assignment7

Chanida Limthamprasert

21/2/2564

#install biobase package

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

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

```
## Bioconductor version 3.12 (BiocManager 1.30.10), R 4.0.3 (2020-10-10)

## Installing package(s) 'Biobase'

## package 'Biobase' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'Biobase'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:
## \Users\Nut\Documents\R\win-library\4.0\00LOCK\Biobase\libs\x64\Biobase.dll to C:
## \Users\Nut\Documents\R\win-library\4.0\Biobase\libs\x64\Biobase.dll: Permission
## denied

## Warning: restored 'Biobase'

##
## The downloaded binary packages are in
##  C:\Users\Nut\AppData\Local\Temp\RtmpEnq2jP\downloaded_packages

## Installation path not writeable, unable to update packages: boot, class,
##   cluster, codetools, foreign, KernSmooth, MASS, Matrix, mgcv, nlme, nnet,
##   spatial

## Old packages: 'cachem', 'data.table', 'kableExtra', 'svglite', 'xfun'
```

##Question7

answer:

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

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: ’BiocGenerics’
```

```
## The following objects are masked from ’package:parallel’:
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

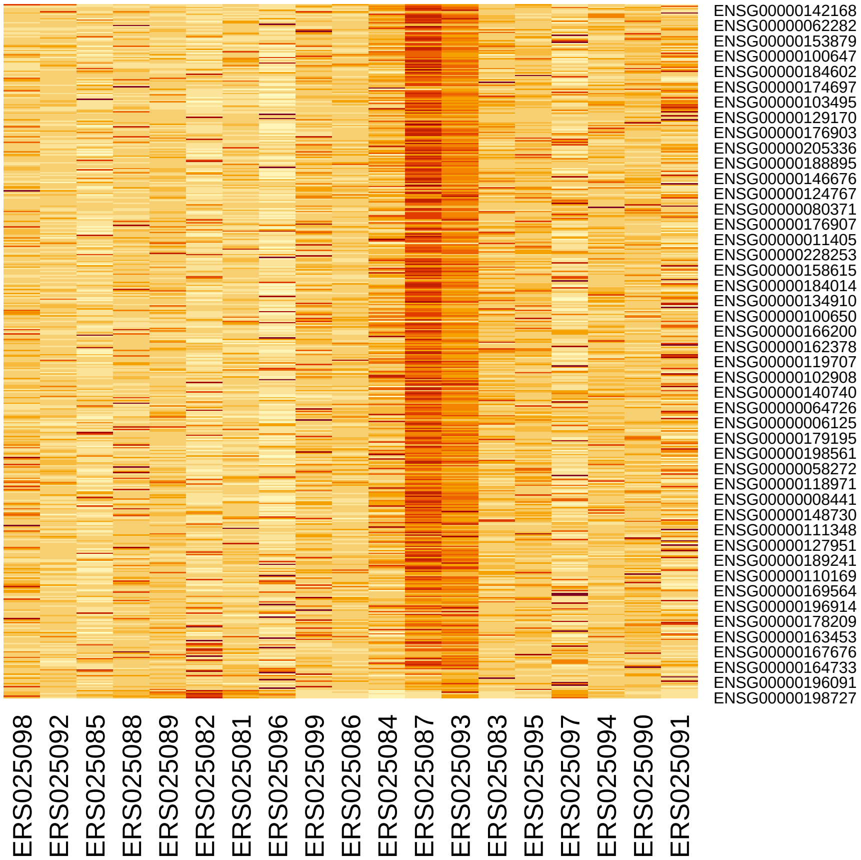
```
## The following objects are masked from ’package:stats’:
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from ’package:base’:
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     ’browseVignettes()’. To cite Bioconductor, see
##     ’citation("Biobase")’, and for packages ’citation("pkgname")’.
```

