# Design of RNA-Seq and Result Interpretation

Sanzhen Liu
Department of Plant Pathology
Kansas State University
@K-State IGF RNA-Seq Workshop 2014

6/6/2014

# Outline

**Design of DE experiments and results**

- Experimental design
- P-values and q-values

**Other analyses**

- Visualization
- GO term enrichment analysis

# An RNA-Seq experiment


Control 1 TRT 1 (salt)

RNA
Library
Sequencing

N1 reads N2 reads

Trimming
Alignment
Read counting per gene

| Gene | Control 1 | TRT 1 |
|---|---|---|
| 1 | C1 | T1 |
| 2 | C2 | T2 |
| … | | |
| Sum | N1’ | N2’ |

Q: If the gene expression of Gene 1 is associated with the treatment?

- Fisher’s Exact Test on Gene 1

2x2 Table for Gene 1

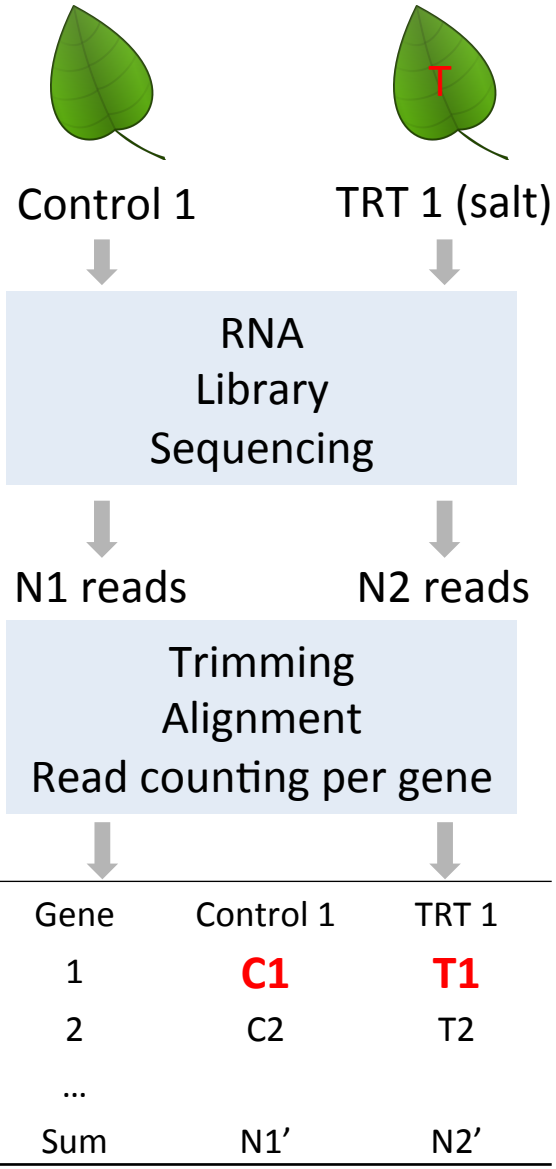
| | Gene 1 | Others |
|---|---|---|
| Control 1 | C1 | N1’ – C1 |
| TRT 1 | T1 | N2’ – T1 |

A p-value for Gene 1

- Repeat on all the genes then perform multiple test correction
- p-values
- q-values

Setup the FDR cutoff and declare that the genes are significant if q-values < FDR

# An RNA-Seq experiment – source of variance

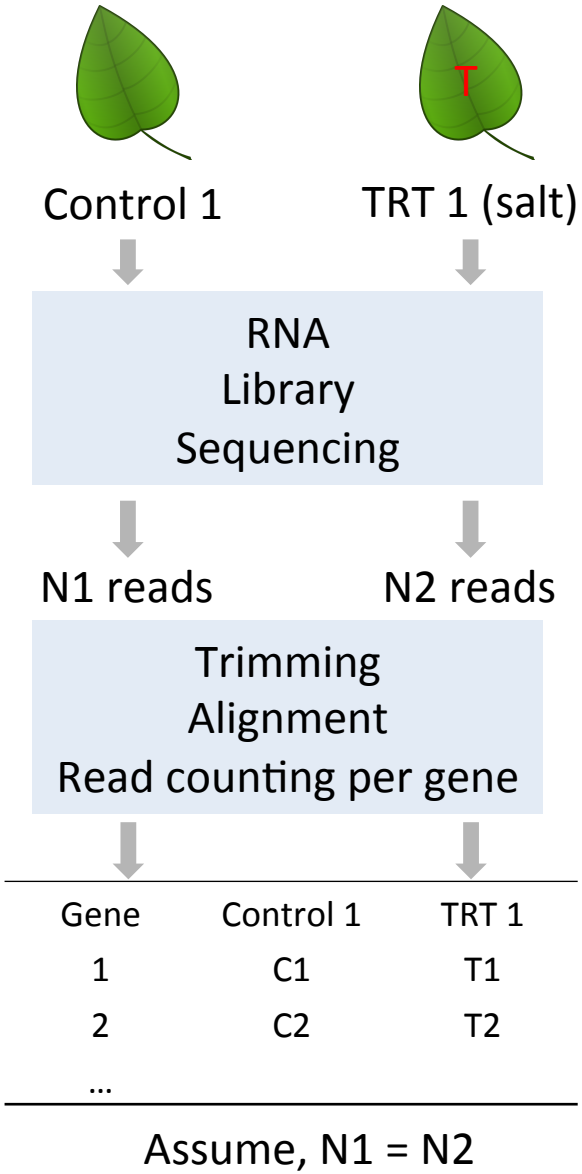



We are interested in compare the proportions of read counts of each gene out of the total reads in Control 1 and TRT 1

For example,
C1 out of N1' vs. T1 out of N2'

Let us not worry about the experimental design. First, can you think about what would cause different proportions of read counts of Gene 1 out of the total reads in Control 1 and TRT 1?

# An RNA-Seq experiment – source of variance




- **Treatment effect**
- Difference between two plants

Biological variance

- RNA quality
- Library preparation
- Sequencing

Technical variance

- Randomly sampling

Sampling variance

# Source of variance in RNA-Seq - sampling

- **Sampling variance** derived from the inherent nature of counting experiments

1,000 molecules of Gene 1 in $10^9$ total molecules

Randomly sample $10^7$

| | |
|---|---|
| First sampling | 6 |
| Second sampling | 13 |
| Third sampling | 8 |

Randomly sample $10^8$

| | |
|---|---|
| First sampling | 107 |
| Second sampling | 93 |
| Third sampling | 97 |

Sequence depth (sampling number) matters.

# Source of variance in RNA-Seq

- **Biological variance**

Biological variance include **biological variance of interest (related to treatment)** and **biological variance of no-interest**

- **Technical variance**
- **Random sampling variance** stemming from the inherent nature of counting experiments

# Fisher's Exact Test (FET)


Control 1 TRT 1 (salt)

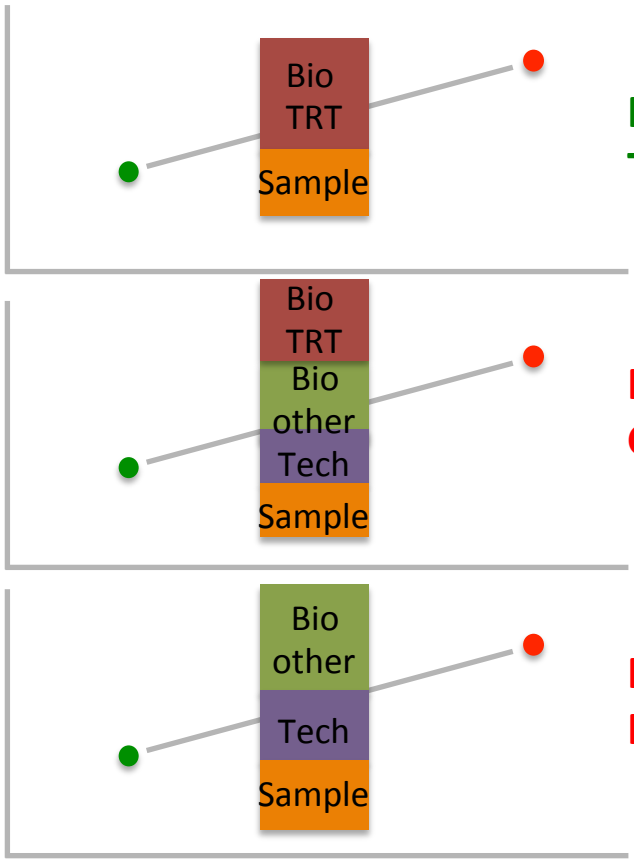
RNA
Library
Sequencing

N1 reads N2 reads

Trimming
Alignment
Read counting per gene

| Gene | Control 1 | TRT 1 |
|---|---|---|
| 1 | C1 | T1 |
| 2 | C2 | T2 |
| … | | |


