the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **112,** 172–192 doi: 10.1111/tpj.15938

Distinct composition and amplification dynamics of transposable elements in sacred lotus (*Nelumbo nucifera* Gaertn.)

Stefan Cerbin[1,†], Shujun Ou[1,‡], Yang Li[2], Yanni Sun[2] and Ning Jiang[1,*]

[1]*Department of Horticulture, Michigan State University, 1066 Bogue Street, East Lansing, MI 48824, USA,*

[2]*Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China*

Received 21 April 2022; revised 19 July 2022; accepted 8 August 2022; published online 12 August 2022.

*For correspondence (e-mail jiangn@msu.edu).

[†]Present address: Department of Ecology & Evolutionary Biology, University of Kansas, 1200 Sunnyside Avenue, Lawrence, KS, 66045, USA

[‡]Present address: Department of Computer Science, Johns Hopkins University, Baltimore, MD, 21218, USA

SUMMARY

Sacred lotus (*Nelumbo nucifera* Gaertn.) is a basal eudicot plant with a unique lifestyle, physiological features, and evolutionary characteristics. Here we report the unique profile of transposable elements (TEs) in the genome, using a manually curated repeat library. TEs account for 59% of the genome, and *hAT* (*Ac*/*Ds*) elements alone represent 8%, more than in any other known plant genome. About 18% of the lotus genome is comprised of *Copia* LTR retrotransposons, and over 25% of them are associated with non-canonical termini (non-TGCA). Such high abundance of non-canonical LTR retrotransposons has not been reported for any other organism. TEs are very abundant in genic regions, with retrotransposons enriched in introns and DNA transposons primarily in flanking regions of genes. The recent insertion of TEs in introns has led to significant intron size expansion, with a total of 200 Mb in the 28 455 genes. This is accompanied by declining TE activity in intergenic regions, suggesting distinct control efficacy of TE amplification in different genomic compartments. Despite the prevalence of TEs in genic regions, some genes are associated with fewer TEs, such as those involved in fruit ripening and stress responses. Other genes are enriched with TEs, and genes in epigenetic pathways are the most associated with TEs in introns, indicating a dynamic interaction between TEs and the host surveillance machinery. The dramatic differential abundance of TEs with genes involved in different biological processes as well as the variation of target preference of different TEs suggests the composition and activity of TEs influence the path of evolution.

Keywords: transposon, retrotransposon, target specificity, *Nelumbo nucifera*, intron, amplification, genes.

INTRODUCTION

The angiosperm lineage is the most dominant plant taxon containing as many as 400 000 species and ranks second to insects in species richness (Jarvis, 2007). The two major angiosperm groups, monocotyledons (monocots) and dicotyledons (dicots), diverged 150–130 million years ago (MYA) (Wikstrom et al., 2001). At present, the eudicot clade represents approximately 75% of the species in angiosperms (Drinnan et al., 1994). Among the dicot plants, sacred lotus (*Nelumbo nucifera*) occupies a key position in studies of angiosperm evolution. Lotus diverged from its closest sister lineage around 137–125 MYA (Wikstrom et al., 2001). Compared to the grape (*Vitis vinifera*) genome that diverged from its sister lineage 118–108 MYA (Wikstrom et al., 2001), the sacred lotus genome is a valuable addition to basal eudicot studies (Ming et al., 2013; Velasco et al., 2007). Phylogenetic comparisons between grape and sacred lotus suggest that sacred lotus is a better model for inferences about the common ancestors of eudicots (Ming et al., 2013). Further, genomic analysis revealed that the γ triplication event which occurred in all core eudicots did not occur in the sacred lotus genome and that the sacred lotus genome shows a remarkably low substitution rate and higher retention of duplicated genes compared with most other angiosperm genomes (Ming et al., 2013).

Sacred lotus is a land plant adapted to an aquatic environment, belonging to the Nelumbonaceae family, and is found throughout Asia and northern Australia (Han et al., 2007; Pan et al., 2010). It is cultivated as an ornamental and food crop; additionally, parts of the lotus plant such

as the flowers, roots, and rhizomes are used for medicinal purposes (Shen-Miller, 2002). Sacred lotus has the longest reported seed viability, up to 1300 years (Shen-Miller, 2002). The genome of the sacred lotus variety 'China Antique' was sequenced in 2013, using Illumina and 454 technologies (Ming et al., 2013), and was improved in 2018 through a linkage map and a BioNano optical map (Gui et al., 2018). Recently, the genome of sacred lotus was further improved through PacBio Sequel subreads (Li et al., 2021; Shi et al., 2020). The sacred lotus genome represents an excellent resource in the evolutionary analysis of eudicots and comparative studies between dicots and monocots. For simplicity, sacred lotus will be referred to as 'lotus' in the remainder of the manuscript.

Transposable elements (TEs) are genetic sequences first discovered over 70 years ago by Barbara McClintock (McClintock, 1950). TEs mobilize from one genomic location into another and in the process may increase their copy number. TEs are classified into two major groups based on the intermediate form of transposition: Class I or RNA retroelements, which transpose via RNA intermediates using a copy-and-paste mechanism, and Class II or DNA elements, which transpose via the excision of their DNA sequences called the cut-and-paste mechanism (Kapitonov & Jurka, 2008; Wicker et al., 2007). Based on their structural features, class I retroelements are further divided into two subclasses: long terminal repeat (LTR) retrotransposons and non-LTR retroelements, which include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (Kumar & Bennetzen, 1999). To date, eight superfamilies of DNA transposons have been identified in plants, including CACTA (*En*/*Spm*/DTC), *hAT* (*Ac*/*Ds*/DTA), *Helitron* (DHH), *Mutator*-like transposable elements (MULEs) (*Mutator*/DTM), *PIF*/*Harbinger* (*Tourist*/DTH), *Tc1*/*Mariner* (*Stowaway*/DTT), *Sola*, and *GingerRoot* (Bao et al., 2009; Cerbin et al., 2019; Wicker et al., 2007). Among these, *Sola* and *GingerRoot* are absent from angiosperm genomes analyzed to date. In addition, the coding capacity of elements for transposition machinery proteins allows for further element classification into autonomous elements, which code for these proteins, or non-autonomous elements, which rely on their cognate autonomous elements for movement within the genome.

Due to their capacity to multiply within a host and their prevalence among plant and animal genomes, TEs contribute significantly to increases in genome size (Agren & Wright, 2011; Bennetzen & Kellogg, 1997; Piegu et al., 2006; SanMiguel et al., 1998). In some instances, TEs may constitute the majority of the genome (Badouin et al., 2017; Bertioli et al., 2019; Mayer et al., 2012; Schnable et al., 2009; Song et al., 2021). Unlike genes, TE turnover is rapid at an evolutionary scale (Ma & Bennetzen, 2004; Wicker et al., 2018); therefore, TEs are only conserved in closely related species. As a result, dramatic differences exist in the content and diversity of TEs between organisms. While the genomes of mammals typically contain a high proportion of non-LTR retrotransposons (Lander et al., 2001; Waterston et al., 2002), the TE composition in arthropods is more heterogeneous (Wu & Lu, 2019). In contrast, LTR retrotransposons largely dominate the TE landscape in plants (Bennetzen & Wang, 2014; International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; Rensing et al., 2008; Schnable et al., 2009; VanBuren et al., 2018; Wicker et al., 2018; Zhou et al., 2021). The majority of LTR retrotransposons in plants are classified into two major superfamilies, *Copia* and *Gypsy*, depending on the arrangement of the genes in the *pol* region (Kumar & Bennetzen, 1999). In both major superfamilies, the canonical LTR found at the ends of these retroelements typically start with 5′-TG and end in CA-3′ (Kumar & Bennetzen, 1999), which forms a short inverted repeat. In fact, many computational tools use these criteria in *de novo* searches for LTRs (Ellinghaus et al., 2008; McCarthy & McDonald, 2003; Xu & Wang, 2007). However, there is a small subset of LTR retrotransposons with atypical ends rather than 5′-TG and CA-3′, called non-canonical or non-TGCA retroelements, accounting for about 1% of the total LTR retroelements in plants (Ou & Jiang, 2018). Despite their low abundance, non-TGCA retroelements have often been involved in gene and genome evolution due to their proximity to genes (Ou & Jiang, 2018). The best-characterized non-TGCA retrotransposon is *Tos17* (ending with TGGA), a *Copia* retrotransposon activated in tissue culture in rice (*Oryza sativa*) (Hirochika et al., 1996). *Tos17* specifically targets genic regions, which has been used to build a transposon tagging population for functional analysis of rice genes (Miyao et al., 2003). *Hopscotch*, a *Copia* retroelement ending with TACA, was discovered as an insertion into the maize (*Zea mays*) gene *Wx* (White et al., 1994). Subsequently, it was reported that an element related to *Hopscotch* was responsible for the elevated expression of *Tb1*, a major domestication gene in maize (Studer et al., 2011). In addition, three *Gypsy* retroelements, ending with TGCT, were detected from the soybean (*Glycine max*) genome (Du et al., 2010), suggesting elements with alternative ends could occur in both *Copia* and *Gypsy* retroelements.

Different compositions of TEs lead to distinct distribution patterns. For example, in humans, a large TE fraction is located in the introns of genes, leading to the expansion of genic regions (Lander et al., 2001; Sela et al., 2007). In plants, the prevalence of large introns is predominantly observed in gymnosperms with very large genome sizes of 10 Gb or more (Guan et al., 2016; Nystedt et al., 2013; Stival Sena et al., 2014; Voronova et al., 2020; Zimin et al., 2017). In contrast, most sequenced angiosperms have relatively compact genic regions. In the small Arabidopsis genome (125 Mb), most TEs are located in pericentromeric

regions (The Arabidopsis Genome Initiative, 2000). Maize has a relatively large genome (2.5 Gb) among angiosperm plants, where TEs account for 85% of the genomic DNA, yet maize TEs are mostly located in pericentromeric regions as well as intergenic heterochromatic islands (Hufford et al., 2021; Schnable et al., 2009). Due to its position in the eudicot phylogeny, lotus may offer important biological insights in terms of its TE content, structure, distribution, and diversity. Here we report the detailed results of a comprehensive computational analysis of the repetitive content in the assembled lotus genome. Our analysis reveals a unique landscape of TEs in lotus. Despite its moderate genome size (less than 1 Gb), the size of genic regions of lotus is much larger than that of most sequenced angiosperm species and comparable to that of gymnosperm species, with half of the gene space filled by TEs. Moreover, we reveal the dramatic difference among lotus genes in their association with TEs and the exceptional abundance of non-TGCA retroelements in lotus.

RESULTS

The content and diversity of TEs in the lotus genome

Characterization of repeats in lotus (cv. China Antique) was performed using the most recent assembly of 821 Mb (Li et al., 2021; Shi et al., 2020), which accounts for approximately 88% of the estimated lotus genome (929 Mb) (Diao et al., 2006). The N50 of the contigs is 1.8 Mb, and the LTR Assembly Index is 11. This indicates the assembly is of reference quality (Ou et al., 2018), so the intergenic regions are reasonably assembled. TEs were mined using a combination of structure-based and homology-based approaches (see the Experimental Procedures section), with manual curation to ensure accuracy. About 59% of the genome is composed of recognizable TEs, with nearly 400 000 copies (Table 1).

Based on genome coverage, the majority of recognizable TE sequences in lotus is contributed by retrotransposons (42% of the genome), a familiar phenomenon across the plant kingdom, where the amplification of LTR retroelements contributes to genome size expansion (Bennetzen et al., 2005; Vitte & Panaud, 2005). In lotus, the LTR retrotransposon content (35% of the genome) is comprised of slightly more *Copia* retroelements than *Gypsy* retroelements in terms of genomic fraction (18.5% versus 16.8%, Table 1) and copy numbers (37 831 versus 33 089, Table 1). Although this pattern is not unique, the *Gypsy* content is considerably higher than the *Copia* content in the majority of examined plant genomes (Table S1). Among the 93 genomes with an available *Gypsy*:*Copia* ratio, *Copia* content is higher than *Gypsy* content in only 13 (14.0%) genomes (Table S1). An alternative possibility is that *Gypsy* retroelements are enriched in sequencing gaps, so the apparent low abundance of those elements in lotus is an artifact of imperfect assembly. In addition, the lotus genome contains a relatively high coverage of non-LTR retrotransposons (6.5%), predominantly contributed by LINEs (Table 1). Only nine other plant genomes analyzed (9.7%) contain a higher fraction of non-LTR retrotransposons, ranging from 7.0% to 21.7% (Table S1). Taken together, these results suggest a higher net accumulation of non-LTR and *Copia* retrotransposons in the lotus genome compared to many other plants.

