DNA Research, 2020, 0(0), 1–13
doi: 10.1093/dnares/dsaa004
Advance Access Publication Date: 9 April 2020
Full Paper

OXFORD

Full Paper

Genome-wide analysis of short interspersed nuclear elements provides insight into gene and genome evolution in citrus

Haijun Meng *, Jiancan Feng, Tuanhui Bai, Zaihai Jian, Yanhui Chen, and Guoliang Wu*

College of Horticulture, Henan Agricultural University, Zhengzhou 450002, China

*To whom correspondence should be addressed. Tel./Fax. +86 371 63579623. Email: hjmeng@foxmail.com (H.M.); wglhnnd@163.com (G.W.)

Received 20 June 2019; Editorial decision 2 April 2020; Accepted 3 April 2020

Abstract

Short interspersed nuclear elements (SINEs) are non-autonomous retrotransposons that are highly abundant, but not well annotated, in plant genomes. In this study, we identified 41,573 copies of SINEs in seven citrus genomes, including 11,275 full-length copies. The citrus SINEs were distributed among 12 families, with an average full-length rate of 0.27, and were dispersed throughout the chromosomes, preferentially in AT-rich areas. Approximately 18.4% of citrus SINEs were found in close proximity (≤1 kb upstream) to genes, indicating a significant enrichment of SINEs in promoter regions. Citrus SINEs promote gene and genome evolution by offering exons as well as splice sites and start and stop codons, creating novel genes and forming tandem and dispersed repeat structures. Comparative analysis of unique homologous SINE-containing loci (HSCLs) revealed chromosome rearrangements in sweet orange, pummelo, and mandarin, suggesting that unique HSCLs might be valuable for understanding chromosomal abnormalities. This study of SINEs provides us with new perspectives and new avenues by which to understand the evolution of citrus genes and genomes.

Key words: citrus, evolution, gene association, genome, short interspersed nuclear elements

1. Introduction

Transposable elements (TEs) are mobile DNA fragments that make up the largest fraction of eukaryotic genomes. For a long time, TEs were called 'junk genes' because they were thought to have no known function. In the past 30 years, research has shown that TEs play important roles in altering gene expression and structure,[1–3] chromosome rearrangement,[4,5] and the variability of genome size.[6,7] Through these and other functions, TE transposition serves as an important source of genetic variation, and thus, TEs have been exploited for the genetic improvement in crops.

Based on their different replication strategies, TEs in eukaryotes are divided into two broad classes: Type I elements (retrotransposons), which use an RNA-mediated mechanism for amplification, and Type II elements (DNA transposons), which use a DNA-mediated mechanism for transposition.[8] Retrotransposons can amplify themselves into thousands or tens of thousands of copies, whereas DNA transposons rarely attain these levels, with the exception of miniature inverted-repeat TEs (MITEs).[9] Each class of TEs contains autonomous elements, which have ORFs encoding the enzymes required for transposition, and non-autonomous elements, which do not encode transposition proteins but are still able to transpose.[10] Short interspersed nuclear elements (SINEs) are non-autonomous retrotransposons and depend on transposition proteins derived from their autonomous partners, long interspersed nuclear elements (LINEs), for amplification.[11]

SINEs range in length from 80 to 500 bp. They are a heterogeneous group of elements derived from a variety of RNA genes (tRNA, 7SL RNA, and 5S RNA),[10,12] but most are derived from tRNA gene sequences.[13] SINEs are characterized by a simple sequence repeat, usually a poly(A) at the 3′ terminus, and an internal RNA Pol III promoter within the 5′ terminus.[14] SINEs, except SINE3 from zebrafish, are flanked by target site duplications (TSDs).[12] These features are weakly conserved in plants, and thus, the annotation of SINEs is difficult and tedious via computer-based methods that are widely used for genome-wide identification of TEs.[15] Recently, Wenke et al. developed an algorithm for the *de novo* identification of SINEs, named 'SINE-Finder'.[14] However, at present, there remains fewer comprehensive studies of SINEs in plants than there are for other classes of TEs.

Although SINEs are short in length and compose small portions of eukaryotic genomes, they are abundant. In the human genome, there are over 1 million copies of the SINE *Alu*, which accounts for ~11% of the genome.[16] On the other hand, SINEs are relatively rare in plants. In potato, there are 2,359 SINE copies comprising 0.15% of the genome, and in sugar beets there are 6,326 SINE copies constituting 0.18% of the genome.[14,17]

Once integrated, SINEs are able to provide regulatory sequences to adjacent genes at a new integration site and thereafter have the potential to influence gene regulation. In mammals, SINEs are located throughout the genome, from intergenic regions to protein-coding genes,[18] and there is evidence that SINEs regulate gene expression. The mouse SINE B1 contains functional TF-binding sites for the carcinogen-activated dioxin receptor Xenobiotic Responsive Element (XRE) and for the epithelial–mesenchymal transition regulator Slug,[19] indicating that B1 acts as a *cis*-regulatory element on neighbouring genes. Human *Alu* RNA blocks the transcription of some protein-coding genes by binding Pol II and entering complexes of promoters during heat shock,[20] demonstrating its *trans*-regulation on distal genes. SINEs also play roles in post-transcriptional gene regulation. In human and mouse myoblasts, SINEs accumulate in 3′-untranslated regions (3′-UTRs) and regulate gene expression by Staufen-mediated mRNA decay.[21] In plants, SINEs are frequently associated with genes. Approximately 38% of the insertions are associated with transcribed regions in wheat, and 30% of SINE copies are associated with genes in Solanaceae,[15,22] indicating that SINEs also have the potential to regulate gene expression and alter gene structure in plants.

There are some reports that TEs are involved in the development of mutations in fruit. A MITE-like insertion in the promoter region of *CitRWP* generates polyembryonic citrus varieties with polyembryonic alleles.[23] In Chinese box orange, a MITE insertion with a variable level of DNA methylation in the promotor of *AbRuby2* affects the accumulation of anthocyanins in leaves.[24] In apple, a columnar mutation is associated with an integration of a *Gypsy*-like retrotransposon.[25] Over the years, the selection of different bud sports (somatic mutation) has given rise to many new fruit cultivars, particularly in seedless fruits such as sweet orange (*Citrus sinensis*) and Clementine mandarin (*Citrus reticulata*). SINEs, which are frequently associated with genes, have the potential to cause somatic mutation. However, the roles of SINEs in plant somatic mutation are rarely reported because SINEs are poorly investigated in plants, especially in citrus.

In this study, we identified 12 SINE families in 7 citrus genomes and analysed their family characteristics, amplification patterns, and copy distribution. We showed that citrus SINEs are inserted preferentially into AT-rich areas and enriched in the promoter regions of genes. In addition, citrus SINEs can create novel genes and be co-opted into genes. These results indicated that SINEs have played important roles in the gene and genome evolution of citrus.

2. Materials and methods

2.1. Data resources

Genome sequences and annotation files of sweet orange (*C. sinensis* cv. Valencia, 319 Mbp), Mangshan wild mandarin (*C. reticulata*, 334 Mbp), haploid pummelo (*Citrus grandis*, 346 Mbp), Ichang papeda (*Citrus ichangensis*, 357 Mbp), and citron (*Citrus medica*, 405 Mbp) were downloaded via the '*Citrus sinensis* annotation project' homepage (http://citrus.hzau.edu.cn/orange/download/data.php).[23,26] Genome sequences and annotation files of Clementine mandarin (*C. reticulata* cv. Clementina de Nules, 301 Mbp) were downloaded via the Citrus Genome Database homepage (https://www.citrusgenomedb.org/species/clementina/genome1.0) (Wu et al., 2014). Genome sequences and annotation files of Satsuma mandarin (*C. reticulata* cv. Miyagawa wase, 360 Mbp) were downloaded via the resources for citrus genomics homepage (http://www.citrusgenome.jp/).[27] The expression data of sweet orange were queried from the '*Citrus sinensis* annotation project' database homepage (http://citrus.hzau.edu.cn).[28]

2.2. Extraction of SINEs

To identify SINE candidates in citrus, we used the SINE-Finder program to search genome sequences using the default setting.[14] In brief, the settings were as follows: a 5′ TSD region of 40 nt, the box A motif RVTGG, a spacer of 25–50 nt, the box B motif GTTCRA, a spacer of 20–500 nt, six adenines or thymines as a poly(A/T) stretch or simple sequence repeats, and a 3′ TSD region of 40 nt.

The candidate sequences were clustered based on similarity using NCBI-BLAST 2.2.31+ toolkits.[29] The procedure consisted of three steps: clustering of candidate sequences, verification of SINE copies, and family assignment.

1. SINE candidates were clustered via all-against-all BLAST searches. In brief, an all-against-all BLAST was performed using the mega-BLAST algorithm with the following settings: qcov_hsp_perc of 80, perc_identity of 80, reward of 2, penalty of −3, gapopen of 5, gapextend of 2, word_size of 11, evalue of 10, no dust, and no soft_masking. The sequence with maximum homologues was retained while its homologues were removed. These two steps were repeated until no sequences had homologues. The remaining sequences were treated as representative copies.
2. Representative copies were used as queries for mega-BLAST to search citrus genome sequences to verify the candidate sequences. Mega-BLAST settings were the same as those in the first step. Hit sequences plus 60 bp flanking sequences were extracted and aligned using MUSCLE.[30] Alignments of each cluster were manually checked, and clusters without characteristics of SINEs or transposition hallmarks[31] (where regions belonging to SINEs are highly similar, but flanking regions are usually unrelated sequences) were discarded.
3. The remaining clusters were assigned into families via the construction of consensus sequences to perform all-against-all BLAST searches using the BLASTn algorithm with default settings, except that qcov_hsp_perc was 80, perc_identity was 80, and evalue was 10. Consensus sequences were constructed using SeaView.[32] SINE clusters with significant similarity (qcov_hsp_perc of 80, perc_identity of 80, and evalue of 10) were combined into a SINE family. Each family was designated as a CitruS (Citrus SINE) with a different number. For verification of the family assignment, 20

full-length copies of the top hits for each family were aligned using MUSCLE[29] prior to manual refinement and curation. Dendrograms were constructed with MEGA 7,[33] applying the neighbour-joining distance method.

