Molecular Genetics and Genomics
https://doi.org/10.1007/s00438-018-1449-y

METHODS PAPER

Identification of transposons near predicted lncRNA and mRNA pools of *Prunus mume* using an integrative transposable element database constructed from Rosaceae plant genomes

Kaifeng Ma[1] · Qixiang Zhang[1,2] · Tangren Cheng[1] · Jia Wang[1]

Received: 15 September 2017 / Accepted: 17 May 2018

Abstract

This study focused on the construction of a database of transposable elements (TEs) from Rosaceae plants, the third most economically important plant family in temperate regions, and its transcriptomics applications. The evolutionary effects of TEs on gene regulation have been explored, and TE insertions can be the molecular bases of changes in gene structure and function. However, a specific Rosaceae plant TE database (RPTEdb) is lacking. The genomes of several Rosaceae plants have been sequenced, providing the opportunity to mine TE data at a whole-genome level. Therefore, we constructed the RPTEdb, a collective and comprehensive database of 19,596 annotated TEs in the genomes of Rosaceae plants using previously described identification and annotation methods and published genome sequences. The user-friendly web-based database provides access to research tools through hyperlinks, including Browse, TE tree, tools, JBrowse, and search sections, and through the inputting of sequences on the main webpage. Next, we performed one advanced application in which TEs near predicted long non-coding RNA (lncRNA) and mRNA domains within white and red petal-tissue transcriptomes of *Prunus mume* 'Fuban Tiaozhi' were identified, revealing 16 TEs that overlapped or were near 16 differentially expressed lncRNA domains, and 54 TEs that overlapped or were near 54 differentially expressed mRNA domains, and the TEs' possible functions were also discussed. We believe that the RPTEdb will contribute to the understanding of TE roles in the structural, functional and evolutionary dynamics of Rosaceae plant genomes.

Keywords Transposons · Annotation · Database · Rosaceae plants · LncRNA and mRNA · Application

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-018-1449-y) contains supplementary material, which is available to authorized users.

✉ Qixiang Zhang
zqxbjfu@126.com

[1] Beijing Key Laboratory of Ornamental Plants Germplasm Innovation and Molecular Breeding, National Engineering Research Center for Floriculture, Beijing Laboratory of Urban and Rural Ecological Environment, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants of Ministry of Education, School of Landscape Architecture, Beijing Forestry University, Beijing 100083, China

[2] Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forestry University, Beijing 100083, China

Abbreviations

TE	Transposable element
lncRNA	Long non-coding RNA
mRNA	Messenger RNA
LTR	Long terminal repeat
MITE	Miniature inverted repeat transposable element
RPTEdb	Rosaceae plant transposable elements database
HMM	Hidden Markov model
WT	White petal tissues
RT	Red petal tissues
FPKM	Fragments per kilobase of transcript per million fragments

Introduction

Rosaceae is the third most economically important plant family in temperate regions worldwide. The Rosaceae family, with over 100 genera and 3000 species, has been traditionally divided into four subfamilies based on fruit type:

Published online: 26 May 2018

Rosoideae, Prunoideae, Spiraeoideae, and Maloideae (Zhang et al. 2012). The genomes of several Rosaceae plants, including *Malus*×*domestica* (Velasco et al. 2010), *Fragaria vesca* (Shulaev et al. 2011), *Prunus mume* (Zhang et al. 2012), *Prunus persica* (Verde et al. 2013) and *Pyrus bretschneideri* (Wu et al. 2013), have now been sequenced, presenting the opportunity to mine transposable elements (TEs or transposons) at the whole-genome level.

TEs, first discovered in maize by McClintock (1950), are DNA sequences that can move from one location to another and self-replicate in eukaryote genomes. TEs play various important roles, including repressing gene expression or prematurely terminating gene transcription by inserting into gene functional domains (Hirochika 2001), affecting post-transcriptional regulation (Macia et al. 2015) and contributing to gene evolution in higher plants (Jiang et al. 2004; Lai et al. 2005). TEs are associated with the origins of long non-coding RNA (lncRNA) molecules and with the lncRNAs' transcribed regions, altering their functions (Ponting et al. 2009; Kapusta et al. 2013; Xu et al. 2015; Wang et al. 2016a, b). Therefore, research that reveals TE functions and explores their evolution is important for genome sequencing projects and has attracted much attention.

TEs accumulate in many species' genomes, like maize, in which ~ 80% of the genomic sequences are TEs (Huang et al. 2012). More TEs have been identified with the development of next-generation whole-genome sequencing and bioinformatics. The systematic classification of TEs is very important, although there are several different criteria for determining their multi-copy numbers and mutation rates (Wicker et al. 2007; Seberg and Petersen 2009). The most widely accepted approach is based on the mode of transposition, which classifies TEs into either retrotransposons (Class I) or DNA transposons (Class II) (Jurka et al. 2005; Wicker et al. 2007). These two classes can be further subdivided into orders, superfamilies, and families based on structural features. Class I elements, divided into long terminal repeat (LTR) retrotransposons (such as *Ty1*/*Copia, Ty3*/*Gypsy, Bel*/*Pao*, and *Dirs*) and non-LTR retrotransposons (such as long and short interspersed nuclear elements), propagate through an RNA intermediate using a 'copy and paste' mechanism to insert themselves into new locations in the genome (Kashkush et al. 2003; Kashkush and Khasdan 2007). Class II elements, which act through a 'cut and paste' mechanism to integrate the DNA sequence into the host genome, are divided into the terminal inverted repeats (TIRs), miniature inverted repeat transposable elements (MITEs) and *Helitron* orders (Feschotte et al. 2002; Bonchev and Parisod 2013; Lisch 2013).

It is difficult to assemble, identify, and annotate TEs during the analyses of complex sequences. Thus, the identification and annotation of TEs, and the construction of a TE-specific database, are of great importance. To our knowledge, specific databases for TEs within the genomes of silkworm (Xu et al. 2013), mulberry (Ma et al. 2015), cotton (Xu et al. 2017), and dioecious plants (Li et al. 2016) have been established. In our study, using these published Rosaceae plant genomes and our own data, we first identified and annotated TEs and then constructed a specific, collective, comprehensive and user-friendly web-based database, called the Rosaceae plant transposable elements database (RPTEdb). In addition, we performed an advanced application to explore TEs and their possible functions near lncRNA and mRNA domains. This database will contribute to our understanding of the roles of TEs in the structural, functional, and evolutionary dynamics of the genomes of Rosaceae plants.

Materials and methods

Construction and content of the RPTEdb

Whole-genome sequence sources

The genome sequences of five Rosaceae plants, *F. vesca, Malus*×*domestica, P. mume, P. bretschneideri*, and *P. persica*, were downloaded from the websites, as shown in Table 1. Three approaches were used to identify TEs within each genome.

De novo identification of TEs in the five Rosaceae plant genomes

The combined PILER (Edgar and Myers 2005) and RepeatModeler analysis method (http://www.repeatmasker.org/

Table 1 Plant species and source websites for their genomic sequences

Species	URLs	References
Fragaria vesca	https://www.rosaceae.org/organism/Fragaria/vesca	Shulaev et al. (2011)
Malus×*domestica*	https://www.rosaceae.org/organism/Malus/x-domestica	Velasco et al. (2010)
Prunus mume	ftp://ftp.ncbi.nlm.nih.gov/genomes/Prunus_mume/	Zhang et al. (2012)
Pyrus bretschneideri	http://peargenome.njau.edu.cn	Wu et al. (2013)
Prunus persica	https://www.rosaceae.org/species/prunus_persica/genome_v1.0	Verde et al. (2013)

RepeatModeler.html, version 1.0.7) was used to perform de novo identifications of TEs. The analysis process was divided into three stages. The genomes were first compared with themselves using the PALS (http://drive5.com/pals/) algorithm with the default parameters. The genomes were divided into chunks small enough for PALS, because the entire sequences were too large to align. The chunks were aligned to themselves and then each pair of chunks was aligned with each other. The comparison data files were merged for each species' genome. Second, we detected repeats having TE characteristics based on the criterion that the dispersed family had at least three members with a maximum length difference of 5% between any two members. Third, MUSCLE (version 3.7) (Edgar 2004) was used to create a consensus sequence for each family after aligning the identified repeat sequences.

