bio-101

An Analysis Pipeline for Identification of RNA Modification, Alternative Splicing and Polyadenylation Using Third Generation Sequencing

Yuxiang Liufu[1, #], Xuqing Liu[1, #], Lin Wu[1], Hangxiao Zhang[2], Yubang Gao[2, *] and Lianfeng Gu[2, *]

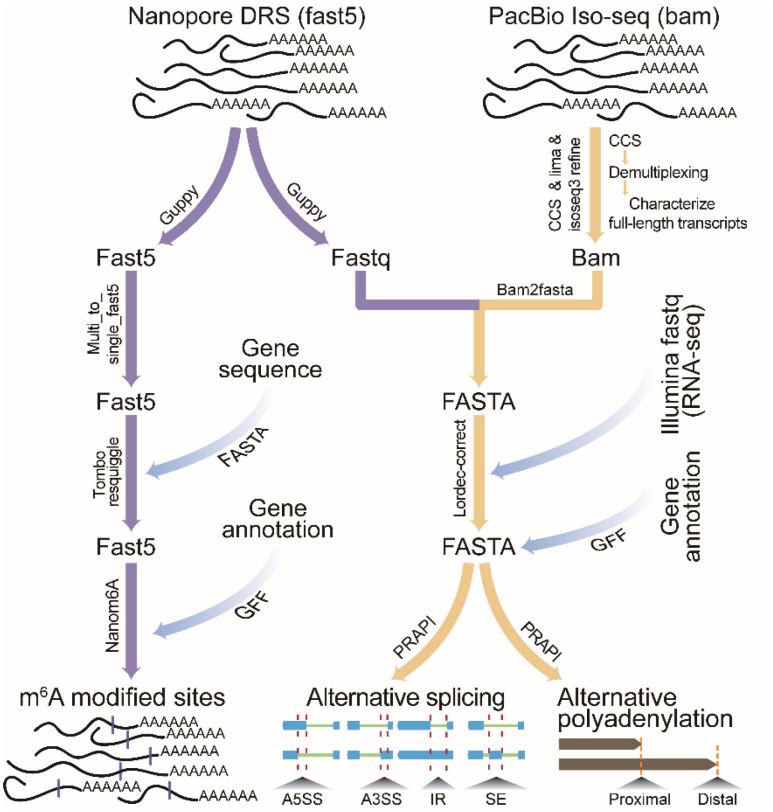
[1]College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[2]Basic Forestry and Proteomics Research Center, College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, China
*For correspondence: yubanggaofafu@gmail.com; lfgu@fafu.edu.cn
[#]Contributed equally to this work

Abstract

Nanopore sequencing based on Oxford Nanopore Technologies (ONT) and Pacific BioSciences (PacBio) single-molecule real-time (SMRT) long-read isoform sequencing (Iso-Seq) have shown great potential in detecting post-transcriptional regulation. Direct RNA sequencing (DRS) has the advantages in capturing RNA modification due to without PCR amplification which is the limitation of the next-generation sequencing (NGS). Here, we provide a comprehensive computational procedure for the quantification of RNA modification in single-base resolution based on DRS data. Moreover, we also provide procedure on the identification of alternative splicing (AS) and alternative polyadenylation (APA) based on both DRS and PacBio Iso-Seq data. The entire step was based on two packages (Nanom6A and PRAPI), which were based on Python language on Linux system.

Keywords: Nanopore direct RNA sequencing, PacBio Iso-Seq, RNA modification, N^6-methyladenosine, Alternative splicing, Alternative polyadenylation

Graphical abstract:



The flowchart for identification of RNA modification, alternative splicing and alternative polyadenylation based on third generation sequencing technology.

