Mol Genet Genomics (2017) 292:741–754
DOI 10.1007/s00438-017-1308-2

ORIGINAL ARTICLE

Distribution of *Divo* in *Coffea* genomes, a poorly described family of angiosperm LTR-Retrotransposons

Mathilde Dupeyron[1,2] · Rogerio Fernandes de Souza[3] · Perla Hamon[1] · Alexandre de Kochko[1] · Dominique Crouzillat[4] · Emmanuel Couturon[1] · Douglas Silva Domingues[5] · Romain Guyot[2]

Received: 26 December 2016 / Accepted: 7 March 2017 / Published online: 17 March 2017

Abstract *Coffea arabica* (the Arabica coffee) is an allotetraploid species originating from a recent hybridization between two diploid species: *C. canephora* and *C. eugenioides*. Transposable elements can drive structural and functional variation during the process of hybridization and allopolyploid formation in plants. To learn more about the evolution of the *C. arabica* genome, we characterized and studied a new *Copia* LTR-Retrotransposon (LTR-RT) family in diploid and allotetraploid Coffea genomes called *Divo*. It is a complete and relatively compact LTR-RT element (~5 kb), carrying typical Gag and Pol *Copia* type domains. Reverse Trancriptase (RT) domain-based phylogeny demonstrated that *Divo* is a new and well-supported family in the *Bianca* lineage, but strictly restricted to dicotyledonous species. In *C. canephora*, *Divo* is expressed and showed a genomic distribution along gene rich and gene poor regions. The copy number, the molecular estimation of insertion time and the analysis at orthologous locations of insertions in diploid and allotetraploid coffee genomes suggest that *Divo* underwent a different and recent transposition activity in *C. arabica* and *C. canephora* when compared to *C. eugenioides*. The analysis of this novel LTR-RT family represents an important step toward uncovering the genome structure and evolution of *C. arabica* allotetraploid genome.

Keywords *Coffea* · *Copia* LTR-Retrotransposons · *Divo* · *Bianca* · Genomic evolution

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-017-1308-2) contains supplementary material, which is available to authorized users.

✉ Romain Guyot
romain.guyot@ird.fr

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

[2] IRD, CIRAD, Univ. Montpellier, IPME, BP 64501, 34394, Montpellier Cedex 5, France

[3] Departamento de Biologia Geral, CCB, Universidade Estadual de Londrina,UEL, Londrina, Brazil

[4] Nestlé R&D Tours, Notre-Dame d'Oé, Tours, France

[5] Department of Botany, Instituto de Biociências, Universidade Estadual Paulista, UNESP, Rio Claro, Brazil

Introduction

Transposable elements (TEs) are mobile genetic elements representing the main components of numerous plant genomes such as rice (35%, International Rice Genome Sequencing Project 2005), grapevine (40%, The French–Italian Public Consortium for Grapevine Genome Characterization 2007), coffee-tree (*Coffea canephora* 50%, Denoeud et al. 2014), orchids (60%, Cai et al. 2015), tomato (60%, Mehra et al. 2015), bread wheat (80%, Brenchley et al. 2012), and maize (80–85%, Schnable et al. 2009). They have the capacity to move from one locus to another within genomes, and for some of them to increase their copy numbers by doing so. Recently, it has been suggested that TEs may also propagate via horizontal transfer mechanisms among genomes of different species or even genera (Feschotte and Pritham 2007; Schaack et al. 2010; Fedoroff 2012; Dias et al. 2015; Gilbert et al. 2016; Lin et al. 2016; Panaud 2016). TEs are also considered as remarkable genome evolution drivers allowing genome adaptation and innovation through chromosome rearrangements, gene expression alterations and sometimes,

generation of new gene functions via molecular domestication of TE domains (Feschotte and Pritham 2007; Fontana 2010). During the allopolyploidy processes, TEs may represent the most dynamic fraction of the genome with major changes in their copy numbers (Parisod et al. 2010).

The faculty of producing large amount of genomic and transcriptomic sequencing data, and the availability of whole-genome sequence data, have promoted the development of bioinformatics tools to identify and to analyze genome components, including TEs (Lerat 2010). The large diversity of TEs led the scientific community to define a hierarchical classification, first separating elements according to their mode of mobility into retrotransposons, or Class 1 elements, and DNA transposons, or Class 2 elements. These classes were further subdivided into orders, super-families, lineages, and families according to their structural features and similarities (Wicker et al. 2007).

Among the class 1 elements, LTR-Retrotransposons (LTR-RTs) are the most abundant TEs in plant genomes. They represent a wide fraction of genomes ranged between 14% in *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000), up to 75% in maize (Schnable et al. 2009). LTR-RTs are divided into two super-families: *Copia* and *Gypsy* that differ mainly in their internal coding regions order (Wicker et al. 2007). *Copia* and *Gypsy* are composed of ancient and conserved lineages in plants (Wicker and Keller 2007) that can be phylogenetically classified based on their RT domain (Eickbush and Jamburuthugoda 2007). *Copia* and *Gypsy* LTR-RT may occupy different chromosomal locations as demonstrated by the available sequences of plant genomes (The Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005; The French–Italian Public Consortium for Grapevine Genome Characterization 2007; Paterson et al. 2009).

The recently released genome of *C. canephora* also contains an important fraction of LTR-RTs of 42% (Denoeud et al. 2014). The *Gypsy* elements clearly outnumber the *Copia* with 24.1% and 6.8% of the genome sequence, respectively. The remaining 11% is composed of unclassified LTR-RTs and classes small in number like *Bel/Pao*, Caulimoviruses, Retroviridae.

Coffea genus belongs to the Rubiaceae family. It contains 124 described species, originating from the intertropic forests of Africa, western Indian Ocean islands, India, Tropical and SouthEast Asia, and Australasia (Davis et al. 2011). All species are diploids with $2n = 2x = 22$ chromosomes (Bouharmont 1959; Louarn 1976), with the exception of the allotetraploid *C. arabica*, one of the two major cultivated species (Carvalho 1952). *C. arabica* has a recent origin (Yu et al. 2011), arising from hybridization between two wild diploid species: *C. canephora*, the other cultivated species (known as Robusta) and *C. eugenioides*, an East African wild species (Lashermes et al. 1999).

Previously, the two first LTR-RT elements identified in sequenced *C. canephora* Bacterial Artificial Clones (BAC) were called *Nana* and *Divo*. They were used to perform RBIP (retrotransposon-based insertion polymorphism) and REMAP (retrotransposon-microsatellite amplified polymorphism) analyses to study the species relationships within *Coffea. Divo* was particularly efficient at a low taxonomic level to resolve the genetic diversity within *C. canephora*, suggesting that the mobility of the *Divo* family participated to the *C. canephora* differentiation (Hamon et al. 2011).

In this study, we describe a genomic overview of the *Divo* family, from the *Bianca* lineage, in *C. arabica* and its two diploid progenitors, *C. canephora* and *C. eugenioides*. The *Bianca* lineage has been described in barley, Arabidopsis, and rice (Wicker and Keller 2007) and mentionned in other plant species (Kolano et al. 2013; Marcon et al. 2015; Yin et al. 2015). *Matita*, an element belonging to the *Bianca* lineage, was described more deeply in *Arachis hypogaea*, the cultivated allotetraploid peanut (Nielen et al. 2012). *Matita* appears to be present in peanut genome for a long time, as its insertions have been dated around 3,5 Mya. Its chromosomal distribution has been investigated by FISH experiments, which showed its presence mainly in distal regions of all the chromosomes. The annotated copies did not contain ORFs (stop codons and frameshifts in the putative coding regions) so the potential activity or non-activity of *Matita* has not been studied (Nielen et al. 2012). Since few data or characterizations of LTR-RTs from the *Bianca* lineage are available so far in plants, except in cultivated peanut, we selected this lineage, represented by the family *Divo* in coffee-trees, for further characterization of LTR-RT families in *Coffea. Divo* have a relatively short size (5 kb) and a moderated copy number. A RT domain-based phylogenetic analysis demonstrated that *Divo* belongs to the dicotyledonous section of the poorly known *Bianca* lineage. These elements are expressed and quite evenly distributed in the *C. canephora* genome. Differences in the abundance and in the insertion chronology of *Divo* elements were observed among *C. canephora, C. arabica*, and *C. eugenioides* genomes, suggesting different dynamics and impact on diploid and allotetraploid genomes structural evolution.

