Group Members: JooYoung Gwag, Karen Wang, Keerthana Ande

Project Title: Comparative Analysis of Telomeric Regions in GRCh38 and T2T-CHM13 Genome Assemblies

Data Source: The data used in this project is sourced from the UCSC Genome Browser. Specifically, we downloaded chromosome sequences in FASTA format from both the GRCh38 and the T2T-CHM13 genome assemblies.

Problem Statement: The GRCh38 human reference genome, despite being widely used in genetic studies and medical diagnostics, contains unresolved regions—particularly at the telomeric ends of chromosomes—where the sequence is represented by long stretches of unknown bases denoted as "N." These unresolved sequences pose a risk of incorrect genetic interpretation or medical diagnostics.

The newer T2T-CHM13 assembly addresses this issue by fully sequencing previously unresolved regions, including the telomeres, and providing accurate telomeric repeat arrays (typically TTAGGG). Despite this improvement, the Genome Browser does not highlight the differences explicitly, making it challenging and time-consuming for researchers to manually identify and compare the changes between the two assemblies.

Solution Overview: To address the issue, our project will focus on comparing the telomeric regions of a selected chromosome (e.g., Chromosome X) from the GRCh38 and T2T-CHM13 assemblies. The aim is to:

1. Identify and measure gaps (i.e., runs of 'N') in the GRCh38 telomeric regions.
2. Detect and count TTAGGG repeats in the corresponding regions of the T2T-CHM13 assembly.
3. Compare these metrics to determine whether the gaps in GRCh38 were fully resolved in the T2T assembly and estimate actual telomere length.

Technical Approach (Code Design):

1. **Data Parsing:**
    - Read the FASTA files containing chromosome sequences from both genome assemblies using a custom parser or Biopython.
2. **Telomere Detection Logic:**
    - For GRCh38:
        - Search the beginning and end of the sequence for long stretches of 'N'.
        - Count the length of these stretches to estimate the gap size.
    - For T2T-CHM13:
        - Search for continuous runs of the telomeric repeat "TTAGGG" (and its reverse complement on the opposite end).
        - Count the number of repeats to estimate telomere length.
3. **Comparison Algorithm:**

    - Compare the length of telomeric repeats in T2T-CHM13 to the gap size in GRCh38.
    - Determine if the gaps are fully resolved and estimate the true length of the telomeres.
4. **Output:**
    - Report the number of 'N's in GRCh38.
    - Report the number of TTAGGG repeats in T2T-CHM13.
    - Display whether the gaps are completely filled and provide a length estimate of the telomeres.

Conclusion: This project will deliver a clear comparison of telomeric regions between GRCh38 and T2T-CHM13, offering insights into how much previously unsequenced telomeric data has been recovered. The methodology is extensible and can be applied to other chromosomes or repeat sequences in future analyses.

Note: We chose to focus on telomeric sequences due to their medical relevance and the striking difference in sequence completeness between the two assemblies. The project simplifies the otherwise complex process of genome comparison using an automated and reproducible computational approach.