LUXEMBOURG
INSTITUTE
OF HEALTH

UNIVERSITÉ DU
LUXEMBOURG

Data Science for Bioinformatics:
Data and Statistics

Petr Nazarov

CANBIO2 course

petr.nazarov@lih.lu

http://edu.modas.lu/dasbit_ds

2025-02-10

Outline of the Course

1. Data overview

- RNA-seq data generation
- File formats, Phred-quality
- Sequence-based QC: FastQC / MultiQC
- Statistical properties of the data

3. Statistical basics

- Hypothesis testing (p-value)
- T-test, Wilcoxon test
- Multiple testing (FDR, FWER)
- Linear models: ANOVA

4. Statistics for RNA-seq

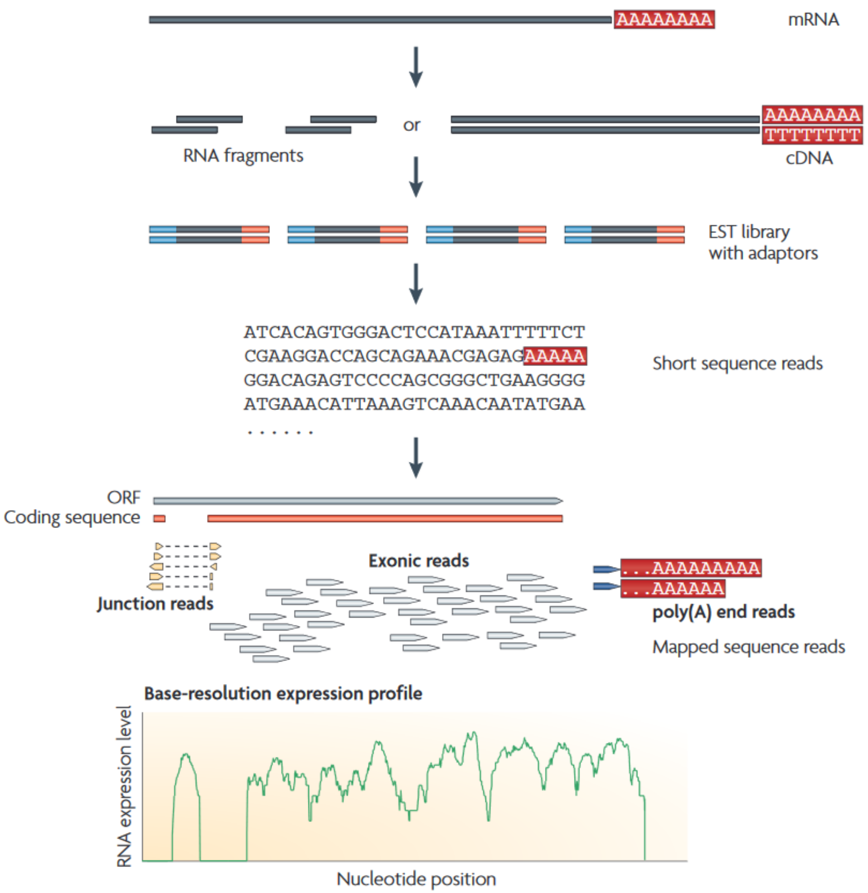
- Differential expression analysis
- EdgeR, DESeq, limma
- Enrichment analysis

Please see scripts and materials online http://edu.modas.lu/transcript-seq

1. Data Overview

http://edu.modas.lu/transcript-seq/part1.html

1.1. RNA-seq Data Generation



Wang Z et al. RNA-Seq: a revolutionary tool for transcriptomics. **Nat Rev Genet. 2009**

read: short fragment detected by RNA-seq

library: collection of all reads from the sample

CPM: counts per million nucleotides

TPM: transcripts per million (proportion)

FPKM: fragments per kilobase of exon per million reads mapped

RPKM: reads per ……. (for single-end)

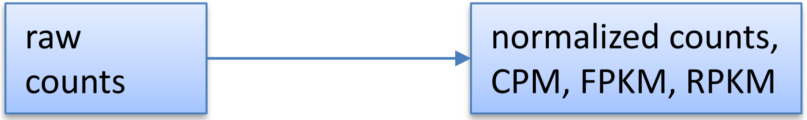
$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6 \qquad \mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6$$

X_i – observed number of reads

N – library size

l_i – length of the gene (transcript)

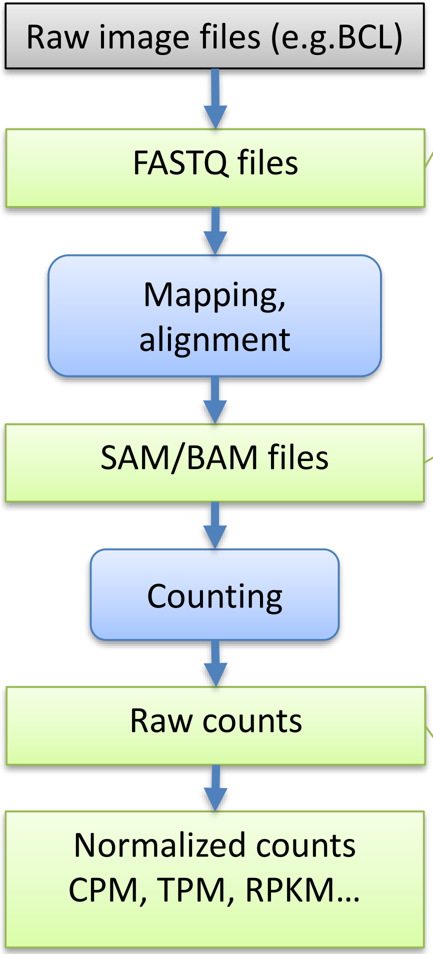
$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$



raw counts → normalized counts, CPM, FPKM, RPKM

10-minute simple explanation of TPM / FPKM
https://www.youtube.com/watch?v=TTUrtCY2k-w

1.2. File Formats

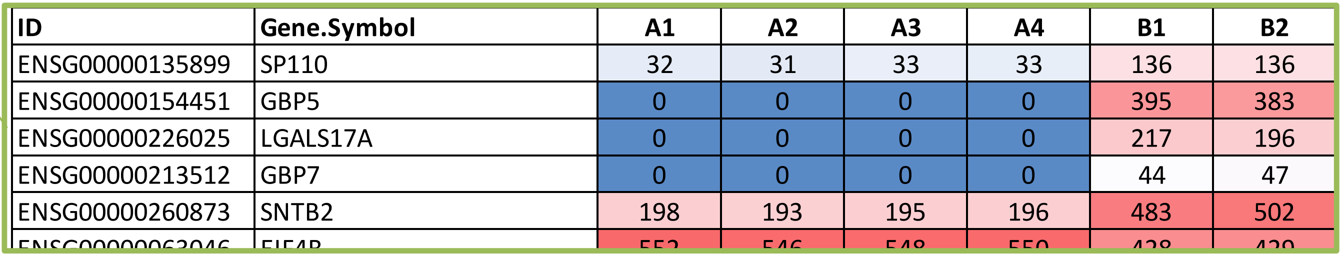


```
@HWI-ST508:152:D06G9ACXX:2:1101:1160:2042 1:Y:0:ATCACG
NAAGACCGAATTCTCCAAGCTATGGTAAACATTGCACTGGCCTTTCATCTG
+
#11??+2<<<CCB4AC?32@+1@AB1**1?AB<4=4>=BB<9=>?######
```

Link with detailed explanati

```
@HD       VN:1.0 SO:coordinate
@SQ       SN:seq1 LN:5000
@SQ       SN:seq2  LN:5000
@CO       Example of SAM/BAM file format.

B7_591:4:96:693:509 73       seq1     1        99        36M       *
          0         0        CACTAGTGGCTCATTGTAAATGTGTGGTTTAACTCG
                             <<<<<<<<<<<<<<<;<<<<<<<<<5<<<<<;:<;7
          MF:i:18   Aq:i:73  NM:i:0    UQ:i:0    H0:i:1

H1:i:0EAS54_65:7:152:368:11373       seq1      3         99        35M       *
          0         0
          CTAGTGGCTCATTGTAAATGTGTGGTTTAACTCGT
          <<<<<<<<<<0<<<<655<<7<<<:9<<3/:<6):   MF:i:18   Aq:i:66   NM:i:0
          UQ:i:0    H0:i:1    H1:i:0
```

Link with detailed explanati



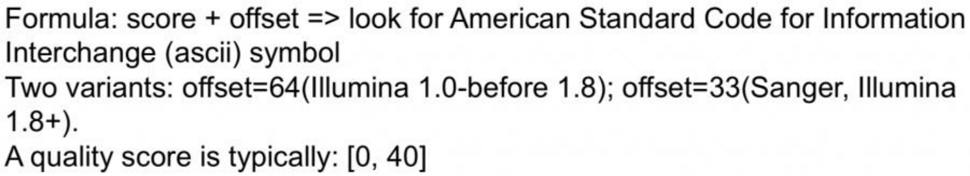
ID	Gene.Symbol	A1	A2	A3	A4	B1	B2
ENSG00000135899	SP110	32	31	33	33	136	136
ENSG00000154451	GBP5	0	0	0	0	395	383
ENSG00000226025	LGALS17A	0	0	0	0	217	196
ENSG00000213512	GBP7	0	0	0	0	44	47
ENSG00000260873	SNTB2	198	193	195	196	483	502
ENSG00000063046	EIF4B	552	546	548	550	428	429

Advantage of RNA-seq: you can repeat the pipeline with new knowledge or questions

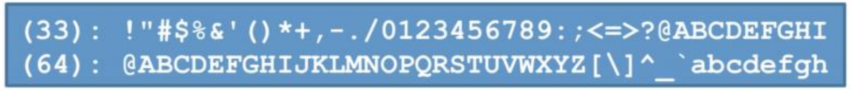
1.2. File Formats

```
@HWI-ST508:152:D06G9ACXX:2:1101:1160:2042 1:Y:0:ATCACG
NAAGACCGAATTCTCCAAGCTATGGTAAACATTGCACTGGCCTTTCATCTG
+
#11??+2<<<CCB4AC?32@+1@AB1**1?AB<4=4>=BB<9=>?######
```

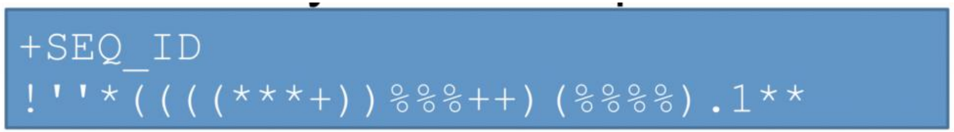
Quality scores started as numbers (0-40) but have since changed to an ASCII encoding to reduce filesize and make working with this format a bit easier, however they still hold the same information. ASCII codes are assigned based on the formula found below. This table can serve as a lookup as you progress through your analysis.



