Gene Expression Analysis Using R - Part I

Case Study: Interferon regulatory factor 6 (*IRF6*)

October 14, 2019

Contents

0.1 Today's Objectives

- Download and load microarray dataset into R
- Explore the dataset with basic visualizations
- Identify differentially expressed genes (DEGs)
- Generate annotation of the DEGs

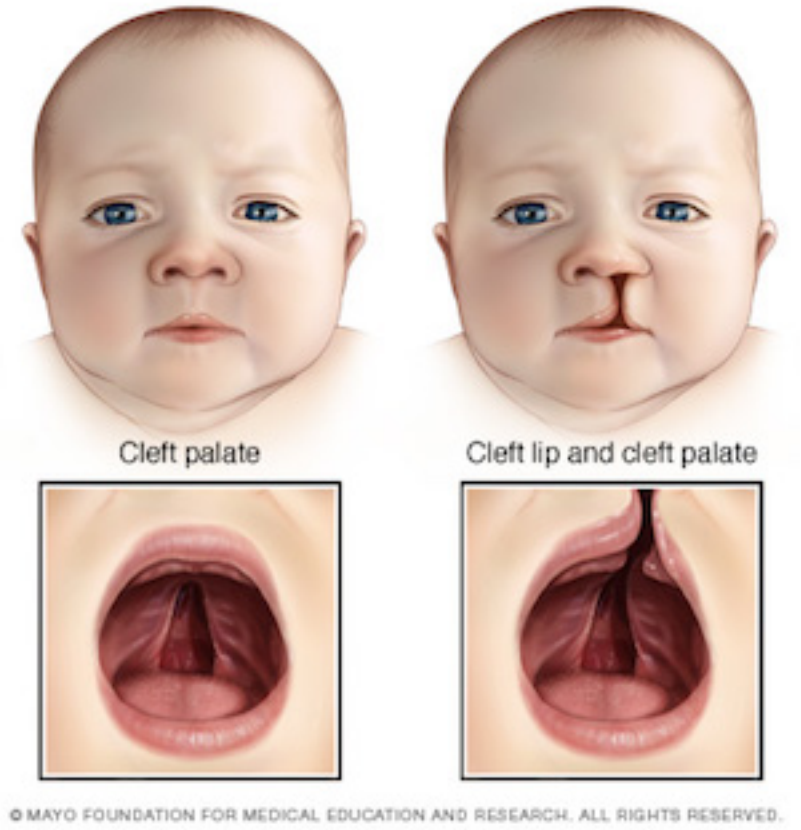


Figure 1: Cleft lip and palate

0.2 Cleft Lip and Palate 1/3

Cleft lip and cleft palate (**CLP**) are splits in the upper lip, the roof of the mouth (palate) or both. They result when facial structures that are developing in an unborn baby do not close completely. CLP is one of the most common birth defects with a frequency of 1/700 live births.

0.3 Cleft Lip and Palate 2/3

Children with cleft lip with or without cleft palate face a variety of challenges, depending on the type and severity of the cleft.

- **Difficulty feeding.** One of the most immediate concerns after birth is feeding.
- **Ear infections and hearing loss.** Babies with cleft palate are especially at risk of developing middle ear fluid and hearing loss.
- **Dental problems.** If the cleft extends through the upper gum, tooth development may be affected.
- **Speech difficulties.** Because the palate is used in forming sounds, the development of normal speech can be affected by a cleft palate. Speech may sound too nasal.

Reference: Mayo Foundation for Medical Education and Research

0.4 Cleft Lip and Palate 3/3

- DNA variation in Interferon Regulatory Factor 6 (**IRF6**) causes Van der Woude syndrome (**VWS**)
- VWS is the most common syndromic form of cleft lip and palate.
- However, the causing variant in IRF6 has been found in *only* 70% of VWS families!
- IRF6 is a **transcription factor** with a conserved helix-loop-helix DNA binding domain and a less well-conserved protein binding domain.

Reference: Hum Mol Genet. 2014 May 15; 23(10): 2711–2720

0.5 Question

Given:

1. The pathogenic variant in IRF6 exists in only 70% of the VWS families
2. IRF6 is a transcription factor

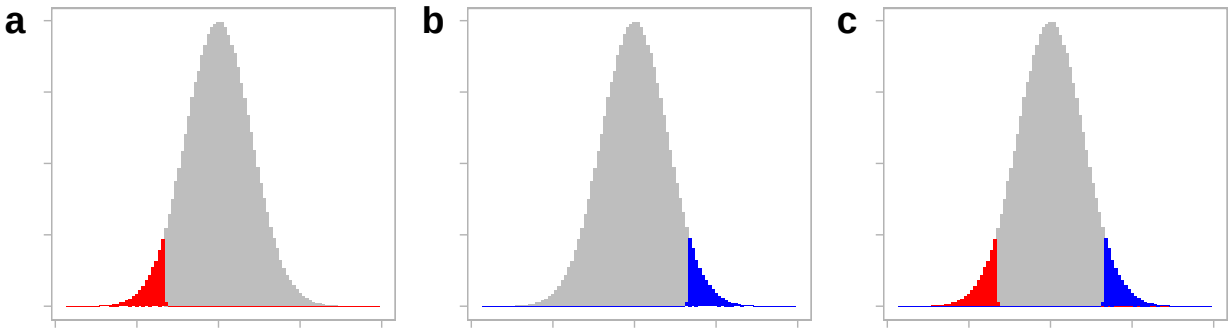
How can we identify other genes that might be involved in the remaining 30% of the VWS families?

