**Homework 4 Solutions**
**BIOS-7659/CPBS-7659**

## 1) RNA-Seq Data Initial Analysis

Introduction: An RNA-Seq experiment results in millions of sequencing reads. There are special formats for storing read data, which will be described. There are also possible biases in nucleotide content and positions across reads, which will be assessed with a sample data set.

a) The fastq format is used to store sequence reads in a text-based format, along with a header and quality information for each read. The header is the first line and is denoted by the @ character. It contains a sequence ID and optional descriptions such as the length of the read. The second line consists of the sequence read, which is a string of nucleotides (A,C,G,T and N for an ambiguous position). The third line starts with the + symbol and is optionally followed by the same sequence ID and description from line 1. The fourth line encodes the quality values of the each base in the sequence in a hexadecimal format. Quality scores range from 30 to 100 and thus using two digits is not a viable option. Therefore, a hexadecimal format is used so that a single symbol corresponds to the quality value of each base.

In the uploaded data set, the first read in the sequence has a symbol of "#" for each base in the sequence. This symbol has corresponding ASCII code of 35, corresponding to quality score of 2, which is considered low quality (https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm). The length of each read in the file is 36 and there 3,614,610 total reads. The first entry is displayed below:

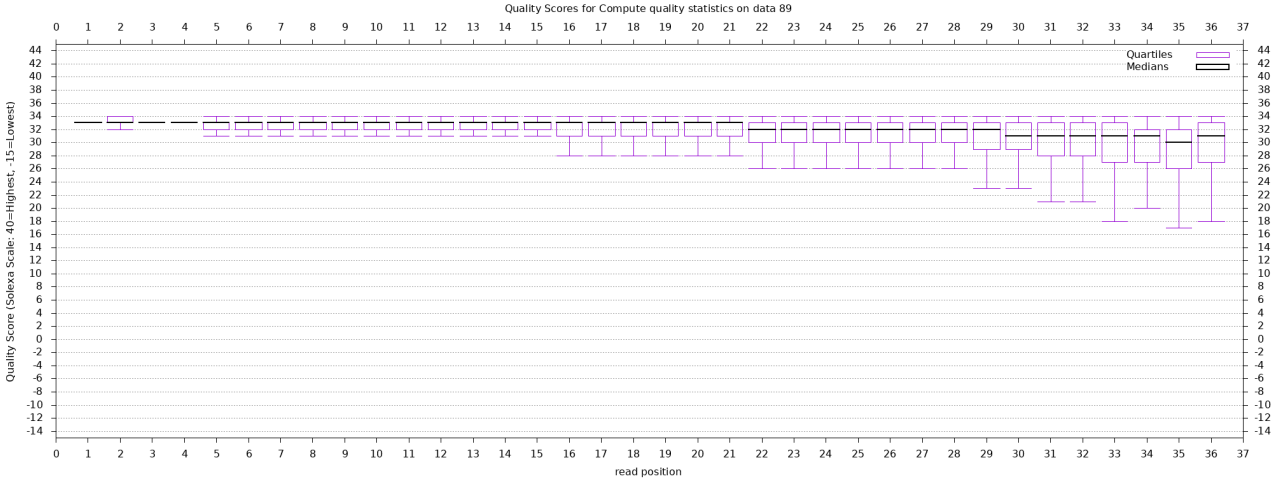
```
@SRR390924.1 COLUMBO:1:1:1:1926 length=36
AAAAAAAANAAAAAAAAAAAAAAAAAAAAAAAAA
+SRR390924.1 COLUMBO:1:1:1:1926 length=36
####################################
```

b) For each position, the fastq summary statistics from Galaxy returns summary statistics for the quality scores (columns 3-13) and nucleotide counts (14-20) that can be used to evaluate for any positional biases. Columns are listed below:

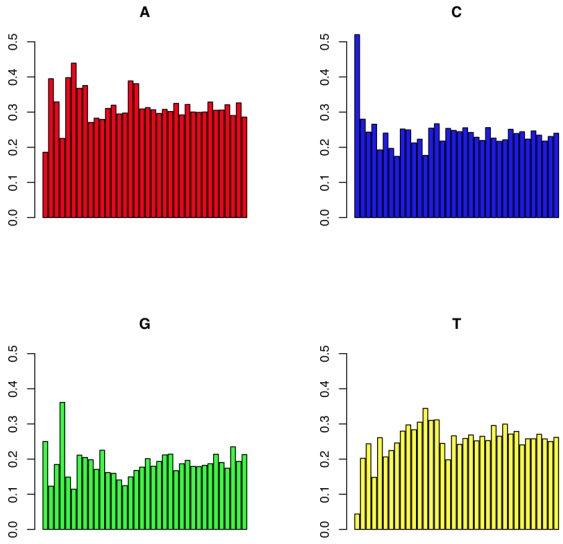
column = column number (1 to 36 for a 36-cycles read Solexa file)
count = number of bases found in this column.
min = Lowest quality score value found in this column.
max = Highest quality score value found in this column.
sum = Sum of quality score values for this column.
mean = Mean quality score value 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.

The boxplot for quality scores across positions is also returned and displayed below. As the position increases, the quality score median decreases and the variance increases, indicating lower and more variable quality for later positions.




