C1\_Project.V.1.1:Human

yejg

2017/12/25

## We will analysis the **Single cell RNA-seq data(gene count) from Human**.There are **1099** sample in the origin file.Before analysis,I had combined them into a dataframe and save it to txt file(*R or shell*).Based on the infornmation from the sample xlxs file(sample message file),we know that there are four control condition in some sample.And we will show them bellow.

### Load the packages

library(Seurat)  
library(data.table)  
library(NMF)  
library(rsvd)  
library(Rtsne)  
library(ggplot2)  
library(cowplot)  
library(sva)  
library(igraph)  
library(cccd)  
library(KernSmooth)  
library(beeswarm)  
library(stringr)  
library(formatR)  
source("tools.R")  
library(DESeq2)

### The function will be used in the follow

## We have two steps to analysis:

Condition message **Step 1: Ignore control condition,use all sample** **Step 2:Under the control condition** *Positive* *Negative* *Tube\_1* *Tube\_2*

## Step 1: All data: Analysis based on sample group

### Read data

### Data QA

human.only.pro <- Load\_data(data\_dir = "../data/human.txt")  
important.genes <- c("ITGB4", "ABCB5", "KRT19", "ACTB", "KRT12", "KRT5", "GAPDH",   
 "KRT3", "PAX6", "WNT7A", "KRT14", "TP63", "KRT10")  
human.only.pro <- human.only.pro[, colnames(human.only.pro)[unlist(lapply(colnames(human.only.pro),   
 function(x) return(str\_split(x, "\_")[[1]][2]))) %in% c("10um", "20um", "6um")]]  
human.only.pro <- human.only.pro[, colnames(human.only.pro)[!unlist(lapply(colnames(human.only.pro),   
 function(x) return(str\_split(x, "\_")[[1]][1]))) %in% c("hc001", "shoutiao")]]  
  
  
table(unlist(lapply(colnames(human.only.pro), function(x) return(str\_split(x,   
 "\_")[[1]][2]))))

##   
## 10um 20um 6um   
## 326 560 159

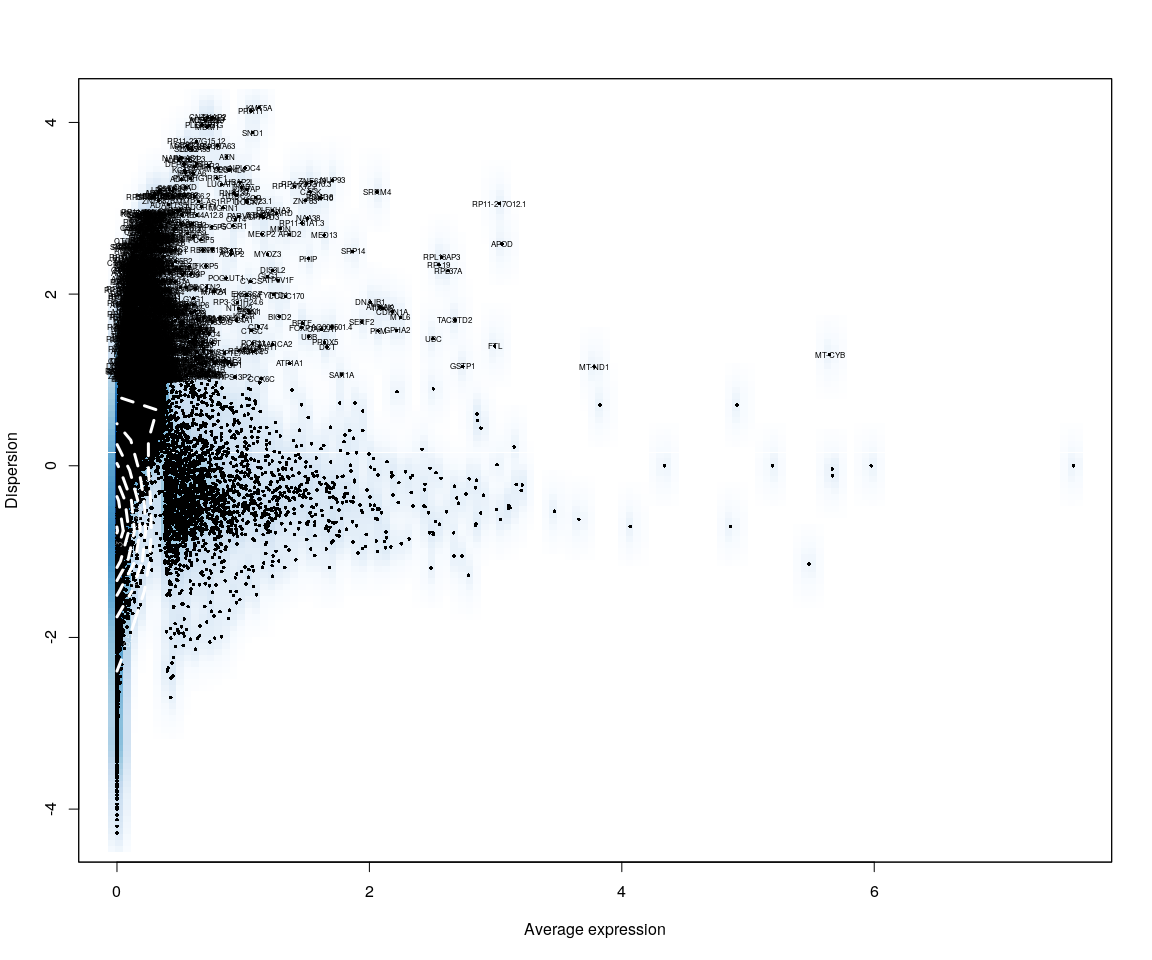
table(unlist(lapply(colnames(human.only.pro), function(x) return(str\_split(x,   
 "\_")[[1]][1]))))

##   
## hc006 hc009 hc012 hc017 hc018 hc020 hc021   
## 92 187 66 170 188 184 158

### Create Seurat object and not caculate DESeq,but not set **min.cells** and **min.genes**

# only select the cells contain 10 genes expressed at least,select the genes  
# must be expressed in two cells at least  
human.all.DESeq <- DESeq\_SeuratObj(X = human.only.pro, DESq = FALSE, min.cells = 10,   
 min.genes = 2)

## [1] "Scaling data matrix"  
##   
 |   
 | | 0%  
 |   
 |=================================================================| 100%

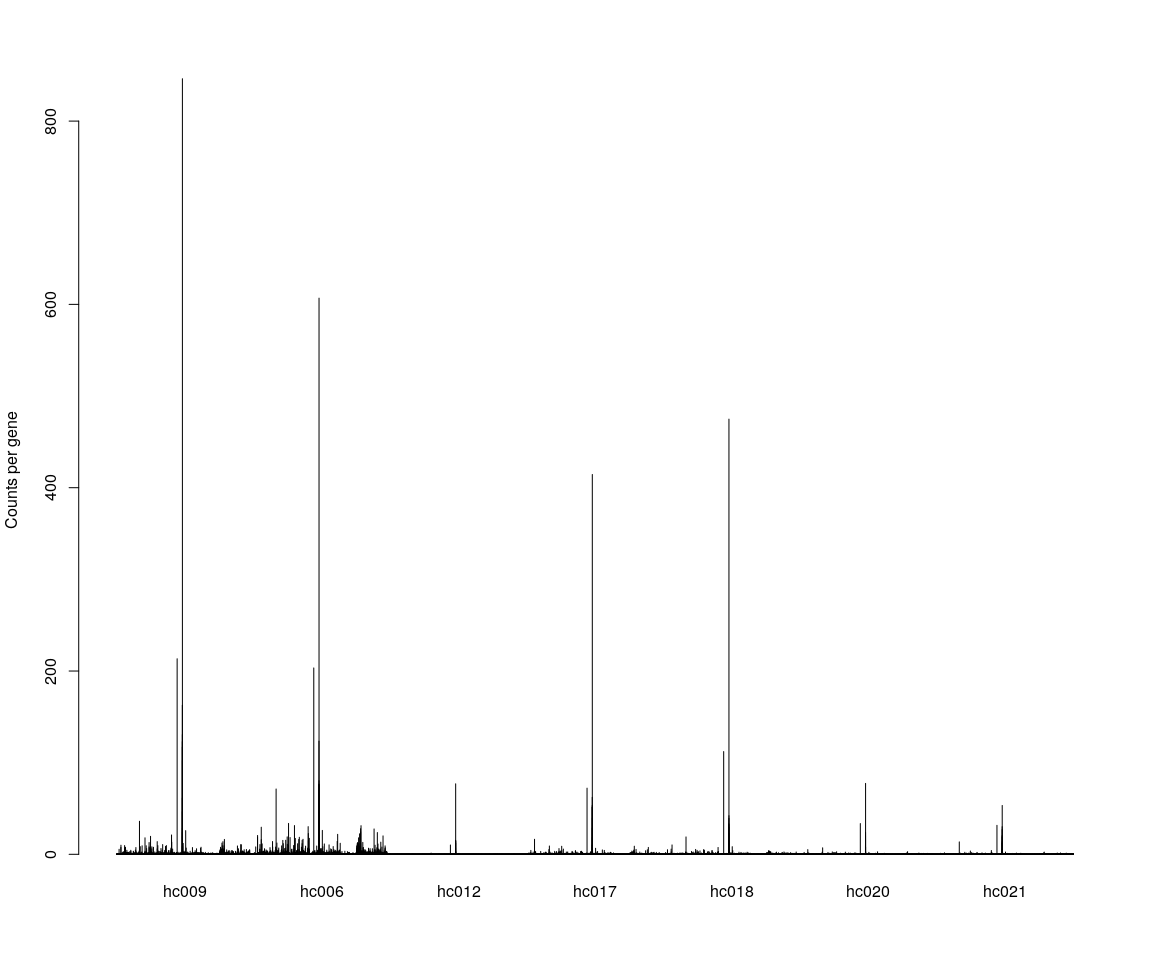


all.sample.group <- unlist(lapply(human.all.DESeq@cell.names, function(x) return(str\_split(x,   
 "\_")[[1]][1])))  
all.sample.size <- unlist(lapply(human.all.DESeq@cell.names, function(x) return(str\_split(x,   
 "\_")[[1]][2])))  
# reset ident human.all.DESeq<-SetIdent(human.all.DESeq,cells.use =  
# human.all.DESeq@cell.names,ident.use = all.sample.size)

