Alexander Gilson

AS.410.635.81

Genomics Spring 2021

Exam 3 - Due 11:59 pm (ET), Sunday, 05/02/2021

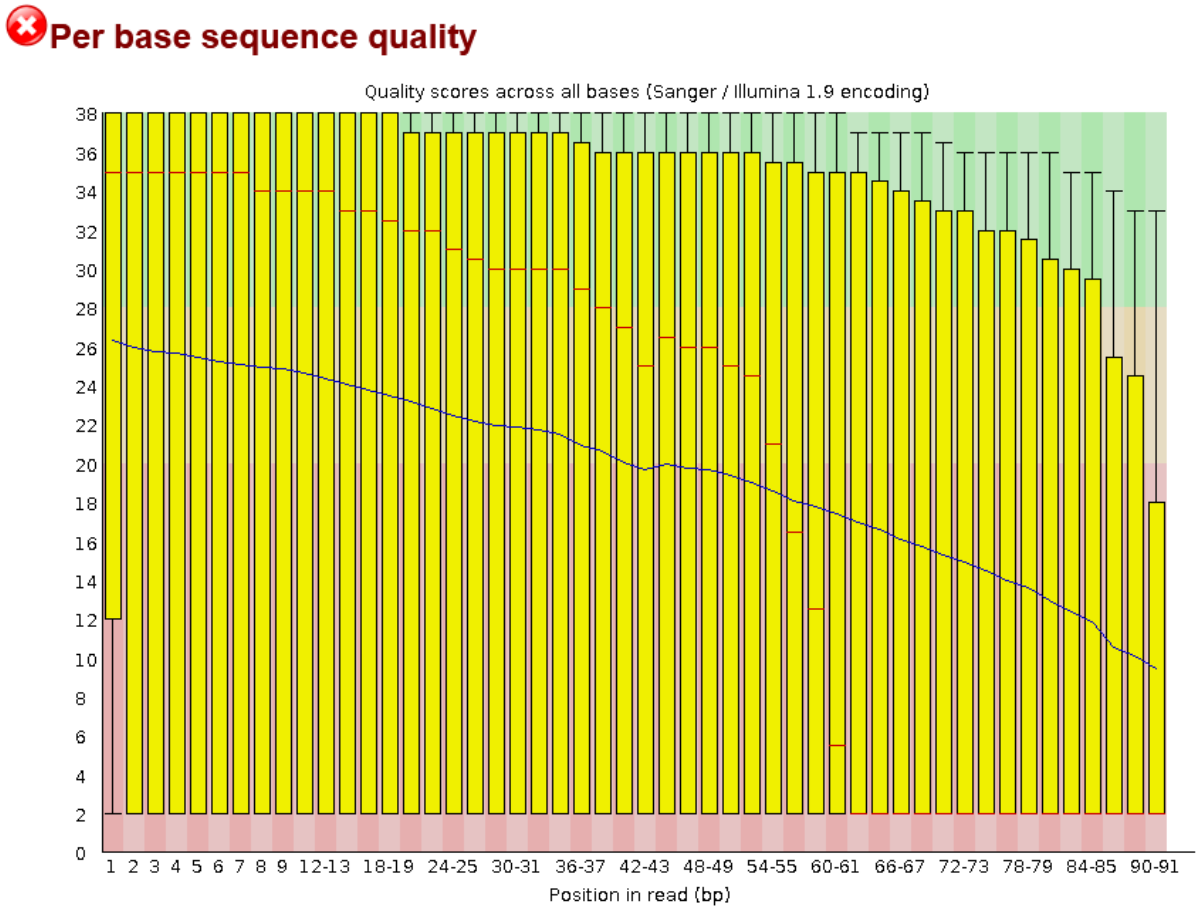
Please work alone and submit through Blackboard. Good luck!

Part 1 - 7 points

In Galaxy, run FASTQC on the following file:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00324/sequence_read/ERR018456.filt.fastq.gz

1. (1 pts) Submit the box plot of quality scores.