Materials and methods

Genomic sources

A total of four coffee genome sequences were used in this study: *C. canephora* DH 200−94 (Denoeud et al. 2014), accounting for 568 Mb of scaffolds and assembled into pseudo-molecules, including chromosome 0 (representing

80% of the estimated genome size i.e., 710 Mb); and three genomes sequenced with the single molecule real-time (SMRT, Pacific Biosciences—PacBio) sequencing technology: *C. canephora* (accession DH 200−94), *C. arabica* (accession Et39), and *C. eugenioides* (BU-A) accounting respectively for 679, 1060, and 789 Mb of unordered contigs. The *C. canephora, C. arabica*, and *C. eugenioides* PacBio genome sequences were generated under the Arabica Coffee Genome Consortium (ACGC 2014).

Identification, classification and annotation of LTR-RTs in *C. canephora, C. arabica* and *C. eugenioides* genomes

Potential LTR-RTs were de novo identified using the LTR_STRUC (McCarthy and McDonald 2003) algorithm against the *C. canephora* published genome, and the *C. canephora, C. arabica*, and *C. eugenioides* PacBio genomes. The predicted elements were classified into *Copia* and *Gypsy* super-families according to BLASTX similarities (Altschul et al. 1990) against a database of Gag and Pol domains (available at GyDB, http://www.gydb.org/ Llorens et al. 2011). LTR-RT predicted elements showing no similarity with any GyDB domain were not retained for further analyses.

Reverse transcriptase-based classification of LTR-RTs

The amino-acid RT domain of all LTR-RTs recovered with LTR_STRUC from each genome was extracted as described in Guyot et al. (2016), with a minimum length of 150 amino-acid residues. RT reference domains from GyDB were added to them to understand *Coffea* LTR-RTs affiliations in the *Copia* lineage. Aligned sequences were used to construct a bootstrapped neighbor-joining (NJ) tree (100 bootstrap replicates) edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Classification, annotation and characterization of the *Bianca* lineage and *Divo* LTR-RT family

The coffee LTR-RTs seqences from the *Bianca* lineage were compared to known LTR-RTs from *C. canephora* and all elements from the *Bianca* lineage in plants (Wicker et al. 2007) using BLASTN. Sequences similar to *Divo*, a previously identified LTR-RT from *C. canephora* (Hamon et al. 2011) were compared using dot-plot (Sonnhammer and Durbin 1996). To search for *Divo* similar elements in publicly available plant genomes, the sequence fragment of *Divo* described in Hamon et al. (2011) (NCBI accession HM755952.1) was used as query for similarity searches on the NCBI website (http://blast.ncbi.nlm.nih.gov/), using a BLASTX and BLASTN e-value cut-off of $1e^{-100}$ and a minimum of 50% of identity over 50% of the query sequence length. Recovered elements were annotated using BLASTX and dot-plot alignments with reference domains (Gypsy Database 2.0 web site) (Sonnhammer and Durbin 1996) and LTR_Finder (Xu and Wang 2007, http://tlife.fudan.edu.cn/ltr_finder/). Final annotations were edited with Artemis (Rutherford et al. 2000). Annotated elements were used for another phylogenetic analysis based on RT amino-acid domains as described in the previous paragraph.

Search for Divo elements in plant genomes

We searched for *Divo* LTR-RTs similar sequences in transposable elements dedicated databases: RepBase (http://www.girinst.org/, (Bao et al. 2015)), the Plant Repeat Database (http://plantrepeats.plantbiology.msu.edu, Ouyang and Buell 2004), and RetrOryza (http://retroryza.fr, Chaparro et al. 2007) using BLASTN. To better understand the evolution of the *Divo* family and its relationships with the *Bianca* lineage, we searched for sequences similar to *Divo* in eukaryote publicly available genome sequences using BLASTN and BLASTX (evalue $< e^{-100}$), using four *Divo* sequences from *C. canephora* (Denoeud et al. 2014 and PacBio), *C. arabica*, and *C. eugenioides* (accessions #: KX767840, KX767841, KX767839 and KX767842). 22 genomic sequences were recovered from 14 angiosperm species and their RT amino-acid domains were used to construct a NJ phylogenetic tree (*Oryza sativa*—accession #AC147802.2, *A. thaliana*—#AP002459, *V. vinifera*—#AM477556.1, *Sorghum bicolor*—#AF466199.1, *Zea mays*—#DQ493648.1, *Rosa rugosa*—#JQ791545.1, *Theobroma cacao* (Jurka 2014—accession #HQ244500), *Fragaria vesca*—#XM_004309244.1, *Ipomoea trifida*—#AY4480105.1, *Beta vulgaris*—#GU057342.1, *Arachis hypogaea*—#HQ637177.1, Oryza rufipogon—#FO681399, *Solanum lycopersicum*—#AAK84483, *M. truncatula*—#CM001223. Additional LTR_RT sequences from TAIR and RetrOryza database and LTR_STRUC output for *A. thaliana, V. vinifera*, and *O. sativa*).

Divo homologous elements were also specifically searched for and characterized from two reference plant genomes: *Arabidopsis thaliana* (GCA_000001735.1) and *Vitis vinifera* (GCA_000003745.2) available from TAIR (https://www.arabidopsis.org) and NCBI (http://www.ncbi.nlm.nih.gov/). First, all potential full-length LTR-RTs were de novo searched with LTR_STRUC and compared by BLASTN with *Divo* elements identified previously. Second, all LTR-RT sequences from Arabidopsis and grapevine previously identified and available in the Plant Repeat Database (http://plantrepeats.plantbiology.msu.edu/search.html) were downloaded and compared by BLASTN with coffee *Divo* elements.

Copy number and insertion time of *Divo* in *C. canephora, C. arabica*, and *C. eugenioides*

Assessment of *Divo* copy number in of *C. canephora* (Denoeud et al. 2014) and the *C. canephora, C. arabica*, and *C. eugenioides* PacBio genomes (ACGC 2014) was carried out with Censor (Kohany et al. 2006). A complete *Divo* element is considered when it contains both ORFs Gag and Pol and a minimum of 99% sequence identity between both LTRs. Such a sequence was found in the *C. canephora* genome and was used as a reference for similarity searches (accession number #KX767841). A copy is considered if it covers a minimum of 80% of the reference sequence with at least 80% of nucleotide identity (Wicker et al. 2007) and a fragmented copy is considered if it covers a minimum of 20% of the reference sequence with at least 80% of nucleotide identity. Full-length copies were also extracted according to the following definition: 80% of nucleotide identity over 100% of the reference sequence length as well as potential solo LTRs (80% of identity over 100% of the LTR sequence length). The genomic distribution of the identified elements in the *C. canephora* pseudochromosomes was established using Circos (Krzywinski et al. 2009).

The insertion time of full-length *Divo* copies was estimated based on the divergence of the 5′- and 3′-LTR sequences of each identified full-length copy. The two LTRs were aligned using Stretcher (EMBOSS), and the divergence (K) was calculated using the Kimura 2-parameter method implemented in Distmat (EMBOSS). The insertion dates (T) were estimated using the formula $T = K/2r$ (SanMiguel et al. 1998) where we used average base substitution rates (r) of $1.3e^{-8}$ established by Ma & Bennetzen (2004).

Presence of *Divo* at orthologous locations in three coffee-trees genomes

Insertion of full-length copies of *Divo* in *C. canephora, C. eugenioides* and *C. arabica* at orthologous locations among the three genomes were compared. As a first step, genomic regions containing full-length *Divo* copies were recovered from the *C. canephora* contigs adding 2 kb upstream and downstream the element. The recovered genomic fragments are then compared as queries using BLASTN (evalue $1e^{-100}$) against the other two genomes. The best results (lowest e-values and highest scores) are then extracted and compared to the queries using dot-plot alignements (Sonnhammer and Durbin 1996). Finally, dot-plot alignments are manually evaluated to classify the orthologous relationships into the following categories: (i) queries are not conserved and so no orthologous regions could be identified; (ii) queries are conserved within an orthologous region but the *Divo* element is not conserved, and (iii) queries are conserved within an orthologous region and the *Divo* element is present at the same insertion site. These steps are repeated for the full-length copies of *Divo* in *C. arabica* and *C. eugenioides*.

