**Day 4 Homework**

Author: Jessica Westfall: jessica.westfall@colorado.edu & Rutendo Sigauke rutendo.sigauke@ucdenver.edu
Edited: Lynn Sanford, 2023

*Introduction: In a future day of the workshop we will go into more details about RNA-seq libraries. This homework will go over the tasks that we did in class and provide more practice. We will be using these files on subsequent days of the workshop.*

Rutendo has provided us a good starting point for the pipeline. Copy the **example_process_rnaseq.sbatch** script from the GitHub repo day04 scripts folder to your home directory and make the necessary edits to do the following tasks.

**FastQC**

1. Evaluate the remaining fastq files in **/scratch/Shares/public/sread2023/homework_data_files/day4/**
How is the quality of these sequence libraries? Things we want to look at are:

- GC Content (Is the library contaminated?)
- Adaptor Content (Did the sequencer read into our adaptors?)
- Read Duplication (Is our sample overamplified? *Depends on library type…*)
- Sequence Quality/N content (How confident was the sequencer in calling each base?)
- Sequence Quality based on flow cell location (Was there a sequencing failure?)
- Base Identity at each location (Was there any bias in amplification/ligation?)

**Trimmomatic**

2. Trim the sequence library to remove adapters. Save the output to have the suffix '_trim.fastq' to track the trimming.

When trimming, consider if you are trimming single-ended (SE) or paired-ended (PE) reads. Consider other parameters listed for trimming:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5' end and clips the read once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length by removing bases from the end
- HEADCROP: Cut the specified number of bases from the start of the read

- MINLEN: Drop the read if it is below a specified length

3. Run FastQC on the trimmed fastq and reevaluate the fastq.

Take a look at the fastQC of the trimmed fastq and ask yourself, are these files trimmed well? Can you adjust the parameters to make the trimming more stringent to remove adapter content?

**HISAT2**

4. Edit the sbatch script to map the two corresponding paired-end fastq files. For example,
chr21Ethan_repA.RNA.end1.fastq and chr21Ethan_repA.RNA.end2.fastq

Mapping has different parameters to change the mapping efficiency. What would happen if you alter the script which currently has --very-fast to --very-sensitive?

**SAMTOOLS**

Now we have a huge SAM file.

5. Convert the SAM file to a BAM file to produce a compressed binary file that takes up less space.

What is the size difference between SAM versus BAM? What is the difference between the two filetypes that contributes to the size difference?

6. Sort and index your BAM files.

**IGV**

7. Transfer your BAM files to your local computer.

8. Open IGV and visualize the sorted BAM files

Do you know which reference genome your sequence reads were aligned and mapped to? Since both hg19 and hg38 are different versions of the human genome, are they interchangeable in IGV?