

Evaluation of Oxford Nanopore Sequencing data classification beyond canonical nucleotides

Bachelor Thesis - Data Science and Artificial Intelligence



Plan

1. Introduction

2. Description of the technology

3. RQ1

- Experiment
- Results
- Conclusion

4. RQ2

- Experiment
- Results
- Conclusion



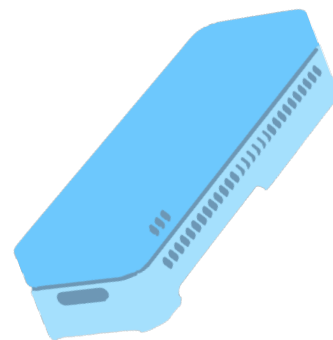


Problem statement



Context:

Sequence RNA or DNA to discover interesting genes that code for a disease for instance



State-of-the-art:

Nanopore Sequencing developed by Oxford Nanopore Technologies (ONT)



Goals:

1. Test the accessibility of data and reusability of the nano-ID software
2. Evaluate the NN's performance to detect the presence of RNA isoforms

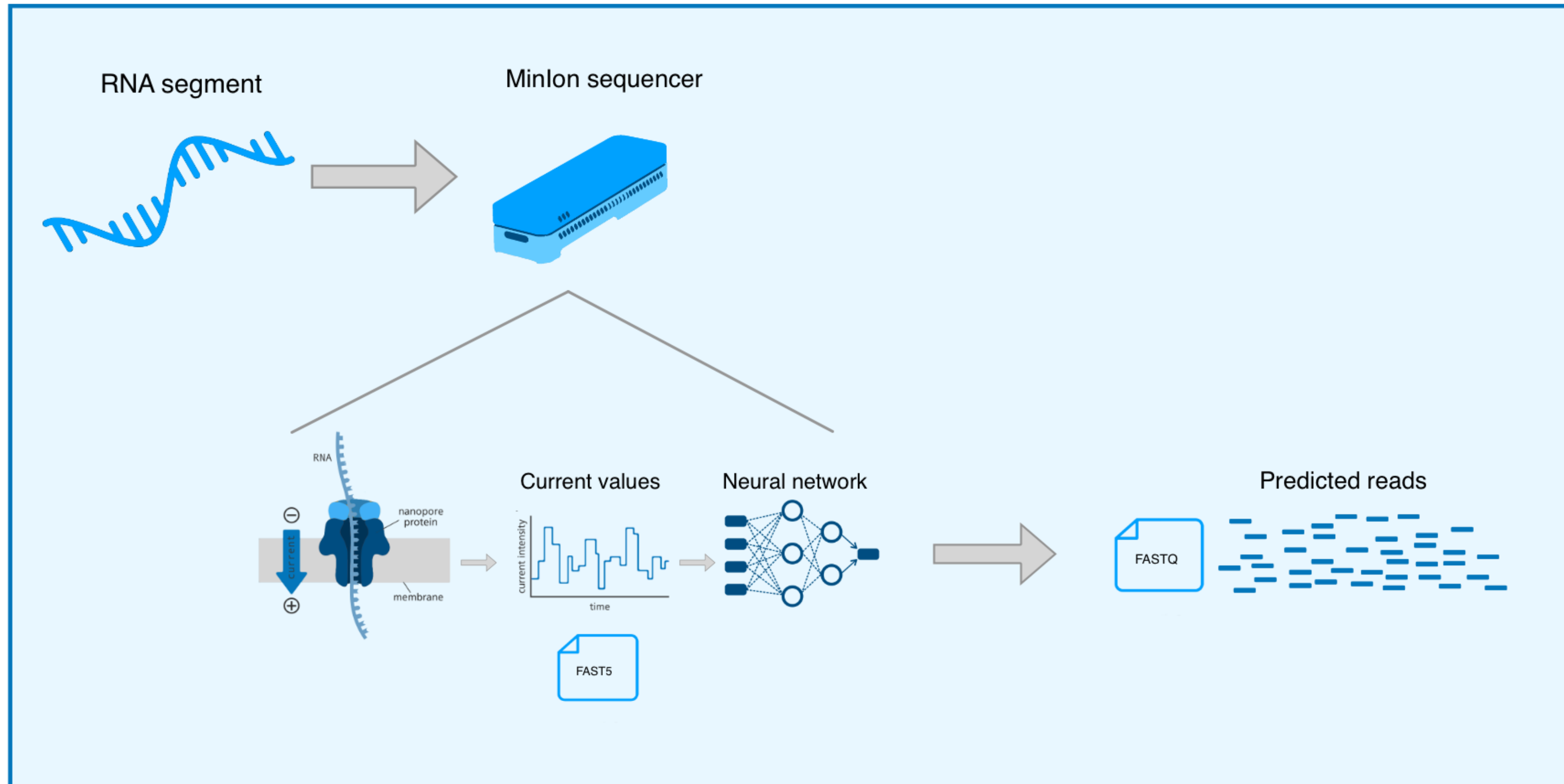
RNA isoforms ?
modified nucleotides
e.g.: Uracil->
5-Ethynyl Uracil

