RESEARCH ARTICLE

De novo assembly, transcriptome characterization, and simple sequence repeat marker development in duckweed *Lemna gibba*

Lili Fu^{1,2} · Zehong Ding^{1,2} · Anuwat Kumpeangkeaw^{3,4} · Deguan Tan^{1,2} ·
Bingying Han^{1,2} · Xuepiao Sun^{1,2} · Jiaming Zhang^{1,2}

Received: 15 March 2019 / Revised: 11 September 2019 / Accepted: 18 October 2019 / Published online: 9 December 2019
© Prof. H.S. Srivastava Foundation for Science and Society 2019

Abstract *Lemna gibba* is a species of duckweed showing great potential in bioenergy production and wastewater treatment. However, the relevant transcriptomic and genomic resources are very limited for this species, which dramatically hinders its genetic diversity and genome mapping researches. In this work, ~ 233.5 million clean reads were generated from *L. gibba* by Illumina paired-end sequencing, and subsequently they were de novo assembled into 131,870 unigenes, of which 61,622 were annotated and 43,319 were expressed with Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) > 5. In total, 19,297 simple sequence repeats (SSRs) were identified from 15,261 SSR-containing unigenes. Dinucleotide (78.4%) were the most abundant SSRs, followed by tri- (14.9%), tetra- (4.1%), and penta-

nucleotides (1.5%). The top three motifs were AG/CT (69.9%), AC/GT (6.5%), and ATC/ATG (4.9%). Further analysis revealed that the presence of SSR motif was independent of the expression level for a given gene. Based on the sequence of these SSR-containing unigenes, a total of 10,292 SSR markers were developed, of which only 2671 were further retained after removing those derived from unannotated or extra-low expressed (e.g., FPKM ≤ 5) unigenes. Finally, a subset of 70 SSR markers was randomly selected and examined in nine diverse *L. gibba* genotypes for the PCR amplification and polymorphism, as well as in other duckweed species for the inter-specifically amplifiability. This work is the first report on the transcriptome-based large-scale SSR markers development and analysis in *L. gibba*. The transcriptome generated and the SSR markers developed in this work will provide a valuable resource for genetic diversity assessment in *L. gibba* and also for species relationship investigation in *Lemnaceae* family.

Lili Fu and Zehong Ding have contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12298-019-00726-9>) contains supplementary material, which is available to authorized users.

✉ Zehong Ding
dingzehong@itbb.org.cn

✉ Jiaming Zhang
zhangjiaming@itbb.org.cn

Lili Fu
fulili@itbb.org.cn

Anuwat Kumpeangkeaw
flaaaay66@hotmail.com

Deguan Tan
tandeguan@itbb.org.cn

Bingying Han
hanbingying@itbb.org.cn

Xuepiao Sun
sunxuepiao@itbb.org.cn

- ¹ Institute of Tropical Bioscience and Biotechnology, MOA Key Laboratory of Tropical Crops Biology and Genetic Resources, Hainan Bioenergy Center, Chinese Academy of Tropical Agricultural Sciences, Xueyuan Road 4, Haikou 571101, China
- ² Hainan Academy of Tropical Agricultural Resource, Chinese Academy of Tropical Agricultural Sciences, Xueyuan Road 4, Haikou 571101, China
- ³ International College, Huazhong Agricultural University, Wuhan 430070, China
- ⁴ Department of Agriculture, Song Khla Agricultural Research and Development Center, Ministry of Agriculture and Cooperatives, Had Yai, Song Khla 90110, Thailand

Keywords *Lemna gibba* · Transcriptome · De novo assembly · SSR marker

Introduction

Lemnaceae, also widely known as duckweed, represents a family of aquatic floating plants showing great potential in modern agriculture and industry. For example, duckweed has good nutritional values and high protein content, thus it has widely been used as food for poultry and livestock (Leng et al. 1995; Anderson et al. 2011). Because of its high starch content and fast growth rate, duckweed is considered as a new type of energy plant for bioethanol and biobutanol production (Qian et al. 2012). Besides, duckweed has good stress-tolerance and can rapidly remove nutrients (such as nitrogen and phosphorus) and heavy metals from sewage (El-Shafai et al. 2007), therefore, it has also been applied for the municipal and industrial wastewater treatment.

Duckweed includes five genera of *Lemna*, *Landoltia*, *Spirodela*, *Wolffiella*, and *Wolffia* with a total of 37 species (Appenroth et al. 2013). With the rapid development of next-generation sequencing technology, many significant progresses have been made in recent decades in duckweed genomics and transcriptomics (Wang and Messing 2015; An et al. 2018). These studies were mainly focused on a few species, including *Spirodela polyrrhiza* (Wang et al. 2014a, b), *Lemna minor* (Van Hoeck et al. 2015, 2017), *Lemna aequinoctialis* (Yu et al. 2017), and *Landoltia punctata* (Tao et al. 2013). However, until now no studies were conducted concerning the genome and transcriptome in *Lemna gibba* that has been widely used in research and application (Fu et al. 2017). To date (Mar 2019), there are only ~ 160 records related to *L. gibba* in NCBI Nucleotide database, which greatly limits the genomic researches in this species (e.g., primer design for hybrid identification) (Fu et al. 2017), and this status needs to be urgently improved.

