Genomics Proteomics Bioinformatics 18 (2020) 321–332

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Genome Size Evolution Mediated by *Gypsy* Retrotransposons in Brassicaceae

Shi-Jian Zhang [1], Lei Liu [2], Ruolin Yang [3,*], Xiangfeng Wang [1,*]

[1] *Department of Crop Genomics and Bioinformatics, College of Agronomy and Biotechnology, National Maize Improvement Center of China, China Agricultural University, Beijing 100094, China*
[2] *Beijing Key Laboratory of Plant Resources Research and Development, School of Sciences, Beijing Technology and Business University, Beijing 100048, China*
[3] *College of Life Sciences, Northwest A&F University, Yangling 712100, China*

Received 9 January 2018; revised 31 May 2018; accepted 16 July 2018
Available online 31 October 2020

Handled by Songnian Hu

KEYWORDS

Brassicaceae;
Comparative genomics;
Gypsy retrotransposon;
Copia retrotransposon;
Genome size evolution

Abstract The dynamic activity of transposable elements (TEs) contributes to the vast diversity of genome size and architecture among plants. Here, we examined the genomic distribution and transposition activity of long terminal repeat retrotransposons (LTR-RTs) in *Arabidopsis thaliana* (*Ath*) and three of its relatives, *Arabidopsis lyrata* (*Aly*), *Eutrema salsugineum* (*Esa*), and *Schrenkiella parvula* (*Spa*), in **Brassicaceae**. Our analyses revealed the distinct evolutionary dynamics of ***Gypsy* retrotransposons**, which reflects the different patterns of genome size changes of the four species over the past million years. The rate of *Gypsy* transposition in *Aly* is approximately five times more rapid than that of *Ath* and *Esa*, suggesting an expanding *Aly* genome. *Gypsy* insertions in *Esa* are strictly confined to pericentromeric heterochromatin and associated with dramatic centromere expansion. In contrast, *Gypsy* insertions in *Spa* have been largely suppressed over the last million years, likely as a result of a combination of an inherent molecular mechanism of preferential DNA removal and purifying selection at *Gypsy* elements. Additionally, species-specific clades of *Gypsy* elements shaped the distinct genome architectures of *Aly* and *Esa*.

Introduction

Transposable elements (TEs) play important roles in shaping genome structure, directly or indirectly influencing gene function and creating novel genetic materials in host genomes [1,2]. In plants, proliferation of long terminal repeat retrotransposons (LTR-RTs) through a "copy-and-paste" mechanism contributes to the vast diversity of genome size in different plant species [3]. LTR-RTs are classified into two superfamilies based on the relative location of the integrase

[*] Corresponding authors.
E-mail: sysbio@cau.edu.cn (Wang X), desert.ruolin@gmail.com (Yang R).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2018.07.009
1672-0229

(INT) encoded by the *pol* gene, namely the *Gypsy*-like and *Copia*-like retrotransposable elements (abbreviated as the *Gypsy* and *Copia* elements, respectively, hereafter), which have had different impacts on the evolution of genome size [4,5]. Comparison of the recent activity of LTR-RTs, including their transposition rate and genomic context, among related plant species may facilitate estimation of trends in genome size changes over short periods of evolutionary time. Previous studies of the activity of LTR-RTs in *Arabidopsis lyrata* (*Aly*) and *Arabidopsis thaliana* (*Ath*) revealed active transposition of LTR-RTs in *Aly* and efficient removal of LTR-RTs in *Ath*, suggesting that the *Aly* genome has likely expanded over the past five million years, while the *Ath* genome has likely shrunk during this time period [6].

TE activity may also influence gene family evolution. Retrotransposons mediate gene family expansion by carrying adjacent genes and incorporating them into other genomic locations during transposition. It has been reported that the genes of rice and sorghum captured by LTR-RTs and fixed in the population exhibit signatures of positive selection [7]. Additional studies revealed the effects of natural selection on LTR-RTs. For example, Baucom and colleagues [8] systematically analyzed rice LTR-RTs and detected strong purifying selection, as well as a few episodes of positive selection. However, the scenarios in which positive selection may occur are poorly understood.

The patterns of LTR-RT integration into the host genome vary greatly among LTR-RT clades and involve specific LTR-RT domains, as well as interactions between LTR-RTs and the host genome [9–11]. The observed distribution pattern of LTR-RTs in a current genome might be a result of the combined effects of purifying selection and preferential insertion of LTR-RTs [12]. A type of LTR-RTs that includes chromoviruses (specific lineages of *Gypsy* elements) and those within the *Athila* family is especially intriguing to researchers because it exhibits strong site-targeting specificity [13,14]. Recently, Weber et al. [10] characterised the chromoviruses of *Beta vulgaris* bearing specific types of chromodomains, facilitating preferential insertion of this type of *Gypsy* elements into designated locations in a host genome.

The Brassicaceae is an economically important family of angiosperms. With the exception of the invaluable model plant *Ath*, many species in this family, including *Eutrema salsugineum* (*Esa*), *Schrenkiella parvula* (*Spa*), *Aly*, *Brassica rapa*, and *Capsella rubella*, have recently been sequenced or are scheduled to be sequenced [6,15–20], providing rich genomic resources suitable for comparative genomics studies. Most importantly, these species are native to highly diversified niches. For example, *Esa* and *Spa* are naturally salt-tolerant species that inhabit harsh environments, including soils of high salt concentration. A majority of LTR-RTs are recognised as transcriptionally quiescent during normal plant development, but are capable of strong activation by a variety of abiotic and biotic stresses [21,22]. Excessive transposition of LTR-RTs may damage genome integrity and lead to impaired gene function. To better understand the dynamics of LTR-RTs, host genomes and their co-evolution, we comprehensively identified and characterized the LTR-RTs of *Esa* and *Spa* along with two related salt-sensitive species, *Ath* and *Aly*. To identify the possible mechanisms underlying TE proliferation defence, we systematically compared the genomic composition, transposition activity, insertion patterns, genomic distribution and selective forces of LTR-RTs among the four genomes.

Results

Distinct activity of *Gypsy* elements in the four Brassicaceae species

The four wild species in the Brassicaceae family used for comparative analysis are of high relatedness, with over 90% of the protein-coding genes positioned in high genomic synteny, as illustrated by the comparison of different genomic elements between *Aly* and *Esa* (**Figure 1**). Because of their similar habitats and life cycles, these species are considered to be good model systems for investigation of miscellaneous non-genic elements in Brassicaceae. The genome sizes of the four Brassicaceae species in this study vary greatly: ~250 Mb for *Esa* (8 chromosomes), ~200 Mb for *Aly* (7 chromosomes), ~140 Mb for *Spa* (8 chromosomes), and ~125 Mb for *Ath* (5 chromosomes).

