Chromosome Res (2021) 29:261–284
https://doi.org/10.1007/s10577-021-09663-4

ORIGINAL ARTICLE

Full-length LTR retroelements in *Capsicum annuum* revealed a few species-specific family bursts with insertional preferences

Anahí Mara Yañez-Santos · Rosalía Cristina Paz · Paula Beatriz Paz-Sepúlveda · Juan Domingo Urdampilleta

Received: 4 November 2020 / Revised: 4 May 2021 / Accepted: 6 May 2021 / Published online: 4 June 2021

Abstract *Capsicum annuum* is a species that has undergone an expansion of the size of its genome caused mainly by the amplification of repetitive DNA sequences, including mobile genetic elements. Based on information obtained from sequencing the genome of pepper, the estimated fraction of retroelements is approximately 81%, and previous results revealed an important contribution of lineages derived from Gypsy superfamily. However, the dynamics of the retroelements in the *C. annuum* genome is poorly understood. In this way, the present work seeks to investigate the phylogenetic diversity and genomic abundance of the families of autonomous (complete and intact) LTR retroelements from *C. annuum* and inspect their distribution along its chromosomes. In total, we identified 1151 structurally full-length retroelements (340 Copia; 811 Gypsy) grouped in 124 phylogenetic families in the base of their retrotranscriptase. All the evolutive lineages of LTR retroelements identified in plants were present in pepper; however, three of them comprise 83% of the entire LTR retroelements population, the lineages Athila, Del/Tekay, and Ale/Retrofit. From them, only three families represent 70.8% of the total number of the identified retroelements. A massive family-specific wave of amplification of two of them occurred in the last 0.5 Mya (GypsyCa_16; CopiaCa_01), whereas the third is more ancient and occurred 3.0 Mya (GypsyCa_13). Fluorescent in situ hybridization performed with family and lineage-specific probes revealed contrasting patterns of chromosomal affinity. Our results provide a database of the populations LTR retroelements specific to *C. annuum* genome. The most abundant families were analyzed according to chromosome insertional preferences, suppling useful tools to the design of retroelement-based markers specific to the species.

Keywords LTR retroelement family · Capsicum annuum · FISH · Family radiation · Recent insertion · Idiogram

Anahí Mara Yañez Santos and Rosalía Cristina Paz contributed equally to this research.

Responsible Editor: Jiming Jiang

A. M. Yañez-Santos · R. C. Paz (✉)
CIGEOBIO (FCEFyN, UNSJ/CONICET), Av. Ignacio de la Roza 590 (Oeste), J5402DCS, Rivadavia, San Juan, Argentina
e-mail: rosaliapaz@gmail.com

A. M. Yañez-Santos
e-mail: anahimyanez@gmail.com

A. M. Yañez-Santos · J. D. Urdampilleta
Instituto Multidisciplinario de Biología Vegetal (IMBIV), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)–Universidad Nacional de Córdoba (UNC), Córdoba, Argentina

J. D. Urdampilleta
e-mail: jurdampilleta@imbiv.unc.edu.ar

P. B. Paz-Sepúlveda
Instituto Multidisciplinario de Biología Celular (IMBICE), Consejo Nacional de Investigaciones Científicas y Técnicas de la República Argentina (CONICET) – Comisión de Investigaciones Científicas (CIC) – Universidad Nacional de La Plata (UNLP), La Plata, Argentina
e-mail: paulabeatrizpaz@gmail.com

Introduction

LTR retroelements are the most abundant transposable elements in plant genomes and constitute an important portion of the dispersed repetitive DNA (Bennetzen 2000; Schnable et al. 2009; Jiang and Ramachandran 2013). These elements are sequences of DNA between 4000 and 30,000 bp that, in their complete and intact state, have all the necessary machinery for their replication and amplification in the host genome (Grandbastien 1998; Hirochika et al. 1996). This is carried out in two structurally different regions: (i) the *coding module*, typically made up of one or two open reading frames: *Gag*, which encodes a structural protein involved in the maturation and packaging of mRNA, while the second is made up of a polycistronic mRNA called *Pol* that encodes a polyprotein that contains a protease A (PR), a retrotranscriptase (RT), an integrase (INT), and an RNase-H (RH) (Kumar and Bennetzen 1999; Havecker et al. 2004); and (ii) the *regulatory module*, made up of two long terminal repeats (LTR) between 100 bp and 5 kbp that flank the coding portion and it carries promoters, regulatory elements, and terminators (Benachenhou et al. 2013; Cavrak et al. 2014; Galindo-González et al. 2017; Neumann et al. 2019). At the time of insertion of an individual retrotransposon, both LTRs are identical; however, as time passes, they accumulate a series of independent mutations that produce divergent sequences, but that maintain a high degree of similarity (SanMiguel et al. 1998; Bowen and McDonald 2001).

Using a taxonomic classification with a phylogenetic criterion, plant retroelements were grouped into two major superfamilies, Copia and Gypsy. Despite both groups being ancestral and sharing common characteristics such as their life cycle, genome organization, and protein functions, their origin is polyphyletic (Kumar and Bennetzen 1999). Thus, significant differences are found in protein organization and their amino acid sequence. Moreover, each Superfamily is derived in different evolutive lineages. In Angiosperms, six canonical lineages in Gypsy (Athila, Tat, Galadriel, Reina, CRM/CR, and Del/Tekay) and seven canonical lineages in Copia (Tar/Tork, Angela/Tork, GMR/Tork, Maximus/Sire, Ivana/Oryco, Ale/Retrofit, and Bianca) were widely accepted by the international community working in the field (Wicker and Keller 2007; Du et al. 2010; Llorens et al. 2011; Domingues et al. 2012). Following the same criterion but on a finer scale, lineages consist of copies of retroelements that share different degrees of sequence similarity. Copies of retroelements with high sequence similarity are considered to belong to the same family, which is the last accepted taxonomic level for retroelements (Wicker and Keller 2007). Whereas canonical lineages derived from Gypsy and Copia LTR retroelements are ancestral and mimics eukaryotic macroevolution (Llorens et al. 2009), the radiation of families of retroelements is involved in microevolutionary processes such as adaptation and speciation (Bourgeois and Boissinot 2019).

At a functional level, LTR retroelements might be classified into two main groups based on their capacity to complete their life cycle, autonomous and non-autonomous. The first group corresponds to fully functional elements and is characterized by carrying on all the essential components for its self-retrotransposition (Schulman and Kalendar 2005). Thus, individual autonomous copies may be, to varying degrees, transcriptionally or translationally competent (translation leading to a functional protein) or active. Contrarily, non-autonomous retroelements are characterized by lacking some of the coding domains for the main proteins required for replication retrotransposons (Sabot and Schulman 2006). This group can be generated by deletion or mutation from autonomous retroelements, and they can gain mobility parasitizing autonomous members of the same or related families (Sabot and Schulman 2006). Despite that some non-autonomous retroelements demonstrate successful radiation with a fairly uniform structure in plant genomes (Myers 2001; Jiang et al. 2002a), the vast majority of them correspond to older or fossil elements that have experienced severe deletions or fragmentation by unequal homologous recombination and illegitimate recombination (Devos et al. 2002; Ma et al. 2004; Du et al. 2010). Hence, they were not able to be dated or delimited precisely at lineage and family level using the traditional procedure, as well as the identification of the autonomous partner families that gave rise to them is very difficult (Jiang et al. 2002b; Kalendar et al. 2004; Kejnovsky et al. 2006). Thus, under a functional and taxonomic perspective, the analysis of autonomous retroelement populations in genomes shows to be a robust tool to understand the dynamics of recently amplified families potentially active nowadays (Du et al. 2010; Marcon et al. 2015; Paz et al. 2017).

Despite the great diversity of retroelement families present in plant genomes, most of the copies are in a quiescent state, highly regulated by genetic and epigenetic host genome mechanisms (Vicient 2010; Beulé et al. 2015). This state can be altered due to different types of stress, promoting the activation and expression of certain copies (Hirochika and Hirochika 1993; Grandbastien 1998; Paz et al. 2015). During this process, a single activated copy of a retroelement can produce a large number of identical copies of itself, which could be inserted at new positions within the host genome (Biémont and Vieira 2006; Feschotte and Pritham 2007; Wicker and Keller 2007; Zhao and Ma 2013). Therefore, depending on the window of deregulation caused by the host genome, waves of family-specific amplification can occur. When the genome regains control and activates the silencing mechanisms, the newly inserted copies occupy a permanent place in the genome, with different genetic consequences. If this phenomenon occurs in the reproductive tissues, genetic modification becomes heritable (Maupetit-Mehouas and Vaury 2020).

Several tools have been developed to study the dynamics of retroelements in genomes. On the one hand, the divergence between the LTR sequences within a retroelement has proven an excellent molecular clock to determine the time of insertion of each copy to identify those that have recently been inserted in a given genome. (Ma et al. 2004; Sharma et al. 2008; Kijima and Innan 2009; Paz et al. 2017). On the other hand, there is evidence that retroelements are integrated into genomic regions in a non-random manner (Gao et al. 2008; Baucom et al. 2009; Nellåker et al. 2012). From that perspective, certain retroelements tend to be inserted in regions that are not silenced and have less competition or regions enriched with other retroelements (SanMiguel et al. 1996; Gao et al. 2008; Naito et al. 2009). In this way, the application of the FISH technique allows us to identify affinities of lineages of retroelements towards particular chromosome regions.

In this study, we analyzed the dynamics of autonomous retroelements in the *Capsicum annuum* genome. This species is diploid (2n = 24) but with the peculiarity of having a relatively large genome (3.26 Gb) if compared to other solanaceous species with the same ploidy level (Bennetzen 2002; Park et al. 2011; Qin et al. 2014). This "genomic obesity" is due mostly to the accumulation of repetitive sequences, especially mobile elements (Bennetzen 2002), so pepper is an ideal model for the study of expansion and distribution of LTR retroelements. Thereby, this research aims to identify recently radiated LTR retroelement lineages in *C. annuum* genome and determine their insertional preferences along chromosomes. Moreover, a database of autonomous and potentially active families of LTR retroelements specific to *C. annuum* genome was obtained. The information provided here is of interest for the study of intraspecific genetic variability in pepper, since most of the retroelement-based markers in the species are heterologous.

Materials and methods

Bioinformatic analysis

Data mining

The nuclear genome sequence of the *Capsicum annuum* cv. Zunla (Ref_v1.0) was obtained from the GenBank database (accession no. ASJU00000000). De novo LTR retroelements were identified with LTR-Finder software (http://tlife.fudan.edu.cn/ltr_finder/; Xu and Wang 2007) with the following parameters: (i) minimal distance between LTRs: 3500 bp; (ii) ps_scan algorithm to detect protein domains of RT, IN, and RH if they are identified; (iii) conserved domain prediction PBS (primer binding sequence) which was conducted assigning as a reference genome the database of "*Arabidopsis thaliana* (2004)"; (iv) presence of conserved sequences, such as conserved endings TG-CA; and (v) contain at least two of the following features: TSR (terminal repeated sequences), PBS, and PPT (polypurine tract terminal).

The sequence between the two putative LTR (internal region) was subsequently analyzed in the Conserved Domains databases at NCBI in the same way as described by Paz et al. (2017). Structurally full-length elements were defined as those containing both LTRs and an internal portion encoding for all the typical proteins of Gypsy and Copia superfamilies (Fig. S1A). Full-length elements were annotated and the amino acid sequences of the RT for phylogenetic analyses were extracted from the list of domain hits provided in the output of the Conserved Domains database in the same manner as described in Fig. S1A. Truncated elements and fragments were not considered in this study (Fig. S1B). Retroelement families were defined by evolutionary relationships based on a phylogeny tree of RT.

Phylogenetic analyses

The evolutionary relationships of all the copies of LTR retroelements annotated were analyzed. Reference sequences from previously characterized LTR retroelements from different plant host organisms including several Solanaceae spp. were included (Table S1). Protein sequences were aligned in Seaview using Muscle (Gouy et al. 2010). Maximum likelihood phylogenetic analyses based on the amino acid sequence of the RT were performed with version 7.2.8 RAxML, under the JTT + Γ model. One hundred rapid bootstrap inferences were done with RAxML.

Retroelement families were defined by LTR sequence clustering and by evolutionary relationships based on a phylogeny tree of RT. This analysis revealed similar results to phylogenetic analysis using RH (Paz et al. 2017). Copies of LTR retroelements with bootstrap values higher than 95% were considered belonging to the same family. Once the families were defined, a reference sequence of one member per family was selected and submitted to the DDBJ database (www.ddbj.nig.ac.jp), accession numbers LC434324–LC434447.

To validate and to determine the physic distribution of each retroelement family in *C. annuum* reference genome, a UGENE BLAST was performed considering 85% of sequence identity with the software Unipro UGENE version 1.31 (Okonechnikov et al. 2012). In addition, Pearson’s correlation analysis between the frequency of each family identified by the LTR finder and the number of hits identified by UGENE was performed.

Estimation of insertion time for LTR retrotransposons in pepper

The insertion time was estimated according to the method described by Ma et al. (2004). The CLUSTAL multiple alignment method from MEGA4 (Tamura et al. 2007) was used to align all LTR pairs. The Kimura two-parameter method was used to calculate the distance (d) estimations and the SE for all LTR pairs, under the complete deletion option (Tamura et al. 2007). The rate variation among sites was modeled with a gamma distribution (shape parameter = 8). SE estimates were obtained by using the analytical formula option in MEGA4. Insertion times were estimated by using the following equation: $t = d/2r$. The rate (r) of neutral evolution of 1.3×10^{-8} substitutions per site per year was used (Ma et al. 2004).