Control 1 TRT 1 (salt)



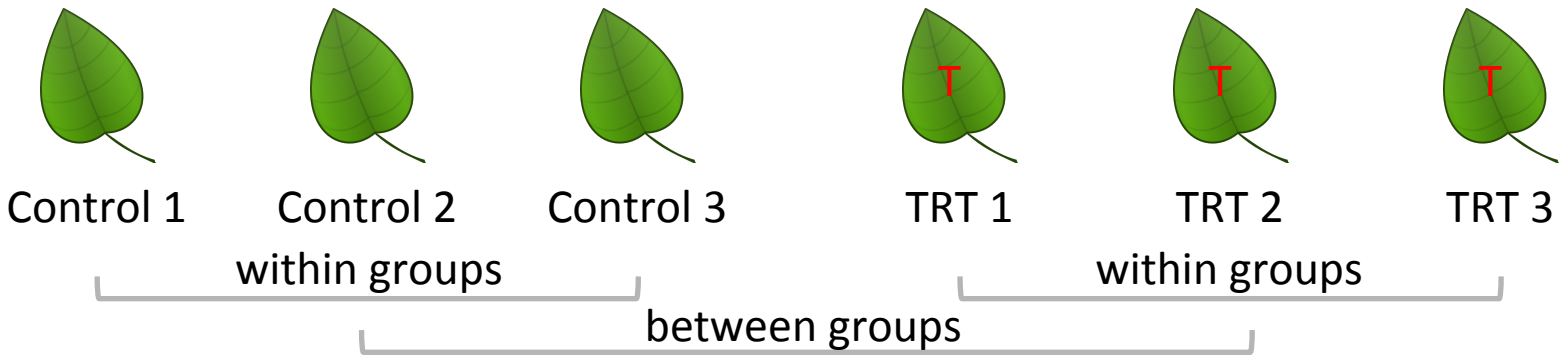

**FET – Significant**
**True Positive**

**FET – Significant**
**Gain false power**

**FET – Significant**
**False Positive**

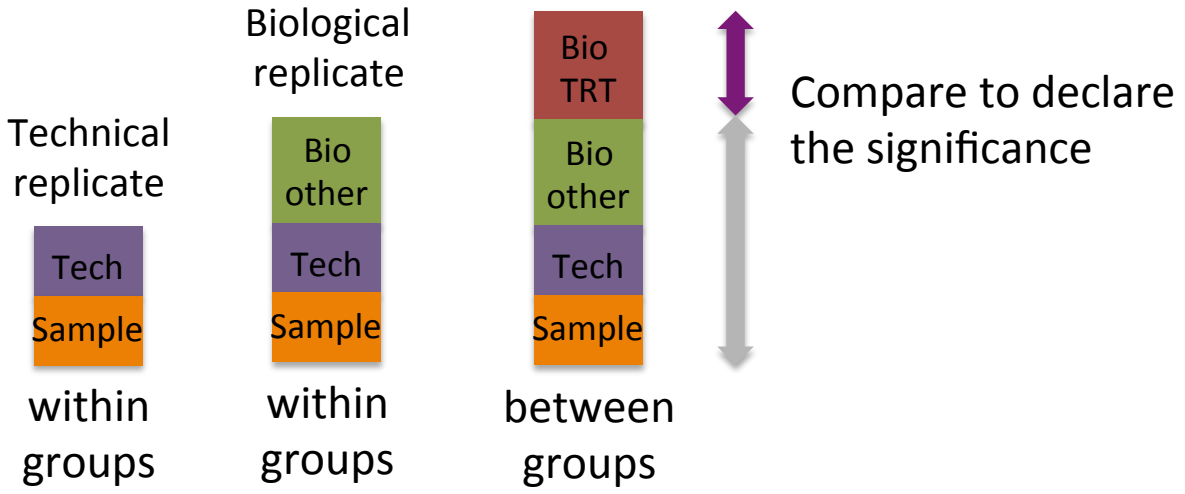
Needs a better experimental design and analysis to distinguish different types of variance and estimate them.

# Replication



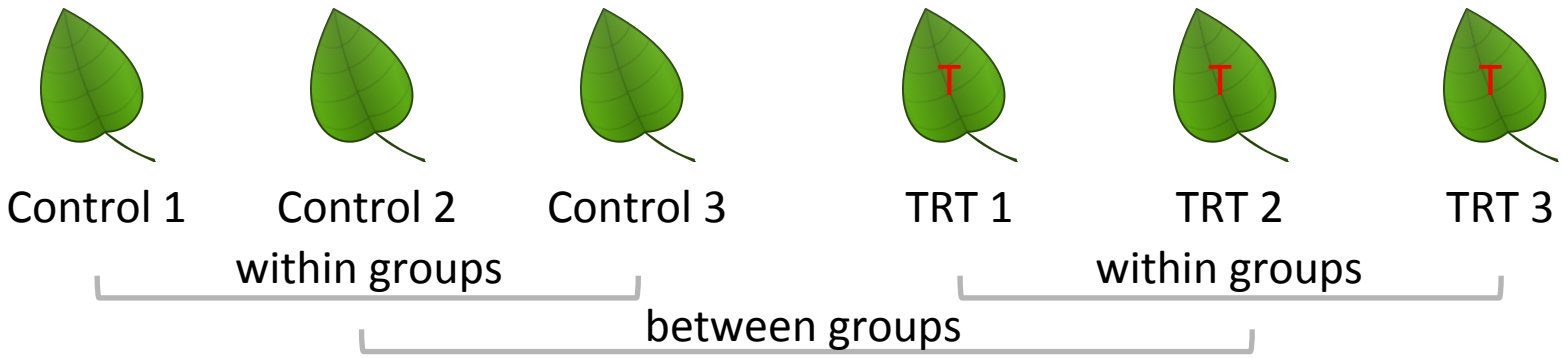

*Technical rep* refers to the sequencing of multiple libraries derived from the same biological sample.

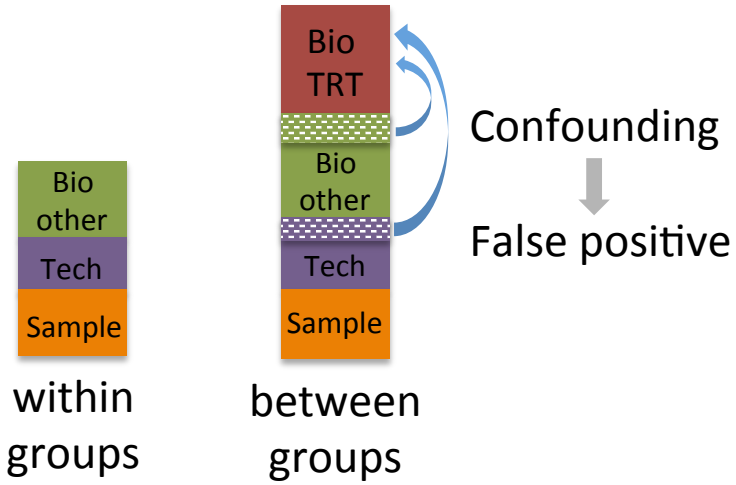
*Biological rep* from different biological samples.




1. Use biological replication instead of technical replication unless you have your own particular interest.
2. More replicates increase the power to detect small treatment effect

# Randomization and Unbiasedness




Seeds, planting, tissue collection, RNA extraction, library preparation, sequencing

Randomization and unbiasedness should be considered during the whole experiment as much as possible.

# Experimental Design

- Sequencing depth

Increasing sequencing depth decreases sampling variance

- Biological replication

Reasonable number of biological replication helps accurately estimate variances to achieve reliable statistical inference.

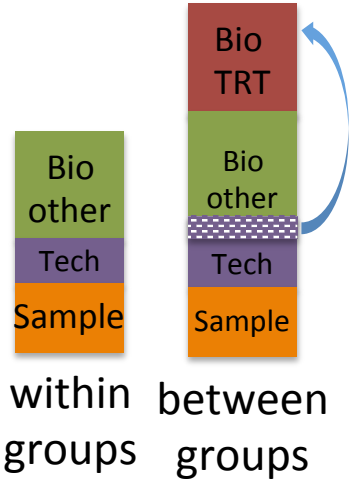
- Randomization and unbiasedness

To avoid confounding effect

# Question I

My lab conducted an RNA-Seq experiment to identify the DEs between two biological groups to examine a treatment of great interest. Each group has five biological replicates. I told my graduate student to perform the experiment of each group separately (then I don't need to worry that the samples from two groups are messed up).

