Genetics and Molecular Research

Development, characterization, and annotation of potential simple sequence repeats by transcriptome sequencing in pears (*Pyrus pyrifolia* Nakai)

H. Zhou, B.H. Cai, Z.Q. Lü, Z.H. Gao and Y.S. Qiao

Laboratory of Fruit Tree Biotechnology, College of Horticulture, Nanjing Agricultural University, Nanjing, China

Corresponding author: Y.S. Qiao
E-mail: qiaoyushan@njau.edu.cn

Genet. Mol. Res. 15 (3): gmr.15038683
Received March 31, 2016
Accepted July 15, 2016
Published September 23, 2016
DOI http://dx.doi.org/10.4238/gmr.15038683

ABSTRACT. Simple sequence repeats (SSRs), one of the most powerful molecular markers, can be used for DNA fingerprinting, variety identification, genetic mapping, and marker-assisted selection. Using the pear's (*Pyrus pyrifolia* Nakai) 75,764 unigenes (55,676,271 bp) obtained by deep transcriptome sequencing, a total of 10,622 novel SSRs were identified in 9154 unigenes, accounting for 14.02% of all unigenes. The average length and distribution of these SSRs was about 16 bp and 5.24 kb, respectively. Dinucleotide repeat motifs were the main type, with a frequency of 55.87%, followed by trinucleotides (24.45%). There were 159 kinds of repeat motifs existing in the pear transcriptome. AG/CT was the most frequent motif, accounting for 49.64%. All 9154 SSR-containing unigenes were functionally annotated using Nr (NCBI non-redundant protein database), Nt (NCBI non-redundant nucleotide

database), and the Swiss-Prot database, and were classified further by Gene Ontology and Clusters of Orthologous Groups. In addition, a total of 4300 primer pairs were designed from all SSR loci obtained. Of these, 40 primers were randomly selected for PCR amplification and polyacrylamide gel (PAGE) analysis. Among the 40 primer pairs, 31 were successfully separated via PAGE. These findings also confirm that mining SSRs using next-generating sequencing technologies is a fast, effective, and reliable approach.

Key words: Simple sequence repeats; *Pyrus pyrifolia* Nakai; Transcriptome; Development of primers; Functional annotation

INTRODUCTION

Pears (*Pyrus* spp) belong to the genus *Pyrus* within the family Rosaceae and are the third most important temperate fruits after grapes and apples (Potter et al., 2007; Wu et al., 2013). *Pyrus* is considered to be typically self-incompatible. Its phylogenetic relationships are complicated and difficult to establish because of high heterozygosity and poor morphological diversity. Hence, the development of highly polymorphic molecular markers is essential for resolving these problems in breeding programs and other genetic research on pears.

SSRs, or microsatellites, are repetitive DNA sequences consisting of motif repeats (1-6 bp) with conservative flanking sequences. SSR markers are among the most valuable and efficient molecular markers because of their good genome coverage, genetic co-dominant inheritance, polymorphism, and reproducibility (Tautz and Renz, 1984; Shen et al., 2015). In addition, since the use of SSRs is based on PCR amplification, the technique is simple and only a small amount of DNA is required. SSRs can be classified into two categories according to their source: genomic SSRs (DNA-derived SSRs) and EST-SSRs/genic SSRs (expressed sequence tag-SSRs). DNA-derived SSRs are derived from SSR-enriched genomic libraries or random genomic sequences. In contrast, EST-SSRs are usually derived from the transcriptome or EST sequences in public databases. Compared with DNA-derived SSRs, EST-SSRs are considered more valuable molecular markers, to a certain extent, because of their functionality, co-specific transferability, and polymorphisms, especially in some specific genetic studies.

Traditional methods for the development of DNA-derived SSRs are expensive, time-consuming, and laborious. SSRs developed from publicly available genetic/genomic information of the species of interest to us, is the most effective method. However, the dearth of available sequences is a serious obstacle in the application of this method. Mining EST-SSR markers based on transcriptome information generated by large-scale sequencing (such as Roche/454, Illumina/Solexa, and ABI/Solid platforms) is an effective strategy to overcome this problem.

The genome of *Pyrus bretschneideri* Rehd. 'Dangshansu' has been previously sequenced (Wu et al., 2013) and used to develop genomic SSRs (Fan et al., 2013). Some developments in pear EST-SSRs have also been recently reported (Yue et al., 2014; Zhang et al., 2014). However, only a few EST-SSRs, which were developed from pear transcriptome sequencing, have been reported. The number of publicly available SSRs is still too low for use in breeding selection, phylogenetic relationship analysis, and other genetic studies in pears. Hence, developing more EST-SSRs from pears would be desirable.

To our knowledge, we here report the first case of functional annotation of pear

genic SSRs-containing sequences, and this is also one of the few studies where EST-SSRs were developed from the pear transcriptome. The frequency, type, distribution, and other features of these EST-SSRs are also reported in our paper. The thousands of EST-SSRs reported in our paper can be used to enrich molecular markers and accelerate genetic research in pears, and they can also be used in genetic studies of other members of Rosaceae, because of its high co-specific transferability and polymorphism, especially in those species with limited available sequences.

MATERIAL AND METHODS

Plant materials and data sources

The transcriptome data used in this study were from high-throughput sequencing of *Pyrus pyrifolia* Nakai 'Huanghua' peels. *P. pyrifolia* 'Huanghua' was grown in the Lishui Orchard (Nanjing, China). Fruits were collected at the following developmental times: 6, 7, 8, and 9 weeks after flowering. Total RNA was isolated by the modified CTAB method (Wang et al., 2010a). RNA quality was confirmed by analyzing samples with a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), with a minimum RIN value of 8.2. In total, 20 μg RNA was equally pooled from the four samples for cDNA library preparation. Using a Solexa/Illumina HiSeq™ 2000 RNA-seq platform, 53,856,740 high-quality reads (more than 4.8 G) were obtained, trimmed, and assembled into 75,764 unigenes.

