Evaluation of Oxford Nanopore Sequencing data classification beyond canonical nucleotides

Adèle Imparato
Department of Data Science and Knowledge Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—Nanopore sequencing is a technology that allows to sequence DNA/RNA sequences. Currently, the detection of modified RNA bases is poorly characterized. This report therefore focuses on a published paper titled "Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms", which presents a tool, nano-ID, allowing to detect one specific RNA isoform: ^{5E}U. This research concentrates on integrity of their data and the reusability of their code, which proved to be neither 'FAIR', nor reusable. Moreover, it aims to evaluate the performance of the neural network developed in the nano-ID tool, showing that it is possible to detect RNA isoforms such as ^{5E}U.**

***Index Terms*—nanopore sequencing, Oxford Nanopore Technologies, nano-ID, RNA-Seq, RNA isoforms, neural network, Open Science, FAIR Data Principles**

I. Introduction

A. Context and history

Sequencing DNA/RNA is the task consisting of determining the order of the bases (A, C, G and T/U) that make up the DNA/RNA molecule. Ever since the structure of DNA was discovered in 1953 [4], scientists aim to sequence DNA for the purpose of detecting novel interesting features in the sequence such as genes coding for a disease [17]. Today, various technologies exist for sequencing DNA/RNA. A trivial example is the four-color fluorescent Sanger sequencing, which consists of reading a sequence using fluorescent markers where each color corresponds to one nucleotide. This technology was created by Frederik Sanger in 1977 and is an example of first-generation sequencing technique. In 1996, second generation sequencing techniques emerged lead by pyrosequencing [3] and dominated by Illumina's sequencer. Subsequently, the third generation arose with the purpose to reduce the sequencing cost. Three major technologies of this generation dominate the market: Pacific Biosciences (PacBio), Complete Genomics and Oxford Nanopore Technology. [7]

B. Object of focus

This research paper focuses on one of the third generation sequencing technologies: Nanopore Sequencing, developed by Oxford Nanopore Technologies (ONT) and current state-of-the-art (https://nanoporetech.com). This technology is based on nanopores, tiny holes through which molecules will pass, generating an electric current that can then be interpreted to determine the bases that compose the DNA/RNA segment. The advantages of this technology is that it offers relatively low-cost genotyping, high mobility for testing, and rapid processing of samples with the ability to display results in real-time. These advantages allow for a wider range of applications than former sequencing techniques. However, many of these applications require the development of fit to purpose advanced data analysis pipelines and AI models, beyond traditional bioinformatics [24]. One of the main advantages of Nanopore Sequencing is the facultative conversion to DNA when targeting RNA molecules [9]. Unlike in usual RNA sequencing technologies, there is no need for amplification and reverse transcription to convert the sequence into DNA. That is why one talks about direct RNA nanopore sequencing, which facilitates and unbiases the experimental sample processing as it has one less step, while allowing for a more detailed analysis, for instance, by allowing the detection of modified nucleotides beyond the canonical four [21] [26].

The synthesis and stability of these RNA isoforms is currently poorly characterized, due to the lack of analytical methods that detect and quantify RNA metabolism by using short-read sequencing or other methodologies. Thus, scientists were incapable of detecting RNA isoforms. With ONT, the detection of modified nucleotides in RNA molecules will have a great impact on RNA biology, directly influencing the studies of eukaryotic, bacterial, and viral organisms as well as the characterization of synthetic RNA [18].

C. Problem statement

In 2011, Cold Spring Harbor Laboratory Press published an article titled "Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms" [6]. The article introduces nanopore sequencing-based Isoform Dynamics (nano-ID), a software that detects newly synthesized RNA isoforms and monitors isoform metabolism using a neural network (NN). In order to validate and critique Maier et al.'s work, it is essential to be able to reproduce the experiments carried out. There exist two sets of principles that can be used to verify the accessibility and reusability of data and software developed within the context of research:

1) the FAIR Data Principles, which act as guidelines for strengthening the reusability of data [25];
2) and Open Science, which acts as guidelines for making software code reproducible and more sustainable [23].

The FAIR Data Principles state that data should be Findable, Accessible, Interoperable and Reusable in order to be called

'FAIR'. These guidelines therefore represent a step towards optimizing the reuse of data such that both humans and machines can retrieve and manipulate the data (see [25] for more information). As for Open Science, it states that hardware and software developed in a research context should be accessible so that it can be reused by other researchers for further work. This approach is intended to promote scientific collaboration. Open Science entails other aspects that are beyond the scope of this study.

This research paper therefore aims to discover whether the data used in Maier et al.'s study are accessible. If so, to test and run the neural network developed through the nano-ID tool in order to verify the reproducibility of results. Subsequently, it intends to evaluate the impact of RNA isoforms on the neural network's performance. Based on these objectives, one can formulate two research questions:

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

Firstly, this report includes a description of the sequencing technology used (i.e., nanopore sequencing), as well as the data used to conduct this study. In addition, the nano-ID neural network is described. Second, the experiments carried out in this research are described, together with the results and discussion interpreting these. Eventually, a conclusion will answer the previously mentioned research questions, as well as the following two extra sections: Limitation & Further Improvement and Acknowledgments.

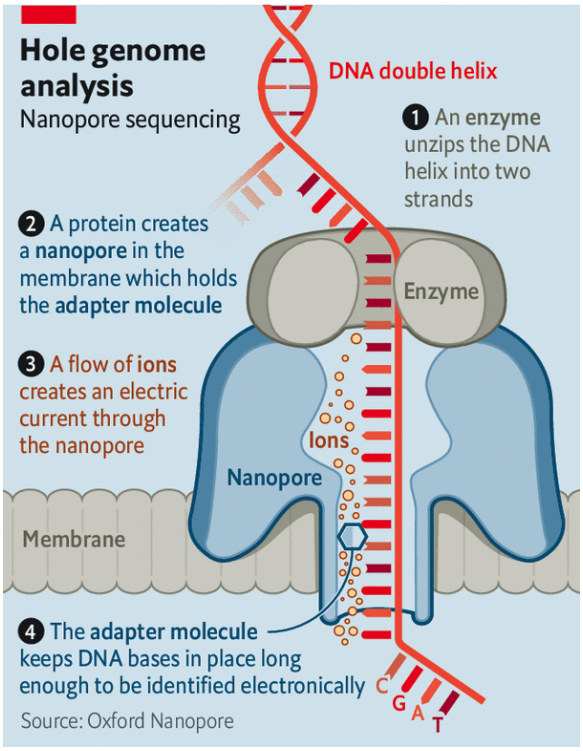


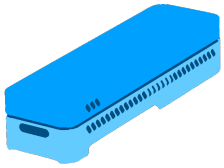
Fig. 1: How nanopore sequencing works. [5]

