Copyright © 2001 by the Genetics Society of America

Expanding the Diversity of the *IS630-Tc1-mariner* Superfamily: Discovery of a Unique DD37E Transposon and Reclassification of the DD37D and DD39D Transposons

Hongguang Shao and Zhijian Tu

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

Manuscript received June 25, 2001
Accepted for publication August 27, 2001

ABSTRACT

A novel transposon named *ITmD37E* was discovered in a wide range of mosquito species. Sequence analysis of multiple copies in three Aedes species showed similar terminal inverted repeats and common putative TA target site duplications. The *ITmD37E* transposases contain a conserved DD37E catalytic motif, which is unique among reported transposons of the *IS630-Tc1-mariner* superfamily. Sequence comparisons and phylogenetic analyses suggest that *ITmD37E* forms a novel family distinct from the widely distributed *Tc1* (DD34E), *mariner* (DD34D), and *pogo* (DDxD) families in the *IS630-Tc1-mariner* superfamily. The inclusion in the phylogenetic analysis of recently reported transposons and transposons uncovered in our database survey provided revisions to previous classifications and identified two additional families, *ITmD37D* and *ITmD39D*, which contain DD37D and DD39D motifs, respectively. The above expansion and reorganization may open the doors to the discovery of related transposons in a broad range of organisms and help illustrate the evolution and structure-function relationships among these distinct transposases in the *IS630-Tc1-mariner* superfamily. The presence of intact open reading frames and highly similar copies in some of the newly characterized transposons suggests recent transposition. Studies of these novel families may add to the limited repertoire of transgenesis and mutagenesis tools for a wide range of organisms, including the medically important mosquitoes.

TRANSPOSABLE elements, or mobile genetic elements, are widely distributed in prokaryotic and eukaryotic genomes (Kidwell and Lisch 2000). They have the ability to replicate and spread in the genomes as primarily "selfish" genetic units (Doolittle and Sapienza 1980). On the other hand, recent evidence suggests that the self-replicating property may have enabled transposable elements to provide the genomes with potent agents to generate tremendous genetic plasticity (Kidwell and Lisch 2000). Transposable elements can be classified by the mechanisms of their transposition as DNA-mediated or RNA-mediated elements. The transposition of RNA elements such as retrotransposons involves a reverse transcription step that generates cDNA from RNA transcripts of the RNA elements. The cDNA molecules are integrated in the genome with the help of the integrase, which is encoded by retrotransposons and retroviruses (*e.g.*, Moore *et al.* 1995). DNA-mediated elements such as *P*, *hobo*, *Tc1*, and *mariner* usually transpose through a cut-and-paste mechanism (*e.g.*, Plasterk *et al.* 1999). They are characterized by terminal inverted repeats (TIRs) flanking a gene encoding a transposase that catalyzes the transposition reaction. Despite the differences in the transposition mechanisms, the integrase of some RNA elements and the transposases of some DNA elements are thought to have evolved from a common origin (Capy *et al.* 1997). Thus prokaryotic *IS* elements, eukaryotic *Tc1* and *mariner* transposons, and eukaryotic retrotransposons and retroviruses form a megafamily that shares similar signature sequences or motifs in the catalytic domain of its respective transposase and integrase (Capy *et al.* 1996, 1997). The common motif for this transposase-integrase megafamily is a conserved D(Asp)DE(Glu) or DDD catalytic triad. The DDE(D) triad is an essential part of the catalytic site and mutations in the triad abolish the transposase activity in *in vivo* excision assays (Lohe *et al.* 1997). The distances between the first two D's are variable while the distances between the last two residues in the catalytic triad are mostly invariable for a given transposon family in eukaryotes, indicating functional importance.

Among these divergent elements, the eukaryotic DNA transposon families *Tc1* and *mariner* and the bacterial *IS630* element and its relatives in prokaryotes and ciliates comprise a superfamily, the *IS630-Tc1-mariner* superfamily, which is based on overall sequence similarities and a common TA dinucleotide insertion target (Henikoff 1992; Doak *et al.* 1994; Robertson 1995; Capy *et al.* 1996). On the basis of the analysis of the more conserved catalytic domain, a group of DNA transposons named *pogo* may also belong to the *IS630-Tc1-mariner*

Corresponding author: Zhijian Tu, Department of Biochemistry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. E-mail: jaketu@vt.edu

superfamily (Smit and Riggs 1996; Capy *et al.* 1997). Within the *IS630-Tc1-mariner* superfamily, the *IS630*-like elements contain a DDxE motif where x indicates variable distance. However, excluding mutations in apparently defective copies, *Tc1*-like elements identified in fungi, invertebrates, and vertebrates all contain a DD34E motif while most *mariner* elements identified in flatworm, insects, and vertebrates contain a DD34D motif (Table 1). There are two reported exceptions to the DD34D motif in *mariners*: the DD37D motif in an insect *mariner Bmmar1* and the DD39D motif in a soy bean *mariner Soymar1* (Robertson and Asplund 1996; Jarvik and Lark 1998). Transposons related to *Bmmar1* and *Soymar1* have also been found in rice and nematodes through database analysis (Tarchini *et al.* 2000; Lampe *et al.* 2001). Here we report the discovery of a novel family of transposons widely distributed in mosquitoes, which contain a unique DD37E motif. We provide a detailed analysis of the characteristics shared by the DD37D and DD39D transposons and suggest that some of the newly discovered DD37D transposons may have been transposing very recently. We present evidence suggesting that the DD37E, DD37D, and DD39D transposons make up three new families in the *IS630-Tc1-mariner* superfamily, defined by their respective catalytic motifs. We therefore name them *I(S630)T(c1)m(ariner)D37E*, *ITmD37D*, and *ITmD39D*. The above classification represents an expansion and reorganization of the *IS630-Tc1-mariner* superfamily. The significance of the discovery of the *ITmD37E* transposons in mosquitoes and the significance of the above classification have been discussed in light of the importance of the *Tc1* and *mariner* families in current genetic and evolutionary studies.

MATERIALS AND METHODS

Polymerase chain reaction and cloning: Genomic DNA of adult *Anopheles gambiae* mosquitoes was isolated using DNAzol from Molecular Research Center (Cincinnati). *An. gambiae* DNA covering the entire open reading frame (ORF) of an *ITmD37E* transposon was obtained using polymerase chain reaction (PCR). The two primers, ATGGAAGCCGAAAGAAG GGA and GCAAATGTAGCGTTTTCTTCAT, were designed according to two short sequences (AL150661 and AL143513) in the *An. gambiae* sequence-tagged site (STS) database that match an *ITmD37E* element in the *Ae. atropalpus* mosquito (AY038030). PCR was performed as previously described (Tu and Hagedorn 1997). The PCR product was separated and cut from an agarose gel and purified using the Sephaglas Bandprep Kit from Amersham Pharmacia Biotech (Arlington Heights, IL). Purified PCR products were cloned in a pCR 2.1 vector using a TA cloning kit from Invitrogen (Carlsbad, CA). Multiple clones were sequenced as described below.

Screening of λ ZapExpress genomic libraries: A genomic library prepared using DNA from the *Ae. atropalpus* mosquito and the λ ZapExpress vector (Stratagene Cloning Systems, La Jolla, CA) was provided by. J. Isoe in the laboratory of H. Hagedorn at the University of Arizona. Genomic libraries prepared using the same λ ZapExpress vector and DNA from the *Ae. epactius* and *Ae. triseriatus* mosquitoes were provided by R. Nussenzveig in the laboratory of M. Wells at the University of Arizona. The above libraries were constructed as described in Tu (2000). The average insert size of these libraries is ~5 kb. The *Ae. atropalpus* and *Ae. epactius* libraries were screened using a PCR fragment probe corresponding to the C-terminal coding region of an *ITmD37E* transposon in *Ae. atropalpus* (AY038030, primers CGACCRTCCMGTAATGYTTTSGCC and CATTAG GCGGCGCACACC). The *Ae. triseriatus* library was screened using a probe corresponding to the entire ORF of an *ITmD37E* transposon in *An. Gambiae*, which was obtained by PCR as described above. The choice of probes was based on a preliminary genomic DNA dot blot analysis (data not shown). Both probes were single stranded as the labeling reactions were performed using asymmetric PCR amplifications. The labeling condition was the same as that described by Tu and Hagedorn (1997), who used a digoxigenin-dUTP labeling mixture. MagnaGraph nylon membranes (Micron Separation, Westborough, MA) were used to lift the plaques. Hybridization was carried out at 55° as described in Tu (2000). The first set of washes was at 55° with 2× SSC and 0.1% SDS. The second set of washes was at 55° with 0.5× SSC and 0.1% SDS. The label was detected using an alkaline phosphatase-linked antidigoxigenin antibody and two phosphatase substrates, X-phosphate and nitroblue tetrazolium salt, following the protocol of Boehringer Mannheim Biochemicals (Indianapolis).

***In vivo* excision and DNA sequencing:** Inserts in λ ZapExpress clones were excised *in vivo* into the pBK-CMV phagemid vector, using the ExAssist helper phage from Stratagene Cloning Systems. Sequencing of the λ ZapExpress clones from genomic libraries and the TA clones from PCR amplification was done either at the sequencing facility at Virginia Tech using an automated sequencer (model 377, Applied Biosystems International, Foster City, CA) or in our laboratory using a 4200S Gene ReadIR sequencing instrument from Li-Cor (Lincoln, NE).

