geneious prime

Application Note

De novo assembly and reconstruction of complete circular chloroplast genomes using Geneious Prime

De novo assembly and reconstruction of complete circular chloroplast genomes using Geneious Prime

M. D. Gibbs, Geneious Application Scientist

Introduction

Geneious Prime[1] contains all of the tools required to do rapid and accurate *de novo* assembly of chloroplast genomes from short-read NGS data. The NGS data may be derived DNA extracted from purified chloroplasts, or "skimmed" from whole-genome sequence of total DNA derived from chloroplast-rich leaf material.

Chloroplast genomes contain large (≈25,000 bp) almost perfect inverted repeats (IR). During *de novo* assembly individual repeats cannot be resolved unless the paired-read insert size is larger than the repeat unit. This means a complete circular plastome cannot be resolved during assembly if using only short-read data. However, the use of paired-read data, combined with identification of the repeat and truncated repeat boundaries, can allow reconstruction of the complete circular plastome.

In this application note we take a short-read NGS data set, available for download from the NCBI Sequence Read Archive (SRA), and describe how to use Geneious Prime to reconstruct a complete, circular, annotated chloroplast genome.

Methods and results

Dataset: Short-read data obtained from the chloroplast of *Sesame indicum* downloaded in FASTQ format from the European nucleotide archive (www.ebi.ac.uk/ena), accession number SRR949054. Paired-end, mate-pair and Roche 454 datasets were generated for this project. Only the paired-end data set was used for assembly in this demonstration (SRR949054). The paired-end data set comprised 1,017,560 reads (length 2 x 100 bp), with an expected insert size of 500 bp, and an average base quality of 34.

The *S. indicum* data were derived using DNA extracted from chloroplasts purified by sucrose gradient. Paired read data were generated by Illumina-Solexa GAIIx, see Zhang et. al. (2013)[2].

Read Processing: The *S. indicum* paired-end fastq dataset was imported into Geneious Prime as two lists. Reads were paired using 'Set paired reads' with an expected insert size of 500 bp. Paired reads were processed using 'Trim using BBDuk'[3], using default settings with the options checked to 'Trim Adapters', 'Trim Low Quality (Q20)', 'Discard short reads (minimum length 10 bp)', and 'Keep original order' - see Fig. 1. Ensure the option to 'Keep original order' is checked otherwise assembly results may differ slightly from those reported here. The trimmed read list (974,422 reads) was then normalized using 'Error correct and normalize reads' (BBNorm[3]), with target coverage level 30, minimum depth 6, and no 'Error correction'. After trimming and normalization, a list of 72,204 paired reads were obtained for *de novo* assembly (note that for some datasets BBNorm results may differ depending on the number of CPU cores used. The advanced setting "threads=N" [where N is some integer], can be used to ensure results are consistent regardless of the hardware used).

Assembly: The normalized paired-read set was assembled using the Geneious *de novo* assembler. Default settings (Medium sensitivity/Fast) were used, with the advanced options to 'Don't merge variants' and 'Produce scaffolds' unchecked. The option to 'Don't merge variants' was de-selected to ensure reads from mixed genotypes (if present) were assembled as a single contig, and so that variant calling could be performed. The 'Produce scaffolds' option was unchecked to prevent non-overlapping read pairs from incorrectly mapping at at terminal repeat regions. *De novo* assembly generated 304 contigs. The largest contig was 129,265 bp (consensus 129,145 bp) with an aerage coverage depth of 47.1 (minimum coverage 20).

Resolution of IR units: Plant chloroplast genomes are circular and approximately 150-160 Kb in size, and the expected large IR unit is around 25 Kb in size. Therefore, it is likely this 129,048 bp consensus represents an entire plastome, with all repeat-derived reads assembled as a single IR unit.

