UNIVERSITÉ DU LUXEMBOURG

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Lecture 13

Data Analysis in Transcriptomics

dr. P. Nazarov

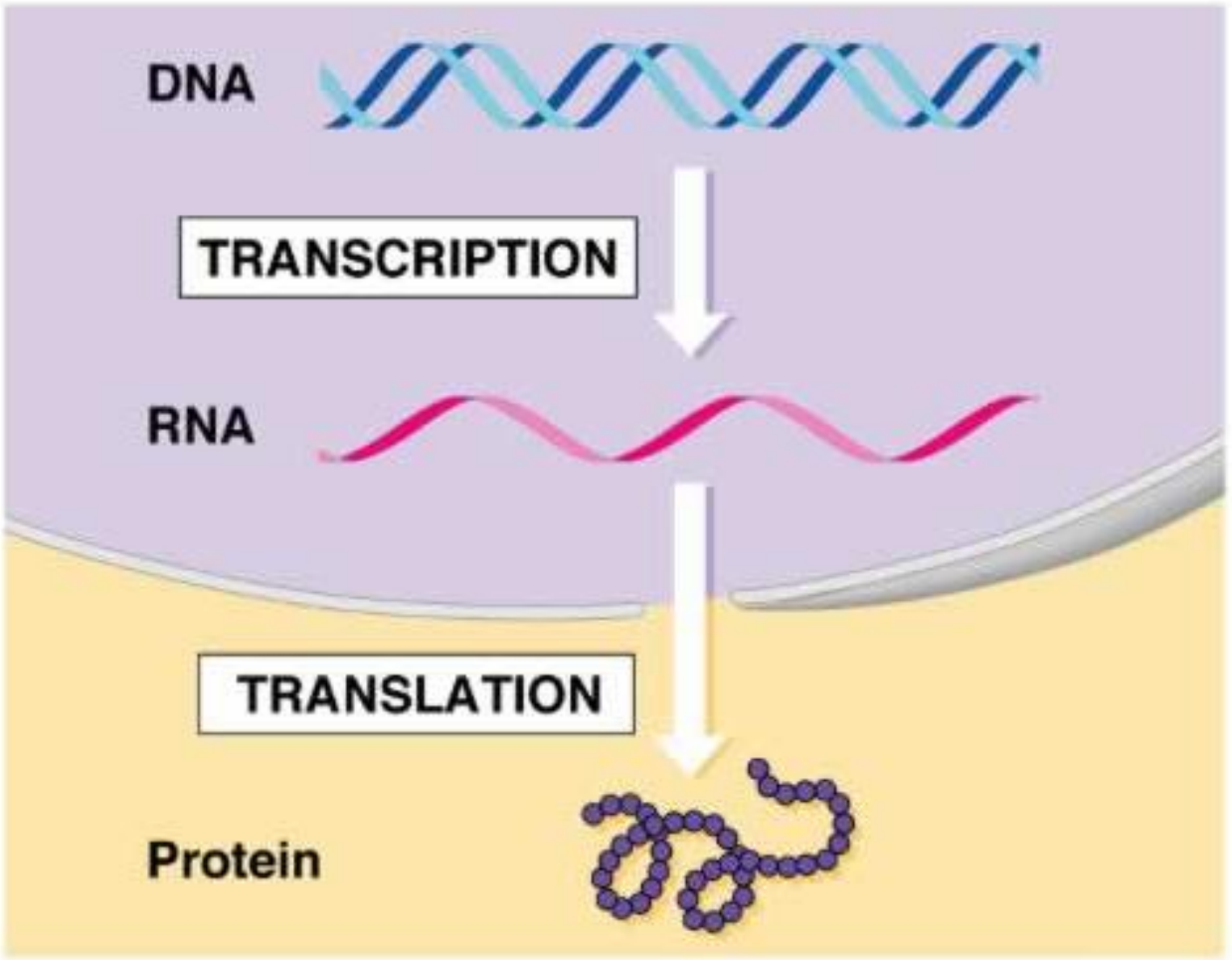
petr.nazarov@lih.lu

26-05-2017

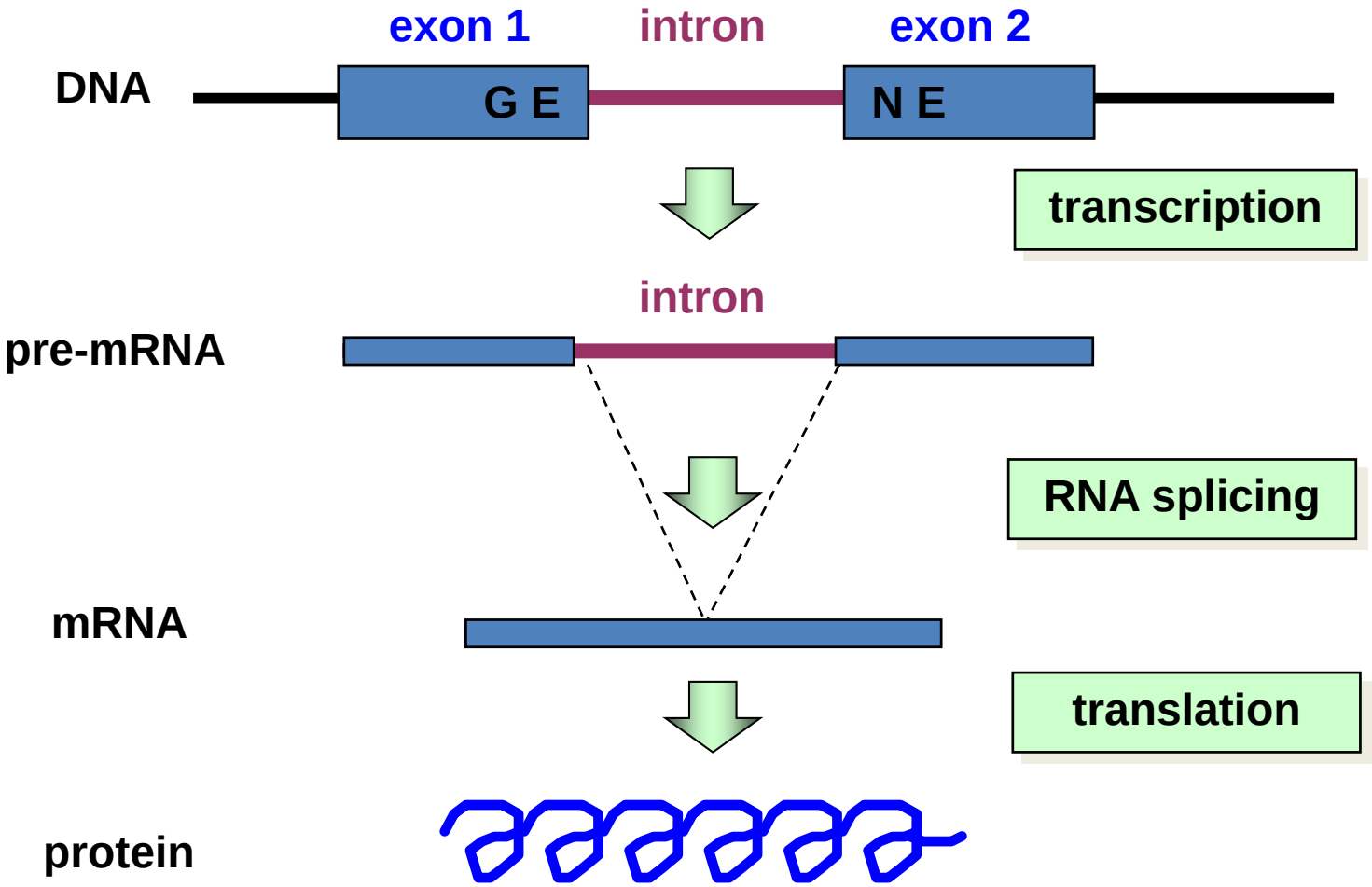
Outline

- **Public data repositories**
  - GEO, ArrayExpress
  - TCGA
- **Microarray data**
  - 2-,1-color arrays
  - normalization
- **RNA-Seq**
- **Exploratory data analysis**
  - PCA
  - clustering
- **Differential expression analysis**
  - multiple hypotheses
  - linear models
- **Classification and marker genes**
- **Enrichment analysis**

Basic Expression Scheme

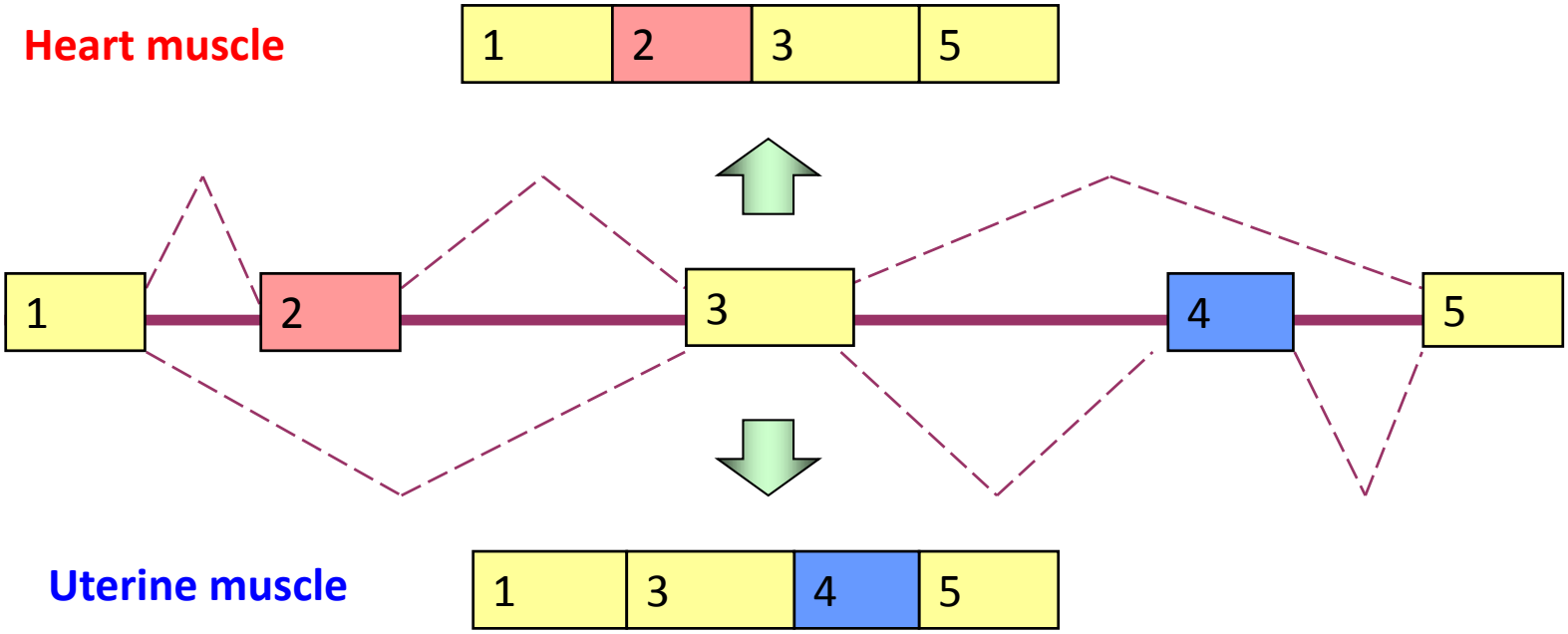


Basic Expression Scheme



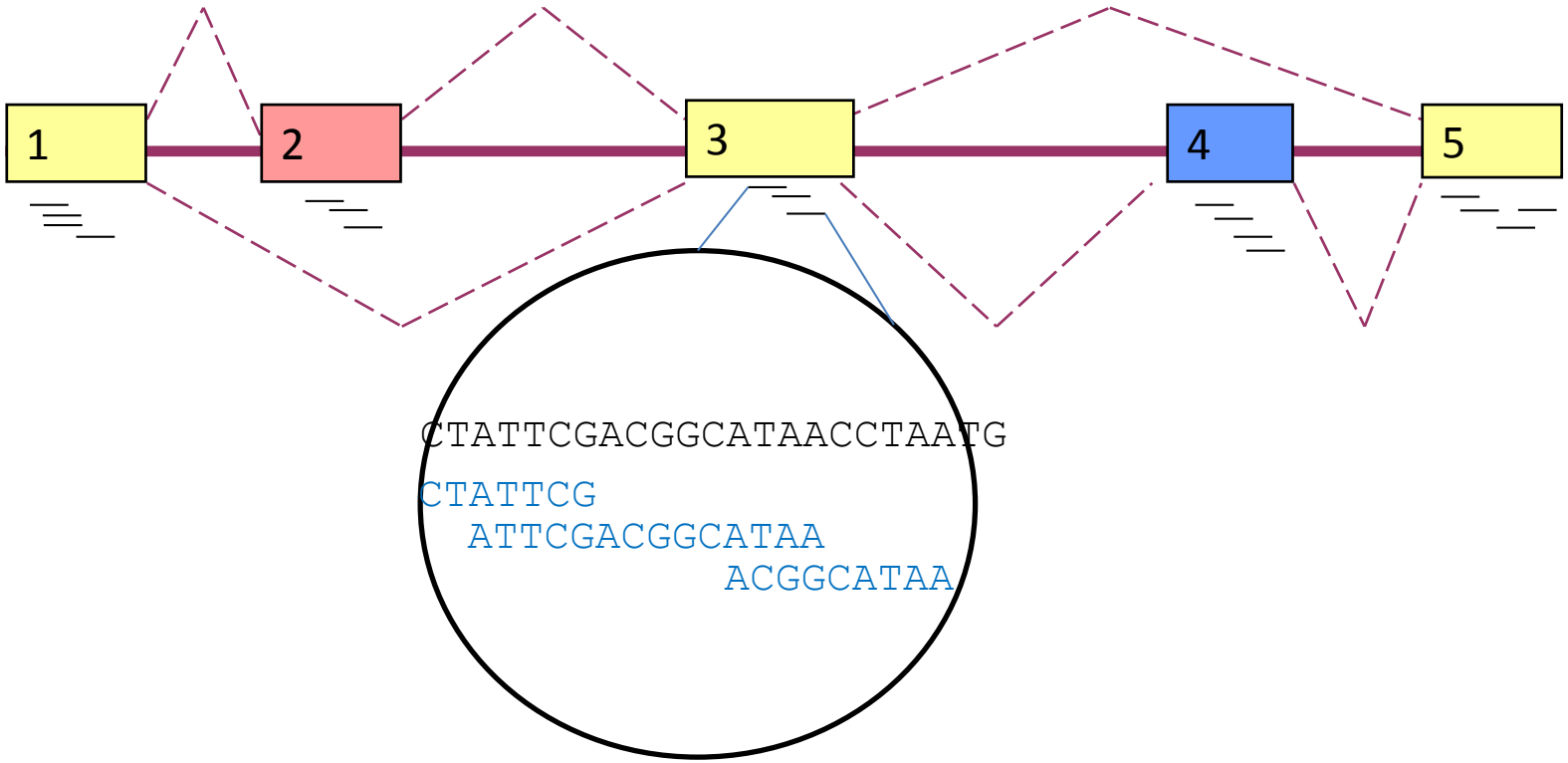
Alternative Splicing

Multiple introns may be spliced differently in different circumstances, for example in different tissues.