Comparative genomic analysis

The dynamics and radiation of retroelements families identified in *C. annuun* in the genomes of another Solanaceae species were determined by performing a BLAST of the complete reference sequence of each family identified against the NCBI database (www.blast.ncbi.nlm.nih.gov). Eight genomic top-level sequences are currently available in the taxid 4070: (i) *Solanum lycopersicum* (SL3.0; GCF_000188115.4; 1350 sequences); (ii) *Solanum pennellii* (SPENNV200 ; GCF_001406875.1; 12 sequences); (iii) *Solanum tuberosum* (SolTub_3.0; GCF_000226075.1; 14,854 sequences); (iv) *Nicotiana tabacum* (Ntab-TN90; GCF_000715135.1; 168,247 sequences); (v) *Nicotiana tomentosiformis* (Ntom_v01; GCF_000390325.2; 159 ,548 sequences); (vi) *Nicotiana attenuata* (NIATTr2; GCF_001879085.1; 37,194 sequences); (vii) *Nicotiana sylvestris* (Nsyl; GCF_000393655.1; 253,918 sequences); (viii) *Capsicum annuum* (Pepper Zunla 1 Ref_v1.0; GCF_000710875.1; 1627 sequences).

BLAST parameters were set as follows: (i) Database, RefSeq Genome Database; (ii) Organisms, Solanaceae (taxid: 4070); (iii) Program selection, Megablast (Highly similar sequence); and (iv) Algorithms general parameters, Max target sequences selected to display among 500 and 1000 aligned sequences according to the number of hits. Additional BLAST analyses were performed using retrotransposons identified from Solanaceae genomes described in Table S1. Results were filtered by Query Coverage range from 80 to 100%, E-value = 0.0 and Score > 200. The number of hits was graphed by species and retroelement family.

Experimental analysis

Plant material

Seeds of *C. annuum* cv. Zunla kindly provided by Dr. Qin Cheng (Zunyi Academy of Agricultural Sciences, China) were used. Twenty plants were pre-germinated in a Petri dish on wet paper for a week. Once they emerged, they were transplanted in pots of 10 cm of diameter filled with sterilized soil as a substrate and maintained in a greenhouse with a photoperiod of 16/8 h at 24/19 °C (day/night), values of relative humidity of

60 ~ 80%, and a light intensity of 200 μmol m^{-1} s^{-1}. DNA was extracted by CTAB II procedure (Weising et al. 2005) from foliar tissue obtained from Zunla plants.

Chromosome preparation

Mitotic chromosomes were examined in root tips obtained from plants grown as previously described. Roots were pretreated in 8-Didroxinonolein 2 mM during 4–5 h at 14 °C and fixed in EtOH:acetic acid (3:1; v/v), washed in distilled water, digested 45 min at 37 °C with Pectinex SP ULTRA® (Novozymes), and squashed in a drop of 45% acetic acid. After coverslip removal in liquid nitrogen, the slides were stored at −20 °C.

LTR retroelement family selection and specific probe design and construction

Specific probes for each of the three most abundant LTR retroelement families identified in *C. annuum* genome, GypsyZla_16; GypsyZla_13, and CopiaZla_01 (see "Results"), were developed as described below. Firstly, a consensus sequence for each family of retroelements was obtained from multiple alignments of at least twenty members of the family using the Muscle tool in the software MEGA4. Then, specific primers were designed over the RT and RH conserved regions of each retroelement family with the primer3 online software (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) with the following selection criteria: (i) primer length among 25 and 29 bp; (ii) CG content among 40 and 60%; (iii) the non-formation of dimers and self-complementarity; (iv) similar Tm among primers; (v) that amplicon must include portions of conserved RT and RH sequences and range from 800 to 1100 pb. Designed primers sequences are detailed in Table S2.

Secondly, probes were amplified with the specific primers described previously using the Zunla DNA as a template. PCR was performed with 25 nmol of the template, 1 μl PCR buffer 10×, 1.4 μl $MgCl_2$ (25 mM), 1.0 μl primer Fw (10 μM), 1.0 μl primer Rv (10 μM), 0.5 μl of dNTPs (10 mM), and 1 unit of Taq DNA polymerase (Invitrogen®), in a final volume of 10 μl. Thermocycler program consisted in 1 cycle of 5 min at 94 °C; 35 cycles of amplification of 45 s at 94 °C, 45 s at 62.5 °C/68.3 °C/65.8 °C (CopiaCa_01/GypsyCa_16/GypsyCa_13, respectively), and 30 s at 72 °C; and a final elongation cycle of 10 min at 72 °C. Amplified fragments were extracted from agarose gels and cloned in the plasmid vector pGEM-T Easy (Promega) in the same way as described in Paz et al. (2015). Single clones positive for inserts were selected for sequencing. The plasmid DNA of individual clones was obtained by the alkaline lysis procedure. The presence of insert in the purified plasmids was verified by PCR reaction with the universal M13F (5′-GTAAAACGACGGCCAG-3′) and M13R (5′-CAGGAAACAGCTATGAC-3′) primers. DNA sequencing was carried out with M13 forward primer by Macrogen Inc. (Seoul, South Korea). All nucleic acid sequences obtained were screened for vector contamination using the Vector Screen program (www.ncbi.nlm.nih.gov/VecScreen) and primer sequences were removed. The obtained nucleotide sequences were deposited in the DDJJ database (DNA Data Bank of Japan, www.ddbj.nig.ac.jp) under accession nos. LC431733–LC431740 (Table S2).

Homology search between the obtained sequences and their respective LTR retrotransposon family was conducted using the tool BLAST2 of the National Center of Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). The homology assignation criterion was based on maximum sequence cover (>90%), maximum identity (>60%), and a minimum E-value of 10^{-20}. Based on this criterion, three clones were selected for probe construction, two family-specific (*P-GypsyCa_16* and *P-CopiaCa_01*), and one clade-specific (*P-[Del/Tekay]-complex*) (see "Results"). The obtained amplicons by PCR with M13 primer and purified plasmids were used as a probe for FISH. Purified DNA was labeled with Digoxigenin-11-dUTP (DIG Nick translation mix, Roche) according to the manufacturer's recommendations.

Fluorescent in situ hybridization

To investigate the chromosomal distribution of specific probes *P-GypsyCa_16*, *P-CopiaCa_01*, and *P-[Del/Tekay]-complex* in *C. annuum* genome, we performed fluorescent in situ hybridization (FISH) on somatic metaphase chromosomes and interphase nuclei. The location and number of specific signals from different probes were determined by FISH, using the protocol described by Schwarzacher and Heslop-Harrison (2000) with minor modifications. The preparations were incubated in 100 μg/ml RNAase, post-fixed in 4% (w/v) paraformaldehyde, dehydrated in a 70–100% graded ethanol series, and air-dried. On each slide, 15 μl of hybridization mixture was added (4–6 ng/μl of the probe, 50% formamide, 10% dextran sulfate, 2 SSC, and 0.3% SDS), previously denatured at 70 °C for 10 min. Chromosome denaturation/hybridization was done at 90 °C for 10 min, 48 °C for 10 min, and 38 °C for 5 min using a thermal cycler (Mastercycler, Eppendorf, Hamburg, Germany), and slides were placed in a humid chamber at 37 °C overnight. Hybridization signals were detected with Avidin-FITC (Sigma) and/or anti-DIG-Rhodamine (Roche) and preparations were mounted with Vectashield-DAPI (Vector Labs).

At least five metaphases were photographed with phase contrast in an Olympus BX61 microscope with a monochromatic CV-M4+ CL model JAI® camera. All the chromosome images were captured in black and white to be subsequently pseudo-colored. Based on the metaphase photographs, the chromosome arm was divided into four regions of equal size to define the chromosome portions according to Roa and Guerra (2012): centromeric (C); proximal (P); interstitial-proximal (IP); interstitial-terminal (IT); and terminal (T). Hybridization signals were considered as dots taking into account the intensity observed in each chromosome portion. Idiogram was constructed in the base of chromosome measurements according to Levan et al. (1964). For the construction of the cytogenetic map, the absolute distance in micrometers from the hybridization signal to the centromere was measured with Adobe Photoshop CS4 (Adobe Systems Inc.) and then located in the idiogram.

Results

Distribution and frequencies of full-length LTR retroelements in pepper genome

The analysis using LTR finder over *C. annuum* genome identified 3522 hits. From this data, the search of complete and intact LTR retroelements on *C. annuum* genome yielded 1151 structurally full-length retroelements distributed across the 12 chromosomes, 340 belonging to Superfamily Copia (30%) and 811 to Superfamily Gypsy (70%) (Table 1; Table S3). Ratios of Copia:Gypsy retroelement were highly variable among chromosomes, ranging from 0.2 to 1.2. In the same way, there was a ~5-fold retroelement density variation among one of the most and least populated chromosomes (Ch06 vs Ch07). This variation in retroelement number was not correlated with chromosome sizes (Pearson's correlation 0.33; $p = 0.30$).

Evolutionary relationships and family radiation of full-length LTR retroelements

Phylogenetic analysis in the base of RT conserved domains revealed the presence of 124 families of LTR retroelements in *C. annuum* genome belonging to all the phylogenetic clades described in plants (Fig. 1; Fig S2; Fig S3; Table 2 and Table S4). However, some differences were observed in frequencies and family

Table 1 Distribution, frequency, and density of the 1151 full-length LTR retroelements identified in the 12 chromosomes of the pepper genome

Chromosome	No. of LTR retrotransposons [Copia;Gypsy]	Ratio Copia:Gypsy	Density per 10 million of bp Total [Copia/Gypsy]	Chromosome size (bp)
Ch01	99 [44;55]	0.8	3.3 [1.5/1.8]	301,019,445
Ch02	102 [25;77]	0.3	6.2 [1.5/4.7]	163,962,470
Ch03	139 [43;96]	0.4	5.3 [1.6/3.7]	261,511,374
Ch04	101 [29;72]	0.4	4.7 [1.3/3.3]	215,701,946
Ch05	77 [15;62]	0.2	3.5 [0.7/2.9]	217,274,494
Ch06	141 [45;96]	0.5	6.4 [2.0/4.4]	219,521,584
Ch07	27 [15;12]	1.2	1.2 [0.7/0.5]	222,112,641
Ch08	63 [18;45]	0.4	4.1 [1.2/2.9]	153,299,543
Ch09	143 [31;112]	0.3	6.0 [1.3/4.7]	238,794,889
Ch10	86 [28;58]	0.5	4.2 [1.4/2.8]	205,736,368
Ch11	60 [26;34]	0.8	2.7 [1.2/1.5]	220,335,243
Ch12	113 [21;92]	0.2	4.9 [0.9/4.0]	229,934,170
Total	1151 [340;811]	0.4	4.3 [1.3/3.1]	2,649,204,167

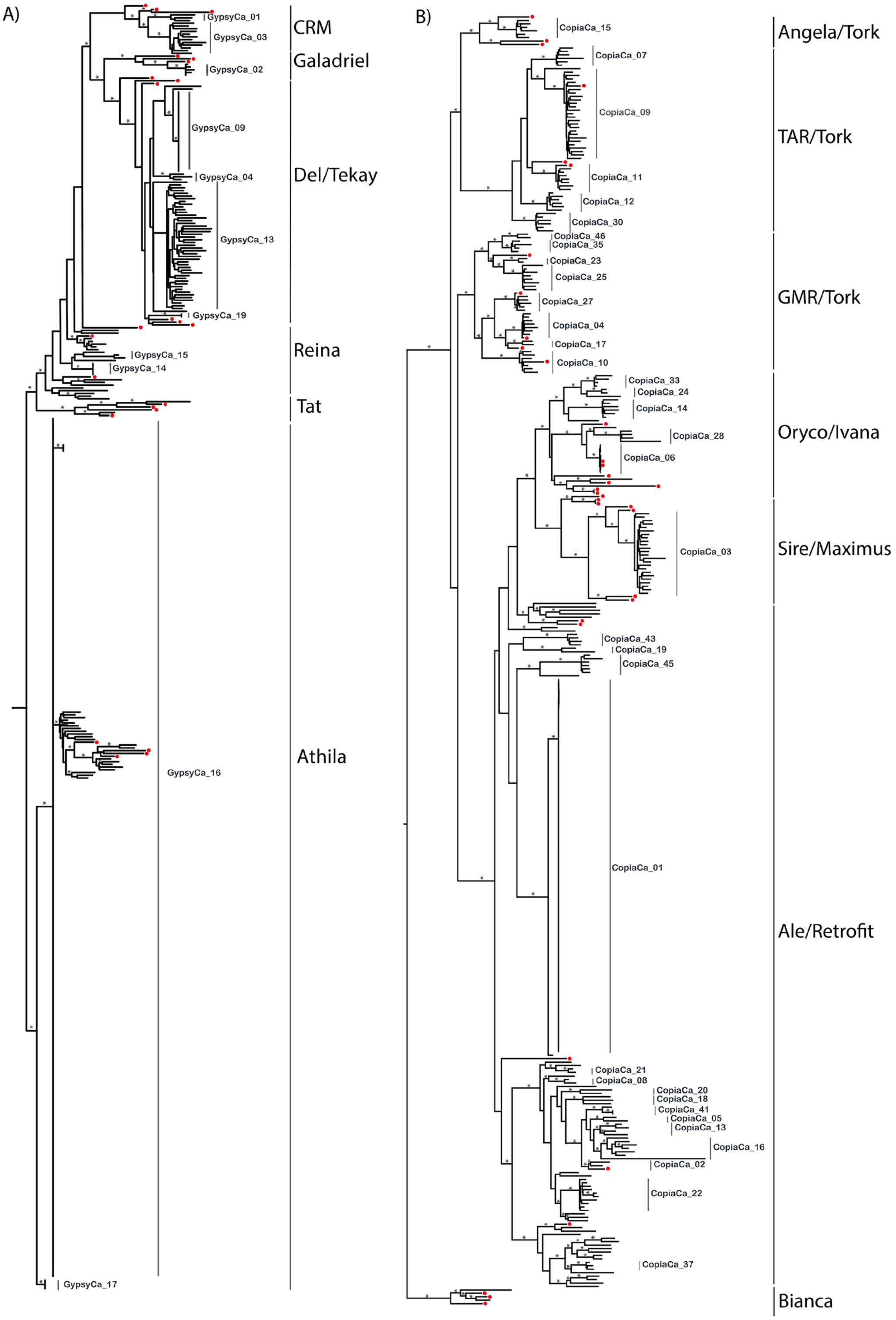


Fig. 1 Phylogenetic trees based on amino acid sequences of the RT from 811 Gypsy (A) and 341 Copia (B) full-length LTR retroelements identified in *C. annuum* genome. Reference sequences are indicated with a red circle, whereas significant bootstrap support values higher than 50% are indicated in the branches with an asterisk. For aesthetic reasons, the branch of the family GypsyCa_16 was reduced nearly to its half. Original trees for both Superfamilies are available in Fig. S2 and Fig S3

