

RESEARCH

Open Access



# *TGTT* and *AACA*: two transcriptionally active LTR retrotransposon subfamilies with a specific LTR structure and horizontal transfer in four Rosaceae species

Hao Yin<sup>1,2</sup>, Xiao Wu<sup>1,2</sup>, Dongqing Shi<sup>1,2</sup>, Yangyang Chen<sup>1,2</sup>, Kaijie Qi<sup>1,2</sup>, Zhengqiang Ma<sup>2,3</sup> and Shaoling Zhang<sup>1,2\*</sup>

## Abstract

**Background:** Long terminal repeat retrotransposons (LTR-RTs) are major components of plant genomes. Common LTR-RTs contain the palindromic dinucleotide 5'-TG'-CA'-3' motif at the ends. Thus, further analyses of non-canonical LTR-RTs with non-palindromic motifs will enhance our understanding of their structures and evolutionary history.

**Results:** Here, we report two new LTR-RT subfamilies (*TGTT* and *AACA*) with atypical dinucleotide ends of 5'-TG'-TT'-3', and 5'-AA'-CA'-3' in pear, apple, peach and mei. In total, 91 intact LTR-RTs were identified and classified into four *TGTT* and four *AACA* families. A structural annotation analysis showed that the four *TGTT* families, together with *AACA1* and *AACA2*, belong to the *Copia*-like superfamily, whereas *AACA3* and *AACA4* appeared to be TRIM elements. The average amplification time frames for the eight families ranged from 0.05 to 2.32 million years. Phylogenetics coupled with sequence analyses revealed that the *TGTT1* elements of peach were horizontally transferred from apple. In addition, 32 elements from two *TGTT* and three *AACA* families had detectable transcriptional activation, and a qRT-PCR analysis indicated that their expression levels varied dramatically in different species, organs and stress treatments.

**Conclusions:** Two novel LTR-RT subfamilies that terminated with non-palindromic dinucleotides at the ends of their LTRs were identified in four Rosaceae species, and a deep analysis showed their recent activity, horizontal transfer and varied transcriptional levels in different species, organs and stress treatments. This work enhances our understanding of the structural variation and evolutionary history of LTR-RTs in plants and also provides a valuable resource for future investigations of LTR-RTs having specific structures in other species.

**Keywords:** LTR retrotransposon, Horizontal transfer, Transcription activity, Pear, Rosaceae

## Background

Long terminal repeat retrotransposons (LTR-RTs) are major components that are widespread in flower plant genomes [1]. They are capable of propagating to reach thousands of copies in a genome using RNA as an intermediate [2, 3]. LTR-RTs are the most significant contributor to genome size, representing 43% of the nuclear

DNA in pear [4], 38% in apple [5], 19% in peach [6], 53% in cotton [7] and over 70% in maize genomes [8]. A representative intact LTR-RT usually contains two highly identical LTRs, which are typically flanked by 2-bp palindromic motifs, commonly 5'-TG-CA-3'. The internal region of an autonomous LTR should contain a primer-binding site (PBS), a polypurine tract (PPT) and two functional genes (*gag*, and *pol*) [9]. Based on the order of Reverse transcriptase (*rt*) and Integrase (*int*) in *pol*, LTR-RTs can be further classed into *Gypsy* and *Copia* superfamilies [9]. In addition, the LTR-RTs also contain two types of non-autonomous groups, large retrotransposon derivatives (LARDs) [10] and terminal-repeat retrotransposons in

\* Correspondence: slzhang@njau.edu.cn

<sup>1</sup>Center of Pear Engineering Technology Research, College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

<sup>2</sup>State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China

Full list of author information is available at the end of the article



miniature (TRIMs) [11]. The insertion of an LTR-RT is accompanied by the duplication of a 4–6-bp sequence immediately flanking with the 5′ and 3′ ends of the element, called target site duplication (TSD).

The most common dinucleotide motif flanking the direct LTR-RT repeat regions is the palindromic 5′-TG-CA-3′ motif. However, several LTR-RT families with non-TGCA motifs have been reported. For example, *Tos17*, a rice LTR-RT that can be activated by tissue culture, has a non-canonical motif of 5′-TG-GA-3′ [12] and *TARE1*, which was identified as intensively amplified in the tomato genome, ends with 5′-TA-CA-3′ motifs [13]. In addition, *AcCOPIA1* that terminated with 5′-‘TG’-‘TA’-3′ at both ends of the LTRs was identified in onion [14]. However, no such non-canonical elements have been identified in the Rosaceae species.

Horizontal transfers (HTs) indicate the transmission of genetic material among sexually isolated species. As a possible dissemination mechanism of transposable elements (TEs) in eukaryotes, the horizontal transfer of TEs (HTTs) into a new organism is an important step for the TE to escape from the silencing machinery of their host genome and obtain a new ‘life cycle’ [15]. The first case of horizontal TE transfer (HTT) was the *P* TE identified between *Drosophila willistoni* and *Drosophila melanogaster* [16]. Recently, with the availability of many plant genome sequences, several HTT cases have been reported mainly through comparative genomic approaches. For example, multiple HTs of the LTR-RT *RIRE1* were identified within the genus *Oryza* [17], and another LTR-RT family *Route66* were found and proven to be HTs among the rice, maize and sorghum genomes through a comparative genomics analysis [18]. In addition, 32 HTs of LTR-RTs were discovered by whole genome surveys and comparative analyses in 46 sequenced plant genomes [19].

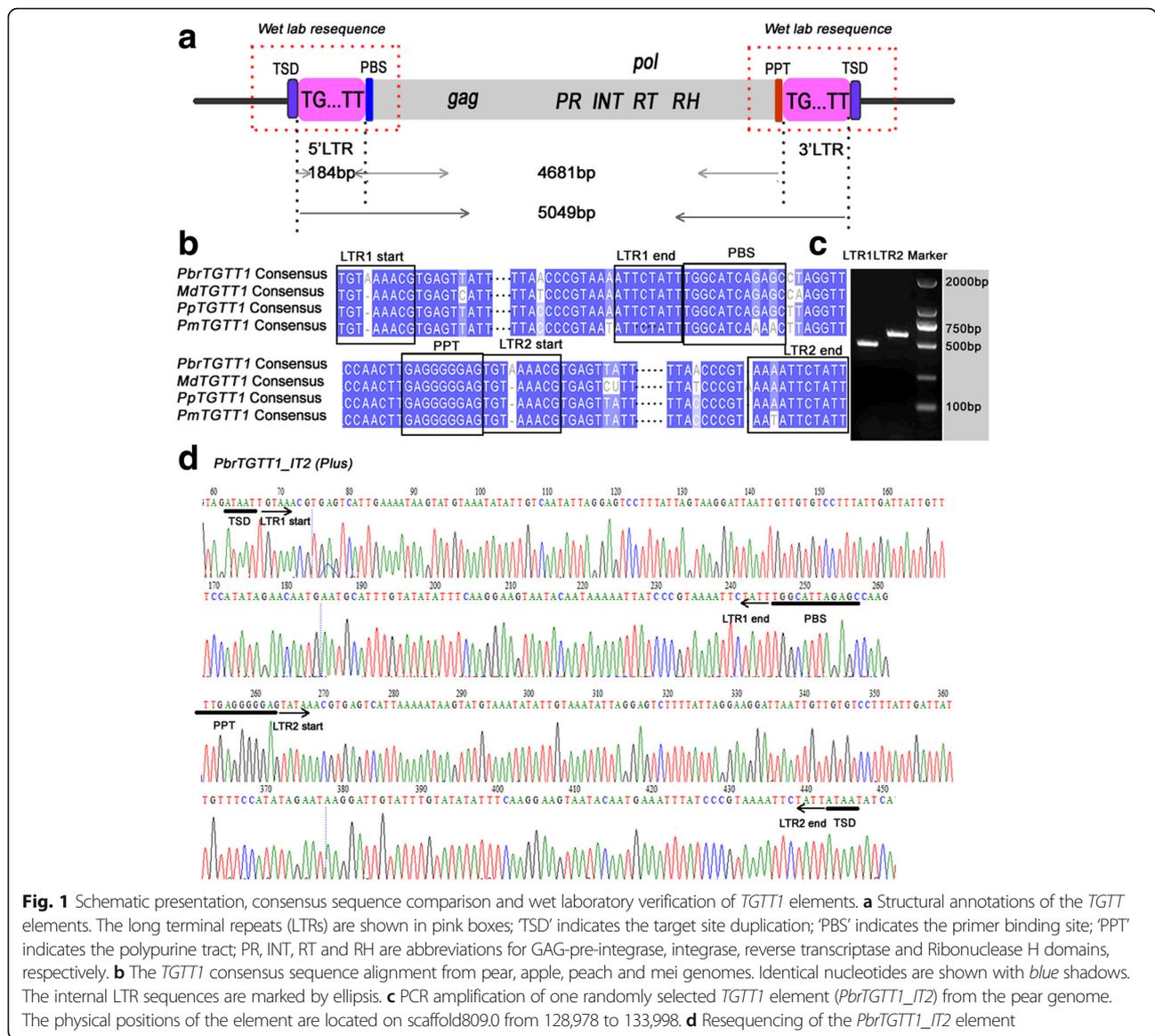
The propensity of LTR-RTs not only contributed to genome size but also resulted in byproducts of gene disruption, expression level alterations and genomic rearrangements by inserting themselves into genes or their promoter regions [20, 21]. In plants, LTR-RTs are usually silent under normal conditions, but some show transcriptional activities and increased accumulations while under stress, potentially triggering the genetic diversity required to evolve adaptations [21, 22]. For example, salt (*AtCopeg1* in *Arabidopsis* [23]), drought (*BARE1* in barley [24]), heat (*ONSEN* in *Arabidopsis* [25, 26]), cold (*mPing* in rice [27, 28]) and wounding (*Corky* from *Quercus* [29]; *CLCoy1* in lemon [30], *OARE1* in oat [31] and *Tnt1* in tobacco [32]). Recently, several LTR-RTs were identified as being expressed in the fruits and buds of pear in the RNA-seq databases [33]. However, their study did not focus on the LTR-RTs’ transcription activities under stress in pear.

