Send Orders for Reprints to reprints@benthamscience.net

Current Bioinformatics, **2019,** *14,* **1-8** **1**

REVIEW ARTICLE

Single-molecule Real-time (SMRT) Isoform Sequencing (Iso-Seq) in Plants: The Status of the Bioinformatics Tools to Unravel the Transcriptome Complexity

Yubang Gao[1], Feihu Xi[1], Hangxiao zhang[1], Xuqing Liu[1], Huiyuan Wang[1], Liangzhen zhao[1], Anireddy S.N. Reddy[2] and Lianfeng Gu[1,*]

[1]*Basic Forestry and Proteomics Research Center, College of Forestry, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, College of life science, Fujian Agriculture and Forestry University, Fuzhou 350002, P.R. China;;* [2]*Department of Biology, Program in Molecular Plant Biology, Program in Cell and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523, USA*

Abstract: ***Background*:** The advent of the Single-molecule Real-time (SMRT) Isoform Sequencing (Iso-Seq) has paved the way to obtain longer full-length transcripts. This method was found to be much superior in identifying full-length splice variants and other post-transcriptional events as compared to the Next Generation Sequencing (NGS)-based short read sequencing (RNA-Seq). Several different bioinformatics tools to analyze the Iso-Seq data have been developed and some of them are still being refined to address different aspects of transcriptome complexity. However, a comprehensive summary of the available tools and their utility is still lacking.

***Objective*:** Here, we summarized the existing Iso-Seq analysis tools and presented an integrated bioinformatics pipeline for Iso-Seq analysis, which overcomes the limitations of NGS and generates long contiguous Full-length Non-chimeric (FLNC) reads for the analysis of post-transcriptional events.

***Results*:** In this review, we summarized recent applications of Iso-Seq in plants, which include improved genome annotations, identification of novel genes and lncRNAs, identification of full-length splice isoforms, detection of novel Alternative Splicing (AS) and alternative polyadenylation (APA) events. In addition, we also discussed the bioinformatics pipeline for comprehensive Iso-Seq data analysis, including how to reduce the error rate in the reads and how to identify and quantify post-transcriptional events. Furthermore, the visualization approach of Iso-Seq was discussed as well. Finally, we discussed methods to combine Iso-Seq data with RNA-Seq for transcriptome quantification.

***Conclusion*:** Overall, this review demonstrates that the Iso-Seq is pivotal for analyzing transcriptome complexity and this new method offers unprecedented opportunities to comprehensively understand transcripts diversity.

A R T I C L E H I S T O R Y

Received: April 04, 2018
Revised: July 20, 2018
Accepted: August 22, 2018

DOI:
10.2174/1574893614666190204151746

Keywords: Pacific Bioscience (PacBio), SMRT isoform sequencing (Iso-Seq), Next-generation sequencing (NGS), Alternative splicing (AS), Alternative polyadenylation (APA), Genome annotation, Novel genes.

1. INTRODUCTION

In the past few years, the single-molecule real-time (SMRT) isoform sequencing (Iso-Seq) method was developed by Pacific BioSciences (PacBio), which is the new Third-Generation Sequencing (TGS) platform, which generates much longer reads and differs enormously from the Next Generation Sequencing (NGS) [1]. Iso-Seq using PacBio technologies stimulated widespread interest and revolutionized genomic and transcriptome studies because of the improvement of reads length [2]. Iso-Seq offers many advantages for a number of different applications, including inferring splice isoforms [3], identification of previously unknown novel genes [3], detecting alternative splicing (AS) and alternative polyadenylation (APA) [3, 4], identifying fusion gene [5] and lncRNA regulation [6], *etc*. Here, we evaluate various applications of Iso-Seq to demonstrate its broader utility in obtaining a complete transcriptomic landscape in plants.

To perform these analyses, various Iso-Seq analysis tools have been established to achieve different goals. In this re-

* Address Correspondence to this author at the Fujian Agriculture and Forestry University, Fuzhou, P.R. China; Tel: 86-591-83590305; Fax: 86-591-88250137; E-mail: lfgu@fafu.edu.cn

1574-8936/19 $58.00+.00

view, we present a survey of the available computational pipelines used for Iso-Seq analysis, including the following aspects: full-length non-chimeric (FLNC) reads extraction, error correction [7], aligning [8], isoform clustering [9], identification of AS and APA events [3]. Equipped with these tools, more insights about transcriptome complexity will be gained from Iso-Seq analysis.

