Tree Genetics & Genomes (2015) 11: 114
DOI 10.1007/s11295-015-0942-2

ORIGINAL ARTICLE

An improved method for chloroplast genome sequencing in non-model forest tree species

Fang K. Du[1] · **Tiange Lang**[2] · **Sihai Lu**[1,3] · **Yuyao Wang**[1] · **Junqing Li**[1] · **Kangquan Yin**[1,4]

Received: 3 June 2015 / Revised: 23 September 2015 / Accepted: 30 September 2015 / Published online: 9 October 2015

Abstract Chloroplast genomes can provide a large amount of information and resources for use in studies on plant evolution and molecular ecology. However, a rapid and efficient method for obtaining chloroplast genome sequences is still lacking. In this study, we report a modified method for the isolation of intact chloroplasts, which needs less than 0.5 g leaf material. Coupled with rolling circle amplification (RCA), next-generation sequencing, and a pipeline combining de novo assembly and reference-guided assembly (RGA), we successfully obtained a complete chloroplast genome for the non-model forest tree species, evergreen oak *Quercus spinosa*, with as many as 36 % of the sequence reads mapped to the chloroplast genome. The *Q. spinosa* cpDNA is 160,825 bp in length and codes for 134 genes (89 protein coding, 8 ribosomal RNAs (rRNAs), and 36 distinct transfer RNAs (tRNAs)). The genome organization and arrangement are similar to those found among most angiosperm chloroplast genomes. Our inexpensive and efficient protocol can be applied to the reconstruction of chloroplast genomes for plant evolutionary studies, especially in non-model tree species.

Keywords Chloroplast · Next-generation sequencing · Rolling circle amplification (RCA) · Oak

Communicated by Y. Tsumura

This article is part of the Topical Collection on *Genome Biology*

Electronic supplementary material The online version of this article (doi:10.1007/s11295-015-0942-2) contains supplementary material, which is available to authorized users.

✉ Fang K. Du
dufang325@bjfu.edu.cn

✉ Kangquan Yin
yinkangquan@pku.edu.cn

[1] College of Forestry, Beijing Forestry University, Beijing 100083, People's Republic of China

[2] Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Menglun, Mengla 666303, Yunnan, People's Republic of China

[3] College of life science, Lanzhou University, Lanzhou 730000, Gansu, People's Republic of China

[4] State Key Laboratory of Plant Genomics, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, People's Republic of China

Introduction

The chloroplast (cp) genome plays an important role in plant molecular ecology and evolution studies because, in contrast to the nuclear (nu) genome, it is haploid, is generally non-recombinant, and evolves relatively rapidly. So far, more than 300 complete chloroplast genome sequences from angiosperm species, covering a wide range of taxa, have been deposited at NCBI (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid). However, plant species whose cp genomes have been sequenced represent only a tiny proportion (0.1 %) of all existing angiosperms (Aslan et al. 2013). Prior to 2006, the complete chloroplast genome sequences of no more than 10 species had been deposited with NCBI, whereas since 2010, the number of chloroplast genomes reported has increased exponentially (Fig. 1a and Supplementary Table S1). Up to 2010, the Sanger sequencing method, often termed the first-generation sequencing approach, was used almost exclusively for obtaining complete chloroplast genome sequences. Next-generation sequencing methods, which were first used in 2006 (Schuster 2008), became predominant from 2011 onward (Fig. 1a).

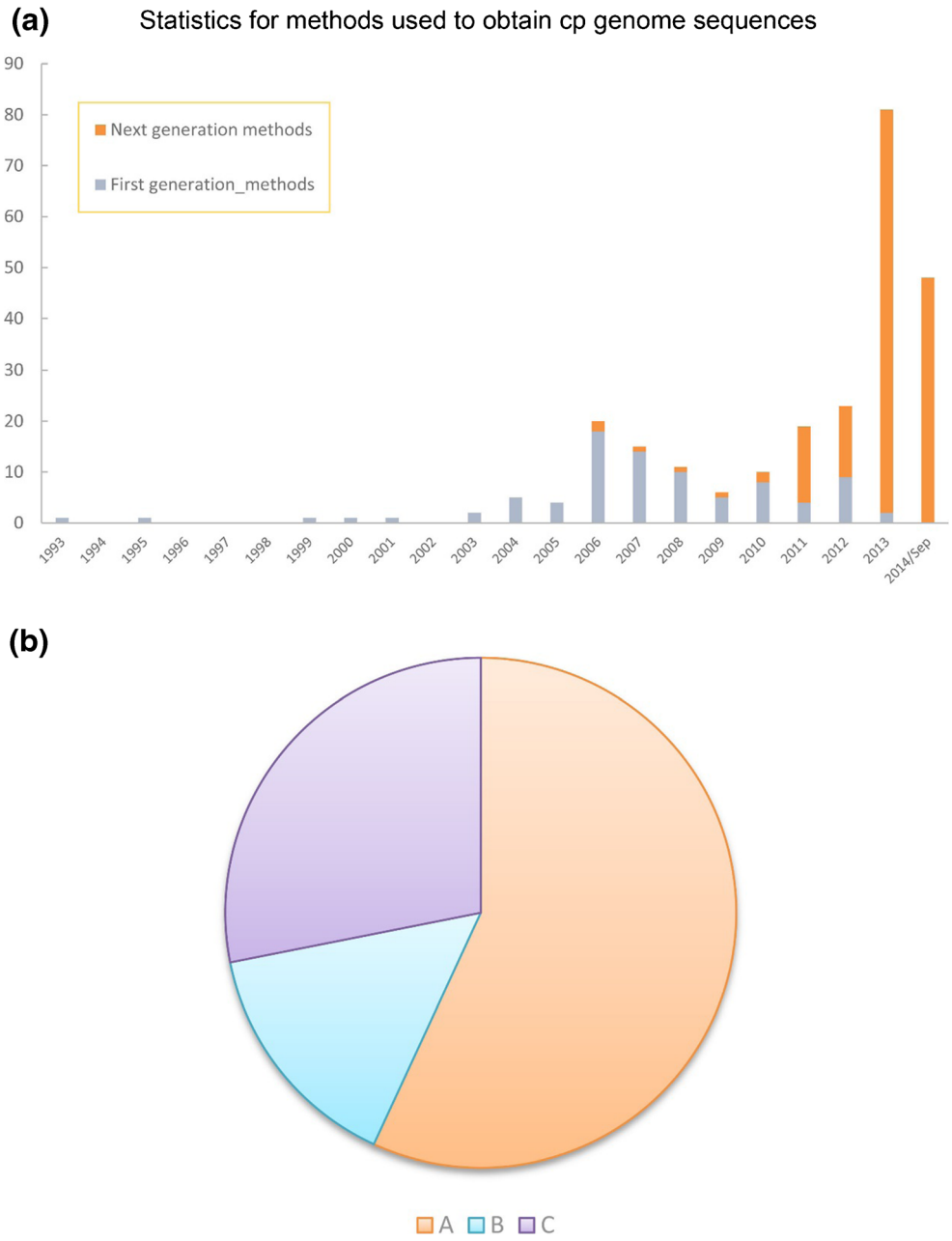

Fig. 1 Flowering plant cp genome sequences. **a** Trends in sequencing methods used for obtaining cp genome data. The *x*-axis shows year and the *y*-axis shows the number of cp genome sequences. *Blue columns* represent cp genome sequences obtained by first-generation sequencing methods and *orange columns* represent cp genome sequences obtained by next-generation sequencing methods. **b** Comparison of three different approaches for obtaining cp genome sequences. *A*: Isolation of pure cpDNA followed by NGS sequencing. *B*: Long-range PCR followed by NGS sequencing. *C*: Extraction of total DNA followed directly by NGS sequencing

Although first-generation sequencing can provide relatively long sequences reliably, the high costs and low throughput of this method limit its application. The large number of reads generated by next-generation sequencing facilitates assembly of cp genomes, given their relatively small size and low complexity (Cronn et al. 2008). For example, an Illumina HiSeq-2500 can produce 600 GB of data per run; this equates to 37, 500 cp genomes with an average size of 160 kb at a coverage of 100×. Next-generation sequencing is undoubtedly a powerful, economical, and time-saving method for cp genome sequencing. To take full advantage of its potential, obtaining high-quality cpDNA prior to sequencing is critical.

