Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families

Elaine Silva Dias, Clémence Hatt, Perla Hamon, Serge Hamon, Michel Rigoreau, Dominique Crouzillat, Claudia Marcia Aparecida Carareto, Alexandre De Kochko, Romain Guyot

Large distribution and high sequence identity of a *Copia*-type retrotransposon in angiosperm families

Elaine Silva Dias[1,3] (elainedias_bio@yahoo.com.br)

Clémence Hatt[1] (clemhatt@gmail.com)

Serge Hamon[1] (serge.hamon@ird.fr)

Perla Hamon[1] (perla.hamon@ird.fr)

Michel Rigoreau[2] (michel.rigoreau@rdto.nestle.com)

Dominique Crouzillat[2](dominique.crouzillat@rdto.nestle.com)

Claudia Marcia Aparecida Carareto[3] (carareto@ibilce.unesp.br)

Alexandre de Kochko[1] (alexandre.dekochko@ird.fr)

Romain Guyot[4]* (romain.guyot@ird.fr)

[1]IRD UMR DIADE, EVODYN, BP 64501, 34394 Montpellier Cedex 5, France

[2]Nestlé R&D Tours, 101 AV. G. Eiffel, Notre Dame d'Oe´, BP 49716 37097, Tours, Cedex 2, France

[3]UNESP – Univ. Estadual Paulista, Department of Biology, São José do Rio Preto, SP, Brazil.

4 IRD UMR IPME, COFFEEADAPT, BP 64501, 34394 Montpellier Cedex 5, France

*Corresponding Author: Romain Guyot, Institut de Recherche pour le Développement (IRD), UMR IPME, BP 64501, 34394 Montpellier Cedex 5, France, +33467416455, romain.guyot@ird.fr

Number of Figures: 5

9,913 words

Number of Tables: 3

1 Supplementary file

Abstract

Retrotransposons are the main component of plant genomes. Recent studies have revealed the complexity of their evolutionary dynamics. Here, we have identified *Copia25* in *Coffea canephora*, a new plant retrotransposon belonging to the *Ty1-Copia* superfamily. In the *Coffea* genomes analyzed, *Copia25* is present in relatively low copy numbers and transcribed. Similarity sequence searches and PCR analyses show that this retrotransposon with LTRs (Long Terminal Repeats) is widely distributed among the Rubiaceae family and that it is also present in other distantly related species belonging to Asterids, Rosids and monocots. A particular situation is the high sequence identity found between the *Copia25* sequences of *Musa*, a monocot, and *Ixora*, a dicot species (Rubiaceae). Our results reveal the complexity of the evolutionary dynamics of the ancient element *Copia25* in angiosperm, involving several processes including sequence conservation, rapid turnover, stochastic losses and horizontal transfer.

Large distribution and high sequence identity of a *Copia*-type retrotransposon in angiosperm families

Authors and Affiliations

Elaine Silva Dias[1,3] (elainedias_bio@yahoo.com.br)

Clémence Hatt[1] (clemhatt@gmail.com)

Serge Hamon[1] (serge.hamon@ird.fr)

Perla Hamon[1] (perla.hamon@ird.fr)

Michel Rigoreau[2] (michel.rigoreau@rdto.nestle.com)

Dominique Crouzillat[2] (dominique.crouzillat@rdto.nestle.com)

Claudia Marcia Aparecida Carareto[3] (carareto@ibilce.unesp.br)

Alexandre de Kochko[1] (alexandre.dekochko@ird.fr)

Romain Guyot[4]* (romain.guyot@ird.fr)

[1]IRD UMR DIADE, EVODYN, BP 64501, 34394 Montpellier Cedex 5, France

[2]Nestlé R&D Tours, 101 AV. G. Eiffel, Notre Dame d'Oe´, BP 49716 37097, Tours, Cedex 2, France

[3]UNESP – Univ. Estadual Paulista, Department of Biology, São José do Rio Preto, SP, Brazil.

[4]IRD UMR IPME, COFFEEADAPT, BP 64501, 34394 Montpellier Cedex 5, France

*Corresponding Author: Romain Guyot, Institut de Recherche pour le Développement (IRD), UMR IPME, BP 64501, 34394 Montpellier Cedex 5, France, +33467416455, romain.guyot@ird.fr

Data deposition: KM439056 to KM439101

Abstract

Retrotransposons are the main component of plant genomes. Recent studies have revealed the complexity of their evolutionary dynamics. Here, we have identified *Copia25* in *Coffea canephora*, a new plant retrotransposon belonging to the *Ty1-Copia* superfamily. In the *Coffea* genomes analyzed, *Copia25* is present in relatively low copy numbers and transcribed. Similarity sequence searches and PCR analyses show that this retrotransposon with LTRs (Long Terminal Repeats) is widely distributed among the Rubiaceae family and that it is also present in other distantly related species belonging to Asterids, Rosids and monocots. A particular situation is the high sequence identity found between the *Copia25* sequences of *Musa*, a monocot, and *Ixora*, a dicot species (Rubiaceae). Our results reveal the complexity of the evolutionary dynamics of the ancient element *Copia25* in angiosperm, involving several processes including sequence conservation, rapid turnover, stochastic losses and horizontal transfer.

Keywords

Copia25, transposable element, genome dynamics, sequence conservation, horizontal transfer, Rubiaceae.

Introduction

Transposable elements (TEs) are the major component of plant genomes. TEs are typically "vertically" transmitted from parent to offspring. If a new insertion occurs in germ cells tissues, the new copy will be transmitted to the progeny. In certain cases, TEs can be horizontally transferred (HT) between reproductively isolated species. Although more than 200 cases of HT have been reported most of them involve animals (Schaack et al. 2010), mainly insects (mostly *Drosophila*), and few potential cases have been reported in plants (Cheng et al. 2009; Diao et al. 2006; Fortune et al. 2008; Roulin et al. 2008) with the exception of a very recent observation (El Baidouri et al. 2014). The HTs concern both Class I (or Retrotransposon) and Class II (or Transposons) elements, and the mechanisms underlying TE HTs remain speculative in most of the cases (vectors could be pathogens, intracellular parasites, insects, etc.). Because TEs play a major role in the dynamics of genomes, their direct introduction into a "naïve" genome through HT may induce important consequences in chromosomal and genomic evolution. However, the detection of potential HT of TEs in complete genomes is relatively complex and requires highly sensitive methods to differentiate between unresolved sequence conservation and HT events (de Carvalho and Loreto 2012). In the absence of a clear mechanism underlying HT, cases of outstanding sequence conservation of TEs between evolutionarily distant plant species living in separate geographical areas have raised questions as to the existence of other mechanisms leading to this conservation (Moisy et al. 2014). The recent availability of plant genome sequences (Michael and Jackson 2013) gave new opportunities to identify and to characterize transposable elements and to gain a higher understanding of the evolutionary dynamics of these elements and their conservation between distantly related species.

The coffee genus (*Coffea*) that belongs to the Rubiaceae family, comprises 124 species, originating from Africa, Madagascar, the Mascarene Islands, Asia and Oceania (Davis 2010; Davis 2011). *Coffea* species are diploids (2n = 2x = 22) and generally allogamous. The notable exception is the self-fertilizing allotetraploid *Coffea arabica* (2n = 4x = 44), native to the Ethiopian highlands and originating from a recent hybridization of two different diploid ancestors, *C. canephora* and *C. eugenioides* (Lashermes et al. 1999; Yu et al. 2011). The current possibility of accessing genomic and transcriptomic sequences of *Coffea* species has made it possible to expand our knowledge of the composition and behavior of TEs in these important species. The analysis of the *C. canephora* genome showed that these sequences contained about 50% of transposable elements (Denoeud et al. 2014). The vast majority of them (85%) are retrotransposons with LTRs (LTR-RTs). The study of TEs in *Coffea* is very recent and the few individual TEs investigated to date show different dynamics between closely related coffee species (Hamon et al. 2011; Yuyama et al. 2012).

In this study, LTR-RTs were identified in the *C. canephora* genome using BAC-end sequences (BESs) and 454 sequences. One of them, a *Ty1-Copia* element named *Copia25*, was characterized and analyzed under different aspects of its evolution because its nucleotide sequence showed unusually high similarities with distantly related plant genomes. Furthermore, *Copia25* was found quite similar to *Rider*, an active retrotransposon identified in the tomato with a rather unique evolutionary history. *Rider* activity has played a role in the origin of at least three different phenotypes of this species (Jiang et al. 2009; Jiang et al. 2012; Xiao et al. 2008). Since it is absent in *Solanum tuberosum*, it has been suggested that *Rider* appeared in the tomato by HT from *Arabidopsis thaliana* (Cheng et al. 2009). The similarity shared between *Copia25* and *Rider* makes the TE identified in *C. canephora* interesting to investigate, particularly for its activity and evolutionary dynamics. In the current study, we show that *Copia25* is an active element in *Coffea*, widely present in Rubiaceae species. In

addition, a phylogenetic analysis indicates outstanding conservation of *Copia25* in coffee trees and in distantly related species, such as banana (*Musa* genus), a monocot. The different processes that can lead to high conservation of *Copia25* in Angiosperms are discussed.

Materials and Methods

Genome sequencing

The Next-Generation Sequencing (NGS – by Genomic 454 Pyrosequencing - GS Junior System Roche) was performed in two accessions of *C. canephora* Pierre ex A. Froehner (HD200-94 a double haploid from the Congolese diversity group, also used for whole genome sequencing – Denoeud et al. 2014, http://coffee-genome.org –, and BUD15 from Uganda), as well as in one accession from each of the following taxa: *C. arabica* L. (ET39 from Ethiopia), *C. eugenioides* S. Moore (DA56 from Kenya), *C. pseudozanguebariae* Bridson (08107 from Kenya), *C. heterocalyx* Stoff (JC65 from Cameroon), *C. racemosa* Lour (IA56 from Mozambique), *C. humblotiana* Baill (A.230 from Comoros), *C. millotii* J.-F. Leroy (ex-*dolichophylla*, A.206 from Madagascar) and *C. tetragona* Jum. & H. Perrier (A.252 from Madagascar), *Coffea* (ex-*Psilanthus*) *horsfieldiana* (Miq.) J.-F. Leroy (HOR from Indonesia) and *Craterispermum* Sp. Novo Kribi (from Cameroon) (Chevalier 1946; Maurin et al. 2007). The cultivars and the above-mentioned sequenced accessions grow in the IRD greenhouses (Montpellier, France), at the Kianjavato research station (Madagascar) or in the Nestlé R&D greenhouses (Tours, France). The total genomic DNA was extracted from young leaves using the Qiagen DNeasy Plant Mini Kit following the manufacturer's protocol. The library and sequencing for the NGS were performed at the Nestlé R&D laboratory according to the

Roche/454 Life Sciences Sequencing Method. Data were submitted to GenBank, BioProject PRJNA242989.

Sequence Analyses

We used 131,412 BAC end sequences (BESs) (Dereeper et al. 2013) obtained by Sanger sequencing and 106,459 sequences obtained by 454 Roche-NGS technology, both derived from the *C. canephora* HD200-94 accession. All sequences (Sanger and 454 Roche) were used for the assembly using AAARF (Assisted Automated Assembler of Repeat Families - DeBarry et al. 2008). The following parameters for the BLAST analyses and the Minimally Covered Sequences (MCS) construction and controlling "build" extensions were applied: minimum hit length: 150; minimum hit identity: 0.89; minimum coverage depth: 4; required MCS length: 150; maximum E-value: $1e^{-25}$; required coverage length: 150; minimum hit number: 2; required overlap between MCS and new query: 90; and maximum times a number sequence is used in each direction: 13. These parameters were those that gave best assembly results after several modification and assembly testing.

AAARF "builds" were analyzed using BLASTx (min E-value $1e^{-4}$) against public protein sequence databases (uniprot_sprot; http://www.uniprot.org/), and transposable element databases available in Repbase (Jurka et al. 2005 – http://www.girinst.org/repbase/) and Gypsy DB 2.0 (http://gydb.org - Llorens et al. 2011). The graphical dot-pot (Dotter - Sonnhammer and Durbin 1995) was also performed. The final annotations of each "build" were edited in Artemis (Carver et al. 2005). Validation of LTR-RT "build" structures was performed by comparative analysis with public Coffee BAC sequences, from the NCBI and the genome of *C. canephora* (Denoeud et al. 2014 - coffee-genome.org). Five BAC clones for *C. canephora* (EU164537, HQ696512, HQ696507, HQ696513 and HM635075) and 12 BAC

clones for *C*. *arabica* (GU123896, GU123899, GU123898, GU123894, GU123897, GU123895, HQ696508, HQ696510, HQ696509, HQ696511, HQ834787 and HQ832564) were downloaded from GenBank, accounting for a total of 3,023 Mb. BLASTN searches (E-value < $1e^{-150}$) against public Expressed Sequenced Tags (ESTs) databases from *C*. *canephora* and *C*. *arabica* were used to evaluate the transcription of the builds.

