DOI: 10.1534/genetics.107.081109

Note

Evolution and Horizontal Transfer of a DD37E DNA Transposon in Mosquitoes

James K. Biedler,[1,2] Hongguang Shao[2,3] and Zhijian Tu[2]

Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Manuscript received August 25, 2007
Accepted for publication September 24, 2007

ABSTRACT

ITmD37E, a unique class II transposable element (TE) with an ancient origin, appears to have been involved in multiple horizontal transfers in mosquitoes as *ITmD37E* sequences from 10 mosquito species of five genera share high nucleotide (nt) identities. For example, *ITmD37E* sequences from *Aedes aegypti* and *Anopheles gambiae*, which have an estimated common ancestor of 145–200 million years ago, display 92% nt identity. The comparison of *ITmD37E* and host mosquito phylogenies shows a lack of congruence. The wide distribution of conserved *ITmD37E*s in mosquitoes and the presence of intact copies suggest that this element may have been recently active.

TRANSPOSABLE elements (TEs) are mobile genetic elements that are able to replicate and increase their copy number in a genome (Doolittle and Sapienza 1980). They are divided into RNA-mediated (class I) and DNA-mediated (class II) elements on the basis of their mechanism of transposition. Class I elements include long terminal repeat (LTR) retrotransposons, non-LTR retrotransposons, and SINEs, all retrotransposing via an RNA intermediate. Class II elements include *Tc1*, *mariner*, and the *P* element, which transpose by a "cut-and-paste" mechanism.

Horizontal transfer has been reported for many class II elements in insects (Kidwell 1992; Robertson 1993, 2002; Bonnivard *et al.* 2000; Handler and McCombs 2000; Silva *et al.* 2004). We have previously described a DNA TE named *ITmD37E* in mosquitoes, which has a unique DD37E catalytic triad (Shao and Tu 2001). Here we report the ancient origin, evolution, and horizontal transfer of *ITmD37E* elements in mosquitoes.

***ITmD37E* in mosquitoes:** A representative *ITmD37E* element from *An. gambiae* is 1.3 kb in length and contains a 1008-bp open reading frame (ORF) and imperfect 27-bp terminal inverted repeats. Insertions are flanked by a TA dinucleotide target-site duplication. A genomic database search revealed 18 and 3 elements, respectively, in the *Ae. aegypti* and *An. gambiae* genomes. Full-length copies with intact ORFs could be found only in *An. gambiae*. Additional *ITmD37E* element copies were obtained from other mosquito species using genomic library screening and polymerase chain reaction and submitted to NCBI (see supplemental File S2 at http://www.genetics.org/supplemental/ for information about sequences used in this study). Interestingly, no representatives could be found in the *Culex pipiens* EST database or in the newly released genomic assembly (http://www.vectorbase.org).

***ITmD37E* TEs found in divergent taxa suggest an ancient origin:** Database searches (NCBI) revealed the existence of *ITmD37E* representatives in divergent freshwater invertebrates such as *Philodina roseola* (Arkhipova and Meselson 2005), *Dugesia ryukyuensis*, and *Hydra magnipapillata* (Figure 1). All three *H. magnipapillata* sequences were found in the EST database and are short sequences. Two of the three sequences include the region spanning the second D and E residues that are a part of the DD37E catalytic triad motif. The only *D. ryukyuensis* sequence that could be found was from an EST database and this sequence contains all three residues of the triad. The *H. magnipapillata* sequences group with mosquito *ITmD37E* sequences, having strong phylogenetic support. The *P. roseola* and *D. ryukyuensis* sequences are more distantly related to the mosquito *ITmD37E* sequences, although they have the DD37E motif. They may simply represent a sampling from a divergent paralogous lineage. In summary, this analysis

[1]*Corresponding author:* Department of Biochemistry, Virginia Polytechnic Institute and State University, 315 Fralin, W. Campus Dr., Blacksburg, VA 24061. E-mail: jbiedler@vt.edu

[2]These authors contributed equally to this article.

[3]*Present address:* Monell Chemical Senses Center, Philadelphia, PA 19104.

