MEDDOCS Open Access Publisher

Journal of Plant Biology and Crop Research

Open Access | Mini Review

Overview of single-molecule long RNA-seq reads and their transcriptomic applications in plant research

***Corresponding Author(s):** Kan Liu,

Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln NE 68588, USA

Email: kliu7@unl.edu

Abstract

Understanding transcriptome regulation and gene functions needs a systematic method to fully unravel the gene expression profiles, including the isoform classification, at both molecule and single-cell levels. Nowadays, alternative splicing and gene-fusion analysis can take advantage of the third-generation sequencing technology such as PacBio ISO-Seq and Oxford Nanopore Technologies (ONT) MinION RNA-Seq [1] strides at the unprecedented speed in plant organs and cells research area.

Received: Jan 24, 2018

Accepted: Mar 24, 2018

Published Online: Apr 10, 2018

Journal: Journal of Plant Biology and Crop Research

Publisher: MedDocs Publishers LLC

Online edition: http://meddocsonline.org/

Keywords: Long read sequencing; Pacbio; Nanopore; Transcriptome

Mini-Review

Both PacBio and ONT utilize the biochemistry/biophysics strategy to capture and direct the very long nucleotide molecule for further base calling. PacBio ISO-Seq uses a similar sequencing-by-synthesis strategy to Illumina sequencing, but the difference is that, PacBio collect individual DNA/RNA molecule for further base calling yet Illumina captures the amplified signals from a clonal cluster of the DNA/RNA fragments. The estimated error rate of PacBio is 13-15% [2]. To increase the quality and accuracy of the read, PacBio adopts the polymerase that repeatedly traversed and replicate the nucleotide. Therefore, the average longest read length can be typically from 10kb to 60kb, which depends on the polymerase lifetime.ONT is a nanopore-based single molecule sequencing technology which was first prototypically released in 2014. ONT developed a portable sequencing device, called MinION, with just a USB flash drive size. In the nanopore-based device, pores are embedded in a membrane which was placed over an electrical detection grid to detect the ionic current. For the library preparation, a hairpin adapter is attached to one end of the DNA/RNA molecule and a protein is used to unwind the DNA/RNA molecule as well as to control the nanopore passing rate of the molecule. ONT directly sequences a single-stranded DNA or RNA molecule by letting it pass through the nanopore and then monitoring the current changes of each type of nucleotide using a base-calling algorithm to obtain the detailed sequence contents.

The long reads of both PacBio and ONT offer more information for the real-time plant transcriptme profile analysis, such as for novel transcript identification, alternative splicing events detection, fusion gene characterization and structural variation.

Cite this article: Liu K. Overview of single-molecule long RNA-seq reads and their transcriptomic applications in plant research. J Plant Biol Crop Res. 2018; 1(1): 1004.

Long read can easily span the repetitive region of the genome that is hard to recover for the short-read data. Thus far, due to the long-read length, high throughput, and low cost, PacBio and ONT data present the opportunity for many important applications in plant genomics and transcriptomics on the horizon.

Both PacBio and ONT showed a high error rate (estimated 10%~30%), which restricts their accuracy and further applications compared with Illumina short reads. PacBio long reads shows an overall higher accuracy and data quality compared with ION based on current data generated in many aspects [3]. PacBio base calling is dependent upon the different signals from the neighborhood background, while ONT relies on the monitoring current changes from the five upstream nucleotides. As a result, both have a pattern that shows the error in base calling being context-specific. Based on the error pattern of long-length reads, SNP and small insertion/deletion calling might be erroneous with a large proportion of false positives. For PacBio and ONT reads, there showed two types of mismatch events CG -> CA and CG ->TG in PacBio and TAG->TGG TAC ->TCG are predominant in ONT. Both PacBio and ONT reads have the same bias pattern of homopolymer to certain nucleotides: A and T in insertion and G and C in deletion [3]. Therefore, long read should be corrected before further analysis. Short read sequencing using the Illumina platform to evaluate and correct the error of long reads is both necessary and powerful in the further transcriptome analysis [4].

Usually the read correction calls for the combination of high quality short reads to computationally polish the error-prone long reads. Some error correction tools and algorithms are available for the read correction. TAPIS [5], LSC [6] and LoRDEC [7] are widely used for PacBio while Nanocorr is developed for ONT data [8]. Nanocorr starts by aligning the short reads to the long nanopore data using sequence aligner such as BLASR [9]. Using the alignment of nearly full-length transcript, and then use the short reads alignment for the correction, which outputs about 67% reduction in the error rate.