RNA-seq is a common tool to quantify gene expression in plants, and it has widely been applied to identify specific trait-associated genes and regulatory networks (Ding et al. 2016; Fu et al. 2016). In addition, it is also used to identify single nucleotide polymorphisms (SNPs) for SNPs-traits association (Chen et al. 2018a) or to detect simple sequence repeats (SSRs) for genetic diversity assessment (Chen et al. 2018b; Deng et al. 2018), especially for the species without available genomes.

In this study, RNA-seq was employed to de novo assemble the transcripts in *L. gibba*, based on which SSRs were identified and the corresponding markers were developed, and finally a subset of 70 SSR markers was tested in nine diverse *L. gibba* accessions as well as four additional accessions from other duckweed species. To our

knowledge, this is the first report on the large-scale development of SSR markers in *L. gibba*. These results will offer a valuable resource for further population genetic assay and germplasm assessment in this species.

Materials and methods

Plant material

Lemna gibba clone 7741 was used and cultured as previously described (Fu et al. 2017) at the Biological Collection Centre of the Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Hainan, China. Generally, three-frond colony was inoculated in a 250 mL flask containing 100 mL MH medium with five replicates, and then cultivated under 16 h daily photoperiod (25 ± 1 °C) at $40 \mu\text{mol m}^{-2} \text{s}^{-1}$ photosynthetic active radiation (PAR). One week later, the whole plants of *L. gibba* were collected and frozen immediately in liquid nitrogen for RNA-seq sequencing.

RNA isolation and sequencing

Total RNA was isolated using RNAiso reagent (OMEGA), and then reversely transcribed into cDNA using PrimeScriptTM RT reagent Kit with gDNA Eraser (Takara, Dalian, China) in accord with the manufacturer's instruction. The RNA quality and integrity were assayed by a Nanodrop-2000 spectrophotometer (Thermo Scientific Inc., USA) and an Agilent-2100 Bioanalyzer (Agilent, USA). RNA-seq libraries (three replicates in total) were generated with Illumina TruSeqTM RNA sample prep Kit (Illumina, San Diego, USA) using Ribo-Zero Magnetic kit to deplete rRNA according to the manufacturer's protocol, and then sequenced on Hiseq 2500 platform in PE150 model.

De novo transcriptome assembly and functional annotation

Adapters were removed from raw reads using trimmomatic software v0.38 (Bolger et al. 2014). Sequence quality was inspected by FastQC program v0.11.4 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), and low quality reads (e.g., the reads with 'N' percentage > 5%) were removed using customized Perl scripts.

The clean reads were de novo assembled using the Trinity program (Grabherr et al. 2011) with following parameters: `-seqType fq -SS_lib_type FR -min_contig_length 200; -min_glue 2 -group_pairs_distance 500; -path_reinforcement_distance 75 -min_kmer_cov 2`. Trinotate software (<http://trinotate.github.io>) was used to annotate the function of assembled unigenes. Besides,

Blast2GO software (Conesa et al. 2005) was also used to assign the unigenes into three gene ontology (GO) categories including cellular component, molecular function, and biological process. The expression levels of unigenes were computed as Fragments Per Kilobase of transcript per Million fragments mapped (FPKM).

Identification and development of SSR markers

The sequences of unigenes were used for SSR detection using MISA software (<http://pgrc.ipk-gatersleben.de/misa/misa.html>). The SSR loci containing two to seven nucleotide motifs with a minimum of 7, 6, 5, 4, 4, and 4 repeats, respectively, were identified. The SSR primers were designed using Primer3 v2.3.6 (Untergasser et al. 2012).

DNA extraction, PCR amplification, and SSR marker validation

A set of nine diverse *L. gibba* accessions (including two from Italy, one each from China, Finland, Germany, India, and Argentina, and two without source information) were used for polymerase chain reaction (PCR) amplification and SSR marker confirmation. In addition, one each accession from *Wolffiella neotropica*, *Wolffia brasiliensis*, *Spirodela polyrhiza*, and *Landoltia punctata* was also used to test the inter-specifically amplifiability. DNA of each accession was extracted using the CTAB method, respectively.

The PCR reaction was conducted in 20 μ L mixture containing $\sim$ 100 ng of DNA template, 1 μ L of each primer (10 μ M), 3 μ L of 10 $\times$ Taq buffer, 1.5 μ L of dNTPs (10 mM), and 1 μ L of Taq DNA polymerase. The PCR amplification conditions for SSR primers were: 94 $^{\circ}$ C for 4 min, followed by 34 cycles of 94 $^{\circ}$ C for 35 s, 55 $^{\circ}$ C for 35 s, and 72 $^{\circ}$ C for 35 s, followed by 72 $^{\circ}$ C for 5 min and stored at 4 $^{\circ}$ C. Amplification products were separated using 4% polyacrylamide gel with a 100-bp DNA ladder. SSR alleles were visualized by silver staining as described previously (Wang et al. 2011). Genetic parameters including the number of alleles (NA), polymorphism information content (PIC), and expected heterozygosity (He) were analyzed using Cervus 3.0 package (Kalinowski et al. 2007).