The Rosaceae family is an economically important angiosperm lineage, containing over 3000 distinct species with chromosome’s numbering from 7 to 17 pairs [34]. Some genera with higher economic values that are widely cultivated have had their whole genomes sequenced in the last decade, including pear (*Pyrus bretschneideri*, *n* = 17, 527 Mb) [4], apple (*Malus domestica*, *n* = 17, 743 Mb) [5], peach (*Prunus persica*, *n* = 8, 265 Mb) [6], mei (*Prunus mume*, *n* = 8, 280 Mb) [35] and woodland strawberry (*Fragaria vesca*, *n* = 7) [36] (Additional file 1: Table S1). Based on DNA sequence data, *Fragaria* belongs to the Rosoideae, supertribe Rosadea, tribe Potentilleae, *Malus* and *Pyrus* occur in the Spiraeoideae, supertribe Pyrodeae, tribe Pyreae and *Prunus* is in the Spiraeoideae, tribe Amygdaleae [37]. The availability of the five Rosaceae genomic sequences provided opportunities to undertake comparative analyses of LTR-RTs in pear and four other genomes [3, 38]. In this study, a genome wide identification of non-typical LTR-RTs in pear genome was conducted. Two new subfamilies of LTR-RTs, *TGTT* and *AACA*, were identified in pear, apple, peach and mei, but not in strawberry. Their structures, abundance levels, insertion time frames, evolution and transcription activities have been comprehensively analyzed between the four Rosaceae species. *TGTT* and *AACA* elements terminate in short inverted repeat dinucleotides, such as ‘TG’ and ‘TT’, ‘AA’ and ‘CA’, and the *AACA1* elements in peach may have been horizontally transferred from apple. In addition, multiple elements from the two subfamilies present differential expression levels in different pear organs and also show different expression levels under heat, cold and salt stress treatments. Our study reveals novel structures, horizontal transfer and the transcription activation of two new LTR-RT subfamilies, providing additional information on, and knowledge of, the structure, evolution and activity of TEs in plants.

## Results

### Identification, structural characterization and sequence analysis of *TGTT* and *AACA* TEs in the pear genome

We started our analyses by focusing on a class of atypical LTR-RTs identified in the pear (*P. bretschneideri*) genome. Initially, 12 intact TEs with atypical characteristics were identified using the LTR\_STRUC program [39]. The LTRs of the 12 intact TEs terminate in the dinucleotide 5′-TG-TT-3′ or 5′-AA-CA-3′ (Figs. 1a,b and 2) instead of 5′-TG-CA-3′ usually found in typical LTR-RTs. Thus, these TEs were classified into two subfamilies, named *TGTT* and *AACA*, based on their terminal dinucleotides. In total, 66 intact TEs with two clearly defined boundaries and TSDs were identified using combined homology-based approaches as previously described [3, 13, 38]. Using the unified classification for eukaryotic

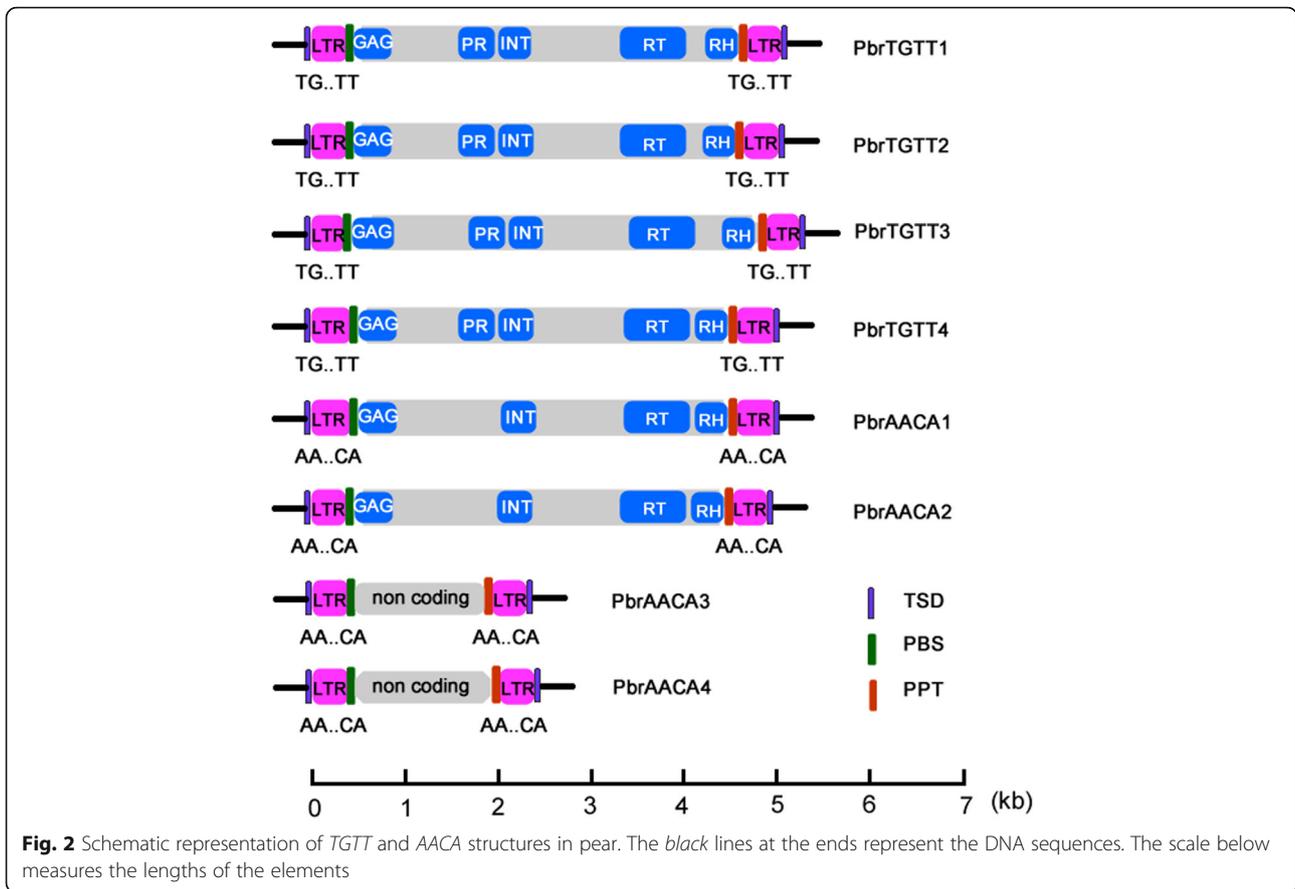


**Fig. 1** Schematic presentation, consensus sequence comparison and wet laboratory verification of *TGTT1* elements. **a** Structural annotations of the *TGTT1* elements. The long terminal repeats (LTRs) are shown in pink boxes; 'TSD' indicates the target site duplication; 'PBS' indicates the primer binding site; 'PPT' indicates the polypurine tract; PR, INT, RT and RH are abbreviations for GAG-pre-integrase, integrase, reverse transcriptase and Ribonuclease H domains, respectively. **b** The *TGTT1* consensus sequence alignment from pear, apple, peach and mei genomes. Identical nucleotides are shown with blue shadows. The internal LTR sequences are marked by ellipsis. **c** PCR amplification of one randomly selected *TGTT1* element (*PbrTGTT1\_IT2*) from the pear genome. The physical positions of the element are located on scaffold809.0 from 128,978 to 133,998. **d** Resequencing of the *PbrTGTT1\_IT2* element

TEs [40], the 66 TEs were grouped into eight distinct families based on an over 80% identity in at least 80% of their LTR regions (Table 1, Additional file 1: Tables S2 and S3 and Additional file 2: Figure S1). Four families, containing 35 TEs, belonged to the *TGTT* subfamily and the other four families, containing 31 TEs, belonged to the *AACA* subfamily. We randomly selected nine elements and confirmed the existence of 5'-TG-3' and 5'-AA-3' terminals in the LTR sequences by PCR and Sanger sequencing (see Methods, Fig. 1c and d, Additional file 2: Figure S2 and Additional file 1: Table S4).

The consensus sequence sizes of the eight families ranged from 2129 (*PbrAACA3*) to 5114 bp (*PbrTGTT3*), and the LTR sequence sizes ranged from 152 (*PbrTGTT3*) to 266 bp (*PbrAACA1*, Table 1). The coding

sequences of the 66 elements indicated that all of the *PbrTGTT* elements contained the *Gag* and *Pol* genes, including the protease (PR), integrase (INT), reverse transcriptase (RT), and RNase H (RH) domains. The *PbrAACA1* and *PbrAACA2* TE elements also contain the *Gag* and *Pol* genes, but the PR domain was absent in their *Pol*s. The order of *int*, *rt*, and *rh* defined the six families (46 elements) as *Copia*-like elements (Fig. 2). Interestingly, no coding sequences were identified in the short internal sequences between the two LTRs (1641–2042-bp) of *PbrAACA3* and *PbrAACA4* (20 elements), indicating that these two *AACA* families were TRIM families (Fig. 2). Notably, the TSD sizes of the *PbrAACA* elements were 6 bp, while those of the *PbrTGTT* family varied from 4 to 6 bp (Table 1).



**Table 1** Summary of *TGTT* and *AACA* elements in four Rosaceae genomes

Family	Superfamily Lineage	No. of intact element	Length of LTR (bp)	Length of intact element (bp)	Start and end of LTR (Plus)	Length of TSDs	Ave. age (mys)
<i>PbrTGTT1</i>	<i>Copia/Ale</i>	8	184	5049	TG..TT	5	2.32
<i>MdTGTT1</i>	<i>Copia/Ale</i>	6	179	5113	TG..TT	5	0.91
<i>PpTGTT1</i>	<i>Copia/Ale</i>	4	179	5009	TG..TT	5	0.16
<i>PmTGTT1</i>	<i>Copia/ Ale</i>	2	180	5055	TG..TT	5	0.65
<i>PbrTGTT2</i>	<i>Copia/ Ale</i>	22	180	5039	TG..TT	5	0.3
<i>MdTGTT2</i>	<i>Copia/ Ale</i>	3	169	5046	TG..TT	5	0.92
<i>PbrTGTT3</i>	<i>Copia/ Ale</i>	2	152	5114	TG..TT	4	0.38
<i>PbrTGTT4</i>	<i>Copia/ Ale</i>	3	229	4999	TG..TT	6	0.28
Subtotal/average		50	180	5053	TG..TT	5	0.74
<i>PbrAACA1</i>	<i>Copia/ Ale</i>	7	266	4924	AA..CA	6	0.94
<i>PmAACA1</i>	<i>Copia/ Ale</i>	2	251	4883	AA..CA	6	0.15
<i>PbrAACA2</i>	<i>Copia/ Ale</i>	4	201	4857	AA..CA	6	0.05
<i>PbrAACA3</i>	<i>TRIM</i>	4	244	2129	AA..CA	6	1.85
<i>PbrAACA4</i>	<i>TRIM</i>	16	203/242	2364/2522	AA..CA	6	1.84
<i>MdAACA4</i>	<i>TRIM</i>	8	207/242	2606/2735	AA..CA	6	1.03
Subtotal/average		41	232	3378	AA..CA	6	1.26

