Long Read Assembly Lab

BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

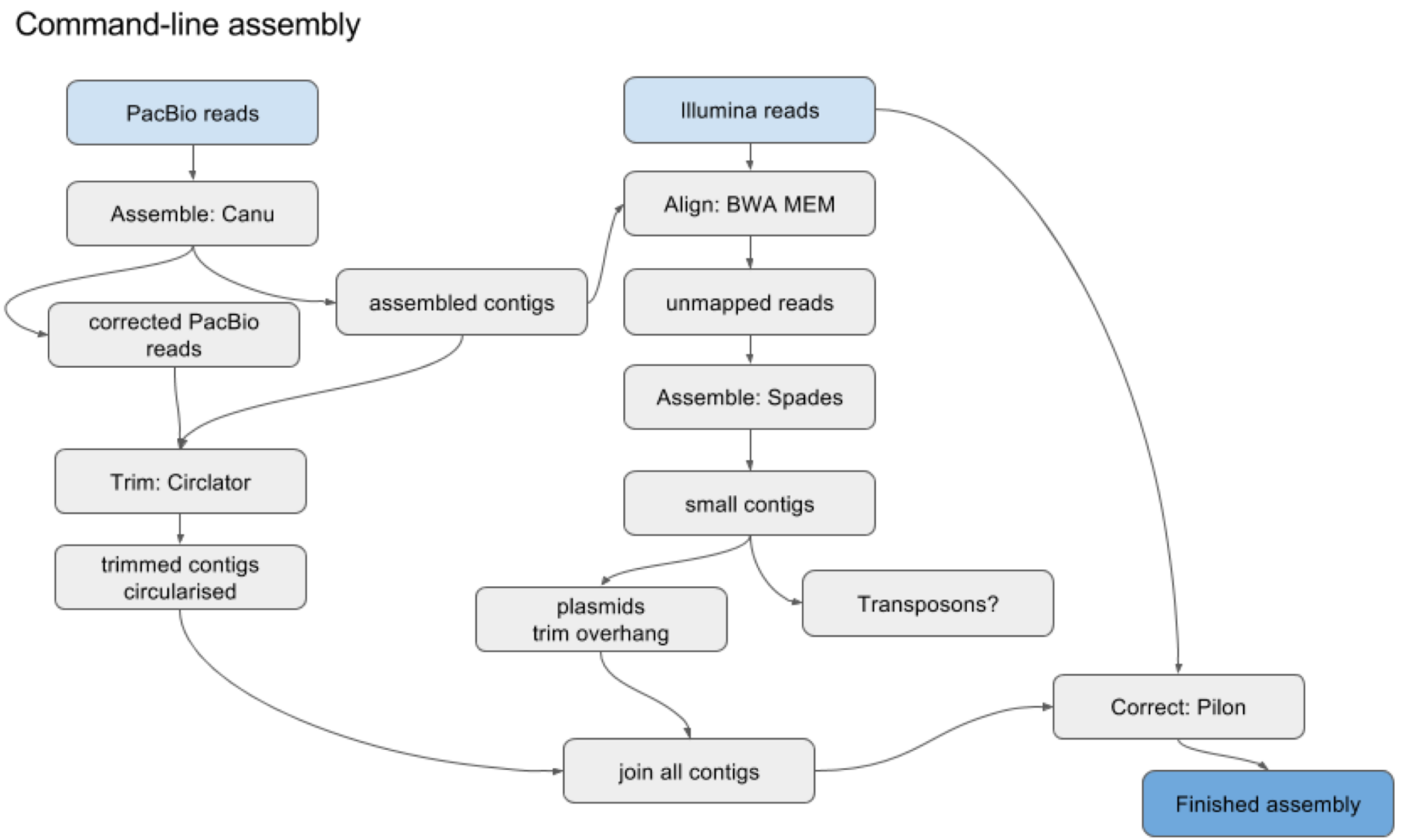
Purpose

- This is a tutorial for long-read (PacBio) genome assembly.

