NASA Gene Lab update

Alex Seo

#Astrobotany project: Finding patterns from space data using statistical method

So far, from first milestone of the project, objective of the project (name of the project) was defined with help from Choi et.al paper. Then, data exploration was done with BRIC19 data by using R. Few of concerns and possible suggestions with data was discussed. Possible statistical method such as unsupervised learning (e.g. clustering) and deep learning (e.g. NLP based neural networks) was suggested.

After formal discussion with first milestone that I have done, it is decided that new data from NASA Gene lab GLDS-120 will be used to perform the analysis. Therefore, process that was done with first milestone, such as, data preprocessing and data exploration will be performed again and relevant statistical model will be built.

#Study Description of GLDS-120 (Can skip if you know about the study)

Experimentation on the International Space Station has reached the stage where repeated and nuanced transcriptome studies are beginning to illuminate the structural and metabolic differences between plants grown in space compared to plants on the Earth.

Genes that are important in setting up the spaceflight responses are being identified; their role in spaceflight physiological adaptation are increasingly understood, and the fact that different genotypes adapt differently is recognized. However, the basic question of whether these spaceflight responses are required for survival has yet to be posed, and the fundamental notion that spaceflight responses may be non-adaptive has yet to be explored.

Therefore, the experiments presented here were designed to ask if portions of the plant spaceflight response can be genetically removed without causing loss of spaceflight survival and without causing increased stress responses.

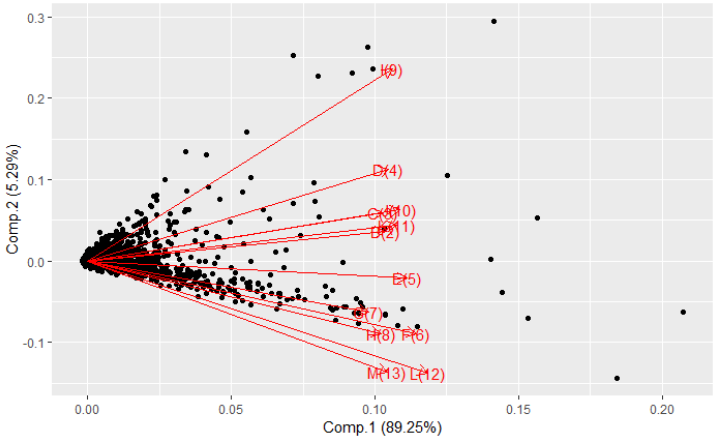
The CARA experiment compared the spaceflight transcriptome responses of two Arabidopsis ecotypes, Col-0 and WS, as well as that of a phyD mutant of Col-0. When grown with the ambient light of the ISS, phyD displayed a significantly reduced spaceflight transcriptome response compared to Col-0, suggesting that altering the activity of a single gene can actually improve spaceflight adaptation by reducing the transcriptome cost of physiological adaptation. The WS genotype showed an even simpler spaceflight transcriptome response in the ambient light of the ISS, more broadly indicating that the plant genotype can be manipulated to reduce the transcriptome cost of plant physiological adaptation to spaceflight and suggesting that genetic manipulation might further reduce, or perhaps eliminate the metabolic cost of spaceflight adaptation.

When plants were germinated and then left in the dark on the ISS, the WS genotype actually mounted a larger transcriptome response than Col-0, suggesting that the in-space light environment affects physiological adaptation, which further implies that manipulating the local habitat can also substantially impact the metabolic cost of spaceflight adaptation.

For this project, ‘Normalized_counts’ data from GLDS-120 will be mainly used and ‘Array_Genediff_pilot', 'RNAseq_Genediff_pilot' and 'RNAseq_Isoformsdiff_pilot' will be used as supplementary data. Normalized counts were expressed in terms of FPKM values (fragments per kilobase of transcript per million mapped fragments). FPKM is directly proportional to abundance of the transcript.

The given dataset is consisted with 2 genotype (WS, col0) and 1 mutant (col0-phyD) Arabidopsis' reaction under ground-control and flight, and under dark and light.

