy

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Biochemistry

Zhijian (Jake) Tu, Chair
Zachary Adelman
Glenda Gillaspy
Peter J Kennelly
John McDowell

(13th of June 2007)
Blacksburg, Virginia

Keywords: DNA transposable element, evolution, horizontal transfer, *gambol,* interspersed repeat, mosquito, *Tc1*, transposition assay

DD34E DNA Transposable Elements of Mosquitoes: Whole-genome survey, evolution, and transposition

Monique Royer Coy

Abstract

Transposable elements (TEs) are mobile genetic elements capable of replicating and spreading within, and in some cases, between genomes. I describe a whole-genome analysis of DD34E TEs, which belong to the *IS630-Tc1-mariner* superfamily of DNA transposable elements, in the African malaria mosquito, *Anopheles gambiae*. Twenty-six new transposons as well as a new family, *gambol*, were identified. The *gambol* family shares the DD34E catalytic motif with *Tc1*-DD34E transposons, but is distinct from these elements in their phylogenetic relationships. Although *gambol* appears to be related to a few DD34E transposons from cyanobacteria and fungi, no *gambol* elements have been reported in any other insects or animals thus far. This discovery expands the already expansive diversity of the *IS630-Tc1-mariner* TEs, and raises interesting questions as to the origin of *gambol* elements and their apparent diversity in *An. gambiae*. Several DD34E transposons discovered in *An. gambiae* possess characteristics that are associated with recent transposition, such as high sequence identity between copies, and intact terminal-inverted repeats and open reading frames. One such element, *AgTango*, was also found in a distantly related mosquito species, *Aedes aegypti*, at high amino acid sequence identity (79.9%). It was discovered that *Tango* transposons have patchy distribution among twelve mosquito species surveyed using PCR as well as genomic searches, suggesting a possible case for horizontal transfer. Additionally, it was discovered that in some mosquito genomes, there are several *Tango* transposons. These observations suggest differential evolutionary scenarios and/or TE-host interaction of *Tango* elements between mosquito species. This strengthened the case that *AgTango* may be a functional transposase, and I sought to test its potential activity in a cell culture-based inter-

plasmid transposition assay using the *Herves* plasmids as a positive control (Arensburger *et al*., 2005). *AgTango* constructs were successfully constructed; however, no transposition events were detected for *Tango* or *Herves*. Because the positive control failed to work, no assessment can be made concerning *Tango's* transposase. Possible causes and solutions for these results, alternative means to detect transposition, as well as future directions with *Tango* are discussed.

Acknowledgements

To Jake, my dissertation advisor:

Thank you for providing me with a nurturing, intellectual environment and giving me freedom and control over my project. You have many qualities that make you an excellent advisor, but most of all I appreciated your security and patience to allow me to make mistakes.

To Dr. Peter Kennelly:

I was fortunate to have not one advisor, but two. I appreciate the dedication you have shown towards my education and development as a scientist. Thank you very much for the times that you were my sounding board and gave me a sense of balance as I worked my way through my graduate education.

To my committee members – past and present:

Thank you for your dedication and time to my education. From the classes I took with you to my committee meetings, it was apparent that you cared about my education. Something I learned early on in my graduate career was that professors have many obligations. Yet, every time we met, it was obvious to me that you had read the materials that I provided and considered carefully them before we met. I realize now how difficult that truly is, and I appreciate it very much.

To Thomas E Coy, my husband:

Tom, this was a joint effort and I could have never accomplished it without you. This Ph.D. is as much yours as it is mine. I appreciate your efforts to understand my ramblings, and your tolerance of my piles of journal articles that litter our home, and the dead bugs in our refrigerator. Never in my wildest dreams did I ever imagine that I would find somebody who would be able to understand and accept my quirkiness, much less love and encourage it.

Most importantly, thank you for cultivating my sense of humor as this would have been impossible otherwise.

To Nicholas D Coy, my son:

Never before have I wanted to get my act together so badly until the day I looked into your eyes for the first time. Your arrival gave me focus and a renewed point to my life. I hope you will never tire of me telling you about the physical laws of nature, teaching you birdcalls, and hugging you.

To David L Royer, my dad and Mona Royer, my mom:

Dad, thank you for grounding me early in the concept that the world is a physical place that obeys laws which can be studied and understood. You taught me that things are knowable, understandable, often predictable, and to a certain degree, controllable. Because of this, I have avoided the pitfall of depending upon superstition to understand the world, which is the greatest gift you could have given me. Mom, thank you for hanging a butterfly net on my wall as a decoration, and didn't kill me later when I used it for its purpose to let loose butterflies in my bedroom and to show you how they roll their tongues in and out. I appreciate your "tolerance" of the Petri dishes under my bed. And, although I'll never admit it in person, thank you for dragging me all over the world to visit art museums. To both of you, thank you both for never, ever, discouraging or limiting me from doing things because I was a "girl".

Attribution

Work presented in Chapters 2 and 3 are reprinted from the journal *Insect Molecular Biology*. My chair and primary advisor, Dr. Zhijian (Jake) Tu, is a coauthor on both publications. He provided guidance of the research presented in these chapters, and edited and revised the drafts prepared for publication. Dr. Tu obtained his PhD in entomology at the University of Arizona. He is currently an associate professor in the biochemistry department at Virginia Tech.

Table of Contents

Abstract ii

Acknowledgements iv

Attribution vi

Table of Contents vii

List of Figures x

List of Tables xii

List of Abbreviations xiii

Chapter 1 Introduction 1

1.1 Transposable Elements

1.1.1 Transposable Elements Defined 1

1.1.2 Classification of Transposable elements 2

1.1.3 *IS630-Tc1-mariner* Superfamily of Class II Transposable Elements 3

1.1.4 Evolutionary Impact of Transposable Elements on Genome Function and Evolution 5

1.1.5 Horizontal Transfer of Transposable Elements 8

1.1.6 Utility of Transposable Elements as Biological Tools 9

1.1.7 Identification of Active Elements 10

1.2 Mosquitoes

1.2.1 *Anopheles gambiae* 11

1.2.2 Mosquito Taxonomy and Taxonomic Relationships 12

1.2.3 Controlling Disease through Genetic Manipulation 14

1.3 Dissertation Scheme

1.3.1 Research Aims and Scheme 16

Chapter 2 *Gambol* and *Tc1* are two distinct families of DD34E transposons:

Analysis of the *Anopheles gambiae* genome expands the diversity of the *IS630-Tc1-mariner* superfamily 17

2.1 Abstract 17

2.2 Introduction 18

2.3 Materials and Methods 21

2.4 Results 25

2.5 Discussion 30

2.6 Acknowledgements 33

Chapter 3 Genomic and evolutionary analyses of *Tango* transposons in *Aedes aegypti, Anopheles gambiae*, and other mosquito species 44

3.1 Abstract 44

3.2 Introduction 45

3.3 Materials and Methods 47

3.4 Results 53

3.5 Discussion 58

3.6 Acknowledgements 62

Chapter 4 Inter-plasmid transposition assay to test the functionality of the *AgTango* transposase 77

4.1 Abstract 77

4.2 Introduction 78

4.3 Materials and Methods 80

4.4 Results 89

4.5 Discussion 92

4.6 Acknowledgements 98

References 111

Appendix A Identities at the amino acid levels of different gene products and *Tango* between *Anopheles gambiae* and *Aedes aegypti* 129

Curriculum vitae 130

List of Figures

Figure 2.1 Select sections of the multiple sequence alignment of the *An. gambiae* DD34E TEs and *IS630-Tc1-mariner* superfamily representatives 35

Figure 2.2 The DNA binding domains of *Tc3* and *Sleeping Beauty* (SB) compared to the predicted HTH motifs of the *gambol* elements and two *mariner* TEs with which the *gambol* elements share a similar N-terminal signature (XXDEDC) 37

Figure 2.3 Phylogenetic relationships between *An. gambiae* DD34E and representative *IS630-Tc1-mariner* superfamily transposases 38

Figure 2.4 TIR consensus sequences in WebLogo format for the first 24 nucleotides of the *gambol* and *Tc1* TEs from *An. gambiae*, and *ITmDD37D* TEs from *Caenorhabditis elegans*, *C. briggsae*, and *An. gambiae* 40

Figure 2.5 Density of *gambol* and *Tc1* TEs per chromosomal arm of the *An. gambiae* genome 41

Figure 3.1 BLASTp alignment showing high sequence similarity between *AgTango1* and *AeTango1* 64

Figure 3.2 Inferred phylogenetic relationships between *Aedes aegypti Tango* elements, *AgTango1* from *Anopheles gambiae*, and representative *Tc1* from *Ae. aegypti* and other organisms 66

Figure 3.3 Conserved features and sequences of the terminal inverted repeats (TIRs) of *Tango* transposons 68

Figure 3.4 Inferred phylogenetic relationship of *Tango* elements 70

Figure 3.5 Comparison of selection pressure on *Tango1* and four host genes of *Anopheles gambiae* and *Aedes aegypti* 73

Figure 4.1 Schematic of *AgTango*, a *Tc1* transposable element from *Anopheles gambiae* 99

Figure 4.2 Maps of pBSHvSacOα (*Donor Plasmid*) and pBSHvOα (*Donor Plasmid – Revised*) 101

Figure 4.3 A reverse-contrast photograph of an ethidium bromide-stained 1% agarose-TAE gel showing the results of *Pst*I restriction digest of *pBSHvSacOα* (*Donor Plasmid*) 104

Figure 4.3 Amino acid translations of the PCR sequence of *AgTango's* ORF (AgTangoAmp) *vs.* the sequence from the *Anopheles gambiae* genome (AgTango) 105

Figure 4.5 Sequences from the 5' and 3' ends of a *Herves* recombinant obtained in the interplasmid transposition assay 106

List of Tables

Table 2.1	The DD34E transposable elements of *Anopheles gambiae*	42
Table 3.1	Percent amino acid identities between *Tango* transposons, three other *Tc1* elements from *Aedes aegypti*, and representative *Tc1* elements *Quetzal* and *Sleeping Beauty*	74
Table 3.2	The molecular characteristics and copy numbers of *Tango* transposons from *Aedes aegypti* and *Anopheles gambiae*, and three other *Tc1* elements from *Ae. aegypti*	75
Table 3.3.	Distribution of *Tango* elements among mosquito species surveyed using degenerate PCR	76
Table 4.1	Oligonucleotide primers used to sequence the ambiguous region of *pBSHvSacKOα*	107
Table 4.2	Oligonucleotide primers to create *Tango* inserts for *Tango Helper* and *Donor* plasmids for inter-plasmid transposition assay, and to sequence recombinants to identify transposition reactions	108
Table 4.3	Oligonucleotide primers used to verify *Herves* transposition events in products of interplasmid assay	109
Table 4.4	Results for inter-plasmid transposition assay for *Herves* and *AgTango*	110

List of abbreviations

Amp	ampicillin
bp	base pair
Cam	chloramphenicol
D	aspartic acid
E	glutamic acid
hsp	heat-shock protein 70 coding region
Hsp	heat-shock protein
HTH	helix-turn-helix
IR/DR	inverted repeat-direct repeat
IPTG	isopropyl-*β*-D-thiogalactopyranoside
ITm	*IS630-Tc1-mariner*
Kan	kanamycin
lacZ	*β*-galactosidase coding region from *E. coli*
LB	Luria broth
uF	micro-Faraday
MITE	miniature inverted repeat transposable element
MYA	million years ago
NLS	nuclear localization signal
Ω	Ohm
ORF	open reading frame
ORI	origin of replication
RT-PCR	reverse transcriptase polymerase chain reaction
s.s.	*sensu stricto*

TE	transposable element
TIR	terminal inverted repeat
TSD	target-site duplication
X-GAL	5-bromo-4-chloro-3-indolyl-β-D-galacto-pyranoside
V	volts

Chapter 1

INTRODUCTION

1.1 Transposable Elements

1.1.1 Transposable Elements Defined

Transposable elements (TEs) are genetic entities that move from one genomic location to another, replicate and spread within a genome, and typically do not serve a function directly to the genome in which they find themselves. In some cases, TEs are capable of crossing species barriers to invade a new genome in a process called horizontal transfer. TEs make up a significant portion of most eukaryotic genomes studied so far (Craig, 2002), and many TEs have been discovered and described. However, little is known about their biological impact, evolution, or influence on their host genomes. TEs can cause coding, regulatory and structural changes within the genome in a number of direct and indirect ways. For example, upon integration, TEs can disrupt coding or regulatory regions. Imprecise excision of a TE can result in portions of host DNA being removed and relocated to a new genomic location, and footprints left behind can result in altered gene products (Colot *et al.*, 1998). Indirectly, TEs can facilitate genomic rearrangements by providing regions of homology for ectopic recombination. Such changes can result in new gene linkages.

With the explosion of available sequenced genomes for analysis, the significance of TEs to genome evolution is beginning to be realized. In the past, TEs have been regarded as "junk" (Doolittle & Sapienza, 1980) and "parasites" (Orgel & Crick, 1980), but these viewpoints are beginning to be tempered. While it is agreed that in general TEs are genomic parasites, using the cellular machinery of their hosts for their own

propagation without providing a direct return, growing evidence suggests that these elements can inadvertently play a positive role in genome evolution by providing function and plasticity to their host genomes (reviewed in Kidwell & Lisch, 2001).

1.1.2 Classification of Transposable Elements

Transposable elements are classified as either Class I or Class II, depending on their mode of transposition (Finnegan, 1992). The members of Class I transpose through an RNA intermediate, and include the long terminal repeat retrotransposons (LTRs), non-LTRs, and short interspersed nuclear elements (SINEs). Class II elements possess terminal inverted repeats (TIRs) and transpose through a DNA intermediate, and include DNA TEs such as *Tc1* and *hobo*, miniature inverted repeat transposable elements (MITEs), and helitrons. All members of this class transpose by a "cut-and-paste" mechanism except for helitrons, which are believed to transpose by a rolling circle mechanism e.g., (Plasterk *et al.*, 1999).

Class II elements are further categorized based upon structural and mechanistic characteristics, presumably reflecting evolutionary relationships. The terms superfamily, family, subfamily, transposon, element, and copy are often used to describe relationships of and between TEs, and in this dissertation will be used as follows: Going from increasing degrees of relatedness, there are superfamilies, further broken down into families and then into subfamilies, if applicable. *IS630-Tc1-mariner* and *hAT* are examples of superfamilies. *IS630*, *Tc1*, and *mariner* are the founding families of the *IS630-Tc1-mariner* superfamily (Plasterk *et al.*, 1999; Shao & Tu, 2001). The members of a family are a group of related transposons found in diverse organisms, and usually

share conserved amino acid sequences in their transposase. For example, the TEs *Impala*, *Sleeping Beauty* and *Topi* are transposons of the *Tc1* family from fungus, fish, and mosquito, respectively (Langin *et al.*, 1995; Ivics *et al.*, 1997; Grossman *et al.*, 1999). Likewise, multiple transposons from a given family can occupy a genome. For example, *Topi*, *Tiang* and *Tsessebe* are *Tc1* transposons that reside in *An. gambiae* (Grossman *et al.*, 1999). The term 'element' is used interchangeably with 'transposon'. After an element invades a genome, it multiplies, resulting in multiple copies. Amplification of DNA TEs is tied to DNA replication and host repair. Copies are referred to as full-length or truncated. Full-length copies contain all features of the transposon, including TIRs and an intact open reading frame (ORF). It is assumed that full-length copies are potentially autonomous, that is, retain the ability to move and be moved. Truncated copies are those missing portions of the transposon, terminally, internally, or both. Those that retain intact TIRs retain the ability to be moved by their cognate transposase *in trans*. Truncated copies are also referred to as non-autonomous.

1.1.3 *IS630-Tc1-mariner* Superfamily of Class II Transposable Elements

The focus of this dissertation is the DD34E TEs from the *IS630-Tc1-mariner* superfamily. Representatives from this superfamily are found in a wide range of organisms including bacteria, plants, fungi and animals, and are probably the most widespread DNA TEs in nature (Plasterk & van Luenen, 2002) and references therein). There is substantial evidence for horizontal transfer for some members of this superfamily, particularly for the *mariner* element (summarized in Robertson *et al.*, 2002), and this process has probably contributed to their widespread distribution and

evolutionary success (Robertson & Lampe, 1995b). These elements typically range from 1.3 to 2.4 kilobase pairs (kb) in length, are flanked by TIRs typically 20-500bp in length, and contain a single gene encoding a transposase (Plasterk *et al.*, 1999). They target 'TA' sequences for integration into the host genome, resulting in 'TA' target-site duplications flanking the integrated transposon (van Luenen *et al.*, 1994).

The catalytic domain of this superfamily contains a conserved motif of D(Asp)DE(Glu) or DDD triad, which has been shown to be necessary for transposition activity through *in vivo* experimentation (Lohe *et al.*, 1997). The triad coordinates divalent metal ions which are thought to assist nucleophilic attacking groups during the cleavage and strand transfer reactions during transposition, and is found in the integrase enzyme of some RNA elements (Class I elements) and retroviruses, thus suggesting a possible common origin between these elements and members of the *IS630-Tc1-mariner* superfamily (Capy *et al.*, 1996).

The members of the *IS630-Tc1-mariner* superfamily can be further classified based on the number of amino acid residues between the second D and the third D/E residue of the catalytic triad (Shao & Tu, 2001). For example, all *Tc1*-like elements contain a DD34E motif in which 34 amino acids separate the second D from the third E residue. To date, the "triads" that have been identified in *An. gambiae* are DD34D, DD34E, DD37D and DD37E (Robertson, 1993; Grossman *et al.*, 1999; Shao & Tu, 2001; Shao and Tu, unpublished). *IS630-Tc1-mariner* members containing variable numbers of amino acid residues between the second and third amino acids (DDxD) have also been identified, namely those belonging to the *pogo* family of TEs (Shao & Tu, 2001).

The N-terminal domain of *Tc1-mariner* transposases is involved in the recognition and binding of its cognate TIRs (reviewed in Plasterk *et al.*, 1999). Binding is mediated by two helix-turn-helix (HTH) motifs, designated the PAI and RED domains (N and C-terminus within the N-terminus, respectively). They, in turn, are connected by a short stretch of conserved amino acids, the GRPR sequence, which is thought to stabilize DNA binding (Izsvak *et al.*, 2002; Yant *et al.*, 2004). These two HTH motifs occupy approximately the first 120 amino acid residues in the transposase. Through binding assays (Yant *et al.*, 2004) and crystallographic studies (Watkins *et al.*, 2004), it has been shown that the recognition and binding are primarily mediated by the PAI domain in the *Tc1* transposons. The RED domain may play a role in stabilizing the DNA/transposase complex through non-specific DNA binding (Vos *et al.*, 1993; Colloms *et al.*, 1994; Pietrokovski & Henikoff, 1997; Watkins *et al.*, 2004).

1.1.4 Evolutionary Impact of Transposable Elements on Genome Function and Evolution

The ways in which TEs are believed to have influenced genome function and evolution can be broken down into three main categories: 1) Molecular domestication, whereby the genome has co-opted the activity of TEs to perform functions within the genome; 2) Accelerated evolution, whereby TEs can bring about gross changes within the genome; 3) Evolutionary drive, whereby the host and TE engage in an evolutionary "arms race".

Throughout evolutionary history, the activity of TEs has been harnessed by and for the benefit of the genome. A familiar case is telomere maintenance in *Drosophila* in

which the non-LTR retrotransposons, TART and HeT-A, maintain telomeres rather than telomerase as in other eukaryotes (Biessmann *et al.*, 1992; Levis *et al.*, 1993). Strikingly, it appears that these two TEs, despite their separate evolutionary origins, work cooperatively, rather than competitively, to maintain telomeres (Pardue & DeBaryshe, 1999). The classic example of molecular domestication, however, is the RAG1/RAG2 enzyme system responsible for the recombination of the V(D)J locus in the vertebrate adaptive immune system. These enzymes are believed to have evolved from an ancient DNA TE, and show remarkable functional and structural similarities to the *Tc1* family (reviewed in Miller & Capy, 2004).

Some researchers argue that transposable elements provide a means for the rapid evolution of proteins by virtue of their ability to produce gross changes within the genome unattainable via simple random drift (reviewed in Kidwell & Lisch, 2001). It has been suggested that the genome can take advantage of the gross genomic changes generated and facilitated by TEs to better adapt the host to stressful environmental conditions (Wessler, 1996). It is argued by some researchers that "TE activity can be seen as an essential component of the hosts' response to stress, facilitating the adaptation of populations and species facing changing environments" (Ashburner *et al.*, 1998). This notion is supported by the increase of TE activity in response to abiotic and biotic stressors (reviewed in Wessler, 1996). Unlike molecular domestication, unambiguous examples of accelerated evolution are few; however chromosomal inversions constitute an excellent example of how TEs can facilitate rapid change. Transposable elements have been shown to be associated with the break-points of chromosomal inversions in fly species, including *hobo* in *Drosophila* (Miller & Capy, 2004) and *Odysseus* in

(Mathiopoulos *et al.*, 1998; Mathiopoulos *et al.*, 1999). Chromosomal inversions in *Anopheles* species have been associated with better adaptation of species to their particular environments through the new gene linkages created by the inversions (Powell *et al.*, 1999; Coluzzi *et al.*, 2002; Ayala & Coluzzi, 2005). It appears that TEs may have played an important, albeit inadvertent, role in better adapting these mosquitoes to their environment.

Perhaps the most controversial theory concerning TE influence in genome evolution is the hypothesis that mechanisms which evolved within the host genome to control TE activity have subsequently lead to the evolution of more complex organisms, much like technology that comes out of an arms race (McDonald, 1998). Specifically, it is proposed that TEs were instrumental in the prokaryotic/eukaryotic and the invertebrate/vertebrate macro-evolutionary transitions (McDonald, 1998). Chromatin formation and methylation, mechanisms believed to be prerequisites for the prokaryotic/eukaryotic and invertebrate/vertebrate transitions, respectively (Bird, 1995), are hypothesized to have originally evolved as mechanisms to squelch TE activity by the genome (McDonald, 1998).

With each newly sequenced genome, becomes increasingly apparent that TEs have played a larger role in genome evolution than previously thought, and as a consequence, the paradigm for TEs is changing. A deeper look at TEs and their relationships with their hosts reveals a subtle interplay that defies a single definition for these genetic elements. It appears that each TE-host case is dependent upon the individual relationship between the two and will have to be evaluated independently to determine where the TE falls in the continuum of parasite to benefactor (Kidwell & Lisch, 2001).

1.1.5 Horizontal Transfer of Transposable Elements

Horizontal transfer refers to the processes by which genetic material crosses species boundaries to invade a naïve genome. The mechanisms by which transposable elements undergo horizontal transfer remain largely unknown. However, there is evidence that it may be mediated by the movement of viruses, as in the case of the Lepidopteran DNA TE *piggyBac* (Fraser *et al.*, 1983; Cary *et al.*, 1989; Zimowska & Handler, 2006); introgressive hybridization as suggested for the distribution of *P* elements among some *Drosophila* species (reviewed in Silva & Kidwell, 2000); or through the intimate relationship between host and parasite, as in the case of the *mariner* element shared between the parasitoid wasp *Ascogaster reticulates* and its host moth *Adoxophyes honmai* (Yoshiyama et al., 2001).

Past horizontal transfer events are typically inferred from three types of circumstantial evidence (reviewed in Silva *et al.*, 2004). The most persuasive cases of horizontal transfer demonstrate all three. The first indicator is high sequence identity of the TE between divergent species, in which case the identity of the TE exceeds that of most host genes. The second is incongruence between the TE phylogeny and that of the host species in which the TE is found. Finally, patchy distribution of the TE among closely related species can also suggest horizontal transfer, although it remains a possibility that the TE was lost from the species in which the element is absent.

Several cases of horizontal transfer of elements in the *IS630-Tc1-mariner* superfamily are known, with the strongest cases for *mariner* elements (Robertson *et al.*, 2002). All three pieces of evidence have been shown for *mariner*, and horizontal transfer

of this element has been well demonstrated (reviewed in Robertson & Lampe, 1995a). It is hypothesized that horizontal transfer is an essential step in the lifecycle of *Tc1-mariner* elements. Existing evidence suggests that horizontal transfer is the only step wherein selection occurs during their lifecycle (Lampe *et al.*, 2003). These elements do not require host factors for transposition, and therefore are not restricted with regard to the genomes which they can invade. This characteristic has probably contributed to their success in invading a wide diversity of organisms (Robertson & Lampe, 1995b). Other examples of horizontal transfer within this superfamily include *Minos* (Arca & Savakis, 2000; de Almeida & Carareto, 2005) and TCp3.2 (Jehle *et al.*, 1998), both of which are *Tc1* elements.

1.1.6 Utility of Transposable Elements as Biological Tools

In addition to being important components of genomes, transposable elements have proven to be valuable biological tools. They are used as genetic markers in population studies (reviewed in Tu & Li, 2005), in the exploration of genomes by promoter and gene trapping (*e.g.*, Springer, 2000), and to introduce transgenes into genomes (reviewed by Handler & O'Brochta, 2005). Because they can rapidly spread through a population, they have been proposed as drive mechanisms to drive transgenes (Carareto *et al.*, 1997). The utility of a given transposon for transgenesis depends on the endogenous TEs residing in the genome to be manipulated, and what host factors, if any, are required for its mobility. Cross-reactivity can occur between endogenous and exogenous TEs (Sundararajan *et al.*, 1999; Jasinskiene *et al.*, 2000), potentially leading to uncontrolled, unwanted transposition events. Problems can also arise if the host genome cannot provide necessary

factors for transposition. This can result in either no transposition activity or uncontrolled, indiscriminate activity, both of which are usually undesired (O'Brochta & Atkinson, 1996). Identification of the TEs resident within a genome will enable better decision-making in selecting TEs for transgenesis.

1.1.7 Identification of Active Elements

Aside from their use as genetic markers, the utility of transposable elements as tools requires them to be active or functional. Active TEs can be identified with varying levels of confidence by a number of means. Most active TEs have been identified from the spontaneous mutations they cause in the host organism, such as *Ac* from *Zea mays* (McClintock, 1948) and *Tc1* from *Caenorhabditis elegans* (Emmons *et al.*, 1983). However, this approach depends on the serendipity of finding such an element in the organism of interest. Therefore circumstantial evidence is often used when an active TE is being sought. Examples of such evidence include high degree of polymorphism of a TE between individuals of the same species. Because such an observation can also result from ancestral polymorphism, corroborating evidence is necessary before any final interpretation can be made. This might include the TE being in one species or in high copy number, but not in another, closely related species, or the identification of footprints left behind after the excision of the TE within the host genome (evidence of past mobility). Stronger evidence of activity includes the identification of transcripts from the TE, or detection of extrachromosomal copies of the TE in the process of transposition by Southern blot or nested, inverse PCR. Finally, TEs can be tested for their ability to carry out transposition in a number of *in vitro* and *in vivo* transposition assays. It is important

to note that although a TE may be able to carry out transposition; this does not necessarily mean that it is active, just functionally competent. It is also important to bear in mind that differences have been observed in TE activity depending on the nature of the system in which the element is being tested (e.g. Pledger *et al.*, 2004), suggesting that TE behaviour is species-dependent, and activity in one system is not a guarantee of activity in another.

1.2 Mosquitoes

1.2.1 *Anopheles gambiae*

The principal organism targeted by this project is the African malaria mosquito, *Anopheles gambiae*. *An. gambiae*, which is highly anthropophilic in nature, is the primary vector of the malaria parasite, *Plasmodium falciparum*. *P. falciparum* is the most lethal of four *Plasmodium* species that causes malaria in humans. *An. gambiae* is also a vector of other important diseases including filariasis, caused by the nematode *Wuchercheria bancrofti*, and O'nyong-nyong Fever, caused by an arbovirus. The *An. gambiae* genome sequence, which was released in 2002 (Holt *et al.*, 2002), is approximately 278 million base pairs (Mbp) in size. It is estimated that 16% of the euchromatic and 60% of the heterochromatic region is comprised of transposable elements. A number of *Tc1* transposable elements from *An. gambiae* have been described (Grossman *et al.*, 1999; Hill *et al.*, 2001; Warren, *et al*., unpublished data), but with the availability of the *An. gambiae* genome it became possible to conduct a systematic, genome-wide search and analysis of these elements. While *An. gambiae* was the focus of this project, other mosquito species were included. Their relationships to *An. gambiae* are described below.

1.2.2 Mosquito Taxonomy and Taxonomic Relationships

Mosquitoes are dipteran insects, also known as "true flies", belonging to the family Culicidae. They can be found all over the world, with exception to those places that are permanently frozen (Clements, 1992). Three subfamilies of mosquitoes exist, the Culicinae, Anophelinae, and Toxorhynchitinae, the first two of which contain important mosquito species involved in disease transmission.

An. gambiae is a member of the *Anopheles gambiae* species complex, which is comprised of at least seven morphologically indistinguishable species (Coetzee *et al.*, 2000 and references therein). The complex also includes *An. arabiensis*, *An. quadriannulatus* A and B, *An. bwambae*, *An. merus* and *An. melas*, all of which belong to the subfamily Anophelinae. Hybrids have been observed in nature between *An. arabiensis* and *An. gambiae*, *An. arabiensis* and *An. quadriannulatus*, and *An. gambiae* and *An. melas* (Powell *et al.*, 1999 and references therein). However hybrids are rare, accounting for less than 0.1% of populations sampled (Coluzzi *et al.*, 1979; Petrarca *et al.*, 1991; Toure *et al.*, 1998). Most experimental crosses between these species produce fertile F_1 females but sterile F_1 males with the exceptions that crosses between *An. quadriannulatus* females against *An. gambiae* males and *An. quadriannulatus* females against *An. bwambae* males yield some fertile F_1 males as well as females (Powell *et al.*, 1999 and references therein). In addition, there is growing evidence that suggests introgression occurs between *An. gambiae* and *An. arabiensis* (Besansky *et al.*, 1994; Powell *et al.*, 1999; Besansky *et al.*, 2003; Donnelly *et al.*, 2004). These results suggest that these species retain enough similarity between their genomes for gene flow to occur, thereby providing an avenue for horizontal transfer, an important consideration when

contemplating the introduction of transgenes. The species *An. gambiae* itself can be divided into two genetically distinct variants, the molecular forms M and S (della Torre *et al.*, 2001). These two variants can interbreed to produce viable offspring, however, hybrids are rarely found in nature. Therefore, it is hypothesized that these two forms are undergoing insipient speciation (della Torre *et al.*, 2001).