According to ONT [22], the technology described above has the advantage of being the only one enabling direct, real-time analysis of short to ultra-long extracts of DNA/RNA in scalable formats. One of the advantage of long-reads (between 5000 and 30000 bases long) over short-reads (between 50 to 350 bases long) is that they allow to get a high-quality alignment with significantly fewer gaps. Nanopore sequencing's numerous advantages allow for instance to classify brain tumors based on modified bases induced by DNA methylation [2]. For this particular example, a MinIon device was used (see Figure 2(a)).

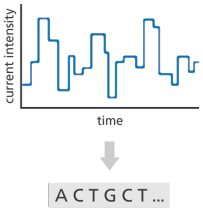
II. Methods

A. The technology

The technology used in this study is nanopore sequencing. This high-throughput sequencing technology consists of a device (here, MinIon [20]) able to decrypt the nucleotides that compose a DNA/RNA extract by means of protein pores (see Figure 1). Sequences are injected into the sequencer, pass through the nanopores and generate an electric current that is directly influenced by the bases. This electric current can then be measured and interpreted by the neural network aiming to predict which base generated which fluctuation, eventually determining the sequence's composition (see Figure 2(b)).



(a) MinIon sequencer. Pocket-sized nanopore sequencing USB device that can contain up to 521 channels (or pores) and is able to generate short to ultra-long reads (>4 Mb). [14]



(b) The sequencer generates an electric current directly influenced by the sequence's composition. This current is interpreted by the neural network so that it predicts the bases. This image was adapted from [11]

Fig. 2

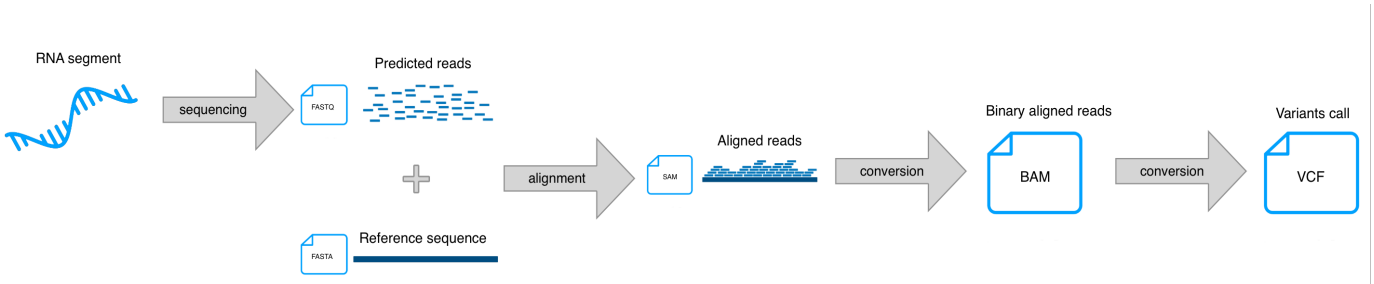


Fig. 3: Summary of the overall process from an RNA sequence to the VCF targeted format. Image components were taken from [11].

The nanopore approach has not been fully explored yet and is still in development. The company business model is to develop the product and software in a agile way including the feedback of costumers and having them sometimes lead the development. For this reason, it is state-of-the-art and has a great potential to make future fourth-generation technologies.

B. The data

The paper provides 20 datasets available for download to the Göttingen Research Online Data Archive (GRO) under the DOI https://doi.org/10.25625/XNSXV6. The dataset used for the first experiment is synthetic RNA labeled with 5-bromouridine (^{5Br}U), ^{4s}U and 6-thioguanine (^{6s}G), three different RNA isoforms. This dataset being synthetic (it does not come from animal molecule), its goal is to investigate the performance of nano-ID at detecting RNA isoforms, in this specific case, ^{5Br}U, ^{4s}U and ^{6s}G. However, Maier et al. observed that thiol-based analogs, ^{4s}U and ^{6s}G, were more difficult to incorporate into synthetic RNA during in vitro transcription (IVT). That is to say that only few of these modified nucleotides are present in this RNA segment. Moreover, they noticed that ^{5Br}U was less easily recognizable compared to ^{5E}U (5-Ethynyluridine) and ^{5l}U, two other isoforms tested in a different synthetic dataset. This leads us to the assumption that the only modified nucleotide that can possibly be detected in this specific dataset is ^{5Br}U, although less easily detectable compared to ^{5E}U.

The dataset contains a large amount of reads in the form of FAST5 files, which are classified into three folders: fail, pass and skip. It should be noted that when a sequence passes through the flow cell, it is translated in current values in the form of long-reads - hence the reads inside FAST5 files. These files are to be used as input for the neural net as it can be seen on Figure 4. On the other side, the output of this one is a single FASTQ file (also provided by the dataset) containing the predicted nucleotides in the form of reads.

FAST5 file format is a type of format that is specific to ONT as it is the file type generated by the sequencer. Indeed, the raw signals that are generated by the sequences passing through the nanopores, measured in pico-amp, are stored in FAST5 files as shown in Figure 4. This format, which is a specification over a HDF5 file, enables to structure the data in a folder-like structure. This way of storing data is supported by numerous common programming languages but however cannot be read using a simple text editor. For more information, please refer to the following article: A Look at the Nanopore fast5 Format, Shian Su [19]. In contrast, FASTQ format is easier to read as it is text-based and therefore it can be opened using a regular text editor. The file consists of several nucleotide sequences (i.e. the predictions) and their corresponding quality score encrypted in ASCII code, which represents the confidence level of each predicted base.