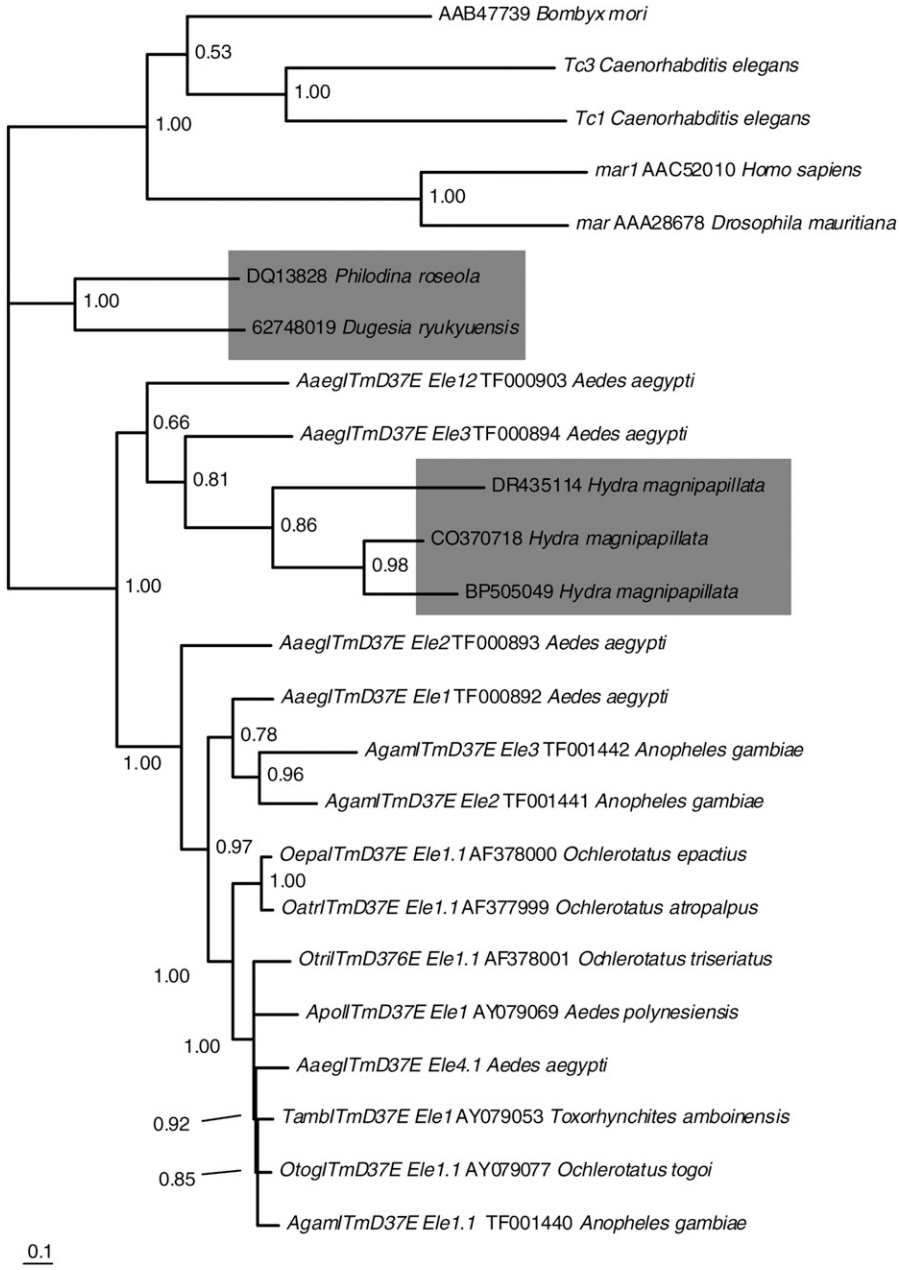

FIGURE 1.—*ITmD37E* TEs are of an ancient origin. Consensus tree (>50%) based on conceptual translations constructed using MrBayes version 3.1.2 (HUELSENBECK and RONQUIST 2001; RONQUIST and HUELSENBECK 2003). *ITmD37E* sequences from freshwater invertebrates are in shaded boxes. *Tc1*, *mariner*, and *DD37D* (*Bmmar1*) TEs are used to root the tree. Clade credibility values are shown at each node. The scale represents substitutions per site. All mosquito elements have intact ORFs except for *AaegITmD37E_Ele4.1*, *AgamITmD37E_Ele2*, and *AgamITmD37E_Ele3*, which were obtained from whole-genome sequence projects. Element name, accession number, and species name are given when applicable. See supplemental File S1 (http://www.genetics.org/supplemental/) for methods. Information pertaining to sequences can be found in supplemental File S2. See supplemental File S3 for alignment.