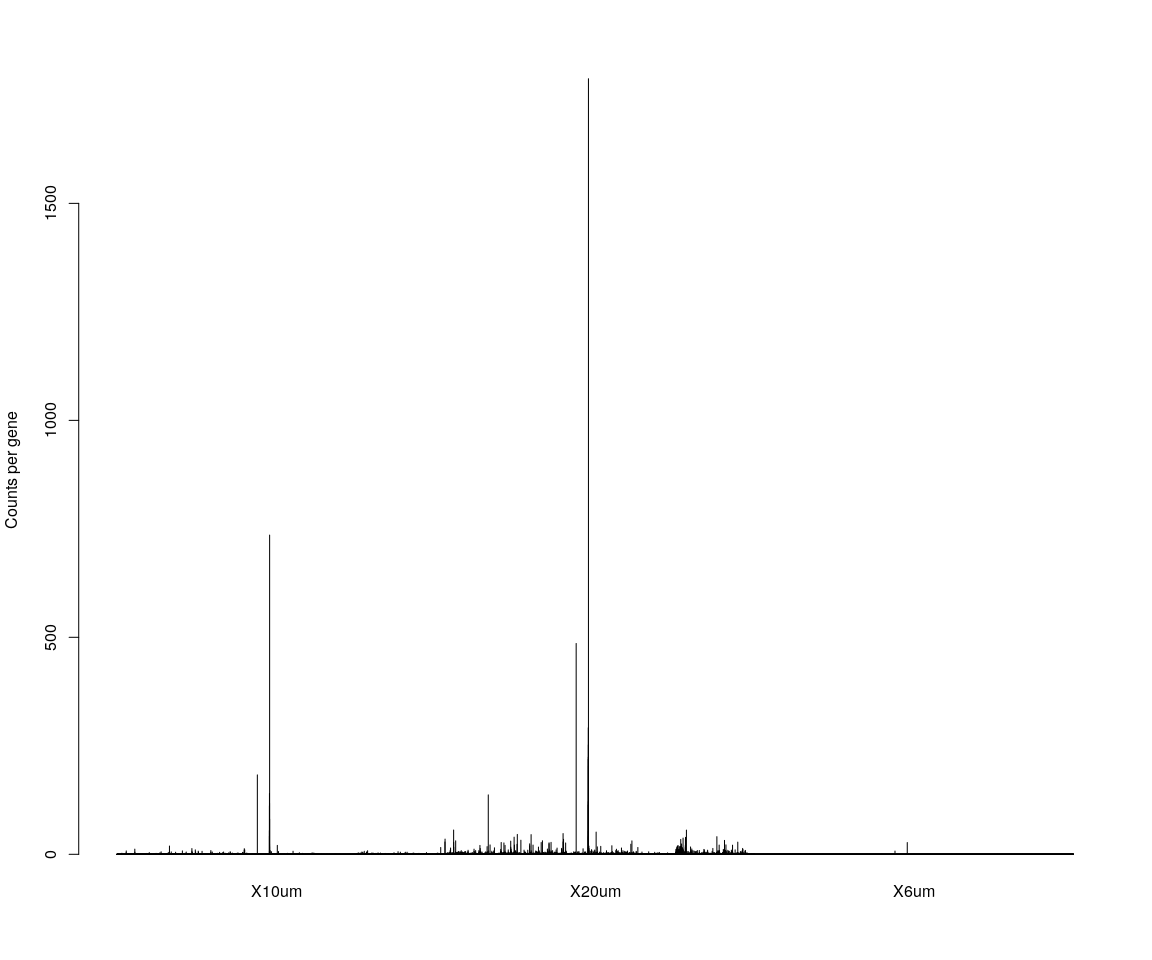
## Figure Explore

### First,use the plot,eg. Barplot,Violin…,we can explore some message from sample

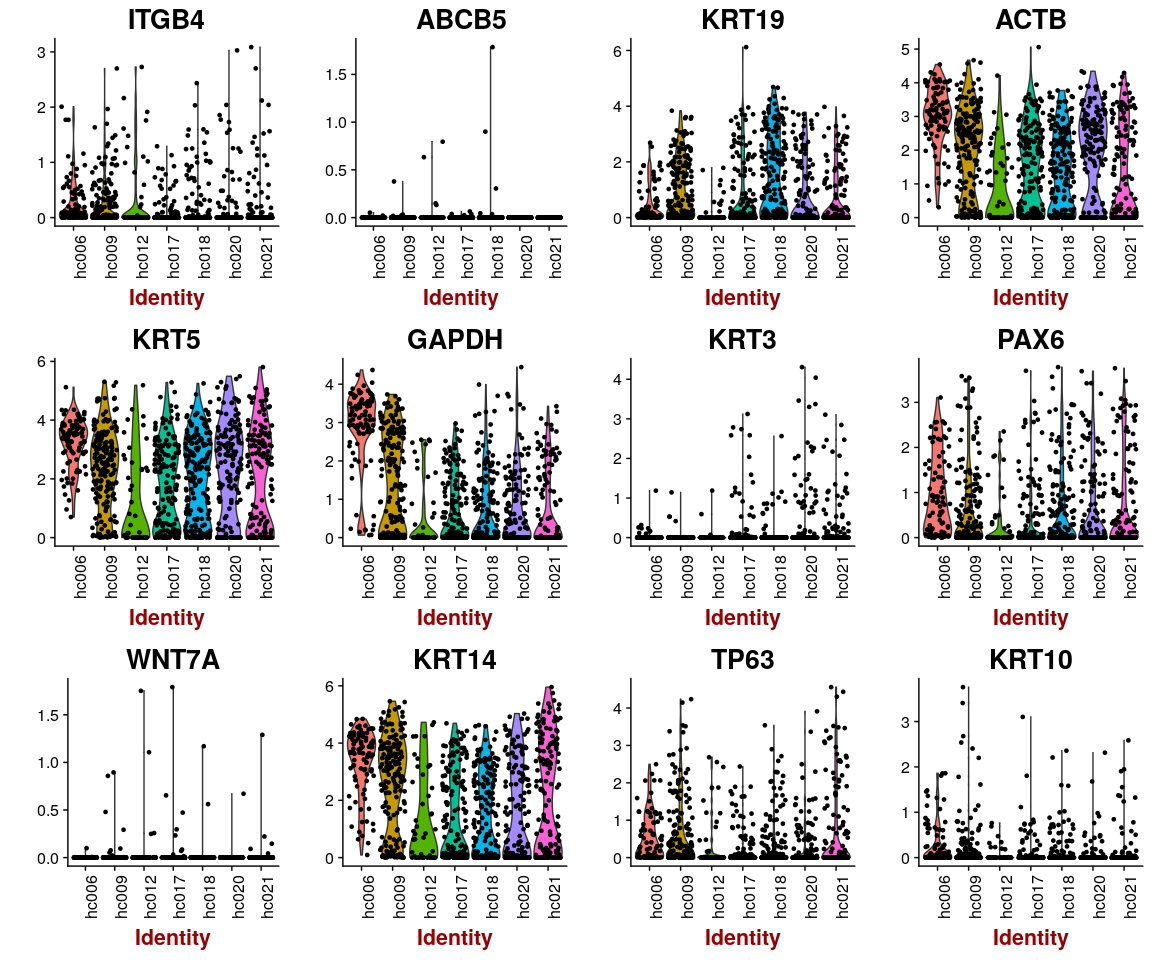
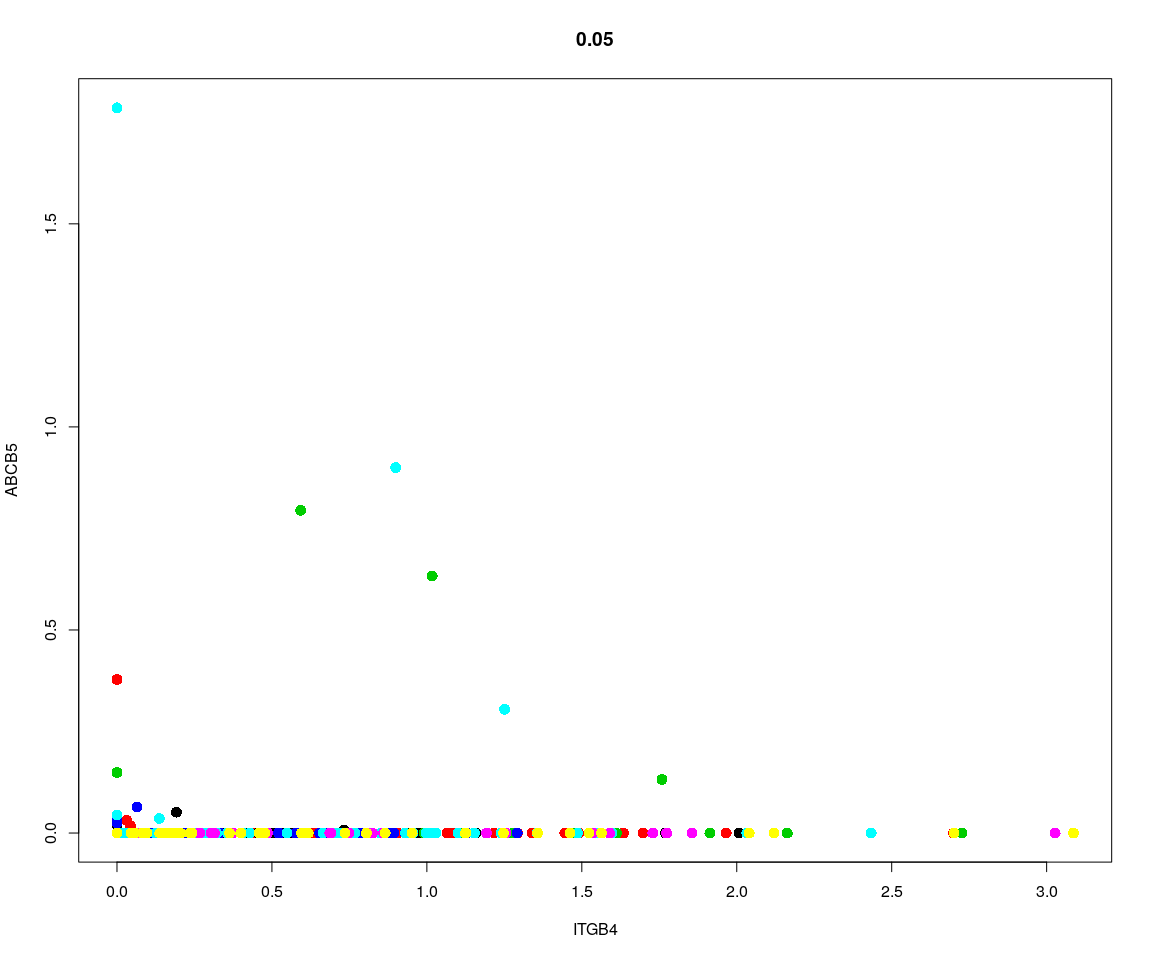
Group\_Bar(human.all.DESeq@raw.data, group = all.sample.group)



Group\_Bar(human.all.DESeq@raw.data, group = all.sample.size)