There are four main strategies used to obtain cpDNA for sequencing (Table 1): First, pure cpDNA can be isolated from chloroplasts. This approach requires isolation of a sufficiently large number of chloroplasts, from which cpDNA is released through lysis. Sucrose gradient centrifugation, an essential technique in chloroplast isolation, requires access to a high-speed refrigerated centrifuge. In addition, this method usually requires the use of about 100 g of fresh leaf starting material (Jansen et al. 2005), although 20 g of fresh leaf tissue is sufficient for a modified protocol (Shi et al. 2012). The second approach is based on the amplification of large fragments of cpDNA by long-range PCR, taking advantage of the

Table 1 Comparison of four methods used for obtaining cp genomes

Method	Advantages	Disadvantages	Reference
Isolation of pure cpDNA and sequencing	High yield of cpDNA; high quality of cpDNA	Substantial amount of leaf material required	Jansen et al. 2005
Long-range PCR and sequencing	Easy to perform; DNA specific to cp genome is enriched; sequence coverage is more likely to be even	Low transferability of primers	Goremykin et al. 2003a, b, 2004, 2005; Chung et al. 2007; Mardanov et al. 2008; Wu et al. 2009; 2010; Leseberg and Duvall 2009
Sequencing genomic DNA (genome skimming)	No need for chloroplast isolation; possible to recover complete cp genome, partial mitochondrial DNA and nuclear ribosomal DNA at the same time; allows for limited multiplexing of cp genomes	High depth of sequencing	Straub et al. 2012; Huang et al. 2014
Targeted enrichment hybridization capture and sequencing	No need for chloroplast isolation; efficient for multiplexing cp genomes of nearly 100 individuals at a time	Problematic for non-model species	Cronn et al. 2012; Stull et al. 2013; Parks et al. 2012; Mariac et al. 2014

conserved order of chloroplast genes across species. This method is simple and attractive. It relies only on conventional PCR and a number of paired universal primers, and it has been successfully applied to many species such as *Amborella* (Goremykin et al. 2003a), sweet shrub (Goremykin et al. 2003b), water lily (Goremykin et al. 2004), sweet flag (Goremykin et al. 2005), cucumber (Chung et al. 2007), duckweed (Mardanov et al. 2008), bamboos (Wu et al. 2009), orchids (Wu et al. 2010), and grasses (Leseberg and Duvall 2009). However, this method also has some limitations: low transferability of primers between species, need for costly high-fidelity DNA polymerase, and possible base errors introduced by the PCR reaction. The third method, which has been introduced recently, relies on whole cellular DNA, without further isolation of cpDNA, for next-generation sequencing. However, it requires sequencing to 15× or 30× coverage of nuclear genome, which increases the cost (e.g., 7.5 to 15 GB for *Populus*, Huang et al. 2014). The last method is based on targeted enrichment hybridization capture (Cronn et al. 2012), which was originally adopted from human genetics research (Ng et al. 2009). This method is promising, but preparing probes for enrichment of cpDNA is expensive and labor intensive. Only four cases applying this method, two at a multiplex level, have so far been published (Ng et al. 2009; Stull et al. 2013; Parks et al. 2012; Mariac et al. 2014). Up to now, of the 255 cp genomes for which the sequencing method is available, 171 were determined through next-generation sequencing (NGS). Of these 171 cp genomes, 57 % were obtained following cpDNA enrichment, 28 % were obtained by relying on whole genomic DNA, only 15 % used long-range PCR, and less than 1 % were obtained by targeted enrichment hybridization capture, suggesting that isolation of chloroplasts before downstream NGS has been the preferred method for obtaining cp genomes (Fig. 1b).

For non-model forest tree species, exploiting cp genome resources is essential for evolutionary dynamics and conservation studies. In this study, we developed an efficient protocol for cp genome sequencing based on cpDNA enrichment from a very limited amount of leaf material. This protocol consists of five major steps: isolation of intact chloroplasts from fresh leaves in 2-ml tubes, isolation of DNA from intact chloroplasts, rolling circle amplification (RCA) of cpDNA, sequencing of the cp genome by NGS, and assembly of the cp genome by means of a reference-guided assembly (RGA) approach. We successfully adopted this protocol to elucidate the chloroplast genome of a non-model forest tree species, leathery-leaved evergreen oak species *Quercus spinosa* David ex Franchet, which lives on the Tibetan Plateau (Wu and Raven 1999), and annotated the genome with features including the locations and descriptions of genes, conserved regions, and repeats. Our results indicate that this is a rapid and reliable method for complete cp genome sequencing.

Materials and Methods

Plant material

Q. spinosa plants less than 3 years old were collected, together with soil, from Xinyi village, Judian town, Yulong County, Lijiang City in Yunan Province, China. *Q. spinosa* is a common, non-endangered tree species in China. No specific permissions were required for these locations/activities, and we did not sample in any protected areas. The plants were then placed in a growth chamber, with 16-h light/8-h dark cycles

and a temperature of 24 °C with a constant humidity of 65 %, in Beijing Forestry University. Voucher specimens were deposited at the herbarium of Beijing Forestry University, Beijing, China.

Chloroplast isolation and DNA extraction

Prior to chloroplast isolation, plants were grown in the dark for 24–36 h to degrade the starch in the chloroplasts, in order to reduce the possibility of chloroplast breakage brought about by excessive starch (Nobel 1974). After dark treatment, 0.3–0.5 g leaf material was excised from each cultivated seedling. De-veined leaves were then homogenized, using a pre-cooled mortar and pestle, in 2–4 ml pre-chilled isolation buffer (0.33 M sorbitol, 5 mM $MgCl_2$, 1 mM DTT, 5 mM monosodium phosphate, 5 mM disodium phosphate, pH 6.8) containing 0.1 % (*w*/*v*) bovine serum albumin (BSA). We used a double layer of Miracloth to filter the homogenates into a 2-ml tube. Next, we removed intact cells and cell debris by centrifuging the filtered homogenates at 200×*g* for 3 min in an angled rotor (Sigma 4 K15). The supernatant was then transferred into a new tube and centrifuged at 1000×*g* for 7 min to pellet the chloroplasts. We discarded the supernatant and resuspended the pellet in 500 μl isolation buffer containing 0.1 % BSA. The supernatant was firstly loaded carefully onto the top of a 40 % (1 ml) and 80 % (0.6 ml) Percoll gradient. Then the Percoll gradient was centrifuged at 1700×*g* for 6 min to separate Intact chloroplasts. We carefully collected the band at the interface using wide bore tips and resuspended it in three volumes of isolation buffer without BSA. We used a DNAsecure Plant Kit (Tiangen, Beijing), following the manufacturer's instructions, to extract the DNA from intact chloroplasts.

RCA

The entire chloroplast genome was amplified by RCA, following the methods described in Hutchison et al. (2005) with some modifications. Chloroplast DNA was first mixed with Exo-Resistant Random Primer (final concentration 200 μM; Thermo, USA) in annealing buffer (final concentration 40 mM Tris–HCl, pH 8.0; 10 mM $MgCl_2$). The mixture was heated at 94 °C for 3 min and then cooled. Reactions were started by adding 10× phi29 buffer (final concentration 1×; NEB), BSA (final concentration 100 μg/ml), dNTPs (final concentration 1 mM), yeast pyrophosphatase (final concentration 1 U/ml), and phi29 DNA polymerase (final concentration 0.25 U/μl) to the mixture. The reactions were incubated in a thermocycler at 30 °C for 18 h, followed by heating at 65 °C for 10 min to inactivate the phi29 DNA polymerase.

