Mini-Project 2

ECE/CS 498DS

Spring 2020

Feiyang Li(feiyang3), Yunhan Wang(ywang530), Jingde Chen(jingdec2)

Task 1 - Question 0

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

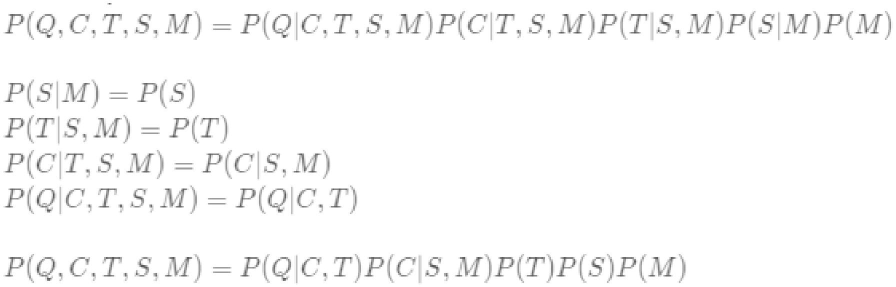
Larger number of samples helps reduce the error caused by the fluctuation in the data. The larger the sample size, the more convincing conclusion we can draw from the data.

2. Number of samples analyzed: A total of 764 samples were analyzed.
3. Number of microbes identified: There were 149 identified microbes.

Task 1 – Question 1

- a. Factorization of joint probability distribution:

Let the variables be denoted as:
S = Storage Temp
M = Collection Method
C = Contamination
T = Lab Time Before Processing
Q = Quality



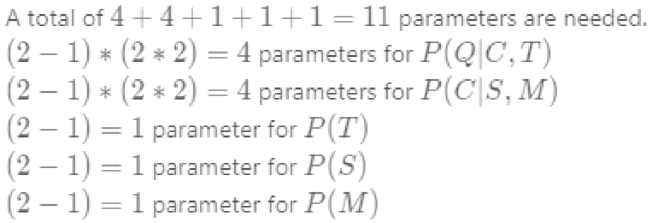
$$P(Q,C,T,S,M) = P(Q|C,T,S,M)P(C|T,S,M)P(T|S,M)P(S|M)P(M)$$

$$P(S|M) = P(S)$$
$$P(T|S,M) = P(T)$$
$$P(C|T,S,M) = P(C|S,M)$$
$$P(Q|C,T,S,M) = P(Q|C,T)$$

$$P(Q,C,T,S,M) = P(Q|C,T)P(C|S,M)P(T)P(S)P(M)$$

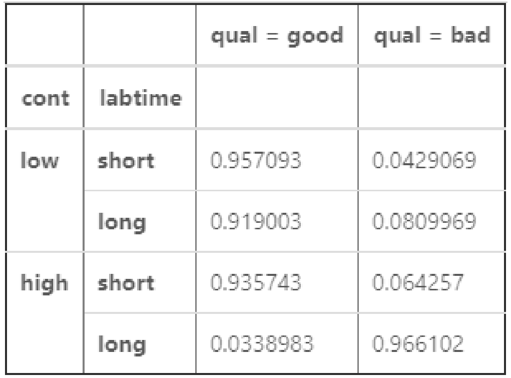
- b. Number of parameters needed to define conditional probability distribution:

11 parameters needed



A total of $4+4+1+1+1=11$ parameters are needed.
$(2-1)*(2*2)=4$ parameters for $P(Q|C,T)$
$(2-1)*(2*2)=4$ parameters for $P(C|S,M)$
$(2-1)=1$ parameter for $P(T)$
$(2-1)=1$ parameter for $P(S)$
$(2-1)=1$ parameter for $P(M)$

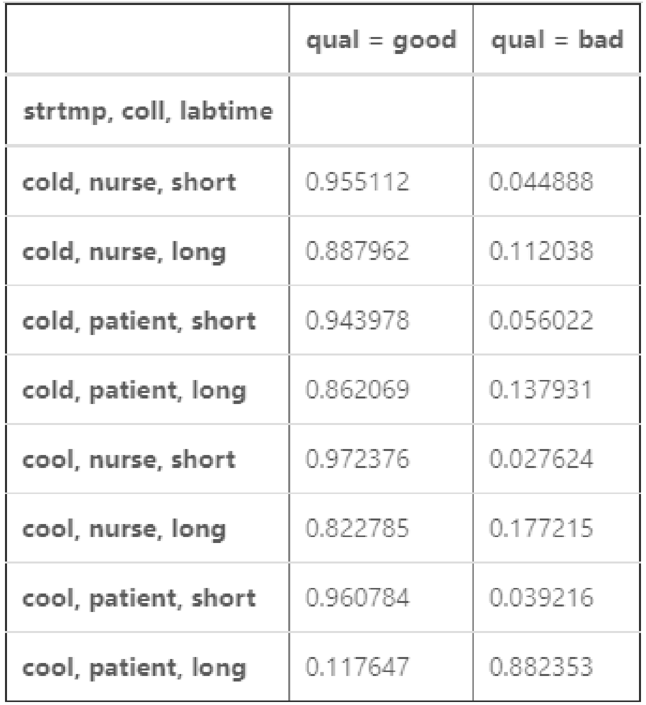
- c. Conditional probability tables:



		qual = good	qual = bad
cont	labtime		
low	short	0.957093	0.0429069
	long	0.919003	0.0809969
high	short	0.935743	0.064257
	long	0.0338983	0.966102

Task 1 – Question 1 (continued)

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



	qual = good	qual = bad
strtmp, coll, labtime		
cold, nurse, short	0.955112	0.044888
cold, nurse, long	0.887962	0.112038
cold, patient, short	0.943978	0.056022
cold, patient, long	0.862069	0.137931
cool, nurse, short	0.972376	0.027624
cool, nurse, long	0.822785	0.177215
cool, patient, short	0.960784	0.039216
cool, patient, long	0.117647	0.882353

- e. Total number of samples dropped:

  130 samples (Sample 699 ~ Sample 763 for both HE0 and HE1)

Task 1 – Question 2

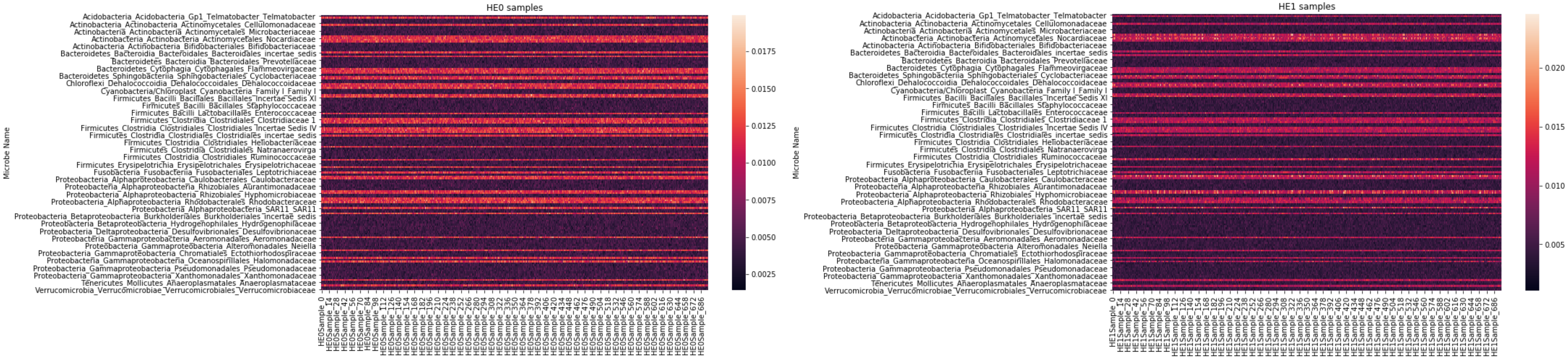
- 1. Number of samples removed: 0

- 2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

  The benefit of using relative abundance is that all data entries we get are on the same scale. But we lose the absolute abundance of the microbes, which is the magnitude of each data.