Note: *PbrAACA4* and *MdAACA4* elements can be separated into two sub-groups based on their sequence length, since their sequence identity and sequence length are still over 80%, the two sub-groups were still classified into one family

**TGTT and AACA TEs are also present in other Rosaceae genomes**

To detect whether the *TGTT* and *AACA* elements are specific to the pear genome, these elements were annotated in other published plant genomes at phytozome (<http://www.phytozome.net>) using the same strategies as described above. Only four families of *TGTT* and *AACA* were identified in three other closely related Rosaceae genomes, apple (*M. domestica*) [5], peach (*P. persica*) [6] and mei (*P. mume*) [35] (Additional file 1: Table S1). To distinguish these TEs in different genomes, we have named them *MdTGTT1*, *PpTGTT1*, *PmTGTT1*, *MdTGTT2*, *PmAACA1* and *MdAACA4* (Table 1, Additional file 1: Table S3). In total, six *MdTGTT1* copies, four *PpTGTT1* copies, two *PmTGTT1* copies, three *MdTGTT2* copies, two *PmAACA1* and eight *MdAACA4* copies, which are all less than the number in pear, were identified. No *TGTT* or *AACA* TEs were identified in the closely related Rosaceae species, woodland strawberry (*F. vesca*) [36] or other published plant genomes.

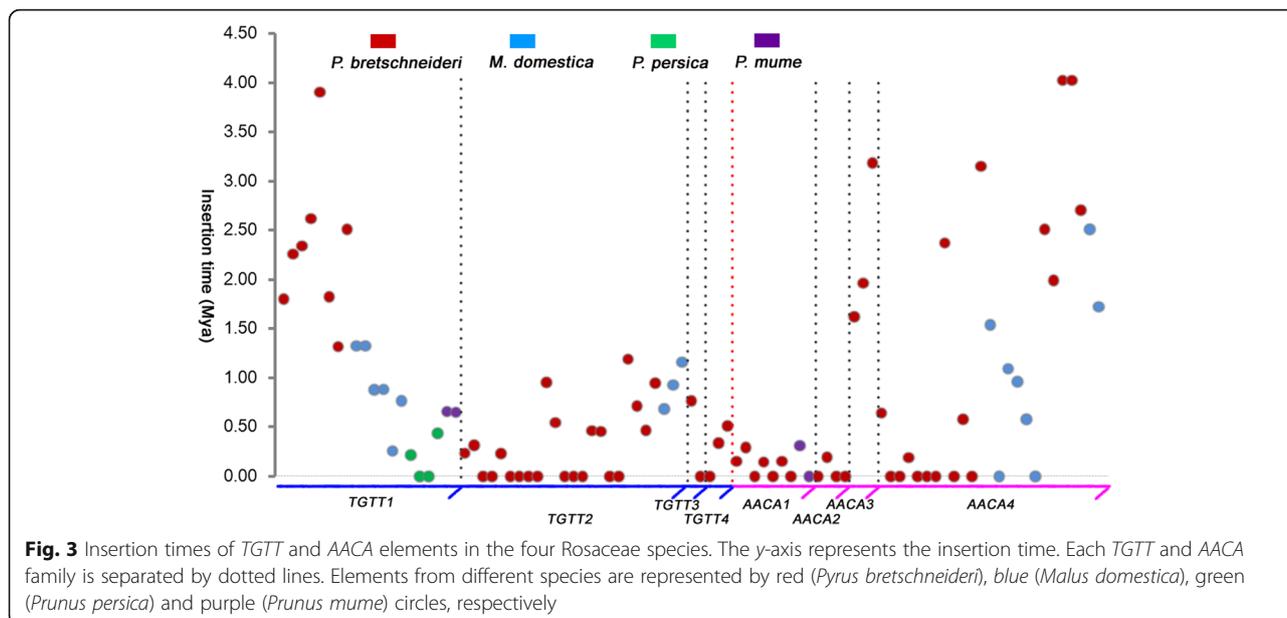
**Variable spectra of activities for amplification of TGTT and AACA elements over evolutionary time**

To compare the activities and amplification time frames of *TGTT* and *AACA* elements among the four Rosaceae species, the full-length TEs with TSDs were dated using a previously described approach [41, 42]. Even though the two LTR sequences of an intact LTR-RT element are identical at the time of insertion, both LTRs accumulate nucleotide substitutions independently over evolutionary time. Thus, when an evolutionary rate is applied to the LTR-RT element, the sequence divergence of two LTRs can be roughly converted into the insertion time.

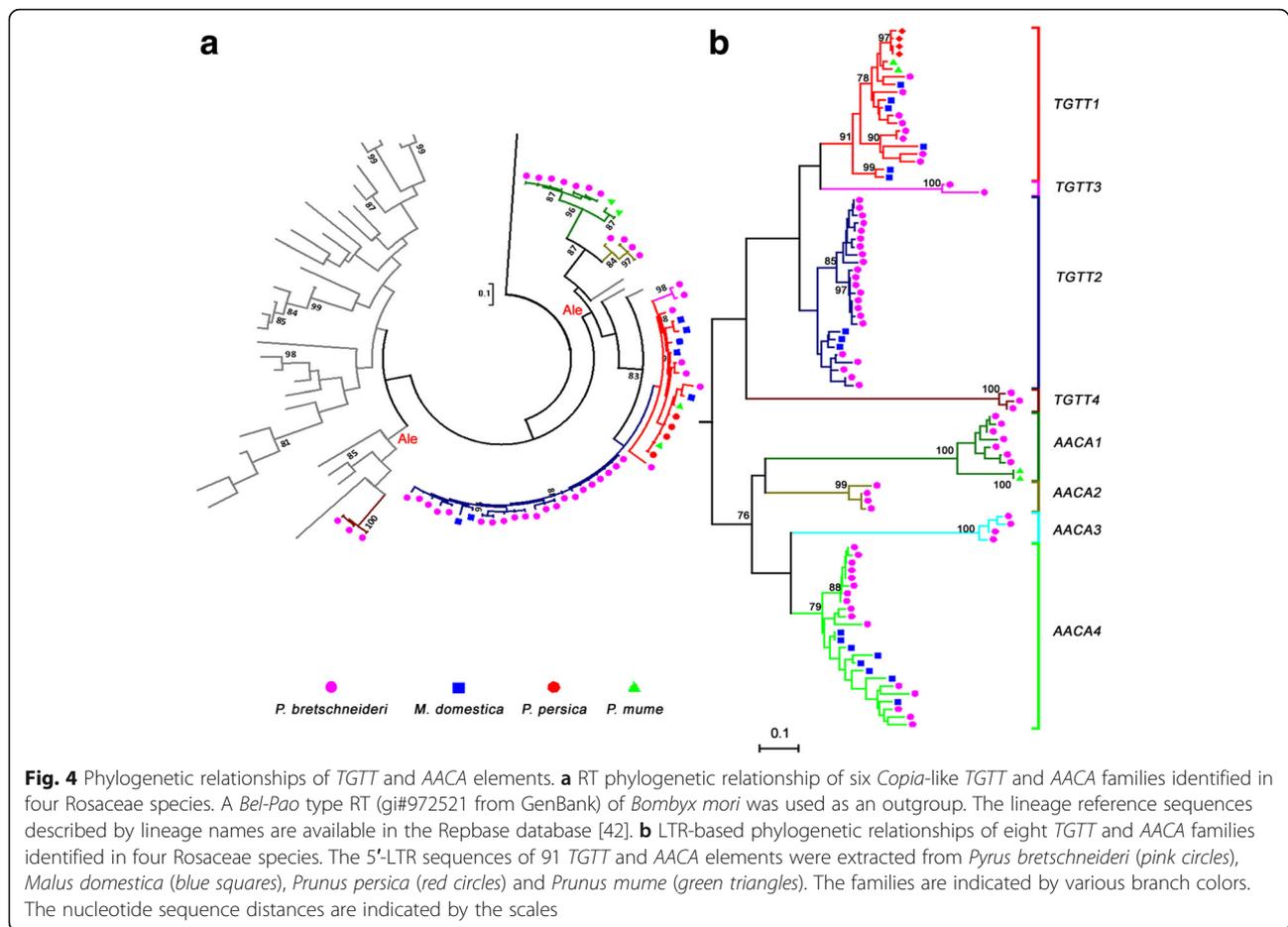
Although the evolution rate of LTR-RTs varies among different loci, families, and lineages [43], an estimation of  $1.3 \times 10^{-8}$  per site per year has been applied in many studies [13, 42, 44]. Using this rate, the insertion times of the 50 *TGTT* and 41 *AACA* intact copies with TSDs from the four Rosaceae species were estimated. The following was observed: 1) the average insertion times of *TGTT* and *AACA* subfamilies are 0.74 and 1.26 million years (Mys), respectively; 2) the average insertion times of the eight families in the four Rosaceae species ranged from 0.05 (*PbrAACA2*) to 2.32 Mys (*PbrTGTT1*). Most of these elements (65, 71.43%) inserted into the genome <1.0 million years ago (Mya), and 21 copies (23.08%) integrated into the genome within 1–3 Mya. In addition, only five copies (5.49%) have been dated >3 Mya; 3) over one third of these *TGTT* and *AACA* elements (31, 34.07%) have two identical LTRs, and the ratio of *TGTT* to *AACA* TEs is almost 1:1 (15:16, Fig. 3, Additional file 1: Table S3); and 4) the average insertion times of *TGTT* and *AACA* TEs varied among pear, apple, peach and mei, at 1.00, 0.95, 0.16 and 0.40 Mya, respectively (Fig. 3).

**The evolutionary relationship between TGTT and AACA TEs**

To understand the evolutionary relationships among *TGTT* and *AACA* TEs in the four Rosaceae species, a phylogenetic tree using the 5' LTR sequences was constructed (Fig. 4a). The 91 *TGTT* and *AACA* TEs can be successfully separated into eight clades. The four *TGTT* families clustered together, and the four *AACA* families were closer to each other, indicating that the *TGTT* and *AACA* TEs evolved independently. Although *TGTT* and *AACA* TEs are separated from each other, the *MdTGTT2*



**Fig. 3** Insertion times of *TGTT* and *AACA* elements in the four Rosaceae species. The y-axis represents the insertion time. Each *TGTT* and *AACA* family is separated by dotted lines. Elements from different species are represented by red (*Pyrus bretschneideri*), blue (*Malus domestica*), green (*Prunus persica*) and purple (*Prunus mume*) circles, respectively



and *MdAACA4* elements were mixed with *PbrTGTT2* and *PbrAACA4* elements, respectively. In addition, the *PpTGTT1* and *PmTGTT1* were also found with *MdTGTT1* and *PbrTGTT1*, indicating that the species may have experienced some introgression in early stages of their evolution or HT events after their divergence.

