Author Version

Nouroz F, Noreen S, Ahmad H, Heslop-Harrison JS. 2017. The landscape and structural diversity of LTR retrotransposons in Musa genome. Molecular Genetics and Genomics First Online: 10 June 2017

DOI: 10.1007/s00438-017-1333-1.

The landscape and structural diversity of LTR Retrotransposons in *Musa* genome

Faisal Nouroz[1,2], Shumaila Noreen[3], Habib Ahmad[4] and JS Pat Heslop Harrison[1]

[1]Molecular Genetics Laboratory, Department of Biology, University of Leicester, UK
[2]Bioinformatics Laboratory, Department of Botany, Hazara University Mansehra, Pakistan
[3]Molecular Genetics Laboratory, Department of Genetics, University of Leicester, UK
[4]Genetics Laboratory, Department of Genetics, Hazara University Mansehra, Pakistan
Corresponding author e-mail: faisalnouroz@gmail.com

Abstract

Long terminal repeat retrotransposons are major drivers of genome evolution and diversity, mostly localized in heterochromatic regions of chromosomes. *Musa* is an important fruit crop and also used as a starchy vegetable in many countries. BAC sequence analysis by dot plot was employed to investigate the LTR retrotransposons from *Musa* genomes. Fifty intact LTR retrotransposons from selected *Musa* BACs were identified by dot plot analysis and further BLASTN searches retrieved 153 intact copies, 61 truncated and a great number of partial copies/remnants from GenBank database. LARD-like elements were also identified with several copies dispersed among the *Musa* genotypes. The predominant elements were the LTR retrotransposons Copia and Gypsy, while Caulimoviridae (pararetrovirus) were rare in *Musa* genome. PCR amplification of reverse transcriptase (RT) sequences revealed their abundance in almost all tested *Musa* accessions and their ancient nature before the divergence of *Musa* species. The phylogenetic analysis based on RT sequences of *Musa* and other retrotransposons clustered them into Gypsy, Caulimoviridae and Copia lineages. Most of the *Musa* related elements clustered in their respective groups, while some grouped with other elements indicating homologous sequences. The present work will be helpful to understand the LTR retrotransposons landscape, their structural features, annotation and evolutionary dynamics in *Musa* genome.

Keywords: *Musa*, LTR retrotransposons, Copia, Gypsy, Biodiversity, Evolutionary relationship.

Introduction

Banana and plantains, the fourth most important tropical crop of the world, are herbaceous monocotyledonous plants of genus *Musa* of family *Musaceae* and order Zingiberales (Tomlinson 1969). There are more than 1000 banana cultivars with a high genomic diversity and variability with most cultivated species as

triploids with few as diploid and tetraploid genotypes (Heslop-Harrison and Schwarzacher 2007). The genus *Musa* is divided into four sections on the basis of morphology and chromosomes numbers as Eumusa (n=11), Rhodochlamys (n=11), Callimusa (n=9/10) and Australimusa (n=10). Edible bananas belong to section Eumusa and are mostly sterile, parthenocarpic, triploid (2n=3x=33) hybrids (along with a few diploids and tetraploids) from *Musa acuminata* (A-genome) alone or in combination with B-genome diploid *M. balbisiana* (Perrier et al. 2011). Cultivars have multiple origins from cultivated and wild cultivars by hybridisation (Hippolyte et al. 2012). Most cooking types are inter-specific hybrids (AAB/ABB), while sweet dessert bananas are triploid *M. acuminata* (AAA) (Pollefeys et al. 2004; Heslop-Harrison 2011).

Transposable elements (TEs) represent a very diverse group of sequences that are classified into two major classes (Class I retrotransposons and Class II DNA transposons) based on their mode of transposition. The Class I elements are further classified into super-families such as Copia, Gypsy, Retroviruses, Caulimovirus and Bel-Pao. Copia and Gypsy super-families are most predominant in plant and fungal genomes. Based on presence or absence of *gag-pol* protein coding domains, they are classified as complete (autonomous) or incomplete (non-autonomous) elements. Large retrotransposon derivatives (LARDs) are non-autonomous elements considered as deletion derivatives of autonomous LTR retrotransposons (Wicker et al. 2007; Defraia and Slotkin 2014; Nouroz 2015). Caulimoviruses (pararetroviruses) belong to Caulimoviridae superfamily, which replicate in plants via an RNA intermediate evolved from LTR retroelements (Bousalem et al. 2008; Llorens et al. 2011). Among TEs, the major proportion in plants is represented by long terminal repeat (LTR) retrotransposons (REs), which reverse transcribe their RNA to generate DNA copy integration to new host sites. The LTR retrotransposons (LTR REs) in plants display 4-6 bp target site duplications (TSDs), few hundred bp to several kilobases LTRs, exhibit primer binding sites (PBS) and polypurine tract (PPT) at 5′ and 3′ respectively (Eickbush and Jamburuthugoda 2008; Wicker et al. 2007; Nouroz et al. 2015). Previous studies revealed that Caulimoviruses/pararetroviruses have evolved from LTR REs.

Genome expansions in various organisms are the consequences of both increase in TEs copy numbers and types from different superfamilies (Du et al. 2010; Zhang et al. 2014). Whole genome sequencing has explored the ways to identify and characterize TEs in sequenced genomes and it was demonstrated that the half of the *Musa* genome is made up of TEs with LTR REs as the most dominant elements (>27.76%), followed by long interspersed elements (LINEs; 5.5%). The class 2 DNA TEs are rare (1.3%) in *Musa* and are mostly represented by hAT, Harbinger and Mutator superfamilies (D'Hont et al. 2012).

The LTR REs were investigated in many eukaryotic genomes and after developing the SSR, SSAP, IRAP, REMAP, ISSR and RAPD techniques, it becomes more feasible to study the diversity and landscape of LTR REs in various organisms (Schulman et al. 2012; Gyorgy et al. 2013; Izzatullayeva et al. 2014). In the recent years these transposon based markers are utilized in several plant genomes to study the biodiversity of plants like wheat (Queen et al. 2004), *Rhodiola rosea* (Gyorgy et al. 2013), Sugar beet (Izzatullayeva et al. 2014), *Brassica* (Nouroz et al. 2015) and several other plants. In the last decade, due to their major role in genome evolution and duplication, the LTR REs were investigated in various genomes (Schulman et al. 2012).

The present study was conducted to identify the LTR retrotransposons in *Musa* BAC sequences, to study their structural diversity, evolutionary relationships and distribution in various *Musa* genotypes.

Material and Methods

Dot plot identification of LTR retrotransposons from *Musa* BACs

The present study involved an approach for the identification of LTR retrotransposons based on the dot plot comparison of BAC/genomic sequences against themselves. The *Musa* BACs were retrieved from NCBI database before June, 2014 and only BACs with LTR REs were selected for further analysis. Initially the candidates of full length elements were identified by running each BAC genomic sequence against itself in dot plot analysis in the Dotter program (Sonnhammer and Durbin 1995). The central diagonal line extending from one corner of the dot plot to the diagonally opposite corner represents the homology of the sequence. The LTRs on both termini were represented by 2 small diagonal lines at opposite corners indicating 5´and 3´ LTRs. The boundaries of the elements were defined in BAC sequences, numbers of nucleotides in LTRs of each element were counted and TSDs were characterized by visual inspection.

Computational analysis and data mining for LTR retrotransposons

The intact or full length elements identified by dot plot analyses were further blasted against the *Musa* Nucleotide Collection (nr/nt) database (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch) available in NCBI. The searches for LTR retrotransposons were performed in several steps to identify the intact, truncated, partial elements, solo LTRs and remnants. The intact or full length elements were defined as elements having both LTRs and internal *gag-pol* genes. In the second step, the complete elements were used as query to find the full length copies, truncated elements, partial or deleted elements and remnants, which were defined with small modifications according to the recommendations of Ma et al. (2004) and Nouroz et al. (2015). For the

identification of conserved *gag-pol* gene encoding proteins, the nucleotide sequences were investigated in 'Conserved Domain Database' (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) implemented in NCBI. The PBS and PPT motifs were detected in the LTR_FINDER by using parameter 'Predict PBS by using *Zea mays* and *Oryza sativa* tRNA database'.

Characterization and naming of LTR retrotransposons

The Repbase (Jurka et al. 2005) and Gypsy databases (Llorens et al. 2011) were used as reference databases to characterize the retrotransposons on homology basis. Elements that failed to be characterized by homology searches against TE databases were characterized by visual inspection on the basis of their structural hallmarks such as TSDs, LTRs and organization of their *gag-pol* encoding protein domains such as integrase (INT), reverse transcriptase (RT), RNaseH (RH) and envelope (ENV). The retrotransposons were classified as Copia, if they displayed *pol* gene as 5'-INT-RT-RH-3', Gypsy as 5'-RT-RH-INT-3', Retroviruses as 5'-RT-RH-INT-ENV-3' and LARDs if they exhibit TSDs, LTRs and large non-coding internal regions without any known *gag-pol* gene domain. The names of the elements were given on the recommendations of Capy (2005) such as ***MaGYP1***, where 1st letter '***M***' indicates genera *Musa*, second letter '***a***' indicates specie *acuminata*, 3 letters '***GYP***' represent the superfamily (Gypsy) and the number '**1**' indicates the number of the identified element.

Polymerase chain reactions (PCRs)

DNA samples from 48 *Musa* accession/genotypes (Table 1) were analysed for presence of LTR REs. The degenerate primer pairs (Table 2) designated as reverse transcriptase amplification polymorphism (RTAP) markers were designed from RT regions with online program Primer3 (http://frodo.wi.mit.edu/primer3/). PCR was conducted in 15 µl reaction mixture with 50-75 ng/µl genomic DNA + 10X buffer A (Kapa Biosystems, UK) + 1.0 mM $MgCl_2$ + 2-2.5 mM dNTP (YORKBIO) + 10 pmoles of each primer (SIGMA-ALDRICH) + 0.5-1 U of 5U/µl Taq polymerase (Kapa Biosystems, UK). The thermal cycling conditions were 3 min denaturation at 94˚C; 35 cycles of 1 min denaturation at 94˚C, 1 min annealing at 52-64˚C (depending on primers) and 1 min extension at 72˚C; final 5 min extension at 72˚C. PCR products were separated by electrophoresis in 1% agarose gel with TAE buffer and gels were stained with 1-2 µl ethidium bromide for the detection of DNA bands under UV illumination.

Multiple sequence alignment and phylogenetic analysis

The RT sequences from 77 known LTR REs (Supplementary Table) of Copia, Gypsy and Pararetroviruses superfamilies were collected from Gypsy database (Llorens et al. 2011). The 33 RT sequences

(~180-220 aa) were taken from identified *Musa* LTR retrotransposons and were aligned with 2 known elements (*Ty1-Copia*, *Ty3-Gypsy*) in CLUSTALW multiple alignment implemented in BioEdit, which were visually inspected and edited manually, if needed. Small insertions and deletions were removed and frame shifts were introduced. All 5´-3´ oriented RT regions were included in alignment, even if they have stop codons or frame shift mutations. The phylogenetic analyses were performed by constructing the un-rooted Neighbour-Joining trees with 1000 bootstrap replicates implemented in MEGA5 program (Tamura et al., 2011). The evolutionary distances were computed using the p-distance method. The methodology is summarized in Fig. 1.

