Gene 440 (2009) 42–49

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*

Crescenzio Francesco Minervini, Luigi Viggiano, Ruggiero Caizzi, Renè Massimiliano Marsano *

Università degli Studi di Bari, Dipartimento di Genetica e Microbiologia — DI.GE.MI. Via Amendola 165/A, Bari, Italy

ARTICLE INFO

Article history:
Received 4 January 2009
Received in revised form 27 March 2009
Accepted 31 March 2009
Available online 9 April 2009

Received by N. Okada

Keywords:
Aedes aegypti
Bel-Pao retrotransposons Ty3-gypsy retrotransposons
In silico genomic analysis
LTR_Struc program
Genes–retrotransposons interaction

ABSTRACT

We have detected seventy-six novel LTR retrotransposons in the genome of the mosquito *Aedes aegypti* by a genome wide analysis using the LTR_STRUC program. We have performed a phylogenetic classification of these novel elements and a distribution analysis in the genome of *A. aegypti*. These mobile elements belong either to the Ty3/gypsy or to the Bel family of retrotransposons and were not annotated in the mosquito LTR retrotransposon database (TEfam). We have found that ~1.8% of the genome is occupied by these newly detected retrotransposons that are distributed predominantly in intergenic genomic sequences and introns. The potential role of retrotransposon insertions linked to host genes is described and discussed. We show that a retrotransposon family belonging to the Osvaldo lineage has peculiar structural features, and its presence is likely to be restricted to the *A. aegypti* and to the *Culex pipiens quinquefasciatus* genomes. Furthermore we show that the ninja-like group of elements lacks the Primer Binding Site (PBS) sequence necessary for the replication of retrotransposons. These results integrate the knowledge on the complicate genomic structure of an important disease vector.

1. Introduction

Eukaryotic genomes are largely composed of transposable elements (TE). These elements are classified in two main classes (class I and class II) according their transposition mechanisms (reviewed in Finnegan, 1992). Class II elements are characterized by DNA to DNA transposition using of a self encoded transposase. Class I elements use an RNA intermediate, which is reverse transcribed into cDNA molecules and then inserted in the genome. Class I elements can be further categorized in LTR- and non-LTR retrotransposons depending on the presence or absence of terminal direct repeats. Completely sequenced genomes facilitate the characterization of the full transposon complement in a genome. This is possible both with a sequence similarity search analyses (extrinsic methods) using characterized mobile elements from related model organisms as query and with the development of *in silico* methods that focus on the structure (intrinsic methods) of TEs rather than the sequence similarity. The latter methods allow a faster identification of mobile elements that have a low sequence similarity with respect to reference elements. This strategy has been successfully applied to the identification of L1 insertions in the human genome (Szak et al., 2002), LTR retrotransposons insertions in *A. gambiae* (Marsano and Caizzi, 2005) and *Mus musculus* genomes (McCarty and McDonald, 2004), and MITEs (*M*iniature *I*nverted repeat *T*ransposable *E*lements) in the *A. gambiae* genome (Tu, 2001).

Aedes aegypti is the primary mosquito vector responsible for the transmission of both the yellow fever and dengue viruses. Recently Nene et al. (2007) have revealed that nearly 50% of its genome consists of transposable elements. LTR retrotransposons built up about 10.5% of the *A. aegypti* genome. Furthermore an extensive compilation of mobile elements has been reported and a relational database called TEfam (http://tefam.biochem.vt.edu/tefam) was released. Here the sequences of more than one thousand of mosquitoes TE families have been annotated. More than 800 families of the TEs reported in the TEfam database are related to *A. aegypti* retrotransposons and 642 belong either to the Ty3/gypsy (179 elements), Ty5/copia (233 elements) or Bel/Pao (230 elements) families. In addition, six distinct phylogenetic lineages can be recognized within the Ty3/gypsy family (namely the gypsy lineage (21 elements), Mag lineage (64 elements), CsRn1 lineage (15 elements), mdg1 lineage (26 elements), Osvaldo lineage (30 elements) and mdg3 lineage (23 elements).

The massive presence of transposable elements in the genome of *A. aegypti* is consistent with two observations. First, *A. aegypti* genome is 4-fold larger than *A. gambiae* genome: this must be taken into account when studying repetitive sequences from *A. aegypti*. Second, *A. aegypti's* introns are on average longer than introns of related species due to the presence of transposable elements (Nene et al., 2007).

Here we report 76 additional LTR retrotransposon elements in the genome of *A. aegypti*, identified using the LTR_STRUC program (McCarthy and McDonald, 2003). We have performed classification on the basis of evolutionary relationships with other LTR

Abbreviations: LTR, long terminal repeat; PBS, Primer Binding Site; TSD, target site duplication; TE, transposable element; GAG, group specific antigen; PRO, protease domain; RT, reverse transcriptase domain; INT, integrase domain; ENV, envelop protein.

* Corresponding author. Tel.: +39 80 5443338; fax: +39 80 5443386.
E-mail address: r.marsano@biologia.uniba.it (R.M. Marsano).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.03.021

retrotransposons. We have also analyzed the structure and the genomic distribution of the new elements detected. A novel family belonging to the Osvaldo lineage with unexpected structural features has been identified. Furthermore all members of the ninja group identified in this study lack a discrete PBS sequence (Primer Binding Site). The results of the genomic distribution analysis are consistent with the presence of retrotransposons preferentially in intergenic regions of the genome of *A. aegypti* or in intron sequences. The possible functional role of some insertions on the host gene organization is also discussed.

2. Materials and methods

2.1. LTR_STRUC analysis and classification of LTR retrotransposons

The entire genome of *A. aegypti* was downloaded from the Broad Institute website (http://www.broad.mit.edu/index.html) and scanned with the LTR_STRUC program (McCarthy and McDonald, 2003) using the default parameters. 4026 putative retrotransposon sequences obtained as output were subjected to an "all against all" BLAST in order to group sequences with % identity greater than 98% over a sequence of at least 1 Kb. Two hundred and seventeen groups (containing at least 2 sequences) and 359 singlets (i.e. containing a single sequence) were obtained after this step. The final subset of LTR retrotransposons was then blasted against the TEfam database in order to define families of elements and to detect previously not annotated sequences.

