Bioinformatics Lab 4, Group 3

Roshni Sundaramurthy, Prudhvi Peddmallu, Jiawei Wu, Zijie Feng

12 December 2018

Question 1

```
# install_github("seandavi/GEOquery")
library(GEOquery)
```

The GEOquery library is loaded to perform the Gene Expression methods.

```
#Get the Gene Expression Omnibus (GEO) data
x = getGEOSuppFiles("GSE20986")
x
```

```
##                                                              size isdir mode
## C:/Users/fengy/Desktop/lab4/GSE20986/GSE20986_RAW.tar 56360960 FALSE  666
##                                                                    mtime
## C:/Users/fengy/Desktop/lab4/GSE20986/GSE20986_RAW.tar 2018-12-13 23:13:03
##                                                                    ctime
## C:/Users/fengy/Desktop/lab4/GSE20986/GSE20986_RAW.tar 2018-12-13 23:12:54
##                                                                    atime
## C:/Users/fengy/Desktop/lab4/GSE20986/GSE20986_RAW.tar 2018-12-13 23:12:54
##                                                                   exe
## C:/Users/fengy/Desktop/lab4/GSE20986/GSE20986_RAW.tar  no
```

Using getGEOSuppFiles (Get Supplemental Files from GEO), the supplemental files based on the GEO accession number (*GSE20986*) is obtained. And it is a dataframe containing 1 object of 7 variables.

```
# untarring the data
untar("GSE20986/GSE20986_RAW.tar", exdir = "data")

# gunzipping the data
cels = list.files("data/", pattern = "[gz]")
sapply(paste("data", cels, sep = "/"), gunzip)
```

```
## data/GSM524662.CEL.gz data/GSM524663.CEL.gz data/GSM524664.CEL.gz
##              13555726              13555055              13555639
## data/GSM524665.CEL.gz data/GSM524666.CEL.gz data/GSM524667.CEL.gz
##              13560122              13555663              13557614
## data/GSM524668.CEL.gz data/GSM524669.CEL.gz data/GSM524670.CEL.gz
##              13556090              13560054              13555971
## data/GSM524671.CEL.gz data/GSM524672.CEL.gz data/GSM524673.CEL.gz
##              13554926              13555042              13555290
```

The file is extracted and the contents of a tar archive is listed using untar function. The data folder has been created. The list.files function produces a character vector of the names of files.

```
# creating your phenodata
phenodata <- matrix(rep(list.files("data"), 2), ncol =2)
class(phenodata)
```

```
## [1] "matrix"
```

Now importing "phenotype" data, describing the experimental design of our dataset. It is of type matrix.

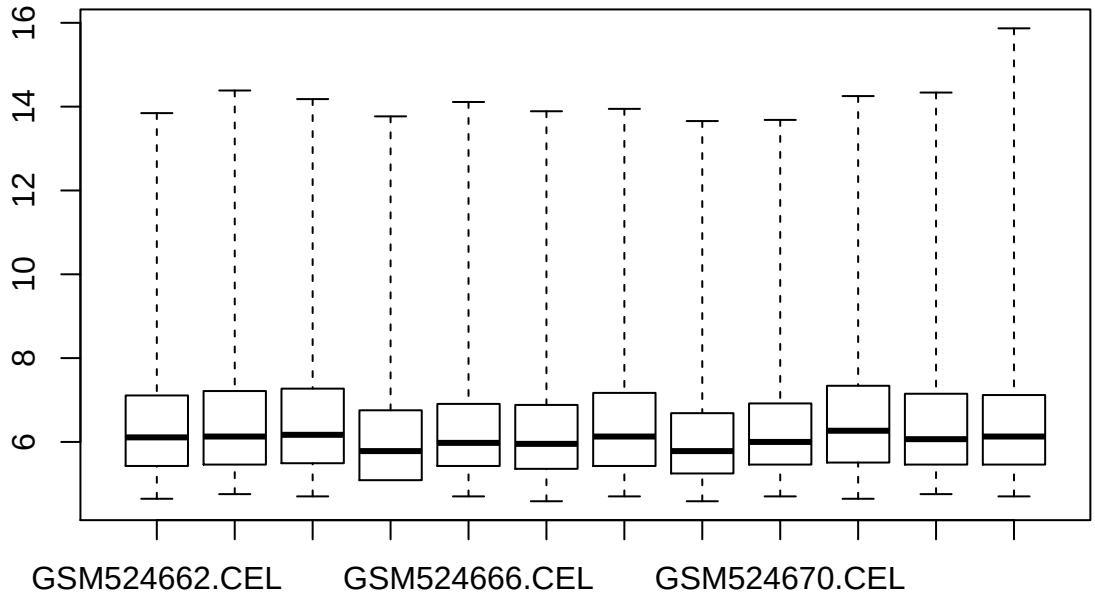
```
phenodata <- as.data.frame(phenodata)
colnames(phenodata) <- c("Name", "FileName")
phenodata$Targets <- c("iris",
                       "retina",
                       "retina",
                       "iris",
                       "retina",
                       "iris",
                       "choroid",
                       "choroid",
                       "choroid",
                       "huvec",
                       "huvec",
                       "huvec")
write.table(phenodata, "data/phenodata.txt", quote = F, sep = "\t", row.names = F)
```

The column names of phenodata has been changed. A new variable "Target" is created consisting of iris, retina,choroid and huvec. The phenodata now consists of 12 objects with 3 variables and it is saved as a text file in data folder for further use.

```
# source("https://bioconductor.org/biocLite.R")
# biocLite("simpleaffy")
library(simpleaffy)
```

Loading the required package *simpleaffy* along with genefilter.

```
celfiles <- read.affy(covdesc = "phenodata.txt", path = "data")
boxplot(celfiles)
```




```
# the median and quartiles are similar in different examples
```

