Genetica
DOI 10.1007/s10709-017-9977-7

ORIGINAL PAPER

Diversity, distribution and dynamics of full-length Copia and Gypsy LTR retroelements in *Solanum lycopersicum*

Rosalía Cristina Paz[1] · Melisa Eliana Kozaczek[2] · Hernán Guillermo Rosli[3] · Natalia Pilar Andino[4] · Maria Virginia Sanchez-Puerta[5]

Received: 5 April 2017 / Accepted: 26 July 2017

Abstract Transposable elements are the most abundant components of plant genomes and can dramatically induce genetic changes and impact genome evolution. In the recently sequenced genome of tomato (*Solanum lycopersicum*), the estimated fraction of elements corresponding to retrotransposons is nearly 62%. Given that tomato is one of the most important vegetable crop cultivated and consumed worldwide, understanding retrotransposon dynamics can provide insight into its evolution and domestication processes. In this study, we performed a genome-wide *in silico* search of full-length LTR retroelements in the tomato nuclear genome and annotated 736 full-length Gypsy and Copia retroelements. The dispersion level across the 12 chromosomes, the diversity and tissue-specific expression of those elements were estimated. Phylogenetic analysis based on the retrotranscriptase region revealed the presence of 12 major lineages of LTR retroelements in the tomato genome. We identified 97 families, of which 77 and 20 belong to the superfamilies Copia and Gypsy, respectively. Each retroelement family was characterized according to their element size, relative frequencies and insertion time. These analyses represent a valuable resource for comparative genomics within the Solanaceae, transposon-tagging and for the design of cultivar-specific molecular markers in tomato.

Keywords *Solanum lycopersicum* · Full length LTR retroelements · Family delimitation · Phylogeny · Insertion time · Expression

Introduction

Autonomous long terminal repeat (LTR) retroelements are mobile genetic entities ranging in size from 3500 to 22,000 bp that multiply via RNA intermediates and inhabit eukaryotic genomes (Kumar and Bennetzen 1999). They

Electronic supplementary material The online version of this article (doi:10.1007/s10709-017-9977-7) contains supplementary material, which is available to authorized users.

✉ Rosalía Cristina Paz
rosaliapaz@gmail.com

Melisa Eliana Kozaczek
kmelisaeliana@hotmail.com

Hernán Guillermo Rosli
hrosli@agro.unlp.edu.ar

Natalia Pilar Andino
nandino@unsj-cuim.edu.ar

Maria Virginia Sanchez-Puerta
sanchezpuerta@gmail.com

[1] CIGEOBIO (FCEFyN, UNSJ/CONICET), Av. Ignacio de la Roza 590 (Oeste), J5402DCS, Rivadavia, San Juan, Argentina

[2] Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones, Posadas, Misiones, Argentina

[3] Instituto de Fisiología Vegetal, INFIVE, Universidad Nacional de La Plata, CONICET, La Plata, Buenos Aires, Argentina

[4] Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, San Juan, Argentina

[5] IBAM, Universidad Nacional de Cuyo, CONICET, FCA and FCEN, Almirante Brown 500, M5528AHB Chacras de Coria, Argentina

Published online: 03 August 2017

are present in high copy number in most plant genomes, encompassing more than 75% of the nuclear genome of some species (Sanmiguel and Bennetzen 1998; Kumar and Bennetzen 1999; Li et al. 2004). Each LTR retroelement is a functional unit composed of two modules, structurally and functionally different: (a) two LTRs carrying regulatory sequences and flanking the coding region; and (b) an internal portion encoding *gag* and *pol* genes required to complete the retrotransposition. The gene *gag* encodes a structural core protein named GAG, whereas the gene *pol* encodes a polyprotein that includes a Proteinase (AP), an RNAse H (RH), a Retrotranscriptase (RT), and an Integrase (IN) (Kumar and Bennetzen 1999; Havecker et al. 2004). LTR retroelement copies encoding the entire repertoire of proteins are denominated "full- length" (Domingues et al. 2012; Gao et al. 2014), being potentially autonomous and able to retrotranspose and insert themselves in new locations in the genome.

In response to internal and/or environmental signals, regulatory sequences in the 5´LTR prompt the transcription of the internal portion and specific sequences of both LTRs into polycistronic mRNA molecules that are processed and exported to the cytoplasm to be translated similarly to other transcripts. In the cytoplasm, cumulus of GAG, GAG-Pol and mRNAs are formed and assembled into Virus-like Particles, which undergo a maturation process where polycistronic proteins are cleaved and activated by the AP and mRNAs are retrotranscribed into cDNAs by the RT. Then, the IN recognizes the LTR extremes in the mRNA and conforms a Pre-integration Complex that is released to the cytoplasm and imported into the nucleus, where it is integrated into a new location in the host genome (Voytas and Boeke 2002; Wilhelm and Wilhelm 2001). This process completes the intracellular life cycle of an LTR retroelement creating a new copy of itself. At insertion time, a retrotransposon is assumed to share high level of similarity with their parental copy. Older insertions might accumulate mutations, undergo homologous recombination, lose their function, and end up as a genomic fossil. The increasing number of sequenced eukaryotic genomes in recent decades revealed a great complexity of LTR retroelement populations and a close relationship with their host genome.

A considerable effort was made over the last years to classify transposable elements. One of the most accepted methods was the hierarchical classification system that subdivided them into Classes, Orders, Superfamilies and Families (Wicker and Keller 2007), although the delimitation of families has been questioned (El Baidouri and Panaud 2013). LTR retroelements belong to Class I, Order LTR retroelements and encompass two evolutionary distinct Superfamilies based on protein sequence and order: Gypsy and Copia (Wicker and Keller 2007). In addition, an intermediate taxonomic category between Superfamily and Family, termed "lineages", was proposed to include those families that were evolutionary related and share structural and functional features. In plants, six lineages (Athila, Tat, Galadriel, Reina, CRM/CR, and Del/Tekay) were identified in the Superfamily Gypsy; and seven lineages (TAR/Tork, Angela/Tork, GMR/Tork, Maximus/Sire, Ivana/Oryco, Ale/Retrofit, and Bianca) in the Superfamily Copia (Wicker and Keller 2007; Du et al. 2010; Llorens et al. 2011). This classification system was successfully used in retroelement classification of several plant species (Wicker and Keller 2007; Du et al. 2010; Llorens et al. 2011; Domingues et al. 2012; Xu et al. 2017).

Nuclear genomes of different organisms contain diverse numbers of transposable element families and copy numbers per family (Hua-Van et al. 2011; Biémont and Vieira 2006; Feschotte and Pritham 2007). Each retroelement family exhibits a differential amplification among genera or even within a single plant species (Du et al. 2010). Currently, the most popular definition of LTR families is based on sequence identity, where two elements belong to the same family if they share > 80% sequence identity in >80% of their coding region, their LTR or both (Wicker et al. 2007). A family can contain various elements that have been rendered defective from point mutations or small insertions or deletions (indels), and may or may not retain sufficient DNA identity for family membership. The high diversification rate of retroelement families with repeated transpositional bursts may lead to an overestimation of the family number when using such definition based on sequence identity (El Baidouri and Panaud 2013). Instead, a classification based on a clustering strategy based on the nucleotide sequence of LTR has been proposed, which may be particularly useful when the analysis includes defective copies that accumulated several types of mutations (El Baidouri and Panaud 2013). In our research, we employed LTR sequences and the phylogenetic relations of amino acid sequences of RT and RH as strategy to delimitate retroelement families.

The availability of a high-quality assembly of the nuclear genome of tomato (*Solanum lycopersicum* cv. Heinz 1706) and transcriptomic data from different tissues enables the characterization and analysis of LTR retroelement families in this economically important species (The Tomato Genome Consortium 2012). We focused on the characterization of full- length Copia and Gypsy LTR retroelements to identify potentially active retroelement families in the tomato genome based on sequence identity, evolutionary affiliations, insertion time and different tissue expression. To date only a few active elements were described in the genus *Solanum* (Pearce et al. 1996; Cheng et al. 2009; Paz et al. 2015). These results provide a valuable resource for the *S. lycopersicum* genome annotation and comparative genomics within the Solanaceae, and may be useful to discover active LTR retroelements for transposon-tagging and molecular marker design in tomato. In addition, the importance of accurate and full-length LTR retroelement annotation is increasingly recognized as a priority in plant genome sequencing projects to minimize the inaccuracy of gene annotation and facilitate functional gene studies.

