Comparative Genomic Analysis Reveals Multiple Long Terminal Repeats, Lineage-Specific Amplification, and Frequent Interelement Recombination for *Cassandra* Retrotransposon in Pear (*Pyrus bretschneideri* Rehd.)

Hao Yin[1,†], Jianchang Du[2,†], Leiting Li[1], Cong Jin[1], Lian Fan[1], Meng Li[1], Jun Wu[1], and Shaoling Zhang[1,*]

[1]State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Horticulture, Nanjing Agricultural University, China

[2]Bioinformatics Group, Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, China

*Corresponding author: E-mail: slzhang@njau.edu.cn.

[†]These authors contributed equally to this work.

Accepted: May 21, 2014

Abstract

Cassandra transposable elements belong to a specific group of terminal-repeat retrotransposons in miniature (TRIM). Although *Cassandra* TRIM elements have been found in almost all vascular plants, detailed investigations on the nature, abundance, amplification timeframe, and evolution have not been performed in an individual genome. We therefore conducted a comprehensive analysis of *Cassandra* retrotransposons using the newly sequenced pear genome along with four other Rosaceae species, including apple, peach, mei, and woodland strawberry. Our data reveal several interesting findings for this particular retrotransposon family: 1) A large number of the intact copies contain three, four, or five long terminal repeats (LTRs) (~20% in pear); 2) intact copies and solo LTRs with or without target site duplications are both common (~80% vs. 20%) in each genome; 3) the elements exhibit an overall unbiased distribution among the chromosomes; 4) the elements are most successfully amplified in pear (5,032 copies); and 5) the evolutionary relationships of these elements vary among different lineages, species, and evolutionary time. These results indicate that *Cassandra* retrotransposons contain more complex structures (elements with multiple LTRs) than what we have known previously, and that frequent interelement unequal recombination followed by transposition may play a critical role in shaping and reshaping host genomes. Thus this study provides insights into the property, propensity, and molecular mechanisms governing the formation and amplification of *Cassandra* retrotransposons, and enhances our understanding of the structural variation, evolutionary history, and transposition process of LTR retrotransposons in plants.

Key words: *Cassandra* retrotransposon, TRIM, amplification, recombination, pear, Rosaceae.

Introduction

Transposable elements (TEs) are DNA components that are capable of moving from one place to another in a genome. Based on their transposition process, TEs can be classified into 1) retrotransposons, which use RNA as an intermediate, and 2) DNA transposons that transpose via a DNA intermediate. It has been well documented that long terminal-repeat retrotransposons (LTR-RTs) are major DNA components in plants. For example, about 19% of peach (Verde et al. 2013), 43% of pear (Wu et al. 2012), 53% of cotton (Paterson et al. 2012) and over 70% of maize genomes (Schnable et al. 2009) are composed of LTR-RTs.

LTR-RTs can be separated into two groups, autonomous and nonautonomous, based on their structural completeness and the capacity of transposition. Both autonomous and nonautonomous LTR-RTs comprise two LTRs, a primer binding site (PBS) and a polypurine tract (PPT) site. Autonomous LTR-RTs usually contain a full set of genes, which encode several proteins related with transposition process, such as *gag* involved in the maturation and packaging of retrotransposon RNA, and *pol* genes comprised protease (*pro*), RNase H (*rt*), reverse transcriptase (*rt*), and integrase (*int*) (Kumar and Bennetzen 1999). In contrast, nonautonomous LTR-RTs usually lack at least one necessary gene, which prevent them from generating proteins required for transposition, and have to rely on their autonomous partners to continue to move in the host genome (Wicker and Keller 2007). It should be noticed that some nonautonomous LTR-RTs, such as *BARE2* elements in barley (Tanskanen et al. 2007), *Dasheng* elements in rice (Jiang, Bao, et al. 2002), and *SNRE* elements in soybean (Du, Tian, Bowen, et al. 2010), can be amplified to up to more than 1,000 copies in their host genomes within a very short timeframe, indicating that structural incompleteness does not affect their moving capability.

Two types of nonautonomous LTR-RTs have been described in plant genomes, including Large Retrotransposon Derivatives (LARDs) (Kalendar et al. 2004) and Terminal-repeat Retrotransposons In Miniature (TRIMs) (Witte et al. 2001). In LARDs elements, the coding region is replaced by a large conserved noncoding DNA sequence (usually >4 kb) whereas in TRIM elements, the internal part between two LTRs is very short and thus the whole element is very small (typically <1 kb). LARDs and TRIMs are presumed to be derivatives of their autonomous copies, but in practice, most such autonomous elements cannot be found, and this makes their origin remain mysterious. *Cassandra* is a particularly interesting group of TRIM elements. Because the elements in this family carry approximately 120-bp conserved 5 S RNA domain within two LTR regions, which is associated with RNA polymerase (pol) III promoters and terminators. As *Cassandra* elements have been found in both monocot and eudicot species, they have been presumed to be ancient and their origin can be traced at least to the Permian, 250 Ma (Antonius-Klemola et al. 2006; Kalendar et al. 2008).

As one of the most economically important angiosperm lineages, the Rosaceae family comprises approximately 90 genera and over 3,000 distinct species with chromosome numbers from 7 to 17 pairs (Kalkman 2004). Some fleshy-fruited genera have been widely cultivated due to economic value, including apple (*Malus*), pear (*Pyurs*), peach (*Prunus*), strawberry (*Fragria*), chokeberry (*Aronia*), loquat (*Eriobotrya*), and quince (*Cydonia*). A previous study based on DNA sequence data has classified the genus into three subfamilies, Dryadoideae, Spiraeoideae, and Rosoideae, and each can be further separated into several supertribes and tribes. For example, *Malus* and *Pyrus* are included in the Spiraeoideae, supertribe Pyrodeae, tribe Pyreae; *Prunus* belongs to the Spiraeoideae, tribe Amygdaleae; and *Fragaria* can be included in the Rosoideae, supertribe Rosadea, tribe Potentilleae (Potter et al. 2007).

Our group led the effort to finish and release the genomic sequence of the third most important temperate fruit species, pear (Wu et al. 2012). The availability of four other Rosaceae genomic sequences, including apple (*Malus domestica*) (Velasco et al. 2010), peach (*Prunus persica*) (Verde et al. 2013), mei (*Pr. mume*) (Zhang et al. 2012), and woodland strawberry (*F. vesca*) (Shulaev et al. 2011), has provided good opportunities to compare the structure, abundance, amplification timeframe of TEs within and between closely related species. In this study, we first conduct a genome-wide identification of *Cassandra* elements in pear using both structure-based and homologous search approaches, and then similarity searches are performed and homology is inferred from four other genomic sequences. We have also investigated the target site specificity, the phylogenetic relationships, and the orthologous copies of *Cassandra* TRIMs. Our data show that many *Cassandra* copies contain three, four, or five LTRs, particularly in pear, and that *Cassandra* copies with or without target site duplications (TSDs) are present very frequently in Rosaceae species. In addition, the pear genome was found to be occupied by more *Cassandra* copies than other Rosaceae species. Thus our new analysis reveals novel structures, differential amplification, and frequent interelement recombination of *Cassandra* elements, providing additional information and knowledge regarding the structure and evolution of TEs in plants.

Materials and Methods

Genome Sequence Data and Annotation of LTR-RTs

Genome sequence data information for the five Rosaceae species is available in supplementary table S1, Supplementary Material online, including genome size, chromosome number, websites, as well as the statistics of genome assemblies (sequencing technology, gene number, repetitive sequence rate, scaffold number, scaffold N50).

Initially, the draft pear genome sequence was scanned by employing the *LTR_STRUC* program to search the relatively young LTR-RT elements, and the annotated *PbrCassandra* elements were manually inspected. To detect the elements missed by the program, the intact elements sequences and LTR sequences of all 27 identified *PbrCassandra* elements were used as queries to scan the whole genome sequences of pear by using the "cross_match" program with default parameters. In order to detect the homologous elements in the other four related genomes, the LTR sequences of *PbrCassandra* elements were used as queries to scan the apple, peach, mei, and woodland strawberry genomes using the cross_match program with default parameters, and then the sequences of intact *MdCassandra*, *PpCassandra*, *PmCassandra,* and *FvCassandra* elements were extracted and used as queries to search their genome sequences again by using the cross_match program with default parameters. In this study, the intact elements represent those elements with two intact LTRs containing identified PPT site and PBS; solo LTRs are elements only containing an intact LTR sequence; and incomplete elements with sequence length over half of the intact elements were defined as truncated elements (Ma et al. 2004). All the intact elements and solo *Cassandra* copies flanking with 5-bp TSDs were computationally verified to make sure that each of them contains the TSDs with one nucleotide mismatch allowed.