Sequence analysis and phylogenetic inference: Searches for matches of either nucleotide or amino acid sequences in the database (nonredundant GenBank + EMBL + DDBJ + PDB) were done using Fasta of GCG (Genetics Computer Group, Madison, WI, version 10, 1999) and BLAST (Altschul *et al.* 1997). Pairwise comparisons were done using Gap or Bestfit of GCG. Multiple sequences were aligned using either ClustalW (Thompson *et al.* 1994) or Pileup of GCG. The parameters such as gap weight and gap length weight are described in the legend of the alignment figures. Profiles of aligned sequences were generated using Profilemake of GCG. *Z* scores of the comparisons between a sequence and a profile were obtained using Profilesearch of GCG with the SWISS-PROT database plus sequences of interest. Specific parameters are described in the footnote of Table 3. Phylogenetic trees were constructed using minimum evolution, neighbor-joining, and maximum-parsimony methods of PAUP* 4.0 b8 (Swofford 2001). Specific parameters are described in the legend to Figure 4. Five hundred bootstrap replicates were used to assess the confidence in the grouping (Felsenstein and Kishino 1993). Pairwise identities of aligned sequences were converted from pairwise differences calculated using PAUP* 4.0 b8.

The sequence data presented in this article have been submitted to the EMBL/GenBank Data Libraries under the accession nos. AF377999-8002 and AY038026-30. The alignment presented in this article has been submitted to the EMBL/GenBank Data Libraries under the accession no. DS47334.

RESULTS

Discovery of *ITmD37E*, a novel transposon containing a unique DD37E catalytic motif, in the rockpool mosquito *Ae. atropalpus*: The first *ITmD37E* element was

discovered fortuitously in a clone isolated from an *Ae. atropalpus* genomic library. The sequence of the entire clone has been deposited in GenBank (AY038030). To further characterize this element, we screened the *Ae. atropalpus* genomic library, using a digoxigenin-labeled single-stranded DNA probe corresponding to the 3′ region of the putative transposon. Approximately 270 positive clones were obtained from a total of 30,000 clones. Five of these positive clones were sequenced and deposited in GenBank (AF377999 and AY038026–29). Figure 1A shows the nucleotide and the deduced amino acid sequence of the consensus of these six *ITmD37E* transposons. It contains an intact ORF encoding 336-amino-acid residues flanked by 27-bp TIRs and a putative TA target site duplication. The six copies showed 97.0–98.6% identity at the nucleotide level (Figure 1B), indicating relatively recent transposition. The boundaries of the *ITmD37E* elements are confirmed on the basis of sequence comparisons between the six clones because there is no sequence similarity outside the predicted transposon (Figure 1B). All six copies are flanked by putative TA target duplications. The similarities to *Tc1*-like transposases (E values $= 1 \times 10^{-7}$ for some matches during BLAST searches) and the putative TA target duplications suggest that *ITmD37E* belongs to the *IS630-Tc1-mariner* superfamily (Doak *et al.* 1994). As shown in Table 1, the *IS630-Tc1-mariner* superfamily of transposons contains either a DDE or a DDD catalytic triad. Excluding the apparently defective copies that accumulated many mutations, most, if not all, eukaryotic DDE transposons in the superfamily contain a DD34E motif (Table 1). However, as shown in Figure 1A, the *ITmD37E* transposon contains a unique DD37E catalytic motif. As shown by further analysis described below, these elements represent a novel family within the *IS630-Tc1-mariner* superfamily. The family is named *ITmD37E* to reflect its unique catalytic motif. The transposons found in *Ae. atropalpus* are named *Ae.atropalpus.ITmD37E1*, following a naming convention proposed for *mariner* (Robertson and Asplund 1996) as described in Table 1. Instead of referring to an individual copy, the Arabic numeral refers to a distinct type of *ITmD37E* transposon in a species.

***ITmD37E* is conserved and widely distributed in mosquitoes:** The *ITmD37E* transposon was also discovered in other species of mosquitoes. Sequences of five and two full-length copies were obtained from clones isolated from *Ae. epactius* and *Ae. triseriatus* genomic libraries, respectively (see Table 1 for accession numbers of the copies included in the analysis). In addition, sequences of the entire ORFs of *ITmD37E* transposons were also obtained from a distantly related mosquito, *An. gambiae*, using PCR designed according to two short sequences (AL150661 and AL143513) in the *An. gambiae* STS database that match the known *ITmD37E* transposons. Eight PCR clones were sequenced, which showed 92.0–98.7% nucleotide identities to each other. These PCR clones are 93.7–97.9% and 89.4–95.0% identical to the two *An. gambiae* STS sequences, AL150661 and AL143513, respectively. The identities to AL143513 are slightly lower because only a 150-bp fragment at the end of AL143513 matches *ITmD37E*. The consensus for each species and a number of individual copies all contain an uninterrupted ORF encoding a 336-amino-acid transposase with the DD37E motif, although some copies contain stop codons while others contain an ORF extended by an extra 16-amino-acid residue. Therefore, the DD37E motif must have been an important functional motif in these mosquito transposons. As shown in Table 2, the full-length copies of these transposons share highly similar TIRs that are all flanked by putative TA target site duplications. Their transposase proteins are 72.8–94.3% identical, which was calculated using PAUP as described in materials and methods. They seem to form two groups on the basis of sequence similarities: the *Ae. atropalpus* and *Ae. epactius* group and the *Ae. triseriatus* and *An. gambiae* group. There is >89% intragroup identity and 72.8–80.6% intergroup identity. Moreover, related transposons have been identified in nine additional species in five mosquito genera including Aedes, Armigeres, Culex, Toxorhynchites, and Anopheles, on the basis of DNA dot blot, genomic library screening, and preliminary sequence analysis (data not shown). Thus *ITmD37E* is widely distributed in mosquitoes.

Relative abundance of *ITmD37E* in different mosquito genomes: There are ~40 and ~20 copies of *ITmD37E* in *An. gambiae* and *Ae. triseriatus*, respectively. The copy number in the *An. gambiae* haploid genome was extrapolated on the basis of the fact that two *ITmD37E* fragments were found in the *An. gambiae* STS database, which contains >14 Mb of genomic sequences (http://bioweb.pasteur.fr/BBMI), and that the *An. gambiae* genome is 270 Mb (Besansky and Powell 1992). The copy number in *Ae. triseriatus* was estimated on the basis of the average insert size of the genomic library, the number of positive plaques, the total plaques screened, and the known genome size of *Ae. triseriatus* (Rai and Black 1999). The detailed calculation method is described in Tu (2000). Although the copy number of *ITmD37E* in *Ae. atropalpus* and *Ae. epactius* cannot be estimated because of unknown genome sizes, it is possible to assess the average frequency of *ITmD37E* in these genomes on the basis of the ratio of positive plaques over total plaques screened. In every 100 Mb of the genomic DNA, there are ~1000 and ~200 copies of *ITmD37E* in *Ae. epactius* and *Ae. atropalpus*, respectively. However, the frequencies of *ITmD37E* per 100 Mb genomic DNA are much lower in *Ae. triseriatus* and *An. gambiae*, ~2 and ~14 copies, respectively. Although the estimation for the three Aedes species may be influenced by the possible bias of the genomic libraries, the >100-fold differences observed here are probably large enough to override the potential bias. Thus the relative abundance of *ITmD37E* appears to be correlated with the

