Phylogenic study of Lemnoideae (duckweeds) through complete chloroplast genomes for eight accessions

Yanqiang Ding[1,2,3,4], Yang Fang[1,4], Ling Guo[1,3], Zhidan Li[1,3], Kaize He[1], Yun Zhao[2] and Hai Zhao[1]

[1] Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, China
[2] Key Laboratory of Bio-resource and Eco-environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, China
[3] University of Chinese Academy of Sciences, Beijing, China
[4] Key Laboratory of Environment and Applied Microbiology, Chinese Academy of Sciences, Chengdu, China

ABSTRACT

Background. Phylogenetic relationship within different genera of Lemnoideae, a kind of small aquatic monocotyledonous plants, was not well resolved, using either morphological characters or traditional markers. Given that rich genetic information in chloroplast genome makes them particularly useful for phylogenetic studies, we used chloroplast genomes to clarify the phylogeny within Lemnoideae.

Methods. DNAs were sequenced with next-generation sequencing. The duckweeds chloroplast genomes were indirectly filtered from the total DNA data, or directly obtained from chloroplast DNA data. To test the reliability of assembling the chloroplast genome based on the filtration of the total DNA, two methods were used to assemble the chloroplast genome of *Landoltia punctata* strain ZH0202. A phylogenetic tree was built on the basis of the whole chloroplast genome sequences using MrBayes v.3.2.6 and PhyML 3.0.

Results. Eight complete duckweeds chloroplast genomes were assembled, with lengths ranging from 165,775 bp to 171,152 bp, and each contains 80 protein-coding sequences, four rRNAs, 30 tRNAs and two pseudogenes. The identity of *L. punctata* strain ZH0202 chloroplast genomes assembled through two methods was 100%, and their sequences and lengths were completely identical. The chloroplast genome comparison demonstrated that the differences in chloroplast genome sizes among the Lemnoideae primarily resulted from variation in non-coding regions, especially from repeat sequence variation. The phylogenetic analysis demonstrated that the different genera of Lemnoideae are derived from each other in the following order: *Spirodela*, *Landoltia*, *Lemna*, *Wolffiella*, and *Wolffia*.

Discussion. This study demonstrates potential of whole chloroplast genome DNA as an effective option for phylogenetic studies of Lemnoideae. It also showed the possibility of using chloroplast DNA data to elucidate those phylogenies which were not yet solved well by traditional methods even in plants other than duckweeds.

Subjects Bioinformatics, Evolutionary Studies, Genomics, Plant Science

Keywords Chloroplast genome, Lemnoideae, Phylogeny, Assembly

Submitted 1 March 2017
Accepted 2 December 2017
Published 22 December 2017

Corresponding authors
Yun Zhao, zhaoyun@scu.edu.cn
Hai Zhao, zhaohai@cib.ac.cn

Academic editor
Marcial Escudero

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj.4186**

Copyright
2017 Ding et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Three kinds of genomes with different evolutionary origins and histories coexist in plant cells: nuclear, chloroplastic and mitochondrial. Generally, mitochondrial genomes are not the best choice for phylogenetic studies in plants, because their rate of rearrangements is extraordinarily fast compared to chloroplast (cp) genomes (cpDNAs) (*Palmer & Herbon, 1988*). Meanwhile, phylogenetic studies using nuclear genomes are restricted by their complex and infeasibility of enough data (*Olsen et al., 2016*; *Wang et al., 2014*). A single, independent genealogical history can be readily obtained from cpDNAs (*Kim et al., 2015*; *Whittall et al., 2010*), since their inheritance differs from that of nuclear genomes, such as vegetative segregation, uniparental inheritance, haploid status, and general absence of recombination (*Birky Jr, 2001*; *Hansen et al., 2007*; *Petit et al., 2003*). Therefore, cpDNAs are particularly useful for phylogenetic and phylogeographic studies.

However, phylogenetic and phylogeographic analyses based on cpDNAs are typically limited by DNA sequencing costs and genomes assembling methods (*Parks, Cronn & Liston, 2009*). In previous studies, the cpDNAs were directly obtained by primer walking (based on closely related cpDNAs) and by shotgun sequencing (*Jansen et al., 2005*; *Shi et al., 2012*; *Whittall et al., 2010*). Recently, owing to the lower costs of next-generation sequencing (NGS), a new cost-effective method has arisen: indirectly assembling complete cpDNA by filtering from total DNA data (including DNA data of nuclei, cps and mitochondria) (*Wang & Messing, 2011*; *Zhang et al., 2011*). However, few studies had compared the chloroplast genomes obtained from this method with those obtained from primer walking or shotgun sequencing.

Lemnoideae (duckweeds), a kind of small aquatic flowering monocotyledonous plants, has got increasingly more attention due to its asexual reproduction, rapid propagation and potential values in eutrophic wastewater treatment, starch production and bioenergy transformation (*Zhao et al., 2012*). Lemnoideae is a subfamily of the Araceae family and includes five genera and 37 species (*Sree, Bog & Appenroth, 2016*). Traditionally, morphological characters are inspected for identifying the taxonomy of the species and phylogenetic studies. The phylogenetic relationship within the different genera of Lemnoideae has not been well resolved mainly because of their small sizes and highly morphological degeneration. In addition, the confidence values of phylogenetic trees were not strongly supported when using traditional markers (*Rothwell et al., 2004*). Given that neither morphological characters nor traditional markers are sufficient for phylogenetic study within Lemnoideae, we used the whole cpDNAs obtained from NGS to clarify the phylogenetic relationships within this group.

In this paper, we compared two different cpDNA extraction and assembly methods and verified that assembling the cpDNA based on the filtration of the total DNA was reliable. Then we built a phylogenetic tree on the basis of the whole cpDNA sequences which clarified the phylogenetic relationship within different genera of Lemnoideae. This study can help to resolve the phylogeny of Lemnoideae. Meanwhile, it highlights that the whole

cpDNA is a feasible and effective option for phylogenetic studies, and demonstrates the possibilities that the NGS can elucidate those phylogenies which traditionally are not well solved.

MATERIALS & METHODS

Duckweeds strains

Eight duckweeds strains were used in this study. They were *Landoltia punctata* strain ZH0202, *Landoltia punctata* strain 0086, *Landoltia punctata* strain 0062, *Lemna minor* strain 9532, *Lemna gibba* strain 9584, *Lemna japonica* strain 0234, *Lemna japonica* strain 8695, and *Wolffia australiana* strain 7317, respectively. All the strains were stored and cultured in the Chengdu Institute of Biology, Chinese Academy of Sciences (Chengdu, China) under controlled conditions in the greenhouse. Four announced duckweeds cpDNAs data (*Spirodela polyrhiza* strain 7498, *Lemna minor* Renner 2188, *Wolffiella lingulata* strain 7289, and *Wolffia australiana* strain 7733) were also included in the study (*Mardanov et al., 2008*; *Wang & Messing, 2011*).

