Mol Genet Genomics (2017) 292:565–583
DOI 10.1007/s00438-017-1291-7

ORIGINAL ARTICLE

Genome-wide analysis of transposable elements in the coffee berry borer *Hypothenemus hampei* (Coleoptera: Curculionidae): description of novel families

Eric M. Hernandez-Hernandez[1] · Rita Daniela Fernández-Medina[2] · Lucio Navarro-Escalante[3,4] · Jonathan Nuñez[5] · Pablo Benavides-Machado[3] · Claudia M. A. Carareto[1]

Received: 14 August 2016 / Accepted: 12 January 2017 / Published online: 15 February 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract The coffee berry borer (CBB) *Hypothenemus hampei* is the most limiting pest of coffee production worldwide. The CBB genome has been recently sequenced; however, information regarding the presence and characteristics of transposable elements (TEs) was not provided. Using systematic searching strategies based on both *de novo* and homology-based approaches, we present a library of TEs from the draft genome of CBB sequenced by the Colombian Coffee Growers Federation. The library consists of 880 sequences classified as 66% Class I (LTRs: 46%, non-LTRs: 20%) and 34% Class II (DNA transposons: 8%, *Helitrons*: 16% and MITEs: 10%) elements, including families of the three main LTR (*Gypsy, Bel-Pao* and *Copia*) and non-LTR (*CR1, Daphne, I/Nimb, Jockey, Kiri, R1, R2* and *R4*) clades and DNA superfamilies (*Tc1-mariner, hAT, Merlin, P, PIF-Harbinger, PiggyBac* and *Helitron*). We propose the existence of novel families: *Hypo*, belonging to the LTR *Gypsy* superfamily; *Hamp*, belonging to non-LTRs; and *rosa*, belonging to Class II or DNA transposons. Although the *rosa* clade has been previously described, it was considered to be a basal subfamily of the *mariner* family. Based on our phylogenetic analysis, including *Tc1, mariner, pogo, rosa* and *Lsra* elements from other insects, we propose that *rosa* and *Lsra* elements are subfamilies of an independent family of Class II elements termed *rosa*. The annotations obtained indicate that a low percentage of the assembled CBB genome (approximately 8.2%) consists of TEs. Although these TEs display high diversity, most sequences are degenerate, with few full-length copies of LTR and DNA transposons and several complete and putatively active copies of non-LTR elements. MITEs constitute approximately 50% of the total TEs content, with a high proportion associated with DNA transposons in the *Tc1-mariner* superfamily.

Keywords De novo TEs annotation · Transposable element library · Coffee berry borer · *Hypothenemus hampei*

Abbreviations

APE	Apurinic/apyrimidinic endonuclease
BLAST	Basic local alignment search tool
CBB	Coffee berry borer
GtRNAdb	Genomic tRNA Database
GyDB	Gypsy database
HSP	High-scoring segment pairs
LINE	Long interspersed element
LTR	Long terminal repeat
MITEs	Miniature inverted–repeat transposable elements
ORFs	Open reading frames
pHMMs	Profile hidden Markov models

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-017-1291-7) contains supplementary material, which is available to authorized users.

✉ Claudia M. A. Carareto
carareto@ibilce.unesp.br

[1] UNESP-Univ. Estadual Paulista, São José do Rio Preto, SP, Brazil

[2] Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

[3] National Coffee Research Center-CENICAFÉ, Manizales, Colombia

[4] Purdue University, West Lafayette, IN, USA

[5] CIAT-International Center for Tropical Agriculture, Cali, Colombia

RNaseH	Ribonuclease H
RT	Reverse transcriptase
TEs	Transposable elements
TIR	Terminal inverted repeat
TSD	Target site duplication
WGS	Whole-genome sequence

Introduction

Hypothenemus hampei Ferrari (Coleoptera: Curculionidae), also known as the coffee berry borer (CBB), is the most devastating insect pest for coffee production worldwide. CBB action has a direct impact on the sustainability of coffee cultivation because the insect infests coffee beans, the marketable product. The damage caused by CBB affects both the quality and market price of affected coffee beans. This beetle species possesses a sexual determination system that is referred to as functional haplodiploidy (Brun et al. 1995). The coffee berry borer was introduced into the Americas through Brazil presumably in 1913 in coffee seeds imported from Democratic Republic of Congo (Infante et al. 2014) and from there it colonized other coffee producing countries in the Americas and the Caribbean region. It is still unclear though, whether a single introduction of several lineages, or multiple introductions were the source of the insect in the Americas. Gauthier (2010) performed a worldwide CBB phylogenetic reconstruction using Bayesian methods and proposed the existence of four genetic populations (K1–K4). All the Americans (except from Jamaica) formed a single genetic group (K4). These findings suggest that few founder CBB genetic lineages were introduced into the Americas via Brazil and later dispersed across the Americas. This possible founder effect, along with an extreme inbreeding behavior (Gingerich et al. 1996), may explain the low genetic diversity of this species in the continent (Andreev et al. 1998; Gauthier 2010; Benavides et al. 2005; Gil et al. 2014). Yet, neither the CBB reproductive behavior nor its low genetic variability has restricted its success in colonizing new crops and its spread among all coffee producing countries. Historically in Colombia, CBB population sizes fluctuate along the year according to local coffee bean densities across the country (Bustillo 2006), but still, however, the highly diverse pattern of coffee bean production in the country ensures a significant overall CBB population size during all the year around.

The efforts to study the population genetics, structure and dynamics of CBB have been limited by the low genetic variation and high inbreeding of the species. Conventional molecular markers like microsatellites, AFLPs, SSCPs among others, used for population structure studies have been of little use due to their low discriminatory power.

Due to the economical relevance of this insect, a sequencing genomic project of the lineage infesting coffee cultures in Hawaii was envisioned in September 2010 by the United States Department of Agriculture (USDA) soon after CBB colonized the Kona region (Vega et al. 2015). The reported results indicated a very low content of identifiable repeats and low-complexity regions (2.7%) and suggested that repeated sequences might be underrepresented in the assembly; however, no further analysis on the repetitive elements in this genome was performed. In addition, the National Center for Coffee Research of Colombia, CENICAFE, has also pursued a genome sequencing project, which is currently in the annotation phase, of the CBB variant in Colombia. As part of the effort to characterize the genome of CBB prevalent in Colombia, we have analyzed the global transposable element (TEs) content, contributing to a better understanding of their distribution and abundance in the genome of this important insect pest.

Due to their ubiquity, diversity and genomic impact, the identification and characterization of TEs in newly sequenced genomes is essential for better understanding genome structure and evolution. TEs are interspersed DNA sequences that move within the genome and have the capacity to propagate, reaching finally high proportions of the genome. In particular, insect genomes show great variability in TE content, varying from 2.7% in some *Drosophila* species (Clark and Eisen 2007) to as much as 47% in *Aedes aegypti* (Nene et al. 2007). Since there is no consensus to date for a universal TE classification system (Piégu et al. 2015), for simplicity, we have followed the Wicker´s hierarchical system for TEs classification in eukaryotes (Wicker et al. 2007). In this system, two classes of elements are defined according to their genetic and structural characteristics. Class I is composed of five orders: LTR retrotransposons, *DIRS*-like elements, *Penelope*-like elements (PLEs), long interspersed elements (LINEs) and short interspersed elements (SINEs). Class II is further divided into two subclasses. Subclass 1 (orders *TIR* and *Crypton*) comprises the classical 'cut-and-paste' TEs, which are characterized by terminal inverted repeats (TIRs) of variable length; Subclass 2 (orders *Helitron* and *Polintons*/*Maverick*) comprises 'copy-and-paste' elements exhibiting a transposition process that entails replication without double-stranded cleavage. In this hierarchical system, TEs are further classified into superfamilies according to widespread large-scale features and families (also called clades or lineages) defined by DNA sequence conservation. Additionally, elements can be autonomous (i.e., able to transpose) or non-autonomous (such as Class I, SINEs and miniature inverted-repeat transposable elements (MITEs) of Class II).

It is well established that TEs have important impacts on host genomes, both at the structural and functional levels (e.g., Pardue et al. 1997; Kidwell and Lisch 2000; van de Lagemaat et al. 2003; Wong and Choo 2004; Levin and Moran 2011). Therefore, it is not surprising that these sequences are considered valuable tools for comparing genomes, elucidating recent genomic dynamics and making evolutionary inferences. In addition, transposon-based genetic approaches are useful in both higher and lower metazoans (e.g., transgenesis, forward genetic approaches, insertional mutagenesis or genetic population analysis), and are thus important tools of the genetic toolkit for several applications (Mátés et al. 2007).

The repetitive nature of TEs is one of the main difficulties for the correct assembly of genomes. In addition, the accurate detection and annotation of TEs is a difficult task because of their great diversity. A number of computational approaches and tools have been developed for identifying TEs in assembled genomes. Two main strategies, homology-based and de novo approaches (typically based either on their repetitive nature or on their structural signatures) are commonly used but the adoption of combined approaches has shown to be the best strategy for obtaining comprehensive and sensitive results (Permal et al. 2012; Platt et al. 2016). Here, we present a detailed approach for the de novo characterization of TEs in the CBB genome aiming at providing basis for future studies on TE insertion polymorphisms that can help to characterize the populations´ genetic diversity.

Methods

Genomic information

The genome analyzed in this study was isolated from a CBB strain derived from a population from Pueblo Bello (Colombia) that has been maintained in a breeding facility of CENICAFE during 9 years. The genomic information used in this project is property of the National Coffee Growers Federation of Colombia and is protected by intellectual property agreements.

Two genome draft versions were used in this study. One version was produced by assembling reads resulting from 20-kb pair-end whole-genome shotgun sequencing (WGS) by 454-FLX Titanium pyrosequencing. This version was used to build the repeat library, yielding a total of 6,910,124 assembled reads and constituting a genome of approximately 197.3 MB. The library contains 75,017 contigs (rank size 200–36,179 bp), with an average length of 2629.5 bp, an N_{50} value of 6931 bp and GC richness of 36%. The second set of sequences includes genomic information from tag-based next-generation Illumina sequencing technology obtained from a bacterial artificial chromosome (BAC) library. The genome constructed from this set of sequences has 25,643 scaffolds for a total consensus of 218.3 MB, with a rank size between 897 and 278,432 bp, an average contig-length of 8511 bp, an N_{50} value of 27,956 bp and a GC content of 38%. The completeness analysis of the genome using annotation of core genes was carried out. The result gave a value around 90% of genome completeness.

Identification and annotation of TEs

We obtained data from combined independent sources, integrating results from multiple homology-based and de novo TE identification methods. The methods and tools utilized for de novo identification of Class I (LTR and non-LTR retrotransposons) and Class II elements in this study are applied as follows (Fig. 1).