Most of the studies are reference-based, however, many species have no well-annotated reference genome. Thus the perspective of Iso-Seq for species without a reference genome is also introduced in this review. With the exponential development of Iso-Seq and the availability of computational pipelines, the applications for different aspects of post-transcriptional events will disclose the majority of transcriptome complexity.

2. SEQUENCING STRATEGY OF ISO-SEQ

2.1. Library Preparation

Generally, PacBio long-read sequencing libraries are constructed using poly(A)-selected RNAs [6]. The libraries are made using the standard PacBio Iso-Seq experimental protocols, which have been described in the manufacturer's instruction in previous studies [3, 4, 10-12]. Briefly, the first-strand cDNA is synthesized using Clontech SMARTer PCR cDNA Synthesis Kit [4, 12]. Then, the second-strand cDNA is amplified with Long-Distance PCR (LD-PCR) using Phusion High-Fidelity DNA Polymerase for BluePippin size selection [13]. The cDNA products with different sizes can be made into SMRTbell template libraries [12].

2.2. PacBio Sequencing

Iso-Seq is performed using a sequencing-by-synthesis method [14]. The sequencing unit of Iso-Seq is called Zero-mode Waveguide (ZMW), which allows the immobilization of DNA template. When SMRTbell cDNA comes into the ZMW, the DNA polymerase at the bottom of ZMW will bind to it through the hairpin adaptor [1]. Then the replication begins, a fluorescent-labeled nucleotide is added during template-directed incorporation and releases a fluorescent tag, which can be captured in real time. The cDNA will be sequenced many times if the lifetime of the polymerase is long enough. The read produced in this situation is called polymerase read, which is further partitioned to form one or more subreads. The subreads taken from a single ZMS determine the Circular Consensus Sequence (CCS) reads [15]. Using scripts provided by SMART Analysis software (http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/), FASTA format files could be generated.

2.3. Comparison with NGS

At present, Illumina HiSeq 2500 takes about 60 h to produce 8 $\times 10^9$ paired reads with 2×125 nts length (https://www.illu-mina.com). PacBio RS II takes about 4 h to generate about 5×10^4 long reads with 1×10^4 nts on average (www.pacb.com). From the comparison, we can see that the advantage of the second and third generation sequencing platform is the high yield and read length, respectively. For example, the length of transcripts generated from PacBio RS II can be 10 kb, which can cover the size distribution of most transcripts. The disadvantage for the second and third generation sequencing platform is the short reads and low yield, respectively. The boundaries of each transcript are hard to resolve using NGS because of short reads [16]. Also it is difficult to detect the actual combinations of splice-site usage by NGS, because it is difficult to reconstruct full-length splice isoforms [10]. However, Iso-Seq can detect splicing isoforms with high confidence since there is no need to assemble sequence, which suggests that Iso-Seq has a great advantage for identification of isoforms when the complexity increased [10]. However, the low sequence depth of Iso-Seq limits downstream quantification analysis. Hybrid sequencing integrates both techniques and complements the strengths and weaknesses of NGS and TGS [1].

3. THE APPLICATIONS OF ISO-SEQ

3.1. Annotating Genome, and Novel Genes

Due to the ability to sequence longer reads by Iso-Seq, it permits direct characterization of transcript diversity without any limitations that are normally encountered with short-read RNA-seq. Thus, the first contribution for Iso-Seq is to accurately infer gene models by generating full-length transcripts without further assembly [17]. Meanwhile, rare transcripts can be detected using FLNC forms. Thus Iso-Seq is suited for identification of novel genes. For example, Iso-Seq studies revealed 2,171, 8,091 and 3,026 novel genes in transcript clusters that did not overlap with any annotated genes in *Sorghum bicolor* [3], *Phyllostachys edulis* [4] and *Triticum aestivum* [17], respectively.

In addition to the identification of novel genes, Iso-Seq can correct misannotated gene models (Fig. **1**). In *Phyllostachys edulis*, for example, we previously described the corrected annotation for 2,241 genes using Iso-Seq [4]. Another study in Reddy's lab uncovered 178 annotated genes that overlap more than one transcript assemblies in *Sorghum bicolor* [3]. Considering the read length improvements, transcriptomes of more and more species might be re-sequenced using Iso-Seq to precisely annotate the gene models.

