

CanBio Course
coordinator: Prof. Thomas Sauter

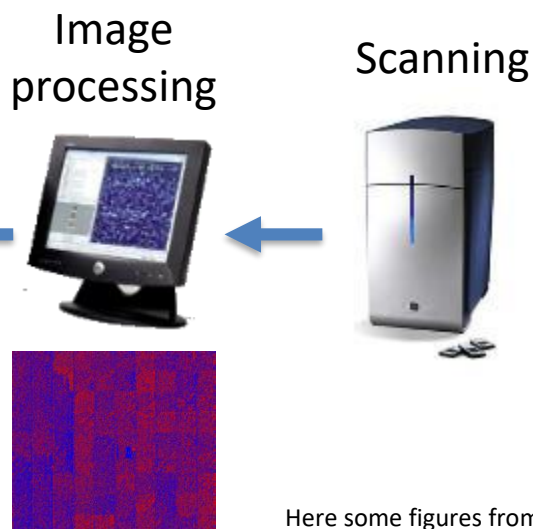
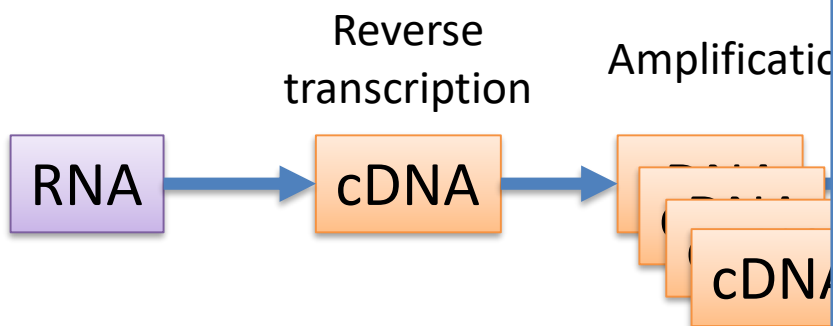
Introduction to Data Analysis

Dr. Petr Nazarov
petr.nazarov@lih.lu

2019-11-04

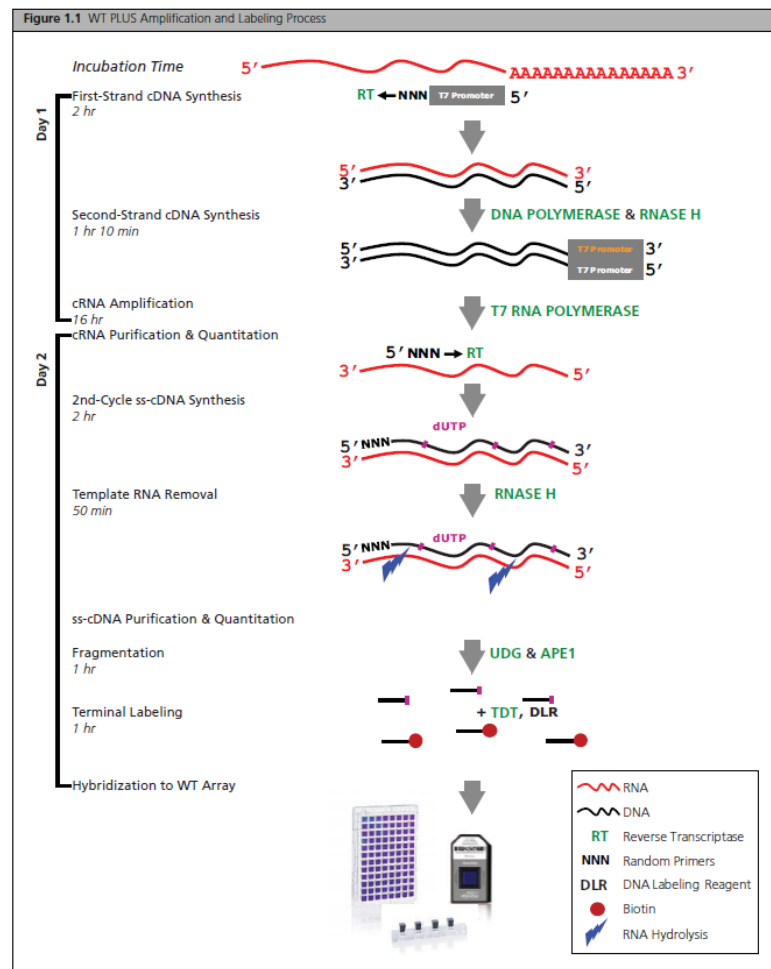
- ◆ **Data overview**
 - ◆ Microarrays
 - ◆ RNA-seq
- ◆ **Dimensionality reduction**
 - ◆ PCA, ICA, NMF
 - ◆ tSNE
- ◆ **Clustering**
 - ◆ k-means, hierarchical, dbscan
- ◆ **Differential expression analysis**
 - ◆ multiple hypotheses
 - ◆ linear models - ANOVA
- ◆ **Enrichment analysis**