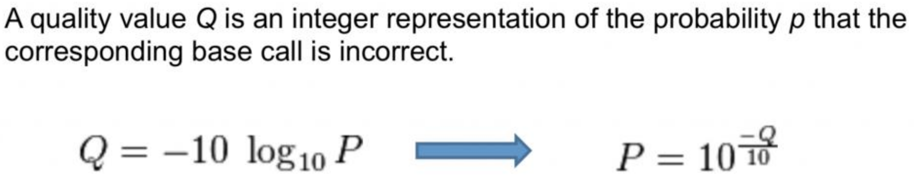
Formula: score + offset => look for American Standard Code for Information Interchange (ascii) symbol
Two variants: offset=64(Illumina 1.0-before 1.8); offset=33(Sanger, Illumina 1.8+).
A quality score is typically: [0, 40]



```
(33): !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
(64): @ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
```

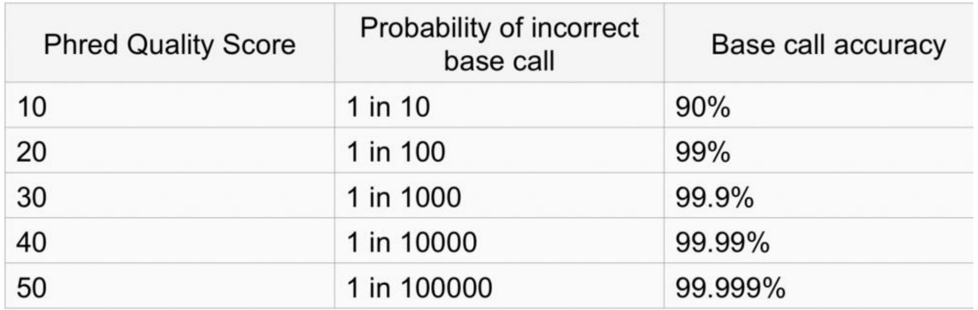


```
+SEQ_ID
!''*((((***+))%%%++)(%%%%).1***
```



A quality value *Q* is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10\ \log_{10} P \longrightarrow P = 10^{\frac{-Q}{10}}$$

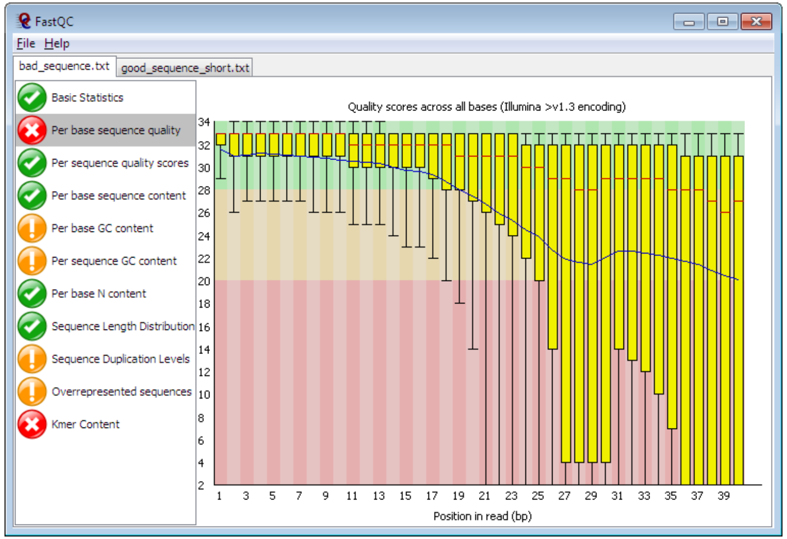


Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/

1.3. Sequence-based QC: FastQC

FastQC – a simple but widely-used Java-based tool for quality control of the experiments at the sequence level. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

- Import of data from BAM, SAM or FastQ files (any variant)
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

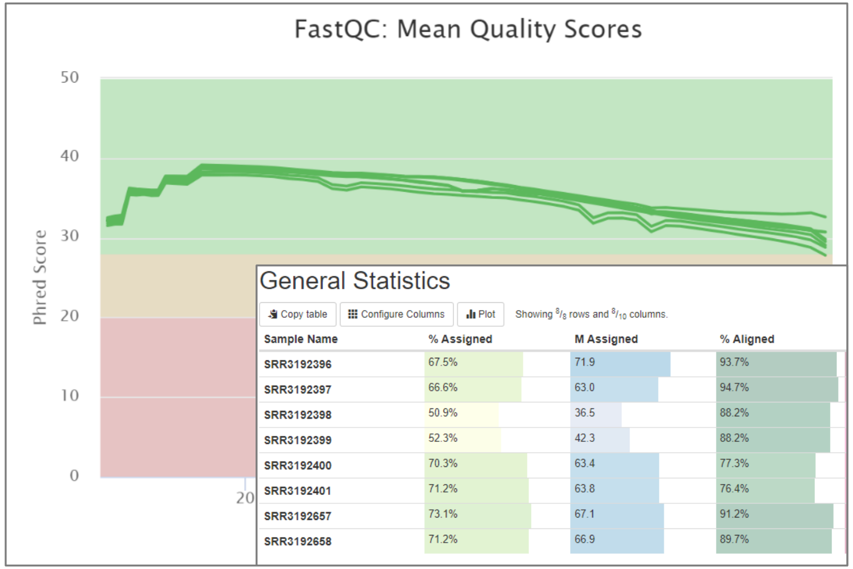
Examples

More detailed explanation & examples:
https://scienceparkstudygroup.github.io/rna-seq-lesson/03-qc-of-sequencing-results/index.html#31-running-fastqc

1.3. Sequence-based QC: MultiQC

A modular tool to aggregate results from bioinformati many samples into a single report. Python-based https://multiqc.info/ - see example online.

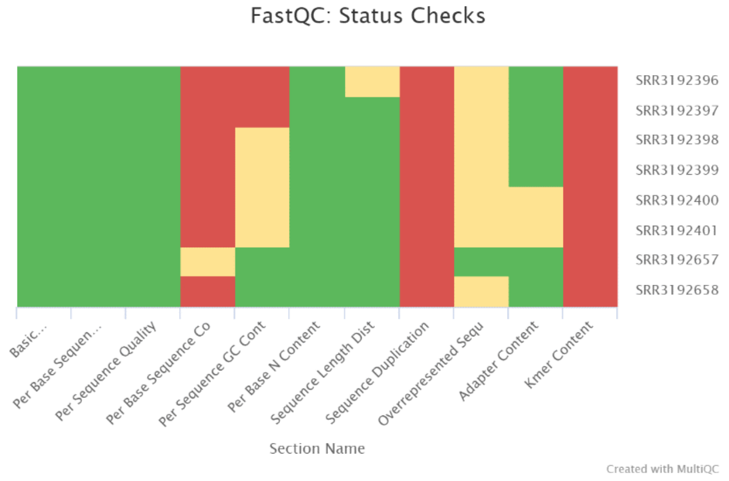
Introduction: https://www.youtube.com/watch?v=BbScv9TcaM



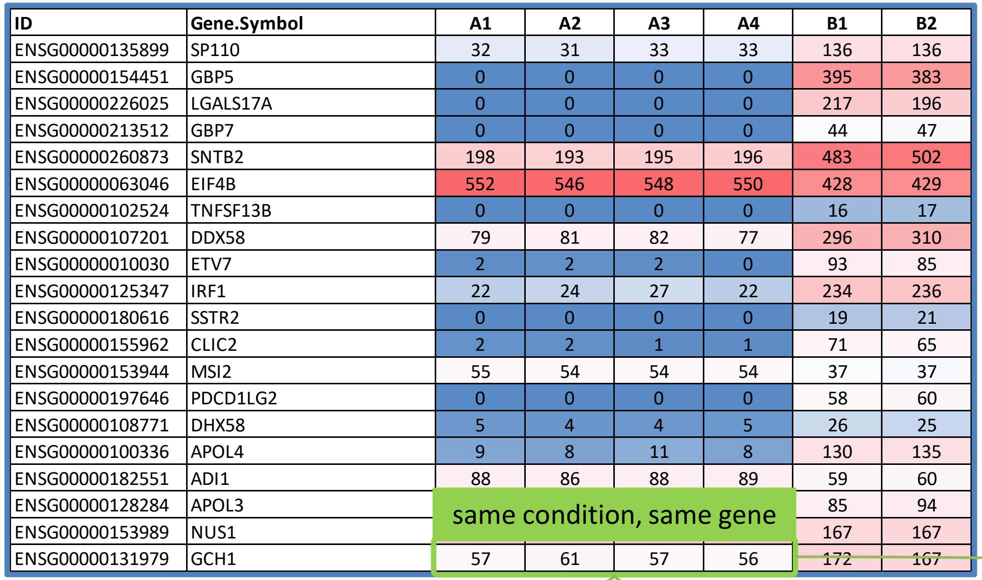
General Statistics

Showing 8/8 rows and 8/10 columns.