Criteria for defining LTR retrotransposons were identical to the previously described criteria adopted during *A. aegypti* TE analysis (Nene et al., 2007). Briefly sequences of the Ty3/gypsy LTR retrotransposons are considered as belonging to the same element if they share at least 85% nucleotide identity along at least 400 bp in their coding region. Ty1/copia sequences that share at least 85% identity at the nucleotide level over at least 1000 bp are considered belonging to the same element. Copies of Pao/Bel retrotransposons are considered as belonging to the same element if they show at least 70% identity at the nucleotide level in their coding sequences.

The names assigned to retrotransposons follow the nomenclature adopted in the Repbase database (Jurka, 2000).

2.2. Analysis of insertions

The ORF finder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) was used to determine the ORF structure and number of each element.

The TSD (Target Site Duplicated upon insertion) and the length of the LTRs of each element obtained were determined by visual inspection of sequences. In absence of a reported list of the tRNA gene sequences in *A. aegypti* the PBS sequences were determined by comparison of a tRNA dataset of *Drosophila melanogaster* at the http://lowelab.ucsc.edu/GtRNAdb/Dmela/ website.

To detect retrotransposon insertions near (or overlapping) host genes, a BLAST search at the Ensembl database (http://www.ensembl.org/Aedes_aegypti/blastview) was performed using the following criteria: 1) only insertions with average similarity greater than 85% were counted; 2) insertions shorter than 180 bp were not counted. These criteria allow the detection of full-length elements and elements carrying deletion without missing solo LTR and prevent misleading results coming from low quality alignments.

2.3. RepeatMasker analysis

RepeatMasker software (version 3.2.5) was used to estimate the retrotransposons occupancy as percent of the genome fraction. Repeats search was performed using Cross_Match as sequence search engine. A repeats library was built for each of the LTR retrotransposon group described in the Results section (gypsy, BEL, ninja) and was used to scan the genome sequence separately. Scanning was carried out using a cutoff value of 250.

2.4. Multiple sequence alignment and phylogenetic analysis

As described previously (Malik and Eickbush, 1999) a better way to reconstruct phylogeny of retroelements is to perform multiple alignment of RT-RnaseH-INT domains. These domains of each putative active element were extracted from the translated ORF encoding the POL polyprotein and used to reconstruct the phylogenetic history of *A. aegypti* gypsy-like retrotransposons. We have no evidence of domain swapping by performing multiple alignment using RnaseH, RT or INT domains (data not shown). ClustalX (Thompson et al., 1997) was used to perform multiple alignments. After a manual check of the alignments Neighbor-joining and bootstrap analyses were performed using Treecon v.1.3b (Van de Peer and De Wachter, 1994). Trees were visualized with Treeview (Page, 1996). As reference, previously described elements in other species (EMBL accession nos.: DS36733 (reverse transcriptase), DS36732 (RNaseH), and DS36734 (integrase), (Malik and Eickbush, 1999)) were used to establish relationships between *A. aegypti* retroelements.

3. Results

We screened the *A. aegypti* genome with the LTR_STRUC program (McCarthy and McDonald, 2003), we have obtained 4026 putative retrotransposon sequences as output. These sequences were arranged in more than 200 groups of sequences sharing 100% identity over at least 99% of the sequence alignment. For each group we chose one representative sequence potentially able to encode protein domains of retrotransposons (GAG PRO RT RH INT). These sequences were used to probe the TEfam and the REPBASE databases (Jurka, 2000).

A small number of transposable element sequences of *A. aegypti* are annotated in the REPBASE database and only two elements match our LTR_STRUC output (namely AACOPIA1 and ZEBEDEE).

The comparison with TEfam database indicated that a great number of the sequences identified in this study have been yet annotated. However 76 sequences did not found a match in TEfam. In both cases the comparison was performed adopting the criteria reported in the Methods section.

Each novel retrotransposon identified was assigned to a specific lineage of LTR retrotransposons on the basis of the evolutionary relationships with known LTR retrotransposons from different organisms. The putative RT-RnaseH-INT domains were aligned with the corresponding domains of reference LTR retrotransposons identified in different organisms.

As shown in Fig. 1, twenty-two out of seventy-six elements fall into the Ty3/gypsy group while fifty-four out of seventy-six elements fall into the BEL/Pao group. No novel Ty1/copia element has been identified in this study.

3.1. Structural features of the Ty3/gypsy group of LTR retrotransposons

The twenty-two Ty3/gypsy retrotransposon elements described in this study fall into four out of the nine lineages described for the Ty3/gypsy group as result of the phylogenetic comparative analysis (Fig. 1). We have performed a structural analysis of the novel elements (Table 1). For a representative element of each family we have determined the LTR length, the sequence of the Primer Binding Site (PBS), the Target Site Duplicated (TSD) upon insertion and the position within the contig where it resides. The PBS analysis was carried out using the *D. melanogaster* tRNA database. Due to the great evolutionary distance between *Drosophila* and *Aedes* we have compared the tRNA genes of *D. melanogaster* with those of *A. aegypti* assessing that they are identical, especially in their 3′ sequence.

