MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) **14**, 1024–1031 doi: 10.1111/1755-0998.12251

Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs

JUN-BO YANG, DE-ZHU LI and HONG-TAO LI
Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China

Abstract

Chloroplast genomes supply indispensable information that helps improve the phylogenetic resolution and even as organelle-scale barcodes. Next-generation sequencing technologies have helped promote sequencing of complete chloroplast genomes, but compared with the number of angiosperms, relatively few chloroplast genomes have been sequenced. There are two major reasons for the paucity of completely sequenced chloroplast genomes: (i) massive amounts of fresh leaves are needed for chloroplast sequencing and (ii) there are considerable gaps in the sequenced chloroplast genomes of many plants because of the difficulty of isolating high-quality chloroplast DNA, preventing complete chloroplast genomes from being assembled. To overcome these obstacles, all known angiosperm chloroplast genomes available to date were analysed, and then we designed nine universal primer pairs corresponding to the highly conserved regions. Using these primers, angiosperm whole chloroplast genomes can be amplified using long-range PCR and sequenced using next-generation sequencing methods. The primers showed high universality, which was tested using 24 species representing major clades of angiosperms. To validate the functionality of the primers, eight species representing major groups of angiosperms, that is, early-diverging angiosperms, magnoliids, monocots, Saxifragales, fabids, malvids and asterids, were sequenced and assembled their complete chloroplast genomes. In our trials, only 100 mg of fresh leaves was used. The results show that the universal primer set provided an easy, effective and feasible approach for sequencing whole chloroplast genomes in angiosperms. The designed universal primer pairs provide a possibility to accelerate genome-scale data acquisition and will therefore magnify the phylogenetic resolution and species identification in angiosperms.

Keywords: angiosperm chloroplast genomes, next-generation sequencing, novel universal primer pairs, phylogenomics, species identification

Received 7 November 2013; revision received 27 February 2014; accepted 6 March 2014

Introduction

Chloroplasts are essential organelles in the plant cells. Molecular differentiation arisen in chloroplast (cp) genomes among plant species and even individuals offer promising tools for phylogenetic reconstruction and species identification. Phylogenomics (O'Brien & Stanyon 1999) has widely employed complete cp genome sequences for studying phylogenetics. For example, phylogenetic relationships among angiosperms were discussed (Samson *et al.* 2007), and relationships between different basal angiosperms have been resolved using genome-scale data from chloroplasts (Moore *et al.* 2007). Alone these same lines, Jansen *et al.* (2007) used 64 cp genomes to infer relationships within angiosperms. Furthermore, Moore *et al.* (2010) helped resolve the early diversification of eudicots, a major clade of angiosperms, using 83 cp genomes. Chloroplast phylogenomics have also suggested that *Ginkgo biloba* is sister to cycads (Wu *et al.* 2013). Similarly, the phylogenetic relationships within the family Pinaceae (Lin *et al.* 2010) and the genus *Pinus* (Parks *et al.* 2009) have been inferred using complete cp genome sequences. Recently, a few studies have discussed using complete cp genomes to identify species (Nock *et al.* 2011) or as organelle-scale barcodes (Yang *et al.* 2013a,b). Complete cp genome sequencing has become a universal method for obtaining evolutionary information that can be used to address questions of species identification and phylogenetic analyses of plants.

Information from chloroplast genomes is indispensable for resolving angiosperm phylogeny as well as identifying angiosperm species. Investigators are constantly seeking out new ways of acquiring complete cp genomes. As a result, many strategies have been proposed to sequence entire cp genomes. These strategies

Correspondence: De-Zhu Li and Hong-Tao Li, Fax: 86 871 521 7791 and 86 871 5217791; E-mails: dzl@mail.kib.ac.cn and lihongtao@mail.kib.ac.cn