The individual *Copia*-like LTR-RT families can be separated into six major evolutionary lineages, *Angela*, *Ale*, *Bianca*, *Ivana*, *Maximus* and *TAR*. To discern the evolutionary history and phylogenetic relationships among the four *TGTT* and two *AACA* *Copia*-like families and the major evolutionary lineages, the conserved RT DNA sequences from each of the *TGTT* and *AACA* elements, as well as *Copia*-like LTR-RTs in *Arabidopsis*, rice and soybean, which were previously identified [42], were used to construct a Maximum Likelihood (ML) phylogenetic tree. As shown in Fig. 4a, the six *Copia*-like *TGTT* and *AACA* families all belong to the *Ale* lineage but formed three distinct sublineages. The two *AACA* families were separated into a sublineage, three *TGTT* families (*TGTT1*–*3*) formed another sublineage, with two sublineages being closer to soybean *Ale* elements, while the *TGTT4* elements, together with the *Arabidopsis* *Ale*

elements, grouped into a distinct sublineage. Because the two *AACA* TRIM families have no coding genes inside the internal regions, their PBS sites were used to make multiple alignments with those of other elements. Interestingly, the PBS sites were highly conserved with those of other *Ale* lineage elements (Additional file 2: Figure S3). Thus, the two *AACA* TRIM families may also originate from the *Ale* lineage.

#### HT of *TGTT1* elements between apple and peach genomes

*TGTT1* is the only family identified in all four Rosaceae species (pear, apple, peach and mei) but not in the woodland strawberry and other species, indicating that these TEs arose after the divergence of strawberry (*F. vesca*,  $n = 7$ ) and the ancestors of pear (*P. bretschneideri*,  $n = 17$ ), apple (*M. domestica*,  $n = 17$ ), peach (*P. persica*,  $n = 8$ ) and mei (*P. mume*,  $n = 8$ ). Previously, the 20 *TGTT1* elements from four Rosaceae species were shown to be mixed in an ML phylogenetic tree based on LTR sequences (Fig. 4b). To understand the evolutionary history of the *TGTT1* elements, the ML phylogenetic tree of *TGTT1* was rebuilt using the whole complete

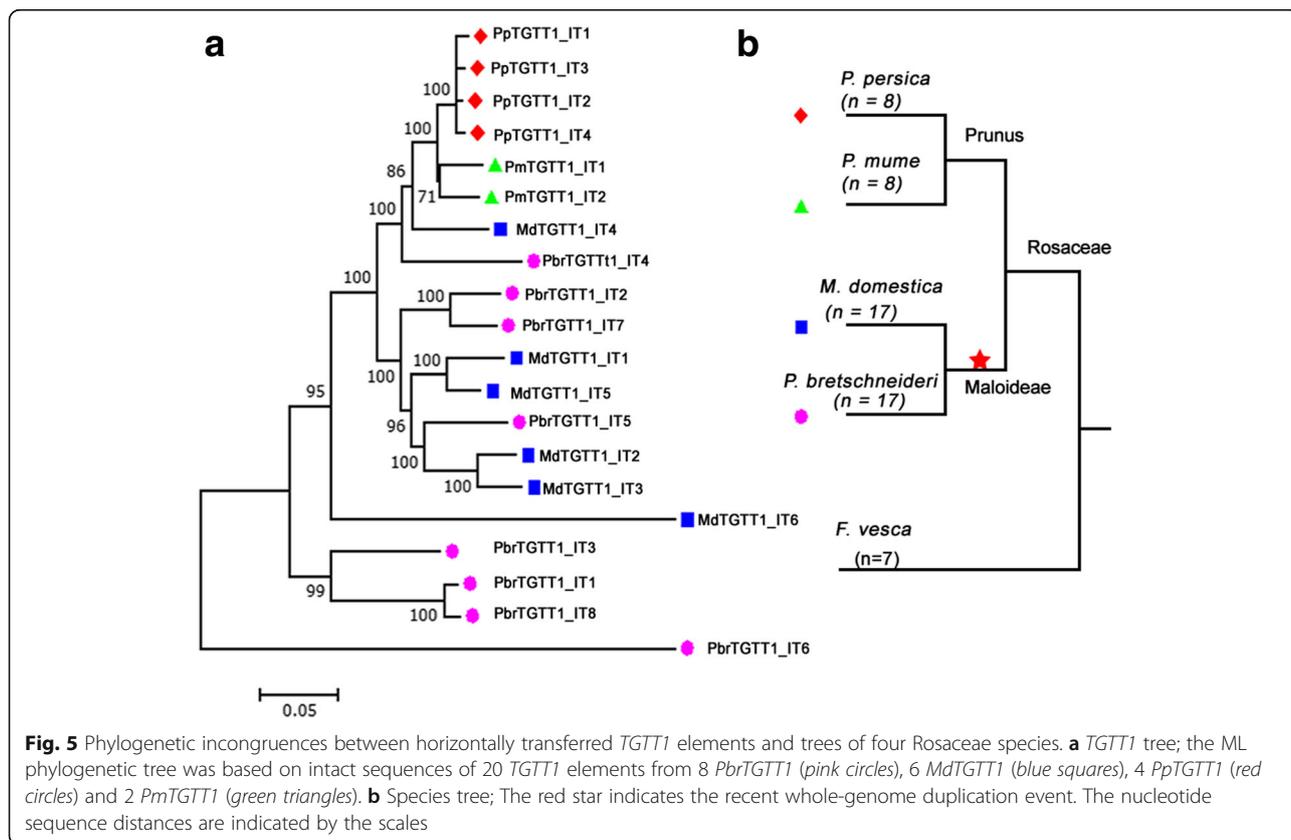
sequences of the 20 TEs, and the data indicated that the *TGTT1* TEs had a patchy distribution in the phylogenies (Fig. 5a). Specifically, four *PpTGTT1* and two *PmTGTT1* TEs were phylogenetically closer to the apple *TGTT1* element (*MdTGTT1\_IT4*). The phylogenetic relationships among these *TGTT1* TEs were not fully congruent with their host species (Fig. 5b), which prompted an investigation into the possibility of HT occurring between distantly related *Prunus* (peach and mei) and *Maloideae* (apple and pear) species.

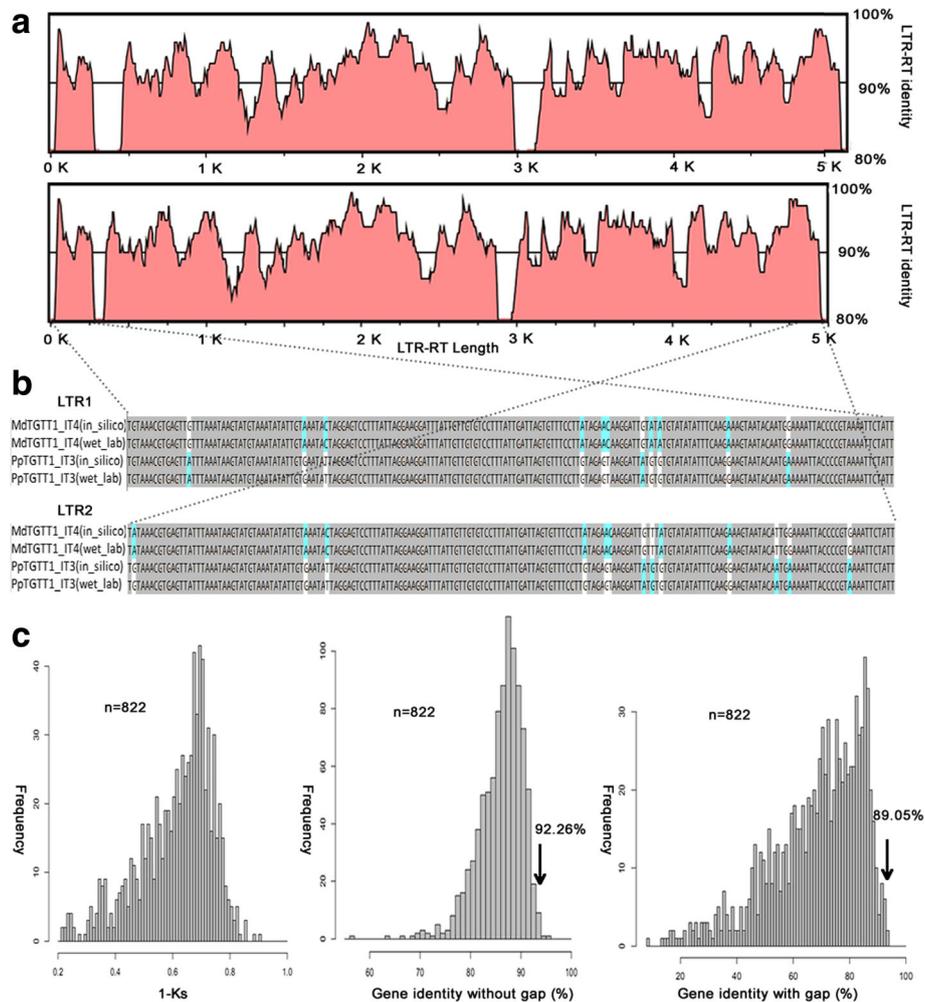
To test this hypothesis, the sequence identities between pairs of *TGTT1* elements were initially analyzed. A higher sequence similarity (92.26%, gaps were excluded) was identified between peach (*PpTGTT1\_IT3*) and apple (*MdTGTT1\_IT4*), and even when taking into account the gaps between the two sequences, the sequence similarity was 89.05% (Additional file 1: Table S5). Additionally, to make a comparison, sequence identities of single orthologous genes were calculated between peach and apple based on their synonymous substitutions per site (Ks; average = 61.29% ± 12.85%) and sequence similarities (gaps calculated average = 69.80% ± 15.56%; gaps excluded average = 86.57% ± 3.59%). Sequence identity at the peak of the distribution should be a good indicator of overall genomic divergence [19]. Here, the sequence identities of *PpTGTT1\_IT3* and *MdTGTT1\_IT4* were always higher

than the sequence identity peak values of 822 orthologous single genes (Fig. 6, Additional file 1: Tables S6 and S7). This hypothesis was also supported by the Ks values of *TGTT1* family integrases, which are much lower for *PpTGTT1* (0.02 ± 0.01) than for *MdTGTT1* (0.78 ± 0.29) (Table 2), suggesting that it recently entered the peach genome. The presence of the two elements with higher sequence identities in apple and peach were tested by PCR amplification of the LTRs and Sanger sequencing (Fig. 6 and Additional file 2: Figure S2). Thus, the HT of *TGTT1* might have occurred between the distantly related apple and peach.

