geneious prime

Application Note

Circular de novo assembly of organelle genomes

Circular de novo assembly of organelle genomes using Geneious Prime

Introduction

Circular genomes, such as viruses, bacteria, mitochondria and plasmids, are common. However, assembly of such genomes can be difficult in the absence of a reference genome, as most *de novo* assemblers do not account for circularity and produce linear sequences with an arbitrarily defined start and end. This can result in repeated sections of sequence at the arbitrary start and end points, and an artificial drop in coverage in these regions which can affect downstream analyses.

The Geneious Prime *de novo* assembler overcomes these issues by allowing contigs to circularise during the assembly process. In this study we assemble two mitochondrial genomes from short-read NGS sequence data using the Geneious Prime *de novo* assembler and compare the results with assembles produced by Velvet, MIRA and SPAdes.

Methods

Datasets for the Asiatic Lion[1] (*Panthera leo persica*) and the Chimpanzee[2] (*Pan troglodytes*) were downloaded from the NCBI Short Read Archive (Accession numbers SRR821548 and ERR032959, respectively).

The *Panthera leo* dataset consists of unpaired Ion Torrent reads from a purified mitochondrial DNA preparation. Prior to assembly adaptors and poor quality bases were trimmed off, and reads less than 50 bp were removed to leave 237,432 reads of 50-367 bp (mean 164).

The *Pan troglodytes* dataset is from whole-genome shotgun sequencing (approximately 1 x coverage), and consists of paired 76bp Illumina GAII reads with 250 bp insert length. Reads were quality trimmed prior to assembly. This dataset contains a total of 57,237,068 reads, but only 5% were assembled as the mitochondrial fraction was expected to be at much higher coverage than the nuclear fraction.

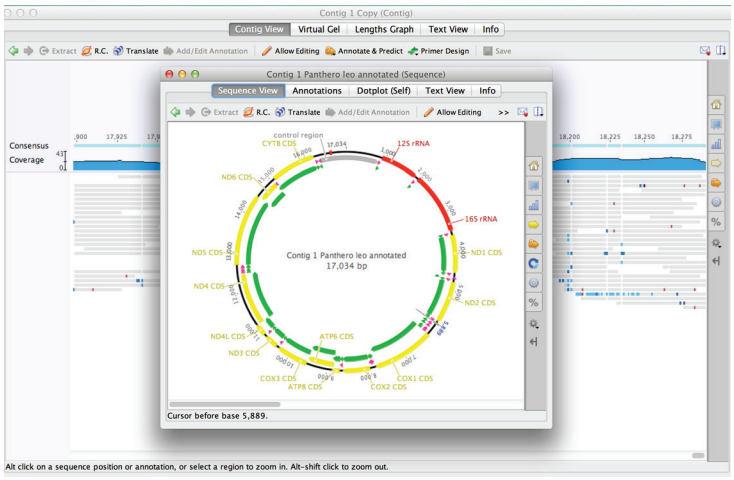
Assemblies were performed using Geneious Prime (version 7.1.5), Velvet[3], MIRA[4] and SPAdes[5]. Velvet and MIRA were run as plugins to Geneious Prime. Optimal parameters for Velvet were chosen by Velvet Optimizer to maximise the length of the longest contig. The following settings were used for each dataset:

Panthera leo:
Geneious Prime: Med/High sensitivity, circularize contigs option on
MIRA: Genome /contiguous sequence, accurate quality, Ion torrent setting
Velvet: Optimal kmer 57
SPAdes: kmers 21, 33, 55, 77, 99, read correction on, Ion torrent setting

Pan troglodytes:
Geneious Prime: Med/Low sensitivity, circularize contigs option on
MIRA: Genome /contiguous sequence, accurate quality, Illumina setting
Velvet: Optimal kmer 47
SPAdes: kmers 21, 33, 55, read correction on

Contigs produced from each assembly were mapped back to published mitochondrial genome sequences for each species using the Geneious Prime read mapper with medium sensitivity settings and no fine tuning.

Figure 1. Circular contig produced by de novo assembly in Geneious Prime



Results

1. Assembly of unpaired Ion Torrent reads (*Panthera leo*)

The Geneious R7 assembler produced a single, circular contig containing the entire mitochondrial genome from a dataset of unpaired Ion torrent reads (Figure 1). Although this dataset was from purified mtDNA, a large number of short linear contigs were also produced (not shown), indicating a significant level of nuclear contamination. By contrast, none of the other assemblers could assemble the mitochondrial genome into a single contig (Table 1).

The Geneious Prime assembly shows good agreement with the published genome, apart from a few positions where it is impossible to call the length of homopolymer runs due to the Ion Torrent error model, and the control region, where low coverage makes it difficult to resolve repetitive regions.

