Published in final edited form as:
Genetica. 2017 June ; 145(3): 275–293. doi:10.1007/s10709-017-9964-z.

Transposable elements in the *Anopheles funestus* transcriptome

Rita D. Fernández-Medina[1,*], **Claudia M.A. Carareto**[2], **Cláudio J. Struchiner**[1], and **José M.C. Ribeiro**[3]

[1]Fundação Oswaldo Cruz, Escola Nacional de Saúde Pública. Av. Brasil, 4365. Rio de Janeiro, Brasil

[2]UNESP- Universidade Estadual Paulista, Departamento de Biologia, Rua Cristóvão Colombo, 2265, São José do Rio Preto, SP, Brasil

[3]Laboratory of Malaria and Vector Research, NIAID/NIH, Rockville, Maryland 20852, USA

Abstract

Transposable elements (TEs) are present in most of the eukaryotic genomes and their impact on genome evolution is increasingly recognized. Although there is extensive information on the TEs present in several eukaryotic genomes, less is known about the expression of these elements at the transcriptome level. Here we present a detailed analysis regarding the expression of TEs in *A. funestus,* the second most important vector of human malaria in Africa. Several transcriptionally active TE families belonging both to Class I and II were identified and characterized. Interestingly, we have identified a full-length putative active element (including the presence of full length TIRs in the genomic sequence) belonging to the *hAT* superfamily, which presents active members in other insect genomes. This work contributes to a comprehensive understanding of the landscape of transposable elements in *A. funestus* transcriptome. Our results reveal that TEs are abundant and diverse in the mosquito and that most of the TE families found in the genome are represented in the mosquito transcriptome, a fact that could indicate activity of these elements. The vast diversity of TEs expressed in *A. funestus* suggests that there is ongoing amplification of several families in this organism.

Keywords

Anopheles funestus; transposable elements; transcriptome

1. Background

Transposable Elements (TEs) are abundant and ancient genetic sequences present in all eukaryotic genomes showing the ability of transposing between different loci. The distribution and abundance of TEs within and between different genomes varies widely,

[*]Corresponding Author: RDFM, dfernandezmedina@gmail.com, Telephone: +5521 97151-3464.

Competing interests: The authors declare no competing financial interests.

Author's contributions: Conceived the study: R.D.F.M, C.M.A.C., J.M.R. and C.J.S Analyzed the data and wrote the paper: R.D.F.M All authors read and approved the final manuscript.

constituting the majority of the DNA content in some species, while in others they represent just a small fraction of the total genomic DNA (Bennetzen and Devos 2005; de Koning *et al.* 2011; Kidwell 2002). In insects, for instance, genome sizes vary from less than 100 Megabases (Mb) to more than 10 Gigabases (Gb) (Gregory *et al.* 2007). The causes and consequences of these differences - sometimes in organisms with very similar degrees of complexity - are yet unknown. However, much of this variation reflects different evolutionary dynamics of TE families present in different genomic contexts and shows the enormous impact that these elements might have had in eukaryotic genomes' evolution.

Although far from a consensus for a universal TE classification system (Piégu et al. 2015), these elements have been classified based on their genetic structure and mode of replication into two main classes - Class I, or retrotransposons, and Class II, or DNA transposons (Finnegan 1989), and further hierarchically into orders, superfamilies, families and subfamilies by Wicker et al. (2007). Accordingly, Class I is composed of five orders: the LTRs, DIRS and Penelope-Like elements and the LINEs and SINEs, or Non-LTRs. These elements replicate through a reverse transcription step of an intermediary RNA and produce new copies in each cycle of replication while Class II transpose directly as a DNA molecule. The last are further divided into two subclasses, subclass 1 the classical ‘cut-and-paste’ elements (characterized by their terminal inverted repeats), and subclass 2 characterized by a transposition process without double-stranded cleavage (Orders Helitron and Polintons/ Maverick). Both Class I and Class II are further classified into several superfamilies, distinguished by large-scale features, such as the structure of protein and noncoding domains, the presence and size of target site duplications (TSD). Superfamilies are further classified into families (also named clades or lineages) defined by DNA sequence conservation and further into subfamilies defined by phylogenetic relationships (Wicker et al., 2007) as well as by the degree of identity among their sequences.

An archetypal TE family can be composed of elements with different degree of activity: some elements having coding capacity, called autonomous elements, and others with inactivating mutations that can still harbor the ability of being mobilized by active and autonomous counterparts, known as non-autonomous elements. Most of the TEs in the present-day genomes are inactive, deteriorated or methylated, a fact that has been related to the evolution of controlling mechanism acting on the TE's mobilization. Probably, due to their mutagenic activity, eukaryote genomes have developed efficient mechanisms to silence them. Inactive elements populate genomes and evolve neutrally until eventually being lost from the genome. TEs can also be co-opted by their host genome; domestication of TE-derived protein coding and regulatory modules has indeed taken place repeatedly in the course of eukaryotic genome evolution (Miller et al, 1997, Casacuberta and Pardue, 2005, Kapitonov and Kunin, 2015).

Active elements have been reported in many genomes including mosquitoes – as *Herves* in *Anopheles* (Arensburguer *et al.* 2005), flies - *Drosophila P* element (Kidwell 1985), and L1 in humans (Sassaman *et al.* 1997).

Numerous studies have been conducted to identify and characterize TEs in insect at the genomic level, such as in *D. melanogaster* (Kaminker *et al.* 2002), *A. gambiae* (Holt *et al.*

2002; Fernández-Medina *et al.* 2011), *Culex quinquefasciatus* (Arensburguer *et al.* 2010; Marsano *et al.* 2012), *Rhodnius prolixus* (Mesquita *et al.* 2015), and *Tribolium castaneum* (Wang *et al.* 2008), among others; however, less is known about these elements at the transcription level (de Araujo *et al.* 2005; Deloger *et al.* 2009; Mourier and Willerslev 2010; Iorizzo *et al.* 2011; Jiang *et al.* 2012; Sze *et al.* 2012). TEs are not only relevant due to the bulk of the genome occupied by them, but also due to the impact they have at the transcription level, by inserting into coding or regulatory regions, by influencing alternative mRNA processing or as sources of small regulatory RNAs (Kines and Belancio 2012; Mourier and Willerslev 2010; Cowley and Oakey 2013; de Araujo *et al.* 2005; Iorizzo *et al.* 2011; Kelley and Rinn 2012).

Here, we present an analysis of the transposable elements present in the *A. funestus* transcriptome, which together with *A. gambiae*, are the most important vectors for malaria transmission in sub-Saharan Africa. Mosquitoes of different species are among the most relevant invertebrate's vectors for veterinary and human vector borne diseases (VBD). Several *Anophelines* species are responsible for the transmission of Malaria, a disease that caused 212 million cases and 429,000 deaths, in 2015 (World Malaria Report 2016). Of the several measures used to control VBD, the mosquito populations are one of the main targets by the use of insecticides against which, many of them have acquired resistance (Chénais *et al.* 2012). TE-mediated mechanisms for developing resistance against insecticides in mosquitoes have been reported previously (Mouches *et al.* 1990, 1991, Darboux *et al.* 2007). Even if the relevance of TEs to insecticide resistance as a rule is not especially strong, they exemplify relations between the 'TEs landscape' and the appearance of adaptive traits, which is of big importance given the fact that vector control is one of the only control measures that show broad efficacy against diseases such as malaria.

In the malaria mosquito, several transcriptionally active TE families belonging both to Class I and II were identified and characterized. A full-length putative active element (with the presence of full-length TIRs in the genomic sequence) was also identified. This element belongs to the *hAT* superfamily, which also presents active members in other insect genomes (*Herves* in *A. gambiae*, *Hermes* in *Musca domestica* and *A. aegypti* and *hobo* in *Drosophila*) (Arensburguer *et al.* 2005, Warren *et al.* 1994, O'Brochta *et al.* 1996, Sarkar *et al.* 1997; Sheen *et al.* 1993). Our data also showed a great diversity of TEs expressed at the transcriptome level in the mosquito.

2. Materials and Methods

2.1 Transcriptome Assemblying

We have assembled the raw data generated in a *de novo* transcriptome Illumina sequencing approach that used RNA from 30 *A. funestus* adult female individuals 3-5 days old (Crawford *et al.* 2010) derived from a newly founded colony in Burkin Faso and mRNA sequences obtained from different stages (pupae, larvae and adults from both sexes) of two strains (a field collected and a laboratory strain) and sequenced by the 454 technique (Gregory *et al.* 2011). The ABySS system -a short-read assembler that can process genome or transcriptome sequence data (Simpson *et al.* 2009; Birol *et al.* 2009; Robertson *et al.* 2010) was then used to assemble the short sequences obtained. As transcriptome samples typically contain transcripts with a wide range of expression levels, and assemblies generated with different k-mer lengths perform differently in capturing transcripts expressed at different levels, it is recommended to use several values for k-mer. We used k values ranging from 25 to 65 and generated a final non-redundant fastA following reassembly of the different fastA files from each k, using a parallelized blast/cap3 pipeline where blastn was used with decreasing word sizes (-w switch) from 300 to 60 to feed sequences to the cap3 assembler (Huang and Madan 1999; Karim et al. 2011). After assembling the sequences the final contig set consisted of 46,398 contigs.

Transposable Elements Identification and Characterization—In order to characterize and identify putative transposable elements sequences, all the expression units generated (46,398 sequences) were used as queries for several BLAST programs performed on different databases (both public and in-house database versions) as follows: (i) NR-LIGHT by Blastx (Altschul et al. 1997): a subset of the non-redundant database (nr-db) from the NCBI, containing approximately 30% of the sequences and based on 106 genera and species, (ii) SWISSP protein database by Blastx, (iii) Gene Ontology database by Blastx (iv) the CDD database by RPSBlast (v) the eukaryotic cluster of orthologous groups (KOG) database by RPSblast, (vi) the PFAM database by RPSblast, (vii) the PRK database for functional annotation of the NCBI by RPSblast, (viii) the SMART database by RPSblast, (ix) an in-house database called “TE-DATABASE” generated by Blastx using transposable elements against a subset of the nr-db having the following keywords related to TEs: “transposase”, “transcriptase”, “transposable element”, “retroposon”, “retrotransposon”, (x) an in-house database “TE-CLASS” generated by psi-blast using TE-class specific proteins as queries on the nr-db using the RPSblast, (xi) an in-house database “TRANSPOSASE” generated by blastX on the nr-db using the keyword “transposase”, (xii) an in-house database “GAG” generated by blastX on the nr-db using the keyword “gag”, (xiii) an in-house database “RRNA” containing rRNA, (xiv) an in-house database “MIT-PLA” containing mitochondrial and plastid DNA sequences and (xv) “REPBASE” a TE reference database for eukaryotic elements: all the protein sequences from TEs deposited in REPBASE (both the entries with translated ORFs as well as the theoretical translations of those ORFs larger than 200 aa that were not presented as translated sequences in Repbase) were used as queries by a Psiblast on the nr-database in order to generate the TE models that were further used to run rps-Blasts against different sets of query sequences. We have previously used a similar approach in order to characterize the TE present in the genome of *A. gambiae* (Fernández-Medina *et al.* 2011).

Our objective was to identify and characterize putative active elements being expressed in the *A. funestus* transcriptome. The criteria for selecting putative active transposable element sequences is presented in the workflow shown in Figure 1. Briefly, 46,398 sequences were screened for identities against the sequences deposited in Repbase (RB) by RPSBlast. 2803 showed e-values <10e^{-15}, of which 926 have more than 50% coverage and 151 more than 20% identity with the Repbase elements. The remaining 45,247 sequences were screened by their identities to a TE-database by Blastx. 1539 were selected of which those covering more than 50% of the elements and presenting amino acid identities higher than 30% were included for further analysis. We totally selected 211 sequences presenting similarities to previously described TEs and further analyzed those sequences. All the remaining sequences were excluded of our analysis.

