Assignment7

Arissara Tubtimyoy SIMB/M 6338008

21 Febuary 2021

7. Which of the following code chunks will make a heatmap of the 500 most highly expressed genes (as defined by total count), without re-ordering due to clustering? Are the highly expressed samples next to each other in sample order?

Solution:

To find the answer, I will try to run all code chunks to see which code chunk makes a heatmap of the 500 most highly expressed genes without clustering.

Install a required package

```
BiocManager::install("Biobase")
```

Load the library

```
library(Biobase)
```

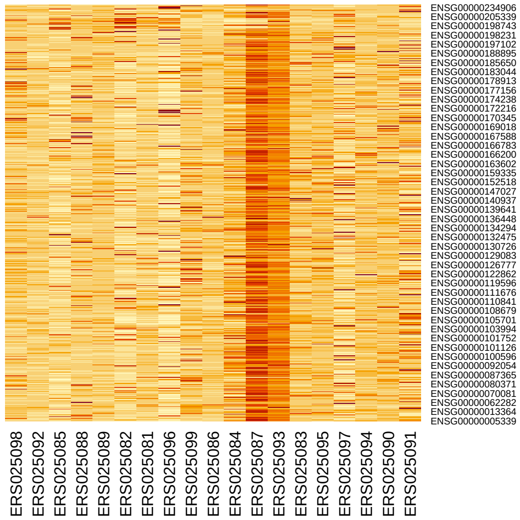
Load the data

```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/bodymap_eset.RData")
load(file=con)
close(con)
bm = bodymap.eset
edata = exprs(bm)
```

Choice 1: The highly expressed samples are next to each other.

Try code chunk1

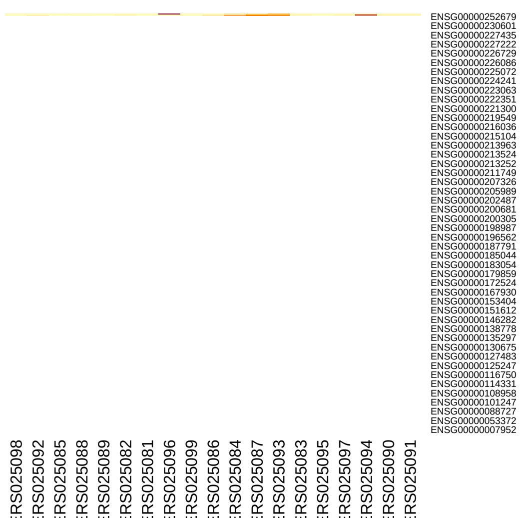
```
row_sums = rowSums(edata)
index = which(rank(-row_sums) < 500 )
heatmap(edata[index,],Rowv=NA,Colv=NA)
```



Choice 2: No they are not next to each other.

Try code chunk2

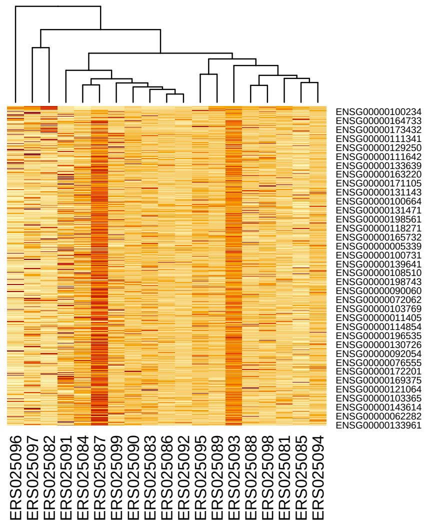
```
row_sums = rowSums(edata)
edata = edata[order(row_sums),]
index = which(rank(-row_sums) < 500 )
heatmap(edata[index,],Rowv=NA,Colv=NA)
```



Choice 3: The highly expressed samples are next to each other.

