GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Development and characterization of genic-SSR markers from different Asia lotus (*Nelumbo nucifera*) types by RNA-seq

X.F. Zheng[1]*, Y.N. You[1]*, Y. Diao[1,2], X.W. Zheng[3], K.Q. Xie[3], M.Q. Zhou[4], Z.L. Hu[1,4] and Y.W. Wang[5]

[1]State Key Laboratory of Hybrid Rice, College of Life Sciences,
Wuhan University, Wuhan, Hubei, China
[2]College of Forestry and Life Sciences,
Chongqing University of Arts and Sciences, Yongchuan, Chongqing, China
[3]Guangchang White Lotus Research Institute, Fuzhou, Jiangxi, China
[4]Lotus Center, Wuhan University, Wuhan, Hubei, China
[5]College of Pharmaceutical Sciences, Wuhan University, Hubei, China

*Theses authors contributed equally to this study.
Corresponding authors: Z.L. Hu
E-mail: huzhongli@whu.edu.cn

Genet. Mol. Res. 14 (3): 11171-11184 (2015)
Received January 6, 2015
Accepted May 25, 2015
Published September 22, 2015
DOI http://dx.doi.org/10.4238/2015.September.22.11

ABSTRACT. *Nelumbo nucifera* is an important economic vegetable and traditional medicine, but available genetic resources remain limited. Next generation sequencing has proven to be a rapid and effective means of identifying genic simple sequence repeat (genic-SSR) markers. This study developed genic-SSRs for *N. nucifera* using Illumina sequencing technology to assess diversity across cultivated and wild lotus. A total of 105,834 uni-contigs were produced with an average read length of 722 bp. Exactly 11,178 genic-SSR loci were identified in 9523 uni-contigs. Di-nucleotide (64.5%) was the most abundant SSR, followed by tri-nucleotide (23%), tetra-nucleotide (8.9%), penta-nucleotide (2.5%), and hexa-nucleotide (1%) repeat types. The most common di- and tri-nucleotide repeat motifs were AG/CT (51%) and AAG/CTT (8%), respectively. Based on these SSRs sequences, 6568 primer pairs were designed, of which 72 primers were randomly selected for synthesis and validation, and 38 *in-silico* polymorphic primers were obtained using in-house perl scripts. A total of 110 primers were screened in the lotus samples and the results showed that 101 primers yielded amplification products, of which 80 were polymorphs. The number of alleles ranged from 2 to 17 and the PIC (polymorphism information content) ranged from 0.19 to 0.87 with a mean value of 0.55. An Unweighted Pair Group Method with Arithmetic Mean (UPGMA) dendrogram based on Jaccard's similarity coefficients showed that the correlation between geographical source and genotype was low. This study describes the distribution of genic-SSRs in the expressed portion of the lotus genome. These genic-SSRs have an important role to play in molecular mapping, diversity analysis, and marker-assisted selection strategies in *Nelumbo*.

Key words: Transcriptome; Genic-SSRs; *Nelumbo nucifera*;
Genetic diversity

INTRODUCTION

The genus *Nelumbo* Adans. only contains *Nelumbo nucifera* and *Nelumbo lutea*, and has an Asian and North American disjunct distribution pattern. *N. nucifera* is well known as an economically important ornamental (i.e., flower types) and dietary plant (i.e. rhizome and seed lotus types). In China, it is also used as an important traditional medicine. In the previous study, almost all parts of *N. nucifera*, including the flower, rhizome, and leaf, possessed anti-obesity, anti-inflammatory, anti-pyretic, anti-oxidant, hepatoprotective and free radical scavenging activities (Sohn et al., 2003).

Previous studies on *Nelumbo* mainly focused on pharmaceutical activities (Liu et al., 2010) and the population genetic diversity of inter simple sequence repeats (ISSR), allozyme, sequence related amplified polymorphism (SRAP), amplified fragment length polymorphism (AFLP), and simple sequence repeat (SSR) markers (Han et al., 2007; Tian et al., 2008b; Hu et al., 2012; Pan et al., 2011; Yang et al., 2012b). Lotus germplasm accessions are not only a useful natural resource, but also an important gene pool. A better evaluation of lotus germplasm genetic diversity is crucial if it is to be utilized in breeding and conservation. Fu et al. (2011) and Hu et al. (2012) reported genetic variation between *N. nucifera* and *N. lutea*. Genetic diversity between China and Thailand lotus plants were evaluated using ISSR by Chen et al. (2008) and Li et al. (2010), which improved understanding of lotus germplasm classification. Although the development of a number of SSRs was reported by Tian et al. (2008a), Kubo et al. (2009), Pan et al. (2010), and Xue et al. (2012), few SSR markers have been exploited, particularly in *Nelumbo* molecular quantity genetic studies and molecular assistance breeding.

SSR makers are randomly repeated 1-6 DNA motifs that are abundant in eukaryote genomes and can mutate rapidly through the loss or gain of repeat units. Thus, microsatellites showing extensive length polymorphism, high polymorphism information content, and co-dominance have been widely used for comparative mapping, DNA fingerprinting, and biodiversity studies (Luro et al., 2008). SSR markers are divided into two types: genomic-SSRs and genic-SSRs. Genic-SSRs are highly transferable compared to genomic-SSRs, are linked with particular function genes that contribute to phenotype formation, and they can have powerful utility for Marker Assisted Selection (MAS) (Varshney et al., 2005). Technological advances in large-scale RNA-seq should lead to cost-effective, fast, and reliable generation of ESTs (Expression Sequence Tags). The large ESTs are a powerful genetic resource that can be used to detect genic-SSR loci with the bio-information software. Therefore, we performed *Nelumbo* RNA-seq to develop large numbers of novel and efficient genic-SSR markers.

This paper reports on the generation of a large expressed sequence dataset based on Illumina HiSeq^{TM} 2000 sequencing technology. The frequency and distribution of genic-SSRs from the rhizome and seed transcriptomes were analyzed. A total of 110 primer pairs were selected and synthesized to validate their amplification effect and the relationships among cultivars and wild lotus plants were assessed.

