Sequencing Report

Kaile Yuan

May 17, 2019

1 Multiqc alignment report

The samples that I have analyzed include WT_05, WT_13, WT_30, WT_33, WT_35, WT_44, and SNF2_02, SNF2_08, SNF2_19, SNF2_20, SNF2_28, SNF2_47. (Figure 1 and Figure 2)
Using STAR to align the sequences and the results are shown below:

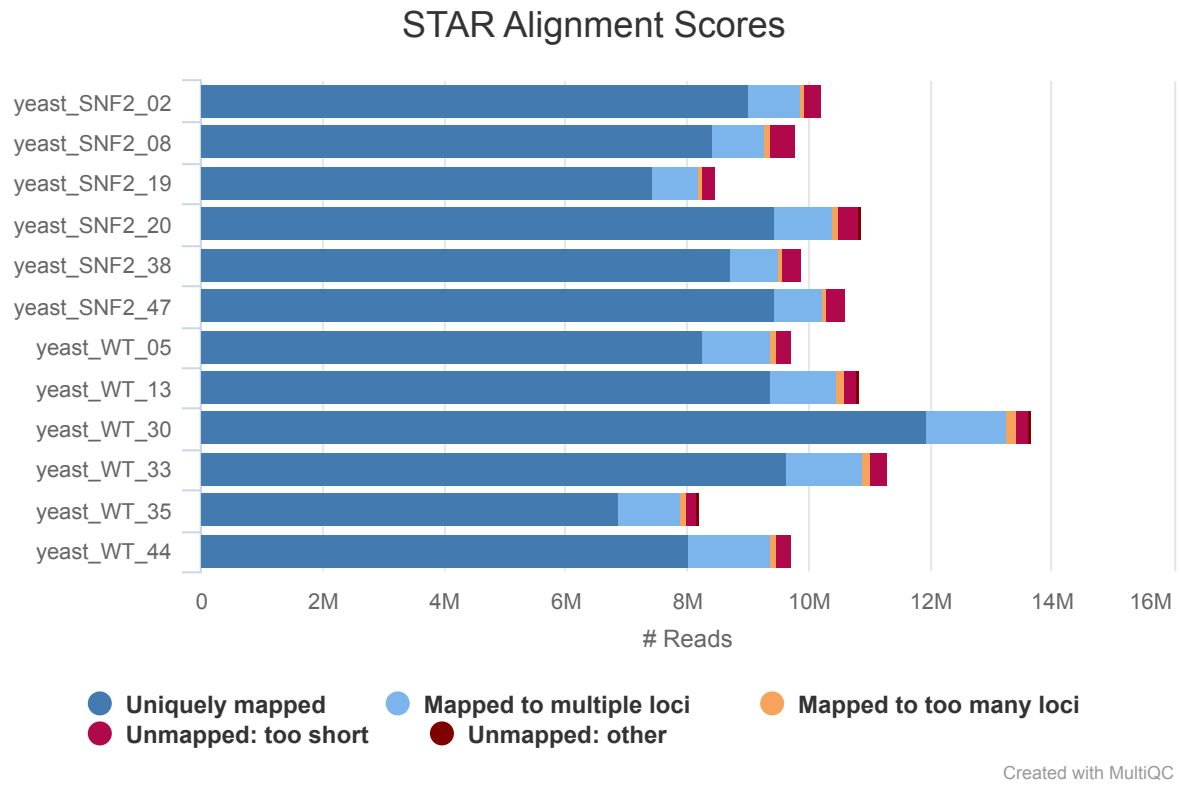

Figure 1: `multiqc` alignment report, alignment score

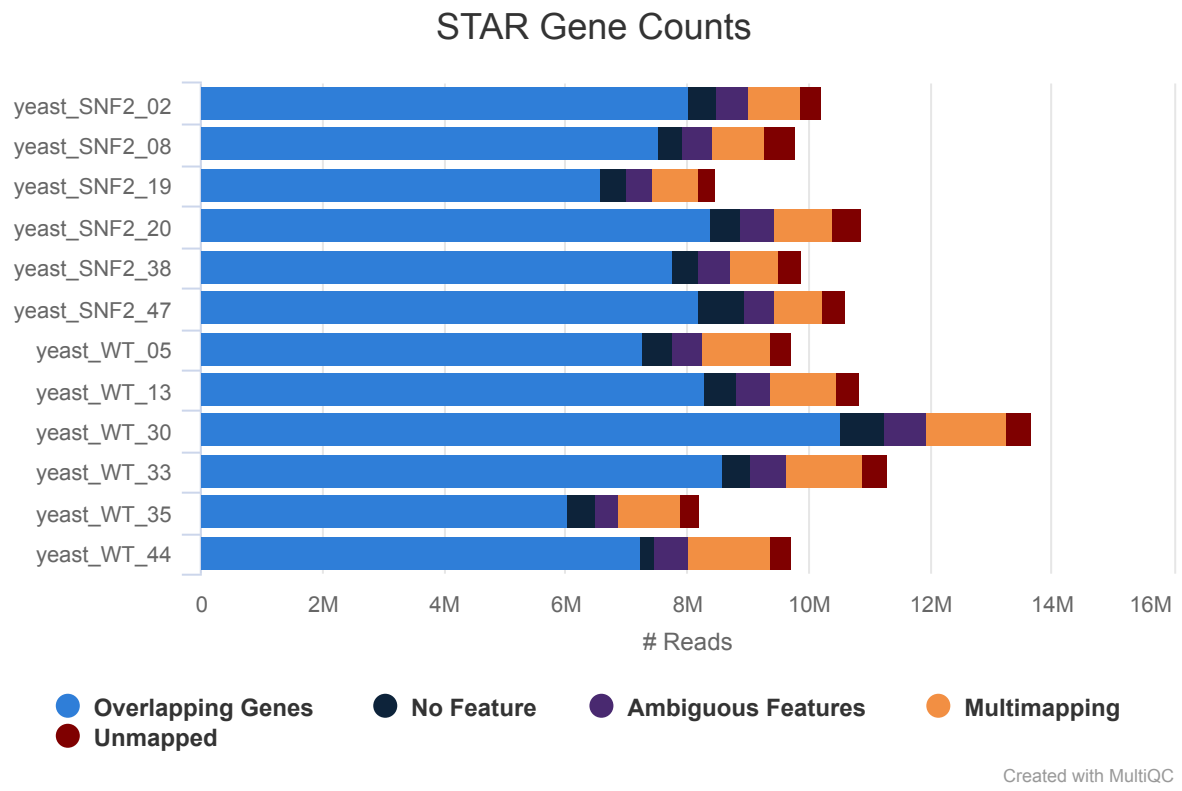

Figure 2: `multiqc` alignment report, gene counts

2 Differential expression analysis

DESeq2 package gives the differential expressed genes based on log2 fold change, and here we can see in the left panel that each gene, as represented by one circular dot in the plot, clusters around one central point. Noticeably all the dots marked red are considered differentially expressed. A shrunken log2 fold change result is provided at the right panel. (Figure 3)

