**BIOS 7659 HW4**

**Guannan Shen**

**(a) Read Information**

The .fastq format contains 4 lines per entry.

Line 1 begins with a “@” is sequence ID and an optional description.

Line 2 is the raw sequence, containing “A”, “T”, “C”, “G”.

Line 3 begins with the “+” character, followed by the same sequence ID and another optional description. Also the whole content can be removed but keep the “+”.

Line 4 encodes quality values in hexadecimal format for the raw sequence in line 2, which is the Phred Score and calculated from error probability (p), q = -10log10(p). Illumina “quality scores greater than 30” are “high quality”.

The first entry is:

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

AAAAAAAANAAAAAAAAAAAAAAAAAAAAAAAAAAA

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

####################################

This is a low quality read, the “#” represents the third lowest quality level. There are dozens of quality levels. In addition, in this case, ASCII range is from # (35) to C (67) and the decimal range is from 2 to 34. Hence, the # represents the lowest quality scores, less than 30 and is very poor quality.

The length is 36, single-end (by the paper).

The total number of reads in the file is 3614610.

**(b) Summary Statistics**

Columns returned by “FASTQ Summary Statistics” are as following:

Column = column number, in this case, 1 to 36 for a 36-cycles read. Count equals 3614610, number of reads in this file.

The following columns: min, max, sum, mean, Q1 (1st quartile), med (median), Q3 (3rd quartile), IQR (Q3 – Q1), lW ('Left-Whisker' value for boxplotting), rW ('Right-Whisker' value for boxplotting) and outliers (Scores falling beyond the left and right whiskers (comma separated list)). These columns are all about quality scores.

In this case, based on the Q1, Q3, lW, rW and outliers, we can tell:

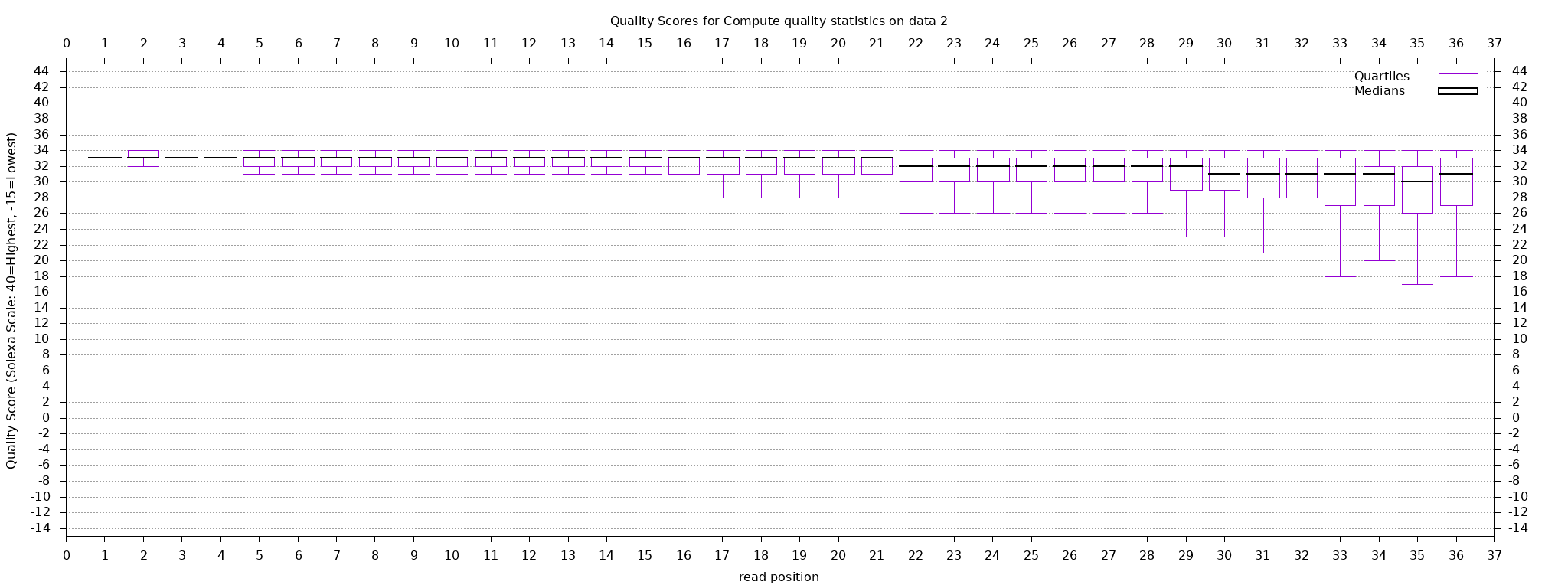
1) As the cycle number increases, the quality score decreases. Start from cycle 16, the lW values fall below 30, from cycle 29, the Q1 values fall below 30.