The full-length copies of each family were retrieved by using consensus sequences as queries to search genome sequences with the BLASTn algorithm following default settings, except qcov_hsp_perc 80, perc_identity 80, and evalue 10. Retrieved sequences were aligned and checked to filter-out truncated SINE copies (the ends of homologous sequences varied by >5 bp). RepeatMasker[34] searches against genome sequences were conducted using consensus sequences as queries to retrieve all copies of each family (including intact and truncated SINE copies) with the following settings: -nolow, -no_is, -par 28, -xsmall, and -gff. Overlapping SINE copies were fused to keep the outermost genomic coordinates using BEDTools[35] under the following conditions: (i) overlapping SINE copies from an identical family were named after the original family and (ii) overlapping SINE copies from different families were named CitruS-composite.

2.3. Identification of TSDs and insertion site preferences

For identification of the TSDs of SINEs, a BLASTn search was performed using each full-length SINE sequence as a query with the default settings, except that word_size = 9, strand: plus, and evalue = 0.1. The results were filtered by size (10–40 nt) and location (upstream and downstream of SINEs).

Five nucleotides of the flanking region upstream of the 5′ TSD (positions −5 to −1) and the first six nucleotides of the 5′ TSD (positions 0–5) of each full-length copy were extracted and aligned to investigate the insertion preferences of the SINEs. Sequence conservation at each position was graphically represented by Weblogo 3.[36]

2.4. Analysis of similarity profiles of SINE families

Similarity profiles of SINE families were calculated with a modified protocol as described by Schwichtenberg et al.,[17] which was based on the sequence identity to the consensus sequence in each SINE family. In brief, a BLAST search was performed using consensus sequences as queries against all full-length copies, and then histograms were created using R script (https://www.R-project.org/) based on the resulting sequence identity data.

2.5. Identification of SINE transposition

To identify polymorphic SINE transposition sites in citrus, the leaves of 14 citrus cultivars were harvested from the collection at Huazhong Agriculture University (Wuhan, China). The DNA samples were isolated using Tiangen DP360 DNA extraction kit (Tiangen Biotech, Beijing, China).

To experimentally validate the insertion polymorphism, 84 SINE copies from the CitruS-I and II families were randomly selected from the genome sequence of sweet orange to design three primers (one primer shared across two pairs) for each locus, with one set designed to amplify the complete SINE copy and the other for the 5′ terminus of the SINE copy (Supplementary Table S1). Resulting polymorphic loci were mapped to the reference sequences of citrus using MUSCLE.[29]

Polymerase chain reaction (PCR) was conducted in a MJ-PTC-200 thermal PCR cycler (MJ Research, Waltham, MA) using the following program: 5 min of 94°C, 32 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 60 s, and a final extension at 72°C for 10 min. The reaction mixture (20 μL) contained about 20 ng of template genomic DNA, 0.2 μM each primer, 200 μM dNTPs, 1x PCR reaction buffer, and 1 unit of TaKaRa Taq™ DNA polymerase (Takara Bio, Beijing, China).

2.6. Chromosomal localization and gene association

We used MAPCHART[37] to display the chromosomal localization of the citrus SINEs. The positional information (GFF) of each SINE was extracted from RepeatMasker outputs and converted to the format required by MAPCHART. The sequence lengths of the chromosomal pseudomolecules of sweet orange, pummelo, Satsuma mandarin, and the longest nine scaffolds of Clementine mandarin were calculated to define the expected length of chromosomes for MAPCHART.

To analyse the association of the citrus SINEs with annotated genes, we compared the sites of SINE integration to the genomic coordinates of genes using BEDTools and the R packages systemPipeR and GenomicFeatures.[38,39] The SINE family specific fraction of genes as well as the distances of intergenic copies to the closest neighbouring gene were determined as described.[15] Genic regions were further distinguished into cds, introns, and UTRs according to the GFF annotation files. Intergenic regions were further distinguished into regions ≤1 kb upstream of genes, 1–2 kb upstream of genes, 2–5 kb upstream of genes, and ≥5 kb upstream of genes. When the downstream region of a gene overlapped with the upstream region of the next gene (0–5 kb), the overlapping region was excluded from the upstream region of the next gene (0–5 kb). The number of SINEs within each region was counted. Exemplary loci harbouring SINEs were manually refined and visualized with the respective annotations using Adobe illustrator CS6 (www.adobe.com). A dot plot of the dispersed duplication derived from truncated SINEs was calculated using the Emboss tool 'Dotmatcher' with the following settings: wordsize 20 and threshold 50.[40]

The Chi-squared test was used to compare the expected and observed values for gene association. The theoretical expectation was calculated based on the portion (%) of features in the genome, which was determined based on the published gene annotations after fusing overlapping annotations.

2.7. Comparative analysis of homologous SINE-containing loci

SINE copies sharing both 5′ and 3′ flanking sequences were isolated from the alignment of each family to extract syntenic blocks. Then, 200 bp sequences flanking both ends of these SINE copies were extracted and aligned. SINE copies sharing identical flanking sequences (identity ≥80%) were referred to as homologous SINE-containing loci (HSCLs). Synteny for HSCLs was illustrated by Circos[41] following the instructions provided by the author.

3. Results

3.1. Mining SINEs in citrus genomes

Using SINE-Finder and the public genome sequences of Clementine mandarin, Mangshan wild mandarin, Satsuma mandarin, pummelo, sweet orange, citron, and Ichang papeda, 10,632 SINE candidates were identified. To exclude false candidates, we developed a three-step pipeline to refine the output from SINE-Finder. In the first step, all-against-all BLAST searches using the mega-BLAST algorithm was performed to cluster the SINE candidates and find representative copies. Secondly, the representative copies were used as queries for BLAST searches to retrieve homologous sequences and flanking

sequences from genome sequences. The resulting sequences were aligned, and 15 clusters were identified as SINEs, each of them harboured typical structural SINE features, such as RNA polymerase III promoter boxes A and B, poly(A/T) tails, TSDs, and transposition hallmarks.[31] Clusters lacking any one of these features were removed. Thirdly, another all-against-all BLAST search was performed between the consensus sequences of 15 clusters using the BLASTn algorithm to find similarities between the clusters. The BLAST results revealed that these SINEs were grouped into 12 distinct families, designated CitruS-I to CitruS-XII (Table 1 and Supplementary Data S1), following Wicker's '80-80-80' criteria.[8]

Consensus sequences of the 12 families were used as queries for BLAST searches against the 7 citrus genomes, resulting in 11,275 full-length copies (Table 1). Pummelo harboured the highest copy number, with 2,118 full-length copies, and citron harboured the lowest copy number, with 1,339 full-length copies. The lengths of the consensus sequences were highly variable, ranging from 192 nt (CitruS-XI) to 335 nt (CitruS-IV). RepeatMasker was utilized to find all of the members, including full-length copies and truncated copies that resulted from genomic rearrangements and/or termination of the RNA intermediate. A total of 41,573 full-length and truncated SINE copies were found among all seven citrus genomes (Table 1 and Supplementary Data S2), in which the number of SINE in each species ranged from 7,097 (citron) to 4,792 (Clementine mandarin). Hundreds of composite SINEs were identified that represented overlapping copies of SINEs from different families. The genome fractions covered by SINEs varied between 0.37%, in pummelo, and 0.33%, in mandarins (Clementine mandarin, Mangshan wild mandarin, and Satsuma mandarin).

The number of full-length SINEs differed widely among the families (Table 1). CitruS-I was the most populous family, with >3,434 full-length copies across the seven citrus genomes, whereas the CitruS-III and CitruS-X families each contained <100 full-length copies. The remaining nine families had >340 full-length members. The abundance of the full-length SINEs differed from the number of all members (truncated and full-length) of each family identified using RepeatMasker.[34] Some families had relatively few full-length copies but abundant truncated members. To shed light on this inconsistency, we calculated the copy number ratios of full-length copies to all members for each family (Fig. 1), referred to as the full-length rate. The average full-length rate of all citrus SINEs was ~0.27, which indicated that most copies of citrus SINEs were truncated during evolution. The average full-length rate varied among the SINE families, ranging from 0.02 in CitruS-X to 0.62 in CitruS-I. The differences in the full-length rate among the SINE families were substantially broader in different genomes, ranging from 0.01 (CitruS-X in citron) to 0.86 (CitruS-III in Ichang papeda). These data suggested that citrus SINEs experienced family specific evolutionary histories.

3.2. Comparative analysis of citrus SINEs

To verify the structural features of citrus SINEs, each full-length SINE copy plus 60 nts of flanking sequences was extracted and compared. SINE copies within each family were highly conserved and shared sequence similarity of 80–100%, but their flanking regions were not similar, confirming the presence of transposition hallmarks.[31] Two conserved motifs (box A and box B) were identified in all 12 SINE families (Supplementary Fig. S1).

