

RESEARCH

Open Access



Retand LTR-retrotransposons in plants: a long way from *pol* to 3′LTR

Carlos M. Viciant^{1*}

Abstract

Background Plant Gypsy LTR-retrotransposons are classified into lineages according to the phylogenetic relationships of the reverse transcriptase. Retand is a lineage of non-chromovirus elements characterized by the presence of a long internal region compared to other lineages.

Results This work focuses on the identification and characterization of Potentially Recently Active Retand Elements (PRAREs) in 617 genomic sequence assemblies of *Viridiplantae* species. The Retand elements were considered PRAREs if their LTRs and insertion sequences were identical, and the sizes of their internal regions and LTRs did not differ by more than 2% from the consensus. A total of 2,735 PRAREs were identified, distributed in 122 clusters corresponding to 34 species, with copy numbers per cluster varying between 1 and 180. They are present in *Eudicotyledons* and *Liliopsida* but not in other groups of plants. Some PRAREs are non-autonomous elements, lacking some of the typical LTR retrotransposon coding domains. The size of the POL-3′LTR regions varies between 2,933 and 6,566 bp, and in all cases, includes potential coding regions oriented antisense to the *gag* and *pol* genes. 97% of the clusters contain antisense ORFs encoding the TRP28 protein domain of unknown function. The analysis of the consensus TRP28 domain indicates that it probably can bind DNA. About half of the PRAREs contain arrays of tandem repeats in the POL-3′LTR region.

Conclusions The large internal region of the Retand elements is due to the presence of a long POL-3′LTR region. This region frequently contains arrays of tandem repeats that contribute to the expansion of this area. The presence of antisense ORFs in the POL-3′LTR region is also a common feature in these elements, many of which encode proteins with conserved domains, especially the TRP28 domain. The possible function of these TRP28-containing proteins is unknown, but their potential DNA binding capacity and the comparison with similar genes in some retroviruses suggest that they may play a regulatory role in the Retand transposition process.

Keywords LTR retrotransposons, Domain annotation, Additional ORF, Antisense, Tandem repeat, Retrovirus

Background

In plant genomes, long terminal repeat (LTR) retrotransposons are the most abundant and widely distributed class of transposable elements (TEs), creating new copies via reverse transcription of the full-length RNA intermediate by the element-encoded reverse transcriptase (RT)

[1]. LTR retrotransposons share many structural features with retroviruses [2], typically consisting of two identical regions at their ends, known as long terminal repeats (LTRs), which range from a few hundred base pairs to several kilobases in length, surrounding an internal region that encodes the essential proteins for their replication and contains a primer binding site (PBS) and a polypurine tract (PPT) at the 5′ and 3′ ends, respectively, both of which are used during the retrotransposition process. The LTR retrotransposons in plants display a target site duplication (TSD) that is usually 5 bp long [3].

*Correspondence:

Carlos M. Viciant
carlos.viciant@cragenomica.es

¹ Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus UAB, Bellaterra, Barcelona 08193, Spain



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

LTR retrotransposons are classified into two superfamilies based on the organization of their coding domains: Ty1/copia (*Pseudoviridae* in the ICTV classification of viruses) and Ty3/gypsy (*Metaviridae*). Phylogenetic analyses of the RT, RH, and INT domains divide each of these superfamilies into different lineages [4]. In plants, the Copia superfamily is divided into Ale, Alesia, Angela, Bianca, Bryco, Lyco, Gymco, Ikeros, Ivana, Osser, SIRE, Tar, and Tork lineages, while the Gypsy superfamily comprises Athila, Clamyvir, CRM, Galadriel, Ogre, Phygy, Reina, Retand, Selgy, Tat, Tcn1, and Tekay. Although the classification is based on phylogenetic analysis, in some cases, the identified lineages are also differentiated by other characteristics shared by most members of a lineage. These features include the presence of an especially long internal region in the cases of the Ogre and Retand lineages [4], the presence of chromodomains, or the ability to encode additional proteins [5].

Typical LTR retrotransposons encode two genes: *gag* and *pol* [6]. The *gag* gene encodes the structural proteins, including capsid (CA) and nucleocapsid (NC), which assemble into virus-like particles (VLPs). The *pol* gene encodes the proteins that provide enzymes involved in reverse transcription and integration into the host genome: aspartic proteinase (AP), RT, RNaseH (RH), and integrase (INT). Both genes can be encoded by the same or two ORFs. Based on the presence or absence of functional gag-pol protein-coding domains, LTR retrotransposons are classified as complete (autonomous) or incomplete (non-autonomous). Non-autonomous LTR retrotransposons must parasitize active ones to propagate [7].

Retroviruses also contain the *gag* and *pol* genes, but they additionally encode the envelope (ENV) protein and, in some cases, other proteins like TAT, REV, NEF, VPR, or VIF. Some of these are encoded in an antisense orientation relative to the *gag-pol-env* genes [8]. The roles of some of these additional proteins are known [9]: HBZ of the Human T-cell Lymphotropic Virus 1 (HTLV-1) exhibits nuclear localization and plays a critical role in the virus lifecycle and the pathogenic process, while the APH-2 from HTLV-2 also shows nuclear localization and interacts with TAX and CREB proteins [10]. Some LTR retrotransposon families also contain ORFs in the internal region encoding additional proteins (aORFs) [5]. The aORF can be found in either sense or antisense orientation with respect to the *gag-pol* genes and upstream or downstream of them [11]. In some cases, the presence of an aORF can be explained by a phenomenon similar to retrovirus gene transduction [12], but in these cases, the aORFs are present in only one or a few copies. In contrast, some families appear to contain aORFs or derivatives in all or most of their copies and in distant species

[13]. Sense aORFs include those encoding ENV-like proteins, so named because they exhibit some structural and functional similarities with retroviral envelope (ENV) proteins [14–16]. Antisense aORFs between the *pol* gene and 3' LTR have also been found in, for example, LTR-retrotransposon families such as maize Grande [13], rice RIRE2 [17], or *Silene Retand* [18]. The roles of the encoded proteins are not known, but in the case of Grande, the encoded protein shows nuclear localization [13].

