


Partial sequencing reveals the transposable element composition of *Coffea* genomes and provides evidence for distinct evolutionary stories

Romain Guyot²  · Thibaud Darré¹ · Mathilde Dupeyron¹ · Alexandre de Kochko¹ · Serge Hamon¹ · Emmanuel Couturon¹ · Dominique Crouzillat³ · Michel Rigoreau³ · Jean-Jacques Rakotomalala⁴ · Nathalie E. Raharimalala⁴ · Sélastique Doffou Akaffou⁵ · Perla Hamon¹

Received: 19 May 2016 / Accepted: 25 July 2016 / Published online: 28 July 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract The *Coffea* genus, 124 described species, has a natural distribution spreading from inter-tropical Africa, to Western Indian Ocean Islands, India, Asia and up to Australasia. Two cultivated species, *C. arabica* and *C. canephora*, are intensively studied while, the breeding potential and the genome composition of all the wild species remained poorly uncharacterized. Here, we report the characterization and comparison of the highly repeated transposable elements content of 11 *Coffea* species representatives of the natural biogeographic distribution. A total of 994 Mb from 454 reads were produced with a genome coverage ranging between 3.2 and 15.7 %. The analyses showed that highly repeated transposable elements, mainly LTR retrotransposons (LTR-RT), represent between 32 and 53 % of *Coffea* genomes depending on their biogeographic location and genome size. Species from West and Central Africa (Eucoffea) contained the highest LTR-RT content but with no strong variation relative to their genome size.

At the opposite, for the insular species (Mascarocoffea), a strong variation of LTR-RT was observed suggesting differential dynamics of these elements in this group. Two LTR-RT lineages, SIRE and Del were clearly differentially accumulated between African and insular species, suggesting these lineages were associated to the genome divergence of *Coffea* species in Africa. Altogether, the information obtained in this study improves our knowledge and brings new data on the composition, the evolution and the divergence of wild *Coffea* genomes.

Keywords LTR retrotransposons · Partial genome sequencing · *Coffea* · Genome size · Geographic divergence

Introduction

Repetitive sequences are major components of plant genomes. Transposable elements (TEs), constituting the mobile part of the genomes, are divided into two main classes (Class I and Class II) according to their mode of transposition. They are hierarchically classified into orders, super-families, lineages, families and individuals within each class (Wicker and Keller 2007; Wicker et al. 2007). Class I elements known as retrotransposons, transpose via an RNA intermediate without movement of the master copy. This ‘copy-and-paste’ mechanism can theoretically lead to a rapid increase of the frequency of the original copy. Class II, or transposons move following a ‘cut-and-paste’ mechanism or through DNA replication, resulting to low or moderate new inserted copies. Plant retrotransposons include two major orders: Long Tandem Repeat retrotransposons (LTR-RT) and non-LTR retrotransposons. The first ones include two super-families: *Copia* and *Gypsy* that differ mainly in their coding region organization and

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-016-1235-7) 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 UMR IPME, CoffeeAdapt, BP 64501, 34394 Montpellier Cedex 5, France

³ Nestlé R&D Tours, 101 AV. G. Eiffel, Notre Dame d’Oe’, BP 49716, 37097 Tours Cedex 2, France

⁴ FOFIFA, Ambatobe, Madagascar

⁵ University Jean Lorougnon Guédé, Daloa, Ivory Coast

are composed of ancient conserved evolutionary lineages in plants (Wicker and Keller 2007). The second ones includes long and short interspersed nuclear element, LINE and SINE, respectively (Kumar and Bennetzen 1999; Wicker et al. 2007).

During the last 10 years, the accumulation of genomic sequencing data (Michael and Jackson 2013) indicated that TEs are the major component of plant genomes and that their accumulation could be correlated with the genome sizes (Ibarra-Laclette et al. 2013; Kumar and Bennetzen 1999; Lisch 2013). LTR-RTs are the most redundant elements and in extreme cases, they can represent up to 80 % of plant genome sequences, suggesting that their propagation mechanisms are directly responsible of the genome size increase (SanMiguel et al. 1998; Schulman et al. 2004; Bennetzen et al. 2005; Hawkins et al. 2006; Dvořák 2009). Sometimes the propagation mechanisms induce a rapid accumulation, called a “burst”, of a few number of LTR-RT families as demonstrated in the wild rice *Oryza australiensis* (Piegu et al. 2006). At the opposite, in maize, the accumulation of LTR-RT families was probably gradual, but leading to a considerable genome size increase when compared to the sorghum or the rice genome. However, a correlation between genome size variations and LTR-RT copy numbers was not established for the *Zea* genus (Meyers et al. 2001), suggesting that the proliferation mechanism of a few LTR-RT families per se cannot explain all genome size variations in plants. The host genome controls the level of transposition of LTR-RTs through epigenetic mechanisms (Bucher et al. 2012; Ito 2013; Ito and Kakutani 2014). This control might be reduced under abiotic stresses (Todorovska 2007; Alzohairy et al. 2014; Kinoshita and Seki 2014), leading to an increase of transposition and suggesting that LTR-RT play a role in the genome adaptation facing environmental changes (Casacuberta and Gonzalez 2013). However, so far no correlation was established between plant genome size and their habitat or phenotypic and life traits (Eilam et al. 2007; Knight and Beaulieu 2008; Slovak et al. 2009; Dušková et al. 2010).

Next generation sequencing technologies provided powerful tools to identify and characterize the repetitive fraction of genomes even in large genomes such as wheat, barley or pea (Macas et al. 2007; Wicker et al. 2009). For these authors, an important advantage of the NGS sequencing lies in the limited bias obtained for the production of the sequences. Low-depth sequencing was effective in identifying the most highly repeated sequences and in estimating their copy numbers in the pea genome (Swaminathan et al. 2007), banana genome (Hribova et al. 2010) and in vesper bats (Pagan et al. 2012) and to study genome evolution at a genus or a family scale (Nystedt et al. 2013). It also allows identifying TEs insertion polymorphism accompanying clonal variation in grape (Carrier et al. 2012). The uneven

distribution of TEs between wheat and barley (Wicker et al. 2009), the genome size variation in the allotetraploid species *Nicotiana tabaccum*, (Renny-Byfield et al. 2011) as well the composition and abundance of highly repeated TEs in ten Triticeae taxa (Middleton et al. 2013) were also studied via a 454 pyrosequencing genomic survey.