Results

The LTR retrotransposons landscape in *Musa*

Fifty elements from 30 *Musa* BACs (column 3 of Table 3) were identified by dot plot analysis by plotting each BAC sequence against itself (Fig. 2). Of the fifty identified elements, 20 belonged to Gypsy, 19 to Copia, 1 to Caulimoviridae (Pararetroviruses) and 10 to LARD-like elements (Table 3). The search was extended by using these elements as query in BLASTN searches against *Musa* Nucleotide Collection (nr/nt) database of NCBI and all full length, truncated and partial copies were counted. A total of 16246 elements and their partial fragments from Copia, Gypsy, Caulimoviridae and LARDs were identified, of which 153 were intact (full length) elements with 58 from Gypsy, 48 from Copia, 1 Pararetrovirus and 46 from LARD-like elements. A total of 61 truncated elements, 635 partial elements, 258 solo LTRs and 15140 remnants were counted from *Musa* Nucleotide Collection (nr/nt) database deposited in NCBI database.

General Characteristics of *Musa* Gypsy retrotransposons

The Gypsy elements ranged in sizes from 3015–17804 bp, where smallest non-autonomous *MaGYP6* was 3015 bp large, while the largest autonomous *MbGYP20* was 17804 bp having nested structure (Table 3). Around 90% elements were found terminated by 5 bp TSDs, while rest (10%) showed 4 bp TSDs. The LTRs of the Gypsy ranged in sizes from 264–1105 bp with average size of 450-550 bp (Table 3). The *M. acuminata* BAC clone sequence 'AC226035.1' showed maximum copy numbers of retrotransposons among the investigated BACs, where 4 Copia, 3 Gypsy and 2 LARD-like elements were identified covering a total size of ~60 kb (58%) of 105 kb long BAC (Fig. 2). Another *M. acuminata* BAC sequence 'AC226048.1' harboured six Gypsy elements (*MaGYP8–MaGYP13*; Table 3) covering a total of ~31 kb (24%) of 134.5 kb BAC sequence. The partial copies or remnants from these elements further increased their size and percentage.

Structural features of the Gypsy Retrotransposons in *Musa*

The structural features of all Gypsy elements identified in present study were analysed in detail. *MaGYP1* (4982 bp) was flanked by 5 bp TSDs and 505 bp LTRs (Table 3). *MaGYP2* was identified as a non-autonomous element (3.8 kb) flanked by 5′-543/527-3′ bp LTRs. *MaGYP3* and *MaGYP4* showed high structural homology with sizes of 4.5 and 4.6 kb respectively and incorporated Transcriptional regulator (TR), Haemthiolate proteins (HP), Tymovirus proteins (TVP) and Hepadnavirus proteins (HVP) like additional proteins (Table 4). *MaGYP5* (6.25 kb) was found to be flanked by LTRs of 586 bp and displayed internally deleted *gag-pol* region with additional domains not common to REs (Table 4). *MaGYP6*, the smallest non-autonomous Gypsy was only 3 kb including 655 bp LTRs (Table 3). *MaGYP7* and *MaGYP9* (5.3 kb) were flanked by LTRs of 411-519 bp with small insertions in 5′LTR. *MaGYP8* (5.9 kb) and *MaGYP10* (5.4 kb) displayed canonical *gag-pol* polyproteins structure (Table 4).

MaGYP12 (5.76 kb) terminated by 671 bp LTRs (Table 3) encoded *gag-pol* gene domains as 5′-AP-RT-RH-INT-3′ with an additional Zinc knuckle (ZK) domain (Fig. 3a). *MaGYP13* (5.4 kb) was flanked by 1062 bp LTRs, displaying only RT domain with homology to the RT of Non-LTR retroelements. A 4.9 kb *MbGYP15* was identified in *M. balbisiana* flanked by 884 bp LTRs and displayed *gag-pol* genes (Table 4). The elements *MbGYP16*, *MaGYP17* and *MbGYP18* were 4.0, 6.1 and 7.4 kb large terminated by 4 bp TSDs, where *MaGYP17* encoded an additional CSP protein domain (Fig. 3a). *MbGYP19* (7.36 kb) was flanked by 1105 bp 5′LTR and 883 bp 3′LTR with AT rich insertion next to the downstream of 5′LTR.

Structural features of *Musa* Copia elements

Nineteen intact Copia elements identified by dot plot analyses of *Musa* BACs were investigated in detail. *MaCOP1* and *MbCOP19* showed homologies in their molecular structures with 5.3 and 5.2 kb sizes, flanked by LTRs of 605 and 592 bp respectively (Table 3) encoding the conserved protein domains of 5′-INT-RT-RH-3′ (Table 4). *MaCOP2* (4.8 kb) displayed a PBS, canonical *gag-pol* genes and PPT upstream to 3′LTR. *MaCOP4* (4.0 kb) was identified from *M. acuminata* with only INT and an additional ZK domain in its structure (Table 4). *MaCOP5* and *MaCOP17* with structural homologies and lengths (8.1 kb) were flanked by LTRs of 5′-1285/1201-3′ and 5′-1000/1324-3′ bp respectively. *MaCOP6* and *MaCOP9* though ~7.0 kb in sizes lacked the *pol* polyproteins except RT. *MaCOP7* (5.0 kb) flanked by 5′-144/149-3′ bp LTRs have shown the shortest LTRs in present study. *MaCOP8* and *MaCOP14* ranged 6.0 kb in size, flanked by 5′-499/500-3′ and 5′-492/548-3′ bp

LTRs respectively. *MaCOP10* and *MaCOP11* were 8.7 and 8.4 kb large elements, where *MaCOP10* was found terminated by 1597 bp LTRs while *MaCOP11* (Fig. 3a) was flanked by 5′-1494/1388-3′ LTRs. *MaCOP12* and *MaCOP13* were 7.1 and 5.9 kb large elements, flanked by 5′-1238/1132-3′ and 5′-573/548-3′ bp respectively. *MbCOP15* and *MaCOP16* displayed the ORF encoding the *gag-pol* products (GAG-INT-RT-RH). *MbCOP18* (9.8 kb) investigated in *M. balbisiana* accession 'AC226052.1' was terminated by long (5′-1415/1396-3′ bp) LTRs (Fig. 3a; Table 3).

Structural features of *Musa* Caulimovirus (Pararetrovirus)

An 11.1 kb large element was investigated from *M. acuminata* BAC sequence (AC226046.1), flanked by largest LTRs (3.8 kb) investigated in present study. The element was named as *MaCVI* (*Musa acuminata* chromovirus). *MaCVI* was characterized by having 3.8 kb LTRs, an internal region containing the PBS, *pol* gene encoding the AP, RT and RH domains and a PPT adjacent to 3′LTR with two additional protein domains (Fig. 3a). The PBS of *MaCVI* was found different from all other Copia and Gypsy elements described here with $tRNA_{Gly}$ (unusual RNA type). A 15 bp PPT upstream to 3′LTR was found with different sequence structure compared to other elements (Table 4).

Structural features of LARD-like elements

Despite of several autonomous retroelements, non-autonomous LARD-like elements were also characterized (Table 3) by having 4-5 bp TSDs, LTRs, exhibiting PBS/PPT motifs and internal non-coding regions. *MaLAR1* harboured in *M. acuminata* accession (AY484588.1) was 4564 bp large, flanked by 4 bp TSD and 447 bp LTRs (Fig. 3a). *MbLAR2* (4428 bp) is another homologue of *MaLAR1* identified from *M. balbisiana* flanked by 445 bp LTRs. *MbLAR3* shared structural homology with *MbLAR5* and *MbLAR6* with a size of 4.4 kb, displaying LTRs of 382-383 bp. *MaLAR4* was 4.3 kb including the LTRs (5′-607/611-3′) and flanked by 5 bp imperfect TSDs. *MaLAR7* and *MaLAR8* showed similar structural features having 4.5 kb size. *MbLAR9* identified from *M. balbisiana* BAC (AC186754.1) was 7.7 kb element with no detectable PBS and PPT motifs. It displayed an unknown insertion and a non-autonomous hAT element with two additional solo LTRs (Fig. 3a). *MaLAR10* was smallest LARD-like element (4 kb) studied here displaying LTRs of 5′-974/984-3′ bp.

Nested LTR retrotransposons structures in *Musa*

Two Gypsy-like LTR REs identified in present study have shown complex nested structures, where 1 or more TEs or unidentifiable sequences were found inserted within the element. *MaGYP14* (11.6 kb) was found flanked by 624 bp LTRs, exhibiting *pol* gene domains as 5′-AP-RT-RH-INT-3′, with ~1.7 kb additional transcriptional regulator protein and a GC rich unknown insertion of ~2.3 kb (Table 3). The most complex structure was observed in *MbGYP20* (17.8 kb), which showed a nested structure of 3 insertions and 2 solo LTRs (Fig. 3b). One insertion was 9.6 kb Gypsy, where another unknown insertion of 4.5 kb was inserted in opposite orientation. *MaCOP3* (16.2 kb) was found inserted in *M. acuminata* BAC (AC226035.1) displaying a complex nested structure of LTR REs (Fig. 3b). A 5.3 kb *MaCOP1* element was inserted in *MaCOP3* starting from 3203-8492 bp. The outer element *MaCOP3* (10.9 kb) showed an insertion towards 5′ LTR and was flanked by 5′-338/299-3′ bp LTRs, while the inserted element *MaCOP1* was terminated by 605 bp LTRs.

The *gag-pol* polyprotein organization in *Musa* LTR retrotransposons

The organization of *gag-pol* protein domains of Gypsy REs revealed 2 patterns (canonical and defective) and 14 sub-patterns of domain organizations (Table 4). The canonical Gypsy domain structure (5′GAG-RT-RH-INT-3′) was observed in 6 elements. A single element *MaGYP6* encoded a *gag* protein only, *MaGYP2* showed *gag* and a transcriptional regulator (TR) domain, *MaGYP13* encoded RT only and *MbGYP18* displayed 5′-AP-RT-3′. The five elements (*MaGYP7, MaGYP9, MaGYP11, MaGYP15* and *MaGYP16)* lacked the INT domain, while RT and RH domains were absent in *MaGYP1*. Three elements *MaGYP8*, *MaGYP10* and *MaGYP17* showed similar domain pattern (5′-GAG-AP-RT-RH-INT-CHR-3′) with one or other extra domain. *MbGYP20* showed a complex organization of protein domains due to nested retrotransposons structures as 5′-GAG-AP-(3′-CMV-RH-DUF-5′)-DUF-CMV-RT-RH-CHR-3′ (Fig. 3b; Table 4).

The Copia *gag-pol* protein structural organization revealed seven sub-patterns of the two main patterns (canonical and defective). The canonical pattern of Copia protein domain organization is 5′-GAG-INT-RT-RH-3′, observed in almost 90% of the elements with one less or extra domain. *MaCOP6* and *MaCOP11* encoded only a RT domain, while *MaCOP17* showed a slightly different pattern 5′-GAG-RT-RH-MT-3′, where additional Mannosyl transferase (MT) protein was replaced with INT. A nested LTR RE *MaCOP3* showed a complex pattern 5′-GAG-AP-INT-RT-RH/GAG-INT-RT-RH-3′, where two sets of proteins domains were detected encoded by 2 different Copia elements. The other 13 elements showed the canonical Copia protein organization (5′-INT-RT-RH-3′) (Table 4). All the LARDs elements were investigated for their *gag-pol* genes but no identifiable *gag-pol* gene protein domains were detected.