RepeatModeler was used to build, adjust, and classify consistent sequences (Bao and Eddy 2002; Price et al. 2005) and to reduce the redundancy of similar predictions (Benson 1999).

Signature-based identification of TEs in the five species

LTR retrotransposons in each plant genome were identified using LTR_STRUC (McCarthy and McDonald 2003) together with the LTR_FINDER method (Xu and Wang 2007). The default parameters were used, except the option -w 2 in LTR_FINDER. Non-LTR retrotransposons were identified using the MGEScan-non-LTR program based on a hidden Markov model (HMM) algorithm (Rho and Tang 2009; Eddy 2011). *Helitron* elements were detected using HelitronScanner (Xiong et al. 2014) with default parameters based on the local combinational variable (LCV) algorithm (Xiong et al. 2009). MITE elements were detected using MITE-Hunter (Han and Wessler 2010) with default parameters.

Similarity-based identification of TEs in the five species

The similarity-based identification of TEs was performed using RepeatMasker (http://www.repeatmasker.org) with the Repbase database to search whole genomes (Jurka et al. 2005; Bao et al. 2015). Results with a score < 250 or with < 40% target coverage were removed.

Definition of superfamilies and families of predicted TEs

In each genome, a repeat-library, used to compare previously characterized elements, was created by combining the predicted TEs generated by the three approaches. The repeat sequences in the repeat-library were compared with Repbase with RepeatMasker, and the best-hit target TE was used to define the superfamily of the putative TEs (Ma et al. 2015; Li et al. 2016). All of the TEs in each of the five genomes were classified into families according to the 80–80–80 rule described by Wicker et al. (2007).

Construction of TE trees

Each superfamily of TEs in the five genomes was used to construct a TE tree. TEs of a similar length and distribution were subjected to multiple sequence alignment using MUSCLE (v3.8.31). Then, the software TreeBeST (version 1.9.2, bootstrap = 1000) was used to construct a neighbor joining phylogenetic tree (http://treesoft.svn.sourceforge.net/viewvc/treesoft/trunk/treebest/).

TE annotation, verification, and website system construction

The repeat annotation of each species was analyzed using RepeatMasker v4.0.3 (http://www.repeatmasker.org/) with the predicted TEs. The annotated TEs' abundance levels and percentages of coverage length within each genome were calculated. To verify the completeness and accuracy of the putative TE library, we extracted previously identified TEs of each corresponding Rosaceae plant from Repbase database and performed a BLAST algorithm-based alignment (*e*-value < 1e−5).

In addition, the open-source software Linux Ubuntu Sever 12.04, Apache 2, MySQL Server 5.5 and Perl 5.16.3/PHP 5.4.30 were used to construct the RPTEdb. We stored all of the predicted TEs' data and information in MySQL tables and developed common gateway interface (CGI) programs using the Perl, JavaScript, and PHP programming languages. For the manipulation and display of positional relationships between genes and TEs in the five Rosaceae genomes, we used JBrowse, which is an embeddable genome browser built with HTML5 and JavaScript (Skinner et al. 2009).

RNA sequencing and annotation of the identified lncRNAs and mRNAs

Plant materials and RNA sequencing

White petal tissues (WT) and red petal tissues (RT) that were collected from an individual ornamental *P. mume* 'Fuban Tiaozhi' tree (Supplementary Fig. S1) were independently used for total RNA isolation, following the instructions of the RNeasy Plant Mini Kit (Qiagen China, Shanghai, China). The extracted products were monitored on 1% agarose gels and evaluated using a NanoPhotometer spectrophotometer (Implen, CA, USA). The RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

Then, the ribosomal RNA within each sample was removed from 3 μg total RNA using an Epicentre Ribo-zer rRNA Removal Kit (Epicentre, USA). The following steps and the Illumina sequencing, including quality control, were carried out by Novogene Corporation (Beijing, China).

Sequences mapping and transcriptome assembly

The sequencing analyses were based on the high-quality clean data obtained from the raw data, and the reference genome files were downloaded from a website (ftp://ftp.ncbi.nlm.nih.gov/genomes/Prunus_mume/). The sequencing data were aligned to the reference genome using TopHat v2.0 (Kim et al. 2013). HTSeq v0.5.3 (https://bioweb.pasteur.fr/docs/modules/HTSeq/0.5.3p9/) (-m union) was used to count the read numbers mapped to each reference gene. In addition, the fragments per kilobase of transcript per million fragments (FPKM) value of each gene were calculated based on the length of the gene and read counts mapped to this gene (Trapnell et al. 2010). Using Scripture (beta2, default parameters) (Guttman et al. 2010) and Cufflinks (v2.1.1, min-frags-per-transfrag = 0) (Trapnell et al. 2010), mapped reads were assembled.

Prediction of lncRNAs and analyses of the differentially expressed lncRNAs' target genes

The assembled transcript sequences were used for lncRNA predictions, and the basic lncRNA screening was performed based on the following criteria: single-exon transcripts (length ≥ 200 bp) within ≥ 500 bp of other transcripts were selected; expression levels of the assembled sequences were tested by cufflinks (Trapnell et al. 2012), with FPKM ≥ 0.5 for multi-exon transcripts and FPKM ≥ 2 for single-exon transcripts; transcripts similar or identical to non-lncRNAs or non-mRNAs (such as rRNA, tRNA, snRNA, snoRNA, pre-miRNA, and pseudogenes) were not acceptable; these reserved transcripts were compared with known mRNAs and the lncRNAs; and intronic lncRNAs and anti-sense lncRNAs were selected online (http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html#transfrag-class-codes). Next, the coding-potential analyses of the selected lncRNAs were performed using the coding-potential calculator (CPC) (0.9-r2, -*e* value 1e−10) (Kong et al. 2007) and Pfam Scan (v1.3; -E 0.001 -domE 0.001 -pfamB) (Bateman et al. 2002; Punta et al. 2012) profiles with default parameters. Transcripts predicted by coding potential using either or both tools were removed, and those without coding potentials were defined as our candidate lncRNAs.

The FPKM values of lncRNAs in each sample provided the statistical bases for determining differential expression using digital transcripts or gene expression data using a model based on the negative binomial distribution (Trapnell et al. 2010). Q value < 0.05 was set as the threshold for being significantly different. In addition, potential target genes, within 100-kB upstream or downstream of each differentially expressed lncRNA, were predicted using a *cis*-acting algorithm based on the genome annotation and a genome browser (Jia et al. 2010). The functional enrichment analyses of the potential target genes were carried out using gene ontology (GO, http://www.geneontology.org) enrichment (Young et al. 2010) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses (KOBAS 2.0) (Mao et al. 2005).

Annotation of differentially expressed mRNAs

Cufflinks v2.1.1 was used to assemble and identify known and novel transcripts, and the gene expression levels (FPKM) were estimated based on HTSeq v0.5.3.1 (-m union) (Trapnell et al. 2010). A differential expression analysis of WT vs. RT groups was performed in DEGSeq v1.10.1 (Wang et al. 2010). Significantly differentially expressed genes (DEGs) were identified using a Q value cutoff of 0.05. The DEGs were annotated according to the genome annotation and were then categorized under the GO terms and KEGG database.

Distribution of TEs related to the lncRNA or mRNA pool

The BAM files of the transcriptomes were imported into the JBrowse of *P. mume*. The lncRNA-related TEs were identified within 5-kB regions upstream and downstream of the lncRNA pool, and similarly, the mRNA-related TEs were detected within 5-kB regions upstream and downstream of the mRNA pool using an in-house Perl script.

