Python Script Output

Summary stastistics of Adjusted P\_value for A vs C padj

count 5.975000e+03

mean 4.543972e-01

std 3.340562e-01

min 1.720720e-19

25% 1.274957e-01

50% 4.361504e-01

75% 7.677395e-01

max 1.000000e+00

Table 1: Summary statistics of A vs C data of the adjusted p values

Summary stastistics of Log fold change for A vs C log2FoldChange

count 5975.000000

mean -0.023737

std 0.998987

min -23.634090

25% -0.223346

50% -0.000873

75% 0.221863

max 20.276410

Table 2: Summary statistics of A vs C data for the Log2FoldChange

Summary stastistics of Adjusted P\_value for A vs E padj

count 6.094000e+03

mean 2.594718e-01

std 3.200505e-01

min 2.343659e-215

25% 1.144667e-03

50% 8.282949e-02

75% 4.889770e-01

max 1.000000e+00

Table 3: Summary statistics of A vs E data of the adjusted p values

Summary stastistics of Log fold change for A vs E log2FoldChange

count 6094.000000

mean -0.131371

std 1.443180

min -24.882900

25% -0.628644

50% -0.032872

75% 0.510509

max 22.732160

Table 4: Summary statistics of A vs E data Log2FoldChange

In regards to the mean adjusted p-values it is lower in the A vs E dataset when compared to A vs C however in both cases the result would be considered insignificant adjusted p-value = 0.05. For the Log2Foldchange the mean for both is a negative value meaning that both the datasets tend towards genes being down rather than upregulated, though for the A vs C dataset the value is very close to zero so it could be argued more genes tend towards no change in expression in that dataset when compared to A vs E.

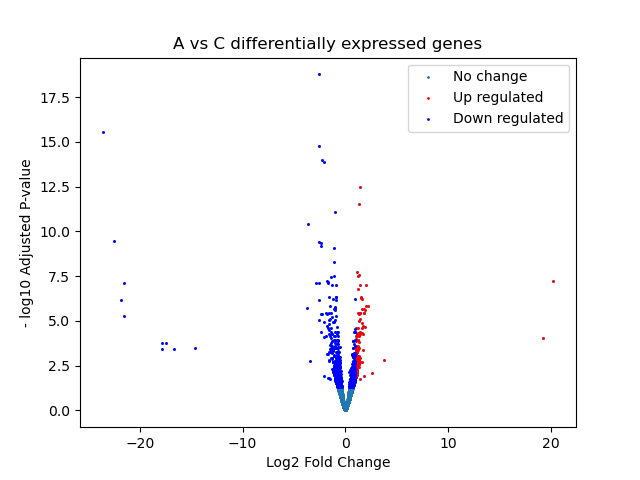


Figure 1: The relationships between log fold change and -log10 adjusted p-value for each of the genes in the A vs C. Up regulated genes are shown in light blue, up regulated genes shown in red and downregulated genes shown in dark blue.

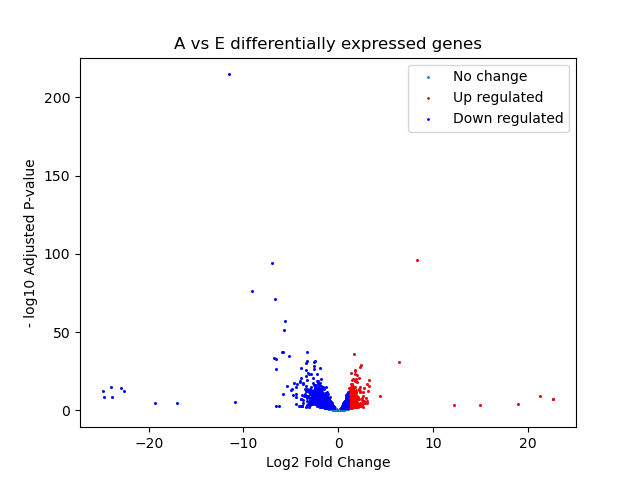


Figure 2: The relationships between log fold change and -log10 adjusted p-value for each of the genes in the A vs E. Up regulated genes are shown in light blue, up regulated genes shown in red and downregulated genes shown in dark blue.

