geneious prime

Application Note

De novo assembly of a bacterial genome from high-throughput sequencing reads

De novo assembly of a bacterial genome from high-throughput sequencing reads

Introduction

Advances in high-throughput sequencing technology mean that sequencing of genomes is now cheap and easy. The Geneious Prime[1] *de novo* assembler is an overlap assembler optimized for genomes of 100 Mbp or less, and is thus ideal for applications to viral and bacterial genomes.

This application note describes a workflow for assembly and annotation of a bacterial genome from Illumina MiSeq data. The Illumina MiSeq platform has become the sequencer of choice for many microbial laboratories due to its long read length (up to 300 bp paired-end), accuracy and cost-effective protocols. Using the independent assembly evaluation tool Quast[2], we compare the assemblies created in Geneious Prime with a published assembly from the same dataset produced by the SPAdes assembler[3].

Data

We used a dataset from a Shiga toxin-producing *Escherichia coli* (STEC) strain O157:H7 ATCC 35150[4], which is available from the NCBI Short Read Archive (SRR1781795). These sequences were prepared using the Nextera XT DNA sample preparation kit and sequenced on the Illumina MiSeq platform using the 500 cycle kit (2 x 250 bp). The dataset contains 1,919,204 paired-end reads, with mean read length of 232 bp. The reads do not appear to contain adapter sequences.

Methods

Reads were downloaded from the NCBI Sequence Read Archive in fastq format, imported into Geneious ver. 9.1 and then processed within the program as follows: Firstly, reads were paired using “Set Paired Reads” with an insert size of 450 bp. Datasets produced by the Nextera XT library kit typically have a range of insert sizes, and 450 bp represents the approximate midpoint of the size range for this dataset*. Secondly, poor quality bases were trimmed from the ends of the reads using the Geneious Prime trimmer (“Trim Ends”) with an Error Probability Limit of 0.05.

The BBDuk[5] plugin was used to further trim and filter the dataset, as shown in Figure 1. Reads shorter than 20 bp or with a minimum average quality score of less than 20 were removed, and paired read overlaps (where a read extends past the start of its mate) were trimmed to ensure complete adapter removal. To test the effect of merging reads, the “Merge Paired Reads” function using BBMerge[5] was run using the lowest merging rate.

De novo assembly was performed using the Geneious Prime assembler. Low, Medium-Low and Medium sensitivity settings were used for comparison. Only contigs greater than 1000 bp were retained for comparison to the published genome. Assemblies were evaluated using Quast ver 3.22, with the *E. coli* O157:H7 reference genome (str. Sakai) and its plasmids used for comparison (NC_002695, NC_002127, NC_002128). Quast identifies breakpoints where the *de novo* assembled contigs do not match the reference. It then calculates an NA50 value, which is the N50 recalculated on contigs which are broken into pieces at misassembly breakpoints. In this case, we do not expect the O157:H7 str. Sakai reference sequence to be identical to the strain assembled here, so many of the “misassemblies” may be real differences between strains, but it still provides a useful standard to compare different assembly methods.

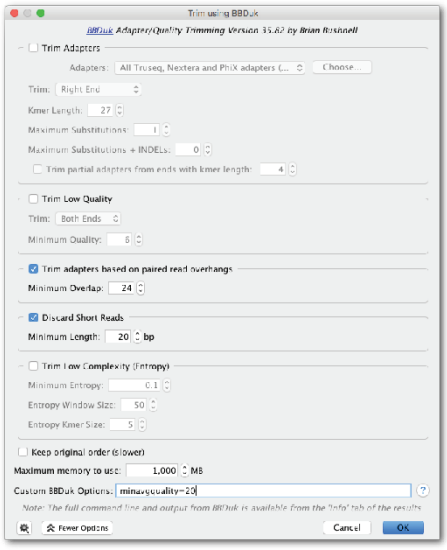

Figure 1. BBDuk options used for quality trimming

*For the purposes of this example, we determined the spread of insert sizes in the dataset by mapping the raw reads to the published assembly for this dataset (NZ_JXUS00000000). This revealed insert sizes of between about 100 to 800 bp. If mapping to a reference is not possible, users should consult their sequencing provider regarding the fragment size of their dataset and use the approximate midpoint when setting paired reads.

Results

The best Geneious Prime assembly was produced using the default "Medium-Low" sensitivity setting. This assembly had 108 contigs >1000 bp, with an N50 of 163,893 and mean coverage of 60, which compares favourably to the published assembly (performed using SPAdes) of 242 contigs with an N50 of 58,735 and mean coverage of 314.

We evaluated the accuracy of the assembly against the reference using Quast. These results showed that the SPAdes assembly had fewer putative misassemblies when compared to the reference sequence than the Geneious assembly, but even once these were taken into account, the Geneious Prime assembly had superior length metrics compared to SPAdes. The Geneious Prime assembly had an NA50 of 114,793 and longest alignment with the reference of 377,879, compared with NA50 of 56,301 and longest alignment of 202,085 for the SPAdes assembly.

Using the Geneious Prime Repeat Finder plugin, and mapping the *de novo*-assembled contigs to the reference sequence, we identified that the majority of misassembly breakpoints between the Geneious Prime assembly and the reference strain occurred at or adjacent to repetitive elements. In contrast, the SPAdes assembler is highly conservative and produced multiple separate contigs through these regions, which reduced the detection of misassemblies by Quast. We expect some divergence between the assembled strain and the reference genome at repetitive regions, given the higher mutation rate of those elements. For this reason, it is not possible to accurately determine the efficacy of the assemblers tested at these particular points. We would recommend the addition of longer reads to the dataset to resolve these regions and produce a complete genome sequence.

