Basecalling by deep learning on MinION nanopore sequencing data

Adrien Jarretier-Yuste
Supervised by Hector Hernandez-Vargas,
Chloé Goldsmith,
Céline Robardet,
Stefan Duffner,
Marc Plantevit

Université Claude Bernard Lyon 1, France

Abstract. DNA is a biological macromolecule present in all known living organisms, the basic unit is the nucleotide, or "base". The sequence of the human genome is essentially known since 2003, however there are still today many unknowns about the structure of our DNA.
In this internship we are interested in detecting the chemical modifications of one of the four bases of DNA, cytosine, with a third-generation sequencer. The MinION is a device for nanopore sequencing whose output is a time series corresponding to a current measurement. Basecalling for this device is the process of translating this measurement into a DNA sequence.
The objective of this internship is first to set up a pipeline for quick and simple analysis of DNA sequenced with the MinION. Subsequently, we propose a method for direct discrimination of non modified and modified cytosines as a fifth base during basecalling. The signal to analyse is a time series, comparable to data from an audio recording. In speech recognition literature, there are powerful models on this kind of data. The chosen and implemented model is a deep learning model, specifically a CNN-RNN-CTC. A convolutional neural network followed by a bidirectional LSTM recurrent neural network is used to perform a connectionist temporal classification. Such a network can be trained to properly label a DNA sequence based on a time series. Results show that our model is able to learn to correctly align and return a sequence from the sequencing of a synthetic DNA strand of a few hundred bases.

Keywords: basecalling, bioinformatics, deep learning,
nanopore sequencing, recurrent neural network

1 Introduction

1.1 DNA methylation

DNA methylation is a type of chemical modification of DNA, it is a process in which a methyl group is added on the fifth carbon of a cytosine, thus converting it into 5-methylcytosine (5mC) (Fig.1).
There are many other possible modifications of DNA [33], we are particularly interested in 5mC and 5hmC (5-hydroxymethylcytosine) in humans.

In humans, DNA methylation usually occurs on cytosines that precede guanine (this is called a CpG site). Between 60 and 90% of the CpG sites are methylated in mammals [21]. Some regions of the DNA have a high concentration of CpG sites, they are called CpG islands and are usually unmethylated. These are most often located in the promoter sequences. A promoter is a region of the DNA located upstream of the start site of the transcription of a gene [10] [30].

When a CpG island present in the promoter is methylated, this is associated with downstream gene inhibition (Fig.2).

Methylation is one of the phenomena determining the identity of cells, so that cells having the same basic DNA sequence are differentiated into specialized types (an example with the differentiation of blood cells in figure 3).

Abnormal methylation is associated with deregulated cellular processes that can lead to cancer [17] [18].

The functions of methylation are still poorly understood. New methods allow us to study the entire genome and paint a clearer picture of the methylome. It seems that the relationship between DNA methylation and gene transcription is more dynamic and nuanced than expected [23].

5-hydroxymethylcytosine (5hmC) is another chemical modification of cytosine, a methyl group on a cytosine (5mC) can be oxidized to 5hmC. 5hmC has been recently described as the "6th base" of DNA, early research indicates that its functional role is distinct from 5mC with a determining role in the differentiation of progenitor cells [11].

1.2 Immune system plasticity

Our immune system is a complex biological system characterized by its flexibility. Cancer cells can be detected and destroyed or left to divide and proliferate. This immune response is coordinated by different cell types, among which are the T cells that are known to play a key role. The balance between T cell subtypes (eg Th0, Tregs, Th17) influences the final result of the response [3] [36]. These T cell subtypes are not fixed, their cell type is known to be "plastic" [38], and their change in proportions is in some cases associated with cancer.

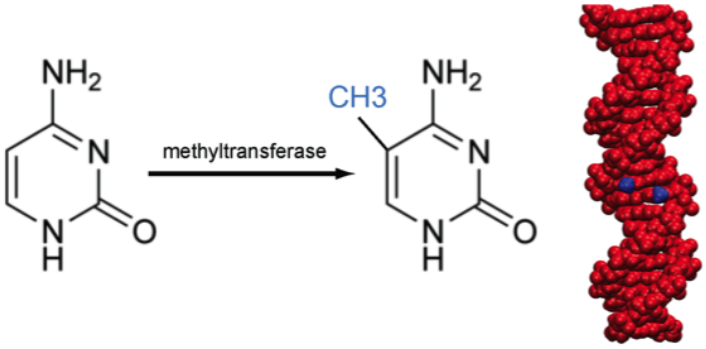

Fig. 1: Methylation process (schema from Mikhail Dozmorov)

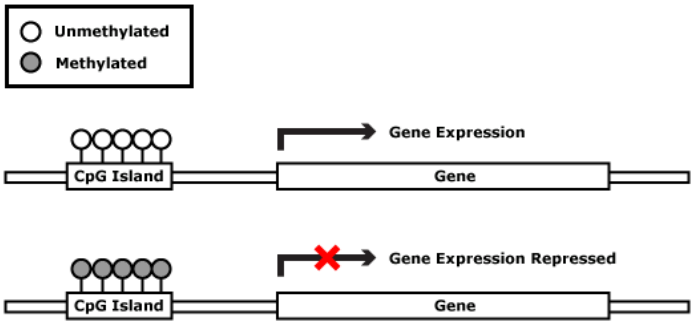

Fig. 2: Methylation in the promoter is associated with inactivation of the gene (schema from Mikhail Dozmorov)

This internship is part of the AEROBICS project (Analysis and Epigenetic Recognition Of dysBalanced Immune Cell plaSticity), which aims to study and measure this imbalance of immune cells as a cancer marker by identifying the methylation modifications. This project is itself part of the larger PLASCAN project (Preventing Tumor Plasticity and Adaptability: Towards the New Generation of Personalized Medicine), a multidisciplinary research project for understanding and modeling cancer.

1.3 DNA sequencing by nanopores

To better understand these mechanisms, we seek to study DNA and its modifications using a third generation sequencer, MinION, developed and distributed by Oxford Nanopore Technologies since 2014 [22].

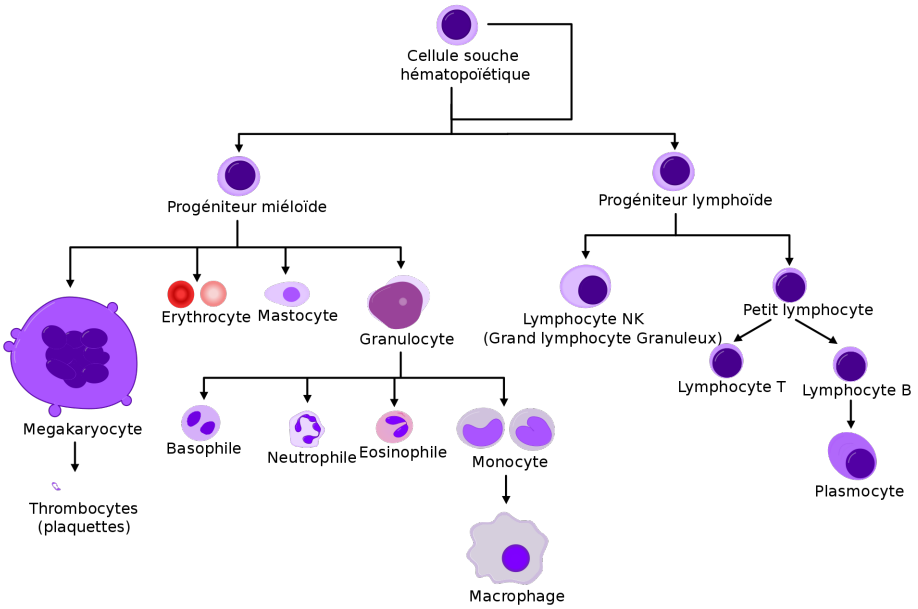

Fig. 3: Simplified model of hematopoiesis. A multipotent hematopoietic stem cell differentiates into a lymphoid or myeloid progenitor. The differentiation of the lymphoid progenitor results in different types of lymphocytes. The differentiation of the myeloid progenitor provides, after several stages, erythrocytes, other leucocytes (granulocytes, monocytes and macrophages) and platelets. (from the original by A. Rad)

Oxford Nanopore Technologies (ONT) is a UK-based company that develops nanopore sequencing products (including the portable DNA sequencer MinION).

