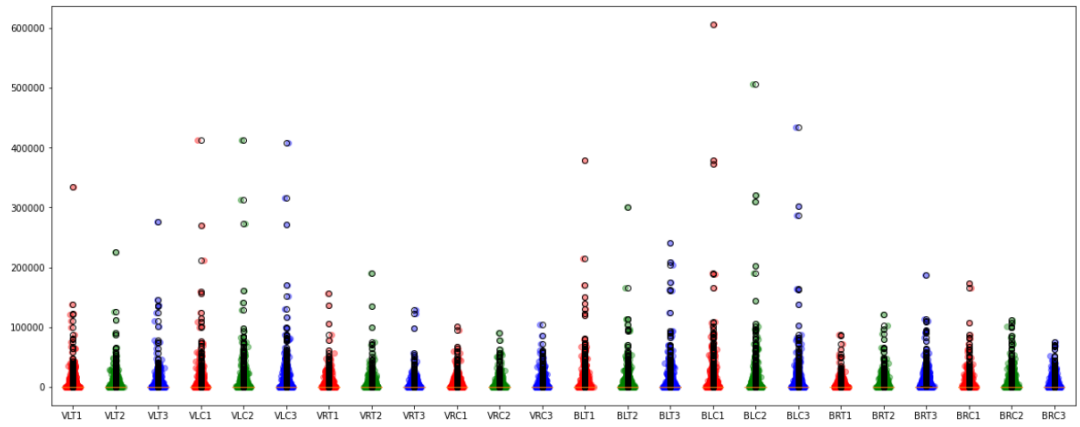
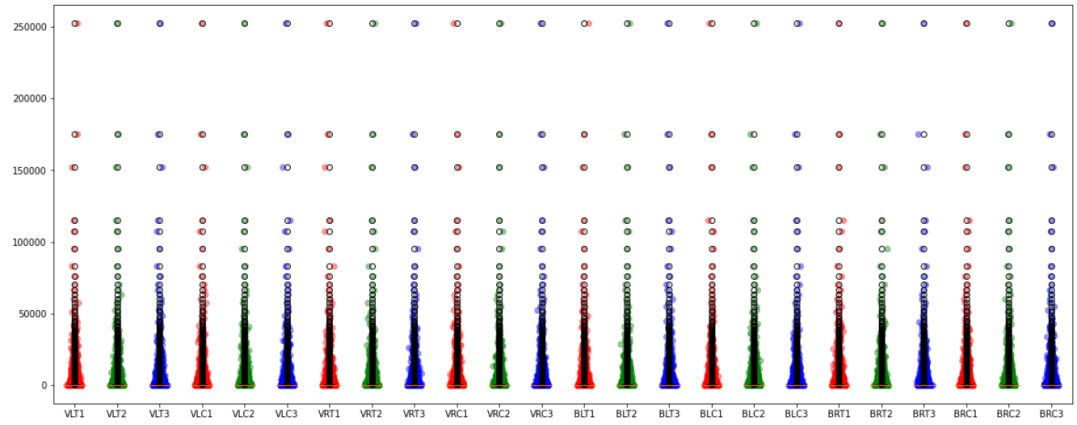
Normalization:

In each experimental condition, due to any parameter, all the data points might shift up/down. So before doing hypothesis testing, we must normalize it. In this case, I used quartile normalization. It forces the distribution of the normalized data to be the same for each sample by replacing each quantile with the average (or median) of that quantile across all samples.[1]. I found a library on github to do it, so it made things easier for me.

Before Normalization:



After Normalization:



Hypothesis testing:

Differentially expressed genes must have a statistically significant change in read-counts. To measure this 'significant' change we need to use p-value. If the p-value is greater than 0.01, that means Null Hypothesis is true (Observations are similar/equal, No difference in genes is observed). I found the p value using scipy.stat.ttest_ind() function.

But this approach can give a 1% False Positive result. To reduce this error I used the Benjamini **and Hochberg** Correction method [2] used in the paper.