0.6 Hint

- Usually, genes that are regulated by a transcription factor belong to the same biological process or pathway.
- Therefore, by comparing the gene expression patterns between wild-type (functional) *Irf6* and knockout (non-functional) *Irf6*, it could be possible to identify genes that are regulated (targeted) by *Irf6*.

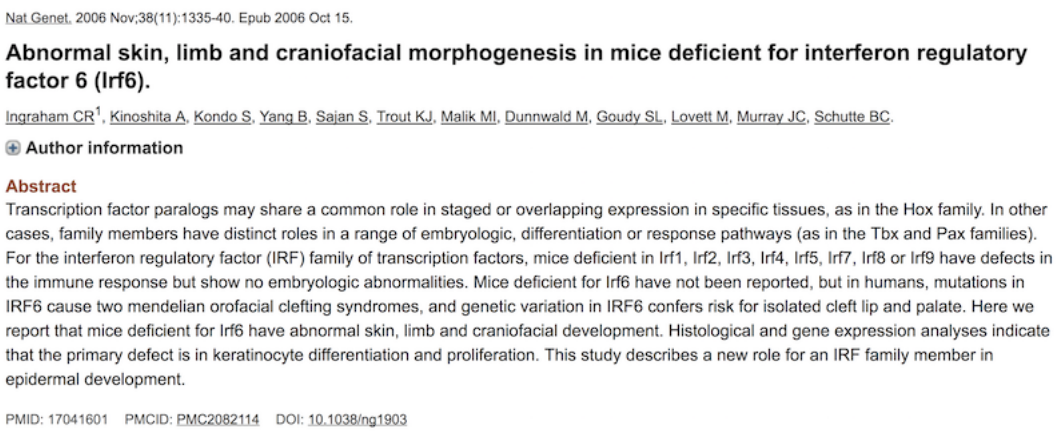
0.7 Hypothesis

- $H_O : \mu_{WT} = \mu_{KO}$
- $H_A : \mu_{WT} \neq \mu_{KO}$
- Where μ is the *mean* of the gene expression values of a gene.
- **One**-sided or **Two**-sided testing?



0.8 Why Microarray?

- It does not require a predefinied set of candidate genes



Nat Genet. 2006 Nov;38(11):1335-40. Epub 2006 Oct 15.

Abnormal skin, limb and craniofacial morphogenesis in mice deficient for interferon regulatory factor 6 (Irf6).

Ingraham CR[1], Kinoshita A, Kondo S, Yang B, Sajan S, Trout KJ, Malik MI, Dunnwald M, Goudy SL, Lovett M, Murray JC, Schutte BC.

Author information

Abstract

Transcription factor paralogs may share a common role in staged or overlapping expression in specific tissues, as in the Hox family. In other cases, family members have distinct roles in a range of embryologic, differentiation or response pathways (as in the Tbx and Pax families). For the interferon regulatory factor (IRF) family of transcription factors, mice deficient in Irf1, Irf2, Irf3, Irf4, Irf5, Irf7, Irf8 or Irf9 have defects in the immune response but show no embryologic abnormalities. Mice deficient for Irf6 have not been reported, but in humans, mutations in IRF6 cause two mendelian orofacial clefting syndromes, and genetic variation in IRF6 confers risk for isolated cleft lip and palate. Here we report that mice deficient for Irf6 have abnormal skin, limb and craniofacial development. Histological and gene expression analyses indicate that the primary defect is in keratinocyte differentiation and proliferation. This study describes a new role for an IRF family member in epidermal development.