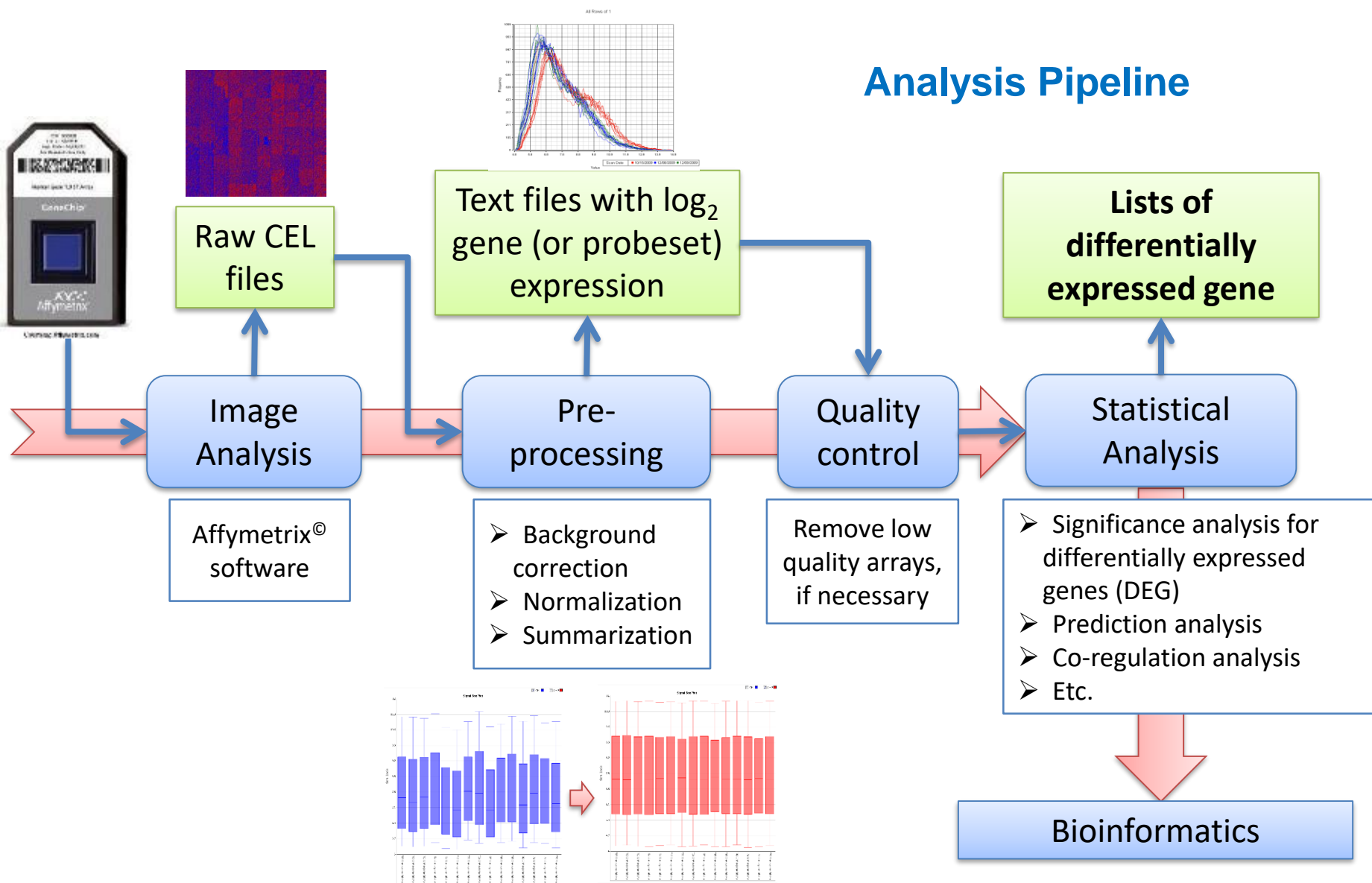
Data overview



Here some figures from Affymetrix

Assay Workflow



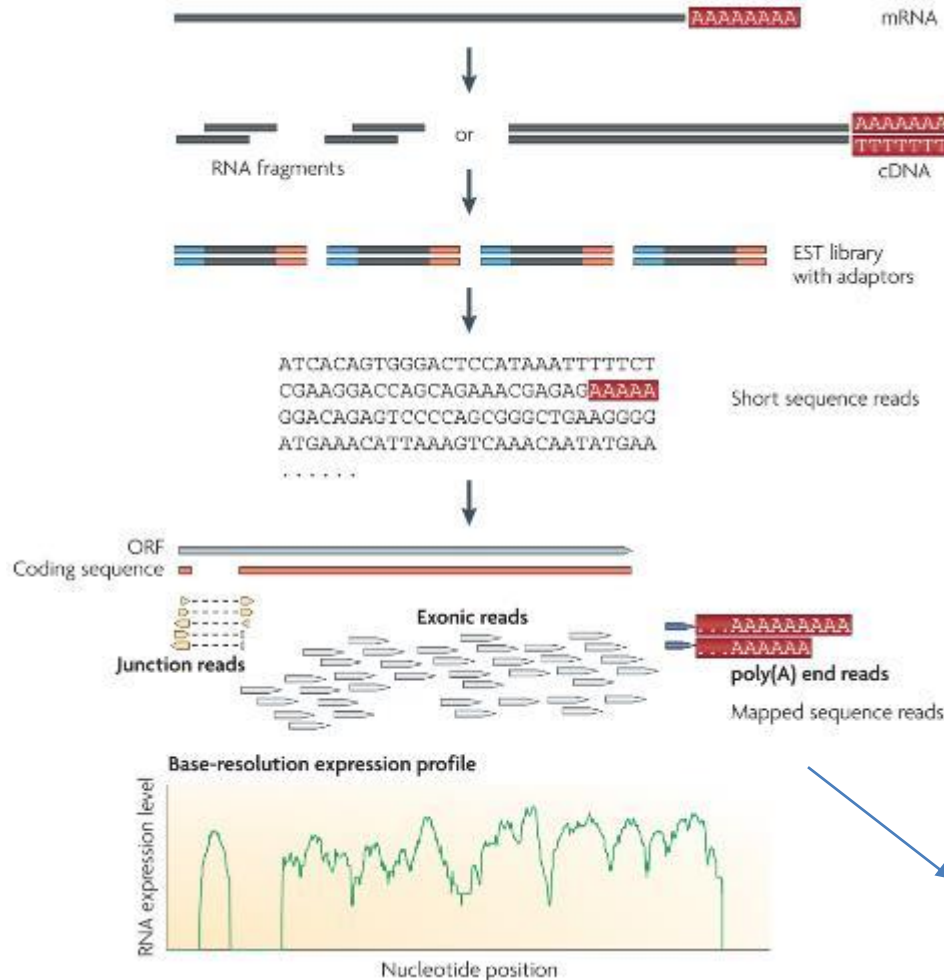


Data Example (*in log scale*)