The sequences were further classified according to their TE class and superfamily, and divided in those presenting conserved domains (according to their "Best matches to the CDD" and "Best matched to pfam" databases) and those representing fragments or not showing the presence of functional domains.

2.2 Genome mapping

To evaluate the degree of identity of the "de novo" assembled sequences to the recently assembled *An. funestus* genome (Neafsey et al. 2015), we compared the assembled sequences to the assembled genome (version 1.0 from Vectorbase) using the blat tool (Li and Durbin 2010), as well as by mapping the raw reads from the Kiribina (NCBI bioproject PRJNA177025) and Folonzo data sets (PRJNA177018) by Blastn (Altschul et al. 1997) (using a word size of 30, allowing for 1 gap, minimum 95% identity and up to 10 mapped reads to different targets if and only if the matches had the same score) to the deducted coding sequences of *An. funestus* added of 211 TE sequences that were "de novo" assembled. The resulting read mapping indicated an average/median coverage depth of 92.7 / 70.8 fold for the Folonzo data set and 68.5 / 47.6 fold for the Kiribina data set. To estimate the copy number for each TE or CDS, we divided the CDS or TE fold coverage obtained for each library by the median coverage of the respective library.

2.3 Phylogenetic Analysis

Sequences belonging to different superfamilies were aligned with canonical sequences using MUSCLE (Edgar 2004). Phylogenetic relationships among the transposable element sequences and canonical sequences from the same superfamily/lineage were explored using neighbor-joining (NJ) and maximum likelihood (ML). The amino acid substitution models were evaluated using MEGA 5.0 (Tamura *et al.* 2011), the models with the lowest BIC scores (Bayesian Information Criterion) were considered the best to describe the substitution pattern (Tamura *et al.* 2011). NJ and ML trees were constructed using MEGA 5.0. Bootstrap values for each branch were assessed from 1000 replicates in both cases.

3. Results and Discussion

3.1 Transposable Elements Identification

In order to study the expression of TEs in *A. funestus*, we assembled the 102.6 M Illumina reads from RNAseq generated by Crawford *et al.* (2010) together with the 375,619 454-pyrosequencing reads generated by Gregory *et al.* (2011) yielding a total of 46,398 contigs. We further used a pipeline that relies on different algorithms based on the Blast programs, against several databases as subjects (see methodology) in order to identify and characterize the TEs expressed in the mosquito genome (Figure 1).

The results were compiled in a database (db) of expressed repetitive elements called Afun-TExcel (Table S1). This database provides information about 211 sequences that were clearly identified as TEs in the mosquito's transcriptome. The information is organized as an

Excel spreadsheet with cells containing, in a hyperlinked format, the results obtained after the various analyses performed in the characterization of each TE-like sequence.

We used stringent criteria for the inclusion of a sequence as a putative TE-like transcript (i.e. sequences with highly significant blast matches to Repbase or to TE-CLASS). Therefore, we have obtained a conservative set of expressed, putative active elements.

Our approach relies on the similarity of the transcript sequences to known proteins derived from TEs already characterized in other genomes. The use of RPSBlast, a position specific search engine based on profiles, against a PSIBlast-generated RB profile database as a first approach facilitated the classification and functional annotation of sequences. We identified 2,803 sequences with significant matches to a set of TE-specific profiles by RPSBlast, based on all the elements deposited in RB, however, most of these sequences present very small coverage or low identities or, in some cases present high identities to nuclear proteins and were consequently discarded. We further restricted the search to those matches presenting at least 50% coverage (926 sequences) and more than 20% amino acid identity (151 sequences) to known TEs. The remaining sequences (46,247) were further classified according to their matches to the “TE-db” by BlastX. This approach allowed the inclusion of 1,539 sequences that were again restricted to those presenting at least 50% coverage (189 sequences) and more than 30% amino acid identity (60 sequences). The resulting 211 sequences (151 plus 60) present e-values in the rps-Blast against RB, smaller than 1×10^{-15} confirming the identity of the sequences to already characterized TEs. However, many of these sequences (118) are not represented by full-length TE-transcripts (i.e gag-pol, for class I, or transposase, for class II), but by truncated sequences of which 77.9% belong to the LTR order, 14.4% to the NLTRs and the remaining 7.6% to the class II. These fragments have been clearly characterized as TEs, however, they have not been further analyzed.

Since most of the TEs in the genomes are represented by inactive or truncated copies, the transcripts identified here constitute an under representation of the total TE content in this genome. However, our study shows that TEs belonging to different families and subfamilies are present in this organism, many of which are transcriptionally active.

3.2. Transposable Elements Characterization

Transposable Elements representing all the classes/orders and most of the superfamilies previously characterized in insects were found in the *A. funestus* transcriptome, contributing to the mRNA and probably to protein diversity in this mosquito. Overall, the TE-related sequences identified in the transcriptome of *A. funestus* belong to Class I (86%), with a great overrepresentation of the LTR retrotransposons (Figure 2). The high diversity of TEs being expressed in this organism is also present in other insect genomes, such as the mosquitoes: *A. gambiae, A. darlingi* and *Culex quinquefasciatus*, or the fly *D. melanogaster* (Holt *et al.* 2002; Marinotti *et al.* 2012; Arensburguer *et al.* 2010; Adams *et al.* 2000). Although these insects harbor relatively small genomes, they contain many different TE families belonging to most of the TE orders/superfamilies so far identified. Not all the TE diversity found in a given genome would be expressed, in most of the cases very few elements are able to be expressed due to a high degree of deterioration of the elements in modern day genomes. In order to correlate our results with those of the reference genome, we have compared the number of TE families resulting from the analysis of the transcriptome (this study) with those identified in the genome (Neafsey et al, 2015) (Table 1). Most of the families we have identified in the transcriptome have also been identified in the genome. There are some discrepancies though, mainly with the number of families identified in each superfamily or lineage but not in the presence/absence of the superfamilies. The exceptions for this are the class II superfamilies, Transib and Academ, with two and four families identified in the genome that were not identified in the transcriptome and the class I, NLTRs, Ingi, L2 and Ouctast. These differences can be due to the differential expression of the families in the transcriptome.

In our data set, composed of 211 TE-like sequences, we have identified 30 sequences corresponding to the expression of full-length elements, 61 sequences spanning over full-length domains, and 120 sequences representing fragmented domains clearly belonging to TEs, as previously mentioned.

In order to confirm the presence of these TE families in the genome, and to validate the “de novo” assembly, we selected the sequences corresponding to full-length domains belonging to all the superfamilies identified in the transcriptome and compared them to the assembled *An. funestus* genome. Sixty two of the 211 sequences had no matches to the assembled genome, while 106 and 133 TE's provided better than 95 and 90% identity to genomic sequences, respectively, indicating that the “de novo” TE assembly appears reasonable. The lack of genomic representation of almost 30% of the identified TE's may reflect the difficulties posed by these repetitive sequences on genome assemblies. To additionally validate the TE sequences, we mapped the genomic raw sequences both from the *Kiribina* and *Folonzo* strains of *A. funestus* (Neafsey *et al.* 2015) to the 211 TE sequences. The median genomic coverage of each dataset (mapped to 13,714 TE sequences+ CDS) was of the same order of magnitude, 47.6 and 70.8 fold, respectively. All the TE-like sequences were identified in the genomic raw data with an average linear coverage of 99.7% in both datasets, including all the TE sequences that had no genomic matches. While we cannot exclude that our “*de novo*” assembly contain chimeric elements, the matches to genomic sequences as well as the raw reads indicate that the majority are representative of sequences corresponding to bona fide elements present in the genome, or the assembly of closely related elements. The element copy number in each library was estimated by considering the coverage of the TE-like sequences in each set (*Kiribina* and *Folonzo* strains) divided by the median library coverage to all TE sequences + CDS in order to obtain an average for both libraries. Overall, at the genome level, the Class I contains superfamilies with the highest estimated copy numbers, especially belonging to the NLTR order. Among them, the R1 superfamily is the most abundant with 315 copies spread in the genome (Figure S1).

The diversity of the TEs in the genomic dataset has been shown to be higher than our estimation departing from expressed sequences (Neafsey *et al.* 2015) an expected fact, since not all the TEs in the genome will be expressed at the transcriptome level. Totally, 4,719 LTR elements were reported in the genome of *A. funestus* of which the *Gypsy* constitute the least represented superfamily in copy number, 786 copies, against 2,129 *Copia* and 1,328 *Bel-Pao* copies (Neafsey *et al.* 2015). We also identified several Class II elements, including many families of *Tc-1/mariner, hAT, PiggyBac* and *Harbinger* superfamilies that were also previously identified at the genomic level (Neafsey *et al.* 2015).

3.2.1 Class I – LTRs—

Four superfamilies within the LTR retrotransposons have been previously described: *Gypsy*, *Bel-Pao, Copia* and *DIRS* (for review see Wicker *et al.* 2007). Each of them has been subsequently classified into different clades/lineages (Coppeland *et al.* 2005; de la Chaux and Wagner 2011; Llorens *et al.* 2009).

In the *A. funestus'* transcriptome sequences belonging to the three main superfamilies (*Gypsy*, *Bel-Pao* and *Copia*) were identified. Transcripts belonging to the *Gypsy* superfamily were the most abundant, even if they have been reported as the less abundant LTR superfamily at the genomic level (Neafsey *et al.* 2015). Of the 78 *Gypsy*-like transcripts, one corresponds to a full-length sequence expressing all the domains in both ORFs 1 and 2: retropepsin, reverse transcriptase (RVT-1), RNAseHI_RT_ty3, and integrase (RVE); 29 contain one or more full-length domains and the remaining transcripts correspond to sequences with truncated domains.

The *Bel-Pao* elements were the second most abundant transcripts identified within the LTR order (49 sequences). Six of them harbor full-length ORFs with all the domains present in a complete element: DUF1758-1759/RVT-1/PeptA17/RVE and twenty seven contain full length domains.

Copia elements, were the least numerous LTR transcripts in the transcriptome, although they represent the most numerous superfamily at the genomic level, with an estimation of 2,000 copies (Neafsey *et al.* 2015). Ten out of 30 *Copia* sequences contain full-length domains, and three have all the domains present in a full-length element (GAG-integrase/RVE/RVT_2/RNAseH_Ty1).

3.2.2 Class I – Non-LTRs—

It is well established that Non-LTR elements create in their replication cycle "Dead-on-arrival" sequences, producing fragments or truncated sequences missing their 5′ends. They have a replicative transposition mechanism that depends on the transcription of the whole element following a reverse transcription step. They are the least represented type of element in the *A. funestus'* transcriptome; 24 different transcripts corresponding to seven different clades, representing 11% of all the TE-like transcripts were identified (Figure 2). However, only the transcripts of the R1 clade present ORFs containing full-length domains and were further analyzed. Sequences representing truncated domains of the *CR1* (4), *I*, *Jockey*, *L1*, *RT2* and *RTE* clades (one sequence in each superfamily) could be identified. The truncated domains correspond to the Exo-Endo-Phosphatase (EEP); the RVT1 and the RH domains. These truncated sequences probably represent incomplete assembled sequences.

3.2.3 Class II – DNA Transposons—

Most of the previously identified DNA transposons belong to the class of cut-and-paste DNA transposons (Subclass 1), currently represented by 15 superfamilies (Kapitonov and Jurka 2008). The transposases encoded by cut-and-paste DNA transposons are also called DDE/DDD Transposases, due to the universal occurrence of three conserved acidic catalytic residues: two aspartates (D) and one glutamate (E), or three aspartates (DDD).