Sample Name	% Assigned	M Assigned	% Aligned
SRR3192396	67.5%	71.9	93.7%
SRR3192397	66.6%	63.0	94.7%
SRR3192398	50.9%	36.5	88.2%
SRR3192399	52.3%	42.3	88.2%
SRR3192400	70.3%	63.4	77.3%
SRR3192401	71.2%	63.8	76.4%
SRR3192657	73.1%	67.1	91.2%
SRR3192658	71.2%	66.9	89.7%

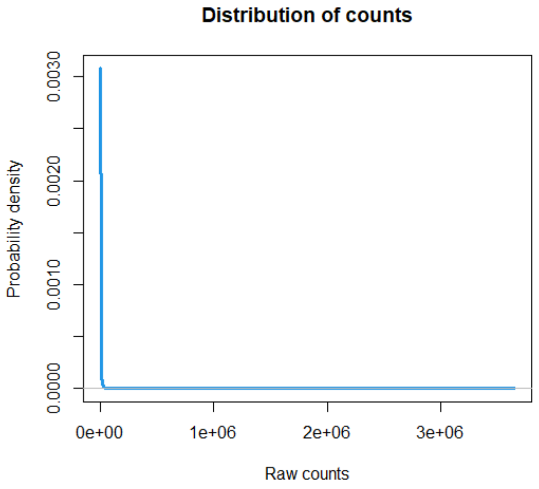


1.4. Statistical Properties of the Data

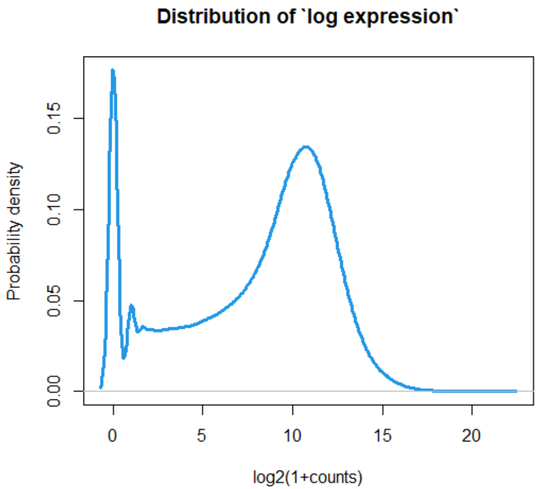


ID	Gene.Symbol	A1	A2	A3	A4	B1	B2
ENSG00000135899	SP110	32	31	33	33	136	136
ENSG00000154451	GBP5	0	0	0	0	395	383
ENSG00000226025	LGALS17A	0	0	0	0	217	196
ENSG00000213512	GBP7	0	0	0	0	44	47
ENSG00000260873	SNTB2	198	193	195	196	483	502
ENSG00000063046	EIF4B	552	546	548	550	428	429
ENSG00000102524	TNFSF13B	0	0	0	0	16	17
ENSG00000107201	DDX58	79	81	82	77	296	310
ENSG00000010030	ETV7	2	2	2	0	93	85
ENSG00000125347	IRF1	22	24	27	22	234	236
ENSG00000180616	SSTR2	0	0	0	0	19	21
ENSG00000155962	CLIC2	2	2	1	1	71	65
ENSG00000153944	MSI2	55	54	54	54	37	37
ENSG00000197646	PDCD1LG2	0	0	0	0	58	60
ENSG00000108771	DHX58	5	4	4	5	26	25
ENSG00000100336	APOL4	9	8	11	8	130	135
ENSG00000182551	ADI1	88	86	88	89	59	60
ENSG00000128284	APOL3	[illegible]	[illegible]	[illegible]	[illegible]	85	94
ENSG00000153989	NUS1	[illegible]	[illegible]	[illegible]	[illegible]	167	167
ENSG00000131979	GCH1	57	61	57	56	172	167

same condition, same gene

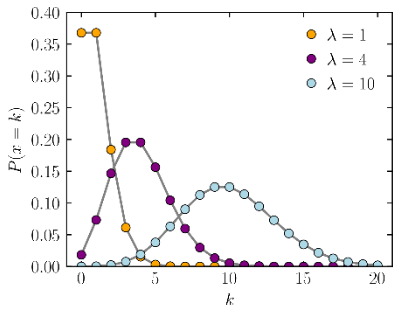


Distribution of counts



Distribution of `log expression`

Poisson distribution

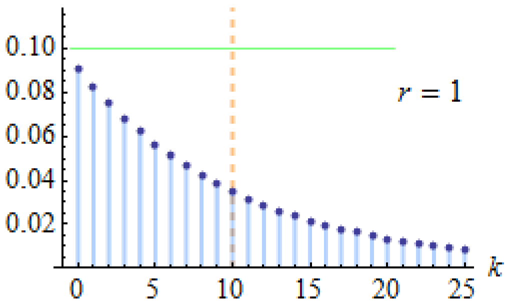


$$\frac{\lambda^k e^{-\lambda}}{k!}$$

1 parameter to fit (λ) =>

too simple!

Negative binomial distribution



$$\binom{k+r-1}{r-1}(1-p)^k p^r$$

2 parameters to fit (p,r) =>

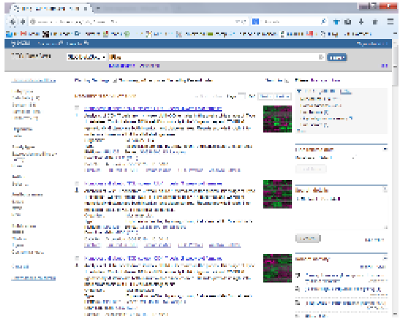
fits biology better!

Normal distribution

Can be used for log(1+k), when k is large, but it is approximate => less power (still usable but may miss interesting cases)

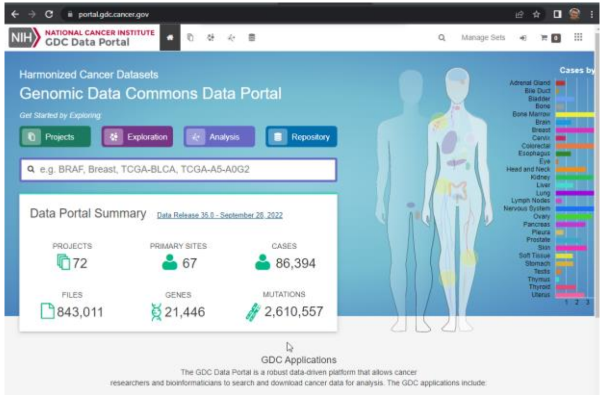
1.5. Data Repositories

GEO: http://www.ncbi.nlm.nih.gov/gds



US-based repository of omics data

TCGA: https://tcga-data.nci.nih.gov/tcga/



portal.gdc.cancer.gov

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Manage Sets

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Exploration | Analysis | Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 35.0 - September 28, 2022

PROJECTS	PRIMARY SITES	CASES
72	67	86,394
FILES	**GENES**	**MUTATIONS**
843,011	21,446	2,610,557

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

~11k tumor samples

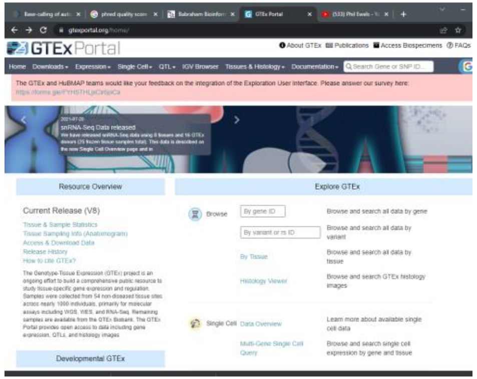
Analysis via:
http://www.cbioportal.org/public-portal/

cBioPortal FOR CANCER GENOMICS

ArrayExpress: http://www.ebi.ac.uk/arrayexpress/

EU-based repository of omics data

GTEx: https://www.gtexportal.org/home/



gtexportal.org/home/

GTEx Portal

About GTEx | Publications | Access Biospecimens | FAQs

Home | Downloads | Expression | Single Cell | QTL | IGV Browser | Tissues & Histology | Documentation | Search Gene or SNP ID...

The GTEx and HuBMAP teams would like your feedback on the integration of the Exploration User Interface. Please answer our survey here: https://forms.gle/FYHSTHLjxCe6pjiCa

2021-07-20
snRNA-Seq Data released
We have released snRNA-Seq data using 8 tissues and 16 GTEx donors (25 frozen tissue samples total). This data is described on the new Single Cell Overview page and in

Resource Overview

Current Release (V8)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

Developmental GTEx

Explore GTEx

Browse	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
	By Tissue	Browse and search all data by tissue
	Histology Viewer	Browse and search GTEx histology images
Single Cell	Data Overview	Learn more about available single cell data
	Multi-Gene Single Cell Query	Browse and search single cell expression by gene and tissue

~17k healthy samples

Take Home Messages 1

- RNA-seq can be used as row counts and normalized (TPM, FPKM). See what you need for a specific algorithm!

- For QC of your samples at the sequence level – use **FastQC**. To combine results - **MultiQC**

- Expression-related data in transcriptomics are **strongly right-skewed**. Therefore:
  - For statistics use either precise distribution (negative binomial for RNA-seq) or work with log-transformed data
  - Use log-transformed data for exploratory analysis and visualization

- Several **large repositories of the data exist**. Before planning your experiments – make a search for existing data

2. Statistical Basics

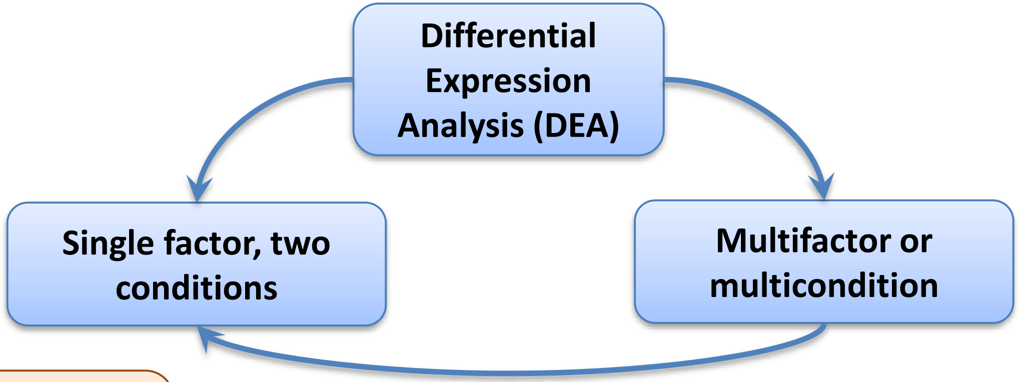
http://edu.modas.lu/transcript-seq/part3.html

see more here: http://edu.modas.lu/modas_dea/index.html

2. Statistical Basics

Questions

- Which genes have changes in **mean** expression level between conditions?
- How reliable are this observations (what is your p-value or FDR?)



