INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
21–0454/2021/26–5–579–586
DOI: 10.17957/IJAB/15.1870
http://www.fspublishers.org

Full Length Article

Identification of Retrotransposons using Transcriptome Data to Analyze Sugarcane Genomes and Chromosome Differentiation

Xiaoyang Wu[1], Dan Chen[1], Xin Lu[2], Wenjie Long[1], Xiongmei Ying[2], Guoyan Zhou[1], Juan Du[1], Shaoyun Wu[1] and Qing Cai[1*]

[1]*Biotechnology and Germplasm Resources Institute, Yunnan Academy of Agricultural Sciences, Kunming 650205, People's Republic of China*

[2]*Sugarcane Research Institute, Yunnan Academy of Agricultural Sciences, Kaiyuan 661699, People's Republic of China*

[*]Correspondence author: caiqingysri@163.com

Received 09 April 2021; Accepted 17 August 2021; Published 15 November 2021

Abstract

Sugarcane belongs to allopolyploid and is an important species for polyploid research. Retrotransposons are an important part of plant genome, however the composition of retrotransposons in sugarcane genome and their effects are not very clear. In this study, six selfing progenies of Yunzhe 07–86 were used as research materials, and the buds of their nodes were used for transcriptome sequencing. More than 80 GB of sequence data were obtained. A total of 97106 unigenes were obtained by sequence assembly without a reference genome. Unigenes were annotated based on the related proteins of retrotransposons, and much sequence information was obtained. One of the unigenes containing nested transposon loci was attracted our attention. Sequence structure analysis showed that this locus was a Line transposon inserted into an LTR transposon. The LTR transposon belongs to the *Ty3-Gypsy* type, and the Line transposon belongs to the *Sof-RTE* type. *Sof-RTE* transposons are a type of transposon with a high copy number in the sugarcane genome. Through analysis of the distribution of *Sof-RTE* transposons in the genomes of *S. spontaneum* and *S. officinarum*, it was found that the distribution of these transposons was site-specific in the genome and chromosome, and there were structural differences among chromosomes and genomes. Therefore, we believe that the activities of transposons such as *Sof-RTE* resulted in the differentiation of sugarcane genomes and chromosomes after polyploidization to a certain extent.

Keywords: *S. spontaneum*; *S. officinarum*; Genome; Retrotransposon; Differentiation

Introduction

Modern cultivated sugarcane (*Saccharum* hybrids spp.) is a perennial crop with aneuploidy and polyploidy. It is usually asexually propagated in production and the most important source of sugar in the world, accounting for 80% of the world's total sugar production (Santchurna *et al*. 2014). The botanical classification of sugarcane is *Poaceae*, *Panicoideae*, *Andropogoneae*, *Saccharinae* and *Saccharum* L. (Fang *et al*. 2014). classified sugarcane into wild species and cultivated species. The wild species were further divided into *S. spontaneum* and *S. robustum*. The cultivated species were divided into *S. sinense*, *S. edule*, *S. bareri* and *S. officinarum* (Yang *et al*. 2014). Some studies suggest that *S. officinarum* evolved or was domesticated from *S. robustum* (Brandes 1956; D'Hont *et al*. 1993, 1998; Racedo *et al*. 2016).

S. spontaneum, *S. robustum* and *S. officinarum* are homologous polyploids, while *S. sinense*, *S. edule*, *S. barberi* and modern sugarcane are interspecific allopolyploid hybrids. Among modern sugarcane, *S. officinarum* and *S. spontaneum* contribute most of the genome components, and ~70 to 80% of the genes come from *S. officinarum*, ~10 to 20% from *S. spontaneum* and 10% from recombinant chromosomes (Hoarau *et al*. 2002; Ming *et al*. 2006; Nishiyama *et al*. 2014). The ploidy level of the sugarcane genome varies from 5 to 16× and the genome size is ~10 Gb. The sugarcane genome is complex and contains 8–12 homologous aneuploid chromosomes. By determining the content of nuclear DNA in a large number of sugarcane samples from different varieties and ploidies, it was found that the genome size of *S. officinarum* was estimated to be ~7.50–8.55 Gb, while that for *S. robustum* was ~7.56–11.78 Gb, and that for *S. spontaneum* was ~3.36–12.64 Gb (Zhang *et al*. 2012). The genome size of sugarcane from different sources depends on the composition of chromosomes. Sugarcane genome composition makes it not only an important cash crop, but also an important plant for genetic and polyploid genome phylogenetic research in plants.

Polyploidy exists widely in nature and is an important driving force for the evolution of eukaryotes (Otto 2007);

To cite this paper: Wu X, D Chen, X Lu, W Long, X Ying, G Zhou, J Du, S Wu, Q Cai (2021). Identification of retrotransposons using transcriptome data to analyze sugarcane genomes and chromosome differentiation. *Intl J Agric Biol* 26:579–586

about 30 ~ 70% of angiosperms are considered to have polyploid ancestors (Masterson 1994). After polyploid formation, it will be further diploidized, its genome and chromosome will differentiate, and its diploid genetic behavior will be restored to maintain genetic stability (Doyle *et al*. 2008; Leitch and Leitch 2008). Transposon activity is considered to be one of the main reasons for diploidization (Feldman and Levy 2012).

Transposons are important genetic elements that can jump between or within chromosomes in the genome (i.e., jumping genes). Genomes of gramineous crops such as rice (Yu *et al*. 2002), maize (Schnable *et al*. 2009) and wheat (IWGSC 2014) have been sequenced. Transposons are one of the main components of these genomes. The classification of transposons can be divided into DNA transposons and RNA transposons (Wicker *et al*. 2007; Bourque *et al*. 2018). RNA transposons are further divided into long terminal repeat (LTR) retrotransposons and non-LTR retrotransposons. LTR retrotransposons can be further classified into *Ty1-Copia* and *Ty3-Gypsy* types, while non-LTR retrotransposons can be further classified into Line and Sine types (Kumar and Bennetzen 1999). Line retrotransposons usually have two large reading frames with a polyA tail at its 3'-end. Transposons with target site repeat (TSD) structure are generally considered as complete transposons, while transposons with autonomous transposon capability are considered as full-length transposons.