In addition to members of the *An. gambiae* species complex, other Anopheline and non-Anopheline mosquitoes were used in this study in order to compare the different magnitudes of evolution within various mosquito lineages. The Anophelines included were, in order of increasing divergence from *An. gambiae*, *An. stephensi*, *An. dirus* and *An. faraurti*, and *An. albimanus*. Mosquitoes outside of the Anophelines included members of the Culicinae subfamily, *Aedes aegypti*, *Ae. albopictus*, *Ochlerotatus atropalpus*, and *Culex pipiens quiquefasciatus*. The divergence time estimated between *An. gambiae* and *An. stephensi* is 5-15 million years ago (MYA) (Gaunt & Miles, 2002) as compared to the estimated 145-200 MYA between *An. gambiae* and *Ae. aegypti* (Krzywinski *et al.*, 2006). In addition to *An. gambiae* genome, there were two other mosquito genomes available for comparative analysis, *Aedes aegypti* (Subfamily Culicinae, approximate genome size 1300Mbp), and *Culex pipiens quinquefasciatus* (Culicinae, 540Mbp). *Aedes aegypti* is an important disease vector involved in the spread of yellow fever. A second draft assembly of this genome has been released (Nene, *et al*., 2007). *Cx. pipens* is a widespread mosquito species involved in the transmission of West Nile virus. This genome project is in its early stages, and during the course of this work, was comprised of trace reads (ftp://ftp.ncbi.nih.gov/pub/TraceDB/).

1.2.3 Controlling Disease through Genetic Manipulation

Insecticides have played a major role in the control of mosquitoes and, by extension, malaria. However, because of growing insecticide resistance among mosquitoes, alternative methods of pest control are being sought, such as the development genetically modified mosquitoes that are refractory to *Plasmodium* or have reduced capabilities to transmit the parasites to human hosts. One of the proposed mechanisms for transforming and driving transgenes into mosquito populations involves the use of TEs.

The concept of the TE-mediated genetic transformation of mosquitoes is straightforward. Once a gene and appropriate control regions are identified, they are incorporated into a plasmid construct, along with the coding region of the transposase of the TE being used, flanked by the cognate TIRs recognized by the TE. This construct is microinjected into embryos, where the transposase is expressed, and transposes the entire construct inserted into the genome. Transformants are chosen using a selection marker that is also included in the construct. The TE chosen for transformation will both introduce transgenes into the genome and subsequently drive the transgene into the mosquito population. In theory this approach is simple, and great strides have been made towards this end. However, there are a number of technical and ethical issues remain to be resolved.

Low embryonic survival from the microinjections has hindered transformation in some mosquito species (reviewed in Atkinson *et al.*, 2001). The ratio of transgenic mosquitoes to the number of injected eggs is often low, and performing injections is laborious and tedious. In addition, a reduction in fitness is often observed in transgenic organisms, rendering them less competitive than their wild-type counterparts (reviewed

in Marrelli *et al.*, 2006). Transgene products can be toxic, especially at high levels, and disruption of host genes or regulatory regions can occur upon insertion of the TE (Marrelli *et al.*, 2006). However, the use of tissue and temporal specific promoters (e.g. Moreira *et al.*, 2000) and the development of chimeric proteins designed to target specific DNA sequences show promise in overcoming these obstacles (Maragathavally *et al.*, 2006).

Transgene expression can be influenced by the regulatory sequences upon insertion surrounding the insertion, sometimes resulting expression levels can be undesirable or insufficient. To circumvent this problem, insulators are being designed to buffer the internal gene (Sarkar *et al.*, 2006). However, implementation of this approach is somewhat constrained by the limited carrying capacity of some TEs (e.g. Lampe *et al.*, 1998; Fischer *et al.*, 1999; Karsi *et al.*, 2001). Problems involving the integration of TEs have also observed, such as with *Hermes*, which often includes flanking DNA from the plasmid construct (reviewed in O'Brochta *et al.*, 2003), and *piggyBac*, which sometimes remobilizes by mechanisms other than cut-and-paste, resulting in multi-copy insertions in the form of tandem arrays of the element (Adelman *et al.*, 2004). However, the most problematic obstacle yet to be overcome is poor post-integration mobility of the TE-transgene in the germ line of transformed mosquitoes (reviewed in O'Brochta *et al.*, 2003). While a handful of DNA TEs are successfully utilized in primary transformation of mosquitoes, including *Minos*, *Mos1*, *Hermes*, and *piggyBac*, all exhibit little or no post-integration mobility in the germ line (reviewed in O'Brochta *et al.*, 2003).

The intrinsic attributes of TEs make them extremely attractive as biological tools – their invasiveness, the structural and functional consequences of their genomic

insertions, and their potential to cross species boundaries (Miller & Capy, 2004). However, these features also make TEs potentially dangerous for transgenic applications. A better understanding of natural TE behaviour in mosquitoes will facilitate overcoming the aforementioned obstacles and help eliminate unwanted consequences of their use in transgenesis. This dissertation project will directly impact that effort by examining the diversity, behaviour, and evolution of DD34E TEs in the *An*. *gambiae* genome, as well their impact upon genome evolution.

1.3 Dissertation Scheme

1.3.1 Research Aims, Scheme, and Significance

This dissertation project was comprised of three distinct, but interrelated aims, designed to maximize the chances of identifying a functional transposase that could then be studied further to better understand TE behaviour in natural mosquito populations. The aims were as follows:

- **Aim 1:** Complete a genome-wide, systematic survey of the DD34E transposable elements in the published genome of *Anopheles gambiae*. Select one or a few DD34E transposable elements with characteristics of recent activity for further study.
- **Aim 2:** Determine the distribution of selected DD34E TE(s) in natural populations of the *An. gambiae* species complex and other mosquito species.
- **Aim 3:** Investigate the potential transposition activity of selected DD34E TEs in an *in vitro* inter-plasmid assay system.

CHAPTER 2

Gambol and *Tc1* are two distinct families of DD34E transposons: Analysis of the *Anopheles gambiae* genome expands the diversity of the *IS630-Tc1-mariner* superfamily

Coy and Tu, 2005

Permission to reprint granted by Blackwell Publishing Company

Insect Molecular Biology

Volume 14 Issue 5 Page 537 - Oct 2005

2.1 Abstract

Tc1 is a family of DNA transposons found in diverse organisms including vertebrates, invertebrates and fungi. *Tc1* belongs to the *IS630-Tc1-mariner* superfamily, which are characterized by common 'TA' target site and conserved D(Asp)DE(Glu) or DDD catalytic triad. All functional *Tc1-like* transposons contain a transposase with a DD34E catalytic triad. We conducted a systematic analysis of DD34E transposons in the African malaria mosquito, *Anopheles gambiae*, using a reiterative and exhaustive search program. In addition to previously described *Tc1*-like elements, we uncovered 26 new DD34E transposons including a novel family that we named *gambol*. Designation of family status to *gambol* is based on phylogenetic analyses of transposase sequences that showed *gambol* and *Tc1* transposons as distinct clades that were separated by *mariner* and other families of the *IS630-Tc1-mariner* superfamily. The distinction between *Tc1* and *gambol* is also consistent with the unique TIRs in *gambol* elements and the presence of a 'W[I/L/V]DEDC' signature near their N-termini. This signature is predicted as part

of the 'RED' domain, a component of the 'PAI' and 'RED' DNA binding domains in *Tc1* and possibly *mariner*. Although *gambol* appears to be related to a few DD34E transposons from cyanobacteria and fungi, no *gambol* has been reported in any other insects or animals thus far. Several *gambol* and *Tc1* elements have intact ORFs and different genomic copies with high sequence identity, which suggests that they may have been recently active.

2.2 Introduction

IS*630* from bacteria, *Tc1* from worm, and *mariner* from fly are the founding members of the *IS630-Tc1-mariner* (*ITm*) superfamily, a group of DNA transposons found in a wide diversity of organisms from bacteria to vertebrates (Doak *et al.*, 1994; Shao & Tu, 2001). These transposons contain a single gene encoding a transposase, which is flanked by terminal inverted repeats (TIRs) that specify their 5' and 3' boundaries (Plasterk *et al.*, 1999). They target 'TA' sequences for integration into the host genome, resulting in 'TA' target-site duplications (TSDs) flanking the integrated transposon (van Luenen *et al.*, 1994). Many elements in this superfamily range between 1.3 – 2.4 kb in length although there are examples of longer elements. The catalytic domain of the transposase is located in the C-terminus, and contains a conserved D(Asp)DE(Glu) or DDD triad, which has been shown to be necessary for transposition (Lohe *et al.*, 1997). The distances between the first two D's are variable while the distances between the last two residues in the triad are conserved within *Tc1* and *mariner* families. Not counting the degenerate copies, all *Tc1*-like transposons are characterized by a DD34E catalytic triad and all *mariner*-like transposons are characterized by a

DD34D catalytic triad. The Arabic numeral here indicates the number of amino acids between the second and third residues of the triad.

Phylogenetic analyses based on transposase sequences from diverse organisms suggest that *IS630-Tc1-mariner* transposons comprise at least six families in eukaryotes: *Tc1, mariner, ITmD37D*, *ITmD39D*, *ITmD41D*, and *ITmD37E* (Doak *et al.*, 1994; Gomulski *et al.*, 2001; Shao & Tu, 2001, Robertson, 2002). As shown for *Tc1* and *mariner*, the above phylogenetic classification is consistent with the profile of the transposase catalytic triad in all six families. Therefore *ITmD37D*, *ITmD39D*, *ITmD41D* and *ITmD37E* represent distinct families of transposons that contain DD37D, DD39D, DD41D, and DD37E catalytic triads respectively. *ITmD37D* and *ITmD39D* were previously considered to be basal subfamilies of *mariner* elements, namely the *Bmmar1* and *Soymar1* subfamiles (Robertson & Asplund, 1996; Jarvik & Lark, 1998). The congruence between phylogenetic classification based on transposase sequences and the classification according to the catalytic triad profile reflects the functional importance of the triad in *IS630-Tc1-mariner* transposons. However, there are two noted exceptions (Shao & Tu, 2001). First, *pogo*-like elements are characterized by the DDxD triad where "x" indicates variable number of residues. *Pogo* has a unique N-terminal DNA-binding domain and a long C-terminal domain rich in acidic residues. *Pogo* may either be a highly divergent group within the *IS630-Tc1-mariner* superfamily, or a related superfamily. The second exception refers to the phylogenetically unresolved elements that comprise the DDxE transposable elements (TEs). They include DDxE transposons in bacteria and archea and a few DD34E transposons from ciliates and fungi (Shao & Tu,

2001). These DDE transposons are phylogenetically distinct from the *Tc1* family that also contains a DD34E catalytic triad.

Transposons in the *Tc1* family are widely distributed in vertebrates, invertebrates, and fungi. Examples include *SleepingBeauty*, an active element reconstructed from several inactive copies from fish genomes (Ivics *et al.*, 1997); *Tc3*, a large group found in nematodes and insects (Collins *et al.*, 1989; Tu & Shao, 2002); and *Impala*, a divergent member in a soil-borne fungus (Langin *et al.*, 1995). In addition to the catalytic domain that is characterized by the DD34E triad, *Tc1* transposases contain a N-terminal region that is involved in the recognition and binding of the transposase to their TIRs (reviewed in Plasterk *et al.*, 1999). These binding activities are mediated by two helix-turn-helix (HTH) motifs, namely the PAI and RED domains (N and C-terminus, respectively) that are connected by a short stretch of amino acids (Izsvak *et al.*, 2002; Yant *et al.*, 2004). These two HTH motifs occupy approximately the first 120 amino acid residues in the transposase. Through binding assays (Yant *et al.*, 2004) and crystallographic studies (Watkins *et al.*, 2004), it has been shown that the recognition and binding is primarily mediated by the PAI domain in the *Tc1* transposons. The RED domain may play a role in stabilizing the DNA/transposase complex through non-specific DNA binding (Vos *et al.*, 1993; Colloms *et al.*, 1994; Pietrokovski & Henikoff, 1997; Watkins *et al.*, 2004). In contrast, the entire region of the N-terminal 140 residues of a *mariner* transposon *Mos1* is required for efficient and specific binding to its TIRs by the transposase, which led to the suggestion that the N-terminal region of *mariner*-like elements is of different origin than that of the *Tc1*-like elements (Auge-Gouillou *et al.*, 2001).

Here we report a systematic analysis of all DD34E transposons in the genome of the African malaria mosquito, *Anopheles gambiae*. In addition to the previously described *Tc1*-like elements (Grossman *et al.*, 1999; Hill *et al.*, 2001), we uncovered 26 new DD34E transposons including a novel group that we named *gambol*. The name is derived from *gambiae* and it also means to "leap about playfully". Although *gambol* elements contain a DD34E catalytic triad, they are distinct from *Tc1* transposons according to phylogenetic analysis of their transposase sequences and additional supporting evidence. *Gambol* appears to be more closely related to a few DD34E transposons from cyanobacteria and fungi, and it represents a new family in the *IS630-Tc1-mariner* superfamily.

2.3 Materials and Methods

2.3.1 Identification of DD34E TEs in *An. gambiae*

The *An. gambiae* genome assembly (Holt *et al.*, 2002) consisting of 8987 scaffolds was downloaded from NCBI on July 30, 2003. Forty-nine full length amino acid sequences representing major clades within the *IS630-Tc1-mariner* superfamily were used as queries to identify DD34E transposons in *An. gambiae*. Sequences from groups other than DD34E were included to insure comprehensive searching of the database, and to identify potentially distant DD34E members.

The search strategy followed that described by Biedler & Tu (2003), using a multi-query Blast search (Altschul *et al.*, 1997) and the computer programs described therein (`TEpost`, `TEcombine`, `FromTEpost`, and `TEmask`). These programs are serially linked together so that the output of one is used as input for the next. The series of

programs is run as one automated program, called `TEpipe`. The strategy is briefly as follows: The above query sequences were blasted against the *An. gambiae* database (chopped up into 100 kb fragments to speed up the search) using a tBlastN search with an E-value cutoff 1e-5. `TEpost` organized the results from this search and `TEcombine` removed redundant hits. The output from `TEcombine` was used by `FromTEpost` to generate a non-redundant list of putative *IS630-Tc1-mariner* TEs in fasta format with scaffold positions. The longest sequence from this list was then used as a query in a BlastN search (E-value: 1e-10) against the *An. gambiae* database. The results of the BlastN search were then used as input for `TEpost`, `TEcombine`, and `FromTEpost` as described above, resulting in a list of sequences in fasta format that represented putative copies that belong to the same transposon as the query copy. `TEmask` masked all copies of a TE from the database, and the search process was repeated until no more potential elements were identified. `TEpipe` provided an alignment of the transposon members using ClustalW (v. 1.81 for Linux, default settings) (Thompson *et al.*, 1994). In essence, `TEpipe` rapidly produced a list of candidate TEs of interest that are subject to further phylogenetic analysis and manual inspection to verify/clarify the classification and boundaries of each TE. Under the settings described above, each transposon is comprised of copies that are less than 20% divergent from each other. Chopping up the *An. gambiae* database did not affect the overall coverage of the survey as no match was found at the boundaries of the fragments. `TEpipe` and individual programs can be downloaded from http://jaketu.biochem.vt.edu/dl_software.htm. `TEpipe` was run on a Dell 530 Linux workstation with twin 2.0 GHz processors, 1.5 Gb RAM, and 80 Gb hard drive.

2.3.2 Characterization of DD34E TEs in *An. gambiae*

Alignments generated with ClustalW for Linux v. 1.81 (Thompson *et al.*, 1994) and/or ClustalX for Windows v. 1.83 (Thompson *et al.*, 1997) were used to identify TE boundaries and TSDs for each transposon. For some TEs, there was only one potentially full-length member, in which case alignments could not be used to determine TE boundaries. In these cases the sequences were run through EINVERTED (EMBOSS: http://ngfnblast.gbf.de/cgi-bin/emboss.pl?_action=input&_app=einverted) in order to locate putative TIRs, which for several transposons, led to the identification of their boundaries and TSDs. Translations were predicted using the Translation Machine at EBI: http://www2.ebi.ac.uk/translate/. TIR consensus sequences were generated using WebLogo (http://weblogo.berkeley.edu/logo.cgi) (Schneider & Stephens, 1990; Crooks *et al.*, 2004). Helix-turn-helix motifs were predicted with PROF predictions (Rost & Sander, 1993; Rost *et al.*, 1996) at http://www.predictprotein.org.

2.3.3 Copy number, genomic distribution, and functional classification of DD34E transposons

To determine copy number, a full-length member (intact ORF and TIRs) was chosen as a representative and used as a BLASTn query (E-value: 1e-10) against the *An. gambiae* genome. For TEs that did not have a full-length representative, the longest member was used as the query sequence. The BLAST results were run through `TEpost`, from which the copy number was determined. All copies of a TE have sequence identities above 80%, and most are above 90%, to the query sequence. TEs were then classified according to the following characteristics: 1) TEs that had TIRs and intact

ORFs coding for a transposase were labeled as potentially 'autonomous'. These are elements that hypothetically could either mobilize themselves or other members of the family with intact TIRs. 2) Those with intact TIRs but not ORFs were categorized as 'non-autonomous with TIRs'. These hypothetically retain the ability to be recognized by the appropriate transposase, and therefore could be mobilized. 3) TEs that no longer have intact TIRs and are therefore incapable of moving themselves or being moved by autonomous elements were classified as 'non-autonomous without TIRs'. If a TE had no TIRs, but did have an ORF, they were placed in this category. A number of hits were rejected and not counted towards copy number, most of which were short matches (on average less than 100bp) at the extreme 5' or 3' end of the query sequence. Hits that corresponded to matches against an inserted element within the query sequence (*i.e.*, a TE within the TE of interest) were rejected as well. Chromosomal locations of TEs were determined through the use of `TEMap` (Mao & Tu, unpublished), which converts the TE positions from the `TEpost` output to chromosomal locations and allows for determination of TE density and their correlations with gene density. The *An. gambiae* genome assembly maintained and updated by the ENSEMBL (EBI) site was downloaded and used for this analysis.

2.3.4 Phylogenetic Analysis

Neighbor-joining (NJ) and maximum-parsimony (MP) analyses, as implemented by PAUP* v. 4.0 b10 for Windows (Swofford, 2003), was used to infer phylogenetic relationships between the *An. gambiae* DD34E and representative *IS630-Tc1-mariner* transposase sequences. MP analysis was conducted using the heuristic search algorithm

with 100 random sequence additions and tree-bisection-and-reconnection branch swapping. Bootstrapping was used to determine confidence of groupings with 2000 replicates for both NJ and MP methods. Whole transposase sequences were used for phylogenetic analyses unless otherwise specified. Alignments of the transposase sequences were generated with ClustalX v. 1.83 (Thompson *et al.*, 1997). Alignment parameters used were a gap opening penalty of 5.0 and a gap extension penalty of 0.5. Minor adjustments to alignments, when necessary, were done with SeaView (Galtier *et al.*, 1996). TreeView was used to visualize the resulting trees (Page, 1996).

2.4 Results

2.4.1 Identification of twenty-six new DD34E transposons in *An. gambiae*

Our systematic survey, using a computer program called `TEpipe`, identified 26 new DD34E transposons in *An. gambiae*, in addition to all previously reported DD34E elements. Among these, *AgH1* was found prior to this analysis during a search for the DD37D and DD37E transposons in *An. gambiae* (Table 2.1; Seok, Shao & Tu, unpublished). Fourteen of the 26 new elements have members with TIRs and/or intact ORFs, nine of which have both. Table 2.1 lists the names, classification, characteristics, and copy number of these 14 new TEs, along with the previously identified *An. gambiae* DD34E elements. The remaining 12 elements are highly degenerate with no TIRs, and were not characterized further. All identified TEs have been reported to ENSEMBL Mosquito gene and TE pages: http://www.ensembl.org/Anopheles_gambiae/submission.

Conceptual translations of the DD34E TEs produce proteins ranging in size between 330-374 amino acids. There is one exception, *Lovejoy* (Table 2.1), which

encodes for 236 amino acids. While this element has an intact ORF, it is unknown whether the protein it codes for would be capable of carrying out transposition. A transposase of 236 amino acids would be the shortest DD34E transposase reported to date, although it is possible that it could result from truncation (see discussion below). Two of the 14 new elements that are characterized are *Tc1*-like, one of which is a degenerate homologue of *Tc3*, while the remaining 12 TEs form a novel group of DD34E transposons that we named *gambol*. As described later, *gambol* is a novel transposon family in the *IS630-Tc1-mariner* superfamily.

2.4.2 *Topi,* a previously identified *Tc1* element, is comprised of three lineages

Topi is a member of the *Tc1* family, and was identified by Grossman *et al.* (1999) through the use of degenerate primers and genomic library screening. They reported two copies of *Topi*, one full length and one truncated copy, which they named *Topi* I and *Topi* II, respectively. During our analysis of the multiple sequence alignment of *Topi* elements, it became apparent that *Topi* comprises three lineages which we named *Topi.1*, *2* and *3*. The division of *Topi* elements is supported by distance data, where the average identity of copies within lineages (98%) is greater than between lineages (91%). The TIRs of the lineages also differ slightly, particularly at the 5' end (Table 2.1). According to our analysis, the two copies identified by Grossman *et al.* are both members of *Topi.1*. *Topi* is among the most abundant DD34E transposons in *An. gambiae,* second only to *Tsessebe*, another *Tc1*-like transposon (Grossman *et al.*, 1999). High sequence identity between members of *Topi.1*, up to 100%, suggests that these elements may have been recently active.

2.4.3 The *gambol* elements

According to the alignment of the transposase sequences, twelve of the new DD34E transposons form the distinct *gambol* group (Fig. 2.1A). The amino acid sequence similarity between members of this group ranges from 38 to 68%. Most of these elements share a conserved 'W[I/L/V]DEDC' (Trp-[Ile/Leu/Val]-Asp-Glu-Asp-Cys) amino acid signature, approximately 80 amino acids upstream from the catalytic domain (Fig 2.1A and B). A similar signature is found in similar positions within the ORFs of two *mariner* (DD34D) TEs, M.destructor.mar1 and D.mauritiana.mar1, with signatures LLDEDC and LLDEDD, respectively (Fig 2.1A). A similar signature was also found in two previously unidentified *mariner* elements in *An. gambiae* (LLDEDD and LLDEDS, data not shown). Although BLAST searches using the *gambol* sequences produced hits against known or putative TEs, no TEs with this signature were identified in any other organism. Figure 2.2 shows that the W[I/L/V]DEDC signature is located in the predicted RED domain of the *gambol* elements, making up the C-terminal end of first helix and the remainder residing between the first and second helix of that domain. While the predicted secondary structures are highly conserved between different *gambol* elements, their overall primary sequences are not except for the W[I/L/V]DEDC signature, suggesting that this primary sequence may be important in the function of the transposase. A few of the *gambol* elements are not represented by potentially autonomous members (Table 2.1). Therefore these elements are probably no longer active, which is why they either lack the conserved W[I/L/V]DEDC signature or the DD34E triad.

2.4.4. Phylogenetic analyses place *gambol* elements in a distinct group separate from *Tc1* transposons

The inferred phylogenetic relationships between the *An. gambiae* DD34E elements and representative transposases from the *IS630-Tc1-mariner* superfamily are depicted in Figure 2.3 as an unrooted NJ phylogram. The major families of the superfamily described in Shao and Tu (2001) form their respective clades. Although *gambol* and *Tc1* are all DD34E transposons, our analysis places *gambol* and *Tc1* transposases in distinct groups that are supported by high bootstrap values (Fig 2.3). We tested the robustness of the relationships shown in Figure 2.3 using different phylogenetic treatments. MP analysis with the same alignment produced a single tree with a similar topology, differing only in the placement of the DD39D TEs (*ITmD39D*). MP places them as a sister group to the DD34D TEs (*mariner*), although this placement was not supported by bootstrap analysis. The other differences were minor changes in the placement of transposons within the *Tc1* and *gambol* groupings. Removal of apparently non-functional *gambol* elements (Table 2.1) did not change the overall tree topology in NJ and MP analyses, neither did the addition of diverse *IS630-Tc1-mariner* elements used in the analysis by Shao and Tu (2001) (data not shown). All above phylogenetic analyses indicate that there are several levels of groupings separating the *gambol* and *Tc1* clades. In other words, *gambol* and *Tc1* are separated by *mariner* (DD34D) and other transposon families in the *IS630-Tc1-mariner* superfamily. For example, *Tc1* elements are more closely related to the DD37D clade than any other clades (Shao and Tu, 2001) including *gambol*. No matter where one roots the tree, *Tc1* and *gambol* are two distinct clades. Therefore we felt justified to designate *gambol* as a transposon family

independent of *Tc1*, *mariner* and other families (e.g., DD37D, DD37E) in the *IS630-Tc1-mariner* superfamily.

Both NJ and MP methods suggest that *gambol* elements may be related to some DD34E TEs from cyanobacteria and fungi (*T31220, Nostdoc,* and *Ant1* in Fig 2.3), although the bootstrap value in the MP analysis was below 50%. Both *gambol* elements and these DD34E transposons are larger than *Tc1*, which usually range between 1.3 and 2.4 kb in size (Table 2.1). No *gambol*-like elements were found so far in any other animal species during BLAST searches using the *gambol* transposase sequences as queries.

2.4.5 Differences between *gambol* and *Tc1* TIR consensus sequences

As seen in Table 2.1, the TIR lengths within *gambol* and *Tc1* vary widely in *An. gambiae*. On average, it appears that the *gambol* TIRs tend to be longer, the longest of which is 1096 bp and extends into the ORF of the element. A comparison of the TIR WebLogo consensus sequences of the *gambol* and *Tc1* TEs from *An. gambiae* (Fig. 2.4A and B) shows that they both invariably begin with a 'CA' sequence, have a preference for a 'G' in the fifth position, and 'CA' in the 11th and 12th position. However, the overall consensus sequences are significantly different with the *gambol* elements beginning with 'CAGGGTTT', and the *Tc1* elements beginning with 'CANTGNC/T'. This evidence is in agreement with the distinction between *gambol* and *Tc1*. Interestingly, the TIR consensus of the *gambol* elements more resembles that of the DD37D transposons (Fig. 2.4C) than that of *Tc1*.

2.4.6 Chromosomal densities of the *gambol* and *Tc1* transposons in *An. gambiae*

Figure 2.5 shows the number of *gambol* and *Tc1* transposons per Mb located on each chromosomal arm, with the highest and lowest densities located on X and 2R, respectively. These results are consistent with those reported by Holt, *et al*. (2002), which suggested that TE densities within the euchromatic component were highest on the X chromosome and lowest on the 2R arm. A large proportion of *gambol* and *Tc1* are located in unmapped scaffolds, which is to be expected if we assume that most of these scaffolds are largely made up of repetitive DNA. While there is approximately double the number of *Tc1* over *gambol* TEs, the percentages of the distributions of the two groups among the chromosomes are approximately the same.

2.5 Discussion

Sequenced genomes are valuable resources for the systematic analysis of the genomic landscape for divergent TEs and the discovery of novel TEs. Whole genome investigation offers unique insight into the evolutionary dynamics of TEs and their contribution to the complex genomic picture we observe today. Our survey of the *An. gambiae* genome using a reiterative and exhaustive search program has revealed a rich diversity of DD34E transposons and uncovered a novel family named *gambol* (Table 2.1). According to sequence comparison and phylogenetic analysis, *gambol* elements are clearly distinct from *Tc1* with which they share the same DD34E catalytic triad (Fig. 2.3). Thus our analysis of the *An. gambiae* genome not only uncovered a number of new DD34E transposons, it adds a new family, *gambol*, to the six defined TE families in the widely distributed *IS630-Tc1-mariner* superfamily. No *gambol*-like elements were found

thus far in any metazoan species other than *An. gambiae*. Among all *IS630-Tc1-mariner* elements, *gambol* transposons appear to be more closely related to some DD34E transposons from cyanobacteria and fungi, which were part of the previously unresolved DDxE elements (Shao & Tu, 2001). The large size of *gambol* is also reminiscent of some of these fungal and archeal DD34E transposons. It is possible that *gambol* may be the founding member of a deep lineage of widely distributed DD34E transposons and additional *gambol*-like transposons from diverse organisms may be discovered as more genome sequences become available. It is worth noting that the *Tc1* family, the other DD34E transposon family, has divergent members in vertebrates, invertebrates, and fungi (Shao & Tu, 2001), indicating a broad distribution.

Despite the current lack of evidence for *gambol* homologues in other metazoans, the existence of many *gambol* elements with divergent transposase sequences in *An. gambiae* (up to 62% divergence at the amino acid level) may suggest a relatively long evolutionary history. Even if we do not include the degenerate elements, *gambol* clearly comprises highly divergent members. Such a genomic landscape may result from multiple invasions (horizontal transfer) by divergent *gambol* transposons during evolution. Intra-genomic diversification, which may result from the evasion of host suppression (Lampe *et al.*, 2001), could also contribute to the divergent composition of the *gambol* family. Such host suppression of TE activity may speed up TE diversification. The above two hypotheses are not mutually exclusive. The observed intra-genomic diversity is not unique to *gambol*. *An. gambiae Tc1* transposons also include a highly divergent group of elements (Fig. 2.3). We have also shown that *Topi*, a previously identified *Tc1* element (Grossman *et al.*, 1999), is comprised of three lineages.

ITmD37E (DD37E), another family of transposons in the *IS630-Tc1-mariner* superfamily, has also only been found in mosquitoes so far. The mosquito-specific distribution of these unique transposon families is intriguing, although there is no obvious reason to suggest that mosquito genomes are particularly fertile grounds for unique families of DNA transposons. In this regard, it is interesting to point out that we have previously noted an unprecedented diversity in non-LTR retrotransposons in the *An. gambiae* genome (Biedler & Tu, 2003).

