Bioinformatics, XXXX, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Subject Section

NanoTrans: an integrated computational framework for comprehensive transcriptome analyses with Nanopore direct-RNA sequencing

Fan Wang[1,^], Xinxin Zhang[1,^], Li Zhang[1,*], Jing Li[1,*], Jia-Xing Yue[1,*]

[1]State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Sun Yat-sen University Cancer Center, Guangzhou, China.

^Contribute equally.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Nanopore direct-RNA sequencing (DRS) provides the direct access to native RNA strands with full-length information, shedding light on rich qualitative and quantitative properties of gene expression profiles. Here with NanoTrans, we present an integrated computational framework that comprehensively covers all major DRS-based application scopes, including isoform clustering and quantification, poly(A) tail length estimation, RNA modification profiling, and fusion gene detection. In addition to its merit in providing such a streamlined one-stop solution, NanoTrans also shines in its workflow-orientated modular design, batch processing capability, rich tabular and graphic report outputs, as well as automatic installation and configuration support. Finally, by applying NanoTrans to real DRS datasets of yeast, *Arabidopsis*, as well as human embryonic kidney and cancer cell lines, we further demonstrated its utility, effectiveness, and efficacy across a wide range of DRS-based application settings.

Availability and implementation: NanoTrans is written in bash, Perl, and R. It is free for use under the MIT license, available at https://github.com/yjx1217/NanoTrans. The key raw data are uploaded to the Research Deposit public platform (www.researchdata.org.cn), with the approval RDD number of RDDXXXXXXXXXXX.

Contact: zhangli@sysucc.org.cn; lijing3@sysucc.org.cn; yuejiaxing@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The recent years have witnessed a rapid advancement of long-read sequencing technologies such as PacBio and Oxford Nanopore in terms of both read length and read accuracy. Combined with their single-molecular nature, this opens up exciting opportunities for many applications in both basic and applied biomedical research. For example, long-read-based DNA sequencing has become the go-to-choice for most genome sequencing projects nowadays, empowering the recent production of chromosome-level telomere-to-telomere (T2T) genome assemblies for diverse organisms (including human) (Yue et al., 2017; Jiao and Schneeberger, 2020; Nurk et al., 2022). Likewise, long-read-based RNA sequencing, best represented by Oxford Nanopore's direct RNA sequencing (DRS), has also become increasingly popular, offering rich biological information while avoiding reverse transcription and amplification biases (Garalde et al., 2018). Accordingly, a number of dedicated bioinformatic tools have recently been developed for analyzing Nanopore DRS data with emphasis on specific applications, such as Mandalorion (Byrne et al., 2017) and Flair (Tang et al., 2020) for full-length isoform clustering, NanoCount (Gleeson et al., 2022) for expression quantification, nanopolish (Loman et al., 2015) and tailfindr (Krause et al., 2019) for poly(A) tail length estimation, EpiNano (Liu et al., 2019) and Xpore (Pratanwanich et al., 2021) for

RNA modification, LongGF (Liu et al., 2020) and JAFFAL (Davidson et al., 2022) for gene fusion detection. While these tools greatly facilitated the Nanopore DRS data analysis in their corresponding application scope, the lack of a unified framework that combines the power of these different tools prevents the broader research community to take full advantage of the Nanopore DRS technology. Furthermore, some of these tools are relatively challenging to set up and run, which imposes a significant technical barrier for general users. Motivated by such practical needs, we developed NanoTrans, an easy-to-use one-stop solution for general users to perform comprehensive Nanopore DRS data analysis along with a streamlined workflow.

2 Descriptions and highlights

NanoTrans is a Linux-based computational framework for automated high-throughput Nanopore DRS data analysis. NanoTrans is self-contained, with all dependencies automatically installed and configured via a pre-shipped installer script. The design of NanoTrans is workflow-oriented, with a series of task-specific modules numbered according to their processing order (Figure 1). Briefly, NanoTrans first performs Nanopore reads basecalling and reference genome preprocessing with its two starting modules numbered with "00". The basecalling step here can be processed either in GPU or CPU mode. Regarding the reference genome setup, NanoTrans supports all organisms with reference genome and annotation retrievable via Ensembl (https://www.ensembl.org) or its sister sites (e.g., Ensembl Fungi, Ensembl Plants, Ensembl Protists, and Ensembl Metazoa). The basecalled fastq reads are subsequently mapped to the preprocessed genome in a splicing-aware manner (module "01"), after which isoform clustering and quantification are further performed accordingly (module "02"). Based on the clustered and quantified isoforms, NanoTrans can perform different application-specific analyses such as isoform expression and splicing comparison (module "03"), isoform RNA modification identification (module "04"), isoform poly(A) tail length profiling (module "05"). In addition, reference-based gene fusion detection (module "06") can be applied as well. A user-defined master sample table is used for specifying sample list, experimental design, and reads locations, based on which automatic batch processing and between group comparison are natively supported.

