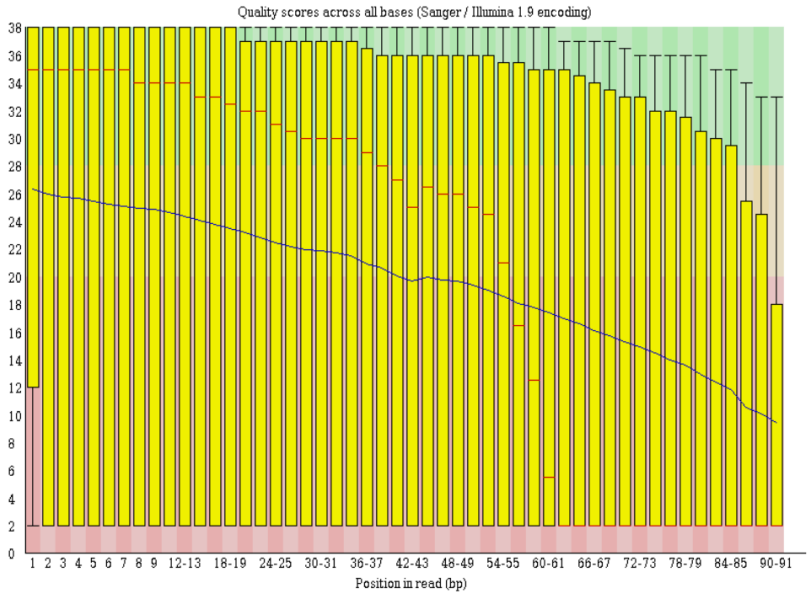
Julie Garcia
Exam #3
April 18, 2018

Part 1 - 8 points

In Galaxy, I ran FASTQC on the following file:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00324/sequence_read/ERR018456.filt.fastq.gz

1. Below is the boxplot of the quality scores. Quality scores are lower and more variable than I am used to seeing.



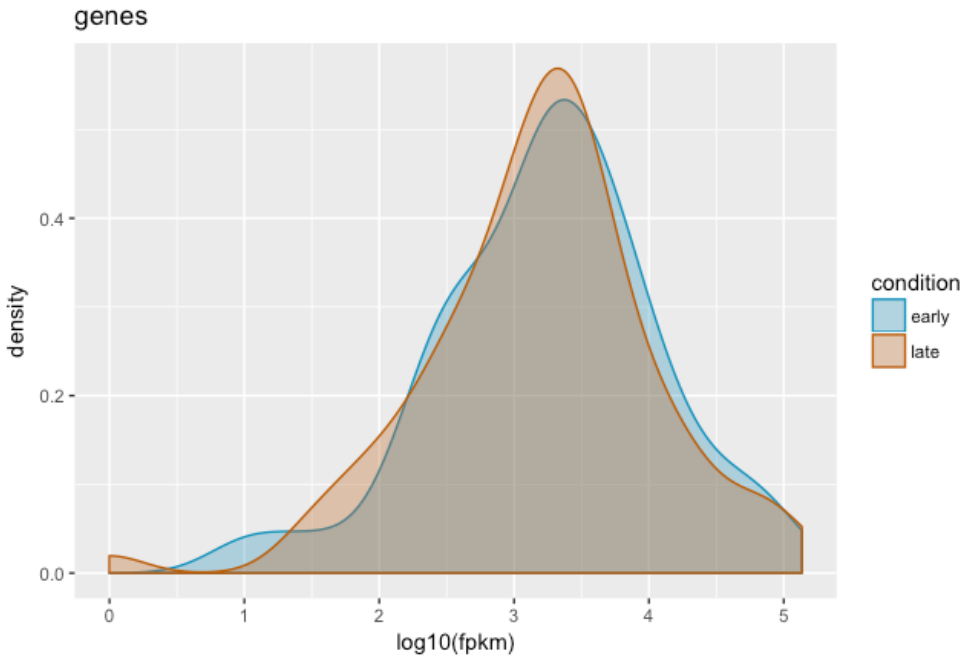
2. The read length is 91.
3. The read length indicates that this is Illumina data, which generally has read lengths up to 150, while Roche 454 can have much longer read lengths.
4. The most variability in sequence quality can be shown in positions 2 through 18 where the quality scores range from 2-38. As you go further along in the read the scores quality goes down, but so does the variability.
5. To get FASTQ Trimmer to work, I had to change the data type to fastqcsanger on the uploaded file. Then I ran FASTQ Trimmer to remove five nucleotides from the 3' ends of all reads by setting it to Absolute values and 5 nucleotides from the 3’ end. I then ran FASTQC on the trimmed data and the boxplot of quality scores can be seen below.