Thus one gene can encode more than one protein. The proteins are similar but not identical and may have distinct properties – an important feature for complex organisms

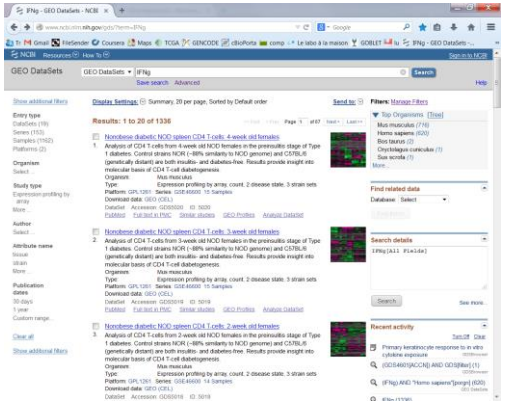
Probes


1
2
3
4
5
CTATTCGACGGCATAACCTAATG
CTATTCG
ATTCGACGGCATAA
ACGGCATAA

Data Overview

Public Repositories

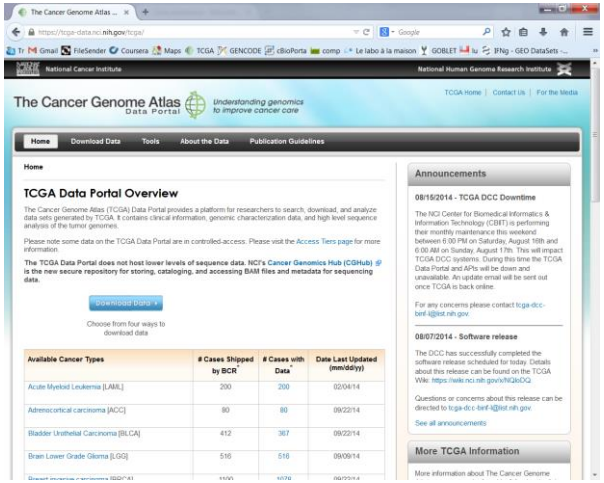
GEO: http://www.ncbi.nlm.nih.gov/gds



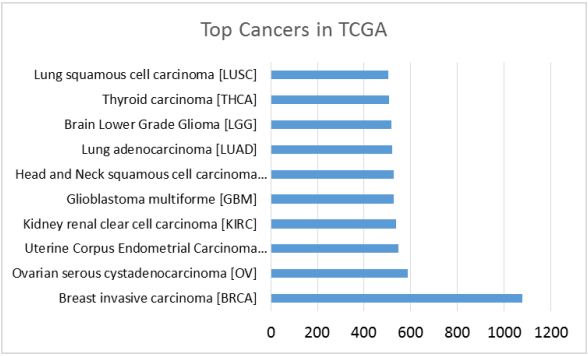
Browse Content

Repository Browser	
DataSets:	3847
Series:	50810
Platforms:	13387
Samples:	1237318

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

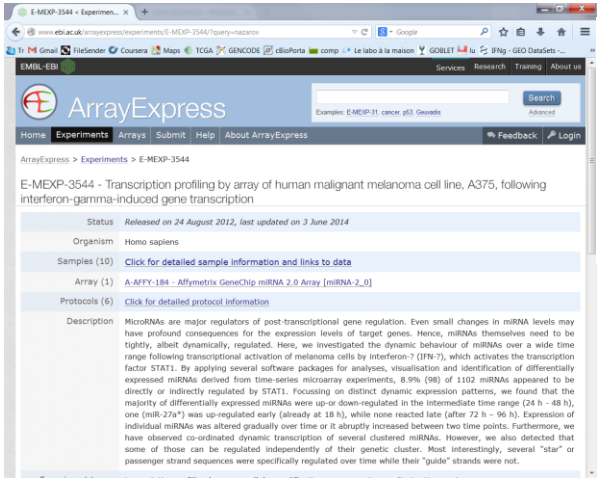


Sep 2014 – more then 10k patients



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

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



Data Content

Updated today at 06:00

- 52801 experiments
- 1555904 assays
- 24.99 TB of archived data

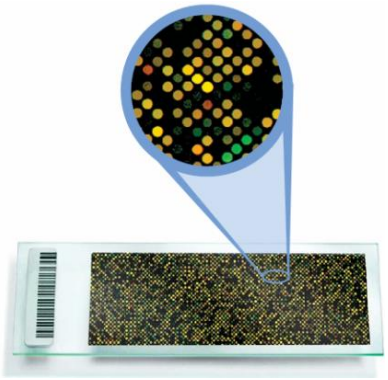
Data for our course: http://edu.sablab.net/transcript

Microarray Data

Types of Microarrays

Two-color Arrays (2C)

- Agilent full genome
- Thematic arrays



Pro

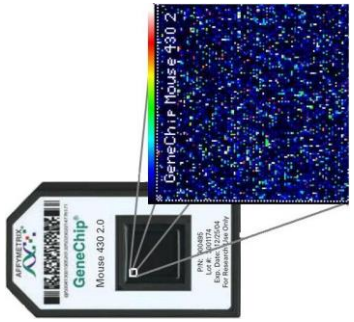
- Direct comparison
- Less sensitive to inaccuracies of spotting

Con

- Dye effects: need for “dye-swaps”
- Non-flexibility in analysis

One-color Arrays (1C)

- Affymetrix GeneChip
- Affymetrix Exon
- Affymetrix mRNA



Pro

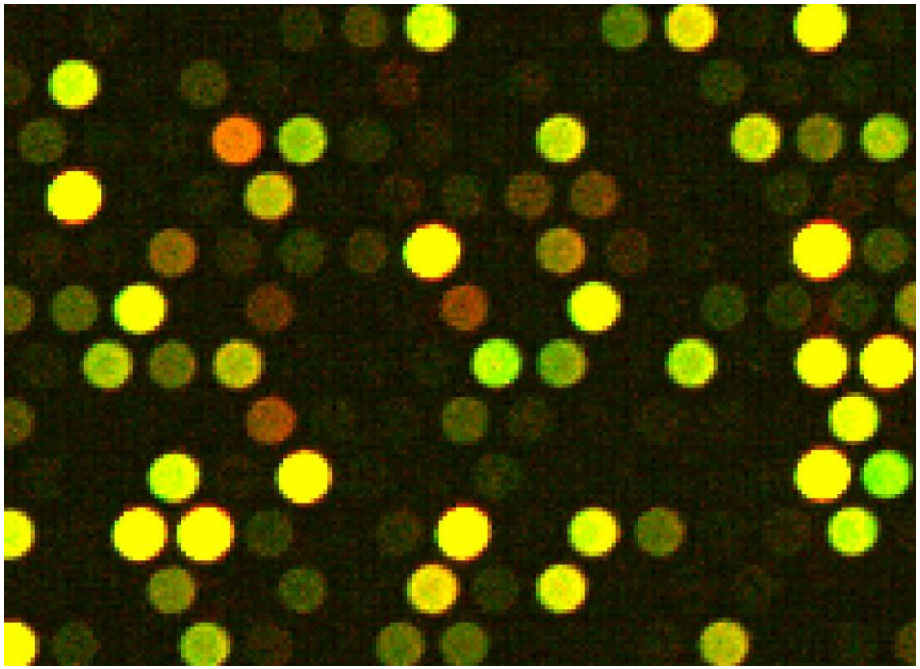
- Flexible analysis
- High level of standardization

Con

- Price

Microarray Data

Two-color Arrays



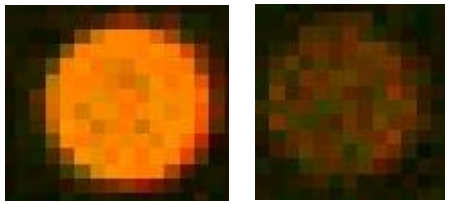
Measure: red (R) and green (G) fluorescence

Estimate: background fluorescence R_{bg} , G_{bg}

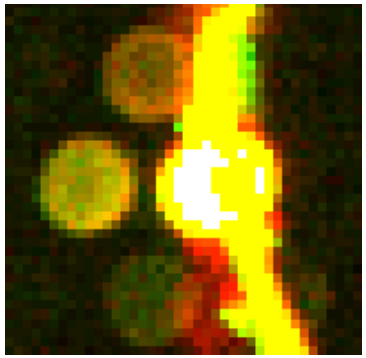
$$Log\ Ratio = logFC = M = log_2\left(\frac{R-R_{bg}}{G-G_{bg}}\right) =$$

$$= log_2(R - R_{bg}) - log_2(G - G_{bg})$$

$$Log\ Intencity = A = \frac{1}{2} log_2\left((R - R_{bg}) + (G - G_{bg})\right)$$



LogFC ≈ 2



Advanced image analysis and corrections are needed MAIA

http://bioinfo-out.curie.fr/projects/maia/index.php

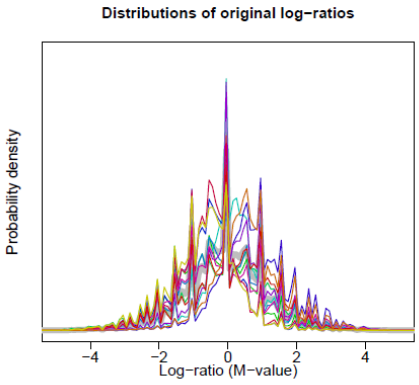
Solutions

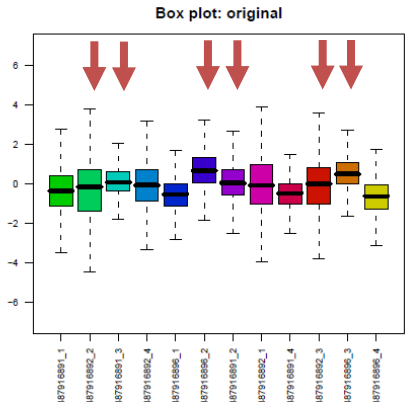
- Several spots with the same probe sequence
- Quantify each spot by a set of parameters
- Find an optimal rule to accept the spots
- Remove bad spots from further analysis

Normalization of Two-color Arrays

Linear effects b/w samples

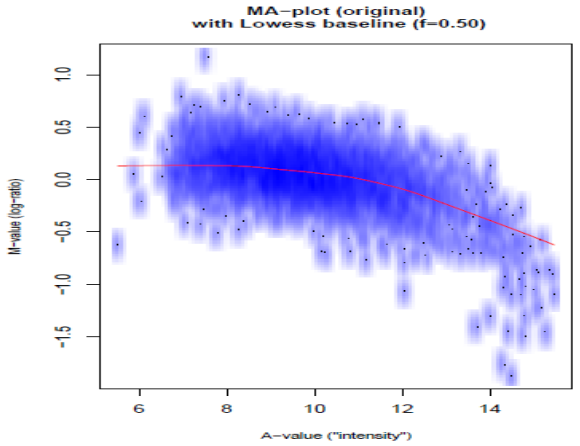
difference in concentration or hybridization efficiency



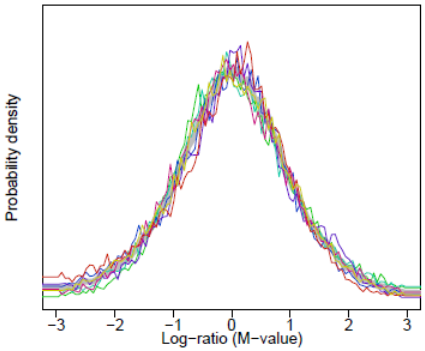


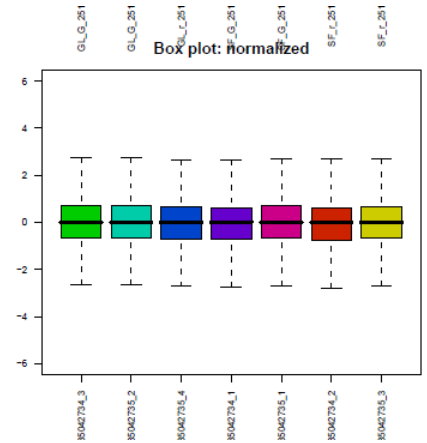
Non-linear dye-effect

photodegradation, radiationless energy transfer, quenching

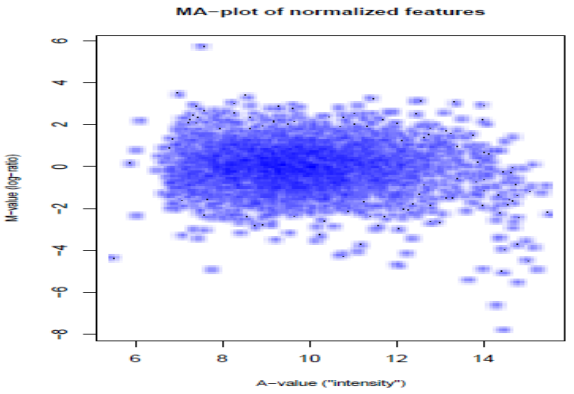


Solution. Linear centring/scaling





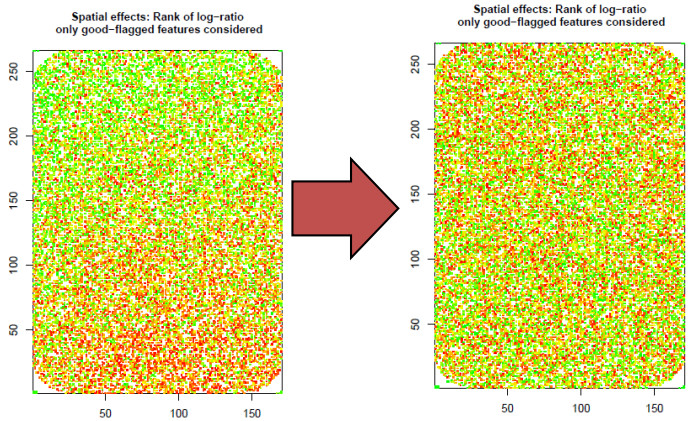
Solution. Lowess (loess) correction



Normalization of Two-color Arrays

Spatial effects

due to technological problems

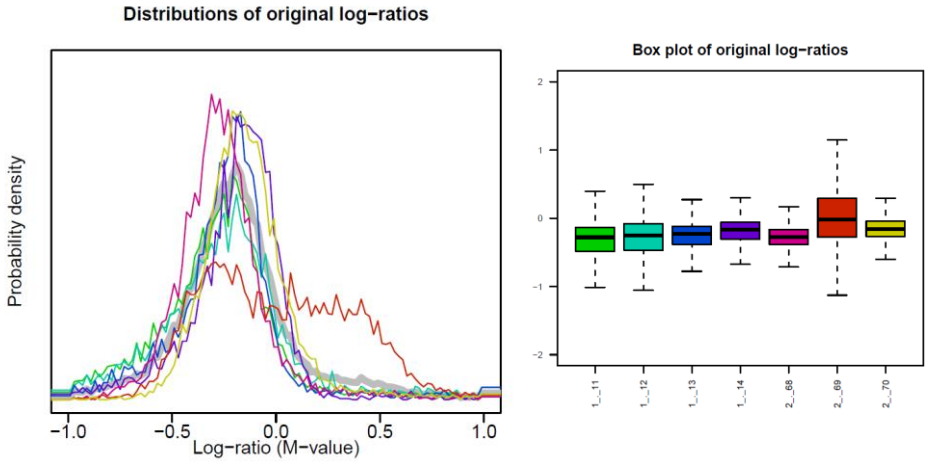


Solution. Spatial normalization

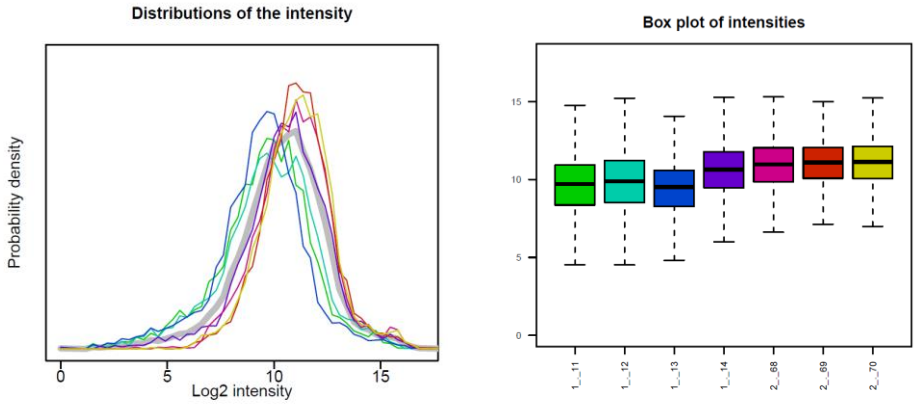
- Using spikes (Agilent)
- Using numerical methods to estimate 2D profile

Two-color Array Data Overview

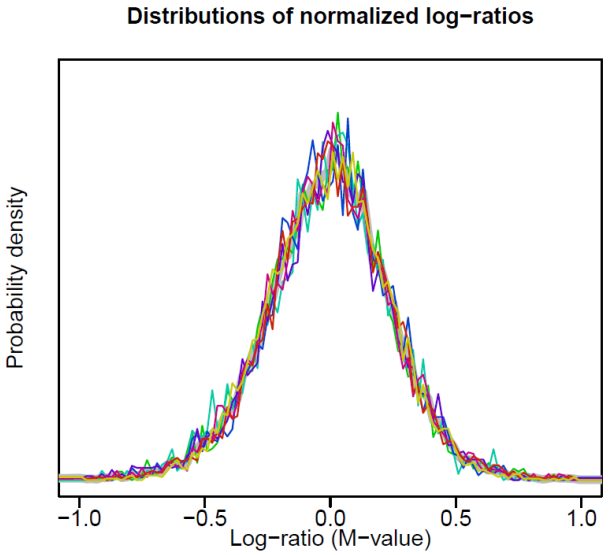
Original LogRatio (logFC)



Original Log Intensity



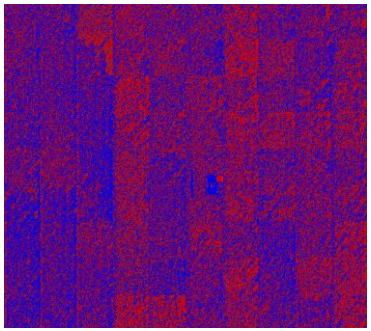
Normalized Log Ratio (logFC)



