the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2010) doi: 10.1111/j.1365-313X.2010.04263.x

Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison

Jianchang Du[1], Zhixi Tian[1], Christian S. Hans[1], Howard M. Laten[2], Steven B. Cannon[3], Scott A. Jackson[1], Randy C. Shoemaker[3,*] and Jianxin Ma[1,*]

[1]*Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA,*

[2]*Department of Biology, Loyola University Chicago, Chicago, IL 60660, USA, and*

[3]*US Department of Agriculture–Agricultural Research Service, Corn Insects, and Crop Genetics Research Unit, Ames, IA 50011, USA*

Received 20 March 2010; revised 14 May 2010; accepted 18 May 2010.

*For correspondence (fax +765 496 7255; e-mail maj@purdue.edu or fax +515 294 2299; e-mail randy.shoemaker@ars.usda.gov).

SUMMARY

The availability of complete or nearly complete genome sequences from several plant species permits detailed discovery and cross-species comparison of transposable elements (TEs) at the whole genome level. We initially investigated 510 long terminal repeat-retrotransposon (LTR-RT) families comprising 32 370 elements in soybean (*Glycine max* (L.) Merr.). Approximately 87% of these elements were located in recombination-suppressed pericentromeric regions, where the ratio (1.26) of solo LTRs to intact elements (S/I) is significantly lower than that of chromosome arms (1.62). Further analysis revealed a significant positive correlation between S/I and LTR sizes, indicating that larger LTRs facilitate solo LTR formation. Phylogenetic analysis revealed seven *Copia* and five *Gypsy* evolutionary lineages that were present before the divergence of eudicot and monocot species, but the scales and timeframes within which they proliferated vary dramatically across families, lineages and species, and notably, a *Copia* lineage has been lost in soybean. Analysis of the physical association of LTR-RTs with centromere satellite repeats identified two putative centromere retrotransposon (*CR*) families of soybean, which were grouped into the *CR* (e.g. *CRR* and *CRM*) lineage found in grasses, indicating that the 'functional specification' of *CR* pre-dates the bifurcation of eudicots and monocots. However, a number of families of the *CR* lineage are not concentrated in centromeres, suggesting that their *CR* roles may now be defunct. Our data also suggest that the envelope-like genes in the putative *Copia* retrovirus-like family are probably derived from the *Gypsy* retrovirus-like lineage, and thus we propose the hypothesis of a single ancient origin of envelope-like genes in flowering plants.

Keywords: centromere retrotransposons, plant retroviruses, LTR-retrotransposons, genome evolution, comparative genomics.

INTRODUCTION

Long terminal repeat-retrotransposons (LTR-RTs) are the most abundant genomic components in flowering plants, making up a large fraction of all plant genomes thus far investigated. For example, approximately one-quarter and three-quarters of the rice and maize genomes, respectively, are composed of LTR-RTs (Ma *et al.*, 2004; International Rice Genome Sequencing Project 2005; Baucom *et al.*, 2009; Schnable *et al.*, 2009). These elements initiate their transposition through a copy/paste mechanism via RNA intermediates. A typical intact element contains two identical LTRs, a primer-binding site (PBS), a polypurine tract (PPT), *gag,* a gene that encodes a polyprotein comprising subcomponents of the virus-like particle (VLP) involved in the maturation and packaging of retrotransposon RNA, and *pol* gene products that encode protease (PR), reverse transcriptase (RT), RNase H (RH) and integrase (IN) that are involved in the synthesis of retrotransposon DNA and integration into the host genome (Kumar and Bennetzen, 1999). Based on the order of RT and IN in POL, LTR-RTs are classified into *Gypsy* and *Copia* types (Xiong and Eickbush,

1990). A few families of plant LTR-RTs were found to contain an open reading frame (ORF) that encodes an envelope (ENV)-like protein that is typically present in infectious retroviruses, leading to the suggestion that these elements may be endogenous plant retroviruses (Laten, 1999; Laten *et al.*, 2003; Wright and Voytas, 2002).

Numerous LTR-RT families have been identified in plants, and their rapid amplification, along with polyploidization, is largely responsible for genome expansion (Bennetzen *et al.*, 2005). However, the transpositional activities of elements vary greatly among families. For example, >80% of LTR-RTs in the maize (*Zea mays*) genome belong to the five largest families (SanMiguel *et al.*, 1996). A recent study shows that the genome size of *Oryza australiensis*, a wild relative of rice, was doubled within the last 3 million years (Myr) by aggressive proliferation of LTR-RTs belonging to three families (Piegu *et al.*, 2006). Indeed, the majority of plant LTR-RTs were estimated to have amplified within the last few Myr (Ma *et al.*, 2004; Vitte and Bennetzen, 2006; SanMiguel and Vitte, 2009). This would be a reasonable expectation, because, in many cases, only intact elements with two LTRs were analyzed. Many older elements have experienced severe deletions or fragmentation by unequal homologous recombination and illegitimate recombination (Devos *et al.*, 2002; Ma *et al.*, 2004), the two major mechanisms that counteract genome expansion, and thus were not able to be dated or identified precisely. The magnitude and pace of elimination of LTR-RT DNA in plants is remarkable, given that few LTR-RT fragments were shared as orthologous copies between closely related species, such as maize and sorghum (Ma *et al.*, 2005), which diverged from each other approximately 12 million years ago (Mya) (Swigonova *et al.*, 2004). It seems clear that the activities of either amplification or elimination of LTR-RTs vary among species (Bennetzen *et al.*, 2005; Wicker and Keller, 2007), but little is known regarding the evolutionary patterns and fates of individual families and their biological propensities and functional diversification in different host genomes.

As well as their impact on genome size variation, LTR-RTs were found to be able to regulate the expression of adjacent genes in their host genomes (Kashkush *et al.*, 2003; Kashkush and Khasdan, 2007). Identification of LTR-RTs is the first step towards the characterization of potential interactions between LTR-RTs and genes in a particular genome. In addition, LTR-RTs, especially uncharacterized low-copy-number elements or fragments, were often mis-annotated as genes (Bennetzen *et al.*, 2004). Thus, accurate and complete annotation of transposable elements, mostly LTR-RTs, has become a priority in most plant genome sequencing projects to minimize the inaccuracy of gene annotations and to facilitate the functional studies of genes.

Soybean (*Glycine max* (L.) Merr.) is one of the world's most economically important crops. It is a member of the Leguminoseae, the third largest family of flowering plants and the family that provides the majority of plant-based protein and more than a quarter of the world's food and animal feed (Graham and Vance, 2003). Previous studies suggest that soybean has undergone two rounds of whole genome duplication (Shoemaker *et al.*, 2006; Schlueter *et al.*, 2007; Gill *et al.*, 2009), thus it is also a good choice for studies of polyploidy and genome evolution. Because of its enormous economic value, soybean has been recently sequenced (Schmutz *et al.*, 2010). The assembled soybean pseudomolecules comprise 955 Mb of DNA, and represent the first completely sequenced legume genome.

Although numerous LTR-RTs have been identified from several sequenced plant genomes (Ma *et al.*, 2004; Pereira, 2004; Baucom *et al.*, 2009; Tian *et al.*, 2009), no previous study has made comprehensive efforts to compare complete sets of LTR-RTs among different plant species at the whole genome level. In this paper we first present the characterization of LTR-RTs in the soybean genome, including structural analysis of LTR-RTs, and comparison of genomic features between recombination-suppressed regions and euchromatic regions. Then we describe the genome-wide or large-scale comparison of LTR-RTs among soybean, rice, Arabidopsis, *Medicago*, and *Lotus*. Finally, we present comparative analysis of putative plant endogenous retrovirus and centromere retrotransposon (*CR*) families in monocot and eudicot species. Our study reveals the dynamics of retrotransposon evolution within a species, among species within a family, and between monocots and eudicots, and provides new insights into the evolutionary dynamics, propensities, and fates (e.g. origin, diversification, and specificity) of *CR* lineages and putative endogenous retroviruses in flowering plants.

RESULTS AND DISCUSSION

Characterization of LTR-retrotransposons in the soybean genome

A combination of structure-based and homology-based approaches (Ma and Bennetzen, 2006; Ma and Jackson, 2006) was employed to identify LTR-RTs in the soybean genome sequence (assembly version Glyma1.01), 950 Mb of which was mapped to the 20 soybean chromosomes (Schmutz *et al.*, 2010). A total of 32 370 elements with clearly defined boundaries were identified and deposited in SoyTEdb, a comprehensive database of transposable elements in the soybean genome (Du *et al.*, 2010a). Of these elements, 14 106 are intact elements and 18 264 are solo LTRs (Figure 1, Table S1 in Supporting Information). All of these elements were manually inspected and defined based on their structures as previously described (Ma *et al.*, 2004). Because the present soybean pseudomolecules contain numerous sequence gaps within and around repetitive sequences, some truncated elements or fragments can be potential products of incomplete assembly or mis-assembly

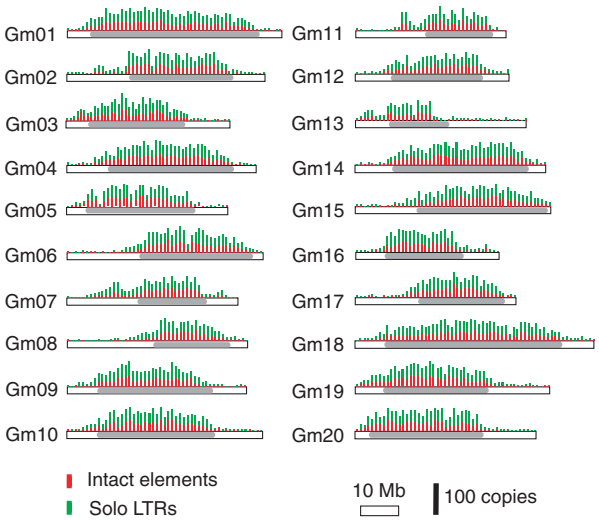

Figure 1. Distribution of long terminal repeat-retrotransposons (LTR-RTs) along the 20 chromosome pseudomolecules (Gm01–Gm20) of soybean.
Intact elements and solo LTRs are shown by the red and green bars, respectively. The recombination-suppressed region of each chromosome is represented by the gray-shadowed area within each box. Intact elements and solo LTRs are plotted along the soybean physical map using 1 Mb per unit.

of the corresponding regions. Therefore, truncated elements without structurally defined termini were not further investigated. Of the 32 370 elements described above, 31 858 (98.4%) were anchored to the currently assembled 20 chromosome pseudomolecules (Schmutz *et al.*, 2010).

