Telomere Detection & Analysis with Nanopore Sequencing Using HMM Model

Lior Paz

Hebrew University of Jerusalem, October 2020

Advisor: Prof. Dudy Tzfati

Abstract

Nanopore sequencing is a novel approach for long reads sequencing of DNA or RNA molecules. It is the only available technology that can read DNA and RNA directly at a single-molecule level, without the need for PCR, reverse transcription, DNA polymerization etc. Most genomes contain significant amounts of repetitive DNA (e.g. telomeres, centromeres, transposons). As the short reads produced by traditional next generation sequencing technology may not span each given repetitive region, the resulting genome assemblies can be highly fragmented. Long-read sequencing technologies have a significant advantage here as the reads generated are more likely to span the full repetitive region, allowing the creation of accurate genome assemblies with minimal gaps.
Telomeres are essential structures that protect the ends of our chromosomes. They are composed of 5-20kb of TTAGGG repeats ending with a 3' overhang, which is critical to their function. Telomeres have important implications in aging, genome stability and cancer. However, no method is currently available that can sequence and reveal the full status of individual single telomeres, including their overall length, the length of the 3' overhang, variations and modifications in the sequence, and the identity of the chromosome end. By creating Telomere Detection Algorithm to go along with Nanopore sequencing Data, we aim to establish such a method.

Introduction

In almost all species that have cells with linear chromosomes, telomeres consist of G-rich repeats and associated proteins. Each of the 92 telomeres in a diploid human cell contain between less than 0.5 kb to more than 20 kb of (TTAGGG)*n* repeats which are in dynamic equilibrium with a specific set of proteins. The G-rich strand is invariably orientated 5′–3′ and the very 3′ end terminates in a 100–200 bp single stranded overhang which is believed to be important in forming a t-loop structure. Telomere length in human cells is strikingly

heterogeneous [1], but at least a few hundred nucleotides of telomere repeats must "cap" each chromosome end to avoid activation of a DNA damage response and DNA repair pathways. Variant telomere repeats are interspersed with pure telomeric repeats in the initial 1 kb of the array; whether this region should be defined as telomeric or subtelomeric, and whether it retains any/all telomere function is still under debate (Fig. 1).

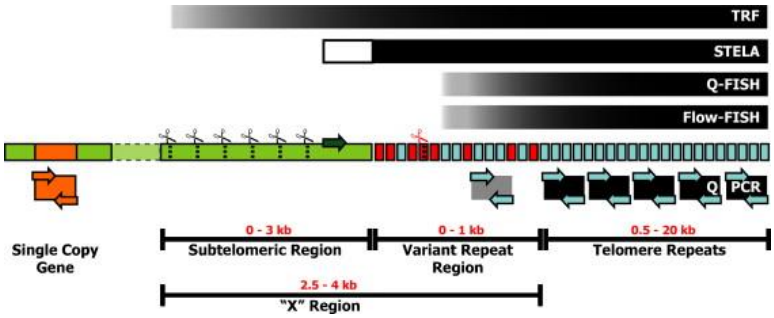


Fig 1. Extra-telomeric regions detected using different telomere measurement techniques [2]

Telomere length measured by TRF includes subtelomeric DNA (green bar) and telomere variant repeats (red bars) which can occupy up to 2 kb of the start of the telomere. This "X" region overestimates the length of pure telomeric repeats (blue bars)

Telomeres have been intensely studied since the proposal that telomere loss could cause cell senescence [3]. That gradual loss of telomere repeats contributes to replicative senescence or apoptosis in human cells was confirmed and loss of telomeres has been implicated in genomic instability and many age-related diseases. The realization that the length of telomere repeats at individual chromosome ends is a critical variable in cell fate decisions has highlighted the need for techniques that can provide accurate information on the length of telomere repeats in different cell types. Measuring telomere length is less important than the distribution of lengths and particularly the length of the shortest telomeres, which are the ones that may trigger genomic instability and checkpoint activation and determine cell fate [4]. There is also evidence that the length of the G-rich 3' telomeric overhang is even more important than the overall length.
Furthermore, we know very little about sequence variations at telomeres (although there is evidence [14]) because the current methods cannot sequence a long fragment of a telomeric sequence. No method is currently available that can provide the full status of single telomeres, including the overall length, the overhang length, the sequence, and the identity of the telomere. Our aim is to establish such a method using Nanopore sequencing. Once established, since only little amount of DNA is required, the application can be combined with single cell methods and population studies by way of barcoding and multiplexing to characterize the telomeres on a single molecule level in single cells or precious clinical samples.

Second generation sequencing has limitations because it requires PCR amplification, and it is based on sequencing of short fragments (at high coverage). The polymerases are sensitive to secondary structures and repeats. The short reads make it difficult or impossible to assemble larger regions and genomes containing repetitive regions. Third generation sequencing (e.g., Pacific Biosciences - PacBio) resolved some of the difficulties by enabling long reads and single molecule sequencing without amplification. However, it is still based on DNA polymerase copying of a template, and not direct sequencing.

