Long read assembly

1. Create a new filder named "**Long_assembly**" inside the folfer "**Day3**" created by you
2. Copy the mouse_demodata.fastq file to your folder

cp /home/nanobioinfo22/shailesh_nipgr/Whole_genome_assembly/mouse_demodata.fastq ./

3. Move to conda environment assembly1

conda activate assembly1

Quality control:

The quality of the input reads are assessed using Porechop that finds and removes adapters from Oxford Nanopore reads.

porechop -i mouse_demodata.fastq -o mouse_demodata_trim.fastq -t 5 --extra_end_trim 0 --extra_middle_trim_good_side 0 --extra_middle_trim_bad_side 0

Genome Assembly:

Flye is a de novo assembler for single-molecule sequencing reads, such as those produced by PacBio and Oxford Nanopore Technologies. It is designed for a wide range of datasets, from small bacterial projects to large mammalian-scale assemblies. The package represents a complete pipeline: it takes raw PacBio / ONT reads as input and outputs polished contigs. Flye also has a special mode for metagenome assembly.

flye --threads 2 --nano-raw mouse_demodata_trim.fastq --genome-size 2g --out-dir mouse_assembly

The assembly.fasta generated from the above step in the directory mouse_assembly is then utilised to close the gaps emerging during the scaffolding process via TGS-GapCloser, further improving the overall quality. It is a gap-closing software tool that uses error-prone long reads generated by third-generation-sequence techniques (Pacbio, Oxford Nanopore, etc.) or preassembled contigs to fill N-gap in the genome assembly.

tgsgapcloser --scaff mouse_assembly/assembly.fasta --reads mouse_demodata_trim.fastq --output tgs_gapcloser_muslong --racon /home/nanobioinfo22/.conda/envs/assembly1/bin/racon

conda deactivate

To describe the completeness and contiguity of a genome assembly, several summary statistics and in-silico validations are performed. Quast i.e Quality Assessment Tool for Genome Assemblies is one such tool for genome assembly evaluation.

quast.py tgs_gapcloser_muslong.scaff_seqs -t 2 -o tgs_gapcloser_quast

```
CWD: /home/nanobioinfo22/naveen/long_assembly
Main parameters:
  MODE: default, threads: 32, min contig length: 500, min alignment length: 65, min alignment IDY: 95.0, \
  ambiguity: one, min local misassembly length: 200, min extensive misassembly length: 1000

Contigs:
  Pre-processing...
  tgs_gapcloser_muslong.scaff_seqs ==> tgs_gapcloser_muslong.scaff_seqs

2024-05-21 15:25:28
Running Basic statistics processor...
  Contig files:
    tgs_gapcloser_muslong.scaff_seqs
  Calculating N50 and L50...
    tgs_gapcloser_muslong.scaff_seqs, N50 = 23001, L50 = 4, auN = 19717.4, Total length = 209827, GC % = 39.37, # N's per 100 kbp =  0.00
  Drawing Nx plot...
    saved to /home/nanobioinfo22/naveen/long_assembly/tgs_gapcloser_quast/basic_stats/Nx_plot.pdf
  Drawing cumulative plot...
    saved to /home/nanobioinfo22/naveen/long_assembly/tgs_gapcloser_quast/basic_stats/cumulative_plot.pdf
  Drawing GC content plot...
    saved to /home/nanobioinfo22/naveen/long_assembly/tgs_gapcloser_quast/basic_stats/GC_content_plot.pdf
  Drawing tgs_gapcloser_muslong.scaff_seqs GC content plot...
    saved to /home/nanobioinfo22/naveen/long_assembly/tgs_gapcloser_quast/basic_stats/tgs_gapcloser_muslong.scaff_seqs_GC_content_plot.pdf
```

BUSCO (Benchmarking Universal Single-Copy Orthologs) is yet another correctness measure that provides measures for quantitative assessment of genome assembly based on evolutionarily informed expectations of gene content from near-universal single-copy orthologs.

busco -f -i tgs_gapcloser_muslong.scaff_seqs -m genome -l metazoa_odb10 -o output_busco -c 124

```
        ---------------------------------------------------
        |Results from dataset metazoa_odb10               |
        ---------------------------------------------------
        |C:0.0%[S:0.0%,D:0.0%],F:0.0%,M:100.0%,n:954      |
        |0       Complete BUSCOs (C)                      |
        |0       Complete and single-copy BUSCOs (S)      |
        |0       Complete and duplicated BUSCOs (D)       |
        |0       Fragmented BUSCOs (F)                    |
        |954     Missing BUSCOs (M)                       |
        |954     Total BUSCO groups searched              |
        ---------------------------------------------------
```