Randomization Analysis of the Genomic Distribution of *PbrCassandra* Elements

We performed a randomization test to analyze whether *PbrCassandra* elements are randomly distributed in the pear genome by using an in-house perl script. It runs in the following way: Initially, the assembled 378 Mb of 17 chromosomes was separated into 378 nonoverlapped 1-Mb windows, and the observed copy number (OCN) of *PbrCassandra* in each window was calculated. Then a total of 3,940 elements were individually reassigned to a randomly selected position in the 378 windows, which was repeated 10,000 times, and the random generated copy number (RGCN) of *PbrCassandra* in each window was also calculated for each time. For each window, the times (n) were counted when RGCN was smaller than OCN, and the formula $P = (n + 1)/(10{,}000 + 1)$ was used to calculate and to test whether the *PbrCassandra* elements were randomly distributed. If the OCN in one window is much smaller than most of the repeated RGCNs ($P < 0.025$) or the opposite ($P > 0.975$), we rejected the hypothesis that the *PbrCassandra* elements are distributed randomly in the window.

PCR and Sequencing Analysis

Total genomic DNA of the pear cultivar "Dangshansuli" (*Py. bretschneideri* Rehd.) was extracted from the young leaves by using the improved CTAB method. Fifteen *PbrCassandra* copies were randomly selected and their 300-bp 5′-flanking sequences and 300-bp 3′-flanking sequences were extracted and used to design primers, respectively (supplementary table S2, Supplementary Material online). Polymerase chain reactions (PCR) were in a total volume of 25 µl, containing: 1 µl of 50 ng/µl genomic DNA template, 2.5 µl of 10× buffer (without $MgCl_2$), 2.5 µl of 2.5 mM dNTP mixture, 2.5 µl of 25 mM $MgCl_2$, 0.8 µl each of forward and reverse primer (10 pmol/µl), and 0.2 µl of 5 U/µl Taq polymerase (Takara Biotechnology Company, Dalian). The reactions were performed with the following conditions: 94 °C for 3 min, then 35 cycles of 94 °C for 30 s, 57 °C for 40 s, and 72 °C for 2 min, and a final step at 72 °C for 10 min. The products were resolved on 1% agarose and detected by EB (ethidium bromide) staining. The analyses were performed three times and loaded on independent gels.

The PCR products of one randomly selected *PbrCassandra* copy with three LTRs (PbrCassandraI_T687) and one copy with four LTRs (PbrCassandraI_T993) were isolated with the DNA Gel Extraction kit AxyPrep (Axygen Inc.). The fragments were cloned into the pMD19-T vector and sequenced by Invitrogen (Shanghai, China).

Estimation of Insertion Time and Clade Time

The intact elements with TSD sequences were used to estimate the insertion time by comparing the divergence of their 5′- and 3′-LTRs, which were believed to be identical at the time of integration (SanMiguel et al. 1998). For each element, to investigate the nucleotide substitution rate, the MUSCLE3.8.31 program was employed to align the two LTR sequences with default parameters (Edgar 2004). The insertion time (T) for each intact element was calculated with the formula: $T = K/2r$, the average number of substitutions per aligned site (K) was corrected by the Jukes–Cantor method (Kimura and Ota 1972), and 1.3×10^{-8} substitution per site per year was used as the average substitution rate of LTRs (r) (Ma and Bennetzen 2006). The age (T) of each phylogenetic clade of elements (fig. 4) was estimated using the formula: $T = K/r$ (Jiang, Jordan, et al. 2002). The average distance (K) was calculated by the alignment of LTR sequence of each intact element in a clade with the consensus LTR sequence of that clade (Kapitonov and Jurka 1996; Costas and Naveira 2000). The consensus sequence of each clade was obtained from the EMBL consensus sequence server (http://coot.embl.de/Alignment//consensus.html, last accessed June 3, 2014) with the cutoff of 50%. The average mutation rate (r) of LTRs is 1.3×10^{-8} substitution per synonymous site per year (Ma and Bennetzen 2006).

Identification of the Insertions of Orthologous *Cassandra* Copies between Species

To identify the insertions of orthologous *Cassandra* copies between species, we employed a modified strategy on the basis of previous studies (Ma and Bennetzen 2006; Tian et al. 2012; Yin et al. 2013). This procedure included three steps: 1) One or two 100-bp (50-bp flanking sequences and 50-bp LTR-RT terminal sequence) sequences from *PbrCassandra, MdCassandra, PpCassandra, PmCassandra,* and *FvCassandra* were extracted as five query databases, including intact elements with TSDs, solo elements with TSDs and truncated elements with one LTR deleted; 2) the five species' *Cassandra* query databases were used to scan each of the other four genome sequences using the cross_match program with the default parameters; 3) if a cross_match query yielded only one hit, it was deemed as a insertion of orthologous *Cassandra* copy, if two or more hits in each genome (indicating this region with query sequence corresponding to a duplication events), the insertions were all excluded from this analysis.

Phylogenetic Analysis

Sequence alignments were performed by MUSCLE3.8.31 program with default options (Edgar 2004). MEGA 5.0 program implemented with P-distance model was employed for the neighbor-joining trees building (Tamura et al. 2011). The analysis was based on 500 bootstrap replicates (Kalendar et al. 2008).

Results

Identification, Structural Characterization, and Sequence Analysis of *Cassandra* Elements in the Pear Genome

By screening the pear genome sequence using "LTR_STRUC" program (McCarthy and McDonald 2003), 1,597 LTR-RTs flanked by perfect 5-bp TSDs were initially identified. Careful examination of these elements reveals some interesting findings. That is, 27 elements have multiple LTRs (each ~268–285 bp; fig. 1*A* and *B*), including 20 elements with three LTRs (3 L-type), 5 elements with four LTRs (4 L-type), and 2 elements with five LTRs (5 L-type). Each internal LTR is flanked by a PBS and a PPT motif (fig. 1*C*). In addition, the terminal two LTRs contain two conserved dinucleotides "TG" and "CA," and 11-bp conserved terminal inverted repeats (5′-TGTAACATCCC . . . GGGATGTGACA-3′), and the internal sequence between two LTRs is very short (~69–75 bp; fig. 1*A*). BLAST searches using these elements as queries against National Center for Biotechnology Information (NCBI) nucleotide database reveal a perfect matches to *Cassandra,* a TRIM family previously reported (Witte et al. 2001; Kalendar et al. 2008). Therefore, these 27 LTR-RTs in pear could be classified as *Cassandra*, and have been named *PbrCassandra*.

To investigate the abundance and the complete picture of *PbrCassandra* elements in the pear genome, a combination of structure-based and similarity-based approaches was employed as previously described (Ma and Bennetzen 2004; Du, Grant, et al. 2010). Overall, 5,032 copies of *PbrCassandra* were identified in the 512 Mb assembled pear genomic sequence, including 1,175 (23.3%) intact copies with TSDs, 198 (3.9%) intact copies without TSDs, 788 (15.7%) solo LTRs with TSDs, 250 (5.0%) solo LTRs without TSDs, and 2,621 (52.1%) truncated copies with at least one LTR partially deleted (table 1 and supplementary table S3, Supplementary Material online). Together with numerous unrecognizable related fragments, *PbrCassandra* elements make up approximately 4.1 Mb of DNA sequence, accounting for approximately 0.8% of the pear genomic sequence. The ratio of solo LTRs to intact elements (with TSDs) was estimated to be approximately 0.88:1 (table 1), which is quite similar to that for *Jinling* elements (0.71:1) and *Rider* elements (0.92:1) in tomato (Jiang et al. 2009). In the 1,175 intact copies with TSDs, many were found to contain multiple LTRs, including 182 copies with three LTRs (15.5%), 22 copies with four LTRs (1.6%), 5 copies with five LTRs (0.4%), and other 966 copies containing two typical LTRs (82.2%). In the 198 intact without TSDs, the corresponding numbers for the above categories are 40 (20.2%), 3 (1.5%), 0 (0%), and 155 (78.3%), respectively (table 2 and supplementary table S8, Supplementary Material online).

To verify the *PbrCassandra* elements with multiple LTRs, we randomly selected 15 elements, including five elements with two LTRs, six elements with three LTRs, two elements with four LTRs, and two elements with five LTRs (see Materials and Methods), and compared the insertion size of each element with that based on a PCR method. Except the two elements with five LTRs, which were not successfully amplified, the actual sizes of other 13 copies were consistent with estimates from the bioinformatics approach. To further verify that elements with multiple LTRs were not caused by an assembly issue, we randomly selected and resequenced one element with three LTRs (PbrCassandraI_T687) and one element harboring four LTRs (PbrCassandraI_T993). As expected, the resequencing data show identical structures of the elements with those predicted via sequence analysis method, suggesting that these elements with multiple LTRs are not caused by a wrong annotation or errors in the assembled sequence (supplementary figs. S1 and S2 and table S2, Supplementary Material online).

