Chapter

Characterization, Comparative, and Phylogenetic Analyses of Retrotransposons in Diverse Plant Genomes

AloysiusBrown, Orlex B. Yllano, Leilani D. Arce, Ephraim A. Evangelista, Ferdinand A. Esplana, Lester Harris R. Catolico and Merbeth Christine L. Pedro

Abstract

Retrotransposons are transposable elements that use reverse transcriptase as an intermediate to copy and paste themselves into a genome via transcription. The presence of retrotransposons is ubiquitous in the genomes of eukaryotic organisms. This study analyzed the structures and determined the comparative distributions and relatedness of retrotransposons across diverse orders (34) and families (58) of kingdom Plantae. *In silico* analyses were conducted on 134 plant retrotransposon sequences using ClustalW, EMBOSS Transeq, Motif Finder, and MEGA X. So far, the analysis of these plant retrotransposons showed a significant genomic relationship among bryophytes and angiosperms (216), bryophytes and gymnosperms (75), pteridophytes and angiosperms (35), pteridophytes and gymnosperms (28), and gymnosperms and angiosperms (70). There were 13 homologous plant retrotransposons, 30 conserved domains, motifs (reverse transcriptase, integrase, and gag domains), and nine significant phylogenetic lineages identified. This study provided comprehensive information on the structures, motifs, domains, and phylogenetic relationships of retrotransposons across diverse orders and families of kingdom Plantae. The ubiquitousness of retrotransposons across diverse taxa makes it an excellent molecular marker to better understand the complexity and dynamics of plant genomes.

Keywords: transposable elements, retrotransposon, genetic polymorphism, phylogenetic analysis, genome

1. Introduction

Retrotransposons can move within genomes due to their highly effective transposition mechanism. Because of this high level of transposition, their presence is a significant feature of plant genomes and other eukaryotic organisms. Since the discovery of transposable elements (TE) by Barbara McClintock more than seven decades ago, there have been several challenges in studying the structures of retrotransposons due to their repetitive structure, diversity in form, their large

number in a genome, and their ability to replicate so frequently [1]. Even studying closely related genomes does not overcome this problem since retrotransposons also tend to be highly species-specific, a trait that makes them difficult to classify. Research has shown that they are not merely transient components of a genome but are instrumental in genomic development and adaptation, influencing these genomes from how chromosomes are structured to helping activate certain genes under certain conditions [2]. The interaction of retrotransposons with a host genome is not a simple one. Pieces of evidence have shown that they have helped shaped genomes for an extended period. In some cases, this has imparted important genetic traits to their host organisms. In others, they have been linked to mutagenesis and disease, prompting their host to develop regulatory safeguards to suppress and limit their activities [3].

Recent advances in sequencing technologies have come a long way in helping unravel the structure of plant genomes. Plant genomes are some of the most complex and diverse among known eukaryotic kingdoms [4] and vary widely in size across kingdom Plantae, with the smallest genomes sequenced so far being from green algae species [5] and the largest being *Pinus taeda*, which is around 22 Gbp in length [6]. A significant portion of the plant genome comprises transposable elements, the so-called "jumping genes" [7]. The diversity and size variation across plant genomes is primarily attributed to the activity of these transposable elements [8]. The transposable elements are known to have viral origins; in particular, retrotransposons structures closely resemble retroviruses without the gene for the viral envelope or with a nonfunctional envelope gene. It is hypothesized that transposable elements enter the genomes of eukaryotes through infection by ancient viruses and remained as parasitic elements in their host genomes [9]. More studies are needed to understand better the complexity of plant retrotransposons and unravel its salient features.

1.1 Classes and types of transposable elements

The complexity and diversity of transposable elements coupled with the availability of recent genomic sequences in the genebanks have generated various groupings of TEs. However, concerted efforts have been made to come up with a generally accepted and unified nomenclature. The replication process employed by transposable elements are used to classify them into two large groups [10]. Retrotransposons or Class I transposable elements use the enzyme reverse transcriptase to copy and paste themselves in the genome and are the most abundant type in plant genomes. DNA transposons or Class II transposable elements use other enzymes, including DNA polymerase and transposase, to copy and insert themselves into genomes [11]. This copy and paste mechanism is responsible for the significant number of transposable elements in eukaryotic genomes.

Class I Transposable Elements or Retrotransposons consists of the long terminal repeats (LTRs) retrotransposons and the non-long terminal repeats (non-LTRs) retrotransposons. These LTR retrotransposons and non-LTR retrotransposons are further subdivided based on their dynamics in the genome. The autonomous retrotransposons can be independently mobile, while the nonautonomous retrotransposons necessitate the presence of TEs for their movement. Some of the LTR retrotransposons in eukaryotes include Gypsy, Copia, BEL, DIRS, ERVI, ERV2, and ERV3 superfamilies. In contrast, superfamilies of non-LTR retrotransposons includes SINE1,2,3, LINES, CR1, CRE, I, RTE, TX1, Jockey, Penelope, R2, R4, RandI, Rex1, L1, and NeSL [12, 13].

A less well-studied class of retrotransposons in plant genomes are non-LTR retrotransposons. These are the LINEs-Long Interspersed Nuclear Elements and

the SINEs-Short Interspersed Nuclear Elements. They do not exhibit much activity in plant genomes and constitute around 33.5% or about one third of the human genome [13]. More so, they contribute to new insertions in the human genome and have been linked to mutagenesis and human diseases [14].

LINEs are considered the oldest class of retrotransposons in plant genomes. Evidence suggests that they are highly regulated or inactive since their transcription is rarely observed in plant genomes [15]. In contrast, studies have shown that the ancient activity of SINEs helped shaped the genomic diversification of some monocot species [16] and the heterogeneity of many eukaryotic genomes, but apart from this, little is known so far of their activity in plant genomes [17]. With this, there is a need to study and characterize the diverse retrotransposons and understand how and to what extent they influence changes in a host genome.

1.2 Characterization of retrotransposons

The presence of transposable elements in an organism has many implications for its genomic activity. Depending on the region of the chromosome they are located on, they may affect what type of genes are expressed in the genome and the functions of these genes [18]. Gypsy retrotransposons have a widespread and more diverse position on the chromosomes in plant genomes, while Copia retrotransposons tend to cluster in proximal regions of the chromosomes they are located on [19]. However, it is worth pointing out that LTR retrotransposons tend to group in different chromosomal regions regardless of their lineages [20]. Research into plant genomic structures has yielded valuable insight into the characterization of retrotransposons due to their ubiquitous presence in plant genomes [21]. They are subclassified into LINES and SINES [22]. The LTR-retrotransposons are further classified into "superfamilies" based on their genetic sequences, namely, the Copia superfamily, the Gypsy superfamily, Bel-Pao, retrovirus, and endogenous retrovirus superfamilies [23]. Of these, the most widespread in plant genomes and the most well studied are the Gypsy and Copia superfamilies. Gypsy retrotransposons are differentiated from Copia retrotransposons by the position of the integrase protein in their genetic sequence. In gypsy retrotransposons, integrase is situated after the reverse transcriptase in the genetic sequence and before the reverse transcriptase in Copia retrotransposons [24]. Phylogenetic analyses and time of divergence are used to further divide these superfamilies into different lineages. The Copia superfamily comprised TORK, Bianca, Ale, Maximus lineages Gypsy superfamily of Attila, CRM, Del, and Galadriel lineages [25]. LTR-retrotransposons showcase such variety in number, position, and distribution in their host genome due to their unique ability to express the independent activity and replicate themselves numerous times on chromosomes [26].