DNA elements comprise about 17% of the genome. This level of DNA TE content is notable, and only two other characterized genomes, those of rice (20%) and red bayberry (*Morella rubra*) (21%), contain more DNA TEs (Jia et al., 2019; Jiang & Panaud, 2013) (Table S2). Moreover, the copy number of DNA transposons is over twice that of retrotransposons (Table 1). The largest contributors are *hAT* elements, with over 122 000 copies accounting for 8% of the genome, which is the highest observed genomic

Table 1 The abundance of different superfamilies of TEs in lotus

Class	Subclass	Superfamily	Length (Mb)	Average element length (bp)	Copy number	Genomic fraction (%)	Percent of total copy number (%)
Class I	LTR	LTR/*Copia*	152.20	4023	37 831	18.53	9.47
		LTR/*Gypsy*	138.21	4177	33 089	16.83	8.29
		LTR/other	1.70	613	2772	0.21	0.69
	Non-LTR	LINE	50.82	2362	21 516	6.19	5.39
		SINE	2.31	156	14 803	0.28	3.71
		Total class I	345.24	3138	110 011	42.04	27.55
Class II	TIR	CACTA	1.41	2655	531	0.17	0.13
		hAT	65.81	536	122 726	8.01	30.73
		MULE	20.95	323	64 867	2.55	16.24
		PIF/*Harbinger*	25.95	324	80 090	3.16	20.06
	Non-TIR	*Helitron*	25.72	1220	21 085	3.13	5.28
		Total class II	139.84	483	289 299	17.02	72.45
Total TEs			485.08	1215	399 310	59.06	100

fraction of *hAT* elements in plants sequenced to date (Table S2). This is followed by *PIF*/*Harbinger* and *Helitron* elements (both occupying approximately 3% of the genome). Although the majority of DNA transposon families present in angiosperms are identified in lotus, the *Tc1*/*Mariner* superfamily is absent. The absence of *Tc1*/*Mariner* elements has been reported in 11 other plant genomes, including two basal angiosperms and grape (Table S2). In addition, CACTA elements are poorly represented (0.2% of the genome) in lotus. New active transposons can be introduced through horizontal transfer, preventing a TE family or superfamily from extinction (Schaack et al., 2010; Wallau et al., 2012). The absence of *Tc1*/*Mariner* elements from multiple plant genomes and the very low abundance of CACTA elements in lotus suggests that horizontal transfer events of both elements in plants are rare or unsuccessful, compared with the frequent horizontal transfer of *Tc1*/*Mariner* elements in animals (Loreto et al., 2008; Robertson & Lampe, 1995; Zhang, Peccoud, et al., 2020). This is also in contrast to LTR retrotransposons, which may have had 2 million horizontal transfer events in angiosperms (El Baidouri et al., 2014). Due to the very low copy number of CACTA and LTR retroelements with unknown classification (LTR/other, Table 1), these two groups of TEs are excluded from the subsequent comparative analysis for different genomic regions.

Abundance and diversity of *hAT* elements

DNA transposable elements belonging to the *hAT* superfamily are widespread in plant and animal genomes and have been widely used in gene tagging and functional genomics studies (Kunze & Weil, 2002; Sundaresan et al., 1995). Although widespread in plants, the contribution of *hAT* elements to genomic repeat content is typically low, as this is typical for most DNA transposons. Among the 85 plant genomes with *hAT* elements detected, these elements represent ≤1% of the genome for 54 (64%) plants (Table S2). The copy number of *hAT* elements (over 122 000) is the highest among all superfamilies of TEs, higher than the total number of copies of all retrotransposons detected (Table 1). In addition, the lotus genome contains about twice the number of *hAT* elements as that in blueberry, which has the second highest *hAT* fraction among all sequenced species (8.0% versus 4.4%, Table S2). There is a correlation between the abundance of *hAT* elements and the total DNA TE fraction among different genomes (Figure S1, Pearson Correlation $r = 0.45$, $P < 0.001$). However, if the *hAT* elements are excluded from total DNA TE content, the correlation between the abundance of *hAT* elements and other DNA transposons is not significant (Figure S1, $r = 0.19$, $P = 0.078$), suggesting independent amplification of different superfamilies of DNA transposons.

To evaluate whether specific families of *hAT* elements have expanded in the genome, the contribution of individual families was determined. Results indicate that overall, the high abundance of *hAT* elements in lotus was not due to the massive amplification of a single or a few families; instead, it was attributed to the amplification of numerous distinct families. Even though some elements are more abundant than others, the most amplified family (in terms of genomic fraction) contributes 0.3% of the genome, and the top 20 families comprise 3.3% of the genome. This contrasts with the LTR retrotransposons in maize, where the top 20 families contribute up to 70% of the maize genome (Baucom et al., 2009). Our repeat library contains a total of 239 families of *hAT* elements from the lotus genome, including 47 families encoding transposases and 192 families of non-autonomous elements. Among the top 20 families of *hAT* elements in lotus, only two potentially encode a transposase. The low abundance of families with transposase sources suggests that either many of these elements are no longer capable of transposition (ancient elements) or they are primarily non-autonomous elements without significant homology to the autonomous partner. As most *hAT* elements are non-autonomous, the average size of all *hAT* elements is only 536 bp (Table 1). This explains why *hAT* elements contribute a much smaller proportion of the genome than retrotransposons despite their higher copy number (Table 1).

To test whether diversity in *hAT* transposases may reflect their successful amplification in the lotus genome, phylogenetic analysis of the most conserved domain (motif 3, which contains the E region of the catalytic DD/E motif) of the *hAT* transposase among autonomous copies was performed (Kempken & Windhofer, 2001; Lazarow et al., 2012). Our analysis indicates substantial diversity among the *hAT* transposases found in the lotus genome, which contains autonomous *hAT*s from the two clades typically found in plants: *Ac*/*Tam3* and *Tag1* (Figure S2) (Kempken & Windhofer, 2001; Robertson, 2002). Despite the expansion of some lineages in lotus, the majority of the *hAT* elements with a recognizable motif 3 show a wide spectrum of diversity wherein various subgroups are more closely related to *hAT* proteins from other plant species than *hAT* transposases of the lotus genome. Overall, these results suggest that diversity within autonomous elements may have contributed to the success of the *hAT* superfamily in lotus.

LTR retroelements with non-canonical ends

Prior to this study, we screened for LTR retroelements with non-canonical ends (non-TGCA LTR retroelements) using an automated tool and revealed their presence in most (42 out of 50) of the sequenced plant genomes (Ou & Jiang, 2018). Nevertheless, overall those elements seemed to only account for about 1% of the total LTR retroelements (Ou & Jiang, 2018). In this study, the improved lotus assembly combined with manual curation of the repeat

library allows more accurate quantification of non-TGCA LTR retroelements.

In lotus, nine different non-canonical LTR ends were found, and the vast majority of retroelements with non-canonical ends (or non-TGCA) are *Copia* elements (Table 2, Table S3). The non-TGCA *Copia* retroelements collectively contribute 4.9% of the genome or over 25% of the total *Copia* retroelements (Table 2), suggesting non-TGCA elements could represent a considerable portion of the *Copia*-like retroelements. Among the eight groups of *Copia* retroelements with non-canonical ends, four (TGCT, TGGA, TACA, and TGTA) harbor mutations in one nucleotide compared to the canonical ends, and the remainder (TGGT, TACT, TATA, and TGTT) harbor two mutations. Those eight ends no longer form a short inverted repeat except TATA. Overall, the groups containing a single mutation are more abundant than those harboring two mutations (4.2% versus 0.7%; Table 2). In addition, variations are observed in mutations at the four sites: (i) no mutation was detected at the first nucleotide (always 'T'); (ii) the second nucleotide is a purine (G or A); (iii) the third nucleotide is the least constrained, and C, G, or T is observed; and (iv) for the last nucleotide 'A' or 'T' is observed. The most abundant non-canonical end type is found in retroelements starting with the canonical 5′-TG but ending in CT-3′ (referred to as TGCT LTR), where the most terminal nucleotide is not inverted. This LTR end type includes an estimated 6203 copies, making up 3.1% of the genome (Table 2). Consistent with our previous study (Ou & Jiang, 2018), the LTR regions of non-TGCA retroelements are about one third of the size of that of the canonical TGCA retroelements (316 bp versus 972 bp, Table 2). On the other hand, the internal region of non-TGCA retroelements is only slightly shorter than that of TGCA retroelements (4446 bp versus 4656 bp).

To determine the relationship between the non-canonical retroelements and canonical retroelements, a phylogenetic tree was constructed using the conserved catalytic domain of integrase of *Copia* LTR retroelements in lotus as well as *Copia* retroelements from other plant species. It is apparent that *Copia* retroelements fall into many clades, yet retroelements with the same type of ends are not always monophyletic (Figure 1). For example, the well-known rice *Tos17* retroelement (with TGGA end) is not clustered with any retroelement with the same end in lotus (Figure 1, red circle). If we consider branches with bootstrap value over 50%, only in one case does a lotus TACA LTR retroelement group with a grape retroelement with the same end (Figure 1, denoted with a green arrow). In addition, a retroelement from Arabidopsis is located on the same branch as two grape retroelements, and they all terminate with TATA (Figure 1, denoted with a purple arrow). This may indicate the termini of these two groups formed in the early stage of the dicot lineage. However, there are six groups of lotus retroelements with different ends clustering together (Figure 1, denoted with red arrows), including retroelements with the canonical end TGCA. This seems to imply that the majority of the non-canonical *Copia* retroelements in lotus may have a relatively recent origin.

The prevalence of TEs in genic regions

To study the distribution of TEs around genes, the gene annotation was downloaded from the Nelumbo Genome Database (Li et al., 2021). We developed a high-confidence gene set by filtering out potential TEs, very short genes, and truncated genes (see Experimental Procedures), leaving 28 455 genes in the dataset. We examined the insertion of TEs within genic regions (from transcription start site [TSS] to transcription termination site [TTS]) as well as 2-kb flanking sequences (upstream and downstream) of genes, resulting in a total of 328 Mb gene space analyzed. Nearly 146 000 TE copies were detected with 152 Mb of TE sequences in these regions, accounting for 37% of total TE

Table 2 The abundance of *Copia* LTR retrotransposons with different termini in lotus

Terminal sequence	Average LTR length (bp)	Average length of internal region (bp)	Copy number	Genome fraction (%)
TGCA	972	4656	28 063 (74.18)	13.648 (73.64)
TGCT	333	4562	6203 (16.40)	3.106 (18.73)
TGGA	266	4384	820 (2.17)	0.407 (2.20)
TGTA	288	4046	941 (2.49)	0.395 (2.13)
TGGT	341	4514	656 (1.73)	0.361 (1.95)
TACA	241	4194	548 (1.45)	0.302 (1.63)
TACT	331	4024	438 (1.16)	0.226 (1.22)
TATA	229	4147	154 (0.41)	0.084 (0.45)
TGTT	299	4249	9 (0.02)	0.004 (0.02)
Total non-TGCA	316	4446	9769 (25.82)	4.885 (26.36)

Numbers in parentheses indicate the percentage of all *Copia* retroelements.

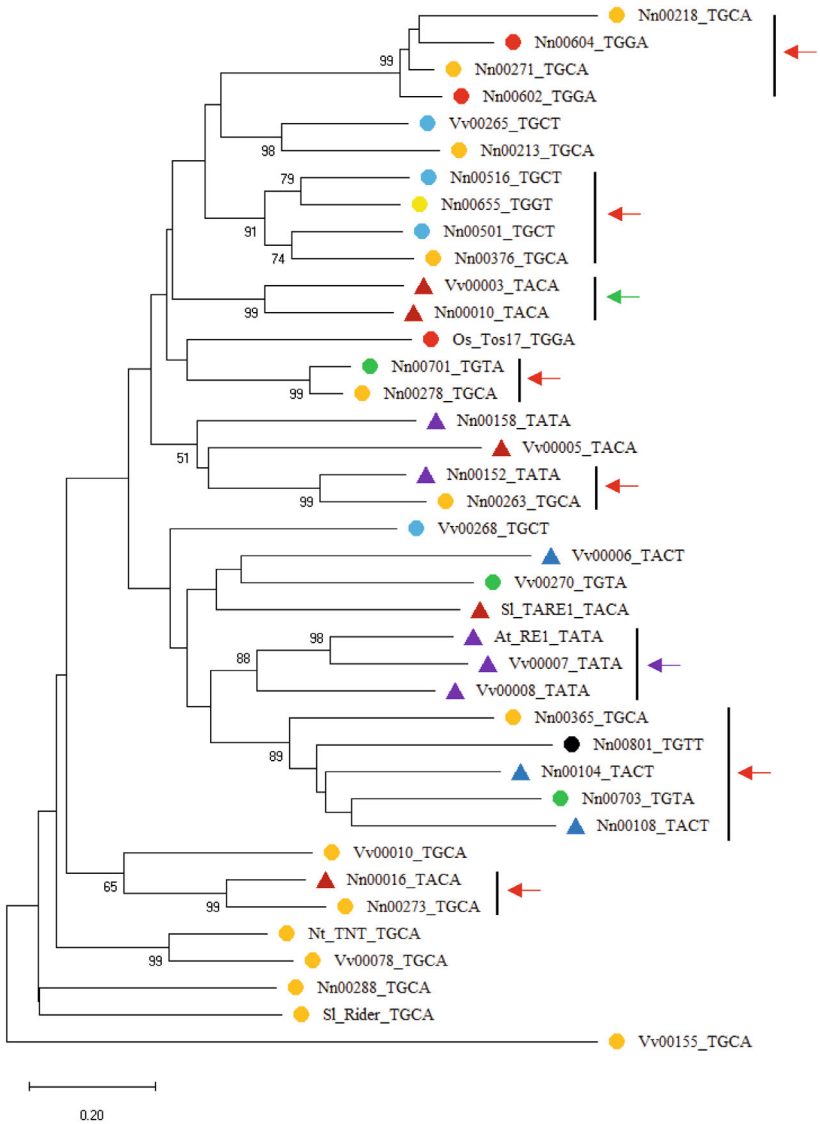

Figure 1. The phylogeny of the core integrase of *Copia* retroelements with different terminal motifs in lotus and other plants. Numbers next to branches indicate the % bootstrap support (1000 replicates, 50% cutoff). Retroelements with the same ends are labeled with the same color and shape. Retroelements starting with 'TG' are shown as dots, while elements starting with 'TA' are shown as triangles. Red arrows and vertical bars indicate branches containing lotus retroelements with different ends. A branch containing lotus and grape retroelements with 'TACA' motif is denoted by a green arrow and a vertical bar. A branch containing Arabidopsis and grape retroelements with 'TATA' motif is denoted by a purple arrow and a vertical bar. The tree is unrooted. Abbreviations for species: At, *Arabidopsis thaliana*; Gr, *Gossypium raimondii*; Nn, *Nelumbo nucifera*; Os, *Oryza sativa*; Sl, *Solanum lycopersicum*; Nt, *Nicotiana tabacum*; Vv, *Vitis vinifera*.

copies and 31% of total TE length, respectively. If we exclude the TEs in the 2-kb flanking regions, about 82 000 copies (20%) with 120 Mb (25%) of TEs are within 15 084 (53%) lotus genes, and the majority are in introns (97% of the TE copies and 99% of the TE length).