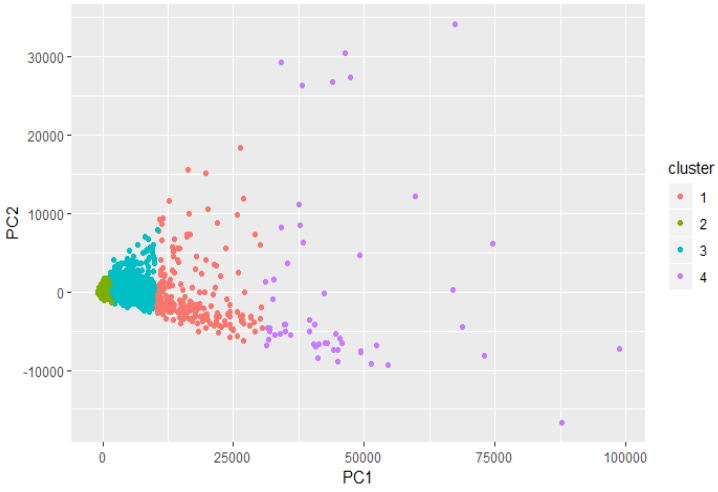
To see if there are distinct patterns from this data PCA and clustering analysis were done



<PCA on col0-phyD>

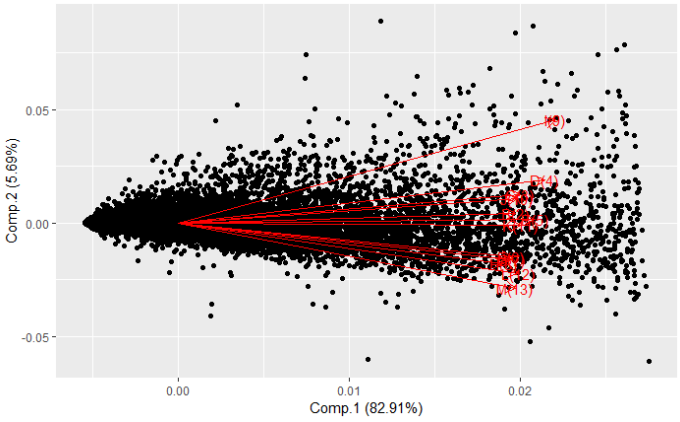
phyD data was first used, and found out that there were 3 common outliers for all genotype.

Row Number of the outliers: 861, 12958, 23875



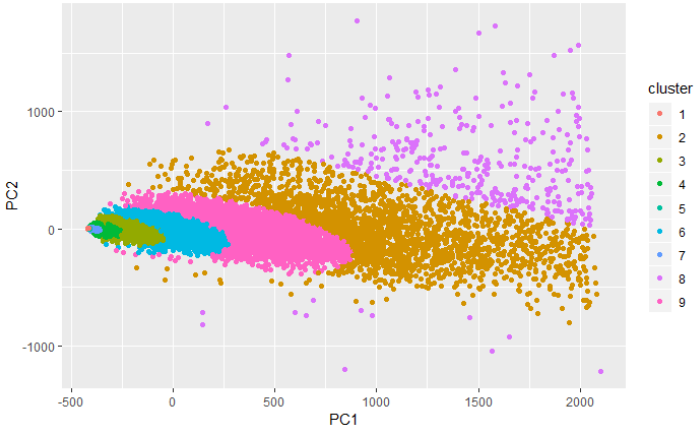
<K-means clustering on col0-phyD>

After discarding these 3 outliers k-means clustering was done with k=4. While it seems each cluster are similarly sized, cluster 2(most inner cluster with green cluster) is highly dense with points with around 90% of the dataset.