A key feature of LTR retrotransposons and the structure that gives them their name is the presence of two homologous structures called long terminal repeats at both ends of their genetic sequence. These DNA sequences can vary in size from a hundred bps to thousands of bps [27]. These LTRs are non-coding regions that bracket the internal coding regions and are also a component of retroviral sequences [28]. LTR retrotransposons vary widely in size and functional characteristics. In plants, they have been documented as short as four kbp in *Helianthus* species [29] to over 23 kbp in *Populus trichocarpa* [30]. The structures of LTR retrotransposons are organized into one or several Open Reading Frames (ORF) [31]. The ORFs contains genetic information for the pol and gag genes and are integral to transcription in the host genome [32]. Like their retroviral counterparts, the gag genes encode functional polyproteins, and the pol gene usually contains the reverse transcriptase. These genes are typically separated by stop codons [33]. The pol gene encodes three

important proteins, each of which has a crucial role in retrotransposal replication in the genome [34]. These proteins are Integrase, Protease, and Reverse Transcriptase [35]. Because retrotransposons replicate similarly to viruses, and their replication can lead to mutations and disrupt DNA repair, there are genomic mechanisms in place to silence their activity [36]. To escape this silencing, LTR retrotransposons may possess another region called the chromodomain. One mechanism the cell uses to silence retrotransposons is the formation of heterochromatin near areas of retrotransposon activity [37]. The presence of heterochromatin makes it difficult for the retrotransposon proteins to access the cell DNA, suppressing replication [38]. The chromodomain region encodes a protein that helps the retrotransposon escape silencing by manipulating these heterochromatins. Chromodomains are found upstream of the 3′ end of the genetic sequence in retrotransposons [39].

1.3 Mechanism of action

Retrotransposons insert and reinsert themselves in a host genome by transcription. This process is accomplished by the reverse transcription of an RNA intermediate transcript. This transcript is the template that is used to generate new copies of the retrotransposon [40]. The reverse transcription of retrotransposons is a complex procedure. In LTR retrotransposon, the process is helped by the long terminal repeats at each end of their structure that acts as start sites for replicating the internal region. The replication of this internal region occurs in opposite directions to produce two DNA strands. At the 3′ end, tRNA binds to the initiation site of the left LTR and replicates one of the two DNA strands. At the right LTR, a Polypurine Tract, which acts as a primer, binds immediately upstream of this region and replicates the second of the two DNA strands [41].

The mRNA template is synthesized first in the replication of retrotransposons. This mRNA template is then translated into proteins utilized in the process. The mRNA template has a U region and a short repeat sequence at each end. tRNA acts as a primer and binds to a primer binding site on the mRNA. This initiates the production of minus (−) strand DNA through the catalyzation of Reverse Transcriptase. The synthesized DNA reaches the U5 region at the 5′ end of the template and pairs with the repeat sequence at the 3′ end of the genomic RNA. Once synthesis of this first DNA strand is complete, the enzyme RNase H deteriorates the genomic RNA template, leaving only fragments. These fragments then prime the synthesis of the second DNA strand. As with the first strand, Reverse Transcriptase synthesizes another DNA strand but uses the first DNA strand as a template. At the end of this process, a linear double-stranded DNA is made with an LTR region (comprised of the repeat sequence, U5, and U3 regions) at each end. The enzyme integrase then inserts this new retrotransposon DNA into the host chromosomal DNA by using the 3’ OH of each strand to integrate at target sites a few base pairs apart in the genome [42].

1.4 Role of retrotransposons

Retrotransposons are known to be major drivers of genomic diversity and homogeneity during the development of eukaryotic genomes. Presently, their activity in plant genomes is regulated by different mechanisms, but they are still capable of bursts of activity when reactivated by mutations, adjacent gene expression, or environmental factors [43]. Grandbastien [44] has noted that all the retrotransposons that are known to be active in plant genomes are usually dormant during their host development but become active in response to environmental stressors. This could be linked to retrotransposons being proliferators of genomic diversity

since their activation by stresses induces survival genes to turn on. The study by Hilbricht et al. [45] on *Craterostigma plantagineum* dehydration led to the isolation and identification of a retroelement gene, the Craterostigma desiccation-tolerant (CDT-1) gene, that is turned on by dehydration and imparts drought-resistant properties to the plant. This is also in line with Zhao et al. [46], which found a potential link of the OAR1 gene to the tolerance of osmotic and alkaline stresses in *Arabidopsis thaliana*. Though often characterized by their propensity to initiate mutagenesis, retrotransposons have been shown to affect the expression of genes they are adjacent to in the genome and even help regulate the structure of centromeres [47], as noted in an investigation of maize species by Gao et al. [48]. Analysis of tomato plants demonstrated that differences in volatile esters between two different colored fruits of different species of these plants are linked to the placement of retrotransposons near the family of esterases that exhibits a high level of enzyme activity. This placement results in a higher expression of the esterase, resulting in the reduced levels of multiple esters [49]. Retrotransposons have also been linked to disease resistance in plants. A study showed that activation of athila LTR retrotransposons led to genome expansion in *Capsicum baccatum* by increasing the number of a disease-resistant gene family [50] and analysis of *Phaeodactylum tricornutum* cells showed the activity of LTR-retrotransposon initiate a plant response to a decrease in nitrate and when exposed to reactive aldehydes that stress diatoms and leads to cell death [51]. Analysis of retrotransposon families in sorghum species shows that their activity influences genomic adaptation and diversity [52]. This finding suggests that retrotransposons play vital roles in regulating genes that encode functional proteins [53]. A study of Thale Cress and Adzuki bean seedlings treated with the DNA methylation inhibitor zebularine increased activity and accumulation of the retrotransposon ONSEN in the seedlings treated than in the control seedlings [54]. These studies point to the pivotal role of retrotransposons in plants' adaptation to their environment and their contribution to genomic diversity.

This study compared, characterized, identified shared patterns, and determined the relationships of different retrotransposons across diverse plant taxa.

2. Materials and methods

To assemble the plant retrotransposon library, we collected genomic DNA sequences deposited at the National Center for Biotechnology Information (NCBI) nucleotide database. These were then further sorted to include only sequences with 300 to 800 base pairs in length. In total, 134 retrotransposon sequences were selected and analyzed in this study. Of these, 54 were angiosperms, 46 were gymnosperms, 11 were pteridophytes, three were liverworts, and 20 were bryophytes. The sequences were downloaded in the FASTA format and saved in a text document for further analyses. To study the characteristics of the plant retrotransposon sequences and identify homogeny, multiple sequence alignment (ClustalW) program was utilized. The parameters of the ClustalW analysis were defined as follows: Pairwise Alignment was set to slow and accurate for DNA sequences only. The Gap Open Penalty was set to 15 and the Gap Extension Penalty to 6.66. The Weight Matrix used was the International Union of Biochemistry (IUB) matrix for DNA sequences. These same parameters were used for the multiple sequence analysis with hydrophilic gaps included in the computation.

Motif analyses were performed on the plant retrotransposon sequences to identify motifs, protein domains, and conserved domains. The nucleotide sequences were translated into their corresponding amino acid (aa) sequences with the EMBOSS Transeq tool developed by the European Bioinformatics Institute. The algorithm was

set to translate the nucleotide sequences into the three possible reading frames using the standard codon table. The translated aa sequences were then analyzed for protein domains, families, and functional sites using the PROSITE tool developed by the Swiss Institute of Bioinformatics [55] and the MOTIF Finder program of the Kyoto University Bioinformatics Center [56]. All three reading frames were analyzed to ensure the proper frame would be used for motif identification. The aligned retrotransposon sequences were analyzed using the MEGA-X. The software was used to construct a maximum likelihood phylogenetic tree with the Tamura-Nei method used to account for the substitution rate differences between nucleotides and the inequality of nucleotide frequencies. The Nearest-Neighbor-Interchange was used as the heuristic method to improve the likelihood of the tree. The phylogenetic tree generated by the MEGA-X program was then modified in the MEGA X Tree Topology Editor to produce a circular phylogenetic diagram for better data visualization.

3. Results and discussion

3.1 Multiple sequence alignment

Figure 1 shows the alignment scores of sequences produced from the multiple sequence alignment analysis performed in the clustalW program. These scores represent the pairwise alignment between each pair of retrotransposon sequences. The cutoff alignment score was set at 50 percent identity between two aligned sequences.

In total, there were 870 pairwise alignments with a 50 to 100 percent alignment score. Fifty-five percent (476) of the alignments had a percent identity in the range of 50 to 59. Thirty-two percent (281) had a percent identity in the range of 60 to 69. Seven percent (65) had a percent identity in the range of 70 to 79, 4% (35) had a percent identity in the range of 80 to 89, and 2% (13) had a percent identity in the range of 90 to 100. The multiple sequence alignment scores of 40% and higher are considered significant. However, an alignment less than 40% is considered too divergent [57]. The alignment score for this multiple sequence analysis was set to 50% to include only highly significant alignments.