Part 2 - 12 points

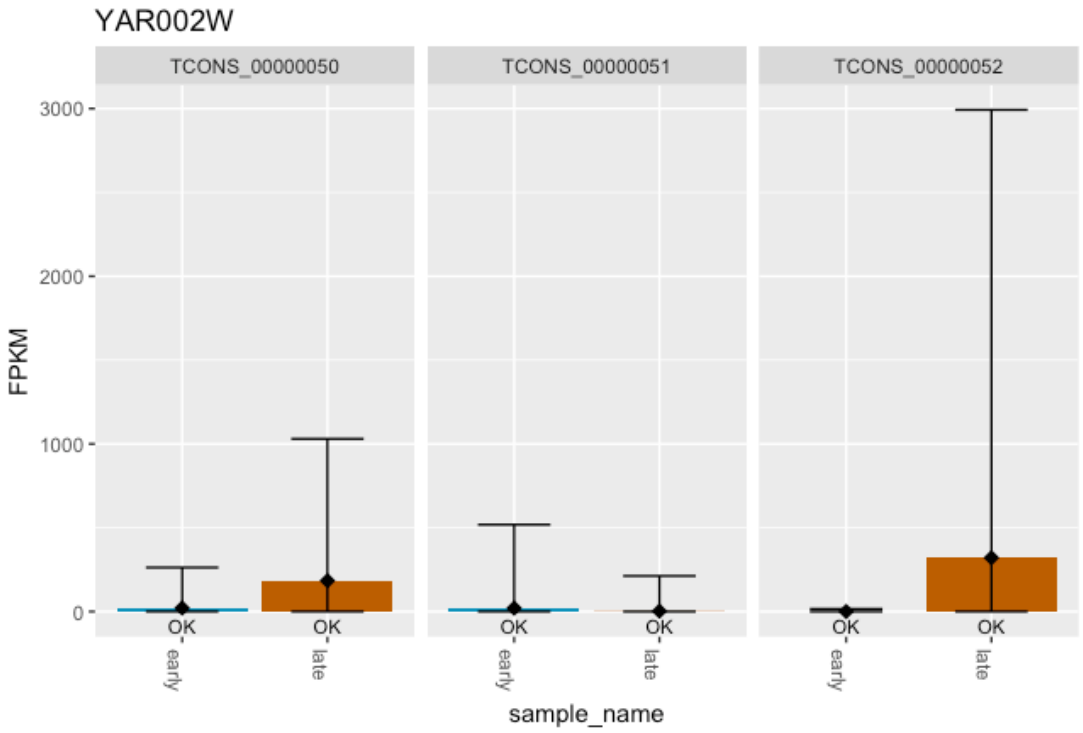
a. Below is an analysis pipeline for two paired-end RNA-seq files. The two cell lines represent Yeast transcriptome profiling in replicative aging with early (26.8 hours) and late (38.4 hours) RNA-seq data. First I copied the files to my user directory.

```
cp ERR905119_1.fq.gz ~/exam3/early_1.fq.gz
cp ERR905119_2.fq.gz ~/exam3/early_2.fq.gz
cp ERR905135_1.fq.gz ~/exam3/late_1.fq.gz
cp ERR905135_2.fq.gz ~/exam3/late_2.fq.gz

bowtie2-build chrI.fa chrI_genome
tophat2 -r 200 -o early chrI_genome early_1.fq.gz early_2.fq.gz
tophat2 -r 200 -o late chrI_genome late_1.fq.gz late_2.fq.gz

cufflinks -g sacCer3_chrI.gtf early/accepted_hits.bam
mv transcripts.gtf early_transcripts.gtf

cufflinks -g sacCer3_chrI.gtf late/accepted_hits.bam
mv transcripts.gtf late_transcripts.gtf

export PATH=/opt/cufflinks:$PATH
source .bashrc
cuffmerge -o merged -s chrI.fa -g sacCer3_chrI.gtf assemblies.txt

cuffdiff -o diff --labels early,late merged/merged.gtf early/accepted_hits.bam late/accepted_hits.bam
```