DNA extraction

Total genomic DNA was extracted from fresh, young *P. pyrifolia* Nakai 'Huanghua' leaves by the modified CTAB method (Pan et al., 2006). DNA quality and quantity was confirmed with 1% agarose E-Gels (Invitrogen, Shanghai, China) and an Eppendorf BioPhotometer (Eppendorf, Hamburg, Germany), respectively. The DNA was diluted in sterilized ddH_2O to a concentration of 100 ng/μL and stored at -20°C for PCR analysis.

SSR locus search

The MIcroSAtellite software (MISA, http://pgrc.ipk-gatersleben.de/misa/), which is based on the Perl language, was performed to detect SSR loci using all the 75,764 unigenes as references. SSRs were defined as mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide, with a minimum of 12, 6, 5, 5, 4, and 4 repetitions for all motifs, respectively. The SSRs detected were classified into two groups: perfect SSRs and compound SSRs. The maximum number of bases interrupting perfect SSRs in a compound SSR was 100.

SSR primer design

Primer3 v2.3.4 was used to design primers in the flanking regions (when greater than 150 bp) of SSR loci on unigenes. The design process was based on the following criteria: 1) The length of primers was between 18-28 bp, with an optimum size of 23 bp; 2) an annealing temperature between 55°-65°C, with a maximum discrepancy within 2°C among forward/reverse primers; and 3) PCR product size ranging from 80 to 160 bp. Other parameters were set at the default setting of the software.

Five pairs of primers were designed for each SSR locus. Then, primers were screened as follows: 1) no SSRs were present within the primers; 2) the primers were aligned to unigene sequences, with the 5'-site containing no more than 3 mismatches; 3) removal of the primers that were aligned to more than one unigene; and 4) SSRs on the product sequences were located with SSR_finder (http://www.fresnostate.edu/ssrfinder/), and the products retained from SSR_finder's results were the same as those of MISA.

Verification of SSR primers

Forty SSR primer pairs were randomly selected from the 4300 primers for PCR amplification. Each PCR was carried out in a total final volume of 25 μL, containing: 1 μL 100 ng/μL genomic DNA, 2.5 μL 10X PCR Buffer, 2.0 μL 2.5 mM dNTP, 1.5 μL $MgCl_2$ (25 mM), 2.0 μL forward and reverse primer (10 pM), and 0.2 μL 5 U/μL Taq DNA polymerase (Takara Biotechnology, Dalian, China). All SSR amplifications were performed under the following conditions: initial denaturation temperature at 94°C for 4 min, followed by 8 cycles at 94°C for 30 s, 59°C for 30 s, and 72°C for 30 s, where the annealing temperature was reduced by 0.5°C per cycle. This was then followed by 30 cycles at 94°C for 30 s, 56°C for 30 s, 72°C for 30 s, and a final extension for 10 min at 72°C. The amplification products and the same amount of ladder markers (20-bp DNA Ladder Marker, D512A) were loaded onto an 8% non-denaturing PAGE. The gel was silver-stained and photographed. Taq DNA polymerase, dNTP, and DNA Ladder Marker were purchased from Takara Biotechnology (Dalian).

Functional annotation for SSR-containing unigenes

Functional annotation was also performed. All SSR-containing unigenes were searched using BLAST against the Nr, Nt, Swiss-prot, and COG, with a cut-off e-value of 10^{-5}. These SSR-containing unigene sequences were first aligned with BLASTx to the protein databases Nr and Swiss-prot, and aligned by BLASTn to the nucleotide database Nt (e-value of 10^{-5}). Functional classification by GO was analyzed with Blast2GO (Conesa et al., 2005). After acquiring a GO annotation for every unigene, the WEGO software was used to perform GO functional classification (Ye et al., 2006). The unigenes were also aligned to the COG database to predict and classify possible functions using the BLAST software (e-value of 10^{-5}).

RESULTS

Frequency, distribution, and other features of SSRs in the pear transcriptome

Of the 75,764 unigenes (55,676,271 bp) derived from the 'Huanghua' pear transcriptome, a total of 10,622 SSRs containing 1-6-bp repeat motifs were identified in 9154 unigenes, of which 1261 unigenes contained more than one SSR (Table 1). Six hundred and forty-four compound SSRs were also detected. In the SSR statistics and primer pairs designed, we separated a compound SSR into several perfect SSRs.

Mononucleotide and dinucleotide were the major types, accounting for 80.32% of all SSR motifs (Table 2). The most abundant motif was dinucleotide (5935, 55.87%), followed by trinucleotide (2597, 24.45%) and mononucleotide (1539, 14.49%). The total number of tetranucleotides, pentanucleotides, and hexanucleotides was only 551 (5.19%). The frequency

and average distribution distance of these 10,622 SSRs was about 14.02% and 5.24 kb, respectively. In our study, the total lengths of all SSRs were up to 173,342 bp, with an average length of 16 bp, including mononucleotides (23,819; 15 bp), dinucleotides (93,506; 16 bp), trinucleotides (43,746, 17 bp), tetranucleotides (3016; 21 bp), pentanucleotides (4455; 21 bp), and hexanucleotides (4800; 25 bp). The incidence of different repeats and frequencies of each motif was also evaluated (Table 2). The most common class was N = 6 (21.34%), which contained mostly dinucleotide repeats, followed by N = 5 (15.96%) and N = 7 (13.67%).

Table 1. SSR search results using MISA.

Item	Number
Total number of unigenes examined	75,764
Total size of examined sequences (bp)	55,676,271
Total number of identified SSRs	10,622
Number of SSR-containing unigenes	9154
Number of unigenes containing more than one SSR	1261
Number of compound SSRs	644

Table 2. Frequency of different SSR repeat motifs.

