

Next-Generation Technologies to Determine Plastid Genome Sequences

Robert J. Henry, Nicole Rice, Mark Edwards, and Catherine J. Nock

Abstract

Sequencing of chloroplast genomes is a key tool for analysis of chloroplasts and the impact of manipulation of chloroplast genomes by biotechnology. Advances in genome sequencing allow the complete sequencing of the chloroplast genome and assessment of variation in the chloroplast genome sequences within a plant. Isolation of chloroplast DNA has been a traditional starting point in these analyses, but the capacity of current sequencing technologies allows effective analysis of the chloroplast genome sequence by shotgun sequencing of a preparation of total DNA from the plant. Chloroplast insertions in the nuclear genome can be distinguished by their much lower copy number. Short-read sequences are best assembled by alignment with a reference chloroplast genome.

Key words DNA sequencing, Genome assembly, Next-generation sequencing, NGS, Plastid genome

1 Introduction

Traditional Sanger sequencing has routinely been applied to the analysis of specific chloroplast sequences following PCR amplification. The whole chloroplast sequence provides a more complete analysis but was beyond the scope of these techniques. Next-generation sequencing allows the generation of far greater volumes of sequence data at low cost making routine whole chloroplast genome sequencing feasible for the first time. Sequencing of the whole chloroplast genome has usually involved isolation of chloroplasts and preparation of DNA from purified chloroplasts [1]. Obtaining highly purified chloroplast DNA samples has been difficult, especially for some species. An alternative simplified strategy to obtain a whole chloroplast genome sequence has been to use whole chloroplast genome amplification to prepare samples for sequencing.

An efficient strategy avoiding both chloroplast isolation and chloroplast genome amplification is extraction of chloroplast genome sequences from shotgun whole genome sequences of total plant DNA [2]. This is a simple and cost-effective option

producing reliable data on chloroplast sequence and variants within the sample.

The chloroplast genome has been sequenced for a growing number of plant species [3]. As of August 2012, the NCBI Organelle Genome Resource contained 285 records of complete eukaryota plastid genomes. These sequences provide reference genome sequences allowing analysis of variants and modified chloroplast genomes by re-sequencing using efficient next-generation technologies. Longer sequence reads may allow de novo assembly of the chloroplast genome as sequencing technology improves [4].

2 Materials

2.1 Sample Preparation

1. QIAGEN TissueLyser.
2. 96 deep well collection plate.

2.2 DNA Preparation

1. Qiagen DNeasy 96 Plant Kit.

2.3 DNA Quality Analysis/Assurance

1. Agarose gel electrophoresis kit.

2.4 Quantification of dsDNA

1. Quant-iT™ PicoGreen(R) dsDNA Reagent and Kits.
2. Qubit® 2.0 Fluorometer.

2.5 DNA Sequencing

1. Covaris S2 DNA shearing device.
2. Agilent 2100 Bioanalyzer and DNA 1000 chip.
3. Agarose gel electrophoresis kit.
4. QIAquick Gel Extraction kit.
5. QIAquick PCR Purification Kit.
6. Illumina Genome Analyzer (GAIIx).
7. Illumina software Pipeline 1.4.

2.6 Data Analyses

1. Computer with internet access.
2. CLC Genomics Workbench v.3.6.5 software package (<http://www.clcbio.com>).
3. DNA sequence editing program (e.g., Sequencher, <http://www.genecodes.com>).

3 Methods

3.1 Sample Preparation

Harvested leaf tissue directly into the sample tubes of a 96 deep well collection plate and grind in a TissueLyser as per the recommendations

in the Qiagen DNeasy 96 Plant Kit (<http://www.qiagen.com/literature/handbooks/literature.aspx?id=1000061>) (see **Note 1**).

3.2 DNA Preparation

A Qiagen DNeasy 96 Plant Kit yields very high-quality DNA; however, the yield is very species specific. The method should be tested for the species you are working with first. For whole genome shotgun sequencing, you may need to increase the number of replicate DNA extractions for each sample to achieve the total mass of DNA required. For species in the genus *Oryza* [2], the manufacturer's method was followed (<http://www.qiagen.com/literature/handbooks/literature.aspx?id=1000061>), and four replicates of each sample were required.