At present, the genome of AP85–441 an *S. spontaneum* line has been sequenced (Zhang *et al*. 2018), which provides basic data for the study of sugarcane genome composition. Ap85–441 is an autotetraploid with a chromosome base of eight. It is composed of four sets of genomes and contains 32 pairs of chromosomes.

Yunzhe07–86 is a new sugarcane germplasm that is generated from crosses between wild species and varieties (Chen *et al*. 2020). Its complex pedigree can be tracked back to *S. spontaneum*, *S. officinarum* and *S. robustum*. Transcriptome sequencing is a commonly used and simplified genome sequencing technology that can be applied to complex genome research. In this study, selfing progenies of Yunzhe07–86 were detected by transcriptome sequencing, and retrotransposons were identified using these transcriptome data. This annotation information was used to further study the genome and chromosome differentiation of sugarcane.

Materials and Methods

Plant material

Plant materials (sugar cane genotypes Yunzhe07–86) were provided by the National Germplasm Repository of Sugarcane, P.R. China (Kaiyuan city, Yunnan province). Yunzhe07-86 is a new sugarcane genotype that generated from the crosses of wild species and varieties. Its complex pedigree can be tracked back to *S. spontaneum*, *S. officinarum* and *S. robustum*. Six selfing progenies of Yunzhe07–86, sister lines12–137, 12–139, 12–149, 12–150, 12–171 and 12–174 were separately used to transcriptome sequencing，and the axillary buds formed in adult stage were selected. The axillary bud tissues were sampled from the cane nodes as experimental samples.

Transcriptome sequencing and sequence assembly

Transcriptome sequencing was performed on samples of axillary bud tissues at the adult plant stage; these samples were used for RNA extraction, reverse transcription and the construction of a DNA library. After the library was qualified, the bidirectional sequencing of the cDNA library was performed using the Illumina HiSeq high-throughput sequencing platform. The reading length was 150-bp. Raw data were filtered to obtain high-quality clean data by removing the connection sequence and low-quality reads. Clean data were used for sequence assembly without a reference genome, and Trinity software (Grabherr *et al*. 2011) was used to perform this work.

Annotation of retrotransposons in unigene

The unigene sequences obtained by sequence assembly were annotated for retrotransposons. The related protein sequences of conserved LTR transposon domains were used for annotation of LTR type retrotransposons (Neumann *et al*. 2019). Line retrotransposons were annotated using 236 related protein sequences of the conserved Line transposon domains in maize genome (Supplementary 1). BlastX, tBlastn and CDD (Marchler-Bauer *et al*. 2017) in NCBI were used to analyse conserved domains of transposons. ORFfinder in NCBI was used to analyse the open reading frame (ORF). Clustal X (1.8) software was used for sequence alignment (Thompson *et al*. 1994). MEGA 5.0 (Kumar *et al*. 1994) was used for phylogenetic tree construction.

Cloning of Line transposons

Genomic DNA was extracted from young leaves and used for transposon cloning (Doyle and Doyle 1990). Based on the identification of the LTR transposon sequence in the AP85–441 genome, molecular markers were developed to clone the full-length sequence of the Line transposon at the nested transposon site. Molecular markers were designed online using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) (Rozen and Skaletsky 2000). Cloning was performed with two fragments and two pairs of primers, F1 5'-TCCAGAGTTTCCAGGGAGTG-3', and R1 5'-TCTTCCTCTCGCGGATTCTA -3'; F2 5'-AAAGGTCAGGCGTATGATGG -3', and R2 5'-TGGTTCAAGATGCAGACCAG -3'.

The PCR system was performed in a final volume of 50 μL containing 50 ng template DNA, 1 U Taq

polymerase, 1× PCR buffer, 0.2 mM dNTPs, and 0.2 μM of each primer. PCR was performed under the following conditions: 5 min at 94°C, 35 cycles of 60 s at 94°C, 60 s at annealing temperature (Tm), and 60 s at 72°C, and a final extension of 5 min at 72°C. The PCR products were then subjected to 1% agarose electrophoresis, cloning, and sequencing. The TA cloning vector was PMD19-T (TaKaRa, Japan). The cloned vector was sequenced using a 3730 DNA Analyzer (Applied Biosystems, America).

Identification of retrotransposons in the sugarcane genome

The genome sequence of *S. spontaneum* line AP85-441 (Zhang *et al*. 2018) with perfect assembly and its gff3 annotation file were downloaded from http://www.life.illinois.edu/ming/downloads/Spontaneum_genome/ and used to identify retrotransposons. Ap85-441 genome was assembled into 32 chromosomes. The transposon sequence was obtained and its chromosome location information was recorded at the same time.

For the identification of LTR retrotransposons, the annotated LTR retrotransposon-related unigene sequence was located in the reference genome by BLASTn; 10-kb sequences from upstream and downstream of this locus were extracted; the LTRs structure of LTR transposon was identified from the extracted sequence; and the TSD structure of the retrotransposon was further identified after its LTR structure was determined.

The original genome sequence of *S. officinarum* line Gp0114240 (sequence information is shown in Supplementary Table 1) was also used to identify retrotransposons. BLASTn (version 2.2.23+; Altschul 1997) was used for sequence blasting against the genome sequence. Blat was used for reads mapping to reference sequence. Perl scripts and regular expressions were used to identify transposons. Line retrotransposons were identified by describing their structure suing regular expressions, with a fragment size from 150 to 4000-bp; a 3'-terminal sequence end of (TTG) n and a 10–20 bp TSD structure. Perl scripts were used to extract sequences with the structural features from the genome sequence.

