Deliverable 2D

All Unit 2 work is due by Mar 10.

Each deliverable is complete when you have:

- answered each question
- saved a terminal session or screenshots demonstrating your performance of the commands (on Discovery or posted on GitHub)
- indicated in a "TOC" (table of contents) file where your work is found (GitHub repo or specific path/name_of_file on the cluster).

Beginning of exercise: You have three data sets consisting of NGS sequencing reads:

- lambda phage from bowtie tutorial
- Ppar reads from Marine Genomics course
- Day lab reads

Perform this first set of operations for all three datasets.

Prepare a separate directory for each project, with subdirectories as needed.

First let's describe the data we have. For each set of fastq files, describe:

1. How many reads are in each file
    a. 1 (F&R) = 32,833,451 reads
    b. 2 (F&R) = 33,738,336 reads
    c. 3 (F&R) = 35,731,214 reads
    d. 4 (F&R) = 36,678,316 reads
    e. 5 (F&R)= 36,972,680 reads
    f. 6 (F&R) = 31,401,357 reads
    g. 7 (F&R) = 35,536,673 reads
    h. 8 (F&R) = 24,498,096 reads
    i. 9 (F&R) = 29,794,050 reads
    j. 10 (F&R) = 28,893,152 reads
    k. 11 (F&R) = 29,291,552 reads
    l. 12 (F&R) = 29,043,844 reads
    m. 13 (F&R) = 29,023,016 reads
    n. 14 (F&R) = 24,730,770 reads
    o. 15 (F&R) = 28,387,419 reads
2. The length of the reads and if they are single or paired-end

a. The sequence length is 150, and they are paired-end (indicated by the R1 and R2 files for each data set)

3. The overall quality of the reads and anything to be concerned about

   a. The overall quality of the reads is pretty good. The per base sequence content is marked as poor, but when examining the graph the results appear pretty consistent for RNAseq data. The beginning of the sequences for RNAseq usually have some specific adapter that causes the change in results as seen in the quality report. The per tile sequence quality, sequence duplication levels, and adapter content were also flagged in most quality reports, but there is not a huge concern for the results in those categories.

4. Whether they appear to have adapter sequences that need to be trimmed

   a. There does not appear to be any adapter sequences that need to be trimmed (there is no common short sequence at the beginning of each read)

Collect quality control data on the reads, in the form of an .html file produced by fastqc.

If the sequences of a project need trimming, perform this step as described in the Marine Genomics tutorial, using cutadapt

There are no sequences that need trimming for the Day Data.

For now, leave the Day data and perform the rest of the operations only on the lambda phage and Ppar (sea cucumber) data.

Index the genome for each species using bowtie.

Map the reads to the genome using bowtie. (How is the command used in the Marine Genomics tutorial different from that used in the bowtie tutorial?)

Convert the files containing mapped reads from sam to bam files using samtools.

There are two programs for determining variants (positions where the read sequences differ from the reference genome) that we were introduced to: bcftools and angsd. Use each of these to call variants for the lambda phage and sea cuke data, and compare the results.

Note that the bcftools protocol requires input files to be in a “sorted.bam” format, whereas angsd takes bam files as input.