Estimation of the *Copia25* copy number using 454 sequencing survey

BLASTN searches were carried out with the full-length *Copia25* sequence (from BAC HQ696507) as query. Reads with more than 90% of nucleotide identity with *Copia25* over a minimum of 80% of the read lengths were considered as potential fragments of the element. Cumulative lengths of aligned reads to *Copia25* were used to extrapolate the contribution of the element to each genome size investigated.

Identification of *Copia25* in plant genomes

The sequence trimmed from AAARF was blasted against the *C*. *canephora* genome, as well as against 40 angiosperm and one non-angiosperm genome sequences available in the public databases of NCBI, Phytozome and Gramene (Table S1). BLASTN was used to search for the complete nucleotide sequence or the coding region of *Copia25* in the genomes. The retrieved sequences were analyzed using LTRharvest (Ellinghaus et al. 2008) in order to recover only the sequences with a structure similar to retrotransposons. These sequences were compared to the amino acid sequence of the *Copia25* reverse transcriptase (RT) using TBLASTN and against the *Ty1-Copia* retrotransposon databases of plants (Repbase http://www.girinst.org) resulting in 98 sequences from 34 species (Table S2).

Molecular analysis

The DNA of 24 Rubiaceae species (Table S3, Fig. S1) was extracted by using DNeasy Plant mini-kit (QIAGEN). The DNA of the *Musa* species was donated by Dr. A. D'hont (CIRAD, France). Primers were designed on intact RT region of *C. canephora Copia25* genomic sequences using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) (*Forward*: 5' GGG GTT GAA GAT GCA AGG TA 3'; *Reverse*: 5' AGC TGC TCC CAA ATC TTT CA 3'). For the reaction, 0.625 unit of Taq polymerase (Invitrogen), 20 ng genomic DNA, 1 mM of $MgCl_2$, 1 X buffer, 0.08 mM of dNTPs and 0.4 mM of each primer were used for a final volume of 25 µL. PCR conditions were as follows: initial denaturation (94 °C, 120 s); followed by 40 cycles of denaturation (94 °C, 30 s), annealing (55 °C, 30 s) and extension (72 °C, 180 s). Each PCR product was analyzed by gel electrophoresis on 1.2% agarose gel, purified (DNA GFX DNA & Gel Band, GE) and cloned (TOPO XL Cloning kit, Invitrogen) according to the manufacturer specifications. The plasmids extracted were sequenced using the specific primers. The *Copia25* sequences were registered under the GenBank Accession Numbers KM439056 to KM439101. For the reverse transcription polymerase chain reaction (RT-PCR) 1 µg of the total RNA from leaves of *C. canephora, C. eugenioides* and *C. arabica* was treated with RQ1 RNase-Free DNase (Promega) and reverse-transcribed using ImProm-II™ Reverse Transcription System (Promega). The synthesized cDNA served as templates for RT-PCR. DNA contamination was checked using the primers of the gene sucrose synthase (SUS10/SUS11 - Marraccini et al. 2011). RT-PCR was performed using the same specific primers according to the protocol described as before, with 50 ng of cDNA.

Evolutionary Analyses

Phylogenetic analyses were performed with MEGA 5.2 (Kumar et al. 2008) on sequence datasets aligned with the MAFFT program. Each phylogeny was reconstructed using the best model using Find Best DNA/Protein Model (Maximum Likelihood) in Mega 6 (Tamura et al. 2013), with 1000 replicates; the bootstrap consensus tree inferred is taken to represent the evolutionary history of the taxa analyzed. All positions containing gaps and missing data were eliminated. As rates of synonymous substitution are not available for Rubiaceae (genes or TEs), and because LTR sequences (non-coding regions) and those from the RT domain (coding region) may evolve differently, two rates, estimated for grasses and palms, were used. The age of insertion of *Copia25* within *C. canephora* genome was estimated using the molecular clock equation, as previously described (Moisy et al. 2014; SanMiguel et al. 1998; Wicker and Keller 2007), where *k* was the Kimura 2-parameter distance between both LTRs of the same copy, and *r* is 1.3 x 10^{-8} base substitutions per site per year (Ma and Bennetzen 2004). The Kimura 2-parameter method of distance estimation of non-coding nucleotide sequences was used for LTR distance estimation (SanMiguel et al. 1996). However, gene conversion between LTR of the same element could be a source of errors in estimating insertion time. This putative error is not taken into account in our analysis since conversion of LTR remains poorly understood in plant genomes. The age of the ancestor of the *Copia25* sequences was also estimated using the molecular clock equation, using *Ks* (number of synonymous substitutions per synonymous site) and the rate of synonymous substitutions as 6.5 x 10^{-9} base substitutions per site per year (Gaut et al. 1996) for the RT domain (Vitte et al. 2007).

In order to investigate whether *Copia25* was under selective pressure a codon substitution model was used to estimate ω (Ka/Ks). The ω ratio measures the direction and the magnitude of selection on amino acid changes, with values of $\omega < 1$, $= 1$, and > 1 indicating negative

purifying selection, neutral evolution, and positive selection, respectively. To estimate ω two approaches were used: (i) the Ka/Ks pairwise ratio for species with the full-length polyprotein sequence available (coffee, potato, tobacco and banana); and (ii) likelihood ratio tests (LRTs) for a simplified phylogeny (Fig. S2) containing species representatives of each of the Rubiaceae tribes and potato, tobacco and banana, using 315 nt of the RT domain. Premature stop codons were removed from the sequences for both analyses. For the pairwise Ka/Ks, the reference sequences of the *Copia25* Subfamilies 1 and 2 (chr7_16264485-16269785 and chr8_8081742-8086630 respectively) were compared with their homologous sequences in potato, tobacco and banana. *Ka* and *Ks* were obtained using DnaSP v5 (Librado and Rozas 2009). Selective pressure acting on COSII (conserved orthologs group) genes of potato, banana and coffee (Wu et al. 2006) was also investigated. The COSII sequences in potato and *C. canephora* are available on the Sol Genomics Network website (http://solgenomics.net). 515 COSII accessions present in single copy in potato and coffee were blasted (BLASTn) against the *Musa acuminata* CDSs (D'Hont et al. 2012 - http://banana-genome.cirad.fr/) in order to obtain the *Musa* COSII sequences. Seven COSII sequences showing the highest sequence identity were used to calculate the Ka/Ks ratio and nucleotide identity (Table S4). The second approach used different ω ratio parameters for different branches on the phylogeny (Anisimova and Ziheng 2007; Yang and Nielsen 1998). To estimate the log likelihood values (LRT), a one-ratio model was used. This model assumes the same ω free or fixed (ω = 1) parameter for the entire tree, Model I and Model II, respectively. A two-ratio model was used to estimate the LRTs for specific clades on the phylogeny, since we assumed that the sequence group of interest (separately for *Ixora*, Model III = ω free, and Model IV = ω fixed; and, for *Musa*, Model V = ω free, and Model VI = ω fixed) has a different ω_F from that of the ω_B background. For the pairs of models (I *vs* II, III *vs* IV, V *vs* VI)), the log

likelihood values were compared in a hypothesis test (X^2). These analyses were implemented using the codeml program in the PAML package (Yang 1997).

Results

Assembly of repeated sequences with BAC-end Sanger sequences and 454 random reads from *C. canephora*

Sanger and 454 sequences from *C. canephora* (accession HD200-94) were used to identify and characterize the TEs. Two bacterial artificial chromosome (BAC) libraries were recently constructed from the same plant and a total of 134,827 Sanger sequences (mean size 683 bp) were generated from BAC-end sequences (BES) and released (Dereeper et al. 2013). In addition, 106,459 random 454 Roche reads (mean size 423 bp) were also generated from the same plant (Table S5).

In all, Sanger and 454 sequences represent 137,104,866 bp (241,286 sequences), giving an estimated coverage of 19.5% of the *C. canephora* genome (710 Mb). They were used together to assemble repeated sequences using the Assisted Automated Assembler of Repeat Families Algorithm (AAARF, DeBarry et al. 2008). A total of 1,306 "builds" (also called contigs) were generated with a length ranging from 135 to 24,745 bp, and a mean length of 1,306 bp. Most of them (45%) have a length comprised between 0.5 and 1 kb. In total, 317 builds showed similarities with TE proteins available in public databases after translating the assembled sequences. Fifty-two of them, showing sizes larger than 3 kb, were selected for the subsequent analysis. Forty-nine out of 52 showed strong similarity to LTR-RT proteins (Table S6 and Table S7). Over the 49 contigs, 12 elements were removed due to non-

canonical (complex) structure, suggesting incorrect assembly, and in a significant number of builds, manual corrections were made (Table S6, 10 builds labeled with ‡), following the same procedure as described in De Barry et al. (2008). The 37 remaining builds with canonical TE structures showed exclusively similarities with LTR-RT proteins, suggesting that it may represent the main abundant transposable element family in the *C*. *canephora* genome (Table 1). These 37 potential retrotransposon builds, were manually annotated, and incomplete structures of all them were found (Fig. S3). According to the structural annotation, the were classified as "LTR-I-LTR" when the internal region and both complete or partial LTRs were present; as, "I" if only an internal region was present, as "LTR-I" with complete or partial 5' LTR with an internal region, and, "I-LTR" with an internal region and complete or partial 3' LTR (Table S6).

The 37 LTR-RT builds were used as query for similarity search (BLASTn) for complete or partial copies present in the available *Coffea* BAC clones sequences (Table S6). Ten LTR-RT builds showed high levels of nucleotide conservation with nine *C*. *canephora* (4) and *C*. *arabica* (5) BAC sequences (BLAST E-value cutoff: $10e^{-100}$; Table S6). Moreover, some builds showed similarities with *Coffea* transcriptomic sequences. Indeed, 15 and four LTR-RT builds were found in *C*. *canephora* and *C*. *arabica* ESTs, respectively (Table S7).

Characterization of *Copia25*, a *Ty1-Copia* LTR retrotransposon in Coffee trees

Among the retrotransposons identified in *C*. *canephora* sequences (accession HD200-94), the sequence of one *Ty1-Copia* element, hereafter named *Copia25*, showed high BLASTN scores across various distantly related plant genomes, suggesting that *Copia25* has a singular evolutionary history. *Copia25* also showed an overall structure similarity to *Rider* (EU195798), an active retrotransposon found conserved between distant dicot species (Cheng

et al. 2009; Jiang et al. 2012), as indicated by dot-plot alignment (not shown). The *Copia25* reassembled contig was blasted (BLASTN $10e^{-100}$, Table S6) against *C. arabica* and *C. canephora* BAC sequences. It was found in *C. canephora* but with an uncommon arrangement, which appears to be a tandem of two elements sharing one LTR sequence in the median of the structure (accession HQ696507). In *C. arabica*, in turn, a complete sequence of 5,382 bp was found. This sequence is flanked by two perfect 5-bp TSDs (5'-GGAAC-3'), and its two LTRs are both 530 bp long and show high sequence identity (99.2%) (accession HQ832564 - Fig. S4). This copy is localized on a homologous region to *C. canephora*, most probably the *C. canephora* sub-genome within *C. arabica*, but it is absent in the syntenic region in both 126 (Moschetto et al. 1996; Yu et al. 2011) and HD200-94 *C. canephora* genotypes (Denoeud et al. 2014).

A search was also made for the *Copia25* contig (using Censor) in the *C. canephora* genome (Denoeud et al. 2014) and 72 full-length copies were identified. All of them showed premature stop codons in the *pol* coding region, indicating that none of them is potentially functional. Nonetheless, similarity searches showed high sequence identity between *Copia25* and Expressed Sequence Tags (98 and 99% of nucleotide identity with DV679393 and GT681881, respectively). In addition, the *Copia25* RT regions were successfully amplified by RT-PCR on RNA extracted from *C. canephora*, *C. arabica* and *C. eugenioides* leaves (Fig. S5).