Task 1 – Question 3

- Heatmaps (HE0 on left HE1 on right):



- Summarize your observations

Since the darker zone refers to low abundance and lighter zone refers to higher abundance, change of color indicates different level of abundance. We can observe that several microbes (rows in the heatmap) have obvious different colors between the two graphs, which indicates altered distribution in HE0 and HE1 samples. Also, from the spectrum we can see that high concentration microbes (lighter color) from HE0 takes higher value (around 0.0125) while high concentration microbes from HE1 takes lower values (around 0.010). We can also observe that the colors on a single row are very similar. This means that the relative abundance level of a single microbe is similar among different samples.

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?

  Heatmaps are good at highlighting the most dense/ most important areas of single metric of our data. Heatmaps highlight the underlying pattern or trend in the data and are suitable for studying the relative magnitude between close areas of data points.

  Heatmap is useful for data that has some sort of trends. However it is less suitable for data that is more random or not having a relative clear trend. For example you can not get any useful information applying heatmap to a white noise signal. Also, Heatmaps are less suitable for studying the relative magnitude of data points that are not adjacent, and it's not great data visualization to look up precise values.

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.

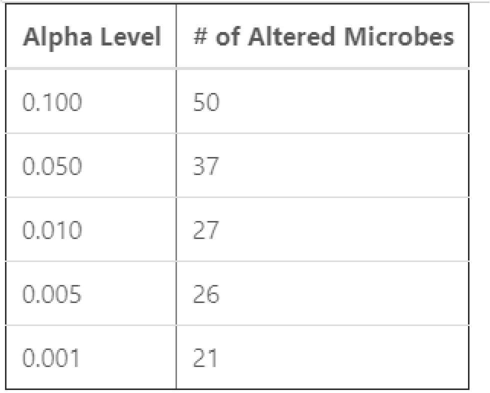
H_0 for the KS Test is that the 2 samples tested are drawn from the same distribution.

Using microbe Acidobacteria_Acidobacteria_Gp3_Gp3_Gp3 as example.

The null hypothesis of the KS test in our context is:

H_0: The relative abundance levels of microbe Acidobacteria_Acidobacteria_Gp3_Gp3_Gp3 from the HE0 samples and the HE1 samples follow the same distribution, which means the microbe's abundance is not altered.

- 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:



Alpha Level	# of Altered Microbes
0.100	50
0.050	37
0.010	27
0.005	26
0.001	21

Task 2 – Question 2

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

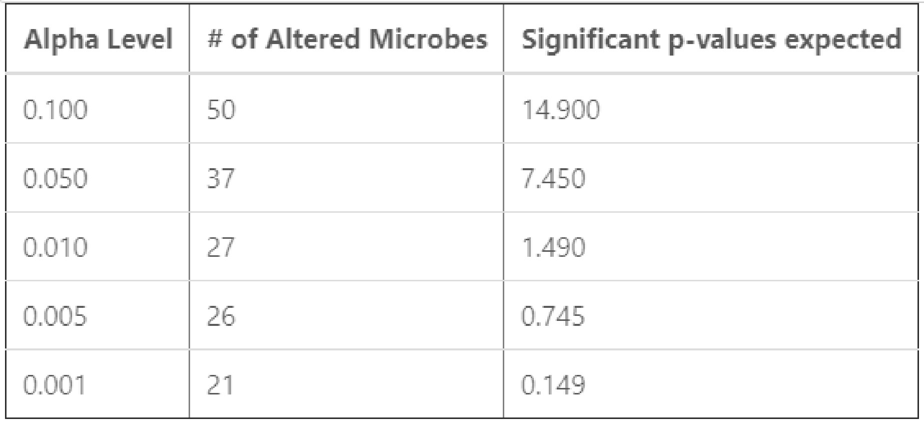
P-value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

In our context, a p-value of 0.05 represents a 5% probability of observing the threshold KS test statistic or a more extreme KS test statistic, given that the microbe's abundance is not altered in the HE0 and HE1 samples.

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

If the null hypothesis is true, then P-value will always be greater than Alpha level. Thus P-value will follow a uniform distribution.

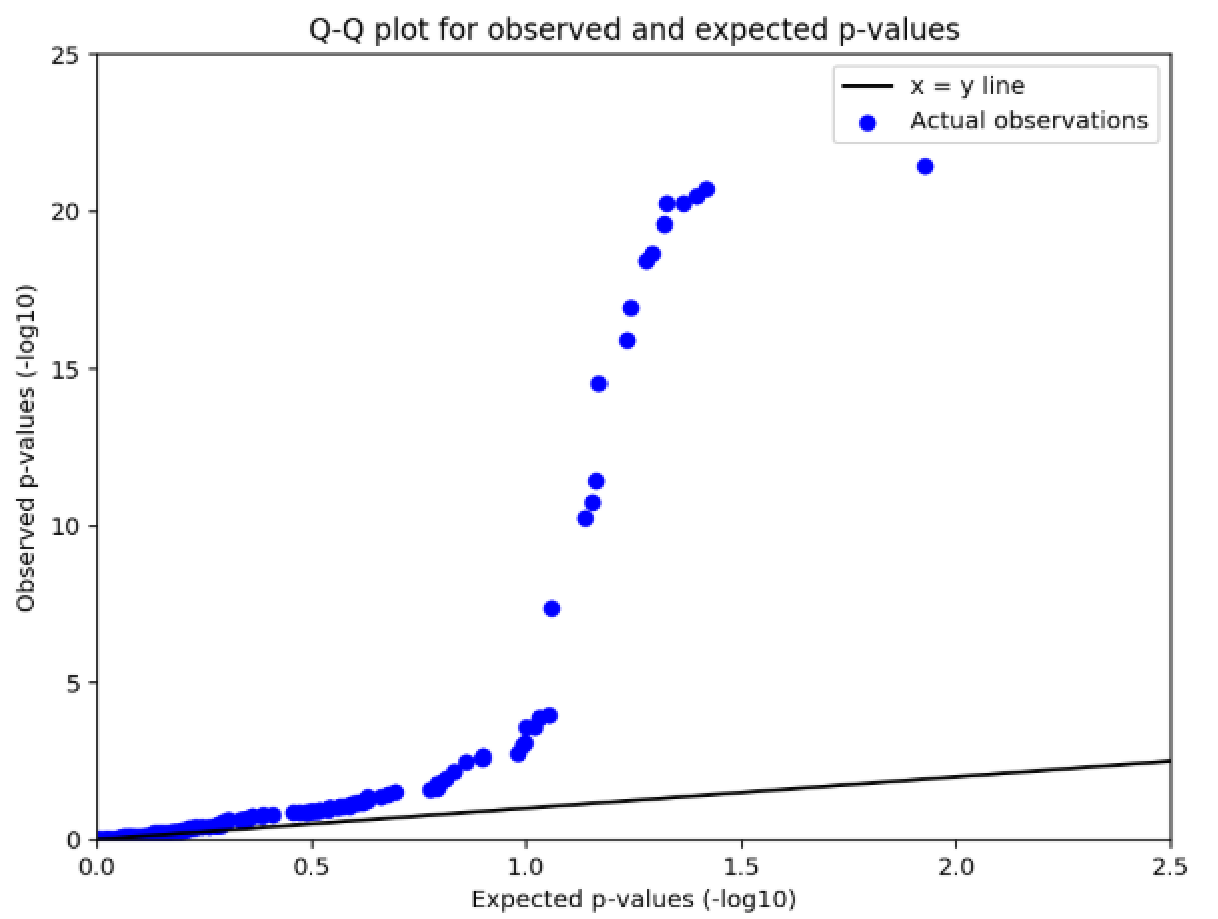
- c. If no microbe's abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.c. Show the comparison in a table below:



Alpha Level	# of Altered Microbes	Significant p-values expected
0.100	50	14.900
0.050	37	7.450
0.010	27	1.490
0.005	26	0.745
0.001	21	0.149

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?

