# Lab 04

Rezaur Rashid

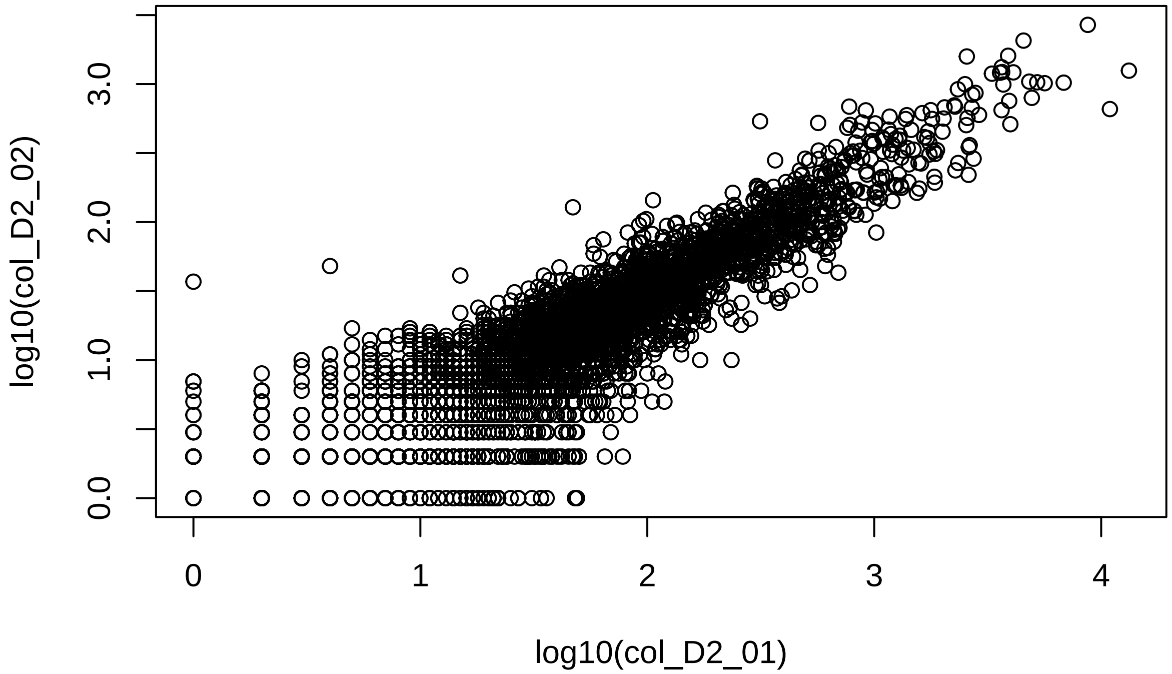
2022-03-03

## Part-01: read the data set

```
myTable = read.table('data/longitdunalRNASeqData/nc101_scaff_dataCounts.txt',
                     header = T, row.names = 1)

# dim(myTable)
# head(myTable)
# colnames(myTable)
```