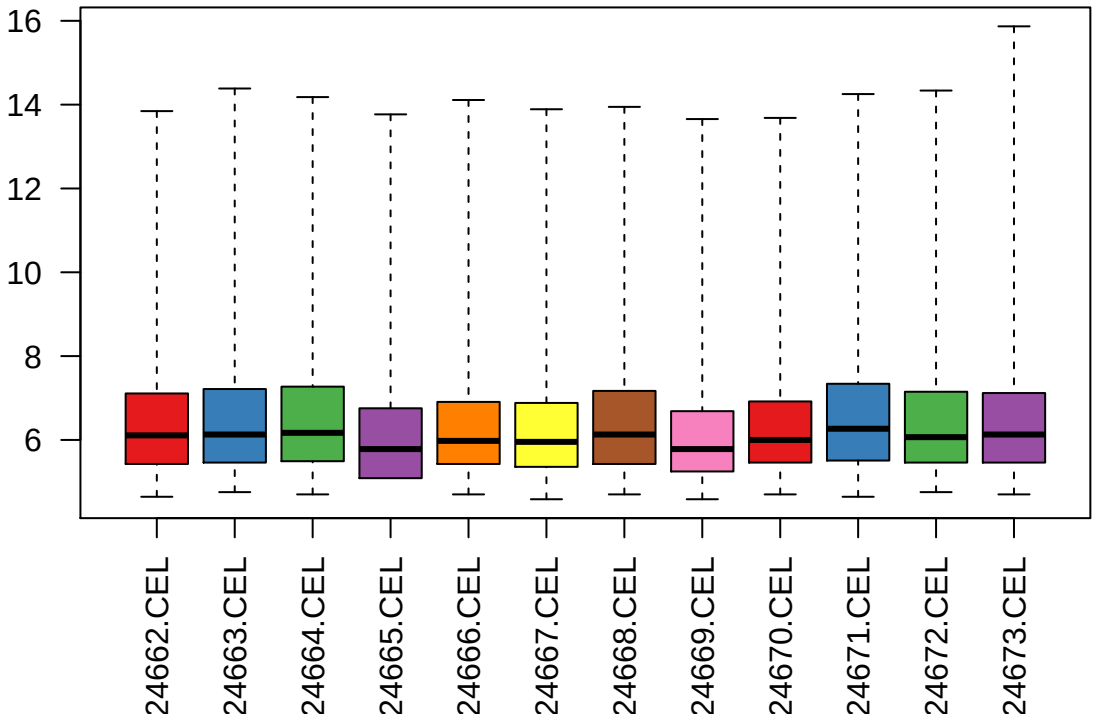
The text file phenodata which defines phenotypic data for a set of .CEL files is read using read.affy to create the AffyBatch object. Now the size of arrays is 1164x1164 features (23 kb), number of samples=12 and number of genes=54675. The boxplots enable us to study the distributional characteristics of a group. From the plot, it is noted that only names of the three tissues are displayed and they are GSM524662.CEL for iris, GSM524666.CEL for retina and GSM524670.CEL for choroid. It is quite hard to find relationship between every objects since it is not well displayed with object names (specified cells).

```
library(RColorBrewer)
cols = brewer.pal(8, "Set1")
eset <- exprs(celfiles)
samples <- celfiles$Targets
colnames(eset)
```

```
##  [1] "GSM524662.CEL" "GSM524663.CEL" "GSM524664.CEL" "GSM524665.CEL"
##  [5] "GSM524666.CEL" "GSM524667.CEL" "GSM524668.CEL" "GSM524669.CEL"
##  [9] "GSM524670.CEL" "GSM524671.CEL" "GSM524672.CEL" "GSM524673.CEL"
```

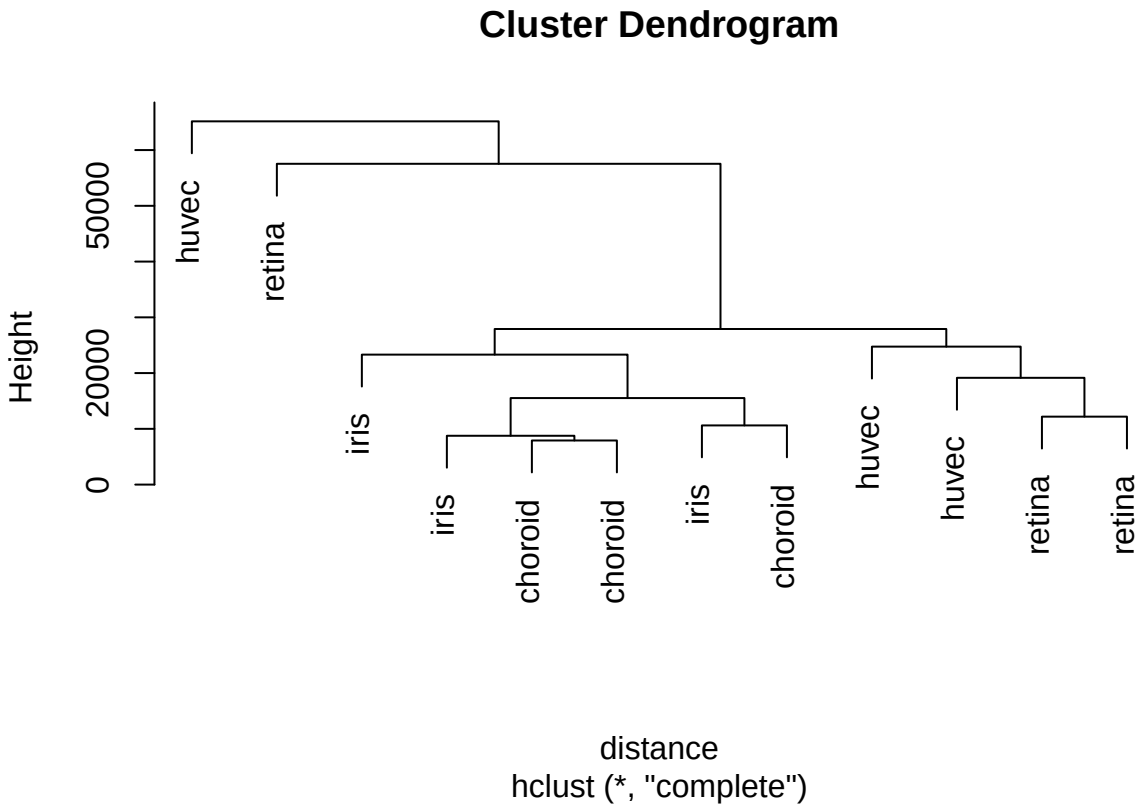
To overcome the above discomfort, the good colour palettes for thematic maps have been created. The generic function *exprs* retrieves the expression data from eSets and the result is a large matrix.

```
colnames(eset) <- samples
boxplot(celfiles, col = cols, las = 2)
```



The names in the samples has been assigned to eset column names for easy view. The colourful boxplot for our data is created. Now the plot is clear enough to find and analyse various cells from 4 tissues of our dataset. The median marks the mid-point of the data and is shown by the line that divides the box into two parts. The upper whisker seems to be longer than the lower whisker. It denotes that the upper whisker stretched over the wide range of values.

```
distance <- dist(t(eset), method = "maximum")
clusters <- hclust(distance)
plot(clusters)
```



The distances between the rows of a data matrix eset has been measured by using maximum distance measure. Then performing the hierarchical cluster analysis for several objects being clustered. The cluster dendogram is plotted with heights in y axis and distance in x axis. The height ranges from 0-60000. The distance is basically the dissimilarities between the clusters. Form the plots, we can observe 2 big clusters. One comprises of iris and choroid, other comprises of huvec and retina. Unfortunately, two outliers huvec and retina are observed.

```
require(simpleaffy)
# devtools::install.github("bmbolstad/affyPLM")
require(affyPLM)
```