Nanopore sequencing [5], considered the fourth-generation sequencing technology, is based on an entirely different approach. In this technology a single molecule of DNA or RNA can be sequenced directly by passing it through a very small pore (nanopore) in a membrane (Fig. 2). Sequencing is made possible because, while passing through the pore, each base cause characteristic changes in the electric current density across the nanopore surface, which is measured by a sensor. Sequencing occurs without the need for PCR amplification, reverse transcription, DNA polymerization or chemical labelling of the sample. Most importantly, this is the only available technology that can read DNA and RNA directly, at a single-molecule level, without the need for any polymerase, and thus it is insensitive to repeats and does not introduce any bias. The computational approach to convert signatures of electric traces to sequence data involves machine learning and recurrent neural networks. The Nanopore has shown impressive results in ultra-long reads and is likely to enable us to sequence the telomeres in full [6].

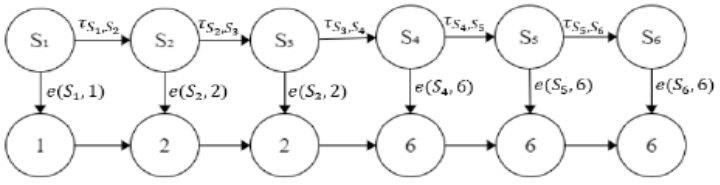
Fig 2. Animation of DNA strand passing through nanopore [17]

On the side we can see the changing electrical current according to sequence

We have chosen to tackle the challenge of detecting telomers repetitions in the Nanopore data with Hidden Markov Model (HMM) [7]. Markov Chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. We can model a first order Markov chain into an automaton, describing the different states, the probability for each event per state, and the transition probabilities between the states. In Fig. 3 we see an example where the state represents type of dice, and the event is the number received from a throw of that dice.

For sequence $(x_1, x_2 \dots x_n)$ and sequence of states $(s_1, s_2 \dots s_n)$,given an emissions matrix e_{s_i,x_i} describing $P(x_i|s_j)$, and transition matrix τ_{s_{i-1},s_i} describing $P(s_i = s \in \{states\}|s_{i-1})$, we can calculate full log likelihood - $P(x_1 \dots x_n, s_1 \dots . . s_n) = \Pi_i e_{s_i,x_i} \cdot \tau_{s_{i-1},s_i}$.

Given $(x_1, x_2 \dots x_n), e_{s_i,x_i}, \tau_{s_{i-1},s_i}$ – we can calculate the sequence $(s_1 \dots . . s_n)$ that will lead to



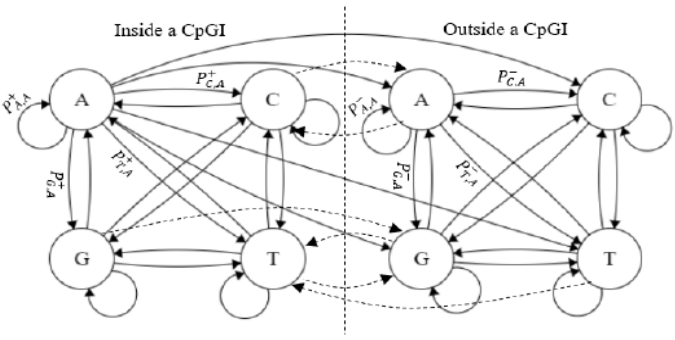


Fig 3. HMM graphic representation

On left - the cheating croupier example. The states (cubes) are hidden, and we can only observe their outcome (number on the dice). e is the emissions matrix, and τ is the transition matrix. **On the right,** automaton describing the states of HMM aimed to detect CpGI islands.

the best log likelihood, and by that try to infer the different parts and motifs of the sequence, or in our case as will be seen later – detecting telomeres.
There are 2 main known Algorithms for that – Viterbi [8] & Posterior. According to [9,10,11] - the HMM is endowed with a decoding algorithm to assign the most probable state path, and in turn the class labelling, to an unknown sequence. The Viterbi and the posterior decoding algorithms are the most common. The former is very efficient when one path dominates, while the latter, even though does not guarantee to preserve the HMM grammar, is more effective when several concurring paths have similar probabilities.
In our research we will use the posterior algorithm. The posterior is the probability the i'th state in the sequence will equal k given the entire sample: $P(S_i = k|(x_1 \dots x_n)$. An efficient way of calculating it will be using the forward-backward algorithm. Forward matrix could be efficiently calculated using dynamic programming, so each cell will contain $F_{k,i} = P(x_1 \dots x_i|s_i = k)$. In same way, we can calculate the backward matrix $B_{k,i} = P(x_{i+1} \dots x_n|s_i = k)$. Using both, we can calculate the posterior – $P\big(S_i = k\big|(x_1 \dots x_n)\big) = \frac{B_{k,i} \cdot F_{k,i}}{P(x_1 \dots x_n)}$.
The Baum-Welch [12] is an Expectation Maximization machine learning algorithm. It is used to understand what the emission & transition probabilities are $\Theta = (e, \tau)$ that will result the best log likelihoods possible for a given set of known sequences $\left\{x_1^j . \dots x_{n_j}^j\right\}_{j=1}^{N}$. This can enhance the model capabilities in case we do not know the probabilities in advance. It is the case in telomere research. In addition, we can infer from these learned probabilities from a set of telomeres significant biological conclusions. This is simply achieved by an iterative process where we find the hidden states in every iteration, and improve Θ according to $\hat{e}_{k,x} = \frac{N_{k,x}}{\Sigma_y N_{k,y}}$, $\hat{\tau}_{k,l} = \frac{N_{k,l}}{\Sigma_m N_{k,m}}$. $k, l \in States, x \in Observations$.

