Exploratory and differential expression analyses

Workshop on RNA-Seq

Ashfaq Ali | 24-Nov-2019
NBIS, SciLifeLab

Exploratory and differential expressio

Workshop on RNA-Seq

Ashfaq Ali | 24-Nov-2019
NBIS, SciLifeLab

Contents

- Exploratory
- DGE

Exploratory Analyses

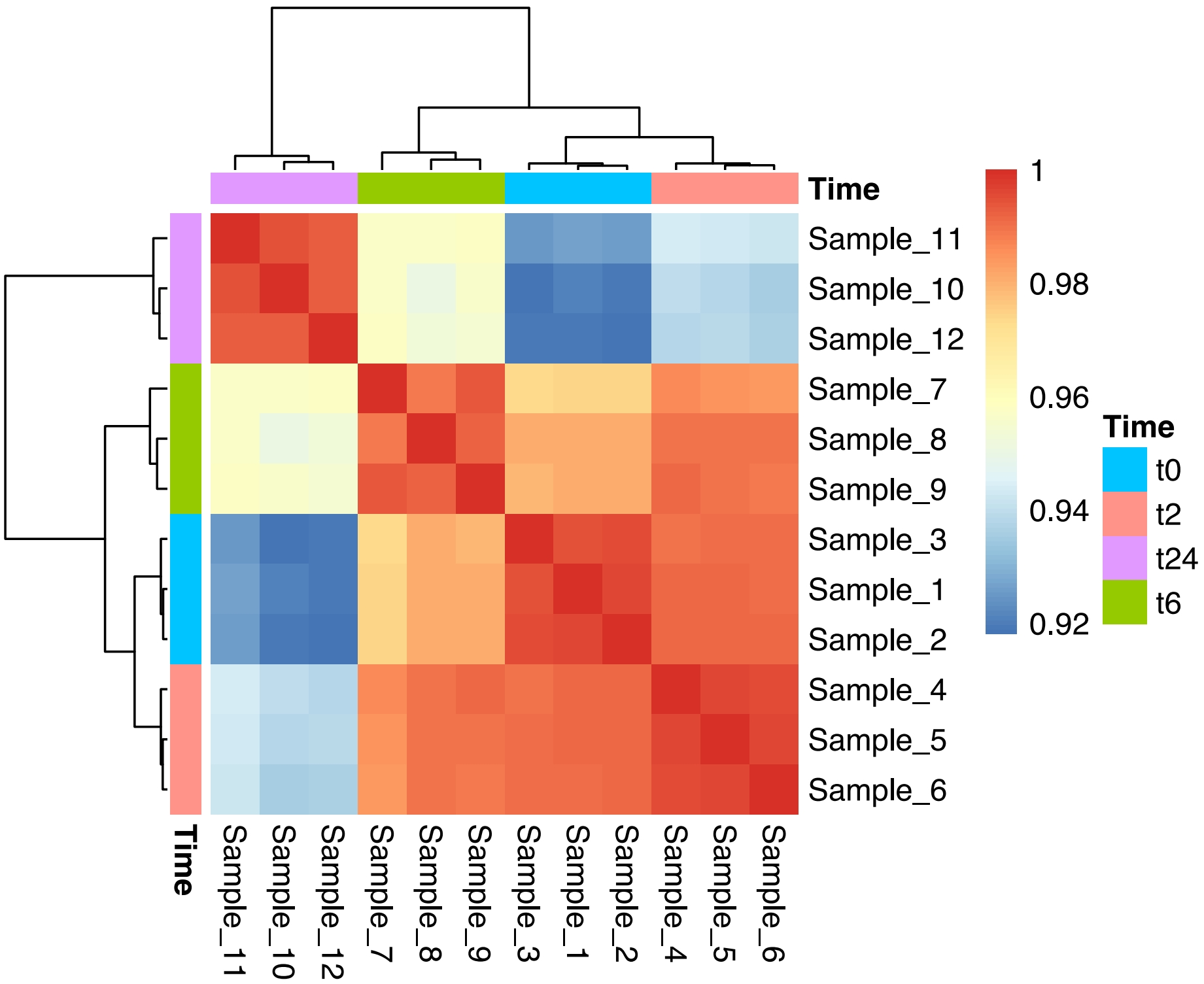
- **Exploratory analyses on samples**
    - Detect outliers using PCA
    - Filter out low quality samples if possible
    - Map meta data on PCA plot to find any confounding factors
    - Sample to sample correlations
- **Exploratory analyses on gene level estimates**
    - Gene counts/library size
    - Check if the treatment groups relate to library size
    - Gene dispersion estimate and visulizations
- **Exploratory analyses on statitics**
    - MA plot of base mean vs. LFC
    - Histogram of p-values
    - Plot counts of individual genes

- Look for warning signs and take action on all levels and use visulizations as m

