Transcriptome analysis and development of simple sequence repeat (SSR) markers in *Zingiber striolatum* Diels

Kuanping Deng¹ · Renju Deng² · Jianxin Fan³ · Enfa Chen²

Received: 24 August 2017 / Revised: 16 October 2017 / Accepted: 6 November 2017 / Published online: 8 December 2017
© Prof. H.S. Srivastava Foundation for Science and Society 2017

Abstract Illumina-based paired-end sequencing technology was used for the high-throughput transcriptome sequencing of combined *Zingiber striolatum* Diels tissues (i.e., root, stem, leaf, flower, and fruit tissues). More than 130 million sequencing reads were generated, and a *de novo* assembly yielded 287,959 contigs and 112,107 unigenes with an average length of 1029 and 28,891 bp, respectively. Similarity searches with known sequences led to the identification of 51,804 (46.21%) genes. Of the annotated unigenes, 6867 and 51,987 were assigned to Gene Ontology and Clusters of Orthologous Groups categories, respectively. Additionally, 8384 simple sequence repeats (SSRs) were identified as potential molecular markers in the unigenes. Thirty pairs of polymerase chain reaction primers were designed and used to validate the unigenes and assess the associated genomic polymorphism. The PCR amplification products for 25 primer pairs were of the expected size. These primers may represent usable molecular markers. The thousands of SSR markers identified in the present study may be useful for analyses of genetic diversity, genetic linkage mapping, and the identification and improvement of varieties during the breeding of *Z. striolatum* Diels. The unigene sequences and SSR markers described herein may serve as valuable resources for future investigations of *Z. striolatum* Diels.

Keywords *Zingiber striolatum* Diels · Transcriptome · SSR · Molecular marker

Introduction

Zingiber striolatum Diels (i.e., white myoga ginger), *Arum sagittifolium*, and *Alpinia galanga* (L.) Willd. are among the perennial vegetables (family: Zingiberaceae) grown in various Chinese provinces (e.g., Guizhou, Sichuan, Guangxi, Hubei, Hunan, Jiangxi, and Guangdong), often at an altitude of about 800 m above sea level (An editorial committee of flora of China 1981). These plants serve as a source of dietary fiber and can be used as medicine as well as food. *Zingiber striolatum* Diels contains protein, many amino acids, polysaccharides, several trace elements, and abundant cellulose. Because of its medicinal properties, this plant species can be used to promote blood circulation for regulating menstrual flow. It may also function as an antitussive expectorant, detumescence agent, and detoxifier (Qu et al. 2015). The food products (e.g., pickles) and beverages prepared from *Z. striolatum* Diels are popular among consumers in major cities around the world. However, there is an insufficient supply of *Z. striolatum* Diels products to satisfy the global demand.

Zingiber striolatum Diels is a unique vegetable grown in China. Although it is sometimes grown under trees in residential areas (e.g., in gardens), *Z. striolatum* Diels is currently primarily harvested from the wild (Qu et al. 2015), with minimal cultivation in artificial plots. However, the diversity in wild varieties as well as the limited growth areas, relatively low root yields, and paucity of related research, have resulted in the inability to grow enough *Z. striolatum* Diels plants to satisfy long-term consumer demands. To date, there have been only a few

✉ Renju Deng
xiaojue63@163.com; 15362220@qq.com

¹ Zunyi Academy of Agricultural Sciences,
Zunyi 563100, Guizhou, China

² Institute of Biotechnology, Guizhou Academy of Agricultural
Sciences, Guiyang 550006, Guizhou, China

³ Institute of Subtropical Crops, Guizhou Academy of
Agricultural Sciences, Guiyang 550006, Guizhou, China

studies on the cultivation of *Z. striolatum* Diels or analyses of plant contents (e.g., polysaccharides, trace elements, water-soluble dietary fiber, and medicinal chemical constituents) (Qu et al. 2015). Additionally, there are no reports describing a molecular-level investigation of this plant species in China or elsewhere.