Methods

Decode HMM states using Posterior Algorithm

According to HMM principles and using posterior algorithm, we developed a tool for Telomere Detection. The DNA was modelled into different sets of Markov states - telomeric parts and non-telomeric, where the telomeric parts where divided to the different suspected telomeric motifs – TTAGGG, TTAAAA, TTGGGG. Every Motif is defined as a set of 6 states, where every motif state for specific index is set with high transition probability to continue to it is following index state. For every specific state, high emission probability was given (Fig. 5) to the matching DNA base (The observation). The fully detailed model is in Fig. 4.

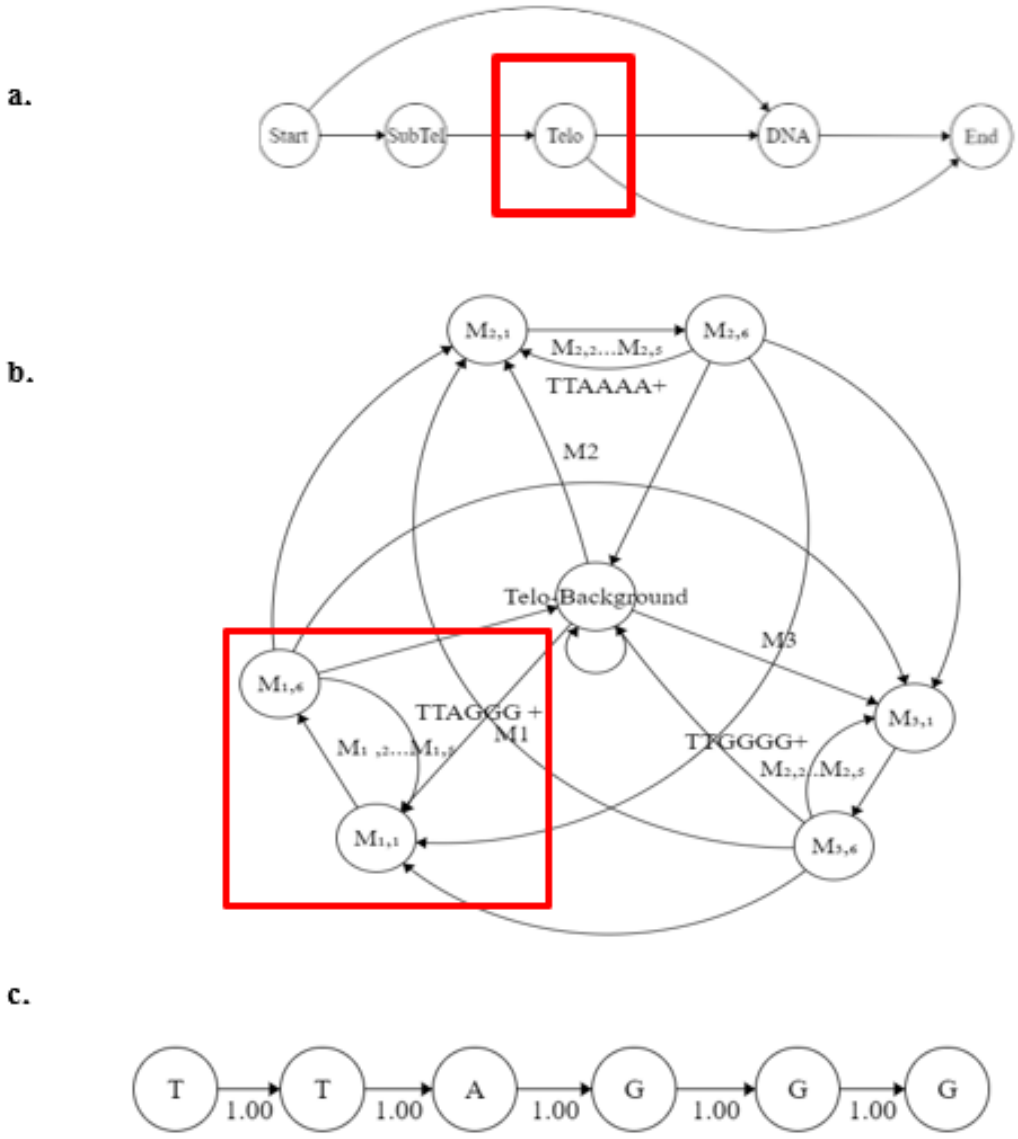


Fig 4. Algorithm Scheme – Division of DNA into states

a. The DNA could be in any state either a chromosomal DNA, a sub-telomere, or a telomere (a collection of states - **b**). The differentiation between DNA to sub-telomere was performed to understand using EM if the bases distributions is different between them. The transition from telomere to DNA is exist because there are observations of telomeric repeats inside the chromosomes, and not only at the edges [13]. Part of our goals is understanding the rate of the phenomenon. Low probability from start to Sub-telomere. Low probability from telomere to DNA.

b. the states combining the Telomere. Any base could be part of a telomeric motif (a collection of states - **c**), or "background" between them – the telomeres are not perfect, and random bases between motifs occur many times, from different reasons. Every motif received certain transition probability from the background, where the primary motif TTAGGG receive higher probability.

c. every base in the motif is represented by its own state. Transition probability from one base state to its sequential is close to 1. The bases emission probabilities in every state is 0.92 to the current base, and 0.08 divided to the other 3 (Nanopore has 8% mistake chance). You can start and end every motif from every other motif or background.