The *gambol* transposases are characterized by three common features in their N-terminal DNA-binding domain: a unique amino acid signature W[I/L/V]DEDC, and two conserved HTH motifs, namely the PAI and RED domains. At present, the function of the signature sequence has not been determined. While the primary sequence of the predicted HTH motifs has undergone numerous amino acid changes, the secondary structure has been conserved between *gambol* elements. On the other hand, no changes appear to have been tolerated in the W[I/L/V]DEDC signature, indicating its functional importance. The signature is located in the RED domain, which is hypothesized to play a role in non-specific DNA binding in *Tc1* transposons (Vos *et al.*, 1993; Colloms *et al.*, 1994; Pietrokovski & Henikoff, 1997; Watkins *et al.*, 2004). The PAI domain, on the other hand, has been shown to mediate the recognition and binding of TIRs in *Tc1*. However, the *gambol* W[I/L/V]DEDC signature sequence is more similar to some *mariner* elements (Fig. 2.2), in which the entire N-terminal region is believed to be important for the efficient and specific binding to TIRs (Auge-Gouillou *et al.*, 2001). Although it is tempting to suggest that *gambol* and *mariner* share similar DNA-binding

domains, we are not able to confidently infer such a relationship because of the lack of overall sequence conservation in this region.

Several DD34E TEs may be active as indicated by high sequence identity between transposon members and the possession of intact TIRs and ORFs. These include *Topi*, *Tsessebe, Parker*, *Mango*, and *AgH1*. If active copies can be found, these elements may provide new transformation tools for the genetic manipulation of mosquito genomes. In addition, systematic analysis of endogenous transposons in mosquito genomes such as the study presented here will lead to better-informed design of transposon-based transformation systems to reduce complication or instability resulting from interactions between transformation vectors and endogenous TEs (Sundararajan *et al.*, 1999; Jasinskiene *et al.*, 2000; Atkinson *et al.*, 2001). Efficient transgenic tools will contribute to the control of mosquito-borne diseases either directly by creating pathogen-refractory mosquitoes, or indirectly by improving our understanding of mosquito-pathogen interactions. The existence of a wide array of TEs in the *An. gambiae* genome is evidence of their successful spread during certain evolutionary time. Although there are considerable challenges, it is conceivable that a well-designed transformation system will be able to help us drive refractory genes into mosquito populations, resulting in reduced vectorial capacity and the control of mosquito-borne diseases.

2.6 Acknowledgements

`TEpost`, `TEcombine`, `FromTEpost`, `TEmask`, and `TEpipe` were implemented by Feng Zhang and Rui Yang of the Department of Computer Science at Virginia Tech. `TEMap` was implemented by Chunhong Mao at Virginia Bioinformatics Institute. We would like

to thank Jim Biedler for development and assistance with `TEpipe`, and for many stimulating conversations. We would also like to thank two anonymous reviewers for constructive comments and suggestions. This work was supported by NIH grants AI42121 and AI053203, and the Virginia Experimental Station.

A

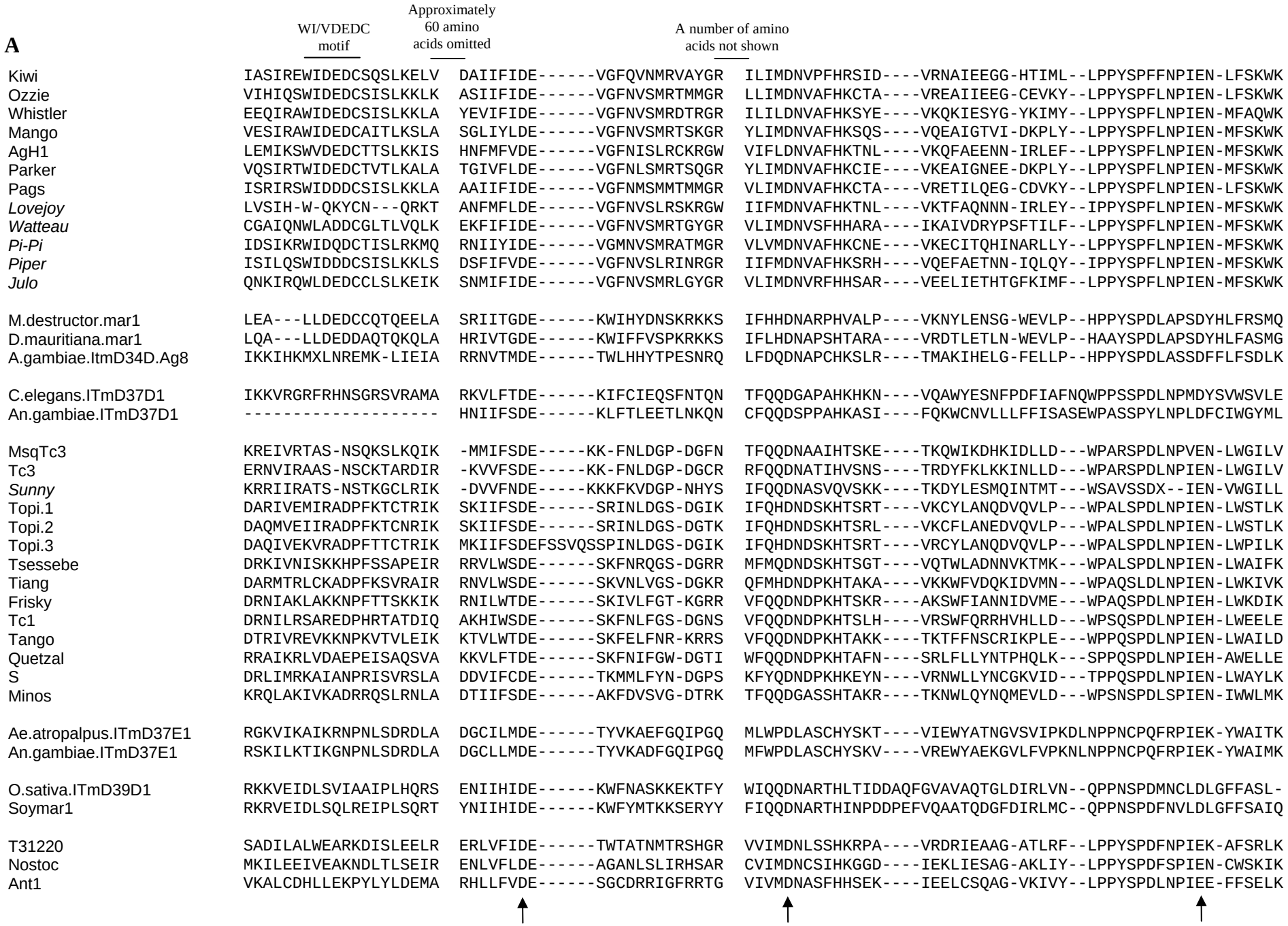


Figure continued…

B

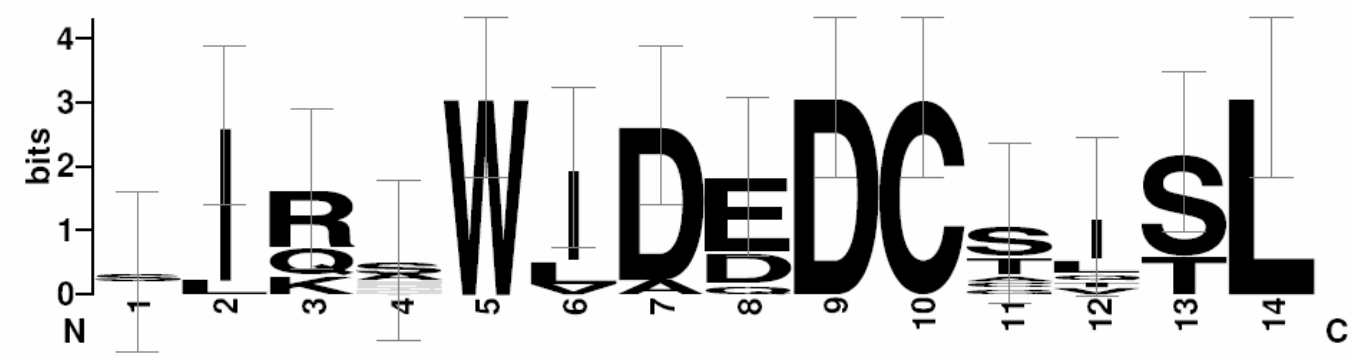


Figure 2.1A. Select sections of the multiple sequence alignment of the *An. gambiae* DD34E TEs and *IS630-Tc1-mariner* superfamily representatives. Underlined text above the alignment is the conserved N-terminal amino acid signature of the *gambol* elements, W[I/L/V]DEDC. Residues of the catalytic triad are marked with arrows. Italicized TE names indicate those elements which are probably inactive due to various mutations in their ORFs. **2.1B.** Consensus sequence of the *gambol* W[I/L/V]DEDC signature presented in WebLogo format (http://weblogo.berkeley.edu/logo.cgi), which shows the frequency and degree of sequence conservation of the amino acids at each position. The most frequent amino acid is on top of the stack, and the height of each letter is proportional to its frequency. The total height of each stack represents the overall conservation at that position. Positions without letters means that there is no consensus at that particular position.

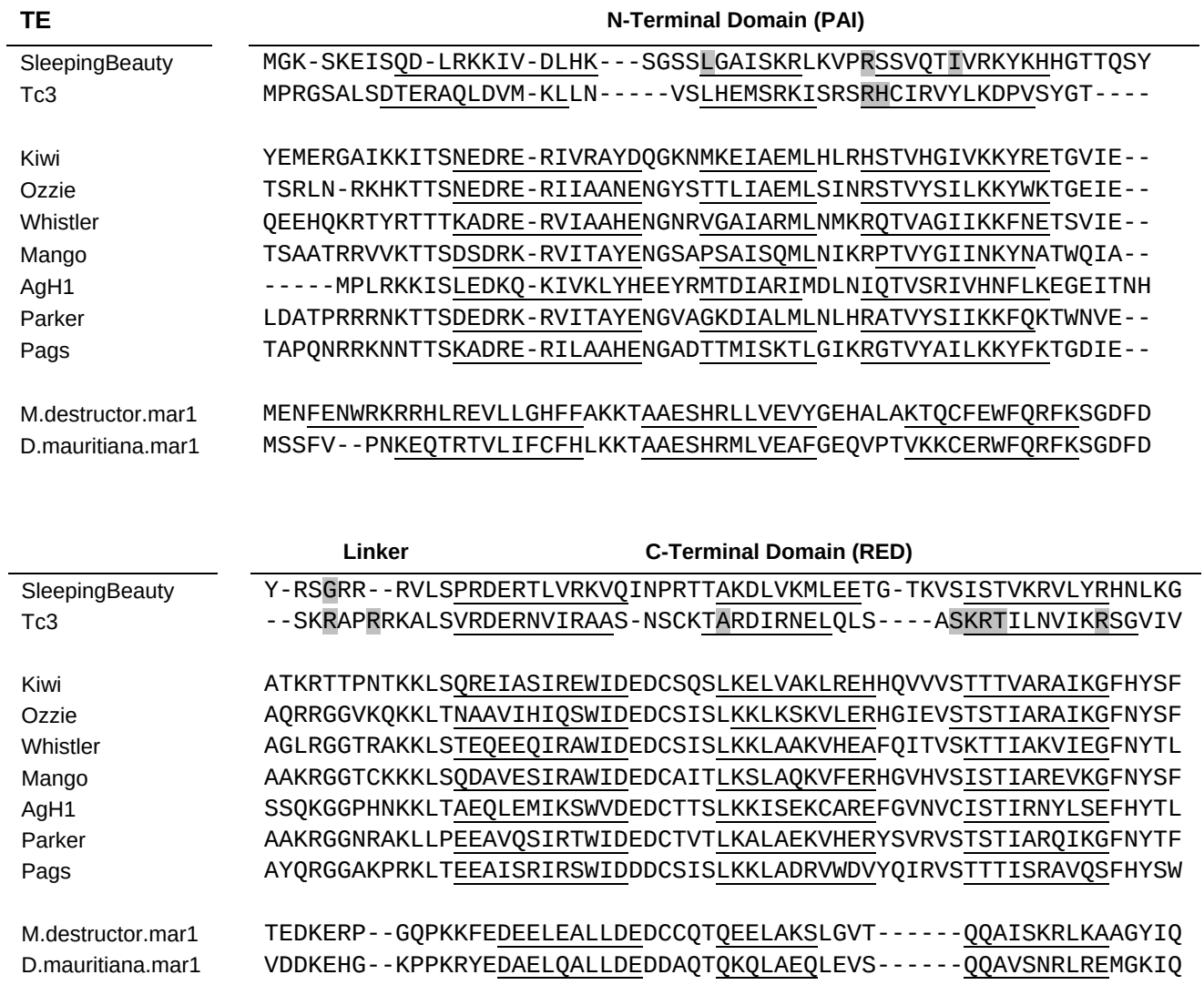


Figure 2.2. The DNA binding domains of *Tc3* and *Sleeping Beauty* (SB) compared to the predicted HTH motifs of the *gambol* elements and two *mariner* TEs with which the *gambol* elements share a similar N-terminal signature (XXDEDC). In grey shading, SB: residues that when changed to alanine destroy transposase/TIR binding (Yant *et al*., 2004); *Tc3*: residues that contact the TIR in a crystal structure (Watkins *et al*., 2004). Underlined are residues predicted to form helix structures according to PROF predictions (http://www.predictprotein.org).

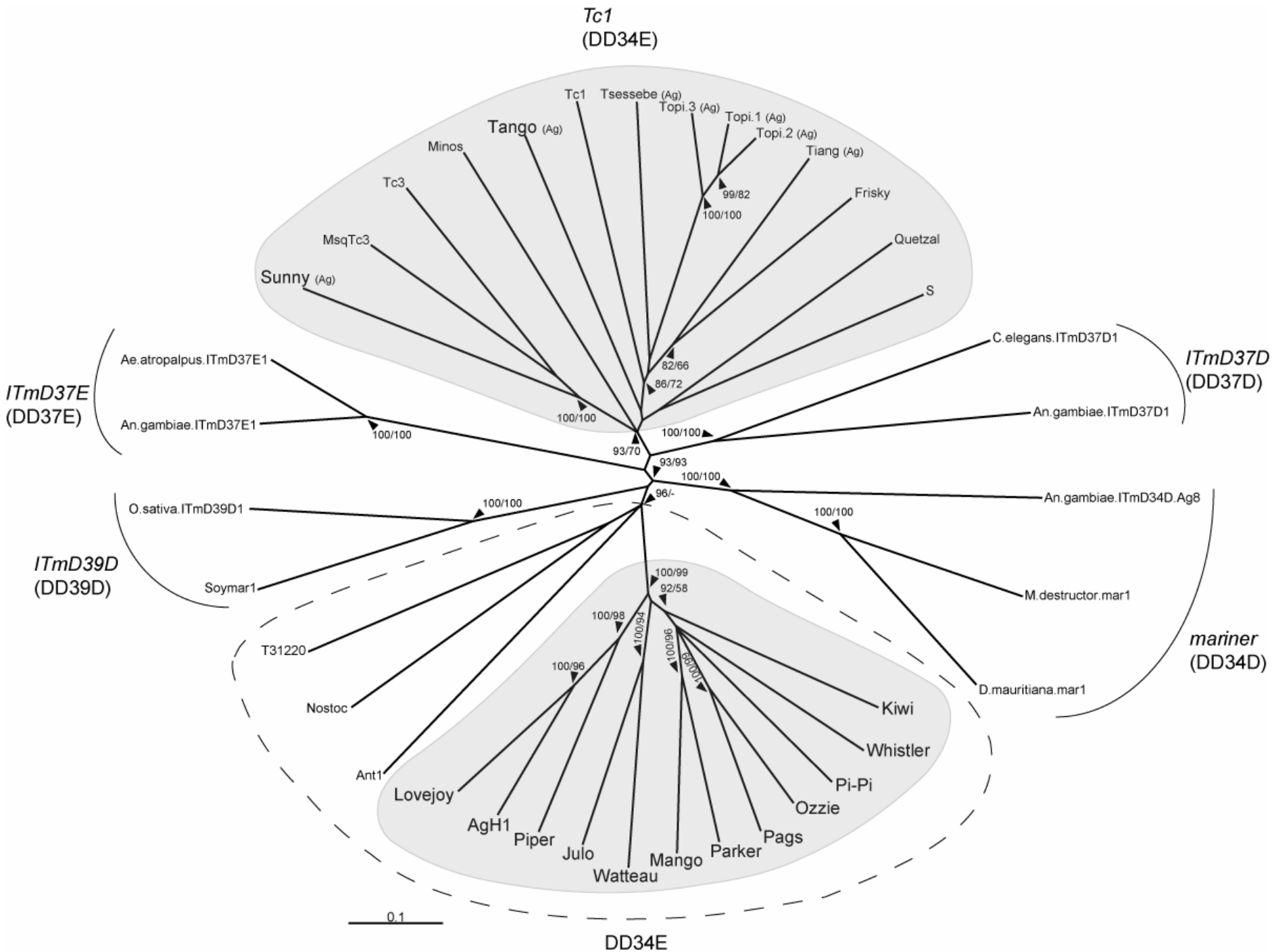
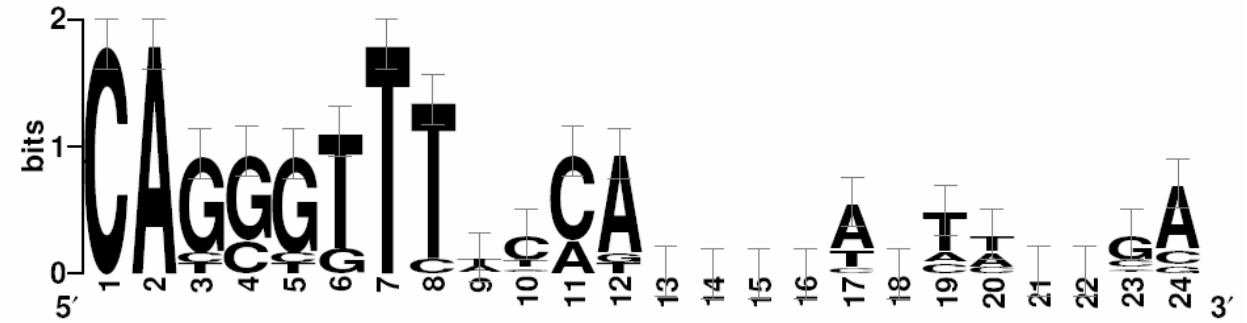
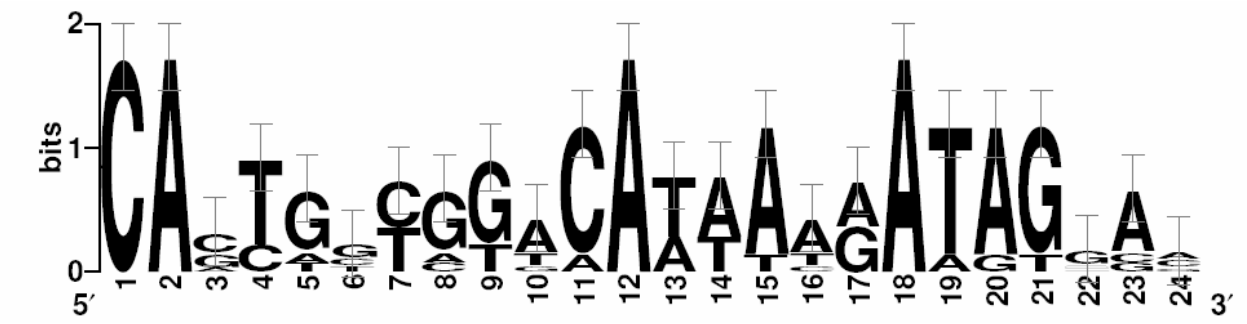

Tc1
(DD34E)
Tc1
Tsessebe (Ag)
Topi.3 (Ag)
Topi.1 (Ag)
Topi.2 (Ag)
Tiang (Ag)
Tango (Ag)
Minos
Tc3
MsqTc3
Sunny (Ag)
Frisky
Quetzal
S
99/82
100/100
82/66
86/72
100/100
ITmD37E
(DD37E)
Ae.atropalpus.ITmD37E1
An.gambiae.ITmD37E1
100/100
C.elegans.ITmD37D1
An.gambiae.ITmD37D1
ITmD37D
(DD37D)
100/100
93/70
93/93
100/100
96/-
An.gambiae.ITmD34D.Ag8
100/100
M.destructor.mar1
D.mauritiana.mar1
mariner
(DD34D)
O.sativa.ITmD39D1
100/100
ITmD39D
(DD39D)
Soymar1
T31220
Nostoc
Ant1
100/99
92/58
100/98
100/94
100/96
100/96
66/100
Kiwi
Whistler
Pi-Pi
Ozzie
Pags
Parker
Mango
Watteau
Julo
Piper
AgH1
Lovejoy
0.1
DD34E

Figure 2.3. Phylogenetic relationships between *An. gambiae* DD34E and representative *IS630-Tc1-mariner* superfamily transposases. Shown is an unrooted neighbor-joining (NJ) phylogram. Maximum parsimony (MP) produced a single tree with similar overall topology. Small numbers are the percent of the time that branches were grouped together at a particular node out of 2000 bootstrap replications for NJ and MP, respectively. Values below 50% are not shown. The alignment used to generate trees is the same as shown in Fig. 2.1. All phylogenetic analyses were carried out using PAUP* 4.0 b10 (Swofford, 2003). New TEs discovered in this study are in larger font. Ag in brackets refers to the *Tc1* elements found in *An. gambiae*.

A.



B.



C.

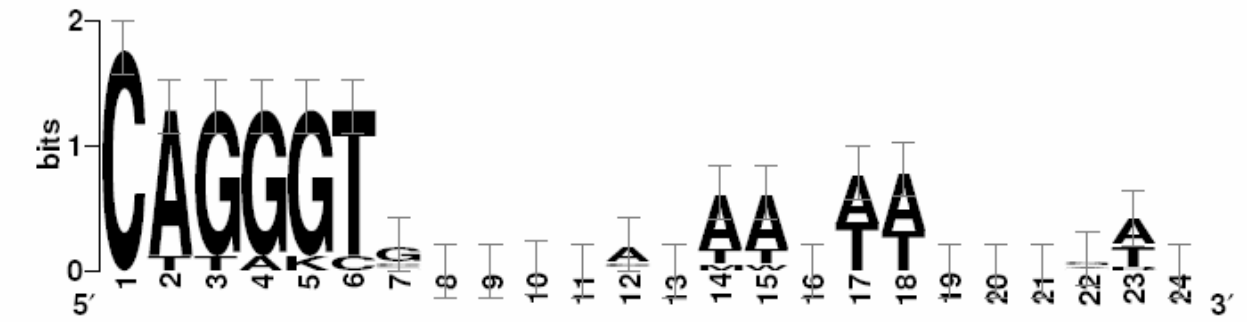


Figure 2.4. TIR consensus sequences in WebLogo format for the first 24 nucleotides of the *gambol* (**A**) and *Tc1* (**B**) TEs from *An. gambiae*, and *ITmDD37D* (**C**) TEs from *Caenorhabditis elegans*, *C. briggsae*, and *An. gambiae* (Shao & Tu, 2001; Shao & Tu, unpublished). The TIRs of *gambol* and *Tc1* are listed in Table 2.1.

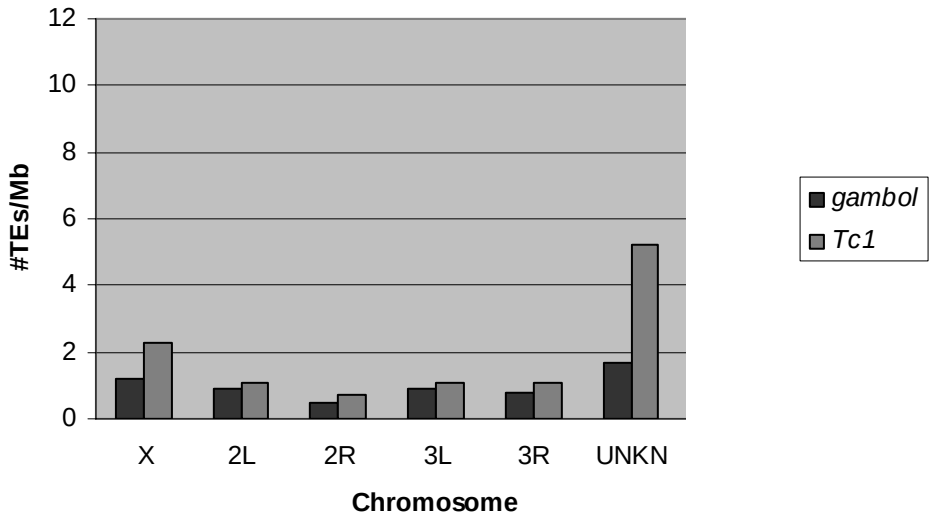


Figure 2.5. Density of *gambol* and *Tc1* TEs per chromosomal arm of the *An. gambiae* genome. For both families of TEs, the highest density is on chromosome X whereas the lowest is on chromosomal arm 2R. UNKN refers to TE copies located in unmapped scaffolds.

TABLE 2.1. The DD34E transposable elements of *Anophles gambiae*.

Transposon		Molecular Characteristics					Copy Number				Position of Representative TE		
								Non-autonomous					
TE Family	Name	TIR Length (bp)	Length (bp)	TSD	ORF (aa)	First 24bp of 5' TIR	"Autonomous"	With TIRs	Without TIRs[a]	Total	Accession #	Start	End
Tc1	Topi.1[b]	24	1413	TA	332	CACTGGTGGACATTAAAATAGGAA	12	15	73(1)	100	AAAB01008944	5222445	5223857
Tc1	Topi.2[b]	24	1556	TA	332	CACCGCTGGACATAAAAATAGGAA	10	43	61(1)	105	AAAB01008849	2027739	2029294
Tc1	Topi.3[b]	24	1446	TA	338	CACCGGTGGACAAAAAGATAGGAA	1	5	11	17	AAAB01008975	147381	148826
Tc1	Tango	224	1670	TA	338	CAGTGGCCGGCAAAATAAAGTGGC	1	26	11	38	AAAB01008960	16265453	16267122
Tc1	Tiang[bc]	260	2003	TA	333	CAGTGTCGGACAAATCGATAGGAC	6	5	33 (1)	44	AAAB01008987	7538596	7540598
Tc1	Tsessebe[b]	120	1983	TA	330	CAGTATCGGACATTAAGATAGAAC	10	143	112 (3)	265	AAAB01008968	2949062	2951044
Tc1	Frisky[d]	25	1690	TA	331	CAATTGTGTTCATTAAAATAGCAG	1	1	9	11	AAAB01008880	3259980	3261669
Tc1	*Sunny*[eh]	65	2547	TA	N/A	CACTGCCATTAATAATGATAGTCG	0	1	14	15	AAAB01008811	1156482	1159028
gambol	Kiwi	335	1880	TA	343	CACCTGTTTCCATCTACGTTCGAA	1	36	7	44	AAAB01008846	12859	14738
gambol	Ozzie[f]	680	5956	TA	374	CAGGGTTTACCAAGTCACTTTGCG	1	15	1	17	AAAB01008815	407785	413740
gambol	Whistler	229	2453	TA	343	CAGGGTTTACCTAGCCAATCGGGC	1	33	6	40	AAAB01008849	597731	600183
gambol	Mango	267	3050	TA	359	CAGGGTTTTCCAGTGGTTCTGATA	1	0	6 (1)	7	AAAB01008879	621801	624850
gambol	AgH1[g]	343	3680	TA	337	CAGGGTTTCGAATCATATAAGGGA	2	24	9	35	AAAB01008960	15668672	15672351
gambol	Parker	1096	3660	TA	343	CAGGGTTTTTCAATTGAGTTTTGA	1	22	1	24	AAAB01008960	17976068	17979727
gambol	Pags	136	3197	TA	348	CAGGGTTTTACAAGTCAGAATCGA	1	0	2	3	AAAB01005441	35464	38660
gambol	*Lovejoy*[h]	39	3652	TA	236	CATCGTTCATAAAAAATGCCAAGA	1	0	43	44	AAAB01008948	247393	251044
gambol	*Watteau*[hj]	ND	ND	ND	339	ND	0	ND	1 (1)	1	AAAB01008906	20898	21914
gambol	*Pi-Pi*[hj]	141	ND[i]	TA	354	CAGGGTTTACCAGCATAATAAACA	0	36	11 (1)	47	AAAB01006919	1	1615
gambol	*Piper*[h]	483	2989	TA	339	CAGGGTTTTCAACGGTTATATTGA	0	2	12 (1)	14	AAAB01008831	161856	164875
gambol	*Julo*[hj]	217	2533	TA	N/A	CAGCCGTTGCCGTCAAATTTTGGC	0	17	21	38	AAAB01008815	22575	25107

[a]The numbers in parentheses under the "Without TIRs" category are the number of TEs with intact ORFs within this category.

[b]*Topi*, *Tsessebe*, and *Tian*g were identified by Grossman, *et al*., 1999.

[c]*Tiang* (Grossman, *et al*., 1999) and *Crusoe* (Hill *et al*., 2001) are truncated and full copies of the same transposon.

[d]Warren, *et al*. identified *Frisky* and reported the sequence under accession #AF298053.

[e]*Sunny* is interrupted by two repetitive elements, one 5' and one 3' of the ORF, increasing the representative TE length by 765bp.

[f]*Ozzie* is interrupted by another TE of approximately 1200 bp 3’ of the ORF.

[g]*AgH1* was identified and characterized by Seok, Shao and Tu, unpublished.

[h]These TEs are probably no longer functional, see text for details.

[i]The total length of *Pi-Pi* cannot be determined because the element is located in a short scaffold which cuts off its 3' end. The TIR length of this element was determined by alignment of non-autonomous transposon members.

[j]*Watteau*, *Pi-Pi*, and *Julo* have DD35E motifs.