Effect of read processing

Our results show that removing poor quality data is important for producing an optimal assembly. We used the BBDuk plugin for Geneious Prime in addition to the Geneious Prime "Trim Ends" function to stringently trim short reads, poor quality reads and adapters. This process removed a total of 437,126 reads from the dataset and improved the mean quality score from 34.1 to 35.6. The assembly produced after this trimming procedure had half the number of contigs with twice the N50 than an assembly performed with only basic trimming of poor quality data from the ends of reads (see Table 1). Performing more extensive quality filtering with BBDuk also reduces the number of mismatches and misassemblies when compared to the reference sequence.

We also tested the effect of merging reads prior to the assembly, as the Nextera XT sample preparation method produced a range of insert sizes of 100-800 bp, meaning that some of the paired reads are overlapping. Using the lowest merge rate under "Merge Paired Reads", which uses the BBMerge algorithm, 454,317 pairs were merged, leaving 573,444 unmerged reads. The Geneious Prime assembly of merged + unmerged reads had fewer short contigs (<1000bp) than the assemblies where all reads were left unmerged. However, merging reads did not improve the overall assembly, as shown in Table 1. The merged reads assemblies also had more misassemblies compared to the reference, which may indicate some incorrect merging of reads, especially around repetitive regions.

Table 1: Effect of read processing on assembly outcome. All assemblies were performed in Geneious Prime with default settings (med/low sensitivity). NA50 is N50 calculated on the basis of aligned blocks rather than contig lengths. Longest alignment is the longest contiguous alignment with the reference sequence. All stats are calculated on contigs >1000 bp

Data processing	No. of contigs (Total, (>1000bp)	Mean coverage	Total length	N50	NA50	Longest contig	Longest alignment
Basic trim	4,811 (202)	67.4	5,586,865	68,413	65,155	225,821	224,665
Trim, filter with BBDuk	**1424 (108)**	**59.6**	**5,482,554**	**163,893**	**114,793**	**381,532**	**377,897**
Trim, filter with BB-Duk, Merge reads	387 (118)	46.4	5,521,328	32,839	109,566	361,460	361,460

Effect of algorithm on assembly outcome

The sensitivity settings in the Geneious Prime assembler allow the user to control the stringency and speed of the assembly. Lower sensitivity settings use longer Word and Index Word lengths (which specify the minimum number of consecutive matching bases between 2 reads), and allow fewer mismatches and gaps per read. Thus, at lower sensitivity settings fewer reads will be assembled, the assembly will be faster and require less RAM than at higher sensitivity settings. Low sensitivity settings produce a more stringent assembly, but if coverage is low, the higher sensitivity settings may improve the assembly without sacrificing accuracy by incorporating more reads.

Geneious Prime automatically chooses the most appropriate sensitivity setting for your dataset based on its size, which for our dataset was Medium-Low. We found that the Medium-Low sensitivity setting produced the highest N50, and longest contig, but the Medium sensitivity setting produced fewer contigs overall, and an NA50 identical to the Medium-Low NA50 (see Table 2).

The Low sensitivity setting produces far more short, low coverage contigs, and the longest contig overall. However the longest contig had two putative misassembly breakpoints so we cannot confirm it was correctly assembled, and both NA50 and the longest alignment were lower than for the higher sensitivity assemblies.

Conclusions

The Geneious Prime assembler can produce good quality assemblies of bacterial genomes from Illumina MiSeq data. For this dataset, trimming the raw reads with BBduk and using the default settings for *de novo* assembly produced the best assembly. This assembly required around 5GB of RAM, and could be run on a standard desktop machine in a few hours.

Geneious Prime generally produced a better assembly than SPAdes. The main difference between the Geneious Prime assembler and SPAdes was in the handling of repetitive regions, where Geneious Prime is less conservative and produces longer contigs. Also, the Geneious Prime assembler can handle data from most types of sequencing machine with reads of any length, including paired-end reads and mixtures of reads from different platforms, making it easy to incorporate additional longer reads into the assembly to aid in resolving repetitive regions and structural variation.

Furthermore, this study shows the importance of stringently trimming and filtering the dataset to remove poor quality data prior to assembly. However, merging overlapping reads did not improve the assembly, and may introduce errors if the merging is not conservative enough. Given that merging tools may incorrectly collapse reads in repetitive regions, we suggest letting the assembler handle the reads as pairs rather than merged sequences.

The BBDuk plugin in Geneious R9 greatly expands the read trimming options available to users, meaning the entire *de novo* assembly workflow from raw data to

Table 2. Effect of algorithm on assembly outcome. Geneious assemblies were performed using reads trimmed and filtered with BBDuk but not merged. The Spades assembly is from Markell et al., 2015[5] (NZ_JXUS00000000). Assembly metrics are as in Table 1.

Algorithm	No. of contigs (Total, (>1000bp))	Mean Coverage	Total length	N50	NA50	Longest contig	Longest alignment
Geneious, Low sens	2,194 (146)	58.4	5,502,679	92,660	88,204	382,625	323,400
Geneious, Med-Low sens	**1,424 (108)**	**59.6**	**5,482,554**	**163,893**	**114,793**	**381,532**	**377,897**
Geneious, Med sens	1,118 (101)	63	5,470,162	127,371	114,793	375,525	375,525
Spades	(242)*	31	5,303,239	58,735	56,301	202,085	202,085

1.Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649 (2012).
2.Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075 (2013).
3.Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477 (2012).
4.Markell, J. A., Koziol, A. G. & Lambert, D. Draft Genome Sequence of *Escherichia coli* O157:H7 ATCC 35150 and a Nalidixic Acid-Resistant Mutant Derivative. Genome Announc. 3, (2015).
5. https://sourceforge.net/projects/bbmap/