Ranked fourth among angiosperms, the young Rubiaceae family [90.4 My divergence time, (Bremer and Eriksson 2009)] comprises ca. 600 genera and ca. 13,600 species. This family includes herbs, shrubs and trees growing naturally in overly diverse habitats (from desert to tropical sempervirent forests via temperate areas), altitudes (from sea level to over 2500 m) and soils. In this plant family, diploids are the most common and share the same basic chromosome number [$x = 11$ (Kiehn 1995)]. The *Coffea* genus, member of Rubiaceae, is the most known genus due to its major socio-economic importance worldwide (producers in Southern countries and consumers in Northern countries). Accounting for 124 described species, all diploids with $2n = 2x = 22$ but *C. arabica* (allotetraploid), the natural distribution in inter-tropical forests of Africa and of Western Indian Ocean Islands was recently extended to India, Asia and Australasia (Davis et al. 2011). The recent sequencing of *C. canephora* (also called Robusta) genome showed that no whole genome duplication has occurred after the Asterid clade divergence, some 110 My ago (Denoëud et al. 2014). Moreover, comparative mapping between two divergent African genomes: *C. canephora* and *C. pseudozanguebariae* did not reveal any major chromosomal rearrangements (unpublished data). Despite structural conservations, a notable variation of genome sizes is observed among *Coffea* species. This variation ranges from 469 to 900 Mb with a general pattern of increasing genome sizes from East to West in Africa (Noirot et al. 2003) and from North to South-East in Madagascar (Razafinarivo et al. 2012), suggesting a gradual accumulation of nuclear DNA, under speciation and adaptive processes of the species. Recently, the *C. canephora* genome sequencing allowed the computational identification of TEs (Denoëud et al. 2014). They represent more than half of the available genome sequence, and among them, LTR-RTs are the most frequent order of elements (42 % of the genome). However, outside the *C. canephora* genome, no wide survey of TE composition has been conducted in the *Coffea* genus.

Here, we used a 454 sequencing survey of one tetraploid and ten diploid species representative of the botanical and geographical diversity of the genus *Coffea* to study and compare the composition and abundance of highly repeated transposable elements in their genomes. Using a genome coverage ranging from 3.2 to 15.7 %, the analysis of LTR-RT composition and dynamics shows a clear difference between African and insular *Coffea* species, suggesting an ancient divergence. Contrary to previous hypotheses and

Table 1 454 sequencing data for 11 *Coffea* Species

Species	Accession	Country of origin	Group	No. of 454 reads	Total (bp)	Mean size (bp)	Coverage (%)	Genome size (Mb)
<i>C. arabica</i>	ET39	Ethiopia	EUC	93,194	41,643,904	446	3.2	1300
<i>C. arabica</i>	ET39	Ethiopia	EUC	112,615	51,892,121	460	3.9	1300
<i>C. arabica</i>	ET39	Ethiopia	EUC	140,976	61,729,478	437	4.7	1300
<i>C. canephora</i>	IF410	Ivory Coast	EUC	186,138	85,292,671	458	12.2	700
<i>C. canephora</i>	DH200-94	D. Republic Congo	EUC	98,017	43,037,451	439	6.1	700
<i>C. canephora</i>	BUD15	Uganda	EUC	140,120	64,290,611	458	9.2	700
<i>C. charrieriana</i>	OA22	Cameroon	EUC ^a	136,518	57,405,992	420	7.9	723
<i>C. eugenoides</i>	OUG14	Uganda	EUC	186,449	85,961,094	461	13.3	645
<i>C. eugenoides</i>	DA56	Kenya	EUC	91,834	39,993,235	435	6.2	645
<i>C. heterocalyx</i>	JC65	Cameroon	EUC	123,119	45,633,337	370	5.2	863
<i>C. pseudozanguebariae</i>	8107	Kenya	MOZ	215,117	91,733,301	426	15.5	593
<i>C. racemosa</i>	IA56	Mozambique	MOZ	173,803	79,199,218	455	15.7	506
<i>C. tetragona</i>	A.252	Madagascar	MAS	147,430	68,881,825	467	13.4	513
<i>C. dolichophylla</i>	A.206	Madagascar	MAS	147,758	70,632,674	478	10.4	682
<i>C. humblotiana</i>	A.230	Comoros	MAS	141,834	62,465,685	440	10.4	469
<i>C. horsfieldiana</i>	HOR	Indonesia	PSI	104,605	44,610,588	426	7.5	593

Botanical groups (Group) are those from Chevalier (1942) with *EUC* Eucoffea (species from West and Central Africa), *MOZ* Mozambicoffea (East Africa), *MAS* Mascarcocoffea (species from Western Indian Ocean Islands), *PSI* Paracoffea

^a The Eucoffea classification for *C. charrieriana* was not established by Chevalier since the species was recently described by Stoffelen et al. (2008). Therefore, its classification was assumed according to its geographical origin. Genome sizes are from Noiroto et al. (2003) and Razafinarivo et al. (2012). The genome coverage is given in %

generally admitted idea, our results suggest that the *Coffea* species from Western Indian Ocean Islands and from Asia have diverged independently from their continental counterparts. Furthermore, no strong activation of LTR-RTs was obvious in any species, whatever their genome size, suggesting that other molecular mechanisms or general but limited variation in TE copy numbers are associated to genome size increases in the *Coffea* genus.

Materials and methods

DNA isolation and 454 sequencing

Leaves from Madagascan and Comorian species were obtained from the Kianjavato Coffee Research Station (KCRS) in Madagascar. The African species were sampled from the *Coffea* collection maintained at IRD (Montpellier, France) or Nestlé R&D (Tours, France) greenhouses. The studied species belong to Chevalier's (Chevalier 1942) botanical sections, i.e., Eucoffea (West and Central African species), Mozambicoffea (East African species), Mascarcocoffea (species from the Western Indian Ocean Islands) and Paracoffea (species belonging to *Psilanthus* subgenus *Afrocoffea*). In total, we used seven Eucoffea, two Mozambicoffea, three Mascarcocoffea and one Paracoffea

accessions. Information on the accessions used, their origin and other used data are given in Table 1.

DNA was isolated from fresh or dried leaves using Qiagen DNeasy Plant Mini extraction kits following the manufacturer protocol. Quantity and quality of DNA was measured using a Nanodrop (ND-1000). The libraries construction and Next Generation sequencing were performed at Nestlé R&D laboratory (Tours, France) according to the Roche/454 Life Sciences Sequencing Method using one Roche 454 GS Junior plate per accession. Data were submitted to GenBank, BioProject PRJNA242989. General information on 454-pyrosequencing is available in Table 1.

Sequences analyses

Quality of 454 reads was checked using FASTQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and cleaned using Prinseq v0.20.4 (Schmieder and Edwards 2011).

BLASTX searches (minimum e-value $10e^{-4}$) were first carried out on 454-reads against the RepBase amino acid sequence dataset (Kohany et al. 2006; Jurka et al. 2005)—<http://www.girinst.org/repbase/>. BLASTN were carried out against *Coffea* coding sequence (CDS, <http://coffee-genome.org>), the *C. arabica* chloroplast genome (EF044213) and rRNA sequence (X52320 and AY083685)

with a minimum e-value of $10e^{-6}$. BLASTN analyses were also performed against the *C. canephora* repeat database built with REPET (<https://urgi.versailles.inra.fr/Tools/REPET>) with an e-value of $10e^{-20}$. The goal was to identify the major TE classes, super-families and lineages reported until today at different scales (amino acid and nucleotide) and to obtain their proportion in the investigated genomes. Given the importance of the Class I/LTR-RT in all genomes, BLASTN similarity searches were conducted between 454 reads and a dataset of LTR retrotransposons consensus sequences from *C. canephora* classified according to their Reverse Transcriptase (RT) amino acid similarities (available at the Gypsy Database 2.0). 454 sequences showing similarities with RT domains were classified by phylogenetic analyses. Identified RT domains from 454 datasets were extracted from the nucleotide sequences and translated into amino acids. Amino acid sequences (with a minimum of 150 residues) were aligned (ClustalW) to construct a bootstrapped neighbor-joining tree, edited with FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>).