A

```
taAAGTCTGCTTCATTTAATATTGGAACAAATTCTTTTGCTCATTGTGGCGCTCCTAGTGGACGGATTTGGAAACTTTTTCACCCACGTGTCGGGAAATTCATTACCTTTCACCATGTA   120
TTTATGTCATAACACCAAACGATAGCATTTTGTAAACAACCGCCATGGAAGCCGAACGGAGAGATAAAATTGTGCACAGTTTTCTTGAAAATCCATTGTTGTCGGCATCTAAGCTAGCTA   240
                                                 M  E  A  E  R  R  D  K  I  V  H  S  F  L  E  N  P  L  L  S  A  S  K  L  A  K
AACAGCTTAAAATGCCCAGAAATACCGTATGGCGTGTTATCAAGCAGTACAAGGAAACATTGACGACGGCTCGGAAGCCGCATTCGAAGCGTCGGAGTGGAACTGTCGACCGGAAACTGC   360
  Q  L  K  M  P  R  N  T  V  W  R  V  I  K  Q  Y  K  E  T  L  T  T  A  R  K  P  H  S  K  R  R  S  G  T  V  D  R  K  L  R
GTGGGAAAGTCATCAAGGCCGTCAAGAGGAATCCCAATCTTTCAGACCGCGATTTGGCCAATAAGTTCCAGGCCGCTCACAGTACGGTGCGACGAATTCGTCTCCGGGAAGGAAGAAGGT   480
  G  K  V  I  K  A  V  K  R  N  P  N  L  S  D  R  D  L  A  N  K  F  Q  A  A  H  S  T  V  R  R  I  R  L  R  E  G  R  R  S
CATTCCGAGCCAGCAAACAGCCAAATCGGACGCTGAAGCAGAACAATGTGGCCAGAATCCGTGCTCGAAAGCTGTACGACCAAGTGCTGACCAAGTTCGACGGATGTATTCTGATGGACG   600
  F  R  A  S  K  Q  P  N  R  T  L  K  Q  N  N  V  A  R  I  R  A  R  K  L  Y  D  Q  V  L  T  K  F  D  G  C  I  L  M  D (D)
ATGAGACCTACGTGAAGGCGGACTTTGGGCAAATCCCAGGTCAAAAATTTTATTTGGCGACGGCTCGGGGGGATGTACCTGCAAAATTTAAGTTCGTGTTTGCGGATAAATTCGCCCGCA   720
  E  T  Y  V  K  A  D  F  G  Q  I  P  G  Q  K  F  Y  L  A  T  A  R  G  D  V  P  A  K  F  K  F  V  F  A  D  K  F  A  R  K
AGTACATGATTTGGCAGGGGATATGCAGTTGTGGACAGAAAACCAAAGTCTTTCTCACTGACAAGACAATGACATCAGAAGTCTACAAAAAAGAGTGCCTACAGAAACGGATTCTGCCGT   840
  Y  M  I  W  Q  G  I  C  S  C  G  Q  K  T  K  V  F  L  T  D  K  T  M  T  S  E  V  Y  K  K  E  C  L  Q  K  R  I  L  P  F
TTATTCGTGCCCACGACCGTCCAGTAATGCTTTGGCCGGACCTCGCAAGCTGCCATTACAGCAAAACGGTTATAGAATGGTACGCAACGAACGGGGTCAGCGTAATACCGAAGGACCTCA   960
  I  R  A  H  D  R  P  V  M  L  W  P (D) L  A  S  C  H  Y  S  K  T  V  I  E  W  Y  A  T  N  G  V  S  V  I  P  K  D  L  N
ACCCCCCAAACTGCCCCCAGTTTCGGCCGATAGAGAAATACTGGGCAATCACGAAGCGGAGGCTTAAGGCAAAGGGAAAACTTGTTAGGAACATGACTCAAATGAAGAACTGGTGGAATC  1080
  P  P  N  C  P  Q  F  R  P  I (E) K  Y  W  A  I  T  K  R  R  L  K  A  K  G  K  L  V  R  N  M  T  Q  M  K  N  W  W  N  Q
AAATCGCCAAAACGGTGGACGAAAAGGGTGTGCGCCGCCTAATGTGCCGTATTACGGGAAAAGTACGAGAATTTCTTCGAAACAGCAATGAATAATTTTTTAATTTTTTTATGAAAGAG  1200
  I  A  K  T  V  D  E  K  G  V  R  R  L  M  C  R  I  T  G  K  V  R  E  F  L  R  N  S  N  E  *
TAAAGAAAATGCTACATTTGTATGAAAAACAAATTATGAATTCGTTAATAAATGACTGAACTACACGCAATTGTTTGTGTTCCAATATTAAATGAAGCAGACTTta                 1306
```

B

```
AY038026   tttttcgggggctaacgctaa ta  AAGTCTGCTTCATTTAATATTGGAACAAATTCTTTTGCT.....CACGCAATTGTTTGTGTTCCAATATTAAATGATGCTGACTT  tacttcacgcgctgtgcctgaca
AY038028   ccacttactttactacgaagttc ta  AAGTCTGCTTCATTTAATATTGGAACAAATTCTTTTGCT.....CACGCAATTGTTTGTGTTCCAATATTAAATGAAGCAGACTT  tacgttaaaacgtatgctatatt
AY038030   tataaatccttgccgggccacag ta  AAGTCTGCTTCATTTAATATTGGAACAAATTATTTTGCT.....CACGCAATTGTTTGTGTTCCAATATTAAATGAAGCAGAC.T  taattatacgggcgcgagaagag
AF377999   atgaatcacccttaatagcagca ta  AAGTCTGCTTCATTTAATCTTGGGACAAATTCTTTTGCT.....CACGCAATTGTTTGTGTTCCAATATTAAATGAAGCAGACTT  taattgcaatgtcatttgaacaa
AY038027   attccgatattagaaacgtgcaa ta  AAGTCTGCTTCATTTAATATTGGAACAAATTCTTTTGCT.....CATGCAATTGTTTGTGTTCCAATATTGAATGAAGCAGACTT  tatattgaaaaaattgcattaaa
AY038029   agagcgagcgttgcaacgcttcc ta  AAGTCTGCTTCATTTAATATTGGAACAAATTCTTTTGCT.....CACGCGATTGTTTGTGTTCCAATATTAAATGAAGCAGACTT  tattctatggaggaaaacttcat
```

Ae.atropalpus.ITmD37E1

Different copies showed 97.0-98.6% nucleotide sequence identity

FIGURE 1.—(A) Nucleotide and deduced amino acid sequence of the consensus of six *ITmD37E* transposons in the rockpool mosquito *Ae. atropalpus* (AF377999 and AY038026–30). The boundaries between the putative TA target site duplication and the transposon are marked by arrows. The TIRs are underlined. The deduced amino acid sequence is shown under the nucleotide sequence. The amino acid residues of the DD37E triad are circled. (B) Multiple sequence alignment of the six *ITmD37E* elements of *Ae. atropalpus* and their flanking sequences. All sequences are identified by their accession numbers shown at the left. The sequences were aligned using ClustalW (THOMPSON *et al.* 1994; gap weight = 10; gap extension weight = 0.05). The majority of the internal sequences of *ITmD37E*, indicated by dots, are not shown. The two-sided arrow indicates the boundaries of *ITmD37E*. Putative TA target duplications are underlined. Nucleotide sequence identities obtained by pairwise comparisons between the six *ITmD37E* transposons are indicated. The identities were calculated using PAUP* 4.0 b8 (SWOFFORD 2001).

TABLE 1

List of transposons analyzed in this study

Element	Motif	Organism	Accession no.
Ae.triseriatus.ITmD37E1	DD37E	*Ae. triseriatus*	AF378001
An.gambiae.ITmD37E1	DD37E	*An. gambiae*	AF378002
Ae.atropalpus.ITmD37E1	DD37E	*Ae. atropalpus*	AF377999
Ae.epactius.ITmD37E1	DD37E	*Ae. epactius*	AF378000
C.elegans.ITmD37D1	DD37D	*C. elegans*	Z93391
C.briggsae.ITmD37D1	DD37D	*C. briggsae*	AC084515
C.briggsae.ITmD37D2	DD37D	*C. briggsae*	AC084524
Bmmar1	DD37D	*Bombyx mori*	U47917
S.peregrina.ITmD37D1	DD37D	*Sarcophaga peregrina*	AB054644
O.sativa.ITmD39D1	DD39D	*Oryza sativa*	AP000836
Soymar1	DD39D	*Glycine max*	AF078934
O.sativa.ITmD39D2	DD39D	*O. sativa*	AF172282
A.thaliana.ITmD39D1	DD39D	*A. thaliana*	AC007662
D.mauritiana.mar1	DD34D	*D. mauritiana*	A26491
M.destructor.mar1	DD34D	*Mayetiola destructor*	U24436
G.tigrina.mar1	DD34D	*Girardia tigrina*	448753
H.sapiens.mar1	DD34D	*Homo sapiens*	U52077
G.palpalis.mar1[a]	AD34D	*Glossina palpalis*	U18308
D.erecta.mar1	DD34D	*D. erecta*	U08094
C.plorabunda.mar1	DD34D	*Chrysoperla plorabunda*	S35520
H.sapiens.mar2	DD34D	*H. sapiens*	U49974
Tc1	DD34E	*C. elegans*	X01005
Quetzal	DD34E	*An. albimanus*	L76231
Tc1-P.platessa	DD34E	*Pleuronectes platessa*	AJ303069
S	DD34E	*D. melanogaster*	U33463
Paris	DD34E	*D. virilis*	U26938
Bari1	DD34E	*D. melanogaster*	S33560
Tc3	DD34E	*C. elegans*	AF025458
MsqTc3	DD34E	*Ae. aegypti*	AF208675
Topi	DD34E	*An. gambiae*	U89799
Minos	DD34E	*D. hydei*	S26856
Impala	DD34E	*Fusarium oxysporum*	AF282722
pogoDm	DD30D	*D. melanogaster*	S20478
Tigger1	DD32D	*H. sapiens*	U49973
Fot1	DD35D	*Fusarium oxysporum*	S20466
Tan1	DD35D	*Aspergillus niger*	U58946
Flipper	DD35D	*Botryotinia fuckeliana*	U74294
TEC1	DD34E	*Euplotes crassus*	L03359
TEC2	DD34E	*E. crassus*	L03360
TBE1	DD34E	*Oxytricha fallax*	L23169
Ant1	DD34E	*Aspergillus niger*	S80872
Hupfer1[b]	DD34E	*Beauveria bassiana*	X84950
IS630Ss	DD35E	*Shigella sonnei*	X05955
IS630Sd	DD35E	*Salmonella dublin*	A43586
H75276	DD35E	*Deinococcus radiodurans*	H75276
IS895[c]	DD43E	*Anabaena sp.*	M67475
RSA	DD33E	*Rhizobium japonicum*	X02581
RIATL	DD35E	*Agrobacterium rhizogenes*	K03313

The three new families of *IS630-Tc1-mariner* transposons are named *ITmD37E, ITmD37D*, and *ITmD39D*, respectively, according to their unique catalytic motifs. The first D of the DDE(D) triad is not included in the name as it is invariable. The naming convention proposed for the *mariner* family (Robertson and Asplund 1996) is used for each of the three families where the host species name precedes the transposon name. Each distinct type of element from a particular host is given a number. Original names are used for previously described transposons *Bmmar1* and *Soymar1*. Accession numbers are from GenBank, EMBL, or PIR. To avoid a long list of references, literature citations that can be identified in database sequence files are not provided. Some of the bacterial transposons were found in plasmids.