Identification and Annotation of the *Cassandra* Elements in Four Other Rosaceae Genomes

Using the same strategies above, we have also annotated *Cassandra* elements in four other Rosaceae genomes such as apple (*M. domestica*) (Velasco et al. 2010), peach (*Pr. persica*) (Verde et al. 2013), mei (*Pr. mume*) (Zhang et al. 2012), and woodland strawberry (*F. vesca*) (Shulaev et al. 2011). To distinguish these elements in different genomes, we have named them *MdCassandra*, *PpCassandra*, *PmCassandra*, and *FvCassandra*, respectively. In total, we have identified 2,041 *MdCassandra* copies, 667 *PpCassandra* copies, 388 *PmCassandra* copies and 132 *FvCassandra* copies, and the ratios of solo LTRs to intact elements are 1.12:1, 0.97:1, 1.67:1 and 1.02:1, respectively (table 1 and supplementary tables S2–S5, Supplementary Material online). These elements, together with numerous *Cassandra* remnants, make up 3.14, 0.39, 0.23 and 0.07 Mb of their host genomic DNA, accounting for 0.63%, 0.17%, 0.1%, and 0.03% of their assembled genomic sequences, respectively. Although *Cassandra* elements with multiple LTRs have also been detected in these four Rosaceae genomes, their copy numbers and frequency are much lower. For example, only 31 copies are present with three LTRs, four copies contain four LTRs, and no elements with five LTRs were detected in these genomes (table 2 and supplementary table S8, Supplementary Material online).

To check the transcriptional activities of the *Cassandra* elements, we performed BLAST searches against the EST (expressed sequence tag) database in NCBI using the elements as queries. We have detected >50 ESTs matching *MdCassandra*, *PpCassandra*, and *FvCassandra* elements, but no *PbrCassandras* or *PmCassandras* (*e* value $< 10^{-10}$) (supplementary table S9, Supplementary Material online). Specifically two ESTs (gi#84629412 and gi#84633895) were found to share high similarity with the *PpCassandras* containing three LTRs and span the two flanking sequences of the internal LTR,

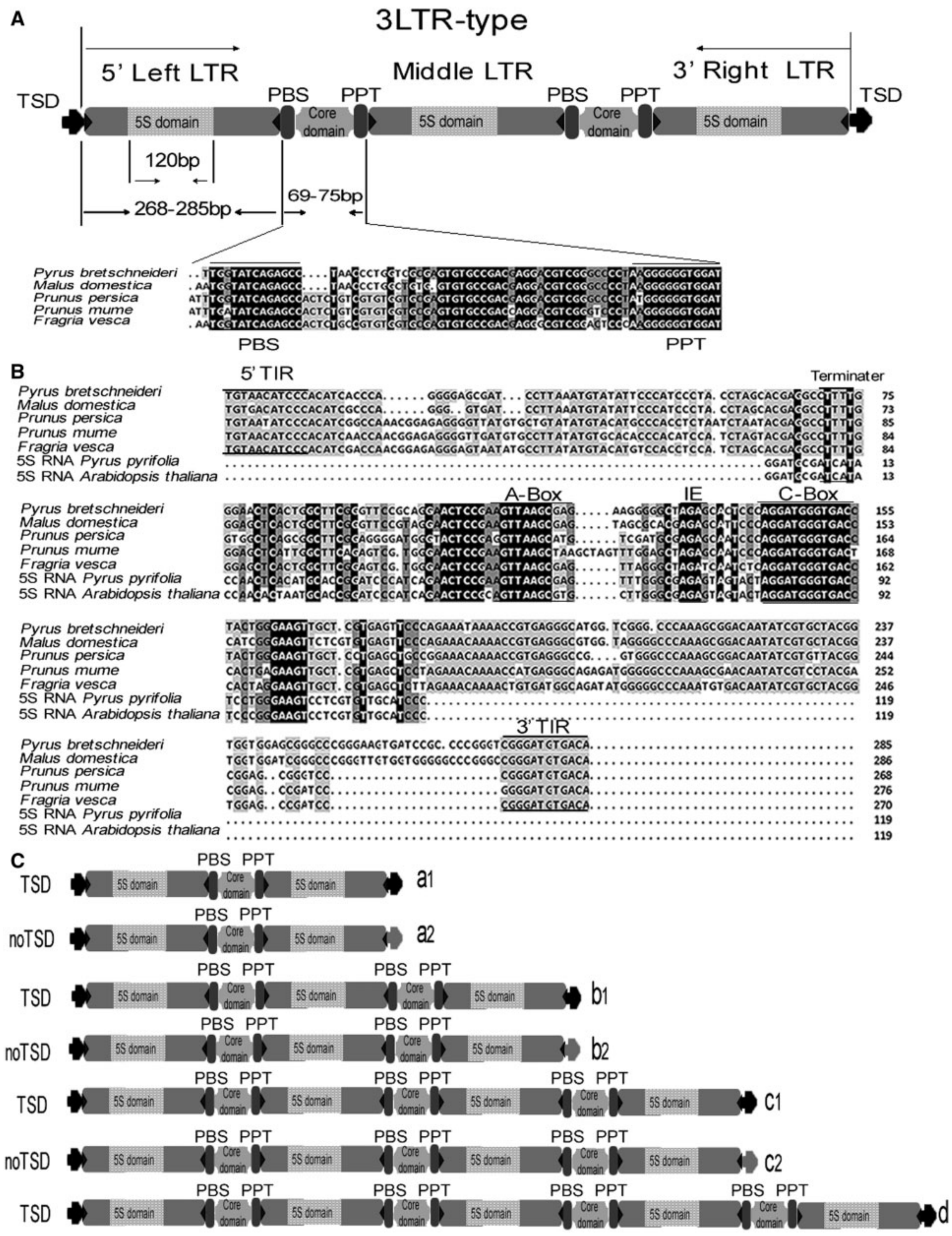

FIG. 1.—Schematic of *Cassandra* element and various structure patterns of intact elements. (*A*) Structure annotation of intact element with three LTRs flanking with TSDs. The 5 S RNA domain regions of LTR are shown in light gray boxes, the terminal repeats (TIRs) are shown as dark blue triangles, "TSD" indicates the 5-bp target site duplication; "Core domain" means the intra-sequence between PBS and PPT with no protein coding functions. (*B*) The

(continued)

indicating that the two copies with three LTRs may be active (supplementary table S8, Supplementary Material online).

Distribution, Target Site Specificity, and Gene Disruption of *Cassandra* Elements

One of the major properties of plant LTR-RTs is the presence of biased insertion into pericentromeric regions (Presting et al. 1998; Jiang, Jordan, et al. 2002; Du, Tian, Hans, et al. 2010; Tian et al. 2012). For example, approximately 87% of the intact elements and solo LTRs were found in the recombination-suppressed pericentromeric regions, which only cover 54% of genomic DNA in soybean (Du et al. 2012). A recently investigated *Copia*-like retrotransposon, *TARE1*, also exhibits enrichment close to centromeres (Yin et al. 2013). In contrast, *SMART*, a highly conserved small LTR-RT in grass species, shows a preferential insertion into or close to genes, a

Table 1
Copy Numbers of Different Types of *Cassandra* Elements Identified in Five Rosaceae Species

	PbrCassandra			***MdCassandra***	***PpCassandra***			***PmCassandra***			***FvCassandra***
Structure	**Unanchored**	**Anchored**	**Total**	**Anchored**	**Unanchored**	**Anchored**	**Total**	**Unanchored**	**Anchored**	**Total**	**Anchored**
Intact elements with TSDs	259	916	1,175	188	7	187	194	10	59	69	36
Intact elements without TSDs	42	156	198	90	2	45	47	4	19	23	11
Solo LTR with TSDs	149	639	788	164	9	150	159	11	74	85	28
Solo LTR without TSDs	37	213	250	47	1	30	31	6	24	30	9
Truncated elements with left side deleted	190	631	821	331	4	89	93	15	62	77	23
Truncated elements with right side deleted	167	645	812	884	2	114	116	12	57	69	23
Truncated elements with both sides deleted	248	740	988	337	5	22	27	6	29	35	2
Total	1,092	3,940	5,032	2,041	30	637	667	64	324	388	132

Note.—Unanchored represents those elements have not been assembled on the chromosomes. Anchored means those elements have been assembled on the chromosomes.

