Gene Analysis of Colorectal Cancer

Yang Xu Stat366 Project

Introduction

Caused by the abnormal growth cells that have the ability to spread all parts of the body, colorectal cancer (CRC) has become the third most frequent cancer for both males and females in the world especially for the western countries. The standardized incidence rates for the age have kept similar during 1970 to 2004 despite major advanced which have been made in understanding its pathogenesis at the molecular level (Warren, 2005). Based on the research, the overall survival is around 40% in 5 years, and more than one-third of the sample patients would die from this cancer (Karin et al., 2002).

Gene expression could improve accurate diagnosis and prediction of survival and bring new insights into underlying molecular mechanisms. Compared with molecular complexity of disease, molecular studies have been largely focused on individual candidate genes. The accumulating effects of a large amount of genetic alterations accompany the multistep progression of CRC, and this process requires years and even decades accomplishing (François et al., 2004). Previous experiment showed a result that some genes were reported as differentially expressed with a statistically significant frequency in cancer compared with normal and adenoma compared with normal comparisons but not in the cancer compared with adenoma comparison (Simon, 2008).

In this project, we will analyze the gene expression for the colorectal cancer to see if the gene expressions are different for tumor tissue and normal tissue, whether colorectal cancer is related to certain genes and if there is a relationship between specific traits and different genes. We will apply RNA sequence analysis, Differential Expression and Network Analysis to this project.

Data

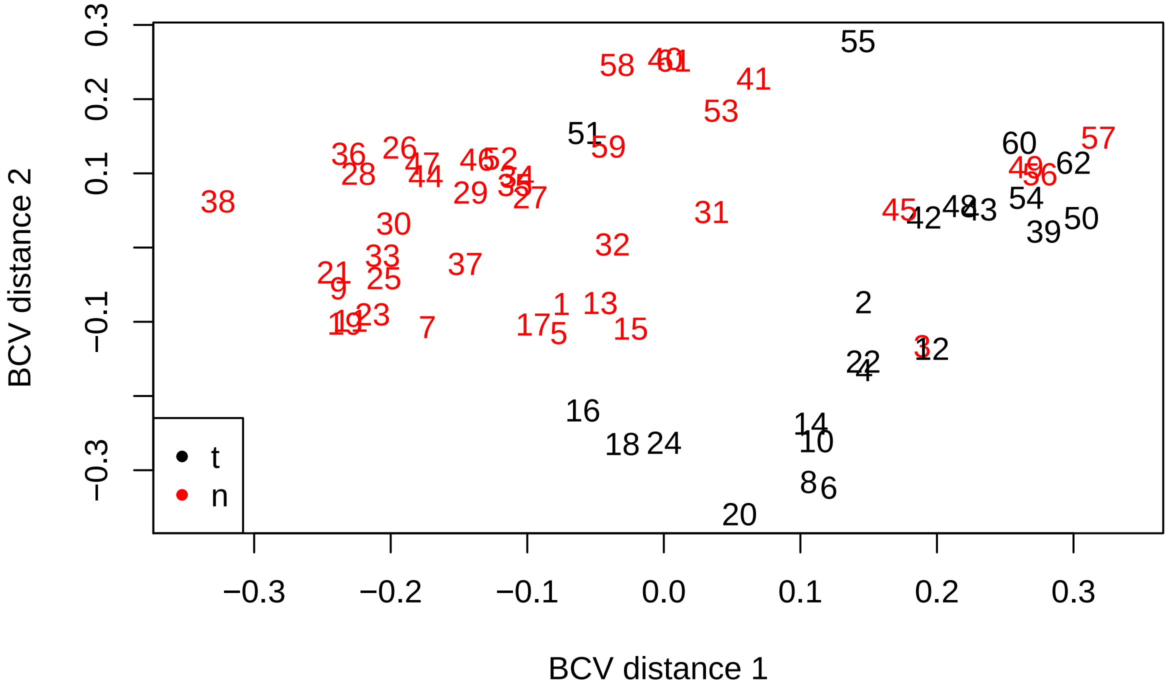
This project will use the dataset `colonCA` from the package **colonCA**, which was firtly published by Alon et al. (1999) and contains 40 tumor samples and 22 normal suamples for colon-cancer patients. These 62 sampeles were analyzed with an Affymetrix oligonucleotide Hum6000 array. The dataset is one expression set with 2000 genes and it was not preprocessed. The following is a list of covariates in this dataset:

Variables	Description	Values
expNr	Number of sample	
samp	Sample code	
class	Tissue identity	n:normal tissue; t:tumor tissue

The dataset is filtered before analysis. The genes which did not occure frequently were removed and those left must have a count per million (cpm) of 100 or greater for at least 25 samples. After filtering, there is little information lost and the number of genes is reduced to 1862.

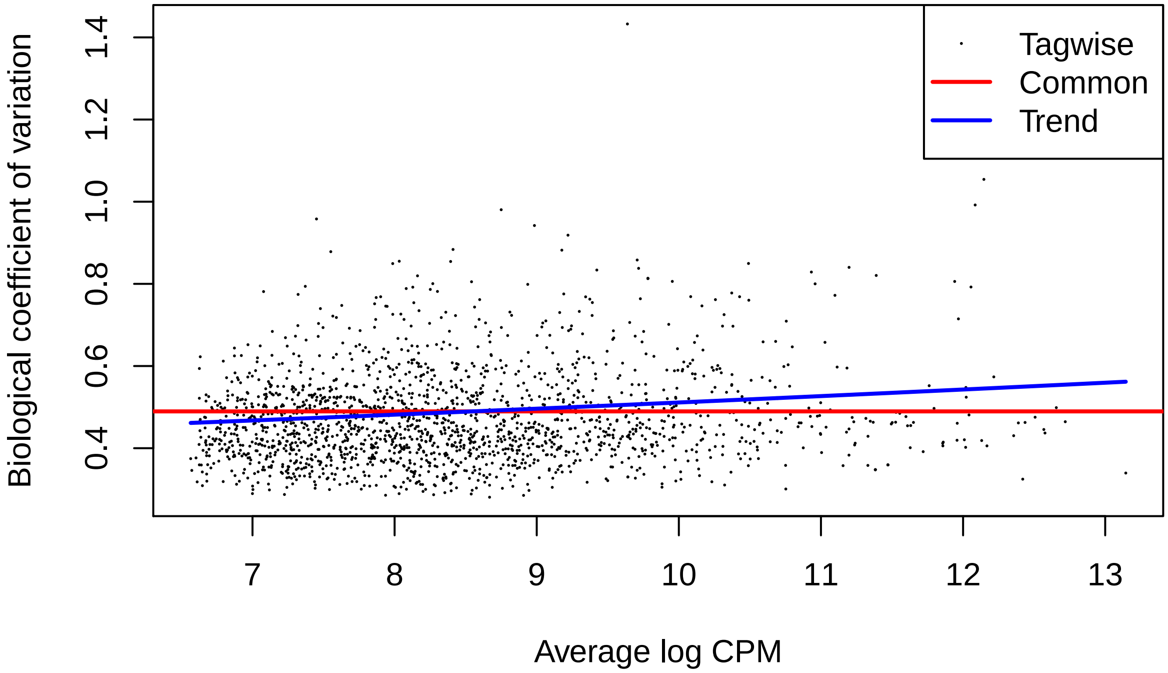
Results

Multidimensional Scaling Analysis



The multidimensional scaling graph is used to provide the first idea about which samples are close to each other in figure. Black number refers to tumor tissue and red number refers to normal tissue. ALthough some samples overlap, most of them separate out nicely. It is an evidence that the gene expression might be different between people with and without tumor. Thus, it is worthwhile for us to do further study.

GLM Estimate of Dispersion



In this graph, `method="power"` is used to fit one trend line (blue line), which shows that the tagwise biological coefficient of variation has one upward sloping trend. Blue line deviates from the red line, which indicates that common estimate of dispersion with naive method is a bad model because the tagwise dispersion does not follow the trend of common dispersion. Thus, it is worthwile to do differential expression.