Based on a unified classification system for eukaryotic transposable elements (Wicker *et al.*, 2007), the 32 370 elements were classified into 510 distinct families, including 353 *Gypsy*-like families (19 052 elements) and 157 *Copia*-like families (13 318 elements), approximately 95% of which were the first reported (Table S1) (Jurka *et al.*, 2005; Du *et al.*, 2010a). The ratio of *Gypsy*-like to *Copia*-like elements in soybean is 1.4:1 (Table S1), slightly lower than in maize (1.6:1) (Baucom *et al.*, 2009; Schnable *et al.*, 2009), much lower than in rice (4.9:1) (International Rice Genome Sequencing Project 2005; Tian *et al.*, 2009) and sorghum (3.7:1) (Paterson *et al.*, 2009), but considerably higher than reported in *Medicago* (0.3:1) (Wang and Liu, 2008). Nevertheless, the ratio of *Gypsy*-like to *Copia*-like elements in *Medicago* may be a biased estimation, as only euchromatic portions of the *Medicago* genome have been sequenced and analyzed (Wang and Liu, 2008).

The length of intact elements in soybean varies from 1 to 20 kb, with LTRs ranging from 0.1 to 4 kb in size (Figure S1). The copy numbers of individual LTR-RT families in soybean vary greatly, ranging from 1 to 4724, with an average number of 63 (Table S1). The three largest families are *Gmr9* [i.e. *SNARE*/*GmOgre* (Laten *et al.*, 2009; Du *et al.*, 2010b)], *Gmr4* [i.e. *GmGypsy10* (Laten *et al.*, 2009)], and *Gmr5*, which have 4724, 3370, and 2925 copies of intact elements and solo LTRs, respectively (Table S1). Overall, the 32 370 intact elements and solo LTRs, together with numerous truncated fragments or remnants measured by the Repeatmasker program (http://www.repeatmasker.org), make up 401 Mb of repetitive DNA, accounting for approximately 42% of the soybean genome (Schmutz *et al.*, 2010). This proportion is lower than estimated in the larger maize genome (79%) (Schnable *et al.*, 2009) and sorghum genome (55%) (Paterson *et al.*, 2009), but higher than the smaller rice genome (26%) (International Rice Genome Sequencing Project 2005). It appears that the two rounds of the whole-genome duplication events that shaped the current soybean genome are mostly responsible for the larger-size genome but with a lower proportion of LTR-RT DNA in soybean in contrast to sorghum.

Structural variation of LTR-RTs according to their ages and distribution in recombination-suppressed pericentromeric regions and chromosome arms

Of the 31 858 elements anchored to the assembled 20 chromosome pseudomolecules, 27 836 (approximately 87%) were found in the recombination-suppressed pericentromeric regions (Schmutz *et al.*, 2010) (Figure 1, Table 1). This is probably an underestimate, given that a large number of assembled scaffolds predominately composed of retrotransposon fragments and centromere satellite repeats

Table 1 Distribution of long terminal repeat-retrotransposons (LTR-RTs) within and outside of recombination-suppressed pericentromeric regions

	No. of intact elements		No. of solo LTRs		S/I ratio[a]	
Chr.	Within	Outside	Within	Outside	Within	Outside
1	949	54	1262	58	1.33	1.07
2	645	82	812	154	1.26	1.88
3	575	136	823	135	1.43	0.99
4	772	75	908	109	1.18	1.45
5	594	36	720	51	1.21	1.42
6	664	46	810	106	1.22	2.30
7	433	148	450	264	1.04	1.78
8	407	60	505	145	1.24	2.42
9	640	52	814	94	1.27	1.81
10	710	71	841	103	1.18	1.45
11	400	103	464	179	1.16	1.74
12	492	62	623	96	1.27	1.55
13	242	104	344	194	1.42	1.87
14	820	70	1054	76	1.29	1.09
15	718	70	866	117	1.21	1.67
16	476	65	648	104	1.36	1.60
17	501	67	600	93	1.20	1.39
18	933	73	1196	114	1.28	1.56
19	647	126	835	179	1.29	1.42
20	705	51	938	100	1.33	1.96
Total	12 323	1551	15 513	2471	1.26	1.62

[a]Ratio of solo LTRs (S) to intact elements (I).

(approximately 17.7 Mb) have not yet been integrated into the 20 chromosomes. By contrast, <18% (2292 out of 12 918) of LTR-RTs (intact elements and solo LTRs) in the rice genome are located in the recombination-suppressed pericentromeric regions (Tian *et al.*, 2009). The densities of LTR-RTs in the recombination-suppressed pericentromeric regions and chromosome arms are 52 Mb^{-1} and 9 Mb^{-1} in soybean (Figure 1, Table 1), and 51 Mb^{-1} and 33 Mb^{-1} in rice (Tian *et al.*, 2009), respectively. When all fragments were included, the proportions of retrotransposon DNA in the recombination-suppressed pericentromeric regions and chromosome arms are 63 and 11% in soybean, and 39 and 17% in rice (Tian *et al.*, 2009), respectively.

The formation of solo LTRs by unequal intra-element homologous recombination is thought to be a major process for removal of LTR-RT DNA in plants (Devos *et al.*, 2002; Ma *et al.*, 2004). Our data show that the ratio of solo LTRs to intact elements (S/I) in soybean is approximately 1.29:1 (Table S1), much higher than the ratio (approximately 0.12:1) suggested previously using limited bacterial artificial chromosome (BAC) sequences (Wawrzynski *et al.*, 2008). This estimate is significantly lower than reported in rice (1.62:1) (Tian *et al.*, 2009) (Fisher's exact test, $P < 10^{-25}$), but significantly higher than that in Arabidopsis (0.50:1; Table S2) (Fisher's exact test, $P < 10^{-24}$). In an attempt to shed light on the potential forces that facilitate the formation of solo LTRs, we investigated the structures of LTR-RTs along chromosomes (Figure 1). Our data reveal that the S/I ratio in pericentromeric regions (1.26:1) is significantly lower than in chromosome arms (1.62:1) (*t*-test, $P < 0.001$; Table 1). This observation, paralleling a recent study in rice, which reported significantly lower S/I ratios in pericentromeric regions (1.36:1) than in chromosome arms (1.68:1) (Tian *et al.*, 2009), suggests that the mechanisms for suppression of genetic recombination in pericentromeric regions may reduce the frequency of formation of solo LTRs by unequal recombination.

Using an approach employed earlier (Ma and Bennetzen, 2004), we estimated the ages of intact elements in soybean. As shown in Figure 2(a), most of the elements (91%) were amplified in the last 3 Myr, and approximately 3248 elements were generated within the last 0.5 Myr. Despite recent amplification, it appears that many families were active and amplified within distinct evolutionary timeframes. For example, *Gmr2/SIRE* (Laten *et al.*, 1998) has the greatest number of copies that arose within the past 0.5 Myr, and this family may be recently or even currently active, given that it contains 75 intact elements each having two identical LTRs (Figure 2b). In contrast, the majority of *Gmr3*, *Gmr19/Diaspora* (Yano *et al.*, 2005), and *Gmr25* elements were amplified within the last 0.5–1.0, 1.5–2.0, and 2.0–2.5 Myr, respectively.

Similar to that previously described in a few other species (Wicker and Keller, 2007), the overall age distribution of

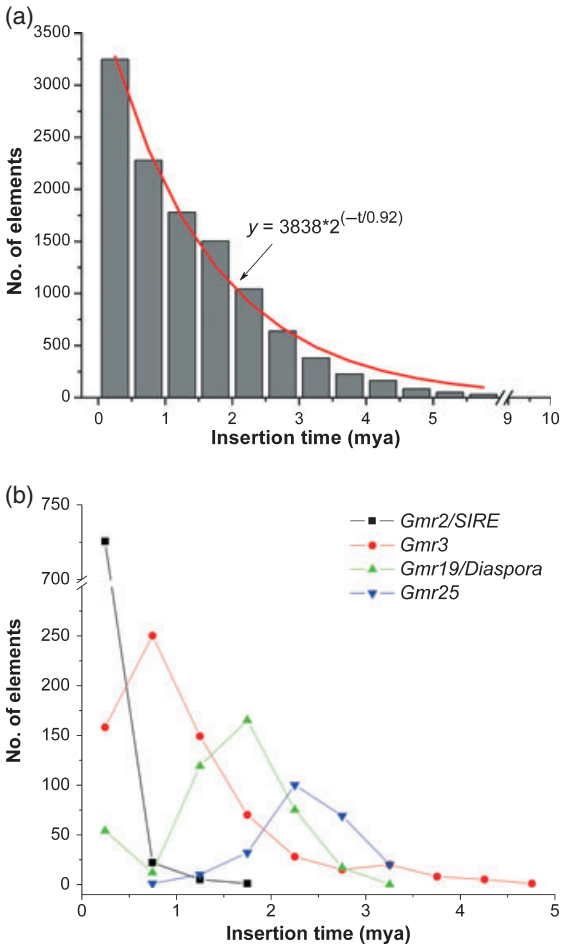

Figure 2. Timing and activities of recent amplification of long terminal repeat-retrotransposons in soybean.
(a) Insertion time of intact elements (mya, million years ago).
(b) Comparisons of activities of different families.

intact elements in soybean fits an exponential decay curve ($r = 0.99$, $P < 0.001$). This pattern is expected, because unequal recombination and illegitimate recombination have been thought to be common mechanisms responsible for rapid elimination of retrotransposon DNA during the evolution of plant genomes. In particular, illegitimate recombination that generates small deletions has been documented to be a major mechanism for elimination of LTR-RT DNA in Arabidopsis (Devos *et al.*, 2002). In this study, we identified numerous truncated LTR-RT fragments in soybean. However, because the assembled soybean genome sequence was generated by the whole genome shotgun (WGS) approach, and unavoidably contains many sequence gaps, it thus does not allow a precise assessment of the effectiveness of illegitimate recombination for the shrinkage of the soybean genome.