Detailed annotation of the SIRE lineage (*Copia*) was performed using LTRFinder (<http://tlife.fudan.edu.cn/ltrfinder/>), (Xu and Wang 2007). LTR domain sequences were aligned with MUSCLE to build a consensus 100 bootstraps neighbor-joining phylogenetic tree with ClustalW. Complete SIRE elements were annotated with Artemis (Rutherford et al. 2000) and used as references. Structural incongruities (InDels and rearrangements) were searched using graphic alignments (dot-plot, (Sonnhammer and Durbin 1995)).

The copy number of SIRE in 454 dataset was estimated as described in (Chaparro et al. 2015) and (Dias et al. 2015). BLASTN searches were carried out with full-length SIRE elements found in the *C. canephora* genome. Reads with more than 90 % of nucleotide identity with the reference sequence over a minimum 90 % 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.

De novo detection of repeated sequences

De novo detection of repeated sequences was carried out using RepeatScout (<http://bix.ucsd.edu/repeatscout/>) (Price et al. 2005) on 454 sequences for each species. The libraries of repeated sequences were used to mask each 454 dataset using RepeatMasker (<http://www.repeatmasker.org>). Repetitions were then filtered out according to their

minimum redundancy in 454 dataset as follow: 20, 100, 500 and 1000 repetitions.

Searches for microsatellites

Microsatellites were detected on 454 sequences using the MicroSATellite identification tools (<http://pgrc.ipk-gatersleben.de/misa/>). The unit size of repetition ranged from 1 to 20 and the number repeated units ranged from 1 to 10.

PCR amplification on *Coffea* DNA

Primers were designed on three full-length SIRE annotated in this analysis (called 36-863, 3-942 and 6-1571) on ENV and LTR domains using Primer3 (<http://primer3.ut.ee>) (Supplemental data 1A). PCR amplifications were performed in a final volume of 20 μ L using the GoTaq DNA polymerase from Promega, according to the manufacturer recommendations: 0.5 ml of dNTP (10 nM), 1 ml of each primer (10 mM), 0.2 U of Taq polymerase (GoTaq, Promega) and 20 ng of DNA matrix. We used the following PCR amplification cycle: 98 °C 2 min.; three steps (98 °C 30 s, 55 °C 30 s, 72 °C 30 s) repeated 35 times followed by a final elongation step (72 °C 5 min). The DNA samples, representative of the biogeographic *Coffea* groups, (Supplemental data 1B) are those used in (Razafinarivo et al. 2013).

Results

454 sequencing in *Coffea*: run reproducibility and characterization of genomes composition

The 454 junior runs were produced for 10 *Coffea* diploid and one tetraploid species. Three independent runs for the same accession (ET39) of the tetraploid species, *C. arabica*, were carried out to check the reproducibility of the runs. In addition, for two diploid species, *C. canephora* and *C. eugenoides*, three (BUD15, HD200 and IF410) and two (DA56 and OUG14) accessions were, respectively, sequenced. The 454 sequencing produced a genome coverage ranging from 3.2 to 4.7 % for *C. arabica* and from 5.2 to 15.7 % for all the diploid species (Table 1). In total, more than 2.2 millions reads, accounting for 994 Mb, were produced and analyzed in this study. The three *C. arabica* replicates gave similar results showing the good reproducibility of the sequencing and enabling to have confidence in the results presented here.

Using BLASTN (CDS, chloroplast genome, rDNA) and BLASTX (transposable elements) we found that protein-coding genes represented between 11 % (*C. heterocalyx*) and 18 % (*C. canephora* acc. DH200-94) of the obtained

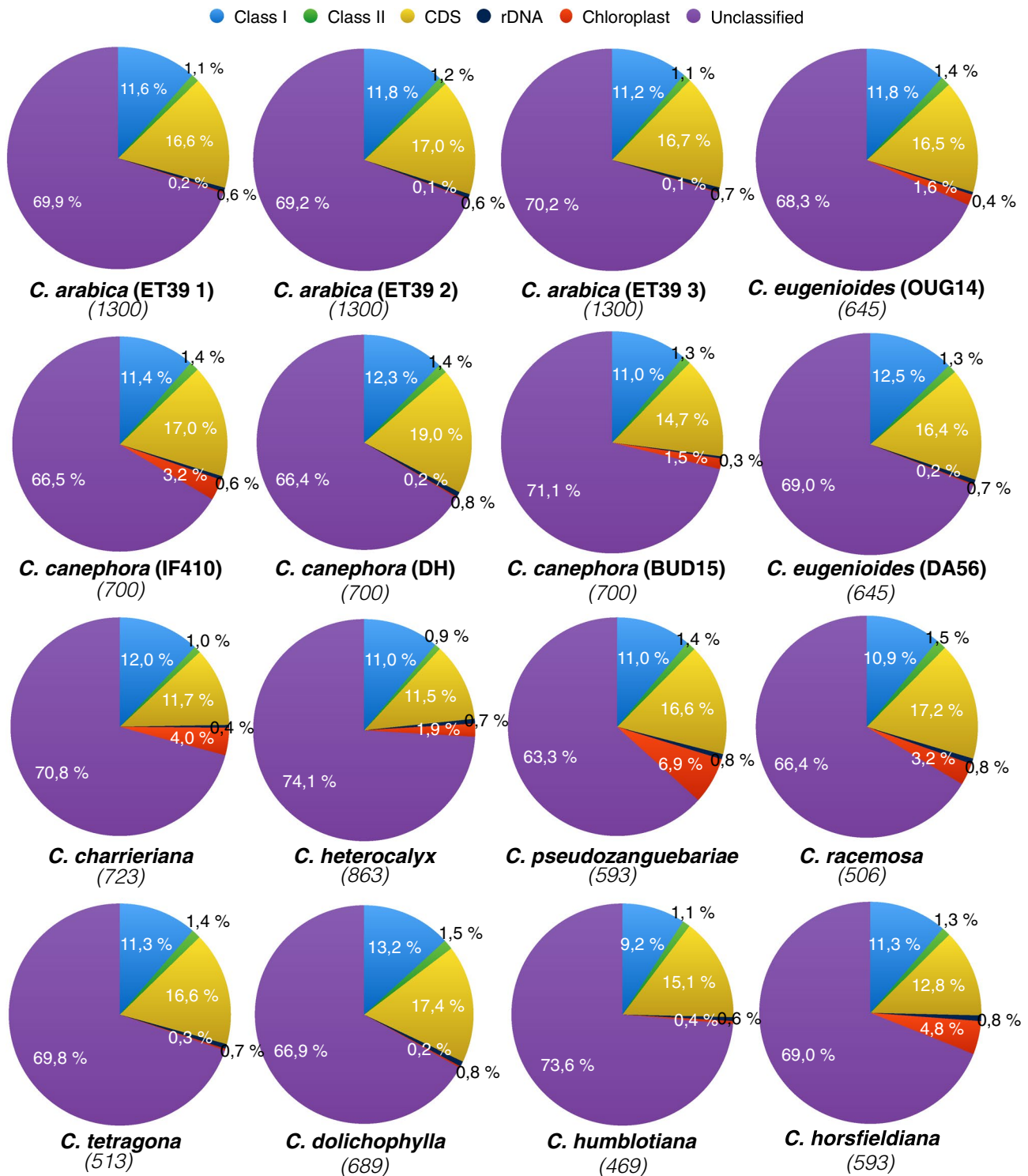


Fig. 1 Composition of 454 reads for 11 *Coffea* species and 14 accessions. *Class I* and *Class II* are known transposable element coding regions, *CDS* cellular coding regions, *rDNAs* ribosomal DNA genes.