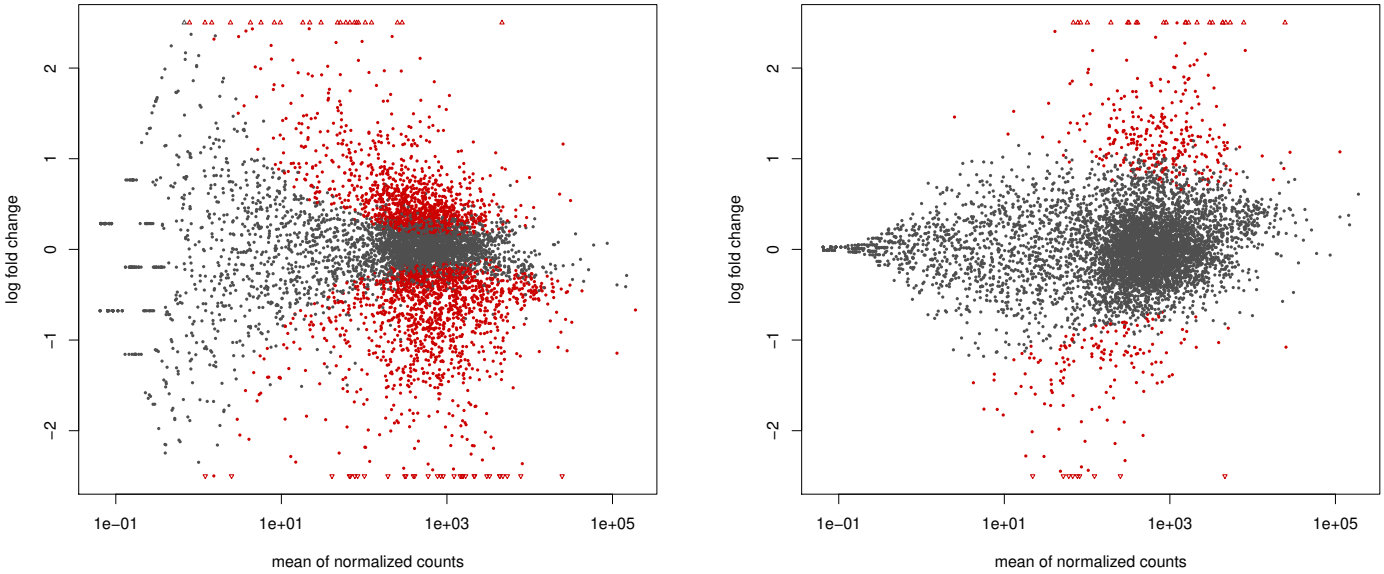

Figure 3: differential expression analysis: normal and shrunken MA plot

Figure 4 has shown the 20 most differentially expressed genes (top to bottom decreasing) within the two conditions. To remove the dependence of variance on the mean, particularly when mean is small, 'regularized log' transform has been employed, as shown in the right panel.

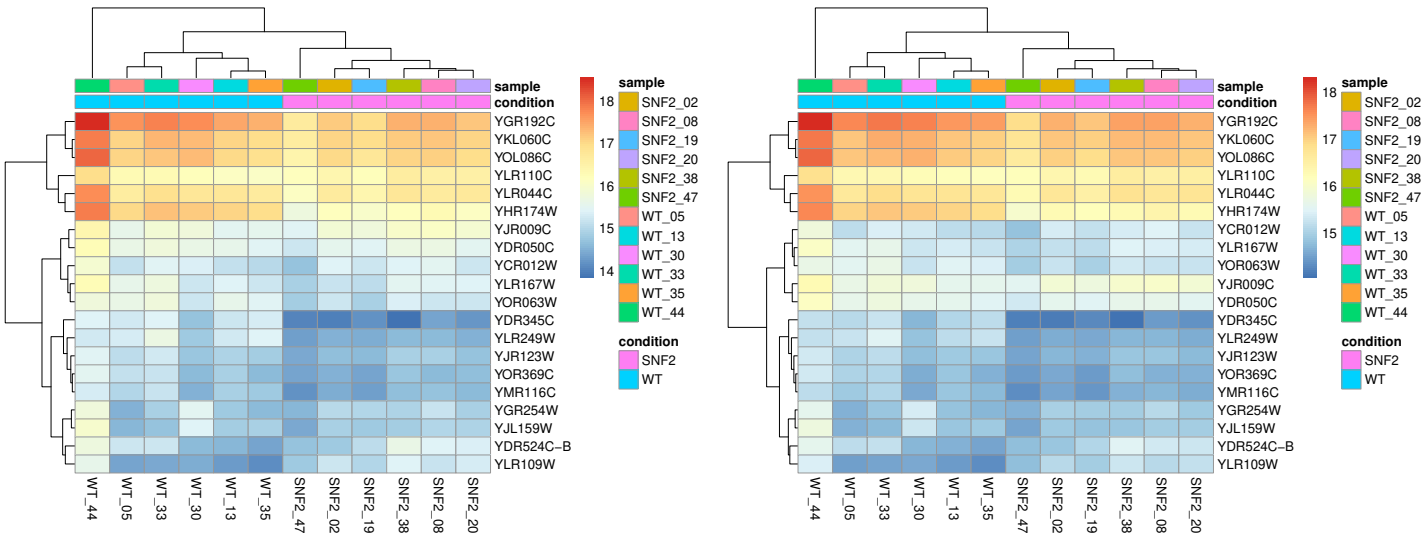

Figure 4: Heat map for 20 most differentially expressed gene

Finally, principal component analysis on the high dimensional data has been used to visualize the distance between different samples in 2d plot, as shown in Figure 5. We see clearly that wild type samples cluster densely in the upper right corner whereas SNF2 mutant samples cluster loosely in the left half of the plot. This confirms that the two groups are different in terms of those most differentially expressed genes.