Background

Transcriptome provides an in-depth perspective for understanding the regulation in gene expression and growth development. With the rapid development of the third-generation sequencing technology, the data based on long-read sequencing presented great advantage in revealing post-transcriptional regulation (Smith *et al*., 2019). Post-transcription regulation has been known as a complex biological process, playing an important part in influencing the splicing, export, localization and translation of RNA (Keene, 2007). For example, alternative polyadenylation (APA) plays a vital role in maintaining the structure and stability of RNA (Di Giammartino *et al*., 2011, Tian and Manley, 2017). Alternative splicing (AS) is implicated in the regulation of rhizome-associated development (Wang *et al*., 2017) and hormone response (Zhang *et al*., 2018). A variety of AS events could be found in key genes related to biosynthesis pathways (Chen *et al*., 2020). PacBio Iso-Seq can capture the full-length mRNA in isoform-wide to analyze the post-transcriptional processing events such as AS and APA (Rhoads and Au, 2015). Accuracy and reliability tools are important to identify post-transcriptional events. AS based on short-read platform can be identified by several excellent software, such as rMATS (Shen *et al*., 2014) and SUPPA2 (Trincado *et al*., 2018). However short-read platforms show disadvantage in detecting full-length isoforms due to the limitation of transcriptome assembly (Rhoads and Au, 2015). TAPIS (Abdel-Ghany *et al*., 2016) and FLAIR (Tang *et al*., 2020) have been developed to identify AS events based on long-read sequencing. However, TAPIS does not provide quantitative analysis of post-transcriptional events (Abdel-Ghany *et al*., 2016). FLAIR can detect isoforms, but lacks APA detecting module (Tang *et al*., 2020). Hence, we provide a comprehensive and user-friendly pipeline (PRAPI) for the identification and visualization of AS and APA (Gao *et al*., 2018). Here we provide detail procedure for data processing, visualization, identification of AS and APA using PRAPI based on Iso-Seq isoforms and short read data.

In addition to PacBio platform, Oxford Nanopore Technologies (ONT) platform applies a unique measurement method to keep the information of RNA molecular passing through a special pore as an electrical signal (Garalde *et al*., 2018), which can identify RNA modifications including N^6-methyladenosine (m^6A) (Linder *et al*., 2015; Zhang *et al*., 2019), pseudouridine (Ψ) and 2′-O-methylation (Nm) (Begik *et al.*, 2021) in single-nucleotide-resolution based on native RNA. Moreover, Nanopore DRS is capable of determining RNA secondary structure (Aw *et al*., 2021), and detecting amino acid from native unfolding protein sequence (Hu *et al*., 2021). Methods for detecting RNA modification based on ONT data have been reported in several software packages including Epinano (Liu *et al*., 2019), xPore (Pratanwanich *et al*., 2021), MINES (Lorenz *et al.*, 2020), NanoDoc (Ueda, 2021), Nanom6A (Gao *et al*., 2021), Tombo (Stoiber *et al*., 2017). Epinano, xPore, Nanom6A, and MINES can achieve accuracy in single-base resolution. However, Epinano cannot distinguish m^6A from other modified bases (such as m^1A) due to depending on base-calling errors (Liu *et al*., 2019). Moreover, xPore requires m^6A writer knockout lines as comparison and may be biased due to its strong enrichment in DRACH motif (Pratanwanich *et al*., 2021). Finally, MINES only detects four RRACH motifs (AGACT, GGACA, GGACC, and GGACT) and overlooked other 8 RRACH motifs (Lorenz *et al*., 2020). Nanom6A provides method using Nanopore direct RNA reads for identification of both qualitative and quantitative m^6A modification in single-base resolution based on XGBoost algorithm, providing a highly precise transcriptome-wide identification, quantification and sequence contexts (RRACH) of m^6A modification (Gao *et al*., 2021). In this study, we provide detail procedure for high accuracy of quantitation of RNA modification in single-base resolution. In brief, we provide comprehensive procedure on the identification of AS, APA and m^6A based on DRS or PacBio Iso-Seq data. The Linux Bash Shells and scripts for Nanom6A and PRAPI analysis are now available in https://github.com/GuInNGS/NanoPrapi and https://github.com/Bio-protocol/Pipeline_for_Identification_m6A_AS_and_APA_with_Long_Read.