[a] The deviation from the DD34D motif is likely caused by neutral mutations of an obviously degenerate copy.

[b] Edited according to Maurer *et al.* (1997).

[c] Edited according to Doak *et al.* (1994).

TABLE 2

Putative target site duplications and terminal inverted repeats

Element	Motif	Target	Terminal inverted repeat[a]
An.gambiae.ITmD37E1	DD37E	ND[b]	ND[b]
Ae.triseriatus.ITmD37E1	DD37E	TA	AAGctTGCTTCRKWTAgWATTGGAACAaa
Ae.atropalpus.ITmD37E1	DD37E	TA	AAGTCTGCTTCATTTAATATTGGAACA
Ae.epactius.ITmD37E1	DD37E	TA	ARGTCTGCTTCATTTAATATTGGAACA
C.elegans.ITmD37D1	DD37D	TA	**CAG**GGTg-aGt**C**-**A**a**AA**t**TAT**gGtAAgt
C.briggsae.ITmD37D1	DD37D	TA	**CAG**GGTc-cGC**C**-**A**a**AA**t**T**AcgTGTGCA
C.briggsae.ITmD37D2	DD37D	TA	**CAG**GGT--GGC**C**g**A**T**AA**g**T**A**T**aGgAcg
Bmmar1	DD37D	TA	**C**ttaGTctGGC**C**-**A**T**AA**a**T**ActGttACAa
S.peregrina.ITmD37D1	DD37D	TA	**CAG**GGT--KGCtg**A**T**A**tm**T**AWtScAACMaa
O.sativa.ITmD39D1	DD39D	TA	CTtCCTCYGTccCAgtAaACgTGWCGtT
Soymar1	DD39D	TA	CTCCCTYCGTTTCAaaATACaTGTCcaTttt...
O.sativa.ITmD39D2	DD39D	TA	CTCCCTCCGTTTCgtttTgttTGTCG
A.thaliana.ITmD39D1	DD39D	ND[b]	ND[b]
Fot1-like consensus	DD35D[c]	TA	AGTnnnnnACnnnnnnnCnGnCCnCnC...
pogo-like consensus	DDxD[c]	TA	CAGnATAnnTCG
Tc1 family consensus	DD34E	TA	**CAG**Tnnnnnn**C**A**AAA**n**T**n**T**nnnnnC[d]
mariner family consensus	DD34D	TA	TnAGGTnGnnnnATAAGT[d]

[a] Conserved nucleotides in each family are shown as uppercase letters. Conserved nucleotides between *Tc1* and DD37D transposons are in boldface type.

[b] The target and TIRs of *An.gambiae.ITmD37E1* and *A.thaliana.ITmD39D1* are not determined because they are not full length.

[c] The five *pogo*-like elements are grouped into two groups that have different terminal repeats. x indicates variable length.

[d] According to Robertson and Asplund (1996).

groupings described above. However, it is not yet clear whether this observation can be applied to *ITmD37E* transposons in other mosquitoes.

Transposons containing conserved DD37D and DD39D motifs: To understand the evolution of the unique *ITmD37E* transposons in the *IS630-Tc1-mariner* superfamily, a database survey that revealed additional diversity was conducted. First, four additional nematode transposons containing a DD37D catalytic motif, similar to a previously described insect *mariner Bmmar1* (Robertson and Asplund 1996), were identified and included in our analyses described below (Tables 1 and 2). Some of these nematode transposons have been noted as members of the *mori* (*Bmmar1*) subfamily of the *mariner* family in a very recent analysis including *Tc1* and different subfamilies of the *mariner* elements (Lampe *et al.* 2001). Similar TIRs (Table 2) and transposases (30.0–47.1% amino acid identities) confirm that these nematode DD37D transposons and *Bmmar1* constitute a distinct group. In addition, a transposon containing a DD39D motif similar to *Soymar1*, a previously described soy bean *mariner* (Jarvik and Lark 1998), was identified in *Arabidopsis thaliana* (Table 1). These transposons and their relatives in rice (Table 1; Tarchini *et al.* 2000) also form a distinct group on the basis of similar TIRs (Table 2) and transposases (42.5–52.7% identities). However, further analyses described below indicate that instead of being divergent subfamilies of the *mariner* family (Robertson and Asplund 1996; Jarvik and Lark 1998; Lampe *et al.* 2001), the DD37D and DD39D transposons are likely two new families distinct from *mariner*. These elements are thus named *ITmD37D* and *ITmD39D*, respectively, using the same naming convention described for *ITmD37E* (Table 1). Like the *ITmD37E* elements, all full-length *ITmD37D* and *ITmD39D* transposons are flanked by putative TA target duplications.

As shown in Figure 2, two of the *ITmD37D* elements, *C.elegans.ITmD37D1* and *C.briggsae.ITmD37D1*, have probably been active very recently because comparisons between different copies within each of the transposons showed >99.5% nucleotide sequence identity. The boundaries of *C.elegans.ITmD37D1* and *C.briggsae.ITmD37D1* were confirmed by multiple sequence alignments of different copies shown in Figure 2, A and B. Evidence of previous mobility of *C.elegans.ITmD37D1* is also presented by the identification of an insertion of *C.elegans.ITmD37D1* in an unknown repetitive sequence (Figure 2C). This also confirmed that the flanking TA nucleotides are indeed the target site duplication.

Comparisons between the *ITmD37E*, *ITmD37D*, and *ITmD39D* transposases and the protein sequence profiles of the *IS630-Tc1-mariner* superfamily: According to BLAST searches, *ITmD37E* and *ITmD37D* are most similar to a Drosophila *Tc1* element *Minos* (Z29098; *E* values are 1×10^{-7} and 1×10^{-27}, respectively), while *ITmD39D* is most similar to a medfly *mariner* (U40493; *E* value, 4×10^{-6}). Systematic analyses have been per-