Try code chunk3

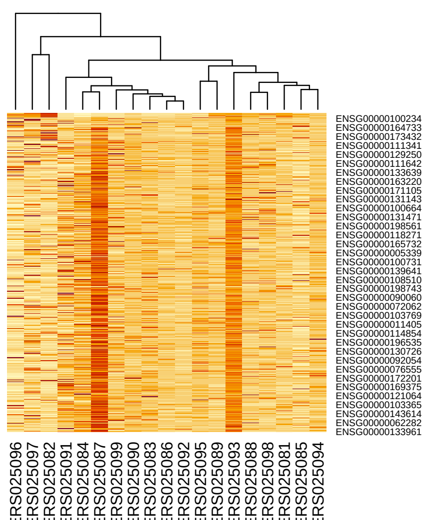
```
row_sums = rowSums(edata)
index = which(rank(-row_sums) < 500 )
heatmap(edata[index,],Rowv=NA)
```



Choice 4: The highly expressed samples are not next to each other.

Try code chunk4

```
heatmap(edata[index,],Rowv=NA)
```



Answer:

From the above heat maps, a heat map provided by code chunk1 shows the 500 most highly expressed genes without clustering. In addition, the highly expressed samples (red color) are next to each other. Therefore, the answer to the question 7 is choice 1.

8. Make an MA-plot of the first sample versus the second sample using the log2 transform (hint: you may have to add 1 first) and the rlog transform from the DESeq2 package. How are the two MA-plots different? Which kind of genes appear most different in each plot?

Solution:

Install a required package

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("DESeq2")
```

Load library

```
library(DESeq2)
```

Load the Bodymap data

```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/bodymap_eset.RData")
load(file=con)
close(con)
bm = bodymap.eset
pdata = pData(bm)
edata = exprs(bm)
```

Check for missing values (NA)

```
sum(is.na(edata))
```

```
## [1] 0
```

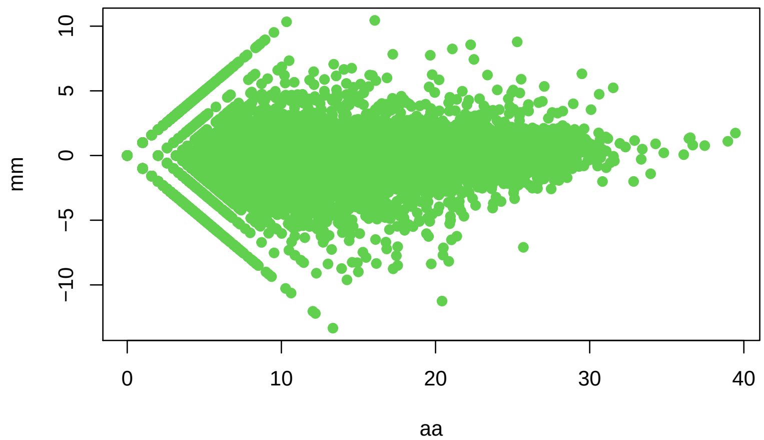
Check for the number of matching rows

```
dim(edata)
```

```
## [1] 52580    19
```

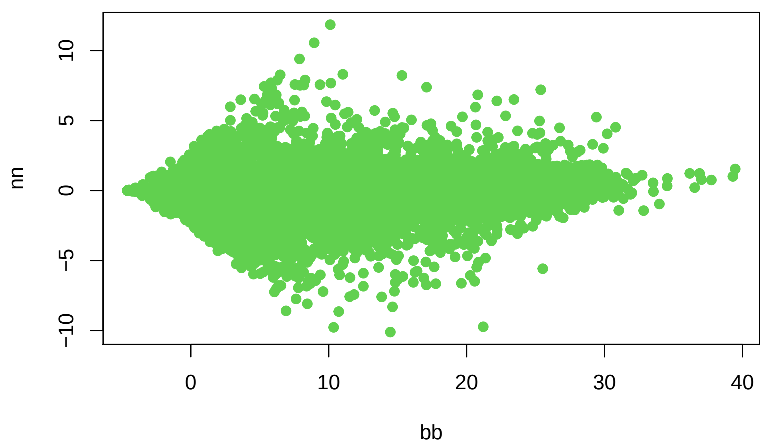
MA-plot using the log2 transform