Exploratory | Correlation

- Correlation between samples

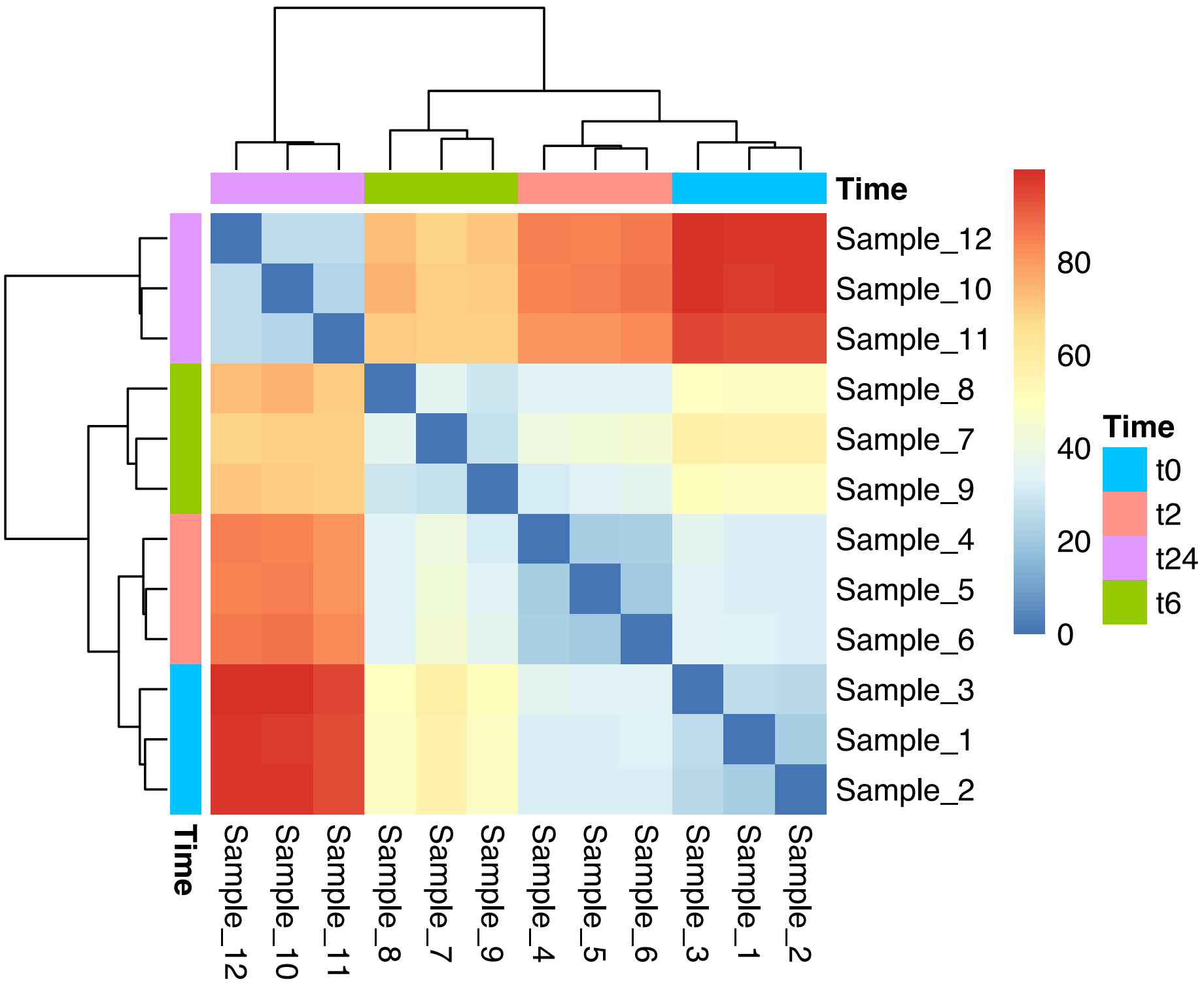
```
dmat <- as.matrix(cor(cv,method="spearman"))
pheatmap::pheatmap(dmat,border_color=NA,annotation_col=mr[,"Time",
         annotation_row=mr[,"Time",drop=F],annotation_legend=T)
```