3.2 Identification of homologous sequences

Table 1 contains the aligned sequences with the highest alignment score. There is a diversity in the relationship of these sequences. *T. pellucida* 1 to *T. pellucida* 2

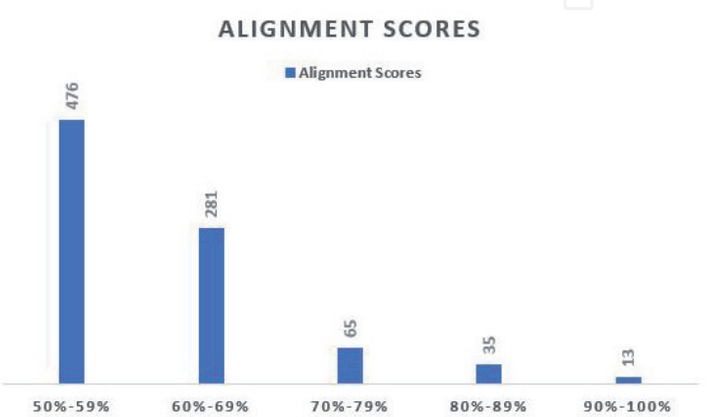

Figure 1.
Significant pairwise alignment scores of 134 plant retrotransposon sequences.

Sequences Aligned	Aligned Score
A. concolor: *A. veitchii*	90
L. saxicola: *P. schreberi*	91
S. cooperi: *D. truncatula*	91
D. polysetum: *L. glaucum*	93
P. cuspidatum: *R. canescens*	93
L. gmelinii: *L. czekanowskii*	94
A. araucana: *A. brownii*	94
A. sativa: *A. sterilis*	94
A. ipaensis: *A. hypogaea*	95
P. patens: *M. polymorpha 1*	99
T. pellucida1: *T. pellucida 2*	100
V. dubyana: *F. antipyretica*	100
N. tetragona: *M. grandiflora 2*	100

Table 1.
Aligned sequences with an alignment score of 90 to 100.

are of the same species but clones. Each plant in the sequence pairs alignments of *A. concolor* to *A. veitchii*, *L. gmelinii* to *L. czekanowskii*, *A. sativa* to *A. sterilis*, *A. ipaensis* to *A. hypogaea*, and *V. dubyana* to *F. antipyretica*, belong to the same genus. *A. araucana* and *A. brownii* belong to the same family. The sequences aligned in each alignment pair of *L. saxicola* to *P. schreberi* and *D. polysetum* to *L. glaucum* belong to the same order, while those in the pairs of *S. cooperi* to *D. truncatula* and *P. cuspidatum* to *R. canescens* belong to the same class. Sequences belonging to only the same division can be closely related, as in the case of *P. patens* to *M. polymorpha* with a 99% identity and *N. tetragona* to *M. grandflora* with a 100% identity. The pair of sequences aligned in the same genus had the highest number of aligned pairs.

The results above confirm the highly conserved nature of retrotransposons. This was supported by the study of retrotransposons in mammals [58]. Despite their enormous size and diversity, it has been noted that similar retrotransposons tend to cluster together in similar genomes of hosts belonging to the same order, family, or class [59]. Specific types of retrotransposons belonging to the same family or lineage can be conserved across a particular kingdom or division [60]. The presence of homologs can be inferred from these aligned sequences considering their high percent identity and their distribution to different species [61]. An alignment of 90 and higher was used as the cutoff value for homolog identification [62].

3.3 Conservation of retrotransposons

Table 2 is a summation of retrotransposons sequences with an alignment score of 80 to 89. This is the pairwise alignment score between pairs of sequences.

Aligned sequence pairs in the same genus were: *L. occidentalis* to *L. sibirica*, *A. concolor* to *A. balsamea*, *L. occidentalis* to *L. kaempferia*, *P. rubens* to *P. schrenkiana*, *A. veitchii* to *A. balsamea*, and *L. kaempferi* to *L. sibirica*. More so, the aligned sequences pairs that had sequences in the same family were: *L. sibirica* to *P. rubens*, *L. sibirica* to *P. schrenkiana*, *L. occidentalis* to *P. contorta*, *P. contorta* to *L. sibirica*, *L. occidentalis* to *P. schrenkiana*, *P. contorta* to *P. schrenkiana*, *L. kaempferi* to *P. rubens*, *P. contorta* to *L. kaempferi*, *P. contorta* to *P. reubens*, *L. occidentalis* to *P. rubens*, and *L. kaempferi* to *P. schrenkiana*. At the same order level, the following were the aligned

Sequences Aligned	Aligned Score	Sequences Aligned	Aligned Score
L. saxicola: *P. polyantha*	80	*P. contorta*: *L. sibirica*	86
L. saxicola: *D. polysetum*	80	*S. obtusum*: *A. rupestris*	87
D. polysetum: *R. canescens*	80	*L. occidentalis*: *L. kaempferi*	87
L. glaucum: *P. cuspidatum*	80	*L. occidentalis*: *P. schrenkiana*	87
P. polyantha: *L. glaucum*	81	*P. contorta*: *P. schrenkiana*	87
P. polyantha: *P. cuspidatum*	81	*L. kaempferi*: *P. rubens*	87
P. polyantha: *R. canescens*	81	*P. rubens*: *P. schrenkiana*	87
P. polyantha: *D. polysetum*	82	*J. communis*: *T. baccata*	87
P. polyantha: *H. ciliata*	82	*S. cooperi*: *N. exaltata*	88
D. polysetum: *P. cuspidatum*	82	*D. truncatula*: *N. exaltata*	88
G. biloba2: *P. rubens*	82	*P. contorta*: *L. kaempferi*	88
P. schreberi: *P. polyantha*	83	*P. contorta*: *P. rubens*	88
G. biloba2: *P. contorta*	83	*A. veitchii*: *A. balsamea*	88
L. occidentalis: *L. sibirica*	83	*L. occidentalis*: *P. rubens*	89
L. sibirica: *P. rubens*	84	*L. kaempferi*: *L. sibirica*	89
L. sibirica: *P. schrenkiana*	84	*L. kaempferi*: *P. schrenkiana*	89
G. biloba2: *P. schrenkiana*	85		
A. concolor: *A. balsamea*	85		
L. occidentalis: *P. contorta*	86		

Table 2.
Aligned sequences with an alignment score of 80 to 89.

sequence pairs: *L. saxicola* to *P. polyantha*, *J. communis* to *T. baccata*, and *D. tuncatula* to *N. exaltata*. Aligned sequence pairs that had sequences in the same class were: *L. saxicola* to *D. polysetum*, *D. polysetum* to *R. canescens*, *L. glaucum* to *P. cuspidatum*, *P. polyantha* to *L. glaucum*, *P. polyantha* to *P. cuspidatum*, *P. polyantha* to *R. canescens*, *P. polyantha* to *D. polysetum*, *P. polyantha* to *H. ciliata*, *D. polysetum* to *P. cuspidatum*, *P. schreberi* to *P. polyantha* and *S. cooperi* to *N. exaltata*. Likewise, the aligned sequence pairs with sequences in the same division were: *G. biloba* to *P. schrenkiana*, *G. biloba* to *P. contorta*, *G. biloba* to *P. rubens*, and *S. obtusum* to *A. rupestris*.

3.4 Motifs and domains

Molecular characterization is important in understanding the nature of any genetic element and its insertion origin in a genome. Molecular characterization provides a detailed description of the structure of a genetic sequence, changes that it induces in a genome, and how it affects genetic expression [63]. Characterization is an important feature in the study of retrotransposons. It is also used for classifying retrotransposons [64], uncovering their associations in a genome [65, 66], and discovering new types of retrotransposons (**Table 3**) [66].

