Monkey summary analysis

yejg

2018/1/26

### 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)

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

### Read data

### Data QA

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

##   
## 10um 20um 6um   
## 124 344 126

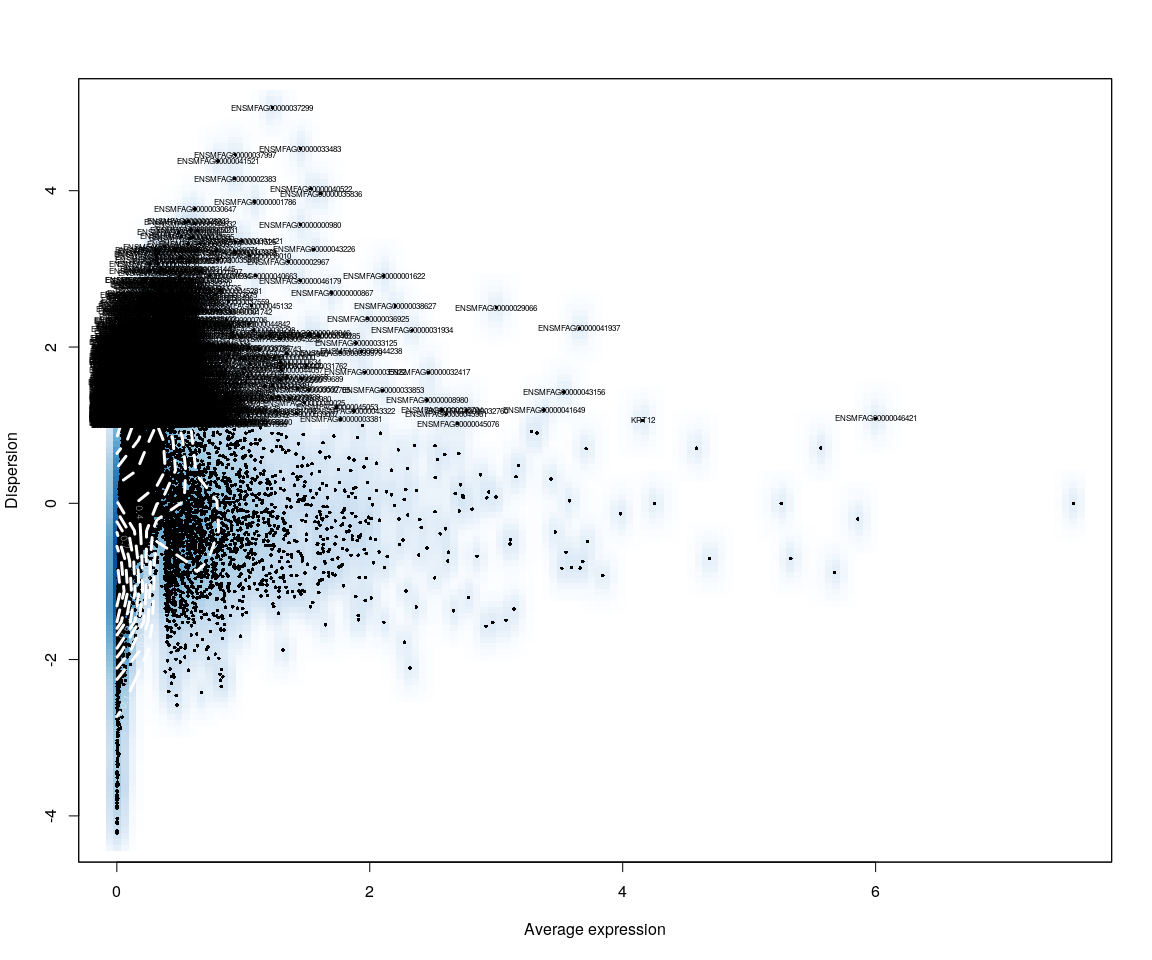
table(unlist(lapply(colnames(monkey.only.pro), function(x) return(str\_split(x,   
 "\_")[[1]][1]))))

##   
## mkc001 mkc003 mkc004 mkc005   
## 156 155 169 114

### Create Seurat object and not caculate DESeq,but set **min.cells=10** and **min.genes=2**

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

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



all.sample.group <- unlist(lapply(monkey.all.pbmc@cell.names, function(x) return(str\_split(x,   
 "\_")[[1]][1])))  
all.sample.size <- unlist(lapply(monkey.all.pbmc@cell.names, function(x) return(str\_split(x,   
 "\_")[[1]][2])))

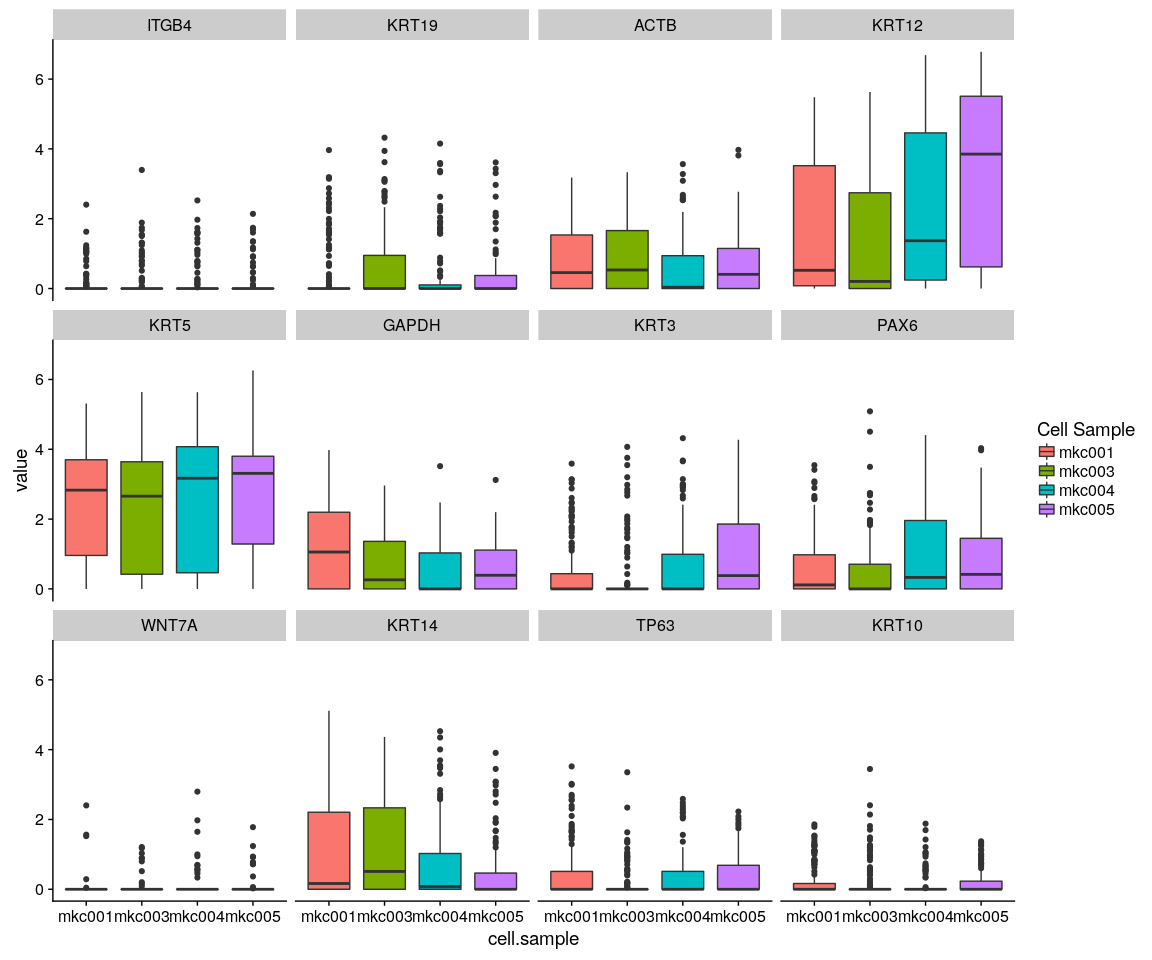
## Figure Explore

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

monkey.imp.lognorm <- data.frame(FetchData(monkey.all.pbmc, vars.all = important.genes[important.genes %in%   
 rownames(monkey.all.pbmc@raw.data)]))  
monkey.imp.lognorm$cell.size <- unlist(lapply(rownames(monkey.imp.lognorm),   
 function(x) return(str\_split(x, "\_")[[1]][2])))  
monkey.imp.lognorm$cell.sample <- unlist(lapply(rownames(monkey.imp.lognorm),   
 function(x) return(str\_split(x, "\_")[[1]][1])))  
monkey.imp.lognorm.melt <- melt(monkey.imp.lognorm)

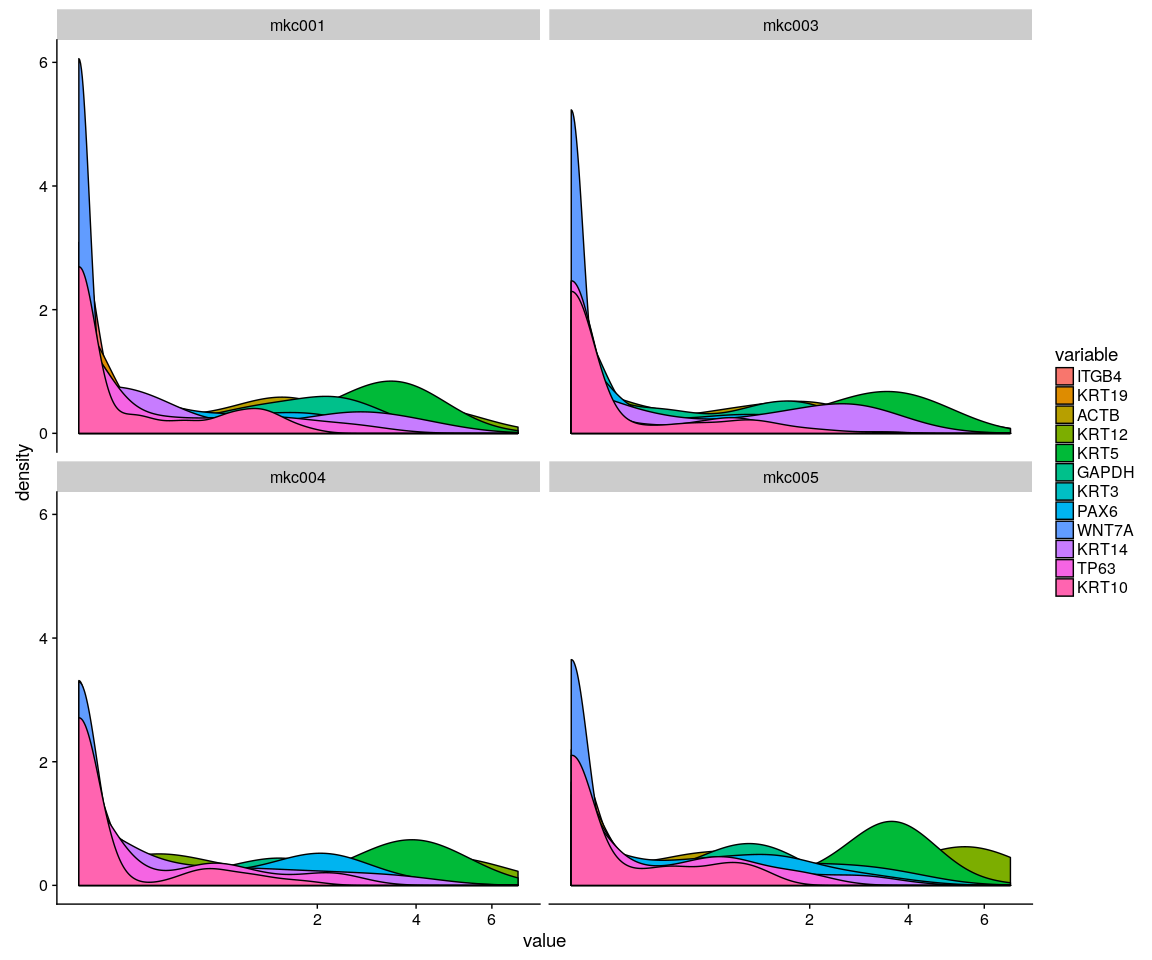
#### Boxplot

p <- ggplot(data = monkey.imp.lognorm.melt, aes(y = value, x = cell.sample,   
 fill = cell.sample))  
p + geom\_boxplot() + guides(fill = guide\_legend(title = "Cell Sample")) + facet\_wrap(~variable)

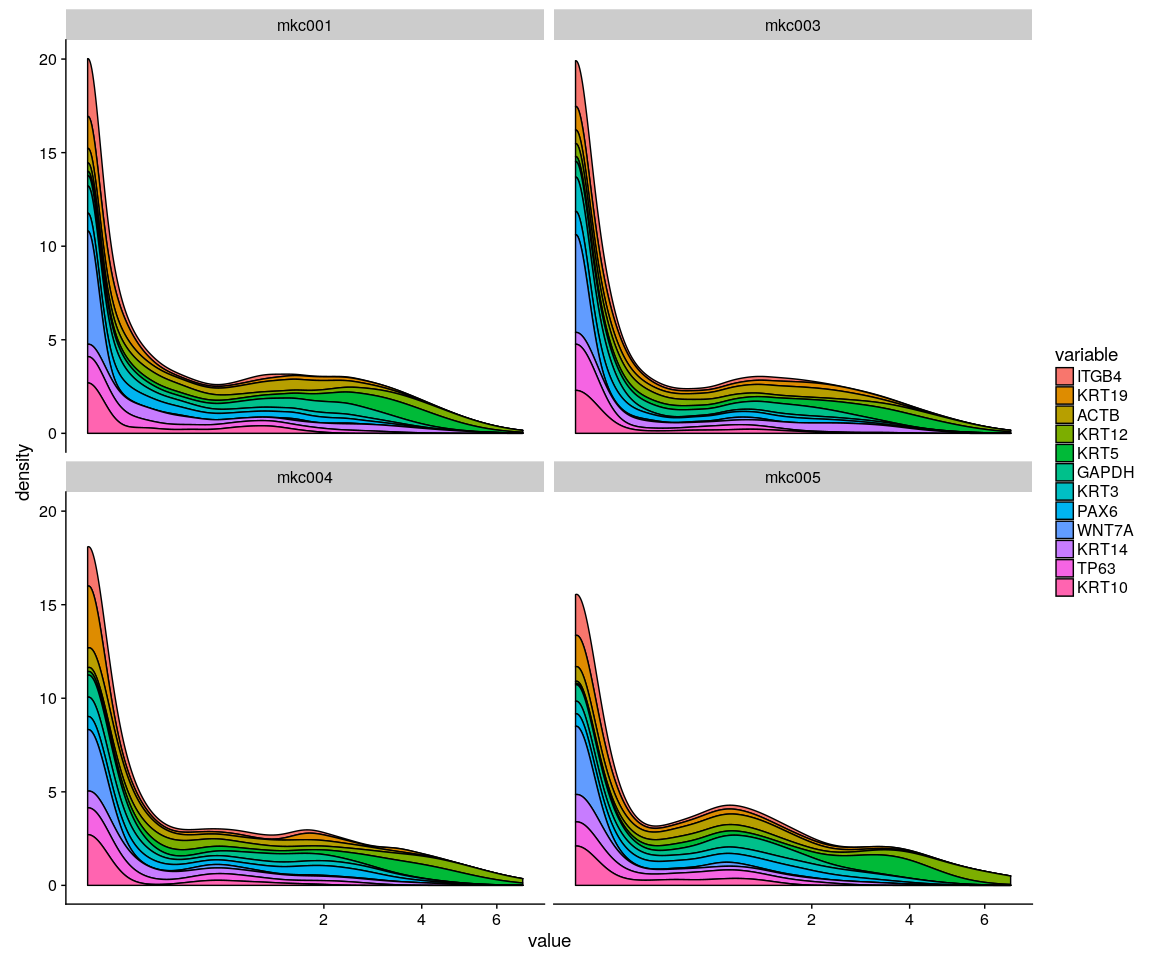


#### Density,histogram

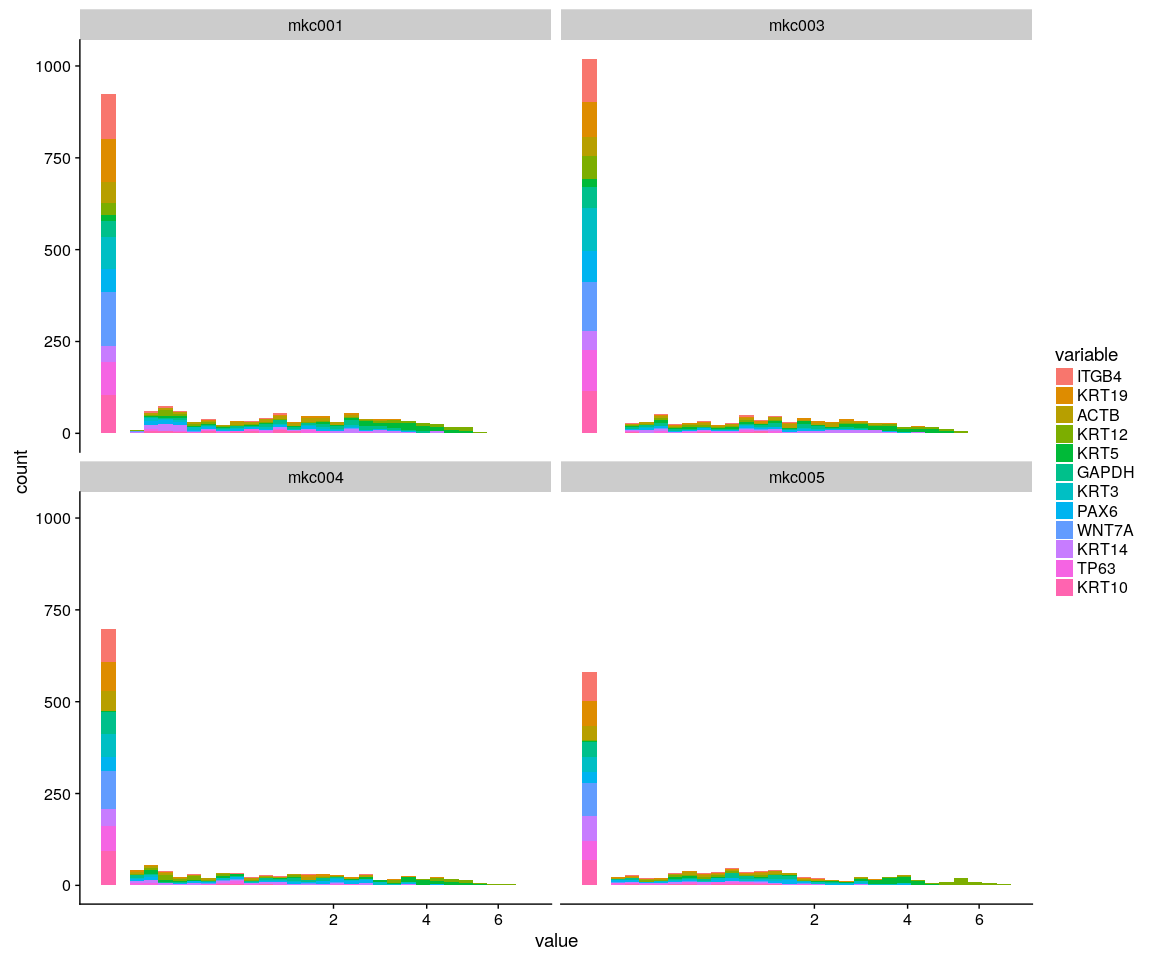
ggplot(data = monkey.imp.lognorm.melt, aes(x = value, fill = variable)) + geom\_density(kernel = "gaussian") +   
 scale\_x\_sqrt() + facet\_wrap(~cell.sample)



