BioE 131/231
Fall 2018

Lab Assignment #6

Overview and references

You will use the provided Illumina MiSeq reads from a sequencing run of *Shewanella oneidensis* to generate a plot showing how many times each position in the reference genome is covered. First, you will have to filter the reads for contamination from human cells. Use bowtie2 for the alignments, samtools to calculate coverage, and matplotlib to plot the results.

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 of next week. You can upload a link to your GitHub repo on bCourses.

Bowtie2
http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

samtools
http://www.htslib.org/doc/samtools-1.2.html

Background

You are studying a strain of *Shewanella oneidensis*, a metal reducing bacterium. Your friend Jamie is learning to sequence genomes using Illumina MiSeq and asked if they could sequence your strain for practice. You get reads back from the sequencer, and Jamie mentions they might have contaminated the run with some of their own DNA during library prep. You don't think it will be a problem, though, since humans and *Shewanella* diverged a long time ago and there aren't likely to be many sequences that match closely.

The first thing you want to do is figure out how the sequencing run went. How many reads map to humans? How many map to *S. oneidensis*? Do the reads cover the *S. oneidensis* genome uniformly?

Open up a terminal on bioe131.com and locate the data

This lab will take place mostly in the terminal. Please connect to bioe131.com via SSH (PuTTY on windows) or Jupyter.

The reads from your sequencing run are at:
/data/lab6/illumina_reads.fastq

A bowtie2 database containing the human genome is at:
/data/hg19/hg19

A bowtie2 database containing the *S. oneidensis MR-1* reference genome is at:
/data/lab6/shewanella_oneidensis_mr-1

Align the reads to the human genome

You're going to want to map these reads to the human genome first to remove any contamination. To do this, check out the sample bowtie2 command from the PowerPoint.

What will you use for the database? Input reads?
If you use the --un unaligned.fastq option, what will unaligned.fastq contain?

When Bowtie2 finishes, copy the output message to your iPython notebook. It's only a few lines and just reports how many reads aligned during the run.

What percentage of your sequencing library came from contaminating human DNA?

Align the reads to the *S. oneidensis* reference genome

Now, let's find out how well your bacterium was sequenced. Run bowtie2 again, using the filtered reads you obtained in the previous step as input and the *Shewanella* reference genome as a database.

Again, copy the output message that lists how many reads aligned into your iPython notebook.

If you use the --un unaligned.fastq option, what will unaligned.fastq contain?

What percentage of your filtered library didn't align to the reference genome? If you use --very-sensitive instead of --very-fast, does this percentage increase or decrease?

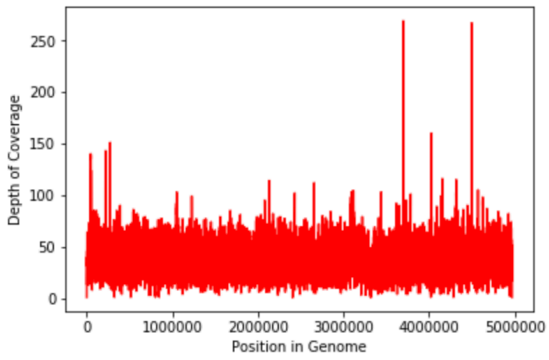
Generate a coverage plot

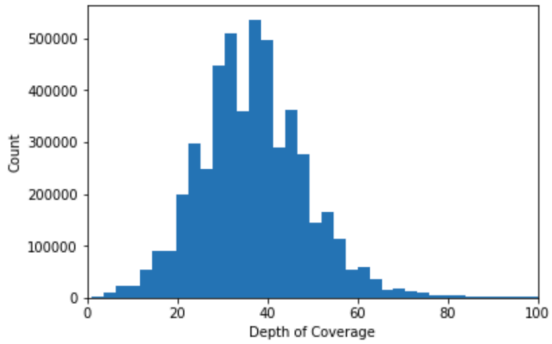
At this point, you should have a SAM file containing the alignment of your filtered reads to the *S. oneidensis* reference genome. Referring to the PowerPoint, please convert this file to BAM, sort it, index it, and use "depth" to generate a file containing the depth of coverage for every position in the reference genome.

What is the min, max, and mean coverage across all positions?

Once you have your depth file, load it into iPython and use matplotlib to generate a scatter plot where the x-axis is position in the genome and the y-axis is depth of coverage. Next, using the same data, generate a coverage histogram. We want to see what the distribution of coverage depth looks like.

Your plots should look something like this:





Extra Credit 1: Generate the same two plots using the SAM file you produced aligning the reads to the human reference genome. Instead of using "Position in Genome" (the plot on the left in red, above), plot the average depth (total depth / length of chromosome) for each human chromosome: chr1-22, X, and Y. What biological sex is Jamie?

Extra Credit 2: Try to "zoom in" on regions that seem to have higher coverage than average. Can you figure out what genes are in those regions? You'll have to extract the sequence from those regions and BLAST it.