TSDs originate from the integration process of SINEs and thereafter retain residual information of the transposition process. Among the 11,275 full-length SINEs, we identified 4,793 with pair-matched TSDs (size threshold: 10–40 nt) (Supplementary Data S3). The TSDs averaged between 15 and 17 nt (Table 1). The length of individual TSDs was variable, with the majority of copies ranging from 10 to 22 nt (Supplementary Fig. S2). To illustrate the insertion preferences of the citrus SINEs, the families that had at least 30 full-length copies with detectable TSDs were further investigated. We examined five nucleotides of the flanking region upstream of the 5′ TSD (positions −5 to −1) and the first six nucleotides of the 5′ TSD (positions 0–5) of each SINE copy (Fig. 2A and Supplementary Fig. S3). Generally, citrus SINEs preferentially integrated upstream of short adenine stretches (positions 0–5), with the first two nucleotides of the TSD (positions 0–1) most likely being adenine. The flanking region upstream of the TSD (positions −4 to −2) was A/T rich. However, the first nucleotide upstream of the TSD (position −1) was rarely adenine. The 5′ termini of TSDs were rich in adenines, which likely overlapped with the poly(A) tails. The variations in the lengths of poly(A) tails were not evaluated to avoid bias derived from overlap between poly(A) tails and TSDs.

To visualize the divergence and grouping of the SINE families, we attempted to construct a dendrogram using the 20 full-length SINE sequences from each family with the highest similarity to the consensus sequence. A trial construction of the dendrogram failed due to the presence of three excessively divergent families. However, a dendrogram was constructed when the remaining nine families that formed separate branches were used. The families CitruS-II and CitruS-VI were assigned to two subfamilies and grouped on separate branches (Fig. 2B). The consensus sequences of CitruS-II a/b and CitruS-VI a/b showed identity values of 82.2 and 82.1%, respectively. In contrast, the consensus sequences of CitruS-VII and IV show identities of 70%, due to which they are classified as distinct SINE families rather than subfamilies.

3.3. Estimation of the age and transposition activity of the SINEs

The lengths of TSDs and poly(A) tails are indicators of the relative insertion time of individual SINE copies.[15,17,42,43] In citrus, the lengths of the TSDs and poly(A) tails were not eligible as an age indicator because more than half of the full-length copies did not have detectable TSDs, and the poly(A) tails were likely to overlap with short adenine stretches of the TSDs. Therefore, these could not be utilized to determine the age of the SINE copies in citrus. Generally, SINE copies were 'copied and pasted' a long time ago, leading to the accumulation of mutations. As the consensus sequence is a suitable approximation of the original source copy, the decreasing identity of a copy to the consensus sequence can serve as an alternative way to estimate its age.[15,44,45]

To estimate the activity of a SINE family, we calculated copy numbers relative to sequence similarity intervals for the families/subfamilies that had at least 30 full-length copies (Fig. 3B and Supplementary Fig. S4). Figure 3 shows typical examples of SINEs with family- and genome-specific transposition patterns. In some families, for example CitruS-I in sweet orange, each similarity interval contained a relatively consistent copy number, suggesting that this family had consistent activity over a long period (Fig. 3A). In pummelo, all CitruS-I copies were at least 90% similar to the consensus sequence, indicating a recent activity of this family in pummelo (Fig. 3B). Some families, for example CitruS-XI in Satsuma mandarin, were predicted to be old, because all SINE copies had similarities of <91% to the consensus sequence (Fig. 3C). Peaks of ~90% similarity were identified in some families, for example CitruS-VII in

Table 1. SINE families in seven citrus genomes

Family	Clementine		Mangshan		Satsuma		Pummelo		Sweet orange		Citron		Ichang papeda		Total			
	Full-length copies	All copies[a]	Full-length copies	All copies	Full-length copies	All copies	Full-length copies	All copies	Full-length copies	All copies	Full-length copies	All copies	Full-length copies	All copies	Full-length copies	All copies	Consensus (nt)[b]	TSD (nt)[c]
CitruS-I	413	603	423	704	461	762	746	1,035	535	841	396	895	460	666	3,434	5,506	204	16
CitruS-IIa	97	256	127	394	96	301	46	183	81	259	105	501	74	249	626	2,143	298	15
CitruS-IIb	107	396	107	472	130	487	98	397	99	418	142	600	101	485	784	3,255	316	15
CitruS-III	15	18	16	22	14	19	10	12	13	18	5	9	12	14	85	112	178	17
CitruS-IV	79	355	66	418	99	436	136	478	103	457	48	446	86	444	617	3,034	335	15
CitruS-V	152	294	167	374	185	353	198	374	178	349	157	382	183	360	1,220	2,486	215	16
CitruS-VIa	73	302	73	329	97	400	248	578	126	452	119	931	180	505	916	3,497	302	16
CitruS-VIb	24	642	17	753	24	740	30	673	33	693	67	864	18	843	213	5,208	297	16
CitruS-VII	113	408	96	459	110	487	215	561	148	501	18	294	140	528	840	3,238	328	15
CitruS-VIII	39	232	45	305	47	303	72	345	54	289	40	622	99	442	396	2,538	258	16
CitruS-IX	92	164	103	199	96	175	95	149	88	175	69	173	138	220	681	1,255	171	16
CitruS-X	4	184	3	251	3	226	5	214	4	217	2	213	12	278	33	1,583	283	11
CitruS-XI	128	413	153	498	161	502	166	481	154	474	127	478	198	621	1,087	3,467	192	16
CitruS-XII	42	115	51	142	41	133	53	136	53	127	44	153	59	160	343	966	123	16
Composite[d]		410		445		461		416		471		536		546		3,285		
Total	1,378	4,792	1,447	5,765	1,564	5,785	2,118	6,032	1,669	5,741	1,339	7,097	1,760	6,361	11,275	41,573		
Genome fraction[e]		0.33%		0.33%		0.33%		0.37%		0.35%		0.36%		0.36%				

nt: nucleotides; Clementine: Clementine mandarin; Mangshan: Mangshan wild mandarin; Satsuma: Satsuma mandarin.

[a]Including full-length copies and truncated copies.

[b]Consensus sequence without poly(A/T).

[c]Averaged length.

[d]CitruS-composite.

[e]Genome fractions covered by SINEs.

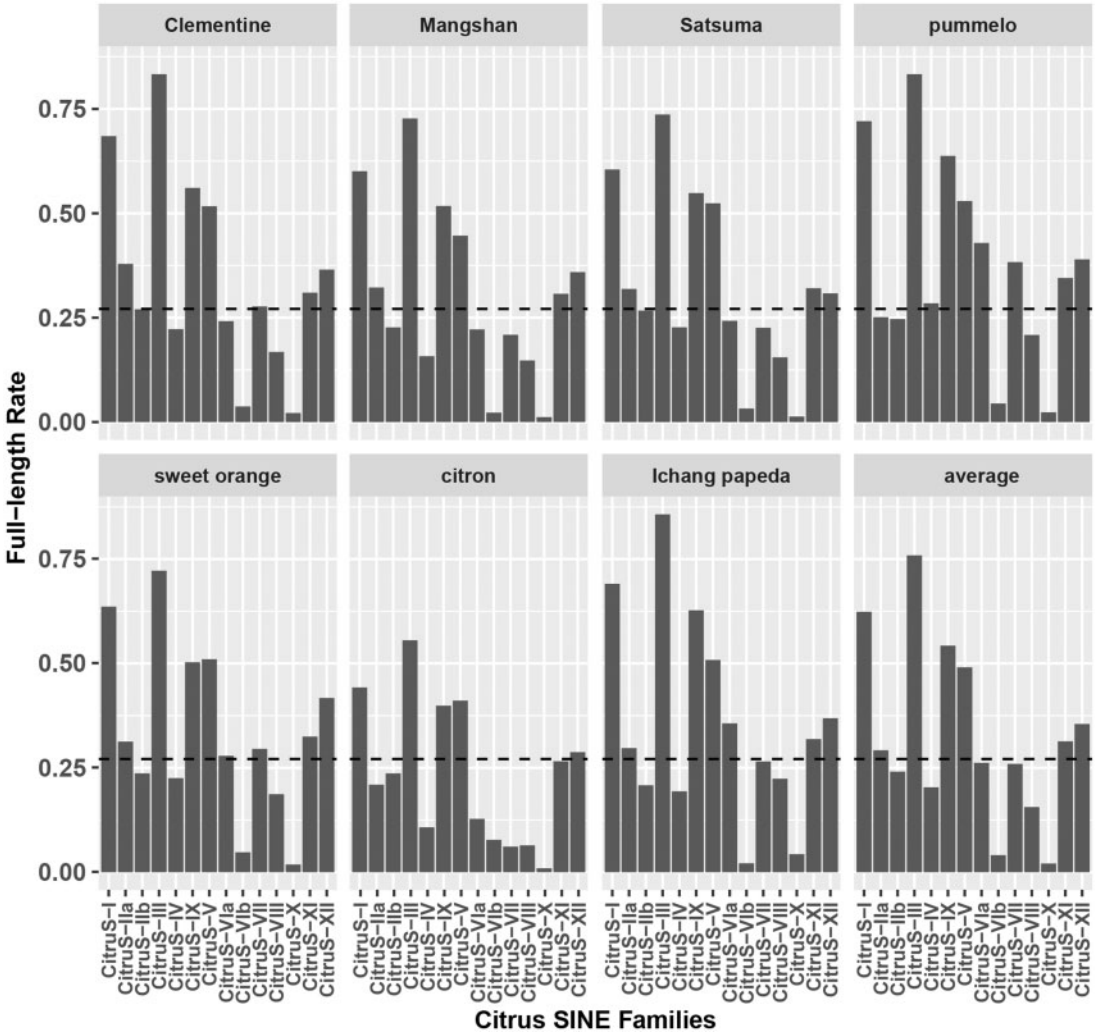

Figure 1. The full-length rate of citrus SINEs characterized on the SINE family and genome level. Full-length rates for each family are presented for Clementine mandarin, Satsuma mandarin, Mangshan wild mandarin, pummelo, sweet orange, citron, Ichang papeda, and the average across all seven genomes (average). The dashed line represents the average full-length rate of all citrus SINEs.

sweet orange, which suggested that there was an ancient rapid amplification of these families (Fig. 3D).