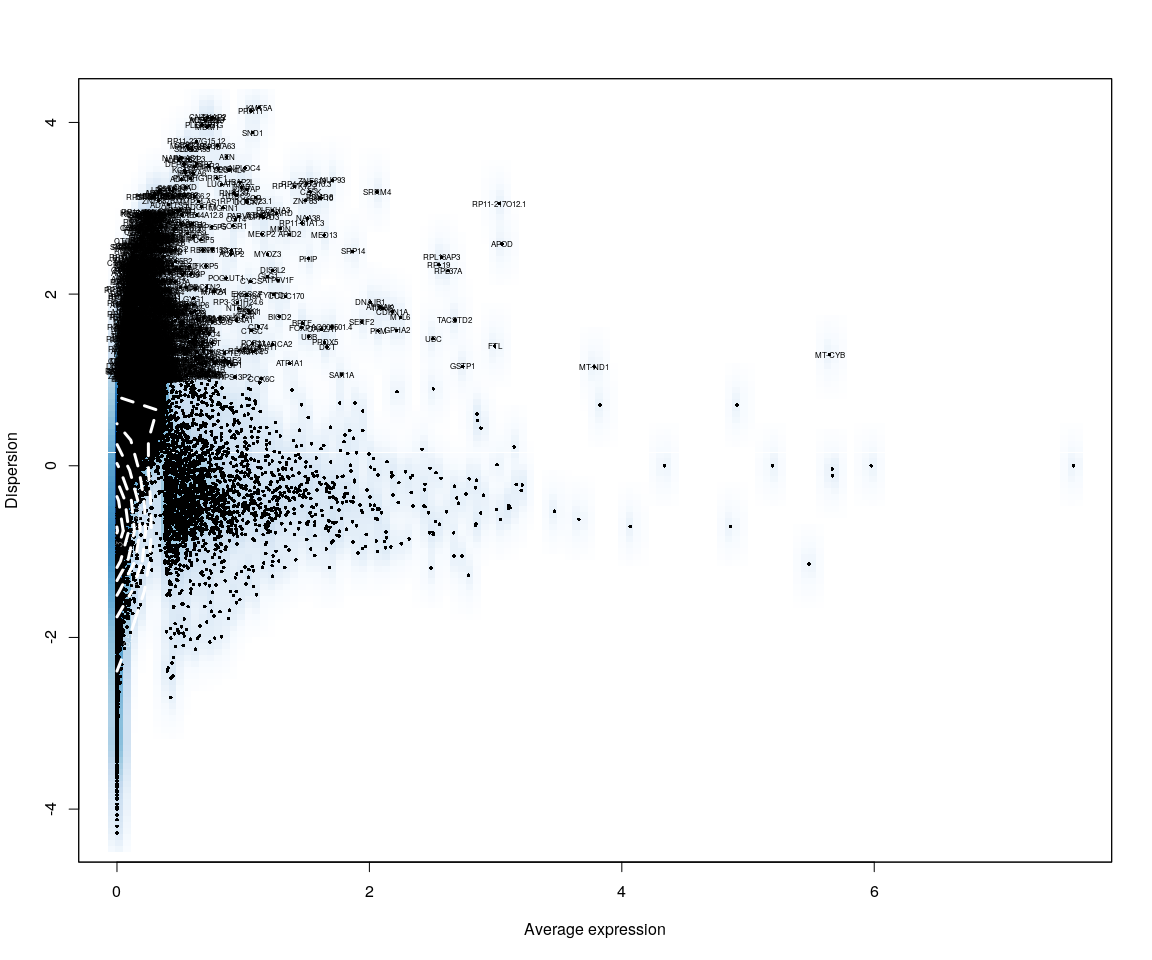
# We are interested in the gene ITGB4  
GenePlot(human.all.DESeq, gene1 = "ITGB4", gene2 = important.genes[2])  
# VlnPlot(human.all.DESeq,features.plot = 'ITGB4',y.lab.rot = 90) # Violinn  
# plot of gene ITGB in all sample  
VlnPlot(human.all.DESeq, features.plot = important.genes[important.genes %in%   
 rownames(human.all.DESeq@raw.data)], y.lab.rot = 90) # Violinn plot of gene ITGB in all sample



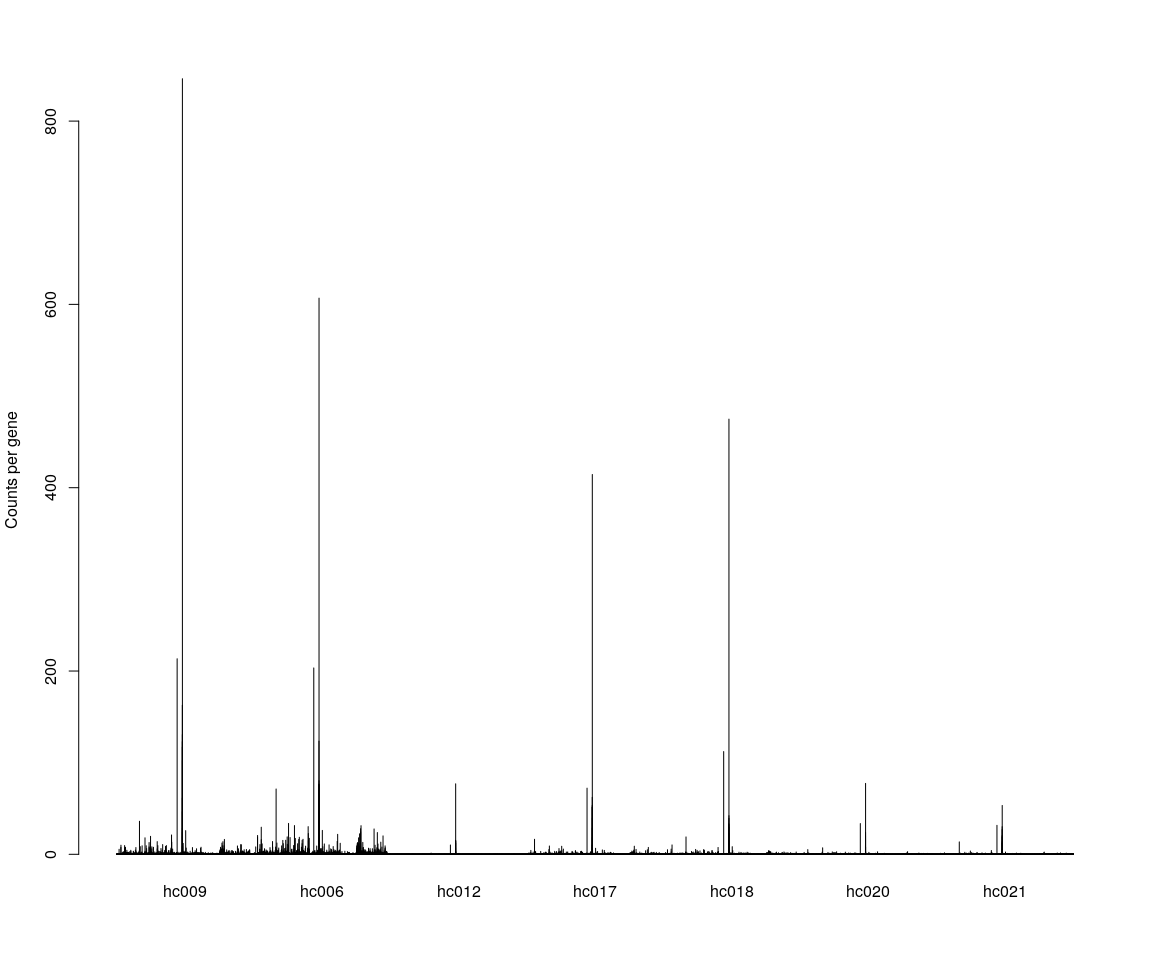
According to the plot of explore data,we can know that the gene ITGB4, ABCB5, KRT19, ACTB, KRT12, KRT5, GAPDH, KRT3, PAX6, WNT7A, KRT14, TP63, KRT10 are expressed differently across the sample group.It is fit to what we are interested.And,across the sample(Barplot),there are many genes that actually expressed different.The barplot of sample group tells us that the gene expression level in *hc001* is almost 0.So we consider remove the sample group *hc001*. Next,based on the explore plot,we will analysis data deeply with other method

### According to barplot across sample group,remove the hc001 group.

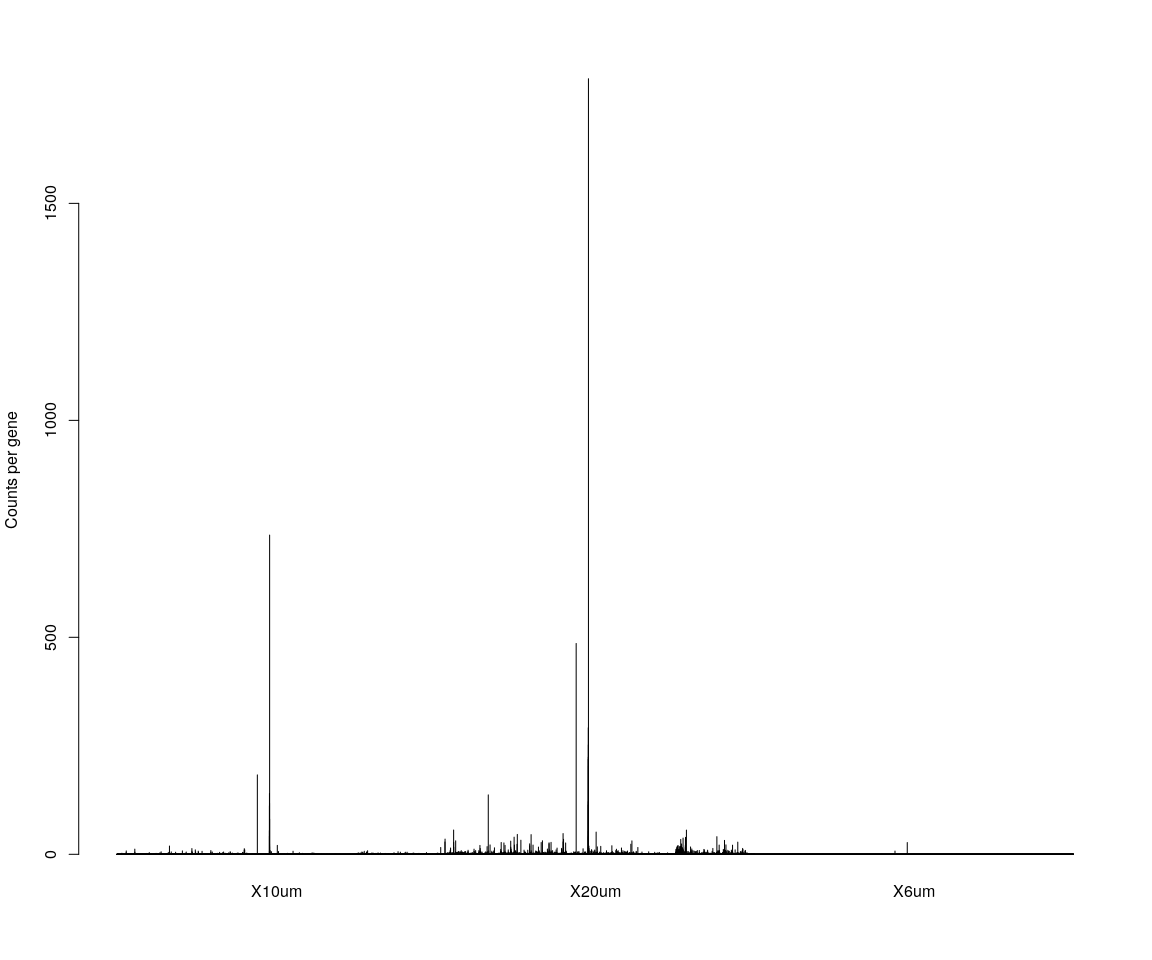
## [1] "Scaling data matrix"  
##   
 |   
 | | 0%  
 |   
 |=================================================================| 100%



