Microarray Data Analysis

Alireza Fathian

Contents

Introduction **1**

Setting up project directory **1**

Import libraries **1**

Load Data **2**

Dataset Description **2**

Extracting Matrices **3**

Correlation HeatMap **6**

Principal Component Analysis **8**

Differential Gene Expression **16**

Introduction

Setting up project directory

```
knitr::opts_knit$set(root.dir = '/home/alireza/Scripts/microarray_data_analysis')
```

Import libraries

```
library(limma)
library(Biobase)
library(GEOquery)
library(pheatmap)
library(ggplot2)
library(plyr)
library(pheatmap)
```

Load Data

Importing GSE52509_series_matrix.txt.gz

```
# import existing data
print(datadir)
```

```
## [1] "/home/alireza/Scripts/microarray_data_analysis/data/raw/GSE52509_series_matrix.txt.gz"
```

```
gset=getGEO(filename=datadir, GSEMatrix = TRUE, AnnotGPL = TRUE)
```

Dataset Description

```
header=gset[[1]]
print(header)
```

```
##                                                                                  V2
## Lung tissue from cigarette smoke-treated mice at 4 months of age, biological replicate 1
##                                                                                  V3
## Lung tissue from cigarette smoke-treated mice at 4 months of age, biological replicate 2
##                                                                                  V4
## Lung tissue from cigarette smoke-treated mice at 4 months of age, biological replicate 3
##                                                                                  V5
##                 Lung tissue from control mice at 4 months of age, biological replicate 1
##                                                                                  V6
##                 Lung tissue from control mice at 4 months of age, biological replicate 2
##                                                                                  V7
##                 Lung tissue from control mice at 4 months of age, biological replicate 3
##                                                                                  V8
## Lung tissue from cigarette smoke-treated mice at 6 months of age, biological replicate 1
##                                                                                  V9
## Lung tissue from cigarette smoke-treated mice at 6 months of age, biological replicate 2
##                                                                                 V10
## Lung tissue from cigarette smoke-treated mice at 6 months of age, biological replicate 3
##                                                                                 V11
##                 Lung tissue from control mice at 6 months of age, biological replicate 1
##                                                                                 V12
##                 Lung tissue from control mice at 6 months of age, biological replicate 2
##                                                                                 V13
##                 Lung tissue from control mice at 6 months of age, biological replicate 3
## 12 Levels: Lung tissue from cigarette smoke-treated mice at 4 months of age, biological replicate 1
```

Choosing shoerter column names:

```
sml=c(rep("smoke_4",3),rep("control_4",3),rep("smoke_6",3),rep("control_6",3))
sml <- factor(sml)
levels(sml)
```

```
## [1] "control_4" "control_6" "smoke_4"   "smoke_6"
```

```
class(sml)
```

```
## [1] "factor"
```

Extracting Matrices

```
ex<-exprs(gset)
class(ex)
```

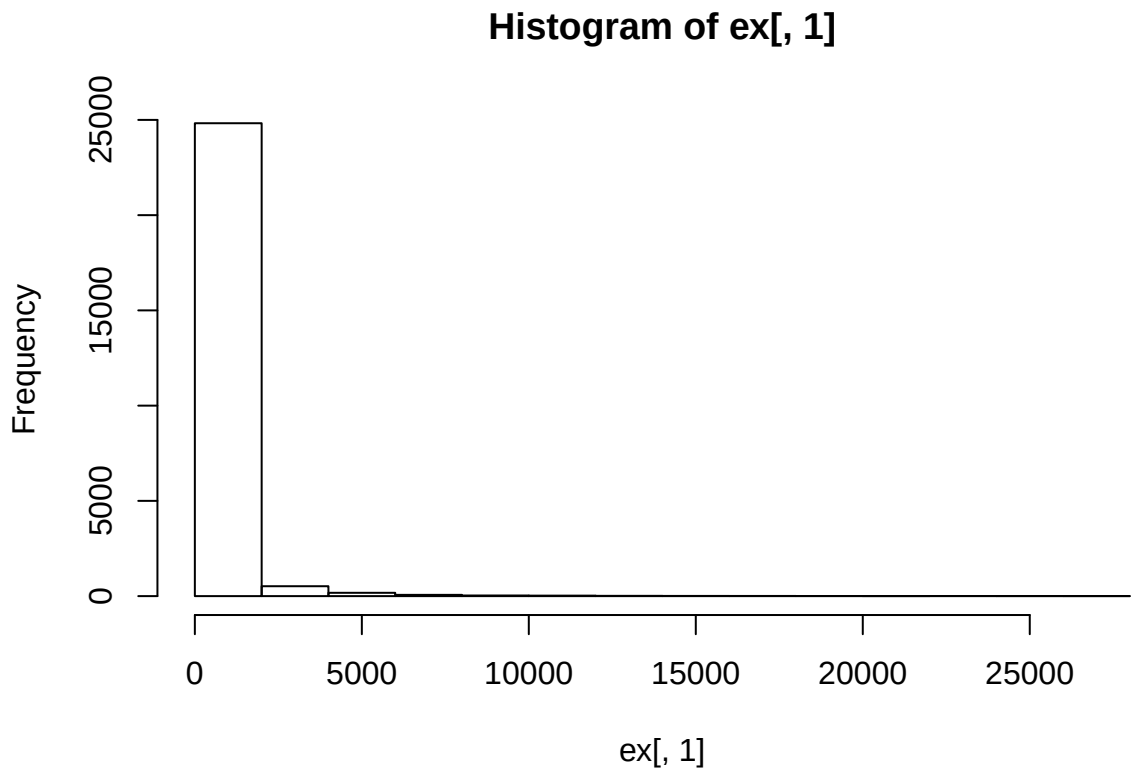
```
## [1] "matrix"
```

```
dim(ex)
```

```
## [1] 25697    12
```

