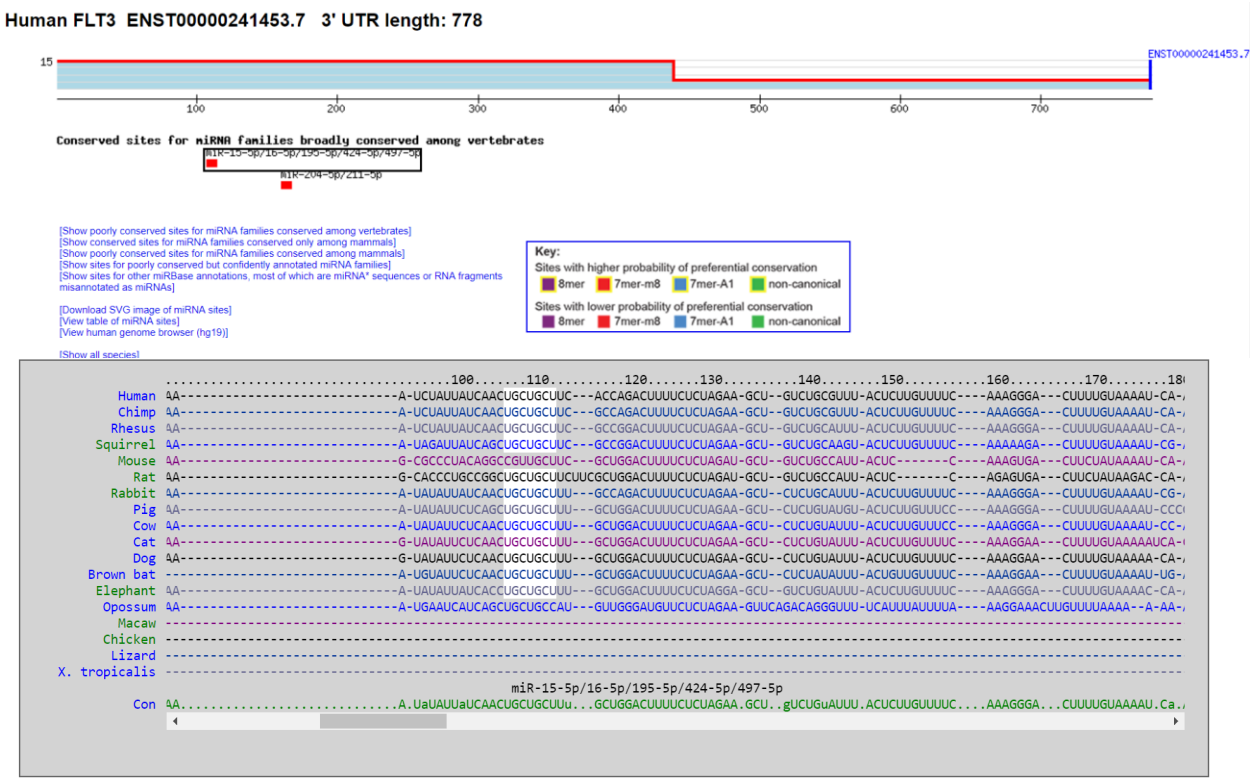
Julie Garcia
Graded Homework
Units 9/10
March 14, 2018

Part 1

I chose the human gene FLT3 (ENSG00000122025), as it is associated with the development of acute myeloid leukemia (AML) and other types of cancer. The results are shown below.