Simple sequence repeats (SSRs) are among the most effective molecular markers in plant genetics (Powell et al. 1996; Zalapa et al. 2012) and breeding because of several factors, including their simplicity of use, specificity, wide genomic distribution, co-dominance, and their ability to screen for multiple alleles (Cheng et al. 2016). Various SSR markers have been used to identify animal and plant varieties (Rongwen et al. 1995), examine hybrids (Provan et al. 1996), assess genetic diversity (Goldstein et al. 1996), map genes (Chen et al. 2014), assign genes (Yu et al. 1993), investigate gene flow (Moe and Weiblen 2011), and characterize molecular evolution (Kelkar et al. 2008; Wang et al. 2014).

In this study, we examined *Z. striolatum* germplasm resource ZSP11 from Guizhou province, China. An Illumina-based paired-end sequencing technique was used to sequence the transcriptome from different *Z. striolatum* tissues. Additionally, SSR markers were identified and developed at the whole genome level. These markers may be useful for identifying agriculturally important genes and for the genetic improvement and molecular breeding of *Z. striolatum*.

Materials and methods

Plant materials and RNA extraction

Zingiber striolatum Diels germplasm ZSP11 plants were grown in an experimental field at the Zunyi Academy of Agricultural Sciences in Xinzhou (town), Xipu (district), Zunyi (city), China. The root, stem, and leaf tissues were collected at the 4-leaf stage, while the flower and fruit tissues were harvested during the flowering and seed-setting periods, respectively. The collected samples were immediately frozen in liquid nitrogen and stored at -80°C .

Total RNA was extracted from each sample using a modified CTAB method (Zong et al. 2012), and then purified with the RNeasy Plant Mini Kit (Qiagen, Valencia, CA). The quality of the purified RNA was assessed using the 2100 Bioanalyzer RNA Nanochip (Agilent, Santa Clara, CA, USA). The RNA Integrity Number was above 8.5 for the five tissues. The RNA concentration was determined with the ND-1000 spectrophotometer (NanoDrop, Wilmington, DE, USA). A 20- μg combined RNA

sample (i.e., 4 μg RNA from each collected tissue) was used as the template for preparing the cDNA library.

Construction and sequencing of the cDNA library

The mRNA in the extracted total RNA sample was enriched using oligo-dT magnetic beads. The mRNA was fragmented in fragmentation buffer, and then used as the template for first-strand cDNA synthesis with random hexamers. The second strand was synthesized after the addition of buffer, dNTPs, RNase H, and DNA polymerase I. The prepared cDNA was purified with the QIAQuick PCR kit and eluted in EB buffer. An A-tail along with a sequencing adapter were added to the end-repaired purified cDNA. The fragments with the expected sizes were purified following agarose gel electrophoresis for use in polymerase chain reaction (PCR) analyses. The Illumina HiSeq^{TM} 2000 PE100 sequencing system was used for the transcriptome sequencing of the completed library.

Data processing and de novo assembly

Some of the original sequencing data included low quality sequences with adapters. Thus, the data had to be filtered to obtain clean reads that were subsequently analyzed. The quality requirements for *de novo* transcriptome sequencing are much higher than those for re-sequencing because errors can introduce several problems for the algorithms used to assemble short fragments. Therefore, we implemented the following rigorous filtration process: (1) reads in which more than 10% of the sequence was N were removed; (2) reads with more than 50% of the base mass values less than 5 were eliminated; and (3) contaminating adapter sequences were removed.

Sequences were assembled to construct the *de novo* transcriptome using the Trinity program (version 3.0; <http://trinity.github.io/>), which was developed by the Broad Institute and the Hebrew University of Jerusalem. This program assembles a full-length transcript according to the de Bruijn graph theory involving the variable transcript splicing characteristics.

Unigenes were annotated using the NCBI non-redundant (NR) protein database (<http://www.ncbi.nlm.nih.gov/>), Swiss-Prot protein database (<http://expasy.ch/sprot>), Kyoto Encyclopedia of Genes and Genomes (KEGG) database (<http://www.genome.jp/kegg>), and the Clusters of Orthologous Groups database (<http://www.ncbi.nlm.nih.gov/COG>). Annotations were also based on BLASTx comparisons (E value $< 10^{-5}$). The closest matches were used to determine the unigene sequence direction. If there were any inconsistencies among the results for different databases, the priority was given to the results based on the NR

database, followed by the Swiss-Prot, KEGG, and COG databases.