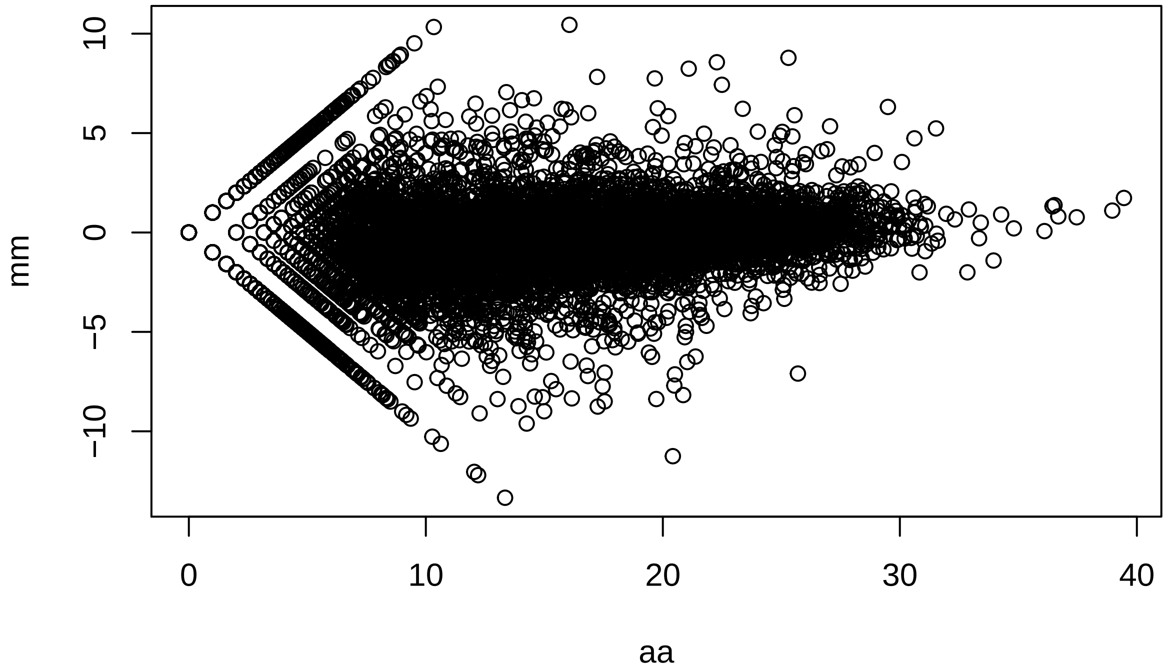
```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 52580 features, 19 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: ERS025098 ERS025092 ... ERS025091 (19 total)
##   varLabels: sample.id num.tech.reps ... race (6 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: ENSG00000000003 ENSG00000000005 ... LRG_99 (52580
##     total)
##   fvarLabels: gene
##   fvarMetadata: labelDescription
## experimentData: use ’experimentData(object)’
## Annotation:
```

```
edata = exprs(bm)
row_sums = rowSums(edata)
edata = edata[order(-row_sums),]
index = 1:500
heatmap(edata[index,],Rowv=NA,Colv=NA)
```



#Question8 ###MA-plot in log2

```
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)

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



###MA-plot in rlog

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

BiocManager::install("DESeq2")
```

```
## Bioconductor version 3.12 (BiocManager 1.30.10), R 4.0.3 (2020-10-10)
```

```
## Installing package(s) 'DESeq2'
```

```
## package 'DESeq2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Nut\AppData\Local\Temp\RtmpEnq2jP\downloaded_packages
```