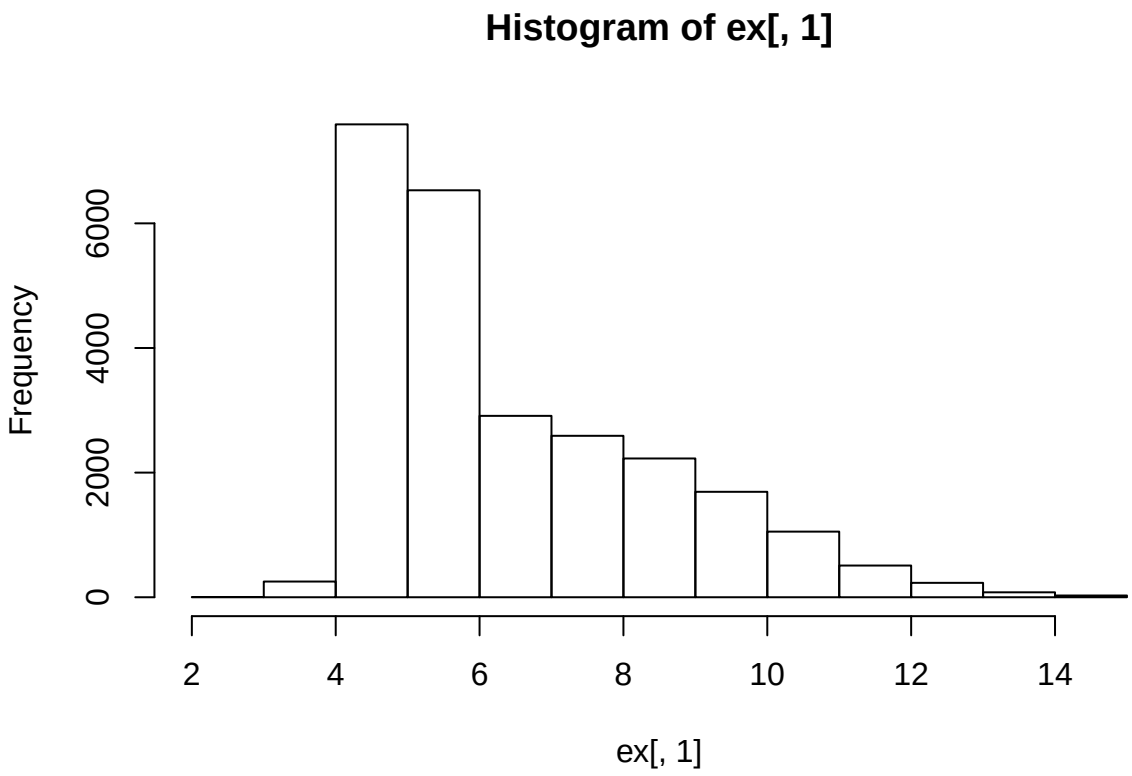
Frequency Histogram

```
hist(ex[,1])
```

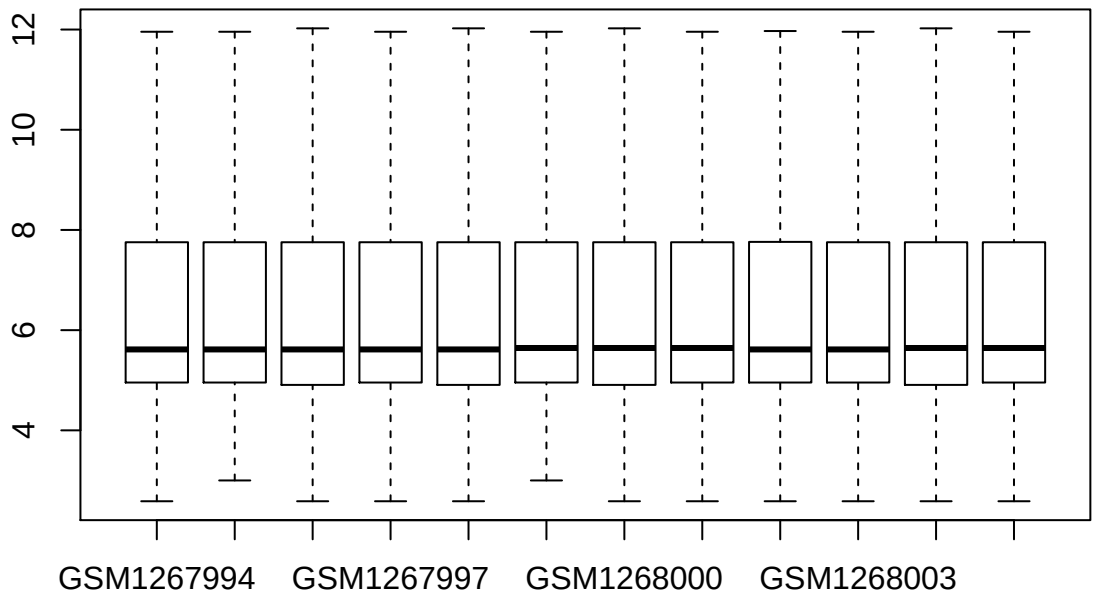


Normalizing Data

```
ex<-log2(ex)
hist(ex[,1])
```

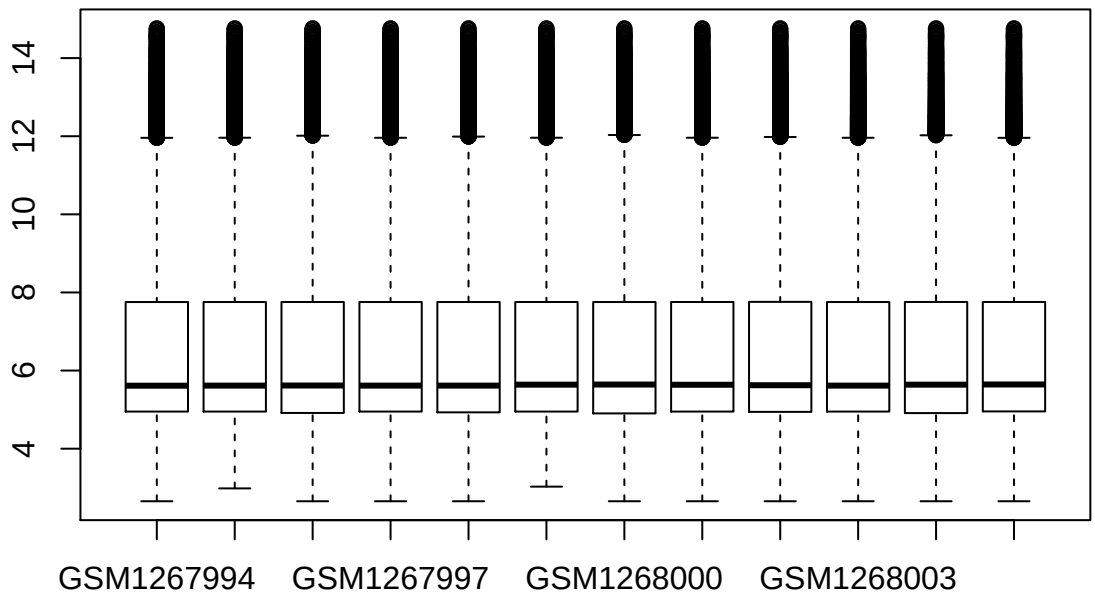



```
exprs(gset)<-ex
boxplot(ex,outline=FALSE)
```



