Front. Agr. Sci. Eng. 2017, **4**(4): 421–432
https://doi.org/10.15302/J-FASE-2017160
Available online at http://engineering.cae.cn

RESEARCH ARTICLE

Repeats in the transcribed regions: comprehensive characterization and comparison of *Citrus* spp.

Manosh Kumar BISWAS[1], Christoph MAYER[2], Xiuxin DENG (✉)[1]

1 Key Laboratory of Horticultural Plant Biology, Ministry of Education/Huazhong Agricultural University, Wuhan 430070, China
2 Center of Molecular Biodiversity, Forschungsmuseum Alexander Koenig, Adenauerallee, Bonn 53113, Germany

Abstract A large number of expressed sequences tags are available for *Citrus* spp., which provides an opportunity to understand genomic organization of the transcribed regions. Here, we report a detailed analysis of repetitive elements including tandem repeats (TRs) and transposable elements (TEs) in the transcribed region of the *Citrus* spp. On average, 22% of the expressed sequence tags (ESTs) contain TRs. The relative density of TR classes is highly taxon-specific. For instance, *Citrus limonia* has a high relative density of mononucleotide repeats, whereas dinucleotide repeats are rare. The proportions of 2–6, 7–30 and 31–50 bp repeats were almost identical in all studied species except for *C. limonia* and *C. limettioides*. We found that $< 1\%$ of the citrus ESTs have a similarity with transposable elements. Transcriptional activity of transposable element families varied even within the same class of elements. A high proportion of transcriptional activity was observed for *gypsy*-like TEs compare to other TE classes. While TEs are relatively rare, TRs are abundant elements in ESTs of citrus. The high proportion of TRs that have a unit size longer than 6 bp raises the question about a possible functional or evolutionary role of these elements.

Keywords *Citrus* spp., tandem repeats, transcribed region, transposable elements

1 Introduction

Repetitive elements are abundant in plant genomes. They can be categorized into three main types; tandem repeats (TRs), transposable elements (TEs) and high-copy number genes. TRs are often divided into three subclasses according to the size of the repetitive unit: satellites (> 100 bp), minisatellites (7–100 bp) and microsatellites (1–6 bp)[1,2]. TRs frequently occur within or close to genes, i.e., in the untranslated regions (UTRs) up- and down-stream of open reading frames, within introns, or in coding regions (CDS)[3]. TRs appear in high densities in the centromeric, telomeric, and subtelomeric regions of many eukaryotes, comprising hundreds or thousands of repeats[4]. They are also found at interspersed positions and in low-recombining regions, such as sex or B chromosomes[1,5]. The reason why TRs are such ubiquitous elements in genomes is still not completely known. While originally classified as nonfunctional or junk DNA, more recent studies strongly hint at either a functional or evolutionary role[3,6–12]. The high mutation rates of TRs lead to their prominent role and their importance in many fields of molecular evolution[13,14]. They are used as informative molecular markers in population genetics and molecular breeding, in plants as well as in animals[15–18]. Besides TRs, TEs are another class of important repetitive elements that are particularly abundant in plant genomes. They are important in genome and gene evolution[19]. Their main characteristic is their ability to move or copy themselves within the genome[20]. They are divided into two classes, RNA-mediated Class I retrotransposons and DNA-mediated Class II transposons. Both classes contain elements that encode functional products required for transposition (autonomous) and elements that only retain the *cis* sequences necessary for recognition by the transposition machinery (non-autonomous). Class I elements can further be divided into several subclasses: SINEs, LINEs, long-terminal repeat (LTR) retrotransposons and terminal-repeat retrotransposons in miniature, which are LTR non-autonomous elements[21]. Class II elements comprise autonomous and non-autonomous transposons, including MITEs (miniature inverted-repeat transposable elements). TEs can serve as a very rich source of identifiable polymorphisms. Some studies suggest that

Received October 11, 2016; accepted April 10, 2017

Correspondence: xxdeng@mail.hzau.edu.cn

TEs might even be more useful as molecular markers (e.g., SSAP, IRAP or REMAP markers), in particular in plant breeding application, than other markers (e.g., SSR and AFLP)[22,23].

Citrus is one of the most popular fruit crops worldwide with great economic and health value. It grows throughout the tropical and subtropical regions of the world. The major citrus producing areas are in south and east Asia (led by China, India and Japan), Americas (led by Brazil, USA, Mexico and Argentina) and the Mediterranean basin (led by Spain, Italy, Egypt and Turkey)[24]. Although citrus is one of the most important fruit crops, its genome has been much less explored than other plant species (e.g., rice, maize and soybean). The knowledge of repetitive sequence elements is essential for understanding the nature and consequences of genome size variation between different species, and the large-scale organization and evolution of plant genomes[1]. Several methods have recently been developed for the analysis of repetitive sequence elements in genomes[1,2,25–30]. Expressed sequence tag (EST) databases are valuable resources for predictions regarding genome structure and genomic organization in the transcribed regions of genomes. The large number of publicly available citrus EST sequences offers the great possibility to study transcribed regions in citrus genomes. The analysis of repetitive elements in citrus ESTs will facilitate and provide valuable information when studying highly important questions concerned with a genetic improvement of citrus. They will also provide a valuable resource for the development of genetic tools such as molecular markers. Several studies have been conducted in the past to find repetitive elements in genomes of many plant species, including papaya[1], maize[31], soybean[32]. Although several studies have already analyzed citrus ESTs and characterized microsatellite to develop SSR[33,34], most details concerning the repeat characteristics such as minisatellite, satellites and TEs found in the transcribed regions of citrus remain unexplored. Thus a detailed structural analysis of the transcribed regions of citrus genomes remains to be performed. In this study we screened clustered non-redundant EST data sets of 11 *Citrus* spp. for TRs and TEs with the aim to understand the genomic organization in the transcribed regions of citrus. For TRs we compared the densities and length characteristics of different repeat types and unit size ranges. TEs were classified and frequencies were computed. For selected TEs, we also estimated phylogenetic distances.

2 Materials and methods

2.1 Sequences retrieved and processing

EST sequences of the 11 *Citrus* spp. were retrieved from NCBI (http://www.ncbi.nlm.nih.gov) on November 14, 2015 (Table 1). A Perl script est_trimmer.pl (http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl) was used to remove unusual EST sequences, vector contamination, poly-A and poly-T bases from the EST sequences. After that, the CAP3 program (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::cap3) was used to obtain non-redundant EST sequences. The 11 sets of non-redundant sequences were used for subsequent data mining.

2.2 Tandem repeat detection

TRs were detected in the citrus EST data sets by using the Tandem Repeats Finder (TRF) software[25] and PHOBOS, version 3.2.6[35]. Both programs have been used to search for imperfect TRs in a unit size range from 1 to 1000 bp without using a pre-specified motif library. TRF was used with default parameters. PHOBOS used the alignment scores 1, – 5, – 5, 0, for match, mismatch, gap and N positions. In every TR, the first repeat unit was not scored. Only a maximum number of four successive Ns were allowed. For a TR to be considered in the analysis it was required to have a minimum repeat alignment score of 12, if its unit size was less or equal to 12 bp, or a score of at least the unit size for unit sizes above 12 bp. As a consequence, mono-, di- and tri-nucleotide repeats were required to have a minimum length of 13, 14 and 15 bp, respectively, to achieve the minimum score. For repeat units above 12 bp, a perfect repeat had to be at least two units long and an imperfect repeat even longer, to obtain the minimum score.