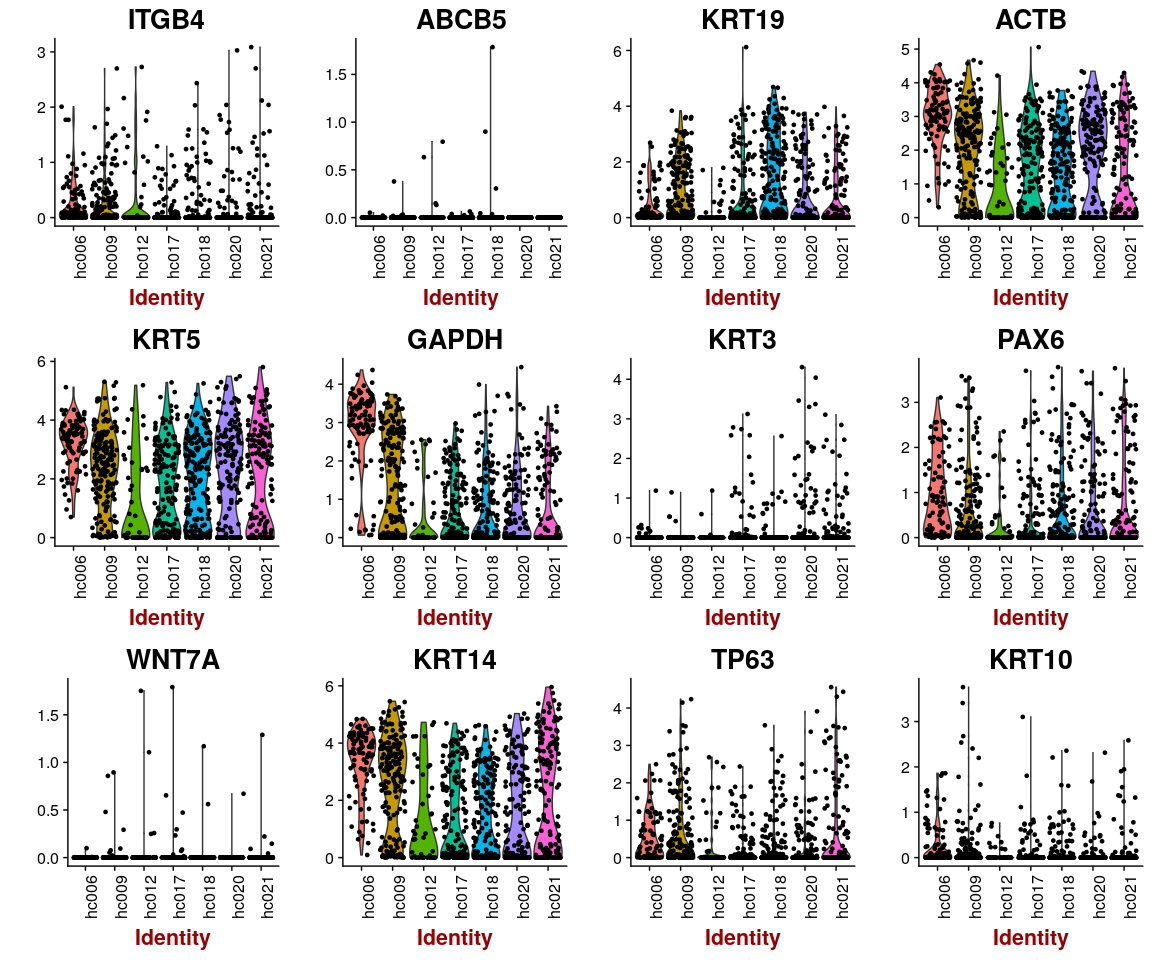
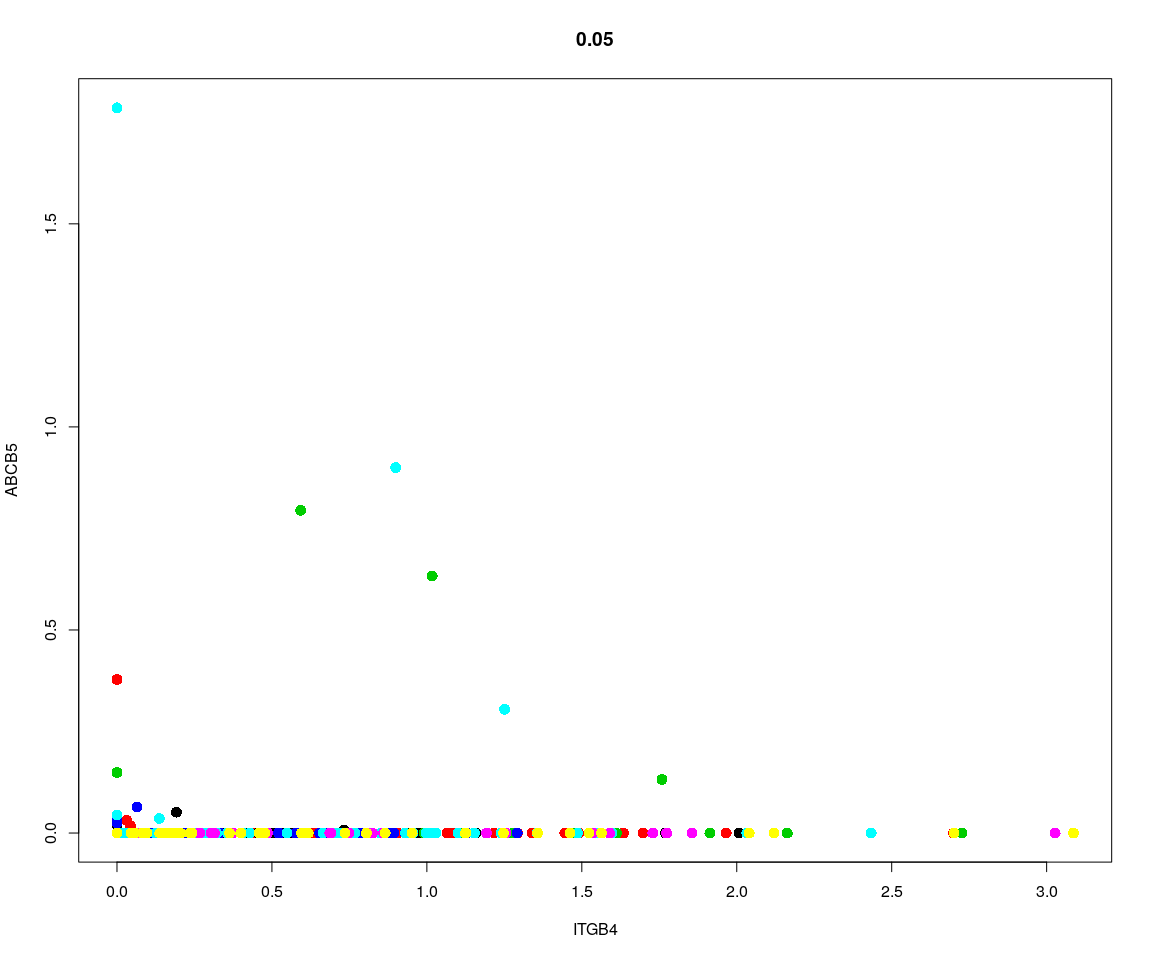
Group\_Bar(human.all.DESeq@raw.data, group = all.sample.group)



Group\_Bar(human.all.DESeq@raw.data, group = all.sample.size)



# We are interested in the gene ITGB4  
GenePlot(human.all.DESeq, gene1 = "ITGB4", gene2 = important.genes[2])  
# VlnPlot(human.all.DESeq,features.plot = 'ITGB4',y.lab.rot = 90) # Violinn  
# plot of gene ITGB in all sample  
VlnPlot(human.all.DESeq, features.plot = important.genes[important.genes %in%   
 rownames(human.all.DESeq@raw.data)], y.lab.rot = 90) # Violinn plot of gene ITGB in all sample