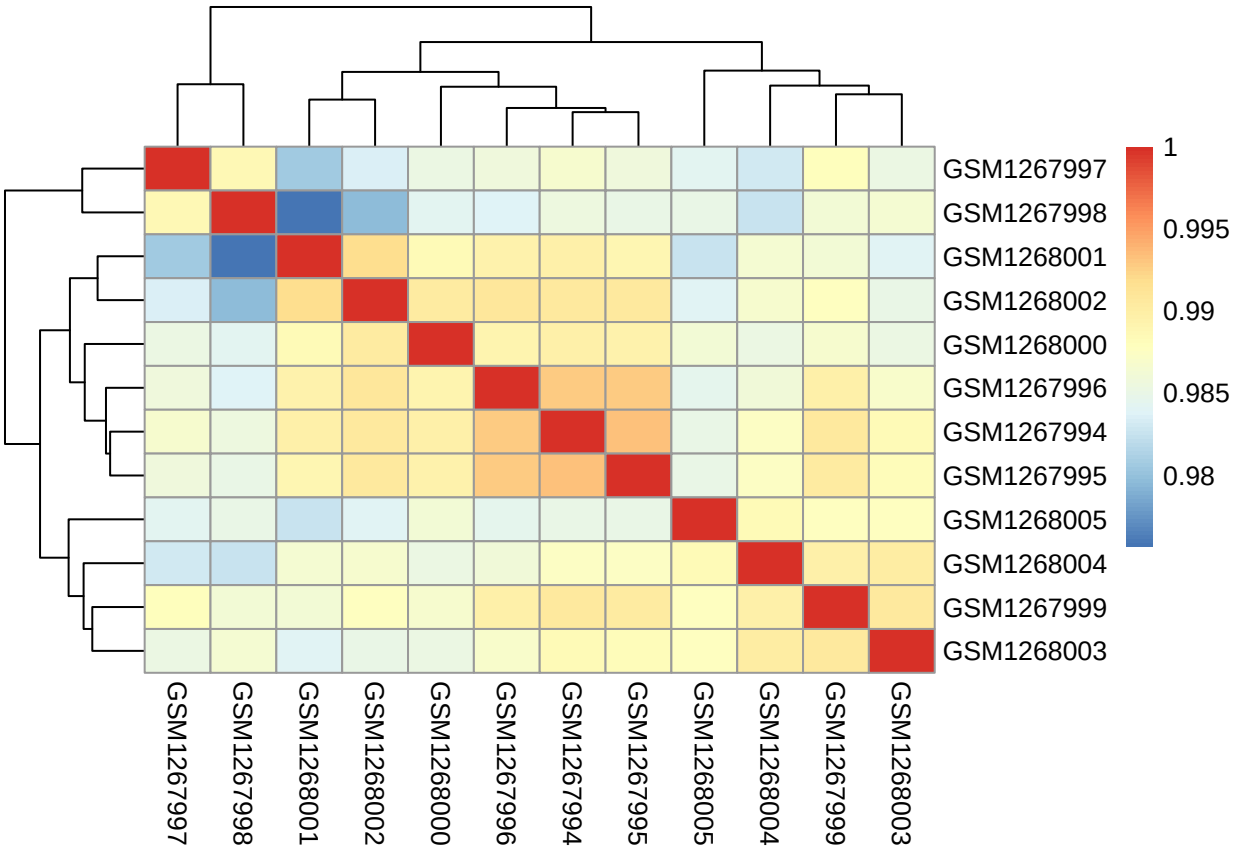
If data was not normal

```
x<-normalizeQuantiles(ex)
boxplot(x)
```

