Mini-Project 2 Checkpoint 1

ECE/CS 498DS

Spring 2020

Jiashuo Tong (jtong8), Yilin Zhu (yilinz10), Rongqi Gao(rongqig2)

Task 1 - Question 0

1. Why do biologists need multiple samples to identify microbes with significantly altered abundance?

Using multiple samples, we will be more confident in rejecting a null hypothesis if a significantly altered abundance occurs. If we use a two-sample Z-test, larger sample sizes will yield a larger Z score, which allows us to reject H_0 with greater confidence level. If we use a two sample K-S test, larger sample sizes will give smaller critical values (on the right hand side). Thus, we can achieve higher confidence level with the same D value (K–S statistic).

2. Number of samples analyzed: 764
3. Number of microbes identified: 149

Task 1 – Question 1

- a. Factorization of joint probability distribution:

$$P(Q, C, LT, ST, CM) = P(Q|C, LT) \times P(C|ST, CM) \times P(ST) \times P(CM) \times P(LT)$$

where Q denotes "Quality", C "Contamination", LT "Lab Time", ST "Storage Temp", and CM "Collection Method"

- b. Number of parameters needed to define conditional probability distribution: 11 parameters
- c. Conditional probability tables:

	qual	bad	good
cont	labtime		
high	long	0.966102	0.033898
	short	0.064257	0.935743
low	long	0.080997	0.919003
	short	0.042907	0.957093

	cont	high	low
strtmp	coll		
cold	nurse	0.043983	0.956017
	patient	0.076577	0.923423
cool	nurse	0.088435	0.911565
	patient	0.838235	0.161765

strtmp	cold	cool
coll	0.8982	0.1018

coll	nurse	patient
strtmp	0.8976	0.1024

labtime	long	short
coll	0.2044	0.7956

Task 1 – Question 1 (continued)

- d. Table of P(Quality|Storage Temp, Collection Method, Lab Time):

	strtmp	coll	labtime	good	bad
0	cold	nurse	long	0.887962	0.112038
1	cold	nurse	short	0.955112	0.044888
2	cold	patient	long	0.862069	0.137931
3	cold	patient	short	0.943978	0.056022
4	cool	nurse	long	0.822785	0.177215
5	cool	nurse	short	0.972376	0.027624
6	cool	patient	long	0.117647	0.882353
7	cool	patient	short	0.960784	0.039216

- e. Total number of samples dropped: We dropped samples from HE0Sample_699 t0 HE0Sample_763, and HE1Sample_699 t0 HE1Sample_763, totally 65 * 2 = 130

Task 1 – Question 2

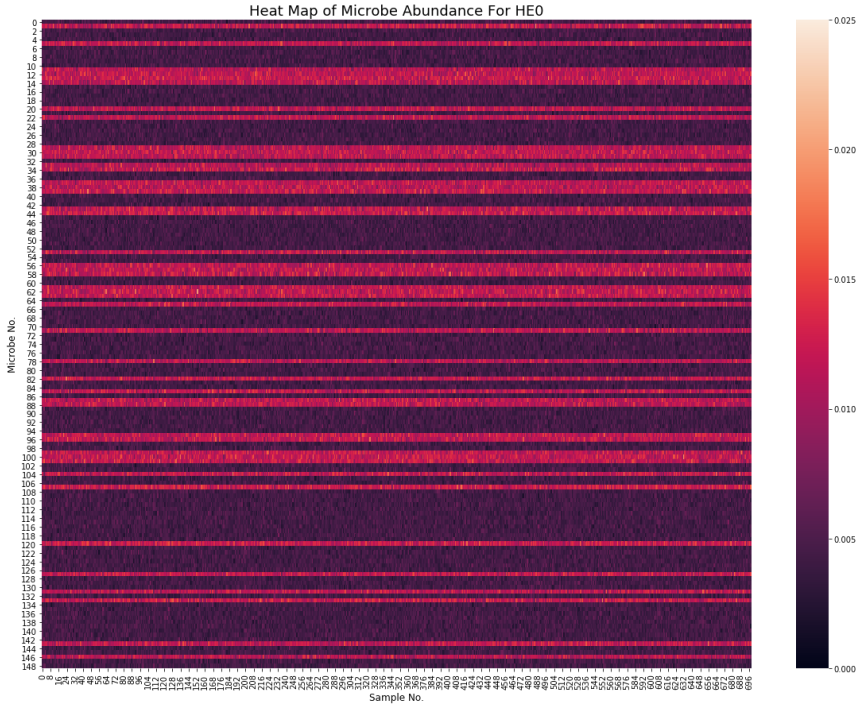
- 1. Number of samples removed: Zero.
- 2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

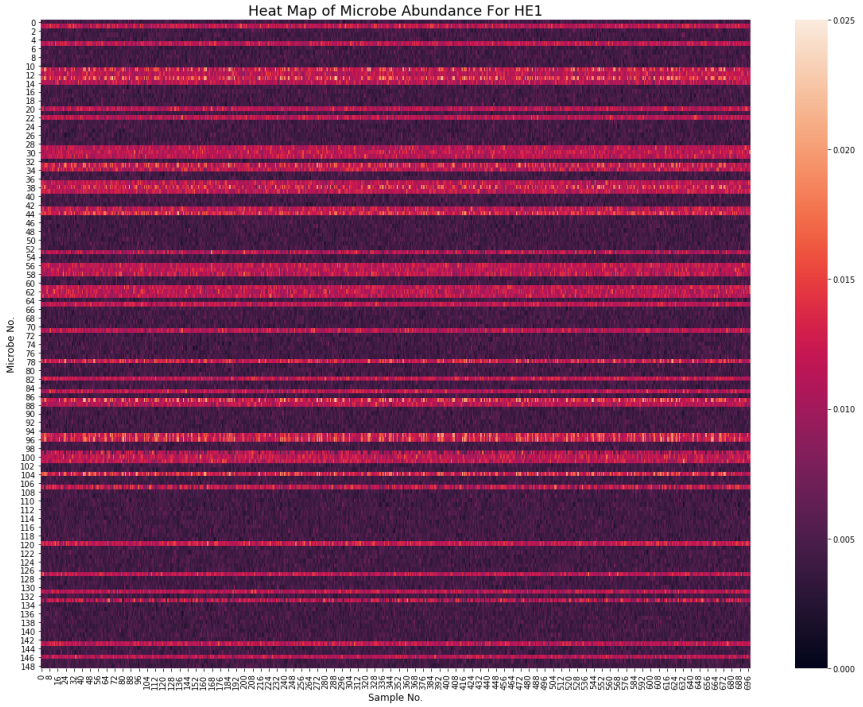
Advantage: The absolute abundance for microbes might vary from person to person, so it cannot be used to measure the abundance level. The relative abundance is a standardized variable, so we can use it to compare the abundance levels of two samples.

Drawback: The relative abundance loses information about how much of each microbe exists. This information can be critical in the sense that the amount of microbes may also change to whether a person has HE or not.

Task 1 – Question 3

- Heatmaps (HE0 on left HE1 on right):





- Summarize your observations

From the two heatmaps, we could find that the pattern of the relative abundance of microbe of HE0 and HE1 samples are approximately the same, for each sample the relative abundance of microbe is similar.

Task 1 – Question 3 (continued)

- Which aspects of the data are the heatmaps good at highlighting? What types of things are heatmaps less suitable for?

Pros: The heatmaps are good at highlighting the data distribute pattern of a two-dimensional data set. Since it uses color to represent the amount of each data point, we can easily find out how the larger and smaller data points distributed. And for the same reason, we can easily get the information from a huge data set.

Cons: The shortcoming of heatmaps is that we are poor at comparing the amount of data points in non-adjacent regions, in some situations we might misread the heatmaps.

Task 2 – Question 1

- b. What is the null hypothesis of the KS test in our context? Use one microbe as an example to explain your answer.

  Our null hypothesis is that the HE0 samples and HE1 samples obey the same distribution.

- c. Count the number of microbes with significantly altered expression at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Summarize your answers in a table below:

	Confidence Level	Number of Significantly Altered Microbes
0	0.100	50
1	0.050	37
2	0.010	27
3	0.005	26
4	0.001	21

Task 2 – Question 2

- a. What does a p-value of 0.05 represent in our context?

  A p-value of 0.05 from KS test on a microbe means the probability of observing the HE0 & HE1 samples is only 5%, given that the null hypothesis (the microbe is not altered) is true.

- b. If the null hypothesis is true, what distribution will the p-values follow?

