BIOS 7659 Homework 4

Tim Vigers

20 October 2020

# 1. RNA-seq Data and QC

## a) Read information

The first entry in the .fastq file is:

`@SRR390924.1.1 COLUMBO:1:1:1:1926 length=36

AAAAAAAANAAAAAAAAAAAAAAAAAAAAAAAAAAA

+SRR390924.1.1 COLUMBO:1:1:1:1926 length=36

####################################`

The first line contains the sequence ID (after the @ symbol) and an optional description. Line 2 contains the read sequence and line 3 has the same sequence ID (this time after the + symbol) followed by another optional description. The final line encodes a quality score for each base pair (BP) in the sequence using the hexadecimal format. In this dataset, reads are 36 BP long.

For more recent Illumina systems, you can find the q-score for each BP by subtracting 33 from the ASCII code. From the q-score you can then calculate the probability that the BP call was incorrect using the formula . For for the first entry in this file, , so 0.631. This is not a high quality read.

According to the SRA entry there are 3,614,610 reads in this file.

## b) Summary statistics

“FASTQ Summary Statistics” returns a table where each row represents a BP position within a read (in this case 1-36). For each BP position, the table includes quality summary statistics such as minimum, maximum, and mean q-score. In general the summary statistics are fairly standard and self-explanatory, but the column descriptions from the usegalaxy.org documentation are as follows:

column = column number (1 to 36 for a 36-cycles read Solexa file)

count = number of bases found in this column.

min = Lowest quality score value found in this column.

max = Highest quality score value found in this column.

sum = Sum of quality score values for this column.

mean = Mean quality score value for this column.

Q1 = 1st quartile quality score.

med = Median quality score.

Q3 = 3rd quartile quality score.

IQR = Inter-Quartile range (Q3-Q1).

lW = ‘Left-Whisker’ value (for boxplotting).

rW = ‘Right-Whisker’ value (for boxplotting).

outliers = Scores falling beyond the left and right whiskers (comma separated list).

A\_Count = Count of ‘A’ nucleotides found in this column.

C\_Count = Count of ‘C’ nucleotides found in this column.

G\_Count = Count of ‘G’ nucleotides found in this column.

T\_Count = Count of ‘T’ nucleotides found in this column.

N\_Count = Count of ‘N’ nucleotides found in this column.

Other\_Nucs = Comma separated list of other nucleotides found in this column.

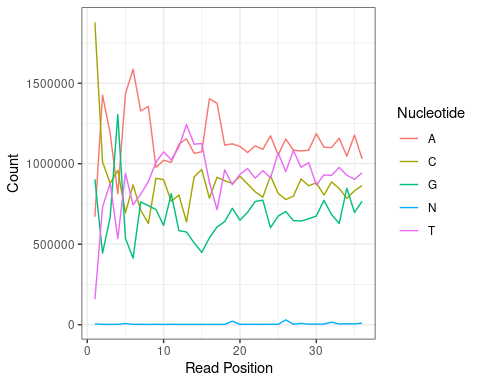
Other\_Count = Comma separated count of other nucleotides found in this column.

sum\_stat <- read.delim("./FASTQ\_Summary\_Statistics.tabular")  
colnames(sum\_stat)[1] <- "position"  
kable(head(sum\_stat,5))

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| position | count | min | max | sum | mean | Q1 | med | Q3 | IQR | lW | rW | outliers | A\_Count | C\_Count | G\_Count | T\_Count | N\_Count | other\_bases | other\_base\_count |
| 1 | 3614610 | 2 | 33 | 113732847 | 31.46476 | 33 | 33 | 33 | 0 | 33 | 33 | 2,4,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32 | 670260 | 1877729 | 903145 | 158171 | 5305 | NA | NA |
| 2 | 3614610 | 2 | 34 | 114293473 | 31.61986 | 33 | 33 | 34 | 1 | 32 | 34 | 2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31 | 1426604 | 1010947 | 444249 | 730545 | 2265 | NA | NA |
| 3 | 3614610 | 2 | 34 | 113169385 | 31.30888 | 33 | 33 | 33 | 0 | 33 | 33 | 2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,34 | 1188225 | 877986 | 667041 | 879162 | 2196 | NA | NA |
| 4 | 3614610 | 2 | 34 | 113623731 | 31.43458 | 33 | 33 | 33 | 0 | 33 | 33 | 2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,34 | 813065 | 959180 | 1305248 | 534975 | 2142 | NA | NA |
| 5 | 3614610 | 2 | 34 | 111459496 | 30.83583 | 32 | 33 | 33 | 1 | 31 | 34 | 2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30 | 1434968 | 694033 | 536934 | 941327 | 7348 | NA | NA |