Our analysis reveals that the intact elements identified in pericentromeric regions and chromosome arms were

amplified at different times in both soybean and rice (average, 1.29 and 1.64 Myr in soybean, and 2.12 and 1.43 Myr in rice). Since the frequencies for removal of LTR-RT DNA by solo LTR formation, as reflected by the S/I ratios in either euchromatic regions or heterochromatic regions, are strikingly similar between rice and soybean, the formation of such large proportions of recombination-suppressed pericentromeric regions in soybean may be mainly caused by preferential insertions of LTR-RTs in the regions, instead of biased removal in gene-rich euchromatic regions. It appears that this deduction reinforces our observation that soybean LTR-RTs in pericentromeric regions are, on average, younger than those in chromosome arms. However, this deduction needs to be made with the caveat that solo LTR formation may be the predominant process for the elimination of LTR-RT DNA in both species. This is true in rice (Ma *et al.*, 2004; Tian *et al.*, 2009), but is less clear in soybean.

Association of the S/I ratios with insertion times and LTR lengths

The S/I ratios vary among families. Among the 53 families each with >50 copies (Table S1), the lowest and highest S/I ratios are 0.11 (*Gmr15*) and 16.39 (*Gmr24*), whereas the average ages of the intact elements of families *Gmr15* and *Gmr24* are 1.60 and 0.87 Myr, respectively (Table S1). No significant correlation between the S/I ratios and the average ages of intact elements was detected among these 53 families ($r = 0.12$, $P = 0.38$; Figure 3a), although families with younger intact elements have a tendency toward lower S/I ratios.

Theoretically, more solo LTRs would be formed from intact elements over evolutionary time. For example, the S/I ratio of elements amplified before the divergence of *indica* and *japonica*, two subspecies of rice, approximately 0.5 Ma, is twice that of the S/I ratio of elements amplified after the divergence of these two subspecies (Ma *et al.*, 2004; Tian *et al.*, 2009). The lack of significant correlation between the S/I ratios and the average ages of intact elements may be due to the distinct waves and scales of LTR-RT amplifications among individual families, as well as the variable degrees and magnitudes of unequal recombination and illegitimate recombination over evolutionary time (Tian *et al.*, 2009). In contrast, our data reveal a significant positive correlation between the S/I ratios and LTR sizes among the 53 soybean LTR-RT families each with >50 copies ($r = 0.63$; $P < 0.0001$; Figure 3b). Such a correlation was also detected in rice ($r = 0.39$; $P = 0.001$) by analyzing the 66 largest rice LTR-RT families (Tian *et al.*, 2009). One possible explanation is that larger LTRs can facilitate unequal homologous recombination for the formation of solo LTRs. Together, these data suggest that the formation of solo LTRs in soybean may be driven by multiple factors, such as chromosomal distribution, local genetic recombination, the lengths of LTRs, and probably also selection against insertions within or adjacent to genes (Tian *et al.*, 2009).

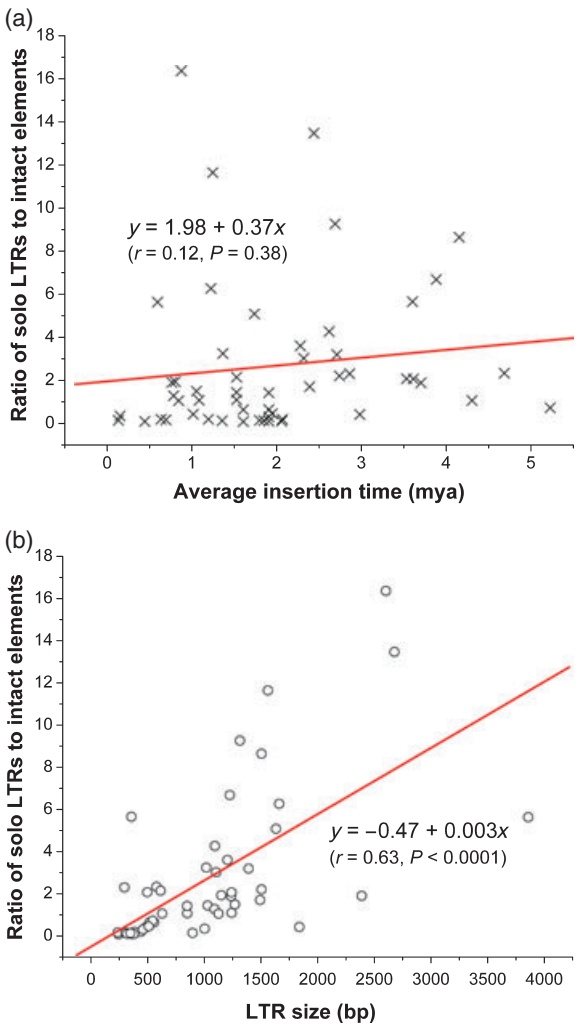

Figure 3. Genetic factors associated with solo long terminal repeat (LTR) formation.
(a) Ratios of solo LTRs to intact elements (S/I) versus average insertion times (mya, million years ago).
(b) S/I ratio versus LTR sizes.

Most evolutionary lineages of LTR-RTs are shared between monocots and eudicots, but one is extinct in soybean

To understand the evolutionary dynamics, history, and fates of LTR-RTs in flowering plants over evolutionary time, we performed comprehensive phylogenetic and comparative analyses of complete sets of LTR-RTs identified in soybean, rice (Tian *et al.*, 2009), and Arabidopsis (Pereira, 2004; Table S2), and large sets of LTR-RTs identified in *Medicago* (Wang and Liu, 2008) and *Lotus* (Holligan *et al.*, 2006). Rice is a monocot, while the other four are eudicots. The divergence

time of these species and two other monocots (maize and sorghum) is shown in Figure S2 (Chaw *et al.*, 2004; Choi *et al.*, 2004; Swigonova *et al.*, 2004; Lavin *et al.*, 2005; Tuskan *et al.*, 2006; The International *Brachypodium* Initiative 2010).

Because many families of LTR-RTs among species or within an individual genome are highly diverged, we only used the relatively conserved RT domains from individual elements for phylogenetic analysis. Of the 510 families identified in the soybean genome, 145 *Copia* and 284 *Gypsy* families each have at least one element that contains a conserved RT domain. The other 81 families are either all non-autonomous elements or contain deletions of RT domains. Five (*Gmr324*, *Gmr40*, *Gmr28*, *Gmr190*, and *Gmr27*) of these 81 families have more than 100 copies (Table S1) in soybean. Most of the non-autonomous families were each found to contain at least two elements with similar structures based on sequence alignments, suggesting that these families were capable of amplification.

A previous survey of 20 *Copia* families from barley and wheat and their homologous elements from Arabidopsis (22 families) and rice (46 families), defined six major common evolutionary *Copia* lineages (Wicker and Keller, 2007). Using a similar approach, we grouped all *Copia* families with conserved RT domains identified in Arabidopsis (33 families), rice (113 families), and soybean (145 families) into seven distinct lineages, *Ivana*, *Maximus*, *Ale*, *Angela*, *TAR*, *GMR*, and *Bianca* (Figure 4a). Six of the seven lineages are shared by these three species, except that the *Bianca* lineage was not found in soybean (Figure 4a). The *Bianca* lineage contains one and seven families that are composed of five and 32 elements in Arabidopsis and rice, respectively (Table 2). These data suggest that the *Bianca* lineage elements are now extinct in the soybean genome. Because this lineage was also identified in both *Lotus* and *Medicago* (Holligan *et al.*, 2006; Wang and Liu, 2008; Figure S3), two model legume species that diverged from soybean about 51 Mya (Lavin *et al.*, 2005), the *Bianca* lineage must have been lost in soybean within the last 51 Myr.

The *Gypsy* families from soybean, Arabidopsis, and rice fell into five previously defined evolutionary lineages: *Reina*, *CR*, *Tekay*, *Athila*, and *Tat* (Figure 4b, Table 2) (Wang and Liu, 2008). These five lineages all existed before the divergence of eudicots and monocots, and are still shared by other two legume species, *Medicago* and *Lotus* (Figure S4).

The spectrum of activity for proliferation of LTR-RTs is highly variable among lineages and species over evolutionary time

Although the 11 evolutionary lineages are shared by soybean, Arabidopsis, and rice, the scales and timeframes of activity for proliferation of LTR-RTs vary tremendously among lineages and species. The numbers of families within each lineage and the copy numbers of elements within each

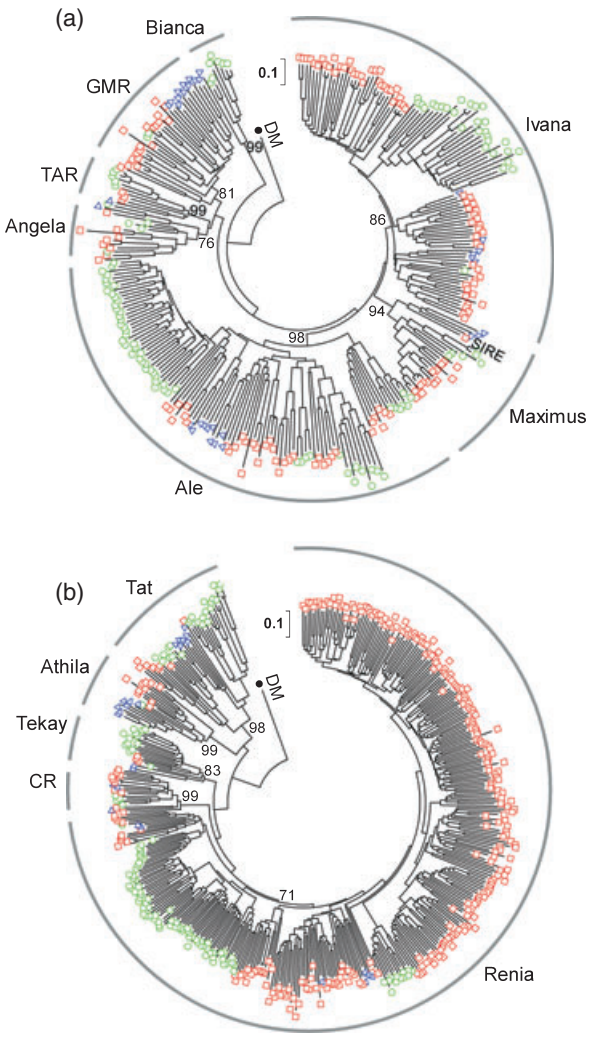

Figure 4. Phylogenetic relationships of long terminal repeat-retrotransposon (LTR-RT) families identified in soybean, rice, and Arabidopsis.
(a) *Copia* families and (b) *Gypsy* families. The RT nucleotide sequences of individual families were used to construct the phylogenetic trees, which were rooted using the RT sequences of the *Copia* element *DM* (a) and *Gypsy* element *INVIDER2* (b) identified in *Drosophila melanogaster* (DM).
Individual LTR-RT families in soybean, rice, and Arabidopsis are indicated by red boxes, green boxes, and blue triangles, respectively.