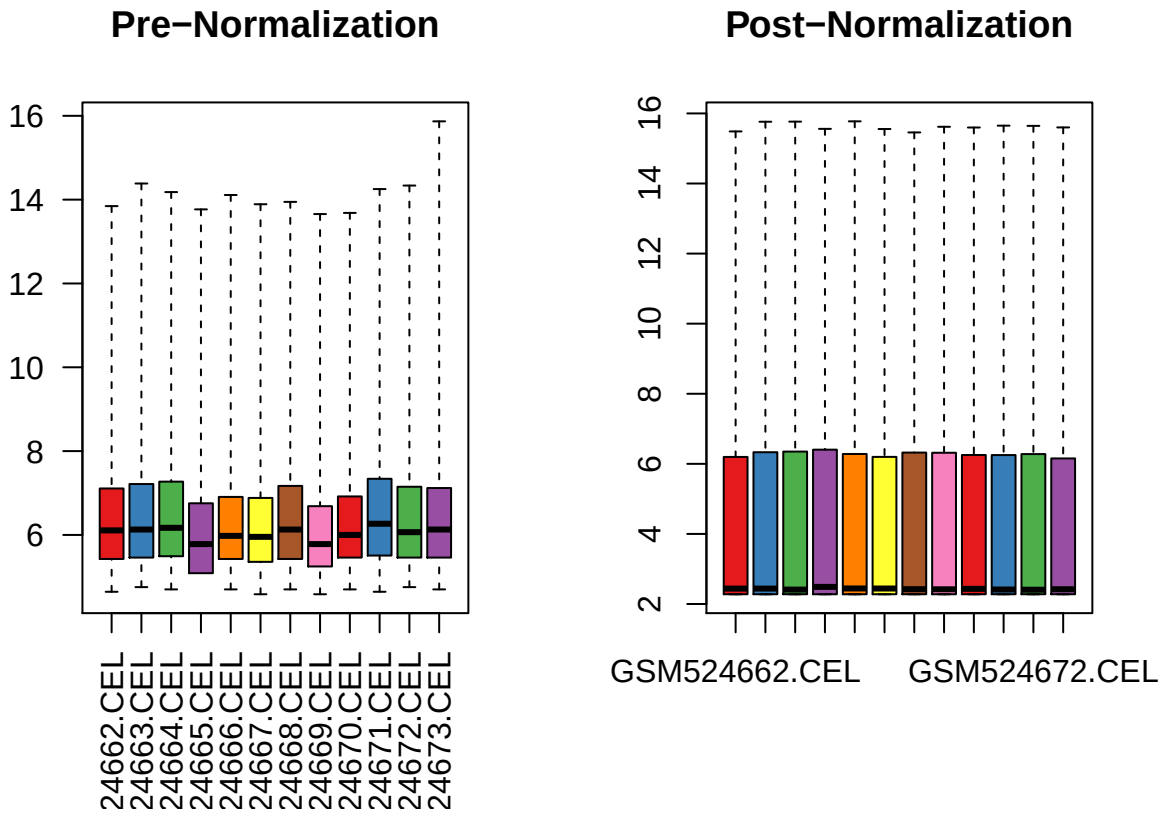
Using the required packages simpleaffy anf affyPLM.

```
celfiles.gcrma = gcrma(celfiles)
```

```
## Adjusting for optical effect.....Done.
## Computing affinities.Done.
## Adjusting for non-specific binding.....Done.
## Normalizing
## Calculating Expression
```

This gcrma function converts an AffyBatch into an ExpressionSet using the robust multi-array average (RMA) expression measure with help of probe sequence. First the affinities have been calculated. Then normalizing and calculating expression.

```
par(mfrow=c(1,2))
boxplot(celfiles, col = cols, las = 2, main = "Pre-Normalization")
boxplot(celfiles.gcrma, col = cols,  main = "Post-Normalization")
```



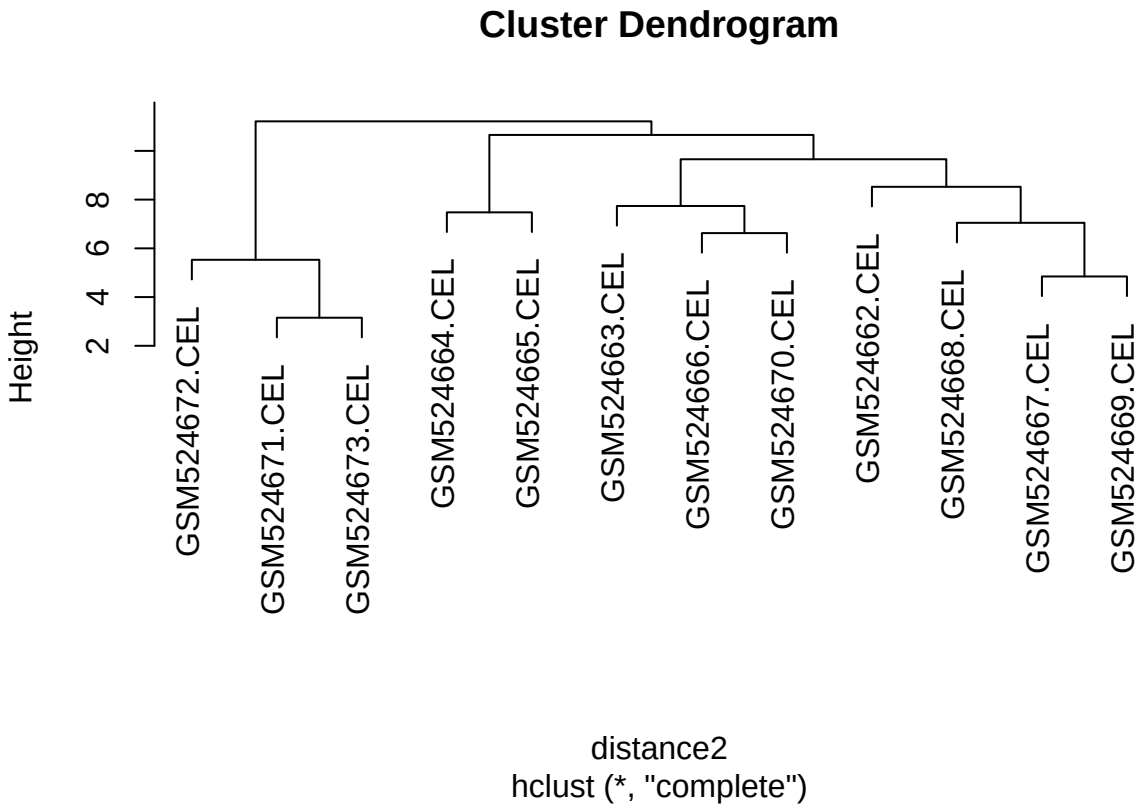
Using par() to create 1 x 2 pictures on one plot. Before applying normalization technique, it seems that the upper whiskers for all objects are uneven and also the 4th box (violet) named 24665.CEL has no lower whisker. After normalization, all boxes seems to have equal length and upper whiskers are evenly distributed with minute height variations. The median seems to be appear at the bottom of the boxes.

```
dev.off()
```

```
## null device
##           1
```

