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**

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)
```

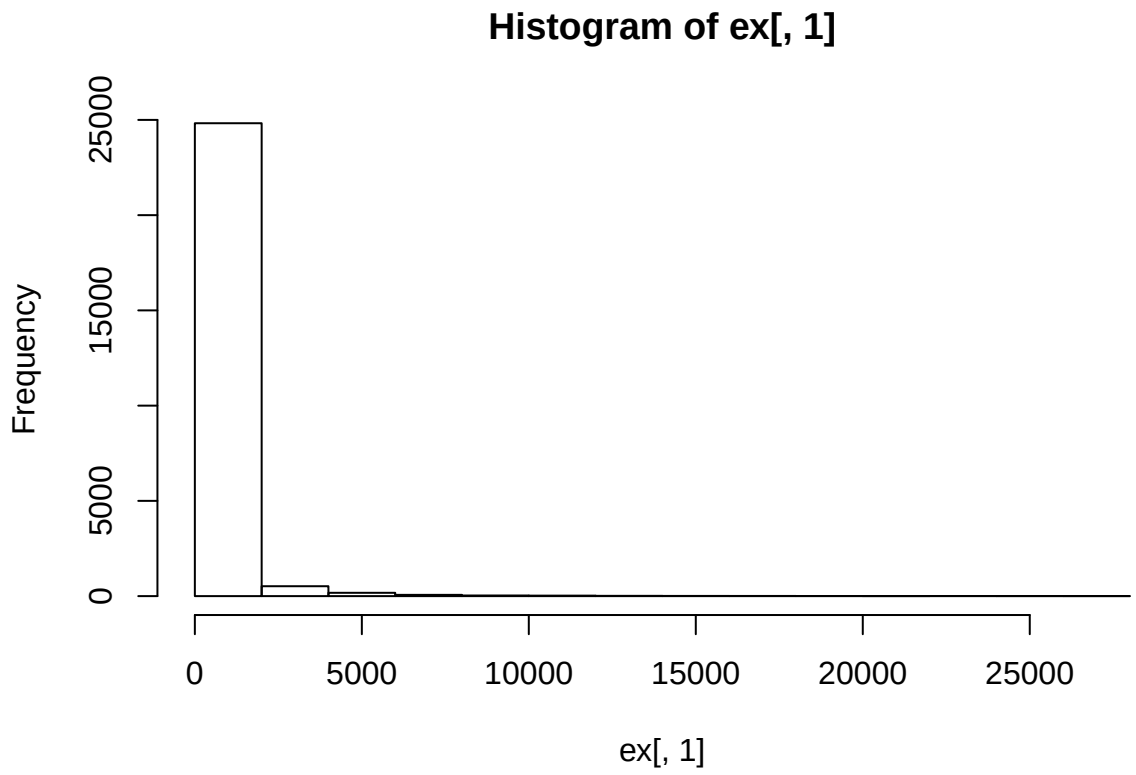
```
## [1] "matrix"
```

```
dim(ex)
```

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

