Screening the data

Sungmin Ji

2023-12-07

```
# Data load
trainad <- anndata::read_h5ad("data/train_study.h5ad")
validad <- anndata::read_h5ad("data/valid_study.h5ad")
```

```
# Training data
cell_type <- c(trainad$obs[,1], validad$obs[,1])
unique(cell_type)
```

```
## [1] NK cells          Dendritic cells   CD4 T cells       B cells
## [5] FCGR3A+ Monocytes CD14+ Monocytes   CD8 T cells
## 7 Levels: CD4 T cells CD14+ Monocytes B cells CD8 T cells ... Dendritic cells
```

```
sample_name <- c(trainad$obs_names, validad$obs_names)
head(sample_name)
```

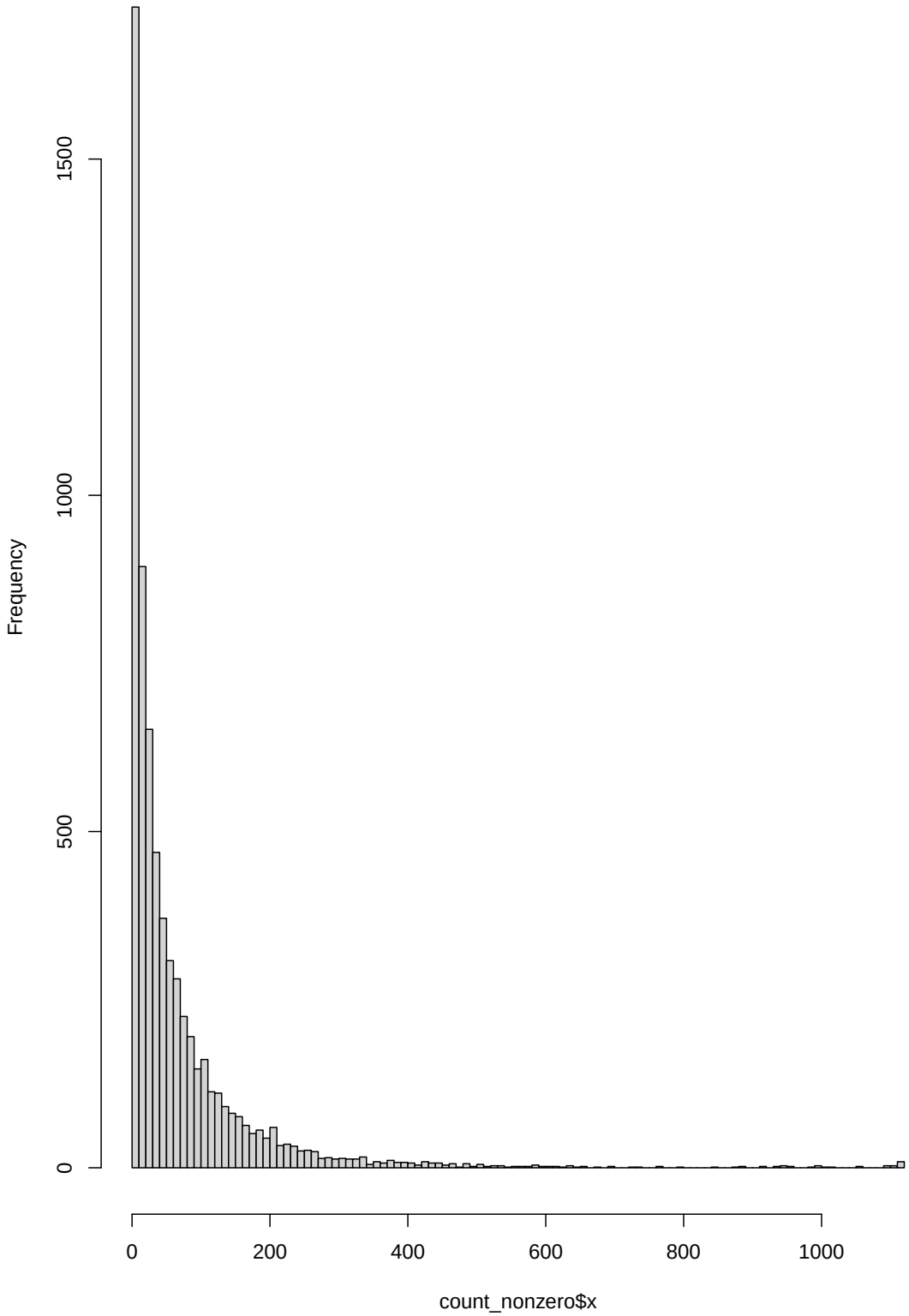
```
## [1] "AAACATACCAAGCT-1-stimulated-0" "AAACATACCCCTAC-1-stimulated-0"
## [3] "AAACATACCCGTAA-1-stimulated-0" "AAACATACCCTCGT-1-stimulated-0"
## [5] "AAACATACGAGGTG-1-stimulated-0" "AAACATACGCGAAG-1-stimulated-0"
```

```
i1=7 ### choose two types of cell