The Retand lineage of plant LTR-retrotransposons belongs to the non-chromovirus gypsy elements and is abundant in some plant genomes, such as *Silene latifolia* (4,800 copies) [18] and maize (15,904 copies) [19]. No systematic studies have been conducted to determine the taxonomic distribution of the Retand elements. Here, I have taken advantage of the abundance of plant genome assembly sequences to determine the presence of PRA-REs and to analyze their structures, paying special attention to the POL-3'LTR region.

Methods

Genomes used

A total of 617 *Viridiplantae* genome assemblies available in the NCBI database (<https://www.ncbi.nlm.nih.gov>) were selected, considering only one assembly per species and only genomes assembled at least at the chromosome level (Additional file 1).

Retand LTR retrotransposon mining

tBLASTn was used to identify the Retand elements with the default parameters (except $-e$ option set to $1e-10$), using the amino acid sequences of the RT central conserved domain of four characterized Retand elements as queries: Tat4-1 from *Arabidopsis* (AAD20101.1; 159 aa), Gret1 from vineyard (AB242301.1; 159 aa), Cinfu1-1 from maize (AF049110.1; 159 aa), and Grande1-4 from teosinte (X97604.1, 160 aa) (Additional file 2). Only hits with complete and uninterrupted RT domains (ranging from 157 to 162 aa) without stop codons or frameshifts were selected. Only RT domains encoded in chromosome assemblies, not in scaffolds, were chosen. For each genome assembly, the selected set of RT sequences was aligned using MAFFT (<https://mafft.cbrc.jp/alignment/server/>) with 62 model RT sequences (Additional file 2), including the four Retand sequences used as queries, 56 RT sequences from other lineages, and two RT sequences from *Caulimoviridae*. A phylogenetic tree was constructed using the NJ method and 100 bootstrap repetitions (<https://mafft.cbrc.jp/alignment/server/>). Only the RT sequences that clustered with the Retand RT sequences were used for further analyses.

To identify the complete LTR-retrotransposon sequences, the 10 kb up- and downstream sequences of each RT domain were selected and compared using BLAST2seq (<http://kinase.com/blast/wblast2.cgi?0>). Only elements with identical LTRs and identical insertion repeats were retained. The LTR sequences identified in each genome were grouped using CD-HIT with a sequence identity cut-off of 90% (https://mafft.cbrc.jp/alignment/server/spool/_ho.250120195231950bzjTZQOq87tUISl5scuRklsfnormal.html). Finally, the elements with significant insertions or deletions were removed (their LTR or internal sequences differed by more than 2% from the consensus size of the cluster).

Sequence analyses

The ORFs were predicted using ORF Finder (<https://www.ncbi.nlm.nih.gov/orffinder/>). Protein domains were identified using MOTIF (<https://www.genome.jp/tools/motif/>). The presence of tandem repeats was determined using Tandem Repeats Finder (<https://tandem.bu.edu/trf/trf.submit.options.html>), and only tandem arrays containing at least three copies were considered. Sequence alignments were performed using the G-INS-I method of MAFFT (<https://mafft.cbrc.jp/alignment/server/>). Amino acid conservation was determined using Weblogo (<https://weblogo.berkeley.edu/logo.cgi>). Protein DNA-binding predictions were performed using hybridDBRpred (hybrid model for DNA-binding residue prediction) (<http://biomine.cs.vcu.edu/servers/hybridDBRpred/>).

Results

Species distribution

Once inserted into the genome, retrotransposons begin to suffer mutations, causing their sequences and structures to diverge from their original active form. Therefore, it is easier to determine LTR-retrotransposon structures in recent elements, meaning that they have identical LTRs and insertion repeats and do not contain significant insertions or deletions. I named them PRAREs (Potentially Recently Active Retand Elements). The presence of PRAREs was examined in 617 genome assemblies of *Viridiplantae* species (Additional file 1), and a total of 2,735 PRARE copies were identified, distributed across 122 clusters corresponding to 34 species, with copy numbers per cluster varying between 1 and 180 (Additional file 3). PRAREs were found in *Eudicotyledons* and *Liliopsida*, but not in other groups of plants (Fig. 1). Among the *Liliopsida*, PRAREs are especially abundant in *Poales*, although they are also found in *Liliales*. Among *Eudicotyledons*, the distribution is much more dispersed. They are abundant in *Gentianidae* (*Apiales*, *Asterales*, and *Dipsacales*), *Ericales*, and *Cornales*, and are also present in

Myrtales, *Lamiales*, *Caryophyllales*, *Brassicales*, *Sapindales*, *Rosales*, and *Fabales*.

Phylogenetic analysis

A consensus amino acid sequence of the RT domain was determined for each cluster containing at least ten copies. The RT sequences were aligned, and a ML phylogenetic tree was constructed (Fig. 2). Sequences are distributed among six clades, three containing sequences only from *Liliopsida*, two only from *Eudicotyledons*, and one from both. Clusters of the same species tend to group together in one or more groups corresponding to the same or different clades. For example, the four sunflower sequences are distributed in two groups, with two sequences each in different clades (Eudicot 1 and 2), whereas the ten *Alopecurus myosuroides* clusters are distributed in three groups, all within the same clade (*Liliopsida* 1).

The *Liliopsida*/Eudicot clade includes two sequences from *Rhynchospora pubera* (*Liliopsida*) and *Linaria vulgaris* (*Eudicotyledon*). The *Liliopsida*/Eudicot clade is closely related to the Eudicot-2 clade. Phylogenetic trees using other domains (GAG-AP, RH-INT, and TRP28) also show that *Rhynchospora pubera* 3 groups with *Eudicotyledons* and not with *Liliopsida* (Additional File 4). The location of *Rhynchospora pubera* 3 suggests that it arose from a horizontal transfer from an Eudicotyledonous plant.

Structural variability and encoded conserved protein domains

The structural characteristics of the elements in the clusters were analyzed, including the sizes of the LTRs and internal regions, the presence of ORFs, and the existence of arrays of tandem repeats (Table 1). The sizes of the LTRs vary between 315 and 1,194 bp, with an average of 635 bp, while the sizes of the internal regions range from 7,380 to 12,869 bp, with an average of 10,364 bp.