MATERIAL AND METHODS

Plant materials

A total of 51 lotus individuals were analyzed in the genetic diversity experiment (Table 1). The group contained 20 cultivated accessions, 29 wild lotus types, and two distance hybrid progeny. The lotus leaves were collected and underwent DNA extraction using the cetrimonium bromide method (CTAB) (Doyle 1987) with a slight modification by adding polyvinylpyrrolidone (PVP) into CTAB extraction buffer (final concentration: 1%). Genomic DNA quality was confirmed by 0.8% agarose gel electrophoresis. Apical buds from wild flower lotus (WFL) plants and cultivated rhizome lotus (CRL) plants were harvested, immersed in liquid nitrogen, and stored at -80°C before RNA extraction.

Construction and solexa sequencing of *Nelumbo* cDNA libraries

Total RNA was isolated from WFL and CRL apical buds using TRIzol (Invitrogen, Carlsbad, CA, USA) following the manufacturer protocols and the total mRNA was then purified using a Micropoly (A) Purist^{TM} mRNA purification kit (Ambion, USA). The cDNAs were synthesized following the modified method reported by Ng et al. (2005) and then shattered into fragments that were 300 to 500 bp long. After purifying them using Ampure beads (Agencourt, Beverly, MA, USA), cDNA libraries were constructed with TruSeq^{TM} DNA Sample Prep Kit Set A (Illumina, USA). After amplification using a TruSeq PE Cluster Kit (Illumina), the products were sequenced on an Illumina sequencing platform (Illumina Inc. San Diego, CA, USA) using a 100-bp paired-end approach.

Genic-SSR identification, primer design, and PCR amplification

A perl program called MISA (Thiel et al., 2003) was used to identify genic-SSR loci from the WFL, CRL, and their combined sequence databases. The search parameters were di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with minimum repeat numbers of 6, 5, 4, 4, and 4, respectively.

Table 1. Information about the lotus samples used in this study.

No.	Name	Species or origin	Sample	Source	Type
1	Thailand lotus-I	*Nelumbo nucifera*	Seed lotus	Thailand	Wild
2	Thailand lotus-II	*N. nucifera*	Seed lotus	Thailand	Wild
3	Thailand lotus-III	*N. nucifera*	Seed lotus	Thailand	Wild
4	Thailand lotus-IV	*N .nucifera*	Seed lotus	Thailand	Wild
5	Thailand lotus-V	*N. nucifera*	Seed lotus	Thailand	Wild
6	Thailand lotus-VI	*N. nucifera*	Seed lotus	Thailand	Wild
7	Cunsan lotus	*N. nucifera*	Seed lotus	Xiangtan, Hunan	Local Variety
8	Furong lotus	*N. nucifera*	Seed lotus	Xiangtan, Hunan	Local Variety
9	Baixiang lotus	*N. nucifera*	Seed lotus	Xiangtan, Hunan	Local Variety
10	Hongxiagn lotus	*N. nucifera*	Seed lotus	Xiangtan, Hunan	Local Variety
11	Yueyang wild lotus	*N. nucifera*	Seed lotus	Yueyang, Hunan	Wild
12	Xuehugong lotus	*N. nucifera*	Rhizome lotus	Anqing, Anhui	Wild
13	Nangeng wild lotus	*N. nucifera*	Rhizome lotus	Ma’anshan, Anhui	Wild
14	Piaohua lotus	*N. nucifera*	Rhizome lotus	Anhui	Wild
15	Lianhu wild lotus	*N. nucifera*	Rhizome lotus	Anhui	Wild
16	Madang lotus	*N. nucifera*	Rhizome lotus	Anhui	Wild
17	Shanmiao wild lotus	*N. nucifera*	Rhizome lotus	Anhui	Wild
18	Qingtang wild lotus	*N. nucifera*	Rhizome lotus	Anhui	Wild
19	Baihu wild lotus	*N. nucifera*	Rhizome lotus	Anhui	Wild
20	Chuzhoubai lotus	*N. nucifera*	Seed lotus	Jinhua, Zhejiang	Local Variety
21	Tuxuan lotus	*N. nucifera*	Seed lotus	Jinhua, Zhejiang	Local Variety
22	Baihuajian lotus	*N. nucifera*	Seed lotus	Jianning, Fujian	Local Variety
23	Honghuajian lotus	*N. nucifera*	Seed lotus	Jianning, Fujian	Local Variety
24	Heilongjiang wild lotus	*N. nucifera*	Seed lotus	Beijing	Wild
25	Dajinhu lotus	*N. nucifera*	Seed lotus	Beijing	Wild
26	Zhaoyuan wild lotus	*N. nucifera*	Seed lotus	Beijing	Wild
27	Guangchangbaiye lotus	*N. nucifera*	Seed lotus	Guangchang, Jiangxi	Local Variety
28	Jingguang-2 hao	*N. nucifera*	Seed lotus	Guangchang, Jiangxi	Local Variety
29	Xingkongmudan	*N. nucifera*	Seed lotus	Guangchang, Jiangxi	Local Variety
30	Gudai lotus	*N. nucifera*	Seed lotus	Wuhan, Hubei	Wild
31	Zhongnanhaigu lotus	*N. nucifera*	Seed lotus	Wuhan, Hubei	Wild
32	Puzheheibai lotus	*N. nucifera*	Seed lotus	Wuhan, Hubei	Wild
33	Wufei lotus	Hybrid	Flower louts	Beijing	Local Variety
34	Baiyangdianhong lotus	*N. nucifera*	Flower lotus	Nanjing, Jiangsu	Wild
35	Zhuanshang lotus	*N. nucifera*	Flower louts	Beijing	Local Variety
36	Jiandehonghua lotus	*N. nucifera*	Flower louts	Jiande, Zhejiang	Local Variety
37	Yixian lotus	Hybrid	Flower louts	Nanjing, Jiangsu	Local Variety
38	Qiushuichagntian	*N. nucifera*	Flower louts	Guangchang, Jiangxi	Local Variety
39	Chongtai lotus	*N. nucifera*	Flower louts	Wuhan, Hubei	Local Variety
40	Donggua lotus	*N. nucifera*	Flower louts	Changsha, Hunan	Local Variety
41	Taikong 36 hao	*N. nucifera*	Seed lotus	Guangchang, Jiangxi	Local Variety
42	American lotus-I	*N. lutea*	Flower louts	America	Wild
43	American lotus-II	*N. lutea*	Flower louts	America	Wild
44	American lotus-III	*N. lutea*	Flower louts	America	Wild
45	American lotus-IV	*N. lutea*	Flower louts	America	Wild
46	American lotus-V	*N. lutea*	Flower louts	America	Wild
47	American lotus-VI	*N. lutea*	Flower louts	America	Wild
48	Riza 3 hao-I	*N. nucifera*	Seed lotus	Wuhan, Hubei	Local Variety
49	Riza 3 hao-II	*N. nucifera*	Seed lotus	Wuhan, Hubei	Local Variety
50	WFL*	*N. nucifera*	Seed lotus	Diaocha Lake, Hubei	Wild
51	CRL*	*N.nucifera*	Rhizome lotus	Wuhan, Hubei	Local Variety