unique(cell_type)[i1]
```

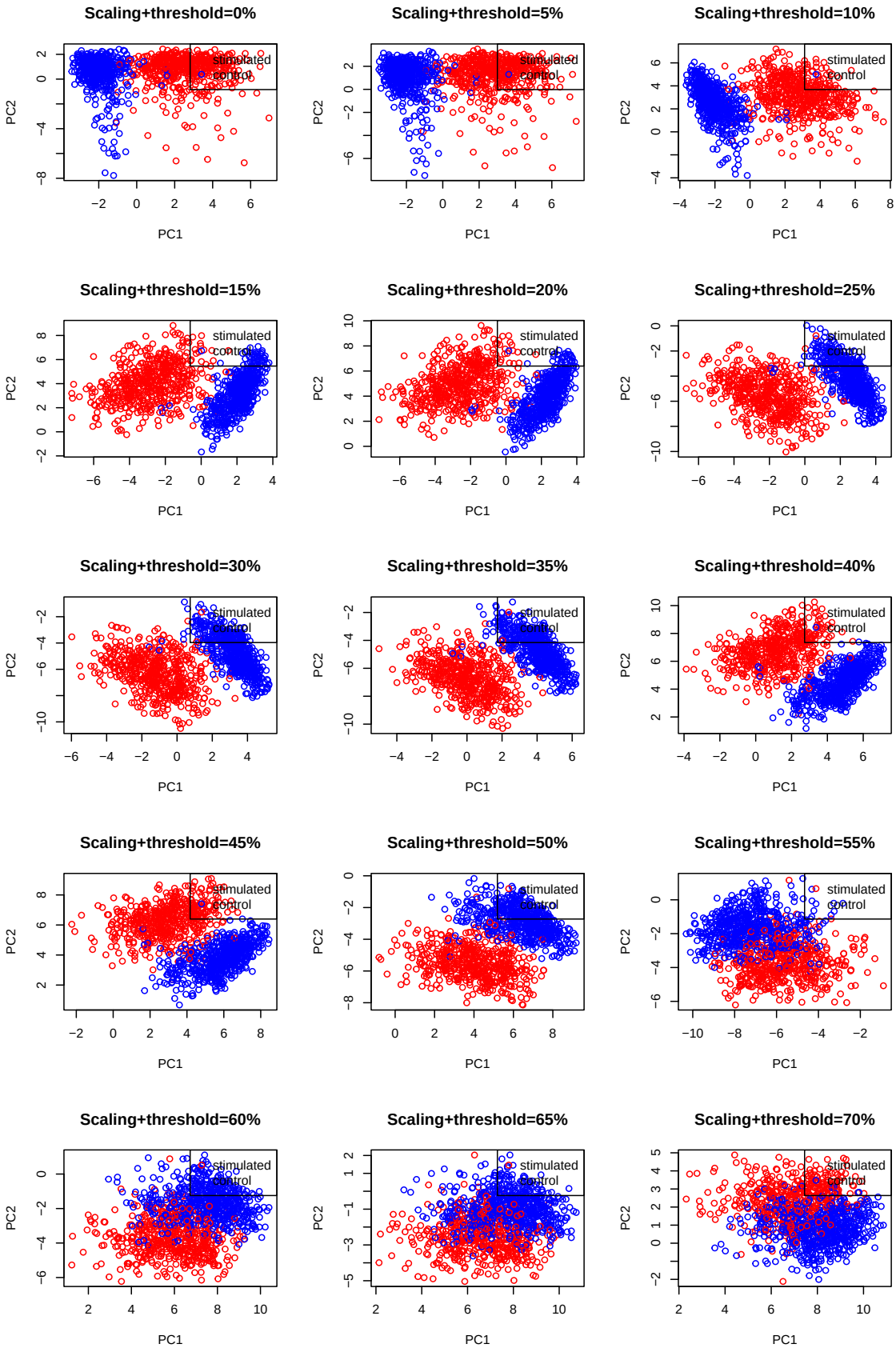
```
## [1] CD8 T cells
## 7 Levels: CD4 T cells CD14+ Monocytes B cells CD8 T cells ... Dendritic cells
```

```
# Screening genes procedure (Reduction of redundant dimensions)
X <- rbind(trainad$X, validad$X)

selected_mat <- X[cell_type == "CD8 T cells", ]

stim_ind <- grepl("stimulated", rownames(selected_mat))
cont_ind <- !stim_ind

## 1, -1 label
Y <- rep(0, nrow(selected_mat))
Y[stim_ind] = 1
Y[cont_ind] = -1
```

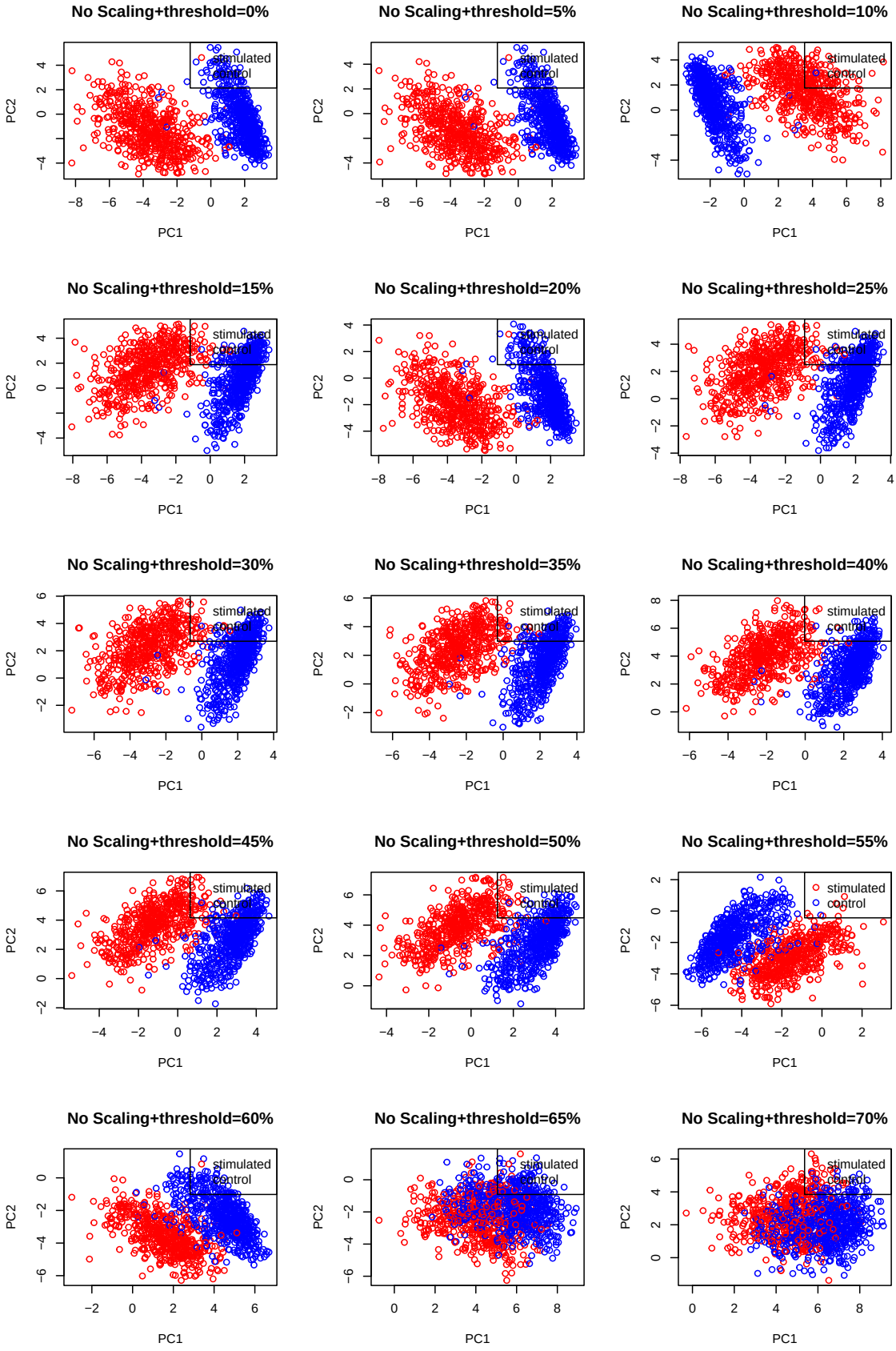
```r
## 1, 0 label
#Y <- rep(0, nrow(selected_mat))
#Y[stim_ind] = 1
#Y[cont_ind] = 0


## Choose genes based on two rules
### First screening the gene that has at least 10% proportion for non-zero individuals. (Remove rare ge
### for each gene, sample standard deviation is greater than 0.2, select the gene (check the histogram)
### find the intersection of two criteria


## First screening
n <- nrow(selected_mat)