Identification of transposable elements

Class I elements

LTR retrotransposons Figure 2a summarizes the methods and tools used for the de novo identification of LTR retrotransposons. The *LTRharvest* (Ellinghaus et al. 2008) program was used for the de novo identification of LTR retrotransposons with the following parameters: seed value 100, minimum LTR size 100, maximum LTR size 5000, minimum distance between LTRs 1000, maximum distance between LTRs 20,000, similarity 85%, overlaps "*best*" option, rank size of target site duplication of 4–8 bp. Automated annotation of the internal features of the identified LTR retrotransposons was performed using the *LTRdigest* software (Steinbiss et al. 2009), adding the eukaryotic-tRNAs set from genomic tRNA database (GtRNAdb) (Chan

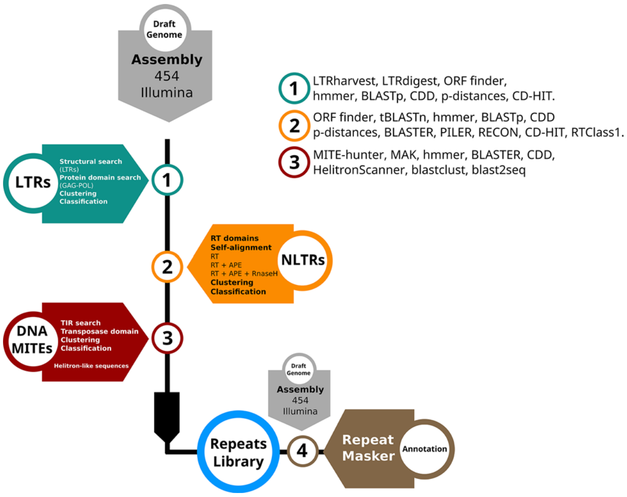


Fig. 1 Flowchart showing the pipeline applied for the identification of TE canonical sequences using structural and homology-based approaches and their classification in the genome of CBB

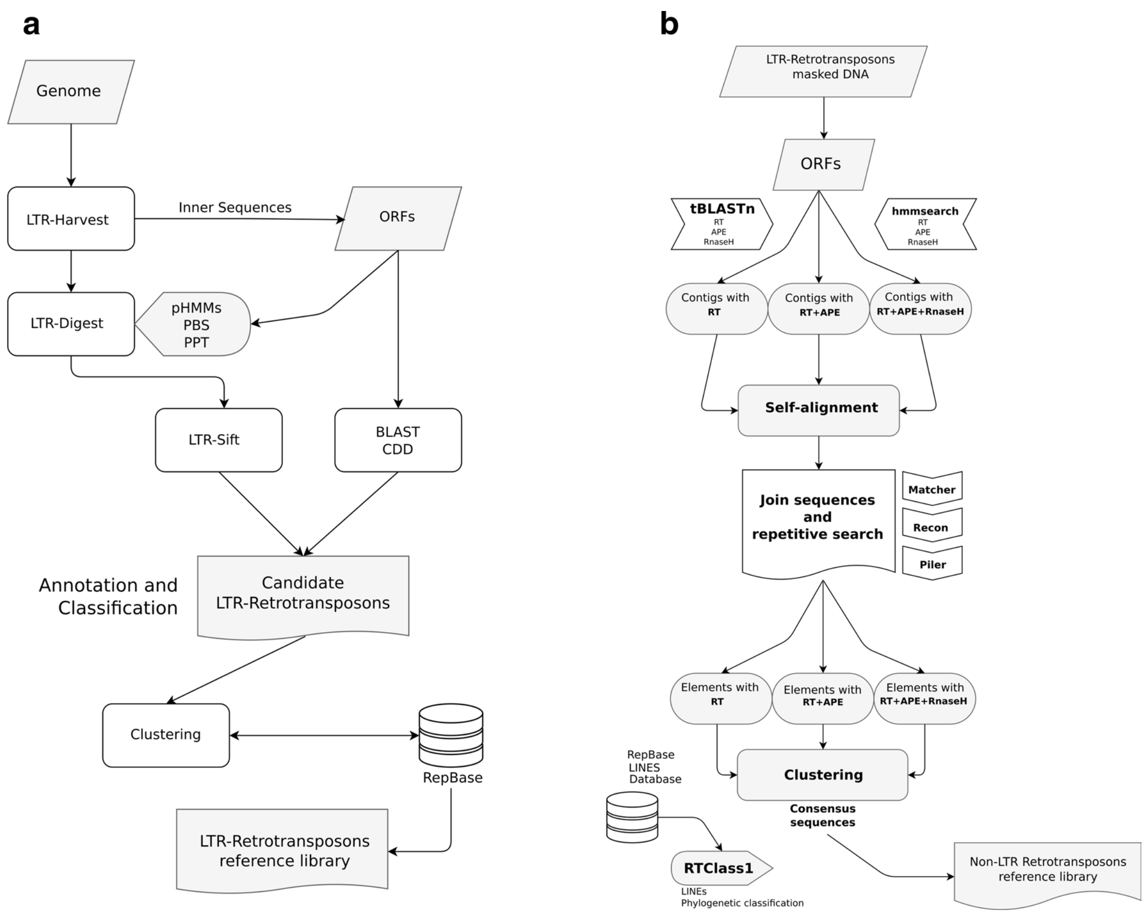


Fig. 2 Flowchart for de novo TE canonical sequences identification using structural and homology-based for approaches: **a** LTR retrotransposons, **b** non-LTR retrotransposons

and Lowe 2009) and replacing its default hidden Markov model (pHMMs) profiles library with the non-redundant database of pHMMs from the Gypsy Database (Llorens et al. 2011; http://gydb.org/index.php/Main_Page), which is larger and more complete in pHMMs than the database associated with *LTRdigest*. Positive hits with pHMMs were visually inspected using the graphical desktop tool *LTRsift* (Steinbiss et al. 2012); this program permits better internal structure inspection of the predicted elements and helps filtering and classifying them in a functional manner. In addition to searching pHMMs in the predicted LTR retrotransposons, open reading frames (ORFs) from internal sequences (between LTRs) were identified using the ORF Finder Tool Server (http://www.ncbi.nlm.nih.gov/orffinder/) and blasted against the non-redundant (nr) database using the BLASTp program. Also, hits with conserved protein domains were detected (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

Clustering of sequences and subsequent generation of consensus was performed using the *CD-HIT* program (Li and Godzik 2006) considering a similarity of 80% following the Wicker's rules for TEs classification (Wicker et al. 2007). Finally, a step for the classification of elements into superfamilies was incorporated based on p-distances and clade bootstrap. The inferred amino acid sequences corresponding to the reverse transcriptase of the identified LTR retrotransposons were grouped with the insects LTR sequences available in GyDB. Multiple sequence alignments of the RT domains with LTR reference sequences of insects (Table S2) were performed to further classify the *H. hampei* elements by phylogenetic analyses.

Non-LTR retrotransposons Figure 2b summarizes the methods and tools used for the de novo identification of non-LTR retrotransposons. First, the location coordinates of the LTR retrotransposons previously identified in the genome were masked to avoid hits with the RT domains already identified. Non-LTRs have been previously classified into clades or lineages (Malik et al. 1999), and subsequently into different families; only two of the clades have a single domain encoding the RT (*R2* and *CRE* clades), the others contain an additional coding region for an apyrimidinic endonuclease domain (APE). Other elements, such as those belonging to clade I, have an additional RNaseH domain; nevertheless, classification of non-LTRs has generally been based on the sequence for RT (Xiong and Eickbush 1990; Kapitonov et al. 2009). We then performed a search with *hmmsearch* (HMMER 3.0) (Eddy 2011) of the genomic sequences that hit the RT hidden Markov model profile EMBL: DS36752 calibrated with the program *MGEScan-nonLTR* (Rho and Tang 2009) as well as to the PF00078 Pfam profile (27.0) (Finn et al. 2014). This pHMMs search was performed against the ORFs extracted from the masked genome using the *getorf* tool from the EMBOSS v6.4.0.0 package. A minimum ORF size of 500 bp was considered to contemplate the APE domain (rank size between 600 and 800 bp in 97% of inspected non-LTR elements).

The *BLASTER* algorithm (Quesneville et al. 2005) was implemented on all the scaffolds showing matches with the query domains already mentioned, allowing all-by-all alignments between sequences. This process helped defining the non-LTR boundaries. The *BLASTER* program can use either the *WU-BLAST* or *NCBI-BLAST* programs to compare a genome with itself to detect repeats; its binaries are distributed with the REPET pipeline (Flutre and Duprat 2011). After clustering, *hmmer* and *tBLASTn* searches were implemented to define the three main groups of non-LTRs according to the presence of specific domains: only RT (*CRE* and *R2* clades), RT + APE (*Rex, L2, Jockey, CR1, L1, Rand* and *RTE* clades) and RT + APE + RNaseH (*Tad, I* and *R1* clades). The *FASTA* sequences and the pHMMs used were those reported by Malik et al. (1999). The initial step in each element classification was based on several *BLAST* searches (*BLASTn, tBLASTx* and *BLASTx)* against the non-LTR retroelements deposited in Repbase (Jurka et al. 2005). For each representative non-LTR clade, the identified RT amino acid sequence was submitted to the Genetic Information Research Institute server (GIRI) that hosts *RTclass1*, a tool that assigns an unclassified non-LTR retrotransposon to one of the known non-LTR retrotransposons clades based on phylogenetic analysis of the RT domain (Kapitonov et al. 2009). Multiple sequence alignments together with non-LTR reference sequences of insects (Table S2) were performed for further non-LTR retrotransposons classification of the *H. hampei* sequences by phylogenetic analyses.

Class II elements

We have used a methodology for the detection of DNA transposons based on the initial identification of non-autonomous elements, such as MITEs, in the genome. The general idea is that non-autonomous elements frequently outnumber their autonomous counterparts and that autonomous elements are lost from the genome, whereas non-autonomous elements persist. In addition, non-autonomous elements may not arise simply by deletions of the autonomous elements (Kapitonov et al. 2009) and strategies based on the detection of non-autonomous elements from their autonomous counterparts would result in an underestimation of non-autonomous elements by its allocation to a specific autonomous element (Jiang 2013).

To identify non-autonomous elements, the program *MITE-Hunter* was used (Han and Wessler 2010). This program was designed for identification of MITEs (*Tourist* and *Stowaway* elements) and other Class 2 non-autonomous elements containing terminal inverted repeats (TIRs). *MITE-Hunter* uses a search mechanism similar to that of other well-known programs such as *FINDMITE* and *MUST* (MITE Uncovering SysTem) that are based on locating of TIRs and TSDs in adjacent regions; however, *MITE-Hunter* has a subsequent filtering step for retrieving MITE candidates and their flanking sequences after distinguishing homology between candidate sequence and flanking sequences, thereby achieving a false positive rate of only 4.4% relative to the 85% or higher false positive rates obtained with the other programs. *MITE-Hunter* was used with default parameters with a maximal length threshold of 800 bp.

The candidates MITEs identified were further classified based on the structural features of DNA transposons. Visual inspection was performed to identify structural features that define the putative associated family, using less conservative judgment because of the draft state of this genome; thus, when a candidate MITE displayed several structural features of a given family (Jiang 2013) it was assigned to that family. Hence, to associate a MITE with the *hAT* elements, its sequence was required to meet one of the following conditions: a TSD = 8 bp; TIR size between 8 and 22 bp; or a terminal sequence (5′ … 3′) with nucleotides [CT]A … T[AG]. In the same way, to be associated with the *CACTA* elements, the criteria were TSD = 3 bp, TIR size between 12 and 28 bp or a terminal sequence (5′ … 3′) with nucleotides CACT[AG] … [CT]AGTG; to be associated with *MULE* elements, TSD = 7–11 bp, TIR size between 0 and 800 bp or a terminal sequence (5′.0.3′) containing nucleotides [GC] … [GC]; to be associated with *Tourist* (*PIF/Pong*) elements, a TSD = TNA (N being any nucleotide), TIR size between 14 and 60 bp or a terminal sequence (5′ … 3′) with nucleotides G[GC][GC] … [GC][GC]C or G[AG]CA … TGC[TC] for *PIF* and *Pong*, respectively.