The basic concept of nanopore sequencing is as follows:
Pass a strand of DNA through a nanoscale pore of a membrane from head to toe and apply a potential difference across the membrane. The nucleotides present in the pore will affect its electrical resistance so that current measurements over time can indicate the base sequence of the DNA passing through the pore [24] [5]. This electrical current signal ("squiggle" due to its appearance when plotted) corresponds to the raw data collected by an ONT sequencer (Fig. 4).

1.4 The basecalling

Basecalling, for ONT devices, is the process of translating this raw signal into a DNA sequence. This process is not a trivial task because the electrical signals come from single molecules and the rate at which the DNA strand passes through the pore is non-uniform, which produces noisy data [29]. In addition,

Fig. 4: The squiggle is the trace of the signal measured by the MinION. For the illustration, a sequence after basecalling was added above, for a visualization of a real signal see figure 11

the electrical resistance of a pore is determined by the bases present in several nucleotides which are in the narrowest point of the pore (about five nucleotides for the pores R9.4), which gives a large number of possible states (or k-mers): $4^5 = 1024$ for a standard four-base model.
When modified bases are present, e.g. 5-methylcytosine, the number of possible states can still increase:
$5^5 = 3125$.
Due to the non-uniformity of the transition rate of the DNA through the nanopore, it is also difficult to detect a transition between two identical k-mers [29]. This makes the basecalling of ONT devices a difficult machine learning problem and a determining factor for the quality and operability of ONT sequencing.

During this internship we have two objectives:

1. Set up a pipeline to quickly and easily analyze the DNA sequenced with the MinION (Fig. 6).
2. Develop a deep learning model to perform basecalling by directly differentiating unmodified and modified cytosines. This model can be specially trained for human DNA with the consideration of methylated cytosines.

2 The analysis pipeline

In order to efficiently analyze the data provided by the MinION (Fig. 6) we are developing a pipeline with different existing tools.

In the remainder of this document we will talk about a "standard DNA", it is a set of three samples of short synthetic DNA (900 bases) provided by Zymo Research (https://www.zymoresearch.com/products/5-methylcytosine-5-hydroxymethylcytosine-dna-standard-set) for which we know perfectly the sequence. We sequenced these samples with the MinION in order to have control data for all the experiments.

These three samples have exactly the same base sequence but the cytosines have been modified differently:

- The first sample contains only cytosines without modification.
- The second sample contains only 5-methylcytosines.
- The third sample contains 100% 5-hydroxymethylcytosines.

2.1 Pipeline Description

The final pipeline that has been developed is shown in figure 7.

- The basecaller performs the basecalling (cf 1.4), the result of this first part of analysis is a set of files in the format fastQ [8], these files contain, for each strand of DNA that went into a nanopore, its sequence accompanied by quality scores phred [13] [12].

- Before sequencing, each DNA sample is "barcoded", an artificial barcode [25] is attached to each strand of DNA. Thus one can sequence several samples at the same time. Once the sequence is identified after the basecalling, we can separate the different samples with the demultiplexer. This step allows us to separate fastQ files in different folders, each corresponding to a specific barcode.

- Then it is necessary to align the sequences with a reference file, for the standard DNA the sequence is provided by Zymo Research, for the human one, we use hg38. This alignment phase gives us Sequence Alignment Map (SAM) files that contain the previously obtained sequences with their position in the genome.

- Finally, we can identify the frequency of methylation of regions of the genome using Nanopolish [32].

2.2 The choice of basecaller

There are several tools capable of performing basecalling:

- Albacore (available on `community.nanoporetech.com`), the former basecaller developed by ONT, now replaced by Guppy
- BasecRAWller [34],
- Chiron [35], a basecaller based on deep learning similar to the model we later implement, a CNN-RNN-CTC.
- DeepNano [4]
- Flappie (`https://github.com/nanoporetech/flappie`) from ONT, an experimental basecaller who uses a CTC, it replaced Scrappie.
- Guppy (available on `community.nanoporetech.com`), ONT's current Gated Recurrent Unit network based basecaller.

- Nanocall [9]
- Scrappie (`https://github.com/nanoporetech/scrappie`) of ONT, formerly an experimental basecaller, now replaced by Flappie

Their performances have been compared in the literature [37] [29]. These results show empirically that the new neural network based basecallers outperform the old HMM (Hidden Markov Models) models. Thus Oxford Nanopore Technologies quickly integrate in their basecaller the most effective models and the most suitable to suit the evolution of nanopores version of MinION.

Their basecaller, Guppy, which uses a recurring GRU network, was therefore our choice for our pipeline.

Guppy is provided with several parts, Guppy_Basecaller, Guppy_Barcoder and Guppy_Aligner.

For the detection of methylation we had the choice between SignalAlign [28] and Nanopolish [32]. The 2 programs are HMM, however, SignalAlign runs under Python 2 and has not been updated for 3 years, so we preferred Nanopolish.

2.3 Pipeline Result

The use of our pipeline on standard DNA samples shows that it is possible for us to differentiate globally methylated sequences from unmodified sequences (Tab. 1, Fig. 5).

Moreover, the analysis of the 5-hydroxymethylated sample (5hmC) allows us to verify that this modification is not detected as a false positive for 5mC (Tab. 1, Fig. 5) . Nanopolish [32] is not trained for 5hmC, it is normal that it does not detect the chemical modification on this sample.

Sample	Unmodified	5-methylated	5-hydroxymethylated
Number of CpG sites	2391329	10521	3485
Number of CpG sites 5-methylated	33348	6939	64
P-value	$< .000001$	$< .000001$	$< .000001$
Average methylation	.014	.660	.019

Table 1: Average methylation levels of the CpG sites of the different standard DNA samples obtained with our pipeline.

Despite the positive results of the execution of our pipeline, two points can be immediately raised by seeing these results:

1. For the average methylation, we expect 0, 1 and 0 but we get 0.014, 0.660 and 0.019, up to 34% error in the detection of actually methylated Cytosines.