DNA extraction and sequencing

The cp DNA of *L. punctata* strain ZH0202 and *L. japonica* strain 8695 was isolated from 1 g of tissue of young duckweeds produced from a single mother, by using a Plant Cp DNA Isolation Kit (Genmed Scientific Inc., Arlington, USA). The DNA concentration and purity were checked with NanoDrop 2000c. Paired-end (PE) libraries with a 300-bp insert size were constructed and sequenced with Illumina HiSeq 2500 platform by the Beijing Genomics Institute (Shenzhen, China).

The total DNA of *L. punctata* strain ZH0202, *L. punctata* strain 0086, *L. punctata* strain 0062, *L. minor* strain 9532, *L. gibba* strain 9584, *L. japonica* strain 0234, and *W. australiana* strain 7317 was extracted by using the CTAB method (*Jansen et al., 2005*). The DNA concentration and purity were also checked with NanoDrop 2000c. PE libraries with 500-bp insert size were constructed and sequenced with Illumina HiSeq 2000 platform by the Beijing Genomics Institute (Shenzhen, China).

CpDNA assembly and annotation

When using the total DNA, we assembled the cpDNAs as follows (Fig. 1): (1) filtering of the data using FastQC 0.11.3 with the default parameters; (2) pre-assembly using SOAPdenovo 2 with diverse K-mer values (23–89) (*Luo et al., 2012*); (3) isolation of the cp contigs; (4) addition of the inverted repeats (IRs, including IRa and IRb) sequence; (5) extension of the contigs with SSPACE 2.02 (*Boetzer et al., 2011*); (6) closing of the gaps with GapCloser 1.12 (*Kim et al., 2015*); (7) mapping of the reads to the draft genome by using SOAP 2.21 to identify and correct any errors (*Zhang et al., 2012*). When isolating the cp contigs (step (3)), the best assembly result (K-mer 87) was first aligned with the most closely related reference genome (*S. polyrhiza* strain 7498, GenBank Accession: JN160603; *L. minor*, GenBank Accession: DQ400350; *W. australiana* strain 7733, GenBank Accession: JN160605) using BLAST to identify contigs with high identity (>95%, *e*-value 10^{-5}) (*McGinnis & Madden, 2004*). Then, the read depth of those high-identity contigs was calculated using SOAP 2.21

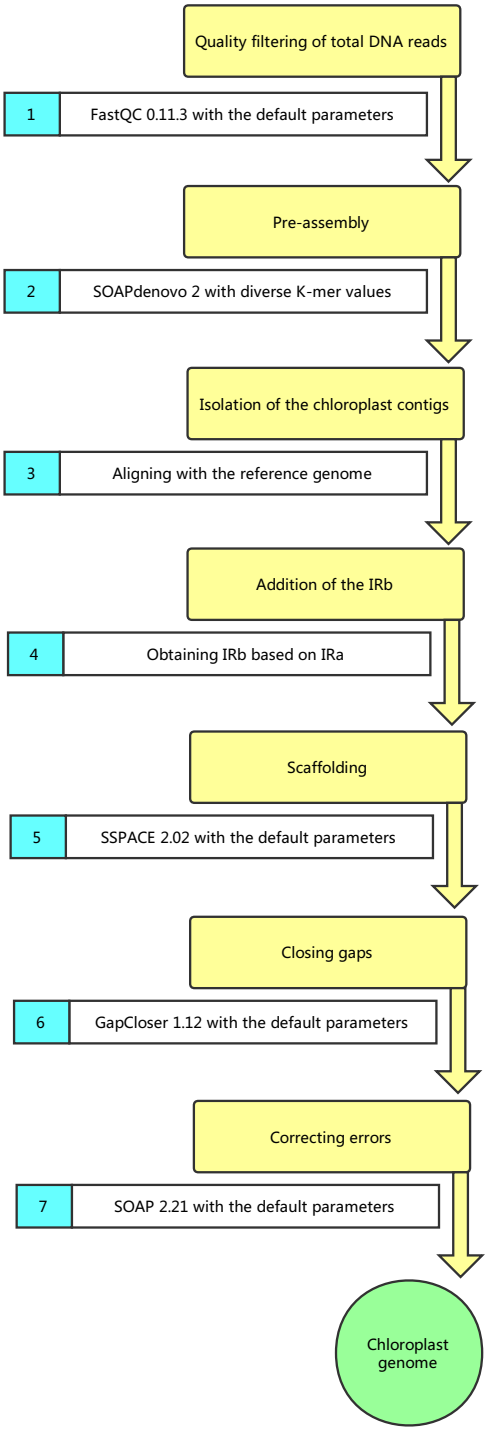

Figure 1 **Pipeline of chloroplast genome assembly.**

Full-size DOI: 10.7717/peerj.4186/fig-1

and SOAPcoverage 2.7.7 (*Li et al., 2009*). The cp contigs with high coverage (more than 1,000×) were isolated. Subsequently, the contigs were aligned with the reference genome by using MAUVE 2.3.1 to ensure that they belonged to the cpDNA (*Bennett & Triemer, 2015*). Meanwhile, when using cp DNA, all the steps were applied to assemble the cpDNAs, except the isolation of cp contigs (step(3)).

CpDNAs were annotated using DOGMA with default parameters (*Wyman, Jansen & Boore, 2004*). The tRNA genes were further identified using tRNAscan-SE under default parameters (*Sedlar et al., 2015*). The single sequence repeats (SSRs) and tandem repeats in the cpDNAs of 12 strains of duckweeds were detected using Phobos 3.3.12 with the default parameters, except that the maximum unit length was set as 30 (*Raman & Park, 2015*). Map of the circular plastome was drawn with OGDraw 1.2 (*Lohse et al., 2013*).

Sequence polymorphisms of duckweeds cpDNAs

DnaSP v5 was used to calculate the DNA polymorphism (*Librado & Rozas, 2009*). To study the sequence divergence in the whole genome level, whole-genome alignments were carried out using mVISTA with "LAGAN" alignment program (Global multiple alignment of finished sequences) (*Kim et al., 2015*). We also compared the large single copy (LSC)/IRa/ small single copy (SSC) /IRb boundary regions of the duckweeds cpDNAs to study IRs contraction or expansion.