suggests that the *ITmD37E* TEs are a longstanding group.

High nucleotide identities between copies from divergent hosts suggest horizontal transfer: Several species from five genera have *ITmD37E* copies that share high nucleotide (nt) identities in their coding regions (Table 1). Sequences from *Ae. aegypti* and *An. gambiae*, which have an estimated common ancestor of 145–200 million years ago (MYA) (KRZYWINSKI *et al.* 2006), have 93% identity in the coding region. The conservation is not restricted to the coding regions, as both 5′- and 3′-untranslated regions (UTRs) have a consistently high degree of conservation. The 162-bp 5′-UTR and the 125-bp 3′-UTR have 90 and 87% nt identity, respectively. Sequence comparisons between other highly divergent species such as *Ae. aegypti* and *Toxorhynchites amboinensis* also demonstrate high nt identities (>95%).

It is important to note that of 300 class II elements from *Ae. aegypti* and *An. gambiae*, there are only 2 elements with copies sharing at least 80% nt identity for 500 bp or more between the two genomes. Two hundred forty-eight *Ae. aegypti* elements and 52 *An. gambiae* elements (http://tefam.biochem.vt.edu/tefam/index.php) were used as queries in BLASTN against *Ae. aegypti* or *An. gambiae* whole-genome sequences and then output was filtered for minimum 80% nt identity and minimum hit lengths of 500 bp. *AaegITmD37E_Ele4* and *AgamITmD37E_Ele1* have the highest identities, with 90% for 1300 bp [alignment by CLUSTAL_X (THOMPSON *et al.* 1997) and distance determination by PAUP

TABLE 1

Pairwise nucleotide identities between select *ITmD37E*-coding sequences from species of five genera

	Ar. subalbatus	*O. atropalpus*	*An. gambiae*	*T. amboinensis*	*O. epactius*	*O. togoi*	*Ae. aegypti*	*Ae. polynesiensis*
O. atropalpus	79.8							
An. gambiae	93.6	77.9						
T. amboinensis	95.5	79.5	93.5					
O. epactius	80.3	97.6	78.4	79.9				
O. togoi	95.8	79.7	94.3	96.2	80.1			
Ae. aegypti	95.2	78.9	93.2	95.3	79.0	95.4		
Ae. polynesiensis	93.2	78.4	91.1	93.0	78.7	93.2	93.2	
O. triseriatus	94.6	80.6	92.1	94.2	81.1	94.6	93.9	92.0

Numbers are percentages. Only coding regions were compared because they were the only sequence available for some copies. Pairwise nt identities were determined by PAUP (Swofford 2002) after alignment with CLUSTAL_X (Thompson *et al.* 1997). *Ae. aegypti*, *AaegITmD37E_Ele 4.1*; *Ae. polynesiensis*, *ApolITmD37E_Ele1*; *Ar. subalbatus*, *AsubITmD37E_Ele1.2*; *O. atropalpus*, *OatrITmD37E_Ele1.1*; *O. epactius*, *OepaITmD37E_Ele1.1*; *O. togoi*, *OtogITmD37E_Ele1.1*; *O. triseriatus*, *OtriITmD37E_Ele1.1*; *An. gambiae*, *AgamITmD37E_Ele1.1*; *T. amboinensis*, *TambITmD37E_Ele1*. See supplemental File S1 at http://www.genetics.org/supplemental/ for methods. Refer to supplemental File S2 for sequence information.

(Swofford 2002) results in 92%]. The next highest identity match was between *Tango* elements (80% identity over a 500-bp fragment), which have been reported to be involved in horizontal transfer (Coy and Tu 2007).