PBS and PPT pattern of *Musa* retrotransposons

The 15-18 bp PBS located downstream to 5´LTR and its reverse compliment PPT located adjacent to the 3´LTR were detected (Table 4) by scanning the LTR RE against *Zea mays* tRNA database. A total of 80% and 75% elements showed the presence of 14-18 bp PBS and 15 bp PPT respectively, 10% showed PPT only, while remaining 10% failed to detect any PBS or PPT by scanning tRNA of *Zea mays*, which were than scanned against *Oryza sativa* tRNA database and their PBS and PPT were obtained (Table 4). *MaGYP2* lacked PPT, while PBS was not detected in *MaGYP6* and *MbGYP19*. Seven different tRNA types were investigated in Gypsy elements with $tRNA_{Met,}$ as most frequent type present in 30% of the elements followed by $tRNA_{Asn,}$ found in 20% elements (Table 4). The PBS and PPT structures of Copia elements revealed that 95% elements showed 14-18 bp PBS except *MaCOP14* (Table 4). Eight different types of tRNA types were observed in all Copia elements investigated, with $tRNA_{Met}$ as most common tRNA type detected in 40% of the elements; followed by $tRNA_{Val,}$ observed in 20% elements. All the other 6 types of tRNA contributed 5% of the tRNA type. PPT adjacent to the 3´LTR was detected in 90% of all Copia elements except *MaCOP6* and *MaCOP14* (Table 4). The PBS and PPT motifs in LARDs revealed that out of 10 individual elements, only 2 elements (20%) *MaLAR7* and *MaLAR8* displayed the PBS and PPT motifs in their 5´and 3´LTRs respectively (Table 4).

PCR amplification of retrotransposons in *Musa* genomes

The presence and abundance of Gypsy elements in 48 diverse *Musa* genotypes were determined by reverse transcriptase amplification polymorphism (RTAP) in using PCR. The primers were designed from conserved RT regions (Table 2). Of the 48 *Musa* genotypes (Table 1), 6 *M. acuminata* (AA), 6 *M. balbisiana* (BB), 3 hybrids (AB), 8 triploid *M. acuminata* (AAA), 19 (AAB) and 6 (ABB) allotriploids were used to analyze the presence of LTR REs in them. The primer pair MaGYP8F/R (Table 2) was used to amplify 684 bp RT region of *MaGYP8* family. The products were amplified from all *M. acuminata* (AA) (Calcutta 4, Sannachenkadali, Pisanglilin, Kadali, Matti, Cherukadali), *M. balbisiana* (BB) (PKW1, PKW2, Javan, Klutuk, Tani, Batu), AB genomes (Njalipovan, Adukkan, Padalamukili), AAA (Manoranjitham, Grand Nain, Gross Michel, Greenred, Red, Monsmari, Robusta, Dwarf Cavendish), AAB (Motta Povan, Karimkadali, Perumadali, Kunoor Ettan, Palyamcodan, Mysoreettan, Krisnavazhai, Poovan, Doothsagar, Charapadati, Kumbillakannan, Velipadati, Vellapalayamcodan, Ettapadati, Padati, Chinali, Nendran, Poomkalli, Kamaramasengi) and ABB genotypes (Kosta Bontha, Peyan, Kanchikela, Boothibale, Monthan, Karpooravali) (Fig. 4a). This showed the ancient nature

of this element, which was present in a common ancestor predating the separation of A and B-genome *Musa*. The RT based amplification polymorphism of *MaGYP12* family revealed its amplification from all 47 *Musa* accessions except *M. acuminata* (Calcutta 4), where no amplification suggests its absence or recent swept from the genome (Fig. 4b). The 835 bp RT regions of *MaGYP17* family were amplified from all the 48 *Musa* accessions (Fig. 4c) by primer pair MaGYP17F/R. The amplification of various Gypsy elements from *Musa* genotypes revealed their distribution in almost all regardless of A or B-genome specificity (Fig. 4a-c).

The availability of various members of Copia superfamilies were investigated in 48 *Musa* accessions (Table 1) by PCR analysis. The primer pair MaCOP5F/R (Table 2) amplified a 744 bp RT region in all *M. acuminata* (AA), *M. balbisiana* (BB), AB, AAA, AAB and ABB genotypes (Fig. 4d). A 964 bp *MaCOP8* RT genomic sequence was amplified in PCR by primer pair MaCOP8F/R from all 48 diploid and triploid *Musa* genotypes with weak and strong signals in various genotypes (Fig. 4e). The abundance of Caulimovirus named *MaCVI* was examined in various *Musa* genotypes by PCR analysis. The primer pair MACVIF/R was designed to amplify a 425 bp RT sequence, which revealed that the product was amplified from all *Musa* genotypes except *M. balbisiana* (BB) accession 'PKW2' (Fig. 4f).

Phylogenetic relationships of *Musa* and other plant LTR retrotransposons

The phylogenetic relationships of 33 RT sequences of *Musa* LTR REs and 2 known elements (Ty1-Copia, Ty3-Gypsy) were performed in MEGA5. Two main lineages separated the Copia, Gypsy/Caulimoviridae (Pararetrovirus) elements with 19 and 16 elements respectively (Fig. 5). The Copia lineage is further resolved into two groups with 2 (*MaCOP6*, *MaCOP9*) and 17 elements in respective groups. Of the 17 Copia, the Ty1-Copia from *Saccharomyces cerevisiae* out-grouped from rest of *Musa* Copia. Some Copia elements clustered on same or sister branches due to high homologies in their RT sequences (Fig. 5). Of the Gypsy lineage, Ty3-Gypsy from *Saccharomyces cerevisiae* out-grouped from *Musa* Gypsy elements. *MACV1* from Caulimoviridae also out-grouped from *Musa* Gypsy elements. The RT sequences of *MACV1* and Gypsy indicated homology in their sequences, yet they are distinct from Copia. Most of the RT sequences from *M. acuminata* and *M. balbisiana* showed homology to each other and are resolved on same or sister branches (Fig. 5).

The evolutionary relationships of *Musa* and other organism based LTR REs were performed by constructing a phylogenetic tree of 110 RT sequences (Fig. 6), of which 33 were collected from *Musa* LTR REs of present study, while other 77 were from various organisms and were collected from Gypsy database (Supplementary Table). The evolutionary history was reconstructed by the un-rooted Neighbor-joining method with 1000 bootstrap replicates,

where strong bootstrap values supported the monophyletic origin of Gypsy and Copia retrotransposons, yet the three main lineages (shown by different colours and shapes in Fig. 6) separate the Gypsy, Copia and Caulimoviruses indicating distinct homology and no recombination between the sequences of these superfamilies. The Gypsy clustered 40, Copia 46 and Caulimoviruses 24 elements in their respective lineages. Of the Gypsy lineage, the *Ty3-Gypsy* from *Saccharomyces cerevisiae* and *Gypsy* element from *Drosophila melanogaster* out-grouped. Most *Musa* Gypsy elements clustered in *Musa* specific groups except few elements (Fig. 6) as *MaGYP12* clustered together with *CRM* of *Zea mays*, *MaGYP14* with *Gloin* of *Arabidopsis thaliana* and *MbGYP19* with *Cereba* of *Hordeum vulgare*. The caulimoviridae constituted close lineage to Gypsy lineage with 24 elements, where *PCSV* out-grouped and misfits near Copia elements. *MaCV1* formed a sister family with other caulimoviridae members as *BSOLV, CSSV, KTSV, BSGFV* and *BSVAV* (Supplementary Table). In Copia lineage, *Ty1-Copia* and *Ty2* from *Saccharomyces cerevisiae* out-grouped. In most cases, the *Musa* specific Copia clustered in their respective families, while others (*MaCOP2, MaCOP15, MaCOP17*) shared families with other members as *MaCOP2* grouped with *Araco* of *Arabidopsis thaliana*, *MaCOP15* with *Tork-4* of *Solanum lycopersicum* and *MaCOP17* with *TSI-9* of *Setaria italica* (Fig. 6).

Discussion

One of the major sources of genomic variations are the repetitive DNA sequences (Bennetzen, 2000; Biscotti et al. 2015). The genome of *Musa* is also rich in LTR REs belonging to Copia, Gypsy and Caulimovirus superfamilies (D'Hont et al. 2012; Davey et al. 2013). As the genome sequencing is progressing and updated, there is a need to discover the TEs especially LTR REs, which are major drivers of gene and genome evolution, and the BAC sequence analysis provides valuable reference sequences, uninfluenced by heterozygosity and presence of multiple copies throughout the genome. D'Hont et al. (2012) identified the composition of TEs in *M. acuminata* genotype DH-Pahang, finding Copia in high proportion followed by Gypsy and LINEs. The DNA transposons were very rare representing Harbinger, Mutator and hAT families; Menzel et al. (2015) found only 70 hAT elements although the related MITEs were amplified to much higher copy numbers. The most active DNA transposons identified from other angiosperms plants like Mariner, Harbinger and CACTA were very rare in *Musa*, although recent studies confirmed the abundance of Harbinger (Nouroz et al., 2016) and CACTA (Nouroz et al., 2017) elements in plants like *Brassica*. A study showed 26.85% of LTR REs in *Musa balbisiana* genotype 'PKW' with Copia and Gypsy as dominant superfamilies. The percentages of LINEs and DNA transposons were less and similar in A and B genome *Musa* (Davey et al. 2013).

The present study involved identification and description of Copia, Gypsy, Caulimoviruses and LARD-like elements in *Musa* BACs, as previous analyses were superficial or have focussed on selected repeats and revealed that 30-50% *Musa* genome is comprised of repetitive sequences (Hribova et al. 2010; D'Hont et al. 2012). The approach of comparative analysis of BAC sequences by dot plot was effective and highly informative to identify the LTR REs in the sequenced genome of *Musa*. This strategy helped in the identification of most of the elements present in *Musa* BAC sequences, which otherwise are not possible to precisely detect with other softwares such as LTR_FINDER, LTR_STRUC and LTR_harvest due to various genomic deformations of these retroelements structures. In the initial effort, 50 intact elements from three main superfamilies (Copia, Gypsy and Caulimoviruses) were identified. Further BLAST analysis using these full length elements retrieved a total of 153 intact elements from 6 Mbp of *Musa* BACs screened. The intact copies (listed in Table 3) covered 15-18% of the genome surveyed, which is further strengthening the investigations revealing high repetitive proportions found in the *Musa* genome analysis using short reads from 454 sequencing (Hribova et al. 2010) and BAC-end sequencing (Cheung and Town 2007). About 61 truncated copies, 635 partial copies, 258 solo LTRs and 16246 small fragments (remnants) were also identified; precise alignment of truncated or partial copies is not possible due to deletions and the high numbers, but their contribution to the *Musa* genome was counted. Such deleted elements and insertions in LTR REs are common in plants like rice (Ma et al. 2004) and *Arabidopsis* (Devos et al. 2002). Most of the deletions or insertions in the intact elements were bounded by terminal duplications of few bp. Such terminal duplications were observed around the deletions within retroelements from *Arabidopsis* (Devos et al. 2002). The numbers of partial copies, truncated elements and remnants (mentioned above) were very high in our study in comparison to the full length copies. The low copy number of full length LTR REs in comparison to partial or deleted copies were obvious from other plants such as only 583 full length LTR REs were identified from *Elaeis guineensis* genome (Beule at al. 2015). The full length elements of current study ranged in sizes from 4-17.8 kb with flanking LTRs of 149 bp to 3.8 kb. These findings are in accordance to the study of LTR REs in *Medicago truncatula*, where the elements ranged in size from 4-18.7 kb with similar sized LTRs (Wang and Liu 2008).

A Caulimovirus element (*MaCV1*) residing in *Musa* genome displayed the structural features common to Caulimoviruses present in many plant genomes including *Musa* and potato. Three families of such Pararetroviruses were isolated from potato genome and their distributions on chromosomes were studied by fluorescent in situ hybridization (Hansen et al. 2005). The RT alignment and phylogenetic analysis revealed that *MaCV1* formed sister lineage with Gypsy elements suggesting homology in their sequences, but detail analysis showed that both

followed different evolutionary pathways. In *Brassica*, the virus-like elements grouped with Gypsy lineage indicating their common ancestral origin but they also followed two different evolutionary pathways (Alix and Heslop-Harrison 2004). The clustering of most of the *Musa* related sequences in their respective families revealed separate line of evolutionary history, while in few cases the elements shared sequence similarity with other elements and thus were clustered with them. The domain organization of the elements also varied, consistent with earlier studies: Copia-like elements were 5′-AP-INT-RT-RH-3′, Gypsy-like elements 5′-AP-RT-RH-INT-3′, and caulimoviruses showed 5′-ORF-AP-RT-RH-3′ domain pattern (Hansen and Heslop-Harrison 2004; Wicker et al. 2007).

