236523 - Introduction to Bioinformatics - Spring 21 - HW2

Alon Hacohen & Roy Bernea

5/27/2021

Roy Bernea - 316315829 ; Alon Hacohen - 311587653 # Question 2

Following is the code and answers for the questions under the “Question 2” Section - as instructed in the assignment:

### Q2.1

Why the mapped reads are mapped preferentially to exons? # TODO - write better The mapped reads are mapped preferentially to exons since during the process of splicing the introns are being removed from the RNA. In the final product (the RNA molecule) there are only exons, therefore the read are mapped to the exons.

### Q2.2

Sometimes if we map RNA-seq reads of certain tissue samples and we look on certain genes we will see that reads are also mapped to the intronic region. What can be biological cause/reason for such an observation? # TODO A biological reason for reads being mapped to the intronic region during the RNA-seq process is that #EDEN’s that when the method is polyA-selection and we also have non-coding RNA in the mRNA then we can see reads that are mapped to the intronic region.

### Q2.3

### Q2.3.1

Additional names (aliases) for the gene *MARCH7* Are: Membrane Associated Ring-CH-Type Finger Axotrophin MARCH-VII RNF177 Membrane-Associated Ring Finger (C3HC4) 7, E3 Ubiquitin Protein Ligase RING-Type E3 Ubiquitin Transferase MARCHF7 Membrane-Associated RING Finger Protein 7 Membrane-Associated RING-CH Protein VII E3 Ubiquitin-Protein Ligase MARCHF7 RING Finger Protein 177 MARCH7 AXOT Membrane-Associated Ring Finger (C3HC4) RING-Type E3 Ubiquitin Transferase MARCH7 E3 Ubiquitin-Protein Ligase MARCH7 Membrane Associated Ring Finger 7 EC 2.3.2.27 EC 6.3.2 MARCHF7 AXO

These gene names were extracted from the following databases: HGNC - the HUGO Gene Nomenclature Committee database Ensembl genome database UniProt database NCBI’s Entrez molecular sequence database OMIM - Online Mendelian Inheritance in Man database

### Q2.3.2

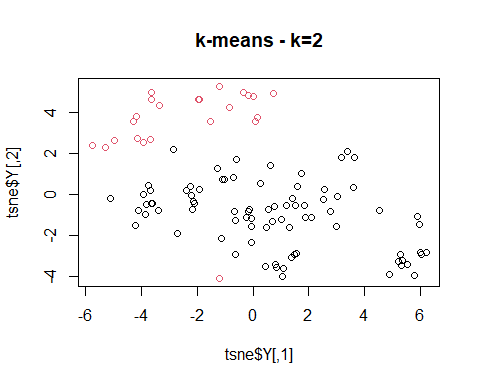
If we paste the name of this gene manually into an Excel sheet, meaning we paste “MARCH7” into a cell inside the sheet, a possible problem we might encounter is that Excel would recognize the name as a date - March 7 - and refer to it as one. In such case, instead of referring to the gene’s name “MARCH7” as its ID (or as a unique identifier), Excel would mistakenly refer to it as a date - which might cause us problem later on in our data analysis process (e.g. if we want to compare the gene in our excel sheet to its appearance in a database - but R would recognize the identifiers as two different ones because of their type etc.).

# Question 3

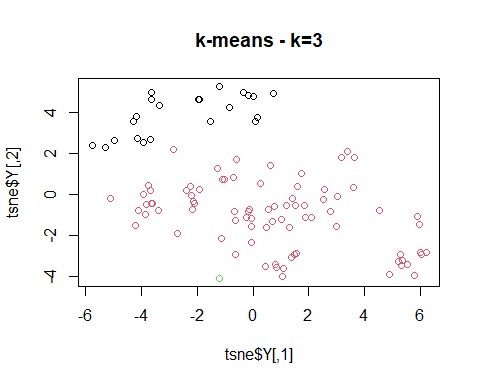
Following is the code and answers for the questions under the “Question 3” Section - as instructed in the assignment:

### Q3.1

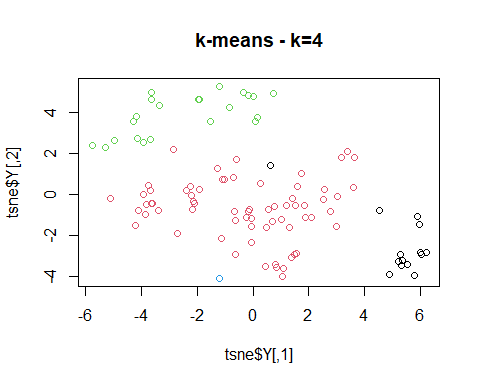
library(ggplot2)  
library(RColorBrewer)  
library(factoextra)  
library(tidyverse)  
library(Rtsne)  
library(dbscan)  
  
# data preperation  
#change to relative path?  
sdy <- readRDS("C:/CS\_Technion/236523-Intro\_to\_Bioinformatics/Spr21/Intro\_to\_BioInf-Spr21-HW2/sdy420.rds")  
#print(t(sdy$expr))  
data <- na.omit(t(sdy$expr))  
data.scaled <- scale(data)  
#view(data.scaled)  
#dist.mat <- dist(data)  
  
# preparing tsne for later  
tsne = Rtsne(data.scaled)  
  
# Clustering methods  
  
set.seed(555)  
  
# kmeans  
km.result <- kmeans(data.scaled, centers = 3, nstart = 25)  
# hierarchical clustering  
hc.result <- hclust(dist(data.scaled), "ward.D")  
# DBSCAN  
dbscan.result <- dbscan(tsne$Y, eps = 1, minPts = 4)  
  
#dbscan.result  
  
  
  
  
#Q3.2+3.3  
  
# Plotting the clusters  
  
# kmeans  
#k=2  
km.result <- kmeans(data.scaled, centers = 2, nstart = 25)  
plot(tsne$Y, col=km.result$cluster, main = "k-means - k=2")



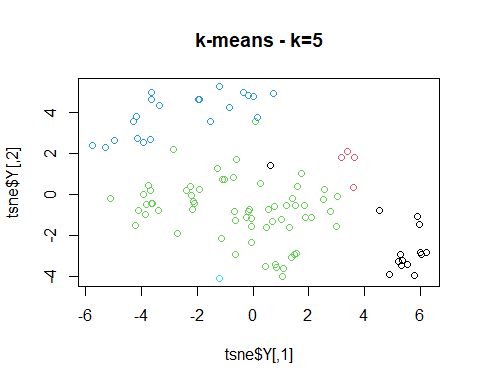
#k=3  
km.result <- kmeans(data.scaled, centers = 3, nstart = 25)  
plot(tsne$Y, col=km.result$cluster, main = "k-means - k=3")



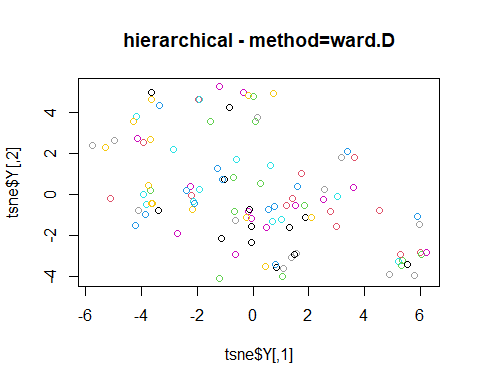
#k=4  
km.result <- kmeans(data.scaled, centers = 4, nstart = 25)  
plot(tsne$Y, col=km.result$cluster, main = "k-means - k=4")