Differential Expression

With tagwise tests, we compared samples with tumor and without tumor, and get a list of top 10 genes that are most likely to be differentially expressed:

```
## Comparison of groups:  t-n
##                  logFC    logCPM       PValue          FDR
## Hsa.37937 -1.720314  9.268847 3.126338e-18 5.821241e-15
## Hsa.692.2 -2.204294  9.543269 7.319316e-18 6.814283e-15
## Hsa.8147  -2.050942 10.592840 5.489167e-17 3.406943e-14
## Hsa.36689 -1.205063  8.704418 4.291004e-16 1.997462e-13
## Hsa.692   -1.670233  9.995630 2.774987e-14 1.033405e-11
## Hsa.692.1 -1.621191 10.026005 1.278006e-13 3.966079e-11
## Hsa.1832  -2.104540  9.189552 3.531000e-13 9.392461e-11
## Hsa.2097  -1.383712  7.403679 3.188226e-11 7.420597e-09
## Hsa.3306   1.227152  9.628334 2.962514e-10 6.129112e-08
## Hsa.549    1.487431  8.526357 3.447391e-10 6.419041e-08
```

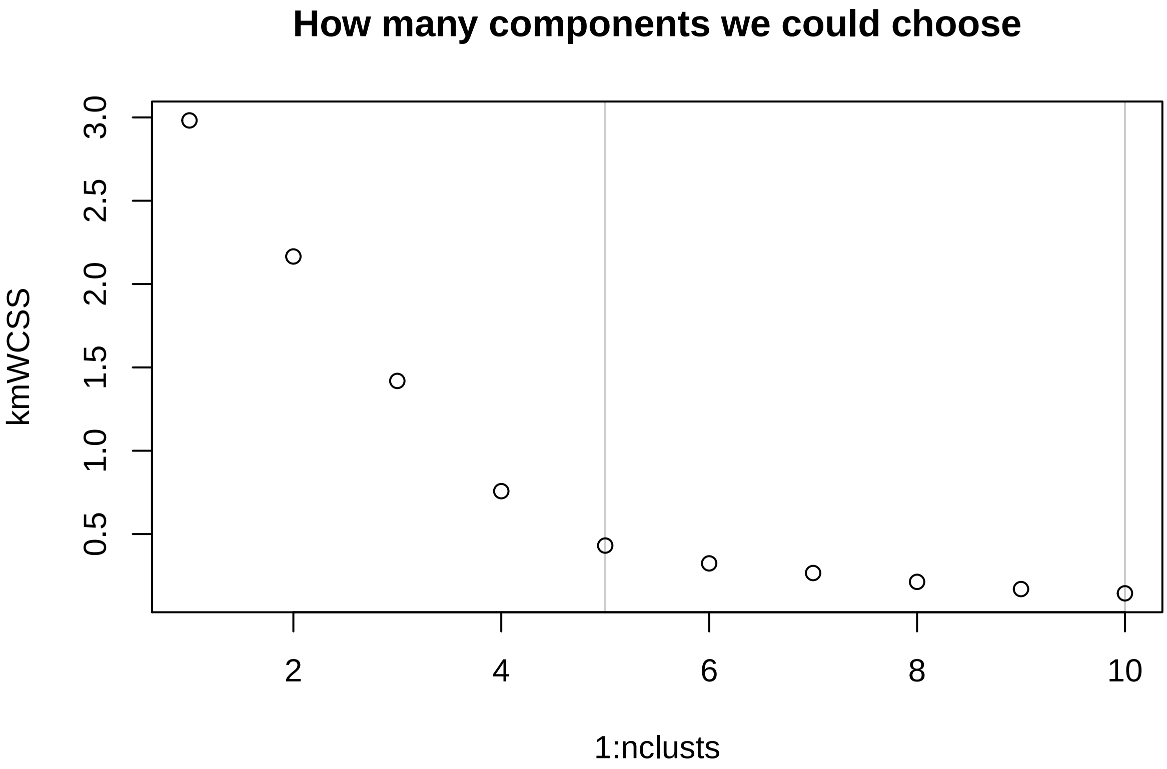
Generally speaking, cancer is caused by certain alters to genes such as genetic mutations and genetic recombination. Normal cells have much less genetic changes than cancer cells, but the combination of genetic alterations is unique for a specific cancer (2015). False Discovery Rate (FDR) is a method used to conceptualize the rate of type I error, we could find whether the gene has been apparently expressed or not by comparing its FDR with p-value. Based on the data we got above, we have that PValue < FDR for those 10 genes, which means these genes are most likely to be differentially expressed and have a relationship with colorectal cancer.