Table 2 Number of retroelements and families classified according to plant phylogenetic clades

Superfamily	Phylogenetic clades described in plants	No. of retroelements	No. of families (*)
Copia	Ale/Retrofit	194	48 (33)
	Angela/Tork	6	1 (0)
	Bianca	2	2 (2)
	GMR/Tork	35	8 (0)
	Oryco/Ivana	29	8 (0)
	Sire/Maximus	24	1 (0)
	Tar/Tork	50	5 (0)
Gypsy	Athila	682	23 (21)
	CRM	15	2 (0)
	Del/Tekay	85	6 (2)
	Galadriel	5	1 (0)
	Reina	22	17 (15)
	Tat	2	2 (2)

*indicates monotypic families

radiation among clades. In this way, the clades Athila, Ale/Retrofit, and Del/Tekay are the most populated, comprising 83% of the entire LTR retroelement population (Athila, 59%; Ale/Retrofit, 17%; and Del/Tekay, 7%) and 62% of all the retroelement families (Athila, 19%; Ale/Retrofit, 39%; and Del/Tekay, 5%). Contrarily, the clades Angela/Tork, Bianca, Galadriel, and Tat are sparsely populated, with less than 5 specimens and very few families. There was a slight but non-significant correlation between the number of retroelements per clade and the number of families (Pearson's correlation: 0.53; $p = 0.0634$).

At family level, only 6 families encompass 77.6% of LTR retroelement population in *C. annuum* genome, with a relative frequency (F_R) higher than 2% (Table 3). Sorted in decreasing F_R order: (i) GypsyCa_16, $F_R = 57.1\%$ (Athila clade); (ii) CopiaCa_01, $F_R = 9.5\%$ (Ale/Retrofit clade); (iii) GypsyCa_13, $F_R = 4.2\%$ and (iv) GypsyCa_09, $F_R = 2.6\%$ (both belonging to Del/Tekay clade); (v) CopiaCa_09, $F_R = 2.2\%$ (TAR/Tork clade); and (vi) CopiaCa_03, $F_R = 2.1\%$ (Sire/Maximus clade). The remaining 22.4% of the identified retroelements are distributed along 118 low populated families, of which 78 families are constituted by only one retroelement (monotypic). These results were validated by UGENE Blast against the reference genome with high positive significative correlation (LTR finder vs UGENE, Pearson's correlation 0.84, $p < 0.0001$; Fig. S4).

All the full-length LTR retroelement families identified in this research have been inserted into *C. annuum* genome in a period shorter than 5.85 Mya (Table 3; Fig. 2; Table S3). The distribution of insertion times of Gypsy and Copia superfamilies has a leptokurtic distribution with asymmetry to the left (Fisher's coefficient, Gypsy: G = 6.2; Copia: G = 2.6), indicating that the vast majority of events of insertion occurred less than 1.0 Mya (84% and 38% of insertions respectively; Fig. 2A, B). It is noteworthy that 22% of the identified complete and intact retroelements were currently inserted (0.0 Ma).

Different waves of amplification were observed (Fig. 2). In the case of Copia, the three clades Tar/Tork, Sire/Maximus, and GMR/Tork exhibited a similar trend, experiencing a gradual increase from 4.5 Mya, with a climax to 3 Mya, followed by a gradual reduction, with very few new insertions in the last period. Another wave of expansion was experimented with more recently by the Copia clades Ale/Retrofit and Oryco/Ivana, with a gradual increase in their population from 3 Mya and a substantial numeric expansion in the last 0.5 Mya, especially in Ale/Retrofit (Fig. 2A). This last clade is the one that experienced the greatest radiation, with a great diversification of families in the first wave of its expansion, dominated by the formation of new monotypic families. In contrast, the second wave was experienced by only one family, CopiaCa_01, which during the last 6 Mya maintained a very low rate of insertion (ranged

Table 3 Frequency, size, and insertion time of LTR retroelement families identified in *C. annuum* genome

Superfam	Clade	Family	F_R [F_A]	Int	LTR	LTR sim	MYA
Gypsy	Galadriel	GypsyCa_02	5 [0.4]	4812±98 [4684;4950]	554±5 [546;558]	0.95±0.04 [0.89;0.98]	1.29±0.61 [0.65;2.12]
	Tat	monotypic	2 [0.2]	9531±1034 [8800;10262]	1159±489 [813;1505]	0.91±0.13 [0.82;1.00]	2.33±3.29 [0;4.65]
	Del/Tekay	GypsyCa_04	3 [0.3]	5272±362 [5039;5689]	1654±770 [765;2118]	0.92±0.03 [0.89;0.94]	2.49±0.50 [1.96;2.96]
		GypsyCa_09	30 [2.6]	6428±2556 [5162;17243]	2449±482 [167;2577]	0.99±0.02 [0.90;1.00]	0.40±0.74 [0.00;4.00]
		GypsyCa_13	**48 [4.2]**	**5383±733 [4276;7899]**	**2104±737 [177;2950]**	**0.9±0.02 [0.86;0.93]**	**3.17±1.25 [0.00;5.46]**
		GypsyCa_19	2 [0.2]	6581±489 [6235;6926]	700±516 [335;1065]	0.91±0.06 [0.87;0.95]	1.89±0.11 [1.81;1.96]
		monotypic	2 [0.2]	6692±1655 [5522;7862]	1220±1527 [140;2299]	0.95±0.00 [0.95;0.95]	2.27±0.49 [1.92;2.62]
	Reina	GypsyCa_14	5 [0.4]	4353±1 [4352;4355]	322±1 [322;323]	1.00±0.00 [0.99;1.00]	0.05±0.10 [0.00;0.23]
		GypsyCa_15	2 [0.2]	4506±28 [4486;4525]	386±1 [385;386]	0.96±0.01 [0.95;0.96]	1.50±0.38 [1.23;1.77]
		monotypic	15 [1.3]	4513±142 [4231;4818]	383±60 [296;462]	0.95±0.02 [0.91;0.98]	1.84±0.70 [0.77;2.85]
	Athila	**GypsyCa_16**	**658 [57.1]**	**9128±707 [1802;17425]**	**1487±58 [539;1548]**	**1.00±0.00 [0.97;1.00]**	**0.15±0.42 [0.00;4.85]**
		GypsyCa_17	3 [0.3]	9332±610 [8627;9684]	1491±2 [1490;1494]	1.00±0.00 [0.99;1.00]	0.13±0.08 [0.04;0.19]
		monotypic	21 [1.8]	9441±1743 [6868;13291]	1370±225 [739;1659]	0.93±0.03 [0.87;0.97]	2.45±1.13 [0.04;4.50]
Copia	CRM	GypsyCa_01	3 [0.3]	5525±91 [5420;5580]	520±3 [517;522]	0.94±0.01 [0.93;0.95]	1.53±1.28 [0.12;2.62]
		GypsyCa_03	12 [1]	6056±262 [5551;6436]	703±152 [258;812]	0.93±0.02 [0.9;0.97]	2.27±0.80 [1.12;3.96]
	Angela/Tork	CopiaCa_15	6 [0.5]	5618±688 [5075;6990]	501±11 [485;513]	0.91±0.04 [0.85;0.95]	2.60±0.46 [1.88;3.19]
	Sire/Maximus	CopiaCa_03	24 [2.1]	7733±603 [6609;9943]	867±70 [566;928]	0.93±0.02 [0.87;0.96]	2.52±0.66 [1.46;3.88]
	Bianca	monotypic	2 [0.2]	5210±185 [5079;5341]	271±69 [222;319]	0.87±0.09 [0.81;0.94]	4.02±2.59 [2.19;5.85]
	Tar/Tork	CopiaCa_07	6 [0.5]	4454±253 [3956;4682]	535±109 [330;630]	0.91±0.03 [0.87;0.96]	2.75±0.91 [1.46;4.19]
		CopiaCa_09	25 [2.2]	5006±352 [4596;6208]	764±305 [222;1608]	0.90±0.03 [0.84;0.95]	2.90±0.95 [0.04;4.42]
		CopiaCa_11	7 [0.6]	5213±1568 [4606;8769]	566±44 [475;607]	0.94±0.02 [0.91;0.97]	2.05±0.79 [1.04;3.38]
		CopiaCa_12	6 [0.5]	4537±236 [4275;4980]	599±76 [482;720]	0.91±0.04 [0.85;0.95]	2.21±1.19 [0.04;3.38]
		CopiaCa_30	6 [0.5]	5244±1379 [4524;8043]	566±234 [215;776]	0.89±0.04 [0.84;0.92]	2.58±1.55 [0.15;4.35]
	GMR/Tork	CopiaCa_04	7 [0.6]	4038±19 [4005;4061]	734±2 [732;737]	0.92±0.01 [0.90;0.94]	3.44±0.50 [2.54;4.00]
		CopiaCa_10	6 [0.5]	4030±39 [3957;4067]	273±44 [205;322]	0.91±0.03 [0.87;0.95]	3.12±1.20 [2.04;4.65]

Table 3 (continued)

Superfam	Clade	Family	F_R [F_A]	Int	LTR	LTR sim	MYA
		CopiaCa_17	2 [0.2]	3954±173 [3831;4076]	561±54 [523;599]	0.90±0.04 [0.87;0.92]	3.33±0.68 [2.85;3.81]
		CopiaCa_23	2 [0.2]	4029±81 [3972;4086]	614±15 [603;624]	0.90±0.01 [0.89;0.90]	2.20±0.11 [2.12;2.27]
		CopiaCa_25	7 [0.6]	4104±25 [4085;4143]	329±70 [270;429]	0.92±0.03 [0.89;0.98]	3.32±0.76 [2.12;4.46]
		CopiaCa_27	5 [0.4]	4073±55 [3978;4111]	567±25 [540;596]	0.91±0.01 [0.90;0.93]	1.98±1.14 [0.12;2.96]
		CopiaCa_35	4 [0.3]	4275±93 [4143;4356]	344±1 [342;345]	0.91±0.05 [0.84;0.95]	2.62±0.36 [2.15;3.04]
		CopiaCa_46	2 [0.2]	4364±43 [4333;4394]	281±3 [279;283]	0.95±0.02 [0.93;0.96]	2.06±0.90 [1.42;2.69]
	Oryco/Ivana	CopiaCa_06	9 [0.8]	4097±31 [4083;4180]	309±8 [289;314]	0.99±0.01 [0.97;1.00]	0.26±0.28 [0.00;0.77]
		CopiaCa_14	6 [0.5]	4407±546 [3929;5270]	384±29 [331;409]	0.93±0.02 [0.89;0.95]	2.47±0.70 [1.58;3.50]
		CopiaCa_24	3 [0.3]	3851±289 [3517;4026]	299±36 [272;340]	0.94±0.05 [0.89;0.99]	1.22±0.74 [0.42;1.88]
		CopiaCa_28	4 [0.3]	4049±41 [4006;4096]	323±2 [320;325]	0.91±0.05 [0.83;0.94]	2.13±0.47 [1.46;2.54]
		CopiaCa_33	4 [0.3]	3988±95 [3905;4113]	213±56 [129;241]	0.92±0.05 [0.85;0.95]	2.71±0.92 [1.85;3.69]
		monotypic	3 [0.3]	4322±259 [4137;4618]	286±6 [279;289]	0.92±0.08 [0.82;0.99]	1.68±1.46 [0.00;2.69]
	Ale/Retrofit	**CopiaCa_01**	**109 [9.5]**	**4538±316 [4226;7633]**	**153±9 [128;244]**	**0.99±0.02 [0.82;1.00]**	**0.22±0.53 [0.00;4.00]**
		CopiaCa_21	2 [0.2]	4374±198 [4234;4514]	266±0 [266;266]	0.95±0.03 [0.93;0.97]	1.06±0.03 [1.04;1.08]
		CopiaCa_22	10 [0.9]	4480±291 [3693;4770]	328±33 [297;406]	0.94±0.02 [0.90;0.96]	1.78±0.82 [0.08;3.04]
		CopiaCa_37	3 [0.3]	4575±31 [4540;4600]	247±89 [144;299]	0.95±0.01 [0.94;0.95]	2.05±0.33 [1.77;2.42]
		CopiaCa_41	3 [0.3]	4892±92 [4786;4945]	306±101 [195;392]	0.92±0.05 [0.89;0.97]	1.46±0.45 [1.08;1.96]
		CopiaCa_43	4 [0.3]	4925±1608 [3884;7276]	276±252 [142;654]	0.94±0.03 [0.91;0.98]	3.17±1.55 [1.42;5.00]
		CopiaCa_45	6 [0.5]	4142±91 [3962;4202]	255±25 [205;267]	0.92±0.03 [0.87;0.95]	2.79±0.92 [0.00;4.04]
		monotypic	33 [2.9]	4616±296 [4006;5838]	299±77 [152;451]	0.94±0.02 [0.90;1.00]	1.93±0.89 [0.00;4.04]
		CopiaCa_02	2 [0.2]	4834±66 [4787;4881]	294±23 [278;310]	0.98±0.01 [0.97;0.98]	0.54±0.22 [0.38;0.69]
		CopiaCa_05	2 [0.2]	4803±267 [4614;4992]	206±122 [120;292]	0.94±0.01 [0.94;0.94]	2.87±0.46 [2.54;3.19]
		CopiaCa_08	2 [0.2]	5757±1814 [4474;7039]	259±0 [259;259]	0.94±0.04 [0.91;0.97]	2.08±1.63 [0.92;3.23]
		CopiaCa_13	4 [0.3]	5506±1984 [4383;8479]	301±15 [289;321]	0.96±0.01 [0.95;0.97]	1.46±0.34 [1.12;1.81]
		CopiaCa_16	7 [0.6]	4457±374 [3736;4820]	324±12 [313;348]	0.95±0.02 [0.91;0.98]	1.64±1.21 [0.04;3.81]