Differential Expression Analysis (DEA)

Similar to t-test with Student's statistics: compare means

Single factor, two conditions

Multifactor or multicondition

Similar to ANOVA with Fisher's statistics: compare variances

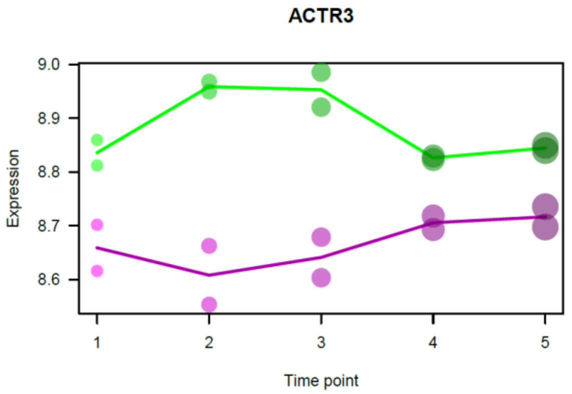
Post-hoc analysis

And do not forget about multiple hypotheses testing

What are those?..

- hypotheses
- p-values
- FDR
- t-test
- ANOVA

Example: two cell lines in time:



2.1. Hypothesis Testing

When statisticians would like to make a claim, they do this in the form of hypothesis testing. In hypothesis testing, we begin by making a tentative assumption about a population parameter, i.e. by formulation of a null hypothesis.

Null hypothesis

The hypothesis tentatively assumed true in the hypothesis testing procedure, H_0 .
For safety reasons, we assume a situation when nothing “interesting” happens as H_0

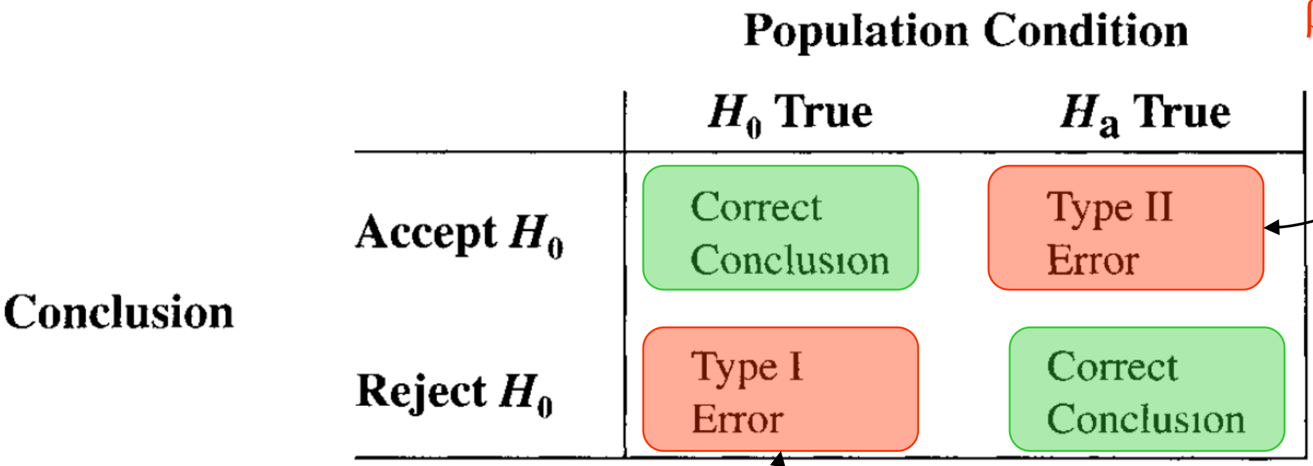
Alternative hypothesis

The hypothesis concluded to be true if the null hypothesis is rejected, H_a
H_a will be a situation when we see something unusual, which requires action

Hypotheses in a simplest case: comparing mean to a constant

One-tailed		Two-tailed
$H_0: \mu \leq$ const $H_a: \mu >$ const	$H_0: \mu \geq$ const $H_a: \mu <$ const	$H_0: \mu =$ const $H_a: \mu \neq$ const

2.1. Hypothesis Testing



		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Negative,
β error

False Positive,
α error

2.1. Hypothesis Testing: p-value

One-tailed test

A hypothesis test in which rejection of the null hypothesis occurs for values of the test statistic in one tail of its sampling distribution

$$H_0: \mu \geq \mu_0$$

$$H_a: \mu < \mu_0$$

A Trade Commission (TC) periodically conducts statistical studies designed to test the claims that manufacturers make about their products. For example, the label on a large can of Hilltop Coffee states that the can contains 3 pounds of coffee. The TC knows that Hilltop's production process cannot place exactly 3 pounds of coffee in each can, even if the mean filling weight for the population of all cans filled is 3 pounds per can. However, as long as the population mean filling weight is at least 3 pounds per can, the rights of consumers will be protected. Thus, the TC interprets the label information on a large can of coffee as a claim by Hilltop that the population mean filling weight is at least 3 pounds per can. We will show how the TC can check Hilltop's claim by conducting a lower tail hypothesis test.

μ_0 = 3 lbm

Suppose a sample of n = 36 coffee cans is selected. From the previous studies, it's known that σ = 0.18 lbm

2.1 Hypothesis Testing: p-value

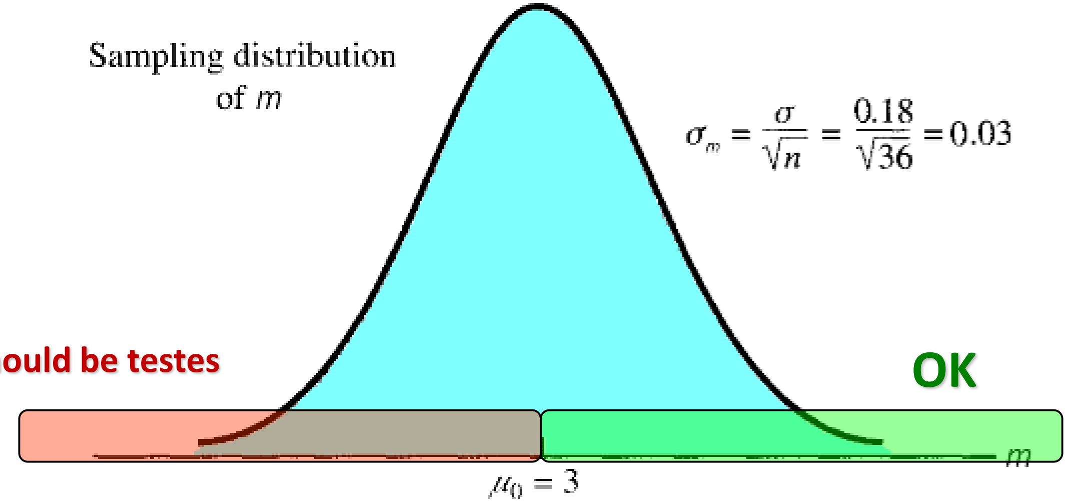
μ_0 = 3 lbm

Suppose a sample of *n* = 36 coffee cans is selected and *m* = 2.92 is observed. From the previous studies, it's known that σ = 0.18 lbm

H_0: $\mu \geq 3$ **no action**

H_a: $\mu < 3$ **legal action**

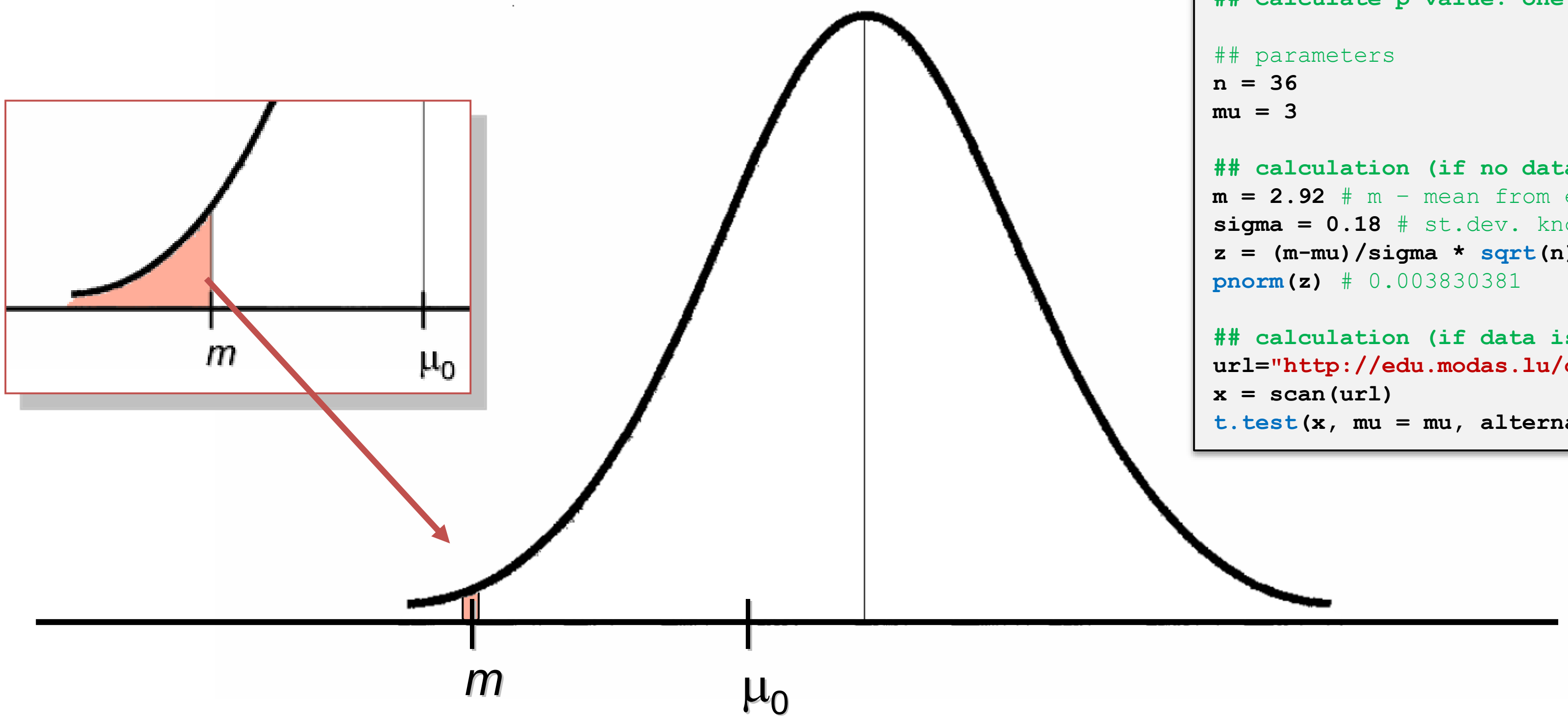
Let's say: in the extreme case, when μ=3, we would like to be 99% sure that we make no mistake, when starting legal actions against Hilltop Coffee. It means that selected significance level is **α = 0.01**