- It demonstrates how to use long PacBio sequencing reads to assemble a bacterial genome, and includes additional steps for circularising, trimming, finding plasmids, and correcting the assembly with short-read Illumina data.

- All resources are from https://www.melbournebioinformatics.org.au/tutorials/tutorials/pacbio/

Workflow Overview



Download Data

- https://doi.org/10.5281/zenodo.1009308
- The downloaded files are located at hopper.slu.edu: /public/ahnt/courses/bcb5250/long_read_assembly_lab/data

Canu Assembler

- Canu install: https://canu.readthedocs.io/en/stable/
- Canu is already installed in my public directory and the binaries are located at hopper.slu.edu: /public/ahnt/courses/bcb5250/long_read_assembly_lab/software/canu/Linux-amd64/bin
  - So, add the path to your environment

Canu, the pipeline

The canu pipeline, that is, what it actually computes, comprises of computing overlaps and processing the overlaps to some result. Each of the three tasks (read correction, read trimming and unitig construction) follow the same pattern:

- Load reads into the read database, gkpStore.
- Compute k-mer counts in preparation for the overlap computation.
- Compute overlaps.
- Load overlaps into the overlap database, ovlStore.
- Do something interesting with the reads and overlaps.
  - The read correction task will replace the original noisy read sequences with consensus sequences computed from overlapping reads.
  - The read trimming task will use overlapping reads to decide what regions of each read are high-quality sequence, and what regions should be trimmed. After trimming, the single largest high-quality chunk of sequence is retained.
  - The unitig construction task finds sets of overlaps that are consistent, and uses those to place reads into a multialignment layout. The layout is then used to generate a consensus sequence for the unitig.

Run Canu

- Run CANU
  - $ canu -p canu -d canu_outdir genomeSize=0.03m corThreads=10 -pacbio-raw ../data/pacbio.fq
    - the first canu tells the program to run
    - -p canu names prefix for output files ("canu")
    - -d canu_outdir names output directory ("canu_outdir")
    - genomeSize only has to be approximate. (In this case we are using a partial genome of expected size 30,000 base pairs).
    - corThreads=10 sets the number of available threads.
    - Canu will correct, trim and assemble the reads.
    - Various output will be displayed on the screen.
    - *Note*: Canu could say "Finished" but may still be running. In this case, type squeue to see if jobs are still running.

Canu Output

- $ cd canu_output
  - The canu.contigs.fasta are the assembled sequences.
  - The canu.unassembled.fasta are the reads that could not be assembled.
  - The canu.correctedReads.fasta.gz are the corrected Pacbio reads that were used in the assembly.
  - The canu.contigs.gfa is the graph of the assembly.
  - The canu.report file is a summary of all of the steps Canu performed with information about the reads used, how they were handled and a whole lot of summary information about the assembly.

Run Quast

- $ quast canu.contigs.fasta

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

06 February 2020, Thursday, 09:46:09

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference	canu.contigs
# contigs	1
# contigs (>= 0 bp)	1
# contigs (>= 1000 bp)	1
# contigs (>= 5000 bp)	1
# contigs (>= 10000 bp)	1
# contigs (>= 25000 bp)	1
# contigs (>= 50000 bp)	0
Largest contig	49 905
Total length	49 905
Total length (>= 0 bp)	49 905
Total length (>= 1000 bp)	49 905
Total length (>= 5000 bp)	49 905
Total length (>= 10000 bp)	49 905
Total length (>= 25000 bp)	49 905
Total length (>= 50000 bp)	0
N50	49 905
N75	49 905
L50	1
L75	1
GC (%)	34.69
Mismatches	
# N's	0
# N's per 100 kbp	0

Additional Work

Follow the steps at https://www.melbournebioinformatics.org.au/tutorials/tutorials/pacbio/

- Trim and circularise
- Find smaller plasmids
- Correct the assembly
- Comparative Genomics