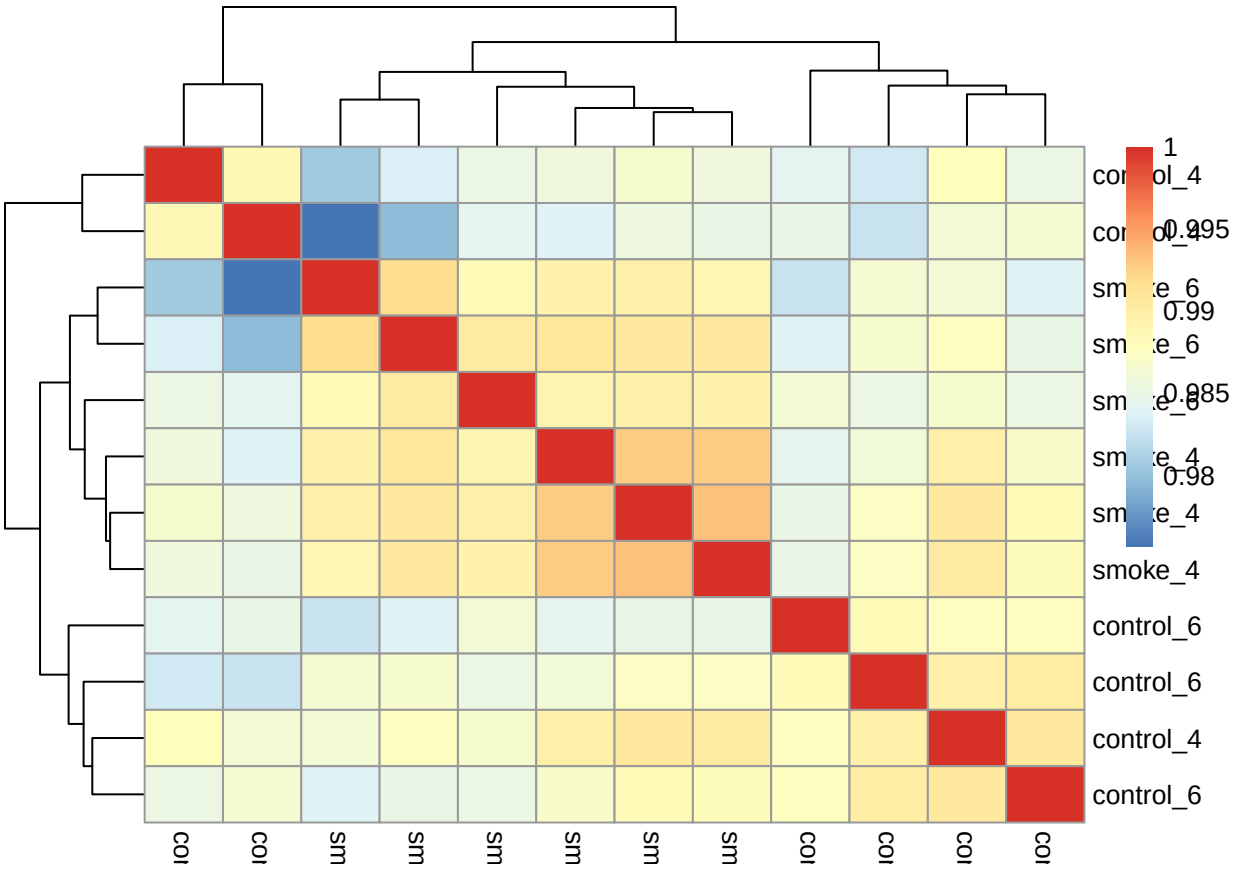


Correlation HeatMap

```
pheatmap(cor(ex))
```

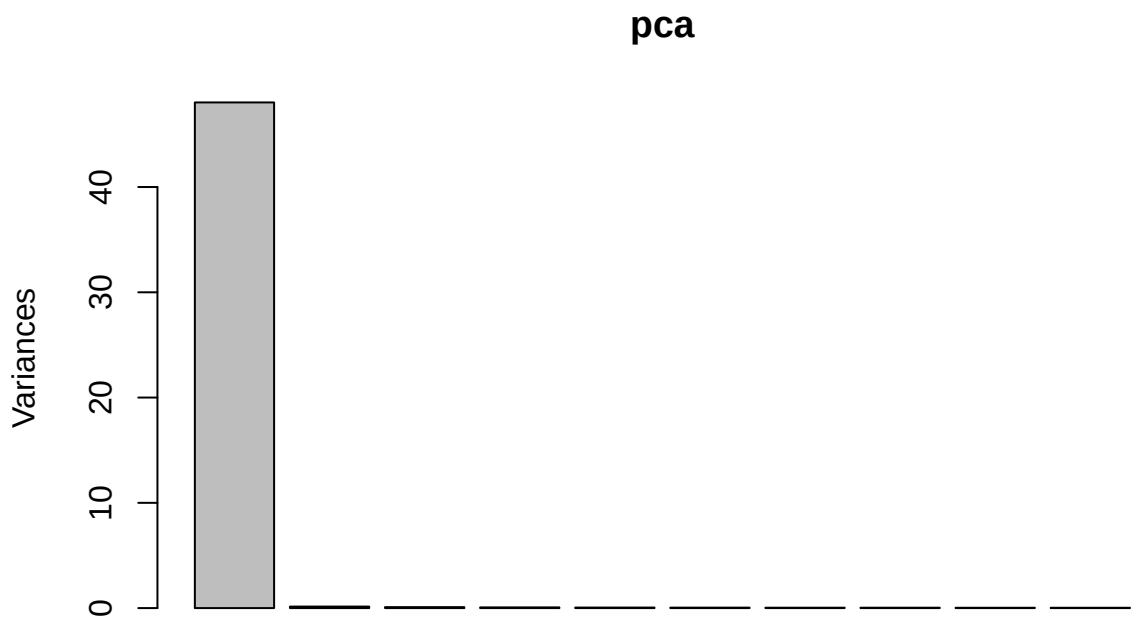



```
pheatmap(cor(ex),labels_row = sml,labels_col = sml,legend = TRUE)
```



Principal Component Analysis

```
pca<-prcomp(ex)
plot(pca)
```




```
names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

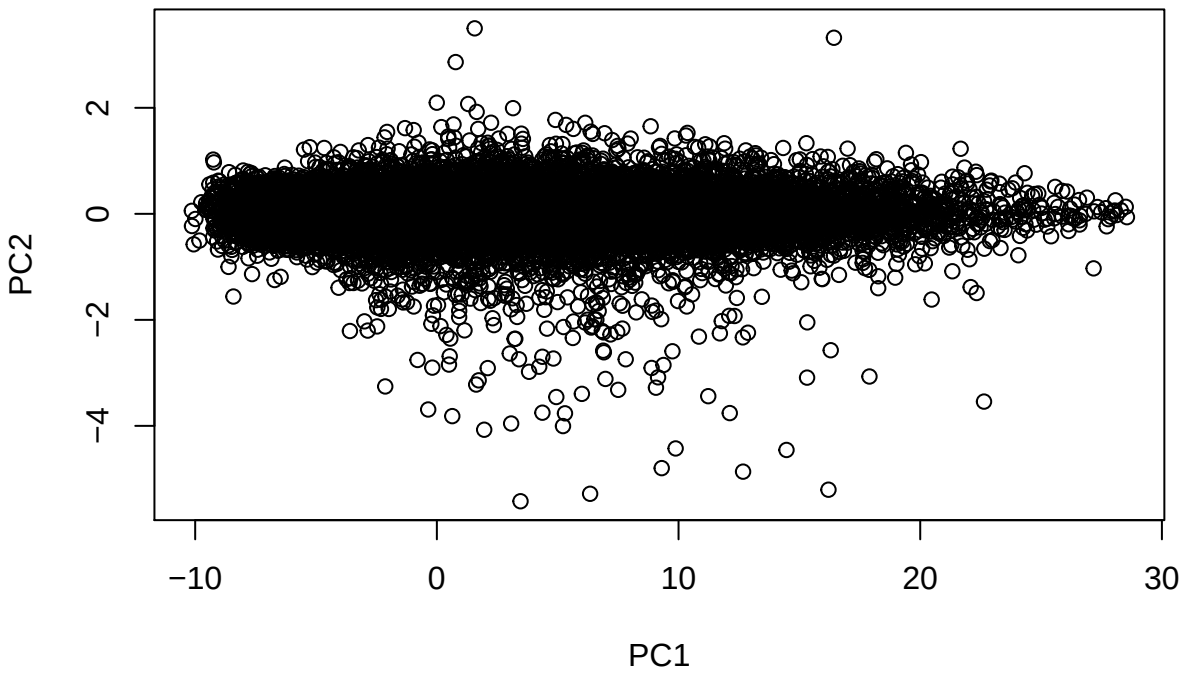
```
pca$sdev
```

```
##  [1] 6.9307237 0.3772307 0.3028166 0.2553910 0.2256388 0.2049636 0.1897303
##  [8] 0.1849727 0.1762470 0.1755605 0.1710861 0.1633337
```