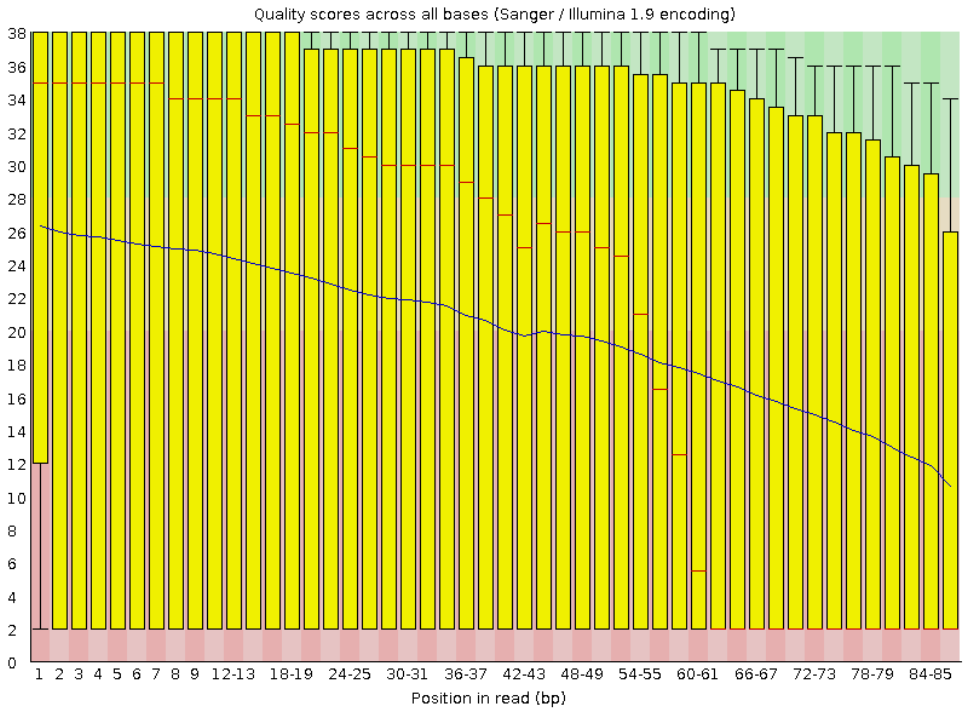
   a.
2. (1 pts) What is the read length?
   a. Read length is 91bp long.

3. (1 pts) Based on the read length, what sequencing technology was likely used: Roche 454 or Illumina? Briefly explain.
    a. Based on the read length, it is most likely that Illumina was used to produce this sequence read. Illumina uses much shorter fragments for reads in order to sequence them (~100-150bp), while Roche uses longer reads for its method.
4. (2 pts) What positions in the sequence have the most variability in sequence quality? Briefly explain.
    a. As the bases are sequenced in a read, the quality declines over the whole course of the read. Therefore, the bases of the read at the end would have the most variability in sequence quality, as their read quality is nowhere near as standard as seen with the beginning of the read.
5. (2 pts) Use the FASTQ Trimmer tool to remove five nucleotides from the 3' ends of all reads. Submit a new box plot of quality scores.

Basic Statistics

Measure	Value
Filename	FASTQ Trimmer on data 1.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	685789
Sequences flagged as poor quality	0
Sequence length	86
%GC	42

Per base sequence quality



a.

Part 2 - 6 points

Open the HIV-1 genome in IGV (Genomes > Load Genome from Server). Create a BED file (0-based start) with the following three intervals:

The gag gene located at positions 336 through 1838.

The vif gene located at positions 4587 through 5165.

A Gag protein potential epitope located at amino acid positions 271 through 285 of the Gag protein. The amino acid sequence is NKIVRMYSPTSILDI.