## Dimensionality reduction

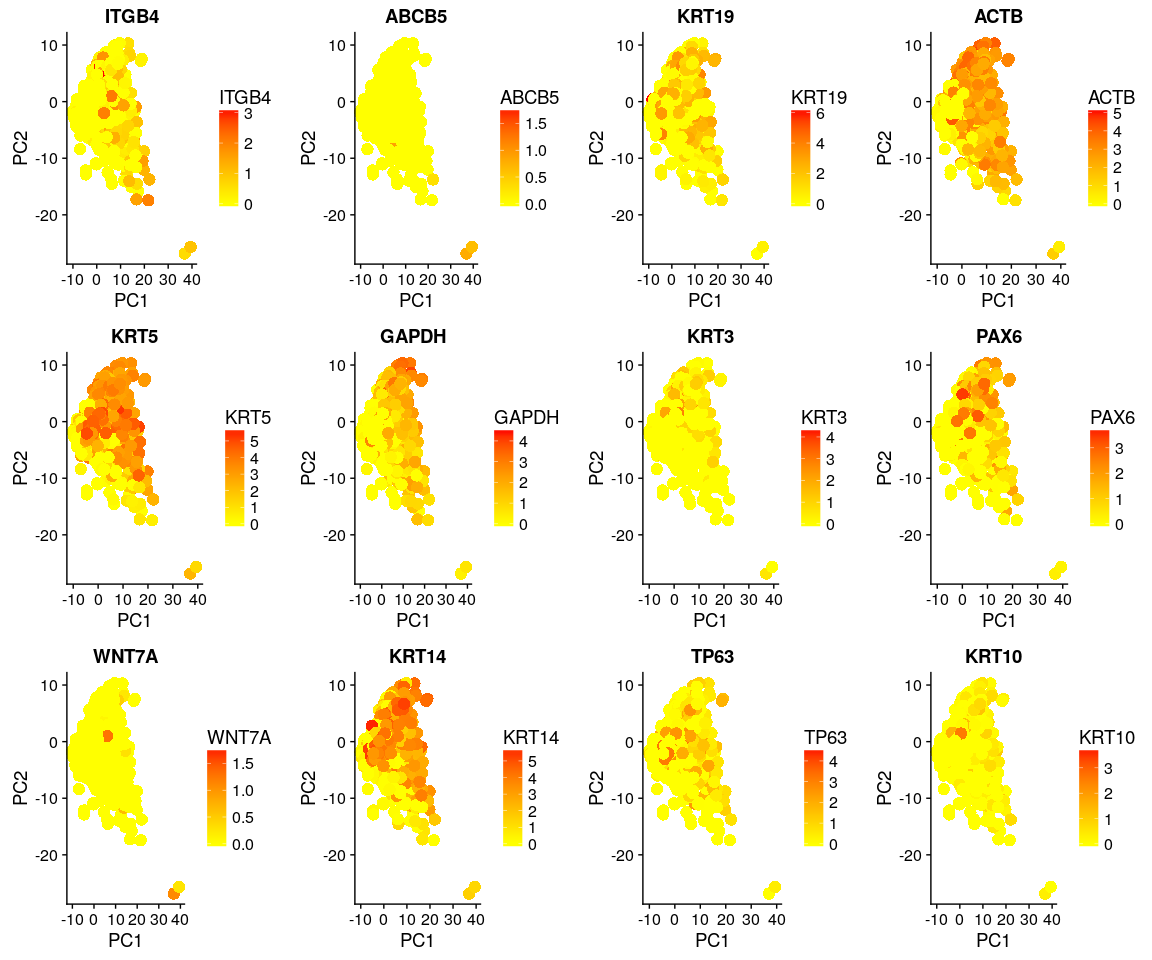
### **PCA** and **tSNE**

Here,do the dimensionality reduction using the PCA, tSNE method

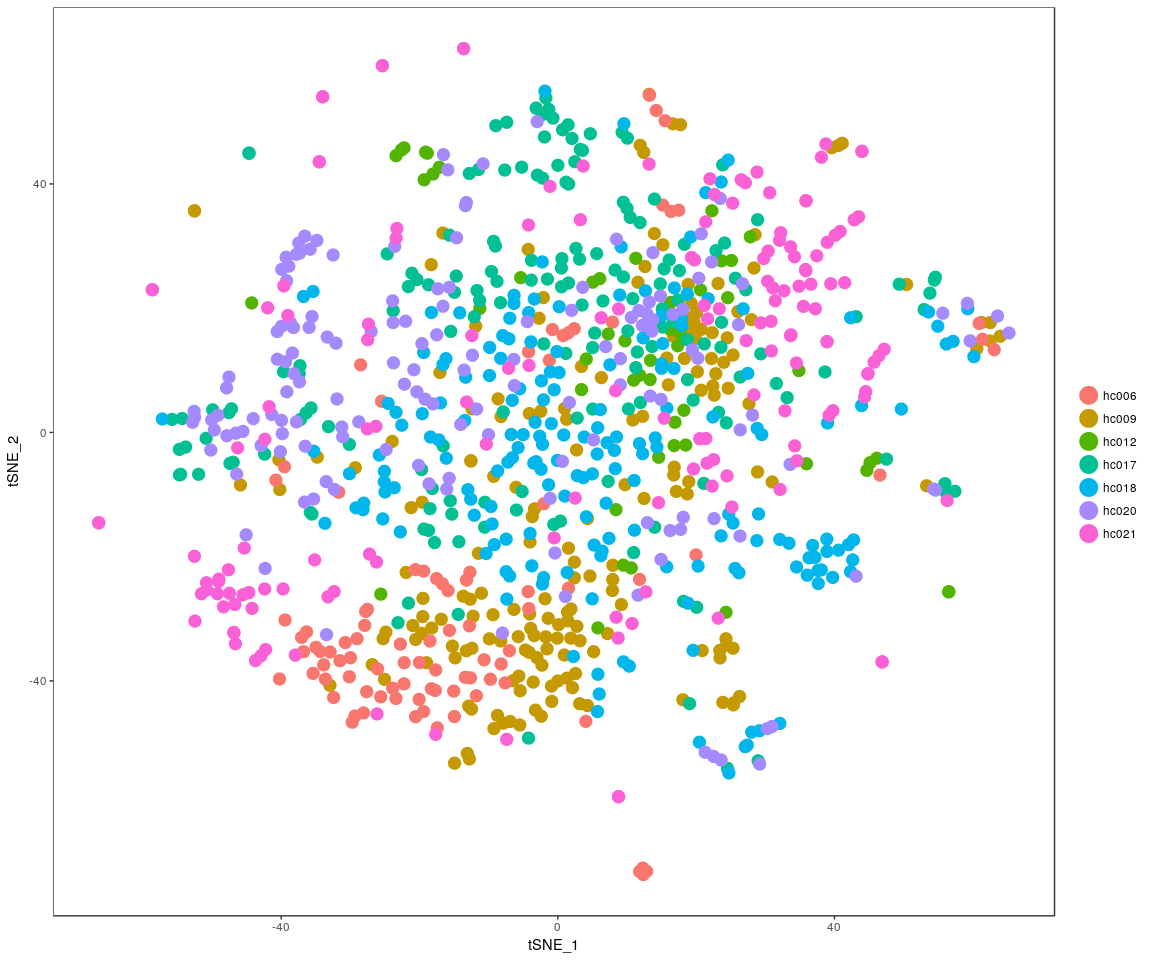
all.pbmc <- PCA.TSNE(object = human.all.DESeq, pcs.compute = FALSE, num.pcs = 28)

### After the PCA and tSNE,try plot: Featureplot of **ITGB4**,four var.genes,PCA plot,tSNE plot…

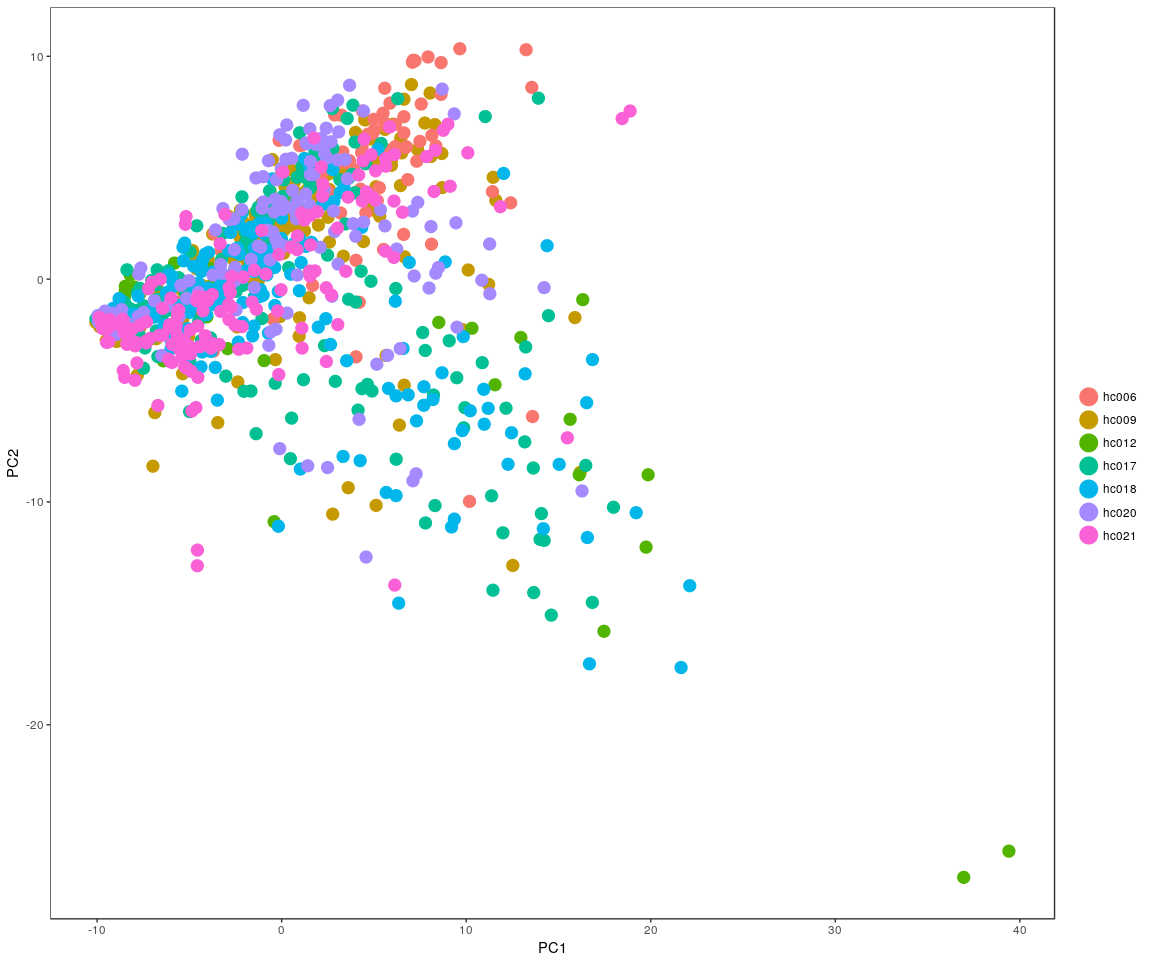
# FeaturePlot(object = all.pbmc,features.plot ='ITGB4',pt.size = 4,no.legend  
# = FALSE) # ITGB4 gene in part dataset  
FeaturePlot(object = all.pbmc, features.plot = important.genes[important.genes %in%   
 rownames(human.all.DESeq@raw.data)], pt.size = 4, no.legend = FALSE, reduction.use = "pca") # ITGB4 gene in part dataset



