



HAL
open science

Simulation of RNA sequencIng with Oxford Nanopore Technologies

Camille Marchet, Leandro Lima

► **To cite this version:**

Camille Marchet, Leandro Lima. Simulation of RNA sequencIng with Oxford Nanopore Technologies. JOBIM 2018 - Journées Ouvertes Biologie, Informatique et Mathématiques, Jul 2018, Marseille, France. pp.1-2. hal-01929917

HAL Id: hal-01929917

<https://hal.science/hal-01929917v1>

Submitted on 21 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Simulation of RNA sequencing with Oxford Nanopore Technologies

Camille Marchet¹ and Leandro Lima²

¹Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

²Laboratoire de Biométrie et Biologie Evolutive (LBBE), CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA 69622 Villeurbanne, France

Introduction

Until recently, transcriptomics applications with long reads were realized with Pacific Biosciences' Iso-seq protocol. Pioneer works start dealing with **characterization of isoforms or gene expression quantification using Oxford Nanopore Technologies (ONT) reads**.

Despite the existence of long reads simulators for genomic data sets, **a current lack is the possibility to adequately simulate long reads from ONT RNA protocols**, which would help with the developments of new tools to handle this kind of data. The simulation of transcriptomics sequencing is a more complex task than in genomics because the **gene expression and transcript variability have to be modeled**.

Methods

Error rates and profiles

- learn error rates and profiles by training using real read data sets such as Nanosim [1]
- compute error rates and percentages of deletion, insertion and substitution, as well as homopolymer errors using AlignQC [2]
- pre-computed error profiles can also be input