Results

Annotation of retrotransposons in unigenes

Through transcriptome sequencing, 84.78-Gb of clean data was obtained after quality control (Table 1). A total of 97106 unigenes were obtained after sequence assembly without a reference genome. By annotating retrotransposons, a series of unigenes related to retrotransposons were found. One of the unigenes (unigene number is c108043) with a special structure attracted our attention. This sequence has both LTR and Line retrotransposon domains in sequence. This unigene originated from a nested transposon site in the sugarcane genome. According to the order of conserved domains on the sequence, we know that this site is a Line retrotransposon inserted into the interior of an LTR retrotransposon (Fig. 1). This locus was located in the *S. officinarum* genome by BLASTn.

Cloning of Line retrotransposons

To understand the sequence characteristics of this nested transposon locus in the sugarcane genome, we first annotated the full-length sequence of this LTR retrotransposon in the AP85-441 genome using the partial sequence of the LTR retrotransposon in c108043. As a result, we obtained three full-length LTR retrotransposon sequences from the AP85-441 genome (Supplementary 2), suggesting that this is an LTR retrotransposon family with low copy number in sugarcane genome. This type of LTR retrotransposon is about ~14 Kb in length; its LTR is ~1.6 Kb in length; and it belongs to *Ty3-Gypsy* type; the ends of the LTR structure have a 4-bp terminal inverted repeats (TIR) structure; which usually has a 5-bp TSD structure, but the TSD sequence is not conserved.

The identification of LTR retrotransposons provides a sequence reference for target sites of Line retrotransposons. We designed primers and cloned the full-length Line retrotransposon at this site. This Line retrotransposon is 3384-bp in length; its TSD sequence is 5'-TGATGTCCCTTATCT-3'; its 3'-terminal ends is 11 5'-TTG-3'; and its sequence contains only one open reading frame; there are two conserved domains in its open reading frame, exonuclease / endonuclease / phosphatase (EEP, cd09076) and reverse transcriptase (RT, cd01650).

This Line retrotransposon belongs to the *Sof-RTE* family. Gao *et al.* (2017) first discovered this type of transposon in the *Arachis duranensis* genome and annotated a family member *Sof-RTE* (KF184817, 90241–93611) in the BAC sequence of the sugarcane genome. Comparing with the works of Gao *et al*. (2017), we found that the transposon we cloned was a full-length *Sof-RTE* transposon. This is the first *Sof-RTE* obtained in this study, which we called *Sof-RTE-ck*.

Annotation of *Sof-RTE* in the sugarcane Genome

Annotation of *Sof-RTE* in the *S. officinarum* genome was performed using the special structure $(TTG)_n$ at the 3'-terminus of *Sof-RTE* to isolate sequences containing the target site. After eliminating duplicates, 64,080 target site of *Sof-RTE* in the *S. officinarum* genome were obtained (Supplementary 4). Thus, *Sof-RTE* is a high-copy transposon in the sugarcane genome.

By annotating *Sof-RTE* in the *S. spontaneum* genome, based on the structural characteristics of the *Sof-RTE* transposon, 2369 sequences were identified (Fig. 2 and Supplementary. 5). By comparison with Line-related

Table 1: The transcriptome sequencing data obtained in this study

Samples	Read number	Base number	GC content (%)	% ⩾ Q30 (%)
12-137	37,021,591	10,898,254,820	54.25	93.36
12-139	47,205,821	13,917,742,462	54.93	93.06
12-149	44,804,301	13,180,928,498	55.33	92.76
12-150	45,589,814	13,409,064,880	57.01	92.86
12-171	45,425,006	13,465,860,422	55.08	93.03
12-174	67,460,254	19,906,175,042	55.81	92.98

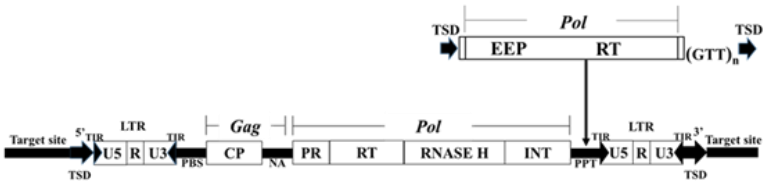

Fig 1: Nested transposon loci in the *S. officinarum* genome
At this locus, a Line retrotransposon was inserted into an LTR retrotransposons. Both transposons have a complete TSD structures. The LTR retrotransposons has two reading frames, *Gag* and *Pol*, *Gag* has a conserved domain CP (capsid-like proteins), and *Pol* has four conserved domains PR (protease), RT (reverse transcriptase), RNASEH (RNase-H) and INT (integrase). According to the order of conserved domains, the LTR retrotransposons belongs to the *Ty3-Gypsy* type. The Line transposon has only one reading frame *Pol*, *Pol* has two conserved domains EEP and RT, and its 3'-terminal ends has a poly 5'-TTG-3' tail

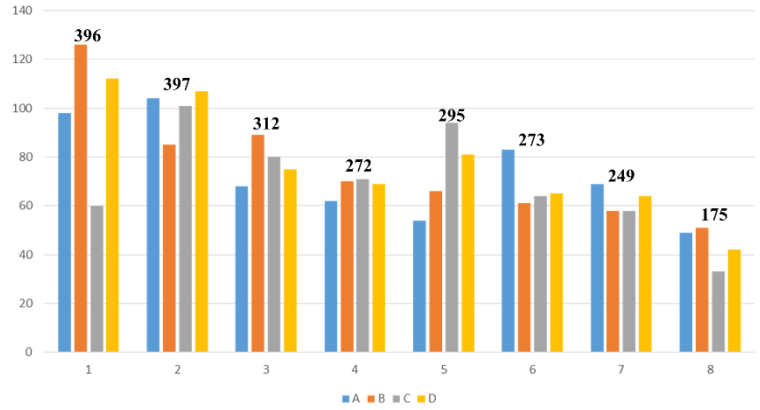

Fig 2: Frequency distribution of *Sof-RTE* on different chromosomes of the *S. spontaneum* genome
In the figure, 1-8 represent the 8 homologous groups of *S. spontaneum*; A to D represent the A to D chromosomes of *S. spontaneum*

proteins, 1655 of these have Line protein coding sequences. Some of them are 5'-terminal truncated *Sof-RTE* transposons. For example, *S. spontaneum-Sof-RTE-1731*, located on chromosome 5B, has a length of 1592 bp, a TSD sequence of 5'-TCAATAGAAGGAAGA-3', a 3'-terminus of six 5'-TTG-3', and a reading frame that covers only part of *Sof-RTE-ck* (Fig. 3. and Supplementary Fig. 1).