The identification of the reverse transcriptase motif in these retrotransposon sequences is significant because it is not only integral to the replication process of retrotransposons but is one of the most significant parts of their structure [67]. The reverse transcriptase type identified in these sequences was only found in LTR retrotransposons and retroviruses. The presence of this reverse transcriptase type

Reverse transcriptase (RNA-dependent DNA polymerase)	**Simian taste bud-specific gene product family**
Reverse transcriptase (RNA-dependent DNA polymerase)	Simian taste bud-specific gene product family
Tsi6	BAFF-R, TALL-1 binding
RNase H-like domain found in reverse transcriptase	Zinc knuckle
Tc5 transposase DNA-binding domain	GAG-polyprotein viral zinc-finger
Peptidase propeptide and YPEB domain	Mis6
Integrase zinc-binding domain	Protein prenyltransferase alpha subunit repeat
Integrase core domain	Chromatin remodeling factor Mit1 C-terminal Zn finger 2
H_2C_2 zinc finger	5′-3′ exonuclease, N-terminal resolvase-like domain
gag-polypeptide of LTR copia-type	Retrotransposon gag protein
Aspartyl protease	C2H2 zinc-finger
gag-polyprotein putative aspartyl protease	GAG-pre-integrase domain
Retroviral aspartyl protease	Eukaryotic translation initiation factor 3 subunit G
Domain of unknown function	3′ exoribonuclease family, domain 2
Putative peptidase (DUF1758)	HicA toxin of bacterial toxin-antitoxin,
Fimbrial assembly protein (PilN)	BRK domain

Table 3.
Motifs and domains identified by the MOTIF finder.

usually indicates that the sequence is a retrotransposon mobile element or a retrovirus [68]. Reverse transcriptase gene identification could be used to identify retrotransposon sequences due to their high specificity. Reverse transcriptases are known to be multidomain enzymes, with notable domains being the catalytic domain and the RNase H domain [69]. The Tc5 transposase DNA-binding domain is a structural motif found in many proteins that regulate gene expression. The RNase H-like domain found in these retrotransposon sequences belongs to a reverse transcriptase subfamily that shares sequence similarity with reverse transcriptases from endogenous retroviruses of the zebrafish and the Moloney mouse leukemia retroviruses [69, 70]. This finding strengthens the viral origins of retrotransposons in eukaryotes.

The presence of the zinc-binding domain indicates the presence of integrase since it is one of the domains in the integrase enzyme. Integrase allows retroviruses and retroelements to insert their DNA into a host genome [71]. The integrase core domain that was also detected in this sequence is one of the three known domains of the integrase enzyme. It is the catalytic domain that catalyzes the transfer of retroviral or retrotransposal DNA made by reverse transcriptase to the site in the genome where it will be inserted [72]. GAG-Pre-Integrase domain lies upstream of the integrase region in retroviral polyproteins. They are usually connected to elements that assist in retroviral insertion [73].

The Copia family of retrotransposons is a large retrotransposon family active in the genomes of plants. It is classified under the long terminal repeats retrotransposons along with the Gypsy family [74]. The GAG Polypeptide of the LTR-Copia type domain is highly conserved and found only in Copia retrotransposons [75]. This domain was identified in seven species: *G. biloba*, *L. occidentalis*, *P. contorta*, *L. kaempferi*, *L. sibirica*, *P. rubens*, *P. schrenkiana*, definitively identifying them as Copia family retrotransposons.

Some domains were identified that are not generally associated with retrotransposons. The Hic A toxin functions as an mRNA interferase in bacteria and archaea species [76], Tsi6 is a bacterial immunity protein, and the Fimbrial Assembly Protein functions in the production of bacterial fimbria used for cellular attachment [77]. The Simian taste-bud specific gene is found in primates, and mutations of this gene have been linked to follicular lymphomas [78]. The Mis6 protein is integral for chromosome segregation during mitosis, and the protein prenyltransferase alpha subunit repeat functions in protein prenylation. In contrast, the eukaryotic translation initiation factor 3 subunit G initiates protein synthesis [79], and the BAFF-R is a polypeptide that binds to the ligands of TALL-1, a tumor necrosis factor that initiates inflammation in humans [80]. Zinger finger proteins are a large family of proteins noted for their role as transcription factors and their ability to bind Zn ions. Several of these protein types were identified from the plant retrotransposon sequences, including: H2C2 zinc finger, zinc knuckle, GAG-polyprotein viral zinc-finger, chromatin remodeling factor Mit1 C-terminal Zn finger 2, and C2H2 zinc-finger. Recent studies revealed that they are highly involved in regulating plant response to abiotic stressors in their environment [81]. Peptidase propeptide and YPEB domain, putative peptidase (DUF1758), 5′-3′ exonuclease, N-terminal resolvase-like domain, and the BRK domain are all hypothetical proteins of which little to nothing is known of their activity presently [82].

3.5 Patterns and profiles

The PROSITE database has an extensive collection of protein families, subfamilies, domains, and motifs managed by the Swiss Institute of Bioinformatics [83]. The database is organized into unique protein profiles and patterns to identify functional sites, domains, and protein families [84].

Table 4 contains the PROSITE patterns of four motifs found in the PROSITE database. IPNS_1 was found in *E. arvense*, ASP_PROTEASE in *G. biloba*, ZINC_PROTEASE in *P. contorta*, and TONB_DEPENDENT REC 1 in *T. aestivum*. Isopenicillin N synthase signature 1 is an enzyme found in bacterial and fungal species instrumental in the production of cephalosporin and penicillin [85]. TonB-dependent receptor proteins signature 1 is a type of protein found in *E. coli* involved in cellular transportation of substrates into the periplasmic space by active transport [86]. The presence of these bacterial domains in plant retrotransposons supports their role as genetic reservoirs. Because of their transposable nature, they can "jump" from bacterial plasmids onto chromosomes, carrying genes with them [87].

Aspartyl proteases are a family of enzymes that hydrolyzes peptide bonds [88]. They are very diverse and can be found in species including humans, retroviruses, plants, and fungi. In retroviruses, they are usually encoded in the pol gene as part of a polypeptide [89]. The zinc protease utilizes zinc in its catalytic function to break down polyproteins. Retrotransposon's polyproteins are very important elements

Found Motif	Description
IPNS_1	PS00185, Isopenicillin N synthase signature 1
ASP_PROTEASE	PS00141, Eukaryotic and viral aspartyl proteases active site
ZINC_PROTEASE	PS00142, Neutral zinc metallopeptidases, zinc-binding region signature
TONB_DEPENDENT_REC_1	PS00430, TonB-dependent receptor proteins signature 1

Table 4.
Patterns identified from plant retrotransposons.

of their replication mechanism, and these proteases enable the hydrolysis of these larger proteins into smaller functional polypeptides [90]. The Pol polyproteins and proteases are needed in retrotransposon replication to form mRNA and its packaging in the transposition of retrotransposons [91].

Table 5 contains the four PROSITE profiles identified in the retrotransposon sequences. The Reverse Transcriptase catalytic domain profile was detected in 25 different species, the Integrase catalytic domain profile in four species, and the zinc finger CCHC-type profile, and the zinc finger SWIM-type profile in one species each. Reverse Transcriptase is a multidomain enzyme consisting of two domains: The Catalytic Domain and the RNase H binding domain. These two domains are used to perform the three enzymatic actions of Reverse Transcriptase [92]. The Catalytic Domain carries out the polymerase activities using DNA-dependent polymerase and RNA-dependent polymerase. The RNase H domain is responsible for the ribonuclease enzymatic activity [93]. Together, these two reverse transcriptase domains enable the "copy" part of the retrotransposon replication mechanism.

The integrase is also a multidomain enzyme (**Table 5**). Its structure consists of three domains integral to its function: An N-terminal zinc finger domain, a C-terminal DNA binding domain, and the Integrase core domain between them [94]. These integrase domains are responsible for the "paste" part of retrotransposon replication, allowing them to transpose themselves into other sites of their host genome [95]. The CCHC zinc finger is associated with retroviruses. They are found in the capsid protein and aids the virus in host infection [96]. The presence of this protein confirms the relationship between retroviruses and retrotransposons. They have developed from retroviruses and still retain proteins for the viral capsids and envelopes [97]. These proteins have been repurposed from aiding in viral infection to assisting in DNA and RNA binding [98].