Name and accession of species were indicated with their respective genome size indicated into brackets (in Mb)

data (Fig. 1). A similar percentage to that of *C. canephora* was found for the three *C. arabica* replicates (17%). However, the proportion of identified chloroplast sequences

between species varies between 0.14 % (*C. arabica*) and 7 % (*C. pseudozanguebariae*). Five species showed a percentage of chloroplast sequences larger than 2 % (Fig. 1).

Chloroplast DNA presence may be attributed to the fact that total DNA was extracted for the sequencing and not just the nuclear fraction as in (Carrier et al. 2011) or to different amount of chloroplast DNA inserted into the nuclear genomes according to the studied taxa, such insertions have been observed in the sequenced *C. canephora* genome (Denoeud et al. 2014). Recognizable coding sequences from transposable elements represented a significant proportion ranging from 10 % for *C. humblotiana*, the smallest genome [469 Mb, (Razafinarivo et al. 2012)] to 14 % for *C. dolichophylla*, an average size genome (689 Mb). Interestingly, the genome of *C. heterocalyx* [the biggest one with 863 Mb, (Noirot et al. 2003)] was containing 12 % of transposable element coding genes.

For *C. canephora*, a similar TE coding sequences proportion (Class I and Class II) was found for the three accessions analyzed (BUD15, IF410 and DH200-94) originating from three different geographical areas (respectively, 12.3, 12.8 and 13.7 %). For all the species, most of the identified coding sequences of transposable elements fell into the Class I, as found for the *C. canephora* genome sequence (Denoeud et al. 2014).

To further investigate the composition of repeated sequences in *Coffea* species, we used as reference the *C. canephora* database of consensus transposable elements that was constructed de novo and annotated using the REPET programs. The *C. canephora* database is composed of 4051 consensus sequences for which 1536 and 2023 belonged to the LTR retrotransposons and non-autonomous LTR retrotransposons, respectively. Using this dataset, the proportion of LTR retrotransposons in the 454 reads reached 32 % for *C. humblotiana* and 53 % for *C. heterocalyx* (Supplemental data 2). Interestingly, the amount of 454 reads similar to *C. canephora* LTR retrotransposon consensus sequences was very similar for Eucoffea species whatever their genome size (*C. arabica*: 50–51 %, *C. eugenoides*: 48–50 %, *C. canephora*: 49–52 %, *C. charrieriana*: 48 % and *C. heterocalyx*: 53 %), while a clear lower amount was observed for the Mozambicoffea species (*C. pseudozanguebariae*: 37 %, *C. racemosa*: 39 %), for Mascarocoffea species (*C. tetragona*: 36 %, *C. dolichophylla*: 40 %, and *C. humblotiana*: 32 %) and for Asian Paracoffea (*C. horsfieldiana*: 34 %). These variations between Eucoffea and the three other botanical groups (Mozambicoffea, Mascarocoffea and Paracoffea), appeared independent from the genomes size, at the exception of *C. humblotiana* that showed both the smallest genome (469 Mb) and the lowest percentage of 454 reads containing sequences similar to *C. canephora* LTR retrotransposons (32 %). Such variation could be attributed to the nucleotide divergence of LTR retrotransposons between Eucoffea and the other botanical groups since the nucleotide database of LTR retrotransposons used as reference was established from *C. canephora*

(Eucoffea). Altogether our data suggest a noticeable variation of the quantitative LTR-RT content in *Coffea* species genomes.

Abundance of LTR-retrotransposon lineages and their contribution to genome size

To further investigate the quantitative variation of LTR-retrotransposon content, we first classified the REPET consensus sequences into *Copia* and *Gypsy* super-families and, thus, into lineages (*Bianca*, *Oryco*, *Retrofit*, *Sire*, *Tork* for *Copia* and *Athila*, *CRM*, *Del*, *Galadriel*, *Reina* and *TAT* for *Gypsy* (Llorens et al. 2009), according to their similarities to reverse transcriptase (RT) reference domains. In total, LTR-retrotransposon consensus sequences were assigned to 877 families containing RT domains, for which 352 and 525 belong to *Copia* and *Gypsy*, respectively. These 877 families belong to all the different LTR-retrotransposon lineages previously discovered in other plant genomes. Using this dataset, all the *Coffea* species analyzed were found to contain a *Gypsy/Copia* ratio ranging from 2.6 to 4.6, suggesting that *Gypsy* represented the most abundant LTR-retrotransposon super-family in *Coffea* species, as previously found in *C. canephora* (Denoeud et al. 2014; Dereeper et al. 2013). The overall proportion of *Copia* and *Gypsy* varied greatly according to Chevalier's botanical classification and increased from Eucoffea to Mascarocoffea (Supplemental data 3). These variations were not noticeable when the 454 reads were translated (using BLASTX analysis against RepBase). Interestingly the *Gypsy/Copia* ratio was clearly heterogeneous among Mascarocoffea species. Indeed the proportion of different lineages also varied according to the botanical classification (Fig. 2). Two lineages, *SIRE* from *Copia* and *Del* from *Gypsy* appeared to differ strongly in the 454 reads between Eucoffea, Mozambicoffea, Mascarocoffea and Paracoffea. In Eucoffea, the *SIRE* lineage is present in 4.5–5.1 % of the 454 reads (identified with BLASTN, value $10e^{-20}$), at the exception of *C. charrieriana* for which 3.2 % of reads contained this lineage. Mozambicoffea species contained a lower percentage of *SIRE*, with 2.1 and 2.2 % for *C. pseudozanguebariae* and *C. racemosa*, while *SIRE* sequences were very rare in Mascarocoffea species and Paracoffea (between 1.1 and 1.5 %). Another important variation between botanical groups is observed for the *Del* fraction; going from 16.2 to 14 % in Eucoffea, 10.7 to 11.6 % in Mozambicoffea, 7.3 to 9.9 % in Mascarocoffea and 7.2 % in Paracoffea (Fig. 2; Supplemental data 4). Here also, the lowest percentage in Eucoffea is observed for *C. charrieriana* (13.1 %), contrasting with the other species of this botanical group. The pattern of LTR retrotransposon identified in *C. charrieriana*, suggests that this species differs from all the other Eucoffea species studied here.