According to Yang and Zhang (2009), Hsa.37937 and Hsa.692 are the most frequently selected genes using different methods, which proves in some ways that our finding in this project is reasonable and meaningful.

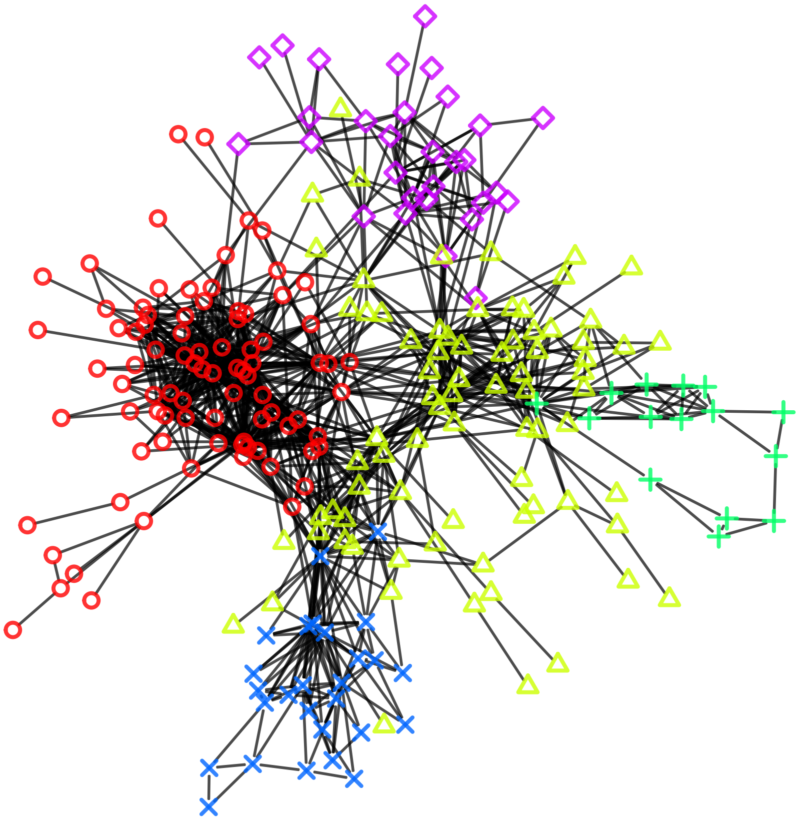
Network Analysis

```
##
##   1   2   3   4   5   9  95 206 212 236
## 926  58   6   6   4   1   1   1   1   1
```

Network analysis is going to focus on clustering and analyse whether some genes are likely to control the same trait together. The table above shows that the connected components of the corresponding graph/network, which means it has 926 isolated vertices and 1 component with 236 vertices. The following analysis will focus on the largest connected component (LCC) and check how genes are likely to work together.



Above graph is one basic step in generating the network graph. The total within sum of squares (TWSS) tells us that we could choose 5 components, and that's because 5 components could almost explain the network in our example.



Above graph tells us that LCC could be docomposed into 5 components, and each component of genes are likely to work together and control same traits. The name of genes are not labelled in this graph, which will overlap each other and make the main shape hard to see. Biologist could validate this finding by biological experiments.