To address the second research question, the required files are VCF files (Variant Call Format), which correspond to a reorganisation of the result of the alignment between reads and their respective reference sequence, in tabular form. Here, each sequenced segment of DNA/RNA has a reference sequence (refseq) used to compare predictions with the actual composition of that segment. This VCF type of format enables to highlight key features of the alignment such as the amount of predicted bases for each position and the distributions of these. Moreover, there is a way filter the VCF file such that the remaining rows of the table correspond only to variants (i.e., the positions where the global prediction does not correspond to the refseq). For this second experiment, one will therefore use complete VCF files and VCF files containing only the variants. See Figures 11, 12 and 13 in the Appendix to get a better understanding of the concept of alignment.

In addition to the previously mentioned synthetic dataset, the second experiment makes use of the file *20180226_1208_K562_5EU_60_labeled_run.bam* available at https://www3.mpibpc.mpg.de/downloads/cramer/illuMinatION/. This file encodes the alignment information of a sample highlighting the presence of the ^{5E}U isoform in K562 human cells[1]. It corresponds to one of the six biological replicates of K652 cells cultured in the presence of ^{5E}U for 60min (^{5E}U 60). See [6] to learn about how these cells were cultured. In fact, ^{5E}U was chosen for more detailed analysis as it was demonstrated to be well suited for nanopore sequencing. For this reason, one uses this dataset in order to compare the level of detectability of ^{5Br}U and ^{5E}U by nano-ID.

[1]Leukemic cells suit for experimentation

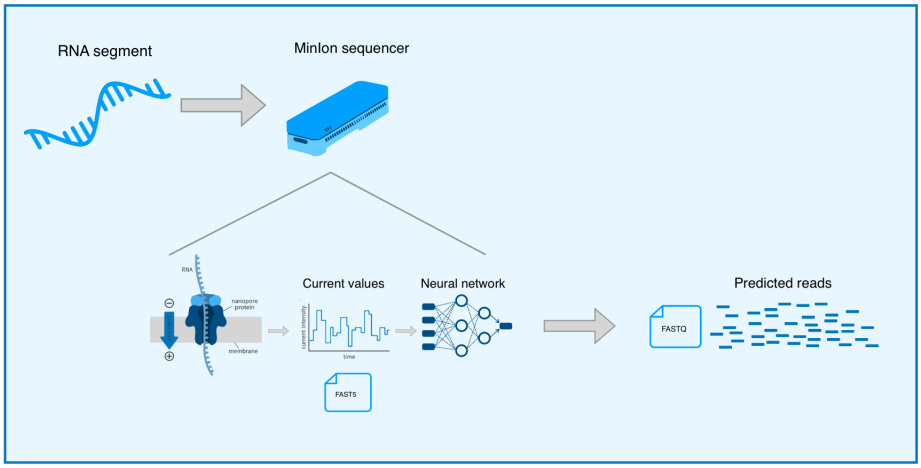


Fig. 4: Summary of the sequencing process.
The RNA segment is incorporated into the MinIon sequencer. This last reads the sequence by means of nanopores generating an electric current encoded as FAST5 files. These files are taken as input for the neural network, which allows to get the predicted reads under FASTQ format. Image components were taken from [11].

To summarize, the files used for this second experiment correspond to:

1) the previously mentioned synthetic dataset mapped with its refseq and then converted into two VCF files (one that is complete and the second one containing only the variants);
2) and the alignment information of the ^{5E}U 60 sample converted in one VCF file containing the variants only (see Limitation & Further Improvement).

1) Preprocessing: Since the targeted files are VCF files, files that are not directly provided, one needs to make the conversion from FASTQ or BAM to VCF. To do so, the reference sequence is required. For the FASTQ file contained in the dataset synthetic RNA labeled with ^{5Br}U, ^{4s}U and ^{6s}G, the reference sequence comes from ERCC RNA spike-in mix (See [6], Supplemental Material, Supplemental Table 3). For the BAM file, the reference sequence is hg38 (UCSC refseq GRCh38), assembly of the human genome released in December 2013 [10] (accessible through [1]. Subsequently, in order to obtain the VCF file, the following steps need to be performed:

1) Map the FASTQ file with its reference sequence using Minimap [16];
2) Convert the output under SAM format into a BAM file using Samtools [12];
3) Sort the BAM file using Samtools;
4) Convert the BAM file into a complete VCF file using Bcftools [13].
5) or Convert the BAM file into a BCF file and then call the variants and encode them in a VCF file using Bcftools [13].

In Figure 3 in the Appendix, the commands used to download the three software and apply these steps are given.

C. The neural network

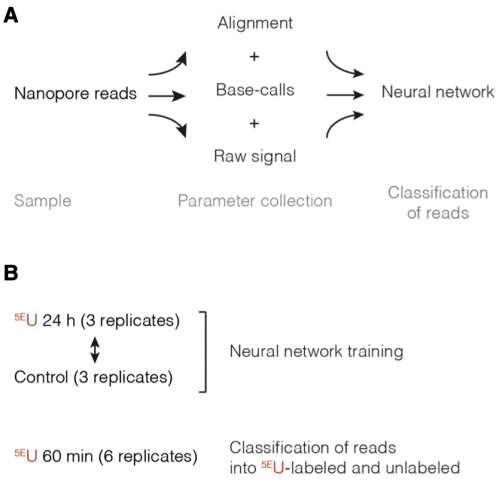


Fig. 5: (A) The parameter collection used for training the nano-ID neural network comes from three layers: the raw signal layer, the base-calls layer, and the alignment layer (mismatch and indel properties). (B) The nano-ID neural network was trained on the ^{5E}U 24 h versus control samples and used to classify reads of the ^{5E}U 60 min samples into ^{5E}U labeled and unlabeled. (This figure was taken from [6], Figure 2 A,B)

As described in [6], the collected data for this study consist of direct RNA nanopore sequencing data taken from human K562 cells. The neural network developed in the nano-ID tool was thus trained on these data, more specifically, on three replicates of cells exposed to ^{5E}U labeling for 24h (^{5E}U 24h) versus three replicates of unlabeled cells (control) as it is assumed the ^{5E}U 24h sample contains only labeled reads while the control sample contains only unlabeled reads.