This functions provide control over multiple graphics devices.

```
# Cluster Dendrogram based on post-normalization
eset2 <- exprs(celfiles.gcrma)
# colnames(eset2) <- samples
distance2 <- dist(t(eset2), method = "maximum")
clusters <- hclust(distance2)
plot(clusters)
```



Again the distance is measured for the normalized data and the cluster dendogram is represented. But here the height ranges from 2-10.

```
library(limma)   #loading package
```

```
phenodata      # displaying phenodata dataset
```

```
##             Name      FileName Targets
## 1  GSM524662.CEL GSM524662.CEL    iris
## 2  GSM524663.CEL GSM524663.CEL  retina
## 3  GSM524664.CEL GSM524664.CEL  retina
## 4  GSM524665.CEL GSM524665.CEL    iris
## 5  GSM524666.CEL GSM524666.CEL  retina
## 6  GSM524667.CEL GSM524667.CEL    iris
## 7  GSM524668.CEL GSM524668.CEL choroid
## 8  GSM524669.CEL GSM524669.CEL choroid
## 9  GSM524670.CEL GSM524670.CEL choroid
## 10 GSM524671.CEL GSM524671.CEL   huvec
## 11 GSM524672.CEL GSM524672.CEL   huvec
## 12 GSM524673.CEL GSM524673.CEL   huvec
```

```
samples <- as.factor(samples)
design <- model.matrix(~0+samples)
colnames(design)
```

```
## [1] "sampleschoroid" "sampleshuvec"   "samplesiris"    "samplesretina"
```

model.matrix creates a design matrix, e.g., by expanding factors to a set of dummy variables (depending on

the contrasts) and expanding interactions similarly. The *design* matrix consists of 0s and 1s for all 4 features.

```
colnames(design) <- c("choroid", "huvec", "iris", "retina")
design
```

```
##    choroid huvec iris retina
## 1        0     0    1      0
## 2        0     0    0      1
## 3        0     0    0      1
## 4        0     0    1      0
## 5        0     0    0      1
## 6        0     0    1      0
## 7        1     0    0      0
## 8        1     0    0      0
## 9        1     0    0      0
## 10       0     1    0      0
## 11       0     1    0      0
## 12       0     1    0      0
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$samples
## [1] "contr.treatment"
```

Changing the column names of design matrix for proper data.

```
contrast.matrix = makeContrasts(
              huvec_choroid = huvec - choroid,
              huvec_retina = huvec - retina,
              huvec_iris <- huvec - iris,
              levels = design)

fit = lmFit(celfiles.gcrma, design)
huvec_fit <- contrasts.fit(fit, contrast.matrix)
huvec_ebay <- eBayes(huvec_fit)
```

Using makeContrasts function from limma package. It construct the contrast matrix corresponding to specified contrasts of a set of parameters. The specified contrasts are huvec_choroid, huvec_retina, huvec_iris. Fitting linear model for normalized data. The estimated coefficients and standard errors for a given set of contrasts are computed from linear model fit. Using eBayes method, the moderated t-statistics, moderated F-statistic, and log-odds of differential expression, p-values are computed.

```
# if (!requireNamespace("BiocManager", quietly = TRUE))
#     install.packages("BiocManager")
# BiocManager::install("hgu133plus2.db", version = "3.8")
# source("https://bioconductor.org/biocLite.R")
# biocLite("hgu133plus2.db")
library(hgu133plus2.db)
```

RNA extracts from endothelial cells were hybridised to Affymetrix HGU133plus2 arrays in triplicate.

```
#biocLite("annotate")
library(annotate)
```

```
probenames.list <- rownames(topTable(huvec_ebay, number = 100000))
getsymbols <- getSYMBOL(probenames.list, "hgu133plus2")
results <- topTable(huvec_ebay, number = 100000, coef = "huvec_choroid")
```

```
results <- cbind(results, getsymbols)
```

The table of the top-ranked genes from a linear model fit is extracted by specifying 100000 maximum number of genes to list. The rownames of this is assigned to probenames.list. Mapping the set of manufacturers identifiers to other identifiers using getSYMBOL().

The statistical values such as logFC, p-value, t, adjusted p values for all top ranked genes for huvec choroid pair are computed and stored in results.

```
summary(results) #To make thresholds
```

```
##      logFC              AveExpr              t                 P.Value
##  Min.   :-9.19111   Min.   : 2.279   Min.   :-39.77473   Min.   :0.0000
##  1st Qu.:-0.05967   1st Qu.: 2.281   1st Qu.: -0.70649   1st Qu.:0.1523
##  Median : 0.00000   Median : 2.480   Median :  0.00000   Median :0.5079
##  Mean   :-0.02353   Mean   : 4.375   Mean   :  0.07441   Mean   :0.5346
##  3rd Qu.: 0.03986   3rd Qu.: 6.241   3rd Qu.:  0.67455   3rd Qu.:1.0000
##  Max.   : 8.67086   Max.   :15.541   Max.   :296.84201   Max.   :1.0000
##
##    adj.P.Val            B             getsymbols
##  Min.   :0.0000   Min.   :-7.710   YME1L1  :   22
##  1st Qu.:0.6036   1st Qu.:-7.710   HFE     :   15
##  Median :1.0000   Median :-7.451   CFLAR   :   14
##  Mean   :0.7436   Mean   :-6.582   NRP2    :   14
##  3rd Qu.:1.0000   3rd Qu.:-6.498   ARHGEF12:   13
##  Max.   :1.0000   Max.   :21.290   (Other) :41859
##                                    NA's    :12738
```

```
results$threshold <- "1"
a <- subset(results, adj.P.Val < 0.05 & logFC > 5)
results[rownames(a), "threshold"] <- "2"
b <- subset(results, adj.P.Val < 0.05 & logFC < -5)
results[rownames(b), "threshold"] <- "3"
table(results$threshold)
```