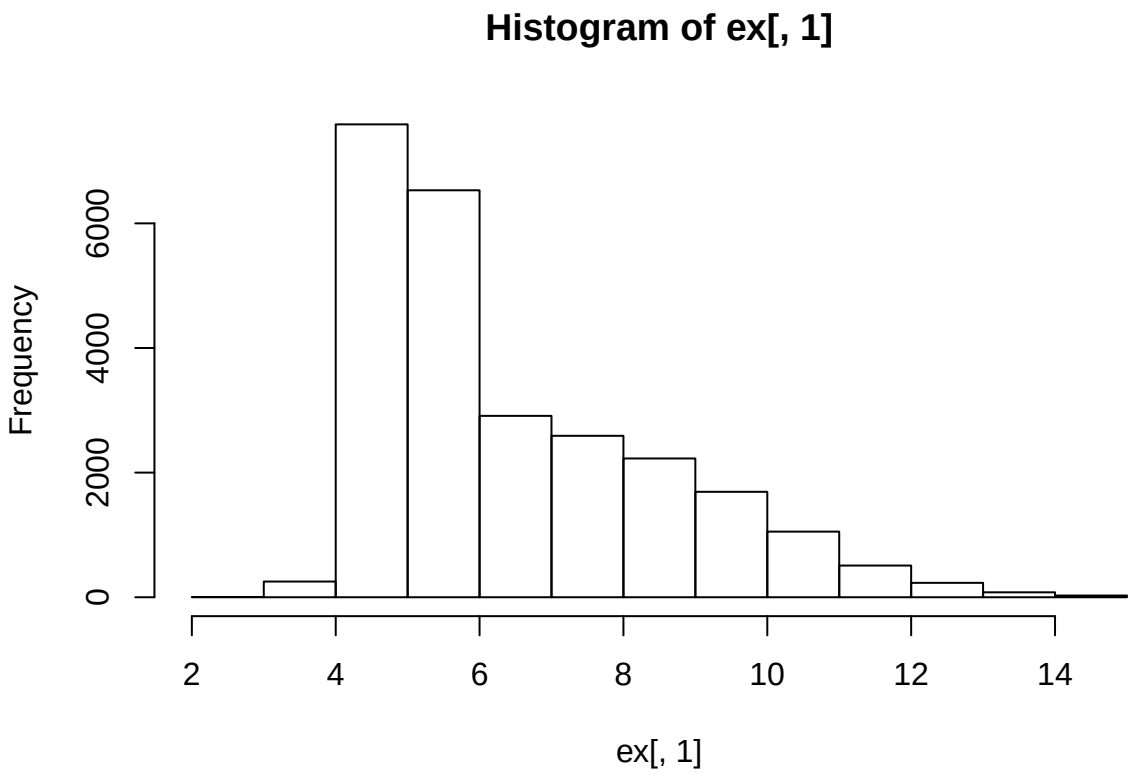
Frequency Histogram

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

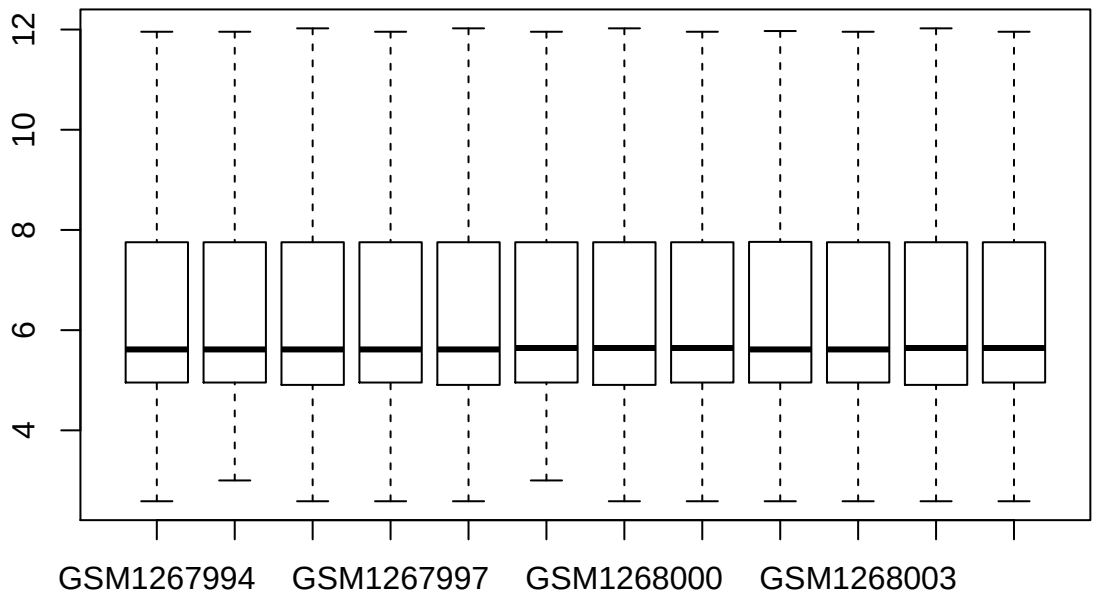


Normalizing Data

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