As genome size diversity in plants is primarily influenced by TEs, we compared the total content of repetitive sequences in the four species [23]. We found that *Esa* and *Spa* contained the highest and lowest proportion of repeats (51.4% and 18.5%), respectively (**Figure 2**A). Accordingly, full-length LTR-RTs, which have retained the paired LTRs at the two ends and may still possess transposition ability, are most abundant in *Esa* and least abundant in *Spa* (Figure 2B). By estimating TE insertion times, we found that the burst (estimated by med-

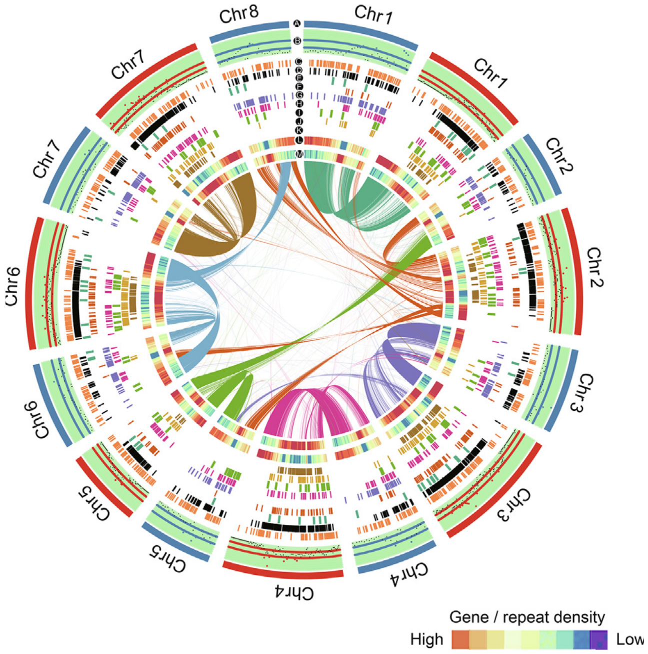

Figure 1 Genomic synteny between *Aly* and *Esa*
The annotations for the 13 tracks showing different genomic elements are: chromosomes of *Aly* and *Esa* represented by the outmost tracks colored in red and blue, respectively (**A**), hotspots of *Gypsy* element (**B**), *Copia* element (**C**), unclassified *Gypsy* element (**D**), g1-g7 families successively (**E**–**K**), gene density (**L**), and repeat density (**M**).

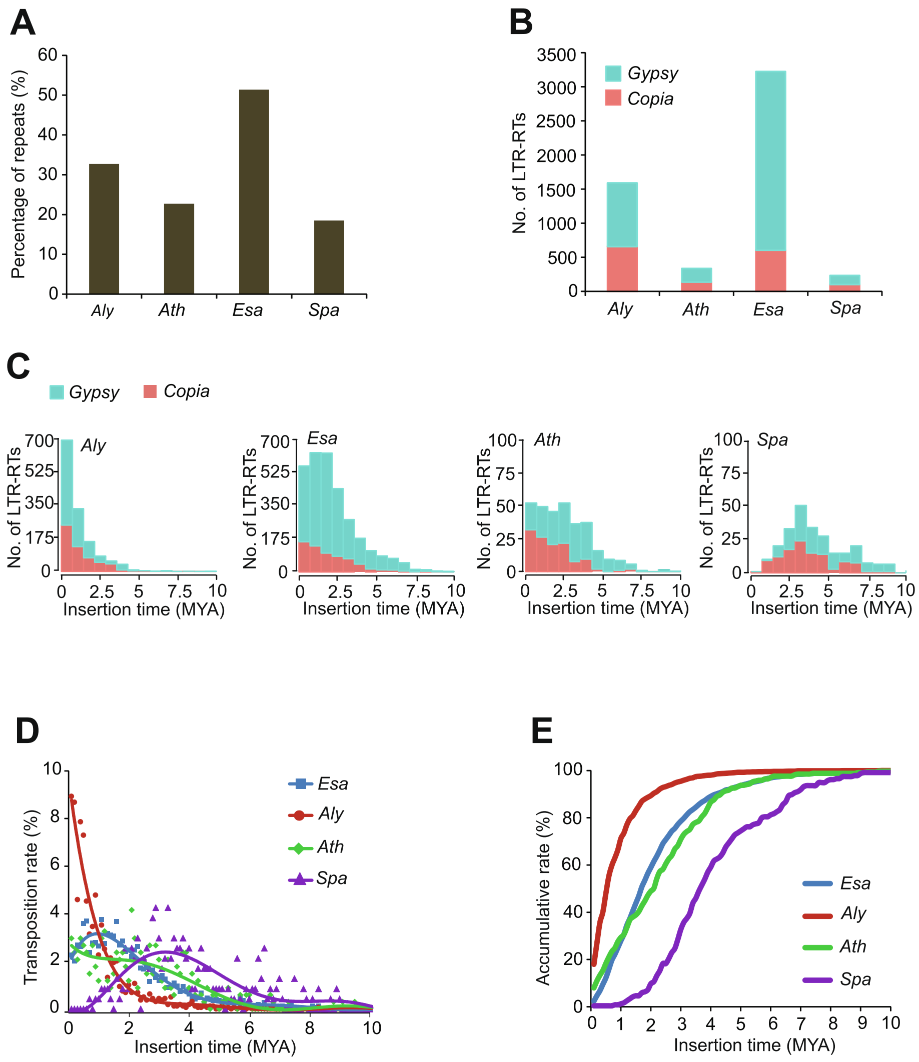

Figure 2 Different activity of LTR-RTs in the four genomes

A. Proportions of repetitive sequences in the four genomes. **B.** Number of full-length *Copia* and *Gypsy* elements in the four genomes. **C.** Distribution of insertion times of *Copia* and *Gypsy* elements in the four species. **D.** Transposition rates of LTR-RTs in the four species, defined as the ratio of the net increase of LTR-RTs every 0.1 MYs relative to the total LTR-RTs over a 10-MY time-scale. **E.** Accumulation rates of LTR-RTs in increments of 0.1 MYs over 10 MYs. LTR-RT, long terminal repeat retrotransposon; MY, million year; MYA, million years ago.

ian age) of LTR-RTs in *Esa* (~2.5 million years ago; MYA) occurred later than in *Spa* (~4.1 MYA), but earlier than in both *Aly* (~0.7 MYA) and *Ath* (~1.9 MYA) (Figure 2C). Additionally, the larger genomes of *Esa* and *Aly* are mainly attributable to *Gypsy* elements that might still actively transpose and are present at proportions significantly higher than those of *Copia* elements (Figure 2B). Whereas most insertions of LTR-RTs in *Esa*, *Ath*, and *Aly* are less than two million years (MYs) old, very few new insertions were found in *Spa*, suggesting very distinct evolutionary dynamics and mobile activity of these retrotransposable elements of the four species over the last few million years.

Genome size evolution, either expansion or shrinkage, is a dynamic process of DNA removal *versus* TE proliferation, with influence from, but not limited to, natural selection and inherent TE activity [23–25]. Although the exact direction of species evolution over a long evolutionary history is difficult to determine, the evolutionary tendency can be inferred from the most recent transposition events of certain TE classes. To compare TE activity, we calculated the transposition rate as the net increase in the number of LTR-RTs within every 0.1 MYs over a 10-MY period (Figure 2D). Within the last three MYs, the transposition rate of *Esa* has been relatively stable, whereas that of *Spa* has been in continuous decline. In comparison, the transposition rate of *Aly* has increased continuously, whereas that of *Ath* has remained relatively stable. The same tendency was also observed in terms of the accumulative rate of LTR-RTs within every 0.1 MYs over a 10-MY scale in the four genomes (Figure 2E).