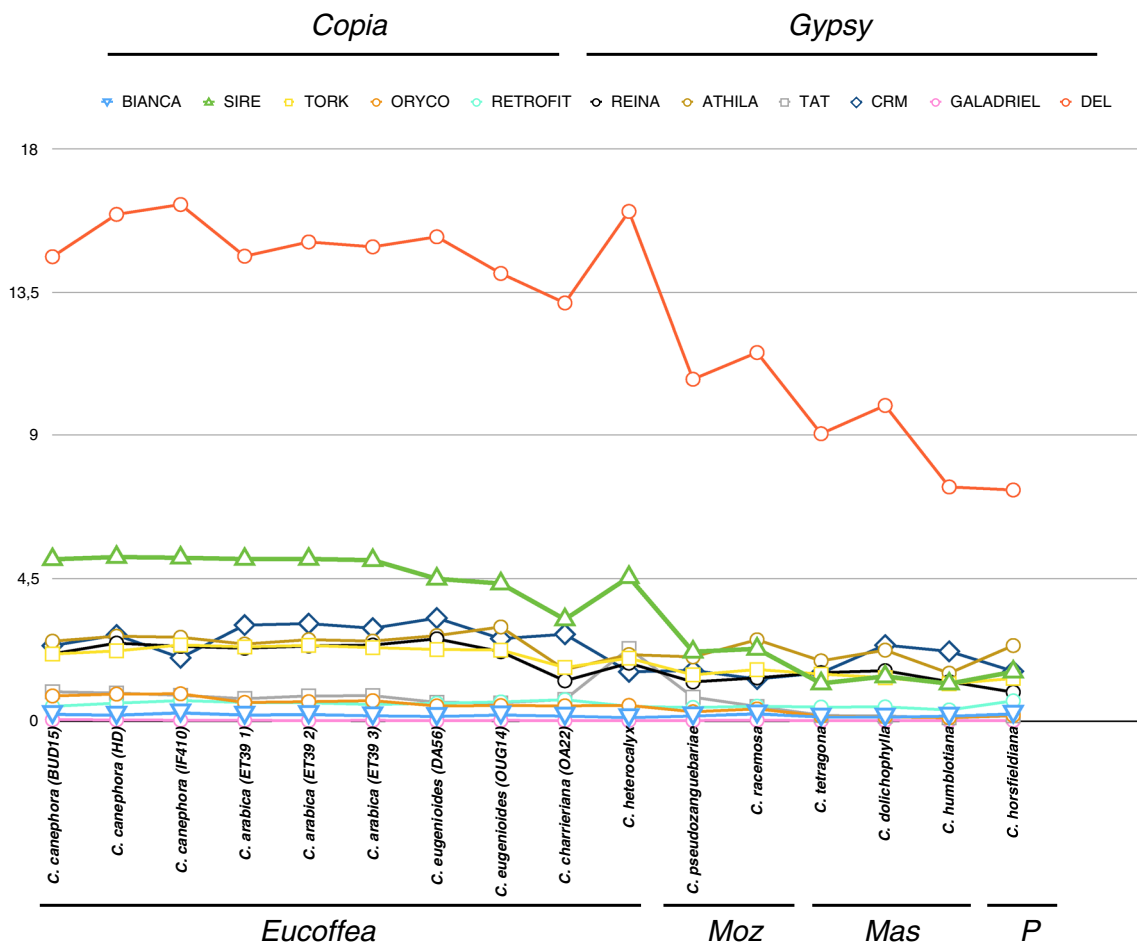


Fig. 2 Composition of 454 reads (in percentage) similar to LTR retrotransposon lineages between 11 *Coffea* species, organized according to their botanical sections: Eucoffea, Mozambicoffea (Moz), Mascarocoffea (Mas) and Paracoffea (P)

Interestingly, no clear relationship was found between the abundance of LTR retrotransposon super-families or lineages and, the genome size variation. However, there is a clear relationship between the abundance of detected elements and the botanical classification of the *Coffea* species.

De novo detection of repeated sequences in *Coffea*

As no clear relationship could be established between the presence of LTR retrotransposons and the genome size variation in *Coffea* genomes, another type of repeated sequences should be involved. For this, we estimated the global number of repeated sequences (excluding microsatellite sequences) presenting more than 20, 100, 500 and 1000 repeats and their proportion in each dataset (Supplemental data 5). Repeated sequences with a minimum of 20 copies represented between 54.4 (*C. heterocalyx*) and 45.6 % (*C. canephora*) of reads for Eucoffea, 46.3 and 41.8 % for Mozambicoffea, 43.4–33.3 % (*C. humblotiana*) for Mascarocoffea and 44.1 % for Paracoffea. A similar

pattern was observed for repeated sequences with more than 100, 500 and 1000 copies. *C. heterocalyx* (863 Mb) and *C. canephora* (IF410; 700 Mb) are the two samples with the highest proportion of repeated sequences, while *C. humblotiana*, the smallest genome, has the lower number of repetitions. Among Mascarocoffea, this percentage differs considerably between *C. humblotiana* (469 Mb) and *C. dolichophylla* (682 Mb). Interestingly, some species appears enriched with highly repeated sequences (>500 and >1000 copies), such as *C. heterocalyx* (10.8 % of sequences were repeated more than 500 times), while *C. humblotiana* and *C. horsfieldiana* contained very few highly repeated sequences (Supplemental data 5).

Microsatellites and genome size variation

Different types of microsatellites were identified and their cumulative length was represented on a histogram (Supplemental data 6). No large variation of the microsatellite content was observed among the species analyzed. Indeed,

the amount of microsatellite is higher in *C. arabica*, which is the allotetraploid species, but for the diploid species it doesn't show any variation corresponding to the genome size, whatever the size of the microsatellite motif (Supplemental data 7).

The SIRE LTR retrotransposon lineage and *Coffea* geographic distribution

As LTR retrotransposons represented a significant but variable part of *Coffea* genomes, we assess their relationships from phylogenetic analysis based on their RT domains at the amino acid level. The tree obtained using 2,325 RT domains (with a minimum length of 150 amino acids) (Supplemental data 8) shows clearly an organization into lineages between the two super-families *Gypsy* and *Copia*. For each lineage, it was possible to observe a combination of RT domains from different botanical groups (Eucoffea, Mozambicoffea, Mascarocoffea and Paracoffea). However, one lineage named SIRE, showed a specific pattern with an over-representation of RT sequences from Eucoffea and Mozambicoffea and, very few from Mascarocoffea and Paracoffea. From the 263 RT belonging to the SIRE lineage, five belong to the Indonesian Paracoffea species, and 21, 49 and 188 belong to Mascarocoffea, Mozambicoffea and Eucoffea, respectively. This observation suggests a different dynamics of SIRE elements depending on the botanical group of the species. An in-depth study of this lineage was performed to confirm our observations.

SIRE LTR retrotransposons were identified, annotated and characterized in the *C. canephora* genome (Chaparro et al. 2015). After detailed analysis, a total of 85 full-length SIRE LTR retrotransposons were selected for further analyses. SIRE elements from this dataset showed strong similarities with the SIRE internal coding domains from the *Gypsy* 2 database, and they had no apparent large insertion. All these predicted SIRE elements showed an overall length around 9–10 kb, with an average LTR length of 1 kb. The internal regions of these sequences included a large open reading frame (ORF1) containing the consensus for the GAG, AP, INT, RT and RNaseH domains. Downstream of ORF1 an additional small ORF (ORF2) showing strong identities with the ENV domain of retroviruses was identified.

These 85 sequences were classified through phylogenetic analysis based on their LTR sequences, into three major clusters (A, B and C) composed of 17, 28 and 40 elements, respectively (Supplemental data 9). For each cluster, one full-length sequence (with highest percentage of LTR identity, and highest overall length) was used as a reference sequences for further analyses (the sequences were named 36-863, 3-942 and 6-1571 for A, B and C cluster, respectively).