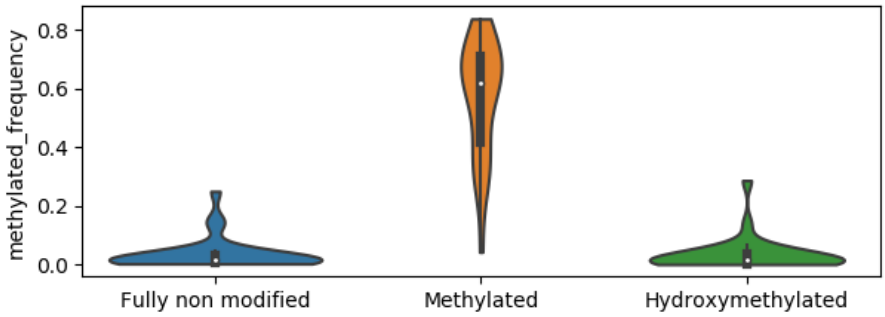

Fig. 5: Frequencies of 5-methylated cytosines detected by Nanopolish at the end of the pipeline on standard DNA samples. The first and third diagrams correspond respectively to the completely unmodified DNA sample and the 5-hydroxymethylated DNA sample, they are correctly detected as very little 5-methylated. The second sample corresponds to the standard DNA containing 5-methylated cytosines, Nanopolish detects it as much more strongly methylated than the others.

2. On the number of CpG sites, between the three synthetic DNA samples, the sequences are exactly the same, so we would expect about 1/3 of CpG sites in each sample, however we get a distribution of the number of sites between samples C, 5mC and 5hmC of 99%, 0.4% and 0.1%.

In our pipeline, the sources of errors during the analysis are numerous and cumulative. During basecalling, sequence prediction errors may occur, such an error may be reflected in the demultiplexing phase if the barcode has been labeled incorrectly. Finally Nanopolish can also make mistakes by classifying a CpG site as methylated or not.

In addition, the model used by Guppy is set and trained for the generation of sequences with 4 labels (A, C, G and T), it is not trained to recognize 5mC or 5hmC. It has been demonstrated that it is possible to detect 5-hydroxymethylcytosine distinctly from 5-methylcytosine [26], the measured electric current is different when the cytosines are unmodified, 5-methylated or 5-hydroxymethylated. This explains the difference in the number of sites between the different samples, the sequences of the modified samples are most likely misclassified and considered as poor quality at the basecalling stage.

In order to study methylation and hydroxymethylation, we propose to develop our own model in order to train it for our basecalling needs. By having the control of the model we will be able to parameterize it and train it to label the methylated and hydroxymethylated sites.

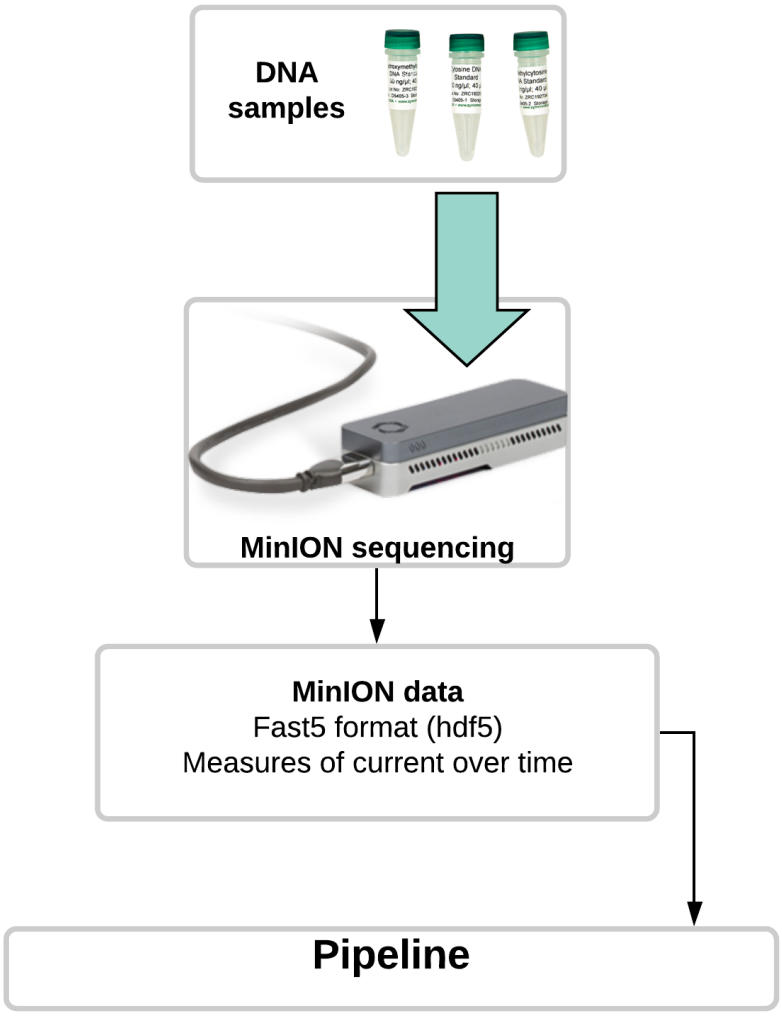

Fig. 6: DNA sequencing with MinION, DNA samples to analyses are prepared, a barcode is added to each (multiplexing). Once the preparation is done, samples are inserted into the sequencer. Reads are analysed with our pipeline (Fig. 7).

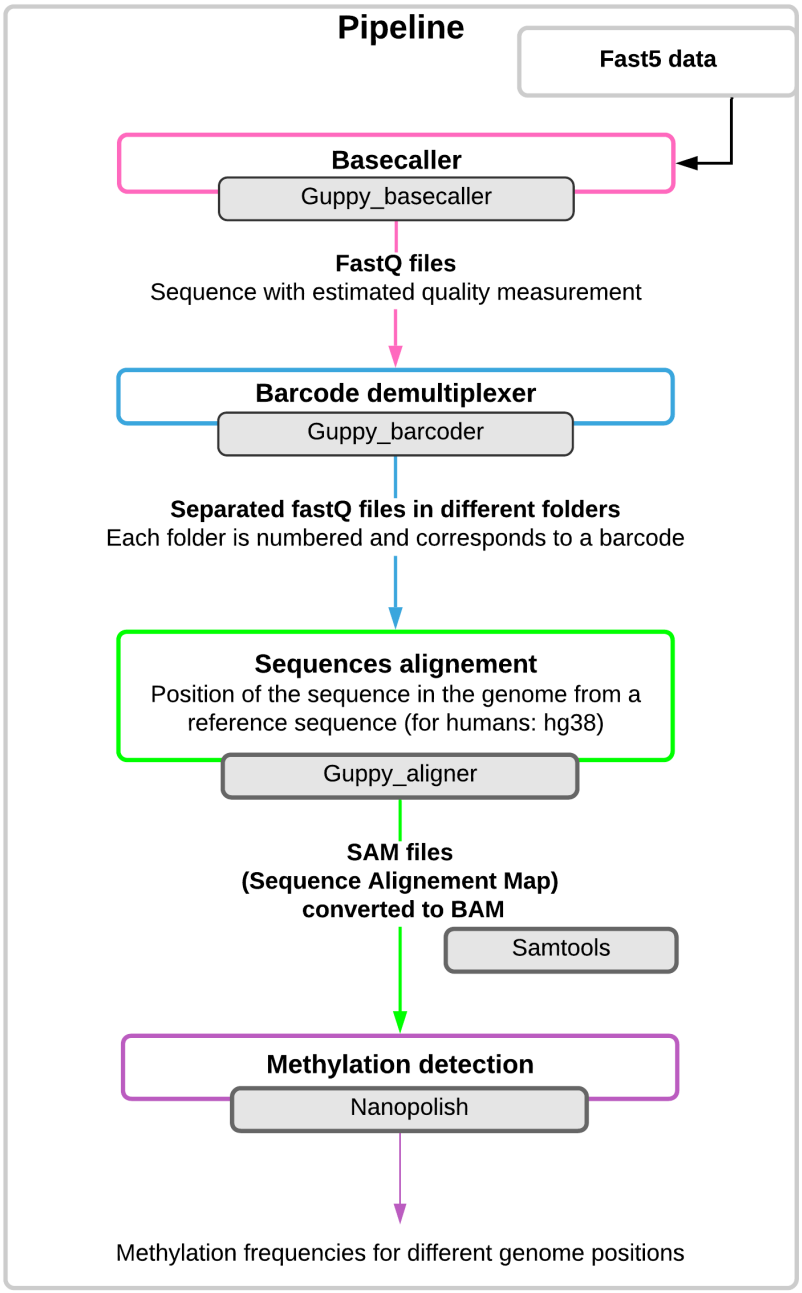

Fig. 7: The MinION data analysis pipeline (cf Fig. 6).

