Pak. J. Bot., 51(5): 1637-1644, 2019. DOI: http://dx.doi.org/10.30848/PJB2019-5(31)

GENOMIC AND EVOLUTIONARY DIVERSITY OF LTR RETROTRANSPOSONS IN DATE PALM (*PHOENIX DACTYLIFERA*)

FAISAL NOUROZ[1,2*] AND MUKARAMIN[1]

[1]Department of Botany, Hazara University Mansehra, Pakistan
[2]Department of Bioinformatics, Hazara University Mansehra, Pakistan
**Corresponding author's e-mail: faisalnouroz@gmail.com*

Abstract

Of the transposable elements (TEs), the retrotransposons are the most copious elements identified from many sequenced genomes. They have played a major role in genome evolution, rearrangement and expansions based on their copy and paste mode of proliferation. They are further divided into LTR and Non-LTR retrotransposons. The purpose of the current study was to identify the LTR REs in sequenced *Phoenix dactylifera* genome and to study their structural diversity. A total of 150 *P. dactylifera* BAC sequences with >60kb sizes were randomly retrieved from NCBI database and screened for the presence of LTR retrotransposons. Seven BAC sequences showed full length LTR Retrotransposons with 4 Copia and 3 Gypsy families having variable copy numbers in respective families. Reverse transcriptase (RT) domain was found as the most conserved domain among Copia and Gypsy superfamilies and was used to deduce evolutionary analysis. The amino acid residues among various RT sequences showed variability in their percentages indicating post divergence evolution. Amino acid Leucine was found in highest proportions followed by Lysine, while Methionine and Tryptophan were in lowest percentages. The phylogenetic analysis based on RT domains confirmed that although having most conserved RT regions, several evolutionary events occurred causing nucleotide polymorphisms and hence clustering of Gypsy and Copia superfamilies into their respective lineages. The study will be helpful in identification and annotation of these elements in other species and genera and their distribution patterns on chromosomes by florescent in situ hybridization techniques.

Key words: Transposable elements, *Phoenix dactylifera*, Retrotransposons, Gypsy, Copia, Phylogenetic analysis.

Introduction

Transposable elements (TEs) or the mobile genetic elements are the segments of DNA, who gradually change their positions on chromosomes due to their mobile nature. They have adopted two different mechanisms for retrotransposition and are divided into two classes as Class I (Retrotransposons) and Class II (DNA transposons) TEs. Retrotransposons (REs) after making their copies are inserted to new sites on chromosomes. Reverse transcriptase (RT) domain acts as enzymatic machinery required for making new copies by adopting the copy and paste mechanism of retrotransposition. In contrast, DNA transposons require transposase domain, which helps in direct mobilization of DNA segment by catalyzing the essential DNA cutting and joining reactions through cut and paste mechanism of proliferation (Feschotte *et al*., 2002; Wicker *et al*., 2007; Kapitonov & Jurka, 2008).

Based on presence and absence of *gag-pol* gene coding polyproteins, the LTR REs are classified as autonomous and non-autonomous elements. The autonomous LTR REs are characterized by the presence of long terminal repeats (LTRs) on terminal ends, primer binding sites (PBS) towards downstream of 5´ LTR, a polypurine tract (PPT) towards upstream of 3´ LTR, internal *gag-pol* genes encoding the proteins as 5´-GAG-INT-RT-RH-3´. Few elements also harbor some additional protein domains of known or unknown functions and nature. In contrast, the non-autonomous elements lack one or more important protein domains necessary for their mobilization and hence are mostly in active or non functional in genomes or utilize the enzymatic machinery of their autonomous partners residing nearby for their retrotransposition.

The major superfamilies of LTR REs are Ty1/Copia, Ty3/Gypsy, Bel/Pao and Retroviridae. Copia, Gypsy and Retroviridae superfamilies are common in all genomes but are most frequently detected in plants, while Bel/Pao are mostly proliferating in animal genomes (Wicker *et al*., 2007). Copia and Gypsy superfamilies can be differentiated based on their degree of sequence similarity and the order of *gag-pol* encoded gene products. In Copia superfamily, the gene order is [5´-LTR-Capsid protein (GAG) -aspartic protease (AP) -integrase (INT) -reverse transcriptase (RT) -RNaseH (RH) -3´LTR], while in Gypsy superfamily, the integrase (INT) is at the end of the open reading frame after RT and RNaseH domains. The retroviruses exhibit an additional domain envelop (ENV) in their structures (Xiong & Eickbush, 1990; Wicker *et al*., 2007; Nouroz *et al*., 2017). The LTR REs are actively proliferating in several plants and recently several autonomous and non-autonomous elements of Copia and Gypsy superfamilies were identified in *Brassica* (Nouroz *et al*., 2015), oil palm (Beule *et al*., 2015), *Musa* (Nouroz *et al*., 2017) and other genomes (Jiang & Ramachandran, 2013; Galindo-González *et al*., 2017).