Genomic distribution of the *Gypsy* elements in the four Brassicaceae species

We next compared the genomic distribution of full-length LTR-RTs among the four species and found that, whereas

Gypsy elements in *Ath* and *Spa* are frequently present in centromeres, they are strongly localized to gene-poor pericentromeric heterochromatin in *Esa* (**Figure 3**A and Figure S1). As the length of euchromatic portions are similar in *Esa* and *Ath*, the larger size of the *Esa* genome, half of which is comprised of pericentromeric heterochromatin, can be attributed to centromere expansion mediated by *Gypsy* proliferation in *Esa*. By comparison, *Gypsy* insertions in *Aly* appear more frequently in euchromatic regions without forming expanded pericentromeric heterochromatin (Figure S1). Additionally, the median distance between *Gypsy* insertions and their nearest genes is longer in *Esa* than in *Aly*, whereas no such differences for *Copia* insertions are apparent among the four species (Figure 3B). Strikingly, the median *Gypsy*-to-gene distance in *Spa* is 0 kb, indicating that most *Gypsy* elements overlap with genes. This finding seemingly contradicts the assumption that TE insertions in genes are more deleterious than intergenic insertions. To investigate this pattern, we added insertion times to the *Gypsy*/*Copia*-to-gene distance distribution in each species. Interestingly, each distribution exhibits primary and

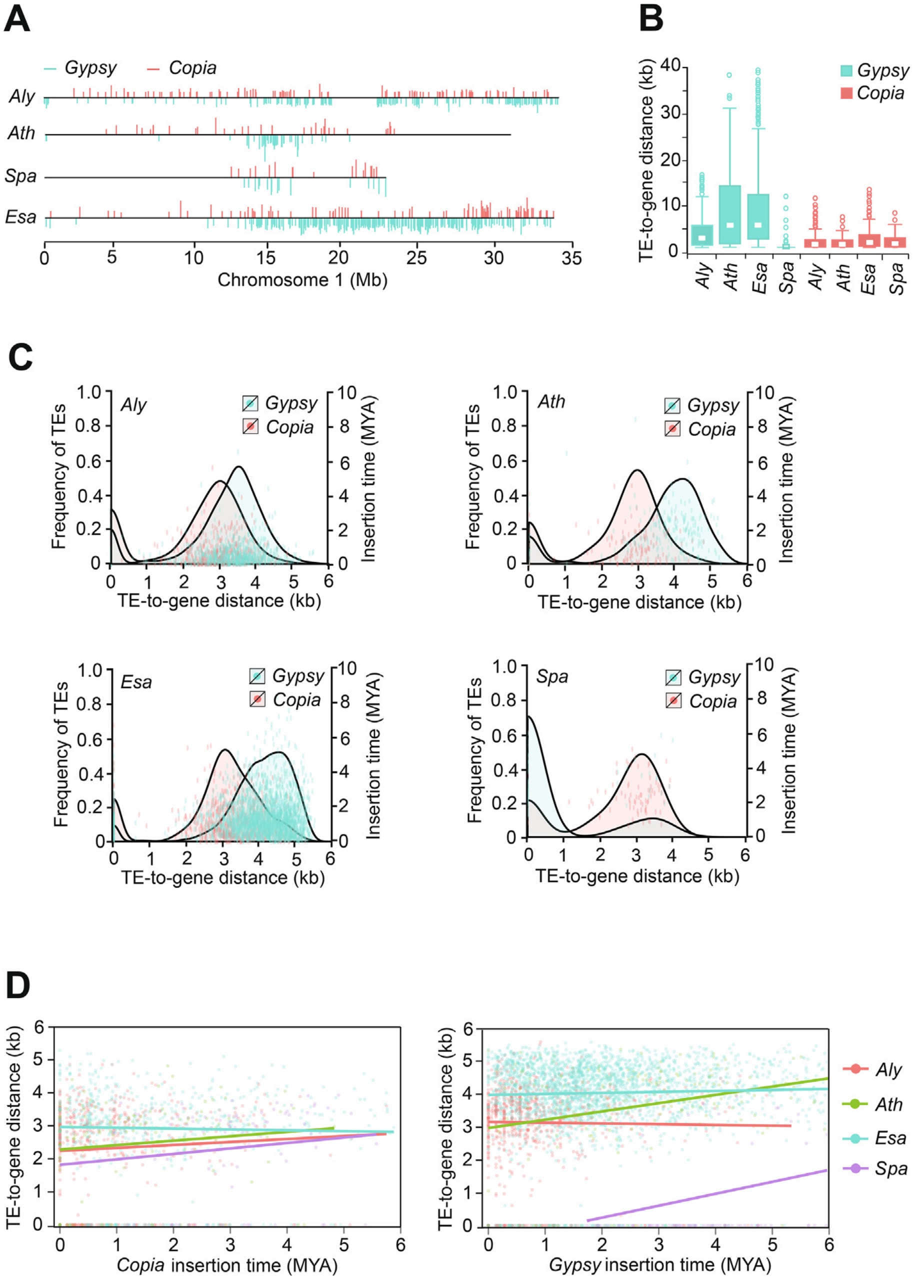

Figure 3 Genomic distribution of LTR-RTs in the four genomes

A. Distributions of *Gypsy* and *Copia* elements on chromosome 1 in the four species. The lengths of the vertical lines indicate the insertion ages: the longer the line, the older the insertion time. **B.** Distributions of *Gypsy*-to-gene and *Copia*-to-gene distances in the four species. **C.** Relationships of insertion times and insertion distances of the *Gypsy* or *Copia* elements to the nearest genes in the four species. TE-to-gene distances were $\log_{10}$ transformed. **D.** Linear regression analysis of insertion times and insertion distances of *Gypsy* and *Copia* elements. The TE-to-gene distanced were $\log_{10}$ transformed.

secondary peaks for both *Gypsy* and *Copia* elements (Figure 3C). In *Esa*, *Aly*, and *Ath*, the primary peaks represent intergenic insertions, and the secondary peaks, centred at 0 kb, are genic insertions that are predominantly younger (< 1 MYA) than the intergenic insertions. In *Spa*, whereas 82% of *Gypsy* elements overlap with genes forming the primary peak, 78% of *Copia*-like elements are located in the intergenic regions forming the secondary peak. The contrast between the depletion of intergenic *Gypsy* elements and the retention of *Copia* elements indicates that a strongly biased elimination of *Gypsy*-related sequences occurs in *Spa* through unknown mechanisms, leading to the overall suppression of *Gypsy* activity. This pattern supports our finding that *Spa* has likely been downsizing its genome over the last two million years.

To determine whether the longer *Gypsy*-to-gene distance in *Esa* is a result of natural selection or inherent transposition bias, we analyzed the correlation between insertion times and insertion distances (Figure 3D). The magnitudes of the slopes of the regression lines are 0.028, 0.033, 0.056 and 0.105 for *Aly*, *Esa*, *Ath*, and *Spa*, respectively. Whereas the most recent (< 1 MYs) *Gypsy* insertions in *Ath* are predominantly located near genes, at distances similar to or slightly shorter than those observed in *Aly*, there is an enrichment of older *Gypsy* insertions at distances far from genes in *Ath*. Interestingly, in *Esa*, we detected a statistically significant, weak correlation between the *Gypsy*-to-gene distance and insertion time ($P = 6.75 \times 10^{-3}$), and insertions in *Esa* are consistently farther away from genes than they are in *Aly*.

Preferential DNA removal of *Gypsy*-related sequences in *Spa*

The absence of intergenic *Gypsy* elements and lack of new insertions (< 2 MYs) in *Spa* suggest an active role for the host genome in aggressively prohibiting *Gypsy* proliferation, thereby leading to disproportionate predominance of *Copia* and *Gypsy* elements. However, further investigation is required to determine whether the observed depletion of *Gypsy* elements in *Spa* is caused by preferential DNA removal of *Gypsy* elements via an inherent molecular mechanism or purifying selection of *Gypsy* elements. DNA removal has been hypothesized to play a major role in preventing TE proliferation-mediated genome expansion [5,24,25]. Full-length LTR-RTs with a pair of identical direct repeats (paired-LTRs) are particularly favoured for DNA removal via unequal homologous recombination (HR) events, because the two LTRs provide homologous sequences to initiate illegitimate recombination [25–27]. Frequent HR-mediated DNA removal may result in a high abundance of solo-LTR remnants in the genome, which can be used as evidence to prove the existence of an inherently efficient DNA removal mechanism.