count_nonzero <- aggregate(x ~ j, summary(selected_mat), length)
hist(count_nonzero$x, breaks=100)
```


Histogram of count_nonzero$x
Frequency
0
500
1000
1500
0
200
400
600
800
1000
count_nonzero$x

```r
ind_seq <- seq(0, 7, by=0.5) ## threshold proportion for non-zero individuals 0% ~ 90%
countid_mat <- matrix(F, nrow=7000, ncol=length(ind_seq))


par(mfrow=c(5,3))
colorsY <- vector("character", length=length(Y))
colorsY[Y==1] = "red"
colorsY[Y==-1] = "blue"

## PCA plot with scaling each column
for(i in 1:length(ind_seq)){
  countid_mat[count_nonzero$j[count_nonzero$x > (n*ind_seq[i])/10], i] <- T
  temp <- as.matrix(selected_mat[, countid_mat[,i]])
  pX_scale <- prcomp(temp, scale.=T)
  pc2 <- temp %*% pX_scale$rotation[,1:2]
  plot(pc2,col=colorsY, main=paste0("Scaling+threshold=", ind_seq[i]*10, "%"))
  legend("topright", pch=1,
       legend=c("stimulated", "control"), col=c("red", "blue"))
}
```


Scaling+threshold=0%
Scaling+threshold=5%
Scaling+threshold=10%
Scaling+threshold=15%
Scaling+threshold=20%
Scaling+threshold=25%
Scaling+threshold=30%
Scaling+threshold=35%
Scaling+threshold=40%
Scaling+threshold=45%
Scaling+threshold=50%
Scaling+threshold=55%
Scaling+threshold=60%
Scaling+threshold=65%
Scaling+threshold=70%
stimulated
control
PC1
PC2

```r
## PCA plot without scaling each column
par(mfrow=c(5,3))

for(i in 1:length(ind_seq)){
  temp <- as.matrix(selected_mat[, countid_mat[,i]])
  pX <- prcomp(temp, scale.=F)
  pc2 <- temp %*% pX$rotation[,1:2]
  plot(pc2,col=colorsY, main=paste0("No Scaling+threshold=", ind_seq[i]*10, "%"))
  legend("topright", pch=1,