	DNA	subTelomere	T	T	A	G	G	G
A	0.30000	0.30000	0.02667	0.02667	0.92000	0.02667	0.02667	0.02667
C	0.20000	0.20000	0.02667	0.02667	0.02667	0.02667	0.02667	0.02667
G	0.20000	0.20000	0.02667	0.02667	0.02667	0.92000	0.92000	0.92000
T	0.30000	0.30000	0.92000	0.92000	0.02667	0.02667	0.02667	0.02667

Fig 5. Algorithm emission state *e*

Columns represent states, rows represents events – 4 DNA bases. The distribution between the bases is different per state. This is only partial matrix, only 1 motif - TTAGGG

After every sequence was annotated with its suspected states using the posterior forward-backward dynamic algorithm, we divided the sequences into groups, whether they contain a telomere. For the sequences containing telomeres, we can define the limit between the telomere and the sub-telomere using the different states definitions (Fig. 6).

The algorithm receives as input the emission & transition probabilities, and the telomeric motifs to look for. As for the motifs, we were looking for TTAGGG known motif, in addition for two variations we were working on in the lab – TTAAAA, TTGGGG. Of course, we can look for any other motif as well – today there are evidence for different telomeric patterns [14]. The input transition parameters are described in the *supplementary materials* section.

The emissions used for the telomere background, sub-telomere and normal DNA states was {A: 0.3, C: 0.2, G: 0.2, T: 0.3} according to latest statistics regarding human genome [15].

One of the reasons that led us to the course of HMM for telomere detection is that a lot is unknown on telomeres exact sequence, and there is a need for a flexible model, probabilistic model, that will enable errors, modifications, and new discoveries regarding telomeric variants. The HMM works great for that purpose, since even if a telomere contains some patterns which do not fit perfectly to our model (Fig. 4), as long as the region is indeed telomeric, and contain enough evidence for that – the best log likelihood will be achieved with telomere states as the hidden states. In addition, the Nanopore is known to have a relatively high error rate (~8%), and this error rate is implemented as part of the emission probabilities, as described above. Furthermore, the HMM model enable you to capture in full the different possibilities of the telomere, and establish multiple thresholds using the different variety of probabilities describing the in-telomere structures. Using these probabilities correctly will help design exactly what we are looking for and will reveal additional patterns we were not aware of.

Telomere length & alignment to the human genome

Once the optimal Markov states path is in place, we define boundaries of telomere start and end point inside the sequence (Fig. 6). This is extremely important for 3 reasons – First, discovery of the actual telomere length, without his sub-telomere, as telomeres length determine cell fate [4]. Second, detecting sequences with telomers which their 3' end isn't telomeric. These sequences are not telomeres, but Interstitial Telomeric Sequences (ITS). ITS are important genomic elements as they confer its karyotype plasticity [13].
Third – after defining the telomere boundaries, we collect from the sequence its non telomeric parts and use them for genome alignment. The alignment is not possible for the actual telomeres, as they are combined from repetitive sequence, and were never fully sequenced and indexed in human genome DB (till today). We cannot use the telomeres, but we can use the sub-telomeric part to try and align to a specific locus at the genome, or for ITS case use the pre

and post chromosomal segments. The alignment part is important - it will help us understand the telomeric variability between the different chromosomes ends and allow us to specify the regions where ITS are located. The alignment to genome is done using python *mappy* package, additional details in *supplementary materials* section.

Fig 6. Algorithm output – detected telomere
An example to one of the detected telomeres. The underlined bases indicate we are at a telomeric region. Each colour represents one of the three motifs we search for. Easy to see the TTAGGG marked yellow. Any bold motif is a motif contain some error.
We can see there is a clear separation between sub-telomeric part and the telomere itself. Also, we can see this is not a text-book telomere, but rather one containing multiple patterns, and not organized perfectly. These are the patterns we wish to learn more about

Inter Telomere alignment

One of our main challenges in telomere mapping using Nanopore is the fact that we don't know which parts of the sequence are exactly as they appear in the genome, and which parts changed due to nanopore error rate, or other modifications during sample processing and

enrichment. Therefore, we align the found telomeres to each other, and try to find matching between them. A high-quality match in telomere and sub-telomere will indicate the two sequences originated from the same genomic location and would verify any telomeric variation findings we can see for that sequence.
The alignment is done using *mappy*.

Baum-Welch EM

The maximization is part of the algorithm process – run once with it, fix probabilities input if needed, and activate the posterior section. By improving emissions & transitions probabilities e, τ we achieve 2 goals – get better results with our algorithm, since log likelihood will increase (need to make sure not to follow blindly, since the learning mechanism might overfit to this specific telomeric set, and not necessarily aim to our overall goal).
In addition, different conclusions can be inferred from the data – transitions probabilities between different motifs, chance to block of motifs – transition from motif to itself, the probability for each stand of every motif (perhaps some bases are less strict than the others).