All TRs with units that differ only by circular permutations and/or the reverse complement are associated to the same repeat type. As a result most tandem repeats and their complementary counterparts can be represented by several different basic unit patterns. Clearly, there are always several repeat units which belong to the same repeat type. For example, the pattern (GCC)*n*, also represents (CCG)*n*, (CGC)*n*, (GGC)*n*, (GCG)*n*, and (CGG)*n*. The convention allows counting and identifying repeat units without reference to the repeat unit phase or strand[2]. In this study, we follow the convention to represent a repeat type by that unit which comes first in an alphabetical ordering of all units that are associated to it[36]. For example, the repeat type represented by the unit AAG incorporates all TRs with units AAG, AGA, GAA, TTC, TCT and CTT. Furthermore, TR patterns are always listed under the smallest possible unit size. For example, patterns like (ACACAC)*n* or (ACAC)*n* were included into the category (AC)*n*. As a result the total number of theoretically possible, non-overlapping patterns was reduced. Finally, the term repeat type is distinguished from the term repeat class which we use to denote the collection of all repeats with the same repeat unit size (e.g., mono-, di-, tri-nucleotide repeats). TR characteristics such as the density and mean length of repeat types were computed using the program Sat-Stat, version 1.3.1 (http://www.ruhr-uni-bochum.de/spezzoo/cm/). Different TR

characteristics have been analyzed in this study. These are (1) the TR density measured in bp/Mbp, which gives the proportion of bp found in repeats with respect to all bp in the sequence, (2) the number of repeats found on average in a sequence of a certain length measure in TR/Mbp, and (3) mean length of repeats measured in bp.

2.3 Transposable element analysis

To find TE in the transcribed region of the citrus genome, we used a combination of homology-based and *de novo* methods. Given that there are many known families of TEs in plants, homology-based methods should be highly effective in identifying and annotating them. We built a custom plant TE library in combination of plant repeats from Repbase[30], plant repeat databases from TIGR (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats) and GeneBank for our initial classification of TEs (Table S1). Repeat elements identified as rRNA sequences, centromere-related sequences, telomere-related and unclassified sequences in the TIGR databases were excluded from our repeat library, leaving a database of 6880 repeat sequences that were used to search the transcribed region of the citrus ESTs. Then customized plant TE databases were compared with the citrus EST data sets using BLASTN analysis. BLASTN analyses were performed using an expected threshold of 10, a word size of 11, a match/mismatch of two to three and gap cost existence of five and extension of two. We only considered search hits with an e-value $< 1\times10^{-5}$. A Perl script was developed for summarizing the results.

2.4 Phylogenetic analysis

Homolog TE sequences were retrieved from citrus EST data sets with the aid of BLASTN searches and using an in house developed Perl script. *Copia*- and *gypsy*-like TE-EST sequences were pooled separately with randomly selected citrus c*opia*- and *gypsy*-like genomic sequences. Sequences were aligned and trees were constructed with MEGA5[37].

3 Results

In this study we analyzed 525510 clean ESTs from 11 *Citrus* spp. After a CAP3 assembly of each data set, the number of sequences reduced to 200968 unigenes. Therefore, about 61% of the citrus ESTs are redundant in the EST databases. Each unigene data set was used for further TR and TE analyses and the main results are summarized in Table 1. The percentage of TR containing ESTs was roughly identical among the studied species except for *Citrus unshiu* and *C. paradisi*. The highest number of TEs was recorded for *Citrus sinensis,* while the lowest was found in *Citrus limittioides*. Overall, less than 1.5% of the ESTs contain a known TE element.

3.1 Tandem repeat analysis

3.1.1 Characteristic of TR in all 11 *Citrus* genomes

Transcribed regions of the 11 *Citrus* spp. were searched for TRs. On average, 22% of the analyzed EST sequences contain one or more TR loci and for most species the fraction of TR containing repeats is relatively close to this value. Details are shown in Table 1. The highest number is found for *Citrus limettioides* (32%) and the lowest for *C. paradisi* (8%). We plotted TR densities against the size of the EST data set (which approximates the size of the

Table 1 List of *Citrus* spp. and the number of sequences analyzed in the present study together with basic characteristics of these sequences

Species	No. of sequences analyzed*	Total length of sequences /bp	CG content /%	No. of TR containing EST/%	No. of TE containing EST/%	TR frequency (TR/Mbp)	TE frequency (TE/Mbp)	TR coverage /%
Citrus sinensis	70917	63428301	45.24	16153 (22.78)	1819 (2.56)	2439	29	6.89
Citrus clementina	24201	22433389	44.62	4930 (20.37)	546 (2.26)	2299	24	5.98
Citrus trifoliata	25388	22799411	46.47	6107 (24.05)	262 (1.03)	2646	11	7.86
Citrus reticulata	29422	26697646	47.51	6736 (22.89)	188 (0.64)	2410	7	6.91
Citrus unshiu	8328	4826095	41.92	861 (10.34)	77 (0.92)	1802	16	4.48
Citrus auruntium	10260	8244717	47.97	2112 (20.58)	30 (0.29)	2469	4	6.90
Citrus limonia	7895	6627791	44.80	1943 (24.61)	60 (0.76)	3030	9	8.32
Citrus latifolia	7173	6313905	48.56	1635 (22.79)	53 (0.74)	2488	8	6.69
Citrus aurantifolia	5977	5093917	49.46	1574 (26.33)	13 (0.22)	3003	3	8.64
Citrus limettioides	7049	6393162	44.62	2257 (32.02)	18 (0.26)	3401	3	9.05
Citrus paradisi	4358	2549165	41.78	385 (8.83)	17 (0.39)	1608	7	4.48
Total	200968	175407499	45.72	44693 (22.24)	3083 (1.53)	2488	18	6.93

Note: *, Unigene sequences (CAP contigs and singlets); TR, tandem repeat; TE, transposable element; TR coverage = number of bp in repeats over number of bp in sequences.

transcribed region). The TR densities vary only slightly among the studied *Citrus* spp. No significant correlation was found between the size of the EST data set and the density of TRs (Fig. 1a, $r = 0.05$, $P < 0.1$). A comparison of the mean lengths of TRs of all 11 genomes shows that TRs are shortest in *C. paradisi* (average length 9.22 bp) and longest in *Citrus trifoliata* (average length 99.74 bp). Again, no significant correlation between the size of the EST data set and the mean length of TRs was found (Fig. 1b, $r = 0.319$, $P < 0.1$). A comparison of TR densities of the different repeat classes is given in Fig. 1c. The result shows that the relative densities of different repeat classes are considerably taxon-specific. For example, *Citrus limonia* has a high relative density of mononucleotide repeats, whereas dinucleotide repeats are rare. The proportion of di-, tetra-, penta-, hexa-nucleotides, 7–30 and 31–50 bp repeats are very similar in all the studied species except for *C. limonia,* and *C. limettioides*.