ID	Gene.Symbol	A1	A2	A3	A4	B1	B2
TC02002853.hg.1	SP110	5.694	5.684	5.719	5.715	7.287	7.288
TC01002850.hg.1	GBP5	3.873	3.839	3.997	3.935	8.699	8.654
TC19000554.hg.1	LGALS17A	3.981	3.967	4.045	4.066	7.887	7.752
TC01006362.hg.1	GBP7	3.862	3.830	3.900	3.881	5.996	6.076
TC16000565.hg.1	SNTB2	7.765	7.734	7.748	7.755	8.973	9.027
TC12000425.hg.1	EIF4B	9.161	9.144	9.150	9.154	8.808	8.811
TC13000383.hg.1	TNFSF13B	3.922	3.890	3.873	3.918	5.151	5.199
TC09000999.hg.1	DDX58	6.629	6.661	6.671	6.598	8.302	8.367
TC06001673.hg.1	ETV7	4.427	4.467	4.434	4.348	6.815	6.713
TC05001767.hg.1	IRF1	5.409	5.470	5.552	5.396	7.988	8.000
TC17000821.hg.1	SSTR2	3.939	3.900	3.922	3.880	5.283	5.360
TC0X001551.hg.1	CLIC2	4.481	4.441	4.388	4.377	6.504	6.416
TC17000705.hg.1	MSI2	6.221	6.201	6.203	6.219	5.832	5.820
TC09000038.hg.1	PDCD1LG2	4.151	4.072	4.219	4.148	6.276	6.330
TC17001523.hg.1	DHX58	4.636	4.581	4.614	4.618	5.526	5.489
TC22000701.hg.1	APOL4	4.866	4.812	4.971	4.828	7.230	7.277
TC02001524.hg.1	ADI1	6.761	6.734	6.760	6.766	6.311	6.313
TC22000700.hg.1	APOL3	5.088	5.080	5.090	5.026	6.715	6.830
TC06000932.hg.1	NUS1	7.870	7.882	7.856	7.871	7.543	7.547
TC14001152.hg.1	GCH1	6.266	6.344	6.268	6.257	7.582	7.551

Here gene expression data are given in \log_2 intensity

Next-Generation Sequencing: RNA-seq



CPM: counts per million nt

TPM: transcripts per million (proportion)

FPKM: fragments per kilobase of exon per million reads mapped

RPKM: reads per (for single-end)

$$CPM_i = \frac{X_i}{N} = \frac{X_i}{N} \cdot 10^6 \quad TPM_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

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

raw counts

normalized counts,
CPM, FPKM, RPKM

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

Data Example (*in linear scale*)

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	14	14	14	13	85	94
ENSG00000153989	NUS1	214	216	212	214	167	167
ENSG00000131979	GCH1	57	61	57	56	172	167

Here gene expression data are given in counts

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 2015 – more than 10k patients

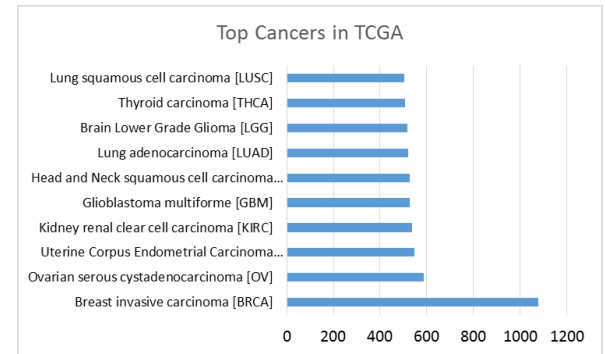
ArrayExpress: <http://www.ebi.ac.uk/arrayexpress/>



Data Content

Updated today at 06:00

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



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

Take Home Messages

- ◆ **Microarrays should be normalized** to remove effects of variable RNA content. **RNA-seq can be normalized as well**, if analyzed using linear models. No need to normalize if special methods are used: edgeR, DESeq2

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

- ◆ Main advantage of RNA-seq data: they **can be reprocessed and reused** taking into account new genomic annotation or asking new questions

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

Dimensionality reduction

Principal Component Analysis (PCA)

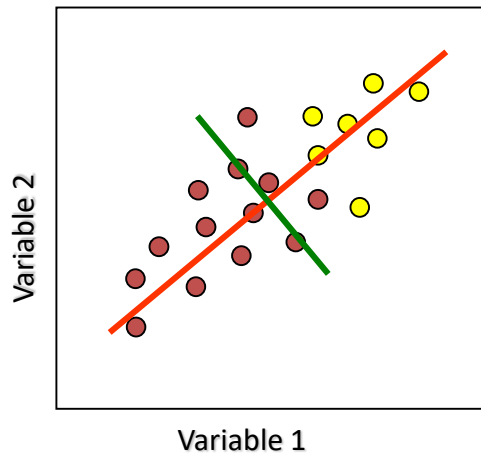
Principal component analysis (PCA)

