Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx830
Advance Access Publication Date: 21 December 2017
Applications Note

OXFORD

Sequence analysis

PRAPI: post-transcriptional regulation analysis pipeline for Iso-Seq

Yubang Gao[1], Huiyuan Wang[1], Hangxiao Zhang[1], Yongsheng Wang[1], Jinfeng Chen[2] and Lianfeng Gu[1,*]

[1]Basic Forestry and Proteomics Research Center, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, College of Life Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China and [2]Department of Plant Pathology and Microbiology, Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on September 4, 2017; revised on October 24, 2017; editorial decision on December 16, 2017; accepted on December 20, 2017

Abstract

Summary: The single-molecule real-time (SMRT) isoform sequencing (Iso-Seq) based on Pacific Bioscience (PacBio) platform has received increasing attention for its ability to explore full-length isoforms. Thus, comprehensive tools for Iso-Seq bioinformatics analysis are extremely useful. Here, we present a one-stop solution for Iso-Seq analysis, called PRAPI to analyze alternative transcription initiation (ATI), alternative splicing (AS), alternative cleavage and polyadenylation (APA), natural antisense transcripts (NAT), and circular RNAs (circRNAs) comprehensively. PRAPI is capable of combining Iso-Seq full-length isoforms with short read data, such as RNA-Seq or polyadenylation site sequencing (PAS-seq) for differential expression analysis of NAT, AS, APA and circRNAs. Furthermore, PRAPI can annotate new genes and correct mis-annotated genes when gene annotation is available. Finally, PRAPI generates high-quality vector graphics to visualize and highlight the Iso-Seq results.

Availability and implementation: The Dockerfile of PRAPI is available at http://www.bioinfor.org/tool/PRAPI.

Contact: lfgu@fafu.edu.cn

1 Introduction

PacBio platform offers a reliable way to identify alternative splicing (AS) and alternative cleavage and polyadenylation (APA) from Iso-Seq (Abdel-Ghany *et al.*, 2016; Wang *et al.*, 2017). To the best of our knowledge, TAPIS (Abdel-Ghany *et al.*, 2016), IDP (Au *et al.*, 2013), and PacBio's SMRT-Analysis are excellent tools for analyzing isoforms from Iso-Seq reads. TAPIS also performs APA analysis since full-length isoforms of Iso-Seq span from the transcription start sites to terminal sites (Abdel-Ghany *et al.*, 2016). At present, transcriptome profiling with next-generation sequencing (NGS) data such as RNA-Seq and polyadenylation site sequencing (PAS-Seq), is an indispensable method for performing quantitative analysis of gene or isoform expression (Wang *et al.*, 2017). However, previous methods lack the module to combine high-accuracy short reads with Iso-Seq long reads. Consequently, a full pipeline consisting of Iso-Seq data process, comprehensive analysis, and visualization, will be necessary when the cost of Iso-Seq is reduced. To meet the requirement of Iso-Seq analysis, we have developed a pipeline called Post-transcriptional Regulation Analysis Pipeline for Iso-Seq (PRAPI), which aims to identify and quantify the post-transcriptional regulation. PRAPI also adds new features to allow users to identify differentially expressed NAT, AS, APA and circRNAs by further combining them with short reads sequencing. Furthermore, the results from PRAPI are highlighted using vector diagrams. In summary, PRAPI is a comprehensive, user-friendly software that facilitates Iso-Seq analysis.

2 Implementation

The PRAPI pipeline is implemented mainly in Python 2.7.13 and can be run on Linux operating systems with multiple threads. The visualization of Iso-Seq is implemented using pycairo (https://cairographics.org/pycairo/). Visualization is generated in vector images including several popular tracks such as Iso-Seq alignment, coordinate regions for AS, APA, ATI, circRNAs and wiggle graphs to represent peak density from NGS. The overall framework of PRAPI is shown in Figure 1. We provide a Docker image that includes the prerequisite tools. Therefore, the pipeline does not require the installation of any third-party dependencies such as the GMAP (Wu and Watanabe, 2005) and SAMtools (Li *et al.*, 2009). PRAPI requires at least two FASTA-formatted input files: reference genome and the first round corrected full-length reads, which can be generated from SMRT-Analysis and LSC (Au *et al.*, 2012). Full-length cDNA isoforms from Oxford Nanopore's third-generation single-molecule sequencing platform are also accepted. Subsequently, PRAPI maps error correcting PacBio reads to reference genome using GMAP (Wu and Watanabe, 2005). All aligned reads are corrected for the second round using reference genome to further remove potential error in PacBio reads by TAPIS module (Abdel-Ghany *et al.*, 2016).

Due to the incomplete genome sequences, a few full-length reads cannot be entirely mapped to genomes. Moreover, both transcript start sites and terminal sites of Iso-Seq reads possessing an end-to-end read alignment with no >5 bp trimmed/clipped region are kept for the identification of ATI and APA with the default option of PRAPI to discard trimmed/clipped alignment. Micro-heterogeneity sites are clustered with 25 nt window and the default number of supporting reads is 2 and 1 for ATI and APA, respectively. If PAS-seq libraries (Wang *et al.*, 2017) are provided by users, the summit peaks from PAS-Seq will be called and expression level for the peaks matched with the poly(A) sites from Iso-Seq will be calculated to identify differential APA sites using Fisher's exact test.

PRAPI identifies four major AS types, including intron retention, alternative 5′ donor, alternative 3′ donor and exon skipping events, according to the exon/intron coordinate characterized by using long-read GMAP alignment or isoforms detection and prediction from IDP (Au *et al.*, 2013). PRAPI uses the statistical model from rMATS (Shen *et al.*, 2014) to identify differential AS events based on RNA-Seq data.