Since different families of TEs vary dramatically in element size (Table 1), a high number of TE insertions is not necessarily correlated with a longer total TE sequence in a certain region. To account for this discrepancy, we use two parameters, the insertion density (copy number per 100 kb) and the genomic fraction (%) of TEs, to indicate the abundance in different regions. Predictably, there are minimal TE insertions into exons (Figure 2, Table S4). The total intron size of lotus genes is about 200 Mb, which is larger than the entire Arabidopsis genome (The Arabidopsis Genome Initiative, 2000), and the average intron size is close to 2 kb (1988 bp). The TE insertion density within introns is slightly lower than the genome-wide level; however, the genomic fraction of TEs in introns is comparable to the genomic average (Figure 2), with approximately 60% of the intron sequences composed of TEs. Unlike all other genomic regions, which contain numerically more DNA TE insertions, the insertion density of the two classes of TEs in introns is similar (Figure 2a). As a result, retrotransposons are most abundant in introns with respect to both insertion density and genomic fraction (Figure 2). The TE density in the immediate 2-kb flanking regions of genes is the highest among all genomic regions, significantly higher than that of the genome-wide level (χ^2 test, $P < 1e{-}10$), yet the genomic fraction of TEs is much lower, due to the enrichment of small DNA TEs upstream and downstream of genes (Figure 2). Upstream of genes,

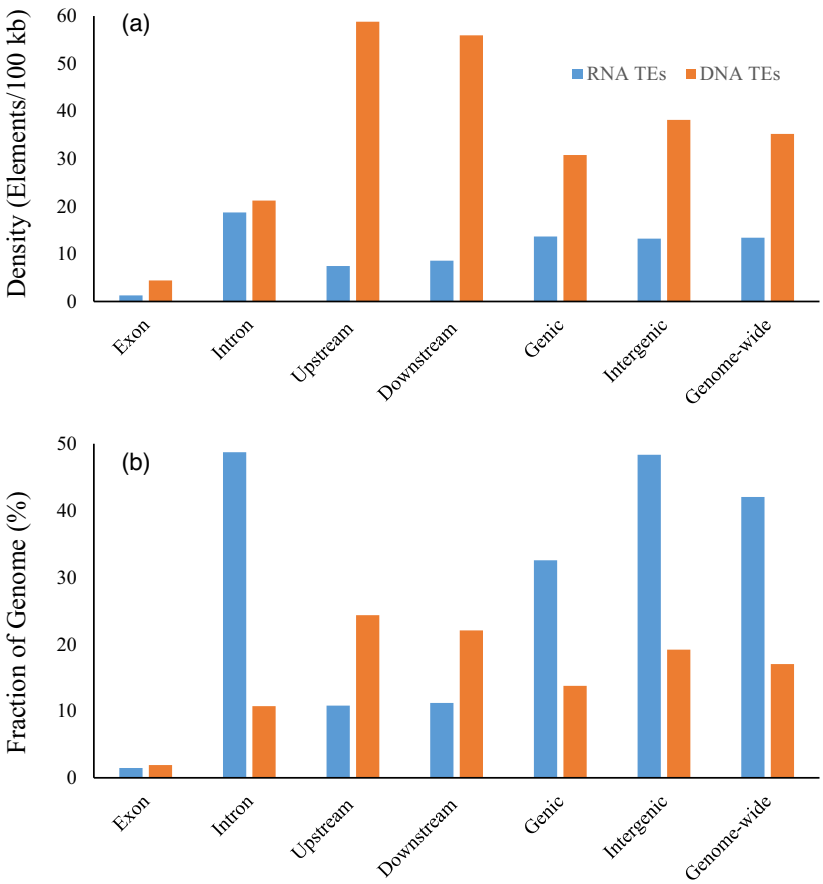

Figure 2. Overview of TE abundance indicated as insertion density (a) and genomic fraction (b) in different genomic regions. Upstream refers to regions 2 kb upstream of the transcription start site, and downstream refers to regions 2 kb downstream of the transcription termination site. Genic regions include upstream regions, downstream regions, exons, and introns. Intergenic refers to the genome excluding gene bodies and 2-kb flanking sequences.

25 341 out of 28 455 (89%) genes harbor TEs within 2 kb and the median distance of those genes to a TE is 511 bp (588 bp for all genes). The TE distribution in the downstream region is similar; 25 239 out of 28 455 (89%) genes harbor TEs within 2 kb and the median distance of those genes to a TE is 498 bp (578 bp for all genes).

Distinct niches for different superfamilies of TEs and selection pressure against large insertions in flanking regions and exons

As mentioned above, most TEs in genic regions are located in introns and flanking regions of genes. The flanking regions are associated with the highest TE insertion density in the genome while introns harbor the highest fraction of TEs among genic regions. The discrepancy between insertion density and genomic fraction is due to the distinct composition of TEs in these regions. To compare differences in TE families across genic regions we analyzed their genomic fraction as well as their enrichment index (EI), shown in Figure 3. We define EI as the percent difference between the genomic fraction (%) of each superfamily in each region and that at the genome-wide level. A positive EI value indicates higher TE abundance, whereas a negative value indicates an underrepresentation of TEs compared to the genome-wide average. Not surprisingly, the fraction of all TEs is very low in exons, with very negative EI values (Figure 3). In introns, the two largest genomic fractions are contributed by *Copia* LTR retrotransposons and LINEs (Figure 3a), and all retrotransposons are enriched except for *Gypsy* LTR retroelements (Figure 3b). The most enriched TEs in introns are LINEs (EI = 148), followed by non-canonical *Copia* retroelements (EI = 127), whereas canonical *Copia* retroelements are only slightly enriched (EI = 19). Overall introns contain 20% of the TE copies in the genome, yet 60% of LINE retroelements are located in introns. In contrast, all DNA TEs, particularly MULEs and *PIF*/*Harbinger* elements, are underrepresented in introns (Figure 3b). On the other hand, all DNA TEs are enriched in flanking regions except *Helitron*s, with MULEs being the most enriched (EI = 123) in upstream regions (Figure 3b). The distribution bias of *hAT* elements is not as strong as that for MULEs and *PIF*/*Harbinger* elements (Figure 3b), yet it contributes to the largest fraction of flanking sequences (Figure 3a) due to the sheer number of elements (Table 1). In the flanking regions of genes, all retrotransposons are underrepresented except SINEs (Figure 3b), which are enriched in downstream regions. As DNA elements and SINEs are both short (Table 1), this

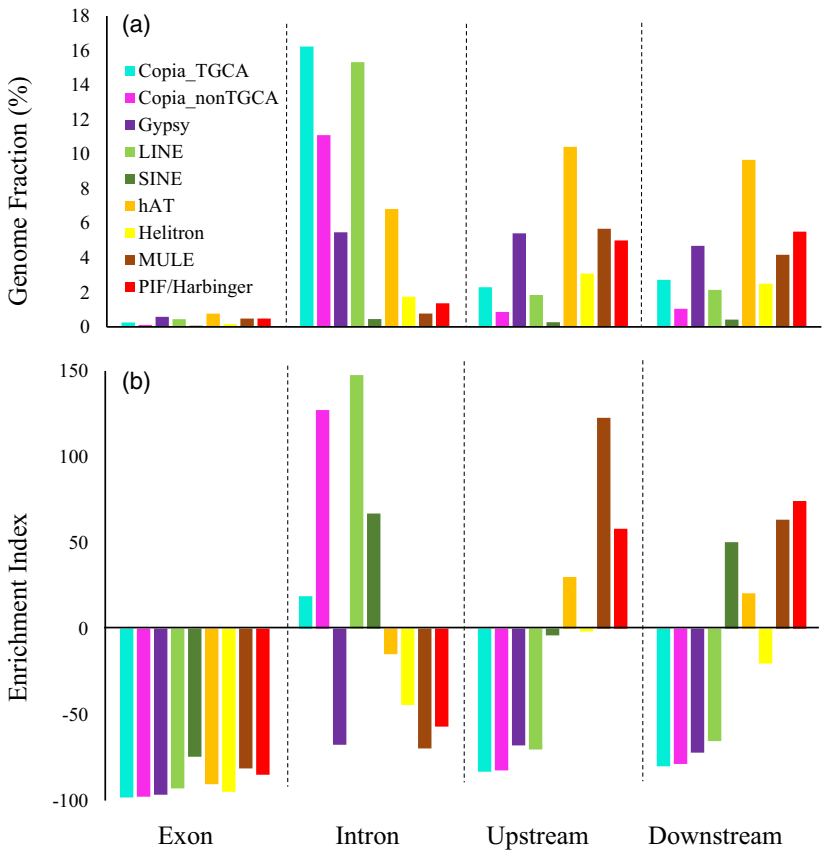

Figure 3. The abundance and enrichment of TE superfamilies in genic regions. (a) The composition of TEs (fraction of genome) in each region. (b) The enrichment/underrepresentation of different TE superfamilies in each region, reflected by the enrichment index, which represents the percent difference between the genomic fraction in each region and that at the genome-wide level. Upstream refers to 2-kb regions upstream of the transcription start site, and downstream refers to 2-kb regions downstream of the transcription termination site.

explains why the flanking regions of genes are associated with the highest insertion density in the genome but a much lower TE fraction than the genomic average (Figure 2). We also calculated EI using the insertion density (copies/100 kb), and the trend is highly similar to that based on the genomic fraction (Table S4).

As flanking regions of genes and introns harbor functional elements for gene expression, termination, and splicing, there is likely a certain level of selective pressure against TE insertions, particularly large insertions. To further explore the role of element size in the distribution of TEs, we conducted Pearson correlation analysis to test the effects of element size on enrichment in a certain region. There are significant negative correlations between EI and element size in downstream regions ($r = -0.93$, $P = 0.0003$), followed by upstream regions ($r = -0.84$, $P = 0.0042$) and exons ($r = -0.82$, $P = 0.0072$) (Figure S3). Nevertheless, there is no significant correlation between EI and element size in introns ($r = 0.31$, $P = 0.4185$, Figure S3). This suggests selection for small TEs may have played significant roles in the TE composition in downstream regions, and it also influences TE distribution in upstream regions and exons.

A summary of the skewed distribution of different superfamilies of TEs is provided in Table 3. Evidently, each superfamily of TEs has a unique distribution pattern. This is even true for LINEs and SINEs, which share transposition machinery (Singer, 1982). Among retrotransposons, LINEs and non-TGCA *Copia* retroelements have the highest specificity as they are only enriched in introns and are underrepresented in the remainder of the genome. No strong preference is observed for canonical *Copia* retroelements except they are underrepresented in flanking regions of genes. Among DNA transposons, MULEs demonstrate the strongest bias, with exceptional enrichment in upstream regions of genes, followed by downstream regions, and are the most underrepresented TEs in introns. *PIF/Harbinger* elements are also enriched in flanking regions but a preference is more evident in downstream regions. In contrast, *hAT* and *Helitron* elements have relatively minor preferences (Figure 3, Table 3), yet *hAT* elements are enriched in flanking regions of genes while *Helitron* elements are not. These distinct TE distribution patterns have shaped the different genomic regions in lotus.