Phylogeny of Lemnoideae based on whole cpDNAs

ClustalW was used to align the cpDNAs sequences under default parameters (*Larkin et al., 2007*), and the alignment was checked manually. The Bayesian Inference (BI) and Maximum-likelihood (ML) methods were performed for the genome-wide phylogenetic analyses using MrBayes v.3.2.6 (*Huelsenbeck & Ronquist, 2001*; *Ronquist & Huelsenbeck, 2003*) and PhyML 3.0 (*Guindon et al., 2010*), respectively. Nucleotide substitution model selection was estimated with jModelTest 2.1.10 (*Darriba et al., 2012*) and Smart Model Selection in PhyML 3.0. The model GTR + G was selected as the best-fitting model for both BI and ML analyses. Bayesian Inference partitioning analysis followed the programs with calculating a majority rule consensus tree with 1×10^7 generations of Markov chain Monte Carlo (MCMC), with frequency of tree sampling every 1,000 generations and the first 2,500 trees discarding as burn-in, and starting from a random tree. After performing two independent runs, the output trees were combined to estimate the Bayesian posterior probabilities (BPP) in 50% majority rule for each node. For ML analysis, PhyML 3.0 was performed with 1,000 bootstrap replicates to calculate the bootstrap values (BS) of the topology. In addition, the significant nodes' supports were considered with 95% BPP and 75% BS (*Hillis & Bull, 1993*) in BI and ML analysis, respectively. The results were treated with iTOL 3.4.3 (*Letunic & Bork, 2016*). *Colocasia esculenta* share the same family as the members of Lemnoideae and was included as an outgroup (*Luo et al., 2016*).

Table 1 Duckweed cp genomes assembly results.

Assembly method	Sample name	Strain	Latin name	CGS (bp)	IRs (bp)	LSC (bp)	SSC (bp)	GC content (%)	Collection area	GenBank accession number
ECD	ZH0051	ZH0202	*L. punctata*	171,013	31,899	92,742	14,473	35.46	Xinjin, China	KY993962
	D0101	8695	*L. japonica*	166,424	31,571	89,277	14,005	35.74	Kyoto, Japan	KY993955
FTD	ZH0051	ZH0202	*L. punctata*	171,013	31,899	92,742	14,473	35.46	Xinjin, China	KY993962
	ZH0086	0086	*L. punctata*	170,994	31,900	92,721	14,473	35.49	Leshan, China	KY993960
	ZH0062	0062	*L. punctata*	171,152	31,894	92,726	14,635	35.44	Yaan, China	KY993959
	D0107	9532	*L. minor*	165,775	31,218	89,735	13,604	35.75	Ohrid, Macedonia	KY993956
	D0289	9584	*L. gibba*	166,553	31,763	89,408	13,619	35.73	Perebel River, Poland	KY993957
	ZH0234	0234	*L. japonica*	165,436	31,468	88,635	13,866	35.74	Kunming, China	KY993961
	M170	7317	*W. australiana*	168,270	31,990	90,871	13,419	35.86	Australia	KY993958

Notes.
CGS, Cp genome size; IRs, Inverted repeats; LSC, Large single copy; SSC, Small single copy; ECD, Extraction of the cp DNA; FTD, Filtering of the total DNA..

RESULTS

Reliability of assembly of the cpDNA on the basis of the total DNA

To test the reliability of assembling the cpDNA based on the filtration of the total DNA, the cpDNA of the same strain (*L. punctata* strain ZH0202) was also assembled by using cp DNA directly. As a result, the identity of *L. punctata* strain ZH0202 cpDNAs assembled through the two methods was 100%, and the sequence and length were completely identical. This experiment revealed no nucleotide variability between the two cpDNA assemblies of *L. punctata* strain ZH0202 (Table 1).

Assembly and annotation of Lemnoideae chloroplast genomes

The Illumina HiSeq system yielded 59 Gb of total clean data (Table S1). After the data were filtered, assembled and validated, eight complete duckweeds cpDNAs were obtained with coverage of over 1,000×. Together with four reported duckweeds cpDNAs, all of the duckweeds cpDNAs were within a range of 165,775 bp to 171,152 bp in length (Table 1), and carried two copies of IRs separated by a SSC and LSC (Fig. 2). Their lengths were variable: IR (312,18 bp–31,990 bp); SSC (13,392 bp–14,635 bp) (*Wang & Messing, 2011*); and LSC (88,635 bp–92,742 bp) (Table 1, Table S2).

Each of the duckweeds cpDNAs contained 80 protein-coding sequences (CDS), four rRNAs, 30 tRNAs and two pseudogenes, including two new annotated genes in duckweeds: *orf* 42 and *orf* 56 (Fig. 2, Table S3). All of the duckweeds cpDNAs had the same gene content and order.

We annotated the repeat sequences and compared their type, length and number among the cpDNAs of 12 strains of duckweeds (Table 2). The total length of repeat sequences was in a range of 10,004 bp (*L. japonica* strain 0234) to 12,832 bp (*L. punctata* strain 0062), and the percentages in cpDNAs were in a range of 6.03% (*W. australiana* strain 7733) to 7.59% (*S. polyrhiza* strain 7498) (Table 2). In all 12 cpDNAs of duckweeds,