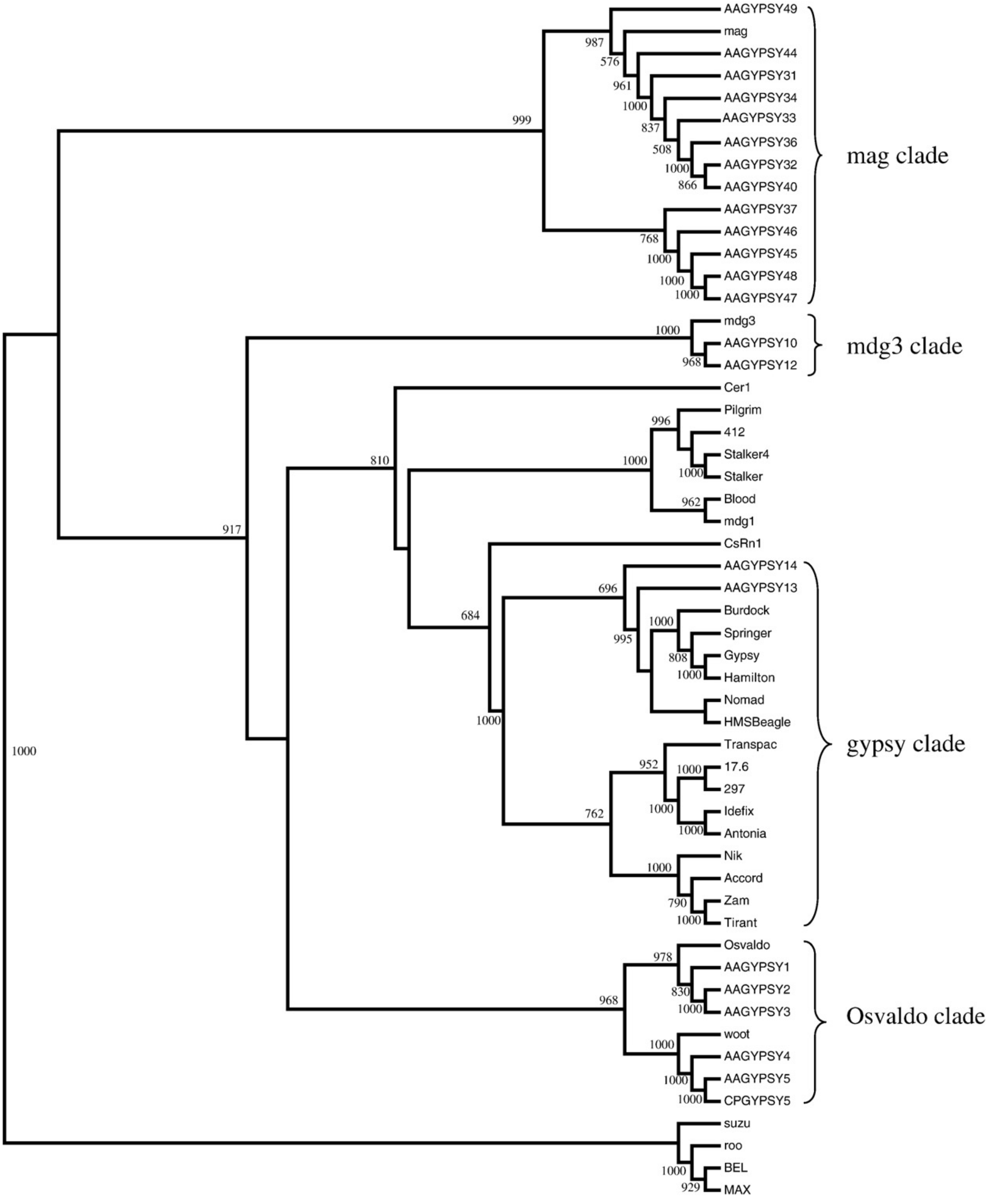

Fig. 1. Phylogenetic tree of the *Ty3*/*gypsy* elements. Phylogenetic relationships of the gypsy-like retrotransposons based on the amino acids alignment of the conserved RT, RNase H and INT domains. The four clades in which fall retrotransposons detected in this paper are indicated. Elements from this study are indicated as AAGYPSY followed by a number. CPGYPSY5 is a LTR retrotransposon identified in the genome of *C. pipiens quinquefasciatus*. The N-J bootstrap values supporting the internal branches are indicated at the nodes. Only bootstrap values greater than 500 are reported. The tree is outgrouped with Bel-like elements.

The structural analysis revealed that the main features of each lineage are conserved in the elements that we have detected in this study, supporting the results of the phylogenetic analysis.

Three of the novel elements identified are related to *Osvaldo*, a retrotransposon presenting very long LTR (Pantazidis et al., 1999). Elements AAGYPSY1, 2 and 3 contain LTRs longer than 1 Kbp. They also contain a putative PBS, similar to the 3′ end of the tRNA^{Lys} of *D. melanogaster*; the same sequence is also conserved in the *Osvaldo* element of *D. buzzatii* (Pantazidis et al., 1999) and *Ulysses* element of *D. virilis* (Evgen'ev et al., 1992). AAGYPSY1 and 2 contain two ORFs while AAGYPSY 3 contains a single ORF which encode for putative GAG and POL proteins.

An unexpected result was obtained from the analysis of elements AAGYPSY4 and AAGYPSY5. They are phylogenetically related to *woot*, a member of the Osvaldo lineage, (Beeman et al., 1996) but they have short LTRs (210 bp and 257 bp long respectively) and a PBS

Table 1
Structural features of the LTR retrotransposons identified in this paper.