Results

The algorithm ran on nanopore data from different samples from Tzfati lab. All the reads we are working on were gathered from patients suffering from Hoyeraal–Hreidarsson syndrome [16], which is caused by mutations in the RTEL1 gene. Some of the samples were transfected with a lentiviral vector expressing WT RTEL1 under an inducible (TET-on) promoter of doxycycline, which rescued the cells and suppressed the telomere defects.
Two main methods were performed to enrich the amount of telomers in the sample. One, using Cas9 to cut the DNA in the sub-telomeric region and by that enabling not too long reads containing telomers, and differentiate them form the rest of the genome. Second approach was to dissect the DNA with restrictions enzymes too small pieces, remove these pieces with beads, and assume that the long segments remained in the samples, along them is the telomeres with no restriction sites. This is a topic for a separate paper – Telomere Enrichment in DNA library preparation.
The purpose of this paper is not to analyse the different telomeres states between these different samples. We only use these nanopore reads as a data set for our algorithm development and assessment.

General Analysis

In total, we have detected **540** suspected telomeres. For the telomere files results see *Supplementary materials*. The average telomere length is **6144**. Keep in mind that big part of the telomeres was not sequenced, and for the sequenced telomeres we can't assure full correction of the algorithm.

From all of the motifs found across all telomeres, **35**% were TTAGGG, **20**% were TTAAAA and 4% were TTGGGG. We also found **9**% of **TTAGGA** – a motif we did not look for initially, but was similar enough to the others, and was detected multiple times. Further research needs to determine whether it is an actual motif or sequencing errors, but it provides a very good lead. The rest of the motifs divided into groups of 1% or less.
In fig. 7 we can see the telomere and sub-telomeres length distribution. As can be seen, we succeeded in mapping some long telomeres, 25k long – as a reminder, there are very few studies that were able to present such significant amount of long telomeres sequences.