3.2. Identification and Experimental Validation of Splice Isoforms

Advances in PacBio long-read sequencing have allowed us to obtain full-length sequences and thus accurately provide isoforms, which presents an appealing application for identification and validation of long splicing isoforms (Fig. **1**). Based on the previous studies, there are 111,151, 42,280, 16,241 non-redundant isoforms, which are detected in *Zea mays* [10], *Phyllostachys edulis* [4] and *Salvia miltiorrhiza* [13], respectively. Iso-Seq analysis of sorghum seedling transcriptome revealed over 11,000 novel splice isoforms [3]. These studies together showed that Iso-Seq is suitable to identify isoforms with subtle variation. More importantly, Iso-Seq has even been adopted to systematically validate the splice isoforms based on short reads assembly, since Iso-Seq can avoid the assembling issue of NGS [6].

3.3. Characterization of AS and APA

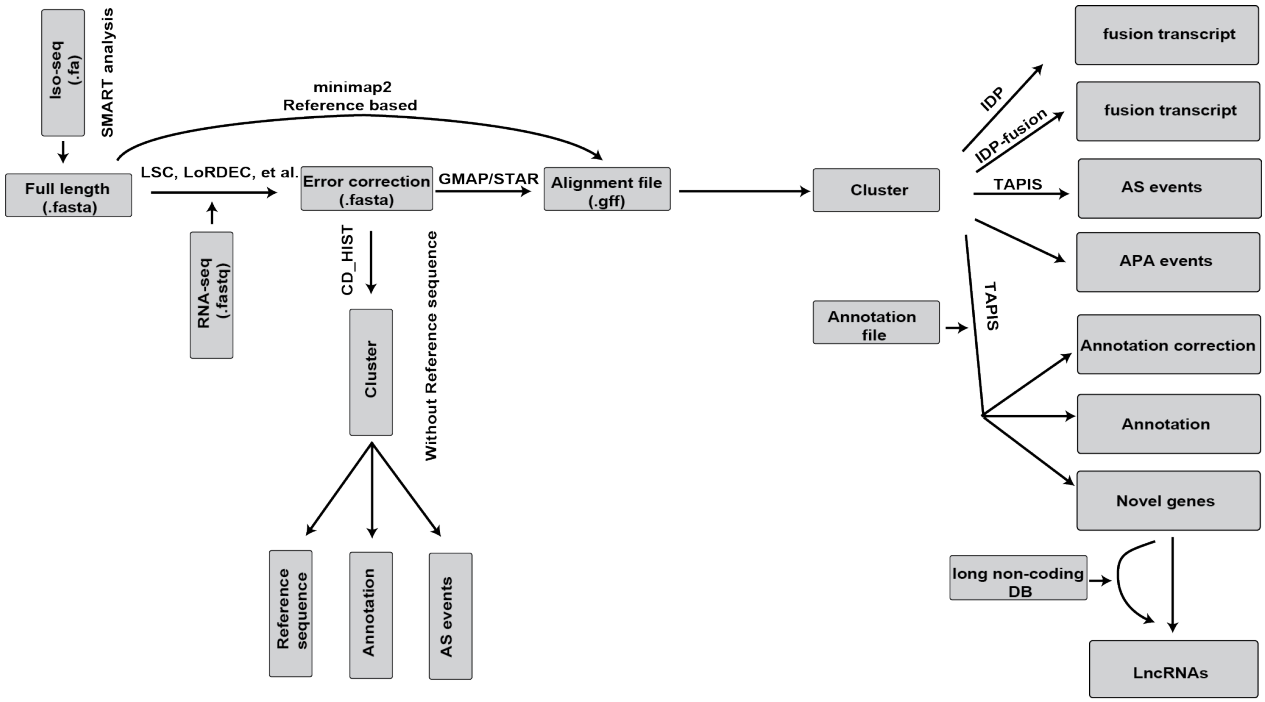

Fig. (1). The scheme of the bioinformatics tool and its biological application for Iso-Seq.
The figure summarized the existing Iso-Seq analysis tools, such as error correction using reference-based method (minimap2) or NGS-based method (LSC). After error correction, long reads were mapped to the genome using GMAP or STAR for downstream analysis. CD-HIT was required for clustering Iso-Seq reads without a reference genome. Iso-seq was widely used in including identification of novel genes and lncRNAs, identification of full-length splice isoforms, detection of AS and APA events.