All 1623 sequences had more than 95% similarity with *Sof-RTE-ck*. In addition, there was a certain proportion of sequences that had less than 80% similarity with *Sof-RTE-ck*. However, they still have the structural characteristics of *Sof-RTE*. By generating a phylogenetic tree, the *Sof-RTE* transposon can be further divided into two subfamilies (Fig. 4).

Distribution of *Sof-RTE* in the sugarcane genome

Analysis of homologous loci in the *S. officinarum* and *S. spontaneum* genomes, revealed 64080 target site sequences of *Sof-RTE* in the *S. officinarum* genome that were used to blasted against the *S. spontaneum* genome sequence. Only 55 common loci were obtained (Table 2), therefore, the distribution of *Sof-RTE* transposon has a strong genome specificity.

The chromosome location of each *Sof-RTE* transposon is shown in Supplementary 6. According to the distribution of *Sof-RTE* acrcoss different chromosomes of the *S. spontaneum* genome, there was no linear relationship between homologous chromosomes, that is to say, there was no common target site of *Sof-RTE* on homologous chromosomes (Fig. 5). It was concluded that the distribution of *Sof-RTE* in the sugarcane genome was the intergenic region, to a certain extent, which resulted in the differentiation of homologous chromosomes in the sugarcane genome.

Discussion

With the development of high-throughput sequencing

Table 2: Homologous loci between the *S. officinarum* and *S. spontaneum* genomes

Number	*S. officinarum*	Chr *in S. spontaneum*	Chromosome segment
1	*S. officinarum-Sof-RTE-19860*	Chr1A	106111357-106111258
2	*S. officinarum-Sof-RTE-27843*	Chr1B	6581776-6581875
3	*S. officinarum-Sof-RTE-25564*	Chr1B	24632359-24632260
4	*S. officinarum-Sof-RTE-40464*	Chr1C	22428752-22428653
5	*S. officinarum-Sof-RTE-1843*	Chr1C	74175599-74175500
6	*S. officinarum-Sof-RTE-2335*	Chr1D	1131556-1131655
7	*S. officinarum-Sof-RTE-41574*	Chr1D	89731189-89731288
8	*S. officinarum-Sof-RTE-41574*	Chr1D	89826842-89826941
9	*S. officinarum-Sof-RTE-41574*	Chr1D	91626648-91626549
10	*S. officinarum-Sof-RTE-22346*	Chr1D	104944989-104944890
11	*S. officinarum-Sof-RTE-45139*	Chr2A	10739590-10739491
12	*S. officinarum-Sof-RTE-56705*	Chr2A	16179220-16179121
13	*S. officinarum-Sof-RTE-28984*	Chr2A	121313131-121313230
14	*S. officinarum-Sof-RTE-22078*	Chr2B	5414006-5413907
15	*S. officinarum-Sof-RTE-21186*	Chr2B	9585567-9585666
16	*S. officinarum-Sof-RTE-9661*	Chr2B	10814620-10814719
17	*S. officinarum-Sof-RTE-61228*	Chr2B	13334082-13334181
18	*S. officinarum-Sof-RTE-61228*	Chr2B	13520903-13520804
19	*S. officinarum-Sof-RTE-9661*	Chr2B	51892192-51892093
20	*S. officinarum-Sof-RTE-50243*	Chr2B	88691390-88691489
21	*S. officinarum-Sof-RTE-13155*	Chr2C	8228582-8228483
22	*S. officinarum-Sof-RTE-13155*	Chr2C	8372924-8372825
23	*S. officinarum-Sof-RTE-23057*	Chr2C	10441229-10441130
24	*S. officinarum-Sof-RTE-28726*	Chr2C	12742638-12742539
25	*S. officinarum-Sof-RTE-57667*	Chr2C	16654576-16654675
26	*S. officinarum-Sof-RTE-24068*	Chr2C	107025007-107024908
27	*S. officinarum-Sof-RTE-15485*	Chr2D	4580685-4580586
28	*S. officinarum-Sof-RTE-15485*	Chr2D	4610527-4610626
29	*S. officinarum-Sof-RTE-19601*	Chr2D	15849006-15848907
30	*S. officinarum-Sof-RTE-45430*	Chr2D	83149575-83149674
31	*S. officinarum-Sof-RTE-31537*	Chr3B	1924081-1924180
32	*S. officinarum-Sof-RTE-31537*	Chr3B	1934881-1934980
33	*S. officinarum-Sof-RTE-53554*	Chr3B	7307962-7307863
34	*S. officinarum-Sof-RTE-19629*	Chr3B	15638469-15638568
35	*S. officinarum-Sof-RTE-52066*	Chr3D	2428956-2429055
36	*S. officinarum-Sof-RTE-47415*	Chr3D	54403261-54403360
37	*S. officinarum-Sof-RTE-33464*	Chr4B	67307808-67307709
38	*S. officinarum-Sof-RTE-54217*	Chr4C	5900120-5900219
39	*S. officinarum-Sof-RTE-21052*	Chr4C	47177877-47177976
40	*S. officinarum-Sof-RTE-47876*	Chr5A	46390937-46391036
41	*S. officinarum-Sof-RTE-54664*	Chr5B	3387867-3387768
42	*S. officinarum-Sof-RTE-16803*	Chr5B	45153628-45153529
43	*S. officinarum-Sof-RTE-16803*	Chr5B	45289674-45289575
44	*S. officinarum-Sof-RTE-48465*	Chr5B	46874481-46874580
45	*S. officinarum-Sof-RTE-29906*	Chr5B	47006227-47006128
46	*S. officinarum-Sof-RTE-17535*	Chr5C	63744421-63744322
47	*S. officinarum-Sof-RTE-17535*	Chr5C	63775422-63775323
48	*S. officinarum-Sof-RTE-20752*	Chr5D	2730626-2730527
49	*S. officinarum-Sof-RTE-20752*	Chr5D	2836968-2836869
50	*S. officinarum-Sof-RTE-17942*	Chr6B	59754200-59754101
51	*S. officinarum-Sof-RTE-44844*	Chr7B	3704118-3704217
52	*S. officinarum-Sof-RTE-51883*	Chr7D	7777073-7776974
53	*S. officinarum-Sof-RTE-33508*	Chr7D	14334279-14334180
54	*S. officinarum-Sof-RTE-50246*	Chr8D	4054789-4054888
55	*S. officinarum-Sof-RTE-25411*	Chr8D	7002287-7002386

