Question9

Chonnikarn Charoenpanich

#Installing Biobase

```
library(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")'.
```

Question 9

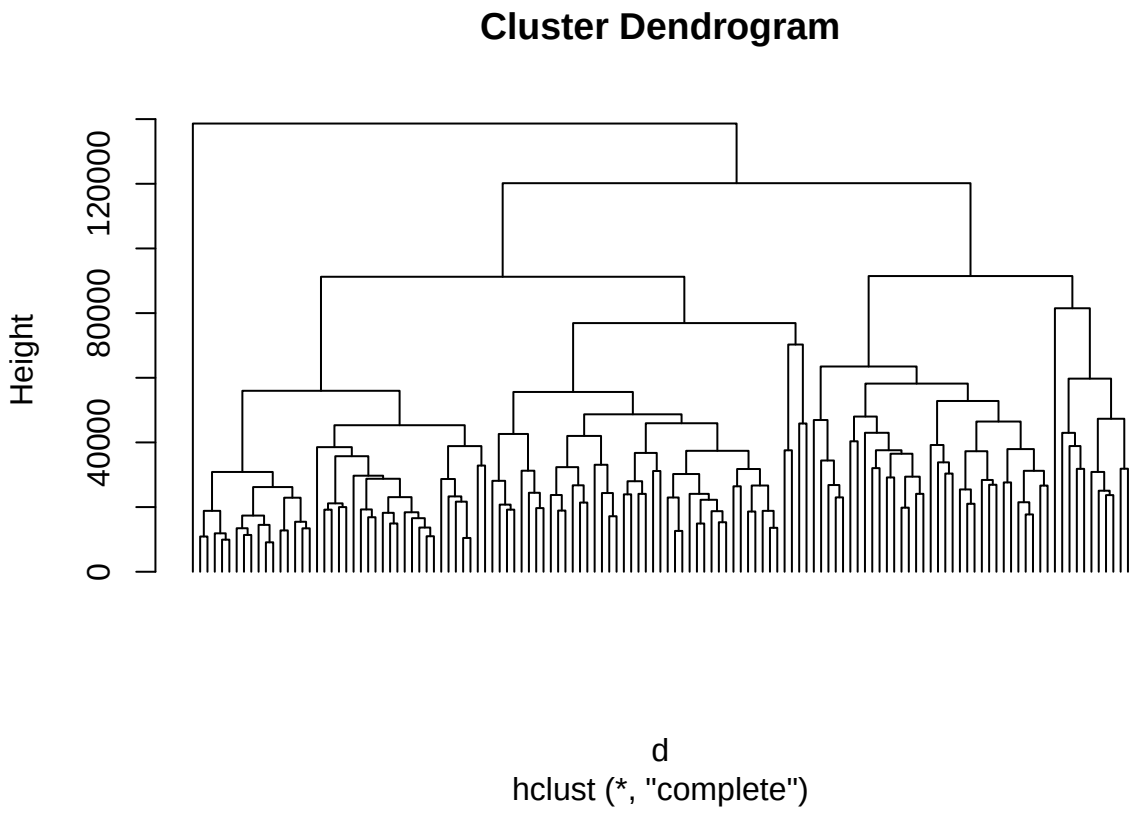
Load the Montgomery and Pickrell eSet:

```
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 data in three ways:

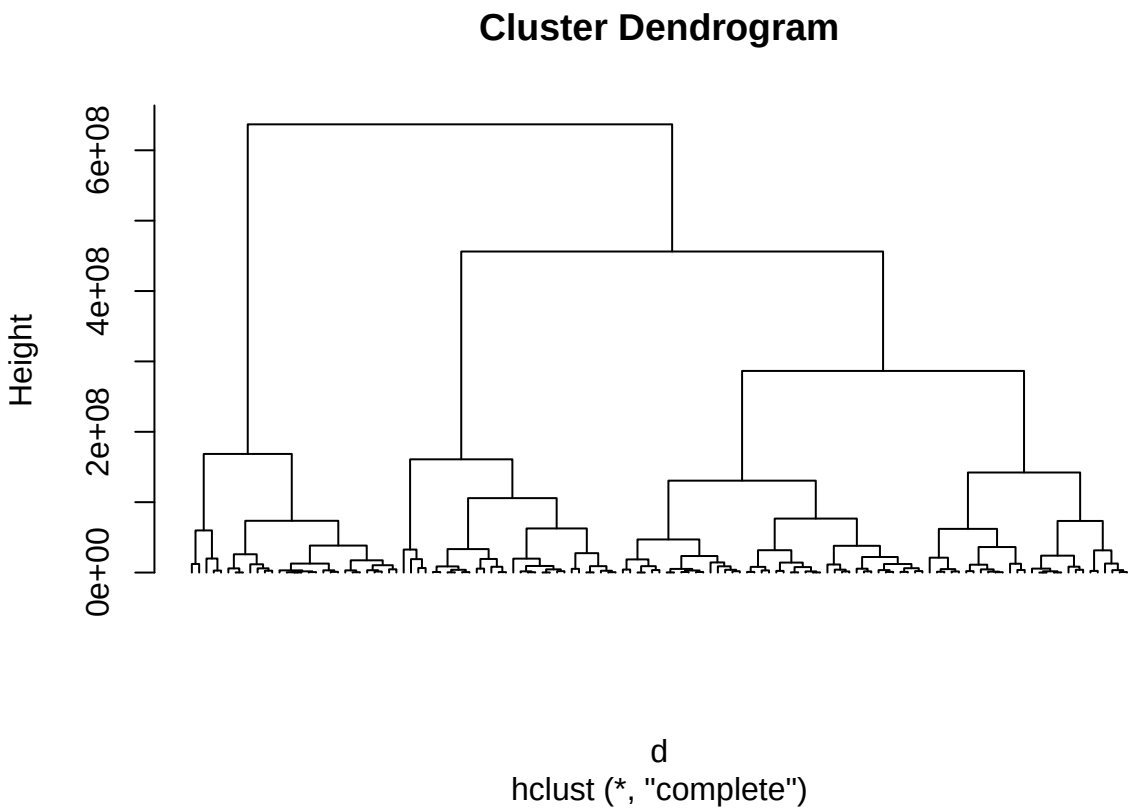
With no changes to the data