In general, most of the introns from transcripts will be removed and the 5' cap and 3' poly (A) tail will be added [18]. AS is a powerful post-transcription mechanism to generate multiple isoforms from a single locus and the resulting isoforms may code for different proteins or may exhibit differential stability due to Nonsense-mediated Decay (NMD) or other mechanisms [19]. APA generates transcripts with different 3' end and the APA, regulates gene expression and has a significant role in RNA transport, localization, stability, and translation [3]. Studies based on NGS have shown that a large number of genes are regulated by AS and APA [20-22].

A recent study has shown that polyadenylation can regulate the splicing of a subset of genes [23]. All possible combinations arise from the AS and APA, resulting in transcript diversity and complexity, which shape the plant transcriptome [3, 4, 18]. Iso-Seq is particularly well suited not only to facilitate the investigation of AS [24], but to address the role of APA as well [3]. Iso-Seq is able to detect AS at a relatively low error rate since it avoids the step of transcriptome assembly, which has been applied widely in AS identification (Fig. **1**). Consequently, by employing Iso-Seq, 10,053, 172,743, 21,154 AS events had been detected in *Sorghum bicolor* [3], *Zea mays* [10], and *Phyllostachys edulis* [4], respectively. Using a hybrid sequencing approach, pre-mRNAs form about 40% of the detected gene loci *in Salvia miltiorrhiza* showed AS events [13].

For APA identification, Poly (A) Site Sequencing (PAS-Seq) is one of the powerful high throughput methods [25]. However artifacts due to internal priming limits this method [26, 27]. The direct RNA sequencing (DRS) technology is another method to detect APA without internal priming issue [28]. However, both DRS and PAS-Seq provide only a part of the full-length transcripts [27]. Iso-Seq presents the whole transcripts and also can avoid the internal priming issues, thus it has been reported by recent studies. For example, using the approach of Iso-Seq, 7,700 genes containing two or more polyadenylation sites have been detected in *Sorghum bicolor* [3]. However, the low sequence depth of Iso-Seq does not permit quantification analysis. A recent study on *Phyllostachys edulis* used a method that combined the NGS with Iso-Seq to identify 1,224 differential APA sites [4].

3.4. Fusion Transcripts Identification

Fusion transcript is a chimeric RNA encoded by two separate genes by subsequent trans-splicing [29]. Fusion transcripts exist in diverse plant species [10, 30], which have been validated by an experiment to confirm the authenticity. In addition to plant fusion transcripts, fusion genes have also been identified and quantified from the MCF-7 breast cancer cells [5], which suggests that fusion genes are a common phenomenon in plant and animal. Iso-Seq is particularly useful in the detection of fusion transcripts with high confidence due to the lack of assembly (Fig. **1**). In total, 1,430 fusion transcripts had been detected in *Zea mays* using Iso-Seq [10].

3.5. LncRNA Identification

Recently, several studies showed the successful identification of long non-coding RNAs (lncRNAs) using Iso-Seq in plants. For example, in *Phyllostachys edulis*, the lncRNAs database, such as GreeeNC and CANTATAdb were adopted to search for the homology of lncRNAs [4]. In *Sorghum bi-*

Table 1. Summary of the software used in Iso-Seq analysis.

Name	Types	Web Sites
SMRT-Analysis	Extract and classify reads of insert	https://github.com/PacificBiosciences/SMRT-Analysis
LSC	Error correction	https://www.healthcare.uiowa.edu/labs/au/LSC
LoRDEC	Error correction	https://github.com/cmonjeau/docker-lordec
PacBioToCA	Error correction	http://wgs-assembler.sourceforge.net/wiki/index.php/PacBioToCA
TAPIS	Error correction	https://bitbucket.org/comp_bio/tapis
LoRMA	Error correction	http://www.cs.helsinki.fi/u/lmsalmel/LoRMA/
IDP	Isoform Detection and Prediction	https://www.healthcare.uiowa.edu/labs/au/IDP/
GMAP	Align	http://research-pub.gene.com/gmap/
BLAT	Align	http://hgwdev.cse.ucsc.edu/~kent/src/
blasr	Align	https://github.com/PacificBiosciences/blasr
minimap2	Align	https://lh3.github.io/minimap2/
minialign	Align	https://github.com/ocxtal/minialign
STAR	Align	https://github.com/alexdobin/STAR
BWA-SW	Align	http://bio-bwa.sourceforge.net/
pbdagcon	Builds consensus sequences	https://github.com/PacificBiosciences/pbdagcon
quiver	Polish consensus isoforms and call variants	https://github.com/PacificBiosciences/GenomicConsensus
IDP-fusion	Isoform Detection and Prediction	https://www.healthcare.uiowa.edu/labs/au/IDP-fusion/
TAPIS	Isoform Detection and Prediction	https://bitbucket.org/comp_bio/tapis
TAPIS	Alternative polyadenylation sites prediction	https://bitbucket.org/comp_bio/tapis
IsoSeq_AS_de_novo	Alternative splicing	https://github.com/liuxiaoxian/IsoSeq_AS_de_novo
pbtranscript-ToFU	Fusion transcript detection	https://github.com/PacificBiosciences/pbtranscript
Iso-Seq Browser	Visualization	https://github.com/goeckslab/isoseq-browser
Iso-View	Visualization	https://github.com/JMF47/IsoView
MatchAnnot	Visualization	https://github.com/TomSkelly/MatchAnnot