Gene annotation and analysis

The coding regions of the assembled transcripts were identified using the TransDecoder program (<https://transdecoder.github.io/>) according to the following criteria: (1) the open reading frame (ORF) needed to be greater than a certain length; (2) the log value for the likelihood function for the sequences needed to be greater than 0; (3) the highest of six ORF scores was used; and (4) if an ORF contained another one, the longer one was used. The ORFs and unigenes were annotated using the Trinotate program (<http://trinotate.github.io/>). Additionally, the predicted sequences were annotated based on the Uniprot, RNAmmer, eggNOG, Gene Ontology (GO), and KEGG databases.

Identification and development of simple sequence repeat markers

The MicroSatellite program (MISA; <http://pgrc.ipk-gatersleben.de/misa/>) was used to detect SSRs in the unigenes. Simple sequence repeats consist of a repeating unit of 1–6 bp, and are divided into three types according to their composition and distribution (i.e., pure, compound, and interrupted). The SSRs in this study comprised at least four consecutive repeats of 2–6 bp. The SSRs occurred at a frequency of one per kb of cDNA. Primer Premier 6.0 (PREMIER Biosoft International, Palo Alto, CA, USA) was used to design PCR primers. The primers were at least 18 bp long with a melting temperature of 46–55 °C, and were designed to produce an amplification product consisting of 100–350 bp.

Results

Paired-end sequencing and de novo assembly

An Illumina-based paired-end sequencing technique was used, and 2 × 100-bp fragments were obtained for the ends of the sequenced DNA fragments. A total of 146,199,712 sequenced 100-bp fragments were obtained for the 200-bp insert library. The Trinity program was used to assemble the sequences for the *de novo* transcriptome. Following a rigorous quality control and data-filtering step, a total of 13,008,709 high-quality fragments were obtained, and the proportion of the bases with a mass value greater than 30 (error rate less than 0.1%) was 95.16%. An analysis of the assembled sequences revealed 287,959 contigs (Table 1), with an average length of 1029 bp, an N50 of 1684 bp, and

a GC content of 43.11%. The contigs were 201–17,062 bp long. Additionally, approximately 38.5% of the contigs were longer than 1000 bp. A total of 112,107 unigenes were *de novo* assembled, with an average length of 28,891 bp, an N50 of 3356 bp, and a length of 328–25,766 bp.

The sequences for assembling the transcriptome were aligned with the unigene sequences (Table 2) using the Burrows–Wheeler Alignment Tool (<http://bio-bwa.sourceforge.net/bwa.shtml>). The alignment ratios revealed that 95.29% of the unigene sequences were aligned to both reads 1 and 2. Additionally, 4.71% of the unigene sequences were aligned to only read 1 or 2. These results indicated that the assembly quality was relatively high.

Functional annotations based on searches of public databases

To validate and annotate the assembled unigenes, a similarity search was used to align the sequences with those in the NCBI Nr and Swiss-Prot databases (see the Materials and Methods for details). A total of 54,865 (48.94%) of the 112,107 unigenes were significantly similar to known sequences in the Nr database. This similarity corresponded to 36,692 unique protein accessions. Additionally, the 51,804 (46.21%) unigenes that matched sequences in the Swiss-Prot database represented 33,561 unique protein accessions. A total of 40,876 unique protein accessions were identified in the sequence alignments, suggesting that the Illumina-based paired-end sequencing likely identified some important *Z. striolatum* Diels genes.

Gene ontology and clusters of orthologous groups classifications

Gene Ontology involves the use of a set of terms to functionally classify eukaryotic genes in cells. These terms are continually accumulated and changed to reflect advances in life science research. Gene Ontology classifications are mainly based on three categories (i.e., biological process, molecular function, and cellular component). Following the annotations based on the Nr database, the Blast2GO program was used to obtain GO annotation details for the unigenes. The WEGO program was then used to functionally classify the unigenes with GO terms. Finally, 6867 unigenes corresponding to known proteins were assigned 18,295 GO terms. Most of the unigenes were associated with the biological process category (7292; 39.86%), followed by the cellular component (6432; 35.16%) and molecular function (4571; 24.98%) categories (Fig. 1).