3 The implementation of a custom basecaller, CNN-RNN-CTC model

In addition to the problems previously seen with MinION signal basecalling (1.4), we have the following knowledge of the electrical signal sampling:

- The sampling frequency is 4 kHz.
- A strand of DNA passes through a nanopore on average at 450 bases / s, this rate varies according to the nucleotide composition, the motor protein can also stop working if it has no more energy (ATP).

So we have on average 9 current measurements per k-mer.

This sampling is similar in shape to that found in automatic speech recognition [16].

3.1 Sequence prediction with the CTC model

The CTC (Connectionist Temporal Classification) model is a modification made to a recurrent neural network during training to allow it to learn how to classify a sequence from a series of time data [15] without a prior segmentation step.

The basic idea of the CTC is as follows: to be able to predict a sequence z with the finite alphabet L from a series of data x with $|z| \leq |x|$, we add to L a label *blank*, $L' = L \cup blank$.

1. For each X_t entry, the recurrent neural network gives a prediction.
2. Each prediction is transformed into a probability vector of size $|L'|$.
3. The resulting matrix is then decoded to obtain a sequence of labels of L'.
4. Consecutive label repetitions are replaced by this label only once, in our case this corresponds to multiple measurements of the same k-mer while the DNA strand has not moved.
5. The *blanks* are deleted, they indicate the changes of k-mer, this step combined with the previous one allows to differentiate the identical consecutive nucleotides.

   Example :
   (a) For a 20 measurements input,
   the decoded sequence is _AAGG_CC_GGTTTGG__GG
   (b) Consecutive repetitions deleted: _AG_C_GTG_G
   (c) *Blanks* deleted : AGCGTGG
6. Compare the sequence obtained with the one expected.

3.2 Matrix decoding, the Beam search algorithm

The step 3 in the prediction with the CTC model, decoding, can be done in two ways. The first method, the simplest and fastest, is greedy, it simply consists in selecting the most probable labels at each time step (Fig. 8).

Algorithm 1 CTC Beam Search

Require: Probabilities matrix mat, $beamWidth$
$beams \leftarrow \{\varnothing\}$
$scores(\varnothing, 0) \leftarrow 1$
for $t = 1...T$ **do**
  $bestBeams \leftarrow bestBeams(beams, beamWidth)$
  $beams \leftarrow \{\}$
  for $b \in bestBeams$ **do**
    $beams \leftarrow beams \cup b$
    $scores(b, t) \leftarrow computeScore(mat, b, t)$
    for $c \in alphabet$ **do**
      $b' \leftarrow b + c$
      $scores(b', t) \leftarrow computeScore(mat, b', t)$
      $beams \leftarrow beams \cup b'$
    end for
  end for
end for
return $bestBeams(beams, 1)$

However, this method has a disadvantage, since it does not take into account all the possible paths in the matrix, it is possible to obtain a sequence with a suboptimal probability, which is why it may be useful to visit different paths (Fig. 9). However, it is impossible to visit exhaustively all paths, the complexity is too important, an algorithm has been proposed, the Beam Search [16] with wich it is possible de decode the matrix by iterating on the time steps while keeping a memory of configurable size on the best paths visited (Alg. 1).

3.3 Training the model

The training of a CTC aims to maximize the logarithm of the probabilities of correct classifications on the training data set, formally defined as follows:

With S the set of training data, $(x, z) \in S$,
x the input data, z the expected sequences, the objective function to minimize is:

$$- \sum_{(x,z) \in S} ln(p(z|x)) \tag{1}$$

We implemented the model with the Keras API [7] and the TensorFlow backend [1]. The model, represented in Figure 10, is composed of a first convolutional network layer (CNN), followed by a recurrent neural network (RNN).

This CNN-RNN architecture is classic in the field of automatic speech recognition [2] when predicting end-to-end sequences without segmentation. In addition, it has been demonstrated on MinION data that this architecture gives better predictions than a CNN or RNN network alone [35].

The RNN is a bidirectional network [31] with LSTM cells [19] [14].

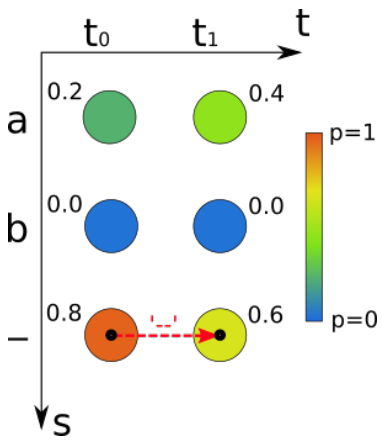

Fig. 8: Decoding the probability matrix with the greedy method, at each time step the label of highest probability is chosen. The sequence is "", the probability of the path is $0.8*0.6 = 0.48$. (Illustration from `https://towardsdatascience.com/`)

The RNN provides a prediction at each time step, the resulting prediction sequence is passed to a perceptron whose activation function is a softmax [6].

The result of this softmax is a prediction matrix whose decoding is described in 3.2.

Preparation of training data Training basecalling models presents a major challenge, it is impossible to provide the model with a training data set S for which we are fully aware of the correspondence between x and z. Indeed, even knowing the sequence of standard DNA samples does not guarantee that a reading obtained by the MinION is not a partial reading, which can happen if the DNA strand breaks.

The method for creating our dataset is therefore somewhat circular, it consists in sequencing DNA samples with a basecaller that we consider reliable, and then using this data as a learning base.

Oxford Nanopore Technologies provides a tool that allows us to process sequencing data, both raw (fastA) and basecalled (fastQ). Taiyaki (`https://github.com/nanoporetech/taiyaki`) allows us to match a measured current signal to the sequence we obtain with our pipeline. This allows us to obtain a file in which a data set x is "mapped" to the corresponding z sequences. A visualization of parts of two sequences aligned in this way on their signal is visible in figure 11.

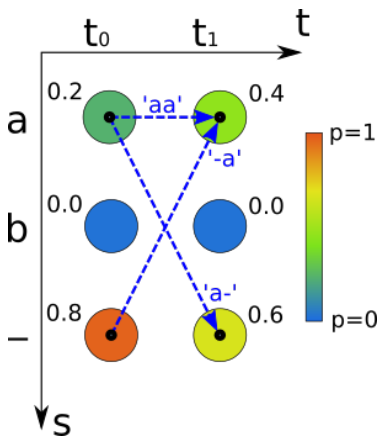

Fig. 9: Decoding by testing all paths, the probabilities of paths that give the same sequence are added together. The sequence with the highest probability is selected. The 3 paths that give "a" have probabilities: $0.2 * 0.4 = 0.08$, $0.2 * 0.6 = 0.12$, $0.8 * 0.4 = 0.32$, their sum is equal to 0.52, $0.52 > 0.48$ so the sequence "a" is more likely than the sequence"". (Illustration from `https://towardsdatascience.com/`)

3.4 Results

We conducted a series of training sessions to assess the learning capabilities of the model and test different hypotheses.