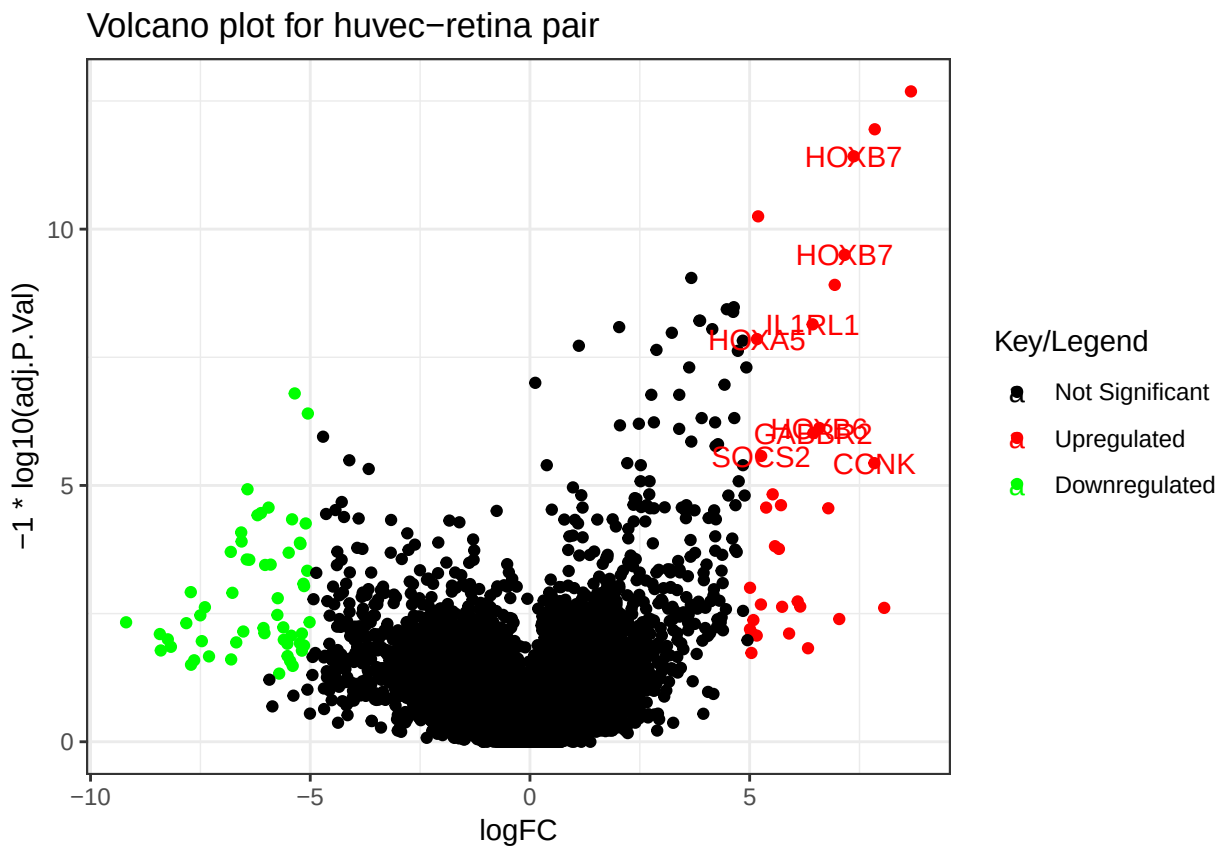
```
##
##     1     2     3
## 54587    33    55
```

Adding threshold column to the *results* dataframe. Subsetting (filtering) the data based on specified conditions such as * adj.P.Val < 0.05 & logFC > 5 *. Changing the threshold values for these specific conditioned data as "2". Again subsetting the datas and changing threshold value as 3 for these datas. The table shows the threshold values such that, number of objects having threshold as 1 is higher than the other 2 threshold values.

```
library(ggplot2)
volcano <- ggplot(data = results,
                  aes(x = logFC, y = -1*log10(adj.P.Val),
                      colour = threshold,
                      label = getsymbols))

volcano <- volcano +
  geom_point() +
  scale_color_manual(values = c("black", "red", "green"),
                     labels = c("Not Significant", "Upregulated", "Downregulated"),
                     name = "Key/Legend")
```

```
volcano +
  geom_text(data = subset(results, logFC > 5 & -1*log10(adj.P.Val) > 5), aes(x = logFC, y = -1*log10(ad
  ggtitle("Volcano plot for huvec-retina pair")+theme_bw()
```



The volcano plot has been created using ggplot(). It plots significance versus fold-change on the y and x axes, respectively. Fold change (x axis) is plotted against statistical significance (y axis) for each set. Genes upregulated with a fold change >=5 and p<0.05 are depicted in red, and those downregulated with a fold change < -5 and p<0.05 are shown in green. Black represents genes in the arrays that were not found to differ significantly.

```
paste("Total genes for huvec-choroid pair:")
```

```
## [1] "Total genes for huvec-choroid pair:"
```

```
table(results$threshold)
```

```
##
##     1     2     3
## 54587    33    55
```

When threshold=2, 33 genes (Red) are differentially expressed ones and When threshold=3, 55 (green) genes are differentially expressed ones. So, totally *88 genes* are classified as differentially expressed genes.

Question 2

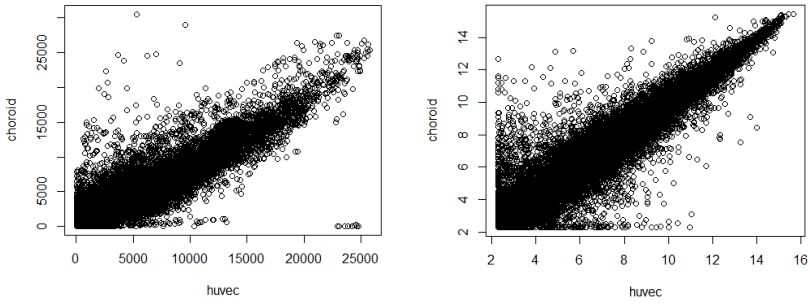
The three contrasts are as follows,

1. huvec_choroid = huvec - choroid
2. huvec_retina = huvec - retina
3. huvec_iris = huvec - iris

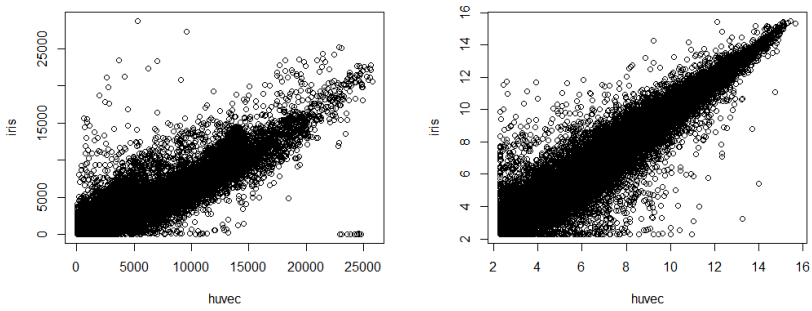
Present the variables versus each other original variables

All the plots on the left are built on the original data, and all the plots on the right are built on the normalized data

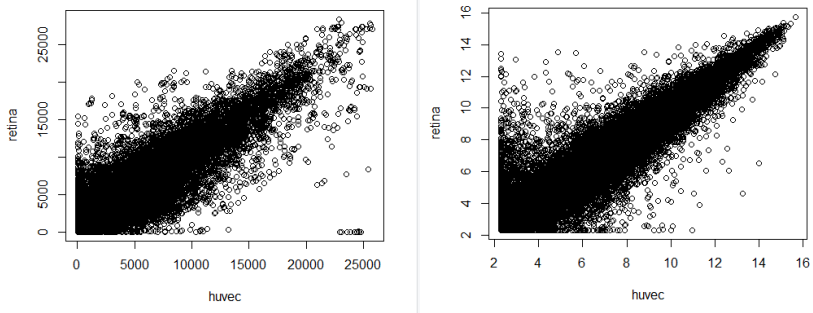
```
## [1] 1 2
```




```
## [1] 3 4
```