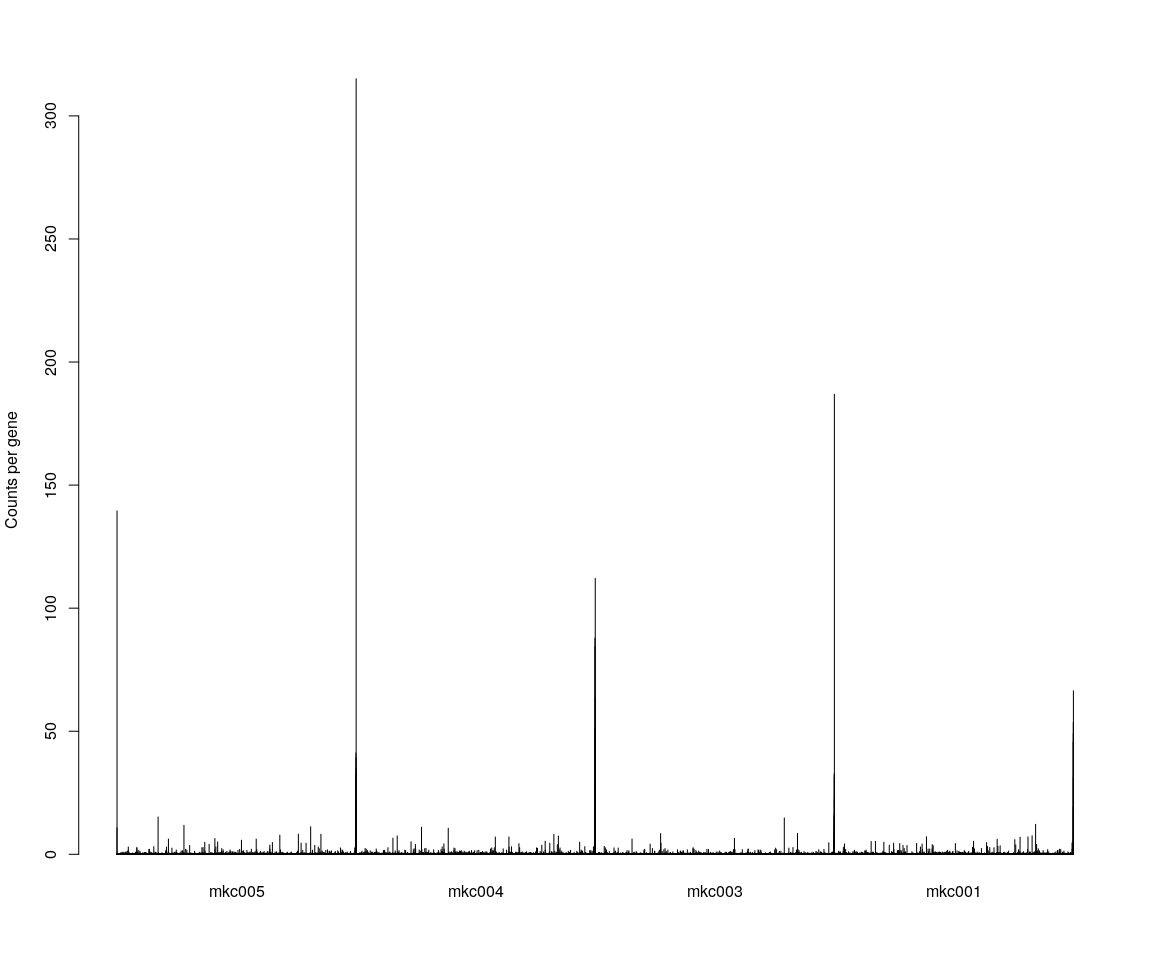
ggplot(data = monkey.imp.lognorm.melt, aes(x = value, fill = variable)) + geom\_density(kernel = "gaussian",   
 position = "stack") + scale\_x\_sqrt() + facet\_wrap(~cell.sample)



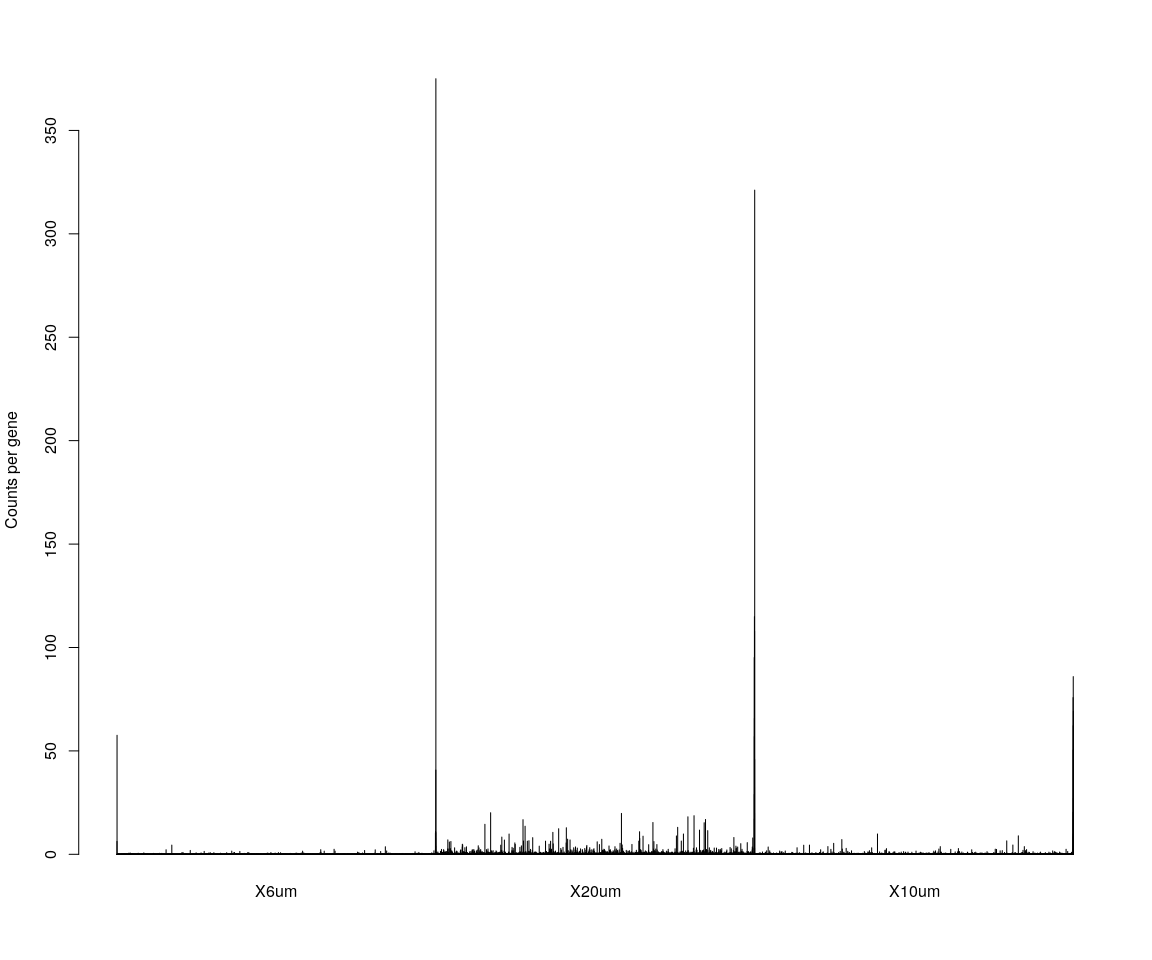
ggplot(data = monkey.imp.lognorm.melt, aes(x = value, fill = variable)) + geom\_histogram() +   
 scale\_x\_sqrt() + facet\_wrap(~cell.sample)



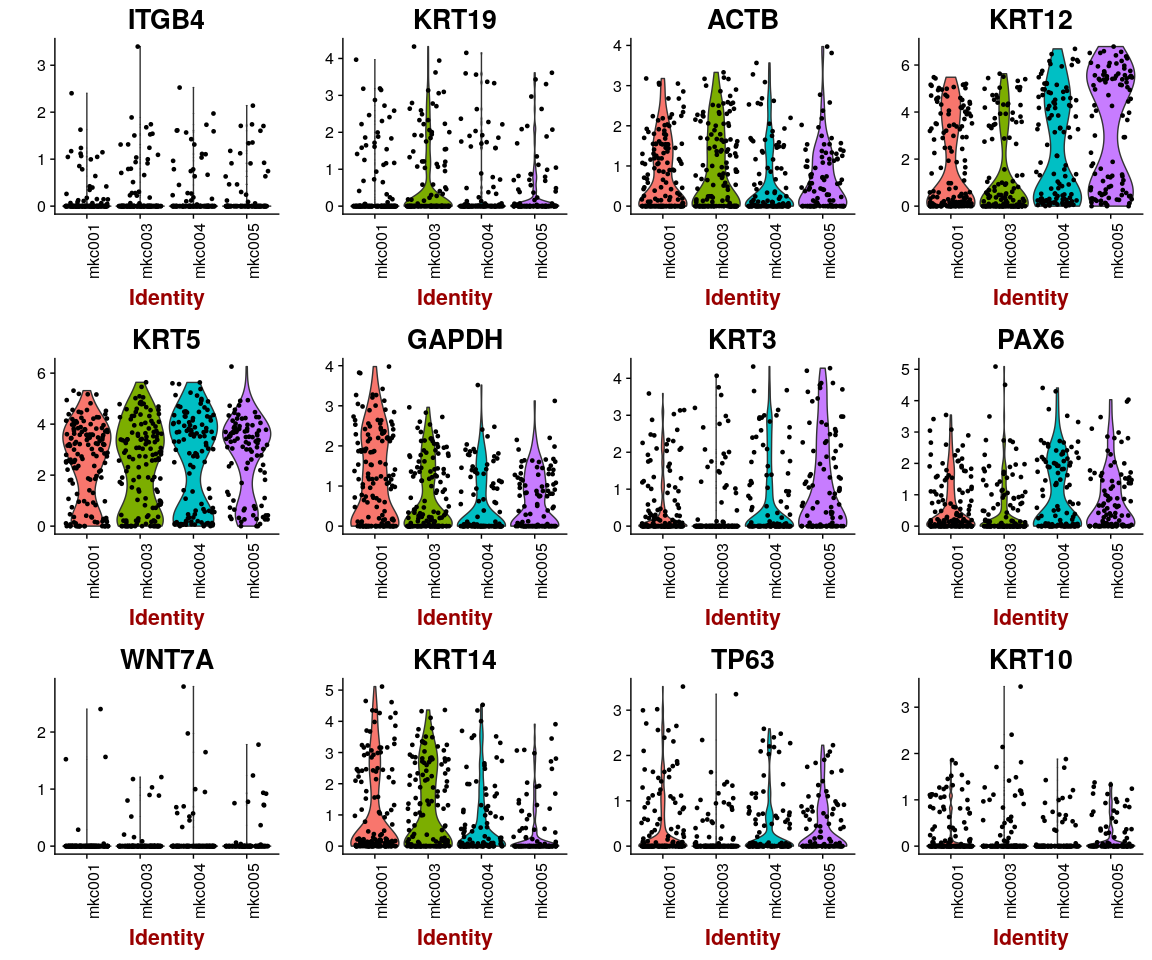
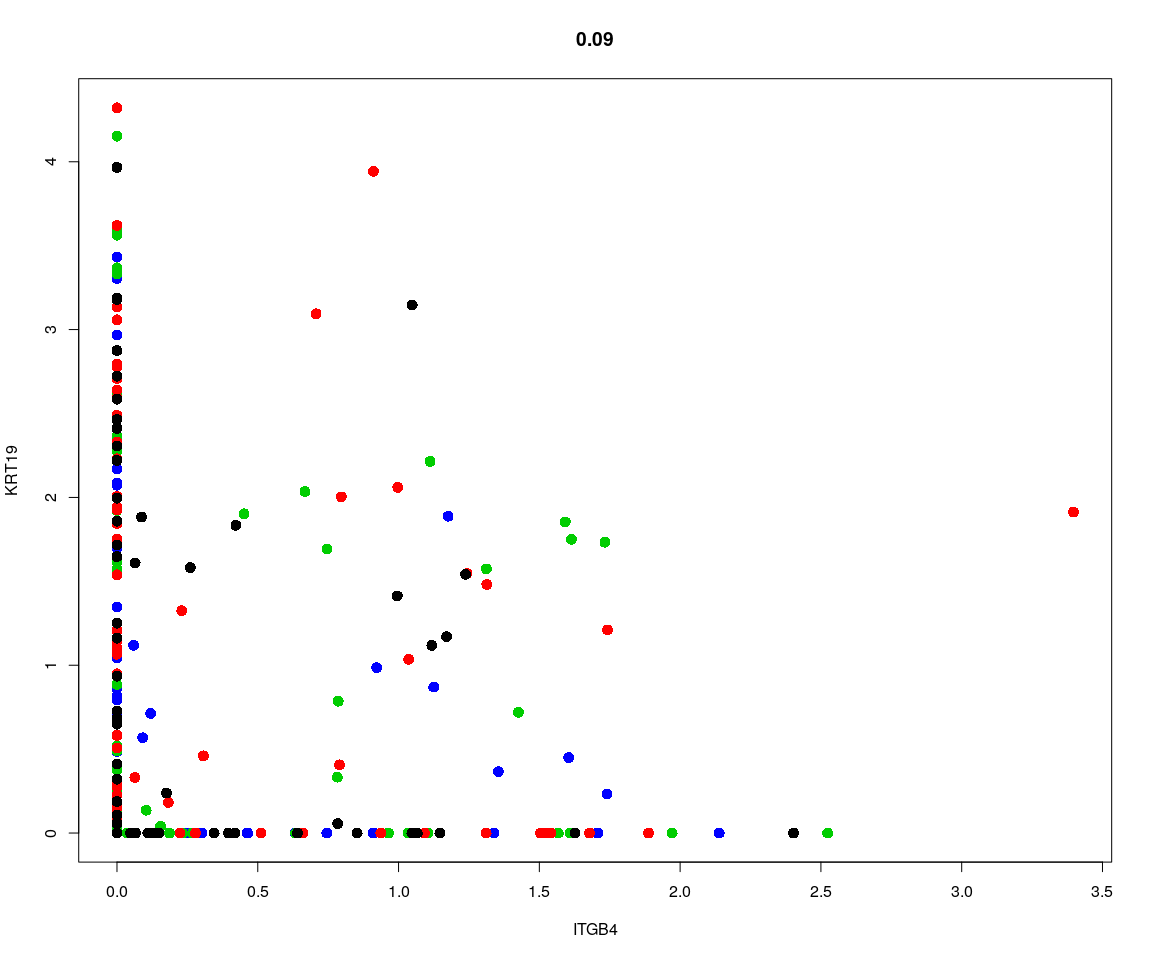
Group\_Bar(monkey.all.pbmc@raw.data, group = all.sample.group)



Group\_Bar(monkey.all.pbmc@raw.data, group = all.sample.size)



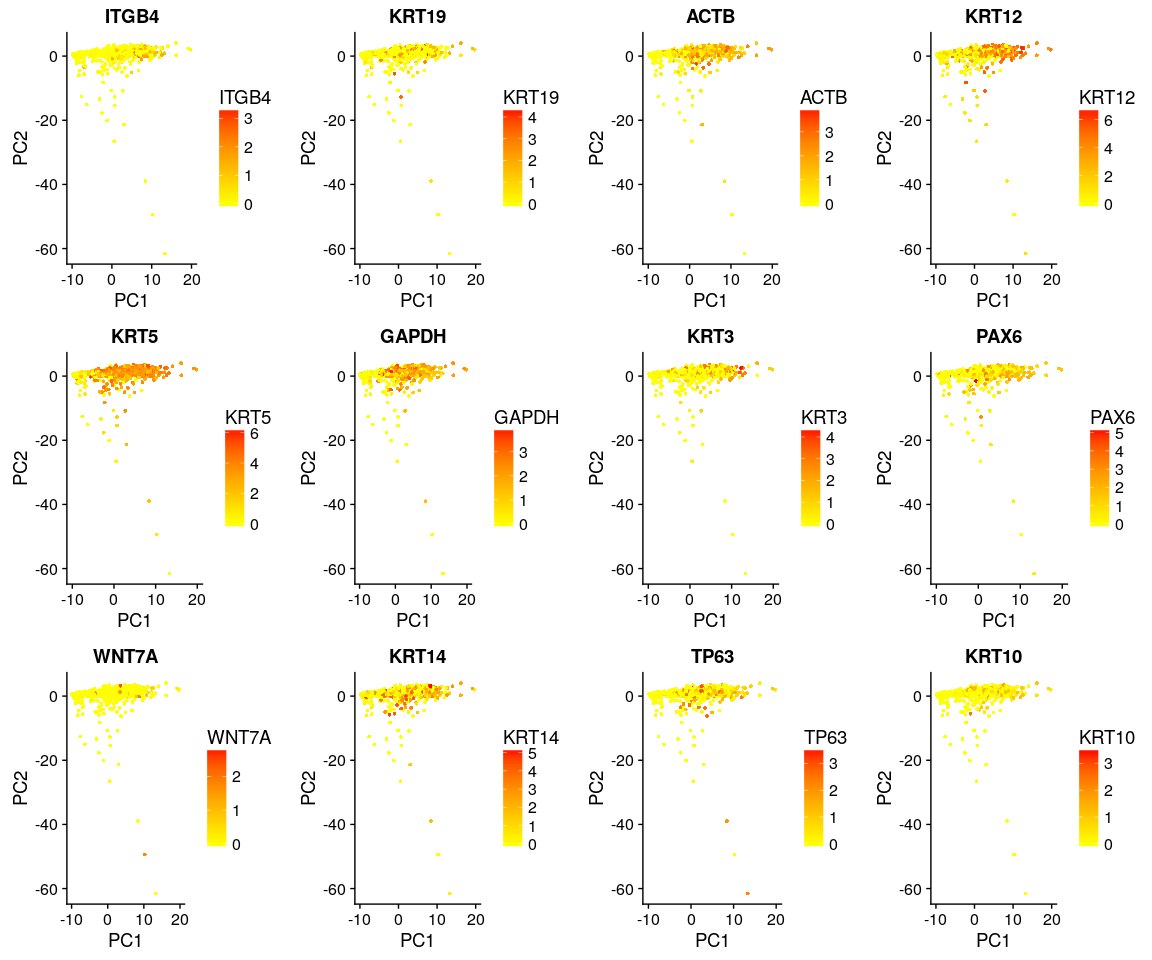
# We are interested in the gene ITGB4  
GenePlot(monkey.all.pbmc, gene1 = "ITGB4", gene2 = important.genes[3])  
  
VlnPlot(monkey.all.pbmc, features.plot = important.genes[important.genes %in%   
 rownames(monkey.all.pbmc@raw.data)], y.lab.rot = 90) # Violinn plot of gene ITGB in all sample

 ## Dimensionality reduction ### **PCA** and **tSNE** Here,do the dimensionality reduction using the PCA, tSNE method

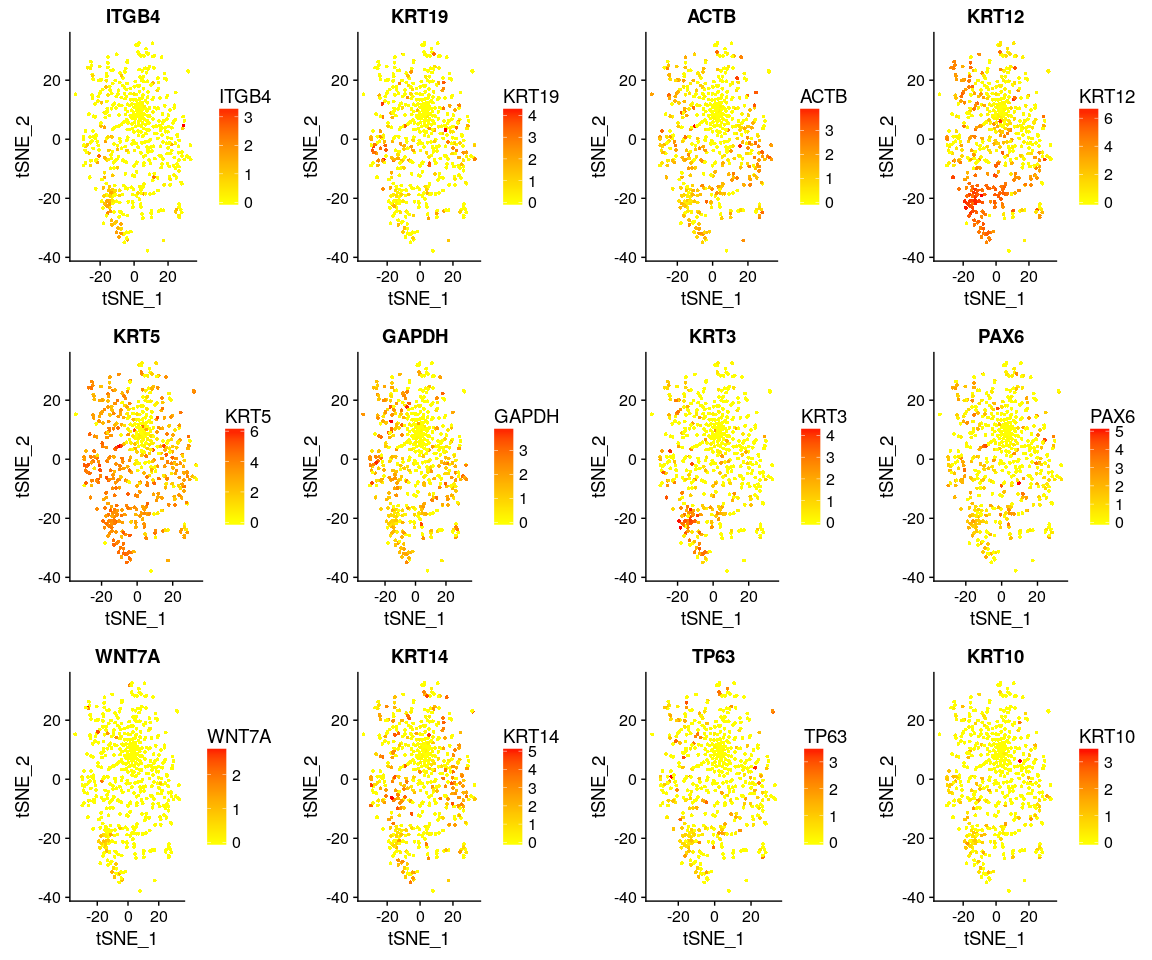
all.pbmc <- PCA.TSNE(object = monkey.all.pbmc, pcs.compute = FALSE, num.pcs = 28)

