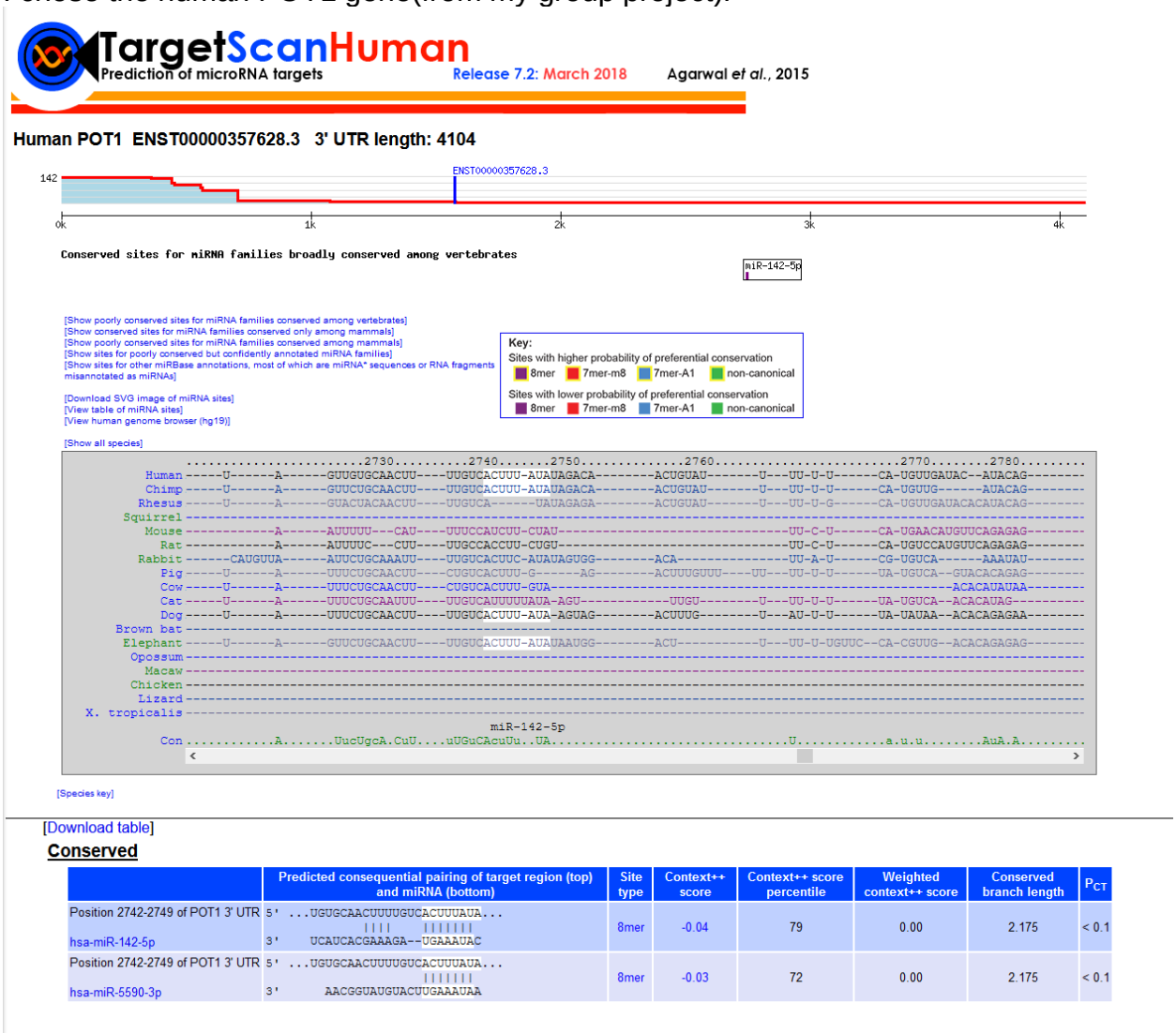
Genomics Spring 2021
Units 9-10 Graded Homework