We then used a searching strategy based on the finding of autonomous and long non-autonomous elements starting from the identification of the TIR homologies with short non-autonomous elements (MITEs) using the *MITE Analysis Tool* kit (*MAK*) (Yang and Hall 2003). The input used corresponds to a set of “*known*” MITEs sequences

(*MITE-Hunter* output); the *-long* and *-anchor* functions were used to identify both "*long elements*" (non-autonomous and non-MITEs > 800 bp) and "*anchor elements*" (autonomous) that were identified by *BLAST* against transposase protein sequences extracted from RepBase (REPET pipeline edition:RepBase 18.08).

Four hundred fifty-one *Pfam* profiles related to transposase were used for *hmmersearch* against the genome; once identified by a hit with the Tpase motif, 5-kb sequence extension to the flanking regions was extracts and subsequently cluster the sequences to search for TIRs of each putative DNA transposon. The clustering process involved the all-by-all alignment of sequences with hits to Tpase previously identified using the *BLASTER* program and *blastclust* with the 80% rule. Finally, TIRs identification was performed by *blast2seq*, and the ORFs between TIRs were blasted against the conserved domain database (CDD) to verify the presence transposase-like domain in each element. The allocation of each type of element to a given superfamily was performed by *CENSOR* in the GIRI server (http://www.girinst.org/censor/), forcing a translated search of the sequences and considering the superfamily of elements in agreement with their masking scores to known repeats in Repbase. Multiple sequence alignments with DNA transposon reference sequences of insects (Table S2) were performed for further DNA transposon classification in the *H. hampei* by phylogenetic analyses.

Finally, for the *Helitrons* identification, a two-layered local combinational variable (LCV) tool *HelitronScanner*, was used. This tool scores 5′ and 3′ termini based on a training set of published *Helitrons* and merges the coordinates and scores for putative *Helitron*-like sequences (Xiong et al. 2014).

Transposable elements annotation

Determination of copy number, frequency, genomic coverage, divergence values and masking of *H. hampei* TE sequences were performed using the repeat library here generated and the *H. hampei* genome by RepeatMasker.

It is well known that events of retroelement amplification and LTR recombination are counteracting mechanisms related to genome size reduction and/or expansion. The unequal homologous recombination that generates the formation of solo LTRs can be identified by RepeatMasker as long as it is guaranteed in the input data. Therefore, each of the LTR elements has their LTR region independently counted of their internal region.

Phylogenetic analyses

Phylogenetic analyses were performed with MEGA5 (Tamura et al. 2011) using MAFFT (Katoh and Standley 2013, http://mafft.cbrc.jp/alignment/server/) for multiple alignments construction using consensus sequences from *H. hampei* and reference sequences representing different lineages or families from other insect genomes. The progressive method with iterative refinement method (parameters G-INS-i) was used.

The MSA were visually inspected and the highly variable regions for which the positional homology could not be determined were manually excluded.

The trees were reconstructed using the neighbor joining (based on pairwise distances) and maximum likelihood (based on the best amino acid substitution models according to the lowest BIC scores—Bayesian Information Criterion) methods, as implemented in MEGA 5.0. The evolutionary distances were computed using the General Reverse Transcriptase + FrEq. method and are in the units of the number of amino acid substitutions per site. The identification of unambiguously aligned position was performed visually to retain the largest number of sites for phylogeny reconstruction. Bootstrap values for each branch were assessed from 1000 to 100 replicates for NJ and ML, respectively.

Results

Using systematic searching strategies based on both de novo and homology-based approaches, a TE reference library was created from the draft genome of the CBB variant prevalent in Colombia. The methodologies implemented led to the construction of a library consisting of 880 TE sequences, including Class I (LTR and non-LTR retrotransposons) and Class II elements (TIR elements, *Helitrons* and MITEs). The TE sequences annotated according to this library comprise 18,098,679 bp, or 8.29%, of the CBB genome (Table 1).

Construction of the coffee berry borer repeat library

The use of multiple methods for identifying and annotating TEs is the only way to obtain reliable results, as different elements vary considerably in genetic structure and sequence (Quesneville et al. 2005; Permal et al. 2012; Hoen et al. 2015; Platt et al. 2016). We used different approaches and methodologies to identify each class and order of TEs present in the genome of this insect (Fig. 1). LTR elements were identified using *LTRharvest* (Ellinghaus et al. 2008), and internal sequences were annotated using *LTRdigest* (Steinbiss et al. 2009) with profile hidden Markov protein models. Simultaneously, open reading frames (ORFs) were identified and confirmed as transposition machinery-related domains (Fig. 2a). Reverse transcriptase (RT), endonuclease and

Table 1 Summary of transposable elements in the *H. hampei* draft genome (238.7 MB) using a species specific de novo library

Class	Order	Superfamily	No. of matches	Total bases	TEs percentage	Percentage of genome	% of divergence mean (min–max)
Class I							
Retrotransposon	LTR	*Gypsy*	5186	236,5071	13.07	1.08	19 (0–50.1)
		Copia	226	137,861	0.76	0.06	18.6 (0–46.1)
		Bel/Pao	711	519,180	2.87	0.24	18 (0–40)
		Total LTR	6123	3,022,112	16.7	1.38	18.3
	DIRS	*DIRS-like*	533	247,802	1.37	0.11	24.4 (0–38.6)
	LINE	*R1*	1246	808,616	4.47	0.37	21.9 (0–40.3)
		Jockey	518	258,341	1.43	0.12	21.5 (0–38.7)
		CR1	103	87,885	0.49	0.04	24 (0.2–37.8)
		I	248	199,151	1.10	0.09	25 (0–39.3)
		Nimb	208	199,172	1.10	0.09	21 (0.1–36.1)
		Daphne	32	28,785	0.16	0.01	26.6 (0–38.2)
		R2	15	6090	0.03	0.00	17.7 (1–31)
		R4	58	40,101	0.22	0.02	8.6 (1.4–18.2)
		Kiri	63	46,213	0.26	0.02	6.6 (0–33.6)
		Unknown	5	4307	0.02	0.00	13.4 (9.6–17.9)
		Total LINE	2496	1,678,661	10.65	0.76	18.6
	Total Class I		9152	4,948,575	27.35	2.25	20.5
Class II							
DNA transposon	TIR	*Tc1-Mariner*	8005	2,083,022	11.51	0.95	24.2 (0–66)
		hAT	255	67,933	0.38	0.03	18.1 (0–35.3)
		P	665	179,390	0.99	0.08	18.5 (0–36)
		PIF-Harbinger	125	35,560	0.20	0.02	27.6 (0–38.2)
		PiggyBAC	158	44,627	0.25	0.02	21.3 (0–36)
		Merlin	331	85,671	0.47	0.04	13 (0–44.7)
		Total TIR	9539	2,496,203	13.8	0.19	20.4
	Helitron	*Helitron-like*	8158	1,741,627	9.62	0.80	22.5 (0–54)
	Total DNA		17,697	4,237,830	23.42	1.94	21.5
	MITE	MITE	48,321	8,912,274	49.24	4.08	14.6 (0–69)
	Total Class II		66,018	13,150,104	72.65	6.02	18.0
Total			75,170	18,098,679	100.00	8.27	19.3

ribonuclease H domains were identified hierarchically, and alignment of the RT region was used to identify non-LTR retrotransposons. Classification of these sequences into different clades was performed using *RTclass1* (Kapitonov et al. 2009), a tool for the automatic assignment of novel non-LTR retrotransposons to known or novel clades using phylogenetic analysis of RT domain protein sequences (Fig. 2b). Non-autonomous DNA elements (MITEs and degraded DNA transposons) were identified by their TIRs and TSDs, which were later used for identification of autonomous counterparts based on the presence of the transposase domain. Specific software (*HelitronScanner*) was used for identifying *Helitron*-like sequences (Xiong et al. 2014).

LTR retrotransposons identification

LTRharvest predicted the presence of 282 sequences harboring both LTRs, covering almost 1.24 Mb and with an average size of 4405 bp. The most frequent target size duplication identified (TSD) was the TTTT motif, which appeared 11 times, followed by ATAT (nine times), TTTA (eight times) and ATTT and TATA (seven times each). LTR retrotransposons superfamilies are defined by the presence of GAG complex plus the presence and order of coding regions for retrotransposition enzymes, and the automated annotation of these internal features inside the identified LTR retrotransposons was performed using *LTRdigest*. A total of 113 sequences that hit at least one hidden Markov model profile (pHMM) were identified. Of those, 105 hit the RT profile; 104 of which hit the integrase enzyme, 102 the RNaseH, 69 the aspartic proteinase (AP), and 70 the GAG domain. Positive hits with pHMMs were visually inspected using the graphical desktop tool *LTRsift* (Steinbiss et al. 2012). A blast search against the CDD database helped to corroborate the sequences with previous pHMM hits; 96 elements contained ORFs that hit this database.

Sequence clustering following Wicker's 80-80-80 rule (sequences longer than 80 bp, sharing more than 80% sequence identity, over 80% of their length) and identification of masked sequences using both RepeatMasker and Censor resulted in 94 representative sequences that were allocated to the three main LTR superfamilies reported in metazoans (65 *Gypsy*, 22 *Bel-Pao* and 7 *Copia*) (Table S1-A).

Phylogenetic analyses by both neighbor joining (data not shown) and maximum likelihood using sequences presenting full-length RTs (42 *Gypsy*, 20 *Bel-Pao* and 6 *Copia*, as one of the identified sequences was truncated at the N terminus, and therefore not used in the alignment), together with published reference sequences (accession numbers are listed in Table S2), were performed to further classify the sequences of each superfamily into different clades, families and subfamilies. This phylogenetic analysis corroborated the classification of the LTR retrotransposons obtained by de novo methodology. The LTR sequences identified in the CBB genome (belonging to *Gypsy, Pao-Bel* and *Copia*) clustered together with the respective reference sequences in insects (Table S2), with highly supported bootstrap values (Figs. 3, 4, 5).