The LARD-like elements were frequent in *Musa* genome and their abundance indicated that like other LTR REs, they are major component of these genomes. Despite of lacking their internal coding domains, several copies were detected in *Musa* BACs. We cannot fully answer the question which LTR REs superfamily these LARDs belong to and who is borrowing them their coding domains (machinery) for transposition and integration to a new site. But comparing the elements with known TE sequences and structural homology, it was revealed that both Gypsy and Copia are the progenitors of these elements. The previous studies revealed that LARDs constitute major proportion of several genomes as identified in *Medicago truncatula* and *Brassica* (Wang and Liu 2008; Nouroz 2015).

In the present work, RT-PCR experiments have demonstrated that Gypsy and Copia retrotransposon families were amplified from all diploids and triploids *Musa* genotypes tested. This confirmed the ancient and conserved nature of RT sequences which evolved pre-separation of A and B *Musa* and transduplication of various *Musa* genomes. Similar ratios of RT were investigated in various members of family Asteraceae (Docking et al. 2006). The present study is evident that very few elements were species specific, either in A-genome *M. acuminata* or B-genome *M. balbisiana*, while the majority were present in both. The PBS and PPT motifs were detected in most elements, while in few either PBS or PPT was missing or might be deleted. The $tRNA_{Met}$ was the most commonly used type in both superfamilies as was investigated in *M. truncatula* (Wang and Liu 2008). Some of the retrotransposons of present study acquired an extra protein domain without any role in their transposition, such domains are harboured by retrotransposons in other plants (Havecker et al. 2004).

Conclusion

The present study described several novel LTR REs in *Musa* genome including their structural features, protein domain organization, pattern of PBS/PPT motifs, evolutionary dynamics and percentage in their host genome by their transduplication. The results indicated that individual LTR REs families have distinct behaviour,

genomic organizations and actively proliferating in their host genomes. This work provided references of single elements, with the description and annotation of major portions of retrotransposons, and a valuable advance in the quest to unravel the genomics of LTR REs in general and their evolutionary dynamics in *Musa* in particular.

Acknowledgements

The study was funded by Post quake Faculty Development Plan of Hazara University and Higher Education Commission of Pakistan. We are thankful to staff at University of Leicester, UK, who provided us technical assistance and all laboratory facilities during this work. The collection of 48 *Musa* genomic DNA was a gift from Professor Ashalatha (Asha) Nair, University of Kerala, India.

Conflict of Interests:

All the authors declared that no financial or other conflict of interests exists in publishing the manuscript.

References

Alix K, Heslop-Harrison JS (2004). The diversity of retroelements in diploid and allotetraploid *Brassica* species. Plant Mol Biol 54: 895-909. doi 10.1111/j.1365-313X.2008.03660.x

Bennetzen JL (2000). Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269. doi: 10.1146/annurev-arplant-050213-035811

Beulé T, Agbessi M, Dussert S, Jaligot E, Guyot R (2015). Genome-wide analysis of LTR-retrotransposons in oil palm. BMC Genomics 16: 795. doi: 10.1186/s12864-015-2023-1

Biscotti MA, Olmo E, Heslop-Harrison JS (2015). Repetitive DNA in eukaryotic genomes. Chromosome Research 23: 415-420.

Bousalem M, Douzery EJ, Seal SE (2008). Taxonomy, molecular phylogeny and evolution of plant reverse transcribing viruses (family Caulimoviridae) inferred from full-length genome and reverse transcriptase sequences. Arch Virol 153: 1085-1102

Capy P (2005). Classification and nomenclature of retrotransposable elements. Cytogenet Genome Res 110: 457-461. doi 10.1159/000084978

Cheung F, Town CD (2007). A BAC end view of the *Musa acuminata* genome. BMC Plant Biol 7: 29. doi: 10.1186/1471-2229-7-29

Davey MW, Gudimella R, Harikrishna JA, Sin LW, Khalid N, Keulemans J. 2013. A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids. BMC Genomics 14:683. doi**:** 10.1186/1471-2164-14-683

Defraia C, Slotkin, RK (2014**).** Analysis of retrotransposon activity in plants. Methods Mol Biol 1112:195-210. doi: 10.1007/978-1-62703-773-0_13

Devos KM, Brown JK, Bennetzen JL (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12: 1075-1079. doi: 10.1101/gr.132102

D'Hont A, Denoeud F, Aury J, Baurens F, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard MCDS et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488: 213-219. doi: 10.1038/nature11241

Docking TR, Saade FE, Elliot MC, Shoen DJ (2006). Retrotransposon sequence variation in four asexual plant species. J Mol Evol 62: 375-387. DOI: 10.1007/s00239-004-0350-y

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J (2010). Evolutionary conservation, diversity and specificity of LTR-Retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J, 63: 584-598. doi: 10.1111/j.1365-313X.2010.04263.x

Eickbush TH, Jamburuthugoda VK (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134: 221-234. doi: 10.1016/j.virusres.2007.12.010.

Gyorgy Z, Fjelldal E, Szabo A, Aspholm PE, Pedryc A (2013). Genetic diversity of golden root (*Rhodiola rosea* L.) in northern Norway based on recently developed SSR markers. Turk J Biol, 37: 655-660. doi: 10.3906/biy-1302-17

Hansen CN, Harper G, Heslop-Harrison JS (2005). Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). Cytogenet Genome Res 110: 559-565. doi:10.1159/000084989

Hansen CN, Heslop-Harrison JS (2004). Sequences and Phylogenies of Plant Pararetroviruses, Viruses and Transposable Elements. Adv Bot Res 41: 165-193.

Havecker ER, Gao X, Voytas DF (2004). The diversity of LTR retrotransposons. Genome Biol 5: 225. doi: 10.1186/gb-2004-5-6-225

Heslop-Harrison JS (2011). Genomics, Banana Breeding and Superdomestication. Proc. Int'l ISHS-Pro*Musa* Symp. on Global Perspectives on Asian Challenges. Acta Hort 897.

Heslop-Harrison JS, Schwarzacher T (2007). Domestication, genomics and the future for banana. Ann Bot 100: 1073-1084. doi: 10.1093/aob/mcm191

Hippolyte I, Jenny C, Gardes L, Bakry F, Riyallan R, Pomies V, Cubry P, Tomekpe K, Risteruci AM, Roux N et al. (2012). Foundation characteristics of edible *Musa* triploids revealed from allelic distribution of SSR markers. Ann Bot 109: 937-951. doi: 10.1093/aob/mcs010

Hribova E, Neumann P, Matsumoto T, Roux N, Macas J, Dolezel J (2010). Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. BMC Plant Biol 10: 204. doi:10.1186/1471-2229-10-204

Izzatullayeva V, Akparov Z, Babayeva S, Ojaghi J, Abbasov M (2014). Efficiency of using RADP and ISSR markers in evaluation of genetic diversity in suger beet. Turk J Biol 38: 429-438. doi:10.3906/biy-1312-35

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. doi 10.1159/000084979

Kapitonov VV, Jurka J (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9: 411-412. doi: 10.1038/nrg2165-c2

Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, Aguilar-Rodriguez J, Vicente-Ripolles M, Fuster G, Bernet GP et al (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39: D70-74. doi: 10.1093/nar/gkq1061

Ma J, Devos KM, Bennetzen JL (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14: 860-869. doi: 10.1101/gr.1466204

Menzel G, Heitkam T, Seibt KM, Nouroz F, Müller-Stoerme M, Heslop-Harrison JS, Schmidt T. (2015). The diversification and activity of hAT transposons in *Musa* genomes. Chromosome Research 22: 559–571. doi: 10.1007/s10577-014-9445-5

Nouroz F (2015). Large retrotransposon derivatives (LARDs) and Terminal repeat retrotransposons in miniature (TRIMs) in *Brassica* genomes. Int J Argric Appl Sci 7: 59-66.

Nouroz F, Noreen S, Heslop-Harrison JS (2015). Identification and characterization of LTR Retrotransposons in *Brassica*. Turk J Biol 39: 740-757. doi: 10.3906/biy-1501-77

Nouroz F, Noreen S, Heslop-Harrison JS (2016). Characterization and diversity of novel *PIF/Harbinger* DNA transposons in *Brassica* genomes. Pak J Bot 48(1): 167-178.

Nouroz F, Noreen S, Heslop-Harrison JS (2017). Identification and evolutionary dynamics of CACTA DNA transposons in *Brassica*. Pak J Bot 49(2): 789-798.

Perrier X, De Langhe E, Donohue M, Lentfer C, Vrydaghs L, Bakry F, Carreel F, Hippolyte I, Horry J-P, Jenny C et al (2011). Multidisciplinary perspectives on banana (*Musa* spp.) domestication. Proc Natl Acad Sci USA 108: 11311-11318. doi: 10.1073/pnas.1102001108

Pollefeys P, Sharrock S, Arnaud E (2004). Preliminary analysis of the literature on the distribution of wild *Musa* species using MGIS and DIVA-GIS. Montpellier, France: INIBAP.

Queen RA, Gribbon BM, James C, Jack P, Flavell AJ (2004). Retrotransposon-based molecular markers for linkage and genetic diversity analysis in wheat. Mol Genet Genomics 271: 91-97. Doi: 10.1007/s00438-003-0960-x

Schulman AH, Flavell AJ, Paux E, Ellis TH (2012). The application of LTR retrotransposons as molecular markers in plants. Methods Mol Biol 859: 115-153. doi: 10.1007/978-1-61779-603-6_7

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

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 28(10): 2731-2739. doi: 10.1093/molbev/msr121

Tomlinson P (1969). Anatomy of the monocotyledons. III. Commelinales–Zingiberales Oxford: Clarendon Press, 1969

Wang H, Liu JS (2008). LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. BMC Genomics 9: 382. doi: 10.1186/1471-2164-9-382

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavel A, Leroy P, Panaud O, Paux E, SanMiguel P, Schulman AH (2007). A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8: 973-982. doi: 10.1038/nrg2165

Zhang L, Yan L, Jiang J, Wang Y, Jiang Y, Yan T, Cao Y (2014). The structure and retrotransposition mechanism of LTR retrotransposons in the asexual yeast *Candida albicans*. Virulence 5: 655-664. doi: 10.4161/viru.32180

LEGENDS TO FIGURES

FIG. 1. Flow chart representing the methodology used in present work from identification to phylogenetic analysis of LTR REs in *Musa*. BACs: Bacterial artificial chromosomes. LTR REs: LTR retrotransposons. CDD: Conserved domain database. LARDs: Large retrotransposons derivatives. PBS: Primer binding site. PPT: Polypurine tract. RT: Reverse transcriptase.

FIG. 2. Dot plot of *Musa acuminata* (AC226035.1) against itself to identify LTR retrotransposons. The central diagonal line running from one corner to other showed the homology of the sequence against itself. The coloured boxes on the diagonal line showed the positions of LTR retrotransposons insertions with LTRs. Four Copia, three Gypsy and two LARD-like elements are inserted with a total size of ~60 kb of the 105 kb BAC size covering 58.5% of total BAC sequence. The nested structure of LTR retrotransposon is also shown in purplesquare.