For each element, the presence in the internal region of conserved ORFs (≥ 600 bp) and the presence of arrays of tandem repeats (≥ 3 repeats) were examined (Additional file 5). A representative element of each cluster was selected, meaning that it possesses the same structural characteristics as most of the elements in the cluster (Additional File 5). DNA sequences of all the elements are available in Additional file 6. All the representative elements contain at least two ORFs: at least one sense ORF (sense with respect to the RT coding region) and at least one antisense ORF (Fig. 3). The *gag* and *pol* genes are encoded by a unique ORF in 37 of the clusters (59%) and by two in the remainder, with 29% considering only *Liliaceae* compared to 83% in *Eudicotyledons*.



Fig. 1 Taxonomic distribution of PRAREs. A black dot indicates an analyzed genome in which no element was detected, while a red circle represents a genome where PRAREs were identified. The diameter of the red circle correlates with the copy number. Orders in red are those in which at least one PRARE is present. The blue box indicates *Eudicotyledons*, and the yellow box represents *Liliaceae*. The phylogenetic tree is based on The Plant Tree of Life [20]

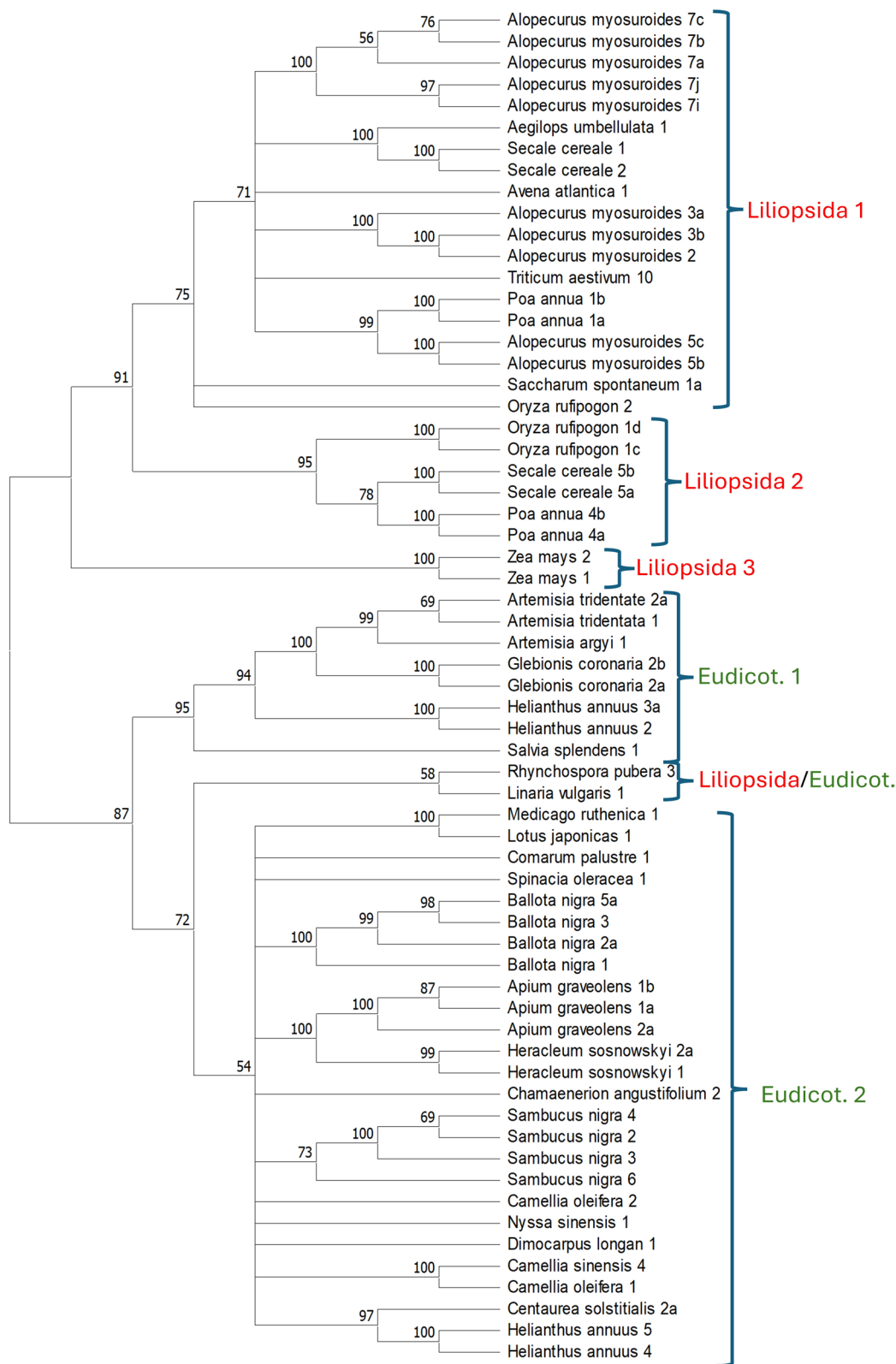


Fig. 2 Phylogeny of PRAREs based on the RT domain. Midpoint-rooted ML phylogenetic tree of the consensus sequences of the RT domain of clusters containing at least 10 copies. Bootstrap values based on 1,000 replicates are shown