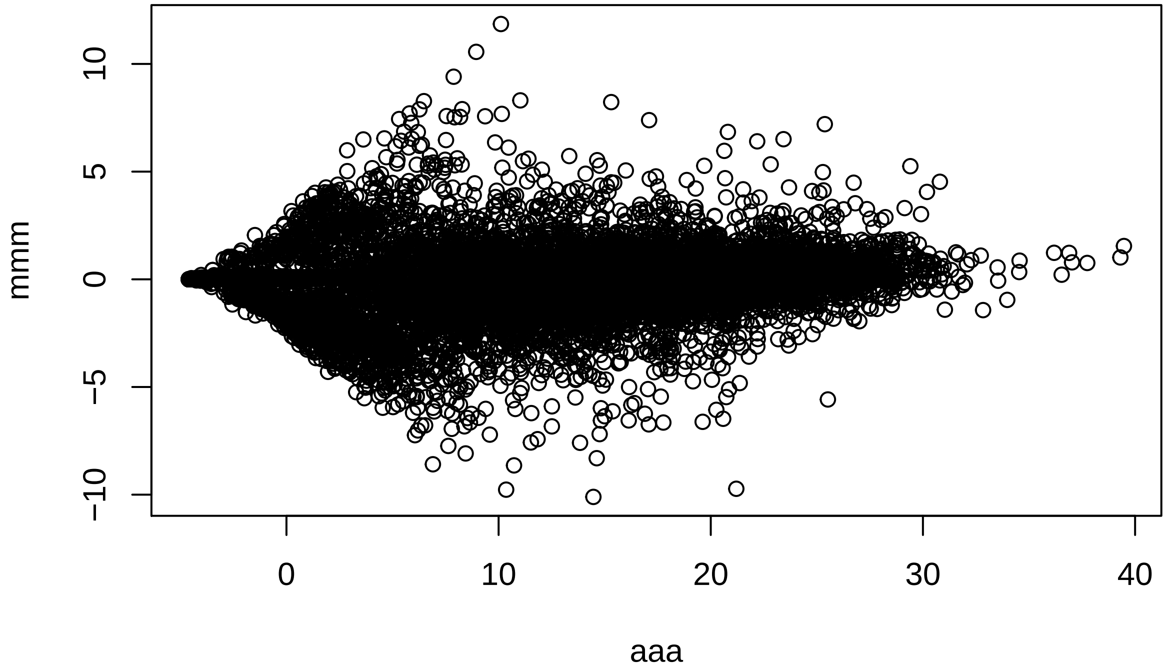
```
## Installation path not writeable, unable to update packages: boot, class,
##   cluster, codetools, foreign, KernSmooth, MASS, Matrix, mgcv, nlme, nnet,
##   spatial
```

```
## Old packages: 'cachem', 'data.table', 'kableExtra', 'svglite', 'xfun'
```

```
library( "DESeq2" )
```

```
## Loading required package: S4Vectors
## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: GenomicRanges

## Loading required package: GenomeInfoDb

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'matrixStats'

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
## The following object is masked from 'package:Biobase':
##
##     rowMedians
```

```
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)

edata1 <-rlog(edata)

mmm = (edata1[,1])-(edata1[,2])
aaa = (edata1[,1])+(edata1[,2])
plot(aaa,mmm)
```



answer: 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.

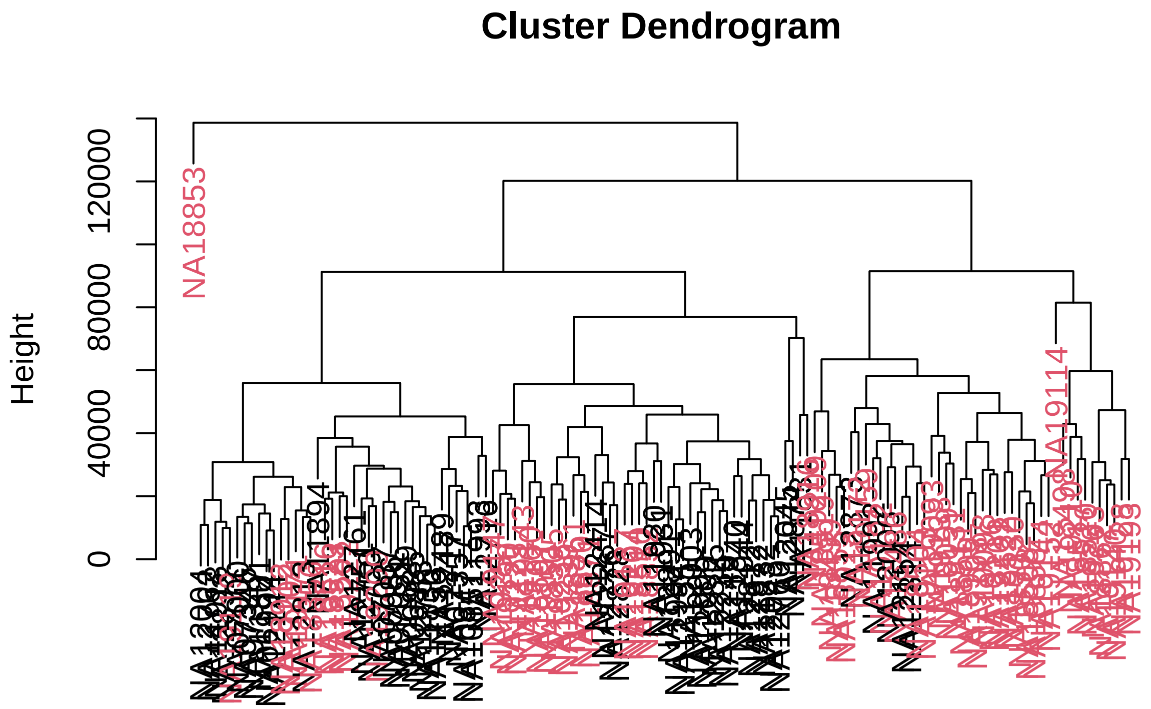
##Question9

###Cluster With no changes to the data

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

```
## Installing package into 'C:/Users/Nut/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## package 'rafalib' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Nut\AppData\Local\Temp\RtmpEnq2jP\downloaded_packages
```