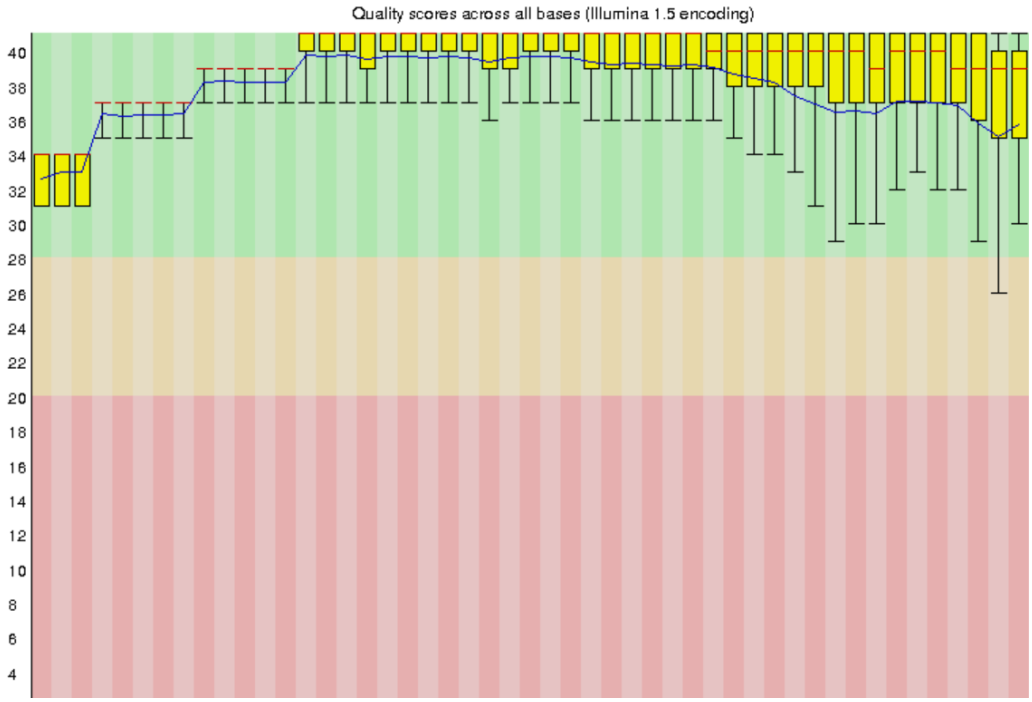
[Species key]

Conserved

	Predicted consequential pairing of target region (top) and miRNA (bottom)	Site type	Context++ score	Context++ score percentile	Weighted context++ score	Conserved branch length	P_{CT}
Position 107-113 of FLT3 3' UTR hsa-miR-16-5p	5' ...AAAUCUAUUAUCAACUGCUGCUU... 3' GCGGUUAUAAAUGC--ACGACGAU	7mer-m8	-0.38	98	-0.38	4.737	0.69
Position 107-113 of FLT3 3' UTR hsa-miR-195-5p	5' ...AAAUCUAUUAUCAACUGCUGCUU... 3' CGGUUAUAAAGAC--ACGACGAU	7mer-m8	-0.38	98	-0.38	4.737	0.69
Position 107-113 of FLT3 3' UTR hsa-miR-15b-5p	5' ...AAAUCUAUUAUCAACUGCUGCUU... 3' ACAUUUGGUACUACACGACGAU	7mer-m8	-0.33	97	-0.33	4.737	0.69
Position 107-113 of FLT3 3' UTR hsa-miR-15a-5p	5' ...AAAUCUAUUAUCAACUGCUGCUU... 3' GUGUUUGGUAAUAC---ACGACGAU	7mer-m8	-0.35	97	-0.35	4.737	0.69
Position 107-113 of FLT3 3' UTR hsa-miR-424-5p	5' ...AAAUCUAUUAUCAACUGCUGCUU... 3' AAGUUUUGUACUUAACGACGAC	7mer-m8	-0.32	96	-0.32	4.737	0.69
Position 107-113 of FLT3 3' UTR hsa-miR-497-5p	5' ...AAAUCUAUUAUCAACUGCUGCUU... 3' UGUUUGGUGUCACACGACGAC	7mer-m8	-0.31	96	-0.31	4.737	0.69
Position 107-113 of FLT3 3' UTR hsa-miR-6838-5p	5' ...AAAUCUAUUAUCAACUGCUGCUU... 3' UCCUCAGAACGGUGACGACGAA	7mer-m8	-0.32	96	-0.32	4.737	0.69

Part 2

a. I ran FASTQC on the uploaded file and the boxplot of quality scores is shown below. Each quality score is in the green region, so I would say, in general, the quality of the data is good.



b. This data uses the Illumina 1.5 phred encoding scheme, which is a previous version of Illumina. The read lengths are 49 and there are 20,000 reads in the file.