Chapter 3

Genomic and evolutionary analyses of *Tango* transposons in *Aedes aegypti, Anopheles gambiae,* and other mosquito species

Permission to reprint granted by Blackwell Publishing Company

Insect Molecular Biology

[Epub ahead of Print - May 2007]

3.1 Abstract

Tango is a transposon of the *Tc1* family and was originally discovered in the African malaria mosquito, *Anopheles gambiae.* Here we report a systematic analysis of the genome sequence of the yellow fever mosquito, *Aedes aegypti*, which uncovered three distinct *Tango* transposons. We name the only *An. gambiae Tango* transposon *AgTango1* and the three *Ae. aegypti Tango* elements *AeTango1-3*. Like *AgTango1*, *AeTango1* and *AeTango2* elements each have members that retain characteristics of autonomous elements such as intact open reading frames and terminal inverted repeats (TIRs). *AeTango3* is a degenerate transposon with no full-length members. All full-length *Tango* transposons contain sub-terminal direct repeats within their TIRs. *AgTango1* and *AeTango1-3* form a single clade among other *Tc1* transposons. Within this clade *AgTango1* and *AeTango1* are closely related and they share approximately 80% identity at the amino acid level, which exceeds the level of similarity of the majority of host genes in the two species. A survey of *Tango* in other mosquito species was carried out using degenerate PCR. *Tango* was isolated and sequenced in all members of the *An. gambiae* species complex, *Ae. albopictus*, and *Ochlerotatus atropalpus*. *Oc. atropalpus* contains a rich diversity of *Tango* elements while *Tango* elements in *Ae. albopictus* and the *An.*

gambiae species complex all belong to *Tango1*. No *Tango* was detected in *Culex pipiens quinquefasciatus, An. stephensi, An. dirus, An. farauti,* or *An. albimanus* using degenerate PCR. Bioinformatic searches of the *Cx. p. quiquefasciatus* (~10 x coverage) and *An. stephensi* (0.33 x coverage) databases also failed to uncover any *Tango* element. Although other evolutionary scenarios can not be ruled out, there are indications that *Tango1* underwent horizontal transfer between divergent mosquito species.

3.2 Introduction

Transposable elements (TEs) are mobile genetic units capable of replicating and spreading in a genome. In some cases, they are capable of escaping their current host genome to invade a naïve genome in a process referred to as horizontal transfer. TEs are inarguably dynamic components of their host genomes, and in many organisms, make up a large proportion of a genome. In the case of *Anopheles gambiae*, for example, it is estimated that TEs make up 16% of the euchromatic and 60% of the heterchromatic regions of the genome (Holt *et al.*, 2002). The interactions between TEs and their host genomes are complex.

Tc1 TEs are DNA transposons that belong to the *IS630-Tc1-mariner* superfamily, a wide-ranging group of transposons found in practically all forms of life, including bacteria, fungi, plants and animals (Shao & Tu, 2001). These transposons contain a single gene encoding a transposase, flanked by terminal inverted repeats (TIRs) that define their 5' and 3' boundaries (Plasterk *et al.*, 1999). They target 'TA' sequences for integration into the host genome, resulting in 'TA' target-site duplications (TSDs) flanking the integrated transposon (van Luenen *et al.*, 1994). *Tc1* elements contain a conserved

DD34E triad within their catalytic domain, where D is aspartic acid, E is glutamic acid and '34' represents the number of amino acids between the second and third residue of the triad. In comparison, *mariner* elements, another founding member of the *IS630-Tc1-mariner* superfamily, contain a DD34D triad within their catalytic domain. This triad has been shown to be necessary for transposition (Lohe *et al.*, 1997).

Several cases of horizontal transfer of elements in the *IS630-Tc1-mariner* superfamily are known, with the best-documented cases for *mariner* elements (see Robertson *et al.*, 2002). It is believed that horizontal transfer is an essential step in the lifecycle of these elements, and that this process has contributed to their widespread distribution (Robertson & Lampe, 1995). These elements do not require host factors for transposition, and therefore are not restricted in which genomes they invade. Evidence suggests that horizontal transfer is the only time upon which selection acts on these elements (Robertson & Lampe, 1995; Lampe *et al.*, 2003). There are also cases of horizontal transfer of *Tc1* transposons, for example *Minos* (Arca & Savakis, 2000; de Almeida & Carareto, 2005) and TCp3.2 (Jehle *et al.*, 1998).

In a recent survey of the DD34E TEs of *An. gambiae*, we uncovered a *Tc1* transposon which we named *Tango* (Coy & Tu, 2005). There are 38 copies of *Tango* within *An. gambiae*, one of which contains an intact open reading frame (ORF) and TIRs. *Tango* is 1670 base pairs (bp) long and contains TIRs of 224 bp. The single ORF codes for a transposase of 338 residues and contains the typical *Tc1* catalytic triad, DD34E. Here we report a systematic analysis of the genome assembly of the yellow fever mosquito, *Aedes aegypti*, which uncovered three distinct *Tango* transposons, *AeTango1-3*. We describe the structural and genomic characteristics of the *AeTango* elements. We

determined the evolutionary relationships of *Tango* elements and found that *AgTango1* and *AeTango1* are most closely related, sharing approximately 80% amino acid identity. PCR surveys and database analyses revealed patchy distribution of *Tango1* in a number of mosquito species. We propose that horizontal transfer between divergent mosquito species best explains the observed species distribution and phylogeny of *Tango1* elements.

3.3 Materials and Methods

3.3.1 Identification of *Tango* elements in the *Ae. aegypti* genome

The *Ae. aegypti* genome assembly was downloaded from The Institute for Genomic Research (TIGR; http://msc.tigr.org/aedes/release.shtml) on October 7, 2005, which is the version for current annotation. The search strategy for *Tango* elements within this genome is similar to the previously described strategy (Biedler & Tu, 2003; Coy & Tu, 2005), which is a reiterative and exhaustive search based on BLAST (Altschul *et al.*, 1997) and implemented using a series of computer programs including `TEpost`, `TEcombine`, `FromTEpost` and `TEmask`. The strategy is briefly as follows: the full-length amino acid sequence of *AgTango1* was used to search the *Ae. aegypti* database during a tBLASTn search with an E-value cutoff of 1e-5. `TEpost` parsed and organized the results from this BLAST search and `TEcombine` removed redundant hits. The output from `TEcombine` was used by `FromTEpost` to generate a non-redundant list of putative *Tango* TE nucleotide sequences in fasta format. Representative sequences from this list were translated to amino acid sequences using the Translate Tool from the ExPASy webserver (http://www.expasy.ch/tools/dna.html), which were included and used in

phylogenetic analysis to verify that they are indeed *Tango* elements. To ensure that we included all *Tango* elements, the selection of representatives was liberal at first, which is why three *Tc1* transposons that were thought to be *Tango* were re-classified as other *Tc1* elements after phylogenetic analysis. We used a setting that each transposon is comprised of copies that are less than 20% divergent from each other. Using the 20% divergence criteria, the three *Ae. aegypti Tango* transposons were able to mask all copies of *Tango* elements in the genome. The boundaries of *Tango* elements were determined by aligning multiple copies of each *Tango* and by identifying TIRs and the TA TSDs. In sum, computer programs described above rapidly produced a list of candidate TEs of interest that were subject to further phylogenetic analysis and manual inspection to verify/clarify the classification and boundaries of each TE. Computer programs described here were run on a Dell 530 Linux workstation with twin 2.0 GHz processors, 1.5 Gb RAM, and 80 Gb hard drive. They can be downloaded from http://jaketu.biochem.vt.edu/dl_software.htm.

3.3.2 DNA Extraction and PCR Amplification

Species analyzed during the PCR survey of *Tango* include members of the *An. gambiae* species complex, (*arabiensis*, *bwambae*, *gambiae*, *melas*, *merus*, *quadriannulatus*) and four Anopheline representatives outside the complex (*albimanus*, *dirus*, *farauti* and *stephensi*). *Ae. albopictus*, *Oc. atropalpus* and *Cx. p. quiquefasciatus* were also surveyed. Genomic DNA was isolated from individual mosquitoes by homogenization in 150ul DNAzol/1.5ul polyacryl carrier in 1.5ml tubes following the manufacturer's instructions (MRC, Cincinnati, OH). DNA was resolubilized in 50ul 0.1

X TE. Eight individuals were used from each species, except in the case of *An. bwambae*, in which three were used. One microliter of genomic DNA from each individual was combined into a pool, one microliter of which was subsequently used as template in degenerate PCR. In the case of *An. albimanus*, *An. dirus*, and *An. farauti*, genomic DNA obtained from Malaria Research and Reference Reagent Resource Center (MR4; http://www.mr4.org/) was used. This genomic DNA was isolated and pooled from an undisclosed number of mosquito individuals by MR4. Hemi-nested PCR was used to amplify *Tango* using degenerate primers, which were designed based on the deduced amino acid sequences of *Tango* elements from *An. gambiae* and *Ae. aegypti*, and span the conserved C terminal domain of the transposase (Fig 3.2). The primer sequences used for the first PCR were TangoDegenF1 5' AARCCNYTNGARTTYTGGAA 3' (KPLEFWK) and TangoDegenR1 5' ABYTTRTTNGTNACNCCNGTYTT 3' (KTGVTNK and the first two nucleotides of N/D/Q at the 5' end of this primer). Reactions were prepared in an AirClean AC600 Series PCR Workstation (AirClean Systems, Raleigh, NC) to reduce chances of cross contamination. Negative controls were used in all reactions which included all reagents for PCR amplification except for template genomic DNA. We also used degenerate primers for glutamate dehydrogenase as positive control to verify that the template DNA was of good quality. Touchdown PCR was used to amplify *Tango* elements from mosquito genomic DNA, using r*Taq* polymerase from TaKaRa Bio, Inc. (Otsu, Japan). Initial denaturation was carried out at 94°C/5 minutes. Subsequent denaturation and extension were at 94°C/30 seconds and 72°C/30 seconds, respectively. An initial annealing temperature of 65°C was used. The annealing temperature was lowered one degree at the end of each cycle until reaching 55°C, at which 14 cycles were

carried out. All annealing steps were for 30 seconds. One microliter of the PCR reaction was used in the second-round PCR using TangoDegenF2 5' AYRYNATGGTNTGGGGNTGYTT 3' (the last two nucleotides of N, V[I]MVWGC, and the first two nucleotides of F) and TangoDegenR1 as primers with an annealing temperature of 55°C and an extension time of 30 seconds.

3.3.3 Cloning and Sequencing of PCR amplified *Tango* sequences

PCR products were gel purified with Sephadex Band Prep (Amersham Biosciences, Piscataway, NJ) and cloned into pGEM T-Easy vector (Promega, Madison, WI), which were then used to transform JM109 bacterial cells (Promega, Madison, WI). Four colonies were chosen for *An. gambiae* molecular form M, five colonies for *An. gambiae* molecular form S, *An. albimanus*, *An. melas*, and *An. merus*, six colonies for *An. bwambae*, *Cx. p. quiquefasciatus*, and *An. quadriannulatus*, seven colonies for *An. arabiensis*, eight colonies for *Oc. atropalpus*, and 13 colonies for *Ae. albopictus* were chosen and grown up overnight using standard protocols. Plasmids were isolated using the Wizard Mini Prep kit (Promega, Madison, WI) and sequenced. Automated sequencing was performed at the Virginia Bioinformatics Institute, Virginia Tech (Blacksburg, VA). In addition to the above, the PCR product from *An. stephensi* was purified directly with GFX PCR DNA and Gel Band Purification kit (Amersham Biosciences, Piscataway, NJ), and treated as above with 23 clones chosen for sequencing and analysis. GenBank accession numbers for sequences obtained from degenerate PCR are (EF423994-EF424048).

3.3.4 Phylogenetic Analysis

Neighbor-joining, minimum evolution, maximum-parsimony, and/or maximum likelihood analyses, as implemented by PAUP* 4.0 b10 (Swofford, 2003), were used to infer phylogenetic relationships between the *Tango* transposons and representative *Tc1* transposase sequences (Fig. 3.2), and for the *Tango* nucleotide sequences obtained from degenerate PCR across mosquito species (Fig. 3.4). Bootstrapping was used to determine confidence of groupings for neighbor-joining, minimum evolution, and maximum parsimony methods. Modeltest (version 3.7, Posada & Crandall, 1998) was performed on nucleotide sequence alignment data to determine the best model for maximum likelihood analysis (Fig 3.4). The number of bootstrap replicates as well as the methods and parameters used to obtain alignment input files and to generate phylogenetic trees are described in respective figure legends. Unless otherwise noted, ClustalX version 1.83 (Thompson *et al.*, 1997) was used to generate alignment input files. Minor adjustments to alignments, when necessary, were made with SeaView (Galtier *et al.*, 1996). Resulting trees were either viewed directly using PAUP* 4.0 b10 (Swofford, 2003) or using TreeView (Page, 1996). Amino acid sequence identities (Table 3.1) were converted from distance data obtained using PAUP* 4.0 b10 (Swofford, 2003).

3.3.5 Analysis of synonymous and non-synonymous substitution rates

SNAP (Synonymous/Non-synonymous Analysis Program) http://hcv.lanl.gov/content/hcv-db/SNAP/SNAP.html (Korber, 2000) was used to determine dN, dS and dN/dS ratios for *AgTango1 vs*. *AeTango1*, and for a number of orthologous host genes shared between *An. gambiae* and *Ae. aegypti* (Severson *et al.*,

2004). Input files for these analyses were generated as follows. Alignments of the amino acid sequences were made with ClustalX version 1.83 (Thompson *et al.*, 1997). These alignments were then used to generate the alignment of the corresponding nucleotide sequences using CodonAlign version 2.0 (Hall, 2004). The resulting alignment was converted into FASTA format using ClustalX version 1.83 (Thompson *et al.*, 1997) and used directly as input for the SNAP program using default parameters. The analysis of *Tango1* was performed with representative sequences from each of the two species as depicted in Table 3.2.

3.3.6 Search for *Tango* in the *Cx. p. quiquefasciatus* and *An. stephensi* databases

The amino acid sequences of *AgTango1* and *AeTango1-3* were used as query sequences in a tBLASTN search against the *Cx. p. quiquefasciatus* trace file, version 3.0. The database was downloaded on March 15th, 2006 from NCBI (ftp://ftp.ncbi.nih.gov/pub/TraceDB/) and has approximately 10x coverage with 5 million sequences and 4.6 billion bases. Also searched was an *An. stephensi* database generated from pyrosequencing by our laboratory with 78 Mbp which represents a 0.33x coverage of the genome. In both cases, a low stringent default e-value cutoff was used which allowed for matches of e-value of 10.

3.4 Results

3.4.1 Discovery, genomic and phylogenetic analysis of *Tango* transposons in *Ae. aegypti*

We have previously described that the *An. gambiae* genome harbors a single *Tango* transposon, *AgTango1*. During our current survey of the *Ae. aegypti* genome, we discovered three distinct *Tango* transposons, namely *AeTango1, AeTango2,* and *AeTango3*. The amino acid sequence identity between *AgTango1* and the three *Ae. aegypti Tango* transposons ranges from 59.1% to 79.9% (Table 3.1). The 79.9% identity is found between *AgTango1* and *AeTango1* (Fig. 3.1A and Table 3.1). The evolutionary scenarios contributing to the high conservation between *Tango1* transposons in two highly divergent mosquitoes are discussed later. The high level of sequence identity (>59%) separates the *Tango* elements from other *Tc1* transposons including the three *Tc1* elements from *Ae. aegypti*, *Tc1_Ele4-6*, that were found during this genome-wide survey (Table 3.1). As shown in Figure 3.1B, all four *Tango* transposons align very well in an alignment that includes other transposons in the *Tc1* family. The classification of the three *AeTango* transposons was confirmed by phylogenetic analysis showing *AgTango1* and *AeTango1-3* form a compact clade that is supported by bootstrap (100%, 99% and 62%) during neighbor-joining, minimum evolution, and maximum parsimony analyses (Fig 3.2). Within the *Tango* clade, *AeTango1* and *AgTango1* are closely related while *AeTango2* and *AeTango3* are closely related. *Tc1_Ele4* is more closely related to the *Tango* elements than other *Tc1*-type elements, as indicated by the high bootstrap values for the clade comprised of *Tc1_Ele4* and all *Tango* elements (96%, 95%, and 84%, Fig 3.2). However, we decide not to classify *Tc1_Ele4* as a *Tango* element because of the low

sequence similarity between *Tc1_Ele4* and *Tango* elements (Table 3.1) and the degenerate nature of the *Tc1_Ele4* sequence (Table 3.2).

Like *AgTango1*, *AeTango1* and *AeTango2* elements each have members which retain characteristics of autonomous elements (intact ORF and TIRs), and therefore could be active (Table 3.2). *AeTango3* is a degenerate transposon with no full-length members and thus its TIRs, TSDs, and length could not be determined. *AeTango1* can be further subdivided into two groups, *AeTango1a* and *AeTango1b*. *AeTango1b* contains elements that may have undergone an internal recombination event that has lead to the generation of long TIRs of approximately 780 bp, and a duplication of the ORF such that half of the ORF is on both the positive and negative strands of the element. Of particular note is that while *AeTango1a* is a low-copy number TE, *AeTango1b* has achieved a comparatively high copy number of approximately 440 copies. Characteristics of the three other *Tc1* transposons discovered during this study are also shown in Table 3.2.

3.4.2 *Tango* from *Ae. aegypti* and *An. gambiae* share structural and molecular characteristics

AgTango1, *AeTango1* and *AeTango2* share a number of conserved molecular characteristics. The overall lengths of the elements, TIRs, and translated ORFs are all comparable between the transposons (Table 3.2). Additionally, all three elements contain imperfect subterminal direct repeats (DRs) within their TIRs (Fig 3.3). Subterminal DRs are also found in the reconstructed and functional DNA transposon *Sleeping Beauty*, and have been shown to be the sites to which the *Sleeping Beauty* transposase binds (Cui *et al.*, 2002). Also conserved are the nucleotide sequences flanking the subterminal

terminal repeats on their 5' ends (Fig 3.3). There is a conserved 'GG' dinucleotide sequence and a conserved 'TGA' trinucleotide sequence that is maintained in all three transposons at the 5' end of the outer and inner DR, respectively. The reason for this conservation is not known, however Cui *et al.* (2002) demonstrated that changes to the nucleotides surrounding the subterminal DRs alter transposition efficiency. There are additional DRs within the TIRs of *AgTango1* (Fig 3.3). These DRs are perfect, but their significance is unknown.

3.4.3 Distribution of *Tango* among mosquito species

We employed a strategy of degenerate PCR followed by sequencing of clones of PCR products to survey *Tango* transposons in a number of mosquito species (results summarized in Table 3.3). The PCR primers were designed according to amino acid sequences that are conserved between *AgTango1*, *AeTango1*, and *AeTango2* (Fig. 3.1B). There is no stretch of amino acid sequence that is conserved between the above three *Tango* elements and the degenerate *AeTango3*. Given the phylogenetic relationship of the *Tango* elements shown in Fig. 3.2, our primer design maximizes the chances for an inclusive amplification of *Tango* elements. *Tango* was found in all members of the *An. gambiae* species complex by degenerate PCR, but not in the other Anophelines tested, namely *An. stephensi, An. dirus, An. farauti,* and *An. albimanus* (Table 3.3). PCR products were not obtained for *An. dirus* and *An. farauti* using *Tango* primers. The integrity of the genomic DNA from these two species was verified using degenerate primers for the glutamate dehydrogenase gene (data not shown). Twenty-three clones were sequenced from *An. stephensi*, a relative of *An. gambiae* that is in the subgenus

Cellia, none of which were found to be *Tango* elements, nor were they from other *Tc1* transposons. These results are consistent with a BLAST search of a pyrosequencing database of *An. stephensi* (0.33x coverage) in which no *Tango* elements were found. Within the Culicine, *Tango* elements were identified in *Ae. albopictus* and *Ochlerotatus atropalpus*, but not in *Culex pipiens quiquefasciatus*. A BLAST search of the trace file of the *Cx. p. quiquefasciatus* genome (~10x coverage) did not reveal any *Tango* elements, consistent with the PCR data. Description of the databases is in the Experimental Procedures section.

3.4.4 Phylogenetic analysis of *Tango* sequences in Anopheline and Culicine species

Maximum likelihood, neighbor-joining, minimum evolution, and maximum parsimony analyses were performed using the nucleotide sequences obtained from the clones from degenerate PCR, plus *AeTango1-3* from *Ae. aegypti* and *AgTango1* from *An. gambiae*, rooted with representative *Tc1* elements. Figure 3.4 shows the maximum likelihood tree obtained using a model selected by Modeltest (version 3.7, Posada & Crandall, 1998). Analyses using neighbor-joining, minimum evolution, and maximum parsimony algorithms produced essentially the same tree as maximum likelihood analysis, differing only in the placement of clone sequences within the major groupings. Bootstrap was performed using neighbor-joining, minimum evolution, and maximum parsimony algorithms and the bootstrap values are overlaid on the maximum likelihood tree (Fig 3.4). As shown in Fig. 3.4, two main *Tango* groups are apparent, one containing *Tango2*/*Tango3*-type elements from *Ae. aegypti* and *Oc. atropalpus*, the other containing *Tango1*-type elements in *Ae. aegypti*, *Ae. albopictus*, *Oc. atropalpus*, and all species in

the *An. gambiae* complex. There are a few interesting points worth mentioning. First, *Oc. atropalpus* harbors highly divergent *Tango* elements of all three types. The fact that we were able to obtain such a diverse range of *Tango* in *Oc. atropalpus* from a sample of only eight PCR clones suggests a possibility of even greater diversity of *Tango* in this species, and testifies to the coverage of our degenerate PCR strategy. All 13 *Tango* sequences from *Ae. albopictus* are *Tango1* elements and they form two distinct clades. All *Tango* sequences from *An. gambiae* complex form a monophyletic group, indicating that only *Tango1*-type elements are found in species of this complex. The majority of the *Tango* sequences isolated from *An. gambiae* are highly similar to *AgTango1*, the putative autonomous element in *An. gambiae*. Anopheline *Tango1* elements are closer to Aedine *Tango1* than to *Oc. atropalpus Tango1*, which is incongruent to the host phylogeny. Although the bootstrap values for the *Tango1* clade (69%/56%/56%) and the clade comprised of *Aedes* and *Anopheles Tango1* (93%/79%/67%) are not high, these clades are supported by all four phylogenetic algorithms including maximum likelihood.

3.4.5 Evolutionary rates of *Tango1* and host genes in *Ae. aegypti* and *An. gambiae*

The rate of nonsynonymous (dN) and synonymous (dS) changes, and the dN/dS ratios were determined for the *AgTango1* and *AeTango1* pair. The same analyses were performed on orthologous gene pairs described in Severson *et al.*, 2004. dS values for only four orthologous gene pairs could be determined as the others were found to be saturated (Fig 3.5). *Tango1* had the highest proportion of synonymous changes out of the genes analyzed. The dN/dS ratio for the *Tango1* pair is 0.07, suggesting that *Tango1* is under purifying selection. The selective constraint on *Tango1* appears to be similar to that

of the host gene *opsin* (dN/dS=0.06) and much higher than *vitellogenin gene 1* (dN/dS=0.22) and *transferrin* (dN/dS=0.29).

3.5 Discussion

The study presented here focuses mainly on a systematic analysis of the *Tango* transposons in the yellow fever mosquito, *Ae. aegypti* and a comparative analysis of the distribution and sequences of *Tango* elements in a number of mosquito species. In addition to providing a systematic view of the diversity, characteristics, genomic distribution, and abundance of *Tango* transposons in *Ae. aegypti*, our analysis points to possible horizontal transfer of *Tango* between highly divergent mosquito species and the possibility of using *Tango* transposons as genetic tools to genetically manipulate mosquitoes.

The first clue suggesting possible horizontal transfer of *Tango* between mosquitoes is that *AgTango1* and *AeTango1* are highly similar in sequence, 79.9% identical at the amino acid level and 70% identical at the nucleotide level. This appears to be relatively high considering the estimated divergence time between *Ae. aegypti* and *An. gambiae* (145-200 Mya; Krzywinski *et al.*, 2006). To get an idea of how this identity compared to other peptides shared between these two species, we randomly surveyed 26 known *An. gambiae* peptides using a tBLASTn search against the *Ae. aegypti* genome. The range of amino acid identities was from 28-96% with an average of 43% and a median value of approximately 61% (Appendix A). The high level of sequence identity between *AeTango1* and *AgTango1* may be explained by either a high selection pressure on *Tango1* sequences due to important functions within these genomes or a case of

horizontal transfer. The analyses of the dN, dS and dN/dS ratios of *Tango1* suggest a relatively high purifying selection pressure, which could result either from purifying selection acting during horizontal transfer of transposons (Lampe *et al.*, 2003) or from *Tango1* being "domesticated" to serve important functions in these mosquitoes. The dN/dS ratio of *Tango1* falls within the range found for orthologous genes from *An. gambiae* and *Ae. aegypti*. Without values from a large number of orthologous gene pairs, we are not able to determine whether the evolutionary constraints on *Tango1* are significantly different from those of the host genes. Therefore, sequence identity data and evolutionary rate analysis by themselves could not distinguish between vertical transmission and horizontal transfer of *Tango1*.

Survey of the distribution and phylogenetic analyses of *Tango* from a wide range of mosquito species provide support for horizontal transfer of *Tango1*. *Tango1* is present in *Ae. aegypti, Ae. albopictus, Oc. atropalpus*, and all species of the *An. gambiae* species complex. No *Tango* was found in the other four Anophelines tested or in *Cx. p. quiquefasciatus* using degenerate PCR. However, we can not rule out the presence of degenerate *Tango* fragments in these species even with the appropriate negative and positive controls as described in experimental procedures. On the other hand, bioinformatic searches of the *Cx. p. quiquefasciatus* (~10 x coverage) and *An. stephensi* (0.33 x coverage) databases also failed to uncover any *Tango* element, consistent with our inability to detect *Tango* in these two species by PCR. Although the inherent limitation of PCR and genome sequencing projects prevents us from drawing a definitive conclusion about the absence of *Tango* in a number of the species surveyed, our current data are consistent with patchy distribution of at least *Tango1*, the element for which we

have the most data. Horizontal transfer of *Tango1* is a more parsimonious explanation than the alternative hypothesis of vertical transmission with multiple losses of *Tango1* (once in *Cx. p. quiquefasciatus* and more than once in the *Anopheline* lineages). The multiple losses theory does not explain the high sequence identity between *AeTango1* and *AgTango1* either. Moreover, although there are only three main clusters within the *Tango1* clade (*An. gambiae* complex cluster, *Aedes* cluster, and *Oc. atropalpus Tango1*), there is incongruence between the host phylogeny and the TE phylogeny between these three clusters (Fig. 3.4). Namely *Anopheline Tango1* is more closely related to *Aedes Tango1* than *Oc. atropalpus Tango1*. Such incongruence is also consistent with horizontal transfer of *Tango1*. It should be noted that we are not suggesting a recent horizontal transfer of *Tango1* between *Ae. aegypti* and *An. gambiae*. We are proposing that there has been a horizontal transfer event of *Tango1* between the ancestors of the two species.

Diverse *Tango* elements (*Tango1-3*) are found in two Culicine species, *Oc. atropalpus* and *Ae. aegypti*, which is in contrast to the tight cluster of only *Tango1* in the *An. gambiae* species complex. A logical hypothesis would be that a common ancestor of the *An. gambiae* complex is the recipient of the horizontal transfer event. If this is the case, we would expect that among *Anopheline* species *Tango* will be found only in members of the *An. gambiae* species complex, and/or its close relatives, depending on when the horizontal transfer had occurred. It is also possible that the diversity of *Tango* elements we observe in the two Culicine species was due to multiple invasions by *Tango* in these genomes. Thus we cannot confidently infer directionality of the horizontal

transfer of *Tango1*. Expanding the survey of *Tango* in an even wider range of mosquito species will shed light on this interesting question.

AeTango1 has achieved a significantly higher copy number in the *Ae. aegypti* genome as compared to *AgTango1* has in *An. gambiae* (3.2). The *Ae. aegypti* genome (1300 Mb) is approximately 4.5x the size of the *An. gambiae* genome (278 Mb). In addition, the organization of single-copy genes relative to repetitive DNA differs between these two species (reviewed in Rai & Black IV, 1999). *Ae. aegypti*'s genome is of the short period interspersion where single copy sequences 1000-2000 bp in length alternate regularly with short (200-600bp) and moderately long (1000-4000) repetitive sequences. The organization of *An. gambiae* is of the long period interspersion where long repeat (>5600bp) alternates with very long (13,000bp) uninterrupted stretches of unique sequences. During our initial survey of the *Ae. aegypti* genome, it appeared that there was a higher diversity and abundance of *Tc1* transposons in *Ae. aegypti* than *Ag. gambiae* (data not shown). It can be speculated that the *Ae. aegypti* genome is more permissive of transposon occupation than *An. gambiae*. Perhaps the TE load in *Ae. aegypti* has contributed to the size and organization of *Ae. aegypti*.

Intact ORFs and the high sequence identity of *AgTango1* and *AeTango1*, along with the low dN value, may suggest a potentially active or recently active element in either *Ae. aegypti* or *An. gambiae*. If *Tango* is active, this would represent the first active TE found in mosquitoes to have potentially undergone horizontal transfer between divergent mosquito species. This scenario would offer a tantalizing opportunity to study the biology of these elements in mosquito species. If not currently active, sequence comparison between the elements may allow for the reconstruction of a functional *Tango*

element, as in the case of *Sleeping Beauty*. Indeed, the ease of gene synthesis and the existence of well established transposition assay systems make reconstruction a feasible option. Active *Tango* may be used as tools to genetically manipulate mosquitoes for basic research and for control of vector-borne diseases.

3.6 Acknowledgements

An. bwambae samples were a kind and generous gift from Dr. R.K. Butlin and Dr. R.E. Harbach. We also thank Jim Biedler for stimulating conversations concerning this work and for reviewing this manuscript. We thank the anonymous reviewers for their comments. This work was supported by an NIH grant AI 042121 to Z. Tu.