technology, the cost of sequencing is decreasing, making this technology became a common method in genome research. In the study of complex genomes, some steps are usually performed before sequencing to remove unwanted DNA information. ChIP-seq is a technology combining chromatin immunoprecipitation (ChIP) with high throughput sequencing. After enrichment for DNA sequences bound to a specific protein, the library is constructed and sequenced. Zhang *et al*. (2017) obtained a large number of retrotransposon sequences using this sequencing technique, and successfully developed sugarcane centromere site-specific fluorescence in situ hybridization (FISH) probes using these sequences. To facilitate transposon sequence assembly without a reference genome, we used transcriptome sequencing to study the sugarcane genome and obtained 97106 unigenes of sugarcane. A large amount of retrotransposon information remains in these unigenes.

Sof-RTE is a member of the *RTE* clade in the Line family, and *RTE* is a type of widespread horizontal transposon that has been transferred across plant and animal genomes (Gao *et al*. 2017). Typical Line transposons usually have two open reading frames and a polyA tail at the 3'-end (Kumar and Bennetzen 1999). However, the end of the *Sof-RTE* is a poly 5'-TTG-3' sequence, with only one open reading frame in its sequence, and the known full-length sequence of *Sof-RTE* is approximately 3-kb. *Cin4* is a type of Line transposon found in the maize genome; its full-length sequence is above 7 kb; and there are 5'-terminal truncated family members (Schwarz-Sommer *et al*. 1987). The 5'-terminal truncated *Cin4* was also found to be inserted in the *waxy* gene (Wu *et al*. 2017). The transposon *Sof-RTE* shows the same pattern. We believe that *Sof-RTE* has completed its jump in the genome, resulting in its 5'-terminal truncation and the 5'-terminal truncated *Sof-RTE* still contains the complete transposon structure.

Polyploidization plays an important role in the genome evolution of eukaryotic organisms, especially plants. The sugarcane genome is very similar to that of sorghum (*Sorghum bicolor* (L.) Moench). Sorghum is a diploid plant. Before Eight million years ago, sugarcane and sorghum shared a common ancestor (Wang *et al*. 2010). Sugarcane ancestors subsequently formed octoploids by double polyploidization of the *S. robustum* and *S. spontaneum* genomes (D'Hont *et al*. 1998; Ha *et al*. 1999; Aitken *et al*. 2014). It is generally accepted that most plants have one or two rounds of genome-wide duplication after diploidization, and the cycle of polyploidization and diploidization forms a stable genome at present (Comai 2005; Jiao *et al*. 2011, 2014). In this study, we found that there were few common transposon target sites between the *S. spontaneum* and *S. officinarum* genomes, which indicates that a large amount of transposon activity occurred in the independent evolution stage of the two genomes. Perhaps it is the activity of transposons such as *Sof-RTE* that results in the differentiation of genomes.

Chromosome pairing during meiosis can be observed and shows the homology and evolution of chromosomes. Although the sugarcane genome is polyploid, meiosis mainly occurs through bivalent pairing, while trivalent and monovalent pairing is rare (Bielig *et al*. 2003). Diploidization after polyploidization has resulted in the existing chromosome behavior in sugarcane. Transposon activity provides an effective pathway for diploidization through chromosome rearrangement (Feldman and Levy

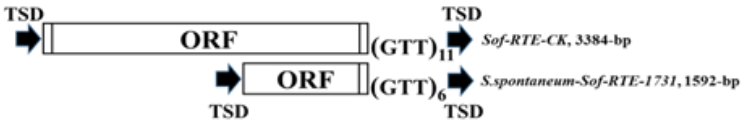

Fig 3: Structural comparison of full-length and 5'-terminal truncated *Sof-RTE*
Taking *Sof-RTE-ck* and *S. spontaneum-Sof-RTE-1731* as examples, the full-length and 5'-terminal truncated *Sof-RTE* transposons were compared. Judging from the transposons with complete TSD structures, both *Sof-RTE-ck* and *S. spontaneum-Sof-RTE-1731* are complete transposons. While *S. spontaneum-Sof-RTE-1731* is a 5'-terminal truncated transposon compared to full-length transposons such as *Sof-RTE-ck*