Gypsy elements in insects have been previously classified into five distinct clades (*Mag, Mdg1, Mdg3, Gypsy* and *CsRn1*). The 42 *Gypsy* sequences spanning the entire RT domain (177 aa) identified in *H. hampei* together with 44 reference sequences representing the *Gypsy* clades described thus far and two outgroup sequences were aligned and used in phylogenetic analyses (Fig. 3). The *Gypsy* sequences in *H. hampei* are highly divergent and could be classified into four of the previously characterized clades: one sequence belongs to the *Mag* clade, four to *Gypsy*, eight to *Mdg3* and 13 to *CsRn1*. We also identified

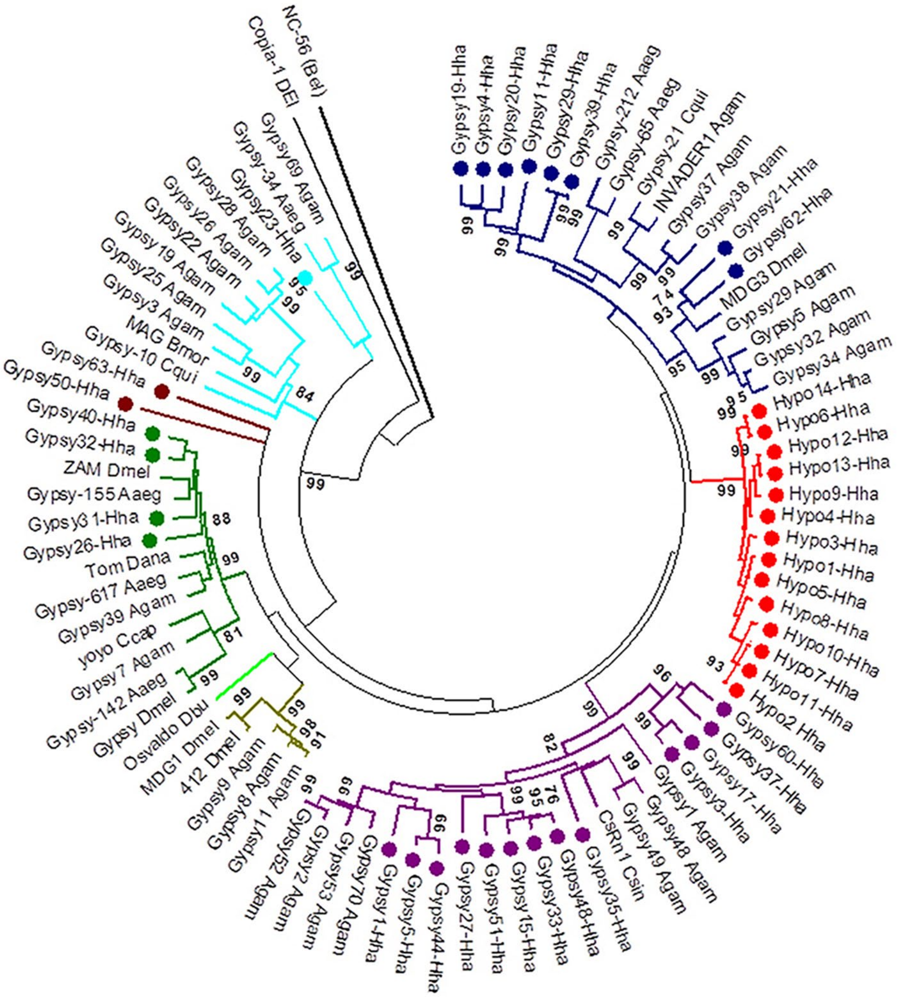


Fig. 3 Phylogenetic relationships of 42 consensus sequences from *H. hampei* spanning the RT domain and 44 reference sequences belonging to the *Gypsy* superfamily plus two outgroup sequences. The MSA were visually inspected and highly variable regions for which the positional homology could not be clearly determined, were excluded. The phylogeny was inferred using maximum likelihood method with 100 bootstrapping resamplings (values over 70% are shown) as implemented in MEGA5. The evolutionary distances were computed using the General Reverse Transcriptase + Freq. method. The analysis involved 88 amino acid sequences. There were a total of 190 positions in the final dataset. *Different color lines* were used to indicate different lineages within the *Gypsy* superfamily as follows: *dark blue: Mdg3; red: Hypo; purple: CsRn1, olive green: Mdg1; light green: Osvaldo; green: Gypsy, brown*: putative novel family, and *light blue: MAG. Colored dots* were used to represent the sequences isolated from *H. hampei*. (Color figure online)

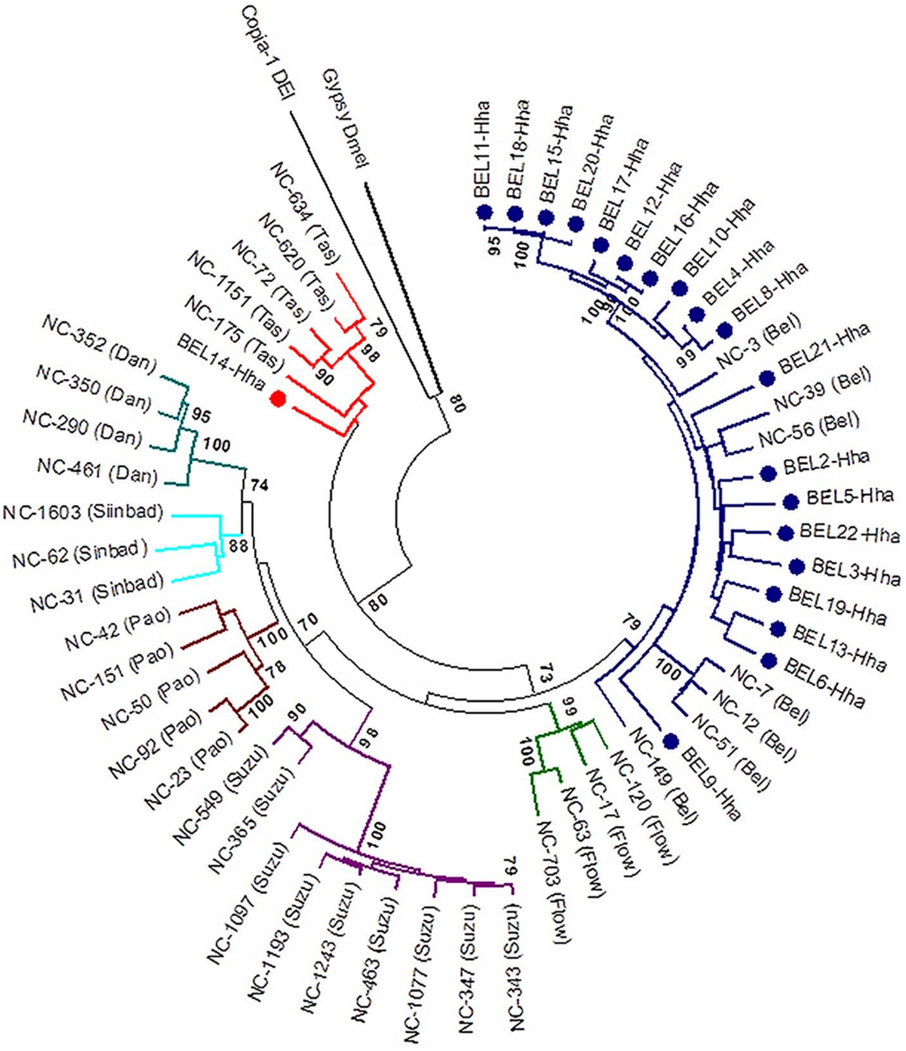


Fig. 4 Phylogenetic relationships of 20 consensus sequences from *H. hampei* spanning the RT domain and 35 reference sequences belonging to the *Bel-Pao* superfamily plus two outgroup sequences. The MSA were visually inspected and highly variable regions for which the positional homology could not be clearly determined, were excluded. The phylogeny was inferred using the maximum likelihood method with 100 bootstrapping resamplings (values over 70% are shown) as implemented in MEGA5. The evolutionary distances were computed using the General Reverse Transcriptase + Freq. method and are in the units of the number of amino acid substitutions per site. The analysis involved 57 amino acid sequences. There were a total of 176 positions in the final dataset. *Different color lines* were used to indicate different lineages within the PaoBel superfamily as follows: *dark blue: Bel*; green: *Flow; purple: Suzu; brown: Pao; light blue: Sinbad; turquoise: Dan* and *red: Tas. Colored dots* were used to represent the sequences isolated from *H. hampei*. (Color figure online)

14 sequences clustering together, with high bootstrap value and separately from any other previously characterized *Gypsy* family (Fig. 3, red dots). The overall average identity in the RT region for all the sequences in this clade is higher than 90%, indicating that they constitute a specific *Gypsy-Hha* lineage that we propose to be called *Hypo*.

Two other sequences (*Gypsy63-Hha* and *Gypsy50-Hha*, brown dots in Fig. 3) do not cluster either with other members of this superfamily or with each other indicating that they also might represent specific *Gypsy* families in the CBB genome, although we cannot rule this possibility out, with such a limited number of sequences.

The *Bel-Pao* superfamily has been previously classified into seven discrete lineages or clades (*Pao, Sinbad, Bel, Tas, Suzu, Flow* and *Dan*), which tend to cluster according to host species phylogeny (Copeland et al. 2005; de la Chaux and Wagner 2011). We identified 20 *Bel-Pao* sequences corresponding to full-length RT (213 aa) that were aligned to 37 reference sequences matching the seven described clades (accession numbers are listed in Table S2). One sequence each of *Gypsy* and *Copia* were used as outgroup retrotransposons. The great majority of the sequences in *H. hampei* belong to the *Bel* clades (19/20), whereas only one corresponds to the *Tas* clade (Fig. 4). Finally, six sequences spanning the full-length RT domain (246 aa) corresponding to the *Copia* superfamily in *H. hampei* were aligned to a set of 37 *Copia* reference sequences from insects (accession numbers are listed in Table S2). The ML tree suggests that the *Copia* elements in *H. hampei* are not monophyletic; these elements do not cluster with significant bootstrap values that allow robust classification with any of the reference sequences (Fig. 5). To our knowledge, the *Copia* superfamily in insects has not been formally classified into different families or clades. In our phylogeny, the elements used as references clustered into at least five different groups, one of which corresponds only to *Drosophila* sequences, whereas the others contain a mixture of sequences belonging to various mosquito species.

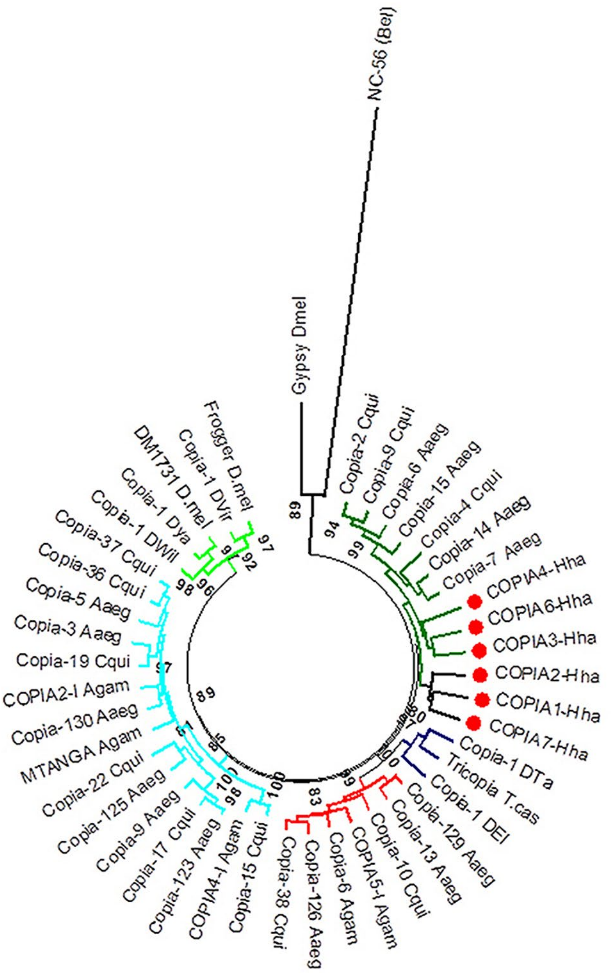


Fig. 5 Phylogenetic relationships of six consensus sequences from *H. hampei* spanning the RT domain and 37 reference sequences belonging to the *Copia* superfamily plus two outgroup sequences. The MSA were visually inspected and highly variable regions for which the positional homology could not be clearly determined, were excluded. The phylogeny was inferred using the maximum likelihood method with 100 bootstrapping resamplings (values over 70% are shown) as implemented in MEGA5. The evolutionary distances were computed using the General Reverse Transcriptase + FrEq. method [2] and are in the units of the number of amino acid substitutions per site. The analysis involved 45 amino acid sequences. All positions with less than 0% site coverage were eliminated. That is, fewer than 100% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 241 positions in the final dataset. *Different color lines* were used to indicate different lineages according to the criteria of bootstrap values higher than 70. *Red dots* were used to represent the sequences isolated from *H. hampei*. (Color figure online)

Some of the LTR elements described herein present extremely long LTRs (>1000 nt) in comparison to the LTR length values found in already characterized LTR elements. This was observed in three *Bel* sequences (Bel15-Hha, Bel16-Hha and Bel21-Hha) with LTRs sizes of 4389, 1261 and 2393 nt, respectively, two *Gypsys* (Gypsy53-Hha and Gypsy60-Hha) belonging to the CsRn1 clade, with LTR sizes of 2174 and 1751 nt, respectively, and five *Hypo* sequences (Hypo11-Hha, Hypo12-Hha, Hypo14-Hha, Hypo17-Hha and Hypo9-Hha) with LTRs ranging from 1019 to 2145 nt (Table S1). A closer analysis of these sequences showed the presence of ORFs and specific TEs conserved domains (Table S1-A). A closer analysis of these sequences showed the presence of ORFs and specific TEs conserved domains within the so-called LTRs (Table S1-A). The presence of PeptidaseA17, RVE, RT and RH domains clearly indicates that these sequences represent nested elements, i.e., pre-existing TEs with secondary TE insertions, and not LTRs as were originally described by our methodology. Although present in many species, nested TEs are relatively less abundant than non-nested TEs. A general TE assessment to detect nested TEs using available databases showed the occurrence of 6% nested TEs: 802 of a total of 11,329 TEs inserted into 690 host TEs (Gao et al. 2012).