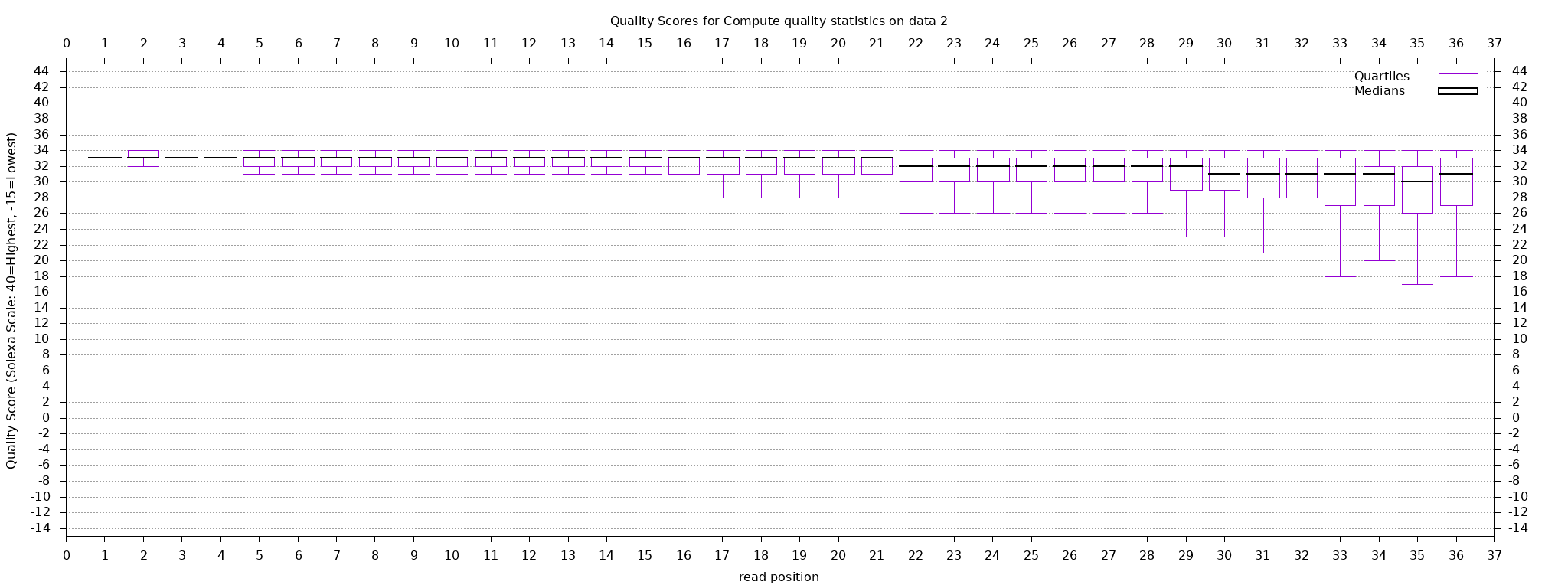
### Nucleotide content by position

plot\_sum\_stat <- sum\_stat %>%   
 pivot\_longer(cols = A\_Count:N\_Count)  
ggplot(plot\_sum\_stat,aes(x=position,y=value,color=name)) +  
 geom\_line() +  
 xlab("Read Position") + ylab("Count") +   
 scale\_color\_discrete(name = "Nucleotide",  
 labels=c("A","C","G","N","T")) +  
 theme\_bw()



Early on in the reads there appear to be a large number of cytosines and not many thymines. However, as the read position increases the proportions seem to stabilize and there are generally more adenosines and thymines, as one might expect. Also, there appears to be more variability early on in the reads, and it decreases as the length of the read increases.

### Quality score boxplot



As the read length increases, the quality tends to deteriorate. Based on visual inspection it seems that the drop in overall quality occurs around position 29. The variability in quality also increases as the read length increases.