Lineage	Name	Length	ORFs	PBS	LTR	TSD	Supercontig	Position
Osvaldo	AAGYPSY1	11062	2	Lys	1837/1831	atat	1.131	882160–893221
Osvaldo	AAGYPSY2	9324	2	Lys	1059	gtattg	1.108	2145350–2154673
Osvaldo	AAGYPSY3	9998	1	Lys	1175	acac	1.134	11880–21877
Osvaldo	AAGYPSY4	5719	1	Pro	210	acca	1.224	1163771–1169489
Osvaldo	AAGYPSY5	5600	1	Pro	257	actg	1.143	702637–708236
Mdg3	AAGYPSY10	4984	1	Ser	267	ctag	1.236	1363511–1368494
Mdg3	AAGYPSY12	5794	1	Leu	245	ttct	1.125	1911380–1917173
Gypsy	AAGYPSY13	6980	3	Thr	304	cgcg	1.345	330425–337404
Gypsy	AAGYPSY14	6423	3	Leu	269	$(tat)_n$	1.447	720621–727039
Mag	AAGYPSY31	5294	1	Ser	234	tatca	1.139	363910–358617
Mag	AAGYPSY32	5519	1	Ser	233	atcac	1.16	995017–1000535
Mag	AAGYPSY33	5458	1	Ser	180	acgcc	1.139	2103418–2097959
Mag	AAGYPSY34	5779	1	Ser	193	gcccc	1.8	486984–492762
Mag	AAGYPSY36	5837	1	Ser	252	gaacg	1.475	94292–88455
Mag	AAGYPSY37	5041	1	Ser	193	actat	1.574	590078–595118
Mag	AAGYPSY40	5951	1	Ser	249	atcag	1.251	921052–915102
Mag	AAGYPSY44	5116	1	Ser	292	gttct	1.110	2347577–2352692
Mag	AAGYPSY45	4662	1	Leu	145	ccacg	1.195	1470848–1475509
Mag	AAGYPSY46	5087	1	Leu	186	gtaca	1.446	797459–792383
Mag	AAGYPSY47	4879	1	Leu	202	ctgag	1.804	198212–193334
Mag	AAGYPSY48	4918	1	Leu	204	catac	1.54	1045256–1050173
Mag	AAGYPSY49	4754	1	ND	235	gactg	1.321	724580–720061
Ninja	AABEL2	7125	1	Leu?	577	taggc	1.932	200044–207168
Ninja	AABEL3	7654	1	ND	501	ctagg	1.120	1590262–1597915
Ninja	AABEL4	7697	1	Ser?	397	cagtg	1.453	479198–486954
Ninja	AABEL7	6928	1	ND	491	aggcc	1.959	118596–111669
Ninja	AABEL8	7634	1	Ala?	698	gctta	1.909	266142–273812
Ninja	AABEL9	7670	1	Arg?	688	accgg	1.118	1504428–1496759
Ninja	AABEL10	7221	1	Ala?	674	acggg	1.744	496002–503222
Ninja	AABEL11	7137	1	Ser?	654	acagg	1.208	1309084–1301946
Ninja	AABEL13	7596	1	Leu?	531	cccat	1.101	1970197–1962636
Ninja	AABEL14	7771	1	ND	626	ttcac	1.198	78406–70636
Ninja	AABEL15	7442	1	ND	648	agtgc	1.127	1983984–1991425
Bel	AABEL19	6007	1	Tyr	296	cttgg	1.223	248512–254518
Bel	AABEL20	5952	1	Tyr	293	gttag	1.1322	9898–15849
Bel	AABEL21	5856	1	Tyr	244	tatgg	1.135	1111684–1117539
Bel	AABEL22	8300	1	Tyr	712	ccagg	1.512	650055–658354
Bel	AABEL23	8339	1	Tyr	811	aatag	1.5	1524724–1533062
Bel	AABEL25	7929	2	Tyr	526	atatc	1.641	546095–554023
Bel	AABEL26	6947	1	Tyr	460/461	aataa	1.431	466092–473038
Bel	AABEL27	6868	1	Tyr	527	cacat	1.61	2269743–2276610
Bel	AABEL28	6796	1	Tyr	518	ggatt	1.29	572774–579569
Bel	AABEL29	6823	1	Tyr	474	gaaat	1.40	1559041–1565863
Bel	AABEL30	7397	1	Tyr	597	cagtt	1.317	121435–128781
Bel	AABEL31	8150	1	Tyr	674/662	acagg	1.67	423351–431500
Bel	AABEL33	7534	1	His	549	atgta	1.273	125574–133107
Bel	AABEL34	8349	1	His	396	gcatt	1.270	659565–667913
Bel	AABEL36	6796	1	His	393	accgc	1.285	1284723–1291517
Bel	AABEL37	7480	1	His	528	atatc	1.131	1622866–1630345
Bel	AABEL39	6588	1	Phe	391	nd	1.134	2086697–2093285
Bel	AABEL40	6546	1	Phe	395	gagtg	1.1	881207–887752
Bel	AABEL41	6596	1	Phe	356	ggttt	1.1267	64708–58113
Bel	AABEL42	6545	1	Phe	408	ND	1.220	1674929–1681473
Bel	AABEL43	8597	1	Tyr	332	gttct	1.250	1135943–1144539
Bel	AABEL45	7913	1	Tyr	336	gtggc	1.26	816914–824826
Bel	AABEL47	7682	1	Tyr	361	tacac	1.134	1618714–1626395
Bel	AABEL49	6282	1	ND	364	gtcag	1.69	1022031–1015750
Bel	AABEL50	6617	1	Tyr	466/473	cagag	1.369	391674–398190
Bel	AABEL51	6578	1	Tyr	426	gtata	1.209	467214–473790
Bel	AABEL52	6842	1	Tyr	433	gcatg	1.30	301192–308033
Bel	AABEL53	6493	1	Tyr	521	ND	1.9	2101819–2108311
Bel	AABEL54	6589	1	Tyr	444	atcgc	1.640	2715–9303
Bel	AABEL56	7085	1	Tyr	624	gtcgt	1.19	3829270–3836354
Bel	AABEL57	6401	1	Tyr	337	acgcc	1.380	1009668–1016068
Bel	AABEL58	6784	1	Tyr	408	ctcgc	1.442	190289–197072
Bel	AABEL59	7040	1	Tyr	752	tcacc	1.28	3296947–3303986
Bel	AABEL60	6519	1	Tyr	473	actta	1.145	789731–796249
Bel	AABEL61	6373	1	Tyr	357	gttc	1.141	378454–384826
Bel	AABEL62	6314	1	Tyr?	225	tgttt	1.621	619698–626011
Bel	AABEL63	6726	1	Tyr	408	ccgtg	1.450	260297–267022
Bel	AABEL64	7405	1	Tyr	207	atatc	1.315	1270790–1278194
Bel	AABEL65	5936	1	Tyr	195	gtcac	1.3	742033–747968
Bel	AABEL66	6611	1	Tyr	517	catat	1.450	668553–675163
Bel	AABEL67	6694	1	Ser	726	ccagg	1.277	1013606–1020299

(*continued on next page*)

Table 1 (*continued*)

Lineage	Name	Length	ORFs	PBS	LTR	TSD	Supercontig	Position
Bel	AABEL68	6232	1	Ser	482	gtgca	1.943	171729–177960
Bel	AABEL69	7604	1	ND	705	ctcac	1.985	43824–51430

For each family are indicated the major lineage they belong to, the element length, the number of ORFs, the target site duplication (TSD), Primer Binding Site (PBS), the LTR length and the position within the supercontig of a representative member for each family.
A question mark denotes an ambiguous PBS sequence.
Two values are reported when the two LTRs differ in size.
ND = not determined.

complementary to the tRNA^{Pro} of *D. melanogaster*. Both structural features are not shared with other known members of the Osvaldo lineage. This unexpected result prompted us to BLAST the genomes of insects species related to *A. aegypti* in order to verify the presence of LTR retrotransposons with the same structural and evolutionary features described for AAGYPSY4 and AAGYPSY5. The top result of the TBLASTX search against the genome of *Culex pipiens* using AAGYPSY5 as query is contig cont3.22711 (GenBank accession number: NZ_AAWU01022711.1). A LTR retrotransposon 5097 bp long was identified (named CPGYPSY5) which has 236 bp long LTRs and has a single ORF encoding a putative GAG-POL polyprotein. Its PBS sequence is complementary to the tRNA^{Pro} of *D. melanogaster*. Full length and defective copies of this element exist in the genome of *C. pipiens*. Cross control TBLASTX analysis against insect genomes using CPGYPSY5 results in the *woot* retrotransposon within the top scoring alignments. The CPGYPSY5 retrotransposon of *C. pipiens* is indeed a member of the woot clade as shown in Fig. 1.