A

```
AeTango1:  MENAVISKIIPLCIRKLVVHDVRNGESHRAVASKYNISKAAVGKILLKQKTFGSVVDRPG 60
           MEN   +K IPL +RKLVV DV+NGESHRAVA KY+ISK+AVGKI  K  T GSVVDRPG
AgTango1:  MENN--TKQIPLAVRKLVVRDVQNGESHRAVAGKYSISKSAVGKIFKKYSTLGSVVDRPG 58

AeTango1   RGRKRKTDARTDAKIMREVKKNPKVTVREIQKTVQLSVSSRTVRRRLVEQGLNSKVARKR 120
           RGRKR TD++TD +I+REVKKNPKVTV EI+ T+QL++S RT+RRR++EQG NSK+A+KR
AgTango1:  RGRKRITDSKTDTRIVREVKKNPKVTVLEIKNTLQLNISDRTIRRRIIEQGYNSKLAKKR 118

AeTango1   PFISKANKAKRLKFAKEHADKPLEFWKTVLWTDESKFELFNQKRRARVWCRSGEELRERH 180
           PFISKANK+KRLKFA+EHADKPLEFWKTVLWTDESKFELFN+KRR+ VWC+ GEEL+ER+
AgTango1:  PFISKANKSKRLKFAREHADKPLEFWKTVLWTDESKFELFNRKRRSHVWCKPGEELQERN 178

AeTango1   IQGTVKHGGGNVMVWGCFSWGGVGSLVKIDGIMTADSYINILRENLEVSLIQTGLEDKFI 240
           IQGTVKHGGGNVMVWGCFSWGG GSLV+I+GIMTAD+YI IL+ENLEVSLI+TGLE+KF+
AgTango1:  IQGTVKHGGGNVMVWGCFSWGGAGSLVRINGIMTADTYITILQENLEVSLIKTGLENKFV 238

AeTango1   FQQDNDPKHTAKKTKSFFRSCRIKPLEWPPQSPDLNPIENLWAILDARVDKTGVTNKNNY 300
           FQQDNDPKHTAKKTK+FF SCRIKPLEWPPQSPDLNPIENLWAILD RV KTGVTNK+ Y
AgTango1:  FQQDNDPKHTAKKTKTFFNSCRIKPLEWPPQSPDLNPIENLWAILDDRVKKTGVTNKDKY 298

AeTango1   FEALERAWEELNPQHLQNLVESMPKRLQQVLKAKGGHINY 340
           FEALE AWE L+P H++NLVESMPKRLQ VL++KGGHI Y
AgTango1:  FEALENAWENLDPNHIENLVESMPKRLQLVLRSKGGHIKY 338
```

B

Figure 3.1. A. BLASTp alignment showing high sequence similarity between *AgTango1* and *AeTango1*. Expect = e-166; Identities = 270/340 (79%), Positives = 309/340 (90%), Gaps = 2/340 (0%). Alignment was made using NCBI blast server with default settings using

the bl2seq program (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi; Tatusova & Madden, 1999). B. Amino acid alignment of the catalytic domain of the *Tango* elements from *Anopheles gambiae* and *Aedes aegypti*, three additional *Tc1* elements from *Ae. aegypti*, and representative *Tc1* elements from other organisms. ClustalX version 1.83 (Thompson *et al.*, 1997) was used to generate the alignment using default settings. Degenerate PCR primers to investigate the distribution of *Tango* in different mosquito species are shaded in grey boxes, which were designed to maximize amplification of *Tango*. Conceptual translations of *Tango* and *Ae. aegypti Tc1* nucleotide sequences were made using ExPASy's translation tool http://www.expasy.ch/tools/dna.html). *Tc1_Ele4* contains an intron, which was removed for this and all subsequent analyses. SB = *Sleeping Beauty*. References for *Tc1*, *Tc3*, *Bari-1*, *Quetzal*, *Sleeping Beauty*, and *Topi* are as follows: Emmons *et al.*, 1983 and Liao *et al.*, 1983; Collins *et al.*, 1989; Caizzi *et al.*, 1993; Ke *et al.*, 1996; Ivics *et al.*, 1997; Grossman *et al.*, 1999, respectively.

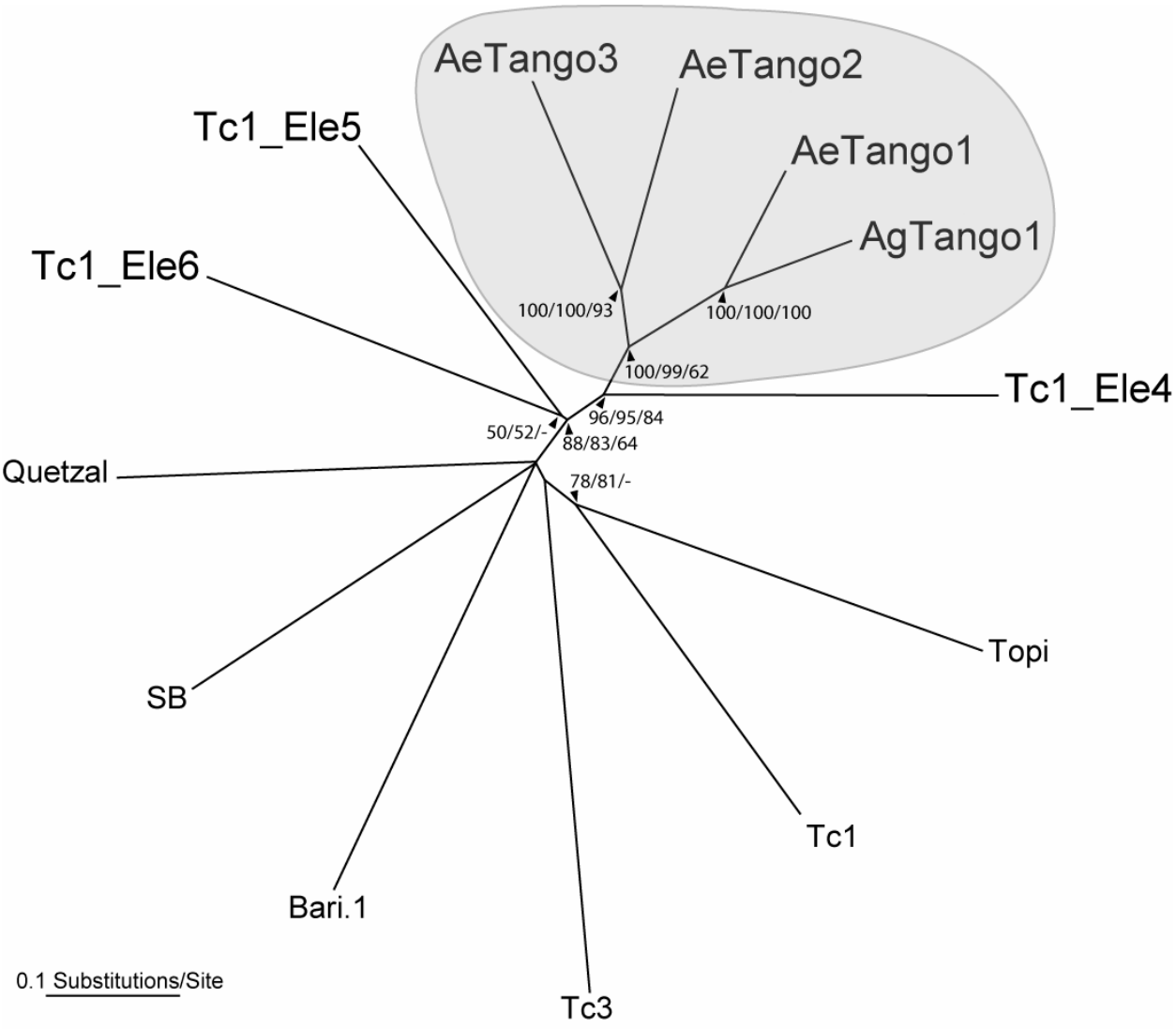


Figure 3.2. Inferred phylogenetic relationships between *Aedes aegypti Tango* elements, *AgTango1* from *Anopheles gambiae*, and representative *Tc1* from *Ae. aegypti* and other organisms. The amino acid alignment used to generate trees was made with ClustalX version 1.83 (Thompson et al. 1997) using default settings and is the same as shown in Fig. 3.1B, except that whole transposase sequences were used. All phylogenetic analyses were carried out using PAUP* 4.0 b10 (Swofford, 2003). Shown is an unrooted neighbor-joining phylogram constructed using mean character difference as distance measure. Bootstrap information from

neighbor-joining as well as minimum evolution and maximum parsimony are overlaid on the neighbor-joining tree. Minimum evolution analysis was performed using the same distance measure and produced a tree identical in topology to the neighbor-joining tree. Unweighted maximum parsimony was performed and produced three most parsimonious trees with similar overall topology to the neighbor-joining tree, differing only in the inter-relationships between *Tc3*, *Bari-1* and *Quetzal*. Small numbers are the percent of the time that branches were grouped together at a particular node out of 2000 bootstrap replications for neighbor-joining, minimum evolution, maximum parsimony, respectively. Values below 50% are not shown. Bootstrap for maximum parsimony analysis was conducted using the heuristic search algorithm with 100 random sequence additions per replicate and tree-bisection-and-reconnection branch swapping. New transposons discovered in this study are in larger font. References for representative *Tc1* transposons are the same as those in Fig. 3.1.

A

CAGTGGCCGGCAAAATAAAGTGGCCACTACTTACATTTTTACATTTTTCTAATTTTTTTCGTTAAAAGTGAAGCTAT

CTTTGTGTTATATTTTTTTAAACAATCTTCATTCTTTGCTCTTAAATGTGCAATCAAAATTTATTGAAAAAATCAAA

GGTTTACCAGTACCAAAAATATTTTTAAAAAAATTAGTCCAAATCACACTGACAAAAAAAAAGTGCCCAC

B

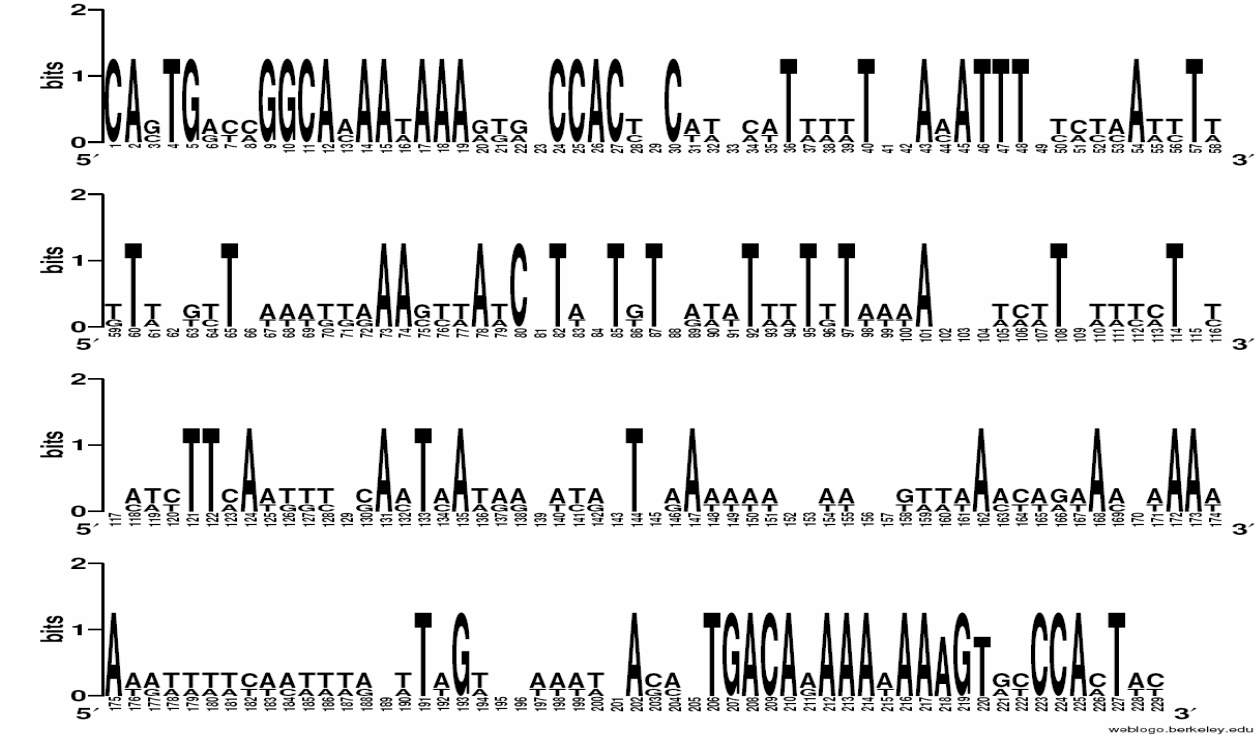


Figure 3.3. Conserved features and sequences of the terminal inverted repeats (TIRs) of *Tango* transposons. A. The 5' TIR of *AgTango1* from *Anopheles gambiae*. The TIRs of *AgTango1*, as well as the full-length *Tango* elements from *Aedes aegypti*, contain subterminal imperfect direct repeats (DRs), shown in grey shading. Underlined are conserved nucleotide sequences flanking the DRs shared by all three *Tango* transposons. *AgTango1* also has two sets of DRs as demarcated with the arrows, not present in the other *Tango* transposons. B. Weblogo consensus (http://weblogo.berkeley.edu/logo.cgi) of the TIRs of *AgTango1*, *AeTango1* and *AeTango2*

showing the regions of conservation of the nucleotide sequences of these transposons. The regions with the greatest conservation are those of the subterminal DRs. In both A and B, only the 5' TIR is shown.

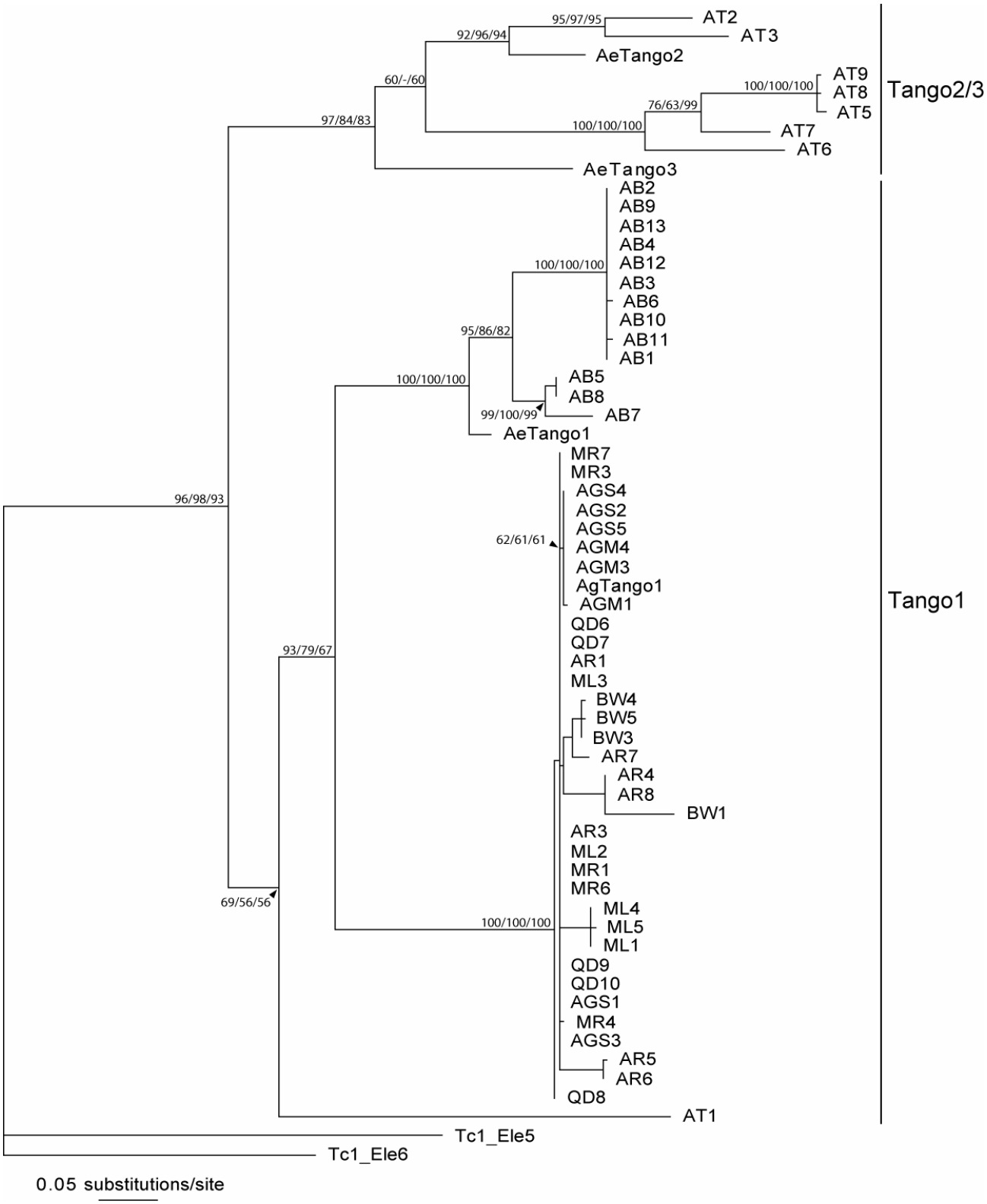

AT2
AT3
AeTango2
AT9
AT8
AT5
AT7
AT6
AeTango3
AB2
AB9
AB13
AB4
AB12
AB3
AB6
AB10
AB11
AB1
AB5
AB8
AB7
AeTango1
MR7
MR3
AGS4
AGS2
AGS5
AGM4
AGM3
AgTango1
AGM1
QD6
QD7
AR1
ML3
BW4
BW5
BW3
AR7
AR4
AR8
BW1
AR3
ML2
MR1
MR6
ML4
ML5
ML1
QD9
QD10
AGS1
MR4
AGS3
AR5
AR6
QD8
AT1
Tc1_Ele5
Tc1_Ele6
95/97/95
92/96/94
60/-/60
100/100/100
76/63/99
97/84/83
100/100/100
100/100/100
95/86/82
100/100/100
99/100/99
96/98/93
62/61/61
93/79/67
69/56/56
100/100/100
Tango2/3
Tango1
0.05 substitutions/site

Figure 3.4. Inferred phylogenetic relationship of *Tango* elements. All phylogenetic analyses were carried out using PAUP* 4.0 b10 (Swofford, 2003). Shown is a maximum likelihood phylogram rooted with related *Tc1* elements. The maximum likelihood analysis was performed using HKY85+G model and the among-site rate variation shape parameter equals 2.2164. These parameters were obtained using Modeltest (version 3.7, Posada & Crandall, 1998). Analysis using neighbor-joining, minimum evolution, and maximum parsimony algorithms produced essentially the same tree as maximum likelihood analysis. Five hundred bootstrap replicates were performed using neighbor-joining, minimum evolution, and maximum parsimony algorithms. The bootstrap values, which are the small numbers at a particular node, are overlaid on the maximum likelihood tree in the order of neighbor-joining, minimum evolution, and maximum parsimony. Values below 50% and values for most groupings within the *An. gambiae* complex clade were not shown. Uncorrected distance was used for neighbor-joining and minimum evolution. Bootstrap of maximum parsimony analysis was conducted using the heuristic search algorithm with 10 random sequence additions per replicate and tree-bisection-and-reconnection branch swapping. Ten random additions per replicate are sufficient because the best tree was identified in the first addition in two separate analyses using 100 random additions. The nucleotide alignment used to generate the trees was based on the corresponding amino acid alignment in the following manner: Amino acid sequences for *AgTango1*, *Tc1_Ele5* and *Tc1_Ele6* corresponding to the region amplified by degenerate PCR were aligned using ClustalX version 1.83 (Thompson *et al.*, 1997) using default settings. The alignment was inspected by eye, and after minor adjustments, was used to align the corresponding nucleotide sequences of these elements using CodonAlign 2.0 (Hall, 2004). Using this alignment as a guide, the nucleotide sequence data for the species surveyed using degenerate PCR were added and aligned. This

resulting alignment was used as input for phylogenetic analyses. AB=*Aedes albopictus*, AGM=*Anopheles gambiae* molecular form M, AGS=*An. gambiae* molecular form S, AR=*An. arabiensis*, AT=*Ochlerotatus atropalpus*, BW=*An. bwambae*, ML=*An. melas*, MR=*An. merus*, QD=*An. quadriannulatus*. Numbers after the species name refer to the clone number. Sequences have been deposited to GenBank under accession numbers EF423994-EF424048.

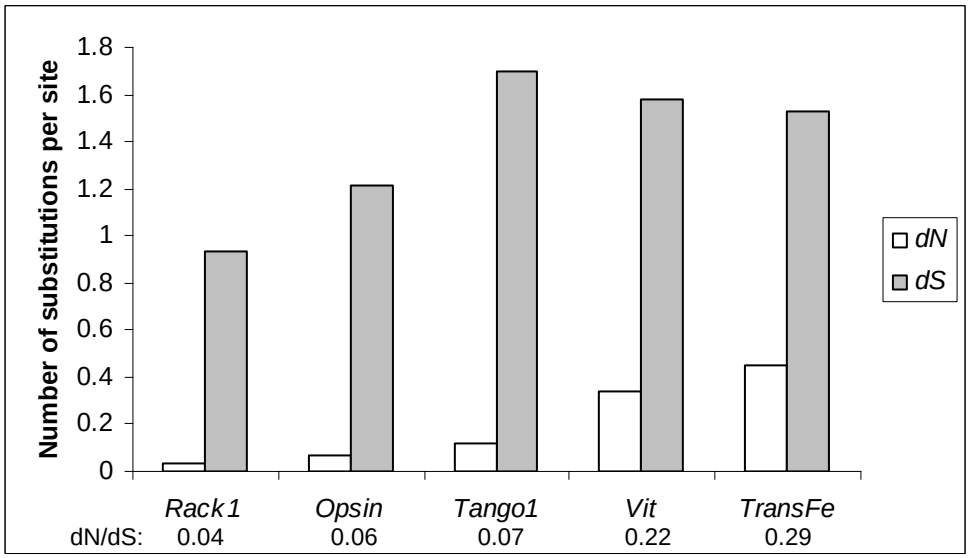


Figure 3.5. Comparison of selection pressure on *Tango1* and four host genes of *Anopheles gambiae* and *Aedes aegypti*. *dN* and *dS* values were determined using SNAP (http://hcv.lanl.gov/content/hcv-db/SNAP/SNAP.html; Korber, 2000). *d*S=mean number of substitutions per synonymous site; *d*N=mean number of substitutions per nonsynonymous site. Vit=Vitellogennin, TransFe=Transferrin.

TABLE 3.1. Percent amino acid identities between *Tango* transposons, three other *Tc1* elements from *Aedes aegypti*, and representative *Tc1* elements *Quetzal* and *Sleeping Beauty*.

	AeTango1	***AeTango2***	***AeTango3***	***Tc1_Ele4***	***Tc1_Ele5***	***Tc1_Ele6***	***Quetzal***[a]	***SB***[b]
AgTango1	79.9	61.3	59.1	48.6	47.7	45.9	37.1	38.0
AeTango1	-	62.5	59.1	48.6	47.1	46.9	38.3	38.6
AeTango2	-	-	66.9	45.0	45.5	45.6	36.6	36.9
AeTango3	-	-	-	43.2	44.8	42.3	36.2	35.6
Tc1_Ele4	-	-	-	-	38.1	36.2	33.8	32.9
Tc1_Ele5	-	-	-	-	-	45.8	37.8	43.2
Tc1_Ele6	-	-	-	-	-	-	34.2	32.4
Quetzal	-	-	-	-	-	-	-	36.2

[a]Ke *et al.*, 1996

[b]SB = *Sleeping Beauty*; Ivics *et al*., 1997

TABLE 3.2. The molecular characteristics and copy numbers of *Tango* transposons from *Aedes aegypti* and *Anopheles gambiae*, and three other *Tc1* elements from *Ae. aegypti*.

Transposable Element	Molecular Characteristics					Copy Number	Position of Rep TE		
Name	TIR Length (bp)	Length (bp)	TSD	ORF (aa)	First 24bp of 5' TIR	Total (potentially active)	Contig/ Scaffold	Start	End
AeTango1a	225	1659	TA	340	CACTGATAGGCAAAATAAAGTGCC	7(2)	6525	23825	25483
AeTango1b	779	1589	TA	NA[a]	CACTGATAGGCAAAATAAAGTGCC	441	15269	43406	44994
AeTango2	224	1665	TA	336	CAGTGACCGGCACAAAAAAAATCC	25 (1)	5660	81	1745
AeTango3[b]	ND	ND	ND	ND	ND	19	22747	6575	7471
AgTango1	224	1670	TA	338	CAGTGGCCGGCAAAATAAAGTGGC	38 (1)	AAAB01008960	16265453	16267122
Tc1_Ele4	26	2406[c]	TA	ND	CAGTACTGGACAAAAAAAAGTACG	4	15410	12848	15253
Tc1_Ele5	21	1660	TA	334	CAGTCAGTGACAAAAGTTAGT	4	11146	68274	69933
Tc1_Ele6[b]	ND	ND	ND	ND	ND	4	8087	102991	103926

[a]Members of *AeTango1b* are degenerate and do not have intact ORFs. See text for details.

[b]No full length copies of *AeTango3* and *Tc1_Ele6* were identified, therefore their molecular characteristics could not be determined.

[c]*Tc1_Ele4* has two insertions from other transposable elements. One disrupts the 5' end of the ORF and the other is 3' of the ORF. Without these insertions, the representative copy would be 1566 bp in length.

TABLE 3.3. Distribution of *Tango* elements among mosquito species surveyed using degenerate PCR.

Subfamily	Species	Subgenus	PCR Product	Tango/Total Clones
Anophelinae	*Anopheles arabiensis* (AR)	*Cellia*	+	6/6
	An. bwambae (BW)	*Cellia*	+	4/6[a]
	An. gambiae M (AGM)	*Cellia*	+	3/4[b]
	An. gambiae S (AGS)	*Cellia*	+	5/5
	An. melas (ML)	*Cellia*	+	4/5[b]
	An. merus (MR)	*Cellia*	+	4/5[b]
	An. quadriannulatus (QD)	*Cellia*	+	5/6[a]
	An. stephensi (ST)	*Cellia*	+	0/23[b]
	An. dirus (DI)	*Cellia*	-[c]	N/A
	An. farauti (FA)	*Cellia*	-[c]	N/A
	An. albimanus (AL)	*Nyssorhynchus*	+	0/5[b]
Culicinae	*Aedes albopictus* (AB)	*Stegomyia*	+	13/13
	Ochlerotatus atropalpus (AT)	*Ochlerotatus*	+	8/8
	Culex pipiens quiquefasciatus (CX)	*Culex*	+	0/6[b]

[a]Two clones from *An. bwambae* were *Quetzal*-like (Ke *et al.*, 1996), and one clone from *An. quandriannulatus* was a *Fot-1*-like (Daboussi *et al.*, 1992) transposon

[b]Clones that were not *Tango* were non-TE sequences

[c]PCR with degenerate glutamate dehydrogenase primers produced positive results, showing that the genomic DNA was intact

CHAPTER 4

Inter-plasmid Transposition Assay to Test Functionality of *AgTango* Transposase

4.1 Abstract

AgTango is a *Tc1* DNA transposable element (TE) from the African malaria mosquito, *Anopheles gambiae*. Recently, we discovered several *Tango* transposons in the distantly related mosquito species, *Aedes aegypti*, at high amino acid sequence identity (79.9%). This observation led us to question whether or not the *AgTango* transposase was functional. An endogenous DNA TE from multiple mosquito species that is capable of transposition in those species would be highly useful for studying TE behaviour, host interaction, and regulation through comparative analyses. A functional copy of *Tango* would enable us to investigate the mechanisms underlying these differences. Therefore, we sought to test *AgTango's* transposase functionality in an inter-plasmid transposition assay in cell culture. While the positive control for this assay system proved non-functional, a number of obstacles were overcome in eventual establishment of this valuable assay system in our laboratory. No transposition recombinants were observed for *AgTango*. Remaining problems concerning the system, and future work with *AgTango* are discussed.

4.2 Introduction

In a survey of the DD34E TEs of the African malaria mosquito, *Anopheles gambiae*, we identified several potentially active DNA transposons using `TEpipe` (Biedler & Tu, 2003; Coy & Tu, 2005). One of these TEs, *AgTango*, was subsequently identified through bioinformatics analyses of the distantly related mosquito species, *Aedes aegypti*, at high sequence identity (79.9%; Coy & Tu, 2007). Given the estimated divergence time between these two mosquito species of 145-200 MYA (Krzywinski *et al.*, 2006), it is unlikely that this TE was transmitted vertically, but rather horizontally, between the ancestors to the extant species at some distant point in evolutionary time. Through a degenerative PCR approach, we surveyed a number of mosquito species, and discovered that *An. gambiae* and *Ae .albopictus* have only one *Tango* transposon, while *Ae. aegypti* and *Ochlerotatus atropalpus* have at several different *Tango* transposons. Furthermore, no *Tango* transposons were identified in *Culex pipiens quinquefasciatus* or in Anophelines outside the *Anopheles gambiae* species complex. These data support the hypothesis that *Tango* has been horizontally transferred between mosquito species, and as such, suggest that *Tango* may retain the capability of transposition. An element that can transpose in numerous mosquito species would facilitate studies concerning differential dynamics of TEs like that observed for *Tango*.

AgTango is a *Tc1* TE that is 1670 basepairs (bp) in length, with an open reading frame (ORF) coding for a putative 338 amino acid transpose in a single reading frame (Fig 4.1). It has 224/223bp imperfect terminal inverted repeats with short direct repeats of approximately 17-18bp within each TIR. This TIR structure, known as indirect repeat/direct repeat (IR/DR), is found in the reconstructed, active *Tc1* TE, *Sleeping*

Beauty (Izsvak *et al.*, 1995; Ivics *et al.*, 1996; reviewed in Plasterk *et al.*, 1999). These direct repeats have been shown to be the cores of the binding sites for the *Sleeping Beauty* transposase (Ivics *et al.*, 1997), and unlike the *Tc3* element in which the internal direct repeats are not necessary for transposition (Fischer *et al.*, 1999), elimination of the internal repeats in *Sleeping Beauty* obliterates transposition activity (Izsvak *et al.*, 2000). The N-terminal domain of the *Tango* transposase has two predicted HTH motifs (Fig. 4.1B and C), termed 'PAI' and 'RED', which are involved in the binding of the DNA substrate (Vos *et al.*, 1993; Colloms *et al.*, 1994; Pietrokovski & Henikoff, 1997; Watkins *et al.*, 2004). The C-terminal domain contains the DD34E catalytic triad which is required for transposition (Vos *et al.*, 1993; van Luenen *et al.*, 1994; Vos & Plasterk, 1994; Lohe *et al.*, 1997). Additional features of the putative *Tango* transposase, common to other *Tc1* TEs, include a nuclear localization signal, a GRPR sequence, and glycine-rich box (Fig 4.1b). The GRPR sequence, an AT-hook-like motif, is a conserved feature of *Tc1* elements, and is believed to mediate substrate binding in coordination with the 'PAI' and 'RED' motifs by contacting the DNA in the minor groove of AT base pairs (Izsvak *et al.*, 2002). The function of the glycine-rich is currently unknown.

