Chloroplast genome of Aconitum barbatum var. puberulum (Ranunculaceae) derived from CCS reads using the PacBio RS platform

Xiaochen Chen, Qiushi Li, Ying Li, Jun Qian and Jianping Han*

Center for Computational Biology and Bioinformatics, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

The chloroplast genome (cp genome) of Aconitum barbatum var. puberulum was sequenced using the third-generation sequencing platform based on the single-molecule real-time (SMRT) sequencing approach. To our knowledge, this is the first reported complete cp genome of Aconitum, and we anticipate that it will have great value for phylogenetic studies of the Ranunculaceae family. In total, 23,498 CCS reads and 20,685,462 base pairs were generated, the mean read length was 880 bp, and the longest read was 2,261 bp. Genome coverage of 100% was achieved with a mean coverage of 132× and no gaps. The accuracy of the assembled genome is 99.973%; the assembly was validated using Sanger sequencing of six selected genes from the cp genome. The complete cp genome of A. barbatum var. puberulum is 156,749 bp in length, including a large single-copy region of 87,630 bp and a small single-copy region of 16,941 bp separated by two inverted repeats of 26,089 bp. The cp genome contains 130 genes, including 84 protein-coding genes, 34 tRNA genes and eight rRNA genes. Four forward, five inverted and eight tandem repeats were identified. According to the SSR analysis, the longest poly structure is a 20-T repeat. Our results presented in this paper will facilitate the phylogenetic studies and molecular authentication on Aconitum.

Keywords: chloroplast genome, Aconitum, circular consensus sequencing, PacBio RS, the third generation sequencing

INTRODUCTION

Aconitum barbatum var. puberulum (Niubian) belongs to the Aconitum subgenus Lycoctonum (Ranunculaceae) and most species of Lycoctonum are low-temperature resistant. However, aconitine is a kind of highly toxic alkaloid, which mainly exists in the plants of Aconitum. Identification and phylogeny studies of Aconitum and the Ranunculaceae family thus are particularly important (Xiao et al., 2005). He et al. (2010) applied the chloroplast genome (cp genome) intergenic region psbA-trnH as a barcode to identify 19 species in Aconitum, and Johansson (Johansson, 1995) used chloroplast DNA restriction site variation among 31 genera of the Ranunculaceae to conduct phylogenetic analyses. However, more in-depth studies of the cp genome are needed.

Chloroplasts possess their own genome and genetic system, which plays an important role in photosynthesis. The first chloroplast genome to be sequenced was that of Nicotiana tabacum, which heralded a new age of chloroplast studies in photosynthesis, phylogenetic biology, evolutionary biology and even chloroplast genetic engineering (Shinozaki et al., 1986; Hiratsuka et al., 1989; Daniell et al., 1998; Pfannschmidt et al., 1999; Moore et al., 2010; Wu et al., 2012). Some researchers (Chen et al., 2014; Li et al., 2014b) advocated the cp genome as a new DNA barcode to distinguish closely related plants. The typical cp genome structure of higher plants is circular with a length of 120–160 kb, containing approximately 130 genes (Sugiura, 1992). Two inverted repeats (IRs), a large single-copy region (LSC), and a small single-copy region (SSC) constitute the complete cp genome (Zhang et al., 2012). With the development of the next-generation sequencing technology, increasing numbers of species have been sequenced, including duckweed, palm, and others (Jordan et al., 1996; Uthaipaisanwong et al., 2012). Although the interest in the cp genome has increased in the past few decades, with 486 complete cp genome sequences deposited in GenBank (By 2014-7-6), there are still challenges and opportunities to develop a simple and rapid method for sequencing cp genomes. One common strategy is the use of a complete set of universal primers to amplify an entire cp genome and then perform the sequencing (Dong et al., 2013). Another frequently used strategy is “whole-genome sequencing”, which uses the total genome DNA to recover the cp genome through massively parallel sequencing (McPherson et al., 2013). This strategy is quite simple and effective, particularly as the cost of high-throughput sequencing decreases. In the present study, we used purified chloroplast DNA as the template for sequencing with the aim of developing a practical strategy involving the use of multiple samples to sequence the cp genome on the PacBio RS platform.