2) The first 15 cycles, overall quality is between 31 – 34, with only few outliers, indicating high quality.

3) The general quality of this experiment.

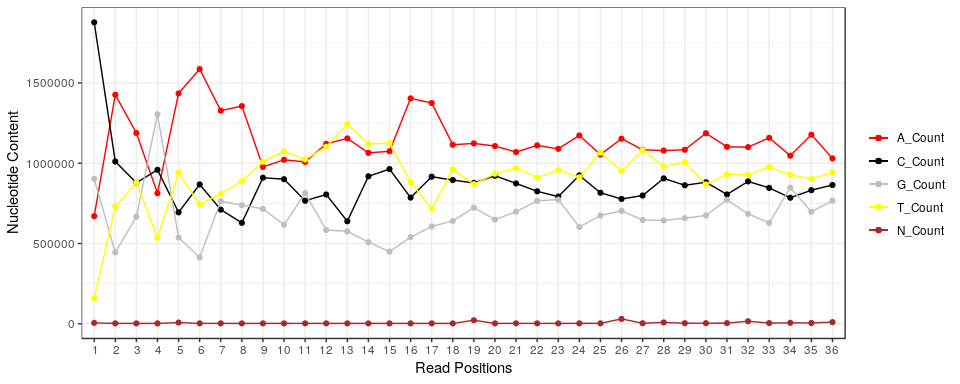
4) 36 cycles represent 36 position, as a biology background student, 36 positions are 36 bases in a read from the beginning to the end as single-end experiment.

The following 5 columns, A\_Count, C\_Count, G\_Count, T\_Count and N\_Count. These are counts of corresponding nucleotide found in this column. The last 2 columns, Other\_Nucs is Comma separated list of other nucleotides found in this column and Other\_Count is Comma separated count of other nucleotides found in this column.



**the quality scores by position in the read**

The trend shown in this plot provides the same information as the quality score columns in the summary statistics table. Across read positions, as the cycle increases, the medians of quality score decrease and the ranges of quality score are more spreading out. In other words, towards the end of the read, the quality drops. Overall, the medians of quality scores for all position are above 30, indicating overall good quality.



The code is on the following page.

Overall, the unspecified base, N, is very low compared with specified A, C, G and T contents. In terms of A, C, G and T, the proportions of these components vary a lot at the beginning and become stable flat lines after around the 18th base. Initially, the G and C counts are higher than A and T, after some fluctuations, the G and C contents fall below A and T.

7659\_HW4

Guannan Shen

October 15, 2018

Table of Contents

## set up workspace  
library(knitr)  
library(tidyverse)  
options(stringsAsFactors = F)  
options(dplyr.width = Inf)  
getwd()

## [1] "/home/guanshim/Documents/Stats/CIDA\_OMICs/7659Stats\_Genetics/HW4"

gal\_sum <- read.table("Galaxy3\_FASTQ\_Summary\_Statistics\_on\_data\_2.tabular")  
ggplot(gal\_sum, aes(x = V1)) + geom\_line(aes(y = V14, color = "A\_Count")) +   
 geom\_point(aes(y = V14, color = "A\_Count")) + geom\_line(aes(y = V15,   
 color = "C\_Count")) + geom\_point(aes(y = V15, color = "C\_Count")) +   
 geom\_line(aes(y = V16, color = "G\_Count")) + geom\_point(aes(y = V16,   
 color = "G\_Count")) + geom\_line(aes(y = V17, color = "T\_Count")) +   
 geom\_point(aes(y = V17, color = "T\_Count")) + geom\_line(aes(y = V18,   
 color = "N\_Count")) + geom\_point(aes(y = V18, color = "N\_Count")) +   
 scale\_x\_discrete(name = "Read Positions", limits = c(1:36)) +   
 scale\_y\_continuous(name = "Nucleotide Content") + theme\_bw() +   
 scale\_colour\_manual("", breaks = c("A\_Count", "C\_Count",   
 "G\_Count", "T\_Count", "N\_Count"), values = c(A\_Count = "red",   
 C\_Count = "black", G\_Count = "grey", T\_Count = "yellow",   
 N\_Count = "brown"))