Search for *Divo* potential expression in *C. canephora* tissues

RNA sequencing (RNA-seq) data generated under the *C. canephora* genome project (Denoeud et al. 2014) from leaves, roots (*C. canephora* accession #T3518), stamen, and pistil (*C. canephora* accession #BP961) were used to identify the transcriptional pattern of reference sequences. The 130.10^6 RNA-Seq reads were cleaned using prinseq (Schmieder and Edwards 2011) and mapped against 18 *Divo* sequences using Bowtie 2 (Langmead and Salzberg 2012). The number of mapped reads per TE sequence was processed and RPKM (reads per kilo base per million) were calculated. A heatmap representing the expression profiles was computed using Heatmap3 package in RStudio (2012). Differential expression among available RNA-seq libraries was detected using Winflat (Audic and Claverie 1997) with significance threshold of 0.05 and Bonferroni correction. This analysis was performed with IDEG6 software (http://telethon.bio.unipd.it/bioinfo/IDEG6_form/) (Romualdi et al. 2003).

Results

Copia LTR-RTs in *C. canephora, C. arabica* and *C. eugenioides* genomes

Since LTR-RTs represent the main part of the TE fraction found in the *C. canephora* genome, we focused our analyses on these elements, and more specifically on *Copia* LTR-RT lineages and families. LTR_STRUC identified 1799 (588 *Gypsy* and 474 *Copia*), 7363 (2010 *Gypsy* and 999 *Copia*), 4346 (2153 *Gypsy* and 1080 *Copia*) and 3591 (1632 *Gypsy* and 913 *Copia*) LTR-RT elements, for *C. canephora* (Denoeud et al. 2014), *and C. canephora, C. arabica*, and *C. eugenioides* (ACGC), respectively. We specifically screened and filtered out LTR_STRUC potentially complete elements according to similarities with the *Copia*-specific domains. The reverse transcriptase (RT) amino-acid domains of *Copia* recovered sequences were extracted and used for a NJ phylogenetic analysis. The analysis of the resulting NJ trees for *C. canephora, C. arabica*, and *C. eugenioides* shows that coffee RT *Copia* domains were classified into all five *Copia* lineages previously described in plants: *Tork, Oryco, SIRE, Retrofit*, and *Bianca* (Llorens et al. 2009; Wicker and

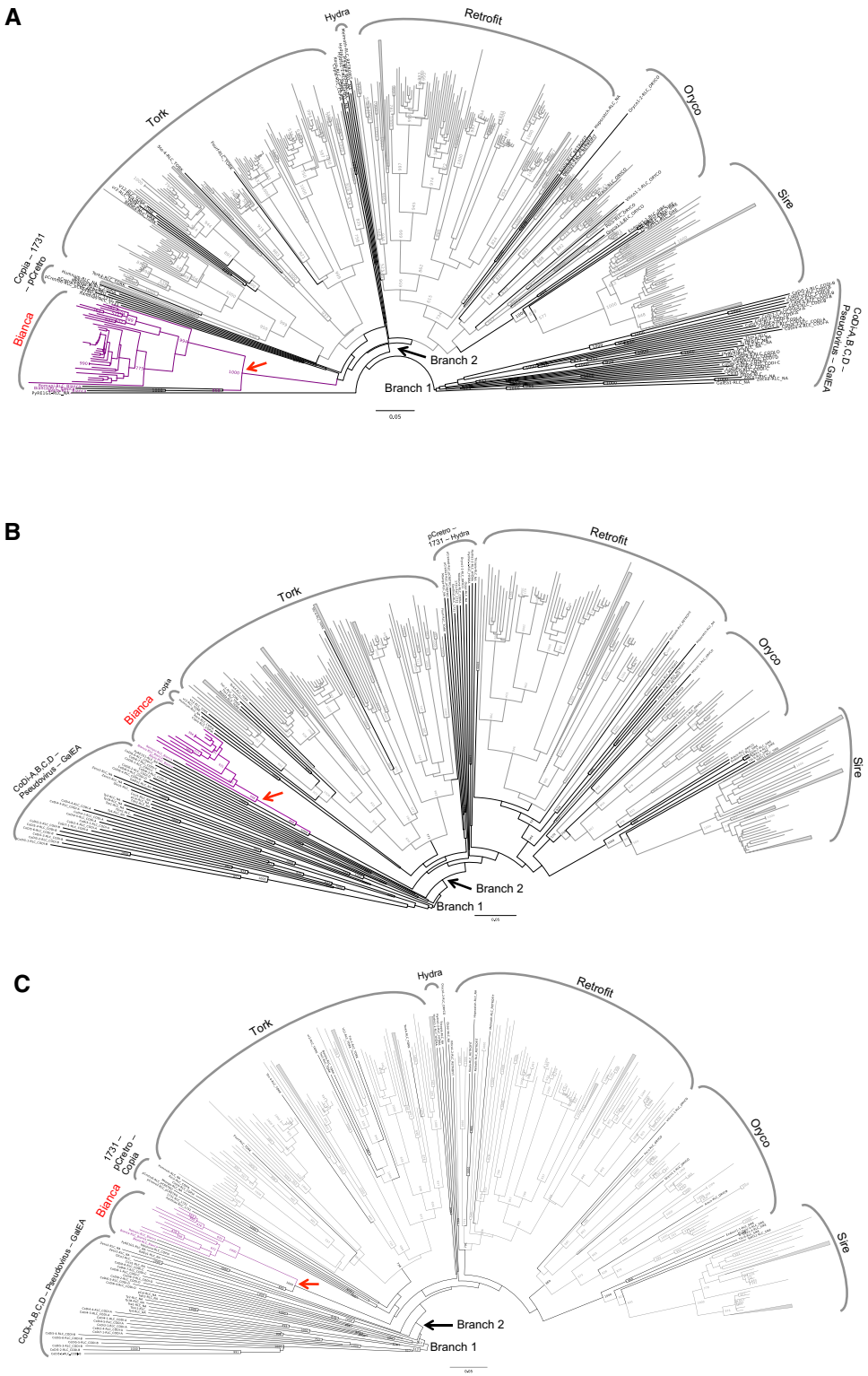

Fig. 1 Phylogenetic analysis of LTR retrotransposons sequences predicted from *C. canephora* (**A**), *C. arabica* (**B**) and *C. eugenioides* (**C**) genomes. Phylogenetic trees were based on amino-acid alignments of the reverse transcriptase (RT) domains; 999, 1080, and 913 amino acids, repectively, from *C. canephora, C. eugenioides*, and *C. arabica* genomes. The classification into lineages was done according to the RT reference domains (*black* branches) downloaded from GyDB. The *Coffea* sequences within the *Bianca* lineage are indicated by a *red arrow*, and lineages are indicated by brackets and names

Keller 2007, Fig. 1). References RT domains from other organisms Diatoms (*CoDI*), Fungi (*Pseudovirus, pCretro*) and Arthropoda (*1731, Hemivirus*) were found clustered outside of plant lineages that include coffee, according to their classification into Branch 1 and 2 (Llorens et al. 2009). The diversity of *Copia* lineages appears very similar between the three species analyzed (Fig. 1). One of the smallest clades called *Bianca* and supported by strong bootstraps (Fig. 1), grouped together 12 sequences from *C. canephora* (Denoeud et al. 2014 and 13 sequences in

PacBio genome), 14 from *C. arabica*, and 12 from *C. eugenioides*.