Table 2
Copy Numbers of Intact *Cassandra* Elements with Multiple LTRs

	PbrCassandra		***MdCassandra***		***PpCassandra***		***PmCassandra***		***FvCassandra***	
Structure	**No. of Element**	**%**	**No. of Element**	**%**	**No. of Element**	**%**	**No. of Element**	**%**	**No. of Element**	**%**
2L-type flanking with TSDs	966	70.4	180	64.8	184	76.4	66	71.7	35	74.5
2L-type flanking without TSDs	155	11.3	85	30.6	47	19.5	20	21.7	11	23.4
3L-type flanking with TSDs	182	13.2	7	2.5	8	3.3	3	3.3	1	2.1
3L-type flanking without TSDs	40	2.9	4	1.4	0	0	3	3.3	0	0
4L-type flanking with TSDs	22	1.6	2	0.7	2	0.8	0	0	0	0
4L-type flanking without TSDs	3	0.2	0	0	0	0	0	0	0	0
5L-type flanking with TSDs	5	0.4	0	0	0	0	0	0	0	0
Total	1,373	100	278	100	241	100	92	100	47	100

Fig. 1.—Continued
sequence alignment of LTRs and cellular 5 S rRNAs. According to the proportion of each position, the nucleotides are shaded different colors in the alignment: White on black, ≥90%; black on dark gray, ≥70%; black on gray, ≥50%; black on white, <50%. The TIRs and motifs identified important to transcription are labeled: A-Box; IE, intermediate element; C-Box; terminator, the predicted poll III terminator. The LTR sequences of five youngest intact elements from Rosaceae species are *PbrCassandraI_T278, MdCassandraI_T113, PpCassandraI_T66, PmCassandraI_T47,* and *FvCassandraI_T1* from top to bottom. The last two cellular 5 S rRNA sequences belong to *Pyrus pyrifolia* and *Arabidopsis* downloaded from NCBI website, from top to bottom, the accession numbers are AB621370.1 and AJ307356.1, respectively. (*C*) Various structure patterns of intact elements. Six different structure patterns of intact elements are shown: "a1" indicates intact elements with two LTRs flanking with TSDs and "a2" indicates intact elements with two LTRs flanking without TSDs; "b1" means intact elements with three LTRs flanking with TSDs and "b2" means intact elements with three LTRs flanking without TSDs; "c1" represents intact elements with four LTRs flanking with TSDs and "c2" means intact elements with four LTRs flanking without TSDs; "d" represents intact elements with five LTRs flanking with TSDs.

characteristic belonging to MITEs (Gao et al. 2012). To understand the distribution pattern of *PbrCassandra* elements in the pear genome, genomic DNA in each chromosome was split using 1-Mb DNA as a window. Thus, the number of the observed *PbrCassandra* elements in each window could be calculated. Overall, we have assigned 3,940 *PbrCassandra* copies into 378 nonoverlapped 1-Mb windows with an average approximately 10 copies in each window (figs. 2 and 6 and supplementary table S3, Supplementary Material online). To check whether these elements have insertion preferences or are randomly distributed in the pear genome, we performed a randomization test by using an in-house perl script (see Materials and Methods). Our data show that the copy number of the observed *PbrCassandra* elements in 345 windows (91%) has no statistical difference with those from a computational simulation, indicating that most *PbrCassandra* elements may insert into the genome randomly rather than having a bias. Furthermore, the pattern of 187 copies harboring three or more LTRs, with or without TSDs, was also shown to be randomly distributed in the genome (fig. 6). In addition, similar analyses for the *MdCassandra* and *PpCassandra* elements were performed on the apple and peach genome, respectively (fig. 6 and supplementary figs. S3 and S4, Supplementary Material online). The distribution patterns for *Cassandra* elements in apple and peach are in agreement with that in pear, indicating that most *Cassandra* elements may be distributed throughout the Rosaceae genomes.

Analysis of the target selection of retrotransposons is important for understanding the structure and evolution of plant genomes (Miyao et al. 2003). In order to examine the target site preference of *Cassandra* elements, the GC content of *Cassandra* insertion sites was analyzed, including the 5-bp TSD sequence and two 20-bp flanking sequences of 1,175 *PbrCassandra*, 188 *MdCassandra,* 194 *PpCassandra,* 69 *PmCassandra,* and 36 *FvCassandra* intact copies with TSDs (fig. 3*A*). The data show that the four base positions at −4, −2, 2, and 4 have higher GC preference ($P < 6.0 \times 10^{-3}$) (fig. 3*A*). In contrast, the five base positions at −5, −3, T3, 3, and 5 have lower GC preference ($P < 5.0 \times 10^{-2}$) (fig. 3*A*). To further understand the target site specificity of *Cassandra* elements, the exact nucleotide compositions at 45 sites surrounding 1,175 *PbrCassandra* copies with TSDs were calculated (fig. 3*B*). As shown in figure 3*B*, higher frequency of A (positions −5 and −3) and higher frequency of T (positions 3 and 5) were observed, indicating the presence of target site preference at these sites. The consensus sequence between position −5 and position 5 is "A(G/C)A(G/C)N-$\underline{\text{NN(A/T)NN}}$-N(C/G)T(C/G)T" (the nucleotides in the middle with underline represent TSD sequences).

In order to evaluate the impact of *Cassandra* elements on their host genes, we performed an association analysis between the chromosome IDs of 3,940 *PbrCassandra* copies and 42,369 annotated genes in pear. The data show that a total of 352 *PbrCassandra* copies (9%) have either integrated into genes (124 copies) or the flanking sequences close to genes (228 copies within 1 kb), indicating their potential influence on the structure and expression of genes (table 3 and supplementary table S10, Supplementary Material online). It is not surprising to see that the majority of 124 copies within genes has been inserted into introns (121 copies), three copies were detected in the UTR regions, and none of these copies is located in protein coding regions, partially because retrotransposon insertions in coding sequences are harmful to the genes. In addition, 106 (out of 145) solo LTRs (73%) were found inserted within the flanking sequences close to genes, much higher than the rate of intact elements (57%) and truncated elements (60%). The rate of solo LTRs to intact elements, which integrated into genes or 1-kb flanking sequences close to genes, is 2.23:1, much higher than that of the genome-wide level (~0.88:1). These results indicate that unequal recombination events occur more frequently within the flanking regions close to genes. Notably, this proportion of *PbrCassandra* elements associated with genes (9%) is much less than that of *SMART* LTR-RT element (53%) (Gao et al. 2012) and *Tos17* elements (26%) (Miyao et al. 2003) in the rice genome. These results further indicate that *Cassandra* elements have weak target site specificity and most of them are distributed randomly throughout the genome.

Dating of *Cassandra* Elements in Five Rosaceae Species

In order to compare the abundance, activity, and amplification timeframe of *Cassandra* elements among different Rosaceae species, the intact elements with TSDs have been dated using the approach previously reported (SanMiguel et al. 1998). This approach is based on the fact that the two LTR sequences of an element are identical at the time of insertion, but both LTRs accumulate nucleotide substitutions over evolutionary time. When an evolutionary rate is applied to LTR-RT elements, the level of nucleotide difference between the two LTRs can be roughly converted into time since a new copy inserted into a genome. Although the rate of LTR-RTs varies among different loci, families, and lineages (Zhao et al. 2013), an estimation of 1.3×10^{-8} per site per year has been applied in many studies (Ma et al. 2004; Du, Tian, Hans, et al. 2010; Yin et al. 2013). Using this rate, we have estimated the insertion time of 1,175 *PbrCassandra*, 188 *MdCassandra,* 194 *PpCassandra,* 69 *PmCassandra,* and 36 *FvCassandra* intact copies with TSDs. In the 1,175 *PbrCassandra* copies, 906 (77.1%) inserted into the genome 1.0–4.5 Ma, and only 127 copies (10.8%) integrated into the genome <1.0 Ma with the youngest one 0.07 Ma. In addition, 142 copies (12.1%) have been dated >4.5 Ma, indicating that they are evolutionarily old in the pear genome (fig. 4 and supplementary table S3, Supplementary Material online). For 188 *EMdCassandra* elements, the spectrum of activities is quite similar to those in pear. A total of 147 copies (78.2%) inserted into the genome 1.5–5.0 Ma, but the peak is much lower. The 194 *PpCassandra* elements