Equipment

1. Linux cluster (We recommend computer with at least 16GB RAM and multiple CPU cores)
   Intensive computing is required for Guppy and Tombo which used fast5 data as input file
2. Server with Linux system (We recommend at least 16GB RAM)
3. Nanopore Sequencing
   MinION or GridION Flow Cell (R9.4.1) (Oxford Nanopore Technologies, Cat. no. FLO-MIN106) (https://store.nanoporetech.com/flow-cell-r9-4-1.html)

Software and Data sets

Software

1. Anaconda (https://www.anaconda.com/products/individual)
   A toolkit included thousands of open-source packages and libraries. User can also use Miniconda (https://docs.conda.io/en/latest/miniconda.html), a free minimal installer for Anaconda, to create environment and install software.
2. Guppy (v3.6.1) (https://community.nanoporetech.com/downloads)
   Guppy from Oxford Nanopore sequencing data processing toolkit can perform basecalling to generate FASTQ file and an additional FAST5 file that contains basecalling information, which is available to ONT customers. Users should be an existing customer or register an account through the Nanopore community to download Guppy.
3. Ont_fast5_api (v0.3.2) (https://github.com/nanoporetech/ont_fast5_api)
   Module of multi_to_single_fast5 was a conversion tool from Ont_fast5_api package to transfer single fast5 with big size into small size.
4. Tombo (Stoiber *et al*., 2017) (v1.5.1) (https://github.com/nanoporetech/tombo)
   The re-squiggle algorithm from Tombo program assigned raw signal and associated base calls to transcriptome reference sequences for downstream analysis.

5. Nanom6A (Gao *et al*., 2021) (2021_3_18 version) (https://github.com/gaoyubang/nanom6A)
   A software for identification and quantification of m^{6}A modification at single-base-resolution base on Nanopore raw data.
6. PRAPI (Gao *et al.*, 2018) (v1.0) (http://forestry.fafu.edu.cn/tool/PRAPI/)
   Iso-Seq reads analysis tool (post-transcriptional regulation analysis pipeline), for identification of post-transcriptional events including AS and APA et. al.
7. LoRDEC (Salmela and Rivals, 2014) (v0.9) (https://gite.lirmm.fr/lordec/lordec-releases/-/wikis/home)
   A tool for error correction of long reads from the third-generation sequencing using the short-sequence reads from second-generation sequencing platform. LoRDEC uses high-accuracy NGS data to construct *de Bruijn* Graph (DBG) for correcting the errors of the third-generation long reads from PacBio or Oxford platforms.
8. Picard (v2.26.8) (https://github.com/broadinstitute/picard)
9. GMAP (Wu and Watanabe, 2005; Wu and Nacu, 2010) (version 2019-12-01) (http://research-pub.gene.com/gmap/)

Input data

1. For m^{6}A identification based on nanom6A, input data should be transcriptome reference sequences and the fast5 file from Nanopore DRS, which included raw signal. Each multi-fast5 file contains 4000 single fast5 reads. Each single fast5 read contains raw signal, and configuration information. The fast5 file is stored in HDF5 format and can be viewed by HDFView (https://www.hdfgroup.org/downloads/hdfview/). One typical HDF5 format looks like this (Figure 1).

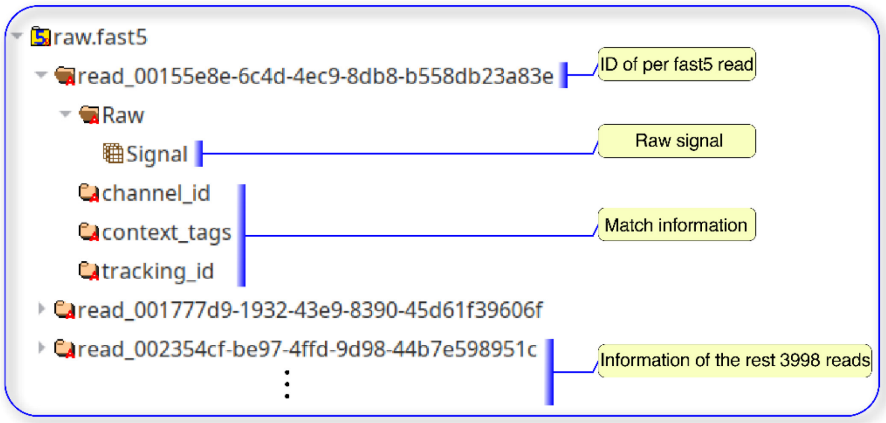


Figure 1. A typical HDF5 format for FAST5 file from DRS.

