Homework 4 – BIOS 7659

Dominic Adducci

Question 1: RNA-seq and QC

Notes on tools:

- Get Data
  - Went to “Get Data” option under tools.
  - Used “Faster Download and Extract Reads in FASTQ” tool.
  - Entered “SRR390924” and pressed “Run Tool” to download data.
  - Used “Email when done” option and let tool run for several hours.
  - Performed this around 11:00 am on a Friday, took a few hours to run. The queue may have been particularly busy given the time of day.
- Collapse Data
  - Went to “Collapse Operations” option under tools.
  - Selected “Collapse Collections” option from the tool bar.
  - Pressed button with ellipses to select the “SRR390924” and selected “Run Tool”.
  - This operation ran very quickly (Ran on a Saturday around 11:30 am).
- Summary Statistics – Part 1B
  - Went to the “Genomic File Manipulation” option under tools.
  - Selected “FASTQ Quality Control”, and then selected “FASTQ Summary Statistics” option.
  - Selected collapsed entry and pressed “Run Tool”.
- Quality Scores – Part 1C
  - Went to “Genomic File Manipulation” option under tools.
  - Selected “FASTQ Quality Control”, and then selected “Compute quality statistics”.
  - After running quality statistics the “Draw quality score boxplot” option was selected.
  - Both operations ran very quickly

Part A

The .fastq format is ASCII text which describes the sequence and quality. The format always includes 4 lines per sequence:

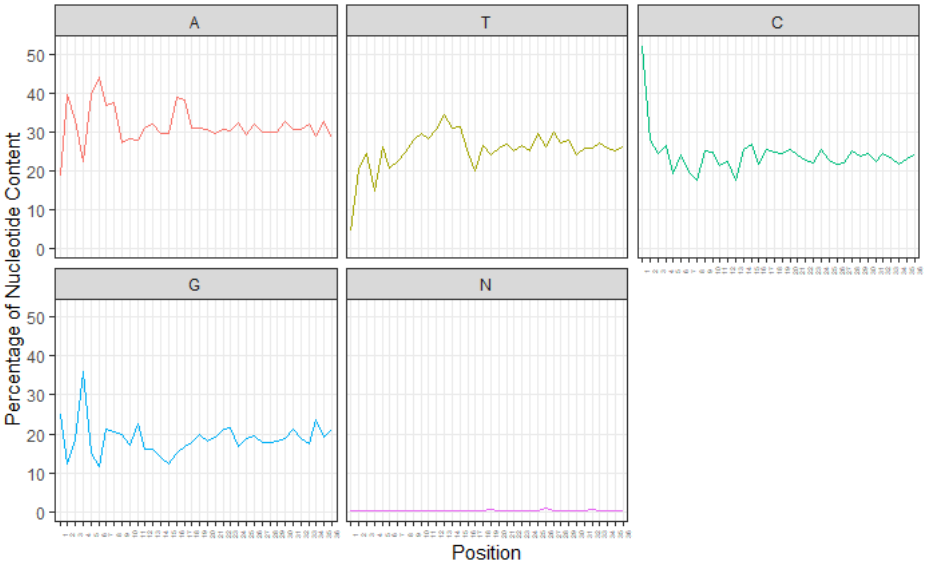
- Line 1: Begins with the ‘@’ character, a sequence ID, and may contain an optional description.
- Line 2: The sequence for the read.

- Line 3: Starts with the '+' character and is followed by the same sequence ID. Additionally, there may be an additional description which may be omitted to save space. If description is omitted the '+' character is kept.
- Line 4: Quality values are encoded in hexadecimal format for the sequence letters described in line 2. The number of hexadecimal characters must equal the number of characters in line 2. Hexadecimal format allows for representation of up to 92 numeric values, saving space.

```
@COLUMBO:1:1:1:1926/1
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
###################################
```

The above shows the first entry from the data set. The first entry has all '#' characters, which is the third lowest quality representation in the FASTQ format, out of 94 options. The length of the reads are 35. From clicking on the eye icon in the 'fasterq-dump log' portion of the history there are 3,614,610 reads.