Table 3 (continued)

Superfam	Clade	Family	F_R [F_A]	Int	LTR	LTR sim	MYA
		CopiaCa_18	3 [0.3]	4753±202 [4520;4880]	284±33 [246;308]	0.94±0.05 [0.88;0.98]	1.22±0.26 [0.92;1.42]
		CopiaCa_19	2 [0.2]	4558±10 [4551;4565]	276±9 [269;282]	0.93±0.07 [0.88;0.98]	2.98±0.74 [2.46;3.5]
		CopiaCa_20	2 [0.2]	4567±47 [4533;4600]	295±32 [272;317]	0.97±0.00 [0.97;0.97]	1.21±0.08 [1.15;1.27]

Bold entries indicate the most populated families of retroelements

between 0 and 4 copies each 0.5 Mya), and in the last 0.5 Mya experienced a 100-fold amplification (Fig. 2C).

In the case of the Gypsy superfamily, although this superfamily is less diverse in family radiation than Copia, two of its clades experienced significant radiation processes, Del/Tekay and Athila (Fig. 2B). In the case of Del/Tekay, their retroelements experienced two expansion waves, the first one in the lapse of between 5.0 and 2.0 Mya, and the second within the last 0.5 Mya. On the other hand, in Athila, a rate of insertions of around 2 and 7 retroelements was inserted each 0.5 Mya. However, in the last 0.5 Mya, this rate has increased to more than 600 new insertions (Fig. 2B). The analysis of the family dynamic in those clades revealed the concordance of these waves with the amplification of only three LTR retroelement families: GypsyCa_09; GypsyCa_13; and GypsyCa_16 (Fig. 2C). The first two families were associated with the first and second waves of expansion observed in Del/Tekay clade respectively, whereas the third is the main family responsible for the expansion observed in Athila in the last 0.5 Mya. Similarly, as observed in Ale/Retrofit, the wave of amplification of Athila families previous to the amplification of GypsyCa_16 was mainly due to the radiation on monotypic families.

Comparative genomic analysis

All the retroelement families identified in this work were subsequently recovered by BLAST on the reference genome of *C. annuum* (Table 3; Table S6). Also, a significant positive correlation was found between the absolute frequency of copies in each family and the number of hits identified by BLAST (LTR finder vs NCBI BLAST hits against *C. annuum* Zunla genome, Pearson correlation's 0.45, $p < 0.0001$). When extending this analysis to the 8 reference Solanaceae genomes available in the RefSeq Genome Database, it was observed that most of the families were exclusive of *C. annuum* (92 families, 74%), this trend being more prominent in Gypsy (48 families, 86% of Gypsy families) than in Copia (44 families, 65% of Copia families) (Fig. 3). Besides, the family CopiaCa_01; and all members of the Athila and Del/Tekay lineages belonged to this group of retroelements specifically radiated in *C. annuum*.

Of the remainder 32 families of retroelements identified in *C. annuum*, only five were universal (they presented copies in the 8 reference genomes), the families CopiaCa_30; CopiaCa_07; CopiaCa_11; monotypic-Ch06a_26; and GypsyCa_15 (Fig. 3B, C). Likewise, some families derived from the GMR/Tork lineage presented high radiation in *Nicotiana* spp. (CopiaCa_23 and CopiaCa_35) while others presented radiation in Solanum spp., particularly *S. tuberosum* (CopiaCa_30 and CopiaCa_45, belonging to the TAR/Tork and Ale/Retrofit lineages respectively).

When comparing these results with the behavior of the 28 families of retroelements identified in other species of Solanaceas, a similar trend was observed. Few families presented a universal radiation (Tnt1; CopiaSL_23; CopiaSL_25; CopiaSL_26), most of them presented a gender-specific radiation (Fig. 3B, C). Thus, a large number of retroelement families were previously identified in Solanum spp. specifically radiated within species of the genus (CopiaSL_05; CopiaSL_15; CopiaSL_17; Tork4/CopiaSL_37; GypsySL_01; GypsySL_03; GypsySL_04; GypsySL_05; GypsySL_07; GypsySL_11; GypsySL_monotypic|Ch03_1s10; GypsySL_monotypic|Ch12_1s55), while others identified in *Nicotiana* spp. behaved similarly (Tto1 and Tntom1) (Fig. 3).

Another aspect to highlight that emanates from this analysis is that the lineages derived from Copia presented a greater degree of conservation among hosts of different Solanaceae genera, while the families derived

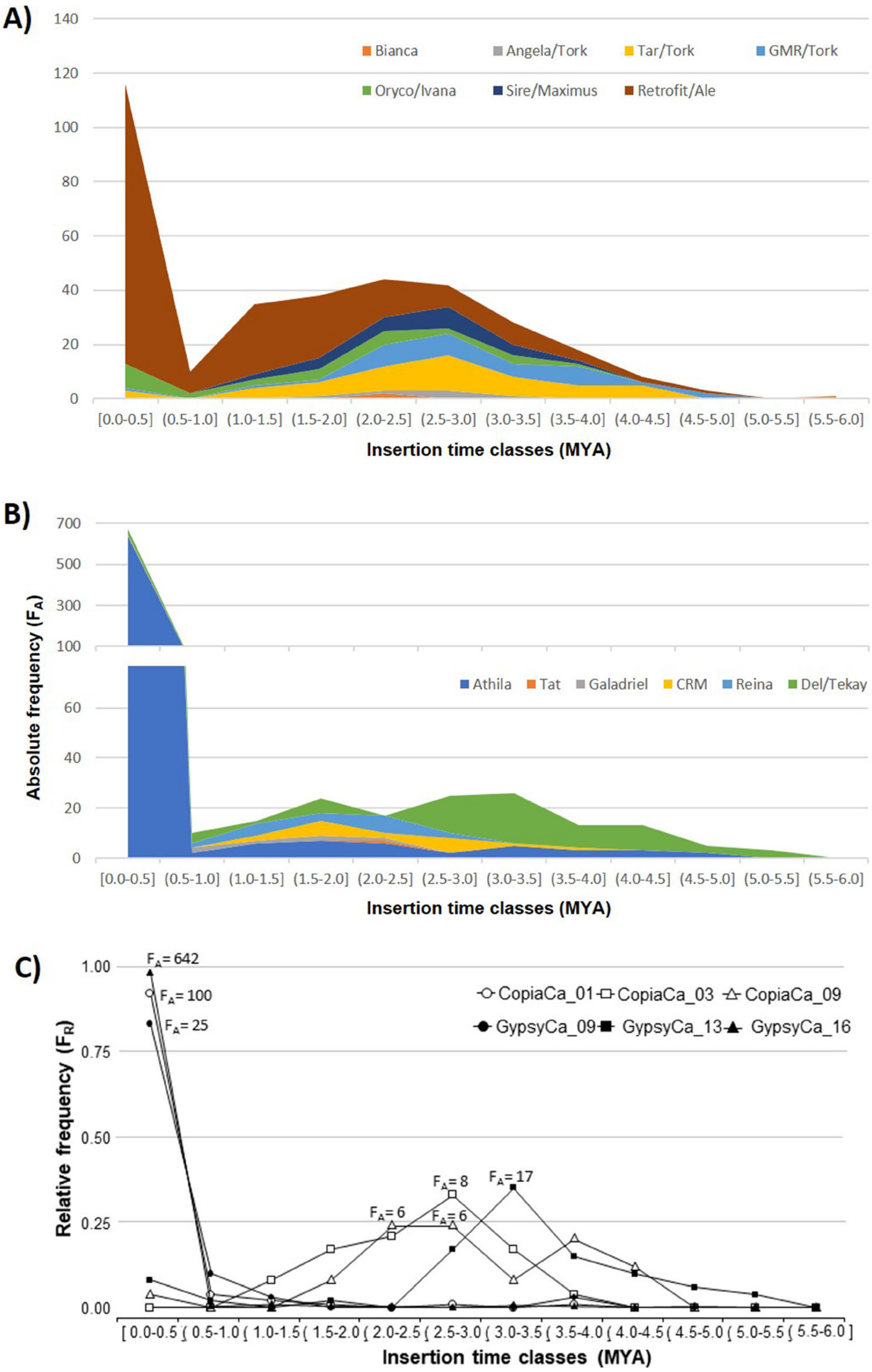


Fig. 2 Radiation waves of different clades and families of full-length LTR retroelements identified in *C. annuum* genome. (A, B) Insertion dynamics each 0.5 Mya of members belonging to the different clades of Copia and Gypsy Superfamilies expressed in absolute frequencies, respectively. (C) Dynamic of most populated families of LTR retroelements identified expressed in relative terms. Estimated insertion times were divided into bins of 100,000 years

from Gypsy have a greater degree of divergence. This is revealed in the fact that only the GypsyCa_15 family was identified in the genomes of other species. This same behavior was observed with the Tntom1 (Galadriel) family identified in *N. tabacum* and with the GypsySL_01, GypsySL_03, and GypsySL_05 families, and Gypsode1/GypsySL_07 (all derived from Del/Tekay), exclusive to *Solanum*.

Distribution of most abundant families of LTR retrotransposons in the genome of *C. annuum*

The three probes showed homology with their respective retroelements families/lineages (Table 4) and had hybridized along all the chromosomes of Zunla, with differences in the number of signals of hybridization relative to each probe (Figs. 4 and 5). Thus, *P-GypsyCa_16* showed a higher number of hybridization signals than *P-CopiaCa_01* and *P-[Del/Tekay]-complex*, with similar values (Fig. 4). These values are proportionally in agreement with the number of retroelements identified by bioinformatic analysis. In some situations, differences were observed in the presence of hybridization signals between the homologous chromosomes (Fig. 5).

A differential retroelement insertion pattern along *C. annuum* chromosome distribution patterns (Fig. 6) was observed. Thus, the distribution of *P-GypsyCa_16* and *P-[Del/Tekay]-complex* probes shares a similar pattern, where the signals were concentrated mainly in interstitial-proximal and proximal regions of the chromosomes, followed by interstitial-terminal regions and practically absent in terminal and centromeric regions. The only exception were those incidences in centromeric regions, where it was related to zero in *P-[Del/Tekay]-complex* whereas it reached a moderate frequency in *P-GypsyCa_16*. Contrarily, in the case of *P-CopiaCa_01*, the pattern of insertion was marked by a high incidence in the terminal region, followed by a moderate incidence in the proximal and interstitial region and lower incidence in centromeric and interstitial-terminal regions.

Discussion

Complete and intact retroelements are a minor fraction in the universe of repetitive sequences

Different studies at the genomic level in Solanaceae species have revealed that LTR retroelements constitute the major fraction of repetitive sequences (between 20 and 80% depending on the species; Qin et al. 2014; Xu et al. 2017; Gaiero et al. 2019). Thus, they have been identified as the main contributors to the variation in the genomic size of this botanical family, being constituted mainly by two fractions: (i) ancestral or fossil retroelements, generated by the gradual loss of the different components of the retroelements throughout evolution, giving rise to truncated, incomplete, and non-autonomous elements (Ma et al. 2004); and (ii) Solo-LTR retroelements (consisting only of LTR-5′ and LTR-3′, lacking the internal coding portion), generated by local homologous recombination between both LTRs of the same element (Vicient et al. 1999; Xu and Du 2014). Those fractions are resulting from the cellular mechanisms that regulate the activity of retroelements, and aim to interrupt their life cycle.