Exploratory | Distance

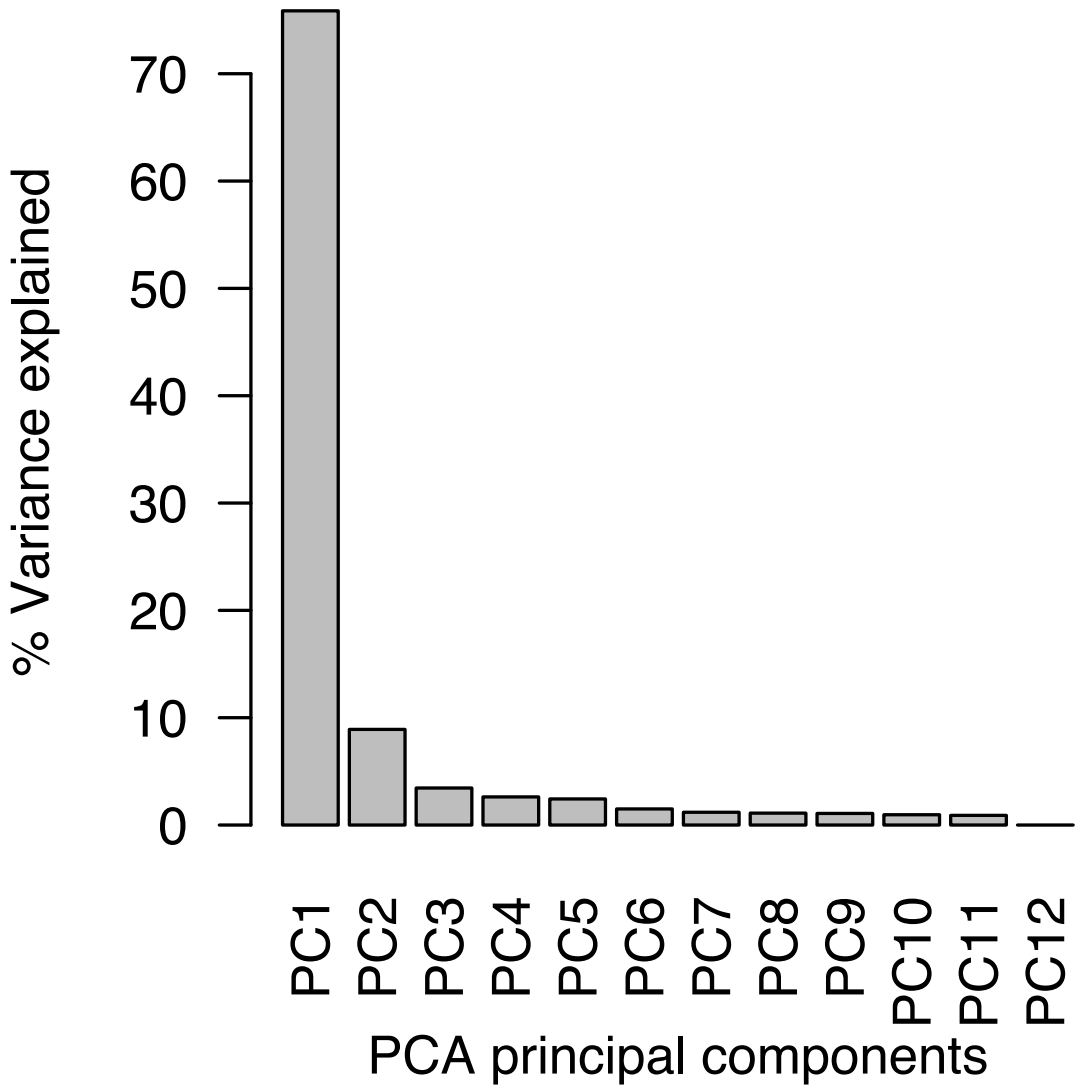
- Similarity between samples

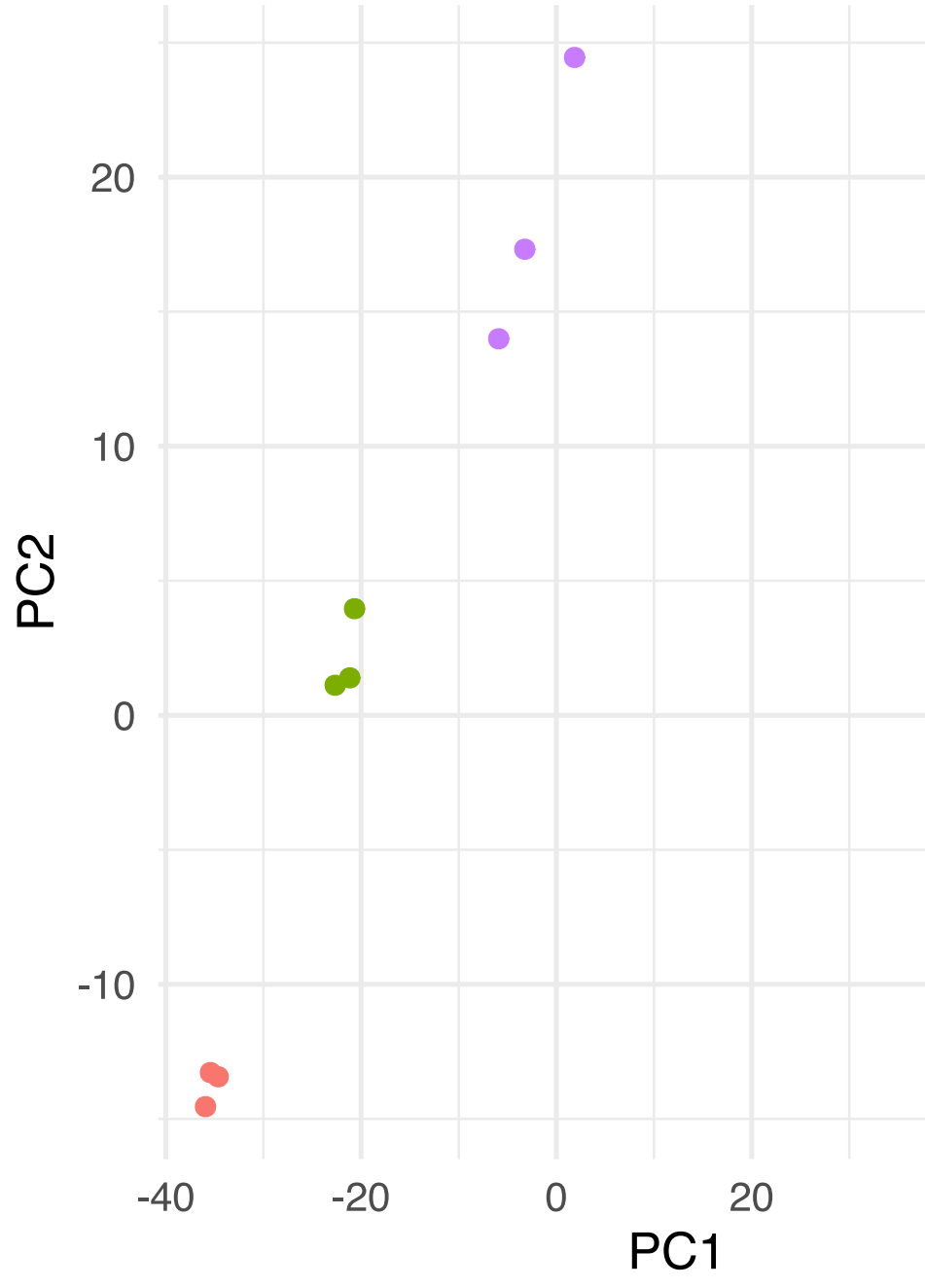
```
dmat <- as.matrix(dist(t(cv)))
pheatmap(dmat,border_color=NA,annotation_col=mr[,"Time",drop=F],
         annotation_row=mr[,"Time",drop=F],annotation_legend=T)
```