CitruS-I and CitruS-II were used to search for polymorphic SINE copies in citrus to verify different amplifications. For each locus, three primers were designed to form two primer pairs in order to amplify two overlapping products, the complete SINE copy and the 5′ terminus of the SINE copy, with one primer upstream of the SINE copy, one primer downstream of the SINE copy and one primer inside the SINE copy. We detected a polymorphic SINE copy that was mainly derived from CitruS-IIa (CitruS-composite) in 14 citrus accessions (Fig. 4A and B). The SINE copy was absent in all 3 pummelo accessions but was present in the remaining 11 accessions, namely, 4 accessions of mandarin and 7 accessions of sweet orange. Moreover, two accessions of mandarin were homozygous for a SINE insertion, and the remaining nine accessions were heterozygous. There were no SINE insertions in a locus located in the sixth intron of a polygalacturonase gene in the published reference genome of Clementine mandarin (cv. Clementina de Nules) (Fig. 4C), but the PCR results showed that the locus was heterozygous for the SINE insertion (cv. Caffin).

3.4. Gene association and chromosomal distribution of citrus SINEs

To shed light on the association of SINEs with genes, we analysed the frequency and position of citrus SINEs relative to annotated genes (Fig. 5A). On average, 7.9% of citrus SINEs were located in genes, of which the majority was found in introns and UTRs. Only 0.5% (0.1–1.4%) of citrus SINEs were found in coding regions of genes (cds). Based on the physical length of the annotated genes,[26] an average of 29% (25–36.4%) of the citrus genome sequences were annotated as genes, which indicated a highly significant depletion of SINEs in genic regions compared with the expectation by random distribution according to Chi-squared tests ($P < 0.001$). Approximately 18.4% of citrus SINEs were found in close proximity ($\leq$1 kb upstream) to genes, which indicated a significant enrichment of SINEs in promoter regions ($P < 0.001$). Approximately 38.7% of citrus SINEs were 1–5 kb upstream of the next gene, and 34.9% were located more distantly (>5 kb). The fact that SINEs were enriched in promoter regions suggested that SINEs have the potential to influence the regulation of the expression of neighbouring genes.

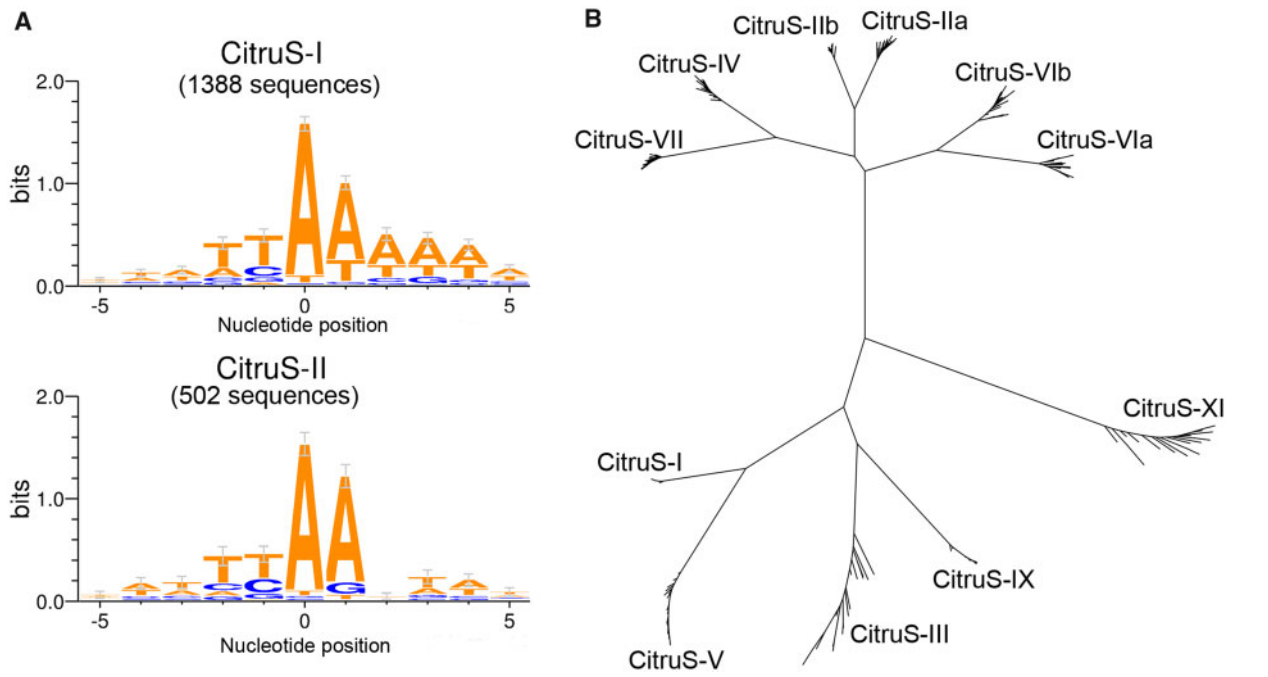

Figure 2. Graphical representation of insertion site preferences and dendrogram showing the divergence and grouping of citrus SINE families. (A) Relative nucleotide frequency at five positions upstream of the 5′ TSD (positions −5 to −1) and the first six positions of the 5′ TSD (positions 0–5). (B) The dendrogram is based on the 20 full-length SINE sequences of each citrus SINE family that had the highest similarity to the consensus sequence. The insertion site preferences of the remaining seven SINE families are shown in Supplementary Fig. S3.

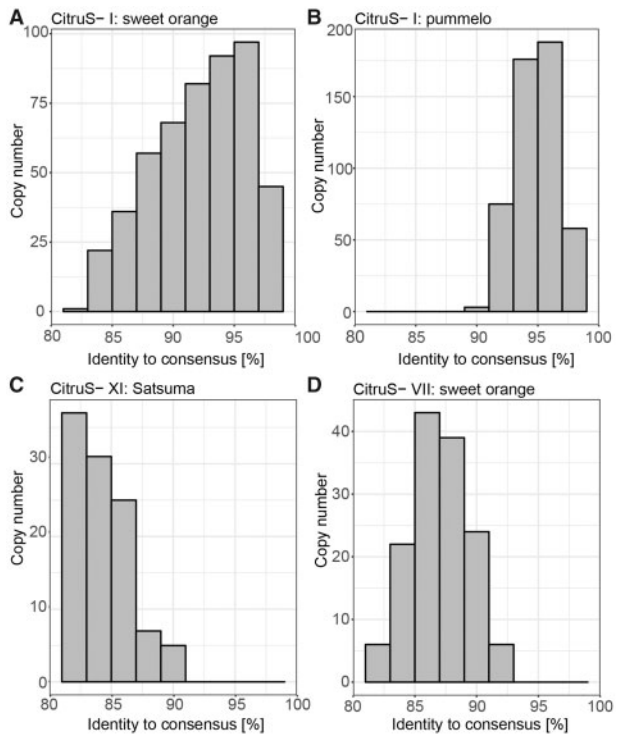

Figure 3. Comparison of copy numbers and sequence similarities among citrus SINE families. For three SINE families, histograms were created based on the identity of each full-length SINE copy within the family to the consensus sequence. Different example patterns are shown, namely, (A) a consistent activity over a long period, (B) a recent activity, (C) transposition activity a long time ago, and (D) an aged, rapid amplification. The complete data for the remaining 74 SINE families by genome are shown in Supplementary Fig. S4.

There were evident differences between the individual SINE families in the association with genes in all citrus genomes. The family CitruS-I accounted for at least 29.8% of SINEs located in cds in sweet orange, Satsuma mandarin, and Clementine mandarin, but no genic SINEs in the remaining four genomes (Supplementary Fig. S5). Overall, the CitruS-I and VIb families were the top two genic SINE fractions in citrus (16.5 and 12.5%, respectively).

The localization of SINE copies was investigated on chromosomal pseudomolecules of Satsuma mandarin, sweet orange, and pummelo to determine the chromosomal distribution of citrus SINEs. The citrus SINEs had a dispersed distribution, with a preference for distal to subterminal regions on some chromosomes (Fig. 5B). Nevertheless, citrus SINEs had small regions of local depletion. The distributions of citrus SINEs also showed family specific differences. For example, the highly abundant CitruS-I and VIb mapped along chromosomes with different patterns in Satsuma mandarin, sweet orange, and pummelo (Supplementary Fig. S6).