Real-time quantitative PCR

To confirm the enrichment of cpDNA, the concentration of genomic DNA (gDNA), cpDNA in isolated chloroplast (cpDNA_iso), and RCA product of cpDNA (RCA_cpDNA) from *Q. spinosa* were quantified by NanoDrop (NanoDrop Technologies) spectrophotometer. The concentrations of above three samples were 14.1, 4.7, and 300 ng/μl, respectively. The enrichment of chloroplastic DNA was determined by tracking the content of chloroplastic DNA with quantitative real-time PCR (qPCR) with gDNA, cpDNA_iso, and RCA_cpDNA templates. The content of chloroplastic DNA of each sample was calculated by dividing the quantity of the chloroplastic *petB* gene by that of input weight of total DNA template (1/(Cq* input weight of total DNA template)). Two specific primers (*PetB*-RT-F ACGTCTTGAGATTCAGGCGATTGC, *PetB*-RT-R CCTCAGTAACCGTTGGGCGATAGT) were designed for *petB* gene. Reactions were performed using a CFX96 cycler (BioRad, USA) with four technical replicates per sample with SYBR Premix Ex Taq (Takara, Japan). Template DNA was diluted 60-fold for RCA_cpDNA sample for qPCR. The cycling conditions were as follows: 95 °C for 30 s, 40 cycles of 95 °C for 5 s, 55 °C for 30 s, and 72 °C for 30 s. The melting curve was constructed for *petB* to verify the presence of gene-specific peak without primer dimer. The relative amount of chloroplastic DNA in gDNA in unit weight of input template DNA was normalized to 1.

NGS sequencing

We used 5 μg of purified RCA DNA product of cpDNA from a single plant for library preparation. A 101-bp paired-end run was performed on an Illumina-HiSeq 1500 (Illumina, USA) at the gene sequencing platform of the School of Life Sciences, Tsinghua University, China. Briefly, library preparation was performed following the manufacturer's instructions, to give an insert size of 500 bp. Base calling was performed with RTA v.1.6 (Illumina, San Diego, CA, US).

Sequence read assembly

Nucleotides of low quality (probability of error >1 %) were removed, together with all subsequent nucleotides up to the end of each read, using an in-house Perl script. Reads of lengths shorter than 25 bp were then filtered out before further analysis. First, the reads were de novo assembled using SOAPdenovo2 (Luo et al. 2012) with default parameters, except that an insert length of 500 bp was set. K-mer size was optimized as follows: assembly was performed individually using 42 K-mers ranging from 17 to 99; the assembly results of each K-mer were blasted to the *Q. rubra* cp genome; blast results of each k-mer was compared each other and the K-mer

with the best blast result was accepted. Second, we extracted contigs longer than 150 nucleotides to construct the consensus sequence based on our reference-guided assembly method.

Identification of homologs and further contig assembly

BLASTN searches (http://www.ncbi.nlm.nih.gov) were performed with the assembled contigs as query sequences and the *Quercus rubra* chloroplast genome (GenBank ID: NC_020152.1) as reference. Contigs with high scoring pair (HSP) values of identities greater than 80 %, E-values less than 1e-30, and HSP lengths of more than 150 were selected for further assembly with the help of the reference genome.

Contig assembly, sequencing depth calculation, and mapping ratio calculation

The selected contigs were then aligned with the reference sequence to assemble a whole contig. If neighbor contigs had overlaps of over 10 bp at both ends, these two contigs were assembled. However, there were gaps among the contigs. We filled the intervals between two neighbor contigs with N. To resolve these gaps, first we used clean reads that contained ends that were homologous with neighboring contigs, with an overlap of more than 15 bp with over 90 % sequence identity, with the help of Burrows–Wheeler Aligner (BWA) (Li and Durbin 2009). Second, for the remaining gaps, we designed primers upstream and downstream of each gap with Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/). Amplified PCR products were sequenced using an ABI 3730 capillary sequencer (ABI USA) at Ruibo Biotech (Beijing, China). The sequencing results were visualized with Chromas (Technelysium Pty Ltd.) and edited using SeqBuilder (DNASTAR, USA). To confirm the accuracy of our gap filling methodology, we randomly selected 10 gaps and designed primers upstream and downstream of the gaps, amplified the gap regions by PCR, and sequenced them using Sanger sequencing.

To calculate the sequencing depth of cpDNA, reads were mapped to the assembled *Q. spinosa* cp genome by BWA (Li and Durbin 2009) with default values. The output SAM file by BWA was converted to binary alignment map (BAM) file by SAMtools (Li et al. 2009). Converted BAM file was used to calculate the sequencing depth by SAMtools with default parameter. Mapping ratio is calculated by dividing number of mapped reads by the total reads.

As an alternative approach to assemble cp genome of *Q. spinosa*, we also performed assembly using MITObim, which employs a baiting and iteration mapping approach with reference (Hahn et al. 2013). The cp genome of *Q. rubra* (Alexander and Woeste 2014) was used as reference and the default parameter of MITObim was used to assemble *Q. spinosa* cp genome.

Annotation of the cp genome

We used CPGAVAS (Liu et al. 2012) for cp genome annotation. This program uses an input cp genome sequence in FASTA format to identify putative protein coding genes by performing BLASTX searches against a custom database of known cp genomes. The program also provides a circular map of the cp genome, showing the protein-coding genes, transfer RNAs (tRNAs), and ribosomal RNAs (rRNAs) according to the annotation. The flowchart of our improved method for cp genome sequencing is illustrated in Fig. 2.

Repeat analysis

Repeats are categorized into three classes: tandem, palindromic, and diverse. Tandem repeat finder (Benson 1999) with default parameters was used to analyze the numbers and locations of tandem repeats within the *Q. spinosa* chloroplast genome. For palindromic and diverse repeats, we used the program REPuter (http://bibiserv.techfak.uni-bielefeld.de/reputer), in which the minimum cutoff identity between two copies was set to 90 %. The minimum repeat size was set to 30 bp for diverse repeats and 20 bp for palindromic repeats. The gap size between palindromic repeats was restricted to a maximum length of 3 kb. Overlapping repeats were merged into one whenever possible. Redundant results were filtered manually.

Results

Isolation of *Q. spinosa* chloroplast DNA for NGS sequencing

To isolate chloroplasts, all protocols recommend using fresh young leaves that contain only small amounts of the secondary metabolites known to interfere with chloroplast isolation. However, it is not always possible to find plants with young leaves in the natural environment and keep them alive until they are brought back to the lab. Moreover, many evergreen plants, such as *Q. spinosa*, have leathery leaves, which make cpDNA isolation difficult. In addition, isolation of chloroplasts usually requires a large amount of leaf tissue and the use of several 50-ml tubes. To overcome these limitations of the commonly used chloroplast isolation protocols, we tried to reduce the amount of leaf material needed for isolation, making it compatible with the use of 2-ml Eppendorf tubes. We isolated intact chloroplasts from only 0.3–0.5 g leaf material by means of our modified Percoll gradient method. To isolate high-quality chloroplasts, it is crucial to prepare a good Percoll gradient by slowly layering 40 % Percoll above the 80 % Percoll layer (Fig. 2) with wide-mouth tips made by cutting off the bottoms of the tips. In addition, it is important

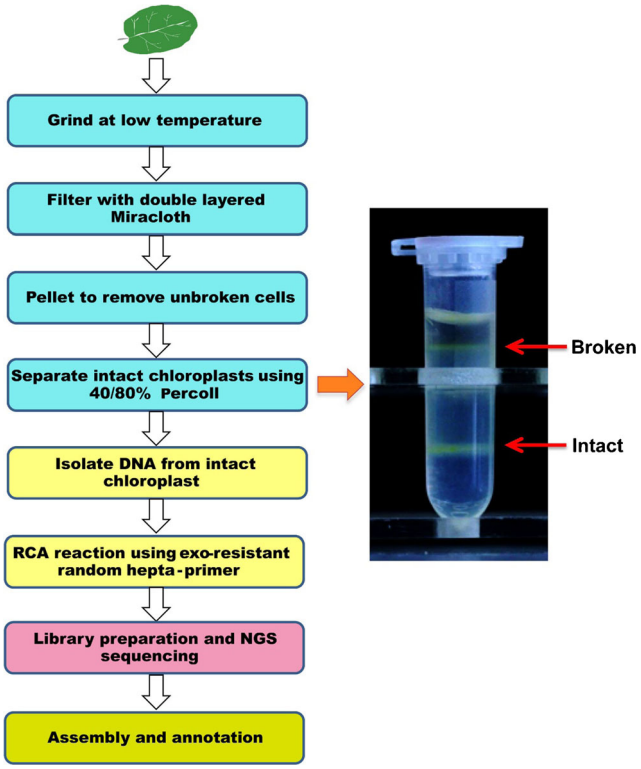

Fig. 2 Flowchart of protocol for isolation of intact chloroplasts followed by NGS sequencing of chloroplast genome. The *right-hand side* of the figure shows the separation of intact chloroplasts from the total chloroplast population by Percoll gradient centrifugation. *Orange arrows* indicate different classes of chloroplast populations