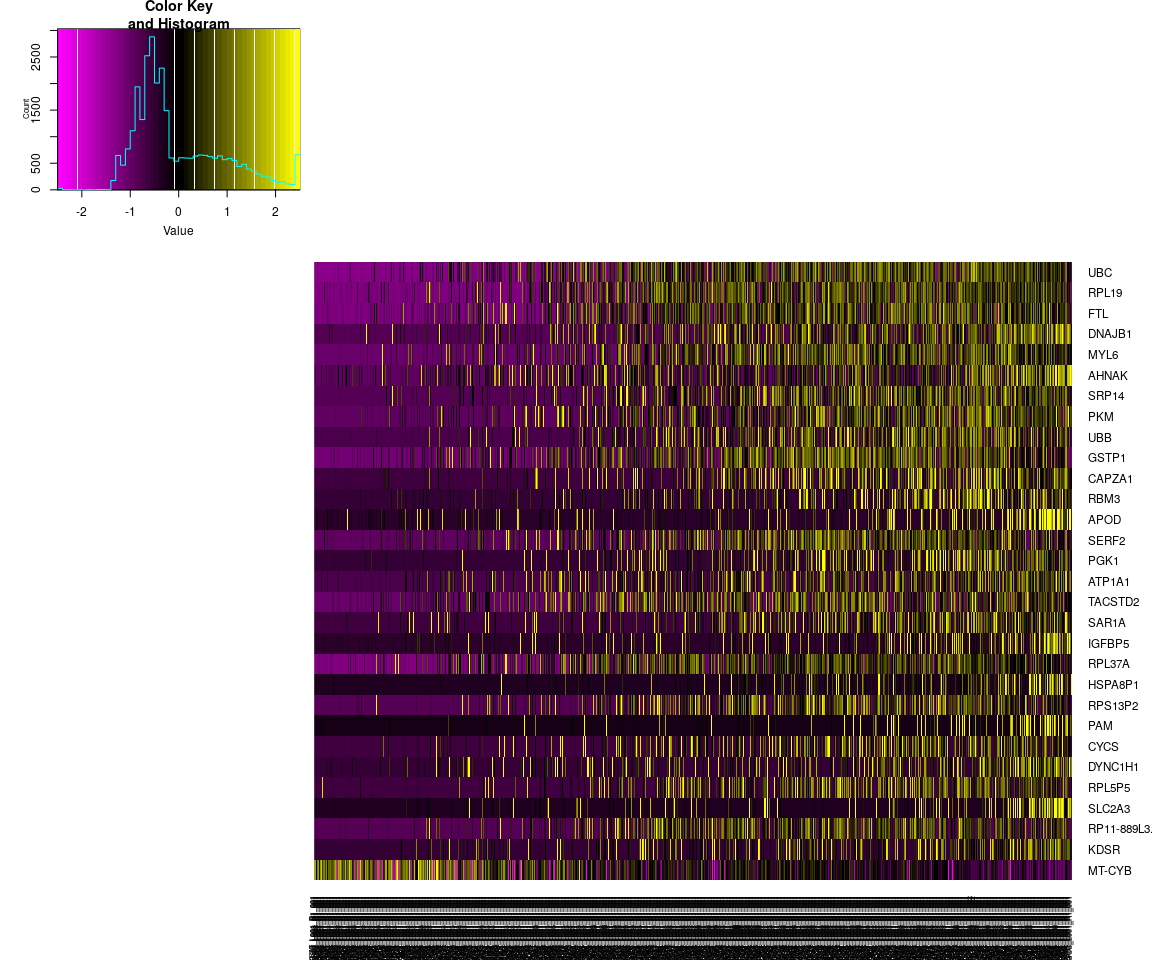
DimPlot(all.pbmc, reduction.use = "tsne", pt.size = 4) # grour by sample



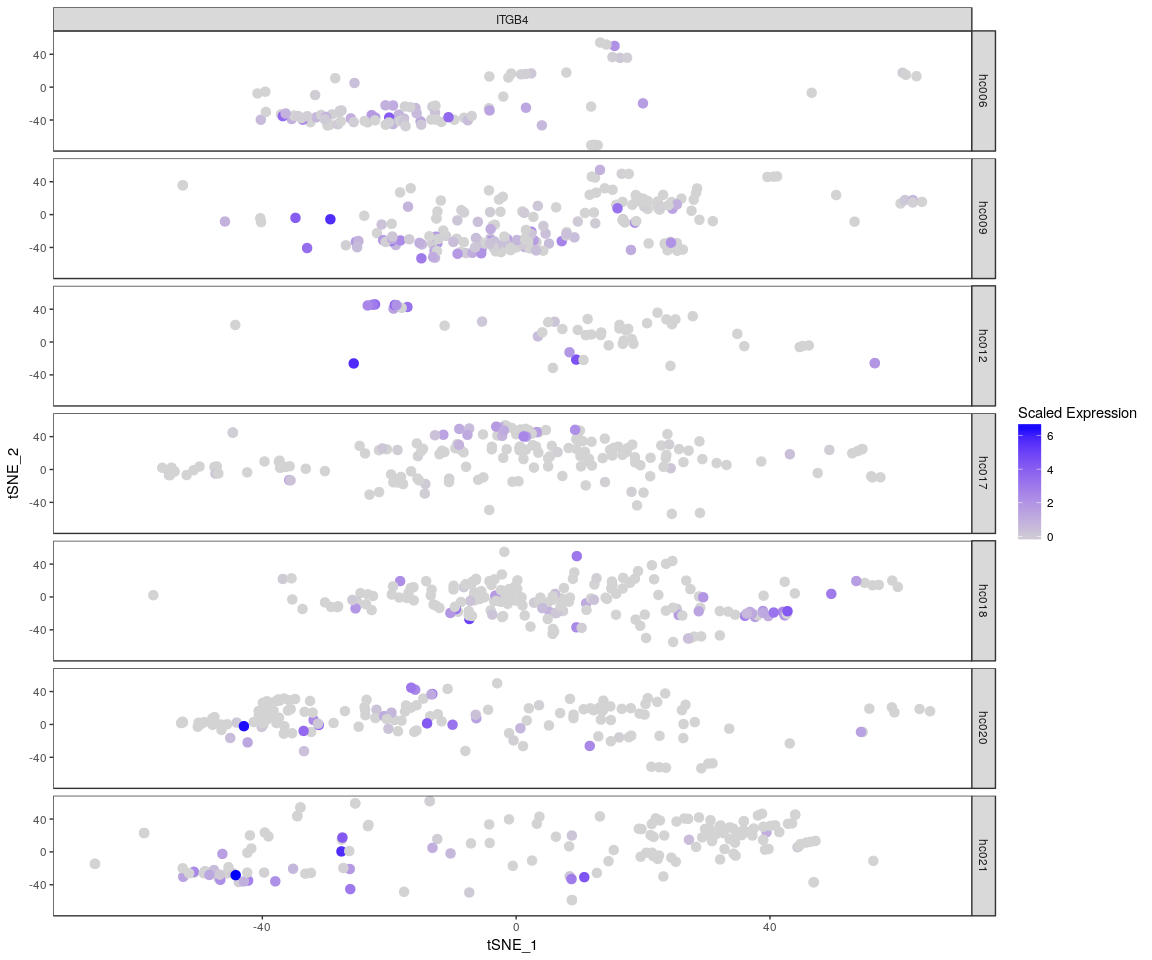
DimPlot(all.pbmc, reduction.use = "pca", pt.size = 4) # grour by sample



DimHeatmap(all.pbmc, reduction.type = "pca", check.plot = FALSE)



FeatureHeatmap(all.pbmc, features.plot = "ITGB4", pt.size = 3, plot.horiz = TRUE,   
 cols.use = c("lightgrey", "blue"))

 The Faetureplot of ITGB4, ABCB5, KRT19, ACTB, KRT12, KRT5, GAPDH, KRT3, PAX6, WNT7A, KRT14, TP63, KRT10based on **PCA** shows that,they only has high expression level in few samples,and expresss lowly in most sample.It means that may be these important genes express differently across sample.The plot also tell us the gene **KRT5,GAPDH,PAXX6,KRT14** have more higher expression level than the other important genes.It is consistent with the result of violin plot. About the heatmap,we only show the gene **ITGB4** And the FeatureHeatmap and Heamap also comfirm this phenomeno.We try the other four variable genes,which has the similar result as gene *ITGB4* But the *tSNE* and \* PCA \* plot show that, the sample can not be split apparently.The result may be is not good based on the PCA and tSNE method.

## Differential expression

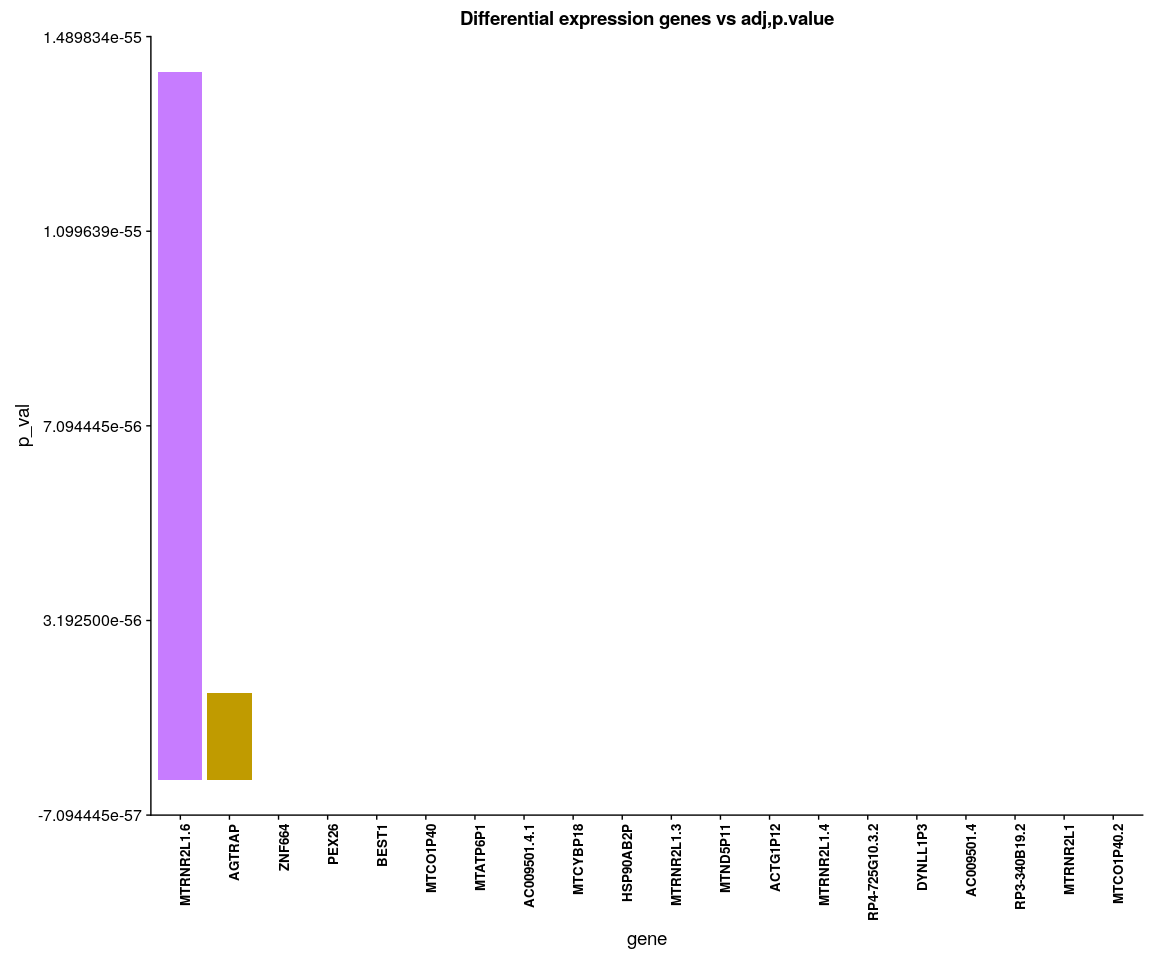
Next,we will have analysis on gene differential expression.Find maker genes across sample.We use the method: \*\*wilcox test\*\*

# Finds markers (differentially expressed genes) for each of the identity  
# classes in a dataset  
all\_markers <- FindAllMarkers(all.pbmc, test.use = "bimod", print.bar = FALSE)  
head(all\_markers)

## p\_val avg\_logFC pct.1 pct.2 p\_val\_adj cluster  
## MTRNR2L1 5.329110e-132 -2.2519643 0.783 0.510 1.085007e-127 hc006  
## AC009501.4 5.811696e-117 -1.7620482 0.717 0.270 1.183261e-112 hc006  
## MTATP6P1 9.133929e-73 -2.1846558 1.000 0.905 1.859668e-68 hc006  
## BEST1 6.321317e-69 0.8036177 0.859 0.050 1.287020e-64 hc006  
## PEX26 7.368432e-66 -0.5092965 0.522 0.083 1.500213e-61 hc006  
## ZNF664 2.940084e-65 0.3984055 0.935 0.120 5.986010e-61 hc006  
## gene  
## MTRNR2L1 MTRNR2L1  
## AC009501.4 AC009501.4  
## MTATP6P1 MTATP6P1  
## BEST1 BEST1  
## PEX26 PEX26  
## ZNF664 ZNF664

We check whether the important genes are still in the marker genes we found from the DESeq analysis. the genes:ITGB4, KRT19, ACTB, KRT5, GAPDH, KRT3, PAX6, KRT14, TP63, KRT10 are still in the marker genes.