TRs were classified into three unit size ranges, namely microsatellites (1–6 bp), minisatellites (7–100 bp) and satellites (>100 bp). Results for the different unit size ranges are given in Table 2. As expected, micro- and minisatellites are more abundant than satellites in the transcribed regions of *Citrus* spp. The highest densities of TRs are recorded in *C. sinensis,* while lowest densities are found in *C. paradisis*. The density of micro- and minisatellites in the transcribed regions are taxon specific. A high abundance of microsatellites was found in the genomes of the *Citrus* spp., *C. sinensis, C. trifoliata, C. reticulata, C. auruntium, C. latifolia, C. aurantifolia* and *C. limettioides*, and a high abundance of minisatellites was found in the ESTs of the *Citrus* spp., *C. clementina, C. unshiu, C. limonia* and *C. paradisi*. In total, minisatellites contribute more to the TR coverage than microsatellites.

3.1.2 Genomic densities of mono- to tri-nucleotide repeat types

Repeat type usage of mono-, di-, and tri-nucleotide repeats in the 11 genomes are summarized in Table 3. It is shown that the repeat type usage in ESTs varies strongly between taxa. Even among more closely related *Citrus* spp., only few common features can be observed. For example, the density of ACT, ACG and CG repeats is consistently low in all species. The repeat types AG, AT, AAG and AAT have high densities in all species. The densities of poly-C repeats are generally high, except for *C. unshiu* and *C. paradisi*, where they are even lower than poly-A repeat densities. Poly-A repeats have the highest density in *C. sinensis* among the 11 species.

3.1.3 Characteristics of tandem repeats with unit sizes 1–50 bp in expressed sequence tags of all 11 *Citrus* spp.

Most previous studies only analyzed TR characteristics in the unit size range 1–6 bp. In this study we compared the TR characteristics in ESTs of 11 species in three unit size ranges, namely 1–6, 7–10, and 11–50 bp. Our results show that the density of TRs with a unit size in the range 7–50 bp contributes significantly to the total repeat density in the unit size range 1–50 bp (Fig. 2). The relative contribution ranges between 17.6% in *C. limonia* and 42.9% in *C. paradisi* with a mean value of 31.9%. Among the 11 EST data sets, strong differences are found also for individual repeat classes (Fig. 2; Fig. S1). TR densities in *C. sinensis*, *C. clementina*, *C. trifoliata* are slightly below average. Mono repeats represent the dominant repeat class followed by tri- and di- repeats in ESTs of all 11 *Citrus* spp. For the longer repeat units, there are usually only very few repeat types which contribute to the density of their repeat classes. A comparison of the longest repeat length and mean repeat length is presented in Fig. S2. This analysis reveals a strong difference between the mean length of TRs among different repeat classes and species. A maximum mean repeat length of 370 bp is found for the 48 bp repeat class in ESTs of *C. aurantium*, which consists of two repeats of length 117 bp and 623 bp. All mean repeat lengths are shorter than 200 bp in the unit size range 1–50 bp for all citrus ESTs except for *C. clementina* and *C. aurantium.*

3.2 Transposable elements

The availability of a large amount of EST sequences provides an opportunity to estimate the transcriptional activity of transposable elements. In this study we used custom TE data sets to query BLASTN against the citrus EST database. The results reveal that 1.53% of the total citrus ESTs (3083 sequences) showed significant sequence homology (e-value $< 1\times10^{-5}$) with one of the TE families (Table 1). It has been found that Class I (RNA-mediated) elements are more abundant than Class II (DNA-mediated) elements (Fig. 3a) in all studied *Citrus* spp.

Among the different TE families, *gypsy*-like elements are most frequent in the citrus EST database (37% of the total TE-ESTs), followed by *copia*-like LTR retrotransposons, while SINE elements are least abundant (0.13%). Comparing TE copy numbers in ESTs with respect to different families, *gypsy* elements are almost four times as frequent as *copia*-like elements (Fig. 3c). There is no significant correlation found between the copy numbers of each TE family and the numbers of ESTs (Fig. 3d). In citrus we found TE families with a large number of family members which were found only in a few ESTs and TE families with a low number of members found in many ESTs. For example, the SINE elements have 664 family members in the database we searched, while we only identified a homology with three ESTs of citrus. In contrast, the 925 different Ty3-*gypsy* elements in the database could be found in 1133 ESTs.