The unigenes covered a broad range of GO functional categories. Under the biological process category, cellular

Table 1 Length distribution of the assembled contigs and unigenes

Nucleotides length (bp)	Number of contigs	Number of unigenes
200–399 ^a	97,600	2690
401–599	35,815	3400
601–799	24,209	4980
800–999	19,349	10,911
1000–1199	17,116	9078
1200–1399	16,015	9832
1400–1599	14,140	8903
1600–1799	12,266	9884
1800–1999	10,294	5671
2000–2199	8881	5624
2200–2399	6993	4905
2400–2599	5571	4992
2600–2799	4471	5663
2800–2999	3409	5902
3000–3199	2628	4633
3200–3399	1874	2462
3400–599	1553	2569
3600–3799	1337	1544
3800–3999	895	1642
> 4000 ^b	3543	6822
Total	287,959	112,107
Minimum length (bp)	201	328
Maximum length (bp)	17,062	25,766
Average length (bp)	1029	28,891
N50 (bp) ^c	1684	3356
Total nucleotides length (bp)	296,520,408	221,876,352

^aNumber of contigs and unigenes longer than 200 bp and shorter than 400 bp^bNumber of contigs and unigenes longer than 4000 bp^cN50 length represents the assembly quality. The N50 length is defined as the shortest contig or unigene length representing 50% of the total assembled length**Table 2** Assessment of assembly quality

Read type	Count	Percentage (%)
Pair mapping	113,767,228	95.29
Right only	2,813,394	2.36
Left only	2,808,662	2.35
Total aligned reads	119,389,284	100

process (589 unigenes) and metabolic process (526 unigenes) were the most abundant sub-categories, suggesting the importance of some metabolic activities for *Z. striolatum* Diels development. Interestingly, 135 genes belonged to the pigmentation sub-category. Additionally, 116 unigenes were involved in different stress responses. Under the molecular function category, binding (648 unigenes) and catalytic (632 unigenes) were the most important sub-

categories. The most important type of binding was to proteins, followed by ions, ATP, DNA, and then RNA. Under the cellular component category, most genes were related to the cell and cell part sub-categories, while others were associated with the organelle and organelle part sub-categories.

The classifications based on the COG database were determined by individually comparing complete protein-encoding genomic sequences (orthologous genes). The proteins corresponding to each COG were all assumed to come from the same ancestral protein. When considering proteins encoded by a given genome, the comparisons identified the most similar proteins encoded by other genomes. Each of the protein-encoding genes were assessed in turn. The best matches between proteins formed a COG class. All of the unigenes identified in this study were aligned to the COG database sequences to predict and

Fig. 1 Gene ontology classifications of the assembled unigenes

Fig. 2 Clusters of orthologous groups (COG) classifications. All unigenes were aligned to COG database sequences to predict and classify possible functions. Of the 102,456 matches to the Nr database sequences, 54,865 sequences were assigned to 25 COG classes

classify their possible functions. A total of 54,865 sequences were attributed to 25 COG classes (Fig. 2).

Of the 25 COG classes, general function prediction only consisted of the most unigenes (9825; 18.90%), followed by replication, recombination, and repair (5966; 11.46%), transcription (4750; 9.14%), translation, ribosomal structure, and biosynthesis (4746; 9.13%), post-translational modification, protein turnover, and chaperones (4280; 8.23%), signal transduction mechanisms (4198; 8.08%), carbohydrate transport and metabolism (3980; 7.66%), and amino acid transport and metabolism (3878; 7.46%). A few genes were related to the nuclear and extracellular structures.

Development and characterization of simple sequence repeat markers

The predicted unigenes were used to further assess assembly quality and develop new molecular markers. The MISA program was used to search for SSR loci in the assembled *Z. striolatum* Diels sequences. A total of 8384 potential SSRs were detected at a frequency of one per 35.4 kb. These SSRs were distributed on 2392 unigenes, with 965 SSRs (11.51%) detected at more than one locus. Additionally, there were 762 compound SSRs (9.09%). The *Z. striolatum* Diels SSRs mainly comprised six repeats (3766; 44.91%), followed by seven, nine, and 10–20 repeats (3850; 45.92%). Only six SSRs consisted of more than 20 repeats (0.07%, Table 3).