Part B

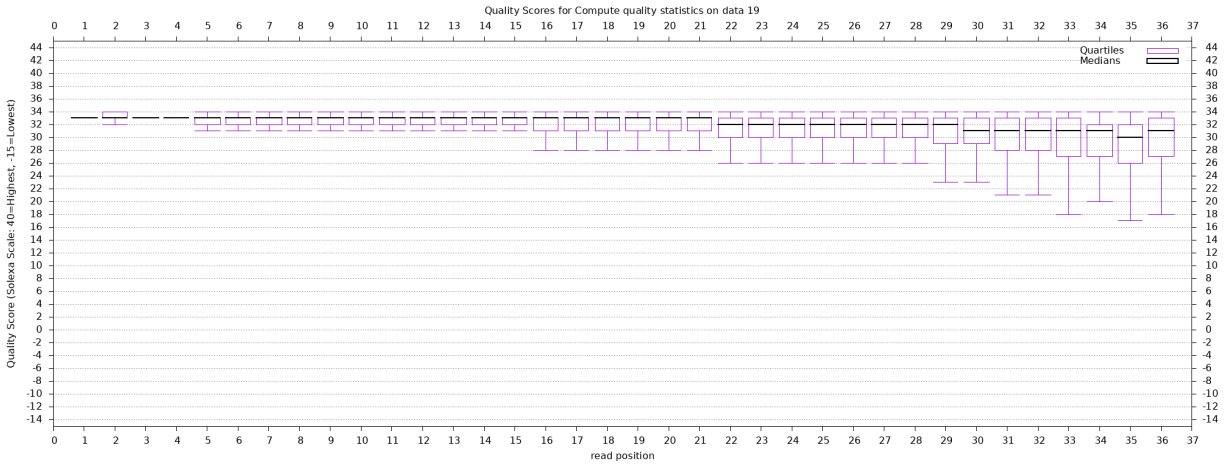


The above plots show the variance across position for each nucleotide, as well as ambiguous calls. Variance stabilizes when moving to the right-hand side of the plot. Adenine has the highest percentage content, while G tends to have lower percentage content. There are relatively very few ambiguous calls.

Columns:

- Column 1: Column number, 1 to 36 due to the 36-cycle read of Solexa files.
- Count: Number of bases in that column.
- min: Lowest quality score for a column.
- max: Highest quality score for a column.
- sum: Sum of quality scores for a column.
- mean: Mean of quality scores for a column.
- Q1: 1st quartile quality score.
- med: Median quality score.
- Q3: 3rd quartile quality score.
- IQR: Inter-Quartile range (Q1-Q3).
- lW: Left-Whisker, used for boxplotting
- rW: Right-Whisker, used for boxplotting
- outliers: Any values that fall beyond the whiskers, used for boxplotting
- A_Count: Count of Adenine nucleotides for a column.
- C_Count: Count of Cytosine nucleotides for a column.
- G_Count: Count of Guanine nucleotides for a column.
- T_Count: Count of Thymine nucleotides for a column.
- N_Count: Count of ambiguous reads for a column.
- other_bases: Other nucleotides found in the column.
- other_base_count: Count of other nucleotides found in the column.

Part C



Quality scores for all genes tend to be high, and variance increases when moving towards the right-hand side of the plot, as indicated by the extended boxplots.