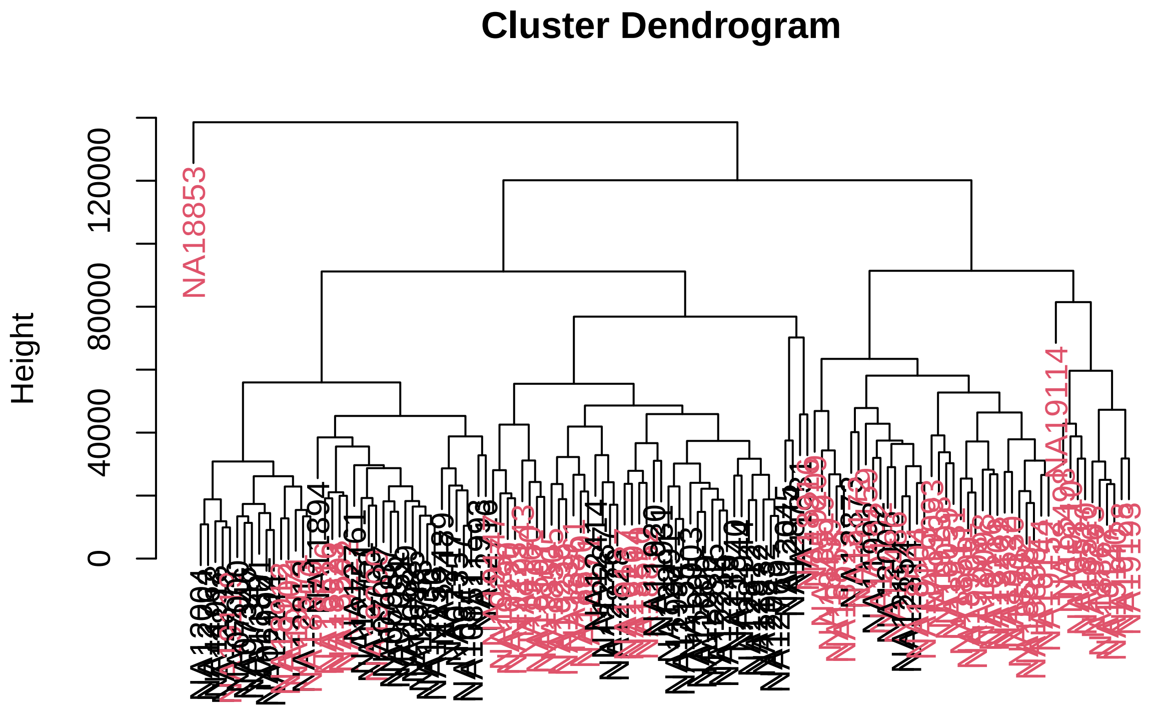
```
library(rafalib)

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)
dist1 = dist(t(edata))
hclust1 = hclust(dist1)
myplclust(hclust1, labels = pdata$sample.id, lab.col = as.numeric(pdata$study), hang = 0.1)
```