**Transcriptional activities of *TGTT* and *AACA* TEs in different organs and under stress treatments in pear**

Because over 70% of *TGTT* and *AACA* TEs were inserted into the four genomes <1.0 Mya, and over one third of these elements (34.07%) contained two identical LTRs, these TEs may still be transcriptionally or even transpositionally active. To detect the transcriptional activity of *TGTT* and *AACA* TEs in the four *Rosaceae* species, HISAT alignments for each were constructed using the RNA-seq data from Sequence Read Archive (SRA) database of NCBI. A total of 12 *AACA* TEs from pear, including 7 *PbrAACA1*, 1 *PbrAACA3* and 4 *PbrAACA4*, were transcriptionally active in fruit and buds of *P.*





**Fig. 6** Comparisons between the sequence identities of HT LTR-RTs and the genomic distances between the two host species involved in the HT. **a** The sequence identities along the complete lengths of the HT LTR-RTs. The black line represents the 90% identity threshold. **b** Wet laboratory validation of the HTTs. Sequenced PCR products of LTRs were aligned with the sequences that were mined from the genome sequence. **c** Histogram representing the distribution of orthologous single-gene identities based on Ks analyses and CDS comparisons with or without gaps (see Methods). The numbers of CDS pairs of orthologous single genes analyzed are as indicated (n). Arrows correspond to average sequence identities between the HT LTR-RTs

*bretshneideri* ('Dangshansuli'; Additional file 1: Table S3 and Additional file 2: Figure S5), whereas none of the *TGTT* TEs were active according to the RNA-seq data. To gain insight into the expression levels of these active TEs, three pair of primers (Additional file 1: Table S8) that corresponded to the transcribed U5 of the 5' LTR or U3 of the 3' LTR and the partial internal regions of three randomly selected elements (*PbrAACA1\_IT5*, *PbrAACA4\_IT11* and *PbrAACA4\_IT14*) were used for qRT-PCR in different pear ('Dangshansuli') organs, including fruit flesh at four developmental periods (32, 65, 99 and 143 d after flower bloom), leaves, pericarp, pollen and stylet. Transcripts of the three elements could be detected in all eight samples (Fig. 7a–c, Additional file 2: Figure S5). The transcriptional levels varied for the three elements in the different organs. For example, the

transcriptional levels of *PbrAACA1\_IT5* (Fig. 7a), *PbrAACA4\_IT11* (Fig. 7b) and *PbrAACA4\_IT14* (Fig. 7c) were highest at 32, 99 and 143 d after flower bloom in fruit flesh, respectively, and they all expressed at their lowest levels in the pollen.

The transcriptional activities of LTR-RTs are usually repressed in plant tissues during normal development, as well as in response to a variety of biotic and abiotic stresses [45]. Thus, the expression patterns of the former three elements in leaves of pear ('Dangshansuli') were examined under various stress treatments, including cold, heat and salt (see Methods). *PbrAACA1\_IT5* was significantly up-regulated in heat (6 h, Fig. 7g) and salt (6 h, Fig. 7j) stress treatments (two-tailed *t*-test,  $P < 0.01$ ), but down-regulated in the cold stress treatments (two-tailed *t*-test,  $P < 0.01$ , Fig. 7d). Elevated

**Table 2** Ka, Ks and Ka/Ks values of integrases in the *TGTT* and *AACA* families

Family	Ka	Ks	Ka/Ks
<i>PbrTGTT1</i>	0.73 ± 0.42	0.56 ± 0.18	1.33 ± 0.65
<i>MdTGTT1</i>	0.74 ± 0.39	0.78 ± 0.29	0.99 ± 0.48
<i>PpTGTT1</i>	0.01 ± 0.00	0.02 ± 0.01	0.51 ± 0.36
<i>PmTGTT1</i>	0.51 ± 0.00	0.50 ± 0.00	1.01 ± 0.00
<i>PbrTGTT2</i>	0.15 ± 0.17	0.32 ± 0.21	0.42 ± 0.38
<i>MdTGTT2</i>	0.42 ± 0.24	0.44 ± 0.20	0.85 ± 0.24
<i>PbrTGTT3</i>	0.12 ± 0.00	0.09 ± 0.00	1.25 ± 0.00
<i>PbrTGTT4</i>	0.58 ± 0.00	0.61 ± 0.00	0.96 ± 0.00
<i>PbrAACA1</i>	0.49 ± 0.73	0.33 ± 0.42	0.63 ± 0.67
<i>PmAACA1</i>	0.00 ± 0.00	0.01 ± 0.00	0.29 ± 0.00
<i>PbrAACA2</i>	0.03 ± 0.00	0.11 ± 0.00	0.25 ± 0.00

levels of *PbrAACA4\_IT11* transcripts were observed in cold (1 h, Fig. 7e) and salt (6 h, Fig. 7k) stress treatments (two-tailed *t*-test,  $P < 0.01$ ) but not in heat. For *PbrAACA4\_IT14*, the up-regulation was also significant in cold (two-tailed *t*-test,  $P < 0.01$ , 1 h, Fig. 7f) and salt stress treatments (two-tailed *t*-test,  $P < 0.01$ , 24 h, Fig. 7l). These data may indicate the constitutive expression of *TGTT* and *AACA* TEs in different pear tissues and the elevation of their expression levels under various stresses.

## Discussion

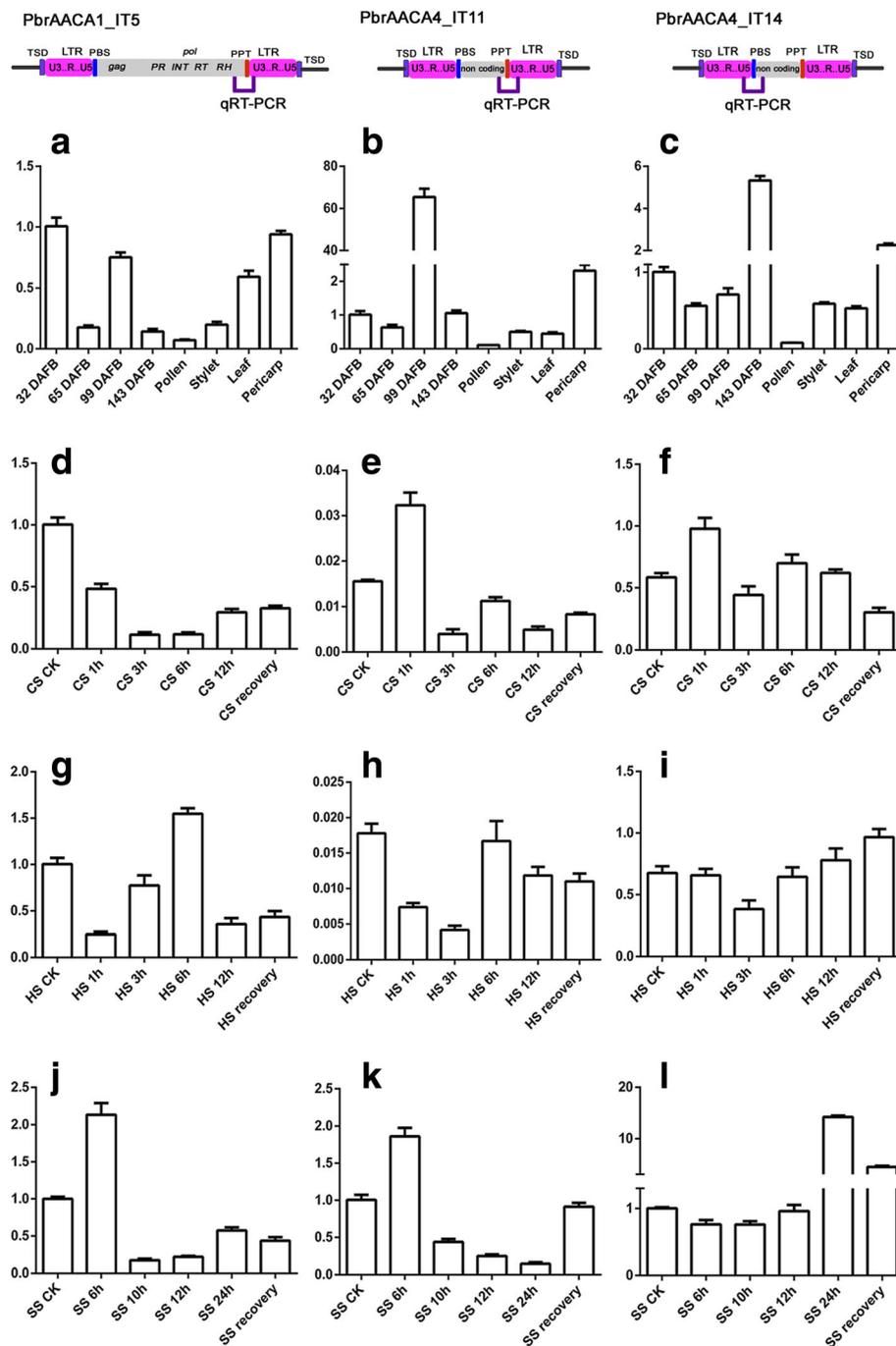
We have isolated two novel LTR-RT subfamilies called *TGTT* and *AACA* in four Rosaceae genomes, and these can be classified into eight families according to the TE classification system proposed by Wicker et al. 2007 [40]. Six *Copia*-like families were classified into the *Ale* lineage using a phylogenetic analysis (Fig. 4a), while the other two *AACA* families were *TRIM* elements owing to their short non-coding internal regions (Fig. 2). However, *TGTT* and *AACA* TEs were restricted to four Rosaceae species (pear, apple, peach and mei), and not even truncated fragments were detected in the closest woodland strawberry genome and other published plant genomes. This suggested that the two subfamilies evolved specifically after the divergence between *F. vesca* and the ancestor of the four Rosaceae species.