Recent activity of LTR retrotransposons in introns and declining activity of *gypsy* elements in intergenic regions

To understand the dynamics of LTR retroelement amplification, the LTR identity of intact LTR retroelements in

Table 3 Summary of distribution preference of TEs in lotus[a]

TE group	Enrichment	Underrepresentation
LTR/*Copia*_TGCA	Intron and intergenic (weak)	Flanking region (very strong)
LTR/*Copia*_nonTGCA	Intron (very strong)	Flanking region (very strong) and intergenic region (strong)
LTR/*Gypsy*	Intergenic (strong)	Flanking region and intron (strong)
LINE	Intron (very strong)	Intergenic, upstream, and downstream (strong)
SINE	Downstream and intron (moderate)	Intergenic (strong)
hAT	Upstream and downstream (weak)	Intron (weak)
Helitron	Intergenic (moderate)	Intron (moderate) and downstream (weak)
MULE	Upstream (very strong), downstream (moderate)	Intron (strong)
PIF	Downstream and upstream (moderate)	Intron (strong)

[a]The copy numbers of CACTA and LTR/other are too low to evaluate preference. All TEs are significantly underrepresented in exon regions. The strength of preference or bias is based on the enrichment index value derived from the genomic fraction (Figure 3, Table S4). For genic regions: preference: 10–40 (weak), 40–80 (moderate), 80–120 (strong), >120 (very strong); bias: −10 to −25 (weak), −25 to −50 (moderate), −50 to −75 (strong), −75 to −100 (very strong). For intergenic regions: preference: 10–25 (moderate), >25 (strong); bias: −10 to −25 (moderate), <−25 (strong).

introns and intergenic regions was examined. LTR identity refers to the sequence identity between the 5′ LTR and 3′ LTR of an individual retroelement. The upstream and downstream regions of genes were not considered since the majority of full-length LTR retroelements are longer than 2 kb, so those regions harbor few intact retroelements. As shown in Figure 4(a), the distribution of intact *Copia* and *Gypsy* retroelements is complimentary as the majority of *Copia* retroelements are located in introns whereas *Gypsy* retroelements are largely located in intergenic regions, consistent with the above analysis using all retroelement-related sequences (Figure 3). Canonical *Copia* retroelements appear to be the youngest, with 50% of the retroelements associated with an LTR identity of >97%. This is followed by non-canonical *Copia* retroelements (33% of elements have an identity of >97%), and *Gypsy* retroelements are the oldest retroelements (17% of elements have an identity of >97%). The age of *Gypsy* retroelements is largely due to the older intergenic retroelements (Figure 4), whereas the distribution of *Gypsy* retroelements in introns is almost identical to that of non-TGCA *Copia* retroelements (Figure 4b). All retroelements in introns (solid lines) are much younger than those in intergenic regions (dashed lines in Figure 4b). Particularly, only 12% of the intergenic *Gypsy* retroelements fall into the 97–100% bin (compared to 40% in introns), suggesting the declining activity of *Gypsy* retroelements in intergenic regions. Combining all LTR retroelements, there is an enrichment of recent insertions (97–100% identity) into introns compared to intergenic regions (104 versus 29 per 10 Mb, 3.6-fold higher, Figure 4a). In contrast, there are nearly twice as many older insertions (<94% identity) into intergenic regions than into introns (54 versus 28 per 10 Mb, 1.9-fold higher, Figure 4a).

Since the non-TGCA *Copia* retrotransposons are older than their TGCA counterparts (Figure 4b), the question arises whether these non-canonical LTR-RTs are simply derived from post-transpositional mutation of the canonical TGCA LTR retroelements. If this is the case, one would not expect any non-TGCA LTR element with identical LTRs due to the time required for mutations. Given the average LTR length for non-TGCA retroelements is 316 bp (Table 2), the probability for an intact retroelement to accumulate mutations at both terminal sequences but not the remainder of the LTR is $1/(316 \times 316)$, or approximately 1 in 100 000 LTR retrotransposons with identical LTRs. In lotus, 129 retrotransposons with identical LTRs were detected, so one would expect 0.0013 [$129/(316 \times 316)$] non-canonical elements with identical LTRs. Nevertheless, 13 non-TGCA retroelements with identical LTRs were detected (Table S5), about 10 000 times the expected value ($P = 1.6 \times 10^{-14}$, χ^2 test). Moreover, one would expect more non-TGCA *Gypsy* retrotransposons than *Copia* retrotransposons since *Gypsy* retroelements are the oldest ones in the genome (Figure 4). However, non-TGCA *Copia* retroelements are much more abundant than non-TGCA *Gypsy* retroelements (9769 versus 184, Table S3). Whereas the above observation does not completely rule out the possibility that some of the old non-TGCA retroelements are mutated forms of the canonical elements, it is clear that the non-TGCA retroelements have evolved into a distinct group, with smaller LTRs, more closely associated with genes, and competent for transposition.

Genes involved in distinct biological processes are differentially enriched with TE insertions

TE target specificity is determined by a variety of genetic and epigenetic features, which vary from gene to gene. To investigate whether different genes have distinct insertion patterns by TEs, we examined the total TE abundance (length) in genes involved in different biological processes according to their GO terms. The abundance of TEs in

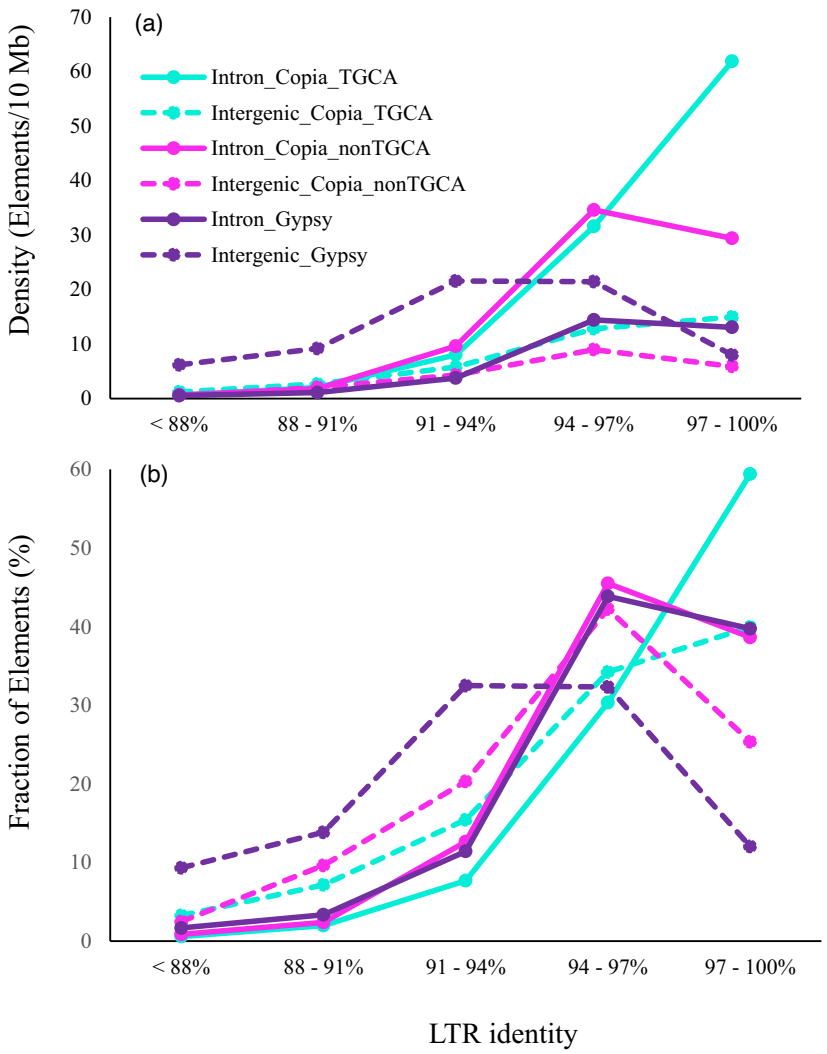

Figure 4. Comparison of abundance of intact LTR retrotransposons with different LTR sequence identity of individual intact retroelements in introns and intergenic regions. (a) The insertion density of retroelements in each identity range. (b) The percentage/fraction of retroelements in each identity range for each group of retroelements. Intergenic regions refer to the genome excluding gene bodies and 2-kb flanking sequences.

introns of different genes varies dramatically, with a 43-fold variation (from 333 bp to over 14 kb per gene, Table S6). Compared to the average level of TE abundance in all biological processes, TEs are underrepresented in genes involved in processes such as fruit ripening, abscission, secondary metabolic processes, pollen–pistil interaction, and responses to various stimuli (endogenous, biotic, abiotic, chemical, light) (Table S6). In contrast, TEs are enriched in genes involved in regulation of gene expression and epigenetics, DNA metabolic processes, cell-to-cell signaling, the cell cycle, etc. The number of introns only varies twofold in different categories of genes (compared to the 43-fold variation of TE abundance), and TE abundance is positively correlated with the number of introns, which is expected (Table S6). However, this relationship does not always hold. For example, genes involved in pollination have fewer introns than those involved in abscission (5.88 versus 6.32), yet the former have over twice the number of TEs (8.41 versus 3.07 kb per gene, Table S6) in introns. This suggests that intron number is not the sole factor that determines the abundance of intronic TEs.

Among the 50 GO terms examined, 26 are associated with significant enrichment/underrepresentation ($P < 0.01$ or $P < 0.001$, Kolmogorov–Smirnov test, Table S6) of TEs in introns, and the gene numbers in those GO terms range from 67 to 4579, with 18 (70%) GO terms containing >500 genes. To test whether the gene number in each GO term influenced the results, a simulation analysis was conducted. We randomly drew 67, 150, 346, and 1000 genes from the gene pool and analyzed TE enrichment/underrepresentation as described above for genes from a GO term. If the amount of TEs within this group of genes was significantly different ($P = 0.01$ or $P = 0.001$) from the actual data analyzed with genes in all biological processes, it was considered a false positive event. The experiment was repeated 1000 times. As shown in Figure S4, the putative false discovery rate (FDR) was indeed slightly higher with small gene numbers. Nonetheless, the overall FDR is very

low (FDR ≤ 0.021 for $P = 0.01$, FDR ≤ 0.004 for $P = 0.001$). The low FDR values and the fact that most GO terms contain over 500 genes suggest that the differential enrichment of TEs is unlikely an artifact of small sample sizes.

As mentioned above, genes involved in the regulation of gene expression and epigenetic pathways are the most enriched with TE insertions in introns (Table S6). Genes involved in this pathway are associated with more introns than average (9.27 versus 6.37, Table S6) as well as large coding regions (Zhao et al., 2017). One example of the genes in the epigenetic pathway is the *Dicer-like 3* gene (*DCL3*), which is known to be responsible for the generation of 24–26-nt small RNAs and silencing of TEs (Mari-Ordonez et al., 2013; Slotkin & Martienssen, 2007; Xie et al., 2004). As shown in Figure 5, the coding sequences of this gene are well conserved in Arabidopsis, grape, and lotus, yet the gene sizes vary more than 10-fold, largely due to the expansion of introns. The introns within the Arabidopsis *DCL3* gene are all very small (<500 bp) and lack TEs, and the gene is 7.3 kb in length. The grape gene is 28.6 kb, largely due to the expansion of four introns (Figure 5), caused by TE insertions. The largest contribution to the intron size in grape *DLC3* is from LINEs (6.0 kb) and a *Copia* retroelement (5.1 kb). The lotus *DCL3* gene is over 100 kb, with numerous TEs in 11 introns, and seven introns are longer than 5 kb. TE sequences account for 76% of the intron sequences of lotus *DCL3*, with the majority from retrotransposons (67% of the intron). The dramatic gene size variation of *DCL3* in the three species demonstrates that amplification of retrotransposons leads to a significant expansion of intron size. Compared with introns, there is little variation in total TE abundance in the flanking regions of genes that are involved in different biological processes (Figures S5 and S6).

Diverged preference for different superfamilies of TEs

In the above analysis, we analyzed the overall abundance of TEs in genes involved in different biological processes. To test whether each superfamily of TEs has specific preferences, a similar analysis was conducted with individual superfamilies. As shown in Figure 6, different superfamilies of TEs vary substantially in terms of the degree of enrichment/underrepresentation in introns of different genes. Again, we used an EI to quantify the degree of enrichment or underrepresentation. Here the EI refers to the percent difference between the average TE length in genes involved in an individual biological process compared with the average of all processes. Only genes involved in two biological processes are enriched in all superfamilies of TEs, with genes involved in two other biological processes associated with underrepresentation of every superfamily of TEs. Particularly, MULEs have unique preferences compared to other TEs. For instance, most TEs are enriched in genes involved in regulation of gene expression and epigenetic pathways (GO:0040029), with LINE retroelements being 147% enriched, whereas MULE elements are not significantly enriched (Figure 6). MULEs are the only elements enriched in genes involved in cellular homeostasis (GO:0019725) and the only elements underrepresented in genes involved in translation (GO:0006412) (Figure 6). *Copia* retroelements are the sole TEs enriched in genes related to pollen–pistil interaction (GO:0009875), with other elements underrepresented in these genes (Figure 6). *Gypsy* retroelements are the only TEs enriched in genes involved in response to extracellular stimulus (GO:0009991) and cell communication (GO:0007154). Although LINEs are much more enriched in introns than SINEs (Figure 3), the two superfamilies of TEs seem to have a common preference for genes in different

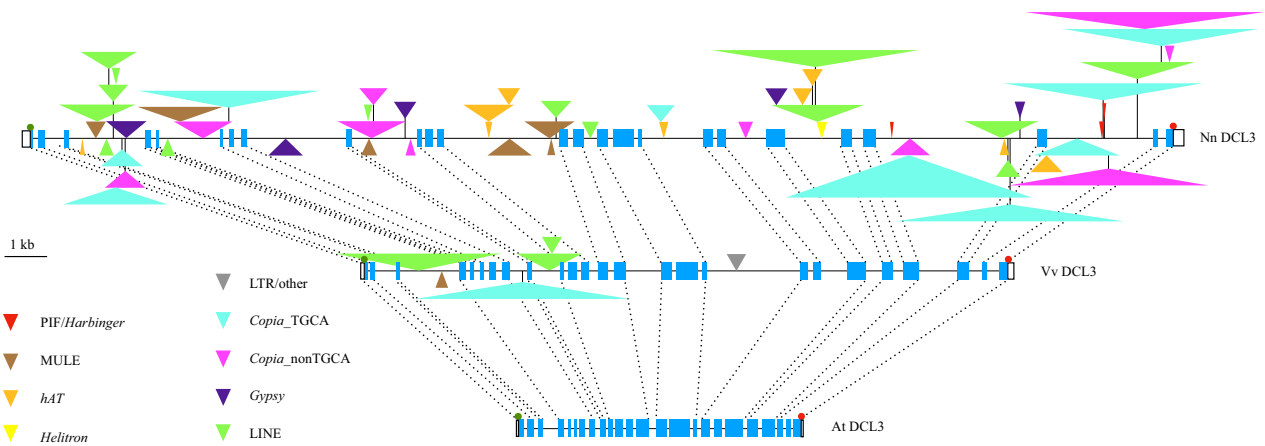

Figure 5. Comparison of the *Dicer-like 3* (*DCL3*) gene structure in Arabidopsis, grape, and sacred lotus. Top, Nn (*Nelumbo nucifera*) *DCL3* (107 kb); bottom, At (*Arabidopsis thaliana*) *DCL3* (7.3 kb, AT3G43920, TAIR10); middle, Vv (*Vitus vinifera*) *DCL3* (28.6 kb, GenBank accession No. NC_012010.3, 1 757 859–1 786 478, GeneID: 100254311). Blue boxes indicate exons of genes, white boxes indicate UTRs, triangles indicate transposons denoted by color, black and horizontal lines represent non-transposon intron sequences. Triangles stacked on top of other triangles signify a nested insertion. Green and red dots signify transcription start and stop sites of the genes, respectively. Dash lines of blue exons denote regions of homology of coding regions between grape and other species.