PMID: 17041601 PMCID: PMC2082114 DOI: 10.1038/ng1903

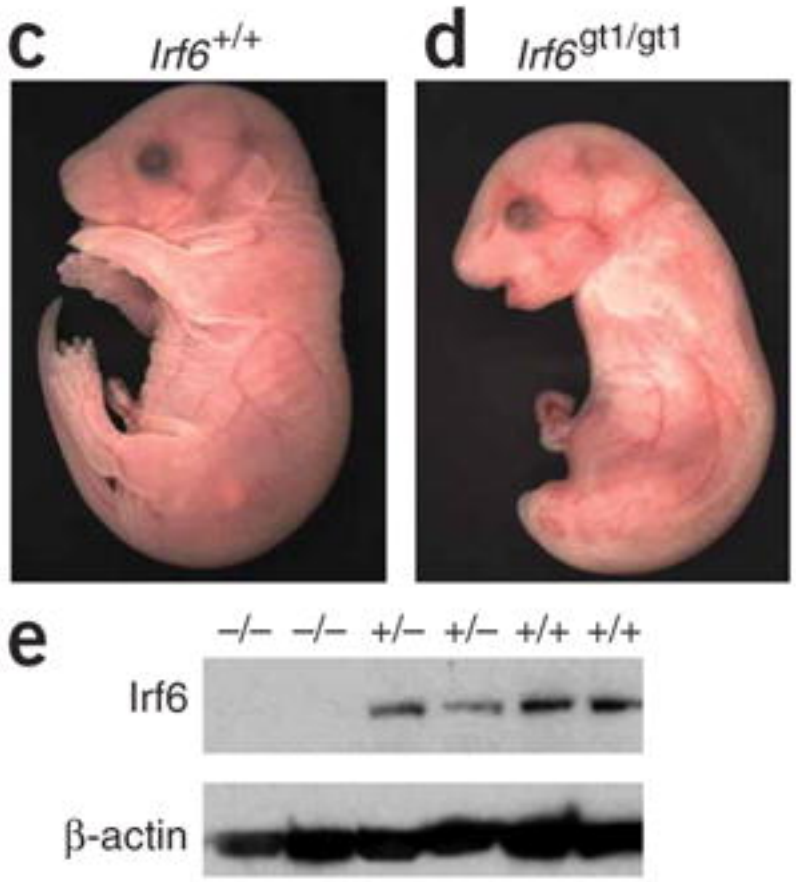
Figure 2: PMID: 17041601

- It requires only a small amount of RNA
- It is a high-throughput protocol - transcriptome-wide analysis
- One experiment can generate multiple hypotheses

0.9 Original Paper

0.10 Experimental Design

- 3 IRF6 wild-type (+/+) and 3 knockout (-/-) mouse embryos.
- E17.5 embryos were removed from euthanized mothers.
- Skin was removed from embryos.
- Total RNA was isolated from the skin.
- Resultant RNA was hybridized to Affymetrix GeneChip Mouse Genome 430 2.0 arrays.



0.11 Dataset

- The original dataset can be obtained from NCBI GEO with accession GSE5800

ID	KO1	KO2	KO3	WT1	WT2	WT3
1415670_at	6531.0	5562.8	6822.4	7732.1	7191.2	7551.9
1415671_at	11486.3	10542.7	10641.4	10408.2	9484.5	7650.2
1415672_at	14339.2	13526.1	14444.7	12936.6	13841.7	13285.7
1415673_at	3156.8	2219.5	3264.4	2374.2	2201.8	2525.3

- Download the dataset from the following link **https://goo.gl/gH7QLM**

0.12 Loading

We are going to load the dataset from a tsv file (`Irf6.tsv`) into a variable called `data` using function `read.table`. `data` here is just an arbitrary **varilable** name to hold the result of `read.table` and it can be called/named *almost* anything. See The State of Naming Conventions in R (Bååth 2012) for more information on naming **varilables** in `R`.

```
# Load data from text file into a varilable
data = as.matrix(read.table("Irf6.tsv", header = TRUE, row.names = 1))
```

Note: the hash sign (`#`) indicates that what comes after is a *comment*. Comments are for documentation and readability of the `R` code and they are not evaluated (or executed).

