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. (**Sanger/Illumina 1.9 encoding**)
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.
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.

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

  There are 32 variants listed in the VCF file. 24 variants were annotated with RefSeq genes.
  (VCF File attached)

- Extract and submit your Galaxy workflow. This is how I will be grading whether you followed the protocol appropriately.

  (Workflow file attached)

- 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?

Choosing a SNP, the position is 18880657 of chromosome 22. The gene affected is , and the nucleotide change is to a T instead of a C. The SNP is located in an intron region of gene FAM230F. Since it is in an intron, it could affect translation and gene function.

* If the analysis with full datasets takes too long (>24 hours), you may do the analysis with one million reads sampled from the full datasets. The samples fastq files are attached here (SRR044234_1.1M.fastq.gz and

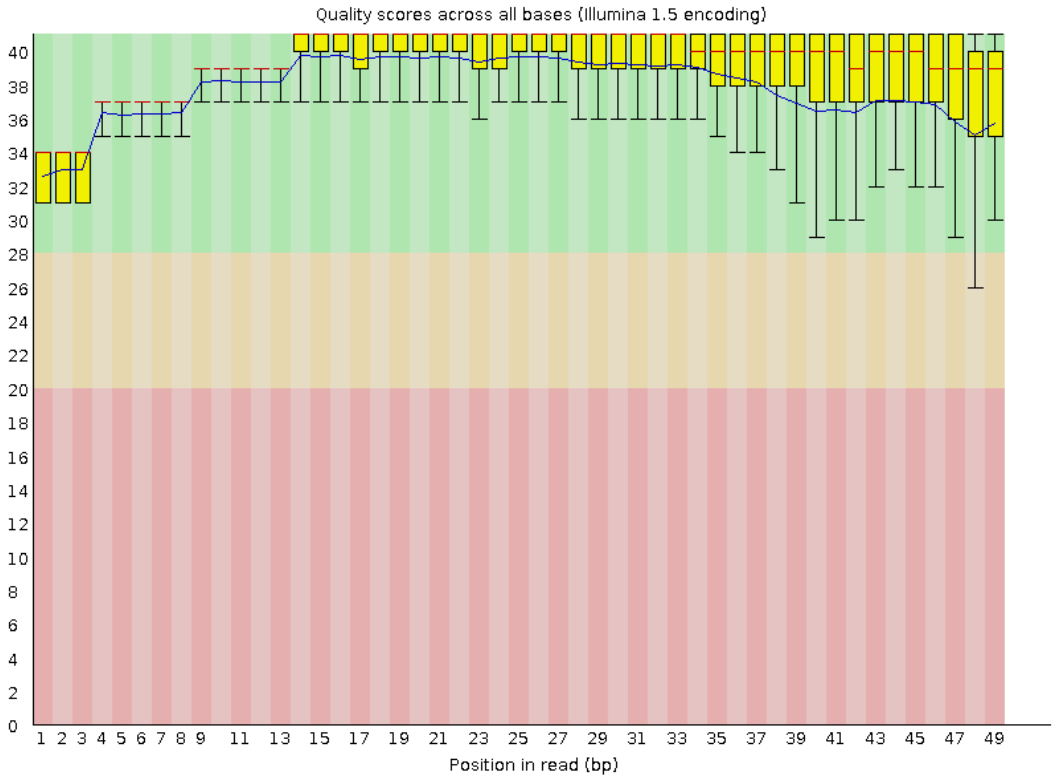
SRR044234_2.1M.fastqq.gz).
For your information, here is how I obtained sampled fastq files using my macbook.
- Install JAVA
- Install BBmap (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/installation-guide/)
- In the terminal, I ran the following command
reformat.sh in=SRR044234_1.filt.fastq.gz in2=SRR044234_2.filt.fastq.gz out=SRR044234_1.1M.fastq.gz out2=SRR044234_2.1M.fastq.gz samplereadstarget=1000000

HW5_Part2_sample_NGS_data.fq.gz
SRR044234_1.1M.fastq.gz
SRR044234_2.1M.fastq.gz

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).

- Run **FASTQC** and submit the boxplot of the quality scores. How would you describe the quality of these data?



I would describe the quality of this data as very good, as all the reads are in the green.

***Saved image also attached. ***

- What phred encoding scheme does this data use? How long are the reads? How many reads are in the file?

The encoding scheme is Illumina 1.5. The sequence length of the reads is 49. There are a total of 20,000 reads.