In the *A. funestus'* transcriptome, the Class II elements correspond to 14% of the TE-like sequences, most of which belong to the TIR order. Two transcripts belonging to the *Helitron* order have also been identified. The identified transposase domains belong to different DDE superfamilies of endonucleases including DDE_1 and DDE_3 (from *Tc1-mariner* elements), DDE_4 (from *Harbinger*), the *hAT* family dimerization domain (from *hAT* elements), DDE_Tnp1_7 (from a *Piggybac* element), Tnp_P (from a *P*-element) and the DEAD-like and C-terminal domains of helicases from two *Academ* elements. The most abundant expressed transposase belongs to the *Tc1/mariner* superfamily.

3.3 Phylogenetic Analyses

3.3.1 Class I – LTRs—The use of the coding region corresponding to the reverse transcriptase to determine phylogenetic relationships has shown to be adequate for classification purposes (Xiong and Eickbush 1988). We performed a phylogenetic analysis of all the sequences presenting full-length ORFs corresponding to the RT domain, i.e 10 *Gypsy*, nine *Bel-Pao* and three *Copia* sequences, together with previously published reference sequences belonging to different insect species. The phylogeny confirmed the sequence classification based in our pipeline for all the LTR sequences at the superfamily level (Figure 3). Additional phylogenetic analyses were performed for the sequences in each superfamily in order to classify them into lineages and further into families.

The Gypsy superfamily: The *Gypsy* superfamily in insects has been classified into five different families also called lineages or clades (i.e *Mag*, *Mdg1*, *Mdg3*, *Gypsy* and *CsRn1*). The 10 *Gypsy* sequences identified in the *A. funestus* transcriptome spanning the whole RT domain (177 aa. positions) were aligned together with reference sequences representing the five *Gypsy* lineages previously described and three *Bel-Pao* elements as outliers (Figure 4). A *Gypsy*-like partial element previously described in the genome of *A. funestus,* deposited in RB and called *Afun1* (Cook *et al.* 2000) was also included in the alignment. This sequence did not cluster with any of the expressed sequences from *A. funestus* and it is 100% identical to *Gypsy35_Agam*, from the *A. gambiae's* genome. Nine sequences from *A. funestus* clustered together with elements belonging to four different lineages, i.e. *gypsy, mag, mdg3* and *CsRn1*; sequences clustering with the *mdg1* lineage, were not identified (Figure 4).

Four sequences belonging to the CsRn1 lineage (*Gypsy1-4_Afun*) belong to different families (mean p-distance of 40.08%, ranging from 26.3 and 47.4%). The sequences *Gypsy1_Afun, Gypsy2_Afun* and *Gypsy4_Afun* share the same family with several sequences from *A. gambiae. Gypsy1* has 94.59% identity with *Gypy49-AG*; *Gypsy2_Afun* has 85.10% identity with *Gypy53-AG*, and *gypsy4* has 91.48 and 86.70% identity with *Gypsy52_AG* and *Gypsy2_AG*, respectively. *Gypy3_Afun*, on the other hand is not clustering with any of the sequences used as references.

*Gypsy*5_*Afun* and *Gypsy*8_*Afun* clustered together with reference sequences from the *mdg3* lineage; however they belong to different families, presenting more than 63% of p-distance among them. *Gypsy*8_*Afun* clustered in the same family with sequences from *A. gambiae*, with a mean p-distance of 27.11%, while *Gypsy*5_Afun clustered with sequences from *A. aegypti* with a mean p-distance of 33.48% constituting the same family. *Gypsy*7_Afun, *Gypsy*9_Afun and *Gypsy*10_Afun clustered within the MAG lineage however; they are clearly related to different families within this lineage (mean p-distance 69.2% ranging from 66.8 to 62.7%).

The Bel-Pao Superfamily: The *Bel-Pao* superfamily has been previously classified into seven discrete lineages (*Pao*, *Sinbad*, *Bel*, *Tas*, *Suzu, Flow and Dan*) which tend to cluster with the host species phylogeny (Copeland *et al.* 2005; de la Chaux and Wagner 2011). These elements appear to have colonized only the genomes of the kingdom *Animalia*. We performed a phylogenetic analysis of the full-length RT region (213 aa) of nine *Bel-Pao* elements expressed in *A. funestus* and a C-terminal truncated sequence together with reference sequences from other insect genomes (Figure 5).

The majority of the elements from *A. funestus* preliminary classified as *Bel-Pao* clustered together with the *bel* lineage while two sequences did it with the *Pao* reference sequences, and none of them clustered with sequences from the other lineages in this superfamily. The p-distance between the sequences *Bel-1_Afun*, *Bel-4_Afun, Bel-7_Afun* and *Bel*-9_*Afun* is 37.4%, (ranging from 0.0% to 47.2%) indicating that they belong to the same family. On the other hand, sequences *Bel-2_Afun* and *Bel*-5_*Afun* (p-distance 29%) belong to the same family than *Bel-6_Afun* (mean p-distance 44.2%).

The Copia superfamily: We identified three sequences in the *A. funestus* transcriptome containing full-length RT sequences (246 aa) and belonging to the *Copia* superfamily. The *Copia* superfamily in insects has not been previously classified into different families or lineages/clades. We used 62 *copia* reference sequences from other insects (including, mosquitoes, flies and bugs) in order to classify the three *copia* sequences expressed in the *A. funestus* transcriptome. In our phylogeny the reference sequences grouped into, at least, five different lineages (I to V) one of them corresponds to *Drosophila*'s sequences only (indicated in blue in Figure 6) while the others contain a mixture of sequences belonging to different mosquitoes' species. These sequences clustered within the same major group of sequences and together with sequences obtained from *A. aegypti* and *C. quinquafasciatus,* but none of the families previously characterized in *A. gambiae*. The p-distance for the three sequences from *A. funestus* is 51.0%, indicating that they belong to different families.

3.3.2 Class I – Non-LTR—Several sequences belonging to different clades belonging to the NLTR order were identified in the *A. funestus* transcriptome, although not all of them correspond to the RT region or represented full-length domains.

Eight sequences (seven belonging to the R, and one to the RTE clade) corresponding to the RT sequence were aligned to reference sequences representing all the clades described in the Non-LTR order. The phylogenetic analysis confirmed the classification obtained after our pipeline was applied (Figure 7). A phylogenetic analysis including reference sequences belonging to the R1 superfamily in insects was performed (data not shown). The *A. funestus* R1 sequences clustered into four different clusters, The sequences *R1-1,2,3,5,8_Anfun* clustered in a node together with sequences from *A. gambiae*. The p-distance among them is 32.5% (ranging from 16.2% to 33.5%), indicating that they belong to the same lineage together with sequences from *A. gambiae*. *R1-9_Afun* (mean p-distance against all the sequences from *A. funestus* 53.5%) and *R1-4_Afun* (mean p-distance against all the sequences from *A. funestus* 55.4%), on the other hand belong to more distant families.

3.3.3 Class II - DNA transposons—Transposases typically contain two domains: the N-terminal DNA-binding domain (an Helix-turn-Helix domain, known as HTH) (Pietrokovski and Henikoff 1997) and the C-terminal catalytic domain, characterized by the presence of a conservative D(Asp)DE(Glu)/D triad (Brillet *et al.* 2007) that has been shown to be essential for transposase activity (Lohe *et al.* 1995). Phylogenetic analyses of the this domain in *mariner* elements permitted the classification of these elements into elven subfamilies: *cecropia*, *irritans*, *mauritiana*, *mellifera, capitata* (Robertson 1993, Robertson and MacLeod 1993), *mori* (Robertson and Asplund 1996), *elegans* and *briggsae* (Robertson 2002), *rosa* (Gomulski *et al.* 2001), *vertumnana* (Green and Frommer 2001), and *marmoratus* (Bui *et al.* 2007).

We have used the full-length transposase sequences of four *Tc1/mariner,* three *Pogo* and four *Harbinger* sequences from the *A. funestus* transcriptome in a phylogenetic analysis including reference sequences representing the different already characterized superfamilies (Figure8). Two of the sequences clustered together with reference sequences from the *mariner*'s family while three sequences clustered within the *Pogo* family, other two sequences clustered with the *Tc1* family, all with significant bootstrap values. Four sequences clustered together with the *Harbinger* family. The p-distances among the sequences belonging to the *Harbinger* (75.37%), *Pogo* (74.31%) and *Tc1* (71.59%) families indicated that they belong to different subfamilies within each family. While the two *mariner* sequences show a lower distance (56.17%) indicating that the sequences do not belong to the same subfamily.

The *Tc1/mariner* elements identified in the *A. funestus* transcriptome contain some divergences from the canonical DDD/E domains (Table 1). Transposase sequences containing the GD34N and the AN45D were identified. We cannot assure that these sequences result in active transposases (Table 1).

In order to further characterize these two sequences within the *mariner* family, a phylogeny including these two sequences and members representing each of the characterized *mariner* subfamilies was performed. The phylogeny indicated that the *mariner* sequences belong both to the *mauritiana* subfamily (data not shown) (Robertson and McLeod 1993; Wallau *et al.* 2014).

<u>The hAT superfamily:</u> The *hAT* elements (by *hobo*, *Ac* and *Tam3*) are present in a wide range of plants and animals, including insects (Kempken and Windhofer 2001; Weil and Kunze 2000). Interestingly, these elements have been found in active forms in insects and also, to be active when introduced into divergent insect species, making them suitable as gene drivers. *Hermes* from the housefly, *Musca domestica* (Atkinson *et al.* 1993), *hobo* from *D. melanogaster* (McGinnis *et al.* 1983), *Herves* from *A. gambiae* (Arensburger *et al.* 2005) also, shown to be present in the genomes of other *Anophelines* (*A. arabiensis* and *A. merus*) and in *A. aegypti* (Arensburger *et al.* 2011) have all been shown to be mobilized in different species.

Sequences belonging to three different *hAT* elements were identified in *A. funestus*. We used the full-length *hAT1_Afun* sequence as a query in a BlastN search in the *A. funestus* sequenced genome. Five *hAT* sequences were identified in the genome. Two pairs of intact TIRs flanking the transposase gene were also identified. A phylogeny of the C-terminal dimerization region of the sequences identified both in the transcriptome and the genome together with reference sequences from *hAt* elements from other insect genomes showed a cluster of three pairs of transcriptome and genome sequences (*hAT1-3_Afun*) (Figure 9). Two other genomic sequences, named *hAT4,5_Afun* presented no expressed counterpart. The p-distance among the genomic sequences corresponding to the three expressed elements is 39.18%, indicating that the three of them belong to the same family. On the other hand, the p-distances between the sequences in the cluster and *hAT4_Afun* and *hAT5_Afun* are 47.89 and 56.96%, respectively.

The finding of an expressed element, presenting full-length transposase gene and intact TIRs at the genomic level is suggestive of recent or ongoing active transposition of this family. The ability to transpose in diverged species from their hosts appears to be common feature of members of the *hAT* superfamily. Consequently, the *hAT* elements have great potential to serve as non-drosophilid insect gene vectors.

4. Concluding Remarks

In this article we have presented data related to the identification and characterization of transposable elements in A. funestus, the second most important vector of malaria in Africa. We have used publicly available data of two whole transcriptome sequencing projects (Crawford et al. 2010, Gregory et al. 2011) to elucidate the extent and character of the repetitive elements being expressed in this mosquito. TEs belonging to all the classes and to most of the TE superfamilies already characterized are present and expressed in this organism. The methodology implemented resulted in the identification of TE superfamilies already identified and characterized in other sequenced genomes avoiding the identification of putative novel elements. The determination of the activity or potential activity of these elements will require further functional verification.