Host and *ITmD37E* phylogenies are incongruent: Comparison of the host species vitellogenin C (Vg-C) and *ITmD37E* phylogenies shows that there is a lack of congruence (Figure 2, A and B). There are three major

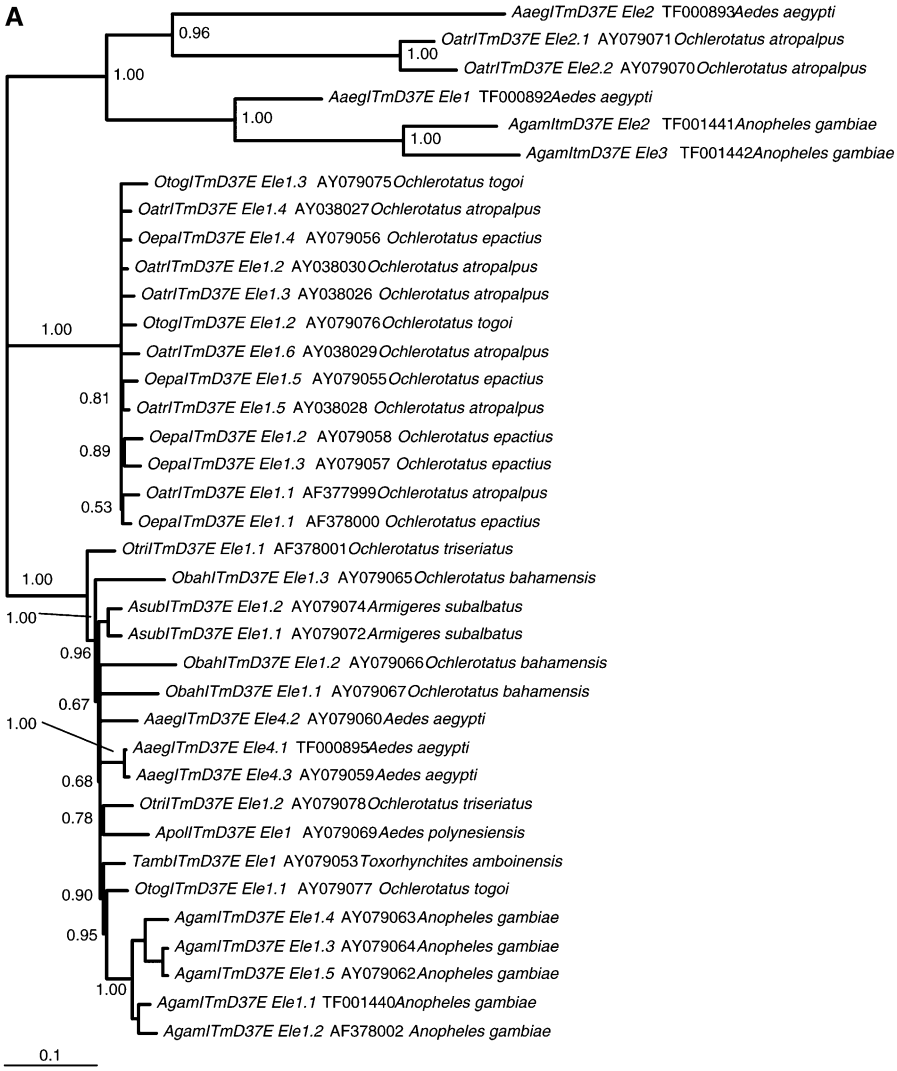

Figure 2.—Comparison of host and *ITmD37E* phylogenies. Both trees shown are consensus trees (>50%) constructed using MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Clade credibility values are shown at each node. The scale represents substitutions per site. (A) *ITmD37E* phylogeny based on nt sequence from ORFs. Tree is rooted with divergent mosquito *ITmD37E* elements seen at the top of the tree. Element name, accession number, and species name are given when applicable. Information regarding sequences used in this study can be found in supplemental File S2 (http://www.genetics.org/supplemental/). (B) Mosquito host phylogeny based on Vg-C nt sequence. Tree is rooted with *An. gambiae* Vg-C. The five genera that have species containing *ITmD37E* sequences are shaded. The *Armigeres subalbatus* sequence is found within the Aedes group, which is consistent with previous analyses (Isoe 2000). Most Vg-C sequences were obtained from Isoe (2000). See supplemental File S1 for methods and supplemental File S3 for alignments.