The third-generation PacBio system is based on the single-molecule real-time (SMRT) sequencing approach (Eid et al., 2009). Second-generation sequencing introduced a novel, rapid method for whole-genome sequencing (Mardis, 2008a,b; Metzker,
A. barbatum var. genomes. sequencing strategy is a viable option for rapidly sequencing cp sequencing platform. Our results demonstrate that the SMRT CCS Aconitum completed cp genome of genome. inate the need to fill in the gaps that exist in the assembled be used for rapid sequencing of whole cp genomes and elim-

rate. A PacBio-only assembly could be completed without the need to construct specialized fosmid libraries or other similar assemblies using second-generation sequencing technologies. We sought to investigate whether third-generation sequencing could be used for rapid sequencing of whole cp genomes and eliminate the need to fill in the gaps that exist in the assembled genome.

In the present study, we report the completed cp genome of A. barbatum var. puberulum. To our knowledge, this is the first completed cp genome of Aconitum using the third-generation sequencing platform. Our results demonstrate that the SMRT CCS sequencing strategy is a viable option for rapidly sequencing cp genomes.

MATERIALS AND METHODS
CHLOROPLAST DNA ISOLATION, SEQUENCING, ASSEMBLY, AND VALIDATION
Fresh leaves were collected from Donglingshan Mountain, Beijing. Total cpDNA was extracted from approximately 100 g fresh leaves using a sucrose gradient centrifugation method that was described by Li et al. (2012). A total of 700 ng cp genomic DNA was sheared to a target size of 2 kb in an AFA clear mini-tube using a Covaris S2-focused ultrasonicator (Covaris Inc.) to construct the libraries according to the Pacific Biosciences SMRT Sequencing Covaris S2-focused ultrasonicator (Covaris Inc.) to construct the libraries according to the Pacific Biosciences SMRT Sequencing

C2 chemistry to sequence the SMRT-bellTM library. Two 45-min windows were captured for sequencing the chloroplast genome. After the CCS reads were derived from the multiple alignments of sub-reads, a quality control step was performed for the downstream assembly: SMRT Portal software (v2.0.0) was used to filter out the sequencing adapters and low-quality sequences (default parameters: sub-read length ≥ 50 bp; polymerase read quality ≥ 0.75; polymerase read length ≥ 50 bp; Li et al., 2014a). The reads were then used to assemble the chloroplast genome according to the strategy described in Qian et al. (2013). First, a workflow was designed to assemble the chloroplast genome: algorithms for greedy assembly, mapping, and consensus calling were used sequentially. Second, BLAST was used to compare the sequences from the greedy workflow, and the results of the alignment were used to construct the raw cp genome. The reads were mapped to the raw cp genome using the BWA tool, and the final cp genome sequence was generated using CAP3-based consensus calling (Altschul et al., 1997; Huang and Madan, 1999; Li and Durbin, 2010). To verify the genome sequence, PCR-based conventional Sanger sequencing was performed on six chloroplast genes (cemA, psbB, psbC, rpoA, rpoC1, and rps4; Cronn et al., 2008). The four junctions between the single-copy regions and the IRs were validated using PCR. The amplified sequences and the SMRT sequencing-based reads were aligned using Mega 5.2.2 (Tamura et al., 2011).