**Native molecule sequencing by
nano-ID reveals synthesis and
stability of RNA isoforms**

© 2020 Maier et al.; Published by Cold Spring Harbor
Laboratory Press




The technology - nanopore sequencing





RQ1

- RQ1 (a) Can the data used in Maier et al.'s research be called 'FAIR' ?
- RQ1 (b) If yes, can the results be reproduced when running the nano-ID neural network over these nanopore data, meaning that this research respects the principles of **Open Science** and **Fair Data** ?



Hardware and software should be **accessible** and **reusable**



Data should be :

- **Findable**
- **Accessible**
- **Interoperable**
- **Reusable**

= 'FAIR'



RQ1 (a) - Are data 'FAIR' ?

nano-ID fast5 files

Version 1.0



Schwalb, Bjoern, 2020, "nano-ID fast5 files", <https://doi.org/10.25625/XNSXV6>, GRO.data, V1

[Cite Dataset](#)

[Learn about Data Citation Standards.](#)

Access Dataset ▾

Contact
Owner

Share

Dataset Metrics ?

0 Views ?

0 Downloads ?

0 Citations ?

Description ?

This dataset contains raw data files for nano-ID. Nanopore sequencing-based Isoform Dynamics (nano-ID), a method that detects newly synthesized RNA isoforms and monitors isoform metabolism. nano-ID combines metabolic RNA labeling, long-read nanopore sequencing of native RNA molecules and machine learning. (2020-01-20)

Subject ?

Medicine, Health and Life Sciences

License/Data Use Agreement



CC0 1.0

Files

Metadata

Terms

Versions

Change View

Table

Tree

Search this dataset...



Filter by

File Type: All ▾

Access: All ▾

File Tag: All ▾

Sort ▾

1 to 10 of 20 Files

Download



20170608_1226_unlabeled_run.tar.gz

nano-ID fast5 files/
Gzip Archive - 44.8 GB
Published 26 févr. 2020
59 Downloads
MD5: 6a4...5e8
Synthetic RNA unlabeled (synthetic control, nano-ID fast5 files).

Data



20170609_1149_labeled_run.tar.gz

nano-ID fast5 files/
Gzip Archive - 2.1 GB
Published 26 févr. 2020
30 Downloads
MD5: 6be...571
Synthetic RNA labeled with 5BrU, 4SU, 6SG (5BrU, 4SU, 6SG, nano-ID fast5 files).

Data



20170912_1101_alternative_run.tar.gz

nano-ID fast5 files/
Gzip Archive - 13.8 GB
Published 26 févr. 2020



Findable ✓

Accessible ✗

Interoperable ✗

Reusable ✗

≠ 'FAIR'