Overall, these graphs are in support of the summary statistics mentioned earlier. For both graphs genes tend towards being downregulated with the A vs E data having more upregulated genes and the A vs C dataset having more genes that have undergone no change. Alongside that A vs E has a much wider spread of log2fold change when compared to the A vs C dataset while the A vs C has a must wider variation in p-values. Admittedly part of the reason for this could be that during the data cleaning the end result was that the two datasets where of uneven sizes so that may contribute in part for the differences between them.

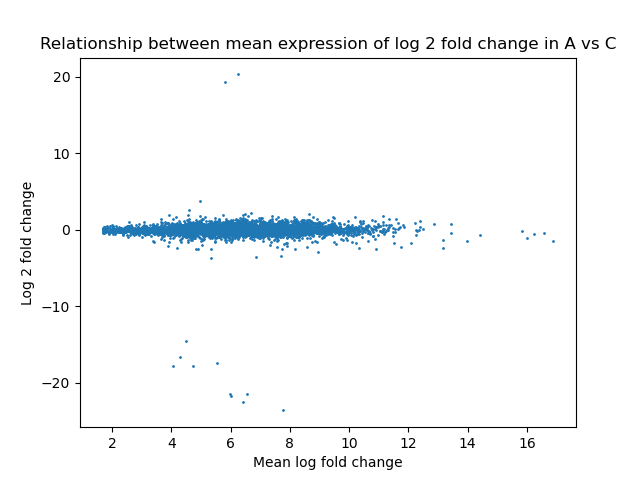
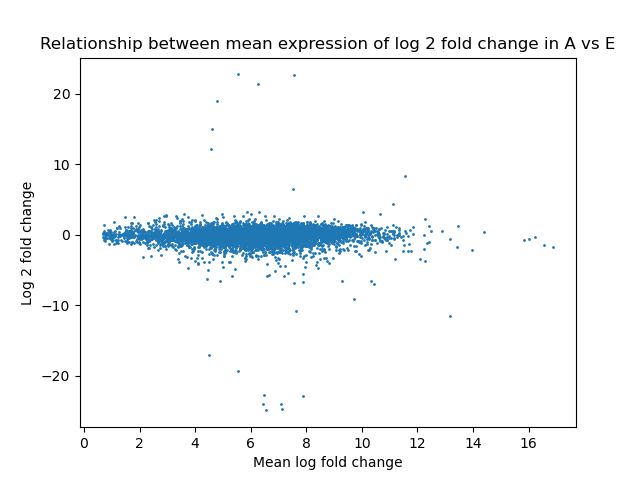


Figure 4: Relationship between mean expression represented as mean log fold change and log2foldchange in the A vs E dataset.

Figure 3: Relationship between mean expression represented as mean log fold change and log2foldchange in the A vs C dataset.



Against similar to the previously mentioned graphs the MA plots support those results. For example, when comparing the MA A vs C plot and the MA A vs E plot, a majority o the points in A vs C place near the 0 log2foldchange value, meaning that 1 more value compared to the A vs E data set experience no change in gene expression and 2, that there is a smaller spread in the magnitude in which genes are up and downregulated when in occurs. This is especially prevalent when comparing the A vs F dataset that shows a wider spread log2foldchange values.