Question 2: RNA-Seq Mapping using Bowtie2

Notes on tools:

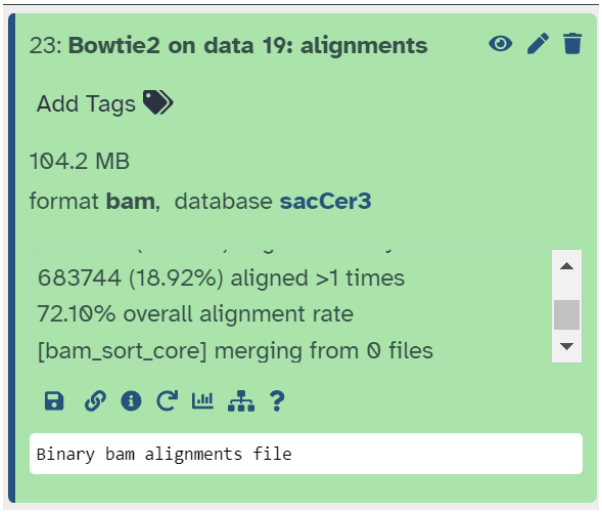
- Setting up Bowtie2
  - Selected “Genomics Analysis” option under tools.
  - Selected “Mapping” and then “Bowtie2”.
  - Collapsed entry was selected, “single-end” library was selected, and “sacCer3” was selected for “Select reference genome”
- Filtering Reads – Part 2B
  - Went to “Genomic File Manipulation” option under tools.
  - Selected “SAM/BAM” option, and then “Filter SAM or BAM, output SAM or BAM”.
  - Bowtie results were selected, and “yes” was set for “Filter on bitwise flag”.
  - Under the “Filter on bitwise flag” menu went to “Skip alignments with any of these flag bits set” and selected “The read is unmapped” and “the read fails platform/vendor quality checks”.
  - Output format was selected as “SAM”.
- Visualization of Mapping – Part 2C
  - Went to “Genomic File Manipulation”, and selected “SAM/BAM” and “SAM-to-BAM”.
  - For “Use a reference sequence” the “Use a built-in genome” option was selected, and Reference was set at “Yeast (Saccharomyces cerevisiae): SacCer3”.
  - Visualization was then performed, and the view was set to “squish”. This portion ran immediately.
- Quatitation for Genes – 2D
  - “Saccharomyces_cerevisiae.R64-1-1.85-v2.gtf” was downloaded from Canvas. The “Get Data” tab under tools was selected, followed by “Upload File”.
  - Next, went to “Genomics Analysis” in the tools bar, followed by selecting the “RNASeq” and “htseq-count” options.
  - This portion ran relatively quickly.

Part A

The above image shows the first read after Bowtie2 mapping. The mapping status of the first read is found in the “FLAG” column, which in this case is number 16. From the reference guide “16” means that SEQ is being reverse complemented.