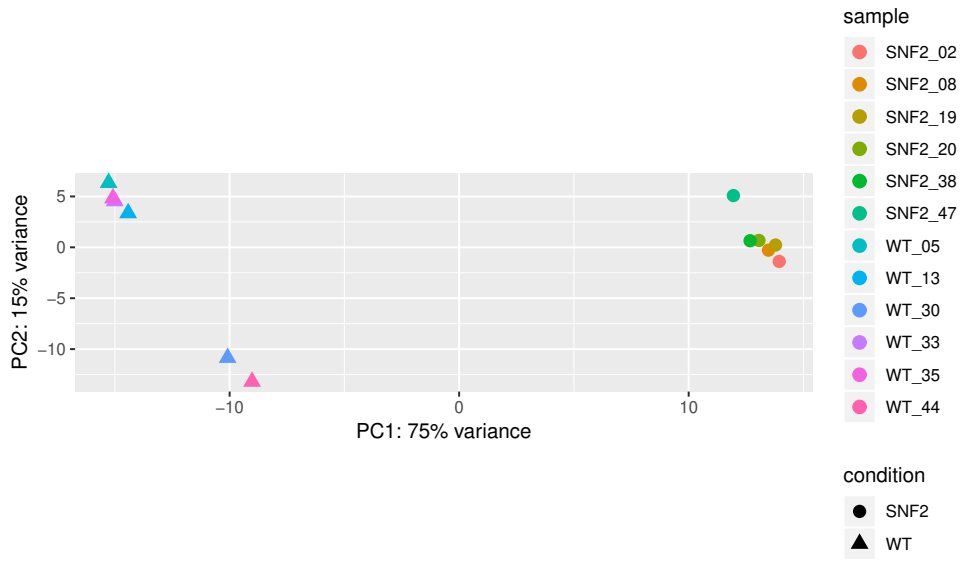

Figure 5: PCA of 10 samples across 2 conditions

3 If to substitute one good SNF2 sample with the bad one, SNF2_06, what difference will there be?

One difference is in Figure 7, the column SNF2_06 exhibits clear anomaly. The presence of bad sample also influence others, that all the differential expression level gets estimated much higher compared to Figure 4. Another difference is presented in Figure 8, where we see that the bad sample doesn't cluster with any other samples in the 2d space.