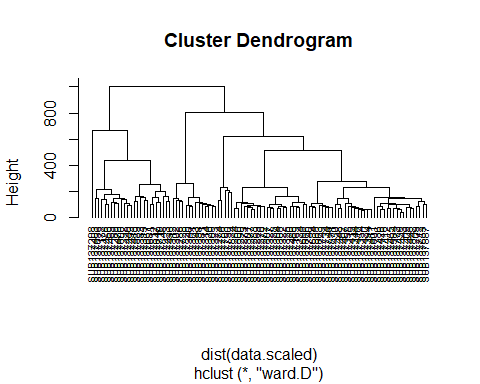
#k=5  
km.result <- kmeans(data.scaled, centers = 5, nstart = 25)  
plot(tsne$Y, col=km.result$cluster, main = "k-means - k=5")



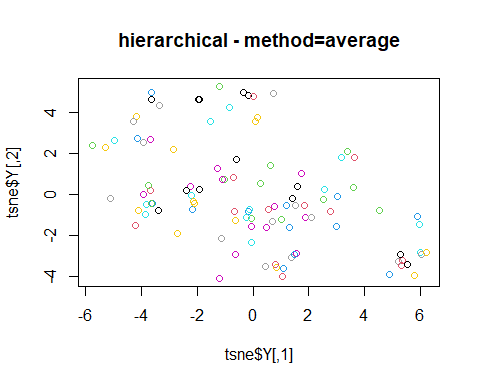
# hierarchical clustering  
#method="ward.D"  
hc.result <- hclust(dist(data.scaled), "ward.D")  
plot(tsne$Y, col=hc.result$order, main = "hierarchical - method=ward.D")



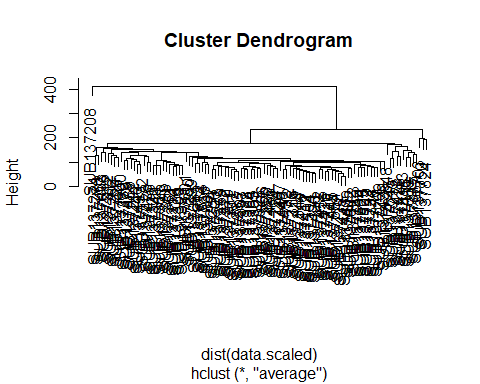
plot(hc.result, cex = 0.6, hang = -1)



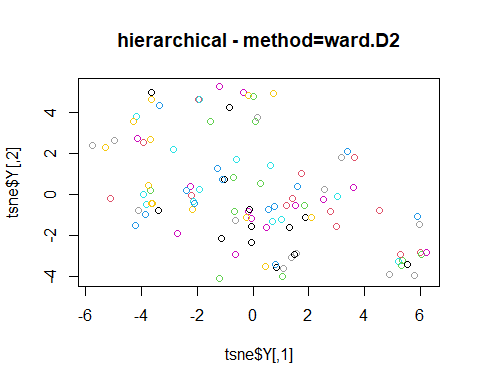
#method="ave"  
hc.result <- hclust(dist(data.scaled), "ave")  
plot(tsne$Y, col=hc.result$order, main = "hierarchical - method=average")



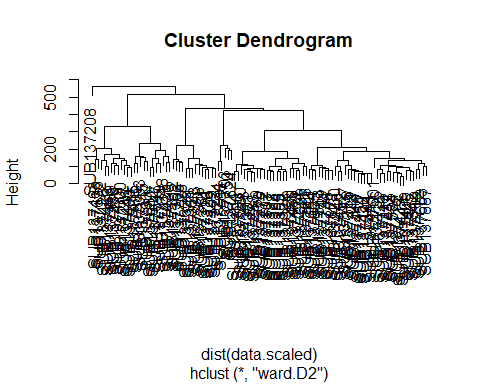
plot(hc.result)