2. In the example for identification of AS and APA based on PRAPI, input file should include genome file(.fasta), annotation file (genePred format), PacBio Iso-Seq read (.fasta) or Nanopore long reads (.fasta), which has been corrected by LoRDEC (Salmela and Rivals, 2014). PRAPI also can take RNA-Seq based on Illumina platform as input using sorted and indexed alignment files (BAM). The annotation file in GenePred table format (https://genome.ucsc.edu/FAQ/FAQformat.html#format9) is used during the process of AS identification. Annotation file with genePred format should have nine columns including name of gene, name of chromosome, strand, transcription start position, transcription end position, coding region start position, coding region end position, number of exons, exon start positions, and exon end positions, respectively.

Procedure

Case study

A. Usage of Nanom6A

Note: Following step is only available for Nanopore DRS data.

1. Install the dependencies
   a. Install miniconda on Linux
      Miniconda is an excellent bio-software installer, which is friendly and convenient for building various software and its dependencies. Following command can be run to install miniconda:
      wget https://mirrors.tuna.tsinghua.edu.cn/anaconda/miniconda/Miniconda3-py37_4.8.3-Linux-x86_64.sh
      ```
      chmod 777 Miniconda3-py37_4.8.3-Linux-x86_64.sh
      ./Miniconda3-py37_4.8.3-Linux-x86_64.sh
      ```
   b. Install Guppy, ont-fast5-api and Tombo
      Guppy is a base caller software, which shows the best performance in accuracy and speed among present long-read base-calling tools (Wick *et al*., 2019). The software can be installed by following command:
      ```
      tar -xvzf ont-guppy_3.1.5_linux64.tar.gz
      export PATH=`pwd`/ont-guppy/bin:$PATH
      pip install ont-fast5-api
      pip install ont-tombo
      ```
2. Installation guide for nanom6A
   a. Download the nanom6A package
      nanom6A_2021_3_18.tar.gz package can be downloaded from following link:
      https://drive.google.com/drive/folders/1Dodt6uJC7lBihSNgT3Mexzpl_uqBagu0?usp=sharing
      This package includes required dependence and one test data. User can create a new conda environment for nanom6A using following command:
      ```
      tar -xvzf nanom6A_2021_3_18.tar.gz
      cd nanom6A_2021_3_18
      conda create -n nanom6A -f conda.yml
      ```
   b. Installation testing
      The script can be tested using following command:
      ```
      sh run_source_code.sh
      ```
3. Running guppy_basecaller
   a. The Guppy is required to run and produce an additional fast5 file with basecalling information with following command:
      ```
      guppy_basecaller -i <raw_fast5_file_path> -s <output path> --num_callers 40 --recursive -fast5_out --config rna_r9.4.1_70bps_hac.cfg
      ```

      In this case, the dataset is generated using FLO-FLG001 flowcell and SQK-RNA002 kit. Thus, we can use the rna_r9.4.1_70bps_hac model. After base-calling, Guppy generates FASTQ files and FAST5 files. FASTQ file contains nucleotide sequence and corresponding quality of each base. FASTQ could be used for downstream AS and APA analysis. Information from basecalling process is added to FAST5 during Tombo analysis.
4. Transferring multi-fast5 file to single-fast5 file
   a. Tombo only supports single fast5 file with size about 150kb as input. However, DRS can generate muti-fast5 file, which generally contains 4000 sequences (about 350MB). Thus ont_fast5_api should be used to transfer it into single-fast5 file using following command:
      ```
      multi_to_single_fast5 --input_path <fast5> --save_path <single_fast5_path> --recursive
      ```

5. Tombo resquiggle
   a. Tombo resquiggle command assigned raw signal through matching the contiguous reference bases with contiguous sections ("Events") of raw signal. Vitables software (https://vitables.org/) can be used for obtaining basecall-group information. Following command is required to perform downstream modification analyses:

```
tombo resquiggle --overwrite --basecall-group Basecall_1D_000
<single_fast5_path> <referance.transcript.fa> --processes 40 --fit-
global-scale -include-event-stdev
```

   CRITICAL: *The reference.transcript.fa file (FASTA) in resquiggle command should be transcriptome file rather than reference genome file.*
   b. Basecall_1D_000 and Segmentation_000 will be added in FAST5 file if raw FAST5 file has never been basecalled. HDF5view can view the output result to obtain Basecall-group information. After base-calling by Guppy and re-squiggle by Tombo, typical FAST5 file looks like this (Figure 2).

