Simulation of RNA sequencing with Oxford Nanopore Technologies
Camille Marchet, Leandro Lima

To cite this version:

HAL Id: hal-01929917
https://hal.archives-ouvertes.fr/hal-01929917
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 Marchet\(^1\) and Leandro Lima\(^2\)

\(^1\)Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France
\(^2\)Laboratoire 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

- **INPUTS**
  - .bam + .bai of reads providing error model
  - reference genome: .gtf
  - .fasta
  - read quantity
- **DETAILS**
  - error profile/rate using AlignQC
  - gene expression levels using Flux Simulator
  - other ONT characteristics (staircase effect...)

1. **error profile computation**
2. **read computation:**
   1.1 sequences extraction from gtf
   2.2 RNA long read simulation

#### 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:

<table>
<thead>
<tr>
<th>Metric</th>
<th>ONT 1D RNA reads</th>
<th>Simulated reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR RATE</td>
<td>13.697%</td>
<td>11.592%</td>
</tr>
<tr>
<td>MISMATCHES</td>
<td>5.089%</td>
<td>5.848%</td>
</tr>
<tr>
<td>DELETION</td>
<td>7.348%</td>
<td>4.85%</td>
</tr>
<tr>
<td>INSERTION</td>
<td>1.26%</td>
<td>0.894%</td>
</tr>
<tr>
<td>NON HOMOPOLYMER DELETION</td>
<td>4.397%</td>
<td>3.069%</td>
</tr>
<tr>
<td>HOMOPOLYMER DELETION</td>
<td>2.951%</td>
<td>1.782%</td>
</tr>
<tr>
<td>NON HOMOPOLYMER INSERTION</td>
<td>0.874%</td>
<td>0.558%</td>
</tr>
<tr>
<td>HOMOPOLYMER INSERTION</td>
<td>0.387%</td>
<td>0.337%</td>
</tr>
</tbody>
</table>

**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:**

#### References


#### 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
- Genomenet
- JOBIM
- CNPq/CsF grant nb 203362/2014-4