Table 1 Structural characteristics of the selected Retand clusters

Cluster	Copy num.	LTR	Internal region	POL-3'LTR	GAG	AP	RH	INT	TRP28	Other domains	Tandem arrays
Liliopsida 1											
<i>Alopecurus myosuroides</i> 7c	26	477	9,671	5,427	NO	YES	YES	YES	YES	NO	0
<i>Alopecurus myosuroides</i> 7b	23	481	9,685	5,433	NO	YES	YES	YES	YES	NO	0
<i>Alopecurus myosuroides</i> 7a	13	479	9,680	5,428	NO	YES	YES	YES	YES	NO	0
<i>Alopecurus myosuroides</i> 7j	180	454	9,480	5,305	NO	YES	YES	YES	YES	NO	0
<i>Alopecurus myosuroides</i> 7i	52	454	9,581	5,279	NO	YES	YES	YES	YES	NO	0
<i>Aegilops umbellulata</i> 1	33	504	12,140	5,663	YES	YES	YES	YES	YES	NO	1
<i>Secale cereale</i> 1	10	506	12,869	6,091	YES	YES	YES	YES	YES	NO	1
<i>Secale cereale</i> 2	26	507	12,726	6,153	YES	YES	YES	YES	YES	NO	1
<i>Avena atlantica</i> 1	37	491	11,237	4,743	YES	YES	YES	YES	YES	DNA topo2	1
<i>Alopecurus myosuroides</i> 3a	46	576	10,798	3,906	YES	NO	YES	YES	YES	NO	0
<i>Alopecurus myosuroides</i> 3b	43	577	8,734	4,079	YES	YES	NO	NO	YES	NO	0
<i>Alopecurus myosuroides</i> 2	55	564	8,734	4,102	YES	YES	NO	NO	YES	NO	0
<i>Triticum aestivum</i> 10	10	528	11,996	5,828	YES	YES	YES	YES	YES	NO	1
<i>Poa annua</i> 1b	10	424	12,250	5,501	YES	YES	YES	YES	YES	NO	2
<i>Poa annua</i> 1a	15	439	12,029	5,301	YES	YES	YES	YES	YES	NO	2
<i>Alopecurus myosuroides</i> 5c	133	491	11,821	5,491	YES	YES	YES	YES	YES	NO	0
<i>Alopecurus myosuroides</i> 5b	15	493	11,834	5,474	YES	NO	YES	YES	YES	NO	0
<i>Saccharum spontaneum</i> 1a	10	1,065	12,304	5,854	YES	YES	YES	YES	YES	polIII	0
<i>Oryza rufipogon</i> 2	18	440	10,428	4,282	YES	YES	YES	YES	YES	NO	0
Liliopsida 2											
<i>Oryza rufipogon</i> 1d	35	1,077	10,727	4,484	YES	YES	YES2	YES	YES	Smc	0
<i>Oryza rufipogon</i> 1c	61	1,174	10,703	4,483	YES	YES	YES2	YES	YES	Smc	0
<i>Secale cereale</i> 5b	21	947	11,752	5,367	YES	YES	YES	YES	YES	Smc/polIII	0
<i>Secale cereale</i> 5a	65	954	11,775	5,359	YES	YES	YES	YES	YES	Smc/polIII	2
<i>Poa annua</i> 4b	21	1,031	10,962	4,981	YES	YES	YES	YES	YES	Smc/polIII	0
<i>Poa annua</i> 4a	24	1,079	10,861	4,914	YES	YES	YES	YES	YES	NO	0
Liliopsida 3											
<i>Zea mays</i> 2	12	585	8,245	4,200	NO	YES	YES	YES	YES	NO	0
<i>Zea mays</i> 1	21	587	8,365	4,313	NO	NO	YES	YES	YES	NO	1
Eudicot 1											
<i>Artemisia tridentata</i> 2a	63	488	9,588	3,868	YES	YES	YES	YES	YES	NO	0
<i>Artemisia tridentata</i> 1	42	488	9,581	3,899	YES	YES	YES	YES	YES	NO	0
<i>Artemisia argyi</i> 1	81	496	9,362	3,665	YES	YES	YES	YES	YES	NO	0
<i>Glebionis coronaria</i> 2b	13	571	9,588	4,035	YES	NO	YES	YES	YES	NO	0
<i>Glebionis coronaria</i> 2a	41	581	9,843	4,108	YES	NO	YES	YES	YES	NO	0
<i>Helianthus annuus</i> 3a	23	880	10,274	4,549	YES	YES	YES	YES	YES	NO	1
<i>Helianthus annuus</i> 2	15	806	11,456	5,899	YES	YES	YES	YES	YES	ToIA	0
<i>Salvia splendens</i> 1	23	597	9,719	4,234	YES	YES	YES	YES	YES	NO	1
Eudicot 2											
<i>Medicago ruthenica</i> 1	19	315	7,380	3,888	NO	NO	YES	YES	YES	NO	0
<i>Lotus japonicus</i> 1	47	378	8,780	3,281	YES	YES	YES	YES	YES	NO	0
<i>Comarum palustre</i> 1	63	483	9,994	4,340	YES	YES	YES	YES	YES	NO	3
<i>Spinacia oleracea</i> 1	54	425	10,719	4,850	YES	YES	YES	YES	YES	NO	1
<i>Ballota nigra</i> 5a	28	775	10,294	4,118	YES	YES	YES	YES	YES	NO	0
<i>Ballota nigra</i> 3	28	685	10,371	4,252	YES	YES	YES	YES	YES	NO	0
<i>Ballota nigra</i> 2a	19	737	10,519	4,282	YES	YES	YES	YES	YES	Smc	0
<i>Ballota nigra</i> 1	12	696	9,557	3,609	YES	YES	YES	YES	YES	Recom.Inh.	1
<i>Apium graveolens</i> 1b	24	569	9,357	3,801	YES	YES	YES	YES	YES	NO	2

Table 1 (continued)