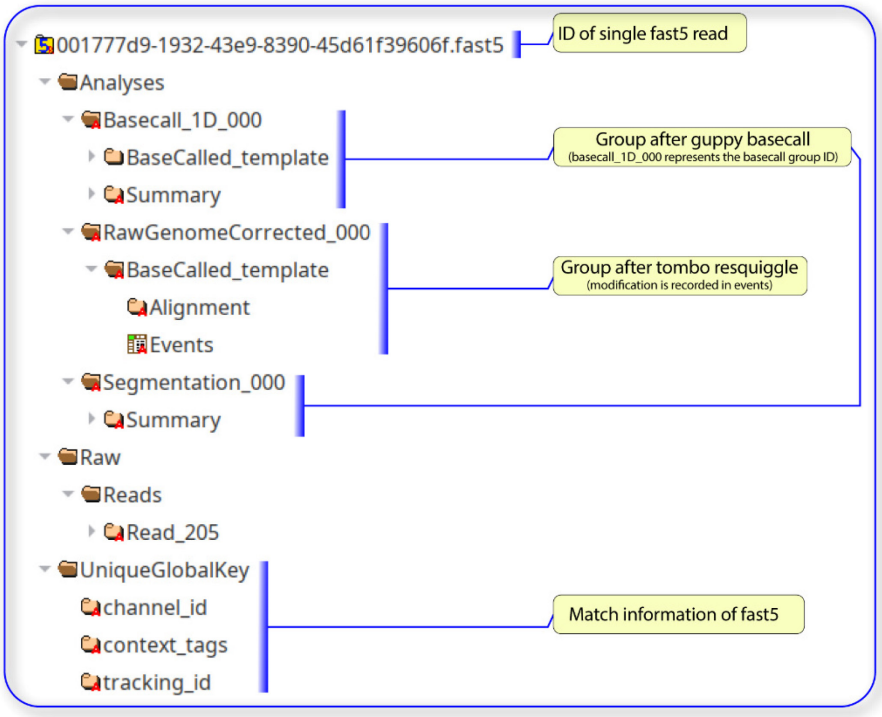


Figure 2. Basecall-group information from HDFview.

6. Identification of modified nucleotide using nanom6A
   a. The path for all FAST5 files from Tombo re-squiggle needs to be collected with the following command:

```
Find <single_fast5_path> -name "*.fast5" > fast5_files.txt
```

   b. Signals should be extracted from FAST5 file before m^{6}A prediction

```
extract_raw_and_feature_fast --cpu=4 --fl=fast5_files.txt -clip=10
-o <name for output file>
```

   This step will generate fa and tsv files, which will be used for downstream analysis for m^{6}A identification.
   c. Identification of m^{6}A site based on nanom6A

```
predict_sites --cpu 4 -i <name of output file from
extract_raw_and_feature_fast> -o <name for output file> -r <BED file
```

```
for gene annotation> -g <genome_file> --model <path of model in nanom6A>
```

Sequence dictionary should be created by picard before running above command:

```
java -jar picard.jar CreateSequenceDictionary -R /path/to/genome.fasta -O genome.dict
```

BED file provides information about gene annotation separated by TAB, which contains chromosome name, start position, end position, gene name, score and transcription direction, respectively.

B. Usage of PRAPI

Note: Following step is available for both Nanopore and Pacbio data.