*Sample was used for Illumina transcriptome sequencing.

The genic-SSR primers were designed using Primer3.0 (Rozen and Skaletsky, 2000) and its default parameters, and the combined sequence database. We selected 72 primers to validate the SSR markers and screened 51 lotus samples. PCR amplification was performed using gradient PCR analysis with annealing temperatures between 54° and 65°C in a 15 µL reaction volume containing 1X buffer, 1.4 mM $MgCl_2$, 0.1 mM dNTPs, 10 pmol of each primer, 0.5 U Taq polymerase, and ~50 ng template DNA. The PCR products were detected as described by Han et al. (2008).

In-silico analysis of genic-SSR polymorphism in lotus flowers and rhizomes

MISA tools was used to locate the position of the SSR loci in the WFL, CRL, and their combined sequence databases. Perl scripts were then compiled and used to search for common and polymorphic genic-SSR markers between WFL and CRL (Figure 1). Firstly, genic-SSR loci from WFL and CRL were identified, and if the forward and rear 15-bp nucleotides beside the SSR locus from WFL and CRL were identical, these SSRs were considered as potential polymorphic loci; secondly, the *in-silico* polymorphic genic-SSRs were confirmed based on the different number of SSR locus repeats; and thirdly, the polymorphic primers were retrieved from a primers database designed using Primer3.0 (Rozen and Skaletsky, 2000) and the combined sequence database. Using this method, 1627 common genic-SSRs loci were found between WFL and CRL, of which only 48 loci were polymorphic and could be used to search the primer databases. Finally, 38 polymorphic primers were identified and synthesized for validation by PCR amplification and electrophonic detection as described above.

Data analysis

The raw reads were first filtered to remove low quality reads (<Q20) and all adaptor sequences, and then assembled with the Trinity software into contigs (Grabherr et al., 2011). Finally, three database sets for *N. nucifera* were constructed (Figure 1).

To estimate the allelic variation for genic-SSRs in the 51 samples, the PIC (Polymorphism Information Content) of each primer was calculated based on the following formula:

$$\mathrm{PIC} = 1 - \sum_{i=0}^{n} \mathrm{Pi}^2$$

where Pi is the frequency of the i^{th} allele and n is the total number of alleles amplified for a given genic-SSR marker. The coefficient of genetic similarity among all samples was calculated using NTSYS-pc Version 2.10 (Rohlf, 2000). A dendrogram was constructed based on the genetic similarity matrix using the UPGMA algorithm.

RESULTS

Assembly of raw reads by Illumina sequencing

A total of 50,578,940 and 49,452,590 filtered sequence reads were obtained from WFL and CRL, respectively (Figure 1). The total length of the reads was over 9 Gbp. The two sets of reads data were subsequently *de novo* assembled by the Trinity software (Grabherr et al., 2011) into 111,925 and 100,016 contigs, respectively, with lengths of over 200 bp, and then further assembled into 105,834 uni-contigs with an average length of 722 bp. The size distribution of these contigs is shown in Figure 2.