➤ stable connection

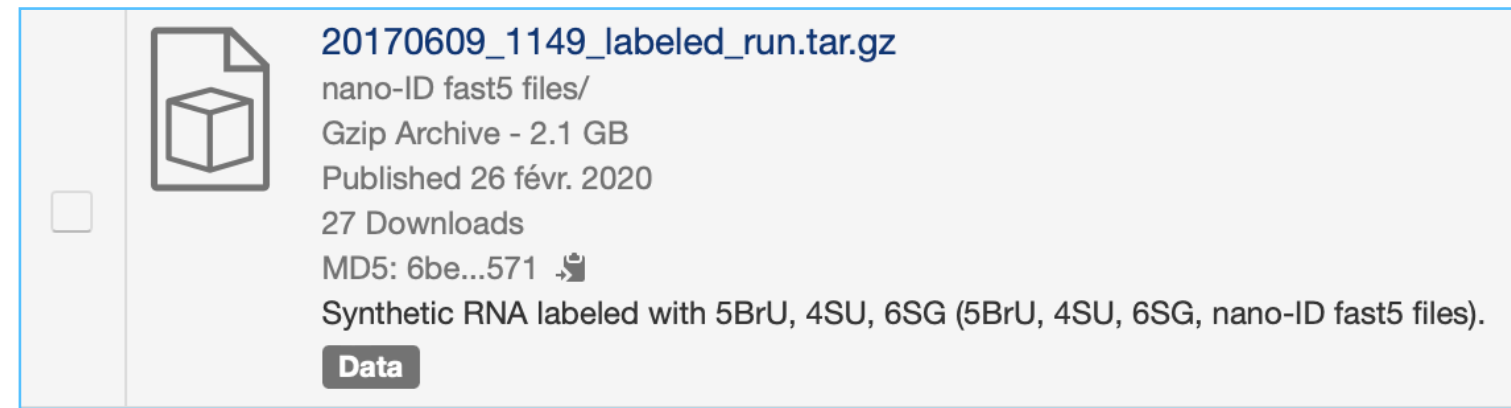
➤ wget

➤ powerful computer

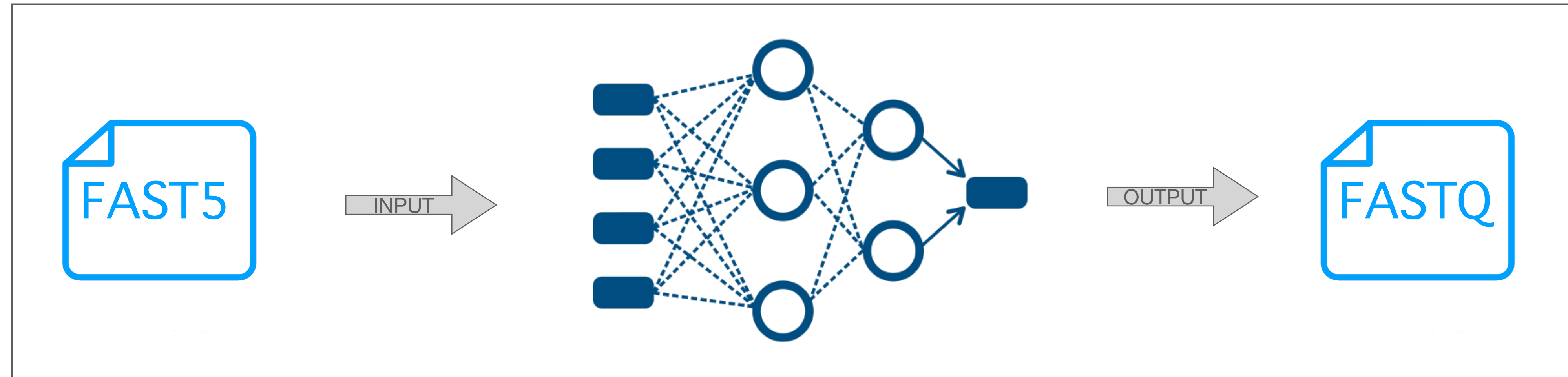


RQ1 (b) - Can the neural network be run ?