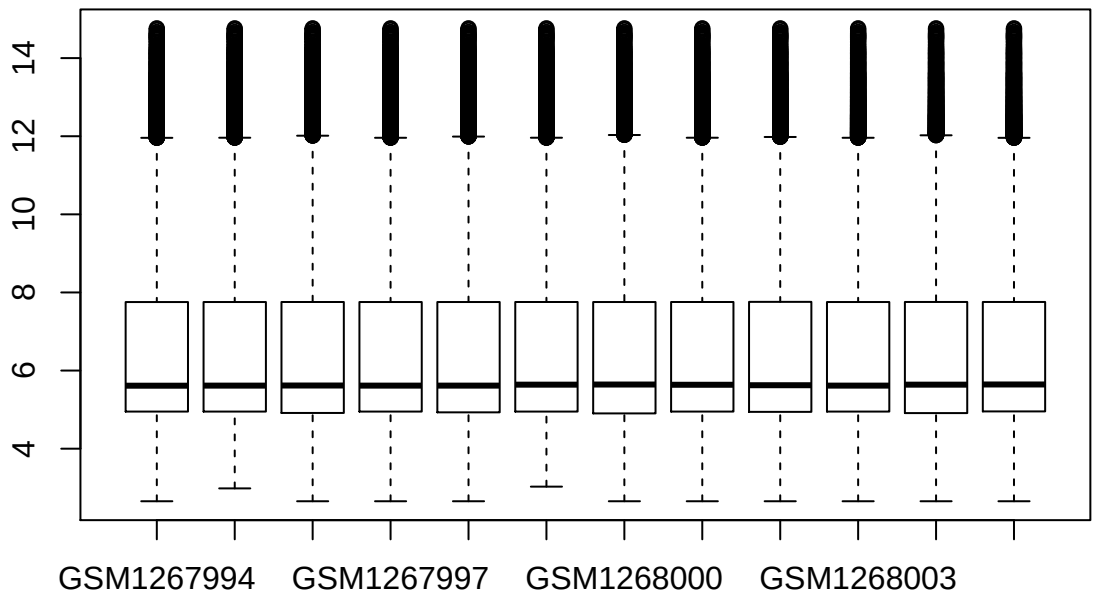



```
exprs(gset)<-ex
boxplot(ex,outline=FALSE)
```



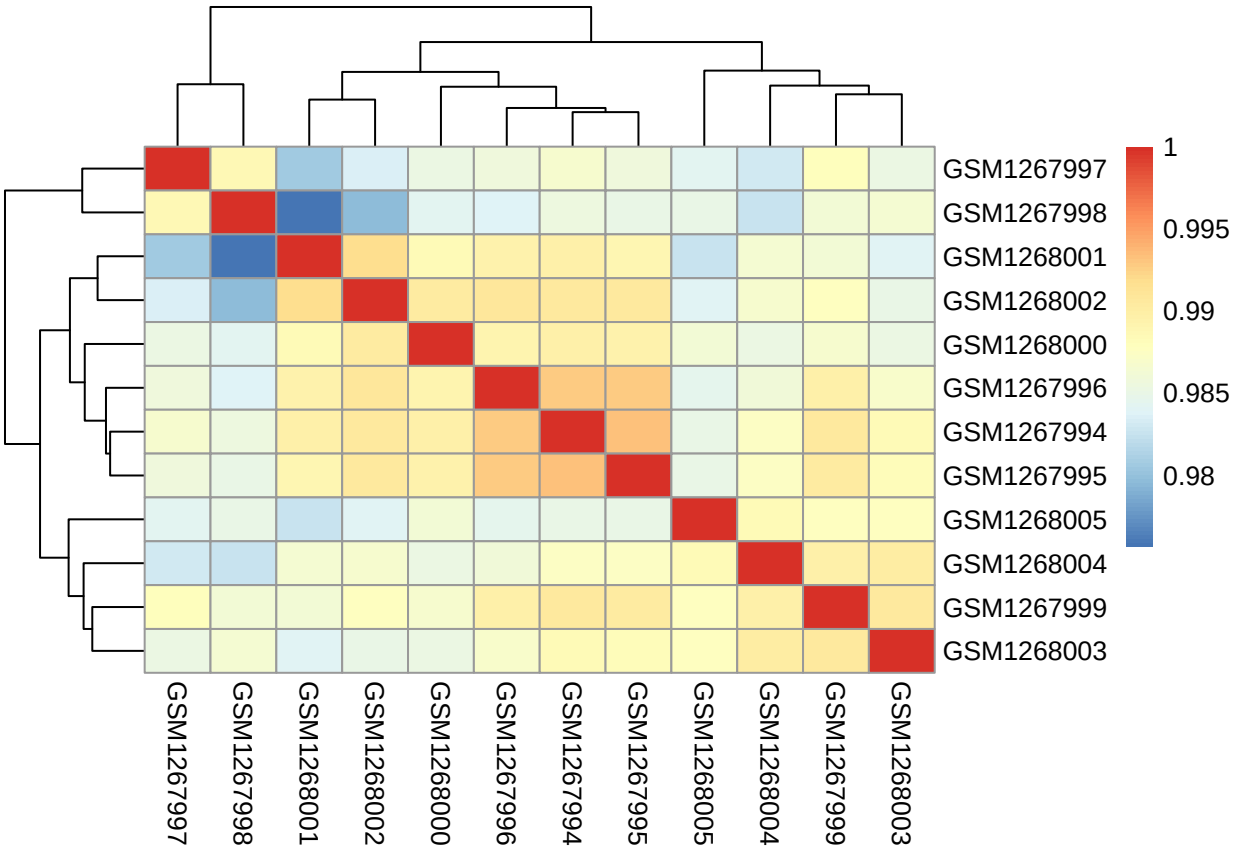
If data was not normal

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