The neural net classifies the predicted sequences of RNA into newly synthesized (^{5E}U-labeled) or pre-existing (unlabeled) RNA. In Maier et al.'s study, the NN was used to classify reads of the six replicates of cells exposed to ^{5E}U labeling for 60 min (^{5E}U 60 min), showing that ^{5E}U-containing RNA isoforms are detectable and therefore determining if the sample was produced before or during ^{5E}U labeling.

About 4700 input parameters in total were estimated in order to train the neural network. These come from three different layers consisting of one raw signal layer, one base-calls layer and one alignment layer. Respectively, the first one handles the electric current generated by the sequences passing through the pores and has in total 2048 parameters. The second layer, the base-calls one, informs on confidence and alternatives of the identified bases in the sequence. It has 2330 parameters. Eventually, the last layer contains information on mismatch and indel[2] properties derived from the mapping phase (see Figure 6). This layer has 147 parameters. These input parameters were estimated through the initial neural network that is inside the MinIon sequencer. Indeed, the latter already considers the raw signal, base-calls it and aligns it with the refseq. The role of the nano-ID NN however is to take these inputs as benchmark in order to classify the sample into newly-synthesized RNA or pre-existing RNA isoforms.

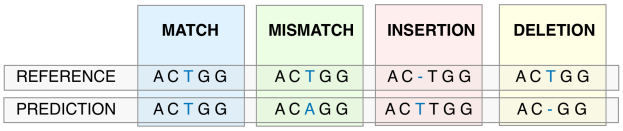


Fig. 6: When a reference sequence is mapped with its prediction, four events might occur: match, mismatch, insertion or deletion. The figure speaks for itself.

Eventually, the trained NN includes eight dense layers and the activation functions that were used are ReLu and Sigmoid. It might be described as a deep feed forward neural network. In order to have more details on its construction, please refer to the previously mentioned paper: Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoform [6].

III. Experiments

A. Accessibility of data and reusability of code

In response to research question 1, the following experiment was conducted. In order to demonstrate the reproducibility of the code and the accessibility of data delivered in Maier et al.'s study, one focuses on two sets of principles previously mentioned: the FAIR Data Principles and Open Science.

In Maier et al's paper, the biological data used and the software at the center of the research (i.e., nano-ID) are provided. One could therefore aim to access these data and then reproduce the results obtained during this study in order to check whether this paper respects the two sets of principles and is fit to further research.

Here, one will only focus on datasets available for download to the GRO under the DOI https://doi.org/10.25625/XNSXV6, as they are the raw data files used as input to run the nano-ID neural network. This experiment therefore consists of aiming to download the previously mentioned files and use them in order to reproduce the results obtained when running the neural net.

Running the neural net requires us to understand the various R files contained in the nano-ID folder, but also, to discover in what order they should be run. In figure 7, the relation between the files is shown, enabling to guess that order.

B. Evaluation of the neural network and detection of RNA isoforms

In order to respond to research question 2, one aims to evaluate the distributions of incorrectly predicted bases by the nano-ID neural network. An incorrectly predicted base may be a hint that an RNA isoform has been detected. This experiment therefore consists of extracting measurements out of the alignment information between sequenced RNA and their corresponding refseq. As described in the Dataset section, the VCF files used enable to easily identify key features of the alignment such as the total allelic depths of high quality bases (AD) but also the amount of variants and their respective alternatives.

In this experiment, 36 ratios were computed in total. On one hand, sixteen ratios corresponding to the average percentage of X bases predicted for a Y position, with X and Y being any pair of canonical bases. These ratios were computed as follow:

$$X\_Y\_ratio = \frac{count\_X}{total\_Y}$$

Where count_X is the total amount of X bases predicted at all Y positions in the refseq and total_Y is the number of Y in the refseq. In addition, four other measures were taken: the percentages of A, C, T and G bases identified in the whole reference sequence. These twenty first metrics were computed on the complete VCF file, that is, the complete alignment information between the synthetic dataset and its refseq. On the other hand, twelve other ratios were computed on the two VCF files containing only the variants, thus for both the synthetic and the human sample. These correspond to the distributions of alternatives for each reference value of variant computed as follow:

$$A\_R\_ratio = \frac{count\_A}{total\_R}$$

Where count_A is the total amount of A alternatives identified at all R reference positions and total_R is the number of variants having R as reference value. In addition, four other percentages were measured indicating the proportion of A/C/T/G among all variants. Eventually, these ratios will allow to highlight unequal distributions over the predicted bases and variants.