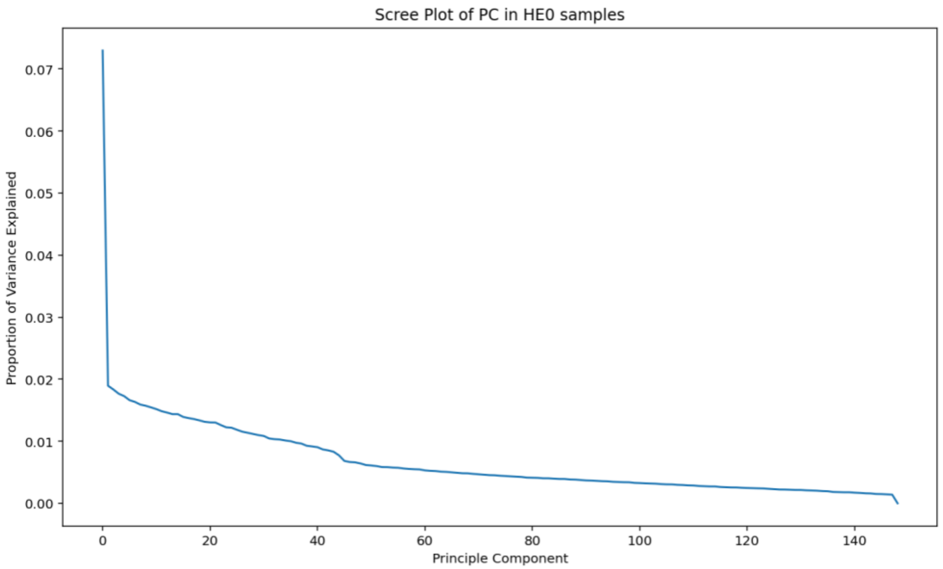
After taking log10(p-values) the ones with the highest values of log10(p-values) are the ones in the tail of the p-value distribution and are more visible.

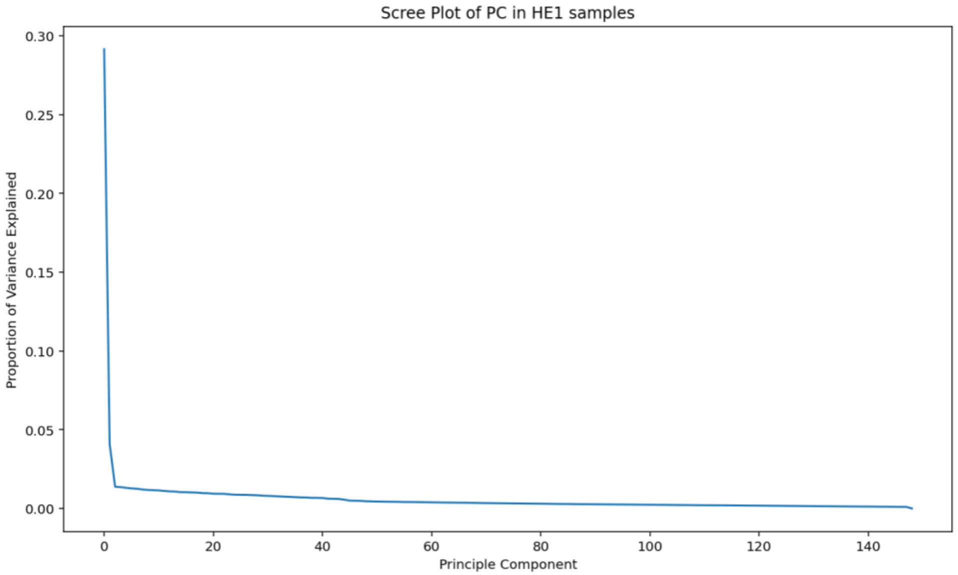
- e.ii. What can you conclude from the Q-Q plot?

The Q-Q plot does NOT align with the x=y, which means that the p-values are not uniformly distributed implying that the null hypothesis is likely not true.

Task 3 – Question 1

- b. Scree plots:





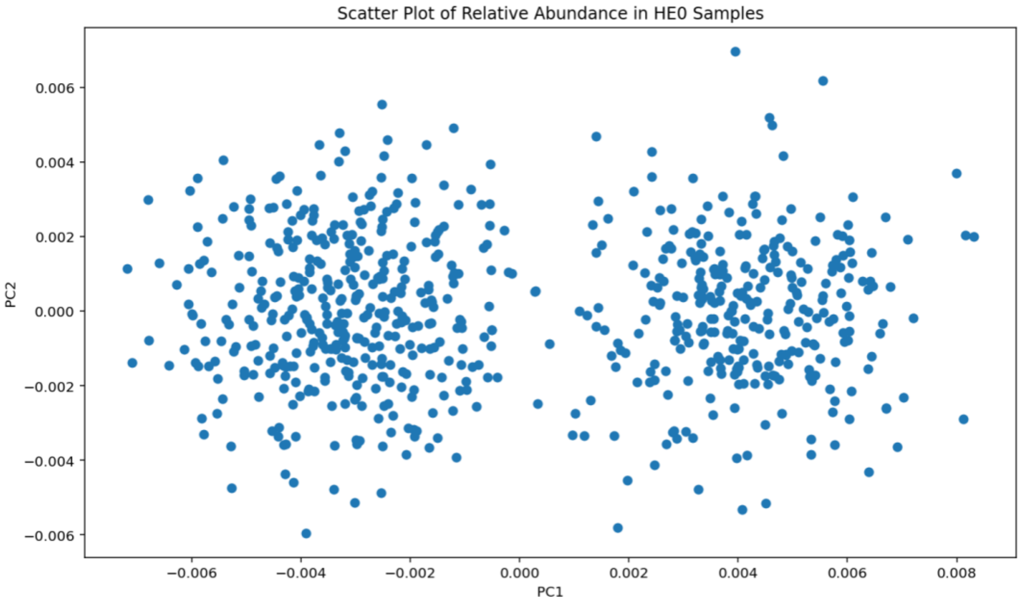
- Number of principal components needed to explain 30% of the total variance (HE0 and HE1):

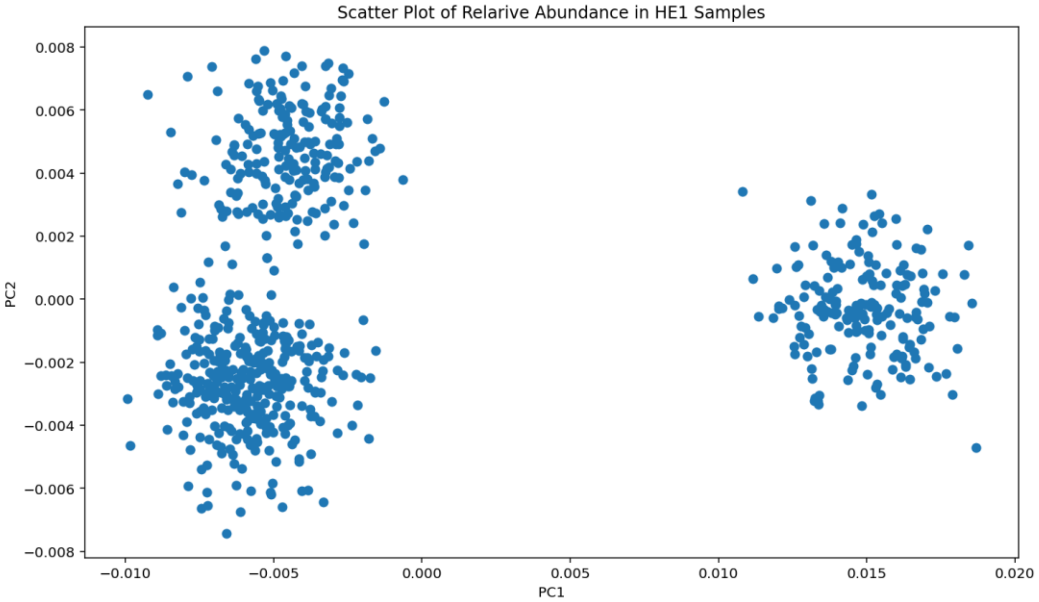
  16 principle components are needed to explain 30% of the total variance in HE0 samples.

  2 principle components are needed to explain 30% of the total variance in HE1 samples.