Materials and methods

Data mining

The nuclear genome sequence of the *Solanum lycopersicum* cv. Heinz 1706 (ITAG2.3) was obtained from the Sol Genomics Network (http://solgenomics.net/). *De novo* LTR retroelements were identified with the online software LTR-finder (http://tlife.fudan.edu.cn/ltr_finder/; Xu and Wang 2007) with the following parameters: (1) Minimal distance between LTRs: 3500 bp; (2) ps_scan algorithm to detect protein domains of RT, IN and RH if they are identified; (3) conserved domain prediction PBS (primer binding sequence) which was conducted assigning as a reference genome the database of "*Arabidopsis thaliana* (2004)"; (4) presence of conserved sequences, such as conserved endings TG-CA; and (5) at least two of the sites conserved TSR (terminal repeated sequences), PBS and PPT (poly purine tract terminal).

The sequence between the two putative LTRs (internal region) was subsequently analyzed by searching for conserved domains in the Conserved Domains databases at NCBI. 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 and RH for phylogenetic analyses were extracted from the domain alignment provided in the output of 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 LTR sequence clustering in a similar manner as described by (El Baidouri and Panaud 2013) and by evolutionary relationships based on a phylogeny tree of RT and RH.

Spatial distribution and diversity of full-length retroelements

The physical locations of Gypsy and Copia retroelements were mapped to the tomato chromosome sequences. To determine the spatial distribution pattern of each Superfamily, the frequency of Gypsy and Copia retroelements was calculated as the number of elements in non-overlapping 50-Mb windows in each chromosome and in the total genome. Spatial patterns of the retroelements were estimated based on the standardized Morisita (Krebs 1999) distribution indexes. This index falls into three critical values between −1 and 1 indicating: uniform (index value <0), random (index value =0), and clustered (index value >0).

The indexes of diversity commonly employed to characterize ecological communities were used to measure the diversity of the identified retroelements. In this case, retroelement families were considered as species, and the host genome as the habitat. The specific diversity index of Shannon–Weaver (Shannon and Weaver 1949) was determined using the formula: $H' = -\Sigma$ Pi ln Pi; with $Pi = S/N$ and where $S =$ number of individuals within a species and $N =$ total number of individuals in the sample. This index ranges between 0.5 and 5 in most natural systems. H′ varies between 2 and 3 in most natural systems; $H' < 2$ indicates low diversity; $H' > 3$, high diversity; while $H' > 5$ is indicative of very high diversity.

To estimate the richness within a retroelement community, the Margalef diversity index was employed (Margalef 1958). This index is based on the numerical distribution of individuals of different species depending on the number of individuals in the sample. This index takes into account the number of different species in a given area and is strongly dependent on sampling size. The formula employed was $Dmg = (S - 1)/\text{In } N$ Values of $Dmg < 2$ are considered low diversity zones, whereas values of $Dmg > 5.0$ are considered high diversity zones. Intermediate ranges are considered normal.

Phylogenetic analyses

The evolutionary relationships of all full-length LTR elements were analyzed. Reference sequences from previously characterized retrotransposons 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 and RH were performed with RAxML version 7.2.8, under the $JTT + \Gamma$ model. A hundred rapid bootstrap inferences were done with RAxML.

The sequence of one member per family was submitted to the DDBJ database (http://www.ddbj.nig.ac.jp), accession numbers LC012610-LC012706. Additionally, full-length LTR retroelement sequences were submitted to Gypsy Database (http://www.gydb.org).

Estimation of insertion time for LTR retrotransposons in tomato

Insertion time was estimated according the method described by Ma et al. (2004). CLUSTAL multiple alignment method from MEGA4 (Tamura et al. 2007) was used to align all LTR pairs. 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).

Expression analysis

To estimate transcript abundance of the retroelements, we analized RNA-seq data publicly available at the NCBI Sequence Read Archive (SRA), accession number SRP068096 that includes data from different tissues, organs and developmental stages from *S. lycopersicum* cv. Micro-Tom. For our analysis we selected root (SRR3095831), leaf (SRR3095793), bud (SRR3095829) and seed from fruits at the following ripening stages: immature green (SRR3095785), mature green (SRR3095826), breaker (SRR3095782), orange (SRR3095790) and red ripe (SRR3095828). Raw reads were first filtered to remove ribosomal RNA contamination using SILVA database (Quast et al. 2013) and HISAT2 (Kim et al. 2015) with-no-spliced-alignment setting. The same program was used, with default settings, to map clean reads to the tomato genome (ITAG2.3) obtained from the Sol Genomics Network (http://solgenomics.net/). The number of reads that mapped to each of the retroelements was used to estimate transcript abundance expressed as FPKM (fragments per kilobase of transcript per million mapped reads).

As a complementary approach to evaluate the expression of the retroelements, BLAST searches against EST nucleotide databases of *S. lycopersicum* (NCBI Taxid: 4081) were performed. We examined 351 independent bioprojects and 303,958 ESTs from *S. lycopersicum* available at NCBI databases (http://blast.ncbi.nlm.nih.gov/BlastAlign.cgi). Analyses were conducted using as a query the full-length LTR sequences from most representative elements of each family (copies with only one open reading frame and with the highest similarity between both LTR). Positive matches were considered when at least one high quality sequence exhibited at least 80% of similarity with the full-length LTR query sequence, score values >200 and E-value $<10^{-32}$ (Wicker et al. 2007; Vicient 2010; Marcon et al. 2015). Correlation analysis between the number of positive matches and frequency of each retroelement family was performed by Pearson correlation analysis using the software InfoStat version 2012 (Di Rienzo et al. 2017).

Results

Number and distribution of full-length LTR retroelements in the tomato genome

The analysis of the *S. lycopersicum* genome using the software LTR-Finder returned 1859 hits of putative LTR retroelements. A further evaluation of their integrity and presence of all constitutive proteins yielded 736 structurally full-length retroelements, 331 (44%) corresponding to the Superfamily Copia and 405 (56%) to the Superfamily Gypsy (Table S2). Across the 12 tomato chromosomes, the Copia:Gypsy ratio varied from 0.4 to 1.3 (Table 1).

Despite the variable number of retroelements per chromosome, ranging from 26 to 134, there were no significant

Table 1 Distribution, frequency, and density of the 736 full-length LTR retroelements identified in the 12 chromosomes of the tomato genome

Chromosome	No. of LTR retrotransposons [Copia; Gypsy]	Ratio Copia:Gypsy	Density per 10 million of bp total [Copia/Gypsy]	Chromosome size (bp)
Ch01	134 [58; 76]	0.8	14.9 [6.5/8.4]	90,304,244
Ch02	37 [20; 17]	1.2	7.4 [4.0/3.4]	49,918,160
Ch03	52 [15; 37]	0.4	8.0 [2.3/5.7]	64,840,714
Ch04	56 [28; 28]	1.0	8.7 [4.4/4.4]	64,064,312
Ch05	44 [25; 19]	1.3	6.8 [3.8/2.9]	65,021,438
Ch06	26 [10; 16]	0.6	5.6 [2.2/3.5]	46,041,636
Ch07	65 [31; 34]	0.9	10.1 [4.9/5.2]	65,268,621
Ch08	81 [40; 41]	1.0	12.9 [6.3/6.5]	63,032,657
Ch09	74 [34; 40]	0.9	10.9 [5.0/5.9]	67,662,091
Ch10	50 [26; 24]	1.1	7.7 [4.0/3.7]	64,834,305
Ch11	53 [21; 32]	0.7	9.9 [3.9/6.0]	53,386,025
Ch12	64 [23; 41]	0.6	9.8 [3.5/6.3]	65,486,253
Total	736 [331; 405]	0.8	9.7 [4.4/5.3]	759,860,456

differences in their density (χ^2 test, df = 1; Table 1). In fact, the number of retroelements was significantly correlated with chromosome length (Pearson correlation r = 0.89; p < 0.0001). However the analysis of spatial pattern of dispersion of retroelements using standardized Morisita Index revealed that, in most cases, both Gypsy and Copia retroelements exhibited a clustered distribution (Fig. 1). Exceptions to this rule were found for either Copia or Gypsy elements in chromosomes with low number of retroelements (chromosomes 5, 6, 10, 11 and 12), and with a spatial pattern of distribution tending to be uniform or random (Fig. 1).

Family delimitation and evolutionary relationships of full-length LTR retroelements

The 736 full-length LTR retroelements identified in the tomato genome were grouped in 97 families based on evolutionary tendencies of the RT sequences (Fig. 2 and Fig. S2 and Fig. S3) and on LTR sequence identity. However, the phylogenetic analyses alone were sufficient to classify the tomato LTR retroelements at the Superfamily, lineage, and Family levels. This analysis was also performed using RH sequences yielding same results (Fig. S4 and S5). Families were also compared with LTR retroelements previously described in tomato (Table S3).Families exhibited differences in frequency (number of elements) and sequence length (Fig. 3; Table S3). A total of 77 and 20 families belonged to the superfamilies Copia and Gypsy, respectively. Most families (49% of Copia and 63% of Gypsy) had multiple members, while the rest were monotypic with only one member (Table S3). Comparisons among the 97 families of LTR retroelements also revealed differences in family diversity, where Copia retroelements (H´ = 3.69; Dm = 13.26) were more diverse than Gypsy ones (H´ = 1.05; Dm = 2.66).