For our first training we use 50 sequences which will be divided into 40 / 10 for the training / validation (Fig. 12). We observe that the training converges very quickly. The error does not go below 800 which is a very high value.

- Hypothesis: The number of sequences has no influence on the value of the error.
- Experiment: Train the model with a very small number of sequences to highlight a difference in the value of the error (Fig. 13).
- Result: With only 4 sequences we already get an error close to 800, so no change from the data set with 40 sequences.

- Hypothesis: The error is not due to a bias in the data set.
- Experiment: Repeat the training with disjointed sets of sequences (Fig. 14).
- Result: We systematically obtain similar errors, which would suggest that the problem does not come from the data itself.

- Hypothesis: The length of the sequences has no influence on the value of the error.

- Experience: We successively train the model with 3 sets of 4 sequences whose length is constrained (Fig. 15).
- Result: There is a strong difference in the error, correlated to the average length of the sequences. The hypothesis can be rejected, the calculation of the error is biased by the length of the sequence.

We finally trained our model on a training data set containing only one sequence for more than 8 hours, i.e. 6600 iterations. As we expected, the error on the training sequence decreases but having only one example to learn from, the model overfits and the error on validation only increases (Fig. 16). The final value of the error is 0.2748, a Levenshtein distance between the predicted sequence and the expected sequence gives us 973, but if we remove the last 973 labels from our prediction, then we have a distance of 0. The predicted sequence is exactly the one expected, the model has learned well by heart, however additional labels are added at the end and distort the prediction.

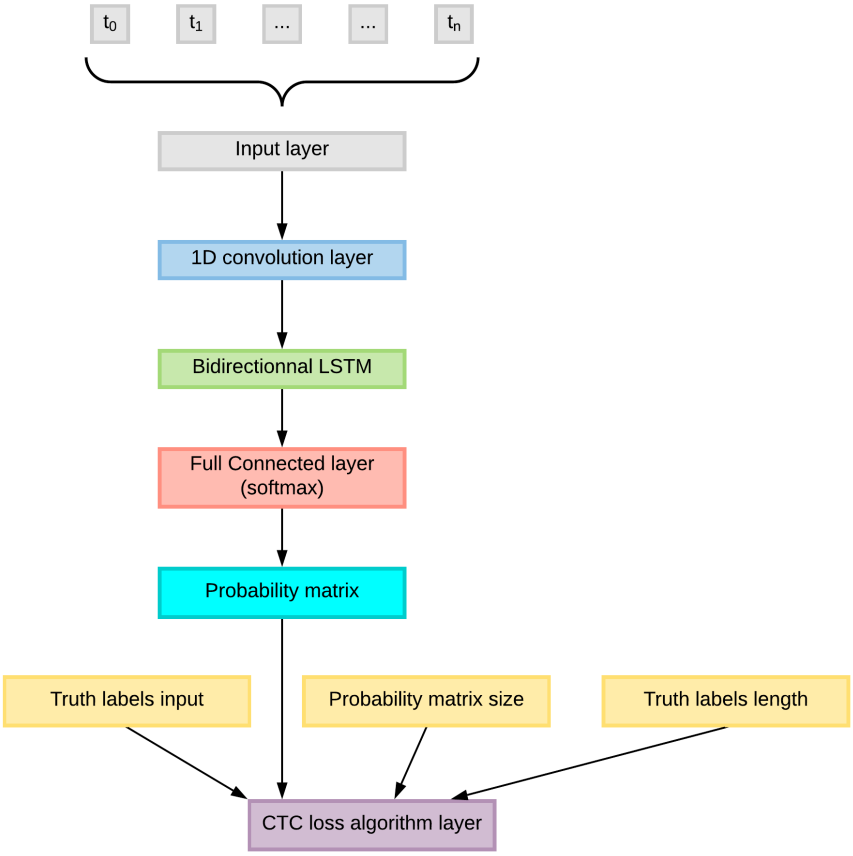

Fig. 10: The implemented model. A time series is given as an input, the first layer, a convolutional network, extracts local patterns. The recurring network makes a prediction for each time step. The next layer is a perceptron whose activation function is a softmax [6], the output is then a matrix of dimensions (n, $|L'|$) containing for each time step the predicted probability of each L' label. Finally, the last layer receives the expected sequence, the probability matrix, the number of data (n), the size of the expected sequence, and calculates the CTC error.

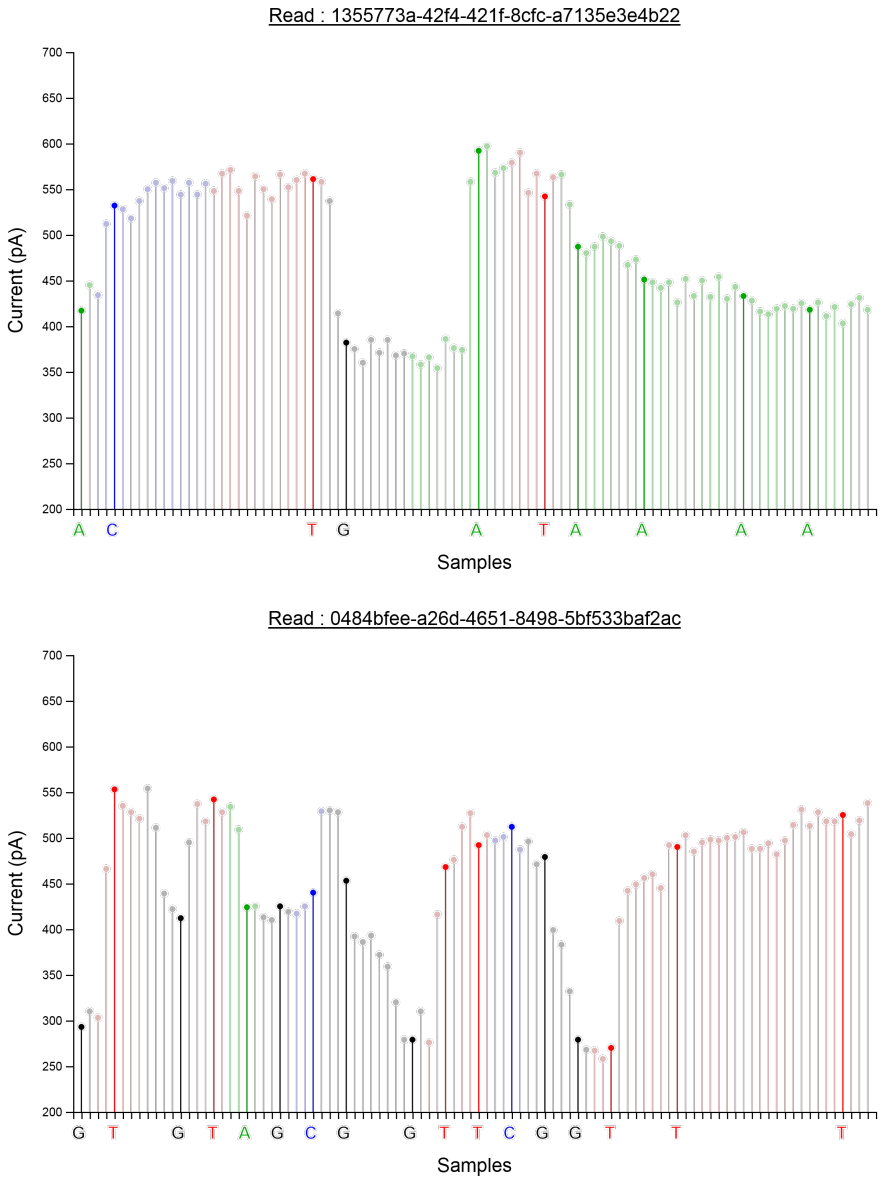

Fig. 11: Labelled non-methylated Standard DNA sequences, aligned with their squiggle, results obtained with Taiyaki after using Guppy for basecalling.