Full-length *Copia25* copies exist throughout the *C. canephora* genome mainly in gene-poor and LTR-RTs rich areas. The majority of them are located in the non-anchored set of scaffolds (pseudo-chromosome "0") (Fig. 1a; Table S8). The sharing of structural characteristics among group of sequences of a TE family might indicate the occurrence of subfamilies. In such cases, the different groups have different most recent ancestral copy – i.e. different mother (or master) copy –, which independently originated copies. A Maximum

Likelihood with the distance corrected by General Time Reversible model and 1000 replicates phylogenetic tree was produced using the *pol* (2,640 nt) nucleotide sequence of the 72 full-length *Copia25* copies. Based on the tree topology, two clusters were segregated (Fig. 1b). Following Wicker's parameters (Wicker et al. 2007) segregating criterion they are hereafter considered as subfamilies, one harboring 44 copies (Subfamily 1) and the other 28 (Subfamily 2). Only one copy did not group with either of the two clusters; this copy was discarded from further analyses. In each subfamily, the sequence with the perfect structure (based on the best conservation of both LTRs and the presence of an intact or few stop codons in the ORF coding for the polyprotein) was chosen as a reference sequence for the subfamily (Subfamily 1: chr7_16264485-16269785; Subfamily 2: chr8_8081742-8086630). These two sequences are 87.8% identical, and have 9.8% of InDels. The differences between them are mainly concentrated in the LTR region, where the identity is only 71%, and InDels reach 15%, resulting in only 59% of overlap. Such difference results in poor LTR alignment of the 72 copies. Additionally, Subfamily 2 presents a 208 bp deletion in the UTL 5' (Untranslated Leader) region. The corrected distances (Tamura-3 parameters) within each subfamily are 0.123 and 0.138 respectively, for Subfamily 1 and 2, and 0.222 between subfamilies (overall mean of 0.174). The divergence between the two LTRs of each copy was calculated and an insertion time was inferred. Subfamily 1 showed a mean time of insertion of 2.97 ± 0.204 Mya (minimum: 0.5, maximum: 5.2 Mya) and Subfamily 2 showed a mean time of insertion of 4.53 ± 0.399 Mya (minimum: 1.3, maximum 10.1 Mya) (Fig. 2, Table S8).

Presence of *Copia25* in the Rubiaceae family

In order to investigate the evolution of *Copia25*, sequence similarity searches and PCR amplifications were used to search for its presence in the *Coffea* genus and in other Rubiaceae

species. First, 11 genotypes representing 10 *Coffea* species (including ex-*Psilanthus*) and *Craterispermum sp*. *Novo kribi* were surveyed using high-throughput 454 Roche sequencing. The number of bases produced for each species and the estimated genome coverage according the genome sizes are shown in the Table 2. The 454 sequences were used to survey the presence of highly conserved *Copia25* sequences, using as criteria: 90% minimal nucleotide identity over 80% of the sequence length. The number of *Copia25* conserved sequences found for each species and their respective cumulative length according to the genome size are available in the Table 2. Sequences fitting these criteria were present in all *Coffea* genomes studied here, but not in *Craterispermum*. The cumulative length of *Copia25* reads was estimated to range from 186 to 1,513 kb of estimated cumulative sequences in diploid species and 842 kb in the allotetraploid *C*. *arabica* (Table 2).

The presence of *Copia25* was also investigated by PCR amplification and sequencing of the product in 13 *Coffea* and 11 other Rubiaceae species (Table S3, Fig. S1). The *Copia25* RT region was amplified and sequenced in 13 *Coffea* species, three from West Africa (*C*. *stenophylla*, *C*. *humilis* and *C*. *ebracteolatus*), one from West/Central Africa (*C*. *canephora*), three from East Africa (*C*. *costatifructa*, *C*. *pseudozanguebariae* and *C*. *eugenioides*), one from Northeast Africa (*C*. *arabica*), and five from Indian Ocean Islands (*C*. *millotii – ex-dolichophylla –*, *C*. *perrieri*, *C*. *resinosa*, *C*. *tetragona* and *C*. *vianneyi*) (Chevalier 1946; Maurin et al. 2007). The same region was also amplified and sequenced in 11 other Rubiaceae species: *Bertiera iturensis*, *Tricalysia congesta*, *Oxyanthus formosus*, *Ixora* sp., *I*. *coccínea*, *I*. *finlaysoniana*, *I*. *foliicalyx*, *Polysphaeria parvifolia*, *Coptosperma* sp., *Pyrostria* sp., and *Craterispermum schwenfurthii*. The final dataset contains 319 nucleotides, and the nucleotide identity varied from 62% to 100% among different sequences comparisons (Table S9).

Copia25 distribution among monocots and dicots

Besides the Rubiaceae species, similar *Copia25* sequences were sought among the 40 available plant sequences representing the angiosperm clades, and one non-angiosperm species using BLASTN. Similar *Copia25* sequences were found in 34 species but not in the remaining eight ones, as follows: *Arabidopsis lyrata*, *Carica papaya*, *Cucumis sativus*, *Fragaria vesca*, *Linum usitatissimum*, *Selaginella moellendorffii*, *Phoenix dactylifera* and *Zea mays* (Table S1).

In the 34 genomes where sequences similar to *Copia25* were found, these latter were extracted for further phylogenetic analysis. Using a fragment of 750 bp from the RT region, a phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter and 1000 replicates in order to investigate the relationships among the *Copia25* sequences (Fig. 3, Fig. S6 and Tables S10 and S11). One well-supported (95% bootstrap value) phylogenetic clade was found to include *C. canephora Copia25* and sequences belonging to four dicotyledonous species: *Nicotiana benthamiana*, *N. tabacum*, *S. tuberosum* (Solanaceae) and *Ricinus communis* (Euphorbiaceae), and more surprisingly, three monocotyledonous species, *Musa accuminata and M. balbisiana* (Musaceae), and, in a basal position, *Eleais guineensis* (Arecaceae). These sequences were considered homologous to *Copia25* because they share over 80% sequence identity over 80% of their length in the reverse transcriptase domain (Wicker et al. 2007), except for *R. communis* and *E. guinensis*. Since these two species cluster within the clade and share, with *Copia25*, over 70% of identity they were considered to belong to the same family.

Besides the *Copia25* clade, additional *Ty1-Copia* sequences related to it, clustered in strongly-supported clades composed of species of the same family, which supports a hypothesis of vertical inheritance (Fig. 3). It is the case of the elements found in the

monocotyledonous family of Poaceae where all of them cluster in a clade with a 94% bootstrap value. A similar occurrence was found in the Malvaceae species (100%) and in Fabaceae (98%) species, but it is also weakly supported among Brassicaceae (79%). The exceptions in this context are the particular strongly-supported relationships between *Medicago truncatula* (Fabaceae) and *Mimulus guttatus* (Phrymaceae) (94%), among *Populus trichocarpa* (Salicaceae), *Gossypium hirsutum* (Malvaceae) and *Malus domestica* (Rosaceae) (100%), and finally between *Solanum lycopersicum* (Solanaceae) and *Arabidopsis thaliana* (Brassicaceae) (100%).

The reconstructed phylogeny using only sequences recovered from public databases (Fig. 3) did not show a clear relationship between the sequences from coffee tree and those from other species in the clade. In an effort to better understand the relationships of *Copia25* among the species present in the *Copia25* clade, we reconstructed a new Maximum Likelihood phylogeny (with the distance corrected by Tamura 3-parameter and 1000 replicates), adding RT sequences obtained from several Rubiaceae species and three Musaceae species (*M*. *accuminata*, *M*. *balbisiana* and *M*. *boman*) (Fig. 4 and Fig. S7). As shown in Fig. 4, the unrooted phylogenetic tree revealed that *Copia25-Musa* is nested into the Rubiaceae species as shown by a closer well-supported relationship (bootstrap value 92%) between *Copia25-Musa* and *Copia25-Ixora* and between *Craterispermum* sp. and all Rubiaceae and Musaceae species (bootstrap value 65%). Rubiaceae and Musaceae *Copia25* are clearly separated from Solanaceae by high bootstrap value (92) and a topology structure. This result suggested that Rubiaceae and Musaceae *Copia25* constitute a unique evolutionary lineage (Fig. 4).

To further confirm the close relationship between *Copia25-Coffea* and *Copia25-Musa*, we first aligned each *Copia25-Coffea* sequence (*Copia25 C*. *canephora* reference sequence of Subfamily 1 and 2) with the *Copia25-Musa* (*M*. *balbisiana* AC186755). The alignments

showed an overall nucleotide identity of 74.1% and 79.6% for Subfamily 1 and 2, respectively, and an overall amino acid sequence identity rate of 81.7% (similarity: 79.8%) with Subfamily 1, and 81.60% (similarity: 80.1%) with Subfamily 2 (Fig. 5a). Their LTRs were also extracted and aligned, showing a high identity rate (53.9% between *Musa* and the reference sequence of Subfamily 1; and 59.4% with the Subfamily 2 reference sequence) (Fig. 5b). This level of identity is indeed quite significant for non-coding regions and considering the species divergence, i.e. about 150 Mya (Chaw et al. 2004; Wikstrom et al. 2001). Homologous sequences to *Copia25-Musa* from the *M. balbisiana* genome (B genome) were also found in the sequenced *M. acuminata* genome (A genome; D'Hont et al. 2012). These homologous sequences show high sequence identity (e.g. Chr9: 16119963-16124880; 91.1% of identity) between the two banana genomes that diverged by about 4.6 Mya (Lescot et al. 2008).

Evolution of *Copia25* in monocots and dicots

To investigate the evolution of *Copia25* in detail, we used the nucleotide sequences of *Copia25* from *M. balbisiana*, *C. canephora*, *S. tuberosum* and *N. benthamiana* for pairwise sequence comparisons. The results summarized in Supplementary Table S12 show higher identity between the *Copia25* of coffee and banana than between all the other species. We compared the identity of *Copia25* with the identities of seven COSII sequences showing the highest sequence identity between banana and coffee. These genes share an average of 74.7% of identity between banana and coffee, while the coding region of *Copia25* shows 85%. For the *Copia25* polyprotein and these seven COSII genes, we performed a pairwise Ka/Ks (non-synonymous per synonymous substitution ratio) analysis by comparison of banana, potato, tobacco and coffee sequences. Both COSII and *Copia25* were under purifying selection,

however they were found more relaxed in *Copia25* (minimum: 0.233, maximum: 0.287) than in COSII (minimum: 0.038, maximum: 0.215) sequences.

The LRT results reinforce the proposition of the purifying selection acting on the *Copia25* sequences (Table 3). The log likelihood values using a one-ratio model (Model I: ω free, and Model II: ω fixed) for the entire phylogenetic tree (Fig S2) were significantly lower than the neutral expectation, indicating purifying selection (0.191, $2\Delta\ell = 239.308$, $p < 0.01$). The LRTs of the *Ixora* and *Musa* clades were estimated separately. For these, a two-ratio model was applied, since we assumed that the sequence group of interest has a different ω_F from that of the ω_B background (Model III: ω free, and Model IV: ω fixed, for *Ixora* clade; and Model V: ω free, and Model VI: ω fixed, for the *Musa* clade). Purifying selection was also detected for *Ixora* clade (0.127, $2\Delta\ell = 33.568$, $p < 0.01$), while for the *Musa* clade the ω value did not differ from neutral evolution (Table 3). The negative selective pressure would explain the narrow relationship between the coffee and banana sequences. However, the negative selection for *Copia25* and COS, and the neutrality for *Copia25* in *Musa* clade indicate that this alone does not explain their clustering in the phylogeny.

The divergence time of two sequences harbored by two species from their common ancestral sequence was estimated by using both COSII and *Copia25*. The estimated divergence time using *Copia25* sequences for *Musa* and *Coffea* is much lower than for COSII sequences. While the latter ones ranged from 94.5 to 181.8 Mya, when using *Copia25* the time was 35.5 and 31.7 Mya. Indeed, the estimated divergence time using the *Copia25* from banana and the Solanaceae species is similar to that found for coffee, tobacco and potato. The high similarity and the *Ks* values for the comparisons between coffee and banana with the other Solanaceae species indicate that the *Copia25* sequence could be a recent guest in banana species genome.