Date palm *(Phoenix dactylifera* L.) is economically an important plant of family Arecaceae or Palmae, which is distributed in Northern Africa, Canary Islands, Pakistan, India and California State of USA. Date palm displays >2,000 varieties having differences in their flavor, colour, size, shape and ripening time of its highly nutritive fruits. The biochemical analysis of date palm fruit revealed that it provided substantial amounts of proteins, fats, fibers and carbohydrates. Previous studies confirmed that fruits are highly nutritive and rich in all the major nutrients required for human body (Al-Farsi *et al*., 2008; Khan *et al*., 2015). The genome sequence analysis indicated that *P.*

dactylifera had a clear genome-wide replication from ancient whole genome replications. On genomic level, limited work is done in *P. dactylifera* and the whole genome sequencing is in progress. One of the most recent report presented by a research team in Qatar based on genome assembly of *Phoenix dactylifera* from the Illumina GAII sequencing platform estimated that of the total genome size (658 Mb), 58% of the sequenced genome (382 Mb) projected around 25,059 genes (Al-Dous *et al*., 2011; Zhang *et al*., 2012). The most biologically defined repeats in *P. dactylifera* genome were retrotransposons, which accounted for 21.99% of the whole genome, of which 14.03% were Ty1/Copia, 4.17% were Ty3/Gypsy and 3.79% were the Non-LTR retrotransposons such as LINEs. The DNA transposons CACTA and MITEs constitute only 0.96% of the total date palm genome, which is very low as compared to the percentages of DNA transposons identified from other plant genomes (Al-Dous *et al*., 2011; Al-Mssallem *et al*., 2013**).** The present study was conducted to identify the LTR REs in *P. dactylifera* genome and to investigate the structural diversity, evolutionary relationships of *P. dactylifera* LTR REs among themselves and across other plant species.

Material and Methods

Bioinformatics and computational analysis: Several computation based programs were used to identify, characterize and study evolutionary relationships of LTR REs in date palm (*P. dactylifera*) genome. In the present study, 150 BAC sequences (>60kb) were retrieved from National Center for Biotechnology Information (NCBI) database to screen LTR REs. LTR_FINDER program (Xu & Wong, 2007) was utilized for the detection of LTR REs in each BAC sequence. The minimum length parameter for LTR REs was selected as 5kb while the maximum was kept as 35kb. The sizes, positions of LTR REs within the BAC sequences and sizes and positions of PBS and PPT motifs within the elements were detected. The tRNA type was also investigated by scanning these sequences against the *Zea mays* tRNA database implemented in LTR_FINDER program.

Structural domains analysis in LTR retrotransposons: The conserved domains among LTR REs were identified by running the sequences in Conserved Domain Database (CDD) of NCBI with default parameters. Each element was subjected to detect its *gag-pol* gene encoding proteins or any additional domain using the CDD and elements were classified as Copia, if they displayed domains as 5′-GAG-AP-INT-RT-RH-3′ and Gypsy having 5′-GAG-AP-RT-RH-INT-3′ protein coding domain organization.

BLAST analysis and amino acid compositions: The RT regions identified by CDD were used as query sequences in BLASTN searches against the *P. dactylifera* Nucleotide Collection (nr/nt) database. Step wise searches were performed to identify Copia or Gypsy based RT domains. Initially, the RT domain was used as a reference query in BLASTN searches and strong hits with >70% query coverage and identity in their entire lengths were collected for further analysis. The compositions or percentages of amino acid (aa) residues from various RT sequences were analyzed in Mega5 software (Tamura *et al*., 2011) using the option "Statistics" and selecting "Amino acid compositions". The amino acid percentages in various RT sequences were also calculated and presented graphically.

Multiple sequence alignment and Phylogenetic analyses: The Copia and Gypsy RT sequences from *P. dactylifera* were obtained from identified LTR REs. Fifty RT sequences (Table 1) from other organisms were collected from Gypsy database (Llorens *et al*., 2011) and were analyzed in BioEdit program (Hall, 1999). The CLASTALW multiple sequence alignment tool available in BioEdit was used to align the sequences. The sequences after alignment were visually inspected and corrected manually, if needed. The frame shifts were introduced and small insertions or deletions were removed to bring the sequences to equal sizes. The aligned sequences were imported to Mega5 (Tamura *et al*., 2011) for phylogenetic analysis. Neighbor-joining method was used to construct the tree with 1000 bootstrap replicates. The genetic distance for the amino acid sequences was computed with p-distance model.

Results

Identification of LTR REs by LTR_FINDER program: The most efficient program LTR_FINDER was used for the identification of the LTR REs from BAC genomic sequences of *P. dactylifera*. Around 150 *P. dactylifera* BAC sequences having a size of >60kb were randomly retrieved from NCBI database. Of them, 21 BACs (KE333218.1, KE333225.1, KE333230.1, KE333238.1, KE333240.1, KE333244.1, KE333250.1, KE333251.1, KE333254.1, KE333258.1, KE333274.1, KE333286.1, KE333296.1, KE333310.1, KE333312.1, KE333319.1, KE333325.1, KE333333.1, KE333334.1, KE333338.1, KE333344.1) showed the presence of retroelements in them. After detecting through LTR_FINDER, it was found that 14 out of 21 BAC sequences showed incomplete or partial LTR REs (non-autonomous), which lack one or few *gag/pol* protein coding domains, while the remaining 07 BAC sequences (KE333244.1, KE333286.1, KE333310.1, KE333319.1, KE333325.1, KE333338.1, KE333344.1) showed full length LTR REs. Of the seven LTR REs identified from these BAC sequences, 04 were Copia and 03 were Gypsy elements (Table 2).

Characterization of *gag-pol* genes polyprotein domains: The 5´-PBS and 3´-PPT motifs in Copia and Gypsy elements were identified and listed (Table 3). The PBS and PPT motifs were detected in all except PdGYP2, where PBS was not identified. The *gag-pol* genes polyprotein domains for various elements were detected by running sequences in CDD implemented in NCBI. The canonical Copia or Gypsy polyproteins were obtained from identified elements, but in rare cases, additional domains of known or unknown functions were also detected. RT and INT was present in almost all sequences, while GAG sequences was present in all identified elements. The tRNA type was also detected by scanning the sequences against *Zea mays* tRNA database implemented in LTR_FINDER program. The most frequent tRNA types were Tyr and -Asp (Table 3).