3.5. Contribution of SINEs to gene and genome evolution

All cds containing SINEs in sweet orange were examined as an example to elucidate the role of citrus SINEs in gene and genome evolution. Comparison of the gene annotations indicated that SINEs might affect transcript structure as they contribute exons as well as splice sites and start and stop codons (Fig. 6). In Fig. 6A, a SINE copy was integrated into the sixth intron as a new exon and contributed two splice sites, maintaining the reading frame of the downstream region. The SINE transcript harboured the 3′ splice site (AG) of the sixth intron and the 5′ splice site (GU) of the seventh intron. In Fig. 6B, a SINE copy provided the first exon, a start codon, and a

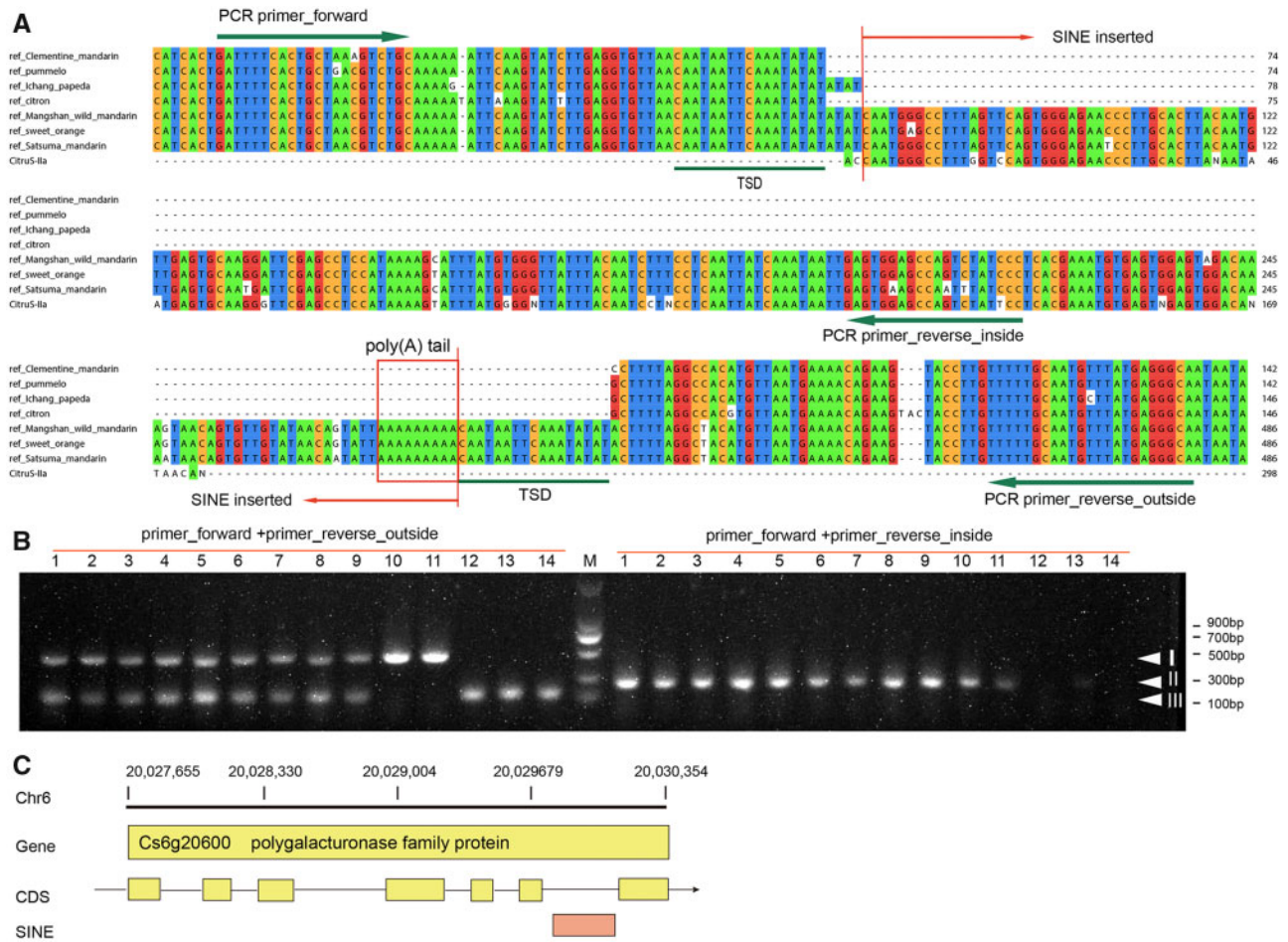

Figure 4. Genomic sequence variations derived from SINE insertion. (A) Alignment of reference sequences of a predicted polygalacturonase gene showing the location of a SINE insertion and PCR primers. The inserted SINEs were mainly derived from CitruS-IIa. Three PCR primers are indicated with green arrowheads. TSDs are indicated with green line, poly(A) tail is indicated with red box. (B) PCR results of two primer pairs using three primers. Lane 1: Clementine mandarin (cultivar 'Caffin'); lanes 2–8: sweet orange (cultivars 'Valencia', 'Qingjia', 'Lunwan', 'Xuecheng', 'Hamlin', 'Newhall', and 'Jinchen', respectively); lanes 9–11: mandarins (cultivars 'Ponkan', 'Guoqing I', and 'Bendizao', respectively); lanes 12–14 : pummelo (cultivars 'Chandler', 'Guanximiyou', and 'Gaoban', respectively); and lane M: DNA ladder. White triangle I indicates PCR products of primer_forward + primer_reverse_outside with a SINE insertion (473 bp). White triangle II indicates PCR products of primer_forward + primer_reverse_inside which suggested a SINE insertion (212 bp). White triangle III indicates PCR products of primer_forward + primer_reverse_outside without a SINE insertion (129 bp). (C) Chromosomal location of the SINE insertion.

5′ splice site. Citrus SINEs donating stop codons (Fig. 6C and D) were also identified in genes.

We identified 10 short genes containing only one exon (Supplementary Table S2), eight of them encoding unrecognized proteins, one encoding a WRKY DNA-binding protein (*Cs8g16370*), and one encoding a retrotransposon protein (*orange1.1t01062.1*). Further investigation revealed that these genes were derived from a SINE and its flanking sequences, in which the SINE provided the start/stop codons or part of the internal codons (Fig. 6E and F). Upon query of the publicly available RNA-seq dataset (citrus.hzau.edu.cn) derived from callus, leaf, flower, and fruit of sweet orange, we found that all 10 genes were expressed at low or moderate levels. The gene *Cs5g04640* was significantly up-regulated ($P < 0.01$) in fruit compared with callus (Supplementary Table S3). These data suggested that these 10 genes were able to transcribe into RNA. In 5 kb of flanking regions, there were no sequences annotated as TEs surrounding the retrotransposon gene *orange1.1t01062.1*, which indicated the gene *orange1.1t01062.1* was not likely a part of another TE. The gene *Cs8g16370*, encoding a homolog of a functional protein, was mainly derived from a SINE. Therefore, we inferred that *orange1.1t01062.1* and *Cs8g16370* might be co-opted from SINEs, and the remaining eight small genes were likely novel genes created by SINEs. In addition, we found tandem amplification and dispersed duplications derived from truncated SINEs (Supplementary Fig. S7), which suggested that SINEs participate in genomic rearrangements.

3.6. Comparative analysis of HSCL

Alignment of full-length SINE copies revealed hundreds of clusters of SINE copies that shared almost identical flanking sequences (Supplementary Fig. S8), indicating that these SINEs were located in homologous loci. Two hundred base pairs of flanking sequences at both ends of each copy in each cluster were extracted and aligned. Loci in which SINE copies shared identical flanking sequences (identity ≥80%) were designated HSCLs. HSCLs were divided into unique HSCLs and multiple HSCLs according to their number of occurrences in each genome. The number of occurrences of HSCLs was not consistent in all seven genomes due to family specific