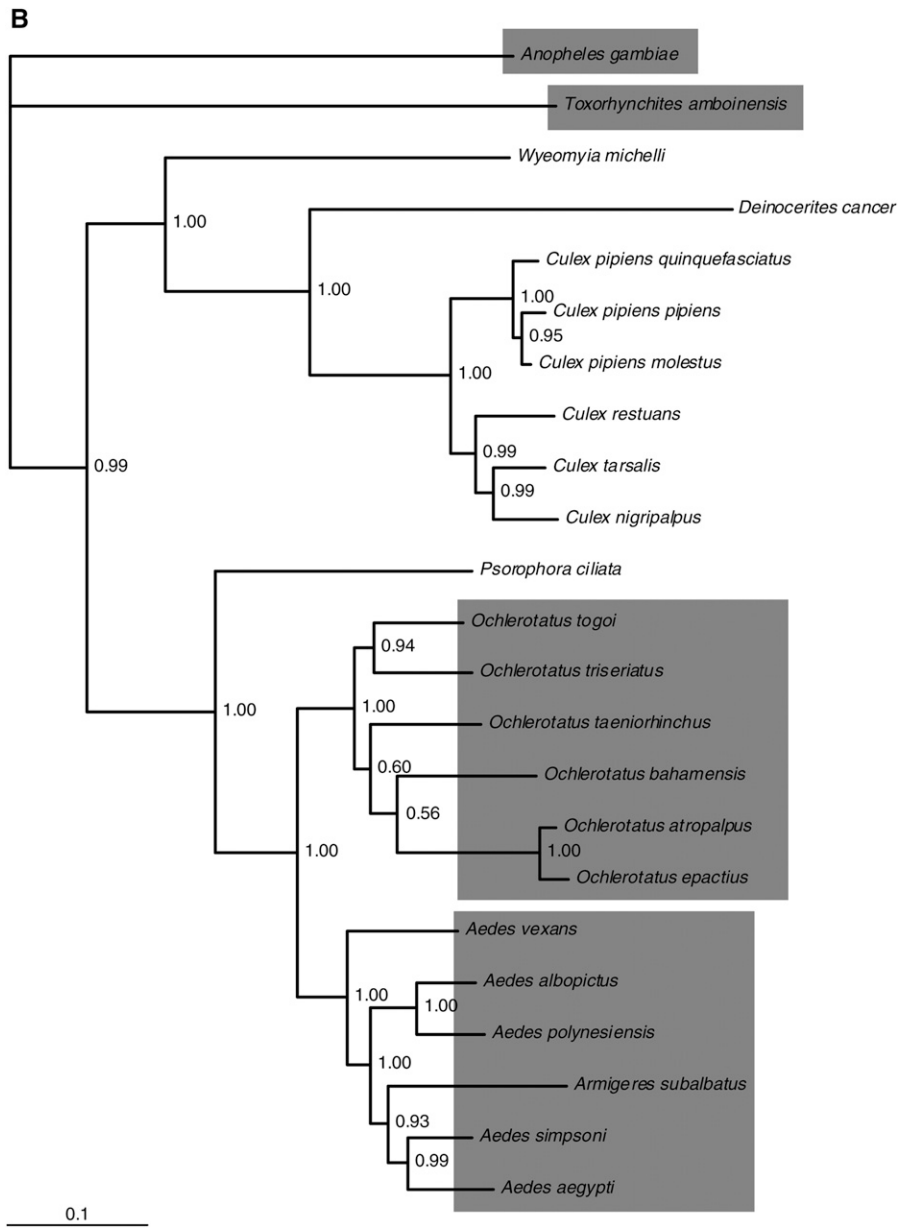

FIGURE 2.—*Continued.*

groups in Figure 2A. The group at the bottom of Figure 2A shows a marked lack of congruence. The *An. gambiae* sequences do form a clade, but nevertheless the phylogeny is perplexing compared to the host phylogeny. For example, the relative distances between sequences of *Ae. aegypti* (from both the library screen and the genomic database) and several other species such as *T. amboinensis* and *An. gambiae* (distant relatives), *Ae. polynesiensis*, *Ochlerotatus triseriatus*, and *O. bahamensis* (closely related species) do not make sense when compared to the host phylogeny. The phylogeny involving these sequences is neither in agreement nor the same compared to the host phylogeny when the two are superimposed.