Divo elements in C. *canephora, C. arabica*, and *C. eugenioides* genomes

In total, 89 full-length elements belonging to the *Bianca* lineage and recognized by LTR_STRUC or BLASTN in the four coffee genome sequences (Supplemental Data 1) were analyzed and annotated. The structure of these elements corresponds to the typical organization of *Copia* elements with two LTRs at each extremity and two ORFs: Gag and Pol containing the protease (PR), integrase (INT), reverse transcriptase (RT), and RNase H (RH) domains, in this specific order. The LTRs were 350 bp long and were terminated with the LTR consensus: 5′TG…CA3′. The overall length of complete elements (i.e., elements carrying two highly conserved LTRs and complete Gag and Pol ORFs) ranged between 5276 bp and 5636 bp (Fig. 2a). The Gag sequence (1065 bp long, separated from Pol by 5 stop codons in all the elements found) presented similarities with the UBN2 family domain (Pfam14223—nucleotide position 718–942). UBN2 is a form of the peptide encoded by the Gag ORF frequently found in the *Copia* LTR-RT superfamily. A Zinc finger amino-acid motif (ZnF_C2HC, nucleotide position 1318–1365), involved in nucleic acids binding, is also found in the peptide encoded by this ORF. The Pol ORF (3501 bp), showed high similarities with Gag_pre-integrase family (Pfam13976, position 2050–2268), Integrase (INT) core domain (Pfam00665, position 2305–2655), Reverse transcriptase (RT) genes (Pfam 07727, position 4645–5085), and RNase-H (RH) domain (position 3637–4374), in this specific order. All these domains show high similarities with *Copia* LTR-RTs.

While the polypurine track motif (PPT, used for the synthesis of the complementary DNA strand) is found upstream the 3′ LTR, the primer-binding site (PBS) presents unusual sequence conservation (Fig. 2b). Among the 89 elements precisely analyzed here, only one (Accession #KX767840) showed a complementary sequence to a tRNA ($tRNA^{Ile}$ (AAT)).

Similarity analyses between coffee full-length elements belonging to the *Bianca* lineage, known *Bianca* elements (Wicker et al. 2007) and known coffee elements showed a relatively good nucleotide conservation with *Divo*, a *Copia* LTR-RT element identified earlier in a *C. canephora* BAC sequence and used to assess insertion site polymorphism (Hamon et al. 2011). Comparisons between *Divo* and a complete and potentially active element in *C. canephora* revealed by LTR_STRUC (Accession #KX767841) indicated an overall percentage of nucleotide identity of 63.6% and a LTR percentage of identity of 58% and 56.7% for the 5′ and 3′ LTR, respectively. This relatively low percentage of nucleotide identity is probably due to the absence of several regions of the *Divo* element identified earlier (Hamon et al. 2011; Fig. 2c). This percentage is similar for all full-length coffee-trees elements. Nevertheless, we named the novel annotated sequences, carrying the new group of RT domains similarly to the initial element discovered earlier: *Divo*. A reference *Divo* element was ascertained for each of the three *Coffea* genomes, based on the most conserved annotated sequence found. These references were used for different analyses when they needed a reference sequence. All the recovered sequences of *Divo* presenting a good conservation and no stop codon in the RT domain were used in RT-based phylogenies, which confirmed their affiliations to the *Bianca* lineage and the *Divo* family (Supplemental data 2).

We also searched for the transcriptional pattern of the *Divo* family using RNAseq reads (Denoeud et al. 2014) from leaves, roots, stamen, and pistil mapped on the 18 *Divo* sequences found in *C. canephora* published genome with LTR_STRUC. Transcriptional pattern suggested transcriptional modulation when vegetative tissues (leaves or roots) are compared to reproductive tissues (stamens or pistils). Seventeen *Divo* exhibited differential expression between leaves or roots versus stamen or pistil, while only seven presented differential expression between leaves and roots and none between pistils and stamen. In addition, a lower degree of expression of these retrotransposons was detected in pistil and stamen when compared to leaves and roots (Supplemental data 3).

Copy number estimation and insertion time of *Divo* elements in *C. canephora, C. arabica* and *C. eugenioides*

One hundred and ninteen, 204, and 132 copies of *Divo* were, respectively, found in *C. canephora* (Denoeud et al. 2014), *C. canephora, C. arabica*, and *C. eugenioides* ACGC sequences (Table 1). Besides looking for highly conserved copies (100% of coverage and ≥80% of identity), less conserved or fragmented copies (80% of identity on at least 20% of the total length) and solo LTRs (Devos et al. 2002) were also detected. Higher copy numbers were obtained for *C. canephora* ACGC sequences, probably due to the completeness of the sequencing technology used. Interestingly, *C. eugenioides* showed a higher *Divo* total copy number when compared to *C. canephora*, but with the notable exception of full-length copies. The allotetraploid genome of *C. arabica* contains the highest total *Divo* copy number. However, for each category, the number of copies in *C. arabica* is lower than the sum of its diploid progenitors. The ratio of solo LTR to full-length or "intact" elements was in a similar order of magnitude for *C. canephora* (4.7:1 and 3.4:1) and *C. arabica* (5.4:1), but three times higher for *C. eugenioides* (16.8:1). In the annotated *C.*