family identified in these three species are listed in Table 2. Among the six *Copia* lineages in soybean, *Ivana* has the largest number of LTR-RT families (63), accounting for 43.4% of the *Copia* families (145) analyzed. However, this lineage only contains 6.3% (788) of the 12 564 *Copia* elements. In contrast, *Maximus* is the *Copia* lineage that contains the highest number of *Copia* elements (8575, 68.3%), but these elements belong to only 15 families. More dramatic differences were observed among the five *Gypsy* lineages in soybean. For example, the lineage *Reina* contains 253 (89.1%) of the 284 *Gypsy* families analyzed, but these families comprise only 892 elements (4.8% of the 18 587 *Gypsy* elements). The largest copy-number family in soybean is

Table 2 Numbers of families and elements of different evolutionary lineages across species

	Arabidopsis				Rice				Soybean			
	Families		Elements		Families		Elements		Families		Elements	
Lineages	No.[a]	%	No.[a]	%	No.[a]	%	No.[a]	%	No.[a]	%	No.[a]	%
Copia												
Ivana	6	18.2	20	25.6	29	25.7	247	13.3	63	43.4	788	6.3
Maximus	3	9.1	15	19.2	6	5.3	576	31.1	15	10.3	8575	68.3
Ale	10	30.3	16	20.5	59	52.2	151	8.1	37	25.5	256	2.0
Angela	1	3.0	4	5.1	4	3.5	100	5.4	9	6.2	690	5.5
TAR	1	3.0	1	1.3	6	5.3	489	26.4	3	2.1	456	3.6
GMR	11	33.3	17	21.8	2	1.8	260	14.0	18	12.4	1799	14.3
Bianca	1	3.0	5	6.4	7	6.2	32	1.7	0	0.0	0	0.0
Subtotal	33	100	78	100	113	100	1855	100	145	100	12 564	100
Gypsy												
Renia	7	26.9	11	3.5	83	66.4	349	8.6	253	89.1	892	4.8
CR	3	11.5	6	1.9	4	3.2	83	2.0	10	3.5	6423	34.6
Tekay	2	7.7	46	14.7	13	10.4	1361	33.5	3	1.1	1385	7.5
Athila	7	26.9	185	59.3	1	0.8	4	0.1	10	3.5	4593	24.7
Tat	7	26.9	64	20.5	24	19.2	2271	55.8	8	2.8	5294	28.5
Subtotal	26	100	312	100	125	100	4068	100	284	100	18 587	100

[a]Only intact elements and solo LTRs were included.

Gmr9/*SNARE*/*GmOgre*, which belongs to the *Tat* lineage and accounts for 15% of all elements identified in the soybean genome. The copy numbers of LTR-RTs within individual families reflect the recent activities for LTR-RT amplification, while the numbers of families within individual lineages record the ancient activities. Hence, the above observations suggest that different lineages and families of LTR-RTs had distinct activities for amplification over evolutionary time.

A similar scenario was seen in rice (Figure 4, Table 2). Notably, the three lineages (*Ivana*, *Ale*, and *Reina*) that contain the highest number of families of LTR-RTs in rice are also the three lineages that contain the highest number of families of LTR-RTs in soybean, although the relative proportions of elements within these three lineages are higher in rice than in soybean. Nevertheless, either the proportions of elements within individual families or the proportions of families within individual lineages show considerable differences between soybean and rice. Compared with soybean and rice, Arabidopsis shows the overall lowest activities for LTR-RT amplification from ancient to recent times (Table 2). In particular, the lineage *TAR* may be facing extinction in Arabidopsis, as only a single intact element of this lineage was found in the entire genome. Our analysis suggests that the consistently low activities of LTR-RTs over evolutionary time are largely responsible for maintaining such a small genome, although illegitimate recombination for rapid accumulation of small deletions was found to be a primary mechanism for counteracting genome expansion during the recent evolution of the Arabidopsis genome (Devos *et al.*, 2002).

'Functional specification' of centromere retrotransposons pre-dates the separation of eudicots and monocots

Among the five *Gypsy* lineages, *CR* (centromeric retrotransposon) shows the highest level of sequence conservation between the monocot and eudicot species (Figure 4b). This lineage contains three Arabidopsis (*Atr39*, *Atr47*, and *Atr48*), four (*Oryza sativa* subspecies, *japonica*) rice (*CRR1*, *CRR2*, *CRR4*, and *rn 417-136*), and 10 soybean (*Gmr3*, *Gmr4*/*GmGypsy10*, *Gmr12*/*GmGypsy11*, *Gmr17*, *Gmr59*, *Gmr102*, *Gmr175*, *Gmr215*, *Gmr235*, *Gmr362*) families (Figures 4b and 5).

The plant *CR* elements were first discovered in cereals cytogenetically anchored to centromeres using fluorescence *in situ* hybridization (Aragon-Alcaide *et al.*, 1996; Jiang *et al.*, 1996; Miller *et al.*, 1998; Presting *et al.*, 1998). Then, two *CR* families (*CRR1* and *CRR2*) in *japonica* rice and three *CR* families (*CRM1*, *CRM2*, and *CRM3*) in maize were isolated from the respective genomes and found to be enriched in the functional centromere by chromatin immunoprecipitation (ChIP)-based analysis with a CENH3-specific antibody (Cheng *et al.*, 2002; Zhong *et al.*, 2002; Nagaki *et al.*, 2005; Sharma and Presting, 2008). *CRR1* versus *CRM3*, *CRR2* versus *CRM2*, and *CRR3* versus *CRM1* are three *CR* pairs that pre-date the divergence of maize and rice (Sharma and Presting, 2008). *CRM1* is a high-copy-number family enriched in maize centromeres, but not a single orthologous copy of *CRR3* was identified in *japonica* rice, and only an incomplete copy was detected in *indica* rice (Sharma and Presting, 2008). Other *CR* families in rice and maize analyzed previously (e.g. *CRR4* and *CRM4*) are not located in

centromeric regions. These observations suggest: (i) the association of *CR* lineages with centromeres was established before the divergence of maize and rice, (ii) some *CR* families are still associated with functional centromeres in both rice and maize, and (iii) some *CR* families have lost their roles as centromere components in these two organisms.

CR families were found to be present in the centromeric regions of most grasses that have been investigated (Aragon-Alcaide *et al.*, 1996; Jiang *et al.*, 1996), but absent in functional centromeres of *Oryza brachyantha* (Lee *et al.*, 2005), a wild species that diverged from rice about 7–9 Mya (Dawe, 2005). Instead, *FRetro3*, a LTR-RT family that belongs to the *Tekay* lineage was found to colonize the *O. brachyantha* centromeres (Gao *et al.*, 2009), representing an exception to the general *CR* conservation of grasses.

Although three families belonging to the *CR* lineages existed in Arabidopsis (Figure 5), they comprise only six elements (Table 2). Thus, it is likely that none of these three families of elements are enriched in Arabidopsis centromeres. This observation seems to echo a previous observation that no significant enrichment of *CR* homologs was detected in functional centromeres of Arabidopsis by ChIP (Nagaki *et al.*, 2003). However, no Arabidopsis centromeres have been fully sequenced; thus we are uncertain whether additional elements belonging to the three *CR* families exist in Arabidopsis. It is also possible that the copy number of *CR* elements in Arabidopsis is below the detection limits of the ChIP assay (Nagaki *et al.*, 2003).

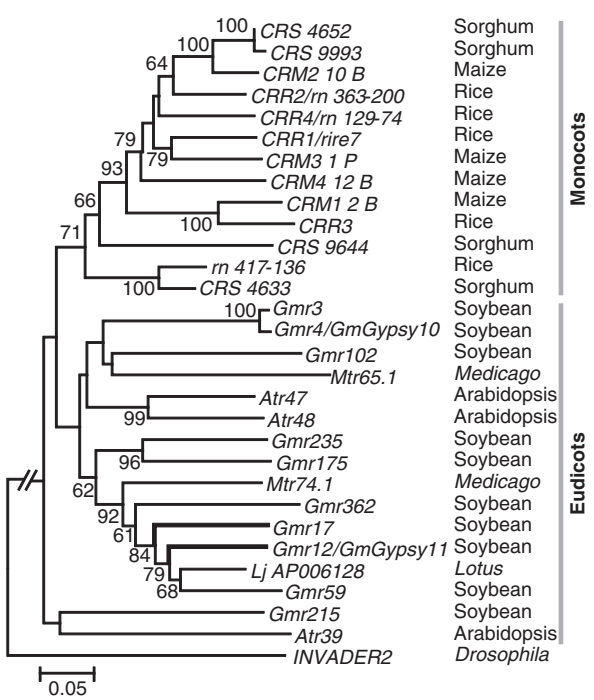

Figure 5. Phylogenetic tree of the *CR* lineage constructed using the conserved retrotransposon (RT) nucleotide sequences.
The putative soybean centromere-enriched retrotransposon families, *Gmr12*/*GmGypsy11* and *Gmr17* are marked by the bold branches. The tree was rooted using the RT sequence of the *INVIDER2* element in *Drosophila*.

