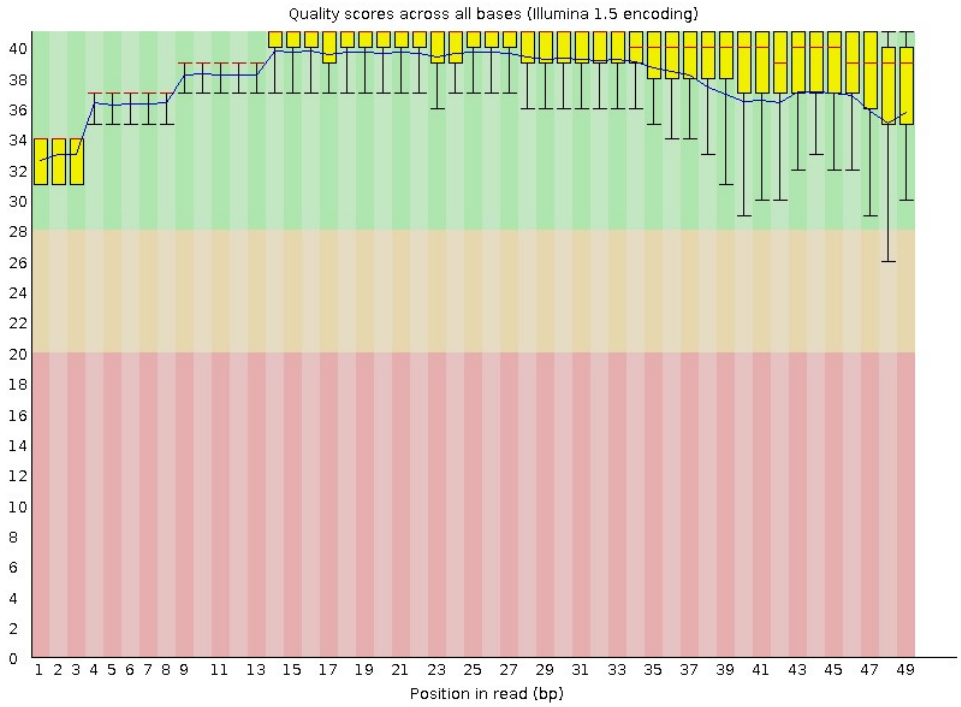
1. Upload provided data from a NGS experiment on C. elegans (genome version WS220/ce10).
   a. Run FASTQC and submit the boxplot of the quality scores. How would you describe the quality of these data?



Generally good quality

   b. 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

Encoding: Sanger/ Illumina 1.5
Sequence Length: 49
How many reads (Total sequences): 20,000

c. 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?

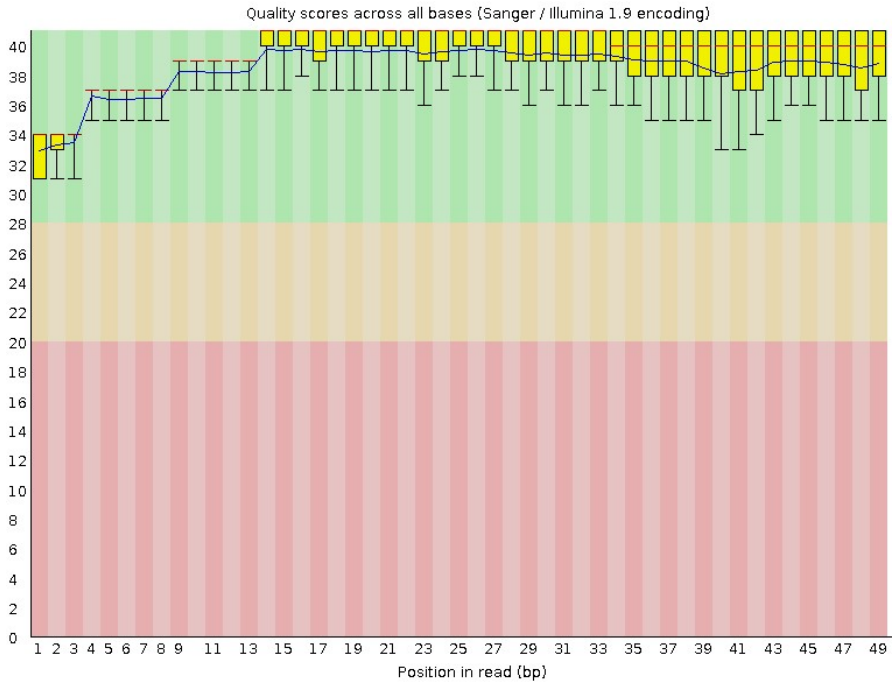
The encoding scheme is Sanger / Illumina 1.9

Basic Statistics

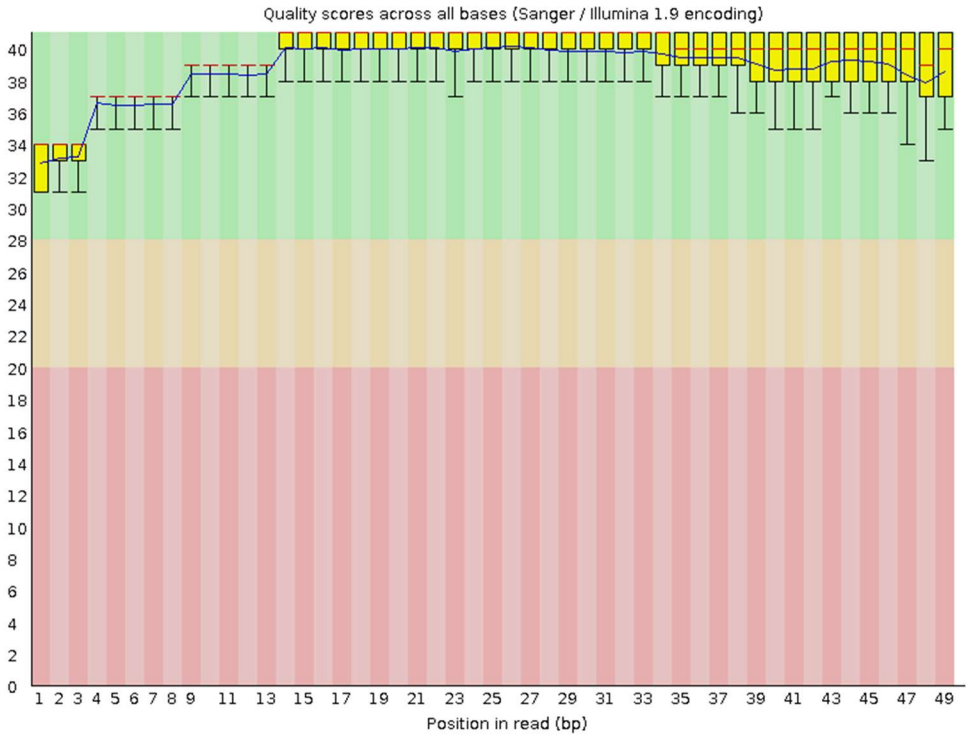
Measure	Value
Filename	FASTQ Groomer on data 1
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	20000
Sequences flagged as poor quality	0
Sequence length	49
%GC	46

2. 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. Run the Trimmomatic tool on the groomed data using the same parameters.
   a. 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.

FastQC on FASTQ Quality Trimmer:



FastQC on Trimmomatic:



3. 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.
   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.
   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.

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

   There are 51 variants listed in the VCF file and annotated with a RefSeq gene.

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

   Workflow can be made available upon request

c. 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?

   The SNP at chr22 16850188 changes a T to a G and is located in an intron