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.**

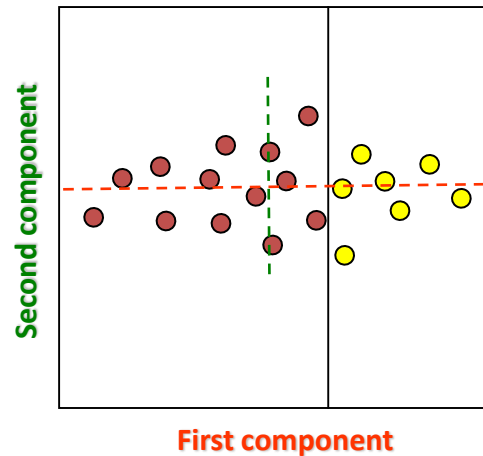
20000 genes →
2 dimensions

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

Scatter plot in
“natural” coordinates



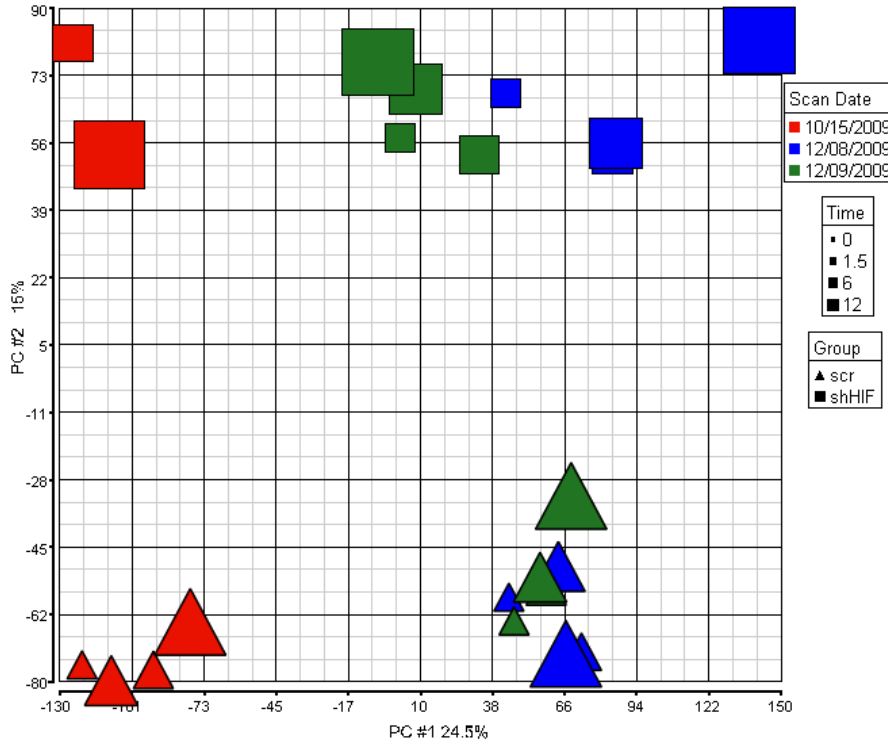
Scatter plot in PC



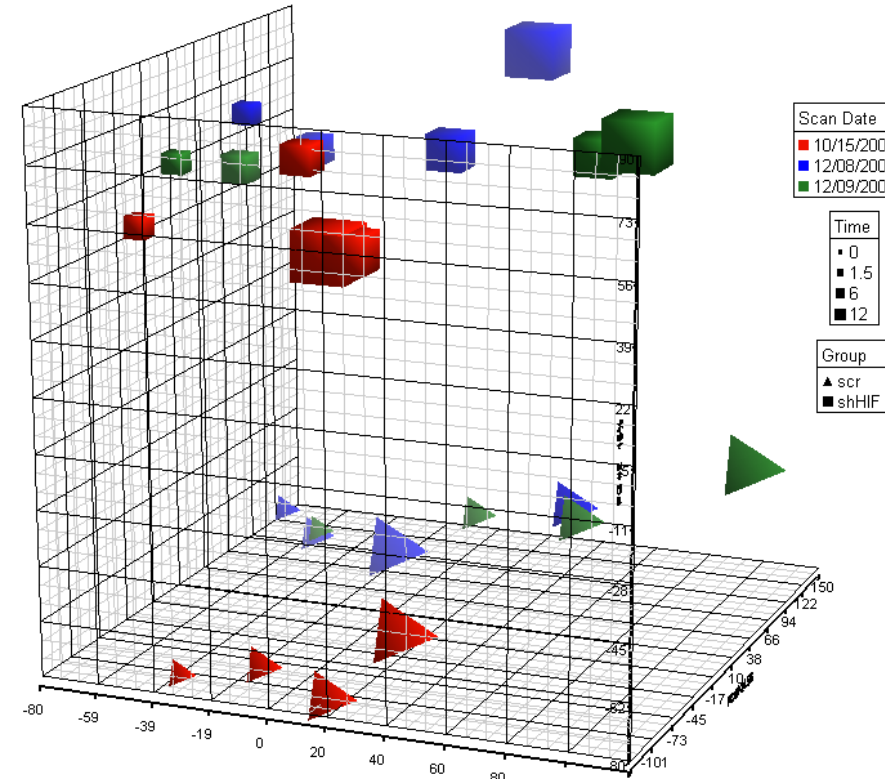
Instead of using 2 “natural” parameters for the classification, we can use the first component!