Phylogenetic trees constructed based on protein alignments of the RT domain revealed that the tomato Gypsy elements belonged to five of the six main lineages defined within Gypsy (Athila, Tat, Galadriel, Reina and Del/Tekay;

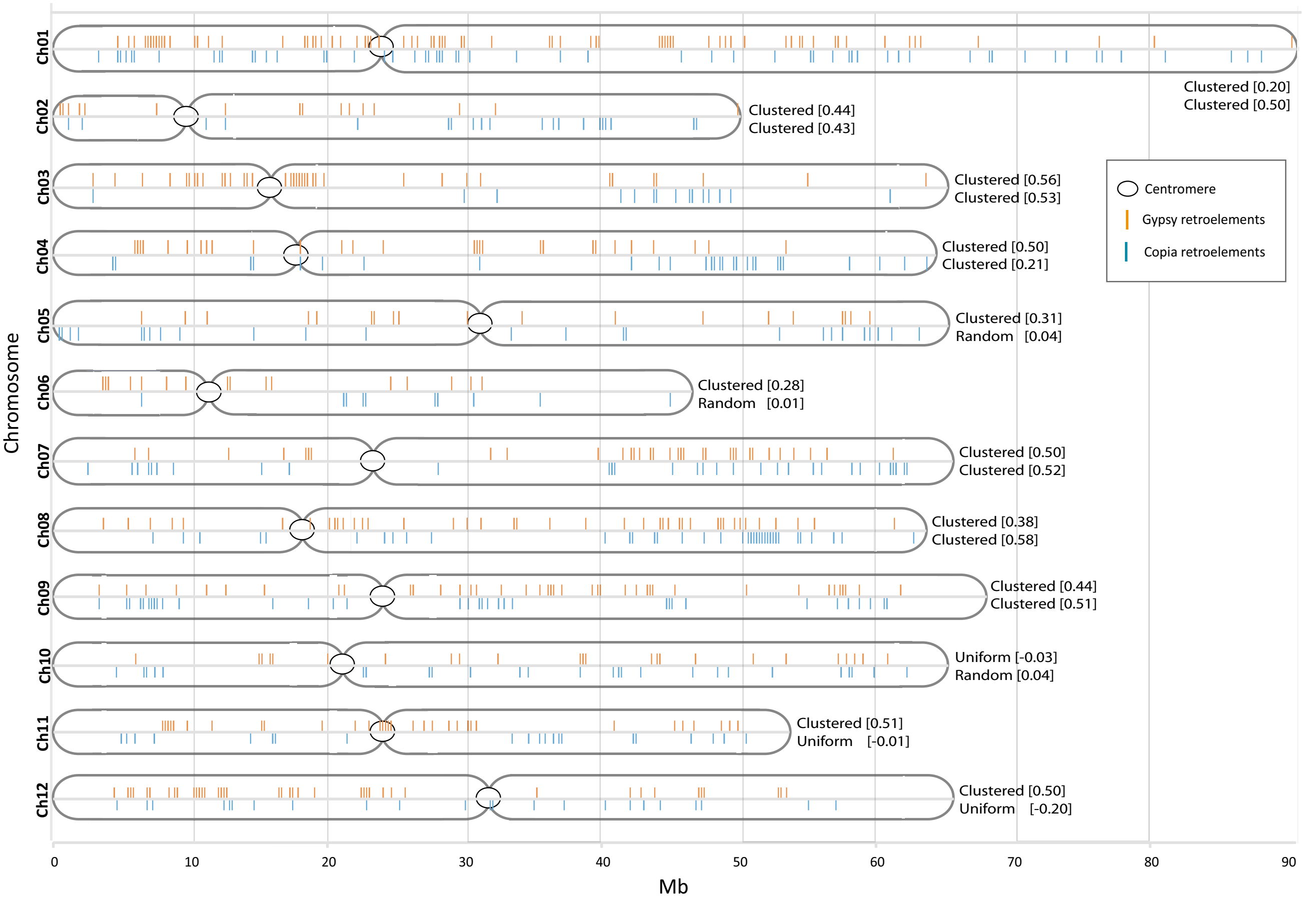
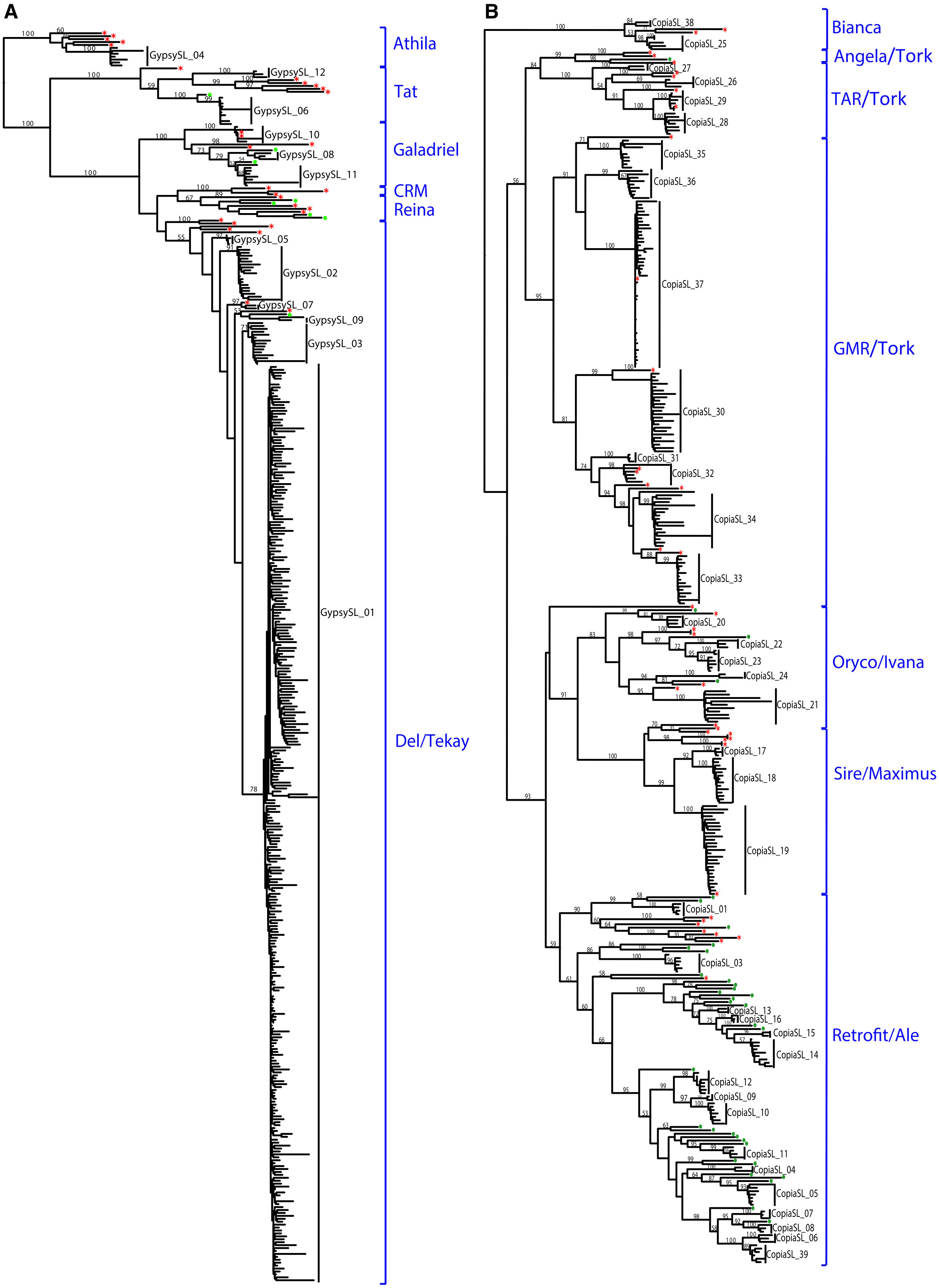


Fig. 1 Chromosomal distribution of complete Copia and Gypsy retroelements in the *S. lycopersicum* genome (*above and below the midline* of each chromosome, respectively). The standardized Morisita Index and the type of distribution are shown on the *right* of each chromosome. Approximate location of the centromeres are indicated according to The Tomato Genome Consortium, (2012)