Selection pressure and codon bias are not responsible for the high conservation found between three *ITmD37E* copies from *Ae. aegypti*, *T. amboinensis*, and *An. gambiae*: It is possible that selection pressure or codon bias could contribute to the high conservation of *ITmD37E* sequences. To investigate these possibilities, *ITmD37E* sequences from each of the three most divergent host species (*Ae. aegypti*, *An. gambiae*, and *T. amboinensis*) were analyzed for selection pressure (Table 2). d_S/d_N, a ratio of the synonymous and nonsynonymous substitutions between two sequences, was determined by SNAP (Nei and Gojobori 1986; Korber 2000). Values ranged from 1.0 to 3.6 for all pairwise comparisons, showing very low to no significant selection. There is no concern for substitution saturation in these cases, as p_s and p_n values are low. p_s and p_n are the proportion of observed synonymous and nonsynonymous substitutions, respectively. d_S/d_N values for Vg-C, a host gene known to be relatively rapidly evolving (Isoe 2000), were also obtained from the same three species. These values are significantly higher than that found for

TABLE 2

Selection pressure analysis of *ITmD37E*- and Vg-C-coding sequences

Sequence 1	Sequence 2	p_s	p_n	d_S	d_N	d_S/d_N	% aa ID
AaegITmD37E_Ele4.1	*AgamITmD37E_Ele1.1*	0.11	0.05	0.11	0.06	2.0	87.5
AaegITmD37E_Ele4.1	*TambITmD37E_Ele1*	0.05	0.05	0.05	0.05	1.0	89.6
AgamITmD37E_Ele1.1	*TambITmD37E_Ele1*	0.14	0.04	0.16	0.04	3.6	90.5
*Aaeg*Vg-C	*Agam*Vg-C	0.65	0.22	1.50	0.26	5.7	61.4
*Aaeg*Vg-C	*Tamb*Vg-C	0.72	0.21	2.52	0.24	10.5	64.4
*Agam*Vg-C	*Tamb*Vg-C	0.80	0.23	NA[a]	0.28	NA[a]	60.4
*Aalb*Vg-C	*Apol*Vg-C	0.30	0.02	0.38	0.02	15.3	94.8
*Oatr*Vg-C	*Oepa*Vg-C	0.12	0.01	0.13	0.01	13.7	97.8

Manually codon-aligned sequences (see supplemental File S4 at http://www.genetics.org/supplemental/) were used to determine selection pressure by SNAP (http://hcv.lanl.gov/content/hcv-db/SNAP/SNAP.html) (Korber 2000), a program that uses the method of Nei and Gojobori (1986). *AgamITmD37E_Ele1.1* and *TambITmD37E_Ele1* have intact coding sequences. *AaegITmD37E_Ele4.1* has multiple frameshifts that were corrected by codon alignment. Most Vg-C sequences were obtained from Isoe (2000). See supplemental File S1 for methods.

[a] SNAP does not give an output for d_S when $p_s > 0.75$, indicating substitution saturation.

ITmD37E sequences. Because these Vg-C comparisons are possibly saturated with respect to synonymous substitution (p_s is near or exceeding 0.75), we also compared Vg-C sequences from closely related species in the Aedes and Ochlerotatus genera, which have p_s values similar to those of the above-mentioned *ITmD37E* sequence comparisons. These Vg-C comparisons also displayed high d_S/d_N values, ranging from 13.7 to 15.3. Therefore, the low d_S/d_N values from the *ITmD37E* comparisons suggest that the high sequence identity between the *ITmD37E* elements does not result from high selection pressure. The effective number of codons (Wright 1990) determined by PDA (http://dpdb.uab.es/pda/pda.asp) (Casillas and Barbadilla 2004) for all three *ITmD37E* sequences is 59–60, demonstrating no codon bias. Possible values range from 20 (one codon per aa—high bias) to 61 (all codons used equally—no bias). These analyses suggest that selection pressure and codon bias are not responsible for the observed high degree of conservation.