<PCA on highly dense cluster>

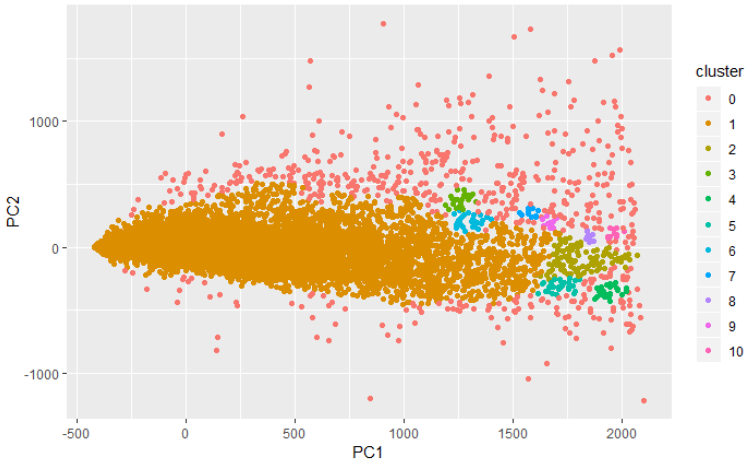
The PCA analysis was done again on col0-phyD dataset, this time only with highly dense cluster from previous k-means clustering analysis.



<Model based clustering>

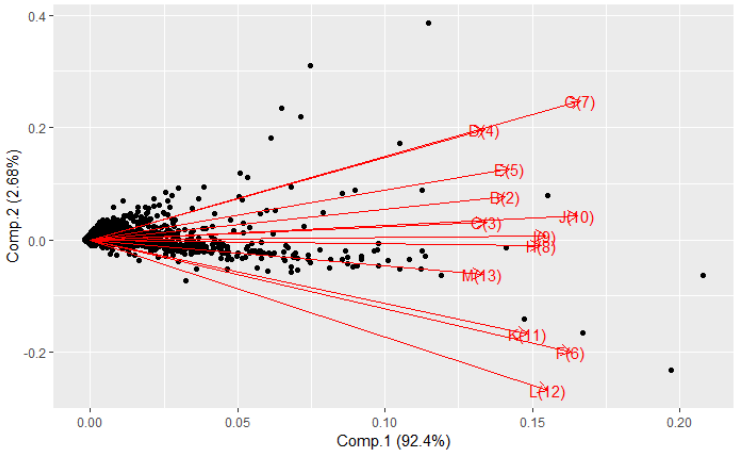
Instead of k-means clustering, model-based clustering was done to divide this data into more detailed clusters. While it seems there is a pattern on how cluster is formed, but we cannot tell anything about it yet.

Then, different clustering method, model that is been used a lot for the clustering recently, DBSCAN was applied. From the result, DBSCAN clustered the dataset less detailed compared to model-based clustering, but it might indicate the one big chunk of cluster (yellow cluster) contains more information than other clusters.



<DBSCAN>

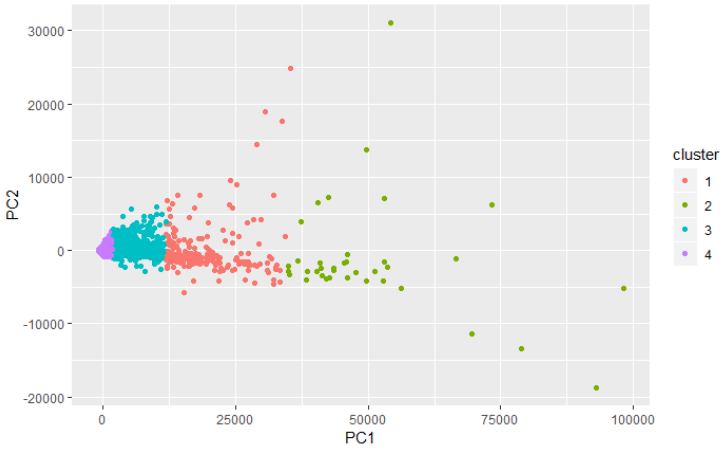
Then, col0 data was used for the same analysis that was done for the col0-phyD data



