Introduction to Galaxy

1) Galaxy FASTQ processing and read mapping
Perform the steps below, answering each ***Question*** as you proceed.

Go to Galaxy at
https://usegalaxy.org or https://usegalaxy.eu/
Register for an account (free) if you do not already have one.

The following file contains 10,000 sequenced reads from an eastern mountain lion museum sample:

- https://github.com/shaunmahony/BMMB554-2022/raw/main/data/RSL813_R1.10K.fastq.gz

This next file contains a complete draft of the *Puma concolor* mitochondrial genome sequence:

- https://github.com/shaunmahony/BMMB554-2022/raw/main/data/puma-concolor-mtDNA.fasta

Download both of these data files to your Desktop before proceeding.

1. Import the data into Galaxy:
   You can import the reads in the following steps.
   Get Data
   Upload file
   Choose local file
   Navigate to your downloaded files
   Start

   Question**:** What are the components of the FASTQ file? What information is given for each read?

2. Examine the FASTQ file
   Let's use a basic text manipulation tool to count the numbers of lines in the file:
   Text manipulation
   Line/Word/Character count
   Run this tool on the data that was uploaded in step 1.

   Question: How many lines and characters does the FASTQ file contain? How many reads does this correspond to?

3. Examine the distribution of quality scores using
   NGS: QC and manipulation
   FASTQC
   Run the tool on the FASTQ reads.

   Questions:
   - Examine the output – how good are the sequences? Do you see any indicators of systematic problems?
   - What sort of result would raise concerns?

- Copy and paste the per sequence quality score histogram that results from FastQC. From what you know about Phred scores, does this distribution indicate that the sequences are good or bad quality (explain your answer)?
- What lengths are the sequences?
- Compute the total length of the sequences in the FASTQ file.

4. Trim sequencing adapters from the FASTQ reads.

NGS: QC and manipulation
Trim Galore!
Ensure that you are using automatic detection of adapter sequences and that the resulting trimmed reads must be a minimum of 20bp in length.
This tool trims sequences matching the Illumina sequencing adapter from the 3' ends of reads, with additional options to remove reads under certain conditions after trimming.
Rename the results to "RSL813_R1.10K.trimmed.fastq"

Question: Using a line count (e.g. as in step 2), how many reads have survived trimming?

5. Find out how much sequence was trimmed in step 4. (convert to FASTA, compute FASTA lengths, compute statistics on FASTA lengths)

Convert to FASTA
Convert Formats
FASTA to FASTQ

Compute FASTA lengths
FASTA manipulation
Compute sequence length

Compute statistics
Statistics
Summary Statistics *(column 2 on sequence length table)*
Graph / Display Data
Histogram *(column 2 on sequence length table)*

This chain of tools illustrates some of the basic text processing and statistical capabilities of Galaxy.

Questions:

- What is the mean length of the trimmed sequences?
- What is the total length of the trimmed sequences? How much sequence has been trimmed in total? (i.e. compare with answer in step 3)
- Copy and paste the length distribution histogram to your answer sheet. What can you conclude from the form of this distribution?

7. Map the trimmed reads to the mountain lion mtDNA sequence.

Map the reads with BWA
NGS: Mapping
Map with BWA (short reads version)
Select "use a genome from history", and select the mitochondrial genome sequence that you uploaded. In "Select input type", select "single fastq" and choose the trimmed FASTQ file from step 4. Execute
This tool uses the Burrows Wheeler Transform to map reads to a genome. It produces output in BAM format.

Get summary statistics from the BAM file
NGS: SAMtools
Stats
Run this tool on the BAM file produced above.

Questions:

- How many reads were mapped? Is this more or less than you expected? Explain why?

2) Motif scanning (25 points)

Your collaborator is investigating the effect of a small molecule on the growth of yeast (*S. cerevisiae*). She has found 10 genes that respond to this small molecule. She believes that these genes are controlled by the REB1 transcription factor, and she has asked you to help her assess that hypothesis. Specifically, she wants you to assess whether there are binding sites for REB1 in the promoter regions of her 10 genes. She has provided you with the promoter sequences for her genes:
http://lugh.bmb.psu.edu/bmb482/hw4-yeast-promoters.txt

1. Use a motif scanning approach to find potential REB1 binding sites in the upstream regions of her target genes.
    a. Download the REB1 binding motif in MEME format from JASPAR (http://jaspar.genereg.net/) – search for "REB1" under the "Fungi" set.
    b. Use FIMO (http://meme-suite.org/tools/fimo) to find motif matches.
        i. Under "Input the motifs", upload the REB1 motif you downloaded.
        ii. Under "Input the sequences", upload the promoter sequences above.
    c. Take a screenshot of the motif scanning results.

2. Do the motif scanning results support your collaborator's hypothesis?
3. What caveats apply to the interpretation of the results?
4. What should your collaborator do next (experimentally and/or computationally) to further assess her hypothesis?