The repeating units of the detected SSRs were mainly dinucleotides (25.99%) and trinucleotides (62.87%, Tables 3 and 4). These sequences were primarily repeated six, seven, or nine times. The dinucleotide repeat sequences were mainly 18 bp long, corresponding to about 40.80% of the dinucleotide repeats. The most common dinucleotide repeats were AG/CT and GA/TC, accounting for about 35.74 and 28.41% of the dinucleotide repeats, respectively. The GC/GC sequence was the least common dinucleotide repeat (0.21%). The trinucleotide repeat sequences were mainly 18–21 bp long, corresponding to about 97.32% of the trinucleotide repeats, with AGG/CCT being the most common (13.82%), followed by GGA/TCC (8.00%). The TAC/GTA sequence was the least common trinucleotide

repeat (0.02%). The tetranucleotide repeats accounted for 11.14% of the repeat sequences.

Primers were designed for 2392 unigenes containing SSR loci. A total of 5623 pairs of SSR-specific primers were designed, accounting for 55.14% of the SSR loci. Thirty primer pairs specific for different repeating units (i.e., dinucleotides, trinucleotides, and tetranucleotides) were randomly selected for PCR amplifications of *Z. striolatum* Diels ZSP11 DNA. Twenty-eight primer pairs (93.33%) amplified clear bands (Fig. 3), including 25 amplification products that were consistent with the expected size. Two amplification products were longer than expected, while one product was smaller than expected. The 25 validated primer pairs may be used to analyze the genetic diversity of *Z. striolatum* germplasms (Fig. 3, Table 5).

Discussion

The rapid development of second-generation sequencing technology continues to expand the available data regarding whole genomes and transcriptomes, which has increased the abundance of genome-wide SSR markers. In recent years, many researchers have studied SSRs using high-throughput sequencing techniques that have been widely applied for genetic diversity analyses, map building, gene mining, and the identification and improvement of plant varieties (Cheng et al. 2016; Chen et al. 2014; Yi et al. 2006; Portis et al. 2007; Tan et al. 2015).

In this study, 287,959 contigs were obtained for *Z. striolatum* Diels following high-throughput sequencing, assembly, and alignments. Additionally, 112,107 unigenes were *de novo* assembled. All of the unigenes were annotated according to alignments with sequences in publically available databases and bioinformatics analyses. A total of 40,876 unique protein accessions were identified. Furthermore, GO and COG functional annotations revealed some important activities related to *Z. striolatum* Diels development.

The assembled *Z. striolatum* Diels sequences were used to analyze the distribution of SSRs, suggesting SSR markers may be relevant to future attempts in identifying

Table 3 Distribution of simple sequence repeats with different motif types and number of repeats in the *Z. striolatum* Diels genome

Motif length/bp	Repeat number						> 20	Total	Ratio (%)
	5	6	7	8	9	10–20			
2	0	0	0	0	889	1285	5	2179	25.99
3	0	3462	1668	135	0	5	1	5271	62.87
4	616	304	1	11	1	1	0	934	11.14
Total	616	3766	1669	146	889	1292	6	8384	–
Ratio /%	7.35	44.92	19.91	1.74	10.6	15.41	0.07	–	–

Table 4 Distribution of simple sequence repeat types in the *Z. striolatum* Diels genome

SSRs motif	Repeat number	SSRs motif	Repeat number	SSRs motif	Repeat number
AC/GT	578	AGA/TCT	2025	CGC/GCG	875
AG/CT	7685	AGC/GCT	1705	CTA/TAG	70
AT/AT	2308	AGG/CCT	4647	CTC/GAG	2444
GC/GC	45	AGT/ACT	122	GAA/TTC	2903
CA/TG	1930	ATA/TAT	553	GAC/GTC	302
GA/TC	6109	ATC/GAT	701	GCA/TGC	1706
TA/TA	2846	ATG/CAT	431	GCC/GGC	2268
AAC/GTT	298	CAA/TTG	339	GGA/TCC	2685
AAG/CTT	2088	CAC/GTG	340	TAA/TTA	742
AAT/ATT	1569	CAG/CTG	1293	TAC/GTA	7
ACA/TGT	202	CCA/TGG	846	TCA/TGA	279
ACC/GGT	468	CCG/CGG	1112		
ACG/CGT	277	CGA/TCG	324		