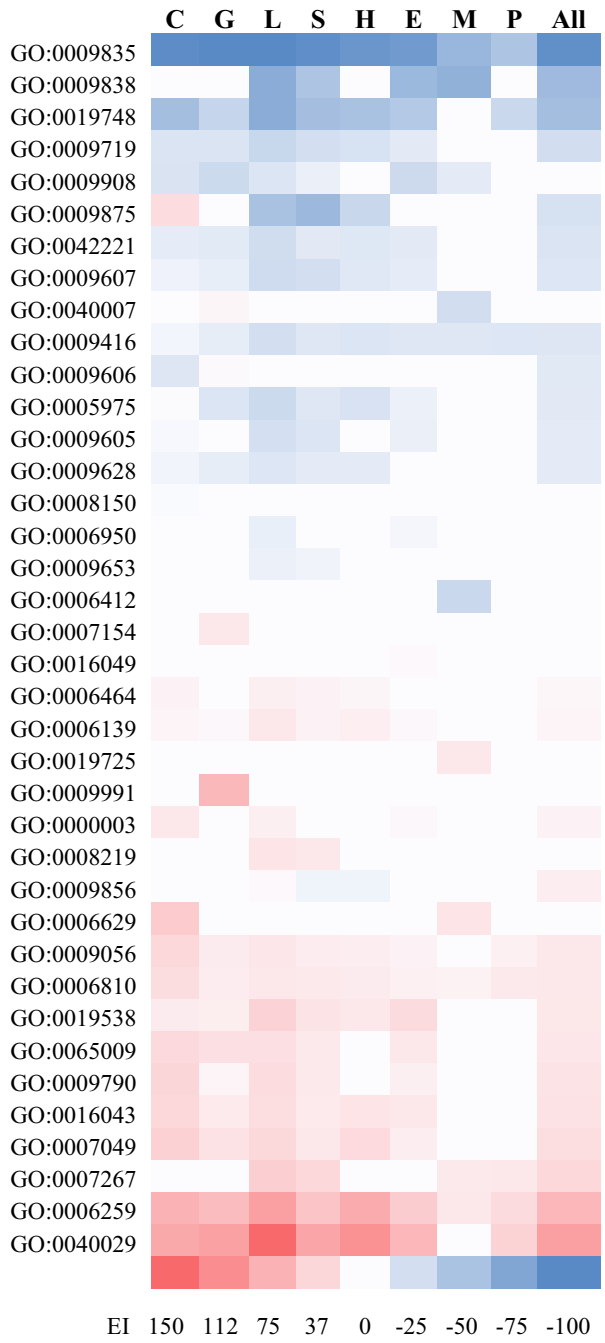

Figure 6. Heatmap of enrichment/underrepresentation of each superfamily of TEs in genes involved in different biological processes. Only significant enrichment/underrepresentation is shown in color. Abbreviations are shown on top. C, *Copia*; G, *Gypsy*; L, LINE; S, SINE; H, *hAT*; E, *Helitron*; M, MULE; P, *PIF*/*Harbinger*; All, all TEs. See Table S6 for GO terms.

biological processes (Figure 6), consistent with the notion that they share transposition machinery.

Although the total TE abundance does not vary dramatically in upstream or downstream regions of genes (see above), individual TE superfamilies exhibit insertion differences, particularly retrotransposons (Figures S5 and S6). Among those, *Gypsy* retroelements are more abundant than other retrotransposons in flanking regions (Figure 3a) and are enriched in the upstream regions of genes involved in pollen–pistil interaction (GO:0009875, EI = 85) (Figure S5). Interestingly, *Gypsy* retroelements are more enriched in the downstream regions of this group of genes (EI = 214, Figure S5), together with *Copia* retroelements (EI = 162, Figure S6). As mentioned above, *Copia* retroelements are also enriched in introns of genes involved in pollen–pistil interaction. Therefore, both gene bodies and the flanking regions of genes involved in pollen–pistil interaction are preferred targets for LTR retrotransposons in lotus.

DISCUSSION

In this study, we analyzed and characterized the TE content and diversity of lotus and showed that 59% of the genome sequence is composed of TEs. Although TEs in lotus are not as abundant as those in plants with larger genomes, the distribution, diversity, and amplification dynamics of some lotus TE families provide new insights into the co-evolution between TEs and other components in the genome. The high-quality repeat library generated by this study represents a useful resource for the community.

DNA transposons have replicated more substantially than retrotransposons in lotus

It has been well established that amplification of TEs, particularly LTR retrotransposons, is responsible for the expansion of plant genomes (Bennetzen & Wang, 2014). The contribution of retrotransposons to the genome size is attributed to their 'copy and paste' transposition mechanism, which leads to a rapid amplification of LTR retroelements (Michael, 2014). Nevertheless, the composition of TEs in lotus challenges the assumption that retrotransposons amplify more rapidly than DNA transposons. Similarly to other plant genomes, a large portion of the lotus genome is contributed by LTR retrotransposons. Solely considering the copy numbers, which reflect the efficacy of replication or retention, the copy number of DNA transposons is much higher than that of retrotransposons (Table 1). Retrotransposons occupy more genomic space due to a larger average element length than DNA transposons (3.14 kb versus 0.48 kb, Table 1), not due to higher copy numbers or more substantial amplification. The higher copy number of DNA transposons could be explained by the observation that small TEs are more competent for transposition (Zhao et al., 2015) or are more likely to be retained due to less deleterious consequences of their insertions. The lotus genome is not an outlier with a high copy number of DNA transposons. In a previous comparison of copy numbers and the genomic fractions of different TEs among 12 angiosperm plant species, five

species were found to have more DNA transposons than retrotransposons in terms of copy numbers (Kejnovsky et al., 2012), suggesting DNA transposons amplify as efficiently as retrotransposons.

The most abundant DNA elements in lotus belong to the *hAT* superfamily. A recent study indicated that the targeting of *hAT* elements is not as precise as MULEs (Zhang, Zhao, et al., 2020), which is consistent with our finding that, among DNA transposons, *hAT* and *Helitron*s are more uniformly distributed than others (Figure 3, Table 3). The relatively low specificity may confer more potential targets for insertions as an adaptive advantage. Unlike MULEs, which target highly expressed genes, *hAT* elements preferentially insert into moderately expressed genes (Zhang, Zhao, et al., 2020). Given that highly expressed genes are often subject to more selective constraints (Davidson et al., 2012; Drummond et al., 2005; Koonin & Wolf, 2010), targeting moderately expressed genes may favor element retention. Moreover, the intensity of epigenetic silencing of a TE family is positively correlated with the family copy number in plants (Cheng et al., 2006; Hirochika et al., 2000; Noreen et al., 2007); thus, the modest amplification of numerous individual *hAT* families instead of the dominance of a few families may prevent complete silencing of transposition activity. Taken together, the size, diversity, and target specificity of *hAT* elements may all contribute to the exceptional abundance of *hAT* elements in lotus.

Both classes of TEs occupy genic regions in lotus but are concentrated in different domains

According to their locations in plant genomes, TEs can be divided into two mutually exclusive groups: (i) one concentrated in large constitutive heterochromatic blocks found in the pericentromeric regions, knobs, and TE islands (heterochromatic TEs) and (ii) one that is found frequently near or inside genes (genic TEs). The activity of those two TE groups is regulated by distinct silencing mechanisms (Sigman & Slotkin, 2016). Among plants, the abundance of TEs in genic regions in the lotus genome is exceptional, including 25% by fraction (20% by copy number) within protein-coding genes, with 53% of the genes influenced. This contrasts with Arabidopsis, where only 3% of TEs (by copy number) are located within genes, including both protein-coding and RNA genes (Le et al., 2015). This is not simply attributed to the overall TE abundance in lotus since the maize genome harbors many more TEs than lotus but TEs are only located within a small subset (15%) of maize genes (Anderson et al., 2019). In addition, TEs in the flanking regions of lotus genes are closely adjacent to genes, with a median distance of approximately 500 bp in both upstream and downstream regions. As a comparison, only a small subset (22%) of Arabidopsis genes have a TE around genes and the median distance is 1089 bp to a DNA TE and 8.6 kb for retrotransposons (Hollister & Gaut, 2009). For bread wheat (*Triticum aestivum L.*), a species with an extraordinary abundance of TEs (85% of the genome), the median distance from a gene to an adjacent TE is 1.52 kb at the 5′ end and 1.55 kb at the 3′ end (Wicker et al., 2018). Accordingly, there is a high prevalence of TEs inside or adjacent to genes in lotus compared to either compact genomes (such as Arabidopsis) or expanded genomes (such as maize and wheat).

In the lotus genome, the variation in the abundance and composition of TEs among different genic regions (even excluding exons) is much more dramatic than that between genic and intergenic regions (Figure 2, Figure 3, Table S4). DNA TEs are predominant in flanking regions, whereas retrotransposons are enriched in introns, suggesting the composition and impact of 'genic TEs' depends on the exact genomic context. Due to selective forces, the distribution pattern does not always reflect TE target specificity. Nevertheless, the distribution pattern of some TEs in lotus seems to be in accordance with their target specificity. For example, the enrichment of MULEs at the upstream regions (Figure 3) is consistent with their target specificity at the 5′ end of genes (Dietrich et al., 2002; Jiang et al., 2011; Liu et al., 2009). Based on our analysis, it is likely that both flanking regions are associated with selection for small TEs, but the correlation between TE size and the degree of enrichment is much more significant in downstream regions than in upstream regions (Figure S3). This may suggest that the distribution pattern of TEs in downstream regions is largely shaped by the constraint for disruption of function (such as the integrity of the poly-A signal), while in upstream regions more targeting specificity by TEs is involved.

Previous studies based on model plants such as Arabidopsis and rice indicated that miniature inverted transposable elements (MITEs) are predominant in genic regions including introns (Bureau et al., 1996; Feschotte et al., 2002; Hua-Van et al., 2005; Kejnovsky et al., 2012). Lotus represents a counterexample to these previous observations with its large introns. Average introns in model plants are rather small (Wendel et al., 2002); accordingly, they contain limited amounts of TEs. Prevalence of large introns has been reported for several gymnosperm species, with giant genomes ranging from 10 to 31 Gb (Guan et al., 2016; Nystedt et al., 2013; Stival Sena et al., 2014; Voronova et al., 2020; Zimin et al., 2014; Zimin et al., 2017). In addition, a few angiosperm plants have been reported with large introns (average size 2 kb or longer), including *Phalaenopsis equestris*, *Liriodendron chinense*, and *Ceratophyllum demersum* (Cai et al., 2015; Chen et al., 2019; Yang et al., 2020), yet the composition and enrichment of TEs were not well studied in these species. In this study, we show that intron size could be significantly expanded in an angiosperm species with a moderate genome size of approximately 900 Mb. As a result, intron and

genome expansion does not have to be correlated with each other. The largest contributions to intron size are from *Copia* LTR retrotransposons and LINEs. Notably, DNA transposons are more underrepresented in introns than in any other regions (except exons) in the genome (Figure 2, Table S4), suggesting introns are unlikely the preferred targets for DNA transposons, at least in lotus.