PCA

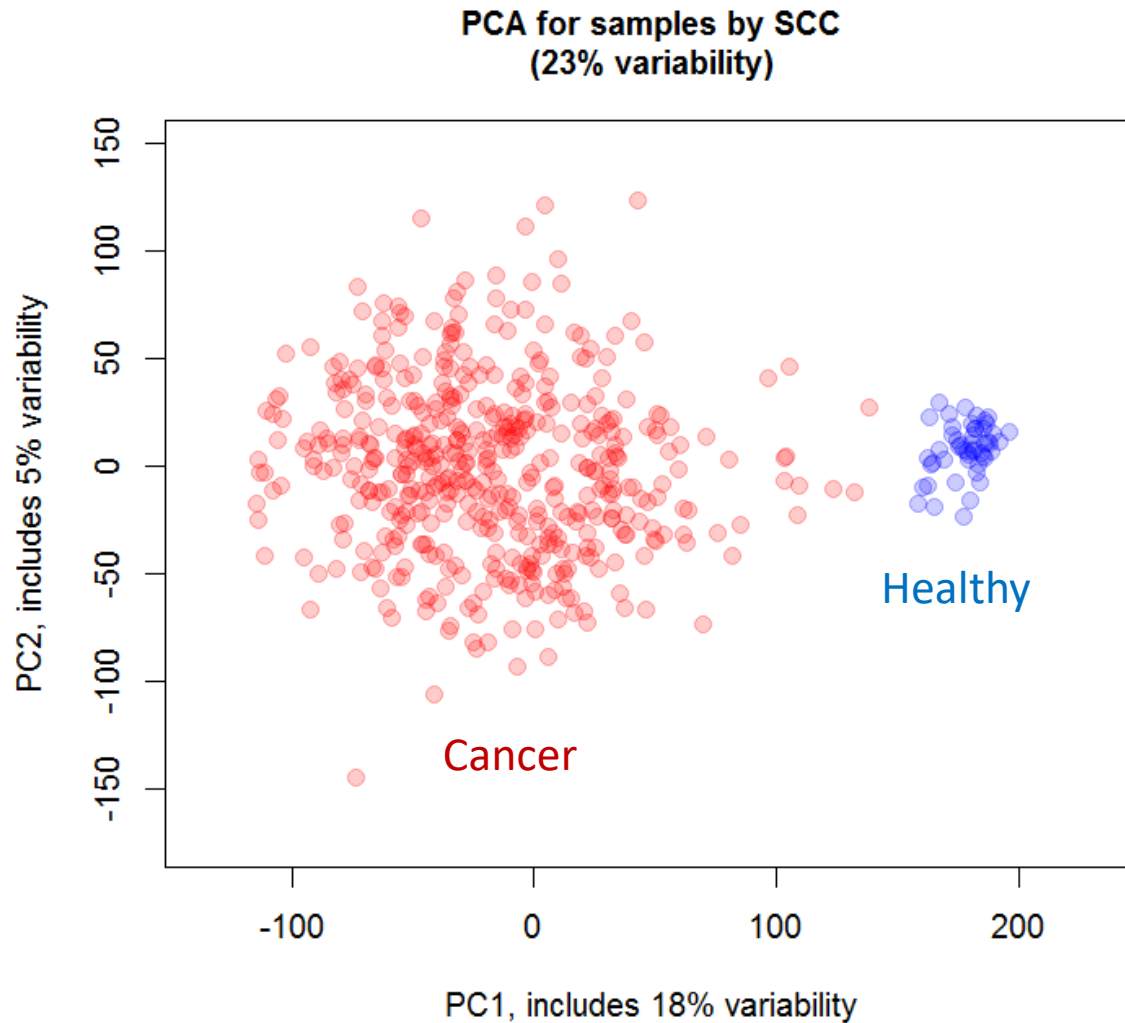
PCA Mapping (39.5%)



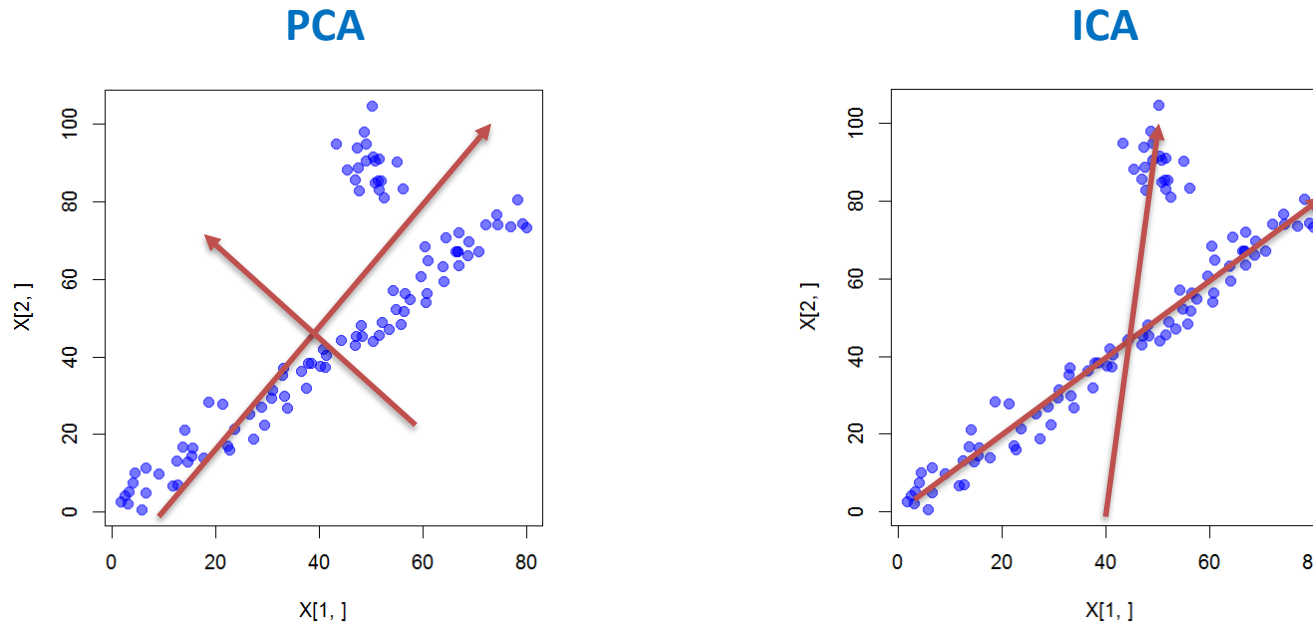
PCA Mapping (48.4%)



PCA in TCGA (LUSC data)



Independent Component Analysis (ICA)

