Beverlie Poblete
PID A11334034

CSE185 PROJECT PROPOSAL

Biological Question

Determining the genomic sequences of microorganisms is the basis and prerequisite for understanding their biology and functional characterization. Liao *et al.* have previously stated that many assembly algorithms, while cost-efficient, often result in unfinished, fragmented draft genomes as a result of short read lengths and long repeats present in multiple copies. Hybrid approaches such as ALLPATHS-LG, PacBio corrected reads pipeline, SPAdes, and SSPACE-LongRead along with non-hybrid approaches such as hierarchical genome-assembly process (HGAP) and PacBio corrected reads pipeline via self-correction have been proposed to use the PacBio long reads to facilitate the assembly of complete microbial genomes. In my study, I aim to re-create and validate the results from Liao *et al.*'s study in comparing the assessment of the above-mentioned assembly approaches in assembling a complete microbial genome. While the authors compared almost 7 methods, I aim to assess one hybrid approach and one non-hybrid approach mentioned above. If time persists, I hope to relate and compare the assessment of these two approaches to the approach taken in our week 3 lab tutorial. If time does not allow me to compare both non-hybrid and hybrid approaches, then I will most likely just compare hybrid approach with the approach taken in class.

Dataset

- Dataset 1: E. coli K-12 MG1655 Fragment Reads 2x101 bp, 180bp insert (SRR447685) [1.6GB – fastq] and Jump Reads 2x93 bp, 3000 bp insert (SRR401827 [132.9Mb – whole genome shotgun sequence reads] and SRR492488 [31Mb – Paired reads]), and 1-3Kbp Long Reads (Ribeiro's ftp[a]) [~600 MB]
- Dataset 2: E. Coli K-12 MG1655 Long Reads 10 Kbp, 17 SMRT cell (SRX255228[c]) [814.2Mb – PacBio RS runs]
- Reference: NC_00913 used for both Datsets 1 and 2

Dataset 1 will be used when using ALLPATHS-LG (v44837) and SPAdes (v3.1.0) to assemble the bacterial genome E. coli K-12 MG1655. Dataset 2 will be used when conducted the non-hybrid approach hierarchical genome-assembly process with HGAP(v2.0) to de novo assemble the PacBio long reads. Dataset 2 consists of SMRT cells.

[a] Long reads were downloaded from
ftp://ftp.broadinstitute.org/pub/papers/assembly/Ribeiro2012/data
[c] PacBio HDF5 files were requested from NCBI Sequence Read Archive (SRA)

- However, 8-10 Kbp, 8 SMRT cells were downloaded from http://files.pacb.com/software/hgap/index.html

Publications linked with Data:

1. Dataset 1: https://www.ncbi.nlm.nih.gov/pubmed/22829535
2. Dataset 2: http://cbcb.umd.edu/software/PBcR/

Bioinformatics Pipeline

I aim to closely follow the pipeline as illustrated in the paper as seen in the link: http://sb.nhri.org.tw/comps/en/Home;jsessionid=94F4B777E7FB60EB08202C2FE727ED22. While not all the specific details have been determined yet, I plan to use ALLPATHS-LG (v44837) and SPAdes (v3.1.0) to assemble the E. coli bacterial genomes. Then I plan to conduct non-hybrid approach hierarchical genome-assembly process HGAP (v2.0) to de novo assemble the PacBio long reads. The studies of each method in the paper used different datasets for the comparative assessment of different assembly approaches, therefore my study may also use different datasets for comparative assessment, although I ideally want to compare assemblies using one dataset throughout. After performing non-hybrid and hybrid assembly approaches, I plan to use QUAST along with NCBI reference sequences to assess the quality of assemblies generated by the various approaches. I also plan on generating assembly dot plots for the sake of comparison against the reference genome to evaluate the assembly's accuracy using r2cat program.

Citations

- Main paper:
  - Liao, Y., Lin, S., & Lin, H. (2015). Completing bacterial genome assemblies: Strategy and performance comparisons. *Scientific Reports*, *5*(1). doi:10.1038/srep08747