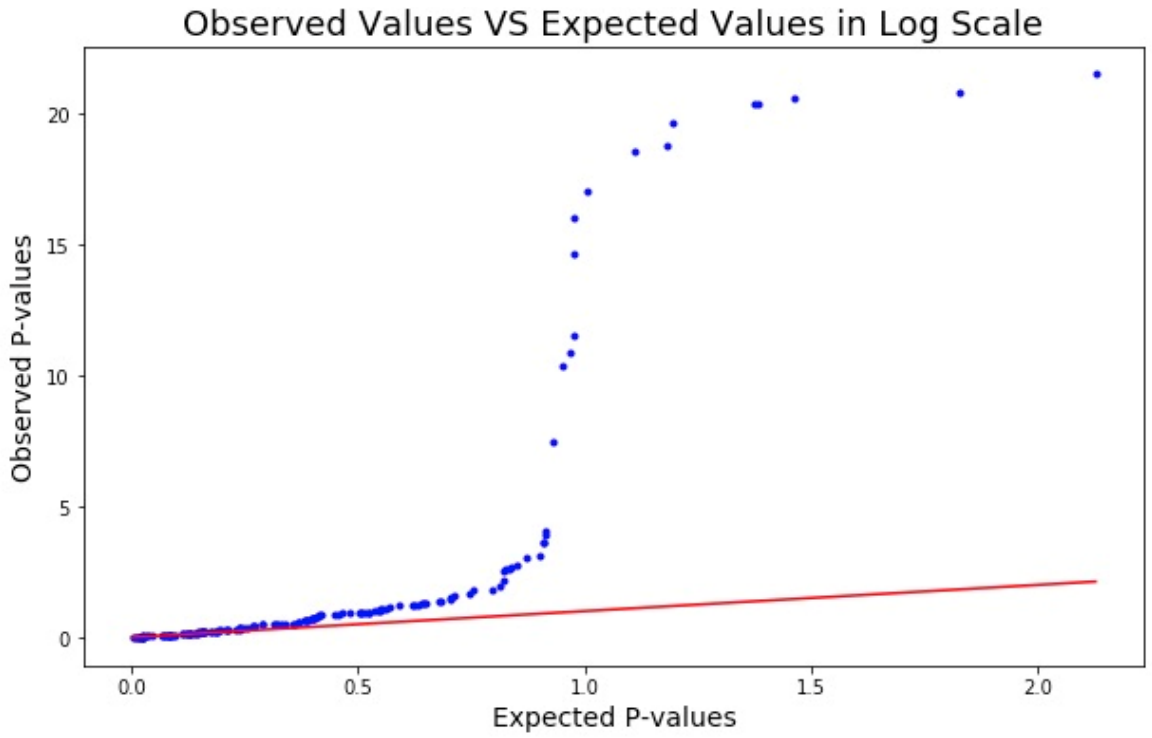
  Uniform distribution. The p-value or probability value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct. Thus, given the null hypothesis is true, the cdf for the random variable P will be F(P) = P. The corresponding pdf will be f(P) = 1, which is Uniform distribution

- c. If no microbe's abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.c. Show the comparison in a table below:

	Confidence Level	Number of Significant P-values	Number of Significant P-values Given H0
0	0.100	50	14
1	0.050	37	7
2	0.010	27	1
3	0.005	26	0
4	0.001	21	0

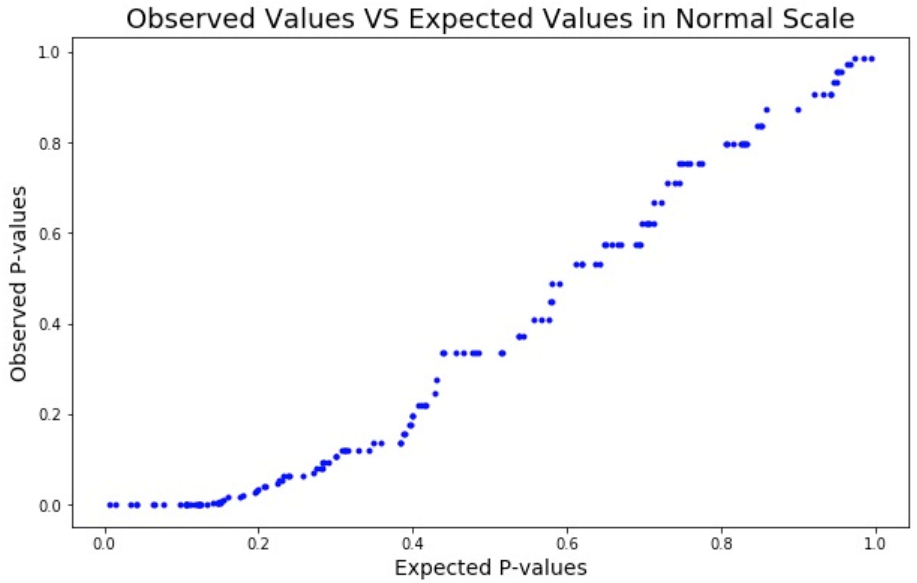
Task 2 – Question 2 (continued)

- d. Q-Q plot:



Task 2 – Question 2 (continued)

- e.i. How does taking the -log10() of the p-values help you visualize the p-value distribution?



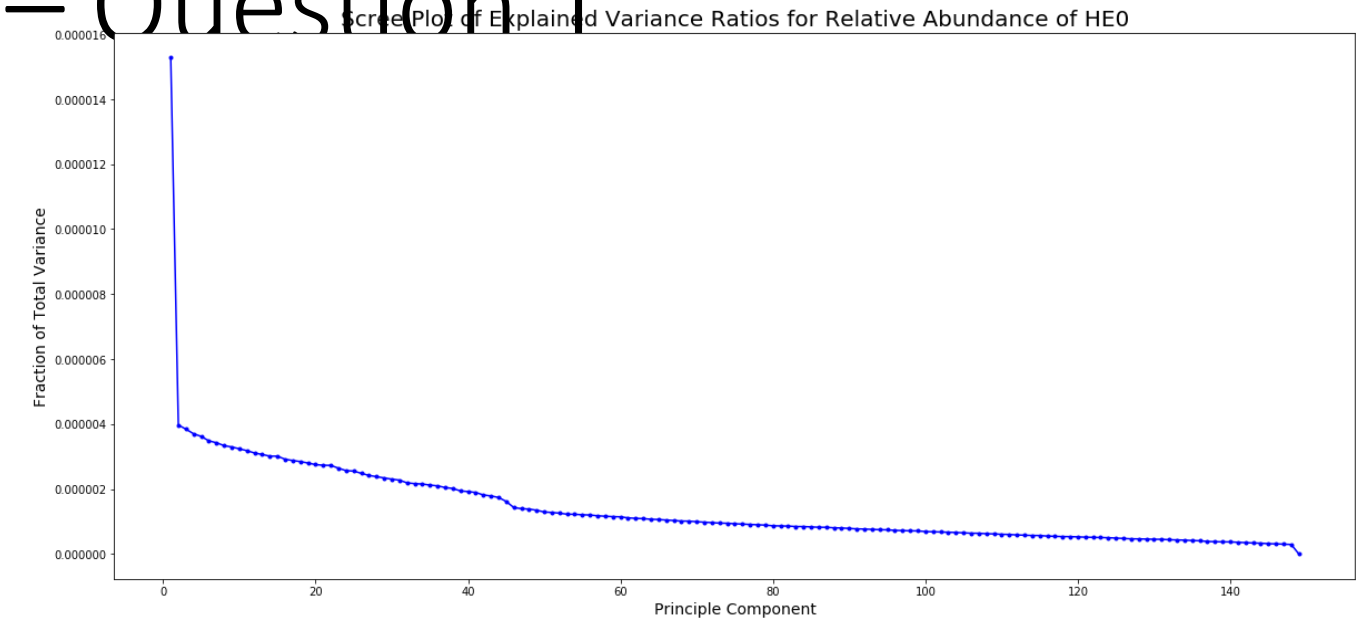
Taking the -log10() of the p-values magnifies the deviation of the observed expected. Say we have a data point (0.001, 0.1). If it is plotted with the orig will be very close to the x=y line. The -log10() values, which are (3,1), stay c margin. Thus, if we used the original p-values, we would have created an a 'linear' curve (as shown in the figure on the right), and we would have drav conclusion. Hence, the -log10() values help us easily identify the non-linear curve which leads to the correct conclusion.

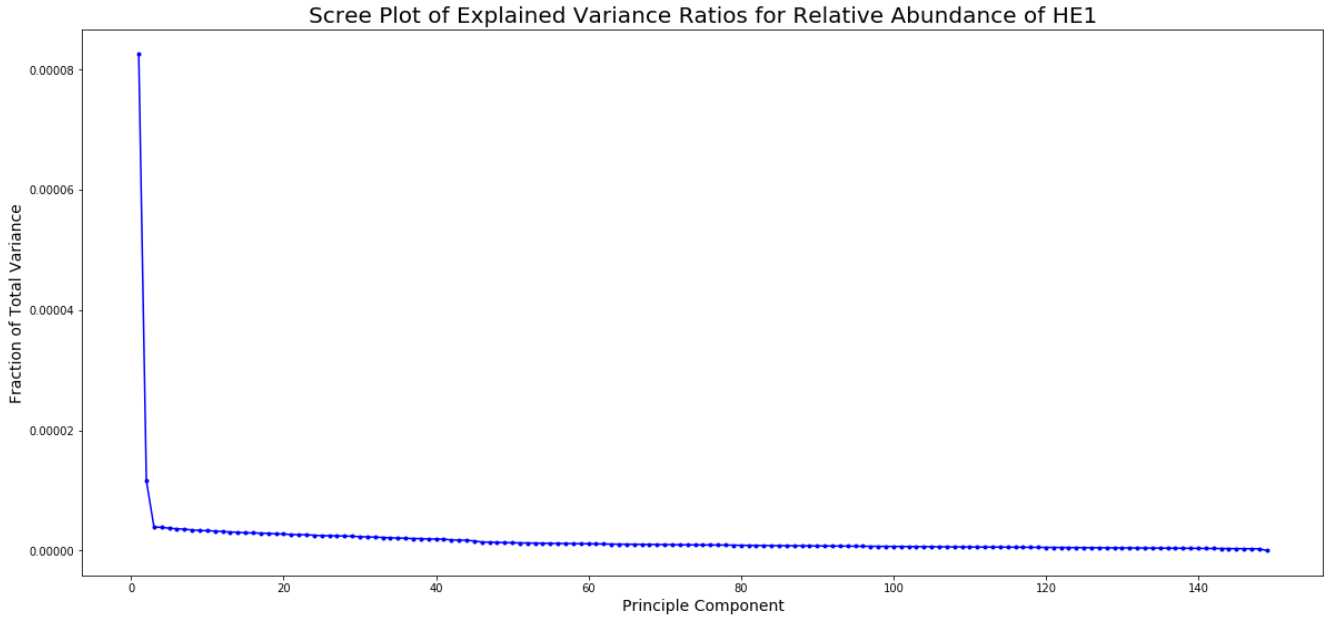
- e.ii. What can you conclude from the Q-Q plot?