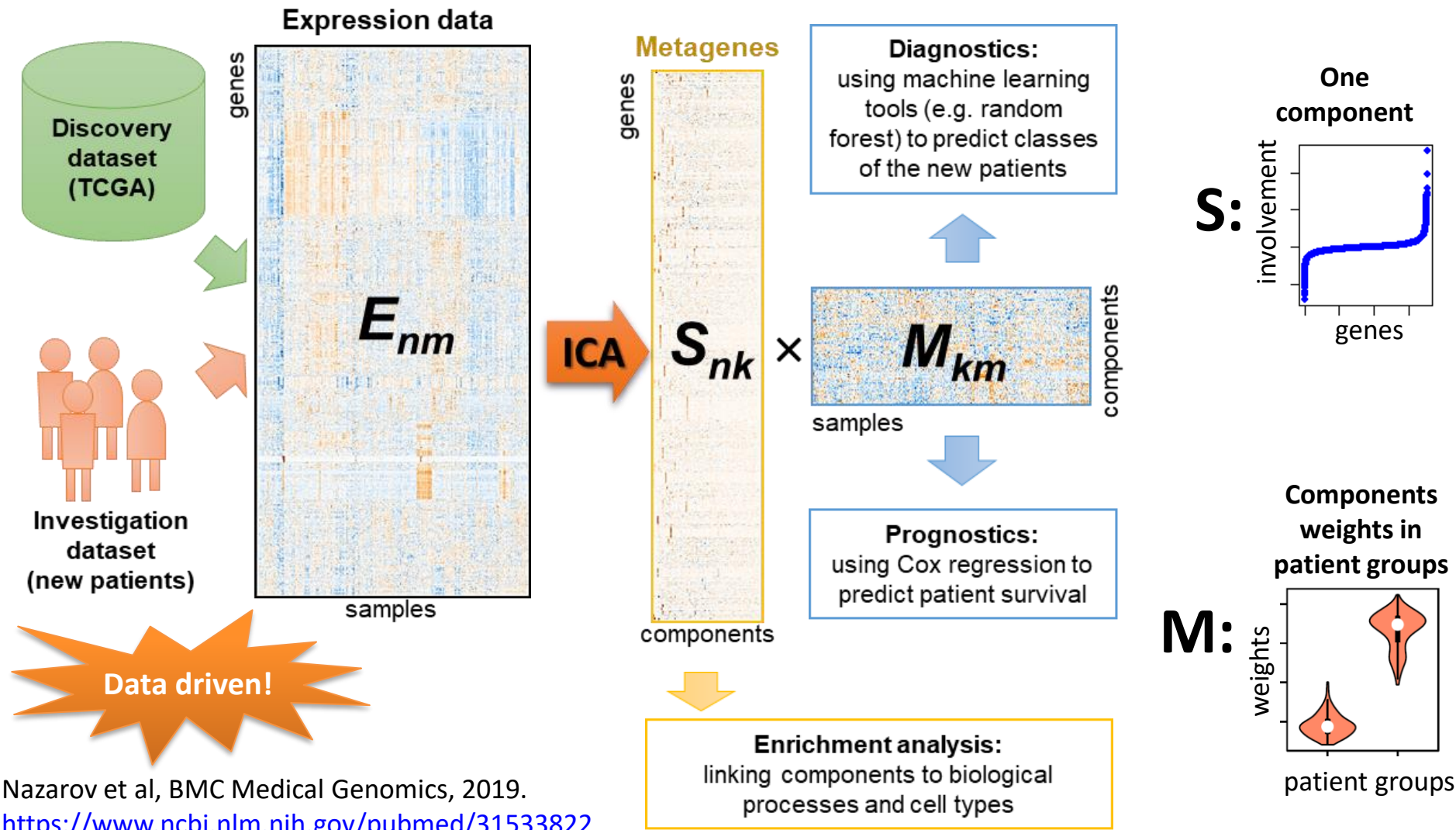


Properties of ICA:

- Separate into **statistically independent** signals
- **Random order and direction** of the components
- **Multiple runs** is advised to ensure stable solution (consensus)
- Usually, **ICs represent biology better than PCs**, esp. when signals are correlated (e.g. transcriptomes of different leukocytes)

Sompairac et al, Int J Mol Sci, 2019. <https://www.ncbi.nlm.nih.gov/pubmed/31500324>

Independent Component Analysis (ICA)

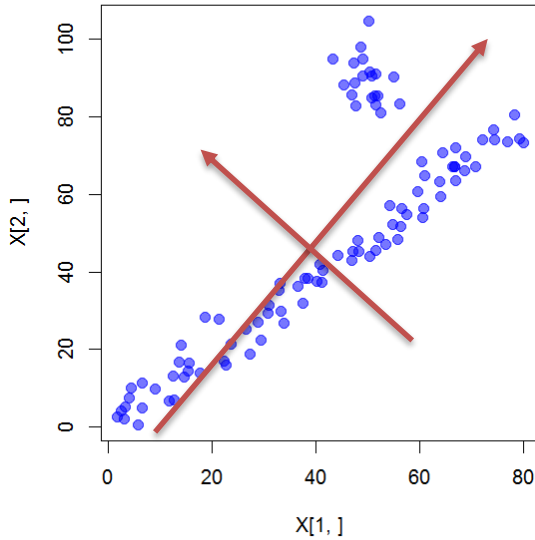


Data driven!