The identification of a vast diversity of TEs expressed in *A. funestus* suggests that there is ongoing amplification of several families in this organism. On the other hand, the lack of genomic representation of many TE's reflects the difficulties related to the correct assembling of transposable elements sequences in genome assemblies.

The data presented here is based primarily on the similarity of the transcript sequences to functional domains for known transposable elements in other species. This might be insufficient support for actual activity of these elements. It is known that the potential for TE activity on a structural level may be restricted by cell type, especially the soma and germ-line, or epigenetic microRNA regulation. And, it is well established that the transcriptional activity of transposable elements is tightly controlled, although some retrotransposons are transcribed under stress conditions such as pathogen infection, physical injuries or abiotic stresses (Grandbastien, 1998; Takeda et al., 1998).

TE-mediated mechanisms for developing resistance against insecticides in mosquitoes have been reported previously. An amplification of at least 250-fold of the esterase gene related to the overproduction and increased activity of esterase B1, involved in the resistant phenotype of Culex species to organophosphorus (OP) insecticides, has been identified in Culex pipiens quienquefasciatus. An amplicon (30kb) in the resistant mosquitoes contained the esterase gene (2.8Kb) framed by DNA sequences homologous to middle or highly repetitive elements present in the genome of susceptible and OP-resistant mosquitoes, which were thought to be of TE origin (Mouches et al. 1990 1991). Microbial larvicides have also commonly been used for controlling mosquitoes-borne diseases. A binary toxin from Bacillus sphericus has a larvicidal property following ingestion by susceptible larvae. However, high levels of resistance has also been reported in field populations of Culex species isolated from different countries where the larvicide has been used extensively (Rao et al. 1995; Yuan et al. 2000; Chevillon et al. 2001). One of the mechanisms of the larvae resistance is related to the insertion of a TE-like element in the coding region of the gene that codifies for the receptor involved in the interaction with the toxin (Darboux et al. 2007). The TE insertion modified a splicing site, creating an intron and leading to the production of a shorter receptor, unable to interact with the toxin and leading to the insect survival. Even if the relevance of TEs to insecticide resistance as a rule is not especially strong, they exemplify relations between the ‘TE landscape’ and the appearance of adaptive traits, which is of big importance given the fact that vector control is one of the only control measures that show broad efficacy against diseases such as malaria.

We used stringent criteria to designate putative functional transcripts, however, we cannot exclude the possibility that some of these sequences might represent read-throughs transcripts. Still, given the importance of the transposable elements and their role in many spontaneous mutations influencing evolution, and the adaptative traits in insects (for instance insecticide resistance) it is important to characterize the presence of this elements at the transcriptome level.

It is therefore not only important to understand the landscape of elements present in any given genome but also the expression of those sequences in transcriptomes. As the examples showed above indicate, TEs might play an important role in the appearance of larvicide- and insecticides- resistant phenotypes, emphasizing the significance of studies aiming at the identification and characterization of TEs in genomes and trancriptomes.

On the other hand, the use of transposable elements as tools for the introduction of desirable genes into target populations has also been pursued as a means for controlling VBD, particularly, the transformation of A. gambiae as a means to control the spread of malaria. Active TEs can be used in genetic engineering as transformation vectors and can be used for gene and enhancer trapping; they also can be used for genome-wide insertional mutagenesis studies (Tu and Li 2009). In this respect, we have identified a full-length putative active element (including the presence of full length TIRs in the genomic sequence) belonging to the hAT superfamily, which presents active members in other insect genomes (Herves in Anopheles gambiae, Hermes in Musca domestica and Aedes aegypti, and hobo in Drosophila) and that have been also used as driver elements. Moreover, a great diversity of active elements is present in A. funestus. Further functionality tests by mobility assays could be of great importance in order to determine the use of these elements as genetic tools in other species. This work contributes overall to a comprehensive understanding of the landscape of transposable elements in this important vector for malaria.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

R.D.F.M. has a postdoctoral position supported by Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Comissão de Aperfeiçoamento de Pessoal do Nível Superior (CAPES). J.M.C.R was supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases. Because JMCR is a government employees and this is a government work, the work is in the public domain in the United States. Notwithstanding any other agreements, the NIH reserves the right to provide the work to PubMedCentral for display and use by the public, and PubMedCentral may tag or modify the work consistent with its customary practices. You can establish rights outside of the U.S. subject to a government use license. C.M.A.C. was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant306493/2013-6 and CJS was partially funded by CNPq and FAPERJ.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, cPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Sidãon-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weisenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. The genome sequence of *Drosophila melanogaster*. Science. 2000 Mar 24; 287(5461):2185–95. [PubMed: 10731132]

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Arensburger P, Hice RH, Zhou L, Smith RC, Tom AC, Wright JA, Knapp J, O'Brochta DA, Craig NL, Atkinson PW. Phylogenetic and functional characterizationof the *hAT* transposon superfamily. Genetics. 2011; 188(1):45–57. [PubMed: 21368277]

Arensburger P, Kim YJ, Orsetti J, Aluvihare C, O'Brochta DA, Atkinson PW. An active transposable element, Herves, from the African malaria mosquito *Anopheles gambiae*. Genetics. 2005; 169(2): 697–708. [PubMed: 15545643]

Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, Bartholomay L, Bidwell S, Caler E, Camara F, Campbell CL, Campbell KS, Casola C, Castro MT, Chandramouliswaran I, Chapman SB, Christley S, Costas J, Eisenstadt E, Feschotte C, Fraser-Liggett C, Guigo R, Haas B, Hammond M, Hansson BS, Hemingway J, Hill SR, Howarth C, Ignell R, Kennedy RC, Kodira CD, Lobo NF, Mao C, Mayhew G, Michel K, Mori A, Liu N, Naveira H, Nene V, Nguyen N, Pearson MD, Pritham EJ, Puiu D, Qi Y, Ranson H, Ribeiro JM, Roberston HM, Severson DW, Shumway M, Stanke M, Strausberg RL, Sun C, Sutton G, Tu ZJ, Tubio JM, Unger MF, Vanlandingham DL, Vilella AJ, White O, White JR, Wondji CS, Wortman J, Zdobnov EM, Birren B, Christensen BM, Collins FH, Cornel A, Dimopoulos G, Hannick LI, Higgs S, Lanzaro GC, Lawson D, Lee NH, Muskavitch MA, Raikhel AS, Atkinson PW. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science. 2010; 330(6000):86–8. [PubMed: 20929810]

Atkinson PW, Warren WD, O'Brochta DA. The hobo transposable element of Drosophila can be cross-mobilized in houseflies and excises like the Ac element of maize. Proc Natl Acad Sci U S A. 1993; 90(20):9693–7. [PubMed: 8415764]

Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. Ann Bot. 2005; 95:127–132. [PubMed: 15596462]

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ. De novo transcriptome assembly with ABySS. Bioinformatics. 2009; 25(21):2872–7. [PubMed: 19528083]

Brillet B, Bigot Y, Augé-Gouillou C. Assembly of the Tc1 and *mariner* transposition initiation complexes depends on the origins of their transposase DNA binding domains. Genetica. 2007; 130(2):105–20. [PubMed: 16912840]

Bui QT, Delaurière L, Casse N, Nicolas V, Laulier M, Chénais B. Molecular characterization and phylogenetic position of a new *mariner*-like element in the coastal crab, *Pachygrapsus marmoratus*. Gene. 2007; 396:248–256. [PubMed: 17490833]

Casacuberta E, Pardue ML. HeT-A and TART, two Drosophila retrotransposons with a bona fide role in Chromosome Structure for more than 60 million years. Cytogenet Genome Res. 2005; 110(1-4): 152–159. [PubMed: 16093667]

Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene. 2012; 509(1):7–15. [PubMed: 22921893]

Chevillon C, Bernard C, Marquine M, Pasteur N. Resistance to *Bacillus sphaericus* in *Culex pipiens* (Diptera: Culicidae): interaction between recessive mutants and evolution in southern France. J Med Entomol. 2001; 38:657–664. [PubMed: 11580038]

Cook MJ, Martin J, Lewin A, Sinden ER, Tristem M. Systematic screening of *Anopheles* mosquito genomes yields evidence for a major clade of Pao-like retrotransposons. Insect Mol Biol. 2000; 9(1):109–117. [PubMed: 10672078]

Copeland Mann VH, Morales ME, Kalinna BH, Brindley PJ. The Sinbad retrotransposon from the genome of the human blood fluke, *Schistosoma mansoni*, and the distribution of related Pao-like elements. BMC Evol Biol. 2005; 5:20. [PubMed: 15725362]

Cowley M, Oakey RJ. Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet. 2013; 9(1):e1003234. [PubMed: 23358118]

Crawford JE, Guelbeogo WM, Sanou A, Traoré A, Vernick KD, Sagnon N, Lazzaro BP. De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology. PLoS One. 2010; 5(12):e14202. [PubMed: 21151993]

Darboux I, Charles JF, Pauchet Y, Warot S, Pauron D. Transposon-mediated resistance to Bacillus sphaericus in a field-evolved population of *Culex pipiens* (Diptera: *Culicidae*). Cell Microbiol. 2007; 9:2022–2029. [PubMed: 17394558] de Araujo PG, Rossi M, de Jesus EM, Saccaro NL Jr, Kajihara D, Massa R, de Felix JM, Drummond RD, Falco MC, Chabregas SM, Ulian EC, Menossi M, Van Sluys MA. Transcriptionally active transposable elements in recent hybrid sugarcane. Plant J. 2005; 44(5):707–17. [PubMed: 16297064]

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011; 7:e1002384. [PubMed: 22144907]

de la Chaux N, Wagner A. BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol. 2011; 11:154. [PubMed: 21639932]

Deloger M, Cavalli FM, Lerat E, Bièmont C, Sagot MF, Vieira C. Identification of expressed transposable element insertions in the sequenced genome of *Drosophila melanogaster*. Gene. 2009; 439(1-2):55–62. [PubMed: 19332112]

Edgar, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32(5):1792–1797. [PubMed: 15034147]

Fernandez-Medina RD, Struchiner CJ, Ribeiro JM. Novel transposable elements from *Anopheles gambiae*. BMC Genomics. 2011; 12:260. [PubMed: 21605407]

Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989; 5:103–107. [PubMed: 2543105]

Gomulski LM, Torti C, Bonizzoni M, Moralli D, Raimondi E, Capy P, Gasperi G, Malacrida AR. A new basal subfamily of *mariner* elements in *Ceratitis rosa* and other tephritid flies. J Mol Evol. 2001; 53(6):597–606. [PubMed: 11677619]

Green CL, Frommer M. The genome of the Queensland fruit fly Bactrocera tryoni contains multiple representatives of the *mariner* family of transposable elements. Insect Mol Biol. 2001; 10:371–386. [PubMed: 11520360]

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. Eukaryotic genome size databases. Nucleic Acids Res. 2007; 35(Database issue):D332–8. [PubMed: 17090588]

Gregory R, Darby AC, Irving H, Coulibaly MB, Hughes M, Koekemoer LL, Coetzee M, Ranson H, Hemingway J, Hall N, Wondji CS. A de novo expression profiling of Anopheles funestus, malaria vector in Africa, using 454 pyrosequencing. PLoS One. 2011; 6(2):e17418. [PubMed: 21364769]

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science. 2002; 298(5591):129–49. [PubMed: 12364791]

Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999; 9:868–877. [PubMed: 10508846]

Iorizzo M, Senalik DA, Grzebelus D, Bowman M, Cavagnaro PF, Matvienko M, Ashrafi H, Van Deynze A, Simon PW. De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics. 2011; 12:389. [PubMed: 21810238]

Jiang F, Yang M, Guo W, Wang X, Kang L. Large-scale transcriptome analysis of retroelements in the migratory locust, Locusta migratoria. PLoS One. 2012; 7(7):e40532. [PubMed: 22792363]

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, Ashburner M, Celniker SE. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 2002; 3(12)

Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008; 9(5):411–2. [PubMed: 18421312]

Kapitonov VV, Koonin EV. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biology Direct. 2015; 10:20. 2015. [PubMed: 25928409]

Karim S, Singh P, Ribeiro JM. A deep insight into the sialotranscriptome of the gulf coast tick, Amblyomma maculatum. PLoS ONE. 2011; 6:e28525. [PubMed: 22216098]

Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012; 13(11):R107. [PubMed: 23181609]

Kempken F, Windhofer F. The *hAT* family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma. 2001; 110(1):1–9. [PubMed: 11398971]

Kidwell MG. Hybrid dysgenesis in *Drosophila melanogaster*: nature and inheritance of P element regulation. Genetics. 1985; 111(2):337–50. [PubMed: 2996978]

Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002; 115(1):49–63. [PubMed: 12188048]

Kines KJ, Belancio VP. Expressing genes do not forget their LINEs: transposable elements and gene expression. Front Biosci (Landmark Ed). 2012; 17:1329–44. [PubMed: 22201807]

LeRouzic A, Boutin TS, Capy P. Long-term evolution of transposable elements. PNAS. 2007; 104(49): 19375–19380. [PubMed: 18040048]

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2010; 26:589–595.

Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct. 2009; 4:41. [PubMed: 19883502]

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL. Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. Mol Biol Evol. 1995; 12(1):62–72. [PubMed: 7877497]

Marinotti O, Cerqueira GC, de Almeida LG, Ferro MI, Loreto EL, Zaha A, Teixeira SM, Wespiser AR, Almeida E, Silva A, Schlindwein AD, Pacheco AC, Silva AL, Graveley BR, Walenz BP, Lima Bde A, Ribeiro CA, Nunes-Silva CG, de Carvalho CR, Soares CM, de Menezes CB, Matiolli C, Caffrey D, Araújo DA, de Oliveira DM, Golenbock D, Grisard EC, Fantinatti-Garboggini F, de Carvalho FM, Barcellos FG, Prosdocimi F, May G, Azevedo Junior GM, Guimarães GM, Goldman GH, Padilha IQ, Batista Jda S, Ferro JA, Ribeiro JM, Fietto JL, Dabbas KM, Cerdeira L, Agnez-Lima LF, Brocchi M, de Carvalho MO, Teixeira Mde M, Diniz Maia Mde M, Goldman MH, Cruz Schneider MP, Felipe MS, Hungria M, Nicolás MF, Pereira M, Montes MA, Cantão ME, Vincentz M, Rafael MS, Silverman N, Stoco PH, Souza RC, Vicentini R, Gazzinelli RT, Neves Rde O, Silva R, Astolfi-Filho S, Maciel TE, Urményi TP, Tadei WP, Camargo EP, de Vasconcelos AT. The genome of Anopheles darlingi, the main neotropical malaria vector. Nucleic Acids Res. 2012; 41(15):7387–400.

Marsano RM, Leronni D, D'Addabbo P, Viggiano L, Tarasco E, Caizzi R. Mosquitoes LTR retrotransposons: a deeper view into the genomic sequence of Culex quinquefasciatus. PLoS One. 2012; 7(2):e30770. [PubMed: 22383973]

McGinnis W, Shermoen AW, Beckendorf SK. A transposable element inserted just 5′ to a Drosophila glue protein gene alters gene expression and chromatin structure. Cell. 1983; 34(1):75–84. [PubMed: 6309414]

Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P, Spieth J, Carvalho AB, Panzera F, Lawson D, Torres AQ, Ribeiro JM, Sorgine MH, Waterhouse RM, Montague MJ, Abad-Franch F, Alves-Bezerra M, Amaral LR, Araujo HM, Araujo RN, Aravind L, Atella GC, Azambuja P, Berni M, Bittencourt-Cunha PR, Braz GR, Calderón-Fernández G, Carareto CM, Christensen MB, Costa IR, Costa SG, Dansa M, Daumas-Filho CR, De-Paula IF, Dias FA, Dimopoulos G, Emrich SJ, Esponda-Behrens N, Fampa P, Fernandez-Medina RD, da Fonseca RN, Fontenele M, Fronick C, Fulton LA, Gandara AC, Garcia ES, Genta FA, Giraldo-Calderón GI, Gomes B, Gondim KC, Granzotto A, Guarneri AA, Guigó R, Harry M, Hughes DS, Jablonka W, Jacquin-Joly E, Juárez MP, Koerich LB, Latorre-Estivalis JM, Lavore A, Lawrence GG, Lazoski C, Lazzari CR, Lopes RR, Lorenzo MG, Lugon MD, Majerowicz D, Marcet PL, Mariotti M, Masuda H, Megy K, Melo AC, Missirlis F, Mota T, Noriega FG, Nouzova M, Nunes RD, Oliveira RL, Oliveira-Silveira G, Ons S, Pagola L, Paiva-Silva GO, Pascual A, Pavan MG, Pedrini N, Peixoto AA, Pereira MH, Pike A, Polycarpo C, Prosdocimi F, Ribeiro-Rodrigues R, Robertson HM, Salerno AP, Salmon D, Santesmasses D, Schama R, Seabra-Junior ES, Silva-Cardoso L, Silva-Neto MA, Souza-Gomes M, Sterkel M, Taracena ML, Tojo M, Tu ZJ, Tubio JM, Ursic-Bedoya R, Venancio TM, Walter-Nuno AB, Wilson D, Warren WC, Wilson RK, Huebner E, Dotson EM, Oliveira PL. Genome of *Rhodnius prolixus*, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc Natl Acad Sci U S A. 2015; 112(48):14936–41. Epub 2015 Nov 16. Erratum in: Proc Natl Acad Sci U S A. 2016 Mar 8;113(10)E1415-6. DOI: 10.1073/pnas.1506226112 [PubMed: 26627243]

Miller WJ, McDonald JF, Pinske W. Molecular domestication of mobile elements. Genetica. 1997; 100:261–270. [PubMed: 9440279]

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008; 5(7):621–8. [PubMed: 18516045]

Mourier T, Willerslev E. Large-scale transcriptome data reveals transcriptional activity of fission yeast LTR retrotransposons. BMC Genomics. 2010; 11:167. [PubMed: 20226011]

Mouches C, Pauplin Y, Agarwal M, Lemieux L, Herzog M, Abadon M, Beyssat-Arnaouty V, Hyrien O, de Saint Vincent BR, Georghiou GP. Characterization of amplification core and esterase B1 gene responsible for insecticide resistance in Culex. Proc Natl Acad Sci U S A. 1990; 87:2574–2578. [PubMed: 2320576]

Mouches C, Agarwal M, Campbell K, Lemieux L, Abadon M. Sequence of a truncated LINE-like retroposon dispersed in the genome of *Culex* mosquitoes. Gene. 1991; 106:279–280. [PubMed: 1657726]

Mourier T, Willerslev E. Large-scale transcriptome data reveals transcriptional activity of fission yeast LTR retrotransposons. BMC Genomics. 2010; 11:167. [PubMed: 20226011]

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcá B, rensburger P, Artemov G, Assour LA, Basseri H, Berlin A, Birren BW, Blandin SA, Brockman AI, Burkot TR, Burt A, Chan CS, Chauve C, Chiu JC, Christensen M, Costantini C, Davidson VL, Deligianni E, Dottorini T, Dritsou V, Gabriel SB, Guelbeogo WM, Hall AB, Han MV, Hlaing T, Hughes DS, Jenkins AM, Jiang X, Jungreis I, Kakani EG, Kamali M, Kemppainen P, Kennedy RC, Kirmitzoglou IK, Koekemoer LL, Laban N, Langridge N, Lawniczak MK, Lirakis M, Lobo NF, Lowy E, MacCallum RM, Mao C, Maslen G, Mbogo C, McCarthy J, Michel K, Mitchell SN, Moore W, Murphy KA, Naumenko AN, Nolan T, Novoa EM, O'Loughlin S, Oringanje C, Oshaghi MA, Pakpour N, Papathanos PA, Peery AN, Povelones M, Prakash A, Price DP, Rajaraman A, Reimer LJ, Rinker DC, Rokas A, Russell TL, Sagnon N, Sharakhova MV, Shea T, Simão FA, Simard F, Slotman MA, Somboon P, Stegniy V, Struchiner CJ, Thomas GW, Tojo M, Topalis P, Tubio JM, Unger MF, Vontas J, Walton C, Wilding CS, Willis JH, Wu YC, Yan G, Zdobnov EM, Zhou X, Catteruccia F, Christophides GK, Collins FH, Cornman RS, Crisanti A, Donnelly MJ, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Hansen IA, Howell PI, Kafatos FC, Kellis M, Lawson D, Louis C, Luckhart S, Muskavitch MA, Ribeiro JM, Riehle MA, Sharakhov IV, Tu Z, Zwiebel LJ, Besansky NJ. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science. 2015; 347(6217):1258522. [PubMed: 25554792]

O'Brochta DA, Warren WD, Saville KJ, Atkinson PW. Hermes, a functional non-drosophilid gene vector from *Musca domestica*. Genetics. 1996; 142:907–914. [PubMed: 8849896]

Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol. 2015; 86:90–109. [PubMed: 25797922]

Pietrokovski S, Henikoff S. A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. Mol Gen Genet. 1997; 254(6):689–95. [PubMed: 9202385]

Rao DR, Mani TR, Rajendran R, Joseph AS, Gajanana A, Reuben R. Development of a high level of resistance to Bacillus sphaericus in a field population of *Culex quinquefasciatus* from Kochi, India. J Am Mosq Control Assoc. 1995; 11:1–5. [PubMed: 7616173]

Robertson HM. The *mariner* transposable element is widespread in insects. Nature. 1993; 362:241–245. [PubMed: 8384700]

Robertson HM, MacLeod EG. Five major subfamilies of *mariner* transposable elements in insects, including the Mediterranean fruit fly, and related arthropods. Insect Mol Biol. 1993; 2:125–139. [PubMed: 9087550]

Robertson HM, Asplund ML. Bmmar1: a basal lineage of the *mariner* family of transposable elements in the silkworm moth, *Bombyx mori*. Insect Biochem Mol Biol. 1996; 26:945–954. [PubMed: 9014339]

Robertson, HM. Evolution of DNA transposons in eukaryotes. In: Craig, NL.Robert Craigie, R.Gellert, M., Lambowitz, A., editors. Mobile DNA II. ASM Press; Washington, DC: 2002. p. 1093-1110.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010; 7(11): 909–12. [PubMed: 20935650]

Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987; 4:406–425. [PubMed: 3447015]

Sarkar A, Yardley K, Atkinson PW, James AA, O'Brochta DA. Transposition of the Hermes element in embryos of the vector mosquito, *Aedes aegypti*. Insect Biochem Mol Biol. 1997; 27(5):359–63. [PubMed: 9219363]

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH Jr. Many human L1 elements are capable of retrotransposition. Nat Genet. 1997; 16(1):37–43. [PubMed: 9140393]

Shao H, Tu Z. Expanding the diversity of the IS630-Tc1-*mariner* superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. Genetics. 2001; 159(3):1103–15. [PubMed: 11729156]

Sheen F, Lim JK, Simmons MJ. Genetic instability in *Drosophila melanogaster* mediated by hobo transposable elements. Genetics. 1993; 133(2):315–34. [PubMed: 8382175]

Siefert JL. Defining the mobilome. Methods Mol Biol. 2009; 532:13–27. [PubMed: 19271177]

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009; 19(6):1117–23. [PubMed: 19251739]

