Terminal-repeat Retrotransposons with GAG domain (TR-GAG) in plant genomes: A new testimony on the complex world of transposable elements

Cristian Chaparro, Thomas Gayraud, Rogerio Fernandes de Souza, Douglas Silva Domingues, Sélastique Doffou Akaffou, Andre Luis Laforga Vanzela, Alexandre de Kochko, Michel Rigoreau, Dominique Crouzillat, Serge Hamon, et al.

Genome Biology and Evolution Advance Access published January 7, 2015 doi:10.1093/gbe/evv001

Terminal-repeat Retrotransposons with GAG domain (TR-GAG) in plant genomes: A new testimony on the complex world of transposable elements.

Authors and affiliations:

Cristian Chaparro [2 co], Thomas Gayraud [1 co], Rogerio Fernandes de Souza [3], Douglas Silva Domingues [4], Sélastique Akaffou [5], Andre Luis Laforga Vanzela [3], Alexandre de Kohcko [1], Michel Rigoreau [6], Dominique Crouzillat [6], Serge Hamon [1], Perla Hamon [1], Romain Guyot [1,7*]

1 Institut de Recherche pour le Développement (IRD), UMR DIADE (CIRAD, IRD, UM2), BP 64501, 34394 Montpellier Cedex 5, France

2 2EI UMR5244 Université de Perpignan Via Domitia, UMR 5244 CNRS Ecologie et Evolution des Interactions (2EI), 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France.

3 Departamento de Biologia Geral, CCB Universidade Estadual de Londrina (UEL), Londrina, PR, Brazil, 86051-990

4 Instituto Agronômico do Paraná, Coordenação de Pesquisa, Laboratório de Biotecnologia Vegetal, 86047-902, Londrina, PR, Brasil

5 Université Jean Lorougnon Guédé, BP 150 Daloa Côte d'Ivoire

6 Nestlé R&D Tours, 101 AV. G. Eiffel, Notre Dame d'Oé, BP 49716 37097, Tours, Cedex 2, France

7 Institut de Recherche pour le Développement (IRD), UMR IPME, BP 64501, 34394 Montpellier Cedex 5, France

*Author for Correspondence: 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: PRJNA242989, KM360147, KM371274, KM371276, KM371277, KM371275

Abstract

A novel structure of non-autonomous LTR retrotransposons called Terminal Repeat with GAG domain (TR-GAG) has been described in plants, both in

monocotyledonous, dicotyledonous and basal angiosperm genomes. TR-GAGs are relatively short elements in length (< 4 kb) showing the typical features of LTR retrotransposons. However, they carry only one Open Reading Frame coding for the GAG precursor protein involved for instance in transposition, the assembly and the packaging of the element into the Virus Like Particle. GAG precursors show similarities with both Copia and Gypsy GAG proteins, suggesting evolutionary relationships of TR-GAG elements with both families. Despite the lack of the enzymatic machinery required for their mobility, strong evidences suggest that TR-GAGs are still active. TR-GAGs represent ubiquitous non-autonomous structures that could be involved in the molecular diversities of plant genomes.

Keywords: Non-autonomous elements, LTR retrotransposons, GAG, conservation in Plant genomes

Introduction

Repeated sequences are the main component of plant genomes, especially those with large C-value. In bread wheat, barley and maize, more than 80% of the sequenced DNA is classified into mobile elements, so called transposable elements (TE) (Schnable, Ware et al. 2009, Wicker, Taudien et al. 2009). TEs were traditionally classified into two main classes according to their lifestyle cycle: Class I, or Retrotransposons, for TEs moving via an RNA intermediate, which use a so called "copy and paste" mechanism, and Class II, or Transposons, for TEs moving via a DNA intermediate, which use a so called "Cut and Paste" mechanisms (Wicker, Sabot et al. 2007). LTR-retrotransposons, that pertain to Class I, are the most abundant TEs identified in plant genomes. The activity of transposable elements has a deep influence on the evolution and function of plant genes and genomes and so contributes to the implementation of molecular diversification and genetic diversity. Their activity is controlled at the transcriptional and post-transcriptional levels by the host. However the high activity of LTR-retrotransposons overtakes occasionally these mechanisms that control TE proliferation leading to sudden accumulation of LTR retrotransposon copies (so called "burst") and, consequently to a rapid genome size increase (Piegu, Guyot et al. 2006).

With the advent of large-scale plant genome sequencing and the advances in TE bioinformatics analysis (Flutre, Duprat et al. 2011), it became clear that most of the TEs identified so far where not able to synthesize the full enzymatic machinery and all the molecules involved in their own mobility and to accomplish their multiplication cycle, disabling their coding capacities, that lead to their inactivation and so counteract their impact on genome size increase (Devos, Brown et al. 2002, Ma, Devos et al. 2004, Vitte and Bennetzen 2006). In some cases, homologous recombination mechanisms occurring between LTR sequences in the same LTR retrotransposon element leads to solo LTR formation implicating the removal of a large internal portion of elements. These altered elements are usually considered as dead elements, which are no longer capable of transcription and mobility.

However, there are reports where elements carrying a defective transposition machinery can get "back to life" and meet again the ability to move and to multiply their copy numbers in the host genome (Kalendar, Vicient et al. 2004)(Witte, Le et al. 2001)(Tanskanem, Sabot et al. 2007). Such elements, often called non-autonomous

elements, are supposed to mobilize via a cross activation (*in trans*) with autonomous and functional partners. This interaction requires that non-autonomous elements still carry recognition domains for proteins encoded by autonomous partners (Wicker, Sabot et al. 2007, Schulman 2013). Two groups of Class I non-autonomous LTR-retrotransposons were identified in numerous plant genomes: TRIM (Terminal-repeat Retrotransposons in Miniature) (Witte, Le et al. 2001), and LARD (Large Retrotransposon Derivative) (Kalendar, Vicient et al. 2004) (Figure 1). TRIMs and LARDs are respectively short (< 2 Kb) and long (> 4 Kb) elements that although they have lost their internal coding regions, they are involved in restructuring plant genomes (Witte, Le et al. 2001, Kalendar, Tanskanen et al. 2008). BARE-2 is another type of active non-autonomous elements found in Barley (Tanskanem, Sabot et al. 2007). BARE-2, that lacks the GAG domain, involved in the packaging of the element into the Virus Like Particule (VLP), remains mobile using the functional GAG capsid protein encoded by the BARE-1 autonomous elements (Tanskanem, Sabot et al. 2007). BARE-2 elements represent the unique described case of cis-parasitims of a LTR-retrotransposons in plants. However, the BARE-2 non-autonomous structure was investigated only in Triticeae genomes (Vicient, Kalendar et al. 2005). The profusion of LTR-retrotransposons within plant genomes, the abundance of structural variation of defective elements and the recent discovery of non-autonomous elements raise the question to know if the whole structural variety of non autonomous LTR-retrotransposons have been really identified or if novel structures remain to be charaterized.

In an attempt to characterize the whole set of mobile elements within the *Coffea* genomes, especially in *C. canephora* (Denoeud et al., 2014), we report here a new group of non-autonomous LTR-retrotransposons, called TR-GAG (Terminal-repeat Retrotransposons with GAG domain) in plants. TR-GAG elements are short LTR-retrotransposons (< 4 Kbp) carrying a unique Open Reading Frame (ORF) coding for a GAG capsid protein. In *C. canephora* genome, five families of TR-GAG elements were described. These elements are expressed and their evolutionary dynamics in the *Coffea* genus indicated different pathways in the copy number variations. Similar structures were found in numerous available sequenced eudicotyledoneous, monocotyledoneous and algae genomes, indicating that TR-GAG elements could be ubiquitous TEs in plants.