Part 1

(1 pt) Pick your favorite protein-coding gene from any of the following organisms: human, mouse, fly, worm. Use TargetScan to find predicted miRNA binding sites on the 3'UTR of your gene. If your gene has multiple 3'UTR isoforms, choose only one to query. Report only 1 prediction by screenshot(s) that covering the whole output in the master pdf document (no need to submit the screenshot as a separate file for this case). Be sure to include all the miR names, the position of the target sites (relative to gene or absolute on chromosome), and some measure of confidence of the predictions in your screenshots.

- I chose the human POT1 gene(from my group project).

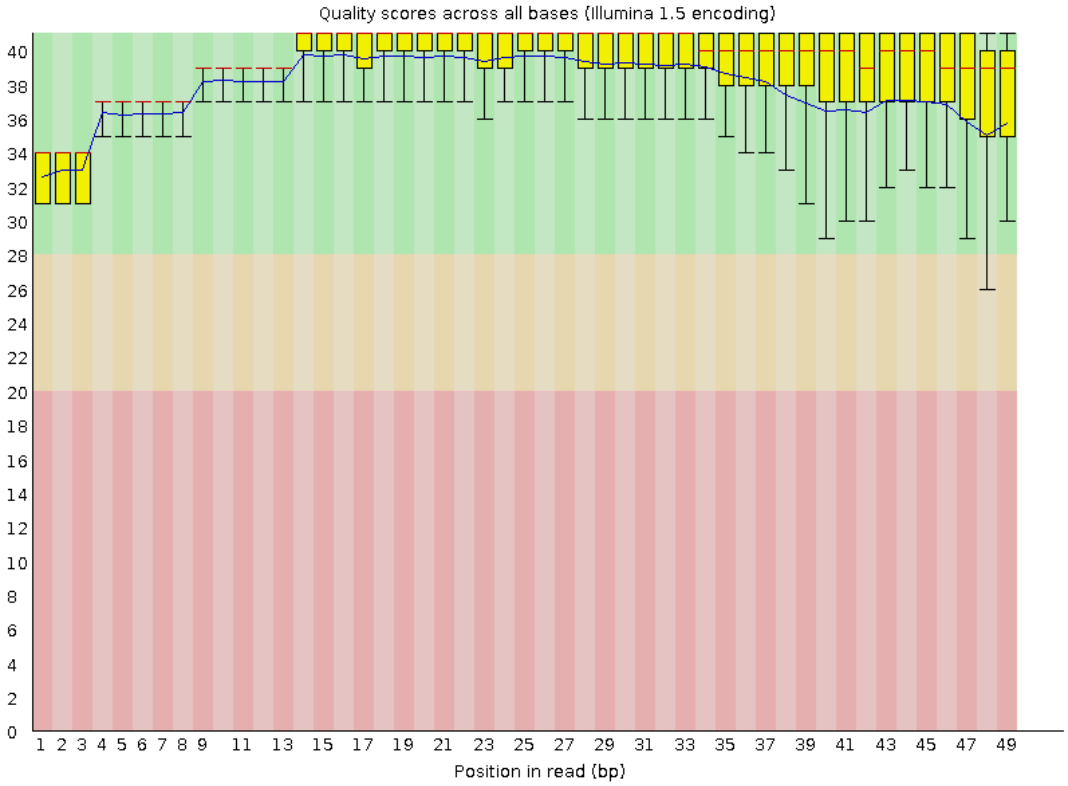


-

Part 2

Upload the attached `HW5_Part2_sample_NGS_data.fq.gz` file to Galaxy. This downsampled file is from a NGS experiment on *C. elegans* (genome version WS220/ce10).

a. (0.25 pts) Run FASTQC and submit the boxplot of the quality scores. How would you describe the quality of these data?



a.

b. (0.25 pts) What phred encoding scheme does this data use? How long are the reads? How many reads are in the file?