Results

Identification and annotation of TEs in five Rosaceae plants

Using various identification processes, 19,596 TEs belonging to 31 superfamilies and 2471 families were found based on the sequences of the five Rosaceae genomes. In total, 1995 (1563 retrotransposons/432 DNA transposons), 4483 (3437 retrotransposons/1046 DNA transposons), 2633 (2149 retrotransposons/484 DNA transposons), 6941 (5914 retrotransposons/1027 DNA transposons), and 3544 (3042 retrotransposons/502 DNA transposons) TEs were identified in the genomes of *F. vesca, M.* × *domestica, P. mume, P. bretschneideri*, and *P. persica*, respectively. There were more Class I TE members than Class II members in each plant species' genome (Table 2). The comparative analysis

Table 2 Statistical information for transposable elements identified in five Rosaceae plant whole-genome sequencing assemblies

Species			*Fragaria vesca*		*Malus × domestica*		*Pyrus bretschneideri*		*Prunus mume*		*Prunus persica*	
Class	Order	Superfamily	Members	Families	Members	Families	Members	Families	Members	Families	Members	Families
Retrotransposons	LINE	*CR1*	1	1	1	1	1	1	0	0	0	0
		DRE	1	1	0	0	0	0	0	0	0	0
		I	1	1	4	4	5	5	0	0	1	1
		Jockey	0	0	0	0	1	1	0	0	0	0
		L1	93	67	105	60	140	65	59	43	45	31
		L2	0	0	0	0	1	1	0	0	0	0
		Penelope	0	0	0	0	1	1	0	0	0	0
		R1	1	1	0	0	0	0	0	0	0	0
		RTE	0	0	113	1	13	1	0	0	0	0
		RTE-BovB	0	0	0	0	1	1	0	0	0	0
		RTE-RTE	0	0	0	0	2	2	0	0	0	0
		RTE-X	0	0	0	0	1	1	0	0	0	0
	LTR	*Caulimovirus*	2	2	3	1	7	3	5	3	12	1
		Copia	710	15	1226	11	2378	10	706	23	1518	24
		DIRS	6	6	3	3	2	2	11	9	5	3
		ERV1	18	16	13	11	52	22	39	12	49	20
		ERV4	2	2	0	0	0	0	2	2	2	2
		ERVK	5	5	5	5	15	9	17	14	12	8
		ERVL	0	0	0	0	0	0	4	3	2	2
		Gypsy	521	26	1680	8	2780	16	861	26	1035	21
		Ngaro	0	0	0	0	0	0	0	0	4	4
		Pao	4	4	3	3	12	10	12	9	14	10
		Unknown	198	105	281	60	502	83	433	98	343	84
Subtotal			1563	252	3437	168	5914	234	2149	242	3042	211
DNA transposons	TIR	*CMC*	25	16	3	3	7	6	9	9	5	5
		Harbinger	1	1	0	0	0	0	0	0	0	0
		hAT	43	35	14	14	54	44	41	29	32	24
		MULE	29	26	27	26	46	44	31	30	27	27
		P	0	0	0	0	3	3	0	0	0	0
		PIF-Harbinger	27	21	7	7	22	16	22	22	12	12
		TcMar	2	2	0	0	5	5	0	0	1	1
		Unknown	25	21	1	1	36	30	10	9	8	7
	Helitron	*Helitron*	136	42	869	56	725	71	276	66	307	70
	MITE	*MITE*	144	135	125	114	129	119	95	92	110	103

Table 2 (continued)

Species			*Fragaria vesca*		*Malus × domestica*		*Pyrus bretschneideri*		*Prunus mume*		*Prunus persica*	
Class	Order	Superfamily	Members	Families	Members	Families	Members	Families	Members	Families	Members	Families
Subtotal			432	299	1046	221	1027	338	484	257	502	249
Total			1995	551	4483	389	6941	572	2633	499	3544	460

of the sequences revealed that 94.84–99.39% of the Repbase-TEs were mapped to the putative TEs within each Rosaceae genome, and the identity ratios > 80% were within the 93.50–98.77% range. Because the classification of Repbase-TEs did not provide enough detail, the matching rates of superfamily and order were moderately low (< 60%) within the *F. vesca* and *M.* × *domestica* genomes. The verification of reliability and integrity demonstrated that most published Repbase-TEs were covered by our putative TE library, which contained a more detailed classification (Supplementary Table S1).

All of these putative TEs were annotated and covered ~ 36.6, 40.0, 61.7, 43.8, and 49.6% of the *F. vesca, M.* × *domestica, P. bretschneideri, P. mume*, and *P. persica* genomes, respectively (Fig. 1a). The percentage of LTR retrotransposon lengths within each of the five Rosaceae genomes (28.25%, *F. vesca*; 30.98%, *M.* × *domestica*; 51.00%, *P. bretschneideri*; 37.92%, *P. mume*; and 43.26%, *P. persica*) was the greatest ($F = 88.91$, $df = 4$, $P < 0.001$; each of the percentage datum underwent square root arcsine transformation) (Fig. 1b).

Description of the RPTEdb

Interface and functional description

RPTEdb, an easy-to-use platform, was built using the 19,596 identified TEs of Rosaceae genomes, and researchers can browse, search, and download TE information by clicking on the webpage's hyperlinks. Tools are also supplied for sequence analyses. The welcome page (Fig. 2a) displays a summary of the database, and the home page is organized with a header (Fig. 2b) and a left-side column (Fig. 2c) for navigation, making access to the TE data efficient and user-friendly. The header contains links to five major sections: Browse, TE tree, tools, JBrowse, and search. The left-side column additionally contains systematics and link sections.

Browse

We have provided a browsing interface to satisfy the user's requirement that the interface reveals detailed information for each superfamily and TE sequence. The 19,596 identified TEs, allocated into 2471 families, can be shown on the interface by clicking Browse hyperlinks. Each of the TE sequences can be freely downloaded (Fig. 2d).

TE tree

In total, 19,503 TEs were used for the phylogenetic analysis of TE families, and 14 TE trees were constructed. RPTEdb also offers a service system to view or download all of the

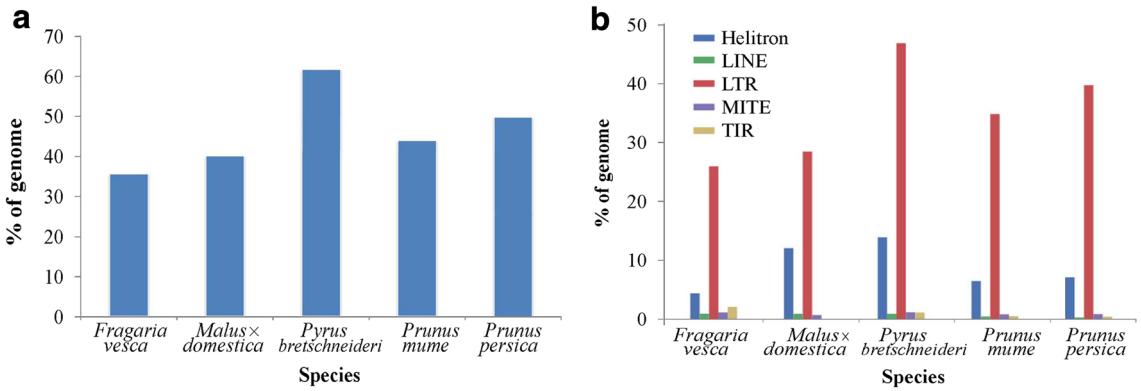


Fig. 1 Annotated transposable elements (TEs) (%) in several species' genomes. **a** Annotated TEs as percentages of each plant species' genome; **b** types of TEs as percentages of each plant species' genome. *LINE* long interspersed nuclear element, *LTR* long terminal repeat, *MITE* miniature inverted repeat transposable element, *TIR* terminal inverted repeat. (Color figure online)

TE trees, which show the diverse evolutionary relationships among different TE families.

Tools

BLAST, HMM, GetORF, and Cut Sequence are included in the tool hyperlinks and embedded in RPTEdb to help users mine, analyze, and visualize the TE information.

BLAST: this section allows homologous comparison analyses between query sequences and the RPTEdb website by inputting specific sequences or profiles in the .txt format (Fig. 3a).

HMM: after entering specific protein sequences, the coding domains collected from the previous studies are used to identify and classify the specific sequences with HMM profiles of LTR and non-LTR retrotransposon (Fig. 3b).

GetORF: the potential ORFs of query sequences can be found using this program according to the parameters set by the user. The ORFs obtained by GetORF can be used as queries to search against the HMM profiles using the HMM package and to classify the sequences into corresponding superfamilies (Fig. 3c).