Hundreds to thousands copies of TEs are found in the

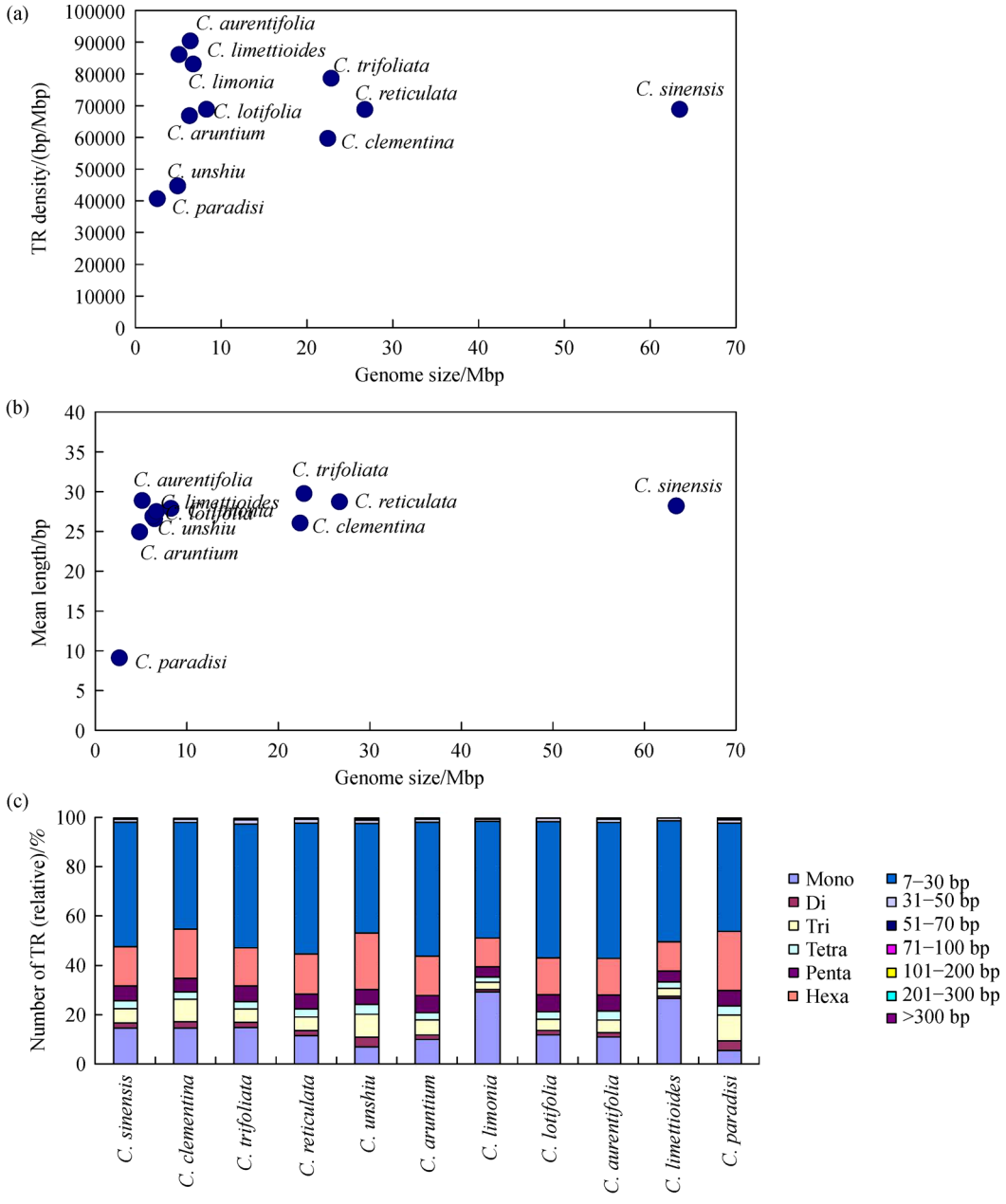

Fig. 1 Density and length characteristics of tandem repeats (TRs) in the transcribed regions of 11 *Citrus* spp. (a) TR density versus size of the expressed sequence tag (EST) data set; (b) mean repeat length versus size of the EST data set; (c) relative frequency of TR classes.

genome; but the question is how many of these are transcriptionally active. To find the answer to this question we constructed a phylogenetic tree based on EST sequences and randomly selected genomic sequences of the TE families of citrus (Fig. 4). The phylogenetic analysis indicated that transcriptionally active TEs are found in distinct clades, and very few are shared with genomic sequence based TEs. These findings indicated that few evolutionary branches of the TE family have retained transcriptional capability.

4 Discussion

4.1 Tandem repetitive elements

Tandem repeats are one of the most common elements in plant genomes and they are key for understanding genome organization and evolution. Available EST or GSS sequences provide an opportunity to study TR elements in transcribed regions of genomes. Although many studies have been conducted for TRs in plant genomes, few have

Table 2 Summary statistics of the tandem repeat in the transcribed regions of 11 citrus genomes

Species	Microsatellite (1–6 bp)					Minisatellite (7–100 bp)					Satellite (> 100)				
	NL	ANRL	Var.	Den.	Cov.	NL	ANRL	Var.	Den.	Cov.	NL	ANRL	Var.	Den.	Cov.
Citrus sinensis	1164	13	502	26611	2.66	1271	3	31609	40558	4.06	4	3	246	1707	0.17
C. clementina	1254	10	484	25738	2.57	1039	3	11414	31960	3.20	5	3	94	2079	0.21
C. trifoliata	1242	14	485	30582	3.06	1396	3	15612	45889	4.59	5	3	101	2120	0.21
C. reticulata	1076	10	491	23002	2.30	1327	3	17248	44461	4.45	3	3	91	1602	0.16
C. unshiu	956	7	405	18391	1.84	840	2	2873	24257	2.43	6	2	29	2134	0.21
C. auruntium	1086	9	457	22786	2.28	1379	3	6865	42890	4.29	6	3	51	3282	0.33
C. limonia	1550	17	394	36353	3.64	1474	3	4942	44002	4.40	6	3	41	2858	0.29
C. latifolia	1073	11	406	22791	2.28	1414	3	5245	42758	4.28	3	2	20	1359	0.14
C. aurantifolia	1291	11	410	29151	2.92	1703	3	5178	55022	5.50	5	3	24	2189	0.22
C. limettioides	1690	14	399	35493	3.55	1710	3	5325	53125	5.31	4	3	24	1873	0.19
C. paradise	866	6	338	5720	1.64	736	2	1511	7413	2.10	7	3	18	1703	0.32
Over all					2.65					4.12					0.19

Note: NL, No. of TR/Mbp; ANRL, average number of repeat units/locus; Var., variants; Den., density (bp/Mbp); Cov., coverage (%).

Table 3 Density (TRs/Mbp) of different TR types from mono- to tri-nucleotides repeats for EST data sets of all 11 *Citrus* spp.

Repeat type	*C. sinensis*	*C. clementina*	*C. trifoliata*	*C. reticulata*	*C. unshiu*	*C. auruntium*	*C. limonia*	*C. latifolia*	*C. aurantifolia*	*C. limettioides*	*C. paradisi*	Min	Max
A	197.1	175.1	217.2	112.4	83.9	101.4	500.3	157.3	130.2	481.9	55.3	55.3	500.3
C	82.1	89.2	99.3	96.7	3.3	78.1	230.8	70.3	107.6	248.4	4.7	3.3	248.4
AC	4.7	5.5	4.4	3.6	6.0	3.9	2.9	3.3	3.1	1.3	7.1	1.3	7.1
AG	17.7	28.1	16.4	17.6	27.1	14.4	7.4	11.4	16.3	13.6	24.7	7.4	28.1
AT	9.4	9.2	9.5	7.5	13.9	6.5	8.0	6.5	11.2	4.1	14.5	4.1	14.5
CG	0.4	0.1	0.4	0.3	0.0	0.7	0.0	0.2	1.0	0.0	0.0	0.0	1.0
AAC	3.2	6.2	3.6	3.0	4.1	3.0	1.7	1.9	4.5	0.8	4.7	0.8	6.2
AAG	12.2	17.3	11.4	11.3	14.9	10.9	8.8	8.4	10.0	8.9	14.5	8.4	17.3
AAT	9.6	12.4	8.2	7.0	11.8	7.4	5.4	4.1	6.7	4.5	14.5	4.1	14.5
ACC	3.4	7.4	3.5	2.9	2.7	3.5	0.6	2.9	3.7	2.7	4.3	0.6	7.4
ACG	1.2	3.2	1.7	0.9	2.3	1.0	0.6	0.5	1.2	1.1	0.8	0.5	3.2
ACT	0.4	1.1	0.6	0.3	0.6	0.4	0.3	1.0	0.6	0.2	1.6	0.2	1.6
AGC	8.0	11.5	6.9	8.5	6.8	10.2	3.8	7.3	8.2	6.1	11.4	3.8	11.5
AGG	3.5	6.9	3.0	4.5	5.2	3.2	0.8	1.4	2.9	3.4	2.4	0.8	6.9
ATC	5.1	8.2	5.0	4.2	7.5	3.8	4.8	2.1	5.7	3.6	8.2	2.1	8.2
CCG	2.2	7.0	3.4	2.7	3.3	2.9	1.5	2.1	1.6	1.1	3.1	1.1	7.0

studied these elements in citrus. Furthermore, most studies are restricted to TRs in the unit size of 1–6 bp. In particular very little is known about TR elements in transcribed regions. In this study, we analyzed and compared the TR content in the transcribed region of 11 *Citrus* spp. in three unit size ranges: 1–6bp (microsatellites), 7–100 bp (minisatellites) and > 100bp (satellites). Our results reveal that on average 6.93% of each EST sequence is covered with TRs and that a significant proportion of this coverage is contributed by minisatellites, with their contribution being almost two-fold the contribution of microsatellites and 22-fold the coverage contribution of satellites (Table 2). This finding suggests that both microsatellites and minisatellites play a role in organization and function of the transcribed regions of citrus.