The copy number of SIREs elements estimated in the set of species analyzed here and using the three references SIRE sequences previously defined, showed a large variation between botanical groups (Supplemental data 10). The highest number was obtained for the Eucoffea with the exception of *C. charrieriana*, while Mascarocoffea species and Paracoffea showed very few SIRE sequences. The Mozambicoffea showed a moderate number of SIRE copies, whose numbers ranged between that of Eucoffea and Mascarocoffea. To confirm these observations at the molecular level, we conducted a PCR amplification survey of LTR and/or ENV domains based on the three SIRE elements reference over a large panel of species (Supplemental data 1). Amplification products were obtained for nearly all the Eucoffea, while amplifications were obtained for few Mozambicoffea species and almost no amplifications were observed for the Mascarocoffea and Paracoffea (Fig. 3; Supplemental data 11).

Discussion

The objective of this study was to investigate the transposable element composition of diploid and allotetraploid genomes from the *Coffea* genus. In some plant genomes, a clear relationship was established between the number of LTR retrotransposons and the variation of genome size (Piegu et al. 2006; Lee and Kim 2014). Considering a relatively short evolutionary divergence time of the *Coffea* genus [~11 MY; (Tosh et al. 2013)] and a significant variation of genome size observed among species (from 469 to 900 Mb), we focused our study on the identification and the characterization of repeated sequences and more particularly the LTR retrotransposons.

We used the 454 Junior apparatus to produce partial genome sequencing, representing genome coverage from 3.2 to 4.7 % for the allotetraploid *C. arabica* and 5.2–15.7 % for ten diploid *Coffea* species. Such “454 whole genome snapshot” approach has been recently used in plant and animal genomes to study and compare their composition in transposable elements, with similar or even lower (Wicker et al. 2009; Middleton et al. 2013; Sergeeva et al. 2014; Swaminathan et al. 2007; Pagan et al. 2012). No bias of genomic sampling for particular sequence type was noted when using the 454 sequencing procedure (Swaminathan et al. 2007). Indeed using a relatively low genome coverage, only highly repeated transposable elements can be accurately studied and low-copy number repeated sequences will not be represented in our dataset (Macas et al. 2007). Despite the 454 sequencing technology is beginning to be outdated; it generates long reads allowing an accurate identification of genes and transposable elements. Other approaches are now possible to study the

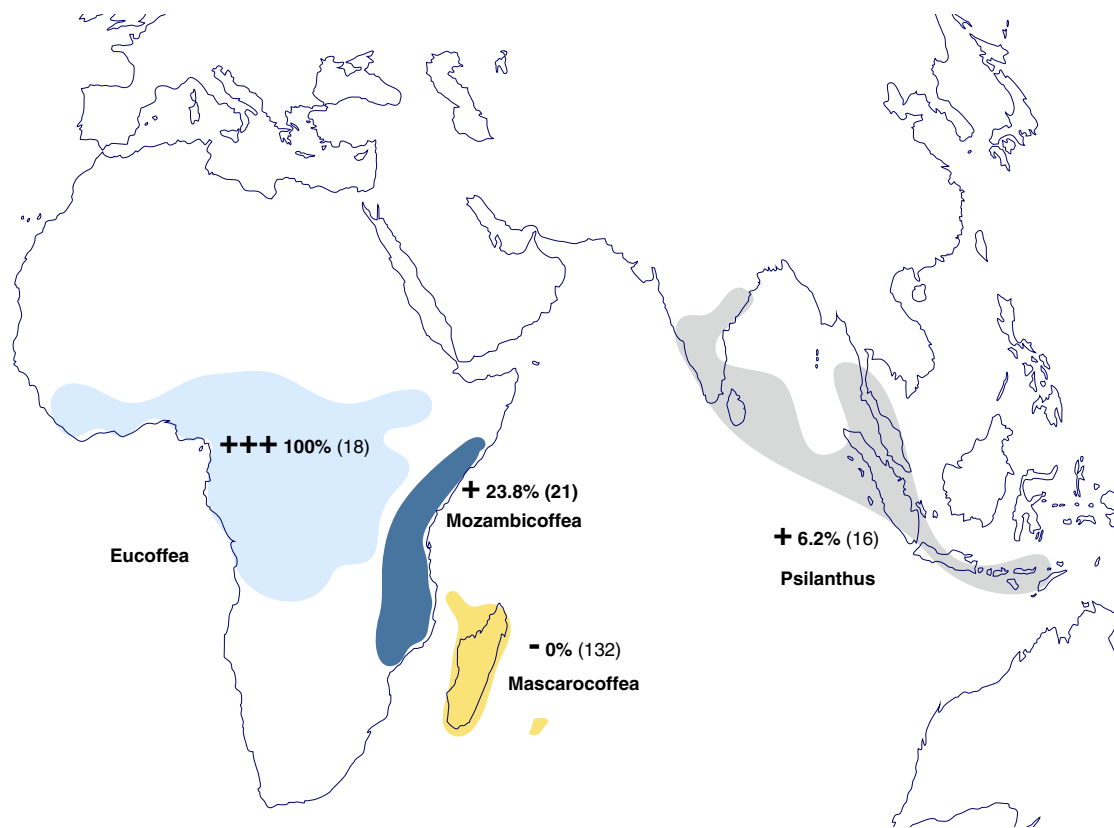


Fig. 3 Geographical distribution of *Coffea* botanical groups and summary of SIRE PCR amplifications. The summary of SIRE PCR amplification is symbolized by the rate of PCR amplification in percentage, and the number of PCR assays performed in parenthesis. 18,

7, 30 and 4 species were used as DNA matrix for, respectively, Eucoffea, Mozambicoffea, Mascarocoffea and Paracoffea (Supplemental data 8, 9 and 10)

transposable element composition and copy numbers using the Illumina platform providing shorter read length but with an unrivaled genome coverage (Ramachandran and Hawkins 2016).

In this study, we confirmed that no bias was observed in the randomness of the sequencing when performing three repetitions of the same accession (*C. arabica*, Et39). So far, few studies concerned the TE abundance and dynamics among species from a single plant genus. Most of them were performed on annual plants, with the exception of the gymnosperm family (Nystedt et al. 2013). However, for perennial angiosperms, the dynamics and the evolutionary history of TE within a genus remain poorly studied.

TE composition reflects the divergence of the botanical groups

For the first time, we conducted a study on TE composition of the genome of eleven *Coffea* species. Our study was based on the analysis of 454 reads for (1) their similarities with known TE proteins in plants and against a library of TE annotated in the *Coffea canephora*

genome; and, (2) an ab initio identification of repeated sequences.

We found that the most repeated order of transposable elements are LTR retrotransposons as found in the *C. canephora* genome and in most plant genomes (Lee and Kim 2014). At the TE amino acid level, as curated in Repbase, we found a similar percentage of TE between the generated *C. canephora* 454 reads (ranged from 12.2 to 13.5 %) and a recent and similar analysis of 131,412 BAC End Sequences (BES) from two *C. canephora* (DH200-94) BAC libraries [11.9 % (Dereeper et al. 2013)]. Surprisingly the percentage of known TE coding sequences remains relatively stable whatever the botanical groups, the species and the genome size. Interestingly, the only notable differences concerned *C. dolichophylla* and *C. humblotiana* species showing, respectively, 14.6 and 10.2 % of detected TE coding sequences. Considering the genome size difference (689 and 469 Mb), these species that belong to the Mascarocoffea may have underwent a different history of TE accumulation. This observation was confirmed at the nucleotide level using a *C. canephora* *de novo* library of TEs using REPET.