Cut Sequence: a sequence or sequences in a position, such as TE_start and TE_end, can be defined by the user as needed.

JBrowse

JBrowse, a genomic visualization tool, was created to display TE densities, coordinates between TEs and nearby genes (Fig. 4a). By clicking the zoom button, an enlarged scale in a graphic interface is displayed, and detailed information for TEs and genes can be visualized conveniently (Fig. 4b). The primary data, attributes, and sequence region of an identified TE can be shown in a rectangle box by clicking the TE's ID label (Fig. 4c).

Search

In this section, users can consult or locate specific TEs using their ID or family. For ID searching, users should enter the single or multiple sequence ID numbers (separated by ',', ';' or '-') without spacing. Users can also obtain the search information in a tabular format. In addition, the information can be acquired by clicking the corresponding download hyperlinks.

Information and links

Clicking on the 'Information' button takes users to a webpage containing a list of the main species of Rosaceae. The left-side column contains many links connecting mainstream TE databases and software websites.

Genome-wide identification and annotation of differentially expressed lncRNAs and mRNAs

In total, approximately 15.09 GB (WT) and 14.51 GB (RT) of clean bases were sequenced (Supplementary Table S2), and the mapping rates of clean reads were 71.61 and 76.01%, respectively (Supplementary Table S3). Non-coding RNAs represented 0.6 and 0.7% of the clean reads from the WT and RT RNA-seq data, respectively (Supplementary Fig. S2). Using prediction and screening software, 16,320 lncRNAs and 28,450 mRNAs were identified (Fig. 5a, Supplementary Data S1–S2). At last, 52 differentially expressed lncRNAs (Supplementary Data S3) and 550 differentially expressed mRNAs (Supplementary Data S4), respectively, were detected (Fig. 5b).

We also performed GO term and KEGG pathway enrichments of the differentially expressed lncRNAs' *cis*-acting target genes and DEGs. In total, 683 target genes were predicted and enriched into 46 GO terms ($P < 0.05$) including

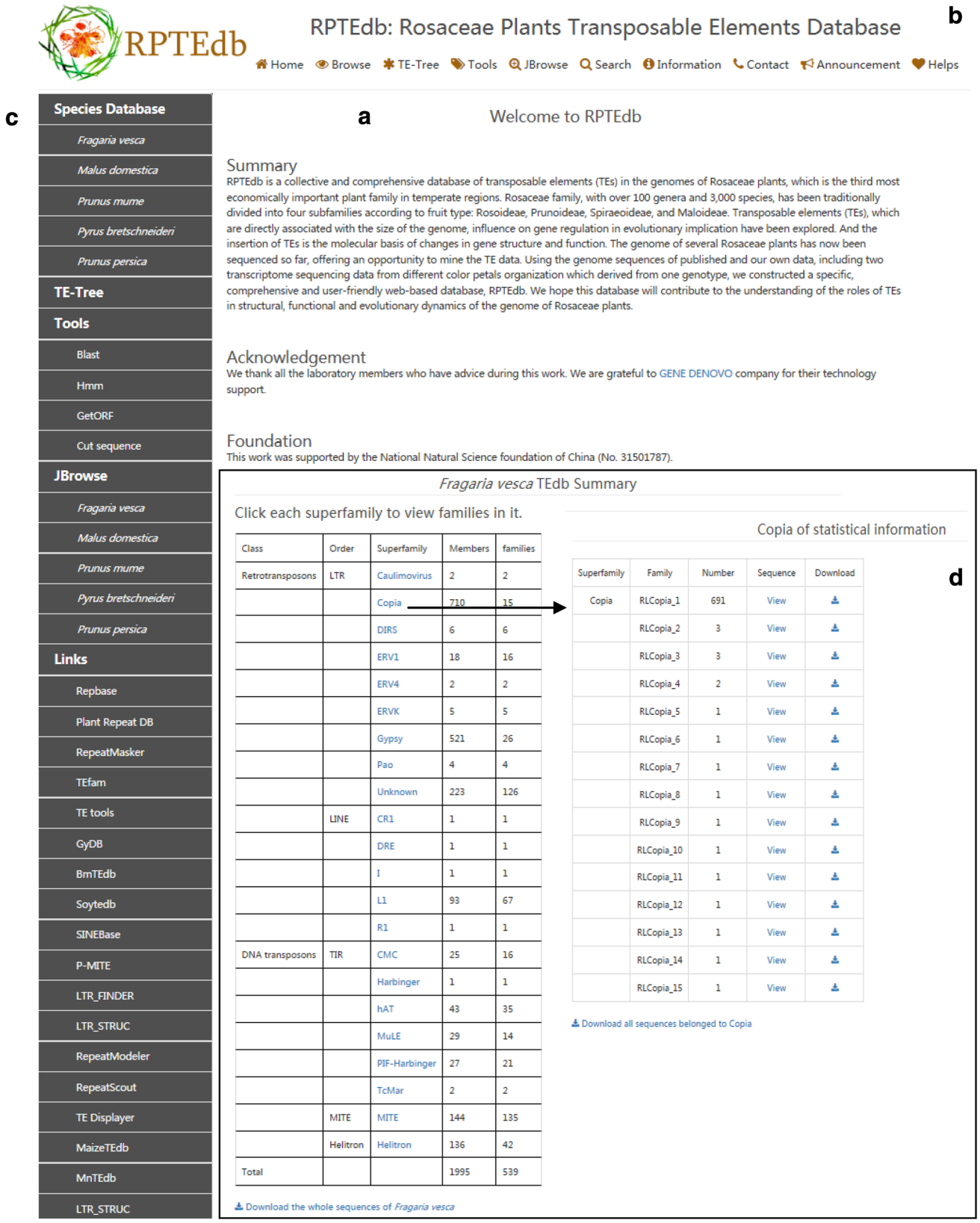


Fig. 2 Home page and browse section of the RPTEdb. **a** Welcome screen and summary of the RPTEdb; **b** the header of the RPTEdb; **c** left-side column of the RPTEdb; **d** browse section of the RPTEdb. (Color figure online)