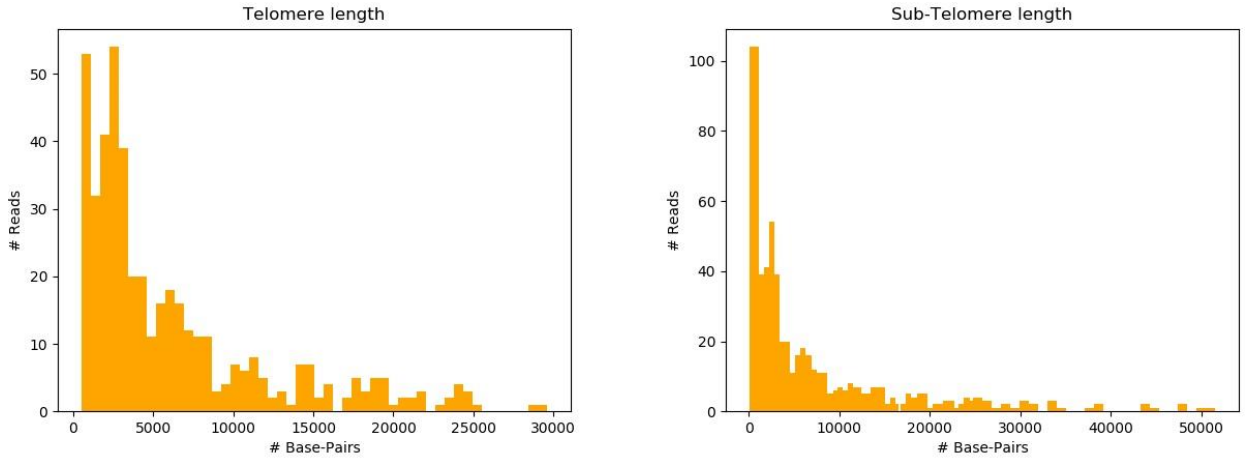


Fig 7. Telomere and sub-telomeres length distribution

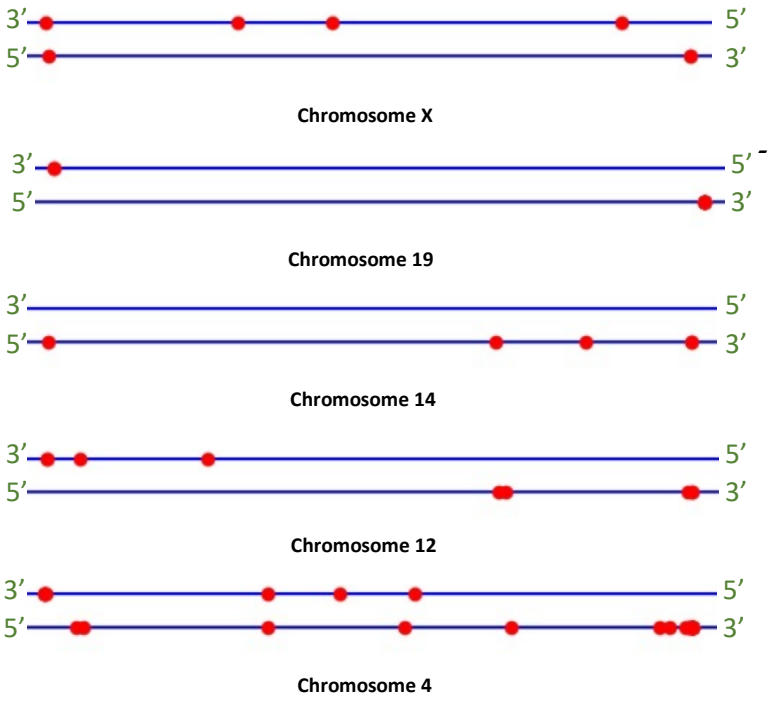
Alignment to Genome

231 telomeres out of these 540 were able to successfully align to the human genome reference using non-telomeric parts of the sequence (Fig. 8). This was one of the research main goals, to match certain telomere patterns to locations of the human genome. Further studies could produce significant research using such data.
As can be seen in Fig. 8, many telomeric sequences are not derived from the chromosomes ends, but from inner chromosomal parts [13]. These are the Interstitial Telomeric Sequences (ITS) mentioned earlier and examining their exact sequence and location could be highly interesting. An interesting statistic that we have discovered – the average telomere length deriving from the ends is **7225**, where the average length of ITS is **2826**. That in fact is evidence that our alignment process using the sub-telomeric parts is working – telomeric sequences on the edges are long, as ITS are much shorter.
In addition, while looking at the table of Fig. 8, There are almost no telomeres derived from the 5' end of the chromosomes (**127** from 3' against **3** from 5') - That is aligned with our expectations. As part of the library processing for the Nanopore sequencing, we mix the DNA sample with adapters [17]. These adapters attach to a double-stranded DNA end. When arriving to the pore, a helicase unwinds the two strands. Out of the 2 strands attached to the adaptor, the 5' will go through sequencing, while the 3' will be released. Our working

hypothesis is that due to the telomeric 3' overhang, the adaptor won't succeed in attaching to the telomeric end in a stable position. It will only attach to an internal end (produced by sample pre-processing), which will lead only the 5′ → 3′ strand to get sequenced.

Inter-telomere alignment

Multiple Telomeres (~50) were able to align between the telomere set, however, most of these alignments were found on the telomeric repeats, with gaps that led us believe these are not true alignments, rather then a partial match in the repeating part of the telomere – not a strong enough alignment to claim they derive from the same location in genome. For these results please look in the *supplementary materials* section. However, an example for a successful alignment is presented in Fig. 9. Unfortunately, we need more data to verify a sequence without sequencing errors.



	Strand +1			Strand -1		
Chr.	5'	3'	Not Edges	'5	3'	Not Edges
1	0	1	3	0	8	6
2	0	10	4	0	2	5
3	0	1	0	0	1	6
4	0	8	11	0	5	3
5	1	3	0	0	3	2
6	0	0	2	0	0	2
7	0	3	2	0	1	2
8	0	2	1	0	3	2
9	0	0	5	0	6	2
10	0	5	1	0	2	2
11	0	2	2	0	2	0
12	0	3	2	0	0	4
13	0	3	1	0	0	5
14	0	2	3	0	0	0
15	0	0	4	0	0	3
16	0	2	1	0	4	0
17	0	1	0	0	2	2
18	0	2	2	0	11	2
19	0	6	0	0	2	0
20	0	2	0	1	0	0
21	0	11	3	0	0	0
22	0	2	3	0	0	0
X	1	2	0	0	1	3
Y	0	2	0	0	1	0

Fig 8. Telomere alignment to human genome
On left – graphic examples to locations on different chromosomes. We chose random chromosomes to present. The chromosomes are normalized such that we see the entire length with fixed width. Many dots contain multiple reads, too small resolution to notice. **On the right,** a table describing all the locations of the aligned telomeres.

EM

As explained, we used the EM algorithm both as means to improve our telomere detection results and increase the sequences log likelihood, and as a tool to better understand the telomere dynamics. The tool worked great, and the log likelihood of the samples only increased with every iteration. Let us examine some of the telomeric features could be learned. Full results could be found at the *supplementary materials* section.

For the primary motif TT**A**GG**G** – seems as the two bold bases tend to switch sometimes – $P(A) = 0.92 -> \mathbf{0.88}, P(G_3) = 0.92 \rightarrow \mathbf{0.78}$, where for the rest of the bases the lowest probability is $\mathbf{0.97}$. For the TTAAAA motif we received only evidence to support its existence – all bases received probabilities $0.92 \rightarrow \sim\mathbf{0.94}$. The TTGGGG motif has not shown any strong evidence, most bases probability dropped $0.92 \rightarrow \sim\mathbf{0.65}$.

Probability for primary motif TTAGGG has dropped $0.8 \rightarrow 0.\mathbf{59}$.

Probability to exit from motif to telomere 'background' increased $0.25 \rightarrow \mathbf{0.37}$. the meaning of that is that the telomeres contains a lot of unordered bases in-between its motifs repeats.