Exploratory | PCA

- Relationship between samples





DGE

- Create the DESeq2 object

```
library(DESeq2)
mr$Time <- factor(mr$Time)
d <- DESeqDataSetFromMatrix(countData=cf,colData=mr,design=~Time)
d
```

```
## class: DESeqDataSet
## dim: 14578 12
## metadata(1): version
## assays(1): counts
## rownames(14578): ENSG00000000003 ENSG00000000419 ... ENSG000000
## rowData names(0):
## colnames(12): Sample_1 Sample_2 ... Sample_11 Sample_12
## colData names(5): Sample_ID Sample_Name Time Replicate Cell
```

- Model must be factors

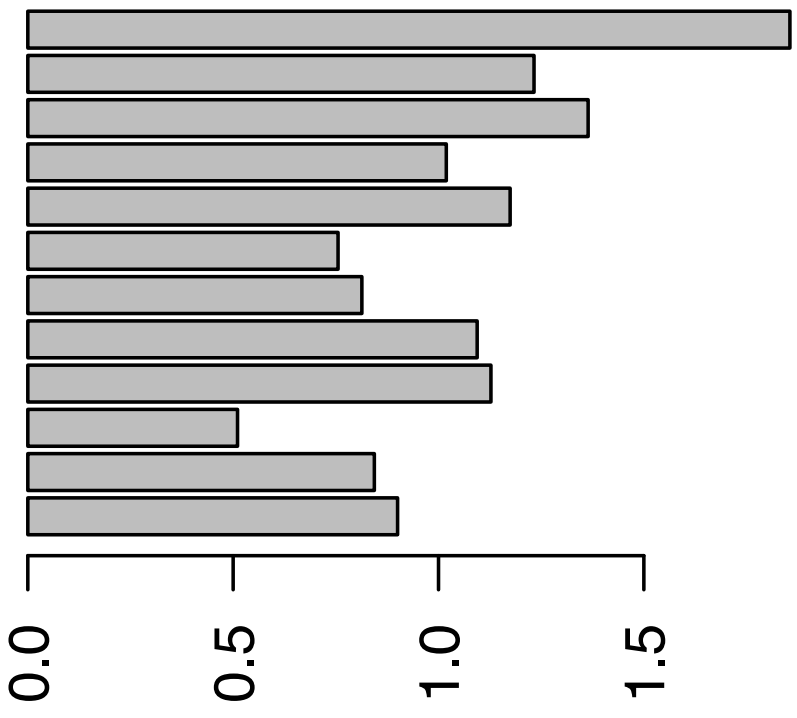
- `~var`
- `~covar+var`