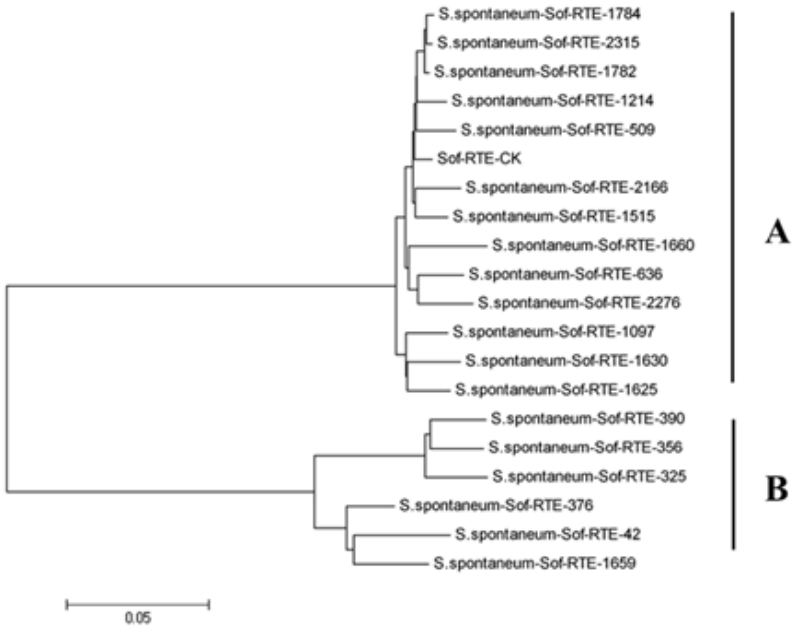

Fig 4: Phylogenetic tree of *Sof-RTE* in the *S. spontaneum* genome
Twenty *Sof-RTE* sequences were used to construct the phylogenetic tree. There are transposons with the same conserved domains but low sequence similarity with *Sof-RTE-CK* in the *S. spontaneum* genome. By establishing phylogenetic trees, these sequences can be divided into two subclasses, A and B

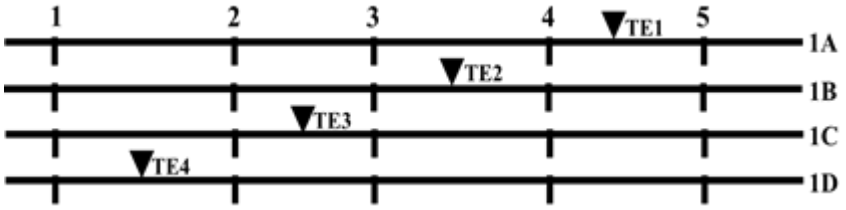

Fig 5: Transposon activity causes homologous chromosome differentiation
In the figure, 1A, 1B, 1C and 1D represent four homologous chromosomes of the sugarcane genome; 1-5 indicates the linear gene loci among the homologous chromosomes; and TE1-TE4 indicates transposon insertion sites. There is no linear relationship among the insertion sites of transposons across homologous chromosomes, so it can be assumed that the activity of transposons causes the differentiation of homologous chromosomes

2012). Transposon activity can lead to the expansion and reduction of genome capacity (Schubert and Vu 2016). Transposon activity is usually not so accurate, which causes the change of its target site sequences (Jiang *et al.* 2004). Even transposons that lose transposable ability still have the potentiality to change chromosome structure (Carvalho and Lupski 2016). Recombination events can occur between homologous regions, even if these are scattered far away in same or different chromosomes, resulting in large-scale deletion, duplication or inversion (Bennetzen and Wang 2014). Transposons also provide micro homologous regions, which predispose to template switching during repair of replication, resulting in chromosome structural variants (Lee *et al.* 2007). In this study, we found that distribution of the *Sof-RTE* transposon has a strong genome specificity and that there is no linear relationship between homologous chromosomes. We believe that transposon like *Sof-RTE* activity is responsible for the differentiation of homologous chromosomes and further results in the diploidization of chromosomal behavior.

Conclusion

In this study, we found the Line transposon *Sof-RTE* through the annotation of sugarcane unigenes. *Sof-RTE* transposons are high copy number transposons in the sugarcane genome. These transposons were site-specific in the genome and chromosome. The activity of *Sof-RTE* transposons in the genome results in the structural differences among homologous chromosomes and genomes of sugarcane. We believe that the transposons activity similar to that of *Sof-RTE* is one of the main driving forces of chromosome and genome differentiation.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (31560416), the Post-doctoral Targeted Funding of Yunnan Province (Yunren Shetong [2018] 168) and the Yunnan Provincial Science and Technology Department Plan (2016FB067).

Conflicts of Interest

The authors declare no competing interests.

Data Availability

Data presented in this study will be available upon reasonable request by the corresponding author.

Ethics Approval

Not applicable to this paper.

References

Aitken KS, MD McNeil, PJ Berkman, S Hermann, A Kilian, PC Bundock, JC Li (2014). Comparative mapping in the Poaceae family reveals translocations in the complex polyploid genome of sugarcane. *BMC Plant Biol* 14; Article 190

Altschul SF, LM Thomas, AS Alejandro, JH Zhang, Z Zhang, M Webb, JL David (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402

Bennetzen JL, H Wang (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol* 65:505–530

Bielig ML, A Mariani, N Berding (2003). Cytological studies of 2n male gamete formation in sugarcane, *Saccharum* L. *Euphytica* 133:137–151

Bourque G, KH Burns, M Gehring, V Gorbunova, A Seluanov, M Hammell, M Imbeault, Z Izsvák, HL Levin, TS Macfarlan, DL Mager, C Feschotte (2018). Ten things you should know about transposable elements. *Genome Biol* 19; Article 199

Brandes EW (1956). Origin, dispersal and use in breeding of the Melanesian Garden sugarcanes and their derivatives, *Saccharum officinarum* L. *Proc Intl Soc Suagr Cane Technol* 9:709–750

Carvalho CM, JR Lupski (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 17:224–238

Chen D, X Lu, XY Wu, XM Ying, WJ Long, HS Su, HB Liu, XQ Lin, CH Xu, Q Cai (2020). Transcriptome analysis of axillary bud differentiation in a new dual-axillary bud genotype of sugarcane. *Genet Resour Crop Evol* 67:685–701

Comai L (2005). The advantages and disadvantages of being polyploidy. *Nat Rev Genet* 6:836–846

D'Hont A, D Ison, K Alix, C Roux, JC Glaszmann (1998). Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. *Genome* 41:221–225

D'Hont A, YH Lu, P Feldmann, JC Glaszmann (1993). Cytoplasmic diversity in sugarcane revealed by heterologous probes. *Sugar Cane* 1:12–15