Results

Paired-end sequencing and de novo assembly

In total, $\sim$ 248.7 million pair-end reads of 150-bp were generated from RNA-seq sequencing. Subsequently, the

Trinity program was used to de novo assemble the transcriptome. Following rigorous steps of quality control and data filtering, totally $\sim$ 233.5 million clean reads were obtained, and they were further assembled into 157,556 transcripts, with an N50 of 1115 bp (Table 1). The average length of these transcripts was 679 bp, and the GC content was 49.93%. These transcripts ranged from 201 bp to 19,825 bp, and most (81.3%) of them were shorter than 1000 bp.

In total, 131,870 unigenes were de novo assembled, with an average length of 533 bp and an N50 of 669 bp. Similar to the transcripts, these unigenes varied from 201 bp to 19,825 bp, and most (88.7%) of them were shorter than 1000 bp (Table 1).

To comprehensively capture the read alignments, clean reads were subsequently aligned back to the assembled transcripts, and the numbers of proper pairs or orphan-read alignments were counted. The results showed that $\sim$ 92.3% reads were mapped as proper pairs and the overall alignment rate was up to 96.9%, suggesting that the transcriptome was properly assembled.

Functional annotation of unigenes

To functionally annotate the assembled unigenes, their sequences were used as queries to BLAST against NCBI non-redundant protein (Nr) and Swiss-Prot databases. A total of 33,210 unigenes were closely matched to the known sequences deposited in Nr database, and the species included most matched sequences were *Mus musculus* (9640), *Elaeis guineensis* (3497), and *Phoenix dactylifera* (2884). Additionally, 58,695 unigenes were annotated by matching the sequences in Swiss-Prot database.

Gene Ontology (GO) applies a series of terms to functionally characterize genes into three major categories of cellular component, biological process, and molecular function. In total, 25,564 unigenes corresponding to known proteins were classified into 15,243 GO terms. Most of unigenes were related to the biological process category (51.6%), then followed by the cellular component (31.0%) and molecular function (17.4%, Fig. 1). By merging the results from different databases, together, a total of 61,622 (46.7%) unigenes were finally annotated, and they will be considered as a valuable source for genetic researches in *L. gibba*.

Expression analysis of unigenes

To investigate the expression characteristics of unigenes, they were classified into five different groups, including extra-low expressed (FPKM $\leq$ 5), low-expressed (5 < FPKM $\leq$ 50), mid-expressed (50 < FPKM $\leq$ 200), high-expressed (200 < FPKM $\leq$ 500), and extra-high

Table 1 Length distribution of the assembled transcripts and unigenes

Nucleotides length (bp)	Number of transcripts	Number of unigenes
Number of transcripts/unigenes	157,556	131,870
of which, 100–500 bp	99,662	94,618
500–1000 bp	28,451	22,374
1000–1500 bp	12,437	7204
1500–2000 bp	7249	3644
2000–3000 bp	6312	2788
≥3000 bp	3445	1242
Min gene length (bp)	201	201
Max gene length (bp)	19,825	19,825
Mean gene length (bp)	679	533
N50 length (bp)	1115	669
GC (%)	49.93	49.82
Total base (Mb)	106.9	70.3

Fig. 1 Classification of unigenes based on the GO term annotation. *BP* biological process, *MF* molecular function, *CC* cellular component

Fig. 2 The frequency of unigenes according to their expression levels. The number above the column indicates the unigenes within each group

expressed ($500 \leq \text{FPKM}$), according to their expression levels. The groups with most genes were extra-low expressed (67.2%), then followed by low-expressed (30.6%) and mid-expressed (1.9%), and only a few genes were high-expressed (0.2%) or extra-high expressed (0.1%), indicating that the majority of *L. gibba* unigenes were extra-low expressed or low-expressed (Fig. 2).

Identification and characterization of SSRs

The predicted unigenes were used to identify repeated motifs for SSR marker development. In total, 19,297 SSRs, including 15,133 dinucleotide repeats, 2872 trinucleotide repeats, 787 tetranucleotide repeats, 281 pentanucleotide repeats, 211 hexanucleotide repeats, and 13 heptanucleotide repeats (Table 2), were identified with an average of one SSR per 3.6 kb. Further analysis showed that these SSRs were distributed within the sequences of 15,261 unigenes, of which 12,095 (79.3%) contained only one SSR while the remaining 3166 (20.7%) unigenes contained more than one SSR (Table 3).

Totally 180 repeated motifs were identified. The motifs with greater frequency were AG/CT (13,486, 69.9%), followed by AC/GT (1249, 6.5%), ATC/ATG (940, 4.9%), AAG/CTT (707, 3.7%), and CCG/CGG (485, 2.5%) (Table 4). Further analysis revealed that dinucleotide repeats were the most abundant (78.4%), followed by trinucleotide (14.9%) and tetranucleotide (4.1%) repeats. It's noteworthy that SSRs mainly comprised seven repeats (20.7%), followed by eight (14.8%), six (8.1%), and nine (7.2%) repeats (Table 2).