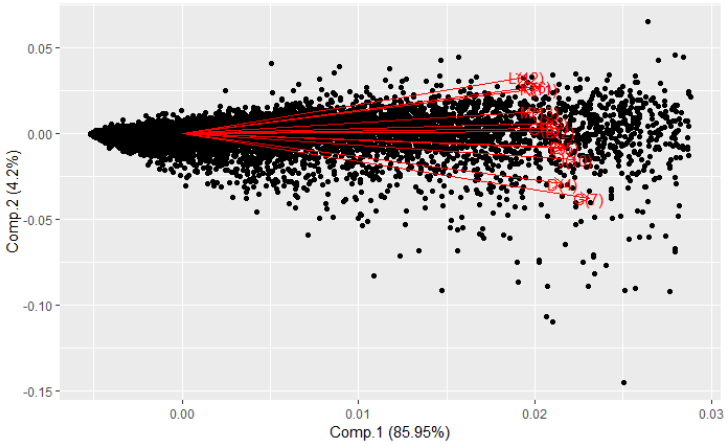
<PCA on col0>

From the PCA analysis on col0 data we can see that the data is distributed similarly to col0-phyD's PCA analysis having expanding variation, which can indicate that the experiment that was done on the different genotype of Arabidopsis has similar result. This might mean that there is no significant difference between genotype on the experiment that has been done.

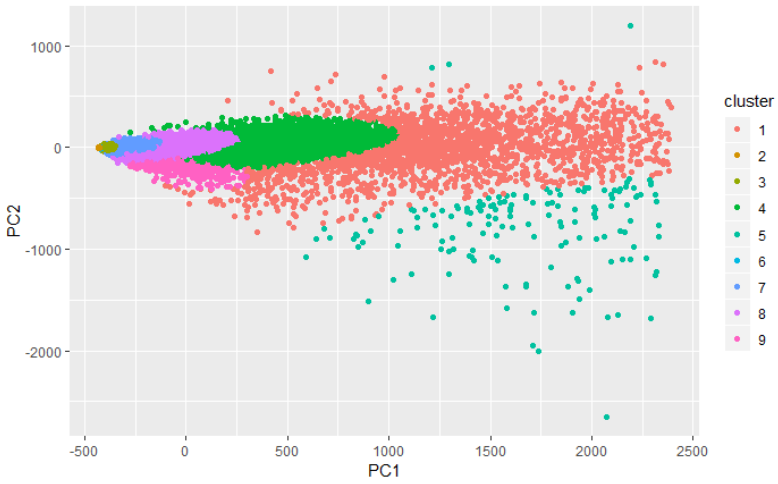
From k-means clustering with k=4, it also gave us similar result by one cluster being highly dense with data compared to other clusters. Therefore, PCA analysis was done again with highly dense cluster similar to what was done with col0-phyD data.



<K-means clustering on col0>

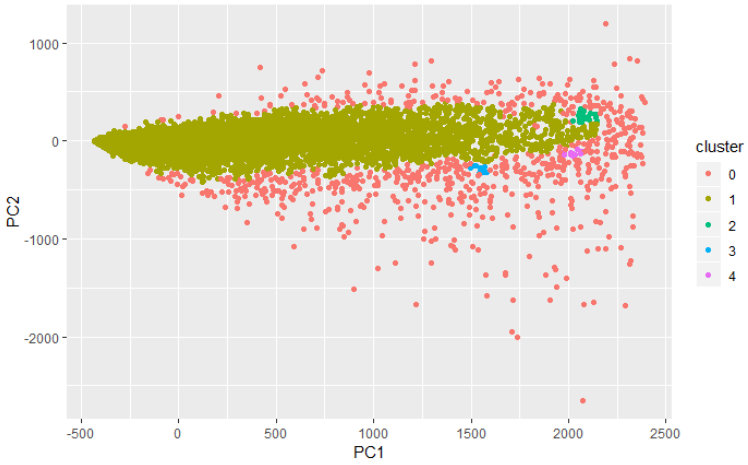


<PCA on highly dense cluster>



<Model-based clustering>

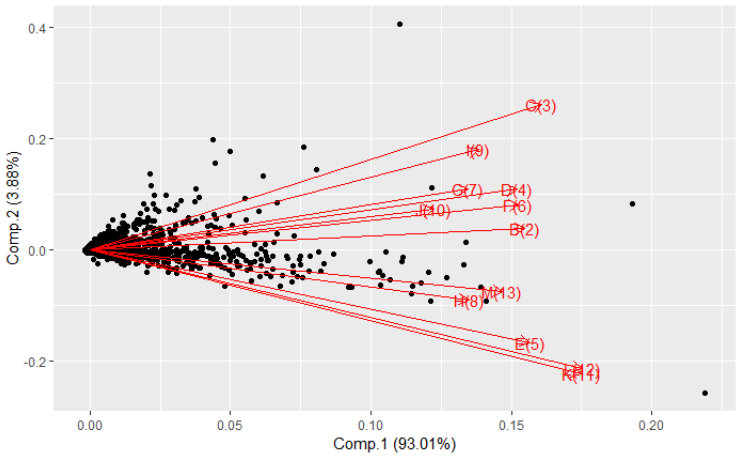
Model based clustering was performed to the data, similar to col0-phyD. Interesting point is that the way how clusters are shaped was similar to the cluster that was shaped on col0-phyD data. From this finding we can dig deep into that pattern and see what was the reason behind.