current values

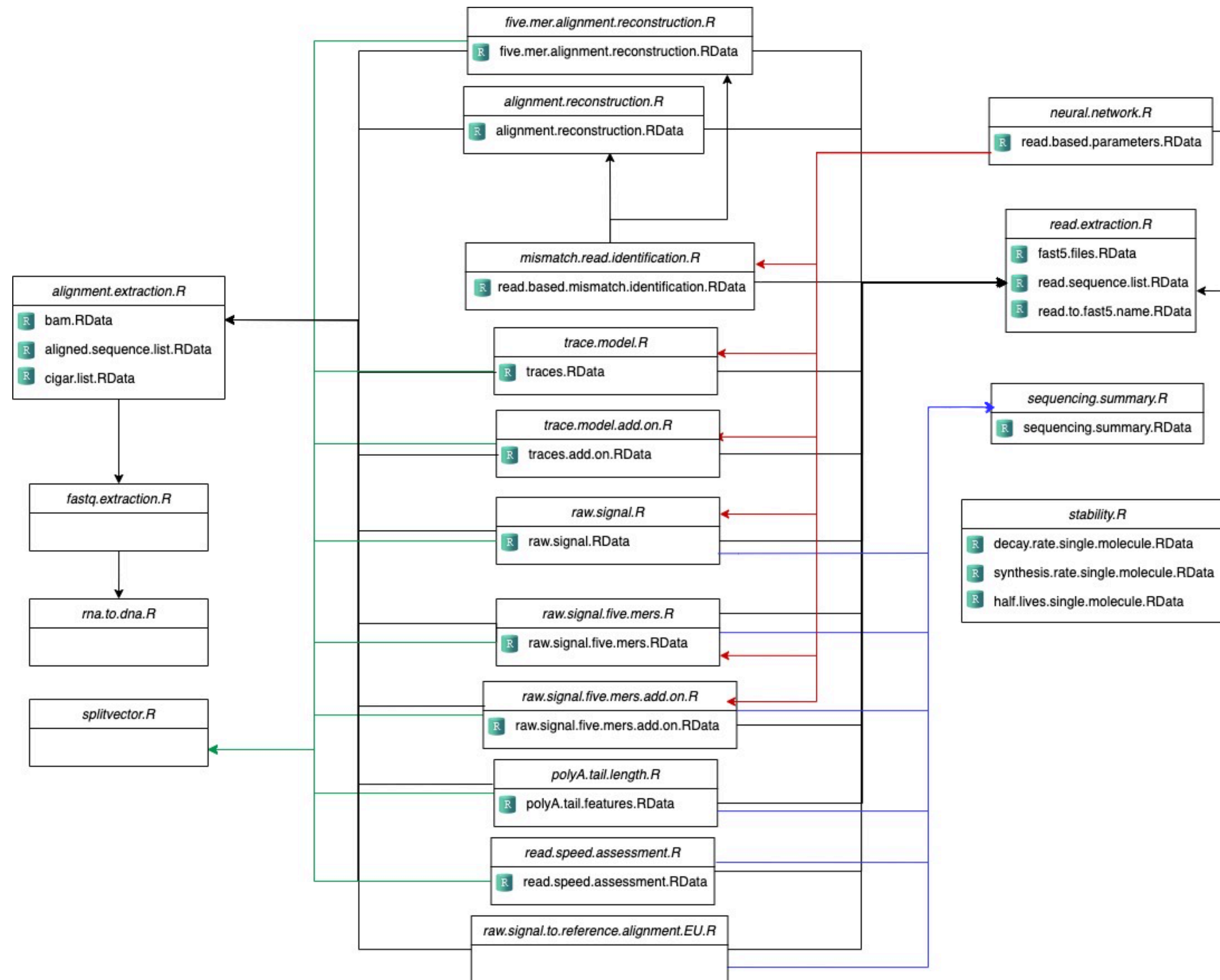


predicted reads





RQ1 (b) - Can the neural network be run ?



➤ poorly documented

➤ NN cannot be run



RQ1

- **RQ1 (a)** Can the data used in Maier et al.'s research be called 'FAIR' ?
 - **RQ1 (b)** If yes, can the results be reproduced when running the nano-ID neural network over these nanopore data, meaning that this research respects the principles of Open Science and Fair Data ?
-
- **RQ1 (a)** The data in Maier et al.'s research are not 'FAIR'.
 - **RQ1 (b)** The results of the neural network cannot be reproduced and thus the overall research does not respect the principles of Fair Data and Open Science.