0.13 Checking

```
dim(data) # Dimension of the dataset
```

```
## [1] 45101     6
```

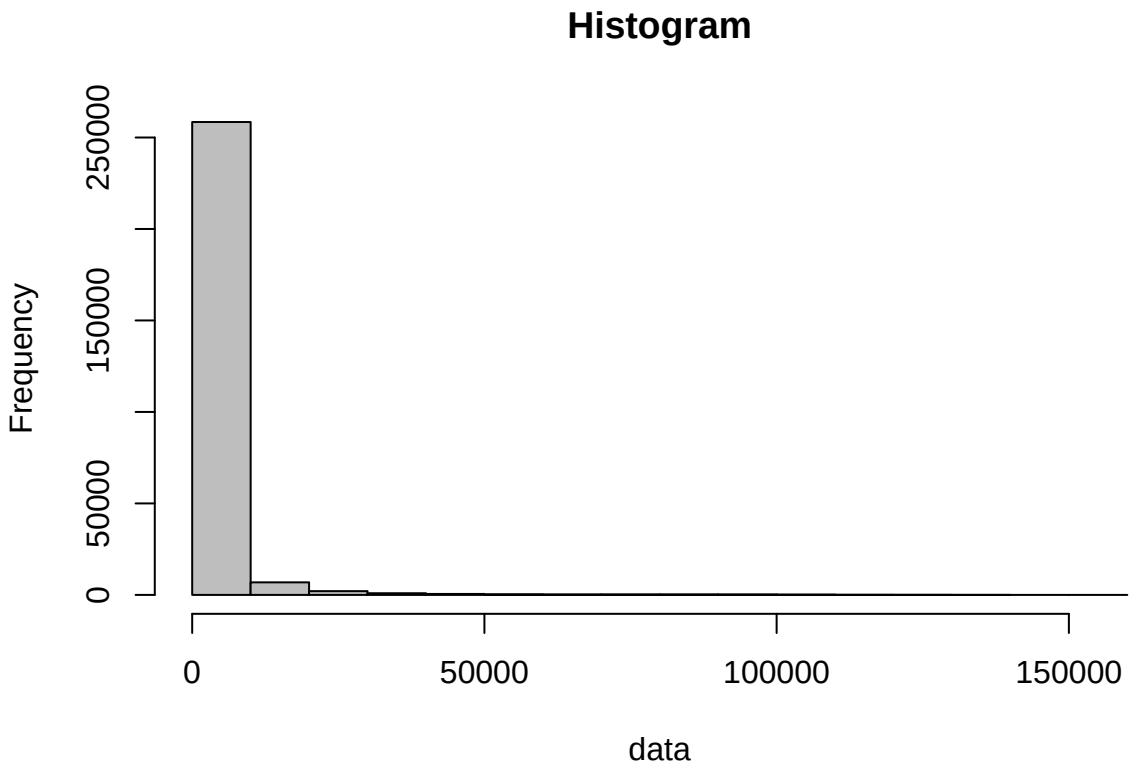
```
head(data) # First few rows
```

	KO1	KO2	KO3	WT1	WT2	WT3
1415670_at	6531.0	5562.8	6822.4	7732.1	7191.2	7551.9
1415671_at	11486.3	10542.7	10641.4	10408.2	9484.5	7650.2
1415672_at	14339.2	13526.1	14444.7	12936.6	13841.7	13285.7
1415673_at	3156.8	2219.5	3264.4	2374.2	2201.8	2525.3
1415674_a_at	4002.0	3306.9	3777.0	3760.6	3137.0	2911.5
1415675_at	3468.4	3347.4	3332.9	3073.5	3046.0	2914.4

0.14 Exploring

Check the behavior of the data (e.g., normal?, skewed?)

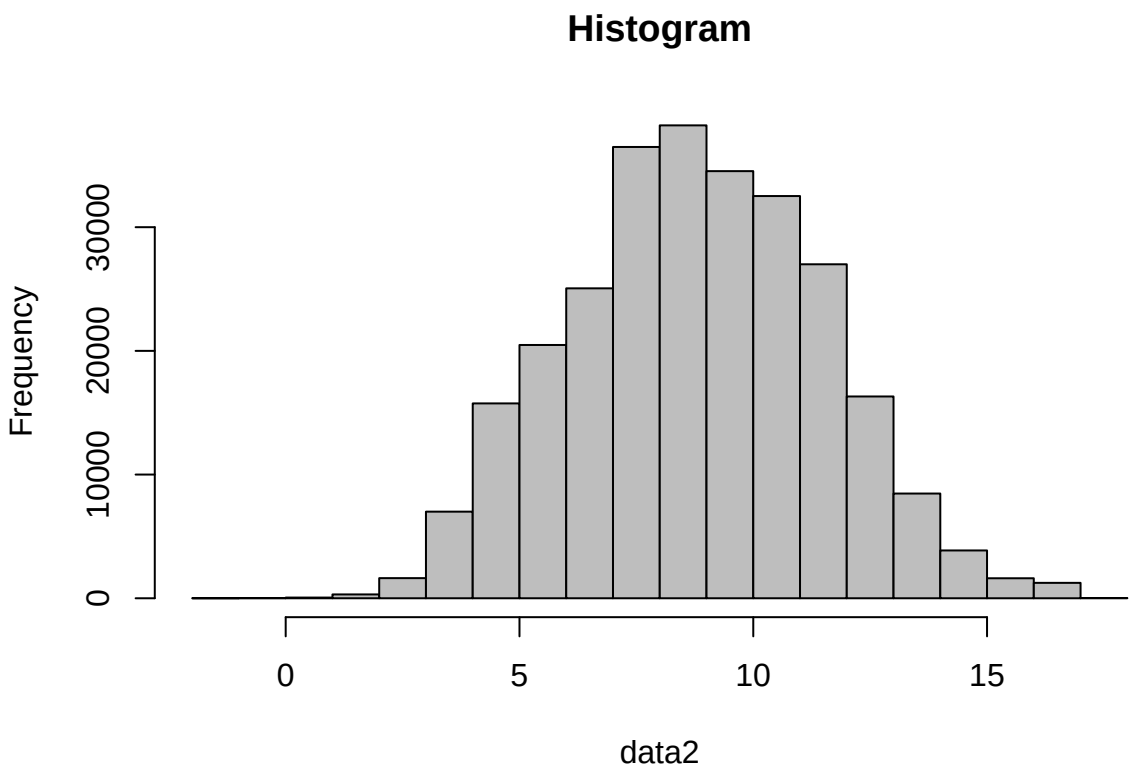
```
hist(data, col = "gray", main="Histogram")
```