color, novel genes without coding potential or exhibiting sequence identity to miRNA were analyzed as candidates of lncRNAs [3]. In *Salvia miltiorrhiza*, Coding Potential Calculator (CPC) was used to predict lncRNAs with the input data of non-redundant sequence [13]. In *Zea mays*, high-confidence known datasets of lncRNAs were identified using modeling of a predictor of long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme (PLEK), and software like EMBOSS to exclude PacBio isoforms with encoding ability [10]. Taken together, these studies show that Iso-Seq is well suited for characterization of lncRNAs also.

4. BIOINFORMATICS TOOLS OF ISO-SEQ

4.1. Reads of Insert Extraction and Reads Classification (Full-length or Non-full-length)

Pacific BioSciences developed SMRT tool to extract 'Reads of Insert'. ToFu script from SMRT Analysis package can search sequencing adapters. If a read contains both 5'- and 3'- cDNA primers, and a poly (A) tail signal preceding the 3' read, then the reads will be regarded as FLNC reads [4, 12]. The entire Iso-Seq flowchart can be found in Fig. (**1**). At present, Iterative Clustering for Error Correction (ICE) algorithm iteratively classifies full-length non-chimeric CCS reads into isoform level clustering to remove redundancy by collapsing redundant transcripts. Then Quiver finds the template sequencing and polishes consensus isoforms from Iso-Seq (https://github.com/ben-lerch/IsoSeq-3.0/blob/master/README.md#cluster-options).

4.2. Error Correction of Iso-Seq Reads

Although the length of Iso-Seq reads is longer than NGS, it still has a limitation due to high error rate [31]. Thus the quality control is a key step for downstream analysis. Three methods to solve this problem have been shown below. One major method is to perform error correction with Illumina short-read data using LSC [7], LoRDEC [32], and PacBi-

oToCA [31]. Specifically, it is worth noting that another method is reference-based error correction. To some extent, Basic Local Alignment with Successive Refinement (BLASR) [33], minimap2 [34], as well as minialign (https://github.com/ocxtal/minialign) are three alignment programs designed for PacBio long reads with high insertion and deletion error rate. However, BLASR is not a splice-aware aligner and cannot be used to align transcript sequence from Iso-Seq to reference genome. Evaluation on these software shows that minimap2 presents higher mapping accuracy than minialign [34]. Thus minimap2 is highly recommended for Iso-Seq alignment. Reference-based error correction was also adopted by Transcriptome Analysis Pipeline from Isoform Sequencing (TAPIS) [3]. Apart from the error correction strategies offered by the reference genome or the short-read data generated by NGS, there is a de novo corrector called LoRMA (Table **1**), which uses de Bruijn graph from long reads only to implement accurate self-correction.

After error correction, long reads can be mapped to the genome using the Genomic Mapping and Alignment Program (GMAP) [8] or STAR [35] for downstream analysis. All the corresponding software information can be found in Table **1**.

GMAP and STAR differ from Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW) [36] and the BLAST-Like Alignment Tool (BLAT) [37]. GMAP and BWA-SW perform better than hashing-based software, such as BLAT [8, 36]. STAR has been developed for mapping full-length RNA sequences and non-canonical splices and fusion transcripts [35], which outperforms BLAT by higher accuracy of alignment and more than two orders speed. BWA-SW and BLAT are not a splice-aware aligners, which cannot be used in Iso-Seq alignment. With this in mind we recommend splice-aware aligner, such as GMAP and STAR for Iso-Seq alignment after error correction. Evaluation of these tools show that STAR failed on highly error-prone long reads. However, GMAP always showed the best alignment results and produced the highest alignment rates using the least memory [38]. Though BLAT is not splice-aware aligner, which cannot be used for Iso-seq alignment to reference genome, BLAT can be used for species without reference genome to find INDEL by pair-wise comparison of FLNC reads from Iso-Seq.