Several studies have shown that TRs are generally non-randomly distributed in genomes[2,3,38]. Exceptions have been reported for example for the papaya genome, where TRs are more or less randomly distributed[1]. We found that several TR characteristics are non-randomly

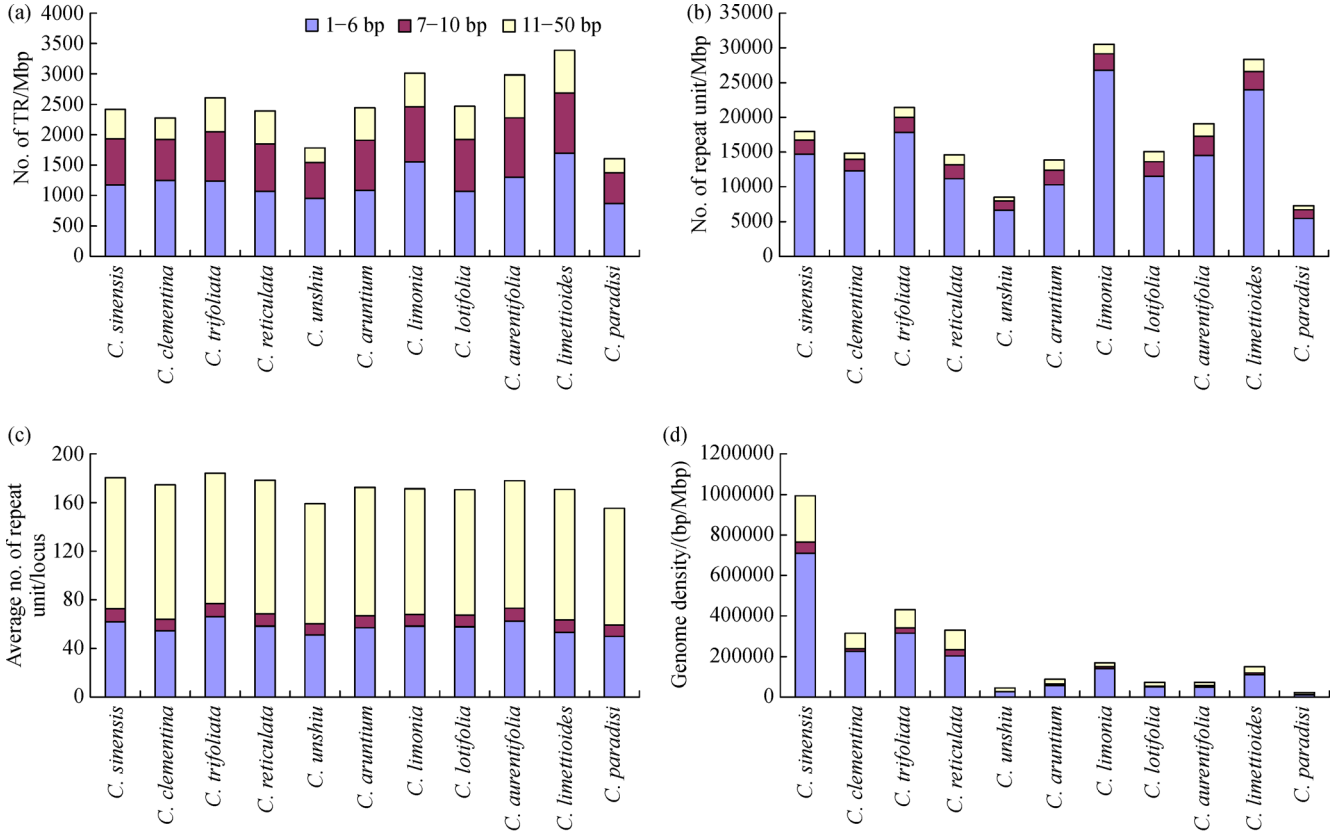

Fig. 2 Comparative features of the distribution of TRs in the transcribed regions of 11 citrus species. (a) Number of Tandem Repeats (TRs) per Mbp; (b) number of different units found per Mbp genome; (c) average no of repeat units per locus; (d) genomic density of TRs in the three different unit size ranges 1–6 bp, 7–10 bp and 11–50 bp for 11 *Citrus* spp.

distributed in the transcribed regions of citrus genomes. No significant correlation was found between the size of the EST data set and TRs densities or length characteristics in this study, which is consistent with results obtained for complete genomes, e.g.. Tautz et al.[39], Tóth et al.[40] and Victoria et al.[41]. Except for the relatively low densities of ACT, ACG and CG repeats and high densities for AG, AT, AAG and AAT repeats, no TR characteristics were found to be common in all 11 citrus genomes. This result is in agreement with the comparative genomic analyses of a wide range of plant groups reported by Tóth et al.[40]. The dominance of taxon rather than group specific characteristics has also been reported for *Arabidopsis*, barley, rice and wheat[42,43] when comparing number counts of satellites or when considering densities[2]. Evidence suggested that ACG repeats are underrepresented in most eukaryotic genomes[40]. The only known counter example among green plants is the algae, *Ostreococcus lucimarinus,* which has a particularly high density of ACG repeats[2]. Usually, CG, ACG and CCG repeat densities are low in higher plants. This is generally attributed to the fact that methylated CpG dinucleotides are highly mutable, which disrupts CpG rich domains on short timescales[2,40]. However, other mechanisms have also been proposed; see Tóth et al.[40]. Low densities of CCG repeats were also found in the genomes of *C. clementina*, *Brassica* and yeast. According to our result CG, ACG and CCG repeats have low abundances in the transcribed regions of all citrus genomes (Table 3).

Notably, the high absolute and relative di- and tri-nucleotide repeat densities found in *C. sinensis* are almost exclusively based on the high densities of the AG, AAG and AGC repeat types that are also common in all other *Citrus* spp. in this study (Table 3). Victoria et al.[41] reported that AG and AAG repeats generally predominate among di- and tri-nucleotide repeats in higher plants. Several studies demonstrated that poly-A repeats are more frequent than poly-C repeats in almost all vascular plants, which was also found in the present study[44]. As a general trend and except for the features just mentioned, we find that common TR characteristics are rare. We also observed that the length of TRs did not correlate with the repeat unit size.

