NEXT GENERATION SEQUENCING

Objective:-To perform transcriptomic analysis of the impact of climate change on mycotoxin production in *Fusarium langsethiae,* under 3 conditions; 1) Normal (20 °C & 0.995 aw) 2) Condition 1 (25 °C & 0.995 aw) and 3) Condition 2 (25°C & 0.98 aw) .

TASK 1

The quality control (QC) for the raw reads was done by illumine encoding 1.9 and it was found out that the: -

(a) quality scores of all the reads fell under the category of good quality with all the scores >= 25, also the

(b) sequence length being 150 for all the sequences. Then instead of per base sequence quality we can see the

(c) per sequence quality scores which will give us the mean scores of all the bases in that sequence and the average quality per read is quite good too.

(d) However, the per base sequence content is varying drastically and this module failed because the difference between A & T or G & C is greater than 20% in any position and the reason for this could be libraries produced by priming using random hexamers, and it is shown only during the starting of the reads afterwards the lines are parallel. Also correcting it is not possible by trimming and it won't affect the downstream analysis.

(e) The GC content of the sequences in overlay with the theoretical distribution.

(f) The per base N content gives us any bases that are uncalled, and we don't have any uncalled bases.

(g) the sequence duplication is high which indicates that there is a possibility of enrichment bias with the PCR over amplification, and in our case the uniquely mapped sequences range from 30% to 40% which is quite low, and the duplication levels are high >50% for which the model failed.

(h) The overrepresented sequences show the sequences that are >0.1% of the total, and one observation that we can deduce is that the samples with 20°C treatment has the overrepresented sequence, which could be of biological significance or it could also be some contamination, but this module cant differentiate between those two.

- By seeing the quality of data which is very good ,the need for trimming /filtering the raw data was not needed and thus was not performed.

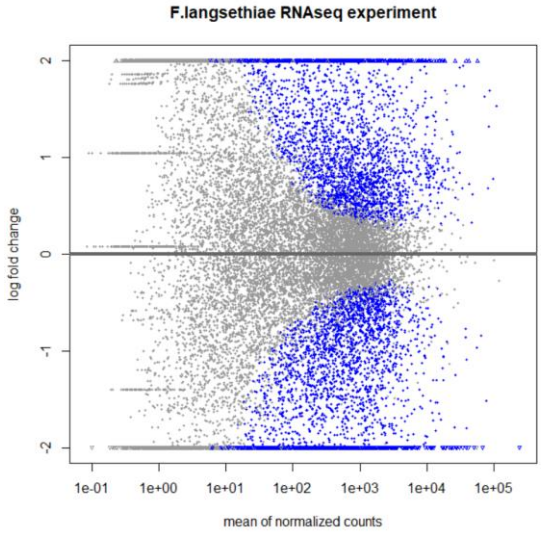
TASK 2

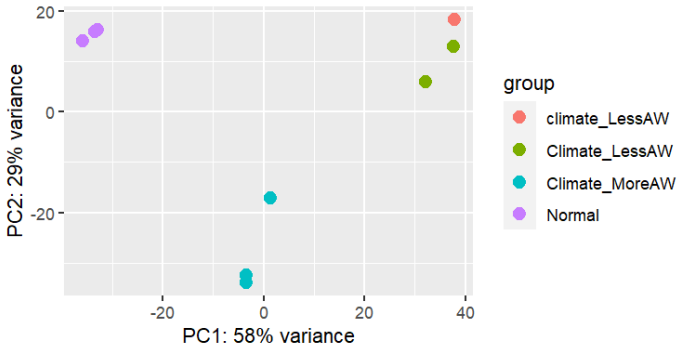
The alignment for our reads was done with two softwares; STAR and BWA. And from percentage of uniquely mapped reads ,the conclusion was made to go with the results of STAR alignment which had % varying from 65-80.While for the BWA software the percentage of unique mapping was too high (>90%),which is much higher than an average threshold. Thus that option was eliminated.

And then after that HTSeq files with the counting reads aligned in each BAM files that overlap with genes were generated.

TASK 3

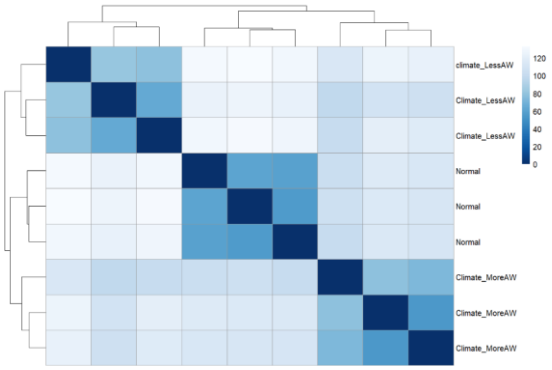
a) From the HTSeq.txt file an overall differential gene expression for three condition was performed;Normal,high temperature (low water activity) and high temperature (high water activity). And the number of genes differentially expressed with p value(adjusted) is 4724. Then we produce a MA plot from which we can see that on y axis is the condition while on x axis we have how strong are the genes expressed in those conditions ,while the blue dots are genes that are significantly different between conditions.