Discussion

Copia25 in the Rubiaceae family

In this study, we identified an expressed *Ty1-Copia* in the *C. canephora* genome, *Copia25*, and analyzed it under various aspects, providing a broad insight into its evolution. *Copia25* was found distributed in several species of the *Coffea* genus from Africa, the Indian Ocean Islands and Indonesia. The occurrence of *Copia25* in these species denotes that it could be present in the ancestor of this phylogenetic group and has been inherited by the derived lineages. Our proposition of its presence in the *Coffea* lineage ancestor is reinforced by the occurrence of *Copia25* in at least two of the three subfamilies of the Rubiaceae family, Rubioideae (*Craterispermum schwenfurthii*) and Ixoroideae (*Coffea* spp., *Ixora* spp., *Bertiera iturensis*, *Coptosperma* sp., *Oxyanthus formosus*, *Polysphaeria parvifolia*, *Pyrostria* sp., *Tricalysia cloneongesta*), also suggesting its ancient evolutionary history in Rubiaceae. Altogether these data suggest the presence of *Copia25* in both of the Rubiaceae subfamilies preceding their ancient divergence.

High sequence identity of *Copia25* of over 150 My of plant genome evolution

Our similarity searches and molecular biology approaches revealed patchy conservation of *Copia25*. They show high sequence identity between a monocot genus of the Musaceae family and two different dicotyledonous families in Asteridae: the Rubiaceae and Solanaceae families. While monocot and dicot species diverged about 150 Mya, the Asteridae and Rosidae lineages diverged ~114 Mya. More recently, Rubiaceae and Solanaceae diverged

from their common ancestor about 83 Mya (Chaw et al. 2004; Wikstrom et al. 2001). This discontinuous and incongruent distribution in dicots and monocots highlights a complex evolutionary history of *Copia25* in plants that could be traced back to the origin of angiosperms.

Copia25-Coffea clusters in a strongly supported clade (100% bootstrap value) with homologous sequences from three Solanaceae species, *S. tuberosum*, *N. tabacum* and *N. benthamiana*, and Musaceae species, *Musa* spp.. However, the nucleotide identity between *Copia25-Coffea* and *Copia25-Musa* is higher than the one observed between *Coffea* and potato and tobacco, and even in the comparison between *Musa* and Solanaceae (*S. tuberosum*: 77.4%; *N. benthamiana*: 77.2%). When the seven orthologous (COSII) genes showing the highest sequence conservation are compared among the same species, the nucleotide identity between *C. canephora* and *M. balbisiana* ranged from 67.8% to 80.2%, less than the *Copia25* polyprotein identity for the same species comparison (Subfamily 1: 84.5% and Subfamily 2: 85.5%). Equivalent identities were also found in the *gag* region. Such outstandingly high conservation raises questions about the molecular mechanisms, which are at its origin.

Conservation of TEs between distantly related genera could be the result of different and non-exclusive processes (Capy et al. 1994; Cummings 1994; Schaack et al. 2010; Wallau et al. 2011) such as: (i) domestication, (ii) conservation of functional sites, (iii) similarity of evolutionary rates, (iv) purifying selection and (v) horizontal transfer. The first two scenarios cannot explain the conservation of *Copia25* across genera, since only portions of the TE are generally domesticated and because the mechanisms of conserving functional sites exclusively involve coding regions. High sequence identity was found for the full-length sequences of *Copia25,* including non-coding LTR regions. Similar TE evolutionary rate in distinct species is an attractive hypothesis to explain the conservation observed in *Copia25*. However, the TE evolutionary rate depends on multiple parameters such as the specific TE

activity and the efficiency of TE host control mechanisms. Such a scenario remains unlikely since these evolutionary mechanisms should be identical in several distantly-related species. The fourth process, a purifying selection, would explain the high identity of a given TE between distantly related species. The Ka/Ks ratio estimated for pairwise comparisons of *Copia25* between *Musa* and *Coffea* sequences is low (< 0.3), denoting purifying selection and explaining the conservation and the activity (at least until very recently) of this particular element. However, the *Ks* values between *Coffea* and Solanaceae, *Musa* and Solanaceae and *Musa* and *Coffea* species are at least twice as low for *Copia25* as for COSII sequences. This observation suggests that other evolutionary processes besides purifying selection might be involved in *Copia25* conservation. Finally, HTs of TEs, an occurrence suggested but rarely confirmed in plants (Diao et al. 2006; El Baidouri et al. 2014; Fortune et al. 2008) may explain the strong conservation level in coding and non-coding regions, and the sparse distribution of TEs. However, HT scenarios first require ecological, chronological, and geographical distribution overlapping between the species involved in the potential transfer to be seriously considered. These requirements are not expected for *Musa* and *Coffea*, but a chronological and geographical distribution overlap might have existed for the *Musa* and *Ixora* species. The *Ixora* genus belongs to the Ixoroideae subfamily of the Rubiaceae family such as the *Coffea* genus, but both belong to different tribes, Ixoreae and Coffeeae (Fig. S1). The genus *Musa* evolved and diversified in tropical Asia (Liu et al. 2010), and the *Musa* lineage ancestor originated ~50 Mya (Christelova et al. 2011). Likewise, the *Ixora* genus originated in South-East Asia, in Borneo in particular (Lorence et al. 2007), and its ancestral lineage originated 30 to 50 Mya (Tosh et al. 2013). Therefore, the ancestors of *Musa* and *Ixora* could have shared the same period and geographical origin. The hypothesis of the HT of *Copia25* between the ancestors of *Ixora* and *Musa* is therefore supported by the chronological and geographical distribution of species. This hypothesis is also supported by the high global

sequence identity as well as by the *Ks* values, which are much lower for *Copia25* than for the COSII, suggesting that its presence is recent in the *Musa* genome. Furthermore, the phylogeny of *Copia25* RT including the *Musa* and Rubiaceae species sequences clearly indicates a strong relationship between *Copia25-Musa* and *Copia25-Ixora* (Fig. 4). This relationship does not result from similar selective pressure acting in both groups (as showed by LRT analyses, which exclude purifying selection as the process responsible for sequence similarity) and thus reinforces the proposition of HT. The putative period of *Copia25* transfer from *Ixora* to *Musa* can be estimated by the molecular clock equation using the RT sequences (375 nt; *Ks* ranged from 0.25 to 0.56). The estimated age range from 19 to 43 Mya is congruent with the period when the ancestors of both genera shared geographical distribution. This estimation must be considered with caution because of the short sequence used for establishing the time of divergence and because the molecular clock used is not calibrated for Rubiaceae. Our results thus suggest a potential and ancestral HT of *Copia25* from *Ixora* to *Musa* (Fig. S8).

With the facility for plants to inter-cross and given the autonomy of their germ line, plant genomes have a natural propensity to transfer genetic material. They also have a high content of LTR-RTs, elements whose cytoplasmic multiplication phase heightens the likelihood of being captured and exchanged among other species, thus favoring potential HT. Thanks to the fast-growing number of data sequences available, more studies are being conducted involving several species. Their results reveal scenarios of complex evolution, particularly those concerning TEs. Here, our detailed analyses of *Copia25* in angiosperms disclose the complexity of the evolutionary dynamics of this ancient element, involving several processes including sequence conservation, rapid turnover, stochastic losses and horizontal transfer. Additional information on the presence and the activity of *Copia25* in angiosperms is required to precisely identify the mechanism involved in such remarkable

conservation of a transposable element harbored by large and divergent groups of plant species.

Acknowledgments

This research was supported Agropolis Fondation through the "Investissement d'avenir" program (ANR-10-LABX-0001-01) under the reference ID 1002-009 and 1102-006, CAPES (Grants 01/2010 to CMAC and fellowship 9127-11-9 to ESD), Brazilian agencies FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo - Grant 2013/15070-4 to CMAC and fellowship 2011/18226-0 to ESD) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Grant 306493/2013-6 to CMAC) and French agency ANR (Agence Nationale de la Recherche; Genoplante ANR-08- GENM-022-001). Acknowledgments to Dr. A. D'Hont for providing *Musa* spp. DNA samples; Herman E. Taedoumg for providing *Craterispermum* samples; Dr. P. De Block for providing Rubiaceae samples; Dr. J-J. Rakotomalala for providing Mascarocoffea samples. Acknowledgements to Philippe Lashermes and the Coffee Genome Consortium for the availability of the *C*. *canephora* BAC-end sequences and draft genome.

Conflict of Interest The authors declare that they have no competing interests.

Electronic supplementary material

The paper contains supplementary material, File 1.

Figure legends

Fig. 1 Distribution and phylogenetic relationship of the copies of *Copia25* identified in the *C. canephora* genome. a Distribution of full-length copies (black lines) and fragmented copies of *Copia25* (red dashes) along the 11 *C. canephora* pseudo-molecules. The gene density along pseudo-molecules is represented in grey while the LTR retrotransposons are represented in red in a separate layer. Fragmented copies are defined as a minimum of 90% nucleotide conservation and 10 to 80% coverage of full-length copies. **b** Phylogeny reconstructed using the *pol* of the full-length copies of *Copia25*. The phylogeny was reconstructed using Neighbor joining, with the distance corrected by General Time Reversible model, and 1000 replicates. All positions containing gaps and missing data were eliminated. There were a total of 2,640 nucleotides in the final dataset. Only the bootstrap values over 70 are shown. Represented in blue are the sequences of Subfamily 1, and in red, Subfamily 2.

Fig. 2 Estimation of the insertion time distribution (in millions of years) of the 72 full-length *Copia25* (Subfamily 1 and 2) copies identified in the *C. canephora* genome. The insertion time was estimated using the Kimura 2-parameter between both LTRs of the same copy and the following molecular clock equation with $r = 1.3 \times 10^{-8}$ (Ma and Bennetzen 2004).

Fig. 3 Phylogeny of the RT domain from sequences similar to the *Copia25* elements in the 29 plant genomes analyzed. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the bootstrap consensus tree inferred is taken to represent the evolutionary history of the taxa analyzed. All positions containing gaps and missing data were eliminated. There were a total of 602 nucleotide sites in the final dataset; and a total of 98 nucleotide sequences. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.7864)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 10.1863% sites). The highlighted clade corresponds to the *Copia25* family; in blue, the monocot species in *Copia25* clade; the number in parentheses is the number of sequences collapsed in the tree. Species abbreviation: *S. tuberosum Solanum tuberosum* (potato), *N. tabacum: Nicotamia tabacum* (tobacco), *N. benthamiana: Nicotamia benthamiana*, *C. canephora*: *Coffea canephora* (coffee), *R. communis; Ricinus communis* (castor oil), *E. guineensis: Elaeis guineensis* (African oil palm), *S. italica: Setaria italica* (Foxtail millet), *S. bicolor: Sorghum bicolor* (sorghum), *O. sativa*: *Oryza sativa* (rice), *T. aestivum: Triticum aestivum* (wheat), *H. vulgare: Hordeum vulgare* (barley), *B. distachyon: Brachypodium distachyon, V. vinifera: Vitis vinifera (grape), Gossypium* (cotton), *A. trichopoda: Amborella trichopoda, G. max: Glycine max* (soybean), *P. vulgaris: Phaseolus vulgaris* (common bean), *C. cajan: Cajanus cajan* (pigeon pea), *L. japonicus: Lotus japonicus*, *M. truncatula: Medicago truncatula*, *E. grandis: Eucalyptus grandis*, *T. cacao: Theobroma cacao*, *F. ananasa: Fragaria x ananasa* (strawberry), *P. trichocarpa: Populus trichocarpa*, *G. hirsutum: Gossypium hirsutum*, *M. domestica: Malus domestica* (apple), *A. thaliana: Arabidopsis thaliana*, *S. lycopersicum: Solanum lycopersicum* (tomato), *M. guttatus: Mimulus guttatus*, *C. sinensis: Clementina sinensis*, *B. rapa: Brassica rapa*.

Fig. 4 Phylogenetic analysis of *Copia25* RT domain homologs. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the tree with the highest log likelihood (-4739.5265) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.1187)). The tree is drawn to scale, with branch lengths measured by number of substitutions per site. All positions containing gaps and missing data were eliminated. There were a total of 313 positions in the final dataset; and a total of 69 nucleotide sequences. Only the bootstrap values over 50% are shown. In green, the clade corresponding to the cluster between *Copia25 Musa* and *Ixora* sequences; in blue, the monocot species. The number of collapsed sequences is indicated in parentheses. Species abbreviation: *S. tuberosum Solanum tuberosum* (potato), *N. tabacum: Nicotamia tabacum* (tobacco), *N. benthamiana Nicotamia benthamiana*, *R. communis; Ricinus communis* (castor oil), *E. guineensis: Elaeis guineensis* (African oil palm) and *C.* means *Coffea*.

Fig. 5 Comparison between *Copia25* and *Copia25-Musa*. a/b Dot plot alignment between the full-length copy of *Copia25* (reference sequences, Subfamilies 1 **(a)** and 2 **(b)**) and the *Copia25-Musa* found in a genomic segment of the *Musa balbisiana* BAC clone (horizontal axis; AC186755 100804-105774). **c** Nucleotide alignment of 5' LTR of *Copia25* Subfamily 2 and *Copia25-Musa*.