Basic Statistics

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

- 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?

Sanger/Illumina 1.9 is now the encoding listed.

Basic Statistics

Measure	Value
Filename	FASTQ Groomer on data 5
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20000
Sequences flagged as poor quality	0
Sequence length	49
%GC	46

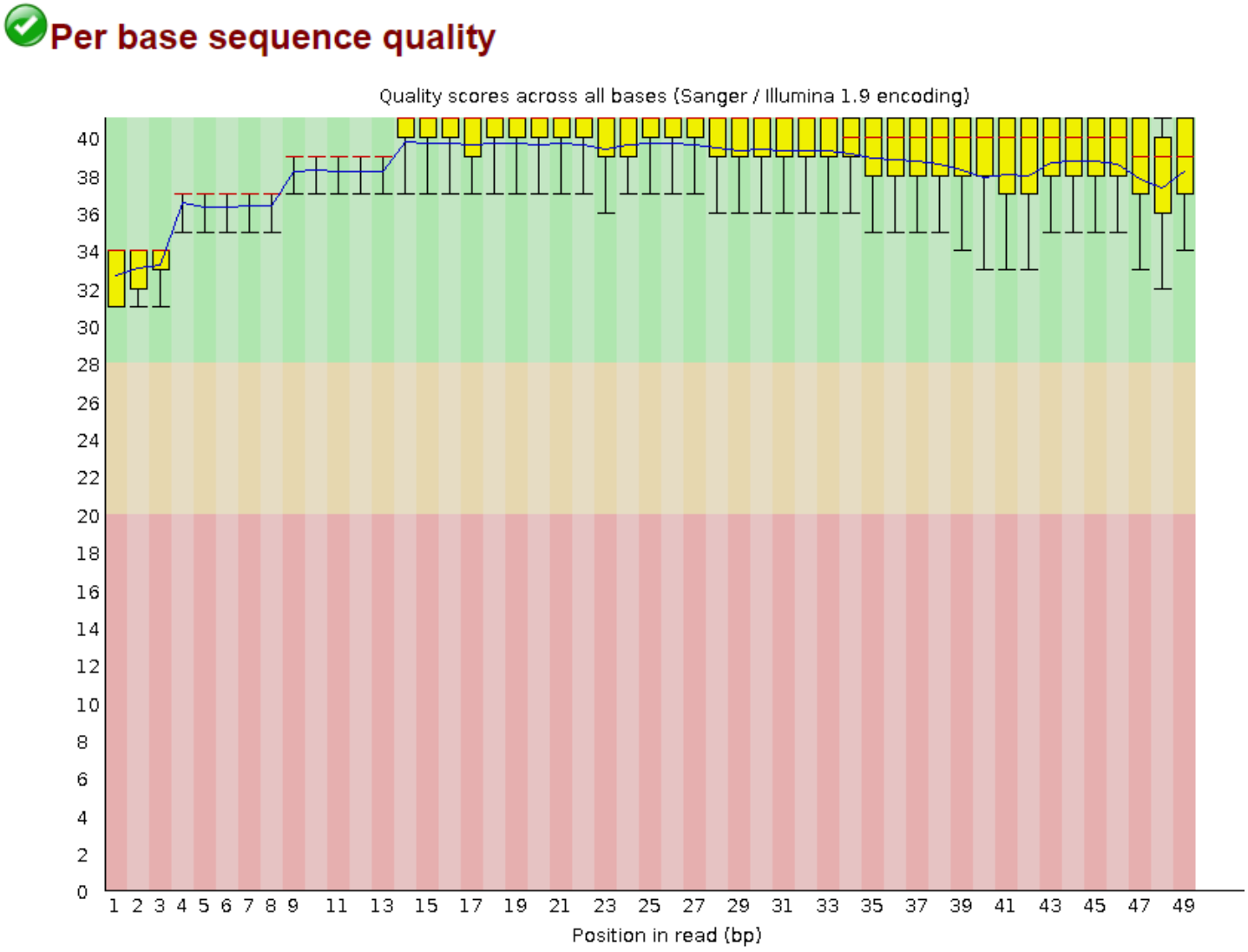
Part 3

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.