```
colnames(pca$x)
```

```
##  [1] "PC1"  "PC2"  "PC3"  "PC4"  "PC5"  "PC6"  "PC7"  "PC8"  "PC9"  "PC10"
## [11] "PC11" "PC12"
```

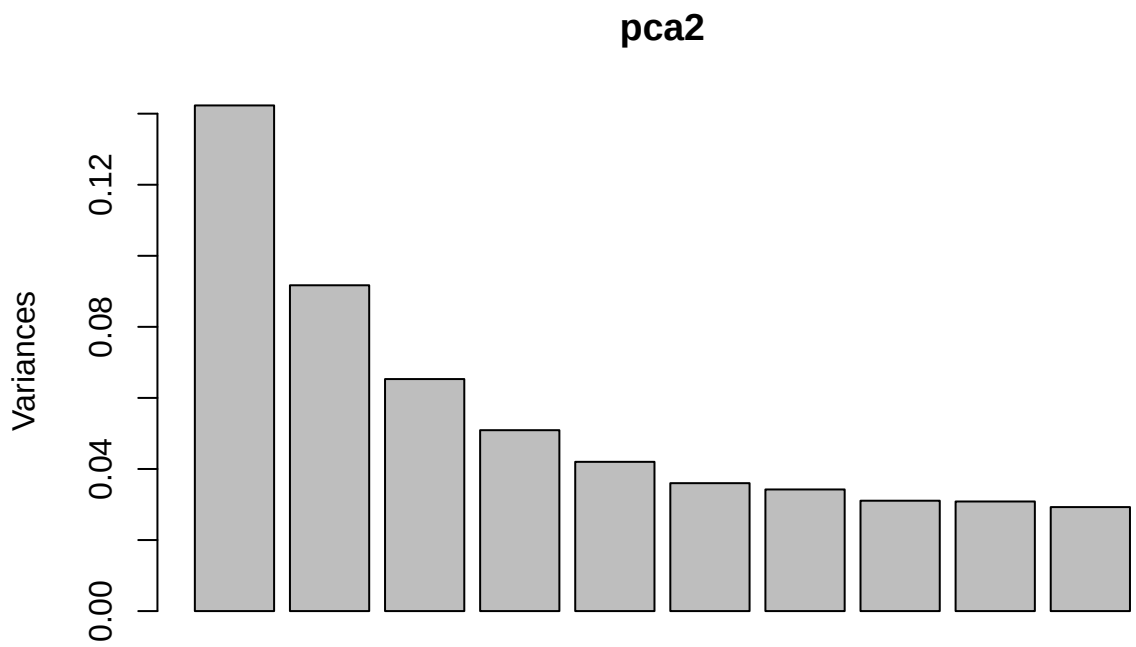
```
plot(pca$x[,1:2])
```



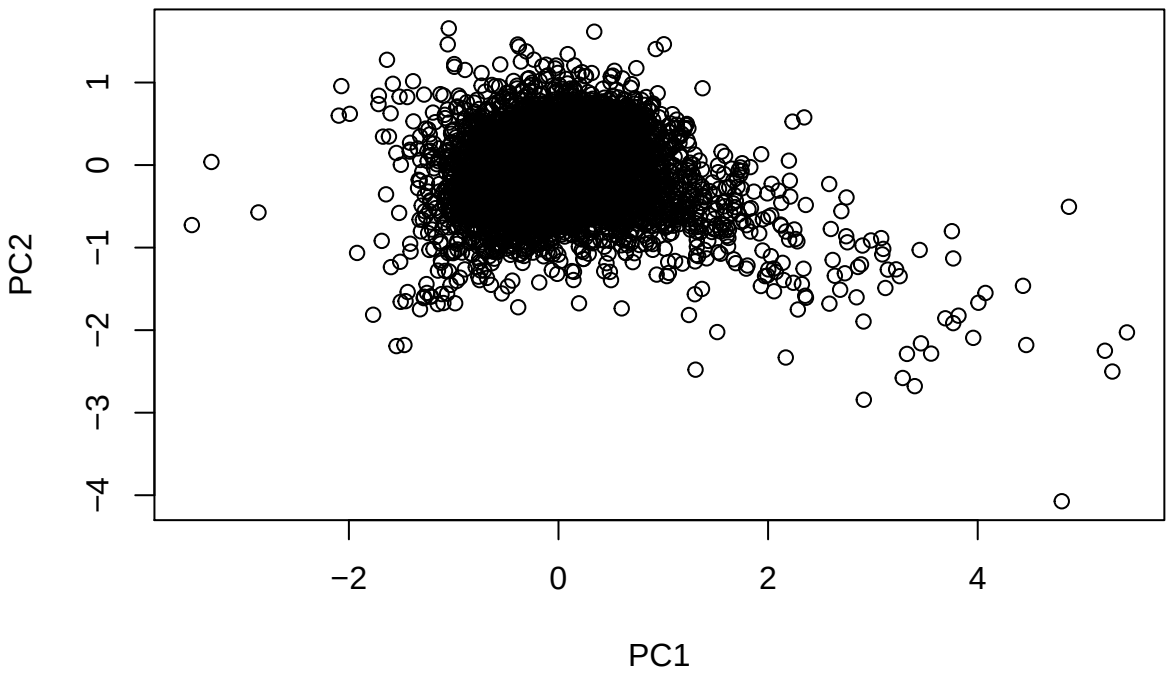

```
ex_scale=t(scale(t(ex),scale=F))
mean(ex_scale[1,])
```

```
## [1] 1.48032e-16
```

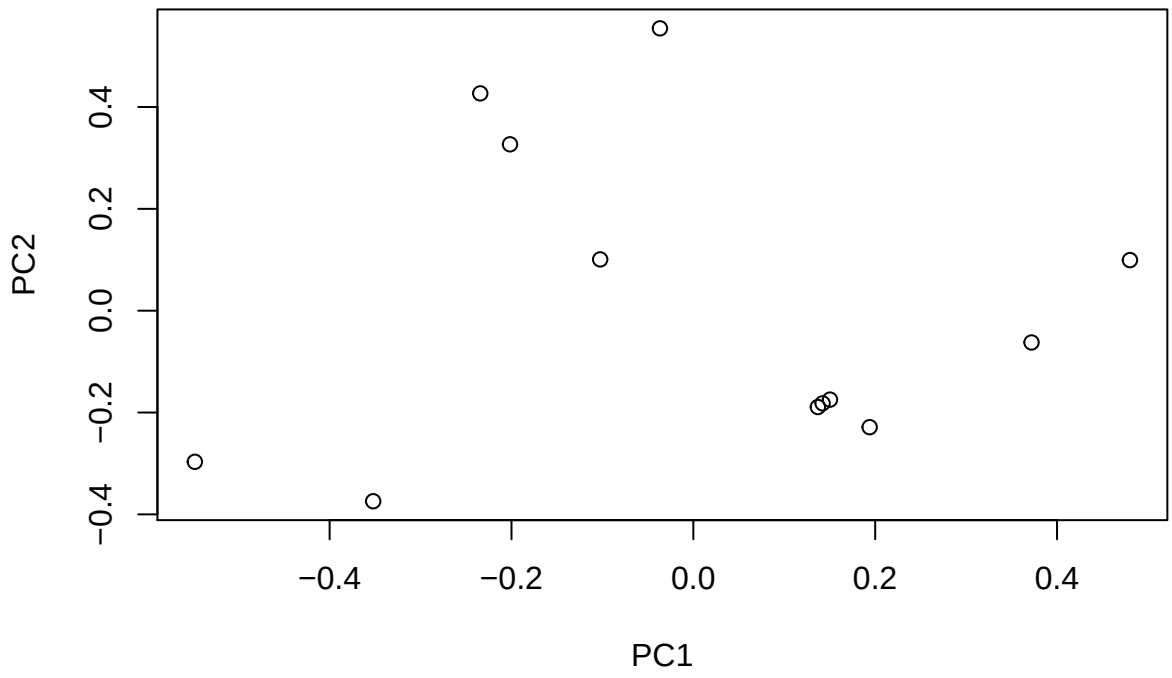
```
pca2<-prcomp(ex_scale)
plot(pca2)
```