```
mm = log2(edata[,1]+1) - log2(edata[,2]+1)
aa = log2(edata[,1]+1) + log2(edata[,2]+1)
plot(aa,mm,col=3,pch = 19)
```



MA-plot using the rlog transform

```
edata_bm <-rlog(edata)

nn = (edata_bm[,1])-(edata_bm[,2])
bb = (edata_bm[,1])+(edata_bm[,2])
plot(bb,nn,col=3,pch = 19)
```



Choice 1:

The plots look **pretty similar**, but there are two strong diagonal stripes (corresponding to the zero count genes) in **the rlog plot**. In both cases, the genes in the middle of the expression distribution show the biggest differences, but the low abundance genes seem to show smaller differences with **the log2 transform**.

Choice 2:

The plots look **pretty similar**, but there are two strong diagonal stripes (corresponding to the zero count genes) in **the log2 plot**. In both cases, the genes in the middle of the expression distribution show the biggest differences, but the low abundance genes seem to show smaller differences with **the rlog transform**.

Choice 3:

The plots are **very different** as **the log2 plot** seems to shrink **low abundance** genes more and **the rlog plot** seems to shrink **high abundance** genes more. The genes in the middle of the distribution show the biggest differences.

Choice 4:

The plots are **nearly identical**. Both transforms seem to deal with the **low abundance** genes, including the zero genes the same way. The high-abundance genes show the most differences.

Answer:

According to the comparison of the log2 plot and the rlog plot, it seems both plots look similar, but the log2 plot has two stron gdiagonal stripes. Moreover, the rlog plot shows smaller differences of the low abundance genes. Therefore, the answer to the question 8 is choice 2.

9. Cluster the data in three ways:

1. With no changes to the data

2. After filtering all genes with rowMeans less than 100

3. After taking the log2 transform of the data without filtering

Color the samples by which study they came from (Hint: consider using the function myplclust.R in the package rafalib available from CRAN and looking at the argument lab.col.)

How do the methods compare in terms of how well they cluster the data by study? Why do you think that is?

Solution:

Install a required package

```
install.packages("rafalib", repos = "http://cran.us.r-project.org")
```

Load library

```
library(rafalib)
```

Load the Montgomery and Pickrell eSet

```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData")
load(file=con)
close(con)
mp = montpick.eset
pdata=pData(mp)
edata=as.data.frame(exprs(mp))
fdata = fData(mp)
```

Check for the study that provides the samples

```
str(pdata)
```

```
## 'data.frame':    129 obs. of  4 variables:
##  $ sample.id    : Factor w/ 129 levels "NA06985","NA06986",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ num.tech.reps: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ population   : Factor w/ 2 levels "CEU","YRI": 1 1 1 1 1 1 1 1 1 1 ...
##  $ study        : Factor w/ 2 levels "Montgomery","Pickrell": 1 1 1 1 1 1 1 1 1 1 ...
```

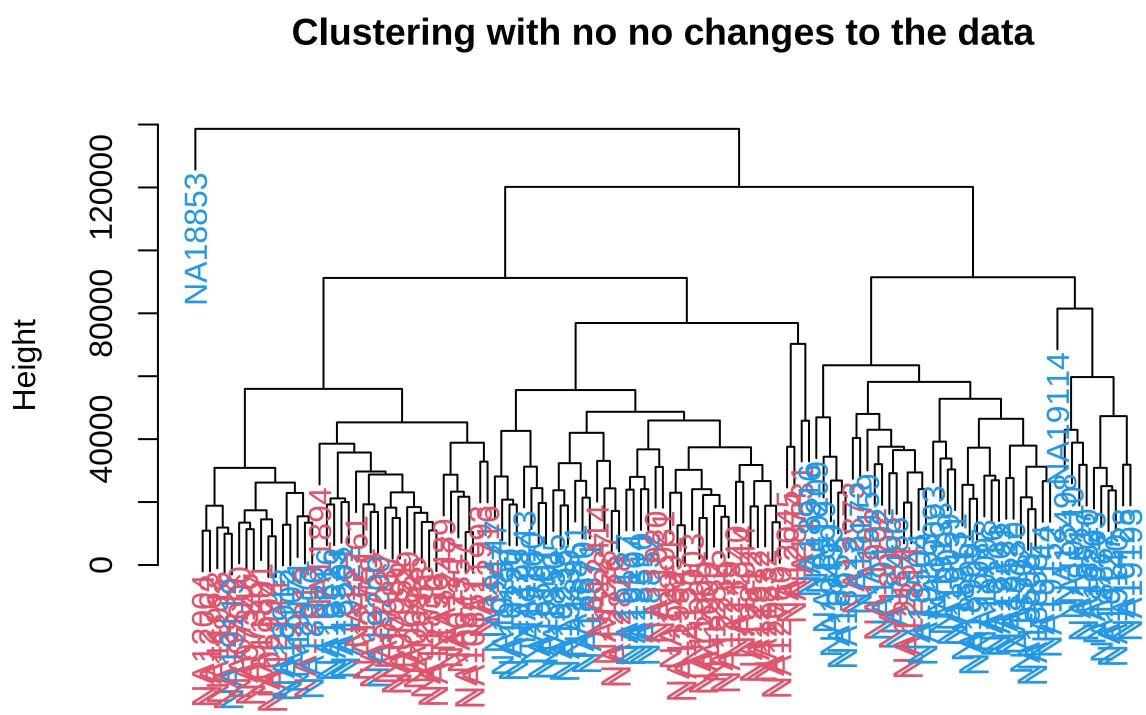
Check for the number of the samples in each study