Table 1 Summary of the AAARF assembly. Only contigs larger than 3 Kb (52 over 317) and with a correct assembly structure (37 over 52) were analyzed.

TE classification	Number of identified contigs (> 3Kb)	Number of contigs with EST similarity (E-value <$10e^{-100}$)
Class I LTR retrotransposons	37	26
Class I LTR retrotransposons, *Ty3-Gypsy*	28	22
Class I LTR retrotransposons, *Ty1-Copia*	9	4
Class II transposons	0	0
Total	37	26

Table 2 Estimation of the *Copia25* copy number in *Coffea* genomes using 454 sequencing survey. Only 454 reads with a minimum of 90% of nucleotide identity and over 80% of the read length were considered.

Species	Ploidy level	Estimated genome size (Mb)	#454 sequences	Produced bases (Mb)	Genome coverage %	# of *Copia25* reads	Cumulative length of aligned reads (Kb)	Estimated length in genomes (Kb)
C. canephora (HD94-200)	2x	710	106459	45.05	6.40	70	31,189	487,3
C. canephora (BUD15)	2x	710	149196	67.08	9,58	102	47,092	491,5
C. arabica	4x	1,240	122258	54.5	4.39	85	36,980	842,3
C. eugenioides	2x	645	101309	42.1	6.52	71	30,171	462,7
C. heterocalyx	2x	863	194300	60.51	2.25	42	13,732	610,3
C. racemosa	2x	506	88498	34.19	5.7	179	86,284	1513,7
C. pseudozanguebariae	2x	593	215117	91.7	15.4	68	28,669	186,1
C. humblotiana	2x	469	160479	67.99	14.49	102	45,373	313,3
C. tetragona	2x	513	160107	72.66	14.10	199	97,927	694,5
C. millotii	2x	682	163873	76.65	11.23	95	43,173	384,4
C. horsfieldiana	2x	593*	112793	46.25	7.8	72	29,593	379,3
Craterispermum sp. Novo Kribi	2x	748	49789	19.44	2.59	0	0	0

*: mean value estimates from other ex-*Psilanthus* accessions in absence of clear data for *C. horsfieldiana*.

Table 3 Likelihood ratio test for testing models of sequence evolution for Copia25 retrotransposons.

Model		Parameter	ℓ	2Δℓ	ω_B	ω_F	Conclusion
One-ratio	Model I	ω free	-2469.160	239.308**	0.191	-	Purifying selection in the *Copia25* tree
	Model II	ω = 1	-2588.814		-	-	
Two-ratio	Model III	ω free	-2468.462	33.568**	0.198	0.127	Purifying selection in the *Ixora Copia25* clade
	Model IV	ω = 1	-2485.246		0.198	1	
	Model V	ω free	-2463.734	2.526	0.169	0.552	Neutral evolution in the *Musa Copia25* clade
	Model VI	ω = 1	-2464.998		0.168	1	

Critical values of X^2, 1 df: *: 3.84; **: 6.63; $2\Delta\ell = 2\ (l_1-l_0)$

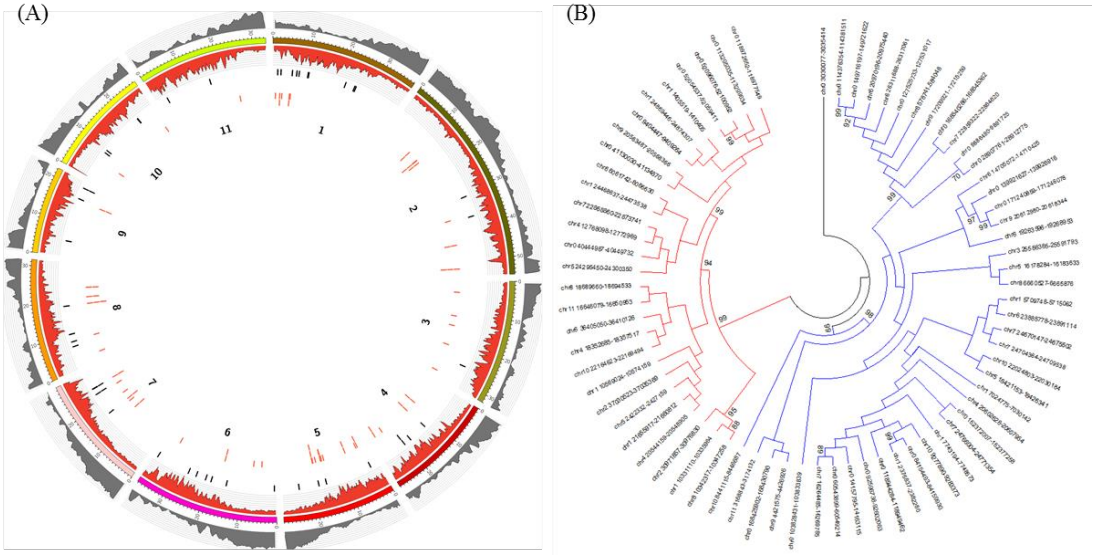


Fig. 1 Distribution and phylogenetic relationship of the copies of *Copia25* identified in the *C. canephora* genome. a Distribution of full-length copies (black lines) and fragmented copies of *Copia25* (red dashes) along the 11 *C. canephora* pseudo-molecules. The gene density along pseudo-molecules is represented in grey while the LTR retrotransposons are represented in red in a separate layer. Fragmented copies are defined as a minimum of 90% nucleotide conservation and 10 to 80% coverage of full-length copies. **b** Phylogeny reconstructed using the *pol* of the full-length copies of *Copia25*. The phylogeny was reconstructed using Neighbor joining, with the distance corrected by General Time Reversible model, and 1000 replicates. All positions containing gaps and missing data were eliminated. There were a total of 2,640 nucleotides in the final dataset. Only the bootstrap values over 70 are shown. Represented in blue are the sequences of Subfamily 1, and in red, Subfamily 2.

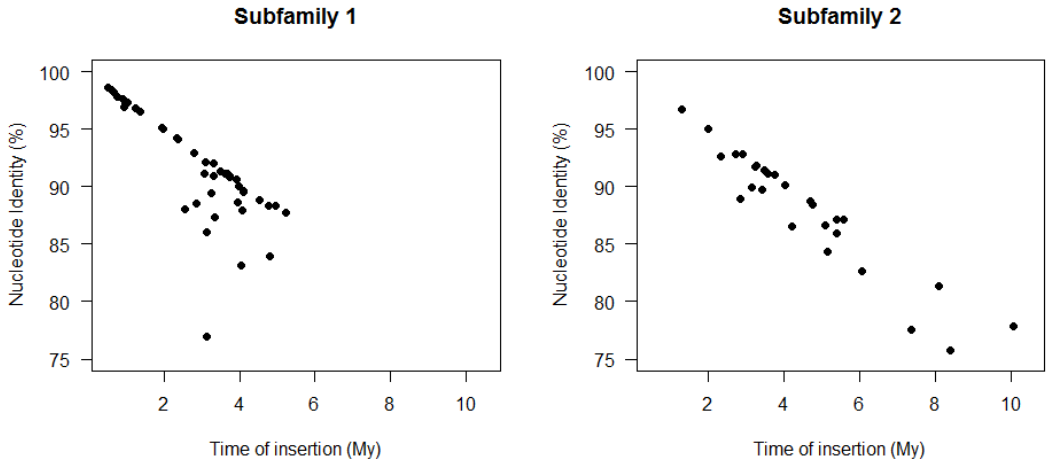


Fig. 2 Estimation of the insertion time distribution (in millions of years) of the 72 full-length *Copia25* (Subfamily 1 and 2) copies identified in the *C. canephora* genome. The insertion time was estimated using the Kimura 2-parameter between both LTRs of the same copy and the following molecular clock equation with $r = 1.3 \times 10^{-8}$ (Ma and Bennetzen 2004).

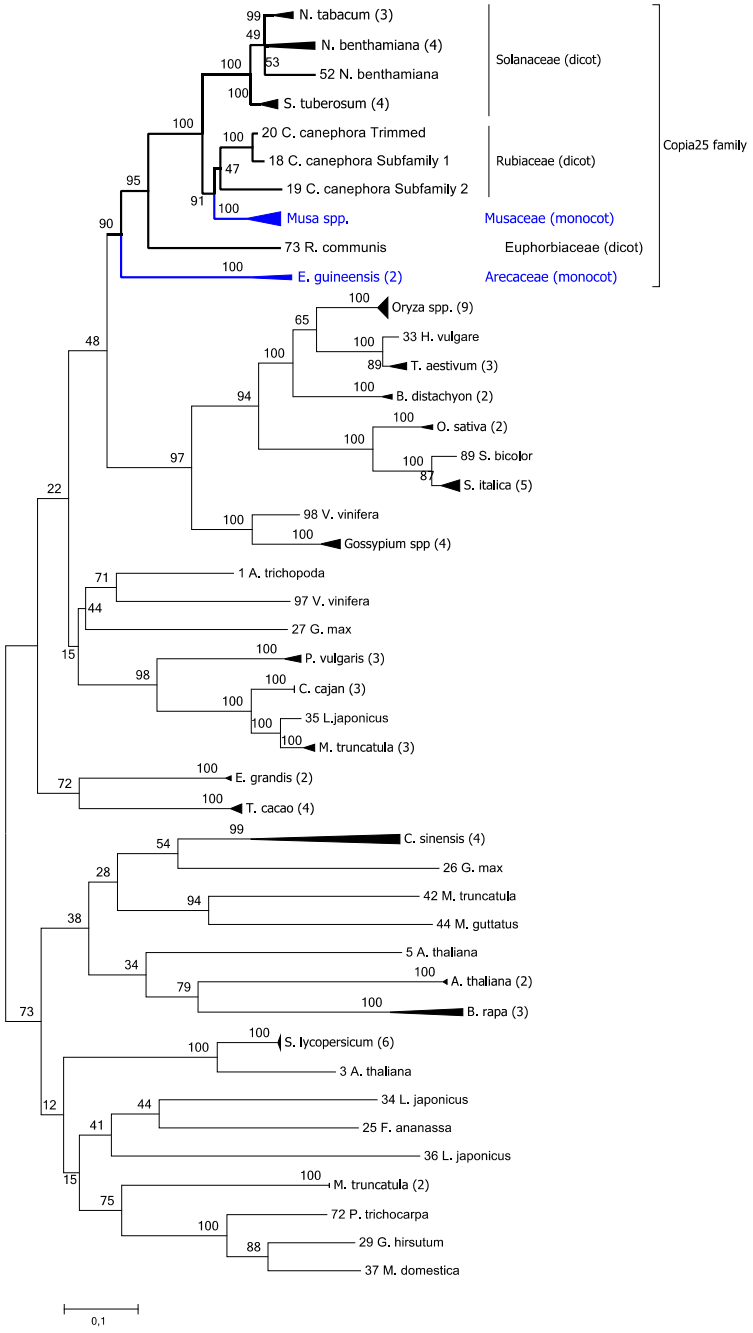


Fig. 3 Phylogeny of the RT domain from sequences similar to the *Copia25* elements in the 29 plant genomes analyzed. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the bootstrap consensus tree inferred is taken to represent the evolutionary history of the taxa analyzed. All positions containing gaps and missing data were eliminated. There were a total of 602 nucleotide sites in the final dataset; and a total of 98 nucleotide sequences. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.7864)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 10.1863% sites). The highlighted clade corresponds to the *Copia25* family; in blue, the monocot species in *Copia25* clade; the number in parentheses is the number of sequences collapsed in the

tree. Species abbreviation: *S. tuberosum Solanum tuberosum* (potato), *N. tabacum: Nicotamia tabacum* (tobacco), *N. benthamiana: Nicotamia benthamiana*, *C. canephora*: *Coffea canephora* (coffee), *R. communis; Ricinus communis* (castor oil), *E. guineensis: Elaeis guineensis* (African oil palm), *S. italica: Setaria italica* (Foxtail millet), *S. bicolor: Sorghum bicolor* (sorghum), *O. sativa*: *Oryza sativa* (rice), *T. aestivum: Triticum aestivum* (wheat), *H. vulgare: Hordeum vulgare* (barley), *B. distachyon: Brachypodium distachyon, V. vinifera: Vitis vinifera (grape), Gossypium* (cotton), *A. trichopoda: Amborella trichopoda*, *G. max: Glycine max* (soybean), *P. vulgaris: Phaseolus vulgaris* (common bean), *C. cajan: Cajanus cajan* (pigeon pea), *L. japonicus: Lotus japonicus*, *M. truncatula: Medicago truncatula*, *E. grandis: Eucalyptus grandis*, *T. cacao: Theobroma cacao*, *F. ananasa: Fragaria x ananasa* (strawberry), *P. trichocarpa: Populus trichocarpa*, *G. hirsutum: Gossypium hirsutum*, *M. domestica: Malus domestica* (apple), *A. thaliana: Arabidopsis thaliana*, *S. lycopersicum: Solanum lycopersicum* (tomato), *M. guttatus: Mimulus guttatus*, *C. sinensis: Clementina sinensis*, *B. rapa: Brassica rapa*.