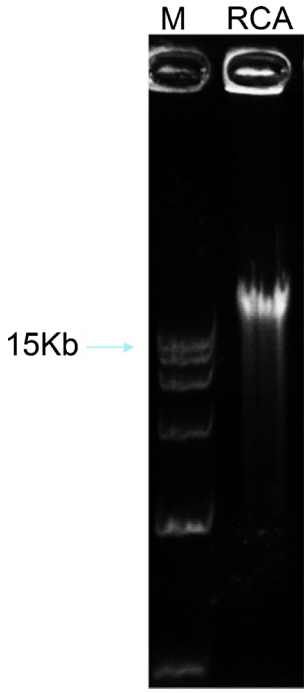

Fig. 3 RCA products of cp genome DNA were assayed by gel electrophoresis. The upper marker band indicates 15 kb. RCA products of cp genome DNA gave a distinct band, indicating large average size and high quantity. "M" represents the 15-kb marker (ladder)

not to disturb the interface. Our intact chloroplast isolation protocol eliminated the need for an expensive high-speed refrigerated centrifuge, which may be unaffordable for small molecular biology labs. This improvement will greatly reduce the cost of instrumentation required for the experiment as a whole. After isolation of intact chloroplasts, we extracted cpDNA using the same method as for extracting whole plant genomic DNA. At the final stage, we used 30 μl distilled water to elute the DNA, producing a solution with a DNA concentration of 7.5 ng/μl. To increase the quantity of DNA while maintaining high fidelity of synthesis, we used phi29 DNA polymerase, which is able to perform strand displacement DNA synthesis from large DNA molecules without dissociating the template (Dean et al. 2002). Starting with 10 ng DNA of intact chloroplast, the concentration was increased to 300 ng/μl after the RCA reaction. We repeated the RCA reaction three times and obtained similar results. Moreover, the majority of amplified DNA fragments were more than 15 kb in length (Fig. 3).

To ensure the enrichment of cpDNA by isolating chloroplast, the relative quantity of cpDNA per unit weight of total DNA in samples of gDNA, cpDNA_iso, and RCA_cpDNA was determined by qPCR (Fig. 4). cpDNA in enriched isolated chloroplast (cpDNA_iso) was about 3.6-fold higher compared to gDNA which was prepared by standard genomic DNA extraction. After RCA, the cpDNA in RCA products (RCA_cpDNA) was 3.1-fold higher compared to gDNA.

For cp genome sequencing, RCA offers several advantages over approaches based on long-range PCR. The phi29 DNA polymerase used in the RCA reaction is extremely processive, enabling the synthesis of long DNA templates up to 70 kb in length (Blanco et al. 1989). In our experiments, the majority of RCA products were over 10 kb. Phi29 DNA polymerase generally has an error rate in the range 10^{-6} to 10^{-8}, which is

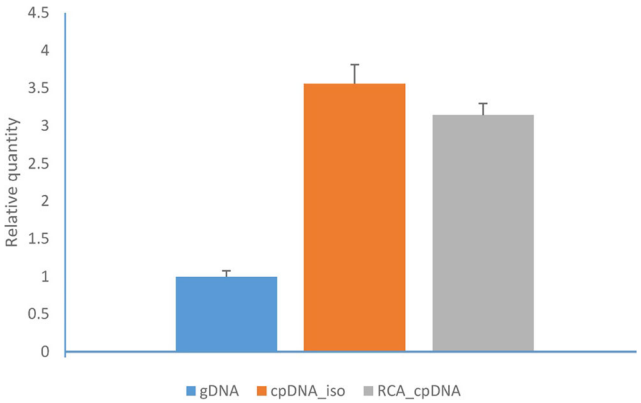

Fig. 4 Enrichment of cpDNA from isolated chloroplast from *Q. spinosa*. The relative quantity of chloroplastic DNA was determined in samples of gDNA, cpDNA_iso, and RCA_cpDNA. qPCR amplification of *PetB* gene was compared to total weight of input DNA template of each sample. The mean of four replicates±SE is shown in fold differences relative to gDNA

comparable to DNA polymerases with 3' proofreading exonuclease activity (Esteban and SalasM 1993; Kunkel 2004). In contrast, long-range PCR usually uses LA *Taq* DNA polymerase with a fidelity of 6.5× as compared to *Taq* DNA polymerase (according to the manufacturer), which has an error rate of 3.7×10^{-5} (Zhou et al. 1991). In addition, RCA can yield a large quantity of DNA regardless of the amount of starting template (Dean et al. 2002).

There have been many attempts to isolate intact chloroplasts for extracting cpDNA, while minimizing contamination with nuclear and mitochondrial DNA, using fresh leaves; typical amounts of tissue used are 20–100 g for grass (Shi et al. 2012; Diekmann et al. 2008), 25 g for three conifer trees (Vieira et al. 2014), 3.8 g for heavenly bamboo, 30.8 g for American sycamore (Moore et al. 2006), and 2.5–5 g for karaka (Atherton et al. 2010). In contrast to the above cases, our modified chloroplast isolation protocol requires only 0.3–0.5 g starting leaf material. To our knowledge, this is the smallest amount of leaf material used to date for intact chloroplast isolation. Although only a very low concentration of cpDNA could be recovered with our protocol, RCA yielded enough DNA for further NGS sequencing.

Sequencing and assembly of the chloroplast genome

Paired-end sequencing of RCA products of *Q. spinosa* chloroplast DNA generated 5,562,830 reads, comprising 561,845,830 bases. After trimming adaptors and removing low-quality sequences with a minimum threshold value of ≤Q20 bases, 5,116,725 reads with an average length of 91 bp were retained for further analysis (Table 2). After de novo assembly by SOAPdenovo2 (Luo et al. 2012) with trimmed reads under a K-mer length of 81, we obtained 2121 contigs with an N50 of 811 bp and mean length of 541 bp; the largest contig was 7173 bp. Following de novo contig assembly, we performed a RGA of the cp genome. Initially, we aligned contigs onto the reference genome (GenBank accession NC_020152) with a cutoff of 80 % identity; 205 contigs met this criterion. Among these contigs, the longest was 4625 bp, the shortest was 152 bp, and the N50 value was 1038 bp with a mean length of 712 bp. For those contigs with overlaps of more than 10 bp, we trimmed the overlaps and filled the gaps with N (the number of N was set according to the reference genome) to build a draft consensus. At this stage, we obtained a draft of the cp genome with 53 gaps. Some gaps were then filled by using clean reads that had at least 15-bp overlap and over 90 % identity with the contig ends. After this process, only one gap remained that was not covered by reads. This single gap was filled by means of PCR. This result suggests that the clean reads obtained from HiSeq1500 sequencing were of high quality and fidelity.

Table 2 Sequencing and assembly results for *Quercus spinosa* cpDNA

Total number of paired-end reads	5562830
Total of nucleotides after cleanup	465,621,980 bp
Mean read length	101 bp
Mean read length (Q20)	91 bp
Average depth of coverage	301×
Total coverage of raw reads on assembled genome	35.5 %
Size of largest contig	7173 bp
N50	1038 bp
Largest contig aligned to cpDNA	4625 bp

Reads were not evenly distributed across the cp genome; there were three regions having a very high coverage of more than 400× (Fig. 5a). One of them (0.5–3.5 kb) is located in the large single-copy (LSC) region (Fig. 5b) and the other two are located in the inverted repeat (IR) regions of cp genome (Fig. 5c). This was partly caused by the fact that there are two repeat regions (inverted repeats) in the cp genome, and reads were used twice in these two regions because we could not distinguish the origin of these reads (Wu et al. 2012). According to our calculations of the overall coverage across the whole cp genome, we achieved an average coverage of 301× (Table 2), which was sufficient to assemble the cp genome (Burger et al. 2007). A total of 1,976,714 reads (35.5 %) could be mapped to the assembled *Q. spinosa* cp genome sequence, suggesting that nearly 65 % of the DNA sequences correspond to nuclear and mitochondrial DNA due to DNA contamination from other cell compartments during chloroplast DNA isolation. The complete cp genome sequence of *Q. spinosa* has been deposited in the GenBank database under the accession number KM841421. The sequencing reads were submitted to SRA in NCBI under the accession number SRP061187.