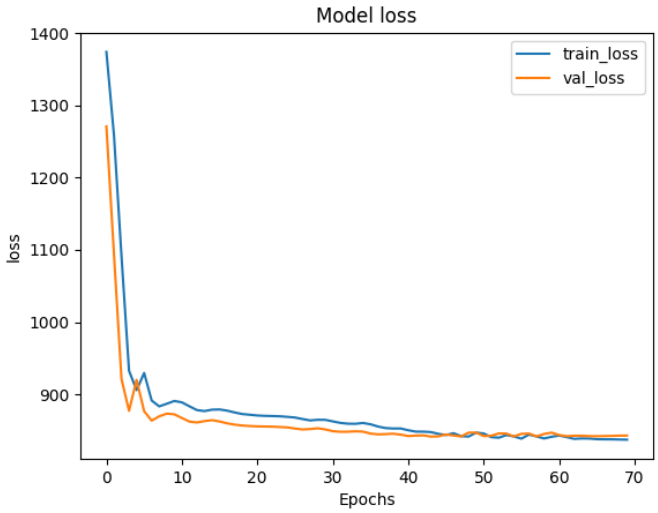

Fig. 12: Error evolution with 40 training sequences and 10 validation sequences.

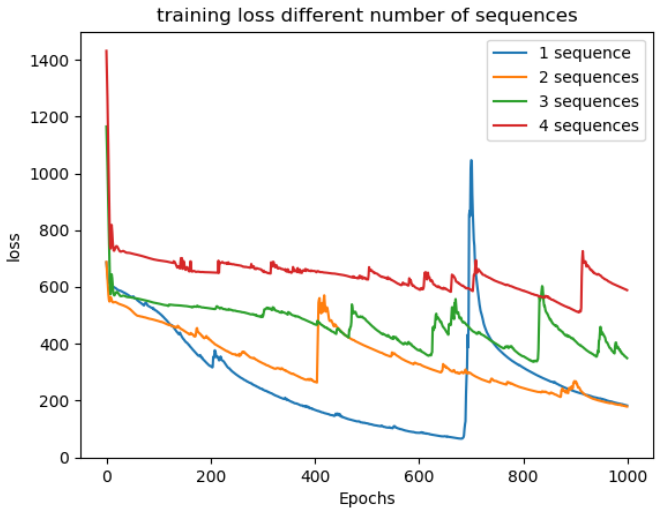

Fig. 13: Error evolution with different number of sequences in the training data sets.

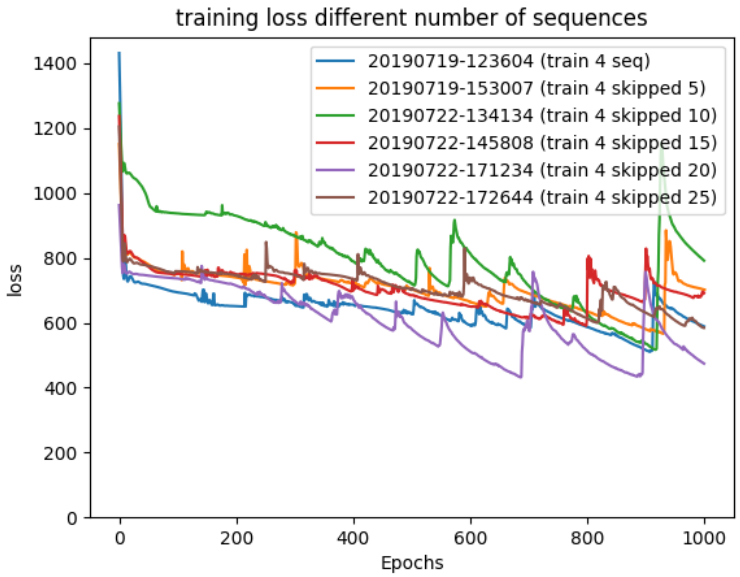

Fig. 14: Evolution of the error with disjointed training sets.

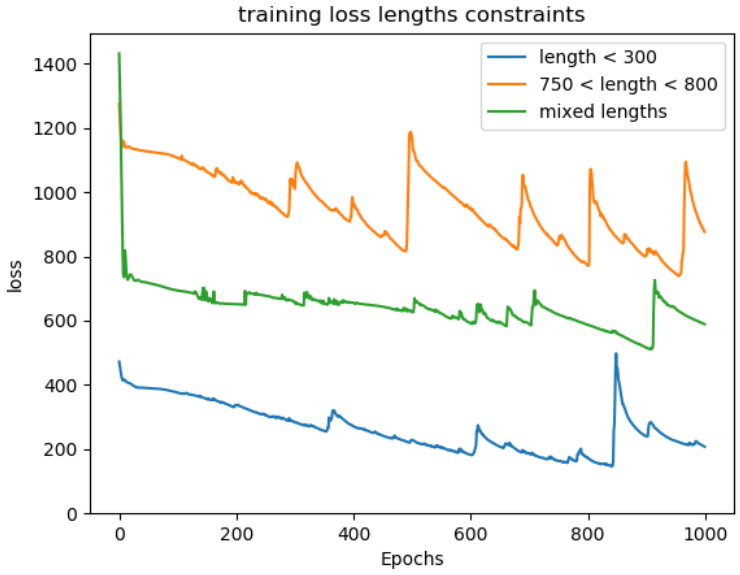

Fig. 15: Error evolution with sequences of different lengths.

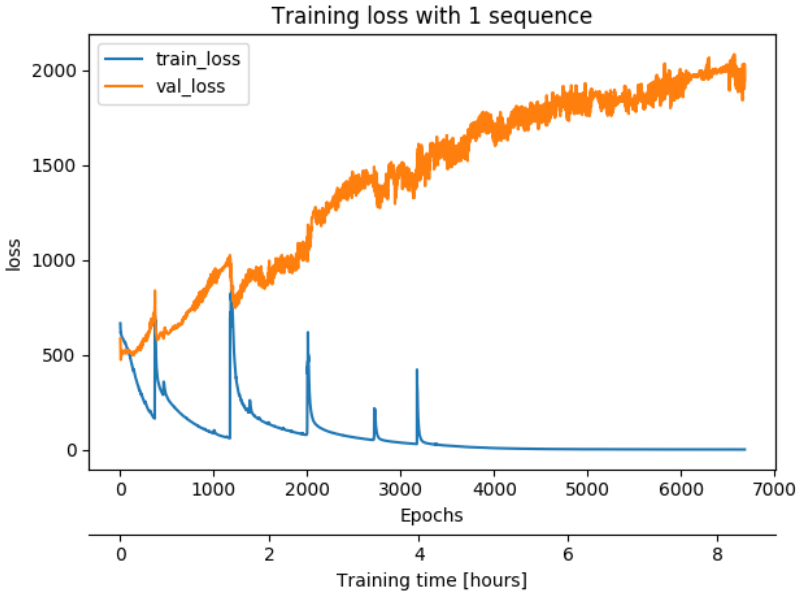

Fig. 16: Evolution of the error during training on a single sequence, validation is performed on a different sequence. There is an expected overfitting.