Table 1. List of various LTR retrotransposons (Copia, Gypsy superfamilies) collected from Gypsy database (Llorens *et al.*, 2011) for various phylogenetic studies. NG: Not given.

No.	Elements name	Element size (Kb)	Identified from	No.	Elements name	Element size (Kb)	Identified from
	Copia Superfamily				Gypsy Superfamily		
1.	*Copia*	5.2	*Drosophila* spp.	26.	*Gypsy*	7.4	*Drosophila melanogaster*
2.	*SIRE1*	9.8	*Glycine max*	27.	*Ty3-1*	5.5	*Saccharomyces cerevisiae*
3.	*Opie-2*	11.7	*Zea mays*	28	*Del*	9.3	*Lilium henryi*
4.	*Endovir1*	9.1	*Arabidopsis thaliana*	29.	*Galadriel*	6.2	*Lycopersicon esculentum*
5.	*ToRTL1*	9.7	*Lycopersicon esculentum*	30.	*Tntom1*	6.0	*Nicotiana tomentosiformis*
6.	*TSI-9*	9.3	*Setaria italic*	31.	*Cereba*	11.6	*Hordeum vulgare*
7.	*Araco*	4.9	*Arabidopsis thaliana*	32.	*CRM*	7.6	*Zea mays*
8.	*Oryco1*	4.9	*Oryza sativa* ssp. *japonica*	33.	*Beetle1*	6.7	*Beta vulgaris*
9.	*Vitico1*	4.6	*Vitis vinifera*	34.	*Tma*	7.3	*Arabidopsis thaliana*
10.	*Poco*	4.3	*Populus trichocarpa*	35.	*Reina*	5.4	*Zea Mays*
11.	*Melmoth*	4.8	*Brassica* spp.	36.	*Gloin*	5.4	*Arabidopsis thaliana*
12.	*Retrofit*	4.9	*Oryza longistaminata*	37.	*Legolas*	7.5	*Arabidopsis thaliana*
13.	*Koala*	5.0	*Oryza australiensis*	38.	*Monkey*	6.3	*Musa acuminate*
14.	*Osser*	4.9	*Volvox carteri*	39.	*Ifg7*	5.9	*Pinus radiate*
15.	*Tto1*	5.3	*Nicotiana tabacum*	40.	*Peabody*	7.9	*Pisum sativum*
16.	*Batata*	4.2	*Ipomoea batatas*	41.	*Retrosat-2*	12.7	*Oryza sativa*
17.	*Sto-4*	7.2	*Zea mays*	42.	*Athila4-1*	14.0	*Arabidopsis thaliana*
18.	*Fourf*	7.0	*Zea mays*	43.	*Diaspora*	11.7	*Glycine max*
19.	*Tork4*	4.9	*Lycopersicon esculentum*	44.	*Ogre*	22.7	*Pisum sativum*
20.	*RTvr2*	7.8	*Vigna radiate*	45.	*Bagy-1*	14.4	*Hordeum vulgare*
21.	*V12*	5.4	*Vitis vinifera*	46.	*RIRE2*	11.2	*Oryza sativa*
22.	*Tnt-1*	5.3	*Nicotiana tabacum*	47.	*Cinful-1*	8.6	*Zea mays*
23.	*Ty1B*	6.0	*Saccharomyces cerevisiae*	48.	*Grande1*	13.8	*Zea diploperennis*
24.	*Ty2*	6.0	*Saccharomyces cerevisiae*	49.	*Tat4-1*	11.9	*Arabidopsis thaliana*
25.	*Ty4*	6.2	*Saccharomyces cerevisiae*	50.	*Tft2*	13.2	*Arabidopsis thaliana*

Table 2. List of Copia and Gypsy elements identified in BAC clone sequences of *P. dactylifera* with their sizes, TSDs, and positions inBACs. ND: not determined.

No.	Element name	Super-family	Accession	Size	TSDs	Position in BACs
1.	PdCOP1	Copia	KE333286.1	5754	CTAGA	32815 – 38568
2.	PdCOP2	Copia	KE333325.1	11515	ND	43995 – 55509
3.	PdCOP3	Copia	KE333338.1	9249	AATT	63741 – 72989
4.	PdCOP4	Copia	KE333344.1	10615	ND	28941 – 39555
5.	PdGYP1	Gypsy	KE333244.1	11817	CTTAA	72368 – 84184
6.	PdGYP2	Gypsy	KE333310.1	11304	ND	37278 – 48581
7.	PdGYP3	Gypsy	KE333319.1	13770	ATTC	74424 – 88193

Table 3. List of Copia and Gypsy elements with their tRNA type, PBS & PPT positions and domain structures.

Element name	tRNA type	PBS (5´-3´)	Position in BACs	PPT (5´-3´)	Position	Domain structure (5´-3´)
PdCOP1	-Asp	TCCCCGGCAACAGCG	82278-82292	CCCCCTAACTAGCTT	74254-74268	GAG, INT, RT, RH
PdCOP2	Tyr	AAATGTGTTGTGGATCC	4718-4734	GAAAACTTAAAGGGG	14927-14941	GAG, RH, RT
PdCOP3	Tyr	TAAGGTTAATTGCTGATCTC	78299-78318	GACAAAAAGACAGAA	84452-84466	GAG, INT, RH, RT,
PdCOP4	Pro	TCAATGT-ACATGGATT	76186-76202	ACTAGGATTAAAAAA	86432-86446	GAG, INT, RH, RT,
PdGYP1	-Asp	AATCTCCGGCAACGGTG	53718-53734	CTTTCATTTTGTTCT	45778-45792	GAG, RH, RT, INT
PdGYP2	-----	-----------	--------	ACTCCCACTTAGCCT	64890-64904	GAG, RH, RT, INT
PdGYP3	-Asn	AACATCCTACTTGGG	37744-37758	ATCTCCCCCTTTTTG	30658-30672	GAG, UBN2, RH, RT, INT