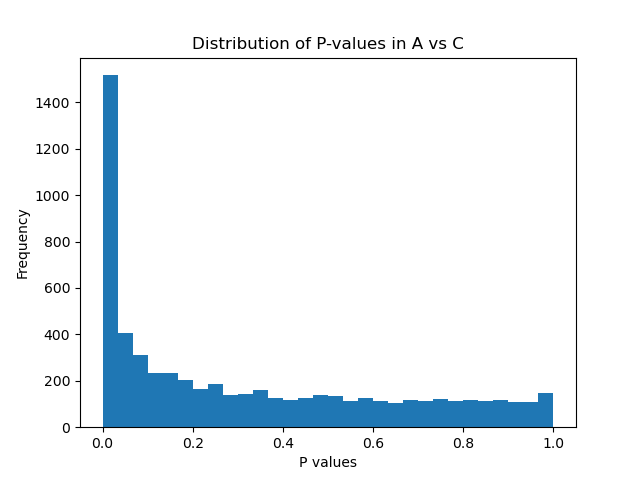


Figure 5: Distribution of p-values in the A vs C dataset when bins is = 30

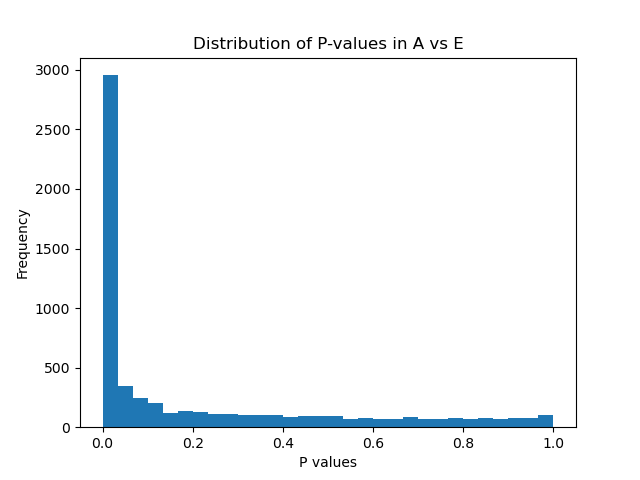


Figure 6: Distribution of p-values in A vs E dataset when bins = 30.

Overall, the two graphs confirm previously mentioned interpretations of the results with the A vs C dataset showing wider spread in p-values when compared to A vs E meaning that less of the results in the A vs C dataset would be considered significant. It should be mentioned however those observations were based on adjusted p-values rather than raw p-values meaning that there could be slight changes in the distribution when using this correction for both of the datasets, likely causing more of the genes in the A vs C dataset to be considered not significantly differentially expressed.

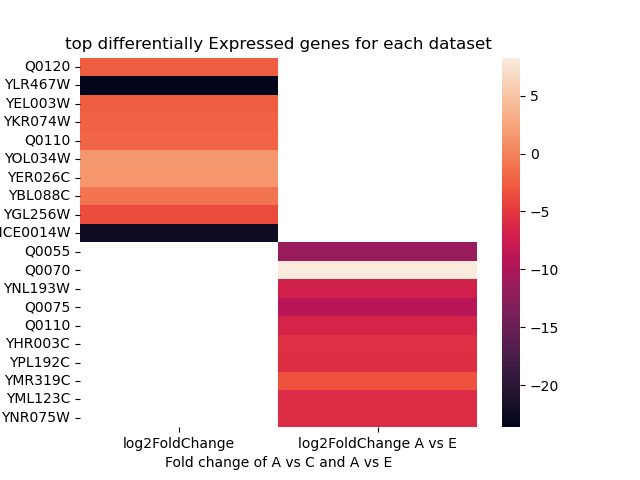


Figure 7: Top 10 differentially expressed genes for the A vs C dataset represented by the “log2Foldchange” vs the 10 10 differentially expressed genes in the A vs E dataset represented by the log2FoldChange “A vs E”. The darker colours such as black represent downregulation while lighter colours represent upregulation of genes. The order the genes are put in goes from most significant at the top then decreasing in significance as you go down.

When comparing the two datasets A vs C has two heavily downregulated genes and several highly upregulated genes while for the A vs E dataset a majority of the genes are downregulated being around -5 to -10. The only exception to that is Q0070 which is the most upregulated gene between the two datasets. Alongside this both conditions share a gene with it being the 5th most significant gene in both the datasets.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| incoming | converted | n\_incoming | n\_converted | name | description | namespaces |
| Q0120 | Q0120 | 1 | 1 | BI4 | Mitochondrial mRNA maturase | ENSG,ENSP |

Table 5: Results of G: convert for the top most differentially expressed gene in the A vs C dataset with the species being used being *saccharomyces cerevisiae.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| incoming | converted | n\_incoming | n\_converted | name | description | namespaces |
| Q0055 | Q0055 | 1 | 1 | AI2 | Reverse Transcriptase | ENSG,ENSP |

Table 6: Results of G: convert for the top most differentially expressed gene in the A vs E dataset with the species being used being *saccharomyces cerevisiae.*