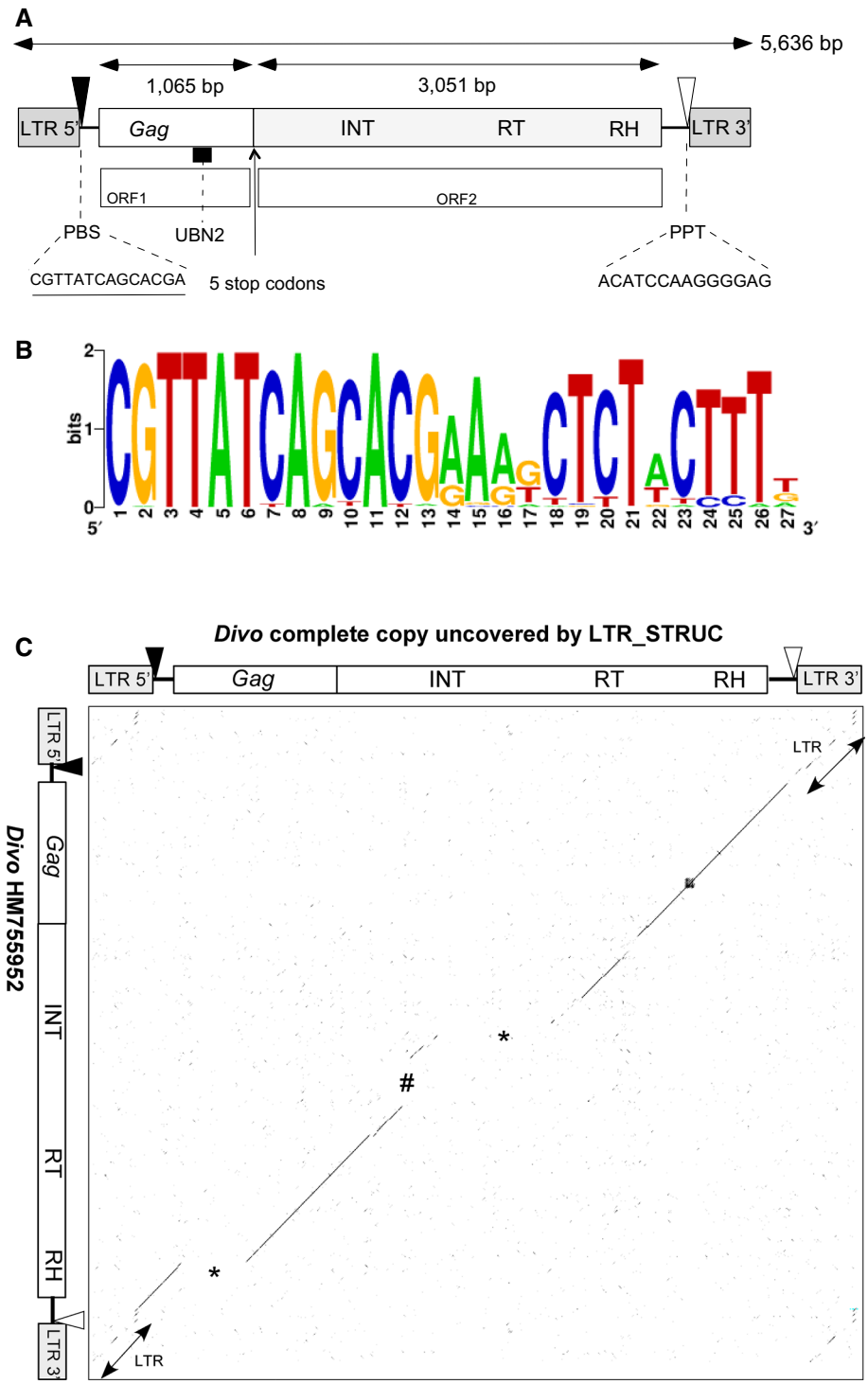

Fig. 2 Structure of the *Copia* LTR-RT *Divo*. **a** Structural features of the *Divo* family. The complete *Divo* element was identified in *C. cenephora* genome (KX767841). *Gag* and *Pol* ORFs are separated by five stop codons. *LTR* long terminal repeats, *PBS* primer-binding site (*black triangle*), *PPT* polypurine tract (*open triangle*), *UBN2* ubinuclein 2 domain, *INT* integrase, *RT* reverse transcriptase, *RH* RNAse H. **b** Web-Logo representation of the PBS of *Divo* full-length copies found in (**c**) *canephora* and *C. arabica*. **c** *Dotter* alignment between the fragmented *Divo (*Hamon et al. 2011, HM755952) and a complete *Divo* element uncovered by LTR_STRUC in *C. canephora* (KX767840). *Asterisks* regions absent in *Divo* but present in the complete element. **#** Regions present in *Divo* but absent in the complete element. The positions of LTR are indicated

canephora pseudo-molecules, the *Divo* family, whatever the status of the copy (full-length, "80–80," fragmented or solo LTR), appears equally distributed along TE-rich and gene-rich regions (Supplemental data 4).

Complete copies LTR sequences (80–100%) were used to calculate their nucleotide divergence and estimate their insertion times in *C. canephora, C. arabica*, and *C. eugenioides* according to the substitution rate established by SanMiguel et al. 1998 (Fig. 3). Our analysis indicates relatively recent insertions of *Divo* in *C. canephora* and in *C. arabica* (at 0–0.5 Mya), while in *C. eugenioides*, two more ancient peaks (at 0.5–1 and 1.5–2 Mya, red line) are detected. Interestingly, the second ancient peak observed in *C. eugenioides* is also detected in a lesser extent in *C.*

Table 1 Estimation of the copy numbers of *Divo* elements in the *C. canephora* genome (*, Denoeud et al. 2014) and *C. canephora, C. arabica*, and *C. eugenioides* genome sequences (§, PacBio)

	Number of intact copies (80–100)	Number of copies (80–80)	Number of partial copies (20–80)	Number of solo LTRs	Solo LTR/ intact copies ratio	Total
*C. canephora**	28	119	199	132	4.7:1	478
C. canephora§	41	129	212	142	3.4:1	524
C. arabica§	37	204	351	201	5.4:1	793
C. eugenioides§	20	132	223	336	16.8:1	711

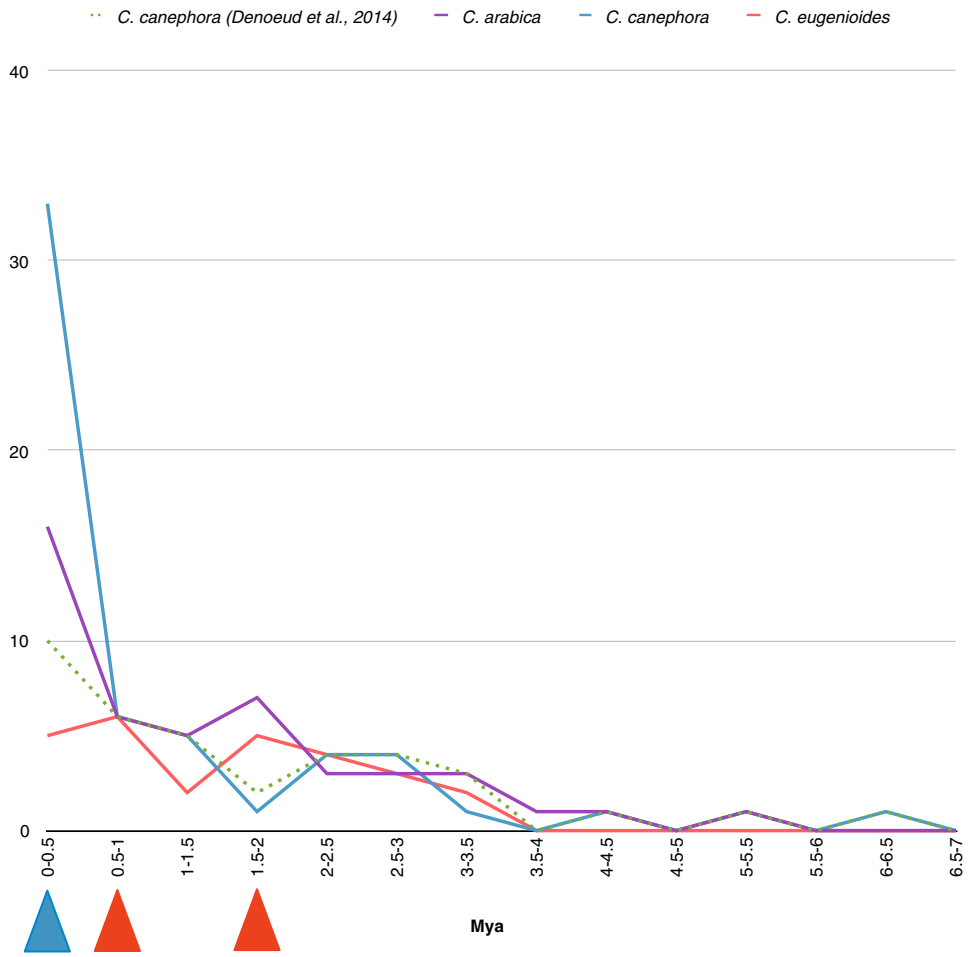

Fig. 3 Estimation of insertion times of *Divo* elements in coffee genome sequences. The LTR sequences of 178 full-length elements uncovered from *C. canephora, C. arabica*, and *C. eugenioides* genomes were used to estimate insertion time using the substitution rate of 1.3×10^{-8} (Ma and Bennetzen 2004). *Blue, red, purple*, and *green lines* represent insertion times respectively in *C. canephora, C. eugenioides, C. arabica*, and *C. canephora* (Denoeud et al. 2014)

arabica (purple line), showing a good conservation of copies from the *C. eugenioides* parental genome in the allotetraploid.

Comparison of orthologous regions of full-length *Divo* insertions between *C. canephora, C. arabica* and *C. eugenioides*

Insertion sites of 39, 37, and 20 *Divo* full-length copies were mined in *C. canephora, C. arabica*, and *C. eugenioides* genomes, respectively, with their location given by Censor (Kohany et al. 2006). 31 specific insertion sites are represented by blue, purple, and red squares, respectively (Fig. 4). In orthologous regions, 16 copy sites are shared between *C. canephora* and *C. arabica* (blue stars), six between *C. arabica* and *C. eugenioides* (red stars), and one between *C. canephora* and *C. eugenioides* (blue and red stars at 0,7 My). Twenty-four copy sites are shared between the three genomes (yellow circles). Copy sites shared between the three genomes are dated from 0.8×10^6 up to 3.2×10^6 years. In *C. canephora*, specific copy insertions are dated from 0 to 3.1×10^6 years.