### Specific dinucleotide termini of two new subfamilies of non-canonical LTR-RTs in the four Rosaceae species

LTR sequences start with 'TG' and end with 'CA' in typical LTR-RTs [9]. In a previous study, our group described a systematic survey of LTR-RTs in the sequenced pear (*P. bretschneideri*) genome, and 3221 full-length LTR-RTs have been found to terminate with 5'-TG-CA-3' [3, 38]. Here, although these new TEs contained most of the typical

features of LTR-RTs, a salient difference was identified in the dinucleotide positions at both ends of the LTRs. When these TEs with their two flanking sequences (500-bp for each site) were combined and aligned, they showed an accurate insertion site terminating with 5'-TG-TT-3' or 5'-AA-CA-3', and flanked by perfect 4 to 6-bp TSDs. Thus, these atypical LTR-RTs were defined as two new subfamilies (*TGTT* and *AACA*). Previously, an atypical LTR-RT family (*TARE1*) in tomato [13] that contained 5'-TA-CA-3' at both ends of the LTRs was reported, and a plausible mutation model explaining the creation of such atypical dinucleotides in the LTRs was proposed. In addition, another exception, *AcCOPIA1*, which terminates with 'TG' and 'TA' at both ends of the LTRs, was identified in onion [14]. Similar to *TARE1*, only one nucleotide changed from 'CA' to 'TA' in *AcCOPIA1*. Compared with the *TARE1* and *AcCOPIA1* TEs, the 'CA' has been changed to 'TT' in the *TGTT* elements and 'TG' turned into 'AA' in the *AACA* elements. Interestingly, the 5'-TG-TT-3' of *TGTT* is the reverse complement of the 5'-AA-CA-3' of *AACA*. The simultaneous mutation of the dinucleotides is a low probability event that cannot be easily explained by the mutation model [13]. A global annotation, structural analysis and phylogenetic study of all the non-TGCA TEs within the eukaryotes is worth performing in the future, to unravel the scale and frequencies of non-canonical LTR-RTs that terminate with non-TGCA motifs and whether they exist naturally or derived from the normal TGCA-containing LTR-RT elements.

Although most of LTR-RTs carry the palindromic dinucleotide motif (5'-TG-CA-3') flanking each LTR, the importance of this conserved motif is still poorly understood. Studies of retrovirus integration indicate that the 3' CA terminal sequences of retroviral LTRs are essential for viral integration [46, 47]. The close relationship between retroviruses and LTR-RT TEs, with an additional envelope protein [46, 47], may explain why most LTR-RTs have the conserved 5'-TG-CA-3' motif. Here, based on the phylogenetic tree of *TGTT*, *AACA* and typical TGCA integrases, most *TGTT* integrases (*TGTT1*, *TGTT2* and *TGTT3* elements) can be differentiated from typical integrases (Additional file 2: Figure S4), whereas the *AACA* elements could not be differentiated. The Ka/Ks values of the INT domains of *TGTT1* and *TGTT3* families (>1) are significantly greater than those of *AACA1* and *AACA2* families (<1) (Table 2). Thus, the functional divergence of the integrase active sites from these 3'-TT ends LTR-RTs has occurred and might result in a novel integration mechanism. Based on sequence comparisons, structural and phylogenetic analyses, the newly identified *TGTT* and *AACA* TEs should provide a valuable resource to study the non-canonical LTR-RTs integration mechanism.



**Fig. 7** Time-course expression levels of active AACA elements in *Pyrus bretschneideri*. The positions of the primers used for transcriptional validation are indicated in the schematic of each element in qRT-PCR region. The expression levels were detected in the fruit flesh of four developmental stages, pollen, stylet, leaf and pericarp. **a–c**, leaves under cold (**d–f**), heat (**g–i**), and salt (**j–l**) treatments for *PbrAACA1\_IT5* (**a, d, g, j**), *PbrAACA4\_IT11* (**b, e, h, k**) and *PbrAACA4\_IT14* (**c, f, i, l**). Error bars indicate the standard deviations of three biological replicates

**The *PpTGTT1* elements originated from *MdTGTT1* through HT**

The transmission of genetic materials among sexually isolated species is usually defined as HTs. HTTs were first proposed as a possible dissemination mechanism of

TEs in eukaryotes [15]. Because TEs could undergo epigenetic-mediated silencing by the host genome [48], HTTs could be the mechanism of escaping the silencing and ensuring the long-term survival of TEs among eukaryotic lineages. Based on this model, most of the

active TEs found in plant and animal genomes may originate from other species through HTT [49]. Owing to the availability of more released eukaryotic genome sequences and standard comparative genomics approaches [50], hundreds of cases of HTTs have been reported over the past years [18, 19, 51]. Recently, Fawcett and Innan (2016) [52] proposed a method to differentiate the HTT and vertical transmission scenarios, which involves testing whether the hypothetical HTT copies are present in the orthologous regions of the two species [52]. If the two species acquired the copies independently by HTs, then the two species should not share any copies and each copy should be species-specific. This theory is reasonable because TE-based recombination and loss occur frequently in host genomes. The same analysis was conducted for the *TGTT1* TEs among the four Rosaceae genomes, however, no shared intact *TGTT1* TEs or degenerated fragments were detected between pairs of the Rosaceae genomes. Combined with the high similarity between *MdTGTT1\_IT4* and *PpTGTT1\_IT3*, the deep divergence time between Maloideae and Prunus (>45 Mys), and their patchy distribution in Rosaceae, indicates that the HT of the *TGTT1* TEs occurred between apple and peach. Recently, 32 clear cases of recent HTTs of LTR-RTs, including 5 HTTs between apple and peach, were detected among 46 sequenced plant genomes [19]. As expected, the HTT of *TGTT1* TEs was not included in the five reported HTTs between apple and peach, possibly because the *TGTT1* TEs were not identified initially through their LTR-RTs annotation method. Thus, the estimation of millions of HTTs occurring among the angiosperms in the recent evolutionary past may have been an underestimation.

The insertion time of the HTT elements have facilitated us to speculate the HTT history and time frame. Based on the sequence divergence of two LTRs of *MdTGTT1\_IT4* and *PpTGTT1\_IT3*, we propose that the presence of *TGTT1* in peach was resulted from HT of *MdTGTT1* between 0.43 Mys and 0.88 Mys. First, the average insertion time of the four *PpTGTT1* elements (0.16 Mys) is much younger than that of the six *MdTGTT1* elements (0.91 Mys, Additional file 1: Table S3), especially, *PpTGTT1\_IT3* (0.43 Mys) aged much younger than *MdTGTT1\_IT4* (0.88 Mys); Second, *PpTGTT1\_IT3* is the oldest element of the four *PpTGTT1* elements (Additional file 1: Table S3); Third, four *MdTGTT1* elements including *MdTGTT1\_IT4* are still transcriptionally active (Additional file 2: Figure S5 and Additional file 1: Table S3). Last, the cluster of the horizontally transferred *PpTGTT1* copies is included in the larger cluster of copies from apple and pear. All of these evidences suggest that peach is the recipient species of the HT event. Although several studies suggest that “host-vector species” interactions may favor HTTs in animals [53, 54]. However, evidence of “host-vector-driven” HTTs has not been provided in plants, and no experimental evidence of this

process has been reported yet. Thus it is unclear how the transfer of *TGTT1* may have occurred between apple and peach. Considering that apple and peach belong to Rosaceae fruit crops with higher economic values, the most plausible explanation is that the *TGTT1* was transmitted by their common pathogen, such as bacteria, fungi and virus often believed to be the vectors of HT [55, 56], and perhaps with the help of a plant cell-piercing insect [54, 55]. Thus *TGTT1* should be an attractive candidate for testing whether similar mechanisms of HTTs exist in plants.

#### Varied transcription activities of these non-canonical TEs in pear and apple

Although the LTR-RTs are less likely to be actively expressed in plant tissues during normal development, several exceptions have been reported in various organs belonging to different species, such as *Ogre* elements in leaves, roots and flowers of pea [57], *Grande* elements in leaves of *Zea* and *Tripsacum* [58], eight LTR-RT families in leaves, stalks and roots of *Eucalyptus* genus [59], and *EARE-1* elements in roots, staminate flowers, pistillate flowers, leaves and seeds of *Excoecaria agallocha*, which were all detected as transcriptionally active [60]. In our study, three families (12 *AACA* TEs) from pear were initially detected with transcriptional activity using the published RNA-seq data in the SRA database of NCBI, whereas no transcripts of the six families from peach (*PpTGTT1*), mei (*PmTGTT1* and *PmAACA1*) and apple (*MdTGTT1*, *MdTGTT2* and *MdAACA1*) were identified. In particular, all of the *TGTT* TEs were silenced in the four species, indicating that the transcriptional activities of the *TGTT* and *AACA* TEs varied in different species. The qRT-PCR analyses of three randomly selected elements from pear (*PbrAACA1\_IT5*, *PbrAACA4\_IT11* and *PbrAACA4\_IT14*) also proved that these *TGTT* and *AACA* TEs are transcriptionally active at different levels in different organs. In addition, various elevated transcript levels of the three pear elements were observed following heat, cold and salt treatments, indicating that the *TGTT* and *AACA* TEs could be activated by stresses. This is coincident with the discovery that several other LTR-RTs are frequently activated under stress conditions [28, 60, 61], and also conforms to McClintock’s theory of genome shock in which the enhanced activities of TEs under stress might represent an evolutionary strategy for plant species to increase the chances of survival under unfavorable conditions [62]. Although the transcriptional activities of *TGTT* and *AACA* TEs were not detected in peach and mei using the SRA and EST databases from NCBI, the *PpTGTT1* family has proliferated into four copies since the HT of *PpTGTT1\_IT3* from apple, indicating that the element was active for a short time after invasion and then the life cycle of *PpTGTT1* may have been firmly controlled at the post-

transcriptional level. Further studies will be conducted in the future.

## Conclusions

*TGTT* and *AACA* are two new types of LTR-RT subfamilies isolated from pear, apple, peach and mei that terminate with atypical dinucleotide structures. Their family and element copy numbers, proliferation time frames and transcriptional activities varied among the four Rosaceae species. HT might have played a significant role in the life cycle of *TGTTI*. These newly identified TEs should be valuable materials for the further investigation of atypical LTR-RTs in other sequenced plant species and will provide interesting insights into their structural evolution and TE-driven genomic evolution.