We also determined the relationships between the SSRs and expression levels of these unigenes, and found that the presentation of SSR motif was independent of the expression level for a given gene, i.e., AG/CT was the most dominant SSR motif presented in unigenes no matter what kinds of expression levels they were.

Development and validation of SSR markers

In total, 10,292 SSR markers (accounting for 67.4% of the SSR loci) were developed based on the sequences of 15,261 unigenes containing repeated motifs. Because the expressed and well-annotated unigenes were more attractive in the genetic application, therefore, these SSR markers, whose corresponding unigenes were unannotated in this work or the expression $\text{FPKM} \leq 5$, were further removed, resulting in a total of 2671 SSR markers for 2378 unigenes (Table S1). This set of primers will provide valuable foundations for genetic researches in *L. gibba*.

To validate the SSR developed, in total 70 SSR markers were randomly selected and tested them in nine genotypes of *L. gibba* for the PCR amplification and polymorphism. Fifty-six (80%) SSR markers amplified clear bands with the expected size, of which 13 markers were monomorphic, and 43 were polymorphic with allele number from two to four with a mean of 2.63 (Table 5). These validated SSR markers can be applied to investigate the genetic diversity of *L. gibba* population and germplasm.

Furthermore, the SSR markers were also examined among different duckweed species, including one sample each from *W. neotropica*, *W. brasiliensis*, *S. polyrhiza*, and *L. punctata*, to test their inter-specifically amplifiability.

Table 2 Length distribution of SSRs based on the number of nucleotide repeat units

No. of repeats	Di-	Tri-	Quad-	Penta-	Hexa-	Hepta-	Total	Percent (%)
4	NA	NA	NA	193	181	8	382	1.98
5	NA	NA	357	48	10	1	416	2.16
6	NA	1237	305	14	11	2	1569	8.13
7	3329	638	23	7	2	0	3999	20.72
8	2314	500	33	7	2	0	2856	14.80
9	1323	48	14	3	2	0	1390	7.20
10	897	81	6	5	0	2	991	5.14
11	1256	70	13	2	1	0	1342	6.95
12	824	68	12	0	0	0	904	4.68
13	308	48	5	1	1	0	363	1.88
14	429	36	5	0	0	0	470	2.44
15	442	27	3	1	0	0	473	2.45
16	457	17	2	0	0	0	476	2.47
17	395	22	0	0	0	0	417	2.16
18	388	16	3	0	1	0	408	2.11
19	378	12	1	0	0	0	391	2.03
20	349	12	1	0	0	0	362	1.88
21	310	6	0	0	0	0	316	1.64
22	286	6	0	0	0	0	292	1.51
23	254	7	0	0	0	0	261	1.35
24	205	3	0	0	0	0	208	1.08
25	156	6	0	0	0	0	162	0.84
26	128	1	0	0	0	0	129	0.67
27	104	2	0	0	0	0	106	0.55
28	89	2	0	0	0	0	91	0.47
29	78	4	1	0	0	0	83	0.43
30	77	1	0	0	0	0	78	0.40
>30	357	2	3	0	0	0	362	1.88
Total	15,133	2872	787	281	211	13	19,297	
Percent (%)	78.42	14.88	4.08	1.46	1.09	0.07		

NA data were not available

Table 3 Summary of SSRs identified in *L. gibba* transcriptome unigenes

Information	Number
Total number of sequences examined	1,31,870
Total size of examined sequences (bp)	7,02,94,856
Total number of identified SSRs	19,297
Number of SSR containing sequences	15,261
Number of sequences containing more than 1 SSR	3166
Number of SSRs present in compound formation	2244

However, only three (4.3%) of them amplified clear bands with the expected PCR-product size, indicating a relatively low cross-species transferability in duckweed.

Discussion

To the best of our knowledge, to date, there was no report concerning SSR marker development in *L. gibba*, which greatly limited the studies on its genetic diversity and genome mapping. Therefore, the SSR markers need to be urgently developed in *L. gibba*.

The discovery of SSR loci can be made by next-generation sequencing technology (e.g., RNA-seq) even without reference genome sequences. For examples, RNA-seq approach has been applied to de novo assemble transcripts, based on which thousands of SSR markers were developed in *Alfalfa* (Liu et al. 2013), sweet potato (Wang et al. 2010), and *Zingiber striolatum* (Deng et al. 2018), respectively. In this study, a total of 157,556 transcripts were obtained in *L. gibba* by RNA-seq sequencing.

