Long read sequencing and local graph assembly reveal heterogeneity of telomeres

Grigorev K, Foox J, Bezdan D, Butler D, Mason C
Institute for Computational Biomedicine, Weill Cornell Medicine

Abstract

Telomeres are regions of repetitive nucleotide sequences capping eukaryotic chromosomes that protect the ends of chromosomes from deterioration. Telomeres are known to generally shorten after each cell replication, eventually blocking somatic cell division and preventing genomic instability. As such, their length is an important marker in senescence, where it can inversely correlate with a subject's age, and in cancers, where both telomere shortening and unrestricted elongation have been suggested as risk factors.

Given their length and repetitive nature, telomeric regions cannot be veritably assembled from reads of kilobase-order lengths (Sanger sequencing) or sub-1Kbp reads (second-generation sequencing), making telomere resolution a very costly and generally intractable problem. Recently, with third-generation technologies like SMRT and nanopore sequencing attaining read lengths on the order of tens and hundreds kilobase pairs, with the longest reads reported as 2Mbp, it became possible to routinely read into the telomeric regions and inspect their structure and length.

We describe a framework for extracting telomeric reads from third-generation sequencing experiments and describing their sequence content and prevalent motifs. We find that human telomeric sequences exhibit surprising heterogeneity, suggesting the possibility of localization of previously reported non-canonical motifs as well as novel sequences.

We also propose a local graph assembly algorithm capable of describing the haplotypic diversity of telomeres. Given the lower complexity of such reads, established methods for long read overlap and assembly that rely on MinHash sketches and minimizers are unsuitable for this problem and fail to detect most correct overlaps when compared to the computationally prohibitive, but mathematically correct Smith-Waterman alignment. We implement a modified method relying on *unimizers* (minimizers occurring once within a given read) that improves overlaps and reduces complexity of the assembly graph, and locally assemble branching telomeric sequences using the computationally efficient A-Bruijn structure.

Background

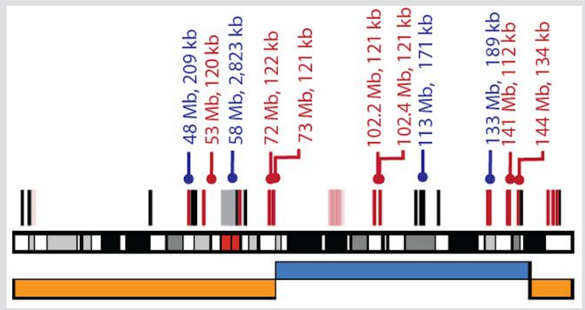

Figure 1: Telomere-to-telomere sequencing of the human X chromosome
(adapted from *Miga et al., 2019* [1]).

Figure 2: Sequencing of human subtelomeres (adapted from *Stong et al., 2014* [2]).

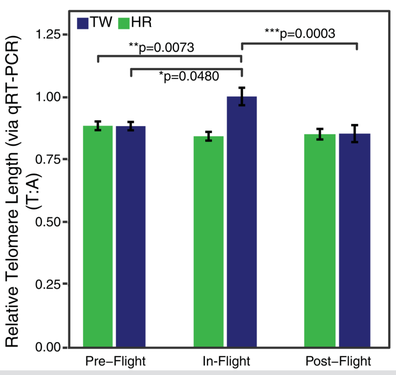

Figure 3: Elongation of telomeres observed during space flight in the NASA Twins Study [3].

Methods

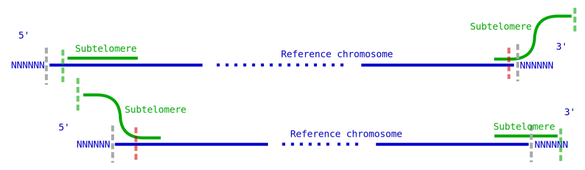

Figure 4: Construction of the extended reference. Subtelomeric sequences were mapped to the hg38 reference and added to the index according to their relationship with the chromosomes.
Annotated: hard-masked (unsequenced) regions (gray dashed line), boundaries between the subtelomere and the telomere (green), origins of divergent sequences (red).