## Methods

### Genome sequence resources, annotation and classification of *TGTT* and *AACA* LTR-RTs

The genome sequence data for the four Rosaceae species are available in Additional file 1: Table S1. The annotation method of *TGTT* and *AACA* LTR-RT elements has been widely used in previous studies [3, 13, 38]. First, based on the structural analysis, several intact elements were identified by the LTR\_STRUC program [39]. Then, all of the identified LTR sequences of the intact elements with clearly defined boundaries were used as queries to detect additional intact elements through sequence homology searches using CROSS\_MATCH and CLUSTALW programs with default parameters. Finally, the structures and boundaries of all of the identified LTR-RTs were manually inspected and confirmed, and the TSD sites were defined with one mismatch allowed. Fragments and truncated elements were not analyzed in this study. The *TGTT* and *AACA* LTR-RTs were classified into superfamilies based on the conserved functional domains detected using the BLASTX tool. The queried domains included GAG (for UBN2 superfamily domain, pfam14223), PR (for GAG-pre-integrase domain, pfam13976), INT (for integrase core domain, pfam00665), RT (for reverse transcriptase domain, pfam07727) and RNase H (for Ribonuclease H domain, cd09272). Each individual family was classified using sequence homology comparisons according to the criteria described previously [40].

### Estimation of insertion time

The insertion time of intact elements with TSD sites was estimated by comparing the divergence of their 5' and 3' LTR sequences because both LTR sequences of a newly proliferated LTR-RT were believed to be identical at the time of integration [41]. To investigate the nucleotide substitution rate for each element, the two LTR sequences were aligned using the MUSCLE program with default

parameters [63]. The insertion time (T) for each intact element was calculated using the formula:  $T = K/2r$ , in which the average number of substitutions per aligned site (K) was corrected using the Jukes–Cantor method [64], and  $1.3 \times 10^{-8}$  substitutions per site per year was used as the average LTR substitution rate (r) [44].

### Phylogenetic analysis

For each *TGTT* and *AACA* TE, the 5' LTR sequence, and RT and INT domains were extracted from the intact sequence using a perl script. Sequence alignments were performed by the MAFFT version 7 program with default options [65]. The MEGA 5.2 program implemented with Jukes–Cantor model was employed for building the Maximum Likelihood trees based on 1000 bootstrap replicates [66]. The taxonomic tree was built using the common tree tool on the NCBI website (<http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi>, last accessed December 25, 2016).

### Identification of orthologous single genes and estimation of genomic sequence divergence

The strategy to identify orthologous single genes between the apple and peach genomes has been used in previous studies [4, 38]. First, the genomic protein and CDS sequences of apple and peach were downloaded from the Phytozome website (<http://www.phytozome.net>) and set as a database. Then, the BLASTP and orthoMCL software [67] were employed to identify all the orthologous single genes in the two genomes using the same parameters in the previous study [38]. All of the identified single-copy orthologous genes were manually inspected, and gene sequences that contained frame-shift mutations or stop codons were excluded from further analysis.

The  $K_a$ ,  $K_s$ , and  $K_a/K_s$  ratio of orthologous single genes and the intra-family INT domains of *AACA* and *TGTT* TEs were calculated using the YN00 program implemented in the PAML software package [68].

The CDS sequence identities of orthologous single genes in apple and peach were computed using an in-house perl script, which ran in the following three steps: (1) All of the identified orthologous single-gene pairs were separately aligned using MUSCLE software; (2) For each orthologous single-gene pair, the numbers of identical nucleotides (I), mismatches (M) and gaps (G) were counted. (3) Gene identities without gaps were calculated using the formula:  $I/(I + M) \times 100$ , and gene identities with gaps were calculated using the formula:  $I/(I + M + G) \times 100$  (Additional file 1: Table S7). The sequence identity analysis between the *TGTTI* TEs were also conducted using the same strategy.

### PCR and sequencing analysis

The total genomic DNA of the four Rosaceae species were extracted from the young leaves using the improved cetyltrimethyl ammonium bromide method. In total, 11 *TGTT* and 5 *AACA* TEs from eight families were randomly selected for validation. For each element, 300-bp 5'-flanking sequences and 300-bp 3'-flanking sequences of both LTR sequences were extracted and used to design primers (Additional file 2: Figure S2 and Additional file 1: Table S5). Polymerase chain reactions (PCR) were performed in a total volume of 25 µl, containing 1 µl of 50 ng/ml genomic DNA template, 2.5 µl of 10× buffer (without MgCl<sub>2</sub>), 2.5 µl of 2.5 mM dNTP mixture, 2.5 µl of 25 mM MgCl<sub>2</sub>, 0.8 µl each of forward and reverse primer (10 pmol/ml) and 0.2 µl of 5 U/ml Taq polymerase (Takara Biotechnology Company, Dalian, China). 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 PCR products were resolved on 1% agarose and detected by ethidium bromide staining. The analyses were performed three times and loaded on independent gels. All of the specific PCR products were isolated with the DNA Gel Extraction kit AxyPrep (Axygen Inc.). The fragments were cloned into the pMD19-T vector (Takara, China), and the plasmids were sequenced by Invitrogen (Shanghai, China).

### Transcriptional activity analysis of *TGTT* and *AACA* elements

The Illumina RNA-Seq data of four samples from the SRA database of NCBI (<https://www.ncbi.nlm.nih.gov/>), including pear fruit peel (*P. bretschneideri* 'Dangshansuli'; SRX298075), pear bud (*P. bretschneideri* 'Dangshansuli'; SRX147917) and apple leaves (*M. domestica* 'Gala'; SRX1150925), were used to identify the transcriptional patterns. For each element, the whole-nucleotide sequences were used as queries to construct HISAT alignments using the default parameters [69].

### Stress treatments and LTR-RT expression analysis by quantitative RT-PCR

The scions of 'Dangshansuli' (*P. bretschneideri*) were grafted to 1-year-old *Pyrus betulifolia* plants and grown in a culture room at 25 °C under long-day conditions (16 h light/8 h dark) for 30 d prior to stresses. For the cold treatment, seedlings were placed in a growth incubator set at 4 °C for 0, 1, 3, 6 and 12 h. For heat stress treatments, plants were transferred to 40 °C for 0, 1, 3, 6 and 12 h. Salt stress was carried out by watering the plants with 1600 mM NaCl solution for 0, 1, 3, 6 and 12 h. All of the samples were recovered for 24 h.

The total RNA from fruit flesh at four developmental periods (32, 65, 99 and 143 d after flower bloom), leaves, pericarp, pollen and stylet were extracted using a

cetyltrimethyl ammonium bromide-based method and digested with RNase-free DNase I (Thermo) to remove DNA contamination. According to the manufacturer's instructions, 1 mg of total RNA was reverse transcribed into cDNA using the ReverTra Ace qPCR RT Kit (Toyobo, Shanghai, China). Specific primers for *PbrAACA1\_IT5*, *PbrAACA4\_IT11* and *PbrAACA4\_IT14* were designed using the Primer 5 software (Additional file 1: Table S7). Quantitative real-time RT-PCR (qRT-PCR) was used for measuring the transcript levels of the three LTR-RTs. The PCR solution (20 µl) contained 10 µl of SYBR-Green PCR Master Mix (SYBR Premix EX Taq™, TaKaRa), 0.25 mM each of forward and reverse primer, 100 ng of cDNA template, and nuclease free water. The qRT-PCR analysis with a SYBR Green PCR kit was performed in a Light Cycler 480 (Roche, USA) Real-Time System. The reactions were conducted under the following conditions: 95 °C for 5 min, then 45 cycles of 94 °C for 10 s, 60 °C for 30 s and 72 °C for 30 s, followed by a final extension at 72 °C for 3 min. The 2<sup>-ΔΔCT</sup> method [70] was used to calculate the relative expression levels of each gene. Each sample was analyzed for three replicates. The mRNA capping enzyme gene (*Pbr035952.1*) and cytochrome B561 gene (*Pbr013721.1*) were used as internal controls for five different tissue and three stress treatments, respectively, and to normalize the relative expression levels of each LTR-RT. The expression analysis of each time point was repeated three times.

### Additional files

**Additional file 1: Table S1.** List of four species and their genomic sequence and DNA material source information used in this study. **Table S2.** Sequence identity between all *TGTT* and *AACA* consensus sequences in the four Rosaceae genomes. **Table S3.** Summary of *TGTT* and *AACA* elements identified in *P. bretschneideri*, *M. domestica*, *P. persica* and *P. mume*. **Table S4.** Primers used for wet lab validations. **Table S5.** Sequence identity between all *TGTT1* elements identified in the four Rosaceae genomes. **Table S6.** Sequence divergence of orthologous singletons between apple and peach. **Table S7.** Sequence identity of the orthologous singletons between apple and peach. **Table S8.** Primers used for Real time Quantitative PCR. (XLSX 142 kb)

**Additional file 2: Figure S1.** Sequence alignment of the *TGTT1* elements. **Figure S2.** Wet laboratory validation of the *TGTT* and *AACA* elements. **Figure S3.** Multiple alignments of PBS sites from *TGTT*, *AACA* and normal TGCA elements of the *Ale* lineage. **Figure S4.** INT phylogenetic relationships among *TGTT*, *AACA* and normal TGCA elements. **Figure S5.** Evidence of transcriptional activity in five *TGTT* and *AACA* elements. (PDF 2824 kb)

### Abbreviations

HTs: Horizontal transfers; HTT: Horizontal TE transfer; HTTEs: Horizontal transfer of transposable elements; int/INT: Integrase; LARDs: Large retrotransposon derivatives; LTR-RTs: Long terminal repeat retrotransposons; Mya: Million years ago; Mys: Million years; PBS: Primer-binding site; PPT: Polypurine tract; PR: Protease; qRT-PCR: Quantitative real time polymerase chain reaction; rh/RH: RNase H; rt/RT: Reverse transcriptase; SRA: Sequence read archive; TEs: Transposable elements; TRIMs: Terminal-repeat retrotransposons in miniature; TSD: Target site duplication

### Acknowledgements

We would like to thank many genome data contributors. The high-quality genome sequences generated by them were instrumental in conducting this study.

### Funding

This study was supported by the Natural Science Foundation of Jiangsu Province in China (BK20160715), China Postdoctoral Science Foundation funded project (2015 M570456) and the National Natural Science Foundation of China (31701890).

### Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

### Author's contributions

Conceived and designed the experiments: HY SZ. Performed the experiments: HY XW. Analyzed the data: HY XW DS YC KQ ZM. Contributed reagents/materials/analysis tools: HY XW ZM. Wrote the paper: HY SZ. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable

### Consent for publication

Not applicable

### Competing interests

The authors have declared that no competing interests exist.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>Center of Pear Engineering Technology Research, College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China. <sup>2</sup>State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China. <sup>3</sup>College of Agricultural Sciences, Nanjing Agricultural University, Nanjing, China.