Thus, we compared the ratios of solo-LTRs *versus* paired-LTRs of *Copia* and *Gypsy* elements among the four genomes. The relative abundances of solo- and paired-LTRs were used to evaluate the propensity of HR-mediated removal of active LTR insertions in the four genomes. We found that the numbers of solo-LTRs of *Copia* elements were 2.98, 1.53, 6.92, and 1.60 times those of paired-LTRs in *Aly*, *Ath*, *Esa*, and *Spa*, respectively; whereas the corresponding ratios for the *Gypsy* elements were 5.09, 8.22, 8.03, and 17.76 times greater (**Table 1**). Because both *Gypsy* and *Copia* elements possess identical direct repeats, the intra-chromosomal recombination mechanism may not distinguish the two superfamilies of the LTRs. Thus, the disproportion of the solo- and paired-LTRs between *Gypsy* and *Copia* indicates that *Gypsy* elements were removed more efficiently than *Copia* elements in all four species, suggesting that the effect of *Gypsy* elements on individual survival is more deleterious than that of *Copia* elements. In addition, *Spa* has a much greater proportion of solo-LTRs of *Gypsy* in comparison with those of the other three species, while the ratio of solo-LTRs *versus* paired-LTRs in *Copia* elements was comparable between *Spa* and *Ath*. In summary, our results suggest that *Spa* might possess a highly efficient, inherent molecular mechanism to purge *Gypsy* elements, probably through HR-mediated DNA removal, to accelerate the processes of depressing deleterious, active *Gypsy* insertions.

Phylogenetic analysis showed diverse LTR-RT clades in the four species

Both *Gypsy* and *Copia* elements are composed of many clades that can be phylogenetically classified based on their reverse transcriptase (RT) sequences [8]. We first categorized all of the full-length LTR-RTs into three groups, namely "active", "inactive" and "fossil", based on the coding integrity of their RT domains. Active LTR-RTs most likely possess the capacity to transpose, while inactive LTR-RTs may have lost this capacity because of premature stop codons or frame-shift mutations in the RT domains; the remaining LTR-RTs are considered to be fossil LTR-RTs because their RT domains are severely fragmented. In all four species, inactive and fossil LTR-RTs are present in high proportions. *Aly* has the highest proportion (58%) of active LTR-RTs, while *Spa* has the lowest (13%); *Esa* and *Ath* have proportions of active LTR-RTs between those of the other two species (**Figure 4**A and B).

We further classified the group of active LTR-RTs into clades based on pairwise similarities in their RT domains using an affinity propagation clustering algorithm. It's worth noting that clade classification based on featured domain similarity is relatively a simplified means of classification, because different classes of LTR-RTs are usually embedded with each other,

Table 1 Comparison of paired- and solo-LTRs of *Copia* and *Gypsy* elements

	Copia elements				*Gypsy* elements			
	Aly	*Ath*	*Esa*	*Spa*	*Aly*	*Ath*	*Esa*	*Spa*
Solo-LTR	1957	207	4173	160	4750	1668	21,043	2398
Paired LTR	656	135	603	100	934	203	2621	135
Solo-LTR (%)	74.9	60.5	87.4	61.5	83.6	89.2	88.9	94.7
Solo-LTR *vs.* Paired-LTR	2.98	1.53	6.92	1.60	5.09	8.22	8.03	17.76

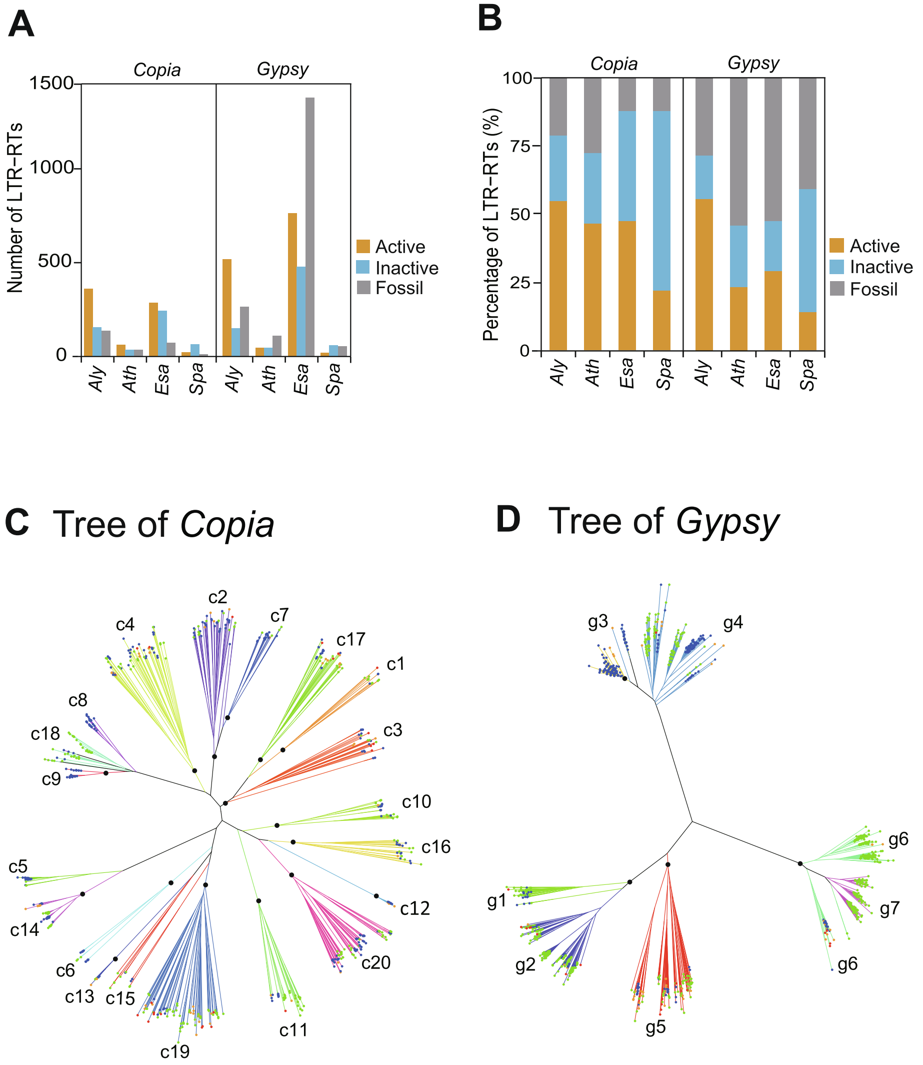

Figure 4 Phylogenetic classification of active LTR-RTs in the four species

A. Numbers of "active", "inactive" and "fossil" LTR-RTs in the four genomes. **B.** Percentages of "active", "inactive" and "fossil" LTR-RTs in the four genomes. **C.** and **D.** Neighbor-joining trees of the 20 clades of active *Copia* elements (C) and the 7 clades of active *Gypsy* elements (D). Leaf nodes in different colors denote the species origin of individual elements. Colored branches indicate the family identity. Major branch nodes at or near the most recent common ancestors of the 20 clades of *Copia* elements and the 7 clades of *Gypsy* elements are labelled with a black-filled circle if the associated bootstrap values exceed 80%.

leading to multiple returns from one query and making family/subfamily assignment difficult. As a result, 20 *Copia* clades (c1–c20) and 7 *Gypsy* clades (g1–g7) were generated among all of the LTR-RT sequences in the four species. These clades were subsequently superimposed, with high consistence onto a phylogenetic tree generated by the neighbor-joining (NJ) method, indicating that the LTR-RT classification is reliable (Figure 4C and D). These clades varied greatly in size: the numbers of elements range from 7 to 79 for the *Copia* clades and from 45 to 280 for the *Gypsy* clades. The LTR-RTs of the four species exhibited different compositions among the clades (**Figure** 5A). For example, *Gypsy* clade g3 and *Copia* clades c2, c5, c8, and c9 mainly occur in *Aly*, which is the youngest of the four species, whereas *Copia* clades c11, c18, and c19, as well as *Gypsy* clades g1, g6, and g7, dominantly occur in *Esa*, indicating species-specific proliferation of LTR-RT clades in *Aly* and *Esa*. In contrast, most of the *Gypsy* elements in *Ath* and *Spa* belong to the clade g5 and clade g2, respectively, indicating that the mobile activity of most LTR-RTs is suppressed in these two species (Figure 5A). The diverse compositions of LTR-RT clades in the four genomes were unexpected.