There are more and more raw reads from released land plant cp genomes; however, computational time cost and hardware requirement have become the main limitations for complete cp genome assembly. To provide alternative approaches for cp genome assembly and compare their features, we also used MITObim (Hahn et al. 2013) to assemble *Q. spinosa* cp genome. When *Q. rubra* cp genome was used as the reference, MITObim gave a 161,464 bp *Q. spinosa* cp genome. This cp genome was compared to that assembled by SOAPdenovo2 using MAFFT web version (http://mafft.cbrc.jp/alignment/server/; Katoh et al. 2002). These two cp genomes have a overall 98.7 % identity with 2159 different sites, of which 1127 sites are SNPs, 61 sites are insertions corresponding to 813 bp and 27 sites are deletions corresponding to 174 bp (Supplementary Figure S1).

General features of the *Q. spinosa* cp genome

The complete cp genome of *Q. spinosa* was determined to be 160,825 bp in length. The gene organization is shown in

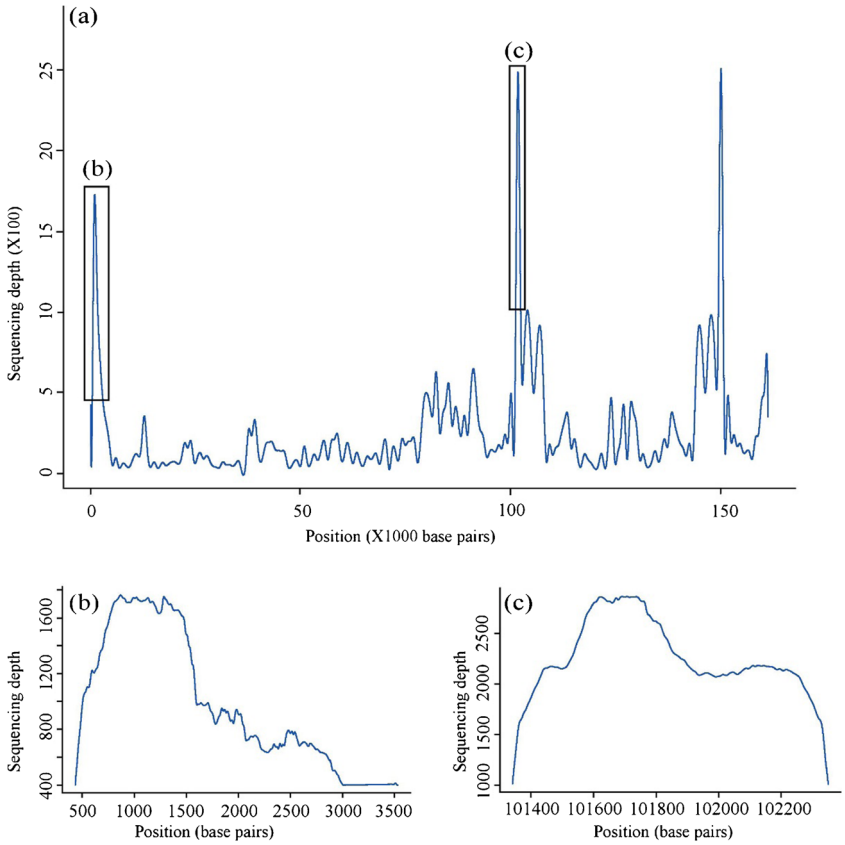

Fig. 5 Sequencing coverage of the *Quercus spinosa* cp genome. **a** Reads were mapped onto the *Quercus spinosa* cp genome sequence. The areas in *two boxes* are shown in detail in (**b**) and (**c**). **b** Enlarged profile of sequencing depth from 0.5 to 3.5 kb. **c** Enlarged profile of sequencing depth from 101.4 to 102.4 kb. The *x*-axis in (**a**), (**b**), and (**c**) represents relative position in the cp genome, and the *y*-axis in (**a**), (**b**), and (**c**) shows read coverage

Fig. 6. It comprises a pair of inverted repeat (IR) regions, each of 25,856 bp, a large single-copy (LSC) region of 90,934 bp, and a small single-copy (SSC) region of 25,854 bp (Table 3). The *Q. spinosa* cp genome encodes 126 genes, of which 16 are in IR regions (duplicated). There are 8 distinct rRNA genes, 29 distinct tRNA genes, and 87 protein coding genes (Table 4). About 57.58 % of the cp genome comprises coding regions. The GC content of the cp genome is 36.89 %. There are 22 intron-containing genes, of which *ycf3* has two introns and the rest have only one intron each (Table 5). Two intron-containing genes (*rpl2* and *ndhB*) are duplicated in the IR regions (Fig. 6). The boundary between the IRa and LSC regions resides between the genes *rps19* and *rpl12*. The IRa/SSC boundary is near the end of the *ndhF* gene. The SSC/IRb boundary resides in the coding region of the *ycf1* gene, resulting in the duplication of the 3' end region of this gene in IRa. The IRb/LSC boundary is between *rpl12* and *trnH* (Fig. 7).

Repeat structure analysis using tandem repeats finder (Benson 1999) identified 23 sets of tandem repeats in the *Q. spinosa* chloroplast genome (Supplementary Table S2). The sizes of these repeats ranged from 26 to 68 bp. Two of the tandem repeats were from the exons of the *ycf2* and *ndhF* genes. The program REPuter (Kurtz and Schleiermacher 1999) identified 109 sets of dispersed repeats and 6 sets of palindromic repeats (Supplementary Table S2). The sizes of these repeats were in the range of 15–30 bp. Of the dispersed and palindromic repeats, only one set of repeats came from the same gene, *ycf1*.

Discussion

Improved methods for the isolation of cpDNA from non-model forest tree species providing extensive plastid genomic resources can be applied to structural, functional, and comparative genomic studies. The method most commonly used for cpDNA isolation is sucrose-gradient-based chloroplast isolation followed by RCA. This requires high-speed centrifugation and is time-consuming, factors which hinder its application for high-throughput analysis. Our modified protocol uses a Percoll gradient and centrifugation at a relatively low speed (<1500*g*), requiring only a standard laboratory centrifuge. Generally, methods for sucrose- or Percoll-based isolation of cpDNA need 20–100 g leaf material. However, we used less than 0.5 g leaf material and obtained enough intact chloroplasts for the subsequent operations. As in all the other methods, we used fresh leaves for cpDNA isolation; dry specimens present a challenge, because their chloroplasts are broken during the drying process and most of the cp-associated DNA is degraded. RCA is a very sensitive way to amplify DNAs from as little as 150 molecules of template