In an effort to determine if *AgTango* was functional, we employed an inter-plasmid transposition assay carried out in *S2* cell culture as described by Arensburger *et al.* (2005). This system employs three plasmids, a helper which provides the transposase, a donor which supplies the substrate, and a target with selectable markers to 'capture' and report transposition events. We obtained the plasmids from Arensburger to serve as a positive control, and as starting material for the *AgTango* constructs. Because of the

ambiguous nature of the pBSHvSacKOα plasmid (*Donor Plasmid* - described below), efforts to characterize this plasmid were also undertaken.

4.3 Materials and Methods

4.3.1 Overview of the Inter-plasmid Transposition Assay

To test *AgTango's* transposition activity, an *in vitro* inter-plasmid assay in cell culture was employed (Sarkar *et al.*, 1997; Arensburger *et al.*, 2005). The system is comprised of three plasmids: a helper, donor and target. The helper provides the transposase being tested under the control of an inducible promoter. The donor provides the substrate for the transposase, namely a cassette with three selectable bacterial markers flanked by the cognate TIRs of the transposase. This cassette contains a gene for kanamycin (Kan) resistance, an *Escherichia coli* origin of replication (ORI), and a coding region for the α-peptide from the β-galactosidase gene of *E. coli* (*lacZ* α). Outside of the cassette resides a gene for ampicillin (Amp) resistance. The target is a plasmid that is unable to replicate in *E. coli* and carries a gene for chloramphenicol (Cam) resistance. These plasmids are transfected into cell culture, and the transposase is expressed by inducing the promoter. The transposase, if functional, should recognize the TIRs of the donor cassette and transpose the entire construct to new locations, one of which being the target. Upon receipt of the donor cassette, the target plasmid becomes capable of amplification in *E. coli*. *E. coli* carrying the target plasmid then become Kan and Cam resistant. The DNA is isolated from the cell culture and is electroporated into *E. coli*, which is then plated onto Kan/Cam plates. A small amount is also plated onto Amp plates to determine the donor 'titer', which is used to calculate transposition rate. Colonies that grow on Kan/Cam

plates are recombinants between the donor cassette and the target. To distinguish transposition from other recombination events, the colonies are subcloned and the plasmids are recovered. A *Pst*I restriction enzyme digest is performed, which produces a characteristic digest pattern for transposition events. Those plasmids with the correct digestion pattern are then submitted for sequencing. Transposition events can be further distinguished by the characteristic 'TA' target-site duplications generated through transposition of *IS630-Tc1-mariner* TEs.

The plasmids used for making the *AgTango* constructs were obtained from Dr. Peter Atkinson (Arensburger *et al.*, 2005), which had been used to test *Herves*, an active Class II TE from *An. gambiae*. In addition to providing starting material for *AgTango* constructs, this system also served as a positive control. Dr. Atkinson generously provided four plasmids, brief descriptions of which are provided below. Detailed descriptions of their construction can be found in Arensburger *et al.* (2005) and Sarkar *et al.* (1997).

4.3.2 *pKhsp70Herves-PEST Helper Plasmid* and *pKhsp70rp Helper Plasmid*

The *pKhsp70Herves-PEST* helper plasmid (*Herves Helper*) was constructed from pK19, and contains the coding region for neomycin phosphotransferase, which confers resistance to neomycin and kanamycin. The consensus sequence for the three intact *Herves* ORFs was used as the coding sequence for the *Herves* transposase (Arensburger, *et al*., 2005). The ORF is flanked by the 5' and 3' regulatory regions of a *Drosophila melanogaster hsp70* (Karch *et al.*, 1981; Knipple & Marsella-Herrick, 1988). The *hsp70* promoter is strongly upregulated by heat stress and other environmental stressors such as

heavy metals, and has been demonstrated to drive heterologous gene expression in S2 cells after being subjected to heat-shock. *pKhsp70rp* helper plasmid (*Helper Empty*) duplicates that sequence of the *Helper Herves,* minus the *Herves* ORF. Instead, between the 5' and 3' regions of *hsp70*, a small multiple-cloning site is inserted to facilitate the cloning of the DNA of interest. Both plasmids were amplified in OneShot Top10 *E. coli* cells (Invitrogen Corp., Carlsbad, CA) following manufacturers instructions.

4.3.3 *pBSHvSacKOα Herves Donor Plasmid*

The *pBSHvSacKOα Herves* donor plasmid (*Herves Donor*) was constructed by amplifying the entire *Herves* sequence and cloning it into pBluescriptSK+ (Stratagene, La Jolla, CA). The *Herves* ORF was then replaced by DNA encoding three genetic markers: a gene conferring kanamycin resistance, a ColE1 ORI, and the *α*-peptide coding region from the *β*-galactosidase gene of *E. coli* (*lacZ*). This produced a *Donor Cassette* containing selectable gene markers flanked by the TIRs of *Herves*. *Herves Donor* also contains a gene outside the *Donor Cassette*, conferring ampicillin resistance to aid in the propagation of the plasmid, and the determination of donor titer in the transposition assay (Arensburger *et al.* 2005, Sarkar *et al.*, 1997). DH5α *E. coli* cells (Invitrogen Corp.) were used for amplification of the donor plasmid per manufacturer's instructions.

The sequence for the *Herves Donor* as reported by Arensburger *et al.* (2005) contained an ambiguous region of approximately 1260bp that started right outside near the 3' end of Herves 5' TIR and continued all the way through the kanamycin resistance gene (Fig 4.2A). To elucidate the nucleotide sequence in this region, three sequential rounds of sequencing were performed, the primers for which are listed in Table 4.1.

Because of the uncertainty of the original sequence, multiple initial primers were designed for initial sequencing (Table 4.1). Subsequent primers were based on sequence data obtained from the previous round. To test the functionality of the *Donor* cassette, the kanamycin resistance and the *lacZ* gene, and the ColE1 ORI, the *Donor Plasmid* was digested with *Spe*I, and the fragment containing the 3' end of *Herves* 5' TIR through the ColE1 ORI (Fig 4.2A) was religated and used in transformation of OneShot Top10 cells (Invitrogen Corp.). The bacterial were plated onto LB/X-GAL/IPTG/kan (25ug/ml) plates.

4.3.4 *pGDV1 Target Plasmid*

pGDV1 is a cloning vector containing an ORI from a plasmid derived from *Corynebacterium xerosis* and a chloramphenicol resistance gene (Bron, 1990). It is replicated at very high copy number in *Bacillus subtilis*, and is incapable of replicating in *E. coli.* Recombination with the *Donor Cassette* should confer the ability of this plasmid to be replicated in *E. coli*. Dr. Atkinson's laboratory provided plated *B. subtilis* colonies on LB plates carrying the *pGDV1* plasmid.

4.3.4 Genomic DNA

Genomic DNA was isolated from eight individual *An. gambiae* mosquitoes by homogenization in 150ul DNAzol/1.5ul polyacryl carrier in 1.5ml tubes following the manufacturer's instructions (MRC; Cincinnati, OH). DNA was resolubilized in 50ul 0.1 X TE. One microlitre from each isolate was combined into a pool which was used as a template in all PCR reactions except where noted. *An. gambiae* mosquitoes (G3 strain)

were obtained from Malaria Research and Reference Reagent Resource Center (MR4; http://www.mr4.org/).

4.3.5 PCR

PCR was carried out using an Eppendorf Mastercycler Thermocycler 5333 (Eppendorf Scientific, Inc., Westbury, NY). TaKaRa *rTaq* DNA polymerase (TaKaRa Bio, Inc.; Otsu, Japan) was used in all reactions. Originally, *Pfu* High Fidelity Polymerase (Stratagene, La Jolla, CA) was used in PCR reactions, but in some cases would not amplify the PCR product from An. gambiae genomic DNA for an unknown reason(s). Sephglas BandPrep (Amersham Biosciences Corp.) was used to gel purify PCR products from agarose gels per manufacturer's instructions. Except where noted, PCR products were ligated into pGEM T-Easy vector (Promega Corp.), and ligation products were then introduced into OneShot Top10 cells (Invitrogen Corp.) following manufacturer's instructions. Plasmids were isolated using the Wizard *Plus* Miniprep DNA Isolation System (Promega Corp.). Cloned PCR products were sequenced by Virginia Bioinformatics Institute, Virginia Tech (Blacksburg, VA). All primers used for making and verifying the *Tango* constructs are listed in Table 4.2.

A nested PCR strategy was employed for amplifying *Tango's* ORF and TIRs in order to insure amplifying the specific genomic copy of *Tango* selected for study, specifically the only copy possessing intact TIRs and ORF. In the first PCR, at least one of the two primers targeted DNA flanking the *Tango* element in scaffold number AAAB01008960. In the second round of PCR, primers were designed that included the appropriate restriction enzyme recognition sites on their termini to clone the fragments

into the backbones of *Herves* plasmids. Templates for the second round of PCR were 25ug of the first PCR product ligated pGEM T-Easy vector. The PCR scheme that was used for genomic DNA templates was two cycles of 94°C for 2min, annealing temperature (T_m) for 45s, extension time (E_T) at 72°C followed by 29 cycles of 94°C for 30s, (T_m) for 45s, and E_T at 72°C with a final extension time of 72°C for 5min. For nested PCR templates in pGEM T-Easy vector, a PCR scheme of 30 cycles of 94°C for 30s, (T_m) for 45s, and E_T at 72°C with a final extension time of 72°C for 5min was employed. T_ms and E_Ts (T_ms / E_Ts) are listed below. Standard molecular cloning techniques were used to create the *Tango* plasmids as described in (Sambrook *et al.*, 1989) unless otherwise noted.

4.3.6 Construction of *Tango Helper Plasmid*

The template for *Tango's* ORF was PCR-amplified using TangoTemp F2/R2 primers (57°C/2min) and 'TA' cloned into pGEM T-Easy (Promega Corp.). After sequence verification, 25ug of the template was used in PCR of *Tango's* ORF (58°C/2min). The 5' primer, TangoORF-F1-*Xma*I, included the recognition sequence for the *Xma*I restriction enzyme and was designed to insure the inclusion of the Kozak sequence (`[G/A]NNATGG`). The 3' primer, TangoORF-R1-Stop-*Bam*HI included a recognition site for *Bam*HI on its 3' end. Once the sequence of this product was verified, it was cloned into the *Helper Empty* plasmid using standard molecular cloning techniques. The construction of the resulting plasmid, *Tango Helper*, was verified with primers pKhsp70-F1/R2.

4.3.7 Construction of the *Tango Donor Plasmid*

Templates for the 5' and 3' ends of *Tango* were PCR-amplified and subcloned separately, with primers TangoProTemp-F1/R2 (59°C/1min30s) and Tango3'Temp-F1/R1 (61°C/45s), respectively. After sequence verification, the left end of *Tango*, including bases 1–385 plus the flanking 'GTTA' from *An. gambiae*, and the right end, including 1374-1670 plus the flanking 'TATACA' on the 3' end, were PCR-amplified with Tango5'TIR-F1-*Kpn*I/ Tango5'TIR-R1-*Xcm*I (61°C/30s) and Tango3'TIR-F1-*Bam*HI/Tango3'TIR-R1-*Rsr*II (62°C/30sec), respectively. The 5' PCR fragment included the TIR and all intervening sequence up to the 'ATG' start codon of the ORF. The 3' end included the last 22 nucleotides of the ORF through to the end of the 3' TIR. *Tango's* 5' and 3' ends were transferred into the *Donor Plasmid* as *Kpn*I and *Xcm*I and *Bam*HI and *Rsr*II fragments, respectively. This resulted in the removal of all of the *Herves'* TIRs except for a short fragment of approximately 45bp of the 3' end of the 5' TIR (Fig. 4.2B). This was due to the fact that there were no unique RE sites in this region that would allow for the complete removal of the 5' TIR of *Herves*. *Tango Donor* was verified using the following primers TangoDonor-F1/F2/F3 and R1/R2.

4.3.8 S2 Cells

Drosophila melanogaster S2 cells (Schneider, 1972) were maintained at 23°C in Schneider's *Drosophila* media supplemented with 10% heat-inactivated FBS, and 0.2U/ml and 100ug/ml of penicillin and streptomycin, respectively (all reagents from Invitrogen Corp). Initial S2 cultures were obtained from ATCC (Manassas, VA) at

passage number 517, and all transposition assays were performed in cell passages of 540 or less.

4.3.9 Transposition Assay

Plasmids were purified using the PureYield Plasmid Midiprep System (Promega Corp.). Prior to the day of transfection, $5x10^6$ S2 cells were plated in each well of a 6 well plate. The next day, transfection mixes were prepared for each treatment as follows: 7.5ul of Cellfectin Transfection Reagent (Invitrogen Corp.) was mixed in 100ul of serum free media (SFM) by inversion and set aside. Helper, donor and target plasmids (2.5, 2.5 and 5.0ug, respectively) were added to 100ul of SFM in a 1.5ml microfuge tube and mixed by gentle inversion. To control for transfection conditions, one well of cells was transfected with a plasmid carrying *lacZ* gene under the control of the baculovirus promoter, *OpIE2*. Cellfectin/media mix (107.5ul) was added to each tube, mixed by inversion, and allowed to sit at room temperature for 20 minutes. Next, 800ul of SFM was added to each tube and mixed by inversion.

Cells were washed once with 2ml SFM, overlaid with transfection mix, and allowed to incubate at 23°C for 10 hours. Afterwards, the transfection mix was removed and replaced with complete media (FBS plus antibiotics – see above). The plate was sealed with parafilm, and placed in the 23°C incubator overnight. The next day, cells were placed in pre-warmed oven set at 42°C for 2 hours. Afterwards, they were placed back in the 23°C incubator overnight. The following day, cells were harvested by scraping, and DNA was isolated using Promega's Wizard Genomic DNA Purification Kit following manufacturer's instructions. Purified DNA was reconstituted in 30ul of sterile,

nuclease-free water. Four microlitres of the DNA preparation were used in the electroporation of 40ul of DH10β *E. coli* bacterial cells (Invitrogen Corp.) in 2mm cuvettes, with 2.5kV, 25uF and 200Ω. Cells were allowed to recover in 1ml of SOC media in a shaking incubator set at 225 rpm and 37°C for 1 hour, after which 5ul was spread onto LB/X-GAL/IPTG/Amp (100ug/ml) plates. The remainder was spun down and all but approximately 100ul of the overlying SOC media was removed. Cells were gently resuspended by vortexing in the remaining SOC media, and spread onto three LB/X-GAL/IPTG/Kan/Cam plates (15ug/ml and 10ug/ml respectively). The Amp plates, which were used to determine donor titer, were incubated overnight and the Kan/Cam plates were incubated for three days. Blue colonies from the Kan/Cam plates were picked and used to inoculate 3ml of LB broth supplemented with 25ug/ml kanamycin. Subcultures were used to inoculate LB broth with 100ug/ml ampicillin for counter-selection. Plasmids were isolated from LB/Kan growth and subjected to *Pst*I restriction digest. Plasmids with the correct digestion pattern were submitted to VBI for sequencing with HervesDonor-F1/F2/F3 and R1/R2 (Table 4.3).

Cells transfected with the *OpIE2:lacZ* construct were assayed for β-galactosidase activity using Invitrogen's B-GAL Staining Kit following manufacturer's instructions.

4.3.10 Bioinformatic Analyses

Helix-turn-helix motifs were predicted with PROF predictions (Rost & Sander, 1993; Rost *et al.*, 1996) at http://www.predictprotein.org. ClustalX for Windows v. 1.83 (Thompson *et al.*, 1997) was used for making alignments. Assembly of the Donor

Plasmid sequence and maps of plasmids were created with SeqMan Pro and SeqBuilder, respectively, from the DNA* Lasergene suite, version 7.1.0.

4.4 Results

4.4.1 Analysis of the *pBSHvSacKOα Herves Donor Plasmid*

The reported sequence of the *Donor Plasmid* indicated that a *Pst*I restriction digest should produce three fragments, 622, 894 and 7626bp (Fig. 4.2). This was confirmed empirically (Fig. 4.3, Lane 2). The 622 and 894bp fragment arise from within the *Donor Cassette*, and should be common to all recombinant plasmids. In Arensburger (2005), only the 622 fragment is mentioned as a diagnostic feature of recombination between the *Donor Cassette* and the target plasmid.

Although the name of the plasmid implies that it contains a coding region for sucrase, no such ORF was found upon the analysis of the supplied sequence. Instead, the sequence for a large portion of the *Hermes* TE was found, including both TIRs and part of the ORF of that element (Fig 4.2A). In addition, the sequence contained an ambiguous region of approximately 1260bp in size which started immediately outside of the 3' end of *Herves*' 5' TIR and continued all the way through the kanamycin resistance gene (Fig 4.2A). A number of preliminary restriction digests produced inconsistent results with the predicted digest patterns based on the sequence (data not shown). Due to these ambiguities and inconsistencies, the region that extends from *Sph*I restriction enzyme site in the middle of the *Herves'* 5' TIR through the *Pst*I site just before the ColE1 origin of replication (ORI) in the *Donor Cassette* (Fig 4.2A) was sequenced. The sequencing of the ambiguous region revealed that it consists mostly of the kanamycin resistance gene (Fig.

4.2B). Moreover, *Hermes* was not present in the plasmid as indicated in the supplied sequence. Lastly, the plasmid is approximately 1.2kbp smaller than reported.

4.4.2 Verification of *Tango* Constructs

Sequencing of the template for the *AgTango* ORF revealed two transitions in nucleotide sequence resulting in two codon changes: GTT→ATT and CGA → CAA. To confirm that these changes were not the result of errors in the amplification process, another PCR reaction was carried out, and the same sequence changes were observed in the product. The nucleotide differences produce two corresponding changes in the predicted amino acid sequence of *AgTango*, V16I and R134Q (Fig. 4.4 and 4.1C). Because valine → isoleucine is a conservative change, and because the site affected is near the N-terminus of the transposase, the probability that it would affect transposition activity appears to be small. However, the arginine → glutamine change lies between the NLS and the first 'D' of the DD34E catalytic domain. Moreover, arginine is a significantly larger residue than glutamine, and if protonated, would introduce a positive charge at a previously neutral position. The corresponding residue in *AeTango1* is a lysine (see Fig. 3.1A), also a positively charged residue. The location and nature of the substitution suggested a strong probability that catalytic function might be affected, therefore, efforts to correct this change for future comparisons between the two versions of the transposase were planned.

The integrity of the *Tango Helper Plasmid* was confirmed by a double digest with *Bam*HI and *Xma*I. The digest produced a fragment slightly larger than 1kb as expected. The plasmid was sequenced with primers (pKhsp70-F1/R2) designed against the *Helper*

Empty backbone facing inwards towards the inserted ORF of *AgTango*. The data produced from this sequencing spanned the entire *AgTango* ORF and the adjacent DNA of the plasmid, confirming that 1) The entire ORF sequence was present and correct and 2) it was in the correct position and orientation.

In a similar fashion, the integrity and sequence of the *Tango Donor Plasmid* was verified by restriction digests and DNA sequencing. To confirm the identity of the 3' TIR insert, *Tango Donor* was digested with *Bam*HI and *Rsr*II restriction enzymes, producing the fragment of approximately 300bp as expected. Likewise, the 5' TIR insertion was verified by digestion with *Xcm*I and *Kpn*I, producing an expected fragment of approximately 400bp. Sequencing showed that both inserts were correct in sequence and in orientation. However, sequence data outside of the 5' end of *AgTango's* 3' TIR suggested that this region is not as reported.

4.4.3 No Transposition Detected for *Herves* – No Positive Control

Only general recombinant products were obtained for the *Herves* system. In three experiments, 9.9×10^5 plasmids were screened, out of which 17 recombinants were recovered (Table 4.4). Any clones passing the ampicillin cross-selection test which had *Pst*I restriction fragments of approximately 600 and 900bp, regardless of the number or size of the other fragments, were sequenced. A total of three clones were sequenced, all of which were determined to be non-transposition recombinants (*e.g.*, Fig 4.5). Arensburger (2005) observed 12 transposition reactions in a screening of 5.5×10^5 plasmids. Based on these results, we would expect to have observed transposition products in our experiments. A number of controls were employed to identify underlying

problems with this system. In particular, cells were transfected with a plasmid construct expressing the *lacZ* gene under the *OpIE2* baculovirus promoter, and an ampicillin resistance gene. This control served two purposes, one to make certain that transfection conditions were appropriate for the uptake of plasmid DNA into S2 cells, and to verify our DNA isolation conditions. In cultures transfected with this plasmid, approximately 8% of the cells were stained (26 blue cells/331 total cells). This is within the typical range of 5-10% observed in Dr. Atkinson's laboratory (Rob Hice, personal communication). No staining was observed in non-transfected cells. These observations indicated that our transfection conditions were appropriate. Furthermore, colonies were obtained on LB/Amp plates, indicating that the plasmids were reisolated back from the S2 cells.

4.4.4 No Transposition Events Detected for *Tango*

An estimated 5.5×10^5 plasmids were screened in our *Tango* transposition assay (Table 4.4). Five recombinants were obtained, one of which was ampicillin sensitive and yielded a passable *Pst*I digest. However, as with *Herves*, sequencing revealed this was a non-transposition recombinant.

4.5 Discussion

4.5.1 Potential Problems with the Transposition Assay

The three most critical steps of the assay are 1) transfecting the plasmids into S2 cells, 2) isolating the plasmids from S2 cells after allowing time for transposition to take place, and 3) transfecting the plasmids into *E. coli* to screen for transposition products. We

know that the initial transfection was successful because 5-10% of the cells transfected with the *OpIE2:lacZ* construct exhibited *β*-galactosidase activity (Rob Hice, personal communication). Moreover, we were able to isolate *Donor Plasmids* from the S2 cells and transfected them into *E. coli* by electroporation. Because of these observations, it seems reasonable to assume that these steps are working properly.

Another critical step in the assay is the heat shock treatment. Induction of a heat-shock response is necessary for the expression of the transposase. No control was included in the assay for verifying this step. The heat shock response in S2 cells, as in whole organisms, is temperature and time dependent. It is also affected by whether heating is gradual or rapid (Lindquist, 1980). Additionally, the optimal temperature and the magnitude of the heat shock response depend upon the temperature at which the cells were previously grown (Lindquist, 1980). Heat shock response is typically measured by one of two methods, by directly monitoring the expression of heat-shock proteins (Hsps) themselves or by the fusion of the *hsp* promoter of interest to a reporter gene. By measuring Hsp expression, Lindquist (1980) determined that the optimal Hsp70 response in S2 cells for gradual temperature increase (2°C/15 minutes) occurred at 36-37°C for 1 hour. When the cells were rapidly exposed to high temperature, the maximum response was observed at 37°C with little response at 39°C and above. The implication of this research on the technical aspects of the transposition assay suggests that, based on the growth conditions being used to maintain the S2 cells (23°C), and the type of heat-shock treatment to which they were subjected (rapid), the optimal heat-shock temperature should be around 37°C for 1-2 hours, not 42°C for 2 hours as noted in Arensburger *et al*. (2005). Because heat-shock response can depend on a number of variables, including

subtle differences between batches of media (summarized in Lindquist 1980), the optimal heat-shock temperature and time should be determined empirically for our laboratory through the use of an *hsp70*:reporter gene construct. This reporter construct should also be included in future experimentation to control for this critical aspect of the assay.

4.5.2 *Donor Plasmid*

A number of ambiguities were resolved and corrections made to the working sequence of the *Donor Plasmid*, although nothing was discovered that would indicate that there was anything functionally wrong with the plasmid itself. However, because the transposition assay did not produce transposition products, sequencing efforts of the plasmid should be continued in order to confirm its integrity. Having the correct sequence of the *Donor Plasmid* is a prerequisite for both successful trouble-shooting and construct design. If continued use of this plasmid is planned a number of changes should be considered, including the removal of homologous regions within the *Donor Plasmid*. This should reduce chances of homologous recombination and would decrease the size of the plasmid. In particular, the extra ColE1 ORI outside the cassette should be excised. There are also several regions outside of the cassette that contain remnants of the *lacZ* gene that could be eliminated (Fig. 4.2B). Insertion of a multiple cloning site at the 3' end of the 5' TIR would make the system more amenable for exchanging of TIRs. Whether it would be more efficient to reengineering the existing *Donor Plasmid*, or to start over from scratch needs to be evaluated.

4.5.3 *Tango* Transposition Assay

Given that the positive control did not yield transposition products, no firm conclusions can be drawn from the failure to isolate transposition products using *AgTango*. It is likely that the problem(s) affecting the positive control are also affecting *AgTango's* transposition assay. However, there are a few points that need to be kept in mind. The choice of what regions to include, beyond that of *Tango's* TIRs, in the donor plasmid must balance size with the risk that the region beyond the TIR may affect transposition, either in a stimulatory or inhibitory fashion. Transposition activity of *Tc1* TEs has been shown to drop off exponentially as the size of the TE increases (Fischer *et al.*, 1999; Izsvak *et al.*, 2000; Karsi *et al.*, 2001). Experimental data from two *Tc1* transposons with the same IR/DR TIR structure, *Minos* and *Sleeping Beauty,* indicate that internal sequences much beyond the TIRs are not required for mobility (Catteruccia *et al.*, 2000; Klinakis *et al.*, 2000). This is in comparison to *PiggyBac* and the *mariner* TE *Mos1*, for which internal sequences are necessary for efficient transposition (Pledger *et al.*, 2004; Li *et al.*, 2005). However, because the regions residing between *AgTango's* TIRs and ORF are short, approximately 200bp total, it was decided to include these regions in the *Tango Donor* construct. *AgTango's* donor cassette, including the TIRs, is approximately 3.7kb, which lies within the practical upper limit of 5kb determined for *Sleeping Beauty* (Izsvak *et al.*, 2000; Karsi *et al.*, 2001). However, based on previous experimentation with other *Tc1* TEs, a lower rate in transposition efficiency would be expected as compared to the native size (Fischer *et al.*, 1999; Izsvak *et al.*, 2000; Karsi *et al.*, 2001). If no transposition products were detected, it may be that the size of the *Donor Cassette* is above *AgTango's* permissive limit, or that transposition rate is too low for detection.

Therefore a negative result cannot be interpreted as evidence of *Tango's* lack of functionality, and an alternative means to test activity may be necessary. In particular, the next step might be to use a system with a smaller substrate (Ivics *et al.*, 1997).

The above brings to point why having a *Donor Plasmid* more amenable to exchange of TIRs is highly desirable. It is not known *a priori* what regions will be required, enhance, or inhibit transposition – in some studies this may actually be the focus. The ultimate goal of the investigation is to determine the functional behaviour of the TE, and elucidating regulatory regions within the element moves towards that goal. A *Donor Plasmid* that has short MCS at each TIR terminus would greatly facilitate these types of studies.

4.5.4 Alternative Methods for Detecting Transposition

The concept behind the inter-plasmid assay is both elegant and simple. This method simplifies efforts in identifying functional transposases through a rapid screening process based on selection markers that significantly reduce the number of false positives, and does not require as many steps for verification of transposition as in other assays. However, if the problems interfering with the inter-plasmid assay cannot be identified and resolved, alternative means to test the functionality *AgTango*'s transposase should be considered. Lack of results in the positive control not withstanding, no transposition products from the *AgTango* transposase in this particular assay system could be for reasons other than a lack of functionality. In particular, because the activity of *Tc1* transposases drops exponentially as the size of the substrate increases, other means of testing activity may be more appropriate for these elements. As stated above, the size

of the *Tango Donor Cassette* falls within the functional range, but activity may be too low for detection. Other transposition assay systems with smaller cassettes might be tried (*e.g.*, Ivics, 1997). Typically, these systems are comprised of two plasmids, one which provides the transposase and one which provides the substrate, which is usually an antibiotic resistant gene flanked by the cognate TIRs. These constructs are transfected into cell culture and transposase expression is induced, as those in the inter-plasmid assay system. Cells are subjected to the appropriate antibiotic, and those surviving the treatment are then assayed for chromosomal insertions by Southern blotting. TE display could also be used in this type of assay. Another means to detect chromosomal insertions is through the use of TE display, an assay system built upon the concept of RFLP (Casa, *et al*., 2000; Biedler, *et al*., 2003). TE display is an attractive choice because it provides a genome-wide analysis of insertions, allows for numerous cell passages or cell types to be analyzed simultaneously, and flanking DNA of the insertions can be determined with relative ease. Indeed, in comparative analyses of the behaviour of a given transposase within different cell types as discussed below, TE display would offer many advantages.

4.5.5 Future Directions for *Tango*

Tango offers a unique opportunity to study TE behaviour in mosquitoes. It resides in divergent mosquito species' genomes and appears to have patchy distribution among mosquito species. Some mosquito genomes have several *Tango* transposons while others only have one (Coy & Tu, 2007). This raises a number of questions. For example, are *Ae. aegypti* and *O. atropalpus* are more 'permissive' of *Tango* transposons? Did one *Tango* give rise to multiple transposons, or were these genomes invaded several times? Did

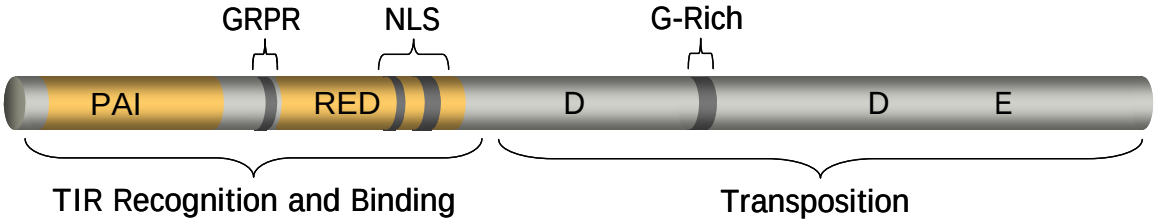
Tango invade *An. gambiae* and was shortly thereafter inactivated before it could diversify? Has the *An. gambiae* genome just not had exposure to other *Tango* transposons? The answer to some of these questions probably lies in the host-TE interaction through host factors and other inhibitory mechanisms. The ideal would be to identify or develop a functional *Tango* transposon that could subsequently be used to investigate the behaviour of the element within these mosquito species to begin to dissect the mechanisms behind the observed differences of this element. We have an abundance of resources at our disposal in terms of live mosquitoes and cell cultures from all angles concerning the *Tango* transposon to begin elucidating the TE-host interaction within mosquito. The key to this research is to get the transposition assay working, and to identify a functional copy of *Tango* that will work in multiple species of mosquito. Doing so will lead to a better understanding of the nature of these elements, and possibly leading to better research and transgenesis tools developed from *Tc1* transposable elements.

