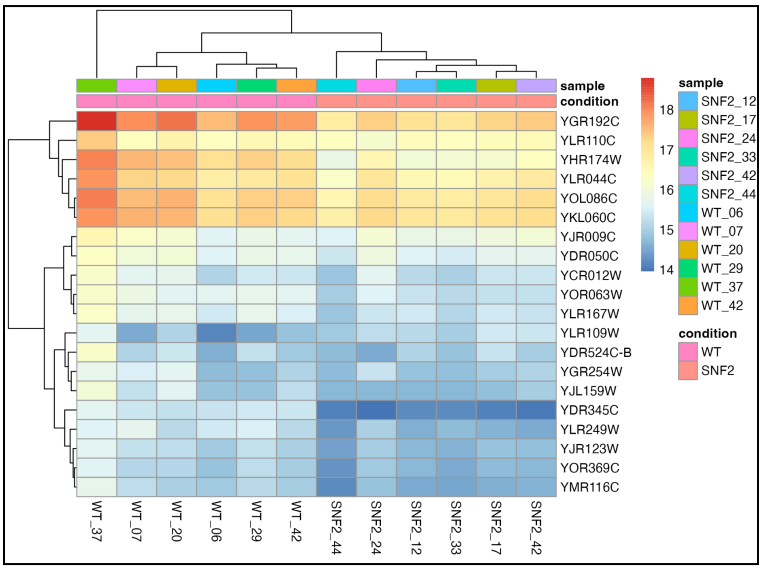
Sammy Mustafa

Differential expression

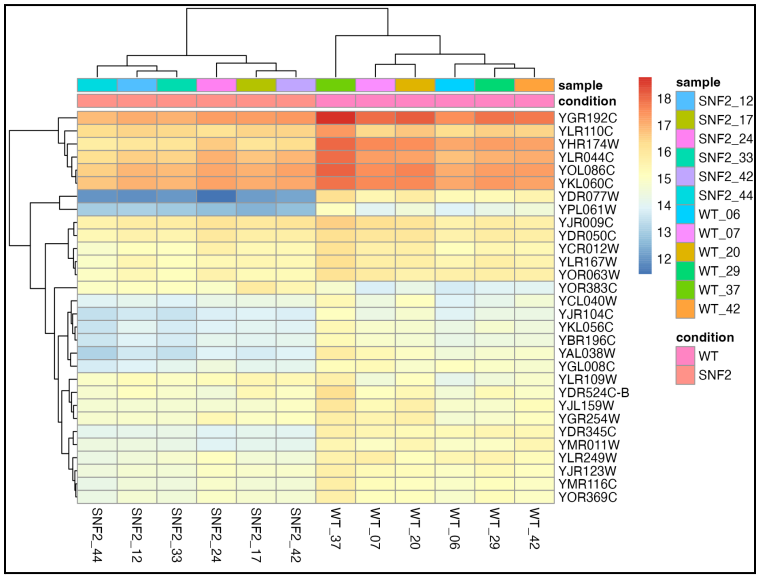
	6 Good Replicates of Each Type	1 Bad and 5 Good Replicates of Each Type	3 Good Replicates of Each Type	1 Bad and 2 Good Replicates of Each Type
# of Differentially Expressed Genes	495	270	301	102

It seems like the differential expression results are sensitive to both the number and quality of replicates used in highlighting the number of differentially expressed genes. While less and worse quality replicates both cause a lower number of differentially expressed genes being highlighted by our results, the quality of the replicates has a more negative effect on the results. Both in combination have the most negative effects on the differential expression analysis.