Using a detailed classification of LTR-RT REPET consensus, we also found that some lineages have varying distribution levels among *Coffea* species and botanical groups. For example the *Gypsy* Del lineage identified in higher abundance in African species, decreases from *Eucoffea* species (14–16 %), to *Mozambicoffea* (10–11 %), *Mascarocoffea* and *Paracoffea* (9–7 %). This suggests an overall increase of the Del LTR-RT westwards; from Indonesian and Malagasy *Coffea* species to eastern and western African species. Another LTR RT lineage, named SIRE (*Copia* super-family) was identified as being significantly numerous in African species (in 5 % of the 454 reads), but almost absent in Indonesian, Madagascan and Comorian species (~1 %), this observation was confirmed by the realization of PCR amplifications (Fig. 3). This indicates that the SIREs proliferated successfully in African species (in *Mozambicoffea* and especially in *Eucoffea*) while the copy number remained low, by lack of activity or elimination, in species from insular species.

These two examples of LTR-RT lineages variation, suggesting different history of TE proliferation, reflect independent genome divergences between *Coffea* botanical groups. This result also suggests that geographical differentiation could be associated to independent niches colonization and speciation in Africa, Madagascar and Indonesia. Therefore, quantitative and qualitative TE composition might be used for performing phylogeny analysis and to reinforce a model for the evolution of plant species.

TE composition reflects a different evolution of species within the botanical groups

It is well established that plant genome sizes are directly linked with the proportion of transposable elements. A large amplification of a small number of LTR retrotransposons lineages may cause a dramatic and sudden genome size increase (Piegu et al. 2006). In our study, we found contrasted results between the genome size of *Coffea* species and their TEs composition.

Few variation of TE composition was related to the genome size in *Eucoffea*, although genome size varies from 645 Mb for *C. eugenoides*, to 863 Mb for *C. heterocalyx* (700 Mb for *C. canephora*). This suggests that no rapid proliferation of few TE families was involved to explain this genome size difference. Particularly the TE proportion is almost identical between *C. canephora* and *C. heterocalyx* with the exception to one *Gypsy* lineage named *TAT*, that varies from 0.9 % in *C. canephora* to 2.2 % in *C. heterocalyx*. However, this recent proliferation in *C. heterocalyx* cannot explain alone the genome size difference between the two species. We, therefore, propose that in *Eucoffea* the genome size variation would result from a differential

accumulation of numerous transposable elements (mainly LTR RT) belonging to a large panel of families.

Similarly, no strong variation of microsatellite copy numbers was detected between species, suggesting that a rapid amplification of some of these simple sequence repeats was not the main mechanisms involved in the *C. heterocalyx* genome size increase as it was observed in *Lupinus* (Martin et al. 2016). Our results are congruent with those of *Pinus* (Morse et al. 2009), *Helianthus* (Cavallini et al. 2010), and *Lupinus* (Martin et al. 2016) both genera showing a large genome size variation (18–40, 3.2–12.3 and, 0.97–2.4 Gb, respectively) but with none element contributing specifically to this variation.

At the opposite, the *Mascarocoffea* species present more important variations of their TEs composition. The strong contrast in TE content between *C. dolichophylla* and *C. humblotiana* is due to an increase/decrease of the amount of the Del LTR retrotransposon lineage (10 vs 7 %) and a smaller increase/decrease for the remaining LTR RT lineages. *C. humblotiana*, has undergone few proliferation of LTR retrotransposons explaining its small genome size (469 Mb) while *C. dolichophylla* has undergone proliferation of mainly *Del* and several other *Copia* and *Gypsy* LTR-RT lineages. The variation of repeated sequences between *C. dolichophylla* and *C. humblotiana* is also clear with the *de novo* analysis showing a clear increase/decrease in repeated sequences. Since the fully resolved phylogenetic analysis of *Mascarocoffea* is not yet available, the time-scale of the LTR RT proliferation in *C. dolichophylla* cannot be estimated.

Altogether, our analysis demonstrated the power of sequencing at low coverage to study the transposable elements composition of genomes at the genus scale for comparative structural genomics of non-model species. The *C. humblotiana* species represents an interesting genomic model, worth to have its genome completely sequenced. This WGS will allow a better understanding of the mechanisms involved in the decrease or in the control of the proliferation of transposable elements in a genome.

Compliance with ethical standards

Conflict of interest All authors declare they have no conflict of interest.

Funding This research was supported Agropolis Fondation through the “Investissement d’avenir” program (ANR-10-LABX-0001-01) under the reference ID 1002-009.

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

Data availability The project has been deposited at DDBJ/EMBL/GenBank BioProject ID PRJNA242989.