**2. RNA-seq Mapping using Bowtie**

**(a) Bowtie Mapping:**

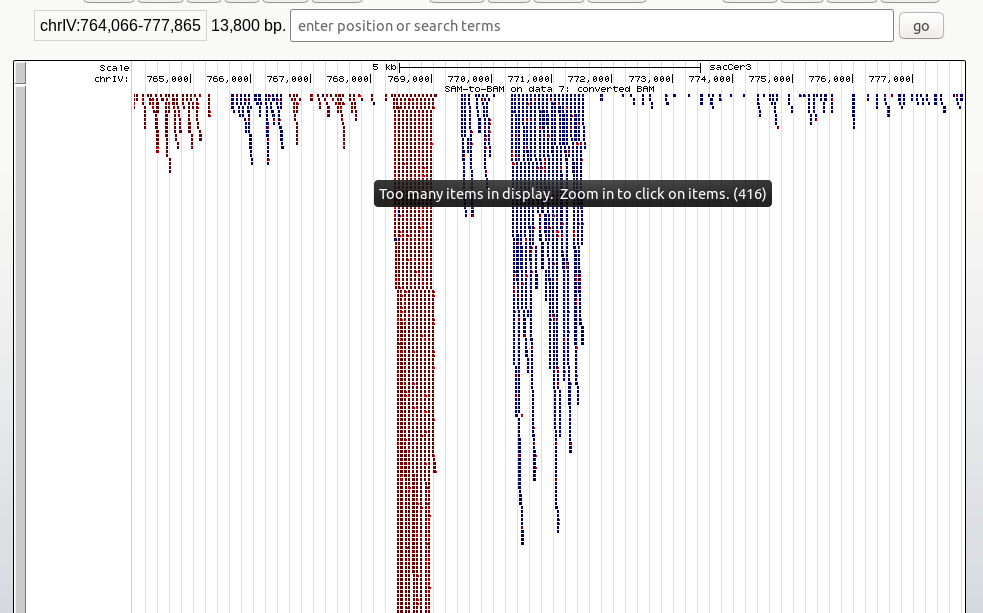
Bowtie: Reads aligned one nucleotide at a time searching against transformed genome. It is faster than seed methods but less sensitive.

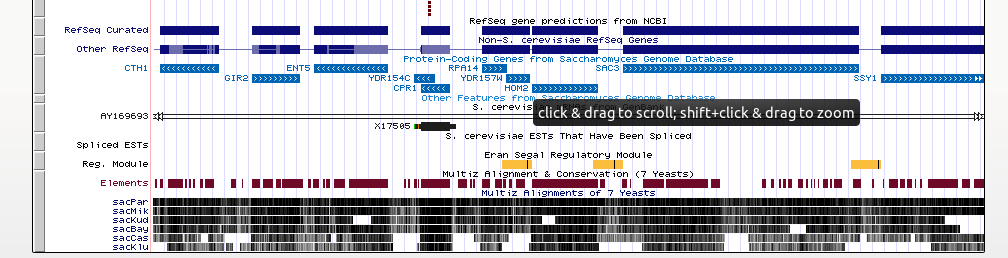
The first read listed in the .sam file, SRR390924.5.1 COLUMBO:1:1:1:764 length=36, is unmapped. The FLAG in the file is the combination of bitwise FLAGs. For this one, FLAG 4 means segment unmapped. The third one SRR390924.1.1 is the first mapped with FLAG 16 and mapped to chromosome XIV.

There are about 3,700,000 lines in the unfiltered .sam file and about 2,600,000 lines in the filtered file. Thus, 1,100,000/3,700,000 about 29.7% of reads were filtered out.

**(b) Visualization of Mapping:**

I was viewing chrIV:765,966-775,965 (10,000 bp). 10 protein coding genes (8 pieces of gene) was on this part of the chromosome: CTH1, GIR2, ENT5, YDR154C, CPR1, RPA14, YDR154W, HOM2, SAC3 and SSY1.





Based on the above plots, YDR154C, CPR1 and HOM2 have much higher coverage.

**(c) Quantitation for Genes:**

The HT-seq returns the list of number of reads in a unit, as shown below. The left column is the gene\_id and the right column is the number of reads. This quantitation was done on the gene as the unit level. Some genes such as RPM1, RPR1 have no mapped reads, while genes such as RDN5-1 to RDN5-6 have more than a thousand reads mapped on each gene.