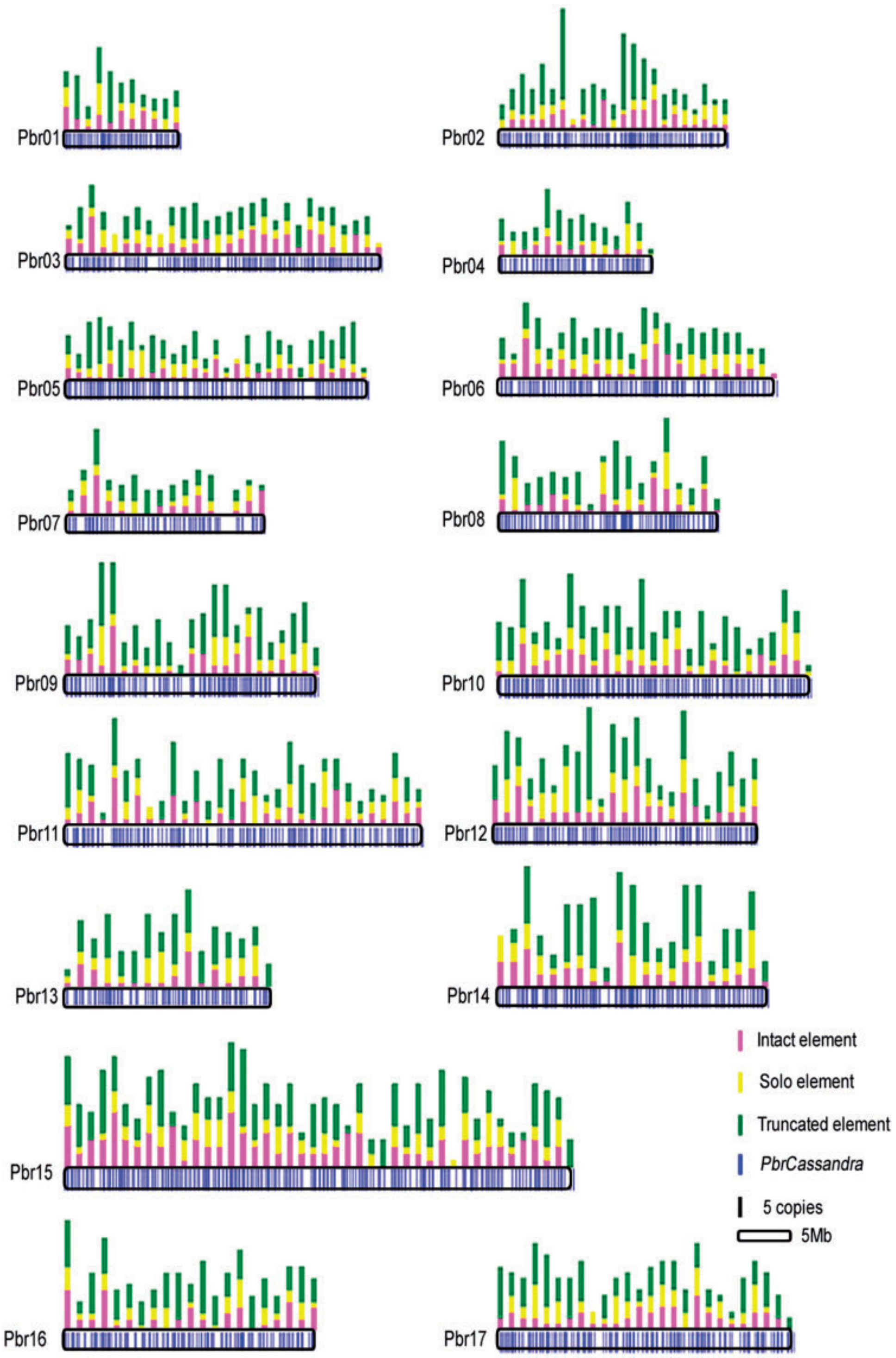

Fig. 2.—Distribution of *PbrCassandra* elements through 17 pear chromosomes. Pear chromosomes and *PbrCassandra* element insertions are represented by black horizontal boxes with blue vertical lines. Histograms over the horizontal boxes indicate the copy number of *PbrCassandra* elements per Mb.

appear to have no obvious peak of activity, and have maintained low capability for proliferation within 0–8 Ma. Different from the activities of *PbrCassandra*, *MdCassandra*, and *PpCassandra* elements, all the 69 *PmCassandra* elements transposed >2.5 Ma, whereas all the 36 *FvCassandra* elements integrated into the genome <3 Ma (fig. 4 and supplementary tables S4–S7, Supplementary Material online). In total, we have only found five copies with two identical

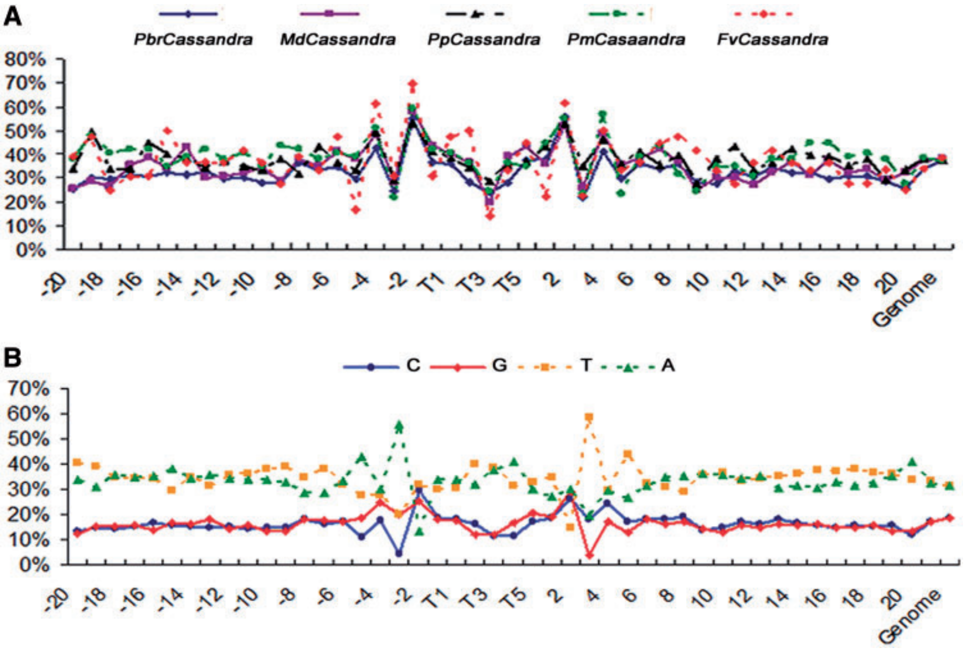

FIG. 3.—Base preferences of *Cassandra* element insertion sites. (*A*) GC contents of *Cassandra* elements insertion sites. Positions from T1 to T5 represent the TSD sequence; numbers from −20 to −1 and 1 to 20 indicate flanking sequence base numbers downstream and upstream from TSD, respectively. The second to last position and the last position represent the average base content for 1,175 *PbrCassandra* intact elements and the whole pear genome sequences, respectively. (*B*) Base preferences of *PbrCassandra* elements insertion sites. Positions on *x* axis represent the same as (*A*).

Table 3
Insertion Sites of *PbrCassandra* Elements in Pear Genome

Location	Intact (No.)	Solo (No.)	Truncated (No.)	Total (No.)
Gene	28	39	57	124
Intron	28	37	56	121
Exon	N	N	N	N
5′-UTR	N	2	1	3
3′-UTR	N	N	N	N
Within 1-kb flanking of gene	37 (11u + 26d)	106 (57u + 49d)	85 (41u + 44d)	228
Total	65	145	142	352
Percent	5.96	17.37	7.09	8.98

NOTE.—N means not present; u and d represent upstream and downstream, respectively.

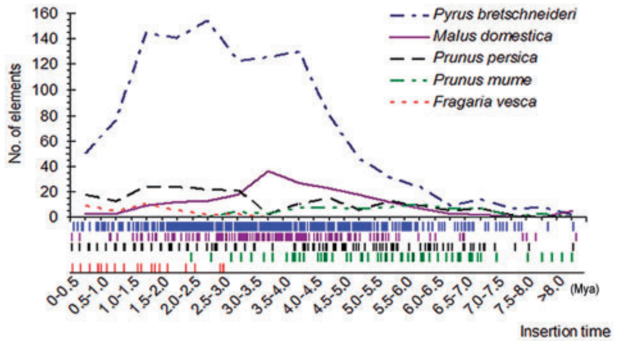

FIG. 4.—Insertion times of *Cassandra* intact elements. The insertion times of 1,175 *PbrCassandra,* 188 *MdCassandra,* 194 *PpCassandra,* 69 *PmCassandra,* and 36 *FvCassandra* intact elements with TSDs were analyzed. Vertical lines under the line graph represent insertion events.