GENOME ANNOTATION AND CODON USAGE
The cp genome was annotated using the program DOGMA (Wyman et al., 2004; default parameters: the percent identity cutoff for protein coding genes = 60%, the percent identity cutoff for RNAs = 80%, the E-value = 1e−5 and the number of blast hits to return = 5.), and the position of each gene was determined using a blast method with the complete cp genome sequence of Ranunculus macranthus (GenBank Acc. No. NC_008796) as a reference sequence. Manual corrections for start and stop codons and for intron/exon boundaries were performed by referencing the Chloroplast Genome Database (ChloroplastDB; Cui et al., 2006). The tRNA genes were identified using DOGMA and tRNAscan-SE (Schattner et al., 2005). The circular cp genome map of A. barbatum var. puberulum was drawn using the OrganellarGenome DRAW tool (ORDRAW; Lohse et al., 2007). Codon usage and GC content were analyzed by Mega 5.2.2.

REPEAT ANALYSIS
REPuter (Kurtz et al., 2001) was used to assess both direct and IRs according to the following criteria: cutoff n ≥ 30 bp and 90% sequence identities (Hamming distance equal to 3). Tandem Repeats Finder (TRF) v4.04 (Benson, 1999) was used to analyze tandem repeats with the settings reported by Nie et al. (2012). Simple sequence repeats (SSRs) were detected using MISA (http://pgrc.ipk-gatersleben.de/misa/), with thresholds of eight, four and three units for mono-, di-, and trinucleotide SSRs and tetra-, penta-, and hexanucleotide SSRs, respectively.
RESULTS

PacBio RS OUTPUT AND GENOME VALIDATION

Quantitative analysis using an Agilent 2100 showed that the average length of the sheared DNA fragments was approximately 1 kb. In total, 23,498 CCS reads and 20,685,462 base pairs were generated, the mean read length was 880 bp, and the longest read was 2,261 bp. Genome coverage of 100% was achieved with a mean coverage of 132× and no gaps. Detailed information is listed in Table 1. Six conserved genes with poly-structures (cemA, psbB, psbC, rpoA, rpoC1, and rpo4) and four junction regions were validated using Sanger sequencing. The validated genes amounted to 7,341 bp, and a comparison of the assembled cp genome sequence dated using Sanger sequencing. The validated genes included all 20 amino acids necessary for biosynthesis. Leucine and serine (three of the 31, respectively) were the two most common amino acids represented by the codons of the tRNA in the cp genome.

GENOME FEATURES

The complete cp genome of *A. barbatum* var. *puberulum* (GenBank acc. No. KC844054) was 156,749 bp in length with the common quadripartite structure found in most land plants (Figure 1), which included a LSC of 87,630 bp and a SSC of 16,941 bp separated by two IRRs of 26,089 bp. In accordance with most chloroplast genomes, the nucleotide composition of *A. barbatum* var. *puberulum* was biased toward A+T (Sato et al., 1999; Nie et al., 2012; Pan et al., 2012; Yi and Kim, 2012). Overall, the *A. barbatum* var. *puberulum* cp genome A+T content was 61.3%, and the LSC and SSC regions (63.9 and 67.3%, respectively) were higher in A+T content than the IR regions (57.0%; Table 2). The *A. barbatum* var. *puberulum* cp genome contained 84 protein-coding regions, including seven genes (*rpl2, rpl23, ycf2, ndhB, rps7, rps12*, and *ycf1*) that were duplicated in the IR regions. In total, 31 unique tRNA genes (including seven tRNA genes located in the IR regions, *trnL-CAU, trnL-CAA, trnV-GAC, trnL-GAU, trnA-UGC, trnR-ACG* and *trnN-GUU*) were distributed throughout the cp genome, and four tRNA genes were duplicated in the IR regions. In summary, the cp genome of *A. barbatum* var. *puberulum* contained 130 genes, 18 of which were intron-containing genes (Table 3). Three of the intron-containing genes (*ycf3, ctp*, and *rps12*) had two introns, and the other 15 had only one intron. The 5′ end of *rps12* was located in the LSC region, and the 3′ end was located in the IR region, which caused trans-splicing in *rps12*. In addition, the sequence of *psbD* in the cp genome of *A. barbatum* var. *puberulum* differed from that in the reference sequence from *R. maccranthus* (GenBank: NC_008796), which was found to be complementary. Moreover, *infA* was not present in the cp genome of *A. barbatum* var. *puberulum*; this gene codes for translation initiation factor 1 and is suspected to be an example of chloroplast-to-nucleus gene transfer (Millen et al., 2001). The codon usage and codon-anticodon recognition pattern of the cp genome are summarized in Table 4. The 31 unique tRNA genes included codons for all 20 amino acids necessary for biosynthesis. Leucine and serine (three of the 31, respectively) were the two most common amino acids represented by the codons of the tRNA in the cp genome.