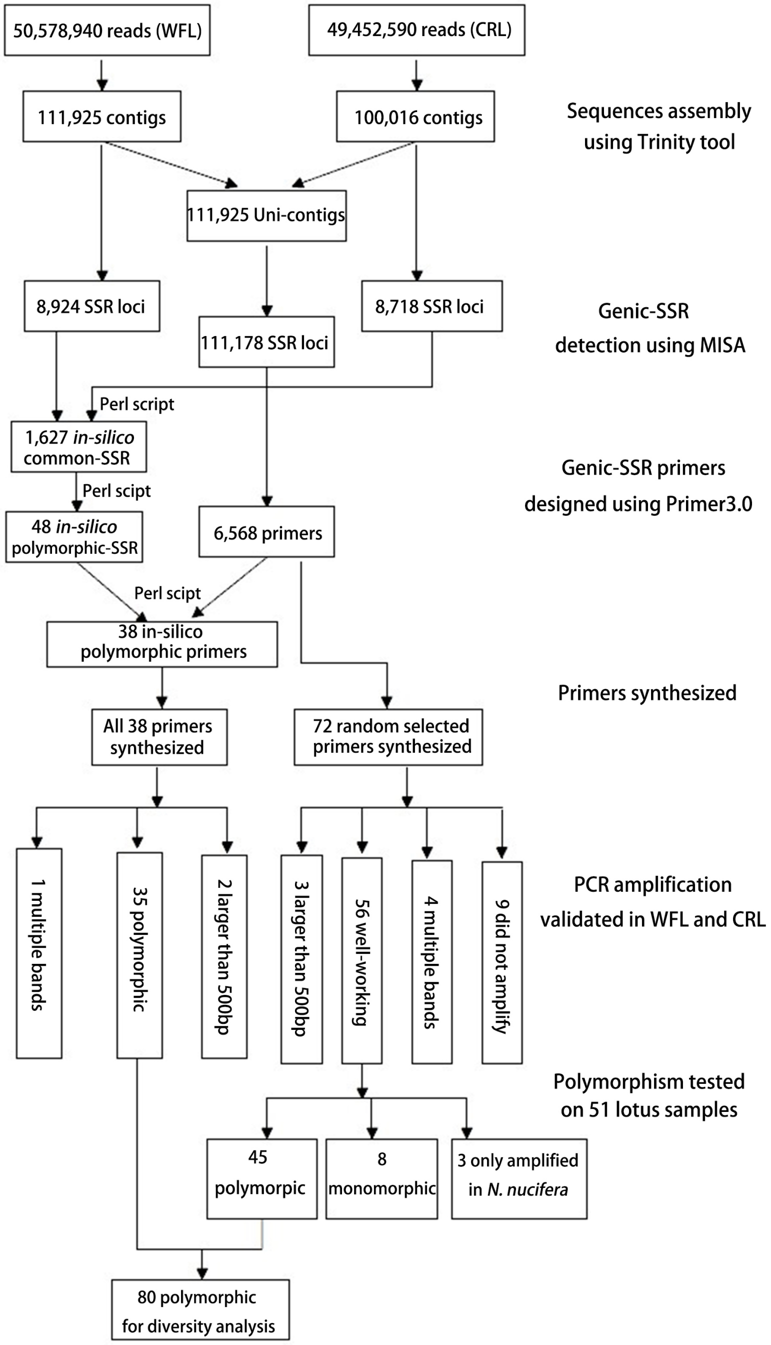
50,578,940 reads (WFL)
49,452,590 reads (CRL)
111,925 contigs
100,016 contigs
Sequences assembly using Trinity tool
111,925 Uni-contigs
8,924 SSR loci
111,178 SSR loci
8,718 SSR loci
Genic-SSR detection using MISA
Perl script
1,627 in-silico common-SSR
Perl scipt
48 in-silico polymorphic-SSR
6,568 primers
Genic-SSR primers designed using Primer3.0
Perl scipt
38 in-silico polymorphic primers
Primers synthesized
All 38 primers synthesized
72 random selected primers synthesized
1 multiple bands
35 polymorphic
2 larger than 500bp
3 larger than 500bp
56 well-working
4 multiple bands
9 did not amplify
PCR amplification validated in WFL and CRL
Polymorphism tested on 51 lotus samples
45 polymorpic
8 monomorphic
3 only amplified in N. nucifera
80 polymorphic for diversity analysis

Figure 1. Flow chart of genic-SSR development and diversity analysis from WFL, CRL, and Uni-contigs, respectively.

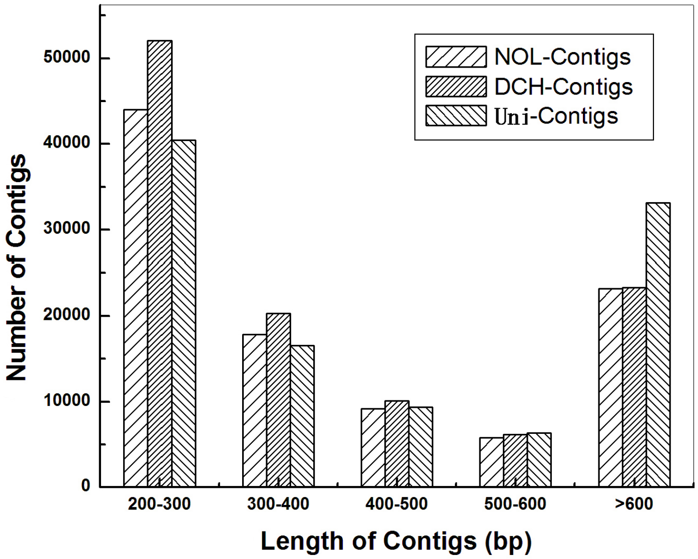

Figure 2. Size distribution of contigs from WFL, CRL, and combined sequences respectively.

Identification, frequency, and distribution of different types of genic-SSR loci

We used MISA tools to detect 11,178 genic-SSRs loci from 9523 of 105,834 uni-contigs, of which 1345 contained more than one SSR. The distribution density was one genic-SSR locus per 6.84 kb. Among the identified SSR repeats, di-nucleotide (7214, 64.5%) was the most abundant repeat unit, followed by tri- (2576, 23%), tetra- (998, 8.9%), penta- (279, 2.5%), and hexa- (111, 1%) nucleotides (Figure 3a). Clearly, there were large proportions of di-nucleotide and tri-nucleotide motifs, whereas the remainder made up less than 12.5% in total. The average frequency of genic-SSR occurrence was about 9%. The number of SSR repeats ranged from 4 to 31, and SSRs with six repeats (2704, 24.2%) were the most abundant (Figure 3b), followed by those with seven tandem repeats (1941, 17.4%), five tandem repeats (1696, 15.2%), and eight tandem repeats (1535, 13.7%). Single sample repeat loci containing 12 bp were the most common (1962, 17.6%) followed by those with 14 bp (2247, 20.1%), 18 bp (2018, 18.1%), 14 bp (1564, 14%), and 15 bp (1454, 13%) (Figure 3c). The longest SSR locus was 102 bp.

A total of 182 repeat motif types were identified, based on sequence complementary. The AG/CT repeat motif (5714, 51.1%) was the most common. The six next most abundant repeat motifs were (AAG/CTT)n, (AC/GT)n, (AT/AT)n, (ATC/ATG)n, (AAAT/ATTT)n, and (AGG/CCT)n, with frequencies of 8.2, 7.1, 6.2, 4, 3, and 2%, respectively. The most common 12 repeat motifs are shown in Table 2.