## Part-02: plot D2_01 and D2_02 on log10-log10 scale

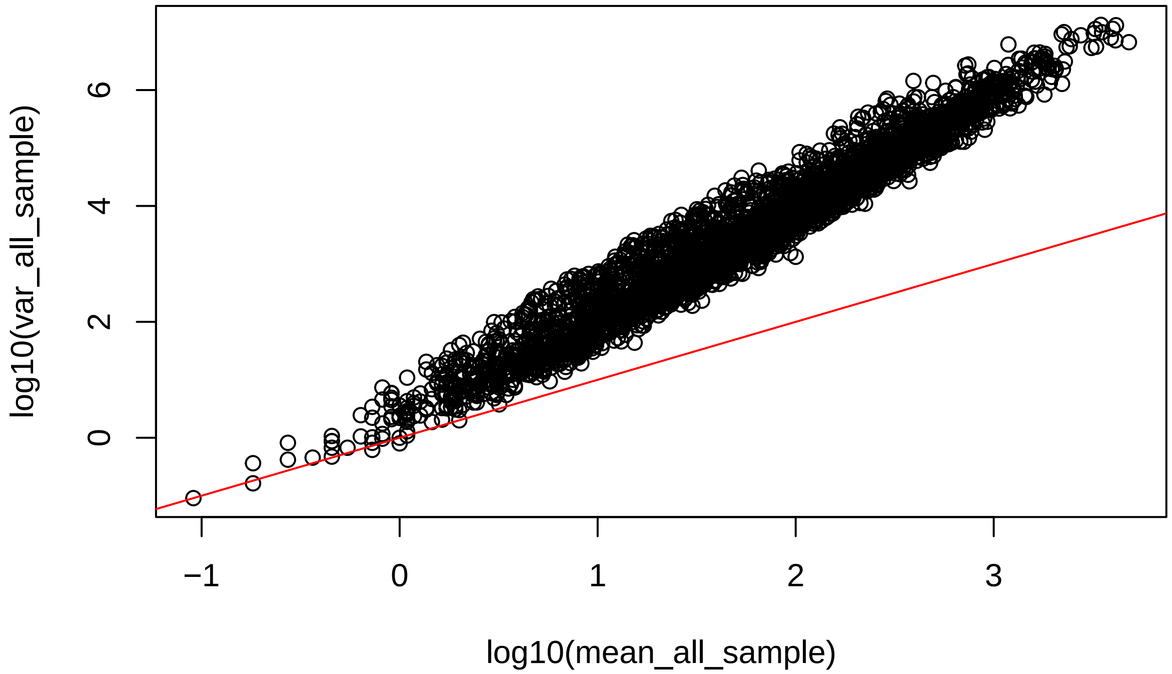
```
col_D2_01 = myTable[ , c("D2_01")]
col_D2_02 = myTable[ , c("D2_02")]

plot(log10(col_D2_01), log10(col_D2_02))
```



Qualitatively, from the plot, we see that both the replicas has a linear relationship among them in gene expression. Thus we can assume that they have a similar pattern.

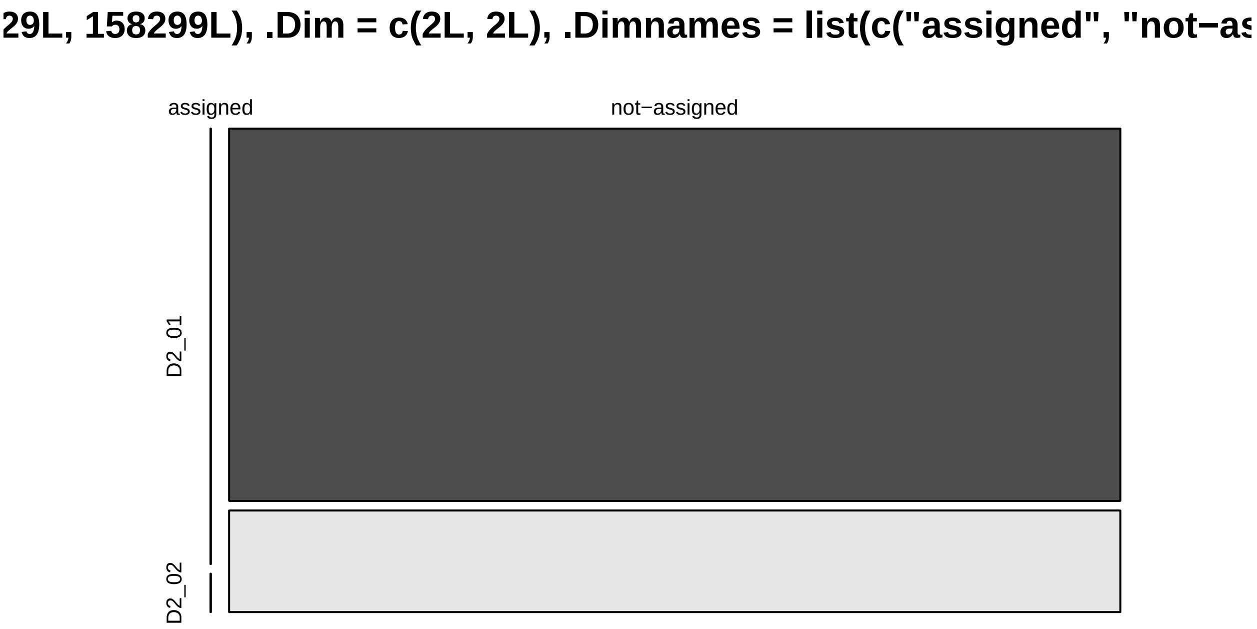
### Part-03: plot var(x-all samples) vs mean(x-all sample) on log10-log10 scale

```
var_all_sample = apply(myTable, 1, var)
mean_all_sample = apply(myTable, 1, mean)

plot(log10(mean_all_sample), log10(var_all_sample))
abline(coef = c(0,1), col = 'red')
```