The average insertion time of all of the actively transposed members in a clade can be used to estimate the age of an LTR-RT clade. The standard deviation (SD) value of the insertion ages in a clade may indicate that the proliferation burst of these active members occurs in a short period time or occurs dispersedly over a long period. Interestingly, we found a significant correlation (Pearson's r = 0.725 and 0.728, P = 8.647 $\times 10^{-10}$ and 2.776 $\times 10^{-4}$ for *Copia* and *Gypsy* elements, respectively) between the SD and mean values of all clade ages (Figure 5B). However, this correlation between SD and mean values was only true for young clades (average age $<$ 2 MYs); a linear regression with an intercept fixed at zero indicated that

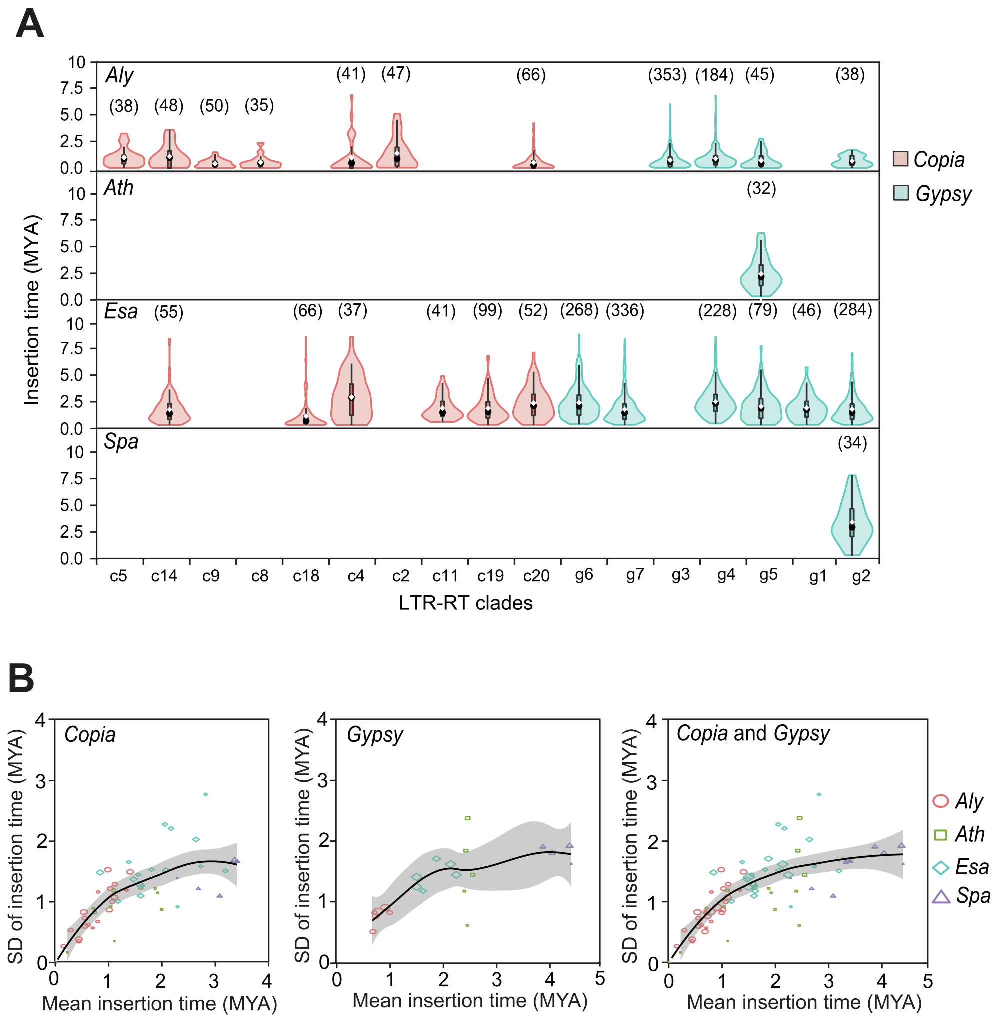

Figure 5 Estimation of the age of the LTR-RT clades in the four species

A. Violin plots representing the age of the 27 clades of LTR-RTs in the four species. Only clades containing more than 30 members are shown. Numbers in the brackets above each violin box indicate the numbers of LTR-RT elements. **B.** Relationship between the mean and standard deviation (SD) values of the age of an LTR-RT clade. The solid fitting curves were generated by a LOESS model, with the grey area showing the 95% confidence interval.

the mean value was nearly equal to the SD value (slope coefficient = 0.878 and 0.889, *F*-test: $P = 2.20 \times 10^{-16}$ and 2.36×10^{-9} for *Copia* and *Gypsy* elements, respectively). In contrast, most of the clades older than 2 MYs displayed a disproportionate drop in their SD values relative to the mean values of insertion time, as evidenced by the gradually flattened fitting curve (Figure 5B). The reason for this phenomenon might be that, comparing to young LTR-RTs, the old ones are more likely purged by negative selection, as mutations have accumulated in the pairs of LTR sequences that have a chance to escape from illegitimate recombination.

Species-specific *Gypsy* proliferation shaped the distinct genome architectures of *Aly* and *Esa*

After the RT domain, the INT domain is the second most critical component of functionally active LTR-RTs, because it plays an important role in directing an LTR-RT to integrate into specific locations of the host genome. We observed a more abundant enrichment of active *Gypsy* elements in *Esa* and *Aly* compared to *Copia* elements, as well as only specific *Gypsy* clades g5 and g2 enriched in *Ath* and *Spa*, respectively (Figure 5A). We speculated that the differences in the genomic distribution patterns of the *Gypsy* elements of *Esa* and *Aly* may be attributed to unknown differences in the features of the INT domains of the two species. To test our hypothesis, we first identified active *Gypsy* elements possessing full-length INT domains with complete coding capacity: 61.6% (319/518) in *Aly*, 27.7% (13/47) in *Ath*, 45.6% (348/763) in *Esa*, and 10.5% (2/19) in *Spa*. These ratios were consistent with the contents of young *Gypsy* insertions found in the four species. Next, we characterized a specific type of INT domain, the chromodomain (chromatin organization modifier domain), composed of a ~50 amino-acid motif that is a signature for chromovirus and can recognize specific genomic locations to be integrated in the host genome [10,14]. In plants, chromoviruses are categorised into five clades, the Tekay, CRM, Galadriel, Reina, and Tcn1 clades, which have different genomic integration behaviors [10,28]. We collected the sequences of the five known types of chromodomains as templates for the HMMER software package to predict the domains in LTR-RTs in the four species. In total, 46 INTs (including 21, 8, 16 and 1 in *Aly*, *Ath*, *Esa*, and *Spa*, respectively) showed the motif features of chromodomains. Interestingly, with only one exception, the 45 *Gypsy* elements bearing either the Tekay or Reina motif belong to the g5 clade (**Figure 6**A and B). However, because the C-terminal sequence of the chromodomain is extremely variable, a considerable number of chromodomains might have been overlooked.

