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 proposal 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 data of the *Vibrio gazogenes* genome generated by the Smithsonian Institution [4] was utilized. Due to the large size of the dataset, Trimmomatic and BBtools were used to trim the dataset based on read quality and genome coverage, respectively.

Many assemblers will be tested as the project continues but at this point, the focus is on SPAdes and SOAPdenovo2. The assembled data was run through QUAST, an assembler evaluation tool.

Paired-End Sequencing

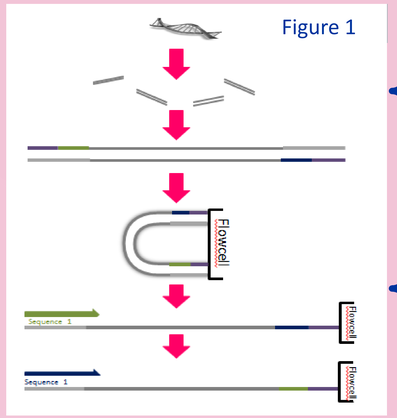

Figure 1: Diagram showing process of collecting paired-end reads. The genomic DNA is sequences into fragments which adaptors and primers are attached to (Green, Blue, and Purple ends). A cluster is formed and the sequences are read starting from both adaptors, producing the paired-end read (Modified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

Assembly Workflow

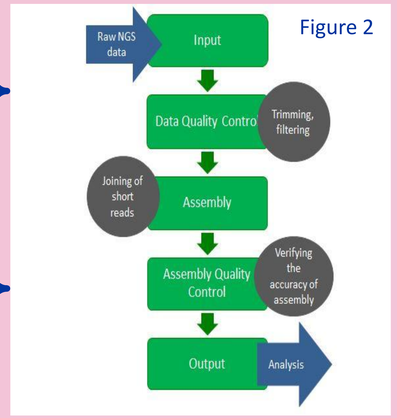

Figure 2: Diagram for the complete assembly process, beginning with raw sequence data. The assembled sequences must be checked for accuracy– a difficult step. Green rectangles are the steps, gray circles a short description. And blue arrows are steps that have their own process.

Results

a.

SPAdes (Trimmomatic)				
Kmer size	# of Contigs	Genome Size	N50	GC %
21	514	4,430,394	17,374	45.27
33	282	4,467,765	54,782	45.27
55	215	4,496,327	68,126	45.27
71	120	4,555,395	246,573	45.32
Subset 51	201	4,468,133	61,386	45.30
Subset 61	193	4,485,523	68,843	45.31
Subset 71	180	4,499,332	79,631	45.32
Subset 81	173	4,510,565	88,093	45.33
Subset 91	88	4,545,153	262,031	45.36

b.

SOAPdenovo2 (Trimmomatic)				
Kmer Size	# of Contigs	Genome Size	N50	GC %
21	16	11,398	690	42.96
33	17	11,766	690	41.00
55	1,385	968,669	685	46.87
71	444	4,448,857	18,563	45.33
Subset 51	1,481	4,321,140	4,296	45.39
Subset 61	309	4,459,372	29,329	45.30
Subset 71	206	4,481,934	55,249	45.30
Subset 81	172	4,499,317	75,768	45.32
Subset 91	159	4,519,076	100,098	45.34

c.

SPAdes (BBtools)				
Kmer Size	# of Contigs	Genome Size	N50	GC %
21	506	4,409,861	17,893	45.29
33	263	4,445,712	49,223	45.28
55	190	4,474,737	65,281	45.30
71	106	4,532,943	167,499	45.31

d.

SOAPdenovo2 (BBtools)				
Kmer Size	# of Contigs	Genome Size	N50	GC %
21	770	4,389,210	9,940	45.29
33	379	4,430,953	24,090	45.30
55	202	4,467,392	62,696	45.30
71	169	4,488,672	81,399	45.35

Tables a-d: (a) is for the assembly of Trimmomatic trimmed data through SPAdes while table (b) is for the same but using SOAPdenovo2. Both table (a) and (b) show number of contigs , genome size , N50 , and GC % 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. (c) is for the assembly of Bbtool trimmed data through SPAdes while (d) is for the same but through SOAPdenovo2. Both tables (c) and (d) show number of contigs, genome size, N50, and GC % statistics for k-mer sizes 21, 33, 55, and 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. Due to the similarity of the assembly statistics between Trimmomatic processed data and Bbtools processed data, we propose that neither tool caused a negative effect on the raw data. A collective script with the ability to run Bbtools (bbnorm and bbtrim), SOAPdenovo2 and/or SPAdes, and QUAST is in development with high hopes of a quick completion.

Future Work

Once the collective assembly script works correctly through a command-line or coding interface, a graphical interface will be integrated. The final script will allow researchers to efficiently and easily complete their assembly projects. While the current collective script is only able to run Bbtools, SOAP or SPAdes, and QUAST, the hope is to give scientists the ability to choose between varying programs and run a job that is specific to their needs. However, a collective script that only runs 3 or 4 programs and contains a mainly static workflow is still beneficial. The script in its current state would significantly decrease the amount of time scientists would spend making each program running smoothly and individually.

References

1. Bankevich, Anton, et al. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology* 19.5 (2012): 455-77. Web.
2. Bubnoff, Andreas Von. "Next-Generation Sequencing: The Race Is On." *Cell* 132.5 (2008): 721-23. Web.
3. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170
4. Dikow, Rebecca B., and William Leo Smith. "Genome-level Homology and Phylogeny of Vibrionaceae (Gammaproteobacteria: Vibrionales) with Three New Complete Genome Sequences." *BMC Microbiology* 13.1 (2013): 80. Web.
5. Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. "QUAST: Quality assessment Tool for Genome Assemblies." *Bioinformatics* 29.8 (2013): 1072-075. Web.
6. Illumina. "Genomic Sequencing." *Data Sheet:* Sequencing (2010): n. pag. Web.
7. Magoc, T., S. Pabinger, S. Canzar, X. Liu, Q. Su, D. Puiu, L. J. Tallon, and S. L. Salzberg. "GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms." *Bioinformatics* 29.14 (2013): 1718-725. Web.
8. Nagarajan, Niranjan, and Mihai Pop. "Sequence Assembly Demystified." *Nature Reviews Genetics* 14.3 (2013): 157-67. Web.
9. NICS. "Nautilus." http://www.nics.utk.edu/computing-resources/nautilus.

Acknowledgements

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

Joint Institute for Computational Sciences — JICS, UT, ORNL, Computational Sciences

OAK RIDGE National Laboratory

THE UNIVERSITY of TENNESSEE — KNOXVILLE, CHATTANOOGA, MARTIN, TULLAHOMA, MEMPHIS

NSF

NICS