A third minor fraction, and study subject of this work, consists of those (iii) complete and intact retroelements characterized by carrying on all the essential components for its retrotransposition. This autonomous and mobile fraction of the genome can impair changes in gene or genome structure, often with accompanying alterations in gene activity, promoting genome divergence and evolution (Bennetzen 1996, 2000; Raskina et al. 2008; Belyayev 2014; Bennetzen and Wang 2014; Anderson et al. 2019). Their action potential can be substantially increased by enhancing the activity of non-autonomous elements that hack them (Sabot and Schulman 2006). In our study, this fraction represents 0.4% of the pepper genome; this value is similar in magnitude range to those found in the genome of other plant species (Vitte et al. 2007; Beulé et al. 2015; Yadav et al. 2015; Paz et al. 2017), as in other species of the reported genus Capsicum (De Assis et al. 2020).

Fig. 3 Comparative genomic analysis of the radiation of LTR retroelement families in Solanaceae. (A) Retroelements radiated exclusively in *C. annuum* genome. (B, C) Radiation of Gypsy and Copia retroelement families in different Solanaceae genomes. The size of the circles is proportional to the number of hits found by BLAST analysis of a reference sequence of each LTR retroelement family against RefSeq genome database (filtered by Solanaceae Taxid: 4070), whereas the color of circles refers to Solanaceae species. Asterisks indicate the families of retroelements selected for fluorescent in situ hybridization (FISH) analysis. Circles below some families of retroelement names indicate reference retroelement families identified in other Solanaceae species ▶

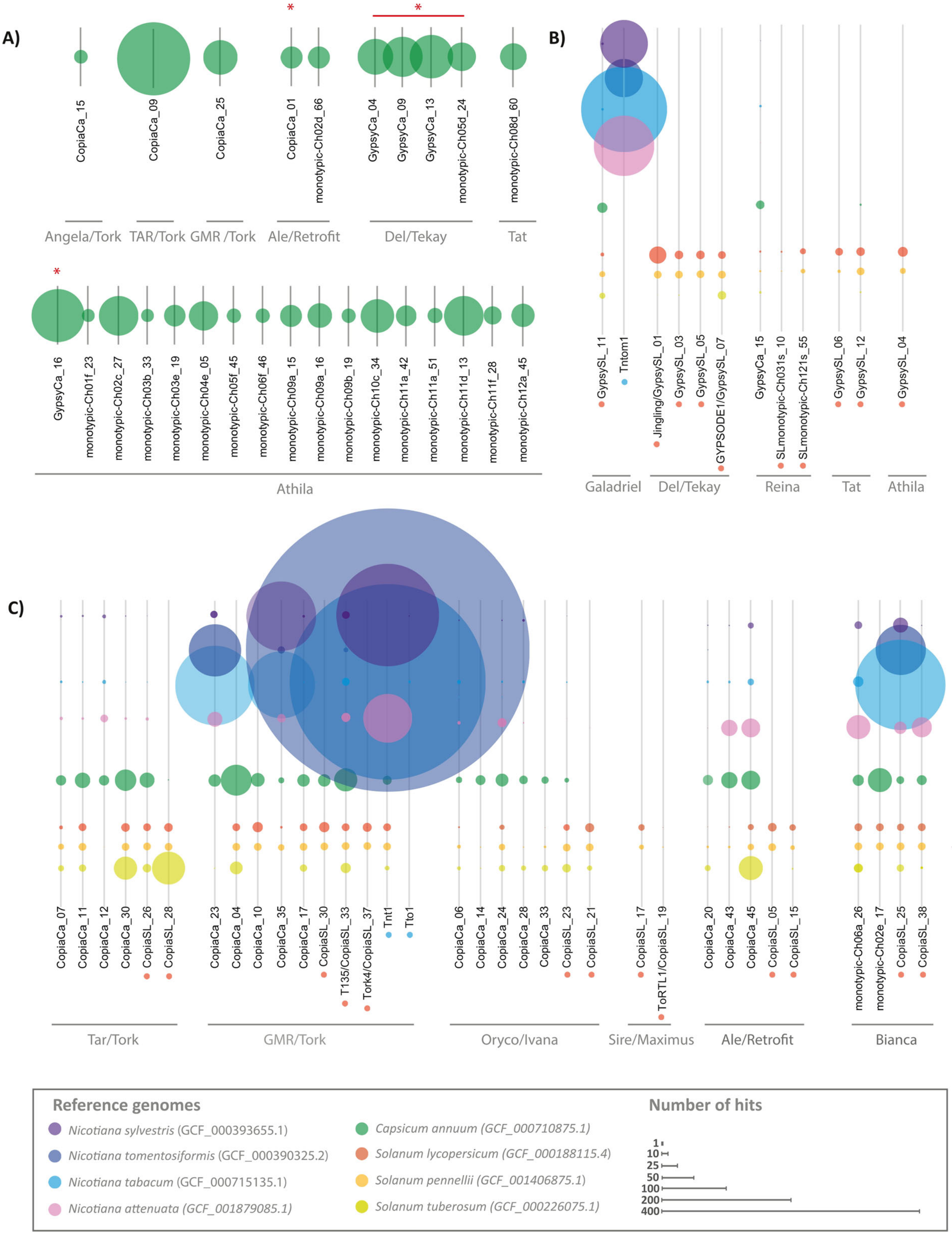

A)
*
*
CopiaCa_15
CopiaCa_09
CopiaCa_25
CopiaCa_01
monotypic-Ch02d_66
GypsyCa_04
GypsyCa_09
GypsyCa_13
monotypic-Ch05d_24
monotypic-Ch08d_60
Angela/Tork
TAR/Tork
GMR /Tork
Ale/Retrofit
Del/Tekay
Tat
*
GypsyCa_16
monotypic-Ch01f_23
monotypic-Ch02c_27
monotypic-Ch03b_33
monotypic-Ch03e_19
monotypic-Ch04e_05
monotypic-Ch05f_45
monotypic-Ch06f_46
monotypic-Ch09a_15
monotypic-Ch09a_16
monotypic-Ch09b_19
monotypic-Ch10c_34
monotypic-Ch11a_42
monotypic-Ch11a_51
monotypic-Ch11d_13
monotypic-Ch11f_28
monotypic-Ch12a_45
Athila
B)
GypsySL_11
Tntom1
Jingling/GypsySL_01
GypsySL_03
GypsySL_05
GYPSODE1/GypsySL_07
GypsyCa_15
SLmonotypic-Ch031s_10
SLmonotypic-Ch121s_55
GypsySL_06
GypsySL_12
GypsySL_04
Galadriel
Del/Tekay
Reina
Tat
Athila
C)
CopiaCa_07
CopiaCa_11
CopiaCa_12
CopiaCa_30
CopiaSL_26
CopiaSL_28
CopiaCa_23
CopiaCa_04
CopiaCa_10
CopiaCa_35
CopiaCa_17
CopiaSL_30
T135/CopiaSL_33
Tork4/CopiaSL_37
Tnt1
Tto1
CopiaCa_06
CopiaCa_14
CopiaCa_24
CopiaCa_28
CopiaCa_33
CopiaSL_23
CopiaSL_21
CopiaSL_17
ToRTL1/CopiaSL_19
CopiaCa_20
CopiaCa_43
CopiaCa_45
CopiaSL_05
CopiaSL_15
monotypic-Ch06a_26
monotypic-Ch02e_17
CopiaSL_25
CopiaSL_38
Tar/Tork
GMR/Tork
Oryco/Ivana
Sire/Maximus
Ale/Retrofit
Bianca
Reference genomes
● Nicotiana sylvestris (GCF_000393655.1)
● Nicotiana tomentosiformis (GCF_000390325.2)
● Nicotiana tabacum (GCF_000715135.1)
● Nicotiana attenuata (GCF_001879085.1)
● Capsicum annuum (GCF_000710875.1)
● Solanum lycopersicum (GCF_000188115.4)
● Solanum pennellii (GCF_001406875.1)
● Solanum tuberosum (GCF_000226075.1)
Number of hits
1
10
25
50
100
200
400

Table 4 Specificity of the probe to selected retroelement families. BLAST results of probe sequences against retroelements families identified in *C. annuum*

Probe specificity	Families	Max score	Query cover	E value	Ident
CopiaCa_01	CopiaCa_01	1779	99%	0.0	99%
	monotypic|Ch02d_66	1779	99%	0.0	99%
GypsyCa_16	GypsyCa_16	1725	99%	0.0	99%
	GypsyCa_17	1725	99%	0.0	99%
[Del/Tekay]-complex	GypsyCa_04	351	98%	2.00E-97	68%
	monotypic|Ch11a_28	233	97%	3.00E-62	66%
	GypsyCa_09	159	99%	6.00E-40	65%
	GypsyCa_13	132	95%	8.00E-32	63%
	GypsyCa_19	113	99	8.00E-26	64%
	monotypic|Ch05d_24	54	11%	7.00E-8	69%

Intra- and inter-specific radiation of retroelement populations in Solanaceae

In our study, we detected that autonomous Gypsy retroelements were ~2.4-fold greater and younger than Copia ones (Tables 1 and 3). The radiation of Gypsy and Copia and their respective lineages have been described as widely variable in the different plant genomes. Thus, in species such as *Vitis vinifera* (Jaillon et al. 2007), *Theobroma cacao* (Argout et al. 2011), and palm species (*Elaeis guineensis* and *E. oleífera*; Beulé et al. 2015), Copia retroelements are preponderant, while the opposite occurs in other species such as *Oryza sativa* (Vitte et al. 2007; Zhang and Gao 2017) and *Helianthus* sp. (Qiu and Ungerer 2018). In Solanaceae species, several studies have revealed that the expansion of Gypsy has been dominant on Copy with different waves of amplification (The Tomato Genome Consortium et al 2012; Bolger et al. 2014; Qin et al. 2014; Paz et al. 2017; Xu et al. 2017; Esposito et al. 2019; Gaiero et al. 2019).

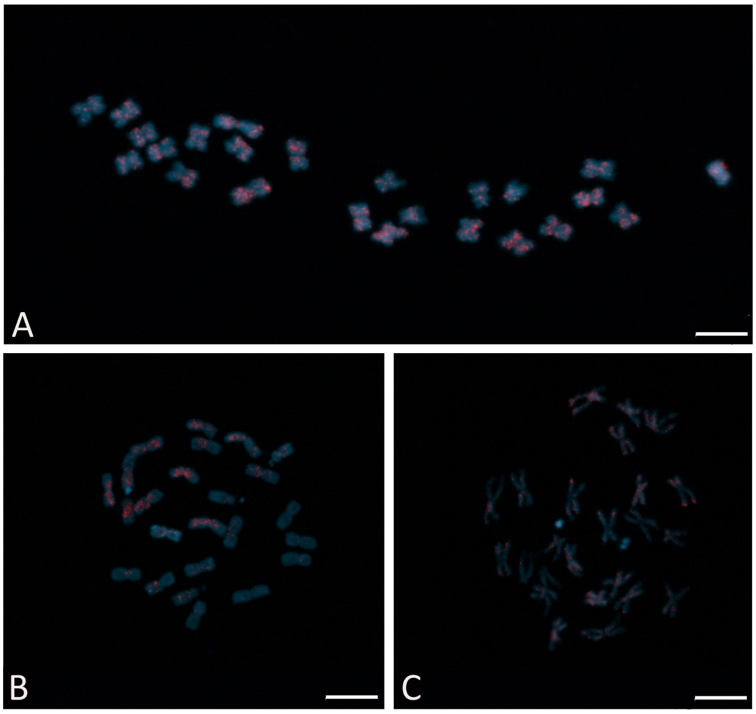


Fig. 4 Fluorescent in situ hybridization (FISH), in *C. annuum* metaphase chromosomes (Zunla). (A) *S-GypsyCa_16*; (B) *S-[Del/Tekay]-Complex*; and (C) *S-CopiaCa_01*. Chromosomes, light blue color, were stained with 4′,6-diamino-2-phenylindole (DAPI), while red fluorescence dots (signal) indicate hybridization of the different probes (built on the conserved RT and RH sequence of the families of the selected retroelements). Labeled with Digoxigenin (DIG) and detected with antibodies conjugated with tetramethylrhodamine isothiocyanate (TRITC). Scale 5 um