Table 4 Distribution of SSR types in the *L. gibba* genome

SSRs_motif	Repeat_number	SSRs_motif	Repeat_number	SSRs_motif	Repeat_number
AG/CT	13,486	ACG/CGT	85	AAAAG/CTTTT	18
AC/GT	1249	AAC/GTT	83	AGAGAT/ATCTCT	16
ATC/ATG	940	AAAT/ATTT	82	AAAAT/ATTTT	15
AAG/CTT	707	AAT/ATT	78	AGGC/CCTG	14
CCG/CGG	485	ATCC/ATGG	64	AAGAG/CTCTT	14
AT/AT	380	ACAT/ATGT	41	ATGC/ATGC	11
AGAT/ATCT	237	AAAG/CTTT	40	AAAGTC/ACTTTG	11
AGC/CTG	214	ACAG/CTGT	38	AACCCT/AGGGTT	11
AGG/CCT	187	AGGG/CCCT	30	AACC/GGTT	10
AAAC/GTTT	132	AAGG/CCTT	26	AATT/AATT	10
ACC/GGT	90	AGAGG/CCTCT	19	ACCCAT/ATGGGT	10
AAAAC/GTTTT	87	CG/CG	18		

SSRs motif with repeat number less than 10 was not shown

Additionally, 131,870 unigenes were further de novo assembled with an average length of 533 bp. This length was longer than that reported in tea (402 bp) (Tan et al. 2013), but shorter than that described in sweet potato (581 bp) (Wang et al. 2010) and *Alfalfa* (803 bp) (Liu et al. 2013). Subsequently, all these unigenes were annotated by several databases such as NCBI-Nr, Swiss-Prot, and GO terms, resulting in a total of 61,622 unigenes corresponding to known proteins. According to the expression levels, most *L. gibba* unigenes were extra-low expressed (67.2%) or low-expressed (30.6%), and very a few were mid-expressed or high-expressed. This subset of unigenes that were well-annotated and not extra-low expressed will provide a valuable resource for functional genomics analysis in *L. gibba*.

The assembled sequences of 131,870 *L. gibba* unigenes were subsequently used to investigate the distribution of SSRs. In total, 19,297 SSR loci were identified with a frequency of 11.5%, which is greater than that of banana (5.3%) (Wang et al. 2008), sorghum (3.6%) (Kantety et al. 2002), and sugarcane (2.9%) (Cordeiro et al. 2001), but much lower than that of *Lonicera caerulea* (32.51%) (Zhang et al. 2016) and coffee (17.3%) (Aggarwal et al. 2007). These differences might reflect the variability in the SSR search criteria, size of sequence data, and the species used. In this work, it was found that di-nucleotide repeats were the most frequent SSRs, in accord with that reported in *Pareuchiloglanis sinensis* (Chen and He 2018), okra (Ravishankar et al. 2018), and *Chlorostoma rustica* (Wang et al. 2018). However, in some species such as *Pogostemon cablin*, trinucleotide repeats were the most frequent SSRs (Ouyang et al. 2018). Notably, AG/CT was the most frequent motif among the dinucleotide repeats in our dataset,

which might indicate that this motif plays an important role in gene expression regulation, since CT motif frequently occurs in the 5' UTRs of a gene (Wei et al. 2011).

Based on the sequences of unigenes containing repeated motifs, a total of 10,292 SSR markers were developed. Of which, the SSR markers derived from unigenes that were expressed greater than a certain level (e.g., FPKM > 5) and well-annotated in this study, were more attractive and further retained, because these markers were more likely embedded in functional gene sequences and showed more transferability between species (Varshney et al. 2005). Moreover, compared with the SSR markers derived from the intron regions and inter-genic regions (usually based on genomic sequences), those detected in the 5'UTR and coding regions (usually based on transcriptomic data) are more practical and might be more closely related to traits because the variations in the 5'UTR and coding regions often result in phenotypic diversity (Chen et al. 2019). Finally, a total of 2671 SSR markers for 2378 unigenes were kept, and 70 of them were randomly picked and tested in nine diverse genotypes of *L. gibba* for the PCR amplification and polymorphism. The number of allele (NA) and polymorphism information content (PIC) ranged from 2 to 4 and from 0.18 to 0.64, with a mean of 2.63 and 0.42, respectively (Table 5). These values were greatly lower than that reported in *Alfalfa* (Liu et al. 2013), *Angelica dahurica* (Chen et al. 2019), and Indian garlic (*Allium sativum*) (Kumar et al. 2019), but showed an equivalent level to that reported in rice bean (*Vigna umbellata*) (Chen et al. 2016) and cowpea (*Vigna unguiculata*) (Chen et al. 2017). Together, these SSR primers generated in this work will provide valuable foundations for genetic diversity