Number of Differentially Expressed Genes:

Comparison	My Findings (using, alpha=0.05)	Table-2 of Paper	Error %
VLT vs VLC	5094	5992	14.98%
VRT vs VRC	5901	7221	18.28%
BLT vs BLC	2117	1572	34.67%
BRT vs BRC	2512	2263	11.00%

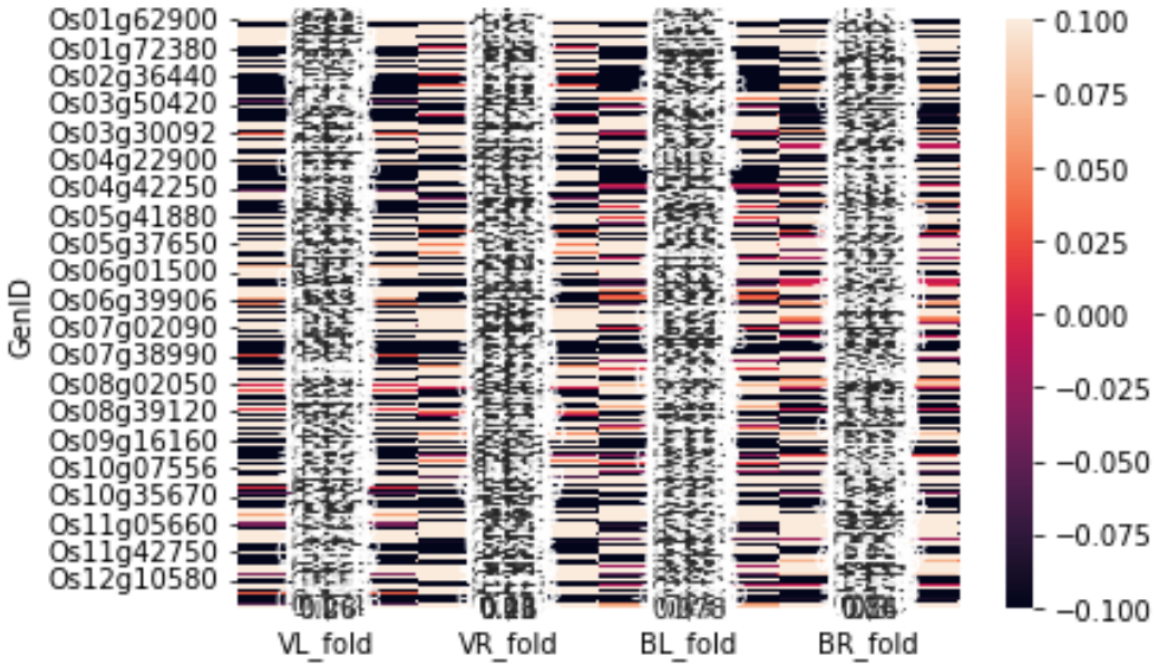
Due to different normalizing and hypothesis testing, my findings and the researcher's findings differ. But, 3/4th time my error is <20%. So, I can conclude that I found a correct way to approximate the answer.

Differentially Expressed Genes:

If any genes showed a 0.01 level significant difference in any of the four experiments, I considered it as a DEG. (though I'm not sure about any/all)

However, all the DEGs and Common_DEGs are stored in a python list for further analysis.

Heatmap:



Bonus Task:

I downloaded the two-pore potassium channel Os03t0752300-01 from the RAP-DB website and the genome of Oryza Sativa from this rice.uga site as a cds file and renamed it as fasta file.

For all the Differentially expressed gene's homology matching, I searched their first occurrence in the OS-genome and extracted their sequence. Aligned them globally using pw2.align.globalxx and calculated homology match. [3] using the formula = score/ max_len

Result:

110 of the 747 genes that were DGE in all 4 studies have over 60% match with K2P.

341 of the 2055 genes that were DGE in any of the 4 studies have over 60% match with K2P.

Submitted By,

Mahdee Mushfique Kamal