PRAPI identifies NAT pairs that have overlapping regions between two oppositely oriented Iso-Seq. The resulting NAT pairs are mainly categorized into three types: head-to-head, tail-to-tail, and fully overlapping. Read counts from NAT pairs are calculated using user-provided strand-specific RNA-Seq libraries to identify differential NATs. The *P* value and FDR are calculated using edgeR (Robinson *et al.*, 2010).

Mis-annotation of gene is identified when Iso-Seq reads span two or more different genes. Novel genes are defined as Iso-Seq clusters that have no corresponding gene annotation. The overlapping intervals in exons of all the Iso-Seq reads are merged into unique coordinates to annotate these novel loci. The identification of circRNAs is highly dependent on the annotation of splice junction (Gao *et al.*, 2015). PacBio Iso-Seq provides a significant advantage in obtaining annotation of reliable isoform splice junctions. PRAPI generates unique splice junctions according to isoforms from Iso-Seq and uses it as the input file for CIRIexplorer (Gao *et al.*, 2015) to identify circRNAs. Read counts for each circRNAs are calculated using user-provided RNase R treatment libraries or non-poly(A)-selected RNA-Seq to identify differential circRNAs. The *P* values and FDR are calculated using DEGseq (Wang *et al.*, 2010).

Time consumed for PRAPI was evaluated using the data from a previous study (Wang *et al.*, 2017), which includes 146 225 error-corrected full-length non-chimeric reads. The time required for processing the whole dataset is 98 min using 18.0 GB Memory on our server (Linux: Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60 GHz; 20 Cores; Memory 256GB).

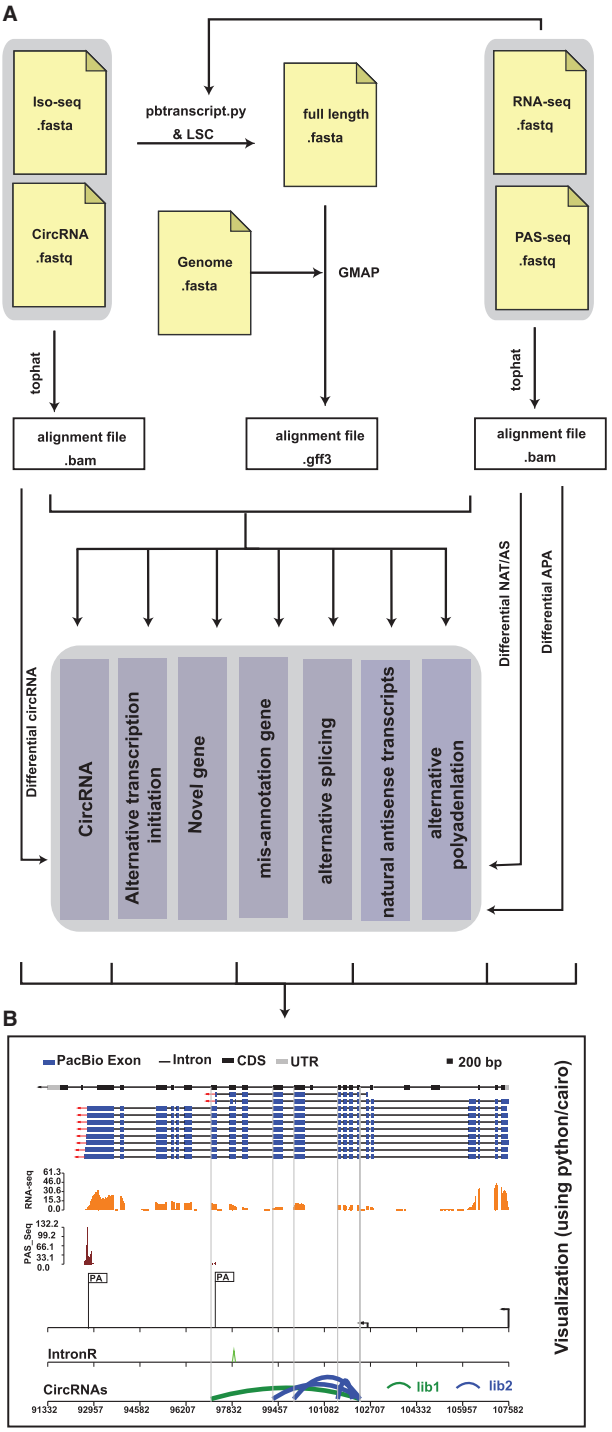

Fig. 1. The overall design and visualization of Iso-Seq

3 Conclusion

In summary, we present PRAPI, a full-function pipeline aimed at not only interpreting post-transcriptional regulation based on PacBio'

full-length reads but also presenting the visualization of the high throughput data. Descriptions of all the input, output, and parameters can be found in our online tutorials and test dataset. The source code with complete tutorial is publicly available.

Acknowledgement

The authors thank Professor Anireddy S.N. Reddy and Dr. Michael Hamilton for advice on TAPIS.

Funding

This work was supported by the National Key Research and Development Program of China (2016YFD0600106), the National Natural Science Foundation of China Grant (31570674).

Conflict of Interest: none declared.

References

Abdel-Ghany,S.E. *et al*. (2016) A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun*., **7**, 11706.

Au,K.F. *et al*. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Nat. Acad. Sci. USA*., **110**, E4821–E4830.

Au,K.F. *et al*. (2012) Improving PacBio long read accuracy by short read alignment. *PLoS One*, **7**, e46679.

Gao,Y. *et al*. (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol*., **16**, 4.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Shen,S. *et al*. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad. Sci. USA*., **111**, E5593–E5601.

Wang,L. *et al*. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.

Wang,T. *et al*. (2017) Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). *Plant J*., **91**, 684–699.

Wu,T.D., and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.