Fig. 3 PCR product of polymorphism of SR primers in different *Z. striolatum* Diels varieties. **a** cSSR16; **b** cSSR21. ZSP1 – ZSP21 represent 1–21 varieties, respectively

and classifying *Z. striolatum* Diels varieties. A search of all unigenes detected 8384 SSR loci (7.48%), with a frequency of one per 35.4 kb. This SSR frequency is higher than that of bananas (5.3%) (Wang et al. 2008), sugarcane (2.9%) (Cordeiro et al. 2001), cotton (4.64%) (Li et al. 2005), rice (4.7%) (Kantety et al. 2002), wheat (3.2%) (Liu et al. 2012), and sorghum (3.6%) (Cordeiro et al. 2001; Kantety et al. 2002). However, it is slightly lower than that of

pepper (7.83%) (Liu et al. 2012), and much lower than that of Chinese cabbage (10.4%) (Ge et al. 2005), tea (21.56%) (Jin et al. 2006), coffee (17.3%) (Aggarwal et al. 2007), and *Lonicera caerulea* (32.51%) (Zhang et al. 2016). These differences may be related to variabilities in the SSR search criteria, database availability, and species.

A previous study concluded that the expressed sequence tag (EST)-SSRs of most plants consist primarily of

Table 5 Details of 25 primer pairs used for polymerase chain reaction analyses

Primer no.	Source	Primer sequence (5'→3')	SSR motif	Length of product
cSSR01	comp101033_c0_seq 3	CGATCGAGGCGTACACAG GAGGAGCGGCTTCTTAGGAT	(AG) ₁₁	224
cSSR02	comp101573_c0_seq 1	AATGGCTCGGGAGTCAAGAT GGCCAGTTTGAGCGTGTC	(AGG) ₆	233
cSSR04	comp102711_c0_seq 1	ATGAAGCCGTGAACGAGAAG TCGATCGTGCTCAGTCTCTG	(AGA) ₇	178
cSSR06	comp103178_c0_seq 2	GCATTGCTGAAGAAGGGAAG TTGTTCTCCATTGGCTTGG	(GAT) ₆	200
cSSR07	comp104297_c0_seq 2	ACCCCTCTCGCCCTCTTAT ATGCGGCAGCAGATCATAG	(GCC) ₆	209
cSSR09	comp105337_c0_seq 4	GTCAGTTCCGGGGAGGTAAT GACCGAAGACGAAGTCGATG	(CAA) ₆	225
cSSR10	comp107044_c0_seq 8	ATTTGTGGACCCGATCTACG GCTTGATGACCCTGTGGAG	(CGC) ₆	233
cSSR11	comp107202_c0_seq 1	CCCACCATAACCCTACGTTGT GCTCTACCTTTAAGTGCCTTGG	(CCT) ₆	208
cSSR13	comp108504_c4_seq 9	TCCCATTTCTCTGTGAGTT GGACGGAAGTCGTAATCTGG	(GG) ₉	245
cSSR14	comp109879_c1_seq 1	CTCCTCTCTTCGCTCCAAAA GCCTCCTCTCCCATGTCTCT	(GA) ₁₀	205
cSSR15	comp111587_c0_seq 6	CGGATCAGAACTTCCCTGAC GGACAATTACGCCGACAAAG	(AC) ₉	204
cSSR16	comp112622_c1_seq 3	ACGAAGCTCCCTAGCTGACA TTTTCTTTGGGTTGCAAGG	(GAA) ₆	241
cSSR17	comp113200_c2_seq 1	CTCTTCATTGGTGGAAAGCA GGCATCCTCAAAGACTGCTG	(CAT) ₆	220
cSSR18	comp27703_c0_seq 1	ATGGACGGCCATGACTATGT TTTGGGTTGGAGAGAGTTGG	(AAC) ₆	214
cSSR20	comp50668_c0_seq 1	CTTACCCACCCTCTCCTTCC ATGAAAGCCCGAGGTCAAG	(TTG) ₆	204
cSSR21	comp55565_c0_seq 1	CGACAATTAAAGATAACATCCCAAC CCACGTTATGATCGAAATGG	(GAT) ₆	223
cSSR22	comp90577_c0_seq 1	GCGTGTACTCGTGAAATTG GGCTCACTTATGCCTTCGTC	(GGA) ₆	273
cSSR23	comp92597_c0_seq 1	TTGAGAAGGCGTCAGGTACA AAGTCCTGCCATCAAAATGC	(CATC) ₅	205
cSSR25	comp95230_c0_seq 1	AGGGAAAGCAAGGAAAGGAA TCGATCCTCTGTTCTGCAAC	(TCC) ₆	190
cSSR26	comp96728_c0_seq 2	AGGAGATTGCCATTGACGAC CGGTTTCGGTAAGTTCACCTG	(TTC) ₆	216
cSSR27	comp97410_c3_seq 1	CGACACGTCTTCATGGATCT TCTATGACGACCCTCGGAAT	(TGC) ₇	187
cSSR28	comp98468_c0_seq 1	GACAGACATTATTGGGGGAAAA GCAGAAAGGCTGCTGGAAT	(TGAT) ₅	218
cSSR29	comp99040_c0_seq 2	CATGCTCCTCTGCTGGTACA TCATCAATTCTTGGGGAAAA	(GCT) ₆	205
cSSR30	comp101873_c0_seq 1	GGGATTGGATTGGTATCTTTGA TGAAGGGTGTTTTAGTCTTTTCC	(ATTT) ₅	280