```
d = dist(t(edata))
h = hclust(d)
plot(h, hang = -1, labels=FALSE)
```



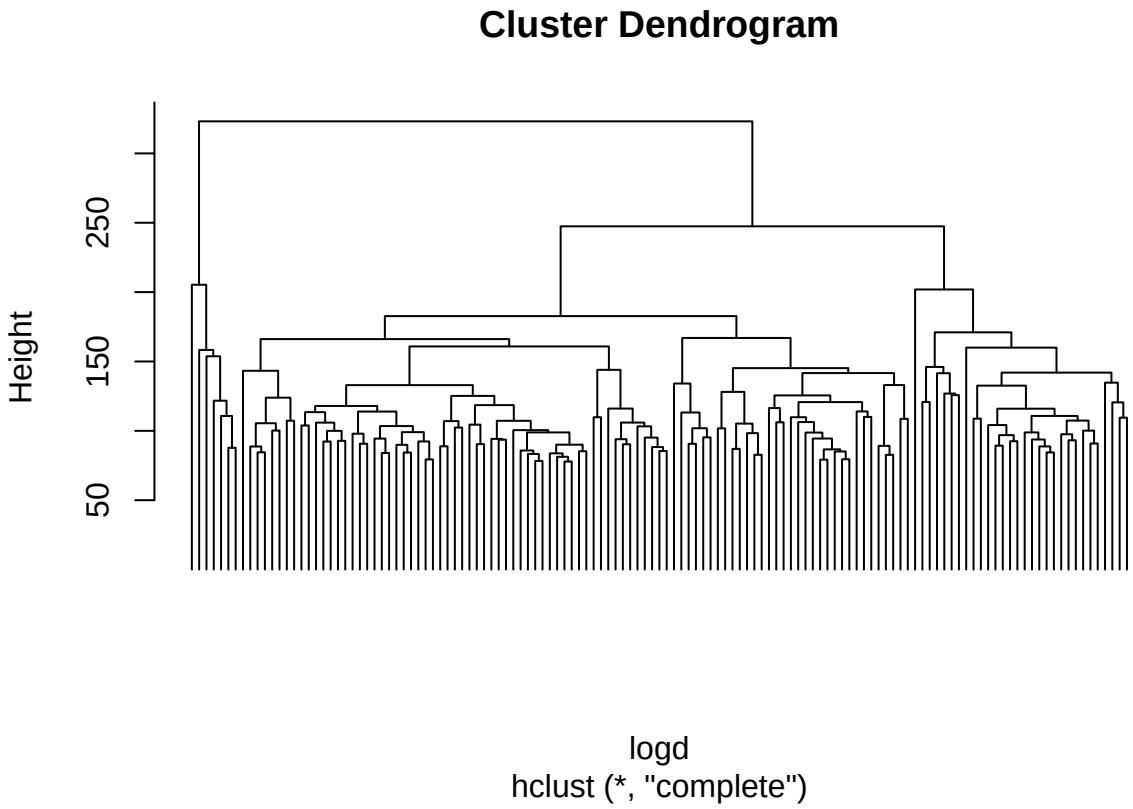
After filtering all genes with `rowMeans`rowMeans less than 100