The SWIM-type zinc finger was isolated from a retrotransposon sequence of *Manihot esculenta* (**Table 5**). The SWIM zinc finger is found in all major eukaryotic groups. It has a strong association with the plant MuDR family of transposases. These enzymes belong to the MuDR transposon, a part of one of the largest families of transposons in plants, the Mu family [99]. They are known mutagens, which is in line with one of the characteristics of transposable elements as instigators of mutagenesis in their host genomes [100].

3.6 Phylogenetic analysis

The phylogenetic analysis uses characters like nucleotide or amino acid sequences to construct a tree to show the relationship among different taxa at the molecular level. This analysis can also investigate domain relationships within an individual taxon [101], and this has become an essential tool for comparing genetic data between different species and groups [102].

Found Motif	Description
RT_POL	PS50878, Reverse transcriptase (RT) catalytic domain profile
INTEGRASE	PS50994, Integrase catalytic domain profile
ZF_CCHC	PS50158, Zinc finger CCHC-type profile
ZF_SWIM	PS50966, Zinc finger SWIM-type profile

Table 5.
Profiles identified from plant retrotransposons.

The history of these retrotransposons was analyzed and created using the Maximum Likelihood method and Tamura-Nei model [103]. The initial tree and guide tree for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Tamura-Nei model. All the codon positions included were 1st+2nd+3rd+noncoding translated proteins. The final dataset consisted of 892 positions. The MEGA X program was used to investigate relationship analyses [104]. The neighbor-joining tree algorithm was tested with bootstrap replicates of 1000 [105] and the resulting bootstrap values displayed above the tree's nodes. The cutoff value for the tree branches was set at 70% [106] to identify lineage clusters. The largest of these clusters with values above the cutoff is the group "C," which contained well-supported branches of retrotransposon lineages. All the plant sequences in this group were from bryophytes. Well-supported groups were group "B" (*M. grandiflora 1* and *M. polymorpha 2*), group "E" (*A. sativa* and *A. sterilis*), group "F" (*S. cooperi* and *D. truncatula*), group "G" (*M. esculenta* and *F. virosa*), and group "I" (*N. tetragona* and *M. grandiflora 2*) (**Figure 2**). Likewise, moderately supported groups (**Figure 3**) were group "A" (*M. polymorpha 3* and *M. notabilis*), group "D" (*V. speciosa 2* and *B. papyrifera*), and group "H" (*P. patens 2* and *L. lagopus 2*) [107].

Figure 4 shows the circular ideogram of diverse retrotransposons across range-wide orders and families of the kingdom Plantae. This ideogram was constructed to ensure holistic visualization of large-scale data and efficiently visualize enormous amounts of genomic information.

The "red" group on the upper right was represented by a cluster of retrotransposons from gymnosperms, while the "blue" group had retrotransposons originating from angiosperm. The "green" group had two novel retrotransposons, namely, Silava and Romani, distinct for gymnosperms. The "yellow" group comprises Gypsy family retrotransposons from angiosperms except for M. polymorpha and *P. massoniana*, a liverwort and gymnosperm, respectively. The "orange" group is the largest cluster composed of Gypsy family retrotransposons from the bryophytes. The "purple" group is a clade of two gymnosperm retrotransposons from the Gypsy and Copia families. In contrast, the "brown" group is a clade of two gymnosperms Copia retrotransposons, and the "pacific blue" group is a clade of non-LTR retrotransposons from two eudicots. The "ruby" group is a cluster of Copia family retrotransposons, and the "Davidson orange" group comprises mostly Gypsy retrotransposons with some notable novel-type families (Cereba, N1, Osr30, and Silava). Osr30 is distinct to *O. sativa*, Cereba to cereal plants, and Silava to gymnosperms. The "pink" group is a cluster of angiosperm Gypsy retrotransposons, the "medium green" is a cluster of giant ferns Cassandra retrotransposons, and the "tyrian purple" is a cluster of Poaceae family retrotransposons. The "lochmara blue" is a cluster of Copia-like retrotransposons, and the "deep red" group is a cluster of

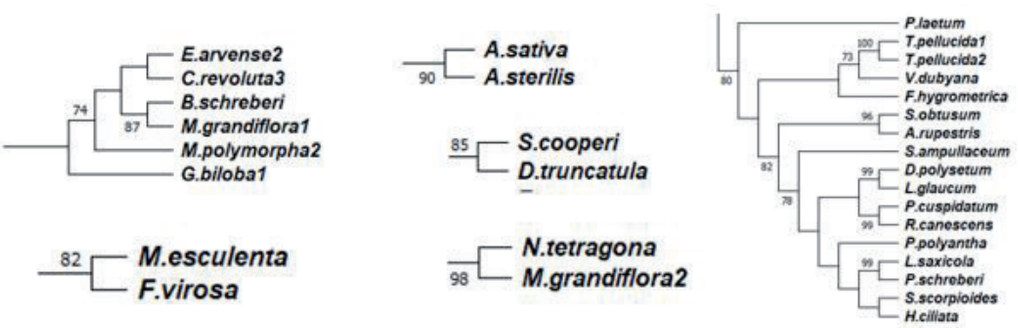

Figure 2.
Well supported bootstrap branches based on the phylogenetic analysis.

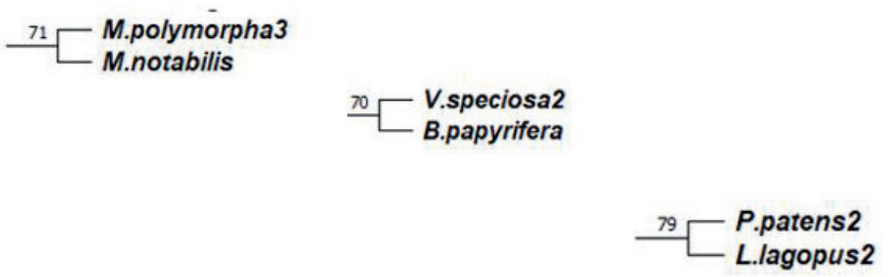

Figure 3.
Moderately supported bootstrap branches based on the phylogenetic analysis.