Analysis of amino acid compositions of RT sequences from various genomes: The amino acid (aa) compositions (% ages) of 36 RT sequences of *P. dactylifera* were investigated (Fig. 1a). The average highest percentage of aa present in all *P. dactylifera* RT sequences was Leucine (9.7%), followed by Alanine (8.4%) and Lysine (8%). The lowest average percentage was that of Histidine (1.6%) and for start codon Methionine (AUG; 2.3%) (Fig. 1a). The composition of highest, medium and minimum amino acids found in RT sequences collected from various LTR REs of *Phoenix* and other organisms were graphically presented (Fig. 1b), which revealed that highest average percentage was found for Leucine (10.7%), followed by Asparaginine and Lysine (7%), while lowest average percentage was observed for Tryptophan (0.8%) (Fig. 1b).

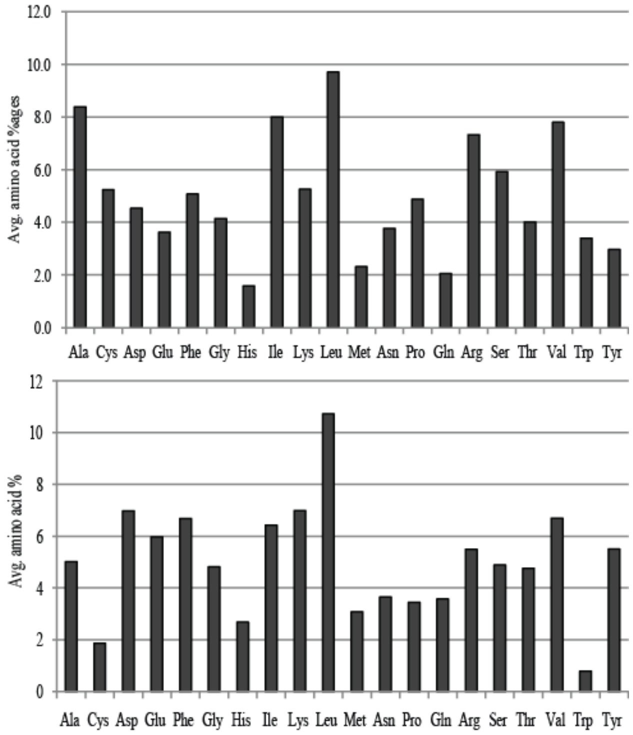

Fig. 1a-b. Average percentages of amino acids in reverse transcriptase (RT) of Gypsy and Copia elements identified from a) *Phoenix dactylifera* b) *P. dactylifera* and other plant RT sequences.

The percentages of amino acid residues across *P. dactylifera* RT sequences revealed that Leucine, Alanine were in higher amounts, and Histidine was in minimum amounts in various Copia and Gypsy elements indicating variations in percentages of amino acids. The highest percentage was observed for Leucine, while the %age of Histidine was lowest (Fig. 2). The composition of aa in *P. dactylifera a*nd other organisms RT sequences revealed (Fig. 3) that the %age of Leucine was highest, while Tryptophan was lowest (0.8%) followed by Cysteine (1.9). The highest percentages of Leucine was observed in Cereba (18%), Gloin (15.7%) and in Osser elements (14.8%). The aa counts confirmed the differences in aa compositions of Copia and Gypsy superfamilies of LTR REs (Fig. 3).

Phylogenetic analysis of Reverse transcriptase domain of LTR Retrotransposons

Phylogenetic investigation of *P. dactylifera* RT domains: Initially, 14 RT sequences from *P. dactylifera* Copia and Gypsy RT domain along with 2 known elements COPIA and GYPSY (identified from *Drosophila*) were aligned and tree was generated. The 16 RT sequences clustered into two main lineages separating the Copia and Gypsy lineage (Fig. 4). The 6 Copia RT sequences clustered into four groups. PdCOP1 formed one, PdCOP2 constituted second, PdCOP4 formed third, while PdCOP3 and COPIA elements clustered in fourth group. The other 10 Gypsy RT sequences from second lineage also segregated in 4 groups. The members of all PdGYP2 clustered in one group, PdGYP1 in second, GYPSY in third and PdGYP3 elements constituted the fourth group. In *P. dactylifera*, our results showed that the Copia and Gypsy elements clustered in two separate lineages with four Copia and three Gypsy families.

Phylogenetic analysis of *P. dactylifera* and other organisms RT domains: To investigate the phylogenetic and evolutionary relationships of Copia and Gypsy RT of *P. dactylifera* with other organisms, 20 RT domains were collected from *P. dactylifera*, while 50 RT domains of various organisms were retrieved from Gypsy database (listed in Table 1). The tree was generated using the neighbor-joining method in Mega5 with 1000 bootstrap values based on ~180aa around the conserved D-DD/E triad of RT. The results showed that the 70 RT sequences clustered into two lineages separating the Copia and Gypsy lineage (Fig. 5). The 30 Copia RT sequences further clustered into five family specific groups, where *P. dactylifera* related Copia elements were clustered in four groups. The first group designated as Copia group is comprised of 14 elements including *P. dactylifera* related PdCOP3 and other elements as V12, Batata, RTvr2, Tto1, Tnt1, Tork4, Sto4, Fourf, Copia, Melmoth, Osser, Koala, and Retrofit. The second group named Araco was comprised of elements as Araco, PdCOP4 (2 elements), Oryco1, Poco, Vitico1, Opie2, SIRE1, Endovir, TSI-9 and ToRTL1. The third group (Ty1B) was represented by 3 elements (Ty4, Ty1B, Ty2) collected from *Saccharomyces cerevisiae*. *P. dactylifera* related Copia PdCOP1 and PdCOP2 formed their respective groups (4, 5) representing single element due to their partial and varied sequences.