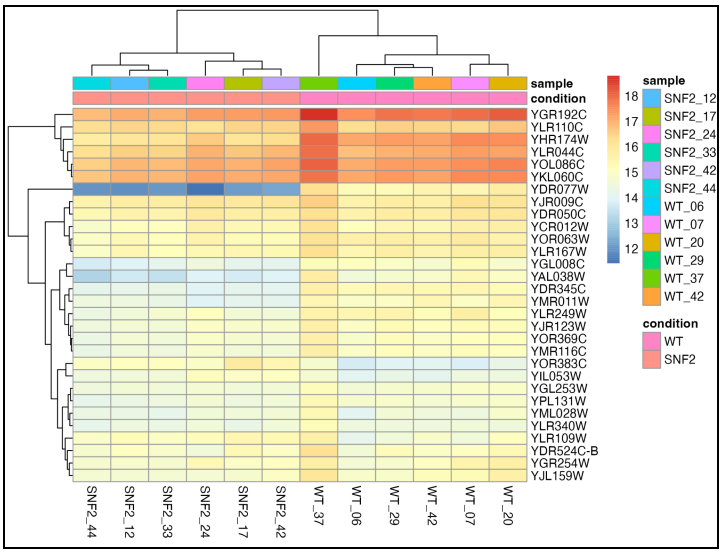
Visualizing DESeq2 results



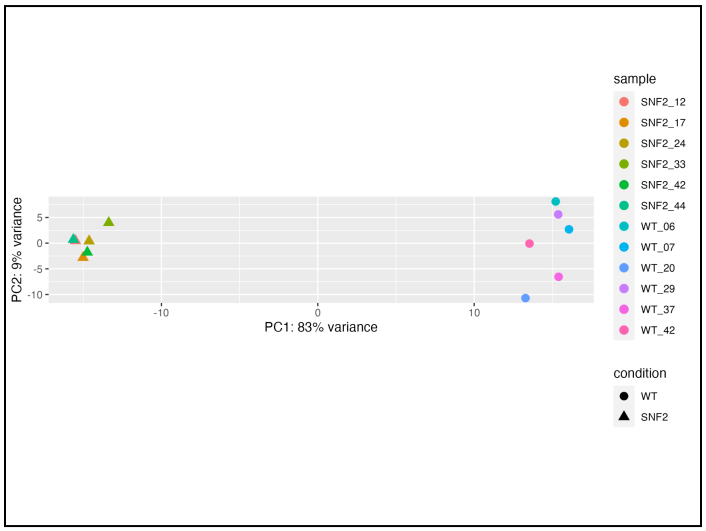
Heatmap of top 20 genes with the greatest mean expression



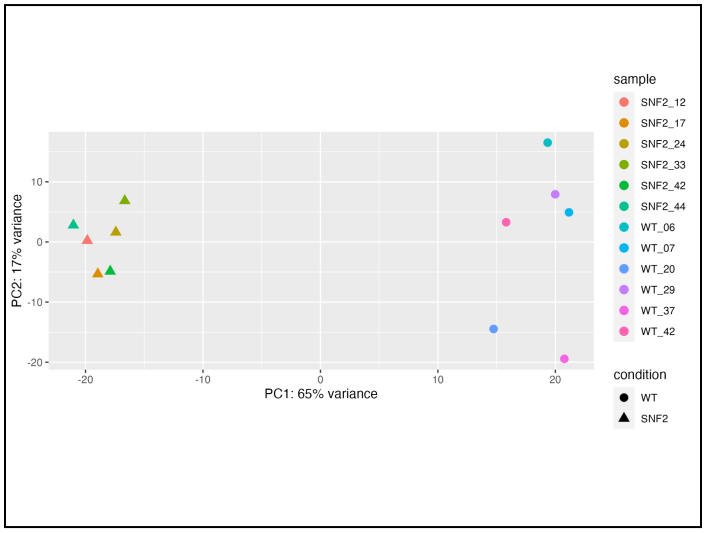
Heatmap of top 30 genes with the greatest mean expression



Heatmap of the top 30 genes with the greatest variance



PCA plot of 500 genes with the greatest variance



PCA plot of 5000 genes with the greatest variance

For the heatmaps, the addition of more genes with highest mean expression caused a change in the hierarchical clustering both by rows and columns, but this can also be attributed to the change in the order of the sample conditions presented in the heatmaps (it would be interesting to see if the first 20 samples studied had relatively the same grouping with each other). Moreover, looking at the genes with the greatest variance in comparison to the greatest mean expression caused a shift in the grouping of samples and replicates, but it seems like this is to a lesser extent than the addition of more genes to the analysis (however this may simply be dependent on the

samples utilized here). In addition, looking at more genes in the heatmap caused a shift in the scale of the heatmap colors (from 14-18 to 12-18) to account for the lower levels of up- and down-regulation from more gene samples; this adds to the visual difference between these analyses as there are less occurrences of blue with more genes studied.

In terms of the PCA plot, adding more of the genes with the greatest variance caused a change in scale to account for the greater values for PC1 and PC2. PC1 had less variance while PC2 had more variance from this addition. The actual position of the points on the plot in respect to the other points remained relatively similar if ignoring the change of the axes scales. However, if we consider the change in scale, it suggests that the addition of the next 4500 genes with the greatest variance did change the calculated similarity of the samples with each other due to this increase in variance.