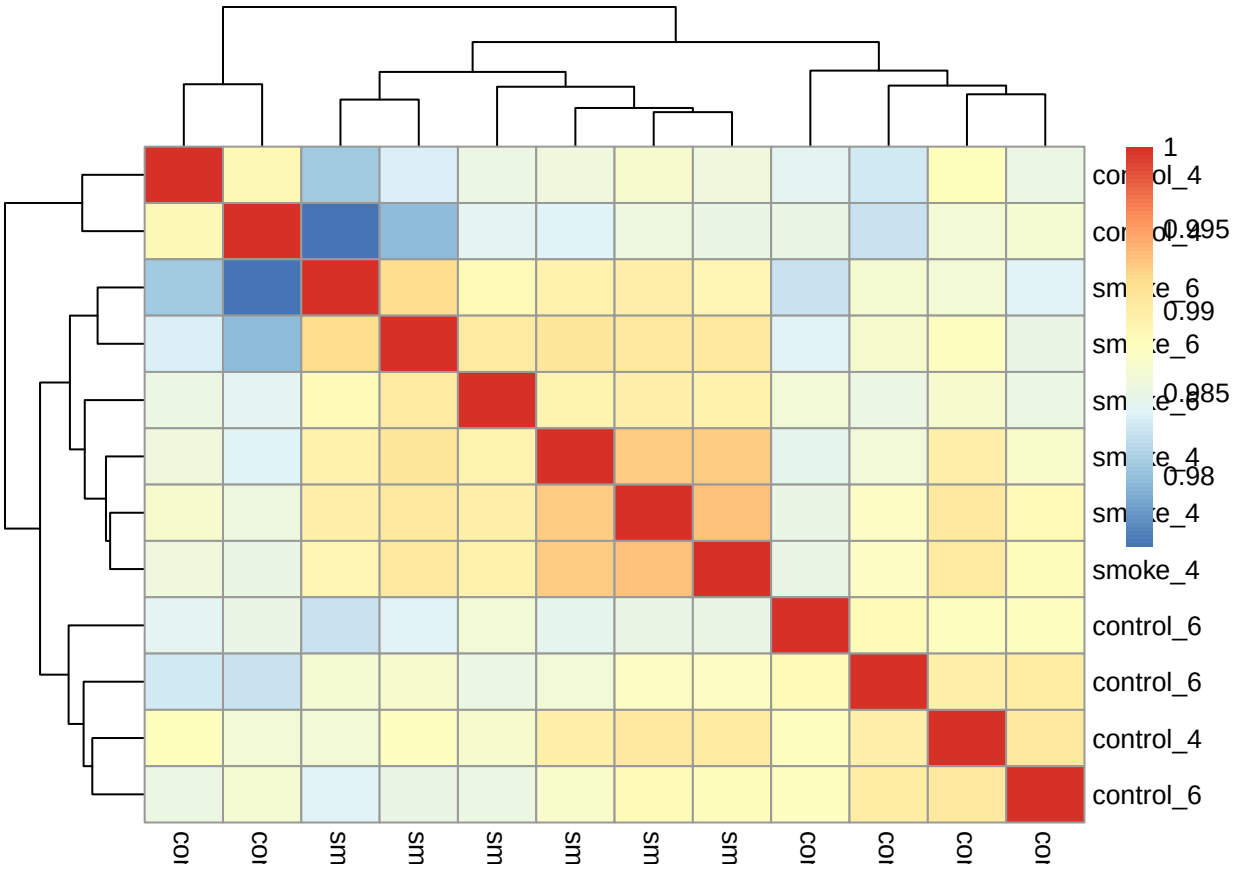


Correlation HeatMap

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

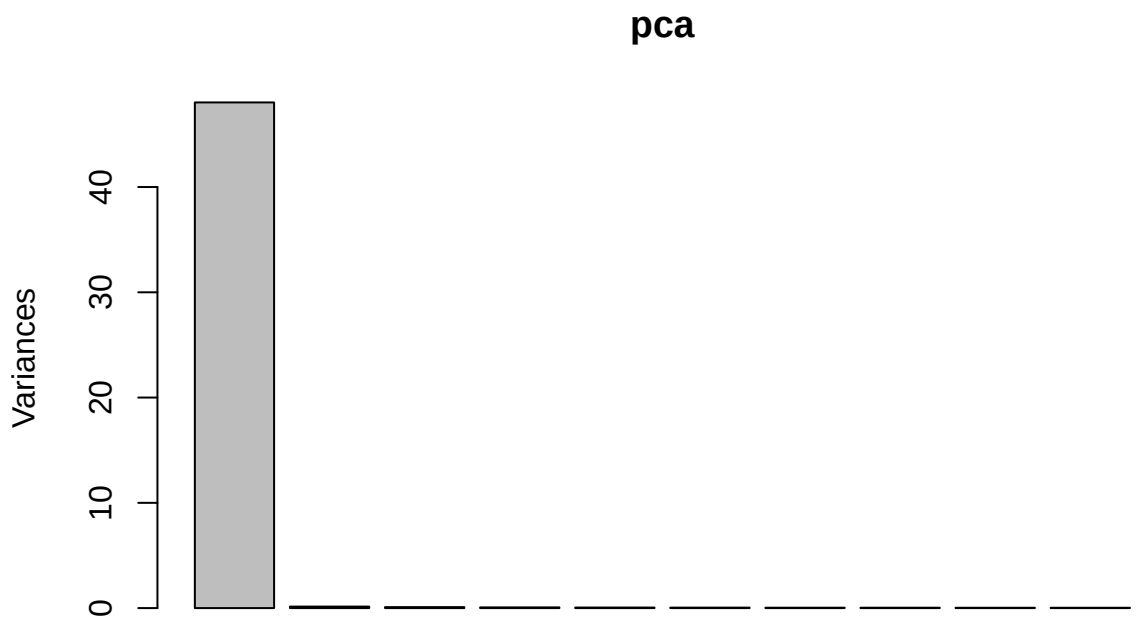



```
pheatmap(cor(ex),labels_row = sml,labels_col = sml,legend = TRUE)