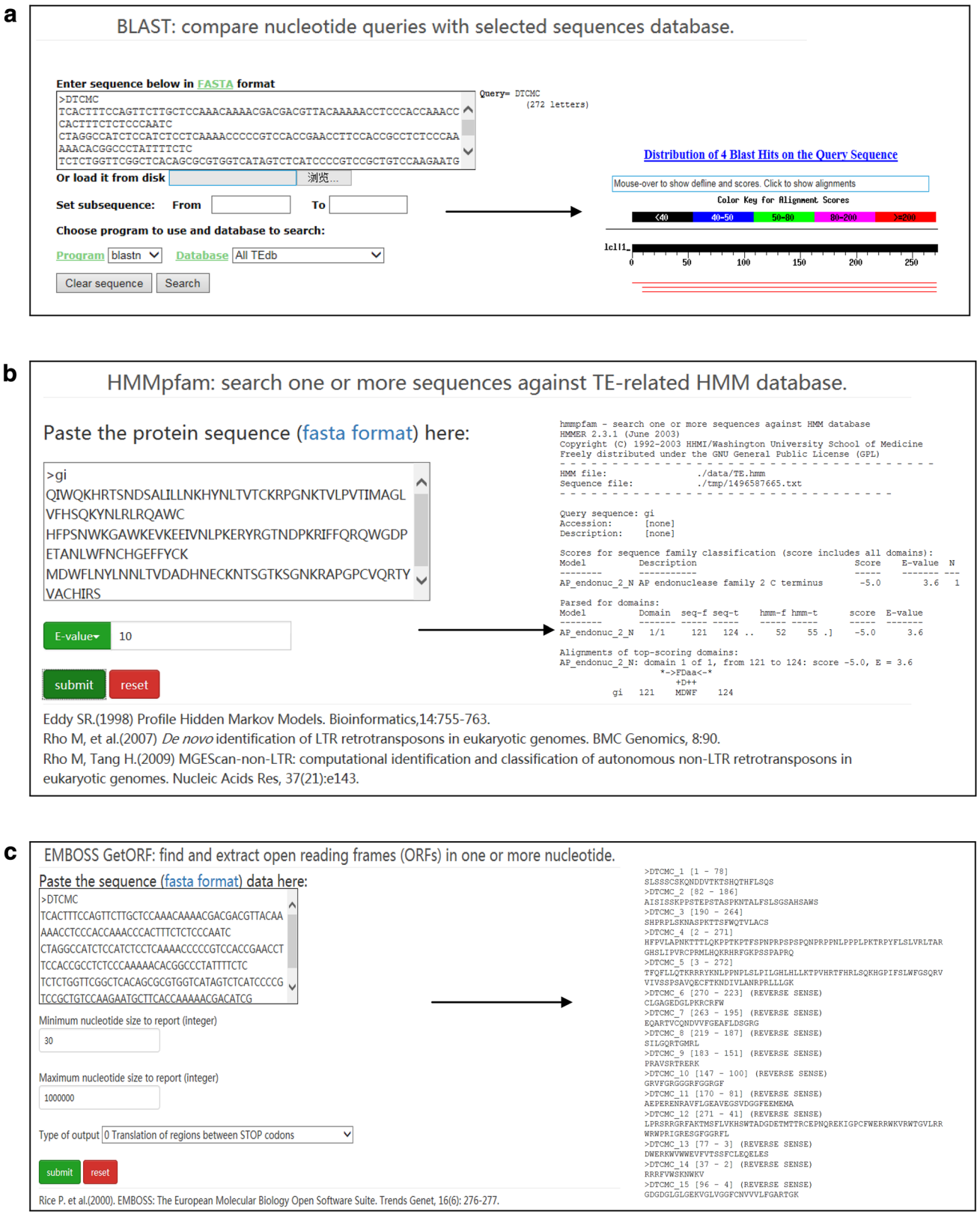


Fig. 3 Functions of tools provided in the RPTEdb. **a** Sample BLAST search; **b** sample HMM search; **c** sample GetORF search. (Color figure online)

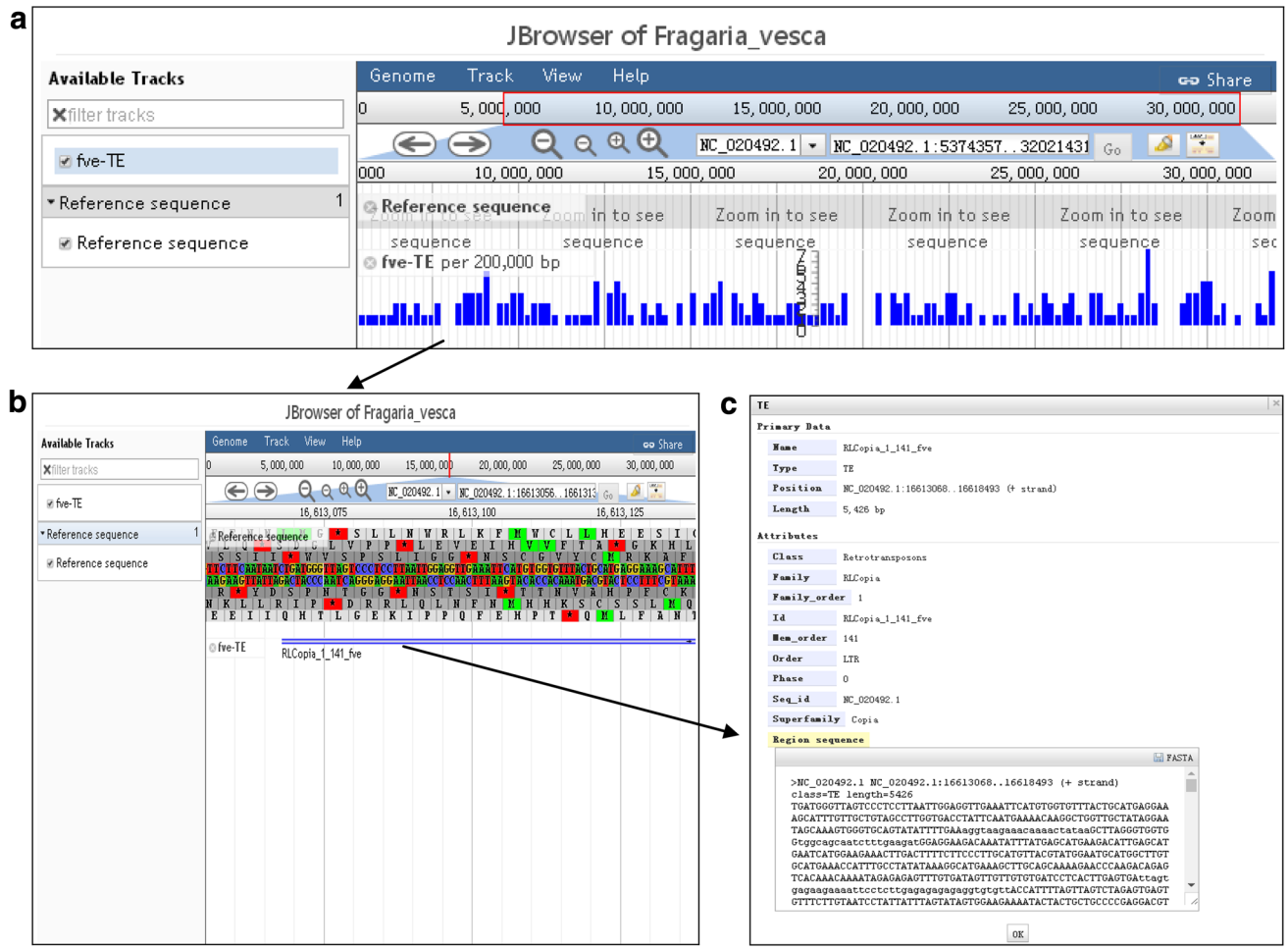


Fig. 4 Demonstration of the JBrowse section of the RPTEdb using *Fragaria vesca*. **a** Genome sequence and displayable fve-TE sites of *F. vesca*. The ruler label in the middle of the web interface indicates the TE number within the range of a certain sequence length; **b** sequence and locational information for RLCopia_1_141_fve were visualized by enlarging the specific region; and **c** detailed information, including the primary data, attributes and sequence region of the identified TE, is shown in a rectangular box when users click the TE's ID label. (Color figure online)

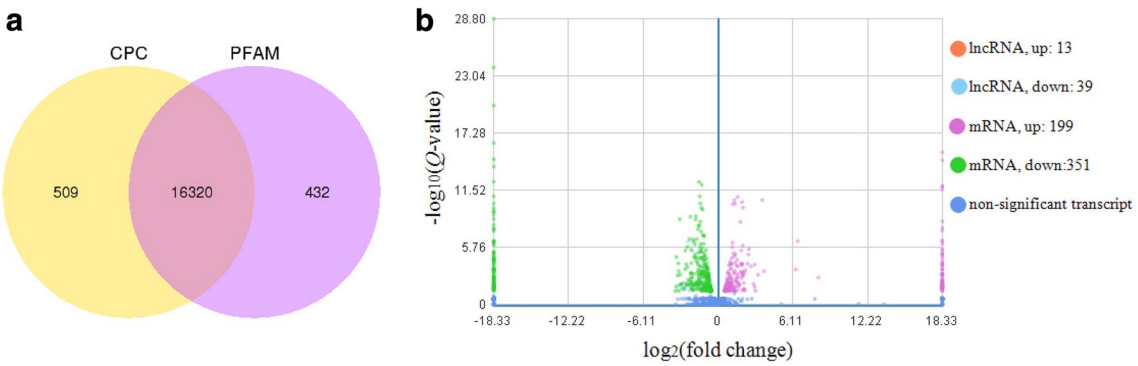


Fig. 5 Total identified and differentially expressed lncRNAs and mRNAs within the white (WT) and red (RT) petal tissue samples. **a** LncRNAs identified with both coding-potential calculator (CPC; left) (0.9-r2) and Pfam Scan v1.3 (PFAM; right) profiles. Each circle represents the number of non-coding transcripts predicted using the relevant software's profile. The overlapping section includes the number of non-coding transcripts predicted by both methods; **b** numbers of differentially expressed lncRNAs and mRNAs. The horizontal axis represents the change in transcript expression levels in different samples (WT vs. RT), and the vertical axis represents the statistical significance. (Color figure online)

15 biological processes, 3 cellular components, and 28 molecular functions. A significant number of the GO terms were associated with rRNA binding, chromatin binding, and UDP-glycosyltransferase activity (Supplementary Data S5). The target genes were also subjected to a KEGG pathway enrichment, although the effective pathway obtained was not involved in flower color regulation (Supplementary Data S6). Similarly, the 550 DEGs were enriched in 125 GO terms ($P < 0.05$), including carotenoid metabolic and biosynthetic processes, oligosaccharyltransferase complex and binding. However, the DEGs were enriched in ineffective KEGG pathway involved in color formation (Supplementary Data S7 – S8).