3 Application demonstration

To demonstrate the application of NanoTrans in real case scenarios, we applied it to four different public Nanopore DRS datasets from the budding yeast Saccharomyces cerevisiae, the mustard plant Arabidopsis thaliana, the human embryonic kidney cell line (HEK293), and the human lung adenocarcinoma (A549) and leukemia (K562) cancer cell lines, respectively (Tudek et al., 2021; Parker et al., 2020; Gewartowska et al., 2021; Chen et al., 2021). These datasets were selected to demonstrate the performance of different functional modules of NanoTrans across different application settings. For each dataset, raw Nanopore fast5 reads were retrieved according to the sequencing run accession of the original study and fed into NanoTrans for basecalling and downstream analysis (Supplementary Note). With the yeast dataset, we quantified the expression level and poly(A) tail length of each isoform and replicated the observation from the original study showing a negative correlation between RNA abundance and poly(A) tail length. With the *Arabidopsis* dataset, we examined differential expression, differential splicing, and RNA modification. With the human embryonic kidney cell line dataset, we verified the finding of the original study that TENT5A is responsible for elongation the poly(A) tail of mRNA molecules. Finally, with the human cancer cell lines, we identified a list of gene fusion alterations for the A549 and K562 cell lines respectively, highlighting the frequent chromosomal rearrangements in cancer genomes.

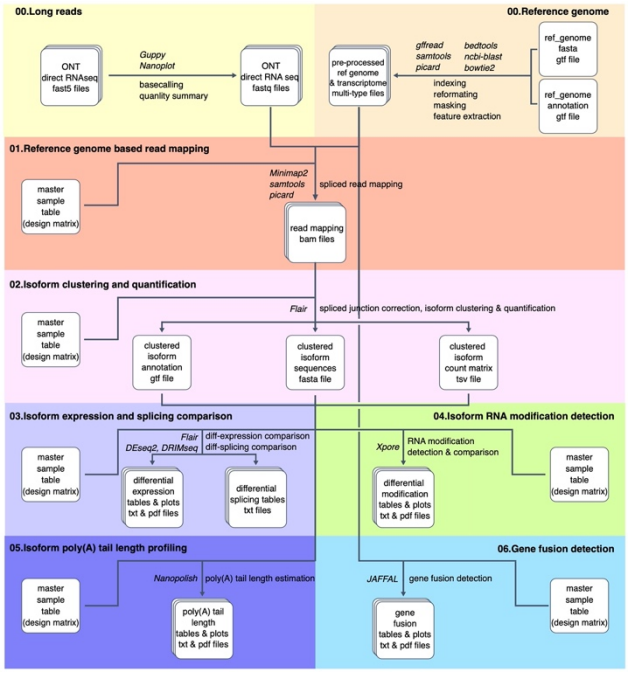

Fig. 1. The framework design and application demonstration of NanoTrans. The names of third-party tools employed in each step are denoted in italics.

4 Conclusions

We developed NanoTrans, an integrated and versatile computational framework that supports comprehensive transcriptome analysis based on Nanopore DRS data. Starting with raw fast5 reads, NanoTrans automates the full workflow of DRS data analysis and covers a wide range of applications settings including isoform clustering and quantification, differential expression and splicing examination, RNA modification identification, poly(A) tail length profiling, and gene fusion detection. Given the increasing adoption of Nanopore DRS technology, we believe NanoTrans will become a highly useful tool to help researchers to fully explore the power of this exciting technology with rich biological insights obtained.

Acknowledgements

We thank Dr. Song Gao from Sun Yat-sen University Cancer Center for inspiring discussion, which motivated the initiation of this work. We thank Dr. Long Wang from Nanjing University for the help in downloading raw fast5 reads of some of the public datasets used in this study for software testing and application demonstration.

Funding

This work is supported by National Natural Science Foundation of China (32070592 to J.-X. Y. and 32000395 to J. L.), Natural Science Foundation of Guangdong Province (2022A1515010717 to J.-X. Y. and 2022A1515011873 to J. L.), Guangdong Basic and Applied Basic Research Foundation (2019A1515110762 to J.-X. Y.), Guangdong Pearl River Talents Program (2019QN01Y183 to J.-X. Y.), Guangzhou Municipal Science and Technology Bureau (202102020938 to J. L.). The funders have not played any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

References

Byrne,A. et al. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun, 8, 16027.

Chen,Y. et al. (2021) A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. BioRxiv, 2021.04.21.440736.

Davidson,N.M. et al. (2022) JAFFAL: detecting fusion genes with long-read transcriptome sequencing. Genome Biology, 23, 10.

Garalde,D.R. et al. (2018) Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods, 15, 201–206.

Gewartowska,O. et al. (2021) Cytoplasmic polyadenylation by TENT5A is required for proper bone formation. Cell Reports, 35.

Gleeson,J. et al. (2022) Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Research, 50, e19.

Jiao,W.-B. and Schneeberger,K. (2020) Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun, 11, 989.

Krause,M. et al. (2019) tailfindr: alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing. RNA, 25, 1229–1241.

Liu,H. et al. (2019) Accurate detection of m6A RNA modifications in native RNA sequences. Nat Commun, 10, 4079.

Liu,Q. et al. (2020) LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. BMC Genomics, 21, 793.

Loman,N.J. et al. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods, 12, 733–735.

Nurk,S. et al. (2022) The complete sequence of a human genome. Science, 376, 44–53.

Parker,M.T. et al. (2020) Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. eLife, 9, e49658.

Pratanwanich,P.N. et al. (2021) Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat Biotechnol, 39, 1394–1402.

Tang,A.D. et al. (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun, 11, 1438.

Tudek,A. et al. (2021) Global view on the metabolism of RNA poly(A) tails in yeast Saccharomyces cerevisiae. Nat Commun, 12, 4951.

Yue,J.-X. et al. (2017) Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nature Genetics, 49, 913–924.