3.3 DNA Quality Analysis/Assurance

Extracted DNA samples should be quantified and the quality checked using an agarose gel checking for excessive smearing or fragmentation. High-quality starting material is important for library generation. The DNA sample should be pure, having an OD 260/280 ratio of between 1.8 and 2. It is important to quantitate the input material carefully, preferably using fluorescence-based quantitation methods (such as Invitrogen PicoGreen or Qubit™ Quantitation Fluorometer assays).

3.4 Quantification of dsDNA Using the Invitrogen PicoGreen Reagent Kit

For quantification, the following protocol has been adapted from the procedure supplied with the Molecular Probes PicoGreen dsDNA quantitation kit. The PicoGreen reagent is used as an ultrasensitive fluorescent stain for quantitating double-stranded DNA (dsDNA) in solution. The protocol described here is suitable for quantitating dsDNA concentrations ranging from 1 ng/μL to 200 ng/μL of DNA sample. For further information on limitations and compounds that may interfere with the assay, please refer to Molecular Probes product information sheets available from the following: <http://probes.invitrogen.com/>.

1. Preparation of 1× TE buffer. The kit is supplied with 20× TE buffer. To prepare a 1 mL working solution of 1× TE buffer, mix 50 μL 20× TE buffer with 950 μL sterile Milli-Q water. To calculate the total volume of 1× TE required, use the following equation:

$$\text{Total volume } 1 \times \text{TE} = 200 \mu\text{L} \times (S_{a_n} + [P_{l_n} \times (St_n + 2)]) + 1,500 \mu\text{L}$$

where St_n = total number of points within the standard curve including the standard blank, S_{a_n} = total number of samples, and P_{l_n} = total number of plates to be run consecutively.

Example, for 50 samples, the total volume of 1× TE required = 200 μL 1× TE working reagent × (50 samples + [1 plate × (7 standards + 2 waste allowance)]) + 1,500 μL) for making up λ DNA working stock solutions = 13.4 mL.

Table 1
High-range λ dsDNA standard curve

TE (μ L)	2 ng/ μ L λ DNA (μ L)	dsDNA ^a (ng)
100	0	Blank
100	0	0
90	10	20
80	20	40
70	30	60
50	50	100
0	100	200

^aQuantity of dsDNA added to 200 μ L assay volume

Table 2
Low-range λ dsDNA standard curve

TE (μ L)	200 pg/ μ L λ DNA (μ L)	dsDNA ^a (ng)
100	0	Blank
100	0	0
90	10	2
80	20	4
70	30	6
50	50	10
0	100	20

^aQuantity of dsDNA added to 200 μ L assay volume

2. Preparation of λ DNA working stock solutions. The lambda DNA stock supplied with the kit (component C) contains 100 μ g λ dsDNA/mL.

To make a 2 ng/ μ L λ DNA working stock for a high-range standard curve (Table 1), combine 10 μ L of the stock λ dsDNA with 490 μ L 1 \times TE.

To make a 200 pg/ μ L λ DNA working stock for the low-range standard curve (Table 2), carry out a two-step dilution as follows:

Step 1—prepare a 2 ng/ μ L λ DNA solution by combining 10 μ L of the stock λ dsDNA with 490 μ L 1 \times TE (final concentration 1 ng/ μ L λ DNA).

Step 2—prepare a 200 pg/ μ L λ DNA working stock by combining 100 μ L of the 2 ng/ μ L λ DNA solution with 900 μ L 1 \times TE.

3. Preparation of PicoGreen working solution. The PicoGreen working solution is prepared by making a 200-fold dilution of the concentrated stock (supplied with kit) in 1× TE. To make 1 mL of PicoGreen working solution, mix 5 µL concentrated stock with 995 µL 1× TE buffer (*see Note 2*).