TEs near the differentially expressed lncRNA and mRNA domains

Within the WT and RT transcriptomes of *P. mume* 'Fuban Tiaozhi', 2958 out of 16,320 lncRNAs were related to 1553 TEs, and 16 TEs were identified as being near or overlapping 16 differentially expressed lncRNAs (15 downregulated; 1 up-regulated) based on the annotational information in RPTEdb and the in-house Perl script. Of these 16 lncRNAs, 11 overlapped 10 TEs, 3 were upstream of 2 TEs, and 3 were downstream of 3 TEs. Another 2942 non-differentially expressed lncRNAs were detected near TEs (Table 3, Supplementary Data S9). In addition, 4244 mRNAs, including 54 differentially and 4190 non-differentially expressed transcripts, respectively, were detected next to 1827 TEs. Of these DEGs, 30 down- and 24 up-regulated genes overlapped or were near 54 TEs (Table 3, Supplementary Data S10). More differentially expressed mRNAs were found than differentially expressed lncRNAs. However, the ratio of the differentially expressed mRNAs near TEs to differentially expressed mRNAs was much less than the ratio of differentially expressed lncRNAs near TEs to differentially expressed lncRNAs (Table 3).

Furthermore, we imported the BAM files of the two transcriptomes into the JBrowse of *P. mume* that helps the users to realize the densities of TEs and RNAs, as well as the numbers of clean reads of WT and RT transcriptomes underneath the gene models (Fig. 6a). By enlarging the window or inputting the starting and ending positions of the gene sequence, a view of the positional relationships between TEs and RNAs, as well as the differential expression between WT and RT samples, can be observed (Fig. 6b).

A functional annotation analysis showed that the target genes of differentially expressed lncRNAs (dependent of their positional relationships to TEs) were annotated as *MYB, bHLH*, and *WD* transcription factor genes or played roles in UDP-glycosyltransferase activity (Supplementary Data S3). We also detected that the annotated DEGs (dependent of their positional relationships to TEs) took an active role in coding zeaxanthin epoxidase, transcription factors (MYB, bHLH, and WD), and UDP-glycosyltransferase (Supplementary Data S4). This suggested that these genes may play important roles in anthocyanins or carotenoids regulatory pathways, resulting in the variety of petal colors in *P. mume* 'Fuban Tiaozhi.'

Discussion

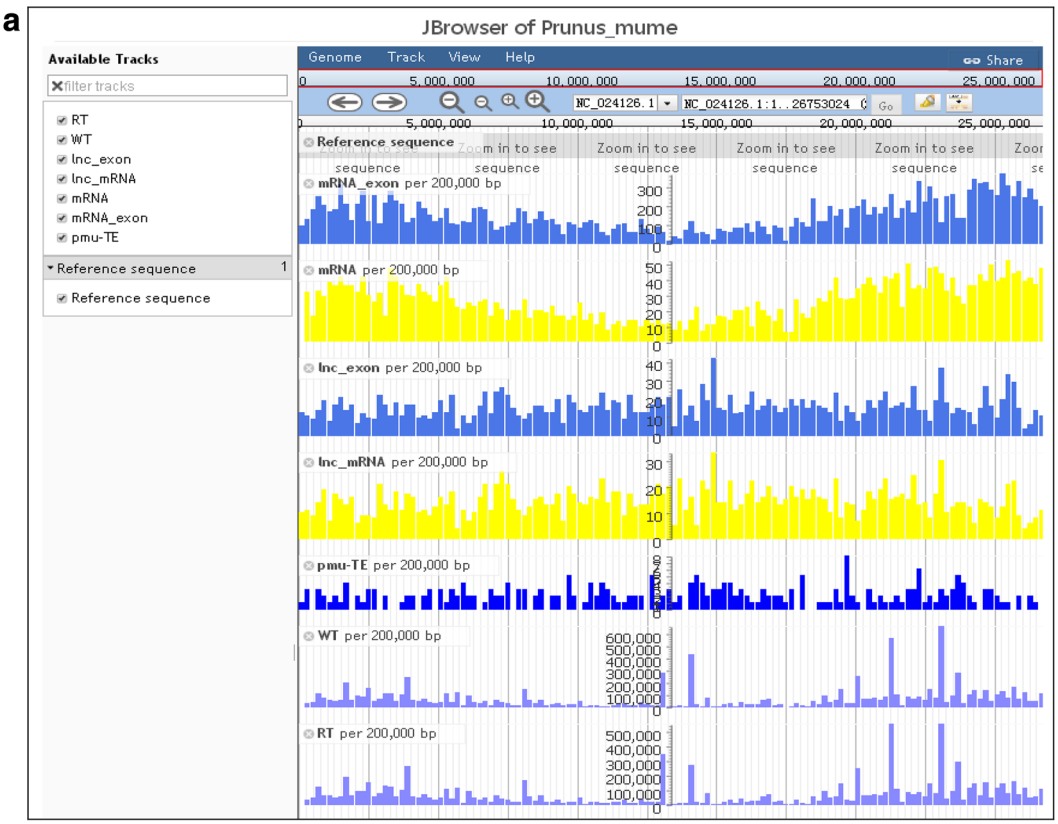
An exponential expansion of bioinformatics is in progress, with massive amounts of data being generated every year. The collection, classification, analysis, and reorganization of the data are necessary to uncover new phenomena and rules. Database construction is one approach to analyze genomes (Dong et al. 2004; Duvick et al. 2007), proteomes (Punta et al. 2012; Komatsu et al. 2017), transcriptomes (Cheng and Stromvik 2008), and TEs (Kronmiller and Wise 2008; Ma et al. 2015; Li et al. 2016; Xu et al. 2017) in plants. TEs account for a large proportion of genomes, as indicated by their ~ 80% coverage of the maize genome (Huang et al. 2012). In the sequenced Rosaceae plants, we found that annotated TE sequences accounted for 36.6–61.7% of the *F. vesca, M.* × *domestica, P. mume, P. bretschneideri*, and *P. persica* genomes. Thus, it is advantageous to construct one database of Rosaceae plants for TE browsing, classification, visualization, and searching.

TEnest, founded on the PlantGDB site (http://www.plantgdb.org/) under the tools section, is one database for the

Table 3 Annotated transposable elements (TEs) identified near the lncRNA or mRNA regions

Number or ratio	Total RNAs/TEs annotated RNAs	Differently expressed RNAs/TEs annotated differently expressed RNAs	Non-differently expressed RNAs/TEs annotated non-differently expressed RNAs
Number of lncRNA	16,320/2958	52/16	16,268/2942
Ratio[a]	0.181	0.308	0.181
Number of mRNA	28,450/4244	550/54	27,900/4190
Ratio[a]	0.149	0.098	0.150

[a]The ratio represents the number of RNAs near annotated TEs to the number of total RNAs or differentially expressed RNAs


a
JBrowser of Prunus_mume
Available Tracks
filter tracks
RT
WT
lnc_exon
lnc_mRNA
mRNA
mRNA_exon
pmu-TE
Reference sequence
Reference sequence
Genome
Track
View
Help
Share
mRNA_exon per 200,000 bp
mRNA per 200,000 bp
lnc_exon per 200,000 bp
lnc_mRNA per 200,000 bp
pmu-TE per 200,000 bp
WT per 200,000 bp
RT per 200,000 bp


b
JBrowser of Prunus_mume
Available Tracks
filter tracks
RT
WT
lnc_exon
lnc_mRNA
mRNA
mRNA_exon
pmu-TE
Reference sequence
Reference sequence
Genome
Track
View
Help
Share
Zoom in to see sequence
mRNA_exon
mRNA
XM_008235106.1
XM_008235005.1
lnc_exon
TCONS_00037555_1
TCONS_00062901_1
TCONS_00044795_1
TCONS_00062906_1
lnc_mRNA
TCONS_00037555
TCONS_00062901
TCONS_00044795
TCONS_00062906
WT per 500 bp
pmu-TE
RLGypsy_1_173_pmu
RT per 500 bp

