NPGREAT: Hybrid Assembly of Human Subtelomeres with the use of Nanopore and Linked-Read datasets

Eleni Adam*, Desh Ranjan*, Harold Riethman†
**Department of Computer Science, Old Dominion University*
†School of Medical Diagnostic & Translational Sciences, Old Dominion University

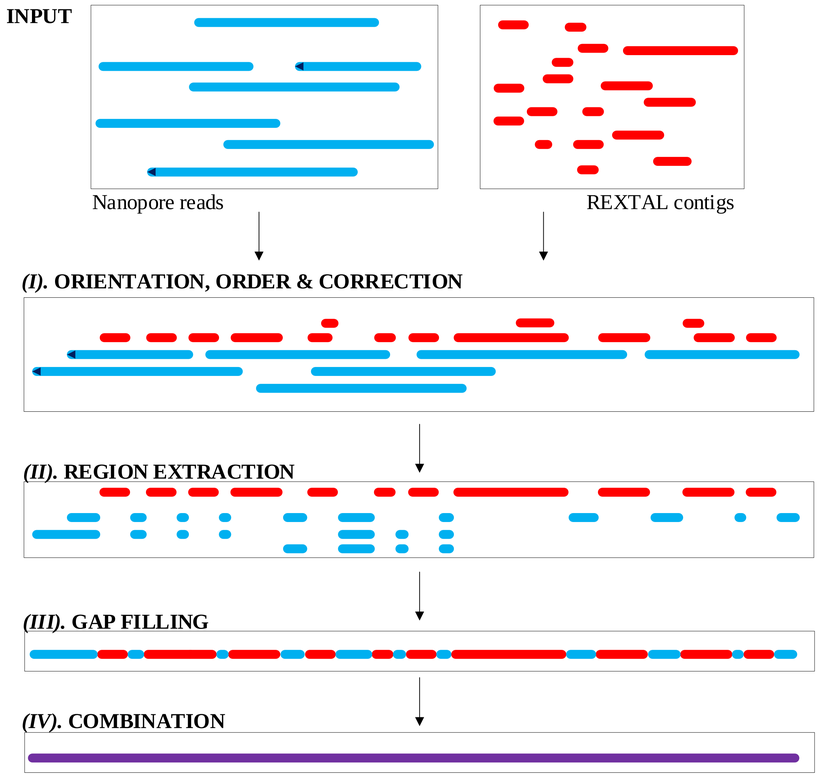
INTRODUCTION

- The **telomeres** are located at the tips of the chromosomes and have the critical role of protecting them. The **subtelomeres**, located next to the telomeres, are vital in regulating the adjacent telomere lengths.
- Age-related diseases, including **cancer** occur due to telomere dysfunction, which is caused by length shortening or other types of telomere rearrangement.
- Subtelomeric regions are hard to investigate due to:
  - The absence of a technique to obtain the entire subtelomere region directly, since the DNA sequence is given in multiple ***pieces***.
  - The difficulty in assembling those pieces accurately due to the **repetitive structure** of the region and the **quality** of the *pieces*.
  - The **high variability** of the subtelomeric regions between different people.

Aim: A method that accurately assembles the DNA *pieces* to obtain the human subtelomeric regions.

METHOD

The NanoPore Guided Regional Assembly Tool (**NPGREAT**) utilizes two of the latest available types of data (*pieces*), which complement each other: **Linked-Reads** and **ultralong Nanopore reads**. Initially the adjacent single-copy region of a telomere is used to select the linked-read and nanopore read data that correspond to the subtelomere region in question. Then the REXTAL computational method is used to create the set of short-read assemblies derived from the selected linked-reads. The selected nanopore reads (*color blue*) and the contigs of the REXTAL short-read assemblies (*color red*) constitute the input data of the **NPGREAT** method.

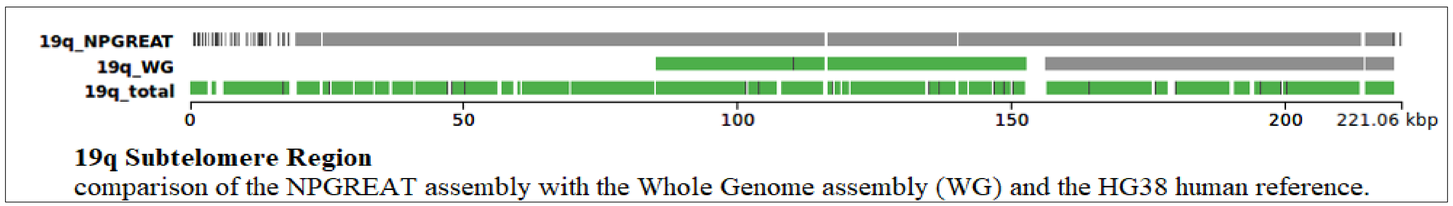


The NPGREAT consists of four main steps:

i. The **Orientation, Order and Correction** of the short *pieces (color red)* is obtained by using the long *pieces (color blue)* as scaffolds, upon which the short *pieces* are mapped to.
ii. In the **Region Extraction**, the segments of the multiple long *pieces* that can be used to connect the short *pieces*, are extracted.
iii. In the **Gap Filling** step, all possible segments are taken into account and one is selected to fill each gap.
iv. In the **Combination** step, the corrected short *pieces* are combined with the connector segments. The output is the subtelomere region of the chromosome (*purple color*).

RESULTS

We tested NPGREAT on the NA12878 human cell line. The output assemblies are of high percent identity with the hg38 reference, with differences only in the variable tandem-repeat regions of the sequence. The hybrid NPGREAT method provides for the first time the high quality continuous assembly of human subtelomeric regions.



19q Subtelomere Region
comparison of the NPGREAT assembly with the Whole Genome assembly (WG) and the HG38 human reference.

Contact Email: eadam002@odu.edu