### Part-04: 2-by-2 contingency table for NC101_00003 (fisher's test)

```
conTab = data.frame('D2_01' = c(col_D2_01[1], (sum(col_D2_01)-col_D2_01[1])),
                    'D2_02' = c(col_D2_02[1], (sum(col_D2_02)-col_D2_02[1])),
                    row.names = c('assigned', 'not-assigned'))

mosaicplot(conTab, color = TRUE)
```



```r
test = fisher.test(conTab)

pVal = test$p.value
print(paste('p-value: ', pVal))
```

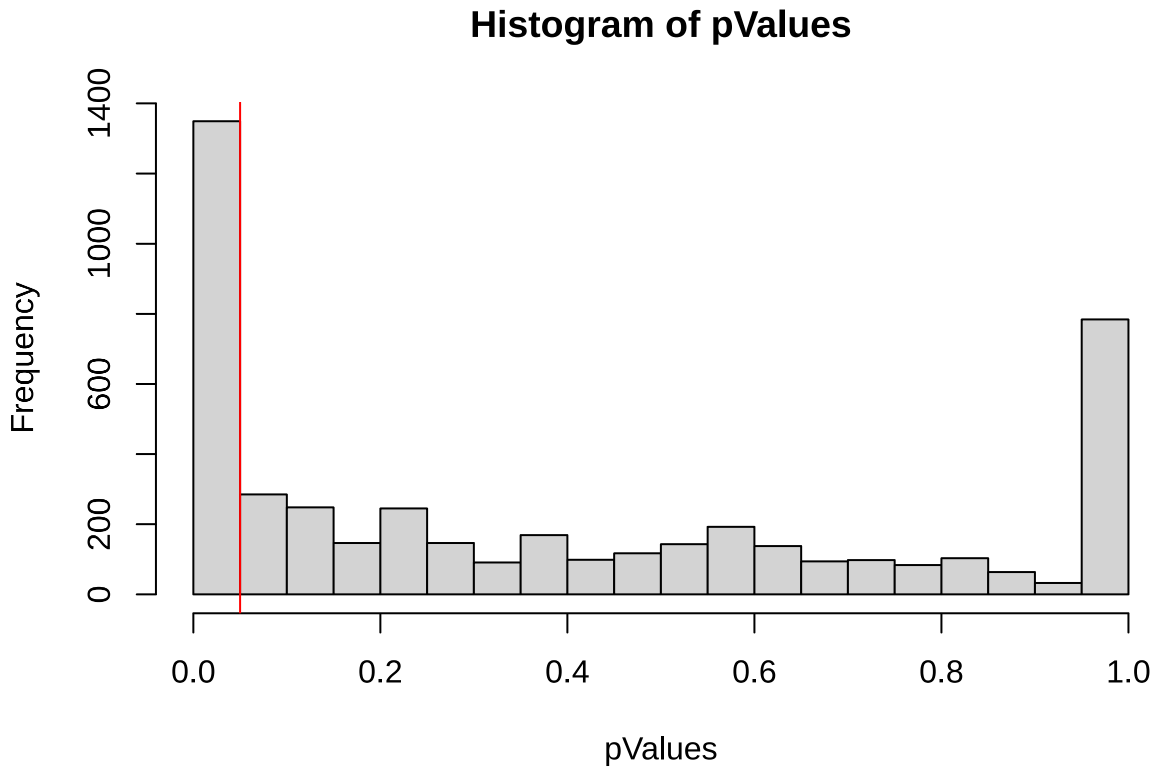
```
## [1] "p-value:  1.67001714123219e-11"
```

Since, the p-value obtained from fisher's test is significant [<0.05], we reject the null hypothesis and conclude that there is association between the column and row variables.

