Comparative analysis of Oxford Nanopore Technologies error-rate during direct-RNA and DNA sequencing

G. Calia[1,2], D. Micheletti[1], M. Moser[1], S. Piazza[1], F. Di Leva[3], A. Cestaro[1]

[1] Research and Innovation Centre, Fondazione Edmund Mach, Italy [2] Faculty of Science and Technology, University of Bolzano, Italy

[3] Biomedicine Institute, EURAC Research, Italy

Background

Direct RNA sequencing is a solution to PCR amplification and fragmentation problems and furnishes a valid alternative to the complementary DNA conversion from other sequencing technologies.

Direct RNA sequencing enables:

- **Sequencing of entire transcripts**
- **Detection of RNA modifications**
- **Identification of isoforms**

The actual limitation of ONT long-reads is the high error rate (~5-15%), and the need to use reads correction tools to decrease the non-biological source of variation.

Can classical correction tools for cDNA sequencing be used also on direct RNA sequencing reads?

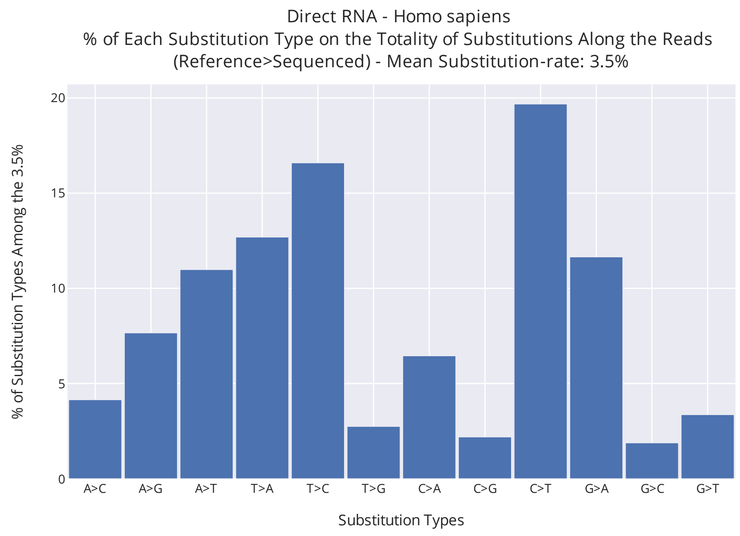
Human transcriptome (direct RNA) vs DNA sequencing

DNA sequencing error rate for insertions, deletions, and substitutions, has been widely investigated so far [1].
However, still little is known about the rate of these errors in direct RNA sequencing.

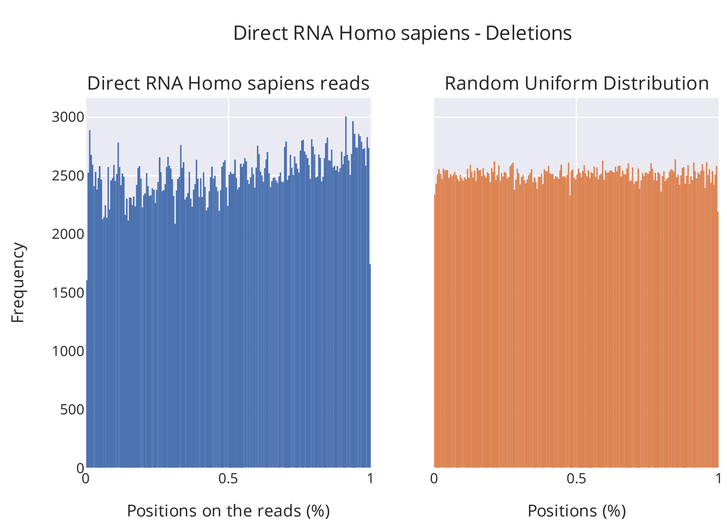
This analysis includes:

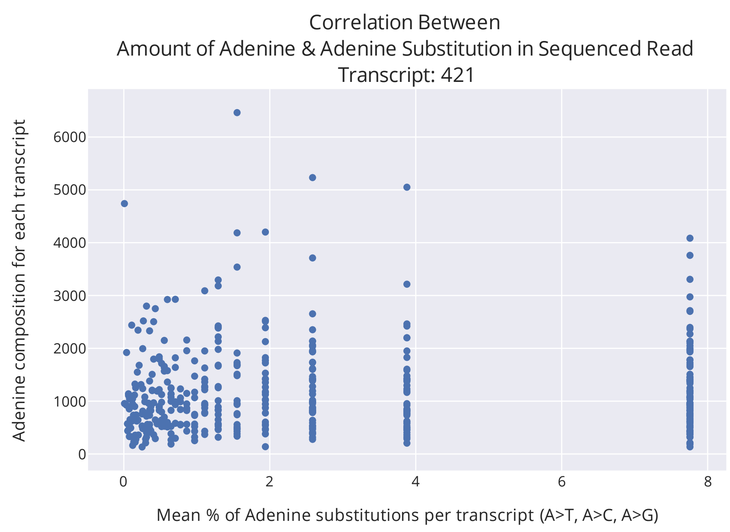
- **Overall error occurrence**
- **Detailed investigation of each error type**
- **Uniformity distribution of errors**
- **Influence of nucleotide composition on error occurrences**

Seq. Method	Insertions (%)	Deletions (%)	Substitutions(%)	Total Error Rate (%)
DNA[1]	3.69	4.54	4.33	12.56
Direct RNA	2.50	4.90	3.50	10.90



	Kolmogorov-Smirnov test p-value
Insertions	1.26e-13
Deletions	1.24e-151
Substitutions	3.11e-06





Analysis methods

BLASTN (.xml output format), is used to map direct RNA reads on a subsample of the reference genome (GRCh38).
Reads are filtered for % identity and query coverage > 80.
Then average percentage of each error type is normalized by read length and the number of mapped reads.

Kolmogorov-Smirnov[2] test (H_0: observed position distribution belongs to a uniform distribution, p-value: 0.05), is used to evaluate the uniformity of error distribution.

Pearson's correlation coefficient[3] is used to evaluate the presence of a correlation between the nucleotide composition of the transcript with the errors rate.

Conclusions

Insertions, deletions, and substitution rates are completely comparable between DNA and direct RNA.

Not uniform distribution along the reads indicating possible biological or technical biases to be further considered.

Transcript nucleotide-composition and the error occurrence do not show a strong relationship.

The error rate estimation are not influenced by this possible analysis-bias.

Software and pipeline for ONT cDNA correction can be valid also for direct RNA reads.

Future Investigations

References

1. Dohm, J. C., Peters, P., Stralis-Pavese, N. & Himmelbauer, H. Benchmarking of long-read correction methods. *NAR Genomics and Bioinformatics* **2**, (2020).
2. Kolmogorov–Smirnov Test. The Concise Encyclopedia of Statistics 283–287 (2008) doi:10.1007/978-0-387-32833-1_214.
3. Pearson's Correlation Coefficient. *Encyclopedia of Public Health* 1090–1091 (2019) doi:10.1007/978-1-4020-5614-7_2569.

FONDAZIONE EDMUND MACH

giulia.calia@fmach.it
https://www.linkedin.com/in/giulia-calia/
https://twitter.com/GiuliaCalia
https://github.com/Giulia-Calia