In fact, chromoviruses of the CRM clade possess a signature chromodomain motif that has been reported to favour recognition of centromere sequences [5]. *Gypsy* elements carrying the chromodomain motif have a strong bias for insertion into centromeric and/or pericentromeric heterochromatin in plants [14]. Thus, we used the HMMER software package to build a HMM profile based on reported CRM *Gypsy* elements in plants to scan the 636 INTs again. The HMM search returned 78 results, including 21 in *Aly*, 0 in *Ath*, 56 in *Esa*, and 1 in *Spa*, showing the signature of a chromodomain motif (Figure 6C). Remarkably, all of the *Gypsy* elements in the results belong to the g2 clade, accounting for 69% of the members in this clade. However, it should be noted that the number of CRM *Gypsy* elements might be underestimated because of the fast-evolving nature of chromodomain motifs.

The genome of the hot pepper also has expanded pericentromeric heterochromatin and is thus four times larger than that of the tomato, a close relative. Kim et al.'s analysis showed that the genome structure of the hot pepper is mainly a result of highly abundant *Athila Gypsy* elements that preferentially accumulated in heterochromatin [9]. Inspired by this work, we collected more *Gypsy* sequences from the GyDB. Using the same HMM prediction procedure, we found that 605 of the 606 members of the g6 and g7 clades, which are specific to *Esa*, show significant similarity to *Athila Gypsy*

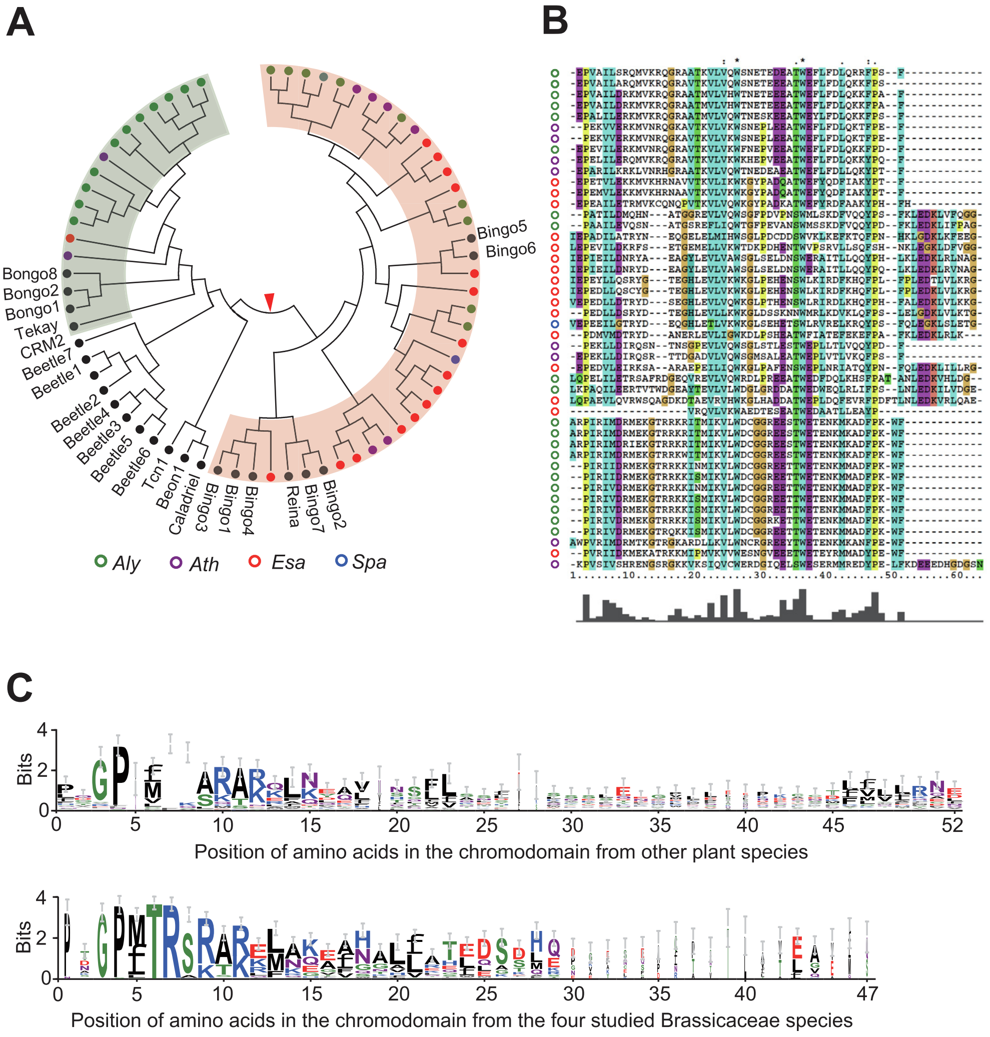

Figure 6 ***Gypsy* elements carrying chromodomains in the four species**

A. Phylogenetic tree of chromodomains across species. Colored nodes represent the chromodomains identified in the four Brassicaceae species. Black nodes with labels indicate previously reported sequences by Weber et al. [10]. The two shaded areas indicate that the sequences were from either Tekay (green) or Reina (orange) clade. **B.** Multiple sequence alignment of the chromodomains of *Gypsy* elements. **C.** Chromodomain motif representation. The upper logo indicates the prevalence of amino acids at specific positions of known chromodomain motifs from other plant species. The lower logo shows the composition of chromodomain motifs identified in the four studied species.

elements, which are completely absent from *Aly*. In contrast, the g3 clade, which is specific to *Aly*, shows significant similarity to *Tft2 Gypsy* elements belonging to the Tat family, which are mainly present in the euchromatic regions of the hot pepper genome [4,16,29]. Therefore, the distinct genomic architectures of *Aly* and *Esa* are highly likely to have been shaped by species-specific clades of active *Gypsy* elements with distinct INT domains causing preferential integration into euchromatic and heterochromatic regions, respectively.

Discussion

The activity of *Gypsy* elements reflects the tendency for genome size evolution

Comparative genomics provides an avenue to infer different evolutionary events during species evolution, such as birth and death of new genes that undergo sub-functionalization and pseudogenization processes, respectively [30]. With more and more plant genome sequences are completed, dynamics of genome evolution of different species is far more diverse than what we previously known in the plant kingdom. In addition to frequent polyploidization of higher plants, TE activity is another avenue contributing to the great diversity of plant genome sizes, especially for those closely related species sharing similar evolutionary history. As the coding portions are similar among plant species, the contexts and proportions of TEs, especially retrotransposable elements harbored in intergenic regions, exhibit great diversity. Our comprehensive analyses of LTR-RTs based on a comparison of four Brassicaceae species, in terms of phylogenetic classification, insertion age, transposition bias, and functional domains, indicate the existence of a strong correlation between the tendency for genome size evolution and the activity of *Gypsy* elements. Nevertheless, it's worth noting that, because the draft genomes of *Aly*, *Esa*, and *Spa* were assembled based on reads from next-generation sequencing, LTR-RTs might not be completely identified from current versions of the three assemblies. Especially for *Aly* and *Esa* genomes enriched with *Gypsy* LTR-RTs, misassembly of repeat-rich sequences may lead to a series of problems when estimating insertion time and performing clade classification. Therefore, our analysis was strictly focused on full-length LTR-RTs with recognizable paired LTRs, in order to minimize possible mistakes due to assembly errors. When new versions of the reference genomes of the three non-model organisms were released, analyses can be redone to validate the conclusion from the current study.