REPEAT ANALYSIS

Four forward, five inverted and eight tandem repeats were identified by REPuter and TRF with a copy size 30 bp or longer (Table 5). Most repeats possessed lengths between 30 and 40 bp, and the longest repeat was 52 bp as a forward repeat located the LSC region (*psaA, psaB, CDS*). All tandem repeats were found to be repeated twice in the whole cp genome, and six of these were located in intergenic spacer regions, with the left two located within *ycf2* (CDS) and *rps16* (intron), respectively.

SSR ANALYSIS

Microsatellites in the chloroplast genome are highly informative about genetic diversity and represent a useful tool for population genetics and evolutionary and ecological studies (Powell et al., 1996; Huang and Sun, 2000; Provan et al., 2001). Thus, the SSRs in the cp genome of *A. barbatum* var. *puberulum* were identified for use in future studies. The total number of the mononucleotides (not shorter than 8 bp) was 131, and T represented the highest portion (53.4%) followed by A, C, and G (44.3%, 1.5%, and 0.8%, respectively). The longest poly structure was a 20 T-repeat. In total, 56 dinucleotides were detected throughout the cp genome, and most of them were present as four repetitions (78.6%), e.g., ATATATAT. The combination of AT/TA was the most prevalent dinucleotide (42.9%). Four types of trinucleotide (ATA/ATT/TAT/TTA) were present as multiple A/T nucleotides. Seven tetranucleotides were detected, but no penta- or hexanucleotides (repeated at least three times) were found in the cp genome of *A. barbatum* var. *puberulum*. It can be inferred that the SSR loci contribute to the A+T richness of the cp genome. The longest poly-T and poly-A structures (20-nucleotide repeats and 14-nucleotide repeats, respectively) were located in IGS (*petA-pabf, ycf3*, and IGS (*psaI-rpl33*).

DISCUSSION

*Aconitum* IS HIGHLY TOXIC, AND CHLOROPLAST GENOME IS INFORMATIVE AND REFERABLE FOR MOLECULAR IDENTIFICATION

In recent years, there have been many reports on the improper use of toxic, aconitine-containing plants, which has led to deaths (Poon et al., 2006; Chen et al., 2012). Therefore, *Aconitum* identification is important. With reductions in sequencing costs, cp genomes could be used as super-barcodes in the near future. After

---

Table 1 | Summary of the sequencing data for PacBio SMRT.

| Raw data | Clean data |
|----------|------------|
| Number of reads | 300,582 | 23,498 |
| Number of nucleotides (bp) | 673,594,247 | 20,685,462 |
| Longest read length (bp) | 11,914 | 2,261 |
| Mean read length (bp) | 2,241 | 880 |
| Genome coverage % | 100% |
| Average depth of coverage | 132× |
| Number of contigs | 1 |
| Whole cp genome length (bp) | 156,749 |
| Run time | 45 min×2 |
| Total DNA requirements (ng) | 700 |

www.frontiersin.org
sequencing and analyzing the cp genomes of 37 different Pinus species, Parks et al. (2009) concluded that cp genomes could be used to improve phylogenetic resolution at lower taxonomic levels and could be thought of as species-level DNA barcodes. Li et al. (2014b) also suggest that complete cp genomes have tremendous potential for the identification of closely related species. Aconitum consists of approximately 300 species and its taxonomy has been complex due to the close relationships among different species (Xiao et al., 2005; Jabbour and Renner, 2012). Cp genome regions such as psbA-trnH have been applied, but it cannot be used to identify all the species of Aconitum (He et al., 2010). Hence, whole cp genomes are thought to have the potential in Aconitum identification studies. In our study, the successful use of a third-generation sequencing platform provides a new, rapid way to sequence the Aconitum cp genomes, which could help to lay the foundation for the molecular identification of Aconitum based on its cp genomes.