0.15 Transforming

log_2 transformation (why?)

```
data2 = log2(data)
hist(data2, col = "gray", main="Histogram")
```



0.16 Multiple Plots 1/2

```
samples = colnames(data2) # Headers (names) of the columns
samples
```

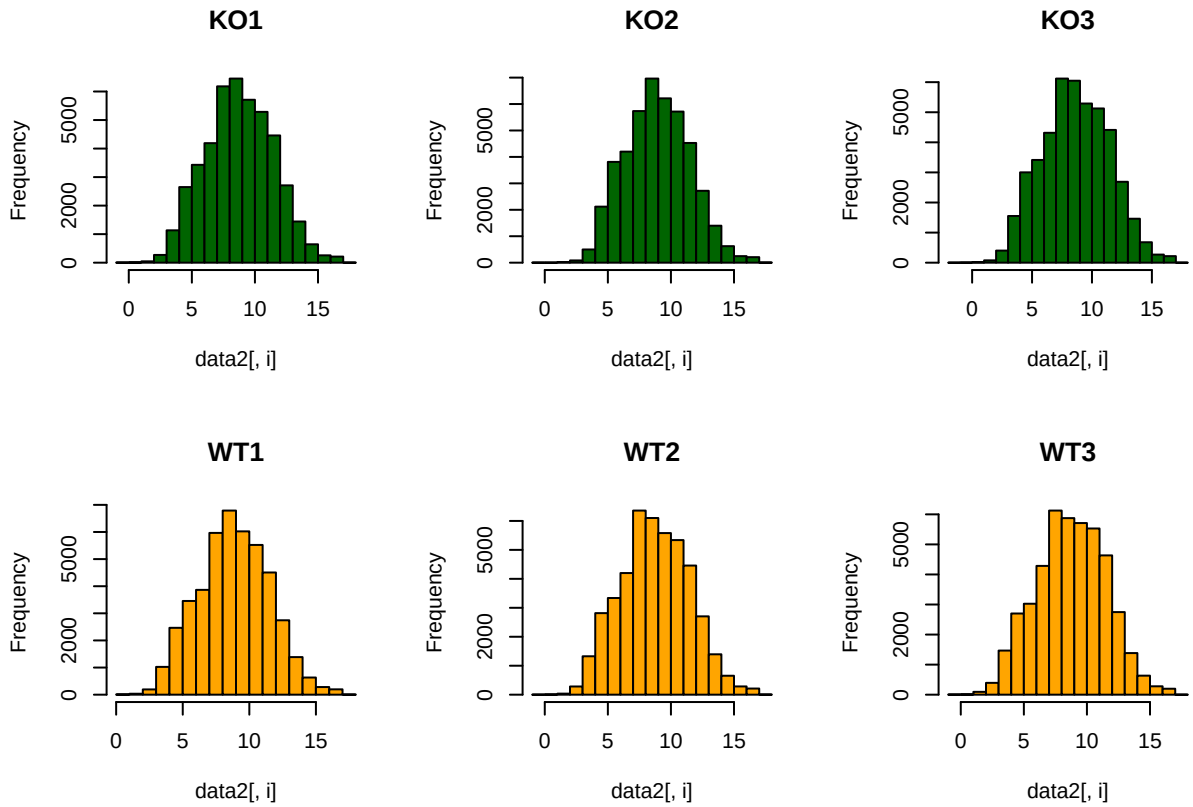
```
## [1] "KO1" "KO2" "KO3" "WT1" "WT2" "WT3"
```

```
par( mfrow = c( 2, 3 ) ) # Split screen into 2 rows x 3 columns paritions

for (i in 1:3) {
  # for each of the first 3 columns in the table
  hist(data2[,i], col = "red", main = samples[i])
}

for (i in 4:6) {
  # for each of the last 3 columns in the table
  hist(data2[,i], col = "green", main = samples[i])
}
```