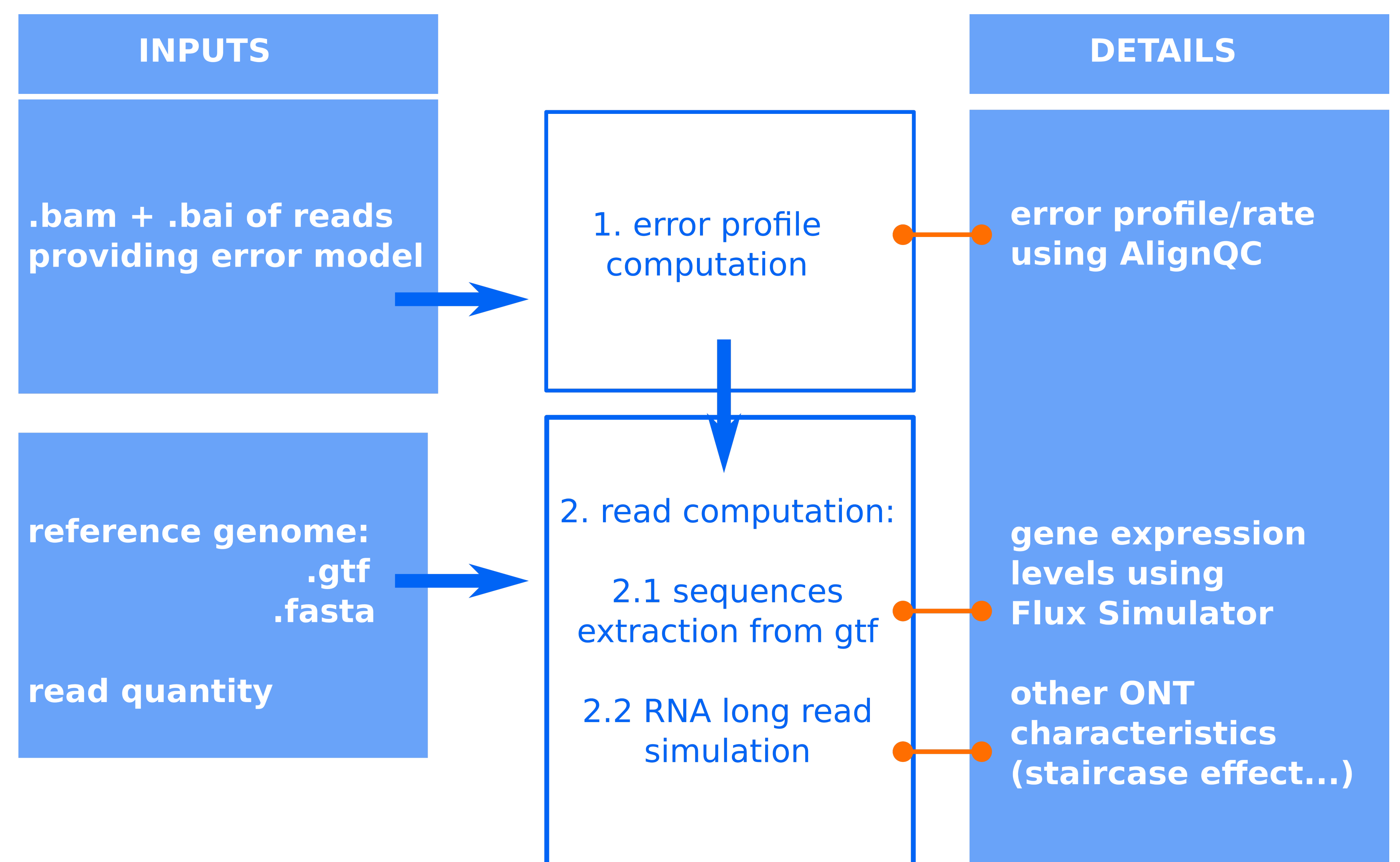
Simulate transcripts

- extract template transcripts from the GTF file of the desired reference using gffreads (<http://ccb.jhu.edu/software/stringtie/gff.shtml>)
- extract realistic expression levels from Flux Simulator [3] results

Integrate errors and transcripts in synthetic reads

- novel implementation in C++ that generates the final reads
- adds the errors to the sequences extracted from the GTF
- deals with regular versus homopolymer errors