Two novel gypsy elements were identified in this study. The structural analysis have revealed that the PBS of these elements follow the general rule of other members of the *gypsy* lineage identified in other organisms: the first base of the PBS overlaps the last base of the 5′ LTR (Inouye et al., 1986).

Several members of the *gypsy* lineage identified so far in other organisms contain an ORF that could potentially encode for the envelope protein (ENV), a typical retrovirus like protein reported to be important in the horizontal transmission process (Syomin et al., 2002). The two gypsy-like elements detected in this study also contain a canonical ENV-coding ORF.

Thirteen retrotransposon families of *A. aegypti* belong to the mag lineage. Members of this lineage have been identified in different insect genomes: *Bombyx mori* (Michaille et al., 1990), *A. gambiae*, *D. melanogaster* (Tubio et al., 2004) and *C. elegans* (Malik et al., 2000; Ganko et al., 2001). Eight elements cluster with the *mag* element and five elements are basal to this clade (Fig. 1). The PBS of the mag-like elements is complementary either to the tRNA^{Leu} or to the tRNA^{Ser}.

Two elements phylogenetically related to the *mdg3* element of *D. melanogaster* have been identified. The structural analysis of these elements reveals that while AAGYPSY10 has a single ORF encoding the GAG-POL polyprotein that is the typical ORF structure of the *mdg3* like elements, AAGYPSY12 encodes GAG and POL proteins on independent ORFs. The PBS sequence of elements belonging to this lineage is complementary to tRNA^{Leu} or tRNA^{Ser} and, as suggested by Saigo (1986), these elements might use a 3′-truncated tRNA (i.e. without terminal CCA) to prime the reverse transcription process.

3.2. The Bel-Pao family

Fifty-four new retrotransposons belonging to the Bel-Pao family were found in the genome of *A. aegypti*. Phylogenetic analysis performed with POL protein of the representative elements allows differentiating two main lineages, *Bel* lineage and *ninja* lineage that have been shown to be distinct within the BEL-Pao family. Our reference elements were already known members of either lineage. We have used Bel-like and ninja-like elements from different eukaryotic organisms retrieved from the RepBase database (including *Anopheles gambiae* and *D. melanogaster*). ninja-like and Bel-like elements selected from *A. gambiae* have been assigned to the ninja or to the BEL lineages in previous analyses (Marsano and Caizzi, 2005).

As shown in Fig. 2 the ninja lineage is clearly distinct from the Bel lineage. The ninja lineage is composed of thirteen LTR retrotransposons identified in this study while the BEL lineage is composed of 41 sequences detected in *A. aegypti*; several clades observed in Fig. 2 are supported by high bootstrap values allowing the determination of the relationships between LTR retrotransposons of *A. aegypti* and *A. gambiae*. All the elements identified in the Bel lineage have a single ORF encoding all the protein domains of LTR retrotransposons (GAG PR-RT-RH-IN). The evolutionary relationships are supported by PBS analysis; indeed, members of each clade observed in the Bel lineage homogeneously use the same initiator primer. Furthermore reference elements of *A. gambiae* and *Drosophila* have the same PBS detected in the retrotransposons identified in *A. aegypti*; this observation suggests a conservation of this structural feature during the evolution of insects.

However, this observation is not true for the elements belonging to the ninja lineage, as their PBS sequence cannot be detected in many cases. Only for seven ninja-like elements it is possible to identify a sequence somehow similar to the 3′ end of a tRNA of *D. melanogaster* (see Table 1). However the similarity observed is weak and do not allow to assign a PBS unequivocally.

3.3. The genomic distribution analysis of the novel retrotransposons

We have performed a genomic distribution analysis using BLAST and RepeatMasker (Smit et al., 1996–2004). RepeatMasker allows a rapid estimation of the genomic fraction occupied by the sequences analyzed. The analysis was performed separately for the gypsy, ninja and Bel-like elements. For each group of retrotransposons (i.e. gypsy, ninja and Bel), software returned the *A. aegypti* sequence masked. The genome fraction occupied by the retrotransposon sequences identified in this study resulted respectively 0.50% (gypsy-like elements), 0.83% (ninja-like elements) and 0.49% (Bel-like elements).

The BLAST search was performed against *A. aegypti* genomic database and the results allow us to discriminate among insertions in gene free (or intergenic) genomic regions and to evaluate the number of full-length elements vs. rearranged elements (i.e. elements carrying disrupting mutations, insertion, deletions) or solo LTR insertions. A great number of insertions are represented by rearranged elements and by solo-LTRs that can be generated by homologous recombination events between the 5′ and 3′ LTRs have been also detected. In order to define the distance of the LTR retrotransposons from genes we performed our analysis using an arbitrary window length of 10 Kb upstream/downstream the genes annotated in Ensembl in which insertions have been searched (Supplemental table).

Our results indicate that 72% of the insertions detected by BLAST analysis lie outside the 10 Kbp window upstream/downstream the genes. Thirteen percent of LTR retrotransposon sequences analyzed lie within 10 Kb upstream/downstream putative or validated mosquito genes. Fifteen percent of the insertions detected fall within genes, but the vast majority of such insertions are localized in intronic sequences.