Task 3 – Question 1 (continued)

- c. Plots:



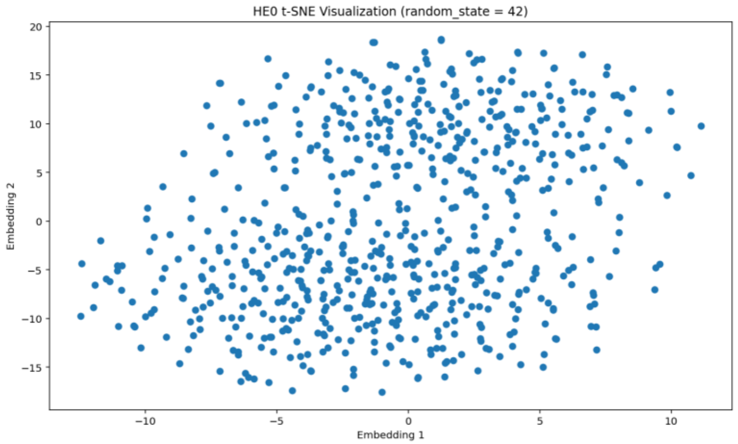


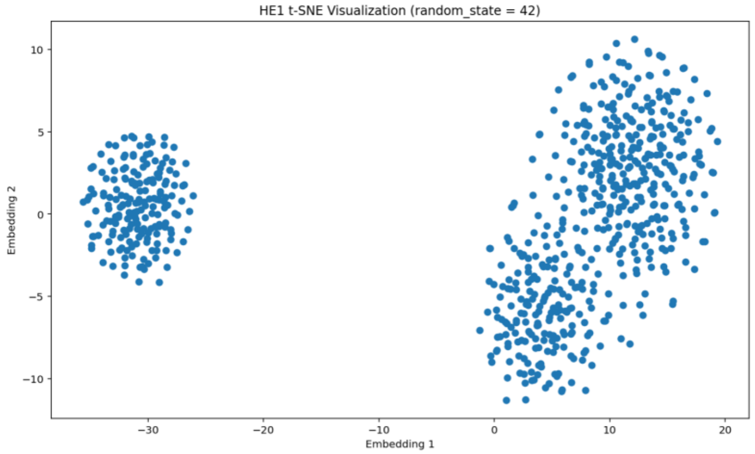
- Observations:

From visual inspection of the 2-D Scatter Plots with the PCA components, we find that there are 2 clusters for HE0 samples, and 3 clusters for HE1 samples, so n_components = 2 for HE0 as well as n_components = 3 for HE1.

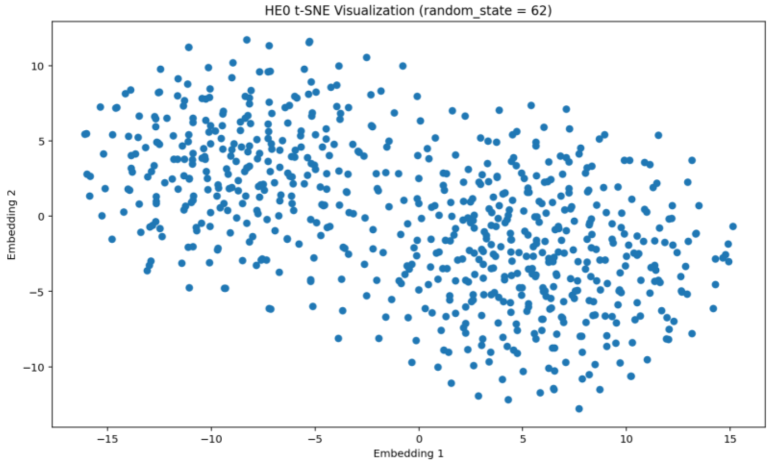
Task 3 – Question 2

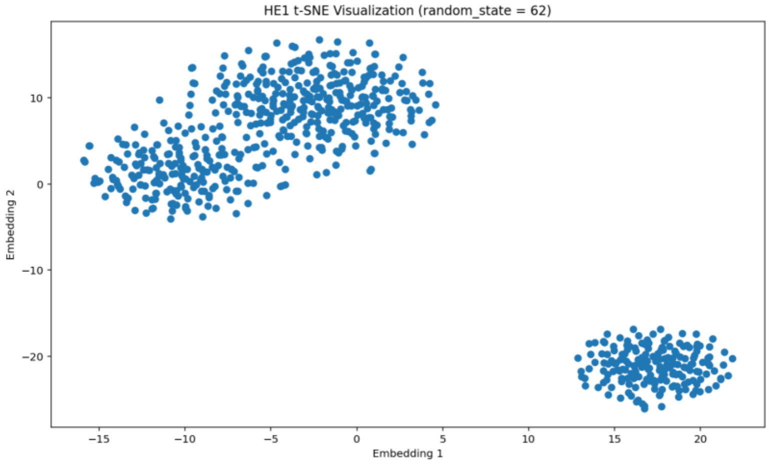
- c. Plots (random_state=42):





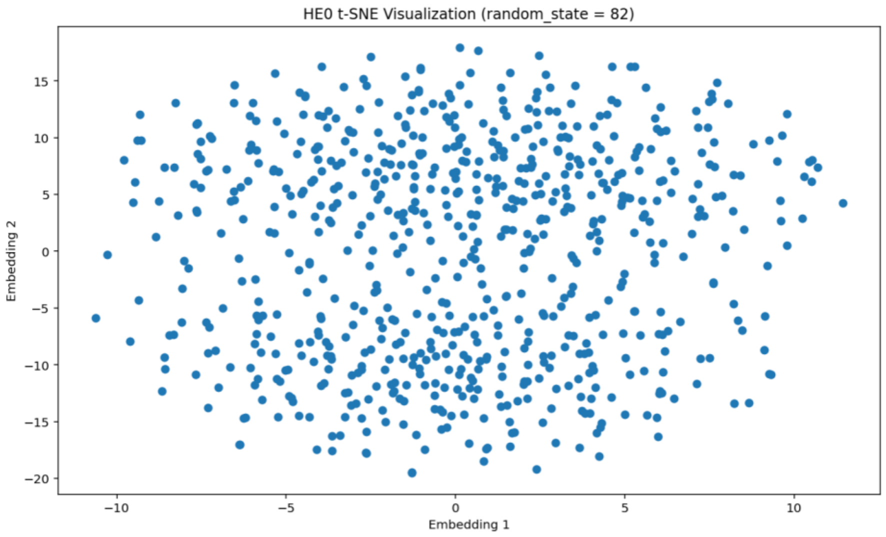
- Plots (random_state=62):

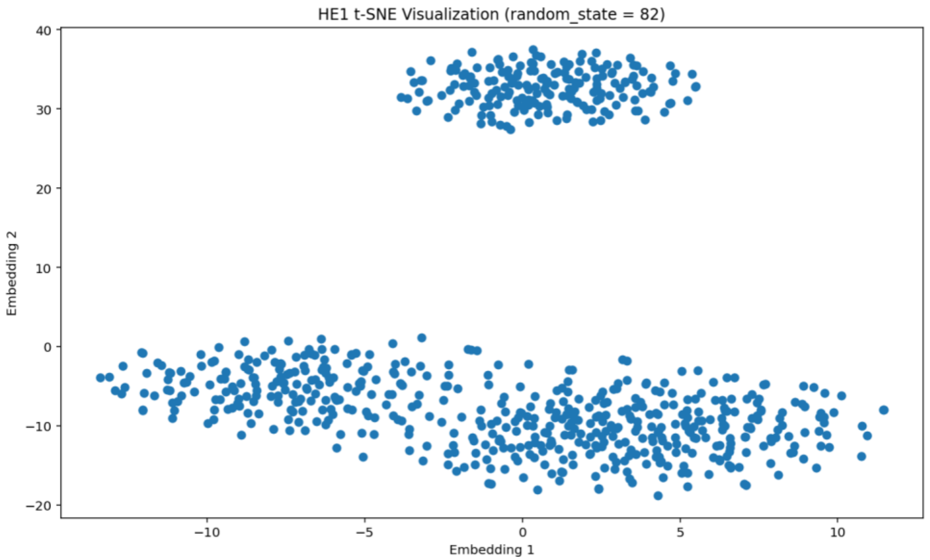




Task 3 – Question 2 (continued)

- c. Plots (random_state=82):





- Observations:

The clusters are more distinguishable and obvious for HE 1 while HE 0 clusters are more adjacent to each other. The clusters are dependent on the random states we selected. Different random states will produce different clusters.