Although *de novo* assembly of short reads will not allow resolution of both large IR units, *de novo* assembly of paired reads will regenerate one full IR, and also, the

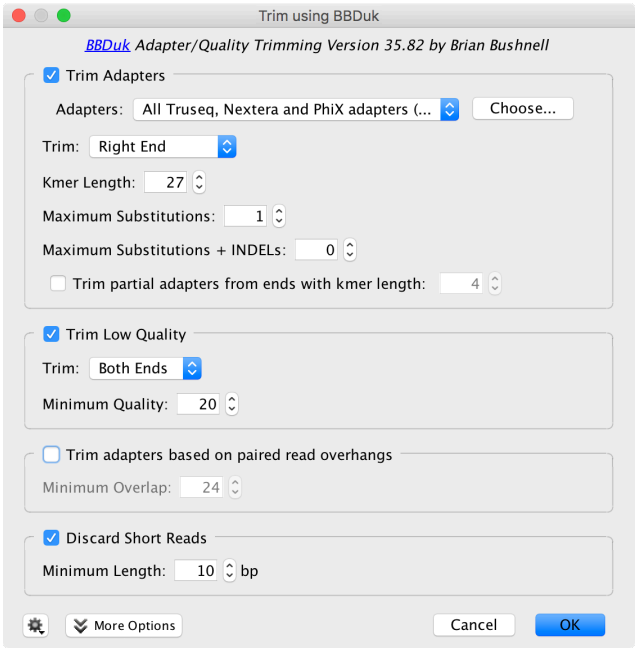

Figure 1. BBDuk settings for trimming of paired read data.

truncated ends of the second IR. These truncated ends will be found at the contig termini. Identification of these truncated ends allows complete regeneration of a draft circular genome via identification and duplication of the full-length IR and concatenation to the consensus.

The following steps were taken to regenerate the circular plastome sequence. These steps are outlined in Fig. 2.

First, the consensus sequence of the 129,145 bp contig 1 was extracted to a new file using the tool 'Generate consensus sequence' (Fig. 2A).

The Geneious 'Find Repeats' plugin was then used to find all perfect repeats with a minimum length of 70 bp. This identified two sets of repeats (Repeat 1 x2 and Repeat 2 x2, see Fig. 2B). The region encompassed by the internal Repeat 1 and repeat 2-pair (indicated by yellow arrow) spans 25,145 bp and represents the single instance of the large IR . The short repeats identified at the ends of the consensus represent truncated boundaries of the second large IR. As can be seen in Fig. 2B, the repeat pairs are inverted in relation to each other which is to be expected if the large IR's are inverted repeats.

The repeats identified on the termini of the 128,145 bp consensus sequence, and any flanking terminal sequence, were then deleted (Fig. 2C, regions deleted 128,703-129,145 and 1-498) to generate a truncated 128,204 bp consensus.

The region encompassed by the internal short Repeat 1 and Repeat 2 repeats (positions 85,186-110,330) was selected and 'Extracted' to a new 25,145 bp sequence. As the second IR is inverted with respect to the first, the extracted IR was then reverse complemented and saved (Fig. 2D).

Next, the extracted inverted repeat and the trimmed consensus sequence were both selected, and concatenated and circularized using the 'Concatenate Sequences' command to create a draft consensus genome of 153,349 bp (Fig. 2E).

Map to reference: The above normalization step has the potential to amplify to "significance" strand-specific sequencing errors and/or miscalls/indels in homopolymer regions. Therefore, the circular consensus derived from *de novo* assembly was then used as the reference for a 'Map to Reference' assembly using the full trimmed paired reads dataset (Fig. 2F). Default settings were used, with the Advanced setting to 'Only map paired reads which map nearby' enabled to ensure read pairs located within or across the IR boundaries mapped correctly to a single IR. The consensus sequence generated from the