A further characterization of the insertions detected within genes has revealed that in few cases sequences belonging to a LTR retrotransposon can be detected in gene transcripts annotated in the

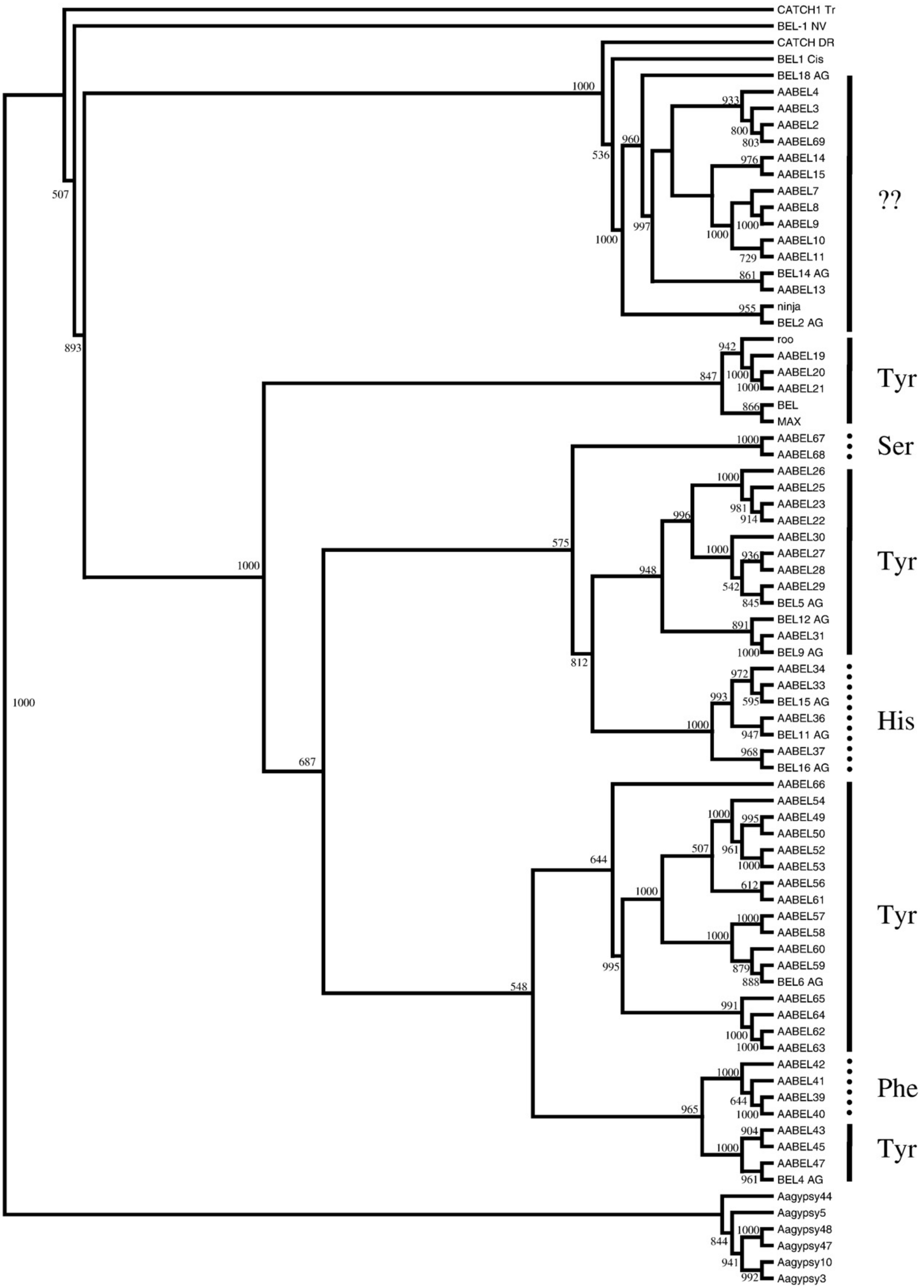

Fig. 2. Phylogenetic tree of the *Bel-Pao* elements. Phylogenetic relationships of the Bel-like retrotransposons based on the amino acids alignment of the conserved RT, RNase H and INT domains. Elements from this study are indicated as AABEL followed by a number, together to other elements retrieved from sequences annotated in Repbase. The PBS used by each group of elements is indicated. The N-J bootstrap values supporting the internal branches are indicated at the nodes. Only bootstrap values greater than 500 are reported. The tree is outgrouped with gypsy-like elements.

Table 2
List of *Aedes aegypti* gene transcripts containing retrotransposon fragment(s).

Element	Transcript ID	Description	% similarity[a]	Exon(s) recruited/total exons	Total exons
AABEL8	XM_001657844	NR	99	1st	2
AABEL8	XM_001664103	NR	98	1st	2
AABEL8	XM_001661571	NR	98	1st	2
AABEL8	XM_001657968	NR	98	1st	2
AABEL8	XM_001661578	Neurogenic locus delta protein	98	4	7
AABEL69	XM_001660875	NR	93	1st	3
AABEL19	XM_001652380	NR	99	1st, 2nd, 3rd	4
AABEL19	XM_001652512	NR	99	2nd, 3rd	3
AABEL51	XM_001660398	DNA polymerase zeta catalytic subunit	97	3	7
AABEL51	XM_001649905	NR	99	4th	4
AABEL54	XM_001660919	Zinc finger protein	97	1st	4
AABEL56	XM_001662457	NR	87	1st	6
AABEL58	XM_001652081	Regulator of sex-limitation	97	1st	8
AABEL62	XM_001652519	nnp-1 protein	96	3rd (part)	3
AABEL62	XM_001653357	NR	90	1st, 2nd	9
AABEL64	XM_001659509	NR	87	3′UTR	3
AABEL64	XM_001660436	Cask	83	3′UTR	5
AABEL64	XM_001660143	NR	97	1st	3
AABEL65	XM_001658735	Trypsin	100	3′UTR	1
AABEL65	XM_001651801	Maltose phosphorylase	100	2nd	3
AABEL66	XM_001648363	Oligophrenin	89	1st	6
AAGYPSY10	XM_001659841	NR	99	1st	2
AAGYPSY13	XM_001663843	Beta-arrestin 1	99	6th	9
AAGYPSY32	XM_001650161	Mediator complex	97	3′UTR	4
AAGYPSY40	XM_001664209	NR	96	1st	3
AAGYPSY40	XM_001648480	NR	77	3′UTR	3
AAGYPSY44	XM_001662251	NR	95	1st	2
AAGYPSY44	XM_001661168	NR	93	2nd	3
AAGYPSY44	XM_001658090	NR	98	1st	2
AAGYPSY48	XM_001660842	Tetratricopeptide repeat protein	99	1st	5

NR = none reported.
[a] The similarity percentage refers to the elements listed in Table 1.