A
GypsySL_04
GypsySL_12
GypsySL_06
GypsySL_10
GypsySL_08
GypsySL_11
GypsySL_05
GypsySL_02
GypsySL_07
GypsySL_09
GypsySL_03
GypsySL_01
Athila
Tat
Galadriel
CRM
Reina
Del/Tekay
B
CopiaSL_38
CopiaSL_25
CopiaSL_27
CopiaSL_26
CopiaSL_29
CopiaSL_28
CopiaSL_35
CopiaSL_36
CopiaSL_37
CopiaSL_30
CopiaSL_31
CopiaSL_32
CopiaSL_34
CopiaSL_33
CopiaSL_20
CopiaSL_22
CopiaSL_23
CopiaSL_24
CopiaSL_21
CopiaSL_17
CopiaSL_18
CopiaSL_19
CopiaSL_01
CopiaSL_03
CopiaSL_13
CopiaSL_16
CopiaSL_15
CopiaSL_14
CopiaSL_12
CopiaSL_09
CopiaSL_10
CopiaSL_11
CopiaSL_04
CopiaSL_05
CopiaSL_07
CopiaSL_08
CopiaSL_06
CopiaSL_39
Bianca
Angela/Tork
TAR/Tork
GMR/Tork
Oryco/Ivana
Sire/Maximus
Retrofit/Ale

◂**Fig. 2** Phylogenetic trees based on amino acid sequences of the RT from 433 Gypsy (**a**) and 370 Copia (**b**) LTR retroelements, respectively, including reference sequences (marked with *asterisks*, Table S1). Retroelement families and known lineages are indicated by *vertical lines* and *brackets*, respectively. Monotypic families are indicated by green circles. *Numbers above the branches* represent bootstrap support values >50%

Fig. 2a); whereas within the Superfamily Copia we identified seven well defined lineages (Sire/Maximus, Oryco/Ivana, Retrofit/Ale, Bianca, TAR/Tork, Angela/Tork, GMR/Tork; Fig. 2b). The number of retroelement families per lineage was variable (Table S3; Fig. 3). Among the Copia retroelements, the most family-rich lineage was Retrofit/Ale with 50 families, followed by GMR, and Oryco/Ivana with 9, and 8 families, respectively (Fig. 3c). Contrarily, all Gypsy lineages included less than 7 families (Fig. 3c). However, the total number of retroelements per Gypsy lineage exhibited a different trend, being Del/Tekay the most numerous (>350 elements), with ~2 and ~threefold more retroelements than the Copia lineages GMR/Tork and Retrofit/Ale, respectively (Fig. 3d). The remaining lineages contained less than 50 retroelements each (Fig. 3d).

The lineages Del/Tekay and GMR/Tork included the two most numerous families: GypsySL_01 and CopiaSL_37 with 318 and 49 copies, respectively (Fig. 3a). Together they accounted for almost half of all full-length retroelements identified in this work (Table S3). In the case of the family-rich lineage Retrofit/Ale, most of the families (68%) were monotypic or poorly populated (less than 10 members). This trend in terms of proportion of monotypic families and limited abundance was observed in other lineages, such as Oryco/Ivana, Galadriel, Reina and Tat (Fig. 3a). The lineage Sire/Maximus presented one family (CopiaSL_19) with more than 20 individuals, whereas the other two families within this lineage were low frequency ones. In contrast, the lineages Athila and Angela/Tork were each represented by a single, scarcely populated family (GypsySL_04 and CopiaSL_monotypic|Ch12_2s29).

Insertion time

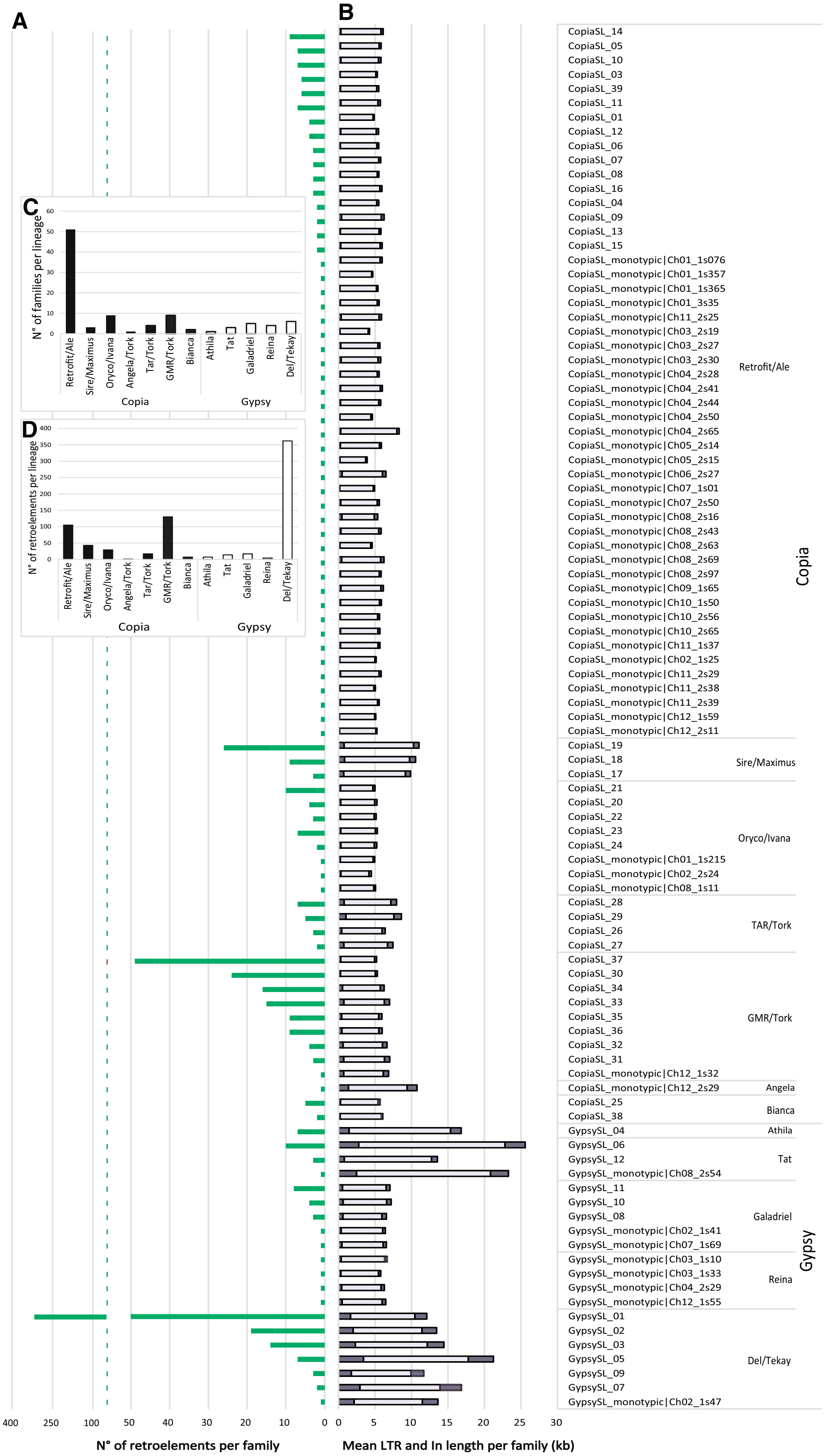
The analysis of insertion time of the 736 retroelements revealed that 73% of them inserted in the last 2 MYA, and 16% within the last 0.5 MYA (Table S2). The comparison of the insertion time in both retroelement superfamilies suggests that most of Copy members are younger than Gypsy ones (ANOVA, $p < 0.0001$). The insertion time of 49% of Copia retroelements were estimated in less than 1.0 MYA (Fig. S6). Contrarily, a small fraction of Gypsy members were inserted within the last 1.0 MYA (26%), while most of them (60%) inserted between 1.0 and 2.5 MYA ago. Likewise the fraction of retroelements with more than 2.5 MYA of insertion time is very small for both Copia and Gypsy superfamilies (12.6 and 13.7% respectively).

The evaluation of the insertion time within each phylogenetic lineage also revealed differences on the dynamics of full-length retroelements (Table 2). Out of 12 lineages of potentially-active LTR retroelement identified in the tomato genome, only 7 contained members inserted less than a 0.5 MYA (Retrofit/Ale, Sire/Maximus, Oryco/Ivana, Tork/GMR, Tat, Galadriel and Del/Tekay). Other lineages exhibited a burst of amplification in more ancient times. This is the case of Sire/Maximus (72% of retroelements inserted between 0.5 and 2.0 MYA); Oryco/Ivana (72%, 1.0–3.0 MYA); Tork/Tar (78% inserted between 0.5 and 2.5 MYA); Galadriel (66% inserted between 1.0 and 2.5 MYA); Reina (100% inserted between 0.5 and 3.0 MYA). In the case of the lineage Del/Tekay, despite the elevated number of retroelements inserted very recently (30 retroelements earlier than 0.5 MYA), the burst of amplification occurred more anciently (79% inserted between 0.5 and 2.5 MYA).