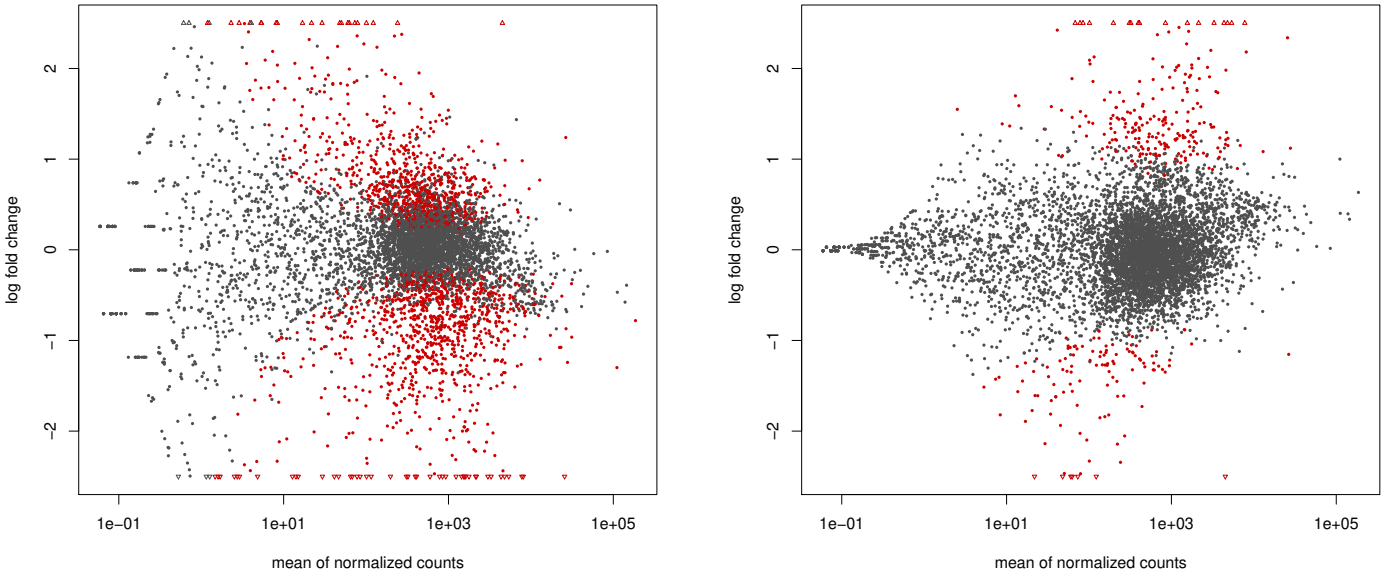

Figure 6: differential expression analysis: normal and shrunken MA plot

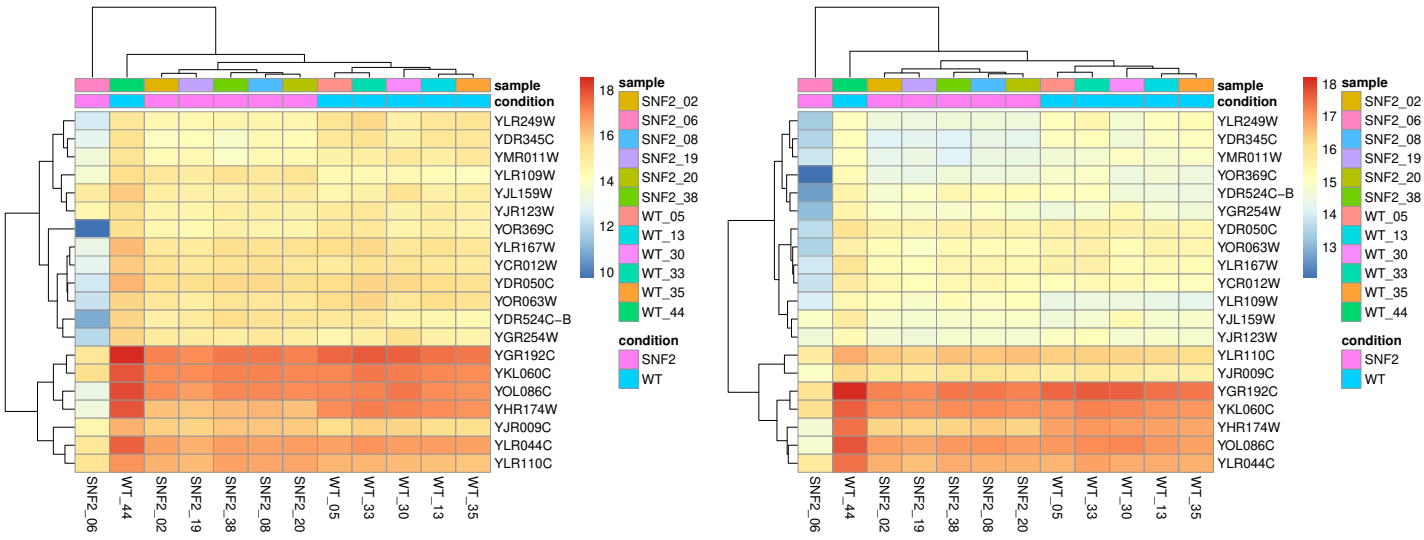

Figure 7: Heat map for 20 most differentially expressed gene

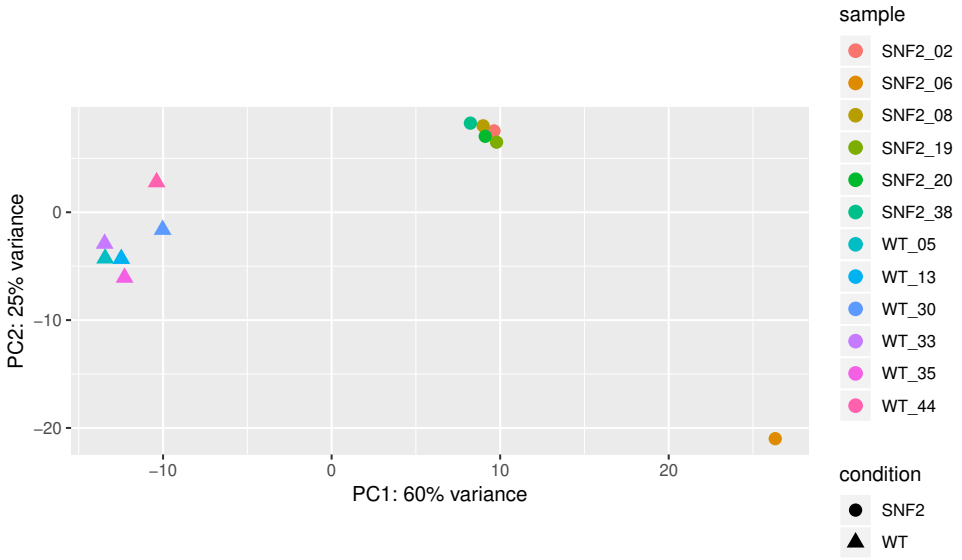

Figure 8: PCA of 10 samples across 2 conditions

4 Use three samples this time

4.1 Use three good samples

As shown in Figure 10, we see that WT_44 sample is one possible outlier in differential expression heat map. Shown in Figure 11, we see that WT_35 is one outlier in terms of expression. Therefore, experimentalist might focus more on the these two sample to look into the anomaly.