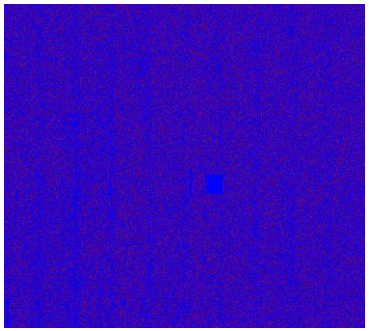
Microarray Data

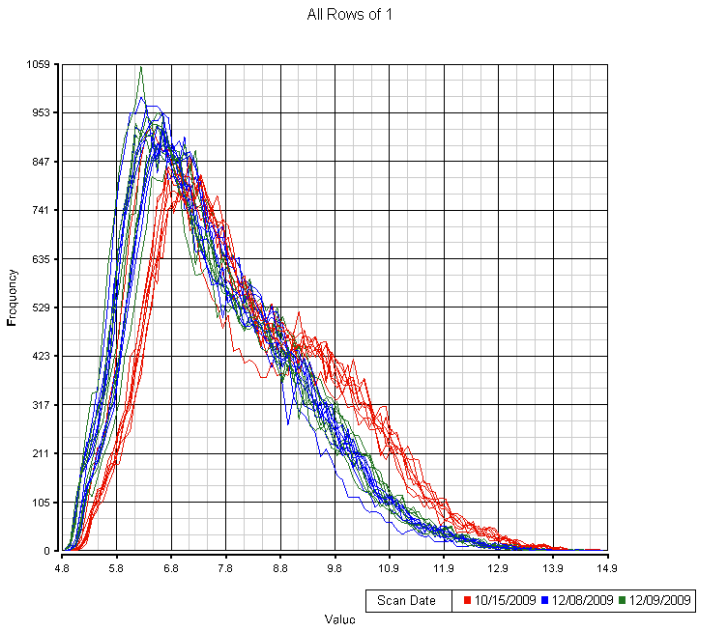
One-color Arrays

Raw

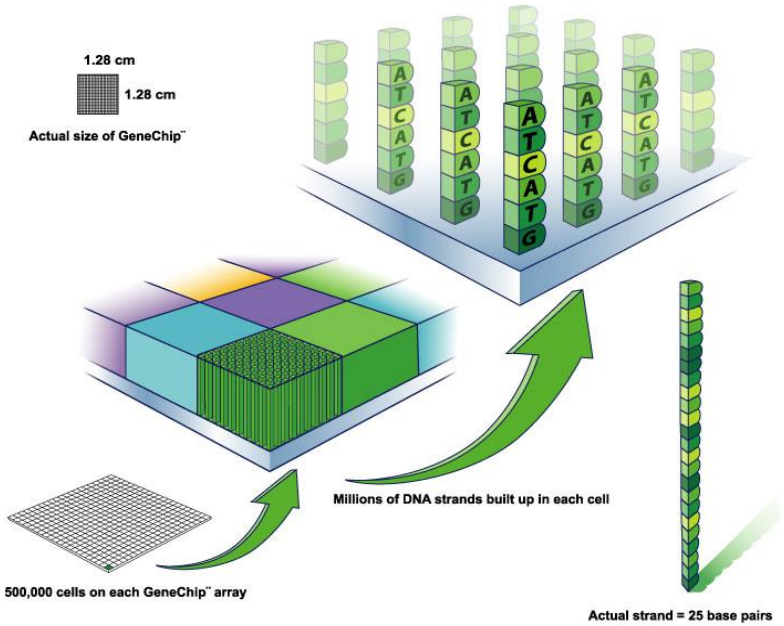


Normalized





High reproducibility and quality of spotting is required.
Affymetrix – "photolithography"-like technique



$$LogIntensity = log_2(I)$$

Background is "removed" during normalization step

Filtering may help removing uninformative features

Microarray Data

Affymetrix: Probes, Probesets and Transcript clusters

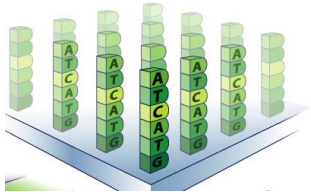
Probes
25-mer sequences targeted on a single region of transcriptome (hopefully)

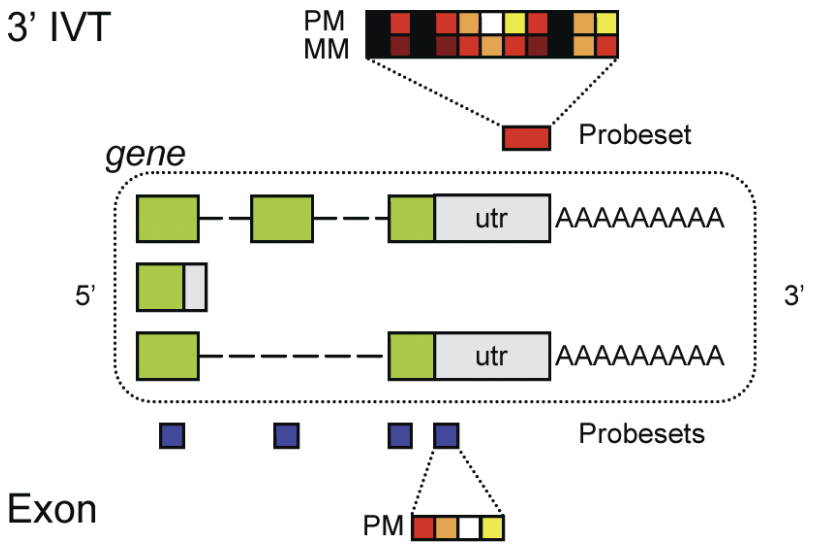
Probesets
groups of closely located or overlapped probes (on average 4 probes)

Exons
HuExon and HTA arrays allow measuring exon expression

Transcript clusters
For majority of features - synonymous to "genes". However, some distinct transcripts of genes are considered as different transcript clusters.

In old versions of Affy arrays (hgu95, hgu133, etc), there were:
PM – perfect match probes
MM – mismatch probes (having replacement in th 13th character)
This was done for background estimation.
But this approach is not used now!!





Okoniewski M, Comprehensive Analysis of Affymetrix Exon Arrays Using BioConductor, PLoS CompBiol, 2008

Normalization of Affymetrix Arrays by RMA

Background correction

↓

Normalization b/w arrays

↓

Estimate expression

Background and signal are strictly positive.
Noise is additive in log scale:

$$PM_{ij} = \underset{\text{exponential}}{Sijn} + \overset{\text{normal}}{Bijn}$$

Quantile **normalization** b/w arrays: makes distribution of probes the same across all arrays

Probeset expression is estimated from a linear model:

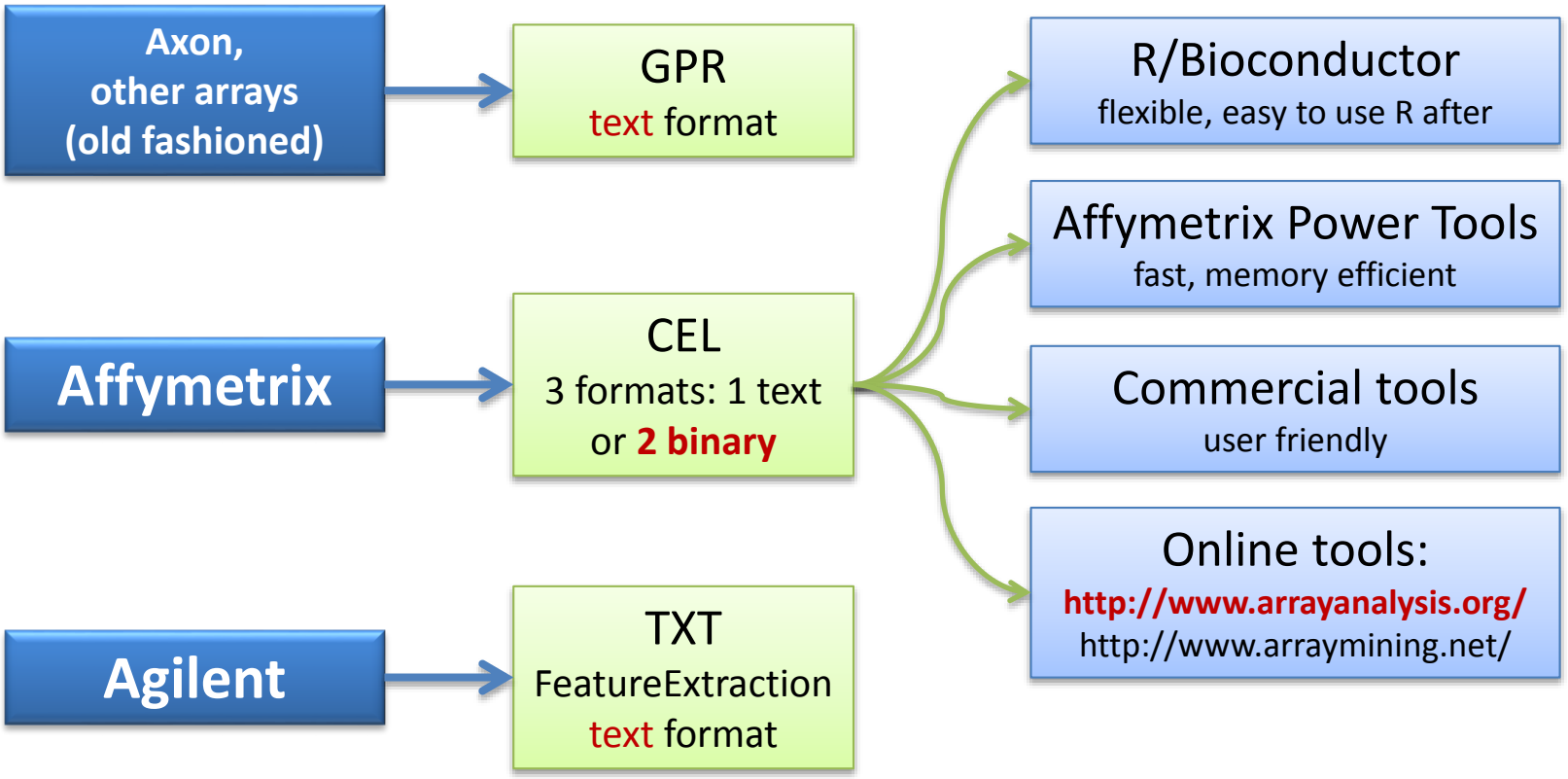
$$\underset{\text{observed}}{Y_{ijn}} = \underset{\text{expression}}{\mu_{in}} + \underset{\text{probe affinity}}{\alpha_{jn}} + \underset{\text{error with 0 mean}}{\varepsilon_{ijn}}$$

i – array
j -- probe
n -- probeset

"Median polish" helps avoid outliers effect

Microarray Data

File Formats



Microarray Data

Analysis Pipeline

Raw CEL files

Text files with $\log_2$ gene (or probeset) expression

Lists of differentially expressed gene

Image Analysis

Pre-processing

Quality control

Statistical Analysis

Affymetrix© software

- Background correction
- Normalization
- Summarization

Remove low quality arrays, if necessary

- Significance analysis for differentially expressed genes (DEG)
- Prediction analysis
- Co-regulation analysis
- Etc.

Bioinformatics

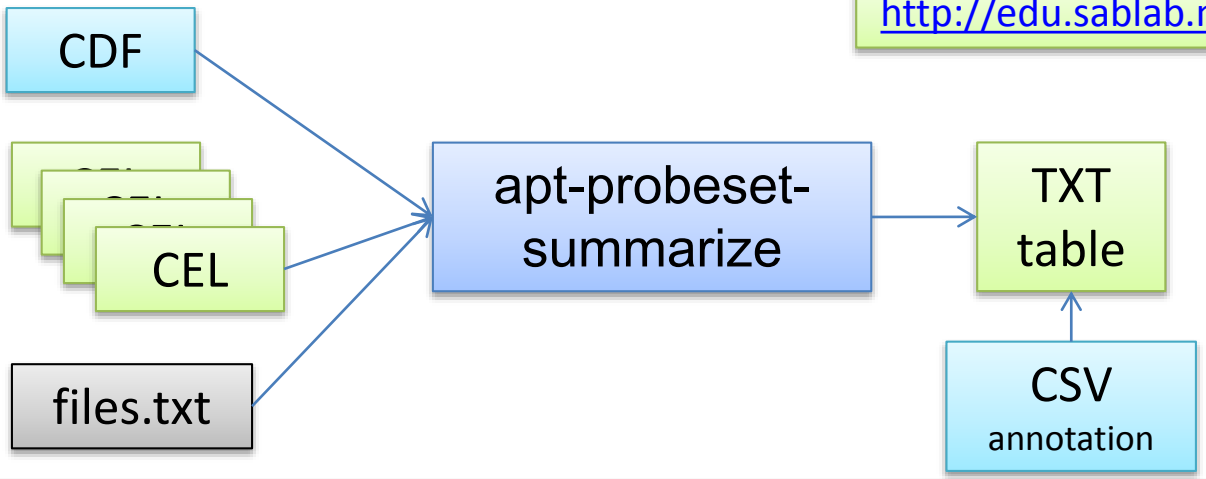
Example: Affymetrix Power Tools

apt-probeset-summarize is a program for doing background subtraction, normalization and summarizing probe sets from Affymetrix expression microarrays. It implements analysis algorithms such as RMA, Plier, and DABG (detected above background).
The main features of **apt-probeset-summarize** not common in other implementations are: Quantile normalization using a subset (sketch) of the data which results in much smaller memory usage.

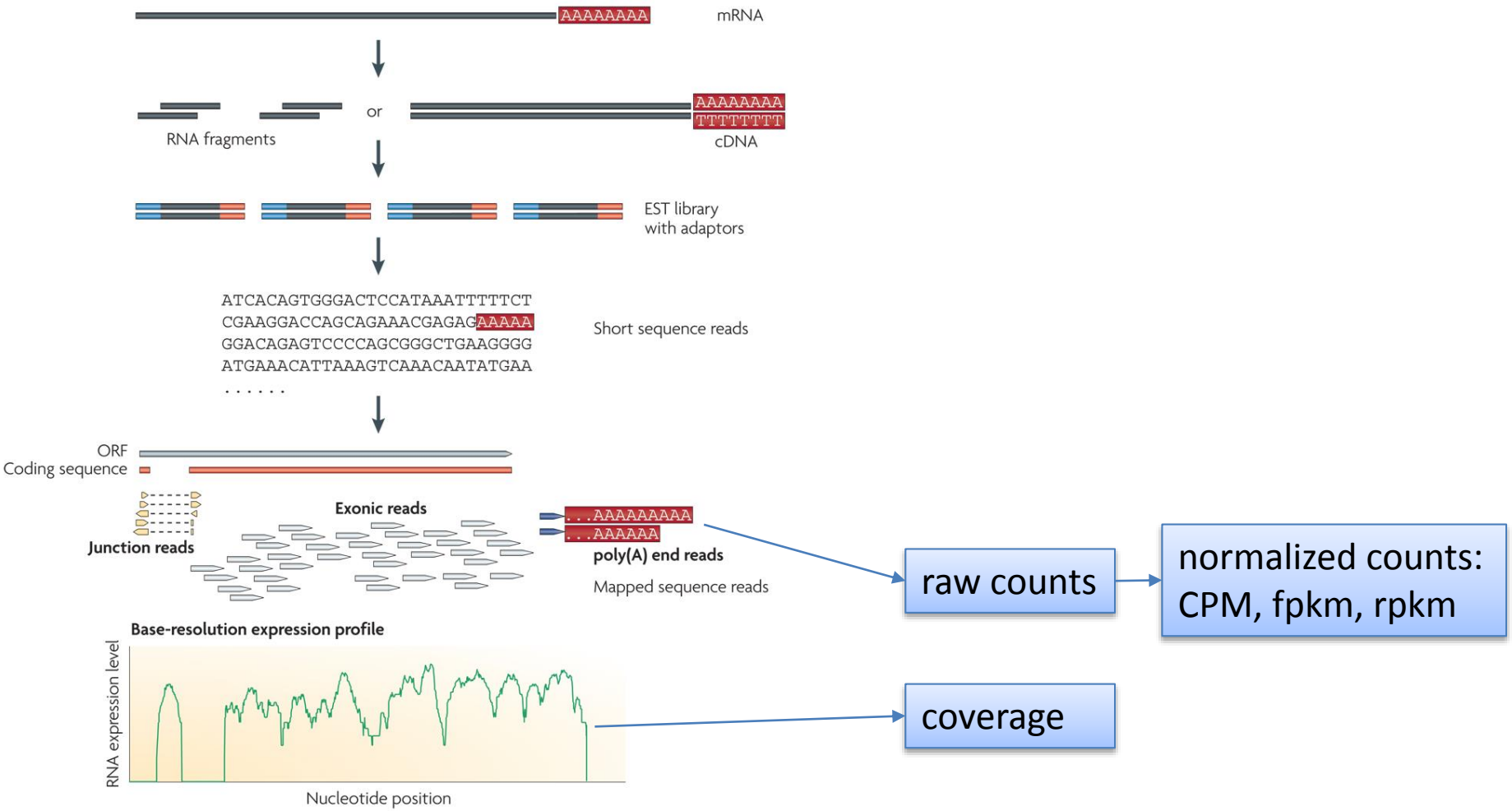
http://www.affymetrix.com/support/developer/powertools/changelog/apt-probeset-summarize.html

```
apt-probeset-summarize
    -a rma-sketch -d chip.cdf -o output-dir  --cel-files  files.txt
```

http://edu.sablab.net/transcript



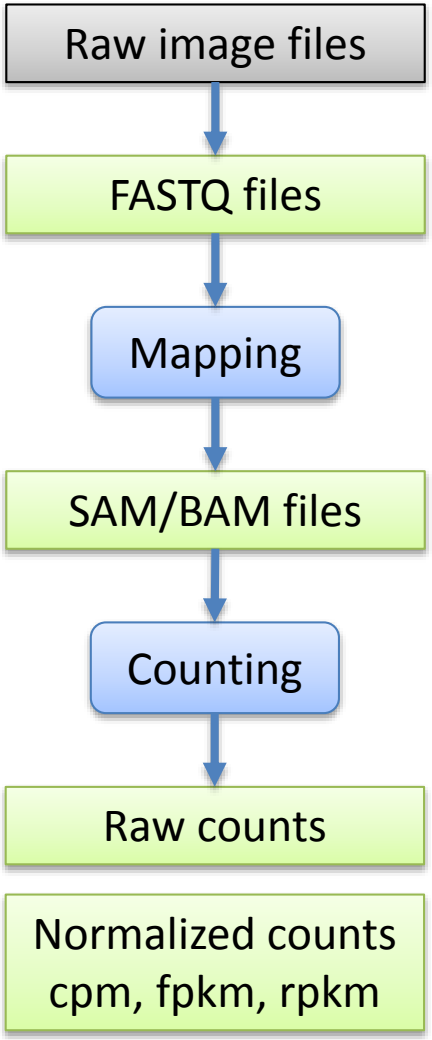
Next Generation Sequencing: RNA-Seq



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

RNA-Seq Data

File Types




```
@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=>?######
```

