Sequence Assembly and the NGS Pipeline

Catherine Eason (Wofford College), Amit Upadhyay (JICS, UT-ORNL), Bhanu Rekepalli (JICS, UT-ORNL)

Introduction

Next Generation Sequencing (NGS) is a term that applies to many new sequencing technologies. New datasets are generated by NGS methods faster than raw data can be thoroughly analyzed [2]. Since no currently developed technology produces the genome as fragments or short, a full genome must be “assembled” through computational methods [8]. Sequence assembly is computationally intensive and it is nearly impossible to verify accuracy. There are three main steps in sequence assembly–data quality control, assembly and assembly verification [7]. Many sequence assembly programs do not conduct all three steps and often, two or three separate programs are required to complete the assembly pipeline. Most of the developed programs are a command-line interface and require significant skills in programming or computational science to run successfully. As one can imagine, this creates a critical need for pipelines and interfaces. The objective of the NGS pipeline project is to create pipelines for four different computationally intensive processes required in scientific studies—genome assembly, genome annotation, RNA-seq, and variant calling. This project focuses on the pipeline intended for genome or sequence assembly and the different ways in which assembly can be refined.

Resources

Currently, jobs are run on Nautilus—a super computer located at Oak Ridge National Laboratory. Nautilus uses an SGI Altix UV system and has one UV1000 node containing 128 Intel processors (1024 cores) with 4 terabytes of global shared memory and 8 GPUs [9]. Nautilus was chosen for the large amount of available memory.

Paired-End Sequencing

Figure 1: Diagram showing process of collecting paired-end reads. The genomic DNA is sequenced into fragments which adaptor and primers are attached to (Green, blue, and purple ends). A cluster is formed and the sequences are read starting from both adaptors, producing the paired-end read (Modified from [6]).

Assembly Workflow

Figure 2: Diagram for the complete assembly process, beginning with raw sequence data. The assembled sequences must be checked for accuracy–a difficult step. Green rectangles are the steps that are checked. Blue arrows are steps that have their own process.

Results

Table 1: Summary of the assembled data from the two assembly pipelines.

| Dataset | Size (MB) | N50 (bp) | GC % | Annotation
|--------|----------|----------|-------|-----------
| SPAdes (Bbtools) | 195,337,876 | 58,093 | 45.31 |
| SOAPdenovo2 (Bbtools) | 165,651,506 | 88,083 | 45.32 |
| subset 1 | 94,766,115 | 66,281 | 45.30 |
| subset 2 | 100,571,761 | 88,083 | 45.32 |

Table 2: Statistics for k-mer sizes 21, 33, 55, 71 and a random 50% subset of data's statistics for k-mer sizes 51, 61, 71, 81, and 91. (a) is for the assembly of Trimmomatic trimmed data through SPAdes while (b) is for the same but using SOAPdenovo2. Both tables (a) and (b) show number of contigs, genome size, N50, and GC % statistics for k-mer sizes 21, 33, 55, 71.

Conclusions

Despite Dikow, et al. [4] reporting a genome size of over 6 million base pairs, we suggest that the genome size of Vibrio gazogenes is ~4 to 4.5 million base pairs. Both normalized and quality trimmed data produced genomes of this size. Species closely related to V. gazogenes typically have genomes of ~4.5 to 5 million base pairs.

References


Acknowledgements

The opportunity to work on this project would have not been possible without the National Science Foundation, Joint Institute for Computational Sciences, University of Tennessee, and Oak Ridge National Laboratory. Thank you.