Results

Annotation and identification of the Non-autonomous LTR retrotransposons TRIM-1 family in the Coffea canephora genome

We used the draft genome sequence of the *Coffea canephora* accession DH 200-94 to annotate transposable elements (Denoeud et al., 2014; http://coffee-genome.org). We first performed a manual annotation of transposable elements using the 10 largest scaffolds from C. canephora genome sequencing project (accounting for 65,698,623 bp, scaffold1 to 10), and an initial database of 948 TEs was produced (Guyot et al. unpublished data). Among the 948 elements, eleven conserved short elements (< 3 Kbp) harboring a typical LTR-retrotransposons structure (two duplicated regions starting by TG, and finishing by CA, flanked by a target site duplication of 5 bp, and a PPT located upstream the 3' duplicated region) were identified using similarity searches (BLASTn). After initial analyses, we found two sequence groups with different lengths. Short sequences (~1,700 bp) were called TRIM-1-S have the typical structure of TRIM (Witte, Le et al. 2001) while long sequences (~2,500 bp) were called TR-GAG1 (Terminal Repeat with GAG domain). They are similar to the TRIM but carry an internal region similar to LTR-retrotransposons GAG capsid domain (Figure 2A and 2B). The last structure was not previously described in plant genomes. The two groups of sequences are conserved except for the presence of the GAG domain in TR-GAG1 (Figure 2C and supplemental data 3). Multiple alignment of the LTR sequences from the TRIM-1-S and TR-GAG1 elements show an overall strong conservation between the two groups as well the presence of a putative TATA box that could intervene in the initiation of the elements' transcriptions (Supplemental data 3). An exhaustive search against the *Coffea canephora* draft genome (568 Mbp) indicated the presence of 71 and 60 complete copies of TRIM-1-S and TR-GAG1, respectively. All complete dispersed copies within the chromosomes with conserved LTR extremities, showed different insertion sites (Supplemental data 4). The complete elements are flanked by 5 bp direct repeats usually generated during the LTR-retrotransposon insertions, suggesting that they are originated from different replications events. Using BLASTN algorithm, we searched in the *C. canephora* genome for autonomous elements

sharing high nucleotide conservation with TRIM-1-S and TR-GAG1, but we did not find any autonomous full-length elements in the available genomic sequences.

Detailed analysis of the TR-GAG1 elements

We detailed the structure of the TR_GAG1 elements (Copy found in *C. canephora* draft genome located on "Chr 0", positions 113,020,990-113,023,502, accession KM360147), since such conserved structure of non-autonomous LTR-retrotransposon was not described yet. TR-GAG1 elements have LTR lengths of ~485 bp. The 5' LTR is flanked downstream by a Primer Binding Site (PBS) complementary to the Leucine tRNA and the 3' LTR is flanked upstream by a PolyPurine Tract (PPT) 5'-AAAAGGCAAATGGAG -3' (Figure 3). Beside LTR regions, no internal duplicated region was found in the TR-GAG1 sequence. The inner region is composed of an ORF of 433 amino acids with strong similarities with GAG (group specific antigens) and more particularly with the UBN2 family domain from Pfam (gag-polypeptide of LTR Copia superfamily). The small structural motif of Zinc finger (Zf-C2HC) is also found at the amino-acid residue 275 the ORF (position 1245-1286 along the full-length TR-GAG nucleotide sequence). At the C terminal part, few similarities were observed with Aspartic proteases from the GyDB but no motif was conserved in Pfam database (Punta, Coggill et al. 2012). The UBN2 Pfam domain (PF14223) from TR-GAG1 is described as associated with Copia Superfamily of complete LTR retrotransposons (http://pfam.xfam.org). No significant RNA secondary structure was found with the putative leader sequence of TR-GAG1, suggesting either absent or labile PSI (Packaging Signal) and DIS (Dimerization Signal) motifs. These motifs were identified in Retroviruses and are involved in the packaging and RNA dimerization (Tanskanem, Sabot et al. 2007).

Transcriptional responses of the TRIM1/TR-GAG family

We analyzed the transcriptional pattern of TRIM-1-S and TR-GAG1 elements in three coffee species. Specific primers were selected in TRIM-1-S and TR-GAG1 to amplify the inner regions. For TR-GAG1, primers amplify a 328 bp product from the GAG precursor. RT-PCR analyses indicate the presence of transcripts for TRIM-1-S and TR-GAG1 originating from mRNA leaves, suggesting that elements are expressed in *C. canephora*, *C. eugenioides* and *C. arabica* (Supplemental data 4 A and B). RNA-seq analysis using 130 millions of Illumina reads shows that 38 complete copies of

TR-GAG1 are expressed at low level in vegetative tissues (leaves and roots) while no or few expression were detected in reproductive tissues (pistil and stamen)(Supplemental data 5).

Characterization of TR-GAG families in *C. canephora*

We searched the presence of other TR-GAG families in the draft genome of *C. canephora*. We used first the results of LTR_STRUC prediction of LTR-retrotransposons. The 1,799 putative LTR-retrotransposons predicted by LTR_STRUC were filtered out according to the features identified for TR-GAG1. Beside an overall structure of elements, such as presence of LTR, PBS and PPT regions, sequences with a maximum length of 4 Kbp, a minimum redundancy of two copies, and with similarities for GAG Capsid proteins but not with aspartic protease, integrase, reverse transcriptase and RNAse H domains were selected for further analysis. On 1,799 predicted elements, 130 were retained. Sequences were compared against themselves using Dot-plot alignments (Figure 4, A.). Sequences were clustered into five groups of sequences according to their similarities and classified into five different families (called TR-GAG1 to TR-GAG5). One family called TR-GAG2, which exhibited a large number of conserved predicted structures (110 elements) as observed in dot-plot and alignment analysis (Supplemental data 6), was analyzed further (Figure 4, A and B). Among the 110 conserved predicted elements, we selected one copy for detailed analysis (located on pseudo-chromosome 4 21,003,142-21,006,851). This element presented an overall similar structure to TR-GAG1 (Figure 4, C). TR-GAG3, TR-GAG4 and TR-GAG5 families were analyzed and also shown a typical structure of TR-GAG non-autonomous elements (Supplemental data 8). While TR-GAG2 shares similarities with the same *Copia* GAG Pfam domain family (UBN2) with TR-GAG1, TR-GAG3 and TR-GAG4 contain the Retrotrans_gag motif (Pfam PF03732) that appears associated with annotated *Copia* and *Gypsy* polyproteins in Uniprot database (http://www.uniprot.org). Phylogenetic analysis with reference GAG domains from GyDB confirmed the similarity of TR-GAG1 and TR-GAG2 GAG domains with Copia and TR-GAG4 with Gypsy sub-family GAG domains (Supplemental data 7). All five TR-GAG families were analyzed using RNA-seq. We observed different pattern of expression according to the four tissues analyzed: leaf, root, pistil and stamen (Supplemental data 5).

Distribution and Copy number estimation of TR-GAG elements in the *Coffea* genus

We first investigate the copy number of the five identified TR-GAG families in the *C. canephora* sequenced genome (Supplemental data 9). Complete copies of TR-GAG1 and TR-GAG2, as defined by 80% of nucleotide identity over 100 % of the reference element length, were used to estimate their insertion times (Supplemental data 10). Our analysis indicates a relatively recent increase of TR-GAG2 elements (highest peak at 0.5-1 million years ago).

The distribution of the five identified TR-GAG families along the reconstructed pseudo-chromosomes in *C. canephora* were also studied. Copies (with two level of conservation: 80-80 and 70-70), solo LTRs and fragmented copies were identified from the *C. canephora* draft genome sequence (Supplemental data 11).

In order to investigate the evolution of TR-GAG families, we used *in silico* approaches to search for its presence in the *Coffea* genus. Nine additional *Coffea* species (including *Coffea horsfieldiana* (ex-*Psilanthus horsfieldiana)*) and an out-group in the Rubiaceae family: *Craterispermum kribi* from Cameroon, were surveyed using a high-throughput 454 sequencing analysis. The *Craterispermum* genus, belonging to the Rubioideae sub-family, diverged early from the *Coffea* genus (Ixoroideae sub-family), about 80 Mya (Bremer and Eriksson 2009).