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

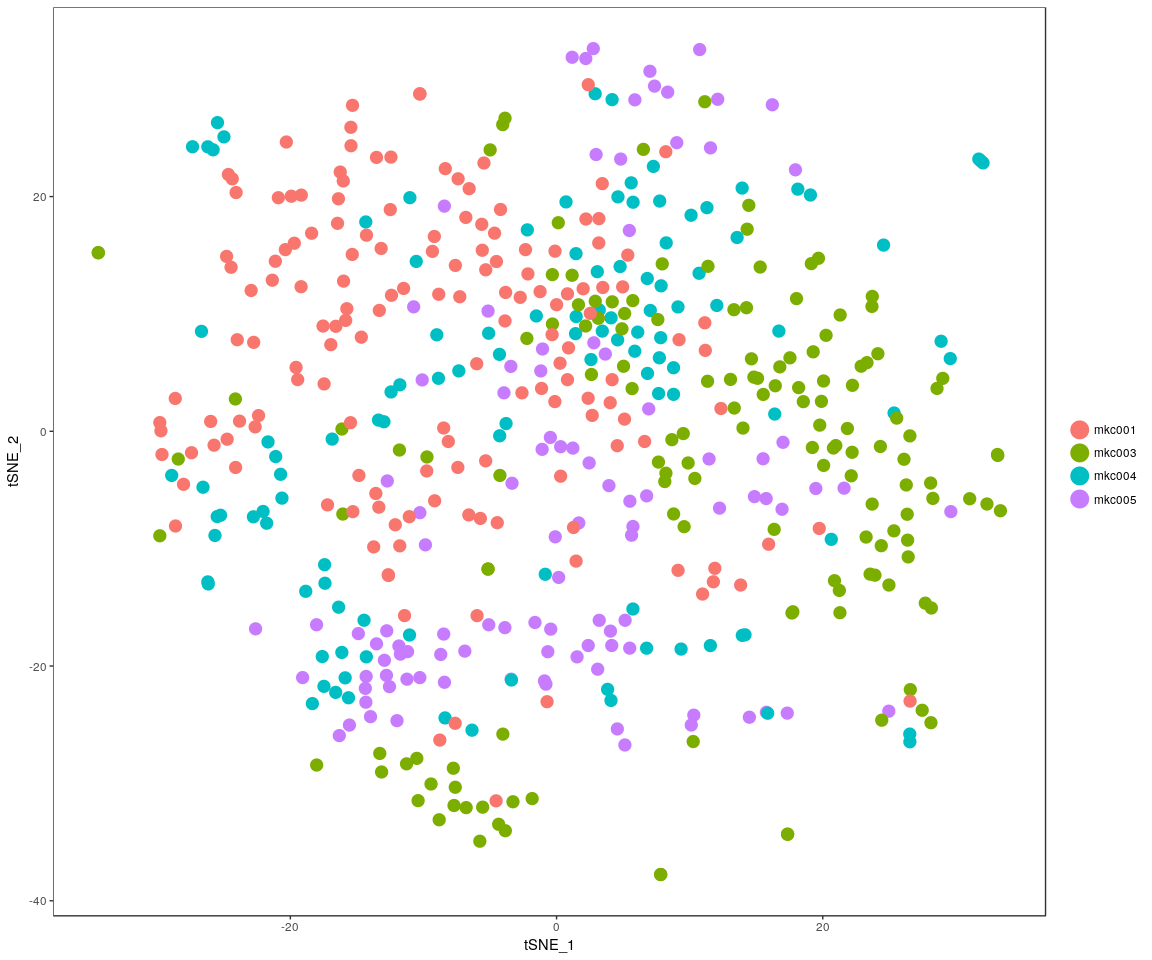
# FeaturePlot(object = all.pbmc,features.plot ='ITGB4',pt.size = 4,no.legend  
# = FALSE) # ITGB4 gene in part dataset  
FeaturePlot(object = all.pbmc, features.plot = important.genes[important.genes %in%   
 rownames(all.pbmc@raw.data)], pt.size = 1, no.legend = FALSE, reduction.use = "pca")



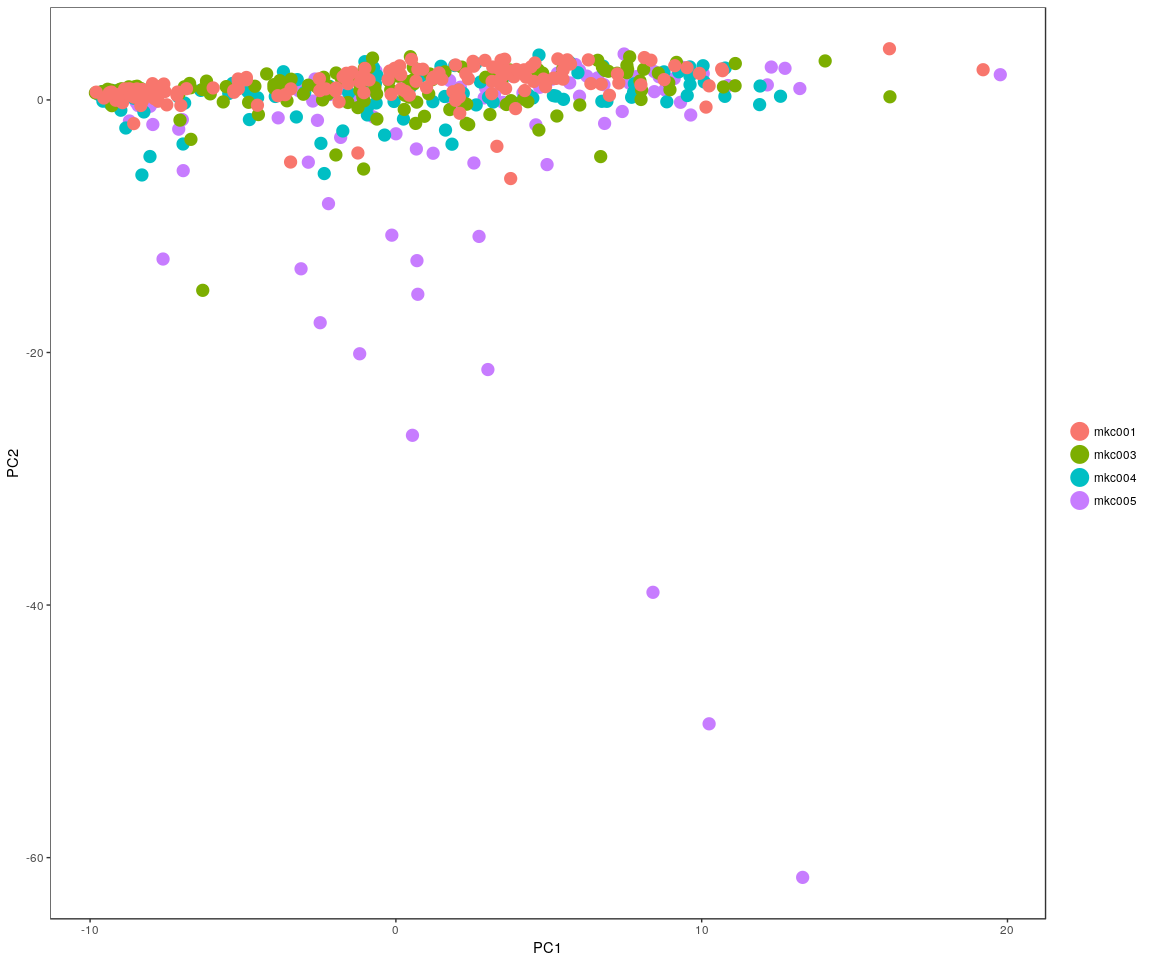
FeaturePlot(object = all.pbmc, features.plot = important.genes[important.genes %in%   
 rownames(all.pbmc@raw.data)], pt.size = 1, no.legend = FALSE, reduction.use = "tsne")



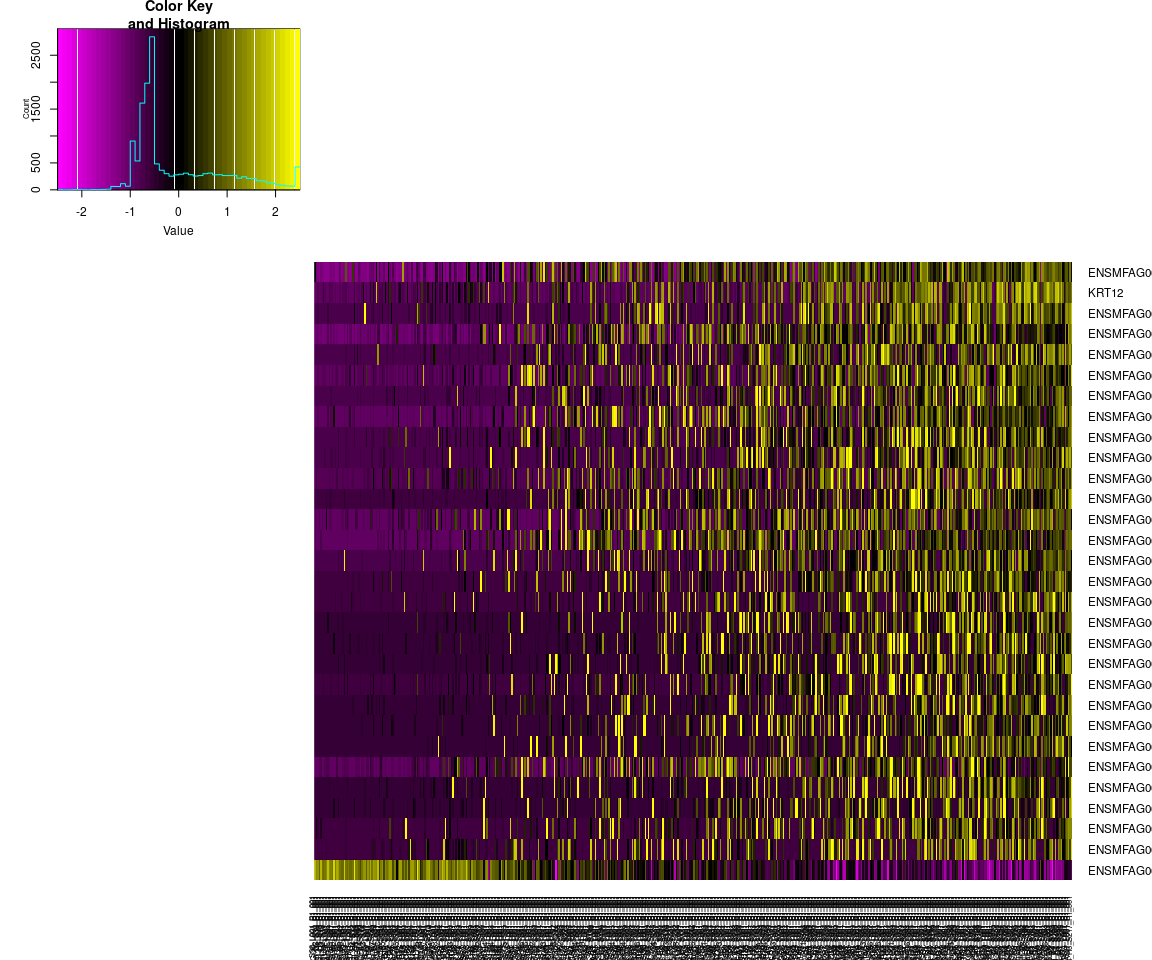
DimPlot(all.pbmc, reduction.use = "tsne", pt.size = 4) # grour by sample



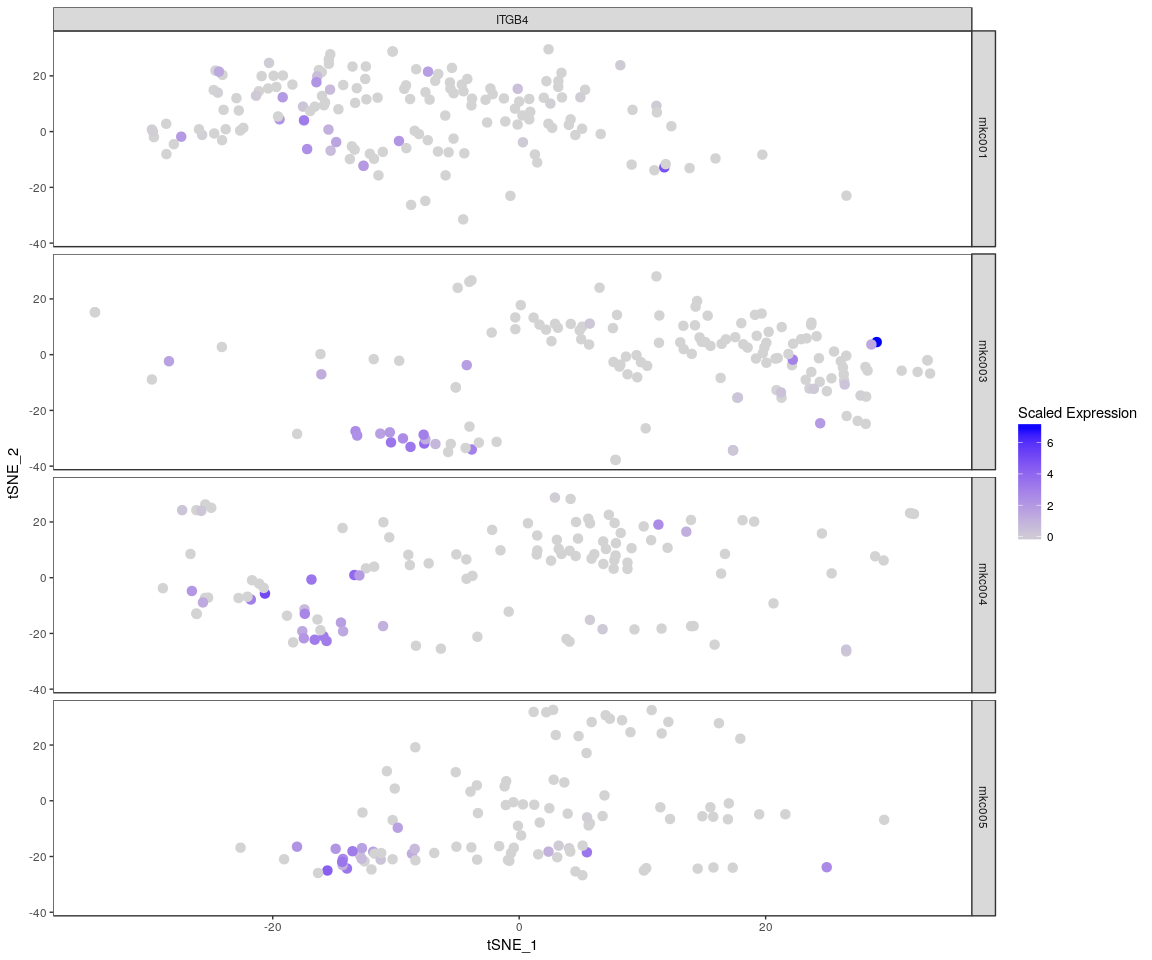
DimPlot(all.pbmc, reduction.use = "pca", pt.size = 4) # grour by sample



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"))

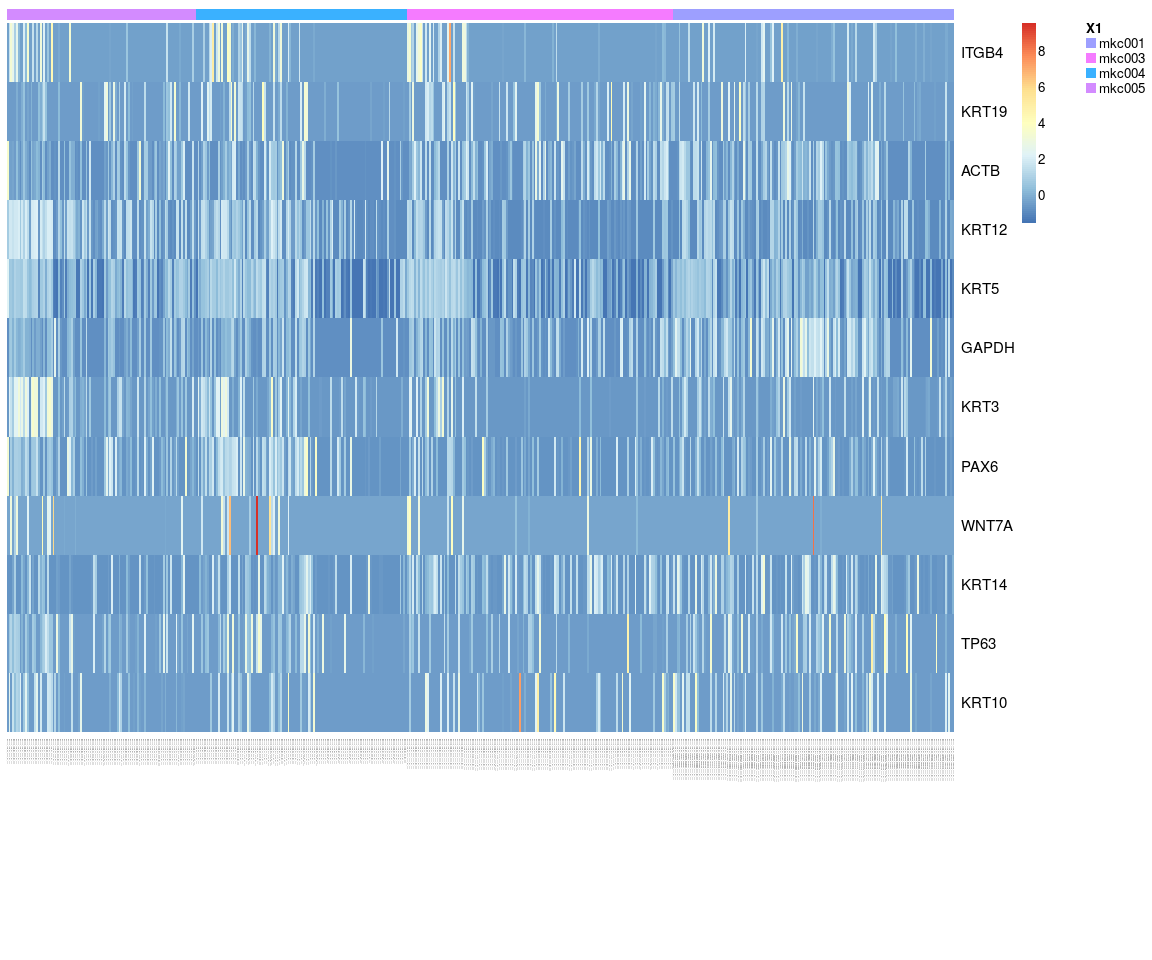


### Heatmap for important genes

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

## There ara 4 conditions  
## Whether creat data accurate 0

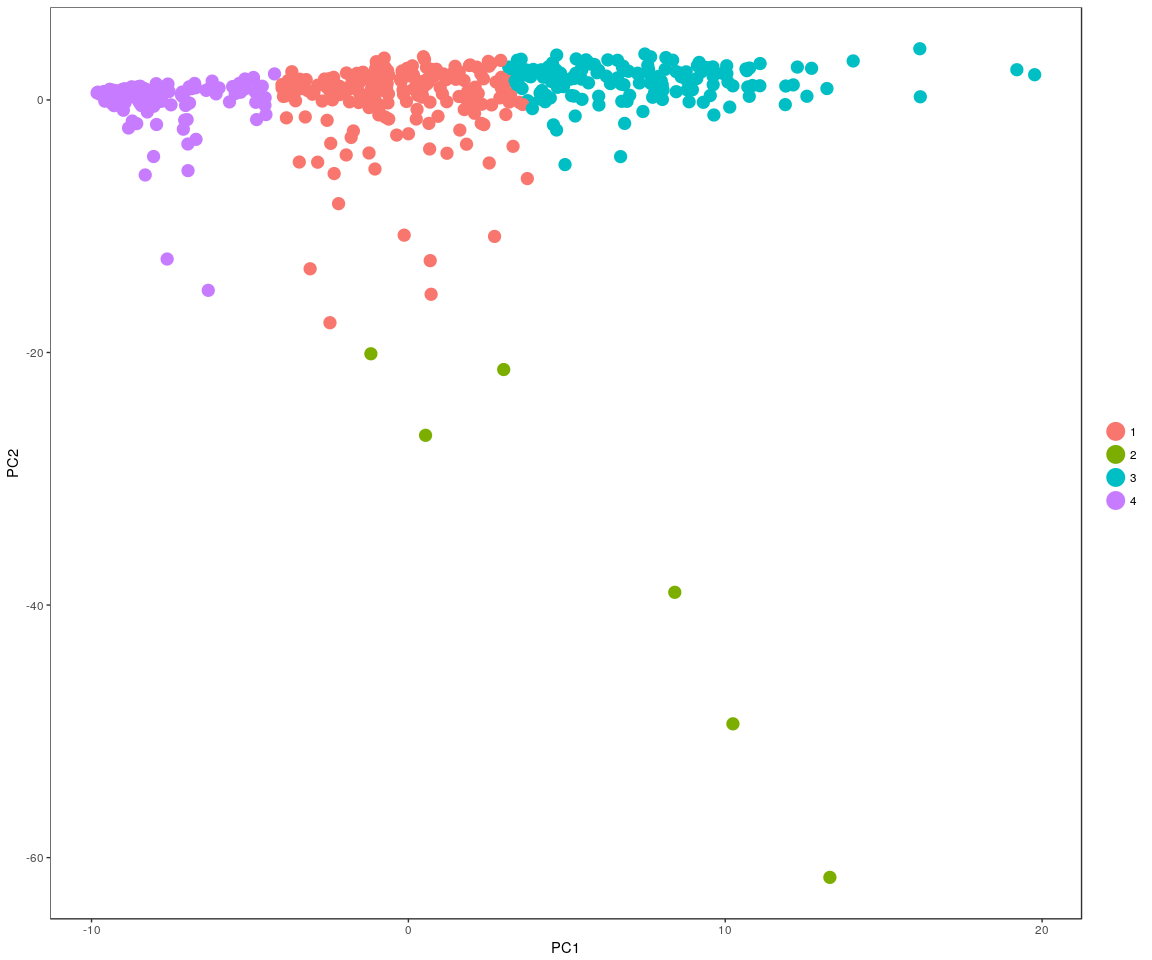
NMF::aheatmap(monkey.heatmap[[2]], Rowv = NA, Colv = NA, annCol = monkey.heatmap[[1]],   
 scale = "none")



The heatmap of genes ITGB4, KRT19, ACTB, KRT12, KRT5, GAPDH, KRT3, PAX6, WNT7A, KRT14, TP63, KRT10 .It tells us that **KRT14,WN&7A,ITGB4** expressed differently across sample,expressed more significant.

