LoRTIA – The Long-read RNA-Seq Transcript Isoform Annotator Toolkit

Zsolt Balázs[1,2,3], Zsolt Boldogkői[3]

[1] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
[2] Medical Informatics, University Hospital Zurich, Zurich, Switzerland
[3] Department of Medical Biology, Faculty of Medicine, University of Szeged, Szeged, Hungary

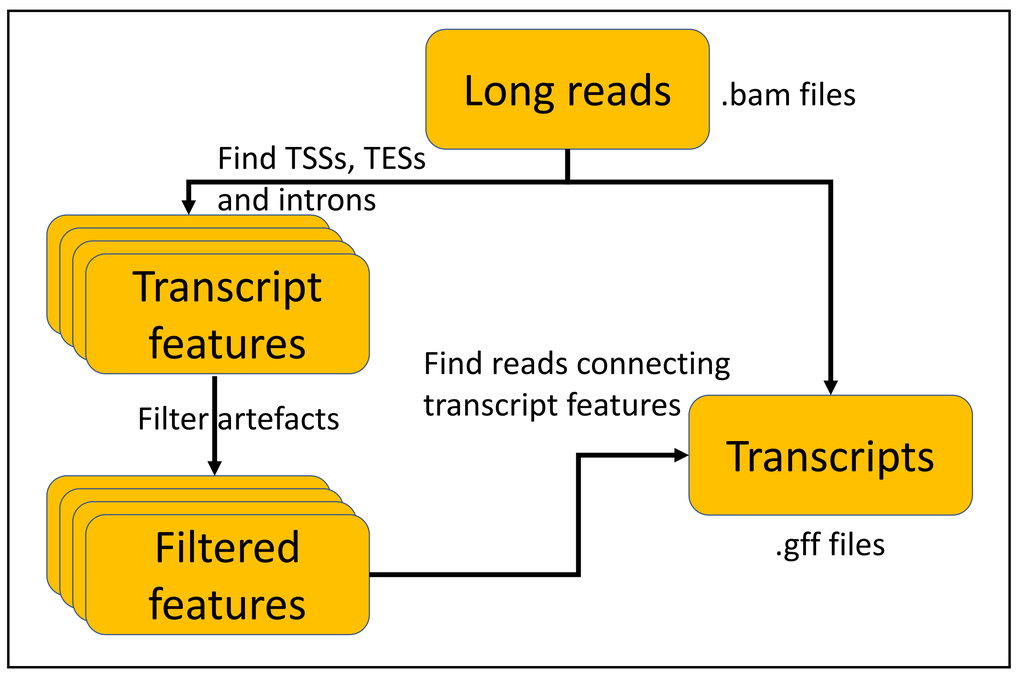
Introduction

The genetic information is first expressed during transcription. Transcripts carry the blueprints for protein synthesis, but transcripts can also regulate protein synthesis or even transcription itself. Knowing the transcriptional repertoire of an organism is crucial to the understanding of its molecular biology. Long-read sequencing, especially Nanopore sequencing has greatly furthered the field of transcriptomics. Previously, transcript models had to be assembled from short-read sequencing data, now, long-read sequencing is capable of capturing full-length transcripts. While transcript discovery based on long-read sequencing is more straight-forward compared to short-read sequencing, RNA fragmentation, RNA degradation and library preparation artefacts still complicate long-read transcriptome analysis[1]. Available transcript annotation tools rely heavily on existing annotations and are likely to label fragmented reads or technical artefacts as full-length transcripts.

Aim

In order to be able to characterize the transcriptomes of model and non-model organisms alike, we developed a software toolkit that does not require existing transcript annotations and still efficiently filters out technical artefacts.