◂**Fig. 6** Detection of positional relationships between the transposable elements (TEs) and RNAs. **a** Landscape of the TE and RNA (lncRNA and mRNA) densities, and clean read numbers of white (WT) and red (RT) petal tissues within the NC_024126.1 linkage group; **b** demonstration of the positional relationships between TE (RLGypsy_1_173_pmu, NC_024126.1 and NC_024126.1:14860144..14871660) and lncRNAs (TCONS_00037555 and TCONS_00044795). Clean read numbers of WT and RT are also displayed by the ruler label or by alignment sequences mapped to the reference genome when the field of view is suitable and large enough. (Color figure online)

identification of TEs (Kronmiller and Wise 2008). It provides full-length repeats with annotations and foci on evolutionary research using identified LTR retrotransposons in common model plants. Recently, species-specific databases were established for silkworm, mulberry, cotton, and other plants (Xu et al. 2013, 2017; Ma et al. 2015; Li et al. 2016). We constructed the classification detailed RPTEdb, which is open to the public, using 19,596 TEs belonging to 31 superfamilies and 2471 families annotated in the genomes of the five Rosaceae plants, to classify TEs and investigate their evolutionary relationships and regulatory functions. The different TE families showed diverse relationships among the five genomes. TE sequences, not only single TEs but multi-TEs, can be obtained conveniently, and sequence alignment analyses described in recent studies (Ma et al. 2015; Li et al. 2016) are offered in RPTEdb. In addition, newly identified insertions and TEs will be uploaded into RPTEdb for scaffold-type genomes, and other Rosaceae plants and different identification methods will also be listed when they are available.

Evidence has demonstrated that TEs are associated with lncRNAs (Kapusta et al. 2013; Sytnikova et al. 2014; Xu et al. 2015), and the characteristics of TEs allow their activation or transposition to result in differential gene expression, producing phenotypic variations (McClintock 1950; Feschotte et al. 2002). Therefore, we explored the annotated TEs adjacent to lncRNA and mRNA pools within transcriptomes of WT and RT to display the application of RPTEdb in comparative genomics research and investigated the TEs' possible functions and features in regulating lncRNAs and mRNAs. The lncRNAs are widely distributed and identified in the genomes of eukaryotes, and they regulate gene expression levels and biological processes (Lv et al. 2015; Wang et al. 2016a, b). Theory, which has been partly proven in the major force shaping the lncRNA repertoire of vertebrates (Kelley and Rinn 2012; Kapusta et al. 2013; Hutchins and Pei 2015; Kannan et al. 2015) and model plants (Wang et al. 2016a, b), suggests that TEs contribute to the origin of lncRNA (Ponting et al. 2009; Johnson and Guigó 2014) and that TEs may act as functional domains, which participate in the regulation of lncRNAs in humans and plants. We obtained 52 differentially expressed lncRNAs, 16 out of which were adjacent to 16 TEs, which had *cis*-target genes that were annotated as *MYB, bHLH* and *WD* genes, or *UDP-glycosyltransferase* genes. It suggested that the differentially expressed lncRNAs themselves and/or the TEs act with the lncRNAs, which leads to the regulation of the chimeric coloration of flowers.

Moreover, because the insertion and excision of the TEs have given rise to color variations in maize kernels (McClintock 1950), sorghum grain pericarp (Chopra et al. 1999), and *Dianthus caryophyllus* flowers (Nishizaki et al. 2011), we also detected that 54 out of 550 differentially expressed mRNAs adjacent to the TEs when comparing WT and RT transcriptomes of *P. mume* 'Fuban Tiaozhi' on the basis of the RPTEdb. It showed that the mRNA distribution density near the TEs appeared to be much greater than the lncRNA distribution density near the TEs. However, the ratio of the differentially expressed mRNAs near TEs to differentially expressed mRNAs was much lower than the ratio of differentially expressed lncRNAs near TEs to differentially expressed lncRNAs, which indicates a close relationship between TEs and lncRNAs (Ponting et al. 2009). Through a gene functional annotation, we found that these genes encode zeaxanthin epoxidase, transcription factors (MYB, bHLH, and WD) and UDP-glycosyltransferase. We assumed that the mRNAs themselves, as well as TEs near the mRNAs, can lead to color variation. Hence, further verification studies that determine whether lncRNAs and/or mRNAs themselves or the TEs adjacent to the molecular sequences play primary functions in the chimeric petal color patterns is required.

In conclusion, we constructed the RPTEdb based on 19,596 annotated TEs, providing more detailed information than Repbase-TEs, and identified them using three approaches. The number of displayed retrotransposons was greater than the number of DNA transposons within the five Rosaceae genomes. The reliable RPTEdb is an easy-to-use website that provides research support through hyperlinks, including Browse, TE tree, tools, JBrowse, and search sections. The users can browse TE data sets of each species, and identify their TEs by inputting a gene (or protein) sequence or ID. Using the annotation information, we also performed an advanced application to identify TEs adjacent to lncRNA and mRNA domains viewed visually based on JBrowse of *P. mume*, and discussed the positional relationships between TEs and the differentially expressed lncRNAs and mRNAs from the transcriptomes of WT and RT samples, although their functions in chimeric flower color formation still need to be explored. At last, the web-based RPTEdb will facilitate the identification, intra- and inter-specific comparisons, evolutionary relationships, and regulatory functions of TEs at the whole-genome level.

Database URL http://genedenovoweb.ticp.net:81/RPTEdb/.

Accessibility The transcriptome sequencing data (accession number: CRA000803; http://bigd.big.ac.cn/gsa/s/Py2vcQfC) are available in the database of Genome Sequence Archive (GSA, http://bigd.big.ac.cn/).

Acknowledgements This study was funded by the National Natural Science Foundation of China (No. 31501787), the Fundamental Research Funds for the Central Universities (No. 2016ZCQ02), and the Special Fund for Beijing Common Construction Project.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Bao ZR, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12(8):1269–1276

Bao WD, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6(1):11

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL (2002) The Pfam protein families database. Nucleic Acids Res 30(1):276–280

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580

Bonchev G, Parisod C (2013) Transposable elements and microevolutionary changes in natural populations. Mol Ecol Resour 13(5):765–775

Cheng KC, Stromvik MV (2008) SoyXpress: A database for exploring the soybean transcriptome. BMC Genom 9(1):368

Chopra S, Brendel V, Zhang JB, Axtell JD, Peterson T (1999) Molecular characterization of a mutable pigmentation phenotype and isolation of the first active transposable element from sorghum bicolor. Proc Natl Acad Sci USA 96(26):15330–15335

Dong Q, Schlueter SD, Brendel V (2004) PlantGDB, plant genome database and analysis tools. Nucleic Acids Res 32(90001):354–359

Duvick J, Fu A, Muppirala UK, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2007) PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res 36(Database issue):959–965

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7(10):e1002195

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797

Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21(Suppl 1):152–158

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3(5):329–341

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson JT, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28(5):503–510

Han YJ, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38(22):e199

Hirochika H (2001) Contribution of the *Tos17* retrotransposon to rice functional genomics. Curr Opin Plant Biol 4(2):118–122

Huang CR, Burns KH, Boeke JD (2012) Active transposition in genomes. Annu Rev Genet 46(1):651–675

Hutchins AP, Pei DQ (2015) Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. Chin Sci Bull 60(20):1722–1733

Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. RNA 16(8):1478–1487

Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431(7008):569–573

Johnson R, Guigó R (2014) The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA 20(7):959–976

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

Kannan S, Chernikova D, Rogozin IB, Poliakov E, Managadze D, Koonin EV, Milanesi L (2015) Transposable element insertions in long intergenic non-coding RNA genes. Front Bioeng Biotechnol 3:71