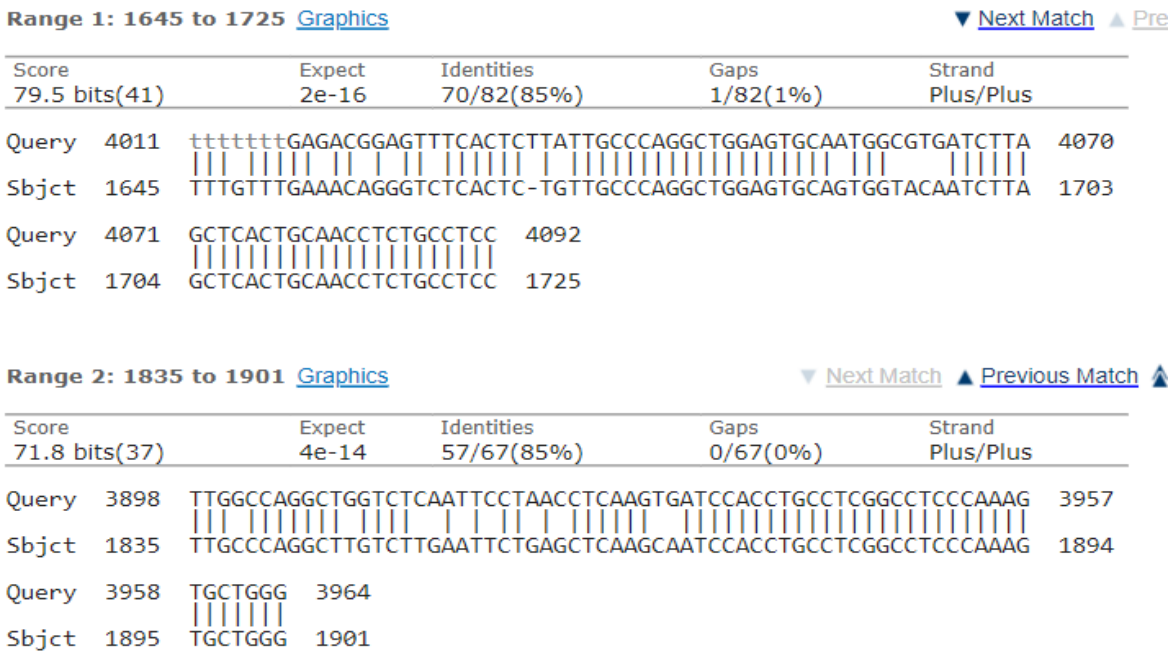


Fig 9. Partial alignment between two telomeres reads
The alignment was found in the algorithm and is graphically shown through Blast [https://blast.ncbi.nlm.nih.gov/Blast.cgi]. We received two significant alignments in the non-repetitive part of the sequence (sub-telomere) – a head start to verify this telomere sequence.

Discussion

We successfully developed a computational tool to help detect, explore, and analyse the telomeres from a Nanopore sequencing run. Many telomeres were detected and aligned to the genome successfully – both from chromosomes ends and ITS. Long reads were found, and reads from variant telomeric 'families' – different motifs, order etc. All is in our results data in *supplementary materials*.

A great achievement of this tool is the ability to define boundaries to the telomere location in the sequence and use the non-telomeric parts for alignment to genome purposes. Also, its

ability to overcome local patterns errors and modifications makes it the perfect fir for the telomere detection problem.

Two main problems raised along our working process – as this is a classification problem, how to verify the detected telomeres (false positive)? How to make sure we did not miss any telomeres (false negative)?
It is crucial to keep in mind that there is no universal truth regarding telomeric identity. As explained before, we are in breaking grounds here, and no current telomeric sequencing is exist. No data base to compare to. That why there is no one answer regarding how many telomeres are verified.
Along the process, we used mainly manual work to go over the results and rule-out false cases, along with different automated analysis to rule out certain groups. For example, we discovered a wide phenomenon of short (~10-20) repeats of different variations of TTAAAA, but certainly not telomeric. There was no organization or pattern, but it confused the algorithm.
we can say that we trust that above 90% of our detections are indeed telomeres. Again, there is no clear method to determine a specific number.
Regarding false negative – there is no manual ways to go over all the not detected sequences. Millions and millions of reads. We tried to perform as many analyses as we could to try and detect more telomeric suspicious groups and try to adjust the algorithm to include them. Again, there is no guarantee, but we are certain that this tool is a great way to start exploring these areas – and discover more features of telomeres that will help to continue enhancing the tool down the road.

There is still a long way ahead – the tool still needs to successfully align the telomeres to themselves, to verify the telomeric sequences. Also, a classification option needs to be added to divide the telomeres to ITS, Telomeres, Telomeres by type and sequence, and more.

In addition, the tool we applied here could be used to find other featured areas in the genomes, such as centromeres, CVN's (copy number variations). It represents the known approach of HMM but exploit the possibility to use it as a full automaton, with different states 'hierarchies' as shown at the paper.

Supplementary material

The algorithm can be found in full here https://github.com/lappazos/Telomere_Project.
Every comment or improvement can be uploaded as an issue in GitHub.
The full data results could be found here. The raw data (pre algorithm process) could be found here.