Sampling distribution
of m
$\sigma_m = \frac{\sigma}{\sqrt{n}} = \frac{0.18}{\sqrt{36}} = 0.03$
should be testes
OK
m
$\mu_0 = 3$

2.1 Hypothesis Testing: p-value

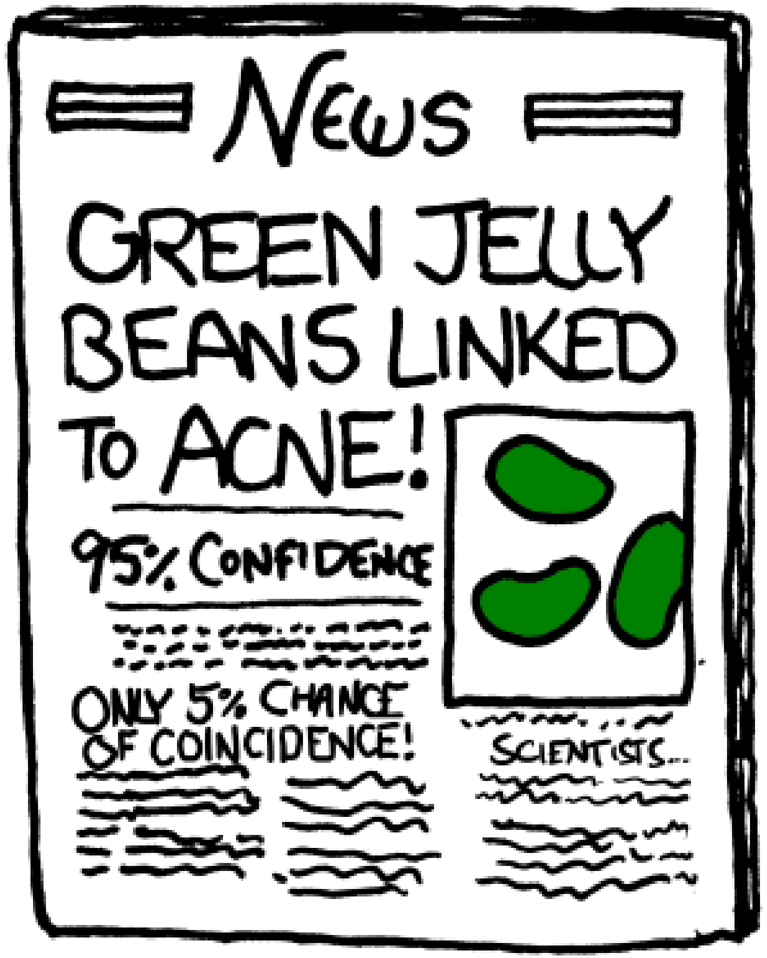
Let’s find the probability of observation *m* for all possible $\mu \geq 3$. We start from an extreme case (μ=3) and then probe all possible $\mu > 3$. See the behavior of the small probability area around measured *m*. What you will get if you summarize its area for all possible $\mu \geq 3$?

```
## Calculate p-value: one sample mean

## parameters
n = 36
mu = 3

## calculation (if no data available)
m = 2.92 # m - mean from experiment
sigma = 0.18 # st.dev. known beforehand
z = (m-mu)/sigma * sqrt(n)
pnorm(z) # 0.003830381

## calculation (if data is available)
url="http://edu.modas.lu/data/txt/coffee.txt"
x = scan(url)
t.test(x, mu = mu, alternative="less")
```



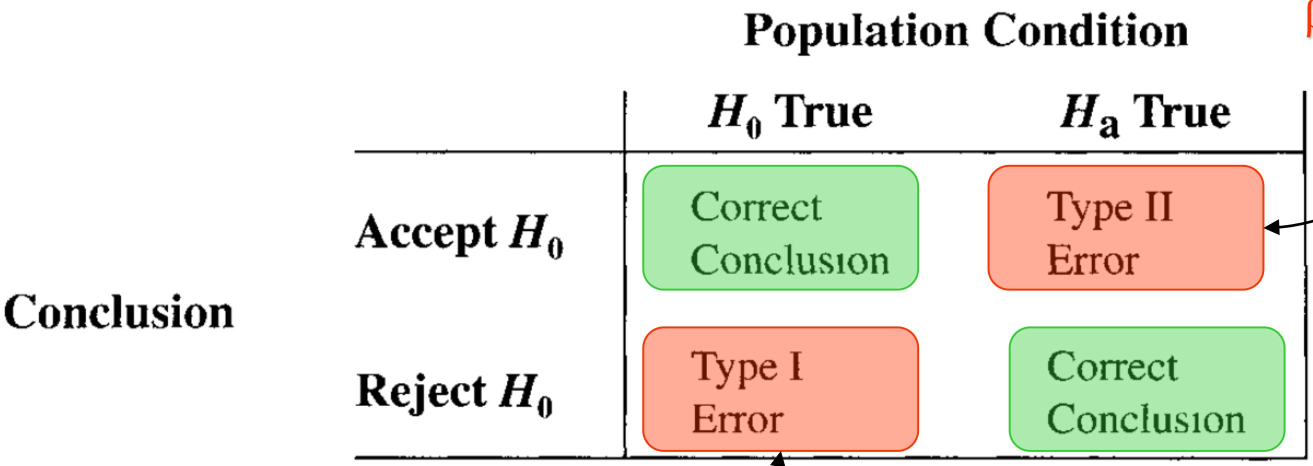
P(*m*) for all possible $\mu \geq \mu_0$ is equal to *P*(*x*<*m*) for an extreme case of $\mu = \mu_0$

2.2. Hypothesis Testing: Multiple Testing


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

```r
## Why do we need multiple testing correction?

## 1. Generate a random matrix: 1000 genes x 6 samples
X = matrix(rnorm(6*1000),nrow=1000,ncol=6)
rownames(X) = paste0("gene",1:1000)

## 2. Assume col 1,2,3 - exp, 4,5,6 - ctrl
colnames(X) = c("exp1","exp2","exp3","ctrl1","ctrl2","ctrl3")

## 3. Do a t.test for each "gene" (slow, but who cares :)
pv = NULL
for (i in 1:nrow(X))
    pv[i] = t.test(X[i,1:3],X[i,4:6])$p.value

table(pv < 0.05) # around 50 false positives are expected

## do FDR adjustment
fdr = p.adjust(pv,"fdr")
table(fdr < 0.05)
```

2.2. Multiple Testing



		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Negative,
β error

False Positive,
α error

Probability of an error in a multiple test, when α=0.05: $1-(0.95)^{\text{number of comparisons}}$

2.2. Multiple Testing: FDR

False discovery rate (FDR)

FDR control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors).

		Population Condition		
		H_0 is TRUE	H_0 is FALSE	Total
Conclusion	Accept H_0 (non-significant)	$\boldsymbol{U}$	$\boldsymbol{T}$	$m - R$
	Reject H_0 (significant)	$\boldsymbol{V}$	$\boldsymbol{S}$	R
	Total	m_0	$m - m_0$	$\boldsymbol{m}$

$$FDR = E\left(\frac{V}{V+S}\right)$$

2.2. Multiple Testing: FDR and FWER

False Discovery Rate: Benjamini & Hochberg

Assume we need to perform *m* = 100 comparisons, and select maximum **FDR = α = 0.05**

$$FDR = E\left(\frac{V}{V+S}\right)$$

Expected value for FDR < α if

$$P_{(k)} < \frac{k}{m}\alpha \Rightarrow \frac{mP_{(k)}}{k} < \alpha$$

```
p.adjust(pv, method="fdr")
```

Theoretically, the sign should be “≤”.
But for practical reasons it is replaced by “<“

Familywise Error Rate (FWER)

Bonferroni – simple, but too stringent, not recommended

$$mP_{(k)} < \alpha$$

Holm-Bonferroni – a more powerful, less stringent but still universal FWER

$$\left(m+1-k\right)P_{(k)} < \alpha$$

```
p.adjust(pv, method="holm")
```

2.3. Linear Models

Many conditions

We have measurements for 5 conditions. Are the means for these conditions equal?

Many factors

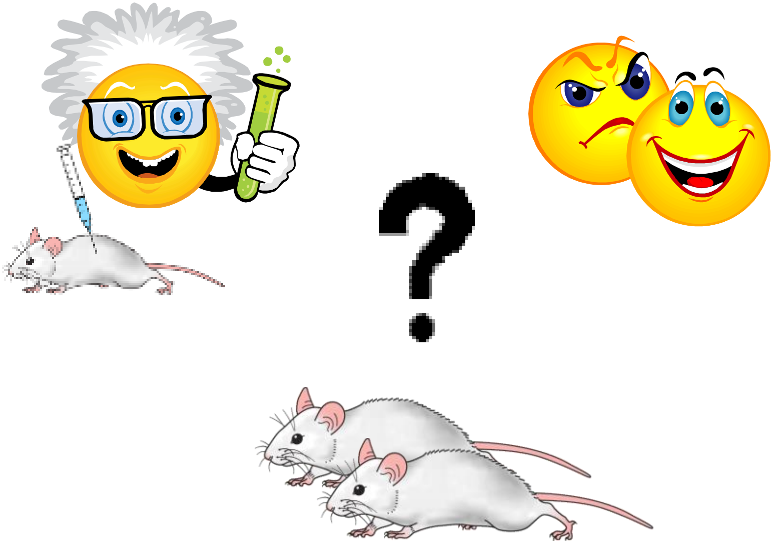
We assume that we have several factors affecting our data. Which factors are most significant? Which can be neglected?

ANOVA
example from Partek™

If we would use pairwise comparisons, what will be the probability of getting error?

Number of comparisons: $C_2^5 = \frac{5!}{2!3!} = 10$

Probability of an error: $1 – (0.95)^{10}$ = 0.4



2.3. Linear Models: ANOVA

As part of a long-term study of individuals 65 years of age or older, sociologists and physicians at the Wentworth Medical Center in upstate New York investigated the relationship between geographic location and depression. A sample of 60 individuals, all in reasonably good health, was selected; 20 individuals were residents of Florida, 20 were residents of New York, and 20 were residents of North Carolina. Each of the individuals sampled was given a standardized test to measure depression. The data collected follow; higher test scores indicate higher levels of depression.