Sze SH, Dunham JP, Carey B, Chang PL, Li F, Edman RM, Fjeldsted C, Scott MJ, Nuzhdin SV, Tarone AM. A de novo transcriptome assembly of *Lucilia sericata* (Diptera: Calliphoridae) with predicted alternative splices, single nucleotide polymorphisms and transcript expression estimates. Insect Mol Biol. 2012; 21(2):205–21. [PubMed: 22283785]

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution. 2011; 28:2731–2739. [PubMed: 21546353]

Tu Z, Li S. Mobile genetic elements of malaria vectors and other mosquitoes. Mobile genetic elements in metazoan parasites Landes Biosci. 2009; 123

Wallau GL, Capy P, Loreto E, Hua-Van A. Genomic landscape and evolutionary dynamics of *mariner* transposable elements within the Drosophila genus. BMC Genomics. 2014; 15:727. [PubMed: 25163909]

Wang S, Lorenzen MD, Beeman RW, Brown SJ. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. Genome Biol. 2008; 9(3):R61. [PubMed: 18366801]

Warren WD, Atkinson PW, O'Brochta DA. The Hermes transposable element from the house fly, Musca domestica, is a short inverted repeat-type element of the hobo, Ac, and Tam3 (hAT) element family. Genet Res. 1994; 64(2):87–97. [PubMed: 7813905]

Weil CF, Kunze R. Transposition of maize Ac/Ds transposable elements in the yeast Saccharomyces cerevisiae. Nat Genet. 2000; 26(2):187–90. [PubMed: 11017074]

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007; 8(12):973–82. [PubMed: 17984973]

World Health Organization. World malaria report. 2015. http://apps.who.int/iris/bitstream/10665/200018/1/9789241565158_eng.pdf?ua=1

Xiong Y, Eickbush TH. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol. 1988; 5(6):675–90. [PubMed: 2464735]

Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci U S A. 2011; 108(19):7884–9. [PubMed: 21518873]

Yuan ZM, Zhang YM, Cai QX, Liu EY. High-level resistance to *Bacillus sphaericus* C3–41 in *Culex quinquefasciatus* from Southern China. Biocontrol Sci Tech. 2000; 10:41–49. 2000.

Abbreviations

CDS	Coding Sequence
Db	Database
LTR	Long Terminal Repeat
MITE	Miniature Inverted Repeat Transposable Elements
NLTR	Non-Long Terminal Repeat
ORF	Open Reading Frame
PSI-Blast	Position-Specific Iterated BLAST
RB	Repbase
RPS-Blast	Reverse Position-Specific BLAST
TE	Transposable Elements
TIR	Terminal Inverted Repeat
TSD	Terminal Site Duplications

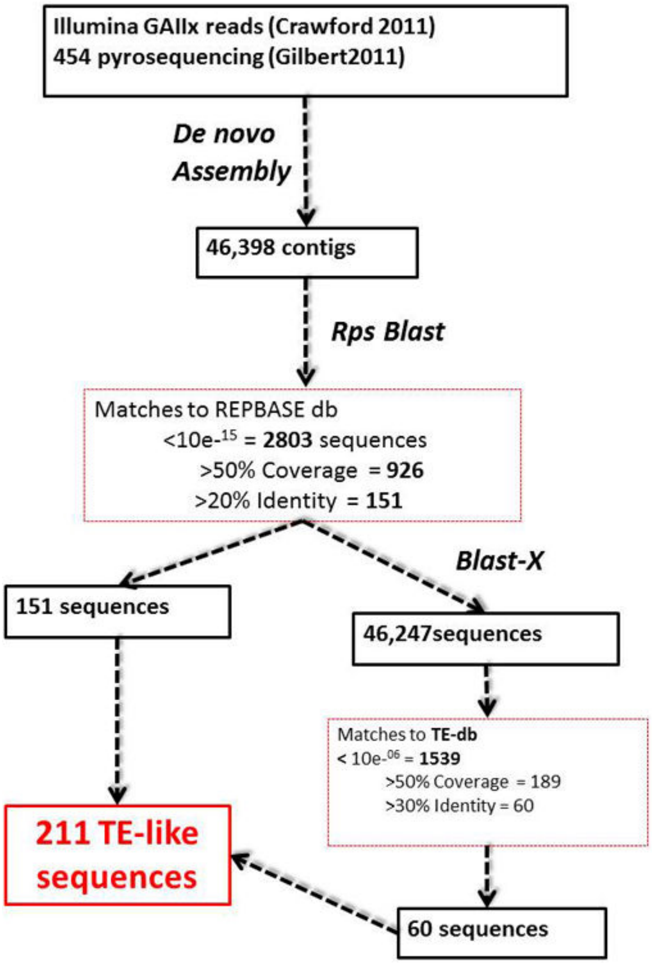

Figure 1.

Pipeline used for the identification of TE-like sequences in the transcriptome of *A. funestus*.

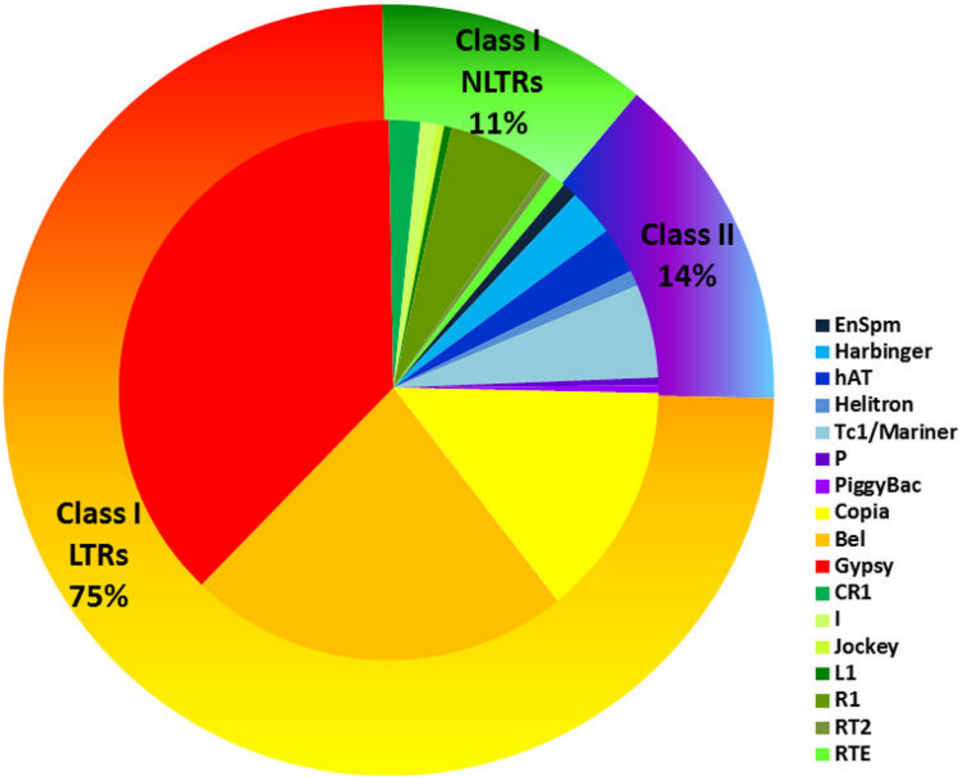

Figure 2. Distribution of TE-like sequences in the *A. funestus* transcriptome

The outer chart represents the three main classes/orders of TEs (LTRs, Non-LTRs and Class II) and the inner chart shows the distribution of TE superfamilies within each class/order. The figures are based on the 211 sequences that were characterized in the transcriptome.

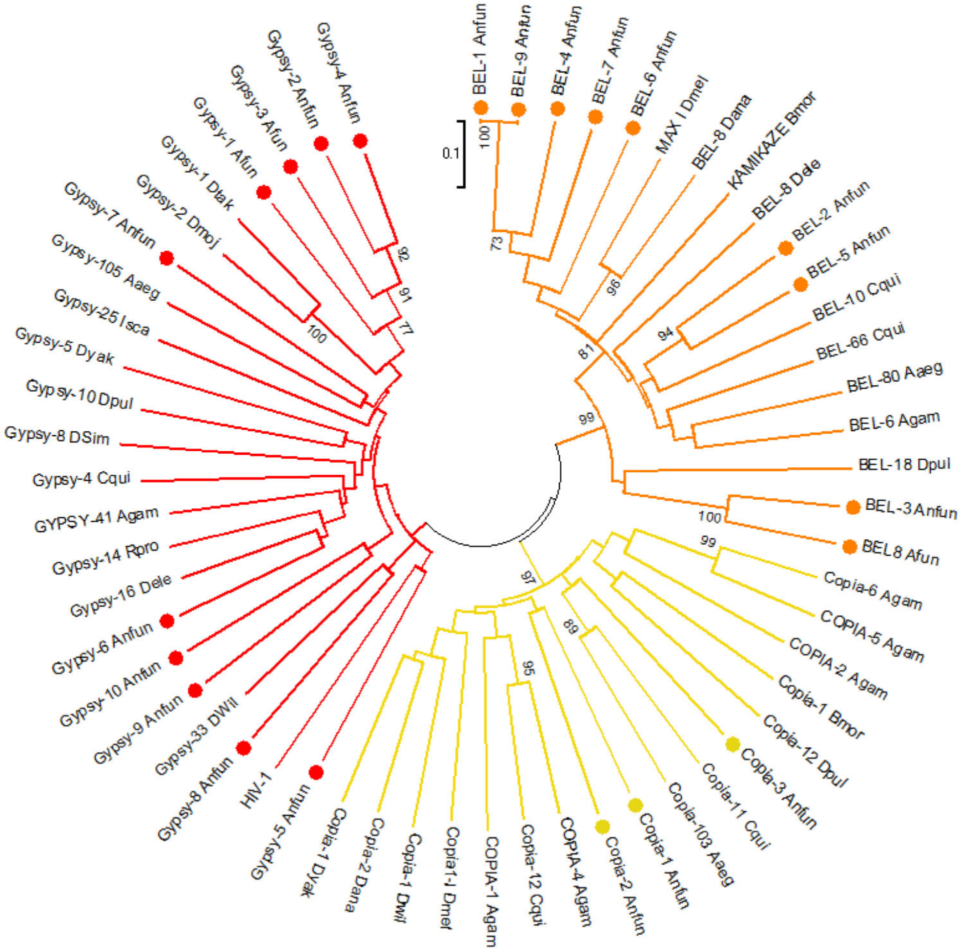

Figure 3. Phylogenetic relationships of LTR sequences from *A. funestus*

The phylogenetic relationships of 22 LTR sequences from *A. funestus* plus 35 reference sequences from other insect genomes (accession numbers in Table S3) including sequences from the *gypsy*, *copia*, *Bel-Pao,* and HIV, spanning the RT domain. The phylogeny was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 15.51243938 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 59 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 302 positions in the final dataset. The analyses were conducted in MEGA5 (Tamura *et al.* 2011). The numbers above the branches indicate the bootstrap value of a total of 1000 resamplings (only values higher than 70 are shown in the Figure). The different *LTR* superfamilies are coloured as follows: *Gypsy* in red, *Pao-Bel* in green and *Copia* in blue. The *A. funestus'* sequences are highlighted with colored dots corresponding to each superfamily.