To calculate the total volume of PicoGreen working solution required, use the following equation:

$$\text{Total volume PicoGreen working solution} = 100 \mu\text{L} \times \left(\text{Sa}_n + \left[\text{Pl}_n \times (\text{St}_n + 2) \right] \right)$$

where Sa_n = total number of samples, St_n = total number of points within the standard curve including the standard blank, and Pl_n = total number of plates to be run consecutively.

Example, for 50 samples, the total volume of PicoGreen working solution = 100 µL PicoGreen working reagent × (50 samples + [1 plate × (7 standards + 2 waste allowance)]) = 5,950 µL.

4. To prepare a dsDNA standard curve, dispense the appropriate volumes (see Tables 1 and 2 above) of 1× TE and λ DNA working stock into each microplate well.
5. To prepare dsDNA samples, add 2 µL DNA sample and 98 µL TE to each well (*see Note 3*).
6. Aliquot 100 µL PicoGreen working solution into each microplate well (standards and samples).
7. Incubate plate for 5 min at RT protected from light.
8. Measure the fluorescence in a spectrophotometer, with fluorescence and microtiter plate capability. Filters with the following wavelength parameters should be used: excitation is 485/20 nm and emission is 528/20 nm.
9. To determine the concentration of dsDNA in 1 µL of the sample, divide the result from the assay by the volume of sample added to the assay. For example, if the DNA mass shown in the assay results is 100 ng and 2 µL was used in the assay, 100 ng/2 = 50 ng dsDNA/µL sample.

3.5 DNA Sequencing

Several platforms for high-throughput sequencing can be applied to chloroplast sequencing. These include those produced by Applied Biosystems (solid), Illumina, and 454 Roche. The protocol presented here utilizes an Illumina sequencing platform. Modifications to sample preparation and data handling would be necessary for other sequencing platforms. A method using the Roche GS FLX platform is presented by Yang et al. [5].

General steps for short-read NGS platforms involve:

1. Random shearing of DNA, either via sonication or nebulization

2. Ligation of universal adapters at both ends of the DNA fragments
3. Immobilization and amplification (to improve signal intensity) of the adapter-flanked fragments

This generates clustered amplicons to serve as templates for the sequencing reactions. Through alternating cycles of base incorporation and image capture, these platforms produce highly accurate short DNA sequences which range in size from 25 to 150 bases.

For Illumina sequencing, around 1–3 µg of total DNA was sheared and prepared following the manufacturer's instructions (Illumina sample preparation for paired-end sequencing) with the following modifications. DNA was sheared using the adaptive focused acoustics method using a Covaris S2 device as follows: duty cycle 10 %, intensity 5, and cycles per burst 200 for 180 s at 6 °C. DNA quantity and quality of fragmentation were assessed with a DNA 1000 chip on an Agilent 2100 Bioanalyzer. Ligation products were purified by agarose gel electrophoresis (2 % agarose, 120 V for 120 min). A narrow size range of predominantly 300-bp fragments was excised from the gel and the products isolated with a QIAquick Gel Extraction kit without heating (alternatively Invitrogen E-Gels SizeSelect™ agarose gels are also effective). PCR products were further purified with a QIAquick PCR Purification Kit and quantified again using a DNA 1000 chip. Approximately 4 pmol per individual and 3 pmol of the PhiX control lane were sequenced for 36×2 cycles on an Illumina Genome Analyzer (GAIIx) following the manufacturer's instructions. Base calling was performed with Illumina software Pipeline 1.4. Multiplexed sequencing using indexing may also be considered as an option for obtaining chloroplast sequences.

3.6 Data Analysis

1. Import FASTQ files [6] from Illumina paired-end sequencing run to CLC Genomics Workbench (*see Note 4*).
2. Trim paired-end and single-read sequences to remove low-quality data and adaptor sequences (trim sequence option). A quality score Q of 20 corresponds to an expected base call error of $P=0.01$. Filter to remove sequence reads below a specified length after trimming.
3. Identify suitable reference chloroplast genome sequence from a close relative (ideally the same species or genus) for read mapping. Where a close relative is unavailable or the target species is unknown, de novo assembly can be used to identify the closest published chloroplast genome sequence for reference mapping (*see Note 5*).
4. Export reference sequence from DNA sequence database (e.g., GenBank) and import into CLC Genomics Workbench.