LTRs, including 2 *MdCassandra* elements, 2 *PpCassandra*, and 1 *FvCassandra* elements (fig. 4 and supplementary tables S4–S7, Supplementary Material online). Consistent with this result, we have identified 41 EST sequences in apple, 11 EST sequences in peach, and 3 EST sequences in strawberry matching *Cassandra* TRIMs, with no EST sequences detected in pear and mei (supplementary table S9, Supplementary Material online). The multispecific comparisons of *Cassandra* elements in different Rosaceae species indicate that 1) the activities of *Cassandra* TRIMs vary greatly depending on different genomes, and *PbrCassandra* elements have been most successfully amplified in pear; 2) the proliferation peak and amplification timeframe of *Cassandra* TRIMs are different in different genomes; and 3) genome size variation is not a determination for the copy number variation of *Cassandra* TRIMs.

Evolutionary Relationships among *Cassandra* Elements, Lineages, and Species

To understand the evolutionary relationships of *Cassandra* elements among different species, we have constructed a

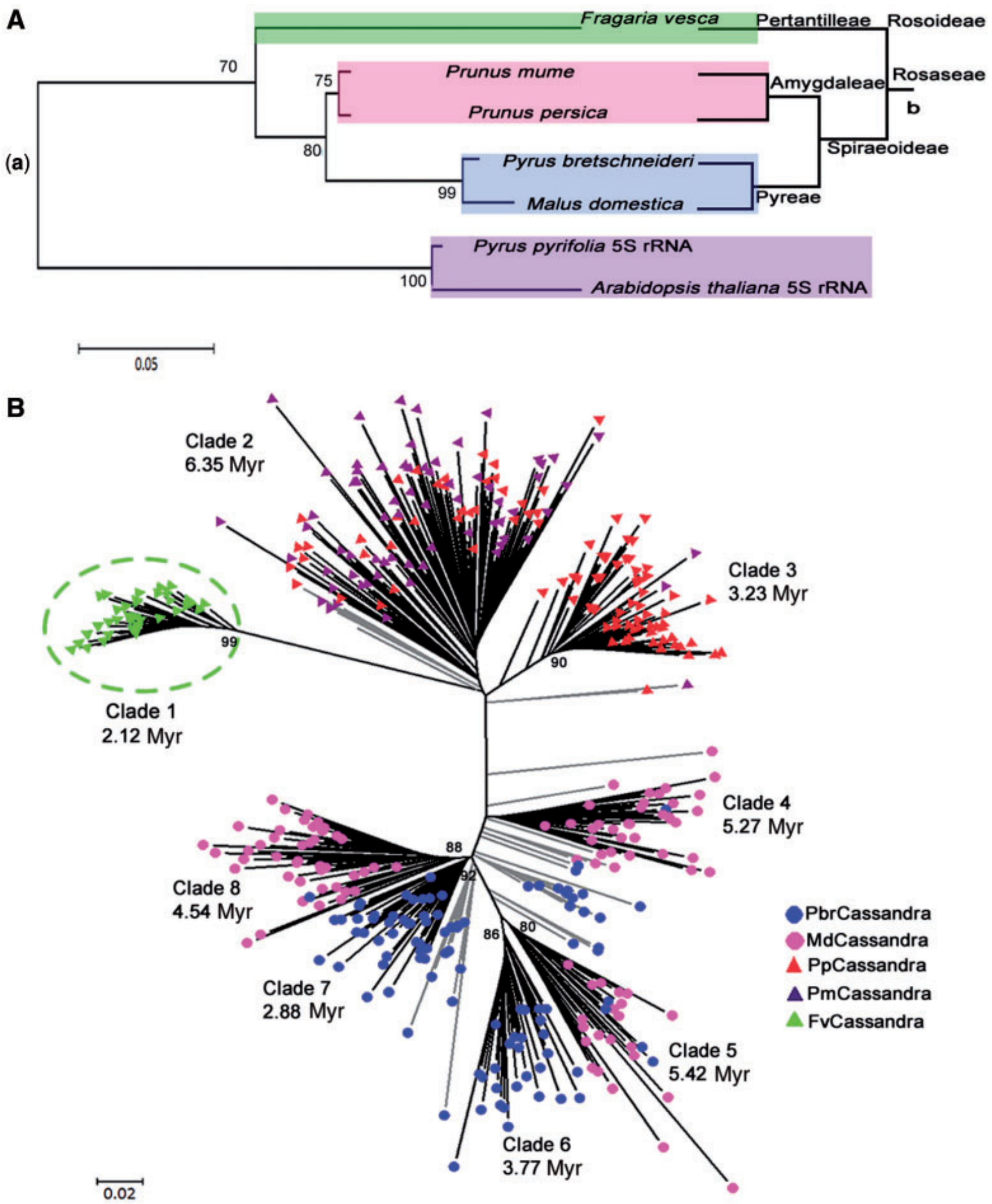

FIG. 5.—Phylogenetic relationships among five Rosaceae species. (*A-a*) Neighbor-joining tree of five *Cassandra* 5 S RNA domains and two cellular 5 S rRNA domains. The five *Cassandra* 5 S RNA domains are extracted from the 5′-LTR consensus sequence of *PbrCassandra, MdCassandra, PpCassandra, PmCassandra,* and *FvCassandra*, respectively. The two cellular 5 S rRNA sequences are the same as figure 1*A*. The level of nucleotide sequence distance is indicated by the scales. (*A-b*) Taxonomy tree of five Rosaceae species. The taxonomy tree was built using the common tree tool on the NCBI website (http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi, last accessed June 3, 2014). (*B*) Neighbor-joining tree of five Rosaceae species. 5′-LTR sequences of 405 *Cassandra* elements are extracted from 36 *FvCassandra* elements (green triangles), 69 *PmCassandra* elements (red triangles), 100 random samples of *PbrCassandra* (blue circles), 100 random samples of *MdCassandra* (pink circles), and 100 random samples of *PpCassandra* (purple triangles) elements. All the *Cassandra* elements used are intact with TSDs. The level of nucleotide sequence distance is indicated by the scales.

phylogenetic tree using the conserved 120-bp 5 S RNA residing in the two LTRs (fig. 5*A*). The data show that two Pyreae species, *Py. bretschneideri* ($n = 17$) and *M. domestica* ($n = 17$), are close to each other, and two Amygdaleae species, *Pr. persica* ($n = 8$) and *Pr. mume* ($n = 8$), are clustered together. In contrast, the Potentilleae species, *F. vesca* ($n = 7$), formed a distinct clade (fig. 5*A-a*), and the phylogeny tree reflects well the species' phylogeny (fig. 5*A-b*). The 5 S RNA sequences carried by the *Cassandra* elements are basically distinguished from the cellular 5 S RNA in *P. pyrifolia* and *Arabidopsis*, further indicating that *Cassandra* TRIMs evolved independently (fig. 5*A*) (Kalendar et al. 2008).