4.6 Acknowledgements

We are indebted to Dr. Peter Atkinson's laboratory for the *Herves* plasmids, and to Rob Hice for much assistance and advice with the transposition assay. We also thank Jim Biedler and Ray Miller for stimulating conversations and scientific camradere.

A.

B.



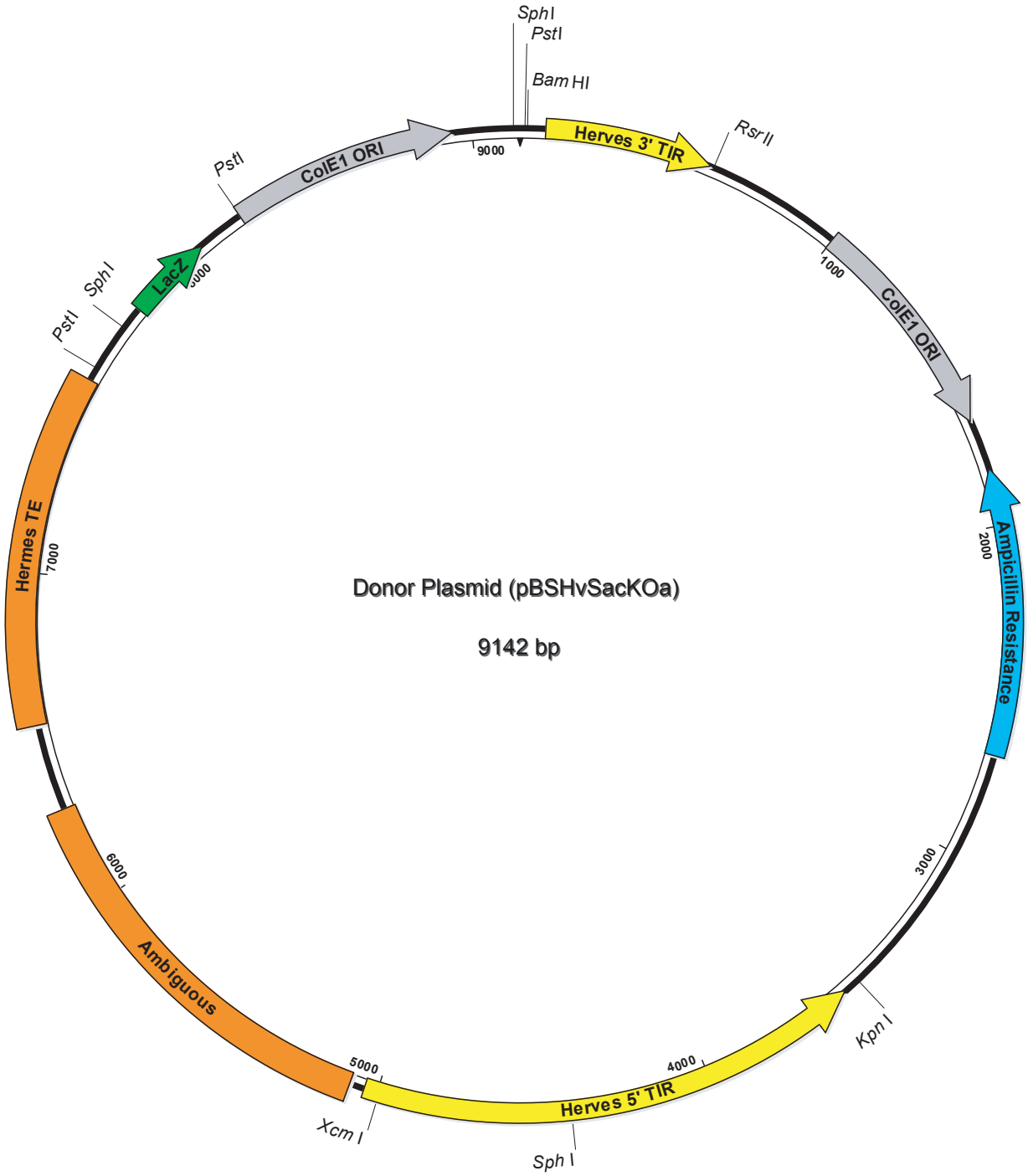
C.

```
                          *                PAI
AgTango1    MENNTKQIPLAVRKLVVRDVQNGESHRAVAGKYSISKSAVGKIFKKYSTLGSVVDRP
SB          M-GKSKEISQDLRKKIVDLHKSGSSLGAISKRLKVPRSSVQTIVRKYKHHGTTQPSY

               GRPR                                RED              NLS
AgTango1    GRGRKRITDSKTDTRIVREVKKNPKVTVLEIKNTLQLN---ISDRTIRRRIIEQGYN
SB          RSGRRRVLSPRDERTLVRKVQINPRTTAKDLVKMLEETGTKVSISTVKRVLYRHNLK

               NLS              *
AgTango1    SKLAKKRPFISKANKSKRLKFAREHADKPLEFWKTVLWTDESKFELFNRKRRSHVWC
SB          GRSARKKPLLQNRHKKARLRFATAHGDKDRTFWRNVLWSDETKIELFGHNDHRYVWR

                                  Glycine-rich Box
AgTango1    KPGEELQERNIQGTVKHGGGNVMVWGCFSWGGAGSLVRINGIMTADTYITILQENLE
SB          KKGEACKPKNTIPTVKHGGGSIMLWCGFAAGGTGALHKIDGIMRKENYVDILKQHLK

AgTango1    VSLIKTGLENKFVFQQDNDPKHTAKKTKTFFNSCRIKPLEWPPQSPDLNPIENLWAI
SB          TSVRKLKLGRKWVFQMDNDPKHTSKVVAKWLKDNKVKVLEWPSQSPDLNPIENLWAE

AgTango1    LDDRVKKTGVTNKDKYFEALENAWENLDPNHIENLVESMPKRLQLVLRSKGGHIKY
SB          LKKRVRARRPTNLTQLHQLCQEEWAKIHPTYCGKLVEGYPKRLTQVKQFKGNATKY
```

Figure 4.1. A. Schematic of *AgTango*, a *Tc1* transposable element from *Anopheles gambiae*. Direct repeats are demarcated by black triangles in the terminal inverted repeats. This terminal-inverted repeat structure, known as an inverted repeat-direct repeat (IR/DR), is also present in the reconstructed and functional *Tc1* transposon,

Sleeping Beauty (Ivics, *et al*., 1997). **B**. *AgTango's* transposase showing predicted functional motifs and domains: GRPR, GRPR box; NLS, nuclear localization signal; G-Rich, glycine-rich box; 'D', 'D', 'E', residues of the catalytic domain. These characteristics are typical features of a *Tc1* transposable element, see text for details. **C.** Amino acid sequence alignment of *AgTango* and *Sleeping Beauty*, showing the above motifs and domains within the sequences demarcated by boxes. In grey are the predicted helix-turn-helix motifs of the paired-like DNA binding domain. In bold are the residues of the DD34E catalytic triad. Asterisks above the sequence alignment indicate residues that were found to be different than that predicted by the sequence found within the *Anopheles gambiae* genome.

A.


SphI
PstI
BamHI
Herves 3' TIR
RsrII
ColE1 ORI
PstI
LacZ
SphI
PstI
Hermes TE
Ambiguous
Xcm I
Herves 5' TIR
Sph I
Kpn I
Ampicillin Resistance
ColE1 ORI
Donor Plasmid (pBSHvSacKOa)
9142 bp

B.

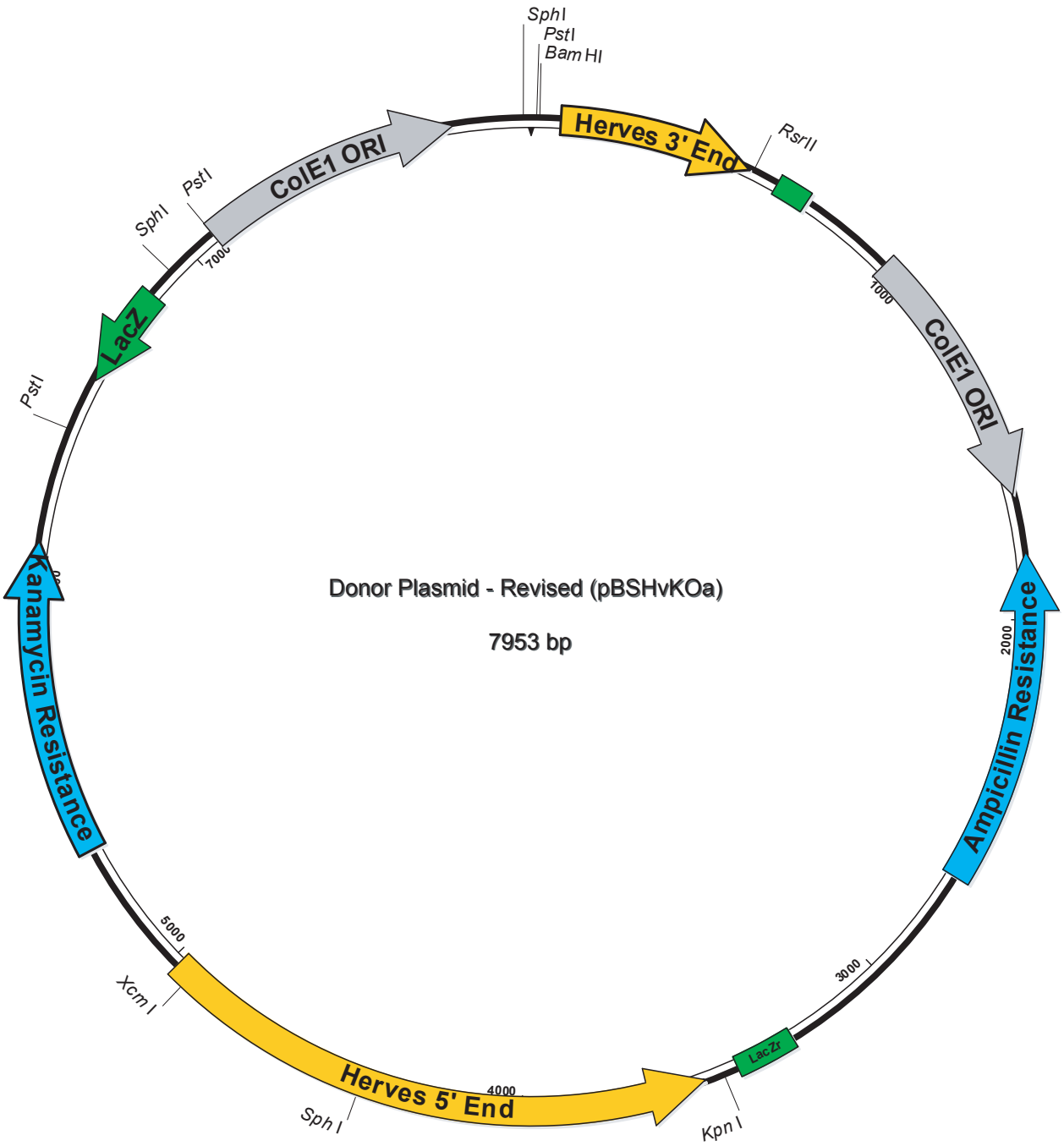

SphI
PstI
BamHI
Herves 3' End
RsrII
ColE1 ORI
1000
ColE1 ORI
SphI
PstI
7000
LacZ
PstI
Donor Plasmid - Revised (pBSHvKOa)
7953 bp
Kanamycin Resistance
2000
Ampicillin Resistance
5000
3000
XcmI
Herves 5' End
4000
LacZr
SphI
KpnI

Figure 4.2. Maps of pBSHvSacOα (*Donor Plasmid*) and pBSHvOα (*Donor Plasmid – Revised*). **A**. Original map was based on the sequence supplied with the plasmid. Sections labeled "Ambiguous" and "Hermes" were the two features that were in question before sequencing. Orientations of features were taken from the map provided with the plasmid. **B**. The revised map of the plasmid after incorporating sequence data. The region that was verified spans from the *Sph*I site in the middle of *Herves* 5' End through the *Pst*I site just outside the ColE1 origin of replication. Other changes include the reversal of the orientation of the *lacZ* gene and the reduction in the overall size of the plasmid from 9142 to 7953bp. Small boxes immediately outside the *Herves* TIRs are partial *lacZ* coding sequences.

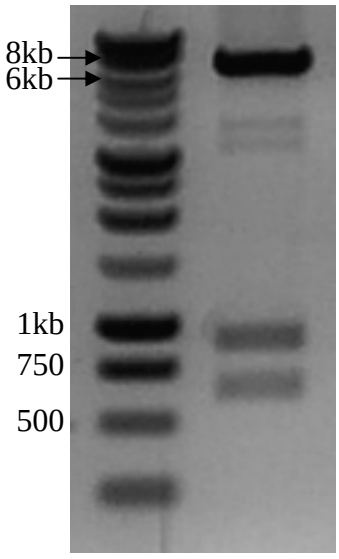


Figure 4.3. A reverse-contrast photograph of an ethidium bromide-stained 1% agarose-TAE gel showing the results of *Pst*I restriction digest of *pBSHvSacOα* (*Donor Plasmid*). First lane is 1kb marker. Expected fragments based on supplied sequence were 622, 894 and 7626, two of which arise from the *Donor Cassette*, and will be consistent in all recombinant plasmids that contain it. No mention of the consistency of the 894bp fragment was mentioned in Arensburger *et al*., (2005). The large fragment is estimated to be 6437bp based upon the revised *Donor Plasmid* sequence.

```
AgTangoAmp      MENNTKQIPLAVRKLIVRDVQNGESHRAVAGKYSISKSAVGKIFKKYSTLGSVVDRPGRG
AgTango         MENNTKQIPLAVRKLVVRDVQNGESHRAVAGKYSISKSAVGKIFKKYSTLGSVVDRPGRG
                ***************:*******************************************

AgTangoAmp      RKRITDSKTDTRIVREVKKNPKVTVLEIKNTLQLNISDRTIRRRIIEQGYNSKLAKKRPF
AgTango         RKRITDSKTDTRIVREVKKNPKVTVLEIKNTLQLNISDRTIRRRIIEQGYNSKLAKKRPF
                ************************************************************

AgTangoAmp      ISKANKSKRLKFAQEHADKPLEFWKTVLWTDESKFELFNRKRRSHVWCKPGEELQERNIQ
AgTango         ISKANKSKRLKFAREHADKPLEFWKTVLWTDESKFELFNRKRRSHVWCKPGEELQERNIQ
                *************:**********************************************

AgTangoAmp      GTVKHGGGNVMVWGCFSWGGAGSLVRINGIMTADTYITILQENLEVSLIKTGLENKFVFQ
AgTango         GTVKHGGGNVMVWGCFSWGGAGSLVRINGIMTADTYITILQENLEVSLIKTGLENKFVFQ
                ************************************************************

AgTangoAmp      QDNDPKHTAKKTKTFFNSCRIKPLEWPPQSPDLNPIENLWAILDDRVKKTGVTNKDKYFE
AgTango         QDNDPKHTAKKTKTFFNSCRIKPLEWPPQSPDLNPIENLWAILDDRVKKTGVTNKDKYFE
                ************************************************************

AgTangoAmp      ALENAWENLDPNHIENLVESMPKRLQLVLRSKGGHIKY
AgTango         ALENAWENLDPNHIENLVESMPKRLQLVLRSKGGHIKY
                ***************************************
```

Figure 4.4 Amino acid translations of the PCR sequence of *AgTango's* ORF (AgTangoAmp) *vs.* the sequence from the *Anopheles gambiae* genome (AgTango). In grey are the changes in predicted amino acid sequence.

A.

>Herves 5’ 2C1+HervesDonorR2 sequence exported from 1_WellA022C1+HervesDonorR2.ab1
GGACACAGGTGTTTGGGGGTGAAGCATAACAGGATGTTAACAGTCTCTGCAGATACAGGTATAGGAAGCACGAGAGCACCCATC
GTCTTTAGCCCCTTCATTCTCCTGCGGCTCTGGTCAGGGAATCTGTAGGCGCAGTCCGCTATTGAGGGAATAAGGAGACTCAGC
AATGTGGGTTCTCATCCCATTCTTCTAAAAGTTCCTCCTGCATTTTCAGGGTTATTGTCTCTTGAGCGGATACTTATTGATGTA
ATAGAAATCATATCAAATAGGGGGGCCGAAACATCACAAGAAAAGTGCCACAAAAGAATAGAGCGTTATTGGGTTGAGAAATTC
GCGTTAAGGATCAAGAGTATCAGCTCATTGAACGTGGAATAGGCCGTCATAGGCAGAATCCCTGTTAAATCAGAAAGAATAGAC
CGAGATAGGGCTGAGTCCCTAATCAGTTTGGAACAAGAGTCCACTATTAAAAGCCGTAAATCGGAACCTCAAAGGGAGAAAAAC
CGTCTATCAGGGCGAGGGAACGCTGCGGGACGCGTCGAGATAGTCAGGTAAGAAAGGGTAGAGAGCCGGTAAAACGGTAAATCC
AAGCGTTAAAGGGAGCCCCCGAGTTAGAGCCACACGGGGAAAGCCGAAGAACGTGGCGCGAAAGGAAGGGCAGAAAGCGAAAGG
AGCGGGCATAGGGGGGAAGCAAGATCGCGGTCACCTTGCGCGTAACCACCCAGCTGGCGAAAGGGGGATGCCGGCACAGGGCGC
GTCCCGGGTACCATCCAGGTTGTGCCAGTCTTGGGAAGGTAAATCGGCGCGGGCCTAGCGCGCGTAATACCAGCTGGCGAAAGG
GGAATTGGGGCAAGGGCCCCCCCTTGGGGTCAATCCGGGATTGACCCGACTTGGTAGCATAAAGAGTCGGCCCTCGAGCGCCAG
TAATGTGACTCTCTATCGGTTGATCTGGGTGCCGGGCCGCCCCTCGAGGTCAATTCACTAGTGATTCGATTTTGTAGCAATTAG
AGTTGTGCCTCAAGAACCAGAAATGTCCGGTTCAGGCAACTCTTGAAAAGGAGCGAGCGACAGTCGTTCATGTAGTTCATTCCC
CCCATTCAAATATGAACGTGAATGGTATGAGCAGGACGTTCACGAAAAGAACGTCCTACCCGTCGAAATGAGAACTCACGGGGT
T

B.

>Herves 3’ 2C1+HervesDonorF1 sequence exported from 10_WellF012C1+HervesDonorF1.ab1
AAGTTAGCGGATCGCTCGTTCTGTTAATGAAAGAACCGGTAAATTCACGTTCATGGTAATGAATCGAATTGCCCAACCCTAGTA
GCAACCGGACCGAATCACTAGTGAATTCGCGGCCCCAGCCGGATCTAATCGCTTTACAAGCTGCCAATCGAAGTTCAATTAGTG
AGTTTGAACTGTACTAGTCGTTTTGTTTGAGAGAGAAATGACCTTCTCGTGCTTTGAGGAAGAGAGGAAAATTACACTAATACG
TTTATGGTCACTAAATCCATATGATAAATACATAGATCACGTGTGTTCGGTTCGTTGTTGTATAGGTGATTTTGACCGTTATTT
TTGAATACGTTCGATTTCTTCAATGTGTTGTAGGCTTATGGTAATAAACATTGCTGAAACCGTGATTGTATTGATAAATTGATC
TTTGATATGCCATTTGTTGTGATATTTTGACTTGTAAACCACAAATTGATCTACGCTCCATCGAAATAAGTGAAAATTTAATAA
TTGCACGGCGATCAAAGGTAACATTAGTCTTGGTAGTTAAAAACAAAAACATAAAGTGTTGTATGAAACATTTCCACAGATATG
ATGGCTCCAACAAACGCAACAACAAGCCCTGTCTGGGATCATAATCGAATTGACGGTATCGATAAGCTTGATCTAGAGTCTGAG
GTCTGCCTCGTGAAGAAGGTGTTGCTGACTCATACCAGGCCTGAATCGCCCCATCATCCAGCCAGAAAGTGAGGGAGCCACGGT
TGATGAGAGCTTTGTTGTAGGTGGACCAGTTGGTGATTTTGAACTTTTGCTTTGCCACGGAACGGTCTGCGTTGTCGGGAAGAT
GCGTGATCTGATCCTTCAACTCAGCAAAAGTTCGATTTATTCAACAAAGCCGCCGTCCCGTCAAGTCAGCGTAATGCTCTGCCA
GTGTTACAACCAATTAACCAATTCTGATTAGAAAAACTCATCGAGCATCAAATGAAACTGCATTTATTCATATCAGATTATCAT
ACCATATTTTGAAAGCCGTTCTGTATGAAGGAGAAACTCACGAGGCAGTTCATAGATGCAGATCTGTATCGTCTGCGATCGACT
CGTCAACATCAATACACTATATTCCCCTCGTCAAATAGTTATCAGGTGAGAAACACCATGATGACG

Figure 4.5 Sequences from the 5’ (A) and 3’ ends (B) of a *Herves* recombinant obtained in the interplasmid transposition assay. These sequences are characteristic of all recombinants obtained passing the selection process. Terminal inverted repeats are underlined and in bold. Target site duplications are in grey, and are the same sequences as those found in the *Donor Plasmid*. All of the sequence obtained for both ends comes from the *Donor Plasmid* except for the region in yellow highlighting, which comes from the *Helper Plasmid*. For this particular transposable element, the target site duplication sequences should be different than those found in the *Donor Plasmid*, and the sequence outside of the terminal inverted repeats should be that of the *Target Plasmid*.

Table 4.1. Oligonucleotide primers to sequence the ambiguous region of *pBSHvSacKOα Donor Plasmid*.

Oligo Name	Sequence 5' to 3'	Worked?
First Round		
pBSHvSacKOα-F1	CTTGATCTAGATAAGCTTCTCG	N
pBSHvSacKOα-F2	CATTTCCACAGATATGATGG	Y
pBSHvSacKOα-F3	CACAGATATGATGGCTCCAAC	Y
pBSHvSacKOα-R1	ATCCAGCCAGAAAGTGAG	Y
pBSHvSacKOα-R2	TGCTGACTCATACCAGGC	Y
Second Round		
Kan-F1	GCTCATAACACCCCTTGTATTACT	Y
Kan-F2	TTCTGGCTGGATGATGG	N
Kan-R1	GCCTCTTCCGACCATC	Y
Kan-R2	TGATGCGAGTGATTTTGAT	Y
Third Set		
Kan-F3	CAGCAACACCTTCTTCACG	Y
Kan-F4	CAGACCTCAGACCTGCAG	Y
Kan-F5	ACAGGAAACAGCTATGACCAT	N

Table 4.2. Oligonucleotide primers to create *Tango* inserts for *Tango Helper* and *Donor* plasmids for inter-plasmid transposition assay, and to sequence recombinants to identify transposition reactions.

Oligo Name	Sequence 5' to 3'	Purpose
Helper		
TangoTemp-F2	GCATTAGCATTTGGCATTG	Create template for Tango ORF
TangoTemp-R2	AATTCACTCCTTATCGTTCG	Create template for Tango ORF
TangoORF-F1-*Xma*I	cccgggAAAATGGAAAACAACACAAAAC†	Amplify Tango ORF with *Xma*I RE site
TangoORF-R1-Stop-*Bam*HI	ggatccTTAATATTTTATGTGCCCCCC†	Amplify Tango ORF with stop codon and *Bam*HI RE site
pKhsp70-F1	CTAAGCGAAAGCTAAGCAAAT	Verify Tango ORF insert in Helper
pKhsp70-R1	AACTTAAGCCAGGAACTGAA	Verify Tango ORF insert in Helper
Donor		
TangoProTemp-F1	AGCGAACAACAGCACAAC	Create template for Tango 5' TIR
TangoProTemp-R1	ACACCCCCAAACCATAAC	Create template for Tango 5' TIR
Tango5'TIR-F1-*Kpn*I	ggtaccGTTACAGTGGCCGGCAA†	Amplify Tango 5' TIR to go into Donor with *Kpn*I RE site
Tango5'TIR-R1-*Xcm*I	ccatcatatctgtggCATTTTGTGCAACTGAGCC†	Amplify Tango 5' TIR to go into Donor with *Xcm*I RE site
Tango3'Temp-F1	CAAAGGGGGGCACATAAA	Create template for Tango 3' TIR
Tango3'Temp-R1	ACTACAAACGAGCACCAACG	Create template for Tango 3'TIR
Tango3'TIR-F1-*Bam*HI	ggatccCAAAGGGGGGCACATAAA†	Amplify Tango 3' TIR to go into Donor with *Bam*HI RE site
Tango3'TIR-R1-*Rsr*II	cggtccgTGTATACAGTGGCCGGCA†	Amplify Tango 3' TIR to go into Donor with *Rsr*II RE site
TangoSeqDonor-F1	CCAAAGGGGGGCACATAA	Verify Tango 3' TIR inserts in Donor
TangoSeqDonor-F2	AGGAAGGGAAGAAAGCGAAA	Verify Tango 3' TIR inserts in Donor
TangoSeqDonor-F3	TTGACATGAGGTGTTTAGGCA	Verify Tango 3' TIR inserts in Donor
TangoSeqDonor-R1	AGCGAGGAAGCGGAAGA	Verify Tango 5' TIR inserts in Donor
TangoSeqDonor-R2	TGCCTAAACACCTCATGTCAA	Verify Tango 5' TIR inserts in Donor
Tango Transposition Reaction		
TangoDonor-F1	CCAAAGGGGGGCACATAA	Verify 3' end of Tango Donor Cassette in Target
TangoDonor-F2	GCCAAACGAATGTCTAGATCGA	Verify 3' end of Tango Donor Cassette in Target
TangoDonor-F3	GATATCGAGCTCGCTTGGAC	Verify 3' end of Tango Donor Cassette in Target
TangoDonor-R1	GATCGATTTGCATAGTGGGC	Verify 5' end of Tango Donor Cassette in Target
TangoDonor-R2	GTTGGAGCCATCATATCTGTGG	Verify 5' end of Tango Donor Cassette in Target

†Lower case font indicates the restriction enzyme recognition site added to the primer to facilitate molecular cloning.

Table 4.3. Oligonucleotide primers used to verify *Herves* transposition events in products of inter-plasmid assay.

Oligo Name	Sequence 5' to 3'
HervesDonor-F1	TGTGGTTAGTTGACGTGAACG
HervesDonor-F2	GTGTTAAAATGTGTTGGTTGCC
HervesDonor-R1	GTAGGTCGTTCGTTCTTTAGGATG
HervesDonor-R2	AATGGGGGGAATGAACGAC
HervesDonor-R3	GGCTGAGTGAGAAAGTATGCTGAT

Table 4.4 Results for inter-plasmid transposition assay for *Herves* and *AgTango*.

TE Construct	Number of Experiments	Number of Plasmids Screened	Number of Recombinants (Kan/Cam Resistant)	Number Passing Amp Cross-Selection	Number Passing *PstI* Digest	Number of Plasmids Sequenced	Number of Transposition Events
Herves	3	987,000	17	3	3	3	0
AgTango	2	553,000	5	0	1	1	0

References

Adelman, Z.N., Jasinskiene, N., Vally, K.J., Peek, C., Travanty, E.A., Olson, K.E., Brown, S.E., Stephens, J.L., Knudson, D.L., Coates, C.J. and James, A.A. (2004) Formation and loss of large, unstable tandem arrays of the *piggyBac* transposable element in the yellow fever mosquito, *Aedes aegypti*. *Transgenic Res* **13:** 411-425.

de Almeida, L.M. and Carareto, C.M. (2005) Multiple events of horizontal transfer of the *Minos* transposable element between *Drosophila* species. *Mol Phylogenet Evol* **35:** 583-594.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Arca, B. and Savakis, C. (2000) Distribution of the transposable element *Minos* in the genus *Drosophila*. *Genetica* **108:** 263-267.

Arensburger, P., Kim, Y.J., Orsetti, J., Aluvihare, C., O'Brochta, D.A. and Atkinson, P.W. (2005) An active transposable element, *Herves*, from the African malaria mosquito *Anopheles gambiae*. *Genetics* **169:** 697-708.

Ashburner, M., Hoy, M.A. and Peloquin, J.J. (1998) Prospects for the genetic transformation of arthropods. *Insect Mol Biol* **7:** 201-213.

Atkinson, P.W., Pinkerton, A.C. and O'Brochta, D.A. (2001) Genetic transformation systems in insects. *Annu Rev Entomol* **46:** 317-346.

Auge-Gouillou, C., Hamelin, M.H., Demattei, M.V., Periquet, G. and Bigot, Y. (2001) The ITR binding domain of the *Mariner Mos-1* transposase. *Mol Genet Genomics* **265:** 58-65.

Ayala, F.J. and Coluzzi, M. (2005) Chromosome speciation: humans, *Drosophila*, and mosquitoes. *Proc Natl Acad Sci U S A* **102 Suppl 1:** 6535-6542.

Besansky, N.J., Krzywinski, J., Lehmann, T., Simard, F., Kern, M., Mukabayire, O., Fontenille, D., Toure, Y. and Sagnon, N. (2003) Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: evidence from multilocus DNA sequence variation. *Proc Natl Acad Sci U S A* **100:** 10818-10823.

Besansky, N.J., Powell, J.R., Caccone, A., Hamm, D.M., Scott, J.A. and Collins, F.H. (1994) Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors. *Proc Natl Acad Sci U S A* **91:** 6885-6888.

Biedler, J., Qi, Y., Holligan, D., Della Torre, A., Wessler, S. and Tu, Z. (2003) Transposable element (TE) display and rapid detection of TE insertion polymorphism in the *Anopheles gambiae* species complex. *Insect Mol Biol* **12:** 211-216.

Biedler, J. and Tu, Z. (2003) Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. *Mol Biol Evol* **20:** 1811-1825.

Biessmann, H., Valgeirsdottir, K., Lofsky, A., Chin, C., Ginther, B., Levis, R.W. and Pardue, M.L. (1992) *HeT-A*, a transposable element specifically involved in

"healing" broken chromosome ends in *Drosophila melanogaster*. *Mol Cell Biol* **12:** 3910-3918.