b. I attached the genes_fpkm_tracking file from my cuffdiff output to this submission.
c. Below is the density plot from the genes from cummeRbund.



d. Below is an expression bar plot for the isoforms of the TSS45 gene.



e. According to this plot, there are 3 isoforms of YAR002W.
f. There are five commands in this RNA-seq analysis pipeline. Bowtie2-build uses the Burrows-Wheeler transformation (BWT) to create an index for the reference genome. This speeds up

alignment in the following steps and allows for low memory (RAM) usage while running the subsequent tools. The paired-end fastq read files are then each run through tophat2 to align them to the reference genome that you supply using the indexes created with bowtie2. Cufflinks creates the GTF file, which is assembled transcript structures. Cuffmerge takes several different gtf files and merges them into one so that cuffdiff can then be run to identify differentially expressed genes. CummeRbund is used to visualize this cuffdiff data in plots using R.

Here is my cummeRbund code to generate the plots from the cuffdiff output:

```
#install and load cummeRbund
source("https://bioconductor.org/biocLite.R")
biocLite("cummeRbund")
library(cummeRbund)

#read all cuffdiff files in working directory
setwd("~/Datasets/diff")
cuff <- readCufflinks()
cuff

#density curve
csDensity(genes(cuff))

#find differentially expressed genes
gene_diff_data <- diffData(genes(cuff))
sig_gene_data <- subset(gene_diff_data, (significant == 'yes'))
nrow(sig_gene_data)
sig_gene_data

#print stats for the gene and count table for the isoforms of the gene
tss45_gene <- getGene(cuff, "TSS45")
tss45_gene
head(fpkm(isoforms(tss45_gene)))

# create an expression bar plot for gene isoforms
igb <- expressionBarplot(isoforms(tss45_gene), replicates=T)
igb
```

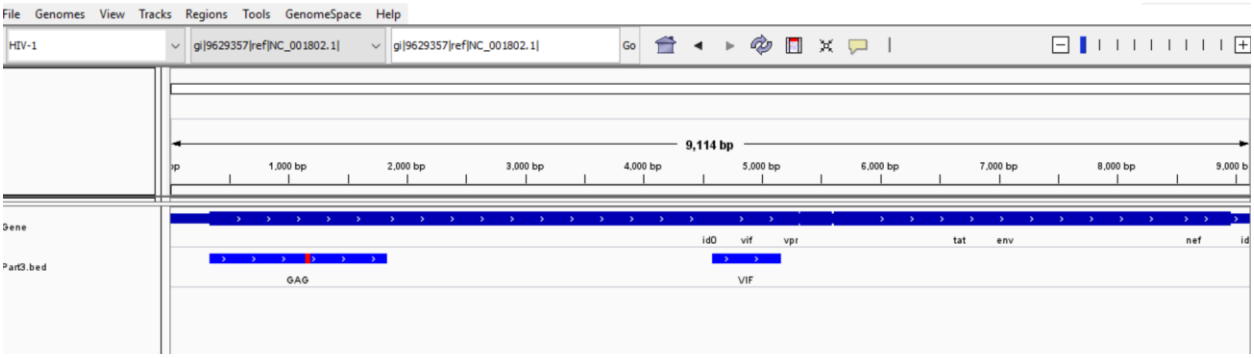
Part 3 - 6 points

I loaded the HIV-1 Genome into IGV and a custom BED file with the following intervals.