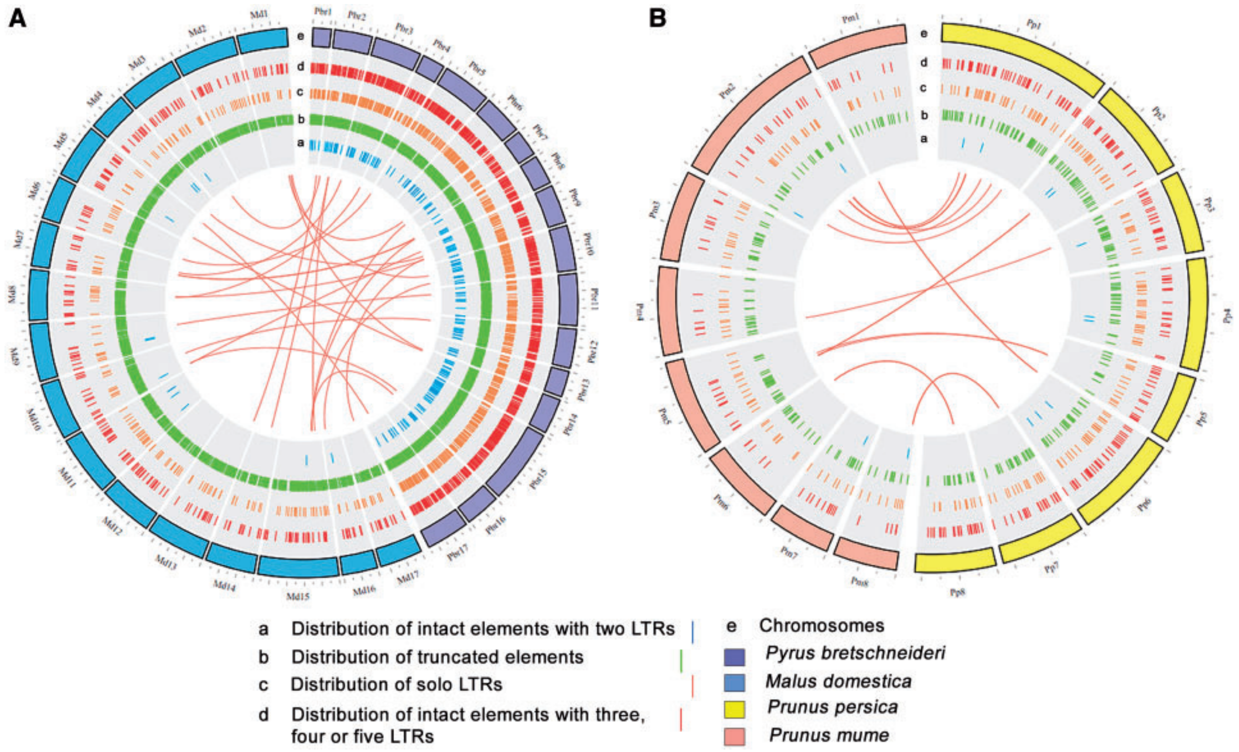

Fig. 6.—Insertions of orthologous copies between *Cassandra* elements. (*A*) Distribution of *PbrCassandra*, *MdCassandra* elements, and their orthologous copies. (*B*) Distribution of *PpCassandra*, *PmCassandra* elements, and their orthologous copies. Orthologous copies are connected by pink red lines. Vertical lines of blue (a), green (b), orange (c) and red (d) indicate intact elements with multiple LTRs, truncated elements, solo LTRs and typical intact elements, respectively. The purple, blue, yellow and orange blocks (e) represent chromosomes of *Pyrus bretschneideri*, *Malus domestica*, *Prunus persica* and *Pr. mume*, respectively. Circos (Krzywinski et al. 2009) (http://circos.ca, last accessed June 3, 2014) was employed for constructing this diagram.

The phylogenetic relationships among the five Rosaceae species are further reflected by the tree generated using 5′-LTR sequences of *Cassandra* elements. As shown in figure 5*B*, this tree can be grouped into eight clades, including one *FvCassandra* clade (clade 1), one *PmCassandra* and *PpCassandra* clade (clade 2), one *PpCassandra* clade (clade 3), and five clades from *PbrCassandra* and *MdCassandra* elements (clade4, clade5, clade6, clade7, and clade8) (fig. 5*B*). These clades can be roughly dated to be 2.12, 6.35, 3.23, 5.27, 5.42, 3.77, 2.88, and 4.54 Myr, respectively (fig. 5*B*), using the approach previously described (Jiang, Jordan, et al. 2002; Yin et al. 2013). Although *PbrCassandra* and *MdCassandra* elements can overall be well separated from each other in each of the three clades (clade 4, clade 5, and clade 6), many *PpCassandra* elements are mixed with *PmCassandra* elements in clade 2, indicating that these two species may have experienced some introgression in early stages of their evolution.

Orthologous Insertions between *M. demestica* and *Py. bretschneideri*, as well as between *Pr. persica* and *Pr. mume*

It has been documented that the origin of LTR-RT elements can be tracked to before the divergence of monocot and eudicot plants (Du, Tian, Hans, et al. 2010), but most recognizable intact elements were inserted in the host genome <5 Ma. This is mainly because many old elements have undergone one or more rounds of recombination, or have been completely deleted from the genome (Ma and Bennetzen 2004). To test if any orthologous *Cassandra* insertion is still present between different species investigated in this study, we have examined the flanking sequences for each element using the method previously described (Yin et al. 2013) (also see Materials and Methods). As expected, we have identified only 26 orthologous LTR pairs between *PbrCassandra* and *MdCassandra* elements (fig. 6*A* and supplementary table S11, Supplementary Material online), and 22 orthologous insertions between *PpCassandra* and *PmCassandra* (fig. 6*A* and supplementary table S12, Supplementary Material online). As shown in figure 6, all 26 copies between *Py. bretschneideri* and *M. domestica* are scattered throughout the genome, indicating that large-scale DNA rearrangements may have occurred after the split of the two species. In contrast, between chromosome 1 in *Pr. prunus* and chromosome 2 in *Pr. mume* six pairs maintain good colinearity, suggesting that this may be the orthologous region between the two genomes.

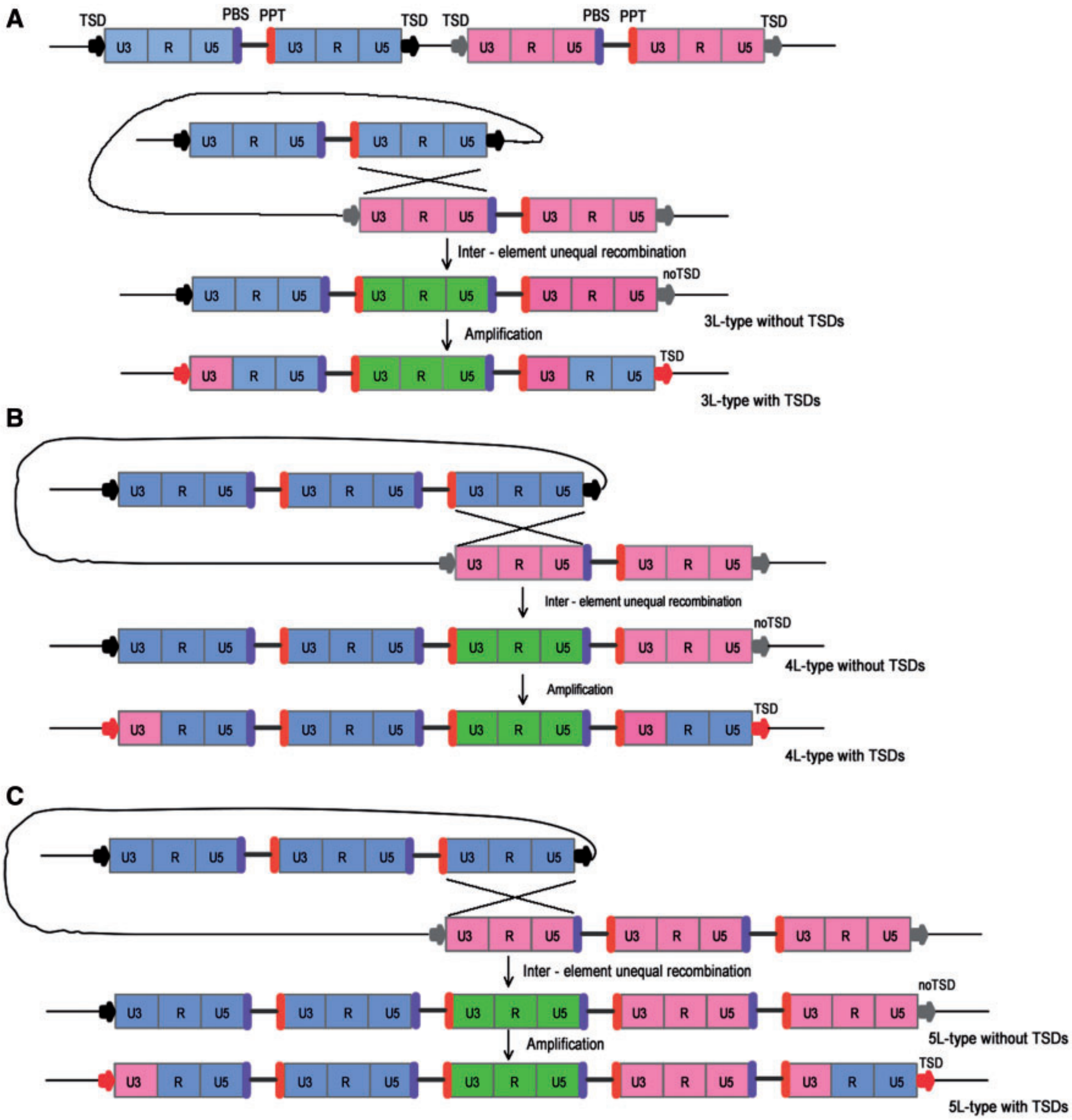

FIG. 7.—Models for the evolution and amplification of *Cassandra* elements with multiple LTRs in five Rosaceae species. Models for evolution and amplification of *Cassandra* elements with three LTRs flanking with TSDs (*A*), *Cassandra* elements with four LTRs flanking with TSDs (*B*), and *Cassandra* elements with five LTRs flanking with TSDs (*C*).