###Cluster After filtering all genes with rowMeans less than 100

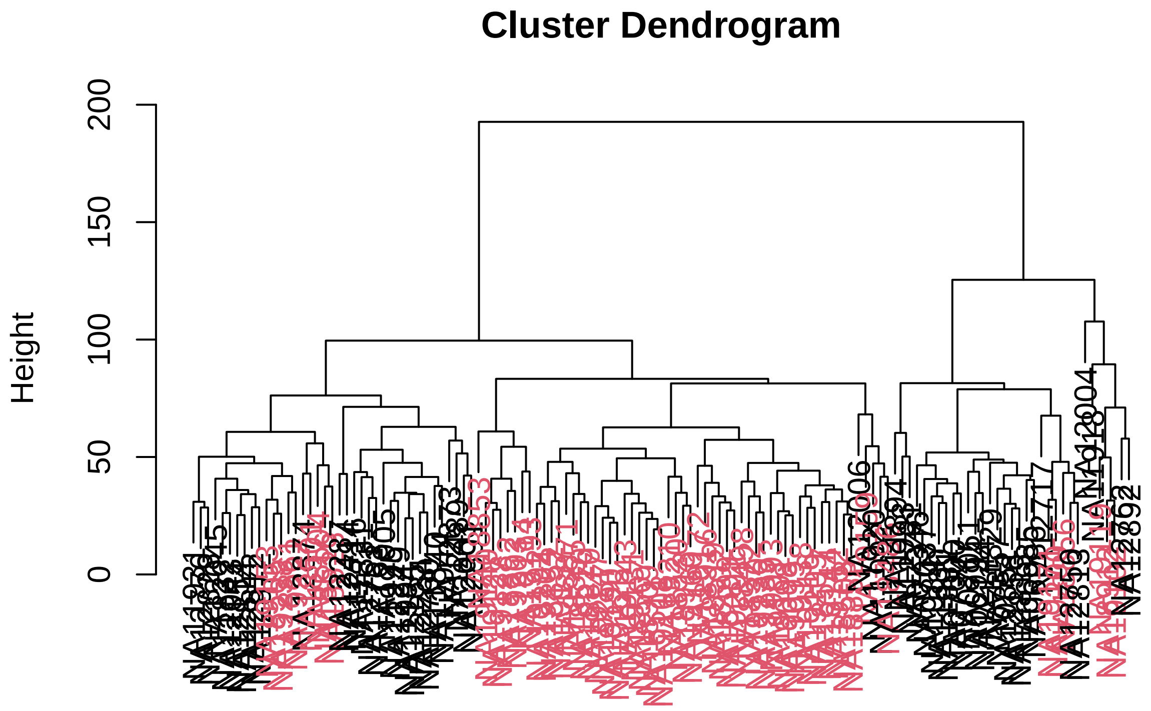
```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData")
load(file=con)
```

```
close(con)
mp = montpick.eset
pdata=pData(mp)
edata=as.data.frame(exprs(mp))
fdata = fData(mp)
edata = edata[rowMeans(edata) > 100,]
dist1 = dist(t(edata))
hclust1 = hclust(dist1)
myplclust(hclust1, labels = pdata$sample.id, lab.col = as.numeric(pdata$study), hang = 0.1)
```



###After taking the log2 transform of the data without filtering

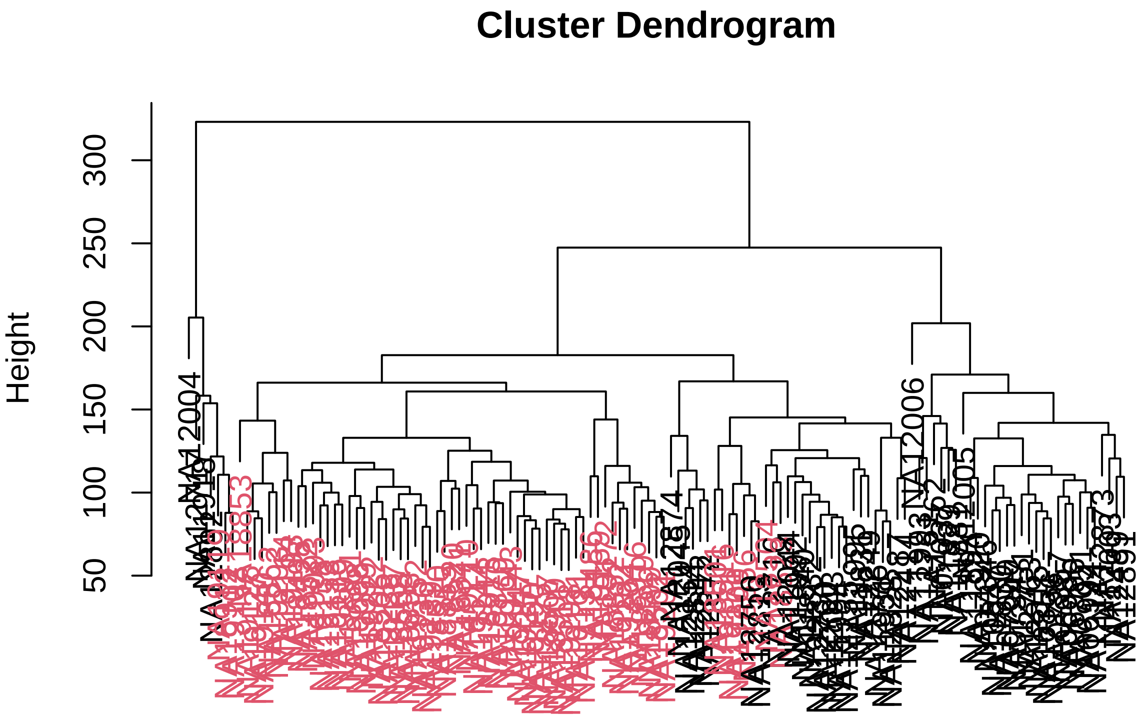
```
con =url("http://bowtie-bio.sourceforge.net/recount/ExpressionSets/montpick_eset.RData")
load(file=con)
close(con)
mp = montpick.eset
pdata=pData(mp)
edata=as.data.frame(exprs(mp))
fdata = fData(mp)
edata = edata[rowMeans(edata) > 100,]
edata = log2(edata + 1)
dist1 = dist(t(edata))
hclust1 = hclust(dist1)
myplclust(hclust1, labels = pdata$sample.id, lab.col = as.numeric(pdata$study), hang = 0.1)
```