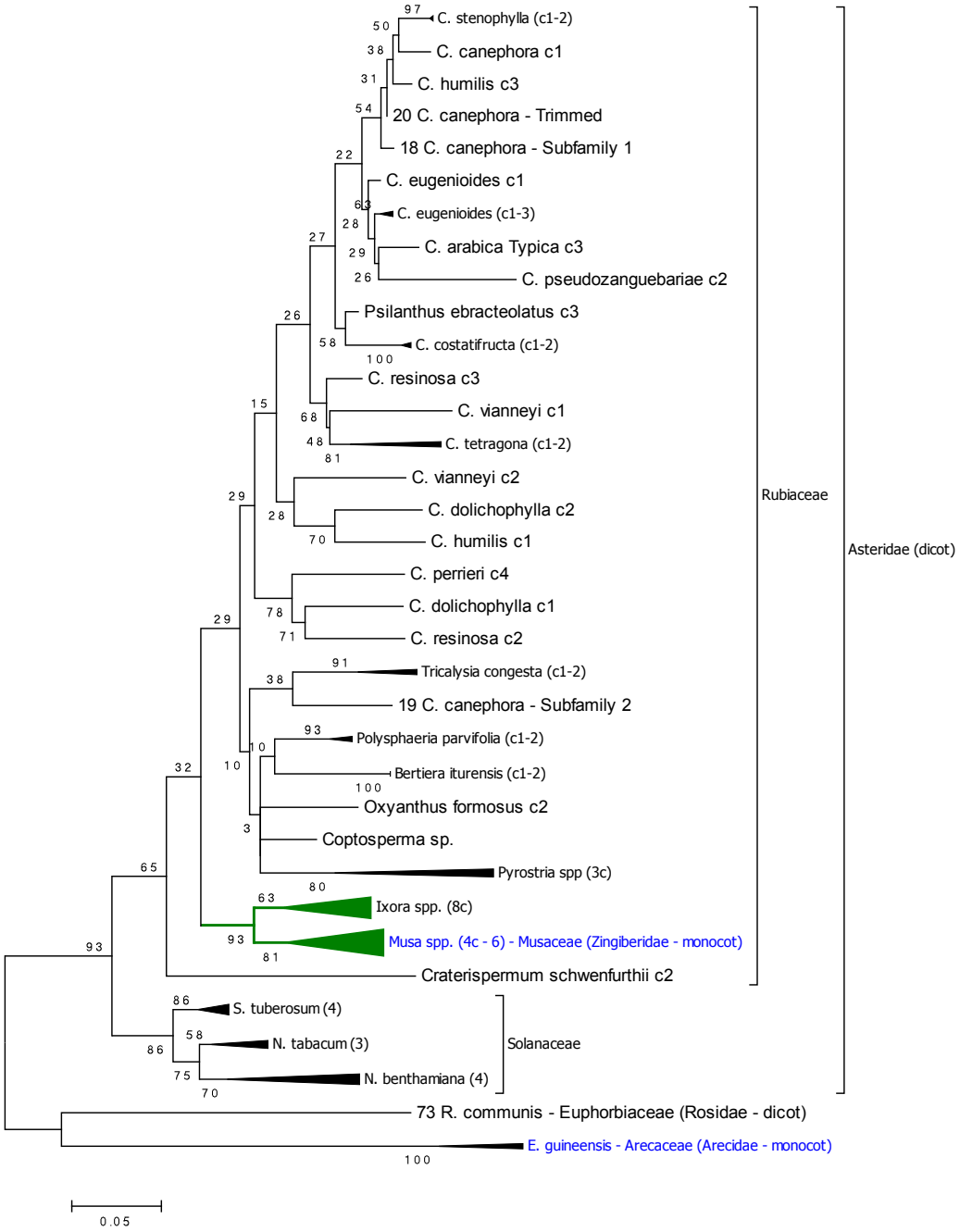


Fig. 4 Phylogenetic analysis of *Copia25* RT domain homologs. The phylogeny was reconstructed using Maximum Likelihood, with the distance corrected by Tamura 3-parameter, and 1000 replicates; the tree with the highest log likelihood (-4739.5265) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+G, parameter = 1.1187)). The tree is drawn to scale, with branch lengths measured by number of substitutions per site. All positions containing gaps and missing data were eliminated. There were a total of 313 positions in the final dataset; and a total of 69 nucleotide sequences. Only the bootstrap values over 50% are shown. In green, the clade corresponding to the cluster between *Copia25 Musa* and *Ixora* sequences; in blue, the monocot species. The number of collapsed sequences is

indicated in parentheses. Species abbreviation: *S*. *tuberosum Solanum tuberosum* (potato), *N*. *tabacum: Nicotamia tabacum* (tobacco), *N*. *benthamiana Nicotamia benthamiana*, *R*. *communis; Ricinus communis* (castor oil), *E*. *guineensis: Elaeis guineensis* (African oil palm) and *C*. means *Coffea*.

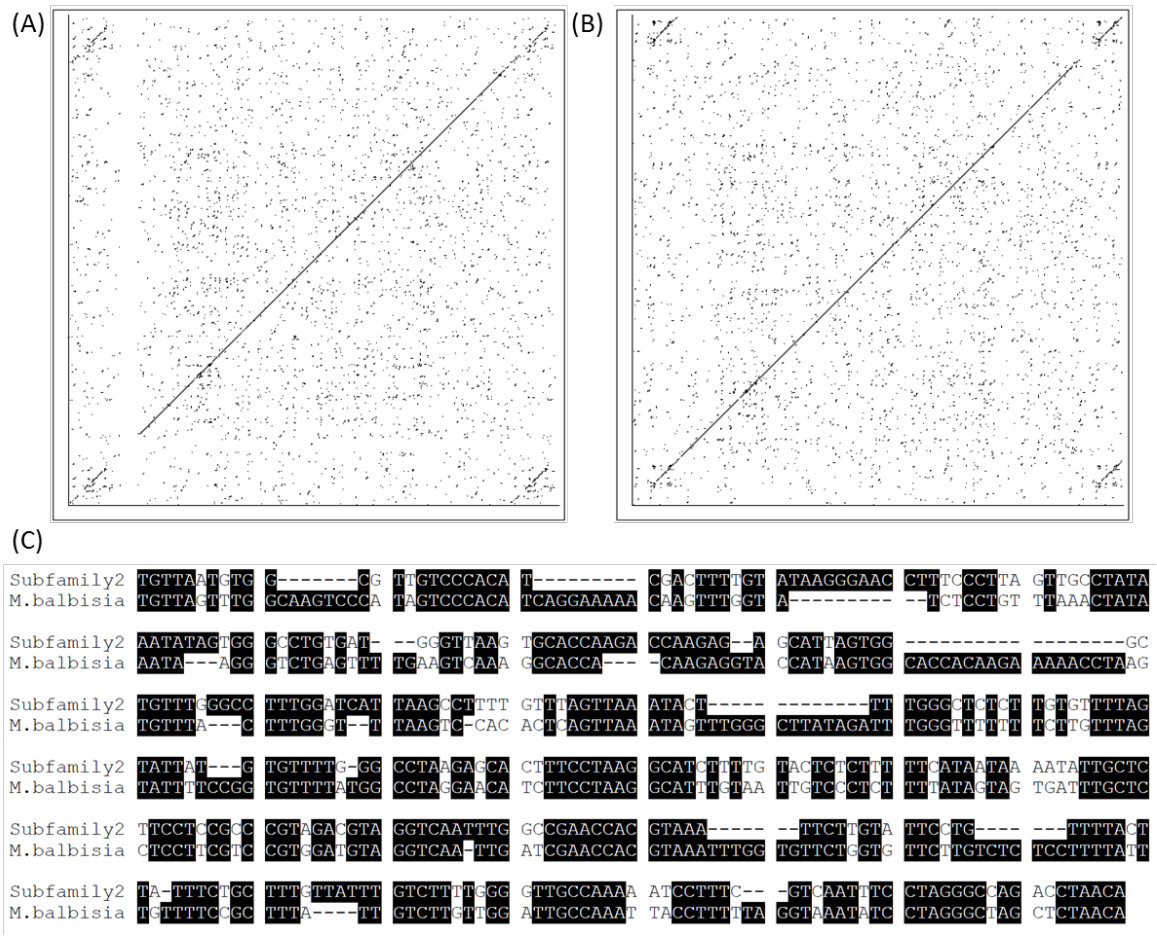


Fig. 5 Comparison between *Copia25* and *Copia25-Musa*. Dot plot alignment between the full-length copy of *Copia25* (reference sequences, **a** Subfamilies 1 and **b** 2) and the *Copia25-Musa* found in a genomic segment of the *Musa balbisiana* BAC clone (horizontal axis; AC186755 100804-105774). **c** Nucleotide alignment of 5' LTR of *Copia25* Subfamily 2 and *Copia25-Musa*.

References

Anisimova M, Ziheng Y (2007) Multiple Hypothesis Testing to Detect Lineages under Positive Selection that Affects Only a Few Sites. Mol Biol Evol 24:1219–1228

Capy P, Anxolabehere D, Langin T (1994) The strange phylogenies of transposable elements: are horizontal transfers the only explantation? Trends Genet 10:7-12

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21:3422-3423

Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58:424-441

Cheng X, Zhang D, Cheng Z, Keller B, Ling HQ (2009) A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. Genetics 181:1183-1193

Chevalier A (1946) Ecologie et distribution géographique des caféiers sauvages et cultivés. Rev Bot Appl Agric Trop 26:81-94

Christelova P, Valarik M, Hribova E, De Langhe E, Dolezel J (2011) A multi gene sequence-based phylogeny of the Musaceae (banana) family. BMC Evol Biol 11:103

Cummings MP (1994) Transmission patterns of eukaryotic transposable elements: arguments for and against horizontal transfer. Trends Ecol Evol 9:141-145

D'Hont A, Denoeud F, Aury J-M, et al (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488:213–217

Davis AP (2010) Six species of Psilanthus transferred to Coffea (Coffeeae, Rubiaceae). Phytotaxa 10:41-45

Davis AP (2011) Psilanthus mannii, the type species of Psilanthus, transferred to Coffea. Nordic Journal of Botany 29:471-472

de Carvalho MO, Loreto EL (2012) Methods for detection of horizontal transfer of transposable elements in complete genomes. Genetics and molecular biology 35:1078-1084

DeBarry JD, Liu R, Bennetzen JL (2008) Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm. BMC Bioinformatics 9:235

Denoeud F, Carretero-Paulet L, Dereeper A, et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–4

Dereeper A, Guyot R, Tranchant-Dubreuil C, et al (2013) BAC-end sequences analysis provides first insights into coffee (Coffea canephora P.) genome composition and evolution. Plant Mol Biol 83:177–89

Diao X, Freeling M, Lisch D (2006) Horizontal transfer of a plant transposon. PLoS Biol 4:e5

El Baidouri M, Carpentier M-CC, Cooke R, et al (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24:831–8

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

Fortune PM, Roulin A, Panaud O (2008) Horizontal transfer of transposable elements in plants. Commun Integr Biol 1:74-77

Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci U S A 93:10274-10279

Hamon P, Duroy P-OO, Dubreuil-Tranchant C, et al (2011) Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae). Mol Genet Genomics 285:447–60

Jiang N, Gao D, Xiao H, van der Knaap E (2009) Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. Plant J 60:181-193

Jiang N, Visa S, Wu S, van der Knaap E (2012) Rider Transposon Insertion and Phenotypic Change in Tomato. Topics in Current Genetics 24:297-312

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

Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9:299-306

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A (1999) Molecular characterisation and origin of the Coffea arabica L. genome. Mol Gen Genet 261:259-266

