Chapter

Transcriptome Atlas by Long-Read RNA Sequencing: Contribution to a Reference Transcriptome

Dong Jin Lee and Chang Pyo Hong

Abstract

The recent emergence of long-read transcriptome sequencing has helped improve the overall accuracy of gene prediction compared with that by short-read RNA-Seq. In addition, the technology can offer a more comprehensive view of functional genomics in uncharacterized species with an efficient full-length unigene build and high-precision gene annotation, thus being efficient in developing transcriptome data resources from useful genetic pools. Hence, I will review the applications of long-read RNA isoform sequencing, including the relative merits of the technology, the improvement of the accuracy in gene prediction and gene annotation, and the full-length unigene builds in a new genome; the limitations of the technology will be also discussed. The review will be valuable in collecting data resources for functional genomic studies.

Keywords: functional genomics, gene prediction, long-read RNA sequencing, transcriptome

1. Introduction

Transcriptomics is the study of transcript catalogs in a cell, tissue, or organism for a given developmental stage or physiological condition [1]. The transcriptome indicates the complete set of transcripts that consists of protein-coding messenger RNA (mRNA) and non-coding RNA (ncRNA), including ribosomal RNA (rRNA), transfer RNA (tRNA), and other ncRNAs [2, 3]. In contrast with the relatively stable genome, various factors such as developmental stage, physiological condition, and external environment influence the changes in the transcriptome. The goals of transcriptomics include the annotation of the transcriptome, and the determination of the functional structure of each gene in the genome and the changes in the expression levels of each gene among different transcriptome samples [1, 4, 5].

Transcriptome analysis depends heavily on the availability of high-throughput tools on account of the complexity of the transcriptome. Thus, RNA sequencing (RNA-Seq) has become an important tool for biological studies. RNA-Seq can quantify gene expression spatially and temporally. Although RNA-Seq has enabled the generation of massive amounts of sequence data due to their high-throughput characteristic, their application of short reads makes them poorly suited for genome and transcriptome assembly, and isoform detection. Single-molecule real-time (SMRT) sequencing, a new method to generate long-read sequences developed by

PacBio platform, provides an alternative approach to overcome these limitations in sequence length and accelerate improving our understanding of the complexity of the transcripts [6].

In general, the read length of Illumina HiSeq platform is about 100–150 bp, which is relatively short compared to that of PacBio platform (around 10 kb). However, Illumina HiSeq platform has the advantage of generating more accurate reads and high-throughput data. On the other hand, even though its accuracy is lower than that of Illumina HiSeq platform, single-molecule real-time (SMRT) sequencing of PacBio platform, a new method of sequence analysis, was developed and applied to elucidate the genomic structures of difficult to sequence organisms [7] because of its long-reads, which results in the improvement of assembly, gene prediction, and annotation. Using this technique, sequences are analyzed from a single strand of DNA without genomic amplification [9]. PCR-free long-read sequencing enables to help to carry out large complex whole-genomes (i.e., hexaploid wheat and maize).

PacBio sequencing captures sequences during the replication process of the target DNA in real-time. The template, also called a SMRTbell, contains a target double-stranded DNA (dsDNA) ligated with hairpin adaptors at both ends, resulting in a closed and single-stranded circular DNA [8]. When the SMRTbell is loaded into a chip called a SMRT cell, diffusion of the SMRTbell into a sequencing unit called a zero-mode wave guide (ZMW) is carried out [10]. In each ZMW, a single polymerase immobilized at the bottom can bind to adaptors of the SMRTbell [11]. Each of the four nucleotides is fluorescent-labeled. As a nucleotide associates with the template in the active site of the polymerase, a light pulse is produced for base detection. A single polymerase read can be generated up to 40 kb, depending on the library size and sequencing time. The closed-circle form of the SMRTbell can make the reaction repeat until the reaction is terminated after the replication of one strand of the target dsDNA or double-stranded complementary DNA by the polymerase. However, the mean length of full transcripts is 1–3 kb in most plant and animal genomes (e.g., 1.6 kb in *Arabidopsis* [12], 1.8 kb in rice [13], 2.3 kb in human [14], and 1.2 kb in mouse [15]); thus, the same transcript can be covered multiple times by the long polymerase read. In this scenario, a few reads (called subreads) can be generated from the polymerase read by trimming adaptor sequences. The consensus sequence of multiple subreads in a single ZMW generates a read of insert (ROI) or a circular consensus sequence (CCS) read with higher accuracy. Hence, a protocol of isoform sequencing (Iso-Seq) for long-read transcriptome sequencing that includes library construction, size selection, sequencing, and data processing was developed by PacBio. Iso-Seq allows the direct sequencing of transcripts up to 10 kb, which is particularly useful for the genomes of uncharacterized species.