```



Principal Component Analysis

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




```
names(pca)
```

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

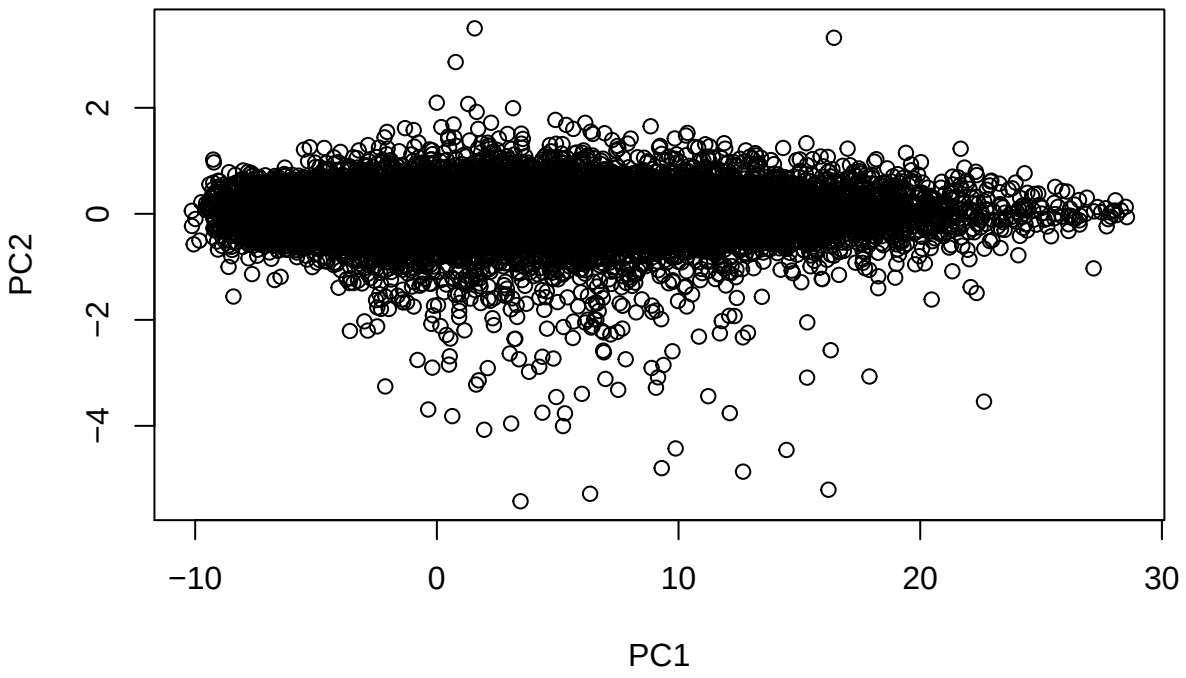
```
pca$sdev
```