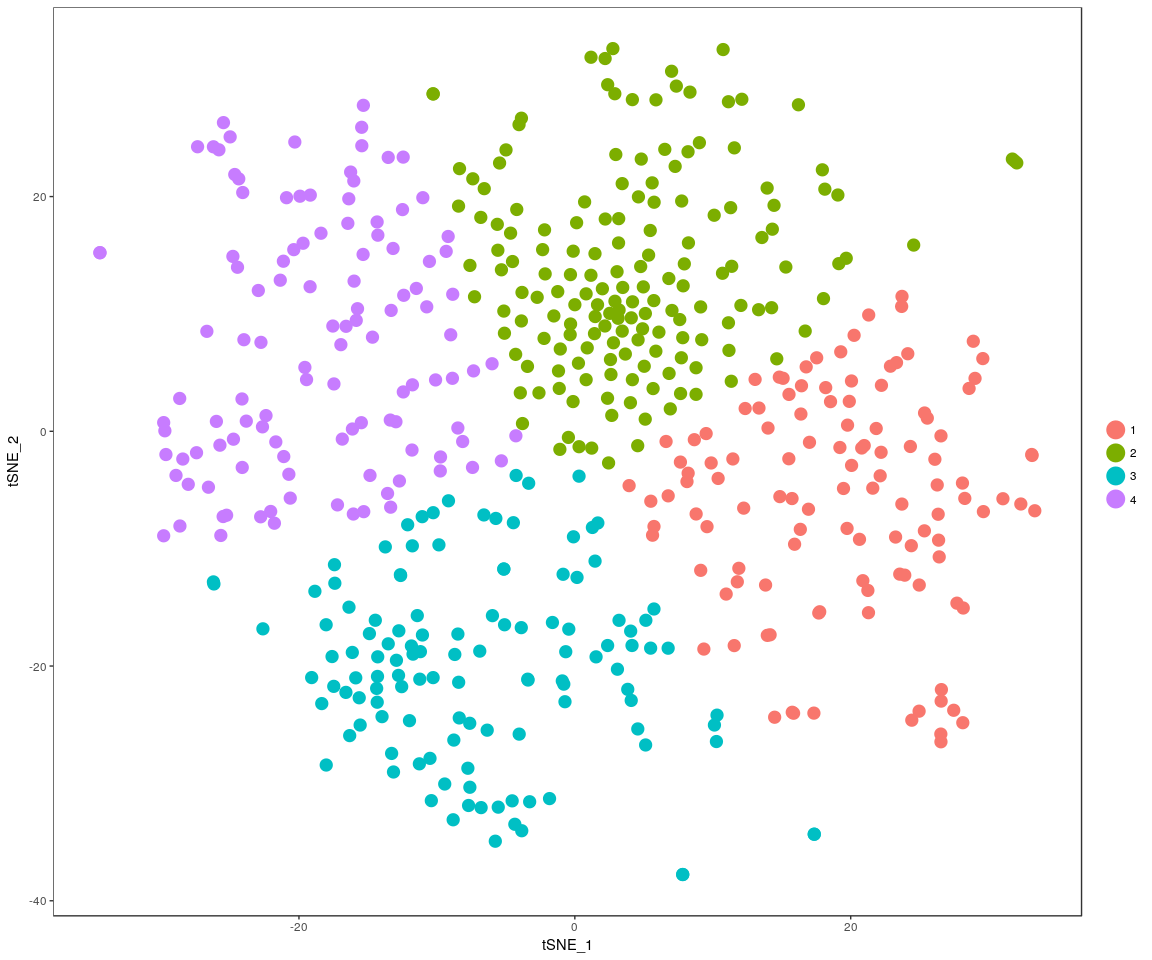
### Next,Spectral k-means clustering on single cells based on PCA

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



## Differential expression

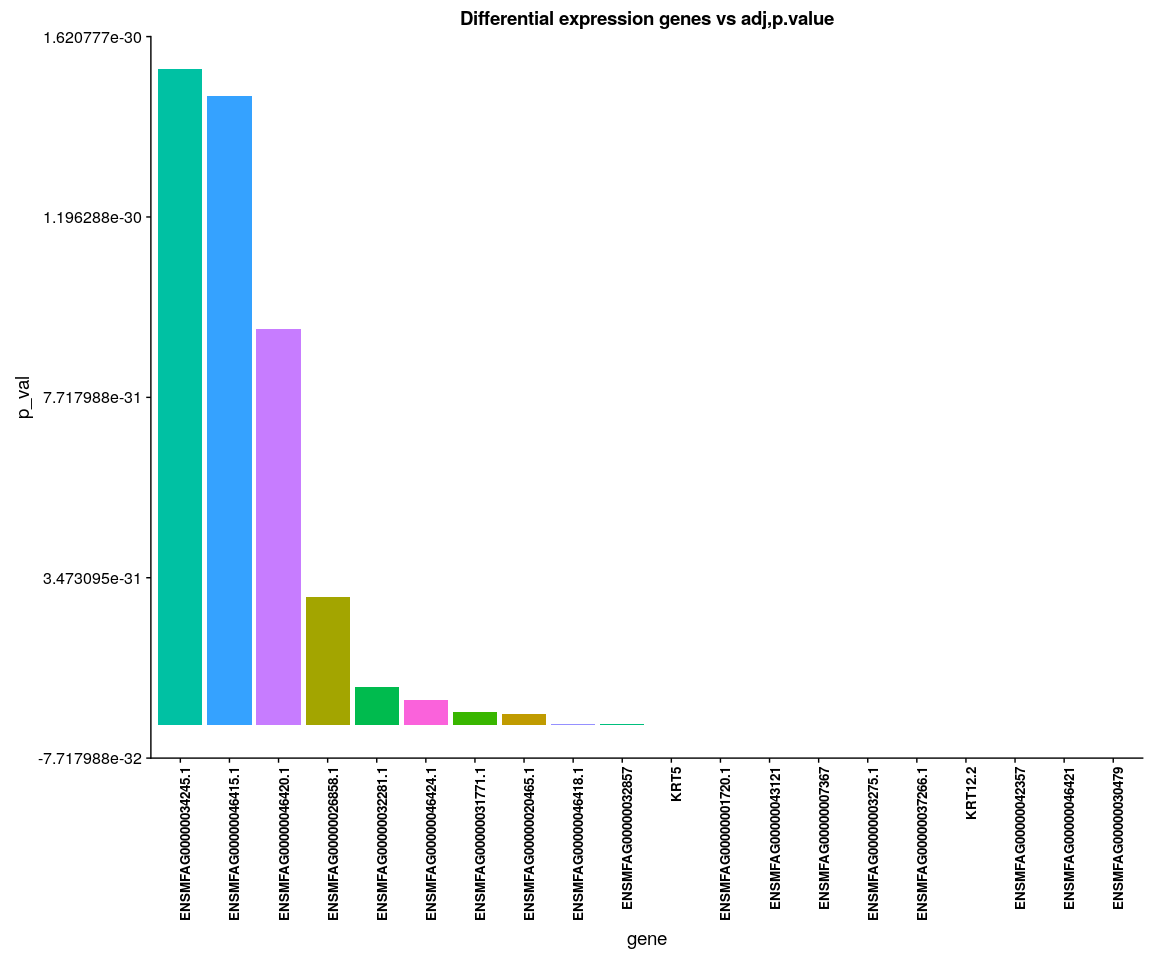
Next,we will have analysis on gene differential expression.Find maker genes across sample.We use the method: \*\*wilcox test\*\*

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

## p\_val avg\_logFC pct.1 pct.2 p\_val\_adj cluster  
## ENSMFAG00000041937 1.816973e-27 2.0113116 0.828 0.512 2.173645e-23 1  
## ENSMFAG00000035453 3.782086e-19 0.8493166 0.612 0.392 4.524509e-15 1  
## ENSMFAG00000044456 1.052680e-17 1.2813744 0.707 0.378 1.259321e-13 1  
## ENSMFAG00000044057 3.323185e-17 1.3091728 0.569 0.238 3.975526e-13 1  
## ENSMFAG00000045961 5.850502e-17 1.3622032 0.897 0.725 6.998955e-13 1  
## ENSMFAG00000028267 9.677991e-16 1.1379582 0.957 0.835 1.157778e-11 1  
## gene  
## ENSMFAG00000041937 ENSMFAG00000041937  
## ENSMFAG00000035453 ENSMFAG00000035453  
## ENSMFAG00000044456 ENSMFAG00000044456  
## ENSMFAG00000044057 ENSMFAG00000044057  
## ENSMFAG00000045961 ENSMFAG00000045961  
## ENSMFAG00000028267 ENSMFAG00000028267

We check whether the important genes are still in the marker genes we found from the DESeq analysis. the genes:ITGB4, KRT19, ACTB, KRT12, KRT5, GAPDH, KRT3, PAX6, KRT14, TP63 are still in the marker genes.

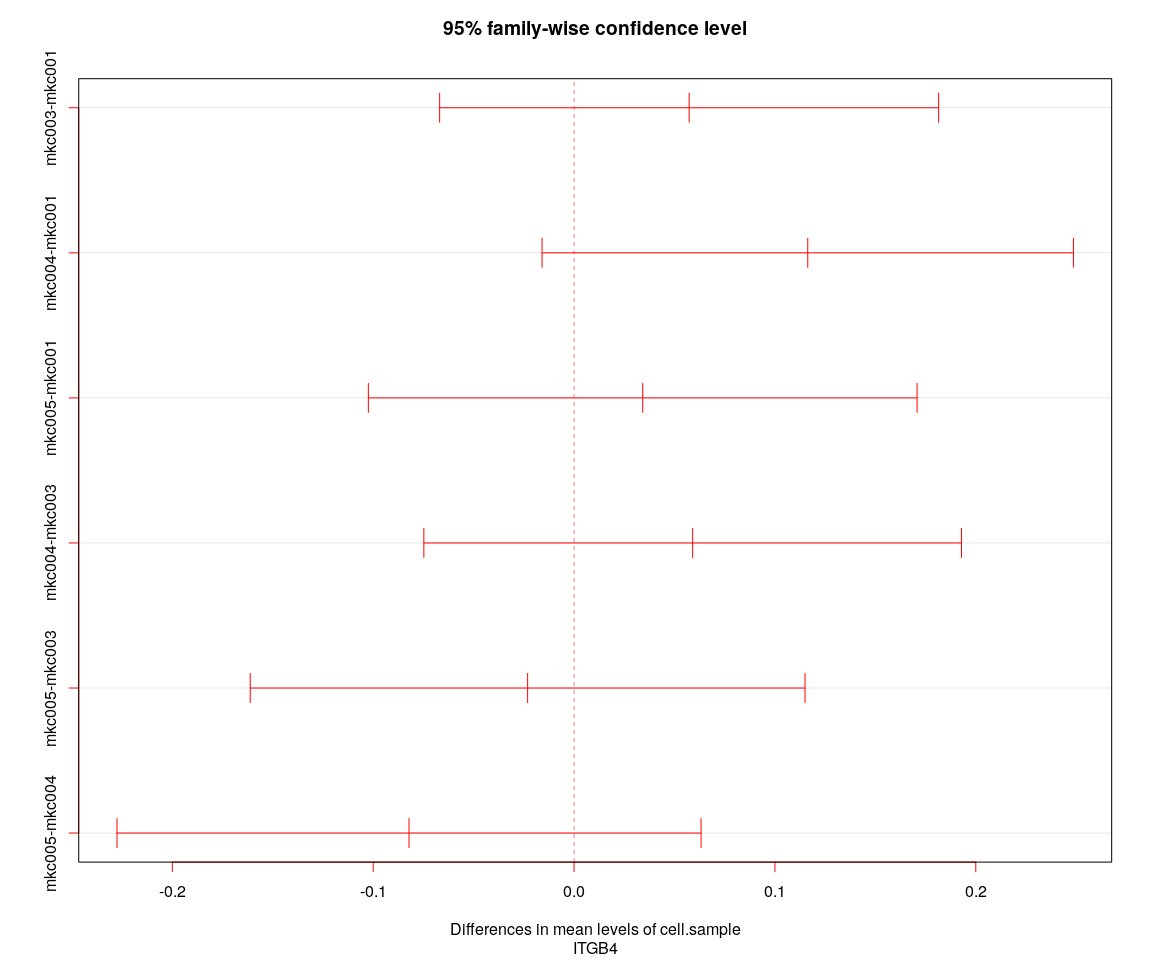
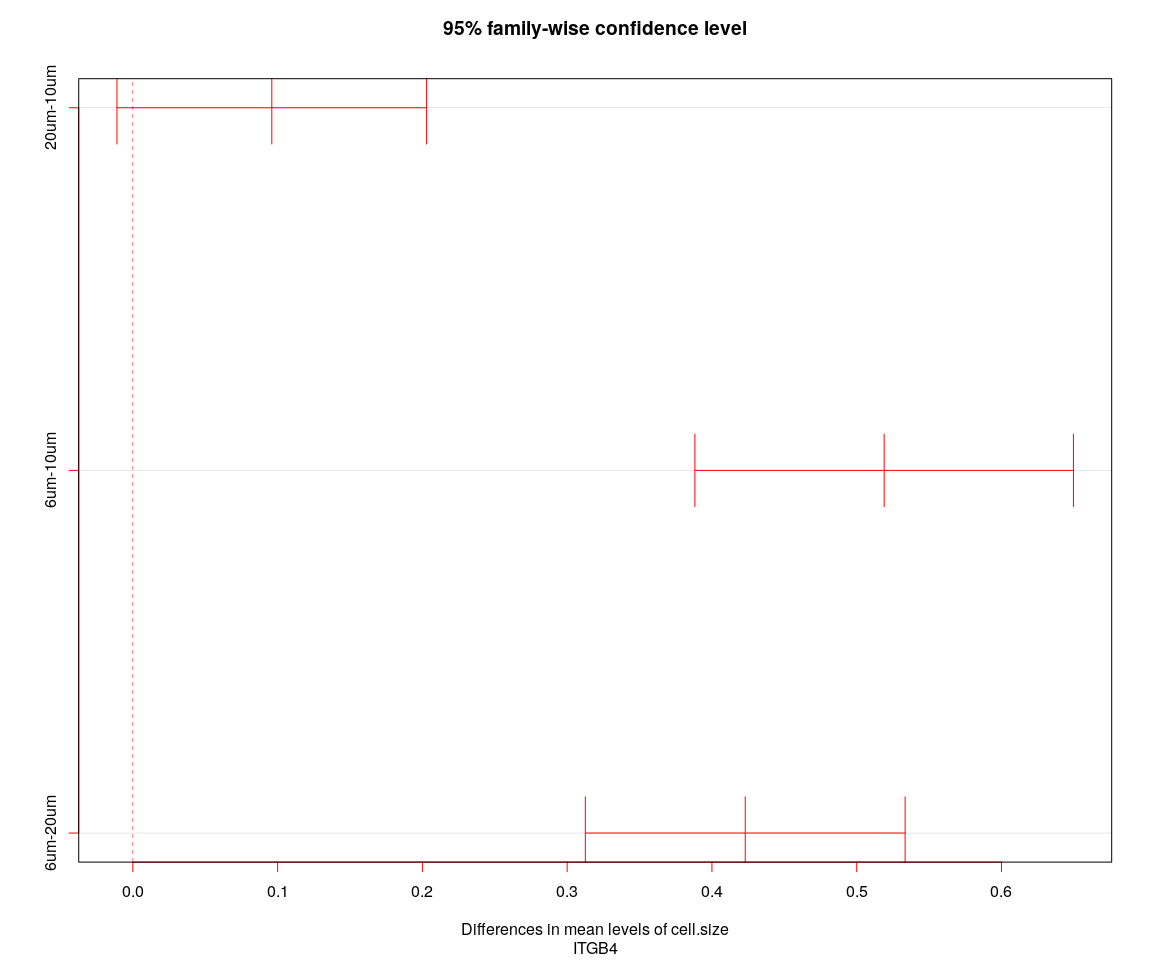
### Bar plot of gene’s p.val



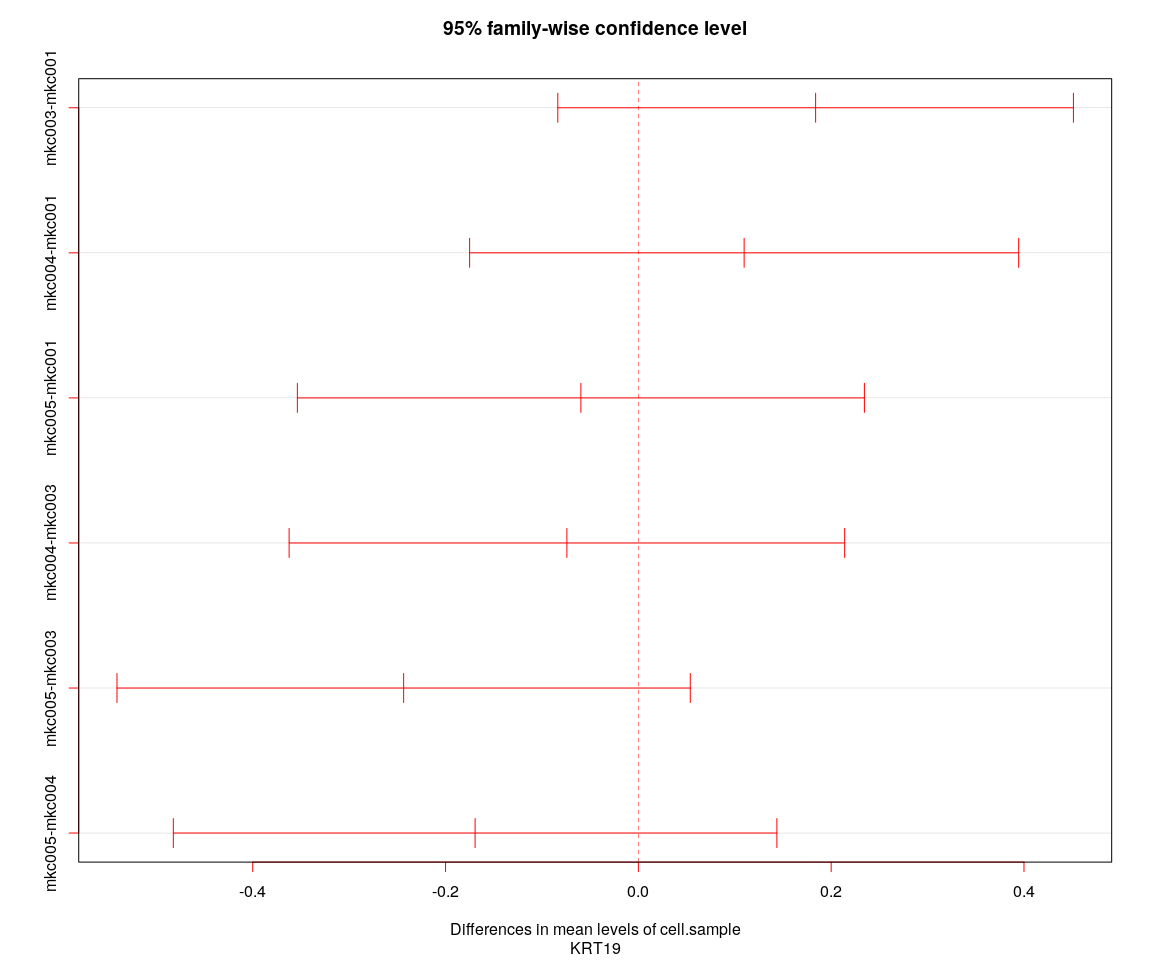
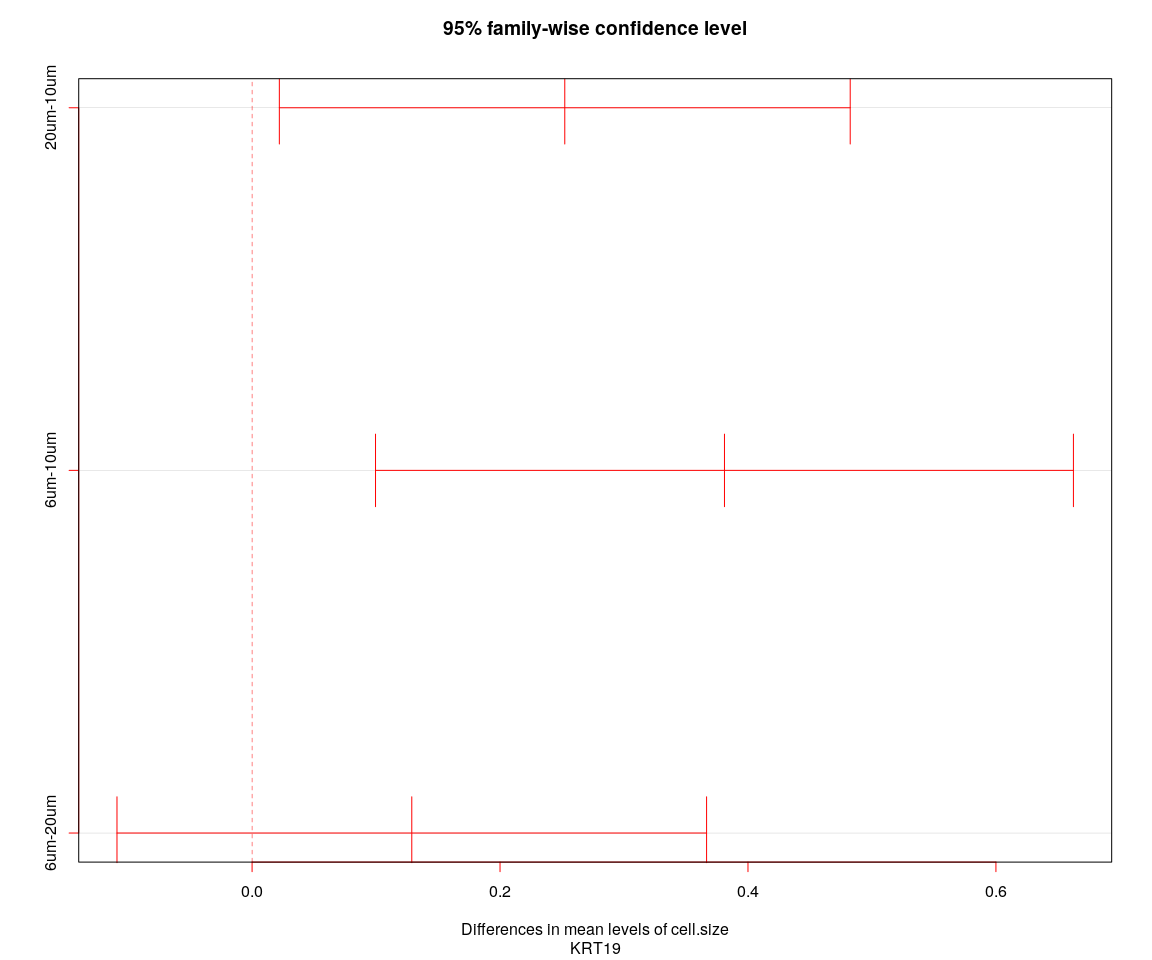
### Anova analsis:test the important genes’mean between cell size,cell groups ,whether is equal