However, even though PacBio sequencing has an advantage in terms of read length over next-generation sequencing, the throughput of PacBio sequencing is relatively low. A single SMRT cell contains 150,000 ZMWs, each of which can produce one polymerase read with a mean length of 1o kb. Typically, only 35,000–70,000 reads of the 150,000 ZMW wells on a SMRT cell can be produced successfully because of the failure of anchoring a polymerase and loading more than one DNA molecule in a ZMW. Consequently, the typical throughput of the PacBio RS II system is around 0.5–1 Gb per SMRT cell [16]. Recently, PacBio developed another system called Sequel that produces over seven times the reads, with 1,000,000 ZMWs, and yields around 3.5–7 Gb per SMRT cell [17]. Sequel is appropriate for projects such as *de novo* genome assembly and isoform sequencing of transcriptomes. Another notable problem of PacBio sequencing is the relatively high error rate (around 11–15%) of polymerase reads [18]. Many hybrid sequencing approaches have been attempted to develop a method that has the accuracy of short reads but with the length of PacBio reads [19].

Long-read transcriptome sequencing generates longer and improved transcripts with a high level of assembly completeness and gene annotation. Moreover, it prevents obtaining artifacts such as chimeras, structural errors, incomplete assembly, and base errors [20].

Here, we review the sample preparation, library construction, analytical pipelines, and the result of isoform sequencing (Iso-Seq), as a long-read transcriptome sequencing, in gene prediction and annotation. Furthermore, we will also discuss the relative merits and the limitations of the Iso-Seq technology.

2. Merits of long-read transcriptome sequencing

Long-read transcriptome sequencing such as Iso-Seq generates longer and improved transcripts from a species with a high level of assembly completeness and gene annotation, enabling a comprehensive view of the transcriptome. Conventional methods, such as cDNA cloning and EST sequencing, have limitations with relatively low data coverage. Although deep short-read sequencing (i.e., RNA-Seq) provides good sequencing depth and coverage for genome-wide transcriptome analysis, their short-read length generates assembly incompleteness of transcripts, resulting in high error rate in assembly and unreliable gene annotation. Long-read transcriptome sequencing can also provide experimental verification of predicted gene models in a genome, enable the quality of gene structures predicted and also give the potential to reduce missing gene annotation. For example, missing gene annotation may lead to false interpretation such as gene loss and errors in gene expression profiles that map and quantify RNA-seq reads using predicted gene models. Thus, this technology can be helpful to find full-length (FL) transcripts harboring complete open reading frames (ORFs) and uncover novel splice isoforms as well as novel genes. This can result in the improvement of accuracy of gene prediction with an experimental verification and annotations for aiding in studying gene regulation.

3. Sample preparation and library construction for isoform sequencing

Iso-Seq with the PacBio platform can generate FL cDNA sequences including the 5′ and 3′-UTRs (untranslated regions), as well as the polyA tails of the transcripts. The whole workflow including the experimental protocol and analytical pipelines is illuminated in **Figure 1** [10].

3.1 Isolation of total RNA

The samples can be collected from various tissues (i.e., blood, gill, skin, muscle, liver, spleen, intestine, ovary, testis, kidney, heart, and brain of an animal) [21], or from certain developmental stages (developing rabbit at 21, 49, and 84 days of age) [22]. The high quality of RNA with enough purity and integrity is critical to reduce the amplification cycles required in large-scale PCR and improve the sequencing diversity. RNA extraction is usually done through an easy-spin RNA extraction kit, or RNAiso Pure RNA Isolation kit [20–22]. In general, 2–5 μg of total RNA with an RNA integrity number (RIN) greater than 7 is required.