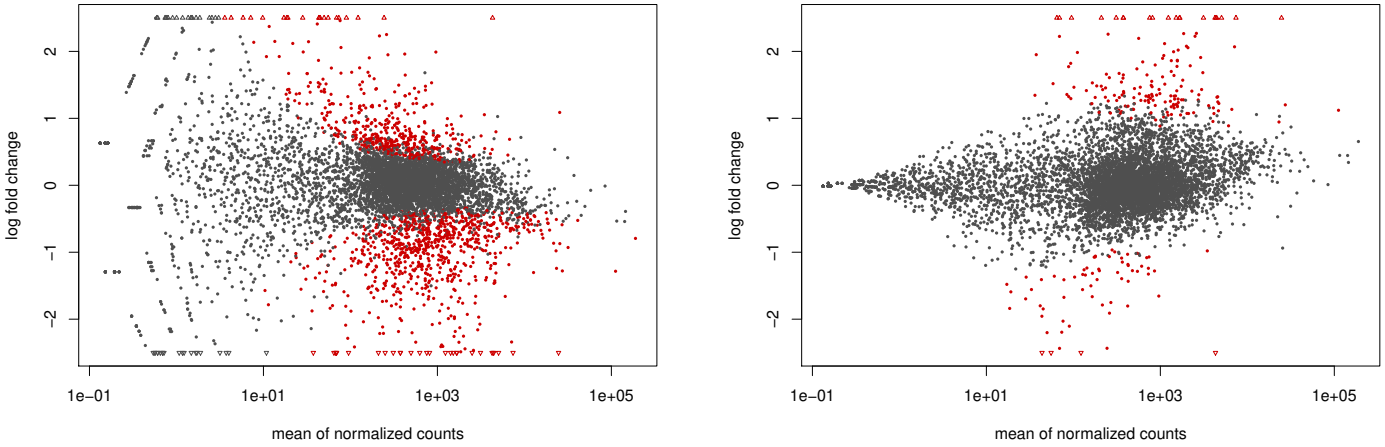

Figure 9: differential expression analysis: normal and shrunken MA plot

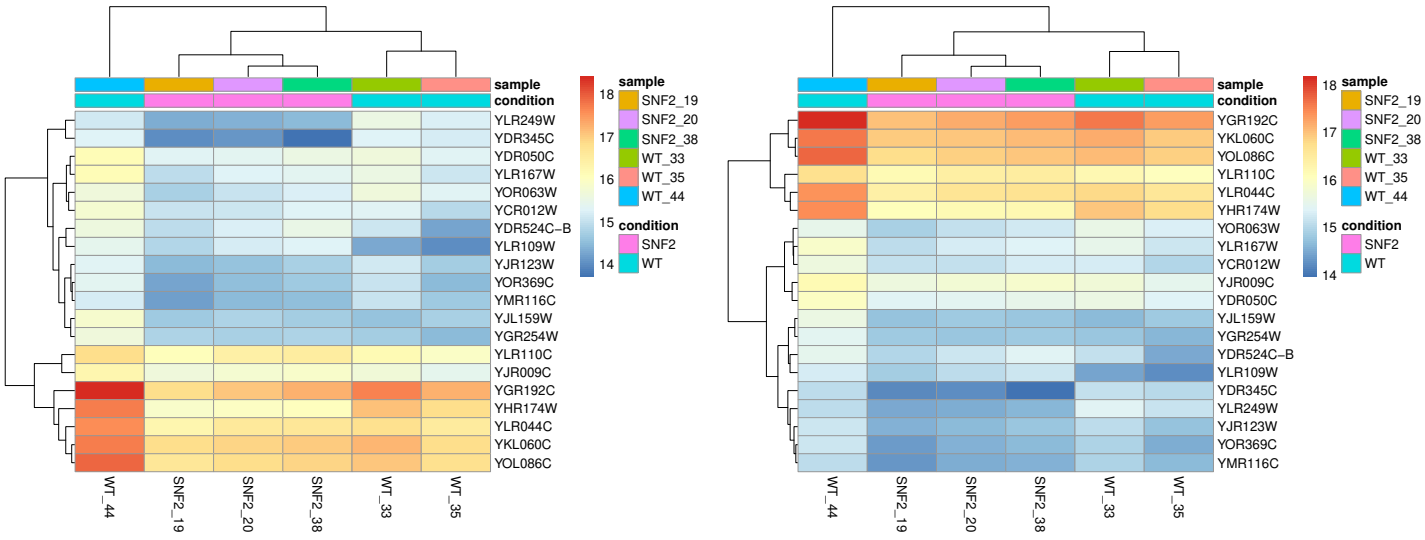

Figure 10: Heat map for 20 most differentially expressed gene

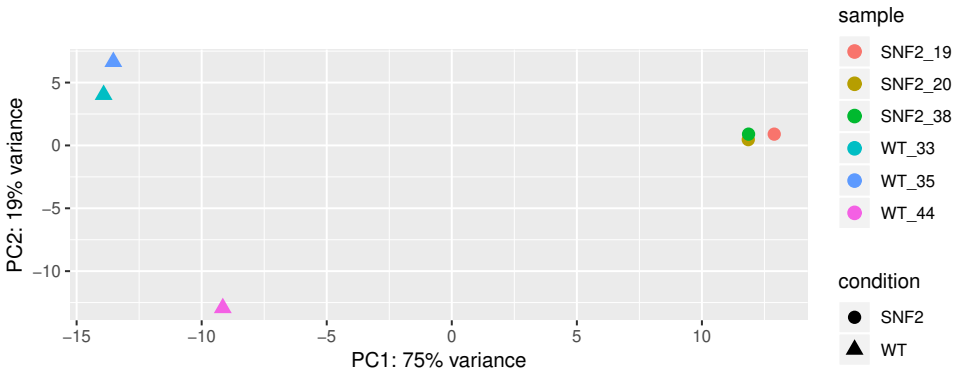

Figure 11: PCA of 10 samples across 2 conditions

4.2 Confound one bad sample

As shown in Figure 13, SNF2_06 sample clearly exhibited anomaly in the plot, and this is confirmed in Figure 14, where it gets separated.