[2]Insertion–deletion mutations

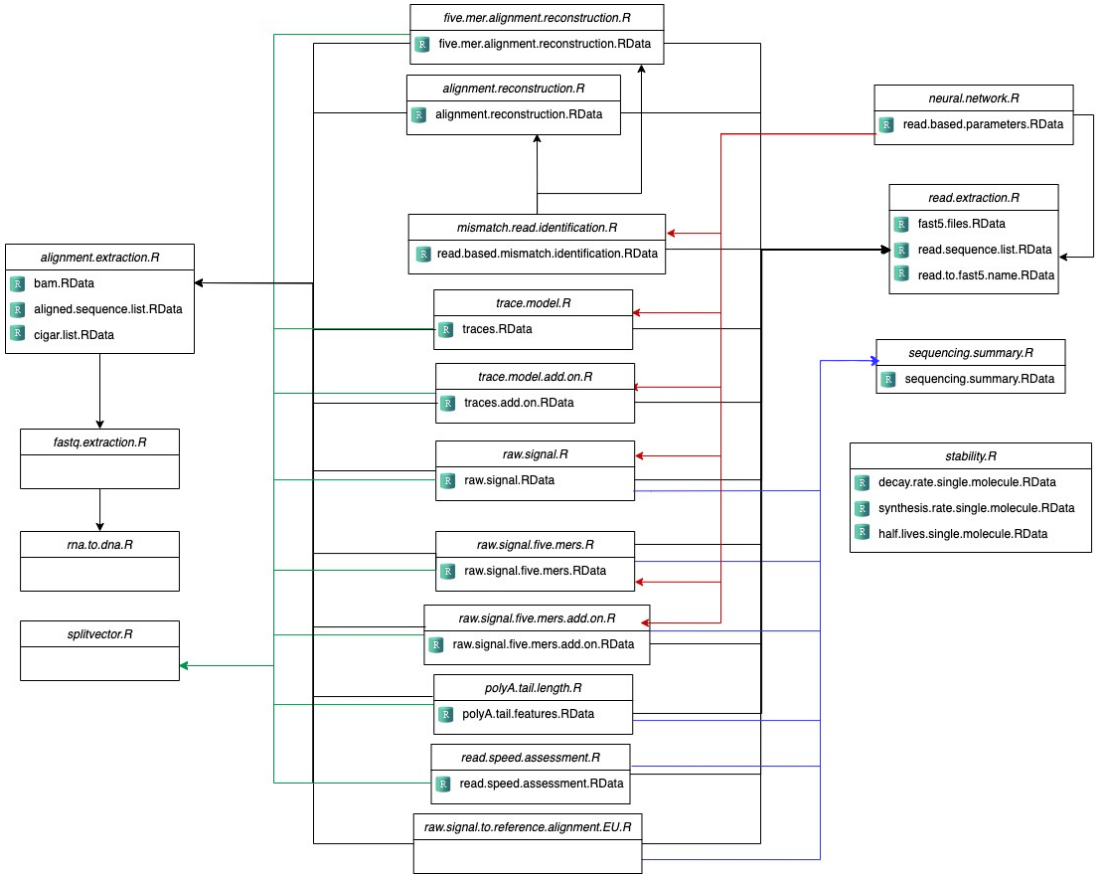


Fig. 7: This figure shows the relations between the different R files of the nano-ID tool. In the boxes, the first line indicates the name of the R code file and the lines below indicate which RData objects are generated in this specific file. The arrows point to the RData objects needed in order to run the source box.

IV. Results

A. Accessibility of data and reusability of code

As mentioned in the Dataset section, Maier et al.'s paper provides twenty datasets available for download to the GRO. However, due to the large size of these files (over 10GB), it is in reality impossible to download them, except for synthetic RNA labeled with ^{5Br}U, ^{4s}U and ^{6s}G which is only about 2GB. While attempting to download the other sets, it always fail sooner or later although the internet connection is stable. The reason behind this is that the downloading is unable to restart where it failed, meaning that if there is any break from the server that provides the data, the download has to restart from the beginning. The webpage also suggests to use Wget [8] in order to avoid interruptions, but the issue remains the same. This one limits us to only being able to download the synthetic RNA dataset mentioned above.

Next, one aims to run the nano-ID neural network using as input the FAST5 files contained in the previously mentioned dataset as it is the only one accessible. The purpose is to get a similar FASTQ output (see Dataset), showing that Maier et al.'s study respects the principle of Open Science, that is, being able to access the nano-ID software and reuse it.

In order to run the neural network, one seeks to run the files contained in the nano-ID folder (https://github.com/birdumbrella/nano-ID) in the correct order, with *neural.network.R* as desired endpoint. From Figure 7, it can be inferred that the first file that needs to be run is *read.extraction.R* since many other files depend on it, in other words, many files require the RData objects created in this piece of code. *Read.extraction.R* reads some FAST5 files, supposedly the input of the neural network. However, it is not indicated which FAST5 files should be taken as input (among the three folders: 'skip', 'fail' and 'pass') in order to get a similar FASTQ file. By making some tests, one can notice that the FAST5 files contained in the 'fail' and 'skip' folder produce an error, suggesting to take the 'pass' folder as input. Then, the next file to be run is *sequencing.summary.R* which simply makes a summary of the sequences' features taken as input by the sequencer. After that, *rna.to.dna.R* needs to be run. This file simply codes for a function that will be then used by *fastq.extraction.R*. *Fastq.extraction.R* is extracting the Fastq content that is inside the FAST5 input files, corresponding to the base-calls provided by the original neural network inside the MinIon device. The next file that can then be run is *alignment.extraction.R*, which, as its name indicates,

extracts the alignment. However, one could ask oneself "what alignment" ? The file takes as input one FASTA file and one BAM file. One might suppose that the Fastq content extracted from the FAST5 files was aligned with its corresponding reference sequence (ERCC spike-in mix here) in order to generate the BAM file. However, for this specific dataset, the given instructions are not explicit enough in such a way that they do not allow to run the next files and eventually reach *neural.network.R*, which supposedly, would be the endpoint. In general, one can notice a lack of comments and instructions in this nano-ID code. Indeed, there is no description of the R files, no input description for each file and no order in which running these files. The code is poorly documented.

B. Evaluation of the neural network and detection of RNA isoforms

For this experiment, it is important to note the values of the current generated by each base, as measured in Maier et al.'s paper. In Figure 8, one can observe the expected ranges of generated current for each nucleotide, including five isoforms.

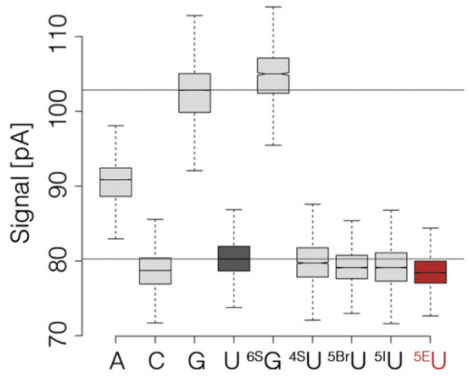


Fig. 8: This figure shows the ranges of current generated by each canonical nucleotide, as well as the five modified nucleotides considered in this study ([6], Figure 1.D).

It can be noticed that both ^{5E}U and^{5Br}U generate an electric current that is the closest to the one generated by U or C[3]. Precisely, the median of the raw signals generated by ^{5Br}U is about 79pA, by ^{5E}U about 78pA, by natural U about 80pA and by C about 78pA. These four medians are relatively really close to each other, making the classification task even harder for the NN. This observation indicates that at each ^{5Br}U or ^{5E}U position in the reference sequence, one expects a lot of U and C predictions and only a few A or G predictions, as their signal is rather far.