Expression analysis

In order to estimate transcript levels of the retroelements, we used RNAseq data publicly available derived from *S. lycopersicum* roots, leaves, buds and seeds from fruits at the following ripening stages: immature green, mature green, breaker, orange and red ripe. Additionally, we employed EST data sets from *S. lycopersicum* available in NCBI. Our analysis revealed potential tissue-specific transcriptional activity in at least 58 of the 97 families of retroelements identified in this research (Table S4, Table S5). We observed a positive correlation between the number of families per lineage identified by both methods (Pearson correlation $= 0.97$; $p < 0.0001$). In fact, the comparison of both expression analyses showed that 29 of the expressed families are common, whereas 8 and 27 families were identified solely by RNA-seq and EST analysis respectively (Table S6).

RNA-seq-based analyses identified 30 families from 9 lineages (families belonging to Angela/Tork, Athila and Reina were absent) with 59 retroelements copies with potential expression activity in tomato (FPKM > 1). Positive correlations between the number of hits (copies with FPKM > 1) and the number of copies within each family (Pearson correlation $r = 0.59$; $p < 0.0001$) and within each lineage (Pearson correlation $r = 0.82$; $p < 0.001$) were observed. Interestingly, most of the monotypic retroelements did not show any expression activity, in contrast to the highly numerous families of retroelements. Also, most of the highly populated lineages such as Retrofit, GMR/Tork, and Del/Tekay exhibited the highest values of hits and expression levels (Table S7). However, not all copies of a family were expressed, with only a few copies


A
B
C
D
N° of families per lineage
60
50
40
30
20
10
0
Retrofit/Ale
Sire/Maximus
Oryco/Ivana
Angela/Tork
Tar/Tork
GMR/Tork
Bianca
Athila
Tat
Galadriel
Reina
Del/Tekay
Copia
Gypsy
N° of retroelements per lineage
400
350
300
250
200
150
100
50
0
CopiaSL_14
CopiaSL_05
CopiaSL_10
CopiaSL_03
CopiaSL_39
CopiaSL_11
CopiaSL_01
CopiaSL_12
CopiaSL_06
CopiaSL_07
CopiaSL_08
CopiaSL_16
CopiaSL_04
CopiaSL_09
CopiaSL_13
CopiaSL_15
CopiaSL_monotypic|Ch01_1s076
CopiaSL_monotypic|Ch01_1s357
CopiaSL_monotypic|Ch01_1s365
CopiaSL_monotypic|Ch01_3s35
CopiaSL_monotypic|Ch11_2s25
CopiaSL_monotypic|Ch03_2s19
CopiaSL_monotypic|Ch03_2s27
CopiaSL_monotypic|Ch03_2s30
CopiaSL_monotypic|Ch04_2s28
CopiaSL_monotypic|Ch04_2s41
CopiaSL_monotypic|Ch04_2s44
CopiaSL_monotypic|Ch04_2s50
CopiaSL_monotypic|Ch04_2s65
CopiaSL_monotypic|Ch05_2s14
CopiaSL_monotypic|Ch05_2s15
CopiaSL_monotypic|Ch06_2s27
CopiaSL_monotypic|Ch07_1s01
CopiaSL_monotypic|Ch07_2s50
CopiaSL_monotypic|Ch08_2s16
CopiaSL_monotypic|Ch08_2s43
CopiaSL_monotypic|Ch08_2s63
CopiaSL_monotypic|Ch08_2s69
CopiaSL_monotypic|Ch08_2s97
CopiaSL_monotypic|Ch09_1s65
CopiaSL_monotypic|Ch10_1s50
CopiaSL_monotypic|Ch10_2s56
CopiaSL_monotypic|Ch10_2s65
CopiaSL_monotypic|Ch11_1s37
CopiaSL_monotypic|Ch02_1s25
CopiaSL_monotypic|Ch11_2s29
CopiaSL_monotypic|Ch11_2s38
CopiaSL_monotypic|Ch11_2s39
CopiaSL_monotypic|Ch12_1s59
CopiaSL_monotypic|Ch12_2s11
Retrofit/Ale
CopiaSL_19
CopiaSL_18
CopiaSL_17
Sire/Maximus
CopiaSL_21
CopiaSL_20
CopiaSL_22
CopiaSL_23
CopiaSL_24
CopiaSL_monotypic|Ch01_1s215
CopiaSL_monotypic|Ch02_2s24
CopiaSL_monotypic|Ch08_1s11
Oryco/Ivana
CopiaSL_28
CopiaSL_29
CopiaSL_26
CopiaSL_27
TAR/Tork
CopiaSL_37
CopiaSL_30
CopiaSL_34
CopiaSL_33
CopiaSL_35
CopiaSL_36
CopiaSL_32
CopiaSL_31
CopiaSL_monotypic|Ch12_1s32
GMR/Tork
CopiaSL_monotypic|Ch12_2s29
Angela
CopiaSL_25
CopiaSL_38
Bianca
Copia
GypsySL_04
Athila
GypsySL_06
GypsySL_12
GypsySL_monotypic|Ch08_2s54
Tat
GypsySL_11
GypsySL_10
GypsySL_08
GypsySL_monotypic|Ch02_1s41
GypsySL_monotypic|Ch07_1s69
Galadriel
GypsySL_monotypic|Ch03_1s10
GypsySL_monotypic|Ch03_1s33
GypsySL_monotypic|Ch04_2s29
GypsySL_monotypic|Ch12_1s55
Reina
GypsySL_01
GypsySL_02
GypsySL_03
GypsySL_05
GypsySL_09
GypsySL_07
GypsySL_monotypic|Ch02_1s47
Del/Tekay
Gypsy
400
250
100
50
40
30
20
10
0
0
5
10
15
20
25
30
N° of retroelements per family
Mean LTR and In length per family (kb)

◂**Fig. 3** Diversity, frequency, and mean lengths of complete Copia and Gypsy retroelements in the *S. lycopersicum* genome. **a** Number of complete retroelements per family. **b** Mean lengths of the LTR (*grey*) and internal portion (In, *white*) of the retroelements in each family. **c** Number of families of retroelements per lineage. **d** Total number of retroelements per lineage

transcriptionally active. For example, the most populated family GypsySL_01 exhibited only one active copy (<1% active copies). Contrarily, families moderately populated such as CopiaSL_37, CopiaSL_30, CopiaSL_19 and GypsySL_05 have 10, 30, 12 and 70% of their copies with some level of expression.

Some of the families exhibited a trendy tissue-specific expression (Fig. 4, Table S7). This is the case of GypsySL_05 (FPKMmax = 2.0 in leaf and bud tissue), GypsySL_09 (FPKMmax = 3.0 and 2.5 in SeedO and Seed RR stages, respectively), CopiaSL_30 (FPKMmax = 1.0, 1.7, 1.6 and 3.9 in Root, Leaf, Bud and Seed RR tissues respectively), CopiaSL_37 (FPKMmax = 1.2, 2.5, 1.9 and 1.6 in Root, Leaf, Bud and Seed RR tissues respectively). Retrofit, the most diverse lineage identified in this research, exhibited 10 families potentially expressed in plant vegetative tissues, being these values noticeably higher in some cases such as CopiaSL_monotypc|Ch03_2s19 (FPKM = 4.06 and 3.34 in bud and root tissues respectively) and CopiaSL_monotypc|Ch10_2s56 (FPKM = 9.63 in bud tissue). The most markedly values of FPKM were registered in a member of GMR/Tork lineage, particularly in one copy of CopiaSL_32 family, with high values in leaf (FPKM = 47.15), bud (FPKM = 5.46), root (FPKM = 3.66) and seed IG and RR stages (FPKM = 3.27 and 3.85 respectively) (Fig. 4).

Finally, a positive correlation was observed between the families that are transcriptionally-active estimated by both methods and the number of retroelements inserted in less than 1.0 Mya (Table 3). The remaining class intervals do not show significant differences.

Discussion

Transposable elements and repetitive DNA derived from these represent one of the most outstanding genomic features of plants (Huang et al. 2012). They are important biological entities because their activation greatly affects the evolution of the host genome (Biémont and Vieira 2006). Genomes change over time through insertions, deletions, and recombination of transposable elements across different chromosomes affecting their size and promoting the generation of genetic diversity and the evolution of new genes and regulatory networks (Bowen et al. 2003; Pritham et al. 2007). Thus, the identification and classification of full-length and potentially-active elements in a particular species, such as tomato, allow comparative and integrative approaches to understand its life cycle and impact on the generation of genetic variability and on plant evolution.

Identification and distribution of full-length LTR retroelements in the *S. lycopersicum* genome

Computational algorithms have been developed to identify transposable elements in genomes and they revealed marked differences between species, in terms of copy number, repertoire and level of breakdown (Le Rouzic and Capy 2006; Bergman and Quesneville 2007; Pritham 2009). Most research has been oriented to identify and classify plant retroelements employing the criterion proposed by Ma et al. (2004) that subdivide LTR retroelements in intact, solo, and truncated populations according to the level of integrity of LTR sequences alone. From a functional point of view, this criterion has limitations because it does not contemplate the functional portion of the retroelement that allows its mobility (El Baidouri and Panaud 2013). Thus, we chose a more stringent criterion of classification based on the presence of both LTR (intact elements according Ma et al. 2004) and the presence of all constitutive structural proteins to define *full-length retroelements* (5′LTR-GagPol-3′LTR). Under this criterion, we identified those retroelements, which are probably autonomous and potentially active within the tomato genome.