5. Assemble trimmed paired-end and single-read sequences by mapping against reference sequence (map reads to reference option).
6. Settings for short and long read parameters, such as mismatch and deletion/insertion cost, will depend on how closely related the target and reference sequences are. To facilitate genome assembly, less stringent setting is required for more distantly related species.
7. Set match mode to random to enable assembly of both inverted repeat regions and repetitive elements.
8. Set conflict resolution mode to vote majority to avoid SNP calling in the consensus sequencing resulting from alignment of nuclear and mitochondrial reads to the chloroplast reference.
9. Examine the consensus sequence produced by read mapping to identify gaps—regions with zero coverage (*see Note 6*).
10. Export the consensus sequence including gaps in FASTA format.
11. Import consensus sequence to a sequence editing program (e.g., Sequencher, <http://www.genecodes.com>). Where gaps cannot be resolved as indels, report gaps as N rather than dash (–) in the sequence.
12. Annotate final chloroplast genome sequence using the program DOGMA [7].

4 Notes

1. Plant tissue was collected preferably as fresh leaf tissue. DNA from collections used for other analyses or from DNA banks may be used as a source of DNA if the quality and quantity of DNA are found to be acceptable for analysis. Samples should be processed while still fresh or extracted from carefully dried or frozen material.
2. The PicoGreen working solution must be used on the day it is prepared, preferably within 4 h. Do not make up in a glass container, as there is a risk that the PicoGreen reagent will bind to the glass.
3. Dilution can be adjusted if DNA concentration is expected to be lower or higher. For example, in high DNA concentration 1 μL + 99 μL TE or 5 μL + 95 μL TE, if DNA concentration is expected to be very low or if the sample volume needs to be conserved, the sample can be diluted to a lower concentration before being assayed.

4. Numerous bioinformatics programs are available for assembly of NGS sequence reads [8]. The method presented in this chapter used the commercial software package CLC Genomics Workbench v.3.6.5 (<http://www.clcbio.com>).
5. When de novo assembly is required to identify a reference chloroplast genome sequence, sort the resulting contigs by length. Select the longest contigs with the highest median coverage. These will invariably include the chloroplast sequence due to the abundance of chloroplast reads in extracts from total DNA. Perform BLAST searches (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) on these contigs to identify the closest available chloroplast genome sequence.
6. Gaps may be due to insertion/deletions, highly divergent sequence regions (particularly where a distant reference genome is used) or inversions. Sanger sequencing can be used to confirm sequence reads in these regions.

Acknowledgments

The authors thank the Australian Plant DNA Bank and Southern Cross Plant Genomics for assistance with samples and sequencing.

References

1. 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
2. Nock CJ, Waters DLE, Edwards MA, Bowen S, Rice N, Cordeiro GM, Henry RJ (2010) Chloroplast genome sequence from total DNA for plant identification. *Plant Biotechnol J* 9:328–333
3. Li X, Gao H, Song JY, Wang JY, Henry R, Wu HZ, Hu ZG, Yao H, Luo HM, Luo K, Pan HL, Chen AL (2012) Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species. *Sci China Ser C* 56:189–198
4. McPherson H, van der Merwe M, Delaney SK, Edwards MA, Henry RJ, McIntosh E, Rymer PD, Milner ML, Siow J, Rossetto M (2013) Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecol* 13:8
5. Yang M, Zhang X, Liu G, Chen K, Yun Q, Zhao D, Al-Mssallem IS, Yu J (2010) The complete chloroplast genome sequence of the date palm (*Phoenix dactylifera* L.). *PLoS One* 5:e12762
6. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38:1767–1771
7. Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252–3255
8. Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95:315–327

Chloroplast Biotechnology

Methods and Protocols

Maliga, P. (Ed.)

2014, XV, 452 p. 90 illus., 48 illus. in color., Hardcover

ISBN: 978-1-62703-994-9

A product of Humana Press