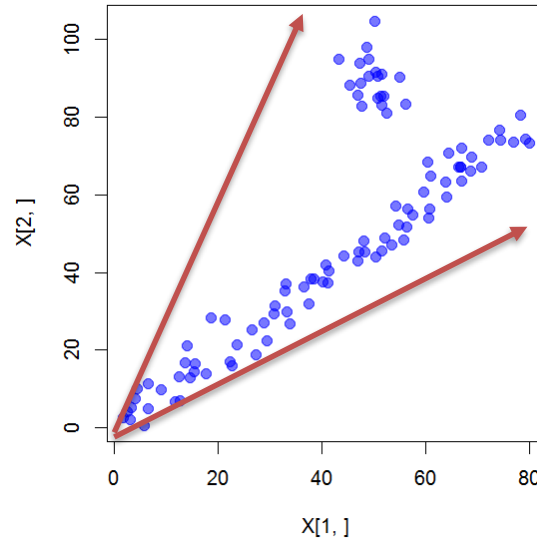
Nazarov et al, BMC Medical Genomics, 2019.
<https://www.ncbi.nlm.nih.gov/pubmed/31533822>

Non-negative Matrix Factorization (NMF)

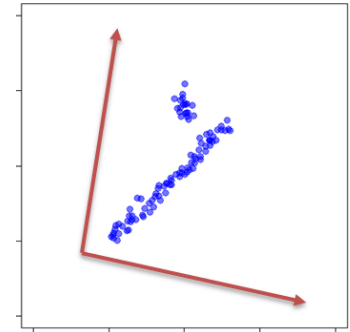
PCA



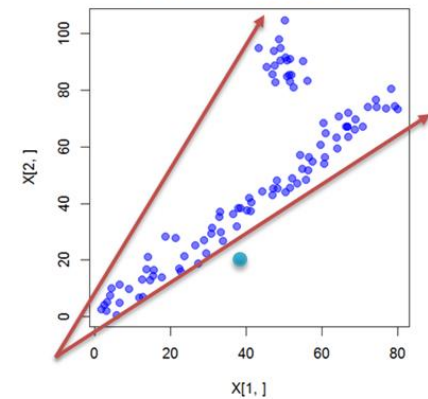
NMF



NMF: issue 1



NMF: issue 2



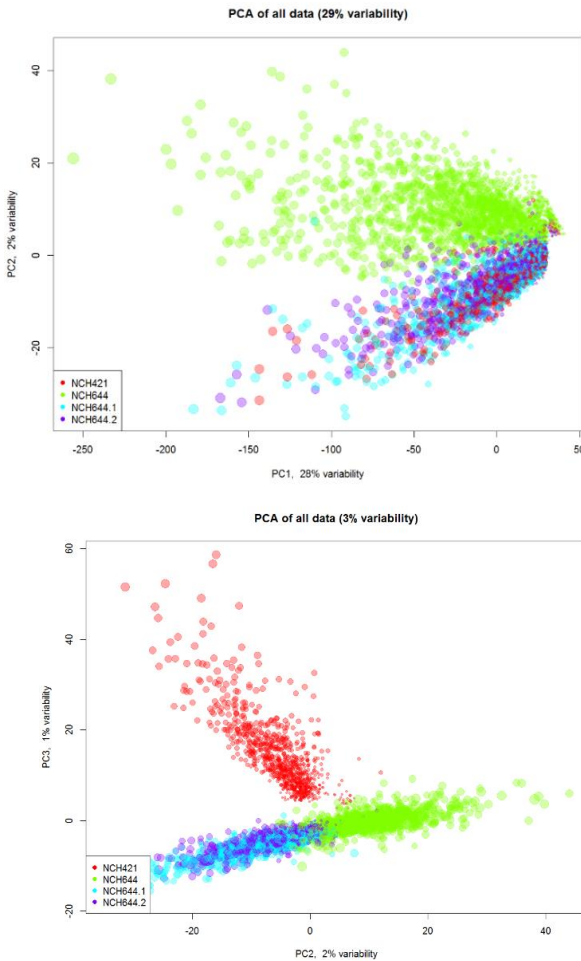
Properties of NMF:

- Separate into **non-negative mixture of non-negative signals**
=> easy to interpret from physical principles
- Solution is **not unique**, additional conditions are required
- **Multiple runs** is advised to ensure stable/best solution (consensus)
- Works perfect for methylation. Not so good (debatable 😊) for transcriptomics

Non-linear Dimensionality Reduction: t-SNE

t-Distributed Stochastic Neighbor Embedding

PCA of single-cell data



PCA captures variability => distant data points have larger effect. This can totally mask small variability in which we are interested

t-SNE is an iterative non-linear transformation that search for objects representation in 2D space by:

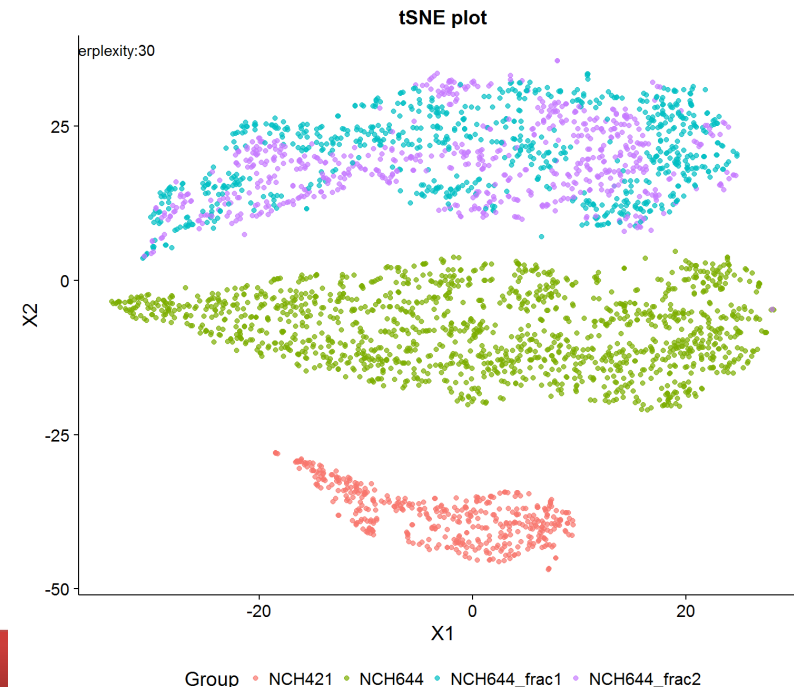
- 1) placing the similar objects together
- 2) controlling the density of the obtained clusters

Unlike PCA, distant objects do not influence t-SNE!