```
filtered = filter(edata, rowMeans(edata) < 100)
d = dist(t(filtered))
h = hclust(d)
plot(h, hang = -1, labels=FALSE)
```



After taking the log2log2 transform of the data without filtering

```
log = log2(edata + 1)
logd = dist(t(log))
logh = hclust(logd)
plot(logh, hang = -1, labels=FALSE)
```



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

```
library(devtools)
```

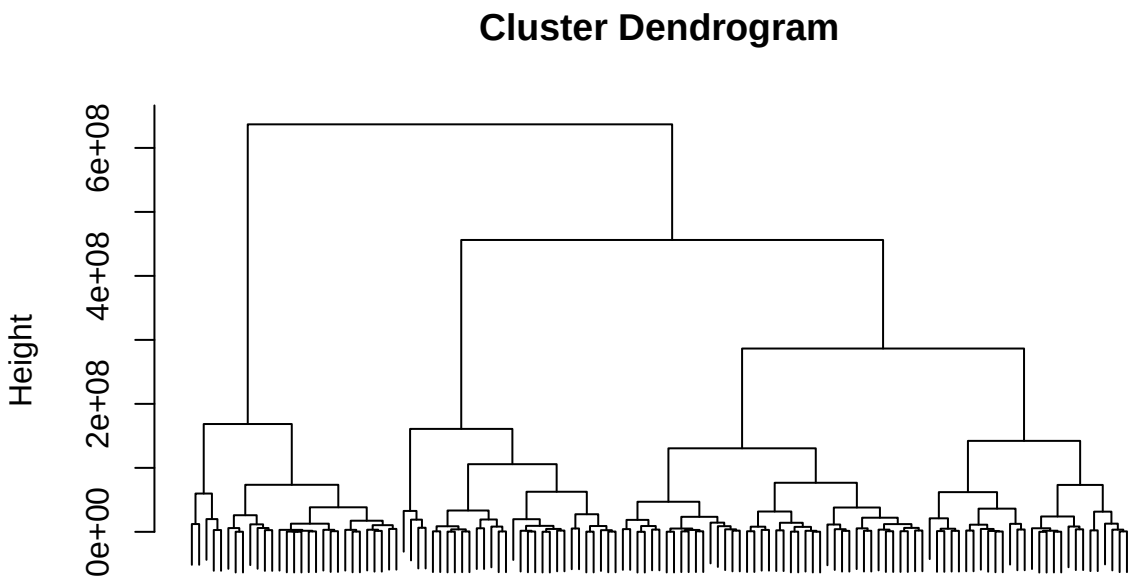
```
## Loading required package: usethis
```

```
install_github("ririzarr/rafalib")
```

```
## WARNING: Rtools is required to build R packages, but is not currently installed.
##
## Please download and install Rtools 4.0 from https://cran.r-project.org/bin/windows/Rtools/.
```

```
## Skipping install of 'rafalib' from a github remote, the SHA1 (6b39b27b) has not changed since last in
##   Use 'force = TRUE' to force installation
```

```
rafalib::myplclust(h,lab.col = d)
```



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

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.