4 Discussion

Our model is able to learn a sequence by heart, however we have seen that when predicting it adds labels after the sequence, which distorts the result, the learning is also extremely long, 6600 iterations and 8 hours to learn a sequence by heart.

These problems are probably related to bugs in the implementation and possibly biases in learning. The implementation was done at a high level in order to make a protoype in a short time using the optimization function present in TensorFlow [1]. We recommend modifying the training loop to have absolute control, especially over the optimization function.

To accelerate learning, it will be necessary to implement batch normalization [20], in a deep learning network a normalization between each layer allows to considerably accelerate the convergence of learning.

In the context of research to understand the mechanisms of cancer in humans, we are particularly interested in the study of human DNA. Control over the model will allow us to train it on sequences of the human genome, which will make it a specialized and non generalist basecaller like existing basecallers. Indeed, the genomic context is completely different between all living species, bacteria, for example, have very different chemical modifications from what is found in humans.

For the detection of 5hmC, it is documented that brain tissues are rich in hydroxymethylated regions [27]. To generate a learning database to train the model to recognize this modification, it will be necessary to use techniques such as cas9-targeting, which makes it possible to target known regions of DNA for nanopore sequencing.

Acknowledgements I would especially like to thank
Hector Hernandez-Vargas and Chloé Goldsmith for their guidance within the laboratory, their availability and all the time they spent
to teach me a lot about biology, DNA sequencing and all the other topics we could discuss. Thanks to them, I was able to integrate myself into a multidisciplinary project and acquire incredible knowledge about a field of research that was unknown to me.

Many thanks to Celine Robardet, Stefan Duffner and Marc Plantevit from the "dm2l" team at LIRIS for their help and advice.

I would also like to thank Anthony Ferrari who took time to provide me with access to the bioinformatics platform's computing cluster.

I am very grateful to the entire "TGF-beta and regulation of the immune response" team at the CRCL for their hospitality, Julien Marie, Saidi Soudja, Vincent Flacher, Alexandra Lainé, Ossama Labiad,
Ramdane Igalouzene, Olivier Fesneau, Apostol Apostolov, Sarah Bettini.

Thanks also to Yenkel Grinberg-Bleyer, Robert Dante and Mounia Defontaine.

My sympathy finally goes to Shiqiang Xu, Inès Nihal El Rifai, Julien Laurencin, Clara Poncet, as well as the other interns with whom I was able to exchange during this internship.

Meeting these people, researchers, post-doctoral fellows, PhD students, interns, taught me a lot of things, beyond the mission of the internship, about the world of research, but also about different cultures.

Finally a thank you to Davide Bolognini, NanoR developer with whom I was able to exchange and who was extremely reactive.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), `https://www.tensorflow.org/`, software available from tensorflow.org
2. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., Zhu, Z.: Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In: International Conference on Machine Learning. pp. 173–182 (Jun 2016), `http://proceedings.mlr.press/v48/amodei16.html`
3. Basu, R., Hatton, R.D., Weaver, C.T.: The Th17 family: flexibility follows function. Immunological reviews 252(1), 89–103 (Mar 2013), `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3607325/`
4. Boža, V., Brejová, B., Vinař, T.: DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLOS ONE 12(6), e0178751 (Jun 2017), `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178751`
5. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., Soni, G.V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., Schloss, J.A.: The potential and challenges of nanopore sequencing. Nature biotechnology 26(10), 1146–1153 (Oct 2008), `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2683588/`
6. Bridle, J.S.: Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In: Soulié, F.F., Hérault, J. (eds.) Neurocomputing. pp. 227–236. NATO ASI Series, Springer Berlin Heidelberg (1990)

7. Chollet, F., et al.: Keras. https://keras.io (2015)
8. Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M.: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38(6), 1767–1771 (Apr 2010), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847217/
9. David, M., Dursi, L.J., Yao, D., Boutros, P.C., Simpson, J.T.: Nanocall: an open source basecaller for Oxford Nanopore sequencing data. Bioinformatics 33(1), 49–55 (Jan 2017), https://academic.oup.com/bioinformatics/article/33/1/49/2525680
10. Deaton, A.M., Bird, A.: CpG islands and the regulation of transcription. Genes & Development 25(10), 1010–1022 (May 2011), http://genesdev.cshlp.org/content/25/10/1010
11. Ecsedi, S., Rodríguez-Aguilera, J.R., Hernandez-Vargas, H.: 5-Hydroxymethylcytosine (5hmc), or How to Identify Your Favorite Cell. Epigenomes 2(1), 3 (Mar 2018), https://www.mdpi.com/2075-4655/2/1/3
12. Ewing, B., Green, P.: Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Research 8(3), 186–194 (Mar 1998), http://genome.cshlp.org/content/8/3/186
13. Ewing, B., Hillier, L., Wendl, M.C., Green, P.: Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. Genome Research 8(3), 175–185 (Mar 1998), http://genome.cshlp.org/content/8/3/175
14. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to Forget: Continual Prediction with LSTM. Neural Computation 12(10), 2451–2471 (Oct 2000), http://www.mitpressjournals.org/doi/10.1162/089976600300015015
15. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 369–376. ICML '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1143844.1143891, event-place: Pittsburgh, Pennsylvania, USA
16. Graves, A., Jaitly, N.: Towards End-to-End Speech Recognitionwith Recurrent Neural Networks. In: International conference on machine learning. p. 9 (2014)
17. Herceg, Z., Ghantous, A., Wild, C.P., Sklias, A., Casati, L., Duthie, S.J., Fry, R., Issa, J.P., Kellermayer, R., Koturbash, I., Kondo, Y., Lepeule, J., Lima, S.C.S., Marsit, C.J., Rakyan, V., Saffery, R., Taylor, J.A., Teschendorff, A.E., Ushijima, T., Vineis, P., Walker, C.L., Waterland, R.A., Wiemels, J., Ambatipudi, S., Esposti, D.D., Hernandez-Vargas, H.: Roadmap for investigating epigenome deregulation and environmental origins of cancer. International Journal of Cancer 142(5), 874–882 (2018), https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.31014
18. Hlady, R.A., Sathyanarayan, A., Thompson, J.J., Zhou, D., Wu, Q., Pham, K., Lee, J.H., Liu, C., Robertson, K.D.: Integrating the Epigenome to Identify Drivers of Hepatocellular Carcinoma. Hepatology 69(2), 639–652 (2019), https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.30211
19. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8), 1735–1780 (Nov 1997), https://doi.org/10.1162/neco.1997.9.8.1735
20. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs] (Feb 2015), http://arxiv.org/abs/1502.03167, arXiv: 1502.03167
21. Jabbari, K., Bernardi, G.: Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene 333, 143–149 (May 2004), https://linkinghub.elsevier.com/retrieve/pii/S0378111904000836