1. Installation of PRAPI
   PRAPI can be installed by conda using following command:
   ```
   wget http://forestry.fafu.edu.cn/tool/PRAPI/prapi_env.yaml
   conda env create prapi_env -f prapi_env.yaml
   conda activate prapi_env
   pip install -i https://pypi.anaconda.org/gaoyubang/simple splicegrapher
   ```
   Note: Above command-line options can be installed in Linux operating system.
2. Preparation of GMAP index files and FASTA input file from long read sequencing
   Genome reference index should be built by GMAP (https://github.com/juliangehring/GMAP-GSNAP/blob/master/README):
   ```
   gmap_build -D <gmap_path> -d <specie_name> <genome file>
   ```

   Input sequences for PRAPI should be FASTA format. Both PacBio and ONT platform always get random error distribution and high error rate (Koren *et al*., 2012; Weirather *et al*., 2017). However, these error could be corrected base on alignment and assembly by providing second-generation sequencing reads (Zhang *et al*., 2020). Thus, long-read file should be corrected by lordec-correct using following command:
   ```
   lordec-correct -T <number of threads> -k 19 -s 3 -S statistics.log (Not necessary) -i <your long read data(Fastq/a)> -2 <short read file(fastq/a)> -o <output data(fasta)>
   ```

   In this command line, parameter "*-k*" indicates k-mer size, which means the piece of sequence of length (k). The option "*-s*" controls the minimum abundance for each kmer and mostly 3 is enough for data correcting.
3. Preparation of configuration file
   Configuration file can be edited by Vim text Editor in Linux system and saved with conf.txt as file name. Configuration file contains several important parameters:
   Long_reads : PacBio Iso-Seq or DRS long reads with FASTA format.
   GMAP_IndexesDir : Directory of genomic index files built by gmap_build program using GMAP
   Genome_Annotation : Reference annotation with GenePred table format (https://genome.ucsc.edu/FAQ/FAQformat.html#format9)
4. Identification of AS and APA using PRAPI
   Finally, following command can be used to generate post-transcriptional events.
   ```
   Pacbio_v16.py -c conf.txt >prapi.log 2>&1
   ```
5. Test file (optional)
   PRAPI also provides one testing script to identify the successful installation of PRAPI using following command:
   ```
   wget http://forestry.fafu.edu.cn/tool/PRAPI/download/v2/test_v2.tar.gz -O - | tar xvzf -
   cd test_v2
   sh run.sh
   ```

Result interpretation

The result directory of nanom6A includes ratio.x.tsv which contains the information of gene name, chromosome, the coordinate site of m^{6}A, the number of m^{6}A modified reads, the number of total reads, and the ratio of the m^{6}A site. The file named genome_abandance.x.bed contains information of name and coordinate information of chromosome, gene name, ID and position of single FAST5 read and motif (kmer). The x in ratio.x.tsv and abandance.x.bed represents the probability of modification. The default probability is 0.5.

Visualization result can be generated by the *nanoplot* command:

```
nanoplot --input <name of output file from predict_sites> -o plot_nano_plot.
```

The output figure provides structure of transcripts and m^{6}A sites highlighted by purple vertical line (Figure 3).

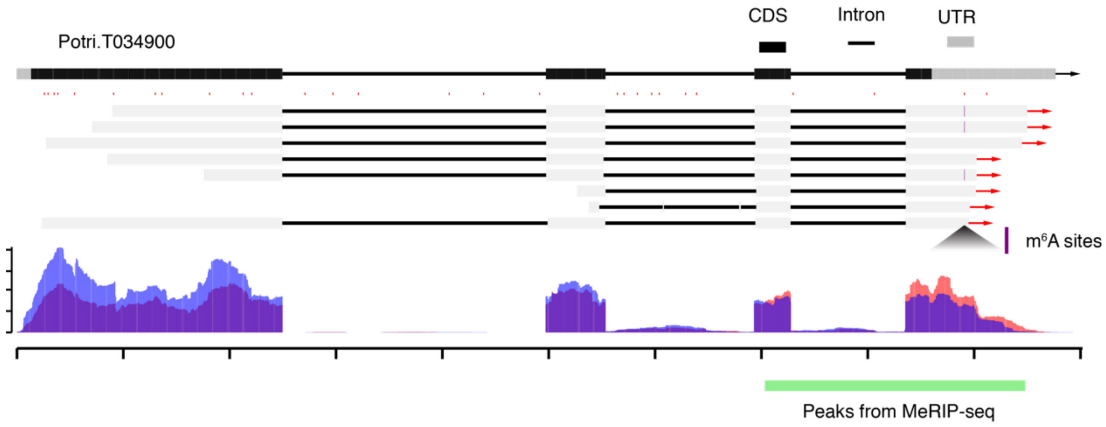


Figure 3. Visualization of m^{6}A sites.
The modified sites in DRS reads were marked with purple vertical line. Red and Blue colors in wiggle plot from MeRIP-seq represented Input and IP library, respectively.

