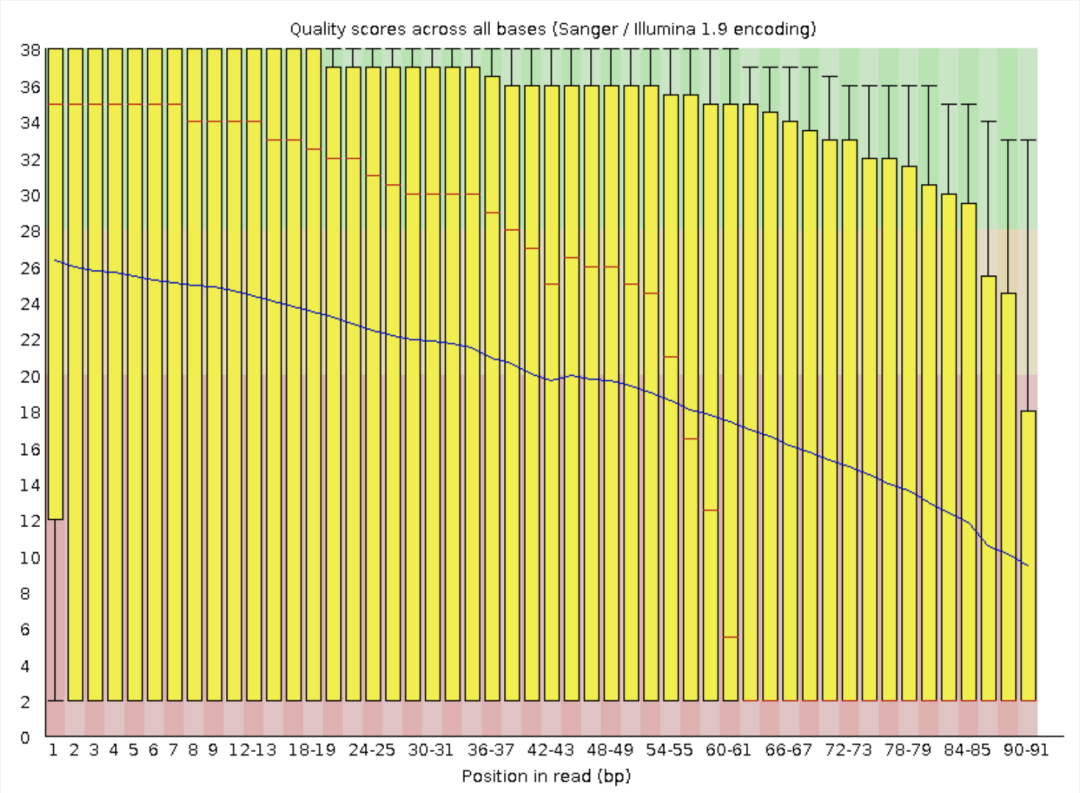
Genomics Spring 2021 Exam 3 - Due 11:59 pm (ET), Sunday, 05/02/2021

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.
See attached - part1_1.png



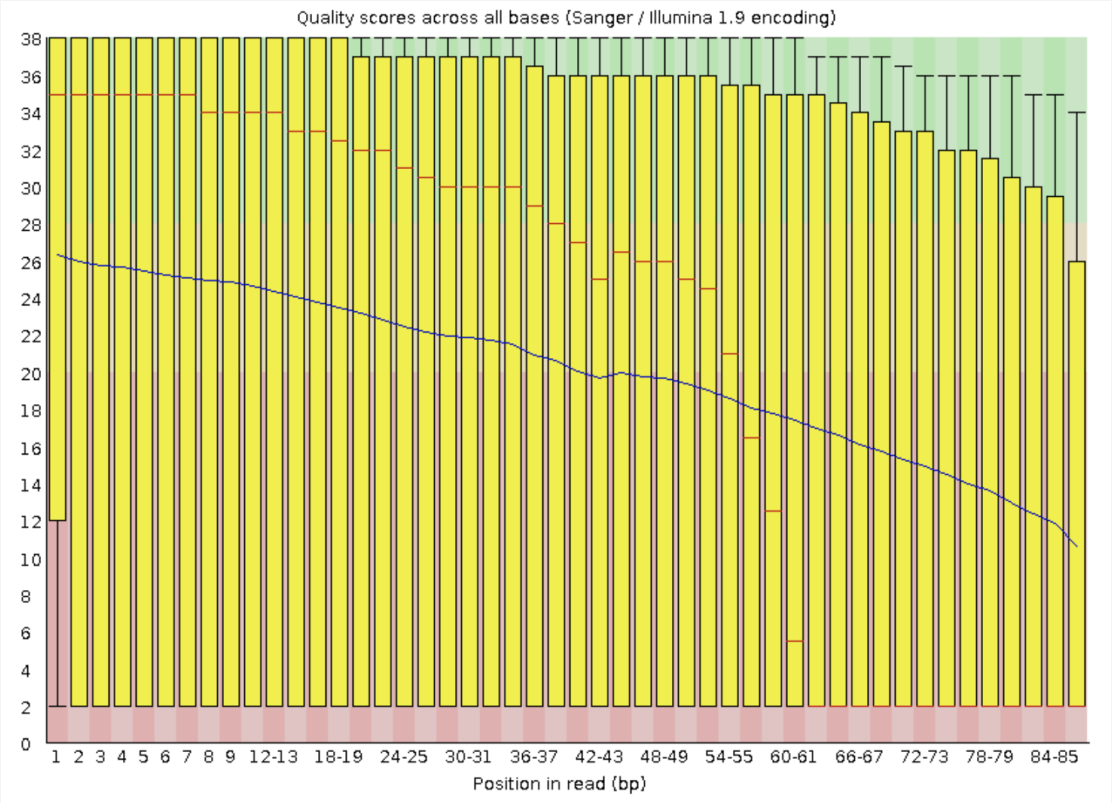
2. (1 pts) What is the read length?
Read length: 91