```
plot(pca2$x[,1:2])
```



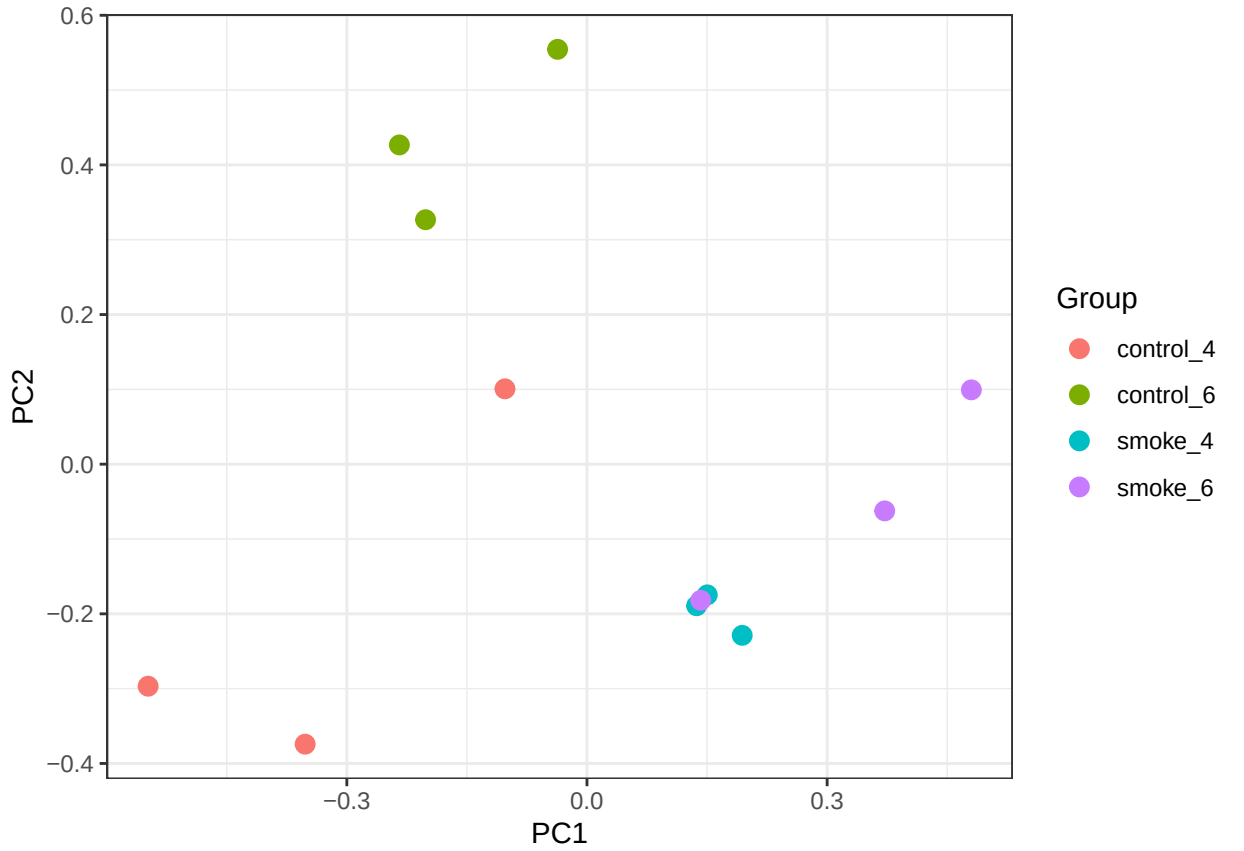

```
plot(pca2$r)
```




```
pc.sample<-data.frame(pca2$r[,1:3],Group=sml)
head(pc.sample)
```

```
##                   PC1        PC2         PC3     Group
## GSM1267994  0.1369598 -0.1891969  0.23801377   smoke_4
## GSM1267995  0.1502831 -0.1746672  0.24786185   smoke_4
## GSM1267996  0.1939202 -0.2287598  0.19455144   smoke_4
## GSM1267997 -0.3520856 -0.3741953 -0.08463492 control_4
## GSM1267998 -0.5482303 -0.2967732 -0.05780631 control_4
## GSM1267999 -0.1024348  0.1007085  0.32999289 control_4
```