4.2 Transposable elements

TEs are a major component and important for the physical structure of many plant genomes. Several studies show that

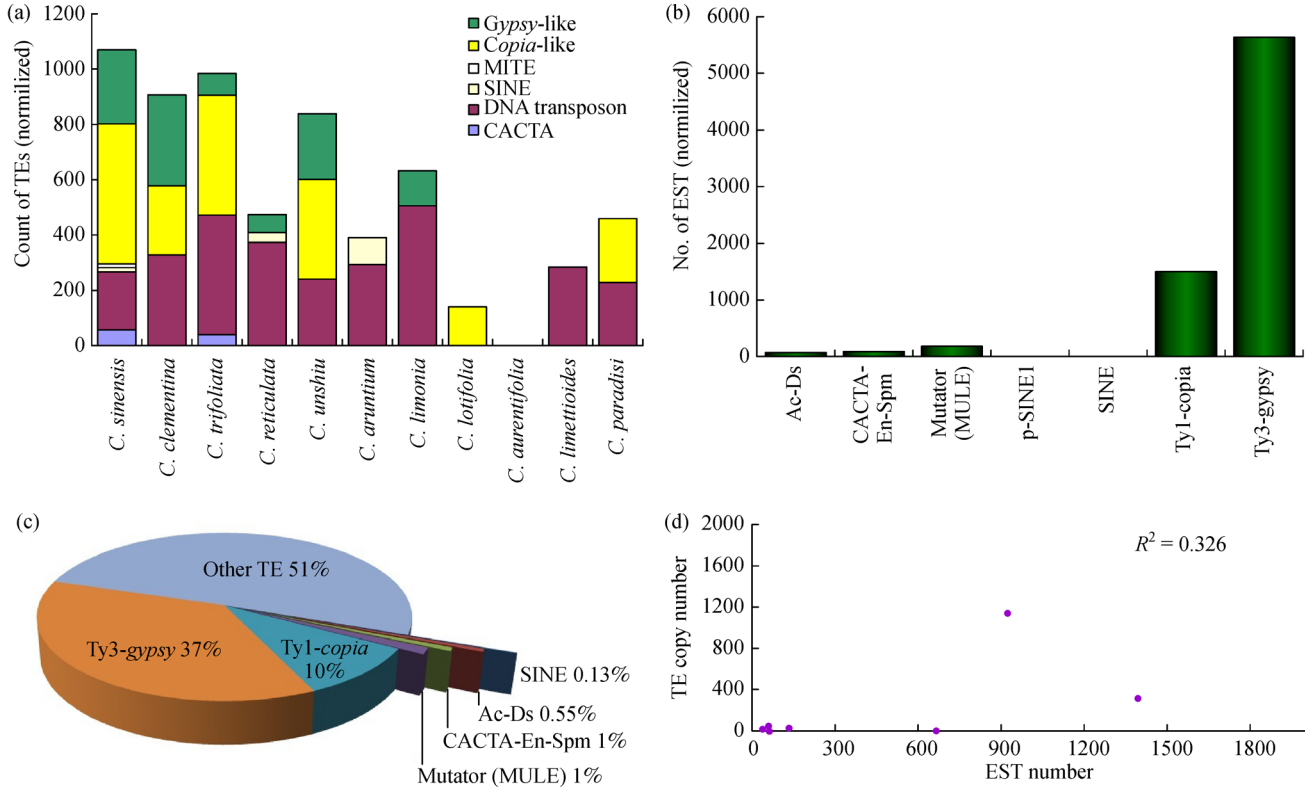

Fig. 3 (a) Number of expressed sequence tags (ESTs) similar to transposable element (TE) families for 11 *Citrus* spp.; (b) number of ESTs similar to TE families in the collection of all ESTs from 11 *Citrus* spp.; (c) percentage distribution of EST among major class of TEs; (d) dot-plot correlation between TE copy number and number of EST similar to the TE.

TEs can account for as much as 80% to 90% of plant genomes[45], and that some TEs are transcriptionally active. To understand the impact of retrotransposition on plant genome evolution, it is important to identify active members of TE families that are present with high copy numbers[45]. Once TEs accumulate and degrade in a genome, they usually become functionally inactive. However, partial or rearranged TE copies may retain their ability to initiate transcription[19]. Cells have active mechanisms to protect the integrity of their genomes against TE activity by transcriptional silencing[46]. Under certain circumstances, some TEs can escape this cell control with the result that they are able to get transcribed and transposed[47]. This phenomenon is frequently observed under biotic or abiotic stress or in cell cultures[48–51]. Consequently, TE transcripts were more abundant in cDNA libraries obtained from stress treated tissue. Thus, the presence of TE transcripts in cDNA libraries can be expected, and EST databases can be used to identify functionally active TEs in genomes. Here we searched citrus EST databases in order to identify transcriptionally active TE families in citrus genomes. We identified Ac-Ds, CACTA, SINE, MITE, *copia*-like and *gypsy*-like TE families as functionally active in citrus genomes. Previous work suggests that *copia*- and *gypsy*-like TE families are highly abundant in citrus genomes and that some of the members of these families are transcriptionally active[52,53]. Our work does not fully support these findings, since the number of ESTs that originated from TEs was low. Since TEs are not necessarily located in transcribed regions, we cannot conclude that the overall TE content in citrus genomes is low. Our analysis suggests that the ratio of *gypsy*- to *copia*-like elements in transcribed regions of the citrus genomes is closer to 3:1. This suggests that *gypsy* TE families are either more frequent or transcriptionally more active than *copia* families. *Gypsy* elements were also found to dominate over other TE families in maize EST data sets[19]. A primary analysis of the sweet orange genome reveals that 20% of the genome contains transposable elements and that *gypsy* elements are predominant[54]. Although we found a low level of transcriptional activity of TEs (about 1.5% of all ESTs) in citrus genomes this is compatible with a similar study in maize[19].

There was no correlation found between the number of TEs of a given family in the database and the number of ESTs they were found in (R^2 = 0.326, Fig. 3d). Meyers et al.[31] reported TE numbers were negatively correlated