Since both soybean and Arabidopsis (eudicots) share the *CR* lineage with rice and many other grass species (monocots), a standing question that intrigued us was whether the *CR* lineage were functionally specified as centromere components before the divergence of eudicots and monocots? To address this question, we developed a computational method to detect putative '*CR*' (referred to as LTR-RT families enriched in centromeres, e.g. *CRR1* and *CRR2* in rice) in plants. Because functional centromeres in most plants that have been investigated are mainly composed of large arrays of centromere satellite repeats and '*CR*' elements (Ma *et al.*, 2007), theoretically, '*CR*' elements should show stronger physical association with centromere satellite repeats than non-'*CR*' elements in a particular genome.

To test this computational method for '*CR*' identification, we first assessed the association of centromere satellite repeats with the *CR* families in *indica* rice and maize. Because the majority of centromeres have not been completely sequenced or accurately assembled by either the BAC-by-BAC approach (for *japonica* rice and maize) or the WGS approach (for soybean and *indica* rice), the assembled genome sequences were not used. Instead, we chose a set of random WGS clones from each genome that make up approximately 1× genome coverage for the association analysis. Association was measured by the percentage of clones that contain both centromere satellite repeats and the terminal sequences of a LTR-RT family versus the clones that contain the terminal sequences of the same LTR-RT family (see Experimental procedures). We found that, as expected, the '*CR*' elements, *CRR1*/*CRM3*, *CRR2*/*CRM2*, and *CRM1*, show strong association with *CentO*/*CentC* satellite repeats (5.2–9.7%) (Table 3). By contrast, *CRR4* and *CRM4* are not associated with the satellite repeats, although they are the largest and the second largest *CR* families in the rice and maize genomes, respectively (Table 3). These results validate the feasibility of the computational approach to prediction of '*CR*' families.

The physical association of centromere satellite repeats, *CentGm-1* and *CentGm-2* (Gill *et al.*, 2009), with all LTR-RT families in soybean was subsequently analyzed. We found that 16 LTR-RT families showed association with the satellite repeats, out of which *Gmr12*/*GmGypsy11* and *Gmr17* showed the strongest association (5.5 and 3.8%, respectively), and thus were considered as putative '*CR*' families in soybean. Both *Gmr12*/*GmGypsy11* and *Gmr17* belong to the *CR* lineage. Additionally, the co-localization of centromere satellite repeats and *Gmr12*/*GmGypsy11* was visualized by fluorescence *in situ* hybridization (FISH) (Figure 6). Small proportions of WGS sequences matching both LTR-RTs and satellites were found for nearly all large-copy-number families (Table 3), but these probably reflect random

Table 3 Physical association of long terminal repeat-retrotransposons (LTR-RTs) with centromere satellite repeats

			Association of LTR-RTs with satellite repeats[a]		
Species	Class	Family	No. of associated LTR-RTs	Total no. of LTR-RTs	Association rate (%)
Soybean (Williams 82)	*Gypsy*	*Gmr1*	14	3756	0.4
	Copia	*Gmr2*	76	4989	1.5
	Gypsy	*Gmr3*	45	6012	0.7
	Gypsy	*Gmr4*	78	13 829	0.6
	Copia	*Gmr5*	29	7508	0.4
	Copia	*Gmr6*	27	8444	0.3
	Gypsy	*Gmr9*	182	30 271	0.6
	Gypsy	*Gmr12*	79	1430	**5.5**[b]
	Gypsy	*Gmr17*	108	2833	**3.8**[b]
	Copia	*Gmr18*	20	4538	0.4
	Gypsy	*Gmr19*	8	4974	0.2
	Gypsy	*Gmr21*	38	2468	1.5
	Gypsy	*Gmr25*	3	4575	0.1
	Gypsy	*Gmr34*	3	2507	0.1
	Copia	*Gmr37*	7	4703	0.1
	Gypsy	*Gmr169*	2	606	0.3
Rice (93–11)	*Gypsy*	*CRR1*	9	93	**9.7**[b]
	Gypsy	*CRR2*	16	306	**5.2**[b]
	Gypsy	*CRR3*	0	1	0.0
	Gypsy	*CRR4*	0	388	0.0
Maize (B73)	*Gypsy*	*CRM1*	325	5536	**5.9**[b]
	Gypsy	*CRM2*	171	2196	**7.8**[b]
	Gypsy	*CRM3*	35	626	**5.6**[b]
	Gypsy	*CRM4*	0	4446	0.0

[a]Referring to the ratio of whole genome shotgun (WGS) sequences containing the ends of a particular family of LTR-RTs and centromere satellite repeats (i.e. *CentGm-1*/*CentGm-2* in soybean, *CentO* in rice, or *CentC* in maize) to the WGS sequences containing the ends of a same family of LTR-RTs.
[b]Centromere-enriched LTR-RT families in soybean, rice, and maize.

insertions of LTR-RTs instead of preferential enrichment of '*CR*' in soybean centromeres. This is consistent with the observations in rice that *non-CRR* families were found in the functional domain of rice centromeres (Nagaki *et al.*, 2004; Wu *et al.*, 2004; Ma and Bennetzen, 2006). If indeed these two *CR* families are associated with the functional centromeres of soybean, then it would be reasonable to deduce that the functional specification of '*CR*' as centromere components pre-dates the divergence of monocots and eudicots about 140–150 Mya (Chaw *et al.*, 2004).

As we mentioned earlier, the *CR* lineage is the most conserved between the eudicot and monocot sublineages (Figures 4b and 5). This suggests that the *CR* lineage evolves more slowly than other lineages, probably because the *CR* families had/have been selected for an important centromere function, and/or were located in a more slowly evolving portion of the genome, or both (Ma *et al.*, 2007). Similar to *CRR4* and *CRM4*, the other eight *CR* families in soybean may have lost their functional roles as centromere components during evolution of the soybean genome. The *CR* lineage was also found in *Medicago* and *Lotus* (Figure 5), but whether they are enriched in the centromeres of these two eudicots remains to be determined.

Presence and absence of the *env*-like genes in the putative plant endogenous retrovirus lineages/sublineages/families

LTR-RTs are thought to proliferate by reverse transcription, a molecular mechanism for replication of retroviruses. Unlike retroviruses from vertebrates, typical LTR-RTs do not contain an *env*-like gene that encodes a transmembrane protein, coiled-coil glycoprotein, which sponsors retroviral infection. Many, non-infectious, mammalian endogenous retroviruses also encode a transmembrane coiled-coil protein, but functional expression of these genes has not been demonstrated. Infectious endogenous retroviruses were found in *Drosophila melanogaster* (Kim *et al.*, 1994; Song *et al.*, 1994; Malik *et al.*, 2000), but whether they exist in plants remains unclear. Nevertheless, the discovery of putative *env*-like genes encoding hypothetical proteins with similar secondary structural elements immediately downstream of *pol* in some plant LTR-RTs, was interpreted as 'indirect' evidence for the discovery of endogenous plant retroviruses (Kumar, 1998; Peterson-Burch *et al.*, 2000; Miguel *et al.*, 2008). To understand the evolution of putative endogenous retroviruses in plants, we first reassessed the lineages/families/elements that have *env*-like genes in soybean, rice, and Arabidopsis in the context of their phylogenies, constructed based on the RT domains from intact elements (Figures 4 and 7).

Maximus is the only *Copia* lineage that has putative endogenous retroviral elements. This lineage contains a sublineage, composed of soybean family *Gmr2* [i.e. *SIRE* (Laten *et al.*, 1998, 2003)] and Arabidopsis family *Atr1* [i.e. *Endovir1* (Laten, 1999; Peterson-Burch *et al.*, 2000)], two putative endogenous retrovirus families previously reported. TBlastN searches using these two putative ENV-proteins as queries against genome sequences investigated in this study identified five additional families with significant matches (e-value < 10^{-6}), including two Arabidopsis families (*Atr37* and *Atr49*) and three *Lotus* families (*Lj1*, *Lj2*, and *Lj3*) (Figure 7a). No rice or *Medicago* family was found in this sublineage based on this analysis.

Athila and *Tat* are the two *Gypsy* lineages that contain putative endogenous retroviral elements based on the presence of a gene encoding a predicted transmembrane protein (Figure 7b). *Athila* contains 10 soybean families, seven Arabidopsis families, and one rice family (Vicient *et al.*, 2001). Of these 18 families, seven from soybean and five from Arabidopsis, and the rice family were found to harbor an *env*-like gene, including the previously identified putative endogenous retrovirus families *Calypso* (i.e. *Gmr11*) in soybean (Peterson-Burch *et al.*, 2000) and *Athila4*