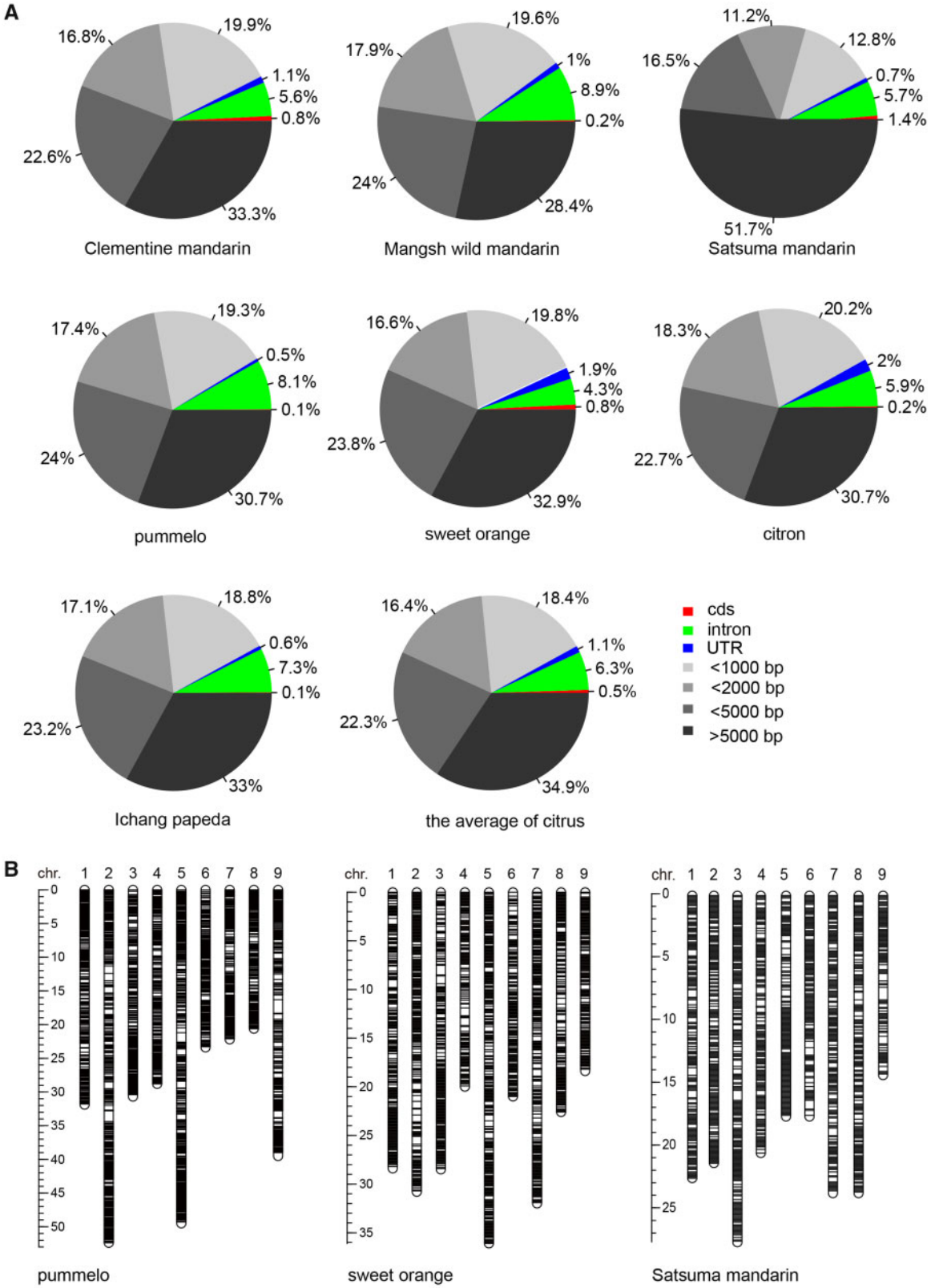

Figure 5. Gene association and chromosomal distribution of citrus SINEs. (A) The frequency and position of citrus SINEs relative to the annotated genes for each genome. The average distribution across all seven genomes is shown in 'the average of citrus'. (B) Chromosomal mapping of all SINEs in pummelo, sweet orange, and Satsuma mandarin. Scale provided in Mbp.

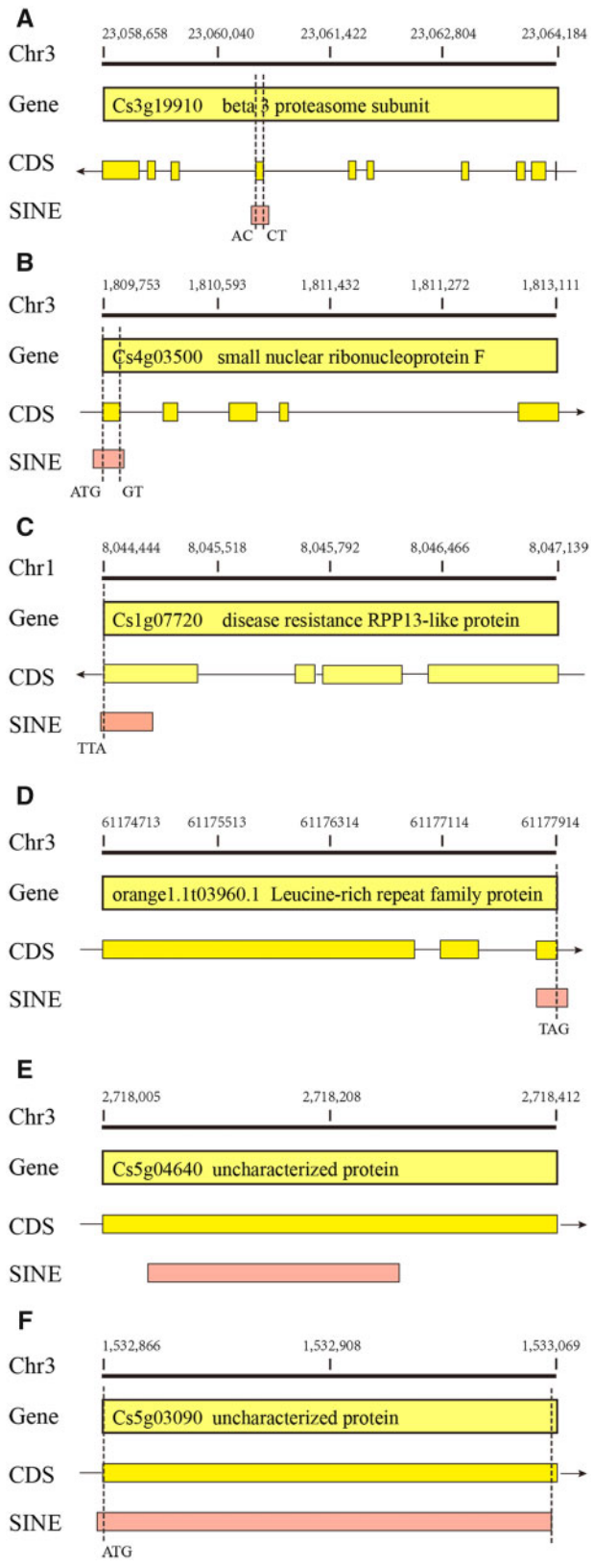

Figure 6. The pattern of SINEs affecting the splicing and translation of annotated genes and creating novel genes in sweet orange. Citrus SINEs were identified that (A) created a new exon and contributed two splice sites, (B and F) provided the first exon donating a start codon and (B) a 5′ splice site, (C and D) donated stop codons, and (E, F) created novel genes.

amplification of citrus SINEs after speciation. We identified 1,038 unique HSCLs shared by sweet orange, pummelo, and Clementine mandarin, of which 234 were shared by all three species, 331 were shared by sweet orange and pummelo, 375 were shared by sweet orange and Clementine mandarin, and 99 were shared by pummelo and Clementine mandarin. Given that sweet orange is a descendant of ancient pummelo and an ancestor of Clementine mandarin,[26,46] those data may reflect the evolutionary relationship between these species.

The chromosomal localization of the SINE copies in the unique HSCLs based on their physical positions was visualized on chromosomal pseudomolecules of sweet orange, pummelo, and the nine longest scaffolds (>10 Mbp) of Clementine mandarin (Supplementary Fig. S9A) to illustrate the synteny of unique HSCLs that amounted to orthologs. Although the majority of unique HSCLs retained substantial synteny, chromosomal rearrangements were observed when we compared the unique HSCL distribution at the chromosome level. For example, while sharing the majority of syntenic blocks with chromosome 4 of sweet orange, scaffold 1 of Clementine mandarin has some blocks homologous to chromosomes 7 and 1 of sweet orange (Fig. 7A). Similar patterns of synteny and chromosomal rearrangements were found among sweet orange, pummelo, and Satsuma mandarin (Supplementary Fig. S9B and Fig. 7B). These syntenic blocks derived from unique HSCLs might provide a robust and precise sequence framework for understanding citrus genome evolution and aid in the assembly of chromosomal pseudomolecules in citrus.

4. Discussion

In many released annotations of plant genomes, SINEs are neglected elements. In current citrus genome annotations, SINEs are either underestimated or absent compared with our results.[23,26,27,47] The difficulty in mining SINEs and their low proportions in the genomes may explain the incomplete annotation of SINEs in these released genomes. The absence of annotated SINEs will hinder the calling of genome structure variants, ultimately compromising all efforts based on next generation sequencing. In the present study, we identified 41,573 SINE copies in 12 families in seven citrus genomes. Some of these SINE copies were associated with genes, which indicate that the insertion of SINEs around genes may be an important source of variations in gene expression and structure.

Our preliminary results are based on the results of using SINE-Finder.[14] Approximately half of the candidates were either parts of other TEs (such as LTR retrotransposons) or not TEs, which diminishes the efficiency of mining for SINEs. In the present study, we developed a pipeline to filter and cluster the SINE candidates produced by SINE-Finder.[14] The pipeline is based on NCBI-BLAST+ tools[29] with the 'qcov_hsp_perc' and 'perc_identity' options, which enabled us to remove false candidates and follow Wicker's '80-80-80' rule[8] for SINE family assignment.

Wenke et al.[14] proposed a new rule for SINE family assignment because they believed that the heterogeneity in the sequences and lengths of SINEs contradicted the criteria for the SINE family assignment suggested by Wicker et al.[8] Thereafter, Schwichtenberg et al.[17] assigned Amaranthaceae SINEs to the same family when they shared at least 60% similarity, resulting in 22 SINE families. This is why the full-length SINE copies in our study showed an average sequence similarity of 84–91% to the consensus, while the average sequence