Basic Statistics

Measure	Value
Filename	HW5_Part2_sample_NGS_data_fq_gz.gz
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	20000
Sequences flagged as poor quality	0
Sequence length	49
%GC	46

a.

b. Illumina 1.9 encooding scheme

c. Reads are 49bp long

d. There are 20,000 reads in the file.

c. (0.25 pts) Run the **FASTQ Groomer** tool to convert the phred quality scores to Sanger/Illumina 1.9. Rerun the **FASTQC** tool on the groomed data. What phred encoding scheme is listed now?

Basic Statistics

Measure	Value
Filename	FASTQ Groomer on data 3
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20000
Sequences flagged as poor quality	0
Sequence length	49
%GC	46

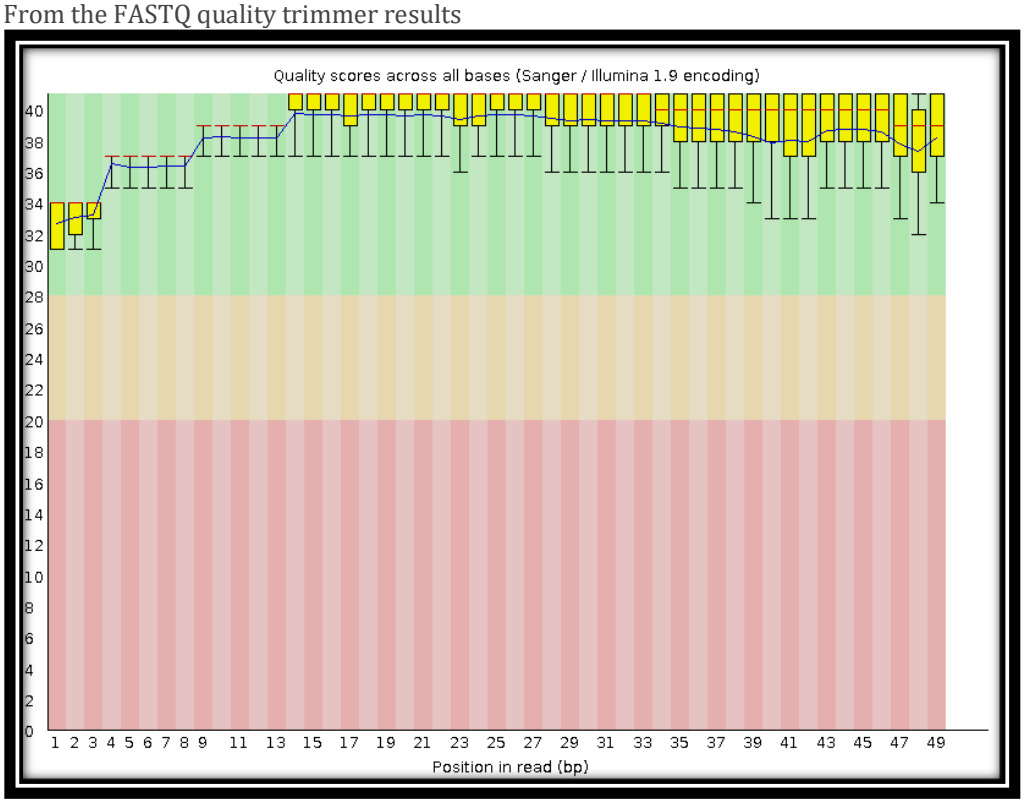
a. It now lists the encoding as Illumina 1.9