Bird, A.P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet* **11:** 94-100.

Bron, S. (1990) Plasmids. In: *Molecular biological methods for Bacillus* (Harwood, C.R. and Cutting, S.M., eds), pp. 75-174. John Wiley & Sons Ltd., West Sussex.

Capy, P., Vitalis, R., Langin, T., Higuet, D. and Bazin, C. (1996) Relationships between transposable elements based upon the integrase-transposase domains: Is there a common ancestor? *J Mol Evol* **42:** 359-368.

Carareto, C.M., Kim, W., Wojciechowski, M.F., O'Grady, P., Prokchorova, A.V., Silva, J.C. and Kidwell, M.G. (1997) Testing transposable elements as genetic drive mechanisms using *Drosophila P* element constructs as a model system. *Genetica* **101:** 13-33.

Caizzi, R., Caggese, C. and Pimpinelli, S. (1993) *Bari-1*, a new transposon-like family in *Drosophila melanogaster* with a unique heterochromatic organization. *Genetics* **133:** 335-345.

Cary, L.C., Goebel, M., Corsaro, B.G., Wang, H.G., Rosen, E. and Fraser, M.J. (1989) Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon *IFP2* insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* **172:** 156-169.

Casa, A.M., Brouwer, C., Nagel, A., Wang, L.J., Zhang, Q., Kresovich, S. and Wessler, S.R. (2000) The MITE family *Heartbreaker* (*Hbr*): Molecular markers in maize. *Proc Natl Acad Sci U S A* **97:** 10083-10089.

Catteruccia, F., Nolan, T., Blass, C., Muller, H.M., Crisanti, A., Kafatos, F.C. and Loukeris, T.G. (2000) Toward *Anopheles* transformation: *Minos* element activity in Anopheline cells, and embryos. *Proc Natl Acad Sci U S A* **97:** 6236.

Clements, A.N. (1992) *The Biology of Mosquitoes,* Chapman & Hall, London.

Coetzee, M., Craig, M. and le Sueur, D. (2000) Distribution of African malaria mosquitoes belonging to the *Anopheles gambiae* complex. *Parasitol Today* **16:** 74-77.

Collins, J., Forbes, E. and Anderson, P. (1989) The *Tc3* family of transposable genetic elements in *Caenorhabditis elegans*. *Genetics* **121:** 47-55.

Colloms, S.D., van Luenen, H.G. and Plasterk, R.H. (1994) DNA binding activities of the *Caenorhabditis elegans Tc3* transposase. *Nucleic Acids Res* **22:** 5548-5554.

Colot, V., Haedens, V. and Rossignol, J.L. (1998) Extensive, nonrandom diversity of excision footprints generated by *Ds*-like transposon *Ascot-1* suggests new parallels with V(D)J recombination. *Mol Cell Biol* **18:** 4337-4346.

Coluzzi, M., Sabatini, A., Petrarca, V. and Di Deco, M.A. (1979) Chromosomal differentiation and adaptation to human environments in the *Anopheles gambiae* complex. *Trans R Soc Trop Med Hyg* **73:** 483-497.

Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M.A. and Petrarca, V. (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* **298:** 1415-1418.

Coy, M.R. and Tu, Z. (2005) *Gambol* and *Tc1* are two distinct families of DD34E transposons: analysis of the *Anopheles gambiae* genome expands the diversity of the *IS630-Tc1-mariner* superfamily. *Insect Mol Biol* **14:** 537-546.

Coy, M.R. and Tu, Z. (2007) Genomic and evolutionary analyses of *Tango* transposons in *Aedes aegypti*, *Anopheles gambiae* and other mosquito species. *Insect Mol Biol*. *In press.*

Craig, N.L. (2002) *Mobile DNA II,* ASM Press, Washington, D.C.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190.

Cui, Z., Geurts, A.M., Liu, G., Kaufman, C.D. and Hackett, P.B. (2002) Structure-function analysis of the inverted terminal repeats of the *Sleeping Beauty* transposon. *J Mol Biol* **318:** 1221-1235.

Daboussi, M.J., Langin, T. and Brygoo, Y. (1992) *Fot1*, a new family of fungal transposable elements. *Mol Gen Genet* **232:** 12-16.

Doak, T.G., Doerder, F.P., Jahn, C.L. and Herrick, G. (1994) A proposed superfamily of transposase genes: Transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proc Natl Acad Sci U S A* **91:** 942-946.

Donnelly, M.J., Pinto, J., Girod, R., Besansky, N.J. and Lehmann, T. (2004) Revisiting the role of introgression *vs* shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the *Anopheles gambiae* complex. *Heredity* **92:** 61-68.

Doolittle, W.F. and Sapienza, C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601-603.

Emmons, S.W., Yesner, L., Ruan, K.S. and Katzenberg, D. (1983) Evidence for a transposon in *Caenorhabditis elegans*. *Cell* **32:** 55-65.

Finnegan, D.J. (1992) Transposable elements. *Curr Opin Genet Dev* **2:** 861-867.

Fischer, S.E., van Luenen, H.G. and Plasterk, R.H. (1999) *Cis* requirements for transposition of *Tc1*-like transposons in C. elegans. *Mol Gen Genet* **262:** 268-274.

Fraser, M.J., Smith, G.E. and Summers, M.D. (1983) 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* **47:** 287-300.

Galtier, N., Gouy, M. and Gautier, C. (1996) SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* **12:** 543-548.

Gaunt, M.W. and Miles, M.A. (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol Biol Evol* **19:** 748-761.

Gomulski, L.M., Torti, C., Bonizzoni, M., Moralli, D., Raimondi, E., Capy, P., Gasperi, G. and Malacrida, A.R. (2001) A new basal subfamily of *mariner* elements in *Ceratitis rosa* and other tephritid flies. *J Mol Evol* **53:** 597-606.

Grossman, G.L., Cornel, A.J., Rafferty, C.S., Robertson, H.M. and Collins, F.H. (1999) *Tsessebe*, *Topi* and *Tiang*: three distinct *Tc1*-like transposable elements in the malaria vector, *Anopheles gambiae*. *Genetica* **105:** 69-80.

Hall, B.G. (2004) *CodonAlign*, version 2.0. Bellingham Research Institute, Bellingham, WA.

Handler, A.M. and O'Brochta, D.A. (2005) Transposable elements for insect transformation. In: *Comprehensive Molecular Insect Science* (Eds, Gilbert, L.I., Iatrou, K. and Gill, S.S.) Elsevier Ltd., Oxford pp. 395-436.

Hill, S.R., Leung, S.S., Quercia, N.L., Vasiliauskas, D., Yu, J., Pasic, I., Leung, D., Tran, A. and Romans, P. (2001) *Ikirara* insertions reveal five new *Anopheles gambiae* transposable elements in islands of repetitious sequence. *J Mol Evol* **52:** 215-231.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M., Majoros, W.H., Rusch, D.B., Lai, Z., Kraft, C.L., Abril, J.F., Anthouard, V., Arensburger, P., Atkinson, P.W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chatuverdi, K., Christophides, G.K., Chrystal, M.A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C.A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M.E., Hladun, S.L., Hogan, J.R., Hong, Y.S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.J., Lobo, N.F., Lopez, J.R., Malek, J.A., McIntosh, T.C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S.D., O'Brochta, D.A., Pfannkoch, C., Qi, R., Regier, M.A., Remington, K., Shao, H., Sharakhova, M.V., Sitter, C.D., Shetty, J., Smith, T.J., Strong, R., Sun, J., Thomasova, D., Ton, L.Q., Topalis, P., Tu, Z., Unger, M.F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K.J., Wortman, J.R., Wu, M., Yao, A., Zdobnov, E.M., Zhang, H., Zhao, Q., et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298:** 129-149.

Ivics, Z., Hackett, P.B., Plasterk, R.H. and Izsvak, Z. (1997) Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon from fish, and its transposition in human cells. *Cell* **91:** 501-510.

Ivics, Z., Izsvak, Z., Minter, A. and Hackett, P.B. (1996) Identification of functional domains and evolution of *Tc1*-like transposable elements. *Proc Natl Acad Sci U S A* **93:** 5008-5013.

Izsvak, Z., Ivics, Z. and Hackett, P.B. (1995) Characterization of a *Tc1*-like transposable element in zebrafish (*Danio rerio*). *Mol Gen Genet* **247:** 312-322.

Izsvak, Z., Ivics, Z. and Plasterk, R.H. (2000) *Sleeping Beauty*, a wide host-range transposon vector for genetic transformation in vertebrates. *J Mol Biol* **302:** 93-102.

Izsvak, Z., Khare, D., Behlke, J., Heinemann, U., Plasterk, R.H. and Ivics, Z. (2002) Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in *Sleeping Beauty* transposition. *J Biol Chem* **277:** 34581-34588.

Jarvik, T. and Lark, K.G. (1998) Characterization of *Soymar1*, a *mariner* element in soybean. *Genetics* **149:** 1569-1574.

Jasinskiene, N., Coates, C.J. and James, A.A. (2000) Structure of *Hermes* integrations in the germ line of the yellow fever mosquito, *Aedes aegypti*. *Insect Mol Biol* **9:** 11-18.

Jehle, J.A., Nickel, A., Vlak, J.M. and Backhaus, H. (1998) Horizontal escape of the novel *Tc1*-like lepidopteran transposon *TCp3.2* into *Cydia pomonella* granulovirus. *J Mol Evol* **46:** 215-224.

Karch, F., Torok, I. and Tissieres, A. (1981) Extensive regions of homology in front of the two hsp70 heat shock variant genes in *Drosophila melanogaster*. *J Mol Biol* **148:** 219-230.

Karsi, A., Moav, B., Hackett, P. and Liu, Z. (2001) Effects of insert size on transposition efficiency of the *Sleeping Beauty* transposon in mouse cells. *Mar Biotechnol (NY)* **3:** 241-245.

Ke, Z., Grossman, G.L., Cornel, A.J. and Collins, F.H. (1996) *Quetzal*: a transposon of the *Tc1* family in the mosquito *Anopheles albimanus*. *Genetica* **98:** 141-147.

Kidwell, M.G. and Lisch, D.R. (2001) Perspective: Transposable elements, parasitic DNA, and genome evolution. *Evolution* **55:** 1-24.

Klinakis, A.G., Loukeris, T.G., Pavlopoulos, A. and Savakis, C. (2000) Mobility assays confirm the broad host-range activity of the *Minos* transposable element and validate new transformation tools. *Insect Mol Biol* **9:** 269-275.

Knipple, D.C. and Marsella-Herrick, P. (1988) Versatile plasmid vectors for the construction, analysis and heat-inducible expression of hybrid genes in eukaryotic cells. *Nucleic Acids Res* **16:** 7748.

Korber, B. (2000) Computational analysis of HIV molecular sequences. In: *Signature and sequence variation analysis* (AG, Rodrigo. and GH, Learn., eds), pp. 55-72. Kluwer Academic Publishers, Dordrecht.

Krzywinski, J., Grushko, O.G. and Besansky, N.J. (2006) Analysis of the complete mitochondrial DNA from *Anopheles funestus*: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. *Mol Phylogenet Evol* **39:** 417-423.

Lampe, D.J., Grant, T.E. and Robertson, H.M. (1998) Factors affecting transposition of the *Himar1 mariner* transposon in vitro. *Genetics* **149:** 179-187.

Lampe, D.J., Walden, K.K. and Robertson, H.M. (2001) Loss of transposase-DNA interaction may underlie the divergence of *mariner* family transposable elements and the ability of more than one *mariner* to occupy the same genome. *Mol Biol Evol* **18:** 954-961.

Lampe, D.J., Witherspoon, D.J., Soto-Adames, F.N. and Robertson, H.M. (2003) Recent horizontal transfer of mellifera subfamily *mariner* transposons into insect lineages representing four different orders shows that selection acts only during horizontal transfer. *Mol Biol Evol* **20:** 554-562.

Langin, T., Capy, P. and Daboussi, M.J. (1995) The transposable element *impala*, a fungal member of the *Tc1-mariner* superfamily. *Mol Gen Genet* **246:** 19-28.

Levis, R.W., Ganesan, R., Houtchens, K., Tolar, L.A. and Sheen, F.M. (1993) Transposons in place of telomeric repeats at a *Drosophila* telomere. *Cell* **75:** 1083-1093.

Li, X., Harrell, R.A., Handler, A.M., Beam, T., Hennessy, K. and Fraser, M.J., Jr. (2005) *piggyBac* internal sequences are necessary for efficient transformation of target genomes. *Insect Mol Biol* **14:** 17-30.

Lindquist, S. (1980) Varying patterns of protein synthesis in *Drosophila* during heat shock: implications for regulation. *Dev Biol* **77:** 463-479.

Liao, L.W., Rosenzweig, B. and Hirsh, D. (1983) Analysis of a transposable element in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* **80:** 3585-3589.

Lohe, A.R., De Aguiar, D. and Hartl, D.L. (1997) Mutations in the *mariner* transposase: the D,D(35)E consensus sequence is nonfunctional. *Proc Natl Acad Sci U S A* **94:** 1293-1297.

van Luenen, H.G., Colloms, S.D. and Plasterk, R.H. (1994) The mechanism of transposition of *Tc3* in *C. elegans*. *Cell* **79:** 293-301.

Maragathavally, K.J., Kaminski, J.M. and Coates, C.J. (2006) Chimeric *Mos1* and *piggyBac* transposases result in site-directed integration. *FASEB J* **20:** 1880-1882.

Marrelli, M.T., Moreira, C.K., Kelly, D., Alphey, L. and Jacobs-Lorena, M. (2006) Mosquito transgenesis: what is the fitness cost? *Trends Parasitol* **22:** 197-202.

Mathiopoulos, K.D., della Torre, A., Predazzi, V., Petrarca, V. and Coluzzi, M. (1998) Cloning of inversion breakpoints in the *Anopheles gambiae* complex traces a transposable element at the inversion junction. *Proc Natl Acad Sci U S A* **95:** 12444-12449.

Mathiopoulos, K.D., della Torre, A., Santolamazza, F., Predazzi, V., Petrarca, V. and Coluzzi, M. (1999) Are chromosomal inversions induced by transposable elements? A paradigm from the malaria mosquito *Anopheles gambiae*. *Parassitologia* **41:** 119-123.

McClintock, B. (1948) Mutable loci in maize. *Carnegie Inst Washington Year Book* **47:** 155-169.

McDonald, J.F. (1998) Transposable elements, gene silencing and macroevolution. *Trends Ecol Evol* **13:** 94-95.

Miller, W.J. and Capy, P. (2004) Mobile genetic elements as natural tools for genome evolution. In: *Mobile Genetic Elements : Protocols and Genomic Applications* (Miller, W.J. and Capy, P., eds), pp. 1-20. Humana Press, Totowa.

Moreira, L.A., Edwards, M.J., Adhami, F., Jasinskiene, N., James, A.A. and Jacobs-Lorena, M. (2000) Robust gut-specific gene expression in transgenic Aedes aegypti mosquitoes. *Proc Natl Acad Sci U S A* **97:** 10895-10898.

Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., Ren, Q., Zdobnov, E.M., Lobo, N.F., Campbell, K.S., Brown, S.E., Bonaldo, M.F., Zhu, J., Sinkins, S.P., Hogenkamp, D.G., Amedo, P., Arsenburger, P., Atkinson, P.W., Bidwell, S., Biedler, J., Birney, E., Bruggner, R.V., Costas, J., Coy, M.R., Crabtree, J., Crawford, M., Debruyn, B., Decaprio, D., Eiglmeier, K., Eisenstadt, E., El-Dorry, H., Gelbart, W.M., Gomes, S.L., Hammond, M., Hannick, L.I., Hogan, J.R., Holmes, M.H., Jaffe, D., Johnston, S.J., Kennedy, R.C., Koo, H., Kravitz, S., Kriventseva, E.V., Kulp, D., Labutti, K., Lee, E., Li, S., Lovin, D.D., Mao, C., Mauceli, E., Menck, C.F., Miller, J.R., Montgomery, P., Mori, A., Nascimento, A.L., Naveira, H.F., Nusbaum, C., O'Leary S, B., Orvis, J., Pertea, M., Quesneville, H., Reidenbach, K.R., Rogers, Y.H., Roth, C.W., Schneider, J.R., Schatz, M., Shumway, M., Stanke, M., Stinson, E.O., Tubio, J.M., Vanzee, J.P., Verjovski-Almeida, S., Werner, D., White, O., Wyder, S., Zeng, Q., Zhao, Q., Zhao, Y., Hill, C.A., Raikhel, A.S., Soares, M.B., Knudson, D.L., Lee, N.H., Galagan, J., Salzberg, S.L., Paulsen, I.T., Dimopoulos, G., Collins, F.H., Bruce, B., Fraser-Liggett, C.M. and Severson,

D.W. (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* [Epub ahead of print].

O'Brochta, D.A. and Atkinson, P.W. (1996) Transposable elements and gene transformation in non-drosophilid insects. *Insect Biochem Mol Biol* **26:** 739-753.

O'Brochta, D.A., Sethuraman, N., Wilson, R., Hice, R.H., Pinkerton, A.C., Levesque, C.S., Bideshi, D.K., Jasinskiene, N., Coates, C.J., James, A.A., Lehane, M.J. and Atkinson, P.W. (2003) Gene vector and transposable element behavior in mosquitoes. *J Exp Biol* **206:** 3823-3834.

Orgel, L.E. and Crick, F.H. (1980) Selfish DNA: the ultimate parasite. *Nature* **284:** 604-607.

Page, R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12:** 357-358.

Pardue, M.L. and DeBaryshe, P.G. (1999) *Drosophila* telomeres: two transposable elements with important roles in chromosomes. *Genetica* **107:** 189-196.

Petrarca, V., Beier, J.C., Onyango, F., Koros, J., Asiago, C., Koech, D.K. and Roberts, C.R. (1991) Species composition of the *Anopheles gambiae* complex (diptera: Culicidae) at two sites in western Kenya. *J Med Entomol* **28:** 307-313.

Pietrokovski, S. and Henikoff, S. (1997) A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. *Mol Gen Genet* **254:** 689-695.

Plasterk, R.H., Izsvak, Z. and Ivics, Z. (1999) Resident aliens: the *Tc1*/*mariner* superfamily of transposable elements. *Trends Genet* **15:** 326-332.

Plasterk, R.H.A. and van Luenen, H.G.A.M. (2002) The *Tc1*/*mariner* family of transposable elements. In: *Mobile DNA II* (Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A., eds), pp. 519-532. ASM Press, Washington, D.C.

Pledger, D.W., Fu, Y.Q. and Coates, C.J. (2004) Analyses of *cis*-acting elements that affect the transposition of *Mos1 mariner* transposons *in vivo*. *Mol Genet Genomics* **272:** 67-75.

Posada, D. and Crandall, K.A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**: 817-818.

Powell, J.R., Petrarca, V., della Torre, A., Caccone, A. and Coluzzi, M. (1999) Population structure, speciation, and introgression in the *Anopheles gambiae* complex. *Parassitologia* **41:** 101-113.

Rai, K.S. and Black IV, W.C. (1999) Mosquito genomes: structure, organization, and evolution. In: *Advances in genetics*, pp. 1-33. Academic Press, San Diego.

Robertson, H.M. (1993) The *mariner* transposable element is widespread in insects. *Nature* **362:** 241-245.

Robertson, H.M. (2002) Evolution of DNA Transposons in Eukaryotes. In *Mobile DNA II* (Eds, Craig, N. L., Craigie, R., Gellert, M. and Lambowitz, A.) ASM Press, Washington, D.C, pp. 1093-1110.

Robertson, H.M. and Asplund, M.L. (1996) *Bmmar1*: a basal lineage of the *mariner* family of transposable elements in the silkworm moth, *Bombyx mori*. *Insect Biochem Mol Biol* **26:** 945-954.

Robertson, H.M. and Lampe, D.J. (1995a) Distribution of transposable elements in arthropods. *Annu Rev Entomol* **40:** 333-357.

Robertson, H.M. and Lampe, D.J. (1995b) Recent horizontal transfer of a *mariner* transposable element among and between Diptera and Neuroptera. *Mol Biol Evol* **12:** 850-862.

Robertson, H.M., Soto-Adames, F.N., Walden, K.K., Avancini, R.M. and Lampe, D.J. (2002) The *mariner* transposons of animals: horizontally jumping genes. In: *Horizontal Gene Transfer* (Syvanen, M. and Kado, C.I., eds), pp. 173-185. Academic Press, San Diego.

Rost, B., Fariselli, P. and Casadio, R. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* **5:** 1704-1718.

Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232:** 584-599.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sarkar, A., Atapattu, A., Belikoff, E.J., Heinrich, J.C., Li, X., Horn, C., Wimmer, E.A. and Scott, M.J. (2006) Insulated piggyBac vectors for insect transgenesis. *BMC Biotechnol* **6:** 27.

Sarkar, A., Coates, C.J., Whyard, S., Willhoeft, U., Atkinson, P.W. and O'Brochta, D.A. (1997) The *Hermes* element from *Musca domestica* can transpose in four families of cyclorrhaphan flies. *Genetica* **99:** 15-29.

Schneider, I. (1972) Cell lines derived from late embryonic stages of *Drosophila melanogaster*. *J Embryol Exp Morphol* **27:** 353-365.

Schneider, T.D. and Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18:** 6097-6100.

Severson, D.W., DeBruyn, B., Lovin, D.D., Brown, S.E., Knudson, D.L. and Morlais, I. (2004) Comparative genome analysis of the yellow fever mosquito *Aedes aegypti* with *Drosophila melanogaster* and the malaria vector mosquito *Anopheles gambiae*. *Journal Hered* **95:** 103-113.

Shao, H. and Tu, Z. (2001) Expanding the diversity of the IS*630-Tc1-mariner* superfamily: Discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics* **159:** 1103-1115.

Silva, J.C. and Kidwell, M.G. (2000) Horizontal transfer and selection in the evolution of *P* elements. *Mol Biol Evol* **17:** 1542-1557.

Silva, J.C., Loreto, E.L. and Clark, J.B. (2004) Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol* **6:** 57-71.

Springer, P.S. (2000) Gene traps: tools for plant development and genomics. *Plant Cell* **12:** 1007-1020.

Sundararajan, P., Atkinson, P.W. and O'Brochta, D.A. (1999) Transposable element interactions in insects: Crossmobilization of *hobo* and *Hermes*. *Insect Mol Biol* **8:** 359-368.

Swofford, D.L. (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, version 4.0. Sinauer Associates, Sunderland, MA.

Tatusova, T.A. and Madden, T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174:** 247-250.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-4882.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

della Torre, A., Fanello, C., Akogbeto, M., Dossou-yovo, J., Favia, G., Petrarca, V. and Coluzzi, M. (2001) Molecular evidence of incipient speciation within *Anopheles gambiae s.s.* in West Africa. *Insect Mol Biol* **10:** 9-18.

Toure, Y.T., Petrarca, V., Traore, S.F., Coulibaly, A., Maiga, H.M., Sankare, O., Sow, M., Di Deco, M.A. and Coluzzi, M. (1998) The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia* **40:** 477-511.

Tu, Z. and Li, S. (2007) Mobile genetic elements of malaria vectors and other mosquitoes. In: *Mobile Genetic Elements in Metazoan Parasites* (Eds Brindley, P.J.) Eurekah/Landes Bioscience, Georgetown, TX. *In press.*

Tu, Z. and Shao, H. (2002) Intra- and inter-specific diversity of *Tc3*-like transposons in nematodes and insects and implications for their evolution and transposition. *Gene* **282:** 133-142.

Vos, J.C., van Luenen, H.G. and Plasterk, R.H. (1993) Characterization of the *Caenorhabditis elegans Tc1* transposase *in vivo* and *in vitro*. *Genes Dev* **7:** 1244-1253.

Vos, J.C. and Plasterk, R.H. (1994) *Tc1* transposase of *Caenorhabditis elegans* is an endonuclease with a bipartite DNA binding domain. *EMBO J* **13:** 6125-6132.

Watkins, S., van Pouderoyen, G. and Sixma, T.K. (2004) Structural analysis of the bipartite DNA-binding domain of *Tc3* transposase bound to transposon DNA. *Nucleic Acids Res* **32:** 4306-4312.

Wessler, S.R. (1996) Turned on by stress. Plant retrotransposons. *Curr Biol* **6:** 959-961.

Yant, S.R., Park, J., Huang, Y., Mikkelsen, J.G. and Kay, M.A. (2004) Mutational analysis of the N-Terminal DNA-binding domain of sleeping beauty transposase: critical residues for DNA binding and hyperactivity in mammalian cells. *Mol Cell Biol* **24:** 9239-9247.

Yoshiyama, M., Tu, Z., Kainoh, Y., Honda, H., Shono, T. and Kimura, K. (2001) Possible horizontal transfer of a transposable element from host to parasitoid. *Mol Biol Evol* **18:** 1952-1958.

Zimowska, G.J. and Handler, A.M. (2006) Highly conserved piggyBac elements in noctuid species of Lepidoptera. *Insect Biochem Mol Biol* **36:** 421-428.

APPENDIX A. Identities at the amino acid levels of different gene products and *Tango* between *Anopheles gambiae* and *Aedes aegypti*

ENSEMBL Protein Family	% AA Identity	ENSEMBL ID
Nanos	28	ENSANGP00000020428
Immune Response Serine Protease Related ISPR5 (Fragment)	37	ENSANGP00000021259
Putative Apyrase Precursor	50	ENSANGP00000015382
Toll Receptor Precursor	50	ENSANGP00000029576
Transferrin	50	ENSANGP00000021949
Caspase 3 Precursor	<53	ENSANGP00000021365
Lectin	<53	ENSANGP00000010670
30E5 11	<56	ENSANGP00000022917
Vitellogenin 1	57	ENSANGP00000012892
Gustatory Receptor 66A	<58	ENSANGP00000021675
Cytochrome P450	60	ENSANGG00000020445
Chymotrypsin 2 Precursor	<61	ENSANGP00000026162
Frizzled Precursor	<61	ENSANGP00000019989
PAP	63	ENSANGP00000002978
Serine Protease Inhibitor	<64	ENSANGP00000023182
Zinc Carboxypeptidase A Precursor	64	ENSANGT00000021052
Transcription Factor	65	ENSANGP00000018140
Scavenger Receptor Class B Member	<67	ENSANGP00000012652
Antennal Carrier AP 2	69	ENSANGP00000027277
25 KDA GTP-Binding Protein (Fragment)	70	ENSANGP00000023777
Beta 1 3 Glucan Binding Precursor BGBP	70	ENSANGP00000016221
Odorant Receptor (GPROR8)	70	ENSANGP00000007927
Tango	**80**	N/A
White Protein	86	ENSANGP00000007325
Glucose 6 Phosphate 1 Dehydrogenase	89	ENSANGP00000018551
Opsin	90	ENSANGP00000011949
Receptor of Activated Protein Kinase C1	96	ENSANGP00000012560

Identities were obtained by randomly surveying 26 known *An. gambiae* peptides using a tBLASTn search against the *Aegypti aegypti* genome. Values proceeded by a '<' indicates that the hit did not span the entire query peptide, and that the identity is therefore presumed to be less than that reported by the BLAST program. In each case, the highest hit was used for comparison.

MONIQUE R COY

Department of Biochemistry
Blacksburg VA 24061
(540) 231-1394
lcoy@vt.edu

755 E Main Street
Christiansburg VA 24073
(540) 381-3495

EDUCATIONAL BACKGROUND

PhD **Biochemistry** June 2007
Virginia Polytechnic Institute and State University, Blacksburg Virginia

Dissertation Title: Analysis of the DD34E transposable elements of the African malaria mosquito *Anopheles gambiae* and the identification and characterization of an active transposase

BS **Zoology** December 1997 Highest Honors
University of Florida, Gainesville Florida
Thesis Title: Phylogenetic relationships of North American woodpeckers

HONORS AND AWARDS

Sigma Xi
Gamma Sigma Delta
Keystone Symposia Travel Award 2003
Eheart Travel Award 2003, 2005
Phi Beta Kappa
Phi Kappa Phi
Golden Key Nat'l Honor Society
Kendall King Award, 2006
Air Force Office of Scientific Research Travel Award 2006
Outstanding Doctoral Student of the Year – Agricultural and Life Sciences 2007

WORK EXPERIENCE

Laboratory Technician 2000-2001
Virginia Commonwealth University, Richmond Virginia
Determined the ability, time, and dose dependence of EtOH, H_2O_2, and deoxycholate to induce apoptosis in AGS and Caco-2 cells as measured by the terminal deoxynucleotidyl transferase dUTP nick-end labeling assay; maintained cell cultures; ordered reagents and supplies

Laboratory Technician 1998-2000
University of Florida, Gainesville Florida
Studied the effect of dietary zinc and immunological challenge on the protein expression levels of zinc transporters, ZnT-1 and ZnT-2, in rats, mice and BHK cell line by western blotting and immunofluorochemistry; designed, developed and purified custom antibodies

TEACHING EXPERIENCE

Graduate Teaching Assistant Spring 2002
Concepts in Biochemistry

UNIVERSITY SERVICE

Graduate Honor System Panel Member May 2004 – June 2006
Graduate Student Mentor Fall 2005

MEETING ABSTRACTS

Fifth International Symposium on Molecular Insect Science 2006
Tango: A DNA transposable element that dances among mosquito species

American Society of Tropical Medicine and Hygiene 2005
Tango: A DNA transposable element that dances among mosquito species

Keystone Symposium 2003
Transposition and Other Genome Rearrangements
Five novel families of *hAT*-like miniature inverted repeat transposable elements

in the *Fugu rubripes* genome

PUBLICATIONS

Cui, L., *et al.* (2000). Prouroguanylin overproduction and localization in the intestine of zinc-deficient rats. *J Nut* 130 (11) 2726-2732.

Coy, M.R. and Z. Tu. (2005). *Gambol* and *Tc1* are two distinct families of DD34E transposons: Analysis of the *Anopheles gambiae* genome expands the diversity of the *IS630-Tc1-mariner* superfamily. *Insect Mol Biol* 14 (5) 537-546.

Coy, M.R. and Z. Tu. *Tango* transposons in *Aedes aegypti*, *Anopheles gambiae*, and other mosquito species: A possible case of horizontal transfer. *Insect Mol Biol In press*.

Nene, V. *et al*., 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Submitted to *Science*. M. Coy is one of the four co-authors from the Tu laboratory.