The frequency of each nucleotide by position is displayed below. There appears to be more variation in the first 5-10 positions. In particular, for position 1, C appears almost 50% of the time, while T appears less than 5%. A and G appear 18-25%. For later positions there seems to be some preference for A (~30%), while there seems to be less G (~20%), and C and T appear at the expected amounts (~25%).



```
Code:
u<-read.table("Galaxy39-FASTQ_Summary_Statistics_on_data_36.tabular",sep= "\t")
ntcounts = u[,14:17]
ntfrac = t(apply(ntcounts,1,function(x) x/sum(x)))
par(mfrow = c(2,2))
barplot(ntfrac[,1], col = "red", main = "A", ylim = c(0,.5))
barplot(ntfrac[,2], col = "blue", main = "C", ylim = c(0,.5))
barplot(ntfrac[,3], col = "green", main = "G", ylim = c(0,.5))
barplot(ntfrac[,4], col = "yellow", main = "T", ylim = c(0,.5))
```

## 2) RNA-Seq Mapping using Bowtie

Introduction: After initial quality control for the raw reads, the next step for an RNA-Seq analysis is mapping reads to the genome. We will map our sample data to the yeast genome and discuss the mapping format file. A visualization of the mappings will also be presented for a representative region of the genome. A final step is quantitation, where counts of reads are assigned to genes (or other genomic features).

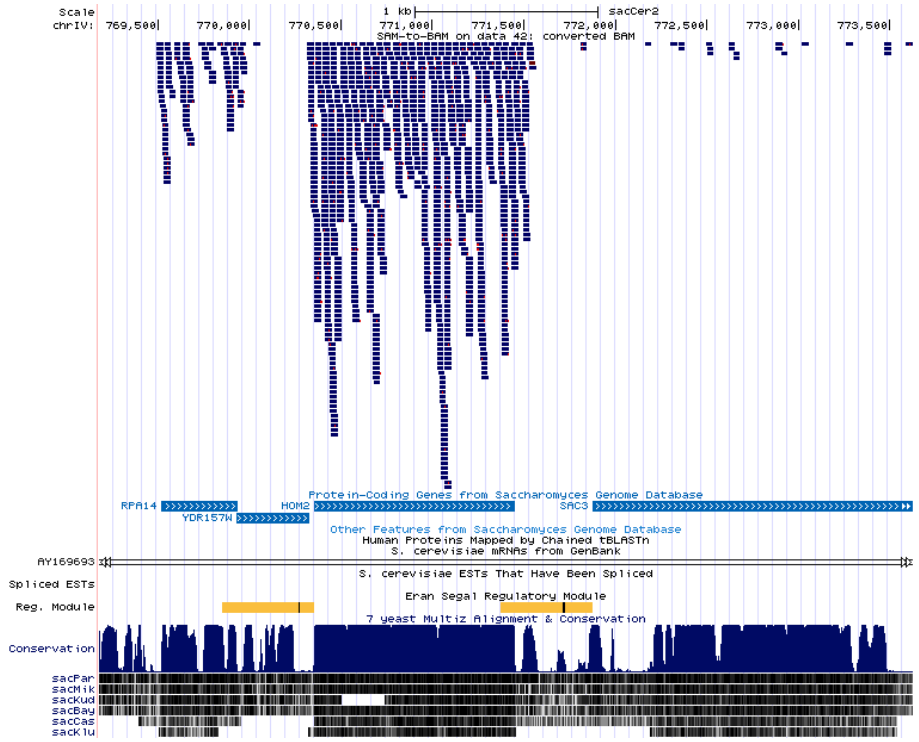
a) After mapping the reads to the yeast genome using Bowtie, the output is in the sam format and the first few lines are displayed below:

| QNAME (read identifier) | Flag | RNAME (chromosome) | POS |
|---|---|---|---|
| SRR390924.1 | 16 | chrVIII | 549764 |
| SRR390924.4 | 16 | chrI | 223119 |
| SRR390924.2 | 16 | chrXIV | 359354 |
| SRR390924.11 | 16 | chrXIV | 359353 |
| SRR390924.10 | 16 | chrXIV | 359354 |
| SRR390924.13 | 16 | chrI | 223119 |
| SRR390924.14 | 16 | chrXIV | 359352 |
| SRR390924.15 | 16 | chrXIV | 359353 |
| SRR390924.3 COLUMBO:1:1:1:1701 length=36 | 4 | * | 0 |
| SRR390924.7 COLUMBO:1:1:1:1609 length=36 | 4 | * | 0 |

This output indicates that the first 8 reads were mapped on the reverse strand (denoted by flag level of 16) and the corresponding chromosome and positions in the yeast genome are displayed. The 9$^{th}$ and 10$^{th}$ reads are not mapped, denoted by flag level of 4. For those entries, neither chromosome nor position is given.

After filtering for unmapped reads, there are about ~2,600,000 lines out of the ~3,600,000 indicating that roughly 28% of reads were filtered.

b) After selecting to view the mapped results at UCSC Genome Browser, a segment of the yeast genome on Chromosome 4 (~769K to ~774K) is displayed (see below) by using the “squish” view for this track. Four genes are contained in this area and one gene (HOM2) has many more reads and deeper coverage than the other genes even though it is not the largest gene segment in this region.




c) Quantitation was performed using HT-Seq. Some example features are displayed. Most have zero counts, but genes such as ICR1 and LSR1 have 38 and 36 read counts respectively. Tables like this can be generated for multiple samples, which can then be used to test for differential expression.

| Feature | Count |
|---|---|
| 15S_rRNA | 0 |
| 21S_rRNA | 0 |
| HRA1 | 0 |
| ICR1 | 38 |
| LSR1 | 36 |
| NME1 | 0 |
| PWR1 | 0 |
| Q0010 | 0 |
| Q0017 | 0 |
| Q0032 | 0 |