pairwise_DEG_withoutpackage

PanZhang

10/12/2019

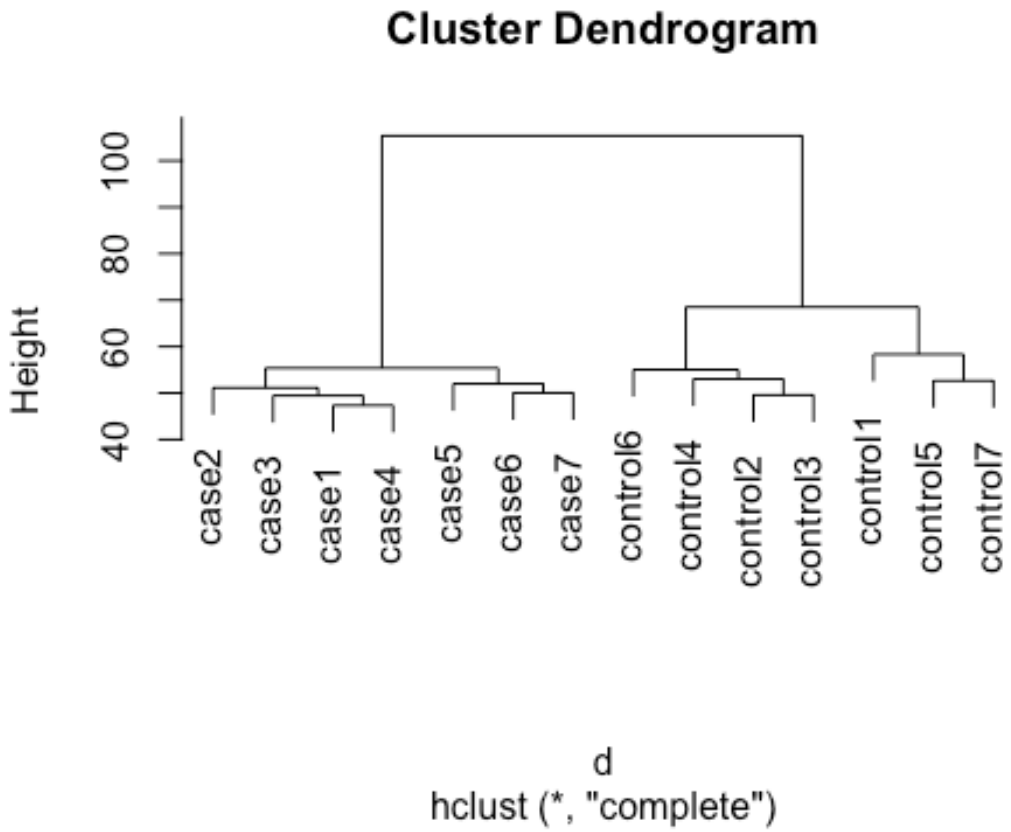
```
setwd("/Users/panzhang/Desktop/basic info/Chicago_senior_bioinformatician")
data <- read.table("htseqcount_combine.txt",header = T, row.names = 1)
dim(data)
```

```
## [1] 26608    14
```

```
# CPM calculation
TN <- apply(data,2,sum)
cpm_data <- data*(1000000)/TN
# Filter genes with no or very low expression
keep <- rowSums(cpm_data>1) >= 7
cpm_data<- cpm_data[keep,]
dim(cpm_data)
```

```
## [1] 11803    14
```

```
#Normalization
normalized_data <- log(cpm_data+1)
d <- dist(t(normalized_data), method = "euclidean", diag = FALSE, upper = FAL
SE, p = 2)
x <- hclust(d)
plot(x, labels = NULL, hang = 0.1, check = TRUE,
     axes = TRUE, frame.plot = FALSE, ann = TRUE,
     main = "Cluster Dendrogram",
     sub = NULL, xlab = NULL, ylab = "Height")
```




```r
# Based on the cluster plot, the samples whithin each group have smaller dist
ance, while the samples between two group have larger distance. So, pass the
samle quality control and move on to differential expression analysis.
n <- dim(normalized_data)[1]
# Differential expression test for each gene
# calculate both p-value and log2(Fold Change)
stat_result <- vector()
for (i in seq(n)){
  result <- t.test(as.numeric(normalized_data[i,1:7]),as.numeric(normalized_d
ata[i,8:14]), paired = TRUE)
  ave_case <- mean(as.numeric(normalized_data[i,1:7]))
  ave_control <- mean(as.numeric(normalized_data[i,8:14]))
  logFC <- log(ave_case/ave_control,base = 2)
  stat_result <- rbind(stat_result, c(result$p.value,logFC))
}
rownames(stat_result) <- rownames(normalized_data)
colnames(stat_result) <- c("p.value", "logFC")
stat_result <- as.data.frame(stat_result)
# Calculate FDR.p.value
FDR.p.value <- p.adjust(as.numeric(stat_result$p.value), method = "BH")
# Summary stat result
stat_result <- cbind(stat_result, FDR.p.value)
# Sort stat_result by FDR.p.value
```

```r
stat_result <-stat_result[order(stat_result$FDR.p.value),]
head(stat_result)
```

```
##                     p.value      logFC  FDR.p.value
## Frmpd3        4.058247e-11  2.9586533 4.789949e-07
## Ddx49         5.043028e-10  0.8790964 2.976143e-06
## Tmem54        2.226903e-09 -0.6215864 8.761380e-06
## 2900053A13Rik 3.664314e-09  3.2251026 1.081247e-05
## 4930429F24Rik 1.049561e-08 -1.0934057 1.166187e-05
## Aqp5          6.668414e-09  4.9901147 1.166187e-05
```

```r
# Visualazed the DEG results by volcano plot
plot(stat_result$logFC, -log10(stat_result$p.value), col = ifelse( stat_resul
t$FDR.p.value < 0.05,'red','green'), xlab="log2(FC)", ylab="-log10(p.value)")

# The red node are differential expressed genes in case group relative to con
trol group
```