Table 5 Summary of 43 polymorphic SSR primers in *L. gibba*

SSR marker	SSR motif	Left-primer	Right-primer	Tm (°C)	Product size (bp)	NA	He	PI _C
ssr01	(CT) ₉	AAACGTTCTAGCTCGACGGG	TCTGGTTTCCTAGGGACGT	55	100	2	0.47	0.35
ssr04	(AG) ₈	CATCTCTGTGGCTCGGTTT	CCAATGCACGTGGTTTCTGG	55	100	3	0.62	0.5
ssr06	(AT) ₁₀	TCGCGGTTACTCTCTCTACCA	GCATCTCGCCGTACTTCTT	55	100	3	0.66	0.53
ssr07	(GA) ₁₈	ACACTCGGACGAGCATTTCTC	CACCGGATCTTCCCGAATA	55	100	4	0.73	0.64
ssr09	(CCG) ₈	GAGCATCGGCATCTTCTCGA	TCTCCAGCGACCTTTC AAC	55	100	3	0.66	0.53
ssr10	(CT) ₆	GAGCTTCTTCCATGGCCACT	GGAAAGCCAGCTCGAAGA	55	100	2	0.26	0.22
ssr14	(GAT) ₈	GGCGATGATGACGACGA	AGCTCAACATAGTAGCCACA	55	102	2	0.5	0.36
ssr16	(GGC) ₆	TATGAAGAAGAGACCGGGC	GCTTCCCTCTGGCTTCAACT	55	103	3	0.63	0.51
ssr20	(ATGG) ₅	AATCGGACGGGATGAATGGG	ATGCACCTCTCTCAACTGC	55	123	3	0.62	0.5
ssr21	(CT) ₁₃	AGTCAACAGCCCGAAGTTCC	CGTGTGCGGCTTAGATCTCT	55	101	3	0.7	0.58
ssr22	(TC) ₇ ...(TC) ₇	GCTGCTCTGCTTAGTCCTCA	GCGGAGATAAGCCCTCTTCC	55	101	2	0.53	0.37
ssr23	(GA) ₁₉	CACTCGGACGAGCATTTCTCA	CACCGCATCTTCCCGAATA	55	101	4	0.73	0.64
ssr24	(AG) ₉	GAGGAATCCCCCTTCTCCCT	TCATGCATCAGCCAACCCCT	55	102	2	0.47	0.35
ssr25	(AG) ₁₁	CGAGATTGACAGGCCATTG	CGTGGGATCCCTAACCACT	55	102	3	0.62	0.5
ssr27	(CT) ₁₄	AATAACCATTTGGGGCCCCA	TTTCTTCTGATTGCCGGCT	55	102	3	0.68	0.57
ssr28	(TC) ₂₀	CATCTCTGACGCTCTCTC	ACATGATCTCTCTGCCCT	55	102	3	0.66	0.53
ssr29	(GA) ₂₁	CACTCGGACGAGCATTTCTCA	CGCGATCTTCCCGAATACCA	55	102	4	0.73	0.64
ssr30	(GA) ₇ ...(AG) ₇	CCTTCTCCCTCTCCCTCC	GAGTCGCTGCAACCGATAGA	55	102	3	0.68	0.57
ssr32	(CT) ₈	ACCACGAATGCCACGAAGAT	CGTCCCGCTTCCCATTTTC	55	103	3	0.7	0.58
ssr33	(AG) ₈	TTCAATCCAACCCAGTCCGG	CTGCCCTTCTCTCTCGTCTG	55	103	3	0.5	0.4
ssr35	(GA) ₉	AGGACAAACGGAGAGATCG	CGGTACGGTTTCTCTCTAT	55	103	3	0.66	0.53
ssr36	(CGC) ₆	GCTTCCCTCTGGCTTCAACT	TATGAAGAAGAGACCGGGC	55	103	2	0.23	0.2
ssr37	(GA) ₁₂	CACCGGGTCTCTCCACTAGA	GAGGAGAGCTTGATCGACGG	55	103	2	0.26	0.22
ssr38	(GA) ₁₂	CCTCTGAGACAGCCCTCTCT	GAATCTCTATCACGCCCCG	55	103	2	0.26	0.22
ssr39	(TCT) ₆	CATGTCTACCGTCTCCTCGC	GTGTGGTTGAGCCCATGGTA	55	104	3	0.63	0.51
ssr41	(TC) ₁₀	GGGTGGCGACTCTTGTTTA	TCTCATCCCCCACCAGTCT	55	104	3	0.48	0.41
ssr42	(TTC) ₇	GATCGCCGGTGAGGGTAAAA	GCTCGACAGGTTGATCAAT	55	104	2	0.55	0.38
ssr43	(TCT) ₇	TGTGTCTTCGCTTCGTCGT	TCCTTGAAGAGGCGGAGAGA	55	104	2	0.49	0.35
ssr44	(CT) ₁₅	AATAACCATTTGGGGCCCCA	TTTCTTCTGATTGCCGGCT	55	104	3	0.68	0.57
ssr45	(CT) ₁₀ ...(CT) ₇	CATGTCAGCTCGCTCTCTCA	TGGGTAAGGTTTCATGCTCA	55	104	2	0.26	0.22
ssr46	(TC) ₉	CGGTGTGTGAAAGGAAGCCCC	CTTCGTACTCCATCTCTCGGC	55	105	2	0.37	0.29
ssr49	(TC) ₁₅	AACTTCTCAGGGATGCCACG	ACAGCGTTCTGCTAGAGAG	55	105	3	0.62	0.5
ssr50	(AGA) ₁₄	CACCGTCACCAAGTCCAC	GTTCATCTCAAGGCCCTCTC	55	105	2	0.37	0.29
ssr51	(CT) ₁₀	TCGAACTCGACATCTGACC	AGTGTCCCAACTGCCACTC	55	106	4	0.75	0.64
ssr52	(CT) ₁₀	TGCGCATCATCTCTTCTCTC	ACTGGTCTCTCTCTCTCAC	55	106	3	0.48	0.41
ssr55	(GA) ₈	GGTTCATGGTAGTTCGGCGA	TGTTTCATCATCGGCACCATCA	55	107	2	0.53	0.37