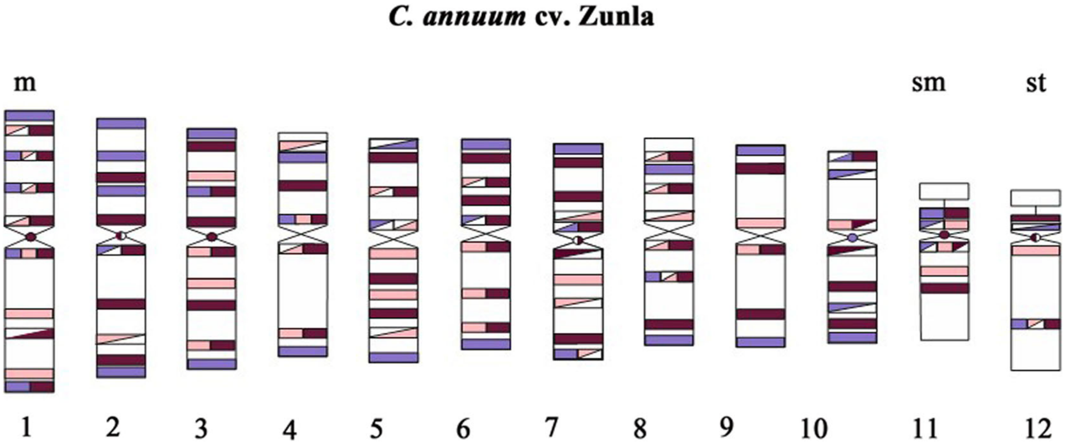
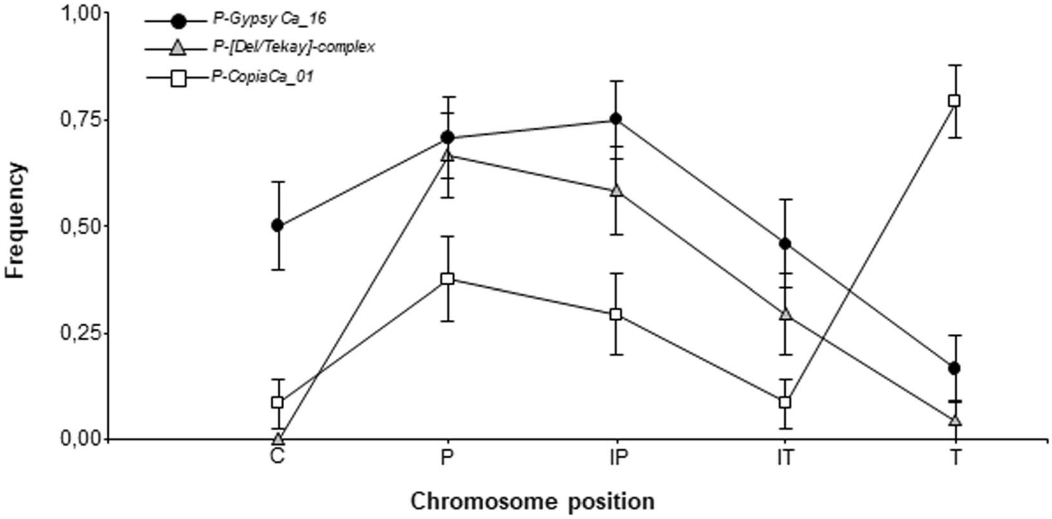


Fig. 5 The cytogenetic map was constructed based on the hybridization sites of the three retroelement probes for the cultivar of Zunla reference. Chromosomes: m, metacentric; sm, submetacentric; sb, subtelocentric. Probes: *S-GypsyCa_16* = violet black color, *S- S-Complex [Del/Tekay]* = pink color, and *S-CopiaCa_01* = violet color. The fully colored rectangles correspond to the detection of hybridization signals in both homologous chromosomes, while the rectangles that have a diagonal line correspond to the detection of the hybridization signal in one of the homologous chromosomes. Scale 5 μm

Although all the evolutionary lineages of LTR retroelements described in Angiosperms were identified in pepper, only three comprised 83% of the retroelements: Athila (60%; Gypsy), Ale/Retrofit (17%; Copia), and Del/Tekay (7%; Gypsy) (Table 3; Fig. 1). This result contrasted with the one observed in the genus Solanum, where a majority of Del/Tekay radiation was detected, with variations in the radiation



Probe	C	P	IP	IT	T
P-CopiaCa_01	0.08 ± 0.06 ^{B}	0.38 ± 0.10 ^{B}	0.29 ± 0.09 ^{B}	0.08 ± 0.06 ^{B}	0.79 ± 0.08 ^{A}
P-GypsyCa_16	0.50 ± 0.10 ^{A}	0.71 ± 0.09 ^{A}	0.75 ± 0.09 ^{A}	0.46 ± 0.10 ^{A}	0.17 ± 0.08 ^{B}
P-Del/Tekay Complex	0.00± 0.00 ^{B}	0.67 ± 0.10 ^{A}	0.58 ± 0.09 ^{A}	0.29 ± 0.09 ^{A,B}	0.04 ± 0.04 ^{B}
$F_{(3,69)}$	15.07 ***	3.43 *	5.83 **	4.58 *	32.41 ***

Fig. 6 Average values of hybridization signal intensity of the three retroelement probes, on different chromosomal regions of the reference cultivar Zunla. Abbreviations: C, centromeric; P, proximal; IP, interstitial-proximal; IT, interstitial-terminal; T, terminal. The asterisk indicates values of statistical significance: ***p < 0,0001; **p < 0.01, and *P < 0.05

of the other lineages depending on whether they were species related to potatoes or tomatoes (Fig. 3; Park et al. 2012; Paz et al. 2017; Esposito et al. 2019; Gaiero et al. 2019). In Nicotiana, important radiation of GMR/Tork lineages was observed (Fig. 3; Melayah et al. 2004; Petit et al. 2007). This would indicate that a differential evolutionary dynamic would be shaping the composition of retroelements in the genomes of this group of species.

These lineage-specific expansion phenomena are due to the massive retrotransposition of a few families of retroelements in the genomes in different plant species at different periods during their evolutionary process (Vicient et al. 2001; Baucom et al. 2009; Beulé et al. 2015; Paz et al. 2017; Zhang and Gao 2017). In our research, we delimited a total of 127 families of complete and intact LTR retroelements in the *C. annuum* genome, of which we were able to verify that only three represented 71% of the total population: GypsyCa_16 (Athila; 57.1%), CopiaCa_01 (Ale/Retrofit; 9.5%), and GypsyCa_13 (Del/Tekay; 4.2%) (Table 3). These highly radiated families in the Capsicum genome were not found in the genomes of other Solanaceae species, belonging to the group of exclusive retroelement families of pepper (78% of the total) (Fig. 3). In the same way, some families identified in Solanum and Nicotiana genera show a similar behavior, especially in Gypsy derivate families (Fig. 3). This retroelement family-specific behavior has been observed in other diploid plant species with a large genome size such as *Hordeum vulgare*—where a single family of retroelements, BARE-1, represents 10% of the genome of the species (Jääskeläinen et al. 2013) and in *Oryza australiensis*—where the amplification of only three families of retroelements produced doubled the size of their genome (Piegu et al. 2006). In the case of *S. lycopersicum*, the survey of the families of retroelements inhabiting its genome revealed that, although there was differential radiation from two families of retroelements derived from Del/Tekay and Tork/GRM (both exclusive of Solanum) and that together they comprised almost 50% of the total of the identified retroelements (Jingling/GypsySL_01 and Trok4/CopiaSl_37; Paz et al. 2017), this radiation was much lower than the one found in pepper (Fig. 3). This behavior could be related to the differential genome expansion between both species in a similar way to that observed in the *Oryza genus* (Piegu et al. 2006; Zhang and Gao 2017). In this regard, the comparison of the dynamics of retroelement populations in related diploid species with different genomic sizes such as the Oryza genus revealed the differential expansion of a few families of retroelements in the species of greater genome size (Zuccolo et al. 2007).

Currently, the best-studied retroelements in plants belong to the GMR/Tork lineage identified and isolated from different Solanaceae species. Interestingly, in *C. annuum*, this lineage was characterized by having a few sparsely populated families but with a high degree of homology and conservation with their relatives inhabiting other Solanaceae genomes that were not observed in the other lineages (Fig. 3). In the case of Tnt1, this retroelement constitutes one of the most abundant families of retroelements in the *N. tabacum* genome and its radiation has been extensively studied in the genomes of *Nicotiana* spp. (Melayah et al. 2004), *Solanum* spp. (Manetti et al. 2009; Paz et al. 2017; Tam et al. 2005), and *Petunia* spp. (Kriedt et al. 2014). In pepper, a single derived family was found with a very low number of copies (CopiaCa_27). This family is ancestral and could be found before *Nicotiana* radiation (occurred 23 Mya ago, Xu et al. 2017). Despite having a very low copy number in *Capsicum* spp. and *Solanum* spp., it is still present in these genomes with a high degree of homology (Fig. 3; Fig. S2; Melayah et al. 2004; Paz et al. 2015, 2017; Tam et al. 2005, 2009). Another similar example is the case of T135/CopiaSl_33, originally identified and isolated from *S. lycopersicum* (Tam et al. 2009), but which is kept in a perfect state of conservation in *C. annuum* (CopiaCa_04; Fig. S2), presenting the unique feature in this study of having a high degree of coverage and identity with its tomato counterpart even at the level of the LTR sequence (results not shown). Although the retroelements derived from Tnt1 and T135/CopiaSL_33 in pepper have a low copy number, their high degree of homology allowed the application of heterologous primers as highly informative genetic markers for phylogenetic inferences in *Capsicum* spp. (Tam et al. 2005). Other well-known and studied families from Tork/GMR lineage in Solanaceae species did not have the same success in the *C. annuum* genome; this is the case of Tto1 and Tork4/CopiaSL_37, the latter very widespread in the tomato genome (Fig. 3; Paz et al. 2017).

A notable feature in *C. annuum* is the fact that a large proportion of the retroelements that inhabit its genome have been inserted recently. In the case of members of the Copia superfamily, at least 35% of the total population did so less than 0.5 Mya, while for Gypsy this number is much higher, an 85% (Fig. 2). These amplification waves are mainly given by three families of retroelements derived from the lineages of Athila, Ale/Retrofit, and Del/Tekay. Evolutionarily, these three lineages are the most likely to radiate in Solanaceae genomes. However, Gypsy-derived lineages are generally more ancient, with little participation in events of recent radiation (Paz et al. 2017; Esposito et al. 2019; Gaiero et al. 2019).

Bridging bioinformatics data with cytogenetics

In our research, we were able to locate cytogenetically the most abundant families identified by bioinformatics tools in the *C. annum* genome. In this sense, it is important to highlight that the challenge of cytogenetically identifying retroelement families in a genome is arduous, not only because they are highly variable sequences, but also because they are dispersed in the genome. The FISH technique is a tool that allows detecting and locating a specific DNA sequence on a chromosome (Kato et al. 2004). The technique relies on exposing chromosomes to a small DNA sequence called a probe that has a fluorescent molecule attached to it. The visualization procedure is indirect, by the analyses of the fluorescent signal intensity. Thus, it is important to note that the technique is qualitative, not quantitative. That is why it is not possible to quantitatively correlate the hybridization signals with what was observed at the bioinformatic level.

In practice, under optimal hybridization and detection conditions, the sensitivity of the FISH technique depends on the accessibility of the probe to the homologous region on the DNA. In turn, this is determined by the degree of condensation of chromosomal DNA. In other words, the less condensed the chromosomes are, the less coiled the DNA molecule will be and, therefore, the accessibility of the probes to chromosomal DNA will be better (Van de Rijke et al. 2000). The degree of chromatin condensation varies substantially, not only between the different phases of cell division but also between the different types of configuration adopted by chromosomal DNA. In this work, we employed metaphase mitotic chromosomes. In this kind of sample, the spatial resolution (minimum physical distance at which two adjacent sequences can be identified under a fluorescence microscope; De Jong et al. 1999) is 5–10 Mb and the sensitivity (minimum size of one DNA sequence that can be unambiguously detected under the microscope; De Jong et al. 1999) is 10 kb (Valárik et al. 2004). Besides, the spatial resolution depends on how the chromosomal material has been previously treated and spread or stretched on the microscope slide, producing some decrease in the hybridization signal (De Jong et al. 1999; Valárik et al. 2004). A decrease in the hybridization signal has been observed in other cultivated Solanaceae genomes (Braz et al. 2018).

Another important factor that defines the success of the technique is the number of Diana sequences. Thus, the more the number of Diana sequences are present in the genome, the more intense is the fluorescent signals found. That is why, to achieve hybridization signals, a large number of copies is required to visualize a chromosomal region, whether it be of short, highly repeated, or long DNA sequences (Boyle et al. 2011; Yamada et al. 2011; Beliveau et al. 2012). Background describes that retroelements are integrated into regions of the genome and can show site-specific preferences (Gao et al. 2008; Baucom et al. 2009; Nellåker et al. 2012), forming groups. The Bare-1 element has been observed to be found in a nested form in the barley genome (Shirasu et al. 2000), a characteristic also observed in the genomes of the Hordeum and Triticeae genera (Vicient et al. 1999; Gribbon et al. 1999) and the families analyzed in this research. This type of insertion would favor detection by FISH.

Finally, our FISH results agree with our bioinformatic family-abundance analysis, whereas *P-GypsyCa_16* showed a higher number of hybridization signals than *P-CopiaCa_01* and P-[Del/Tekay]-complex, both with similar values (Fig. 4). GypsyCa_16 was the most abundant family observed in family-abundance analyses, followed by CopiaCa_01 and Del/Tekay lineages, both with a similar retroelement number (Table 3). In this way, our FISH analysis not only validates the results obtained at the bioinformatic and taxonomic level but also provides information about the distribution of these families/lineages along the chromosomes of *C. annuum.*

Different retroelement lineages, different affinity to chromatin

The genomic environment is highly heterogeneous and zoned. This characteristic is defined at different levels: (i) the complexity of the DNA sequence (coding or noncoding); (ii) the spatial configuration that this sequence adopts in space; (iii) the epigenetic setting; (iv) association with other molecules (Jarillo et al. 2009). Experimental evidence suggests that the different retroelement lineages have an affinity for different types of heterochromatin as a strategy to evade the genome's activity regulation mechanisms of silencing and/or non-homologous recombination. In our work, we identified that the three evaluated probes hybridized in all the *C. annuum* chromosomes and that they also presented differential affinity towards the different chromosomal regions.