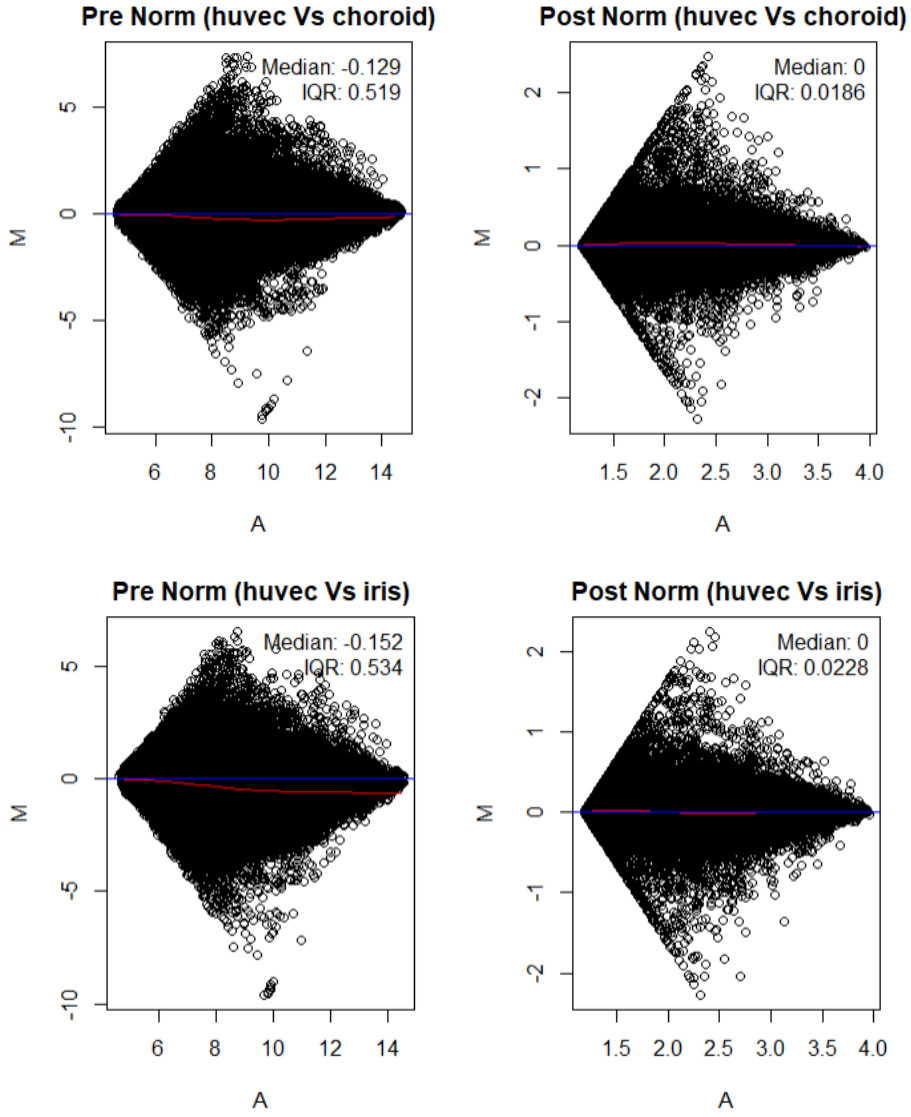


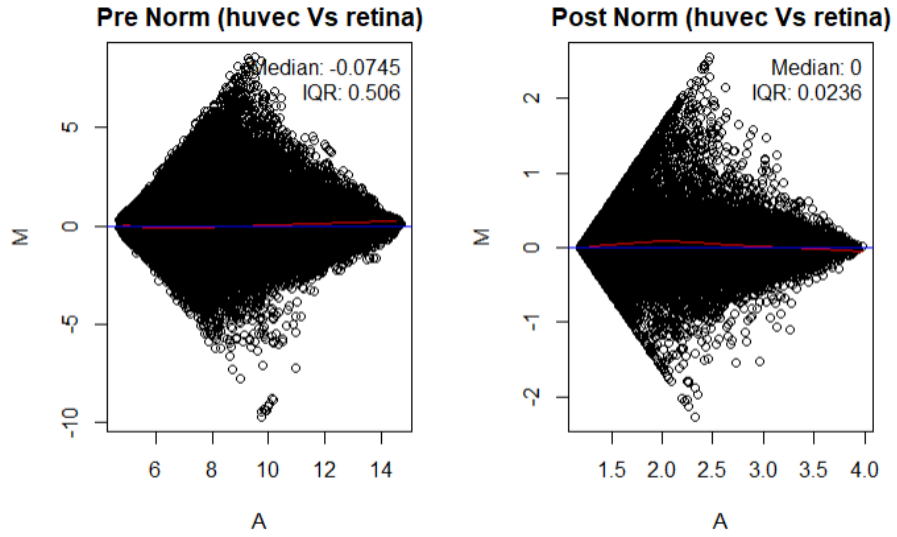

```
## [1] 5 6
```



MA plots

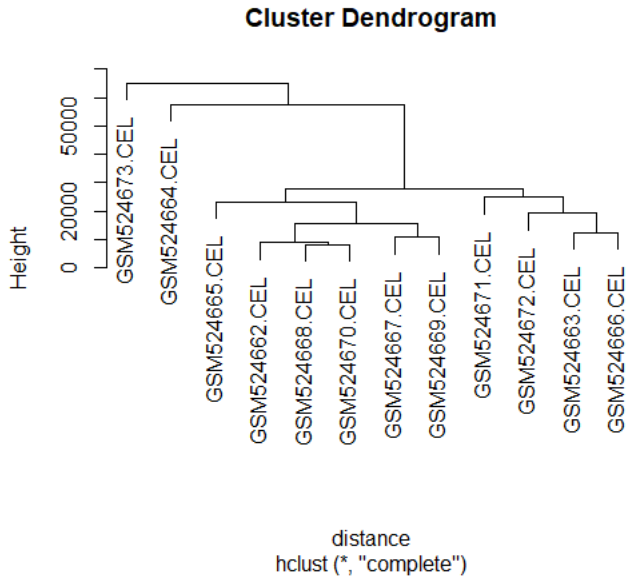
(The first is pre normalized and the second is post normalized)