3. (1 pts) Based on the read length, what sequencing technology was likely used: Roche 454 or Illumina? Briefly explain.
Roche 454 usually produces reads around 400bp in length, while Illumina usually produces reads around 100-250bp in length.

4. (2 pts) What positions in the sequence have the most variability in sequence quality? Briefly explain.
Positions 1 to 60-61 all have a range of 36, the largest range and therefore the most variability in sequence quality.

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.

See attached - part1_5.png



Part 2 - 6 points

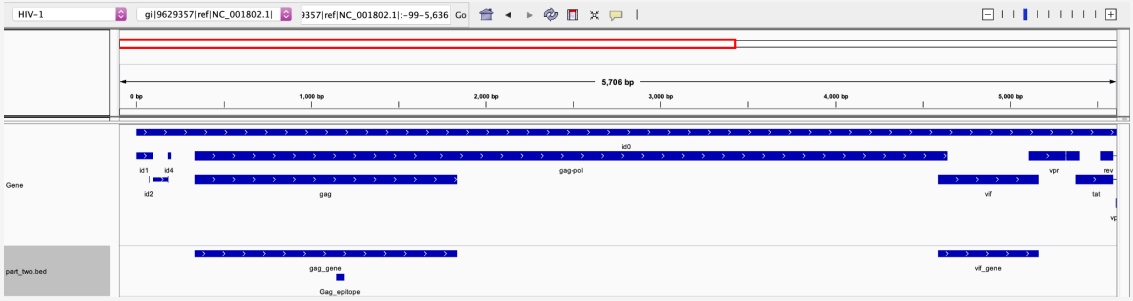
Open the HIV-1 genome in IGV (Genomes > Load Genome from Server). Create a BED file (0-based start) with the following three intervals:

- 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.

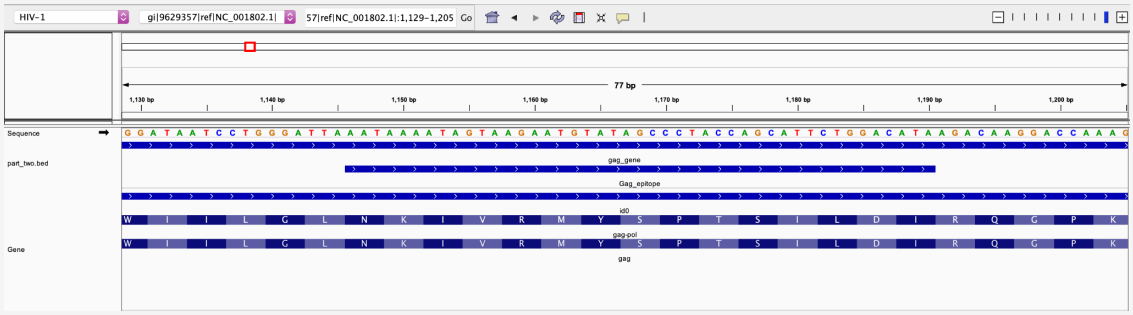
Create the BED file with NC_001802.1 in column one. Load it to IGV.

- (2 pts) Submit the BED file.
  See attached - part_two.bed

- (2 pts) Submit a screenshot that shows all three intervals in IGV.
  See attached - part2_2



- (2 pts) Submit a zoomed in screenshot that shows the epitope and the amino acid sequences.
  See attached - part2_3



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.

- (1 pts) Retrieve the peaks in tablular format. Find the interval chr19:37,340,169-37,340,716. List the value in the fold_enrichment column.
  fold_enrichment: 27.13470
- (2 pts) Load both bedgraph files into IGV, mm9. Go to the interval from Part 3a. What is the nearest transcript?
  4931408D14Rik
- (2 pts) Relative to the nearest two genes, where (upstream, exon, intron, downstream) is the MACS peak?
  upstream
- (2 pts) Submit a screenshot from IGV showing both the MACS peak and a small portion of the nearest two genes.

See attached - part3_4

Treatment and control bedgraphs with autoscale. Top = Treatment; Bottom = Control