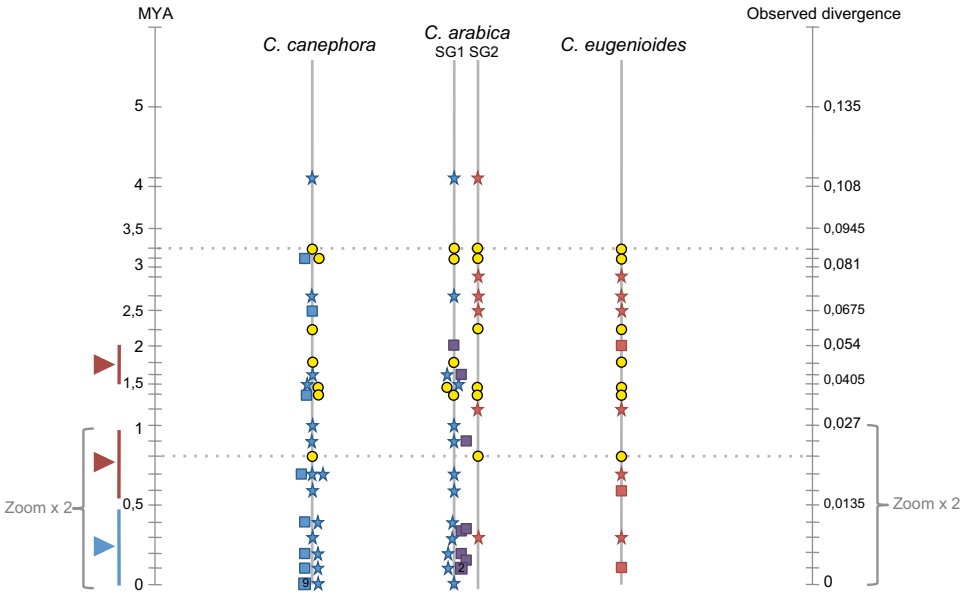

Fig. 4 Timing of insertion of *Divo* and comparative orthologous analysis in *C. canephora, C. eugenioides*, and *C. arabica* ACGC genomes. The *vertical line* on the right shows the divergence scale of LTRs for each element. The *vertical line* on the left shows the insertion times in Mya estimated with the molecular clock of Ma and Bennetzen (2004) (1.3 e^{-8} substitution per site and per year). Peaks of insertions observed in *C. canephora* (0–0.5 Mya) and *C. eugenioides* (0.5–1 and 1.5–2 Mya) relating to Fig. 3 are symbolized by the *blue* and *red triangles*, respectively. The insertion sites are located according to their estimated insertional time. *Yellow circles* represent *Divo* insertions at orthologous sites in the three species. The two horizontal *gray dashed lines* indicate the most recent (0.7 Mya) and the oldest (3.3 Mya) *Divo* elements present in the three species. Noted that for *C. arabica*, the most recent insertion is absent from one sub-genome. Insertions shared between two species are represented in *blue* or *red stars* according to the species involved. The most recent copies shared by *C. eugenioides* and one sub-genome of *C. arabica* in one hand, and *C. canephora* and the other sub-genome of *C. arabica* in the other hand both dated from 0.3 Mya. The oldest copies shared by *C. canephora* and *C arabica* on one hand, and *C. eugeniodies* and *C. arabica* on the other hand, both dated from 2.6 Mya. *Divo* insertions present in only one species are represented by *blue*, *purple*, and *red boxes* respectively for *C. canephora, C. eugenioides*, and *C. arabica* (represented by its two sub-genomes SG1 and SG2). Numbers in *boxes* indicate copy numbers at the site. *Purple boxes* between the two sub-genomes for *C. arabica* indicate unknown sub-genome identification for these insertions

Divo elements in plant genomes

Only one element found in Repbase called *Copia_12* (http://www.girinst.org/2014/vol14/issue9/Copia-12_TC-I.html), showed significant similarity with *Divo* (76% of nucleotide similarity between internal regions and 48.1% between the LTRs). *Copia_12* was annotated in the *Theobroma cacao* genome (Argout et al. 2011), but the element was neither characterized nor classified. Dotplot alignment between *Divo* (Accession #KX767841) and *Copia_12* confirmed the overall conservation of the elements structure with the exception of the LTR regions (only 52% of identity), suggesting that *Copia_12* may belong to the *Divo* family and so that the *Divo* family is not restricted to the *Coffea* genus (Supplemental data 5). We also checked the identity between our sequences of *Divo* from *Coffea* and the *Matita* element from *Arachis duranensis* (accession #JQ040302). The identity between *Matita* and the reference copies of *C. canephora* (Denoeud et al. 2014) and *C. canephora, C. arabica*, and *C. eugenioides* PacBio is of 53.7, 57, 57.2 and 57.1%, respectively. These percentages of identity indicate that *Matita* could effectively be a *Divo* element, but with a different history in *Arachis* genomes, leading to a significant sequence divergence with the *Divo* family from *Coffea*. Moreover, *Matita* is not complete and probably quite degenerate, explaining the weak percentages of identity with complete *Divo* elements.

Using four *Divo* sequences from *C. canephora* (Denoeud et al. 2014 and PacBio), *C. arabica* and *C. eugenioides* (accessions #: KX767840, KX767841, KX767839, and KX767842) as references (best intra-LTR sequence conservation: 97.4, 99.4, 99.4, and 99.7%, respectively, and longest ORF for Pol region). We searched for *Divo* in publicly available plant genomes. 22 genomic sequences were recovered from 14 angiosperm species and their RT amino-acid domains were used to construct a NJ phylogenetic tree (Supplemental data 5). *Divo* from *Coffea* form one monophyletic group inside the *Bianca* lineage. Interestingly, similar sequences to *Divo* found in the previously mentioned plant genomes were separated into two clear clades, corresponding to monocots and dicots, suggesting the *Bianca*

lineage is composed of two families: one for monocots and the other named *Divo* for dicots.

To further characterize *Divo* in dicots, we decided to annotate these elements in two reference genomes: *A. thaliana* (140 Mb) and *V. vinifera* (~500 Mb). A total of 197 and 1,384 potential LTR-RTs were detected in these genomes by LTR_STRUC. Out of these, seven and 44 sequences similar to *Divo* were recovered from the *A. thaliana* and *V. vinifera* genomes, respectively. The overall structure of these sequences is strictly similar to that of the complete *Divo* sequence (#KX767841) (Supplemental data 6), including the total length of the elements (an average of 6,071 bp for *A. thaliana* and 5,824 bp for *V. vinifera*) and the length of LTRs (335 bp on average for *A. thaliana* and 314 bp on average for *V. vinifera*).

In *A. thaliana*, four copies are potentially functional since no frame-shift was present in the ORFs of these elements. One of these (called L34-161, LTRs identity of 98.2%), displays a unique large ORF including the Gag and the Pol regions, as found frequently for *Copia* LTR-RTs (Peterson-Burch and Voytas 2002), but so far unique for all the *Divo* sequences analyzed. In grapevine, three sequences appeared potentially functional. One of them, called L107-1314 (LTRs identity of 96.8%), seems the most conserved as it carries only one stop codon between the Gag and Pol regions, contrary to the two others.

Finally, an analysis of the putative PBS region in 120 *Divo* sequences (from the copies of *C. canephora, C. arabica, C. eugenioides, Arabidopsis thaliana, Vitis vinifera, Brassica rapa, Medicago truncatula* and *Matita*) indicated that only the first 14 bp of the PBS region is conserved, particularly the four nucleotides "TTAT," while the 3′ ends were found more diverse (Fig. 2b).

Altogether these results suggest that *Divo*, the family of LTR-RTs described for the first time from complete elements, is actually conserved among a large panel of dicot plants.

Discussion

A novel LTR-RT family conserved among dicotylodenous plants

We uncovered a novel LTR-RT family called *Divo* in diploid and allotetraploid coffee-tree genomes. This family is related to a degenerated element previously annotated in a *C. canephora* BAC clone and used to study the relationships between 32 *Coffea* species (Hamon et al. 2011). *Divo* was classified into the *Bianca* lineage using a phylogenetic analysis (Fig. 1 and Supplemental data 2) and because it shares the same key structural features with elements from this lineage such as the overall length of the element and LTR sizes (Wicker and Keller 2007; Nielen et al. 2012). However, *Divo*-like homologous sequences were restricted to dicots, suggesting that the *Divo* family evolved specifically since the divergence between dicots and monocots.

Bianca is the most ancient *Copia* lineage as showed by our RT-based phylogenetic analysis (see also Piednoël et al. 2013). *Bianca* elements have been initially detected in Triticeae, rice, Arabidopsis and alfalfa (Wicker and Keller 2007; Wang and Liu 2008). Whereas the *Bianca* lineage was not found in soybean (Du et al. 2010), sugarcane (Domingues et al. 2012) or quinoa (Kolano et al. 2013), it was frequently found in Angiosperm genomes (Piednoël et al. 2013), confirming that this ancient lineage was spread along the Angiosperms divergence. The *Bianca* lineage was also frequently found with a moderated copy number, such as in Arabidopis, rice, peanut, eucalyptus, and poplar (Wicker and Keller 2007; Nielen et al. 2012; Marcon et al. 2015; Natali et al. 2015), with the exception in the pear genome, where *Bianca* represents the highest copy number lineage of all *Copia* elements (Yin et al. 2015).