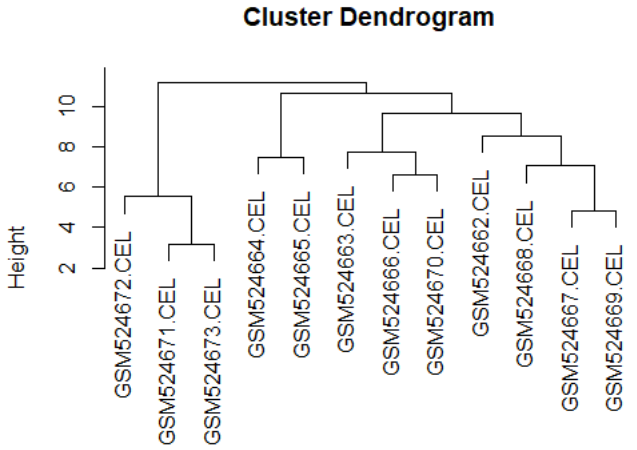


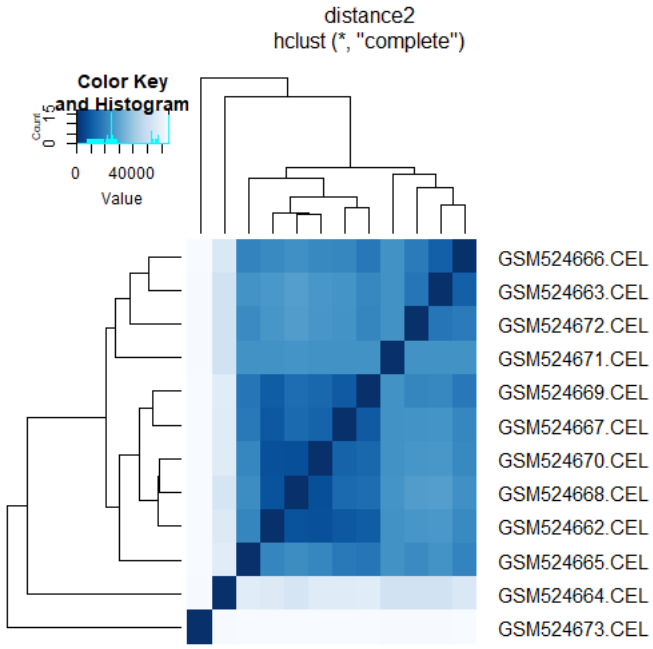


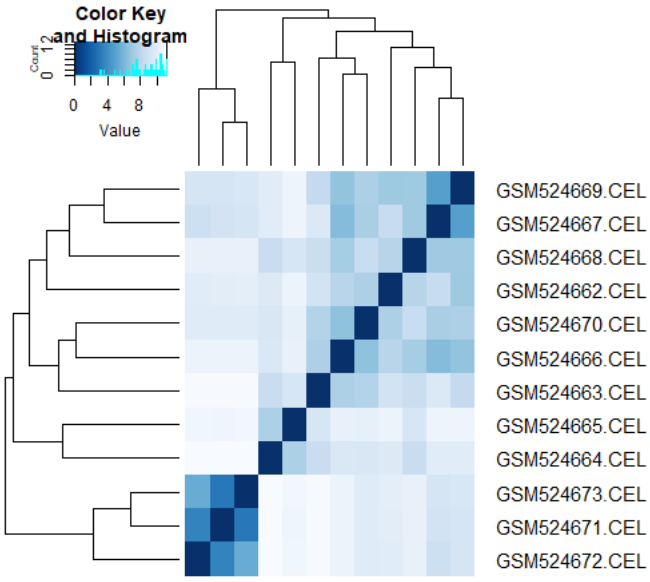
It seems that all the MA plots are symmetric on the mean of normalized counts (x-axis) for most of points. We can also observe that the lines in the plots refer to the medians and IQR values. According to such six plots, the medians of all the post normalization are 0, which are straight lines in plots and are bigger than the medians of pre normalization. Then the inter quantile ranges (IQR) of post normalization are smaller than the ones of pre normalization. This might be the reason why the red and blue are nearly collide in the post normalization.

Cluster Dendrograms and Heat maps




Cluster Dendrogram
Height
2
4
6
8
10
GSM524672.CEL
GSM524671.CEL
GSM524673.CEL
GSM524664.CEL
GSM524665.CEL
GSM524663.CEL
GSM524666.CEL
GSM524670.CEL
GSM524662.CEL
GSM524668.CEL
GSM524667.CEL
GSM524669.CEL

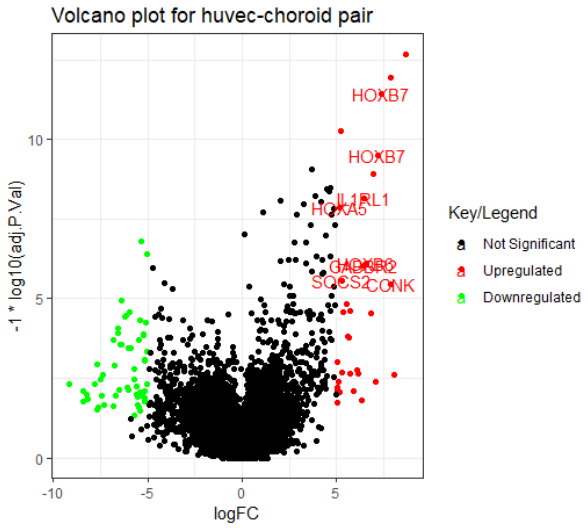

distance2
hclust (*, "complete")
Color Key
and Histogram
Count
0
15
0
40000
Value
GSM524666.CEL
GSM524663.CEL
GSM524672.CEL
GSM524671.CEL
GSM524669.CEL
GSM524667.CEL
GSM524670.CEL
GSM524668.CEL
GSM524662.CEL
GSM524665.CEL
GSM524664.CEL
GSM524673.CEL

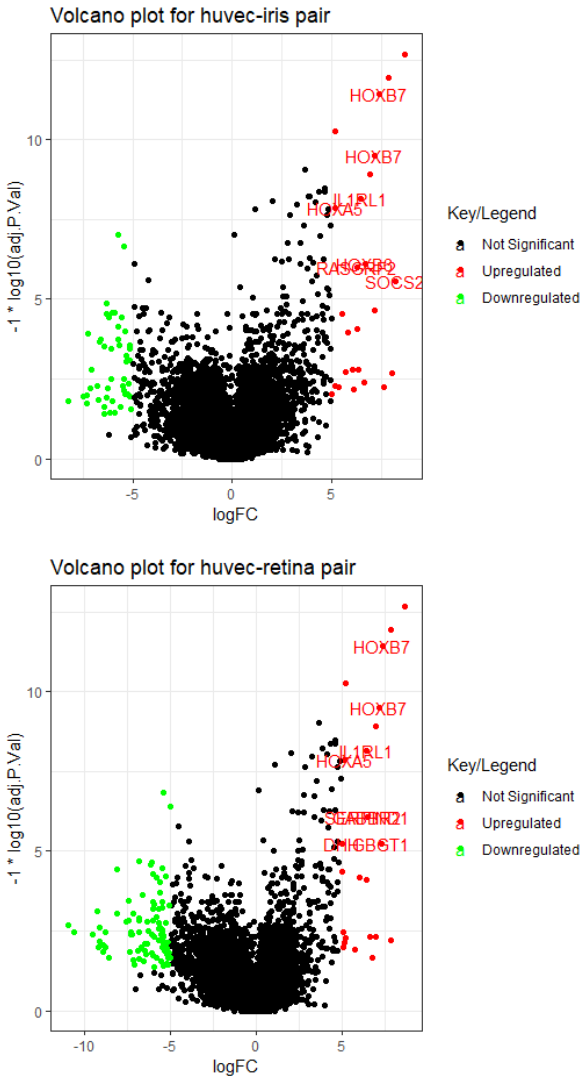


In first heat map, we can see a cluster range including GSM524669, GSM524667, GSM524670, GSM524668 and GSM524662 with dark blue in color and also see the cluster on the right top including GSM524666 and GSM524663. The big cluster represents a lot of relations between the genes. In second heat map we can see a small cluster range containing GSM524673, GSM524672 and GSM524671 with dark blue on the left bottom. And we can see a small cluster including GSM524669 and GSM524667 on the right top.