#### anavo analysis

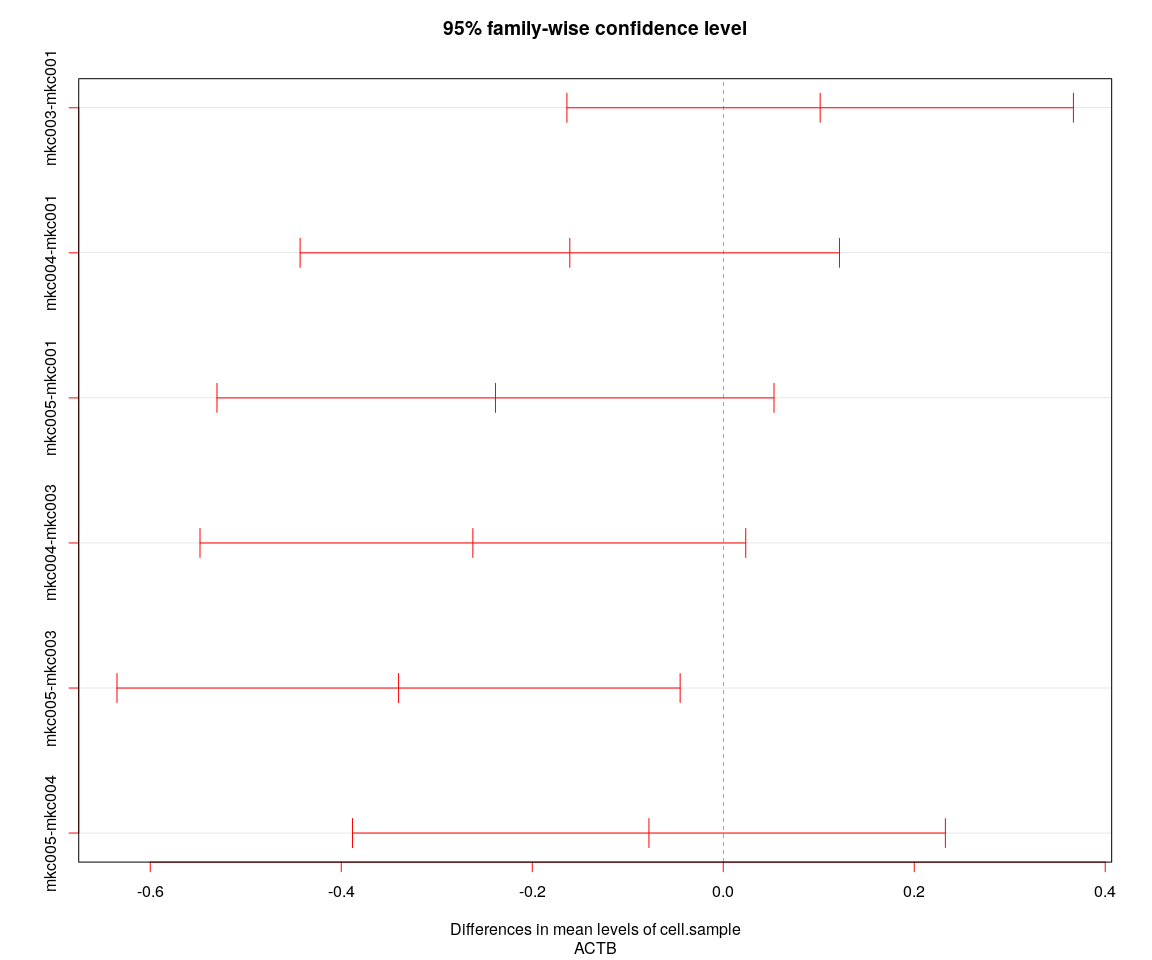
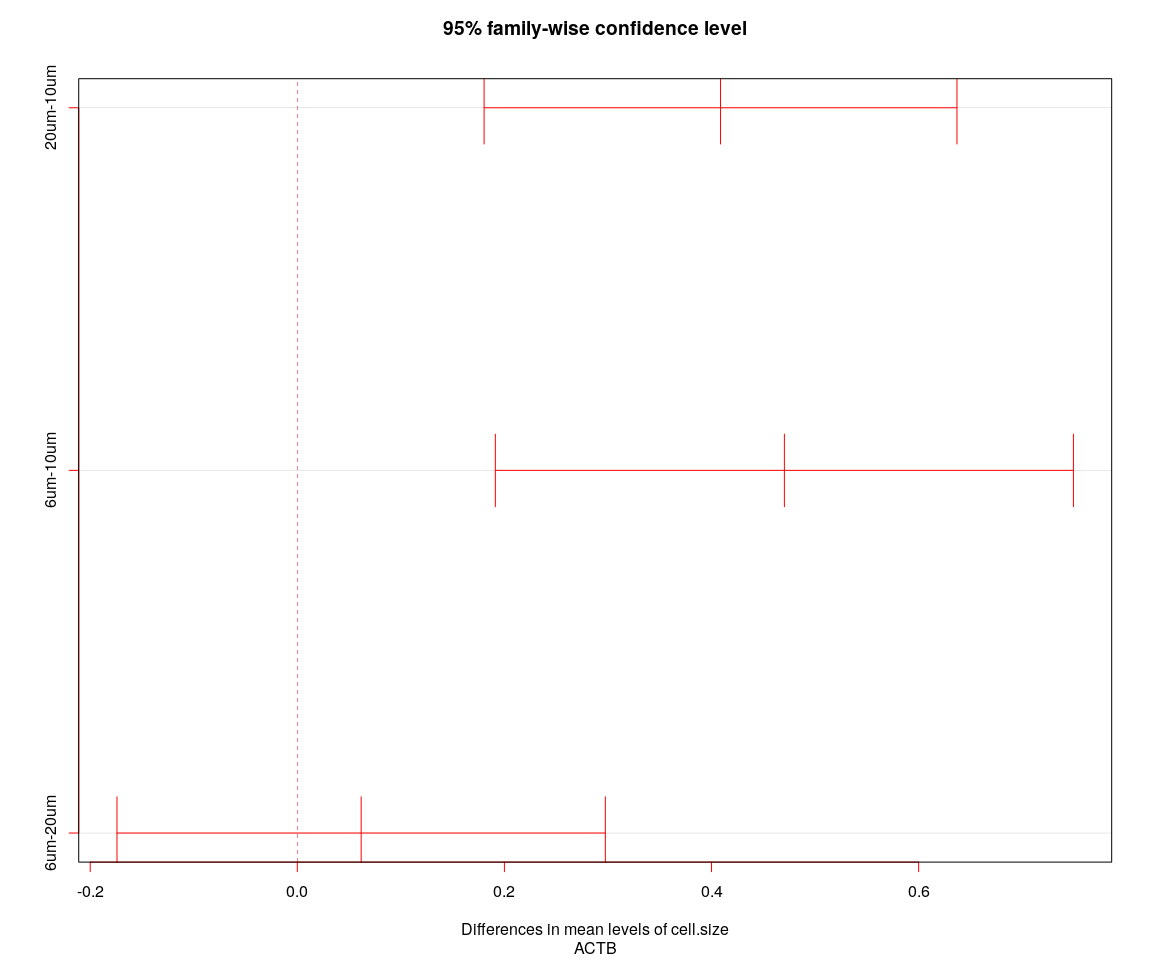
for (gene in unique(monkey.imp.lognorm.melt$variable)) {  
 aov.geen <- Anova.gene(data = monkey.imp.lognorm.melt, gene = gene, inter = FALSE,   
 plot.aov = FALSE)  
}



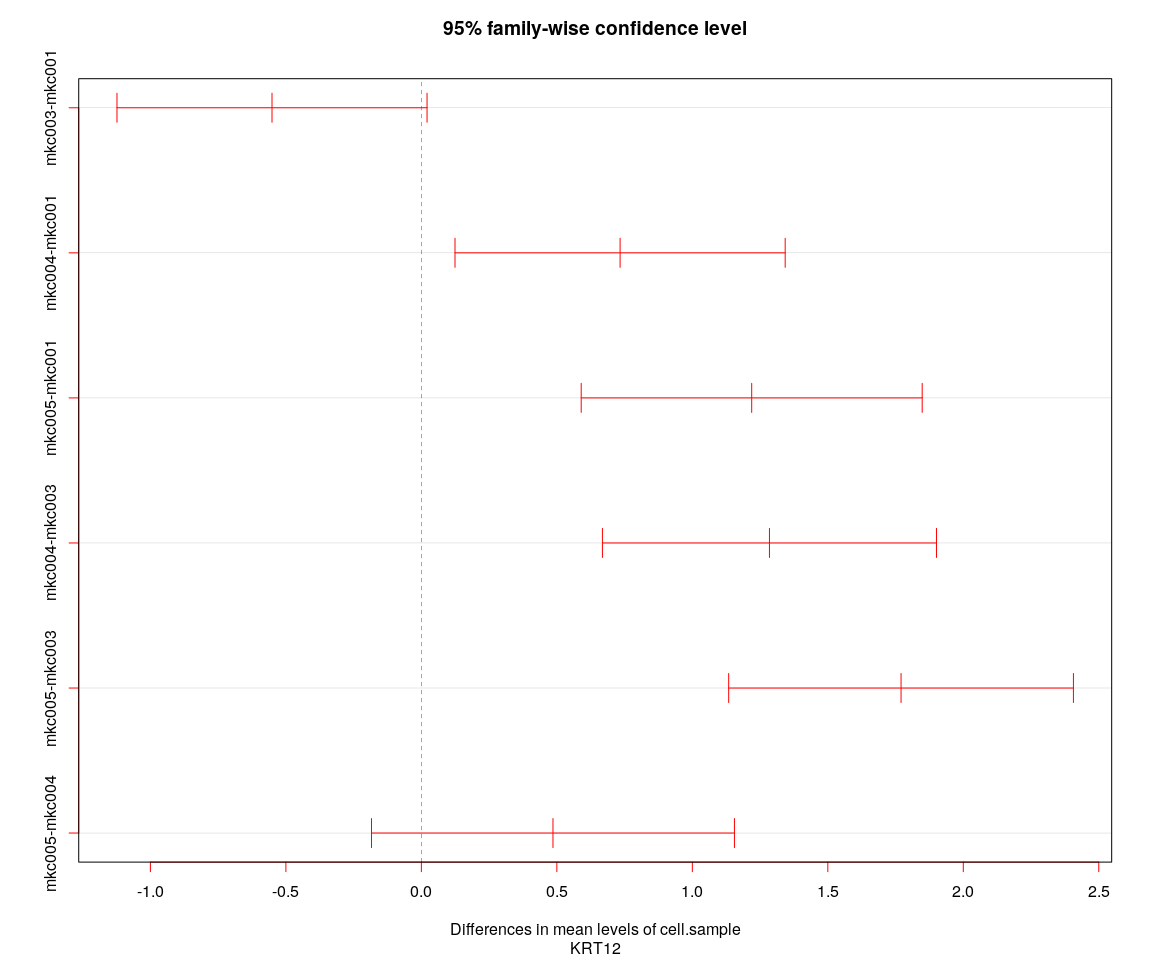
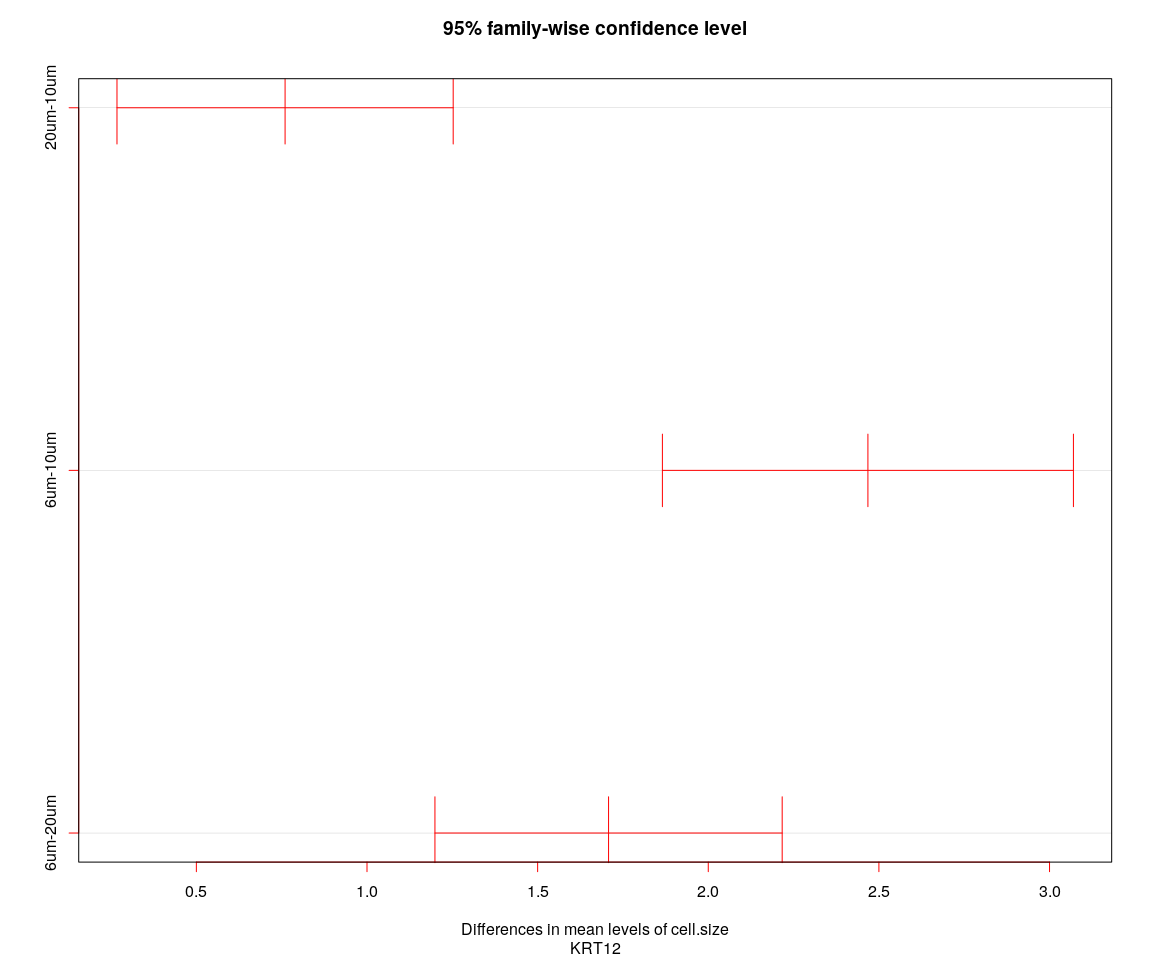
## [1] "----Gene:ITGB4----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.09596351 -0.01097439 0.2029014 0.0889305  
## 6um-10um 0.51896017 0.38822466 0.6496957 0.0000000  
## 6um-20um 0.42299666 0.31254875 0.5334446 0.0000000  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 0.05730000 -0.06691128 0.18151127 0.6340634  
## mkc004-mkc001 0.11634743 -0.01592075 0.24861560 0.1071268  
## mkc005-mkc001 0.03418516 -0.10241093 0.17078124 0.9172524  
## mkc004-mkc003 0.05904743 -0.07477729 0.19287216 0.6666093  
## mkc005-mkc003 -0.02311484 -0.16121870 0.11498903 0.9730629  
## mkc005-mkc004 -0.08216227 -0.22755521 0.06323067 0.4647106