Major output of PRAPI included AS and APA events:
apa.txt_gene.csv provides APA information, which includes the gene name, reads, and positions for cleavage sites.
as.*.txt_gene.csv provides AS information, which includes the gene name, id, gene symbol, chromosome, strand, and types of AS events, including intron retention, exon skipping, alternative 5′donor and alternative 3′donor events.
In addition to AS and APA, PRAPI also provides several other information:
ati.txt_gene.csv presents the alternative transcript start site information.
nat.txt_gene.csv presents information about nature antisense transcripts including gene name, plus read, and minus read that formed NATs.
Lib*_circle.txt presents the basic information of circle RNAs including start site of junction, end site of junction, name of read, etc.

The visualization of AS and APA contains two categories in the output directories: Annotation_Gene and Novel_Gene, which represented long reads located in annotated region and unannotated region, respectively. For example, the graph from Potri.002G178700 shows AS and APA events, which are marked in the figure (Figure 4).

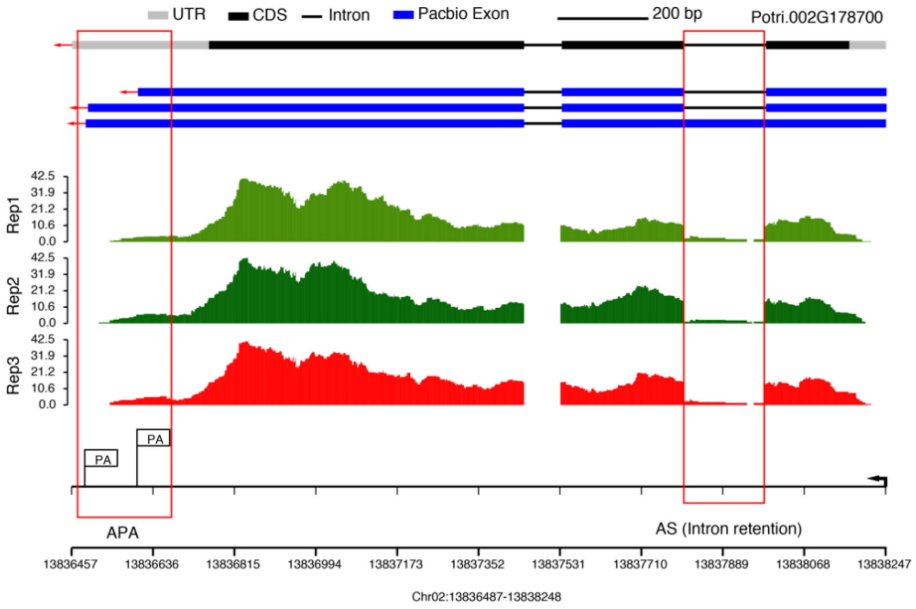


Figure 4. Visualization of AS and APA from PRAPI based on PacBio Iso-Seq.
Reads with blue color represented transcripts from Iso-Seq. Wiggle plot represented the transcripts abundance from RNA-Seq.

Acknowledgments

This work was supported by the National Natural Science Foundation of China Grant (Grant No. 31971734, 31570674, and 31800566), Innovation Fund for Science & Technology Project from Fujian Agriculture and Forestry University (CXZX2020093A), the Natural Science Foundation of Fujian Province (Grant No. 2021J02027), Fujian Forest Seedling Technology Project, and Distinguished Young Scholar Program of Fujian Agriculture and Forestry University (xjq202017).

Competing interests

The authors declare that they have no competing interests.

References

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A. and Reddy, A. S. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun* 7: 11706.

Aw, J. G. A., Lim, S. W., Wang, J. X., Lambert, F. R. P., Tan, W. T., Shen, Y., Zhang, Y., Kaewsapsak, P., Li, C., Ng, S. B., *et al.* (2021). Determination of isoform-specific RNA structure with nanopore long reads. *Nat Biotechnol* 39(3): 336-346.