References

Chan S, Griffith O, Tai I, Jones S: Meta-analysis of Colorectal Cancer Gene Expression Profiling Studies Identifies Consistently Reported Candidate Biomarkers. Cancer Epidemiol Biomarkers Prev 2008;17(3). March 2008

Shih W, Chetty R, Tsao M: Expression profiling by microarrays in colorectal cancer (Review). Oncology Reports 13: 517-524, 2005

Bertucci F, Salas S, Eysteries S, et al: Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene (2004) 23, 1377–1391

Birkenkamp-Demtroder K, Christensen L, Olesen S, et al: Gene Expression in Colorectal Cancer. Cancer Research 62, 4352– 4363, August 1, 2002

National Cancer Institute: The Genetics of Cancer, http://www.cancer.gov/about-cancer/causes-prevention/genetics , April 22, 2015

Lab 4, https://www.stanford.edu/class/bios221/labs/rnaseq/lab_4_rnaseq.html

Lab 7, https://web.stanford.edu/class/bios221/labs/networks/lab_7_networks.html

U. Alon et al: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96, 6745-6750, 1999

Yang, Pengyi and Zhang, Zili: An embedded two-layer featureselection approach for microarray data analysis, IEEE intelligent informatics bulletin, vol. 10, no. 1, pp. 24-32, Dec, 2009

Appendix

```r
#load libraries
library(colonCA)
library(Biobase)
library(edgeR)

data(colonCA)
dataset = colonCA
mat = as.matrix(exprs(dataset) )
mobDataGroups = dataset@phenoData@data$class
#n:normal tissue; t:tumor tissue

#create DGEList
d = DGEList(counts=mat,group=factor(mobDataGroups))
#filter the gene
keep <- rowSums(cpm(d)>100) >= 30
d <- d[keep,]
dim(d)
#reset library size
d$samples$lib.size = colSums(d$counts)
#normalize the data
d = calcNormFactors(d)

plotMDS(d, method = 'bcv', col = as.numeric(d$samples$group))
legend('bottomleft', as.character(unique(d$samples$group)), col = 1:2, pch =20)

design.mat = model.matrix(~ 0 + d$samples$group)
colnames(design.mat) = levels(d$samples$group)
d2 = estimateGLMCommonDisp(d,design.mat)
d2 = estimateGLMTrendedDisp(d2,design.mat, method="power")
d2 <- estimateGLMTagwiseDisp(d2,design.mat)
plotBCV(d2)
```

```r
et12 <- exactTest(d1, pair=c(1,2)) # compare groups 1 and 2
topTags(et12, n=10)


require(ade4) # multivariate analysis
require(grid) # has the viewport function

X = mat
keep= rowSums(cpm(X)>100) >= 25
X= X[keep,]
#dim(X)
XT = t(X)
cor.ecoli= cor(XT)

Sfull <- Matrix(1*(cor.ecoli > 0.9))
Scd <- component.dist(as.matrix(Sfull))
table(Scd$csize)

lcc.ind <- which(Scd$membership == which.max(Scd$csize))
S <- Sfull[lcc.ind,lcc.ind]
n <- dim(S)[1]
d <- as.vector(S %*% rep(1,n))
Lsym <- Diagonal(n) - (Diagonal(x=d^(-1/2)) %*% S %*% Diagonal(x=d^(-1/2)))
eLsym <- eigen(Lsym)
eLV <- eLsym$vectors[,n-1:3]

set.seed(2)
kmWCSS <- sum(sweep(eLV,2,apply(eLV,2,mean))^2)
nclusts <- 10
nstart <- 100
kmWCSS[2:nclusts] <- unlist(lapply(2:nclusts, function(i){
  sum(kmeans(eLV, i, nstart=nstart)$withinss)}))
plot(1:nclusts,kmWCSS)
abline(v=seq(0,nclusts,by=5),col=rgb(0,0,0,0.2))

#5 clusters.
my.nclusts <- 5
my.km <- kmeans(eLV, my.nclusts, nstart=nstart)
#LCC can be docomposed int 5 components, which may have similar function in biology an
Snet <- network(as.matrix(S), dir=F)
set.seed(1)
par(mar=rep(0,4)+0.1) # makes the margins smaller
coords <- plot(Snet, vertex.col=0, vertex.border=0, edge.col=rgb(0,0,0,0.7))
plot(Snet, coord=coords, vertex.col=0, vertex.border=0, edge.col=rgb(0,0,0,0.7))
```

```r
my.cols <- rainbow(my.nclusts, alpha=0.8)
points(coords, col=my.cols[my.km$cluster], pch=my.km$cluster, lwd=2)
#biological validation: validate they have similar function
```