The 454 sequences (Table 2) were firstly used to survey the presence of highly conserved reads of TR-GAG, using the criteria of 80% minimum nucleotide identity with over 80% of the read length. Sequences fitting these criteria show a large variation of reads for the TR-GAG2 family in *Coffea* and *Craterispermum kribi* genomes. Additionally, with this approach we could estimate the copy number of TR-GAG elements in several genomes. Using these conserved reads, TR-GAG was estimated to range from 0 to 696.7 copies in diploid species and from 10.2 to 1,168.7 copies in *C. arabica*. However, in almost all cases (at the exception of *Craterispermum* and *C. tetragona*), the highest copy numbers were obtained for TR-GAG-2. Only few copies (respectively 5 and 7 copies) of TR-GAG-2 and TR-GAG-1 were detected for the *Craterispermum* outgroup (Rubiaceae) (Supplemental data 11). The TR-GAG-2 family contributes to the genome size of diploid species, but with a relatively weak intensity (Supplemental data 12). However the genome size contribution of TR-GAG-2 appears to decrease in species going from West to East in species belonging to Eucoffea (*C. canephora, C. heterocalyx, C. eugenioides*, *C.*

arabica), Mozambicoffea (*C. pseudozanguebarie, C. racemosa*) and Mascarocoffea (*C. humblotiana, C. millotii ex-dolichophylla and C. tetragona*). The Indonesian *Coffea horsfieldiana* appears intermediate between Eucoffea and Mozambicoffea or Mascarocoffea botanical groups. Only traces of TR-GAG2 and TR-GAG1 were detected in *Craterispermum* (Rubiaceae).

Characterization of TR-GAG families in genomes using LTR_STRUC algorithm

We searched TR-GAG element structures in 33 available plant genomes. In total more than 18 Gbp of genomic sequences were processed with LTR_STRUC and a total of 38,772 predicted LTR-retrotransposons were found (Supplemental data 13). After filtering, a total of 373 candidates were found distributed among 23 different monocotyledonous and dicotyledonous plant genomes (Figure 5). Detailed analysis of candidates TR-GAG elements confirmed the structures previously discovered in the *C. canephora* genome.

Detection of TR-GAG families in genomes using HMM

In order to validate the detection of TR-GAG by LTR_STRUC, we developed Hidden Markov Models (HMM) to recognize GAG motifs (retrotrans_gag, UBN2, UBN22, UBN23) surrounded by direct repeats. The new model was used in Banana (*Musa acuminata*, angiosperm, monocots), Cacao (*Theobroma cacao*, angiosperm, dicots), coffee (*C. canephora*, angiosperm, dicots), Ectocarpus (*Ectocarpus siliculosis*, brown algae;(Cock, Coelho et al. 2010)), Chondrus (*Chondrus crispus*, red algae;(Collen, Porcel et al. 2013)) and Drosophila (*Drosophila melanogaster*, insect) genomes. While TR-GAG elements were found in all angiosperm and brown algae genomes, no potential candidate was predicted in red algae and Drosophila genomes. Twenty-five TR-GAG families were detected for Banana and one of then shows a high copy number (~700 copies, Supplemental data 14). In brown algae (*Ectocarpus)*, the presence of one TR-GAG like sequence, were previously reported (Cock, Coelho et al. 2010, in Supplemental material). Using our detection approach, 4 TR-GAG families were finally predicted in this genome (Cock, Coelho et al. 2010, in Supplemental material).

Discussion

The identification and classification of the whole spectrum of LTR-retrotransposon structures is particularly a complex process in plant genomes due to the huge number and variety of defective LTR-retrotransposon structures. Although most of the defective structures, deriving from a wide variety of rearrangement mechanisms, lead to inactive elements, some of them remain mobile like TRIM, LARD and BARE2 elements (Witte, Le et al. 2001, Kalendar, Vicient et al. 2004, Tanskanem, Sabot et al. 2007). These known non-autonomous LTR-retrotransposon structures redefined our view of the definition of what is really an active element in genomes, and raised new questions about their precise classification and their mechanisms of mobility. The discovery of such exceptional diversity of non-autonomous structures opened the door to the large-scale *in silico* exploitation of plant genome sequences to seek novel non-autonomous structures.

The novel element called TR-GAG belongs to such type of non-autonomous structures and brings new insight on transposable element and genome evolution. TR-GAG elements clearly belong to LTR-Retrotransposons order of transposable elements (Wicker, Sabot et al. 2007). TR-GAG can be identified using *de novo* LTR-retrotransposons finding programs like LTR_STRUC (McCarthy and McDonald 2003), since they share key structural features with them, like LTR domains, PPT and PBS motifs and a 5-bp TSD (Target Site Duplication) at their insertion sites in the host genome. TR-GAGs appear generally smaller (< 4 kb) than typical full-length Copia and Gypsy LTR-Retrotransposons (5 to 20 kb) in plants. Several signs suggest that TR-GAGs are active elements in *Coffea* species in spite of the absence of an internal polyprotein domain: (i) RT-PCR and RNA-seq data show the transcription of TR-GAG families. While TR-GAG1 is mainly expressed at a low level in vegetative tissues, other families (TR-GAG2 and TR-GAG3) show a significant expression in reproductive tissues suggesting that new insertions could be vertically transmitted to the progeny; (ii) the copy number of TR-GAG elements in *C. canephora* and the different TSD motifs found for each copy suggests an amplification mechanism that can be achieved by the lifestyle cycle of mobile LTR-Retrotransposons; (iii) the high conservation of sequence and structure between each TR-GAG copy in the *C. canephora* genome; (iv) their insertion time patterns.

TR-GAG elements lack a polyprotein domain involved in the mobility, but carry a GAG precursor, which is usually processes by protease into protein subunits (Matrix, Capsid and Nucleocapsid) (Freed 1998). This structure is the strict opposite of the

described BARE-2 non-autonomous elements in barley, lacking only the GAG domain. It remains mobile as a two-components system: A non-autonomous elements (BARE2) and an autonomous counterpart (BARE-1) providing by complementation-like a functional GAG precursor (Tanskanem, Sabot et al. 2007). For TR-GAG1 elements, no full-length autonomous element similar to the TR-GAG1 sequence was found in the draft genome sequence of *C. canephora*, suggesting that either the mobility of TR-GAG1 is driven *in trans* by a compatible but different full-length autonomous elements, or the complete element appears as absent due to incompleteness of the sequenced genome or it has been specifically lost in the studied and sequenced genotype. The presence of a functional GAG precursor in TR-GAG elements also raises the question to know their potential role in the cycle of other LTR retrotransposon elements. The capsid (CA) and nucleocapsid (NC) protein subunits of GAG precursors are respectively implicated in the transposition and in the assembly packaging, reverse transcription and integration mechanisms. More generally GAG proteins appears to be able to engage interactions with a wide spectrum of molecules such as proteins, DNA, RNA and lipids (Freed 1998).

The GAG peptides encoded by TR-GAG elements may drive *in trans* the mobility of a variety of other LTR retrotransposons that lack functional GAG domain similarly to the BARE2. Additional molecular experimental data will be required to precisely understand the function of GAG domain in TR-GAG elements.

Five different families of TR-GAG were identified in *C. canephora.* They carry GAG domains that show similarities with both Copia and Gypsy superfamily related GAG domains suggesting that TR-GAG structures have been generated with a frequent and common mechanism for all LTR retrotransposons super-families certainly involving unequal recombination events (Ma, Devos et al. 2004). In *C. canephora*, all five TR-GAG families show different complete, fragmented and solo LTR copy numbers, suggesting distinct levels of proliferation control by the host genome. Interestingly, TR-GAG2 that shows the highest copy number, are non-randomly distributed along the *C. canephora* pseudo-molecules and targets preferentially TE - rich regions.

The TR-GAG2 family shows high variation in copy number among the ten *Coffea* species we analyzed. These variations are in agreement with the three botanical sections (or groups) defined by Chevalier (Chevalier 1942), strongly suggesting that TR-GAG2 copy number proliferation is associated with the evolution of botanical

groups of *Coffea*. These botanical sections correspond also to genetically differentiated groups as obvious from fertility of FI inter-specific hybrids (Louarn 1993), mean genome sizes (Noirot, Poncet et al. 2003)(Razafinarivo, Rakotomalala et al. 2012) and from genetic diversity revealed by SSR markers (Razafinarivo, Guyot et al. 2013).