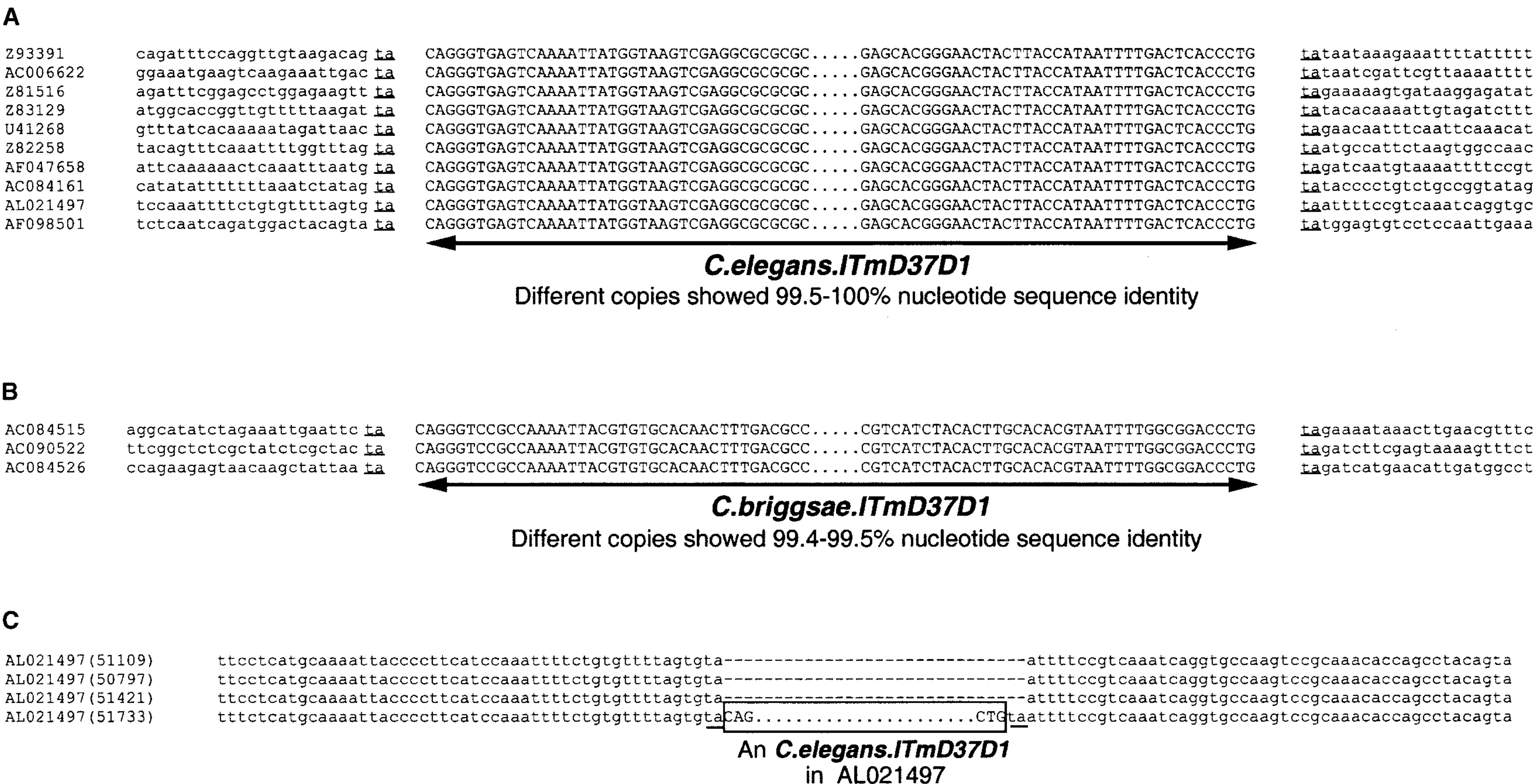


FIGURE 2.—(A) Multiple sequence alignment of the full-length *C.elegans.ITmD37D1* transposons and their flanking sequences. The sequence alignment method and symbols are as described in Figure 1B. (B) Multiple sequence alignment of the full-length *C.briggsae.ITmD37D1* transposons and their flanking sequences. The sequence alignment method and symbols are as described in Figure 1B. (C) Evidence of past mobility of *C.elegans.ITmD37D1*. Shown at the bottom is a repetitive sequence that contains an insertion of a *C.elegans.ITmD37D1*. The top sequences correspond to an unknown repetitive element. These repeats were found in the same cosmid sequence. The numbers in parentheses indicate the positions of the starting nucleotide. Dashed lines indicate gaps. Dots represent nucleotides that are not shown. *C.elegansITmD37D1* is shown in the open box. The TA target duplication flanking the *C.elegansITmD37D1* is underlined. The consensus of the nematode transposons has been used in a recent phylogenetic analysis (LAMPE *et al.* 2001). A–C establish the common characteristics of these transposons (*e.g.*, TIRs and TA target duplications), the high sequence identities between different copies, and the evidence of past mobility.

TABLE 3

Profilesearch scores (*Z* scores) between different transposons and the profiles of various families in the *IS630-Tc1-mariner* superfamily

Family	*Z* scores obtained using the following profiles			
	IS630-Tc1-mariner	*Tc1*	*mariner*	*IS630*
ITmD37E	6.8–8.6	9.6–10.9	<2.5	4.0–4.9
ITmD37D	13.5–17.5	11.7–19.5	7.0–12.4	4.2–6.2
ITmD39D	8.1–10.8	5.3–7.5	5.3–8.1	4.7–7.7
Tc1	18.8–25.4[b]	42.3–54.9[b]	6.6–13.3	8.0–11.6
mariner	15.3–20.8[b]	10.7–13.9	67.5–74.8[b]	<2.5–5.1
pogo[a]	<2.5–3.55	<2.5	<2.5–3.82	<2.5–3.35
IS630	6.1–16.9[b]	<2.5–16.2	<2.5–5.9	25.6–53.3[b]

Profilesearch scores (*Z* scores) were obtained using Profilesearch (gap weight = 10; gap length weight = 0.1) of GCG (version 10, 1999). Profiles used include *IS630-Tc1-mariner* (all previously known members of the *IS630-Tc1-mariner* superfamily shown in Figure 3, *Tc1, mariner*, and *IS630* (all six prokaryotic *IS630*-like elements shown in Figure 3). The alignment used to generate the profiles were constructed using ClustalW (Thompson *et al.* 1994) with the following parameters: gap weight = 10 (*IS630*-like elements) or 5 (the other three groups), and gap length weight = 0.05. The database used for the profilesearch is the Swiss-Prot database plus the proteins being studied. That a sequence gives a *Z* score of 3 indicates that the similarity between this sequence and the search profile is 3 standard deviations above the average scores of all sequences in the database (Doak *et al.* 1994).

[a] Although *Z* scores of the comparisons between full-length *pogo* and the four profiles are relatively low, the catalytic domains of *pogo* elements have been shown to be related to the *IS630-Tc1-mariner* elements (Smit and Riggs 1996; Capy *et al.* 1997). In our analysis, the *Z* scores of *pogo* elements are higher (up to 4.64) when comparing the catalytic domains only. Under less stringent Profilesearch conditions, the comparison between *pogo* and the *IS630-Tc1-mariner* profile gave *Z* scores >7.5 (Smit and Riggs 1996).

[b] *Z* scores from matches between elements and a profile that contains the elements.

formed between individual sequences of the three new groups and the profiles of the *IS630-Tc1-mariner* superfamily and its member families, using a Profilesearch approach described by Doak *et al.* (1994). Comparisons between the profile of the *IS630-Tc1-mariner* superfamily and the *ITmD37E*, *ITmD37D*, and *ITmD39D* elements gave *Z* scores from 6.8 to 17.5, which are within the range of scores of comparisons between the superfamily profile and known *IS630-Tc1-mariner* elements (Table 3). These scores are all higher than the scores of the comparisons between different profiles and *pogo* transposons (Table 3), which are possibly divergent members of the *IS630-Tc1-mariner* superfamily (Capy *et al.* 1996, 1997; Smit and Riggs 1996). The profile analysis also indicated that both *ITmD37E* and *ITmD37D* are probably more similar to *Tc1* than to *mariner* or *IS630* (Table 3). It is clear that all of the three new groups are members of the *IS630-Tc1-mariner* superfamily.