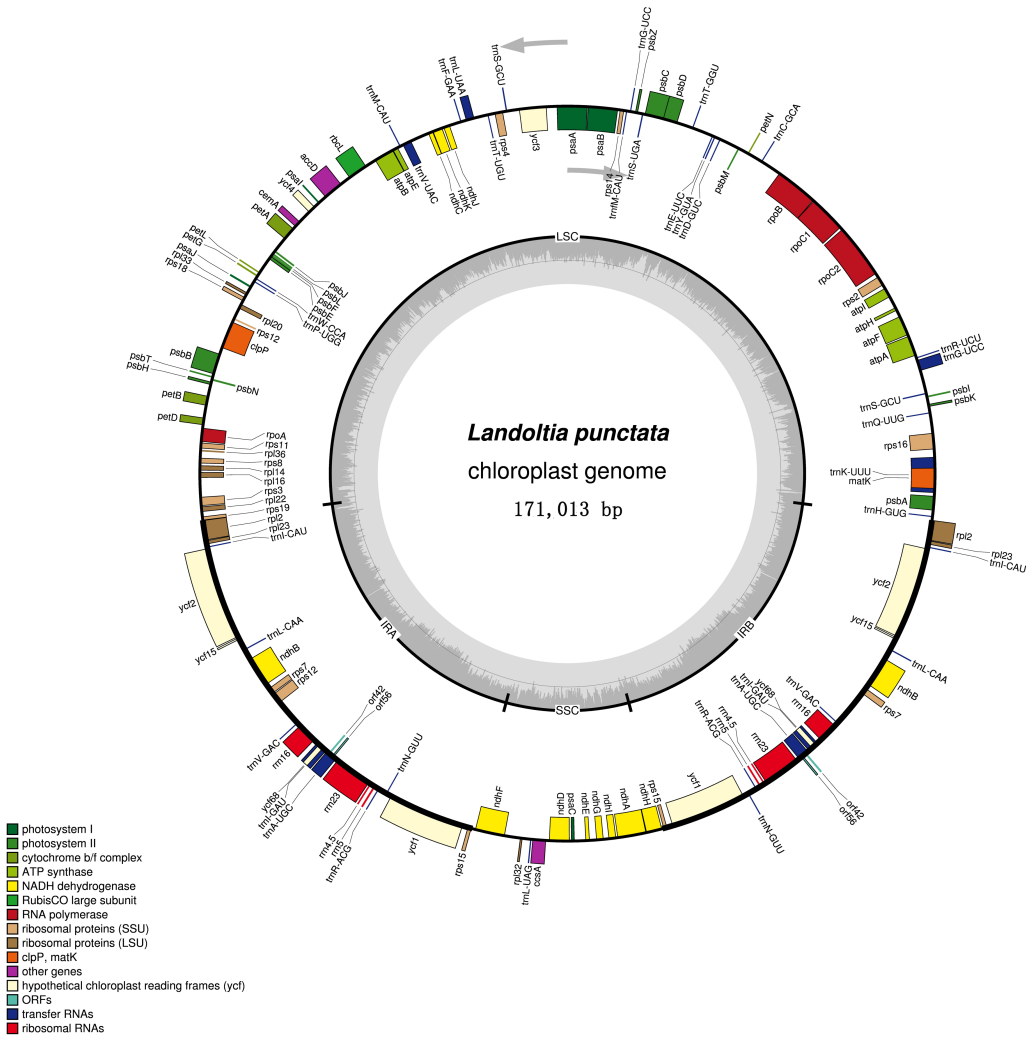

Figure 2 Chloroplast genome of *Landoltia punctata* strain ZH0202. The outer circle shows positions of genes in the large single copy (LSC), small single copy (SSC), and two inverted repeat (IRa and IRb) regions. The inner circle is a graph depicting GC content across the genome. Plastome maps were generated in OGDraw 1.2.

Full-size DOI: 10.7717/peerj.4186/fig-2

homopolymers were the most frequent, followed by hexa-, penta-, and tetrapolymers. However, there were relatively more pentapolymers (approximately 180) compared to hexapolymers (152) in *L. punctata* (Table 2). Interestingly, the longest tandem repeats (AAAAATATATAATAATATTAATAAAAT × 2) in the known duckweeds cpDNAs were found in *L. japonica* strain 0234, which had the shortest total length of repeat sequences and the smallest total number of repeat sequences among the duckweeds (Table 2).

Sequence polymorphisms of duckweeds cpDNAs

A total of 17,438 polymorphic sites were found among duckweeds cpDNAs by using ClustalW and DnaSP. Most of the divergent sequences were in intergenic regions, whereas some were in introns (Fig. S1). For instance, *L. punctata* strain ZH0202 has a 564-bp

Table 2 The type, length and number of repeat sequence in the cp genomes of 12 strains of duckweed.

	S. polyrhiz a strain 7498[a]	*L. punctata* strain ZH0202	*L. punctata* strain 0086	*L. punctata* strain 0062	*L. japonica* strain 8695	*L. minor* strain 9532	*L. minor* Renner2188[a]	*L. gibba* strain 9584	*L. japonica* strain 0234	*W. lingulata* strain 7289[a]	*W. australiana* strain 7317	*W. australiana* strain 7733[a]
Repeats sequence (bp)	12,810	12,810	12,719	12,832	11,119	10,966	11,085	11,222	10,004	11,558	10,424	10,178
Percent(%)	7.59	7.49	7.44	7.50	6.68	6.61	6.68	6.74	6.05	6.83	6.19	6.03
Mono-	335	373	373	373	356	355	355	357	355	354	363	359
Di-	64	71	74	74	63	62	63	63	63	71	70	71
Tri-	83	70	69	69	69	69	69	69	65	74	59	60
Tetra-	123	121	121	121	106	104	105	105	101	112	115	117
Penta-	178	180	178	180	125	122	124	125	117	138	123	124
Hexa-	183	152	152	152	155	155	155	157	136	161	154	155
7	45	38	38	38	40	38	38	37	32	35	23	24
8	21	22	22	22	22	23	23	23	15	20	15	14
9	21	15	15	15	22	22	22	23	17	19	17	15
10	9	6	6	6	6	5	6	6	7	10	7	8
11	3	2	2	2	3	3	3	3	2	3	1	1
12	1	2	2	2	4	3	4	4	1	2	0	0
13	1	7	6	7	0	0	0	1	1	1	1	1
14	1	4	4	4	2	3	3	3	1	2	0	0
15	2	4	3	3	2	1	1	1	2	3	1	1
16	0	1	1	1	3	3	3	3	0	0	0	0
17	1	3	3	3	1	1	1	1	0	0	0	0
18	1	1	1	1	1	2	2	2	1	0	0	0
19	3	1	1	1	1	1	1	1	3	0	1	0
20	0	2	2	2	1	1	1	1	0	0	0	0
21	0	0	1	1	0	0	0	0	1	1	0	0
22	2	2	2	2	0	0	0	0	0	0	0	0
23	1	1	1	1	1	1	1	1	0	1	0	0
24	0	4	4	4	1	1	1	1	0	2	2	0
25	1	1	1	1	1	1	1	0	0	0	1	0
26	0	0	0	0	0	0	0	1	1	0	1	0
27	0	0	0	0	0	0	0	0	1	0	0	0

Notes.

[a]Duckweed cp genomes from previous studies.