Finally, the presence of TR-GAG structures in twenty-three different plant genomes from dicotyledonous and monocotyledonous species, as well as in basal Angiosperms (*Amborella*) and one algae species, indicate that these elements are ubiquitous mobile elements. Comparisons between all predicted TR-GAG elements in plants (Figure 5) show the absence of conservation between species suggesting that TR-GAG elements were originated from distinct pool of full-length autonomous LTR-retrotransposons. The notable exception is the conservation of one TR-GAG family between *Cicer arietinum* and *Lotus japonicus* genomes (Figure 5). Such significant conservation of transposable elements over different plant families suggests that TR-GAG elements could also be subjected to events of horizontal transfer like LTR-retrotransposons (Fortune, Roulin et al. 2008, Roulin, Piegu et al. 2008, Roulin, Piegu et al. 2009).

Conclusions.

In conclusion, TR-GAG elements are a new non-autonomous element ubiquitous in plant genomes. TR-GAG elements are potentially active indicating they are associated to functional full-length LTR retrotransposons to achieve their life cycle. Considering their significant copy numbers TR-GAG elements could play an important role in chromosome structure, alteration of coding region expression and genome evolution in plants.

Material and methods

Plant material, DNA and RNA preparation

Three coffee species were used in our analyses: *C. arabica* (accessions AR52 and ET39), *C. eugenioides* (accession DA71) and *C. canephora* (accessions BA58, BB60, BD69 and DH 200-94). All plants were growing in the greenhouses at the IRD center, Montpellier (France). Leaves were harvested and stored at -80 °C prior to DNA extraction, using Qiagen DNeasy Plant Mini extraction kits. Quantity and quality of DNA was measured using a Nanodrop (ND-1000). RNA preparations were

obtained from leaves of *C. arabica* (accessions ET39), *C. eugenioides* (accession DA71) and *C. canephora* (accessions DH 200-94), using the SV Total RNA Isolation System (Promega).

Identification, classification and annotation of LTR-retrotransposons

A manual annotation procedure was undertaken on 17 publicly available *C. canephora* and *C. arabica* BAC sequences (accounting for 3,023,472 bp) and from the ten largest *C. canephora* scaffolds (accounting for 65,698,623 bp, from the *C. canephora* draft genome generated by the Coffee Genome Consortium) to build an initial database. A total of 948 elements were finally annotated as follows and classified according to the universal classification of TEs (Wicker, Sabot et al. 2007): 516 transposons (DTX), 7 helitrons (DHX), 14 LINE (RIX), 330 LTR-retrotransposons (RLX), 1 Retrovirus (RTX), 61 SINE (RSX) and 19 Unclassified (XXX, noCat). This manually curated database was enriched by a *de novo* detection of LTR-retrotransposons using the LTR_STRUC algorithm (McCarthy and McDonald 2003) against 568 Mbp of the *Coffea canephora* draft genome (Coffee genome project; http://coffee-genome.org; Denoeud et al., 2014). A total of 1,799 full-length LTR-retrotransposons were detected from *Coffea canephora* scaffolds with a size larger than 5 Kbp. This dataset was classified into *Gypsy* (RLG) and *Copia* (RLC) according to their similarity matches against the GyDB domain libraries (http://www.gydb.org/index.php/Main_Page) (Llorens, Futami et al. 2011). Sequences were classified into the RXX category if no conserved domains were found or if only a GAG domain was identified. The LTR_STRUC dataset was composed of 745 RXX (41%), 580 RLG (32%) and 474 RLC (26%).

***In silico* characterization of non-autonomous elements**

The identification of complete, and fragmented copies of elements was done using Censor (Kohany, Gentles et al. 2006) against the 568 Mbp of the *Coffea canephora* draft genome. A complete copy is considered if it covers a minimum of 80% of the reference sequence with a minimum of 80% of nucleotide identity, a distantly complete copy is considered if it covers a minimum of 70% of the reference sequence with a minimum of 70% of nucleotide identity. The genomic distribution of

elements was plotted using CIRCOS (http://circos.ca). The insertion sites of complete copies were identified using the best-conserved sequence considered as reference to extract complete copies with 100% of coverage against the reference sequences. Sequence of ten bp downstream and upstream the insertion sites were extracted and analyzed using WebLogo (http://weblogo.berkeley.edu/logo.cgi).

Characterization of TR-GAG families in *C. canephora draft genome*

Raw results from LTR_STRUC were filtered to retrieve putative TR-GAG families, according to the following parameters; (1) a maximum length of 4 Kbp for each predicted element, (2) Similarity (e-value < $10e^{-4}$ on BLASTx) with only the GAG capsid domains downloaded from the GyDB database (http://www.gydb.org/index.php/Main_Page), and (3) a redundancy of a minimum of two copies within the genome. Sequence of TR-GAGs were submitted to GenBank: TR-GAG1: KM360147, TR-GAG2: KM371274, TR-GAG3: KM371276, TR-GAG4: KM371277, TR-GAG5: KM371275.

Estimation of TR-GAG copy number using 454 sequencing survey

One plate of 454 Pyrosequencing (GS Junior System Roche) was performed for each *Coffea* species classified early by (Chevalier 1942) into Eucoffea such as: two *C. canephora* Pierre ex A.Froehner accessions (DH200-94 from Congo Democratic Republic and BUD15 from Uganda), *C. heterocalyx* Stoff. (JC62) from Cameroon, *C. arabica* L. (ET39) from Ethiopia, *C. eugenioides* S. Moore (DA59) from Kenya, Mozambicoffea such as *C. pseudozanguebarie* Bridson (H52) from Kenya, *C. racemosa* Lour. (IA56) from Mozambique, Mascarocoffea such as *C. humblotiana* Baill. (A230) from Comoro Islands, *C. tetragona* Jum. & H.Perrier (A252) and *C. dolichophylla* J.-F.Leroy (A206) from Madagascar (Supplemental data 1) and *Coffea horsfieldiana* (Miq.) J.-F. Leroy from Indonesia, formerly classified as *Psilanthus* and recently placed into *Coffea* (Davis, Tosh et al. 2011), and *Craterispermum* Sp. *Novo kribi* (Rubiaceae) from Cameroon. The cultivars and accessions used grow in the IRD greenhouses (Montpellier, France) and FOFIFA research station (Kianjavato, Madagascar).

Total genomic DNA was extracted from young leaves using the Qiagen DNeasy Plant Mini Kit following the manufacturer protocol. The library construction and NGS sequencing were performed at Nestlé R&D laboratory according to the Roche/454 Life Sciences Sequencing protocol. In total, 1,624,178 sequences were generated accounting for 678 Mbp. Data were submitted to GenBank, BioProject PRJNA242989.

BLASTn searches were carried out with the five TR-GAG families found previously in the *C. canephora* genome. Reads with more than 80% of nucleotide identity with the reference sequence over a minimum 80% of the read lengths were considered as potential fragments of the element. Cumulative lengths of aligned reads were used to extrapolate the contribution of the element to each genome size investigated. For each element family, the potential number of full-length copies is estimated by the division of the estimated size of total members of the element in the genome by the reference sequence length.