Ensembl. This would mean that, occasionally, LTR retrotransposons fragments could be recruited as exons in the mature mRNA of host genes.

The results of a BLAST analysis against the ESTs database are reported in Table 2. Thirty-four *A. aegypti* transcripts show local sequence similarity with LTR retrotransposons. The aligned regions correspond precisely to predicted exons of the genes and both donor and acceptor splicing sites can be detected. Although many of the genes detected are hypothetical genes a function has been attributed to 15 of them. It is noteworthy that in many cases the contribution of the retrotransposon insertion in terms of exons to the mature mRNA is given by fragmented elements or by retrotransposon relics.

A careful comparative analysis has revealed that the retrotransposon-gene configurations observed in *A. aegypti* are not conserved in the genome of *C. pipiens* with the exception of the tetratricopeptide repeat protein (accession no. XM_001660842), which has a gene organization that mirror the one found in *A. aegypti* (data not shown).

4. Discussion

The genome of *A. aegypti* is particularly rich in transposable elements and this abundance has probably masked the presence of several elements in the initial genomic analyses (Nene et al., 2007). The genomic sequence of *A. aegypti* is continuously updated which means that other transposable element sequences are probably to be discovered and characterized. Moreover the overall organization of the *A. aegypti* genome into euchromatin and heterochromatin is poorly understood at the molecular level; consequently little is known about telomeric, centromeric and pericentromeric regions, which are massively enriched in repeated and mobile sequences (Severson et al., 2001).

We have used an intrinsic approach to detect LTR retrotransposons in the sequenced genome of *A. aegypti*. Such strategy allows a rapid and massive detection of LTR retrotransposons in completely sequenced genomes.

We have identified 76 novel LTR retrotransposon sequences, which can account for nearly 2% of the genome of *A. aegypti*. These elements form two distinct phylogenetic groups corresponding to the Ty3/gypsy family and the Bel family of the LTR retrotransposon. The structural features of each of the new elements described are consistent with the features of well-characterized relatives in other species supporting the evolutionary relationships highlighted by phylogenetic analyses.

Two main results can be highlighted from the structural analysis performed. The first observation concerns the unusually short LTRs of two elements related to the *woot* retrotransposon (AAGYPSY4 and AAGYPSY5). To our knowledge this is the first report of retrotransposons related to *Osvaldo* possessing such short LTRs. We were able to identify retrotransposons with similar characteristics only in the genome of *Culex pipiens quinquefasciatus*. This data indicate that a new LTR retrotransposon element family with unique features is present in the genomes of at least two Culicinae species. The genomic data of Culicinae are unfortunately restricted to *A. aegypti* and *C. pipiens quinquefasciatus* genomes so that further genomic and molecular evidences are necessary to assign specifically this novel family to the Culicinae genomes.

The second observation concerns the ninja-like elements. The analysis performed in the genome of *A. aegypti* confirmed previous results obtained on the genome of *A. gambiae* (Marsano and Caizzi, 2005). We showed that the Primer Binding Site of several LTR retrotransposons belonging to the ninja lineage is highly corrupted and nearly impossible to detect. The *ninja* element itself does not have a well-defined PBS, having “weak complementarity with the tRNA^{Ser}” (Ogura et al., 1996).

The PSB is essential for the replication mechanism of LTR retrotransposons that is very similar to the life cycle of lentiviruses. The “ninja paradox” could find an explanation if we consider that, like other LTR retrotransposons, these elements might not need a fully conserved and functional PBS sequence for their transposition because they may use an alternative mode of replication (Levin, 1995) or the may use truncated forms of tRNA to initiate the replication (Saigo, 1986).

The genomic distribution of transposable elements has been widely studied in completely sequenced genomes aiming to investigate the relationships with the host genes. These studies revealed that most of the elements lie in non-coding regions, intergenic and intron sequences (Griffiths et al., 1999), or in large gene-poor genomic regions (e.g. heterochromatin) (Pimpinelli et al., 1995). However, it has been reported that insertion of transposable elements near genes can influence gene expression (Desset and Vaury, 2005). This influence is significantly position dependent, and is due to the interference over the gene's regulatory elements circuitry (i.e. promoters, enhancer) or to the introduction of new regulatory elements carried by the retrotransposon itself. In particular, *gypsy* elements are known to affect the gene expression of nearby genes by specialized sequences that they harbor (Melnikova et al., 2002; Gause et al., 2001).

The results of the genomic distribution analysis are consistent with previously reported observations that mobile elements can be found in high percentage outside genes contributing to the expansion of intergenic regions in the genome of *A. aegypti*.

On the other hand insertions of mobile elements in introns are also frequent, thus explaining the 4 to 6 fold increase in the average gene length relative to *A. gambiae* (Nene et al., 2007).

The observation that a small fraction of insertions can contribute to the intron/exon organization of host genes is also intriguing.

It has been shown in recent papers that in *D. melanogaster* (Ganko et al., 2006) and *M. musculus* (DeBarry et al., 2006) LTR retrotransposon sequences are often associated with host genes and, in the case of *M. musculus*, they can be recruited as novel or spliced exons.

Although we are showing the presence of transposable elements in the proximity or within genes, this is not an exhaustive analysis, and it should be extended to the entire TEfam database where the bulk of transposable elements are annotated. Our data do not demonstrate that the associations detected are evolutionary fixed at least over long evolutionary time scale. In fact a single gene has been detected with the same configuration in *A. aegypti* and in *C. pipiens quinquefasciatus*. Furthermore gene expression analyses of different *A. aegypti* populations are necessary to assess that the associations observed are fixed at the specie level demonstrating the functional and evolutionary importance of the insertions.

5. Conclusions

We think that the results obtained integrate the already large amount of data concerning the mobile elements of *A. aegypti*. At the same time this is an example of how difficult can be the identification of the complete TE repertoire in a eukaryotic genome. The identification of the complete transposon set in a genome is otherwise essential to understand the evolution and the expression of a genome.