In addition to sequences containing direct repeats, two DIRS-like elements were identified and included in the CBB repeat library. Members of this order contain the unusual features of a tyrosine recombinase instead of an integrase, and their termini resemble either split direct repeats (SDRs) or inverted repeats, indicating an integration mechanism different from that of LTRs and LINEs. Nevertheless, their reverse transcriptase gene places them in Class I. Members of this order have been detected in diverse species, ranging from green algae to animals and fungi (Goodwin and Poulter 2004).

Non-LTR retrotransposons identification

In an initial search, protein sequences homologous to reverse transcriptase (pHMMs PF00078 and DS36752) were found in 1038 scaffolds. *MATCHER* software performed defragmentation of 1,045 sequences that together with 1456 sequences recovered by *PILER* and *RECON*, were saved in a *FASTA* file. Finally, a clustering process of the sequences identified by *MATCHER, RECON* and *PILER* resulted in 1,903 sequences with 100% identity. These sequences were further filtered with a 1-kb threshold, resulting in 995 representative sequences that correspond to non-LTRs. A hierarchical *tBLASTn* search for additional domains (APE and RNaseH) in scaffolds with homology to RT identified 273 scaffolds with APE hits, resulting in 259 unique sequences in the genome with RT + APE structure, which is characteristic of non-LTR retrotransposons. A total of 171 assembled sequences were discarded because they only hit the APE and/or the RNaseH domain, and therefore lack the main protein domain corresponding to reverse transcriptase.

A restricted hidden Markov model search (*e* value $<1\times10^{-03}$) implemented with model profiles for RT (DS36752) and APE (DS36736) enzymes resulted in 138 sequences that hit only the RT domain and 106 sequences that hit RT + APE domains. These sequences were searched against non-LTR elements in Repbase using *tBLASTx, tBLASTn* and *BLASTx*. This approach enabled a preliminary classification of the elements into clades based on the three types of BLAST evidence. Finally, 80 representative sequences of non-LTR retrotransposons distributed in the *Jockey*, I/*Nimb, CR1, Kiri, Daphne, R2* and *R4* clades were clustered using Wicker's criterion.

The results of classification based on the phylogenetic relationships among 65 sequences spanning full-length (196 aa) or nearly full-length RT from *H. hampei* and 54 reference sequences (Table S2) permitted the identification of sequences belonging to most of the previously identified clades described in insects (eight out of twelve, i.e., *R1*: 26 sequences; *Jockey*: 11 sequences; I/*Nimb*: 6 sequences; *CR1*: 6 sequences; *Kiri*: 6 sequences; *Daphne, R2* and *R4*: 1 sequence each) (Biedler and Tu 2006) (Fig. 6; Table S1-B), indicating a high diversity of non-LTR retrotransposons in this species. Moreover, we identified seven sequences (classified as clade I according to *tBLASTn*) representing a novel non-LTR clade, which we named *Hamp*. These sequences, shown in orange in Fig. 6, cluster together in a monophyletic clade (bootstrap value 100%) apart from the sequences clearly classified in previously described non-LTR clades.

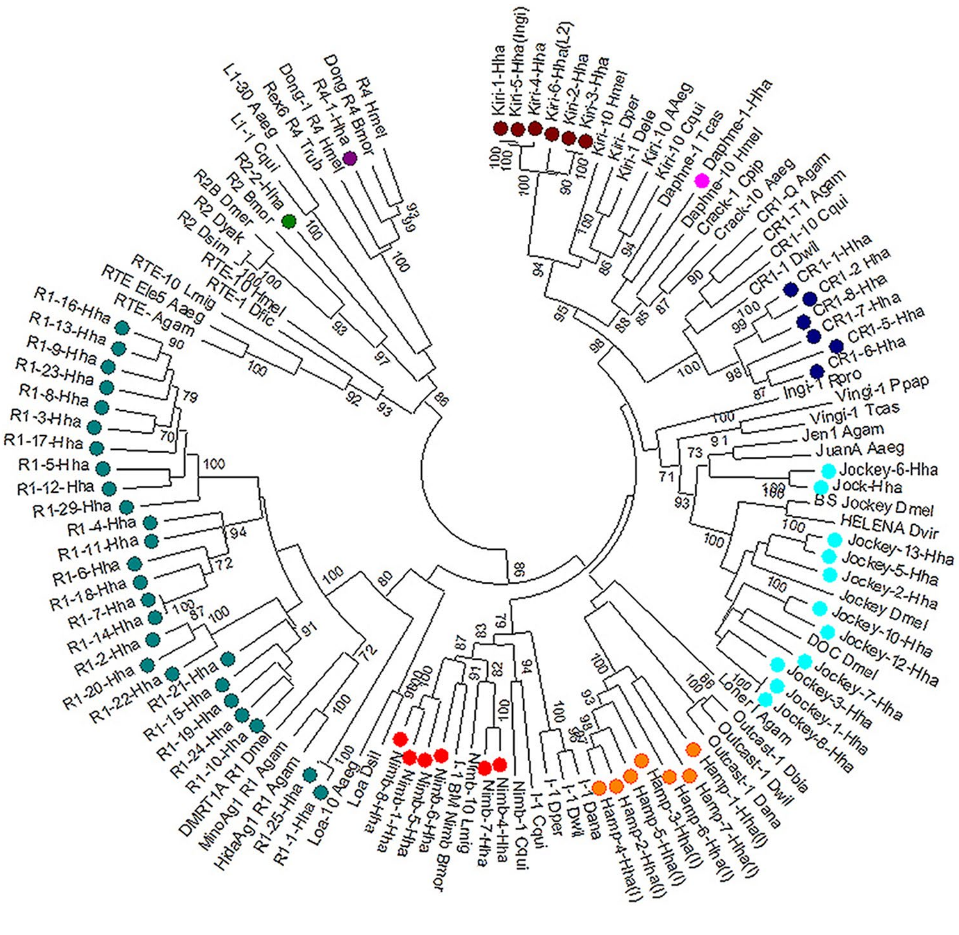


Fig. 6 Phylogenetic relationships of 65 consensus sequences from *H. hampei* spanning the RT domain and 54 reference sequences from the most representative superfamilies within the class I, non-LTR order. The MSA were visually inspected and highly variable regions for which the positional homology could not be clearly determined, were excluded. The phylogeny was inferred using the maximum likelihood method with 100 bootstrapping resamplings (values over 70% are shown) as implemented in MEGA5. The evolutionary distances were computed using the General Reverse Transcriptase + Freq. method and are in the units of the number of amino acid substitutions per site. The analysis involved 118 amino acid sequences. There were a total of 337 positions in the final dataset. *Different color dots* were used to indicate the different lineages to which the *H. hampei* sequences belong to. *Brown: Kiri; pink: Daphne; blue: CR1; light blue: Jockey; orange: Hamp; red: Nimb; turquoise: R1; green: R2* and *purple: R4*. (Color figure online)

DNA transposons identification

The search for DNA transposons resulted in 48 sequences classified as TIR elements (DNA transposon subclass 1), 66 as *Helitron*-like sequences and 488 as MITEs (Table S1D). Most of the DNA transposons from the TIR order belong to the superfamily *Tc1-mariner*, a result that is not surprising considering that this superfamily is present in all eukaryotes and can be successfully horizontally transferred between species, thereby ensuring its persistence and ubiquity. Other identified elements correspond to the *Merlin* (3), *P*-element (3), *PiggyBAC* (2), *hAT* (2) and *PIF-Harbinger* (1) superfamilies (Table S1-C). The structure-based approach to identify *Helitrons* resulted in the inclusion of 66 sequences in the repeat library (Table S1-E).

A multiple sequence alignment was performed with 33 *Tc1-mariner*-like sequences covering the full-length transposase domain (114 aa) and 81 reference sequences corresponding to *Tc1, mariner, Pogo and rosa* families previously described in this superfamily (Table S2). The alignment spanning over 180 aa positions was used for neighbor joining (data not shown) and maximum likelihood phylogenetic (Fig. 7) analyses. The resulting phylogenies indicated the presence of three groups of sequences corresponding to the *Tc1* (14 sequences), *Pogo* (5 sequences) and *rosa* (14 sequences) families. Conversely, no sequences