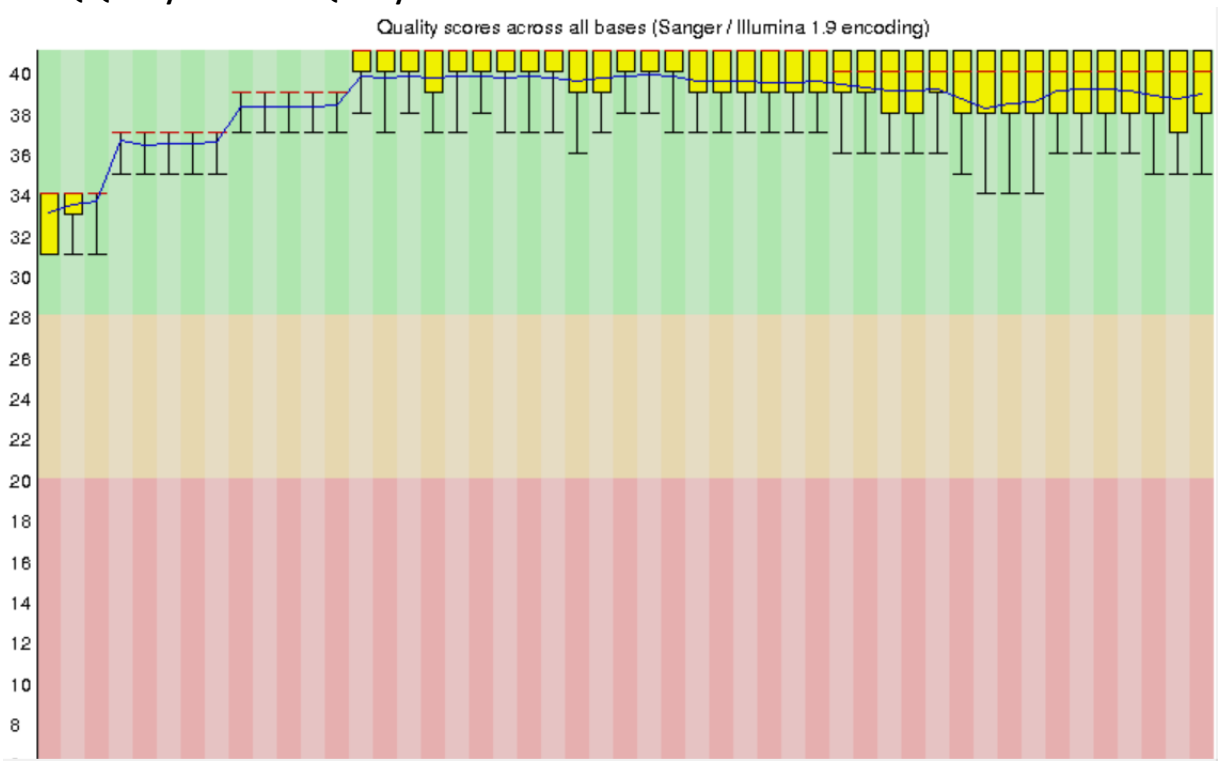
c. Running FASTQ Groomer to conver the phred quality scores and then rerunning FASTQC shows that the phred encoding scheme was converted to Sanger / Ilumina 1.9. The boxplot look almost exactly the same as the one in Part 2a.

Part 3

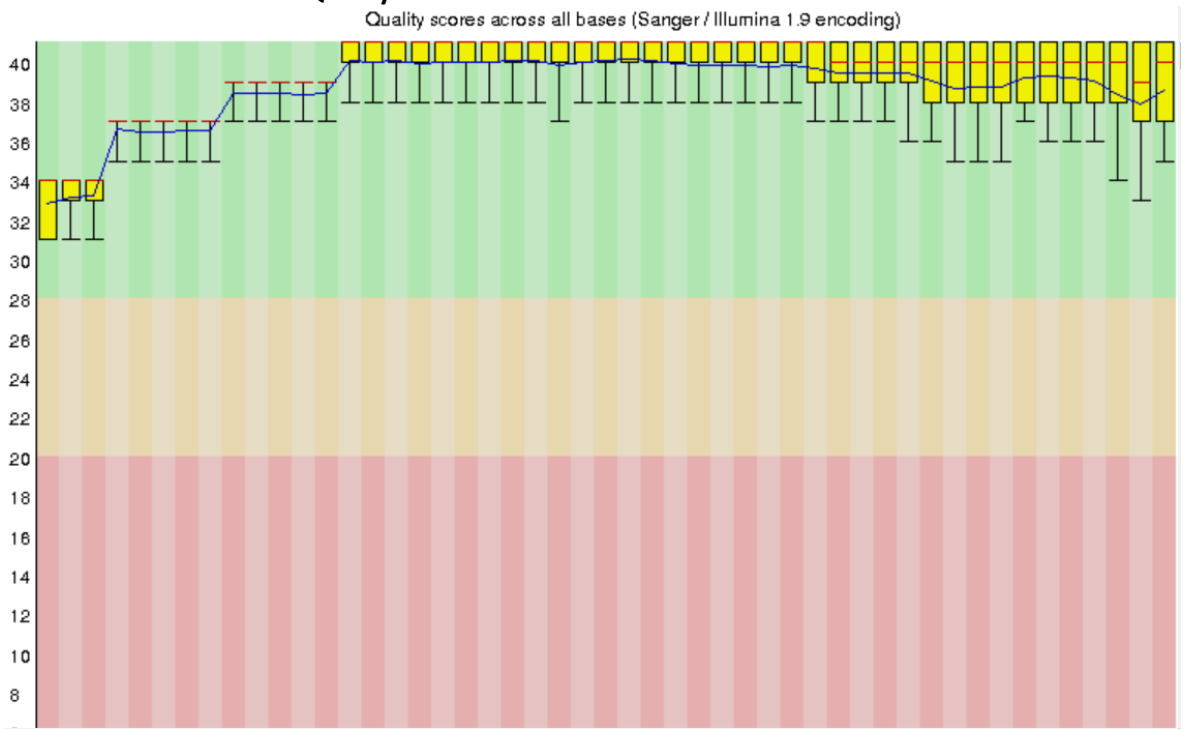
In order to compare the two trimmer tools in Galaxy, I ran FASTQ Quallity Trimmer to trim the data with a sliding window of 4 bases and trimmed the reads until the average quality score of the window is greater than 30. I then ran the Trimmomatic tool on the groomed data and used the same parameters.