FIGURE 1 | Gene map of the Aconitum barbatum var. puberulum chloroplast genome.
Table 2 | General information of the *Aconitum barbatum* var. *puberulum* chloroplast genome.

| Length and GC content of the four regions | Whole genome | LSC region | SSC region | IR region |
|------------------------------------------|--------------|------------|------------|-----------|
| Length (bp)                              | 156,749      | 87,630     | 16,941     | 26,089    |
| GC content (%)                           | 38.7         | 36.1       | 32.7       | 43.0      |

| Number of total genes and intron-containing genes
|------------------------------------------|--------------|
| Protein-coding                          |              |
| Total genes regions                     | Protein-coding |
| rRNA                                     | tRNA          |
| Number of genes                         | 130           | 84          | 31         | 4          |
| Number of intron-containing             | 18            | 12          | 6          | 0          |

**CS READS PROVED TO BE RELIABLE VIA SANGER SEQUENCING VALIDATION**

In this study, we demonstrated the feasibility of sequencing a cp genome using the PacBio SMRT third-generation sequencing platform; use of this platform has been shown to be a rapid approach for sequencing small genomes, such as microbial and plasmid genomes (Chin et al., 2013). We evaluated the error-rate of the PacBio RS data by comparing its results with those obtained by Sanger sequencing. The CCS reads generated in our study had an error rate of approximately 0.027%, which was lower than the rate reported by Cronn et al. (2008) for Illumina sequencing-by-synthesis technology (0.056%). However, some questions remain regarding the error rate of the PacBio system. The observed raw error rate was 12.86%, which was much higher than that of other platforms, such as Illumina MiSeq and Ion Torrent PGM (Quail et al., 2012). To improve this situation, CCS is thought to be an effective approach. CCS is one of the PacBio RS sequencing protocols that performs multiple passes on each molecule that is sequenced. After the application of the necessary QC filters, the result is an error-corrected consensus read with a higher

Table 3 | The genes with introns in the *A. barbatum* var. *puberulum* chloroplast genome and the length of the exons and introns.

| Gene | Location | Exon (bp) | Intron (bp) | Exon (bp) | Intron (bp) | Exon (bp) |
|------|----------|-----------|-------------|-----------|-------------|-----------|
| trnK-UUU | LSC | 35 | 2528 | 37 |
| rps16 | LSC | 251 | 887 | 40 |
| trnG-GCC | LSC | 23 | 738 | 48 |
| atpF | LSC | 407 | 779 | 124 |
| rpoC1 | LSC | 1612 | 758 | 428 |
| ycf3 | LSC | 155 | 746 | 240 | 735 | 124 |
| trnL-UAA | LSC | 35 | 495 | 50 |
| trnV-UAC | LSC | 37 | 597 | 39 |
| cipP | LSC | 246 | 663 | 289 | 833 | 71 |
| petB | LSC | 6 | 948 | 489 |
| petD | LSC | 8 | 704 | 496 |
| rpl16 | LSC | 399 | 1069 | 9 |
| rpl2 | IR | 434 | 667 | 391 |
| ndhB | IR | 756 | 702 | 777 |
| trnI-GAU | IR | 42 | 937 | 35 |
| trnA-UGC | IR | 38 | 802 | 35 |
| ndhA | SSC | 539 | 1003 | 553 |
| rps12 | LSC | 114 | 232 | 544 | 26 |
Table 4 | The codon–anticodon recognition pattern and codon usage for the *A. barbatum* var. *puberulum* chloroplast genome.