[3]N.B.: We are only dealing with one isoform at a time, meaning that the current value cannot be mistaken with the one from another isoform.

PRED \ 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

TABLE I: Summary of the average distributions for each pair of predicted base - reference base. These measures were taken from the complete VCF file corresponding to the alignment information between the synthetic RNA dataset and its refseq.

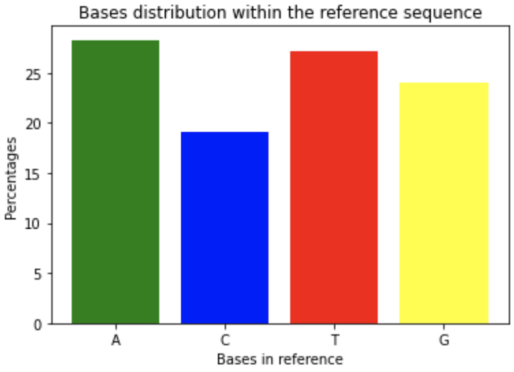


Fig. 9: Proportions of bases in the reference sequence of the synthetic sample.

In Table I, it can be noted that each base in the reference sequence is most of time predicted by itself, e.g. when there is an A in the refseq, 94% of the time, an A will be predicted. The same phenomenon occurs with C, G and T for which the respective percentages are 93%, 95% and 85%, although 85% is proportionately lower. Next, one observes that T is predicted as C in 9% of the cases, which is a relatively high percentage compared to the rest of the table. In addition, on Figure 9 one can observe that the percentages of bases in the reference sequence are roughly fair.

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

TABLE II: Summary of the percentages of alternative base for each possible base in the reference sequence and proportions of T, A, C and G variants among all variants. These measures were taken from the VCF file corresponding to the variants of the alignment between the synthetic RNA dataset and its refseq.

In Table II, it may be observed that the majority of variants, 49 in total, are T bases. Indeed, 37 of them are T's, or 74%. These variants mismatch in 91% of the cases with a C base.

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

TABLE III: Summary of the percentages of alternative base for each possible base in the reference sequence and percentages of T, A, C and G variants. These measures were taken from the variants VCF file corresponding to the variants of the alignment between the human K562 dataset and its refseq.

In this third table, Table III, one can note that 35% of the variants are T's, this percentage is higher than for the other bases.

V. Discussion

A. Accessibility of data and reusability of code

As a reminder, the FAIR Data Principles state that data should be Findable, Accessible, Interoperable and Reusable in order to be called 'FAIR'. As for Open Science, it states that hardware and software developed in a research context should be accessible so that it can be reused by other researchers for further work.

The results of this first experiment show that the input data required to run the nano-ID neural network are Findable as a link to download them is provided. However, although the possibility to use Wget to avoid interrupted downloads, time outs or other failures is presented, it remains impossible to download datasets larger than 3GB. These specific data are therefore not Accessible and thus neither Interoperable, nor Reusable.

As for the software (i.e., nano-ID), it is not documented in a way that enables a new researcher to run it. Indeed, the order of the files to be run is not given and although it could be guessed it is not explicit what the input data must be and with what these last were generated.

B. Evaluation of the neural network and detection of RNA isoforms

The results summarized in Table I allow to measure the percentages of correctly/incorrectly predicted bases for each of the four canonical nucleotides. These percentages indicate that on average, 94% of the predictions are correct for an A position in the reference sequence, 93% for a C, 95% for a G and 85% for a T. Nevertheless, one can notice that this percentage is relatively low for the nucleotide T, meaning that T is more incorrectly predicted by comparison with the other nucleotides. T being the representation of U, that difference is probably the result of natural U bases exchanged for modified ^{5Br}U bases in the reference sequence. In fact, when observing Figure 8, one can note that the signal generated by ^{5Br}U is really close to the one generated by U and C, as previously stated. Therefore, it might be inferred that when this isoform is to be detected by the neural net, this one tend to predict a U or a C. This observation is reflected in Table I. Indeed, 85% of the time, a T base (coding in reality for either natural U or ^{5Br}U) is predicted as a T (coding in reality for a natural U) and 9% of the time as a C, which is at least three times as much as for A and G. In addition, if the percentage of T predictions seems higher than the one of C predictions, it is due to the fact that not all T positions code for ^{5Br}U.

The second table of results, Table II, highlights the fact that most of the variants are T's and are incorrectly predicted as C. This is probably due to the fact that an unknown amount of U positions in the sample were exchanged for a ^{5Br}U, meaning that the nano-ID neural network is expected to fail at predicting these as it only knows four classes (A, C, T and G). These T variants being most of time incorrectly predicted as a C reflects the fact that the raw signal generated by C is real close to the one generated by ^{5Br}U - hence, the confusion.

In the last table of results, Table III, it was noted that the percentage of T variants is relatively higher than the one for A, C or G. This shows that the neural network struggles more at predicting T than it does at predicting A, C or G.

VI. Conclusions

The purpose of this section is to attempt to provide answers to the two research questions previously formulated. To do so, it is proper to consider the results acquired in the two previously conducted experiments, as well as their interpretation.

A. Accessibility of data and reusability of code

As a reminder, the aim is to figure out whether Maier et al's research respect the principles of FAIR Data and Open Science. To do so, RQ1 was formulated as follow: (a) Can the data used in Maier et al.'s research be called 'FAIR' ? (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 ?

In the first experiment, the results obtained enable to declare that Maier et al.'s work does not respect neither the FAIR Data Principles, nor the Open Source principle. Indeed, in the discussion section it has been shown that the input data required to run the nano-ID neural network (except for one) are not Accessible and thus neither Interoperable, nor Reusable. In their research, Maier et al. use various other datasets for different tasks. However, the fact that most of the data cannot be downloaded breaks the FAIR Data Principles. As for the principle of Open Science, it has been demonstrated that the nano-ID software provided in the previously mentioned research is not documented in a way that enables a new researcher to use it. For this reason, one can state that the present research does not respect the principle of Open Science meaning that the conducted experiments cannot be validated and/or improved by a new researcher. Therefore, the data in Maier et al.'s research cannot be considered 'FAIR', 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.