The abundance of TEs in genic regions may be influenced by the asexual propagation and growth behavior of lotus. Our analysis indicates that elements in introns are much younger than those in intergenic regions (Figure 4), suggesting accumulation of TEs in introns is due to recent insertions, not because of low efficacy of exclusion compared to intergenic regions. Lotus propagates through rhizomes, which permits the genome to carry a masked deleterious TE insertion over time instead of being selected out immediately in the gametes. This provides more opportunity for a TE insertion to be retained. Certainly, a heterozygous TE insertion will require sexual reproduction to be fixed in the genome, and in general, selfing or inbreeding favors the retention of a TE insertion due to the small effective population size and limited recombination (Charlesworth & Charlesworth, 1995; Glemin et al., 2019; Wright et al., 2001). Lotus is considered to be an outcrossing plant as an individual lotus flower is not self-fertile because of a lag time between the maturation of stigmas and that of stamens in the same flower (Shen-Miller, 2002). Nonetheless, the rhizomes of lotus plants can generate secondary and tertiary rhizomes. Over time, a single plant forms a network of rhizomes, and their aerial apices, leaves, and flowers can occupy an entire pond (Shen-Miller, 2002). As a result, the probability of pollination between lotus flowers from the same plant is high, which favors the fixation of TE insertions even if they are slightly deleterious. This is consistent with the rather low heterozygosity (0.03%) of the lotus genome (Ming et al., 2013), suggesting the true outcrossing rate is low for lotus.

Distinct composition and amplification dynamics of TEs within a single genome

Given the possible selection against large elements in flanking regions of genes (Figure S3), one question is whether the complimentary enrichment of DNA transposons and retrotransposons in introns and flanking regions simply reflects the differential selection pressure against insertions in different regions. In this scenario, LINEs and *Copia* retroelements prefer genic regions, perhaps due to the accessibility of open chromatin, and uniformly target flanking sequences and introns. Thereafter, most of the insertions of LINEs and *Copia* retroelements in flanking regions are purged out due to their large size, whereas introns could tolerate large insertions so most are retained. According to this hypothesis and logic, *Gypsy* retroelements should be most underrepresented in flanking regions since they are slightly larger than *Copia* retroelements, yet they are more abundant in upstream regions than *Copia* and LINE retroelements (Figure 3). Moreover, the non-TGCA *Copia* retroelements are only slightly shorter than the TGCA *Copia* retroelements (Table 2) and the two types of retroelements are similarly underrepresented in the flanking regions of genes (Figure 3b), indicating similar selection pressure against these two groups of elements. Nevertheless, the non-TGCA retroelements are much more enriched in introns than the TGCA elements (Figure 3b), suggesting *Copia* retroelements with non-canonical ends preferentially target introns compared with their counterparts with canonical ends. This is consistent with the observation that *Tos17*, a non-TGCA *Copia* retroelement in rice, is enriched in gene bodies but not upstream regions (Zhang, Zhao, et al., 2020). Finally, SINEs are the smallest elements (Table 1) in the genome. If the observed distribution is simply due to selection in flanking regions, one would expect SINEs to be most abundant in flanking regions, yet SINEs are enriched in introns but not upstream of genes (Figure 3). Accordingly, while selection may play a role in the composition of TEs in flanking regions especially downstream regions (Figure S3), the enrichment of LINEs and non-TGCA *Copia* retroelements in introns likely reflects their target specificity. Moreover, it is likely TEs employ different strategies to target introns than flanking regions.

The comparison of retroelement ages between introns and intergenic regions represents another piece of evidence that the enrichment of LINEs and non-TGCA *Copia* retroelements in introns is unlikely due to retention. As shown in Figure 4, all LTR retroelements in introns are much younger than their counterparts in intergenic regions. Certainly, this observation does not exclude the possibility that intronic elements are retained longer than those in flanking regions, but it is evident that TEs in introns are not retained for an extended time, and the turnover of elements in introns is likely more rapid than that in intergenic regions (Figure 4). Moreover, it indicates that the overall TE activity could vary dramatically within a single genome; whereas introns harbor many recent insertions of LTR retroelements, few have been inserted into intergenic regions in the same time frame (Figure 4). The dramatic differential insertion density (3.6-fold) of recent elements (>97% LTR identity) between the intergenic regions and introns could have a profound impact on genome structure. Due to the prevalence of old elements in intergenic regions, it is possible that the TE retention rate is higher in intergenic regions than in genic regions (Figure 4). However, it is unclear whether the retention rate (if it is higher) in intergenic regions is sufficient to compensate for the low insertion frequency. As a consequence, if the differential insertion rate persists, the fraction of

intergenic regions might shrink and genic regions would further expand.

The differential amplification dynamics of LTR retroelements between intergenic and genic regions are consistent with the notion that the regulation of the activity of TEs is dependent on their genomic locations (Sigman & Slotkin, 2016). Apparently, intergenic retroelements experienced a certain level of amplification in the past (Figure 4a); thus, it is possible that the increased copy number of individual families has triggered or enhanced silencing mechanisms such as chromatin condensation in intergenic regions. Upon loss of activity, the intergenic/heterochromatic TEs are usually in a deep silenced status (Sigman & Slotkin, 2016); thus, they are less likely to be awakened by environmental factors than TEs in genic regions. The enhanced silencing for existing elements combined with a lack of new active TEs through true outcrossing and horizontal transfer may have resulted in attenuated TE activity in intergenic regions in lotus.

A burst of *Copia* retroelements with non-canonical ends in lotus

In this study, we detected eight different *Copia* retroelements with non-canonical ends (Table 2, Table S3) comprising 26% of the total *Copia* retroelements. Such an abundance of LTR elements with non-canonical ends has not been reported for any other organisms. Certainly, this is related to the relatively high abundance of *Copia* retroelements in lotus. However, there are 12 additional plant genomes associated with a higher fraction of *Copia* retroelements than lotus (Table S1), so an abundance of *Copia* retroelements is not always associated with the prevalence of retroelements with non-canonical ends. Alternatively, the termini of the retroelements in those species were not carefully examined.

If the presence of retroelements with non-canonical ends is due to the long-term co-evolution between the retroelements and the transposition machinery (Du et al., 1997), it is possible that some of the ancient lineages of *Copia* retroelements coding for integrases have higher affinities to non-TGCA ends than to TGCA ends. If that is the case, one would expect retroelements to group monophyletically with similar elements in other species, and it does occur in two cases (Figure 1), so these two groups of non-TGCA retroelements may have ancient origins. However, more commonly lotus retroelements with different ends cluster together, including some TGCA retroelements. This may indicate that most non-TGCA retroelements in lotus are derived from relatively recent mutations and have been maintained since. Furthermore, it is known that sequence swapping occurs among related LTR retrotransposons (Du et al., 2010); this precludes us from rejecting the possibility that sequence swapping is responsible for the intermingled phylogeny of retroelements with different ends in lotus.

The abundance and composition of TEs may influence the direction of evolution

Lotus TEs show variation in composition and amplification dynamics in different genomic regions. Furthermore, these TEs show various preferences for the genomic environments of genes in different biological processes. This implies that genes involved in a certain biological process are more permissive to TE insertions and that the ultimate insertion spectrum depends on which TEs are active. If *Copia* elements are mobilized, genes related to pollen–pistil interaction (GO:0009875) would likely be targeted (Figure 6). If *Gypsy* retroelements are activated, they are preferentially inserted into genes involved in response to extracellular stimulus (GO:0009991) and cell communication (GO:0007154). In the case of active LINEs, genes involved in the epigenetic pathway would be predicted to accumulate the most insertions (Figure 6).

TEs in genic regions could have genetic and epigenetic impacts. Insertions in coding sequences (CDSs) are high-probability candidates for deleterious effects. Even if the insertion is located in a non-coding region, it is often consequential. For example, untranslated regions (UTRs) may contain motifs important for transcription or translation (Juntawong et al., 2014; Srivastava et al., 2018), and a TE insertion may disrupt such regulatory elements. TEs upstream (where promoters of genes are located) may influence the transcription of the genes. In Arabidopsis, the average gene promoter is within 500 bp from the TSS (Korkuc et al., 2014). As the median distance from TEs to the TSS is about 500 bp in lotus (see the Results section), it implies that there are one or more TE insertions in promoter regions of a large portion of lotus genes, which may influence the function of these promoters. TEs in introns may interfere with splicing, and a well-known example is the LINE retroelement *Karma* in African oil palm (*Elaeis guineensis*). This retroelement is located in a large intron of an *AP3*-like B-class MADS-box gene in the flowering pathway (Ong-Abdullah et al., 2015). When *Karma* is hypomethylated, an alternative acceptor site inside the retroelement is utilized, which causes mis-splicing of the gene and leads to infertile fruits. In humans and *Caenorhabditis elegans*, longer introns are associated with a reduced level of expression (Castillo-Davis et al., 2002). In plants, the relationship between intron size and expression level is controversial. In Arabidopsis, rice, and *Picea glauca*, highly expressed genes have longer intron sequences (Ren et al., 2006; Stival Sena et al., 2014). In contrast, grape genes with large introns demonstrate reduced expression levels compared to average genes (Jiang & Goertzen, 2011). In *Phalaenopsis equestris*, the expression levels of genes with TE insertions were lower than those of their paralogs (Cai et al., 2015). Despite the discrepancy, it is apparent that TE insertions in introns

may influence both the quantity and structure of transcripts. Furthermore, a recent study indicated that genes with more and longer introns are associated with lower mutation rates in Arabidopsis (Monroe et al., 2022), so TE insertions in introns may have a long-term impact on gene evolution.

DNA methylation is one of the most important means to control TE activity. In Arabidopsis, there is a negative correlation between gene expression level and the density of methylated TEs (Hollister & Gaut, 2009). Accordingly, there is a 'trade-off' between the control of TE activity and gene expression (Choi & Lee, 2020). If we assume an additive relationship of TE insertions to their impact on genes, it appears that genes in different biological processes are influenced to a different degree. TEs are enriched within genes involved in regulation of gene expression, epigenetics, DNA metabolic processes, and the cell cycle. Some features (such as expression level or epigenetic status) of those genes may make them more attractive for TE targeting. Alternatively, these genes could be more tolerant to TE insertions given the presence of more and longer introns. The underlying mechanism for differential TE abundance among different genes requires further investigation. Ironically, genes involved in epigenetic pathways and DNA metabolic processes are the key to genome integrity, yet it seems they are poor 'sentinels' for themselves. If TEs preferentially target the genetic or epigenetic features associated with those genes, this implies that TEs are not only controlled by the genome surveillance machine, but may co-evolve. On the other hand, if selection leads to genes harboring large amounts of TE insertions, it may suggest the activity of TEs is neutral or provides benefits to the organism, such as novel splicing forms or suppression of mutation. Alternatively, if the cost of controlling TE activity exceeds the benefit of such action, such as the negative impact on gene expression, TE activity will continue. From this point of view, the relationship between TEs and genes in the epigenetic pathway may represent a feedback loop to maintain the activity of silencing at an optimal level. Taken together, the composition and abundance of transposons may not only influence genome size and genome structure, but also the path of evolution of non-TE protein-coding genes.

EXPERIMENTAL PROCEDURES

Construction of repeat library

The lotus repeat library was built using the initial version of the lotus assembly (Ming et al., 2013) and supplemented with the latest assemblies when they were available (Gui et al., 2018; Li et al., 2021). Repetitive sequences were mined using a variety of approaches. LTR retrotransposons were collected using LTR_retriever (Ou & Jiang, 2018). SINE retroelements were collected using AnnoSINE (Li et al., 2022). Non-autonomous DNA elements were mined using the MITE-Hunter package with parameters as recommended (Han & Wessler, 2010). Redundancy in the output of LTR_retriever, AnnoSINE, and MITE-Hunter was reduced based on the definition of family proposed by Wicker et al. (Wicker et al., 2007). In this case, if two elements share 80% or higher identity at the nucleotide level for over 80% of the length of the element, the two elements were considered to belong to the same family and only one sequence was retained. Subsequently, all entries (exemplars) in the non-redundant library were manually verified for their boundary, terminal sequences, and target site duplications (TSDs). See the next paragraph for details of manual curation.

The verified exemplars of LTR retroelements, SINE retroelements, and non-autonomous DNA elements were then used to mask the genomic sequence using RepeatMasker (http://www.repeatmasker.org/) and the repetitive sequences in the unmasked portion of the genomic DNA were further identified in a second mining step using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html). The output of RepeatModeler contains both known and unknown repeats. The resulting sequences were first filtered to remove putative gene families using BLASTX (Altschul et al., 1990) and sequences matching non-TE proteins ($E < 10^{-5}$) were removed. The remaining sequences where the genome coverage is ≥0.05% were manually curated to determine their identity and 5′ and 3′ boundaries. This was done in a stepwise process. First, the relevant sequences collected through RepeatModeler were used to search and retrieve at least 10 hits (BLASTN, $E < 10^{-10}$) (Altschul et al., 1990) with the corresponding 100 bp of 5′ and 3′ flanking sequences. Second, recovered sequences were aligned using DIALIGN2 (Morgenstern, 1999) to determine the possible boundary between elements and their flanking sequences. In this case, a boundary was defined as the position to which sequence homology stops for over half of the aligned sequences. Finally, sequences with defined boundaries were examined for the presence of TSDs. To classify the relevant TEs, features in the terminal ends and TSD were used. Each transposon family is associated with distinct features in its terminal sequences and TSD, which can be utilized to identify the element (Wicker et al., 2007). The identification of putative autonomous elements was assisted by their homology to known transposases from Repbase (Bao et al., 2015). For intact LTR retroelements in the repeat library, the 5′ and 3′ LTR sequences as well as 50-bp flanking sequences of a single retroelement were aligned to examine the exact boundary of LTRs and TSDs.