```
table(pdata$study)
```

```
##
## Montgomery   Pickrell
##         60         69
```

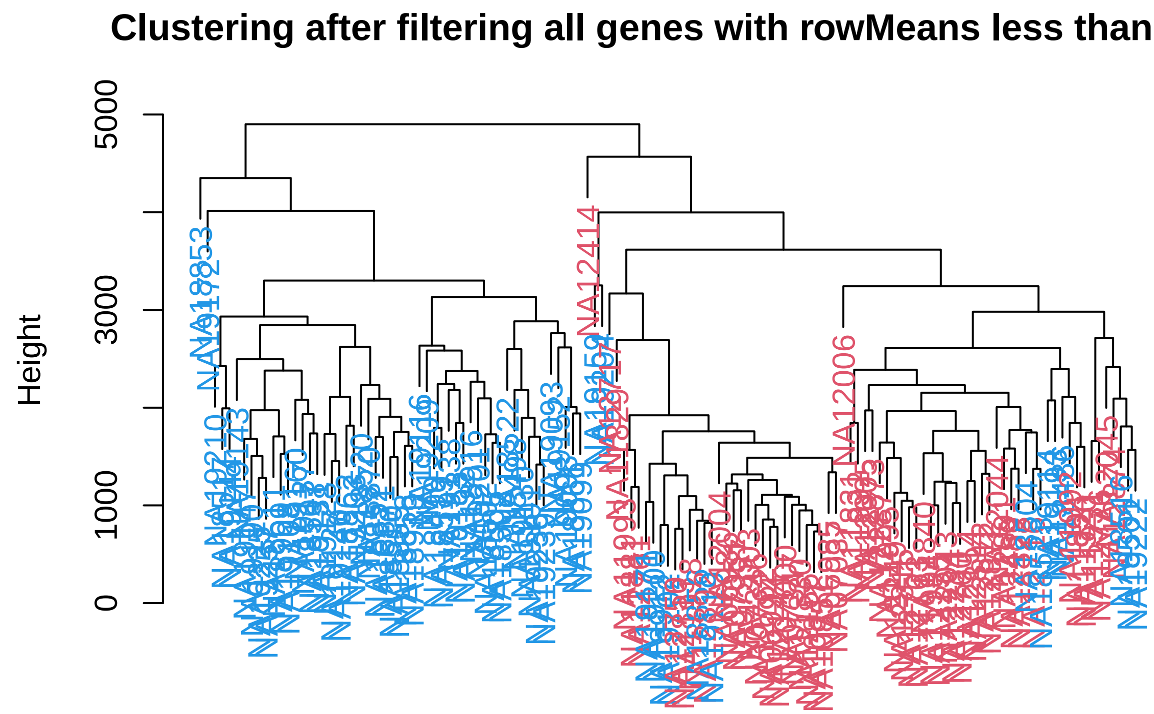
1. Cluster the data with no changes to the data