included the following: (i) isolating chloroplasts using the sucrose-gradient method, enriching the chloroplast DNA by rolling circle amplification, long-range PCR and BAC or fosmid library constructions, and then fragmenting DNA and cloned for sequencing (Goremykin *et al.* 2005; Wu *et al.* 2007; Cattolico *et al.* 2008), (ii) designing short-range PCR primers for conserved cp regions for Sanger DNA sequencing (Taberlet *et al.* 1991; Dumolin-Lapegue *et al.* 1997; Grivet *et al.* 2001; Provan *et al.* 2004; Heinze 2007; Haider 2011; Scarcelli *et al.* 2011; Dong *et al.* 2013; Li *et al.* 2013), and (iii) using next-generation sequencing (NGS), such as approach of constructing a DNA library after isolating total genomic DNA (Moore *et al.* 2006; Atherton *et al.* 2010), the targeted enrichment approach (Stull *et al.* 2013) and the 'genome skimming' approach (Steele *et al.* 2012; Straub *et al.* 2012). At present, acquiring cp genome sequences through NGS technologies (Moore *et al.* 2006; Mardis 2008; Shendure & Ji 2008) has become increasingly popular because DNA sequencing costs have fallen dramatically as NGS technologies are rapidly developed, the power of bioinformatics is exploited, and high-throughput data processing is readily available. However, to date, <300 cp genomes of angiosperms have been sequenced. Compared with the more than 270 000 flowering plant species that exist (Mabberley 2008; Joppa *et al.* 2011; Mora *et al.* 2011), we have barely begun to explore cp genome sequencing in angiosperms. Two major reasons contribute to the current low numbers of completely sequenced angiosperm cp genomes. First, a large quantity of fresh leaves is needed for chloroplast DNA extraction (Jansen *et al.* 2005). Second, it is difficult in many plants to isolate high-quality cp DNA, and considerable gaps were produced using low-quality cp DNA, which made it troublesome to assemble complete cp genome. Owing to these difficulties, obtaining complete cp genome sequences has been limited. These limitations severely restrict the extent to which investigators can analyse complete cp genome data.

A strategy for obtaining sufficient amounts of high quality, pure and complete cp genome DNA from a small number of fresh leaves and acquiring higher coverage of sequencing is urgently needed. The technologies involved in long-range PCR amplification (Barnes 1994; Cheng *et al.* 1994) and NGS methods make it possible to amplify whole cp genomes using several pairs of primers and then sequencing. Angiosperm cp genomes are about 160 kb, and long-range PCR can amplify more than 20 kb. Consequently, eight or nine pairs of primers are enough to cover an entire cp genome. Universal primers are the key for amplifying whole cp genomes of angiosperms.

In vascular plants, cp genomes have a conserved quadripartite structure composed of two copies of a large inverted repeat (IR) and two sections of unique DNA, which are referred to as the large single-copy (LSC) regions and small single-copy (SSC) regions, respectively (Palmer & Stein 1986; Jansen *et al.* 2005). Many highly conserved regions exist simultaneously in angiosperm cp genomes, such as tRNA genes, which are highly conserved in structure, content and location. The similarities of the conserved sequences provide us opportunities to amplify complete cp genomes of angiosperms by designing fewer than 10 pairs of universal primers, and the resulting fragments of PCR products can be sequenced using NGS technologies. Here, we designed nine pairs of universal primers that cover the entire cp genome, and then we tested the universality of these primers by performing PCR on the cp genomes of 24 species from 24 different angiosperm families that represented all the major clades designated by the Angiosperm Phylogeny Group III (APG III) system (Bremer *et al.* 2009). In this study, we validated our designed primers by sequencing the cp fragments of eight angiosperm species, which represented the major groups of angiosperms. After sequencing the fragments, we then assembled the eight corresponding complete cp genomes. We expect that these universal primers will effectively increase the efficiency and feasibility of complete cp genome sequencing. These primers can potentially offer genome-scale data acquisition, improve the phylogenetic resolution of angiosperms and aid in identifying angiosperm species, especially at the lower taxonomic levels (i.e. genera, species and populations).

Materials and methods

Plant materials

Twenty-four species of angiosperms from 24 different groups were field-collected. The species sampled were from the early-diverging angiosperms, magnoliids, Asparagales, commelinids, basal eudicots, Saxifragales, fabids, malvids, Ericales, lamiids and campanulids. Clean, healthy, fresh green leaves from the 24 plants were collected. The voucher herbarium specimens for the 24 sampled plants were deposited in the herbarium of Kunming Institute of Botany, Chinese Academy of Sciences (KUN) (Table 1).

Analysing angiosperm chloroplast genomes and designing the primers

All known angiosperm cp genomes available on GenBank were analysed (Table S1, Supporting information). We discovered many highly conserved regions that had similar structures, content and locations. For example, tRNA genes are highly conserved, and they