Read – a short sequence identified in RNA-Seq experiment
Library – set (10^5 – 10^8) of reads from a single sample

```
@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_65:7:152:368:113        73          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
```

For the list of tools see:
http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools

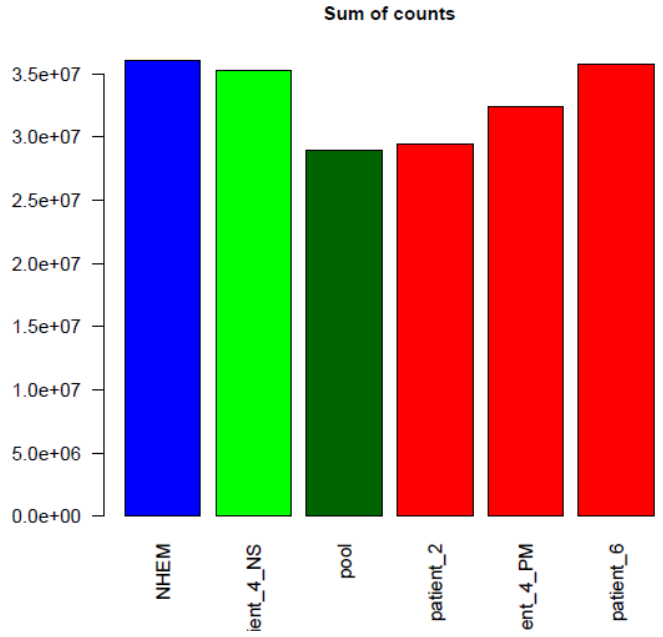
Normalization

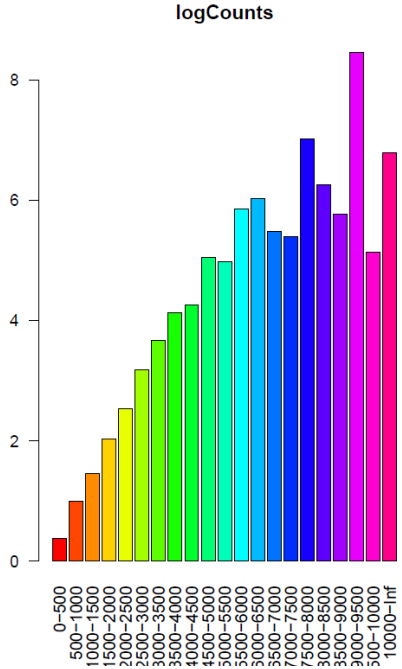
Problems:

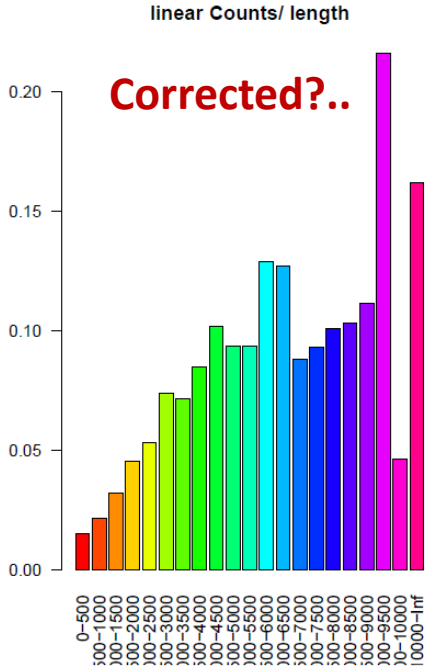
- Libraries has different size (different number of reads from samples)
- Long transcripts produce more reads

Solutions (?) :

- Accounting for library size during analysis (standard) or direct correction for it
- Correction for transcript size (but which transcript is expressed?)







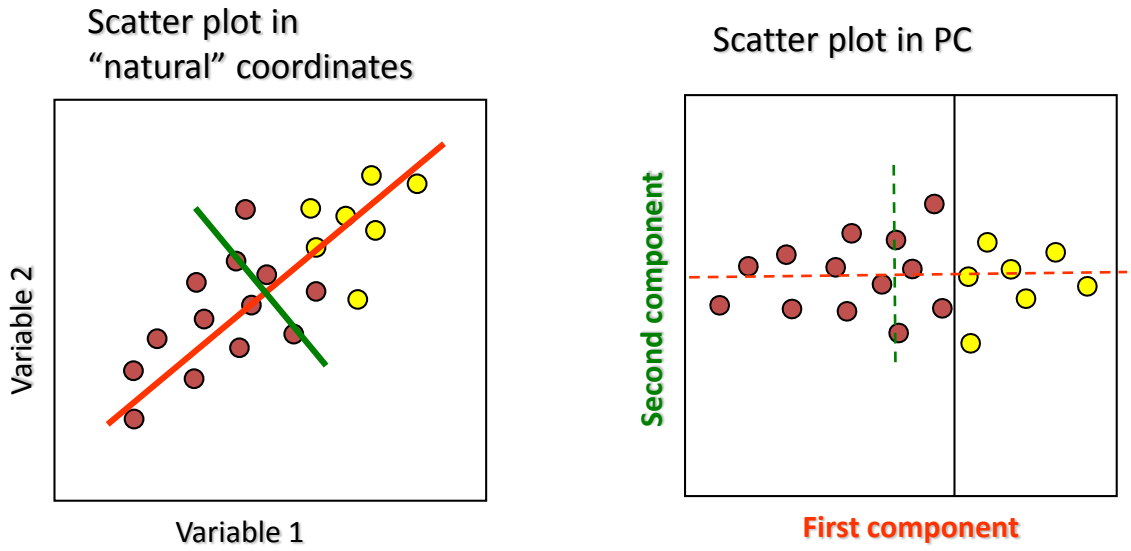
Exploratory Analysis

Principal Component Analysis (PCA)

Principal component analysis (PCA)
is a vector space transform used to reduce multidimensional data sets to lower dimensions for analysis. It selects the **coordinates along which the variation of the data is bigger**.

20000 genes → 2 dimensions

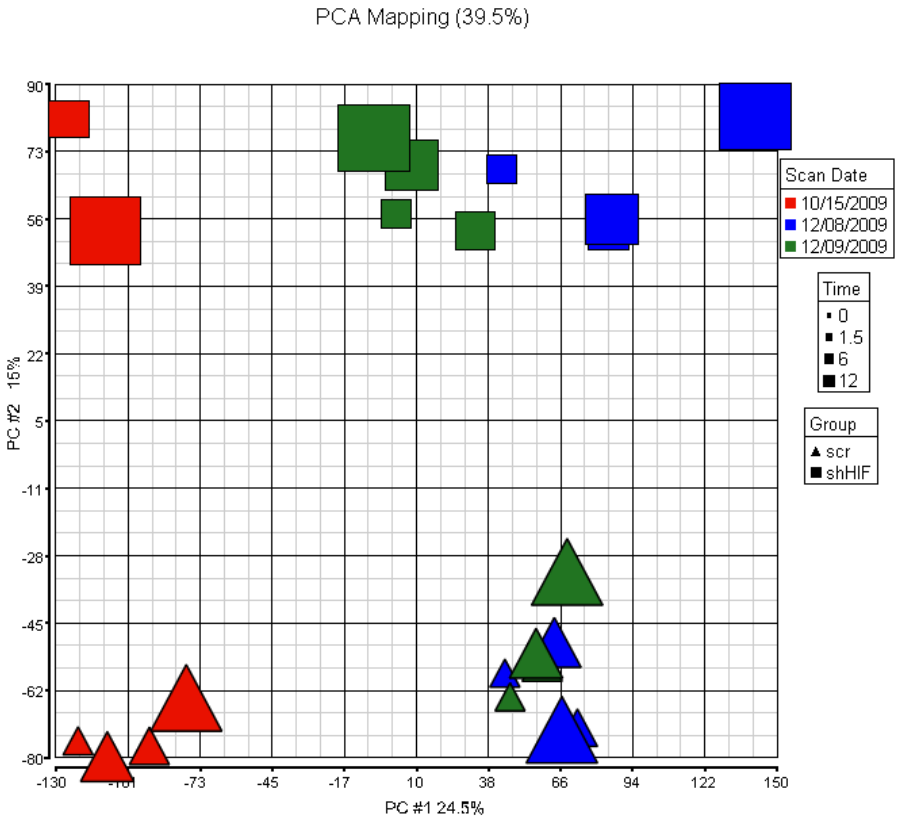
For the simplicity let us consider 2 parametric situation both in terms of data and resulting PCA.

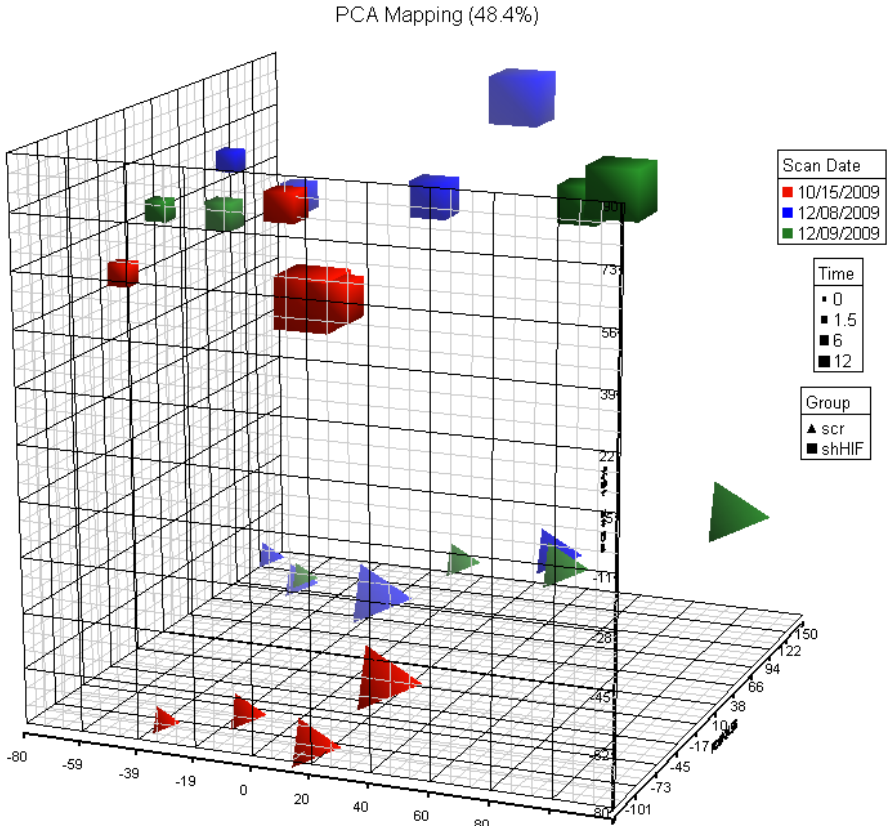


Instead of using 2 "natural" parameters for the classification, we can use the first component!

Exploratory Data Analysis

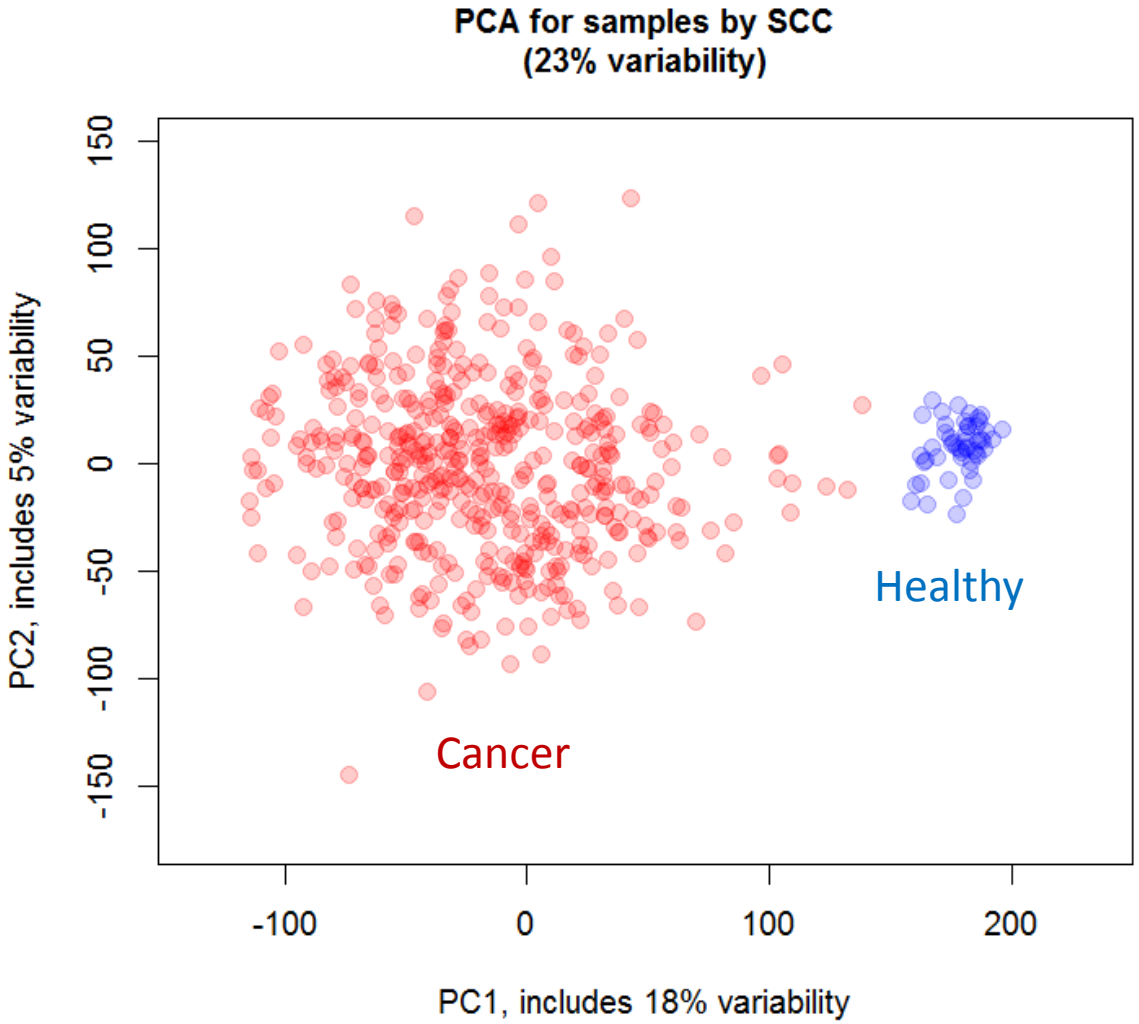
PCA





Exploratory Data Analysis

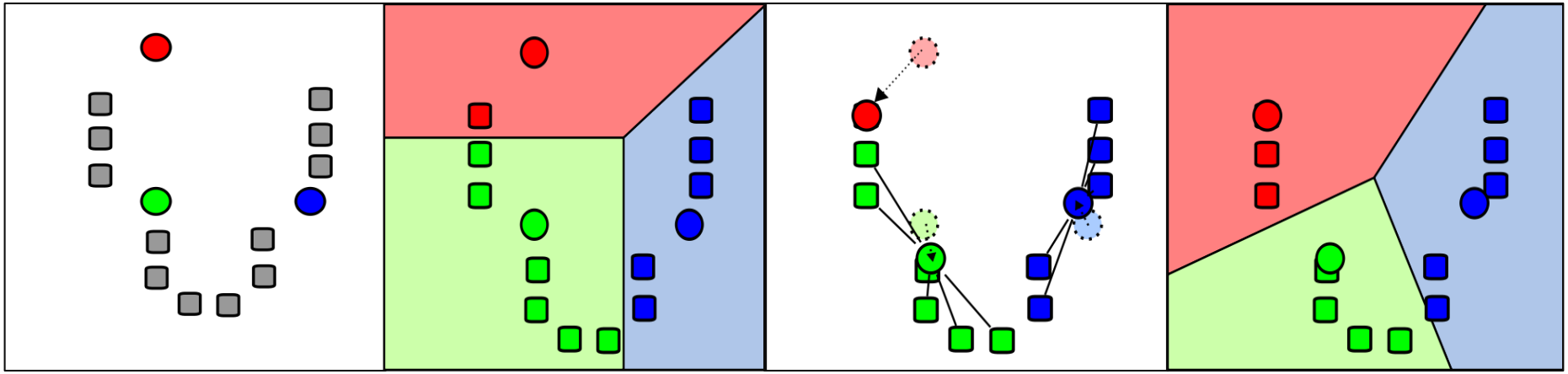
PCA in TCGA (LUSC data)



k-Means Clustering

k-Means Clustering

k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.