Q: Is the depression level same in all 3 locations?

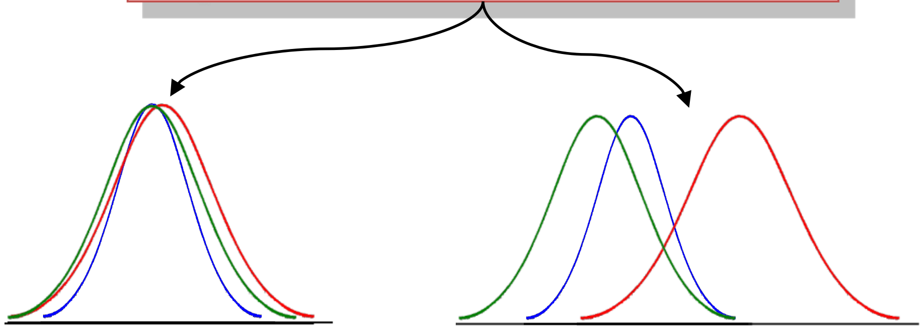
`depression.txt`

1. Good health respondents

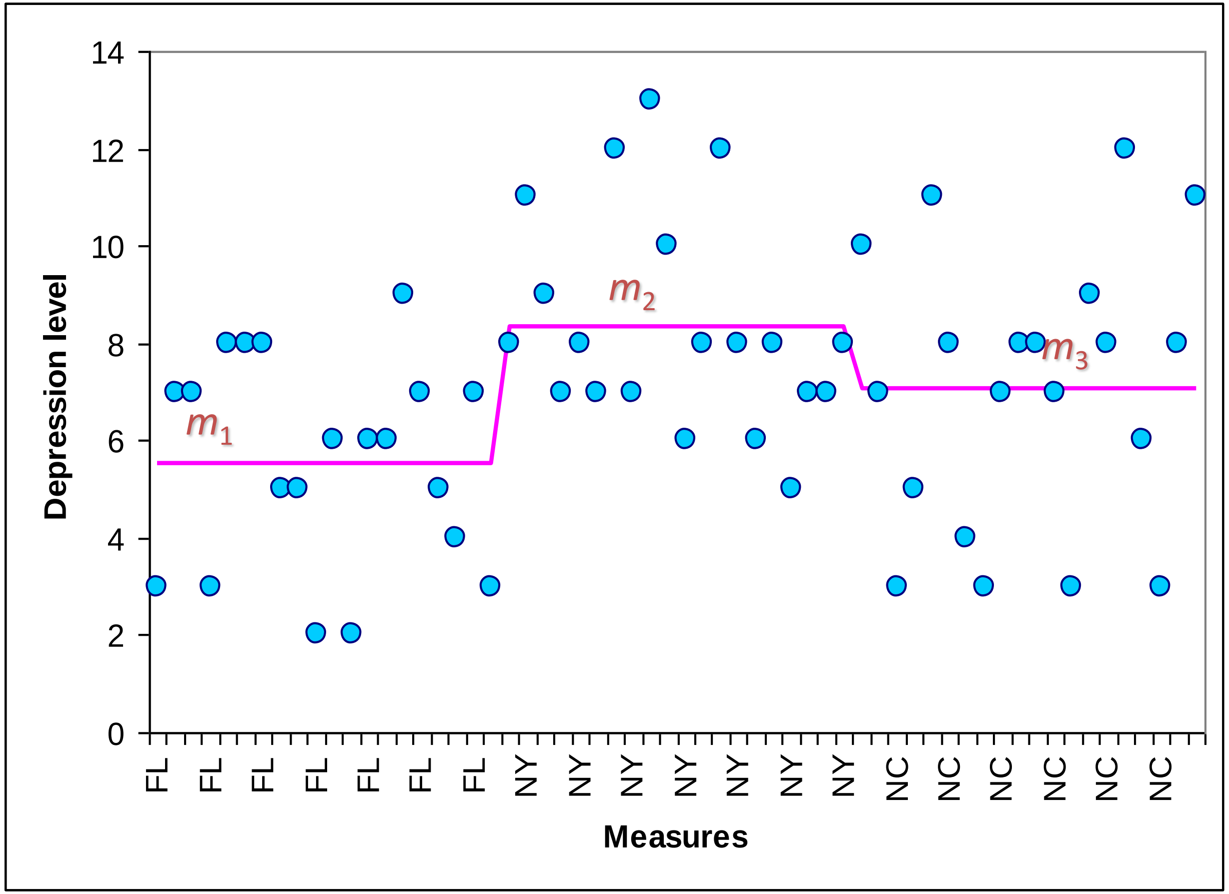
Florida	New York	N. Carolina
3	8	10
7	11	7
7	9	3
3	7	5
8	8	11
8	7	8
…	…	…

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal



2.3. Linear Models: ANOVA



H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal

Please see the code and explanation online: http://edu.modas.lu/transcript-seq/part3.html

```
## load data (*)
Dep = read.table("depression2.txt",
          header=T, sep="\t", as.is=FALSE)
str(Dep)

## run 1-factor ANOVA
DepGH = Dep[Dep$Health == "good",]
res1 = aov(Depression ~  Location, DepGH)
summary(res1)
TukeyHSD(res1)

## run 2-factor ANOVA
res2 = aov( Depression ~
            Location + Health + Location*Health,
            Dep)
summary(res2)
TukeyHSD(res2)
```

(*) http://edu.modas.lu/data/txt/depression2.txt

Take Home Messages 2

- When doing **multiple hypothesis testing** and selecting only those elements which are significant – always use FDR (or other, like FWER) correction!
  - the simplest correction – multiply the p-value by the number of genes. Is it still significant? Use FDR (Benjamini-Hochberg) or FWER (Holm)

- DEA detects the genes which have **changed mean** gene expression between condition
  - => The more data you have, the smaller differences you will be able to see

- Several factors can be taken into account in **ANOVA** approach. This will give you insight into the significance of each experimental factor but at the same time will correct batch effects and allow you to answer complex questions (remember shoes affecting ladies…).

3. Statistics for RNA-seq

http://edu.modas.lu/transcript-seq/part4.html

see more here: http://edu.modas.lu/modas_dea/index.html

3.1. Linear Models for Transcriptomics Data

$$Y_{ij} = \mu_i + A_j + B_j + A_j * B_j + \epsilon_{ij}$$

i – gene index
j – sample index

$A_j * B_j$ – effect which cannot be explained by superposition A and B

Limma – R package for DEA in microarrays or RNA-seq based on linear models.
It is similar to t-test / ANOVA but uses all available data for variance estimation, thus it has higher power when the number of replicates is limited. It assumes a normal distribution of values for the gene between replicates. **Apply it to normalized, log-transformed counts.**

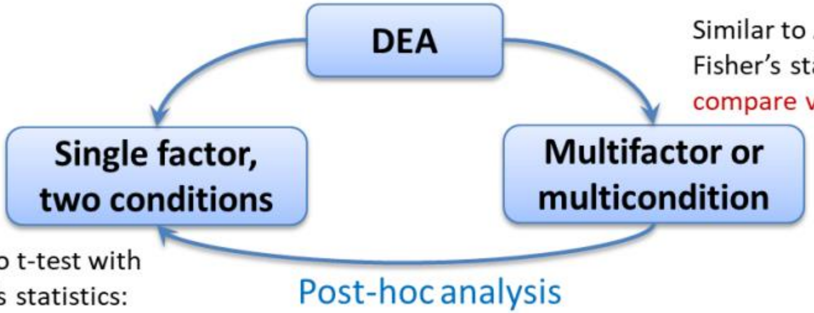
edgeR – R package for DEA in RNA-Seq, based on linear models and negative binomial distribution of counts. Apply to raw counts!
Better noise model results in higher power detecting differentially expressed genes. It assumes a negative-binomial distribution of values for the gene between replicates.

DESeq2 – another R package for DEA in RNA-Seq, based on the negative binomial distribution of counts. DESeq2 is the most sensitive among others. Apply to raw counts!
Better noise model results in higher power detecting differentially expressed genes. It assumes a negative-binomial distribution of values for the gene between replicates.

3.2. DEA: Preparing for *limma, edgeR, DESeq2*

Questions

- Which genes have changes in **mean** expression level between conditions?
- How reliable are this observations



Similar to ANOVA with Fisher's statistics: compare variances

Similar to t-test with Student's statistics: compare means

Here we should define contrasts:
"condition1 – condition2"

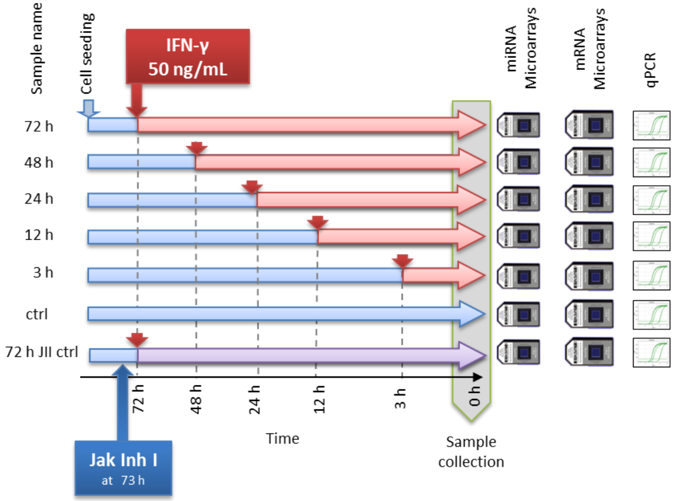
condition 1 – experimental group
condition 2 – control group

```
## install packages
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("limma")
BiocManager::install("edgeR")
BiocManager::install("DESeq2")

## if you wish, you can use my simple warp-up
source("http://r.modas.lu/LibDEA.r")
DEA.limma
DEA.edgeR
DEA.DESeq
```