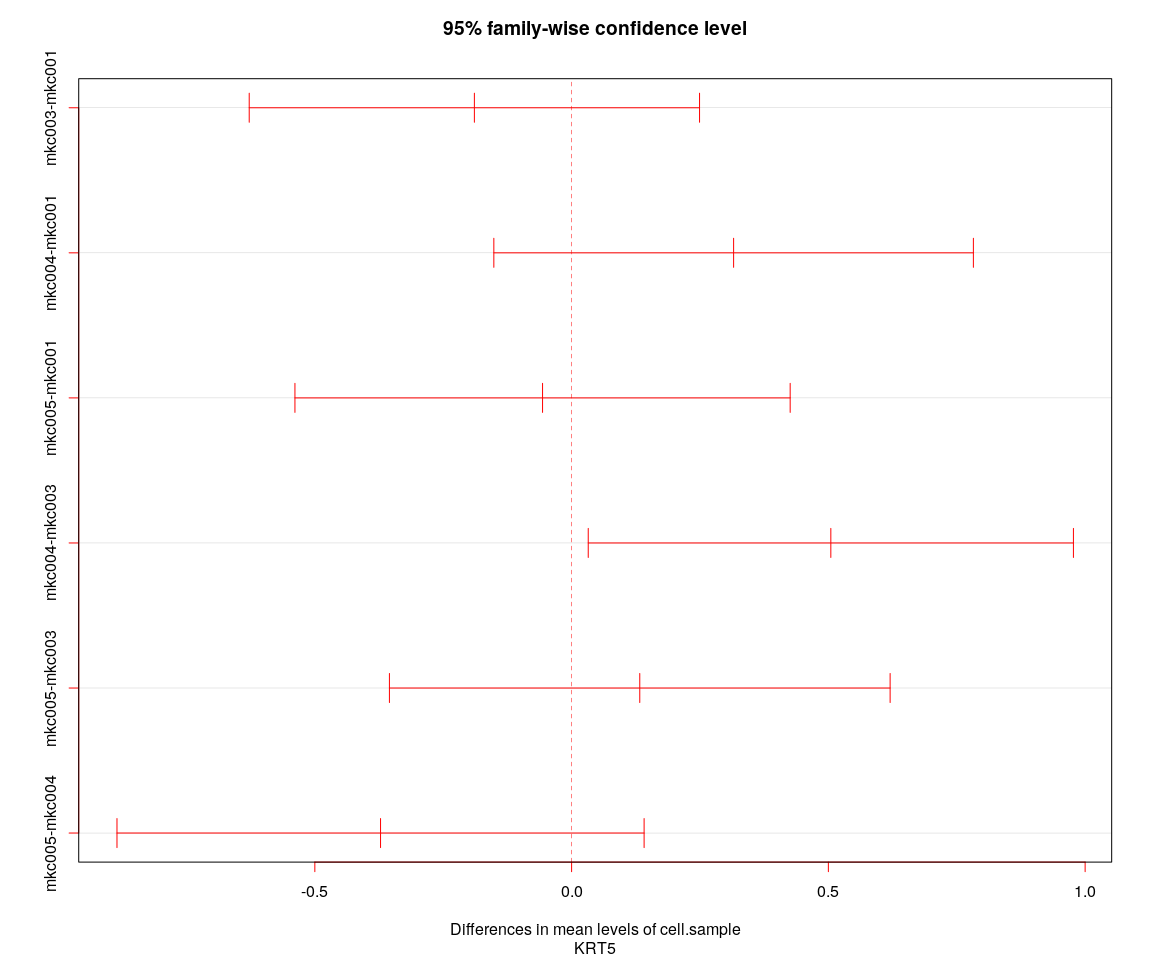
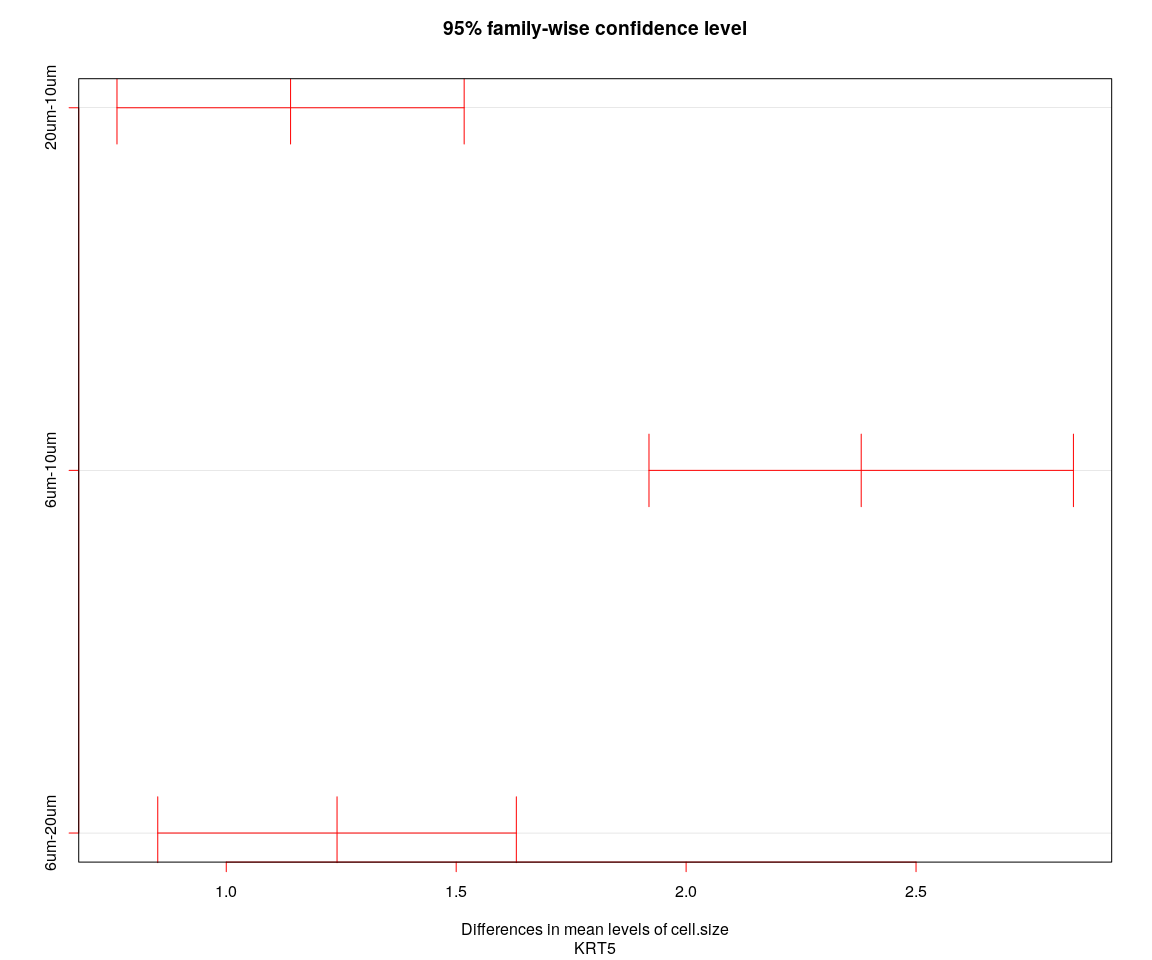
## [1] "----Gene:KRT19----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.2522231 0.02198440 0.4824619 0.0277647  
## 6um-10um 0.3810371 0.09956179 0.6625124 0.0044184  
## 6um-20um 0.1288140 -0.10898187 0.3666098 0.4109353  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 0.18385937 -0.08356916 0.45128790 0.2880294  
## mkc004-mkc001 0.10972452 -0.17505062 0.39449967 0.7534197  
## mkc005-mkc001 -0.05961805 -0.35371124 0.23447515 0.9536544  
## mkc004-mkc003 -0.07413485 -0.36226126 0.21399156 0.9108489  
## mkc005-mkc003 -0.24347742 -0.54081688 0.05386204 0.1510606  
## mkc005-mkc004 -0.16934257 -0.48237550 0.14369037 0.5034343



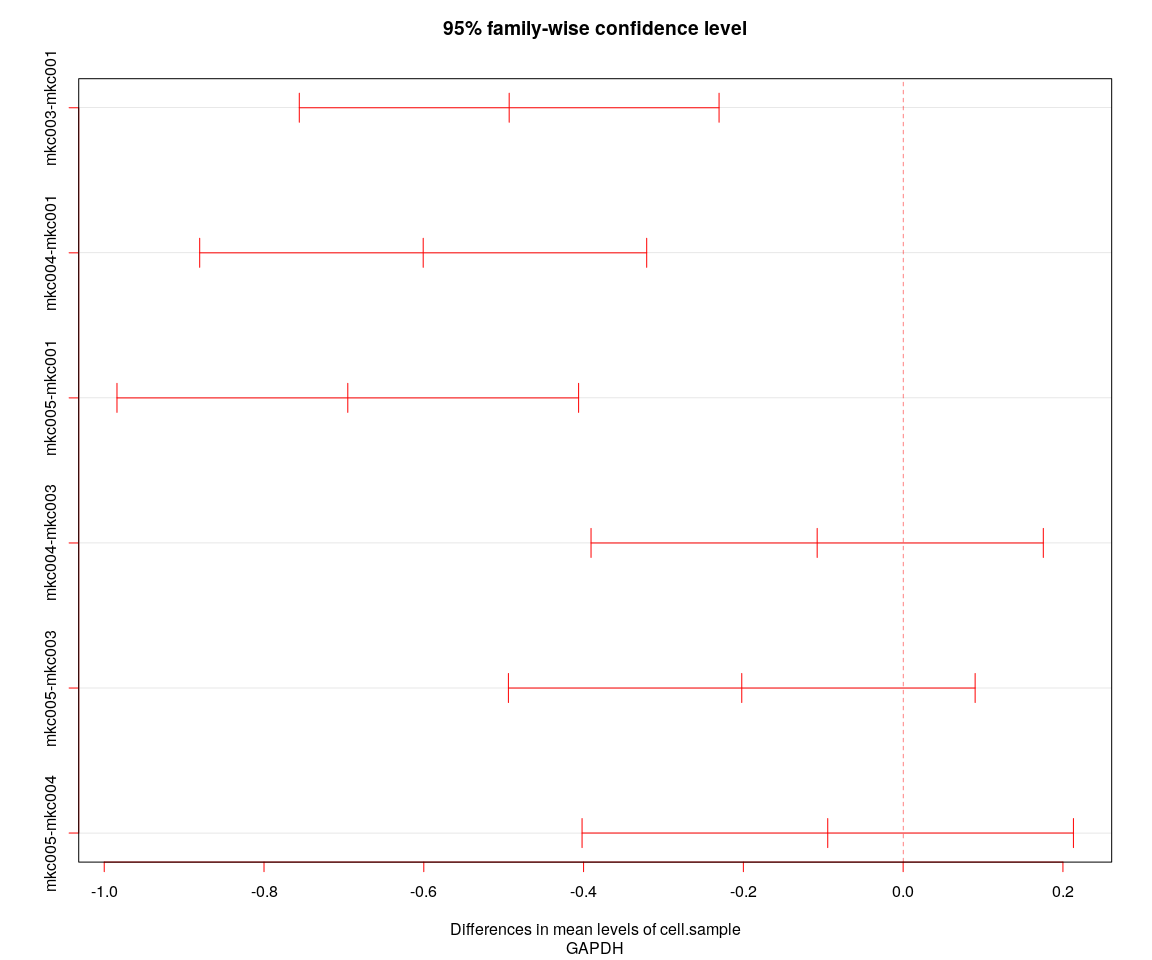
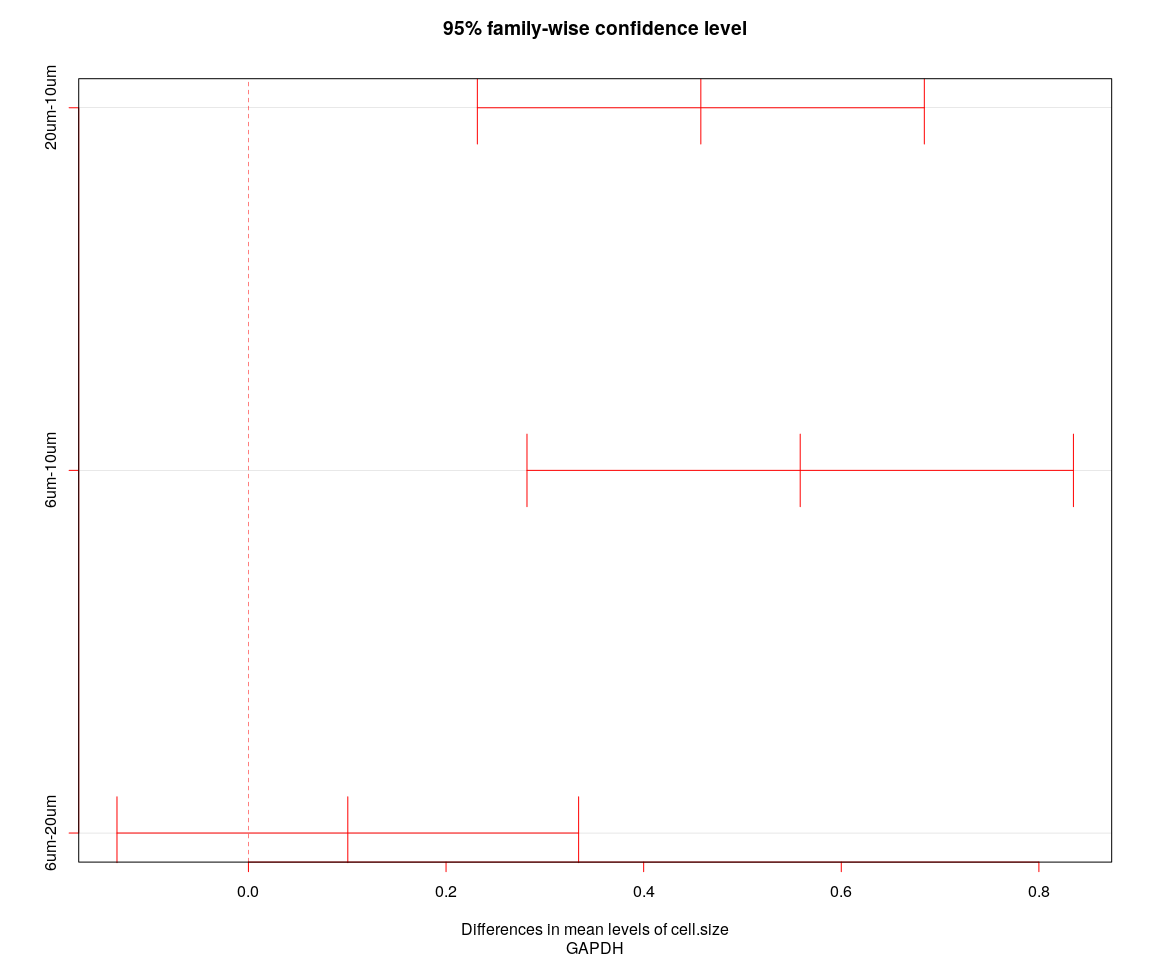
## [1] "----Gene:ACTB----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.40870123 0.1803834 0.6370190 0.0000903  
## 6um-10um 0.47036049 0.1912336 0.7494874 0.0002511  
## 6um-20um 0.06165926 -0.1741526 0.2974711 0.8121925  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 0.1014917 -0.1637056 0.36668903 0.7572631  
## mkc004-mkc001 -0.1606860 -0.4430852 0.12171317 0.4585378  
## mkc005-mkc001 -0.2384984 -0.5301379 0.05314108 0.1519376  
## mkc004-mkc003 -0.2621777 -0.5479003 0.02354476 0.0852204  
## mkc005-mkc003 -0.3399902 -0.6348488 -0.04513147 0.0163201  
## mkc005-mkc004 -0.0778124 -0.3882336 0.23260882 0.9168930



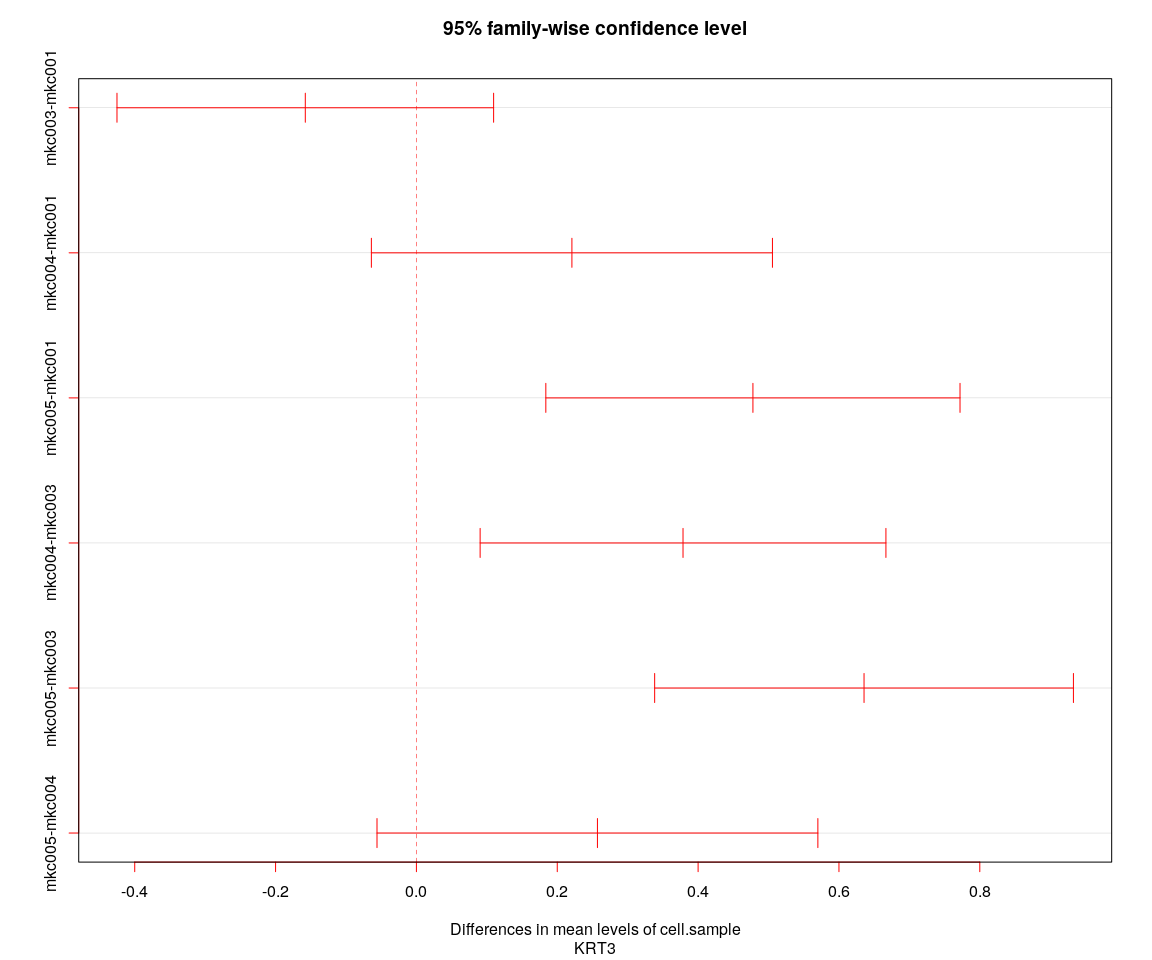
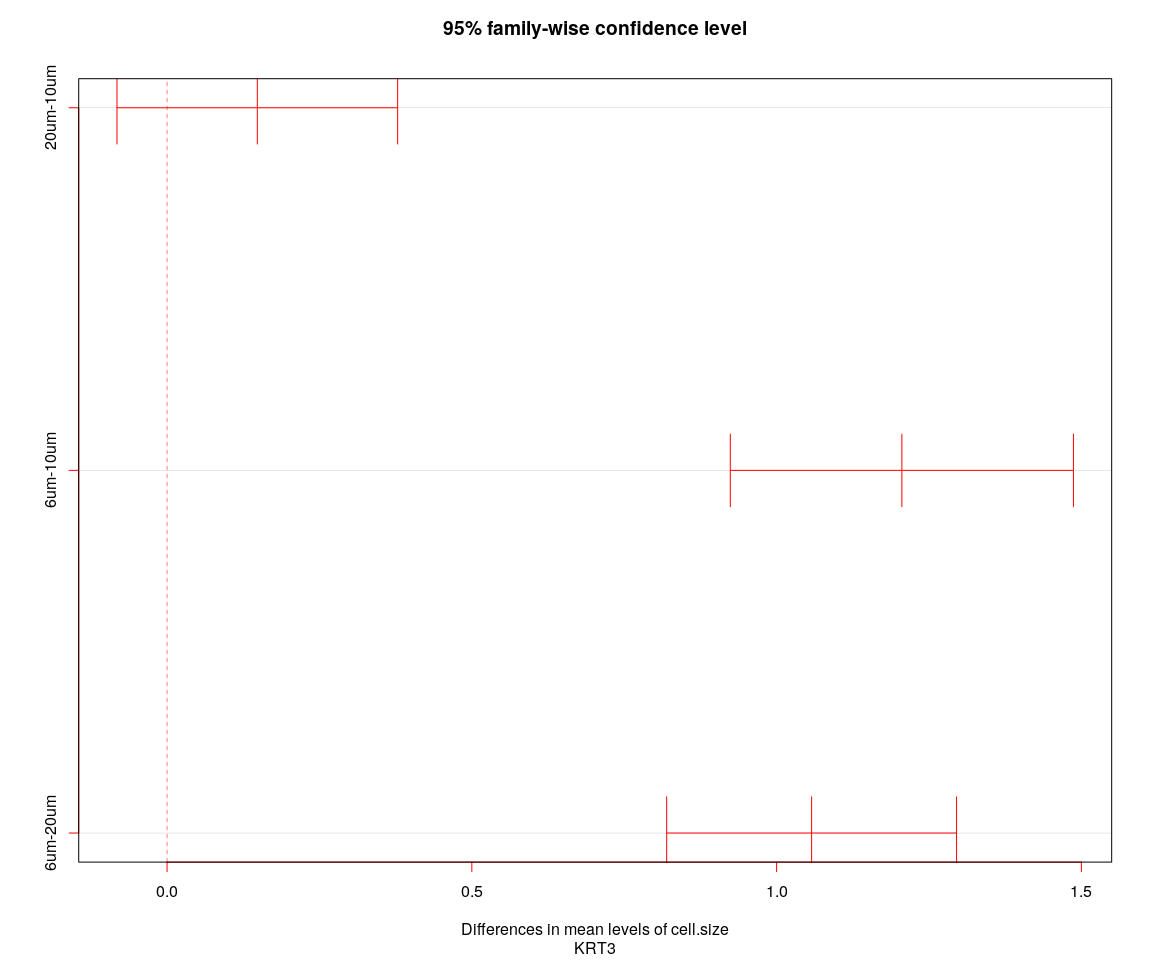
## [1] "----Gene:KRT12----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.7599665 0.2673791 1.252554 0.0009212  
## 6um-10um 2.4677949 1.8655889 3.070001 0.0000000  
## 6um-20um 1.7078285 1.1990730 2.216584 0.0000000  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 -0.5512164 -1.1233700 0.02093715 0.0637485  
## mkc004-mkc001 0.7334023 0.1241363 1.34266830 0.0108672  
## mkc005-mkc001 1.2190785 0.5898769 1.84828012 0.0000048  
## mkc004-mkc003 1.2846187 0.6681828 1.90105462 0.0000007  
## mkc005-mkc003 1.7702949 1.1341480 2.40644179 0.0000000  
## mkc005-mkc004 0.4856762 -0.1840463 1.15539871 0.2426737