- adds supplementary characteristics such as the staircase effect

Pipeline



Experiment

- 300k reads simulation
- Genome reference: *Mus musculus* GRCm38
- Annotation: Ensembl *Mus musculus* GRCm38 release 87
- Template dataset for error training: 740k *Mus musculus* RNA ONT 1D reads
- Comparison of raw reads and simulated reads statistics using AlignQC [2]

Results

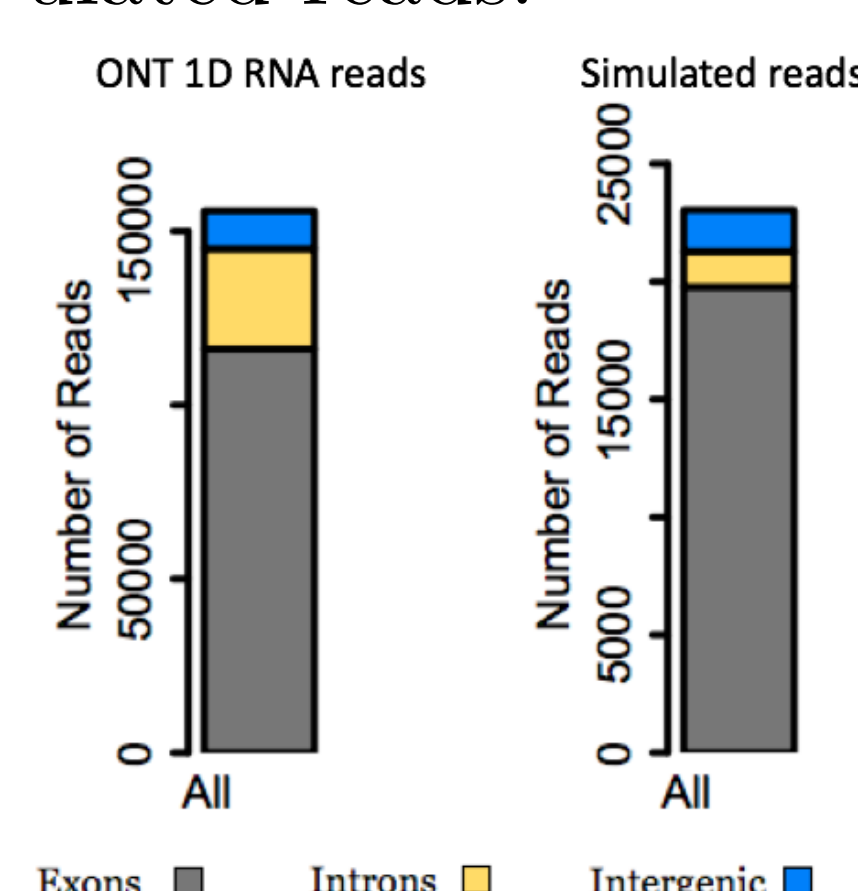
Comparing error rates between ONT 1D RNA reads and the simulated reads shows that the simulator produces reads with realistic error profiles:

Metric	ONT 1D RNA reads	Simulated reads
ERROR RATE	13.697%	11.592%
MISMATCHES	5.089%	5.848%
DELETION	7.348%	4.85%
INSERTION	1.26%	0.894%
NON HOMOPOLYMER DELETION	4.397%	3.069%
HOMOPOLYMER DELETION	2.951%	1.782%
NON HOMOPOLYMER INSERTION	0.874%	0.558%
HOMOPOLYMER INSERTION	0.387%	0.337%

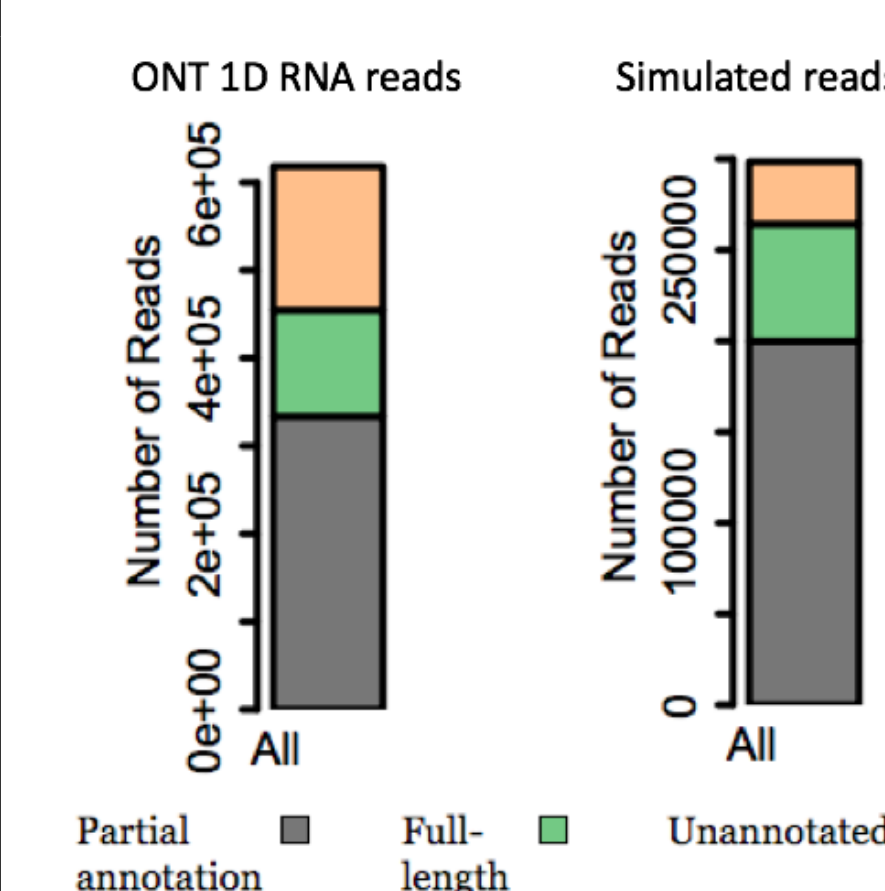
References

- [1] Chen Yang, Justin Chu, René L Warren, and Inanç Birol. Nanosim: nanopore sequence read simulator based on statistical characterization. *GigaScience*, 6(4):1–6, 2017.
- [2] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. *F1000Research*, 6, 2017.
- [3] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic acids research*, 40(20):10073–10083, 2012.

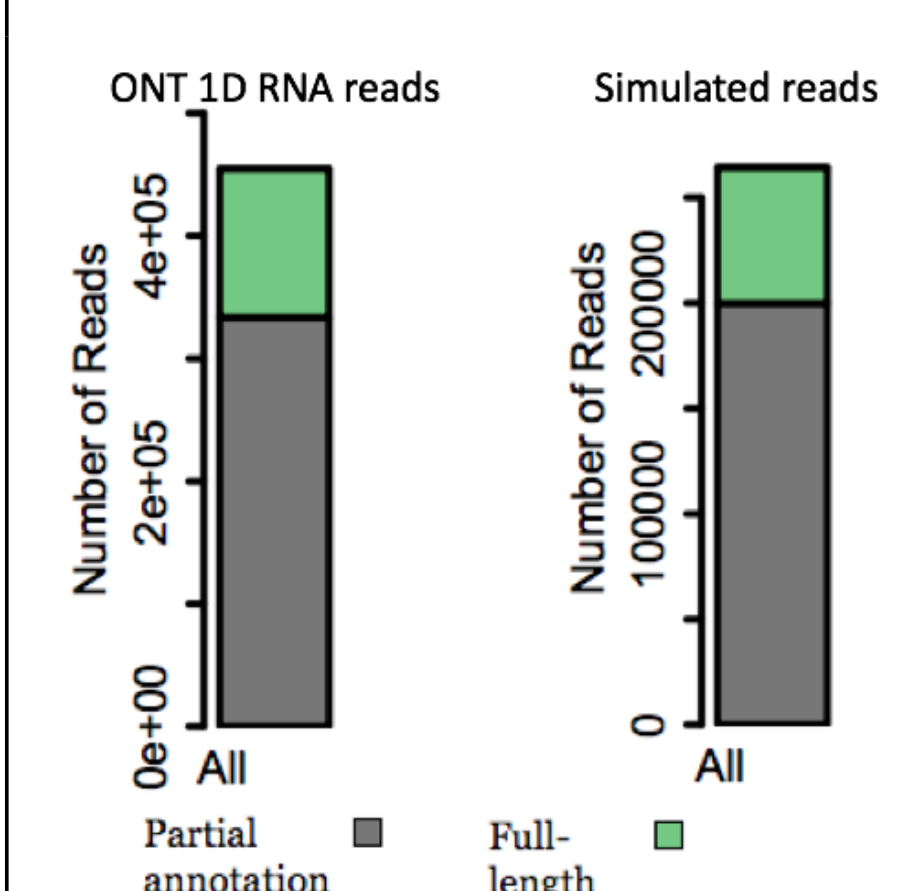
The distribution of reads among genomic features is similar between the real and simulated reads:



As well as the distribution of annotated reads:



And the distribution of identified reference transcripts:



Conclusion

Main messages

- Work in progress, **first tool** to simulate transcriptome sequencing with ONT
- Availability: github.com/kamimrcht/RNA-long-reads-Simulator

Acknowledgments

- ASTER ANR - DS0705
- Genoscope
- Genouest
- JOBIM
- CNPq/CsF grant nb 203362/2014-4