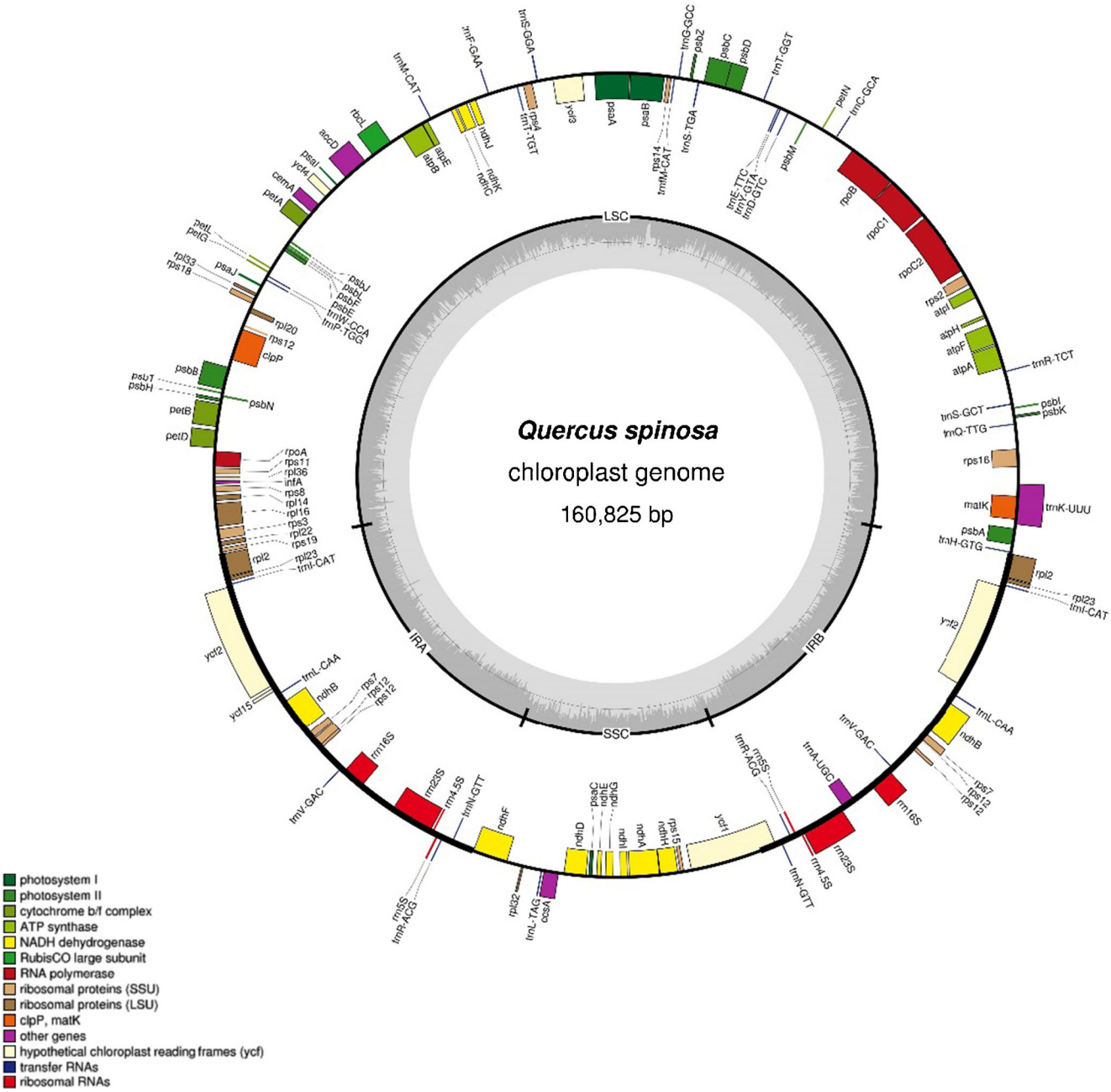

Fig. 6 Annotation of the *Quercus spinosa* cp genome. The annotated 160-KB cp genome of *Quercus spinosa* is represented as concentric *circles*. *Boxes* along the outer circle represent different genes. Two IRs are shown in *black* in the inner circle. LSC and SSC are shown at opposite positions in the inner circle

(Richardson et al. 2003). It should thus be possible to amplify cpDNA from total DNA by RCA using cpDNA-specific primers. However, this hypothesis needs further experimental investigation.

Our results confirmed that this modified protocol is a cost-effective way to sequence whole plastid genomes at the simplex level. Excluding the cost of next-generation sequencing, in this study, we spent a total of just $3 per individual (including $2 per individual for each RCA reaction). Furthermore, it takes only about 3 h to process six samples from the homogenization step up to the RCA reaction preparation step (the RCA reaction in the PCR cycler is not included in this time estimate). One limitation of our protocol is the relatively high cost of next-generation sequencing, although this is decreasing year by year. To overcome this limitation, multiplexing is a good choice, since one lane/run of HiSeq 2500 or MiSeq will generate enough data to determine the chloroplast genomes of many individuals.

Contamination problems are often associated with RCA-based cpDNA amplification (Alexander and Woeste 2014).

Table 3 Summary of features in the complete *Quercus spinosa* cp genome

Feature	Value
Total cpDNA size	160,825 bp
Size of inverted repeat (IR) region	51,722 bp
Size of large single-copy (LSC) region	90,371 bp
Size of small single-copy (SSC) region	18,732 bp
Total number of genes	134
Total number of gene types	113
Number of protein coding genes	87
Number of tRNA genes	36
Number of rRNA genes	8
Number of pseudogenes	2
GC content	~36.87 %
Coding regions (as proportion of whole genome)	~49.7 % (79888)

Compared with Atherton et al.'s (2010) work on the karaka cp genome, which also used the RCA method to enrich cpDNA, we greatly improved the cp read ratio (from 19.6 to 35.5 %). In contrast, studies employing high-throughput sequencing of un-enriched libraries have achieved only 1–12 % of reads mapped to cpDNA (Nock et al. 2011). Although we isolated intact chloroplasts, there was still some nuclear DNA and mitochondrial DNA bound to fragmented chloroplast membranes (Jansen et al. 2005). Some of these contaminating DNA sequences were amplified during the subsequent RCA process, and this reduced the ratio of cpDNA to total DNA. However, due to the larger sizes of the mitochondrial and nuclear genomes, reads from these two genomes were at much lower coverage than the reads from cpDNA. This was an advantage for cp genome assembly. Even though our cpDNA contained high levels of other DNA types, the mean 300× coverage we obtained was much greater than the 30× coverage recommended for cp genome assembly (Straub et al.

Table 4 Genes in the *Quercus spinosa* cp genome

Classification	Gene
RNA genes	
Transfer RNA genes (37)	*trnA-UGC*(x2), *trnC-GCA*, *trnD-GTC*, *trnE-TTC*, *trnF-GAA*, *trnfM-CAT*, *trnG-GCC*(x2), *trnH-GTG*, *trnI-CAT*(x2), *trnK-UUU*, *trnL-CAA*(x2), *trnL-TAG*, *trnM-CAT*, *trnN-GTT*(x2), *trnP-TGG*, *trnQ-TTG*, *trnR-ACG*(x2), *trnR-TCT*, *trnS-GCT*, *trnS-GGA*, *trnS-TGA*, *trnT-GGT*, *trnT-TGT*, *trnV-GAC*(x2), *trnW-CCA*, *trnY-GTA*, *trnI-GAU*(x2), *trnL-UAA*, *trnV-UAC*
Ribosomal RNAs genes (8)	*rrn16S*(x2), *rrn23S*(x2), *rrn4.5S*(x2), *rrn5S*(x2)
Transcription and translation-related genes	
RNA polymerase and related genes (4)	*rpoA*, *rpoB*, *rpoC1*, *rpoC2*
Translation-related gene (1)	*infA*
Ribosomal protein genes	
Large subunit (11)	*rpl14*, *rpl16*, *rpl2*(x2), *rpl20*, *rpl22*, *rpl23*(x2), *rpl32*, *rpl33*, *rpl36*
Small subunit (15)	*rps11*, *rps12*(x3), *rps14*, *rps15*, *rps16*, *rps18*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*(x2), *rps8*
Photosystem-related genes	
Rubisco large subunit gene (1)	*rbcL*
Photosystem I genes (5)	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
Photosystem II genes (15)	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ(ycf9*, *lhb1)*
Assembly/stability of photosystem I (2)	*ycf3*, *ycf4*
Cytochrome b/f complex genes (6)	*petA*, *petB*, *petD*, *petG*, *petL*, *petN*
c-type Cytochrome genes (1)	*ccsA(ycf5)*
Proteasome-like synthase genes (6)	*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*
NADH dehydrogenase genes (12)	*ndhA*, *ndhB*(x2), *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Envelope membrane protein (1)	*cemA(ycf10)*
Acetyl-CoA carboxylase gene (1)	*accD*
ATP-dependent protease subunit (1)	*clpP*
Others	
Maturase (1)	*matK*
Conserved reading frames (ycfs) (4)	*ycf1*(x2), *ycf2*(x2)
Pseudogenes (2)	*ycf15*(x2)