Development and validation of genic-SSRs markers

Primer3.0 (Rozen and Skaletsky, 2000) and the non-redundant uni-contig sequences, which contained 11,178 genic-SSR loci in total, were used to design 6568 primer pairs (**Table S1**), and 72 primer pairs were synthesized to validate their amplification effect (**Table S2**). Among these primer pairs, PCR successfully amplified 63 (87.5%) primer pairs using

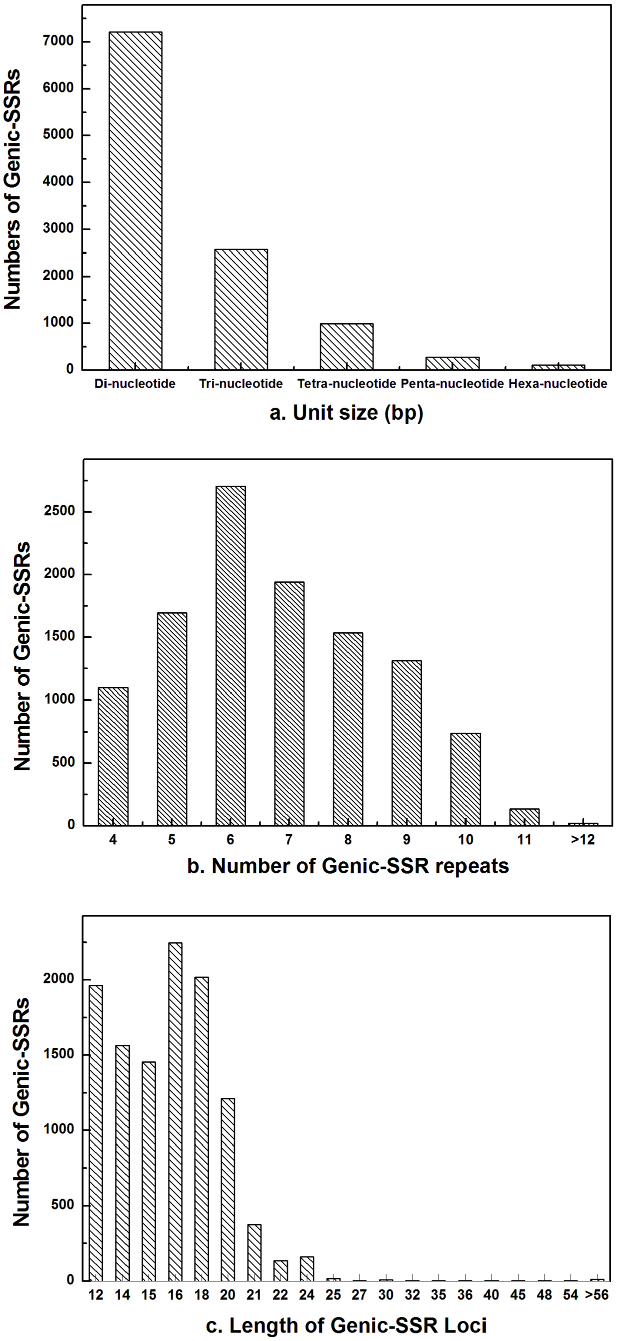

Figure 3. Frequency distribution of the genic-SSRs of differents sizes from the lotus uni-contigs. **a.** Unit size: genic-SSR motif length. **b.** Number of genic-SSR repeat: the number of genic-SSR repeat unit. **c.** Length of genic-SSR loci the length of genic-SSR repeat unit.

Table 2. Frequency distribution of the 12 most frequent genic-SSR repeat motifs in the lotus transcriptome and the number of repeats within each motif.

Order	Repeats	4	5	6	7	8	9	10	11	12	>13	Total	%
1	AG/CT	-	-	1417	1219	1215	1153	618	89	3	-	5714	51%
2	AAG/CTT	-	462	295	161	2	-	-	-	-	1	921	8%
3	AC/GT	-	-	309	187	137	75	60	25	2	-	795	7%
4	AT/AT	-	-	232	156	148	80	59	21	1	1	698	6%
5	ATC/ATG	-	261	127	60	3	-	-	-	-	1	452	4%
6	AAAT/ATTT	287	60	5	-	-	-	-	-	-	-	352	3%
7	AGG/CCT	-	137	64	25	4	-	-	-	-	-	230	2%
8	AAT/ATT	-	118	57	50	3	-	-	-	-	1	229	2%
9	ACC/GGT	-	144	48	29	4	1	-	-	-	-	226	2%
10	AGC/CTG	-	150	47	18	2	-	-	-	1	-	218	2%
11	AAAG/CTTT	135	55	7	-	-	-	-	-	-	-	197	2%
12	AAC/GTT	-	96	37	18	-	1	-	-	-	-	152	1%
13	Others	675	213	59	18	17	2	1	-	1	8	994	9%
	Total	1097	1696	2704	1941	1535	1312	738	135	8	12	11,178	100%

genomic DNA from WFL and CRL, of which four (5.6%) amplified many bands with non-specific products, and the remaining 9-pair primers (12.5%) produced no products, even when the annealing temperature was reduced by 7°C. A total of 56 functioning primer pairs, not including three primers that amplified PCR products more than 500 bp long, were screened in the 51 samples, of which 45 (80.4%) showed polymorphisms and three primers (5.4%) only amplified PCR bands in *N. nucifera*. The other eight primers (14.3%) did not show polymorphism. A total of 189 alleles were identified across 45 polymorphic genic-SSR loci, and the number of alleles ranged from 2 to 10 with an average of 4.2 alleles per locus. In addition, two primers amplified fragments that were larger than the expected sizes, probably because of the presence of an insertion mutation. Only one primer pair amplified a shorter than expected fragment, suggesting that a deletion mutation had occurred in the amplified region.

In order to evaluate and characterize these polymorphisms so that they can be potentially used for assessing molecular diversity or fingerprinting analysis, the PIC values of these polymorphism primers were calculated, based on the allelic variation shown in the 51 lotus accessions. The PIC values across 45 loci ranged from 0.19 for NL-60 to 0.87 for NL-7 with a mean value of 0.52. At the same time, the likely functions of these genic-SSR loci were deduced by BLAST analysis, and this showed that 16 transcriptomic sequences shared clear homology to other functional loci in plants (**Table S3**).