3.2. DEA: Time Series Experiment

```
## Let's use limma for a time-series experiment
## load the data that are in annotated text format
source("http://r.modas.lu/readAMD.r")
mRNA = readAMD("http://edu.modas.lu/data/txt/mrna_ifng.amd.txt",
               stringsAsFactors=TRUE,
               index.column="GeneSymbol",
               sum.func="mean")
str(mRNA)

## attach library with warp-up functions
source("http://r.modas.lu/LibDEA.r")

## DEA: the most variable genes (by F-statistics)
ResF = DEA.limma(data = mRNA$X, group = mRNA$meta$time)
genes = order(ResF$FDR)[1:100]  ## select top 100 genes
pheatmap(mRNA$X[genes,], cluster_col=FALSE, scale="row",
  fontsize_row=2, fontsize_col=10, cellwidth=15,
  main="Top 100 significant genes (F-stat)")

## DEA: genes differentially expressed (by moderated t-test)
Res24 = DEA.limma(data = mRNA$X,
                  group = mRNA$meta$time,
                  key0="T00",key1="T24")
## volcano plot
plotVolcano(Res24,thr.fdr=0.01,thr.lfc=1)
genes = order(Res24$FDR)[1:100]  ## select top 100 genes
samples = grep("T00|T24",mRNA$meta$time) ## select T00,T24 sampl.
pheatmap(mRNA$X[genes,samples],cluster_col=FALSE,scale="row",
  fontsize_row=2, fontsize_col=10, cellwidth=15,
  main="Top 100 significant genes T24-T00 (moderated t-stat)")
```

Experiment: A375 cells stimulated by IFNg

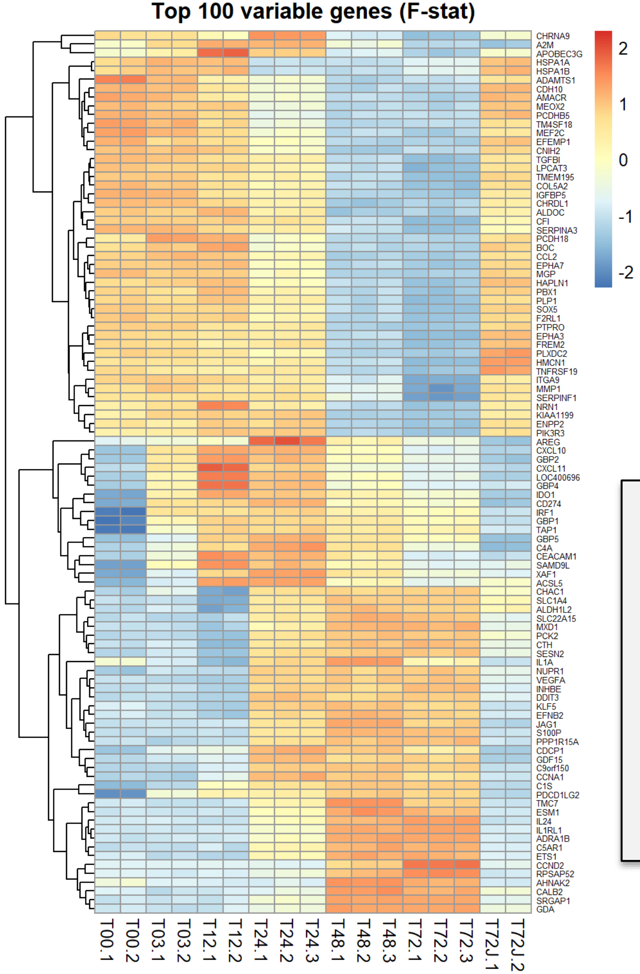


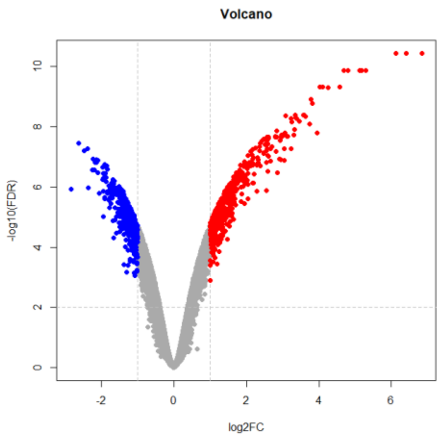
Annotation – Metadata – Data format

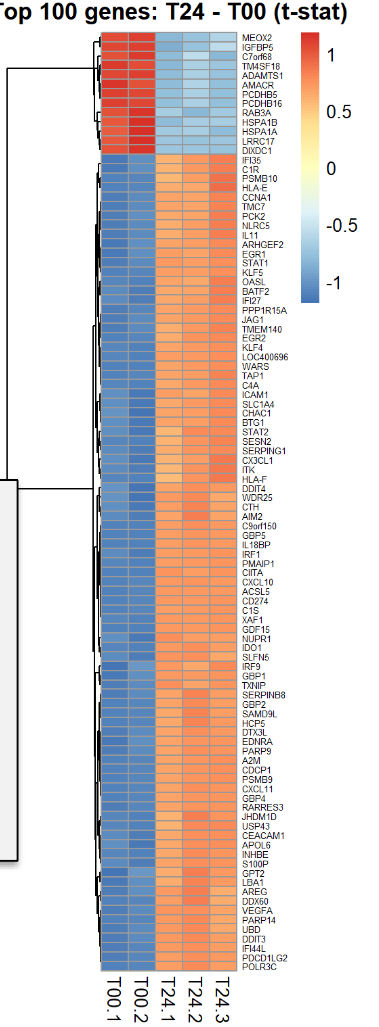
#factor1			control	treated	control	treated
#factor2			rep1	rep1	rep2	rep2
feature_id	**anno1**	**anno2**	**sample_1**	**sample_2**	**sample_3**	**sample_4**
ENSG00000141510	TP53	coding	7.3	7.5	6.8	7.4
ENSG00000115415	STAT1	coding	5.3	8.2	4.9	7.6
ENSG00000229807	XIST	non-coding	3.1	3.5	3.2	3.3
…	…	…	…	…	…	…

See more at http://edu.modas.lu/modas_dea/part3.html

3.2. DEA: Time Series Experiment


Top 100 variable genes (F-stat)



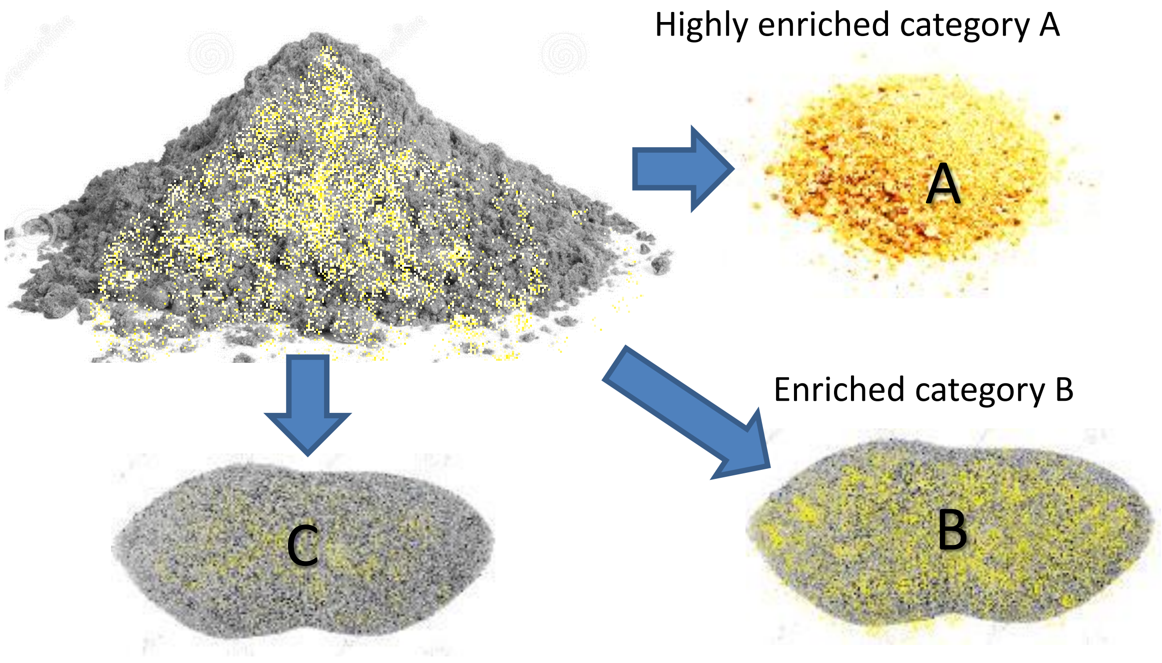

Top 100 genes: T24 - T00 (t-stat)

```
## Save results

## save the most variable genes (by F-statistics)
write.table(ResF[ResF$FDR<0.0001,],file = "DEA_F.txt",
            col.names=NA, sep="\t", quote=FALSE)
## save significant genes T24-vs-T00
write.table(Res24[Res24$FDR<0.001 & abs(Res24$logFC)>1,],
            file = "DEA_T24-T00.txt",
            col.names=NA, sep="\t", quote=FALSE)
## save gene list (response at 24 h of IFNg treatment)
write(Res24[Res24$FDR<0.0001,1],file="genes24.txt")
```

Please, investigate the results. Submit any list to the functional annotation tool **Enrichr**
https://maayanlab.cloud/Enrichr/

3.3. Functional Annotation: Enrichment

Are interesting genes over-represented in a subset corresponding to some biological process?