Discussion: There are generally three lines of evidence used to make a case for horizontal transfer: the discovery of sequences from divergent taxa having high nt identities, incongruence between TE and host phylogenies, and a "patchy TE distribution" among related host taxa. We have shown the first two types of evidence here, and the support for the third is weak but worth mentioning (see below). First, high nt identities have been found between *ITmD37E* sequences from 10 species of five genera (Table 1, Figure 2A). Particularly noteworthy is the comparison of copies from *Ae. aegypti* and *An. gambiae*, species that are estimated to have diverged between 145 and 200 MYA. Second, host and TE phylogenies are clearly incongruent (Figure 2). Here we have many copies with such high nt identities that branch lengths are very short and phylogenetic resolution is low for some sequences. Finally, regarding patchy distribution, no *ITmD37E* representatives could be found in the *C. pipiens* genome assembly (http://www.vectorbase.org) by database search or by genomic library screen (not shown). While this is consistent with patchy distribution resulting from horizontal transfer, it is possible that *ITmD37E* was simply lost from this lineage.

When all these evidences are taken together, the case for horizontal transfer of *ITmD37E* in mosquitoes is strong. While alternative explanations can be provided, they are not likely. It could be argued that *ITmD37E* copies have inserted into genomic regions having a low substitution rate. However, this argument is hard to make, given the nt identities of 92% between copies from *Ae. aegypti* and *An. gambiae*, species with a common ancestor from 145 to 200 MYA. This argument requires that conserved *ITmD37E* copies from all species be inserted into locations with low substitution rates. We have determined the location of *AgamITmD37E_Ele1* copies because chromosomal assignment is available only for *An. gambiae*. Copies are present on all chromosome arms (determined by BLAST using Ensembl; not shown). Although we cannot completely rule out the low-substitution-rate hypothesis, it seems unlikely.

ITmD37E could have been "co-opted" for a host function and therefore be highly conserved. First, it is unlikely that such an indispensable function in so many divergent mosquitoes would have been lost from *C. pipiens*. Second, substitution analysis performed to detect selection pressure by SNAP (Nei and Gojobori 1986; Korber 2000) (http://hcv.lanl.gov/content/hcv-db/SNAP/SNAP.html) using sequences from the three divergent species *Ae. aegypti*, *An. gambiae*, and *T. amboinensis* shows a rather weak selection for sequences with this degree of conservation (Table 2). Even when we compare *ITmD37E* sequences to Vg-C, a gene under moderate selection, *ITmD37E* sequences still demonstrate a much lower selection pressure yet have much higher sequence conservation. Therefore, there is no indication that high conservation of *ITmD37E* sequences is due to selection. A low d_S for *ITmD37E* sequences is

consistent with the horizontal transfer hypothesis, as there has not been enough time for the accumulation of substitutions. This makes sense if horizontal transfer was recent and therefore not enough time had elapsed for selection to become evident by d_S/d_N analysis. Third, we have also shown that the conservation is not from codon bias, as the effective number of codons (Wright 1990) for these sequences is very high, demonstrating effectively no bias.

It is difficult to determine the direction of horizontal transfer, but the relatively high nt identities among copies in both divergent and closely related taxa suggest that *ITmD37E* has been introduced into mosquitoes recently in evolution. It is interesting that both *An. gambiae* and *Ae. aegypti* have long coexisted in parts of Africa, providing a possible ecological connection for horizontal transfer, although we can only speculate about the mechanism. A common virus is a likely candidate, where TE insertions that do not inactivate the viral copy are excised and inserted into a naive genome after infection. Although the *piggybac* TE from a lepidopteran cell line has been found to be responsible for insertional mutations in baculoviruses (Fraser *et al.* 1983), no direct evidence for viral transmission as a mechanism for horizontal transfer has been demonstrated.