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo XY, Ramsay L, Bourque G, Yandell M, Feschotte C (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9(4):e1003470

Kashkush K, Khasdan V (2007) Large-scale survey of cytosine methylation of retrotransposons and the impact of readout transcription from long terminal repeats on expression of adjacent rice genes. Genetics 177(4):1975–1985

Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33(1):102–106

Kelley D, Rinn J (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol 13(11):R107

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36

Komatsu S, Wang X, Yin XJ, Nanjo Y, Ohyanagi H, Sakata K (2017) Integration of gel-based and gel-free proteomic data for functional analysis of proteins through Soybean Proteome Database. J Proteom 163:52–66

Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei LP, Gao G (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35(Web Server issue):345–349

Kronmiller BA, Wise RP (2008) TEnest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146(1):45–59

Lai JS, Li YB, Messing J, Dooner HK (2005) Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA 102(25):9068–9073

Li SF, Zhang GJ, Zhang XJ, Yuan JH, Deng CL, Gu LF, Gao WJ (2016) DPTEdb, an integrative database of transposable elements in dioecious plants. Database 2016:baw078

Lisch D (2013) How important are transposons for plant evolution. Nat Rev Genet 14(1):49–61

Lv J, Liu H, Yu SH, Liu HB, Cui W, Gao Y, Zheng T, Qin G, Guo J, Zeng TB, Han ZB, Zhang Y, Wu Q (2015) Identification of 4438 novel lincRNAs involved in mouse pre-implantation embryonic development. Mol Genet Genom 290(2):685–697

Ma B, Li T, Xiang ZH, He NJ (2015) MnTEdb, a collective resource for mulberry transposable elements. Database 2015:bav004

Macia A, Blanco-Jimenez E, García-Pérez JL (2015) Retrotransposons in pluripotent cells: impact and new roles in cellular plasticity. BBA-Gene Regul Mech 1849(4):417–426

Mao XZ, Cai T, Olyarchuk JG, Wei Liping (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21(19):3787–3793

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19(3):362–367

McClintock B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA 36(6):344–355

Nishizaki Y, Matsuba Y, Okamoto E, Okamura M, Ozeki Y, Sasaki N (2011) Structure of the acyl-glucose-dependent anthocyanin 5-*O*-glucosyltransferase gene in carnations and its disruption by transposable elements in some varieties. Mol Genet Genom 286(5–6):383–394

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136(4):629–641

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):351–358

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang NZ, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40(1):290–301

Rho M, Tang HX (2009) MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Res 37(21):e143

Seberg O, Petersen G (2009) A unified classification system for eukaryotic transposable elements should reflect their phylogeny. Nat Rev Genet 10(4):276

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton JM, Rees DJG, Williams KP, Holt SH, Rojas JJR, Chatterjee M, Liu B, Silva H, Meisel L, Adato A, Filichkin SA, Troggio M, Viola R, Ashman TL, Wang H, Dharmawardhana P, Elser J, Raja R, Priest HD, Bryant DW, Fox SE, Givan SA, Wilhelm LJ, Naithani S, Christoffels A, Salama DY, Carter J, Girona EL, Zdepski A, Wang WQ, Kerstetter RA, Schwab W, Korban SS, Davik J, Monfort A, Denoyes-Rothan B, Arus P, Mittler R, Flinn B, Aharoni A, Bennetzen JL, Salzberg SL, Dickerman AW, Velasco R, Borodovsky M, Veilleux RE, Folta KM (2011) The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43(2):109–116

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: A next-generation genome browser. Genome Res 19(9):1630–1638

Sytnikova YA, Rahman R, Chirn G, Clark JP, Lau NC (2014) Transposable element dynamics and PIWI regulation impacts lncRNA and gene expression diversity in *Drosophila* ovarian cell cultures. Genome Res 24(12):1977–1990

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578

Velasco R, Zharkikh A, Affourtit JP, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury JS, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira CD, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagné D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouzé P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick MH, Egholm M, van de Peer Y, Salamini F, Viola R (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet 42(10):833–839

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, Zuccolo A, Rossini L, Jenkins J, Vendramin E, Meisel LA, Decroocq V, Sosinski B, Prochnik S, Mitros T, Policriti A, Cipriani G, Dondini L, Ficklin S, Goodstein DM, Xuan PF, Fabbro CD, Aramini V, Copetti D, Gonzalez S, Horner DS, Falchi R, Lucas S, Mica E, Maldonado J, Lazzari B, Bielenberg DG, Pirona R, Miculan M, Barakat A, Testolin R, Stella A, Tartarini S, Tonutti P, Arús P, Orellana A, Wells CE, Main D, Vizzotto G, Silva H, Salamini F, Schmutz J, Morgante M, Rokhsar DS (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45(5):487–494

Wang LK, Feng ZX, Wang X, Wang XW, Zhang XG (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26(1):136–138

Wang X, Ai G, Zhang CL, Cui L, Wang JF, Li HX, Zhang JH, Ye ZB (2016a) Expression and diversification analysis reveals transposable elements play important roles in the origin of *Lycopersicon*-specific lncRNAs in tomato. New Phytol 209(4):1442–1455

Wang ZX, Schwacke R, Kunze R (2016b) DNA damage-induced transcription of transposable elements and long non-coding RNAs in *Arabidopsis* is rare and ATM-dependent. Mol Plant 9(8):1142–1155

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell AJ, Leroy P, Morgante M, Panaud O, Paux E, Sanmiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8(12):973–982

Wu J, Wang ZW, Shi ZB, Zhang S, Ming R, Zhu SL, Khan MA, Tao ST, Korban SS, Wang H, Chen NJ, Nishio T, Xu X, Cong L, Qi KJ, Huang XS, Wang YT, Zhao X, Wu JY, Deng C, Gou CY, Zhou WL, Yin H, Qin GH, Sha YH, Tao Y, Chen H, Yang YN, Song Y, Zhan DL, Wang J, Li LT, Dai MS, Gu C, Wang YZ, Shi DH, Wang XW, Zhang HP, Zeng L, Zheng DM, Wang CL, Chen MS, Wang GB, Xie L, Sovero V, Sha SF, Huang WJ, Zhang SJ, Zhang MY, Sun JM, Xu LL, Li Y, Liu X, Li QS, Shen JH, Wang JY, Paull RE, Bennetzen JL, Wang J, Zhang SL (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res 23(2):396–408

Xiong WW, Li TH, Chen K, Tang KL (2009) Local combinational variables: an approach used in DNA-binding helix-turn-helix motif prediction with sequence information. Nucleic Acids Res 37(17):5632–5640

Xiong WW, He LM, Lai JS, Dooner HK, Du CG (2014) HelitronScanner uncovers a large overlooked cache of *Helitron*

transposons in many plant genomes. Proc Natl Acad Sci USA 111(28):10263–10268

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35(Suppl 2):265–268

Xu HE, Zhang HH, Xia T, Han MJ, Shen YH, Zhang Z (2013) BmTEdb: a collective database of transposable elements in the silkworm genome. Database 2013:bat055

Xu HY, Nelson ADL, Shippen DE (2015) A transposable element within the non-canonical telomerase RNA of *Arabidopsis thaliana* modulates telomerase in response to DNA damage. PLoS Genet 11(6):e1005281

Xu ZZ, Liu J, Ni WC, Peng Z, Guo Y, Ye WW, Huang F, Zhang XG, Xu P, Guo Q, Shen XL, Du JC (2017) GrTEdb: the first web-based database of transposable elements in cotton (*Gossypium raimondii*). Database 2017:bax013

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11(2):1–12

Zhang QX, Chen WB, Sun LD, Zhao FY, Huang BQ, Yang WR, Tao Y, Wang J, Yuan ZQ, Fan GY, Xing Z, Han CL, Pan HT, Zhong X, Shi WF, Liang XM, Du DL, Sun FM, Xu ZD, Hao RJ, Lv T, Lv YM, Zheng ZQ, Sun M, Luo L, Cai M, Gao YK, Wang J, Yin Y, Xu X, Cheng TR, Wang J (2012) The genome of *Prunus mume*. Nat Commun 3(176):1318