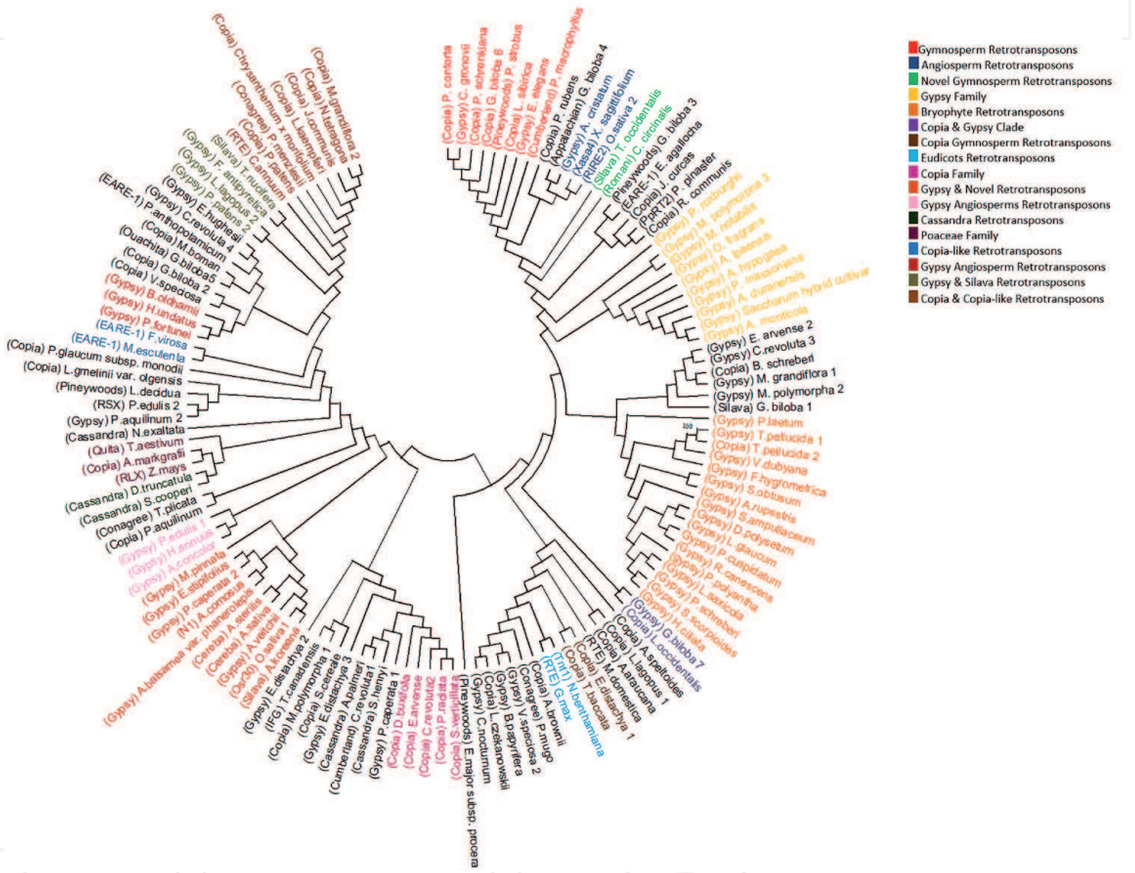

Figure 4.
Circular ideogram of retrotransposons across diverse plant genomes.

angiosperm retrotransposons [108]. The "verdigris" group is a cluster of Gypsy retrotransposons with the inclusion of Silava retrotransposons. It was noted that Silava retrotransposons tend to cluster with Gypsy retrotransposons. The "saddle brown" group is a cluster of Copia retrotransposons with two novel Copia-like retrotransposons (RTE & Conagree). All black clusters formed the mixed groups.

Retrotransposons of the gypsy family tend to cluster together (**Figure 4**). The Gypsy family is the largest group, forming a large cluster of bryophyte sequences and eudicot sequences with few liverworts and gymnosperms sequences forming outgroups of these clades. Gypsy retrotransposons are very diversified and more widespread in plant genomes than Copia retrotransposons [109]. Retrotransposons of the Copia family tend to be grouped based on the plant group they belong to. These retrotransposons are interspersed with novel families of retrotransposons that are Copia-like in structure. Copia-like retrotransposons are common in plant genomes and are identified by their reverse transcriptase, similar in structure to the

Copia family retrotransposons [110]. Gymnosperm retrotransposons are grouped together regardless of family, and they are associated with monocot retrotransposons. Possibly, this attribute could be the result of retrotransposal duplication events in these genomes [111]. Notably, retrotransposons are more active in the Poaceae family [112], leading to the genesis of more unique and novel retrotransposon families.

4. Conclusions

Retrotransposons are such a significant part of plant genomes that they warrant more studies to understand them better. Retrotransposons were conserved in nature, tended to cluster in different plant families and classes, and revealed significant genome relationships between different families within a plant division. Retrotransposons were characterized by certain motifs and domains useful in classifying them and helping understand their role in plant genomes. Plant retrotransposons exhibited much diversification while also retaining the conservation of certain parts of their structures. Retrotransposons in plant genomes retained genes from other life domains, just as they reserved harmful genes. They can also keep useful genes essential in helping their hosts survive adverse conditions. Findings in the PROSITE amino acid patterns and profiles found that some of these plant retrotransposons contain viral, bacterial, fungal, and mammalian genes. The high specificity of retrotransposal Reverse Transcriptase could be used as an important tool in identifying retrotransposons. More so, phylogenetic analysis revealed the relationships of the retrotransposons and unveiled their diversification into several lineages. This study provided valuable information on the characteristics, patterns, profiles, diversity, and phylogenetic relationship of retrotransposons across the range-wide plant orders and families and are necessary in understanding the functions, complexity, and dynamics of plant genomes.

Acknowledgements

We would like to thank the faculty members of the Department of Biology, College of Science and Technology, Adventist University of the Philippines, and reviewers for the valuable comments. The National Center for Biotechnology Information, Bethesda MD, USA for the DNA sequences. We are grateful to Sir Owen E. Pitakia, Dr. Edwin Balila, and Dr. Lorcelie Taclan for their indispensable counsels and support.

Conflict of interest

The authors declare no conflict of interest.

Author details

Aloysius Brown[1], Orlex B. Yllano[1,2*], Leilani D. Arce[3], Ephraim A. Evangelista[4], Ferdinand A. Esplana[4], Lester Harris R. Catolico[5] and Merbeth Christine L. Pedro[5]

1 Department of Biology, College of Science and Technology, Adventist University of the Philippines, Silang, Cavite, Philippines, Brewerville City, Liberia

2 Cell and Molecular Biology Laboratory, Department of Biology, CST, Adventist University of the Philippines, Silang Cavite, Philippines

3 Botany and Systematics Laboratory, Department of Biology, CST, Adventist University of the Philippines, Silang Cavite, Philippines

4 Microbiology Laboratory, Department of Biology, CST, Adventist University of the Philippines, Silang Cavite, Philippines

5 Anatomy and Physiology Laboratory, Department of Biology, CST, Adventist University of the Philippines, Silang Cavite, Philippines

*Address all correspondence to: obyllano@aup.edu.ph

References

[1] Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. 2018;46(21):e126.

[2] Mustafin RN, Khusnutdinova EK. The Role of Transposable Elements in Emergence of Metazoa. Biochemistry (Mosc). 2018;83(3):185-99.

[3] Mita P, Boeke JD. How retrotransposons shape genome regulation. Curr Opin Genet Dev. 2016;37:90-100.

[4] Bennett MD. Variation in Genomic Form in Plants and Its Ecological Implications. New Phytologist. 1987;106(s1):177-200.

[5] Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in plants. Curr Opin Genet Dev. 2015;35:119-25.

[6] Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15(3):R59.

[7] Gao L, McCarthy EM, Ganko EW, McDonald JF. Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequences. BMC Genomics. 2004;5(1):18.

[8] Orozco-Arias S, Isaza G, Guyot R. Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning. Int J Mol Sci. 2019 Aug 6;20(15).

[9] Malik HS, Henikoff S, Eickbush TH. Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. Genome Res. 2000;10(9):1307-18.

[10] Kazazian HH. Mobile elements: drivers of genome evolution. Science. 2004;303(5664):1626-32.

[11] Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification [Internet]. Vol. 10, Mobile DNA. Mob DNA; 2019. Available from: https://pubmed.ncbi.nlm.nih.gov/30622655/

[12] Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet. 2008 May;9(5):411-2; author reply 414.

[13] Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009;10(10):691-703.

[14] Schmidt T. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol Biol. 1999. 40(6):903-10.

[15] Mao H, Wang H. Distribution, Diversity, and Long-Term Retention of Grass Short Interspersed Nuclear Elements (SINEs). Genome Biol Evol. 2017;9(8):2048-56.

[16] Wenke T, Döbel T, Sörensen TR, Junghans H, Weisshaar B, Schmidt T. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. Plant Cell. 2011;23(9):3117-28.

[17] Sahebi M, Hanafi MM, van Wijnen AJ, Rice D, Rafii MY, Azizi P, et al. Contribution of transposable elements in the plant's genome. Gene. 2018;665:155-66.

[18] Nagaki K, Shibata F, Kanatani A, Kashihara K, Murata M. Isolation of centromeric-tandem repetitive DNA sequences by chromatin affinity purification using a HaloTag7-fused centromere-specific histone H3 in tobacco. Plant Cell Rep. 2012;31(4):771-9.

[19] Joly-Lopez Z, Bureau TE. Diversity and evolution of transposable elements in Arabidopsis. Chromosome Res. 2014;22(2):203-16.

[20] Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326(5956):1112-5.

[21] Xiong Y, Eickbush TH. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol. 1988;5(6):675-90.

[22] Zhang L, Yan L, Jiang J, Wang Y, Jiang Y, Yan T, et al. The structure and retrotransposition mechanism of LTR-retrotransposons in the asexual yeast *Candida albicans*. Virulence. 2014;5(6):655-64.

[23] Janicki M, Rooke R, Yang G. Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. Chromosome Res. 2011;19(6):787-808.