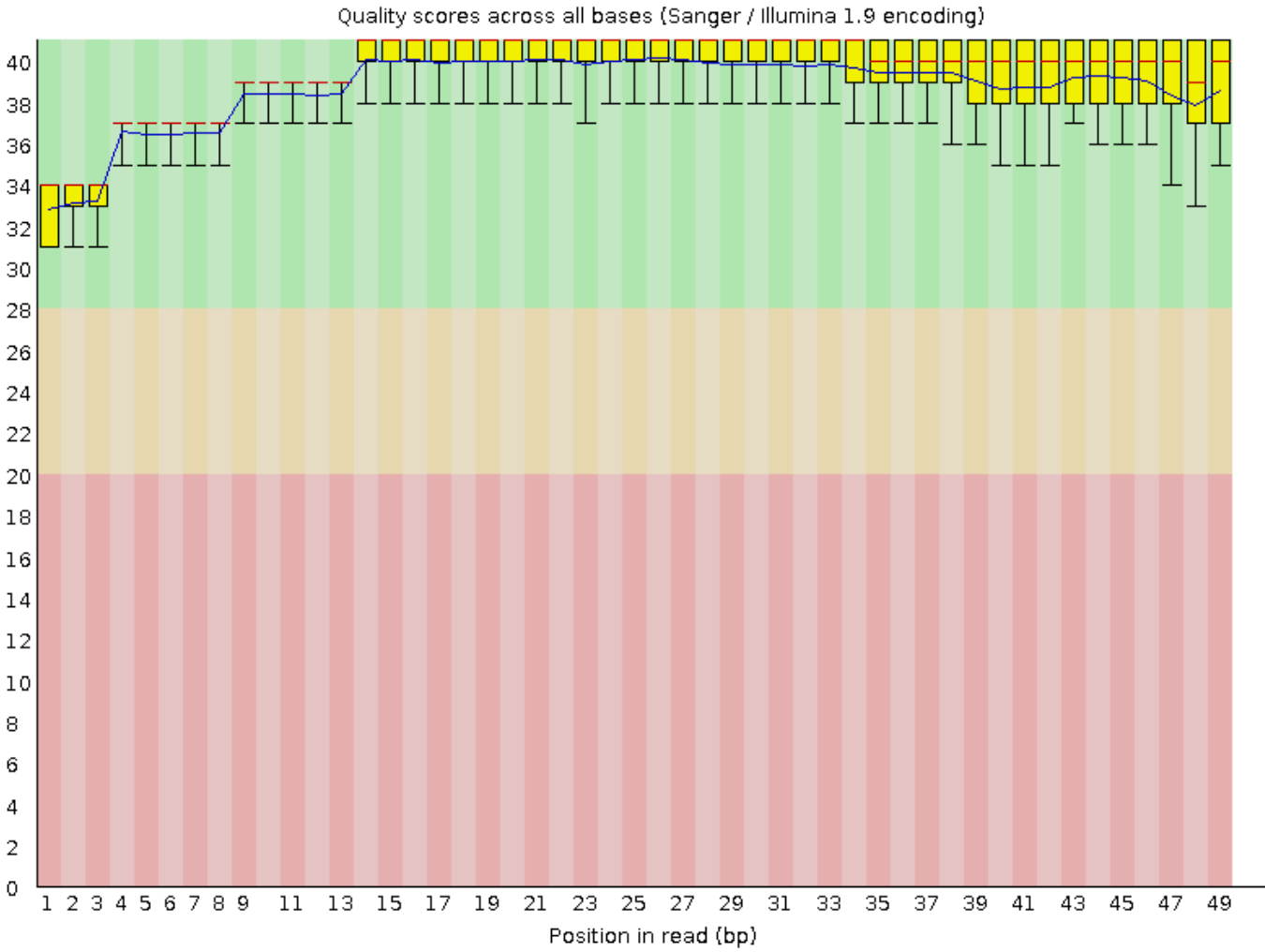
- Run the **FASTQC** tool on each of the FASTQ Quality Trimmer and Trimmomatic outputs. Submit both boxplots of quality scores. Be sure to label which boxplot is for data from which trimming tool.

FASTQ Quality Trimmer:



Trimmomatic:

Per base sequence quality



- 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?

The differences make sense. In general, the overall quality scores, including averages, are higher in the trimmed data than untrimmed. At particular position 39-49, in the untrimmed data there were some quality scores that were at or below a score of 30. With our settings, we wouldn't expect any below that average, and that is exactly what we see. We not only do not see averages below 30, but there aren't any scores below 30 within the range of any base pair.