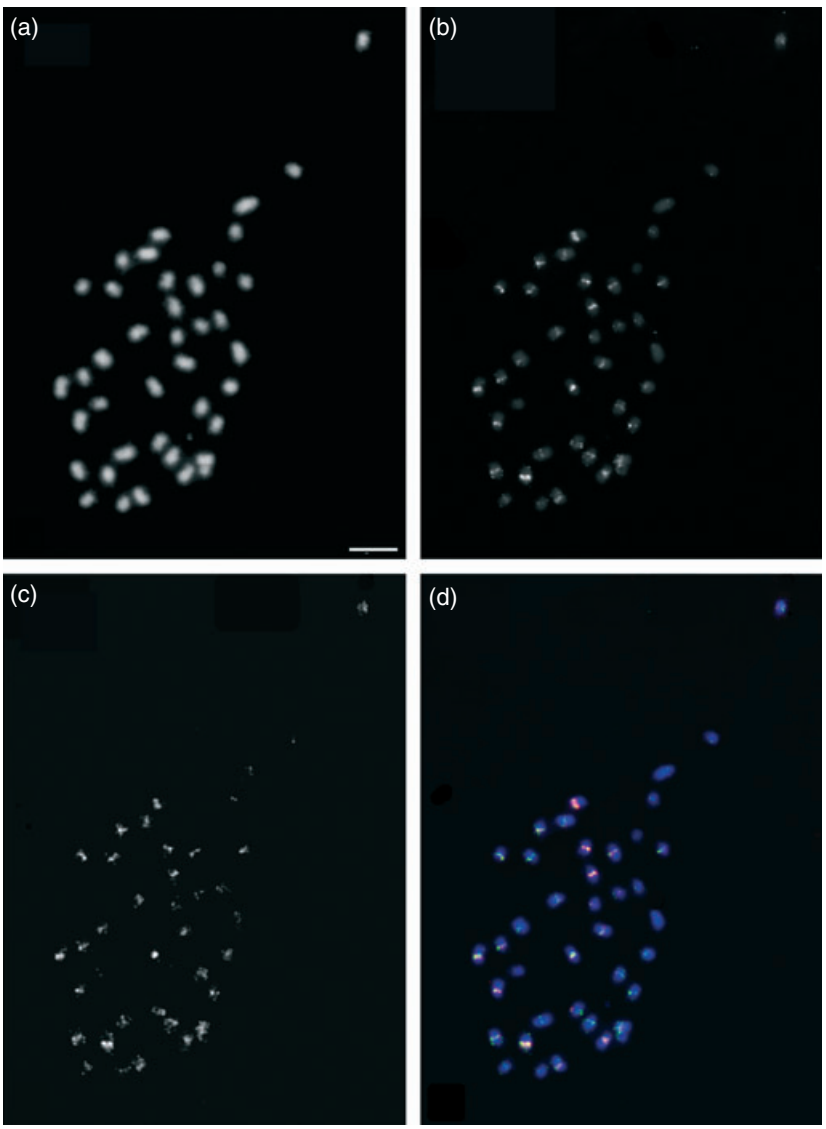

Figure 6. Co-localization of centromere satellite repeats and a putative centromere retrotransposon on mitotic chromosomes.
(a) 4′,6-diamidino-2-phenylindole (DAPI)-stained chromosomes, blue channel. Bar = 5 μm.
(b) Location of the centromeres using *CentGm-1* and *CentGm-2* as probes, red channel.
(c) *Gmr12*/*GmGypsy11* localized to chromosomes, green channel.
(d) Merged imaged: *CentGm-1* and *CentGm-2* (red), with *Gmr12*/*GmGypsy11* (green) on DAPI stained chromosomes (blue).

(i.e. *Atr9*) in Arabidopsis (Wright and Voytas, 2002). In addition, putative endogenous retroviral elements belonging to this lineage were also found in *Lotus* (*Lj18*), *Medicago* (*Mtr60* and *Mtr64*) (Figure 7b), and barley (Vicient *et al.*, 2001). If one presumes that the *env*-like genes in this lineage have a common origin, the absence of the *env*-like genes in *Gmr43*, *Gmr19*/*Diaspora* and *Gmr1* may be interpreted as the outcome of deletion of the gene, as has previously been proposed (Yano *et al.*, 2005).

The *Tat* lineage is evolutionarily close to *Athila* and contains two putative soybean endogenous retrovirus families *Gmr9*/*SNARE*/*GmOgre* and *Gmr338*. *Gmr9*/*SNARE*/*GmOgre* is the largest family in soybean, and its ENV-like protein shares approximately 31% identity with that of *Gmr11*/*Calypso*. The two largest OGRE families in *Medicago* (*Mtr57* and *Mtr59*) were found to be evolutionarily close to *Gmr9*/*SNARE*/*GmOgre*, but these OGRE families lack the *env*-like gene, indicating that either the *Gmr9*/*SNARE*/*GmOgre* family captured an *env*-like gene or the OGRE families lost the *env*-like gene after the divergence of soybean and *Medicago*. Given that the two soybean families (e.g. *Gmr9*/*SNARE*/*GmOgre* and *Gmr338*) that contain the *env*-like genes are, on average, younger than many other families that do not contain this gene (Figure 7b), it is most likely that the former captured the *env*-like genes from the *Athila* lineage.

It should be mentioned that a large number of families within two putative endogenous retroviral lineages (*Maximus* and *Tat*) possess a third ORF between the *pol* and PPT regions. Although the hypothetical proteins encoded by these extra ORFs do not have significant matches with previously described plant ENV-like proteins, some do have strong signatures of transmembrane domains, similar to retrovirus *env* genes (Figure S5; Hofmann and Stoffel, 1993). The nature and origin of the third ORF remain unclear.

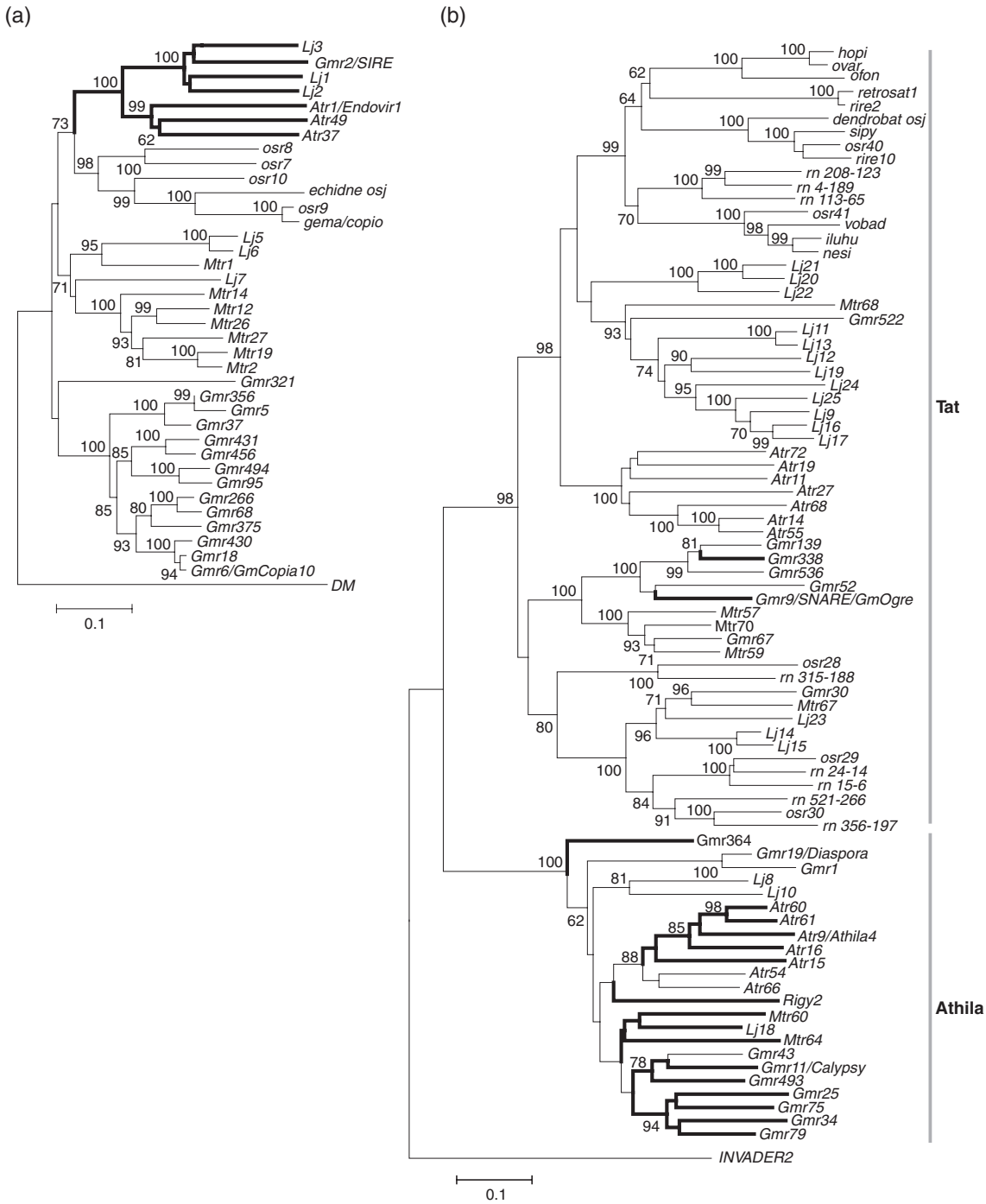

Figure 7. Phylogenetic trees of the putative plant retrovirus lineages.
(a) *Copia* families and (b) *Gypsy* families. The trees were constructed using the conserved retrotransposon (RT) nucleotide sequences, and rooted using the RT sequences of *DM* and *INVIDER2* elements in *Drosophila*. The putative retrovirus families are marked by bold branches.

Origin of *env*-like genes in retrovirus-like families in plants

Given that the *env*-like genes were found in *SIRE1* and *Athila* elements that belong to *Copia* and *Gypsy* superfamilies, respectively, it was proposed that the putative plant endogenous retroviruses have had at least two independent origins (Peterson-Burch *et al.*, 2000). However, this hypothesis needs to be made with the caveat that the putative *Copia* and *Gypsy* retrovirus lineages evolved independently. To shed light on the origin and evolution of the *env*-like genes,

we performed phylogenetic analysis of the putative endogenous retroviral elements using the conserved RT proteins and ENV domains, respectively, and compared the PBS and PPT motifs from these elements. As expected, the *Copia* families and *Gypsy* families were separated into distinct lineages, and the soybean elements and Arabidopsis elements were clearly distinguished based on their RT domains (Figure 8a). Furthermore, the nucleotide similarities of the PBS and PPT motifs (Figure 8b) perfectly reflect the relationships among these families as revealed by the RT domains. By contrast, the ENV-like domains between the *Copia* and *Gypsy* families were not distinguished by two distinct groups (Figure 8c, Figure S6), indicating that the *env*-like genes may not be the components of the common ancestor of the *Copia* and *Gypsy* superfamilies. In other words, either *Copia* or *Gypsy*, or both superfamilies, may