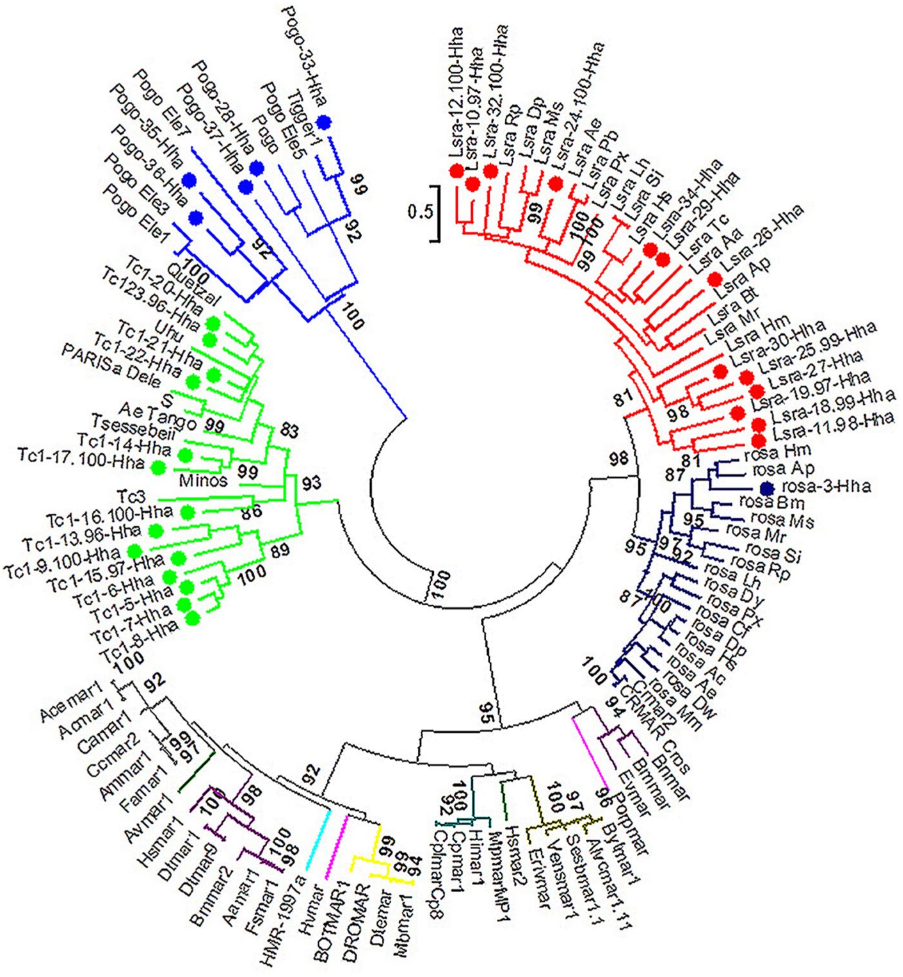


Fig. 7 Phylogenetic relationships of 33 consensus sequences spanning the transposase domain from *H. hampei* and 81 reference sequences from the *Pogo, Tc1, mariner, rosa* and *Lsra* families. The MSA were visually inspected and highly variable regions for which the positional homology could not be clearly determined, were excluded. The phylogeny was inferred using the maximum likelihood method with 100 bootstrapping resamplings (values over 70% are shown) as implemented in MEGA5. The evolutionary distances were computed using the General Reverse Transcriptase + Freq. method and are in the units of the number of amino acid substitutions per site. The analysis involved 114 amino acid sequences. There were a total of 195 positions in the final dataset. *Different color lines* were used to indicate different families within the Class II as follows: *dark blue: Pogo; light green: Tc1; red: Lsra; dark blue: rosa. Colored dots* were used to represent the sequences isolated from *H. hampei*. All the other sequences correspond to the subfamilies within the *mariner* family. (Color figure online)

belonging to the *mariner* family were identified. The *rosa* monophyletic group has been proposed as a basal subfamily of the *mariner* family (Gomulski et al. 2001).

Using a broader set of reference sequences from other insects, we identified a unique sequence, *rosa-3-Hha*, clustering with the *rosa* reference sequences, with a high bootstrap value (95%), and 13 sequences clustering together with *Lsra* sequences, recently described as a novel sister clade of the *rosa* subfamily (Zhang and Shen 2016). The main distinguishing characteristics of the *Lsra* elements are the D,D41D motif, a variant of the D,D34D catalytic domain of *mariner* elements, and large TIRs (ranging from 35 to 1878 bp); in contrast, *rosa* elements have shorter TIRs (28–45 bp). The *Lsra* sequences in *H. hampei* harbor the D,D41D motif (Fig. 8) and long TIRs ranging from 133 to 708 nt. For comparison, the only *rosa* sequence identified in this study contains TIRs of 33 nt, within the range of the reference *rosa* TIRs. The TIR sequences of the *Lsra* elements in *H. hampei* do not show sequence conservation.

A search for MITEs performed with MITE-Hunter resulted in 104 sequences grouped into 38 families of 2–12 members. In addition, 384 singlet sequences were found (Table S1-D). Most of the families were identified as related to the *Tc1-mariner* superfamily (Charlesworth and Langley 1986). Three families are related to *MULE*, one to *Pif-Pong-Tourist* and one to *CATCA*. Finally, one family was classified as unknown. A total of 46 long elements share the same terminal inverted repeat with a particular MITE; 19 of these match the transposase domain and most likely represent autonomous elements associated with only five of the MITE families identified here. In total, 13 long coding elements related to MITEs do not belong to any known family or singlet sequence.

Annotation of TEs

A total of 75,170 matches against the repeat library were detected in the CBB genome, occupying at least 18.1 Mb, or 8.27%, of the total draft sequence (Fig. 9; Table 1). For Class I elements, LTRs correspond to 16.7% of the annotated TEs, DIRS to only 1.37% and LINEs to 10.65%, for a total TE content of 27.5%. For Class II elements, TIRs correspond to 13.8%, *Helitrons* to 9.62% and MITES to 49.24%, for a total TE content of 76.25%.

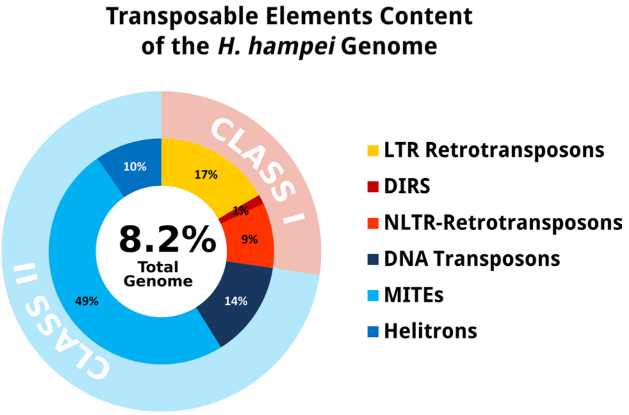
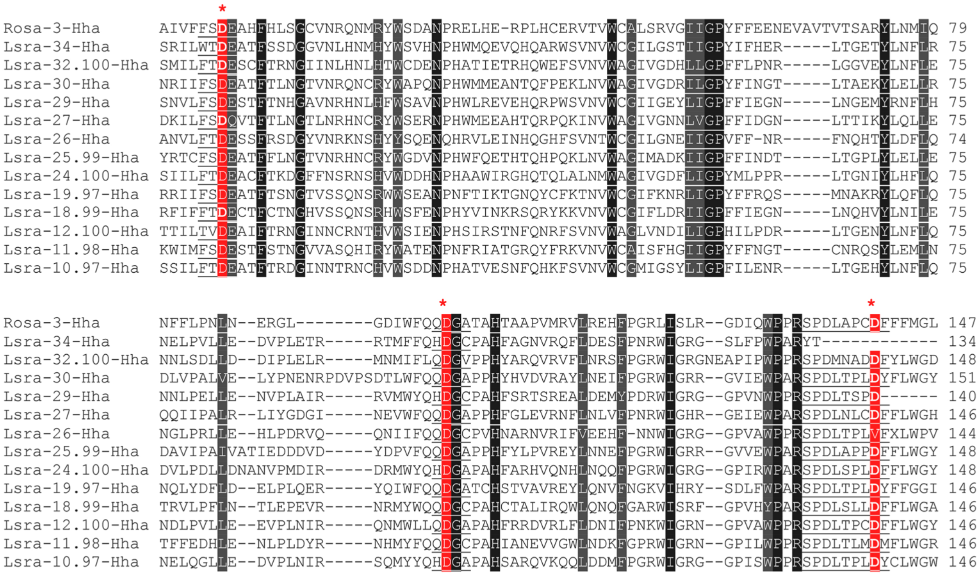


Fig. 9 Transposable elements content of the *H. hampei* draft genome



```
                      *
Rosa-3-Hha      AIVFFSDEAHFHLSGCVNRQNMRYWSDANPRELHE-RPLHCERVTVWCALSRVGIIGPYFFEENEVAVTVTSARYLNMIQ 79
Lsra-34-Hha     SRILWTDEATFSSDGGVNLHNMHYWSVHNPHWMQEVQHQARWSVNVWCGILGSTIIGPYIFHER-----LTGETYLNFLR 75
Lsra-32.100-Hha SMILFTDESCFTRNGIINLHNLHTWCDENPHATIETRHQWEFSVNVWAGIVGDHLLGPFFLPNR-----LGGVEYLNFLE 75
Lsra-30-Hha     NRIIFSDEATFTLNGTVNRQNCRYWAPQNPHWMMEANTQFPEKLNVWAGIVGDRILGPYFINGT-----LTAEKYLELLR 75
Lsra-29-Hha     SNVLFSDESTFTNHGAVNRHNLHFWSAVNPHWLREVEHQRPWSVNVWCGIIGEYLIGPFFIEGN-----LNGEMYRNFLH 75
Lsra-27-Hha     DKILFSDQVTFTLNGTLNRHNCRYWSERNPHWMEEAHTQRPQKINVWAGIVGNNLVGPFFIDGN-----LTTIKYLQLLE 75
Lsra-26-Hha     ANVLFTDESSFRSDGYVNRKNSHYYSQENQHRVLEINHQGHFSVNTWCGILGNEIIGPVFF-NR-----FNQHTYLDFLQ 74
Lsra-25.99-Hha  YRTCFSDEATFFLNGTVNRHNCRYWGDVNPHWFQETHTQHPQKLNVWAGIMADKIIGPFFINDT-----LTGPLYLELLE 75
Lsra-24.100-Hha SIILFTDEACFTKDGFFNSRNSHVWDDHNPHAAWIRGHQTQLALNMWAGIVGDFLIGPYMLPPR-----LTGNIYLHFLQ 75
Lsra-19.97-Hha  RRIIFSDEATFTSNGTVSSQNSRWWSEANPNFTIKTGNQYCFKTNVWCGIFKNRLIGPYFFRQS-----MNAKRYLQFLQ 75
Lsra-18.99-Hha  RFIFFTDECTFCTNGHVSSQNSRHWSFENPHYVINKRSQRYKKVNVWCGIFLDRIIGPFFIEGN-----LNQHVYLNILE 75
Lsra-12.100-Hha TTILTVDEAIFTRNGINNCRNTHVWSIENPHSIRSTNFQNRFSVNVWAGLVNDILIGPHILPDR-----LTGENYLNFLQ 75
Lsra-11.98-Hha  KWIMFSDESTFSTNGVVASQHIRYWATENPNFRIATGRQYFRKVNVWCAISFHGIIGPYFFNGT-----CNRQSYLEMLN 75
Lsra-10.97-Hha  SSILFTDEATFTRDGINNTRNCHVWSDDNPHATVESNFQHKFSVNVWCGMIGSYLIGPFILENR-----LTGEHYLNFLQ 75

                                             *                                           *
Rosa-3-Hha      NFFLPNLN--ERGL--------GDIWFQQDGATAHTAAPVMRVLREHFPGRLISLR--GDIQWPPRSPDLAPCDFFFMGL 147
Lsra-34-Hha     NELPVLLE--DVPLETR-----RTMFFQHDGCPAHFAGNVRQFLDESFPNRWIGRG--SLFPWPARYT------------ 134
Lsra-32.100-Hha NNLSDLLD--DIPLELR-----MNMIFLQDGVPPHYARQVRVFLNRSFPGRWIGRGNEAPIPWPPRSPDMNADDFYLWGD 148
Lsra-30-Hha     DLVPALVE--LYPNENRPDVPSDTLWFQQDGAPPHYHVDVRAYLNEIFPGRWIGRR--GVIEWPARSPDLTPLDYFLWGY 151
Lsra-29-Hha     NNLPELLE--NVPLAIR-----RVMWYQHDGCPAHFSRTSREALDEMYPDRWIGRG--GPVNWPPRSPDLTSPD------ 140
Lsra-27-Hha     QQIIPALR--LIYGDGI-----NEVWFQQDGAPPHFGLEVRNFLNLVFPNRWIGHR--GEIEWPARSPDLNLCDFFLWGH 146
Lsra-26-Hha     NGLPRLLE--HLPDRVQ-----QNIIFQQDGCPVHNARNVRIFVEEHF-NNWIGRG--GPVAWPPRSPDLTPLVFXLWPV 144
Lsra-25.99-Hha  DAVIPAIVATIEDDDVD-----YDPVFQQDGAPPHFYLPVREYLNNEFPGRWIGRR--GVVEWPARSPDLAPPDFFLWGY 148
Lsra-24.100-Hha DVLPDLLDNANVPMDIR-----DRMWYQHDGAPAHFARHVQNHLNQQFPGRWIGRG--GPIRWPARSPDLSPLDFFLWGY 148
Lsra-19.97-Hha  NQLYDFLD--ELPLQER-----YQIWFQQDGATCHSTVAVREYLQNVFNGKVIHRY--SDLFWPARSPDLTPLDYFFGGI 146
Lsra-18.99-Hha  TRVLPFLN--TLEPEVR-----NRMYWQQDGCPAHCTALIRQWLQNQFGARWISRF--GPVHYPARSPDLSLLDFFLWGA 146
Lsra-12.100-Hha NDLPVLLE--EVPLNIR-----QNMWLLQDGAPAHFRRDVRLFLDNIFPNKWIGRN--GPVAWPPRSPDLTPCDFFLWGY 146
Lsra-11.98-Hha  TFFEDHLE--NLPLDYR-----NHMYFQQDGCPAHIANEVVGWLNDKFGPRWIGRN--GPILWPPRSPDLTLMDMFLWGR 146
Lsra-10.97-Hha  NELQGLLE--DVPLNIR-----SQMYYQHDGAPAHSARQVKQQLDDMFPGRWIGRG--GPISWPPRSPDLTPLDYCLWGW 146
```