Table 1 Sampled species and voucher specimens used in this study

Species	Family	Order	Clade	Geographical origin	Voucher	GenBank accession
Acer buergerianum Miq. var. *ningpoensis* (Hce.) Rehd.	Sapindaceae	Sapindales	Malvids	KIB	Sd0060	KF753631
Calanthe triplicata (Willem.) Ames	Orchidaceae	Asparagales	Asparagales	KIB	Sd0053	KF753635
Camellia crapnelliana Tutch.	Theaceae	Ericales	Ericales	KIB	J.B.Yang 2013	KF753632
Nymphaea mexicana Zucc.	Nymphaeaceae	Nymphaeales	Early-diverging	KIB	Sd0050	KF753633
Paeonia sp.	Paeoniaceae	Saxifragales	Saxifragales	KIB	Sd0052	KF753636
Parakmeria yunnanensis Hu	Magnoliaceae	Magnoliales	Magnoliids	KIB	Sd0051	KF753638
Primula poissonii Franch.	Primulaceae	Ericales	Ericales	KIB	Sd0057	KF753634
Rosa odorata (Andr.) Sweet var. *gigantea* (Crep.) Rehd. et Wils.	Rosaceae	Rosales	Fabids	KIB	Sd0064	KF753637
Bauhinia brachycarfa Wall.	Fabaceae	Fabales	Fabids	KIB	Sd0062	
Bigonia cavaleriei Levl.	Begoniaceae	Cucurbitales	Fabids	Malipo, Yunnan	SYMB2013-086	
Cerciphyllum japonicum Sieb. et Zucc.	Cercidiphyllaceae	Saxifragales	Saxifragales	KIB	Sd0055	
Chimonocalamus longiusculus Hsueh et Yi	Poaceae	Poales	Commelinids	Xichou, Yunnan	MPF10182	
Hemsleya dipterygia Kuang et A. M. Lu	Cucurbitaceae	Cucurbitales	Fabids	KIB	LiHT13063	
Hovenia acerba Lindl.	Rhamnaceae	Rosales	Fabids	KIB	Sd0065	
Juglans sigillata Dode	Juglandaceae	Fagales	Fabids	KIB	Sd0059	
Lythrum salicaria var. *tomentosum* (Mill.) DC.	Lythraceae	Myrtales	Malvids	KIB	Sd0058	
Pedicularis densispica Franch. ex Maxim.	Orobanchaceae	Lamiales	Lamiids	Lijiang, Yunnan	YWB2013058	
Rhododendron simsii Planch.	Ericaceae	Ericales	Ericales	KIB	Sd0056	
Sabia sp.	Sabiaceae	Sabiales	Basal eudicots	Gongshan, Yunnan	GmT02-12-1011	
Securinega suffruticosa (Pall.) Rehd.	Phyllanthaceae	Malpighiales	Fabids	KIB	Sd0061	
Senecio scandens Buch.-Ham. ex D. Don	Asteraceae	Asterales	Campanulids	KIB	Sd0066	
Tetracentron sinensis Oliv.	Trochodendraceae	Trochodendrales	Basal eudicots	KIB	Sd0054	
Tilia paucicostata Maxim.	Malvaceae	Malvales	Malvids	KIB	13CS6898	
Viburnum macrocephalum Fort.	Adoxaceae	Dipsacales	Campanulids	KIB	Sd0063	

KIB, Kunming Botanical Garden of the Kunming Institute of Botany.

are located on the LSC, IR and SSC regions. Furthermore, the distance between adjacent tRNA genes is no longer than 20 kb. These regions therefore provide suitable targets for designing long-range PCR primers. We used Geneious software (Biomatters Ltd.; Auckland, New Zealand) (Meintjes *et al.* 2012) to select regions between two tRNAs, which ranged from 15 kb to 23 kb, and by default settings to search primers by binding the tRNAs in the following regions, such that they covered the entire angiosperm cp genomes: (i) *trn*G-UCC, *trn*R-UCU, *trn*C-GCA, *trn*Y-GUA, *trn*S-GGA, *trn*T-GGU, *psa*I, *trn*W-CCA and *pet*B in the LSC region; (ii) *trn*I-CAU and *trn*L-CAA in the IRb region; (iii) *trn*L-UAG in the SSC region; and (iv) *trn*N-GUU and *trn*I-CAU in the IRa region. The resulting PCR products ranged from 15 to 23 kb. Because the family Poaceae shows structural differences in their cp genomes, namely the three inversions located in the LSC region (Doyle *et al.* 1992), three special primer pairs were designed for the LSC region of Poaceae.

Extracting DNA and testing the universality of the primers

Total genomic DNA was extracted from 100 mg of actively growing fresh leaves using a modified CTAB (hexadecyltrimethylammonium bromide) method (Doyle & Doyle 1987), in which 4% CTAB was used instead of

2% CTAB, and adding approximately 1% polyvinyl polypyrrolidone (PVP) and 0.2% DL-dithiothreitol (DTT).

Twenty-four samples were used to test the universality of the primer set. Amplification through long-range PCR was performed using Takara PrimeSTAR GXL DNA polymerase (TAKARA BIO INC.; Dalian, China). Amplifications were performed in 25-μL reaction mixtures containing 1×PrimeSTAR GXL buffer (10 mM Tris-HCl (pH 8.2), 1 mM magnesium chloride ($MgCl_2$), 20 mM sodium chloride (NaCl), 0.02 mM ethylenediaminetetraacetic acid (EDTA), 0.02 mM DTT; 0.02% Tween 20, 0.02% Nonidet P-40, and 10% glycerol); 1.6 mM of dNTPs, 0.5 μM of each primer; 1.25 U of PrimeSTAR GXL DNA polymerase, and 30–100 ng of DNA template. PCR amplifications were conducted under the following conditions: 94 °C for 1 min, 30 cycles at 98 °C for 10 s and 68 °C for 15 min, followed by a final extension step at 72 °C for 10 min.