Is this a sound experimental design? Why?



# Outline

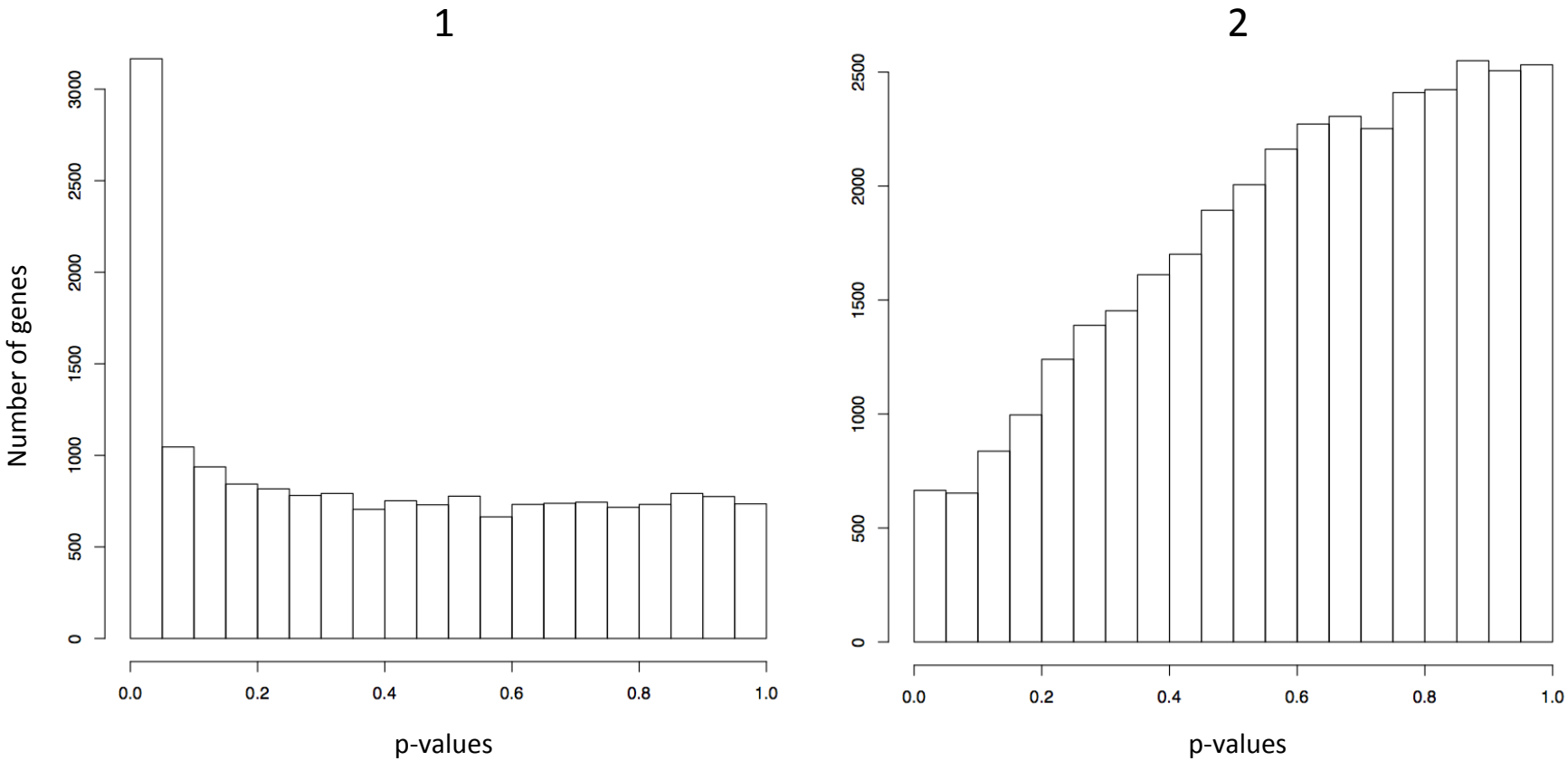
**Design of DE experiments and results**

- Experimental design
- P-values and q-values

**Other analyses**

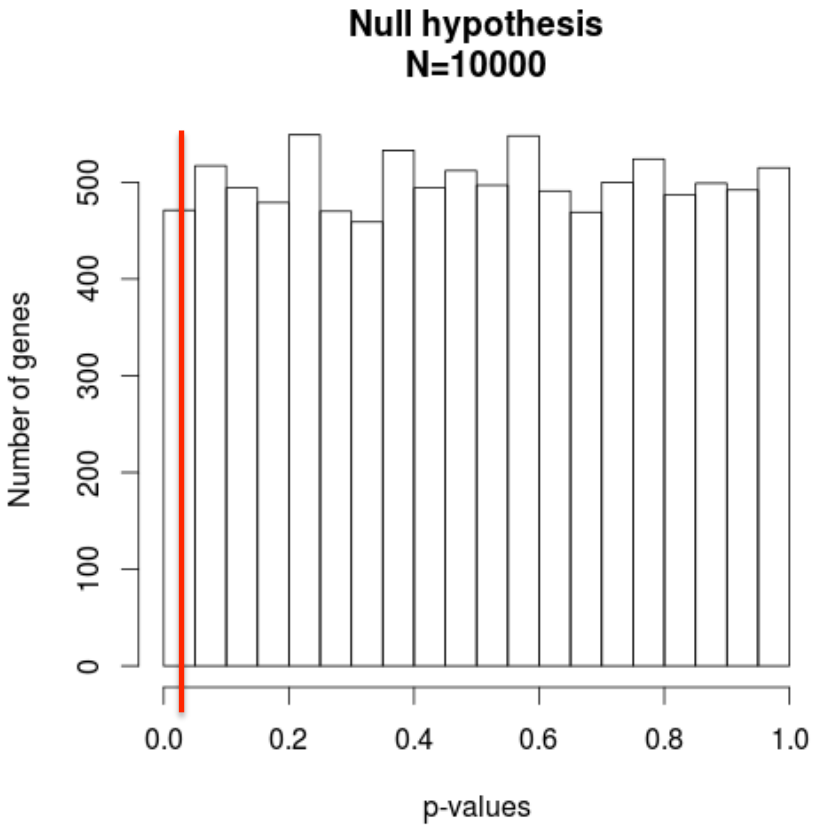
- Visualization
- GO term enrichment analysis

# P-value histograms from real studies




If you perform an RNA-Seq experiment, which one would you like to obtain?

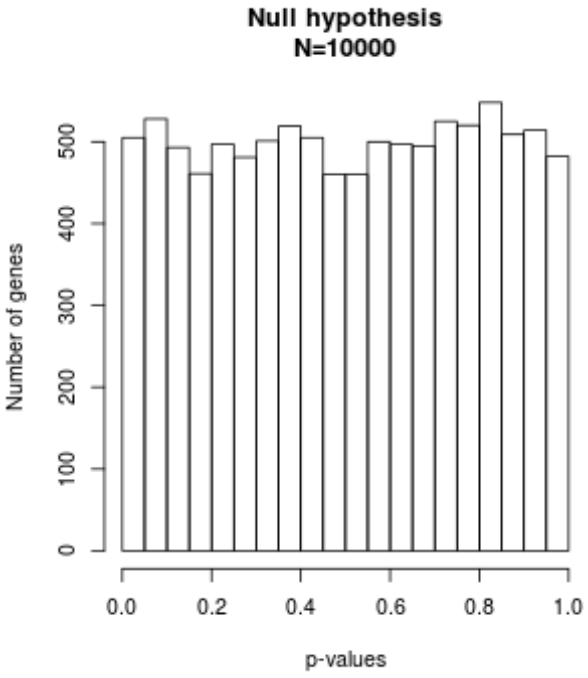
# P-value distribution under the null hypothesis (e.g., no treatment effect)