## Part-05: 2-by-2 contingency table for all the genes (fisher's test)

```r
pValues = vector()

for (i in 1:length(col_D2_01)) {
  conTab = data.frame('D2_01' = c(col_D2_01[i], (sum(col_D2_01)-col_D2_01[i])),
                      'D2_02' = c(col_D2_02[i], (sum(col_D2_02)-col_D2_02[i])),
                      row.names = c('assigned', 'not-assigned'))

  test = fisher.test(conTab)

  pVal = test$p.value
  pValues[i] = pVal
}

hist(pValues, breaks = 25)
```

```
abline(v=0.05, col='red')
```

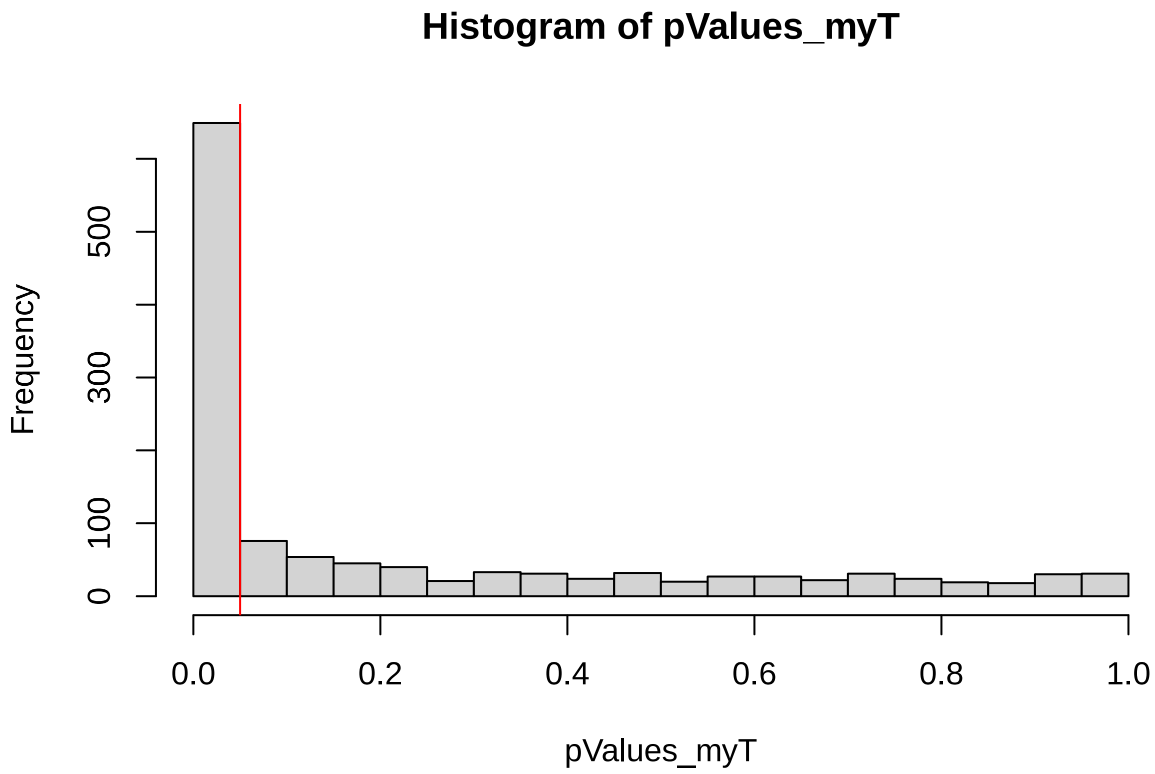


From the plot we can see that, the p-values are not entirely uniformly distributed. We can have p-values to be uniformly distributed when the null hypothesis is true. But we see that, the hypothesis is false for around 30% of the genes (left of red line) since we expect to have dependencies between the rows and columns for these.

Also, we expected to see more significant p-values since we are considering samples from replicas having similar gene patterns.

```
myT <- myTable[ (myTable$D2_01 + myTable$D2_02 > 100),]

col_D2_01_myT = myT[ , c("D2_01")]
col_D2_02_myT = myT[ , c("D2_02")]

pValues_myT = vector()

for (i in 1:length(col_D2_01_myT)) {
  conTab = data.frame('D2_01' = c(col_D2_01_myT[i], (sum(col_D2_01_myT)-col_D2_01_myT[i])),
                      'D2_02' = c(col_D2_02_myT[i], (sum(col_D2_02_myT)-col_D2_02_myT[i])),
                      row.names = c('assigned', 'not-assigned'))

  test = fisher.test(conTab)

  pVal = test$p.value
  pValues_myT[i] = pVal
}
```

```
hist(pValues_myT, breaks = 25)
abline(v=0.05, col='red')
```



When we remove the low abundance genes the p-values starts to become more significant, thus distribution becoming more weighted towards zero.