Repeat number	SSR motif						Total	%
	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-		
4	-	-	-	-	192	165	357	3.36
5	-	-	1539	123	21	12	1695	15.96
6	-	1571	664	22	3	7	2267	21.34
7	-	1117	331	1	0	3	1452	13.67
8	-	1065	43	0	0	1	1109	10.44
9	-	1116	0	0	0	1	1117	10.52
10	-	857	3	0	0	0	860	8.10
11	-	186	2	0	0	0	188	1.77
12	353	19	10	0	0	0	382	3.60
13	271	0	0	0	0	0	271	2.55
14	178	0	5	0	0	0	183	1.72
>14	737	4	0	0	0	0	741	6.98
Total	1539	5935	2597	146	216	189	10,622	100
%	14.49	55.87	24.45	1.37	2.03	1.78	100	

Within the detected SSRs, 159 kinds of repeat motifs were identified, of which mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeat motifs had 2, 4, 10, 23, 45, and 75 types, respectively (Table 3). AG/CT was the most frequent motif, accounting for 49.64% of all SSRs, followed by A/T (1515, 14.26%), AAG/CTT (662, 6.23%), AGG/CCT (544, 5.12%), AGC/CTG (444, 4.18%), AC/GT (391, 3.68%), ACC/GGT (335, 3.15%), AT/AT (260, 2.45%), and ATC/ATG (204, 1.92%). The frequency of the remaining 150 types of motifs accounted for 9.36% (Table 4).

AG/CT comprised 88.85% of all dinucleotide motifs and was the most common type among all SSR repeats (Figure 1A). The predominant trinucleotide motifs were AAG/CTT (25.49%) and AGG/CCT (20.95%) (Figure 1B).

Development and verification of SSR primers

All 10,622 SSRs were screened, and the SSRs wherein the lengths of both flanking regions on the unigene greater than 150 bp were kept to design primers. Finally, 2526 (23.78%) SSRs were retained, and 12,630 primer pairs were designed.

Table 3. Type and number of repeat motifs.

Repeat type	Repeat motif	Number
Mono-	A/T, C/G	2
Di-	AC/GT, AG/CT, AT/AT, CG/CG	4
Tri-	AAC/GTT, AAG/CTT, AAT/ATT, ACC/GGT, ACG/CGT, ACT/AGT, AGC/CTG, AGG/CCT, ATC/ATG, CCG/CGG	10
Tetra-	AAAC/GTTT, AAAG/CTTT, AAAT/ATTT, AACC/GGTT, AACT/AGTT, AAGC/CTTG, AAGG/CCTT, AATC/ATTG, AATG/ATTC, AATT/AATT, ACAG/CTGT, ACAT/ATGT, ACCG/CGGT, ACGT/ACGT, ACTC/AGTG, ACTG/AGTC, AGAT/ATCT, AGCC/CTGG, AGCG/CGCT, AGGC/CCTG, AGGG/CCCT, ATCC/ATGG, ATCG/ATCG	23
Penta-	AAAAC/GTTTT, AAAAG/CTTTT, AAAAT/ATTTT, AAACC/GGTTT, AAACG/CGTTT, AAAGC/CTTTG, AAAGG/CCTTT, AAATC/ATTTG, AAATT/AATTT, AACAC/GTGTT, AACAG/CTGTT, AACCC/GGGTT, AACCG/CGGTT, AACGG/CCGTT, AAGAG/CTCTT, AAGAT/ATCTT, AAGGG/CCCTT, AATAC/ATTGT, AATCC/ATTGG, AATCG/ATTCG, AATCT/AGATT, AATGG/ATTCC, AATTC/AATTG, ACACG/CGTGT, ACAGC/CTGTG, ACAGT/ACTGT, ACCAG/CTGGT, ACCAT/ATGGT, ACCTC/AGGTG, ACGCC/CGTGG, ACGGG/CCCGT, ACTAG/AGTCT, ACTCC/AGTGG, ACTCG/AGTCG, ACTGG/AGTCC, AGAGC/CTCTG, AGAGG/CCTCT, AGATC/ATCTG, AGATG/ATCTC, AGCTC/AGCTG, AGGCC/CCTGG, AGGGG/CCCCT, ATCCC/ATGGG, ATCCG/ATCGG, ATCGC/ATGCG	45
Hexa-	AAAAAC/GTTTTT, AAAAAG/CTTTTT, AAAAAT/ATTTTT, AAAACC/GGTTTT, AAAAGC/CTTTTG, AAAATT/AATTTT, AAACCG/CGGTTT, AAACGG/CCGTTT, AAAGAG/CTCTTT, AAAGCC/CTTTGG, AAAGGG/CCCTTT, AAATAC/ATTTGT, AAATCG/ATTTCG, AAATGG/ATTTCC, AAATTC/AATTTG, AACACC/GGTGTT, AACCAC/GGTTGT, AACCAG/CTGGTT, AACCCC/GGGGTT, AACCCG/CGGGTT, AACCCT/AGGGTT, AACCGC/CGGTTG, AACTCC/AGTTGG, AACTTG/AAGTTC, AAGACG/CGTCTT, AAGAGG/CCTCTT, AAGATG/ATCTTC, AAGCAC/CTTGTG, AAGCAG/CTGCTT, AAGGAG/CCTTCT, AAGGGG/CCCCTT, AAGGTC/ACCTTG, AAGTCC/ACTTGG, AATCCC/ATTGGG, AATCTC/AGATTG, AATGGG/ATTCCC, AATTCC/AATTGG, ACACAG/CTGTGT, ACACCC/GGGTGT, ACACCT/AGGTGT, ACAGAG/CTCTGT, ACAGGC/CCTGTG, ACAGGG/CCCTGT, ACATCC/ATGTGG, ACATGG/ATGTCC, ACCACG/CGTGGT, ACCAGC/CTGGTG, ACCATC/ATGGTG, ACCATG/ATGGTC, ACCGAG/CGGTCT, ACCGCC/CGGTGG, ACCGTC/ACGGTG, ACCTCC/AGGTGG, ACGAGG/CCTCGT, ACGATC/ATCGTG, ACGGAG/CCGTCT, ACGGCG/CCGTCG, ACGGGC/CCCGTG, ACTAGC/AGTGCT, ACTCCC/AGTGGG, ACTCTC/AGAGTG, ACTGCG/AGTCGC, ACTGCT/AGCAGT, AGAGCC/CTCTGG, AGAGGC/CCTCTG, AGAGGG/CCCTCT, AGATCT/AGATCT, AGCAGG/CCTGCT, AGCCTC/AGGCTG, AGGATG/ATCCTC, AGGCGG/CCGCCT, AGGGAT/ATCCCT, AGGGGG/CCCCCT, ATCCCC/ATGGGG, ATCGCC/ATGGCG	75

Table 4. Frequency distribution of different SSR motifs.