The 40 Gypsy RT sequences from second lineage splits into 5 groups with 19, 8, 2, 5, and 6 elements in each group. The first group designated as Del clustered 19 elements, and was comprised of 7 *P. dactylifera* PdGYP2 sequences along with Del, Antom1, Tma, Legol, Peabody, Retrosat, Bagy, Tntom1, Galadriel, Monkey, Beetle, Cereba and CRM. The second group named as Gloin clustered Gloin, Ifg7, Reina and 5 *P. dactylifera* sequences. Third group of Gypsy lineage was comprised of Gypsy and Ty3 elements. Fourth group Athila clustered PdGYP3, Diaspora, Athila and PdGYP1 sequences. Fifth group clustered the Tat4, Tft2, Ogre, Cinful, RIRE2 and Grande. Our results confirmed that the Copia and Gypsy elements were clustered in two separate lineages indicating their separate line of evolution. Further analysis confirmed that the Copia and Gypsy elements from animals and plants showed a separate line of evolution.

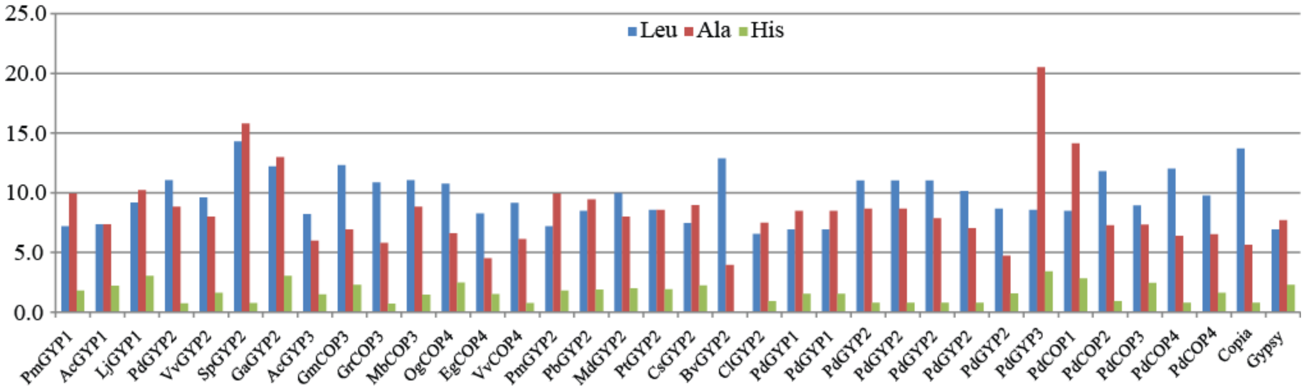

Fig. 2. Graphical presentations of aa compositions in Reverse transcriptase of *P. dactylifera* and other plants Gypsy, Copia elements.

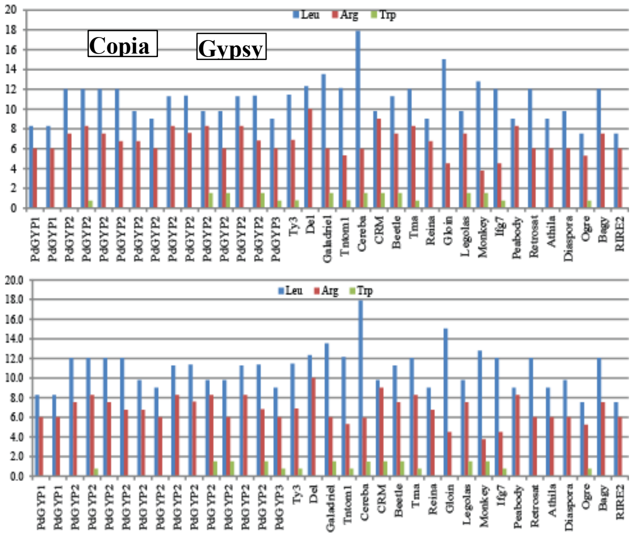

Fig. 3. Graphical presentation of amino acid residues composition in various Copia (above) and Gypsy (below) RT sequences of various LTR retrotransposons.

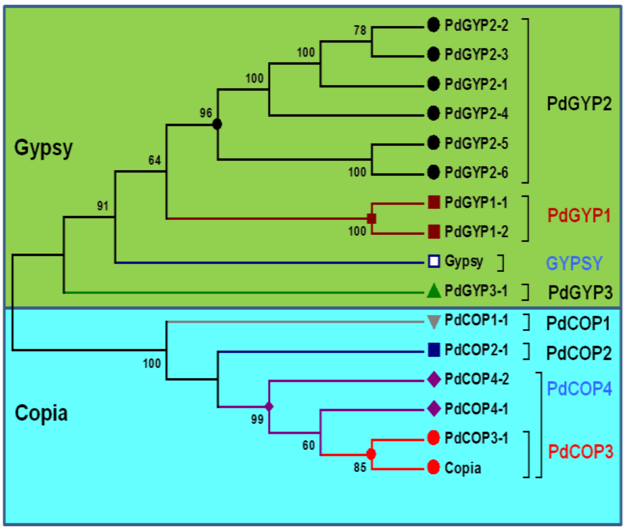

Fig. 4. Phylogenetic analysis of Copia and Gypsy RT sequences from *P. dactylifera*. A Copia and a Gypsy element of *Drosophila* was used as reference sequence. The tree was generated in Mega5 with neighbor joining method with p-distance model to calculate the genetic distance. The tree was based on 1000 bootstrap replicates. The tree clearly separated the sequences into two lineages representing Copia (blue) and Gypsy (green) further clustering them into families specific groups.