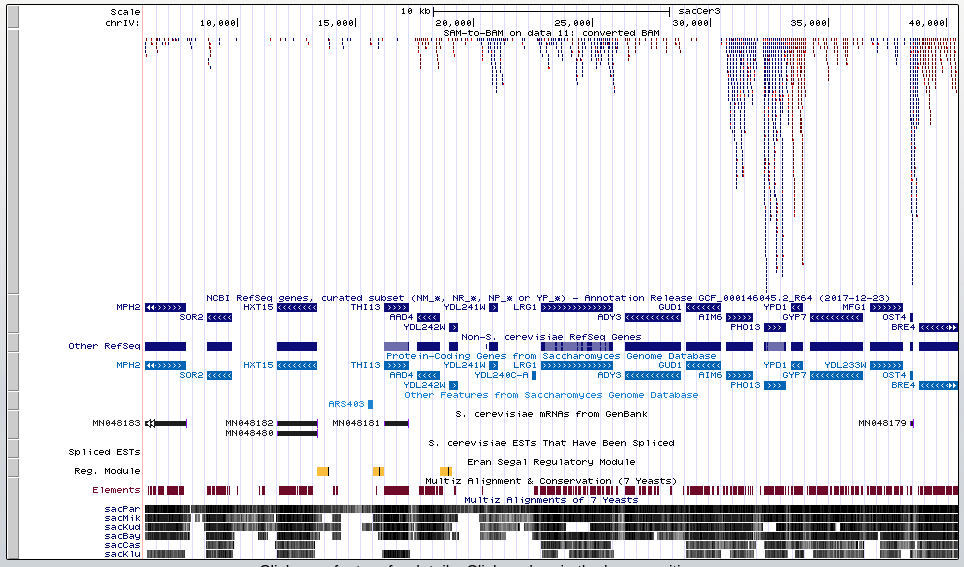
# 2. RNA-seq Mapping using Bowtie2

## a) Bowtie2 Mapping

The first two reads have the flag 16, which corresponds to “SEQ being reverse complemented” (i.e. mapped to the reverse strand). However, the third read has flag 4 which corresponds to “segment unmapped.” Filtering out the unmapped reads and reads which failed quality control checks results in ~2,600,000 reads (which is about 72% of the original number of reads).

## b) Visualization of Mapping

The UCSC squish view is showing chromosome IV. Based on the RefSeq track (the one directly below our data track), there appear to be 16 yeast genes in this section of the chromosome. The genes are MPH2 through BRE4, and it appears that the genes toward the right of the visualization have the highest coverage, particularly AIM6, PHO13, and YPD1. OST4 and BRE4 also appear to have good coverage, although not quite as much as the genes between 30kbp and 35kbp.



Screenshot of UCSC Genome Browser on S. cerevisiae Apr. 2011

## c) Quantitation

The htseq function returns two tables, one with a summary of the number of aligned or ambiguous reads and another with read counts for specific features (in this case gene IDs).

# Import no features  
no\_feat <- read.delim("./htseq-count\_no\_feature.tabular",header = F)  
kable(no\_feat,col.names = c("Category","SAM-to-BAM on data 11: converted BAM"),  
 caption = "htseq Summary Table")

htseq Summary Table

|  |  |
| --- | --- |
| Category | SAM-to-BAM on data 11: converted BAM |
| \_\_no\_feature | 197763 |
| \_\_ambiguous | 7719 |
| \_\_too\_low\_aQual | 582447 |
| \_\_not\_aligned | 0 |
| \_\_alignment\_not\_unique | 0 |

# Counts per features  
feat <- read.delim("./htseq-count.tabular",header = F)  
kable(head(feat[order(feat$V2,decreasing = T),],10),  
 caption = "htseq Top 10 Features by Read Count",  
 col.names = c("Geneid","SAM-to-BAM on data 11: converted BAM"))

htseq Top 10 Features by Read Count

|  |  |  |
| --- | --- | --- |
|  | Geneid | SAM-to-BAM on data 11: converted BAM |
| 3013 | YHR174W | 39192 |
| 2640 | YGR192C | 37184 |
| 3794 | YKL060C | 34937 |
| 107 | YAL038W | 28748 |
| 4423 | YLR249W | 25601 |
| 4195 | YLR044C | 22660 |
| 729 | YCR012W | 20836 |
| 5710 | YOL086C | 19958 |
| 2166 | YGL008C | 19017 |
| 3894 | YKL152C | 16013 |