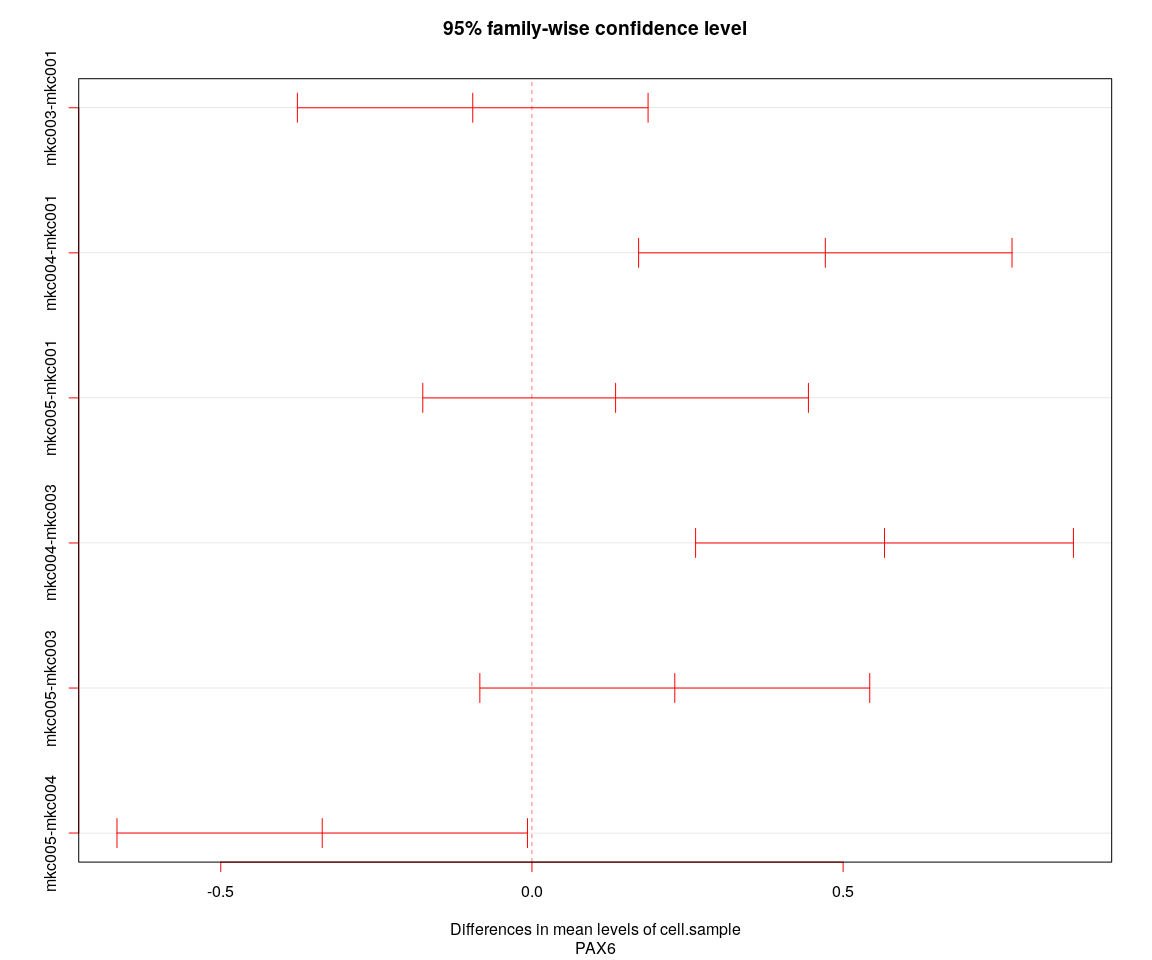
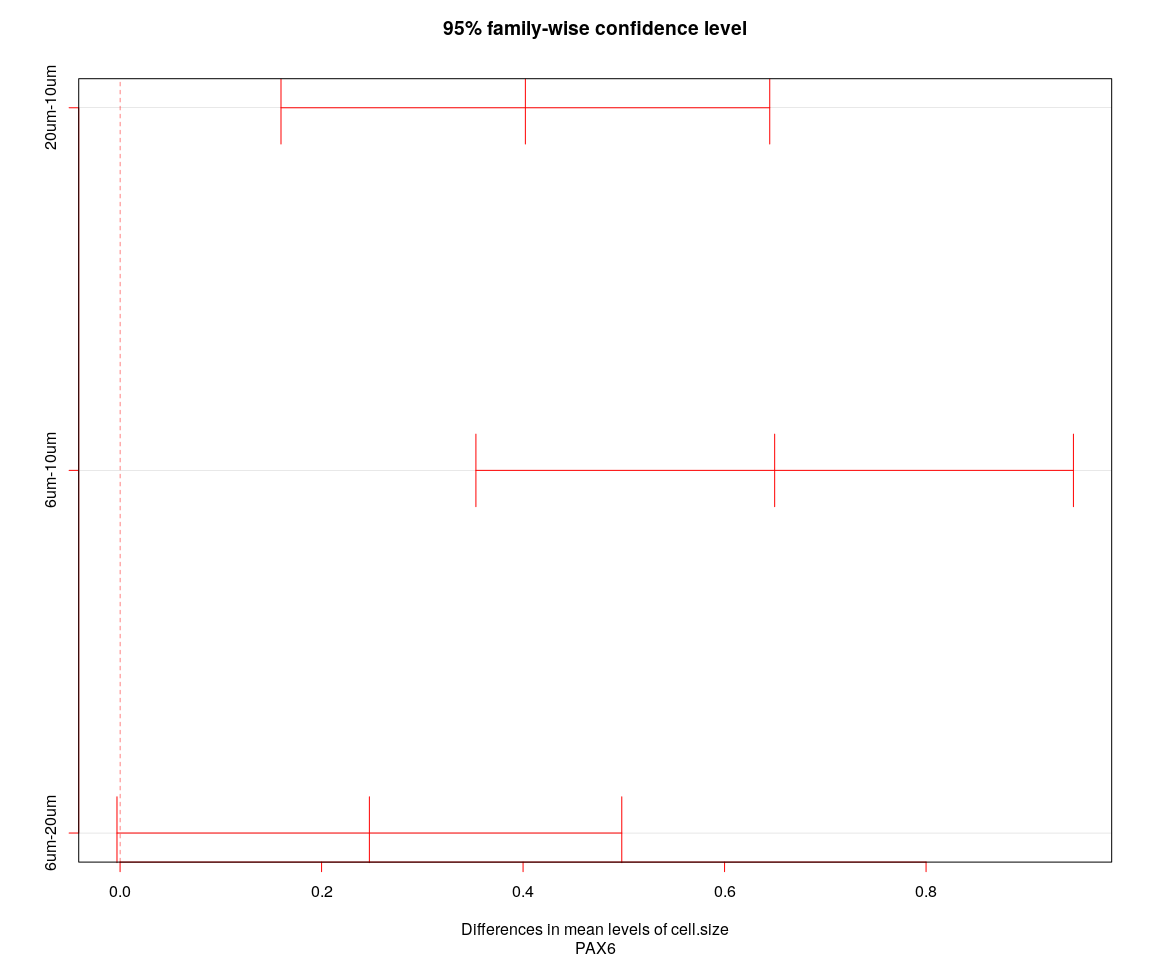
## [1] "----Gene:KRT5----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 1.139845 0.7623508 1.517340 0  
## 6um-10um 2.380757 1.9192563 2.842258 0  
## 6um-20um 1.240912 0.8510268 1.630797 0  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 -0.18931896 -0.6277892 0.2491513 0.6817464  
## mkc004-mkc001 0.31548074 -0.1514307 0.7823922 0.3032467  
## mkc005-mkc001 -0.05657273 -0.5387618 0.4256164 0.9903884  
## mkc004-mkc003 0.50479970 0.0323936 0.9772058 0.0309032  
## mkc005-mkc003 0.13274623 -0.3547654 0.6202579 0.8963400  
## mkc005-mkc004 -0.37205347 -0.8852958 0.1411889 0.2429954



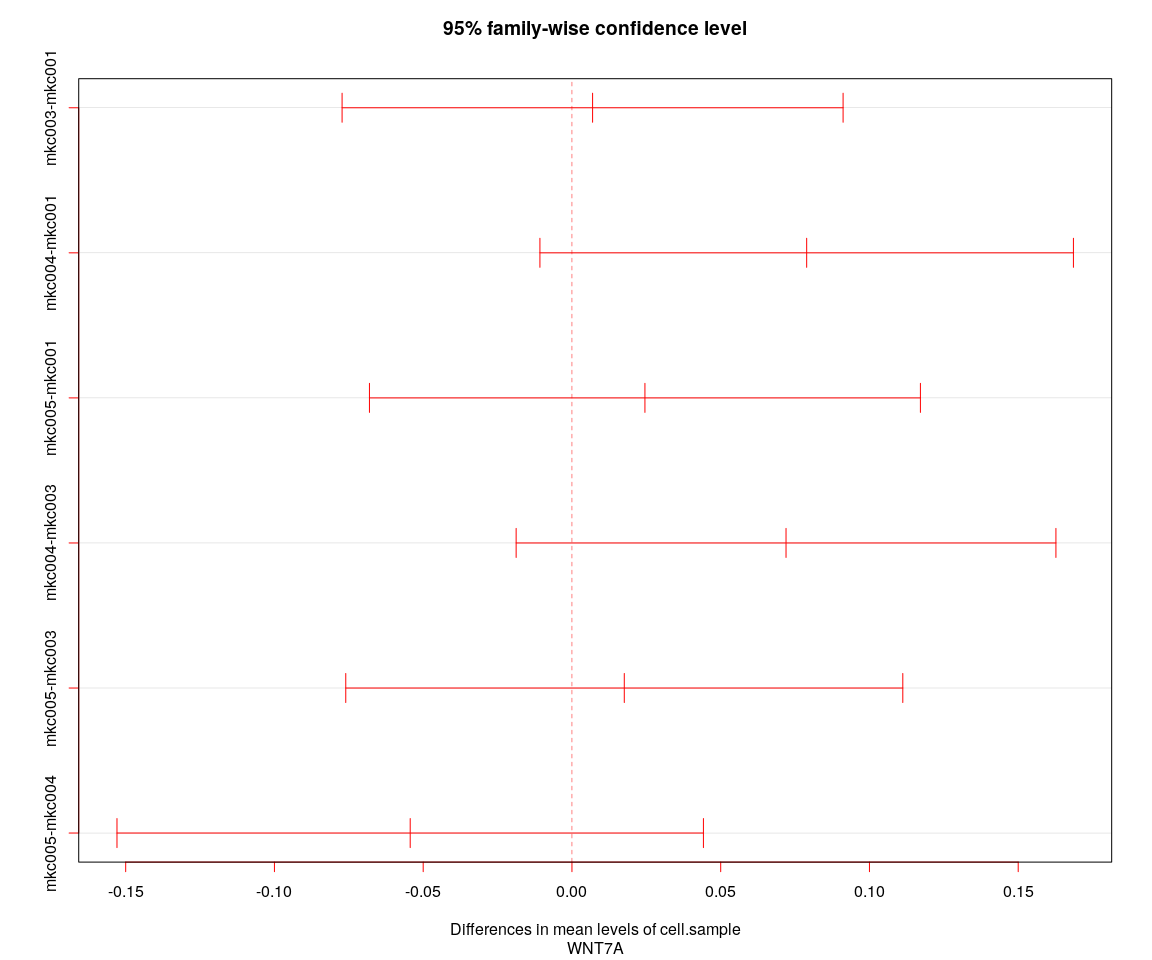
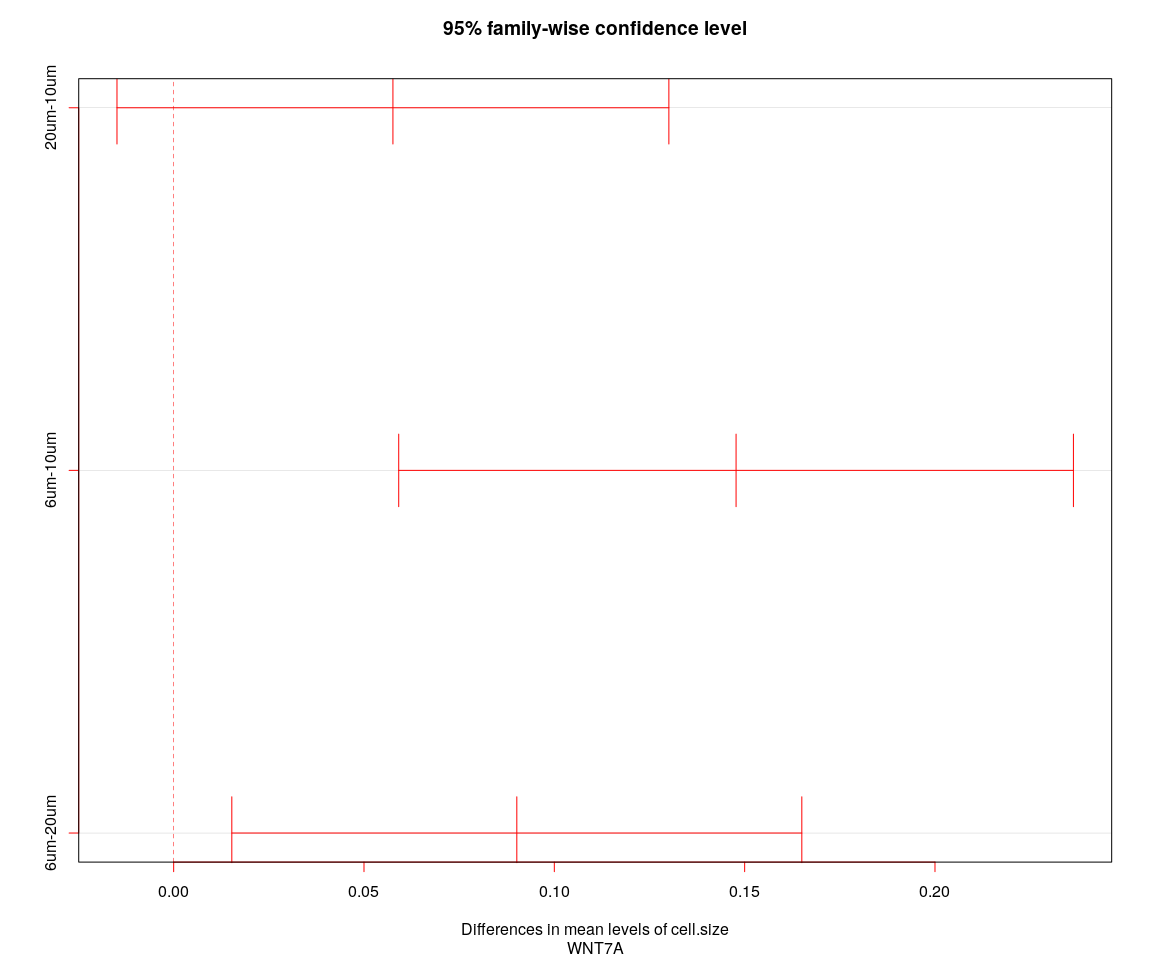
## [1] "----Gene:GAPDH----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.4578612 0.2316786 0.6840438 0.0000076  
## 6um-10um 0.5584298 0.2819133 0.8349463 0.0000080  
## 6um-20um 0.1005686 -0.1330379 0.3341751 0.5696489  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 -0.49313268 -0.7558499 -0.23041548 0.0000103  
## mkc004-mkc001 -0.60075025 -0.8805085 -0.32099202 0.0000003  
## mkc005-mkc001 -0.69513261 -0.9840447 -0.40622050 0.0000000  
## mkc004-mkc003 -0.10761756 -0.3906680 0.17543289 0.7609346  
## mkc005-mkc003 -0.20199993 -0.4941011 0.09010126 0.2829480  
## mkc005-mkc004 -0.09438237 -0.4019006 0.21313582 0.8585046



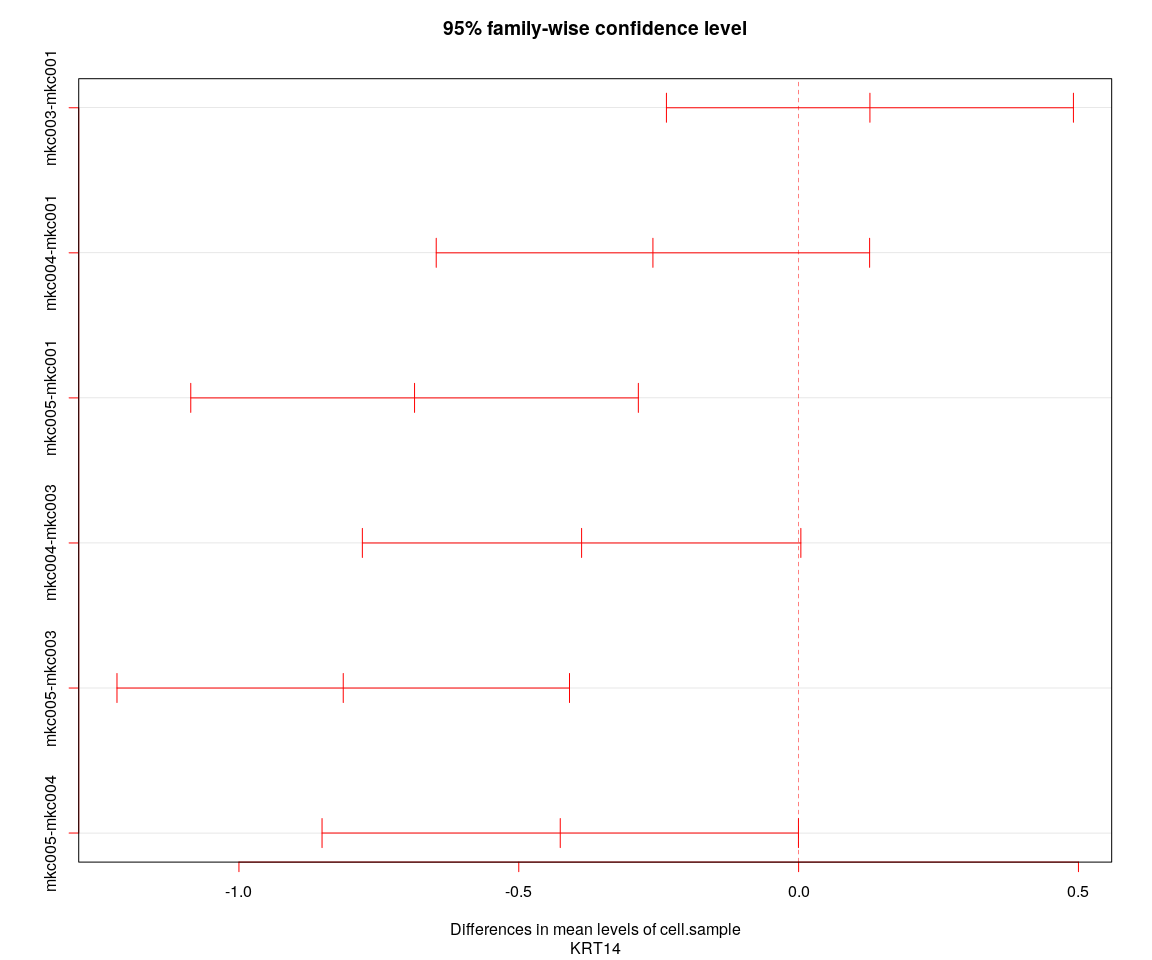
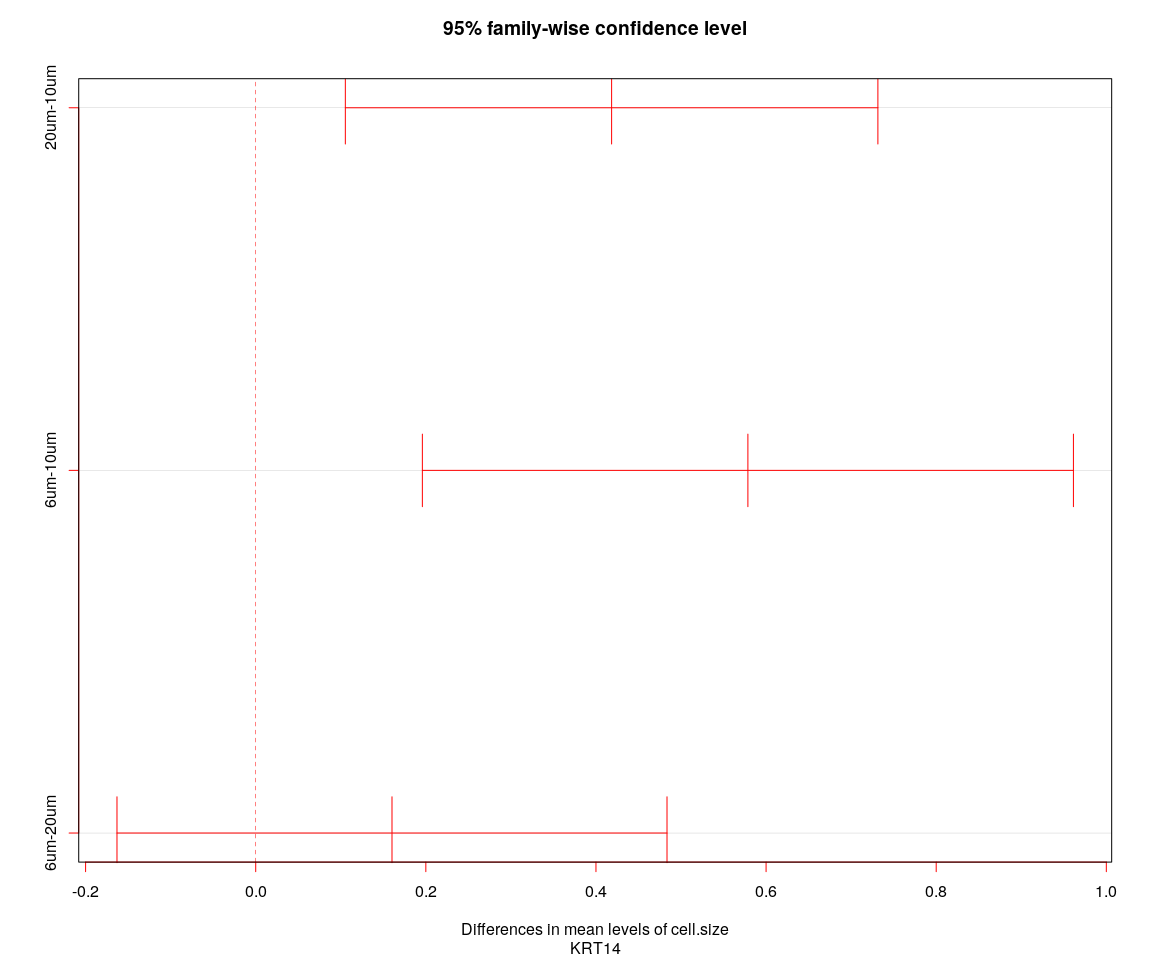
## [1] "----Gene:KRT3----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.1480594 -0.08216785 0.3782867 0.2862556  
## 6um-10um 1.2055059 0.92404463 1.4869672 0.0000000  
## 6um-20um 1.0574465 0.81966248 1.2952305 0.0000000  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 -0.1577871 -0.42520232 0.1096282 0.4257004  
## mkc004-mkc001 0.2207948 -0.06396622 0.5055558 0.1898489  
## mkc005-mkc001 0.4778406 0.18376206 0.7719192 0.0001937  
## mkc004-mkc003 0.3785819 0.09046975 0.6666940 0.0042269  
## mkc005-mkc003 0.6356277 0.33830302 0.9329524 0.0000003  
## mkc005-mkc004 0.2570459 -0.05597153 0.5700633 0.1491576