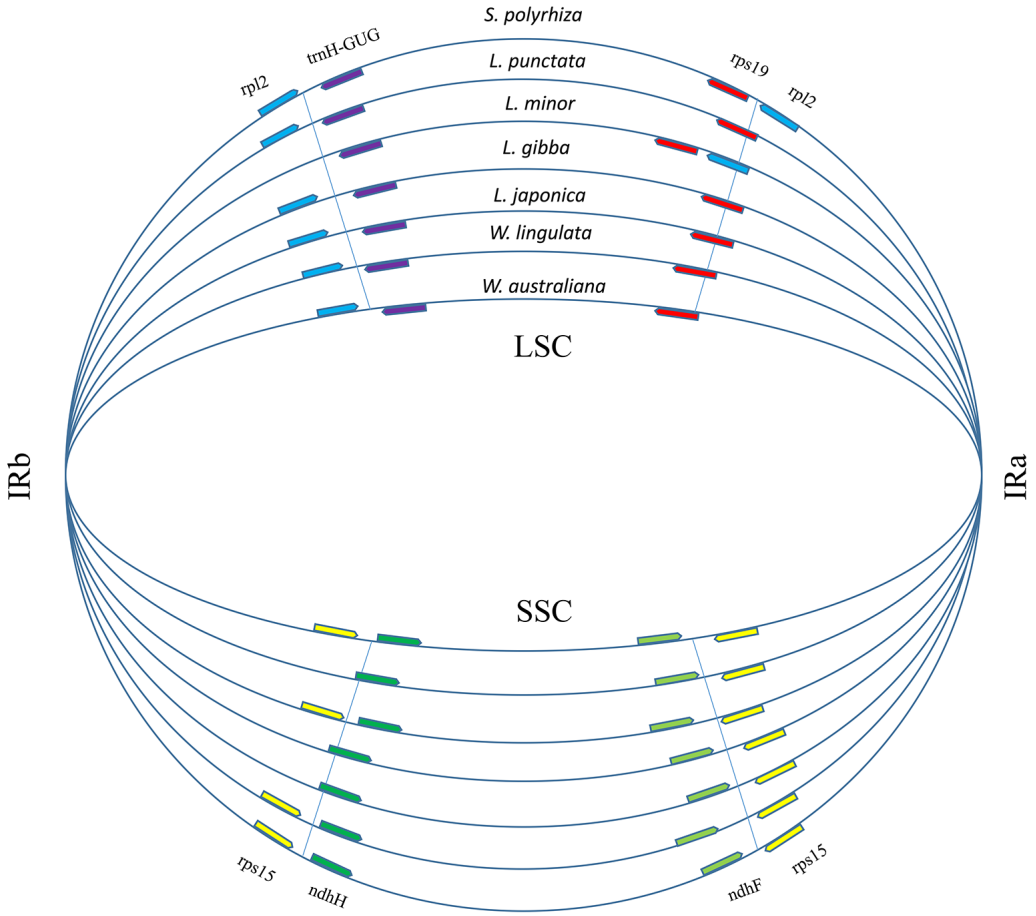

Figure 3 Boundary gene-flow and IR region expansion/contraction events. Comparison of the junction positions of IR boundaries among 7 duckweed cp genomes. The genes near the junction positions are identified by color: red, rps19; light blue, rpl2; yellow, rps15; light green, ndhF; green, ndhH; purple, trnH-GUG.

Full-size DOI: 10.7717/peerj.4186/fig-3

insertion in the 32-kb *petN-psbM* region and an 88-bp insertion in the intron of *atp*F. Five strains of *Lemna* all had approximately 300-bp deletions at the 105-kb regions (Figs. S1; S2). In comparison to intergenic regions, the sequence divergence frequency for regions of coding genes was low. IRs were more conserved than LSC and SSC. Although there was less sequence divergence in IRs, more sequence polymorphisms appeared at the junctions of IRs and LSC/SSC (Fig. S1). We furthermore found that the locations of some genes in the LSC/IRa/SSC/IRb boundary regions were different (Fig. 3), although the gene content and order were the same. The most comprehensive variation was found in the boundary of the LSC and IRa regions, where the *rps*19 sequence completely shifted position towards the LSC region in *L. minor* and *S. polyrhiza*. Additionally, a 385 bp of the *rpl*2 sequence relocated from the IRa toward the LSC region. Furthermore, the *rpl*2 gene at the end of the IRb region was incomplete in *L. minor*, thus annotated as a pseudogene (Fig. 3).

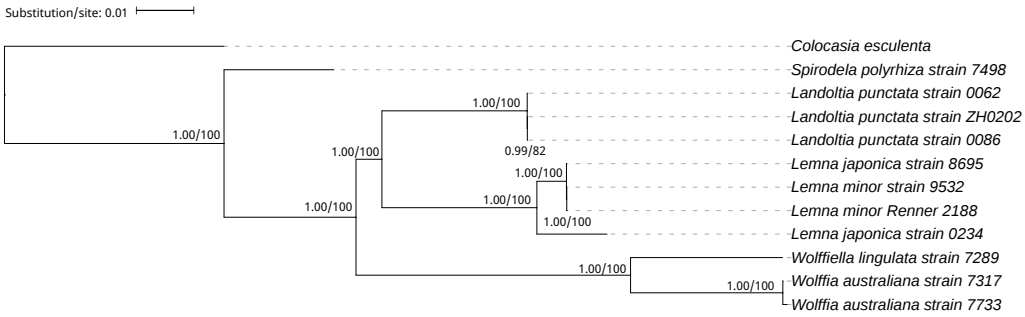

Figure 4 The Bayesian Inference (BI) and Maximum likelihood (ML) tree on the basis of cp genome sequences. The Numbers upon each node indicate Bayesian posterior probabilities and ML Bootstrap, respectively (showed in BPP/BS).

Full-size DOI: 10.7717/peerj.4186/fig-4

Phylogeny of Lemnoideae based on whole cpDNAs

A phylogenetic tree was generated using BI and ML methods, which consistently supported the uniform topology. The topology was reliable with 1.00 BPP and 100% BS for nine out of ten nodes (Fig. 4). The phylogenetic analysis demonstrated that *Spirodela* was derived first from the lineage of the remaining members of the subfamily, then *Landoltia*, *Lemna*, *Wolffiella* and *Wolffia* (Fig. 4). Surprisingly, we found that *L. japonica* strain 8695 and *L. japonica* strain 0234 were in separate branches.

DISCUSSION

Reliability of assembling the cpDNA based on the filtration of the total DNA

For previous studies, the cpDNAs were assembled from independently extracted cp DNA (*Jansen et al., 2005*; *Shinozaki et al., 1986*). Assembling the cpDNA based on indirect filtration of the total DNA was first described in 2011 (*Zhang et al., 2011*). To our knowledge, no study had compared these two assembly approaches until now. In this study, we assembled the cpDNA of *L. punctata* strain ZH0202 with the two methods simultaneously. The sequence and length of *L. punctata* strain ZH0202 cpDNAs assembled through the two methods were completely identical (Table 1). This result verified that indirectly assembling the cpDNA based on the filtration of the total DNA is reliable. Directly assembling cpDNAs from cp still has advantages especially when the reference genomes are unavailable (*Jansen et al., 2005*). Meanwhile, the sequencing costs decrease every year, indirectly assembling cpDNAs from the total DNA becomes more and more attractive (*Wang & Messing, 2011*).