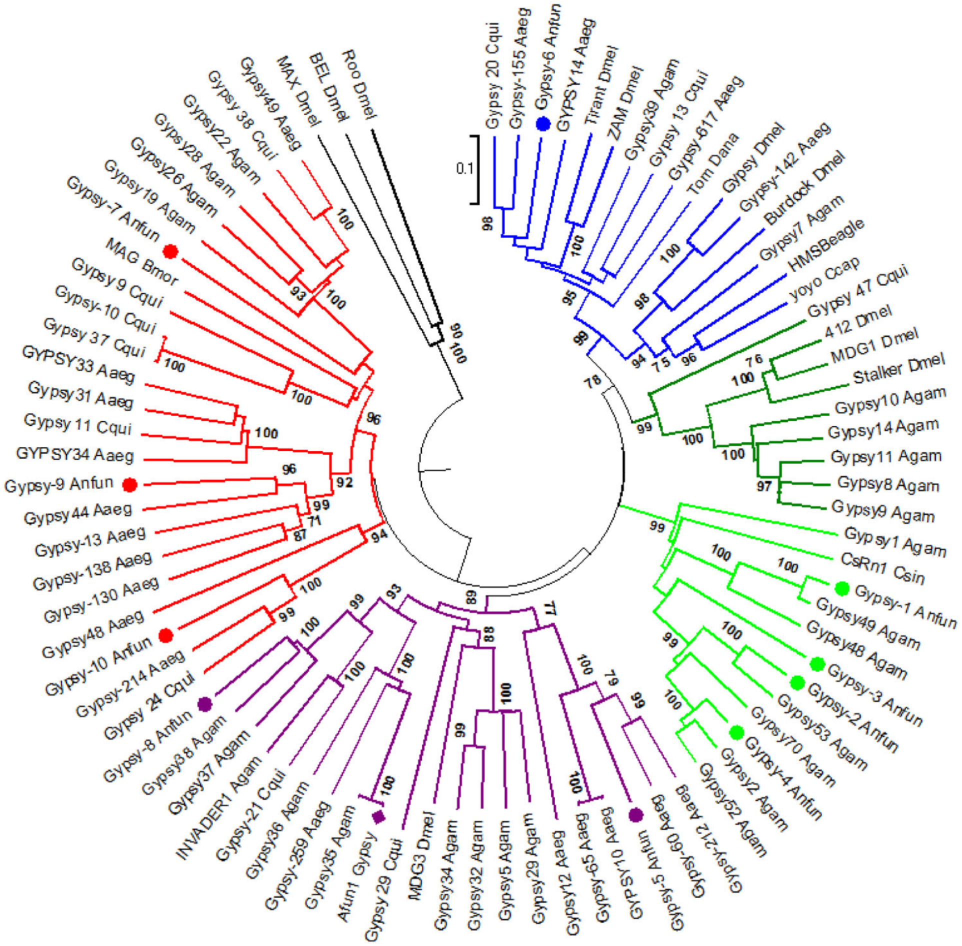

Figure 4. Phylogenetic relationships of *Gypsy* sequences from *A. funestus*

The phylogenetic relationships of ten sequences from *A. funestus* and 56 reference sequences from other insect genomes (accession numbers in Table S3). The phylogeny was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 15.77766579 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 85 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 334 positions in the final dataset. The analyses were conducted in MEGA5 (Tamura *et al.* 2011). The numbers above the branches indicate the bootstrap value of a total of 1000 resamplings (only values higher than 70 are shown in the Figure). The different *Gypsy* lineages are coloured as follows: blue for *Gypsy*, olive-green for *Mdg1*, light-green for *CsRn1*, purple for *Mdg3*, and red for the *Mag lineage*. The *A. funestus'* sequences are highlighted with colored dots corresponding to each lineage.

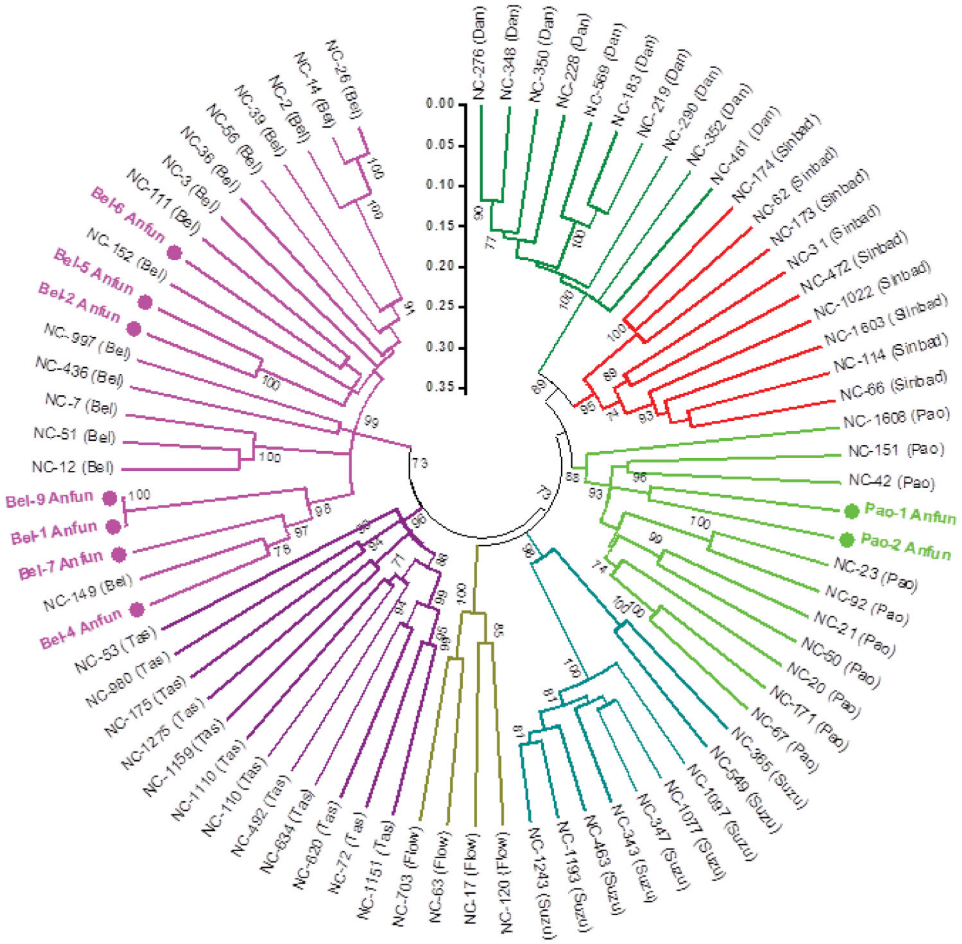

Figure 5. Phylogenetic relationships of *Bel/Pao* sequences from *A. funestus*

Phylogenetic relationships of nine sequences from *A. funestus* and 69 reference sequences from other insect genomes (accession numbers in Table S3). The phylogeny was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 17.56064770 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 78 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 241 positions in the final dataset. The analyses were conducted in MEGA5 (Tamura *et al.* 2011). The numbers above the branches indicate the bootstrap value of a total of 1000 resamplings (only values higher than 70 are shown in the Figure). The different *Pao/Bel* lineages are coloured as follows: green for *Dan*, red for *Simbad*, light-green for *Pao*, turquoise for *Suzu*, olive-green for Flow, purple for Tas and pink for Bel. The *A. funestus'* sequences are highlighted with colored dots corresponding to each lineage.

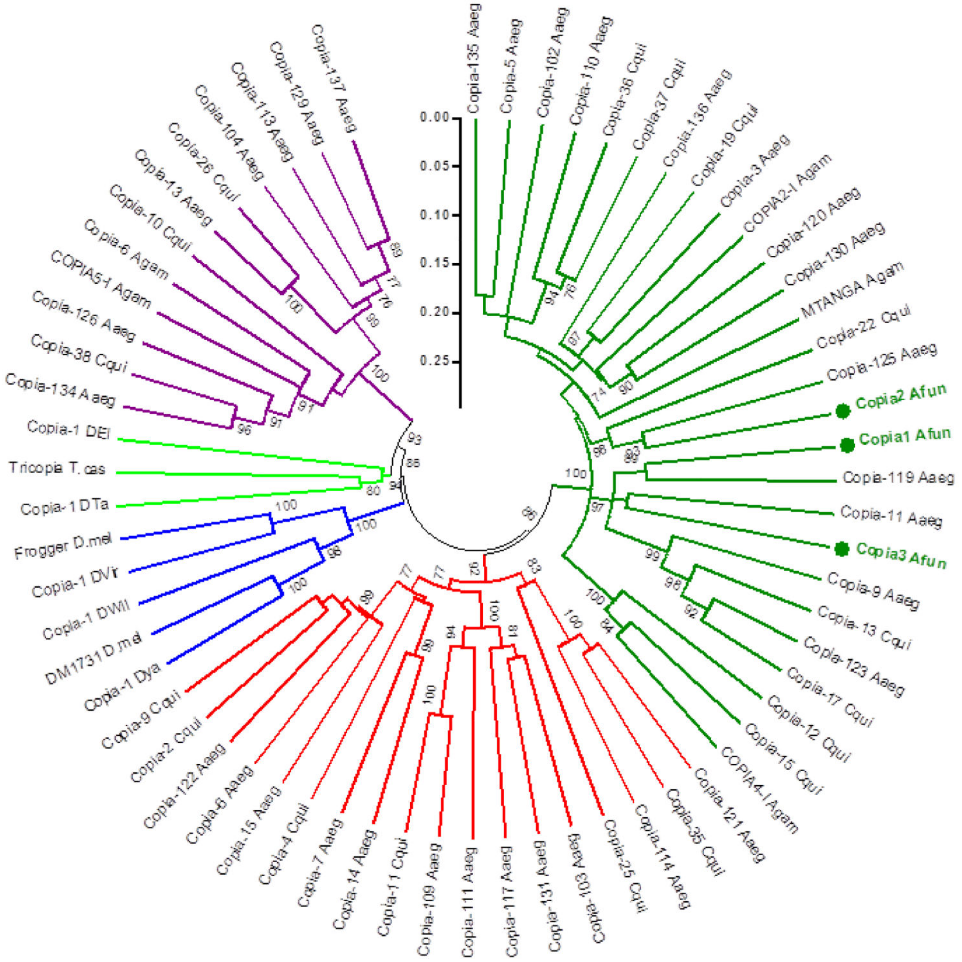

Figure 6. Phylogenetic relationships of *Copia* sequences from *A. funestus*

Phylogenetic relationships of three sequences from *A. funestus* and 62 reference sequences from other insect genomes (accession numbers in Table S3). The phylogeny was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 12.87295711 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 65 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 268 positions in the final dataset. The analyses were conducted in MEGA5 (Tamura *et al.* 2011). The numbers above the branches indicate the bootstrap value of a total of 1000 resamplings (only values higher than 70 are shown in the Figure). The different Copia lineages are colored. The *A. funestus'* sequences are highlighted with green colored dots.