Assembling the genomes of eight angiosperm species

Eight species representing the major angiosperm groups (Table 1) were used to verify that complete cp genomes could be assembled from the amplified PCR products. Purified DNA (6 μg) from the amplified PCR products was fragmented and used to construct short-insert (500 bp) libraries according to the manufacturer's manual (Illumina). DNA from each individual was indexed using tags and pooled together in one lane of a Genome Analyzer (Illumina Hiseq 2000) for sequencing at Beijing Genomics Institute (BGI) in Shenzhen, China.

First, various quality control checks on the short reads were performed using NGS QC Tool Kit (Patel & Jain 2012) to filter the Illumina data for high-quality (cut-off value for percentage of read length = 80, cut-off value for PHRED quality score = 30) and vector- and adaptor-free reads. High-quality short reads were assembled into contigs using *de novo* assembler of CLC Genomics Workbench v. 6.5 (CLC Bio), a *de novo* sequence assembly software, using a k-mer of 63 and a minimum contig length of 1 kb. Then, using the Basic Local Alignment Search Tool (BLAST: http://blast.ncbi.nlm.nih.gov/) with the default search parameters, we identified highly similar genome sequences. Next, we determined the proper orders of our aligned contigs using the highly similar genome sequences identified in the BLAST search as references. At this point, the *de novo* contigs were assembled into complete cp genomes. Finally, we validated the four junctions between LSC/IRs and SSC/IRs using Sanger sequencing of the PCR-based products. The final complete cp genome sequences were deposited into GenBank (Table 1).

We annotated the sequenced cp genomes using the Dual Organellar GenoMe Annotator (DOGMA) tool (Wyman *et al.* 2004), after which we manually corrected start and stop codons and intron/exon boundaries so that they matched their gene predictions. The sequences of identified tRNA genes were obtained using DOGMA and tRNAscan-SE (version 1.23) (Lowe & Eddy 1997). The functional classification of cp genes was carried out by referring to CpBase (http://chloroplast.ocean.washington.edu/). The annotated GenBank files of the cp genomes were used to draw gene maps using the OrganellarGenomeDRAW tool (OGDRAW) (Lohse *et al.* 2013).

Results

Primers and their universality

We designed nine pairs of primers (Table 2) covering whole cp genomes of angiosperms (Fig. 1), except for the cp genomes in the family Poaceae. For Poaceae, three universal primer pairs CP_1, CP_2 and CP_3 were replaced by three special pairs of primers CP_POA_1, CP_POA_2 and CP_POA_3 in the LSC region (Fig. 2), respectively. Among the nine universal primer pairs, five primer pairs were designed to bridge the four junctions between LSC/IRs and SSC/IRs and four other primer pairs were located in the LSC region (Fig. 1). We designed the primers such that the resulting PCR fragments would range from 15 to 23 kb.

Most of the primer pairs amplified their target sequences well for the 24 species we tested, which represented the major clades of APG III (Fig.S1, Supporting information) system. As designed, the size of the products ranged from 15 to 23 kb, and we only found instances of poor amplification for *Rhododendron simsii* and *Senecio scandens*, when primer pairs CP_1, CP_2 and CP_1, CP_3, CP_7 were used (Fig. S1, Supporting information), respectively.

Validation complete chloroplast genomes of eight species

Using the Illumina Hiseq 2000 system (Illumina; San Diego, CA, USA), eight species representing the major angiosperm groups, that is, early-diverging angiosperms, magnoliids, monocots, Saxifragales, fabids, malvids and asterids, were sequenced to produce 5715 356–5820 668 paired-end reads (average reads length = 90 bp). After filtering these paired-end reads using NGS QC Toolkit, a total of 5029 378–5105 998 high-quality paired-end reads were obtained for *de novo* assembly into contigs by the CLC Genomics Work1bench v. 6.5 (CLC Bio). Then, we employed the highly similar genome sequences identified in the BLAST search as references screening these contigs to obtain four to eight contigs covering 4596 630–5015 055 reads and finally assembled the contigs into cp

Table 2 Universal primers for amplifying complete chloroplast genomes of angiosperms