Discussion

It has been well documented that LTR-RTs are ubiquitous in plants (Kumar and Bennetzen 1999). However, most LTR-RT elements investigated so far contain two LTRs (Kumar and Bennetzen 1999; Zhao and Ma 2013). The intact elements with multiple LTRs have not been reported often (Devos et al. 2002; Sabot and Schulman 2007; Tian et al. 2009), indicating that this type of element may be present in plants at very low frequency. Indeed, by searching the *Arabidopsis* genome, only one element was found to carry the third LTR flanked by both PBS and PPT (Devos et al. 2002). Even in some grass species with a large genome size such as *japonica* rice, only a total of five elements have been defined as "complex" with a "LTR–internal–LTR–internal–LTR" structure (Tian et al. 2009). In this study, in order to annotate the *Cassandra* elements with multiple LTRs, not only the LTR sequences but also the intact elements sequences of all 27 *PbrCassandra* elements initially identified were used as queries to scan the whole-genome sequences of pear by using the cross_match program with default parameters. We have identified >8,000 *Cassandra* TRIMs in five Rosaceae species, and found that 282 copies contain at least three LTRs. In particular, we have found 252 intact *PbrCassandras* (~20%) containing three, four, or five LTRs in the pear genome, indicating that LTR-RTs with multiple LTRs may not be rare, but are present very frequently in particular plant genomes.

Regarding the mechanisms generating elements with three LTRs, a possible hypothesis may be as follows: 1) two adjacent normal elements are close together, and experience a DNA recombination event between 5′- and 3′-LTRs; 2) a recombinant element with three LTRs but no TSDs is formed (Devos et al. 2002; Vincent et al. 2005); and 3) many elements with three LTRs and TSDs have been generated through RNA-mediated transposition process (fig. 7*A*). From this, the structure of elements with four or five LTRs can also be explained (fig. 7*B* and *C*). Because the elements with four or five LTRs have experienced two or more recombination events, it is not surprising to see the element number decrease dramatically with more LTRs in an element (table 2).

It should be pointed out that template switching between two molecular RNAs can also create a "complex" with three LTRs, which has been proposed as another potential hypothesis in Triticeae genomes (Sabot and Schulman 2007), but the recombinants generated through this way all contain TSDs. In the pear genome, elements with or without TSDs are both very common (table 2). The phylogenetic relationship among the *PbrCassandra* copies harboring three LTRs also indicates that many interelement recombination events, but not only one single recombination, might have occurred independently (supplementary fig. S5, Supplementary Material online), and may play a critical role in the formation of *PbrCassandras*, particularly at the early stage when the ancestor copies were generated.

It is not clear why *Cassandra* elements have been so successfully amplified in the pear genome. It is equally interesting that 252 *PbrCassandra*s contain multiple LTRs but only 13 *MdCassandra*s were found to have such novel structures. Like the apple genome (assembled genome size 603.9 Mb), the pear genome contains 17 chromosomes (assembled genome size 512 Mb) (supplementary table S1, Supplementary Material online). The *PbrCassandras* and *MdCassandras* share similar structural variation, target site specificity, and insertion time pattern. The pear and apple diverged from their common ancestor >5 Ma (Wu et al. 2012), suggesting that most intact elements were amplified after speciation. Indeed, we have detected a few insertions of orthologous *Cassandra* copies between pear and apple (fig. 6). We thus speculate that this lineage-specific amplification may be caused by the differential regulation of epigenetic silencing on LTR-RTs between pear and apple, and that more copies of *Cassandra* in pear increase the possibility of unequal recombination between two adjacent elements and facilitate the formation of elements with multiple LTRs. Further examination and comparison of complete genomic components and genomic property between the two genomes will help uncover the genetic and epigenetic basis underlying the unique features of *PbrCassandra*s in the pear genome.

Supplementary Material

Supplementary tables S1–S12 and figures S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank many genome data contributors, the high-quality genome sequences generated by them were instrumental in conducting this study. This study was supported by the National High-tech R&D Program (863) of China (2011AA10020602, 2013AA102606-02), the National Natural Science Foundation of China (31230063, 31370266), and Jiangsu Agricultural Science and Technology Innovation Funds (CX(12)5036). The authors have declared that no competing interests exist. H.Y. and S.Z. conceived and designed the experiments; H.Y., C.J., and L.F. performed the experiments; H.Y., J.D., L.L., C.J., L.F., M.L., and J.W. analyzed the data; H.Y., J.D., and L.L. contributed reagents/materials/analysis tools; H.Y., J.D., and S.Z. wrote the paper.

Literature Cited

Antonius-Klemola K, Kalendar R, Schulman AH. 2006. TRIM retrotransposons occur in apple and are polymorphic between varieties but not sports. Theor Appl Genet. 112:999–1008.

Costas J, Naveira H. 2000. Evolutionary history of the human endogenous retrovirus family *ERV9*. Mol Biol Evol. 17:320–330.

Devos KM, Brown JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res. 12:1075–1079.

Du J, et al. 2012. Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. Plant Cell 24:21–32.

Du J, Grant D, et al. 2010. SoyTEdb: a comprehensive database of transposable elements in the soybean genome. BMC Genomics 11:113.

Du J, Tian Z, Bowen NJ, et al. 2010. Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR Swapping in soybean. Plant Cell 22:48–61.

Du J, Tian Z, Hans CS, et al. 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J. 63: 584–598.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Gao D, Chen J, Chen M, Meyers BC, Jackson S. 2012. A highly conserved, small LTR retrotransposon that preferentially targets genes in grass genomes. PLoS One 7:e32010.

Jiang N, Bao Z, et al. 2002. Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. Genetics 161:1293–1305.

Jiang N, Gao D, Xiao H, van der Knaap E. 2009. Genome organization of the tomato sun locus and characterization of the unusual retrotransposon *Rider*. Plant J. 60:181–193.

Jiang N, Jordan IK, Wessler SR. 2002. *Dasheng* and *RIRE2*. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. Plant Physiol. 130:1697–1705.

Kalendar R, et al. 2008. Cassandra retrotransposons carry independently transcribed 5S RNA. Proc Natl Acad Sci U S A. 105:5833–5838.

Kalendar R, et al. 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166:1437–1450.

Kalkman C. 2004. Rosaceae. In: Kubitzki K, editor. Flowering plants—Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales. Berlin (Germany): Springer. p. 343–386.

Kapitonov V, Jurka J. 1996. The age of Alu subfamilies. J Mol Evol. 42: 59–65.

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

Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

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

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci U S A. 101:12404–12410.

Ma J, Bennetzen JL. 2006. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. Proc Natl Acad Sci U S A. 103:383–388.

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

McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19: 362–367.

Miyao A, 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.

Paterson AH, et al. 2012. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492: 423–427.

Potter D, et al. 2007. Phylogeny and classification of Rosaceae. Plant Syst Evol. 266:5–43.

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

Sabot F, Schulman AH. 2007. Template switching can create complex LTR retrotransposon insertions in Triticeae genomes. BMC Genomics 8:247.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nat Genet. 20: 43–45.

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

Shulaev V, et al. 2011. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet. 43:109–116.

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

Tanskanen JA, Sabot F, Vicient C, Schulman AH. 2007. Life without GAG: the BARE-2 retrotransposon as a parasite's parasite. Gene 390: 166–174.

Tian Z, et al. 2009. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res. 19:2221–2230.

Tian Z, et al. 2012. Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. Plant Cell 24:4422–4436.

Velasco R, et al. 2010. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet. 42:833–839.

Verde I, et al. 2013. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 45:487–494.

Vincent C, et al. 2005. Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. J Mol Evol. 61:275–291.

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

Witte CP, Le QH, Bureau T, Kumar A. 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci U S A. 98:13778–13783.

Wu J, et al. 2012. The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res. 23:396–408.

Yin H, et al. 2013. *TARE1*, a mutated *Copia*-like LTR retrotransposon followed by recent massive amplification in tomato. PLoS One 8: e68587.

Zhang Q, et al. 2012. The genome of *Prunus* mume. Nat Commun. 3: 1318.

Zhao M, et al. 2013. Shifts in the evolutionary rate and intensity of purifying selection between two *Brassica* genomes revealed by analyses of orthologous transposons and relics of a whole genome triplication. Plant J. 76:211–222.

Zhao MX, Ma JX. 2013. Co-evolution of plant LTR-retrotransposons and their host genomes. Protein Cell 4:493–501.

Associate editor: Esther Betran