How to read the results:

- Inside any folder you will find the folder Docs - containing all telomeres found.
- Inside any folder, you will find two images with Telos and Sub-telos length distribution. In addition, you will find image per chromosome showing the telomere alignment to genome.
- Inside any folder, you will find a Slurm.txt file - VERY IMPORTANT FILE. In it you'll find 3 sections - Telo_Analyzer, Telomere_DNA_Aligner, Inter_Telo_Aligner.
  In Telo_Analyzer, you'll find the amount each motif has appeared in all of that folder telomeres, and next to amount - percentage.
  In DNA_Aligner, you will find reads which successfully aligned to the human genome. At the end of this section, very interesting - the distribution of the telomeres in that folder across the different chromosomes. The way to read it is –[strand -1][strand +1] -> [Not at the chromosome ends, Begin , End].
  In Inter_Telo, you'll find Telomeres that were able to get aligned with another telomere in the same folder.
- A motif_profile.txt file at the main folder will present the EM results for the entire set of telomeric data sets.

Transition Probabilities

- P (telomere exist in sequence) = 0.0005
- P (enter telomere from sub-telomere) = 0.1
- P (primary motif TTAGGG) = 0.8
- P (same motif block) = 0.65
- P (exit from motif to telomere background) = 0.25
- P (telomere background to motif) = 0.6

Alignment details

The Alignment was performed using python Mappy package. Preset argument was set to '**map-ont**' (Nanopore data). The reference used was **GRCh38_latest_genomic** [https://www.ncbi.nlm.nih.gov/genome/guide/human/]. For the inter-telomeric alignment, the reference was one telomere, and the others mapped to it.

References

[1] Lansdorp, P. M., N. P. Verwoerd, F. M. van de Rijke, V. Dragowska, M. T. Little, R. W. Dirks, A. K. Raap, and H. J. Tanke. 1996. "Heterogeneity in Telomere Length of Human Chromosomes." *Human Molecular Genetics* 5 (5): 685–91.

[2] Aubert, Geraldine, Mark Hills, and Peter M. Lansdorp. 2012. "Telomere Length Measurement-Caveats and a Critical Assessment of the Available Technologies and Tools." *Mutation Research* 730 (1-2): 59–67.

[3] Harley, C. B., A. B. Futcher, and C. W. Greider. 1990. "Telomeres Shorten during Ageing of Human Fibroblasts." *Nature* 345 (6274): 458–60.

[4] Hemann, M. T., M. A. Strong, L. Y. Hao, and C. W. Greider. 2001. "The Shortest Telomere, Not Average Telomere Length, Is Critical for Cell Viability and Chromosome Stability." *Cell* 107 (1): 67–77.

[5] Feng, Yanxiao, Yuechuan Zhang, Cuifeng Ying, Deqiang Wang, and Chunlei Du. 2015. "Nanopore-Based Fourth-Generation DNA Sequencing Technology." *Genomics, Proteomics & Bioinformatics* 13 (1): 4–16.

[6] Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2019. "Telomere-to-Telomere Assembly of a Complete Human X Chromosome." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/735928.

[7] Mor, Bhavya, Sunita Garhwal, and Ajay Kumar. 2020. "A Systematic Review of Hidden Markov Models and Their Applications." *Archives of Computational Methods in Engineering. State of the Art Reviews*, May. https://doi.org/10.1007/s11831-020-09422-4.

[8] Forney, G. D. 1973. "The Viterbi Algorithm." *Proceedings of the IEEE* 61 (3): 268–78.

[9] Krogh, A. 1997. "Two Methods for Improving Performance of an HMM and Their Application for Gene Finding." *Proceedings /... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 5: 179–86.

[10] Fariselli, Piero, Pier Luigi Martelli, and Rita Casadio. 2005. "The Posterior-Viterbi: A New Decoding Algorithm for Hidden Markov Models." *arXiv [q-bio.BM]*. arXiv. http://arxiv.org/abs/q-bio/0501006.

[11] Fariselli, Piero, Pier Luigi Martelli, and Rita Casadio. 2005. "A New Decoding Algorithm for Hidden Markov Models Improves the Prediction of the Topology of All-Beta Membrane Proteins." *BMC Bioinformatics* 6 Suppl 4 (December): S12.

[12] Miklós, István, and Irmtraud M. Meyer. 2005. "A Linear Memory Algorithm for Baum-Welch Training." *BMC Bioinformatics* 6 (September): 231.

[13] Aksenova, Anna Y., and Sergei M. Mirkin. 2019. "At the Beginning of the End and in the Middle of the Beginning: Structure and Maintenance of Telomeric DNA Repeats and Interstitial Telomeric Sequences." *Genes* 10 (2). https://doi.org/10.3390/genes10020118.

[14] Alaguponniah, Sathyalakshmi, Deepa Velayudhan Krishna, Sayan Paul, Johnson Retnaraj Samuel Selvan Christyraj, Krishnan Nallaperumal, and Sudhakar Sivasubramaniam. 2020. "Finding of Novel Telomeric Repeats and Their Distribution in the Human Genome." *Genomics* 112 (5): 3565–70.

[15] Li, Wentian. 2011. "On Parameters of the Human Genome." *Journal of Theoretical Biology* 288 (November): 92–104.

[16] Deng, Zhong, Galina Glousker, Aliah Molczan, Alan J. Fox, Noa Lamm, Jayaraju Dheekollu, Orr-El Weizman, et al. 2013. "Inherited Mutations in the Helicase RTEL1 Cause Telomere Dysfunction and Hoyeraal-Hreidarsson Syndrome." *Proceedings of the National Academy of Sciences of the United States of America* 110 (36): E3408–16.

[17] **https://nanoporetech.com/how-it-works**