Comparative Analysis of Repeat Expansion Callers

Oumnia Chellah, Stuart Scott, Yao Yang

Introduction

Repeat expansions are critical genetic features implicated in a wide array of genetic disorders. The precise detection of these expansions is vital for accurate diagnosis and for advancing our understanding of the underlying biological mechanisms. This research project aims to systematically benchmark the performance of currently available repeat expansion callers on both short and long read sequencing data. By comparing accuracy and efficiency across different types of short tandem repeats (STRs) and evaluating various repeat expansion callers methodologies, our study will identify the strengths and limitations of these tools, ultimately guiding their optimal application in both clinical and research settings.

Goals

- **Compare Performance Across Sequencing Methods:** Evaluate and compare the performance of repeat expansion callers on short read versus long read sequencing data to determine their efficacy and reliability in different sequencing contexts.
- **Analyze Caller Performance for Similar Data:** Conduct a comparative analysis of various repeat expansion callers when applied to similar data sets to identify the most effective tools.
- **Assess Performance Across Different Samples:** Investigate how different repeat expansion callers perform across a diverse range of genetic samples to ensure broad applicability of findings.
- **Identify and Document Shortcomings:** Identify any shortcomings or limitations in
- current repeat expansion detection technologies and document these issues to provide insights for future improvements and research.

Progress

- Created a pipeline to process genomic samples by aligning reads to the GRCh38 reference genome, indexing aligned sequences, and conducting variant calling to identify genetic variations.
- Ran repeat expansion callers TRGT, ExpansionHunter, and STRling, and detected repeat expansions in both short-read and long-read sequencing data of the HG002 genome.
- Conducted initial comparisons between repeat expansion caller results and existing benchmarks.
- Used Truvari to analyze and compare the performance metrics of the repeat expansion callers.
- Developed and executed scripts on Stanford's Sherlock High-Performance Computing (HPC) cluster to perform alignment, repeat expansion calling, and benchmarking procedures.

Bioinformatics tools

- **Repeat Expansion Callers:** TRGT, ExpansionHunter, STRling, LongTR.
- **Comparison algorithms:** Truvari, hap.py.
- **Alignment and Indexing:** minimap2, pbmm2.

Data

- **Sequencing data:**
  - Genome Reference Consortium Human GRCh38
  - WGS of HG002 with PacBio HiFi
  - HG002 Illumina HiSeq 2x250
- Tandem Repeat Benchmark
- **Tandem repeat catalogs:**
  - Project Adotto Tandem-Repeat Regions and Annotations
  - Tandem Repeat Catalog & Variation Clusters