In-silico analysis of genic-SSRs polymorphism between WFL and CRL

In order to identify highly polymorphic genic-SSR markers between WFL and CRL, *in-silico* polymorphism analysis of genic-SSRs was undertaken using in-house perl scripts. Contig databases were obtained from WFL and CRL using Trinity tool, and positional information for the genic-SSRs was found using MISA. The in-house perl scripts showed that 1627 SSR loci were common between WFL and CRL, of which only 48 SSR loci were polymorphic. Furthermore, 38 of these *in-silico* polymorphic SSRs could be used to design primers. Therefore, these primers, named NL-P1 to NL-P38, were synthesized and validated in 51 lotus samples (**Table S2**). The results showed that the NL-P28 primer produced multiple bands, but no specific product, and the size amplification of two primers (NL-P11 and NL-P19) were larger than 500 bp in size. The results for the other 35 primers showed that all 35 primers could successfully amplify polymorphic bands. Finally, 198 alleles were produced from 35 loci, which was consistent with the *in-silico* analysis results. The number of alleles ranged from 2 to 17 with an average of 5.66 alleles per locus. Across these 35 loci, PIC values ranged from 0.19 for NL-P16 and NL-P27 to 0.87 for NL-P4 with a mean value of 0.57. The amplification details of these polymorphic primers were listed in **Table S3**.

Assessment of genetic diversity among cultivated and wild lotus accessions

The genetic diversity of different varieties and wild lotus were analyzed using 80 polymorphic primers. An UPGMA dendrogram based on Jaccard's similarity coefficients was constructed with three distinct clusters at a cut-off similarity index of 0.64 (Figure 4). The genetic similarity among the 51 lotus types ranged from 0.48 to 1.00. Cluster I consisted all *N. nucifera* accessions, which was divided into two sub-clusters with a similarity coefficient of 0.69. Sub-cluster Ia included eight seed lotus accessions, nine rhizome lotus accessions and one flower lotus accession. Sub-cluster Ib consisted of 19 seed lotus, five flower lotus and one rhizome lotus accession. Cluster II was composed of two hybrid lotuses that were produced by hybridization between *N. nucifera* and *N. lutea*. All six American lotus accessions that belonged to *N. lutea* were grouped into Cluster III.

In addition, the six lotus accessions from Thailand were divided into two sub-clusters containing lotus accessions from China and lotus accessions from different provinces in China were grouped together. Wild lotus and local cultivated varieties were also grouped into one cluster. This clustering was based on their appearance and utilization value. In the contrast, the geographic sources of the samples were not consistent with the genetic distances among lotus individuals. These results were consistent with previous studies (Pan et al., 2010, 2011; Yang et al., 2012b).

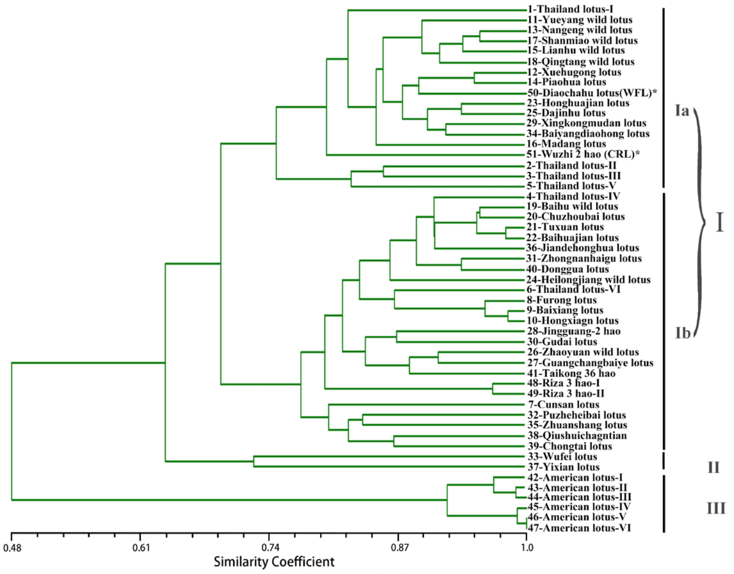

Figure 4. An UPGMA dendrogram of 51 lotus samples based on 80 polymorphic genic-SSR markers. Dendrogram showing similary relationship among 20 cultivated accessions, 29 wild lotus and 2 hybrid progeny, the number 1-51 corresponded to the samples listed in Table 1.

DISCUSSION

Frequency and distribution of different types of genic-SSRs loci

Genic-SSRs have considerable genetic analysis and linkage map construction potential in plants due to their high transferability and conservation. After Pan et al. (2010) reported the development of 23 EST-SSR markers using *Nelumbo* EST sequences in 2010, there have been no more reports of EST sequences being used for EST-SSRs development in the NCBI database. This study identified a total of 11,178 genic-SSR loci based on 105,834 non-redundant contigs. About 9% (9523 contigs) of the transcriptomic sequences possessed SSR loci. This rate was higher than for *Epimedium sagittatum* (3.67%)*,* which were identified using the same search parameters (Zeng et al., 2010). The distribution density in *N. nucifera* is one microsatellite loci per 6.84 kb. This *Nelumbo* genic-SSR frequency occurrence is relatively low compared to some plants, such as 3.4 kb in rice and 5.4 kb in wheat, but higher than the 7.4 kb in soybean, 14 kb in *Arabidopsis*, and 20 kb in cotton (Peng and Lapitan, 2005). Interesting, the SSR frequency (9.33%) (Yang et al., 2012a) in the *Nelumbo* genome sequences was lower than the SSR frequency in the transcriptomic sequences. Furthermore, SSR frequency is affected by many factors, such as the search parameters for exploring microsatellite markers, the determination tools, and species type.