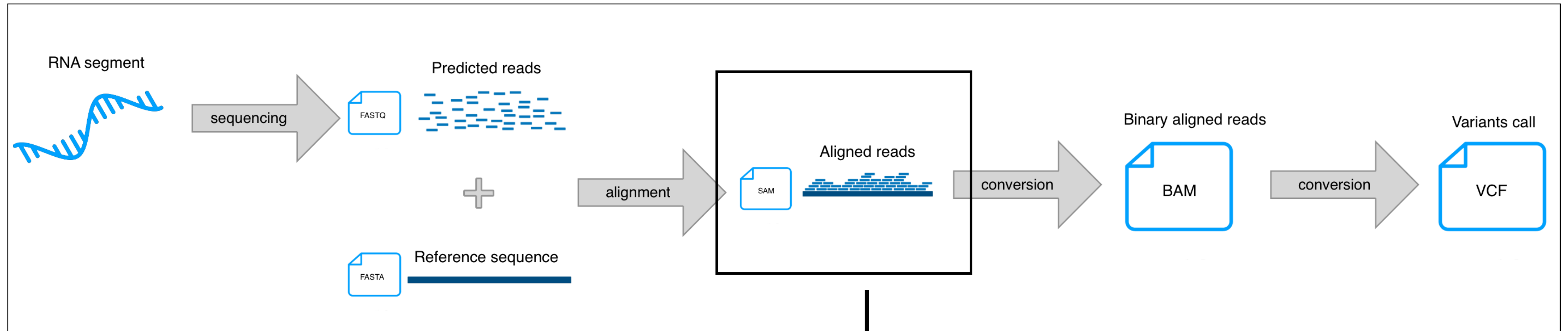


RQ2

- RQ2 What is the performance of the nano-ID neural network in terms of misclassification ? Does it enable one to detect RNA isoforms ?



How do we evaluate the performance ?



Interactive tool
for visualization

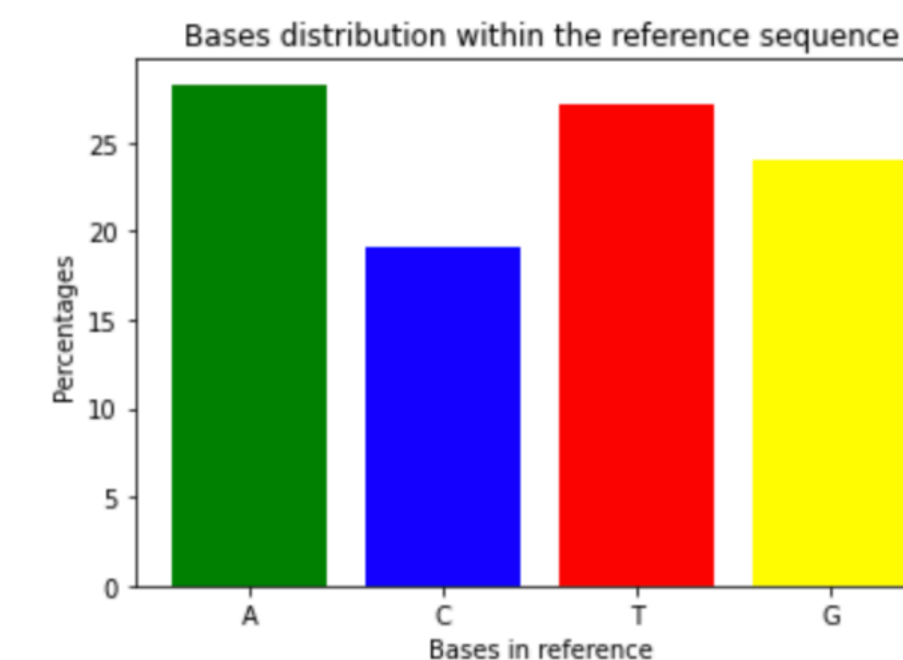


Results

synthetic RNA
labeled with 5BrU,
4SU and 6SG

Complete
alignment
info

PRED \ REF	REF			
	A	C	T	G
A	0.94	0.01	0.03	0.03
C	0.01	0.93	0.09	0.01
T	0.02	0.05	0.85	0.01
G	0.02	0.01	0.02	0.95



only
variants

VARIANTS	A	C	T	G
A	/	0	0.05	1
C	0.14	/	0.92	0
T	0	0	/	0
G	0.86	1	0.03	/
proportion in REF	0.14	0.04	0.74	0.06