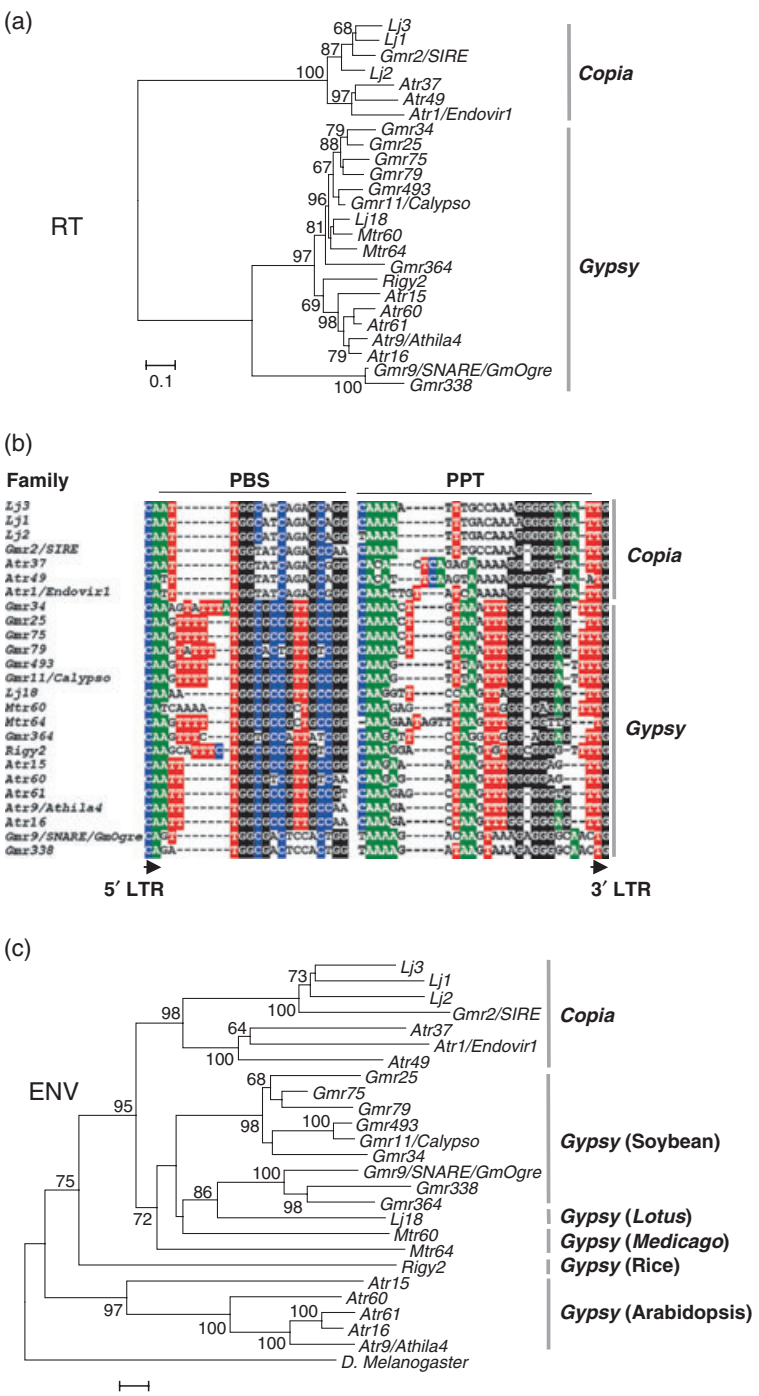

Figure 8. Evolutionary relationships of the putative plant retroviruses.
(a) Phylogenetic tree constructed using the retrotransposon (RT) protein domains.
(b) The relationship of the putative retroviruses reflected by the primer-binding site (PBS) and polypurine tract (PPT) sites.
(c) Phylogenetic tree constructed using the ENV-like protein domains, which was rooted using a putative retrovirus element in *Drosophila* (accession number gi140836).

have captured the *env*-like genes after their bifurcation. Nevertheless, the RT and ENV-like domains from the putative *Copia* endogenous retrovirus families of soybean, *Lotus*, and Arabidopsis exhibit consistent phylogenetic relationships, suggesting that the *env*-like genes in these *Copia* elements have a common origin. The average similarity among the ENV-like domains included in Figure 8(c) is 21%. If the phylogeny of the ENV domains shown in Figure 8(c) reflects the origin and evolution of the *env*-like genes, it would be reasonable to deduce that the putative *Copia* endogenous retroviral elements captured the *env*-like genes from the putative *Gypsy* endogenous retroviral elements. Together, these data are in favor of the hypothesis that the *env*-like genes in the *Copia* and *Gypsy* endogenous retroviral elements in plants may have a common origin.

Rigy2 is the only putative endogenous retrovirus family identified in rice from this analysis. However, neither the RT nor the ENV-like domains separate this family from the putative endogenous retrovirus families in soybean and Arabidopsis (Figures 7b and 8b). On the other hand, the relationships between *Rigy2* and the putative *Gypsy* endogenous retrovirus families in soybean and Arabidopsis reflected by both domains appear to be consistent. Whether *Rigy2* represents an ancient horizontal transfer from eudicots to rice remains unclear.

EXPERIMENTAL PROCEDURES

Identification and classification of LTR-RTs

A combination of structural analysis and sequence homology comparisons were used to identify LTR-RTs in the assembled soybean genome (pseudomolecule assembly version Glyma1.01). Initially, the *LTR_STRUC* program was employed for the identification of intact elements (McCarthy and McDonald, 2003). The intact elements missed by the program and solo LTRs were identified by methods previously described (Ma and Bennetzen, 2004; Ma *et al.*, 2004). Truncated elements and fragments were not considered in this study. The structures and boundaries of all of the identified LTR-RTs were confirmed by manual inspection. The LTR-RTs were classified into *Copia*-like (INT-RT-RH) and *Gypsy*-like (RT-RH-INT) superfamilies, and individual families were defined by the criteria described previously (Wicker *et al.*, 2007).

Following the above approach, we mined the LTR-RTs from the latest annotated Arabidopsis genome (TAIR 9; http://www.arabidopsis.org). The Arabidopsis families identified in this study were named as '*Atr*' (Arabidopsis retrotransposon). *CR* elements in sorghum were identified and classified using the same approach (Ma and Bennetzen, 2004; Ma *et al.*, 2004; Paterson *et al.*, 2009). The LTR-RTs from rice, *Medicago*, and *Lotus* were obtained from previous studies (Holligan *et al.*, 2006; Wang and Liu, 2008; Tian *et al.*, 2009).

Estimation of insertion time

Intact elements with two available LTR sequences were aged by comparing their 5′ and 3′ LTRs. Two LTRs were aligned using the MUSCLE program (Edgar, 2004). If needed, the alignments were manually inspected and corrected using the BioEdit program. The distance (*K*) between two LTRs was corrected by the Jukes–Cantor method (Kimura and Ota, 1972). An average substitution rate (*r*) of 1.3×10^{-8} substitutions per synonymous site per year (Ma and Jackson, 2006) was used for calculations. The time (*T*) since intact element insertion was estimated using the formula $T = k/2r$.

Phylogenetic analysis

Typical *Copia*-like or *Gypsy*-like conserved RT protein sequences were set as queries, to search against soybean LTR-RT database. For the elements with significant hits (e-value < 10^{-9}), translated RT cDNA sequences and RT protein sequences were extracted, aligned, and manually inspected. One typical element from each family was chosen for phylogenetic analysis. ENV-like protein domains were predicted using ORF finder (http://www.ncbi.nlm.nih.gov/projects/gorf) and TBlastN searches. The bootstrap neighbor-joining trees were built using the Kimura two-parameter method integrated in the MEGA4 program (Tamura *et al.*, 2007).

Physical association analysis between LTR-RTs and satellite repeats

Initially intact elements and solo LTRs from individual families were identified from the assembled soybean genome and maize BAC sequences (Schnable *et al.*, 2009). Then trimmed WGS sequences (each >300 bp) from soybean (c.v. Williams 82), rice (c.v. 93-11), and maize (c.v. B73), were extracted to perform physical association analysis (ftp://ftp.ncbi.nlm.nih.gov/pub/TraceDB). A total of 150 bp from both ends of each intact element were extracted, and set as queries to search against WGS sequences. An association was counted if a contig contained both element and centromere satellite repeats *CentGm-1*/*CentGm-2*, *CentO*, or *CenC*.

Fluorescence *in situ* hybridization of LTR-RTs

Seeds (c.v. Williams 82) were germinated in moist filter paper in a 37°C incubator in the dark for 3 days. Chromosomes were prepared according to published protocols (Walling *et al.*, 2005) and FISH was done according to Walling *et al.* (2006) and Gill *et al.* (2009). Chromosome spreads were located by scanning the slides with a 60× oil immersion lens on a Nikon Eclipse 80*i* microscope (http://www.nikon.com/) and single-channel (red, green, and blue) images were taken with a 100× objective using a Photometrics CoolSnap HQ CCD camera (http://www.photomet.com/). Color overlays were made using the Color Combine function from MetaVue software from Molecular Devices (http://www.moleculardevices.com/).

ACKNOWLEDGEMENTS

We thank Dr Vini Pereira for providing an Arabidopsis transposable elements dataset. This study was supported by USDA-ARS Specific Cooperative Agreement to JM, Purdue University faculty startup funds to JM, and National Science Foundation Plant Genome Research Program (DBI-0822258) to SAJ, RCS, and JM.

SUPPORTING INFORMATION

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

Figure S1. Size distribution of intact elements (A) and solo long terminal repeats (LTRs) (B) identified in the soybean genome.

Figure S2. Phylogenetic relationship and divergence time of eudicot and monocot species investigated in this study.

Figure S3. Phylogenetic tree constructed using *Copia* conserved retrotransposon (RT) nucleotide sequences.

Figure S4. Phylogenetic tree constructed using *Gypsy* conserved retrotransposon (RT) sequences.

Figure S5. TMpred output for the third open reading frame (ORF) of retrovirus related families.

Figure S6. Alignment of putative endogenous ENV-like proteins used in Figure 8c.
Table S1. Summary of long terminal repeat-retrotransposon (LTR-RT) families in soybean.
Table S2. Summary of long terminal repeat-retrotransposon (LTR-RT) families in Arabidopsis.
Please note: As a service to our authors and readers, this journal provides Supporting Information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from Supporting Information (other than missing files) should be addressed to the authors.