Highly enriched category A

Enriched category B

No enrichment in C

Method of the analysis:
Fisher's exact test

Someone grabs "randomly" 20 balls from a box with 50x 🔴 and 50x 🟢

🔴🔴🔴🔴🔴🔴🔴🔴🔴🔴🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢
🔴🔴🔴🔴🔴🔴🔴🔴🔴🔴🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢
🔴🔴🔴🔴🔴🔴🔴🔴🔴🔴🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢
🔴🔴🔴🔴🔴🔴🔴🔴🔴🔴🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢
🔴🔴🔴🔴🔴🔴🔴🔴🔴🔴🟢🟢🟢🟢🟢🟢🟢🟢🟢🟢

How surprised will you be if he grabbed

🔴🔴🔴🟢🔴🔴🔴🟢🔴🔴🔴🔴🔴🔴🔴🔴🔴🟢🔴🔴

(17 red , 3 green)

3.3. Gene Over-representation Analysis

Fisher’s exact test: based on hypergeometrical distributions

Hypergeometrical: distribution of objects taken from a “box”, without putting them back

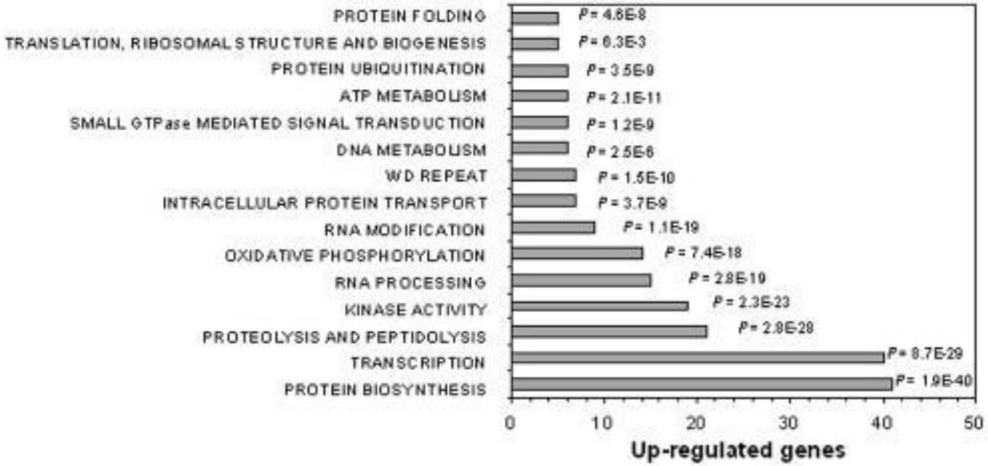
$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N: total number of genes

M: total number of genes annotated with this term

n: number of genes in the list

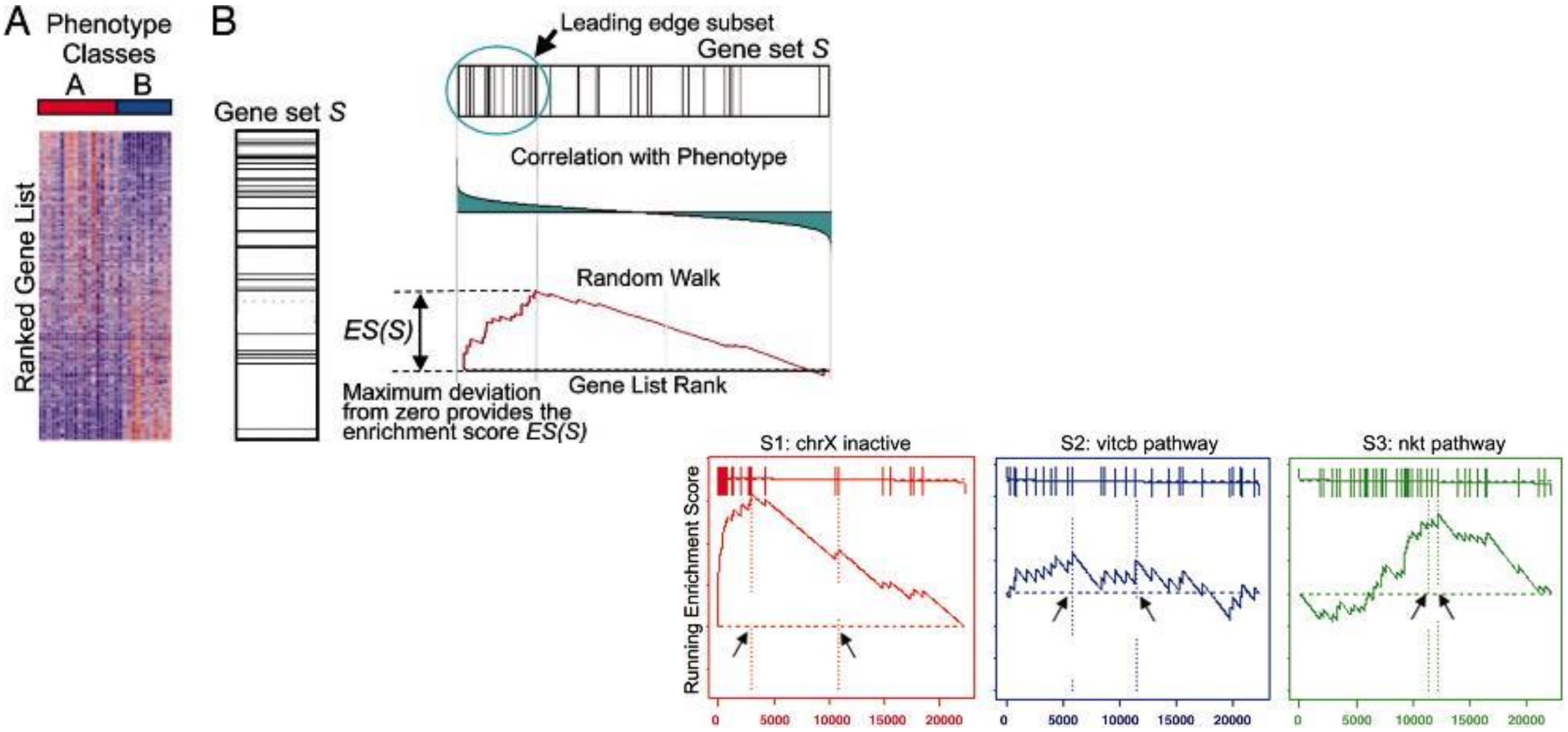
k: number of genes in the list annotated with this term



$$C_k^n = C_n^k = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

3.3. Gene Set Enrichment Analysis

Is the direction of all genes in a category random?



Take Home Messages 3

- If you are looking at a multi-factor / multi-treatment experiment, you may check the variable genes (F-statistics based) first, and then go for the contrasts.

- To find the biological meaning of the significantly regulated genes, please use enrichment analysis methods linking known functional groups of genes to DEA results.

- Enriched categories are usually more robust than individual genes. If you have no significant genes – check gene sets by GSEA.

Enrichr
https://maayanlab.cloud/Enrichr/

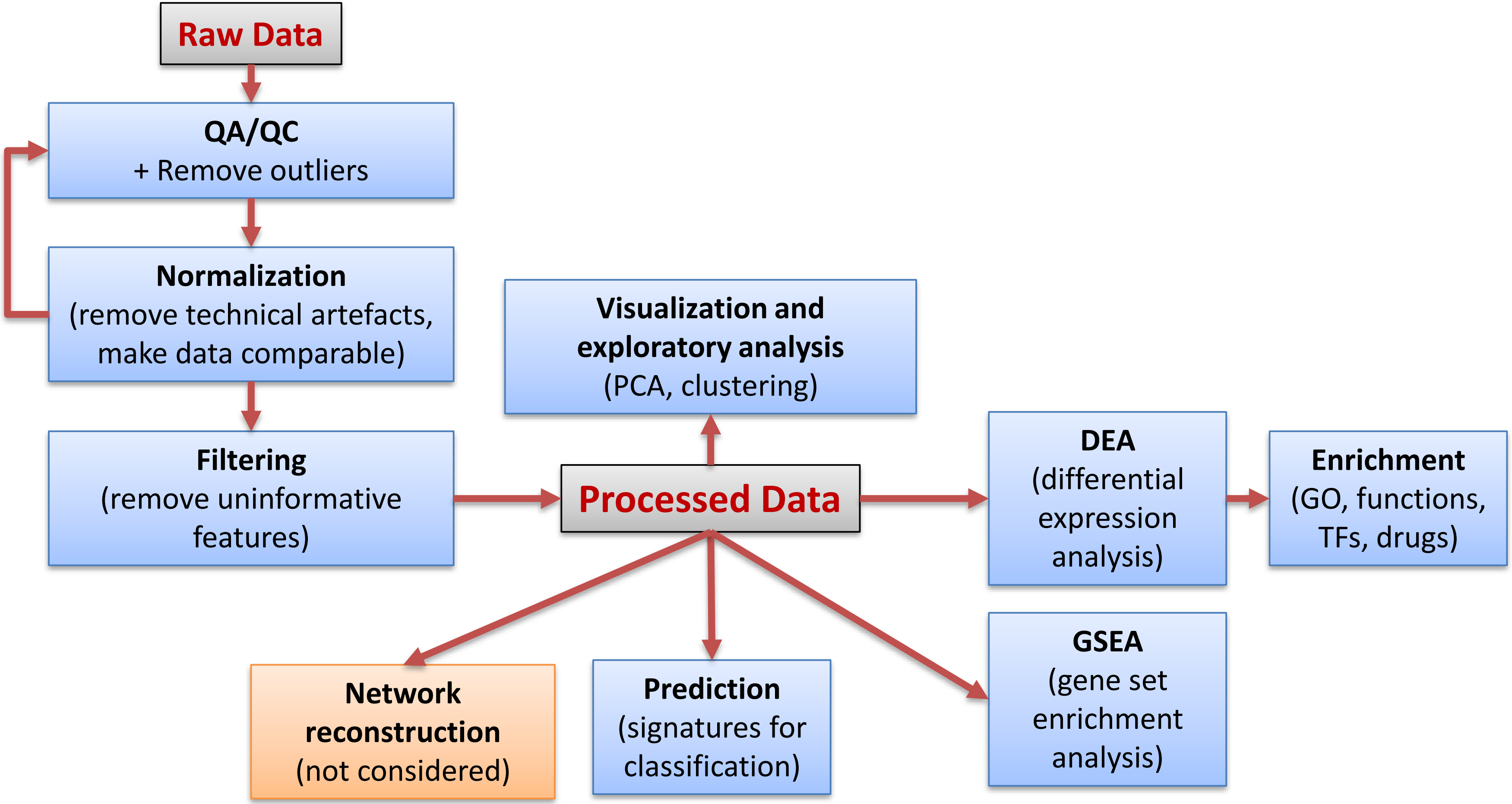
David
https://david.ncifcrf.gov/

Reactome
https://reactome.org/

String
https://string-db.org/

WikiPathways
https://wikipathways.org/

Summary



Raw Data

QA/QC
+ Remove outliers

Normalization
(remove technical artefacts, make data comparable)

Filtering
(remove uninformative features)

Processed Data

Visualization and exploratory analysis
(PCA, clustering)

DEA
(differential expression analysis)

Enrichment
(GO, functions, TFs, drugs)

GSEA
(gene set enrichment analysis)

Prediction
(signatures for classification)

Network reconstruction
(not considered)