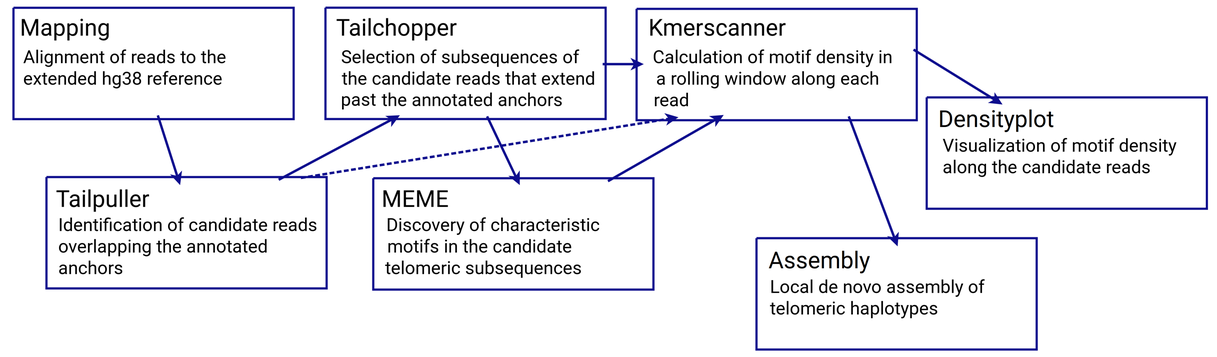

Figure 5: The edgeCase pipeline

Results

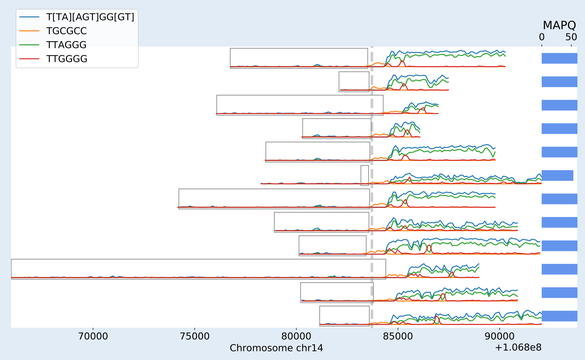

Figure 6: Distribution of major motifs in the telomeric reads from the SMRT hg002 Genome In a Bottle [4] sample.
The density plots suggest that two haplotypes are present, with a shorter and a longer distance between the non-canonical TTGGGG runs. The similarity of overall distributions suggests that the deviation from the canonical motif is unlikely to be due to random sequencing errors.
Local assembly is required to describe these haplotypes.

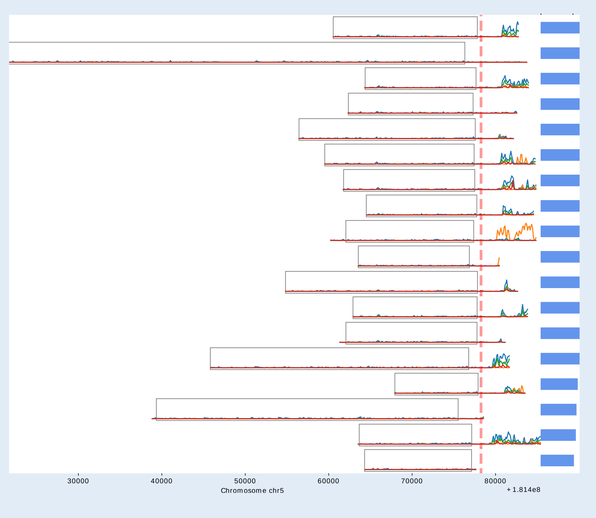

Figure 7: Distribution of major motifs in the telomeric reads from an internal nanopore sequencing dataset, basecalled with *guppy* without the flip-flop model.
The offending motif (orange) is GAA.