References

- Alzohairy A, Sabir J, Gyulai G, Younis R, Jansen RK, Bahieldin A (2014) Environmental stress activation of plant long-terminal repeat retrotransposons. *Funct Plant Biol* 41:557–567
- Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. *Ann Bot* 95:127–132
- Bremer B, Eriksson T (2009) Time tree of Rubiaceae: phylogeny and dating the family, subfamilies, and tribes. *Int J Plant Sci* 170:766–793
- Bucher E, Reinders J, Mirouze M (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. *Curr Opin Plant Biol* 15:503–510
- Carrier G, Santoni S, Rodier-Goud M, Canaguier A, Kochko A, Dubreuil-Tranchant C, This P, Boursiquot JM, Le Cunff L (2011) An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. *Am J Bot* 98:e13–e15
- Carrier G, Le Cunff L, Dereeper A, Legrand D, Sabot F, Bouchez O, Audeguin L, Boursiquot JM, This P (2012) Transposable elements are a major cause of somatic polymorphism in *Vitis vinifera* L. *PLoS One* 7:10
- Casacuberta E, Gonzalez J (2013) The impact of transposable elements in environmental adaptation. *Mol Ecol* 22:1503–1517
- Cavallini A, Natali L, Zuccolo A, Giordani T, Jurman I, Ferrillo V, Vitacolonna N, Sarri V, Cattonaro F, Ceccarelli M, Cionini PG, Morgante M (2010) Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. *Theor Appl Genet* 120:491–508
- Chaparro C, Gayraud T, de Souza RF, Domingues DS, Akaffou S, Laforga Vanzela AL, Kochko A, Rigoreau M, Crouzillat D, Hamon S, Hamon P, Guyot R (2015) Terminal-repeat retrotransposons with GAG domain in plant genomes: a new testimony on the complex world of transposable elements. *Genome Biol Evol* 7:493–504
- Chevalier A (1942) Les caféiers du globe II: Iconographie des caféiers sauvages et cultivés et des Rubiacées prises pour des caféiers. In: Lechevalier P (ed) *Encyclopédie Biologique*, Paris
- 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
- Denoëud 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 I, Leclercq J, Lepellet 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, Victor A, Albert V, Wincker P, Lashermes P (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345:1181–1184
- Dereeper A, Guyot R, Tranchant-Dubreuil C, Anthony F, Argout X, de Bellis F, Combes MC, Gavory F, de Kochko A, Kudrna D, Leroy T, Poulain J, Rondeau M, Song X, Wing R, Lashermes P (2013) BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution. *Plant Mol Biol* 83:177–189
- Dias ES, Hatt C, Hamon S, Hamon P, Rigoreau M, Crouzillat D, Carareto CM, 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
- Dušková E, Kolář F, Sklenář P, Rauchová J, Kubešová M, Fér T, Suda J, Marhold K (2010) Genome size correlates with growth form, habitat and phylogeny in the Andean genus *Lasiocephalus* (Asteraceae). *Preslia* 82:127–148
- Dvořák J (2009) Triticeae genome structure and evolution. In: Muehlbauer JG, Feuillet C (eds) *Genetics and genomics of the Triticeae*. Springer, New York, pp 685–711
- Eilam T, Anikster Y, Millet E, Manisterski J, Sag-Assif O, Feldman M (2007) Genome size and genome evolution in diploid Triticeae species. *Genome* 50:1029–1037
- Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res* 16:1252–1261
- Hřibová E, Neumann P, Matsumoto T, Roux N, Macas J, Doležel J (2010) Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *BMC Plant Biol* 10:204
- Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Juarez MJA, Simpson J, Fernandez-Cortes A, Arteaga-Vazquez M, Gongora-Castillo E, Acevedo-Hernandez G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Perez SA, de Jesus Ortega-Estrada M, Cervantes-Luevano JL, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L (2013) Architecture and evolution of a minute plant genome. *Nature* 498:94–98
- Ito H (2013) Small RNAs and regulation of transposons in plants. *Genes Genet Syst* 88:3–7
- Ito H, Kakutani T (2014) Control of transposable elements in *Arabidopsis thaliana*. *Chromosome Res* 22:217–223
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467
- Kiehn M (1995) Chromosome survey of the Rubiaceae. *Ann Mo Bot Gard* 82:398–408
- Kinoshita T, Seki M (2014) Epigenetic memory for stress response and adaptation in plants. *Plant Cell Physiol* 55:1859–1863
- Knight CA, Beaulieu JM (2008) Genome size scaling through phenotype space. *Ann Bot* 101:759–766
- Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. *BMC Bioinform* 7:474
- Kumar A, Bennetzen JL (1999) Plant retrotransposons. *Annu Rev Genet* 33:479–532
- Lee SI, Kim NS (2014) Transposable elements and genome size variations in plants. *Genomics Inform* 12:87–97
- Lisch D (2013) How important are transposons for plant evolution? *Nat Rev Genet* 14:49–61
- Llorens C, Munoz-Pomer A, Bernad L, Botella H, Moya A (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol Direct* 4:41
- Macas J, Neumann P, Navratilova A (2007) Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genom* 8:427
- Martin G, Paris A, Samar M, Keller J, Salmon A, Novak P, Macas J, Ainouche A (2016) Dramatic lineage-specific accumulation of retrotransposons versus Simple Sequence Repeats across the last 10 million years in Mediterranean and African lupin genomes (*Lupinus*; Fabaceae). In: *International Congress on Transposable elements*, Saint Malo, France
- Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res* 11:1660–1676

- Michael TP, Jackson S (2013) The first 50 plant genomes. *Plant Genome* 6:1–7
- Middleton CP, Stein N, Keller B, Kilian B, Wicker T (2013) Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. *Plant J* 73:347–356
- Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carlson JE, Nelson CD, Davis JM (2009) Evolution of genome size and complexity in *Pinus*. *PLoS One* 4:e4332
- Noirot M, Poncet V, Barre P, Hamon P, Hamon S, De Kochko A (2003) Genome size variations in diploid African *Coffea* species. *Ann Bot (Lond)* 92:709–714
- Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hallman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Kaller M, Luthman J, Lysholm F, Niittyla T, Olson A, Rilakovic N, Ritland C, Rossello JA, Sena J, Svensson T, Talavera-Lopez C, Theissen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature* 497:579–584
- Pagan HJ, Macas J, Novak P, McCulloch ES, Stevens RD, Ray DA (2012) Survey sequencing reveals elevated DNA transposon activity, novel elements, and variation in repetitive landscapes among vesper bats. *Genome Biol Evol* 4:575–585
- Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposon-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* 16:1262–1269
- Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21:i351–i358
- Ramachandran D, Hawkins JS (2016) Methods for accurate quantification of LTR-retrotransposon copy number using short-read sequence data: a case study in Sorghum. *Mol Genet Genomics*
- Razafinarivo N, Rakotomalala JJ, Brown SC, Bourge M, Hamon S, De Kochko A, Poncet V, Dubreuil-Tranchant C, Couturon E, Guyot R, Hamon P (2012) Geographical gradients in the genome size variation of wild coffee trees (*Coffea*) native to Africa and Indian Ocean islands. *Tree Genet Genomes* 8:1345–1358
- Razafinarivo NJ, Guyot R, Davis AP, Couturon E, Hamon S, Crouzilat D, Rigoreau M, Dubreuil-Tranchant C, Poncet V, De Kochko A, Rakotomalala JJ, Hamon P (2013) Genetic structure and diversity of coffee (*Coffea*) across Africa and the Indian Ocean islands revealed using microsatellites. *Ann Bot* 111:229–248
- Renny-Byfield S, Chester M, Kovarik A, Le Comber SC, Grandbastien M-A, Deloger M, Nichols RA, Macas J, Novak P, Chase MW, Leitch AR (2011) Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Mol Biol Evol* 28:2843–2854
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell 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
- Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863–864
- Schulman AH, Gupta PK, Varshney RK (2004) Organization of retrotransposons and microsatellites in cereal genomes. In: Gupta PK, Varshney VR (eds) *Cereal genomics*. Kluwer Academic, Dordrecht, pp 83–118
- Sergeeva EM, Afonnikov DA, Koltunova MK, Gusev VD, Miroshnichenko LA, Vrána J, Kubaláková M, Poncet C, Sourdille P, Feuillet C, Doležel J, Salina EA (2014) Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. *Plant Genome* 7:1–16
- Slovak M, Vit P, Urfus T, Suda J (2009) Complex pattern of genome size variation in a polymorphic member of the Asteraceae. *J Biogeogr* 36:372–384
- Sonnhammer ELL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis (reprinted from *Gene Combis*, vol 167, pg GC1-GC10, 1995). *Gene* 167:GC1–GC10
- Stoffelen P, Noirot M, Couturon E, Anthony F (2008) A new caffeine-free coffee from Cameroon. *Bot J Linn Soc* 158:67–72
- Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genom* 8:132
- Todorovska E (2007) Retrotransposons and their role in plant-Genome evolution. *Biotechnol Bioelectron Equip* 21:294–305
- Tosh J, Dessein S, Buerki S, Groeninckx I, Mouly A, Bremer B, Smets EF, De Block P (2013) Evolutionary history of the Afro-Madagascan *Ixora* species (Rubiaceae): species diversification and distribution of key morphological traits inferred from dated molecular phylogenetic trees. *Ann Bot* 112:1723–1742
- 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
- Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N (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* 59:712–722
- Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35:W265–W268