Primer	Primer sequence	Product size
CP_1F_*trn*I	GGCTGAATGGTTAAAGCGCCCA	15 kb
CP_1R_*trn*R	TTGCGTCCAATAGGATTTGAACCTATACC	
CP_2F_*trn*G	GGTTCGATTCCCGCTACCCGC	23 kb
CP_2R_*trn*Y	TGGTTCAAATCCAGCTCGGCCC	
CP_3F_*trn*C	CCCCGGTTCAAATCTGGGTGTCG	18 kb
CP_3R_*trn*S	CGCCTTGAACCACTCGGCCA	
CP_4F_*trn*S	TGTAGGAGAGATGGCCGAGTGG	15 kb
CP_4R_*psa*I	CCATTGCAATTGCCGGAAATACTAGGC	
CP_5F_*trn*T	ACGGCGGGAGTCATTGGTTCA	15 kb
CP_5R_*trn*W	AGTTCGGTAGAACGTGGGTCTCCA	
CP_6F_*trn*W	TGAACCTACGACATCGGGTTTTGGAGA	20 kb
CP_6R_*trn*I	ATGTACGAGGATCCCCGCTAAGCATC	
CP_7F_*pet*B	GCTTGAGCTGTACGAGATGAAAGTCT	18 kb
CP_7R_*trn*L	AGAGCGTGGAGGTTCGAGTCC	
CP_8F_*trn*L	GGACTCGAACCTCCACGCTCT	20 kb
CP_8R_*trn*L	GCCGCTACTCGGACTCGAACC	
CP_9F_*trn*L	GGTTCGAGTCCGAGTAGCGGC	16 kb
CP_9R_*trn*N	ACAGCCGACCGCTCTACCAC	
CP_POA_1F_*trn*I*	ATTTGCGGGTTCAATTCCTGCTGGATG	20 kb
CP_POA_1R_*trn*D*	CTTGACAGGGCGGTGCTCTGACCAAT	
CP_POA_2F_*trn*D*	AATTGGTCAGAGCACCGCCCTGTCAAG	21 kb
CP_POA_2R_*trn*fM*	AGCTGTTTGGTAGCTCACAAGGCTCAT	
CP_POA_3F_*trn*fM*	CCCCAAGGTTATGAGCCTTGTGAGCTA	9 kb
CP_POA_3R_*trn*S*	CTACATAGCAGTTCCAATGCTACGCCT	

*Three special primer pairs for the family Poaceae.

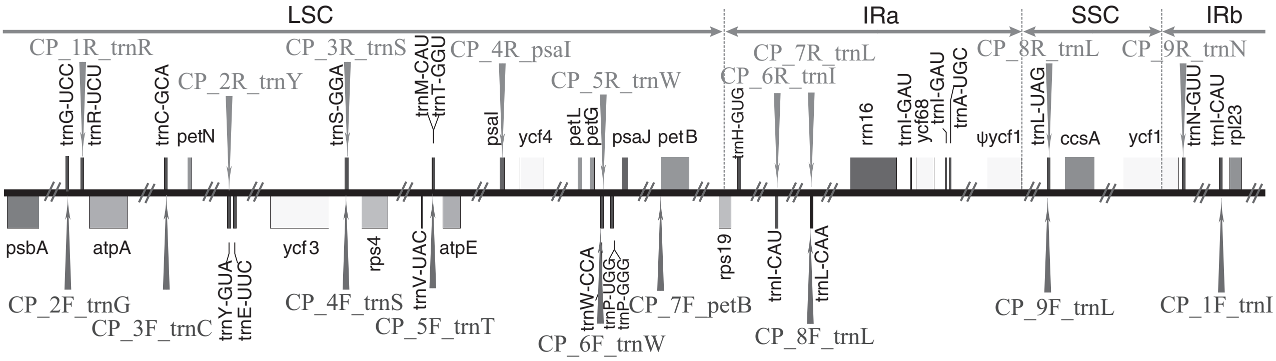

Fig. 1 Distribution of the universal primers designed to cover entire angiosperm chloroplast genomes. The blue and green arrows indicate F and R primers, respectively.

genome. The sequencing depth of coverage was > 2000× for the cp genome, and the utilization ratio of the reads was 90% to 98%. After *de novo* assembly, we obtained eight complete cp genomes. In addition, we simultaneously validated the four junction regions in each cp genome using PCR-based sequencing.