3.2 cDNA synthesis and size partitioning

Isolation of polyA mRNA is required for analyzing the transcripts of protein-coding genes. The Iso-Seq method is flexible and allows different types of RNA

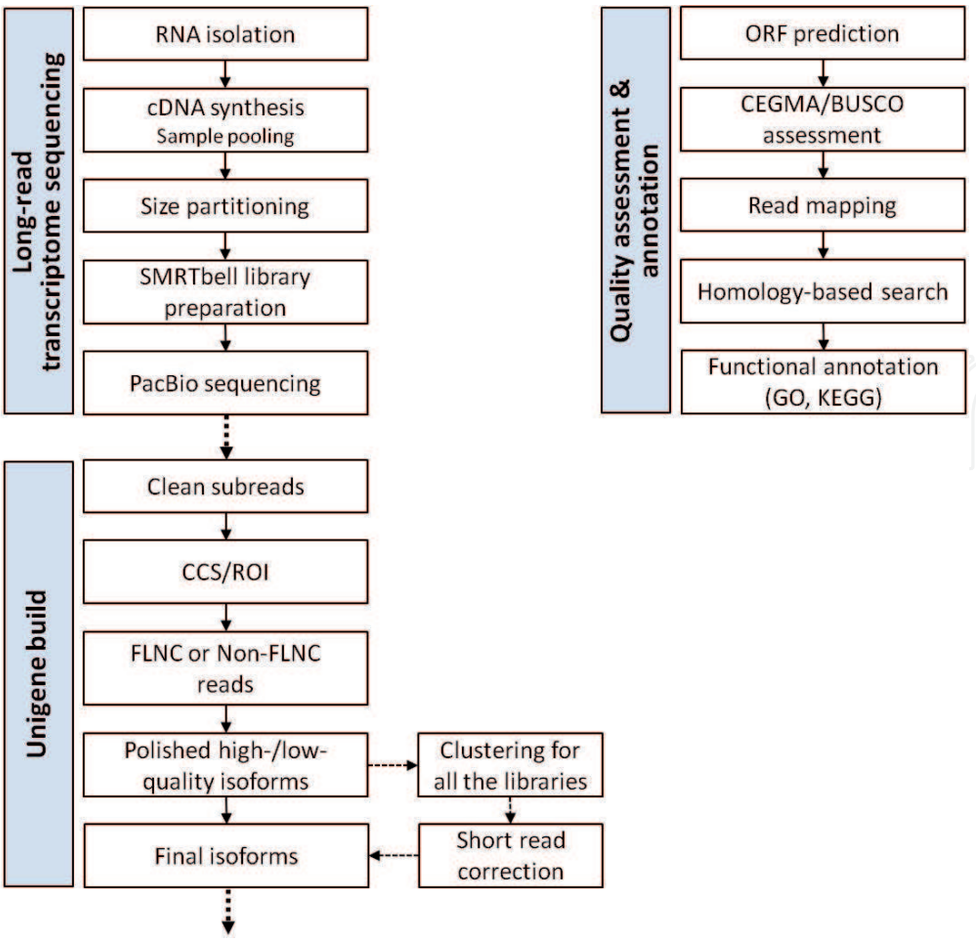

Figure 1.
Schematic workflow of isoform sequencing.

to be sequenced. Alternatively, mRNAs can be selected by polyA enrichment. The first-strand cDNA is amplified with oligo(dT) to enrich RNAs with a polyA tail, including mRNAs and long noncoding RNAs (lncRNAs) for further analysis.

For parallel analysis of RNA samples derived from various tissues, barcode for each sample with unique sequences is alternatively used. For instance, multiplex sequencing was performed to construct a maize transcriptome library from various tissues [23]. However, barcoding samples is not always desired because sequencing efficiency may be reduced by the barcode sequence.