       legend=c("stimulated", "control"), col=c("red", "blue"))
}
```


No Scaling+threshold=0%
No Scaling+threshold=5%
No Scaling+threshold=10%
No Scaling+threshold=15%
No Scaling+threshold=20%
No Scaling+threshold=25%
No Scaling+threshold=30%
No Scaling+threshold=35%
No Scaling+threshold=40%
No Scaling+threshold=45%
No Scaling+threshold=50%
No Scaling+threshold=55%
No Scaling+threshold=60%
No Scaling+threshold=65%
No Scaling+threshold=70%
stimulated
control
PC1
PC2

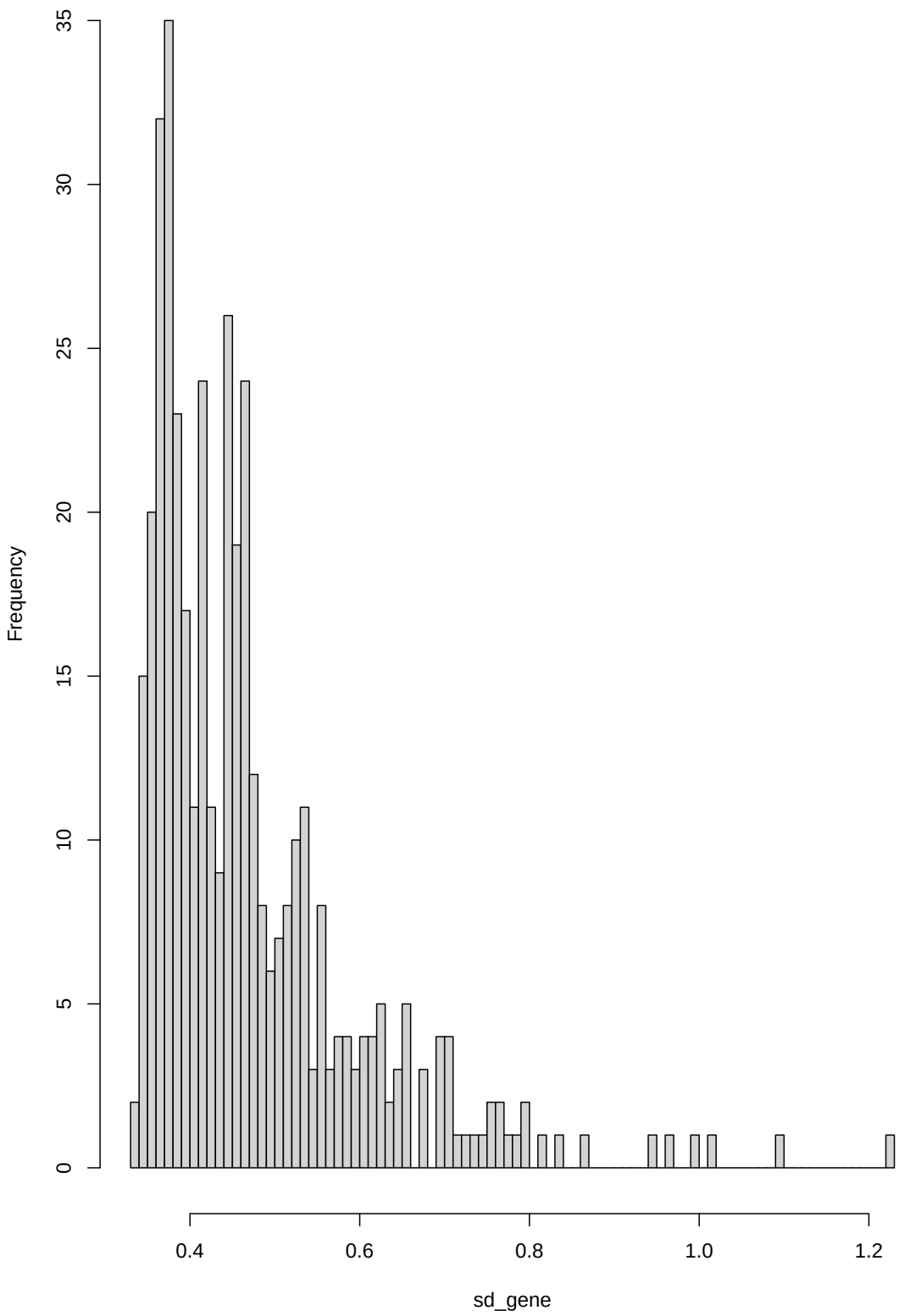
```r
count_id <- countid_mat[,5] ## I chose 25% as a threshold for the first screening based on the separati

screen_mat1 <- as.matrix(selected_mat[, count_id])
dim(screen_mat1)
```

```
## [1] 1115  410
```

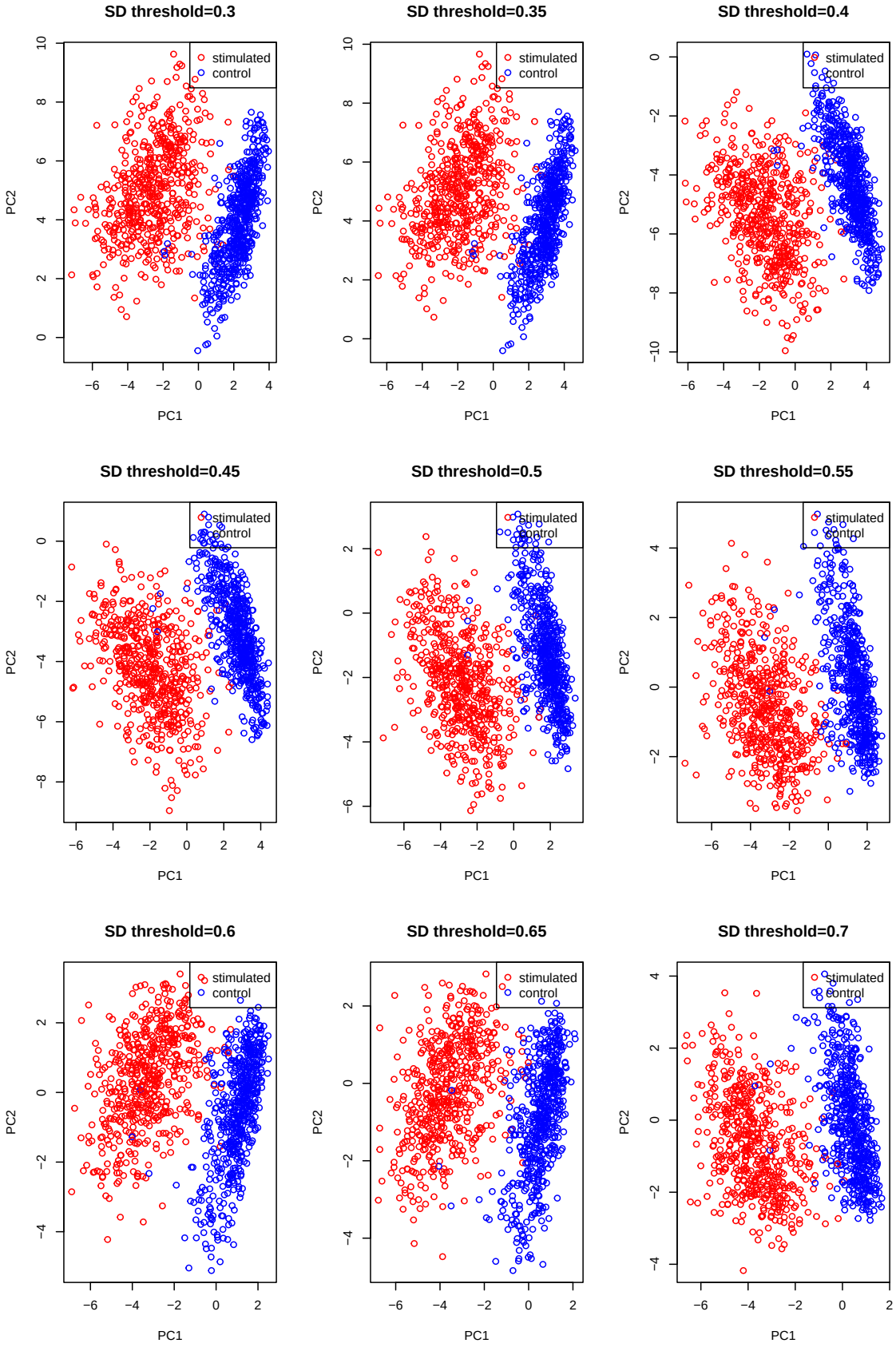
```r
## Randomly separate train and test data with a random seed
n_partition <- floor(nrow(screen_mat1)/5)
set.seed(2023)