4.3. Identification of Alternative Splicing and Alternative Polyadenylation

At present, Isoform Detection and Prediction (IDP) [39] is one wonderful tool to combine the second and third generation sequencing data to identify the transcriptional isoforms by genome-wide relatively long reads alignments. TAPIS is another tool that clusters Iso-Seq reads into unique splice isoforms by searching for the reads with the overlapped genome coordinates [3].

Though the long isoforms are quite informative, the user may also want to get the coordinate information of the major AS events, specifically for the exon skipping events, intron retention, alternative 5' donor and alternative 3' donor sites in plants, which is the foundation of subsequent downstream differential analysis. In order to achieve this ultimate goal, several tools, including TAPIS [3], rMATS [40], Program to Assemble Spliced Alignments (PASA) pipeline [41], and ASTALAVISTA [42] are available for the identification of AS events using both the GMAP alignment of TGS and StringTie [43] assemblies of NGS. TAPIS [3], can identify both AS events and APA only from TGS.

4.4. Fusion Transcript Detection

The principles used to identify fusion transcripts in a previous study are as follows: a) Full-length transcripts map to two or more loci in the genome; b) each mapped locus must be at least 100kb apart and must align with at least 10% of the transcript; c) the combined alignment coverage must be at least 99% [10]. The script of fusion_finder.py from PacBio pbtranscript-ToFU package provides fusion transcript detection (https://github.com/PacificBiosciences). IDP-fusion also provides another candidate to detect gene fusion events by integrating NGS and PacBio long reads sequencing [5].

4.5. Visualization

The track files for genome browser, such as GFF files, can be used for the visualizing splicing isoforms of Iso-Seq. Comparing with the long-read RNA-seq data visualization bioinformatics tools, MatchAnnot, Iso-View and Iso-Seq Browser (ISB) are three specialized tools (Table **1**) to zoom in or out, highlight exon, dynamic clustering of isoforms, and generating publication-ready figures of Iso-Seq [44].

4.6. Analysis of Iso-Seq Reads without a Reference Genome

Although there are several tools for Iso-Seq analysis, these tools can only be applied for reference-based analysis. There is only one Iso-Seq analysis pipeline, which is specifically designed for AS analysis in species without reference genomes [19]. The method originated from the identification of AS events by searching for the insertion in the clustering transcripts from each gene locus [19, 45, 46]. Thus clustering program, such as widely used CD-HIT [9] is required for the obtaining final unique transcripts (Fig. **1**). Then all-VS-all BLAT or BLAST will be used for the identification of highly similar region to infer the insertion segmentation for AS identification [19].

Thought part analysis of Iso-Seq is reference independent, the genome sequence is still valuable for most of the downstream analysis including promoter and intron. In plant, high-quality de novo assembling of grass *Oropetium thomaeum* [47], sunflower [48] and citrus [49] have been reported using the PacBio RS II platform for genome assemble and obtained high quality reference sequence [47]. For example, the assembly of *Oropetium thomaeum* obtains high quality contigs with an N50 length of 2.4 megabases using hierarchical genome-assembly process (HGAP) [50], which suggests that sequencing of massive plant genomes makes great leaps with the advent of TGS.

5. DISCUSSION

Compared with NGS, the great advantage of PacBio long-read sequencing is that it generates superior contiguous long reads. The disadvantage of PacBio long-read sequenc-

ing is that it is expensive and most of the studies cannot reach the necessary sequence depth for statistical quantification, particularly for low abundance transcripts [13]. Currently, Illumina Platform is still necessary for the quantification. Only a few studies have a massive amount of long reads to do quantification analysis [10]. Most of the present studies used different sequencing strategies to obtain optimal result [4]. Nevertheless, emerging studies use a combination of PacBio long-read sequencing and NGS for qualitative and quantitative research [4, 5]. Also, several software, such as IDP-fusion [5] and IDP [39], can use the hybrid sequencing method to identify gene fusion events and transcript isoforms, which are more accurate than with either NGS or NGS method. However, the newest PacBio sequencing system, known as Sequel, can obtain better sequence depth than the PacBio RS instrument and generate reads with average length more than 10 kb. Thus, it can be anticipated that Iso-Seq will be affordable and also suitable for quantification in a routine manner at a reasonable cost in the future.