FIG. 3. a) Schematic representation of few retrotransposons in *Musa*. The red arrowheads at the corners represent the TSDs, while blue arrows indicate TIRs. The *gag-pol* regions are drawn with their protein domains. Scale is measuring the lengths of the elements (bp). Additional insertions or unknown sequences are represented by different colours. b) A 17.8 kb large *MaGYP20* is drawn with other Gypsy and a DNA transposon inserted in it. A 16.2 kb *MaCOP3* is shown with 5.2 kb inserted Copia element in it. AP: Aspartic protease. RT: Reverse transcriptase. INT: Integrase. GAG: gag-nucleocapsid. ZK: Zinc knuckle. DUF: Domain of unknown function. CHR: Chromatin organization modifier. CMV. Cauliflower mosaic virus. PR: Hypothetical protein. UN: Unknown.

FIG. 4. PCR analysis for the detection of retrotransposons RT polymorphisms across 48 cultivars in *Musa*. Dark bands are indicating the expected products. The amplification of a) *MaGYP8* b) *MaGYP12* c) *MaGYP17* d) *MaCOP5* e) *MaCOP8* f) *MaCV1*. PCR figures show reversed images of size-separated ethidium bromide-stained DNA on agarose gels after electrophoresis. Ladders (HP-I) show fragment sizes in base pairs; the diploid and triploid *Musa* genomes represented on the top of Lanes (AA, BB, AB, AAA, AAB, ABB) and the numbers at the base are given in Table 1.

FIG. 5. Phylogenetic analysis of *Musa* LTR retrotransposon. The 33 RT sequences from intact elements from *Musa* and 2 known sequences (*Ty1-Copia, Ty3-Gypsy*) from *Saccharomyces cerevisiae* were used to construct the phylogenetic tree. Neighbor-joining tree was constructed with 1000 bootstrap replicates in MEGA5 program. The p distance model was used to calculate the genetic distance. The two major lineages separate the Gypsy (represented by black rhombus)/ Pararetrovirus elements (green square) and Copia (blue circles). The detailed

descriptions of the elements are given in the Table 1. Ma: *Musa acuminata*. Mb: *Musa blabisiana*. COP: Copia. GYP: Gypsy. CV: Chromoviridae *MaCVI*: *Musa acuminata* chromovirideae.

FIG. 6. Phylogenetic tree showing relationship between RT nucleotide sequences of *Musa* and other plants. Of the 110 RT sequences, 33 sequences are from *Musa* and remaining 77 (Supplementary Table) are from known plant retrotransposons collected from Gypsy database. The tree was inferred by using Neighbor-joining method in MEGA5, where p distance model was used to calculate the genetic distance. The 1000 bootstrap replicates were used and the values <50% are not shown. The three main lineages separate the Gypsy (represented by black filled and open rhombus), Copia (blue filled and open circles) and Caulimoviruses (blue filled and open squares). The *Musa* specific elements from Gypsy, Copia and Caulimoviruses are represented by filled shapes, while open shapes are representing elements from other organisms. The details of the *Musa* elements are given in the Table 1 and other plant elements in supplementary Table. Ma: *Musa acuminata*. Mb: *Musa balbisiana*. COP: Copia. GYP: Gypsy. *MaCVI*: *Musa acuminata* chromovirideae.

TABLE 1. *Musa* accessions used with their names and genomic compositions. Sr. No. Serial Number

Sr. No.	Reference	Genome	Accession Name	Sr. No.	Reference	Genome	Accession Name
1	Eumusa	AA	Calcutta 4	25	Eumusa	AAB	Karimkadali
2	Eumusa	AA	Sannachenkadali	26	Eumusa	AAB	Perumadali
3	Eumusa	AA	Pisanglilin	27	Eumusa	AAB	Kunoor ettan
4	Eumusa	AA	Kadali	28	Eumusa	AAB	Palyamcodan
5	Eumusa	AA	Matti	29	Eumusa	AAB	Mysoreettan
6	Eumusa	AA	Cherukadali	30	Eumusa	AAB	Krisnavazhai
7	Eumusa	BB	Pisang Klutuk Wulung 1	31	Eumusa	AAB	Poovan
8	Eumusa	BB	Pisang Klutuk Wulung 2	32	Eumusa	AAB	Doothsagar
9	Eumusa	BB	Javan	33	Eumusa	AAB	Charapadati
10	Eumusa	BB	Klutuk	34	Eumusa	AAB	Kumbillakannan
11	Eumusa	BB	Tani	35	Eumusa	AAB	Velipadati
12	Eumusa	BB	Batu	36	Eumusa	AAB	Vellapalayamcodan
13	Eumusa	AB	Njalipovan	37	Eumusa	AAB	Ettapadati
14	Eumusa	AB	Adukkan	38	Eumusa	AAB	Padati
15	Eumusa	AB	Padalamukili	39	Eumusa	AAB	Chinali
16	Eumusa	AAA	Manoranjitham	40	Eumusa	AAB	Nendran
17	Eumusa	AAA	Grandnain	41	Eumusa	AAB	Poomkalli
18	Eumusa	AAA	Gross-michel	42	Eumusa	AAB	Kamaramasengi
19	Eumusa	AAA	Greenred	43	Eumusa	ABB	Kosta bontha
20	Eumusa	AAA	Red	44	Eumusa	ABB	Peyan
21	Eumusa	AAA	Monsmari	45	Eumusa	ABB	Kanchikela
22	Eumusa	AAA	Robusta	46	Eumusa	ABB	Boothibale
23	Eumusa	AAA	Dwarf Cavendish	47	Eumusa	ABB	Monthan
24	Eumusa	AAB	Motta povan	48	Eumusa	ABB	Karpooravali

TABLE 2. PCR primer pairs to amplify the RT region of Gypsy, Copia and Caulimovirus (CV) elements.

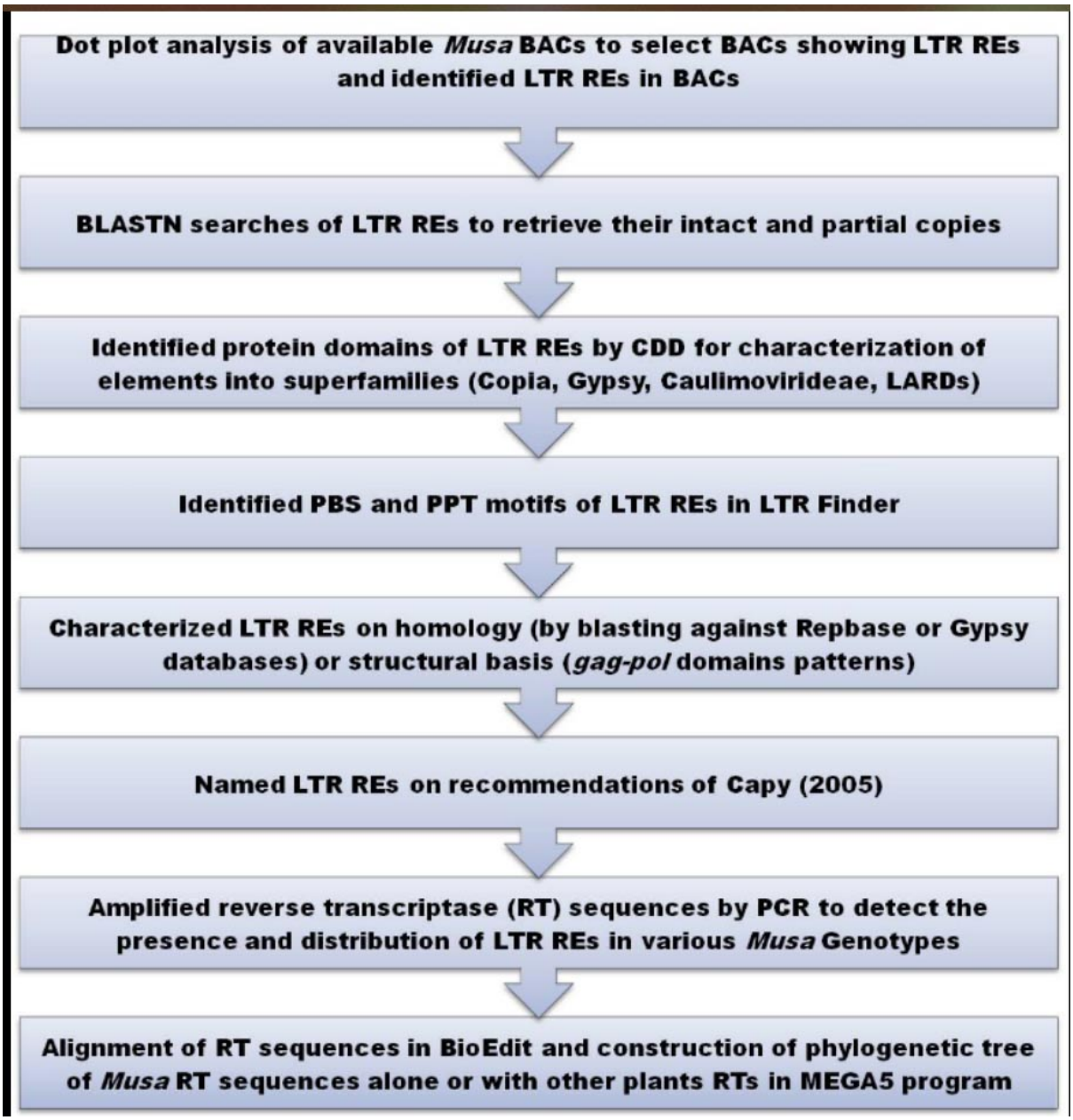
Sr. No.	Super-Family	TE Family	Product size (bp)	Primer name	Primer Sequence
1	Gypsy	*MaGYP8*	684	MaGYP8F MaGYP8R	CTTCTCGGCAACATGACCA GGTCTACCGCCACTCCTTC
2	Gypsy	*MaGYP12*	830	MaGYP12F MaGYP12R	CCAATTCCCACATTAGATGC GAGAGCATGAGTCATTGTGC
3	Gypsy	*MaGYP17*	835	MaGYP17F MaGYP17R	GCAGCTCAAAAGCACCTTTC CCAATAGCAAAGTCCGAAGC
4	Copia	*MaCOP5*	744	MaCOP5F MaCOP5R	CTTAGTCGCAGTACTCATAG TGGAAGCTTGTTCCTAGACC
5	Copia	*MaCOP8*	964	MaCOP8F MaCOP8R	CTTTCACAATGGGAGCAACA GTTGAACCACAAGTTCCTCA
6	CV	*MACV1*	425	MACV1F MACV1R	CAACTACAAGAGGCTGAACG CTATTTCCTTGACTGCTATC