```
##  [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"
```

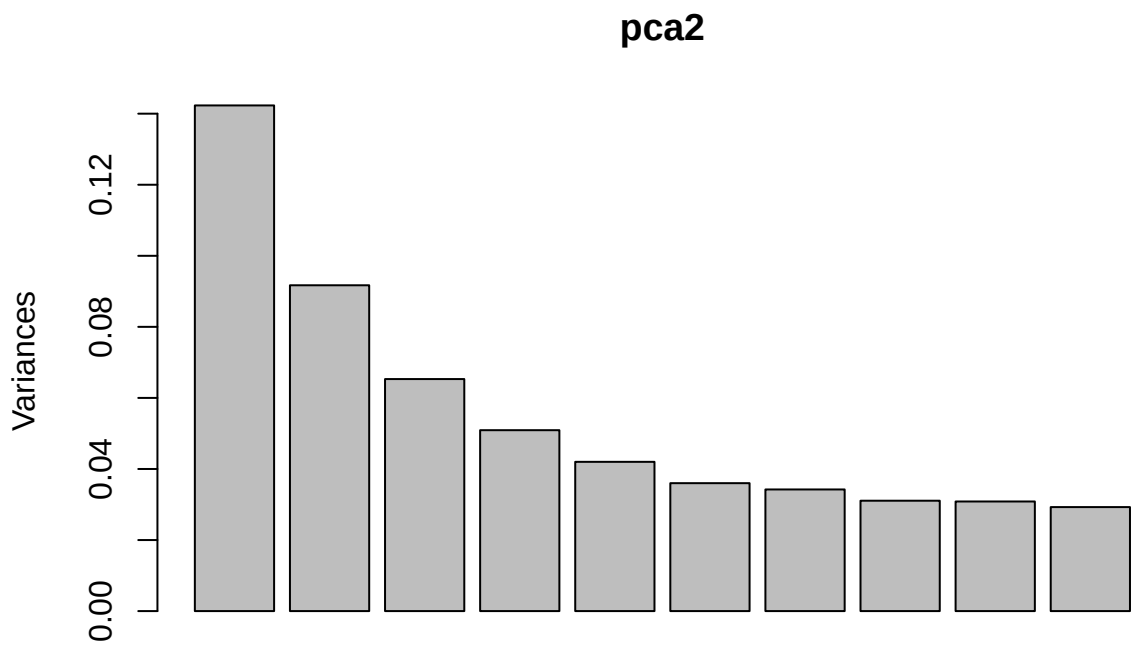
```
plot(pca$x[,1:2])
```



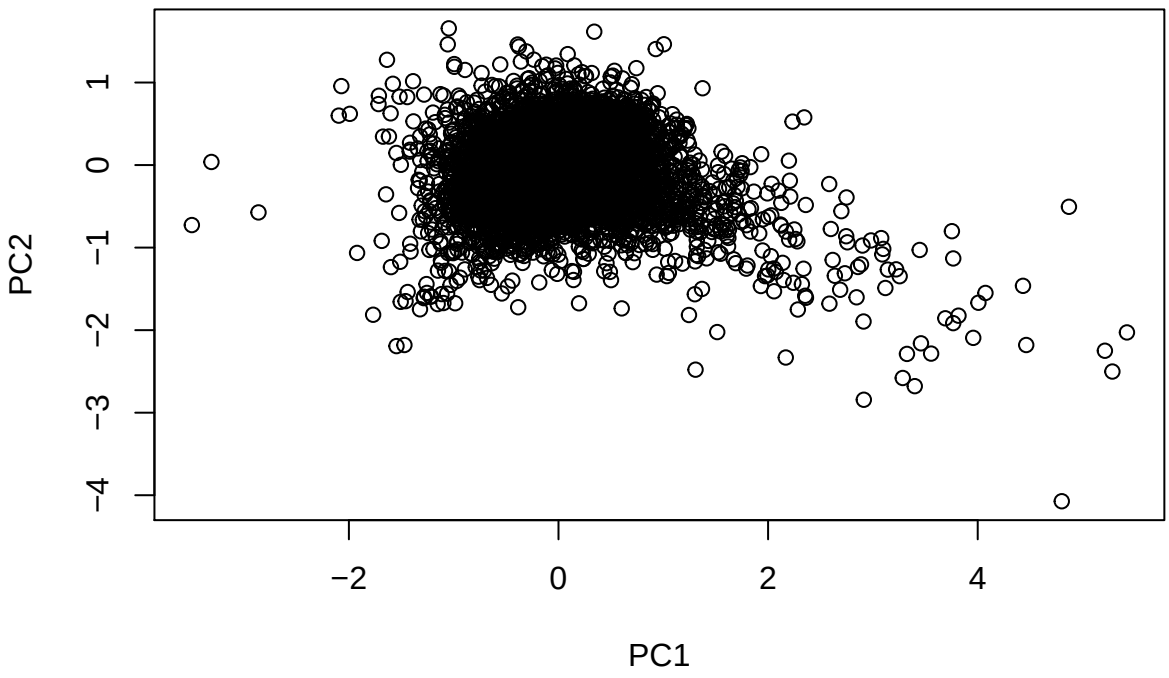

```
ex_scale=t(scale(t(ex),scale=F))
mean(ex_scale[1,])