In this study, the di-nucleotide repeat was most common in the *N. nucifera* transcriptome, just as in many crops, including Rosaceae (Dutta et al., 2011), sesame (Jung et al., 2005), *Pinus contorta* (Parchman et al., 2010), sweet potato (Wang et al., 2010), and pigeonpea (Jung et al., 2005; Wei et al., 2008; Parchman et al., 2010; Wang et al., 2010; Dutta et al., 2011). A total of 182 genic-SSR motif types were detected, of which the most dominant di- and tri-nucleotide repeat motifs were AG/CT (51%) and AAG/CTT (8%), respectively. The same motif proportions have been observed in many plants, e.g. pigeonpea (Wang et al., 2010; Dutta et al., 2011). Interestingly, the CCG/CGG and CG/CG motifs have the lowest dominant repeat type in *N. nucifera*, which is consistent with the results from dicotyledonous plants (Kumpatla and Mukhopadhyay 2005), such as *Epimedium sagittatum* (Jiang et al., 2012) and radish (Zeng et al., 2010; Jiang et al., 2012). These results showed that the CCG/CGG motif is the rarest motif in a large number of dicotyledonous plants. Interestingly the CCG/CGG motif proportion was high in monocots. This might be due to the high GC content and consequent codon usage bias in monocots (Morgante et al., 2002; La Rota et al., 2005). Moreover, the sequences containing CCG/CGG repeats might form potential higher structure, such as hairpins, and thus influence the efficiency and accuracy of RNA splicing (Coleman and Roesser, 1998; Zeng et al., 2010).

Development and validation of genic-SSRs markers

We synthesized and validated 110 primer pairs in order to evaluate the level of polymorphism so that new genic-SSR markers can be developed. A total of 101 primer pairs (91.8%) successfully yielded PCR products, which was in line with previous reports that generally found that 60-92% of primer pairs would be successfully amplified (Xin et al., 2005; Cloutier et al., 2009; Zhang et al., 2012). A total of 72 primers were randomly selected and synthesized, of which, 56 functioning primers were identified in 51 samples. The number of polymorphic primers was 45 (80.4%). This ratio was consistent with that for EST-SSRs in other plants, which range from 40 to 89% (Zhang et al., 2012). Genic-SSRs are generally thought to be less polymorphic than genomic-SSRs. This study supported this, and the polymorphism percentage (80.4%) was consistent with the 87% reported by Pan et al. (2010) for lotus. The *in-silico* polymorphism analysis between WFL and CRL found 1627 common genic-SSRs, of which 48 (3%) possessed polymorphic features. This indicated that there was a close relationship between WFL and CRL, which was supported by the dendrogram based on SSR genotyping data. However, seven polymorphic primers (9.7%, data not shown) were detected in both WFL and CRL from 72 randomly selected primers. This percentage is obviously higher than those obtained from the *in-silico* polymorphism analysis (3%). The differences between the two percentages may be caused by sequencing biases, in that high conserved sequences may be easier to detect simultaneously in WFL and CRL due to their relatively high abundance.

A total of 387 alleles were detected of 80 polymorphic primer pairs. These ranged from 2 to 17 per locus, which is higher than for the *N. nucifera* genomic-SSRs, which ranged from 2 to 5 per locus according to Kubo et al. (2009) and from 2 to 7 according to Tian et al. (2008a). The average number of alleles from the 80 loci (4.8 alleles) was also higher than the 3.88 (Tian et al., 2008a) and 3.9 alleles (Kubo et al., 2009) reported for genomic-SSR loci. The PIC values for each SSR locus ranged from 0.19 to 0.87 with a mean value of 0.55. The informativeness value for these EST-SSR markers (0.55) was a little higher than the genomic SSR marker value (0.51) for *N. nucifera* reported by Kubo et al. (2009); Pan et al. (2010) also reported 23 genic-SSRs from EST sequences found in public databases. The PIC values ranged from 0.02 to 0.61 with a mean of 0.33 ± 0.17, and the number of alleles per locus ranged from two to five, with an average of 2.65 alleles. The genic-SSRs obtained in this study have large PIC values and a considerable number of alleles. A preliminary linkage map of *Nelumbo* using a pseudo-testcross strategy has been constructed by Yang et al. (2012a). The development of these genic-SSRs will strongly support quantitative genetics research and molecular assisted selection, especially the functional gene mapping of lotus.

Assessment of genetic diversity among varieties and wild lotus

An UPGMA dendrogram was constructed based on genetic similarity results for 80 polymorphic genic-SSR markers. A total of 51 lotus plants were divided into three clear groups. *N. nucifera* (Cluster I) and *N. lutea* (Cluster III) were sub-clustered into two genetically distinct groups (Figure 4). All *N. nucifera* accessions were divided into two sub-clusters (Clusters Ia and Ib) based on their appearance and utilization value. The same results were reported by Yang et al. (2012b) and Li et al. (2010). Interestingly, the genetic relationship between these lotus accessions was not based on their geographical sources, as was found in previous reports (Guo et al., 2007). Currently, lotuses were classified into three types based on their morphological features and usage values during breeding (Nguyen, 2001; Guo et al., 2007; Guo, 2009). The cluster results showed that there was considerable genetic differentiation between the lotus rhizome (mainly Cluster Ia) and flower (mainly Cluster Ib), except for two accessions (19-Baihu wild lotus and 34-Baiyangdianhong). Wild lotuses have similar clustering to local varieties, which indicated that these local lotus varieties may be selected from wild lotus and gradually formed into cultivated resources without the use of systemic breeding technology. In summary, the relationships among lotuses gained in this diversity evaluation study will improve future breeding research efforts using cultivated varieties and wild lotus resources.

ACKNOWLEDGMENTS

We thank Chen Dong for help in improving the English grammar. Research supported by a grant from the National Science and Technology Supporting Program (#2012BAD27B01).

Supplementary material

REFERENCES

Chen YY, Zhou RC, Lin XD, Wu KQ, et al. (2008). ISSR analysis of genetic diversity in sacred lotus cultivars. *Aquat. Bot*. 89: 311-316.

Cloutier S, Niu Z, Datla R and Duguid S (2009). Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). *Theor. Appl. Genet*. 119: 53-63.

