Final Report

Student: Purav Biyani

Mentor: Leanne Haggerty, Thiago Genez, Francesca Tricomi

Project: A Nextflow Pipeline for Repeat Annotation

Code: https://github.com/Ensembl/repeat_nf

GSoC'23 profile:

https://summerofcode.withgoogle.com/programs/2023/projects/aFLAmznr

GSoC'23 proposal:

https://drive.google.com/file/d/17j3x1iiF7lwsjnjeGjd_orcTXO-25VFX/view?usp=sharing

Organization: Genome Assembly and Annotation

Description: The "Nextflow Pipeline for Repeat Annotation" project aimed to redesign and implement a new pipeline for finding and annotating repetitive DNA sequences in eukaryotic genomes using Nextflow. The existing infrastructure, which used eHive, was reaching the end of its life, and this project focused on transitioning the workflow to Nextflow. The pipeline outputs a masked genome sequence incorporating comprehensive annotations for repeats, low complexity regions, and tandem repeats. Furthermore, an annotated repeats file in GTF format provides detailed information about the identified repeat elements.

What work was done:

Throughout the duration of the project, the following tasks were accomplished:

- **Workflow Implementation in Nextflow**: Developed a robust Nextflow pipeline for repetitive element annotation. This involved integrating various components such as Repeatmodeler, Repeatmasker, DUST, and TRF. Ensured proper data flow and dependencies between the processes.
- **Singularity Integration:** Incorporated Singularity support into the pipeline to containerize the required tools, ensuring a consistent and isolated environment for execution.
- **Installation and Usage Documentation**: Created detailed installation and usage instructions for running the pipeline. Documented the steps to install necessary dependencies, download required tools, and execute the pipeline effectively.
- **Testing**: Tested the pipeline locally and on cluster to ensure its functionality.

Commits:

- All the commits to the projects can be found here.

What's left to do:

The project achieved significant milestones, but a few tasks remain:

- **Genome Chunking Implementation**:
    - Implement genome chunking strategy to divide large genomes into smaller, manageable segments.
    - Each genome chunk will be processed independently, allowing for parallel execution and improved performance.

- **RepeatModeler Library Check**:
    - Prior to initiating the RepeatModeler process, implement a library existence check.
    - Verify whether the required RepeatModeler library is available on the designated site.
    - If the RepeatModeler library does not exist, proceed with the RepeatModeler process.
    - If the library is present, we can skip the RepeatModeler step and proceed with the rest of the pipeline.
- **Deployment:**
    - Deploy and validate the pipeline on different cloud platforms to verify its scalability and reliability.

Acknowledgement:

I extend my gratitude to my mentors, Leanne, Thiago and Francesca, for their unwavering guidance and support throughout the project. Their expertise and insights greatly contributed to my learning experience. Working with the Genome Assembly and Annotation community and being part of this project has been an amazing journey into software development and teamwork. The challenges faced and milestones achieved have enhanced my skills and will undoubtedly shape my future endeavors.

Thanks