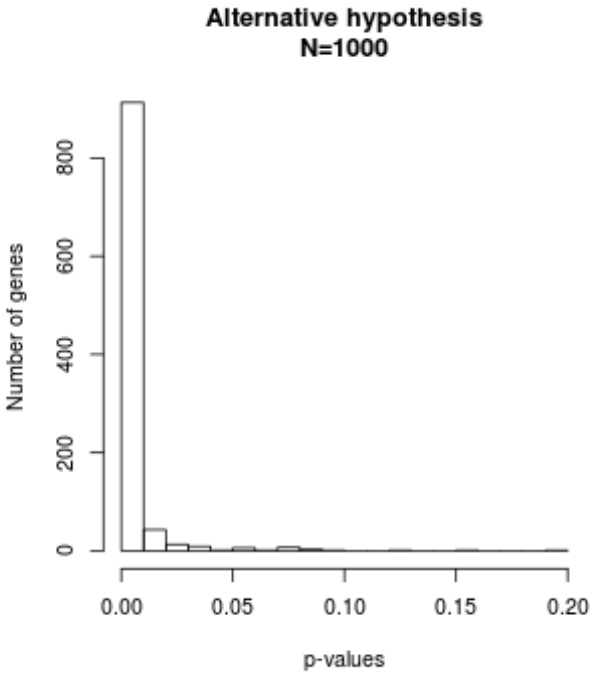


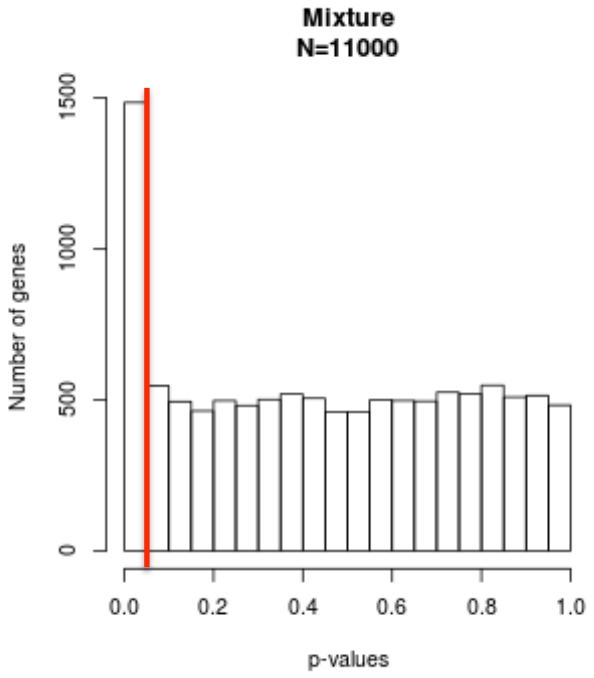

When the null hypothesis is true, a P-value is distributed uniformly.

No matter how stringent the criteria are, you'll identify genes with very small p-values and the false discovery rate (FDR) is 100%.

# P-value distribution under both the null and non-null hypotheses



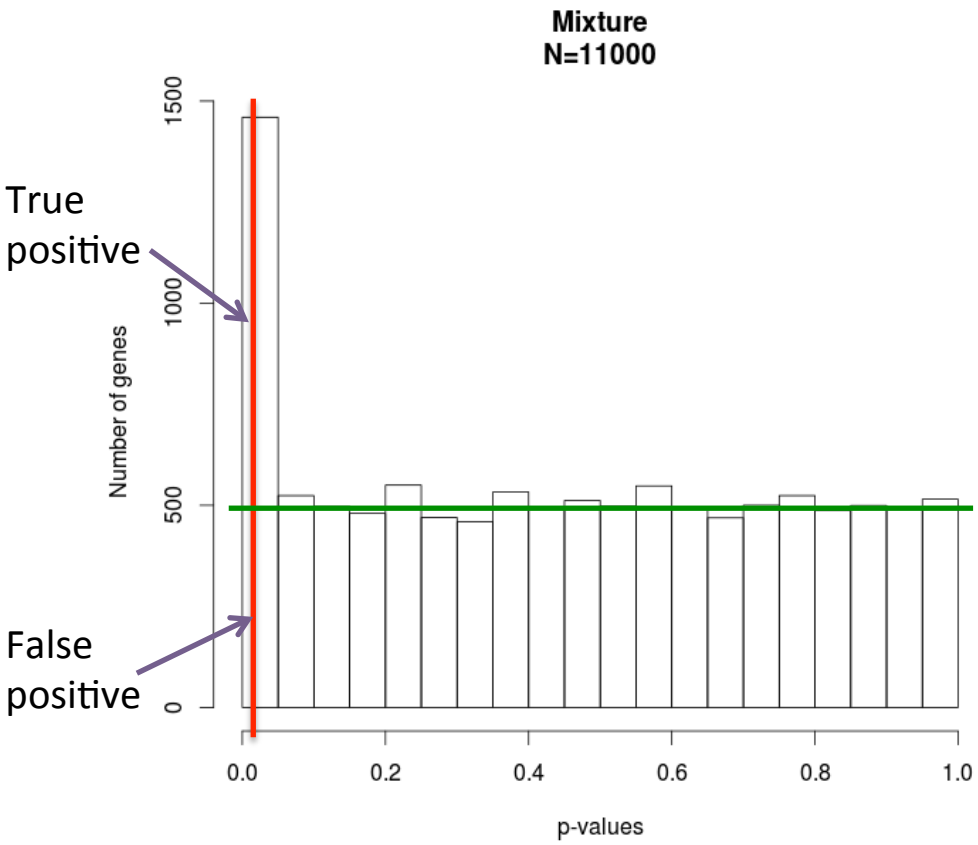

When the null hypothesis is true, a P-value is distributed uniformly.

When the null hypothesis is false, the P-value distribution is skewed toward 0.