Fig. 8 Identification of DD41D in one *rosa* and 12 *Lsra* sequences according to the Color Align Conservation results (http://www.bioinformatics.org/sms2/color_align_cons.html). The *asterisks* correspond to the predicted DDD sites

LTR retrotransposons

We analyzed the proportion of the genome corresponding to complete or fragmented LTR retrotransposons; this portion corresponded to ~3.3 Mb, equivalent to 1.38% of the total genome assembly, distributed among 2,823 scaffolds. The divergence levels of the LTR retrotransposon copies, which were approximately 19% (substitutions in matching regions compared to the consensus), were similar for the three superfamilies (*Gypsy, Bel-Pao* and *Copia*), and indels were close to 2% (Table 1). The *Gypsy* elements account for the most important LTR fraction, corresponding to 1.08% of the genome; *Bel-Pao* sequences account for 0.24% and *Copia* only 0.06%. The most numerous families are *Copia*-6, with 21 fragments detected, *Gypsy*-41 and *Gypsy*-31, with 17 copies each, followed by *Gypsy*-3 with 15 copies, *Bel*-3 with 13 and *Gypsy*-52 and *Gypsy*-53 with 12 copies each. Forty-two solo LTRs, i.e., sequences representing one or both flanking LTRs, were identified; 35 of these correspond to *Gypsy*, four to *Bel-Pao* and three to *Copia*.

Non-LTR retrotransposons

H. hampei harbors a high diversity of elements belonging to the main clades of non-LTR retrotransposons (approximately 1.7 Mb distributed among 972 scaffolds corresponding to 0.76% of the genome). The most abundant elements belong to the *R1* and *Jockey* clades, with 301 and 110 copies, respectively. The copies show divergence values of up to 40%, with *R1* and *Jockey* elements being the most divergent and *R4* and *Kiri* less divergent elements (less than 20%). As expected for dead-on-arrival elements, most of the non-LTR families/clades identified are fragmented, and therefore inactive. Nevertheless, several representative sequences of various clades were found in more than one complete copy, with low divergence. *Jockey* elements are represented by 13 putative complete copies corresponding to different families; also, 47 complete copies corresponding to the *R1* clade were identified, with the most abundant families corresponding to *R1*-29 (eight copies) and *R1*-18 (five copies). Another clade displaying a high number of complete copies was *Nimb*, with 14 sequences distributed among different families. Considering that the diversity of non-LTR families in *Anopheles gambiae*, including 31 families in clade *CR1*, 25 in *Jockey* and 16 in *R1*, is reported as being unprecedented (Biedler and Tu 2003), the non-LTR diversity in the CBB genome is outstanding. Our approach also allowed the identification of a novel non-LTR clade called *Hamp*. It is important to note that the complete copies mentioned above display remarkable ORF integrity and low divergence (less than 20% for most of them). These findings suggest that transpositional activity from a source element may have occurred recently in these clades. Our analysis provided no evidence of the existence of SINEs in the genome of *H. hampei*.

DNA transposons

DNA transposons consist of 9499 regions distributed among 3955 different scaffolds (approximately 2.5 Mb of the CBB genome assembly), occupying 1.94% of the genome. The divergence levels are 20.5% for TIR elements and 22.5% for *Helitrons* (Table 1). The *Tc1-mariner* superfamily exceeds the other superfamilies by fivefold. The large content of highly fragmented Class II elements is an expected finding because nucleotide loss is the main cause of vertical inactivation in this class. The great diversity of elements and the gradient in the degree of conservation and/or deterioration suggest a distant ancestry for many *Tc1-mariner* elements. The content of *Helitron*-like sequence covers approximately 1.4 Mb of the *H. hampei* genome assembly, occupying 3551 scaffolds and corresponding to 9% of the total TE content. Of the 8,158 total matches, only 879 could be considered putative *Helitron-like* sequences.

MITEs

MITEs constitute a large proportion of the total TE content of the *H. hampei* genome and globally represents 4.08% of the genome. The total number of hits was 48,321, distributed among 9605 scaffolds. Two superfamilies are represented by more than 1000 fragments (F29 and F16), with the F29.2 family being the most abundant (2018 fragments) and classified as unknown. The F16 family, which is associated with *Tc1-mariner*, is the second most abundant. A total of 172 complete copies (>90% of the reference) of the F16.7 family were found throughout the genome, and more than 100 of these are 100% intact copies. This may suggest high cross-mobilization inside this superfamily; however, for the DNA transposon repeat library, no associated full-length copies were found as counterparts that could be responsible for their mobilization.

Discussion

Implementation of a systematic analysis for estimating the total content of TEs resulted in an efficient strategy. To date, this is the most complete characterization of the TEs content of a coleopteran genome. A total of 75,170 matches against the TE repeat library were detected in the *H. hampei* genome, occupying at least 18.1 Mb, or 8.2%, of the total draft sequence. The TE content varies among insect genomes by more than an order of magnitude. In *Drosophila*, the values range from ~2.7% in *D. simulans,* *D. pseudoobscura* and *D. grimshawi* to 25% in *D. ananassae* (Clark and Eisen 2007). For coleopteran sequenced genomes, the TE content is equally variable, ranging from 5 to 6% (Richards and Gibbs 2008; Wang and Lorenzen 2008) in *Tribolium castaneum*, 15.4% in *Dendroctonus ponderosae* (Keeling et al. 2013) and to 29.2% in *Oryctes borbonicus* (Meyer and Markov 2016). It should be noted that these TE annotations did not include MITEs. Moreover, TE annotation of 16 Anopheline mosquito genomes, including MITE annotation, showed variation of the same order of magnitude for the total TE content, from 2.29% (*Anopheles darlingi*, Marinotti et al. 2013) to 17.78% (*A. gambiae*, Neafsey et al. 2014). Hence, the TE occupancy of the *H. hampei* genome is within the lower range of that of other insects.

Why species harbor different proportions of TEs has been a matter of debate, but it has been proposed to be mainly related to the reproductive characteristics and population size of the host (Kidwell 1977; Wrigth and Finnegan 2001). Sexual reproduction and outcrossing provide TEs with a means of spreading to all individuals in a population, but in asexual organisms, the rates of infection between different lines are reduced preventing or hampering the spread of TEs (Hickey 1982; Arkhipova and Meselson 2000). Due to the negative fitness consequences associated with TE insertions in coding and regulatory regions, selection against TE may be less effective in smaller populations (Brookfield and Badge 1997) and the fraction of TEs capable of drifting to fixation must decline with increasing N_e. Nevertheless, specialized species with small population size may also harbor low TE copy numbers (Capy et al. 1992; Granzotto et al. 2009). The coffee berry borer CBB presents particularities regarding both the reproductive characteristics and the population size that may be related to low TE content. Its functional haplodiploidy, in which males have a condensed, probably non-functional chromosome set (Brun et al. 1995) that is not transmitted to the next generation (Borsa and Kjellberg 1996) could partly prevent the spread of transposable elements, likewise in asexual populations. Population size variation along the CBB evolutionary history, associated to its introduction into the Americas, with possible founder effect, as well as to fluctuations along the years according to local coffee bean densities (Bustillo 2006) can have also played a role, reducing the total amount of TEs in nowadays CBB populations.

Novel TE families

Based on our library, we propose the existence of three novel families in the CBB genome, two belonging to Class I, *Hypo* and *Hamp* and *rosa* belonging to Class II. The two Class I families consist of completely new, undescribed sequences. *Hypo*, represented in our library by 14 sequences, belongs to the LTR *Gypsy* superfamily, and *Hamp*, represented by seven sequences, belongs to the non-LTR order. Each group of sequences cluster together in monophyletic clades (bootstrap values over 99%) apart from sequences clearly classified within the *Gypsy* superfamily and non-LTRs, respectively.

We also propose the creation of a family named *rosa*. The clade *rosa* was originally created using sequences of *Tc1-mariner* elements found in Tephritidae fruit flies (*Ceratitis rosa, C. capitata, Anastrepha ludens, A. suspense* and *Bactrocera tryoni*) as well as in *D. melanogaster* (Gomulski et al. 2001). Since this group of sequences clusters basally to *mariner* sequences, the authors proposed this clade as a divergent subfamily of the *mariner* family. Here, using a broader set of reference sequences from other insects, we identified two reciprocally monophyletic groups of *rosa*-like sequences (bootstrap: 98%): one composed of sequences previously identified as *rosa* and the CBB *rosa-3-Hha* sequence (bootstrap: 95%) and other composed of 14 *Lsra* sequences recently described in insects (Zhang and Shen 2016) and 13 *Lsra* sequences of *H. hampei* (bootstrap: 81%). The proposal to consider *rosa* as a bona fide family of *Tc1-mariner* is supported by its monophyly and due to the presence of conserved transposase catalytic domain with the D,D41D motif. Moreover, the sharing of short TIRs by the *rosa* clade and long TIRs by the *Lsra* clade and their reciprocal monophyly support this proposal as two bona fide subfamilies within the *rosa* family. We followed the law of priority for scientific classification and maintain the name *rosa* for the new family as well as for the previously described subfamily.

Rank-order abundances of TEs

In insect genomes, the proportion of different TE classes, orders and superfamilies varies as much as the total TE content. Using again the example of the rank order of TE abundance in genome with the two extreme TE contents in Anophelines, *A. darlingi* (Marinotti et al. 2013) and *A. gambiae* (Neafsey et al. 2014), both contain more Class I than Class II elements. However, these genomes differ in the order of their abundances (*A. darlingi*: non-LTRs > MITEs > LTRs > DNA transposons; *A. gambiae*: LTRs > MITEs > non-LTRs > DNA transposons). The content of MITEs in particular is also highly variable. As two extreme examples, Class II elements (3.55%) surpass Class I (2.07%) in the genome of the kissing bug *Rhodnius prolixus* (5.62% total TE content), and the majority of Class II elements belong to the *mariner* family (2.66%) of the *Tc1-mariner* superfamily, with MITEs corresponding to only 14% of this percentage (Fernandez-Medina et al. 2016; Mesquita et al. 2015). In contrast, although

Class II elements (19.5%) also surpass Class I (8.5%) in the mosquito *Culex quinquefasciatus* (28% of TEs), MITEs correspond to 87% of the Class II content (Arensburger et al. 2010).