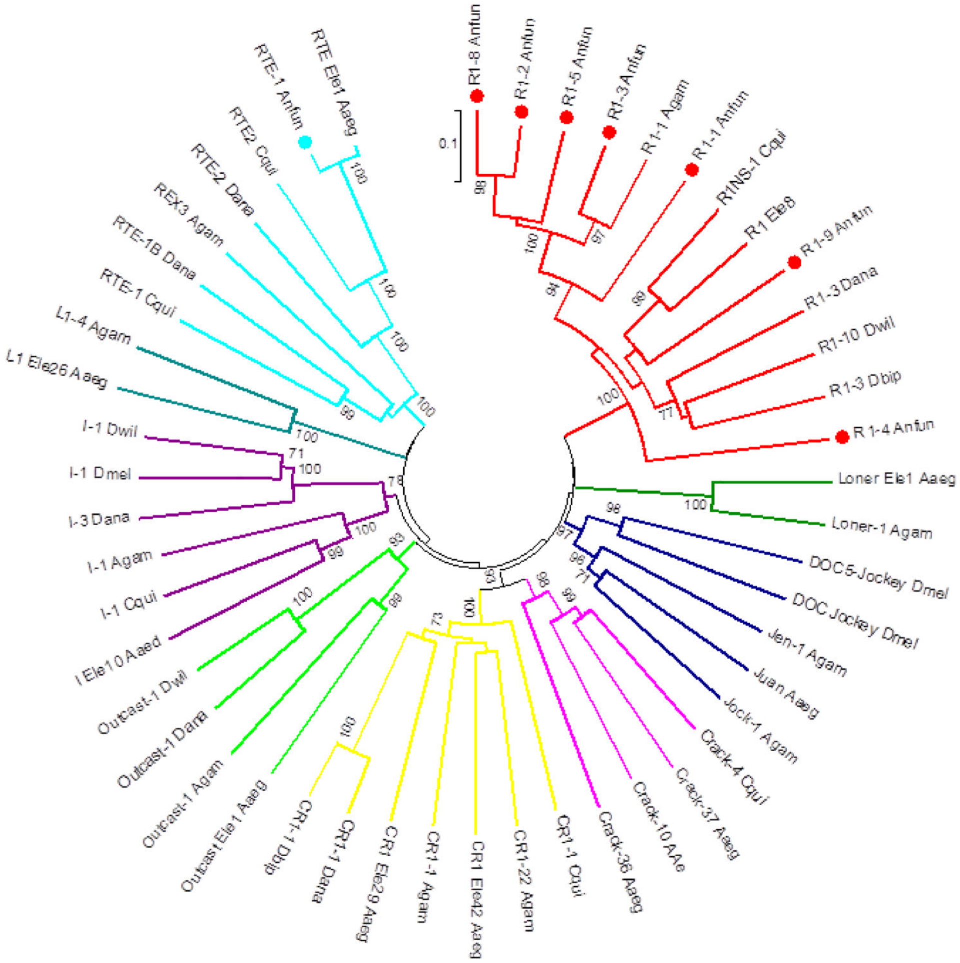

Figure 7. Phylogenetic relationships of *NLTRs* sequences from *A. funestus*

Phylogenetic relationships of seven sequences from *A. funestus* and 42 reference sequences from other insect genomes (accession numbers in Table S3). The phylogeny was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 12.45383147 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 50 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 316 positions in the final dataset. The analyses were conducted in MEGA5 (Tamura *et al.* 2011). The numbers above the branches indicate the bootstrap value of a total of 1000 resamplings (only values higher than 70 are shown in the Figure). The different *NLTR* superfmilies are coloured as follows: red for *R1*, green for *Lones*, blue for *Jockey*, pink for *Crack*, yellow for CR1, light-green for *Outcast,* purple for I, turquoise for L1 and light-blue for *RTE*. The *A. funestus'* sequences are highlighted with colored dots corresponding to each superfamily.

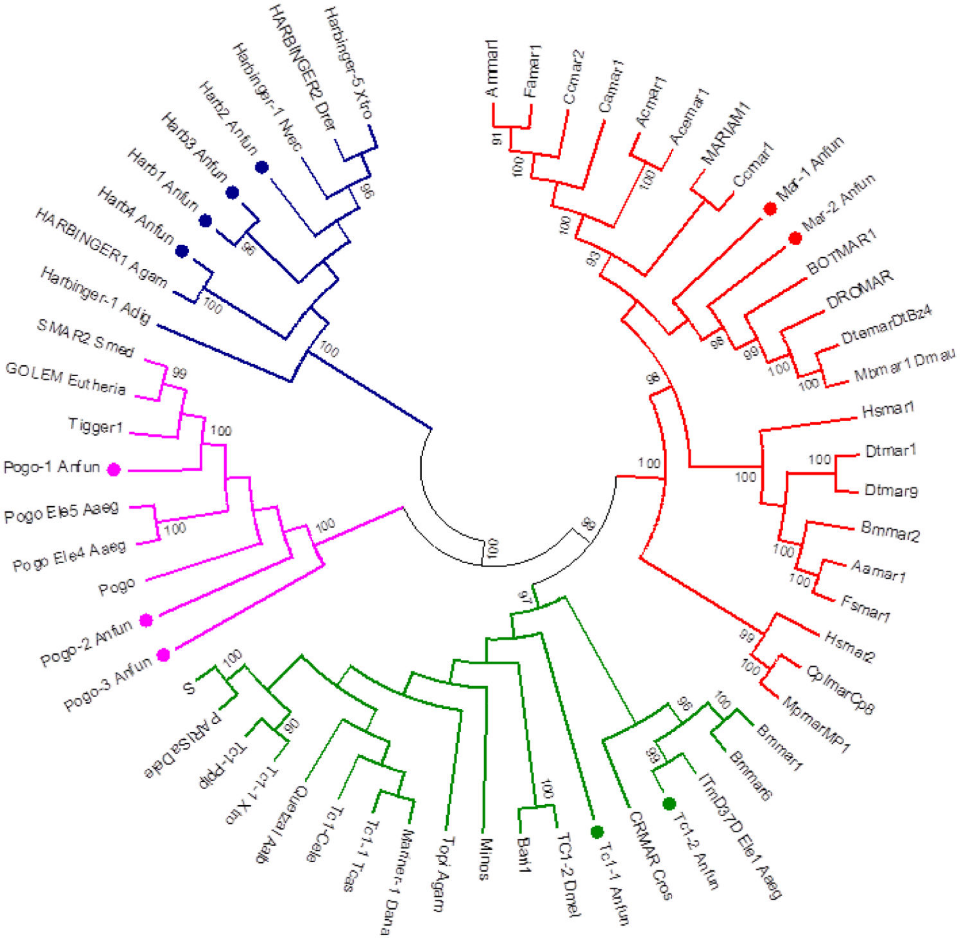

Figure 8. Phylogentic relationships of *DDE/D* sequences from *A. funestus*

Phylogenetic relationships of eleven sequences from *A. funestus* and 48 reference sequences from other insect genomes (accession numbers in Table S3). The phylogeny was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 14.37694765 is shown. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 59 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 285 positions in the final dataset. The analyses were conducted in MEGA5 (Tamura *et al.* 2011). The numbers above the branches indicate the bootstrap value of a total of 1000 resamplings (only values higher than 70 are shown in the Figure). The different *DDE/D* superfamilies are coloured as follows: red for *Mariners*, green for *Tc1*, pink for *pogo*, and blue for *Harbinger*. The *A. funestus'* sequences are highlighted with colored dots corresponding to each lineage.

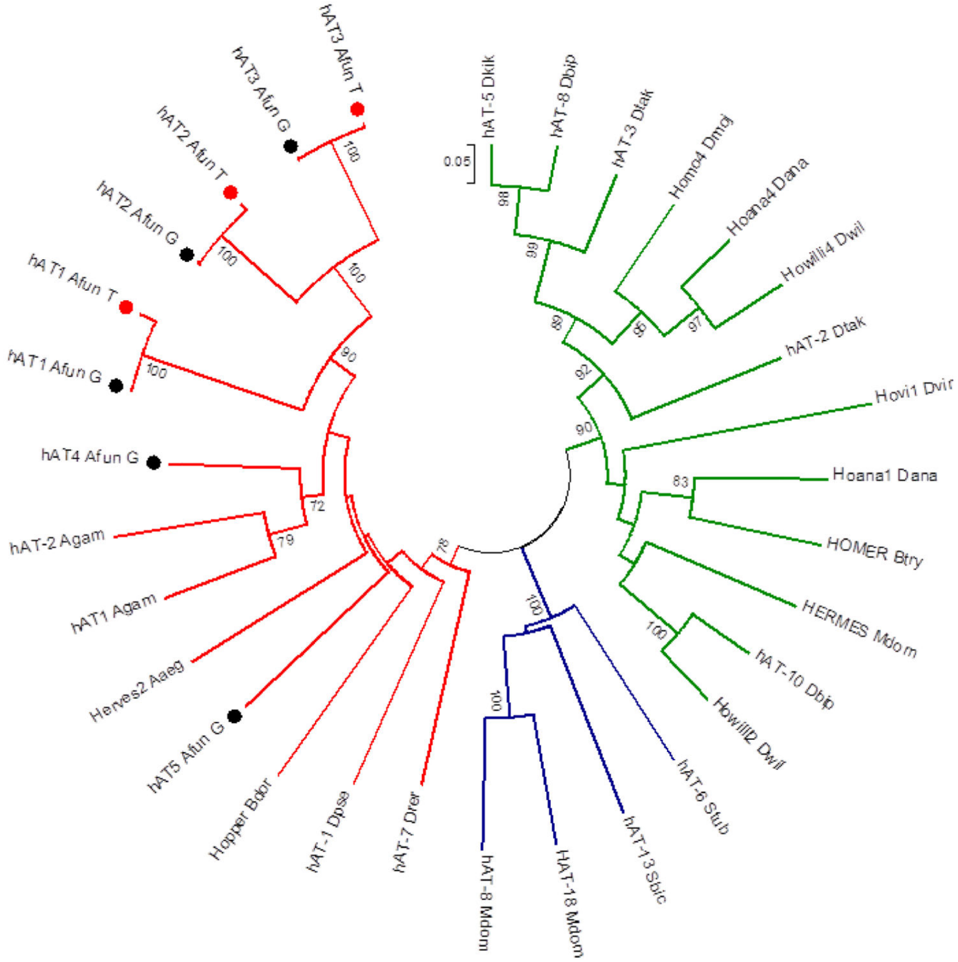

Figure 9. Phylogenetic relationships of *hAT* sequences from *A. funestus*

Phylogenetic relationships of three sequences from the *A. funestus* transcriptome and five sequences from the genome together with 23 sequences from other insect genomes (accession numbers in Table S3). The phylogeny was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 6.28002189 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 31 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 107 positions in the final dataset. The analyses were conducted in MEGA5 (Tamura *et al.* 2011). The numbers above the branches indicate the bootstrap value of a total of 1000 resamplings (only values higher than 70 are shown in the Figure). Different *hAT* lineages are coloured in red, blue and green. The *A. funestus'* sequences are highlighted with colored dots corresponding to each lineage, in red the sequenes isolated from the transcriptome (T) and in black the sequences obtained from the genome (G).

Table 1

Transposable Elements families identified in the transcriptome (this work) and genome (Neafsey et al. 2015) of *A. funestus*

	***A. funestus* transcriptome**	***A. funestus* genome**
Class II		
EnSpm	2	12
Harbinger	6	1
hAT	6	9
Helintron	2	3
Tc1/Mariner	12	10
P	1	3
PiggyBac	1	4
Transib	0	2
Tsessebeii	0	0
Pegassus	0	0
Academ	0	4
Class I		
LTRs		
Copia	30	26
Bel	49	103
Gypsy	78	77
NLTRs		
CR1	4	17
I	2	8
Ingi	0	1
Jockey	1	9
L1	1	7
L2	0	1
Loner	0	0
Outcast	0	7
R1	13	24
RT2	1	0
R4	0	0
RTE	2	6
TOTAL	211	334

Table 2

Amino-acids present in the DDE/D motif of the *Tc1/mariner* and *pogo* elements

Mar1_Afun	DD34D	$D_{92}D_{4}H_{27}P_{1}D$
Mar2_Afun	GD34N	$G_{92}D_{4}H_{27}P_{1}N$
Tc1-1_Afun	DD34E	$D_{85}D_{4}H_{27}P_{1}E$
Tc1-2_Afun	DD37D	$D_{91}D_{4}H_{30}P_{1}D$
Pogo1_Afun	DD32D	$D_{71}D_{4}H_{25}P_{1}D$
Pogo2_Afun	AN45D	$A_{104}N_{4}H_{24}P_{15}D$
Pogo3_Afun	DD45D	$D_{109}D_{4}H_{27}P_{15}D$