Cluster	Copy num.	LTR	Internal region	POL-3'LTR	GAG	AP	RH	INT	TRP28	Other domains	Tandem arrays
<i>Apium graveolens</i> 1a	41	548	9,753	4,121	YES	YES	YES	YES	YES	NO	2
<i>Apium graveolens</i> 2a	10	570	9,626	3,997	YES	YES	YES	YES	YES	NO	3
<i>Heracleum sosnowskyi</i> 2a	16	504	10,937	3,492	YES	YES	YES	YES	YES	NO	1
<i>Heracleum sosnowskyi</i> 1	38	504	9198	3,459	YES	YES	YES	YES	YES	NO	2
<i>Chamaenerion angustifolium</i> 2	171	664	10,051	4,522	YES	YES	YES	YES	YES	NO	1
<i>Sambucus nigra</i> 4	76	586	10,964	4,895	YES	YES	YES	YES	YES ^a	NO	1
<i>Sambucus nigra</i> 2	29	586	10,938	4,954	YES	YES	YES	YES	YES ^a	NO	1
<i>Sambucus nigra</i> 3	68	586	10,943	4,953	YES	YES	YES	YES	YES ^a	NO	1
<i>Sambucus nigra</i> 6	39	857	10,861	4,771	YES	YES	YES	YES	YES	NO	1
<i>Camellia oleifera</i> 2	13	551	10,377	4,812	YES	YES	YES	YES	YES	Smc	2
<i>Nyssa sinensis</i> 1	15	627	10,468	4,582	YES	YES	YES	YES	YES	NO	2
<i>Dimocarpus longan</i> 1	10	713	9,540	3,682	YES	YES	YES	YES	YES	Smc	0
<i>Camellia sinensis</i> 4	18	824	10,483	4,472	YES	YES	YES	YES	YES ^a	NO	1
<i>Camellia oleifera</i> 1	23	584	9,983	4,290	YES	NO	YES	YES	YES ^a	NO	1
<i>Centaurea solstitialis</i> 2a	58	900	8,485	2,933	YES	YES	YES	YES	YES	NO	1
<i>Helianthus annuus</i> 5	121	719	9,313	3,343	YES	YES	YES	YES	NO	NO	2
<i>Helianthus annuus</i> 4	71	720	9,310	3,327	YES	YES	YES	YES	NO	NO	2
Liliopsida/Eudicot											
<i>Rhynchospora pubera</i> 3	12	432	11,818	5,552	YES	YES	YES	YES	YES	NO	3
<i>Linaria vulgaris</i> 1	22	1,194	12,114	6,556	YES	YES	YES	YES	YES	NO	0

The grouping of the clusters is based on Fig 2. The sizes in base pairs (bp) correspond to the average values of the elements in each cluster. TRP28 YES^a indicates that the domain is present, but in an ORF shorter than 600 nt; DNA topo 2 = DNA topoisomerase 2 (PLN03237, PLN03237); Recomb.Inh. = recombination and DNA strand exchange inhibitor protein (PRK00409, PRK00409); Smc = COG1106, smc, chromosome segregation ATPase (COG1196); TolA = cell envelope integrity inner membrane protein TolA (PRK09510); PolIII = DNA polymerase III subunits gamma and tau (PRK07764, PRK07764). Tandem arrays refer to the number of tandem arrays detected containing three or more repeats

Ten of the 63 clusters are defective in at least one of the typical retrotransposon protein domains. Seven clusters do not encode the GAG domain, two do not encode both the GAG and the AP domains, and one cluster does not encode the RH and INT domains. Most of them (9) correspond to *Poales* (Table 1).

The size of the region between the *pol* gene and the 3'LTR varies between 2,933 and 6,566 bp, with an average of 4,616 bp. All clusters contain at least one ORF longer than 600 bp in this region, most of which are antisense with respect to the *gag* and *pol* genes (Fig. 3). None of the sense ORFs in this region encode known protein

domains or putative transmembrane regions, as would be expected in the case of ENV proteins.

Of the 104 antisense ORFs detected, 72 encode known protein domains. 57 of the antisense ORFs encode a protein with the Transposase 28 domain (TRP28, PF04195). When also considering ORFs shorter than 600 bp, all clusters except two encode a protein with a TRP28 domain, all in the antisense direction relative to the *gag* and *pol* genes. The only two exceptions are *Helianthus annuus_4* and 5. The role of the TRP28 domain is unknown, and it appears to be present only in proteins of flowering plants (InterPro database).

(See figure on next page.)

Fig. 3 Structural organization of PRAREs. A schematic representation shows the structure and major features of the representative sequences of PRARE clusters that contain at least 10 elements. The representative sequence was chosen as one of the sequences in the cluster that exhibited the same structural characteristics as most of the elements within it (see Additional file 5 for more details). Clusters are grouped according to the phylogenetic analysis in Fig. 2. The tree on the left is based on Fig. 2. The shown structures include: LTRs (empty boxes), ORFs longer than 600 bp (black arrows), ORFs shorter than 600 bp but encoding a TRP28 domain (grey arrows), regions encoding known protein domains (colored boxes), and arrays of tandem repeats (boxes with vertical lines). Protein domains: Pink, GAG; Yellow, Aspartic proteinase; Red, Reverse transcriptase; Violet, RNaseH; Blue, Integrase; Green, TRP28, transposase 28 (PF04195); Orange, DNA topoisomerase 2 (PLN03237); Dark blue, Mis12 protein (PF05859); Dark grey, Mitotic checkpoint protein (PF05557); Light blue, Smc chromosome segregation ATPases (OG1196); DNA topoisomerase 2 (PLN03237, PLN03237); Light grey, DNA polymerase III subunits gamma and tau (PRK07764, PRK07764). **A** Liliopsida. **B** Eudicotyledons and Eudicotyledons/Liliopsida