1) *k* initial "means" (in this case k=3) are randomly selected from the data set (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean.

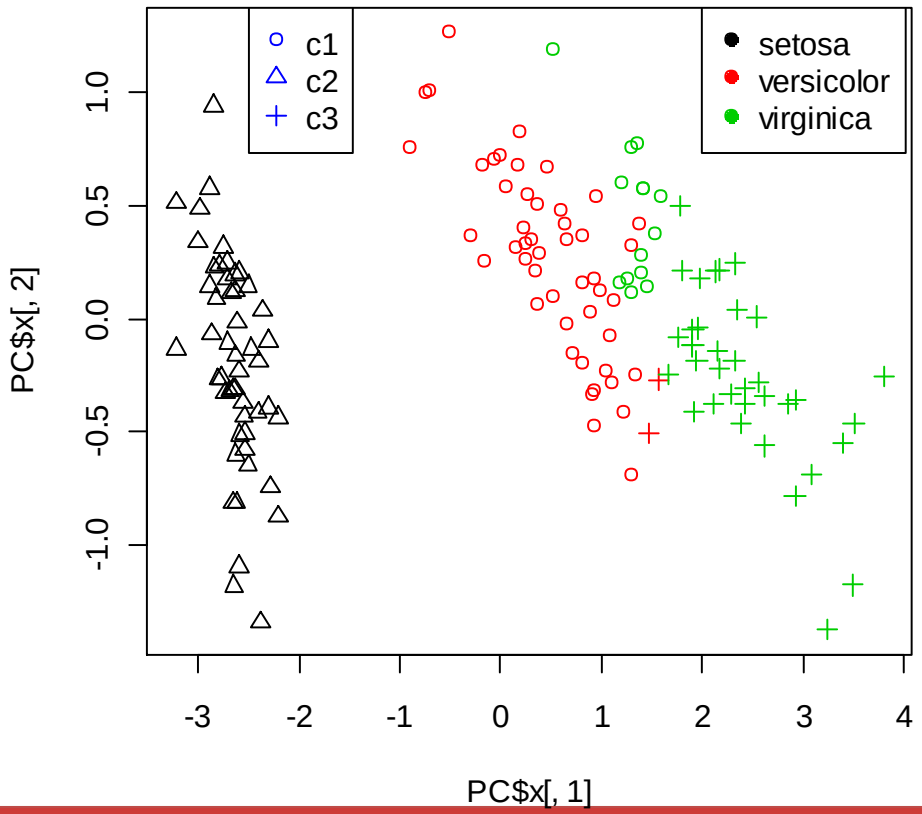
3) *The centroid of each of the k clusters becomes the new means.*

4) Steps 2 and 3 are repeated until convergence has been reached.

http://wikipedia.org

k-Means Clustering: Iris Dataset (Fisher)

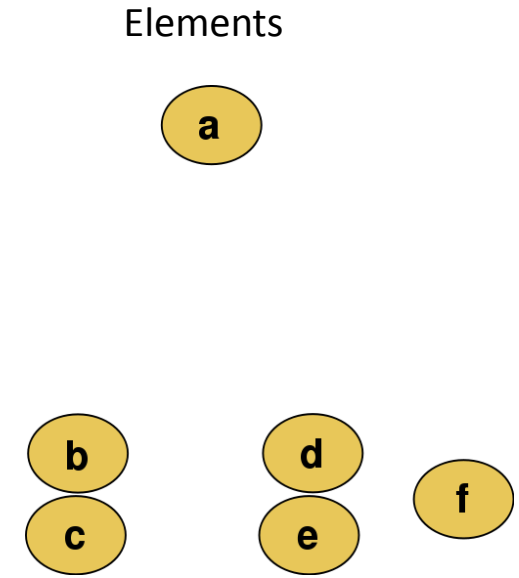
```
clusters = kmeans(x=Data,centers=3,nstart=10)$cluster
plot(PC$x[,1],PC$x[,2],col = classes,pch=clusters)
legend(2,1.4,levels(iris$Species),col=c(1,2,3),pch=19)
legend(-2.5,1.4,c("c1","c2","c3"),col=4,pch=c(1,2,3))
```

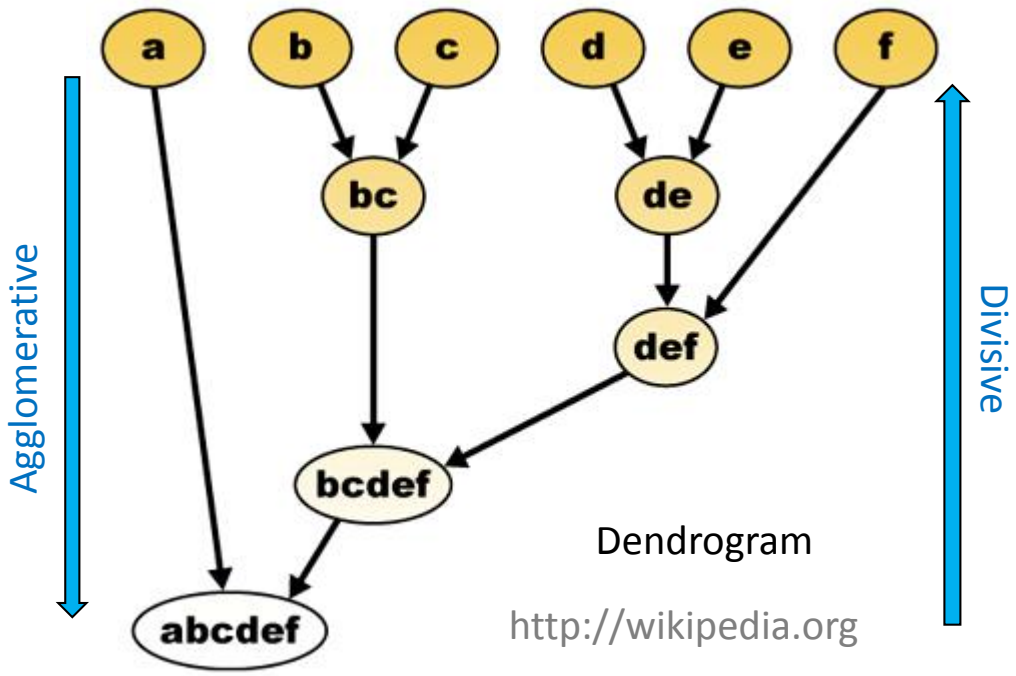


Hierarchical Clustering

Hierarchical Clustering

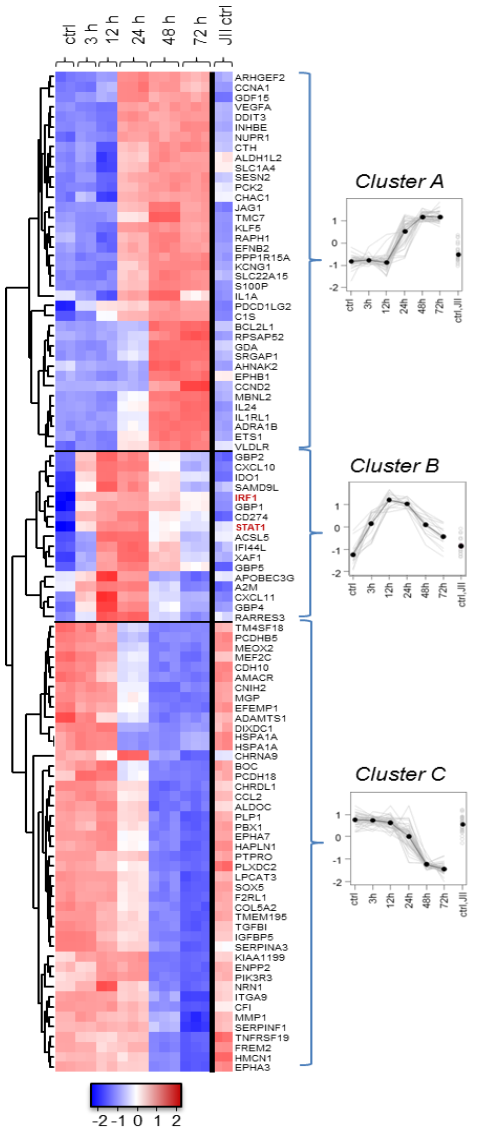
Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations.
Algorithms for hierarchical clustering are generally either agglomerative, in which one starts at the leaves and successively merges clusters together; or divisive, in which one starts at the root and recursively splits the clusters.