Imperfect read alignments of error-rich long reads will also detriment the precise pinpoint of splice site/exon intron boundaries for the multi-exon isoforms. When evaluating the performance (accuracy) of PacBio and ONT long reads, one needs short reads such as Illumina for the identification of the isoform characterization, and reference-guided gene annotation. Using reference library and short reads as the normal, about 15%~30% and 7%~20% splice sites are mistakenly annotated by PacBio and ONT data before and after reads correction [3]. And some quality control tools such as Align QC [3] using the alignment information (BAM file), results from the sequence alignment by mapping tools, such as StringTie [10] to extract statistics of the long-read sequencing including read length, alignment and coverage of the reference genome and gene annotation by generating the XHTML format report to access the analysis results.

Applications in post-transcriptional regulation, such as Alternative Splicing (AS) using long read, have increased the observed frequency of AS in plants. From the pre-Next Generation Sequencing (NGS) era, only about 30% of the plant transcriptome was identified as alternative splicing products, yet for the NGS era it boosts up more than 60% [11]. Researchers have also proven that even in a highly characterized transcriptome such as sorghum [5] and maize [12], identification of genes expression and isoform regulation is still beyond completion. Bo Wang et al. [12] used PacBio data to characterize the maize transcriptome. They found that using short reads only, two-thirds of the splice isoforms can be neglected without the incorporation of PacBio reads. Also, alternative polyadenylation of the 3' end of the sorghum transcriptome showed a vital co-transcriptional modification in different biological conditions. Bing Cheng et al. [13] analyzed the transcriptome diversity and complexity of the tetraploid Arabica coffee (*Coffea arabica*) bean using PacBio ISO-Seq data to uncover the regulation on gene expression based on genome polyploidization, which offered a new horizon of the improvement on plant gene annotation using long read RNA-Seq. Case studies in caffeine and sucrose also proved that the transcriptome diversity and complexity were a result of the novel genes, alternative splicing, alternative polyadenylation, 5'UTR extension, and sub-genome copies. Using Pacbio ISO-Seq, Qiangshan Xu et al. [14] discovered a high proportion of regulated gene isoforms in *Camellia sinensis*, which demonstrates that detection of alternative splicing is far from perfection using short length RNA-Seq data.

Understanding gene expression regulation and corresponding function dynamics needs a genome-wide method suitable for cover both gene expression coverage and complexity characterization. Both PacBio and ONT long reads showed technical breakthroughs and are suitable for full length transcripts analysis and single cell transcriptome investigation. In addition, hybrid-Seq strategies in transcriptome analysis provided a new potential in most transcriptome analyses. Jason L Weirather et al. [3] showed two hybrid-Seq datasets to compare the correctness of isoforms detection: PacBio + Illumina and ONT + Illumina. Although PacBio and ONT long RNA-Seq offer a large quantity of novel transcripts, further validation of the alternative splicing events by PCR is mandatory to get the conclusion [15]. Robust quality/accuracy evaluation and systematic analytical tools for the rapidly-evolving long read technology are greatly needed and should be on the horizon to achieve better applications.

References

1. Jain M, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016; 17: 239.
2. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics. 2015; 13: 278-289.
3. Weirather J, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. 2017; 6.
4. Koren S, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012; 30: 693-700.
5. Abdel-Ghany SE, et al. A survey of the sorghum transcriptome using single-molecule long reads. Nature Communications. 2016; 7: 11706.
6. Au KF, et al. Improving PacBio long read accuracy by short read alignment. PLoS One. 2012; 7: e46679.
7. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014; 30: 3506-3514.
8. Goodwin S, et al. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res. 2015; 25: 1750-1756.
9. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012; 13: 238.

10. Pertea M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015; 33: 290-295.

11. Syed NH, et al. Alternative splicing in plants--coming of age. Trends Plant Sci. 2012; 17: 616-623.

12. Wang B, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 2016; 7: 11708.

13. Cheng B, Furtado A, Henry RJ. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. Gigascience. 2017; 6: 1-13.

14. Xu Q, et al. Transcriptome Profiling Using Single-Molecule Direct RNA Sequencing Approach for In-depth Understanding of Genes in Secondary Metabolism Pathways of Camellia sinensis. Frontiers in Plant Science. 2017; 8.

15. Marinov GK. On the design and prospects of direct RNA sequencing. Brief Funct Genomics. 2017; 16: 326-335.