B. Evaluation of the neural network and detection of RNA isoforms

Secondly, an answer to the second research question may be formulated, namely: What is the performance of the nano-ID neural network in terms of misclassification ? Does it enable one to detect RNA isoforms ?

In the second experiment, the results obtained are not sufficient in order to evaluate the overall performance of the nano-ID neural network. However, it provides a first step towards that goal. Indeed, this experiment provides minimal results in the sense that it measures the proportions of misclassified bases but only on two different datasets. Moreover, the synthetic sample used was not made to evaluate the neural network but rather to investigate if ^{5Br}U is well suited for nanopore sequencing. Still, the second dataset, human cells exposed to ^{5E}U labeling, may be used to evaluate the performance of the neural network. However, the issue remains that this dataset represents only one of the replicates and is therefore not representative - hence, the limited results.

Nonetheless, one can already interpret these results. Indeed, Table I shows that, in this specific sequence of RNA, T is often incorrectly predicted as C and otherwise, predicted as T but some exceptions. This shows that the nano-ID neural net seems to struggle at predicting U bases as some of them code for ^{5Br}U. The same finding might be observed in Table II. Finally, Table III also shows that T is the most common variant although it does not have C as most common alternative. Therefore, the answer to RQ2 is that the nano-ID neural network seems to perform well enough in order to detect the presence of ^{5Br}U and ^{5E}U despite the reduced amount of conducted experiments.

VII. Limitation & Further Improvement

The challenges encountered in this research were mainly the limited access to data and the difficulty to run the nano-ID neural network. These two obstacles have prevented the realisation of other experiments based on running the NN, getting outputs, aligning them and analysing the results. Moreover, the VCF file containing the complete alignment information of the human K562 cells dataset (generated with bcftools) turned out to be more than 40GB. The large size of that file combined with limited resources and time has lead to the impossibility to retrieve metrics out of that file in a way that is similar to that of the synthetic RNA dataset.

For further work, it would be necessary to manage to download all data available to the GRO, as well as being able to run the nano-ID neural network in order to test it on these input and verify whether the same output may be obtained. Moreover, it would be interesting to align these output with their reference sequence and generate the complete VCF files for all alignment obtained. This way, it would be possible to measure misclassification and evaluate the neural network in a more reliable manner as supplemental experiments may be conducted.

Additionally, it would be interesting to consider other datasets highlighting the presence of different isoforms in order to use similar alignment methods to see how the presence of these is revealed.

VIII. Acknowledgments

I thank Rachel Cavill (Assistant Professor in Maastricht University) and Rui Portela for supervising me. In addition, I thank those who supported me during this entire research and the reviewers, in particular, UM Peer Point for their precious help with academic writing.

References

[1] Güngör Budak. How to Download hg38/GRCh38 FASTA Human Reference Genome. https://www.gungorbudak.com/blog/2018/05/16/how-to-download-hg38-grch38-fasta-human-reference-genome/, 05 2018.

[2] L. Djirackor, S. Halldorsson, P. Niehusmann, H. Leske, L. Kuschel, J. Pahnke, B. Due-Tønnessen, I. Langmoen, C.J. Sandberg, P. Euskirchen, and E.O. Vik-Mo. Intraoperative DNA methylation classification of brain tumors impacts neurosurgical strategy. *Brain and Spine*, 1:100547, 2021.

[3] Dr. A. Ebertz. A journey through the history of dna sequencing. https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/.

[4] Dr. A. Ebertz. Who won the race to solve the dna structure? https://the-dna-universe.com/2020/06/25/who-won-the-race-to-solve-the-dna-structure/.

[5] The economist. An ambitious unicorn hopes to up-end DNA analysis. https://www.economist.com/img/b/608/818/90/sites/default/files/images/print-edition/20211002_STC967.png, 2021. [Online; accessed May 10, 2022].

[6] Maier et al. Native molecule sequencing by nano-id reveals synthesis and stability of rna isoforms. 2021.

[7] Thomas P et al. Landscape of next-generation sequencing technologies. 2011.

[8] Free Software Foundation, Inc. GNU Wget 1.21.1-dirty Manual. https://www.gnu.org/software/wget/manual/wget.html, 1996.

[9] Mattia Furlan, Anna Delgado-Tejedor, Logan Mulroney, Mattia Pelizzola, Eva Maria Novoa, and Tommaso Leonardi. Computational methods for rna modification detection from nanopore direct rna sequencing data. *RNA Biology*, 18(sup1):31–40, 2021. PMID: 34559589.

[10] GATK Team. Human genome reference builds - grch38 or hg38. https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19.

[11] Genome Research Limited. . https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing, 2021.

[12] J. Ruan C. Hercus P. Danecek H. Li, B. Handsaker and Genome Research Ltd. Samtools. http://www.htslib.org.

[13] J. Ruan C. Hercus P. Danecek H. Li, B. Handsaker and Genome Research Ltd. bcftools(1). https://samtools.github.io/bcftools/bcftools.html, 2008.

[14] Camilla L.C. Ip, Matthew Loose, John R. Tyson, Mariateresa de Cesare, Bonnie L. Brown, Miten Jain, Richard M. Leggett, David A. Eccles, Vadim Zalunin, John M. Urban, Paolo Piazza, Rory J. Bowden, Benedict Paten, Solomon Mwaigwisya, Elizabeth M. Batty, Jared T. Simpson, Terrance P. Snutch, Ewan Birney, David Buck, Sara Goodwin, Hans J. Jansen, Justin O'Grady, and Hugh E. Olsen. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, 4:1075, 2015.

[15] Douglass Turner Jill P. Mesirov. James T. Robinson, Helga Thorvaldsdóttir. igv.js: an embeddable javascript implementation of the integrative genomics viewer (igv), 2020.

[16] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 05 2018.

[17] NHGRI. Dna sequencing fact sheet. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet.

[18] S. Schwartz and Y. Motorin. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. *RNA biology*, page 1124–1137, 2017.