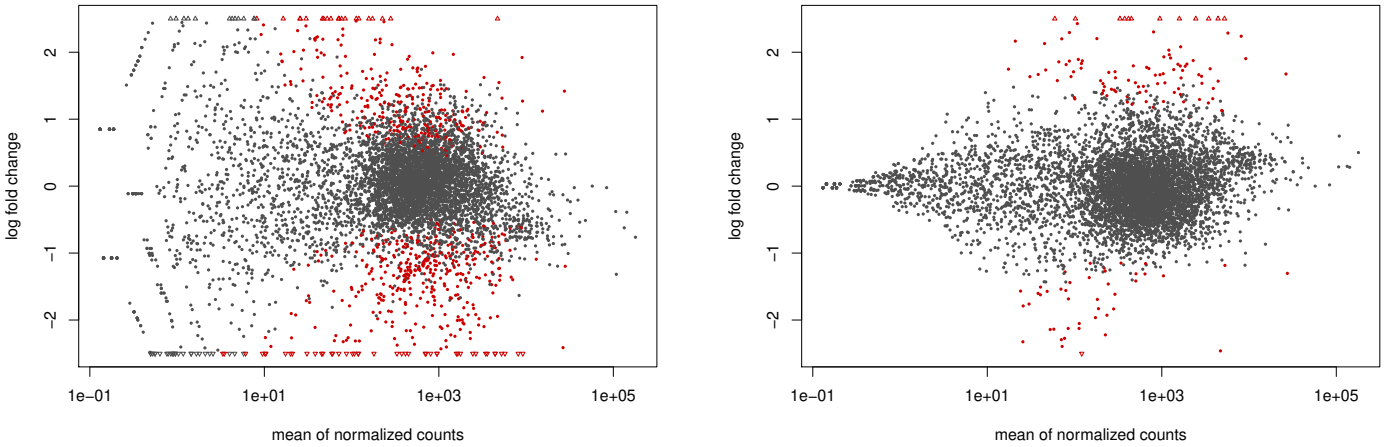

Figure 12: differential expression analysis: normal and shrunken MA plot

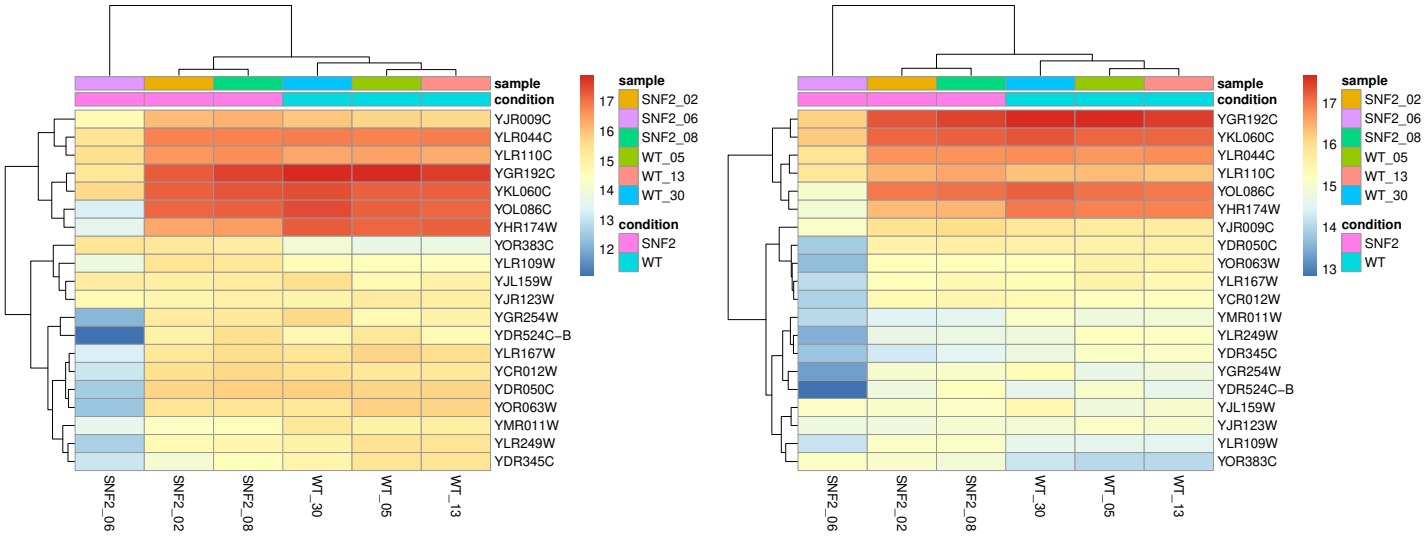

Figure 13: Heat map for 20 most differentially expressed gene

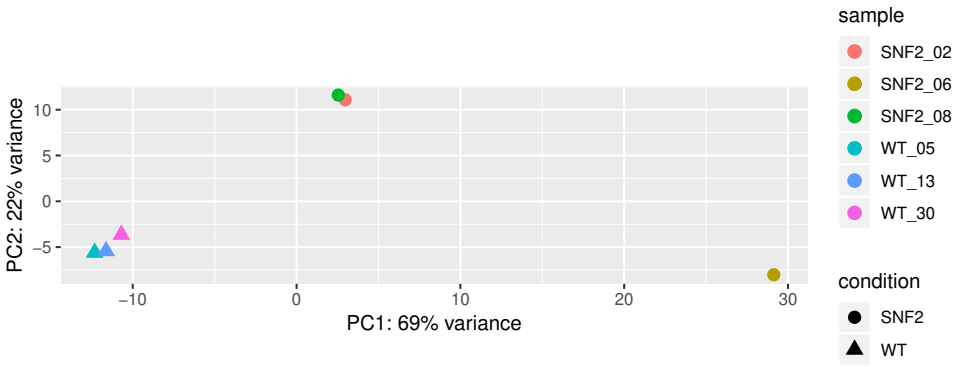

Figure 14: PCA of 10 samples across 2 conditions