Table 5 Intron-containing genes in the *Q. spinosa* cp genome

Gene	Region	No. of introns	No. of exons	Gene length (bp)	Intron length (bp)	Exon length (bp)
trnK-UUU	complement(join(1853..1887, 4395..4431))	1	2	2579	2507	72
trnG-GCC	join(10549..10571, 11306..11342)	1	2	794	734	60
trnL-UAA	join(52497..52531, 53016..53065)	1	2	569	484	85
trnV-UAC	complement(join(56569..56603, 57234..57271))	1	2	703	630	73
trnI-GAU	join(108331..108367, 109323..109357)	1	2	1027	955	72
trnA-UGC	join(109422..109459, 110261..110295)	1	2	874	801	73
trnI-GAU	complement(join(141840..141874, 142825..142866))	1	2	1027	950	77
rps12	complement(join(103737..103766, 104304..104534, 75526..75639))	2	3	912	537	375
ycf3	complement(join(47325..47477, 48245..48474, 49193..49319))	2	3	1995	1485	510
rps12	join(complement(75526..75639), 146663..146895, 147432..147457)	2	3	373	0	373
clpP	complement(join(75803..76030, 76676..76969, 77821..77889))	2	3	2087	1496	591
petB	join(80836..80841, 81674..82315)	1	2	1480	832	648
petD	join(82524..82532, 83173..83646)	1	2	1144	661	483
rpl16	complement(join(87159..87557, 88661..88669))	1	2	1511	1103	408
rpl2	complement(join(90410..90842, 91544..91937))	1	2	1528	701	827
ndhB	complement(join(100681..101437, 102119..102895))	1	2	2215	681	1534
ndhA	complement(join(126494..126952, 127823..128374))	1	2	1962	951	1011
ndhB	join(148302..149078, 149760..150392)	1	2	2215	805	1410
rps16	complement(join(5484..5711, 6615..6656))	1	2	1173	903	270
atpF	complement(join(13533..13942, 14724..14871))	1	2	1339	781	558
rpoC1	complement(join(23020..24636, 25473..25907))	1	2	2888	836	2052
rpl2	join(159260..159652, 160318..160599)	1	2	1528	853	675

2012). Interestingly, sequencing read depth distribution is not uniform along the whole *Q. spinosa* cp genome; rather, it contains three high-coverage regions more than 400×, of which two typical regions are located in the IR regions. Because IR regions are usually identical and hard to distinguish, reads mapped to the IR regions showed higher coverage than other regions. However, there is one region in the LSC region, although short, 0.5–1.5 kb, that showed extreme high coverage of more than 400×. One possibility is that this region shared very high similarity with fragments from mitochondrial genome or nuclear genome. We checked the presence of this region in mitochondrial genome by blasting it to all the mitochondrial genomes and found that only one mitochondrial genome of *Vitis vinifera* contains a 1-kb region that has a high similarity with this region of *Q. spinosa* cp genome. Since only one land plant mitochondrial genome contains part of this region and without the mitochondrial genome of *Quercus* species, we cannot conclude that reads from mitochondrial genome account for the high coverage of this specific region. Recently, the first oak genome of *Quercus robur* has been released. We blasted this specific region to the genome scaffolds (https://urgi.versailles.inra.fr/blast/) and retrieved three scaffolds containing this region with 98–99 % identity. Since *Q. robur* cp genome has been assembled, we blasted the whole *Q. robur* cp genome to the oak genome scaffolds as a control and retrieved 19 scaffolds containing fragments 4–18 kb with 95–99 % identity with the whole *Q. robur* cp genome. This results suggested that this

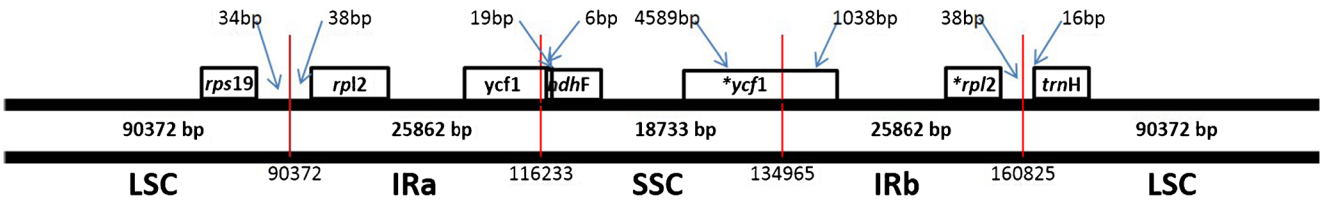

Fig. 7 Detailed views of the positions of the borders of LSC, SSC, and IR in the *Quercus spinosa* cp genome with respect to genes located at or near the boundaries. *Boxes* above the *black line* indicate predicted genes. *Red lines* indicate border positions. *Numbers* on *arrows* indicate the distance between a border and the neighboring gene. The figure shows only the relative positions of borders and neighboring genes; it is not to scale

version of scaffold of oak genome did not exclude the chlroplastic DNA and assembled quality should be improved because it is not likely that a large fragment of cp genome up to 18 kb moves into the nuclear genome. Thus, we still cannot conclude that reads from nuclear genome account for the high coverage of this region with the current oak genome information. Next, we checked whether GC content associated with the coverage differences between regions. We found that the regions in the three peaks showed higher GC content (40.5 % for 0.5–1.5 kb, 42.7 % for two IR regions) than other regions (34.7 % for LSC and 31.2 % for SSC). It has been reported that read coverage was positively related to GC content (when GC content <45 %) when using phi29 DNA polymerase and random primers to perform whole-genome amplification (WGA) (Xu et al. 2014). Our results also support this discovery and could partially explain the high coverage of the three regions.

Assembly method is very important for construction of genomes from raw reads. There are many software and pipelines for assembly with different hardware requirements and output accuracies. In this study, we have developed a pipeline mainly based on SOAPdenovo2 to assemble *Q. spinosa* cp genome. Although SOAPdenovo2 is very popular for genome assemblies, it requires a computer with high-level hardware which costs a lot. To explore the possibilities of using other software to assemble cp genome, we used MITObim (Hahn et al. 2013) to perform the assembly of *Q. spinosa* cp genome with *Q. rubra* cp genome as the reference. These two assembled cp genomes have an overall 98.7 % identity, suggesting that they give a very similar result. However, they also have 2159 different sites including insertions, deletions, and other SNPs. As we extracted chloroplast from leaves of a single plant for HiSeq run, we do not think these are the real SNPs; rather, they represent the assembly differences which may be caused by different algorithms by two pipelines.

The *Q. spinosa* cp genome is 160,825 bp in length, smaller than the cp genomes of *Q. rubra* (161,304 bp; Alexander and Woeste 2014), *Q. robur* (161,295 bp; Alexander and Woeste 2014), and *Quercus aliena* (160,921 bp; Lu et al. 2015); thus, to date, it is the smallest oak cp genome reported. Overall, the structure, gene order, GC composition, and intron content of the *Q. spinosa* cp genome are similar to those of other land plant cp genomes (Table 3). Unlike many angiosperm taxa, *ycf1* was found to be functional in the *Q. spinosa* cp genome. A functional *ycf1* was also identified in *Castanea mollissima*, *Prunus persica*, and *Q. rubra* (Jansen et al. 2011; Alexander and Woeste 2014). *ycf15* was first identified as ORF87 in tobacco (Shinozaki et al. 1986) and appears to be potentially functional as a protein-coding gene. However, with the increasing number of complete cp genomes being released, more and more *ycf15* has been annotated in land plant cp genomes; some of them have multiple premature stop codon that would lead to truncated proteins or even short peptides (Goremykin et al. 2003a; 2003a; b; Steane 2005; Schmitz-Linneweber et al. 2001). Currently, there are only few cp genomes from *Quercus* genus that have been released. Previous report showed that *ycf15* may be functional in *Q. rubra* (Alexander and Woeste 2014); however, *ycf15* has become a pseudogene in *Q. spinosa* (this study) and *Q. aliena* (Lu et al. 2015) by incorporate multiple premature stop codons. This feature suggests that *ycf15* is highly variable in *Quercus* species and has potential to be used as a barcode gene.