```

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

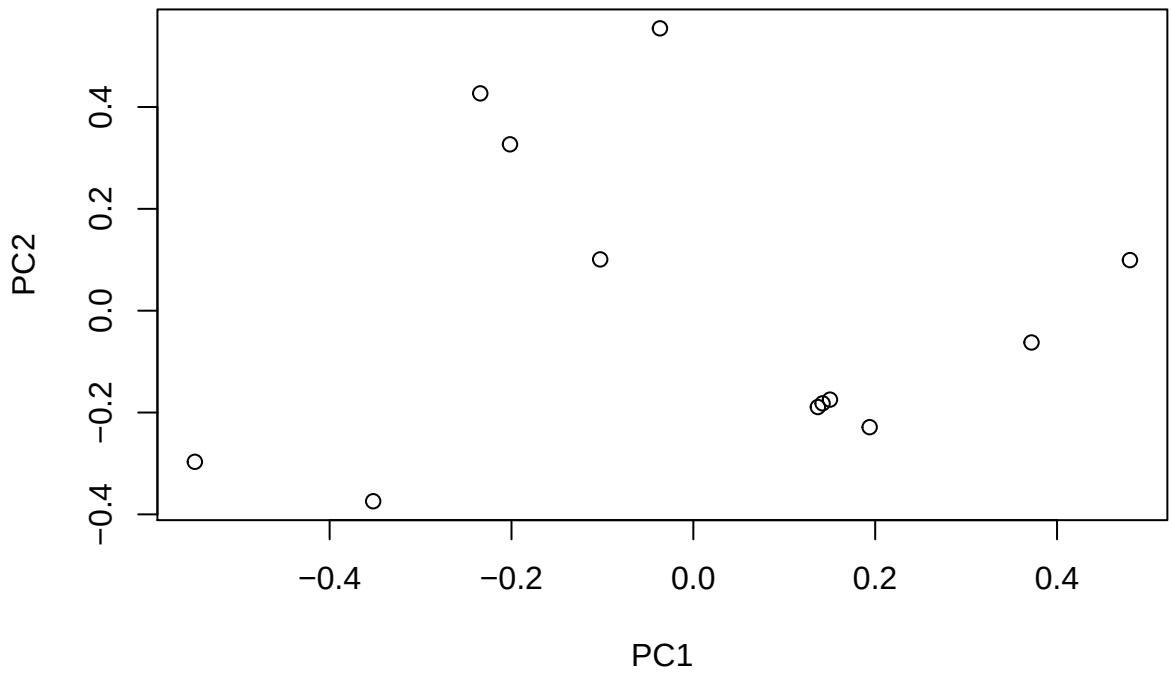
```
pca2<-prcomp(ex_scale)
plot(pca2)
```




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



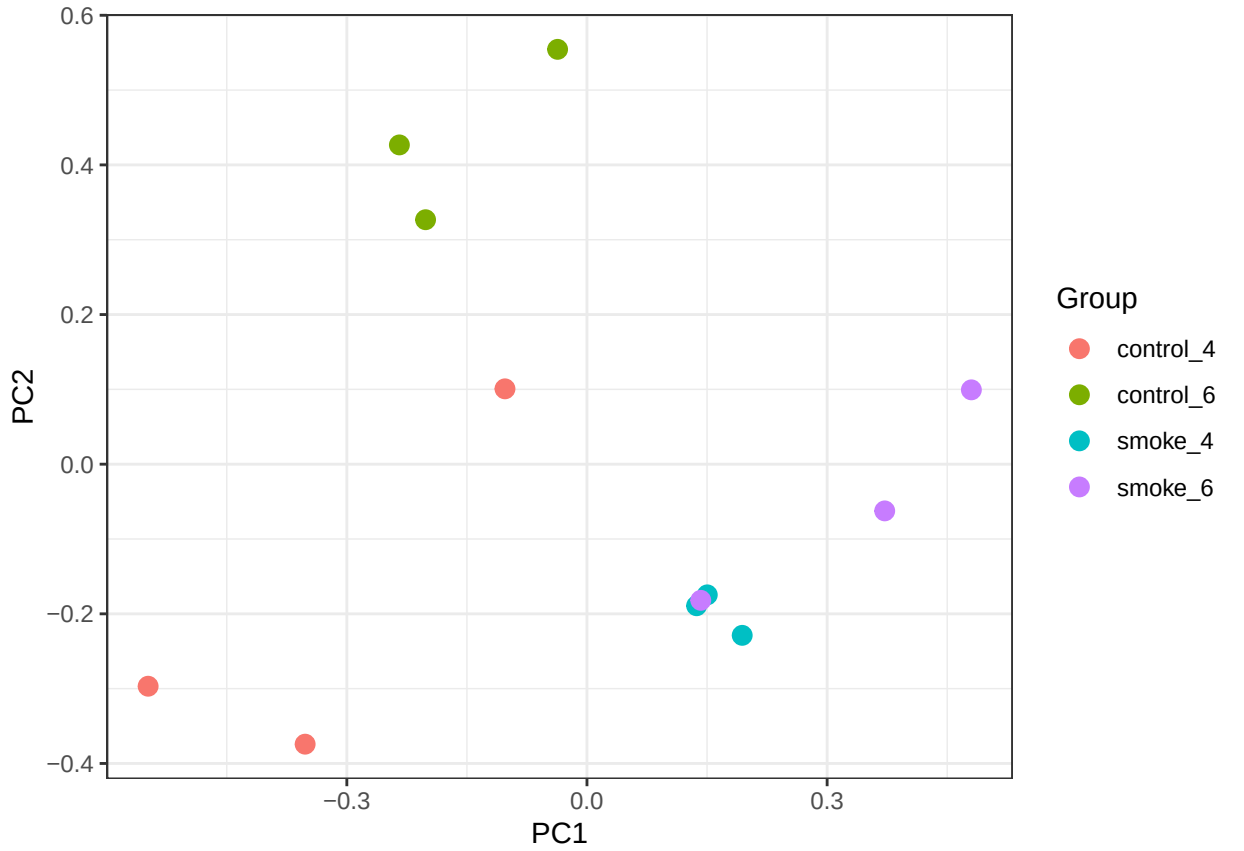