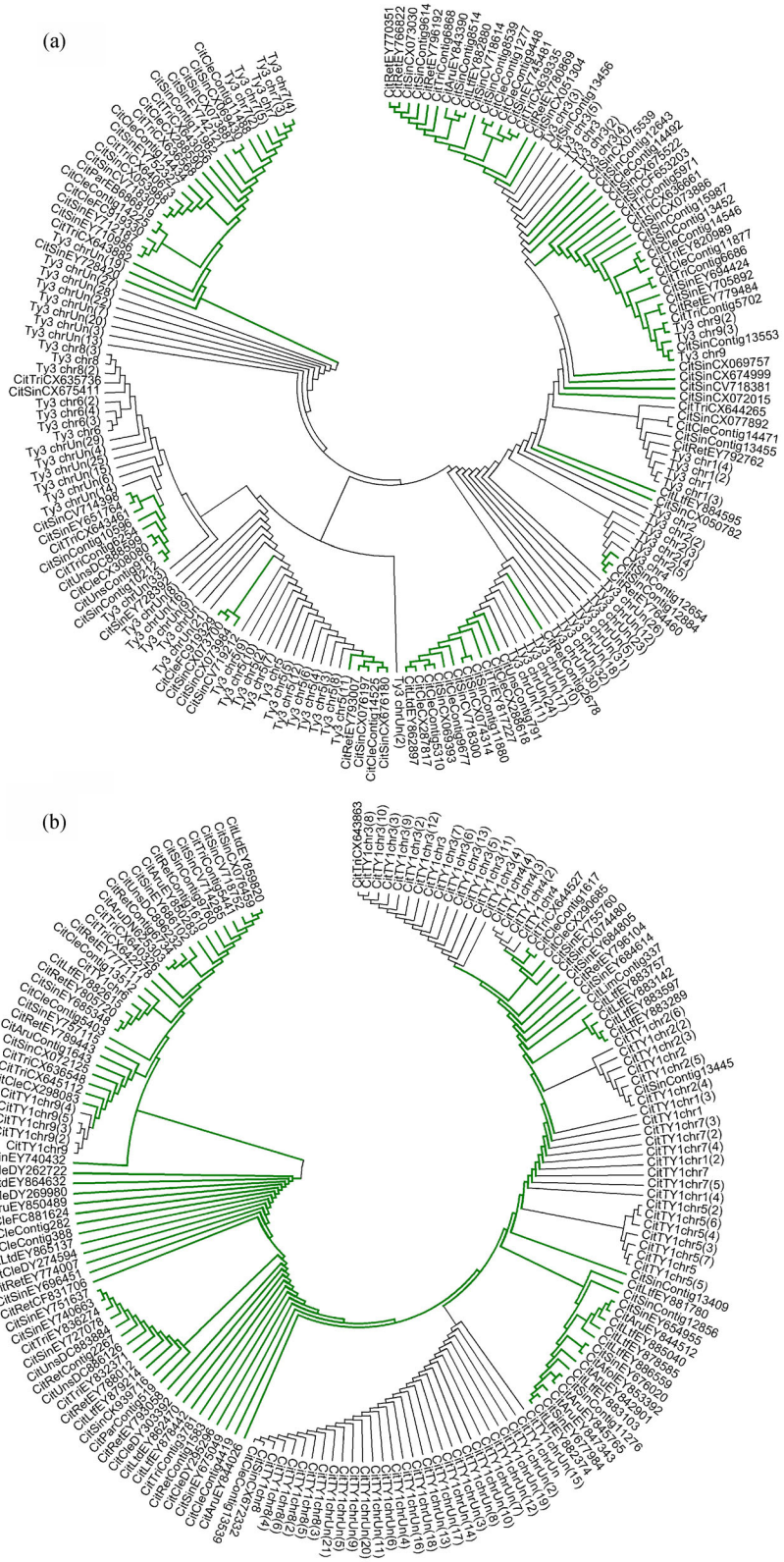

Fig. 4 Phylogenetic analysis of long-terminal repeat-retrotransposon sequences in citrus expressed sequence tag and genomic sequences. (a) Phylogenetic tree based on *copia*-like sequences; (b) phylogenetic tree based on *gypsy*-like sequences.

with the EST database size in Maize, while Vicient[19] did not find any correlation in maize ESTs and TEs. Comparison of TE family members in the database with the number of occurrences of TEs in the EST database and phylogenetic analysis of TEs suggested that high-copy retroelements are transcriptionally less active than low-copy number retroelements in *Citrus* spp. Similar findings were also reported for maize. Rabinowicz et al.[55] found that high-copy retroelements were frequently located outside of hypomethylated regions of the genome, while low-copy were located inside the hypomethylated region of the genome. Consequently low-copy retroelements families may escape methylation and therefore they are transcriptionally active.

5 Conclusions

TRs are abundant element in transcribed regions of citrus genomes, with 22% of the citrus ESTs containing a TR and 7% of the EST sequences covered by TRs. Notably, TRs with a unit size longer than 6 bp contributed significantly to the TR content. TE abundance is rather low in ESTs of citrus; where on average 1.5% of the transcripts are derived from citrus transposable elements. TEs found in ESTs are assumed to be transcriptionally active members of TE families and it would be worthwhile to study their role in gene expression and citrus genome evolution. TR and TE abundance varies strongly from species to species, with very minor common features among species.

Supplementary materials The online version of this article at http://dx.doi.org/10.15302/J-FASE-2017160 contains supplementary materials (Table S1; Fig. S1; Fig. S2).

Acknowledgements This research was financially supported by the Ministry of Science and Technology of China (2011CB100600, 2011AA100205) and the National Natural Science Foundation of China (NSFC). Authors are grateful to the CSC (China Scholarship Council) and China Post Doc Council for providing the Fellowships.

Compliance with ethics guidelines Manosh Kumar Biswas, Christoph Mayer, and Xiuxin Deng declare that they have no conflicts of interest or financial conflicts to disclose.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. Niranjan N, Navajas-Pérez R, Mihai P, Alam M, Ming R, Andrew H P, Steven L S. Genome-wide analysis of repetitive elements in papaya. *Tropical Plant Biology*, 2008, **1**(3): 191–201
2. Mayer C, Leese F, Tollrian R. Genome-wide analysis of tandem repeats in Daphnia pulex—a comparative approach. *BMC Genomics*, 2010, **11**(1): 277
3. Li Y C, Korol A B, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. *Molecular Biology and Evolution*, 2004, **21**(6): 991–1007
4. Ugarković D, Plohl M. Variation in satellite DNA profiles–causes and effects. *EMBO Journal*, 2002, **21**(22): 5955–5959
5. Camacho J P, Sharbel T F, Beukeboom L W. B-chromosome evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 2000, **355**(1394): 163–178
6. Buard J, Jeffreys A J. Big, bad minisatellites. *Nature Genetics*, 1997, **15**(4): 327–328
7. Kashi Y, King D, Soller M. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics*, 1997, **13**(2): 74–78
8. Schlötterer C. Evolutionary dynamics of microsatellite DNA. *Chromosoma*, 2000, **109**(6): 365–371
9. Li Y C, Korol A B, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, 2002, **11**(12): 2453–2465
10. Riley D E, Krieger J N. Diverse eukaryotic transcripts suggest short tandem repeats have cellular functions. *Biochemical and Biophysical Research Communications*, 2002, **298**(4): 581–586
11. Riley D E, Krieger J N. Short tandem repeats are associated with diverse mRNAs encoding membrane-targeted proteins. *BioEssays*, 2004, **26**(4): 434–444
12. Kashi Y, King D G. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics* , 2006, **22**(5): 253–259
13. Dieringer D, Schlötterer C. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Research*, 2003, **13**(10): 2242–2251
14. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, 2004, **5**(6): 435–445
15. Jeffreys A J, Neumann R, Wilson V. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell*, 1990, **60**(3): 473–485
16. Bonhomme F, Rivals E, Orth A, Grant G R, Jeffreys A J, Bois P R. Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. *Genome Biology*, 2007, **8**(5): R80
17. Qiu L, Yang C, Tian B, Yang J B, Liu A. Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). *BMC Plant Biology*, 2010, **10**(1): 278
18. Studer B, Kölliker R, Muylle H, Asp T, Frei U, Roldán-Ruiz I, Barre P, Tomaszewski C, Meally H, Barth S, Skøt L, Armstead I P, Dolstra O, Lübberstedt T. EST-derived SSR markers used as anchor loci for the construction of a consensus linkage map in ryegrass (*Lolium* spp.). *BMC Plant Biology*, 2010, **10**(1): 177
19. Vicient C M. Transcriptional activity of transposable elements in maize. *BMC Genomics*, 2010, **11**(1): 601
20. Kidwell M G, Lisch D. Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences of the United States of America*, 1997, **94**(15): 7704–7711
21. Wicker T, Sabot F, Hua-Van A, Bennetzen J L, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P,