We can conclude from the Q-Q plot that the null hypothesis should be rejected. Reason: If the null hypothesis is true, the Q-Q plot should approximately align with the x=y line. However, our Q-Q plot obviously deviates from the x=y line. Therefore, the null hypothesis is likely to be false.

Task 3 – Question 1

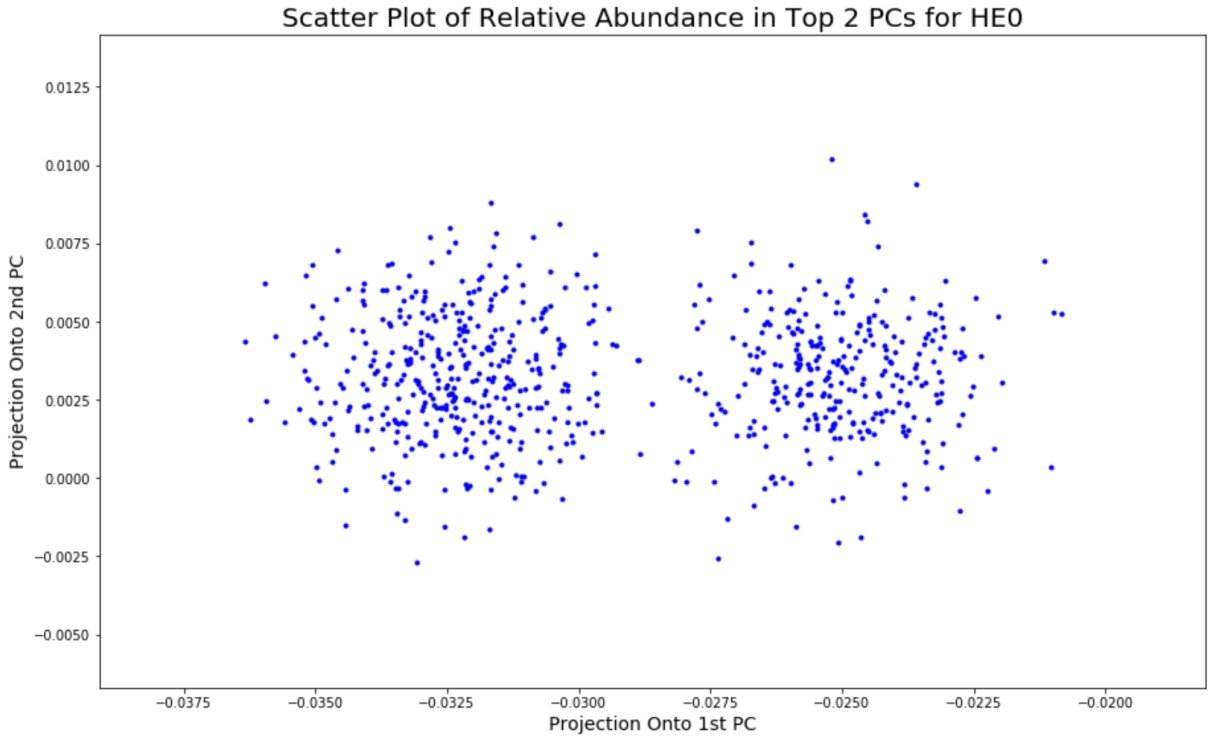
- b. Scree plots:

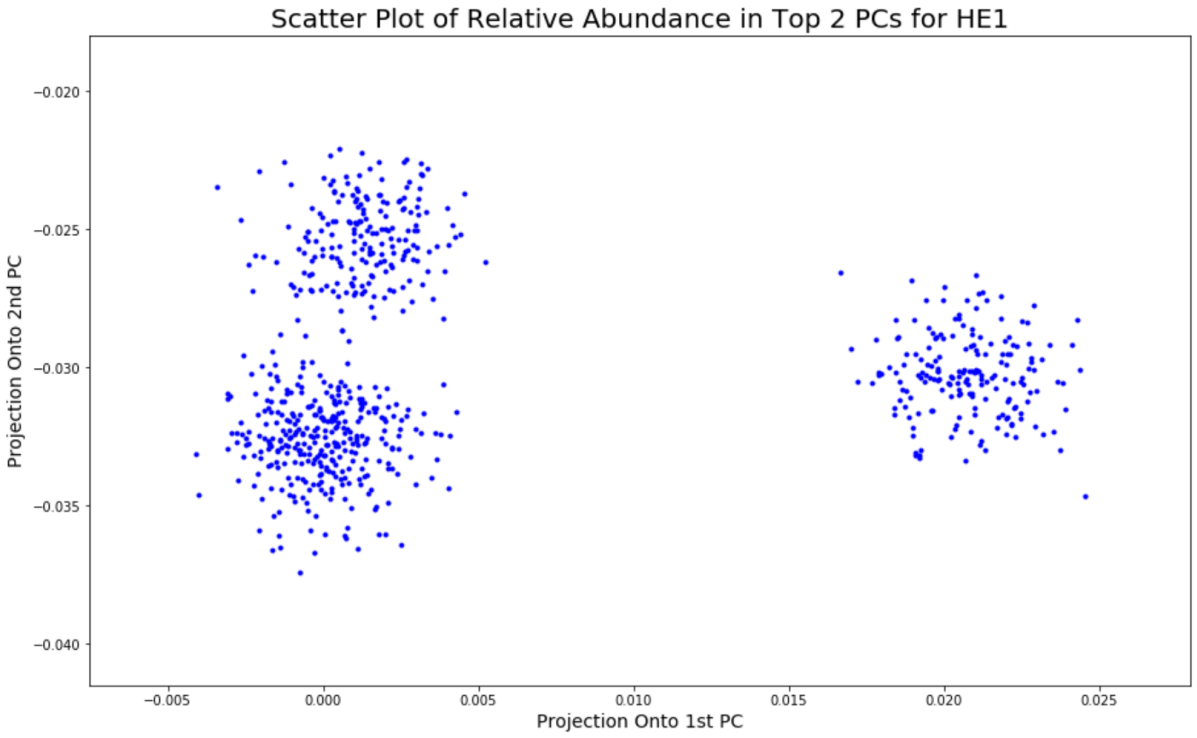




Task 3 – Question 1 (continued)

- c. Plots:



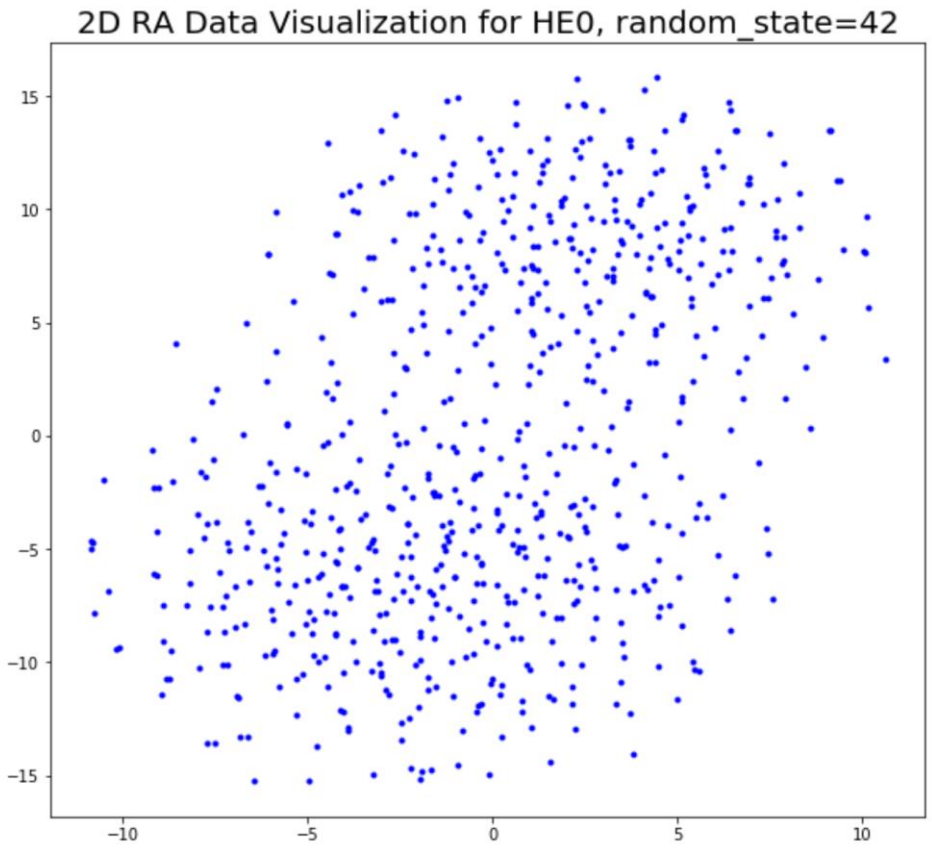


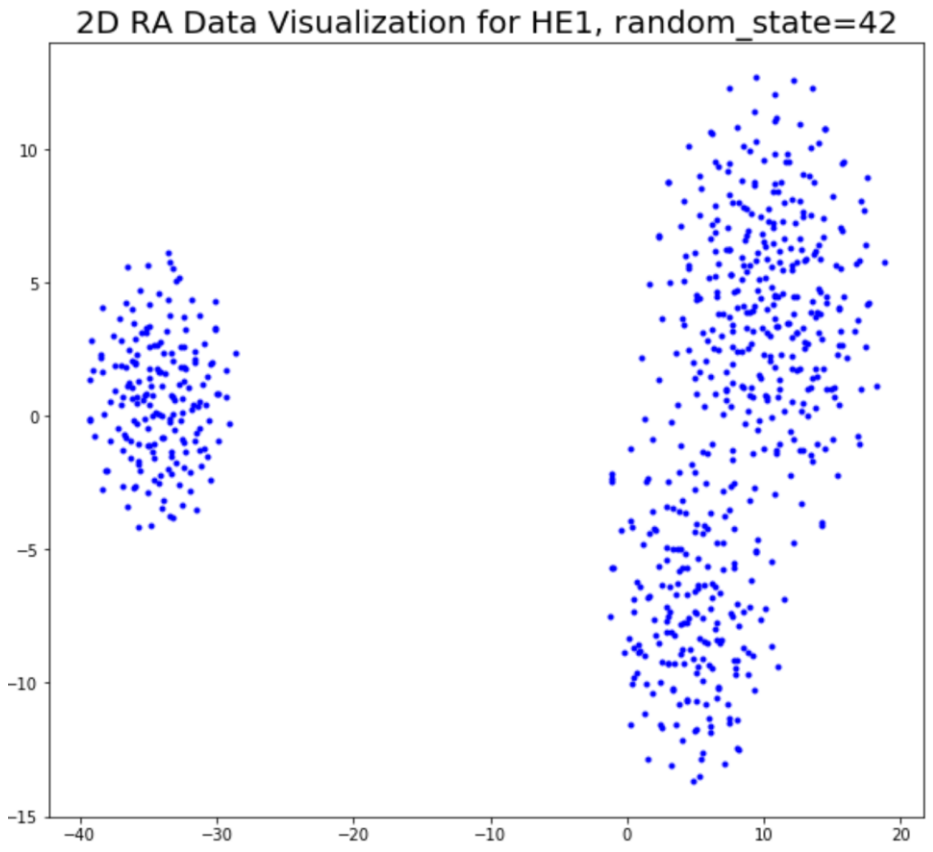
- Observations:

  After projecting the original data into 2 dimensional space (PC1 and PC2), we observed 2 clusters for HE0, and 3 clusters for HE1.

Task 3 – Question 2

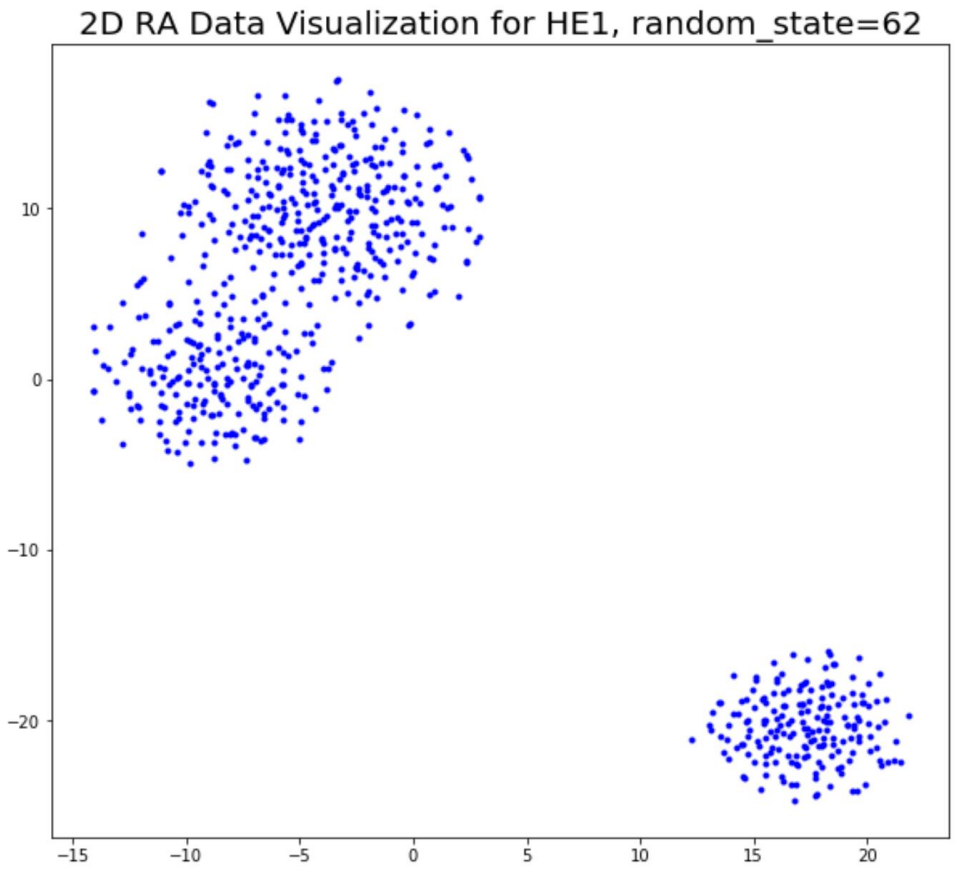
- c. Plots (random_state=42):

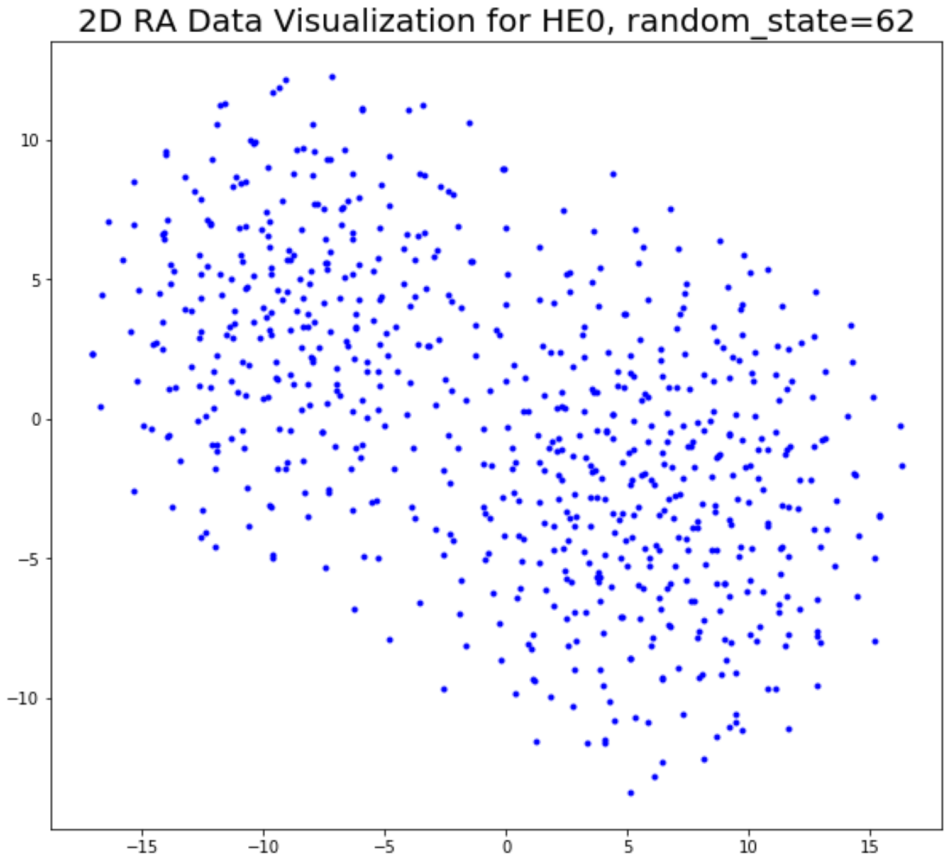




Task 3 – Question 2 (continued)