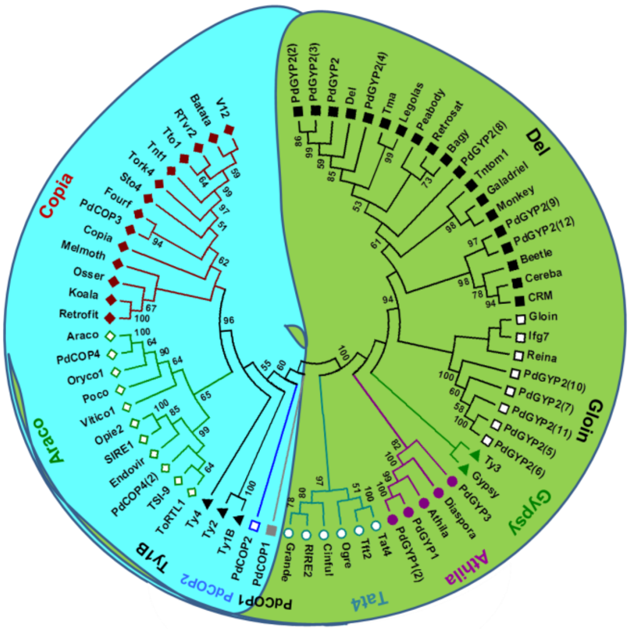

Fig. 5. Phylogenetic relationships of 70 RT sequences collected from P. *dactylifera* (20) and other plants (50). The tree was generated in Mega5 program using the neighbor joining method with p-distance model to calculate the genetic distance. The tree was based on 1000 bootstrap replicates. The tree clearly separated the sequences into two lineages representing Copia (blue) and Gypsy (green) further clustering them into families specific groups. The names of the elements are listed in Table 1.

Multiple sequence alignments of Gypsy and Copia RT sequences: Fourty Gypsy RT sequences from *P. dactylifera* and other organisms were aligned to detect homology and variations within sequences. The 180 aa residues were collected around D-DD/E triad and aligned together in BioEdit program. Several homologous motif like PCDYPN, VMPFGL-NAP, DGYQ and HHLVL were observed amoung aligned sequences (data not shown). The D-DD/E triads were found in almost all (99.9%) Gypsy RT sequences and was most conserved amoug all sequences. In contrast, several variable regions were also detected in Gypsy RT sequences. This variable regions indicate the evolution, after their seperation from a common ancester. Similarly, thirty Copia RT sequences from *P. dactylifera* and other organisms were compared to identify similarities and variations within the sequences. The most conserved regions were GLKQAP and EE-F-L. The D-DD/E triad was found in almost all (98%) Copia sequences. Highly variable motifs were also observed with aa polymorphisms in varios RT domains analysed from various organisms and *P. dactylifera*. These variable motifs showed their post-separation evolution.

Discussion

Among the TEs, the LTR REs are more copious in almost all eukaryotic genomes investigated. They can not be easily differentiated based on their sequences, but their multiple sequence alignments and phylogenetic analysis clearly distinguish them into superfamily specific lineages. The LTR REs can only be differentiated based on the arrangement of their *gag-pol* gene coding domains as Copia (5′-GAG-AP-INT-RT-RH-3′), Gypsy (5′-GAG-AP-RT-RH-INT-3′) and Retroviruses (5′-GAG-AP-RT-RH-INT-ENV-3′). RT is the major domain necessary for their mobilization and retrotransposition to a new site, so is common to all superfamilies of LTR REs and extensively used for phylogenetic analysis to distinguish the elements (Wicker *et al*., 2007; Nouroz *et al*., 2015; Nouroz *et al*., 2017). In *P. dactylifera*, retrotransposons displayed the major portion of the whole genome (21.99%), of which 14.03% are Copia, 4.17% are Gypsy and remaining 3.79% are non-LTR retrotransposons (Al-Mssallem *et al*., 2013). Several previous studies confirmed the high proportions and dominancy of Copia over Gypsy, but in few cases, the Gypsy elements predominated over Copia. The identification and characterization of Copia and Gypsy superfamilies in various plants like *Brassica*, Oil palm and *Musa* revealed the dominancy of Copia superfamily over Gypsy. It was further observed that not all the elements from these superfamilies were canonical (full length) and active, but partial, internally deleted and inactive copies were also dispersed (Al-Mssallem *et al*., 2013; Nouroz *et al*., 2015; Beule *at al*., 2015; Nouroz *et al*., 2017). In the current work, most of the elements were deleted internally or terminally and were inactive or non functional in the genome. The LTR_FINDER program also failed to identify some of these elements, which were deleted from 5′ or 3′ terminus, as the program is based to collect and identify the sequences having both 5′ and 3′ LTRs and internal *gag-pol* gene polyproteins.

With the progress in genome sequencing projects, it was demonstrated that retroelements were responsible for extensive changes in genome structure, expansion and function. Remarkable dramatic differences were reported due to the activity of LTR REs insertions among individuals belonging to the same species. An example of REs dynamics is an evolutionary adaptive mechanism within an ecological system presented by BARE1 elements in wild barley (Kalendar, 2004; Scherrer, 2005). It was demonstrated that Gypsy and Copia superfamilies were profoundly proliferating in genus *Lilium* and have paved a major role in *Lilium* genome expansion. Active REs interrupt the function of gene or modify the expression of gene and are mutagenic. The insertions of LTR REs to new sites not only hinder the regulation of nearby genes but also cause the phenotypic variations. The reactivation of LTR REs sometimes causes the somaclonal variations and produced the altered phenotypes (Lee *et al*., 2013). LTR REs proliferation not only yielded the clues about evolution of the host genomes but also drawn a line on the development of modern day crops through speciation, domestication and hybridization. Such investigations have demonstrated the differential TE proliferation in one or other genome before or after their allotetraploidization (Vukich *et al*., 2009; Beule *et al*., 2015).