Characterization of TR-GAG families in 33 plant genomes

LTR_STRUC (McCarthy and McDonald 2003) was used to predict LTR-retrotranposons in 33 available plant genomes retrieved from specific sites and the Phytozome web site (http://www.phytozome.net; Supplemental data 2) as follow: 24 dicotyledonous genomes - *Nicotiana sylvestris*, *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Mimulus guttatus*, *Uticularia gibba* (bladderwort), *Vitis vinifera* (grape), *Cucumis sativus, Citrullus lanatus* (watermelon), *Fragaria vesca* (strawberry), *Prunus persica* (peach), *Malus domestica* (apple), *Medicago truncatula, Cicer arietinum* (chickpea), *Lotus japonicus, Glycine max* (soybean), *Phaseolus vulgaris* (common bean) *Populus trichocarpa* (poplar), *Manihot esculenta* (cassava), *Ricinus communis, Theobroma cacao* (cacao), *Carica papaya* (papaya), *Arabidopsis thaliana, Brassica rapa* (rapeseed), and *Citrus clementina* (clementine); 7 monocotyledonous genomes - *Phoenix dactylifera* (date palm), *Elaeis oleifera* (oil palm), *Musa acuminata* (banana), *Zea mays* (maize), *Sorghum bicolor* (sorghum), *Brachypodium distachyon* (false brome), and *Oryza sativa* (rice), and two other genomes: *Amborella trichopoda* (angiosperm) and *Selaginella moellendorffii* (non-angiosperm). A total of 18.9 Gbp of sequence was downloaded, processed with LTR_STRUC, and filtered out as described above.

Search for TR-GAG pattern in genomes

We developed an algorithm to automatically detect TR-GAG elements in genomes. The algorithm consists in translating the 6 frames for every "pseudo-molecule" present in the target genome, followed by a search for HMM motifs using the hmmer package (http://hmmer.org). The Retrotrans_gag, UBN2, UBN2_2 and UBN2_3 motifs were used to detect GAG protein signatures. Flanking regions of 5Kb are extracted for all hits with E-value < $1e^{-5}$ and direct repeats greater than 200 bases are searched by dividing the sequence in two and using BLASTn alignment. The region including the direct repeats and the GAG motif are extracted, translated and searched for reverse transcriptase motifs and only the candidates that present no *Copia* or *Gypsy* reverse transcriptase motifs are retained. These candidates are further filtered by size, keeping those sequences between 1 and 6 Kbp while redundant candidates are eliminated.

Transcriptional analysis of the TRIM-1-S and TR-GAG1 elements by RT-PCR

RT-PCR was done using cDNA from *C. arabica* (ET39), *C. eugenioides* (DA71) and *C. canephora* (DH 200-94). cDNA was synthetized from 250 ng of total RNA using the ImProm-II Reverse transcription System Kit (Promega). Primers were selected using Primer3 (http://frodo.wi.mit.edu) on TR-GAG1 and TRIM-1-S sequences (Table 1). PCR were performed in a final volume of 20 µL as follow: 0.5 µL of dNTP (10 nM), 1 µL of each primer (10 mM), 0.2 µL of Taq polymerase (GoTaq, Promega), 4 µL of buffer and 2 µL of cDNA. We used the following PCR amplification cycle: 98°C 5 min; and three steps (98°C 30 sec, 55°C 30 sec, 72°C 30 sec) repeated 35 times followed by a final elongation step (72°C 5 min).

Transcriptional analysis of TR-GAG elements using RNAseq

RNA-seq data generated under the *C. canephora* genome project (coffee-genome.org) from leaves, roots (*C. canephora* accession T3518), stamen and pistil (*C. canephora* accession BP961) were used to identify the transcriptional pattern of reference sequences (http://coffee-genome.org; Denoeud et al., 2014). Nearly 130 millions of Illumina reads (2 x 100 bp) were cleaned using prinseq (Schmieder and

Edwards 2011) and mapped against reference TR-GAG sequences using bowtie 2 (Langmead and Salzberg 2012). Number of mapped reads per reference sequence was processed and RPKM (Reads Per Kilo base per Million) was calculated. Differential expression among RNA-seq libraries were detected from variation of mapped reads and all sequenced reads using Winflat (Audic and Claverie 1997).

Phylogenetic analyses and TR-GAG insertion times

The classification of GAG domains from TR-GAG elements found in the Coffee genome was confirmed by phylogenetic analyses. GAG domains were first identified by similarity against the GAG domains from the Gypsy Database 2.0 (290 domains as in August 2014), extracted from the nucleotide sequence of TR-GAG, and translated into amino acids. Amino acid sequences (with a minimum of 200 residues) were aligned (ClustalW) to construct a bootstrapped neighbour-joining tree, edited with FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

The insertion times of full-length copies, as defined by a minimum of 80% of nucleotide identity over 100% of the reference element length, were dated. Timing of insertion was based on the divergence of the 5' and 3' LTR sequences of each copy. The two LTRs were aligned using stretcher (EMBOSS), and the divergence was calculated using the Kimura 2-parameter method implemented in distmat (EMBOSS). The insertion dates were estimated using an average base substitution rate of 1.3E-8 (Ma and Bennetzen 2004).

Acknowledgements

This research was supported by Agropolis Fondation through the "Investissement d'avenir" program (ANR-10-LABX-0001-01) under the reference ID 1102-006 (Retro-cof). CAPES, through the "CAPES/Agropolis" program, partially funded the work, supported RFS post-doc fellowships and ALLV and DSD working missions. RG is also supported by a « Special Visiting Scientist » grant from the « Ciência sem Fronteiras » program under the reference ID 84/2013 (Cnpq/CAPES). We thank Philippe Lashermes and the Coffee Genome Consortium for the availability of the *Coffea canephora* genome sequence and the South Green Bioinformatics Platform, (www.southgreen.fr) for providing computational resources.

Literature Cited

Audic, S. and J. M. Claverie (1997). "The significance of digital gene expression profiles." Genome Res **7**(10): 986-995.

Chevalier, A. (1942). "Les caféiers du globe, fasc II. Iconographie des caféiers sauvages et cultivés et des Rubiacées prises pour des caféiers." Lechevallier P, ed. Encyclopédie biologique, Tome XXII. Paris.

Cock, J. M., S. M. Coelho, C. Brownlee and A. R. Taylor (2010). "The Ectocarpus genome sequence: insights into brown algal biology and the evolutionary diversity of the eukaryotes." New Phytol **188**(1): 1-4.

Collen, J., B. Porcel, W. Carre, S. G. Ball, C. Chaparro, T. Tonon, T. Barbeyron, G. Michel, B. Noel, K. Valentin, M. Elias, F. Artiguenave, A. Arun, J. M. Aury, J. F. Barbosa-Neto, J. H. Bothwell, F. Y. Bouget, L. Brillet, F. Cabello-Hurtado, S. Capella-Gutierrez, B. Charrier, L. Cladiere, J. M. Cock, S. M. Coelho, C. Colleoni, M. Czjzek, C. Da Silva, L. Delage, F. Denoeud, P. Deschamps, S. M. Dittami, T. Gabaldon, C. M. Gachon, A. Groisillier, C. Herve, K. Jabbari, M. Katinka, B. Kloareg, N. Kowalczyk, K. Labadie, C. Leblanc, P. J. Lopez, D. H. McLachlan, L. Meslet-Cladiere, A. Moustafa, Z. Nehr, P. Nyvall Collen, O. Panaud, F. Partensky, J. Poulain, S. A. Rensing, S. Rousvoal, G. Samson, A. Symeonidi, J. Weissenbach, A. Zambounis, P. Wincker and C. Boyen (2013). "Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida." Proc Natl Acad Sci U S A **110**(13): 5247-5252.

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**(6201):1181-4

Devos, K. M., J. K. Brown and J. L. Bennetzen (2002). "Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis." Genome Res **12**(7): 1075-1079.

Flutre, T., E. Duprat, C. Feuillet and H. Quesneville (2011). "Considering transposable element diversification in de novo annotation approaches." PLoS One **6**(1): e16526.

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

Freed, E. O. (1998). "HIV-1 Gag Proteins: Diverse Functions in the Virus Life Cycle." Virology **251**: 1-15.

Guyot, R., M. de la Mare, V. Viader, P. Hamon, O. Coriton, J. Bustamante-Porras, V. Poncet, C. Campa, S. Hamon and A. de Kochko (2009). "Microcollinearity in an ethylene receptor coding gene region of the Coffea canephora genome is extensively conserved with Vitis vinifera and other distant dicotyledonous sequenced genomes." BMC Plant Biol **9**(1): 22.

Kalendar, R., J. Tanskanen, W. Chang, K. Antonius, H. Sela, O. Peleg and A. H. Schulman (2008). "Cassandra retrotransposons carry independently transcribed 5S RNA." Proc Natl Acad Sci U S A **105**(15): 5833-5838.