Similarly to other Angiosperm genomes, *Divo* was found in coffee-trees with a moderate copy number, suggesting that coffee host genomes may apply a control of the copy number of this family,

One of the main characteristics of the *Divo* family is an atypical PBS that did not show any strong complementary sequence to host tRNAs (Fig. 2). A PBS is usually composed of 11 to 18 nucleotides complementary to a host tRNA that primes the reverse transcription of the element (Le Grice 2003). However, the detection of recent *Divo* element insertions based on the LTR divergence suggests potential recent mobility. Further studies, including the detection of circular dsDNA molecules, suggesting replicative forms of the elements (Mirouze et al. 2009), might bring more evidence about the actual transpositional activity of *Divo*.

The comparison of the *Divo* alleged PBS (Fig. 2, CGT TATCAGCACGA) with those of the families *Romani* in Arabidopsis (GTTTATCAGCAC, Wicker and Keller 2007), *Matita* in peanut (TGTTATCAGCAC, Nielen et al. 2012) and *Mtr13* in *Medicago* (CGTTATCAGCACGC, Wang and Liu 2008) suggest that it could be conserved in different families from the *Bianca* lineage. Other groups of LTR-RTs lacking PBS identification were previously characterized in *Aedes aegypti* (Minervini et al. 2009) and in *Dictyostelium*, (Leng et al. 1998), suggesting that these LTR-RTs may not need a functional PBS and/or that they could use another primer to accomplish their replication cycle.

Divo in diploid and allotetraploid coffee-trees genomes

The time of LTR-RTs insertions in genomic sequences can be roughly estimated using the divergence between

LTR sequences of each element, as these regions are supposed to be strictly identical in an active copy at the time of each insertion. Since no specific substitution rate is available for *Coffea*, we used the one estimated by Ma & Bennetzen (2004) for rice LTR-RTs ($1.3e^{-8}$ substitution per site per year), and often applied to other dicots and monocots LTRs divergence analyses (Vitte and Bennetzen 2006). Estimation of LTR-RTs time of insertions in the studied *Coffea* species showed that these elements were differentially amplified in the last 2.5 My. The *C. canephora* ACGC genome contains more recent *Divo* copies than the other genomes and more than the published *C. canephora* genome (Denoeud et al. 2014), which is probably a consequence of the higher quality and completeness reached by the sequencing technology (Fig. 3). Particularly, 18 recent insertions (100% of nucleotide conservation between their LTRs) were observed in *C. canephora*, suggesting that *Divo* was amplified and activated recently in this species, and with a lesser extent in *C. arabica*. On the contrary, almost no recent insertions were detected in *C. eugenioides* (Fig. 4). This result is in agreement with the data obtained by Hamon et al. (2011), where they showed that *Divo* is accompanying the *C. canephora* diversification but not that of the genus *Coffea*, including *C. eugenioides*. As we can observe recent and specific insertion sites in *C. arabica* (Fig. 4), *Divo* could yet also be active or would have been active in the actual *C. canephora* ancestor of *C. arabica*. On the contrary, *C. eugenioides* did not show recent transpositions, while two discrete periods of activity at 0.5–1 and 1.5–2 Mya were evidenced. Furthermore, a high number of solo LTRs were detected in *C. eugenioides*, suggesting that the control of *Divo* copy number may be more efficient in this genome via unequal homologous recombination mechanisms (Bennetzen and Kellogg 1997). The distinct periods of transposition and removal activities of *Divo* between *C. canephora* and *C. eugenioides* indicate a different evolution of the genome structural dynamics of these two diploids. As expected, the insertion periodes of *Divo* elements within *C. arabica* genome share the pattern of both *C. canephora*, with a recent activity (0–0.5 Mya) and *C. eugenioides*, with a secondary and more ancient peak of insertions (1.5–2 Mya; Fig. 3). This pattern (common timing insertion with diploid ancestor, conservation of orthologous copies, and copy number estimation) suggests that the allotetraploid genome of *C. arabica* did not suffer of strong elimination or increase of *Divo* copy number following the allopolyploidization. This result differs from other LTR-RT families in allopolyploid genomes that underwent modifications of their copy numbers after polyploidization (Ainouche et al. 2009; Parisod et al. 2010). Further and wider comparative analysis of LTR-RTs between the *C. arabica* genome and its two diploid progenitors will bring interesting information concerning the consequences of the polyploidization on the LTR-RTs dynamics and control in this model.

An evolutionary scenario for diploid and allotetraploid genomes divergence

We used the complete copies of *Divo* conserved in orthologous regions between the *C. arabica* genome and its two diploid progenitors, *C. canephora* and *C. eugenioides*, to better understand the evolution of their genomes. The relative time of insertion of *Divo* copies allowed us to propose an evolutionary scenario for the divergence time between *C. canephora, C. eugenioides*, and for the formation of *C. arabica*.

The relative time for the *C. canephora* and *C. eugenioides* radiation can be investigated thanks to the conservation of *Divo* copies at orthologous sites, corresponding to *Divo* copies likely inserted in the common ancestor of the two diploid genomes. Such orthologous copies had an estimated time of insertion ranging between 3.1 and 0.8 Mya, suggesting that the two species completely diverged at least 0.8 Mya. However, *Divo* copies were also found specificaly inserted in *C. canpehora* or in *C. eugenioides* in the same time interval, suggesting a long period of radiation into two gene pools to give rise to the two species. The analysis of all *Divo* copies (conserved and non-conserved) that inserted between 3.1 and 0.8 Mya in the two diploid ancestors, showed two waves of insertion (two peaks at 1.5–2 Mya and 0.5–1 Mya) that occurred in *C. eugenioides* but not in *C. canephora*, suggesting a divergence in the activity of *Divo* during the process of radiation. Finally the clear amplification of *Divo* observed in *C. canephora* but not in *C. eugenioides* in the time interval of 0 to 0.5 Mya confirmed that the two species were already differentiated.

The relative time of *C. arabica* polyploidization event may be also estimated using the insertion time of conserved *Divo* at orthologous locations in the two sub-genomes. Since the last common *Divo* insertions at orthologous sites between *C. arabica* and *C. eugenioides* and between *C. arabica* and *C. canephora* were observed in the last 0.3 Mya, we concluded that *C. arabica* is originated from a very recent hybridization, confirming previous estimation (Yu et al. 2011). Interestingly, *Divo* copies showing 100% of identity between the two LTRs (nine copies) were only found in the *C. canephora* genome strongly suggesting that *Divo* remains active in that species in a very recent time.

Coffea arabica is an allotetraploid species originated from a hybridization event that occurred between diploid species and taking place 46,000–665,000 years ago (Yu et al. 2011). Understanding the mechanisms of genome modifications during the allotetraploidization may be of interest. *Divo*, a novel family of the *Bianca* lineage among

the superfamily *Copia*, is present in moderated copy numbers in dicots. Complete and potentially functional *Divo* copies were detected in *C. arabica* and its diploid *C. canephora* and *C. eugenioides* progenitors. The activity of the *Divo* family, and the mechanisms of control of its copy number played certainly a role in the differentiation of *C. canephora* and *C. eugenioides* genomes. Beside strucural impacts on genomes, its precise functional role remains to be elucidated. In the near future, a complete characterization of active transposable elements in *C. arabica* and its diploid progenitors will bring more insights into plant genomes divergence and evolution.

Funding Information R.G. was supported by a Special Visiting Scientist grant from the Ciência sem Fronteiras program under the reference ID 84/2013 (Cnpq/CAPES).

