BIN508 Assignment 2

Due Date: 24 March, Monday, 13:30 **Cut-off:** 24 March, Monday, 19:30

Late Policy: After the due time, 5 point deduction will be applied for each extra hour.

- **Uploading your assignment as a single PDF is mandatory.**.
- You are free to use Galaxy servers (USA, EU) or a Linux/Unix environment. If you plan to use Linux/Unix, you need to download the human reference genome, unzip and index it with "bwa" first:

```
bwa index Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

- Useful links: 1) Information on mapping 2) Fastq manipulation & SAM/BAM 3) SAM format
- **Starting from this assignment, you do not have to show every QC plot. Share the important parts only (adapter contamination, unexpected GC distribution, quality drops, etc. if any).**
- **Don't forget to go over other reports (such as the trimming report from `cutadapt`, etc.).**
- **Make your Galaxy history and steps visible, and add a link to your Galaxy history at the top of the assignment. If you use a Linux environment, add your script to the end of the assignment with a mono-space font and include comments.**

1) Preprocess `sample1_1.fq.gz` and `sample1_2.fq.gz`.
   a) Use FastQC and then MultiQC. Trim the data using **cutadapt** (remember that this is a paired-end data). Set the "**quality cutoff**" and "**minimum length**" values to **20**. Rerun FastQC and MultiQC on the trimmed data.
      - Compare the quality of preprocessed and postprocessed data using the MultiQC plots. Add screenshots for the plots you comment on.
      - Comment on the cutadapt results.
   b) Use the raw `sample1_1.fq.gz` and `sample1_2.fq.gz` again, but this time, use **fastp** instead (remember that this is paired-end data). In the tool parameters, under "**Filter Options**," set "**Qualified quality phred**" and "**Length required**" to **20**.
      - Comment on the fastp results. Add screenshots for the plots you comment on.
   c) Draw two workflow diagrams showing the inputs and outputs for both 1a and 1b.
   d) Share your thoughts about these two workflows. Which one seems more useful to you? Why?
      - If some of the fastp plots are not visible, please try using Mozilla Firefox.

2) Preprocess and align `sample2_1.fastqsanger` and `sample2_1.fastqsanger`.
    a) The usual: QC -> trim -> QC; add explanations as needed.
    b) Then, use **BWA-MEM** ("Map with BWA-MEM") to align the trimmed reads (The parameters to use are shared on the last page). When it is done, click on the eye icon to see the contents of the output of BWA-MEM. Scroll down to see the first read (the first line starting with "HWUSI"). Which chromosome is it aligned to, and at what position on the chromosome? Add a screenshot showing that line.
        - If using the terminal, your output will be a SAM file. You can view its contents as: `more output_name.sam` and use Space Bar to scroll down.
        - Galaxy directly gives the output as a BAM file as it runs `samtools view` in the background. On the terminal, you have to run `samtools view` manually to convert your SAM file to a BAM file (not necessary for this assignment).
    c) Use **samtools depth** on the output of BWA-MEM. It generates a table where Column 1 is the chromosome, column 2 is the position on the chromosome, and column 3 is the number of reads aligned to that position. Use "**sort**" tool on this table (**NOT** "samtools sort", simply **sort**), and sort by column 3 in descending order. On which chromosome are the reads aligned the most? Please share a screenshot showing the first several lines of the table.
    d) Draw a diagram showing the inputs and the outputs for question 2.

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome (Galaxy Version 0.7.17.2)

Run Tool

Tool Parameters

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome *

Human (Homo sapiens) (b38): hg38

Select genome from the list

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads *

26: ZR751 paired-end RNA-seq subsampled (end 1)

Specify dataset with forward reads

Select second set of reads *

27: ZR751 paired-end RNA-seq subsampled (end 2)

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths. - optional

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

Set read groups (SAM/BAM specification)

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Auto-assign

Yes

Use dataset name or collection information to automatically assign this value

Auto-assign

Yes

Use dataset name or collection information to automatically assign this value

Platform/technology used to produce the reads (PL) *

ILLUMINA

Auto-assign

Yes

Use dataset name or collection information to automatically assign this value