Two genes in duckweeds cpDNAs: *orf*42 and *orf*56

Here, two genes were annotated: *orf* 42 and *orf* 56, which were not found in previous studies of duckweeds cpDNAs (*Mardanov et al., 2008*; *Wang & Messing, 2011*). These two genes are located 200 bp apart in the intron of *trn*A-UGC. Their sequences are conserved,

and it has previously been reported that they are related to mitochondrial genes (*Do, Kim & Kim, 2013*). However, functions of *orf* 42 and *orf* 56 are still unknown (*Bodin, Kim & Kim, 2013*).

Differences in cpDNA sizes among the genera of Lemnoideae

The known cpDNA sizes are conservative within some subfamilies of higher plants. For instance, all the members of the Maloideae have a similar cpDNA size, ranging from 15,9161 bp in *Pydus spinosa* to 16,0041 bp in *Malus pdunifolia* voucher MPRUN20160302 (*Korotkova et al., 2014*). However, the cpDNA sizes are more variable within Lemnoideae, ranging from 166 kb in *Lemna* to 171 kb in *Landoltia*. In the former (*Lemna*), the IR size was the smallest among the known duckweeds cpDNAs, indicating the IRs have contracted in this species (Fig. 3). While in the latter (*Landoltia*), it was found that the lengths of the IRs, LSC and SSC regions were longer than those of the other genera except IRs of *Wolffia*. Our results contrast with a previous study carried out on species of *Gossypium* genus, in which most of the cp size differences reflected indels in the LSC region (*Chen et al., 2017*). The results of this study supported the previous finding that changes in the length of the IR can account for the size variation among plant cpDNAs (*Jansen et al., 2005*; *Wakasugi, Tsudzuki & Sugiura, 2001*).

We furthermore found the differences in the cpDNA sizes among the genera of Lemnoideae resulted primarily from variation in the non-coding regions, while the lengths of coding regions were almost the same. The results of the whole cpDNA alignments also supported this (Fig. S1). This finding was consistent with Zheng's study in seed plants, in which their results pointing to intergenic regions having a great role in the variations in chloroplast genome size among closely related species (*Zheng et al., 2017*).

In addition, variation in the repeat sequences were found to partially account for the difference in cpDNA sizes. For example, *Landoltia* had 12.8 kb of repeat sequences, which were approximately 2.2 kb longer than those of *Lemna*. Moreover, 5.0 kb of sequence in *L. minor* strain 9532 did not exist in the cpDNA of *L. punctata* strain ZH0202, 25.57% of which comprised repeat sequences. This percentage was greater than the average percentage (7.49%) of repeat sequences in *L. punctata* strain ZH0202. Similar as above, 1.9 kb of sequence in *W. australiana* strain 7317 did not exist in the cpDNA of *Wolffiella lingulata* strain 7289, 38.10% of which consisted of repeat sequences. This percentage was greater than the average percentage (6.83%) of repeat sequences in *W. lingulata* strain 7289. There were more repeat sequences in the expanded portion, meaning that repeat sequences influenced the expansion or contraction of the cpDNA and partially led to the difference in cpDNA sizes. Our results were consistent with Wu's study in rice, which indicates the main source of cp length variation is coming from mononucleotide SSRs (*Wu et al., 2017*). The variation in cpDNA size may influence energy generation and ecological strategy (*Zheng et al., 2017*) and provide more information for phylogenic study.

Phylogeny of lemnoideae

It was surprising that *L. japonica* strain 8695 and *L. japonica* strain 0234 are in separate branches (Fig. 4). However, similar cases were found in the phylogenetic study of apples:

Malus sieversii were scattered across branches containing other wild species, which justified its splitting into at least two species (*Nikiforova et al., 2013*). In our study, *L. japonica* strain 8695 and *L. japonica* strain 0234 were scattered across branches containing *L.minor*, indicating that genetic diversity of cpDNAs within *L. japonica* exceeds that between other species. In addition, they have phenotypic differences: *L. japonica* strain 0234 cultured in Kunming have air spaces on the back of the fronds, whereas the air spaces were not found on the fronds of *L. japonica* strain 8695 (Fig. S3). Moreover, their collection areas are far apart (Table 1). Considering all these evidences, we think the most probable reasons of why two strains of *L. japonica* are in separate branches may be a wrong identification or they are two different species. Further research would be necessary to elucidate the cause of why these two strains are present in separate branches.

On the basis of the duckweeds cpDNAs, we confirmed that the evolutionary branching order of Lemnoideae was as follows: *Spirodela*, *Landoltia*, *Lemna*, *Wolffiella*, *Wolffia* (Fig. 4). The evolutionary divergence of *Landoltia* came after that of *Spirodela* and before that of *Lemna*. These results were consistent with a previous finding by *Les et al. (2002)*. In this study, authors applied more than 4,700 characters including data on morphology and anatomy, flavonoids, allozymes, and DNA sequences from cp genes (*rbc*L, *mat*K) and introns (*trn*K, *rpl*16). In our study, higher bootstrap values were obtained that support the evolutionary order of Lemnoideae (Fig. 4). However, our results were different from those of *Rothwell et al. (2004)*. In this study, authors applied the cp *trn*L - *trn* F intergenic spacer and found the evolutionary order of Lemnoideae was *Spirodela*, *Landoltia*, *Wolffia*, *Wolffiella*, and *Lemna*. Our results were also different from Lidia I. Cabrera's study. In this study, authors applied coding regions (*rbc*L, *mat*K) and non-coding plastid DNA (partial *trn*K intron, *trn*L intron, *trn*L - *trn*F spacer), and found the evolutionary order of Lemnoideae was *Spirodela*, *Lemna*, *Landoltia*, *Wolffia*, and *Wolffiella* (*Cabrera et al., 2008*). It was suggested that the evolutionary studies based on whole cpDNAs were more reliable than those based on several cp coding and/or non-coding regions, because the whole cpDNAs contain more rich genetic information (*Hansen et al., 2007*; *Sree, Bog & Appenroth, 2016*). Our finding supported Matthew Parks' study (*Parks, Cronn & Liston, 2009*), in which they compared the resolution obtained when using the cpDNAs and two cp markers, and found the increase in phylogenetic resolution is primarily due to the increase in data matrix length. Therefore, our results indicated that the whole cpDNA is a feasible and effective option for phylogenetic studies, especially for inferring phylogenies at low taxonomic levels (*Parks, Cronn & Liston, 2009*; *Whittall et al., 2010*).

