Day 5 Worksheet – Assessment

Task 1: Obtain fastq files

1. Create a new directory in your directory ~/**sread2021/** directory named **day5/**. Remember that the " ~ " indicates your home directory. You may also put these files into your scratch home directory **(/scratch/Users/<your_id>/)** which is good practice for use on Fiji. Within the **day5/** directory make the following subdirectories:
    - **fastq/**
    - **fastqc/**
    - **hisat2/**
    - **scripts/**
    - **eofiles/** **NOTE:** all sbatch error and output should go to this directory
    - **trimmomatic/**

2. Copy the script **template.sbatch** from **/scratch/Shares/public/sread2021/scripts/** to ~/**sread2021/day5/scripts/.**

3. Make another copy of the **template.sbatch** in ~/**sread2021/day5/scripts/** and rename it to **rsync.sbatch.**

4. Edit the **rsync.sbatch** script to create a job which will rsync all fastq files from **/scratch/Shares/public/sread2021/data_files/day5/** to ~/**sread2021/day5/fastq/**. The error and log files from sbatch should be stored in the /**eofiles/** directory. Note how you can copy multiple files within a directory.

   Once the rsync job is complete, you should have an entire set of RNAseq data files in your fastq folder. This represents a standard RNAseq experimental setup to compare two conditions (Eric vs. Ethan). Three biological replicates were included for each condition, and a pair of sequencing reads (end1 and end2) for each replicate:

Condition	Biological Replicates	Read Pairs
Chr21 Eric	RepA	End1
		End2
	RepB	End1
		End2
	RepC	End1
		End2
Chr21 Ethan	RepA	End1
		End2
	RepB	End1
		End2
	RepC	End1
		End2

Note: For the following **Task 2 - 5**, we will focus on two of the files **chr21Ethan_repA.RNA.end1.fastq** and **chr21Ethan_repA.RNA.end2.fastq**

Task 2: Quick evaluation of fastq file

5. Determine how many reads are in the fastq files. You can use the ***wc -l*** command to determine line counts. Review fastq format to determine from the line counts how many reads are in your files.

   Fill in the blanks:

   There are _____ reads in the file ***chr21Ethan_repA.RNA.end1.fastq***, and _____ reads in the file ***chr21Ethan_repA.RNA.end2.fastq***

Task 3: FastQC

6. Create a new sbatch file for running FastQC:
   a. Make another copy of the **template.sbatch** file, and rename it **fastqc.sbatch**
   b. Make any necessary edits to **fastqc.sbatch** script to run FastqQC on the fastq file and direct the output to **~/sread2021/day5/fastqc**

   Fill in the blanks:

   The read length for the file ***chr21Ethan_repA.RNA.end1.fastq*** is _____. In this data set, there are _NO/YES_ adapters and (if yes) they are _____ adapters.

   The read length for the file ***chr21Ethan_repA.RNA.end2.fastq*** is _____. In this data set, there are _NO/YES_ adapters and (if yes) they are _____ adapters.

Task 4: Trimming your dataset and QC via FastQC

7. Copy and create a new sbatch script named **trim.sbatch** and edit the file to run the program Trimmomatic on both fastq files. Direct the output to **~/sread2021/day5/trimmomatic,** and save the file names as **chr21Ethan_repA.RNA.end1.trimmed.fastq** and **chr21Ethan_repA.RNA.end2.trimmed.fastq.**

8. Edit your **fastqc.sbatch** script, edit it to again run FastQC on the trimmed files. The output should go to **~/sread2021/day5/fastqc**.

   Fill in the blanks:

   "For **chr21Ethan_repA.RNA.end1.trimmed.fastq**:
   There are _____ reads in the fastq file after trimming.
   The minimum read length is _____ and the maximum read length is _____.

   For **chr21Ethan_repA.RNA.end2.trimmed.fastq**:
   There are _____ reads in the fastq file after trimming.
   The minimum read length is _____ and the maximum read length is _____.

   Did you notice any change before vs. after trimming? _____.

Task 5: Mapping to genome.

9. Create a new sbatch script named **mapping.sbatch** and edit the file to run the program HISAT2 followed by Samtools using both fastq files. Map the reads to human **hg38 reference genome**. The HISAT2 genome index can be found in **/scratch/Shares/public/sread2021/data_files/genome/hg38/HISAT2/genome**. Direct the output to ~/**sread2021/day5/hisat2/**.

   Fill in the blanks:
   For **chr21Ethan_repA.RNA**, a total of ____ (__%) of reads mapped to hg38 reference genome, of which ___ reads are pair-end.

Task 6: Combine job scripts and process the remaining RNAseq data files.

10. Now that we are familiar with each individual step required to process a pair-end sequencing read file pair, let's finish processing the remaining data files to get them ready for downstream RNAseq analysis.

    To speed up the process. Let's combine all the commands into a single sbatch script, **rna_read_process.sbatch**. You can copy and paste the commands from previous scripts.

    Hint: Define the file name as a variable in the new sbatch script, so you don't have to re-edit command for every sequencing data.

    Fill in the blanks:
    For **chr21Eric_repA.RNA**, a total of ____ (__%) of reads mapped to hg38 reference genome, of which ___ reads are pair-end.

    For **chr21Eric _repB.RNA**, a total of ____ (__%) of reads mapped to hg38 reference genome, of which ___ reads are pair-end.

    For **chr21Eric _repC.RNA**, a total of ____ (__%) of reads mapped to hg38 reference genome, of which ___ reads are pair-end.

    For **chr21Ethan_repB.RNA**, a total of ____ (__%) of reads mapped to hg38 reference genome, of which ___ reads are pair-end.

    For **chr21Ethan_repC.RNA**, a total of ____ (__%) of reads mapped to hg38 reference genome, of which ___ reads are pair-end.

Task 7: View the files in IGV (Optional)

After processing all fastq files in the datasets, we will obtain a set of sorted bam files. These can be directly loaded onto IGV for comparison. Load all bam files onto IGV and scroll through genes on chromosome 21, do you notice any gene expression differences between Ethan and Eric?