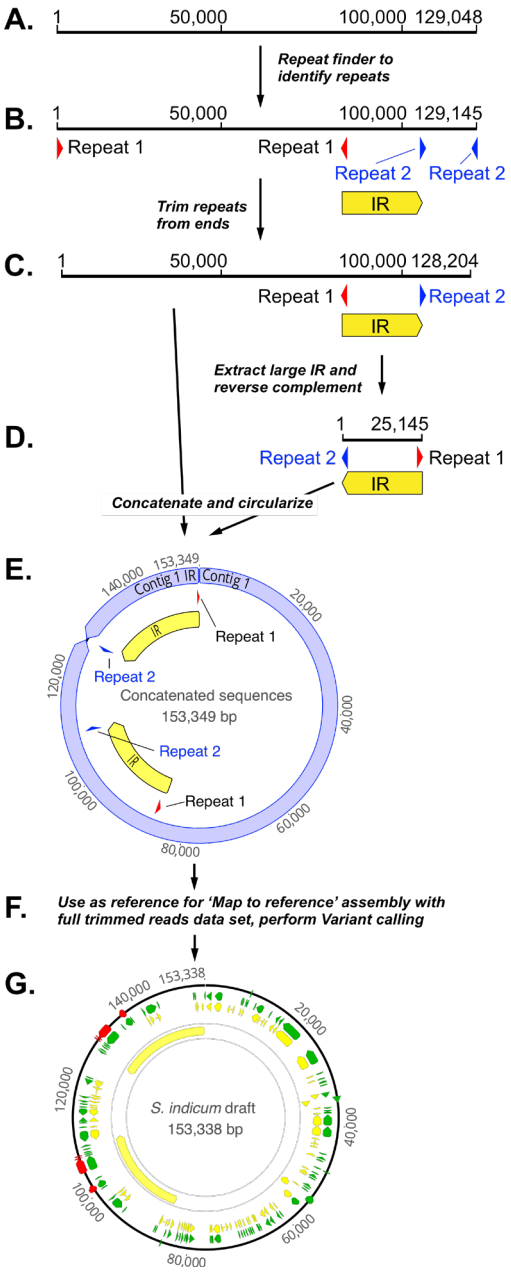

Figure 2. Workflow to assemble complete plastome sequence. A. Generate consensus sequence from largest contig. B. Use Geneious Repeat Finder to identify perfect repeats >70 nucleotides. C. Trim terminal repeats and any flanking sequence. D. Select Repeat 2, hold down shift to select Repeat 1 (and all intervening sequence) and Extract to a new file. Reverse-complement the new file. E. Use Concatenate to join and circularize the two sequences. F. Annotate the sequence using 'Annotate from Database' using an appropriate annotated homolog.

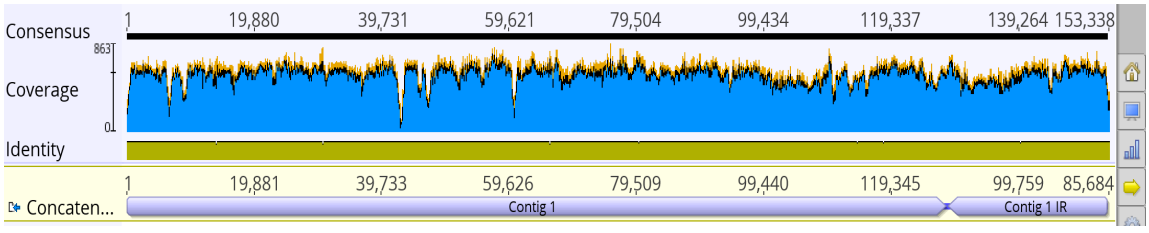

Figure 3. Coverage graph showing trimmed reads (974,422 paired reads) mapped to de novo-derived consensus sequence. Areas of High %A+T correspond to significant dips in the coverage graph, with the minimum coverage (32) occurring at position 42,720 bp.

Map to reference assembly was the same length (153,338 bp) as the published *S. indicum* plastome[2] (Accession no. KC569603, also derived from this data), with a mean coverage depth of 577, minimum coverage depth of 32 (see Fig. 3). A Map to reference alignment of the *de novo* consensus and the Map to reference consensus showed that the difference in lengths (153,349 bp vs 153,338 bp respectively) were primarily due to errors in the *de novo* assembly in homopolymeric regions (data not shown).