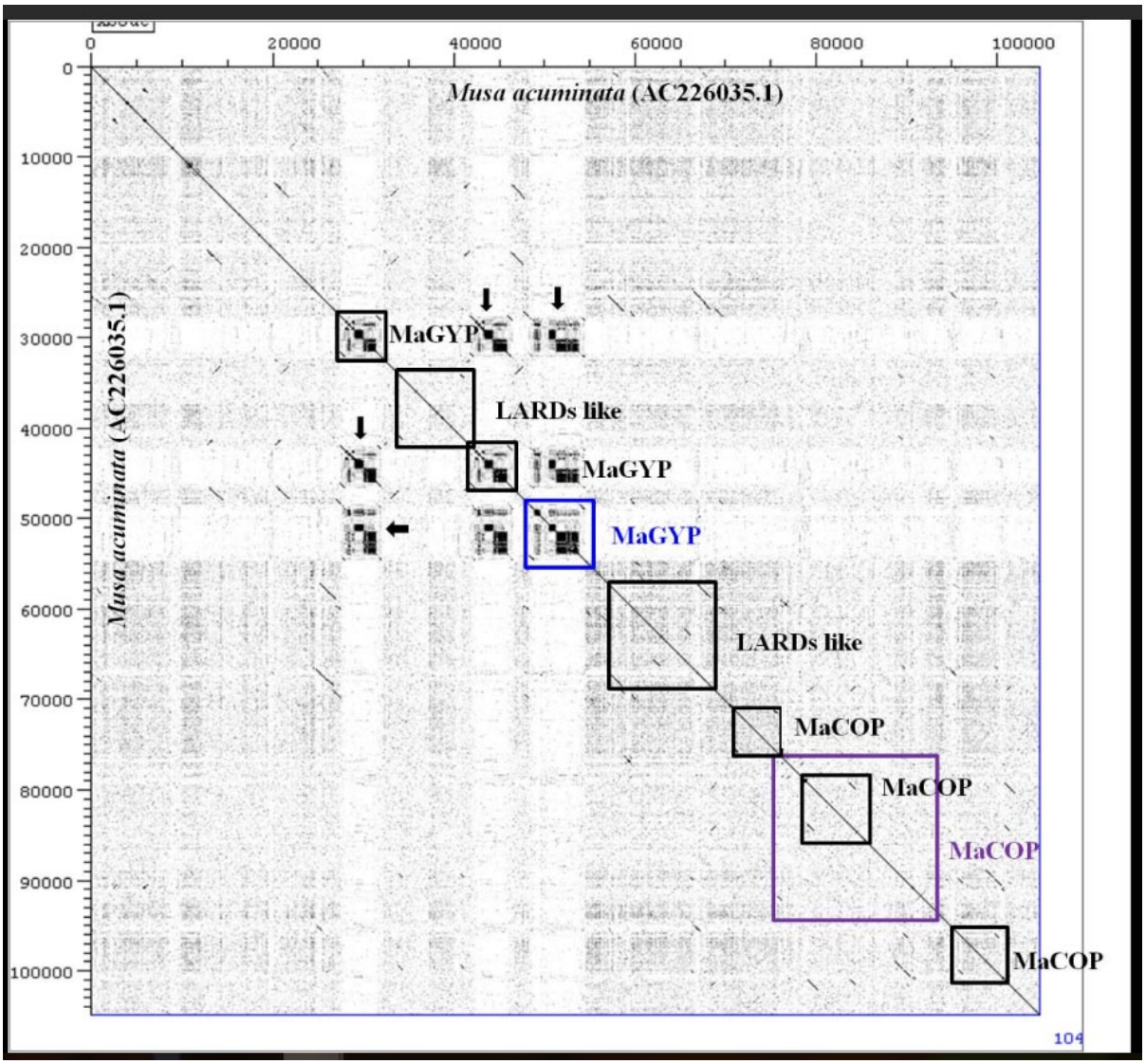
TABLE 3. Superfamilies of LTR retrotransposons identified from *Musa* with their sizes, TSDs, LTRs, positions and orientations in BAC clone sequences. Asterisks after TSD show variable TSDs at 5′-3′. ND: Not determined.

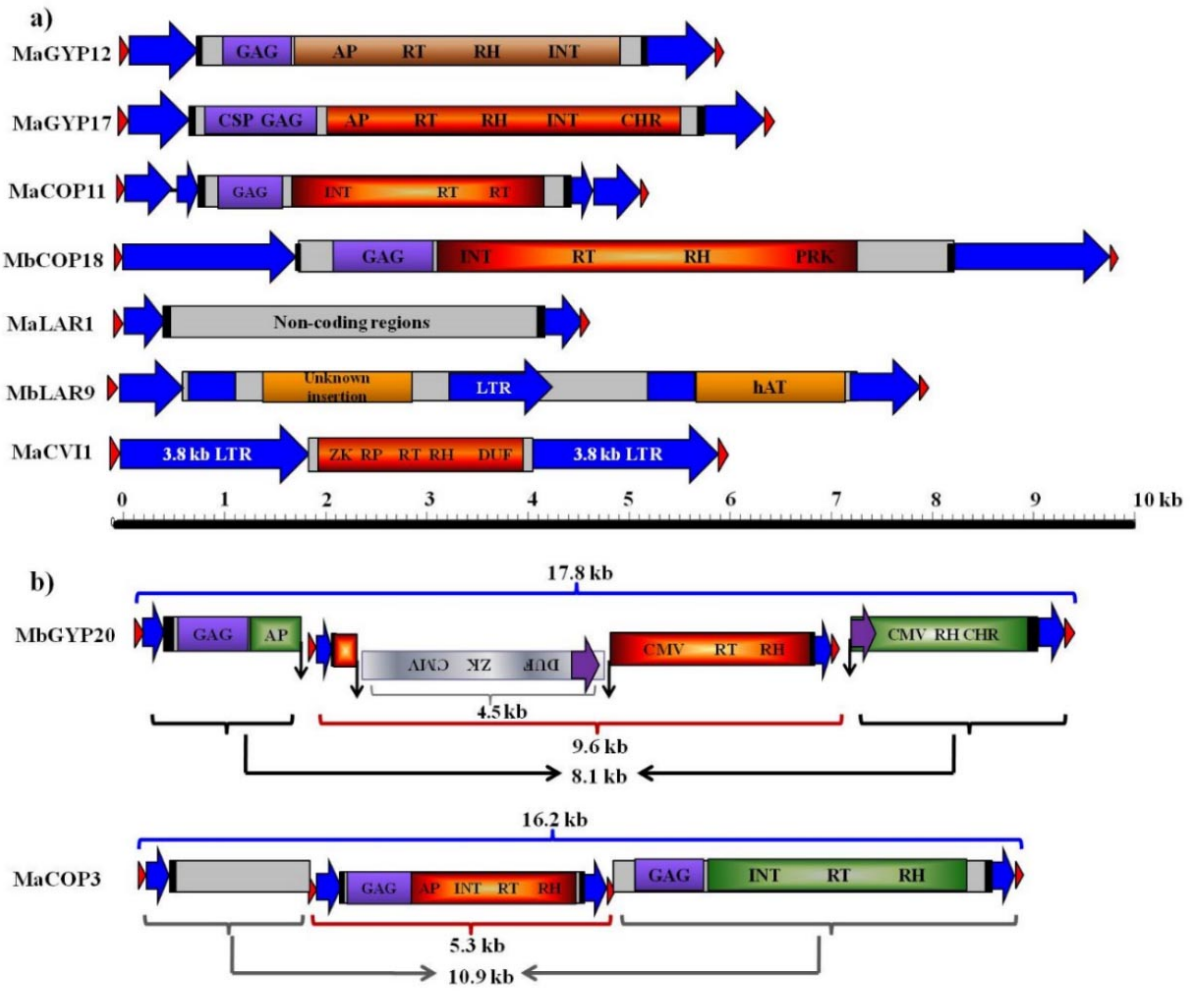
Element name	Super-family	BAC Accession	Species	Size	TSD	LTR	Position in BAC	Orient-ation
MaGYP1	Gypsy	AC226032.1	*M. acuminate*	4982	CCCGG	505/505	65772-70752	5′-3′
MaGYP2	Gypsy	AC226033.1	*M. acuminate*	3802	AGATG	543/527	28867-32673	5′-3′
MaGYP3	Gypsy	AC226035.1	*M. acuminate*	4567	ATGAG	458/458	27408-31974	5′-3′
MaGYP4	Gypsy	AC226035.1	*M. acuminate*	4627	TAGGA	458/458	41793-46419	5′-3′
MaGYP5	Gypsy	AC226035.1	*M. acuminate*	6245	ACTTC	586/586	48237-54481	5′-3′
MaGYP6	Gypsy	AC186752.1	*M. acuminate*	3015	TATGT*	655/655	62134-65148	5′-3′
MaGYP7	Gypsy	AC226046.1	*M. acuminate*	5326	AATAT	462/411	116297-121622	3′-5′
MaGYP8	Gypsy	AC226048.1	*M. acuminate*	5907	TGTTT	473/473	1861-7767	5′-3′
MaGYP9	Gypsy	AC226048.1	*M. acuminate*	5318	AGACG	481/506	12597-17915	5′-3′
MaGYP10	Gypsy	AC226048.1	*M. acuminate*	5435	GAGAT	438/438	24477-29911	3′-5′
MaGYP11	Gypsy	AC226048.1	*M. acuminate*	5319	AGACG	481/519	12597-17915	5′-3′
MaGYP12	Gypsy	AC226048.1	*M. acuminate*	5760	CTGAC	671/671	36420-42179	3′-5′
MaGYP13	Gypsy	AC226048.1	*M. acuminate*	5418	AAACT*	1062/1063	123405-128822	3′-5′
MaGYP14	Gypsy	AC186950.2	*M. acuminate*	11605	CCAGT	624/624	9314-20918	3′-5′
MbGYP15	Gypsy	AC226053.1	*M. balbisiana*	4940	GTTAA*	883/884	121237-126176	5′-3′
MbGYP16	Gypsy	AC226051.1	*M. balbisiana*	4014	TAAA	265/264	125881-129840	3′-5′
MaGYP17	Gypsy	AC226196.1	*M. acuminate*	6436	TCCT	792/792	10198-16633	5′-3′
MbGYP18	Gypsy	AP009325.2	*M. balbisiana*	7108	GCACC*	374/383	45693-52799	5′-3′
MbGYP19	Gypsy	AP009334.1	*M. balbisiana*	7368	GGTAT	1105/883	44828-52195	5′-3′
MbGYP20	Gypsy	AP009325.2	*M. balbisiana*	17804	GCCAC	393/351	79527-97303	3′-5′
MaCOP1	Copia	AC226035.1	*M. acuminate*	5290	CTGCA	605/605	79036-84325	5′-3′
MaCOP2	Copia	AC226035.1	*M. acuminate*	4808	TCTCT	358/360	70977-75784	5′-3′
MaCOP3	Copia	AC226035.1	*M. acuminate*	16242	GAGG	338/299	75834-92033	5′-3′
MaCOP4	Copia	AC226038.1	*M. acuminate*	4022	CCATA	261/263	35057-39078	3′-5′
MaCOP5	Copia	AC226038.1	*M. acuminate*	8158	CATAA	1201/1285	63867-72094	3′-5′
MaCOP6	Copia	AC226038.1	*M. acuminate*	7036	GAATC	452/472	100169-107204	3′-5′
MaCOP7	Copia	AC226041.1	*M. acuminate*	5012	ACTAA	149/144	2059-7070	3′-5′
MaCOP8	Copia	AC226044.1	*M. acuminate*	6019	GGATT	499/500	40359-46377	3′-5′
MaCOP9	Copia	AC226047.1	*M. acuminate*	6959	GGTTT	526/530	68302-75260	5′-3′
MaCOP10	Copia	AC226047.1	*M. acuminate*	8767	TGTAT	1597/1597	15958-24725	3′-5′
MaCOP11	Copia	AC226051.1	*M. acuminate*	8478	AAAG	1494/1388	34505-42982	5′-3′
MaCOP12	Copia	AC226051.1	*M. acuminate*	7176	AGCGA*	1132/1238	118482-125657	3′-5′
MaCOP13	Copia	AC226051.1	*M. acuminate*	5938	ND	548/573	119071-125008	3′-5′
MaCOP14	Copia	AC186753.1	*M. acuminate*	6054	GAAAT*	492/548	28872-34925	3′-5′
MbCOP15	Copia	AC226053.1	*M. balbisiana*	4980	ACCTT	449/449	100799-105778	3′-5′
MaCOP16	Copia	AC226040.1	*M. acuminate*	5424	GCAAC	438/406	40145-45568	3′-5′
MaCOP17	Copia	AC226196.1	*M. acuminate*	8084	GATAT*	1024/1000	50839-58921	3′-5′
MbCOP18	Copia	AC226052.1	*M. balbisiana*	9878	TGTC*	1396/1415	184519-194396	3′-5′
MbCOP19	Copia	AC226055.1	*M. balbisiana*	5203	TTCA	592/590	26728-31930	5′-3′
MaCVI	(CV)	AC226046.1	*M. acuminate*	11077	CTCT	3866/3813	160034-1711104	5′-3′
MaLAR1	LARDs	AY484588.1	*M. acuminate*	4564	GGTT	447/447	48330-52793	5′-3′
MbLAR2	LARDs	AC226055.1	*M. balbisiana*	4428	ATAT	445/445	9329–13756	5′-3′
MbLAR3	LARDs	AP009334.1	*M. balbisiana*	4452	ATGC	383/383	20981-25432	3′-5′
MaLAR4	LARDs	AC186955.1	*M. acuminate*	4318	GTATT*	607/611	47077-51394	3′-5′
MbLAR5	LARDs	FN396604.1	*M. balbisiana*	4449	ATAC	382/382	28462-32910	5′-3′
MbLAR6	LARDs	FN396605.1	*M. balbisiana*	4449	GGAG	382/382	36620-41068	5′-3′
MaLAR7	LARDs	AC186951.1	*M. acuminate*	4571	ATAT	446/446	92933-97503	3′-5′
MaLAR8	LARDs	AC186753.1	*M. acuminate*	4550	GTAG	434/437	15832-20381	5′-3′
MbLAR9	LARDs	AC186754.1	*M. balbisiana*	7712	ATTGT*	626/635	72565-80276	3′-5′
MaLAR10	LARDs	AC226051.1	*M. acuminate*	4005	TTTC*	974/984	129126-133076	5′-3′