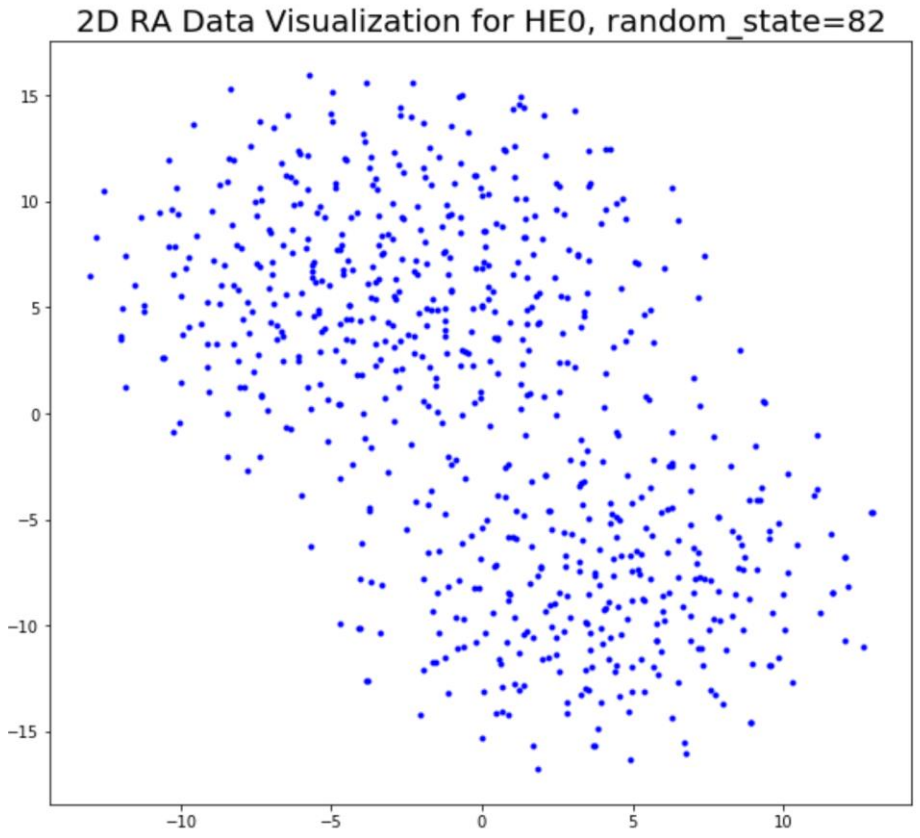
- c. Plots (random_state=62):

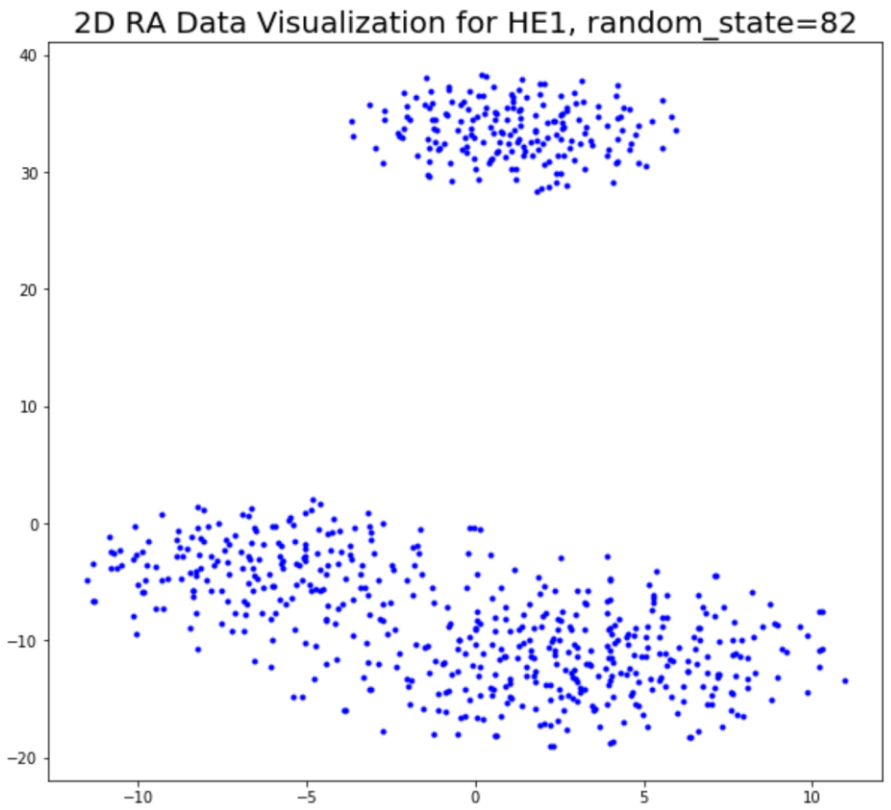




Task 3 – Question 2 (continued)

- c. Plots (random_state=82):





- Observations: We can identify 2 clusters for HE0, but the boundary between them is not so clear. For HE1, we identify one large cluster sitting on bottom, and one small cluster sitting on top.")

Task 3 – Question 2 (continued)

- d. Discussion of similarities and differences between PCA and t-SNE results:

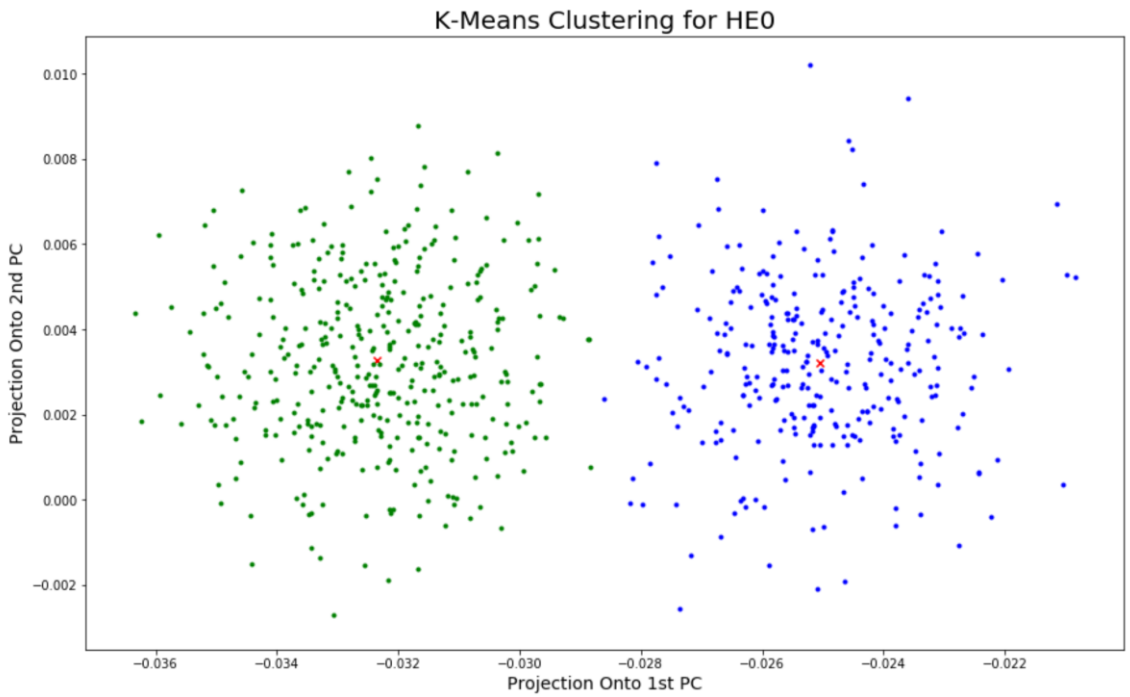
Answer: Similarities and differences: PCA is linear dimensionality reduction while TSNE is nonlinear dimensionality reduction. Comparing the graphs, we can see clusters have formed for both computing techniques. The HE1 graphs show that the clusters from PCA are more 'circular' than those from TSNE.

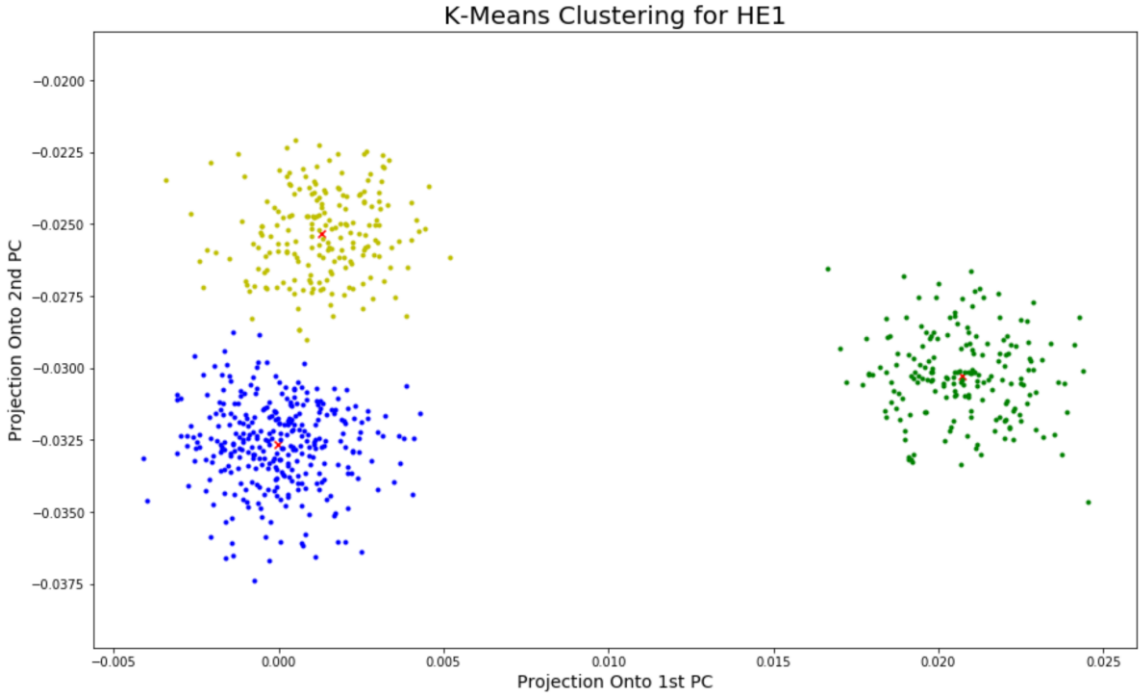
Task 3 – Question 3

- From Task 3.1.c we could easily see that there are 2 clusters in HE0 and 3 clusters in HE1, which gives us a basic judgement of the number of clusters, but we also want to choose the k reasonably. Therefore, we used the Calinski-Harabasz Index as our metric to evaluate the clustering results.
- Calinski-Harabasz Index is also known as Variance ratio criterion, it follows: CH(k) = [B(k) / W(k)] \times [(n-k)/(k-1)], where n = # data points k = # clusters W(k) = within cluster sum-of-squares B(k) = between cluster sum-of-squares.
- Evaluating the ratio for the different models with increasing k, the optimal clustering should be given by the first local maximum of the ratios (Calinski and Harabasz, 1974).
- We will provide scores for each method on both HE0 and HE1 in Task 3.3.e