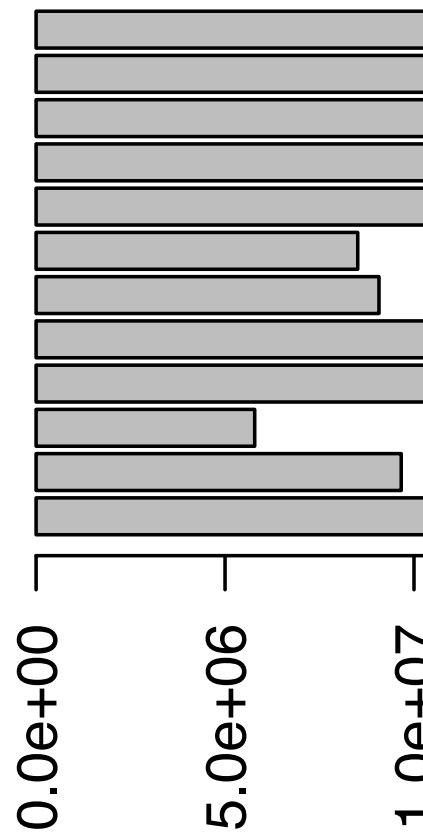
DGE | Size factors

- Normalisation factors are computed

```
d <- DESeq2::estimateSizeFactors(d,type="ratio")
sizeFactors(d)
```

```
##  Sample_1  Sample_2  Sample_3  Sample_4  Sample_5  Sample_6  Sa
## 0.9003753 0.8437393 0.5106445 1.1276451 1.0941383 0.8133849 0.1
## Sample_10 Sample_11 Sample_12
## 1.3642797 1.2325485 1.8555904
```

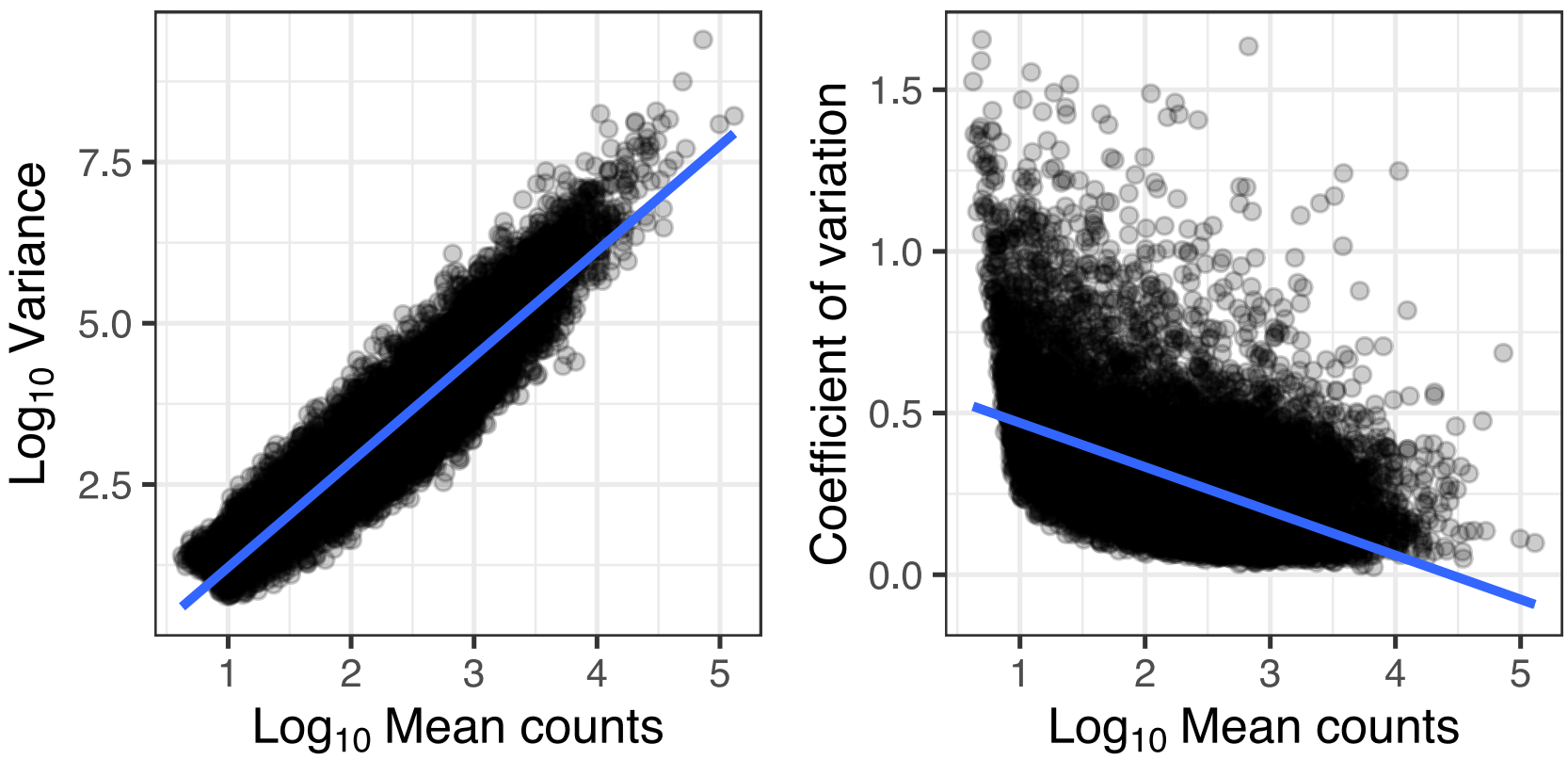




DGE | Dispersion

- We need a measure variability of gene counts

```
dm <- apply(cd,1,mean)
dv <- apply(cd,1,var)
cva <- function(x) sd(x)/mean(x)
dc <- apply(cd,1,cva)
```