The present study involved the identification of various LTR REs by LTR_FINDER program, which confirmed that LTR REs are the major component of *P. dactylifera* genome, contributing its genome duplication and evolution. Four Copia and three Gypsy families were identified with few to many copies in respective families. We presumed that several other LTR REs remained undetected due to their partial sequences or deletions at their terminal ends, which otherwise further increase the LTR REs percentage in *P. dactylifera* genome. Such deleted or partial copies are mostly residing as non active elements and mostly are non functional in genome. Both Copia and Gypsy superfamilies predominated in several plant genomes including important crops (Jiang & Ramachandran, 2013; Galindo-González *et al*., 2017). In sunflower (*Helianthus annus*) *Brassica* and oil palm genome, high proportions of Copia and Gypsy elements were reported, which indicated that LTR REs were present in almost all plants investigated but the ratio of proliferation of these elements varied (Vukich *et al*., 2009; Nouroz *et al*., 2015; Beule *et al*., 2015).

The conserved RT domains of Copia and Gypsy superfamilies amplified from mungbean genomes were cloned and sequenced. These RT sequences were detected from other papilionoid legumes of the same and different tribes (Loteae, Trifoleae, Cicereae), which revealed their presence prior to the radiation of papilionoid legumes and also supported the close relationship of Phaseoleae and Loteae tribes of Papilionoideae subfamily analyzed recently. In contrast, mungbean Copia and Gypsy RT sequences showed significant homologies to those of unrelated plant species revealing their origin from different plant ancestries and also demonstrated that the heterogeneous population of related elements already existed during the evolution of these genera from their common ancestor (Dixit *et al*., 2006). In the present study, three Gypsy and four Copia families were identified with few to several copies in respective families. The phylogenetic tree clearly clustered the sequences into Copia and Gypsy lineages indicating a separate line of evolution of both superfamilies. Furthermore the plants, animals and fungal based elements clustered in their respective groups indicating their separate evolutionary line. Several previous investigations also confirmed the segregation of Copia and Gypsy lineages. In *Helianthus*, the phylogenetic analysis of Copia and Gypsy REs clearly distinguished them in Copia and Gypsy based lineages. Moreover the Copia elements were more uniform in comparison to Gypsy (Vukich, 2009; Nouroz *et al*., 2017). The present study included the evolutionary and comparative analysis of Gypsy and Copia RT sequences, as RT was the most suitable domain in LTR REs to deduce the phylogeny. A total of 70 RT sequences

(20 from *P. dactylifera* and 50 from other genomes) were analyzed for evolutionary relationships. The current study indicated the clustering of Gypsy and Copia elements in their particular lineages, groups and subgroups. The Gypsy lineage was divided into four groups as Gypsy, PdGYP1, PdGYP2, PdGYP3 and its related sequences. Similarly Copia lineage was further classified into Copia, PdCOP1, PdCOP2, PdCOP4, PdCOP3 and its homologous copies (Figs. 4, 5). The position of PdCOP1 and PdCOP2 as out-group of Copia lineage (Fig. 5) is skeptical, as both families clustered on margin of Copia lineage. This might be due to RT sequences having gaps generated by deletions, frame shifts mutations and pre mature stop codons in them. It was further strengthen by the fact that no more copies of both elements were detected from *P. dactylifera* genome indicating they were no longer functional in the genome.

The aa residues among various RT sequences were also investigated to detect the aa polymorphism among these domains. High variability among compositions and percentages of aa revealed post divergence evolution. The detailed analysis confirmed that amino acid Leucine was highest in proportions followed by Lysine and Methionine and Tryptophan were in lowest percentages. The nucleotide polymorphisms are the source of evolution in these elements. The multiple sequence alignment of RT sequences from 70 Copia and Gypsy elements predicted several conserved and variable regions (motifs). The study confirmed that, in plants and animals LTR REs, RT is the most conserved domain with many conserved motifs in it. The D-DD/E triads detected in almost 98-99% of RT sequences indicate the most conserved motif. Our results are accordance to the results of Flavell *et al*., (1992), who detected D-DD triad motifs from almost all aligned RT sequences. The detailed evolutionary analysis of Copia and Gypsy RT sequences collected from *Brassica*, *Musa* and several other plants species also revealed the presence of such conserved motifs in RT domain. The D-DD triad was most conserved region, while YVDD or YNDD were the most conserved motifs in almost all RT sequences (Nouroz *et al*., 2015; Nouroz *et al*., 2017).

Conclusions

The present study presents the comprehensive analysis and description of LTR retrotransposon actively proliferating or silently residing in the date palm genome. Although rare copies of these elements were detected but structural and genomic diversity existed among these sequences representing several events of nucleotide polymorphisms. The structural and phylogenetic analysis segregate them into 4 Copia and 3 Gypsy related families and we assume that many more copies from diverse families are residing in the date palm genome. The segregation of Copia and Gypsy lineages clearly indicate separate line of evolution of both superfamilies. Our identification and annotation of LTR REs in date palm genome will be helpful in future studies aiming to identify such more families of REs, their major role in date palm genome plasticity and localization on chromosomes.