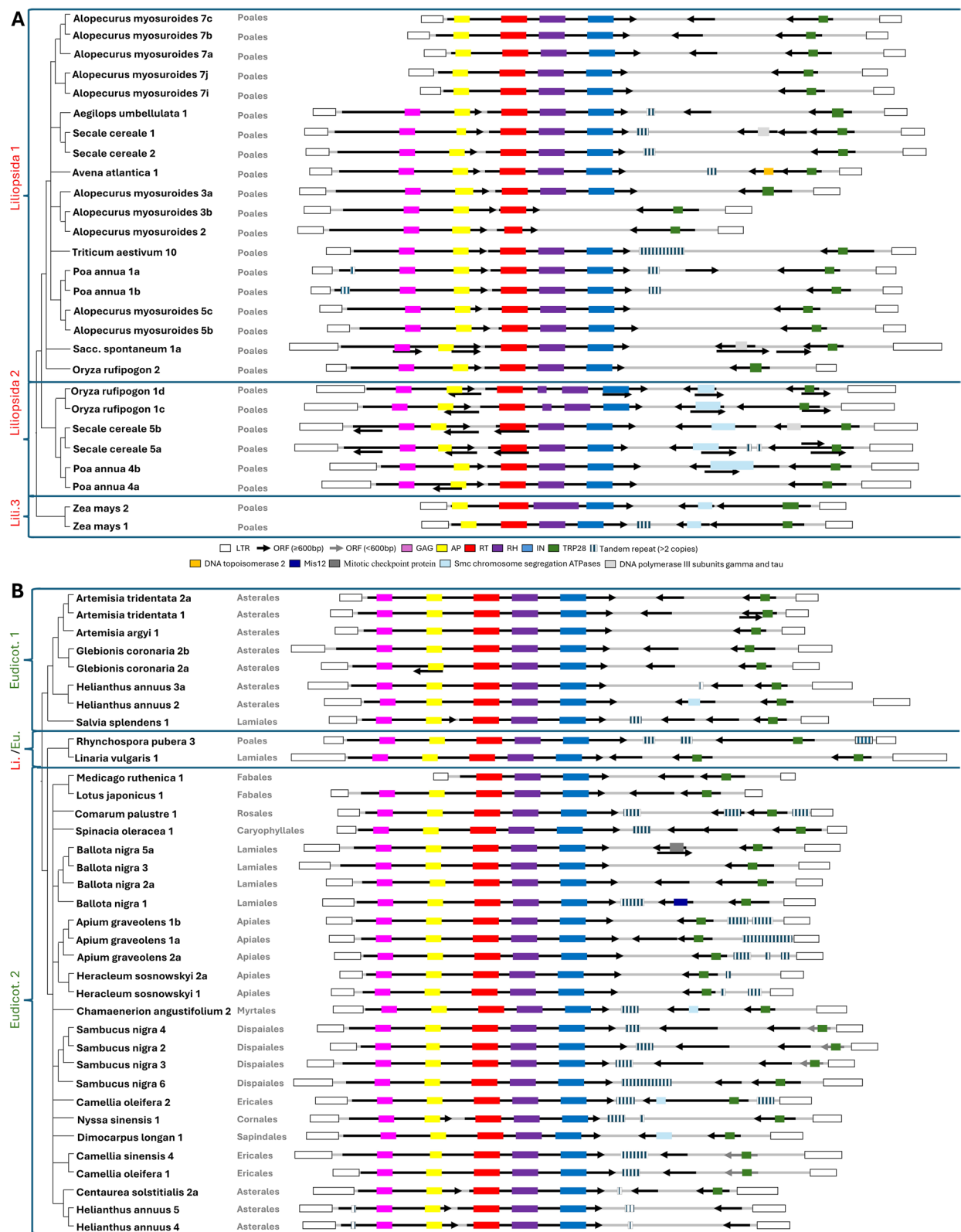


Fig. 3 (See legend on previous page.)

The cluster consensus sequences of the TRP28 domain were aligned (Fig. 4A) and showed several well-conserved positions (G21, P25, L34, P46, N47, and F55). A prediction program indicates that this region has DNA-binding capacity. The regions with a higher probability of binding DNA coincide with some of the most conserved amino acids (Fig. 4B). This information suggests that the function of TRP28 is related to DNA binding.

Five additional known protein domains are encoded by antisense ORFs located in the POL-3'LTR region (Fig. 3): 11 Smc domains, 2 mitotic checkpoint proteins, 2 DNA polymerase III subunits gamma and tau, one DNA topoisomerase 2, and one Mis12.

The analysis of the internal regions of the PRAREs also revealed the presence of arrays of tandemly repeated sequences in 32 of them (Fig. 3). The localization of the tandem arrays varies, but the most frequent is in the POL-3'LTR region, adjacent to the *pol* gene or next to the 3'LTR. They are highly variable in sequence and copy number. The size of the repeated elements ranges from 8 to 100 bp, and the maximum number of repetitions is 85. The number of repetitions varies among copies of the same cluster.

Discussion

A systematic survey of recently inserted Retand elements in plant genomes revealed their presence in *Eudicotyledons* and *Liliaceae* but absence in other plant groups. This data must be interpreted with caution since the number of genomes available for some taxa is very low. In some cases, the results differ among the genomes of closely related species. For example, of the ten species of the genus *Oryza* analyzed, only one (*Oryza rufipogon*) contains PRIMEs. Burst-like amplification appears to be a common process in plants and has been described in different species, including *Arabidopsis thaliana* [21] and *Oryza sativa* [22]. The phylogenetic analysis shows a potential case of horizontal transfer of a Retand element. HTs of transposable elements between plants are known [23]. Although the phylogenetic distribution can be explained by alternative mechanisms to HTs [24], I believe that, in this case, they can be excluded because it involves taxonomically very divergent plant species (*Eudicotyledons* and *Liliaceae*).

The structural analysis of the PRIMEs showed that elements lacking some of the typical LTR-retrotransposon protein domains have been recently active. The amplification of defective LTR-retrotransposons has been described, and they are presumed to constitute non-autonomous elements whose transposition is dependent on complementation by other elements of the same or another related family [7, 25]. Transposition of non-autonomous endogenous retroviruses, dependent on a full-length copy, has also been described in endogenous retroviruses [26].

One of the characteristics of the Retand elements that has been confirmed here is the presence of a large internal region compared to other Gypsy lineages. As observed here, this large size is due to the region between the *pol* gene and the 3'LTR. This region contains ORFs, mainly antisense, some of which encode known protein domains, and frequently arrays of tandem repeats. Many active Retand elements carry tandem repeats in their 3' internal region between the gag-pol and the 3' LTR. The length of the repeat arrays varies, and the repeated sequences are family-specific. Similar structures have been observed in retrotransposons [27–30] and in retroviruses [31]. Their origin and eventual function remain unknown [32].

Plant retrotransposons capable of encoding additional proteins in antisense have been described [5, 11]. Here, I show that all PRIMEs contain antisense ORFs in the POL-3'LTR region, and that 97% of them include antisense ORFs encoding proteins with a TRP28 domain. Therefore, this structural element must be considered a characteristic of the Retand elements. Two fundamental questions arise: origin and function.