Table 5 continued

SSR marker	SSR motif	Left-primer	Right-primer	T _m (°C)	Product size (bp)	NA	He	PIC
ssr57	(TC) ₁₀	TAGCCACGTTTGACAGGAGAG	TCCTTCTCATCATCTCTTTTCCA	55	107	2	0.53	0.37
ssr59	(CT) ₁₂	GCCTAACGGGAGAGGGAGTA	TCAGCCCTCTTTCACGAGAT	55	107	2	0.26	0.22
ssr60	(AG) ₁₄	TTTGGGGGTCAAGGTTGGAC	ACCTATCAACACCTAGGCTCA	55	107	2	0.44	0.33
ssr64	(AG) ₁₀	GCGAATAGCGTGAATAACGTGA	GTCGAGGGTTTGCTGTCTCA	55	108	3	0.63	0.53
ssr65	(CGC) ₇	GCTCGATCGCCGATTCTTA	TCTAGCGGAAGAGGATCGGA	55	108	2	0.26	0.22
ssr68	(AG) ₈	AAGCTCACCATTTGGGGGAC	TCTCTCTCGCCCTCTTCAA	55	109	2	0.21	0.18
ssr69	(TCA) ₆	CGGTCACTAAGAACATGCCCT	AGATGAGTTTGAGTCGCCCG	55	109	2	0.26	0.22

NA number of alleles, He expected heterozygosity, PIC polymorphism information content

assessment in *L. gibba* and also for species relationship investigation in *Lemnaceae* family.

Acknowledgements This project was funded by the Natural Science Foundation of China (31800301), Central Public-interest Scientific Institution Basal Research Fund for Chinese Academy of Tropical Agricultural Sciences, and the International Science and Technology Cooperation Program of China (2014DFA30680).

Authors' contribution ZD and JZ designed the project. LF, ZD, KA, DT, BH, and XS performed the experiments and analyzed the data. ZD performed the bioinformatic analysis. LF and ZD drafted the manuscript, and ZD and JZ revised the paper. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Lalji S (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
- An D, Li C, Zhou Y, Wu Y, Wang W (2018) Genomes and transcriptomes of duckweeds. *Front Chem* 6:230
- Anderson KE, Lowman Z, Stomp AM, Chang J (2011) Duckweed as a feed ingredient in laying hen diets and its effect on egg production and composition. *J Poult Sci* 10:4–7
- Appenroth KJ, Borisjuk N, Lam E (2013) Telling duckweed apart: genotyping technologies for the *Lemnaceae*. *Chin J Appl Environ Biol* 19:1–10
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120
- Chen W, He S (2018) Isolation and characterization of microsatellite markers in a highland fish, *Pareuchiloglanis sinensis* (Siluriformes: Sisoridae) by next-generation sequencing. *J Genet* 97:e111–e116
- Chen H, Chen X, Tian J, Yang Y, Liu Z, Hao X et al (2016) Development of gene-based SSR markers in rice bean (*Vigna umbellata* L.) based on transcriptome data. *PLoS One* 11:e0151040
- Chen H, Wang L, Liu X, Hu L, Wang S, Cheng X (2017) De novo transcriptomic analysis of cowpea (*Vigna unguiculata* L. Walp.) for genic SSR marker development. *BMC Genet* 18:65
- Chen X, Liu X, Zhu S, Tang S, Mei S, Chen J et al (2018a) Transcriptome-referenced association study of clove shape traits in garlic. *DNA Res* 25:587–596
- Chen Z, Ding G, Wang Y, Xu J, Lin Z (2018b) Development and characterization of 33 microsatellite loci for the tiger frog *Hoplobatrachus rugulosus* (Wiegmann 1834) through transcriptome sequencing. *J Genet* 97:e147–e151
- Chen C, Chen Y, Huang W, Jiang Y, Zhang H, Wu W (2019) Mining of simple sequence repeats (SSRs) loci and development of novel transferability-across EST-SSR markers from de novo transcriptome assembly of *Angelica dahurica*. *PLoS One* 14:e0221040
- Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization

- and analysis in functional genomics research. *Bioinformatics* 21:3674–3676
- 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* 160:1115–1123
- Deng K, Deng R, Fan J, Chen E (2018) Transcriptome analysis and development of simple sequence repeat (SSR) markers in *Zingiber striolatum* Diels. *Physiol Mol Biol Plants* 24:125–134
- Ding Z, Zhang Y, Xiao Y, Liu F, Wang M, Zhu X et al (2016) Transcriptome response of cassava leaves under natural shade. *Sci Rep* 6:31673
- El-Shafai SA, El-Gohary FA, Nasr FA, van der Steen NP, Gijzen HJ (2007) Nutrient recovery from domestic wastewater using a UASB-duckweed ponds system. *Bioresour Technol* 98:798–807
- Fu L, Ding Z, Han B, Hu W, Li Y, Zhang J (2016) Physiological investigation and transcriptome analysis of polyethylene glycol (PEG)-induced dehydration stress in cassava. *Int J Mol Sci* 17:283
- Fu L, Huang M, Han B, Sun X, Sree KS, Appenroth KJ et al (2017) Flower induction, microscope-aided cross-pollination, and seed production in the duckweed *Lemna gibba* with discovery of a male-sterile clone. *Sci Rep* 7:3047
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652
- Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol Ecol* 16:1099–1106
- 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
- Kumar M, Rakesh Sharma V, Kumar V, Sirohi U, Chaudhary V, Sharma S et al (2019) Genetic diversity and population structure analysis of Indian garlic (*Allium sativum* L.) collection using SSR markers. *Physiol Mol Biol Plants* 25:377–386
- Leng RA, Stambolie JH, Bell R (1995) Duckweed—a potential high-protein feed resource for domestic animals and fish. *Livestock Res Rural Dev* 7:1–12
- Liu Z, Chen T, Ma L, Zhao Z, Zhao PX, Nan Z et al (2013) Global transcriptome sequencing using the Illumina platform and the development of EST-SSR markers in autotetraploid alfalfa. *PLoS One* 8:e83549
- Ouyang P, Kang D, Mo X, Tian E, Hu Y, Huang R (2018) Development and characterization of high-throughput EST-based SSR markers for *Pogostemon cablin* using transcriptome sequencing. *Molecules* 23:2014
- Qian C, Jin YL, Zhang GH, Yang F, Yao X, Hai Z (2012) Improving production of bioethanol from duckweed (*Landoltia punctata*) by pectinase pretreatment. *Energies* 5:3019–3032
- Ravishankar KV, Muthaiah G, Mottaiyan P, Gundale SK (2018) Identification of novel microsatellite markers in okra (*Abelmoschus esculentus* (L.) Moench) through next-generation sequencing and their utilization in analysis of genetic relatedness studies and cross-species transferability. *J Genet* 97:e39–e47
- Tan LQ, Wang LY, Wei K, Zhang CC, Wu LY, Qi GN et al (2013) Floral transcriptome sequencing for SSR marker development and linkage map construction in the tea plant (*Camellia sinensis*). *PLoS One* 8:e81611
- Tao X, Fang Y, Xiao Y, Jin YL, Ma XR, Zhao Y et al (2013) Comparative transcriptome analysis to investigate the high starch accumulation of duckweed (*Landoltia punctata*) under nutrient starvation. *Biotechnol Biofuels* 6:72
- Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M et al (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res* 40:e115
- Van Hoeck A, Horemans N, Monsieurs P, Cao HX, Vandenhove H, Blust R (2015) The first draft genome of the aquatic model plant *Lemna minor* opens the route for future stress physiology research and biotechnological applications. *Biotechnol Biofuels* 8:188
- Van Hoeck A, Horemans N, Nauts R, Van Hees M, Vandenhove H, Blust R (2017) *Lemna minor* plants chronically exposed to ionising radiation: RNA-seq analysis indicates a dose rate dependent shift from acclimation to survival strategies. *Plant Sci* 257:84–95
- Varshney RK, Sigmund R, Börner A, Korzun V, Stein N, Sorrells ME et al (2005) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci* 168:195–202
- Wang W, Messing J (2015) Status of duckweed genomics and transcriptomics. *Plant Biol* 17:10–15
- Wang JY, Chen YY, Liu WL, Wu YT (2008) Development and application of EST-derived SSR markers for bananas (*Musa nana* Lour.). *Yi Chuan* 30:933–940
- Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L et al (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). *BMC Genom* 11:726
- Wang C, Chen S, Yu S (2011) Functional markers developed from multiple loci in GS3 for fine marker-assisted selection of grain length in rice. *Theor Appl Genet* 122:905–913
- Wang W, Haberer G, Gundlach H, Glasser C, Nussbaumer T, Luo MC et al (2014a) The *Spirodela polyrrhiza* genome reveals insights into its neotenus reduction fast growth and aquatic lifestyle. *Nat Commun* 5:3311
- Wang W, Wu Y, Messing J (2014b) RNA-Seq transcriptome analysis of *Spirodela* dormancy without reproduction. *BMC Genom* 15:60
- Wang L, Yu H, Li Q (2018) Identification and characterization of 23 microsatellite loci for *Chlorostoma rustica* based on RAD-seq. *J Genet* 97:e61–e65
- Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D et al (2011) Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genom* 12:451
- Yu C, Zhao X, Qi G, Bai Z, Wang Y, Wang S et al (2017) Integrated analysis of transcriptome and metabolites reveals an essential role of metabolic flux in starch accumulation under nitrogen starvation in duckweed. *Biotechnol Biofuels* 10:167
- Zhang Q, Li X, Yang Y, Fan S, Ai J (2016) Analysis on SSR information in transcriptome and development of molecular markers in *Lonicera caerulea*. *Acta Horti Sin* 43:557–563

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.