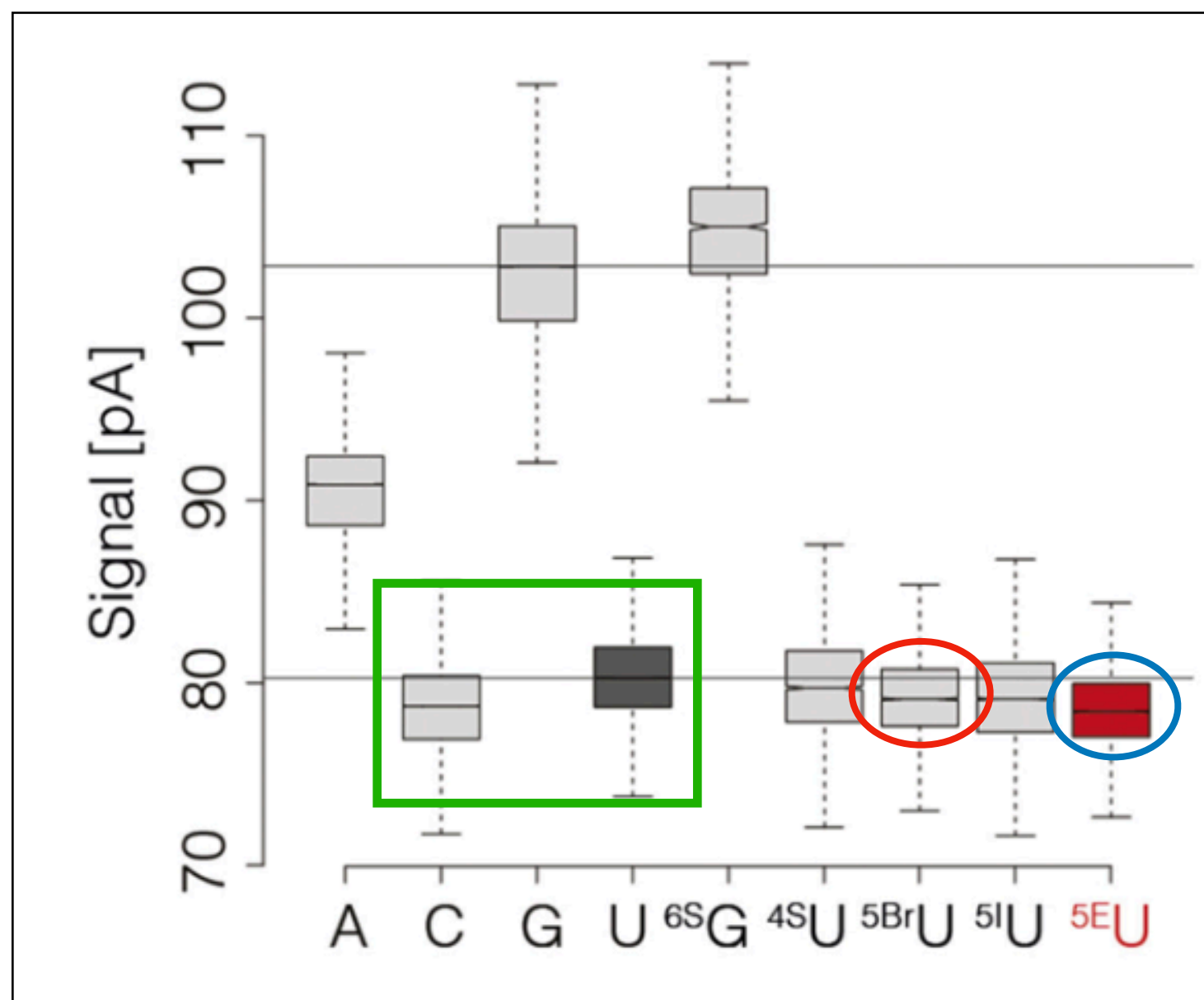
human K562 cells
cultured in the
presence of 5EU for
60min

only
variants

VARIANTS	A	C	T	G
A	/	0.60	0.71	0.82
C	0.11	/	0.22	0.05
T	0.45	0.35	/	0.12
G	0.43	0.05	0.06	/
proportion in REF	0.17	0.24	0.35	0.24



Results



<div>PRED \ REF</div>	A	C	T	G
A	0.94	0.01	0.03	0.03
C	0.01	0.93	0.09	0.01
T	0.02	0.05	0.85	0.01
G	0.02	0.01	0.02	0.95

VARIANTS	A	C	T	G
A	/	0	0.05	1
C	0.14	/	0.92	0
T	0	0	/	0
G	0.86	1	0.03	/
proportion in REF	0.14	0.04	0.74	0.06

VARIANTS	A	C	T	G
A	/	0.60	0.71	0.82
C	0.11	/	0.22	0.05
T	0.45	0.35	/	0.12
G	0.43	0.05	0.06	/
proportion in REF	0.17	0.24	0.35	0.24



RQ2

- **RQ2** What is the performance of the nano-ID neural network in terms of misclassification ? Does it enable one to detect RNA isoforms ?

- **RQ2** The nano-ID neural network seems to perform well enough in order to detect the presence of 5BrU and 5EU despite the reduced amount of conducted experiments.



Thank you for your attention !