Variant calling: The resulting Map to reference assembly was selected and the 'Find Variations/SNPs' tool was used with default settings except Minimum Variant Frequency of 0.1 was used, and 'Ignore reference sequence' checked. A single variant position was identified at consensus position 36,964. However, this potential variant lies at the boundary of a substantial homopolymeric region and is likely an artifact. Interestingly, the consensus for the assembly contained a single ambiguity (G or T at position 116,112). However, this position was not called as a variant position unless the Variant finder was run with the option to screen based on 'Minimum Strand bias' unchecked. The strand bias statistics for this position indicate that the 'true' call at this position is T, and that G calls were due to a strand-specific sequencing error, with 100% of G calls occurring on reverse reads only. This call represents the only difference compared to the published sequence.

Annotation: The final circular consensus was annotated using the Geneious 'Annotate from Database' tool, % Identity set to 50%, using the published annotated genome of *Ageratina adenophora* (Accession NC_015621)[4] as the Source (Fig. 2G). All corresponding CDS/gene and RNA genes were identified and annotated onto the *S. indicum* plastome. See our tutorial on 'Transferring annotations' (www.geneious.com/tutorials) for detailed instructions on how to use 'Annotate from database'.

Discussion

Reconstruction of the *S. indicum* genome as described above resulted in a generation of a complete draft circular genome. The final circular consensus of 153,338 bp was identical in length to that published by Zhang et al (2013)[2] and differed at only a single nucleotide position. The use of the Geneious Variant Finder tool provided evidence that this difference was incorrectly called in the published sequence.

Normalization of the *S. indicum* data was found to be essential for obtaining proper *de novo* assembly of the *S. indicum* plastome. Normalization removes over-represented reads from the dataset and can even out datasets with irregular coverage. For this dataset, normalization removed ≈93% reads, which simplified the dataset to the point where assembly took just a few minutes on a desktop computer. Analysis of a number of chloroplast NGS data sets (from publicly available repositories) typically revealed very poor coverage in several AT-rich regions that are conserved in most plastomes (data not shown). In these AT-rich regions (with %A+T of greater than 90% over a span of 100-200 bp), read depth was often less that 7% of the average coverage, and as a result, generation of a complete assembly was difficult or impossible for many data sets if normalization was not performed. Library preparation techniques that reduce bias against AT-rich regions may improve the coverage of an NGS dataset and increase the chances that a chloroplast genome will be assembled in full[5].

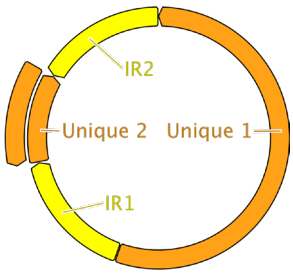

Figure 4. Relative orientation of unique LSC and SSC regions. Due to the presence of the large inverted repeats, de novo assembly has a 50% chance of assembling the LSC and SSC region with relative orientations matching the published sequence (KC569603).

It should be noted that *de novo* assembly of plastomes has a 50% chance of assembling the unique regions (large single copy, LSC and small single copy, SSC) in the same relative orientations as observed in the published KC569603 sequence (See Fig. 4). It has been found that both orientations of the SSC region occur regularly during the course of chloroplast replication within individual plants cells, so both orientations should be regarded as equally correct[6].

Conclusions

Geneious Prime provides all of the tools required for the preprocessing, *de novo* assembly, variant analysis and annotation of plastome genomes. Use of normalization (BBNorm) can improve assembly across poorly represented AT-rich regions, and the use of the Variant finder allows rapid identification of potential heteroplasmy and correction of strand-specific sequencing errors.

References

1. Kearse, M., et al. (2012) Bioinformatics 28:1647–1649
2. Zhang H., et al. (2013) PLOS ONE 8:E80508
3. Bushnell B. (2016) BBMap - sourceforge.net/projects/bbmap/
4. Nie, X., et al. (2012) PLoS ONE 7:E36869
5. van Dijk et al. (2014) Exp. Cell Res. 322:12-20
6. Walker el al. (2015) American J. Bot. 102:1-2

www.geneious.com • support.geneious.com
Published Oct 2016. Revised Jan 2019.