In the case of the *P-[Del/Tekay]-complex* lineage-specific probe, its predominant presence is towards the proximal, interstitial-proximal, and interstitial-proximal regions and with little or no presence in the centromeres and telomeres (Fig. 5). This lineage is characterized by presenting an additional Chromo domain that confers an affinity towards heterochromatin (Neumann et al. 2011). In this sense, in different plant species, it has been demonstrated that families of retroelements belonging to the Del/Tekay lineage have an affinity towards heterochromatic regions but with reduced hybridization towards centromeric regions, secondary constrictions, and major heterochromatic blocks (Wang et al. 2006; Neumann et al. 2011; Park et al. 2011; Domingues et al. 2012; Weber et al. 2013; Yang et al. 2020). In the specific case of Solanaceae, this lineage is quite ancient and has exhibited different waves of amplification before, during, and after speciation between tomato and pepper (Park et al. 2011; Paz et al. 2017). Its insertion in *C. annuum* has been associated with the heterochromatinization of euchromatic regions (Park et al. 2012) and may affect the expression of neighboring genes.

The family-specific probe derived from the Athila lineage, *P-GypsyCa_16*, exhibited hybridization signals in all chromosomal regions, with a preponderance towards the proximal and interstitial-proximal regions, and to a lesser extent in the centromeres and the interstitial-terminal region. This behavior was described for the Athila lineage in other plant species (De Souza et al. 2018; Li et al. 2019). In the case of *Capsicum*, the distribution of potentially autonomous retroelement lineages has recently been described in different species of the genus, including *C. annuum* (De Assis et al. 2020). In this study, the Athila distribution shows a trend towards the interstitial chromosomal regions. These differences with our results could be attributed to different criteria for choosing the retroelement, probe design, and/or particularities of plant material. Likewise, another study revealed an accumulation of Athila lineage in the pericentromeric to interstitial regions of all *C. annuum* chromosomes with a marked affinity for regions rich in genes (Park et al. 2011).

Concerning the CopiaCa_01 family, this work presents a majority distribution pattern towards terminal regions and little or no signal in the interstitial-terminal and centromeric regions in most of the pepper chromosomes. This lineage-specific preference towards telomeres has also been observed in *Erianthus arundinaceus* (Huang et al. 2017) and *Allium cepa* (Pearce et al. 1996). However, other works report a more heterogeneous distribution (Li et al. 2019; Yang et al. 2020) even in *C. annuum* (Park et al. 2011). Various studies have suggested that there is an association between transposable elements and rDNA, affecting their distribution, abundance, and expression (Dubcovsky and Dvorák 1995; Raskina et al. 2004; Datson and Murray 2006). This association has evolutionary implications for plant genomes. The presence of CopiaCa_01 has been associated with polymorphisms of rDNA sites in *C. annuum* (unpublished results, Yañez Santos, AM; Paz RC; Urdampilleta JD). Despite being preliminary, these results could suggest that the activity of CopiaCa_01 could be related to the generation of this type of polymorphism through the generation of non-homologous recombination sites. However, further studies are necessary to obtain more conclusive data in this regard.

Conclusions

In this work, we demonstrate that there are a large number of families of autonomous LTR retroelements that have been inserted in the last 6 Mya in the *C. annuum* genome. All the LTR retroelement lineages described in plants are present in pepper. While the lineage families derived from Del/Tekay (closely associated with speciation events in Solanaceae) exhibited different waves of amplification in this period, two families derived from Athila and Ale/Retrofit have experienced a significant wave of amplification in the last 0.5 Mya. The FISH analysis of the insertion preferences of the majority elements identified in this work revealed significant differences: (i) GypsyCa_16 exhibited a wide insertion profile with a preponderance of signals from the centromere towards the interstitial-proximal region; (ii) CopiaCa_01 exhibited a marked insertion preference towards telomeres; (iii) the Del/Tekay lineage was limited to the proximal to interstitial-terminal regions, with little or no presence in telomeres and centromeres. Knowing these particularities within a species may be of interest in the development of molecular markers since insertional polymorphisms can be detected in different genomic regions within the same species.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10577-021-09663-4.

Acknowledgements The authors acknowledge Dr. Maria Virginia Sanchez-Puerta for her valuable help with phylogenetic tree construction, Dr. Carlos Llorens for his support in retroelements classification, Dr. Thomas Wicker for providing retroelement reference sequences, and to Dr. Alejandra Trenchi for sharing her experience during experiments FISH. We also appreciate the valuable language assistance of Lic. Vanesa Heredia and Dr. Hernan G. Rosli. The authors are also grateful for anonymous reviewers' and the editor Dr. Jiming Jiang's comments which improved substantially this work. This work was part of the PhD thesis of YSA, which benefited from AGENCIA and CONICET fellowship.

Funding This work was funded by Agencia Nacional de Promoción Científica y Tecnológica, Argentina (AGENCIA), D-TEC 0008/13; PROJOVI 2015 and 2018 (Universidad Nacional de San Juan); and FONCyT, PICT-2018-01691 (AGENCIA).

Abbreviations *C,* centromeric chromosome portion; *CTAB,* cetyl trimethyl ammonium bromide; *FISH,* fluorescent in situ hybridization; *IP,* interstitial-proximal chromosome portion; *IT,* interstitial-terminal chromosome portion; *LTR,* long terminal repeats; *Mya,* millions of years ago; *P,* proximal chromosome portion; *PCR,* polymerase chain reaction; *RH,* RNase-H; *RT,* retrotranscriptase; *T,* terminal chromosome portion

References

Anderson, SN, Stitzer, MC, Brohammer, AB, Zhou, P, Noshay, JM, O'Connor, CH, … Springer, NM (2019) Transposable elements contribute to dynamic genome content in maize. Plant J, 100(5), 1052–1065. https://doi.org/10.1111/tpj.14489

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J et al (2011) The genome of Theobroma cacao. Nat Genet 43(2): 101–108. https://doi.org/10.1038/ng.736

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, … Bennetzen JL (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genetics 5(11). https://doi.org/10.1371/journal.pgen.1000732

Beliveau, BJ, Joyce, EF, Apostolopoulos, N, Yilmaz, F, Fonseka, CY, McCole, RB, ... & Wu, CT (2012) Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc Natl Acad Sci, 109(52), 21301–21306. https://doi.org/10.1073/pnas.1213818110

Belyayev A (2014) Bursts of transposable elements as an evolutionary driving force. J Evol Biol 27(12):2573–2584. https://doi.org/10.1111/jeb.12513

Benachenhou F, Sperber GO, Bongcam-Rudloff E, Andersson G, Boeke JD, Blomberg J (2013) Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). Mob DNA 4(1):1–16. https://doi.org/10.1186/1759-8753-4-5

Bennetzen JL (1996) The Mutator transposable element family of maize. Genet Eng 13:1–37. https://doi.org/10.1016/0966-842X(96)10042-1

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42(1):251–269. https://doi.org/10.1023/A:1006344508454

Bennetzen JL (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115(1):29–36. https://doi.org/10.1023/A:1016015913350

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol 65(1):505–530. https://doi.org/10.1146/annurev-arplant-050213-035811

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

Biémont C, Vieira C (2006) Genetics: junk DNA as an evolutionary force. Nature 443(7111):521–524. https://doi.org/10.1038/443521a

Bolger AM, Lohse M, Usadel B (2014) Genome analysis Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bourgeois Y, Boissinot S (2019) On the population dynamics of junk: a review on the population genomics of transposable elements. Genes. 10:419. https://doi.org/10.3390/genes10060419

Bowen NJ, McDonald JF (2001) Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res* 11(9):1527–1540. https://doi.org/10.1101/gr.164201

Boyle S, Rodesch MJ, Halvensleben HA, Jeddeloh JA, Bickmore WA (2011) Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosom Res 19(7):901–909. https://doi.org/10.1007/s10577-011-9245-0

Braz GT, He L, Zhao H, Zhang T, Semrau K, Rouillard JM et al (2018) Comparative oligo-FISH mapping: an efficient and powerful methodology to reveal karyotypic and chromosomal evolution. Genetics 208(2):513–523. https://doi.org/10.1534/genetics.117.300344

Cavrak VV, Lettner N, Jamge S, Kosarewicz A, Bayer LM, Scheid OM (2014) How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet 10(1): e1004115. https://doi.org/10.1371/journal.pgen.1004115

Datson PM, Murray BG (2006) Ribosomal DNA locus evolution in Nemesia: transposition rather than structural rearrangement as the key mechanism? Chromosome. https://doi.org/10.1007/s10577-006-1092-z

De Assis R, Baba VY, Cintra LA, Goncalves LSA, Rodrigues R, Vanzela ALL (2020) Genome relationships and ltr-retrotransposon diversity in three cultivated Capsicum L. (Solanaceae) species. *BMC Genomics 21*(1):237. https://doi.org/10.1186/s12864-020-6618-9

De Jong JH, Fransz P, Zabel P (1999) High resolution FISH in plants–techniques and applications. Trends Plant Sci 4(7): 258–263. https://doi.org/10.1016/S1360-1385(99)01436-3

De Souza TB, Chaluvadi SR, Johnen L, Marques A, González-Elizondo MS, Bennetzen JL, Vanzela ALL (2018) Analysis of retrotransposon abundance, diversity and distribution in holocentric *Eleocharis* (Cyperaceae) genomes. Ann Bot 122(2):279–290. https://doi.org/10.1093/aob/mcy066

Devos KM, Brown JK, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* 12(7):1075–1079. https://doi.org/10.1101/gr.132102

Domingues DS, Cruz GMQ, Metcalfe CJ, Nogueira FTS, Vicentini R, …, Van Sluys, MA (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics, 13(1), 137. https://doi.org/10.1186/1471-2164-13-137

Du, J, Tian, Z, Hans, CS, Laten, HM, Cannon, SB, Jackson, SA., … Ma, J (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J, 63(4), 584–598. https://doi.org/10.1111/j.1365-313X.2010.04263.x

Dubcovsky J, Dvorák J (1995) Ribosomal RNA multigene loci: nomads of the Triticeae genomes. Genetics 140(4):367–377

Esposito S, Barteri F, Casacuberta J, Mirouze M, Carputo D, Aversano R (2019) LTR-TEs abundance, timing and mobility in *Solanum commersonii* and *S. tuberosum* genomes following cold-stress conditions. Planta 1:1781–1787. https://doi.org/10.1007/s00425-019-03283-3

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41(1):331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448

Gaiero P, Vaio M, Peters SA, Eric Schranz M, de Jong H, Speranza PR (2019) Comparative analysis of repetitive sequences among species from the potato and the tomato clades. Ann Bot 123:521–532. https://doi.org/10.1093/aob/mcy186

Galindo-González L, Mhiri C, Deyholos MK, Grandbastien MA (2017) LTR-retrotransposons in plants: engines of evolution. Gene 626:14–25. https://doi.org/10.1016/j.gene.2017.04.051

Gao X, Hou Y, Ebina H, Levin HL, Voytas DF (2008) Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res 18(3):359–369. https://doi.org/10.1101/gr.7146408

Gouy M, Guindon S, Gascuel O (2010) SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27(2): 221–224. https://doi.org/10.1093/molbev/msp259

Grandbastien MA (1998) Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3(5):181–187. https://doi.org/10.1016/S1360-1385(98)01232-1

Gribbon, BM, Pearce, SR, Kalendar, R, Schulman, AH, Paulin, L, Jack, P, ... & Flavell, AJ (1999) Phylogeny and transpositional activity of Ty1-copia group retrotransposons in cereal genomes. Mol Gen Genet MGG*, *261*(6), *883–891. https://doi.org/10.1007/PL00008635

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

Hirochika H, Hirochika R (1993) Tyl-copia group retrotransposons as ubiquitous components of plant genomes. Jpn J Genet. https://doi.org/10.1266/jjg.68.35

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci 93(15):7783–7788. https://doi.org/10.1073/pnas.93.15.7783

Huang Y, Luo L, Hu X, Yu F, Yang Y, Deng Z et al (2017) Characterization, genomic organization, abundance, and chromosomal distribution of Ty1-copia retrotransposons in *Erianthus arundinaceus*. Front Plant Sci 8:924. https://doi.org/10.3389/fpls.2017.00924

Jarillo, J. A., Pineiro, M., Cubas, P., Martinez-Zapater, J. M., Jarillo, J. A., Pineiro, M., … Martinez-Zapater, J. M. (2009). Chromatin remodeling in plant development. *Int J Dev Biol* 53(8-9–10), 1581–1596. https://doi.org/10.1387/ijdb.072460jj

Jääskeläinen M, Chang W, Moisy C, Schulman AH (2013) Retrotransposon BARE displays strong tissue-specific differences in expression. New Phytol 200(4):1000–1008. https://doi.org/10.1111/nph.12470

Jaillon, O, Aury, JM, Noel, B, Policriti, A, Clepet, C, Casagrande, A, … Wincker, P (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449(7161), 463–467. https://doi.org/10.1038/nature06148

Jiang, N, Bao, Z, Temnykh, S, Cheng, Z, Jiang, J, Wing, RA, ... & Wessler, SR (2002a) Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. Genetics, 161(3), 1293–1305. https://doi.org/10.1093/genetics/161.3.1293

Jiang N, Jordan IK, Wessler SR (2002b) Dasheng and RIRE2. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. Plant Physiol 130(4): 1697–1705. https://doi.org/10.1104/pp.015412

Jiang SY, Ramachandran S (2013) Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. PLoS One 8(7):e71118. https://doi.org/10.1371/journal.pone.0071118

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166(3):1437–1450. https://doi.org/10.1534/genetics.166.3.1437

