BioE 131/231
Fall 2018

Lab Assignment #8

Overview and references

Using SPAdes, you will assemble a bacterial genome *de novo* using a combination of long PacBio reads and short Illumina reads. Next week, you will analyze your genome to determine the species and obtain an overview of its metabolism.

You are expected to keep a thorough record of everything you did in your notebook. Create a folder in your home directory for each lab, and keep all your files there. Try to create a directory hierarchy that makes sense, like the one we went over in Lab 1. Copy and paste any terminal commands you used into a Markdown section and explain what the input was, what the tool did, and what the output was. Plot any results in-line and explain them.

An iPython notebook containing your analysis is due at midnight on Wednesday **two weeks from now**, spanning both this lab session and next week's. You can upload a link to your GitHub repo on bCourses.

Sequence assembly
https://en.wikipedia.org/wiki/Sequence_assembly

SPAdes
http://cab.spbu.ru/files/release3.12.0/manual.html

PacBio SMRT Sequencing
https://en.wikipedia.org/wiki/Single_molecule_real_time_sequencing

Illumina sequencing
https://en.wikipedia.org/wiki/Illumina_dye_sequencing

Background

Genome sequencing and assembly are common techniques in biology. To obtain the sequence of a long genome, DNA must be chopped into small pieces that can be read by a sequencer. These short reads must then be stitched back together to form a complete genome. Often, the genome cannot be fully assembled because there are multiple equally plausible ways of stitching the reads together. Ideally, each chromosome is assembled into a single, long sequence. In practice, chromosomes are often assembled into multiple "contigs," or contiguous sequences. A genome assembly is generally considered complete only when all (or nearly all) the sequences are accounted for. Otherwise, it is considered a draft genome.

In a previous lab, you filtered human reads from a bacterial genome by alignment using Bowtie2. In this lab, you will continue your analysis, albeit with a different set of reads. First, you must take the reads and combine them into a complete genome. Next week, you will analyze the contents of your genome.

Locating the data

DNA from an unknown bacterium was sequenced using PacBio and Illumina technologies. The resulting reads are in **/data/lab8**. You should not have to copy them to your home directory.

illumina_reads_R1.fastq – first paired-end read
illumina_reads_R2.fastq – second paired-end read
pacbio_reads.fastq – long PacBio reads

Running SPAdes

SPAdes is a hybrid genome assembler, meaning that it takes multiple sources of information as input and combines them to produce an optimal assembly. Assemblies using only short reads tend to be highly fragmented (i.e., many contigs). Assemblies using a high-quality short read set and a higher error rate long-read set (like PacBio) tend to be the best.

Why do we expect short reads to produce a more fragmented assembly than long reads?
Why does a single-molecule sequencing like PacBio have a higher error rate than Illumina?

SPAdes must be run from the command line, and can take a while. Please run it inside a "screen" so that it continues running even if you disconnect from the server. Check out this quick tutorial on how to use screen: https://www.rackaid.com/blog/linux-screen-tutorial-and-how-to/

Next, come up with your SPAdes command. At a minimum, you will need to specify the output directory with --o, the path to the first Illumina read with -1, the path to the second Illumina read with -2, and the path to your PacBio reads with --pacbio. **Please add -t 1 to your command so that it uses only 1 core on the system rather than all 16. Gotta share with everyone.**

This will take a while. Leave it running overnight to finish and we'll pick up next week. Good luck on the midterm if you're on Thursday and enjoy the weekend if you're on Friday!