Begik, O., Lucas, M. C., Pryszcz, L. P., Ramirez, J. M., Medina, R., Milenkovic, I., Cruciani, S., Liu, H., Vieira, H. G. S., *et al.* (2021). Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol* 39(10): 1278-1291.

Chen, L., Shi, X., Nian, B., Duan, S., Jiang, B., Wang, X., Lv, C., Zhang, G., Ma, Y. and Zhao, M. (2020). Alternative Splicing Regulation of Anthocyanin Biosynthesis in Camellia sinensis var. assamica Unveiled by PacBio Iso-Seq. *G3 (Bethesda)* 10(8): 2713-2723.

Di Giammartino, D. C., Nishida, K. and Manley, J. L. (2011). Mechanisms and consequences of alternative polyadenylation. *Mol Cell* 43(6): 853-866.

Gao, Y., Liu, X., Wu, B., Wang, H., Xi, F., Kohnen, M. V., Reddy, A. S. N. and Gu, L. (2021). Quantitative profiling of N^6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *Genome Biol* 22(1): 22.

Gao, Y., Wang, H., Zhang, H., Wang, Y., Chen, J. and Gu, L. (2018). PRAPI: post-transcriptional regulation analysis pipeline for Iso-Seq. *Bioinformatics* 34(9): 1580-1582.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., *et al*. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 15(3): 201-206.

Hu, Z. L., Huo, M. Z., Ying, Y. L. and Long, Y. T. (2021). Biological Nanopore Approach for Single-Molecule Protein Sequencing. *Angew Chem Int Ed Engl* 60(27): 14738-14749.

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* 8(7): 533-543.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., *et al*. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol* 30(7): 693-700.

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E. and Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 12(8): 767-772.

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., Schwartz, S., Mattick, J. S., Smith, M. A. and Novoa, E. M. (2019). Accurate detection of m(6)A RNA modifications in native RNA sequences. *Nat Commun* 10(1): 4079.

Lorenz, D. A., Sathe, S., Einstein, J. M. and Yeo, G. W. (2020). Direct RNA sequencing enables m(6)A detection in endogenous transcript isoforms at base-specific resolution. *RNA* 26(1): 19-28.

Pratanwanich, P. N., Yao, F., Chen, Y., Koh, C. W. Q., Wan, Y. K., Hendra, C., Poon, P., Goh, Y. T., Yap, P. M. L., Chooi, J. Y., *et al*. (2021). Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat Biotechnol* 39(11): 1394-1402.

Rhoads, A. and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 13(5): 278-289.

Salmela, L. and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30(24): 3506-3514.

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q. and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* 111(51): E5593-5601.

Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. and Akeson, M. (2019). Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One* 14(5): e0216709.

Stoiber, M., Quick, J., Egan, R., Eun Lee, J., Celniker, S., Neely, R. K., Loman, N., Pennacchio, L. A. and Brown, J. (2017). *De novo* Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *BioRxiv*: 094672.

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J. and Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* 11(1): 1438.

Tian, B. and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* 18(1): 18-30.

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J. and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* 19(1): 40.

Ueda, H. (2021). nanoDoc: RNA modification detection using Nanopore raw reads with Deep One-Class Classification. *BioRxiv*: 295089.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X. J., Buck, D. and Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* 6: 100.

Wick, R. R., Judd, L. M. and Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* 20(1): 129.

Wu, T. D. and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26(7): 873-881.
Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9): 1859-1875.
Zhang, H., Jain, C. and Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. *BMC Genomics* 21(Suppl 6): 889.
Zhang, Z., Chen, L. Q., Zhao, Y. L., Yang, C. G., Roundtree, I. A., Zhang, Z., Ren, J., Xie, W., He, C. and Luo, G. Z. (2019). Single-base mapping of m(6)A by an antibody-independent method. *Sci Adv* 5(7): eaax0250.

Supplementary information

1. Data and code availability: All data and code have been deposited to GitHub: https://github.com/ Bio-protocol/Pipeline_for_Identification_m6A_AS_and_APA_with_Long_Read.