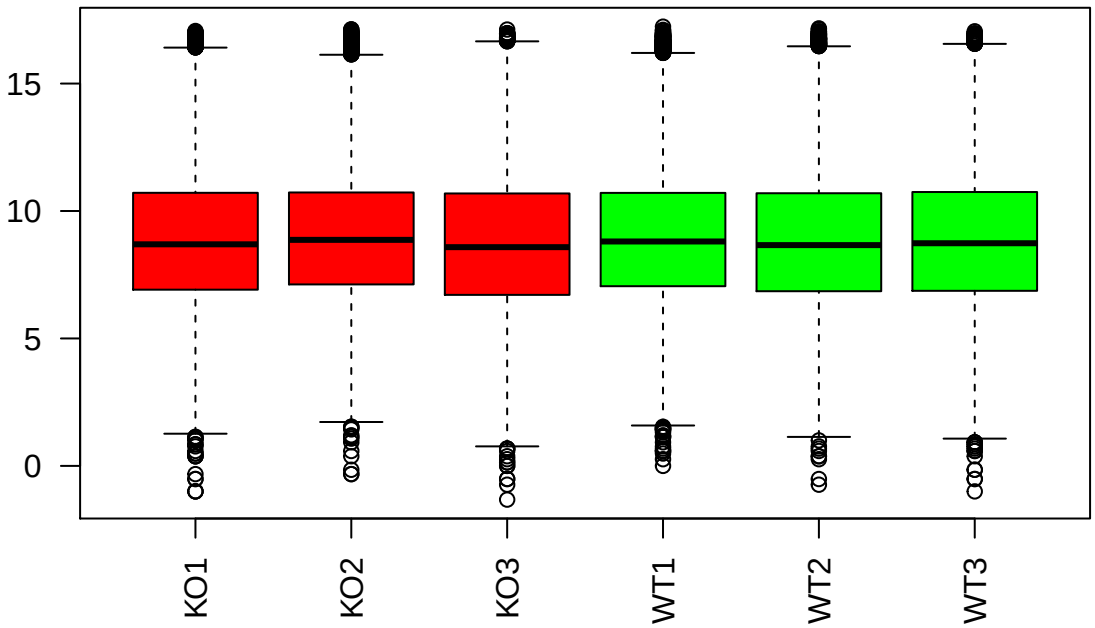
0.17 Multiple Plots 2/2




```
par( mfrow = c( 1, 1 ) ) # Just to set screen back to 1 partition
```

0.18 Boxplot

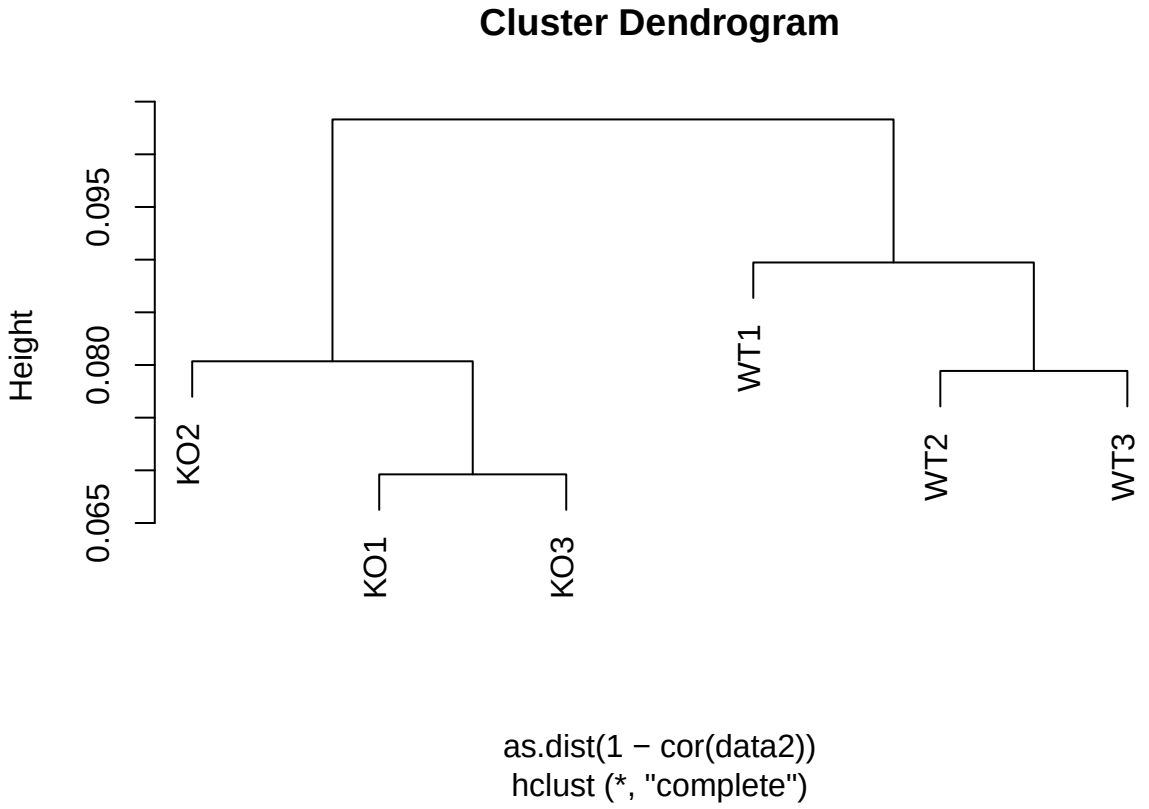
```
colors = c(rep("red", 3), rep("green", 3))
boxplot(data2, col = colors, las = 2)
```



0.19 Clustering 1/2

Hierarchical clustering of the **samples** (i.e., columns) based on the correlation coefficients of the expression values

```
hc = hclust(as.dist(1 - cor(data2)))
plot(hc)
```



0.20 Clustering 2/2

To learn about a function (e.g., `hclust`), you may type `?function` (e.g., `?hclust`) in the `console` to launch `R` documentation on that function:

0.21 Comparing

We are going to compare the **means** of the replicates of the two conditions

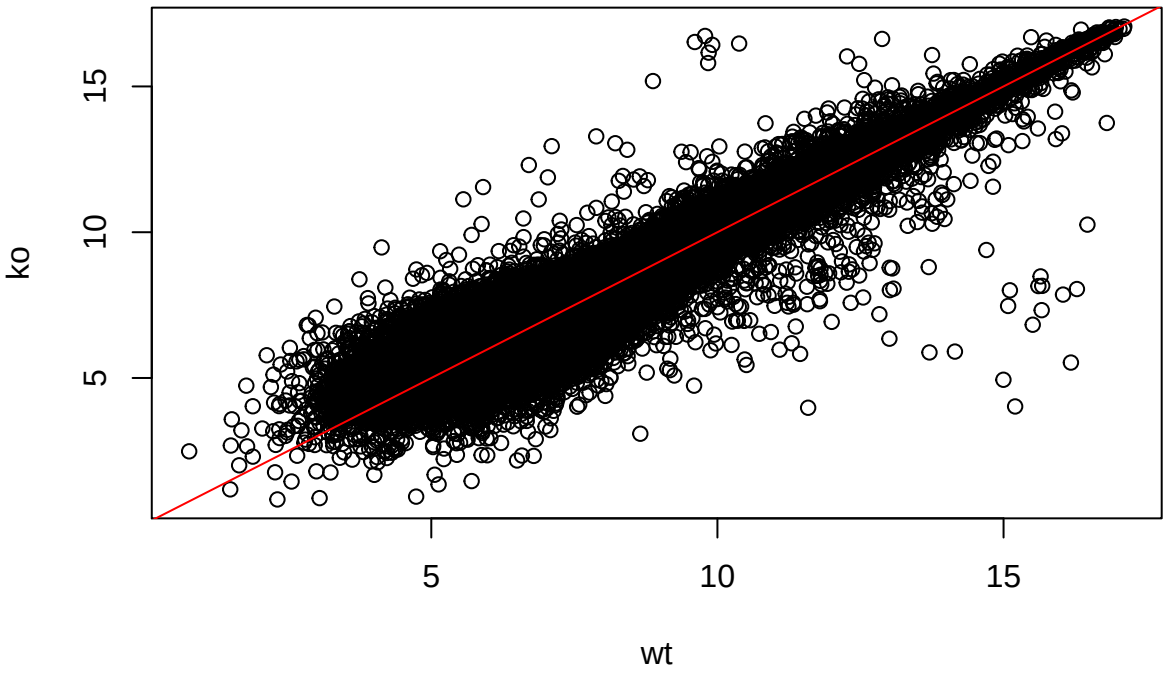
```
# Compute the means of the samples of each condition
ko = apply(data2[, 1:3], 1, mean)
head(ko)
```

```
##   1415670_at   1415671_at   1415672_at   1415673_at 1415674_a_at
##     12.61692     13.40966     13.78313     11.47096     11.84693
##   1415675_at
##     11.72381
```

```
wt = apply(data2[, 4:6], 1, mean)
head(wt)
```

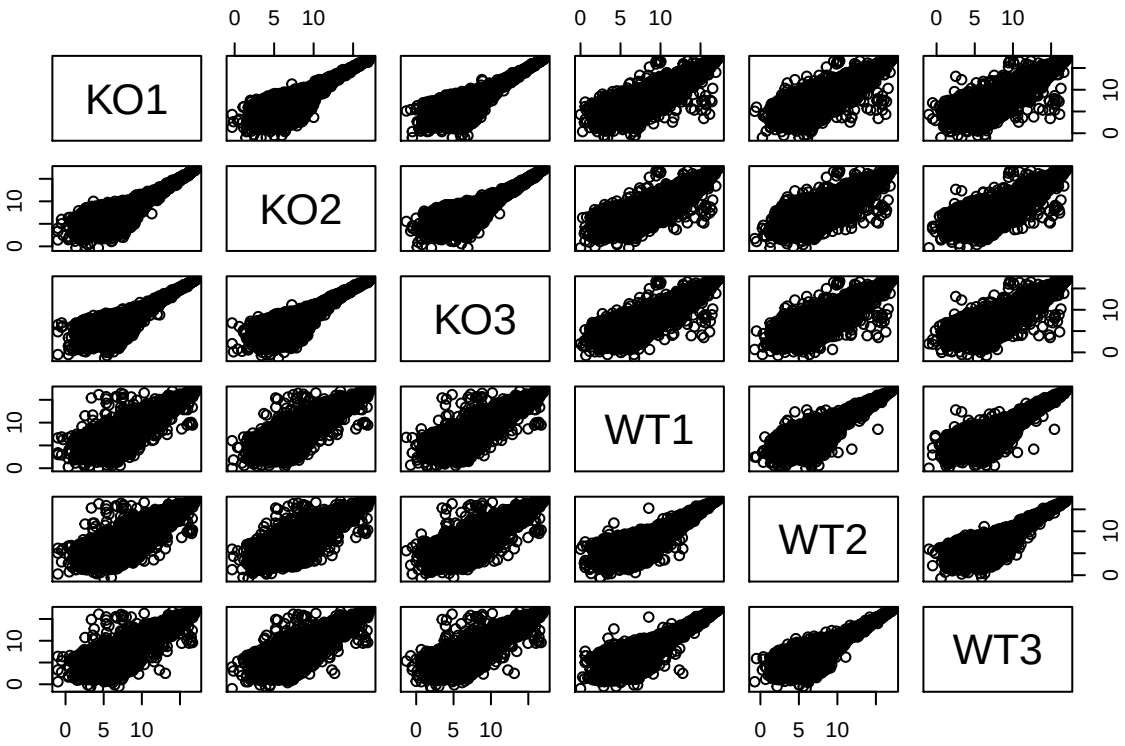
```
##   1415670_at   1415671_at   1415672_at   1415673_at 1415674_a_at
##     12.87043     13.15269     13.70450     11.20664     11.66649
##   1415675_at
##     11.55578
```