```
ggplot(pc.sample,aes(PC1,PC2,color=Group))+geom_point(size=3)+theme_bw()
```



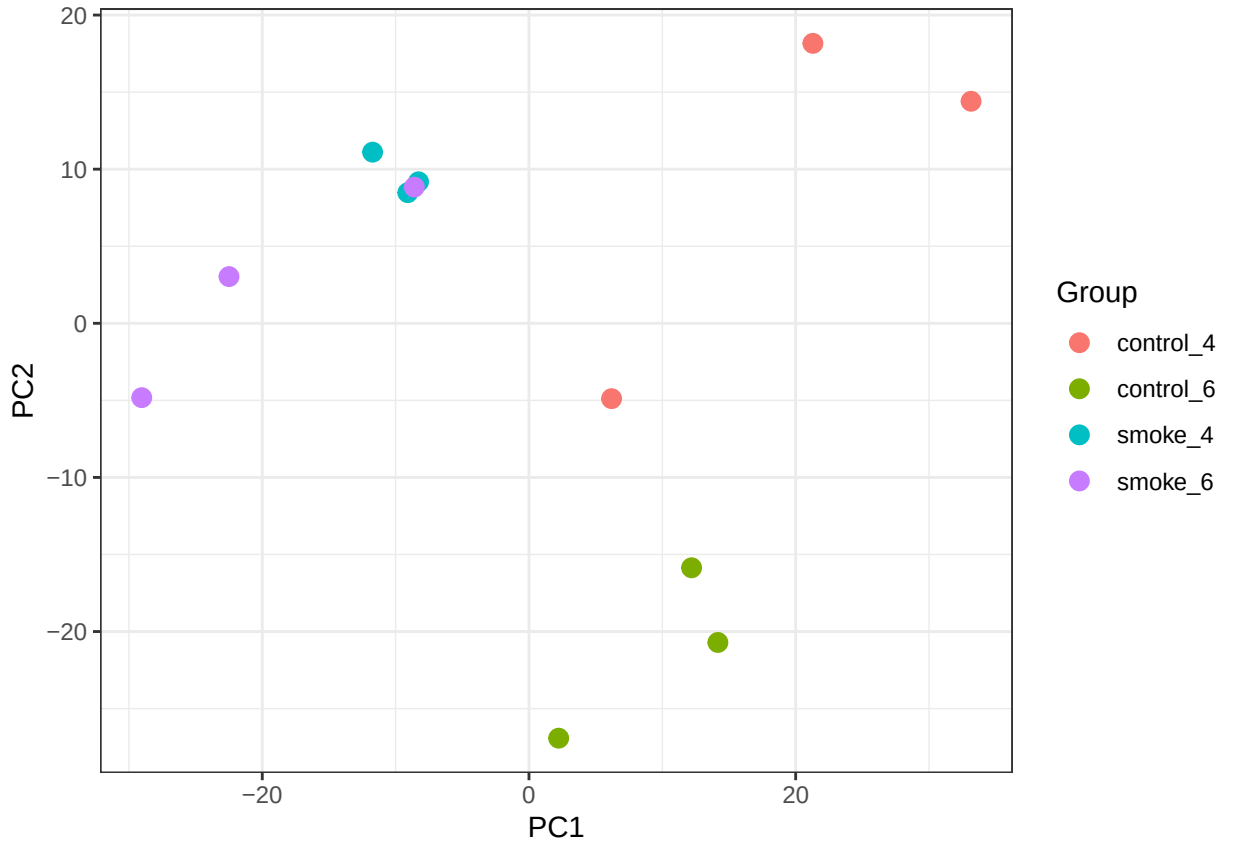

```
dev.off()
```

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

```
pca2<-prcomp(t(ex_scale))
pc.sample<-data.frame(pca2$x[,1:3],Group=sml)
head(pc.sample)
```

```
##                   PC1       PC2        PC3     Group
## GSM1267994  -8.280666  9.181957  -9.754685   smoke_4
## GSM1267995  -9.086877  8.476906 -10.155385   smoke_4
## GSM1267996 -11.726375 11.103007  -7.970225   smoke_4
## GSM1267997  21.292070 18.164345   3.459367 control_4
## GSM1267998  33.153703 14.406556   2.365096 control_4
## GSM1267999   6.194222 -4.888719 -13.504524 control_4
```

```
ggplot(pc.sample,aes(PC1,PC2,color=Group))+geom_point(size=3)+theme_bw()
```




```
dev.off()
```

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

```
sml <- factor(sml)
levels(sml)
```

```
## [1] "control_4" "control_6" "smoke_4"   "smoke_6"
```

```
sml
```

```
##  [1] smoke_4   smoke_4   smoke_4   control_4 control_4 control_4 smoke_6
##  [8] smoke_6   smoke_6   control_6 control_6 control_6
## Levels: control_4 control_6 smoke_4 smoke_6
```

```
gset$description <- sml
design <- model.matrix(~ description + 0, gset) #112
colnames(design) <- levels(sml)
head(design)
```

```
##            control_4 control_6 smoke_4 smoke_6
## GSM1267994         0         0       1       0
## GSM1267995         0         0       1       0
## GSM1267996         0         0       1       0
## GSM1267997         1         0       0       0
## GSM1267998         1         0       0       0
## GSM1267999         1         0       0       0
```

```
design
```

```
##            control_4 control_6 smoke_4 smoke_6
## GSM1267994         0         0       1       0
## GSM1267995         0         0       1       0
## GSM1267996         0         0       1       0
## GSM1267997         1         0       0       0
## GSM1267998         1         0       0       0
## GSM1267999         1         0       0       0
## GSM1268000         0         0       0       1
## GSM1268001         0         0       0       1
## GSM1268002         0         0       0       1
## GSM1268003         0         1       0       0
## GSM1268004         0         1       0       0
## GSM1268005         0         1       0       0
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$description
## [1] "contr.treatment"
```

Differential Gene Expression

```
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(smoke_4-control_4, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
```

Selecting first 250 genes with highest adjusted p-value

```
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
colnames(tT)
```

```
##  [1] "ID"                    "Gene.title"            "Gene.symbol"
##  [4] "Gene.ID"               "UniGene.title"         "UniGene.symbol"
##  [7] "UniGene.ID"            "Nucleotide.Title"      "GI"
## [10] "GenBank.Accession"     "Platform_CLONEID"      "Platform_ORF"
## [13] "Platform_SPOTID"       "Chromosome.location"   "Chromosome.annotation"
## [16] "GO.Function"           "GO.Process"            "GO.Component"
## [19] "GO.Function.ID"        "GO.Process.ID"         "GO.Component.ID"
## [22] "Platform_SEQUENCE"     "logFC"                 "AveExpr"
## [25] "t"                     "P.Value"               "adj.P.Val"
## [28] "B"
```

```
head(ex)
```

```
##              GSM1267994 GSM1267995 GSM1267996 GSM1267997 GSM1267998 GSM1267999
## ILMN_1212607   4.700440   5.044394   5.209453   4.857981   5.044394   4.523562
## ILMN_1212612   7.209453   7.076816   7.129283   7.159871   7.475733   7.219169
## ILMN_1212619   5.857981   5.906891   5.832890   5.426265   5.857981   5.906891
## ILMN_1212628   5.044394   5.044394   5.129283   5.209453   4.700440   5.087463
## ILMN_1212632   5.614710   5.321928   5.781360   5.357552   5.491853   5.754888
## ILMN_1212636   9.988685   9.972980  10.051209   9.804131   9.834471   9.954196
```

```
##              GSM1268000 GSM1268001 GSM1268002 GSM1268003 GSM1268004 GSM1268005
## ILMN_1212607   5.000000   4.954196   4.807355   5.169925   5.000000   4.807355
## ILMN_1212612   7.266787   7.087463   6.918863   6.965784   7.238405   7.417853
## ILMN_1212619   5.954196   6.044394   5.700440   5.781360   6.247928   6.000000
## ILMN_1212628   5.087463   5.392317   4.754888   5.087463   4.643856   5.247928
## ILMN_1212632   5.614710   5.426265   5.700440   5.614710   5.754888   5.954196
## ILMN_1212636   9.840778  10.049849   9.831307   9.870365  10.036174   9.828136
```

```
tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC",
                          "Gene.symbol","Gene.title"))
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=Inf)
colnames(tT)
```