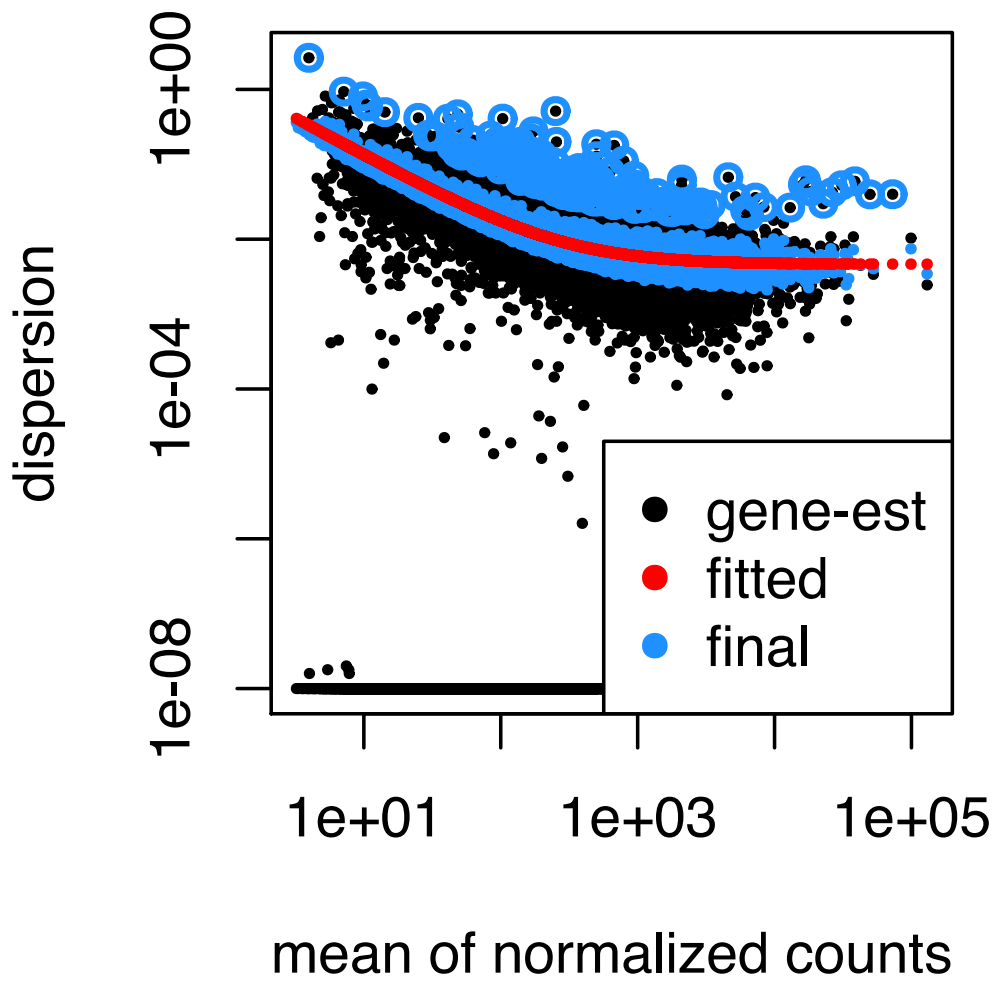


- Dispersion is a measure of variability in gene expression for a given

DGE | Dispersion

- D is unreliable for low mean counts
- Genes with similar mean values must have similar dispersion
- Estimate likely (ML) dispersion for each gene based on counts
- Fit a curve through the gene-wise estimates
- Shrink dispersion towards the curve

```
d <- DESeq2::estimateDispersions(d)
{par(mar=c(4,4,1,1))
plotDispEsts(d)}
```



DGE | Test

- Log2 fold changes changes are computed after GLM fitting

```
dg <- nbinomWaldTest(d)
resultsNames(dg)
```

```
## [1] "Intercept"      "Time_t2_vs_t0"  "Time_t24_vs_t0" "Time_t
```

- Use `results()` to customise/return results
    - Set coefficients using `contrast` or `name`
    - Filtering by fold change using `lfcThreshold`
    - `cooksCutoff` removes outliers
    - `independentFiltering`
    - `pAdjustMethod`
    - `alpha`

DGE | Test

```r
res1 <- results(dg,name="Time_t2_vs_t0",alpha=0.05)
summary(res1)
```

```
##
## out of 14578 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 413, 2.8%
## LFC < 0 (down)     : 696, 4.8%
## outliers [1]       : 0, 0%
## low counts [2]     : 2261, 16%
## (mean count < 26)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

DGE | Test