```
dist1 = dist(t(edata))
hclust1 = hclust(dist1)
col<-c(rep(2,60), rep(4, 69))

myplclust(hclust1,
          labels = hclust1$labels,
          lab.col = col , hang = 0.1,
          main = "Clustering with no no changes to the data")
```



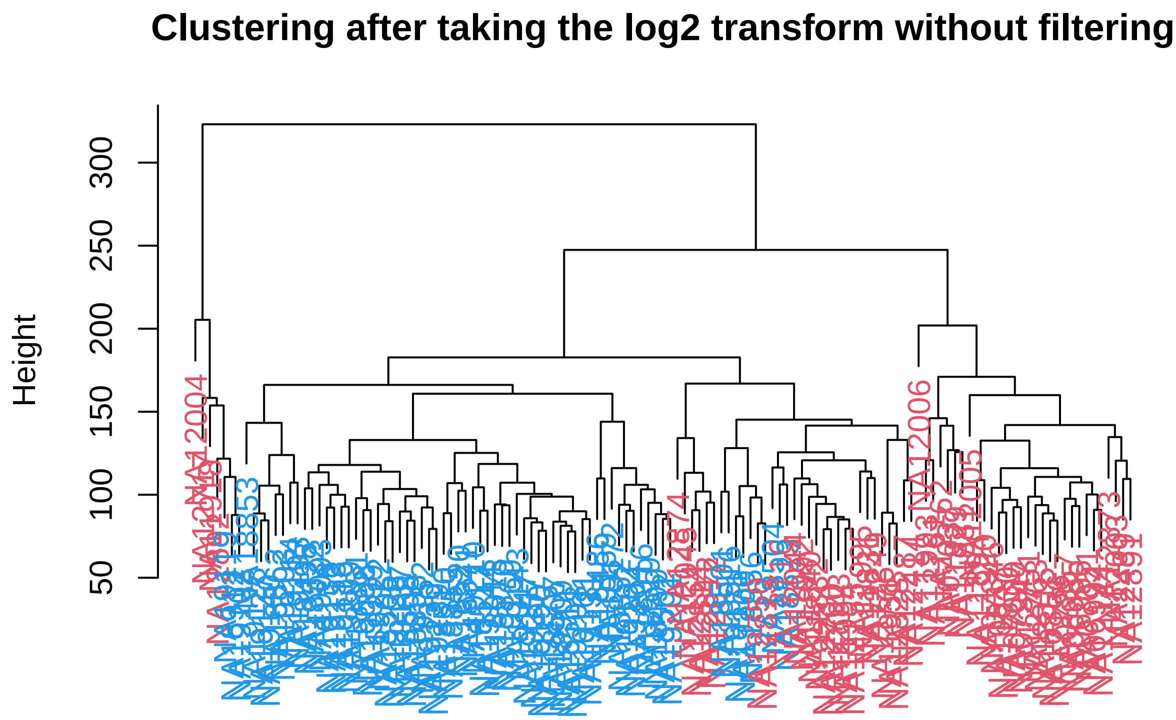
2. Cluster the data after filtering all genes with rowMeans less than 100

```
edata2 = edata[rowMeans(edata) < 100,]
dist2 = dist(t(edata2))
hclust2 = hclust(dist2)
col<-c(rep(2,60), rep(4, 69))

myplclust(hclust2,
          labels = hclust2$labels,
          lab.col = col , hang = 0.1,
          main = "Clustering after filtering all genes with rowMeans less than 100")
```



3. Cluster the data after taking the log2 transform of the data without filtering