Doyle JJ, LE Flage, AH Paterson, RA Rapp, DE Soltis, PS Soltis, JF Wendel (2008). Evolutionary genetics of genome merger and doubling in plants. *Annu Rev Genet* 42:443–461

Doyle JJ, JL Doyle (1990). Isolation of plant DNA from fresh tissue. *Focus* 12:13–15

Fang JP, YX Que, RK Chen (2014). A Review of *Saccharum* Origin and its Evolutionary Relationship with Related Genera. *Chin J Trop Crops* 35:816–822 (In Chinese)

Feldman M, AA Levy (2012). Genome evolution due to allopolyploidization in wheat. *Genetics* 192:763–774

Gao DY, Y Chu, H Xia, CM Xu, K Heyduk, B Abernathy, P Ozias-Akins, HJ Leebens-Mack, AS Jackson (2017). Horizontal transfer of non-LTR retrotransposons from arthropods to flowering plants. *Mol Biol Evol* 35:354–364

Grabherr MG, BJ Haas, M Yassour, JZ Levin, DA Thompson, I Amit, X Adiconis, L Fan, R Raychowdhury, QD Zeng, ZH Chen, E Mauceli, N Hacohen, A Gnirke, N Rhind, F Palma, BW Birren, C Nusbaum, K Lindblad-Toh, N Friedman, A Regev (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652

Ha S, HP Moore, D Heinz, S Kato, N Ohmido, K Fukui (1999). Quantitative chromosome map of the polyploid Saccharum spontaneum by multicolor fluorescence in situ hybridization and imaging methods. *Plant Mol Biol* 38:1165–1173

Hoarau JY, L Grivet, B Offmann, LM Raboin, JP Diorflar, J Payet, M Hellmann, A D'Hont, JC Glaszmann (2002). Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). II Detection of QTLs for yield components. *Theor Appl Genet* 105:1027–1037

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

Jiao YN, JP Li, HB Tang, HA Patersona (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* 26:2792–2802

Jiao YN, JN Wickett, S Ayyampalayam, SA Chanderbali, L Landherr, EP Ralph, PL Tomsho, Y Hu, HY Liang, SP Soltis, ED Soltis, WS Clifton, ES Schlarbaum, CS Schuster, H Ma, J Leebens-Mack, WC dePamphilis (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97–100

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

Kumar S, K Tamura, M Nei (1994). Mega: Molecular evolutionary genetic analysis software for microcomputers. *Comput Appl Biosci* 10:189–191

Lee JA, CM Carvalho, JR Lupski (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131:1235–1247

Leitch A, I Leitch (2008). Genomic plasticity and the diversity of polyploid plants. *Science* 320:481–483

Masterson J (1994). Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science* 264:421–424

Marchler-Bauer A, Y Bo, LY Han, JN He, JC Lanczycki, SN Lu, F Chitsaz, KM Derbyshire, CR Geer, RN Gonzales, M Gwadz, ID Hurwitz, F Lu, Marchler G H., SJ Song, N Thanki, ZX Wang, AR Yamashita, DC Zhang, CJ Zheng, YL Geer, HS Bryant (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 45:D200–D203

Ming R, HP Moore, KK Wu, A D'hont, CJ Glaszmann, LT Tew, TE Mirkov, J Silva, J Jifon, M Rai, JR Schnell, MS Brumbley, P Lakshmanan, CJ Comstock, HA Paterson (2006). Sugarcane improvement through breeding and biotechnology. *Plant Breed Rev* 27:15–118

Neumann P, P Novák, N Hoštáková, J Macas (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* 10; Article 1

Nishiyama MYJ, SS Ferreira, PZ Tang, S Becker, A Pörtner-Taliana, GM Souza (2014). Full-length enriched cDNA libraries and ORFeome analysis of sugarcane hybrid and ancestor genotypes. *PloS One* 9; Article e107351

Otto SP (2007). The evolutionary consequences of polyploidy. *Cell* 131:452–462

Racedo J, L Gutiérrez, MF Perera, Ostengo S Santiago, EM Pardo, MI Cuenya, B Welin, AP Castagnaro (2016) Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biol* 16; Article 142

Rozen S, H Skaletsky (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386

Santchurna D, K Ramdoyala, MGH Badalooa, MT Labuschagneb (2014). From sugar industry to cane industry: Evaluation and simultaneous selection of different types of high biomass canes. *Biomass Bioener* 61:82–92

Schnable SP, D Ware, SR Fulton, CJ Stein, FS Wei, S Pasternak, CZ Liang, JW Zhang, L Fulton, AT Graves, P Minx, AD Reily, L Courtney, SS Kruchowski, C Tomlinson, C Strong, K Delehaunty, C Fronick, B Courtney, MS Rock, E Belter, FY Du, K Kim, MR Abbott, M Cotton, A Levy, P Marchetto, K Ochoa, MS Jackson, B Gillam, WZ Chen, Yan L, J Higginbotham, M Cardenas, J Waligorski, E Applebaum, L Phelps, J Falcone, K Kanchi, T Thane, A Scimone, N Thane, J Henke, T Wang, J Ruppert, N Shah, K Rotter, J Hodges, E Ingenthron, M Cordes, S Kohlberg, J Sgro, B Delgado, K Mead, A Chinwalla, S Leonard, K Crouse, K Collura, D Kudrna, J Currie, RF He, A Angelova, S Rajasekar, T Mueller, R Lomeli, G Scara, A Ko, K Delaney, M Wissotski, G Lopez, D Campos, M Braidotti, E Ashley, W Golser, H Kim, S Lee, JK Lin, Z Dujmic, W Kim, J Talag, A Zuccolo, CZ Fan, A Sebastian, M Kramer, L Spiegel, L Nascimento, T Zutavern, B Miller, C Ambroise, S Muller, W Spooner, A Narechania, LY Ren, SR Wei, S Kumari, B Faga, JM Levy, L McMahan, PV Buren, WM Vaughn, K Ying, CT Ye, JS Emrich, Y Jia, A Kalyanaraman, AP Hsia, WB Brad, SR Baucom, PT Brutnell, CN Carpita, C Chaparro, JM Chia, JM Deragon, CJ Estill, Y Fu, AJ Jeddeloh, YJ Han, H Lee, PH Li, RD Lisch, SZ Liu, ZJ Liu, DH Nagel, CM McCann, P SanMiguel, MA Myers, D Nettleton, J Nguyen, WB Penning, L Ponnala, LK Schneider, CD Schwartz, A Sharma, C Soderlund, MN Springer, Q Sun, H Wang, M Waterman, R Westerman, KT Wolfgruber, LX Yang, YS Yu, LF Zhang, SG Zhou, QH Zhu, LJ Bennetzen, RD Kelly, JM Jiang, N Jiang, GG Presting, RS Wessler, S Aluru, AR Martienssen, WS Clifton, WM Richard, AR Wing, KR Wilson (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326:1112–1115