### part-06: poisson.test() for NC101_00003

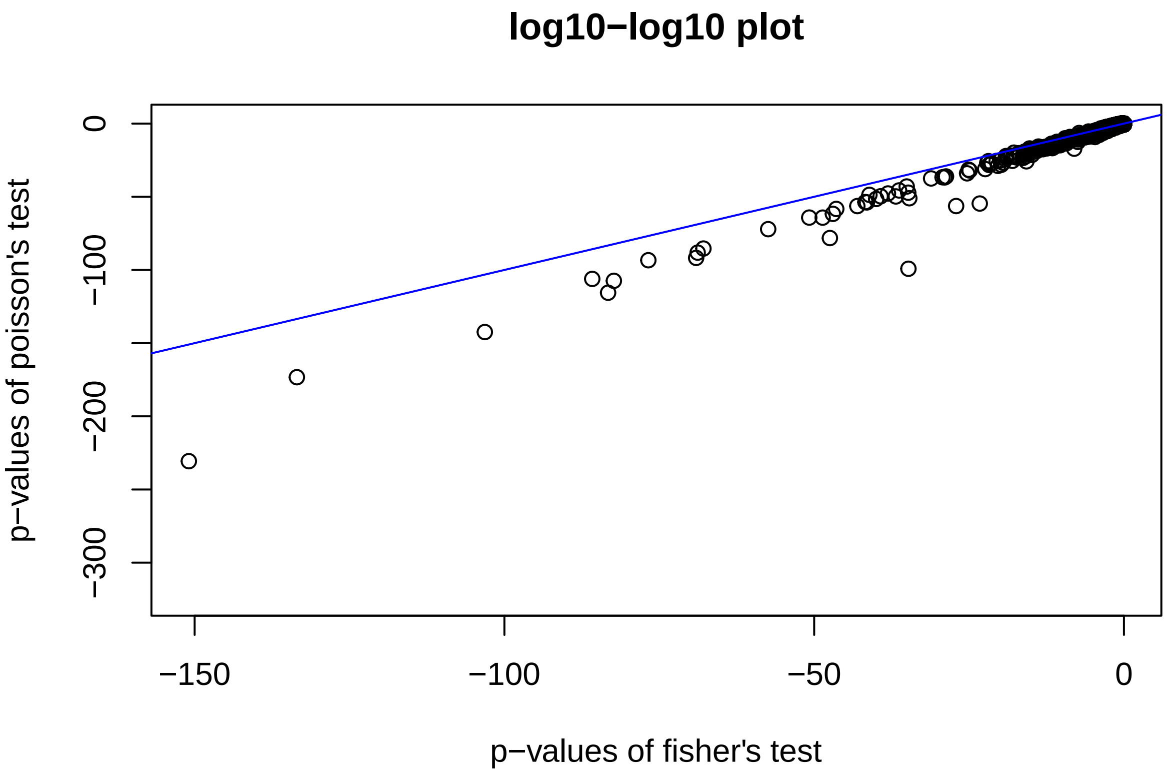
```
newT = myTable+1

col_D2_01_newT = newT[ , c("D2_01")]
col_D2_02_newT = newT[ , c("D2_02")]

p = col_D2_01_newT[1]/sum(col_D2_01_newT)

test = poisson.test(col_D2_02_newT[1], sum(col_D2_02_newT), r=p)

p_val = test$p.value
print(paste('p-value: ', p_val))
```

```
## [1] "p-value:  1.13934089393884e-13"
```

Since, the p-value obtained from poisson's test is significant [<0.05], we reject the null hypothesis and conclude that the expected frequencies are not similar.

### part-07: poisson.test() for all the genes

```
pValues_pos = vector()

for (i in 1:length(col_D2_01_newT)) {

  p = col_D2_01_newT[i]/sum(col_D2_01_newT)

  test = poisson.test(col_D2_02_newT[i], sum(col_D2_02_newT), r=p)

  p_val = test$p.value
  pValues_pos[i] = p_val
}

plot(log10(pValues), log10(pValues_pos), xlab = 'p-values of fisher\'s test',
     ylab = 'p-values of poisson\'s test', main = 'log10-log10 plot')
abline(coef = c(0,1), col = 'blue')
```



From the plot, we see that the p-values from fisher's test and poisson's test has almost a linear relationship. Therefore, we can assume they agrees.