```
edata3 = log2(edata + 1)
dist3 = dist(t(edata3))
hclust3 = hclust(dist3)
col<-c(rep(2,60), rep(4, 69))

myplclust(hclust3,
          labels = hclust3$labels,
          lab.col = col , hang = 0.1,
          main = "Clustering after taking the log2 transform without filtering")
```



Choice 1:

Clustering with or without **log2 transform** is about **the same. Clustering after filtering** shows **better** clustering with respect to the study variable. The reason is that the lowly expressed genes have some **extreme outliers that skew the calculation**.

Choice 2:

Clustering with or without **filtering** is about **the same. Clustering after the log2 transform** shows **better** clustering with respect to the study variable. The likely reason is that the highly skewed distribution doesn't match the Euclidean distance metric being used in the clustering example.

Choice 3:

Clustering is **identical with all three approaches** and they show equal clustering. The distance is an average over all the dimensions so it doesn't change.

Choice 4:

Clustering with or without **log2 transform** is about **the same. Clustering after filtering** shows **better** clustering with respect to the study variable. The reason is that it is just the lowly expressed genes that make **the distance hard to calculate**.

Answer:

From the above plots, I think clustering with or without filtering is about the same and clustering after the log2 transform looks better for clustering. Therefore, the answer to the question 9 is 2.

10. Cluster the samples using k-means clustering after applying the log2 transform (be sure to add 1). Set a seed for reproducible results (use set.seed(1235)). If you choose two clusters, do you get the same two clusters as you get if you use the cutree function to cluster the samples into two groups? Which cluster matches most closely to the study labels?

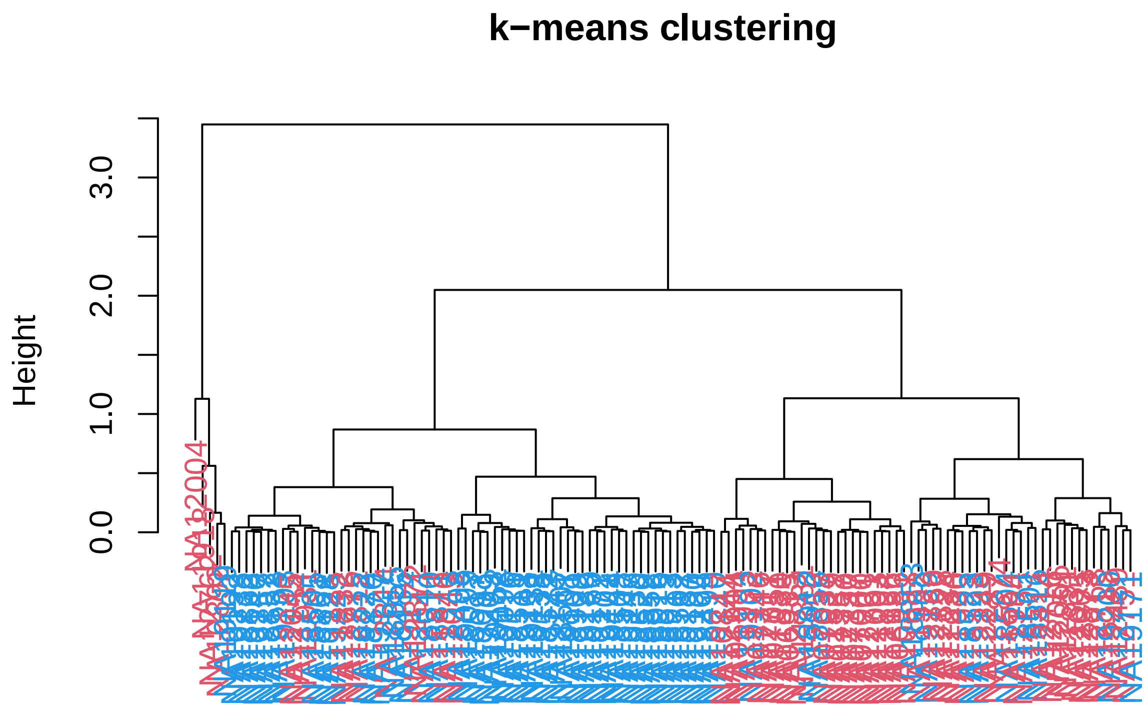
Solution:

Load the Montgomery and Pickrell eSet