And t-SNE is very slow.

Task 3 – Question 2 (continued)

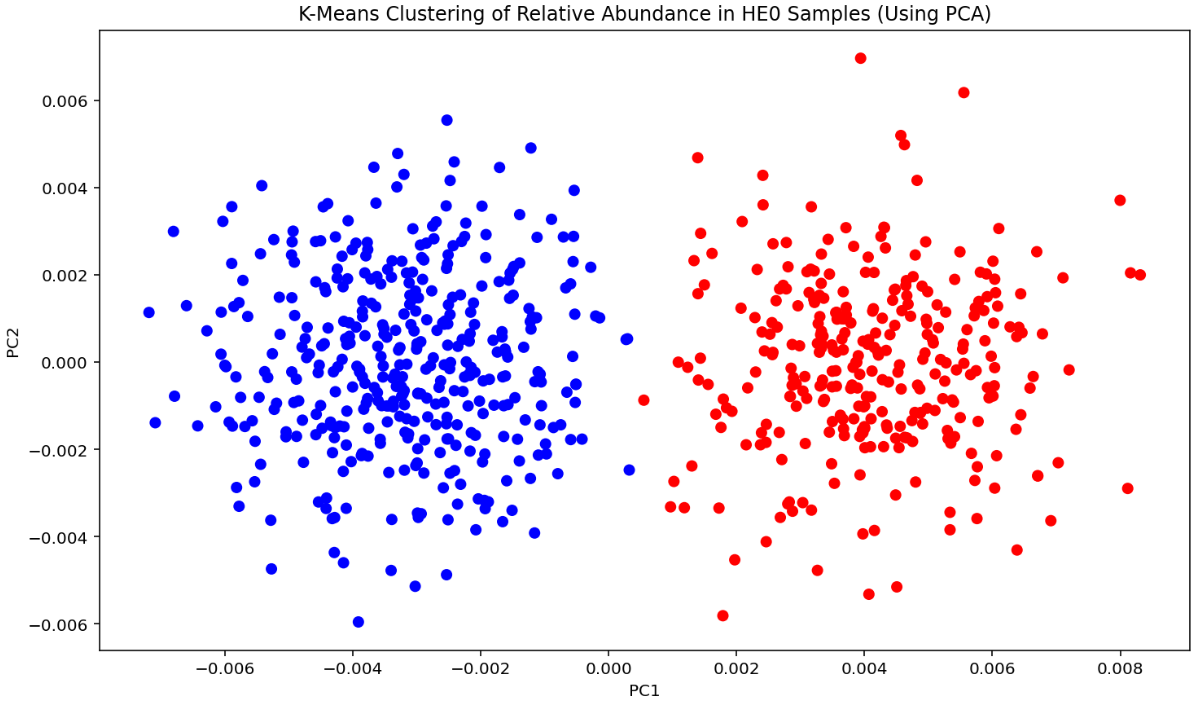
- d. Discussion of similarities and differences between PCA and t-SNE results:

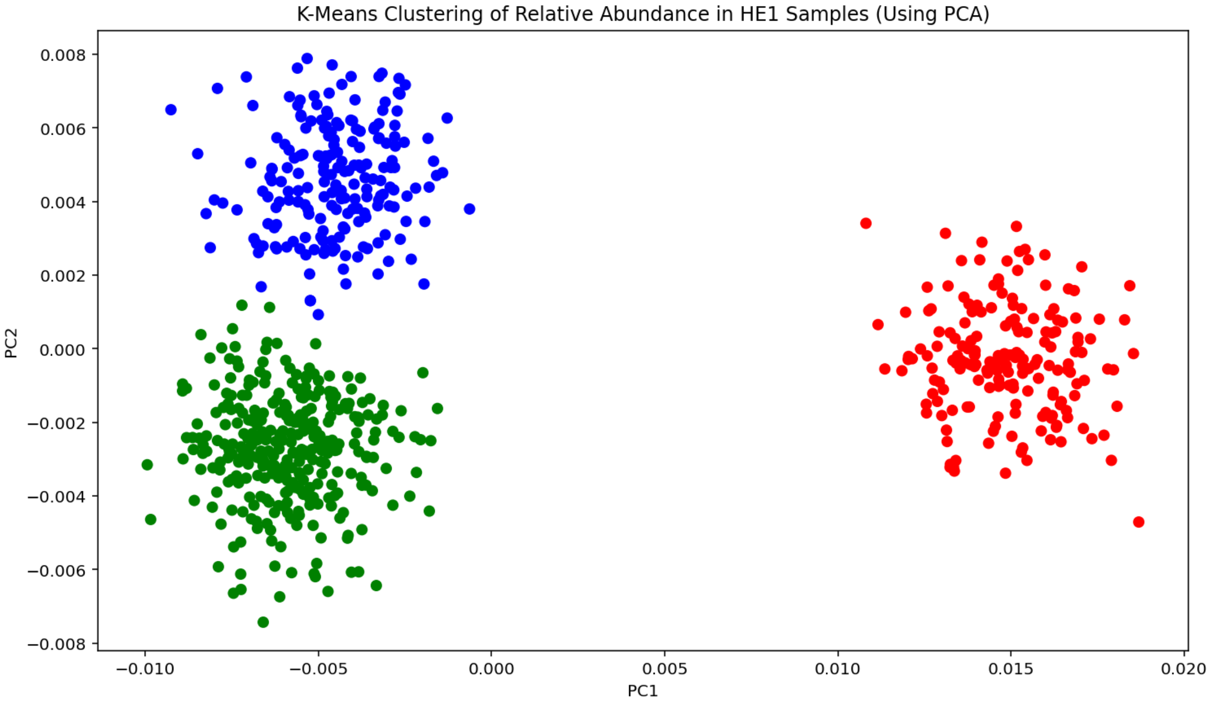
  One key difference is the runtime. The computation time for t-SNE is longer than PCA. And t-SNE is dependent on the random states selected, as different random states will result differently.

  Similarity is both algorithms are used to reduce the dimension of the data while PCA maximizes the variance, t-SNE constructs a probability distribution for the high-dimensional samples in such a way that similar samples have a high likelihood of being picked and dissimilar points have an small likelihood of being picked.

Task 3 – Question 3

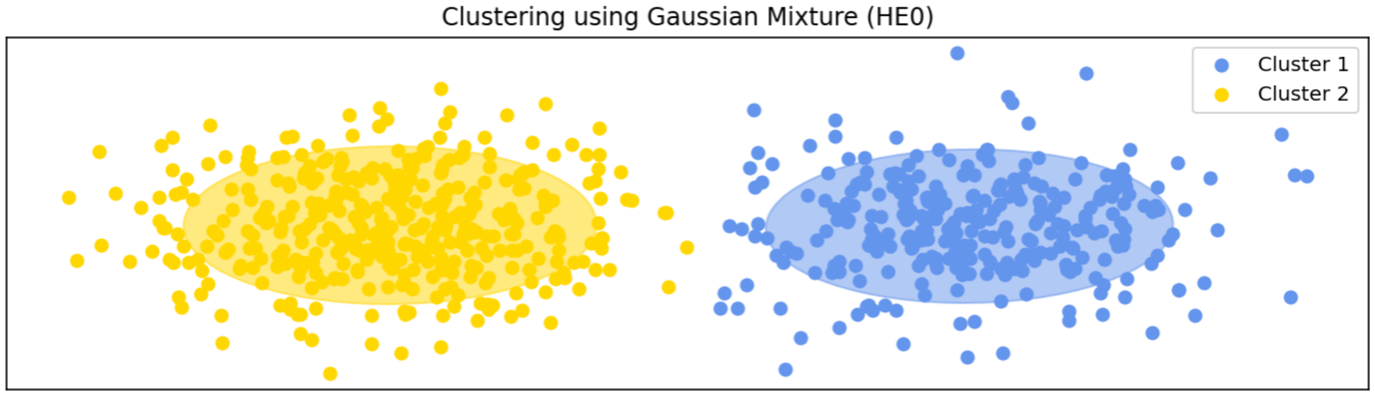
- a. K-means:



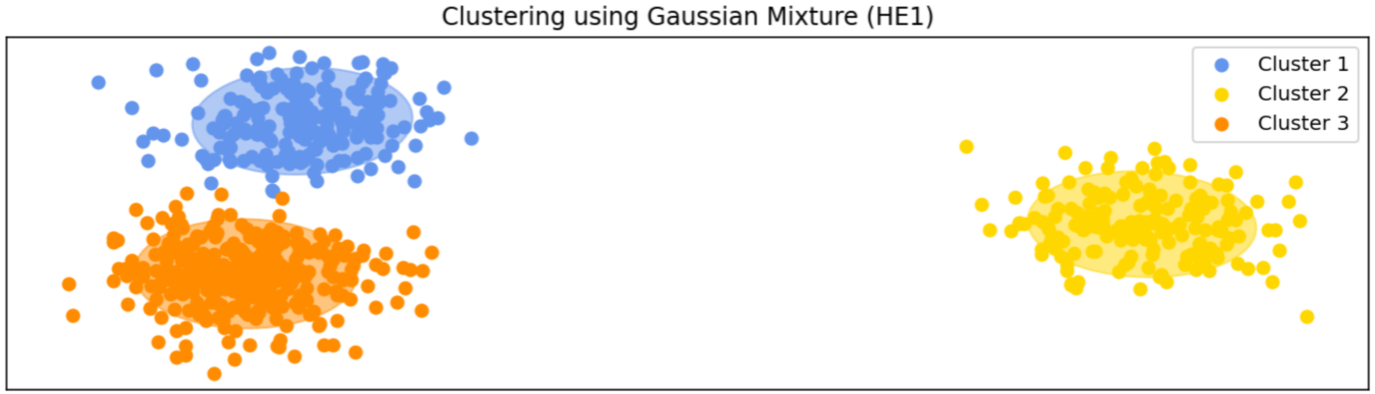


Task 3 – Question 3 (continued)

- b. Gaussian mixture model:



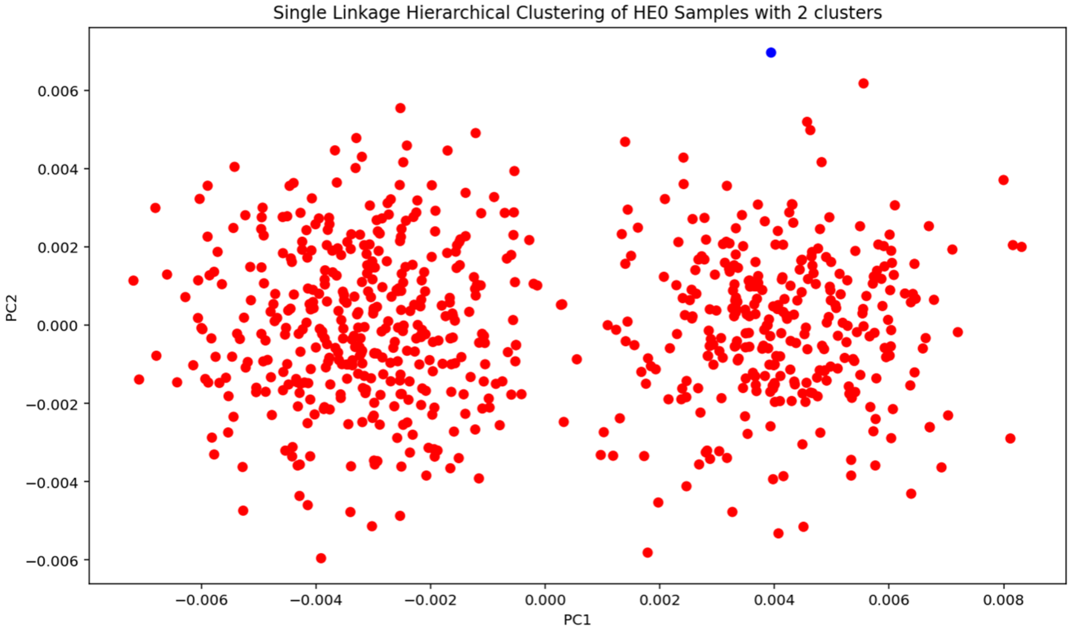
Clustering using Gaussian Mixture (HE0)

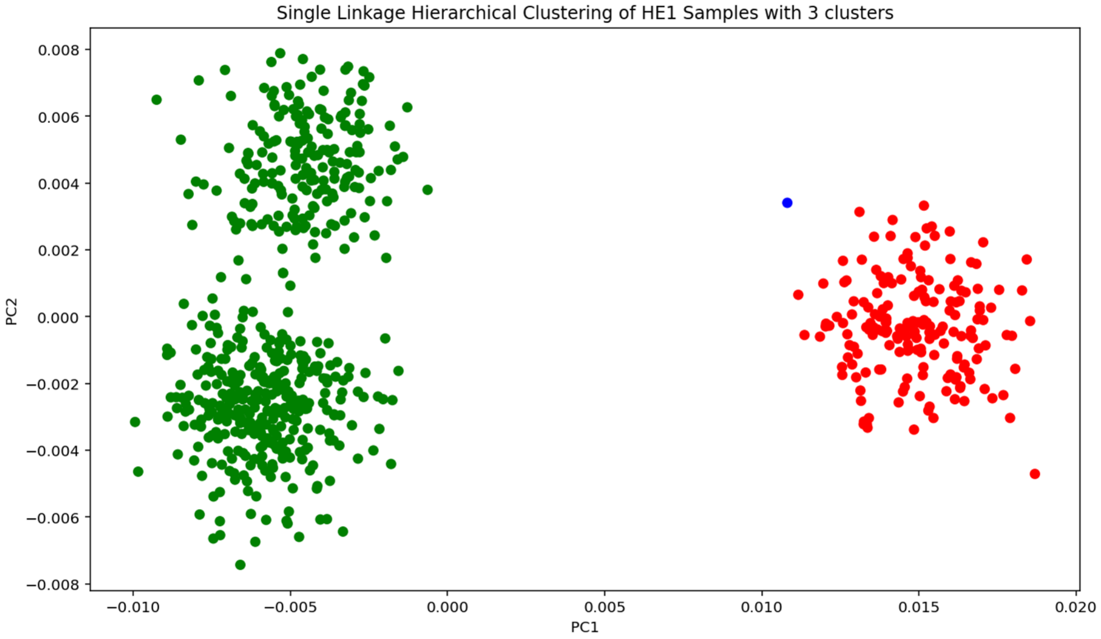


Clustering using Gaussian Mixture (HE1)

Task 3 – Question 3 (continued)

