SciMedCentral

International Journal of Plant Biology & Research

Editorial

Characterization of the *De novo* Assembly Using Oxford Nanopore Sequencing Data

Kan Liu*

Department of Computer Science and Engineering, University of Nebraska-Lincoln, USA

*Corresponding author

Kan Liu, Department of Computer Science and Engineering, University of Nebraska-Lincoln, USA, Email: kliu7@unl.edu

Submitted: 22 March 2018

Accepted: 24 March 2018

Published: 27 March 2018

ISSN: 2333-6668

EDITORIAL

The MinION of Oxford Nanopore Technology (ONT) is a portable single-molecule sequencer released in 2014. As a portable nanopore-based sequencer [1], with just a USB drive size which can be conveniently used connected to a desktop or laptop using the USB 3.0 or higher interface, MinION contains pores embedded on a membrane that is placed over an electric grid. DNA/RNA molecules sequenced by MinION are basically measured by the ionic flow changes through the pores. As a third-generation commercial sequencing platform product, Nanopore sequencing data can reach up to hundreds of thousands of nucleotides in a single run. Usually an average of 5,000 bp of the product long reads are expected for the DNA sequencing. When the two strands of the target sequence are both successfully base-called, a consensus will be generated as a more accurate output (the “2D” sequence) compared with forward strand only output (“1D” sequence). Such long read sequencing technology enables great capacity in assembling large and complex genome data compared with using short reads only.

However, such long reads are generally much more error-prone (10~30% error rate) than short-read sequencing technologies such as Illumina, which generally makes it less competent in direct usage such as small structural variations (SNP and InDel) detections and other sequence analysis and applications. Nanopore sequencing shows a pattern of error in base calling being context-specific, for example, small variations of SNP such as TAG->TGG, TAC ->TCG are predominant in ONT. Early reports [2], showed about ~35% error rate in ONT data significantly hindered the wide application of this new technology. Therefore, Nanopore data usually needs to be corrected for the preprocessing. With the development of the base calling software and nanopore chemistry, a significant drop in error rate has been achieved [3]. Both short reads-based and long reads self-based correction strategies can be conducted based on the input data sets. Common error correction tools as nanocorrect [4], nanocorr [5], can be used for the correction of Nanopore data.

The basic data processing pipeline using ONT sequencing data contains: preprocessing, error correction, sequence assembly. For genome assembly, it is reported that using nanopore data only at about 30X genome coverage can be sufficient for assembling some small genomes such as *E. coli* [4]. Even after error correction, the assembly using Nanopore long reads is still quite challenging now since many existing assemblers were not designed to implement long reads with high error rate. Therefore, some hybrid methods incorporating both long and short reads for genome assembly are developed. The hybrid and non-hybrid *de novo* assembly strategies are both important depend on the data sets: short read will offer high quality for small region of sequence assembly as well as the scaffolding using paired end information, while Nanopore can recover long repetitive regions which cannot be fully reconstructed using short reads only. The selection between hybrid and non-hybrid assembly methods also depends on other considerations such as the size of the target genome, G+C content bias of the genome composition, the genome complexity (repetitive ratio and multiploidy, etc), also the costs and the bioinformatics analysis to be implemented for the project.

For *de novo* assembly tools can be used similar to PacBio reads such as PBc R and canu [6]. PBc R is an excellent assembler used in PacBio small and large genome *de novo* assembly using either hybrid or self-based methods. Canu can also help effectively assemble MinION data into genomes with high sequence accuracy. SPAdes [7] is another assembler used for both simple and multi-chromosome genomes. Other assemblers using *de Bruijn* graph-based methods such as Velvet [8] and ABySS [9], as well as using Overlap Layout Consensus (OLC)-based methods such as Celera assembler named CABOG [10]. Also other greedy graph-based package such as SSAKE [11], can also be applied in Nanopore reads *de novo* assembly. A survey [12] of the benchmarking of those assemblers in their performance using Nanopore long genome reads showed that an ideal strategy of long read assembly should first choose the OLC-based assemblers for a higher initial N50 value & mean and then use the *de Bruijn* graph-based algorithms for the accuracy improvement. NaS pipeline [13] combines both *de Bruijn* graphs and OLC approach for the error-free DNA reads assembly.

For plant genome sequence research, breakthroughs using Nanopore technology have proved the competence of long read in genome assembly of generating very large and long contiguous sequences (contigs) of complex genomes. For *Oryza coarctata* [14], a tetraploid Asian wild rice of approximately 660 Mb in size, a draft genome of multi-chromosome assembled using both hybrid and non-hybrid assembly strategies demonstrated the ability of *de novo* assembly using Nanopore high-noise single molecule sequencing reads. *Arabidopsis thaliana* genome is also successfully reconstructed using Nanopore reads in a fast and cost-effective implementation. Other plant pathogen genome assemblies using Nanopore genomic dataset such as *Agrobacterium tumefaciens* [15], also provided more evidences on long reads being effectively assembled into multi-chromosomal genomes with small number of contigs and high accuracy.

REFERENCES

1. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. Nat Biotechnol. 2012; 30: 344-348.
2. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, et al., Assessing the Performance of the Oxford Nanopore Technologies Minion. Biomol Detect Quantif. 2015; 3: 1-8.
3. Jain M, Hugh E. Olsen, Benedict Paten, Mark Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016; 17: 239.
4. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015; 12: 733-735.
5. Goodwin S, Gurtowski J, Ethe Sayers S, Deshpande P, Schatz MC, McCombie WR. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. 2015; 25: 1750-1756.
6. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017; 27: 722-736.
7. Bankevich Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, et al. SPAdes: A New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19: 455-477.
8. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18: 821-829.
9. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009; 19: 1117-1123.
10. Myers EW. Toward simplifying and accurately formulating fragment assembly. J Comput Biol. 1995; 2: 275-290.
11. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010; 95: 315-327.
12. Cherukuri Y, Janga SC. Benchmarking of de novo assembly algorithms for Nanopore data reveals optimal performance of OLC approaches. BMC Genomics. 2016; 17: 507.
13. Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, et al. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics. 2015; 16: 327.
14. Mondal TK, Rawal HC, Gaikwad K, Sharma TR, Singh NK. First de novo draft genome sequence of Oryza coarctata, the only halophytic species in the genus Oryza. F1000Res. 2017; 6: 1750.
15. Deschamps S, Mudge J, Cameron C, Ramaraj T, Anand A, Fengler K, et al. Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from Agrobacterium tumefaciens. Sci Rep. 2016; 6: 28625.

Cite this article

Liu K (2018) Characterization of the De novo Assembly Using Oxford Nanopore Sequencing Data. Int J Plant Biol Res 6(2): 1087.