All eight cp genomes were composed of a single circular double-stranded DNA molecule, and they displayed the typical quadripartite structure of angiosperms, consisting of a pair of IRs (25 199–26 752 bp) separated by the LSC (83 444–90 019 bp) and SSC (16 969–19 533 bp) regions (Fig. S2, Supporting information). The eight cp genomes were also AT rich, and their overall AT content ranges from 60.7% to 63.3%. They were found to encode a set of 136–139 predicted functional genes, of which 116–117 were unique and 20–23 were duplicated in the IR regions. The 116–117 unique genes comprised 79–80 protein-coding genes, 33 transfer RNA genes and four ribosomal RNA genes (Table S2, Fig. S2, Supporting information). We found that 15 distinct genes, namely

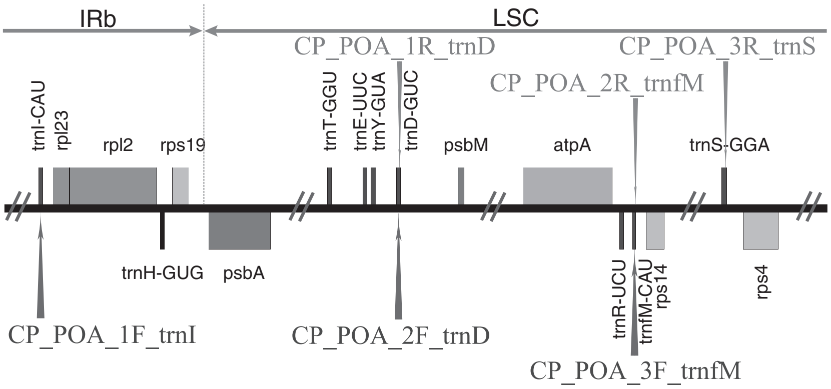

Fig. 2 Distribution of the three special primer pairs for plants in the family Poaceae. The blue and green arrows indicate F and R primers, respectively.

*ndh*A, *ndh*B, *pet*B, *pet*D, *rpl*16, *rpl*2, *rpo*C1, *rps*12, *rps*16, *trn*A-UGC, *trn*G-GCC, *trn*I-GAU, *trn*K-UUU, *trn*L-UAA and *trn*V-UAC, contained one intron and while two genes (*clp*P and *ycf*3) contained two introns. Minor rearranged and gene loss-and-gain were checked, such as inversion of *ycf*15 gene in *Acer buergerianum var. ningpoensis*, gene loss of *rpl*32 gene in *Paeonia lactiflora* and intron loss of *atp*F gene in *Rosa odorata var. gigantean*. The eight cp genomes also revealed that various degrees of IR to LSC contraction and expansion were identified (Fig. S3, Supporting information).

Discussion

Significance of the designed universal primer set

With the rapid development of technologies and bioinformatics, NGS has become a most popular method for genome sequencing. However, some factors still limit the complete sequencing of cp genomes using NGS technologies. First, it is difficult to collect enough fresh leaves at a time for sequencing. This task is nearly impossible when collecting some herbaceous plants because the plants are so small. Second, it is also difficult to isolate high-quality cp DNA in many plants as the plant tissue contains plenty of secondary metabolites such as polysaccharide, polyphenols and esters. In general, sequencing 500 Mb of DNA can produce approximately 3000 average-sized cp genomes at a depth of coverage 100×, but the utilization ratio of sequenced reads is usually below 4%. However, the depth of coverage of sequenced using low-quality cp DNA would be far <100×. As a result, dozens or even hundreds of gaps exist in sequenced cp genomes, rendering it impossible to assemble complete cp genomes. In this study, however, we designed a set of nine universal primers based long-range PCR to overcome these limitations. A few fresh leaves (100 mg) provided enough material to extract and amplify high-quality DNA using long-range PCR, which gave enough pure cp DNA to sequence the entire cp genome. Encouragingly, the utilization ratio of sequenced reads was > 90%, and the depth of coverage was > 2000×. Because of these high coverage values, the assembled cp genomes had only a few, if any, gaps. On the whole, a depth of coverage 200× is enough to assemble complete cp genome. Consequently, sequencing 100 Mb of DNA or even less can assemble complete cp genome using this method. In other words, it will greatly improve the multiplexing of NGS using this method, given that it can sequence at least 5 times samples in one Illumina lane.

Reliability of the universal primer set

Determining the sequencing reliability of complete genomes is crucial for phylogenomic studies, and it is directly related to the reliability of the primers. Using the designed universal primers, the resulting PCR amplifications were found to cover the entire cp genomes of the sampled angiosperms. Consequently, the PCR products contained the pure DNA of the complete cp genomes. To validate the accuracy of this approach, we resequenced the species *Camellia crapnelliana* using two other methods according our earlier study (Yang *et al.* 2013a), which used 123 pairs of primers for PCR and extracted cp DNA for NGS, respectively. Using three sequencing methods independently, the assembled cp genomes of *Camellia crapnelliana* were completely identical. In addition, we simultaneously verified the four junction regions of the eight sequenced angiosperm species using direct PCR sequencing, and the resulting sequences were identical to those obtained using long-range PCR and NGS techniques.