The *ITmD37E* family has been very successful in mosquitoes evidenced by its widespread distribution in 10 species of five genera. Except for those found in a few other invertebrate species (Figure 1), to our knowledge no other *ITmD37E* representatives are known at this time. It will be interesting to see if this group is restricted to invertebrates.

The horizontal transfer of *ITmD37E* and the presence of several copies with intact ORFs indicate that *ITmD37E* has been recently active. Of the 39 mosquito copies used in this study, 9 from five species of four genera have intact ORFs. This may suggest that *ITmD37E* elements are good candidates for developing molecular tools for transgenesis, as there has been significant success in using class II elements as molecular tools in mosquitoes (Adelman *et al.* 2002; Perera *et al.* 2002; O'Brochta *et al.* 2003). Efforts to determine the transposition activity of *ITmD37E* are currently underway.

We thank Jun Isoe from the laboratory of Henry Hagedorn for providing genomic libraries and Vg-C sequences and Shirley Luckhart for providing the *An. gambiae* genomic library. This work was supported by a grant from the National Institutes of Health (AI42121) to Z.T.

LITERATURE CITED

Adelman, Z. N., N. Jasinskiene and A. A. James, 2002 Development and applications of transgenesis in the yellow fever mosquito, Aedes aegypti. Mol. Biochem. Parasitol. **121:** 1–10.

Arkhipova, I. R., and M. Meselson, 2005 Diverse DNA transposons in rotifers of the class Bdelloidea. Proc. Natl. Acad. Sci. USA **102:** 11781–11786.

Bonnivard, E., C. Bazin, B. Denis and D. Higuet, 2000 A scenario for the hobo transposable element invasion, deduced from the structure of natural populations of Drosophila melanogaster using tandem TPE repeats. Genet. Res. **75:** 13–23.

Casillas, S., and A. Barbadilla, 2004 PDA: a pipeline to explore and estimate polymorphism in large DNA databases. Nucleic Acids Res. **32:** W166–W169.

Coy, M. R., and Z. Tu, 2007 Genomic and evolutionary analyses of Tango transposons in Aedes aegypti, Anopheles gambiae and other mosquito species. Insect Mol. Biol. **16:** 411–421.

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

Fraser, M. J., G. E. Smith and M. D. Summers, 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.

Handler, A. M., and S. D. McCombs, 2000 The piggyBac transposon mediates germ-line transformation in the Oriental fruit fly and closely related elements exist in its genome. Insect Mol. Biol. **9:** 605–612.

Huelsenbeck, J. P., and F. Ronquist, 2001 MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:** 754–755.

Isoe, J., 2000 Comparative analysis of the vitellogenin genes of the Culicidae. Ph.D. Thesis, University of Arizona, Tucson, AZ.

Kidwell, M. G., 1992 Horizontal transfer of P elements and other short inverted repeat transposons. Genetica **86:** 275–286.

Korber, B., 2000 Computational analysis of HIV molecular sequences, pp. 55–72 in *HIV Signature and Sequence Variation Analyisis*, edited by A. G. Rodrigo and G. H. Learn. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Krzywinski, J., O. G. Grushko and N. J. Besansky, 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.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

O'Brochta, D. A., N. Sethuraman, R. Wilson, R. H. Hice, A. C. Pinkerton *et al.*, 2003 Gene vector and transposable element behavior in mosquitoes. J. Exp. Biol. **206:** 3823–3834.

Perera, O. P., I. R. Harrell and A. M. Handler, 2002 Germ-line transformation of the South American malaria vector, Anopheles albimanus, with a piggyBac/EGFP transposon vector is routine and highly efficient. Insect Mol. Biol. **11:** 291–297.

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, pp. 1093–1110 in *Mobile DNA II*, edited by N. L. Craig, R. Craigie, M. Gellert and A. M. Lambowitz. American Society for Microbiology, Washington, DC.

Ronquist, F., and J. P. Huelsenbeck, 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572–1574.

Shao, H., and Z. Tu, 2001 Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics **159:** 1103–1115.

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

Swofford, D. L., 2002 *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

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

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Communicating editor: D. Voytas