3.3 Size partitioning

Size selection for size partitioning, which is the most commonly used method to avoid over-representation of smaller transcripts in sequencing data, allows for more even representation of cDNA of different size ranges, since smaller fragments may load preferentially on the sequencer. Furthermore, the process of second fractionation is recommended to remove any smaller fractions from the first size selection. To enhance PCR amplification, different sizes of the cDNA libraries including <1, 1–2, 2–3, and 3–6 kb are generally constructed to maximally recover transcript diversity and sequence. However, such size selection may bring about missing small size transcripts less than approximately 1 kb. This problem appears to result from technical limitation by size selection in the construction of mRNA sequencing libraries. This can get solved by combinatorial use with short-read RNA-Seq data that are very effective for transcriptome coverage, especially small size of transcripts.

3.4 Library preparation and sequencing

Double-stranded cDNA is not enough for SMRTbell library construction following size selection. PacBio suggests PCR amplification using the KAPA HiFi Enzyme [24] with about 10 cycles. Then, a circularized molecule called a SMRTbell template is transformed from the amplified cDNAs by the SMRTbell Template Prep kit. After the step is completed, the library is ready to be loaded into a SMRT cell and subjected to sequencing on the PacBio platform. There is a compromise between SMRT cell numbers and the sequencing cost. In general, the Iso-Seq protocol recommends 8–50 SMRT cells to retrieve diversity in a tissue.

4. Building full-length transcripts in a genome

Error correction of the raw reads is necessary to improve the assembly quality of the FL transcripts. PacBio provides the Iso-Seq analysis software to perform the procedure by iterative clustering for error correction (ICE) and the Quiver algorithm (https://www.pacb.com/applications/rna-sequencing). Then, various analysis approaches can be applied to overcome the limitation of Iso-Seq, improve assembly quality, and evaluate the quality assessment of the unigenes.

The Iso-Seq raw reads are usually called polymerase reads or continuous long reads (CLRs) and have an average length of 10 kb (**Figure 1**). Considering the average length of a transcript is 1–2 kb, the same copies of the inserts are contained in a single polymerase that could be split into several subreads by removing the adaptor sequences by PacBio SMRT link analysis [20]. The circular consensus sequences or ROIs are generated from several subreads. The full-length non-chimeric read (FLNC) is defined not only when the polyA tail signal preceding the 30-primer is present, but also when both 50- and 30-cDNA primers are present. To enhance consensus accuracy and remove the redundancy of FLNC without any additional sequence data, ICE and Quiver can be applied [20]. The Iso-Seq classify tool is used for classifying the ROIs into full-length nonchimeric and non-full-length reads by identifying the 50 and 30 adapters used in library preparation. Then, the Iso-Seq cluster tool is used for clustering all the full-length reads, and the consensus sequences produced by the cluster tool are polished using the non-full-length reads through the Quiver algorithm [25]. Additionally, the CD-HIT program [26] is likely to be helpful to cluster the high and low quiver consensus isoforms from ROIs with high sequence identity threshold (i.e. 0.98–0.99) [20, 21].

Iso-Seq reads present a disadvantage with the high frequency of errors of nucleotide indels and mismatches. Thus, the procedure of correcting InDels and mismatches is performed via alignment with reference genomes [27]. To overcome this, a viable alternative approach is to integrate short reads with long reads via hybrid sequencing. For instance, RNA samples prepared from the same samples are sequenced by both PacBio and Illumina HiSeq. The short reads from the Illumina HiSeq are applied to correct the transcript isoforms using LoRDEC tool v0.6 [28]. Then, the corrected isoform sequences are aligned against a reference genome by GMAP aligner [29]. The following analyses are recommended to exclude the sequences with multiple and chimeric alignments. To assess quality of the unigenes, some software such as CEGMA [30] and BUSCO [31] can be applied [20, 21, 32, 33]. The percentages of the transcripts that fully and partially aligned to the conserved proteins are calculated.

FL or longer transcriptome data have been mostly published from large complex or uncharacterized genomes of plant species (**Table 1**). Although deep short-read transcriptome sequencing (i.e., RNA-Seq) have accumulated over recent year, they are likely to generate low-quality transcripts with a small portion of FL transcripts, prohibiting accurate transcript reconstruction and leading incorrect annotation.