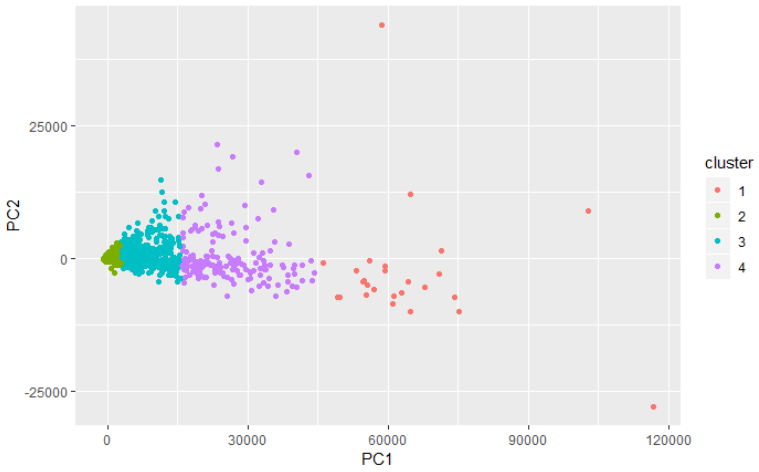
<DBSCAN>

DBSCAN's result on col0 had also similar result.

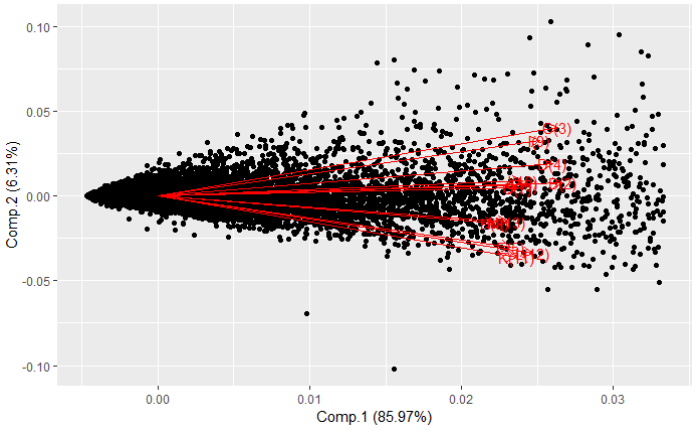
Then, lastly, WS data was used for the same analysis that was done so far.