Part 3

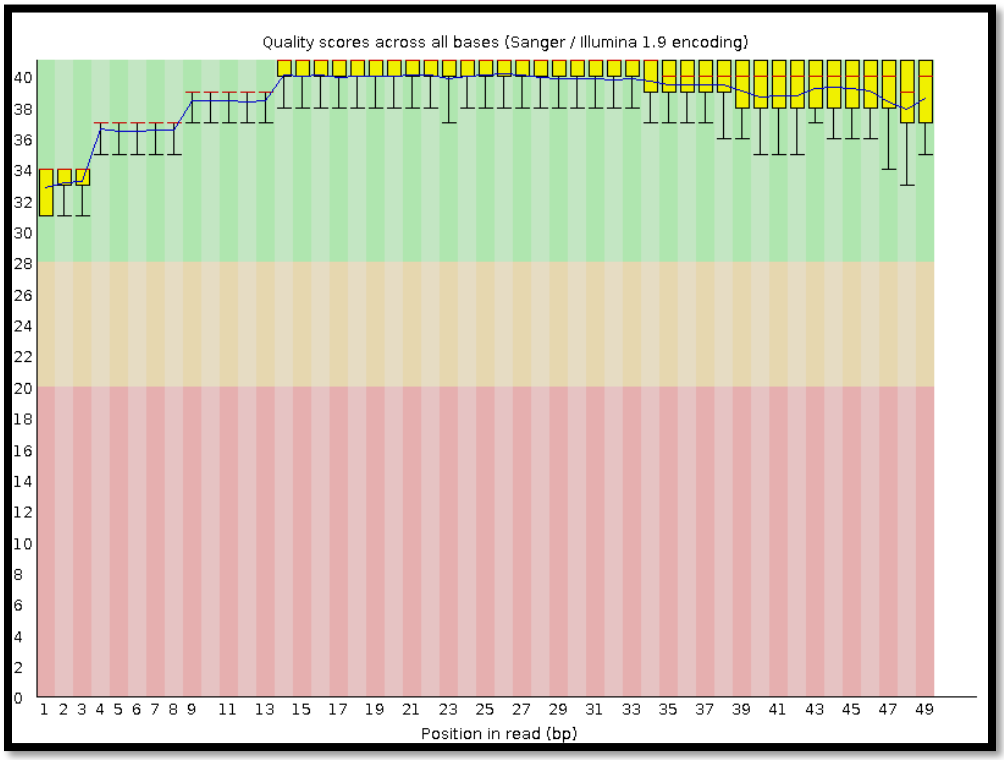
Using the groomed fastq file from Part 2 (be sure that this file is Type 'fastqsanger'), complete the following two trimming steps: (1) Run the **FASTQ Quality Trimmer** tool on the groomed data to trim the data with a sliding window of 4 bases. Trim the reads until the average quality score of the window is greater than 30. (2) Run the **Trimmomatic** tool on the groomed data using the same parameters. Although the tool forms are different, the same parameters can be set for each tool.

a. (0.25 pts) Run the **FASTQC** tool on each of the FASTQ Quality Trimmer and Trimmomatic outputs. Submit both boxplots of quality scores. Be sure to label which boxplot is for data from which trimming tool.

a. From the FASTQ quality trimmer results



b. From the Trimmomatic results



b. (0.5 pts) In a short paragraph, explain any differences you see between the quality score report of the untrimmed data from Part 2 and the trimmed data from Part 3. Do the differences make sense? Why or why not?

a. There is a noticeable difference in the quality score averages for the data seen in part 2 and part 3, particularly at the end of the sequence. The average scores in the untrimmed data can reach to the "yellow" section of the quality scores, indicating there is a position of the untrimmed data that is not as good quality as the other base pair quality scores around it. For this reason, the trimmed data has much more reliable data within it to be used for analysis.

c. (0.5 pts) In a short paragraph, explain any differences you see between the output of the two trimming tools. Be sure to include references to read lengths and number of reads. Which tool do you prefer, and why?

a. There are a few noticeable differences between the results of the two trimming tools. Notably, the trimmomatic results appear much "tighter" than the FastQ Quality trimmer, but this may be because the trimmomatic method removed more outlier data than the Quality trimmer. For the trimmomatic method, the sequence lengths were 4-49bp long, with a total amount of sequences being 19,367 instead of the original 20k. The quality trimmer had sequence lengths of 5-49 bp and kept the total sequence length. Therefore, if I were looking to get the most trimmed data without losing any of the sequence, I would prefer the FastQ Quality Trimmer over trimmomatic.