trinucleotide repeats (Liang et al. 2009). In contrast, the EST-SSRs for a few dicotyledonous plants mainly comprise dinucleotide repeating units (Kumaputra and Mukhopadhyay 2005). Most of the SSRs identified in the current study contained a trinucleotide repeat, followed by a dinucleotide repeating unit. These observations are consistent with the previously reported EST-SSR results for rice, maize, soybean, tomato, cotton, poplar, and *Arabidopsis thaliana* (Cardle et al. 2000). Similar findings were described for EST analyses of major cereal crops (Varshney et al. 2002) and peppers (Liu et al. 2012). The main trinucleotide repeats in *Z. striolatum* Diels are AGG/CCT and GGA/TCC/AGG (21.82%), which is in contrast to the dominant trinucleotide repeats in pepper (AAC/GTT), corn (CCG/GGC), rice (AGG/TCC), sorghum and soybean (AAG/TTC), tomato (AAT/ATT), and banana (AAG/CTT). These differences may be related to variations in EST-SSR features, and EST data sources and quantity.

A total of 5623 SSR-specific primer pairs were designed according to the 2392 unigene sequences. Of the 30 randomly selected primer pairs, 25 amplified products were of the expected size. These primers may be useful for identifying and improving *Z. striolatum* Diels varieties. They may also have applications related to resource analyses, construction of genetic maps, and the functional characterization of genes.

Acknowledgements Professor William Yajima is gratefully acknowledged for correction to the manuscript. This research was supported by the Guizhou innovation talent base construction of potato industry technology ([2016]22), the technical innovation fund for small and medium-sized enterprises funded projects (13C26215205306) and Guizhou science and technology project ([2016]2554).