train <- sample(c(rep(F, n_partition),
                  rep(T,nrow(screen_mat1)-n_partition)), replace=F)

result_mat1 <- cbind(Y, screen_mat1, train)
dim(result_mat1)
```

```
## [1] 1115  412
```

```r
write.csv(result_mat1, file="CD8Tcell_screened1.csv")


## Second Screening genes using sd
sd_gene <- apply(screen_mat1, 2, sd)
par(mfrow=c(1,1))
hist(sd_gene, breaks=100)
```


Histogram of sd_gene
Frequency
0
5
10
15
20
25
30
35
0.4
0.6
0.8
1.0
1.2
sd_gene

```r
sd_seq <- seq(0.30, 0.7, length=9)

par(mfrow=c(3,3))
sd_id_mat <- matrix(F, nrow=sum(count_id), ncol=length(sd_seq))

for(i in 1:length(sd_seq)){
  sd_id_mat[, i] <- apply(screen_mat1, 2, sd) > sd_seq[i]
  temp <- screen_mat1[, sd_id_mat[,i]]
  pX_scale <- prcomp(temp, scale.=T)
  pc2 <- temp %*% pX_scale$rotation[,1:2]
  plot(pc2,col=colorsY, main=paste0("SD threshold=", sd_seq[i]))
  legend("topright", pch=1,
       legend=c("stimulated", "control"), col=c("red", "blue"))
}
```

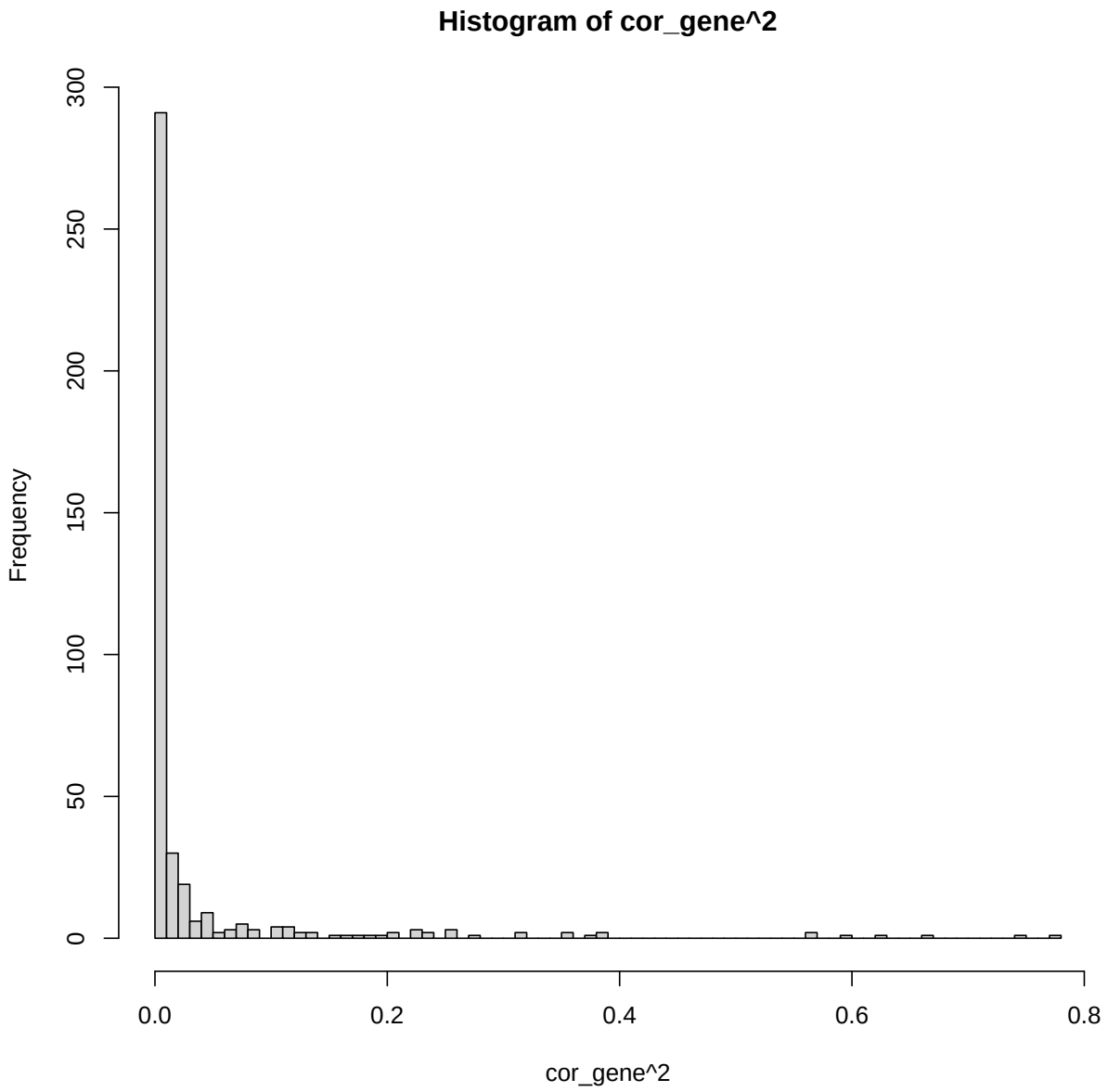

SD threshold=0.3
SD threshold=0.35
SD threshold=0.4
SD threshold=0.45
SD threshold=0.5
SD threshold=0.55
SD threshold=0.6
SD threshold=0.65
SD threshold=0.7
stimulated
control
PC1
PC2

```r
sd_id <- sd_id_mat[, 2] ## choose sd=0.35
screen_mat2 <- screen_mat1[, sd_id]

result_mat2 <- cbind(Y, screen_mat2, train)
dim(result_mat2)
```

```
## [1] 1115  395
```

```r
write.csv(result_mat2, file="CD8Tcell_screened2sd.csv")
```

```r
## Second Screening genes using correlation between Y
cor_seq <- seq(0.15, 0.35, length=5)
cor_gene <- apply(screen_mat1, 2, FUN=function(x) cor(Y,x))
par(mfrow=c(1,1))
hist(cor_gene^2, breaks=100)
```



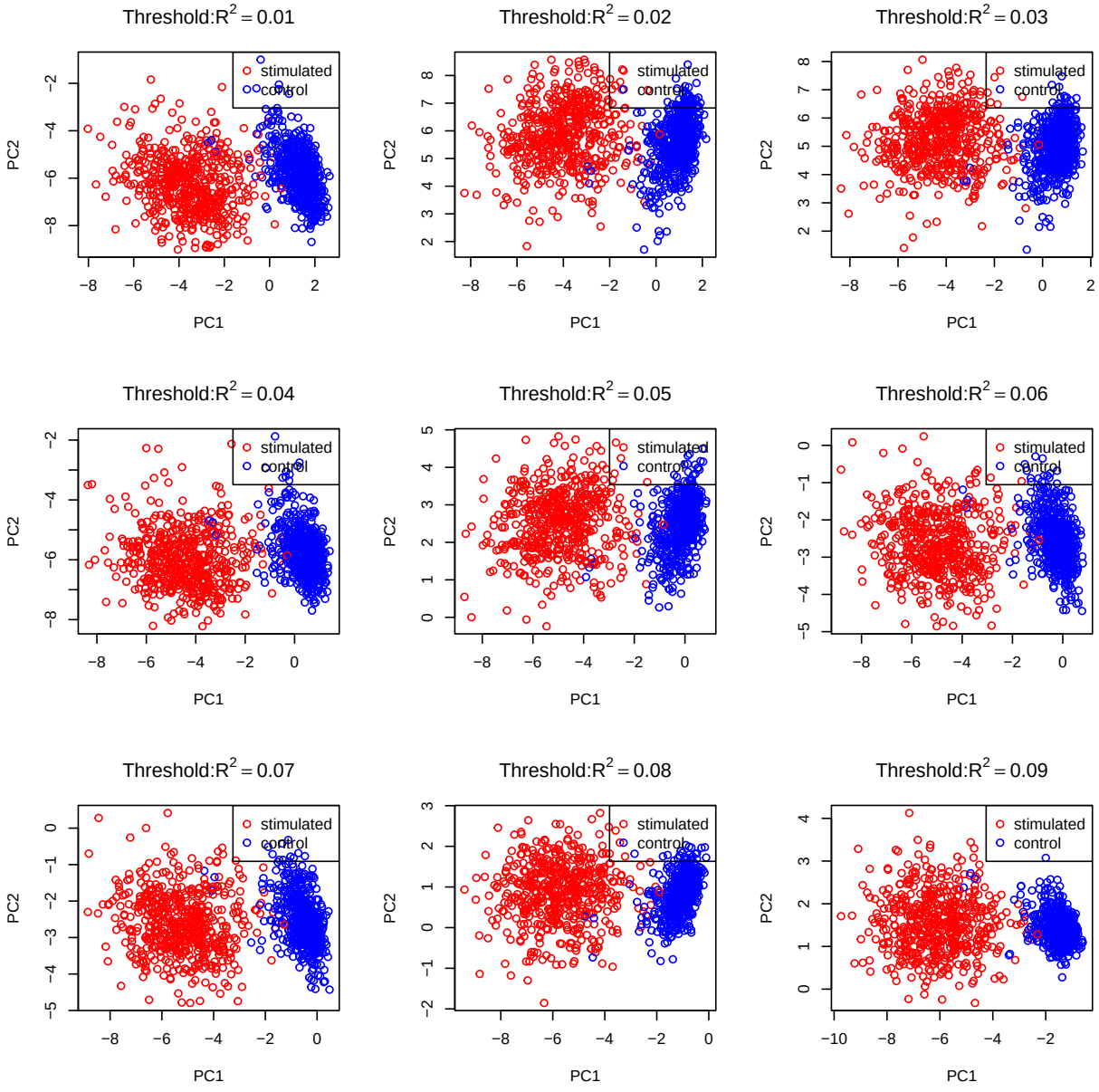

```
sum(cor_gene^2 > 0.05)
```

```
## [1] 55
```

```
cor_seq <- seq(0.01, 0.09, by=0.01)

par(mfrow=c(3,3))
cor_id_mat <- matrix(F, nrow=sum(count_id), ncol=length(cor_seq))
length(cor_seq)
```

```
## [1] 9
```

```
for(i in 1:length(cor_seq)){
  cor_id_mat[, i] <- apply(screen_mat1, 2, FUN=function(x) cor(Y, x)^2) > cor_seq[i]
  temp <- screen_mat1[, cor_id_mat[, i]]
  pX_scale <- prcomp(temp, scale.=T)
  pc2 <- temp %*% pX_scale$rotation[,1:2]
  plot(pc2,col=colorsY, main=bquote(.("Threshold:R")^2==.(cor_seq[i])))
  legend("topright", pch=1,
       legend=c("stimulated", "control"), col=c("red", "blue"))
}
```




```
cor_id <- cor_id_mat[, 6] ## choose R^2=0.06
screen_mat2_2 <- screen_mat1[, cor_id]
result_mat2_2 <- cbind(Y, screen_mat2_2, train)
```

```r
dim(result_mat2_2)
```

```
## [1] 1115   55
```

```r
write.csv(result_mat2_2, file="CD8Tcell_screened2cor.csv")
```