The origin of the POL-3'LTR region must be quite ancient since it is found in the major taxa of flowering plants. One possible origin is the transduction of cellular sequences. Transduction has been well documented for retroviruses [33] and has also been described for LTR-retrotransposons like maize Bs1 [34], various LTR retrotransposons in rice [35], sorghum [36], and in animals [12]. However, no sequence similarities have been found with TRP28 outside TEs, and the retrovirus-transduced genes are usually in the same sense as the retrotransposon *gag-pol* genes. Another possible origin is the insertion of another TE. Nested insertions of TEs are frequent

(See figure on next page.)

Fig. 4 Sequence conservation of the TRP28 domain encoded by PRAREs. **A** Multiple sequence alignment of the consensus sequences of the TRP28 domains. The alignment is colored according to: dark red > 45%; red > 35%; pink > 25%. **B** Sequence logo of the TRP28 domain encoded by PRAREs. The logo is based on the aligned amino acid sequences. Amino acids are represented in a single-letter code, and the height of each letter depicts the frequency of the corresponding residues at that position, with the most frequent at the top. The red boxes represent the predicted DNA-binding residues according to hybridDBRpred

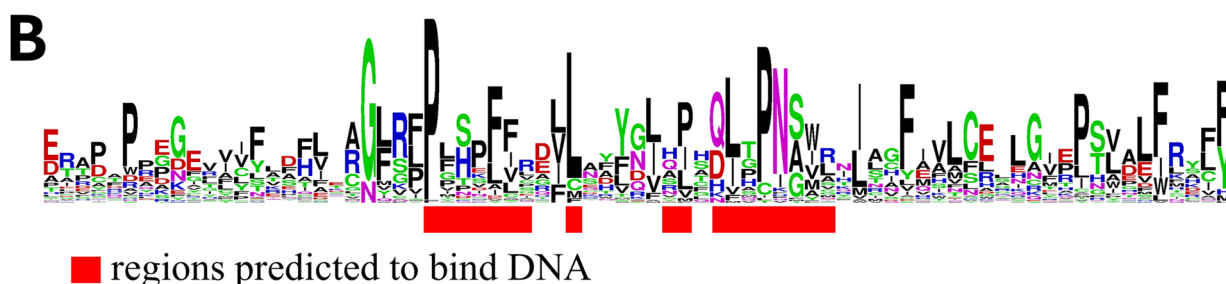


Fig. 4 (See legend on previous page.)

in plant genomes [37]; however, the lack of sequence similarities with other known TEs does not support this hypothesis. Another possible explanation is that they originate from recombination with the genetic material of an unknown virus, which would explain the lack of similarities with available sequences.

Regardless of its origin, the fact that it is present in almost all recently active Retand elements suggests that it must serve some function. Some retroviral genomes encode antisense genes (HTLV-1,2,3 and STLV-2, 3, 4) [9, 38]. Some of these proteins are capable of binding nucleic acids and are involved in, for example, the regulation of transcription (TAT) or the transport of viral RNAs (REV). The TRP28 domain contains regions predicted to bind DNA, and the other protein domains encoded by antisense ORFs have functional categories involved in interactions with DNA. All this data suggest that TRP28 and other additional proteins encoded in the Retand elements may be involved in transcriptional regulatory roles. Further analyses must confirm this hypothesis.

Conclusions

The Retand lineage of plant LTR-retrotransposons is present in *Eudicotyledons* and *Liliopsida*, but not in other groups of plants. They are especially abundant in *Poales* and *Gentianidae*. The horizontal transmission of Retand elements between species appears to be possible. Some of these elements are non-autonomous. Retand elements contain a large region between the *pol* gene and the 3'LTR. This region usually includes arrays of short tandem repeats and antisense ORFs that, in many cases, encode proteins with DNA-binding capabilities. Comparison with retroviruses suggests that they may play regulatory roles.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13100-025-00354-z>.

Additional file 1. Species and genome assembly accession numbers used. Taxonomic information is retrieved from The Plant Tree of Life.

Additional file 2. Model RT sequences.

Additional file 3. Recently inserted Retand elements found.

Additional file 4. Three phylogenies of PRAREs based on the GAG-AP, RH-INT, and TRP28 domains. Midpoint-rooted ML phylogenetic tree of the consensus sequences of the indicated domains of clusters containing at least 10 copies. Bootstrap values are based on 1,000 replicates.

Additional file 5. Schematic representation of the ORFs and tandem repeats found in the PRARE elements. At the top, the consensus structure of the cluster. The meanings of the boxes and colors are as in Fig. 3. The number in red corresponds to the element chosen as representative of the cluster. Only clusters with ten or more elements are included.

Additional file 6. DNA sequences of the PRARE elements identified in this study. LTRs are in red. Flanking sequences are also included.

Acknowledgements

I thank Dr. J.M. Casacuberta for comments on this manuscript.

Authors' contributions

CMV did all the work.

Funding

Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. I acknowledge the financial support from the Spanish Ministerio de Ciencia e Innovación (Grant PID2022-143167NB-I00), Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0533) and CERCA Program / Generalitat de Catalunya (CEX2019-000902-S).

Data availability

Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 8 November 2024 Accepted: 17 March 2025