Manually curated TE sequences from RepeatModeler were supplemented with verified exemplars from LTR_retriever, AnnoSINE, and MITE-Hunter to form the final repeat library, which was used for subsequent analysis, and each sequence in the library was considered as one family. The lotus repeat library is available as Data S1.

Estimation of copy number and genomic fraction

The lotus repeat library was used to mask the genomic sequence to determine TE coverage and copy number. RepeatMasker tends to break down large TEs into multiple segments, so the following procedures were developed to ascertain more accurate estimation of copy numbers. If an element in the genomic sequence matched a sequence in the repeat library over the entire sequence or if the truncation was less than 20 bp on both ends, this copy was considered to be intact. If the element contains one end (truncation less than 20 bp) it was considered truncated. If no end was detected, it was considered a fragment, and the copy number was estimated by comparing the length of the fragment or truncated sequence to the full length of the element. For example, if a fragment of the element was 200 bp and the intact element was 1 kb,

this fragment was considered to be 0.2 copies. The sole exception was for LINEs, which tend to be truncated at the 5′ end upon insertion (Zingler et al., 2005). Since it was unclear whether a 5′ truncated LINE was born truncated (copy number = 1) or truncated after transposition, the copy number was considered to be the mean between 1 and a real truncated element. For LTR retroelements, a soloLTR, which refers to an individual LTR sequence without being attached to the internal region of the same element, was considered as one copy because it has been derived from one intact element. For LTRs associated with internal regions, one LTR was considered as 0.25 copies, so two LTRs represented 0.5 copies. A complete internal region of an LTR retroelement was considered as 0.5 copies. If the status of the LTR sequence was unclear, for example, a fragment, the copy number was the mean between a soloLTR and LTR in a retroelement and normalized by the full length of the LTR. For example, if the full-length LTR (not including the internal region) was 1 kb in length, a 200-bp fragment of LTR was considered $(1 + 0.25) \times 0.2/2 = 0.125$ copies. The genomic fraction of TEs was estimated using the total sequence masked by each superfamily with overlapping regions between different entries only calculated once.

Phylogenetic analysis

The conserved motif 3 of the *hAT* transposase was defined following Lazarow et al. (Lazarow et al., 2012), corresponding to *Ac* transposase amino acids 682–751 (GenBank accession number P08770.2). The curated nucleotide sequences for autonomous *hAT* families in the lotus repeat library or from Repbase (RepBase23.04.fasta) were translated and aligned with motif 3 of the *Ac* protein sequence to obtain the corresponding regions. Frameshifts were manually corrected and premature stop codons were excluded. For *hAT* elements from other plants, sequences from GenBank were also aligned to the *Ac* transposase to identify regions containing the conserved motif 3. For LTR retroelements, the conserved integrase core domain, corresponding to TNT amino acids 481–592 (GenBank accession number P10978.1), of representative LTR retroelements was retrieved similarly. LTR retroelements with various ends from grape were collected by LTR_retriever (Ou & Jiang, 2018). The detailed information of elements used in the phylogenetic analysis is shown in Tables S7–S9.

Sequences of the conserved integrase core domain from LTR retroelements and motif 3 from *hAT* transposase were used to generate multiple alignments and resolved into lineages by generating phylogenetic trees. Multiple sequence alignment was performed by MUSCLE with default parameters (Edgar, 2004). Phylogenetic trees were generated using the neighbor-joining method with MEGA (Kumar et al., 2018). Support for the internal branches of the phylogeny was assessed using 1000 bootstrap replicates.

The abundance of insertion density and fraction in genic regions

The gff files of gene sequences (the NNU-MBE dataset) were downloaded from the Nelumbo Genome Database (Li et al., 2021). For each locus, the gene model with the longest CDSs was considered. Thereafter, genes with one or more of the following features were excluded from subsequent analysis: (i) the CDS encodes less than 50 amino acids; (ii) either no start codon or stop codon; (iii) there is a stop codon within (not at the end of) the CDS; (iv) ≥50% of the CDS is masked by the repeat library or the CDS is homologous to a known transposase (e = 1e−5); (v) ≥50% of the CDS is inside the intact TEs identified in this study; (vi) ≥50% of the introns are 10 bp or smaller. After filtering, a total of 28 455 genes were retained. Thereafter the genic regions were divided into 2 kb upstream (of the TSS), exons, introns, and 2 kb downstream (of the TTS) based on the gff files of those genes. The remainder of the genome was considered intergenic regions. See Data S2 for the list of 28 455 genes used in this study.

The copy number and length of TEs in each genic region were obtained by comparison between the coordinates of the genic regions and the coordinates of the TE in the genome, based on the RepeatMasker output of the assembly. If a TE was located at the boundary of two regions, the copy number was calculated based on the length distribution on each site. For example, if a TE that is 1 kb in length is located at the boundary between an intergenic region and the adjacent upstream region, with 200 bp inside the upstream region, it was considered as 0.2 copies for the upstream region and 0.8 copies for the intergenic region. The TE fraction in each region was calculated by dividing the total TE length by the total length of the relevant region. The insertion density was calculated by dividing the total copy number in each region by the total length. The expected insertions and abundance in a region were calculated using the average density of a superfamily multiplied by the length of the region. The percent difference from expected (EI) was calculated as (observed insertions − expected insertions) ÷ (expected insertions) × 100% if it was density-based. Fraction-based EI was calculated as follows: (genome fraction in a specific region − genome-wide average) ÷ genome-wide average × 100%.

The enrichment and underrepresentation of TEs in genes involved in different biological processes

Using the CDSs of the 28 455 filtered genes a pipeline was developed in Blast2GO to assign GO terms (Conesa et al., 2005). The CDSs were searched using BLASTX with default setting (E value = 0.001), and the top three hits were retained using the UniProtKB protein database (https://www.uniprot.org/help/uniprotkb). GO-Slim plant terms were assigned based on the top three hits, resulting in 23 465 genes with GO-Slim plant GO terms. The average TE abundance (bp/gene) for a certain GO term was obtained by dividing the total TE length in all genes by the gene number in this category. Similar to the last section, EI here refers to the percent difference between the average TE abundance of an individual GO category and that of all GO categories, calculated as (TE abundance of an individual GO category − TE abundance of all GO categories) ÷ (TE abundance of all GO categories) × 100%. The significance of the difference was tested using the Kolmogorov–Smirnov test at the 0.01 and 0.001 significance levels. To determine the possible FDR, groups of genes (with 67, 150, 346, and 1000 genes) were randomly drawn from the gene pool and analyzed the same way as genes in a real GO term. If the result turned out to be significant, it was considered a false positive event. The simulation was repeated 1000 times for each gene group. The FDR was calculated as the number of false positive events in each experiment divided by 1000.

AUTHORS′ CONTRIBUTIONS

NJ designed the research and conducted manual curation of TEs. OS collected LTR retrotransposon candidates and calculated LAI. YL and YS collected SINE candidates. SC and NJ conducted the remainder of analyses. NJ and SC drafted the manuscript. All authors reviewed and edited the manuscript.

ACKNOWLEDGMENTS

We wish to thank Ann Ferguson for her work on a previous version of this manuscript. We thank Drs. Cornelius Barry and Eva Farre for critical reading of the manuscript. This study was

supported by the National Science Foundation (MCB-1121650 and IOS-1740874 to NJ) and the United States Department of Agriculture National Institute of Food and Agriculture and AgBioResearch at Michigan State University (Hatch grant MICL2707 to NJ).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data underlying this article are available in the article and in its online supplementary material.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. Retrotransposon information of various sequenced plant genomes.

Table S2. Genome information and DNA transposon content of various sequenced plant genomes.

Table S3. The abundance of LTR retrotransposons with non-canonical termini in lotus.

Table S4. Abundance and degree of enrichment of individual TE superfamilies in different genomic regions.

Table S5. Non-TGCA LTR retroelements with identical LTR in lotus.

Table S6. GO terms in the biological process category in which genes are associated with enrichment or underrepresentation of TE insertions in introns.

Table S7. GenBank sequences used in phylogeny analysis.

Table S8. Repbase *hAT* elements used in phylogeny analysis.

Table S9. Grape *Copia* LTR retroelements used in phylogeny analysis.

Figure S1. Correlation of genomic fraction between *hAT* elements and DNA transposons.

Figure S2. The phylogeny of motif 3 of *hAT*-like transposase in lotus and other plants.

Figure S3. Correlation between element size and enrichment index (based on genomic fraction) in different genic regions.

Figure S4. The relationship between gene number and false discovery rate using 1000 random gene permutations.

Figure S5. The enrichment/underrepresentation of TEs in upstream regions of genes.

Figure S6. The enrichment/underrepresentation of TEs in downstream regions of genes.

Data S1. Manually curated lotus repeat library used in this study.

Data S2. List of lotus protein-coding genes used in this study.

REFERENCES

Agren, J.A. & Wright, S.I. (2011) Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? *Chromosome Research*, **19**, 777–786.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Anderson, S.N., Stitzer, M.C., Brohammer, A.B., Zhou, P., Noshay, J.M., O'Connor, C.H. *et al.* (2019) Transposable elements contribute to dynamic genome content in maize. *The Plant Journal*, **100**, 1052–1065.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L. *et al.* (2017) The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. *Nature*, **546**, 148–152.

Bao, W., Jurka, M.G., Kapitonov, V.V. & Jurka, J. (2009) New superfamilies of eukaryotic DNA transposons and their internal divisions. *Molecular Biology and Evolution*, **26**, 983–993.

Bao, W., Kojima, K.K. & Kohany, O. (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6**, 11.

Baucom, R.S., Estill, J.C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.M. *et al.* (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genetics*, **5**, e1000732.

Bennetzen, J.L. & Kellogg, E.A. (1997) Do plants have a one-way ticket to genomic obesity? *Plant Cell*, **9**, 1509–1514.

Bennetzen, J.L., Ma, J. & Devos, K.M. (2005) Mechanisms of recent genome size variation in flowering plants. *Annals of Botany*, **95**, 127–132.

Bennetzen, J.L. & Wang, H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology*, **65**, 505–530.

Bertioli, D.J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G. *et al.* (2019) The genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nature Genetics*, **51**, 877–884.

Bureau, T.E., Ronald, P.C. & Wessler, S.R. (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 8524–8529.

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.C., Liu, K.W. *et al.* (2015) The genome sequence of the orchid Phalaenopsis equestris. *Nature Genetics*, **47**, 65–72.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. & Kondrashov, F.A. (2002) Selection for short introns in highly expressed genes. *Nature Genetics*, **31**, 415–418.

Cerbin, S., Wai, C.M., VanBuren, R. & Jiang, N. (2019) GingerRoot: a novel DNA transposon encoding integrase-related transposase in plants and animals. *Genome Biology and Evolution*, **11**, 3181–3193.

Charlesworth, D. & Charlesworth, B. (1995) Transposable elements in inbreeding and outbreeding populations. *Genetics*, **140**, 415–417.

Chen, J., Hao, Z., Guang, X., Zhao, C., Wang, P., Xue, L. *et al.* (2019) Liriodendron genome sheds light on angiosperm phylogeny and species-pair differentiation. *Nature Plants*, **5**, 18–25.

Cheng, C., Daigen, M. & Hirochika, H. (2006) Epigenetic regulation of the rice retrotransposon Tos17. *Molecular Genetics and Genomics*, **276**, 378–390.

Choi, J.Y. & Lee, Y.C.G. (2020) Double-edged sword: the evolutionary consequences of the epigenetic silencing of transposable elements. *PLoS Genetics*, **16**, e1008872.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. & Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.H. *et al.* (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *The Plant Journal*, **71**, 492–502.

Diao, Y., Chen, L., Yang, G., Zhou, M., Song, Y., Hu, Z. *et al.* (2006) Nuclear DNA C-values in 12 species in nymphales. *Caryologia*, **59**, 25–30.

Dietrich, C.R., Cui, F., Packila, M.L., Li, J., Ashlock, D.A., Nikolau, B.J. *et al.* (2002) Maize mu transposons are targeted to the 5′ untranslated region of the gl8 gene and sequences flanking mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. *Genetics*, **160**, 697–716.

Drinnan, A.N., Crane, P.R. & Hoot, S.B. (1994) Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). In: Endress, P.K. & Friis, E.M. (Eds.) *Early evolution of flowers*. Springer Vienna: Vienna, pp. 93–122.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O. & Arnold, F.H. (2005) Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 14338–14343.

Du, J., Tian, Z., Bowen, N.J., Schmutz, J., Shoemaker, R.C. & Ma, J. (2010) Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR swapping in soybean. *Plant Cell*, **22**, 48–61.

Du, Z., Ilyinskii, P.O., Lally, K., Desrosiers, R.C. & Engelman, A. (1997) A mutation in integrase can compensate for mutations in the simian immunodeficiency virus att site. *Journal of Virology*, **71**, 8124–8132.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

El Baidouri, M., Carpentier, M.C., Cooke, R., Gao, D., Lasserre, E., Llauro, C. *et al.* (2014) Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research*, **24**, 831–838.

Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

Feschotte, C., Jiang, N. & Wessler, S.R. (2002) Plant transposable elements: where genetics meets genomics. *Nature Reviews. Genetics*, **3**, 329–341.