2. Assembly of paired Illumina reads from WGS sequencing (*Pan troglodytes*)

Geneious Prime, Velvet and SPAdes produced a single contiguous fragment representing the mitochondria (Table 2). However, the Velvet and SPAdes contigs are not circular and are 45bp and 61bp longer respectively, than the Geneious Prime contig because of a repeated section of sequence at the start and end. When mapped to the circular reference genome this produces a region of double coverage (Figure 2).

Table 1. Comparison of *Panthera leo* Assemblies

Assembler	mt genome coverage (no. of contigs)
Geneious Prime	100% (1 contig)
Velvet	84.5% (48 contigs)
MIRA	99.6% (4 contigs)
SPAdes	99.7% (3 contigs)

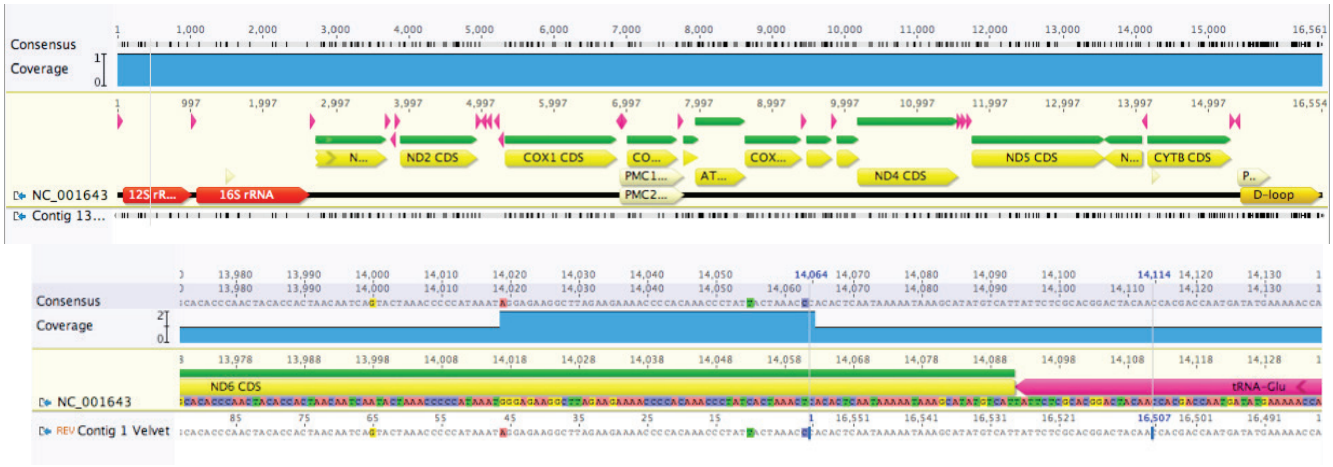
Table 1. Comparison of *Pan troglodytes* Assemblies

Assembler	mt genome coverage (no. of contigs)
Geneious Prime	100% (1 contig)
Velvet	100% (1 contig)
MIRA	99.9% (2 contigs)
SPAdes	100% (1 contig)

Conclusion

The Geneious Prime *de novo* assembler is the only assembler able to produce circular contigs as part of the assembly process. This facilitates efficient assembly of organelle genomes such as mitochondria and chloroplasts, even when whole-genome shotgun sequence reads are used as input. Because of the high copy-number of organelle genomes, they can be assembled directly from whole-genome shotgun data by assembling only a small percentage of the data, without the need to filter out the nuclear genome reads first. As demonstrated here with the Pan troglodytes WGS dataset, the mitochondrial genome is easily identifiable in the assembled contigs, as it is the largest and only circular contig. Geneious Prime also contains a circular mapper, which allows easy comparison of *de novo* assembly results with published genomes.

Figure 2. Mapping of *Pan troglodytes* contigs to published genomes. (A) shows the circular contig produced by Geneious Prime, and (B) shows the overlapping regions of Velvet contig (where coverage =2) caused by linear assembly.



References

1. Bagatharia SB, Joshi MN, Pandya RV et al., (2013) Complete mitogenome of asiatic lion resolves phylogenetic status within Panthera. BMC Genomics 14: 572.
2. Prüfer K, Munch K, Hellmann I et al., (2012). The bonobo genome compared with the chimpanzee and human genomes. Nature 486(7404):527-31.
3. Zerbino DR and Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18:821-829.
4. Chevreux et al. (1999) Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.
5. Nurk S., Bankevic A., Antipov D., (2013) Assembling genomes and mini-metagenomes from highly chimeric reads. Lecture Notes in Computer Science Volume 7821, 2013, pp 158-170.