Published online: 02 April 2025

References

- Pulido M, Casacuberta JM. Transposable element evolution in plant genome ecosystems. *Curr Opin Plant Biol*. 2023;75: 102418.
- Eickbush TH, Jamburuthugoda VK. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res*. 2008;134(1–2):221–34.
- Jedlicka P, Lexa M, Vanat I, Hobza R, Kejnovsky E. Nested plant LTR retrotransposons target specific regions of other elements, while all LTR retrotransposons often target palindromes and nucleosome-occupied regions: in silico study. *Mob DNA*. 2019;10:50.
- Neumann P, Novák P, Hošťáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their poly-protein domains and provides a reference for element classification. *Mob DNA*. 2019;10: 1.
- Viciant CM, Casacuberta JM. Additional ORFs in plant LTR-retrotransposons. *Front Plant Sci*. 2020;11:555.
- 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*. 2019;116(20):10048–57.
- Tanskanen JA, Sabot F, Viciant C, Schulman AH. Life without GAG: the BARE-2 retrotransposon as a parasite's parasite. *Gene*. 2007;390(1–2):166–74.
- Anderson JL, Hope TJ. HIV accessory proteins and surviving the host cell. *Curr HIV/AIDS Rep*. 2004;1(1):47–53.
- Romerio F. Origin and functional role of antisense transcription in endogenous and exogenous retroviruses. *Retrovirology*. 2023;20(1):6.
- Barbeau B, Mesnard JM. Making sense out of antisense transcription in human T-cell lymphotropic viruses (HTLVs). *Viruses*. 2011;3(5):456–68.
- Steinbauerová V, Neumann P, Novák P, Macas J. A widespread occurrence of extra open reading frames in plant Ty3/gypsy retrotransposons. *Genetica*. 2011;139(11–12):1543–55.
- Alvarado-Marchena L, Martínez-Pérez M, Aparicio F, Pallas V, Maumus F. Recent acquisition of functional m6A RNA demethylase domain in orchid Ty3/Gypsy elements. *Front Plant Sci*. 2022;13: 939843.
- Gómez-Orte E, Viciant CM, Martínez-Izquierdo JA. Grande retrotransposons contain an accessory gene in the unusually long 3'-internal region

- that encodes a nuclear protein transcribed from its own promoter. *Plant Mol Biol*. 2013;81(6):541–51.
14. Vicient CM, Kalendar R, Schulman AH. Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Res*. 2001;11(12):2041–9.
 15. Miguel C, Simões M, Oliveira MM, Rocheta M. Envelope-like retrotransposons in the plant kingdom: evidence of their presence in gymnosperms (*Pinus pinaster*). *J Mol Evol*. 2008;67(5):517–25.
 16. Carvalho M, Ribeiro T, Viegas W, Morais-Cecilio L, Rocheta M. Presence of env-like sequences in *Quercus suber* retrotransposons. *J Appl Genet*. 2010;51(4):461–7.
 17. Ohtsubo H, Kume-kawa N, Ohtsubo E. RIRE2, a novel gypsy-type retrotransposon from rice. *Genes Genet Syst*. 1999;74(3):83–91.
 18. Kejnovsky E, Kubat Z, Macas J, Hobza R, Mracek J, Vyskot B. Retand: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat. *Mol Genet Genom*. 2006;276(3):254–63.
 19. Novák P, Hošťáková N, Neumann P, Macas J. DANTE and DANTE_LTR: lineage-centric annotation pipelines for long terminal repeat retrotransposons in plant genomes. *NAR Genom Bioinform*. 2024;6(3):lqae113.
 20. Soltis DE, Moore MJ, Sessa EB, Smith SA, Soltis PS. Using and navigating the plant tree of life. *Am J Botany*. 2018;05(3):287–90.
 21. Tsukahara S, Kobayashi A, Kawabe A, et al. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature*. 2009;461(7262):423–6.
 22. Vitte C, Panaud O, Quesneville H. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics*. 2008;8:218.
 23. Park M, Christin PA, Bennetzen JL. Sample sequence analysis uncovers recurrent horizontal transfers of transposable elements among grasses. *Mol Biol Evol*. 2021;38(9):3664–75.
 24. Orozco-Arias S, Dupeyron M, Gutiérrez-Duque D, Tabares-Soto R, Guyot R. High nucleotide similarity of three *Copia* lineage LTR retrotransposons among plant genomes. *Genome*. 2023;66:51–61.
 25. Kirov I, Merkulov P, Dudnikov M, et al. Transposons hidden in *Arabidopsis thaliana* genome assembly gaps and mobilization of non-autonomous LTR retrotransposons unravelled by nanotei pipeline. *Plants (Basel)*. 2021;10(12):2681.
 26. Rebollo R, Galvão-Ferrari M, Gagnier L, Zhang Y, Ferraj A, Beck CR, Lorincz MC, Mager DL. Inter-strain epigenomic profiling reveals a candidate IAP master copy in C3H mice. *Viruses*. 2020;12(7): 783.
 27. Monfort A, Vicient CM, Raz R, Puigdomènech P, Martínez-Izquierdo JA. Molecular analysis of a putative transposable retroelement from the *Zea* genus with internal clusters of tandem repeats. *DNA Res*. 1995;2(6):255–61.
 28. Vondrak T, Ávila Robledillo L, Novák P, Koblížková A, Neumann P, Macas J. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *The Plant J*. 2020;101(2):484–500.
 29. Adabo Cintra L, Boldieri de Souza T, Parteka LM, et al. An 82 bp tandem repeat family typical of 3' non-coding end of Gypsy/TAT LTR retrotransposons is conserved in *Coffea* spp. Pericentromeres. *Genome*. 2022;65(3):137–51.
 30. Zattera ML, Pacheco BD. Transposable elements as a source of novel repetitive DNA in the eukaryote genome. *Cells*. 2022;11(21):3373.
 31. Shimode S, Nakagawa S, Yoshikawa R, Shojima T, Miyazawa T. Heterogeneity of koala retrovirus isolates. *FEBS Lett*. 2014;588(1):41–6.
 32. Macas J, Koblížková A, Navrátilová A, Neumann P. Hypervariable 3' UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. *Gene*. 2009;448(2):198–206.
 33. Rosenberg N, Jolicoeur P. Retroviral pathogenesis. In: Coffin JM, Hughes SH, Varmus HE, editors. *Retroviruses*. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1997.
 34. Elrouby N, Bureau TE. Bs1, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. *Plant Physiol*. 2010;53(3):1413–24.
 35. Wang W, Zheng H, Fan C, et al. High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*. 2006;18(8):1791–802.
 36. Jiang SY, Ramachandran S. Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. *PLoS One*. 2013;8(7): e71118.
 37. Bennetzen JL, SanMiguel P, Chen M, Tikhonov A, Francki M, Avramova Z. Grass genomes. *Proc Natl Acad Sci*. 1998;95(5):1975–8.
 38. Sauter D, Kirchhoff F. Multilayered and versatile inhibition of cellular antiviral factors by HIV and SIV accessory proteins. *Cytokine Growth Factor Rev*. 2018;40:3–12.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.