Cutoff: p=0.05
FDR=471/(471+989)=32%

Cutoff: p=0.01
FDR=102/(102+912)=10%

# Multiple test correction – FDR method




FDR method (BH) to calculate q-values/adjusted p-values/corrected p-values*

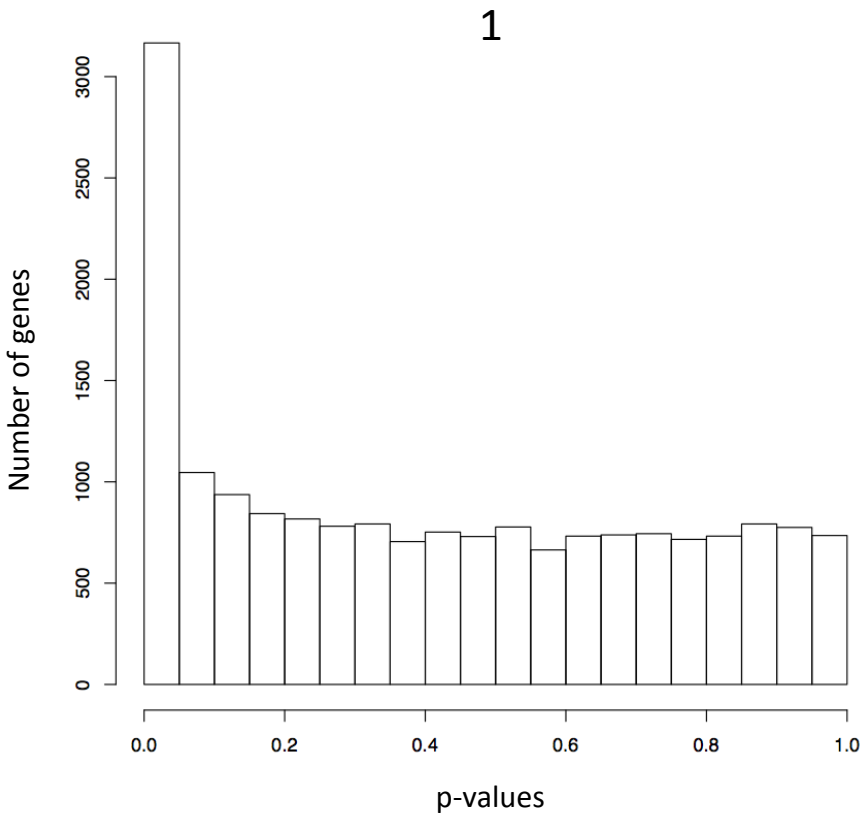
**q-values < 0.1**
FDR 10%

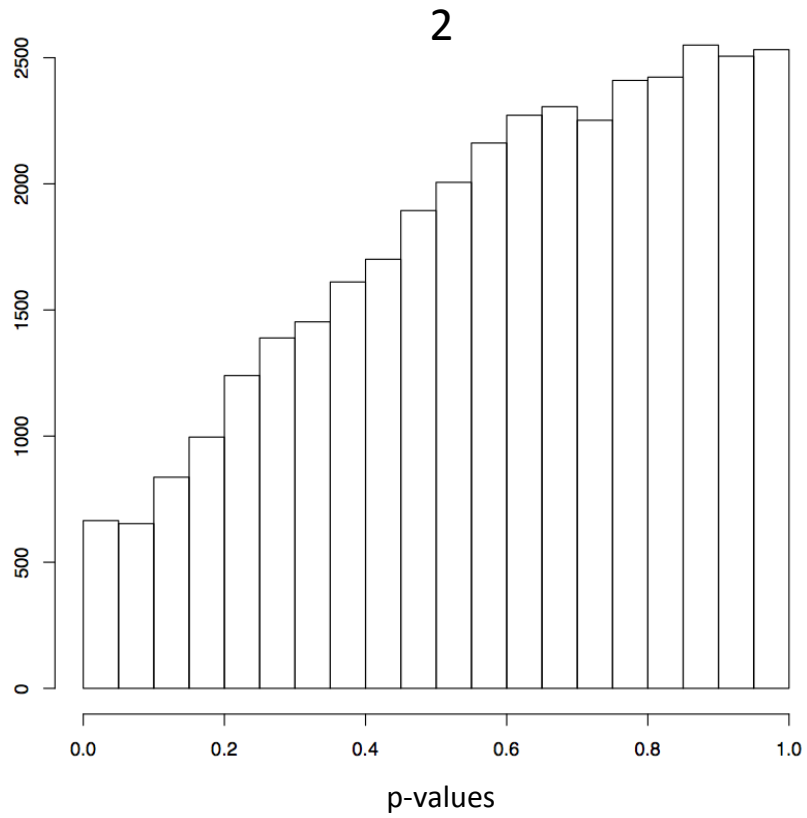
P-values < 0.00009
DE=992
False DE=99

| p-value | q-value |
|---|---|
| 0.846 | 0.984 |
| 0.957 | 0.996 |
| 0.598 | 0.939 |
| 0.112 | 0.586 |
| 0.574 | 0.937 |
| 0.583 | 0.938 |
| 0.178 | 0.712 |
| 0.025 | 0.228 |
| 0.241 | 0.776 |
| 0.832 | 0.983 |
| 0.269 | 0.803 |
| 0.983 | 0.998 |
| 0.917 | 0.993 |
| **0.001** | **0.015** |
| 0.109 | 0.582 |
| 0.585 | 0.938 |
| 0.871 | 0.987 |
| 0.932 | 0.993 |
| … | … |

* p.adjust in R, Journal of the Royal Statistical Society, Series B, 1995, 57:289–300

# P-value histograms from real studies




DE = 1,370, FDR=5%

DE = 0, FDR=20%

# Significant genes

| | Group 1 | | | Group 2 | | | DE Result | | |
|---|---|---|---|---|---|---|---|---|---|
| GeneID | Rep1 | Rep2 | Rep3 | Rep1 | Rep2 | Rep3 | Log2FC | p-vauel | q-value |
| 1 | 2679 | 2360 | 2573 | 2563 | 3398 | 3012 | -0.40 | 0.037 | 0.097 |
| 2 | 177 | 161 | 171 | 154 | 137 | 152 | 0.03 | 0.916 | 0.956 |
| 3 | 381 | 371 | 397 | 541 | 723 | 635 | -0.89 | 2.42E-05 | 0.00017 |
| 4 | 990 | 1073 | 1236 | 850 | 672 | 859 | 0.30 | 0.130 | 0.256 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | NA | NA | NA |
| 6 | 203 | 310 | 306 | 272 | 220 | 259 | -0.07 | 0.811 | 0.892 |
| … | | | | | | | | | |

Which genes can be called significant genes?

**Arbitrary criteria**

5% FDR, q-values < 0.05

10% FDR, q-values < 0.1

20% FDR, q-values < 0.2

5% FDR & (Log2FC > 1 or Log2FC < -1)

# Question II

If we identify 500 differential expression (DE) genes using the 5% FDR to account for multiple tests. Which one below is a better description?

1. I am 95% confident that 500 genes are DE.

2. The 5% genes (25 genes) in the set are not true DE genes.

# Outline

**Design of DE experiments and results**

- Experimental design
- P-values and q-values

**Other analyses**

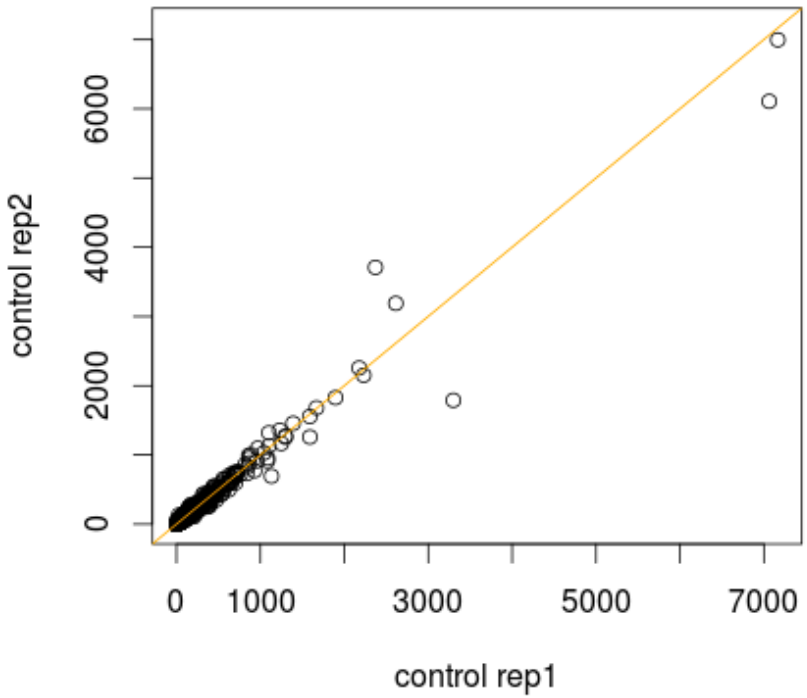
- Visualization
- GO term enrichment analysis

# Scatter plot

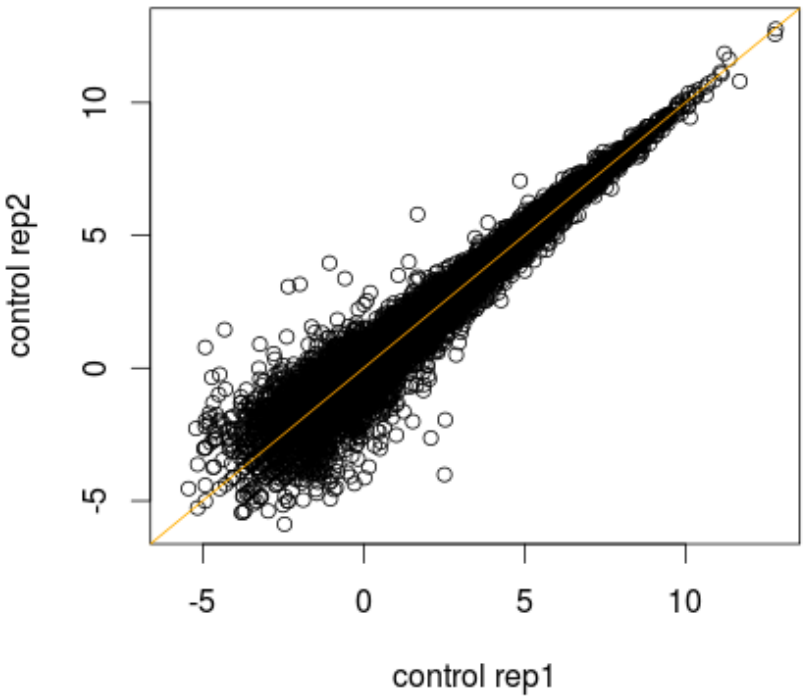
| Gene | Control rep1 | Control rep2 |
|---|---|---|
| 1 | 2679 | 2360 |
| 2 | 177 | 161 |
| 3 | 381 | 371 |
| … | | |

| Gene | Control rep1 RPKM | Control rep2 RPKM |
|---|---|---|
| 1 | 3.4 | 3.3 |
| 2 | 1.3 | 1.2 |
| 3 | 2.0 | 2.0 |
| … | | |