answer: 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.

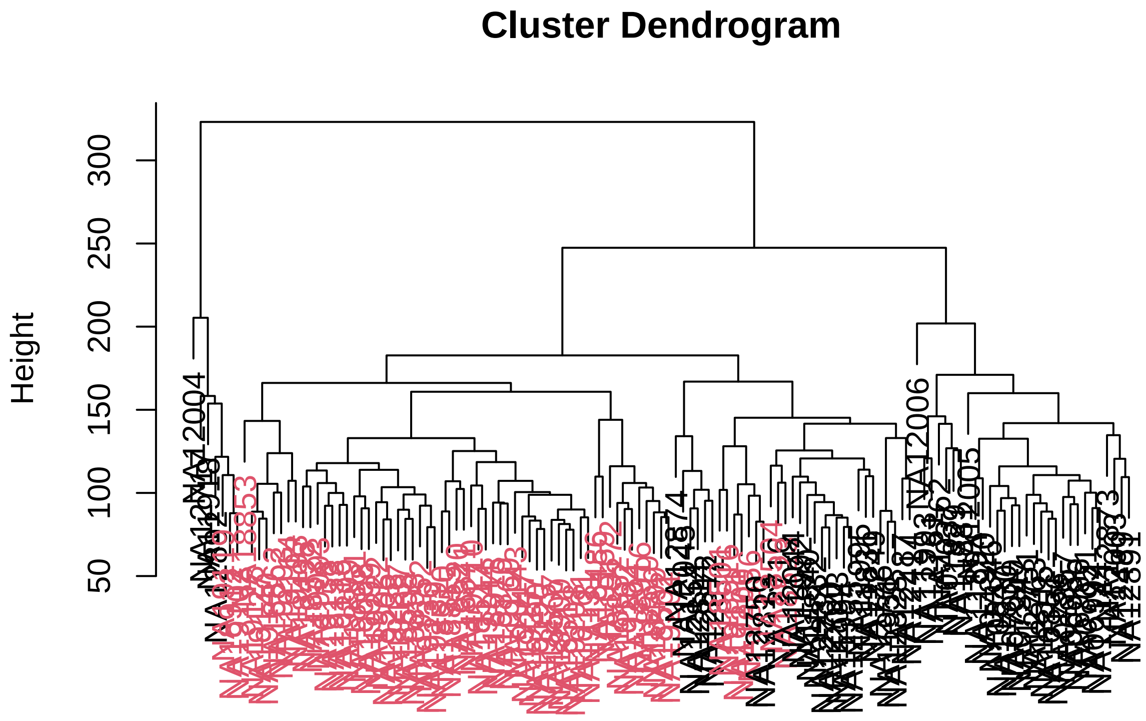
##Question10 ### k-means clustering

```
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)
edata = log2(edata + 1)
set.seed(1235)

kmeans1 = kmeans(edata,center=2)
newdata = as.matrix(edata)[order(kmeans1$cluster),]
dist1 = dist(t(edata))
hclust1 = hclust(dist1)
myplclust(hclust1, labels = pdata$sample.id, lab.col = as.numeric(pdata$study), hang = 0.1)
```



###cutree

```
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)
edata = log2(edata + 1)
set.seed(1235)

dist1 = dist(t(edata))
hclust1 = hclust(dist1)
cutree1 = cutree((hclust1), k =2)

myplclust(hclust1, labels = pdata$sample.id, lab.col = as.numeric(pdata$study), hang = 0.1)
```



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