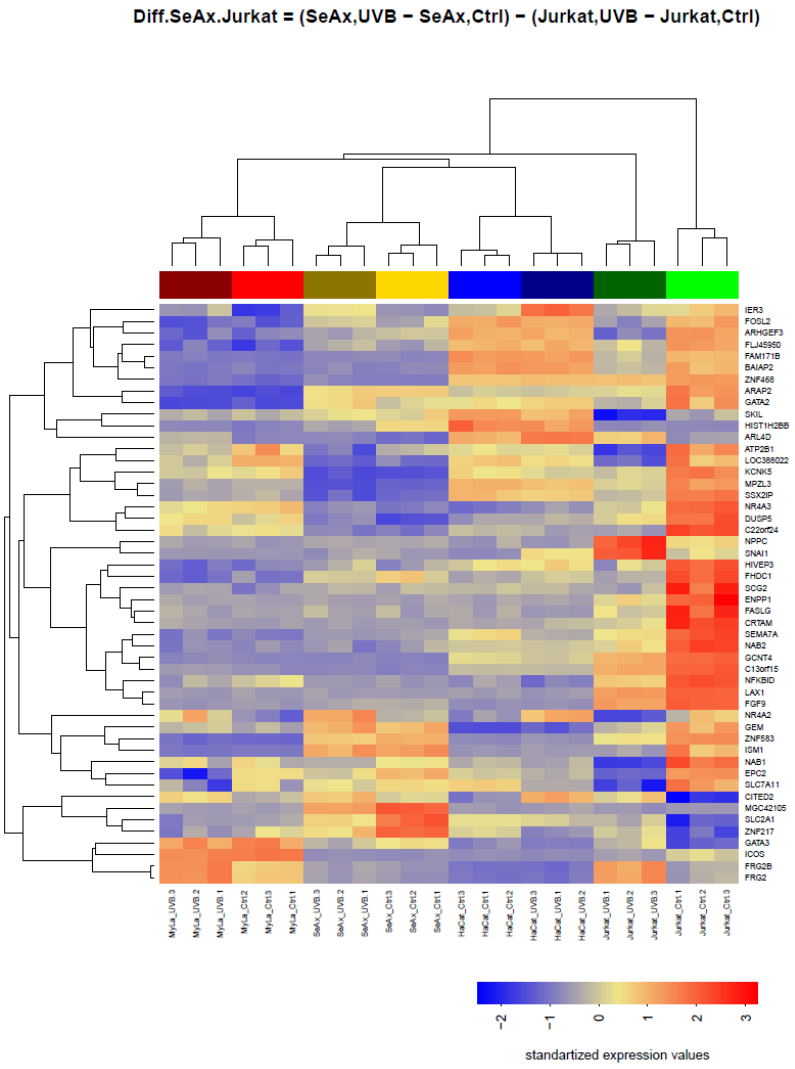




Exploratory Data Analysis

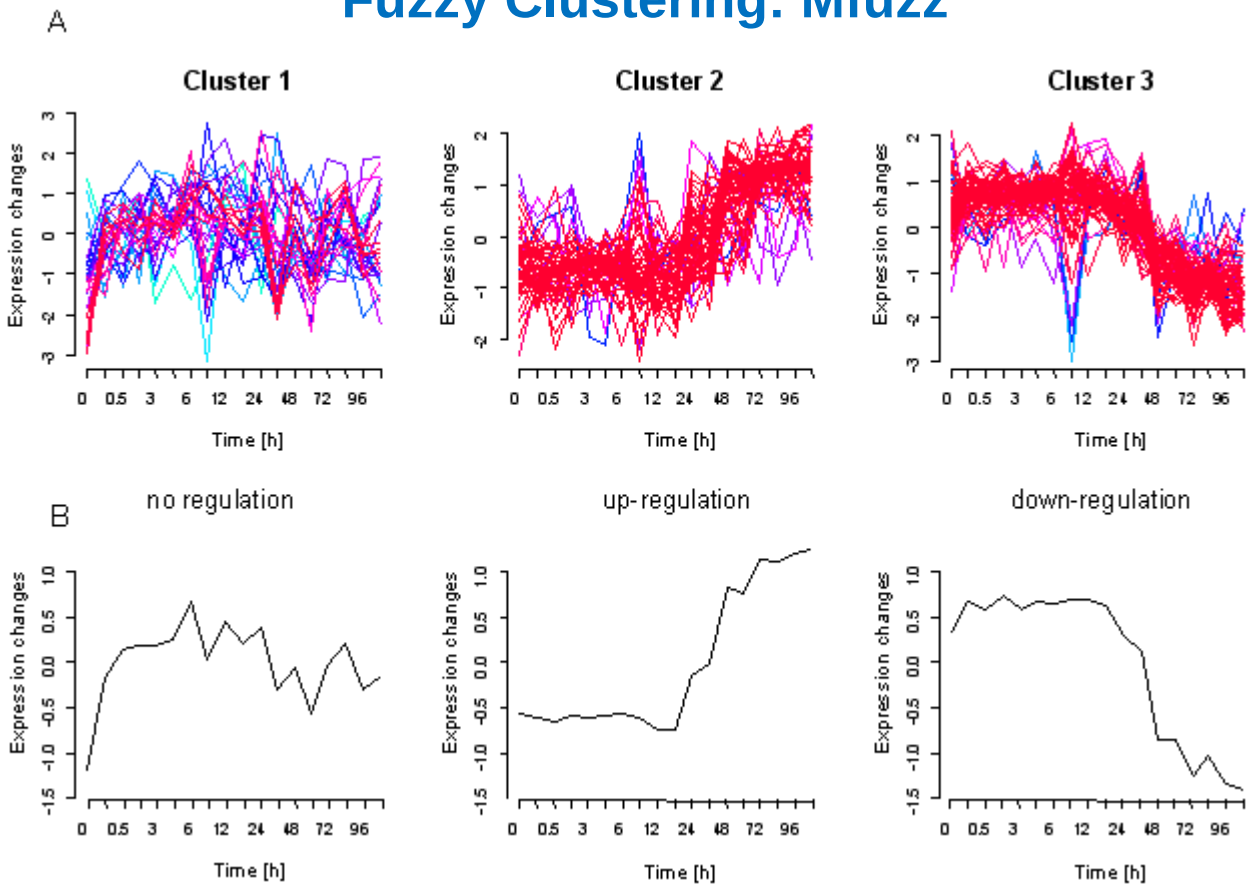
Heatmaps





Exploratory Data Analysis

Fuzzy Clustering: Mfuzz



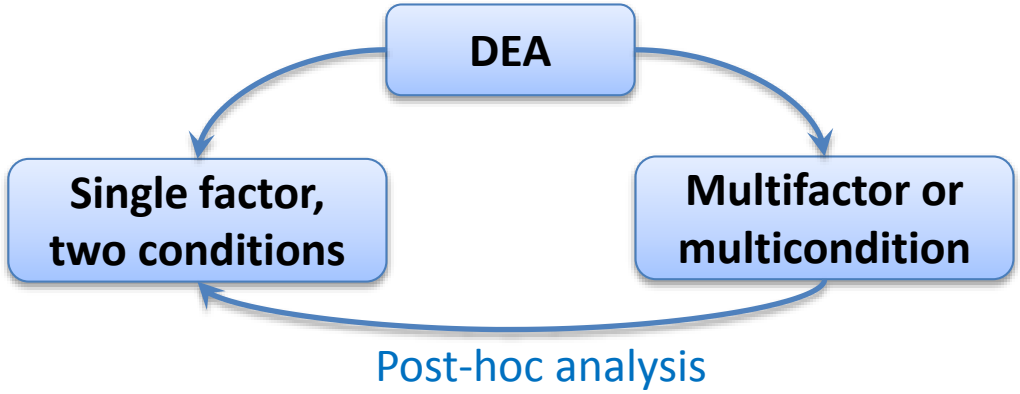
Differential Expression Analysis

Differential Expression Analysis

Basics

Questions

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

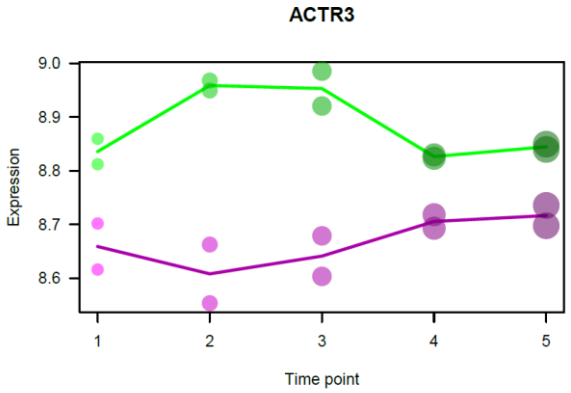


Similar to t-test with Student's statistics: compare means

Similar to ANOVA with Fisher's statistics: compare variances

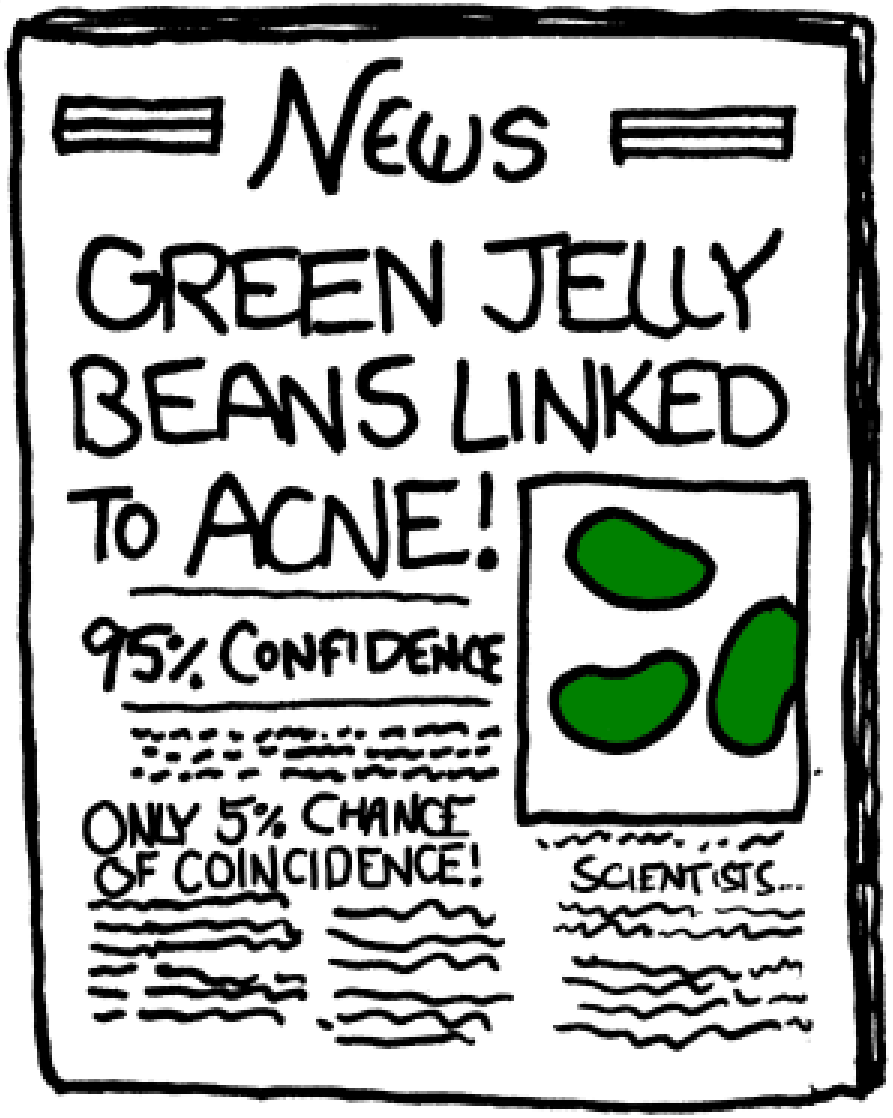
And do not forget about multiple hypotheses testing

Example: 2 cell lines in time:



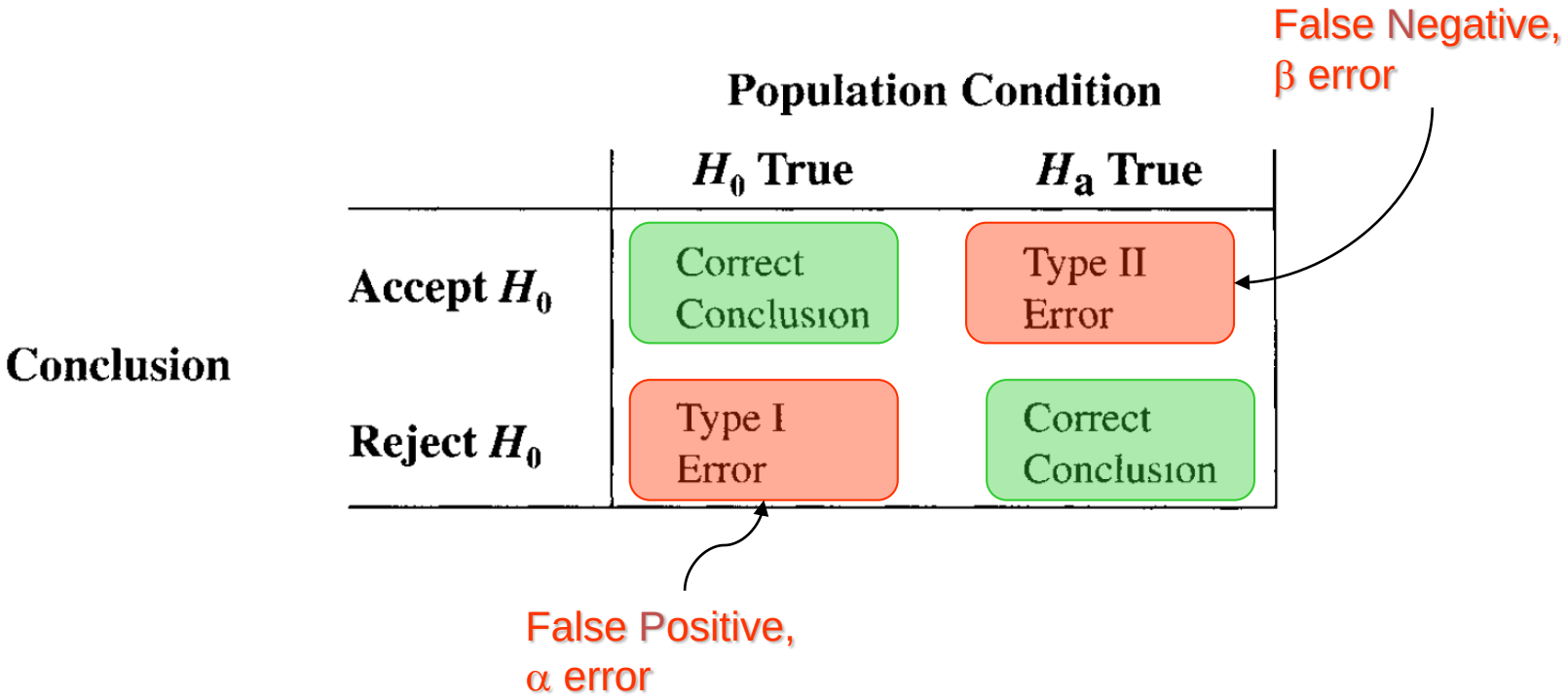
Differential Expression Analysis

Example



http://www.xkcd.com/882/

Multiple Hypotheses



Probability of an error in a multiple test:

$1-(0.95)^{\text{number of comparisons}}$

Multiple Hypotheses: False Discovery Rate

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)	U	T	$m - R$
	Reject H_0 (significant)	V	S	R
	Total	m_0	$m - m_0$	m