In this review, we mainly focused on the current major applications of Iso-Seq (Fig. **1**). Additionally, we believe that Iso-Seq will have broader application to many aspects of the complex transcriptome in the future, such as Natural antisense transcripts and alternative transcript initiation and so on, since it represents full-length transcripts after filtering with SMART analysis pipeline. As the Iso-Seq method continues to evolve, the applications will also grow. We are confident that plant researchers will benefit enormously from taking advantages of this method, which can finally unravel the complexity of transcriptomes under normal growth and developmental conditions and in response to diverse stresses in different plant species.

At present, different bioinformatics tools are available for different data processing steps. The linking of multiple programs allows users to start with a FASTA file obtained from Iso-Seq and end up with results pertinent to many aspects of the transcriptome. To accomplish this, various tools for Iso-Seq analysis, such as FLNC reads extraction, error correction, aligning, isoform clustering, have been linked to achieve these goals (Fig. **1**). However, it will be a great challenge for end-users, typically biological scientists, to use these different tools without knowledge about scripting language, such as Unix shell, Perl and Python *et al.* Thus one promising trend for Iso-Seq pipeline is the containerization of all available Iso-Seq analysis software together with third party dependencies to provide a one-stop bioinformatics pipeline for Iso-Seq analysis.

LIST OF ABBREVIATIONS

AS	=	Alternative splicing
APA	=	Alternative polyadenylation
BLAT	=	BLAST-Like Alignment Tool
CCS	=	Circular consensus sequence
CPC	=	Coding potential calculator
DRS	=	Direct RNA sequencing
FLNC	=	Full-length non-chimeric
GMAP	=	Genomic Mapping and Alignment Program
IDP	=	Isoform Detection and Prediction
ICE	=	Iterative Clustering for Error Correction
Iso-Seq	=	Isoform Sequencing
lncRNAs	=	Long non-coding RNAs
NGS	=	Next generation sequencing
PASA	=	Program to Assemble Spliced Alignments
PLEK	=	Predictor of Long Non-coding RNAs and Messenger RNAs based on an improved *k*-mer scheme
PAS-Seq	=	Poly (A) Site Sequencing
NGS	=	Next Generation Sequencing
SMRT	=	Single-molecule real-time
TAPIS	=	Transcriptome Analysis Pipeline from Isoform Sequencing
TGS	=	Third generation sequencing

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work has been supported by the National Key Research and Development Program of China (2018YFD0600101 and 2016YFD0600106), the National Natural Science Foundation of China Grant (Grant No. 31570674 and 31800566), International Science and Technology Cooperation and Exchange Fund from Fujian Agriculture and Forestry University (KXGH17016), the Natural Science Foundation of Fujian Province (Grant No. 2018J01608) and the Department of Energy Office of Science, Office of Biological and Environmental Research (Grant No. DE-SC0010733).

REFERENCES

[1] Rhoads A, Au KF. PacBio sequencing and its applications. Genomics, proteomics & bioinformatics. 2015;13(5):278-89.

[2] Gonzalez-Garay ML. Introduction to isoform sequencing using pacific biosciences technology (Iso-Seq). Transcriptomics and Gene Regulation: Springer; 2016. p. 141-60.

[3] Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, *et al.* A survey of the sorghum transcriptome using single-molecule long reads. Nature communications. 2016;7.

[4] Wang T, Wang H, Cai D, Gao Y, Zhang H, Wang Y, *et al.* Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (Phyllostachys edulis). The Plant Journal. 2017.

[5] Weirather JL, Afshar PT, Clark TA, Tseng E, Powers LS, Underwood JG, *et al.* Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. Nucleic acids research. 2015;43(18):e116-e.

[6] Li S, Yamada M, Han X, Ohler U, Benfey PN. High-resolution expression map of the Arabidopsis root reveals alternative splicing and lincRNA regulation. Developmental cell. 2016;39(4):508-22.

[7] Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio long read accuracy by short read alignment. PloS one. 2012;7(10):e46679.

[8] Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859-75.

[9] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-2.

[10] Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature communications. 2016;7.

[11] Zhu FY, Chen MX, Ye NH, Shi L, Ma KL, Yang JF, *et al.* Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings. The Plant Journal. 2017.