**Raw counts scatter plot**

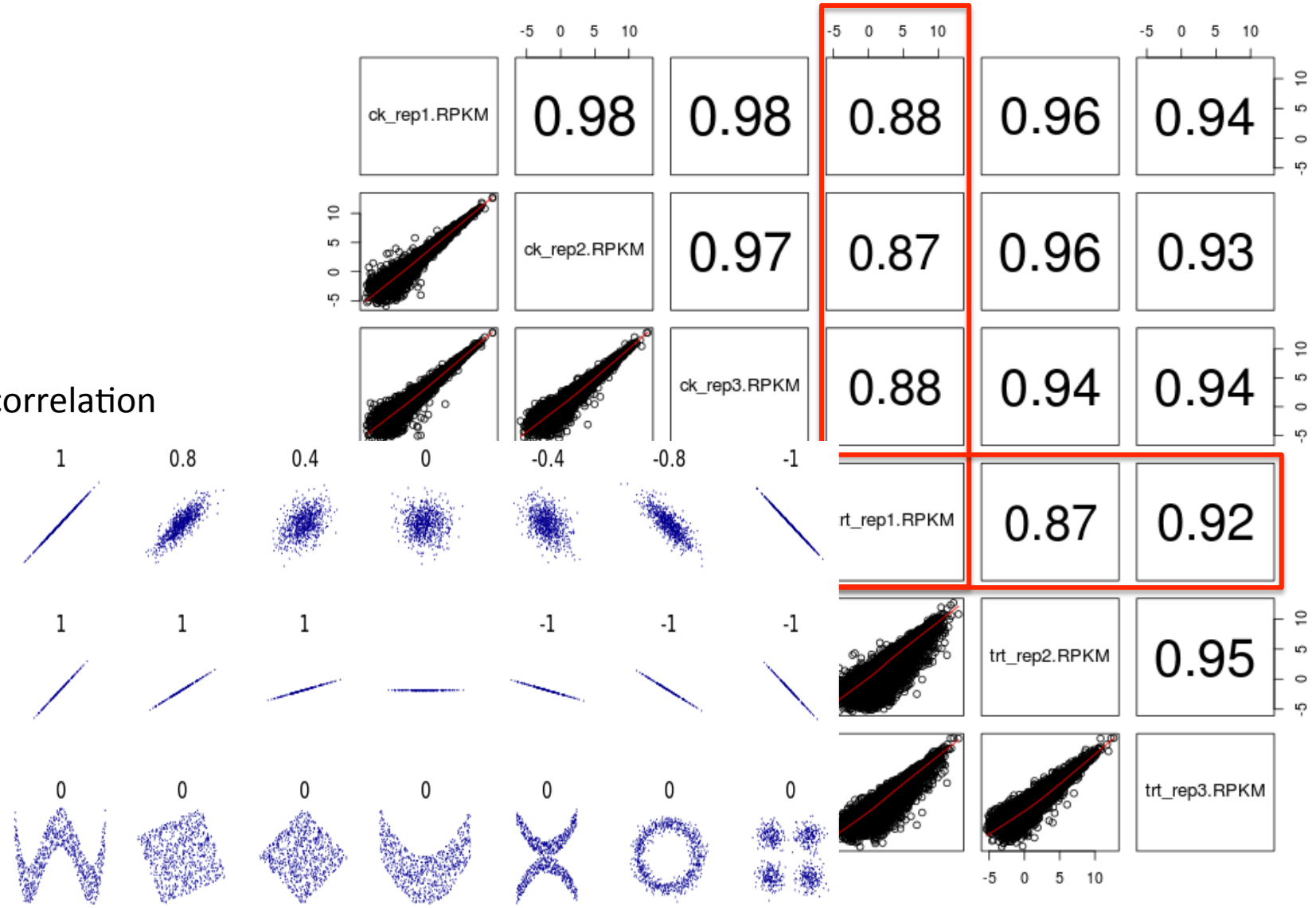



**RPKM scatter plot**



# Pair-wise scatter plot

correlation



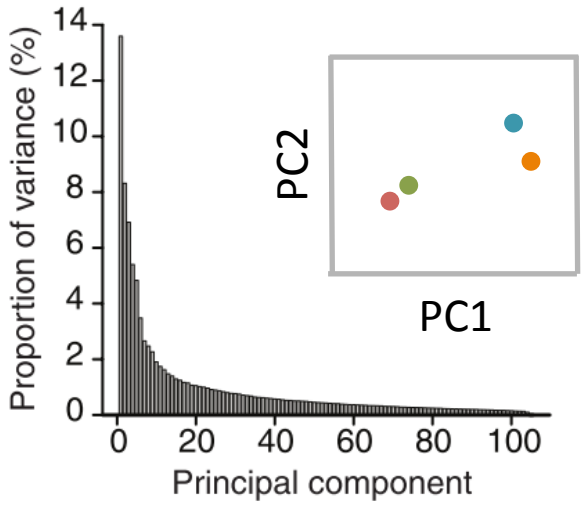

wikipedia

# Principal Component Analysis (PCA)

PCA is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set.

Could you use one sentence to summarize what you said in the last 30 minutes?

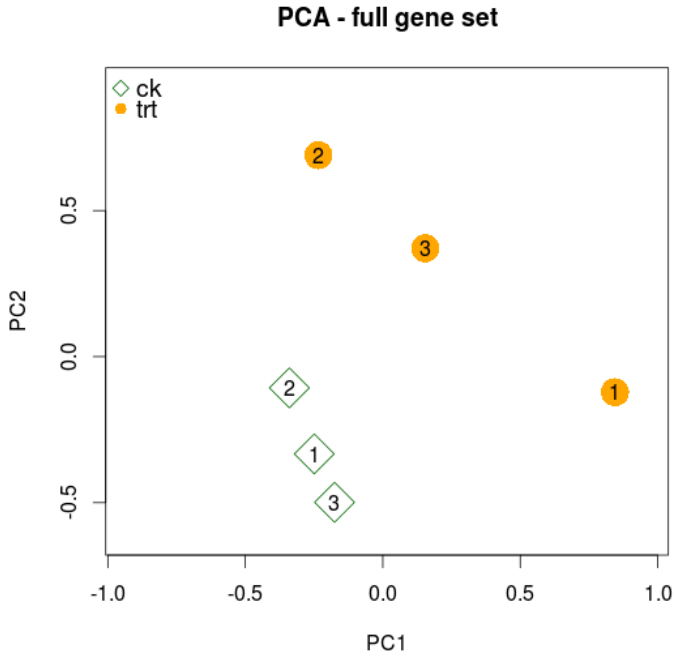
| Feature/ variable | John | Mike | Jack | Justin |
|---|---|---|---|---|
| Weight (lb) | 150 | 243 | 186 | 128 |
| Height (cm) | 171 | 190 | 178 | 175 |
| … | | | | |




Nature Biotech, 2008, 26:303-4

| | Control | | | Treatment | | |
|---|---|---|---|---|---|---|
| GeneID | Rep1 | Rep2 | Rep3 | Rep1 | Rep2 | Rep3 |
| 1 | 2679 | 2360 | 2573 | 2563 | 3398 | 3012 |
| 2 | 177 | 161 | 171 | 154 | 137 | 152 |
| 3 | 381 | 371 | 397 | 541 | 723 | 635 |
| … | | | | | | |
| 30000 | 990 | 1073 | 1236 | 850 | 672 | 859 |

Normalized and standardized data



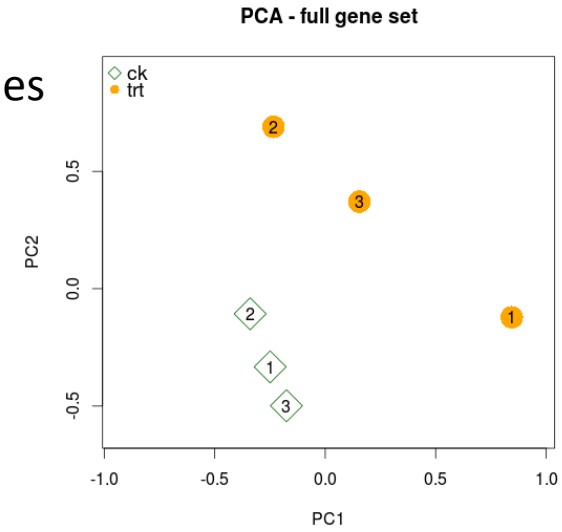
# PCA Plot using different inputs

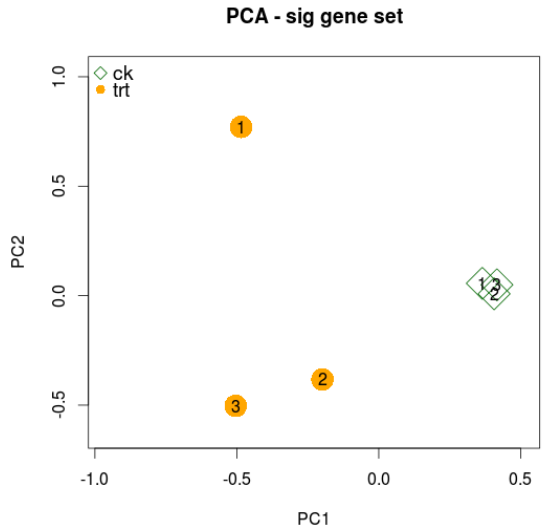
Standardization within genes

| Raw counts | Standarded |
|---|---|
| 3 | -0.48 |
| 6 | 0.02 |
| 8 | 0.36 |
| 4 | -0.31 |
| 6 | 0.02 |
| 9 | 0.52 |
| 5 | -0.14 |