TABLE 4. *Musa* LTR retrotransposons with PBS/PPT motifs and *gag-pol* gene protein domains. UD: undetermined. AP: aspartic protease. RT: reverse transcriptase. INT: integrase. ZK: zinc knuckle. ZF: zinc finger. CHR: chromodomain. CHR: chromatin organization modifier. DUF: protein domain of unknown function.

Element name	tRNA type	PBS sequence (5´-3´)	Position	PPT sequence (5´-3´)	Position	Domain organization (5´-3´)
MaGYP1	Met*	TATCAGAGCAGCGATCTT	516-533	ATGAGGAGCTGAAGA	4394-4408	GAG, AP, INT,CHR
MaGYP2	Asn	CGCTAGAAGGAGGGC	560-574	UD	---	GAG,TR
MaGYP3	Asn	CGCTAGAAGGAGGGC	470-484	ACGGACCAGGGAGAA	4012-4026	HP,TR,TVP,HVP
MaGYP4	Asn	CACTAGAAGGAGGGC	472-486	ACGGACCAGGGAGAA	4072-4086	DUF,HP,APC,HVP
MaGYP5	Ala	GGAGCTATGCGTCGGTTC	612-619	AGGAGAAAGCTAACG	5605-5619	MFS,TR,TVP,
MaGYP6	UD	---	---	GGGGGGGGGGGGGGG	2331-2345	GAG
MaGYP7	Pro	TCGAGGCTGACGATTC	497-512	GGAAGGGCAGCGAGA	4869-4883	GAG,AP,RT,RH
MaGYP8	Met	TATCAGAGCAGCGTT	484-499	ATGAGGAGCTGAAGA	5351-5365	GAG,AP,RT,RH,INT,CHR
MaGYP9	Met	TATCAGAGCAGCGTTCTTG	492-511	TGAAGAGGGCGGGTT	4794-4808	GAG,AP,RT,RH
MaGYP10	Met	TATCAGAGCAGCGTT	468-483	TGAAGAGGGCGGGTC	4977-4991	GAG,TIM,AP,RT,RH,INT,CHR
MaGYP11	Leu	TCATGAATTTTGGGAATTTG	555-574	GGAAGGGCAGCGAGA	4792-4806	GAG,AP,RT,RH
MaGYP12	Ala	TGGAGATGACGCTGAGTCG	754-772	AGACTTGAGGACAAG	5049-5063	GAG,ZK,AP,RT,RH,INT
MaGYP13	Leu	AACATACCACTCTGCAGC	1076-1093	TCATTCTTCTATGTT	4334-4348	RT
MaGYP14	Asn	CGCTAGAAGGAGGGCCT	636-652	TTCAGGGGGGGAATA	10962-10976	GAG,AP,RT,RH,INT
MbGYP15	Thr*	CCAACTAAGTTAGGAATTG	893-911	GCATGAAGAAGGAGA	3968-3982	GAG,AP,RT,RH
MbGYP16	Lys	TTCACCATGGCAAAGCATTG	349-368	TGAGTAATTGTTTAT	3729-3744	GAG,AP,RT,RH
MaGYP17	Met	TATCAGAGCCAGGTT	803-817	GACATGAAGAAGAAG	5568-5582	CSP,GAG,AP,RT,RH,INT,CHR
MbGYP18	Lys	TCTCACCATGCGAAGCACCT	431-452	AAGTTGGGGAGAATA	6673-6687	AP,RT
MbGYP19	UD	---	---	CGAGGAAAGAGGGAA	6516-6530	GAG,AP,RT,RH,INT
MbGYP20	Met	TATCAGAGCAGCGTT	362-376	TGAAGAGGACGGGTC	17392-17406	GAG,AP,(CMV,RH,DUF)*DUF, CMV, RH, CHR
MaCOP1	Met	TATCAGAGCGGGGTTTTG	616-633	AAGAAAGACAGGAGA	4589-4603	GAG,AP,INT,RT,RH
MaCOP2	Met	TATCCAGCATGTCAAGTTTC	388-407	AGGAAGAGGCCATAG	4407-4421	GAG,INT,RT,RH
MaCOP3	Arg*	CGACCTTGCATATGATCG	311-328	AAGAGAAAGGAAGAA	15883-15897	(GAG,AP,INT,RT,RH)*, GAG,INT,RT, RH
MaCOP4	Met*	ATCTGATCTAAGAGTTTTG	262-280	GGAAGAACAAGAAAA	3706-3720	GAG,ZK,INT
MaCOP5	Met	TATCAGAGCAAGGTTATC	1296-1313	CAAAAAGGGGGAGAT	6938-6952	GAG,INT,RT,RH
MaCOP6	Met	TATCAGAGCCAAGTTATT	486-503	UD	---	RT
MaCOP7	Thr*	AGGCTTCGTGAGTGAGTCG	229-247	GGGGTTGGAGAGGGA	4779-4793	GAG,INT,RT,RH
MaCOP8	Cys	TGCCATGAAAATGATTTG	561-579	GACCAAGTGGGAGAA	5501-5515	GAG,INT,RT,RH
MaCOP9	Ser	GATGCCTGAATGATTCG	585-601	GGCCAAGTGGGAGAA	6410-6424	RT
MaCOP10	Met	TATCAAAGCCAAGTTGTTCG	1609-1628	AGGTCAAGTGGGAGA	7150-7164	GAG,INT,RT,RH
MaCOP11	Met	TATCAGAGCCAGGTT	1504-1518	UD	---	GAG,INT,RT,RH
MaCOP12	Val*	TATTAAATATGACATACAAA	1207-1226	AGAAAAAAGCTTAAA	5903-5917	GAG,INT,RT,RH
MaCOP13	Val*	TATTAAATATGACATACAAA	618-637	AGAAAAAAGCTTAAA	5314-5328	GAG,INT,RT,RH
MaCOP14	UD	---	---	AAGAAGAAACCAAAA	5703-5417	GAG,INT,RT,RH
MbCOP15	Met	TATCAGAGCCTAGTTTCG	461-478	AGAAGGTGGAGCAAG	4483-4497	GAG,INT,RT,RH
MaCOP16	Val*	ATTCACCATAGAGGCCACAA	443-463	GAACAAGTGGGGGAT	4967-4981	GAG,RT,RH,MT*
MaCOP17	Val*	TATTGAGATAAAGCAAA	1398-1414	AAATCAAATTGAGAG	7043-7057	GAG,INT,RT,RH
MbCOP18	Sup	GTATCAGAGTGAGGCTC	1424-1440	CAAAAAGGAGAAGAT	8464-8478	GAG,INT,RT,RH,PRK
MbCOP19	Lys	GCCCACAAGGGAGGCT	625-640	AAATACAAAATTAAA	4571-4585	GAG,INT,RT,RH
MaCVI	Gly	TGCAAAAGGCCAAGGAATT	3918-3937	GAGCTGGGTAGCGGA	7172-7186	ZK,AP,RT,RH,DUF
MaLAR1	UD	---	---	ATAAGTGGGGGAGAA	4561-4564	UD
MbLAR2	UD	---	---	ATAAGTGGGGGAGAA	3965-3979	UD
MbLAR3	UD	---	---	ATAAGTGGGGGAGAA	4051-4065	UD
MaLAR4	UD	---	---	UD	---	UD
MbLAR5	UD	---	---	ATAAGTGGGGGAGAA	4049-4063	UD
MbLAR6	UD	---	---	ATAAGTGGGGGAGAA	4049-4063	UD
MaLAR7	Asp	GGGACCTAACGGGGCTGCG	505-523	ATAAGTGGGGGAGAA	4107-4121	UD
MaLAR8	Leu*	TGGTATCAGAGTGGGAT	442-458	AATAAGTGAGGGAGA	4088-4102	UD
MbLAR9	UD	---	---	UD	---	UD
MaLAR10	UD	---	---	UD	---	ADM


Dot plot analysis of available *Musa* BACs to select BACs showing LTR REs and identified LTR REs in BACs
BLASTN searches of LTR REs to retrieve their intact and partial copies
Identified protein domains of LTR REs by CDD for characterization of elements into superfamilies (Copia, Gypsy, Caulimovirideae, LARDs)
Identified PBS and PPT motifs of LTR REs in LTR Finder
Characterized LTR REs on homology (by blasting against Repbase or Gypsy databases) or structural basis (*gag-pol* domains patterns)
Named LTR REs on recommendations of Capy (2005)
Amplified reverse transcriptase (RT) sequences by PCR to detect the presence and distribution of LTR REs in various *Musa* Genotypes
Alignment of RT sequences in BioEdit and construction of phylogenetic tree of *Musa* RT sequences alone or with other plants RTs in MEGA5 program


Musa acuminata (AC226035.1)
Musa acuminata (AC226035.1)
0
20000
40000
60000
80000
100000
10000
30000
50000
70000
90000
MaGYP
LARDs like
MaGYP
MaGYP
LARDs like
MaCOP
MaCOP
MaCOP
MaCOP



Nouroz F, Noreen S, Ahmad H, Heslop-Harrison JS. 2017.
The landscape and structural diversity of LTR retrotransposons in Musa genome.
Molecular Genetics and Genomics First Online: 10 June 2017 DOI: 10.1007/s00438-017-1333-1.