[19] Shian Su. A Look at the Nanopore fast5 Format - Shian Su. https://medium.com/@shiansu/a-look-at-the-nanopore-fast5-format-f711999e2ff6, 12 2021.

[20] Oxford Nanopore Technologies. Product specifications. https://nanoporetech.com/products/specifications. [Online; accessed May 31, 2022].
[21] Oxford Nanopore Technologies. Rna sequencing. https://nanoporetech.com/applications/rna-sequencing. [Online; accessed June 13, 2022].
[22] Oxford Nanopore Technologies. How nanopore sequencing works. https://nanoporetech.com/how-it-works, 2022. [Online; accessed May 7, 2022].
[23] Maastricht University. Open science - research - maastricht university. https://www.maastrichtuniversity.nl/research/open-science.
[24] Yuk Kei Wan, Christopher Hendra, Ploy N. Pratanwanich, and Jonathan Göke. Beyond sequencing: machine learning algorithms extract biology hidden in nanopore signal data. *Trends in Genetics*, 38(3):246–257, 2022.
[25] Aalbersberg IJ et al. Wilkinson MD, Dumontier M. The fair guiding principles for scientific data management and stewardship. 2016. [published correction appears in Sci Data. 2019 Mar 19;6(1):6].
[26] Liu Xu and Masahide Seki. Recent advances in the detection of base modifications using the nanopore sequencer. 65(1):25–33.

APPENDIX

A. Software commands and parameters used

```
# Open terminal

# Select path
cd computer/path

# Download Minimap2
git clone https://github.com/lh3/minimap2
cd minimap2 && make

# Generate an index file for the reference sequence
./minimap2 -x map-ont -d <refseq>.mmi <refseq>.fa

# Generate the alignment under SAM format
./minimap2 -ax map-ont <refseq>.fa <fastq_file>.fastq > <alignment>.sam

# Install Samtools on Conda
conda config --add channels bioconda
conda config --add channels conda-forge
conda create -n samtools samtools

# Activate environment
conda activate samtools

# Convert SAM file into BAM file
samtools view <alignment>.sam -b > <alignment>.bam

# Sort BAM file
samtools sort <alignment>.bam -o <alignment_sorted>.bam

# Install Bcftools on Conda
conda create -n bcftools -c bioconda bcftools -y

# Activate environment
conda activate bcftools

# Generate VCF file
bcftools mpileup -a INFO/AD,FORMAT/DP4 -o <alignment_sorted>.vcf -f <refseq>.fa <alignment_sorted>.bam

# OR generate BCF file and call variants to generate VCF file
bcftools mpileup -Ob -o <alignment_sorted>.bcf -f <refseq>.fa <alignment_sorted>.bam
bcftools call -a INFO/ADF -vmO z -o <alignment_sorted>.vcf <alignment_sorted>.bcf
```

Fig. 10: Commands used to run the software in order to convert files into VCF format.
The parameters used are the following: a) For the minimap2 map-ont command: -d for reducing the indexing time. b) For the samtools view command: -b for binary. c) For the samtools sort command: -o for specifying the output type/name. d) For the bcftools mpileup command: -f for specifying the reference sequence in FASTA format, -a for specifying output options and -o for specifying the output type/name.

B. Visualizations of the alignment using IGV

Fig. 11: Visualization of the alignment between a reference sequence and its prediction reads.
Each nucleotide is encoded by a color. The grey histogram represents the coverage track, which shows the read depth for each position (i.e., the amount of predictions at that specific position). Most of the time, one row codes for one read. Here, the coverage track follows a distribution that is skewed to the left, this is simply due to a technical issue related to the sequencer. The third track is the variant track. This one was fed with the VCF file of the synthetic RNA sample containing only the variants of the alignment. Each variant position is represented by a blue icon as it can seen at position 531. This figure was realized by means of the IGV tool [15] using the synthetic RNA dataset and its reference sequence.

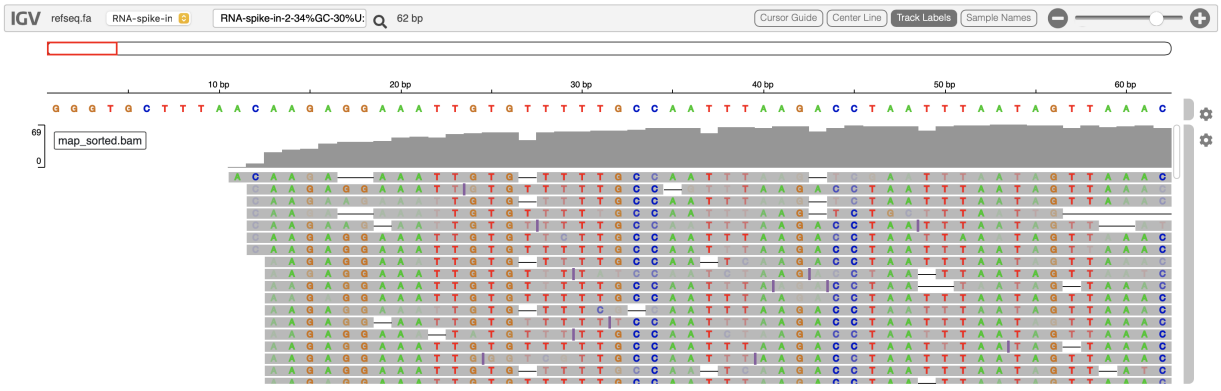


Fig. 12: This figure is a zoom of the the above Figure 11 allowing to see the predictions' values for positions 0 to 62. The opacity of each prediction reflects the confidence of that prediction. Here, the black horizontal line represented a deletion and the purple vertical line an insertion. This figure was realized by means of the IGV tool [15] using the synthetic RNA dataset and its reference sequence.

Fig. 13: Visualization of the alignment between the hg38 and the ^{5E}U 60min sample. The histogram in the coverage track shows several peaks, reflecting the fact that it is not a continuous sequence that was read by the sequencer but rather several sub-sequences characterised by several skewed left histograms. This visualization shows the mitochondrial chromosome of the hg38 (chrM). This figure was realized by means of the IGV tool [15].