22. Jain, M., Olsen, H.E., Paten, B., Akeson, M.: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology 17(1), 239 (Nov 2016), https://doi.org/10.1186/s13059-016-1103-0
23. Jones, P.A.: Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics 13(7), 484–492 (Jul 2012), http://www.nature.com/articles/nrg3230
24. Kasianowicz, J.J., Brandin, E., Branton, D., Deamer, D.W.: Characterization of individual polynucleotide molecules using a membrane channel. Proceedings of the National Academy of Sciences 93(24), 13770–13773 (Nov 1996), https://www.pnas.org/content/93/24/13770
25. Kress, W.J., Erickson, D.L.: DNA barcodes: Genes, genomics, and bioinformatics. Proceedings of the National Academy of Sciences of the United States of America 105(8), 2761–2762 (Feb 2008), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2268532/
26. Laszlo, A.H., Derrington, I.M., Brinkerhoff, H., Langford, K.W., Nova, I.C., Samson, J.M., Bartlett, J.J., Pavlenok, M., Gundlach, J.H.: Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. Proceedings of the National Academy of Sciences 110(47), 18904–18909 (Nov 2013), http://www.pnas.org/cgi/doi/10.1073/pnas.1310240110
27. Ma, Q., Xu, Z., Lu, H., Xu, Z., Zhou, Y., Yuan, B., Ci, W.: Distal regulatory elements identified by methylation and hydroxymethylation haplotype blocks from mouse brain. Epigenetics & Chromatin 11 (Dec 2018), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6311040/
28. Rand, A.C., Jain, M., Eizenga, J.M., Musselman-Brown, A., Olsen, H.E., Akeson, M., Paten, B.: Mapping DNA methylation with high-throughput nanopore sequencing. Nature Methods 14(4), 411–413 (Apr 2017), https://www.nature.com/articles/nmeth.4189
29. Rang, F.J., Kloosterman, W.P., de Ridder, J.: From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biology 19(1), 90 (Jul 2018), https://doi.org/10.1186/s13059-018-1462-9
30. Saxonov, S., Berg, P., Brutlag, D.L.: A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences 103(5), 1412–1417 (Jan 2006), https://www.pnas.org/content/103/5/1412
31. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45(11), 2673–2681 (Nov 1997)
32. Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., Timp, W.: Detecting DNA cytosine methylation using nanopore sequencing. Nature Methods 14(4), 407–410 (Apr 2017), https://www.nature.com/articles/nmeth.4184
33. Sood, A.J., Viner, C., Hoffman, M.M.: DNAmod: the DNA modification database. Journal of Cheminformatics 11(1), 30 (Apr 2019), https://doi.org/10.1186/s13321-019-0349-4
34. Stoiber, M., Brown, J.: BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal. bioRxiv p. 133058 (May 2017), https://www.biorxiv.org/content/10.1101/133058v1
35. Teng, H., Cao, M.D., Hall, M.B., Duarte, T., Wang, S., Coin, L.J.M.: Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience 7(5) (May 2018), https://academic.oup.com/gigascience/article/7/5/giy037/4966989

36. Vesely, M.D., Kershaw, M.H., Schreiber, R.D., Smyth, M.J.: Natural Innate and Adaptive Immunity to Cancer. Annual Review of Immunology 29(1), 235–271 (Apr 2011), http://www.annualreviews.org/doi/10.1146/annurev-immunol-031210-101324
37. Wick, R.R., Judd, L.M., Holt, K.E.: Performance of neural network basecalling tools for Oxford Nanopore sequencing. bioRxiv p. 543439 (Feb 2019), https://www.biorxiv.org/content/10.1101/543439v1
38. Ziegler, S.F., Buckner, J.H.: FOXP3 and the Regulation of Treg/Th17 Differentiation. Microbes and infection / Institut Pasteur 11(5), 594–598 (Apr 2009), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2728495/

Appendices

.1 Deregulated cellular processes

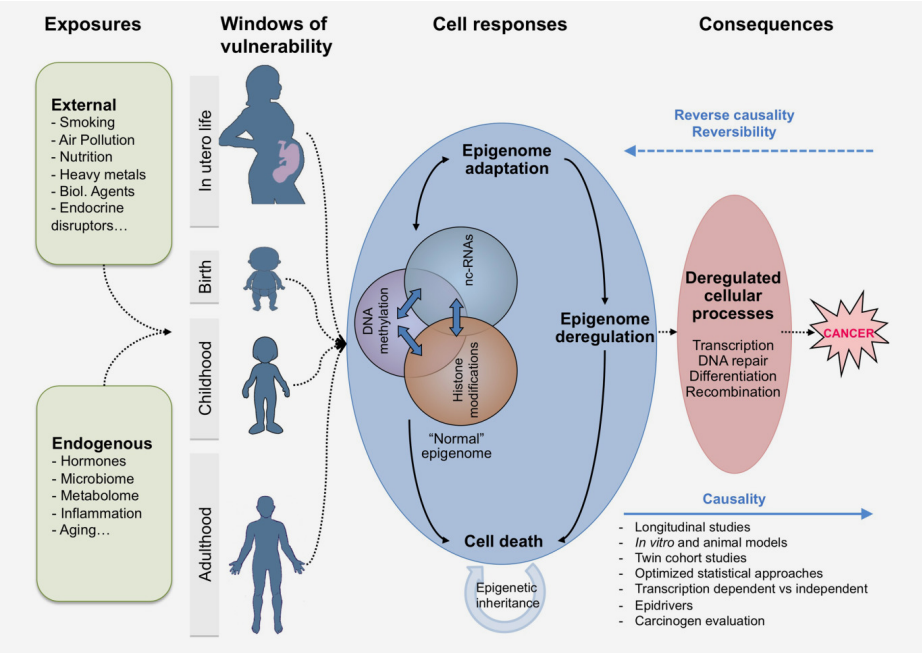

Fig. 17: Exposure to external sources and internal processes can induce stable and potentially reversible changes in the epigenome. The "signatures" and persistence of these alterations depend on multiple factors such as the duration of exposure, the type of tissue and the stage of development (The periods of puberty or intrauterine life can be particularly sensitive). (from [17])

.2 Length distribution of standard DNA sequences completely unmethylated

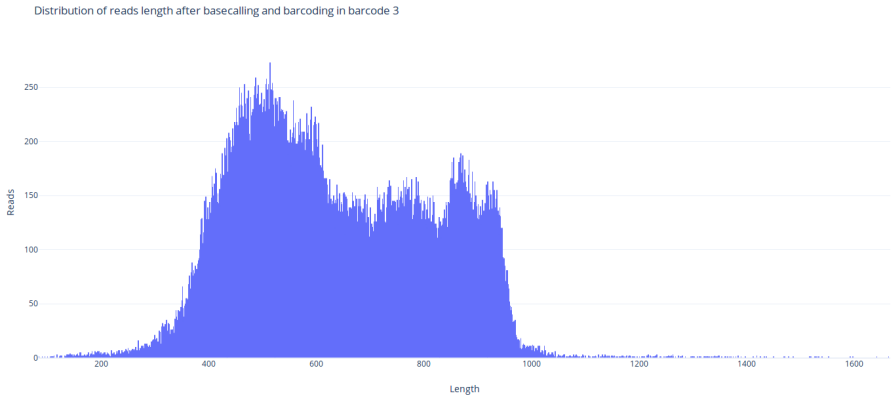

Fig. 18: The sequences with barcode 3 are part of the non-methylated sample. We know that the sequences make 900 bases, however after basecalling we observe a distribution of length mainly between 350 and 900, with two modes at 500 and 900. During DNA preparation, manipulations can damage it, sometimes breaking the strand, this explains the presence of a mode between 450 and 500, the DNA strands can be cut in half which gives a large number of sequences of 450 length. Longer sequences can be contamination, the DNA of a bacterium that has slipped into the sample, for example, or basecaller prediction errors. The large variation in sequence sizes can also be attributed to prediction errors.