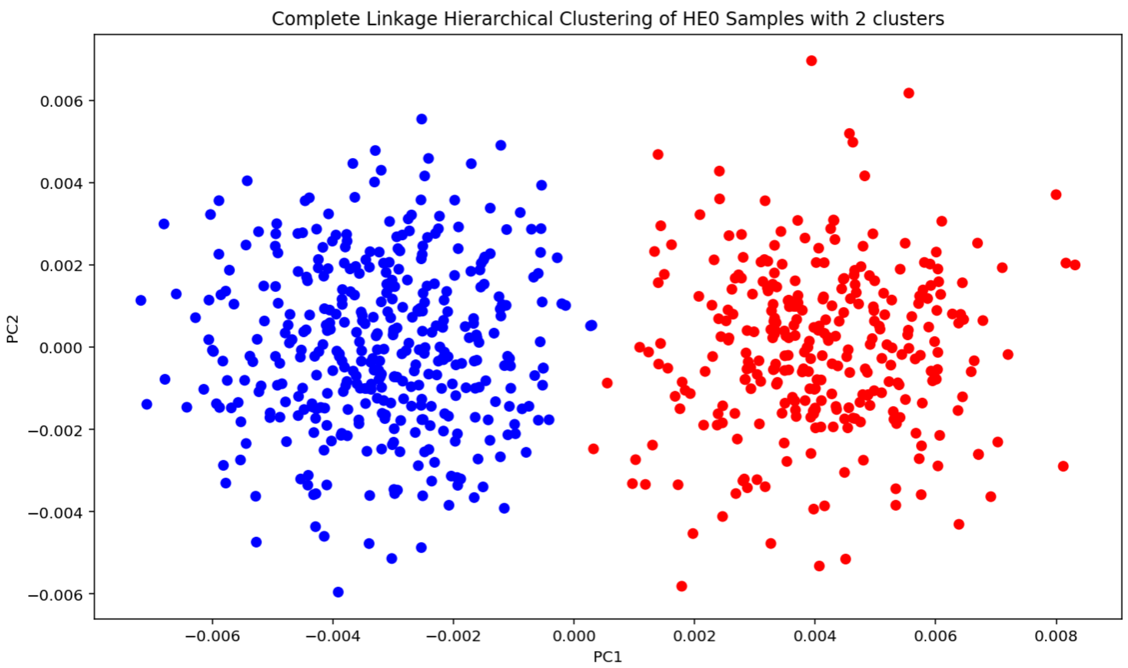
- c. single linkage hierarchical:

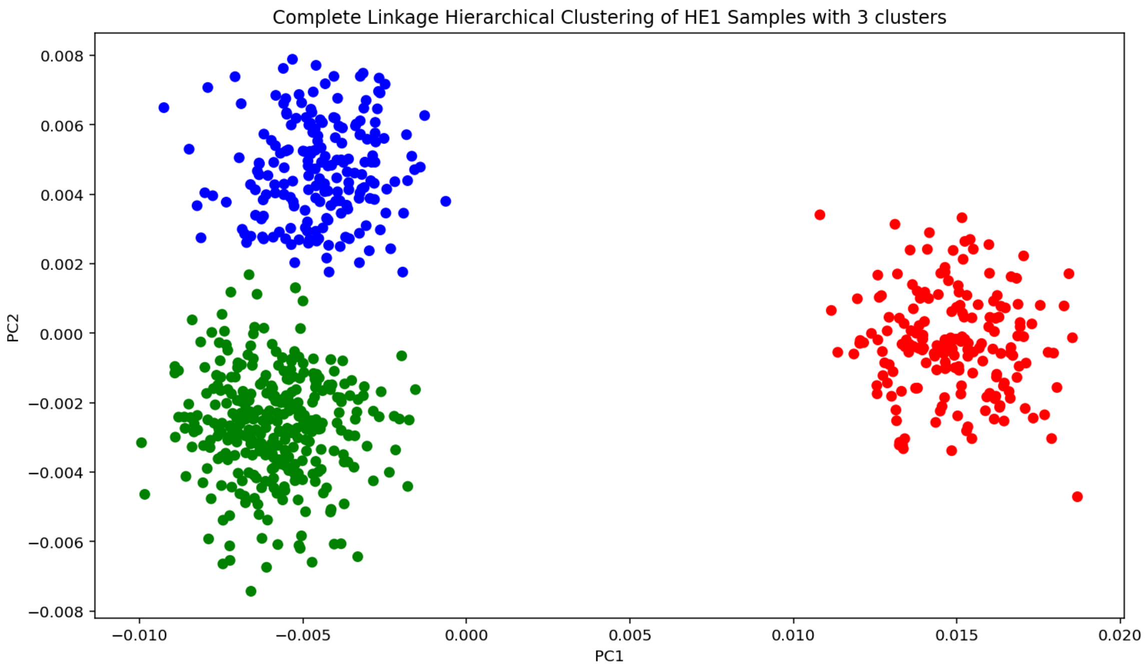




Task 3 – Question 3 (continued)

- c. complete linkage hierarchical:





Task 3 – Question 3 (continued)

- d. Discussion on single vs. complete linkage hierarchical methods:
  Single Link: HE0: There are 2 clusters, but one cluster is formed with only one data point. The reason for this weird clustering is that the data for HE0 has many outliers. Single link is not good at clustering datasets with outliers. HE1: There are 3 clusters, but one cluster is formed with only one data point. The reason for this weird clustering is that the data for HE0 has many outliers. Single link is not good at clustering datasets with outliers.

  Complete Link: HE0: There are 2 well partitioned clusters. Complete link is robust to outliers, and thus it could cluster such a dataset with outliers. HE1: There are 3 well partitioned clusters. Complete link is robust to outliers, and thus it could cluster such a dataset with outliers.

  We observe the common problem of single link clustering being too sensitive to outliers. The outlier in both cases (HE0 and HE1) forms its own cluster (the blue dot). #The complete like clustering performed better than the single link clustering in such a dataset with outliers. Single linkage hierarchical clustering maximizes the minimum distance between data points in the clusters while complete linkage clustering maximizes the maximum distance between data points in the clusters.

  Single linkage and complete linkage clusterings have different behavior on our dataset. Complete linkage method was able to identify the clusters correctly but single linkage method failed to do so. Single linkage method labelled a data point that was far from other points in the cluster as a single cluster and wasn't able to separate clusters with points that are close. The noise and outliers in our data caused the single linkage method to behave differently.

- e. Interpretation and comparison of the different methods:
  We can observe that K-Means, GMM and Complete-link clustering performed well and had similar assignments of the data points, clearly forming 2 clusters for HE0 and 3 clusters for HE1. Single-link clustering assigned single-point clusters and thus is not prefered.

  Since the clustering groups of K-Means, GMM and Complete-link clustering are very similar (expect a few points lying at inter-cluster boundaries), we choose K-means clustering because it's easier to compute.

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?

  In context, the clusters represent groups of samples within HE0 or HE1 which are "closely related" based on a linear combination of relative abundance of microbio. The clusters here represent patterns which are observed after performing PCA on a large dataset, and different PCs cover different linear combination of microbiome along which the variance of the data is maximum.

  One possible factor is some other undiscovered condition that follows with liver cirrhosis just like HE.

- 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?

  We picked the number of clusters based on the 2-d visualization of our data points. We are visualizing the data points only with the first 2 PCs. For HE1 dataset, this explains more than 30% variance in the data and the decision made on visualization may be accurate enough. However in the HE0 dataset, using the first 2 PCs only explains about 10% of the total variance and thus partitioning data into 2 clusters might inaccurate. Also, if the clusters are being formed based on external factors after the sample was being taken, the clusters may not be accurate as well.