Acknowledgements

We are thankful to Professor Dr. J.S. (Pat) Heslop-Harrison for his valuable comments in experimental design. Special thanks to Professor Dr. Manzoor Hussain for providing the research facilities at Department of Botany, Hazara University Mansehra, Pakistan.

References

Al-Dous, E.K., B. George, M.E. Al-Mahmoud, M.Y. Al-Jaber, H Wang, Y.M. Salameh, E.K. Al-Azwani, S. Chaluvadi, A.C. Pontaroli, J. DeBarry, V. Arondel, J. Ohlrogge, I.J. Saie, K.M. Suliman-Elmeer, J.L Bennetzen, R.R. Kruegger, J.A. Malek and J.A. 2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat. Biotechnol.*, 29: 521-527.

Al-Farsi, M.A. and C.Y. Lee. 2008. Nutritional and functional properties of dates: a review. *Crit. Rev. Food Sci. Nutr.*, 48: 877-887.

Al-Mssallem, I.S., S. Hu, X. Zhang, Q. Lin, W. Liu, J. Tan, X. Yu, J. Liu, L. Pan, T. Zhang, Y. Yin, C. Xin, H. Wu, G. Zhang, M.M. Ba Abdullah, D. Huang, Y. Fang, Y.O. Alnakhli, S. Jia, A. Yin, E.M. Alhuzimi, B.A. Alsaihati, S.A. Al-Owayyed, D. Zhao, S. Zhang, N.A. Al-Otaibi, G. Sun, M.A. Majrashi, F. Li, Tala, J. Wang, Q. Yun, N.A. Alnassar, L. Wang, M. Yang, R.F. Al-Jelaify, K. Liu, S. Gao, K. Chen, S.R. Alkhaldi, G. Liu, M. Zhang, H. Guo and J. Yu. 2013. Genome sequence of the date palm *Phoenix dactylifera* L. *Nat. Commun*., 4: 2274.

Beulé, T., M.D. Agbessi, S. Dussert, E. Jaligot and R. Guyot 2015. Genome-wide analysis of LTR-retrotransposons in oil palm. *BMC Genomics*, 16: 795.

Dixit, A., K.H. Ma, J.W. Yu, E.G. Cho and Y.J. Park. 2006. Reverse transcriptase domain sequences from Mungbean (*Vigna radiata*) LTR retrotransposons: sequence characterization and phylogenetic analysis. *Plant Cell Rep*., 25(2): 100-111.

Feschotte, N., N. Giang and S.R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet*., 3: 329-341.

Flavell, A.J., D.B. Smith and A Kumar 1992. Extreme heterogeneity of Ty1-copia group retrotransposons in plants. *Mol. Gen. Genet*., 231: 233-242.

Galindo-González, L., C. Mhiri, M.K. Deyholos, M-A. Grandbastien. 2017. LTR-retrotransposons in plants: Engines of evolution. *Gene*, 626: 14-25.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucl. Acids Symp. Ser.*, 41: 95-98.

Jiang, S-Y., and S. Ramachandran. 2013. Genome-wide survey and comparative analysis of LTR Retrotransposons and their captured genes in rice and sorghum. *PLoS ONE*, 8(7): e71118.

Kalendar, R., C.M. Vicient, O. Peleg, K. Anamthawat-jonsson, A. Bolshoy and A.H. Schulman. 2004. Large retrotransposon derivatives: abundant, conserved but non-autonomous retroelements of barley and related genomes. *Genetics*, 166: 1437-1450.

Kapitonov, V.V., and J. Jurka 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* 9: 411-412.

Khan, H., F. Nouroz, M.F. Khan and S. Rizwan. 2015. Nutritional values of selected Date palm varieties in Pakistan. *American-Eurasian J. Agric. Environ. Sci.*, 15(5): 764-768.

Lee, S.I., K.C. Park, J.H. Son, Y.J. Hwang, K.B. Lim, Y.S. Song, J.H. Kim and N.S Kim. 2013. Isolation and characterization of novel Ty1-copia-like retrotransposons from lily. *Genome*, 56(9): 495-503.

Llorens, C., R. Futami and L. Covelli. 2011. The Gypsy Database (GyDB) of Mobile Genetic Elements: Release 2.0. *Nucl. Acids Res.*, 39: 70-74.

Nouroz, F., S. Noreen and J.S. Heslop-Harrison. 2015. Identification and evolutionary genomics of novel LTR retrotransposons in *Brassica*. *Turk. J. Biol*., 39: 740-757.

Nouroz, F., S. Noreen, H. Ahmed and J.S. Heslop-Harrison. 2017. The LTR Retrotransposons landscape in *Musa* genome. *Mol. Gen. Genomics*, 292: 1051-1067.

Scherrer, B., E. Isidore and P. Klein. 2005. Large Intraspecific haplotype variability at the *Rph7* locus results from rapid and recent divergence in the barley genome. *The Plant Cell*, 17(2): 361-374.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and M Kumar. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol*., 28(10): 2731-2739.

Vukich, M., A.H. Schulman, T. Giordani, L. Natali, R. Kalender and A. Cavallini. 2009. Genetic variability in sunflower (*Helianthus annuus L.*) and in the *Helianthus* genus as assessed by retrotransposon-based molecular markers. *Theor. Appl. Genet*., 119(6): 1027-1038.

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

Xiong, Y. and T.H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J*., 9: 3353-3362.

Xu, Z. and H. Wang. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucl. Acids Res*., 35: 265-268.

Zhang, G.Y., L. Pan and Y. Yin. 2012. Large-scale collection and annotation of gene models for date palm (*Phoenix dactylifera* L.). *Plant Mol. Biol.,* 79: 521-536.

(Received for publication 17 March 2018)