A preliminary analysis of the genome of *S. lycopersicum* cv. Heinz 1706 (ITAG version 2.40) revealed that the retroelements accounted for approximately 62% (460 Mb) of the genome and were mostly incomplete or truncated (The Tomato Genome Consortium 2012). Recently, an in-depth genome analysis reported the presence of 15,134 L retroelements (Xu and Du 2014) of which 2086 were found to be intact according Ma et al. (2004). In our study, each retroelement identified was thoroughly analyzed to identify all constitutive parts of a full-length retroelements described above and found 736 full-length retroelements. Thus, only a small portion of the LTR retroelements inhabiting the tomato genome contains all the structural and functional components that allow their autonomous mobilization.

Of those full-length elements identified here, the overall ratio Copia:Gypsy was 0.8, indicating that both superfamilies are similarly populated in this species. If the analysis is performed including intact, solo and truncated retroelements, this ratio decreases to 0.58 (Xu and Du 2014). This variation might be attributed to the ancient radiation of Gypsy superfamily in the tomato genome (Fig. S6). A comparison with other angiosperms shows that this ratio is variable across plants: *Solanum phureja*, 3.5; *Capsicum annum* 2.3 (Paz RC, Yañez Santos AM, Andino NP; unpublished results); *Glycine max*, 1.4 (Du et al. 2010); *Zea mays*, 1.6 (Schnable et al. 2009); *Oryza*

Table 2 Estimated insertion times of different lineages of full-length Copia and Gypsy retroelements in tomato

Mya	Absolute frequency (relative frequency)											
Class (IL–SL]	Retrofit/ Ale	Sire/ Maximus	Oryco/ Ivana	Angela/ Tork	Tork/TAR	Tork/GMR	Bianca	Athila	Tat	Galadriel	Reina	Del/Tekay
[0.00–0.50]	**39 (0.38)**	3 (0.07)	4 (0.14)	0 (0.00)	0 (0.00)	**33 (0.25)**	0 (0.00)	0 (0.00)	**5 (0.36)**	2 (0.12)	0 (0.00)	30 (0.08)
(0.50–1.00]	**33 (0.32)**	**11 (0.26)**	3 (0.1)	0 (0.00)	3 (0.18)	**30 (0.23)**	2 (0.29)	**2 (0.29)**	**4 (0.29)**	2 (0.12)	**2 (0.50)**	57 (0.16)
(1.00–1.50]	15 (0.14)	10 (0.23)	2 (0.07)	0 (0.00)	**4 (0.24)**	21 (0.16)	1 (0.14)	1 (0.14)	2 (0.14)	4 (0.24)	0 (0.00)	**89 (0.25)**
(1.50–2.00]	6 (0.06)	10 (0.23)	**11 (0.38)**	**1 (1.00)**	3 (0.18)	18 (0.14)	0 (0.00)	1 (0.14)	2 (0.14)	3 (0.18)	**1 (0.25)**	66 (0.18)
(2.00–2.50]	4 (0.04)	3 (0.07)	1 (0.03)	0 (0.00)	3 (0.18)	16 (0.12)	1 (0.14)	**2 (0.29)**	0 (0.00)	4 (0.24)	0 (0.00)	71 (0.2)
(2.50–3.00]	4 (0.04)	4 (0.09)	7 (0.24)	0 (0.00)	1 (0.06)	4 (0.03)	3 (0.43)	0 (0.00)	0 (0.00)	1 (0.06)	**1 (0.25)**	22 (0.06)
(3.00–3.50]	1 (0.01)	1 (0.02)	1 (0.03)	0 (0.00)	1 (0.06)	3 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.06)	0 (0.00)	18 (0.05)
(3.50–4.00]	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	2 (0.12)	3 (0.02)	0 (0.00)	1 (0.14)	1 (0.07)	0 (0.00)	0 (0.00)	3 (0.01)
(4.00–4.50]	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (0.01)
(4.50–5.00]	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	2 (0.01)
(5.00–5.50]	1 (0.01)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.003)
(5.50–6.00]	0 (0.00)	1 (0.02)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.003)
(6.00–6.50]	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.003)

Bold numbers reflect classes of MYA with a frequency higher than 0.20 within each phylogenetic clade

sativa, 4.9 (Tian et al. 2009); and *Sorghum bicolor*, 3.7 (Paterson et al. 2009). However, with the exception of the first two species analysis, the remainder studies were performed analyzing intact, truncated and solo retroelement copies. This reveals that retroelements expansion within a genome is unique, and depends on the evolutionary process within each different plant species.

Our results suggested that the distribution of full-length LTR retroelements across the tomato chromosomes is clustered, in agreement with previous research proposing that the distribution of retroelements in plant genomes is non-random and they are commonly found in clusters (Hua-Van et al. 2011). This skewed distribution appears to respond to the integration affinity of different types of retroelements within specific genomic environments under neutral selection and where they can escape regions with genomic recombination and the epigenetic control exerted by the host genome (Hua-Van et al. 2011; Pereira 2004).

Diversity and phylogenetic analyses of full-length retroelements in the tomato genome

Delimitation of retroelement families represents a great challenge given their high rates of evolution, particularly when truncated or incomplete copies are included in the analyses. In this study, we identified 97 families of retroelements based on phylogenetic analyses of the RT protein. Most families in the tomato genome belonged to Copia (80%) and the remainder to Gypsy (20%). Retroelements from the same family showed high levels of LTR identity among them. In contrast, elements from different families did not share any similarity in their LTR sequences. Similar results were previously reported (Ma et al. 2004; Nagaki et al. 2004).

Ample variation in the number of retroelements per family was observed in this study. The most extreme case is the family GypsySL_01, which represents 40% of all the identified full-length LTR retroelements, in contrast to the 47 monotypic families (48%). Analyses of the copy number within each retroelement family in other plants species also showed great variability because some families suffered dramatic amplification in a similar manner to GypsySL_01. This is the case of SNARE with more than 5000 copies and five other families of retrotransposons with 100 copies in the *Glycine max* genome (Du et al. 2010). In *Zea mays*, five families of LTR retrotransposons represent ~80% of the maize retrotransposon repertoire (Sanmiguel and Bennetzen 1998; Schnable et al. 2009). An extreme case has been reported in the wild rice *Oryza australiensis*, where the amplification of only three LTR retrotransposon families doubled the size of its genome (Piegu et al. 2006).

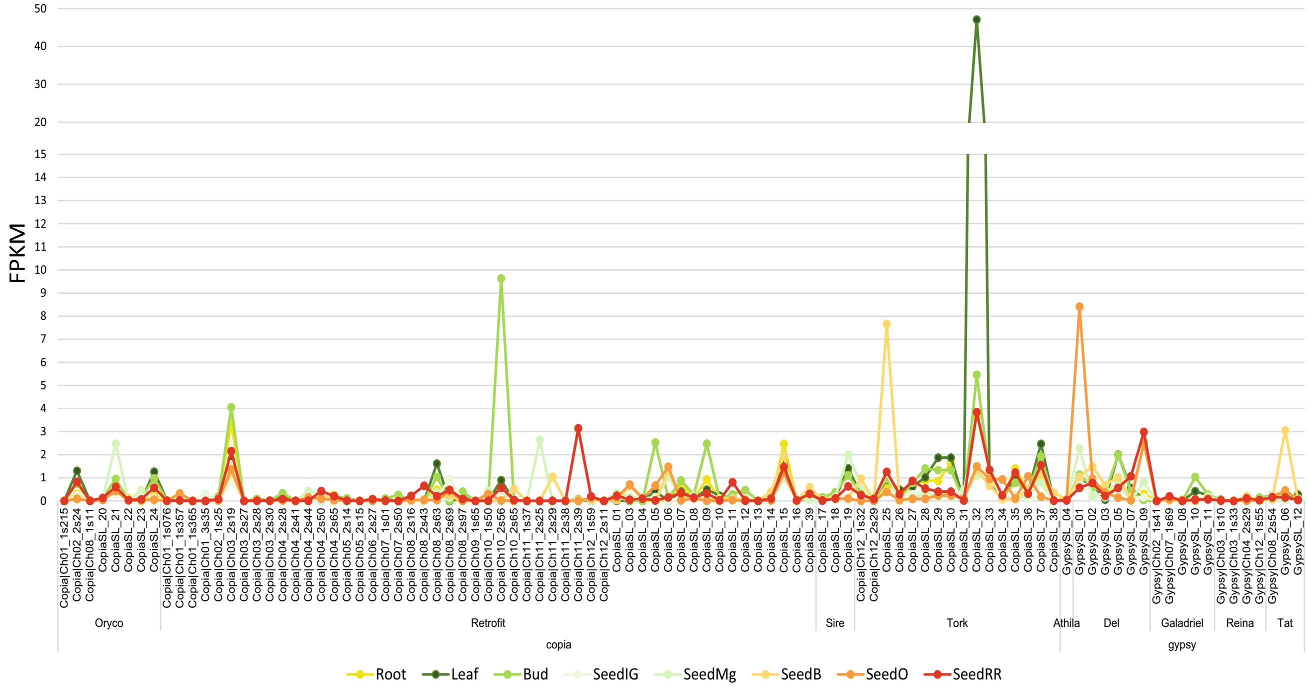


Fig. 4 Tissue-specific expression analysis based on RNA-seq analysis of each retroelement family identified in tomato. Only maximum values of FPKM (fragments per kilobase of transcript per million mapped reads) per family were expressed. *SeedIG* seeds from fruits in immature green stage, *SeedMg* seeds from fruits in mature green stage, *SeedB* seeds from fruits in breaker stage, *SeedO* seeds from fruits in orange stage, *SeedRR* seeds from fruits in red ripe stage