References

- Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V et al (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor Appl Genet* 114(2):359–372
- An editorial committee of flora of China (1981) *Zingiber striolatum* Diels. Flora of China. Science Press, Beijing, p 146
- Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D et al (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156(2):847–854
- Chen H, Song Y, Li LT, Khan MA, Li XG et al (2014) Construction of a high-density simple sequence repeat consensus genetic map for pear (*Pyrus* spp.). *Plant Mol Biol Rep* 33(2):1–10
- Cheng JW, Zhao ZC, Li B, Qin C, Wu ZM et al (2016) A comprehensive characterization of simple sequence repeats in pepper genomes provides valuable resources for marker development in *Capsicum*. *Sci Rep* 6:189–190
- Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci* 160(6):1115–1123
- Ge J, Xie H, Chui CS, Hong JM, Ma RC (2005) Analysis of expressed sequence tags (ESTs) derived SSR markers in Chinese cabbage (*Brassica campestris* L. ssp. *pekinensis*). *J Agric Biotechnol* 13(4):423–428
- Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1996) An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139(1):463–471
- Jin JQ, Cui HR, Chen WY, Lu MZ, Yao YL et al (2006) Data mining for SSRs in ESTs and development of EST-SSR marker in tea plant (*Camellia sinensis*). *J Tea Sci* 26(1):17–23
- Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol* 48:501–510
- Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* 18(1):30–38
- Kumaputra SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48:985–998
- Li HS, Fan SL, Sheng FF (2005) Screening of microsatellite markers from cotton ESTs. *Cotton Sci* 17(4):211–216
- Liang X, Chen X, Hong Y, Liu H, Zhou G et al (2009) Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species. *BMC Plant Biol* 9:35–48
- Liu F, Wang YS, Tian XL, Mao ZC, Zou XX et al (2012) SSR mining in pepper (*Capsicum annuum* L.) transcriptome and the polymorphism analysis. *Acta Horti Sin* 39(1):168–174
- Moe AM, Weiblen GD (2011) Development and characterization of microsatellite loci in dioecious figs (*Ficus*, Moraceae). *Am J Bot* 98(2):e25–e27
- Portis E, Nagy I, Sasvari Z, Stägel A, Barchi L et al (2007) The design of *Capsicum* spp. SSR assays via analysis of in silico DNA sequence, and their potential utility for genetic mapping. *Plant Sci* 172:640–648
- Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci* 1(7):215–222
- Provan J, Kumar A, Shepherd L, Powell W, Waugh R (1996) Analysis of intra-specific somatic hybrids of potato (*Solanum tuberosum*) using simple sequence repeats. *Plant Cell Rep* 16(3–4):196–199
- Qu L, Xia LS, Liu D, Feng WY (2015) Progress in *Zingiber striolatum* Diels. *Yunnan J Tradit Chin Medi Mater Med* 5:111–113
- Rongwen J, Akkaya M, Bhagwat A, Lavi U, Cregan P (1995) The use of microsatellite DNA markers for soybean genotype identification. *Theor Appl Genet* 90(1):43–48
- Tan S, Cheng JW, Zhang L, Qin C, Nong DG et al (2015) Construction of an interspecific genetic map based on InDel and SSR for mapping the QTLs affecting the initiation of flower primordia in pepper (*Capsicum* spp.). *PLoS ONE* 10:1–15
- Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol Biol Lett* 7(2A):537–546
- Wang JY, Chen YY, Liu WL, Wu YT (2008) Development and application of EST-derived SSR markers for bananas (*Musa nana* Lour.). *Hereditas* 30(7):933–940
- Wang HL, Yang J, Boykin LM, Zhao QY, Wang YJ et al (2014) Developing converted microsatellite markers and their implications in evolutionary analysis of the *Bemisia tabaci* complex. *Sci Rep* 4:1–10
- Yi G, Lee JM, Lee S, Choi D, Kim BD (2006) Exploitation of pepper EST-SSRs and an SSR-based linkage map. *Theor Appl Genet* 114:113–130

- Yu Y, Saghai Maroof M, Buss G, Maughan P, Tolin S (1993) RFLP and microsatellite mapping of a gene for soybean mosaic virus resistance. *Phytopathology* 84(1):60–64
- Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D et al (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am J Bot* 99:193–208
- Zhang QT, Li XY, Yang YM, Fan ST, Ai J (2016) Analysis on SSR information in transcriptome and development of molecular markers in *Lonicera caerulea*. *Acta Horti Sin* 43(3):557–563
- Zong XJ, Wang WW, Wang JW, Wei HR, Yan XR et al (2012) The application of SYBR Green I real-time quantitative RT-PCR in quantitative analysis of sweet cherry viruses in different tissues. *Acta Phytolacica Sin* 39(6):497–502