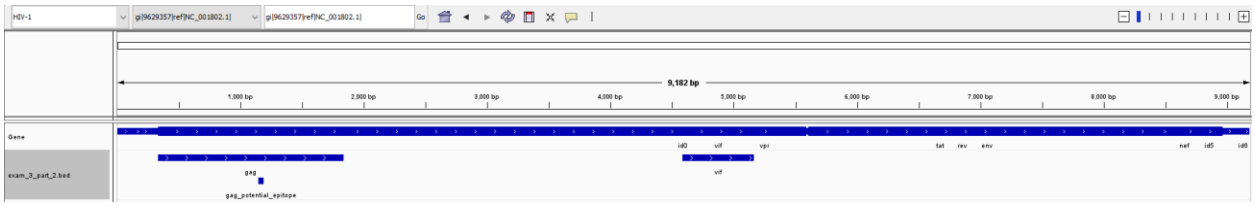
Create the BED file with NC_001802.1 in column one. Load it to IGV.

(2 pts) Submit the BED file.

- File name: exam_3_part_2.bed

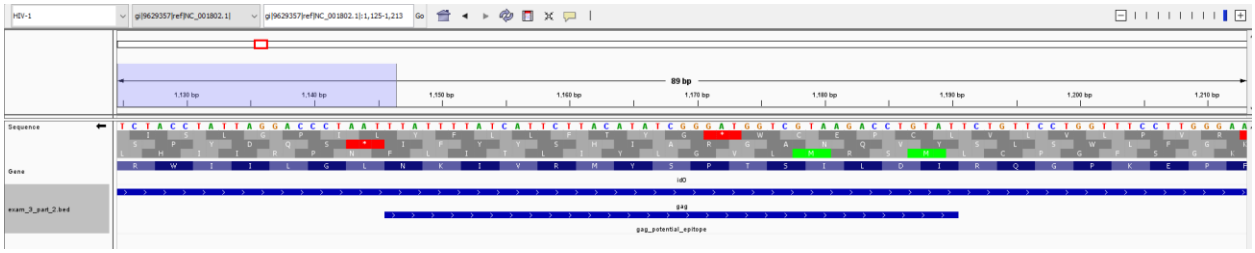
(2 pts) Submit a screenshot that shows all three intervals in IGV.

-



(2 pts) Submit a zoomed in screenshot that shows the epitope and the amino acid sequences.

-



Part 3 - 7 points

Load the attached mouse files to Galaxy. They are ungroomed single-end FASTQ files with Illumina 1.5 phred encoding from a ChIP-seq experiment and downsampled to a part of chromosome 19. In Galaxy, run the FASTQ Groomer tool to convert the reads to fastqsanger format. Then, use Trimmomatic to require a phred score greater than or equal to 20. Align the trimmed reads to the mm9 reference with Map with BWA. Finally, run MACS2 callpeak on the experimental ChIP-seq with the control output as the control.

- First ran FASTQC to check datatype before conversion to fastqsanger. Both were illumina 1.5, so fastq groomer was set accordingly in order to convert reads to fastqsanger format. Trimmomatic parameters were kept as they are originally set as the average quality score required was already set to 20. Map with BWA with single fastq selected for both experimental and control(originally thought experimental was the first uploaded, needed to redo and switch the two)
    - I wanted to keep track of the steps I took to make sure I ran this correctly, and since I am not submitting a workflow I thought it best to include my thought process.

(1 pts) Retrieve the peaks in tabular format. Find the interval chr19:37,340,169-37,340,716. List the value in the fold_enrichment column.

- 27.13470
- chr19 37340169 37340716 548 37340441 75.00 92.63109 27.13470 85.97186 Map_with_BWA_on_data_18__mapped_reads_in_BAM_format__peak_227

(2 pts) Load both bedgraph files into IGV, mm9. Go to the interval from Part 3a. What is the nearest transcript?

- The closest transcript is 4931408D14Rik

(2 pts) Relative to the nearest two genes, where (upstream, exon, intron, downstream) is the MACS peak?

- The MACS3 peak in the treatment group is DOWNSTREAM from the IDE gene and UPSTREAM from the 4931408D14Rik gene.

(2 pts) Submit a screenshot from IGV showing both the MACS peak and a small portion of the nearest two genes.

-