Kalendar, R., C. M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy and A. Schulman (2004). "Large Retrotransposon Derivatives: Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes." Genetics **166**: 1437–1450

Kohany, O., A. J. Gentles, L. Hankus and J. Jurka (2006). "Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor." BMC Bioinformatics **7**: 474.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Llorens, C., R. Futami, L. Covelli, L. Dominguez-Escriba, J. M. Viu, D. Tamarit, J. Aguilar-Rodriguez, M. Vicente-Ripolles, G. Fuster, G. P. Bernet, F. Maumus, A. Munoz-Pomer, J. M. Sempere, A. Latorre and A. Moya (2011). "The Gypsy Database (GyDB) of mobile genetic elements: release 2.0." Nucleic Acids Res **39**(Database issue): D70-74.

Louarn, J. (1993). "Structure génétique des caféiers Africains diploïdes basée sur la fertilité des hybrides interspécifiques. ." ASIC, 15e Colloque, Montpellier, 1993.

Ma, J and . L. Bennetzen (2004). "Rapid recent growth and divergence of rice nuclear genomes." *Proc Natl Acad Sci USA* 101:12404–12410.

Ma, J., K. M. Devos and J. L. Bennetzen (2004). "Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice." Genome Res **14**(5): 860-869.

McCarthy, E. M. and J. F. McDonald (2003). "LTR_STRUC: a novel search and identification program for LTR retrotransposons." Bioinformatics **19**(3): 362-367.

Noirot, M., V. Poncet, P. Barre, P. Hamon, S. Hamon and A. De Kochko (2003). "Genome size variations in diploid African Coffea species." Ann Bot (Lond) **92**(5): 709-714.

Noirot, M., V. Poncet, P. Barre, P. Hamon, S. Hamon and A. De Kochko (2003). "Genome size variations in diploid African Coffea species." Annals of Botany **92**(5): 709-714.

Piegu, B., R. Guyot, N. Picault, A. Roulin, A. Saniyal, H. Kim, K. Collura, D. S. Brar, S. Jackson, R. A. Wing and O. Panaud (2006). "Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice." Genome Res **16**(10): 1262-1269.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn (2012). "The Pfam protein families database." Nucleic Acids Res **40**(Database issue): D290-301.

Razafinarivo, N. J., R. Guyot, A. P. Davis, E. Couturon, S. Hamon, D. Crouzillat, M. Rigoreau, C. Dubreuil-Tranchant, V. Poncet, A. De Kochko, J. J. Rakotomalala and P. Hamon (2013). "Genetic structure and diversity of coffee (Coffea) across Africa and the Indian Ocean islands revealed using microsatellites." Ann Bot **111**(2): 229-248.

Razafinarivo, N. J., J.-J. Rakotomalala, S. C. Brown, M. Bourge, S. Hamon, A. de Kochko, V. Poncet, C. Dubreuil-Tranchant, E. Couturon, R. Guyot and P. Hamon (2012). "Geographical gradients in the genome size variation of wild coffee trees (Coffea) native to Africa and Indian Ocean islands." Tree Genetics & Genomes **8**(6): 1345-1358.

Roulin, A., B. Piegu, P. M. Fortune, F. Sabot, A. D'Hont, D. Manicacci and O. Panaud (2009). "Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae." BMC Evol Biol **9**: 58.

Roulin, A., B. Piegu, R. A. Wing and O. Panaud (2008). "Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza." Plant J **53**(6): 950-959.

Schmieder, R. and R. Edwards (2011). "Quality control and preprocessing of metagenomic datasets." Bioinformatics **27**(6): 863-864.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C. T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A. P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J. M. Chia, J. M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing and R. K. Wilson (2009). "The B73 maize genome: complexity, diversity, and dynamics." Science **326**(5956): 1112-1115.

Sonnhammer, E. L. and R. Durbin (1995). "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis." Gene **167**(1-2): GC1-10.

Tanskanem, J. A., F. Sabot, C. M. Vicient and A. Schulman (2007). "Life without GAG: The BARE-2 retrotransposon as a parasite's parasite." Gene **390**: 166-174.

Vicient, C. M., R. Kalendar and A. H. Schulman (2005). "Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon." J Mol Evol **61**(3): 275-291.

Vitte, C. and J. L. Bennetzen (2006). "Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution." Proc Natl Acad Sci U S A **103**(47): 17638-17643.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel and A. H. Schulman (2007). "A unified classification system for eukaryotic transposable elements." Nat Rev Genet **8**(12): 973-982.

Wicker, T., S. Taudien, A. Houben, B. Keller, A. Graner, M. Platzer and N. Stein (2009). "A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley." Plant J **59**(5): 712-722.

Witte, C. P., Q. H. Le, T. Bureau and A. Kumar (2001). "Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes." Proc Natl Acad Sci U S A **98**(24): 13778-13783.

Figure legends

Fig 1. Conserved structures of Non autonomous LTR retrotransposons documented in plant genomes

Autonomous refers to the structure of complete LTR retrotransposons (here Copia – like): The coding regions are in grey; the Primer Binding Site motif (PBS) is represented as a black triangle and the PolyPurine Track (PPT) is represented as a white triangle; LTR: Long Terminal Repeat, GAG: Capsid, AP: Aspartic protease, INT: Integrase, RT: Reverse transcriptase, RNAse: RNAse H. TRIM refers to the structure of Terminal-repeat Retrotransposons in Miniature (Witte, Le et al. 2001). LARD refers to the Large Retrotransposon Derivative element (Kalendar, Vicient et al. 2004). BARE-2 refers to the BARE-2 non-autonomous found in barley (Tanskanem, Sabot et al. 2007).

Fig 2. Structure and graphical alignments of the non-autonomous LTR Retrotransposons TRIM-1 family

A. Schematic representation of the TRIM-1-S element and alignment of five different *C. canephora* TRIM-1-S genomic copies against themselves using Dotter (Sonnhammer and Durbin 1995). B. Schematic representation of the TR-GAG1 element and alignment of five different *C. canephora* TR-GAG1 genomic copies against themselves using Dotter. C. Dotter alignment between TR-GAG1 (horizontal sequence) and TRIM-1-S (vertical sequence).

Fig 3. Schematic representation of the TR-GAG1 structure

The TR-GAG1 element contains the following sequence characteristics: LTR, PBS, PPT and an ORF harboring known GAG motifs (here UBN2 and Zf-C2HC motifs). The element shown is located on “Chr. 0” positions 113,020,990-113,023,502 from the *C. canephora* draft genome (http://coffee-genome.org).

Fig 4. Characterization of TR-GAG families in the *C. canephora* draft genome

A. Dot-plot of 130 predicted TR-GAG sequences against themselves. TR-GAGs were predicted by LTR_STRUC and filter out according to features described for TR-GAG1. Sequences were clustered by similarity. B. Detailed structure of one copy (Chr. 4, positions 21,003,142-21,006,851) of the TR-GAG2 family.

Fig 5. Identification of TR-GAG families in available plant genomes.

A. Dot-plot of predicted TR-GAG sequences from 23 plant genomes against themselves. TR-GAGs were predicted by LTR_STRUC and filter out according to features described fro TR-GAG1. Sequences were clustered by plant genomes. B. Detailed structure of one TR-GAG family for seven different plant genomes.