| Amino acid | Codon No. | RSCU | tRNA | Amino acid | Codon No. | RSCU | tRNA |
|------------|-----------|------|------|------------|-----------|------|------|
| Phe        | UUU 834   | 1.23 |      | Tyr        | UAU 714   | 1.59 |      |
| Phe        | UUC 519   | 0.77 | tmF-GAA | Tyr       | UAC 185   | 0.41 | tmY-GUA |
| Leu        | UUA 765   | 1.77 | tmL-UAA | Stop       | UAG 26    | 0.93 |      |
| Leu        | UUG 554   | 1.29 | tmL-CAA | Stop       | UAG 26    | 0.93 |      |
| Leu        | CUU 542   | 1.26 |      | His        | CAU 488   | 1.52 |      |
| Leu        | CUC 192   | 0.45 |      | His        | CAC 153   | 0.48 | tmH-GUG |
| Leu        | CUA 356   | 0.83 | tmL-UAG | Gln        |CAA 640| 1.5 | tmG-UUC |
| Leu        | CUG 177   | 0.41 |      | Gln        | CAG 215   | 0.5  |      |
| Lle        | AUU 1017  | 1.46 |      | Asn        | AAC 901   | 1.54 |      |
| Lle        | AUC 430   | 0.62 | tmL-GAU | Asn        | AAC 901   | 1.54 |      |
| Lle        | AUA 646   | 0.93 | tmL-CAU | Lys        | AAA 889   | 1.44 | tmK-UUU |
| Met        | AUG 602   | 1    | tmM-CAU | Lys        | AAG 345   | 0.56 |      |
| Val        | GUU 507   | 1.45 |      | Asp        | GAU 809   | 1.6  |      |
| Val        | GUC 157   | 0.45 | tmV-GAC | Asp        | GAC 200   | 0.4  | tmD-GUC |
| Val        | GUA 527   | 1.51 | tmV-UAC | Glu        | GAA 924   | 1.45 | tmE-UUC |
| Val        | GUG 208   | 0.59 |      | Glu        | GAG 348   | 0.55 |      |
| Ser        | UCU 538   | 1.67 |      | Cys        | UGU 220   | 1.5  |      |
| Ser        | UCC 335   | 1.04 | tmS-GGA | Cys        | UGC 73    | 0.5  | tmC-GCA |
| Ser        | UCA 390   | 1.21 | tmS-UGA | Stop       | UGA 22    | 0.79 |      |
| Ser        | UCG 191   | 0.59 | Trp   | UGG 432   | 1      | tmW-CCA |
| Pro        | CCU 406   | 1.52 |      | Arg        | CGU 348   | 1.38 | tmR-AGC |
| Pro        | CCC 214   | 0.8  | tmP-GGG | Arg        | CGC 86    | 0.34 |      |
| Pro        | CCA 311   | 1.16 | tmP-UGG | Arg        | CGA 346   | 1.38 |      |
| Pro        | CGG 138   | 0.52 |      | Arg        | CGG 114   | 0.45 |      |
| Thr        | ACU 502   | 1.57 |      | Arg        | AGA 451   | 1.79 | tmR-UCU |
| Thr        | ACC 240   | 0.75 | tmT-GGU | Arg        | AGG 163   | 0.65 |      |
| Thr        | ACA 393   | 1.23 | tmT-UGU | Ser        | AGU 372   | 1.15 |      |
| Thr        | ACG 140   | 0.44 |      | Ser        | AGC 108   | 0.34 | tmS-GCU |
| Ala        | GCU 589   | 1.73 |      | Gly        | GGU 592   | 1.36 |      |
| Ala        | GCC 220   | 0.65 |      | Gly        | GGC 177   | 0.41 | tmG-GCC |
| Ala        | GCA 382   | 1.12 | tmA-UGC | Gly        | GGA 694   | 1.59 | tmG-UCC |
| Ala        | GCG 171   | 0.5  |      | Gly        | GGG 282   | 0.65 |      |