Implementation

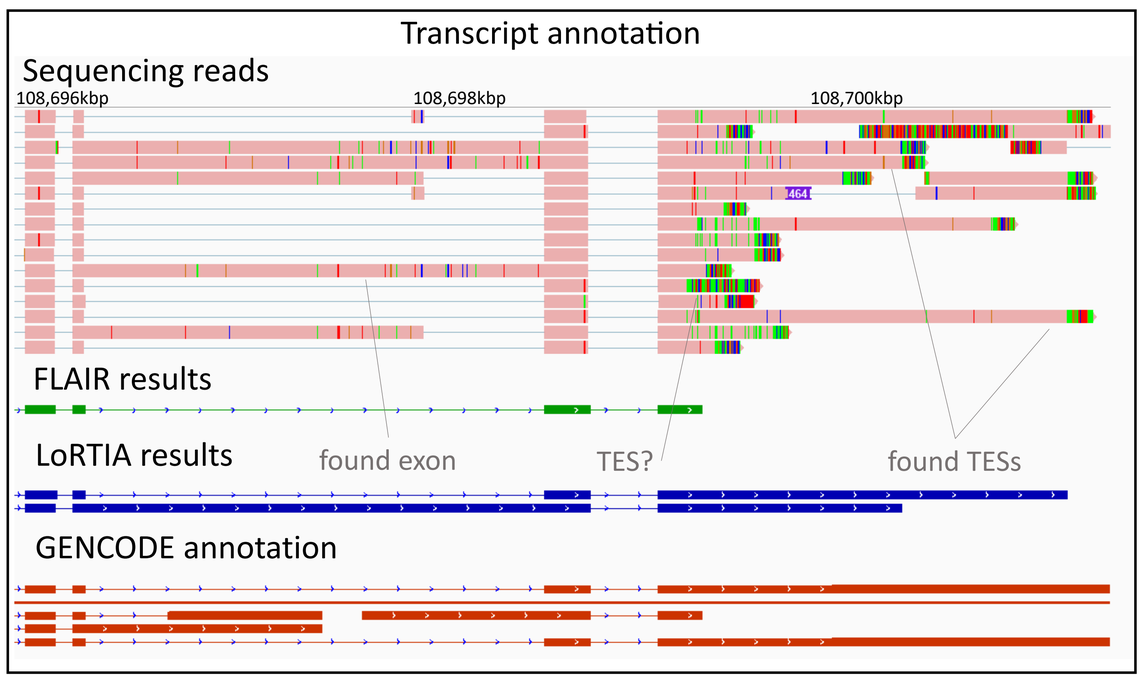


Using genome-aligned reads, the software first identifies transcript features such as transcription start and end sites (TSSs and TESs) and introns. These features are then filtered based on sequence context and read coverage. Finally, full-length transcript isoforms are identified by reads which connect high-confidence transcript features.

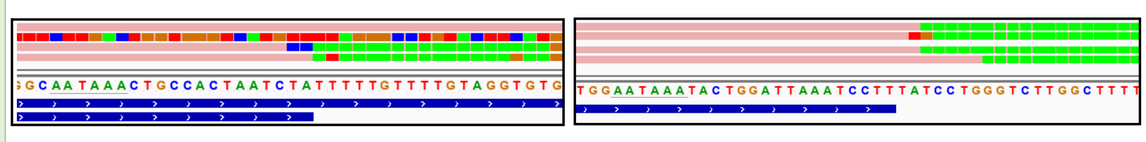
Methods

Two features of the LoRTIA toolkit are showcased on this poster: transcript discovery and the filtering of artefacts. The LoRTIA toolkit was tested on the Oxford Nanopore Human Reference Datasets generated by the Nanopore WGS Consortium[2]. Sensitivity in transcript discovery is demonstrated by comparison to a collapsing-based transcript annotation tool FLAIR[3] (v1.5) which was supplemented with the GENCODE[4] (v34) database. The accuracy of polyA cleavage site detection in cDNA sequencing was tested using dRNA sequencing of the same cell-line. The dRNA sequencing dataset was used as ground truth. The accuracy of the LoRTIA cleavage site detection was compared to using the polyA_DB[5] (v3) dataset, the SQANTI[6] algorithm as well as the commonly applied internal priming filter[7] (where every site with at least 6 consecutive As directly preceding or 13 As out of the last 20 nucleotides preceding the cleavage site is marked as an internal priming artefact).

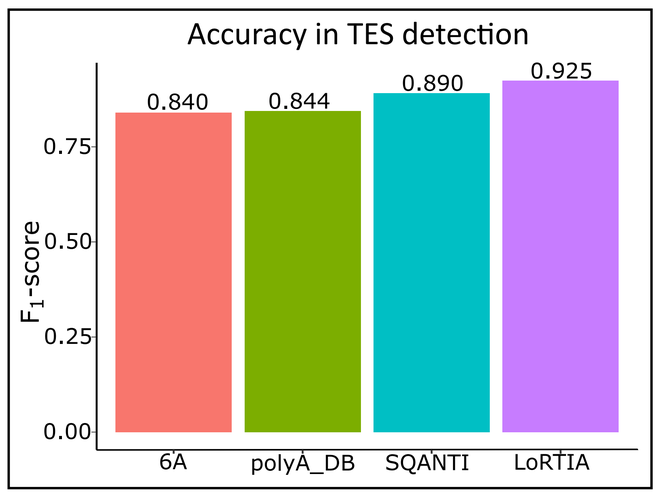
Results



The IGV screenshot demonstrates the advantages of LoRTIA over collapse-based transcript annotators. LoRTIA finds the unannotated exon missed by FLAIR. It doesn't annotate the TES from the GENCODE database when it is not supported by the sequencing data but detects two unannotated distal TESs.



The distal TESs are supported by polyA signals and polyA$^+$ reads.



The TS-filter used by LoRTIA detected TESs with a higher accuracy than conventionally used polyA cleavage-site filtering methods. It outperformed the most commonly used internal priming filtering method (6A), was more sensitive than using the polyA_DB and it was slightly more accurate than SQANTI.[8]

Conclusions

The LoRTIA toolkit is a reference-genome based transcript annotator which identifies transcript features as a basis for transcript annotation. Due to the toolkit's feature-based transcript discovery algorithm, technical artefacts can be filtered more thoroughly than with other long-read RNA sequencing tools. LoRTIA allows for specific transcript detection even in the absence of prior annotations, thereby facilitating the transcriptome analysis of non-model organisms.

References

1 Boldogkői, Z. *et al. Trends Microbiol.* 27, 578–592 (2019)
2 Workman, R. E. *et al. Nat. Methods.* 16, 1297–1305 (2019)
3 Tang, A. D. *et al. Nat. Comm.* 11, 1438 (2020)
4 Harrow, J. *et al. Genome Res.* 22, 1760–74 (2012)
5 Wang, R. *et al. Nucleic Acids Res.* 46, D315–D319 (2018)
6 Tardaguila, M. *et al. Genome Res.* (2018)
7 Gautheret, D. *et al. Genome Res.* 8, 524–30 (1998)
8 Balázs, Z. *et al. BMC Genomics.* 20, 824 (2019)

https://github.com/zsolt-balazs/LoRTIA

The study was supported by the New National Excellence Programme UNKP-18-3-IV-SZTE-2 grant of the Hungarian Ministry of Human Capacities awarded to Zsolt Balázs.