Kato A, Lamb JC, Birchler JA (2004) Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc Natl Acad Sci U S A 101:13554–13559. https://doi.org/10.1073/pnas.0403659101

Kejnovsky E, Kubat Z, Macas J, Hobza R, Mracek J, Vyskot B (2006) Retand: a novel family of gypsy-like retrotransposons harboring an amplified tandem repeat. Mol Gen Genomics

276(3):254–263. https://doi.org/10.1007/s00438-006-0140-x

Kijima TE, Innan H (2009) On the estimation of the insertion time of LTR retrotransposable elements. Mol Biol Evol 27(4): 896–904. https://doi.org/10.1093/molbev/msp295

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33(1):479–532. https://doi.org/10.1146/annurev.genet.33.1.479

Kriedt RA, Cruz GMQ, Bonatto SL, Freitas LB (2014) Novel transposable elements in Solanaceae: evolutionary relationships among Tnt1-related sequences in wild *Petunia* species. Plant Mol Biol Report 32(1):142–152. https://doi.org/10.1007/s11105-013-0626-8

Levan A, Fredga K, Sandberg AA (1964) Nomenclature for centromeric position on chromosomes. Hereditas 52(2): 201–220. https://doi.org/10.1111/j.1601-5223.1964.tb01953.x

Li, SF, Guo, YJ, Li, JR, Zhang, DX, Wang, B X, Li, N, …, Gao, WJ (2019) The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (*Spinacia oleracea* L.). *Mobile DNA*, 10(1), 3. https://doi.org/10.1186/s13100-019-0147-6

Llorens C, Muñoz-Pomer A, Bernad L, Botella H, Moya A (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41. https://doi.org/10.1186/1745-6150-4-41

Llorens, C, Futami, R, Covelli, L, Domínguez-Escribá, L, Viu, JM, Tamarit, D, … Moya, A (2011) The Gypsy Database (GyDB) of Mobile Genetic Elements: Release 2.0. Nucleic Acids Res, 39(SUPPL. 1), 38–46. https://doi.org/10.1093/nar/gkq1061

Ma J, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14(5):860–869. https://doi.org/10.1101/gr.1466204

Manetti ME, Rossi M, Nakabashi M, Grandbastien MA, Van Sluys MA (2009) The Tnt1 family member Retrosol copy number and structure disclose retrotransposon diversification in different *Solanum* species. Mol Gen Genomics 281(3): 261–271. https://doi.org/10.1007/s00438-008-0408-4

Marcon HS, Domingues DS, Silva JC, Borges RJ, Matioli FF, de Mattos Fontes MR, Marino CL (2015) Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic. BMC Plant Biol 15(1):1–16. https://doi.org/10.1186/s12870-015-0550-1

Maupetit-Mehouas S, Vaury C (2020) Transposon reactivation in the germline may be useful for both transposons and their host genomes. Cells. 9(5):1172. https://doi.org/10.3390/cells9051172

Melayah, D, Lim, KY, Bonnivard, E, Chalhoub, B, Dorlhac De Borne, F, Mhiri, C, … Grandbastien, MA (2004) Distribution of the Tnt1 retrotransposon family in the amphidiploid tobacco (*Nicotiana tabacum*) and its wild *Nicotiana relatives*. Biol J Linn Soc, 82(4), 639–649. https://doi.org/10.1111/j.1095-8312.2004.00348.x

Myers T (2001) Rice genome consortium will finish ahead of schedule. Nature 409:752. https://doi.org/10.1038/35057489

Naito, K, Zhang, F, Tsukiyama, T, Saito, H, Hancock, CN, Richardson, AO, … Wessler, SR (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature, 461(7267), 1130–1134. https://doi.org/10.1038/nature08479

Nellåker, C, Keane, T M, Yalcin, B, Wong, K, Agam, A, Belgard, TG, … Ponting, CP (2012) The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol, 13(6). https://doi.org/10.1186/gb-2012-13-6-r45

Neumann, P, Navrátilová, A, Koblížková, A, Kejnovsk, E, Hřibová, E, Hobza, R, … MacAs, J (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA, 2(1), 1–16. https://doi.org/10.1186/1759-8753-2-4

Neumann P, Novák P, Hoštáková N, Macas J (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA 10(1):1–17. https://doi.org/10.1186/s13100-018-0144-1

Okonechnikov K, Golosova O, Fursov M, Team U (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28(8):1166–1167. https://doi.org/10.1093/bioinformatics/bts091

Park, M, Jo, SH, Kwon, JK, Park, J, Ahn, JH, Kim, S, … Choi, D (2011) Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. BMC Genomics, 12(1), 85. https://doi.org/10.1186/1471-2164-12-85

Park, M, Park, J, Kim, S, Kwon, JK, Park, HM, Bae, IH, … Choi, D (2012) Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. Plant J, 69(6), 1018–1029. https://doi.org/10.1111/j.1365-313X.2011.04851.x

Paz RC, Rendina González AP, Ferrer MS, Masuelli RW (2015) Short-term hybridisation activates Tnt1 and Tto1 Copia retrotransposons in wild tuber-bearing *Solanum* species. Plant Biol 17(4):860–869. https://doi.org/10.1111/plb.12301

Paz RC, Kozaczek ME, Rosli HG, Andino NP, Sanchez-Puerta MV (2017) Diversity, distribution and dynamics of full-length Copia and Gypsy LTR retroelements in *Solanum lycopersicum*. Genetica 145(4–5):417–430. https://doi.org/10.1007/s10709-017-9977-7

Pearce SR, Pich U, Harrison G, Flavell AJ, Heslop-Harrison JS, Schubert I, Kumar A (1996) The Ty1-copia group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal heterochromatin. Chromosom Res 4(5):357–364. https://doi.org/10.1007/BF02257271

Petit, M, Lim, KY, Julio, E, Poncet, C, De Borne, FD, Kovarik, A, ... & Mhiri, C (2007) Differential impact of retrotransposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). Mol Gen Genomics, 278(1), 1–15. https://doi.org/10.1007/s00438-007-0226-0

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., … Panaud, O. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res, 21(7), 1201. https://doi.org/10.1101/gr.5290206.of

Qin, C, Yu, C, Shen, Y, Fang, X, Chen, L, Min, J & Cheng, J (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc Natl Acad Sci 111(14). https://doi.org/10.1073/pnas.1400975111

Qiu F, Ungerer MC (2018) Genomic abundance and transcriptional activity of diverse gypsy and copia long terminal repeat retrotransposons in three wild sunflower species. BMC Plant Biol 18(1):1–8. https://doi.org/10.1186/s12870-017-1223-z

Raskina O, Belyayev A, Nevo E (2004) Quantum speciation in Aegilops: molecular cytogenetic evidence from rDNA cluster variability in natural populations. Proc Natl Acad Sci U S A 101(41):14818–14823. https://doi.org/10.1073/pnas.0405817101

Raskina O, Barber JC, Nevo E, Belyayev A (2008) Repetitive DNA and chromosomal rearrangements: speciation-related events in plant genomes. Cytogen Genome Res 120(3–4): 351–357. https://doi.org/10.1159/000121084

Roa F, Guerra M (2012) Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. BMC Evol Biol 12(1):1. https://doi.org/10.1186/1471-2148-12-225

Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97(6):381–388. https://doi.org/10.1038/sj.hdy.6800903

SanMiguel, P, Tikhonov, A, Jin, YK, Motchoulskaia, N, Zakharov, D, Melake-Berhan, A, ... & Bennetzen, JL (1996) Nested retrotransposons in the intergenic regions of the maize genome. Science, 274(5288), 765–768

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20(1):43–45. https://doi.org/10.1038/1695

Schnable, PS, Pasternak, S, Liang, C, Zhang, J, Fulton, L, Graves, TA, … Kumari, S (2009) The B73 maize genome: complexity, diversity, and dynamics. Science, 326(June), 1112–1115. https://doi.org/10.1126/science.1178534

Schulman AH, Kalendar R (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cytogen Genome Res 110(1–4):598–605. https://doi.org/10.1159/000084993

Schwarzacher T, Heslop-Harrison P (2000) Practical in situ hybridization. BIOS Scientific Publishers, Oxford. https://doi.org/10.1023/A:1026756103545

Sharma A, Schneider KL, Presting GG (2008) Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations. Proc Natl Acad Sci U S A 105(40):15470–15474. https://doi.org/10.1073/pnas.0805694105

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res 10(7):908–915. https://doi.org/10.1101/gr.10.7.908

Tam SM, Mhiri C, Vogelaar A, Kerkveld M, Pearce SR, Grandbastien MA (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. Theor Appl Genet 110(5):819–831. https://doi.org/10.1007/s00122-004-1837-z

Tam SM, Lefebvre V, Palloix A, Sage-Palloix AM, Mhiri C, Grandbastien MA (2009) LTR-retrotransposons Tnt1 and T135 markers reveal genetic diversity and evolutionary relationships of domesticated peppers. Theor Appl Genet 119(6):973–989. https://doi.org/10.1007/s00122-009-1102-6

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24(8):1596–1599. https://doi.org/10.1093/molbev/msm092

Tomato Genome Consortium, T, DNA Research Institute, K, Sciences, L, company, R., Express Inc, A., Academy of Agriculture, B, …, SrL, Y (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Tomato Genome Consortium https://doi.org/10.1038/nature11119

Valárik M, Bartoš J, Kovářová P, Kubaláková M, De Jong JH, Doležel J (2004) High-resolution FISH on super-stretched flow-sorted plant chromosomes. Plant J 37(6):940–950. https://doi.org/10.1111/j.1365-313X.2003.02010.x

Van de Rijke FM, Florijn RJ, Tanke HJ, Raap K (2000) DNA fiber-FISH staining mechanism. J Histochem Cytochem 48(6):743–745. https://doi.org/10.1177/002215540004800602

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. Plant Cell 11(9):1769–1784. https://doi.org/10.1105/tpc.11.9.1769

Vicient CM, Jääskeläinen MJ, Kalendar R, Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125(3):1283–1292. https://doi.org/10.1104/pp.125.3.1283

Vicient CM (2010) Transcriptional activity of transposable elements in maize. BMC Genomics 11(1). https://doi.org/10.1186/1471-2164-11-601

Vitte C, Panaud O, Quesneville H (2007) LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. BMC Genomics 8:17–21. https://doi.org/10.1186/1471-2164-8-218

Wang, W, Zheng, H, Fan, C, Li, J, Shi, J, Cai, Z, … Wang, J (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell, 18(8), 1791–1802. https://doi.org/10.1105/tpc.106.041905

Weber, B, Heitkam, T, Holtgräwe, D, Weisshaar, B, Minoche, AE, Dohm, JC, … Schmidt, T (2013) Highly diverse chromoviruses of *Beta vulgaris* are classified by chromodomains and chromosomal integration. Mob DNA, 4(1), 1–16. https://doi.org/10.1186/1759-8753-4-8

Wicker, T & Keller, B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient.pdf, 1072–1081. https://doi.org/10.1101/gr.6214107

Weising K, Nybom H, Wolff K, Kahl G (2005) DNA fingerprinting in plants, 2nd end: principles, methods, and applications. Ann Bot 97(3):476–477. https://doi.org/10.1093/aob/mcj057

Xu Z, Wang H (2007) LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35(SUPPL.2):265–268. https://doi.org/10.1093/nar/gkm286

Xu Y, Du J (2014) Young but not relatively old retrotransposons are preferentially located in gene-rich euchromatic regions in tomato (*Solanum lycopersicum*) plants. Plant J 80(4):582–591. https://doi.org/10.1111/tpj.12656

Xu, Z, Liu, J, Ni, W, Peng, Z, Guo, Y, Ye, W, … Du, J (2017) GrTEdb: the first web-based database of transposable elements in cotton (*Gossypium raimondii*). Database, 2017(1), 1–7. https://doi.org/10.1093/database/bax013

Yadav CB, Bonthala VS, Muthamilarasan M, Pandey G, Khan Y, Prasad M (2015) Genome-wide development of transposable elements-based markers in foxtail millet and construction of an integrated database. DNA Res 22(1):79–90. https://doi.org/10.1093/dnares/dsu039

Yamada, NA, Rector, LS, Tsang, P, Carr, E, Scheffer, A, Sederberg, MC, ... & Brothman, AR (2011) Visualization of fine-scale genomic structure by oligonucleotide-based high-resolution FISH. Cytogen Genome Res 132(4), 248–254. https://doi.org/10.1159/000322717

Yang, S., Zeng, K., Chen, K., Zhao, X., Wu, J., Huang, Y., … Deng, Z. (2020). Sequence evolution, abundance, and chromosomal distribution of Ty1-copia retrotransposons in the *Saccharum spontaneum* genome. Cytogen Genome Res 160(5), 272–282. https://doi.org/10.1159/000506222

Zhang QJ, Gao LZ (2017) Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA- genome *Oryza* species. *Genes|Genomes|Gen*. https://doi.org/10.1534/g3.116.037572

Zhao M, Ma J (2013) Co-evolution of plant LTR-retrotransposons and their host genomes. Protein Cell 4(7):493–501. https://doi.org/10.1007/s13238-013-3037-6

Zuccolo, A, Sebastian, A, Talag, J, Yu, Y, Kim, HR, Collura, K, … Wing, RA (2007) Transposable element distribution, abundance and role in genome size variation in the genus Oryza. BMC Evol Biol, 7, 1–15. https://doi.org/10.1186/1471-2148-7-152