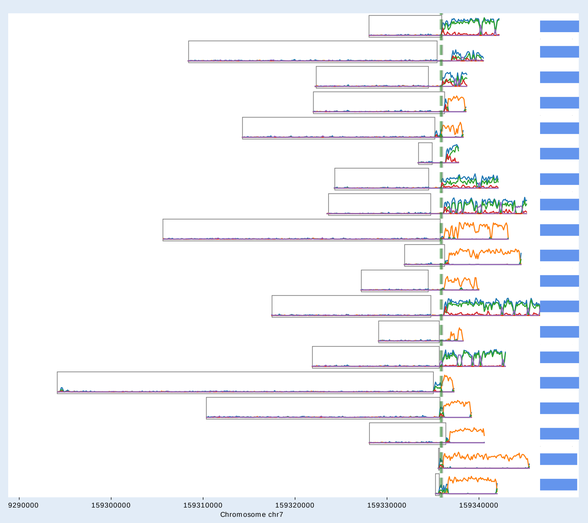

Figure 8: Distribution of major motifs in the telomeric reads from an internal nanopore sequencing dataset, basecalled with *guppy* **with** the flip-flop model. The main offending motif (orange) is CCAGG. The additional offending motif (purple) is TTAAAA. The same motifs are present in Genome In a Bottle nanopore data.

Further work

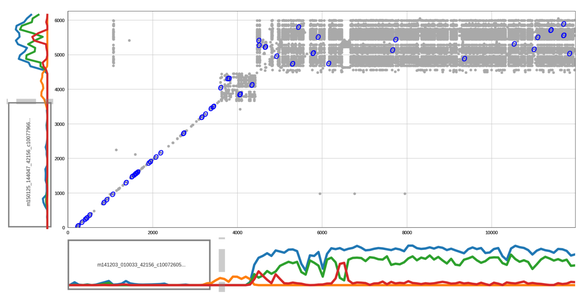

Figure 10: Comparison between *minimizers* (gray) and ***unimizers*** (blue), both with k=16 and w=11.
Unimizers are minimizers that occur only once per given read. The same values of k and w in this figure are used for a strict comparison; more permissive values result in a bigger number of unimizers, especially useful for low-compexity regions.
Arrays of unimizers have an advantage of ***directionality*** compared to minimizers and can identify overlaps of low-complexity reads
while minimizers fail to do so.

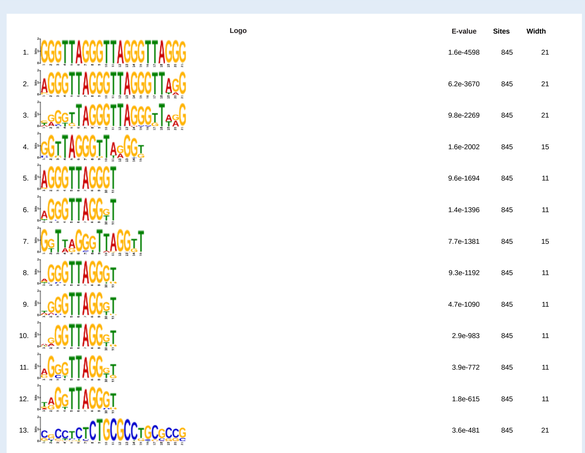

Figure 9: Motifs in segments of SMRT reads extending past the annotated subtelomere.
Motif discovery was performed with MEME [5].

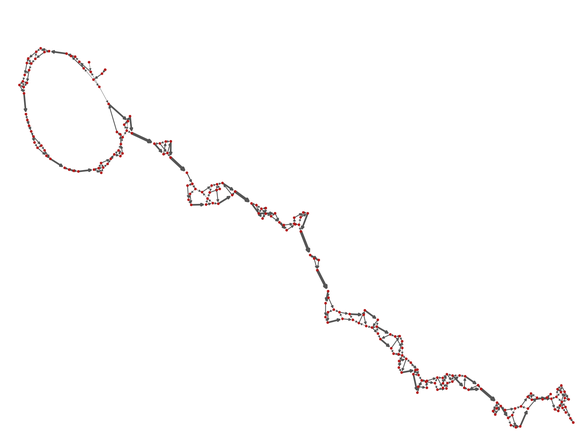

Figure 11: The directionality of unimizers allows to more easily construct de Bruijn-like graphs (A-Bruijn graphs originally introduced in a different framework [6]).

References

1. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome (Aug. 2019).
2. Stong, N. *et al.* Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline. *Genome Research* **24** (Mar. 2014).
3. Garrett-Bakelman, F. E. *et al.* The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **364** (2019).
4. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data* **3** (June 2016).
5. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37** (May 2009).
6. Pevzner, P. A. De Novo Repeat Classification and Fragment Assembly. *Genome Research* **14** (Sept. 2004).

Acknowledgements

Susan Bailey (Colorado State University)