Species	No. of transcripts	Mean length (bp)	Discovery					Reference
			Identification of novel gene isoforms	Isoform annotation	Alternative splicing events	Gene prediction	Other	
Panax ginseng	135,317	3178	Y	Y	Y	—	—	[20]
Triticum aestivum	91,881	2388	Y	Y	—	—	—	[45]
Zea mays B73	111,151	3372	Y	Y	Y	Y	Fusion transcripts	[23]
Sorghum bicolor	27,860	1042 (full-length ROI)	Y	Y	Y	Y	—	[27]
Trifolium pratense	206,465	2789	Y	Y	Y	—	—	[34]
Zea mays W64A	166,693	2715	Y	Y	Y	—	Fusion transcripts	[35]
Allium sativum	36,321	1500	Y	Y	—	—	Association study	[36]
Populus (*P. deltoides* × *P. euramericana* cv. 'Nanlin895')	87,150	2417	Y	Y	Y	—	Fusion transcripts	[37]
Coffea arabica	95,995	3236	Y	Y	Y	Y	—	[38]

Table 1.
Transcriptomics studies in plants by isoform sequencing.

vUnlike RNA-Seq data, Iso-Seq data, which are derived from various tissues as many as possible, harbor a large portion of unique FL transcripts. For example, Wang et al. [23] reported that maize yielded 111,151 non-redundant FL transcript isoforms, corresponding to approximately 26,946 genes. In addition, genome coverage of Iso-Seq data is achieved near-saturation. Ultimately, cost-effective long-read transcriptome sequencing can be the gold standard for transcript completeness, characterization of transcriptome, and draft genome annotation. To identify trait-associated transcripts in species for which a reference genome is lacking (i.e., garlic), this approach was used as a reference sequence for scoring the variation in both SNP and expression level in the population [36], reporting the characterization of transcripts (lncRNAs) associated with garlic clove shape traits.

5. Improvement of the efficiency of functional gene prediction and annotation

Completeness of assembled transcripts is closely related to the efficiency of functional gene prediction or annotation, especially in the absence of reference genome information. Because of such advantage, Iso-Seq has been applied in a variety of species [20–22, 32, 33]. In addition, optimized training and prediction settings on the basis of short- and long-read transcriptome data in gene prediction results in increased their sensitivity and precision [39]. In particular, the method is helpful for obtaining comprehensive gene sets for newly sequenced genomes of non-model eukaryotes [39].

To identify the protein coding potential of transcripts, Transdecoder (https://transdecoder.github.io) is generally applied [20, 21, 32, 40]. For example, even though the number of transcripts using Iso-Seq is much smaller than those *de novo* assembled in previous RNA-seq studies, the transcripts from Iso-Seq show high efficiency in recovering full-length transcripts. ESTScan [41], in addition to Transdecoder, is used to predict coding DNA sequences (CDSs) unless isoforms are annotated in the databases. For example, in the study of *Halogeton glomeratus* [42], the CDS prediction ratio of transcripts using Iso-Seq (95.09%) is much higher than that of transcripts using Illumina RNA-Seq data (66.86%).

For functional annotation, isoform sequences are used as queries for sequence homology searches in Blast, Blast2GO [43], and InterProScan5 [44] to identify functional annotation terms from the nonredundant protein (NR), non-redundant nucleotide (NT), Gene Ontology (GO), Clusters of Orthologous Groups (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), SwissProt, and Interpro databases. For example, when the RNA-Seq data of *H. glomeratus* were re-annotated with Iso-Seq transcriptome data, the length distribution, functional annotation, and coding sequence quantity of the Iso-Seq transcripts were significantly improved [42]. In particular, with respect to the species distribution of the annotation from the NR database, 98.31% of the annotated isoforms showed the highest similarity to sequences from the three most prevalent species. In addition, Illumina RNA-Seq data were highly mapped to the Iso-Seq transcripts (unigenes). This suggests that long-read, full-length or partial-unigene data with high-quality assemblies are invaluable resources as transcriptomic references in a genome and can be used for comparative analyses in closely related medicinal plants.