Schulman A H. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 2007, **8**(12): 973–982

22. Biswas M K, Chai L, Amar M H, Zhang X, Deng X X. Comparative analysis of genetic diversity in *Citrus* germplasm collection using AFLP, SSAP, SAMPL and SSR markers. *Scientia Horticulturae*, 2011, **129**(4): 798–803
23. Biswas M K, Xu Q, Deng X. Utility of RAPD, ISSR, IRAP and REMAP markers for the genetic analysis of *Citrus* spp. *Scientia Horticulturae*, 2010, **124**(2): 254–261
24. Talon M, Gmitter Jr. F G.Citrus genomics. *International Journal of Plant Genomics*, 2008, **2008**: 528361
25. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 1999, **27**(2): 573–580
26. Volfovsky N, Haas B J, Salzberg S L. A clustering method for repeat analysis in DNA sequences. *Genome Biology*, 2001, **2**(8): RESEARCH0027
27. Macas J, Mészáros T, Nouzová M. PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*, 2002, **18**(1): 28–35
28. Wicker T, Matthews D E, Keller B. TREP: a database for Triticeae repetitive elements. *Trends in Plant Science*, 2002, **7**(12): 561–562
29. Messing J, Bharti A K, Karlowski W M, Gundlach H, Kim H R, Yu Y, Wei F, Fuks G, Soderlund C A, Mayer K F, Wing R A. Sequence composition and genome organization of maize. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**(40):14349–14354
30. Jurka J, Kapitonov V V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 2005, **110**(1–4): 462–467
31. Meyers B C, Tingey S V, Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Research*, 2001, **11**(10): 1660–1676
32. Du J, Tian Z, Hans C S, Laten H M, Cannon S B, Jackson S A, Shoemaker R C, Ma J. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. *Plant Journal*, 2010, **63**(4): 584–598
33. Chen C, Zhou P, Choi Y A, Huang S, Gmitter F G Jr. Mining and characterizing microsatellites from citrus ESTs. *Theoretical and Applied Genetics*, 2006, **112**(7): 1248–1257
34. Cheng Y, de Vicente M C, Meng H, Guo W, Tao N, Deng X. A set of primers for analyzing chloroplast DNA diversity in *Citrus* and related genera. *Tree Physiology*, 2005, **25**(6): 661–672
35. Mayer C. Phobos: a tandem repeat search tool. Distributed by the author, http://www.rub.de/spezzoo/cm, 2007
36. Jurka J, Pethiyagoda C. Simple repetitive DNA sequences from primates: compilation and analysis. *Journal of Molecular Evolution*, 1995, **40**(2): 120–126
37. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 2011, **28**(10): 2731–2739
38. Kim T S, Booth J G, Gauch H G Jr, Sun Q, Park J, Lee Y H, Lee K. Simple sequence repeats in Neurospora crassa: distribution, polymorphism and evolutionary inference. *BMC Genomics*, 2008, **9**(1): 31
39. Tautz D, Renz M. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research*, 1984, **12**(10): 4127–4138
40. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, 2000, **10**(7): 967–981
41. Victoria F C, da Maia L C, de Oliveira A C. In silico comparative analysis of SSR markers in plants. *BMC Plant Biology*, 2011, **11**(1): 15
42. La Rota M, Kantety R V, Yu J K, Sorrells M E. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics*, 2005, **6**(1): 23
43. Lawson M J, Zhang L. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biology*, 2006, **7**(2): R14
44. Crane C F. Patterned sequence in the transcriptome of vascular plants. *BMC Genomics*, 2007, **8**(1): 173
45. Feschotte C, Jiang N, Wessler S R. Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics*, 2002, **3**(5): 329–341
46. Tanurdzic M, Vaughn M W, Jiang H, Lee T J, Slotkin R K, Sosinski B, Thompson W F, Doerge R W, Martienssen R A. Epigenomic consequences of immortalized plant cell suspension culture. *PLoS Biology*, 2008, **6**(12): 2880–2895
47. Picault N, Chaparro C, Piegu B, Stenger W, Formey D, Llauro C, Descombin J, Sabot F, Lasserre E, Meynard D, Guiderdoni E, Panaud O. Identification of an active LTR retrotransposon in rice. *Plant Journal*, 2009, **58**(5): 754–765
48. Pouteau S, Huttner E, Grandbastien M A, Caboche M. Specific expression of the tobacco Tnt1 retrotransposon in protoplasts. *EMBO Journal*, 1991, **10**(7): 1911–1918
49. Hirochika H. Activation of tobacco retrotransposons during tissue culture. *EMBO Journal*, 1993, **12**(6): 2521–2528
50. Mhiri C, Morel J B, Vernhettes S, Casacuberta J M, Lucas H, Grandbastien M A. The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Molecular Biology*, 1997, **33**(2): 257–266
51. Ramallo E, Kalendar R, Schulman A H, Martínez-Izquierdo J A. Reme1, a Copia retrotransposon in melon, is transcriptionally induced by UV light. *Plant Molecular Biology*, 2008, **66**(1–2): 137–150
52. Asíns M J, Monforte A J, Mestre P F, Carbonell E A. Citrus and Prunuscopia-like retrotransposons. *TAG Theoretical and Applied Genetics*, 1999, **99**(3–4): 503–510
53. Bernet G P, Asíns M J. Identification and genomic distribution of gypsy like retrotransposons in *Citrus* and *Poncirus*. *Theoretical and Applied Genetics*, 2003, **108**(1): 121–130
54. Xu Q, Chen L L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao W B, Hao B H, Lyon M P, Chen J, Gao S, Xing F, Lan H, Chang J W, Ge X, Lei Y, Hu Q, Miao Y, Wang L, Xiao S, Biswas M K, Zeng

W, Guo F, Cao H, Yang X, Xu X W, Cheng Y J, Xu J, Liu J H, Luo O J, Tang Z, Guo W W, Kuang H, Zhang H Y, Roose M L, Nagarajan N, Deng X X, Ruan Y. The draft genome of sweet orange (*Citrus sinensis*). *Nature Genetics*, 2013, **45**(1): 59–66

55. Rabinowicz P D, Schutz K, Dedhia N, Yordan C, Parnell L D, Stein L, McCombie W R, Martienssen R A. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nature Genetics*, 1999, **23**(3): 305–308