```
##  [1] "ID"                    "Gene.title"            "Gene.symbol"
##  [4] "Gene.ID"               "UniGene.title"         "UniGene.symbol"
##  [7] "UniGene.ID"            "Nucleotide.Title"      "GI"
## [10] "GenBank.Accession"     "Platform_CLONEID"      "Platform_ORF"
## [13] "Platform_SPOTID"       "Chromosome.location"   "Chromosome.annotation"
## [16] "GO.Function"           "GO.Process"            "GO.Component"
## [19] "GO.Function.ID"        "GO.Process.ID"         "GO.Component.ID"
## [22] "Platform_SEQUENCE"     "logFC"                 "AveExpr"
## [25] "t"                     "P.Value"               "adj.P.Val"
## [28] "B"
```

```
head(ex)
```

```
##              GSM1267994 GSM1267995 GSM1267996 GSM1267997 GSM1267998 GSM1267999
## ILMN_1212607   4.700440   5.044394   5.209453   4.857981   5.044394   4.523562
## ILMN_1212612   7.209453   7.076816   7.129283   7.159871   7.475733   7.219169
## ILMN_1212619   5.857981   5.906891   5.832890   5.426265   5.857981   5.906891
## ILMN_1212628   5.044394   5.044394   5.129283   5.209453   4.700440   5.087463
## ILMN_1212632   5.614710   5.321928   5.781360   5.357552   5.491853   5.754888
## ILMN_1212636   9.988685   9.972980  10.051209   9.804131   9.834471   9.954196
##              GSM1268000 GSM1268001 GSM1268002 GSM1268003 GSM1268004 GSM1268005
## ILMN_1212607   5.000000   4.954196   4.807355   5.169925   5.000000   4.807355
## ILMN_1212612   7.266787   7.087463   6.918863   6.965784   7.238405   7.417853
## ILMN_1212619   5.954196   6.044394   5.700440   5.781360   6.247928   6.000000
## ILMN_1212628   5.087463   5.392317   4.754888   5.087463   4.643856   5.247928
## ILMN_1212632   5.614710   5.426265   5.700440   5.614710   5.754888   5.954196
## ILMN_1212636   9.840778  10.049849   9.831307   9.870365  10.036174   9.828136
```

```
tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC",
                          "Gene.symbol","Gene.title"))
```

```
ms.up=subset(tT,select=c("ID","adj.P.Val","logFC","Gene.symbol"))
ms.up=subset(tT,tT$logFC>1 & tT$adj.P.Val<0.05, select=c( "ID", "adj.P.Val", "logFC", "Gene.symbol"))
```

```
#write.table(ms.up.genenames,"/data/processed/up_ptA_ptB.txt", quote = F,row.names = F, col.names = F)
```

Finding out the up-regulated genes

```
ms.up.genenames<- sub("///.*","",ms.up$Gene.symbol)
ms.up.genenames<- ms.up.genenames[ms.up.genenames!= ""]
ms.up.genenames<- strsplit2(ms.up.genenames,"///")
ms.up.genenames<- unique(ms.up.genenames)

ms.up.genenames
```

```
##       [,1]
##  [1,] "Zranb3"
##  [2,] "Gpnmb"
##  [3,] "Ctsk"
##  [4,] "Trem2"
##  [5,] "Lhfpl2"
##  [6,] "Spp1"
##  [7,] "Syngr1"
##  [8,] "Gdf15"
##  [9,] "Ch25h"
## [10,] "Dbp"
## [11,] "Lrp12"
## [12,] "Npy"
## [13,] "Ccl3"
## [14,] "Cxcr1"
## [15,] "Marco"
## [16,] "Ccl6"
## [17,] "Lilr4b"
## [18,] "Igf1"
## [19,] "Ccl7"
## [20,] "Mmp12"
## [21,] "Saa3"
## [22,] "Clec4n"
## [23,] "Mreg"
## [24,] "Ms4a7"
## [25,] "Lilrb4a"
## [26,] "Tnfrsf9"
## [27,] "Pld3"
## [28,] "Wfdc17"
## [29,] "Itih4"
## [30,] "Inhba"
## [31,] "Ccl4"
## [32,] "Adgre1"
## [33,] "Ccl2"
## [34,] "Nr1d2"
## [35,] "Ms4a6d"
## [36,] "Ccl9"
## [37,] "Orm2"
## [38,] "Apol7c"
## [39,] "Mcoln3"
## [40,] "Ear6"
## [41,] "Il1rn"
## [42,] "Slc7a11"
## [43,] "Cd84"
## [44,] "Fabp5"
## [45,] "Itih1"
## [46,] "Mamdc2"
## [47,] "Csf2rb2"
## [48,] "Ctss"
## [49,] "C1qc"
## [50,] "Orm1"
## [51,] "C1qb"
## [52,] "Ccl12"
## [53,] "Fcgr3"
```

```
## [54,] "Clec5a"
## [55,] "Per2"
## [56,] "Nr1d1"
## [57,] "Slc11a1"
## [58,] "Ctsa"
## [59,] "Cxcl1"
```

```
ms.up$Gene.symbol
```

```
##  [1] "Zranb3"  "Gpnmb"   "Ctsk"    "Trem2"   "Lhfpl2"  "Gpnmb"   "Spp1"
##  [8] "Syngr1"  "Gdf15"   "Ch25h"   "Dbp"     "Lrp12"   "Npy"     "Lrp12"
## [15] "Ccl3"    "Cxcr1"   "Marco"   "Syngr1"  "Ccl6"    "Lilr4b"  "Igf1"
## [22] "Ccl7"    "Mmp12"   ""        ""        "Saa3"    "Clec4n"  "Mreg"
## [29] "Ms4a7"   "Lilrb4a" "Tnfrsf9" "Pld3"    "Wfdc17"  "Itih4"   "Inhba"
## [36] "Itih4"   "Ccl4"    "Adgre1"  "Gpnmb"   "Ccl2"    "Pld3"    "Nr1d2"
## [43] "Ms4a6d"  "Itih4"   "Ccl9"    "Orm2"    "Apol7c"  "Mcoln3"  "Ear6"
## [50] "Il1rn"   "Slc7a11" "Cd84"    "Fabp5"   "Adgre1"  "Itih1"   "Mamdc2"
## [57] "Csf2rb2" "Ctss"    "C1qc"    "Orm1"    "C1qb"    "Ccl12"   "Fcgr3"
## [64] "Clec5a"  "Ms4a6d"  "Per2"    "Nr1d1"   "Slc11a1" "Ctsa"    "Per2"
## [71] "Cxcl1"   "Ccl12"   "Nr1d1"
```

```
write.table(ms.up.genenames,"./data/processed/up_ptA_ptB.txt", quote = F,row.names = F, col.names = F)
```