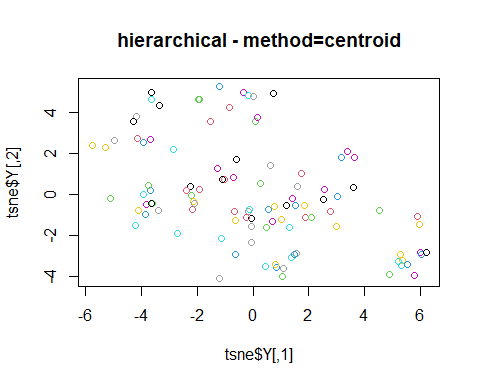
#method="ward.D2"  
hc.result <- hclust(dist(data.scaled), "ward.D2")  
plot(tsne$Y, col=hc.result$order, main = "hierarchical - method=ward.D2")



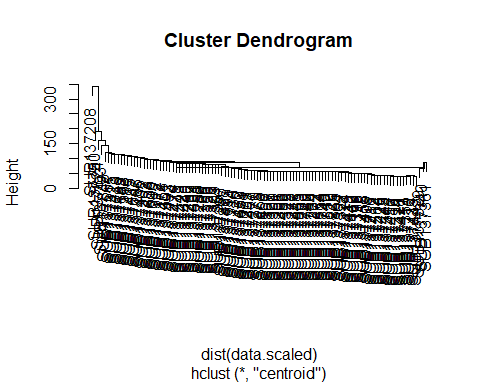
plot(hc.result)



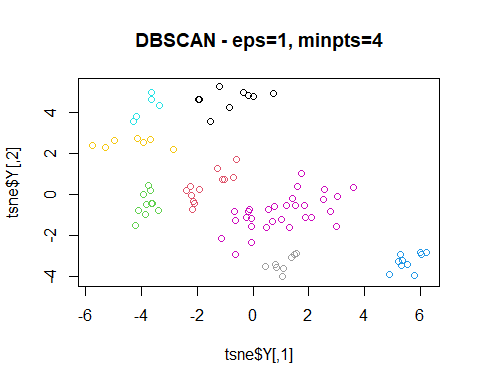
#method="centroid"  
hc.result <- hclust(dist(data.scaled), "centroid")  
plot(tsne$Y, col=hc.result$order, main = "hierarchical - method=centroid")



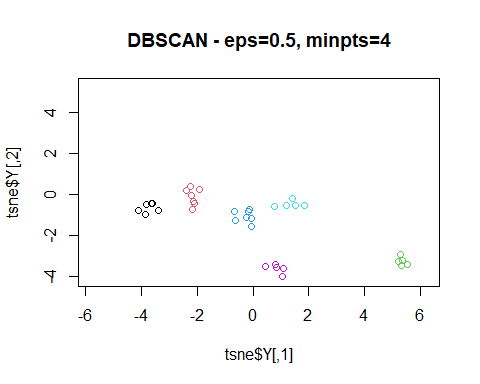
plot(hc.result)



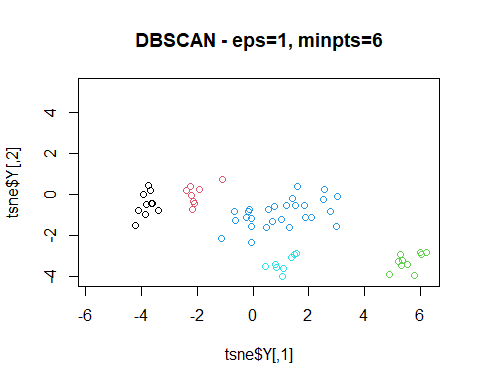
# DBSCAN  
#eps=1,minpts=4  
dbscan.result <- dbscan(tsne$Y, eps = 1, minPts = 4)  
plot(tsne$Y, col=dbscan.result$cluster, main = "DBSCAN - eps=1, minpts=4")