Coleman TP and Roesser JR (1998). RNA secondary structure: an important cis-element in rat calcitonin/CGRP pre-messenger RNA splicing. *Biochemistry* 37: 15941-15950.

Doyle JJ (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19: 11-15.

Dutta S, Kumawat G, Singh BP, Gupta DK, et al. (2011). Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus Cajan* (L.) Millspaugh]. *BMC Plant Biol*. 11: 17.

Fu J, Xiang QY, Zeng XB, Yang M, et al. (2011). Assessment of the genetic diversity and population structure of lotus cultivars grown in China by amplified fragment length polymorphism. *J. Am. Soc. Hortic. Sci*. 136: 339-349.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*. 29: 644-652.

Guo HB (2009). Cultivation of lotus (*Nelumbo nucifera Gaertn*. ssp *nucifera*) and its utilization in China. *Genet. Resour. Crop Evol.* 56: 323-330.

Guo HB, Li SM, Peng J and Ke WD (2007). Genetic diversity of *Nelumbo* accessions revealed by RAPD. *Genet. Resour. Crop Evol.* 54: 741-748.

Han YC, Teng CZ, Zhong S, Zhou MQ, et al. (2007). Genetic variation and clonal diversity in populations of *Nelumbo nucifera* (Nelumbonaceae) in central China detected by ISSR markers. *Aquat. Bot*. 86: 69-75.

Han YC, Teng CZ, Hu Z and Song YC (2008). An optimal method of DNA silver staining in polyacrylamide gels. *Electrophoresis* 29: 1355-1358.

Hu J, Pan L, Liu H, Wang S, et al. (2012). Comparative analysis of genetic diversity in sacred lotus (*Nelumbo nucifera* Gaertn.) using AFLP and SSR markers. *Mol. Biol. Rep.* 39: 3637-3647.

Jiang L, Wang L, Liu L, Zhu X, et al. (2012). Development and characterization of cDNA library based novel EST-SSR marker in radish (*Raphanus sativus* L.). *Sci. Hortic.* 140: 164-172.

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.

Kubo N, Hirai M, Kaneko A, Tanaka D, et al. (2009). Development and characterization of simple sequence repeat (SSR) markers in the water lotus (*Nelumbo nucifera*). *Aquat. Bot.* 90: 191-194.

Kumpatla SP and Mukhopadhyay S (2005). Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome* 48: 985-998.

La Rota MR, Kantety V, Yu JK and Sorrels ME (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 6: 23.

Li Z, Liu XQ, Gituru RW, Juntawong N, et al. (2010). Genetic diversity and classification of *Nelumbo* germplasm of different origins by RAPD and ISSR analysis. *Sci. Hortic.* 125: 724-732.

Liu X, Qu W and Liang J (2010). Research progress of *Nelumbo nucefera* Gaertn. *Strait Pharm. J*. 3: 003.

Luro FL, Costantino G, Terol J, Argout X, et al. (2008). Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina* Hort ex Tan) to other *Citrus* species and their effectiveness for genetic mapping. *BMC Genomics* 9: 287.

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

Ng P, Wei CL, Sung WK, Chiu KP, et al. (2005). Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2: 105-111.

Nguyen Q (2001). Lotus for export to Asia: An agronomic and physiological study. Publication number 1/32, Rural Industries Research and Development Corporation, Barton, ACT 2600.

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.

Pan L, Quan Z, Hu J, Wang G, et al. (2011). Genetic diversity and differentiation of lotus (*Nelumbo nucifera*) accessions assessed by simple sequence repeats. *Ann. Appl. Biol.* 159: 428-441.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, et al. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* 11: 180.

Peng J and Lapitan NL (2005). Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. *Funct. Integr. Genomics* 5: 80-96.

Rohlf F (2000). NTSYS-pc version 2.10 m, Numerical taxonomy and multivariate analysis system, computer software, Exeter Software, New York.

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

Sohn D-H, Kim Y-C, Oh S-H, Park E-J, et al. (2003). Hepatoprotective and free radical scavenging effects of *Nelumbo nucifera*. *Phytomedicine* 10: 165-169.

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.

Tian HL, Chen XQ, Wang JX, Xue JH, et al. (2008a). Development and characterization of microsatellite loci for lotus (*Nelumbo nucifera*). *Conserv. Genet*. 9: 1385-1388.

Tian HL, Xue JH, Wen J, Mitchell G, et al. (2008b). Genetic diversity and relationships of lotus *Nelumbo* cultivars based on allozyme and ISSR markers. *Sci. Hortic.* 116: 421-429.

Varshney RK, Graner A and Sorrells ME (2005). Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* 23: 48-55.

Wang Z, Fang B, Chen J, Zhang X, et al. (2010). *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* 11: 726.

Wei LB, Zhang HY, Zheng YZ, Guo WZ, et al. (2008). Developing EST-derived microsatellites in sesame (*Sesamum indicum* L.). *Acta Agron. Sin*. 34: 2077-2084.

Xin Y, Cui H, Zhang M, Lin R, et al. (2005). Development of EST (expressed sequence tags) marker in Chinese cabbage and its transferability to repeseed. *Hereditas* 27: 410-416.

Xue J, Wang S and Zhou SL (2012). Polymorphic chloroplast microsatellite loci in *Nelumbo* (Nelumbonaceae). *Am. J. Bot*. 99: e240-e244.

Yang M, Han Y, Vanburen R, Ming R, et al. (2012a). Genetic linkage maps for Asian and American lotus constructed using novel SSR markers derived from the genome of sequenced cultivar. *BMC Genomics* 13: 653.

Yang, M, Han Y, Xu L, Zhao J, et al. (2012b). Comparative analysis of genetic diversity of lotus (*Nelumbo*) using SSR and SRAP markers. *Sci. Hortic*. 142: 185-195.

Zeng S, Xiao G, Guo J, Fei Z, et al. (2010). Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* 11: 94.

Zhang H, Wei L, Miao H, Zhang T, et al. (2012). Development and validation of genic-SSR markers in sesame by RNA-seq. *BMC Genomics* 13: 316.