Dynamics of full-length retroelements in the tomato genome

The nucleotide identity between both LTR of a retroelement is helpful to estimate its insertion time because two LTRs of a single LTR retrotransposon are usually identical at the nucleotide sequence level upon integration (Sanmiguel and Bennetzen 1998). In fact, it is widely accepted that copies of retroelements inserted recently usually are more structurally intact and complete, whereas older ones contain a higher percentage of truncated elements and solo LTRs (Ma et al. 2004). Thus, the more recent the integration, the more likely it is active today. All the copies of the retroelements identified in this study are intact and full-length, with a high level of similarity between both LTR, being hence potentially active. In fact, more than 90% of them were inserted in the tomato genome within the last 2.5 Mya, 36% within the last 1 Mya and 16% within 0.5 Mya. Twenty-two copies of retroelements, mostly of the Superfamily Copia, showed an insertion time of 0.0 Mya, suggesting their very recent activation (Supplementary Table 2). The relatively recent insertion times of the full-length retroelements compared to the estimated insertion times of intact retroelements (up to 5 MYA; Ma et al. 2004; Du et al. 2010; Vitte et al. 2013), could be explained by the fact that potentially autonomous full-length copies had a more recent activation than those non-autonomous elements that only conserve intact LTR sequences.

Table 3 Pearson correlation between insertion times of retroelements and RNA-seq and EST expression analyses

MYA classes (IL–SL]	Number of families with positive hits per lineage	
	RNA-seq	EST
[0.00–0.50]	0.94 ($p<0.0001$)	0.93 ($p<0.0001$)
(0.50–1.00]	0.75 ($p<0.05$)	0.72 ($p<0.01$)
(>1.00]	*ns*	*ns*

The values in parentheses correspond to the significance level of Pearson statistic, whereas *ns* indicate not significant correlation

The amplification timeframe of identified retroelements varies dramatically in different, superfamilies, lineages and families, having Copia retroelements higher activity in the recent 0.5 Mya than Gypsy (Table 2). In this sense, we observed several waves of LTR retrotransposon amplification events, with different temporal patterns of transpositional activity among lineages, being Retrofit/Ale, Tork/GMR and Tat the most active recently (<1.0 Mya). Contrarily, despite several members of Del/Tekay lineage were inserted recently, the peak of activity occurred within the 1.0–2.5 Mya ago. This suggests that retroelements were active in waves, rather than continuously within the tomato genome. Similar behavior of LTR retroelements were reported in different plant species (Wicker and Keller 2007; Du et al. 2010; Wollrab et al. 2012; Beulé et al. 2015; Marcon et al. 2015; Yin et al. 2015).

The mechanisms behind bursts of amplification of only a few retrotransposon families are poorly understood, and the most accepted explanation is that those families were able to escape stochastically from silencing cellular mechanisms of the host genome (Lucas et al. 1995; Hirochika et al. 2000; Paz et al. 2015). Currently, several reports suggest that most of the repertoire of retroelements from a particular genome are in a quiescent state (Picault et al. 2009; Vicient 2010; Beulé et al. 2015). Several mechanisms, in two major pathways, can control the activity of retroelements in plants (Feschotte and Pritham 2007). Those mechanisms include methylation of both DNA and histones, as well as post-transcriptional silencing through the siRNA pathway (Slotkin and Martienssen 2007). In addition, RNA-directed DNA methylation (RdDM), a process that leads to chromatin modification through siRNA, is also dependent on the number of copies of a given element and the level of its transcription (Perez-Hormaeche et al. 2008; Picault et al. 2009). In this sense, only 30% of the retroelement families in tomato exhibited some level of transcriptional activity (Table 3). Of those, only a few copies within each family were potentially expressed, mostly with low to moderate levels of transcription (FPKM values from 1 to 10) and occasionally with high levels of transcription (FPKM > 10) (Fig. 4). Even though we observed a positive correlation between the number of copies within a particular family and the level of expression, the number of positive hits are very low in high-copy number families (1–10%) and high in low-copy number families (30–70%).

Probably, the highly repeated tomato families of LTR retrotransposons such as GypsySL_01 or CopiaSL_37 may be the target of such silencing pathways, in contrast to low-copy families such as GypsySL_05, CopiaSL_30 and some monotypic families. In this sense, some experimental evidence suggests that the family GypsySL_01 is highly methylated in sites CG and CNG and consequently silenced (Wang et al. 2006).

Previously reported LTR retroelement families in tomato genome

Nearly 64% of the retroelements described in this research had been previously described (Table S3; Ganal et al. 1988; Parniske et al. 1999; Araujo et al. 2001; Yang et al. 2005; Wang et al. 2006; Tam et al. 2007; Salazar et al. 2007; Jiang et al. 2009; Yin et al. 2013; Xu and Du 2014). The remaining retroelements (~36%) had not been previously described, being this research its first characterization and description.

Even though Gypsy retroelements are the most abundant in the tomato genome, the first elements identified and the most widely studied belong to the Superfamily Copia. The family CopiaSL_32/Retrolyc contains the first retrotransposon whose activation induced by stress was reported and it is one of the best-characterized plant retroelements at the structural and functional levels. It was first discovered in *Nicotiana tabacum* (Tnt1; Grandbastien et al. 1989). Later, its presence was demonstrated in genomes of other Solanaceae and it is actually considered an ancestral retroelement family widely dispersed in the family Solanaceae but with low copy numbers (Manetti et al. 2007, 2009; and this study), and several active members in different species of Solanaceae (Grandbastien et al. 1989; Paz et al. 2015; Tam et al. 2007). The family CopiaSL_37/Rider has also been previously described and, in agreement with our findings, it was found to be a high-copy number retrotransposon family in the tomato genome that may still be active (Jiang et al. 2009).

Another family previously studied was the CopiaSL_monotypic|Chr03_2s19/ TARE1 that exhibited a rare single nucleotide mutation from ‘G’ to ‘A’ in the typical ‘TG’ at the ends of the two LTRs (Yin et al. 2013). In contrast to our results, this family showed high copy numbers (354 copies) distributed along all tomato chromosomes with a relative short burst of activity in the recent past (<1.7 Mya). This discrepancy between studies, in which 1 (this work) or 354 (Yin et al. 2013) copies were identified, can be explained by differences in the search strategy and constraints. Our search strategy excluded all LTR retroelements that lack the typical TG at the end of the LTRs.

In the case of Gypsy retroelements, the family GypsySL_01 has been described as the largest family of retrotransposons identified in tomato constituting about 2.5% of its nuclear genome and a copy number ranging between 2000 and 4000, with most copies being truncated or incomplete (Ganal et al. 1988; Wang et al. 2006; Xu and Du 2014). Different terminology has been used to name this family: PCRT1a (Ganal et al. 1988), TGRII (Yang et al. 2005), Jingling (Wang et al. 2006) and six different names for subsets of the family SL_RT_F319/SL_RT_F322/SL_RT_F324/SL_RT_F325/SL_RT_F329/SL_RT_F159 (Xu and Du 2014). The distribution of the members of this family in the tomato genome constituted pericentromeric heterochromatin (Park et al. 2011), and experimental data suggested that this family is silenced by methylation (Wang et al. 2006).

Acknowledgements This work was funded by Agencia Nacional de Promoción Científica y Tecnológica, Argentina (ANPCyT), D-TEC 0008/13 and PROJOVI 2015 (Universidad Nacional de San Juan). We thank C. Llorens for his valuable help with the Gypsy Database. This work was part of the undergraduate thesis of MEK.