Lescot M, Piffanelli P, Ciampi A, et al (2008) Insights into the Musa genome: Syntenic relationships to rice and between Musa species. BMC Genomics 9:58.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452

Liu A, Kress W, Li D (2010) Phylogenetic analyses of the banana family (Musaceae) based on nuclear ribosomal (ITS) and chloroplast (trnL-F) evidence. Taxon 59:20-28

Llorens C, Futami R, Covelli L, et al. (2010) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70-74

Lorence D, Wagner W, Mouly A, Florence J (2007) Revision of Ixora (Rubiaceae) in the Marquesas Islands (French Polynesia). Botanical Journal of The Linnean Society 155:581–597

Ma J, Bennetzen J (2004) Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences of the United States of America 101:12404–12410

Marraccini P, Freire LP, Alves GS, et al (2011) RBCS1 expression in coffee: Coffea orthologs, Coffea arabica homeologs, and expression variability between genotypes and under drought stress. BMC Plant Biol 11:85

Maurin O, Davis AP, Chester M, Mvungi EF, Jaufeerally-Fakim Y, Fay MF (2007) Towards a Phylogeny for Coffea (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. Ann Bot (Lond) 100:1565-1583

Michael TP, Jackson S (2013) The First 50 Plant Genomes. The Plant Genome 6:1-7

Moisy C, Schulman AH, Kalendar R, et al (2014) The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. Theor Appl Genet 127:1223-35

Moschetto D, Montagnon C, Guyot B, Perriot JJ, Leroy T, Eskes A (1996) Studies on the effect of genotype on cup quality of Coffea canephora. Tropical Science 36:18-31

Roulin A, Piegu B, Wing RA, Panaud O (2008) Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza. Plant J 53:950-959

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43-45

SanMiguel P, Tikhonov A, Jin YK, et al. (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537-546

Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1-10

Tamura K, Stecher G, Peterson D, Filipski A (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Tosh J, Dessein S, Buerki S, et al (2013) Evolutionary history of the Afro-Madagascan Ixora species (Rubiaceae): species diversification and distribution of key morphological traits inferred from dated molecular phylogenetic trees. Annals of Botany 112:1723-1742

Vitte C, Panaud O, Quesneville H (2007) LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8:218

Wallau GL, Hua-Van A, Capy P, Loreto EL (2011) The evolutionary history of mariner-like elements in Neotropical drosophilids. Genetica 139:327-338

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17:1072-1081

Wicker T, Sabot F, Hua-Van A, et al (2007) A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8:973–982

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc Biol Sci 268:2211-2220

Wu F, Mueller LA, Crouzillat D, Petiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174:1407-1420

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319:1527-1530

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409-418

Yu Q, Guyot R, Kochko A de, et al (2011) Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). Plant J 67:305–17

Yuyama PM, Pereira LF, Santos TB dos, et al (2012) FISH using a gag-like fragment probe reveals a common Ty3-gypsy-like retrotransposon in genome of Coffea species. Genome 55:825–33

Table
Click here to download Table: Tables.docx

Table 1 Summary of the AAARF assembly. Only contigs larger than 3 Kb (52 over 317) and with a correct assembly structure (37 over 52) were analyzed.

TE classification	Number of identified contigs (> 3Kb)	Number of contigs with EST similarity (E-value <$10e^{-100}$)
Class I LTR retrotransposons	37	26
Class I LTR retrotransposons, *Ty3/Gypsy*	28	22
Class I LTR retrotransposons, *Ty1/Copia*	9	4
Class II transposons	0	0
Total	37	26

Table 2 Estimation of the *Copia25* copy number in *Coffea* genomes using 454 sequencing survey. Only 454 reads with a minimum of 90% of nucleotide identity and over 80% of the read length were considered.

Species	Ploidy level	Estimated genome size (Mb)	#454 sequences	Produced bases (Mb)	Genome coverage %	# of *Copia25* reads	Cumulative length of aligned reads (Kb)	Estimated length in genomes (Kb)
C. canephora (HD94-200)	2x	710	106459	45.05	6.40	70	31,189	487,3
C. canephora (BUD15)	2x	710	149196	67.08	9,58	102	47,092	491,5
C. arabica	4x	1,240	122258	54.5	4.39	85	36,980	842,3
C. eugenioides	2x	645	101309	42.1	6.52	71	30,171	462,7
C. heterocalyx	2x	863	194300	60.51	2.25	42	13,732	610,3
C. racemosa	2x	506	88498	34.19	5.7	179	86,284	1513,7
C. pseudozanguebariae	2x	593	215117	91.7	15.4	68	28,669	186,1
C. humblotiana	2x	469	160479	67.99	14.49	102	45,373	313,3
C. tetragona	2x	513	160107	72.66	14.10	199	97,927	694,5
C. millotii	2x	682	163873	76.65	11.23	95	43,173	384,4
C. horsfieldiana	2x	593*	112793	46.25	7.8	72	29,593	379,3
Craterispermum sp. Novo Kribi	2x	748	49789	19.44	2.59	0	0	0

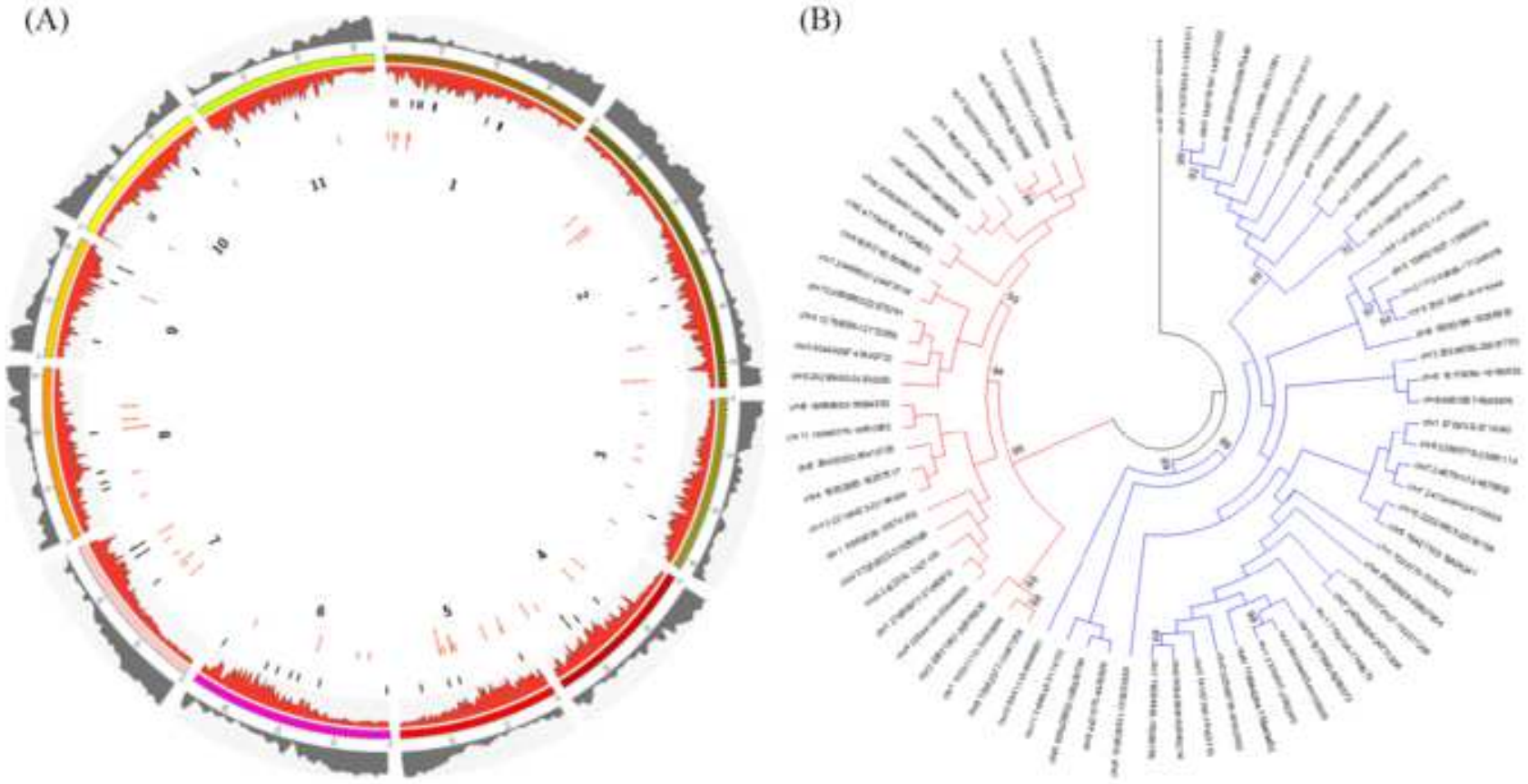
*: mean value estimates from other ex-*Psilanthus* accessions in absence of clear data for *C. horsfieldiana*.

Table 3 Likelihood ratio test for testing models of sequence evolution for Copia25 retrotransposons.

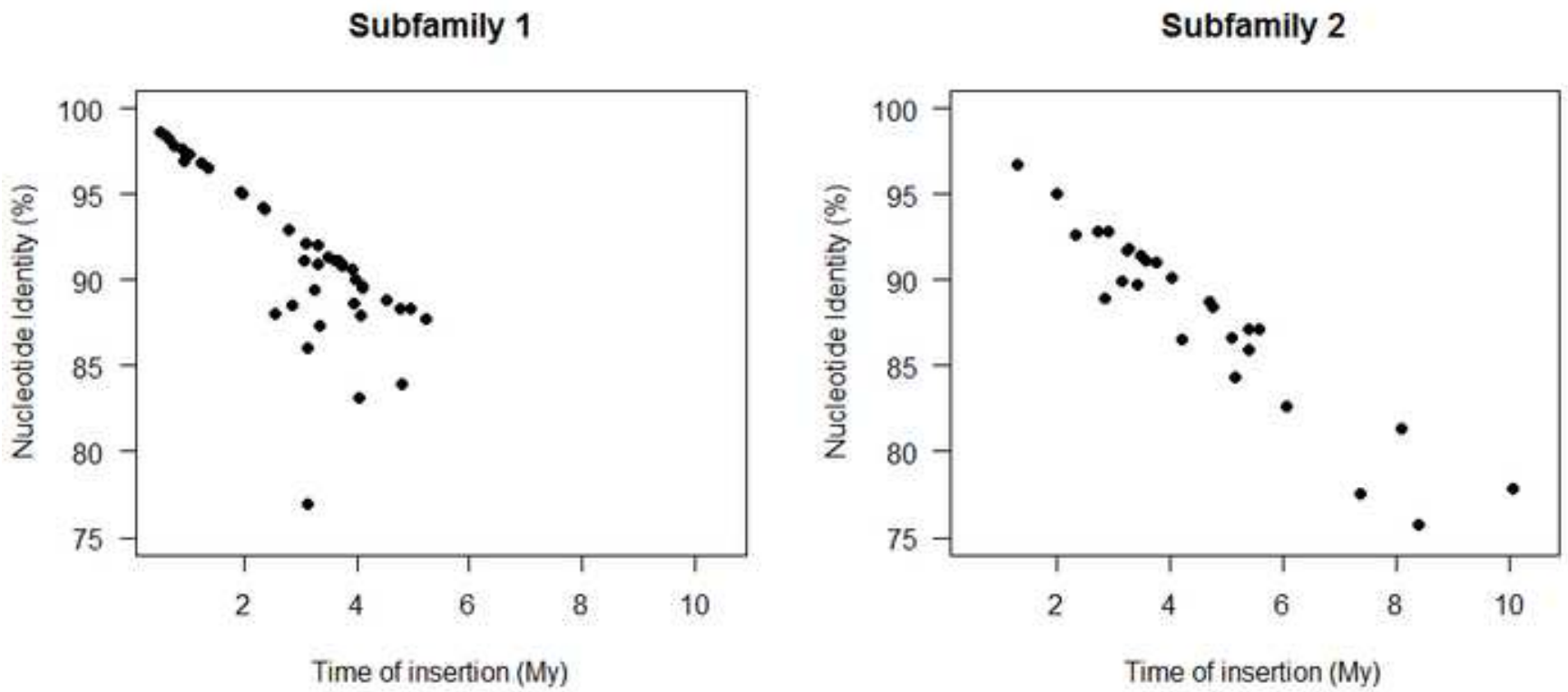
Model		Parameter	ℓ	2Δℓ	ω_B	ω_F	Conclusion
One-ratio	Model I	ω free	-2469.160	239.308**	0.191	-	Purifying selection in the *Copia25* tree
	Model II	ω = 1	-2588.814		-	-	
Two-ratio	Model III	ω free	-2468.462	33.568**	0.198	0.127	Purifying selection in the *Ixora Copia25* clade
	Model IV	ω = 1	-2485.246		0.198	1	
	Model V	ω free	-2463.734	2.526	0.169	0.552	Neutral evolution in the *Musa Copia25* clade
	Model VI	ω = 1	-2464.998		0.168	1	