[12] Xu Q, Zhu J, Zhao S, Hou Y, Li F, Tai Y, *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;8:1205.

[13] Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, *et al.* Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. The Plant Journal. 2015;82(6):951-61.

[14] Metzker ML. Sequencing technologies—the next generation. Nature reviews genetics. 2010;11(1):31-46.

[15] Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic acids research. 2010;38(15):e159-e.

[16] Pelechano V, Wei W, Jakob P, Steinmetz LM. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. Nature protocols. 2014;9(7):1740-59.

[17] Dong L, Liu H, Zhang J, Yang S, Kong G, Chu JS, *et al.* Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. BMC genomics. 2015;16(1):1039.

[18] Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nature biotechnology. 2015;33(7):736-42.

[19] Liu X, Mei W, Soltis PS, Soltis DE, Barbazuk WB. Detecting Alternatively Spliced Transcript Isoforms from Single-Molecule Long-Read Sequences without a Reference Genome. Molecular Ecology Resources. 2017.

[20] Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, *et al.* Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proceedings of the National Academy of Sciences. 2011;108(30):12533-8.

[21] Zhang Y, Gu L, Hou Y, Wang L, Deng X, Hang R, *et al.* Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. Cell research. 2015;25(7):864-76.

[22] Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, *et al.* Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome research. 2010;20(1):45-58.

[23] Muniz L, Davidson L, West S. Poly (A) polymerase and the nuclear poly (A) binding protein, PABPN1, coordinate the splicing and degradation of a subset of human pre-mRNAs. Molecular and cellular biology. 2015;35(13):2218-30.

[24] Li Y, Dai C, Hu C, Liu Z, Kang C. Global identification of alternative splicing via comparative analysis of SMRT-and Illumina-based RNA-seq in strawberry. The Plant Journal. 2017;90(1):164-76.

[25] Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA. 2011;17:761-72.

[26] Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, *et al.* Oligo (dT) primer generates a high frequency of truncated cDNAs through internal poly (A) priming during reverse transcription. Proceedings of the National Academy of Sciences of the United States of America. 2002;99:6152-6.

[27] Sherstnev A, Duc C, Cole C, Zacharaki V, Hornyik C, Ozsolak F, *et al.* Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. Nat Struct Mol Biol. 2012;19:845-52.

[28] Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, *et al.* Direct RNA sequencing. Nature. 2009;461:814-8.

[29] Li H, Wang J, Mor G, Sklar J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. Science. 2008;321(5894):1357-61.

[30] Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, *et al.* Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome research. 2010;20(5):646-54.

[31] Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology. 2012;30(7):693-700.

[32] Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30(24):3506-14.

[33] Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics. 2012;13(1):238.

[34] Li H. Minimap2: versatile pairwise alignment for nucleotide sequences. arXiv. 2017;1708.

[35] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

[36] Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26(5):589-95.

[37] Kent WJ. BLAT—the BLAST-like alignment tool. Genome research. 2002;12(4):656-64.

[38] Krizanovic K, Echchiki A, Roux J, Sikic M. Evaluation of tools for long read RNA-seq splice-aware alignment. bioRxiv. 2017:126656.

[39] Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, Williams BA, *et al.* Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences. 2013;110(50):E4821-E30.

[40] Shen S, Park JW, Lu Z-x, Lin L, Henry MD, Wu YN, *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences. 2014;111(51):E5593-E601.

[41] Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC genomics. 2006;7(1):327.

[42] Foissac S, Sammeth M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Res. 2007;35:W297-W9.

[43] Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015;33(3):290-5.

[44] Hu J, Uapinyoying P, Goecks J. Interactive analysis of Long-read RNA isoforms with Iso-Seq Browser. bioRxiv. 2017:102905.

[45] Zhou R, Moshgabadi N, Adams KL. Extensive changes to alternative splicing patterns following allopolyploidy in natural and

resynthesized polyploids. Proceedings of the National Academy of Sciences. 2011;108(38):16122-7.

[46] Ner-Gaon H, Leviatan N, Rubin E, Fluhr R. Comparative cross-species alternative splicing in plants. Plant physiology. 2007;144(3):1632-41.

[47] VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, *et al.* Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature. 2015;527(7579):508-11.

[48] Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, *et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017;546(7656):148-52.

[49] Wang X, Xu Y, Zhang S, Cao L, Huang Y, Cheng J, *et al.* Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. Nature Genetics. 2017;49(5):765-72.

[50] Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods. 2013;10(6):563-9.