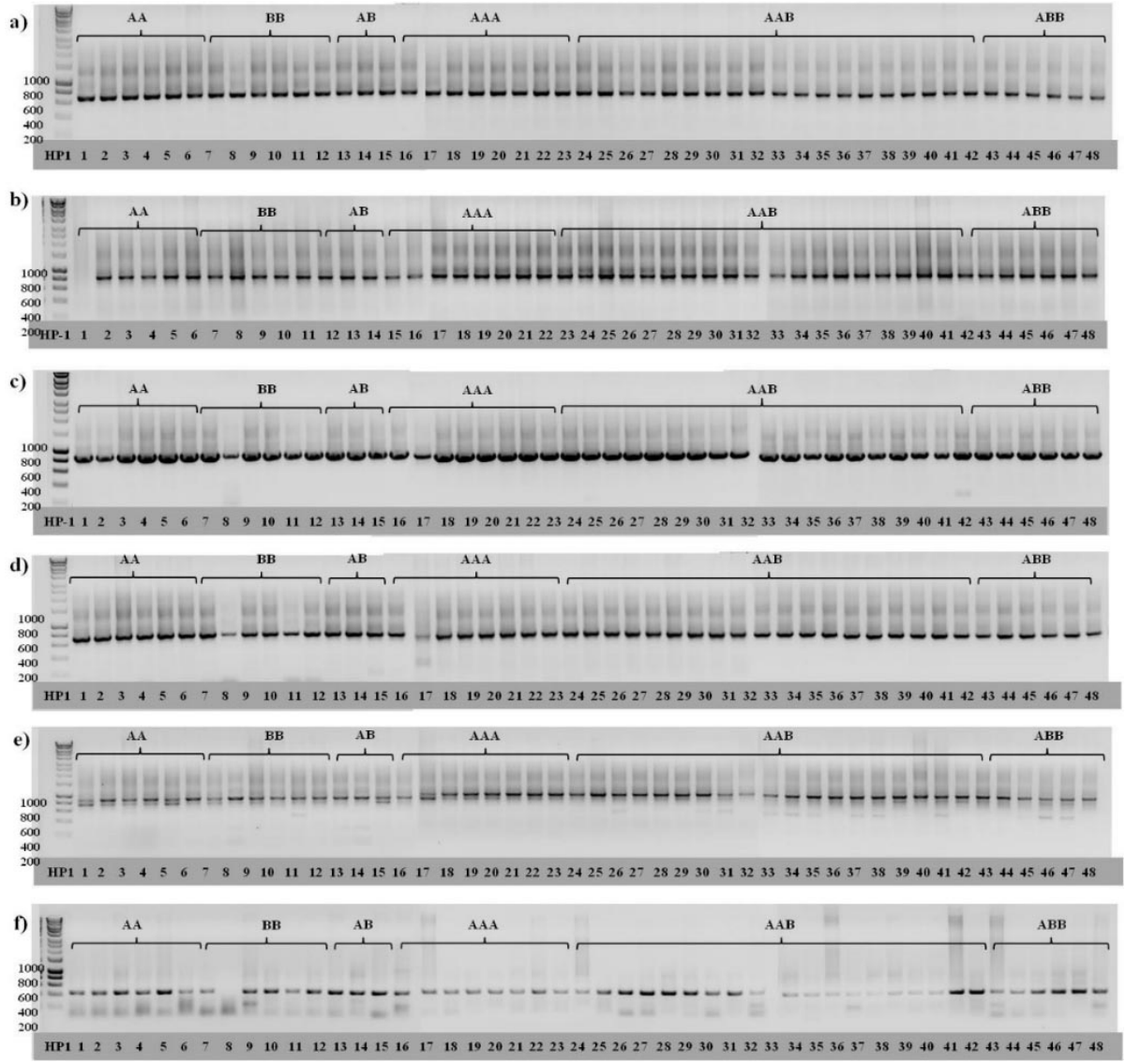

a)
AA
BB
AB
AAA
AAB
ABB
1000
800
600
400
200
HP1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
b)
AA
BB
AB
AAA
AAB
ABB
1000
800
600
400
200
HP-1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
c)
AA
BB
AB
AAA
AAB
ABB
1000
800
600
400
200
HP-1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
d)
AA
BB
AB
AAA
AAB
ABB
1000
800
600
400
200
HP1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
e)
AA
BB
AB
AAA
AAB
ABB
1000
800
600
400
200
HP1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
f)
AA
BB
AB
AAA
AAB
ABB
1000
800
600
400
200
HP1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48

Figure 5

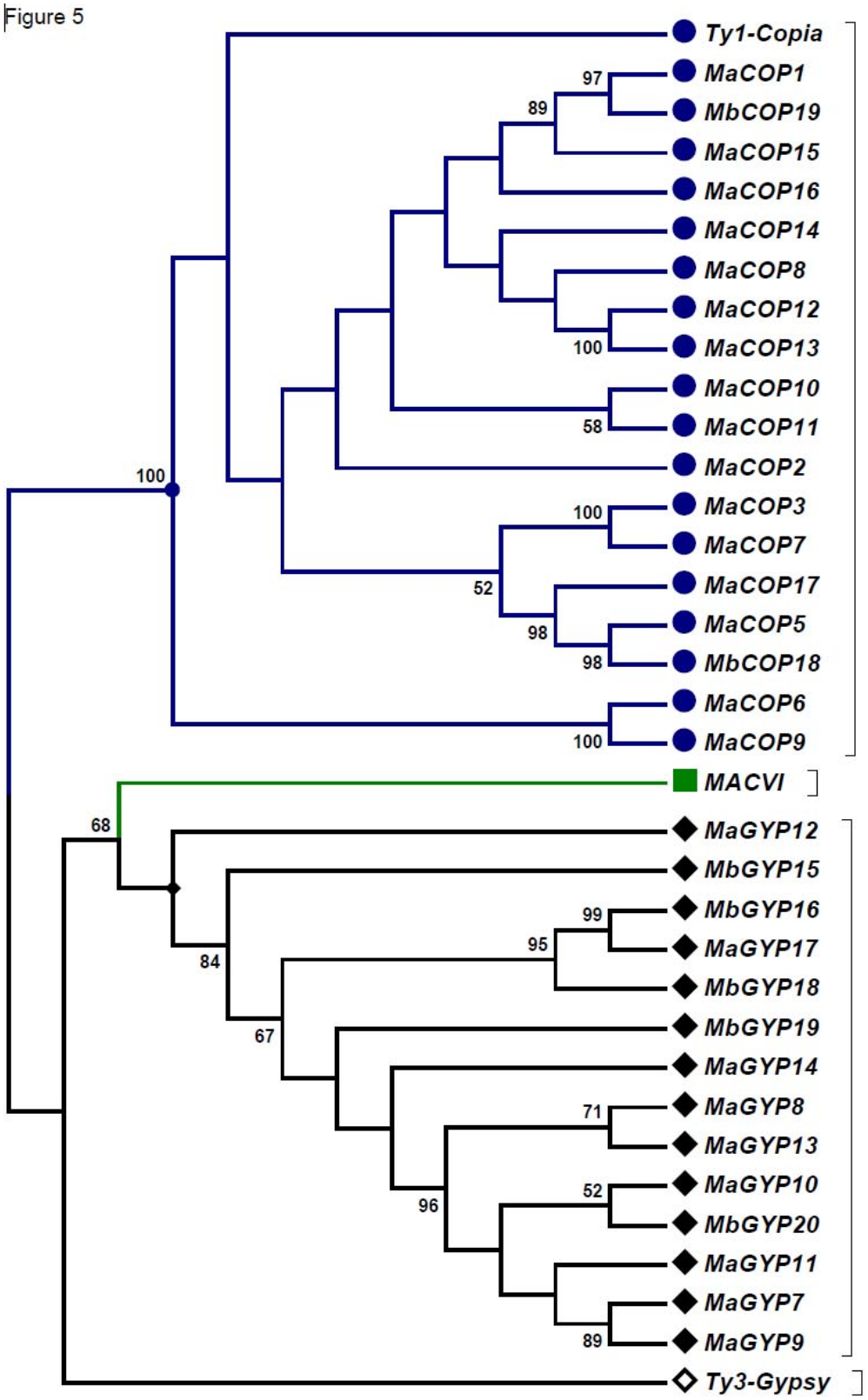


Figure 5
Nouroz F, Noreen S, Ahmad H, Heslop-Harrison JS. 2017.
The landscape and structural diversity of LTR retrotransposons in Musa genomes.
Molecular Genetics and Genomics Online: 10 June 2017 DOI: 10.1007/s00438-017-1333-1.

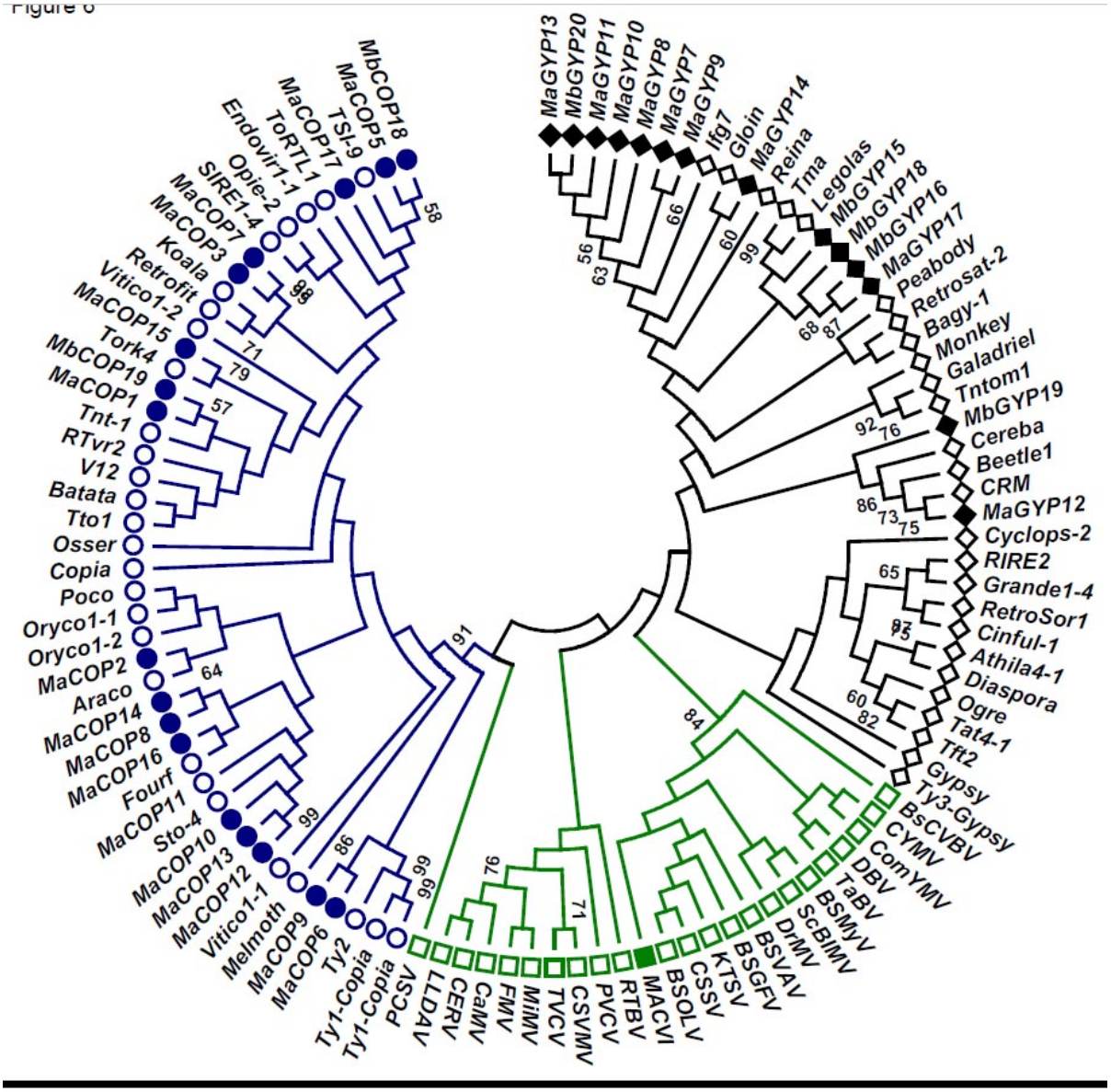


Figure 6
Nouroz F, Noreen S, Ahmad H, Heslop-Harrison JS. 2017.
The landscape and structural diversity of LTR retrotransposons in Musa genomes.
Molecular Genetics and Genomics Online: 10 June 2017 DOI: 10.1007/s00438-017-1333-1.