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

Differential Expression Analysis

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

```
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

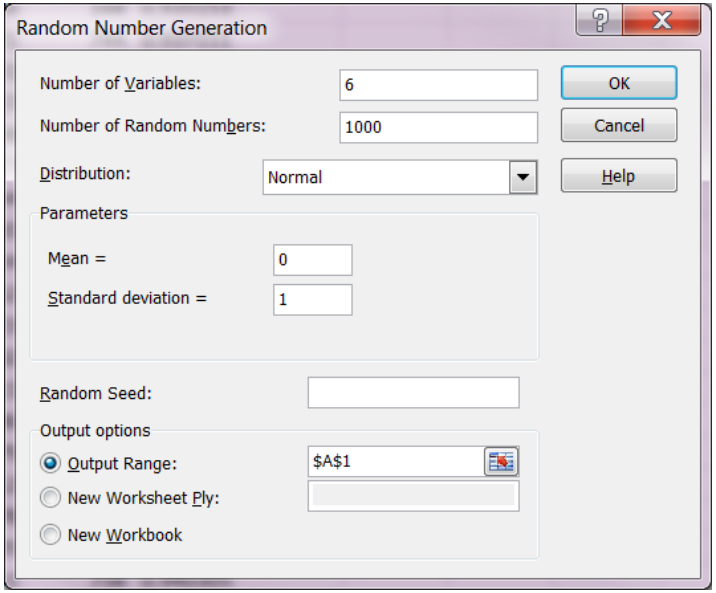
$$(m+1-k)P_{(k)} < \alpha$$

Differential Expression Analysis

Why is it so important to correct p-values?..

Let's generate a completely random experiment (Excel)

- Generate 6 columns of normal random variables (1000 points/candidates in each).
- Consider the first 3 columns as "treatment", and the next 3 columns as "control".
- Using t-test calculate p-values b/w "treatment" and "control" group. How many candidates have p-value<0.05 ?
- Calculate FDR. How many candidates you have now?



Differential Expression Analysis

Linear Models

Many conditions

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

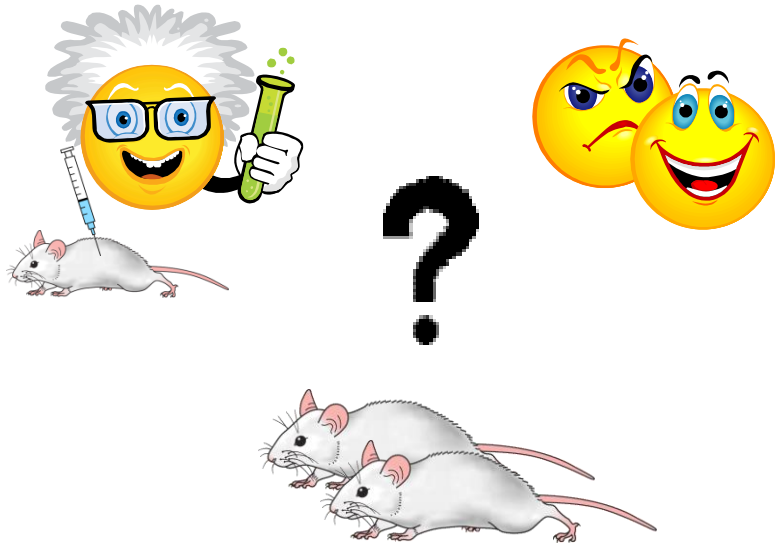
Many factors

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

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: $1-(0.95)^{10} = 0.4$



ANOVA
example from Partek™

http://easylink.playstream.com/affymetrix/ambsymposium/partek_08.wvx

Linear Models

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?

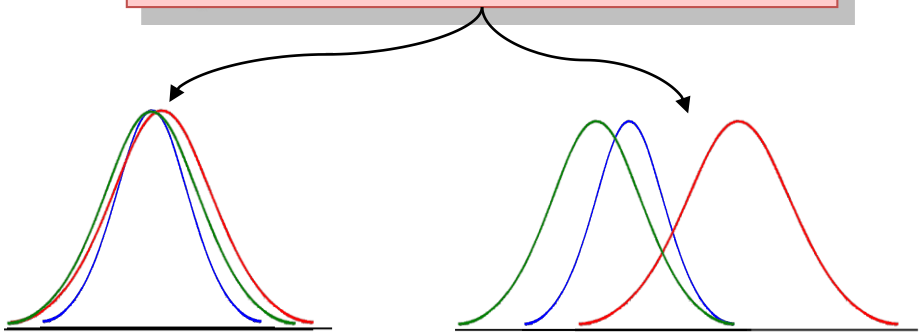
depression.txt

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

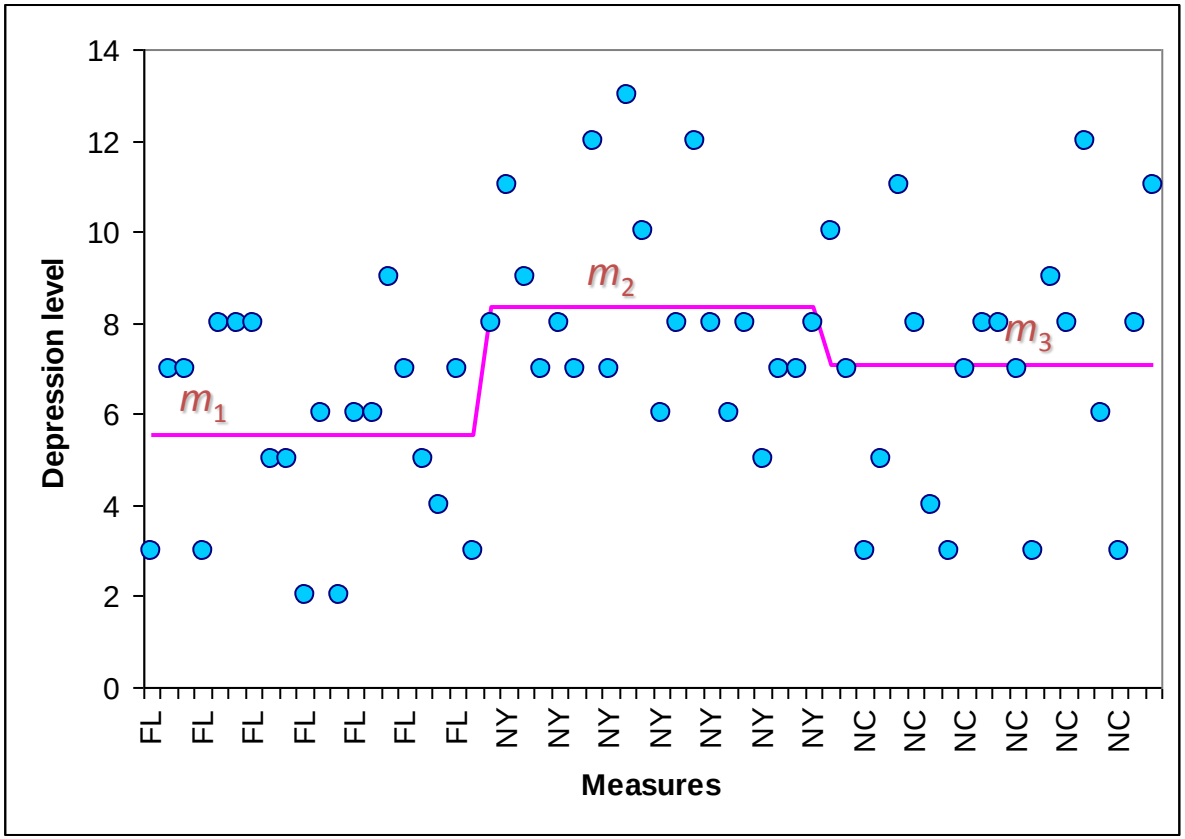


Differential Expression Analysis

Linear Models

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

H_a: not all 3 means are equal



Differential Expression Analysis

LIMMA & EdgeR : Linear Models for Microarrays

$$Y_{ij} = \mu_i + Aj + Bj + Aj * Bj + \varepsilon_{ij}$$

i – gene index
j – sample index

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

Limma – R package for DEA in <u>microarrays</u> based on linear models.

It is similar to t-test / ANOVA but using all available data for variance estimation, thus it has higher power when number of replicates is limited

edgeR – R package for DEA in <u>RNA-Seq</u>, based on linear models and negative binomial distribution of counts.

Better noise model results in higher power detecting differentially expressed genes

negative binomial process – number of tries before success: rolling a die until you get 6

Differential Expression Analysis

Example: let's make it easy

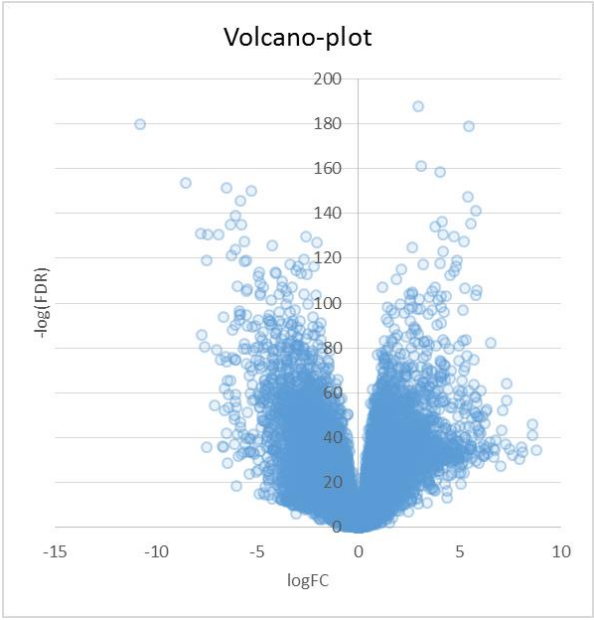
http://edu.sablab.net/transcript/lusc.zip

1. Find genes significantly differentially expressed in SCC vs normal tissue
   - apply t-test. Same or different variance?
   - perform FDR correction
   - Keep genes with FDR > 0.001

2. Calculate mean logFC and keep only genes with |logFC| > 2

3. Make a "volcano plot":
   -log10(FDR) vs LogFC

4. Save lists of up and down regulate genes –
we shall need them



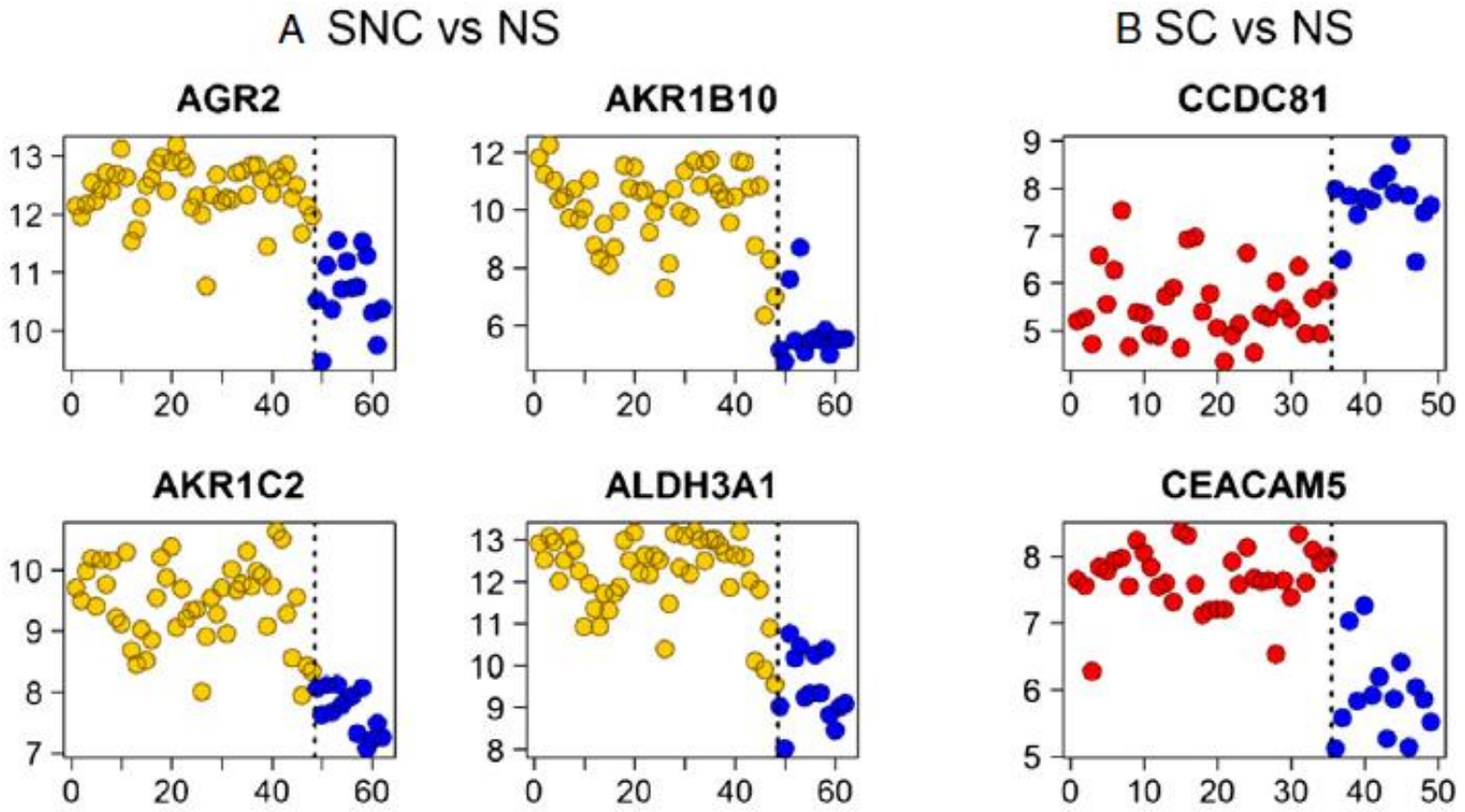
Classification

Classification and Marker Genes

Gene Markers

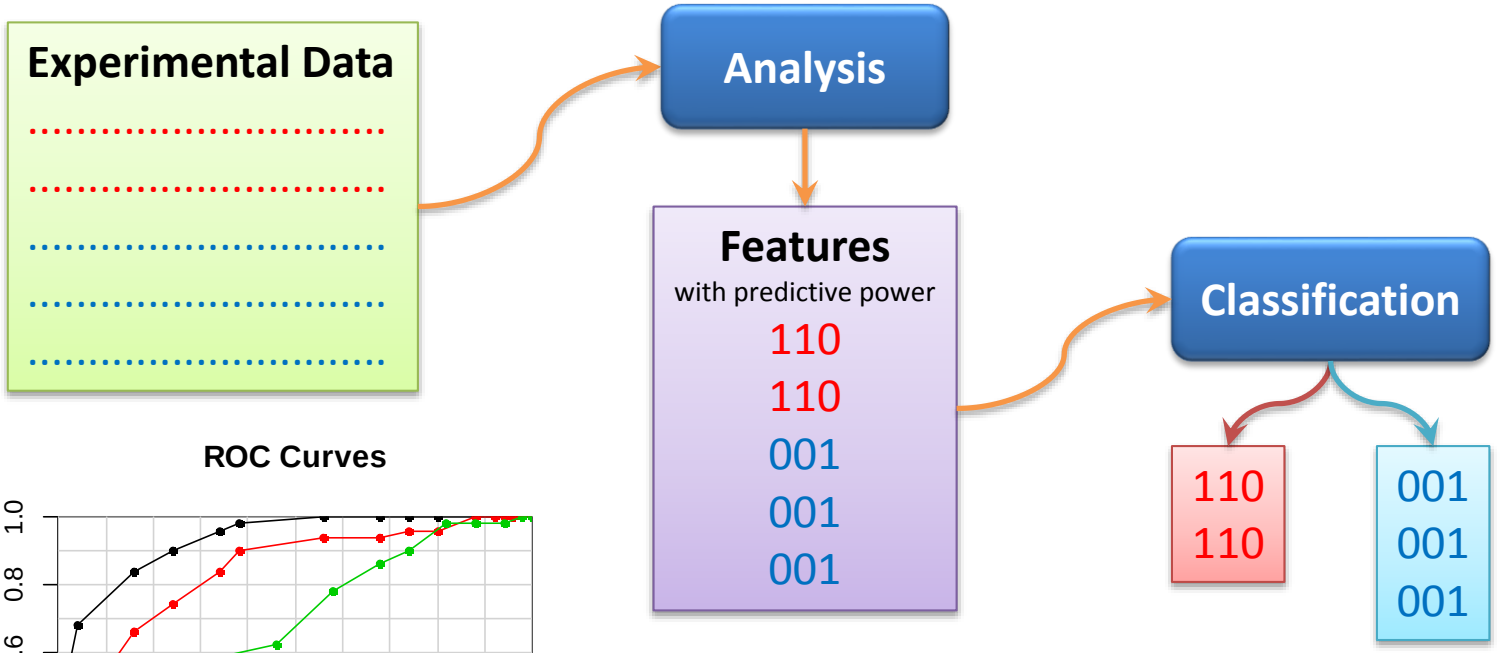
Questions

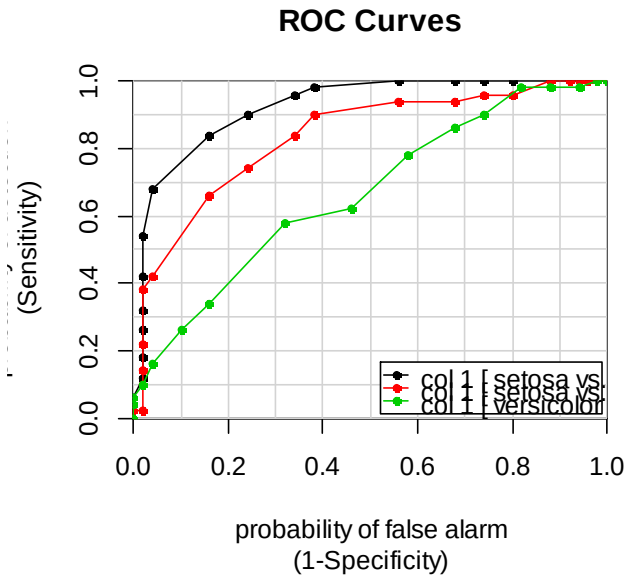
- Based on which genes or gene sets we can **predict** the group of the samples?
- How reliable is this prediction?



Classification and Marker Genes

General Scheme

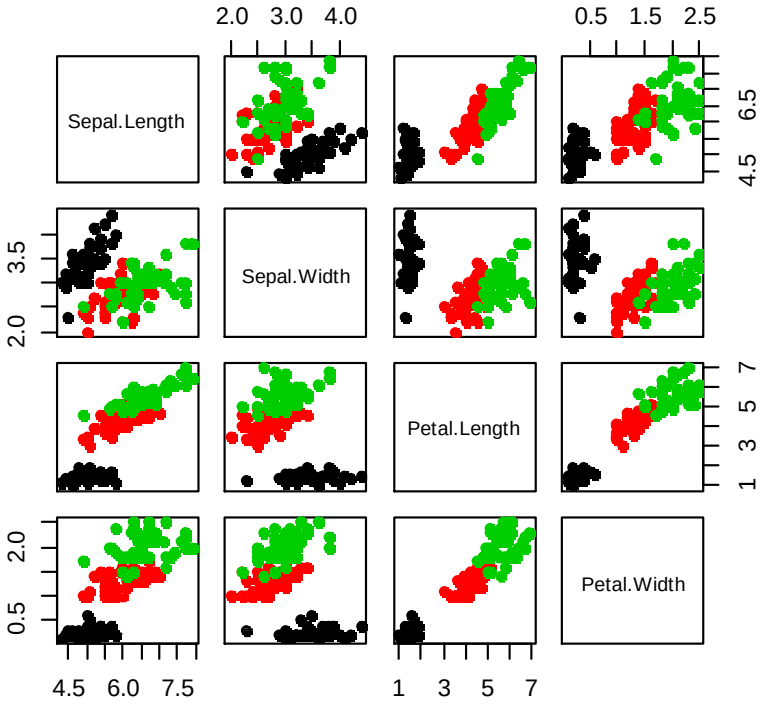




Confusion Matrix

	A	B	C
pred.A	50	0	0
pred.B	0	48	2
pred.C	0	2	48

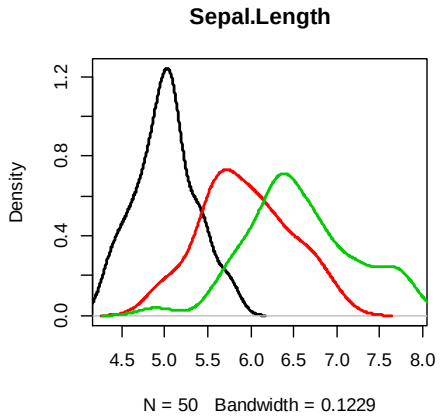
Selection of Features: Iris Dataset (Fisher)

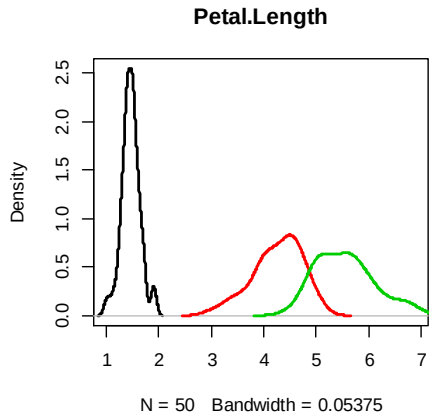


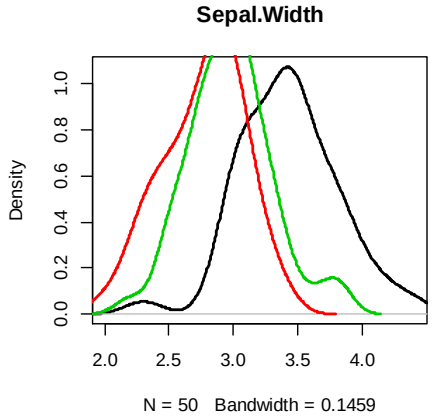
Iris setosa

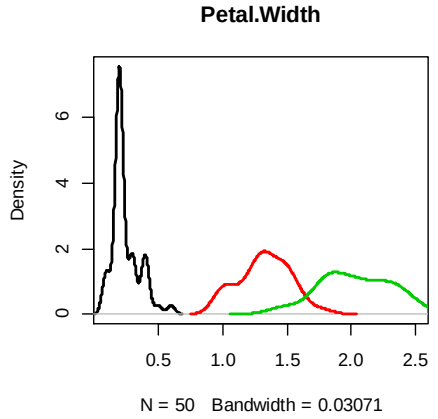
Iris versicolor

Iris virginica







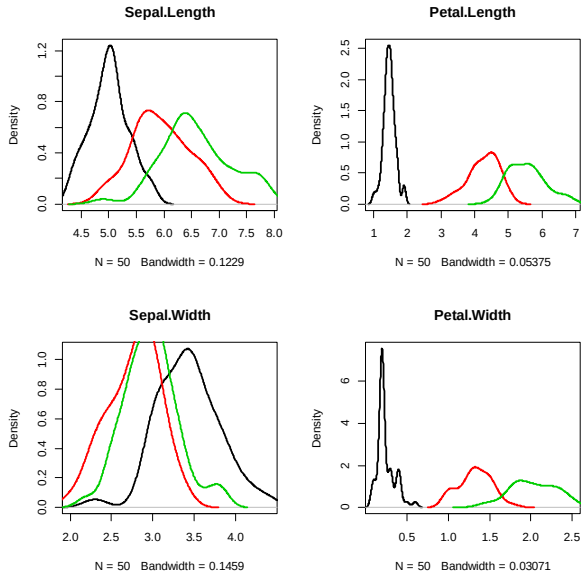


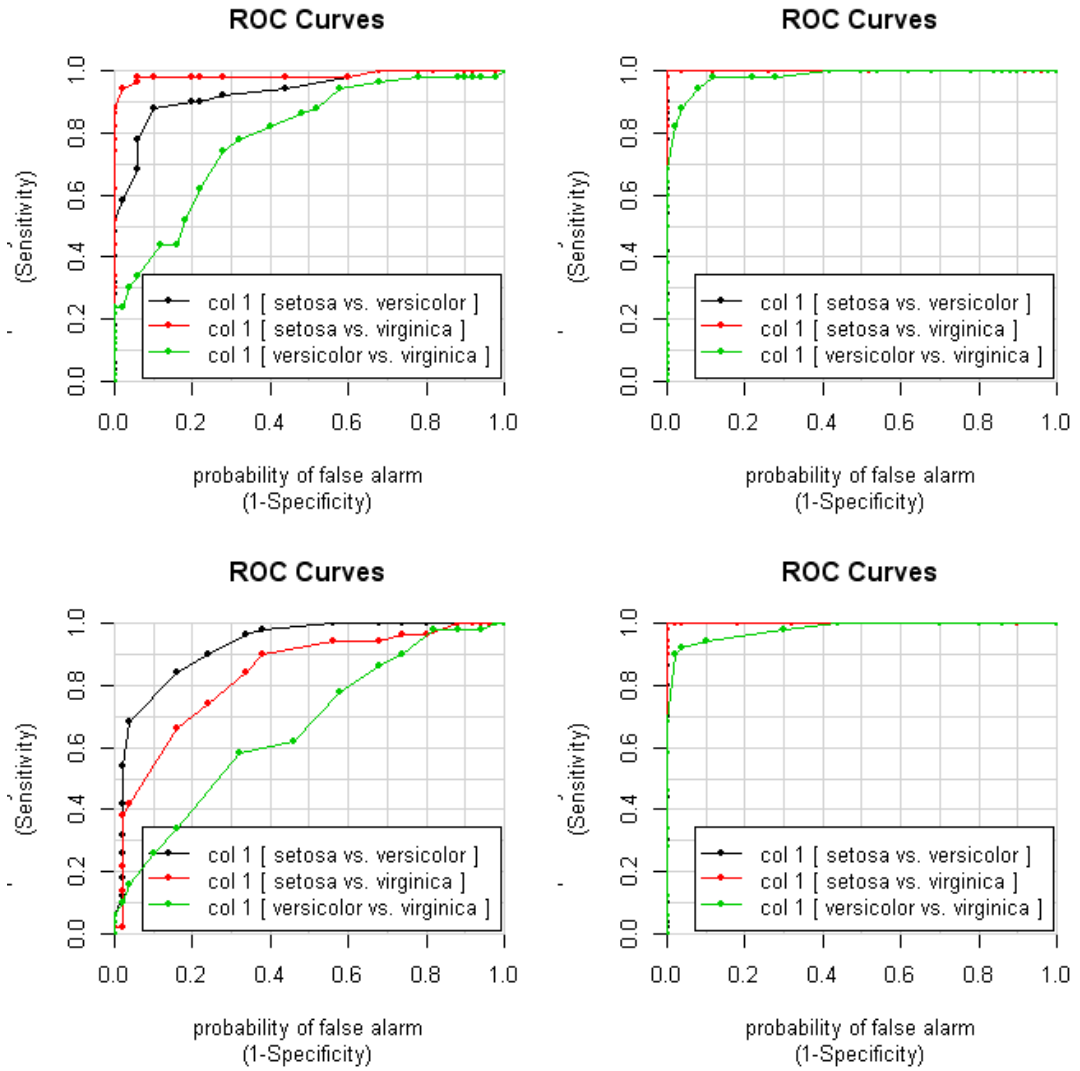
Classification and Marker Genes

Selection of Features: Iris Dataset

ROC curve
is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate (one minus the specificity or true negative rate)

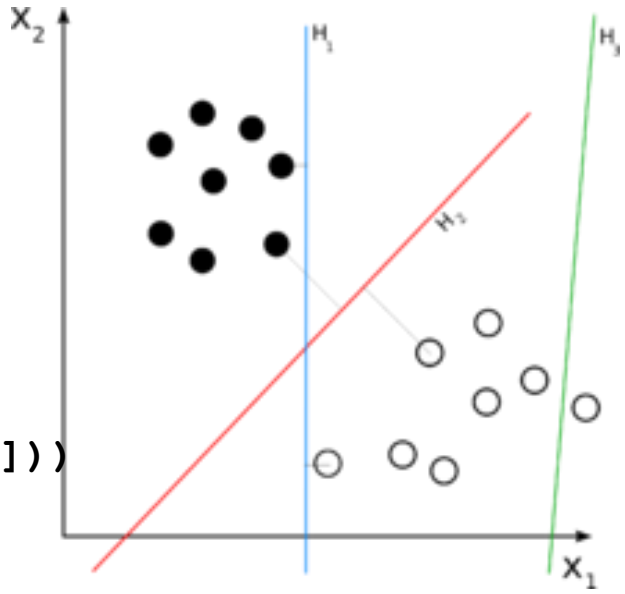
AUC
area under ROC curve: 1 – ideal separation, 0.5 – random separation.





SVM Classification

Support vector machine (SVM)
is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis.



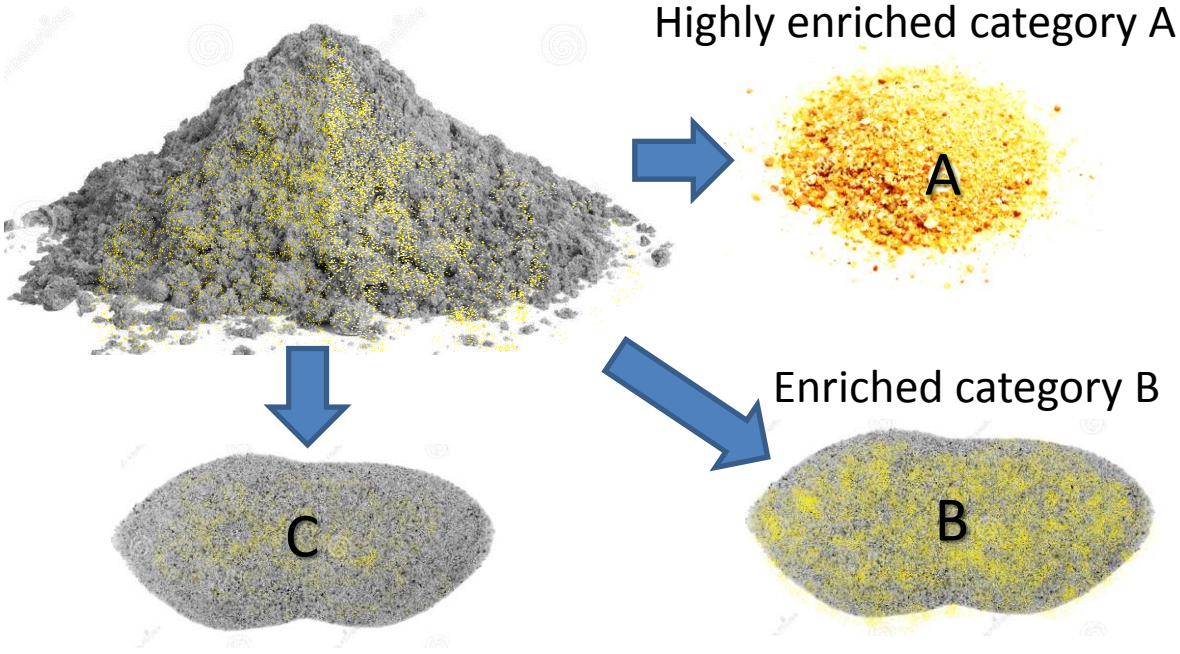