Compliance with ethical standards

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ainouche ML, Fortune PM, Salmon A, Parisod C, Grandbastien MA, Fukunaga K, Ricou M, Misset MT (2009) Hybridization, polyploidy and invasion: lessons from Spartina (Poaceae). Biol Invasions 11:1159–1173

Allaire JJ (2012) RStudio: Integrated development environment for R. J Wildl Manage 75:1

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215:403–410

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Gout M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JSS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, Boccara M, Resterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C (2011) The genome of Theobroma cacao. Nat Genet 43:101–109

Audic S, Claverie J (1997) The Significance of Digital Gene Expression Profiles. Genome Res 7:986–995.

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:1–6. doi:10.1186/s13100-015-0041-9

Bennetzen JL, Kellogg E (1997) Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9:1509–1514

Bouharmont J (1959) Recherches sur les affinités chromosomiques dans le genre Coffea. I.N.É.A.C., Montpellier

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou Y, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Kianian S, Gill B, Anderson O, Kersey P, Dvorak J, McCombie R, Hall A, Mayer KFX, Edwards KJ, Bevan M, Hall N (2012) Analysis of the bread wheat genome using whole genome shotgun sequencing. Nature 491:705–710

Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Chen LJ, He Q, Xu Q, Bian C, Zheng Z, Sun F, Liu W, Hsiao YY, Pan ZJ, Hsu CC, Yang YP, Hsu YC, Chuang YC, Dievart A, Dufayard JF, Xu X, Wang JY, Wang J, Xiao XJ, Zhao XM, Du R, Zhang GQ, Wang M, Su YY, Xie GC, Liu GH, Li LQ, Huang LQ, Luo YB, Chen HH, Van de Peer Y, Liu ZJ (2015) The genome sequence of the orchid Phalaenopsis equestris. Nat Am 47:65–76.

Carvalho A (1952) Taxonomia de Coffea arabica L. VI - Caracteres morfologicos dos haploides. Bragantia 12:201–212.

Chaparro C, Guyot R, Zuccolo A, Piégu B, Panaud O (2007) RetrOryza: A database of the rice LTR-retrotransposons. Nucleic Acids Res 35:66–70

Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. Bot J Linn Soc 167:357–377

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G, Aury JM, Bento P, Bernard M, Bocs S, Campa C, Cenci A, Combes MC, Crouzillat D, Da Silva C, Daddiego L, De Bellis F, Dussert S, Garsmeur O, Gayraud T, Guignon V, Jahn K, Jamilloux V, Joët T, Labadie K, Lan T, Leclercq J, Lepelley M, Leroy T, Li LT, Librado P, Lopez L, Muñoz A, Noel B, Pallavicini A, Perrotta G, Poncet V, Pot D, Priyono, Rigoreau M, Rouard M, Rozas J, Tranchant-Dubreuil C, VanBuren R, Zhang Q, Andrade AC, Argout X, Bertrand B, de Kochko A, Graziosi G, Henry RJ, Jayarama, Ming R, Nagai C, Rounsley S, Sankoff D, Giuliano G, Albert VA, Wincker P, Lashermes P (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1180–1184

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in arabidopsis. Genome Res 12:1075–1079

Dias ES, Hatt C, Hamon S, Hamon P, Rigoreau M, Crouzillat D, Carareto CMA, de Kochko A, Guyot R (2015) Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. Plant Mol Biol 89:83–97

Domingues DS, Cruz GMQ, Metcalfe CJ, Nogueira FTS, Vicentini R, Alves CS, Van Sluys MA (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics 13:1–13. doi:10.1186/1471-2164-13-137

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63:584–598

Eickbush TH, Jamburuthugoda VK (2007) The diversity of retrotransposons and the properties of their reverse transcriptases. Mol Cell Biol 134:221–234

Fedoroff NV (2012) Transposable elements, epigenetics, and genome evolution. Science 338:758–767

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368

Fontana A (2010) A hypothesis on the role of transposons. Biosystems 101:187–193

Gilbert C, Peccoud J, Chateigner A, Moumen B, Cordaux R, Herniou EA (2016) Continuous influx of genetic material from host to virus populations. PLoS Genet 12:1–21

Guyot R, Darré T, Dupeyron M, de Kochko A, Hamon S, Couturon E, Crouzillat D, Rigoreau M, Rakotomalala JJ, Raharimalala NE, Akaffou SD, Hamon P (2016) Partial sequencing reveals the transposable element composition of Coffea genomes and provides evidence for distinct evolutionary stories. Mol Genet Genomics 291:1979–1990

Hamon P, Duroy PO, Dubreuil-Tranchant C, Costa PMD, Duret C, Razafinarivo NJ, Couturon E, Hamon S, e Kochko A, Poncet V, Guyot R (2011) Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae). Mol Genet Genomics 285:447–460

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: Repbase-Submitter and Censor. BMC Bioinformatics 7:474

Kolano B, Bednara E, Weiss-Schneeweiss H (2013) Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of *Chenopodium quinoa* (Amaranthaceae). Plant Cell Rep 32:1575–1588

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic fo comparative genomics. Genome Res 19:1639–1645

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 72:181–204

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

Le Grice SFJ (2003) "In the beginning": initiation of minus strand DNA synthesis in retroviruses and LTR-containing retrotransposons. BioChemistry 42:14349–14355

Leng P, Klatte DH, Schumann G, Boeke JD, Steck TL (1998) Skipper, an LTR retrotransposon of Dictyostelium. Nucleic Acids Res 26:2008–2015

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104:520–533

Lin X, Faridi N, Casola C (2016) An ancient transkingdom horizontal transfer of penelopelike retroelements from arthropods to conifers. Genome Biol Evol 8:1252–1266

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

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, Moya A (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70–D74

Louarn J (1976) Hybrides interspécifiques entre Coffea canephora Pierre et C. eugenioides Moore. Café Cacao Thé 20:33–52

Ma J, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. PNAS 101:12404–12410

Marcon HS, Domingues DS, Silva JC, Borges RJ, Matioli FF, Fonter MRM, Marino CL (2015) Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic. BMC Plant Biol 15:198–214

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367

Mehra M, Gangwar I, Shankar R (2015) A deluge of complex repeats: the solanum genome. PLoS One 10:1–38

Minervini CF, Viggiano L, Caizzi R, Marsano RM (2009) Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*. Gene 440:42–49

Mirouze M, Reinders J, Bucher E, Nashimura T, Schneeberger K, Ossowki S, Cao J, Weigel D, Paszkowski J, Mathieu O (2009) Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461:1–5

Nielen S, Vidigal BS, Leal-Bertioli SCM, Ratnaparkhe M, Paterson AH, Garsmeur O, D'Hont A, Guimarães PM, Bertioli DJ (2012) Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the Arachis A–B genome divergence. Mol Genet Genomics 287:21–38

Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:360–363

Panaud O (2016) Horizontal transfers of transposable elements in eukaryotes: The flying genes. Comptes rendus Biol. doi:10.1016/j.crvi.2016.04.013

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA (2010) Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol 186:37–45

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Fletus FA, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556

Peterson-Burch BD, Voytas DF (2002) Genes of the Pseudoviridae (Ty1/copia Retrotransposons). Mol Biol Evol 19:1832–1845

Piednoël M, Carrete-Vega G, Renner SS (2013) Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. Plant J 75:699–709

Romualdi C, Bortoluzzi S, D'Alessi F, Danieli GA (2003) IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. Physiol Genomics 12:159–162

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrel B (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945

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

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

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternk S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlison C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado BMead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lish DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR,

Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1116

Sonnhammer ELL, Durbin R (1996) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:1–10

The Arabica Coffee Genome Consortium (2014) Towards a Better Understanding of the Coffea Arabica Genome Structure. In: Association for Science and Information on Coffee (ed) International Conference on Coffee Science. Cogito, Armenia, pp 42–45

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

The French–Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. doi:10.1038/nature6148

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Nat Acad Sci USA 103:17638–17643.

Wang H, Liu J-S (2008) LTR retrotransposon landscape in Medicago truncatula: more rapid removal than in rice. BMC Genomics 9:382–395

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, 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

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268

Yin H, Du J, Wu J, Wei S, Xu Y, Tao S, Wu J, Zhang S (2015) Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. Sci Rep 5:1–15.

Yu Q, Guyot R, de Kochko A, Byers A, Navajas-Pérez R, Langston BJ, Dubreuil-Tranchant C, Paterson AH, Poncet V, Nagai C, Ming R (2011) Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). Plant J 67:305–317