Other transposable elements are likely to be identified as the *A. aegypti* genome will become further assembled and annotated, or if novel approaches will be used. Moreover, the results presented in this paper indicate that the contribution of LTR retrotransposon insertion to the evolution of gene structure and function in *A. aegypti* may be not completely absent.

Acknowledgments

This work was supported by Istituto Superiore di Sanita` (contract no. OAD/F2) to R. C.

We thank Konstantinos Lefkimmiatis for critical reading and English improvement of the manuscript, Pietro D'Addabbo for helping us with the RepeatMasker program and for computer assistance. We would also like to thank the anonymous reviewers of the previous version of the paper for their useful comments.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.03.021.

References

Beeman, R.W., Thomson, M.S., Clark, J.M., DeCamillis, M.A., Brown, S.J., Denell, R.E., 1996. *Woot*, an active gypsy-class retrotransposon in the flour beetle, *Tribolium castaneum*, is associated with a recent mutation. Genetics 143, 417–426.

DeBarry, J.D., Ganko, E.W., McCarthy, E.M., McDonald, J.F., 2006. The contribution of LTR retrotransposon sequences to gene evolution in *Mus musculus*. Mol. Biol. Evol. 23, 479–481.

Desset, S., Vaury, C., 2005. Transcriptional interference mediated by retrotransposons within the genome of their host: lessons from alleles of the white gene from *Drosophila melanogaster*. Cytogenet. Genome Res. 110, 209–214.

Evgen'ev, M.B., Corces, V.G., Lankenau, D.H., 1992. *Ulysses* transposable element of *Drosophila* shows high structural similarities to functional domains of retroviruses. J. Mol. Biol. 225, 917–924.

Finnegan, D.J., 1992. Transposable elements. Curr. Opin. Genet. Dev. 2, 861–867.

Ganko, E.W., Fielman, K.T., McDonald, J.F., 2001. Evolutionary history of *Cer* elements and their impact on the *C. elegans* genome. Genome Res. 11, 2066–2074.

Ganko, E.W., Greene, C.S., Lewis, J.A., Bhattacharjee, V., McDonald, J.F., 2006. LTR retrotransposon–gene associations in *Drosophila melanogaster*. J. Mol. Evol. 62, 111–120.

Gause, M., Morcillo, P., Dorsett, D., 2001. Insulation of enhancer-promoter communication by a *gypsy* transposon insert in the *Drosophila* cut gene: cooperation between suppressor of hairy-wing and modifier of mdg4 proteins. Mol. Cell. Biol. 21, 4807–4817.

Griffiths, A.J.F., Gelbart, W.M., Miller, J.H., Lewontin, R.C., 1999. Modern Genetic Analysis. W. H. Freeman & Co, New York.

Inouye, S., Saigo, K., Yamada, K., Kuchino, Y., 1986. Identification and nucleotide sequence determination of a potential primer tRNA for reverse transcription of a *Drosophila* retrotransposon, 297. Nucleic Acids Res. 14, 3031–3043.

Jurka, J., 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 16, 418–420.

Levin, H.L., 1995. A novel mechanism of self-primed reverse transcription defines a new family of retroelements. Mol. Cell. Biol. 15 (6), 3310–3317.

Malik, H.S., Eickbush, T.H., 1999. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J. Virol. 73, 5186–5190.

Malik, H.S., Henikoff, S., Eickbush, T.H., 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. 10, 1307–1318.

Marsano, R.M., Caizzi, R., 2005. A genome-wide screening of BEL-Pao like retrotransposons in *Anopheles gambiae* by the LTR_STRUC program. Gene 357, 115–121.

McCarthy, E.M., McDonald, J.F., 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19, 362–367.

McCarthy, E.M., McDonald, J.F., 2004. Long terminal repeat retrotransposons of *Mus musculus*. Genome Biol. 5 (3), R14.

Melnikova, L., Gause, M., Georgiev, P., 2002. The gypsy insulators flanking yellow enhancers do not form a separate transcriptional domain in *Drosophila melanogaster*: the enhancers can activate an isolated yellow promoter. Genetics 160, 1549–1560.

Michaille, J.J., Mathavan, S., Gaillard, J., Garel, A., 1990. The complete sequence of *mag*, a new retrotransposon in *Bombyx mori*. Nucleic Acids Res. 18, 674.

Nene, V., et al., 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316 (5832), 1718–1723.

Ogura, K., Takechi, S., Nakayama, T., Yamamoto, M.T., 1996. Molecular structure of the transposable element ninja in *Drosophila simulans*. Genes Genet. Syst. 71 (1), 1–8.

Page RDM, 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12, 357–358.

Pantazidis, A., Labrador, M., Fontdevila, A., 1999. The retrotransposon *Osvaldo* from *Drosophila buzzatii* displays all structural features of a functional retrovirus. Mol. Biol. Evol. 16, 909–921.

Pimpinelli, S., et al., 1995. Transposable elements are stable structural components of *Drosophila melanogaster* heterochromatin. Proc. Natl. Acad. Sci. U. S. A. 92, 3804–3808.

Saigo, K., 1986. A potential primer for reverse transcription of mdg3, a *Drosophila* copia-like element, is a leucine tRNA lacking its 3′ terminal 5 bases. Nucleic Acids Res. 14, 4370.

Severson, D.W., Brown, S.E., Knudson, D.L., 2001. Genetic and physical mapping in mosquitoes: molecular approaches. Annu. Rev. Entomol. 46, 183–219.

Smit, A.F.A., Hubley, R. & Green, P., 1996–2004 Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*.1996–2004 http://www.repeatmasker.org>.

Syomin, B.V., Leonova, T.Y., Ilyin, Y.V., 2002. Evidence for horizontal transfer of the LTR retrotransposon mdg3, which lacks an env gene. Mol. Genet. Genomics 267, 418–423.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., Boeke, J.D., 2002. Molecular archeology of L1 insertions in the human genome. Genome Biol. 3 research0052.1-0052.18.

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

Tu, Z., 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. Proc. Natl. Acad. Sci. U. S. A. 98, 1699–1704.

Tubio, J.M.C., Costas, J.C., Naveira, H.F., 2004. Evolution of the *mgd1* lineage of the Ty3/gypsy group of LTR retrotransposons in *Anopheles gambiae*. Gene 330, 123–131.

Van de Peer, Y., De Wachter, Y., 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Applic. Biosci. 10, 569–570.