6. Conclusion

Transcriptome data generated by Iso-Seq generate longer and improved unigenes with a high level of assembly completeness and gene annotation, enabling a

comprehensive view of the transcriptome. In particular, compared with conventional methods, long-read transcriptome sequencing seems to improve misassembly rate and unreliable gene annotation, thus enabling to elucidate the function of genes associated with traits of interest as well as novel transcripts. A hybrid approach that combines isoform sequencing with full-length transcripts and RNA-Seq capable of fixing sequence error and quantifying gene expression is the optimal solution to study transcriptomes for improving completeness of transcripts, data coverage, and gene annotation.

Acknowledgements

This work was supported by grants from the National Agricultural Genome Center (project No. PJ01349002), Rural Development Administration, Republic of Korea.

Conflict of interest

The author declares no conflict of interest to disclose.

Author details

Dong Jin Lee and Chang Pyo Hong*
Theragen Etex Bio Institute, Suwon, Republic of Korea

*Address all correspondence to: changpyo.hong@theragenetex.com

References

[1] Tang F et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods. 2009;**6**:377-382. DOI: 10.1038/nmeth.1315

[2] Lindberg J, Lundeberg J. The plasticity of the mammalian transcriptome. Genomics. 2010;**95**:1-6. DOI: 10.1016/j.ygeno.2009.08.010

[3] Okazaki Y et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002;**420**:563-573. DOI: 10.1038/nature01266

[4] Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. Journal of Biomedicine & Biotechnology. 2010:19. DOI: 10.1155/2010/853916 Article ID 853916

[5] Ruan Y, Le Ber P, Ng HH, Liu ET. Interrogating the transcriptome. Trends in Biotechnology. 2004;**22**(1):23-30. DOI: 10.1016/j.tibtech.2003.11.002

[6] Rhoads A, Au KF. PacBio sequencing and its applications. Genomics, Proteomics & Bioinformatics. 2015;**13**:278-289. DOI: 10.1016/j.gpb.2015.08.002

[7] Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nature Biotechnology. 2013;**31**:1009-1014

[8] Travers KJ et al. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Research. 2010;**38**(15):e159. DOI: 10.1093/nar/gkq543

[9] Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. Genome Biology. 2013;**14**:405

[10] Gonzalez-Garay ML. Introduction to isoform sequencing using Pacific Biosciences technology (Iso-Seq). Vol. 9. Dordrecht, The Netherlands: Springer; 2015. pp. 141-160

[11] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;**323**:133-138. DOI: 10.1126/science.1162986

[12] Swarbreck D et al. The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. Nucleic Acids Research. 2008;**36**(Database issue):D1009-D1014. DOI: 10.1093/nar/gkm965

[13] Ouyang S et al. The TIGR Rice Genome Annotation Resource: Improvements and new features. Nucleic Acids Research. 2007;**35**(Database issue):D883-D887. DOI: 10.1093/nar/gkl976

[14] Ota T et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nature Genetics. 2004;**36**(1):40-45. DOI: 10.1038/ng1285

[15] Kawai J et al. Functional annotation of a full-length mouse cDNA collection. Nature. 2001;**409**(6821):685-690. DOI: 10.1038/35055500

[16] PacBio RS II System. Available online: http://dnatech.genomecenter.ucdavis.edu/pacbio-library-prepsequencing [Accessed: 1 November 2017]

[17] PacBio Sequel System. Available online: http://www.pacb.com/products-and-services/pacbio-systems/sequel [Accessed: 12 July 2017]

[18] Korlach J. Understanding accuracy in SMRT® Sequencing. Available online: https://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing.pdf

[19] Koren S et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology. 2012;**30**:693-700. DOI: 10.1038/nbt.2280

[20] Jo IH, Lee J, Hong CE, Lee DJ, et al. Isoform sequencing provides a more comprehensive view of the *Panax ginseng* transcriptome. Genes. 2017;**8**:228. DOI: 10.3390/genes8090228

[21] Yi S, Zhou X, Li J, Zhang M, Luo S. Full-length transcriptome of *Misgurnus anguillicaudatus* provides insights into evolution of genus *Misgurnus*. Scientific Reports. 2018;**8**(1):11699. DOI: 10.1038/s41598-018-29991-6

[22] Chen SY, Deng F, Jia X, Li C, Lai SJ. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Scientific Reports. 2017;**7**(1):7648. DOI: 10.1038/s41598-017-08138-z