```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData")
load(file=con)
close(con)
mp = montpick.eset
pdata=pData(mp)
edata=as.data.frame(exprs(mp))
fdata = fData(mp)
```

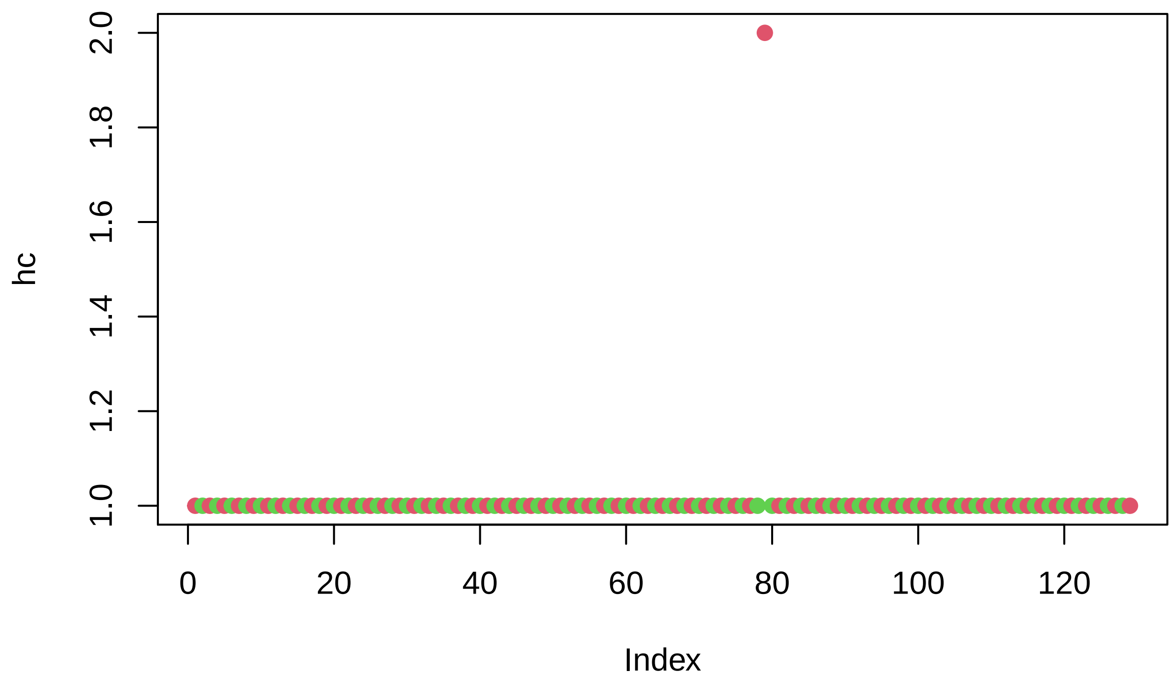
Cluster the samples using k-means clustering after applying the log2 transform

```
edata_kmeans = log2(edata + 1)
set.seed(1235)
kmeans1 = kmeans(edata_kmeans,centers=2)
distk = dist(t(kmeans1$centers))
hclustk = hclust(distk)
col<-c(rep(2,60), rep(4, 69))

myplclust(hclustk, labels = hclustk$labels,
          lab.col = col , hang = 0.1,
          main = "k-means clustering")
```



Cluster the samples using hierarchical clustering after applying the log2 transform

```
disth = dist(t(edata))
hclusth = hclust(disth)
hc <- cutree(hclusth, k=2)
plot(hc,col=2:3,pch = 19)
```



Choice 1:

They produce **different answers**. The **k-means** clustering matches study **better**. Hierarchical clustering would look better if we went farther down the tree but the top split doesn't perfectly describe the study variable.

Choice 2:

They produce **the same answers** and match the study variable **equally** well.

Choice 3:

They produce **the same answers** except for three samples that **hierarchical** clustering **correctly** assigns to the right study but **k-means** does **not.**

Choice 4:

They produce **different answers**, with **k-means** clustering giving a much more **unbalanced clustering**. The **hierarchical** clustering matches study **better**.

Answer:

Due to I cannot create a dendrogram using cutree function, I guess the answer to the question 10 is choice 1