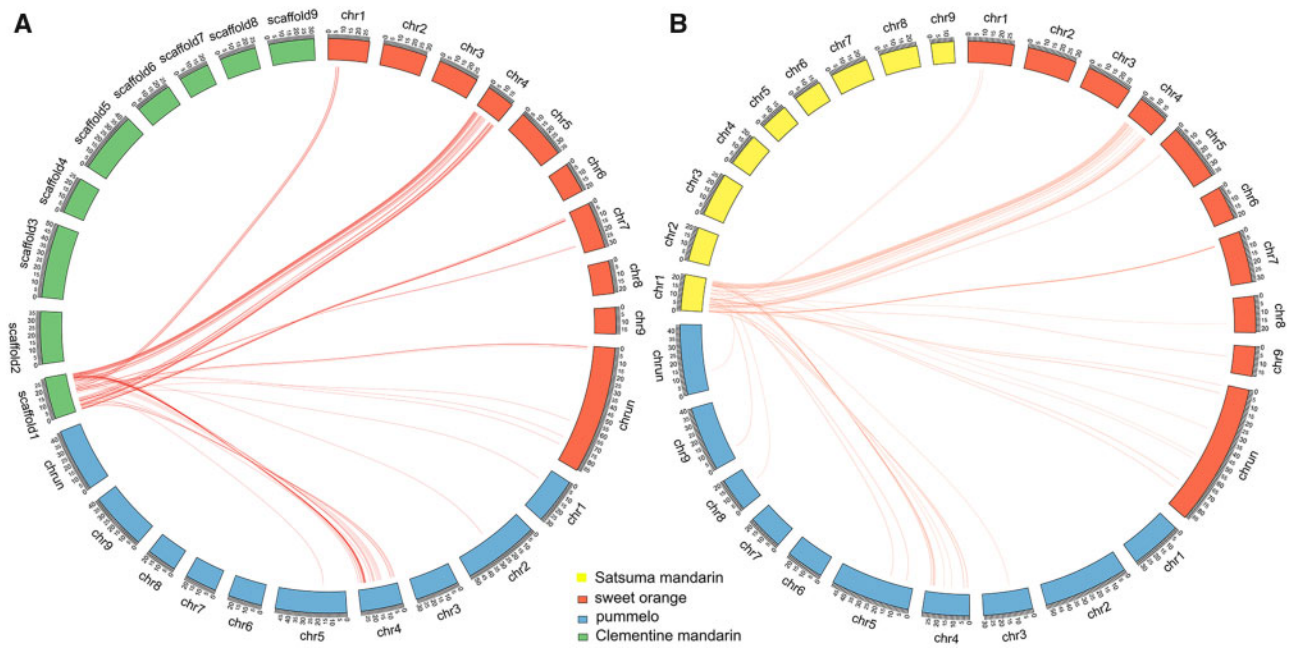

Figure 7. Chromosomal rearrangements revealed by unique HSCLs. Distribution of syntenic blocks linked to scaffold 1 of Clementine mandarin (A) and chromosome 1 of Satsuma mandarin (B). All 12 SINE families were included. Sweet orange is a descendant of ancient pummelo and an ancestor of Clementine mandarin.

similarity ranged from 60 to 100% in Wenke et al.[14] and Schwichtenberg et al.[17]

The copy number of TEs is the result of the balance between amplification and partial or complete loss.[47] Amplification and loss lead to lineage-specific TE copy numbers and copy number variation across families.[10,15,48,49] In the present report, we found that the copy number of full-length SINEs varied significantly across SINE families and that the variations had family specific patterns in citrus. The full-length SINE copies represented recent amplifications or retentions of the intact sequences, while truncated SINEs may result from an incomplete reverse transcription during SINE transposition or partial removal of genomic SINE copies. The full-length rate of SINE families might be an indicator of the balance between SINE amplification and partial loss. Therefore, we inferred that many members of the CitruS-I and CitruS-III families might be newly formed or stably intact based on their full-length rates and activity estimates.

Site-preferential insertion is observed in many classes of TEs.[50–53] Our results showed that SINEs have a very strong preference to integrate upstream of short adenine stretches, which strongly resemble the cleavage site specificity of human LINE L1 endonuclease.[54] SINEs are non-coding and require both active LINEs and sequence-dependent recognition of the SINE 3′ end by the LINE reverse transcriptase for retrotransposition.[14] Our results indicate that citrus SINEs might use reverse transcriptional machinery of LINE L1. The strong preference for AT-rich areas also indicates that citrus SINEs preferentially insert into areas with light cytosine methylation, compared with GC-rich areas. As inferred from previous reports,[14,17] site-preferential insertion may be a reason for the enrichment of citrus SINEs near genes. While LTR retrotransposons, the most abundant TEs in plants, are enriched in heterochromatin and silenced by DNA methylation of cytosine nucleotides and histone methylation,[55–58] SINEs are not enriched in telomeric and centromeric heterochromatin.[59] In citrus, SINEs also have a dispersed distribution pattern along chromosomes. Site-preferential insertion might be one possible explanation for the distribution patterns of SINEs on chromosomes.

The 12 citrus SINE families compose 0.35% of the citrus genome on average, which is much less than the content of SINEs in the human genome.[16] These results also show that SINEs are not as abundant as other types of TEs in plants.[47,60,61] However, >28% of the SINE copies were located in genic and adjacent areas (cds, introns, UTRs, and 1 kb upstream of gene) in the citrus genome, which is consistent with observations in Amaranthaceae and Solanaceae species.[14,17] These observations suggest that SINEs are often associated with genes.

Our results revealed that citrus SINEs promoted gene evolution by their insertion into promoter regions, by increasing UTR and intron lengths, by providing splice sites, exons, and start and stop codons, and by creating novel genes. These events are called TE co-options. TE co-options are involved in changing the patterns of gene expression, changing the functions of the proteins they encode, or both.[62] We believe that in citrus, SINEs were co-opted into genes by providing new potential regulatory sequences to adjacent genes, altering gene structure, and creating novel genes. Interestingly, new SINE integrations may create novel genes.

Our results showed that there was a highly significant depletion of SINEs in genic regions, which might be the result of the transposition of SINEs under selective constraints. This finding suggests that SINE insertions in genic regions might account for only a small fraction of alterations in gene structure. However, SINEs are enriched in promotor regions (≤1 kb upstream of genes), indicating that SINEs might serve as regulatory sequences of neighbouring genes in citrus. TEs respond to biotic and abiotic stresses, including pathogens, extreme environmental conditions, polyploidization, and interspecies hybridization.[15,63–65] DNA methylation of inserted TEs also affects gene expression.[24] Therefore, we speculate that SINEs might play

roles in regulating gene expression through insertion into the regulatory sequences and dynamic changes in DNA methylation.

The insertion of a SINE creates a new allele at each specific locus, which can be inherited by descendants to form HSCLs. HSCLs were unique in each genome except those located inside duplicated genomic regions. The distribution and abundance of unique HSCLs can function as indictors of chromosomal rearrangements. There is a complicated phylogenetic relationship among citrus species. For example, sweet orange, the most widely cultivated citrus, is the offspring of previously admixed individuals derived from pummelo and mandarin, while Clementine mandarin (also known as Algerian tangerine) is a hybrid of Mediterranean mandarin (Willowleaf) and sweet orange.[45] Comparative analysis of unique HSCLs revealed some chromosome rearrangements, suggesting that unique HSCLs are useful tools for ancestry research. Similar patterns of chromosome rearrangements were found between mandarins (Clementine and Satsuma) and sweet orange, which reflected the similar genetic background between Clementine mandarin and Satsuma mandarin. With the widespread use of next generation sequencing, we now have the chance to observe a different aspect of the forces that shape molecular evolution by studying gene sequence variation within a species.[66] Unique HSCLs might be useful tools for comparison of intraspecies variation, too. Nevertheless, the use of unique HSCLs in evolution research has limitations, because the mining of syntenic HSCLs depends on high-quality genome assembly.

The accuracy of our results relied on the quality of the *de novo* sequencing of the genome. Constant improvement in genome sequences due to ongoing assembly, anchoring and orientation of scaffolds and annotation will improve the identification and annotation of SINEs.[17] Our results that SINEs play an important role in gene and genome evolution will serve as a foundation for future investigations of TEs in citrus.

Supplementary data

Supplementary data are available at DNARES online.

Acknowledgements

We would like to thank Anita K. Snyder for English editing of the manuscript.

Accession numbers

BCWF01000001-BCWF01000044

Funding

This work was supported by the Science and Technology Project in Henan Province (102102110178 and 151100110900), the National Natural Science Foundation of China (31171943), and the Natural Science Foundation of Henan Education (2011B210019).

Conflict of interest

None declared.

References

1. Harris, C.J., Scheibe, M., Wongpalee, S.P., et al. 2018, A DNA methylation reader complex that enhances gene transcription, *Science*, **362**, 1182–6.
2. Naito, K., Zhang, F., Tsukiyama, T., et al. 2009, Unexpected consequences of a sudden and massive transposon amplification on rice gene expression, *Nature*, **461**, 1130–4.
3. Wang, L., He, F., Huang, Y., et al. 2018, Genome of wild mandarin and domestication history of mandarin, *Mol. Plant*, **11**, 1024–37.
4. Zhang, J., Yu, C., Krishnaswamy, L. and Peterson, T. 2011, Transposable elements as catalysts for chromosome rearrangements. In: Birchler, J.A., ed., *Plant Chromosome Engineering: Methods and Protocols*, Humana Press: Totowa, NJ, pp. 315–26.
5. Feinberg, A.P. 2007, Phenotypic plasticity and the epigenetics of human disease, *Nature*, **447**, 433–40.
6. Tenaillon, M.I., Hufford, M.B., Gaut, B.S. and Ross-Ibarra, J. 2011, Genome size and transposable element content as determined by high-throughput sequencing in maize and zea luxurians, *Genome Biol. Evol.*, **3**, 219–29.
7. Kapusta, A., Suh, A. and Feschotte, C. 2017, Dynamics of genome size evolution in birds and mammals, *Proc. Natl. Acad. Sci. USA*, **114**, E1460–9.
8. Wicker, T., Sabot, F., Hua-Van, A., et al. 2007, A unified classification system for eukaryotic transposable elements, *Nat. Rev. Genet.* ., **8**, 973–82.
9. Yang, G., Zhang, F., Hancock, C.N. and Wessler, S.R. 2007, Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana, *Proc. Natl. Acad. Sci. USA*, **104**, 10962–7.
10. Wessler, S.R. 2006, Transposable elements and the evolution of eukaryotic genomes, *Proc. Natl. Acad. Sci. USA*, **103**, 17600–1.
11. Okada, N., Hamada, M., Ogiwara, I. and Ohshima, K. 1997, SINEs and LINEs share common 3' sequences: a review, *Gene*, **205**, 229–43.
12. Kapitonov, V.V. and Jurka, J. 2003, A novel class of SINE elements derived from 5S rRNA, *Mol. Biol. Evol.*, **20**, 694–702.
13. Ogiwara, I., Miya, M., Ohshima, K. and Okada, N. 1999, Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs, *Mol. Biol. Evol.*, **16**, 1238–50.
14. Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B. and Schmidt, T. 2011, Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes, *Plant Cell*, **23**, 3117–28.
15. Seibt, K.M., Wenke, T., Muders, K., Truberg, B. and Schmidt, T. 2016, Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization, *Plant J.*., **86**, 268–85.
16. Batzer, M.A. and Deininger, P.L. 2002, Alu repeats and human genomic diversity, *Nat. Rev. Genet.*., **3**, 370–9.
17. Schwichtenberg, K., Wenke, T., Zakrzewski, F., et al. 2016, Diversification, evolution and methylation of short interspersed nuclear element families in sugar beet and related Amaranthaceae species, *Plant J.*., **85**, 229–44.
18. Ponicsan, S.L., Kugel, J.F. and Goodrich, J.A. 2010, Genomic gems: SINE RNAs regulate mRNA production, *Curr. Opin. Genet. Dev.*, **20**, 149–55.
19. Roman, A.C., Benitez, D.A., Carvajal-Gonzalez, J.M. and Fernandez-Salguero, P.M. 2008, Genome-wide B1 retrotransposon binds the transcription factors dioxin receptor and Slug and regulates gene expression, *Proc. Natl. Acad. Sci. USA*, **105**, 1632–7.
20. Mariner, P.D., Walters, R.D., Espinoza, C.A., et al. 2008, Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock, *Mol. Cell.*, **29**, 499–509.
21. Lucas, B.A., Lavi, E., Shiue, L., et al. 2018, Evidence for convergent evolution of SINE-directed Staufen-mediated mRNA decay, *Proc. Natl. Acad. Sci. USA*, **115**, 968–73.
22. Ben-David, S., Yaakov, B. and Kashkush, K. 2013, Genome-wide analysis of short interspersed nuclear elements SINES revealed high sequence conservation, gene association and retrotranspositional activity in wheat, *Plant J.*., **76**, 201–10.
23. Wang, X., Xu, Y., Zhang, S., et al. 2017, Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction, *Nat. Genet.*., **49**, 765–72.