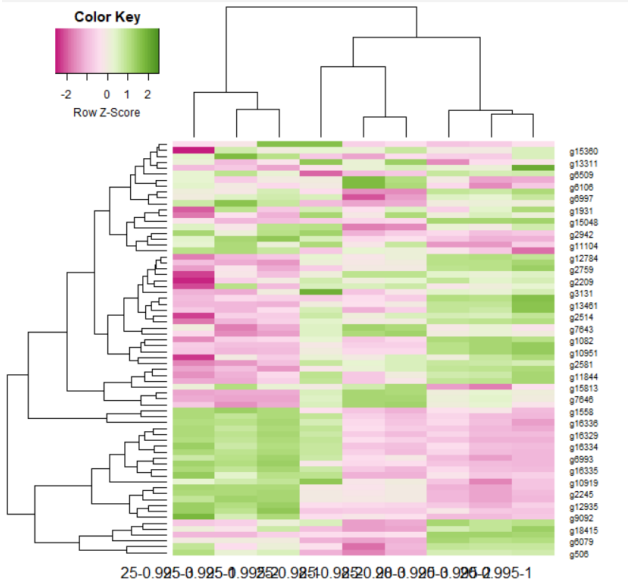




b) From the PCA plot we can reduce the data and understand that the PC1 had approx. 58 % variance while the second most PC2 had 29% variance and the samples with more water activity, tends to have more importance than less water activity as it is closer to PC1 axis. So

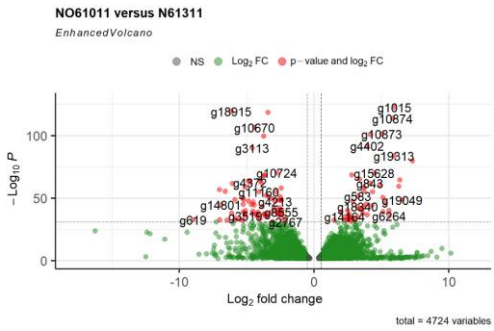
we can say that the data for all the different conditions are clustered together and to find the similarity between the clusters we have to perform HCA.

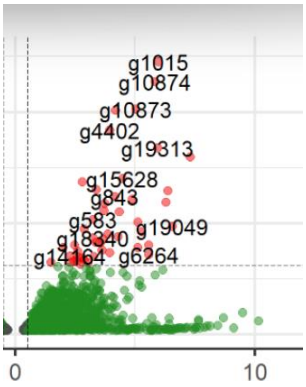




From the HCA plot we can see that the ones in grey are under expressed genes while the ones in the blue are the overexpressed genes. Also in the volcano plot we can see the downregulated genes at the left sides while the upregulated genes are at the right side. Also the the most statistically significant genes are at labelled and are in red colour.(I tried doing the venn diagram but wasn't able to place the contrasting conditions,so transferred it at the end of script with #)

c) From the volcano plot we can see that the genes that are statistically significant and also that are upregulated genes are at the right side of 0. And these are the genes that are of Interest to us.





d) Now our approach should be to see these transcripts and check from the columns of the given annotation files ending with gff for the gene that is responsible for it. We can also see the downregulated genes that are responsible for the mycotoxin production(if),However these are the transcripts that we will be looking at first:-

g1015,g10874,g10873,g4402,g19313,g15628,g843,g583,g18340,g19049,g14164,g6264.

g10874:-These are the transcripts that produces an unknown protein product

g10873:-it is a hypothetical protein which is also found in Fusarium pseudograminearum CS3096.

g4402:- it is a cyclin protein which is responsible for controlling the progression of a cell through the cell cycle by activating cyclin-dependent kinase enzymes.

g19313:- it is responsible for P-loop containing nucleoside triphosphate hydrolase making protein

g15628:-it is also an unnamed protein

g843:- hypothetical protein HYE68_004307 [Fusarium pseudograminearum

g583:-it is a transcript for transmembrane protein

Now we open the research paper given to us in the assignment and search for the trichothecene genes which we can find in table 6 (Gene position, for the TRI gene cluster) and then we can read the description of protein activity and find trichoethecene as the protein activity and then see the corresponding gene so we come up with:-

trichothecene-4-O-acetyltransferase: *TRI7*
trichothecene c-15 hydroxylase: *TRI11*
trichothecene c-3 deacetylase: *TRI8*
trichothecene efflux pump: *TRI12*

trichothecene C-8 acyl transferase: *Tri16*

So now we will try to find these genes in the annotation files of flang_functional_annotation.gff
Now for gene TRI7 the transcripts that we were able to found out are; g6457, g3240

For the TRI11 gene the transcripts are; g3249,

For the TRI8 gene the transcripts are; g3239,

For the TRI12 gene the transcripts are; g3250

For the Tri16 gene the transcripts are;g10260