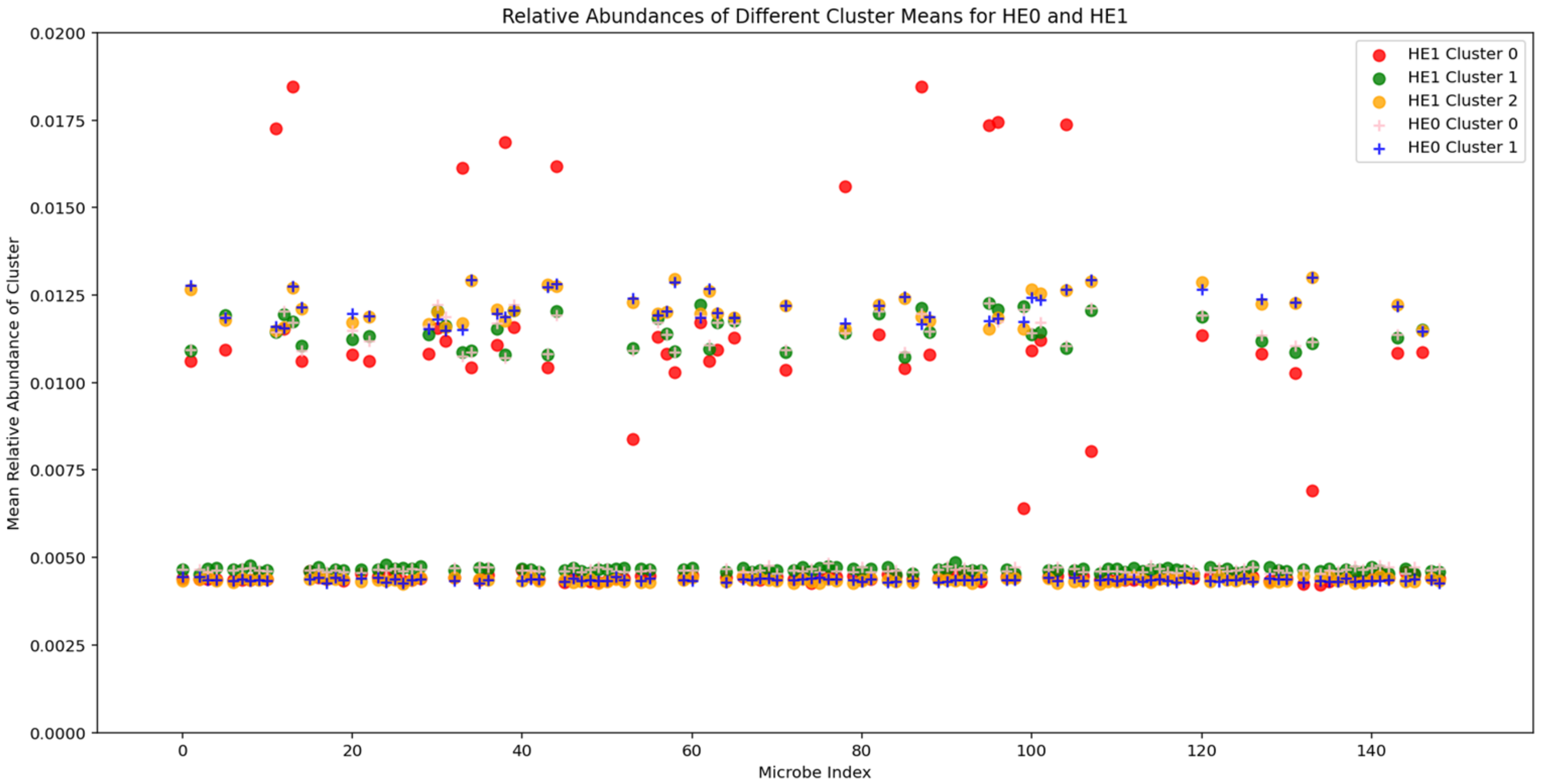
Task 4 – Question 1

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

Analysis Method:
We are comparing the three clusters formed in HE1 with the two clusters formed in HE0. Since it is difficult to compare each individual point in a cluster we use the mean of the points in the cluster to represent each subpopulation.
Step 1: Use the clustering result from the K-Means clustering to identify the subpopulations. For each clustered subpopulation in both HE0 and HE1, calculate the mean of the samples in the subpopulation for all 149 microbes.
Step 2: Visualize the clusters and compare the relative abundances of the cluster means.
Step 3: For each HE1 cluster, observe if any of its microbe abundance levels is significantly different than all of the HE0 clusters. Identify the subpopulations with significantly altered microbes abundances.

For each cluster in HE1, check to see if it has microbes with relative abundance level significantly different than any of the HE0 clusters. If the abundances vary by 10% we say they are significantly different.

The analysis shows that cluster 0 in HE1 samples has significantly altered microbe abundance, as shown in the figure below.



Plotting the cluster means of each microbe gives the graph above.

From the plot we can tell that:

HE1 Cluster 1 (green circle) maps to HE0 Cluster 0 (pink cross)

HE1 Cluster 2 (orange circle) maps to HE0 Cluster 1 (blue cross)

HE1 Cluster 0 (red circle) has microbes with significantly altered abundances that lie far from the HE0 clusters

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.

To see which HE0 cluster HE1 Cluster0 is more similar to, we calculate the distances using the sum of squared error of the abundance means between it and the two HE0 clusters.

```
d(HE1 Cluster 0, HE0 Cluster 0) = 0.0004113865419626202
d(HE1 Cluster 0, HE0 Cluster 1) = 0.0004364532630837121
```

Comparing the distances between HE1 cluster 0 and the two clusters in HE0, we see that HE1 Cluster 0 (red circle) with significantly altered microbes is more similar to HE0 Cluster 0 (pink cross).

Task 4 – Question 1 (continued)

- c. Microbes with significantly altered abundance based on KS test:

```
The following are identified microbes with significantly altered abundance (total 19 microbes):
Actinobacteria_Actinobacteria_Actinomycetales_Corynebacteriaceae
Actinobacteria_Actinobacteria_Actinomycetales_Nakamurellaceae
Actinobacteria_Actinobacteria_Actinomycetales_Propionibacteriaceae
Bacteroidetes_Bacteroidia_Bacteroidales_Bacteroidales_incertae_sedis
Bacteroidetes_Flavobacteriia_Flavobacteriales_Cryomorphaceae
Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae
Chrysiogenetes_Chrysiogenetes_Chrysiogenales_Chrysiogenaceae
Firmicutes_Bacilli_Bacillales_Bacillales_Incertae Sedis XI
Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae
Firmicutes_Clostridia_Clostridiales_Clostridiales_Incertae Sedis XIII
Firmicutes_Clostridia_Halanaerobiales_Halanaerobiaceae
Firmicutes_Negativicutes_Selenomonadales_Veillonellaceae
Parvarchaeota_Candidatus Parvarchaeum_Candidatus Parvarchaeum_Candidatus Parvarchaeum
Proteobacteria_Alphaproteobacteria_Rhizobiales_Brucellaceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae
Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae
Proteobacteria_Alphaproteobacteria_SAR11_SAR11
Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae
Proteobacteria_Gammaproteobacteria_Orbales_Orbaceae
```

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?

  The microbes that show an increase of relative abundance in the HE1 sample are (total 10):

  - Actinobacteria_Actinobacteria_Actinomycetales_Nakamurellaceae
  - Actinobacteria_Actinobacteria_Actinomycetales_Propionibacteriaceae
  - Bacteroidetes_Sphingobacteriia_Sphingobacteriales_Sphingobacteriaceae
  - Chrysiogenetes_Chrysiogenetes_Chrysiogenales_Chrysiogenaceae
  - Firmicutes_Bacilli_Bacillales_Bacillales_Incertae Sedis XI
  - Firmicutes_Clostridia_Halanaerobiales_Halanaerobiaceae
  - Parvarchaeota_Candidatus Parvarchaeum_Candidatus Parvarchaeum_Candidatus Parvarchaeum
  - Proteobacteria_Alphaproteobacteria_Rhizobiales_Brucellaceae
  - Proteobacteria_Alphaproteobacteria_Rhizobiales_Hyphomicrobiaceae
  - Proteobacteria_Alphaproteobacteria_SAR11_SAR11

  The microbes that show an decrease of relative abundance in the HE1 sample are (total 9):

  - Actinobacteria_Actinobacteria_Actinomycetales_Corynebacteriaceae
  - Bacteroidetes_Bacteroidia_Bacteroidales_Bacteroidales_incertae_sedis
  - Bacteroidetes_Flavobacteriia_Flavobacteriales_Cryomorphaceae
  - Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae
  - Firmicutes_Clostridia_Clostridiales_Clostridiales_Incertae Sedis XIII
  - Firmicutes_Negativicutes_Selenomonadales_Veillonellaceae
  - Proteobacteria_Alphaproteobacteria_Rhizobiales_Rhizobiaceae
  - Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae
  - Proteobacteria_Gammaproteobacteria_Orbales_Orbaceae

- Taxonomical relationships and groups among microbes with altered abundance:

  These 19 families of altered microbes come from 6 phylums, 12 classes, 15 orders. 3 are from the Actinobacteria phylum, 3 are from the Bacteroidetes phylum, 1 is from the Chrysiogenetes, 5 are from the Firmicutes phylum, 1 is from the Parvarchaeota phylum, 6 are from the Proteobacteria phylum.