Usefulness of the universal primer set

The nine universal primer pairs were tested using 24 species, and eight of them were sequenced. The amplification and sequencing results showed that the primers were useful and efficient. All eight complete cp genomes were assembled without gaps. Furthermore, small structure variations in cp genomes did not affect the usefulness of the nine universal primer pairs because the primers corresponded to highly conserved regions of the genome. While we did find some small structural variations in the eight sequenced cp genomes, such as gene loss and inversion, these assembled cp genomes were nonetheless complete.

Stemming from the efficiency and ease of using the nine universal primer pairs, we have been engaged in related phylogenomics research in our laboratory. At present, we have investigated dozens of angiosperm families, including hundreds of species. Except for some species of the family such as Asteraceae and Ericaceae, results from using these primers have been promising.

Although the nine universal primer pairs showed high universality in the present study, they did not fully resolve issues surrounding sequencing angiosperm cp genomes. Some plants of Asteraceae and Ericaceae were the other groups of angiosperms that did not respond to the nine universal primer pairs well because of possible structural differences in their cp genomes. In the case of parasitic angiosperms, they could not respond to the nine universal primer pairs owing to plenty of genes loss (Revill *et al.* 2005; Funk *et al.* 2007; McNeal *et al.* 2007; Logacheva *et al.* 2011; Delannoy *et al.* 2011; Li *et al.* 2013). Perhaps, it will be necessary to design special primer sets for these exceptional angiosperms with structural cp genome differences that distinguish them from 'standard' angiosperms.

The nine universal primer pairs provided an effective and feasible approach for sequencing complete cp genomes of angiosperms. Compared with previous methods, these universal primers afford significant advantages in cost, reliability and integrity of the sequenced genomes, and efficient experimentation, especially because only a small amount of sample is required and acquiring higher coverage of sequencing. Once adopted, these universal primers will boost phylogenomics research on angiosperms and amplify the phylogenetic resolution of these very important plants.

Acknowledgements

We are very grateful to Ms. Juan-Hong Zhang, Dr. Shu-Dong Zhang, Peng-Fei Ma, Jie Cai, Wen-Bin Yu, Yu-Min Shui and Xiang Zhao of Kunming Institute of Botany for their help with the experiments and materials. This work was funded by the National Natural Science Foundation of China (31170182, 30870169), the National 863 Project of China (2012AA021801), Instruments Development and Technology Innovation Project of the Chinese Academy of Sciences (Integrated Development System for Acquiring Plant Genetic Information Efficiency) and Yunnan Provincial Natural Science Foundation (2010CD108).

References

Atherton RA, McComish BJ, Shepherd LD *et al.* (2010) Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. *Plant Methods*, **6**, 22.

Barnes WM (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 2216–2220.

Bremer B, Bremer K, Chase MW *et al.* (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society*, **161**, 105–121.

Cattolico RA, Jacobs MA, Zhou Y *et al.* (2008) Chloroplast genome sequencing analysis of *Heterosigma akashiwo* CCMP452 (West Atlantic) and NIES293 (West Pacific) strains. *BMC Genomics*, **9**, 211.

Cheng S, Fockler C, Barnes WM, Higuchi R (1994) Effective amplification of long targets from cloned inserts and human genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 5695–5699.

Delannoy E, Fujii S, Colas des Francs-Small C, Brundrett M, Small I (2011) Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlights evolutionary constraints on plastid genomes. *Molecular Biology and Evolution*, **28**, 2077–2086.

Dong W, Xu C, Cheng T, Lin K, Zhou S (2013) Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of saxifragales. *Genome Biology and Evolution*, **5**, 989–997.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, **19**, 11–15.

Doyle JJ, Davis JI, Soreng RJ, Garvin D, Anderson MJ (1992) Chloroplast DNA inversions and the origin of the grass family (*Poaceae*). *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 7722–7726.

Dumolin-Lapegue S, Pemonge MH, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. *Molecular Ecology*, **6**, 393–397.

Funk HT, Berg S, Krupinska K, Maier UG, Krause K (2007) Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. *BMC Plant Biology*, **7**, 45.

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Molecular Biology and Evolution*, **22**, 1813–1822.

Grivet D, Heinze B, Vendramin GG, Petit RJ (2001) Genome walking with consensus primers: application to the large single copy region of chloroplast DNA. *Molecular Ecology Notes*, **1**, 345–349.

Haider N (2011) Chloroplast-specific universal primers and their uses in plant studies. *Biologia Plantarum*, **55**, 225–236.

Heinze B (2007) A database of PCR primers for the chloroplast genomes of higher plants. *Plant Methods*, **3**, 4.

Jansen RK, Raubeson LA, Boore JL *et al.* (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology*, **395**, 348–384.

Jansen RK, Cai Z, Raubeson LA *et al.* (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19369–19374.

Joppa LN, Roberts DL, Pimm SL (2011) How many species of flowering plants are there? *Proceedings of the Royal Society B-Biological Sciences*, **278**, 554–559.