0.22 Scatter 1/2



0.23 Scatter 2/2

```
pairs(data2) # All pairwise comparisons
```



0.24 Differentially Expressed Genes (DEGs)

To identify DEGs, we will identify:

- **Biologically** significantly differentially expressed
- **Statistically** significantly differentially expressed

Then, we will take the **overlap** (**intersection**) of the two sets

0.25 Biological Significance (fold-change) 1/2

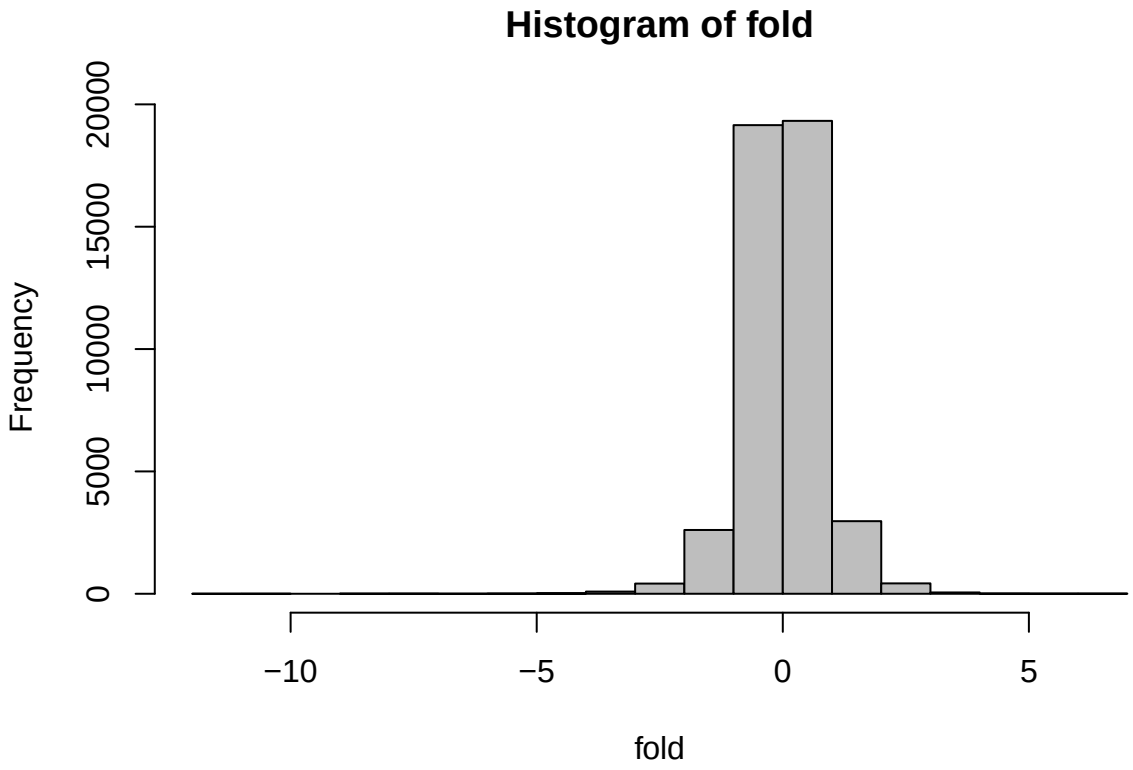
```
fold = ko - wt # Difference between means
head(fold)
```

```
##   1415670_at   1415671_at   1415672_at   1415673_at 1415674_a_at
##  -0.25351267   0.25697097   0.07863227   0.26431191   0.18044345
##   1415675_at
##   0.16803065
```

- What do the positive and negative values of the fold-change indicate? Considering the `WT` condition is the **reference** (or **control**)
  - **+ve** → **Up**-regulation ↑
  - **-ve** → **Down**-regulation ↓

0.26 Biological Significance (fold-change) 2/2

```
hist(fold, col = "gray") # Histogram of the fold
```



0.27 Homework

- Identify the top 10 *biologically* significant genes (i.e., by fold-change)