Task 3 – Question 3

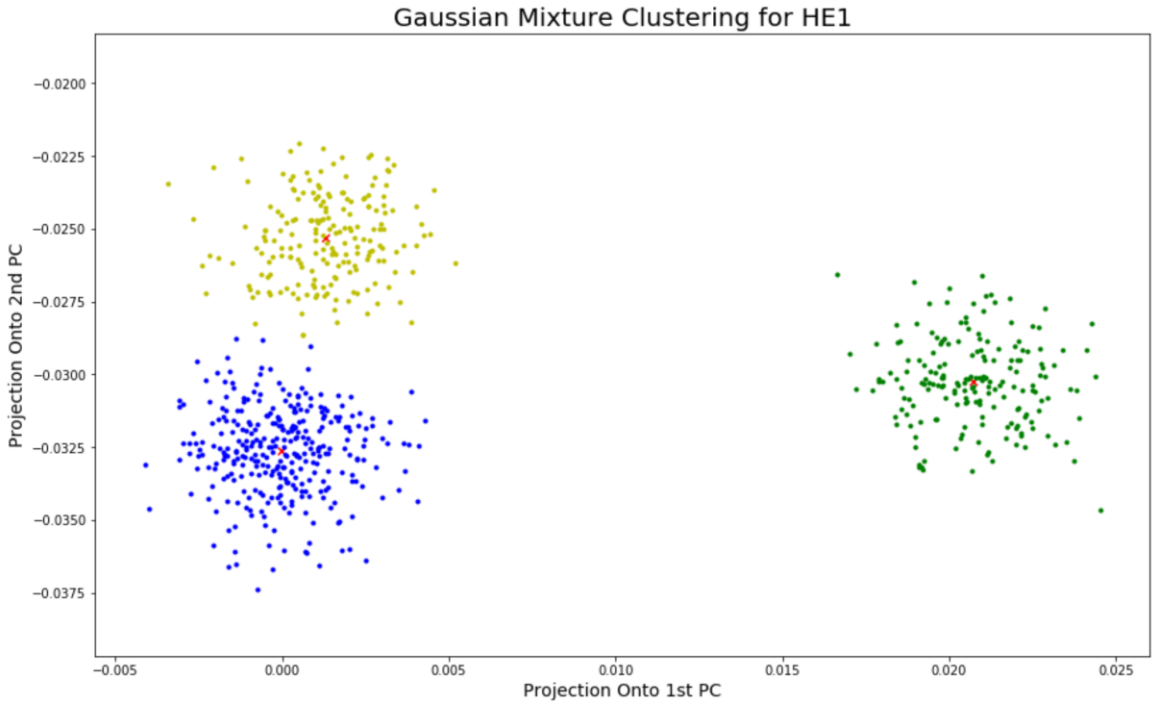
- a. K-means: (From pure inspection of the PCA graphs, we choose k=2 for HE0 and k=3 for HE1.)

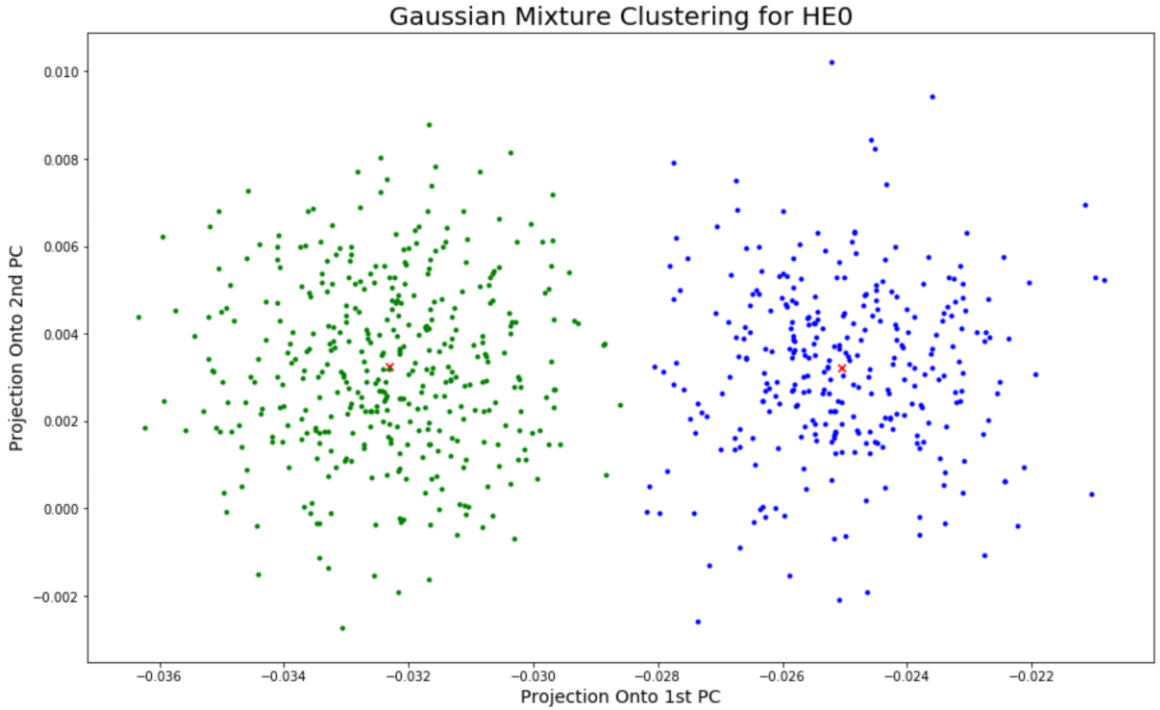




Task 3 – Question 3 (continued)

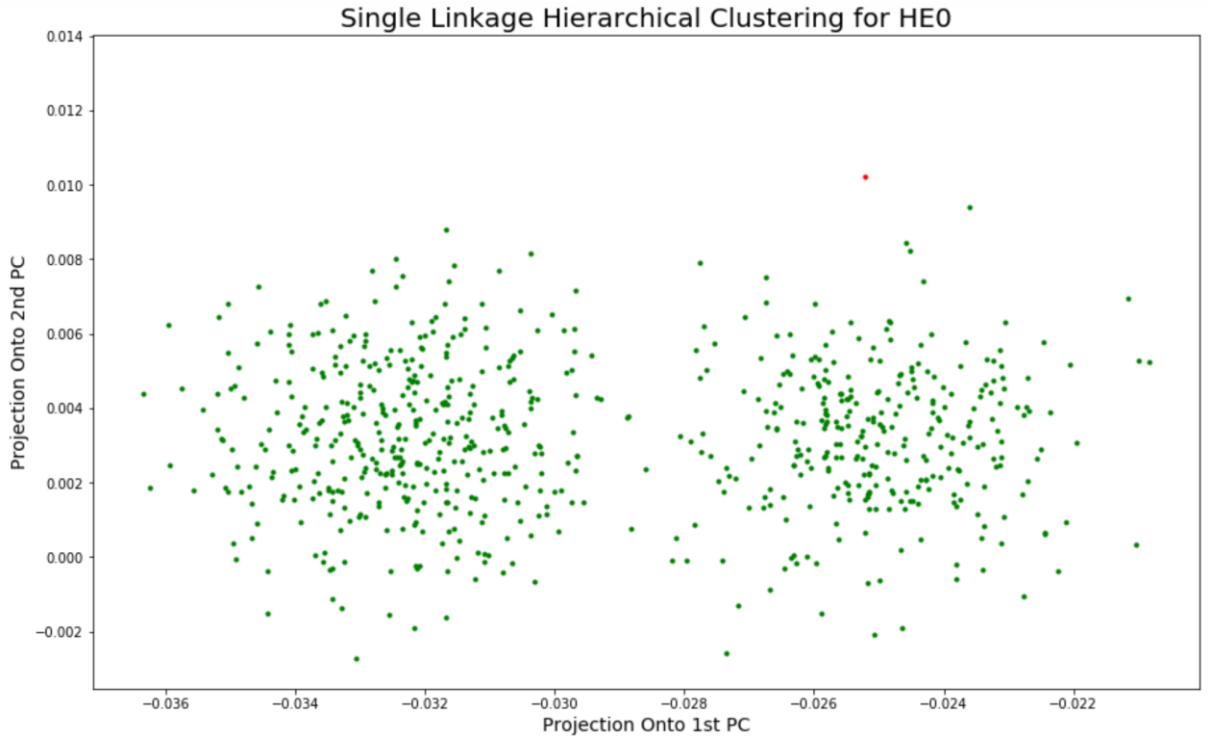
- b. Gaussian mixture model:

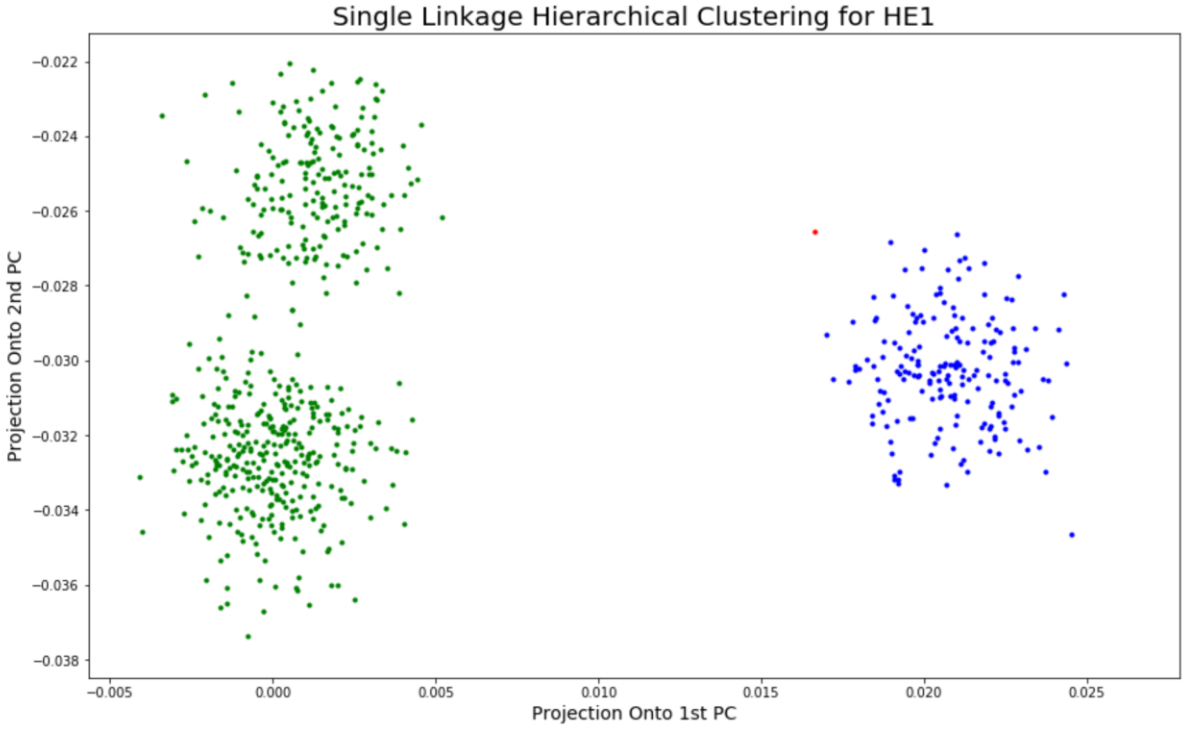




Task 3 – Question 3 (continued)

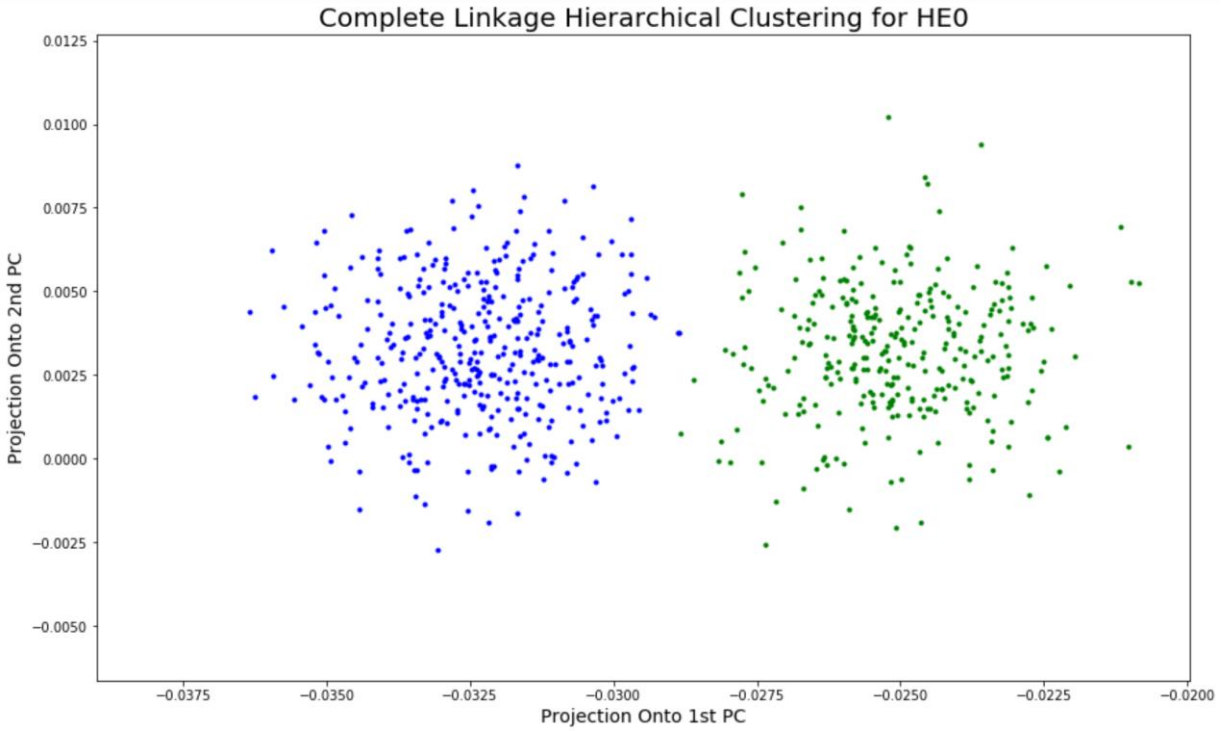
- c. single linkage hierarchical:

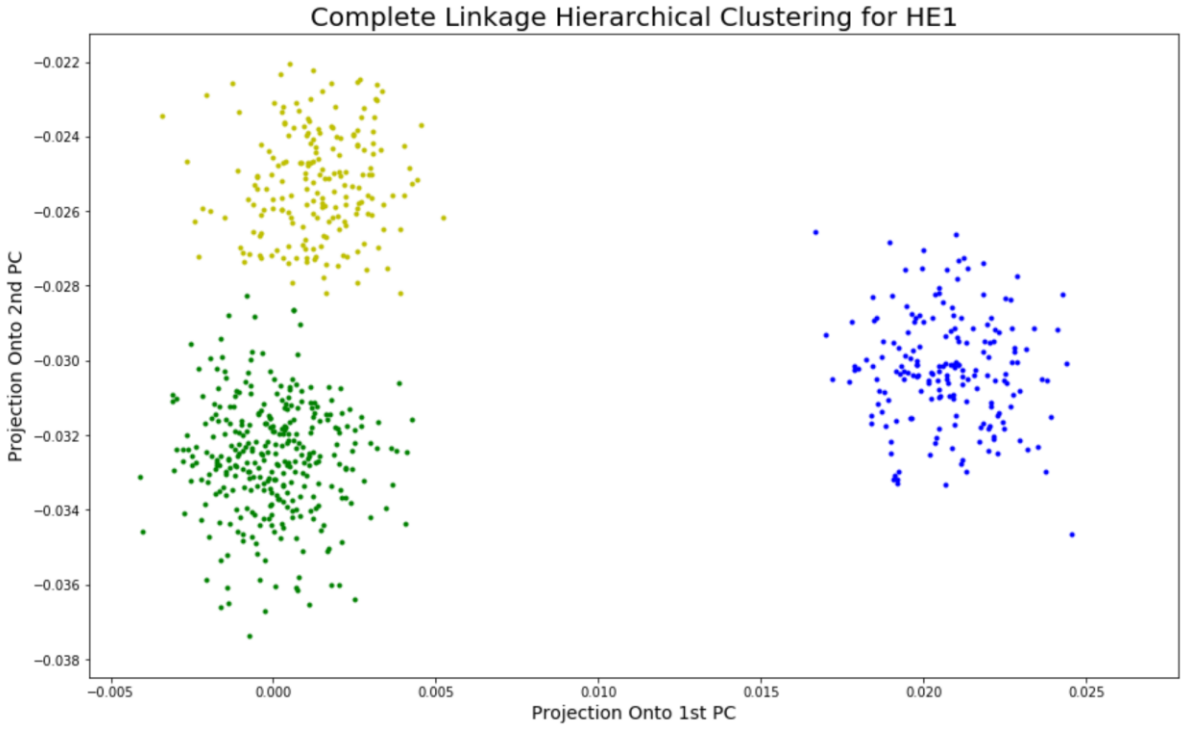




Task 3 – Question 3 (continued)

- c. complete linkage hierarchical:





Task 3 – Question 3 (continued)

- d. Discuss the differences between the single and complete linkage hierarchical clustering methods. Do you see any major differences in the generated clusters? Is there anything about our data which affects if we see a difference between the two linkage options?