a. To access the quality of each dataset, I ran FASTQC on both of the created datasets above. The box plots of quality scores are shown here:

FASTQ Quality Trimmer – Quality Score Box Plot



Trimmomatic Box Plot – Quality Score Box Plot



b. The data quality scores after trimmer are better than the ones that we saw in the boxplot from Part 2. For example, as the reads got longer in the untrimmed data, you can see the quality of the scores went down. The trimmed data has a more consistent quality score across the top of the boxplot. This does make sense to me, because we are trimming the low quality portion of the reads and leaving only high quality data.

c. The tool trimming tools seemed to give almost the same result, however quality scores seemed to be slightly better in the mid-length reads from Trimmomatic, while the longer length reads seemed to have better quality of FASTQ Quality Trimmer. Based on these boxplots alone, I would use either tool, but if I had to choose one it would be Trimmomatic because the the quality scores seemed to be slightly more consistent across all read lengths.

Part 4

To identify SNPs in two FASTQ files with paired-end sequencing data from the 1000 Genomes Project (reference genome hg19), I read the following data into Galaxy in the 'fastq' format with genome build 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

To determine quality encoding I ran FASTQC on each file. The encoding said Sanger/Illumina 1.9, but when I tried running Trimmomatic, it could not find any fastsanger datasets, so I ran FASTA Groomer on both files and then ran FASTQC on both groomed files.

Next, I trimmed the data using Trimmomatic to remove the low-quality bases from each file. I used a window size of 4 bases and a required average quality score of > 20. Then I reran FASTQC on the trimmed data to ensure that low quality bases were removed.

On the trimmed data, I ran HISAT2 to align the files to the hg19 reference genome, using paired-end reads and outputting in sorted BAM format. I attached the BAM file to my submission. I also wanted to note that I tried Bowtie2, and it took a lot longer, as suspected, since HISAT2 is advertised as using a fast algorithm.

I then ran the FreeBayes variant calling program, limiting the output to chr22:0-51304566. This algorithm showed 329 variants in the resulting VCF file, which I attached to my submission. I filtered the variants down to those that show heterozygosity with estimated allele frequency of 0.5 and read depth > 10 with VCFfilter. This narrowed the results down to 16. The filtered VCF file is also attached.

I downloaded RefSeq known genes from the UCSC Table Browser in the region of chr22:0-51304566. The BED file is attached to my submission. I uploaded the BED file to Galaxy and ran VCFannotate to intersect this BED file with the filtered VCF file. The goal was to locate variants within genes on chromosome 22. The final filtered, annotated VCF file was is attached to this submission.

a. There are 16 variants listed in the VCF file. Four of them were annotated with a RefSeq gene. Some of them were annotated with multiple genes.
b. I extracted my Galaxy Workflow and attached it to this submission.
c. I chose a SNP at postion chr22:39040545 and loaded it into the UCSC Genome Browser. The nucleotide change in the VCF file indicates a T to C polymorphism. T is shown as the nucleotide in the genome browser for reference genome hg19. It is located in the third intron of the FAM227A gene. Since it is in and intronic region within the gene, there may be no effect on the gene itself, unless it is in a regulatory region or in a splicing region, which it does not appear to be. There was some ENCODE data indicating a possible CTCF or RNAPol II binding site where the SNP was located. If this the case, the SNP could interfere with these proteins binding and alter the regulation of the gene.