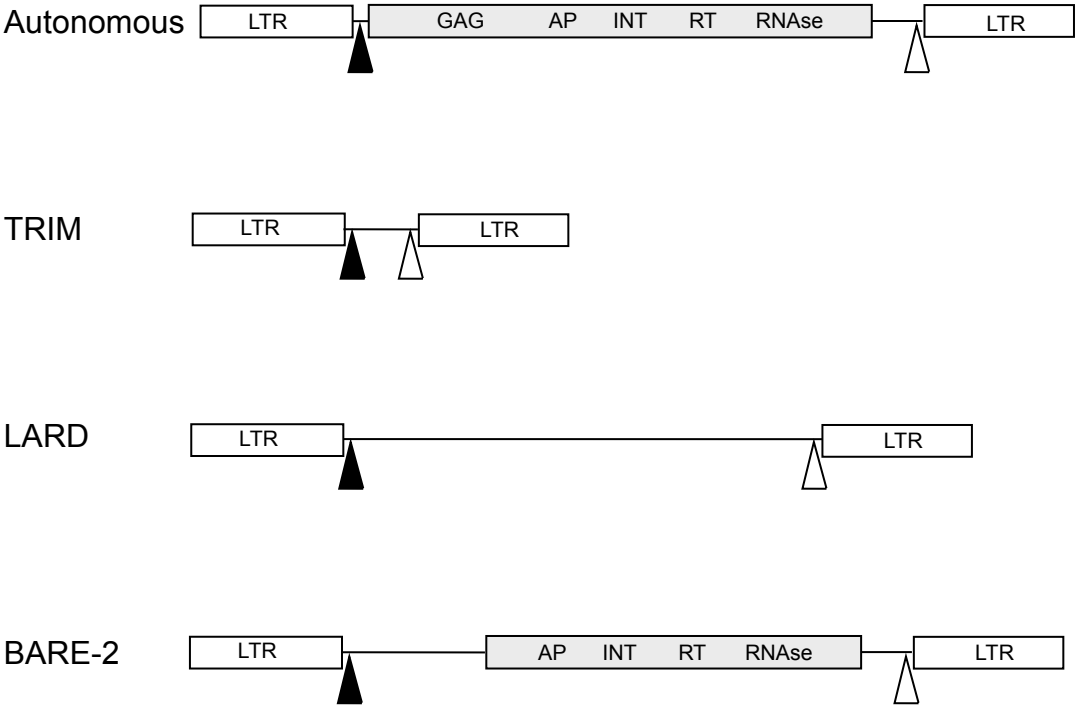
Primers	Sequences (5'-3')	Product size
TRIM-1-S-F	CACCTCCAACGGTTGATTCT	361 bp
TRIM-1-S-R	ATGTGTAGTTGCCCCGAGTC	
TR-GAG1-F	GCAGCAGACCTCTGGAAAAA	328 bp
TR-GAG1-R	TGGTTTGCCTTCCTTTGTTT	
G3-F	ACGAGTGGGTTTCCTGAGTG	‡
G3-R	TGGGTCTCTGGAACTTACCG	

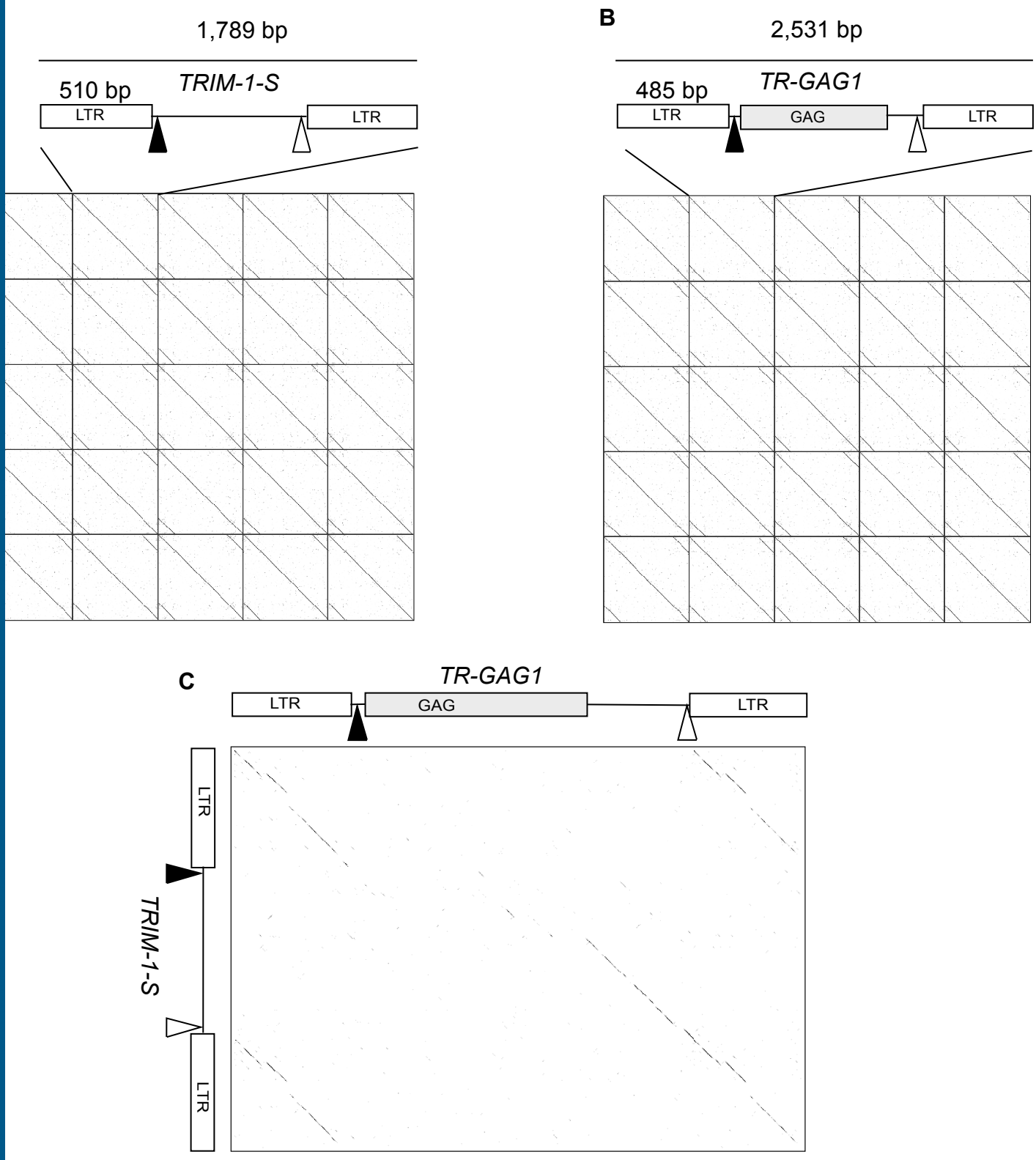
Table 1. List of Primers used for RT-PCR analysis. (‡) Control primers used as in (Guyot, de la Mare et al. 2009).