```
head(res1)
```

```
## log2 fold change (MLE): Time t2 vs t0
## Wald test p-value: Time t2 vs t0
## DataFrame with 6 rows and 6 columns
##                          baseMean     log2FoldChange                 <
##                         <numeric>          <numeric>             <
## ENSG00000000003 490.017213130775  0.220619809383817  0.1127610
## ENSG00000000419 817.780657499214 0.0592719659603648  0.1014812
## ENSG00000000457  82.078767893562  0.207748623173336  0.2204048
## ENSG00000000460  356.07160020304 -0.129186381676168  0.1151391
## ENSG00000001036 919.606750211436 0.0288826917736049  0.0851500
## ENSG00000001084 529.593965301543  0.211964772147083 0.0929810
##                            pvalue                padj
##                         <numeric>           <numeric>
## ENSG00000000003 0.0504034135059863 0.263505451677943
## ENSG00000000419   0.55917459631142 0.830262316253793
## ENSG00000000457   0.34589722283033 0.689945926089258
## ENSG00000000460  0.261861630821574 0.612624613827207
## ENSG00000001036  0.734460942031804 0.909638554349496
## ENSG00000001084 0.0226281312354094 0.159263252815164
```

- Use `lfcShrink()` to correct fold changes for high dispersion genes

DGE | LFC Shrink

```r
# Perform LFC shrinkage.

dg2<- lfcShrink(dg, res = res1, coef ="Time_t2_vs_t0", quiet = TRU
res2 <- dg2[which(dg2$padj<0.05 & abs(dg2$log2FoldChange) >0.58),
```

```r
summary(res2)
```

```
##
## out of 126 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 46, 37%
## LFC < 0 (down)     : 80, 63%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 26)
## [1] see 'cooksCutoff' argument of ?resul
## [2] see 'independentFiltering' argument
```

```r
res11 <- res1[which(
# Order based on ori
summary(res11)
```

```
##
## out of 382 with n
## adjusted p-value
## LFC > 0 (up)
## LFC < 0 (down)
## outliers [1]
## low counts [2]
## (mean count < 26)
## [1] see 'cooksCut
## [2] see 'independ
```

DGE | LFC Shrink

```r
# Order based on shrinked LFC
res2[order(abs(res2$log2FoldChange), decreasing = TRUE),c(1:2)][1:
```

```
## log2 fold change (MAP): Time t2 vs t0
##
## DataFrame with 3 rows and 2 columns
##                          baseMean    log2FoldChange
##                         <numeric>         <numeric>
## ENSG00000120875 712.536673053803 -2.30638870245528
## ENSG00000142627 1572.22058641611 -1.98604476740172
## ENSG00000139289 1300.91800509536 -1.73454657526826
```

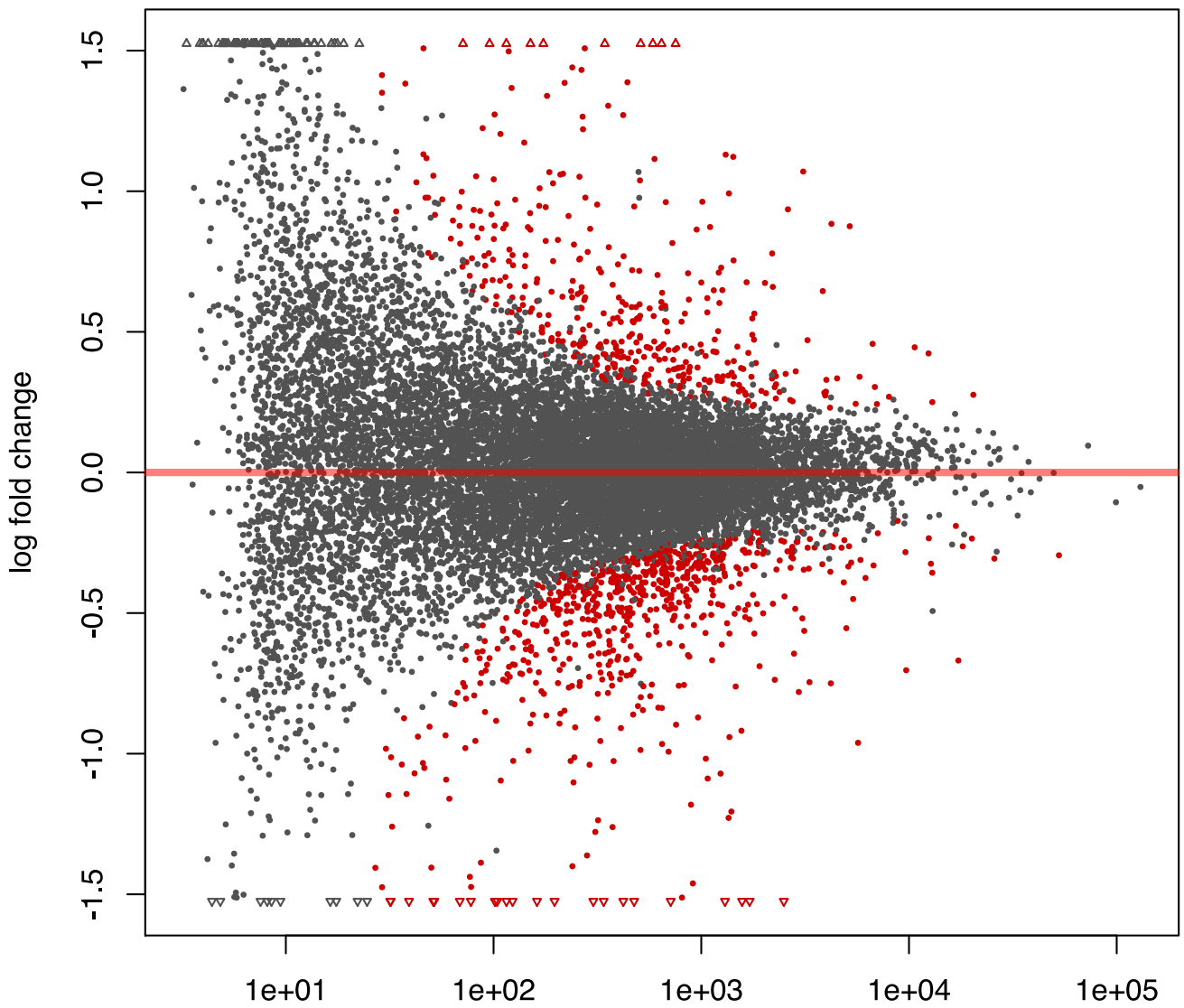
```r
# Order based on original LFC
res11[order(abs(res11$log2FoldChange),decreasing =
                TRUE),c(1:2)][1:3,]
```