<PCA on WS>

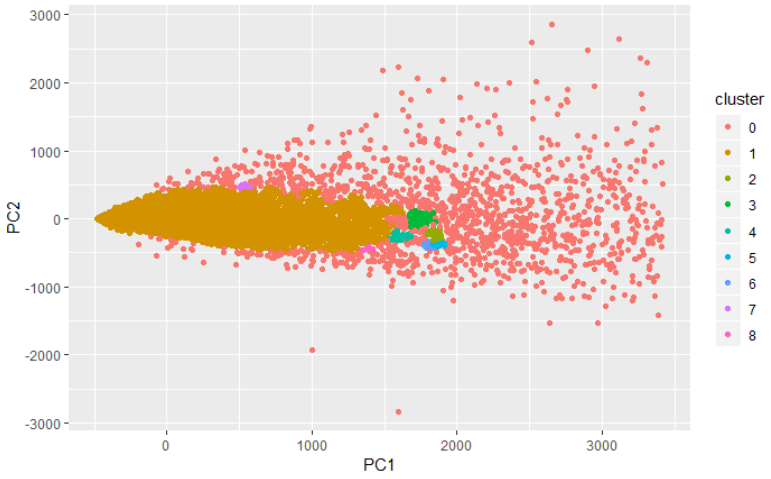


<K-means clustering on WS>

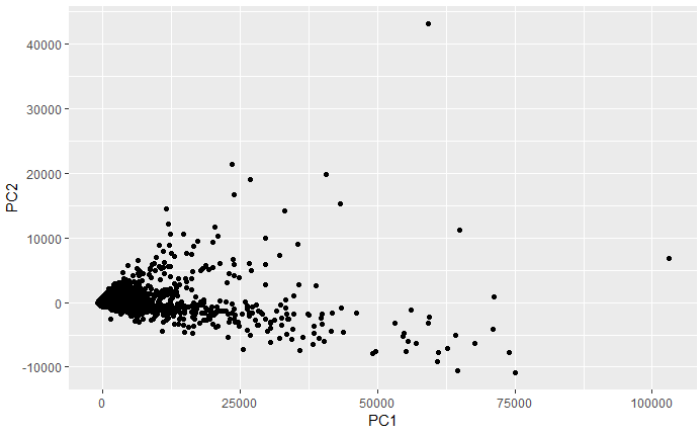


<PCA on highly dense cluster>

The result with PCA analysis and k-means clustering was yet again consistent. With the PCA analysis on highly dense cluster from k-means clustering on WS data, DBSCAN clustering was performed. This also showed some similar result with other genotypes. To go step further, I performed PCA analysis on highly dense cluster from DBSCAN clustering



<DBSCAN>



<PCA on highly dense cluster>

As you can see from the plot, it was interesting that PCA analysis on highly dense cluster from DBSCAN clustering seems similar to original PCA analysis on WS data having expanding variation with shape <.

From all the analysis that was done I come up with 3 suggestion for the following experiment.

1. Dig deep on the pattern that was found from this analysis. Clustering analysis on all the genotype has been consistent. K-means clustering having on highly dense cluster, model based clustering having same cluster shape and with PCA analysis on highly dense cluster from DBSCAN being similarly shaped to PCA analysis on original dataset. There could be some indication behind this cluster on why the cluster was shaped like this. I could perform analysis on each cluster and check the characteristic on the cluster. However, it would be time consuming and will need insight from biologist to catch that characteristic from the pattern. It might not be suitable to find difference between flight and ground control experiment.

2. Use different method on pattern recognition analysis, such as kernel method to see if the similar patterns from this cluster analysis appear.

3. This suggestion is most promising than others above in my opinion,

   Change the structure of the dataset in order to make it possible to perform different type of analysis. Given dataset is constructed like,

RNA_seq	Col0_FLT_light	Col0_FLT_dark	Col0_GC_light	Col0_GC_dark
AT1G01010	#	#	#	#
AT1G01020	#	#	#	#

   This dataset is an unlabeled dataset without target data(y_data). By re-constructing

this dataset to labeled dataset, it will allow us to perform various analysis such as classification and prediction. However, by performing classification it does not lead us to anything, therefore it is important to implement a classification model that is easy to interpret and get useful result.

The proposed reconstructed dataset will look like,

RNA_seq	Genotype	Light setting	NC_rep1	NC_rep2	NC_rep3	Location
AT1G01010	Col0	Light	#	#	#	FLT
AT1G01010	Col0	Dark	#	#	#	FLR
AT1G01010	Col0	Light	#	#	#	GC
AT1G01010	Col0	Dark	#	#	#	GC
AT1G01010	WS	Light	#	#	#	FLT
AT1G01010	WS	Dark	#	#	#	FLT
AT1G01010	WS	Light	#	#	#	GC

This proposed dataset will have labels, unlike given data, with Location variable. We can construct classification model to predict if the data's experiment was done in flight or from the ground using variables such as, RNA_seq, Genotype, Light setting and normalized counts of fpkm. We can use model that is easy to interpret so we can figure out which variables contributed and affected most to predict if the data is from flight or not.

In order to do this, I propose deep factorization machine based neural network which is commonly used for the recommendation system for YouTube and such. This model will help us achieve high prediction accuracy and interpretability using feature interaction that is embedded internally in the model. We can also change the target variable to genotype or light setting to get different perspective of the dataset.