24. Huang, D., Wang, X., Tang, Z., et al. 2018, Subfunctionalization of the Ruby2-Ruby1 gene cluster during the domestication of citrus, *Nat. Plants*, **4**, 930–41.
25. Otto, D., Petersen, R., Brauksiepe, B., Braun, P. and Schmidt, E.R. 2014, The columnar mutation ("Co gene") of apple (Malus × domestica) is associated with an integration of a Gypsy-like retrotransposon, *Mol. Breeding*, **33**, 863–80.
26. Xu, Q., Chen, L.L., Ruan, X., et al. 2013, The draft genome of sweet orange (Citrus sinensis), *Nat. Genet.*, **45**, 59–66.
27. Shimizu, T., Tanizawa, Y., Mochizuki, T., et al. 2017, Draft sequencing of the heterozygous diploid genome of Satsuma (Citrus unshiu Marc.) using a hybrid assembly approach, *Front. Genet.*, **8**, 180.
28. Jiao, W.B., Huang, D., Xing, F., et al. 2013, Genome-wide characterization and expression analysis of genetic variants in sweet orange, *Plant J.*, **75**, 954–64.
29. Camacho, C., Coulouris, G., Avagyan, V., et al. 2009, BLAST+: architecture and applications, *BMC Bioinformatics*, **10**, 421.
30. Edgar, R.C. 2004, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, **5**, 113.
31. Mao, H. and Wang, H. 2017, SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets, *Bioinformatics*, **33**, 743–5.
32. Gouy, M., Guindon, S. and Gascuel, O. 2010, SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building, *Mol. Biol. Evol.*, **27**, 221–4.
33. Kumar, S., Stecher, G. and Tamura, K. 2016, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, *Mol. Biol. Evol..*, **33**, 1870–4.
34. Smit, A., Hubley, R. and Green, P. 2015, RepeatMasker Open-4.0. 2013–2015, *Institute for Systems Biology*, http://repeatmasker.org (18 April 2020, date last accessed).
35. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
36. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. 2004, WebLogo: a sequence logo generator, *Genome Res.*, **14**, 1188–90.
37. Voorrips, R.E. 2002, MapChart: software for the graphical presentation of linkage maps and QTLs, *J Hered.*, **93**, 77–8.
38. Lawrence, M., Huber, W., Pages, H., et al. 2013, Software for computing and annotating genomic ranges, *PLoS Comput. Biol.*, **9**, e1003118.
39. Backman, T. and Girke, T. 2016, systemPipeR: NGS workflow and report generation environment, *BMC Bioinformatics*, **17**, 388.
40. Rice, P., Longden, I. and Bleasby, A. 2000, EMBOSS: the European molecular biology open software suite, *Trends Genet.*, **16**, 276–7.
41. Krzywinski, M., Schein, J., Birol, I., et al. 2009, Circos: an information aesthetic for comparative genomics, *Genome Res.*, **19**, 1639–45.
42. Roy-Engel, A.M., Salem, A.H., Oyeniran, O.O., et al. 2002, Active Alu element "A-tails": size does matter, *Genome Res.*, **12**, 1333–44.
43. Gentles, A.J., Kohany, O. and Jurka, J. 2005, Evolutionary diversity and potential recombinogenic role of integration targets of non-LTR retrotransposons, *Mol. Biol. Evol.*, **22**, 1983–91.
44. Jurka, J. 1998, Repeats in genomic DNA: mining and meaning, *Curr. Opin. Struct. Biol.*, **8**, 333–7.
45. Huang, C.R., Burns, K.H. and Boeke, J.D. 2012, Active transposition in genomes, *Annu. Rev. Genet..*, **46**, 651–75.
46. Wu, G.A., Prochnik, S., Jenkins, J., et al. 2014, Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication, *Nat. Biotechnol..*, **32**, 656–62.
47. Grover, C.E. and Wendel, J.F. 2010, Recent insights into mechanisms of genome size change in Plants, *J. Bot.*, **2010**, 1–8.
48. Werren, J.H. 2011, Selfish genetic elements, genetic conflict, and evolutionary innovation, *Proc. Natl. Acad. Sci. USA*, **108**, 10863–70.
49. Hawkins, J.S., Proulx, S.R., Rapp, R.A. and Wendel, J.F. 2009, Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants, *Proc. Natl. Acad. Sci. USA*, **106**, 17811–6.
50. Zhang, Q., Arbuckle, J. and Wessler, S.R. 2000, Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize, *Proc. Natl. Acad. Sci. USA*, **97**, 1160–5.
51. Steinemann, M. and Steinemann, S. 1991, Preferential Y chromosomal location of TRIM, a novel transposable element of Drosophila miranda, obscura group, *Chromosoma*, **101**, 169–79.
52. Eibel, H. and Philippsen, P. 1984, Preferential integration of yeast transposable element Ty into a promoter region, *Nature*, **307**, 386–8.
53. Cappello, J., Cohen, S.M. and Lodish, H.F. 1984, Dictyostelium transposable element DIRS-1 preferentially inserts into DIRS-1 sequences, *Mol. Cell. Biol.*, **4**, 2207–13.
54. Feng, Q., Moran, J.V., Kazazian, H.H. and Boeke, J.D. 1996, Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition, *Cell*, **87**, 905–16.
55. Varshney, D., Vavrova-Anderson, J., Oler, A.J., Cowling, V.H., Cairns, B.R. and White, R.J. 2015, SINE transcription by RNA polymerase III is suppressed by histone methylation but not by DNA methylation, *Nat. Commun.*, **6**, 6569.
56. Jullien, P.E., Kinoshita, T., Ohad, N. and Berger, F. 2006, Maintenance of DNA methylation during the Arabidopsis life cycle is essential for parental imprinting, *Plant Cell*, **18**, 1360–72.
57. Fedoroff, N.V. 1995, DNA methylation and activity of the maize Spm transposable element, *Curr. Top. Microbiol. Immunol.*, **197**, 143–64.
58. Ding, Y., Wang, X., Su, L., et al. 2007, SDG714, a histone H3K9 methyltransferase, is involved in Tos17 DNA methylation and transposition in rice, *Plant Cell*, **19**, 9–22.
59. Peters, S.A., Datema, E., Szinay, D., et al. 2009, Solanum lycopersicum cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements, *Plant J.*, **58**, 857–69.
60. Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A. and Wendel, J.F. 2006, Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium, *Genome Res.*, **16**, 1252–61.
61. Devos, K.M., Brown, J.K. and Bennetzen, J.L. 2002, Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis, *Genome Res.*, **12**, 1075–9.
62. Hilgers, L., Hartmann, S., Hofreiter, M. and von Rintelen, T. 2018, Novel genes, ancient genes, and gene co-option contributed to the genetic basis of the radula, a molluscan innovation, *Mol. Biol. Evol.*, **35**, 1638–52.
63. Yan, Y., Zhang, Y., Yang, K., et al. 2011, Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice, *Plant J.*, **65**, 820–8.
64. Makarevitch, I., Waters, J., West, T., et al. 2015, Transposable elements contribute to activation of maize genes in response to abiotic stress, *PLoS Genet.*, **11**, e1004915.
65. Le, T., Schumann, U., Smith, N.A., et al. 2014, DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis, *Genome Biol.*, **15**, 458.
66. Shih, J., Hodge, R. and Andrade-Navarro, M.A. 2015, Comparison of inter- and intraspecies variation in humans and fruit flies, *Genomics Data*, **3**, 49–54.