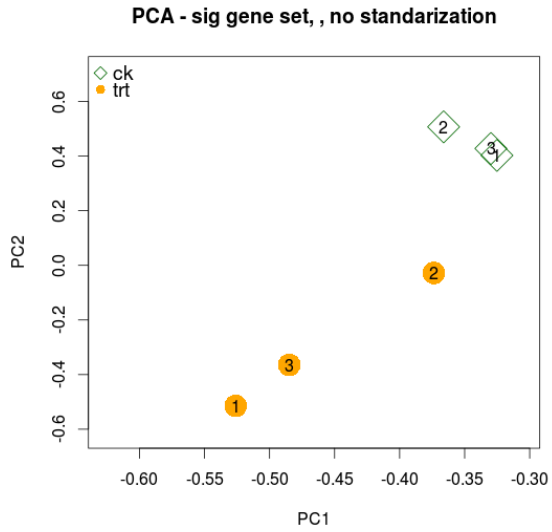
Mean = 5.8
Range = 9 – 3 = 6

x’ = (x-mean)/range



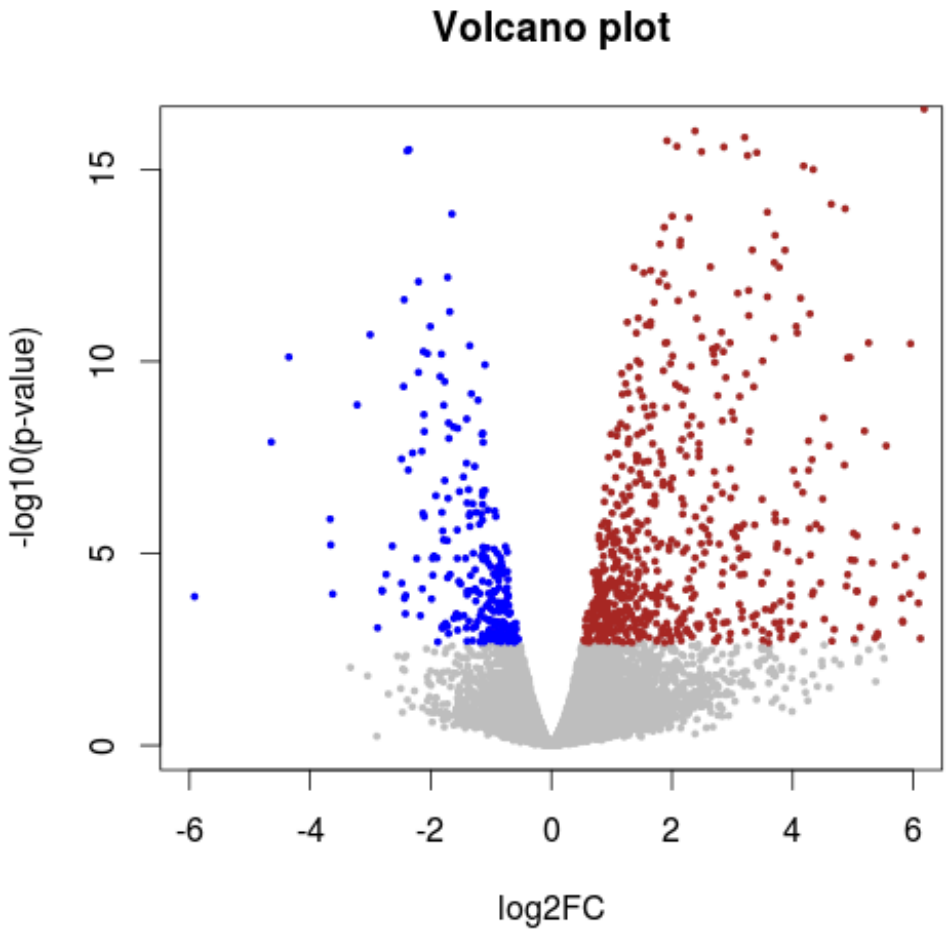







# Vocalno plot

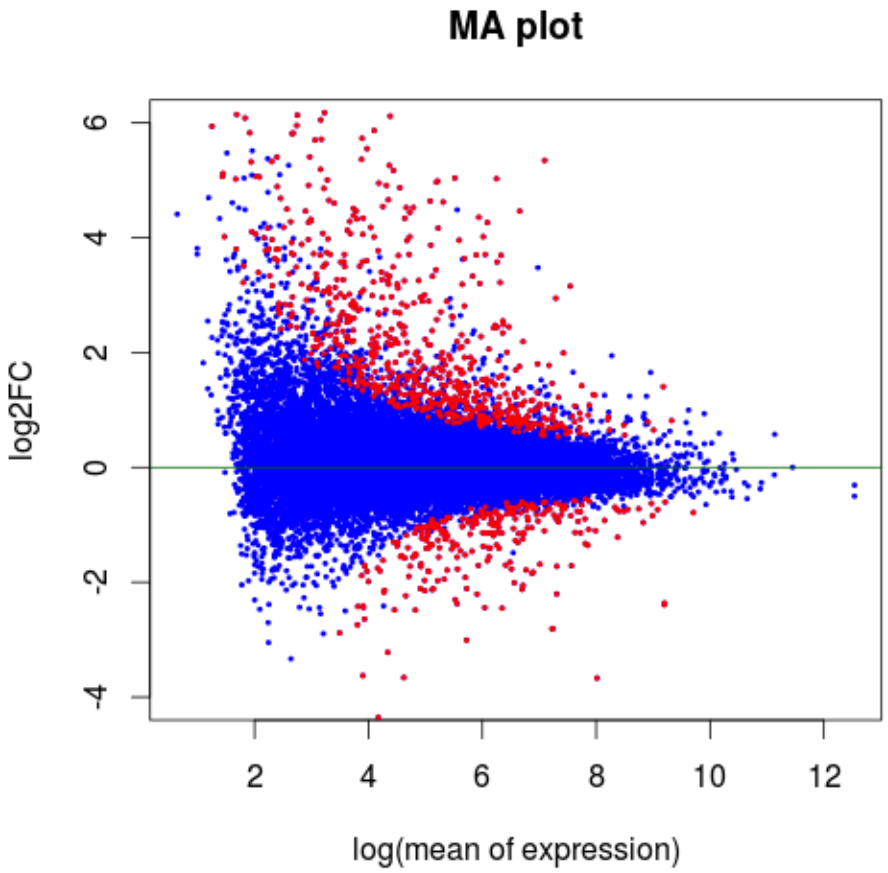

| DE Result | | | |
|---|---|---|---|
| GeneID | Log2FC | p-value | -log10(pvalue) |
| 1 | -0.40 | 0.037 | 1.43 |
| 2 | 0.03 | 0.916 | 0.04 |
| 3 | -0.89 | 2.42E-05 | 4.62 |
| 4 | 0.30 | 0.130 | 0.89 |
| 5 | -0.36 | 0.140 | 0.85 |
| 6 | -0.07 | 0.811 | 0.09 |
| … | | | |



# MA plot

M (log ratios) and A (mean average)

| GeneID | Mean RPKM | log mean | log2FC |
|---|---|---|---|
| 1 | 0.51 | -0.29 | -0.40 |
| 2 | 1.25 | 0.10 | 0.03 |
| 3 | 3.52 | 0.55 | -0.89 |
| 4 | 0.19 | -0.72 | 0.30 |
| 5 | 2.34 | 0.37 | -0.36 |
| 6 | 6.14 | 0.79 | -0.07 |
| … | | | |