[23] Wang B et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature Communications. 2016;**7**(11708). DOI: 10.1038/ncomms11708

[24] PacBio SMRTbell library construction. Available online: http://www.pacb.com/products-and-services/analytical-software/devnet [Accessed: 10 May 2017]

[25] Gordon SP et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS One. 2015;**10**(7):e0132628. DOI: 10.1371/journal.pone.0132628

[26] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: A web server for clustering and comparing biological sequences. Bioinformatics. 2010;**26**(5):680-682. DOI: 10.1093/bioinformatics/btq003

[27] Abdel-Ghany SE et al. A survey of the sorghum transcriptome using single-molecule long reads. Nature Communications. 2016;**7**:11706. DOI: 10.1038/ncomms11706

[28] Salmela L, Rivals E. LoRDEC: Accurate and efficient long read error correction. Bioinformatics. 2014;**30**(24):3506-3514. DOI: 10.1093/bioinformatics/btu538

[29] Wu TD, Watanabe CK. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;**21**(9):1859-1875. DOI: 10.1093/bioinformatics/bti310

[30] Parra G, Bradnam K, Korf I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;**23**(9):1061-1067. DOI: 10.1093/bioinformatics/btm071

[31] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;**31**(19):3210-3212. DOI: 10.1093/bioinformatics/btv351

[32] Zeng D et al. Single-molecule long-read sequencing facilitates shrimp transcriptome research. Scientific Reports. 2018;**8**(1):16920. DOI: 10.1038/s41598-018-35066-3

[33] Pootakham W et al. Development of a novel reference transcriptome for scleractinian coral *Porites lutea* using single-molecule long-read isoform sequencing (Iso-Seq). Frontiers in Marine Science. 2018;**5**(122). DOI: 10.3389/fmars.2018.00122

[34] Chao Y, Yuan J, Li S, Jia S, Han L, Xu L. Analysis of transcripts and splice isoforms in red clover (*Trifolium pratense* L.) by single-molecule long-read sequencing. BMC Plant Biology. 2018;**18**(1):300. DOI: 10.1186/s12870-018-1534-8

[35] Zhou Y, Zhao Z, Zhang Z, Fu M, Wu Y, Wang W. Isoform sequencing provides insight into natural genetic diversity in maize. Plant Biotechnology Journal [Epub ahead of print. 2018. DOI: 10.1111/pbi.13063

[36] Chen X, Liu X, Zhu S, Tang S, Mei S, Chen J, et al. Transcriptome-referenced association study of clove shape traits in garlic. DNA Research. 2018;**25**(6):587-596. DOI: 10.1093/dnares/dsy027

[37] Chao Q, Gao ZF, Zhang D, Zhao BG, Dong FQ, Fu CX, et al. The developmental dynamics of the Populus stem transcriptome. Plant Biotechnology Journal. 2019;**17**(1): 206-219. DOI: 10.1111/pbi.12958

[38] Cheng B, Furtado A, Henry RJ. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. Gigascience. 2017;**6**(11):1-13. DOI: 10.1093/gigascience/gix086

[39] Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biology. 2015;**16**(184). DOI: 10.1186/s13059-015-0729-7

[40] Haas BJ, Papanicolaou A, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;**8**(8):1494-1512. DOI: 10.1038/nprot.2013.084

[41] Iseli C, Jongeneel CV, Bucher P. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proceedings. International Conference on Intelligent Systems for Molecular Biology. 1999:138-148

[42] Wang J et al. Single-molecule long-read transcriptome dataset of halophyte *Halogeton glomeratus*. Frontiers in Genetics. 2017;**8**(197). DOI: 10.3389/fgene.2017.00197

[43] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;**21**(18):3674-3676. DOI: 10.1093/bioinformatics/bti610

[44] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: Protein domains identifier. Nucleic Acids Research. 2005;**33**(Web Server issue):W116-W120. DOI: 10.1093/nar/gki4

[45] Dong L et al. Single-molecule realtime transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. BMC Genomics. 2015;**16**(1039). DOI: 10.1186/s12864-015-2257-y