Li X, Zhang T-C, Qiao Q *et al.* (2013) Complete chloroplast genome sequence of holoparasite *Cistanche deserticola* (*Orobanchaceae*) reveals gene Loss and horizontal gene transfer from its host *Haloxylon ammodendron* (*Chenopodiaceae*). *PLoS ONE*, **8**, e58747.

Lin CP, Huang JP, Wu CS, Hsu CY, Chaw SM (2010) Comparative chloroplast genomics reveals the evolution of *Pinaceae* genera and subfamilies. *Genome Biology and Evolution*, **2**, 504–517.

Logacheva MD, Schelkunov MI, Penin AA (2011) Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis*. *Genome Biology and Evolution*, **3**, 1296–1303.

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.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**, 955–964.

Mabberley DJ (2008) *Mabberley's Plant-book: A Portable Dictionary of Plants, Their Classifications, and Uses*. Cambridge University Press, Cambridge, UK.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133–141.

McNeal JR, Kuehl JV, Boore JL, de Pamphilis CW (2007) Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biology*, **7**, 57.

Meintjes P, Duran C, Kearse M *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Moore MJ, Dhingra A, Soltis PS *et al.* (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 17.

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19363–19368.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 4623–4628.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean? *PLoS Biology*, **9**, e1001127.

Nock CJ, Waters DL, Edwards MA *et al.* (2011) Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal*, **9**, 328–333.

O'Brien SJ, Stanyon R (1999) Phylogenomics - Ancestral primate viewed. *Nature*, **402**, 365–366.

Palmer JD, Stein DB (1986) Conservation of chloroplast genome structure among vascular plants. *Current Genetics*, **10**, 823–833.

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.

Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE*, **7**, e30619.

Provan J, Murphy S, Maggs CA (2004) Universal plastid primers for *Chlorophyta* and *Rhodophyta*. *European Journal of Phycology*, **39**, 43–50.

Revill MJ, Stanley S, Hibberd JM (2005) Plastid genome structure and loss of photosynthetic ability in the parasitic genus *Cuscuta*. *Journal of Experimental Botany*, **56**, 2477–2486.

Samson N, Bausher MG, Lee SB, Jansen RK, Daniell H (2007) The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. *Plant Biotechnology Journal*, **5**, 339–353.

Scarcelli N, Barnaud A, Eiserhardt W *et al.* (2011) A set of 100 chloroplast DNA primer pairs to study population genetics and phylogeny in monocotyledons. *PLoS ONE*, **6**, e19954.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Steele PR, Hertweck KL, Mayfield D *et al.* (2012) Quality and quantity of data recovered from massively parallel sequencing: Examples in *Asparagales* and *Poaceae*. *American Journal of Botany*, **99**, 330–348.

Straub SCK, Parks M, Weitemier K *et al.* (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–364.

Stull GW, Moore MJ, Mandala VS *et al.* (2013) A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences*, **1**, 1200497.

Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology*, **17**, 1105–1109.

Wu CS, Wang YN, Liu SM, Chaw SM (2007) Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: Insights into cpDNA evolution and phylogeny of extant seed plants. *Molecular Biology and Evolution*, **24**, 1366–1379.

Wu CS, Chaw SM, Huang YY (2013) Chloroplast phylogenomics indicates that *Ginkgo biloba* is sister to *Cycads*. *Genome Biology and Evolution*, **5**, 243–254.

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.

Yang J-B, Yang S-X, Li H-T, Yang J, Li D-Z (2013a) Comparative chloroplast genomes of *Camellia species*. *PLoS ONE*, **8**, e73053.

Yang JB, Tang M, Li HT, Zhang ZR, Li DZ (2013b) Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology*, **13**, 84.

J.B.Y., D.Z.L. and H.T.L. designed research and wrote the paper. J.B.Y. and H.T.L. performed research. H.T.L. analysed data.

Data Accessibility

DNA sequences: GenBank accessions KF753631- KF753 638; NCBI SRA: SRP037787.

GenBank accession no. is listed in Table 1.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 PCR gel profiles of the 24 plant species used to test the universality of the primers. Lanes from 1 to 9 correspond to primers CP_1, CP_2, CP_3, CP_4, CP_5, CP_7, CP_8, CP_9 and CP_6, respectively.

Fig. S2 Gene maps of the eight sequenced chloroplast genomes. Genes shown outside of the outer circle are transcribed clockwise, and those inside are transcribed counterclockwise. Color coded genes represent their corresponding functional group. The dashed area in the inner circle indicates the GC content of the chloroplast genome.

Fig. S3 Contraction and expansion of IRs.

Table S1 List of analyzing angiosperm chloroplast genomes for designing the primers

Table S2 Characteristics of the eight sequenced chloroplast genomes