[24] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973-82.

[25] Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. Ann Bot. 2005;95(1):127-32.

[26] Du D, Du X, Mattia MR, Wang Y, Yu Q, Huang M, et al. LTR retrotransposons from the *Citrus x clementina* genome: characterization and application. Tree Genetics and Genomes. 2018;14(4):43.

[27] Rho M, Choi J-H, Kim S, Lynch M, Tang H. De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics. 2007;8:90.

[28] Mascagni F, Giordani T, Ceccarelli M, Cavallini A, Natali L. Genome-wide analysis of LTR-retrotransposon diversity and its impact on the evolution of the genus *Helianthus* (L.). BMC Genomics. 2017;18(1):634.

[29] Cossu RM, Buti M, Giordani T, Natali L, Cavallini A. A computational study of the dynamics of LTR retrotransposons in the *Populus trichocarpa* genome. Tree Genetics and Genomes. 2012;8(1):61-75.

[30] Chang W, Jääskeläinen M, Li S, Schulman AH. BARE retrotransposons are translated and replicated via distinct RNA pools. PLoS One. 2013;8(8):e72270.

[31] Mascagni F, Barghini E, Giordani T, Rieseberg LH, Cavallini A, Natali L. Repetitive DNA and Plant Domestication: Variation in Copy Number and Proximity to Genes of LTR-Retrotransposons among Wild and Cultivated Sunflower (*Helianthus annuus*) Genotypes. Genome Biol Evol. 2015;7(12):3368-82.

[32] Joly-Lopez Z, Bureau TE. Exaptation of transposable element coding sequences. Curr Opin Genet Dev. 2018;49:34-42.

[33] Piednoël M, Carrete-Vega G, Renner SS. Characterization of the LTR retrotransposon repertoire of a plant clade of six diploid and one tetraploid species. Plant J. 2013;75(4):699-709.

[34] Paz RC, Kozaczek ME, Rosli HG, Andino NP, Sanchez-Puerta MV. Diversity, distribution and dynamics of

full-length Copia and Gypsy LTR retroelements in *Solanum lycopersicum*. Genetica. 2017;145(4-5):417-30.

[35] Usai G, Mascagni F, Natali L, Giordani T, Cavallini A. Comparative genome-wide analysis of repetitive DNA in the genus *Populus* L. Tree Genetics & Genomes. 2017;13(5):96.

[36] Sanchez DH, Gaubert H, Drost H-G, Zabet NR, Paszkowski J. High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. Nat Commun. 2017;8(1):1283.

[37] Ma J, Devos KM, Bennetzen JL. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. 2004;14(5):860-9.

[38] Ragupathy R, Banks T, Cloutier S. Molecular characterization of the Sasanda LTR copia retrotransposon family uncovers their recent amplification in *Triticum aestivum* (L.) genome. Mol Genet Genomics. 2010;283(3):255-71.

[39] Curcio MJ, Garfinkel DJ. Regulation of retrotransposition in *Saccharomyces cerevisiae*. Mol Microbiol. 1991;5(8):1823-9.

[40] Boeke JD, Corces VG. Transcription and reverse transcription of retrotransposons. Annu Rev Microbiol. 1989;43:403-34.

[41] Nishihara H. Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. Genes Genet Syst. 2020 Jan 30;94(6):269-81.

[42] Finnegan DJ. Retrotransposons. Current Biology. 2012;22(11):R432-7.

[43] Hirochika H, Okamoto H, Kakutani T. Silencing of retrotransposons in Arabidopsis and reactivation by the ddm1 mutation. Plant Cell. 2000;12(3):357-69.

[44] Grandbastien M-A. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim Biophys Acta. 2015;1849(4):403-16.

[45] Hilbricht T, Varotto S, Sgaramella V, Bartels D, Salamini F, Furini A. Retrotransposons and siRNA have a role in the evolution of desiccation tolerance leading to resurrection of the plant *Craterostigma plantagineum*. New Phytol. 2008;179(3):877-87.

[46] Zhao Y, Xu T, Shen C-Y, Xu G-H, Chen S-X, Song L-Z, et al. Identification of a retroelement from the resurrection plant *Boea hygrometrica* that confers osmotic and alkaline tolerance in *Arabidopsis thaliana*. PLoS One. 2014;9(5):e98098.

[47] Defraia C, Slotkin RK. Analysis of retrotransposon activity in plants. Methods Mol Biol. 2014;1112:195-210.

[48] Gao D, Jiang N, Wing RA, Jiang J, Jackson SA. Transposons play an important role in the evolution and diversification of centromeres among closely related species. Front Plant Sci. 2015;6:216.

[49] Goulet C, Mageroy MH, Lam NB, Floystad A, Tieman DM, Klee HJ. Role of an esterase in flavor volatile variation within the tomato clade. Proc Natl Acad Sci U S A. 2012;109(46):19009-14.

[50] Kim S, Park J, Yeom S-I, Kim Y-M, Seo E, Kim K-T, et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome Biol. 2017;18(1):210.

[51] Maumus F, Allen AE, Mhiri C, Hu H, Jabbari K, Vardi A, et al. Potential impact of stress activated

retrotransposons on genome evolution in a marine diatom. BMC Genomics. 2009;10:624.

[52] Guo H, Jiao Y, Tan X, Wang X, Huang X, Jin H, et al. Gene duplication and genetic innovation in cereal genomes. Genome Res. 2019;29(2):261-9.

[53] Lin J, Cai Y, Huang G, Yang Y, Li Y, Wang K, et al. Analysis of the chromatin binding affinity of retrotransposases reveals novel roles in diploid and tetraploid cotton. J Integr Plant Biol. 2019;61(1):32-44.

[54] Boonjing P, Masuta Y, Nozawa K, Kato A, Ito H. The effect of zebularine on the heat-activated retrotransposon ONSEN in *Arabidopsis thaliana* and *Vigna angularis*. Genes Genet Syst. 2020;95(4):165-72.

[55] ExPASy – PROSITE [Internet]. [cited 2021 Apr 8]. Available from: https://prosite.expasy.org/

[56] MOTIF: Searching Protein Sequence Motifs [Internet]. [cited 2021 Apr 8]. Available from: https://www.genome.jp/tools/motif/

[57] Thompson JD, Gibson TJ, Higgins DG. Multiple Sequence Alignment Using ClustalW and ClustalX. Current Protocols in Bioinformatics. 2003;00(1):2.3.1-2.3.22.

[58] Buckley RM, Kortschak RD, Raison JM, Adelson DL. Similar Evolutionary Trajectories for Retrotransposon Accumulation in Mammals. Genome Biol Evol. 2017;9(9):2336-53.

[59] Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst. 2020;94(6):233-52.

[60] Suguiyama VF, Vasconcelos LAB, Rossi MM, Biondo C, de Setta N. The population genetic structure approach adds new insights into the evolution of plant LTR retrotransposon lineages. PLoS One. 2019;14(5):e0214542.

[61] Masuta Y, Kawabe A, Nozawa K, Naito K, Kato A, Ito H. Characterization of a heat-activated retrotransposon in *Vigna angularis*. Breed Sci. 2018;68(2):168-76.

[62] Reznikoff WS, Bordenstein SR, Apodaca J. Comparative sequence analysis of IS50/Tn5 transposase. J Bacteriol. 2004;186(24):8240-7.

[63] Oliva N, Florida Cueto-Reaño M, Trijatmiko KR, Samia M, Welsch R, Schaub P, et al. Molecular characterization and safety assessment of biofortified provitamin A rice. Scientific Reports. 2020;10(1):1376.

[64] Cao Y, Jiang Y, Ding M, He S, Zhang H, Lin L, et al. Molecular characterization of a transcriptionally active Ty1/copia-like retrotransposon in Gossypium. Plant Cell Rep. 2015;34(6):1037-47.

[65] Cavalcante MG, Souza LF, Vicari MR, de Bastos CEM, de Sousa JV, Nagamachi CY, et al. Molecular cytogenetics characterization of *Rhinoclemmys punctularia* (Testudines, Geoemydidae) and description of a Gypsy-H3 association in its genome. Gene. 2020;738:144477.