Recent interspecific comparisons in both *Oryza* and *Helianthus* genera similarly present a strong correlation between the activity of retrotransposable elements and genome size evolution [31,32]. Both work indicated that estimated insertion times of LTR-TEs may help infer evolutionary tendency, namely expansion or shrinkage of plant genomes [31,32]. Based on the rates of transposition and accumulation of LTR-RTs in the four genomes (Figure 2), and assuming that the rate of DNA loss counteracting TE insertions in *Ath* and *Esa* represents a common pace of genome evolution in the Brassicaceae family, it is very likely that the lack of LTR-RT accumulation in *Spa* over the last two MYs represents a severe depression of TE activity that has facilitated genome shrinkage. In sharp contrast, the *Aly* genome possesses much more abundant active LTR-RTs with young insertion ages, indicating a recent burst of LTR-RT proliferation within the past 1.5 MYs that led to rapid expansion of its genome. Our work demonstrates that, by analyzing the recent activity of LTR-RTs, particularly the superfamily of *Gypsy* elements, we may infer the tendency for genome size evolution in plants within a short period of evolutionary time.

Contrasting strategies utilized by two halophytes to reconcile stress-induced TE proliferation

In this work, we selected four fully sequenced species of high relatedness in the Brassicaceae family to investigate the correlation between genome architecture and TEs. While *Esa* and *Spa* are two natural salt-tolerant halophytic species that inhabit harsh environments that may have high salinity, *Ath* and *Aly* are salt-sensitive species. Severe abiotic stress can induce activation of LTR-RTs which might be the primary form of long non-coding RNAs functionally involved in stress response [21–33]. We initially expected to detect more abundant active LTR-RTs in salt-tolerant species *Esa* and *Spa* than in salt-sensitive species *Ath* and *Aly*. However, we failed to find such a correlation. Despite similar extremophile lifestyles, *Esa* and *Spa* display dramatic differences in TE content and activity. Nevertheless, this difference may imply that *Esa* and *Spa* might have adopted different molecular mechanisms to reconcile intensive TE transposition under severe environmental stress: *Spa* has seemingly adopted an active mechanism that rapidly purges deleterious *Gypsy* insertions via preferential DNA removal; *Esa* has seemingly adopted a passive mechanism that confines harmless *Gypsy* insertions to gene-poor heterochromatin regions. These distinct strategies of TE defence have resulted in a large *Esa* genome that has been subject to intensive centromere expansion, and a small *Spa* genome that has been subject to rapid genome downsizing.

Effect of purifying selection on *Gypsy* insertion in *Spa*

Genome-wide depletion of the *Gypsy* elements in *Spa* might be due to efficient DNA removal by innate molecular mechanisms or elimination of individuals carrying *Gypsy*-related sequences from the population by purifying selection. By comparing the ratios of solo-LTRs *versus* paired-LTRs of *Copia* and *Gypsy* elements among the four species, we provide evidence that a more efficient DNA removal mechanism mediated by illegitimate recombination may contribute to rapid elimination of *Gypsy* elements than it does to *Copia* elements, indicating the existence of an intrinsic molecular machinery monitoring and controlling *Gypsy* activities. However, the disproportionate composition and LTR-to-gene distances of *Gypsy* and *Copia* elements suggest that strong purifying selection may also play a role in eliminating *Gypsy* elements. A fundamental question is why *Gypsy* elements are more likely to be recognized by the innate DNA removal mechanism if both *Gypsy* and *Copia* elements possess paired identical, direct repeats. One possible explanation for this difference is that *Gypsy* insertions might be more toxic than *Copia* insertions to the host genome, and thus subject to strong purifying selection. Two lines of evidence may support this assumption. First, *Copia*-to-gene distance is much shorter than *Gypsy*-to-gene distance, suggesting that *Copia* insertion might be less deleterious than

Gypsy insertion, even if *Copia* elements are inserted very close to genes. Second, no significant correlation between *Copia*-to-gene distance and insertion age was observed in *Ath*, suggesting that *Copia* insertions close to genes might be less harmful than *Gypsy* insertions, as *Gypsy* insertions that occurred close to genes might have been removed by purifying selection. Therefore, the genome contraction process of *Spa*, which carries fewer *Gypsy* elements than *Esa,* might have been accelerated by strong selection necessitated by the requirement for greater fitness in a harsh environment. Another interesting finding from the analysis of *Spa* is that a group of *Gypsy* elements that successfully escaped either DNA removal or purifying selection were primarily young, genic insertions (less than 2 MYs) in protein-coding genes. These insertions may be neutral, less harmful, or even beneficial to genomic adaption to a stressful environment, as TE activity has been hypothesized to be a rapid method of creating raw genetic substrate for natural selection [34].

The transposition preference of specific *Gypsy* clades leads to centromere expansion in *Esa*

Studies on multiple plant genomes showed significant enrichment of TEs in pericentromere and centromere regions, mostly epigenetically silenced by DNA methylation and repressive histone marks [35,36]. Comparison of the four Brassicaceae genomes showed that *Esa* has a unique feature of expanded centromere and pericentromeric heterochromatin because of preferential transposition into gene-poor regions. The transposition bias of *Gypsy* elements in *Esa* is likely due to an innate mechanism of targeted integration, as reported previously for *Ath* [4,37]. This mechanism was inferred from the approximately one-order greater *Gypsy*-to-gene distances of young insertions (< 1 MYs) in *Esa* in comparison with those in *Aly*, as selection pressure may not have had an effect within a short period of evolutionary time [4]. However, although the majority of *Gypsy* insertions are also distal from genes in *Ath*, this characteristic is likely a result of more efficient elimination of individuals carrying *Gypsy* insertions close to genes than those carrying *Gypsy* insertions distal from genes under purifying selection; as apparent from the fitted trend line, the older a *Gypsy* element, the more likely is it to lie far from genes. Thus, our results provide evidence that natural selection has preferentially purged TEs located near genes more efficiently in *Ath*. Compared to *Ath* and *Aly*, another unique feature of *Esa* is lacking of *CHROMOMETHYLASE 3* (*CMT3*) gene responsible for gene body methylation, according to a recent report in *Esa* methylome analysis [38]. However, whether missing *CMT3* has a connection to hugely expanded pericentromere in *Esa* requires further study.

Another intriguing question is why centromere expansion was not observed in *Aly*, although the *Aly* genome possesses more active *Gypsy* elements with young age than does that of *Esa*. We reasoned that the distinct architectures of the two genomes are not a result of natural selection, but are instead most likely due to specific clades of *Gypsy* elements which contain different types of INT domains with different, preferential sites for transpositions. The g6 and g7 clades specifically found in *Esa*, which account for over 50% of the LTR-RTs in *Esa*, showed high homology to the *Athila* family, which has been previously reported to preferentially insert into heterochromatin [29]. In contrast, the g3 clade specific to *Aly* showed significant similarity to the *Tat* family which have been reported to be mainly harbored in the euchromatic regions [9].

Materials and methods

Identification of repeat sequences by RepeatMasker

To provide an unbiased estimate of the number of repetitive elements in the four species, RepeatModeler was first applied to construct *de novo* repeat libraries for the genomes of *Ath*, *Esa*, *Aly*, and *Spa*, respectively. The four *de novo* repeat libraries were then combined with the *Ath* repeat library, which was downloaded from RepBase. RepeatMasker was used to classify the repeats in the four species based on the combined library. RepeatMasker, RepBase and RepeatModeler were all obtained from (http://www.repeatmasker.org).

Identification of full-length LTR-RTs and solo-LTRs and estimation of insertion time

Full-length LTR-RTs with paired, near-identical, long terminal repeats were predicted using LTR-finder. Because LTR-RTs can frequently insert with each other, forming nested patterns, we only used the innermost LTR-RTs to infer insertion times. The insertion times of LTR-RTs were estimated using the Kimura two-parameter distance of paired LTR segments [39]. Specifically, the LTR pairs were aligned by MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/), and the evolutionary distances were computed by the DISTMAT program implemented in the EMBOSS package (http://emboss.sourceforge.net/). Evolutionary distances were converted into insertion times, assuming an equal neutral substitution rate of 7.0×10^{-9} per site per generation [40]. All LTR segments from the full-length LTR-RTs were used as queries to blast (e-value: 1.0×10^{-6}) against the genome sequences to identify homologous fragments. Next, we screened for solo-LTRs that did not overlap with any full-length LTR-RTs.