Glemin, S., Francois, C.M. & Galtier, N. (2019) Genome evolution in outcrossing vs. selfing vs. asexual species. *Methods in Molecular Biology*, **1910**, 331–369.

Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W. *et al.* (2016) Draft genome of the living fossil Ginkgo biloba. *Gigascience*, **5**, 49.

Gui, S., Peng, J., Wang, X., Wu, Z., Cao, R., Salse, J. *et al.* (2018) Improving Nelumbo nucifera genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements. *The Plant Journal*, **94**, 721–734.

Han, Y. & Wessler, S.R. (2010) MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, **38**, e199.

Han, Y.C., Teng, C.Z., Zhong, S., Zhou, M.Q., Hu, Z.L. & Song, Y.C. (2007) Genetic variation and clonal diversity in populations of *Nelumbo nucifera* (Nelumbonaceae) in Central China detected by ISSR markers. *Aquatic Botany*, **86**(1), 69–75.

Hirochika, H., Okamoto, H. & Kakutani, T. (2000) Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. *Plant Cell*, **12**, 357–369.

Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H. & Kanda, M. (1996) Retrotransposons of rice involved in mutations induced by tissue culture. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 7783–7788.

Hollister, J.D. & Gaut, B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, **19**, 1419–1428.

Hua-Van, A., Le Rouzic, A., Maisonhaute, C. & Capy, P. (2005) Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenetic and Genome Research*, **110**, 426–440.

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J. *et al.* (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**, 655–662.

International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Jarvis, C.E. (2007) *Order out of Chaos: Linnaean plant names and their types*. London, England: Linnean Society of London.

Jia, H.M., Jia, H.J., Cai, Q.L., Wang, Y., Zhao, H.B., Yang, W.F. *et al.* (2019) The red bayberry genome and genetic basis of sex determination. *Plant Biotechnology Journal*, **17**, 397–409.

Jiang, K. & Goertzen, L.R. (2011) Spliceosomal intron size expansion in domesticated grapevine (Vitis vinifera). *BMC Research Notes*, **4**, 52.

Jiang, N., Ferguson, A.A., Slotkin, R.K. & Lisch, D. (2011) Pack-Mutator-like transposable elements (pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 1537–1542.

Jiang, N. & Panaud, O. (2013) Transposable element dynamics in rice and its wild relatives. In: Zhang, Q. & Wing, R.A. (Eds.) *Genetics and genomics of rice*. New York: Springer New York, pp. 55–69.

Juntawong, P., Girke, T., Bazin, J. & Bailey-Serres, J. (2014) Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, E203–E212.

Kapitonov, V.V. & Jurka, J. (2008) A universal classification of eukaryotic transposable elements implemented in repbase. *Nature Reviews. Genetics*, **9**, 411–412 author reply 414.

Kejnovsky, E., Hawkins, J.S. & Feschotte, C. (2012) Plant transposable elements: biology and evolution. In: Wendel, J.F., Greilhuber, J., Dolezel, J. & Leitch, I.J. (Eds.) *Plant genome diversity*. Wien: Springer-Verlag, pp. 17–33.

Kempken, F. & Windhofer, F. (2001) The hAT family: a versatile transposon group common to plants, fungi, animals, and man. *Chromosoma*, **110**, 1–9.

Koonin, E.V. & Wolf, Y.I. (2010) Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews. Genetics*, **11**, 487–498.

Korkuc, P., Schippers, J.H. & Walther, D. (2014) Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiology*, **164**, 181–200.

Kumar, A. & Bennetzen, J.L. (1999) Plant retrotransposons. *Annual Review of Genetics*, **33**, 479–532.

Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, **35**, 1547–1549.

Kunze, R. & Weil, C. (2002) The *hAT* and CACTA superfamilies of plant transposons. In: Craig, N.L., Craigie, R., Gellert, M. & Lambowitz, A.M. (Eds.) *Mobile DNA II*. Washington, DC: ASM Press, pp. 565–610.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lazarow, K., Du, M.L., Weimer, R. & Kunze, R. (2012) A hyperactive transposase of the maize transposable element activator (ac). *Genetics*, **191**, 747–756.

Le, T.N., Miyazaki, Y., Takuno, S. & Saze, H. (2015) Epigenetic regulation of intragenic transposable elements impacts gene transcription in Arabidopsis thaliana. *Nucleic Acids Research*, **43**, 3911–3921.

Li, H., Yang, X., Zhang, Y., Gao, Z., Liang, Y., Chen, J. *et al.* (2021) Nelumbo genome database, an integrative resource for gene expression and variants of Nelumbo nucifera. *Scientific Data*, **8**, 38.

Li, Y., Jiang, N. & Sun, Y. (2022) AnnoSINE: a short interspersed nuclear elements annotation tool for plant genomes. *Plant Physiology*, **188**, 955–970.

Liu, S., Yeh, C.T., Ji, T., Ying, K., Wu, H., Tang, H.M. *et al.* (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genetics*, **5**, e1000733.

Loreto, E.L., Carareto, C.M. & Capy, P. (2008) Revisiting horizontal transfer of transposable elements in drosophila. *Heredity (Edinburgh)*, **100**, 545–554.

Ma, J. & Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 12404–12410.

Mari-Ordonez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V. & Voinnet, O. (2013) Reconstructing de novo silencing of an active plant retrotransposon. *Nature Genetics*, **45**, 1029–1039.

Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., Platzer, M. *et al.* (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.

McCarthy, E.M. & McDonald, J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362–367.

McClintock, B. (1950) The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, **36**, 344–355.

Michael, T.P. (2014) Plant genome size variation: bloating and purging DNA. *Briefings in Functional Genomics*, **13**, 308–317.

Ming, R., VanBuren, R., Liu, Y., Yang, M., Han, Y., Li, L.T. *et al.* (2013) Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). *Genome Biology*, **14**, R41.

Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K. *et al.* (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell*, **15**, 1771–1780.

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M. *et al.* (2022) Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, **602**, 101–105.

Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.

Noreen, F., Akbergenov, R., Hohn, T. & Richert-Poggeler, K.R. (2007) Distinct expression of endogenous petunia vein clearing virus and the DNA transposon dTph1 in two Petunia hybrida lines is correlated with

differences in histone modification and siRNA production. *The Plant Journal*, **50**, 219–229.

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G. *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579–584.

Ong-Abdullah, M., Ordway, J.M., Jiang, N., Ooi, S.E., Kok, S.Y., Sarpan, N. *et al.* (2015) Loss of karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, **525**, 533–537.

Ou, S., Chen, J. & Jiang, N. (2018) Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research*, **46**, e126.

Ou, S. & Jiang, N. (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, **176**, 1410–1422.

Pan, L., Xia, Q., Quan, Z., Liu, H., Ke, W. & Ding, Y. (2010) Development of novel EST–SSRs from sacred lotus (*Nelumbo nucifera* Gaertn) and their utilization for the genetic diversity analysis of *N. nucifera*. *Journal of Heredity*, **101**, 71–82.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H. *et al.* (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Research*, **16**, 1262–1269.

Ren, X.Y., Vorst, O., Fiers, M.W., Stiekema, W.J. & Nap, J.P. (2006) In plants, highly expressed genes are the least compact. *Trends in Genetics*, **22**, 528–532.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

Robertson, H.M. (2002) Evolution of DNA transposons. In: Craig, N.L., Craigie, R., Gellert, M. & Lambowitz, A.M. (Eds.) *Mobile DNA II*. Washington, DC: ASM Press, pp. 1093–1110.

Robertson, H.M. & Lampe, D.J. (1995) Recent horizontal transfer of a mariner transposable element among and between diptera and neuroptera. *Molecular Biology and Evolution*, **12**, 850–862.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nature Genetics*, **20**, 43–45.

Schaack, S., Gilbert, C. & Feschotte, C. (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, **25**, 537–546.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Sela, N., Mersch, B., Gal-Mark, N., Lev-Maor, G., Hotz-Wagenblatt, A. & Ast, G. (2007) Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biology*, **8**, R127.

Shen-Miller, J. (2002) Sacred lotus, the long-living fruits of China antique. *Seed Science Research*, **12**, 131–143.

Shi, T., Rahmani, R.S., Gugger, P.F., Wang, M., Li, H., Zhang, Y. *et al.* (2020) Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants. *Molecular Biology and Evolution*, **37**, 2394–2413.

Sigman, M.J. & Slotkin, R.K. (2016) The first rule of plant transposable element silencing: location, location, location. *Plant Cell*, **28**, 304–313.

Singer, M.F. (1982) SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell*, **28**, 433–434.

Slotkin, R.K. & Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nature Reviews. Genetics*, **8**, 272–285.

Song, X., Sun, P., Yuan, J., Gong, K., Li, N., Meng, F. *et al.* (2021) The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. *Plant Biotechnology Journal*, **19**, 731–744.

Srivastava, A.K., Lu, Y., Zinta, G., Lang, Z. & Zhu, J.K. (2018) UTR-dependent control of gene expression in plants. *Trends in Plant Science*, **23**, 248–259.

Stival Sena, J., Giguere, I., Boyle, B., Rigault, P., Birol, I., Zuccolo, A. *et al.* (2014) Evolution of gene structure in the conifer Picea glauca: a comparative analysis of the impact of intron size. *BMC Plant Biology*, **14**, 95.

Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. (2011) Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics*, **43**, 1160–1163.

Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C. *et al.* (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes & Development*, **9**, 1797–1810.

The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

VanBuren, R., Wai, C.M., Ou, S., Pardo, J., Bryant, D., Jiang, N. *et al.* (2018) Extreme haplotype variation in the desiccation-tolerant clubmoss Selaginella lepidophylla. *Nature Communications*, **9**, 13.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D. *et al.* (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, **2**, e1326.

Vitte, C. & Panaud, O. (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic and Genome Research*, **110**, 91–107.

Voronova, A., Rendon-Anaya, M., Ingvarsson, P., Kalendar, R. & Rungis, D. (2020) Comparative study of pine reference genomes reveals transposable element interconnected gene networks. *Genes (Basel)*, **11**, 1216.

Wallau, G.L., Ortiz, M.F. & Loreto, E.L. (2012) Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome Biology and Evolution*, **4**, 689–699.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Wendel, J.F., Cronn, R.C., Alvarez, I., Liu, B., Small, R.L. & Senchina, D.S. (2002) Intron size and genome size in plants. *Molecular Biology and Evolution*, **19**, 2346–2352.

White, S.E., Habera, L.F. & Wessler, S.R. (1994) Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 11792–11796.

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-Gonzalez, R.H. *et al.* (2018) Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology*, **19**, 103.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews. Genetics*, **8**, 973–982.

Wikstrom, N., Savolainen, V. & Chase, M.W. (2001) Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Biological Sciences*, **268**, 2211–2220.

Wright, S.I., Le, Q.H., Schoen, D.J. & Bureau, T.E. (2001) Population dynamics of an ac-like transposable element in self- and cross-pollinating arabidopsis. *Genetics*, **158**, 1279–1288.

Wu, C. & Lu, J. (2019) Diversification of transposable elements in arthropods and its impact on genome evolution. *Genes (Basel)*, **10**, 338.

Xie, Z., Johansen, L.K., Gustafson, A.M., Kasschau, K.D., Lellis, A.D., Zilberman, D. *et al.* (2004) Genetic and functional diversification of small RNA pathways in plants. *PLoS Biology*, **2**, E104.

Xu, Z. & Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**, W265–W268.

Yang, Y., Sun, P., Lv, L., Wang, D., Ru, D., Li, Y. *et al.* (2020) Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nature Plants*, **6**, 215–222.

Zhang, H.H., Peccoud, J., Xu, M.R., Zhang, X.G. & Gilbert, C. (2020) Horizontal transfer and evolution of transposable elements in vertebrates. *Nature Communications*, **11**, 1362.

Zhang, X., Zhao, M., McCarty, D.R. & Lisch, D. (2020) Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. *Nucleic Acids Research*, **48**, 6685–6698.

Zhao, D., Ferguson, A. & Jiang, N. (2015) Transposition of a rice mutator-like element in the yeast Saccharomyces cerevisiae. *Plant Cell*, **27**, 132–148.

Zhao, D., Hamilton, J.P., Hardigan, M., Yin, D., He, T., Vaillancourt, B. *et al.* (2017) Analysis of ribosome-associated mRNAs in rice reveals the

importance of transcript size and GC content in translation. *G3: Genes, Genomes, Genetics*, **7**, 203–219.

Zhou, S.S., Yan, X.M., Zhang, K.F., Liu, H., Xu, J., Nie, S. *et al.* (2021) A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. *Scientific Data*, **8**, 174.

Zimin, A., Stevens, K.A., Crepeau, M.W., Holtz-Morris, A., Koriabine, M., Marcais, G. *et al.* (2014) Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics*, **196**, 875–890.

Zimin, A.V., Stevens, K.A., Crepeau, M.W., Puiu, D., Wegrzyn, J.L., Yorke, J.A. *et al.* (2017) An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience*, **6**, 1–4.

Zingler, N., Willhoeft, U., Brose, H.P., Schoder, V., Jahns, T., Hanschmann, K.M. *et al.* (2005) Analysis of 5′ junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5′-end attachment requiring microhomology-mediated end-joining. *Genome Research*, **15**, 780–789.