References

Araujo PG, Casacuberta JM, Costa APP et al (2001) Retrolyc1 subfamilies defined by different U3 regulatory regions in the *Lycopersicon* genus. Mol Gen Genom 266:35–41

Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8(6):382–392. doi:10.1093/bib/bbm048

Beulé T, Agbessi MD, Dussert S et al (2015) Genome-wide analysis of LTR-retrotransposons in oil palm. BMC Genom 16:1–14. doi:10.1186/s12864-015-2023-1

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

Bowen NJ, Jordan IK, Epstein J a et al (2003) Retrotransposons and their recognition of pol II promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*. Genome Res 13:1984–1997. doi:10.1101/gr.1191603

Cheng X, Zhang D, Cheng Z, Keller B, Ling HQ (2009) A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. Genetics 181:1183–1193

Di Rienzo JA, Casanoves F, Balzarini MG et al (2017) InfoStat versión 2017. Grupo InfoStat, FCA. Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

Domingues DS, Cruz GMQ, Metcalfe CJ et al (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genom 13:137. doi:10.1186/1471-2164-13-137

Du J, Tian Z, Hans CS et al (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63:584–598. doi:10.1111/j.1365-313X.2010.04263.x

El Baidouri M, Panaud O (2013) Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol Evol 5:954–965. doi:10.1093/gbe/evt025

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

Ganal MW, Lapitan NLV, Tanksley SD (1988) A molecular and cytogenetic survey of major repeated DNA sequences in tomato (*Lycopersicon esculentum*). Mol Gen Genet 213:262–268

Gao D, Abernathy B, Rohksar D et al (2014) Annotation and sequence diversity of transposable elements in common bean (*Phaseolus vulgaris*). Front Plant Sci 5:339. doi:10.3389/fpls.2014.00339

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 27:221–224. doi:10.1093/molbev/msp259

Grandbastien MA, Spielmann A, Caboche C (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature 337:376–380

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

Hirochika H, Okamoto H, Kakutani T (2000) Silencing of retrotransposons in Arabidopsis and reactivation by the ddm1 mutation. Plant Cell 12:357–368

Huang CRL, Burns KH, Boeke JD (2012) Active transposition in genomes. Annu Rev Genet 46:651–675. doi:10.1146/annurev-genet-110711-155616

Hua-Van A, Le Rouzic A, Boutin TS et al (2011) The struggle for life of the genome's selfish architects. Biol Direct 6:19. doi:10.1186/1745-6150-6-19

Jiang N, Gao D, Xiao H, van der Knaap E (2009) Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. Plant J 60:181–193. doi:10.1111/j.1365-313X.2009.03946.x

Kim D, Langmead B, Salzberg S (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–360. doi:10.1038/nmeth.3317

Krebs CJ (1999) Estimation of Survival Rates. In: Ecological Methodology. pp 499–539

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

Le Rouzic A, Capy P (2006) Population genetics models of competition between transposable element subfamilies. Genetics 174:785–793. doi:10.1534/genetics.105.052241

Li W, Zhang P, Fellers JP et al (2004) Sequence composition, organization, and evolution of the core *Triticeae* genome. Plant J 40:500–511. doi:10.1111/j.1365-313X.2004.02228.x

Llorens C, Futami R, Covelli L et al (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res 39:D70–D74. doi:10.1093/nar/gkq1061

Lucas H, Feuerbach F, Grandbastien MA, Caboche M (1995) The tobacco retrotransposon Tnt1 transposes in *Arabidopsis thaliana*. EMBO J 14:2364–2373

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

Manetti ME, Rossi M, Costa APP et al (2007) Radiation of the Tnt1 retrotransposon superfamily in three *Solanaceae* genera. BMC Evol Biol 7:34. doi:10.1186/1471-2148-7-34

Manetti ME, Rossi M, Nakabashi M et al (2009) The Tnt1 family member Retrosol copy number and structure disclose retrotransposon diversification in different *Solanum* species. Mol Genet Genomics 281:261–271. doi:10.1007/s00438-008-0408-4

Marcon HS, Domingues DS, Silva JC et al (2015) Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic. BMC Plant Biol 15:198. doi:10.1186/s12870-015-0550-1

Margalef DR (1958) Information theory in ecology. Gen Syst 3:36–71

Nagaki K, Cheng Z, Ouyang S et al (2004) Sequencing of a rice centromere uncovers active genes. Nat Genet 36:138–145

Park M, Jo S, Kwon J-K et al (2011) Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. BMC Genom 12:85. doi:10.1186/1471-2164-12-85

Parniske M, Wulff BBH, Bonnema G et al (1999) Homologues of the Cf-9 disease resistance gene (Hcr9s) are present at multiple loci on the short arm of tomato chromosome 1. Mol Plant Microbe Interact 12:93–102

Paterson AH, Bowers JE, Bruggmann R et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

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

Pearce SR, Pich U, Harrison G et al (1996) The Ty1-Copia group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal hetero- chromatin. Chromosome Res 4(5):357–364

Pereira V (2004) Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol 5:R79

Perez-Hormaeche J, Potet F, Beauclair L et al (2008) Invasion of the Arabidopsis genome by the tobacco retrotransposon Tnt1 is controlled by reversible transcriptional gene silencing. Plant Physiol 147:1264–1278

Picault N, Chaparro C, Piegu B et al (2009) Identification of an active LTR retrotransposon in rice. Plant J 58:754–765. doi:10.1111/j.1365-313X.2009.03813.x

Piegu B, Guyot R, Picault N et al (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262–1269. doi:10.1101/gr.5290206

Pritham EJ (2009) Transposable elements and factors influencing their success in eukaryotes. J Hered 100:648–655. doi:10.1093/jhered/esp065

Pritham EJ, Putliwala T, Feschotte C (2007) Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390:3–17. doi:10.1016/j.gene.2006.08.008

Quast C, Pruesse E, Yilmaz P et al (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41(D1):D590–D596. doi:10.1093/nar/gks1219

Salazar M, González E, Casaretto J a et al (2007) The promoter of the TLC1.1 retrotransposon from *Solanum chilense* is activated by multiple stress-related signaling molecules. Plant Cell Rep 26:1861–1868. doi:10.1007/s00299-007-0375-y

Sanmiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot Lon 82:37–44

Schnable PS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112–1115

Shannon CE, Weaver W (1949) The mathematical theory of communication. University Illinois Press, Urbana

Slotkin R, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272–285

Tam SM, Causse M, Garchery C et al (2007) The distribution of copia-type retrotransposons and the evolutionary history of tomato and related wild species. J Evol Biol 20:1056–1072. doi:10.1111/j.1420-9101.2007.01293.x

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

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641. doi:10.1038/nature

Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL et al (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19:2221–2230

Vicient CM (2010) Transcriptional activity of transposable elements in maize. BMC Genomics 11:1–10. doi:10.1186/1471-2164-11-601

Vitte C, Estep MC, Leebens-Mack J, Bennetzen JL (2013) Young, intact and nested retrotransposons are abundant in the onion and asparagus genomes. Ann Bot 112:881–889. doi:10.1093/aob/mct155

Voytas DF, Boeke JD (2002) Ty1 and Ty5 of *Saccharomyces cerevisiae*. In: Craig NL et al (eds) Mobile DNA II. ASM, Washington, DC, pp 631–683

Wang Y, Tang X, Cheng Z et al (2006) Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome. Genetics 172:2529–2540. doi:10.1534/genetics.106.055772

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17:1072–1081. doi:10.1101/gr.6214107

Wicker T, Sabot F, Hua-Van A et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. doi:10.1038/nrg2165

Wilhelm M, Wilhelm FX (2001) Reverse transcription of retroviruses and LTR retrotransposons. Cell Mol Life Sci 58(9):1246–1262. doi:10.1007/PL00000937

Wollrab C, Heitkam T, Holtgräwe D et al (2012) Evolutionary reshuffling in the Errantivirus lineage Elbe within the *Beta vulgaris* genome. Plant J 72:636–651. doi:10.1111/j.1365-313X.2012.05107.x

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

Xu Z, Wang H (2007) "LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons". Nucleic Acids Res 35(web server issue):W265–W268

Xu Z, Liu J, Ni W et al (2017) GrTEdb: the first web-based database of transposable elements in cotton (*Gossypium raimondii*). Database (Oxford) 2017:1–7. doi:10.1093/database/bax013

Yang T-J, Lee S, Chang S-B et al (2005) In-depth sequence analysis of the tomato chromosome 12 centromeric region: identification of a large CAA block and characterization of pericentromere retrotranposons. Chromosoma 114:103–117. doi:10.1007/s00412-005-0342-8

Yin H, Liu J, Xu Y et al (2013) TARE1, a mutated copia-like LTR retrotransposon followed by recent massive amplification in tomato. PLoS One. doi:10.1371/journal.pone.0068587

Yin H, Du J, Wu J et al (2015) Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. Sci Rep 5:1–15. doi:10.1038/srep17644