Part 4

Follow the protocol below to identify SNPs in NGS data from the 1000 Genomes Project (reference genome hg19). This part uses two FASTQ files from the 1000 Genomes Project that represent a paired-end sequencing experiment. The forward reads are in the file ending in '_1', and the reverse reads are in the file ending in '_2'. Load both files into Galaxy using the **Upload file** tool, choosing **Paste/Fetch data**, and pasting in the given ftp links. The data type should be set to 'fastq' and the genome should be set to 'hg19'.

Forward reads: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence_read/SRR044234_1.filt.fastq.gz

Reverse reads: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00117/sequence_read/SRR044234_2.filt.fastq.gz

Determine quality encoding: Run **FASTQC** on both files. If the quality encoding is found to be Sanger/Illumina 1.9, update the file type to 'fastqsanger'. If the quality encoding is found to be something other than Sanger/Illumina 1.9, use **FASTQ Groomer** to convert the files to Sanger/Illumina 1.9 encoding. At the end, you should have two FASTQ files in Sanger/Illumina 1.9 encoding. **BOTH ARE 1.9**

Trim low-quality bases: Use either the **FASTQ Quality Trimmer** or **Trimmomatic** tool to remove low quality bases from each file. Use a window of size 4 bases and require the average quality in the window to be at least 20. Rerun **FASTQC** on the trimmed data to ensure that low quality bases were removed. **Data WAS removed.**

Align reads to reference genome: Choose either **BWA**, **Bowtie2**, or **HISAT** to align both files to the reference genome hg19. Be sure to align the reads as paired-end. Whichever aligner you choose, get the alignments into BAM format.

Identify variants: Run the **FreeBayes** tool to identify variants. Limit the output to chr22:0-51304566 (for a more manageable file).

Filter and annotate variants: Use the **VCFfilter** tool to filter for variants that show heterozygosity (estimated allele frequency = 0.5) and have more than 10 reads covering them (total read depth > 10). The tag IDs for these parameters can be found in the header of the VCF file. To annotate which genes the variants are in, first bring in RefSeq genes in BED format from **UCSC Main**. Then, use the **VCFannotate** tool to intersect the filtered VCF file with the BED annotations.

I USED THE 1MIL DATA SETS DOWNLOADED FROM BLACKBOARD INSTEAD, THE FETCH USING THE URLs TOOK FAR TOO LONG.

a. (0.5 pts) Submit the filtered, annotated VCF file. How many variants are listed in the VCF file? How many variants were annotated with a RefSeq gene?

   a. Only one variant is listed in the final VCF file. Its location is chr22:18883995
   b. There are no genes listed in the BED file that can be found in that region denoted by RefSeq.
   c. File name: Galaxy20-[VCFannotate__on_data_19_and_data_18].vcf

b. (1.0 pt) Extract and submit your Galaxy workflow. This is how I will be grading whether you followed the protocol appropriately.

   a. File name: Galaxy-Workflow-Workflow_constructed_from_history__Homework_part_4_.ga

c. (0.5 pts) Choose any SNP in the filtered, annotated VCF file that overlaps a gene. View that position in any genome browser. What is the nucleotide change and the gene that is affected? In which part of the gene is the SNP located? What effect might the SNP have on the gene function, if any?

   a. The variant I found at location chr22:18883995 is not associated with a gene, most likely because I was using the 1million subset file of the region rather than the whole region. In order find a gene to use for this part of the homework, I filtered the free bayes file again with no allele frequency filter, but kept the read depth filter. I chose the first variant that was annotated to be associated with a non-coding gene:NR_136571.1 at position chr22:18880041(G→A). This gene is FAM230F, a lncRNA associated with gene expression that could cause major disruptions should mutations of the lncRNA inhibit binding of the target.