SSR motif	Frequency of SSR motif	%	SSR motif	Frequency of SSR motif	%
A/T	1515	14.26	AGC/CTG	444	4.18
C/G	24	0.23	ACC/GGT	335	3.15
AG/CT	5273	49.64	ATC/ATG	204	1.92
AC/GT	391	3.68	CCG/CGG	114	1.07
AT/TA	260	2.45	AAC/GTT	109	1.03
CG/GC	11	1.10	ACG/CGT	107	1.01
AAG/CTT	662	6.23	AAT/ATT	46	0.43
AGG/CCT	544	5.12	ACT/AGT	32	0.30
Others	551	5.19			

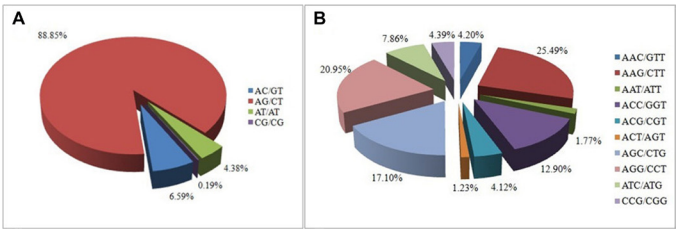

Figure 1. Percentage of different dinucleotide and trinucleotide motifs.

These primers were filtered by the previously mentioned criteria, and 4300 primers were obtained, including 30 (0.70%) for mononucleotide repeats, 1388 (32.28%) for dinucleotide repeats, 2623 (61.00%) for trinucleotide repeats, 49 (1.14%) for tetranucleotide repeats, 97 (2.26%) for pentanucleotide repeats, and 113 (2.63%) for hexanucleotide repeats (**Table S1**). AT/TA did not exist among these 4300 primer pairs. AT/AT is not usually used to develop markers because of its self-complementary nature, which can lead to palindrome formation (Wang et al., 2011). A total of 40 primer pairs were randomly selected for PCR amplification and PAGE analysis (Table 5).

Among the 40 primer pairs, 31 were successfully amplified by PCR and PAGE (Figure 2). The remaining 9 primers failed to generate PCR products at various annealing temperatures and Mg^{2+} concentrations. Among the 31 successfully amplified primer pairs, 29 met the expected size and 2 were smaller products than expected.

Functional annotation for SSR-containing unigene sequences

All 9154 SSR-containing unigenes were functionally annotated by searches against the Nr, Nt, and Swiss-prot databases. All SSR-containing unigenes were searched using BLASTx against the Nr database, using a cut-off e-value of 10^{-5}, and 6485 unigenes (70.84% of all SSR-containing unigenes) were annotated within the database. These unigenes were also aligned by BLASTx to the protein database Swiss-prot, and BLASTn to the Nt database, using a cut-off e-value of 10^{-5}. From the Swiss-prot and Nt databases, a total of 4571 (49.93%) and 6870 (75.05%) unigenes were annotated, respectively.

These unigenes were classified further using GO terms and COG categories. GO has three major categories (ontologies): molecular function, cellular component, and biological process. The basic unit of GO is the GO-term. Every GO-term belongs to a type of ontology. With Nr annotation, we used Blast2GO to generate GO annotations of the unigenes. After acquiring a GO annotation for every unigene, we used WEGO to perform a GO functional classification for all SSR-containing unigenes. Based on sequence homology, 5339 (58.32%) unigenes were assigned to 35,934 GO-term annotations. These 35,934 GO-terms were summarized into the three major categories. Of them, biological process (18,679, 51.98%) comprised the majority of the GO annotations, followed by cellular component (9151, 25.47%) and molecular function (8104, 22.55%).

COG is a database where orthologous gene products are classified. Every protein in COG was assumed to have evolved from an ancestral protein, and the whole database was built on coding proteins with complete genomes as well as system evolution relationships of bacteria, algae, and eukaryotic organisms. In our study, unigenes were aligned to the COG database to predict and classify their possible functions. In total, 2770 (30.26%) unigenes were assigned to the COG classification, with 3924 COG functional annotations (Table 6). Among 25 COG categories, the cluster for ‘General function prediction only’ (759, 27.40%) represented the largest group, followed by ‘Transcription’ (513, 18.52%), ‘Replication, recombination and repair’ (338, 12.20%), and ‘Signal transduction mechanisms’ (328, 11.84%). The following clusters: ‘Cell motility’, ‘Extracellular structures’, and ‘Nuclear structure’ (no unigenes were assigned to this group) were the smallest groups.

Table 5. Forty validated primer pairs used in the study.