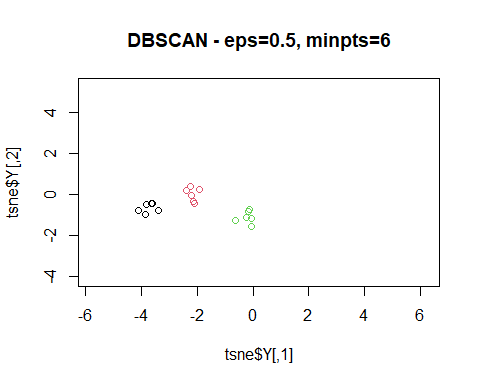
#eps=0.5,minpts=4  
dbscan.result <- dbscan(tsne$Y, eps = 0.5, minPts = 4)  
plot(tsne$Y, col=dbscan.result$cluster, main = "DBSCAN - eps=0.5, minpts=4")



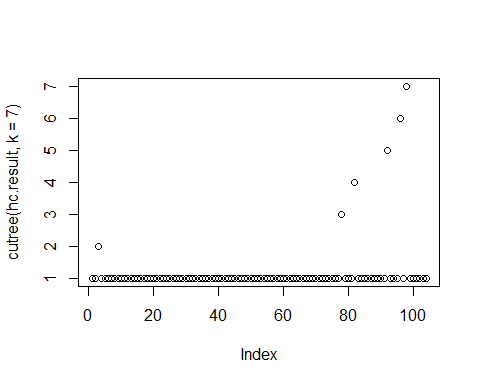
#eps=1,minpts=6  
dbscan.result <- dbscan(tsne$Y, eps = 1, minPts = 6)  
plot(tsne$Y, col=dbscan.result$cluster, main = "DBSCAN - eps=1, minpts=6")



#eps=0.5,minpts=6  
dbscan.result <- dbscan(tsne$Y, eps = 0.5, minPts = 6)  
plot(tsne$Y, col=dbscan.result$cluster, main = "DBSCAN - eps=0.5, minpts=6")



plot(cutree(hc.result, k=7))



# plotting to hclust dendrogram  
#plot(hc.result)

# Q3.3

The method in hclust() that seems to give a clustering that is more accurate to the tSNE plot is ????.

# Q3.4

# TODO - write the code for rand.index

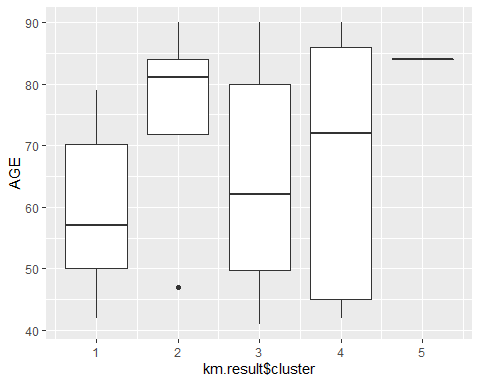
Did you get the same or different clusters using the three techniques? Find ways to present the similarities and differences between the clusterings. If the clusters are different what can explain the difference?

Answer: the clusters are different. reasons- the method that each clustering technique uses is different way to cluster the similarity is because the data is small

# Q3.5

find who belongs to each cluster show it on a graph

sample\_info <- read.csv("C:/CS\_Technion/236523-Intro\_to\_Bioinformatics/Spr21/Intro\_to\_BioInf-Spr21-HW2/sample\_info.csv")  
  
# new kmeans?  
  
# maybe with boxplot? and not ggplot?  
#AGE  
ggplot(sample\_info, aes(x=km.result$cluster, y=AGE, group = km.result$cluster)) + geom\_boxplot()



#RACE  
table(sample\_info$RACE, km.result$cluster)

##   
## 1 2 3 4 5  
## Asian 3 1 11 4 0  
## Black or African American 0 0 0 1 0  
## Native Hawaiian or Other Pacific Islander 0 0 0 1 0  
## Not Specified 0 1 2 1 0  
## Other 1 0 3 0 0  
## White 10 2 48 14 1

#GENDER  
table(sample\_info$GENDER, km.result$cluster)

##   
## 1 2 3 4 5  
## Female 7 0 38 14 0  
## Male 7 4 26 7 1

# Q3.5

explain the results