The Fagaceae is a large angiosperm family comprising more than 900 species mainly distributed in the Northern Hemisphere across Europe, North America, and Asia (Kremer et al. 2012). The Fagaceae species are considered as keystone species in forest ecosystem and are important drivers of terrestrial biodiversity. *Quercus* L., the largest genus in Fagaceae family, comprises ~400 species; most of the diversity of the genus lies in Central America and Southeast Asia (Denk and Grimm 2010). However, till now, extensive genetic and genomic studies are conducted mainly on less species-rich European oaks (e.g., Kremer and Petit 1993; Kremer et al. 2010; Plomion et al. 2015). The annotated cp genome reported here is a valuable genetic resource for providing new insights into the population genetics, conservation genetics, and evolutionary biology of oaks, a non-model forest tree species in large.

Conclusions

We have developed a protocol for improved chloroplast isolation and cpDNA extraction followed by amplification and used it to determine the genome sequence of a non-model forest tree species. This protocol necessitates reduced amounts of leaf material (down to less than 0.5 g) for chloroplast isolation while maintaining high sequencing quality, providing a useful tool for the sequencing of complete cp genomes. This fast, efficient, and economical method should be useful for exploiting complete cp genome from non-model forest tree species.

Acknowledgments The authors thank Dr. Rémy J. Petit, Dr. Antoine Kremer working in INRA Pierroton, France, Dr. Liuyang Wang working in Duke University, USA, and Dr. Saneyoshi Ueno working in Forestry and Forest Products Research Institute, Japan, for revision of, and suggestions about, the preliminary version of this paper. The authors thanks the comments and suggestions from three anonymous reviewers. The research was funded by Beijing Nova Program (grant number: Z151100000315056), National Natural Science Foundation of China (grant number 41201051; 41430749), 111 Project (grant number B13007), and Program for Changjiang Scholars, Innovative Research Team in University (grant number IRT13047) to FKD and the Major projects on control and rectification of water body pollution (2012ZX07105-002-03) to JL

Data Archiving Statement The *Q. spinosa* cp genome sequence data has been deposited into GenBank and released to public under the accession number KM841421.1. The sequencing reads were submitted to SRA in NCBI under the accession number SRP061187.

References

Alexander LW, Woeste KE (2014) Pyrosequencing of the northern red oak (*Quercus rubra* L.) chloroplast genome reveals high quality polymorphisms for population management. Tree Genet Genomes 10:803–12

Aslan CE, Zavaleta ES, Tershy B, Croll D (2013) Mutualism disruption threatens global plant biodiversity: a systematic review. PLoS One 8:e66993. doi:10.1371/journal.pone.0066993

Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ (2010) Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods 6:22

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucl Acids Res 27:573

Blanco L, Bernad A, Lázaro JM, Martin G, Garmendia C, Salas M (1989) Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. J Biol Chem 264:8935–8940

Burger G, Lavrov DV, Forget L, Lang BF (2007) Sequencing complete mitochondrial and plastid genomes. Nat Protoc 2:603–614

Chung SM, Gordon VS, Staub JE (2007) Sequencing cucumber (*Cucumis sativus* L.) chloroplast genomes identifies differences between chilling-tolerant and -susceptible cucumber lines. Genome 50:215–225

Cronn R, Liston A, Parks M et al (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36:122–122

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV et al (2012) Targeted enrichment strategies for next-generation plant biology. Am J Bot 99:291–311

Dean FB, Hosono S, Fang L, Wu X, Faruqi FA, Bray-Ward P et al (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 99:5261–5266

Denk T, Grimm GW (2010) The oaks of western Eurasia: traditional classifications and evidence from two nuclear markers. Taxon 59: 351–366

Diekmann K, Hodkinson TR, Fricke E, Barth S (2008) An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. PLoS One 3:e2813. doi:10.1371/journal.pone.0002813

Esteban JA, SalasM BL (1993) Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. J Biol Chem 268:2719–2726

Goremykin VV, Hirsch-Ernst KI, Wölfl S et al (2003a) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol Biol Evol 20:1499–1505

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003b) The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis*—structural and phylogenetic analyses. Plt Syst Evol 242:119–135

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol 21:1445–1454

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

Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. Nucleic Acids Res 41:9

Huang DI, Hefer CA, Kolosova N, Douglas CJ, Cronk QCB (2014) Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). New Phytol 204:693–703

Hutchison CA, Smith HO, Pfannkoch C, Venter JC (2005) Cell-free cloning using φ29 DNA polymerase. Proc Natl Acad Sci U S A 102:17332–17336

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

Jansen RK, Saski C, Lee SB, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of rpl22 to the nucleus. Mol Biol Evol 28:835–847

Katoh K, Misawa K, Kuma KI, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066

Kremer A, Petit RJ (1993) Gene diversity in natural populations of oak species. Ann Sci For 50:186s–203s

Kremer A, Sederoff R, Wheeler N (2010) Genomics of forest and ecosystem health in the Fagaceae: meeting report. Tree Genet Genome 6:815–820

Kremer A, Abbott AG, Carlson JE et al (2012) Genomics of Fagaceae. Tree Genet Genomes 8:583–610

Kunkel TA (2004) DNA replication fidelity. J Biol Chem 279:16895–16898

Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426–427

Leseberg CH, Duvall MR (2009) The complete chloroplast genome of *Coix lacryma-jobi* and a comparative molecular evolutionary analysis of plastomes in cereals. J Mol Evol 69:311–318

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079

Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13:715

Lu S, Hou M, Du FK, Li J, Yin K (2015) Complete chloroplast genome of the Oriental white oak: *Quercus aliena* Blume. Mitochondrial DNA (ahead-of-print) 1–3

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18

Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS, Kolganova TV et al (2008) Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. J Mol Evol 66:555–564

Mariac C, Scarcelli N, Pouzadou J, Barnaud A, Billot C, Faye A et al (2014) Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour 14:1103–1113

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM et al (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol 6:17

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276

Nobel PS (1974) Rapid isolation techniques for chloroplasts. Meth Enzym 31:600–606

Nock CJ, Waters DL, Edwards MA, Bowen SG, Rice N, Cordeiro GM et al (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9:328–333

Parks M, Cronn R, Liston A (2012) Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). BMC Evol Biol 12:100

Plomion C, Aury JM, Amselem J et al (2015) Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Mol Ecol Resour. doi:10.1111/1755-0998.12425

Richardson PM, Detter C, Schweitzer B et al (2003) Practical applications of rolling circle amplification of DNA templates. Genet Eng 25:51–63

Schmitz-Linneweber C, Maier RM, Alcaraz JP, Cottet A, Herrmann RG, Mache R (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plt Mol Biol 45:307–315

Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5:16–18

Shi C, Hu N, Huang H, Gao J, Zhao YJ, Gao LZ (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, Hayashida N, Matsubayashi T et al (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Steane DA (2005) Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae). DNA Res 12:215–220

Straub SC, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot 99:349–364

Stull GW, Moore MJ, Mandala VS et al. (2013) A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. Appl Plant Sci 1:apps.1200497. doi:10.3732/apps.1200497

Vieira Ldo N, Faoro H, Fraga HP, Rogalski M, de Souza EM, de Oliveira Pedrosa F et al (2014) An improved protocol for intact chloroplasts and cpDNA isolation in conifers. PLoS ONE 9:e84792. doi:10.1371/journal.pone.0084792

Wu Z, Raven P (1999) Flora of China. Vol. 4 (Cycadaceae through Fagaceae). Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis

Wu FH, Kan DP, Lee SB, Daniell H, Lee YW, Lin CC et al (2009) Complete nucleotide sequence of *Dendrocalamus latiflorus* and *Bambusa oldhamii* chloroplast genomes. Tree Physiol 29:847–856

Wu FH, Chan MT, Liao DC, Hsu CT, Lee YW, Daniell H et al (2010) Complete chloroplast genome of *Oncidium Gower* Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol 10:68

Wu J, Liu B, Cheng F, Ramchiary N, Choi SR, Lim YP, Wang XW (2012) Sequencing of chloroplast genome using whole cellular DNA and solexa sequencing technology. Front Plant Sci 3:243

Xu B, Li T, Luo Y, Xu R, Cai H (2014) An Empirical Algorithm for Bias Correction Based on GC Estimation for Single Cell Sequencing. Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing: 15–21

Zhou YH, Zhang XP, Ebright RH (1991) Random mutagenesis of gene-sized DNA molecules by use of PCR with *Taq* DNA polymerase. Nucleic Acids Res 19:6052