CONCLUSION

We indicated that assembly of the cp genome based on the filtration of the total DNA was reliable. Our study suggested that the whole cpDNA was appropriate for phylogenetic studies, especially for inferring phylogenies at low taxonomic levels, and it showed the possibilities that the NGS can offer to elucidate those phylogenies that traditionally have not been well solved. In this study, we demonstrated the evolutionary order of Lemnoideae was as follows: *Spirodela*, *Landoltia*, *Lemna*, *Wolffiella*, *Wolffia*.

ACKNOWLEDGEMENTS

We thank the two reviewers and the editors for their insightful suggestions and comments on the paper. We thank Xiang Tao, Li Tan, Yanling Jin, Yang Liu, Mengjun Huang, Weiliang Shen at Chengdu Institute of Biology for their assistance in the laboratory work. We also thank Kristian Barrett and Zhuolin Yi for advice and comments for writing of the article.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was funded by the National Key Technology R&D Program of China (No. 2015BAD15B01); Projects of International Cooperation of the Ministry of Science and Technology of China (No. 2014DFA30680); Science and Technology Service Network Initiative (No. KFJ-EW-STS-121); Science & Technology Program of Sichuan Province (No. 2016SZ0070; No.2017NZ0018; No.2017HH0077); Funds for Advanced Manufacturing Innovation Education of De yang, Chinese Academy of Sciences (No. YC-2015-QC01); Key Laboratory of Environmental and Applied Microbiology, Chengdu Institute of Biology, Chinese Academy of Sciences (No. KLEAMCAS201501; No. KLCAS-2014-02); Environmental Protection Program of Yunnan Province (2014BI008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Key Technology R&D Program of China: 2015BAD15B01.
Ministry of Science and Technology of China: 2014DFA30680.
Science and Technology Service Network Initiative: KFJ-EW-STS-121.
Science & Technology Program of Sichuan Province: 2016SZ0070, 2017NZ0018, 2017HH0077.
Chinese Academy of Sciences: YC-2015-QC01.
Key Laboratory of Environmental and Applied Microbiology, Chengdu Institute of Biology. Chengdu Institute of Biology, Chinese Academy of Sciences: KLEAMCAS201501, KLCAS-2014-02.
Environmental Protection Program of Yunnan Province: 2014BI008.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Yanqiang Ding conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

- Yang Fang conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Ling Guo performed the experiments, contributed reagents/materials/analysis tools, prepared figures and/or tables, reviewed drafts of the paper.
- Zhidan Li performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Kaize He conceived and designed the experiments, wrote the paper, reviewed drafts of the paper.
- Yun Zhao and Hai Zhao conceived and designed the experiments, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

GenBank BioProject Accession: PRJNA374932; PRJNA374937; PRJNA374938; PRJNA374939; PRJNA374941; PRJNA374942; PRJNA374943; PRJNA374944.

Data Availability

The following information was supplied regarding data availability:

The raw data has been provided as Supplemental File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.4186#supplemental-information.

REFERENCES

Bennett MS, Triemer RE. 2015. Chloroplast genome evolution in the euglenaceae. *Journal of Eukaryotic Microbiology* **62**:773–785 DOI 10.1111/jeu.12235.

Birky Jr CW. 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics* **35**:125–148 DOI 10.1146/annurev.genet.35.102401.090231.

Bodin SS, Kim JS, Kim J-H. 2013. Complete chloroplast genome of Chionographis japonica (Willd) Maxim. (Melanthiaceae): comparative genomics and evaluation of universal primers for Liliales. *Plant Molecular Biology Reporter* **31**:1407–1421 DOI 10.1007/s11105-013-0616-x.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**:578–579 DOI 10.1093/bioinformatics/btq683.

Cabrera LI, Salazar GA, Chase MW, Mayo SJ, Bogner J, Dávila P. 2008. Phylogenetic relationships of aroids and duckweeds (Araceae) inferred from coding and noncoding plastid DNA. *American Journal of Botany* **95**:1153–1165 DOI 10.3732/ajb.0800073.

Chen Z, Grover CE, Li P, Wang Y, Nie H, Zhao Y, Wang M, Liu F, Zhou Z, Wang X, Cai X, Wang K, Wendel JF, Hua J. 2017. Molecular evolution of the plastid genome

during diversification of the cotton genus. *Molecular Phylogenetics and Evolution* **112**:268–276 DOI 10.1016/j.ympev.2017.04.014.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**:772–772 DOI 10.1038/nmeth.2109.

Do HDK, Kim JS, Kim J-H. 2013. Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of Veratrum patulum O. Loes. (Melanthiaceae). *Gene* **530**:229–235 DOI 10.1016/j.gene.2013.07.100.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**:307–321 DOI 10.1093/sysbio/syq010.

Hansen AK, Escobar LK, Gilbert LE, Jansen RK. 2007. Paternal, maternal, and biparental inheritance of the chloroplast genome in Passiflora (Passifloraceae): implications for phylogenetic studies. *American Journal of Botany* **94**:42–46 DOI 10.3732/ajb.94.1.42.

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42**:182–192 DOI 10.1093/sysbio/42.2.182.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754–755 DOI 10.1093/bioinformatics/17.8.754.

Jansen RK, Raubeson LA, Boore JL, DePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ, Fourcade HM, Kuehl JV, McNeal JR, Leebens-Mack J, Cui L. 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology* **395**:348–384 DOI 10.1016/S0076-6879(05)95020-9.

Kim K, Lee SC, Lee J, Yu Y, Yang K, Choi BS, Koh HJ, Waminal NE, Choi HI, Kim NH, Jang W, Park HS, Lee J, Lee HO, Joh HJ, Lee HJ, Park JY, Perumal S, Jayakodi M, Lee YS, Kim B, Copetti D, Kim S, Kim S, Lim KB, Kim YD, Lee J, Cho KS, Park BS, Wing RA, Yang TJ. 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. *Scientific Reports* **5**:15655 DOI 10.1038/srep15655.

Korotkova N, Nauheimer L, Ter-Voskanyan H, Allgaier M, Borsch T. 2014. Variability among the most rapidly evolving plastid genomic regions is lineage-specific: implications of pairwise genome comparisons in Pyrus (Rosaceae) and other angiosperms for marker choice. *PLOS ONE* **9**:e112998 DOI 10.1371/journal.pone.0112998.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**:2947–2948 DOI 10.1093/bioinformatics/btm404.

Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT. 2002. Phylogeny and systematics of Lemnaceae, the duckweed family. *Systematic Botany* **27**:221–240 DOI 10.1043/0363-6445-27.2.221.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**:W242–W245 DOI 10.1093/nar/gkw290.

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**:1966–1967 DOI 10.1093/bioinformatics/btp336.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**:1451–1452 DOI 10.1093/bioinformatics/btp187.

Lohse M, Drechsel O, Kahlau S, Bock R. 2013. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* **41**:W575–W581 DOI 10.1093/nar/gkt289.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**:1–6 DOI 10.1186/2047-217X-1-18.

Luo Y, Ma PF, Li HT, Yang JB, Wang H, Li DZ. 2016. Plastid phylogenomic analyses resolve tofieldiaceae as the root of the early diverging monocot order alismatales. *Genome Biology and Evolution* **8**:932–945 DOI 10.1093/gbe/evv260.

Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS, Kolganova TV, Skyabin KG. 2008. Complete sequence of the duckweed (Lemna minor) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. *Journal of Molecular Evolution* **66**:555–564 DOI 10.1007/s00239-008-9091-7.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**:W20–W25 DOI 10.1093/nar/gkh435.

Nikiforova SV, Cavalieri D, Velasco R, Goremykin V. 2013. Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. *Molecular Biology and Evolution* **30**:1751–1760 DOI 10.1093/molbev/mst092.

Olsen JL, Rouze P, Verhelst B, Lin Y-C, Bayer T, Collen J, Dattolo E, De Paoli E, Dittami S, Maumus F, Michel G, Kersting A, Lauritano C, Lohaus R, Topel M, Tonon T, Vanneste K, Amirebrahimi M, Brakel J, Bostrom C, Chovatia M, Grimwood J, Jenkins JW, Jueterbock A, Mraz A, Stam WT, Tice H, Bornberg-Bauer E, Green PJ, Pearson GA, Procaccini G, Duarte CM, Schmutz J, Reusch TBH, Van de Peer Y. 2016. The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature* **530**:331–335 DOI 10.1038/nature16548.

Palmer JD, Herbon LA. 1988. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution* **28**:87–97 DOI 10.1007/bf02143500.

Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* **7**:84 DOI 10.1186/1741-7007-7-84.

Petit RJ, Aguinagalde I, De Beaulieu J-L, Bittkau C, Brewer S, Cheddadi R, Ennos R, Fineschi S, Grivet D, Lascoux M, Mohanty A, Müller-Starck G, Demesure-Musch B, Palmé A, Martín JP, Rendell S, Vendramin GG. 2003. Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* **300**:1563–1565 DOI 10.1126/science.1083264.

Raman G, Park S. 2015. Analysis of the complete chloroplast genome of a medicinal plant, Dianthus superbus var. longicalyncinus, from a comparative genomics perspective. *PLOS ONE* **10**:e0141329 DOI 10.1371/journal.pone.0141329.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572–1574 DOI 10.1093/bioinformatics/btg180.

Rothwell GW, Van Atta MR, Ballard HE, Stockey RA. 2004. Molecular phylogenetic relationships among Lemnaceae and Araceae using the chloroplast trnL–trnF intergenic spacer. *Molecular Phylogenetics and Evolution* **30**:378–385 DOI 10.1016/S1055-7903(03)00205-7.

Sedlar K, Kolek J, Skutkova H, Branska B, Provaznik I, Patakova P. 2015. Complete genome sequence of Clostridium pasteurianum NRRL B-598, a non-type strain producing butanol. *Journal of Biotechnology* **214**:113–114 DOI 10.1016/j.jbiotec.2015.09.022.

Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z. 2012. An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLOS ONE* **7**:e31468 DOI 10.1371/journal.pone.0031468.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayshida N, Matsubayasha T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kata A, Tohdoh N, Shimada H, Sugiura M. 1986. The complete nucleotide sequence of the tobacco chloroplast genome. *Plant Molecular Biology Reporter* **4**:111–148 DOI 10.1007/bf02669253.

Sree KS, Bog M, Appenroth K-J. 2016. Taxonomy of duckweeds (Lemnaceae), potential new crop plants. *Emirates Journal of Food and Agriculture* **28**:291–302 DOI 10.9755/ejfa.2016-01-038.

Wakasugi T, Tsudzuki T, Sugiura M. 2001. The genomics of land plant chloroplasts: gene content and alteration of genomic information by RNA editing. *Photosynthesis Research* **70**:107–118 DOI 10.1023/a:1013892009589.

Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo MC, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J, Byrant DW, Mockler TC, Appenroth KJ, Grimwood J, Jenkins J, Chow J, Choi C, Adam C, Cao XH, Fuchs J, Schubert I, Rokhsar D, Schmutz J, Michael TP, Mayer KFX, Messing J. 2014. The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nature Communications* **5**:Article 3311 DOI 10.1038/ncomms4311.

Wang W, Messing J. 2011. High-throughput sequencing of three lemnoideae (duckweeds) chloroplast genomes from total DNA. *PLOS ONE* **6(9)**:e24670 DOI 10.1371/journal.pone.0024670.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* **19**:100–114 DOI 10.1111/j.1365-294X.2009.04474.x.

Wu Z, Gu C, Tembrock LR, Zhang D, Ge S. 2017. Characterization of the whole chloroplast genome of Chikusichloa mutica and its comparison with other rice tribe (Oryzeae) species. *PLOS ONE* **12**:e0177553 DOI 10.1371/journal.pone.0177553.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**:3252–3255 DOI 10.1093/bioinformatics/bth352.

Zhang T, Fang Y, Wang X, Deng X, Zhang X, Hu S, Yu J. 2012. The complete chloroplast and mitochondrial genome sequences of Boea hygrometrica: insights into the evolution of plant organellar genomes. *PLOS ONE* **7**:e30531 DOI 10.1371/journal.pone.0030531.

Zhang T, Zhang X, Hu S, Yu J. 2011. An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. *Plant Methods* **7**:Article 38 DOI 10.1186/1746-4811-7-38.

Zhao H, Appenroth K, Landesman L, Salmeán AA, Lam E. 2012. Duckweed rising at Chengdu: summary of the 1st international conference on duckweed application and research. *Plant Molecular Biology* **78**:627–632 DOI 10.1007/s11103-012-9889-y.

Zheng X, Wang J, Feng L, Liu S, Pang H, Qi L, Li J, Sun Y, Qiao W, Zhang L, Cheng Y, Yang Q. 2017. Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. *Scientific Reports* **7**:Article 1555 DOI 10.1038/s41598-017-01518-5.