```
library(e1071)
model = svm(Species ~ ., data = iris)
svm.res = as.character(predict(model, iris[,-5]))

## creat a confusion matrix
ConTab = data.frame(matrix(nr=3,nc=3))
rownames(ConTab) = paste("pred.",levels(iris$Species),sep="")
names(ConTab) = levels(iris$Species)
for (ic in 1:3){
  for (ir in 1:3){
    ConTab[ir,ic] = sum(iris$Species == levels(iris$Species)[ic] &
            svm.res == levels(iris$Species)[ir])
  }
}
```

Enrichment Analysis

Enrichment Analysis

1. Category Enrichment Analysis

Are interesting genes overrepresented in a subset corresponding to some biological process?



Method of the analysis:
Fisher's exact test

Someone grabs "randomly" 20 balls from a box with 100x ● and 100x ●

How surprised will you be if he grabbed

●●●●●●●●●●●●●●●●●●●●

(17 red , 3 green)

sand belongs to: http://www.dreamstime.com/photos-images/pile-sand.html ;)))

Enrichment Analysis

1. Category Enrichment Analysis

Fisher's exact test: based on hypergeometrical distributions

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

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

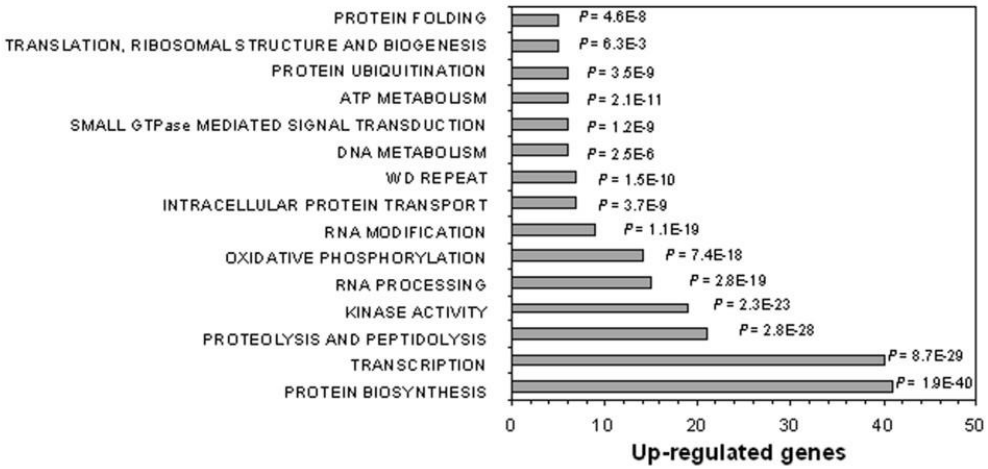
N: total number of genes

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

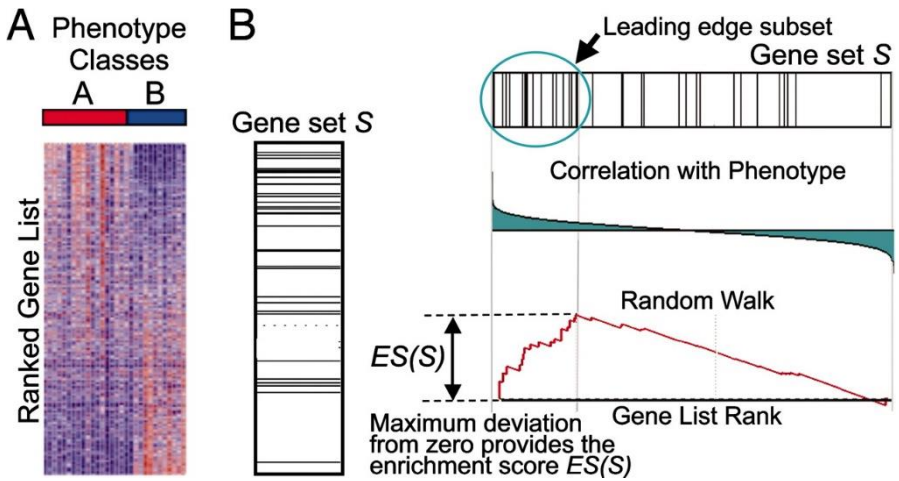
$$C_k^n = C_n^k = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

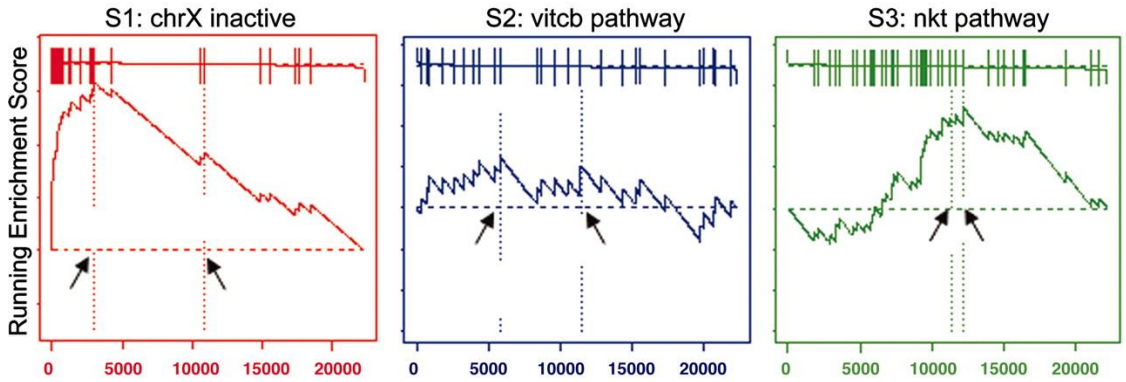


Okamoto et al. Cancer Cell International 2007 7:11 doi:10.1186/1475-2867-7-11

2. Gene Set Enrichment Analysis (GSEA)

Is direction of genes in a category random?





A. Subramanian et al. PNAS **2005**,102,43

Enrichment Analysis

Example: GO enrichment

http://edu.sablab.net/transcript

Strategy 1:
Take all DEG and use them in enrichment.

- Safe
- No additional assumptions
- Cannot distinguish ↑ and ↓ functions

Strategy 2:
Separate DEG to down- and up- regulated genes. Then perform independent enrichment by these 2 groups

- Can be biased (gene can be ↑↓)
- Assume ↑gene => ↑function
- Can distinguish ↑ and ↓ functions

Enrichr
http://amp.pharm.mssm.edu/Enrichr/

BioCompendium
http://biocompendium.embl.de/

Enrichment Analysis

LUSC Example

http://edu.sablab.net/data/txt/lusc.zip

http://amp.pharm.mssm.edu/Enrichr/

http://biocompendium.embl.de/

http://string-db.org

0. Prepare lists of DE genes…

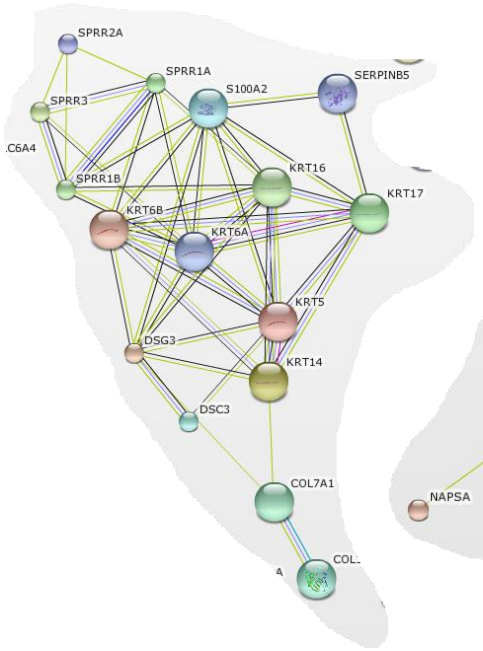
1. Put up-regulated into **enrich**

3. Check: Down CMAP, Disease Signatures from GEO up,

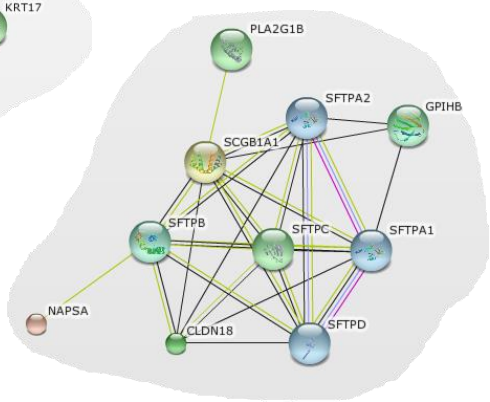
4. Try **biocompendium**

5. Put top 100 genes into String to see PP-interactions

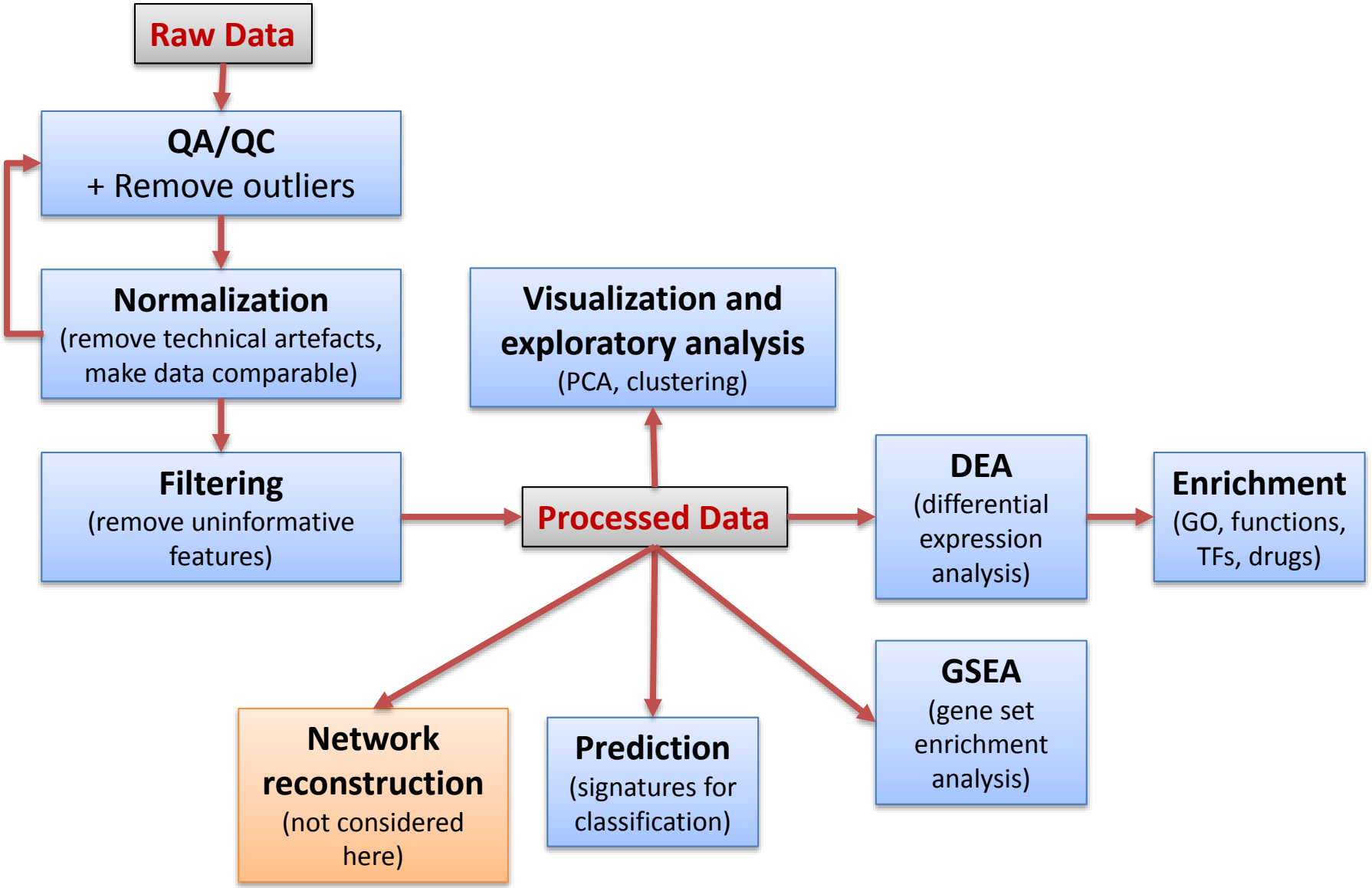
Up regulated



Down regulated



Summary


Raw Data
QA/QC
+ Remove outliers
Normalization
(remove technical artefacts, make data comparable)
Filtering
(remove uninformative features)
Visualization and exploratory analysis
(PCA, clustering)
Processed Data
DEA
(differential expression analysis)
Enrichment
(GO, functions, TFs, drugs)
GSEA
(gene set enrichment analysis)
Network reconstruction
(not considered here)
Prediction
(signatures for classification)

Questions ?

Thank you for your attention !