Code	Primer	SSR motif	Number of repeats	Primer sequence (5'→ 3')	Except size/bp	Product
1	NAUPP_0014	GAT	7	GTTCTCTTCTGTTTGAGAAGTGCT	113	ES
				CTCACTCAGGATCTCAGTCAGGT		
2	NAUPP_0034	AC	11	ACACTGTTTTGGTGTTATGCTTG	132	NA
				CTCCAAAGCTCTCCTCTTTCTTT		
3	NAUPP_0041	AG	9	GTTTTCGAGAACAAGACGAAGAC	143	ES
				TGCCCCTTCTTAATTCCTTCTTA		
4	NAUPP_0051	GGT	6	GTCCAAACAGGAAGATGTGAAAC	154	ES
				ATCTGAAACTCCATCACCCCTAT		
5	NAUPP_0066	GGA	6	ATCGAAATTAGGTTTGTGGGTTT	158	NA
				CATCTTCTTCCCTATCCCCATAC		
6	NAUPP_0181	TGAGGA	6	GTAGTAGTAAGTGGGCTGGGAGG	143	ES
				CTTCTCCGCCTTCATCTCC		
7	NAUPP_0211	GAGC	6	TACATGGACATTCTGAGAGACCC	130	NA
				AAATCAAAACCCCGAATTAAGAG		
8	NAUPP_0298	GAC	6	ATGACAGCGTAGACGAAGAGGA	136	ES
				GACTGTACGACCTCCTGAACATC		
9	NAUPP_0490	CGC	7	TATGGTACTTTTTCTTCCTCGGC	146	NA
				TTGCTCATATCTTCCTGTTGGTT		
10	NAUPP_0613	GAG	5	TTAGGACAGGACTAACTGGAGCA	151	ES
				CAGAATTCCCATAAATCTGCATC		
11	NAUPP_0905	AGTAC	5	ACTCGATGAAATGGAAGTTCGT	134	ES
				ACGCGCAATTATTGACCTTATAC		
12	NAUPP_1129	GGAAA	4	TTTGTTTTTGGATGGAAAGATTG	81	ES
				CTTTCAAGAATCCCATTCCTCTC		
13	NAUPP_1148	TA	7	AATTAATAAGCCCAGCCCTACAT	131	ES
				CCCACCTCTCAATGGTAAGCTA		
14	NAUPP_1298	AG	7	CCTTATAAAGTTGGTGATGTGGG	130	ES
				CGAGGAACTCCATCCTCTACATA		
15	NAUPP_1552	CTT	7	CGAGAGGTCATAGTCATCCATGT	150	ES
				GATTCAGTCGATTGTTGAGGAAG		
16	NAUPP_1639	CAAA	5	AGAGAGTAGAAAACCGAAGCCTG	117	ES
				TCGGTAACTTTGCAATCTTTCAT		
17	NAUPP_1659	T	14	AACCCACTAAGCAAAAGGTGAAT	149	ES
				GGACTCAAATTCCCATCATAACA		
18	NAUPP_1731	CTT	5	GCCCAAAAGACCTCCTTCTC	137	ES
				CTGAACAGCACCAATGTTTTTG		
19	NAUPP_1846	TTTTC	4	AAACTGATGTCAGGGCTCAAGTA	128	ES
				TGAAACCAACATGTCACACCTTA		
20	NAUPP_2033	AT	8	CATCTCCATTTCCTGTGATTTTC	133	ES
				TTTGAGGGTGATGATGTAAAAGC		
21	NAUPP_2098	CT	6	AGTTTTTGGATTTGCTGGTTGT	150	ES
				CAACAAATCCCTACTTTGAGAAATC		
22	NAUPP_2184	AAGCAG	4	AGAAGGAAAGATGGGGTTTGAT	138	NA
				TCCTCTTTCTTGAGTTTGTCAGC		
23	NAUPP_2292	CT	8	GAGTCCCAGACTGTCTCTCTGTC	117	ES
				ATGATGGGATCGTTATCAGTGTC		
24	NAUPP_2390	AG	6	GTTTTCTCACCATAACTGCCTGT	155	ES
				TTGCATTTTAATTTTGGGTTTTG		
25	NAUPP_2446	TC	8	ATCATTCATAAGAAGGACCCCAT	157	NA
				AGAAGTGGCAATATGGAATGTCT		
26	NAUPP_2569	AG	7	TGATTTCAGCATAATAAATGCCC	157	ES
				ACAGAGAGCCAAAAGTAGCACAG		
27	NAUPP_2936	T	15	AGATCTCCAACAAGAAAGAACCC	141	ES
				CCGGATTGTTCATCCAATATAAG		
28	NAUPP_2974	CTA	5	TACCCGTTTTCTGAGTAACCAAA	114	ES
				GACGAAAATCGAACCTAAGACCT		
29	NAUPP_3021	AC	7	CGAGTAATAAAGACGTCCAGGG	160	ES
				GCACTCCAAAAATCTAATTGTGC		

Continued on next page

Table 5. Continued.

Code	Primer	SSR motif	Number of repeats	Primer sequence (5'→ 3')	Except size/bp	Product
30	NAUPP_3106	CT	7	ACTCGCAAAAATACACAAAAAGC	133	NA
				CAACTCTCTCACTCTCTCCCCTT		
31	NAUPP_3153	AG	9	AAGAGCCATTGAGGAGCAATAG	139	ES
				TTCTCTTCTATTTCTTTCGCCG		
32	NAUPP_3207	CTC	5	CTTCCTAACTCCGACCAAAACTC	116	NA
				CCTCAGAAGAAGCATCAACAAAC		
33	NAUPP_3297	AGC	5	ATATGAACCGTAACGACGATCC	146	-
				GAACAGCTGCTCATACTTCATCC		
34	NAUPP_3375	TA	9	GAAGTGCGTGTAAAATCACACAA	150	ES
				CAAACCAACCCTAGCTACTCTCC		
35	NAUPP_3484	CAC	6	CAAACACCAATTTCCTCAAAAAC	160	-
				GATCGATCTTGGATGTGGGT		
36	NAUPP_3502	AG	7	TTGGTTTCTGAGACTTGCTTTGT	120	ES
				ACTTCTCCCTCCGTTGTTTAGTC		
37	NAUPP_3580	CT	6	TTTAATTACAGCACTCGCCTCTC	124	ES
				GAAAATAACATGGGTTTTCTGGG		
38	NAUPP_3638	AG	6	GTACTTGTTGATGCTTGTTGCAG	138	NA
				ATCATAGCAGTAATTCTGGCAGC		
39	NAUPP_3815	G	19	AAGAGTGGGGGTTTGTATTGG	147	ES
				ACCCCGAATTAAGAGTTTTAGCA		
40	NAUPP_3982	CT	6	ACACAAGGTTATTGATATGCGCT	152	ES
				CTCGACATCGATACAATTCAACA		

ES = met expected size; NA = no amplification product; - = smaller than expected size.