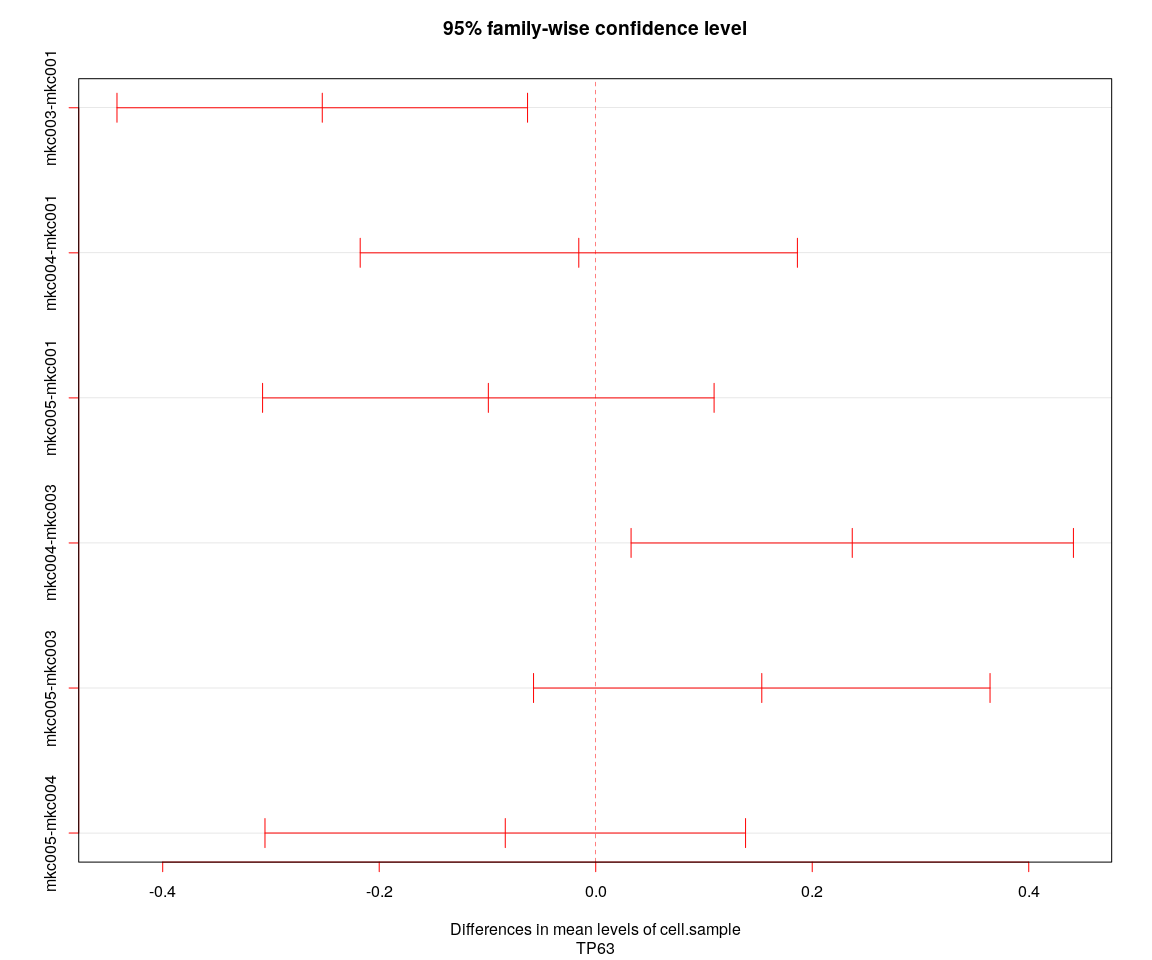
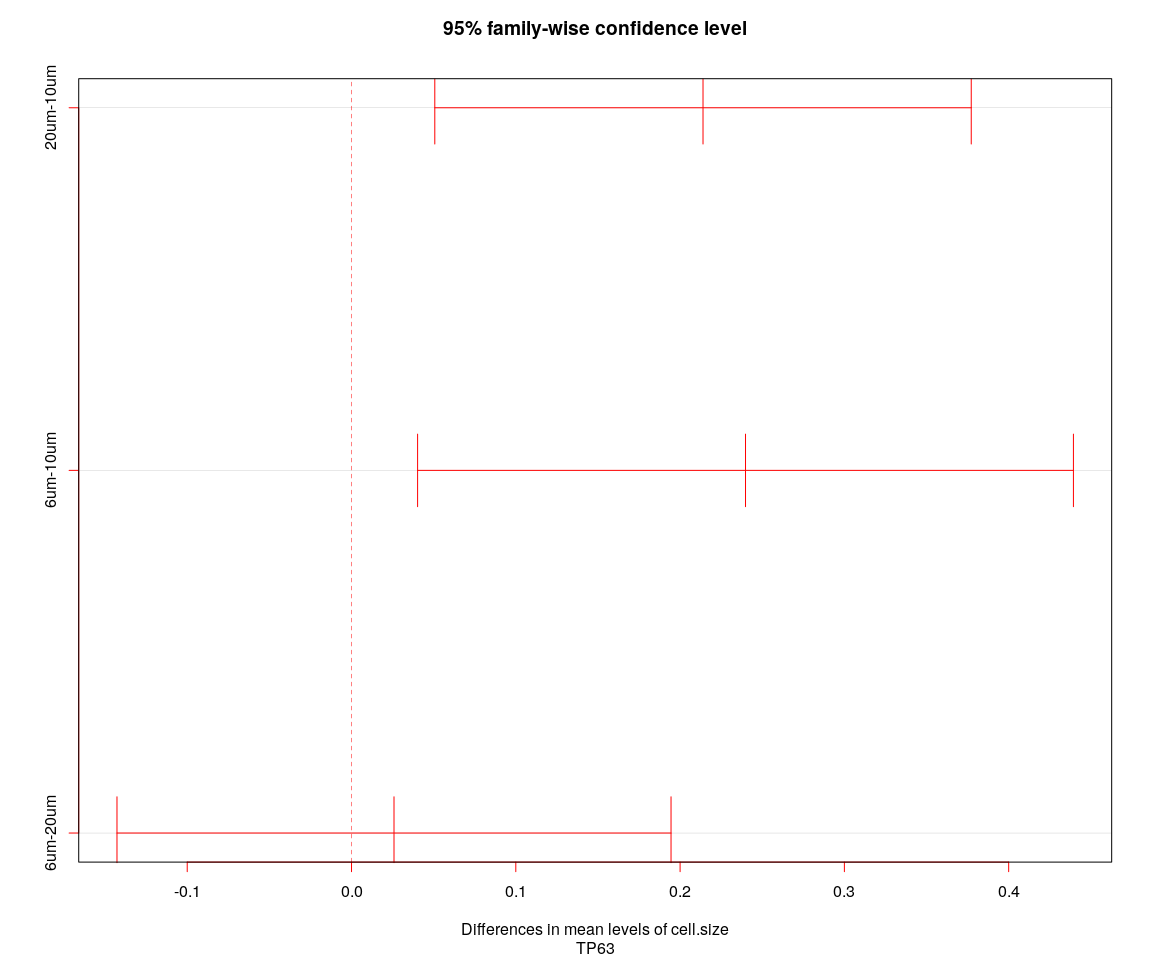
## [1] "----Gene:PAX6----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.4022790 0.159717581 0.6448404 0.0003228  
## 6um-10um 0.6497248 0.353184593 0.9462651 0.0000011  
## 6um-20um 0.2474458 -0.003077163 0.4979688 0.0537337  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 -0.09512192 -0.37686360 0.186619758 0.8202531  
## mkc004-mkc001 0.47140106 0.17138434 0.771417768 0.0003442  
## mkc005-mkc001 0.13442630 -0.17540718 0.444259776 0.6784063  
## mkc004-mkc003 0.56652298 0.26297563 0.870070324 0.0000118  
## mkc005-mkc003 0.22954822 -0.08370526 0.542801708 0.2341157  
## mkc005-mkc004 -0.33697476 -0.66676166 -0.007187856 0.0430659



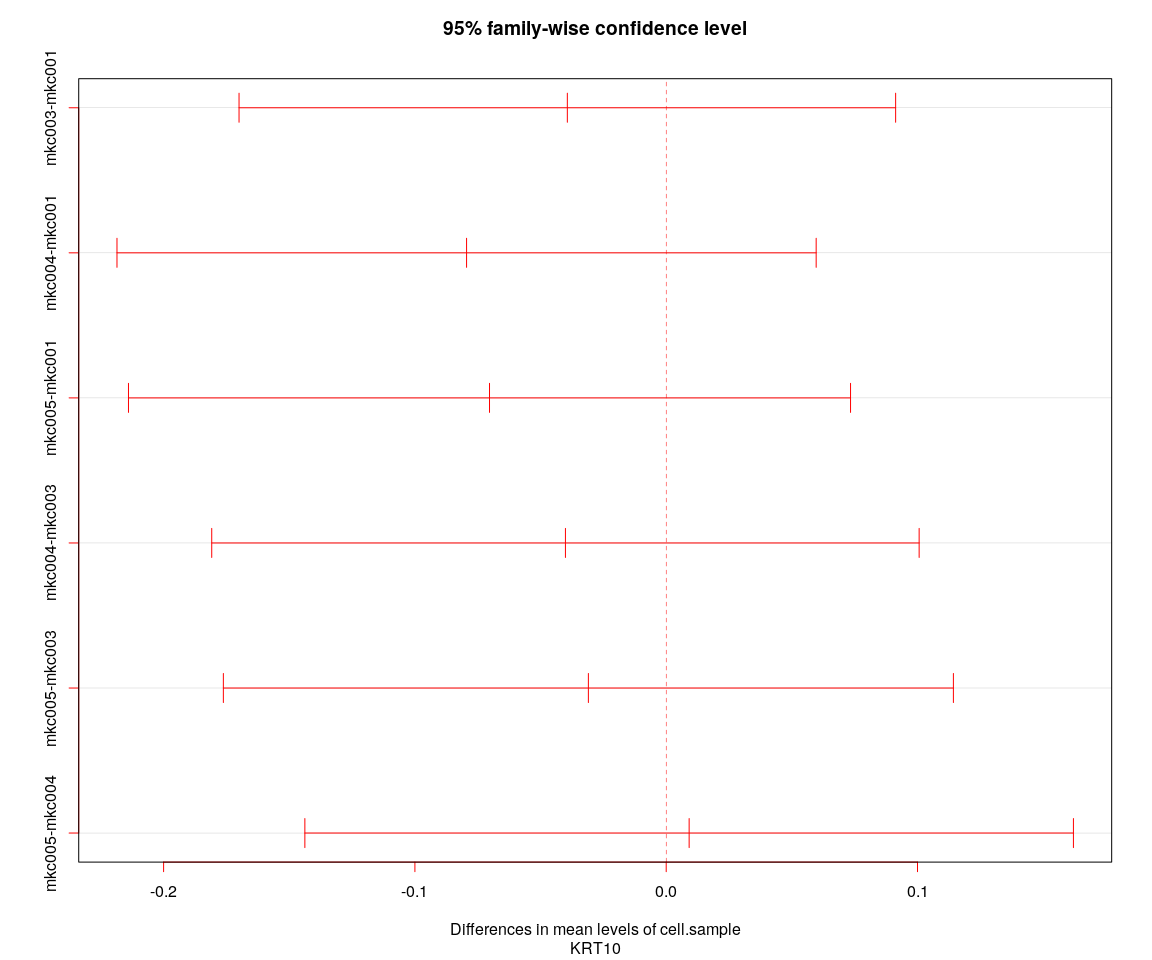
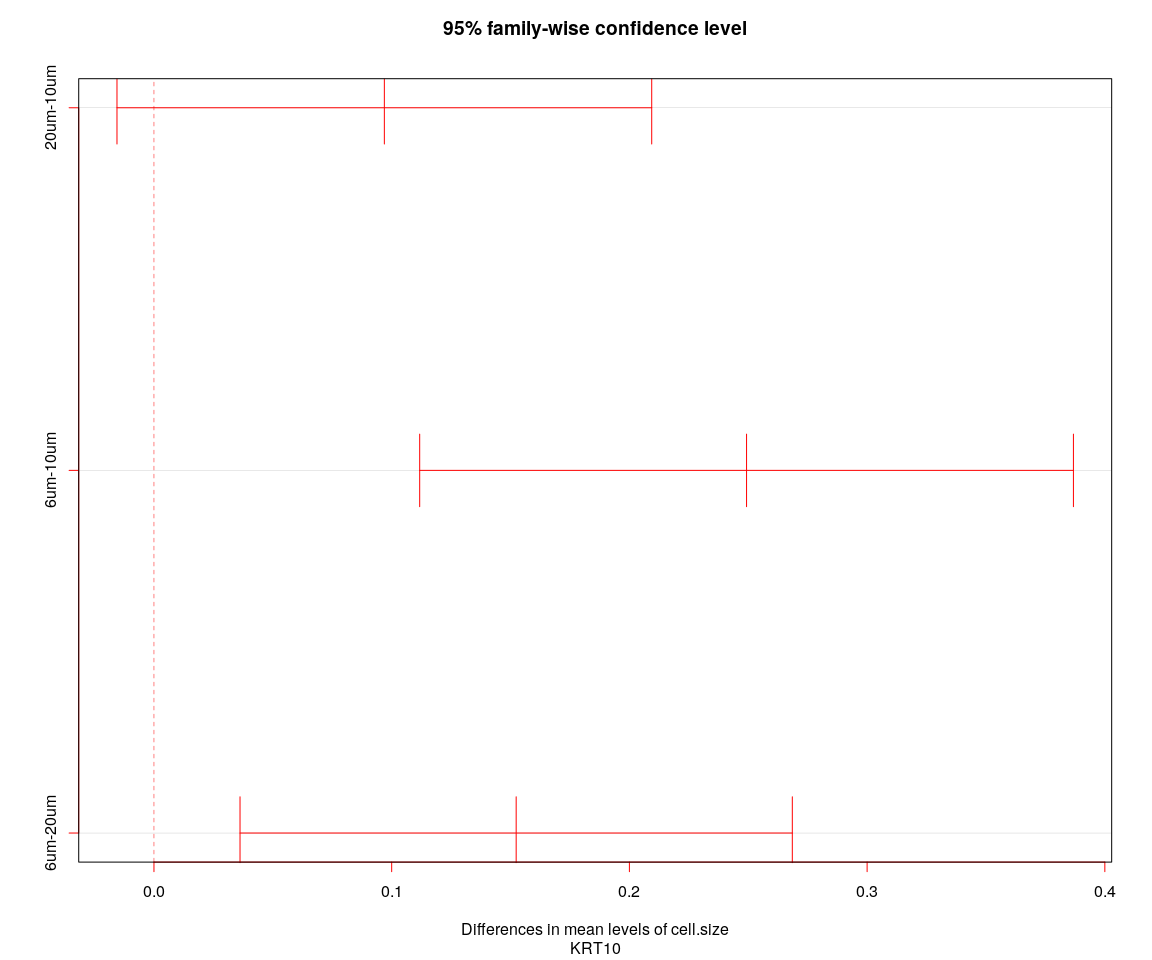
## [1] "----Gene:WNT7A----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.05759956 -0.01489191 0.1300910 0.1491836  
## 6um-10um 0.14774590 0.05912243 0.2363694 0.0002976  
## 6um-20um 0.09014634 0.01527550 0.1650172 0.0133922  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 0.006939123 -0.07726168 0.09113992 0.9966105  
## mkc004-mkc001 0.078897912 -0.01076453 0.16856036 0.1069309  
## mkc005-mkc001 0.024530975 -0.06806529 0.11712724 0.9036168  
## mkc004-mkc003 0.071958789 -0.01875881 0.16267639 0.1731543  
## mkc005-mkc003 0.017591852 -0.07602651 0.11121021 0.9625592  
## mkc005-mkc004 -0.054366937 -0.15292644 0.04419257 0.4862579



## [1] "----Gene:KRT14----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.4184057 0.1053651 0.7314463 0.0050463  
## 6um-10um 0.5786206 0.1959169 0.9613242 0.0012050  
## 6um-20um 0.1602149 -0.1631007 0.4835304 0.4748209  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 0.1274225 -0.2361827 0.4910276778 0.8031013  
## mkc004-mkc001 -0.2603194 -0.6475097 0.1268708474 0.3076320  
## mkc005-mkc001 -0.6862238 -1.0860832 -0.2863644093 0.0000700  
## mkc004-mkc003 -0.3877419 -0.7794887 0.0040048720 0.0535635  
## mkc005-mkc003 -0.8136463 -1.2179195 -0.4093731515 0.0000018  
## mkc005-mkc004 -0.4259044 -0.8515150 -0.0002938533 0.0497667



## [1] "----Gene:TP63----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.21395293 0.05069083 0.3772150 0.0061605  
## 6um-10um 0.23981738 0.04022348 0.4394113 0.0136286  
## 6um-20um 0.02586445 -0.14275637 0.1944853 0.9308649  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 -0.25260451 -0.44223787 -0.06297116 0.0035936  
## mkc004-mkc001 -0.01563896 -0.21757278 0.18629487 0.9971814  
## mkc005-mkc001 -0.09919678 -0.30773803 0.10934446 0.6105520  
## mkc004-mkc003 0.23696556 0.03265535 0.44127576 0.0154729  
## mkc005-mkc003 0.15340773 -0.05743544 0.36425090 0.2399470  
## mkc005-mkc004 -0.08355783 -0.30552923 0.13841357 0.7664438



## [1] "----Gene:KRT10----"  
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = value ~ cell.size + cell.sample, data = genes.data)  
##   
## $cell.size  
## diff lwr upr p adj  
## 20um-10um 0.09691618 -0.01555617 0.2093885 0.1071504  
## 6um-10um 0.24928448 0.11178291 0.3867860 0.0000717  
## 6um-20um 0.15236830 0.03620429 0.2685323 0.0061058  
##   
## $cell.sample  
## diff lwr upr p adj  
## mkc003-mkc001 -0.039341344 -0.1699810 0.09129834 0.8652184  
## mkc004-mkc001 -0.079439464 -0.2185530 0.05967410 0.4553221  
## mkc005-mkc001 -0.070311608 -0.2139771 0.07335385 0.5879490  
## mkc004-mkc003 -0.040098120 -0.1808488 0.10065255 0.8832630  
## mkc005-mkc003 -0.030970264 -0.1762215 0.11428101 0.9466463  
## mkc005-mkc004 0.009127856 -0.1437897 0.16204544 0.9987016