Critical values of X^2, 1 df: *: 3.84; **: 6.63; $2\Delta\ell = 2\ (l_1-l_0)$

Colour figure
Click here to download Colour figure: Figure 1.tif

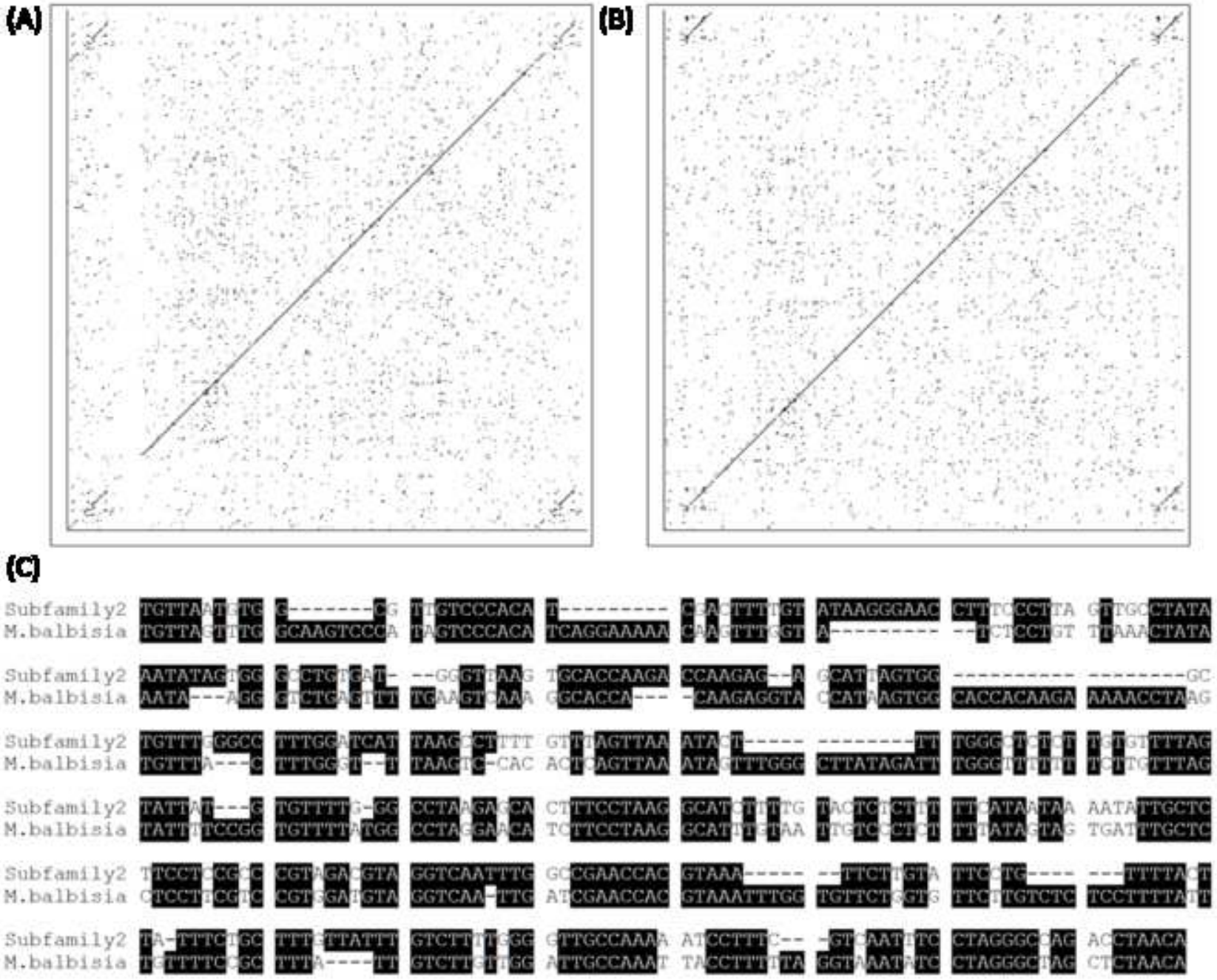


Screen/black and white figure
Click here to download Screen/black and white figure: Figure 2.tif



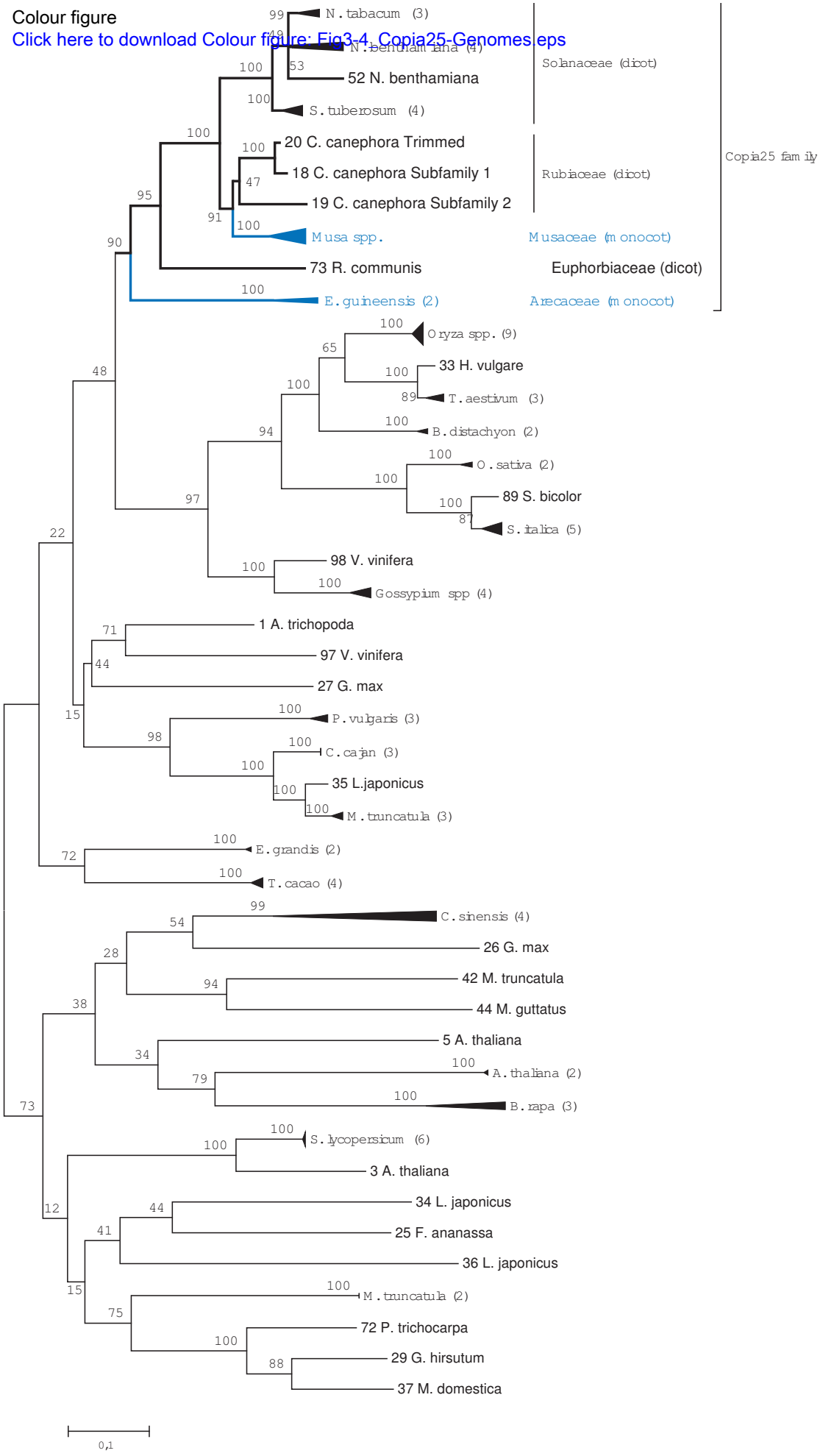
Screen/black and white figure

Click here to download Screen/black and white figure: Figure 5.tif



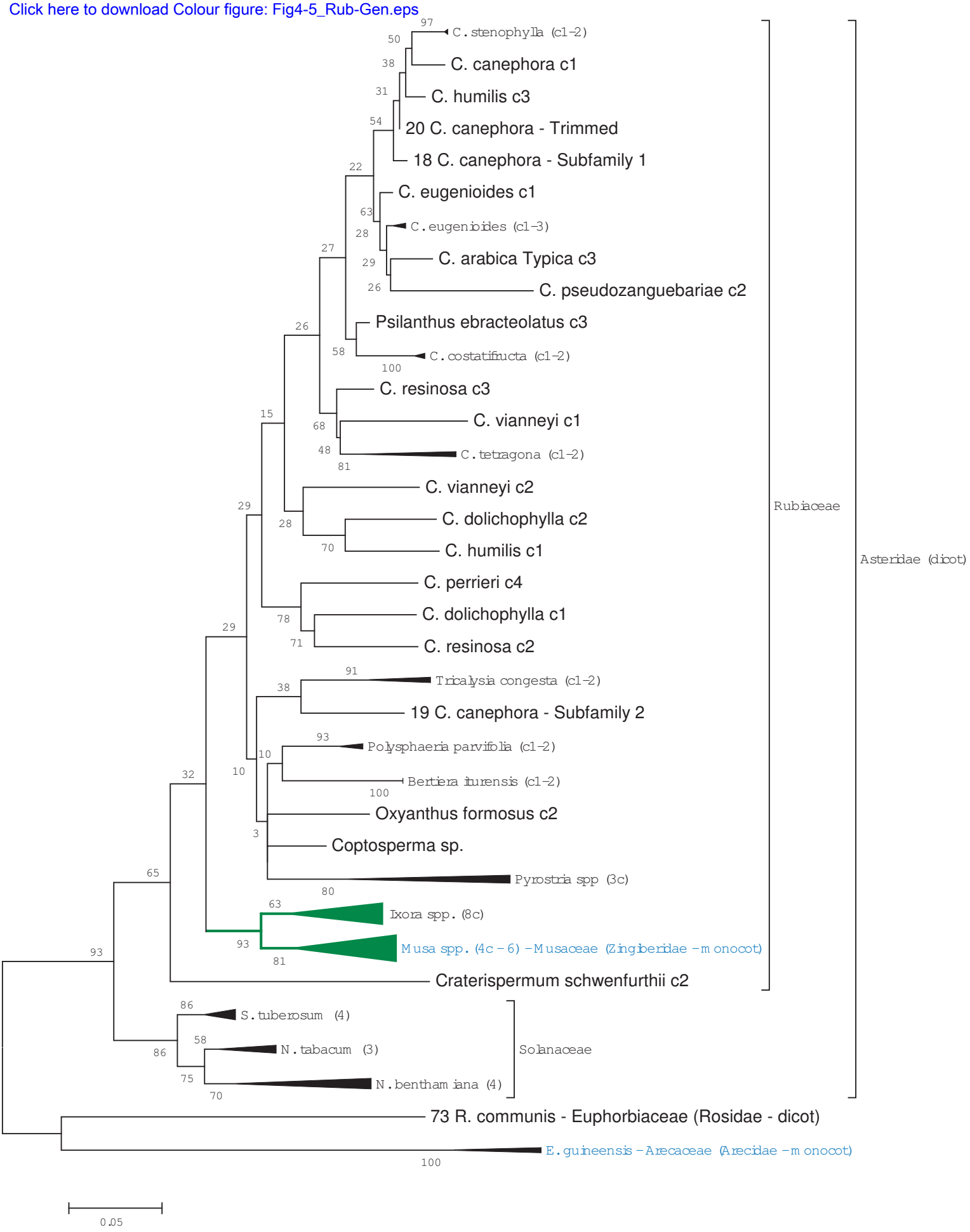
Colour figure

Click here to download Colour figure: Fig3-4_Copia25-Genomes.eps



Colour figure

Click here to download Colour figure: Fig4-5_Rub-Gen.eps



Supplementary material
Click here to download Supplementary material: Final_PMB_SM.docx