Wed Feb 21: BIOL 3411

Toolbox maintenance: What are all the programs we have used so far? Where are they located?

Three data sets consisting of NGS sequencing reads:

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

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. reads_1.fq and reads_1.fq have 10000 reads each
    b. longreads.fq has 6000 reads
2. The length of the reads and if they are single or paired-end
    a. Heading of each read does not say its length
    b. They are most likely paired-ends
3. The overall quality of the reads and anything to be concerned about
    a. They have a pretty poor quality based on the "per sequence base quality"
4. Whether they appear to have adapter sequences that need to be trimmed
    a. No – there are no repeating sequences at the beginning of each read that look like they need to be trimmed.

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.

****

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?)

This command in the bowtie tutorial did not use a shell script, whereas the command in the Marine Genomics tutorial did.

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.

Skills you need to master to work with sequencing data

C V
V R Center for Virology and Vaccine Research

Congratulations to Dr. Barouch for Receiving the 2023 Paragon Award for Research Excellence from the Doris Duke Foundation

2023-11-01 /

Help wanted: data scientists

Supply of people who can work with data has not kept up with amassing of data

(Kuo-Esser et al., 2024)