Received: 19 June 2017 Accepted: 18 October 2017

Published online: 27 October 2017

### References

- Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. *Annu Bot*. 2005;95(1):127–32.
- GT V, Schmutzer T, Bull F, Cao HX, Fuchs J, Tran TD, et al. Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *Plant Genome*. 2015;8(3)
- Yin H, Du J, Li L, Jin C, Fan L, Li M, et al. Comparative genomic analysis reveals multiple long terminal repeats, lineage-specific amplification, and frequent Interelement recombination for *Cassandra* Retrotransposon in pear (*Pyrus bretschneideri* Rehd.). *Genome Biol Evol*. 2014;6(6):1423–36.
- Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, et al. The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res*. 2013;23(2):396–408.
- Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nature Genet*. 2010;42(10):833–9.
- Verde I, Abbott AG, Scalabrini S, Jung S, Shu S, Marroni F, et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genet*. 2013;45(5):487–94.
- Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*. 2012;492(7429):423–7.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009; 326(5956):1112–5.
- Kumar A, Bennetzen JL. Plant retrotransposons. *Annu Rev Genet*. 1999;33(1): 479–532.
- Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics*. 2004;166(3):1437–50.
- Witte C-P, Le QH, Bureau T, Kumar A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *P Natl Acad Sci USA*. 2001;98(24):13778–83.
- Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. Retrotransposons of rice involved in mutations induced by tissue culture. *P Natl Acad Sci USA*. 1996;93(15):7783–8.
- Yin H, Liu J, Xu Y, Liu X, Zhang S, Ma J, et al. *TARE1*, a mutated *Copia*-like LTR retrotransposon followed by recent massive amplification in tomato. *PLoS One*. 2013;8(7):e68587.
- Kim S, Park JY, Yang T-J. Characterization of three active transposable elements recently inserted in three independent *DFR-A* alleles and one high-copy DNA transposon isolated from the pink allele of the *ANS* gene in onion (*Allium cepa* L.). *Mol Gen Genomics*. 2015; 290(3):1027–37.
- Hartl DL, Lohe AR, Lozovskaya ER. Regulation of the transposable element *mariner*. *Genetica*. 1997;100(1):177–84.
- Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics*. 1990;124(2):339–55.
- Roulin A, Piegu B, Wing RA, Panaud O. Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon *RIRE1* within the genus *Oryza*. *Plant J*. 2008;53(6):950–9.
- Roulin A, Piegu B, Fortune PM, Sabot F, D'Hont A, Manicacci D, et al. Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon *Route66* in Poaceae. *BMC Evol Biol*. 2009;9(1):1–10.
- El Baidouri M, Carpentier M-C, Cooke R, Gao D, Lasserre E, Llauro C, et al. Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Res*. 2014;24(5):831–8.
- Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*. 2002;3(5):329–41.
- Paszowski J. Controlled activation of retrotransposition for plant breeding. *Curr Opin Biotech*. 2015;32:200–6.
- Grandbastien M-A. Activation of plant retrotransposons under stress conditions. *Trends Plant Sci*. 1998;3(5):181–7.
- Duan K, Ding X, Zhang Q, Zhu H, Pan A, Huang J. *AtCopeg1*, the unique gene originated from *AtCopia95* retrotransposon family, is sensitive to external hormones and abiotic stresses. *Plant Cell Rep*. 2008;27(6):1065–73.
- Manninen O, Kalendar R, Robinson J, Schulman A. Application of *BARE-1* retrotransposon markers to the mapping of a major resistance gene for net blotch in barley. *Mol Gen Genet*. 2000;264(3):325–34.
- Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*. 2011;472(7341):115–9.
- Ito H, Yoshida T, Tsukahara S, Kawabe A. Evolution of the *ONSEN* retrotransposon family activated upon heat stress in Brassicaceae. *Gene*. 2013;518(2):256–61.
- Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. *Nature*. 2003;421(6919):163–7.
- Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell*. 2012;24(3):1242–55.
- Rocheta M, Carvalho L, Viegas W, Morais-Cecílio L. *Corky*, a *gypsy*-like retrotransposon is differentially transcribed in *Quercus* Suber tissues. *BMC Res Notes*. 2012;5(1):432.
- Felice B, Wilson RR, Argenziano C, Kafantaris I, Conicella C. A transcriptionally active *cop*ia-like retroelement in *Citrus limon*. *Cell Mol Biol Lett*. 2008;14(2):289.
- Kimura Y, Tosa Y, Shimada S, Sogo R, Kusaba M, Sunaga T, et al. *OARE-1*, a *Ty1-copia* retrotransposon in oat activated by abiotic and biotic stresses. *Plant Cell Physiol*. 2001;42(12):1345–54.
- Mhiri C, Morel J-B, Vernhettes S, Casacuberta JM, Lucas H, Grandbastien M-A. The promoter of the tobacco *Tnt1* retrotransposon is induced by wounding and by abiotic stress. *Plant Mol Biol*. 1997;33(2):257–66.
- Jiang S, Cai D, Sun Y, Teng Y. Isolation and characterization of putative functional long terminal repeat retrotransposons in the *Pyrus* genome. *Mob DNA*. 2016;7(1):1.
- Rosaceae KC. In: Kubitzki K. (eds) Flowering Plants · Dicotyledons. The Families and Genera of Vascular Plants. Berlin, Heidelberg: Springer; 2004. p. 343–86.

35. Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, et al. The genome of *Prunus mume*. *Nat Commun*. 2012;3(4):1318.
36. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet*. 2011; 43(2):109–16.
37. Potter D, Eriksson T, Evans RC, Oh S, Smedmark J, Morgan DR, et al. Phylogeny and classification of Rosaceae. *Plant Syst Evol*. 2007;266(1-2):5–43.
38. Yin H, Du J, Wu J, Wei S, Xu Y, Tao S, et al. Genome-wide annotation and comparative analysis of long terminal repeat Retrotransposons between pear species of *P. bretschneideri* and *P. Communis*. *Sci Rep*. 2015;5
39. McCarthy EM, JF MD. LTR\_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*. 2003;19(3):362–7.
40. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8(12):973–82.
41. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. *Nat Genet*. 1998;20(1):43–5.
42. Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, et al. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant J*. 2010;63(4):584–98.
43. Zhao M, Ma J. Co-evolution of plant LTR-retrotransposons and their host genomes. *Protein Cell*. 2013;4(7):493–501.
44. Ma J, Bennetzen JL. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *P Natl Acad Sci USA*. 2006;103(2):383–8.
45. Wessler SR. Plant retrotransposons: turned on by stress. *Curr Biol*. 1996; 6(8):959–61.
46. Hobaika Z, Zargarian L, Boulard Y, Maroun RG, Mauffret O, Femandjian S. Specificity of LTR DNA recognition by a peptide mimicking the HIV-1 integrase  $\alpha 4$  helix. *Nucleic Acids Res*. 2009;37(22):7691–700.
47. Zhou H, Rainey GJ, Wong S-K, Coffin JM. Substrate sequence selection by retroviral integrase. *J Virol*. 2001;75(3):1359–70.
48. Schaack S, Gilbert C, Feschotte C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol*. 2010;25(9):537–46.
49. El Baidouri M, Panaud O. Horizontal transfers and the new model of TE-driven genome evolution in eukaryotes. In: Pontarotti P. (eds) *Evolutionary biology: biodiversity from genotype to phenotype*: Springer Cham. 2015. p. 77–92.
50. Walsh AM, Kortschak RD, Gardner MG, Bertozzi T, Adelson DL. Widespread horizontal transfer of retrotransposons. *P Natl Acad Sci USA*. 2013;110(3):1012–6.
51. Wallau GL, Ortiz MF, Loreto ELS. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome Biol Evol*. 2012;4(8):689–99.
52. Fawcett JA, Innan H. High similarity between distantly related species of a plant SINE family is consistent with a scenario of vertical transmission without horizontal transfers. *Mol Biol Evol*. 2016;33(10):2593–604.
53. Gilbert C, Chateigner A, Ernenwein L, Barbe V, Bézier A, Herniou EA, et al. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. *Nat Commun*. 2014;5
54. Gilbert C, Schaack S, Pace IJ, Brindley PJ, Feschotte C. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature*. 2010;464(7293):1347–50.
55. Won H, Renner SS. Horizontal gene transfer from flowering plants to Gnetum. *P Natl Acad Sci USA*. 2003;100(19):10824–9.
56. Dupuy C, Periquet G, Serbielle C, Bézier A, Louis F, Drezen J-M. Transfer of a chromosomal maverick to endogenous bracovirus in a parasitoid wasp. *Genetica*. 2011;139(4):489–96.
57. Neumann P, Požárková D, Macas J. Highly abundant pea LTR retrotransposon *Ogre* is constitutively transcribed and partially spliced. *Plant Mol Biol*. 2003;53(3):399–410.
58. Gómez E, Schulman AH, Martínez-Izquierdo JA, Vicent CM. Integrase diversity and transcription of the maize retrotransposon Grande. *Genome*. 2006;49(5):558–62.
59. Marcon HS, Domingues DS, Silva JC, Borges RJ, Matioli FF, de Mattos Fontes MR, et al. Transcriptionally active LTR retrotransposons in *Eucalyptus* genus are differentially expressed and insertionally polymorphic. *BMC Plant Biol*. 2015;15(1):198.
60. Huang J, Wang Y, Liu W, Shen X, Fan Q, Jian S, et al. *EARE-1*, a transcriptionally active *Ty1/copia*-like retrotransposon has colonized the genome of *Excoecaria agallocha* through horizontal transfer. *Front Plant Sci*. 2017;8
61. Pietzenk B, Markus C, Gaubert H, Bagwan N, Merotto A, Bucher E, et al. Recurrent evolution of heat-responsiveness in Brassicaceae *COPIA* elements. *Genome Biol*. 2016;17(1):209.
62. McClintock B. The significance of responses of the genome to challenge. Singapore: World Scientific Pub. Co; 1993.
63. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7.
64. Jukes TH, Cantor CR. Evolution of protein molecules. *Mamm Protein Metab*. 1969;3(21):132.
65. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–80.
66. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 2011;28(10):2731–9.
67. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13(9):2178–89.
68. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*. 1997;13(5):555–6.
69. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357–60.
70. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the  $2^{-\Delta\Delta CT}$  method. *Methods*. 2001;25(4):402–8.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