### Bar plot of gene’s p.val for the most significant expressed genes

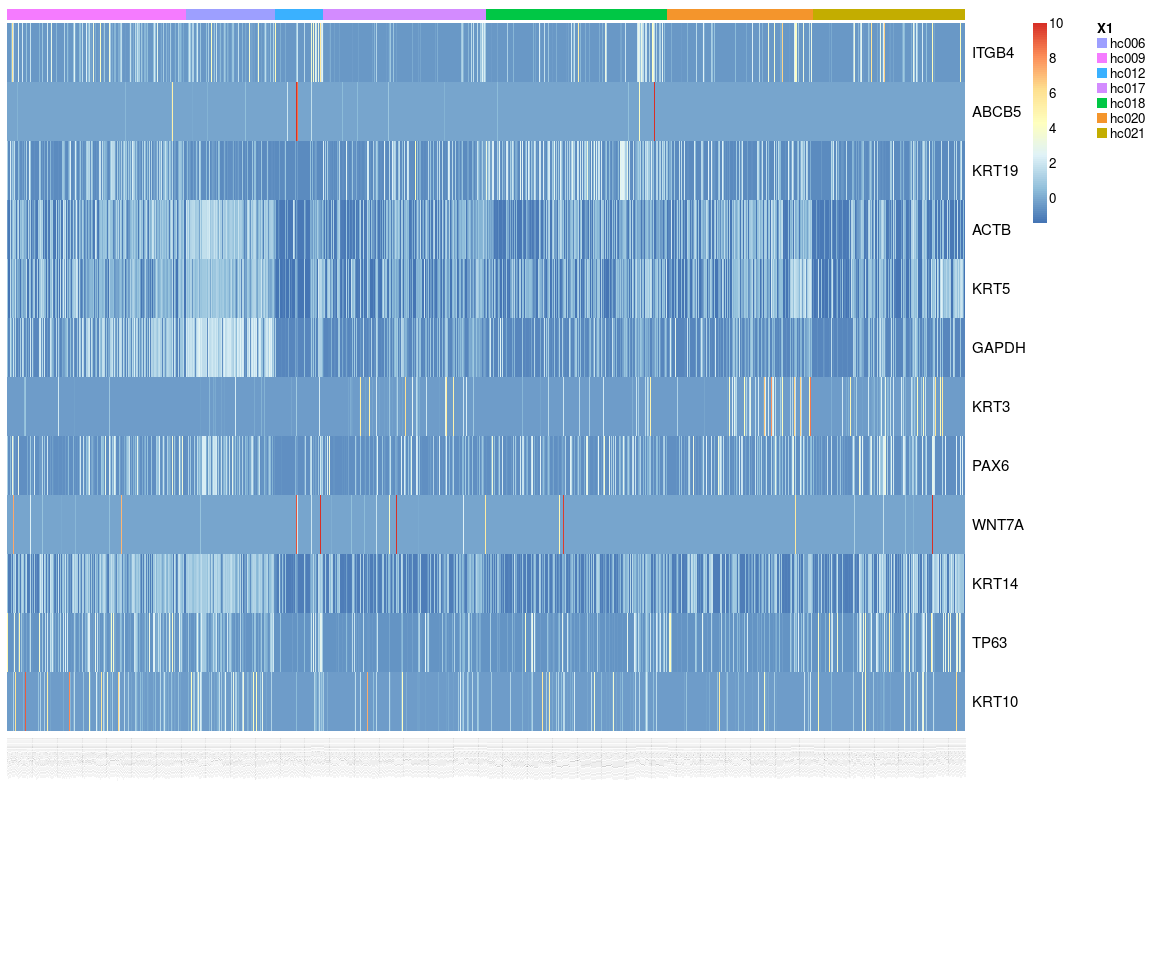


### Heatmap for important genes group by sample batch

human.heatmap <- Heatmap\_fun(genes = important.genes[important.genes %in% rownames(human.all.DESeq@raw.data)],   
 tpm.data = all.pbmc@scale.data, condition = unique(as.character(all.pbmc@ident)),   
 all.condition = as.character(all.pbmc@ident))

## There ara 7 conditions  
## Whether creat data accurate 0

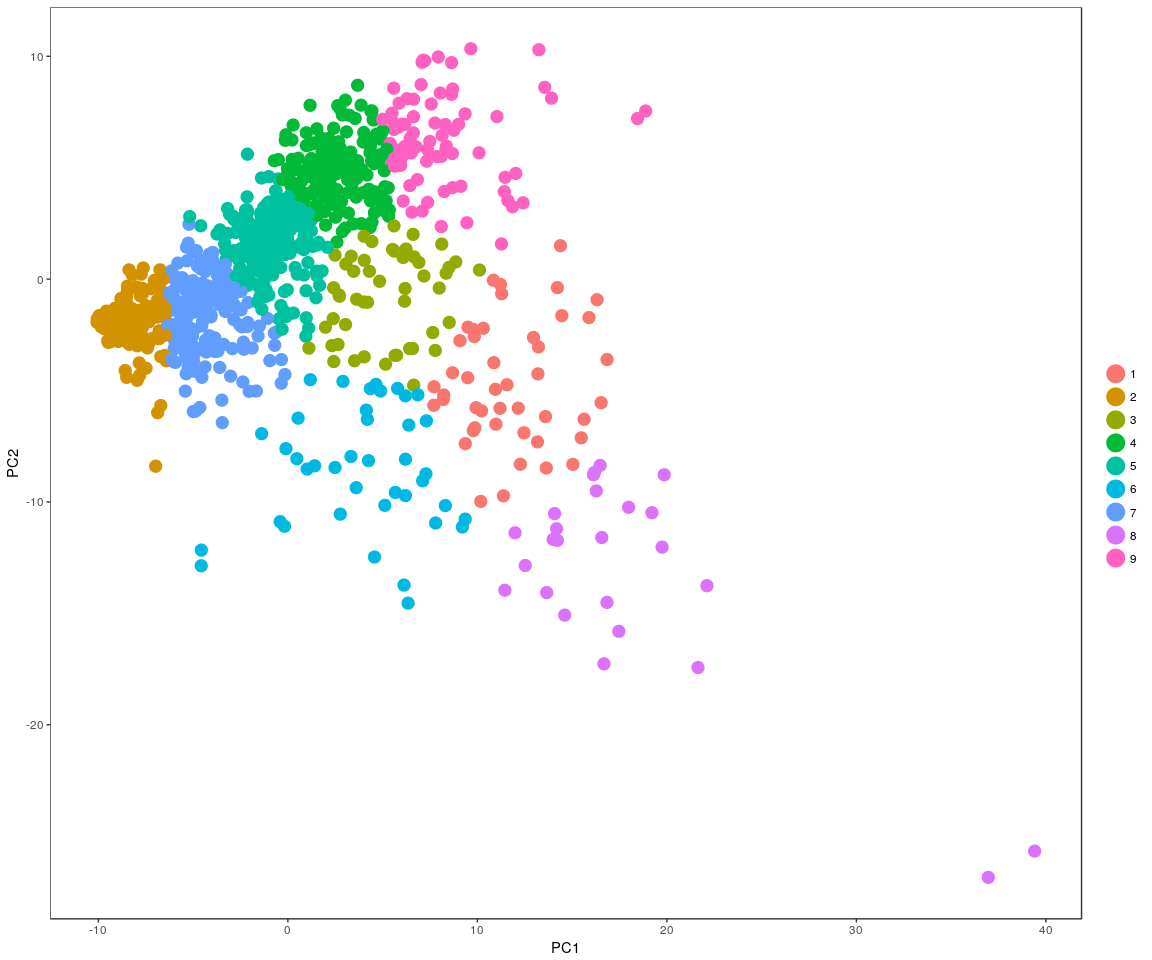
NMF::aheatmap(human.heatmap[[2]], Rowv = NA, Colv = NA, annCol = human.heatmap[[1]],   
 scale = "none")



We have find all marker genes across sample,there are 4780 significant genes(adjust p-value <0.05) in all marker genes.

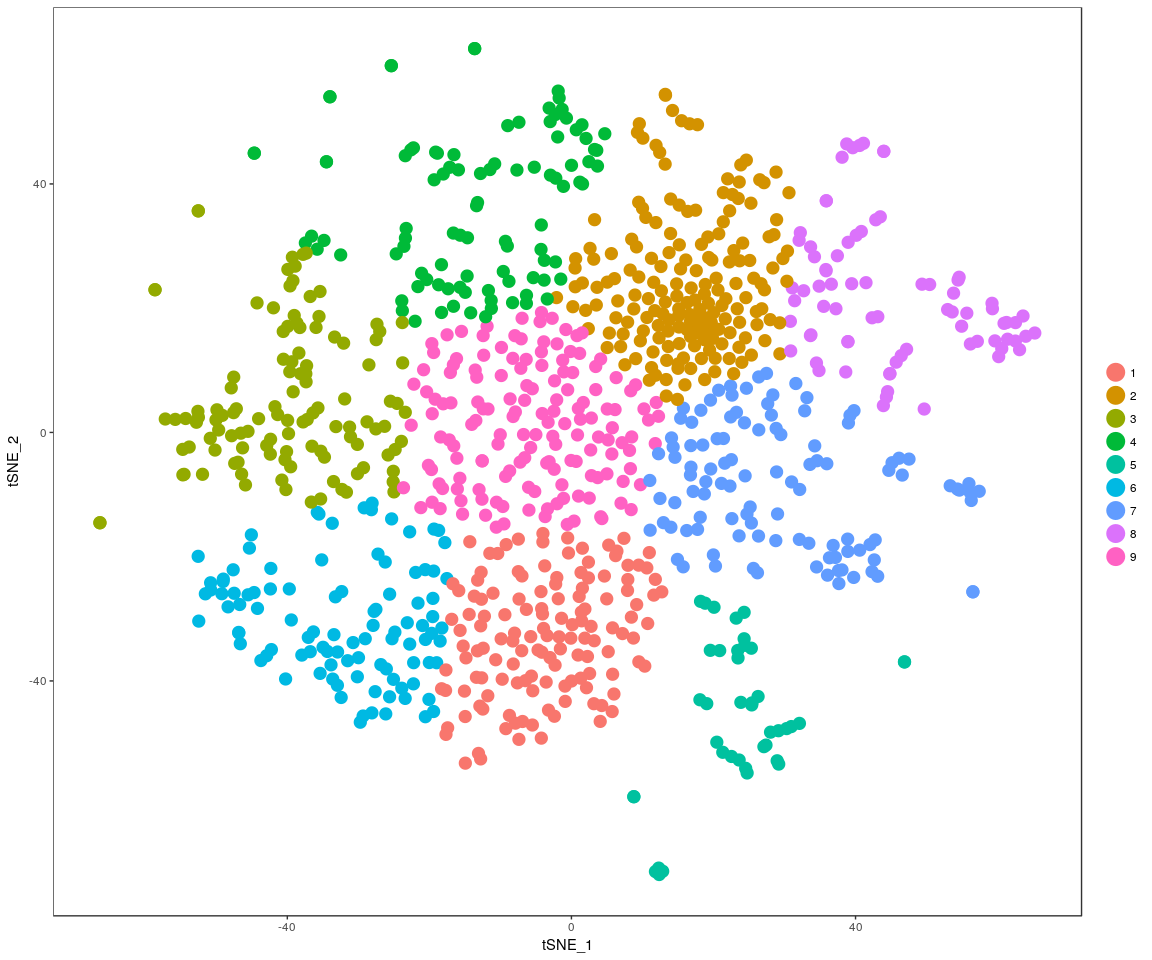
### Next,Spectral k-means clustering on single cells based on PCA