***ITmD37E*, *ITmD37D*, and *ITmD39D* transposons form three distinct families in the *IS630-Tc1-mariner* superfamily:** As described above, *ITmD37E*, *ITmD37D*, and *ITmD39D* transposons form their respective groups on the basis of similar transposases and relatively conserved TIRs. Phylogenetic analysis on the basis of the full-length alignment of representatives of all known *IS630-Tc1-mariner* transposons (Figure 3) strongly supports the idea that *ITmD37E*, *ITmD37D*, and *ITmD39D* form three clades distinct from *Tc1*, *mariner*, *Ant1-Hupfer1*, ciliate transposons, and the *IS630*-like elements (Figure 4A). The bootstrap values obtained using three different methods for the *ITmD37E*, *ITmD37D*, and *ITmD39D* clades are all >96%. While *Tc1* (DD34E) and *mariner* (DD34D) also form their respective clades well supported by bootstrap analysis, *Ant1-Hupfer1* (DD34E), ciliate transposons (DD34E), and the *IS630*-like elements (DDxE) form a less robust clade (the DDxE clade and clade VI) as indicated by low bootstrap values. All of the above clades except DDxE are also supported by bootstrap when *pogo* elements are included in an analysis based on the more conserved catalytic domains (Figure 4B). To study the relationship between the different clades, these two trees are rooted using the prokaryotic *IS* elements. In both trees, *ITmD37E* is a distinct clade. The status of *ITmD37E* is also reflected by its unique TIRs. As shown in Figure 4A, *ITmD37D* is grouped with *Tc1* instead of *mariner*. This relationship is supported by bootstrap in both neighbor-joining and minimum evolution analysis. *ITmD37D* and *Tc1* are also grouped together in the most parsimonious trees (data not shown), although the bootstrap value derived from the parsimony analysis is <50%. In addition, the *ITmD37D* and *Tc1* grouping is supported by bootstrap analyses using all three methods in the tree, including the *pogo* elements (Figure 4B). The relationship between *ITmD37D* and *Tc1* is also consistent with the similarity between their TIRs (Table 2). Thus we propose that the *ITmD37D* transposons, including *Bmmar1*, are not a basal subfamily of *mariner* (Robertson and Asplund 1996). Instead,

```
                                                   A number of amino    Extra 5 amino     Extra 3 amino acids
                                                   acids not shown      acids in DD39D    in DD37E, DD37D                  Distribution
DD37E
   An.gambiae.ITmD37E1   dgclLmd DEtYvkaDf   Vmfwp DlASc HySKv.....VREWyaE.........kgVlfvpknlnPPncPqfrPI Ek.yWAIM
 Ae.triseriatus.ITmD37E1 dgclLmd DEtYvkaDf   Vmfwp DlASc HySKv.....VqEWyaE.........kgVQfvpkhlnPPkcPqfrPI Ek.yWAIM       Mosquitoes
  Ae.atropalpus.ITmD37E1 dgcILmd DEtYvkaEf   VmLwp DlASc HySKt.....ViEWyat.........ngVsvIpkdlnPPncPqfrPI Ek.yWAIt
    Ae.epactius.ITmD37E1 dgcILmd DEtYvkaDf   VmLwp DlASc HySKt.....ViEWyat.........ngVsvIpkdlnPPncPqfrPm Ek.yWvIt
DD37D
     C.elegans.ITmD37D1  hrKVLFt DEKiFcIEq   wtfQQ DgApa HKhKn.....VqaWyEs.........nfpDfIafnqWPPsSPDLNPm DYsvWsVL
   C.briggsae.ITmD37D1   hlitIFs DEKlFsvEa   yIfQQ DgApa HtaKm.....aqQWckD.........hfpafIpkdeWPasSPDLNPL DYsvWgVL      Nematodes
                 Bmmar1  hrKILFt DEnfFtIEq   wsfQQ DsApg HKaRs.....tqsWLEt.........nvsDfIraedWPssSPDLNPL DYdLWsVL      and Insects
  S.peregrina.ITmD37D1   vtefgFp DkKpFqIkq   wtfQQ DsApp HskRi.....nqEWIkk.........evsrfIstaqWPPkSPDLNPL DFcaWgfL
   C.briggsae.ITmD37D2   vsdVIwt DEKiFtIEp   fILQQ DwAps HgSRs.....tlavLEa.........hfpgfLdknlWPasSPDLNPm DFsvWgmL
DD39D
                Soymar1  mynIIhI DEKWFyMtk   IfIQQ DNArt HinpDdpefvqaatqDg.........fdIrLMc...qPPnSPDfNvL DlgffsaI
   A.thaliana.ITmD39D1   mynVVhI DEKWFyMtr   IfIQQ DNvrt HvdlrdeefrVvashhg.........fdIhLMc...qPPnlPDLNiL DlgffsaI      Flowering
     O.sativa.ITmD39D1   menIIhI DEKWFNask   IwIQQ DNArt HltiDdaqfgVavaqtg.........LdIrLvn...qPPnSPDMNcL DlgffAsL      Plants
     O.sativa.ITmD39D2   lfdfIFI DEKWFNItr   IyIQQ DNArp HiapDdrmfceaakqDg.........fnIkLvc...qPanSPDLNvL DlgffnsI
mariner: DD34D
      D.mauritiana.mar1  lhRIVtg DEKWiffvs   VIflh DNAps HtaRa.....VRdtLEt.........LnwEvLp...haaYSPDLaPs DYhLfAsM
      M.destructor.mar1  lsRIItg DEKWihyDn   VIfhh DNArp Hvalp.....VKnYLEn.........sgwEvLp...hPPYSPDLaPs DYhLfrsM
         G.tigrina.mar1  lhRIVtc DEKWimyDn   pILlh DNArp HsaKn.....tvakLQQ.........LglEtLr...hPPYSPDLaPt DYhffqsL      Flatworm,
         H.sapiens.mar1  ldRIVtc DEKWilyDn   pILlh DNArp Hvaqp.....tlQkLnE.........LgyEvLp...hPPYSPDLsPt DYhffkhL      Nematodes,
        G.palpalis.mar1  lkRmVtea EKWityDn   IVfhQ DNsrp HtStv.....tRQkLrE.........LgwEvLm...hPPYSPDLaPs DYhLflaL      Insects and
          D.erecta.mar1  lkRmVtg DEKWityDn   VVfhQ DNArp HtSlm.....tRQkLrE.........LgwEvLs...hPPYSPDIaPs DYhLflsM      Vertebrates
      C.plorabunda.mar1  lrRyVtm DEtWlhhyt   .LfhQ DNApc HKSlr.....tmakIhE.........LgfELLp...hPPYSPDLaPs DFfLfsdL
         H.sapiens.mar2  lrRIVtg DEtWlyqyd   .LLhh DNApa HsShQ.....tRaiLrE.........fRwEIIr...hPPYSPDLaPs DFfLfpnL
Tc1: DD34E
                    Tc1  WaKhIws DEskFNLfG   fVfQQ DNdpk HtSlh.....VRsWfQr.........rhVhLLd...WPsqSPDLNPI EH.LWeeL
                Quetzal  WkKVLFt DEskFNIfG   ywfQQ DNdpk Htafn.....sRlfLly.........ntphqLk...sPPqSPDLNPI EH.aWeLL
          Tc1-P.platessa WkKVLws DEtkieLfG   fIfQQ DNdpk HtaRa.....tKEWfgl.........knVnvLk...WPsqSPDLNPI En.LWqdL
                      S  WddVIFc DEtkmmLfy   fkfyQ DNdpk HKeyn.....VRnWLly.........ncgkvId...tPPqSPDLNPI En.LWAyL      Nematodes,
                  Paris  WddVIFc DEtkimLyy   ykLyQ DNdpk HKSfl.....cRtWLly.........ncskvId...tPaqSPDLNPI En.LWAfL      Insects,
                  Baril  WfnILwt DEsaFqyqG   wILQQ DNApc HKgRi.....ptkfLnD.........LnlavLp...WPPqSPDLNiI En.vWAfI      Vertebrates,
                    Tc3  WsKVVFs DEKkFNLDG   frfQQ DNAtI HvSns.....tRdYfkl.........kKInLLd...WParSPDLNPI En.LWgIL      and Fungi
                 MsqTc3  WtmmIFs DEKkFNLDG   ftfQQ DNAaI HtSKE.....tKQWIkD.........hKIDLLd...WParSPDLNPV En.LWgIL
                   Topi  WsKIIFs DEsriNLDG   yIfQh DNdSk HtSRt.....VKcYLan.........qdVQvLp...WPalSPDLNPI En.LWstL
                  Minos  .dtIIFs DEakFdvsv   ftfQQ DgASs HtaKr.....tKnWLQy.........nqmEvLd...WPsnSPDLsPI En.IWwLM
                 Impala  WrRVkws DEcmvrrgq   dIfmh DNASV HtaRi.....VKalLEE.........LgVDLMt...WPPYSPDLNPI En.LWALM
TEC1: DD34E
                   TEC1  ...VVYI DEcsFN.rs   tIyvf DNASI Hltqk.....VvkcvtD.........rKmvcft...iPPYcPELNkV EH.tfgLL
                   TEC2  ...IVYI DEcsFN.as   tVyvf DNASI Hstek.....VvkaItg.........MKmvcft...iPPYSPELNkI EH.tfgtL      Ciliates
                   Ant1  ....LFV DEsgcdrri   sVIvm DNASf HhSek.....IeElcsQ.........agVkIvY...lPPYSPDLNPI Ee.ffseL      and Fungi
                Hupfer1  ....VFV DEsgcdkra   sVIvm DNASw HhSdE.....lRQmcED.........agVkLMY...lPPYSPDfNPI Ee.ffsVL
                   TBE1  .iKfIha DEavFtfst   IILfv DNlSV HKtKE.....tKksyEQ.........LRItpvF...nvPYSPqfNgI EF.yWgIL
IS630: DDxE
                IS630Ss  ...VfYe DEvd.ihln   ItLiv DNyiI HKSRE.....tqsWLkEn........pKfrvIY...qPvYSPwvNhV Er.LWqaL
                IS630Sd  ...VfYq DEvd.idln   ItLva DNyiI HKSRk.....VerWLEEn........pKfrLLF...lPmYSPwLNPI Er.LWlsL
                    RSA  ...VLcV DEKpsiqal   Lhvil DNlnt HK..k.....nedWLkah........pnVQfhF...tPtsaPwLNqV Ev.wfsIL      Prokaryotes
                  RIATL  ...VLsV DEKsqiqal   Vhvil DNyat HKqpk.....VRaWLarh........pRwtfhF...vPtscswLNaV Eg.ffAkL
                 H75276  ...lVYL DEvgFsLkG   LVvvl DNAgI HRSKa.....tqafvEth........eRlsLvF...lPPYaPELNPI El.vWAyv
                  IS895  ...VLlI DEchlmwgd   LLIfw DgASy HRSKE.....IRgfLDsvnqslpteqwKIhcvr...faPncPvqNPI Ed.IWlqa
              Consensus  W-KVIFI DEKWFNLDG   IILQQ DNASI HKSKE-----VRQWLEE---------LKVELLF---WPPYSPDLNPI EY-LWAIL
                          ↑                        ↑                                                         ↑
```

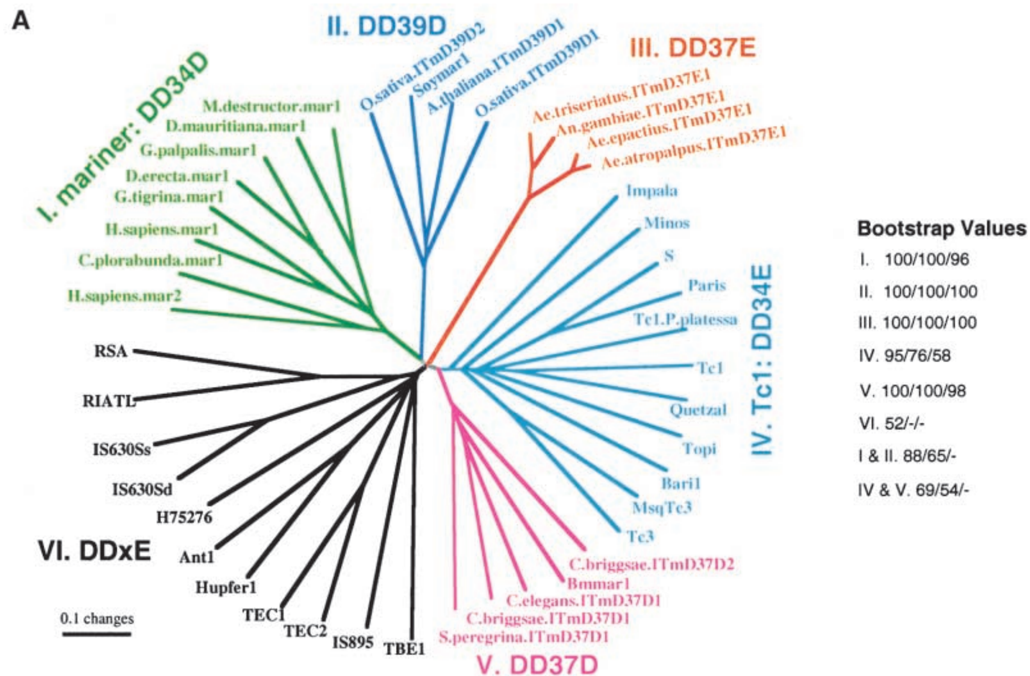
FIGURE 3.—Multiple sequence alignment of *ITmD37E*, *ITmD37D*, and *ITmD39D* transposons and representatives of other families of the *IS630-Tc1-mariner* transposons (DOAK *et al.* 1994; CAPY *et al.* 1996, 1997; ROBERTSON and ASPLUND 1996). The accession numbers of all sequences are listed in Table 1. All transposase sequences were aligned using ClustalW (THOMPSON *et al.* 1994; gap weight = 5; gap extension weight = 0.05). The full alignment has been deposited in an EMBL alignment database (DS47334). Only regions surrounding the DDE(D) catalytic triad are shown here. The DDE(D) triad and other invariable residues are in boldface type while the triad is also marked by arrows. Note the change from D to A in *G.palpalis.mar1*, which is a defective element (ROBERTSON and ASPLUND 1996). *pogo* transposons were not included in this alignment because they cannot be reliably aligned with the *IS630-Tc1-mariner* elements when the entire transposase sequences are included. However, the more conserved catalytic domains of *pogo* have been successfully aligned to the catalytic domains of *IS630-Tc1-mariner* elements, which are described in the legend of Figure 4B. A highly defective *Tc1*-like transposon in the Pacific hagfish *Eptatretus stouti* named *Tes1* was not included in the alignment because the N terminus cannot be reliably aligned and there is no alignable stop codon at the C terminus as indicated by ROBERTSON (1995). Instead of having a DD34E motif as do the rest of *Tc1* transposons, the defective *Tes1* element contains a DD38E catalytic triad, which may be due to neutral mutations after the inactivation of this transposon.