REFERENCES

Aragon-Alcaide, L., Miller, T., Schwarzacher, T., Reader, S. and Moore, G. (1996) A cereal centromeric sequence. *Chromosoma*, **105**, 261–268.

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

Bennetzen, J.L., Coleman, C., Liu, R., Ma, J. and Ramakrishna, W. (2004) Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7**, 732–736.

Bennetzen, J.L., Ma, J. and Devos, K.M. (2005) Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond)*, **95**, 127–132.

Chaw, S.M., Chang, C.C., Chen, H.L. and Li, W.H. (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J. Mol. Evol.* **58**, 424–441.

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R. and Jiang, J. (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell*, **14**, 1691–1704.

Choi, H.K., Mun, J.H., Kim, D.J. *et al.* (2004) Estimating genome conservation between crop and model legume species. *Proc. Natl Acad. Sci. USA*, **101**, 15289–15294.

Dawe, R.K. (2005) Centromere renewal and replacement in the plant kingdom. *Proc. Natl Acad. Sci. USA*, **102**, 11573–11574.

Devos, K.M., Brown, J.K. and Bennetzen, J.L. (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* **12**, 1075–1079.

Du, J., Grant, D., Tian, Z., Nelson, R.T., Zhu, L., Shoemaker, R.C. and Ma, J. (2010a) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. *BMC Genomics*, **11**, 113.

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

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

Gao, D., Gill, N., Kim, H.R. *et al.* (2009) A lineage-specific centromere retrotransposon in *Oryza brachyantha*. *Plant J.* **60**, 820–831.

Gill, N., Findley, S., Walling, J.G., Hans, C., Ma, J., Doyle, J., Stacey, G. and Jackson, S.A. (2009) Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol.* **151**, 1167–1174.

Graham, P.H. and Vance, C.P. (2003) Legumes: importance and constraints to greater use. *Plant Physiol.* **131**, 872–877.

Hofmann, K. and Stoffel, W. (1993) A database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler*, **374**, 166.

Holligan, D., Zhang, X., Jiang, N., Pritham, E.J. and Wessler, S.R. (2006) The transposable element landscape of the model legume *Lotus japonicus*. *Genetics*, **174**, 2215–2228.

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

Jiang, J., Nasuda, S., Dong, F., Scherrer, C.W., Woo, S.S., Wing, R.A., Gill, B.S. and Ward, D.C. (1996) A conserved repetitive DNA element located in the centromeres of cereal chromosomes. *Proc. Natl Acad. Sci. USA*, **93**, 14210–14213.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Kashkush, K. and Khasdan, V. (2007) Large-scale survey of cytosine methylation of retrotransposons and the impact of readout transcription from long terminal repeats on expression of adjacent rice genes. *Genetics*, **177**, 1975–1985.

Kashkush, K., Feldman, M. and Levy, A.A. (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* **33**, 102–106.

Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Purd'homme, N. and Bucheton, A. (1994) Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **91**, 1285–1289.

Kimura, M. and Ota, T. (1972) On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.* **2**, 87–90.

Kumar, A. (1998) The evolution of plant retroviruses: moving to green pastures. *Trends Plant Sci.* **3**, 371–374.

Kumar, A. and Bennetzen, J.L. (1999) Plant retrotransposons. *Annu. Rev. Genet.* **33**, 479–532.

Laten, H.M. (1999) Phylogenetic evidence for *Ty1-copia*-like endogenous retroviruses in plant genomes. *Genetica*, **107**, 87–93.

Laten, H.M., Majumdar, A. and Gaucher, E.A. (1998) SIRE-1, a *copia/Ty1*-like retroelement from soybean, encodes a retroviral envelope-like protein. *Proc. Natl Acad. Sci. USA*, **95**, 6897–6902.

Laten, H.M., Havecker, E.R., Farmer, L.M. and Voytas, D.F. (2003) SIRE1, an endogenous retrovirus family from *Glycine max*, is highly homogeneous and evolutionarily young. *Mol. Biol. Evol.* **20**, 1222–1230.

Laten, H.M., Mogil, L.S. and Wright, L.N. (2009) A shotgun approach to discovering and reconstructing consensus retrotransposons ex novo from dense contigs of short sequences derived from Genbank Genome Survey Sequence database records. *Gene*, **448**, 168–173.

Lavin, M., Herendeen, P.S. and Wojciechowski, M.F. (2005) Evolutionary rates analysis of *Leguminosae* implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* **54**, 575–594.

Lee, H.R., Zhang, W., Langdon, T., Jin, W., Yan, H., Cheng, Z. and Jiang, J. (2005) Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. *Proc. Natl Acad. Sci. USA*, **102**, 11793–11798.

Ma, J. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA*, **101**, 12404–12410.

Ma, J. and Bennetzen, J.L. (2006) Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc. Natl Acad. Sci. USA*, **103**, 383–388.

Ma, J. and Jackson, S.A. (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res.* **16**, 251–259.

Ma, J., Devos, K.M. and Bennetzen, J.L. (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869.

Ma, J., SanMiguel, P., Lai, J., Messing, J. and Bennetzen, J.L. (2005) DNA rearrangement in orthologous orp regions of the maize, rice and sorghum genomes. *Genetics*, **170**, 1209–1220.

Ma, J., Wing, R.A., Bennetzen, J.L. and Jackson, S.A. (2007) Evolutionary history and positional shift of a rice centromere. *Genetics*, **177**, 1217–1220.

Malik, H.S., Henikoff, S. and Eickbush, T.H. (2000) Poised for contagion, evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10**, 1307–1318.

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

Miguel, C., Simoes, M., Oliveira, M.M. and Rocheta, M. (2008) Envelope-like retrotransposons in the plant kingdom: evidence of their presence in gymnosperms (*Pinus pinaster*). *J. Mol. Evol.* **67**, 517–525.

Miller, J.T., Dong, F., Jackson, S.A., Song, J. and Jiang, J. (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. *Genetics*, **150**, 1615–1623.

Nagaki, K., Talbert, P.B., Zhong, C.X., Dawe, R.K., Henikoff, S. and Jiang, J. (2003) Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of *Arabidopsis thaliana* centromeres. *Genetics*, **163**, 1221–1225.

Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P.B., Kim, M., Jones, K.M., Henikoff, S., Buell, C.R. and Jiang, J. (2004) Sequencing of a rice centromere uncovers active genes. *Nat. Genet.* **36**, 138–145.

Nagaki, K., Neumann, P., Zhang, D., Ouyang, S., Buell, C.R., Cheng, Z. and Jiang, J. (2005) Structure, divergence, and distribution of the *CRR* centromeric retrotransposon family in rice. *Mol. Biol. Evol.* **22**, 845–855.

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

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

Peterson-Burch, B.D., Wright, D.A., Laten, H.M. and Voytas, D.F. (2000) Retroviruses in plants? *Trends Genet.* **16**, 151–152.

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

Presting, G.G., Malysheva, L., Fuchs, J. and Schubert, I. (1998) A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant J.* **16**, 721–728.

SanMiguel, P. and Vitte, C. (2009) The LTR retrotransposons of maize. In *The Maize Handbook* – Volume II: Domestication, Genetics and Genomics (Bennetzen, J.L. and Hake, S., eds). New York: Springer, pp. 307–329.

SanMiguel, P., Tikhonov, A., Jin, Y.K. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.

Schlueter, J.A., Lin, J.Y., Schlueter, S.D. *et al.* (2007) Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. *BMC Genomics*, **8**, 330.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

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

Sharma, A. and Presting, G.G. (2008) Centromeric retrotransposon lineages predate the maize/rice divergence and differ in abundance and activity. *Mol. Genet. Genomics*, **279**, 133–147.

Shoemaker, R.C., Schlueter, J. and Doyle, J.J. (2006) Paleopolyploidy and gene duplication in soybean and other legumes. *Curr. Opin. Plant Biol.* **9**, 104–109.

Song, S.U., Gerasimova, T., Kurkulos, M., Boeke, J.D. and Corces, V.G. (1994) An env-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. *Genes Dev.* **8**, 2046–2057.

Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L. and Messing, J. (2004) Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599.

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Tian, Z., Rizzon, C., Du, J., Zhu, L., Bennetzen, J.L., Jackson, S.A., Gaut, B.S. and Ma, J. (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* **19**, 2221–2230.

Tuskan, G.A., Difazio, S., Jansson, S. *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Vicient, C.M., Kalendar, R. and Schulman, A.H. (2001) Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Res.* **11**, 2041–2049.

Vitte, C. and Bennetzen, J.L. (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc. Natl Acad. Sci. USA*, **103**, 17638–17643.

Walling, J.G., Pires, J.C. and Jackson, S.A. (2005) Preparation of samples for comparative studies of plant chromosomes using in situ hybridization methods. *Methods Enzymol.* **395**, 443–460.

Walling, J.G., Shoemaker, R., Young, N., Mudge, J. and Jackson, S. (2006) Chromosome-level homeology in paleopolyploid soybean (*Glycine max*) revealed through integration of genetic and chromosome maps. *Genetics*, **172**, 1893–1900.

Wang, H. and Liu, J.S. (2008) LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. *BMC Genomics*, **9**, 382.

Wawrzynski, A., Ashfield, T., Chen, N.W. *et al.* (2008) Replication of nonautonomous retroelements in soybean appears to be both recent and common. *Plant Physiol.* **148**, 1760–1771.

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

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

Wright, D.A. and Voytas, D.F. (2002) *Athila4* of Arabidopsis and *Calypso* of soybean define a lineage of endogenous plant retroviruses. *Genome Res.* **12**, 122–131.

Wu, J., Yamagata, H., Hayashi-Tsugane, M. *et al.* (2004) Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell*, **16**, 967–976.

Xiong, Y. and Eickbush, T.H. (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**, 3353–3362.

Yano, S.T., Panbehi, B., Das, A. and Laten, H.M. (2005) Diaspora, a large family of *Ty3-gypsy* retrotransposons in *Glycine max*, is an envelope-less member of an endogenous plant retrovirus lineage. *BMC Evol. Biol.* **5**, 30.

Zhong, C.X., Marshall, J.B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J.A., Jiang, J. and Dawe, R.K. (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. *Plant Cell*, **14**, 2825–2836.

