|  |
| --- |
| Figure 1- GSE12221 |
| (A) | Stddev:  Min-0.01224  1st Qu- 0.03622  Median-0.05093  Mean-0.06122  3rd Qu-0.07366  Max-0.28834  CV  Min.- 0.003360  1st Qu.-0.009714  Median- 0.013776  Mean- 0.016811  3rd Qu.- 0.020289  Max. -0.079133  Mean  Min.- 3.271  1st Qu.-3.558  Median- 3.687  Mean - 3.690  3rd Qu.-3.824  Max. - 4.176 |
| (B) | ORF: "YKR009C" "YDL157C" "YER059W" "YGL059W" ...  CV : 0.00785 0.01638 0.01574 0.01826 0.00734 ... |
| (C) | Stddev: 0.0285 0.0639 0.0583 0.0668 0.0282 ... |
| (D) | Mean : 3.63 3.9 3.7 3.66 3.85 ... |
| Figure 2-GSE12221 |
| (A)  Min 1Q Median 3Q Max  -20.1202 -2.3201 0.1785 3.1752 13.0004  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 24.1357 0.4309 56.014 <2e-16 \*\*\*  RLS.tb2$myCV -1.1915 22.6039 -0.053 0.958  Residual standard error: 4.672 on 495 degrees of freedom  Multiple R-squared: 5.614e-06, Adjusted R-squared: -0.002015  F-statistic: 0.002779 on 1 and 495 DF, p-value: 0.958 |
| (B)      Min 1Q Median 3Q Max  -15.1148 -2.9337 0.3343 3.2409 15.9587  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 18.188 5.208 3.492 0.00069  YPD 7.118 5.318 1.339 0.18345  Residual standard error: 5.264 on 110 degrees of freedom  Multiple R-squared: 0.01603, Adjusted R-squared: 0.007084  F-statistic: 1.792 on 1 and 110 DF, p-value: 0.1835 |
| Figure 3-GSE12221 |
| (A) | p-value- 0.001 |
| (B) | diff.CV-  0.008569  1649582743 |
| (D) | diff.CV.obs-  0.00762428  517538344 |
| (E) | p-value- 0 |
| (F) | diff.CV.obs- 0.008187  13958574378 |

Discussion

The GEO data set that was assigned for this project was GSE12221. It is responsible for decaying the profiles of Saccharomyces Cerevisiae mRNAs following oxidative stress and DNA damage. For the first question on the exam the standard deviation, mean, and CV were observed for the gene expression. Along with those results several histograms were created. This data was used to organize analyzing each of the probes. Then for question two the correlation between the data that had been used in the past projects were related to one another. The following is the name of the files: fitness, interaction degree of protein network, interaction degree of positive genetic interaction degree, interaction degree of negative genetic interaction degree, and replicative life span. By running this data on R studio the GSE 12221 data set was found to be significant to this data. The significant of this data set was observed by using the p-value from the various graphs. In the next question permutation was used to calculate the difference in expression of the data sets. The p -values for GSE12221 were small for all the results by observing the negative and positive genetic networks and protein networks. With these observations it was concluded that something in GSE12221 is blocking the results. By looking at someone else’s results from the data set GSE7645 it was the exact opposite of GSE12221 and can hold the answer. The results for GSE12221 may be clearer if the data that represents DNA stress was removed, because the GSE7645 does not contain this set and it contained very significantly higher results. There is still some investigation that needs to be done when it comes to GSE12221 by making comparison of the data that has been given.