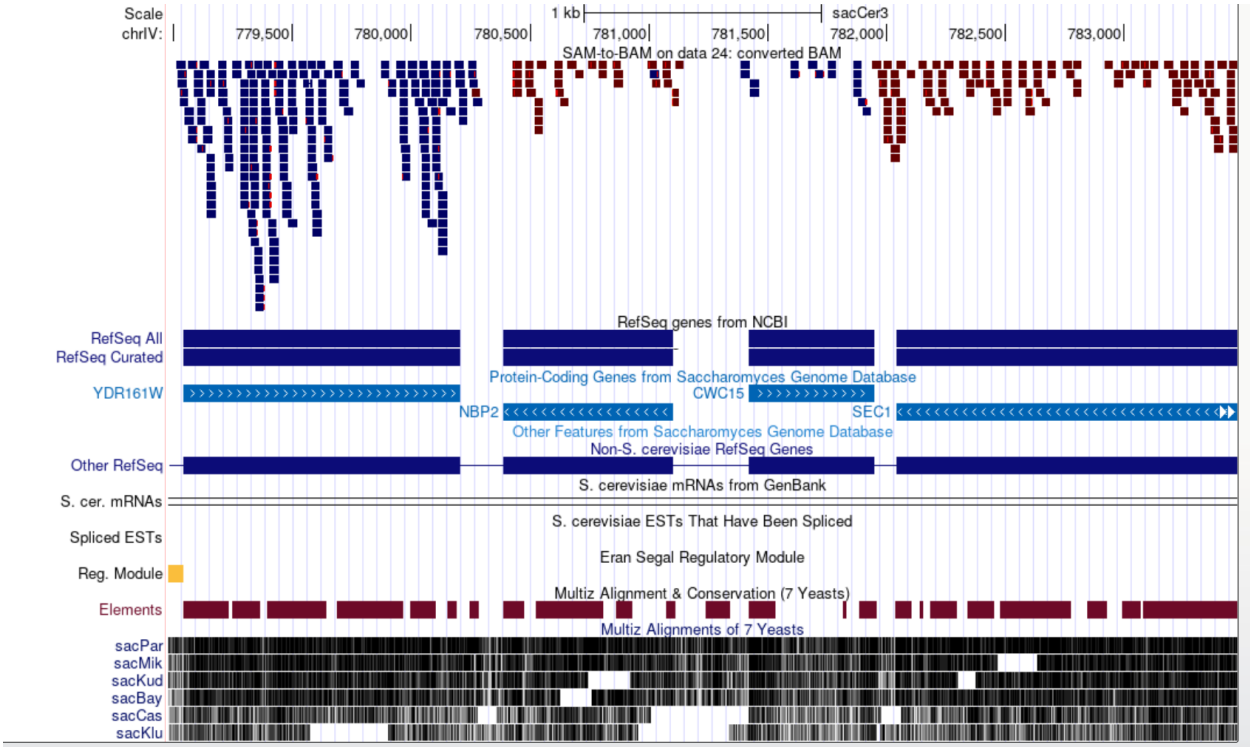
Bowtie2 is a short-read sequence aligner used for mapping sequence reads to reference genomes. The benefits of Bowtie2 are that it is fast, accurate, and less computationally demanding than direct alignment. Bowtie2 works by preprocessing the reference genome into an FM-index data structure, allowing for faster searching. This process is done using the Burrows-Wheeler transformation (BWT).

Part B



Around 28% of the genes were filtered out. This was determined from the overall alignment rate of 72.10% from the above image.

Part C



The above image shows the results of the gene visualization. From the upper portion of the image, we can tell that chromosome 4 (chrIV) is being viewed. In this part of the data there are 4 different genes shown (YDR161W, NBP2, CWC15, SEC1). Genes YDR161W and SEC1 have the highest coverage.

Part D

Geneid	SAM-to-BAM on data 24: converted BAM
15S_rRNA	0
21S_rRNA	0
HRA1	0
ICR1	38
LSR1	36
NME1	0
PWR1	0
Q0010	0
Q0017	0
Q0032	0
Q0045	0

The above image shows the first few results returned from the HT-seq analysis. The right column indicates the number of counts for each gene given the number of aligned reads that overlap its exons.

CODE

```
library(tidyverse)


### START QUESTION 1 CODE ###

## START QUESTION 1 PART B CODE ##

# Loading in summary statistic data
summary_stat_data <- read.delim("C:/Users/domin/Documents/Biostatistics Masters Program/Spring

# First column name came in as X.column. Changing that
colnames(summary_stat_data)[1] <- "Column"


# Transforming data into long format
summary_stat_data_2 <- summary_stat_data %>%
  select(Column, A_Count, C_Count, G_Count, T_Count, N_Count) %>%
  pivot_longer(cols = ends_with("Count"),
               values_to = "Counts",
               names_to = "Nucleotide",
               names_pattern = "(.)_Count") %>%
  mutate(Column = factor(Column),
         Nucleotide = factor(Nucleotide, levels = c("A","T","C","G","N")))


# Making spaghetti plots
summary_plots <- summary_stat_data_2 %>%
  ggplot(aes(x = Column, y = (Counts/3614610)*100,group = Nucleotide,
             col = Nucleotide)) +
  geom_line() +
  facet_wrap(~ Nucleotide) +
  labs(x = "Position",y = "Percentage of Nucleotide Content") +
  theme_bw() +
  theme(legend.position = "none", panel.grid.minor = element_blank(),
        axis.text.x = element_text(angle = 90, size = 4))

summary_plots

## FINISH QUESTION 1 PART B CODE ##

### FINISH QUESTION 1 CODE ###
```