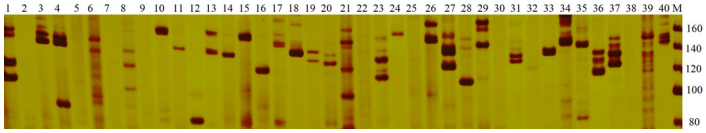

Figure 2. PAGE results from the PCR amplification of 40 primer pairs. Primer codes are the same as those listed in Table 5; *Lane M*: 20-bp DNA Ladder Marker (TaKaRa).

Table 6. COG classification of SSR-containing unigenes.

Code	Functional categories of COG	Gene number
A	RNA processing and modification	52
B	Chromatin structure and dynamics	54
C	Energy production and conversion	103
D	Cell cycle control, cell division, chromosome partitioning	152
E	Amino acid transport and metabolism	152
F	Nucleotide transport and metabolism	45
G	Carbohydrate transport and metabolism	228
H	Coenzyme transport and metabolism	59
I	Lipid transport and metabolism	110
J	Translation, ribosomal structure, and biogenesis	186
K	Transcription	513
L	Replication, recombination, and repair	338
M	Cell wall/membrane/envelope biogenesis	112
N	Cell motility	9
O	Posttranslational modification, protein turnover, chaperones	250
P	Inorganic ion transport and metabolism	111
Q	Secondary metabolites biosynthesis, transport, and catabolism	87
R	General function prediction only	759
S	Function unknown	142
T	Signal transduction mechanisms	328
U	Intracellular trafficking, secretion, and vesicular transport	66
V	Defense mechanisms	31
W	Extracellular structures	2
Y	Nuclear structure	0
Z	Cytoskeleton	35

DISCUSSION

Development and characterization of EST-SSRs in *P. pyrifolia*

A total of 10,622 SSRs were obtained from 75,764 unigenes, with a frequency distribution of 1/5.24 kb. The frequency distribution in the *P. pyrifolia* Nakai 'Huanghua' transcriptome was lower than that of *P. bretschneideri* Rehd. 'Dangshansu' (1/2.78kb) (Zhang et al., 2014). This finding could be due to differences between these two types of pear transcriptome, which may also be greatly affected by SSR-detecting criteria. In SSR development from the transcriptome, the first step was to define SSRs as a certain criterion, which decided the parameters of the SSR search algorithm, thereby directly affecting the final results. An exact comparison of SSR features between different species was complicated due to multiple factors. The frequency of pear EST-SSRs was higher than that of citrus (1/5.7 kb) (Chen et al., 2006), soybean (1/23.8 kb) (Gao et al., 2003), and barley (*Hordeum vulgare* L., 1/6.3 kb) (Thiel et al., 2003), but lower than kiwifruit (*Actinidia* spp, 1/2.48 kb) (Jiang et al., 2009) and tea (*Camellia sinensis*, 1/3.68 kb) (Yang et al., 2011). These differences could be the real cause of SSR frequency between different species, or it may be due to differences in the database used and SSR-detecting criteria. Overall, the SSRs of *P. pyrifolia* Nakai were plentiful.

Repeat motifs among these SSRs consisted of 159 types. A/T motifs occurred much more frequently than C/G motifs, which is in agreement with studies of other plants (Morgante et al., 2002; Qiu et al., 2010; Guan et al., 2013). AG/CT was the most abundant motif and showed a striking dominance of all SSRs. The same results were reported in strawberry (Bombarely et al., 2010), lotus (Pan et al., 2010), and cassava (Raji et al., 2009). This is probably because the AG/CT dinucleotide motif can represent multiple codons resting within the reading frame that can be translated into different amino acids, and AG/CT could be present in codons for alanine and leucine, which have the highest frequency in proteins. The frequency of the GC motif (N = 11) was rare in our study, and similar results have been reported in previous studies. There were no GC motifs in *Arabidopsis thaliana*, apricot, peach, rice, maize, wheat, and soybean (Gao et al., 2003; Nicot et al., 2004; Jung et al., 2005), otherwise only two and one GC motifs were identified in coffee and kiwifruit, respectively (Aggarwal et al., 2007; Jiang et al., 2009). These reports have shown that plants may have an obvious GC bias, and the reasons for this phenomenon require further study. Among the trinucleotide motifs, AAG/CTT (662, 6.23%) occurred most frequently, agreeing with results in coffee (Aggarwal et al., 2007), *P. bretschneideri* Rehd. (Zhang et al., 2014), *Jatropha curcas* L. (Yadav et al., 2011), and many other dicotyledonous plants (Kumpatla and Mukhopadhyay, 2005; Raji et al., 2009; Pan et al., 2010).

Co-specific transferability and amplification rate

Transcriptome sequencing has provided the unprecedented opportunity to mine EST-SSRs by generating massive amounts of available sequence data. Compared with DNA-derived SSRs, EST-SSRs have several obvious advantages. High co-specific transferability may be the most important, which has been demonstrated. Due to the source of EST-SSRs (directly derived from transcripts), they are more conservative and have greater transferability between species than genomic SSRs, which has been demonstrated in previous studies (Decroocq et al., 2003; Castillo et al., 2008; Shirasawa et al., 2011). These EST-SSRs, which have been mined from the pear transcriptome and are reported in our paper, surely can be used in other species

of Rosaceae, especially in those with limited molecular markers.

In total, 4300 primer pairs were obtained (**Table S1**), and 40 primer pairs were randomly selected for PCR amplification and PAGE analysis. Of these primers, 31 primer pairs (77.50%) were successfully amplified. The success rate of amplification for these EST-SSRs was relatively high. The remaining 9 primers could not be amplified, because they might extend across a splice site or be derived from a chimeric cDNA sequence. Another reason might be that these primers were designed from questionable sequences based on transcriptome sequencing and would have inevitably failed when authenticated with genomic DNA. Among the 31 successfully amplified primer pairs, 29 were the expected size and 2 were smaller products than expected. Compared with DNA-derived SSRs, the amplification size of EST-SSRs tends to be more frequently derived from the expected size, which has been demonstrated in previous studies (Cordeiro et al., 2001; Thiel et al., 2003; Yu et al., 2004). This result might be due to introns or insertions-deletions within the amplicons. Moreover, the possibility of assembly errors, which occurred during transcriptome sequencing, cannot be excluded. In addition, the high success rate of amplification also confirmed that transcriptome sequencing is a fast, effective, and reliable approach to develop an abundance of available EST-SSRs.