[66] Rezende-Teixeira P, Siviero F, Brandão AS, Santelli RV, Machado-Santelli GM. Molecular characterization of a retrotransposon in the *Rhynchosciara americana* genome and its association with telomere. Chromosome Res. 2008;16(5):729-42.

[67] Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 1990 Oct;9(10):3353-62.

[68] CDD Conserved Protein Domain Family: RT_LTR [Internet]. [cited 2021

Apr 18]. Available from: https://www.ncbi.nlm.nih.gov/Structure/cdd/cd01647

[69] Das D, Georgiadis MM. The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. Structure. 2004;12(5):819-29.

[70] Nowak E, Potrzebowski W, Konarev PV, Rausch JW, Bona MK, Svergun DI, et al. Structural analysis of monomeric retroviral reverse transcriptase in complex with an RNA/DNA hybrid. Nucleic Acids Res. 2013;41(6):3874-87.

[71] Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P. Retroviral intasome assembly and inhibition of DNA strand transfer. Nature. 2010;464(7286):232-6.

[72] Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science. 1994;266(5193):1981-6.

[73] CDD Conserved Protein Domain Family: gag_pre-integrs [Internet]. [cited 2021 Apr 19]. Available from: https://www.ncbi.nlm.nih.gov/Structure/cdd/cl16514

[74] Yang S, Zeng K, Chen K, Zhao X, Wu J, Huang Y, et al. Sequence Evolution, Abundance, and Chromosomal Distribution of Ty1-copia Retrotransposons in the *Saccharum spontaneum* Genome. Cytogenet Genome Res. 2020;160(5):272-82.

[75] Miller K, Rosenbaum J, Zbrzezna V, Pogo AO. The nucleotide sequence of *Drosophila melanogaster* copia-specific 2.1-kb mRNA. Nucleic Acids Res. 1989;17(5):2134.

[76] Jørgensen MG, Pandey DP, Jaskolska M, Gerdes K. HicA of *Escherichia coli* defines a novel family of translation-independent mRNA interferases in bacteria and archaea. J Bacteriol. 2009;191(4):1191-9.

[77] Martin PR, Watson AA, McCaul TF, Mattick JS. Characterization of a five-gene cluster required for the biogenesis of type 4 fimbriae in *Pseudomonas aeruginosa*. Mol Microbiol. 1995;16(3):497-508.

[78] Skibola CF, Bracci PM, Halperin E, Conde L, Craig DW, Agana L, et al. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. Nat Genet. 2009;41(8):873-5.

[79] Boguski M, Murray A, Powers S. Novel repetitive sequence motifs in the alpha and beta subunits of prenyl-protein transferases and homology of the alpha subunit to the MAD2 gene product of yeast. New Biol. 1992 Apr 1; 4(4):408-11.

[80] Liu Y, Hong X, Kappler J, Jiang L, Zhang R, Xu L, et al. Ligand-receptor binding revealed by the TNF family member TALL-1. Nature. 2003;423(6935):49-56.

[81] Han G, Lu C, Guo J, Qiao Z, Sui N, Qiu N, et al. C2H2 Zinc Finger Proteins: Master Regulators of Abiotic Stress Responses in Plants. Front Plant Sci [Internet]. 2020 [cited 2021 May 30];11. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2020.00115/full

[82] InterPro [Internet]. [cited 2021 May 30]. Available from: https://www.ebi.ac.uk/interpro/

[83] Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, et al. The PROSITE database. Nucleic Acids Res. 2006;34(Database issue):D227-230.

[84] Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V,

Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010;38(Database issue):D161-166.

[85] Cohen G, Shiffman D, Mevarech M, Aharonowitz Y. Microbial isopenicillin N synthase genes: Structure, function, diversity and evolution. Trends in Biotechnology. 1990;8:105-11.

[86] TonB-dependent receptor, conserved site (IPR010917) – InterPro entry – InterPro [Internet]. [cited 2021 Apr 18]. Available from: https://www.ebi.ac.uk/interpro/entry/InterPro/IPR010917/

[87] Babakhani S, Oloomi M. Transposons: the agents of antibiotic resistance in bacteria. J Basic Microbiol. 2018;58(11):905-17.

[88] Rao JK, Erickson JW, Wlodawer A. Structural and evolutionary relationships between retroviral and eucaryotic aspartic proteinases. Biochemistry. 1991;30(19):4663-71.

[89] Davies DR. The structure and function of the aspartic proteinases. Annu Rev Biophys Biophys Chem. 1990;19:189-215.

[90] Gazda LD, Joóné Matúz K, Nagy T, Mótyán JA, Tőzsér J. Biochemical characterization of Ty1 retrotransposon protease. PLoS One. 2020;15(1):e0227062.

[91] Checkley MA, Mitchell JA, Eizenstat LD, Lockett SJ, Garfinkel DJ. Ty1 gag enhances the stability and nuclear export of Ty1 mRNA. Traffic. 2013;14(1):57-69.

[92] Katz RA, Skalka AM. The retroviral enzymes. Annu Rev Biochem. 1994;63:133-73.

[93] Herschhorn A, Hizi A. Retroviral reverse transcriptases. Cell Mol Life Sci. 2010;67(16):2717-47.

[94] Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. Annu Rev Biochem. 1998;67:1-25.

[95] Chen JC, Krucinski J, Miercke LJ, Finer-Moore JS, Tang AH, Leavitt AD, et al. Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. Proc Natl Acad Sci U S A. 2000;97(15):8233-8.

[96] Katz RA, Jentoft JE. What is the role of the cys-his motif in retroviral nucleocapsid (NC) proteins? Bioessays. 1989;11(6):176-81.

[97] Dodonova SO, Prinz S, Bilanchone V, Sandmeyer S, Briggs JAG. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. Proc Natl Acad Sci U S A. 2019;116(20):10048-57.

[98] Sandmeyer SB, Clemens KA. Function of a retrotransposon nucleocapsid protein. RNA Biol. 2010;7(6):642-54.

[99] Makarova KS, Aravind L, Koonin EV. SWIM, a novel Zn-chelating domain present in bacteria, archaea and eukaryotes. Trends Biochem Sci. 2002;27(8):384-6.

[100] Sabot F, Schulman AH. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity (Edinb). 2006;97(6):381-8.

[101] Zhang D, Kan X, Huss SE, Jiang L, Chen L-Q, Hu Y. Using Phylogenetic Analysis to Investigate Eukaryotic Gene Origin. J Vis Exp. 2018;(138).

[102] Hillis DM. Phylogenetic analysis. Current Biology. 1997;7(3):R129-31.

[103] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial

DNA in humans and chimpanzees. Mol Biol Evol. 1993;10(3):512-26.

[104] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-9.

[105] Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. How many bootstrap replicates are necessary? J Comput Biol. 2010;17(3):337-54.

[106] Sacks-Davis R, Daraganova G, Aitken C, Higgs P, Tracy L, Bowden S, et al. Hepatitis C virus phylogenetic clustering is associated with the social-injecting network in a cohort of people who inject drugs. PLoS One. 2012;7(10):e47335.

[107] Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. PNAS. 1996;93(23):13429-13429.

[108] Huang J, Wang Y, Liu W, Shen X, Fan Q, Jian S, et al. EARE-1, a Transcriptionally Active Ty1/Copia-Like Retrotransposon Has Colonized the Genome of *Excoecaria agallocha* through Horizontal Transfer. Front Plant Sci [Internet]. 2017 [cited 2021 Jun 4];8. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2017.00045/full

[109] Suoniemi A, Tanskanen J, Schulman AH. Gypsy-like retrotransposons are widespread in the plant kingdom. Plant J. 1998;13(5):699-705.

[110] Voytas DF, Cummings MP, Koniczny A, Ausubel FM, Rodermel SR. copia-like retrotransposons are ubiquitous among plants. Proc Natl Acad Sci U S A. 1992;89(15):7124-8.

[111] Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. BMC Biology. 2012;10(1):84.

[112] Vicient CM, Jääskeläinen MJ, Kalendar R, Schulman AH. Active Retrotransposons Are a Common Feature of Grass Genomes. Plant Physiology. 2001;125(3):1283-92.