Schubert I, GTH Vu (2016). Genome stability and evolution: Attempting a holistic view. *Trends Plant Sci* 21:749–757

Schwarz-Sommer Z, L Leclercq, E Göbel, H Saedler (1987). Cin4, an insert altering the structure of the *Al* gene in *Zea mays*, exhibits properties of nonviral retrotransposons. *EMBO J* 6:3873–3880

The International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:1–14

Thompson DJ, GD Higgins, JT Gibson (1994). CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* 22:4673–4680

Wang JP, B Roe, S Macmil, QY Yu, JE Murray, HB Tang, CX Chen, F Najar, G Wiley, J Bowers, MV Sluys, DS Rokhsar, ME Hudson, SP Moose, AH Paterson, R Ming (2010). Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11; Article 261

Wicker T, R Sabot, A Hua-Van, LJ Bennetzen, P Capy, B Chalhoub, A Flavell, P Leroy, M Morgante, O Panaud, E Paux, P SanMiguel, HA Schulman (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8:973–982

Wu XY, D Chen, YQ Lu, WH Liu, XM Yang, XQ Li, J Du, LH Li (2017). Molecular characteristics of two new *waxy* mutations in China waxy maize. *Mol Breed* 37:1–7

Yang CF, LT Yang, YR Li (2014) Origins and evolution of sugarcane. *J South Agric* 45:1744–1750 (In Chinese)

Yu J, SN Hu, J Wong GKS Wang, SG Li, B Liu, YJ Deng, L Dai, Y Zhou, XQ Zhang, ML Cao, J Liu, JD Sun, JB Tang, YJ Chen, XB Huang, W Lin, C Ye, W Tong, LJ Cong, JN Geng, YJ Han, L Li, W Li, GQ Hu, XG Huang, WJ Li, J Li, ZW Liu, L Li, JP Liu, QH Qi, JS Liu, L Li, T Li, XG Wang, H Lu, TT Wu, M Zhu, PX Ni, H Han, W Dong, XY Ren, XL Feng, P Cui, XR Li, H Wang, X Xu, WX Zhai, Z Xu, JS Zhang, SJ He, JG Zhang, JC Xu, KL Zhang, XW Zheng, JH Dong, WY Zeng, L Tao, J Ye, J Tan, XD Ren, XW Chen, J He, DF Liu, W Tian, CG Tian, HG Xia, QY Bao, G Li, H Gao, T Cao, J Wang, WM Zhao, P Li, W Chen, XD Wang, Y Zhang, JF Hu, J Wang, S Liu, J Yang, GY Zhang, YQ Xiong, ZJ Li, L Mao, CS Zhou, Z Zhu, RS Chen, BL Hao, WM Zheng, SY Chen, W Guo, GJ Li, SQ Liu, M Tao, J Wang, LH Zhu, LP Yuan, HM Yang (2002). A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*). *Science* 296:79–92

Zhang JS, C Nagai, QY Yu, YB Pan, T Ayala-Silva, JR Schnell, CJ Comstock, KA Arumuganathan, R Ming (2012). Genome size variation in three *Saccharum* species. *Euphytica* 185:511–519

Zhang JS, XT Zhang, HT Tang, Q Hua XT Zhang, XK Ma, F Zhu, T Jones, XG Zhu, J Bowers, CM Wai, CF Zheng, Y Shi, S Chen, XM Xu, JJ Yue, RD Nelson, LX Huang, Li Z Zhen, HM Xu, D Zhou, YJ Wang, WC Hu, JS Lin, YJ Deng, N Pandey, M Mancini, D Zerpa, KJ Nguyen, LM Wang, L Yu, YG Xin, LF Ge, J Arro, OJ Han, S Chakrabarty, M Pushko, WP Zhang, YH Ma, PP Ma, MJ Lv, FM Chen, GY Zheng, JS Xu, ZH Yang, F Deng, XQ Chen, ZY Liao, XX Zhang, ZC Lin, H Lin, HS Yan, Z Kuang, WM Zhong, PP Liang, GF Wang, Y Yuan, JX Shi, JX Hou, JX Lin, JJ Jin, PJ Cao, QC Shen, Q Jiang, P Zhou, YY Ma, XD Zhang, RR Xu, J Liu, YM Zhou, HF Jia, Q Ma, R Qi, ZL Zhang, JP Fang, HK Fang, JJ Song, MJ Wang, GG Dong, G Wang, Z Chen, T Ma, H Liu, RS Dhungana, ES Huss, XP Yang, A Sharma, HJ Trujillo, CM Martinez, M Hudson, JJ Riascos, M Schuler, LQ Chen, MD Braun, L Li, QY Yu, JP Wang, K Wang, CM Schatz, D Heckerman, MV Sluys, GM Souza, HP Moore, D Sankoff, R VanBuren, HA Paterson, C Nagai, R Ming (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat Genet* 50:1565–1573

Zhang WP, S Zuo, ZJ Li, Z Meng, JL Han, JP Song, YB Pan, K Wang (2017). Isolation and characterization of centromeric repetitive DNA sequences in *Saccharum spontaneum*. *Sci Rep* 7; Article 41659