```
plot(pca2$r)
```




```
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()
```



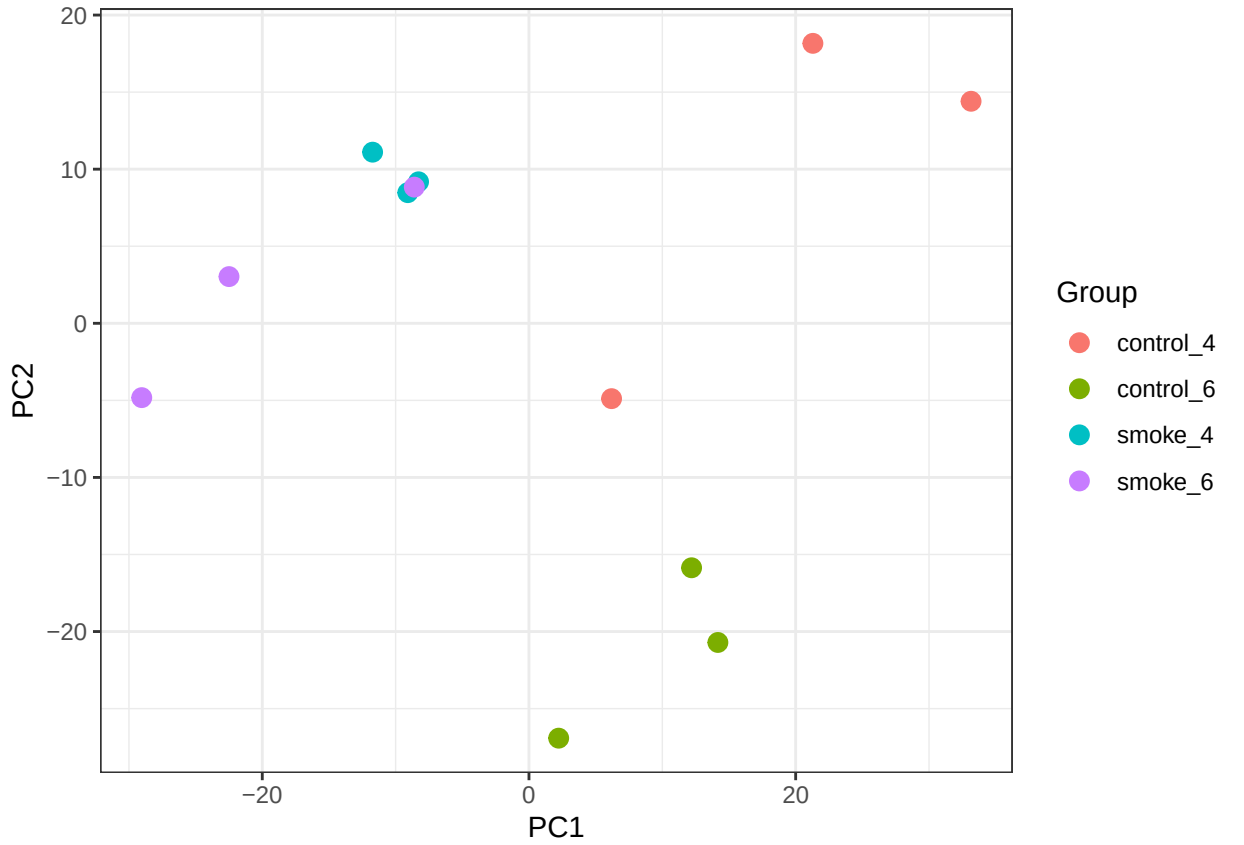

```
dev.off()
```

```
## 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"
```