More at: en.wikipedia.org/wiki/MA_plot

# Outline

**Design of DE experiments and results**

- Experimental design
- P-values and q-values

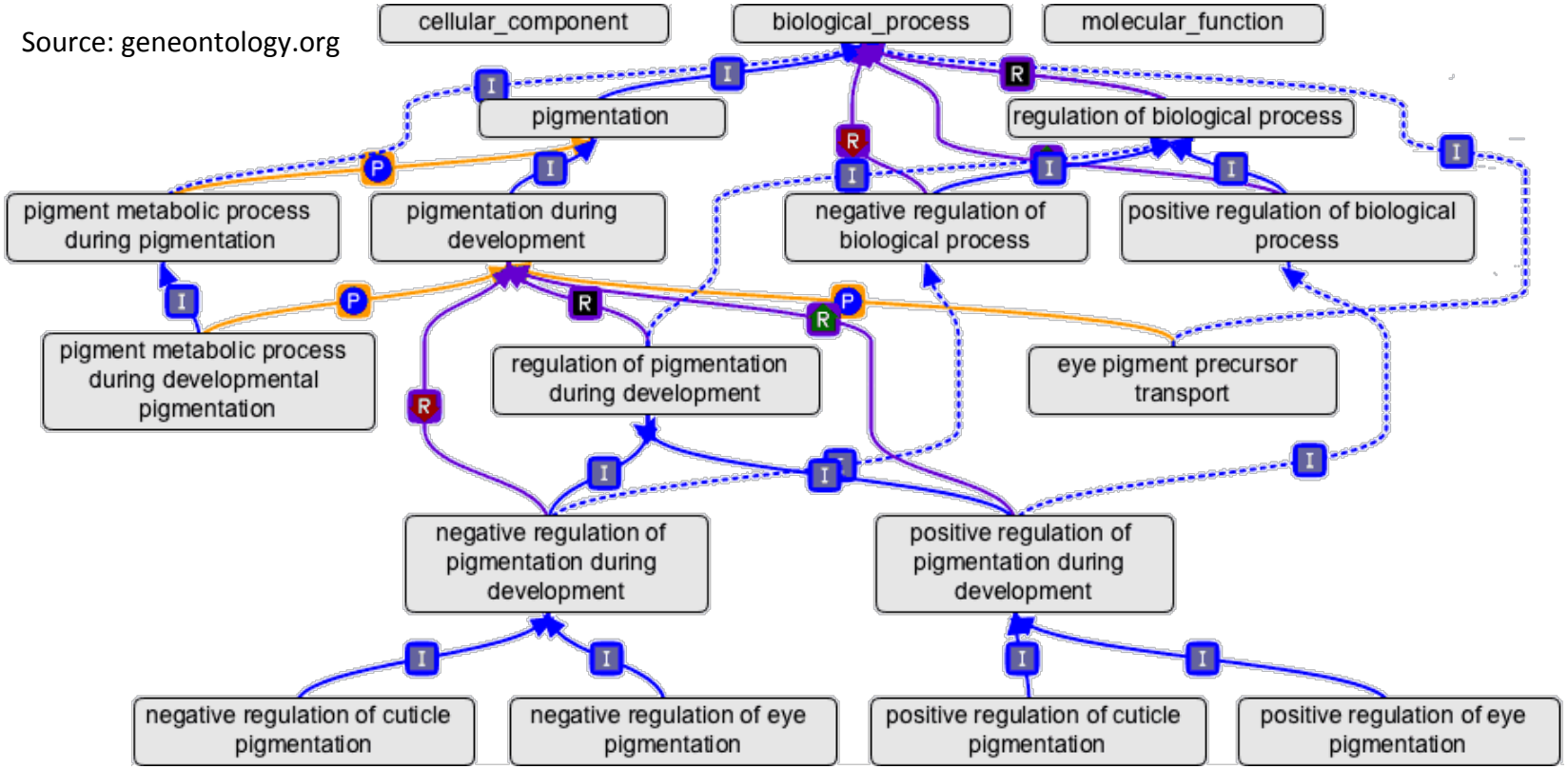
**Other analyses**

- Visualization
- GO term enrichment analysis

# Gene ontology (GO)

An ontology is a representation of a body of knowledge, within a given domain. Ontologies usually consist of a set of classes or terms with relations that operate between them.

Source: geneontology.org




Three domains (roots)

Node: GO term (e.g., cell growth, GO:0016049, biological process)

Edge: term-term connection

Each GO term can be traced back to a root

# GO enrichment test

| Gene | GO accession |
|---|---|
| GRMZM2G001475 | **GO:0006519** |
| GRMZM2G001475 | GO:0016831 |
| GRMZM2G001500 | GO:0005524 |
| GRMZM2G001500 | GO:0006457 |
| GRMZM2G001500 | GO:0051082 |
| GRMZM2G001508 | GO:0003993 |
| GRMZM2G001514 | GO:0003677 |
| GRMZM2G001514 | GO:0004879 |
| GRMZM2G001514 | GO:0005634 |
| GRMZM2G001514 | GO:0006355 |
| … | … |

| | |
|---|---|
| GRMZM2G001475 | 1 |
| GRMZM2G002652 | 2 |
| GRMZM2G006480 | 3 |
| … | … |
| GRMZM5G868038 | 40 |

| Gene | Significant? |
|---|---|
| GRMZM2G001475 | no |
| GRMZM2G002652 | no |
| GRMZM2G006480 | yes |
| … | … |
| GRMZM5G868038 | no |

Question: Are the genes of this GO term enriched in the significant gene set?

Assumption: all genes are independent and equally likely to be selected as DEs.

2x2 Table for GO:0006519

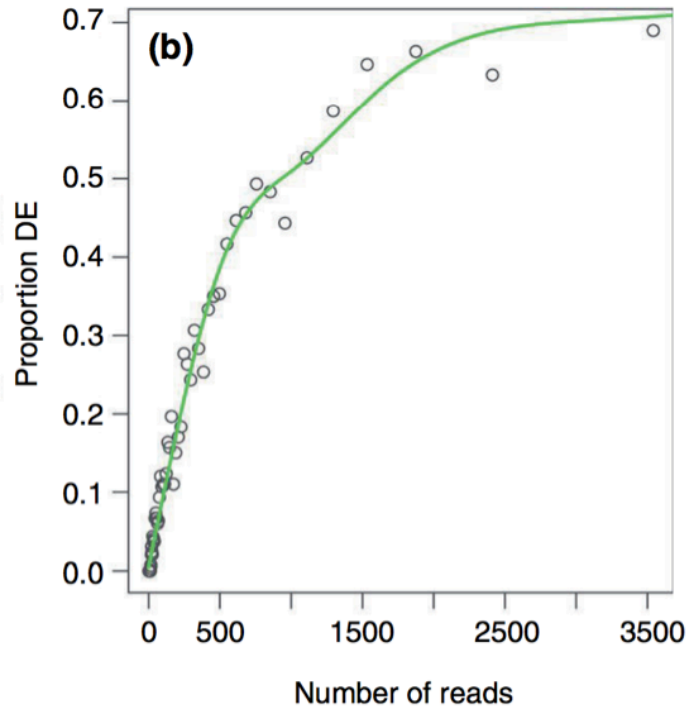
| | **GO:0006519** | Others |
|---|---|---|
| Significant | 5 | 210 |
| Not significant | 35 | 39416 |

Fisher's Exact Test:
p-value = 2.518e-06

| | |
|---|---|
| **Name** | cellular amino acid metabolic process |
| **Ontology** | Biological Process |
| **Definition** | The chemical reactions and pathways involving amino acids, carboxylic acids containing one or more amino groups, as carried out by individual cells. |

# GOSeq

Not all genes are equally likely to be selected as DEs.




Young MD, et al., (2010). Genome Biology, 11: R14.

1. The likelihood of DE as a function of number of reads is quantified through fitting a monotonic function to "proportion of DE" versus "number of reads".

2. The function is incorporated into the enrichment statistical test

| Gene | Significant? | Read counts | Proportion |
|---|---|---|---|
| GRMZM2G001475 | no | 224 | 0.16 |
| GRMZM2G002652 | no | 51 | 0.05 |
| GRMZM2G006480 | yes | 536 | 0.38 |
| … | … | … | … |
| GRMZM5G868038 | no | 0 | 0 |

3. Weighted sampling to perform enrichment test

| GO:0006519 | # DE |
|---|---|
| Obs (from the DE analysis) | 5 |
| 1st weighted sampling | 1 |
| 2nd weighted sampling | 0 |
| 3rd weighted sampling | 2 |
| … | … |

→ p-value

# Summary

- Biological replication and randomization need to be considered during the experimental design
- Multiple test correction is critical for any analyses with large number of statistical tests
- Use a proper approach for the GO enrichment test
- R is an excellent tool to visualize the data

# My contact information

liu3zhen@ksu.edu
4022B Throckmorton Plant Sciences Center
Manhattan, KS 66506-5502
phone: 785-532-1379
twitter: liu3zhen