RSCU, Relative Synonymous Codon Usage.

intra-molecular accuracy. This approach results in higher per-base quality and reduced concerns about suspicious results. By generating multiple reads from the same molecule and eliminating errors resulting from single reads, the PacBio system’s inherent error rate can be bypassed. For that reason, in this study, we used the CCS protocol to sequence the *A. barbatum* var. *puberulum* cp genome and obtain high-quality reads. The data presented here show that SMRT sequencing using the CCS strategy is a powerful tool for sequencing cp genomes. In addition, in some extreme situations, we suggest completing genome assembly by combining CCS reads with regular long reads. We believe that this strategy would be an effective way to solve the problems associated with assembling large genomes or genomes that contain special structures.

**THE LONG READS DERIVED FROM PacBio IMPROVE GENOME ASSEMBLY**

The long read lengths undoubtedly provide a number of benefits in genome sequencing and assembly. The most obvious benefit is for *de novo* assemblies. Previous studies have shown that, compared with Illumina data, chloroplast genome assembly using the PacBio RS sequencer generated longer contigs and fewer unresolved gaps (Ferrarini et al., 2013). In this study, we constructed the draft sequence in a step-by-step manner by extending two
seed reads on both the 5′ and 3′ ends until they overlapped at the two IR regions. For all CCS sub-reads, the top BLASTn hit for the seed sequence was selected and used to extend the read. The longer reads (an average of 880 bp) made our assembly and analysis more effective. We encountered no problems mapping the seed sequence reads to the repeat regions of the *A. barbatum* var. *puberulum* cp genome, which are listed in Table 5. Even without any other biological or phytological information about the target species, it took less than half an hour to finish the genome assembly step. This strategy is clearly a highly effective and accurate method for obtaining plant cp genomes. In addition, one of the features of the cp genome, the two long IR (regions), is also a valuable target for evaluating the PacBio system. As mentioned above, the comparatively longer CCS reads provided more conveniences on dealing with those special structures.

### ELIMINATING THE PCR AMPLIFICATION STEP SAVES TIME

The SMRT method does not require PCR amplification, which reduces the time required for sequencing. In our study, the sequencing reaction time was 90 min, which streamlined the sequencing process by reducing the overall time in the lab.

In addition, eliminating the PCR amplification step alleviated the sequencing bias. In some extreme situations, e.g., AT-rich, GC-rich, and repeat-rich regions, the results are unsatisfactory due to the loss of DNA during amplification (Bashir et al., 2012). The sequencing of unamplified molecules will improve genome assembly and allow the detection unique and informative structures.

### REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402. doi: 10.1093/nar/25.17.3389

---

**Table 5 | Repeated sequences in the *A. barbatum* var. *puberulum* chloroplast genome.**

| Repeat number | Size (bp) | Type | Location | Repeat unit | Region |
|---------------|-----------|------|----------|-------------|--------|
| 1             | 52        | F    | psaB(CDS), psaA(CDS) | AGAAAAAGAATTGCAATAGCTAAATGG(T)GA(G)TGA(C)GCAATATCGGTCAGCCATA | LSC |
| 2             | 39        | F    | ycf3(CDS), IGS(trnV-GAC, rps12) | CAGAACCGTACATGAGATTTTCACCTCATACGGCTCCTC | LSC, Ira |
| 3             | 33        | F    | IGS(psbi, tmS-GCU), IGS(psbC, tmS-UGA) | TAAAC(A)GGAA(G)AGAGAGGGATTCGAACCCTCGGTA | LSC |
| 4             | 32        | F    | IGS(rps15, ycf1) | TTT(G)TTCT(A)T(A)CTTGATTTAGATTCTCTAATTCAA | SSC |
| 5             | 39        | P    | ycf3(intron), IGS(tmV-GAC, rps12) | CAGAACCGTACATGAGATTTTCACCTCATACGGCTCCTC | LSC, Irb |
| 6             | 33        | P    | IGS(perA, psbJ), psbJ(CDS) | GTAAGAATAAGAACTCAATGGACCCTGTCCCTC | LSC |
| 7             | 30        | P    | tmS-GCU, tmS-GGA | ACGAAAGAGAGGGATCGAACCCTCGGTA | LSC |
| 8             | 31        | P    | IGS(tmE-UUC, tmF-GGU) | TCTGATTATTCAATCTATATATCTATATGAT | LSC |
| 9             | 30        | P    | tmS-UGA, tmS-GGA | TACACGAGGGTTCAATCCCTCTCTT(G)TCGGAAT | LSC |
| 10            | 30        | T    | rps16(intron) | TAATAGTATATTAG(×2) | LSC |
| 11            | 34        | T    | IGS(rps16, tmQ-UUU) | TTTTATCTATTATTA(×2) | LSC |
| 12            | 40        | T    | IGS(atpF-atpH) | GTTATGTAGAGATGAAATC(×2) | LSC |
| 13            | 30        | T    | IGS(tmT-GGU, psbD) | ATAGCTATTATAATG(×2) | LSC |
| 14            | 36        | T    | IGS(ribC, accD) | TTTTATTTATGATCA(×2) | LSC |
| 15            | 40        | T    | IGS(tmP-GGG, psaJ) | TTTTATATATAATATTAA(×2) | LSC |
| 16            | 42        | T    | ycf2(CDS) | AGATAATGAACTATTCAAAGA(2) | IRa,b |
| 17            | 30        | T    | IGS(indE, ndhG) | TATTACCTATTATAT(×2) | SSC |
ion torrent, pacific biosciences and illumina miSeq sequencers. BMC Genomics 13:341. doi: 10.1186/1471-2164-13-341

Rasko, D. A., Webster, D. R., Bashir, A., Boisen, N., Scheutz, F., et al. (2011). Origins of the E. coli strain causing an outbreak of hemolytic–uremic syndrome in Germany. N. Engl. J. Med. 365, 709–717. doi: 10.1056/NEJMoa1106920

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. Genome Biol. 14, 405. doi: 10.1186/gb-2013-14-6-405

Rusk, N. (2009). Cheap third-generation sequencing. Nat. Methods 6, 244–244. doi: 10.1038/nmeth0409-244a

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999). Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Res. 6, 283–290. doi: 10.1093/dnares/6.5.283

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. Hum. Mol. Genet. 19, R227–R240. doi: 10.1093/hmg/ddq416

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33, W686–W689. doi: 10.1093/nar/gki366

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.

Sugiura, M. (1992). The chloroplast genome. Plant Mol. Biol. 19, 149–168. doi: 10.1007/BF00015612

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGAS: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739. doi: 10.1093/molbev/msr121

Uthaipaisanwong, P., Chanprasert, J., Shearman, J., Sangsrukru, D., Yoocha, T., Jomchai, N., et al. (2012). Characterization of the chloroplast genome sequence of oil palm (Elaeis guineensis jacq.). Gene 500, 172–180. doi: 10.1016/j.gene.2012.03.061

Wu, J., Liu, B., Cheng, F., Ramchiar, N., Choi, S. R., Lim, Y. P., et al. (2012). Sequencing of chloroplast genome using whole cellular DNA and Solexa sequencing technology. Front. Plant Sci. 3:243. doi: 10.3389/fpls.2012.00243

Yi, D.-K., and Kim, K.-J. (2012). Complete chloroplast genome sequences of important oilseed crop Sesamum indicum L. PLoS ONE 7:e35872. doi: 10.1371/journal.pone.0035872

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 October 2014; accepted: 15 January 2015; published online: 06 February 2015.

Copyright © 2015 Chen, Li, Qian and Han. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

www.frontiersin.org
February 2015 | Volume 6 | Article 42 | 9