In the cluster dendrogram we can see these clusters hierarchically, but heat map can provide us more details among each pair of two genes. Additionally, the pre normalized cluster dendrogram shows us more relations among different genes except GSM524664 and GSM524673. However, most of such relation disappears in post dendrogram, that is why most of parts are colored in light blue here.

Question 3





The volcano plots of the differentially expressed genes. Differentially expressed genes were treated with red dots (up-regulated) or green dots (down-regulated), others indicated with blue dots. The red dots are for threshold = 2 and the green dots are for threshold = 3.

Volcano plot of huvec Vs Retina

Significantly differentially expressed genes were observed as HOXB7, HOXA5, SOCS2, HOXB6, IL1RL1, DHH, GBGT1. Total genes for huvec-retina pair:

1	2	3
54557	24	94

Volcano plot of huvec Vs Iris

Significantly differentially expressed genes were observed as HOXB7, HOXA5, SOCS2, HOXB6, IL1RL1, RASORP2. Total genes for huvec-iris pair:

1	2	3
54601	25	49

Question 4

Reporting all the Gene Ontology (GO) terms associated with each gene and describing them:

Gene 1: *HOXB7*

Official Symbol : HOXB7
Official Full Name : homeobox B7
Other names : HOX2; HOX2C; HHO.C1; Hox-2.3
Summary : This gene is a member of Antp homeobox family and encodes a protein with a homeobox DNA-binding domain. It is included in a cluster of homeobox B genes located on chromosome 17. The encoded nuclear protein functions as a sequence-specific transcription factor which is involved in cell proliferation and differentiation. The increased expression of this gene can result in some cases of melanoma and ovarian carcinoma.

GO terms:

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0000978	RNA polymerase II proximal promoter sequence-specific DNA binding	IDA	8756643
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	ISA	
GO:0001077	transcriptional activator activity, RNA polymerase II proximal promoter sequence-specific DNA binding	IDA	8756643
GO:0003677	DNA binding	IEA	
GO:0003700	DNA binding transcription factor activity	NAS	1678287

Gene 2: *SOCS2*

Official Symbol : SOCS2
Official Full Name : suppressor of cytokine signaling 2
Other names : CIS2; SSI2; Cish2; SSI-2; SOCS-2; STATI2
Summary : The SOCS2 gene encodes a member of the suppressor of cytokine signaling (SOCS) family. This family members are cytokine-inducible negative regulators of cytokine receptor signaling via the Janus kinase/signal transducer and activation of transcription pathway (the JAK/STAT pathway). These proteins interact with major molecules of signaling complexes to block further signal transduction by proteasomal depletion of receptors or signal-transducing proteins via ubiquitination. This gene has pseudogenes on chromosomes 20 and 22. Alternative splicing results in multiple transcript variants.

GO terms:

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0004860	protein kinase inhibitor activity	IBA	
GO:0005070	SH3/SH2 adaptor activity	TAS	9344848
GO:0005131	growth hormone receptor binding	NAS	12135564
GO:0005159	insulin-like growth factor receptor binding	IPI	9727029
GO:0005515	protein binding	IPI	11781573

Gene 3: *HOXA5*

Official Symbol : HOXA5
Official Full Name : homeobox A5
Other names : HOX1; HOX1C; HOX1.3
Summary : The genes encoding the class of transcription factors called homeobox genes are found in clusters named A, B, C, and D on four separate chromosomes. Expression of these proteins is spatially and temporally regulated during embryonic development. This gene is part of the A cluster on chromosome 7 and encodes a DNA-binding transcription factor which may regulate gene expression, morphogenesis, and differentiation. Methylation of this gene may result in the loss of its expression and, since the encoded protein upregulates the tumor suppressor p53, this protein may play an important role in tumorigenesis.

GO terms:

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0000978	RNA polymerase II proximal promoter sequence-specific DNA binding	IDA	10879542
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	NAS	19274049
GO:0001077	transcriptional activator activity, RNA polymerase II proximal promoter sequence-specific DNA binding	IDA	10879542
GO:0003677	DNA binding	IDA	8657138
GO:0003700	DNA binding transcription factor activity	IDA	10879542

Gene 4: *IL1RL1*

Official Symbol : IL1RL1
Official Full Name : interleukin 1 receptor like 1
Other names : T1; ST2; DER4; ST2L; ST2V; FIT-1; IL33R
Summary : This gene is a member of the interleukin 1 receptor family. Studies of the similar gene in *mouse* suggested that this receptor can be induced by proinflammatory stimuli, and may be involved in the function of helper T cells. This gene, interleukin 1 receptor, type I (IL1R1), interleukin 1 receptor, type II (IL1R2) and interleukin 1 receptor-like 2 (IL1RL2) form a cytokine receptor gene cluster in a region mapped to chromosome 2q12. Alternative splicing of this gene results in multiple transcript variants.

GO terms:

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0002113	interleukin-33 binding	IEA	
GO:0002114	interleukin-33 receptor activity	IEA	
GO:0004896	cytokine receptor activity	TAS	10191101
GO:0004908	interleukin-1 receptor activity	IEA	
GO:0005057	obsolete signal transducer activity, downstream of receptor	TAS	

Gene 5: *GBGT1*

Official Symbol : GBGT1
Official Full Name : globoside alpha-1,3-N-acetylgalactosaminyltransferase 1 (FORS blood group)
Other names : FS; A3GALNT; UNQ2513
Summary : This gene encodes a glycosyltransferase that is significant for the synthesis of Forssman glycolipid (FG), a member of the globoseries glycolipid family. Glycolipids such as FG form attachment sites for the binding of pathogens to cells. The expression of this protein may determine host tropism to microorganisms. Alternative splicing results in multiple transcript variants.

GO terms:

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0016740	transferase activity	IEA	
GO:0016757	transferase activity, transferring glycosyl groups	IBA,IEA	
GO:0016758	transferase activity, transferring hexosyl groups	IEA	
GO:0046872	metal ion binding	IEA	
GO:0047277	globoside alpha-N-acetylgalactosaminyltransferase activity	IBA	

From the above description, it is observed that the genes *HOXB7* and *HOXA5* have same GO IDs and GO terms. Both of them are Homeobox protein, but the evidence and pubmed IDs of these genes differ. And also, The GBGT1 has no pubmed IDs for any of its GO IDs.