Categorization of active/inactive/fossil LTR-RTs

We first collected candidate reverse transcriptase (RT) sequences from multiple databases, followed by validation of the structure and completeness of the domain using SMART (http://smart.embl-heidelberg.de/). Then, we selected representative RT domain sequences of *Gypsy* and *Copia* elements to identify the RT regions of the detected full-length LTR-TEs by BLASTX search (e-value: 0.1). Finally, each RT domain was manually inspected for any signature that may cause loss of function, such as premature stop codons, frame-shifts, and large insertions and deletions. The LTR-RTs showing complete coding capacity were denoted as "active"; those showing loss-of-function signatures were denoted as "inactive"; the remaining LTR-RTs, which were severely fragmented, were denoted as "fossil".

Classification of active LTR-RTs into clades

The active LTR-RTs were clustered into clades based on a matrix derived from pairwise alignments of RT domain sequences. To measure the similarities among LTR-RTs, we first conducted all-to-all alignments of the LTR-RTs using the *needleall* command in the EMBOSS package. Then, for each pair of LTRs, within which one LTR at the 5′ end of the element was denoted "5′" and the other "3′", we calculated the normalized similarity score as $S_{5', 3'}/\min(S_{5', 5'}, S_{3', 3'})$, where $S_{5', 3'}$, $S_{5', 5'}$, and $S_{3', 3'}$ are the raw scores corresponding to sequence pairs 5′-LTR *versus* 3′-LTR, 5′-LTR *versus* 5′-LTR, and 3′-LTR *versus* 3′-LTR, whereas $\min(S_{5', 5'}, S_{3', 3'})$ denotes the minimal value of $S_{5', 5'}$ and $S_{3', 3'}$. Finally, we performed unsupervised clustering based on the score matrices (M) using the Affinity Propagation (AP) clustering algorithm with the *apcluster* function (negDistMat(r = 2), M, details = TRUE) of the apcluster package in R.

Phylogenetic tree construction

The amino acids of RT domains extracted from *Gypsy* and *Copia* elements were aligned by the MUSCLE program. The output files were imported into CLUSTALW to produce the neighbor-joining (NJ) trees. We ran RAxML to build a set of phylogenetic trees under a JTT + gamma model, with 100 rapid bootstraps employed to assess the branch reliability of the NJ tree.

CRediT author statement

Shi-Jian Zhang: Formal analysis, Writing - review & editing. **Lei Liu:** Resources, Writing - review & editing. **Ruolin Yang:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Xiangfeng Wang:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the "Innovation Project for the Postdoctoral Researchers" in China, and the "Open Research Fund Program" of Beijing Key Lab of Plant Resource Research and Development, Beijing Technology and Business University.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2018.07.009.

ORCID

0000-0003-4474-4709 (Shi-Jian Zhang)
0000-0003-3943-6381 (Lei Liu)
0000-0001-6241-6317 (Ruolin Yang)
0000-0002-6406-5597 (Xiangfeng Wang)

References

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

[2] Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 2007;41:331–68.

[3] Kumar A, Bennetzen JL. Plant retrotransposons. Annu Rev Genet 1999;33:479–532.

[4] Pereira V. Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol 2004;5:R79.

[5] Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res 2008;18:359–69.

[6] Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 2011;43:476–81.

[7] Jiang SY, Ramachandran S. Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. PLoS One 2013;8:e71118.

[8] Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res 2009;19:243–54.

[9] Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet 2014;46:270–8.

[10] Weber B, Heitkam T, Holtgrawe D, Weisshaar B, Minoche AE, Dohm JC, et al. Highly diverse chromoviruses of beta vulgaris are classified by chromodomains and chromosomal integration. Mob DNA 2013;4:8.

[11] Sandmeyer SB, Hansen LJ, Chalker DL. Integration specificity of retrotransposons and retroviruses. Annu Rev Genet 1990;24:491–518.

[12] Zhao M, Ma J. Co-evolution of plant LTR-retrotransposons and their host genomes. Protein Cell 2013;4:493–501.

[13] Slotkin RK. The epigenetic control of the *Athila* family of retrotransposons in *Arabidopsis*. Epigenetics 2010;5:483–90.

[14] Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E, Hobza R, et al. Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA 2011;2:4.

[15] Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, et al. The reference genome of the halophytic plant *Eutrema salsugineum*. Front Plant Sci 2013;4:46.

[16] Arabidopsis GI. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 2000;408:796–815.

[17] Wu HJ, Zhang ZH, Wang JY, Oh DH, Dassanayake M, Liu BH, et al. Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proc Natl Acad Sci U S A 2012;109:12219–24.

[18] Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, et al. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet 2011;43:913–8.

[19] Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet 2013;45:831–5.

[20] Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, et al. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 2011;43:1035–9.

[21] Wessler SR. Turned on by stress. Plant retrotransposons. Curr Biol 1996;6:959–61.

[22] Tittel-Elmer M, Bucher E, Broger L, Mathieu O, Paszkowski J, Vaillant I. Stress-induced activation of heterochromatic transcription. PLoS Genet 2010;6:e1001175.

[23] Wright SI, Agren JA. The *Arabidopsis lyrata* genome sequence sizing up *Arabidopsis* genome evolution. Heredity 2011;107:509–10.

[24] Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. Evidence for DNA loss as a determinant of genome size. Science 2000;287:1060–2.

[25] Hawkins JS, Proulx SR, Rapp RA, Wendel JF. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc Natl Acad Sci U S A 2009;106:17811–6.

[26] Devos KM, Brown JKM, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 2002;12:1075–9.

[27] Vitte C, Panaud O. Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. Mol Biol Evol 2003;20:528–40.

[28] Novikov A, Smyshlyaev G, Novikova O. Evolutionary history of LTR retrotransposon chromodomains in plants. Int J Plant Genomics 2012;2012:874743.

[29] Pelissier T, Tutois S, Tourmente S, Deragon JM, Picard G. DNA regions flanking the major *Arabidopsis thaliana* satellite are principally enriched in *Athila* retroelement sequences. Genetica 1996;97:141–51.

[30] Rutter MT, Cross KV, Van Woert PA. Birth, death and subfunctionalization in the *Arabidopsis* genome. Trends Plant Sci 2012;17:204–12.

[31] Zhang QJ, Gao LZ. Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA-genome *Oryza* species. G3 (Bethesda) 2017;7:1875–85.

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

[33] Wang J, Meng X, Dobrovolskaya OB, Orlov YL, Chen M. Non-coding RNAs and their roles in stress response in plants. Genomics Proteomics Bioinformatics 2017;15:301–12.

[34] Capy P, Gasperi G, Biemont C, Bazin C. Stress and transposable elements: co-evolution or useful parasites?. Heredity 2000;85:101–6.

[35] Maumus F, Quesneville H. Impact and insights from ancient repetitive elements in plant genomes. Curr Opin Plant Biol 2016;30:41–6.

[36] Maumus F, Quesneville H. Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. Nat Commun 2014;5:4104.

[37] Peterson-Burch BD, Nettleton D, Voytas DF. Genomic neighborhoods for *Arabidopsis retrotransposons*: a role for targeted integration in the distribution of the Metaviridae. Genome Biol 2004;5:R78.

[38] Bewick AJ, Ji LX, Niederhuth CE, Willing EM, Hofmeister BT, Shi XL, et al. On the origin and evolutionary consequences of gene body DNA methylation. Proc Natl Acad Sci U S A 2016;113:9111–6.

[39] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.

[40] Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 2010;327:92–4.