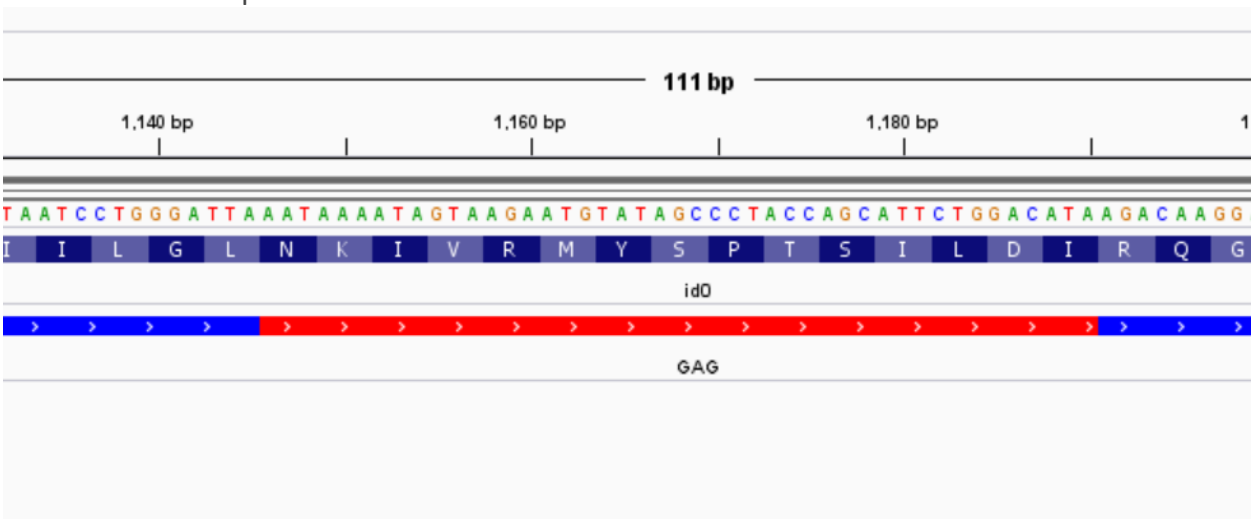
1. The gag gene located at positions 336 through 1838.
2. The vif gene located at positions 4587 through 5165.
3. A Gag protein potential epitope located at amino acid positions 271 through 285 of the Gag protein. The amino acid sequence is NKIVRMYSPTSILDI.

a. The BED file containing the 3 items above is Part3.bed and is attached to my submission.

b. The screenshot below shows all three intervals in IGV. GAG and VIF are in blue, while the potential Gag epitope is shown in red.



c. This screenshot shows IGV zoomed into the show the potential epitope of the Gag protein with the amino acid sequence NKIVRMYSPTSILDI.



Part 4 - 8 points

I loaded two ungroomed single-end FASTQ files with Illumina 1.5 phred encoding from a ChIP-seq mouse (mm9, chr19 only) experiment into Galaxy. In Galaxy, I ran the **FASTQ Groomer** tool to convert the reads to fastqsanger format. I then ran **Trimmomatic** on the groomed data to trim bases to equal a phred score 20 or greater. I aligned the trimmed reads to the mm9 reference with **Map with BWA** and ran **MACS2 callpeak** on the experimental ChIP-seq with the control output as the control.

a. In the tabular format at position chr19:37,340,170-37,340,721 (I did not have the exact interval in my file that was asked for, but this was very close), the fold enrichment is 27.50843.
b. After loading the Control and Treatment Bedgraph files into IGV and zooming to the interval above, you can see that there are two gene transcripts on the minus strand near this interval. The gene transcripts are 4931408D14Rik (NR_040298, NR_04299) and Ide (NM_031156) as shown in the image in Part 4d.
c. There is a MACS peak in the Treatment Bedgraph Track that is upstream of the gene 4931408D14Rik and downstream of the Ide gene.

d. Below is a screenshot from IGV showing both the MACS peak and the two genes described above.