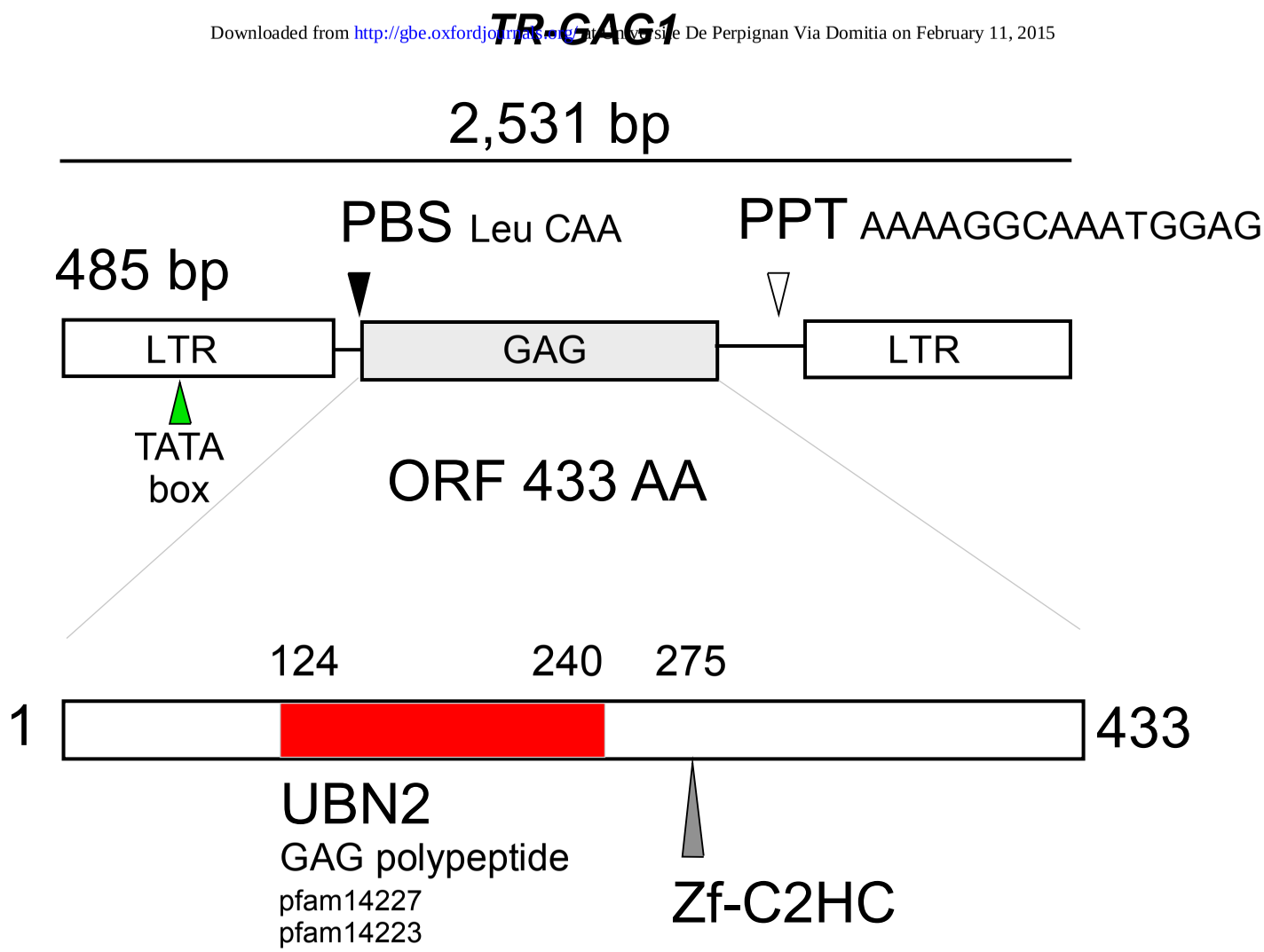
Species	Ploidy level	Estimated genome size (Mb)	#454 sequences	Produced bases (Mb)	Genome coverage	TR-GAG1 copies	TR-GAG2 copies	TR-GAG3 copies	TR-GAG4 copies	TR-GAG5 copies
C. canephora (HD94-200)	2n	700	106,459	45.05	6.40%	172,48	563,07	6,74	8,18	27,28
C. canephora (BUD15)	2n	700	149,196	67.08	9,58%	69,61	390,62	14,85	22,20	44,88
C. arabica	4n	1240	122,258	54.5	4.39	111,55	1168,72	55,40	10,21	35,21
C. eugenioides	2n	645	101,309	42.1	6.52%	62,56	659,44	28,64	26,14	22,42
C. heterocalyx	2n	863	194,3	60.511	2.25%	97,94	696,71	13,97	9,00	24,68
C. racemosa	2n	506	88,498	34.19	5.7%	54,02	103,02	2,96	0,00	16,04
C. pseudozanguerarie	2n	593	215,117	91.7	15.4%	59,76	157,79	1,12	7,34	13,67
C. humblotiana	2n	469	160,479	67.99	14.49%	26,77	80,00	0,00	0,00	13,64
C. tetragona	2n	513	160,107	72.66	14.10%	48,45	34,35	0,92	0,00	21,63
C. dolichophylla	2n	682	163,873	76.65	11.23%	61,91	144,93	0,00	0,00	18,40
Psilanthus horsfieldiana	2n	593	112,793	46.25	7.8%	43,56	336,74	1,35	0,00	24,50
Craterispermum kribi	2n	748	49,789	19.44	2,94%	5,07	6,96	0,00	0,00	0,00

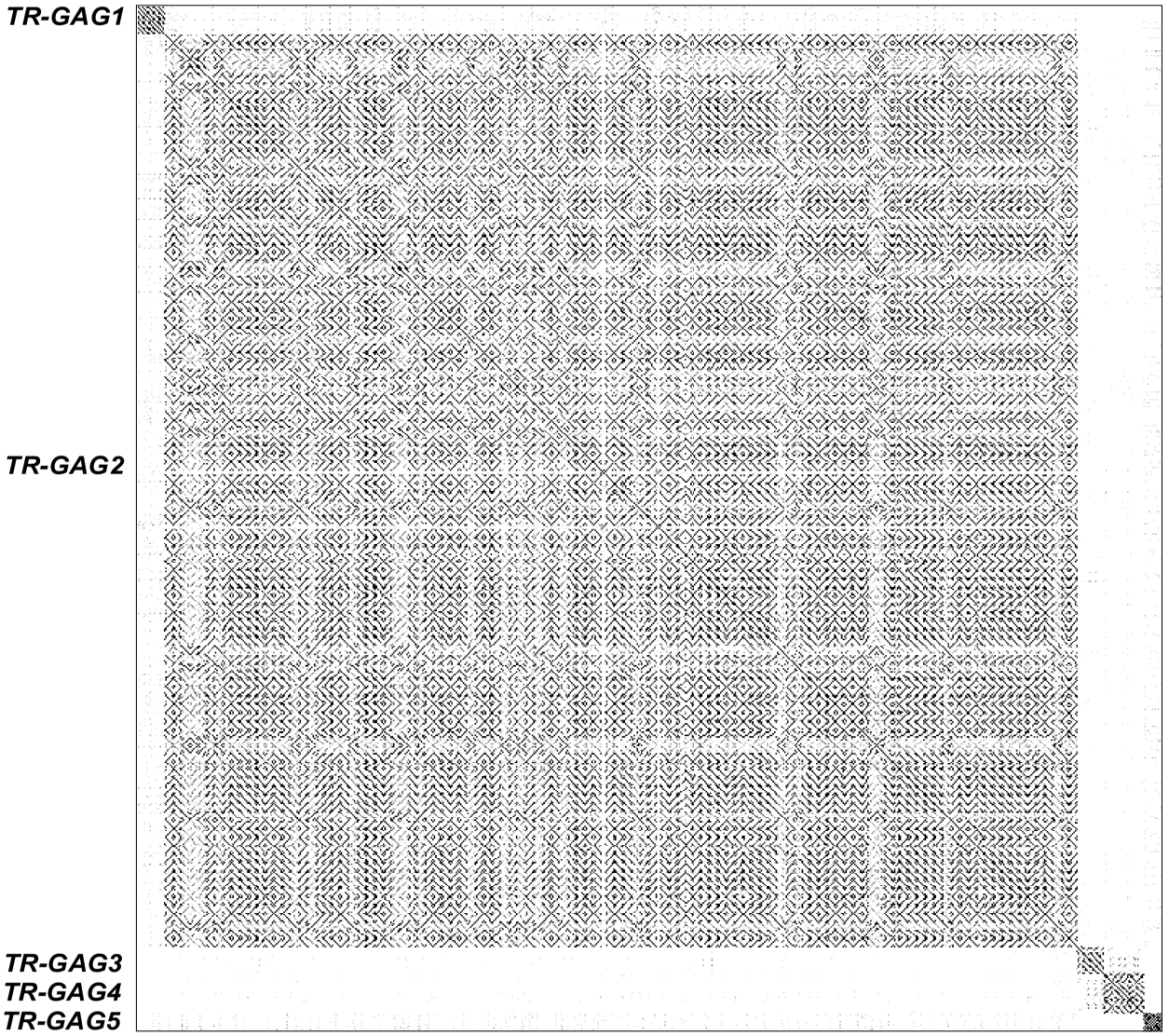
Table 2. Estimation of the TR-GAG family's copy number in *Coffea* genomes using 454 sequencing survey.

Only 454 reads with a minimum of 80% of nucleotide identity over 80% of the read length were considered. Genome sizes were listed in (Noirot, Poncet et al. 2003) and (Razafinarivo, Rakotomalala et al. 2012).









A

B