all.pbmc <- KClustDimension(all.pbmc, reduction.use = "pca", k.use = 9)  
clusters.pca <- all.pbmc@meta.data$kdimension.ident  
DimPlot(all.pbmc, pt.size = 4, group.by = "kdimension.ident")



### Spectral k-means clustering on single cells based on tSNE

all.pbmc <- KClustDimension(all.pbmc, reduction.use = "tsne", k.use = 9)  
clusters.tsne <- all.pbmc@meta.data$kdimension.ident  
DimPlot(all.pbmc, pt.size = 4, group.by = "kdimension.ident", reduction.use = "tsne")



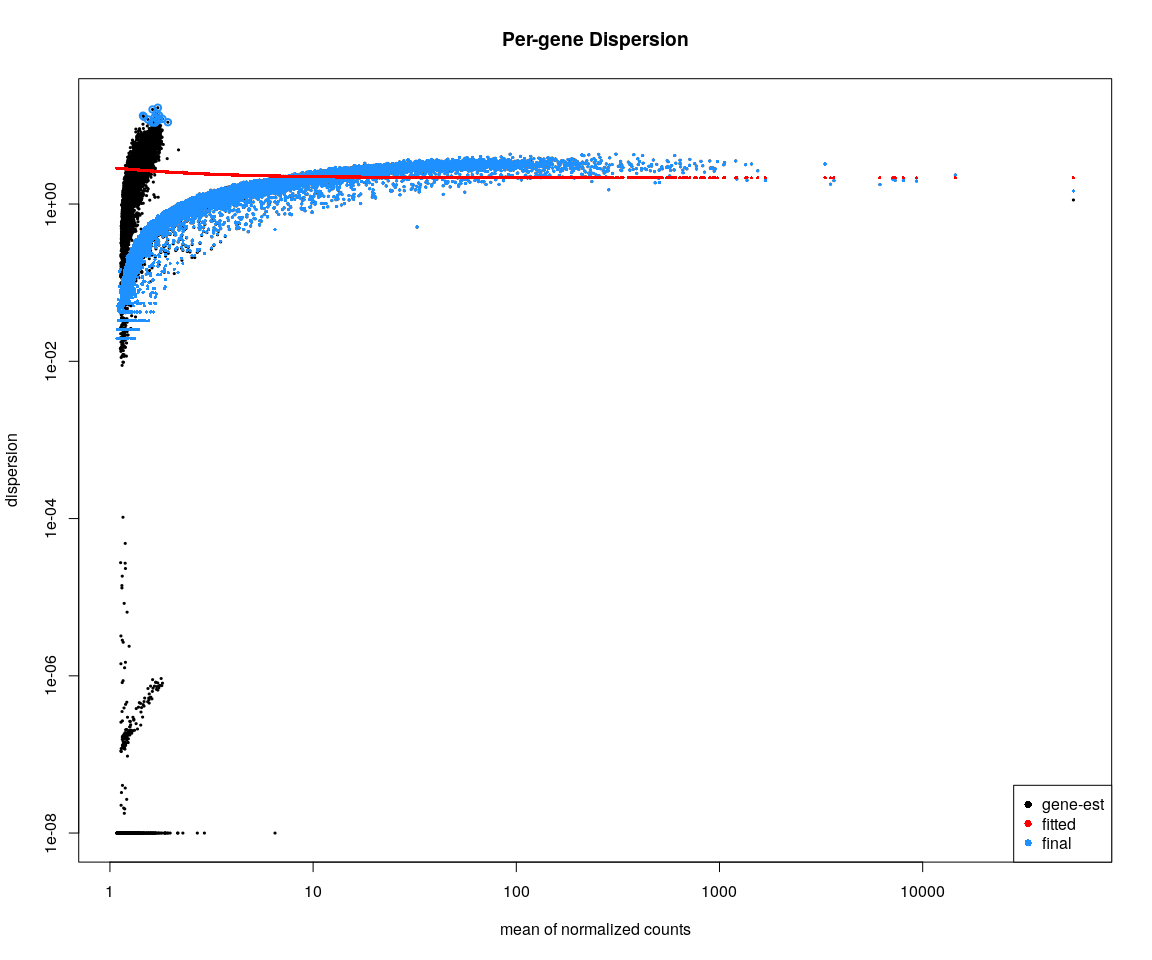
## Step 3: another approach to handle gene count data and use the DESeq testto detect genes Differential expression.

When use the DESeq,it must require the gene count matrix satisify that: **every gene contains at least one zero, cannot compute log geometric means**. So have to take another method to handle data,but I do not know whether it is reasonable.Just try!!!

## DESeq test

### it will take a long time

# sample group  
condition.1 <- unlist(lapply(all.pbmc@cell.names, function(x) return(str\_split(x,   
 "\_")[[1]][1])))  
# condition.2<-unlist(lapply(colnames(human.only.pro),function(x)return(str\_split(x,'\_')[[1]][2])))  
load("Human.sample.RData")  
plotDispEsts(xdds, main = "Per-gene Dispersion")



#### Show the DESeq result between group hc006 and hc009: the differently expressed genes

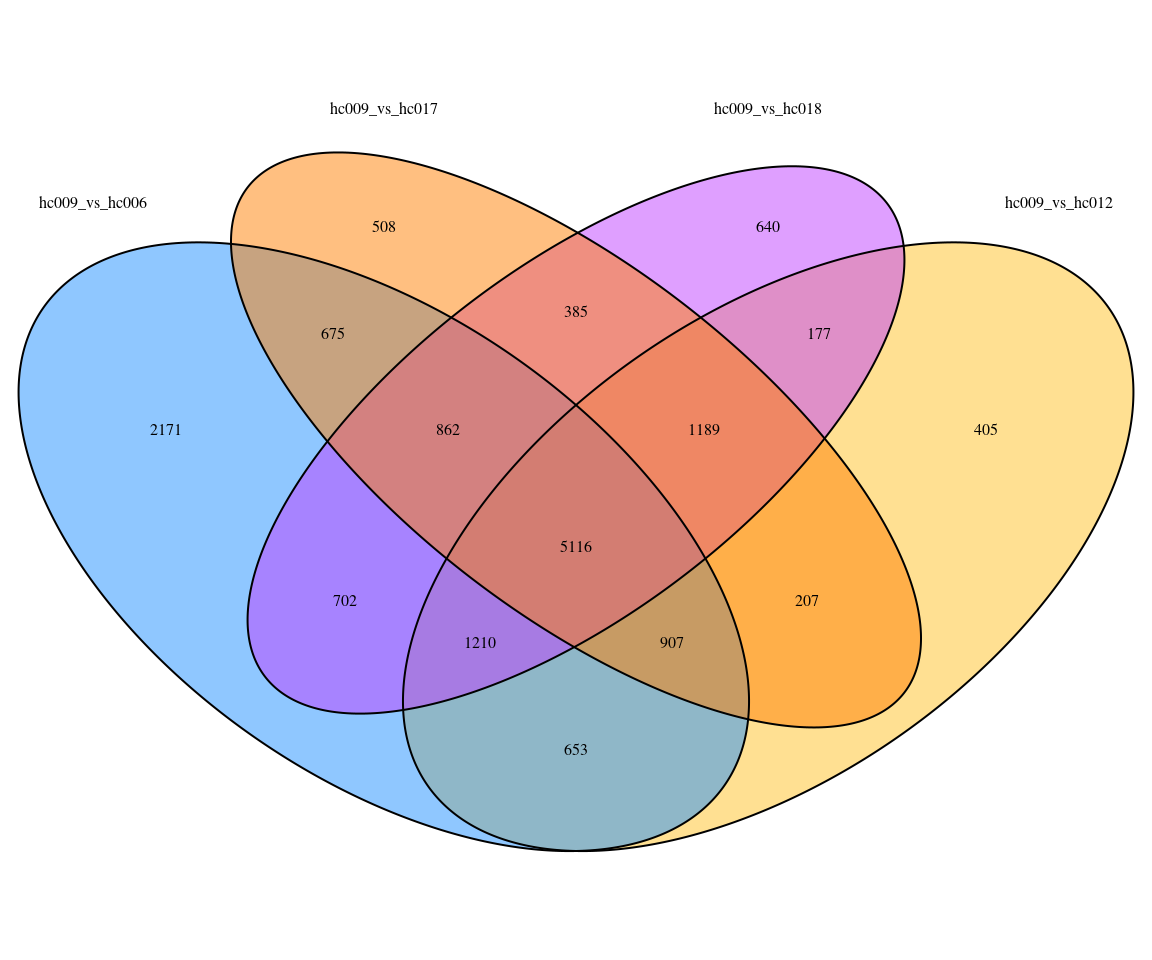
res <- results(xdds, contrast = c("condition.1", "hc009", "hc006"))  
res[which(res$padj < 0.05), ]

the table are the result which gene’s padj value<0.05.Include the variable:**baseMean,log2FoldChange lfcSE,stat,pvalue,padj**.

## Do the DESeq test across all cells with sample group.And get all the significant genes between two groups(p.value < 0.05)

## Venn diagram

library(VennDiagram)  
grid.draw(venn.diagram(x[1:4], filename = NULL, fill = c("dodgerblue", "goldenrod1",   
 "darkorange1", "darkorchid1")))



Venn diagram show that the common differentially expressing genes across gorups.The number represents the the number of common genes.Here just show the randomly selected groups