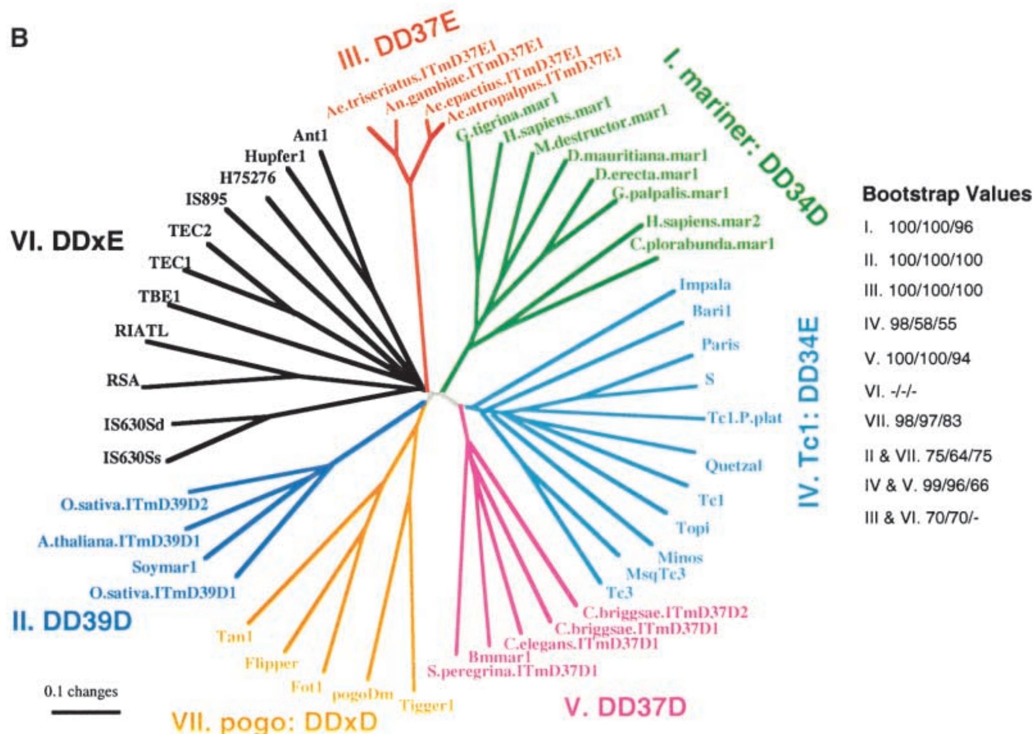
they are a distinct family related to the *Tc1* transposons. The relationship between *ITmD39D* and other transposons is less certain. Although *ITmD39D* is closer to *mariner* as shown in Figure 4A, the relationship was not supported by parsimony analysis. Moreover, as shown in the tree including the *pogo* elements (Figure 4B), the grouping between *ITmD39D* and *mariner* collapsed and *ITmD39D* is related to *pogo* instead. Moreover, the TIRs of *ITmD39D* are completely different from the TIRs of *mariner*. In light of the uncertainty of the phylogenetic relationship between *ITmD39D* and *mariner* transposases and their different catalytic motifs and TIRs, we propose that *ITmD39D* may also be considered as a distinct family.

DISCUSSION

Expansion and reorganization of the *IS630-Tc1-mariner* superfamily on the basis of conserved catalytic motifs: In this study we reported the discovery of *ITmD37E*, a novel family of transposons in the *IS630-Tc1-mariner* superfamily, which is characterized by its unique DD37E catalytic motif and TIRs. We also identified two additional families, *ITmD37D* and *ITmD39D*, which contain

their respective conserved DD37D and DD39D catalytic motifs and TIRs (Table 2 and Figures 3 and 4). The classification of *ITmD37D* and *ITmD39D* as two distinct families represents a revision of earlier studies that reported the original discoveries of the founding members of the two families *Bmmar1* (DD37D; Robertson and Asplund 1996) and *Soymar1* (DD39D; Jarvik and Lark 1998). It was suggested that *Bmmar1* and *Soymar1* may be subfamilies of *mariner*, although the distinction between these elements and the rest of the *mariner* elements was clearly noted (Robertson and Asplund 1996; Jarvik and Lark 1998). According to the analysis presented here, the *IS630-Tc1-mariner* superfamily can be organized in six families that include *ITmD37E*, *ITmD37D*, *ITmD39D*, *Tc1*, *mariner*, and *pogo* and an unresolved clade VI that includes bacterial *IS630*-like ele-





ments and some fungal and ciliate transposons (Figure 4). Further determination of the status and relationship of the transposons in clade VI may require accumulation of more related transposon sequences in a diverse range of organisms. The above classification is supported by bootstrap analysis with or without *pogo* elements (Figure 4). *pogo* is an interesting case as it has a unique N-terminal DNA-binding domain and a long C-terminal domain rich in acidic residues, although its catalytic domain is related to *IS630-Tc1-mariner* transposons (Smit and Riggs 1996). In addition, the distance between the last two D's of the catalytic triad of the *pogo* family is variable. The classification of *Tc1*, *mariner*, and the three new families based on phylogenetic analysis is consistent with the grouping by their respective catalytic triad. It should be noted, however, not all DD34E transposons belong to the *Tc1* family and phylogenetic evidence always needs to be included in the classification of a new transposon. The close relationship between *ITmD37D* and *Tc1* suggests a more complex evolutionary process from their common ancestor to these two current motifs, which may require more than a simple change from a D to an E in the catalytic triad or vice versa. This is consistent with the result that a point mutation from DD34D to DD34E abolished the transposase activity of a *mariner* in *in vivo* excision assays (Lohe *et al.* 1997). Although a few other interfamily relationships were indicated, none was supported in both phylogenetic trees (Figure 4). Finally, the diversity of the catalytic motifs and their conservation within each family raised an interesting question about the structure-function relationship of these distinct, yet related, transposases. Answers to such a biophysical question may help illustrate the evolutionary process that leads to the expansion of such a diverse group of transposons in the *IS630-Tc1-mariner* superfamily.

Distribution and evolutionary implications: *ITmD37E* transposons have been found in all 13 species in five genera of mosquitoes examined, including the Aedes and Anopheles genera. Although the evolutionary distance between Aedes and Anopheles is among the longest in mosquitoes (Isoe 2000), *Ae.triseriatus.ITmD37E1* is more closely related to *An.gambiae.ITmD37E1* than the *ITmD37E* transposons in two other Aedes species, *Ae.atropalpus.ITmD37E1* and *Ae.epactius.ITmD37E1* (Figure 4). We suggest that this may reflect the existence of two subfamilies because our preliminary data suggest the existence of a second *ITmD37E* transposon more closely related to *Ae.triseriatus.ITmD37E1* in both *Ae. atropalpus* and *Ae. epactius* (data not shown). It remains to be determined whether the relatively high sequence similarity between *Ae.triseriatus.ITmD37E1* and *An.gambiae.ITmD37E1* reflects high selection pressure on the transposase proteins or possible horizontal transfer events. Such questions may be addressed by a systematic survey of the *ITmD37E* transposons in a wide range of mosquitoes, which may shed light on the evolutionary dynamics of these transposons in this medically important insect family. The independent family status of *ITmD37E* in a superfamily that has a broad host range suggests that *ITmD37E* is likely a family of ancient origin. It may be reasonable to expect a broad distribution of *ITmD37E* in other insects and perhaps other invertebrates. Although no similar transposons have been found in the *Drosophila melanogaster* genome database, we cannot exclude their existence in *D. melanogaster* because a large fraction of the repeat-rich regions has not been sequenced (Myers *et al.* 2000). In addition, given the nature of transposon-host interaction, we do not expect ubiquitous distribution of transposons in a particular taxonomic group. According to the current compilation, the *ITmD39D* family is limited to the flowering plants while the *ITmD37D* family is found in both nematodes and insects, two highly divergent invertebrate groups. As data from genomic analysis accumulate, this distribution could be expanded. Nonetheless, the identification of related transposons in *ITmD37D* and *ITmD39D* families in diverse organisms is opening the door for a broad survey based on the identification of conserved amino acid residues that can be used to design primers for PCR (Robertson and MacLeod 1993). In fact, stretches of conserved amino acid residues do exist in both families (Figure 3). Evolution of *Tc1* and *mariner* has been a topic of extensive studies including

