DNA sequence assembly project

March 26, 2012

Assignment: for an Illumina ecoli dataset, a Roche 454 ecoli dataset and a mixed 454/Illumina Toxoplasma gondii dataset, clean the sequencing reads, assemble the genomes, map the reads back to genomes, find variants (if any), determine closest relative (by sequence similarity), annotate (computationally). Data can be found on the central server:

`/mnt/home/uirig/user_data/BCB504`

Submit one latex/sweave report for the entire project (ie. part 2 should continue part 1, etc.). The report should include full details of the analysis and the results, as if written in a publication. Finally, include a section that details individual contribution to the analysis/report.

Part 1: Clean the reads

Due Wed April 2nd

- Perform Quality Assurance of the reads
- Removing, or trim, poor quality reads
- Use of R/Python, sfffile, sffinfo

Part 2: Assemble the genomes

Due Fri April 13th

- Assembly and optimize the assembly (maximize the metrics)
- What are the details about the assembly (genome size, coverage, illumina insert sizes, etc.)

Part 3: Map the reads, find variants, annotate

Due Fri April 22nd

- Map the reads back to the assembly.
- report the details (% of reads that map).
- are there variants.
- What is the closest public genome.
- annotate the genomes.