1. Differences: In complete linkage hierarchical clustering, we merge in each step the two clusters with the smallest **maximum** pairwise distance. In single linkage hierarchical clustering, we merge in each step the two clusters with the smallest minimum pairwise distance.

2. The complete linkage hierarchical method performs better than single linkage hierarchical method for our data. In single linkage, we see a single point is mistaken for a cluster, while two separate clusters are grouped into one for both HE0 and HE1. It occurs because both HE0 and HE1 data contain a point that is far from its nearest cluster (, or rather, outlier), and the distance between this point and the cluster is larger than the smallest distance between two clusters.

3. In complete linkage method, however, the outlier will have lesser effect on clustering. With the distance defined as the maximum pairwise distance, as long as the distance between the outlier and its nearest cluster does not exceed the minimum distance between two clusters, the outlier will not make a change.

- e. Compare your results for different clustering methods and interpret them. Select the results from one of the clusters for the following analyses. Pay close attention to the generated clusters when choosing which results to use.

- The K-Means, Gaussian Mixture and Complete Linkage Hierarchical methods all generate similar clustering results. Since the shapes of the clusters are circles, what K-Means algorithm does intuitively is place a circle at the center of each cluster.

- Gaussian Mixture involves the mixture of multiple Gaussian models. It applies the EM algorithm to find the maximum likelihood estimates for the parameters of the models. A point will be assigned to a cluster when the probability of it coming from the corresponding Gaussian model is larger than it coming from the other Gaussian models. The centers of the computed Gaussian models will be very close to the centers of the clusters.

- The Complete Linkage Hierarchical method uses the maximum pairwise distance as the distance between clusters. It avoids a drawback of the single linkage method, i.e., the chaining phenomenon where clusters formed via single linkage clustering may be forced together due to single elements being close to each other, even though many of the elements in each cluster may be very distant to each other.

Task 3 – Question 3 (continued)

We select the results based on the following Calinski Harabasz score. K-Means has the best performance for clustering.

Best Calinski Harabasz Score				
Samples	HE0		HE1	
Method	k	Score	k	Score
K-Means	2	1499.5317	3	6821.4362
Gaussian Mixture Model	2	1498.8232	3	6821.3566
Hierarchy Single Linkage	17	114.2425	2	3913.7859
Hierarchy Complete Linkage	2	1498.7063	3	6787.0441

Task 3 – Question 3 (continued)

- f. In context, what do the clusters you have found represent? What are some factors which could account for this type of clustering pattern?
- Answer: The clusters represent the subpopulations with different microbe for the HE0 population (people with cirrhosis who do not have HE) and HE1 population (people with cirrhosis who have developed HE). From the clusters graph, we identify 2 subpopulations for the HE0 population, and 3 subpopulations for the HE1 population.

- g. Based on your process for deciding the number of clusters to partition the data into, what situations or factors might result in your decision being inaccurate?
- Answer: Firstly, we want to utilize within cluster distance as our metric, it turns out that we always encountered the problem of “overfitting” i.e. you will always get the optimal when you every points assign every point to a cluster. Therefore, by considering this situation, we want some kind of trade-off between cluster distance and within cluster distance. Calinski Harabasz score was chosen to be the final metric.

Task 4 – Question 1

- a. Determining which HE1 subpopulations had a significantly different microbiome than the HE0 samples. Explain your decision process and provide evidence supporting your conclusions.

Percentage Difference Between each cluster in HE1 and HE0	
Cluster1	0.02030477872138440 6
Cluster2	0.00324124377635638 98
Cluster3	0.00442822218410953

- For each cluster of the HE1, we compute (in the feature space) the average vector of the data points from each cluster. We also compute the average vector of all the data points from HE0. Then, we compute the percentage difference between each average vector of HE1 and the average vector of HE0. Since the percentage difference is equal to the 2-norm (length of vector) of the difference vector, we use np.linalg.norm to compute it. We make the above table for percentage differences. Since % difference for Cluster 1 is significantly higher than those for Cluster 2 and Cluster 3, we conclude **that Cluster 1 of HE1 has a significantly different microbiome than the HE0 samples.**

Task 4 – Question 1 (continued)

- b. Determining the HE0 subpopulation most similar to each HE1 subpopulation with a significantly different microbiome. Explain the decision process and provide evidence to support your conclusions.

- Answer:
- In order to determine the similarity between a pair of one HE0 subpopulation and one HE1 subpopulation. We decided to apply two-sample t-test on each microbe of each pair. By setting significant level being 0.05, we could determine the number of microbes which have significantly altered abundance. Upon obtaining these number of different pairs respectively, we could just compare them, to decide which pair shares more similarity. The null hypothesis for each test is that: Average abundance of the microbe in two sample are the same. Under the two-sample t-test, The number of microbes have a altered abundance in the pair of group1HE1 and group1HE0 is: 22 and The number of microbes have a altered abundance in the pair of group1HE1 and group2HE0 is: 38. Therefore, we could say that the group1HE1 is more similar to group1HE0

Task 4 – Question 1 (continued)

- c. Microbes with significantly altered abundance based on KS test:
- We conducted a KS test on each of the microbes from Cluster 1 of HE1 and Cluster 1 of HE0 with alpha level = 0.0000025. The results show that 19 microbes have significantly altered abundance. See the following table for the microbe names and their indices.

	Microbes Index	Microbes' Name
0	5	Actinobacteria_Actinobacteria_Actinomycetales_...
1	11	Actinobacteria_Actinobacteria_Actinomycetales_...
2	13	Actinobacteria_Actinobacteria_Actinomycetales_...
3	20	Bacteroidetes_Bacteroidia_Bacteroidales_Bacter...
4	29	Bacteroidetes_Flavobacteriia_Flavobacteriales_...
5	33	Bacteroidetes_Sphingobacteriia_Sphingobacteria...
6	38	Chrysiogenetes_Chrysiogenetes_Chrysiogenales_C...
7	44	Firmicutes_Bacilli_Bacillales_Bacillales_Incer...
8	53	Firmicutes_Bacilli_Lactobacillales_Lactobacill...
9	63	Firmicutes_Clostridia_Clostridiales_Clostridia...
10	78	Firmicutes_Clostridia_Halanaerobiales_Halanaer...
11	82	Firmicutes_Negativicutes_Selenomonadales_Veill...
12	87	Parvarchaeota_Candidatus Parvarchaeum_Candidat...
13	95	Proteobacteria_Alphaproteobacteria_Rhizobiales...
14	96	Proteobacteria_Alphaproteobacteria_Rhizobiales...
15	99	Proteobacteria_Alphaproteobacteria_Rhizobiales...
16	104	Proteobacteria_Alphaproteobacteria_SAR11_SAR11
17	107	Proteobacteria_Betaproteobacteria_Burkholderia...
18	133	Proteobacteria_Gammaproteobacteria_Orbales_Orb...

Task 4 – Question 2

- a. Which of the microbes that you identified show an increase of relative abundance in the HE1 sample? Do any show a decrease?

Microbes With Increased RA in HE1

	Index	The names of Microbes showing an increase in HE1 sample
0	11	Actinobacteria_Actinobacteria_Actinomycetales_...
1	13	Actinobacteria_Actinobacteria_Actinomycetales_...
2	33	Bacteroidetes_Sphingobacteriia_Sphingobacteria...
3	38	Chrysiogenetes_Chrysiogenetes_Chrysiogenales_C...
4	44	Firmicutes_Bacilli_Bacillales_Bacillales_Incer...
5	78	Firmicutes_Clostridia_Halanaerobiales_Halanaer...
6	87	Parvarchaeota_Candidatus Parvarchaeum_Candidat...
7	95	Proteobacteria_Alphaproteobacteria_Rhizobiales...
8	96	Proteobacteria_Alphaproteobacteria_Rhizobiales...
9	104	Proteobacteria_Alphaproteobacteria_SAR11_SAR11

Microbes With Decreased RA in HE1

	Index	The names of Microbes showing a decrease in HE1 sample
0	5	Actinobacteria_Actinobacteria_Actinomycetales_...
1	20	Bacteroidetes_Bacteroidia_Bacteroidales_Bacter...
2	29	Bacteroidetes_Flavobacteriia_Flavobacteriales_...
3	53	Firmicutes_Bacilli_Lactobacillales_Lactobacill...
4	63	Firmicutes_Clostridia_Clostridiales_Clostridia...
5	82	Firmicutes_Negativicutes_Selenomonadales_Veill...
6	99	Proteobacteria_Alphaproteobacteria_Rhizobiales...
7	107	Proteobacteria_Betaproteobacteria_Burkholderia...
8	133	Proteobacteria_Gammaproteobacteria_Orbales_Orb...

- Taxonomical relationships and groups among microbes with altered abundance:
- Yes, there are. We could conclude from the table in 1.3 by roughly looking at their first word. Therefore, there should be six taxonomical relationships among the microbes with altered abundance. They could be categorized according to the table below.

	Taxonomy
0	Actinobacteria
1	Bacteroidetes
2	Chrysiogenetes
3	Firmicutes
4	Parvarchaeota