In the CBB genome, the Class II content is threefold higher (6.02%) than the Class I (2.25%). The Class I elements surpass Class II when MITE elements are not considered (Fig. S1). Thus, for comparison with published genomes, we might analyze the rank-order abundance of TEs in CBB in two ways, including and not including MITES. When including MITEs, the order is MITEs > DNA transposons > LTRS > non-LTRs (Fig. 9); when not including MITEs, the order is DNA transposons > LTRs > non-LTRS (Fig. S1). Regardless, Class II elements are the main bulk of the *H. hampei* TE content, as in the above comparisons, and members of the *Tc1-mariner* superfamily together with MITEs comprise the main fraction.

Our strategy of first identifying non-autonomous DNA elements (according to the criteria of their TIRs and TSD sequences) gains importance because the number of MITEs present in this genome would have otherwise been seriously underestimated. Although the MITE sequences identified in the CBB genome do not appear to have been derived from autonomous DNA transposons, they have clearly been successful in their persistence in this genome. It is known that co-evolution between transposases and TIR sequences occurs, and any changes in transposase sequence are most likely accompanied by changes in TIR sequences (Feschotte et al. 2003; Lampe et al. 2001). Therefore, it is suggested that cross-mobilization is inefficient to a certain degree of identity (Lampe et al. 2001). Despite the great diversity of *Tc1-mariner* elements in the *H. hampei* genome, it is possible that ancient DNA transposons have been lost or inactivated due to loss of some of the TIRs (a possible reason why they were not structurally identified). Nevertheless, the existence of some transposases belonging to unknown elements that target sequence motifs in MITEs might enhance their cross-mobilization and amplification (Fattash et al. 2013). Finally, the large proportion of MITEs compared to the TE content found in the *H. hampei* genome could have originated de novo from rearrangements of palindromic sequences, which are common in eukaryotes (Feschotte et al. 2003).

The causes (or even the consequences) of the high degree of variability in the distribution, amount and relative proportion of TEs in different genomes are not well understood. However, it is still important to keep on characterizing this important fraction of eukaryotic genomes, since they can bring light into evolutionary phenomena and genomic rearrangements that have occurred in the past.

Concluding remarks

Despite its low TE content, the *H. hampei* genome presents a high diversity of TE superfamilies. Although most of the sequences appear to be degenerate, a few elements are present in at least one copy with an intact structure, suggesting recent transposition and consequently activity. Our study is the first contribution to the knowledge of the composition of this genomic fraction of the coffee berry borer. However, further studies are necessary to elucidate the functional relationship of TEs to the evolution of the *H. hampei* genome and to determine whether insertional variation of TEs at the inter-population level exists and whether these sequences can be used as new genetic markers in innovative pest control strategies.

Acknowledgements We gratefully acknowledge funding from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Grant No. 2013/15070-4) and from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Grant No. 306493/2013-6) to CMAC, FAPESP fellowship Grant No. 12/24813-8 to EHH, fellowship from Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (CAPES/FAPERJ) to RDFM and Ministerio de Agricultura y Desarrollo Rural de Colombia to PBM, JN and LNE.

Compliance with ethical standards

Conflict of interest Author EMHH, author RDFM, author LNE, author JN, author PBM and author CMAC declare that they have no conflict of interest.

Ethical standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with human participants or animals performed by any of the authors.

References

Andreev D, Breilid H, Kirkendall L, Brun LO, ffrench-Constant RH (1998) Lack of nucleotide variability in a beetle pest with extreme inbreeding. Insect Mol Biol 7:197–200

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 (2010)

Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science 330:86–88

Arkhipova I, Meselson M (2000) Transposable elements in sexual and ancient asexual taxa. Proc Natl Acad Sci USA 97:14473–14477

Benavides P, Vega FE, Romero-Severson J, Bustillo AE, Stuart JJ (2005) Biodiversity and biogeography of an important inbred pest of coffee, coffee berry borer (Coleoptera: Curculionidae: Scolytinae). Ann Entomol Soc Am 98:359–366

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

Borsa P, Kjellberg F (1996) Secondary sex ratio adjustment in a Pseudo-arrenotokous insect, *Hypothenemus hampei* (Coleoptera: Scolytidae). CRASP 319:1159–1166

Brun LO, Stuart J, Gaudichon V, Aronstein K, French-Constant RH (1995) Functional haplodiploidy: a mechanism for the spread of insecticide resistance in an important international insect pest. Proc Natl Acad Sci USA 92:9861–9865

Bustillo AE (2006) Una revisión sobre la broca del café, *Hypothenemus hampei* (Coleoptera: Curculionidae: Scolytinae), en Colombia. Revista Colombiana de Entomología 32:101–116

Capy P, David JR, Hartl DL (1992) Evolution of the transposable element *mariner* in the *Drosophila melanogaster* species group. Genetica 86:37–46

Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37(Database issue):D93–D97

Charlesworth B, Langley CH (1986) The evolution of self-regulated transposition of transposable elements. Genetics 112:359–383

Clark AG, Eisen MB (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218

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

de la Chaux N, Wagner A (2011) BEL/Pao retrotransposons in metazoan genomes. BMC Evol Biol 11:154

Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18

Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P, Yang G (2013) Miniature inverted-repeat transposable elements: discovery, distribution, and activity. Genome 56:475–486

Fernández-Medina RD, Granzotto A, Ribeiro JM, Carareto CM (2016) Transposition burst of mariner-like elements in the sequenced genome of *Rhodnius prolixus*. Insect Biochem Mol Biol 69:14–24

Feschotte C, Swamy L, Wessler SR (2003) Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with Stowaway miniature inverted repeat transposable elements (MITEs). Genetics 163(2):747–758

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue):D222–D230

Flutre T, Duprat E (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6:e16526

Gao C, Xiao M, Ren X, Hayward A, Yin J, Wu L, Fu D, Li J (2012) Characterization and functional annotation of nested transposable elements in eukaryotic genomes. Genomics 100:222–230

Gauthier N (2010) Multiple cryptic genetic units in *Hypothenemus hampei* (Coleoptera: Scolytinae): evidence from microsatellite and mitochondrial DNA sequence data. Biol J Linn Soc 101:113–129

Gingerich DP, Borsa P, Suckling DM, Brun L-O (1996) Inbreeding in the coffee berry borer, *Hypothenemus hampei* (Coleoptera: Scolytidae) estimated from endosulfan resistance phenotype frequencies. Bull Entomol Res 86:667–674

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

Goodwin T, Poulter R (2004) A new group of tyrosine recombinase-encoding retrotransposons. Mol Biol Evol 21:746–759

Granzotto A, Lopes FR, Lerat E, Vieira C, Carareto C (2009) The evolutionary dynamics of the *Helena* retrotransposon revealed by sequenced *Drosophila* genomes. BMC Evol Biol 9:174

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199

Hickey D (1982) Selfish DNA: a sexually transmitted nuclear parasite. Genetics 101:519–531

Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, Fiston-Lavier AS, Hua-Van A, Hubley R, Kapusta A, Lerat E, Maumus F, Pollock DD, Quesneville H, Smit A, Wheeler TJ, Bureau TE, Blanchette M. (2015) A call for benchmarking transposable element annotation methods. Mob DNA 6:13

Infante F, Pérez J, Vega FE (2014) The coffee berry borer: the centenary of a biological invasion in Brazil. Braz J Biol 74:S125–S126

Jiang N (2013) Computational methods for identification of DNA transposons. In: Peterson T (ed) Plant transposable elements SE—21, vol 1057. Human Press, pp 289–304. *Methods in Molecular Biology*

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kapitonov VV, Tempel S, Jurka J (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene 448:207–213

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780

Keeling CI, Yuen MM, Liao NY, Docking TR, Chan SK, Taylor GA, Palmquist DL, Jackman SD, Nguyen A, Li M, Henderson H, Janes JK, Zhao Y, Pandoh P, Moore R, Sperling FA, Huber DP, Birol I, Jones SJ, Bohlmann J (2013) Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. BMC Genomics 21(14):198

Kidwell MG (1977) Reciprocal differences in female recombination associated with hybrid dysgenesis in *Drosophila melanogaster*. Genet Res 30:77–88

Kidwell MG, Lisch DR (2000) Transposable elements, parasitic DNA, and genome evolution. Evolution (NY) 55:1–24

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

Levin HL, Moran JV (2011) Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12:615–627

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, Aguilar-Rodríguez J, Vicente-Ripolles M, Fuster G, Bernet GP, Maumus F, Munoz-Pomer A, Sempere JM, Latorre A, Moya A (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39(Database issue):D70–D74

Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16:793–805

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 (2013) The genome of *Anopheles darlingi*, the main neotropical malaria vector. Nucleic Acids Res 41:7387–7400

Mátés L, Izsvák Z, Ivics Z (2007) Technology transfer from worms and flies to vertebrates: transposition-based genome manipulations and their future perspectives. Genome Biol 8(Suppl 1):S1

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, Lange AB, Latorre-Estivalis JM, Lavor 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, Orchard I, 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, Huebne E, Dotson EM, Oliveira PL (2015) Genome of Rhodnius prolixus, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc Natl Acad Sci USA 112:14936–14941

Meyer JM, Markov GV (2016) Draft Genome of the Scarab Beetle *Oryctes borbonicus* on La Réunion Island. Genome Biol Evol 8(7):2093–2105

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger 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 (2015) Mosquito genomics. Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. Science 347:1258522

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyn B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316:1718–1723

Pardue ML, Danilevskaya ON, Traverse KL, Lowenhaupt K (1997) Evolutionary links between telomeres and transposable elements. Genetica 100:73–84

Permal E, Flutre T, Quesneville H (2012) Roadmap for annotating transposable elements in eukaryote genomes. Methods Mol Biol 859:53–68

Piégu B, Bire S, Arensburger P (2015) 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 86:90–109

Platt RN 2nd, Blanco-Berdugo L, Ray DA (2016) Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol 8(2):403–410

Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D (2005) Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol 1:166–175

Rho M, Tang H (2009) MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Res 37:e143

Richards S, Gibbs RA (2008) The genome of the model beetle and pest *Tribolium castaneum*. Nature 452:949–955

Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res 37:7002–7013

Steinbiss S, Kastens S, Kurtz S (2012) LTRsift: a graphical user interface for semi-automatic classification and postprocessing of de novo detected LTR retrotransposons. Mob DNA 3:18

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739 van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530–536

Vega FE, Brown SM, Chen H, Shen E, Nair MB, Ceja-Navarro JA, Brodie EL, Infante F, Dowd PF, Pain A. (2015) Draft genome of the most devastating insect pest of coffee worldwide: the coffee berry borer, *Hypothenemus hampei*. Sci Rep 5:12525

Wang S, Lorenzen MD (2008) Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. Genome Biol 9(3):R61

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 (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wong LH, Choo KH (2004) Evolutionary dynamics of transposable elements at the centromere. Trends Genet 20:611–616

Wright S, Finnegan D (2001) Genome evolution: sex and the transposable element. Curr Biol 11(8):R296–R299

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9(10):3353–3362

Xiong W, He L, Lai J, Dooner HK, Du C (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci USA 111:10263–10268

Yang G, Hall TC (2003) MAK, a computational tool kit for automated MITE analysis. Nucleic Acids Res 31:3659–3665

Zhang H-H, Shen Y-H (2016) Identification and evolutionary history of the DD41D transposons in insects. Genes Genomics 38(2):109–117