Level of potential polymorphism

High polymorphism is one of the most important features of an effective SSR. Generally, perfect SSRs have higher polymorphisms than imperfect or compound SSRs (Weber, 1990). SSR polymorphisms are also affected by location within DNA. SSRs harbored within UTRs (untranslated regions) are more polymorphic than those occurring within exon regions (Qiu et al., 2010). This phenomenon might be caused by selection and evolution. SSRs harbored within UTRs would be easier to change than those within exon regions. Of the six types of SSR motifs, mononucleotide, dinucleotide, tetranucleotide, and pentanucleotide motifs mainly occurred within UTRs and had greater potential of being polymorphic. In contrast, trinucleotide and hexanucleotide repeat loci mainly occurred within exon regions and had lower potential of being polymorphic.

The length of SSRs is also an important factor regarding polymorphisms. Longer SSRs tend to be highly polymorphic and have a higher potential to serve as effective genetic markers, which have been shown in many organisms (Weber, 1990; Cho et al., 2000). According to the length of SSRs, Temnykh et al. (2001) categorized SSRs into two groups: Class I microsatellites, containing perfect SSRs $\geq$ 20 nucleotides in length, and Class II microsatellites, containing perfect SSRs > 12 and < 20 nucleotides in length. Class I SSRs are deemed as having a high chance of showing a polymorphism.

In our report, a total of 10,622 SSRs were identified, of which 2293 (21.59%) were $\geq$20 nucleotides in length. These 2293 SSRs included mononucleotide (282), dinucleotide (1066), tetranucleotide (146), and pentanucleotide (216) motifs, which totaled 1710. These 1710 highly polymorphic SSRs, which were identified and designed in this study, will provide plenty of molecular markers for further genetic analyses of pears and other species in the Rosaceae.

Annotation of SSR-containing unigenes

Strong evidence supports the hypothesis that EST-SSRs play functional roles in organisms, some of which are regulatory (Li et al., 2004). In this study, functional annotations

of SSR-containing sequences were reported, and a relatively higher percentage of unigenes (70.84%) had BLAST hits in the Nr database, which was partially due to the high percentage frequency of long sequences in our assembled unigenes (nearly half of them were longer than 500 bp). As reported in some previous studies (Parchman et al., 2010; Wang et al., 2010b), longer assembled sequences have a higher probability of being annotated in public databases. EST-SSRs, which were mined from the transcriptome, have the potential to serve as functional markers, because they were directly derived from coding regions and some of them are linked with important candidate genes. Through these functional annotations, we can understand the distribution and function of SSR-containing unigenes at the macrolevel, and they are also very valuable for screening SSR markers, which are linked to functional genes.

In conclusion, our paper is the first to annotate pear genic SSR-containing sequences using public databases, GO terms, and COG categories. It is also one of the few studies that have developed EST-SSRs from the pear transcriptome. Pear EST-SSRs exhibited a relatively high frequency of SSR loci, which provide substantial resources for the development of EST-SSR markers. Considering self-incompatibility, high heterozygosity, and poorly differentiating characters of pears, these EST-SSRs should be particularly valuable for use in genomic mapping, breeding selection, phylogenetic relationship analysis, DNA fingerprinting, and other genetic studies. They can also be used in other relevant species because of their high co-specific transferability.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#31272140).

REFERENCES

Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, 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: 359-372. http://dx.doi.org/10.1007/s00122-006-0440-x.

Bombarely A, Merchante C, Csukasi F, Cruz-Rus E, et al. (2010). Generation and analysis of ESTs from strawberry (*Fragaria xananassa*) fruits and evaluation of their utility in genetic and molecular studies. *BMC Genomics* 11: 503. http://dx.doi.org/10.1186/1471-2164-11-503.

Castillo A, Budak H, Varshney RK, Dorado G, et al. (2008). Transferability and polymorphism of barley EST-SSR markers used for phylogenetic analysis in *Hordeum chilense. BMC Plant Biol.* 8: 97 http://dx.doi.org/10.1186/1471-2229-8-97.

Chen C, Zhou P, Choi YA, Huang S, et al. (2006). Mining and characterizing microsatellites from citrus ESTs. *Theor. Appl. Genet.* 112: 1248-1257 http://dx.doi.org/10.1007/s00122-006-0226-1.

Cho YG, Ishii T, Temnykh S, Chen X, et al. (2000). Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 100: 713-722 http://dx.doi.org/10.1007/s001220051343.

Conesa A, Götz S, García-Gómez JM, Terol J, et al. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676. http://dx.doi.org/10.1093/bioinformatics/bti610.

Cordeiro GM, Casu R, McIntyre CL, Manners JM, et al. (2001). Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci.* 160: 1115-1123. http://dx.doi.org/10.1016/S0168-9452(01)00365-X.

Decroocq V, Fave MG, Hagen L, Bordenave L, et al. (2003). Development and transferability of apricot and grape EST microsatellite markers across taxa. *Theor. Appl. Genet.* 106: 912-922.

Fan L, Zhang MY, Liu QZ, Li LT, et al. (2013). Transferability of newly developed pear SSR markers to other Rosaceae species. *Plant Mol. Biol. Report.* 31: 1271-1282. http://dx.doi.org/10.1007/s11105-013-0586-z.

Gao L, Tang J, Li H and Jia J (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol. Breed.* 12: 245-261. http://dx.doi.org/10.1023/A:1026346121217

Guan L, Huang JF, Feng GQ, Wang XW, et al. (2013). Survey of simple sequence repeats in woodland strawberry (*Fragaria vesca*). *Genet. Mol. Res.* 12: 2637-2651. http://dx.doi.org/10.4238/2013.July.30.3.