Figure 4.—(A) Phylogenetic relationship between *ITmD37E*, *ITmD37D*, and *ITmD39D* transposons and representatives of other families of the *IS630-Tc1-mariner* transposons. The tree shown here is an unrooted phylogram constructed using a minimum evolution algorithm based on the full alignment described in Figure 3. Two additional methods, neighbor-joining and maximum parsimony, were also used. Confidence of the groupings was estimated using 500 bootstrap replications. The bootstrap values represent the percentage of times out of 500 bootstrap resamplings that branches were grouped together at a particular node. The first, second, and third numbers represent the bootstrap values derived from minimum evolution, neighbor-joining, and maximum-parsimony analyses, respectively. Only the values for major groupings are shown. Various colors indicate different clades. All phylogenetic analyses were conducted using PAUP 4.0 b8 (Swofford 2001). Detailed methods are described in Tu and Hill (1999). (B) Phylogenetic relationship between different families of the *IS630-Tc1-mariner* transposons, including *pogo*, on the basis of the catalytic domains. The alignment used here was obtained using the catalytic domains of *pogo* and transposons shown in Figures 3 and 4A. Although the precise boundaries of the catalytic domains have not been clearly defined, we used the C-terminal half of most of the transposases starting 20–30 amino acid residues upstream of the first D of the DDE(D) triad. The alignment method was the same as that described in Figure 3. All symbols and phylogenetic analysis methods are the same as those described in A.

their regulation, horizontal transmission, and interactions with the host genomes. It is foreseeable that the rapid accumulation of genomic sequence data from a wide range of organisms in combination with the deliberate PCR and genomic survey will greatly facilitate the discovery of many more transposons in the three distinct families described here. Such expansion will provide wonderful opportunities in a wide range of organisms to study the evolutionary dynamics of these individual families as well as the *IS630-Tc1-mariner* superfamily in general.

Potential applications: The identification of a widespread DNA transposon in mosquitoes may have potentially important applications. Transformation tools are being developed for the genetic manipulation of mosquitoes. Such tools are critical components of the genetic strategy to control mosquito-transmitted diseases by creating disease-resistant mosquitoes through the introduction of refractory genes using DNA transposons (Ashburner *et al.* 1998). A few mosquito species have been successfully transformed using exogenous DNA transposons including *mariner* and *Tc1*-like transposons (Atkinson *et al.* 2001). Although no evidence of active transposition has been obtained for *ITmD37E*, the relatively high sequence similarity within some of the species and the identification of intact ORFs suggest that it may be a worthy effort to identify or to screen for active *ITmD37E* transposons, using methods described by Lampe *et al.* (1999). The seemingly ubiquitous distribution of *ITmD37E* in mosquitoes indicates its potentially broad utility as a transformation tool once active elements are obtained. Although interaction between endogenous elements and the *ITmD37E* vector may be a potential problem, a recent study on the lack of interaction between relatively closely related *mariner* elements (Lampe *et al.* 2001), and the fact that different types of *ITmD37E* can coexist in the same mosquito species, indicate that the problem can be mitigated. Multiple subfamilies of relatively closely related *P* elements have also been found to coexist in a few Drosophila species (Clark and Kidwell 1997). In addition to the potential for the direct application described above, studies of these mosquito transposons may help understand their regulation, long-term activity, and spread in mosquito populations, which will be important for the long-term success of the genetic strategy to control mosquito-transmitted diseases.

Two other transposons described in this study also showed promise as candidates for the development of transformation tools. Ten different copies of *C.elegans. ITmD37D1* showed 99.3–100% identity while three different copies of *C.briggsae.ITmD37D1* showed 99.4–99.5% identity. This level of sequence identity is similar to the 99.6–100% identity that we found for the 19 full-length copies of a currently active *Caenorhabditis elegans* transposon *Tc3* (Collins *et al.* 1989). Thus it is possible that these two *ITmD37D* transposons are either still active or at least recently active. Given the possible broad distribution of *ITmD37D* transposons in nematodes and insects, transformation tools developed on the basis of the above two transposons have the potential to be widely applicable. Continued discovery and analysis of transposons in the three distinct families described here may provide additional transformation systems to the current repertoire of transposon-based broad-range transgenesis and mutagenesis tools, a large fraction of which has been derived from *Tc1* and *mariner* transposons (*e.g.*, Gueiros-Filho and Beverley 1997; Raz *et al.* 1997; Izsvak *et al.* 2000).

We thank Pierre Capy and Hugh Robertson for insightful comments. This work was supported by a National Institutes of Health grant AI-42121 to Z. Tu and by the Agricultural Experimental Station at Virginia Tech.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

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

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

Besansky, N. J., and J. R. Powell, 1992 Reassociation kinetics of *Anopheles gambiae* (Diptera: Culicidae) DNA. J. Med. Entomol. **29:** 125–128.

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

Capy, P., T. Langin, D. Higuet, P. Maurer and C. Bazin, 1997 Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica **100:** 63–72.

Clark, J. B., and M. G. Kidwell, 1997 A phylogenetic perspective on *P* transposable element evolution in *Drosophila*. Proc. Natl. Acad. Sci. USA **94:** 11428–11433.

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

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

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

Felsenstein, J., and H. Kishino, 1993 Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. **42:** 193–200.

Gueiros-Filho, F. J., and S. M. Beverley, 1997 Trans-kingdom transposition of the *Drosophila* element *mariner* within the protozoan *Leishmania*. Science **276:** 1716–1719.

Henikoff, S., 1992 Detection of *Caenorhabditis* transposon homologs in diverse organisms. New Biol. **4:** 382–388.

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

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

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

Kidwell, M. G., and D. M. Lisch, 2000 Transposable elements and host genome evolution. Trends Ecol. Evol. **15:** 95–99.

Lampe, D. J., B. J. Akerley, E. J. Rubin, J. J. Mekalanos and H. M.

Robertson, 1999 Hyperactive transposase mutants of the *Himar1 mariner* transposon. Proc. Natl. Acad. Sci. USA **96:** 11428–11433.

Lampe, D. J., K. K. Walden and H. M. Robertson, 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.

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

Maurer, P., A. Rejasse, P. Capy, T. Langin and G. Riba, 1997 Isolation of the transposable element *hupfer* from the entomopathogenic fungus *Beauveria bassiana* by insertion mutagenesis of the nitrate reductase structural gene. Mol. Gen. Genet. **256:** 195–202.

Moore, S. P., M. Powers and D. J. Garfinkel, 1995 Substrate specificity of Ty1 integrase. J. Virol. **69:** 4683–4692.

Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo *et al.*, 2000 A whole-genome assembly of *Drosophila*. Science **287:** 2196–2204.

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

Rai, K. S., and W. C. I. V. Black, 1999 Mosquito genomes: structure, organization, and evolution. Adv. Genet. **41:** 1–33.

Raz, E., H. G. A. M. Van Luenen, B. Schaerringer, R. H. A. Plasterk and W. Driever, 1997 Transposition of the nematode *Caenorhabditis elegans Tc3* element in the zebrafish *Danio rerio*. Curr. Biol. **8:** 82–88.

Robertson, H., 1995 The *Tc1-mariner* superfamily of transposons in animals. J. Insect Physiol. **41:** 99–105.

Robertson, H. M., and M. L. Asplund, 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 E. G. Macleod, 1993 Five major subfamilies of *mariner* transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect Mol. Biol. **2:** 125–139.

Smit, A. F. A., and A. D. Riggs, 1996 *Tiggers* and other DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA **93:** 1443–1448.

Swofford, D. L., 2001 PAUP*, version 4.0 b8 (a commercial test version; completed version 4.0 to be distributed by Sinauer, Sunderland, MA).

Tarchini, R., P. Biddle, R. Wineland, S. Tingey and A. Rafalski, 2000 The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. Plant Cell **12:** 381–391.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTALW: 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.

Tu, Z., 2000 Molecular and evolutionary analysis of two divergent subfamilies of a novel miniature inverted repeat transposable element in the yellow fever mosquito, *Aedes aegypti*. Mol. Biol. Evol. **17:** 1313–1325.

Tu, Z., and H. H. Hagedorn, 1997 Biochemical, molecular, and phylogenetic analysis of pyruvate carboxylase in the yellow fever mosquito, *Aedes aegypti*. Insect Biochem. Mol. Biol. **27:** 133–147.

Tu, Z., and J. J. Hill, 1999 *MosquI*, a novel family of mosquito retrotransposons distantly related to the *Drosophila* I factors, may consist of elements of more than one origin. Mol. Biol. Evol. **16:** 1675–1686.

Communicating editor: H. Ochman