```
## log2 fold change (MLE): Time t2 vs t0
##
## DataFrame with 3 rows and 2 columns
##                          baseMean    log2FoldChange
##                         <numeric>         <numeric>
## ENSG00000120875 712.536673053803 -2.76414118057782
## ENSG00000142627 1572.22058641611 -2.38780612309795
## ENSG00000139318 101.530389195382 -2.38675628664027
```

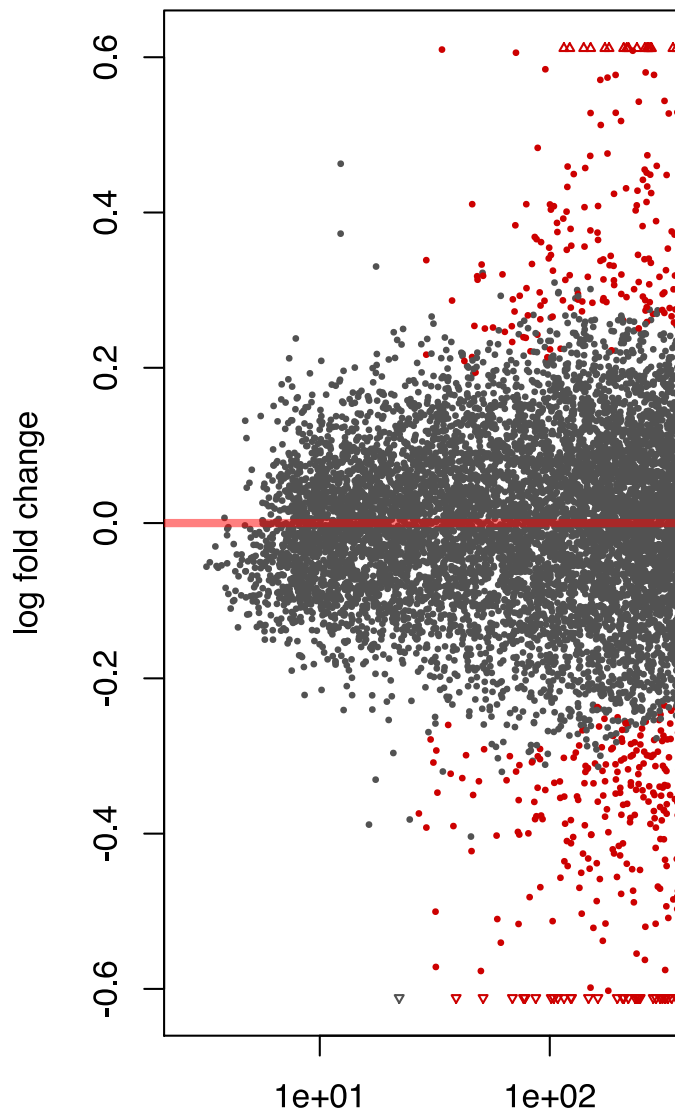
The ranks of top DE genes changes after `lfcSrink`, note the **base means.**

DGE | PlotMA

```r
# Plot non shrinked estimates
DESeq2::plotMA(res1)
```




```r
# Plot shrinked  LFC
DESeq2::plotMA(dg2)
```



Thank you. Questions?

R version 3.6.1 (2019-07-05)

Platform: x86_64-apple-darwin15.6.0 (64-bit)

OS: macOS Catalina 10.15.1

Built on : 24-Nov-2019 at 19:44:10