Jiang CY, Xu XB, Liao J, Ni ZH, et al. (2009). Analysis of SSR information in EST resources of kiwifruit (*Actinidia* ssp.). *Zhongguo Nongxue Tongbao* 13: 37-39.

Jung S, Abbott A, Jesudurai C, Tomkins J, et al. (2005). Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. *Funct. Integr. Genomics* 5: 136-143. http://dx.doi.org/10.1007/s10142-005-0139-0.

Kumpatla SP and Mukhopadhyay S (2005). Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48: 985-998. http://dx.doi.org/10.1139/g05-060.

Li YC, Korol AB, Fahima T and Nevo E (2004). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 21: 991-1007. http://dx.doi.org/10.1093/molbev/msh073

Morgante M, Hanafey M and Powell W (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30: 194-200. http://dx.doi.org/10.1038/ng822.

Nicot N, Chiquet V, Gandon B, Amilhat L, et al. (2004). Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). *Theor. Appl. Genet.* 109: 800-805. http://dx.doi.org/10.1007/s00122-004-1685-x

Pan H, Yang CP, Wei ZG and Jiang J (2006). DNA extraction of birch leaves by improved CTAB method and optimization of its ISSR system. *J. For. Res.* 17: 298-300. http://dx.doi.org/10.1007/s11676-006-0068-3.

Pan L, Xia Q, Quan Z, Liu H, et al. (2010). Development of novel EST-SSRs from sacred lotus (*Nelumbo nucifera* Gaertn) and their utilization for the genetic diversity analysis of *N. nucifera*. *J. Hered.* 101: 71-82. http://dx.doi.org/10.1093/jhered/esp070

Parchman TL, Geist KS, Grahnen JA, Benkman CW, et al. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and discovery. *BMC Genomics* 11: 180. http://dx.doi.org/10.1186/1471-2164-11-180.

Potter D, Eriksson T, Evans RC, Oh S, et al. (2007). Phylogeny and classification of Rosaceae. *Plant Syst. Evol.* 266: 5-43. http://dx.doi.org/10.1007/s00606-007-0539-9.

Qiu L, Yang C, Tian B, Yang JB, et al. (2010). Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). *BMC Plant Biol.* 10: 278. http://dx.doi.org/10.1186/1471-2229-10-278.

Raji AA, Anderson JV, Kolade OA, Ugwu CD, et al. (2009). Gene-based microsatellites for cassava (*Manihot esculenta* Crantz): prevalence, polymorphisms, and cross-taxa utility. *BMC Plant Biol.* 9: 118. http://dx.doi.org/10.1186/1471-2229-9-118.

Shen ZJ, Ma RJ, Cai ZX, Yu ML, et al. (2015). Diversity, population structure, and evolution of local peach cultivars in China identified by simple sequence repeats. *Genet. Mol. Res.* 14: 101-117. http://dx.doi.org/10.4238/2015.January.15.13.

Shirasawa K, Oyama M, Hirakawa H, Sato S, et al. (2011). An EST-SSR linkage map of *Raphanus sativus* and comparative genomics of the Brassicaceae. *DNA Res.* 18: 221-232. http://dx.doi.org/10.1093/dnares/dsr013.

Tautz D and Renz M (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res.* 12: 4127-4138. http://dx.doi.org/10.1093/nar/12.10.4127.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, et al. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* 11: 1441-1452. http://dx.doi.org/10.1101/gr.184001.

Thiel T, Michalek W, Varshney R and Graner A (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106: 411-422.

Wang XC, Qiao YS, Wang F, Ma L, et al. (2010a). Extraction of total RNA from pericarp of pears with three methods. *J. Gansu Agric. Univ.* 45: 91-94.

Wang XW, Luan JB, Li JM, Bao YY, et al. (2010b). *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* 11: 400. http://dx.doi.org/10.1186/1471-2164-11-400.

Wang YW, Samuels TD and Wu YQ (2011). Development of 1,030 genomic SSR markers in switchgrass. *Theor. Appl. Genet.* 122: 677-686. http://dx.doi.org/10.1007/s00122-010-1477-4.

Weber JL (1990). Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. *Genomics* 7: 524-530. http://dx.doi.org/10.1016/0888-7543(90)90195-Z

Wu J, Wang Z, Shi Z, Zhang S, et al. (2013). The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* 23: 396-408. http://dx.doi.org/10.1101/gr.144311.112.

Yadav HK, Ranjan A, Asif MH, Mantri S, et al. (2011). EST-derived SSR markers in Jatropha curcas L.: development, characterization, polymorphism, and transferability across the species/genera. *Tree Genet. Genomes* 7: 207-219. http://dx.doi.org/10.1007/s11295-010-0326-6

Yang H, Chen Q, Wei CL, Shi CY, et al. (2011). Analysis on SSR information in *Camellia sinensis* transcriptome. *J. Anhui Agric. Univ.* 38: 882-886.

Ye J, Fang L, Zheng HK, Zhang Y, et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34: 293-297 http://dx.doi.org/10.1093/nar/gkl031.

Yu JK, Dake TM, Singh S, Benscher D, et al. (2004). Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. *Genome* 47: 805-818. http://dx.doi.org/10.1139/g04-057

Yue XY, Liu GQ, Zong Y, Teng YW, et al. (2014). Development of genic SSR markers from transcriptome sequencing of pear buds. *J. Zhejiang Univ. Sci.* 15: 303-312. http://dx.doi.org/10.1631/jzus.B1300240.

Zhang MY, Fan L, Liu QZ, Song Y, et al. (2014). A novel set of EST-derived SSR markers for pear and cross-species transferability in Rosaceae. *Plant Mol. Biol. Report.* 32: 290-302. http://dx.doi.org/10.1007/s11105-013-0638-4.

Supplementary material

Table S1. Primer results after filtering (4300 pairs of primers).