Modern alternative:

UMAP

<https://arxiv.org/abs/1802.03426>



t-SNE for single cell transcriptomics

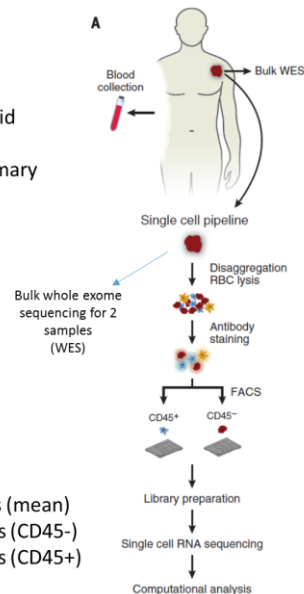
RESEARCH ARTICLES

CANCER GENOMICS

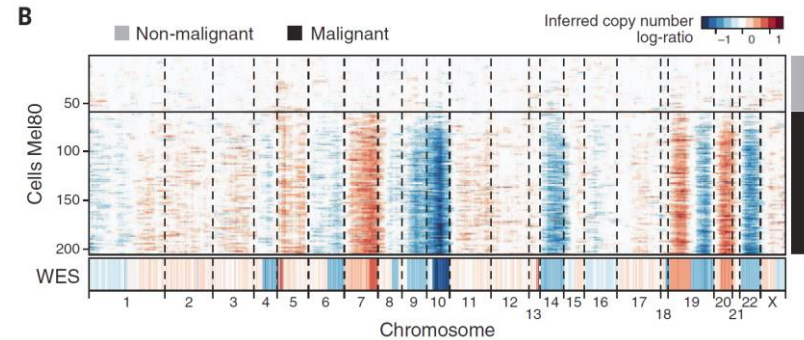
Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq

Itay Tirosh,^{1,*} Benjamin Izar,^{1,2,3,*} Sanjay M. Prakadan,^{1,4,5,6}
 Mara H. Wadsworth II,^{1,4,5,6} Daniel Treacy,¹ John J. Tomsho,¹ Asaf Rotem,^{1,2,3}

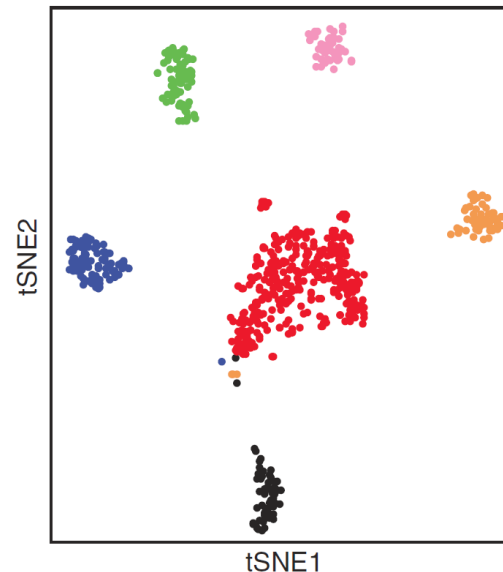
- 19 tumors:
 • 10 lymphoid
 • 8 distant
 • 1 acral primary



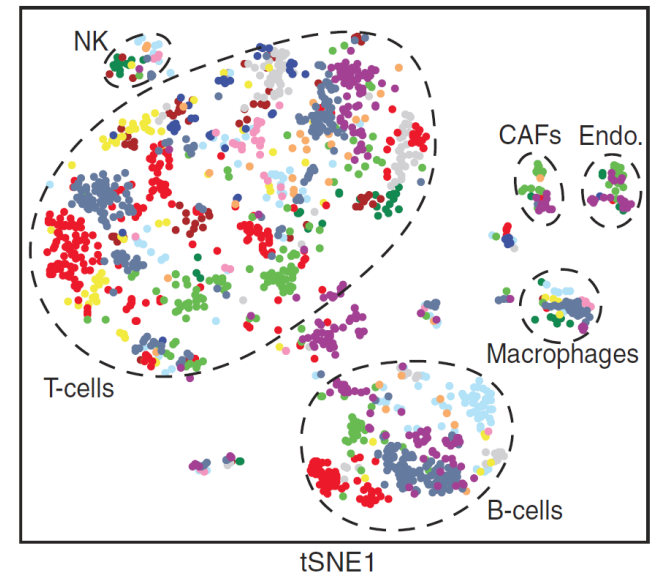
- 4645 cells:
 • 150k reads (mean)
 • 4659 genes (CD45-)
 • 3438 genes (CD45+)



C Malignant cells



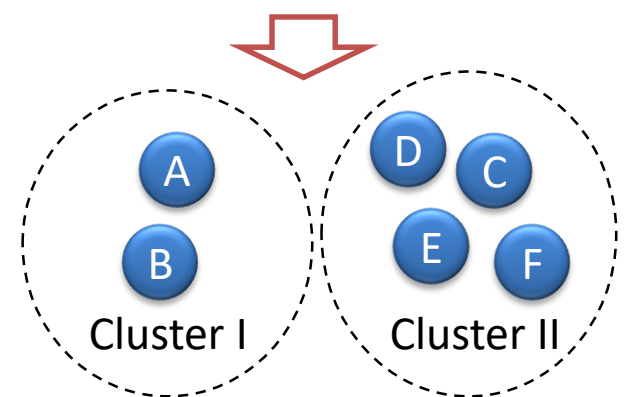
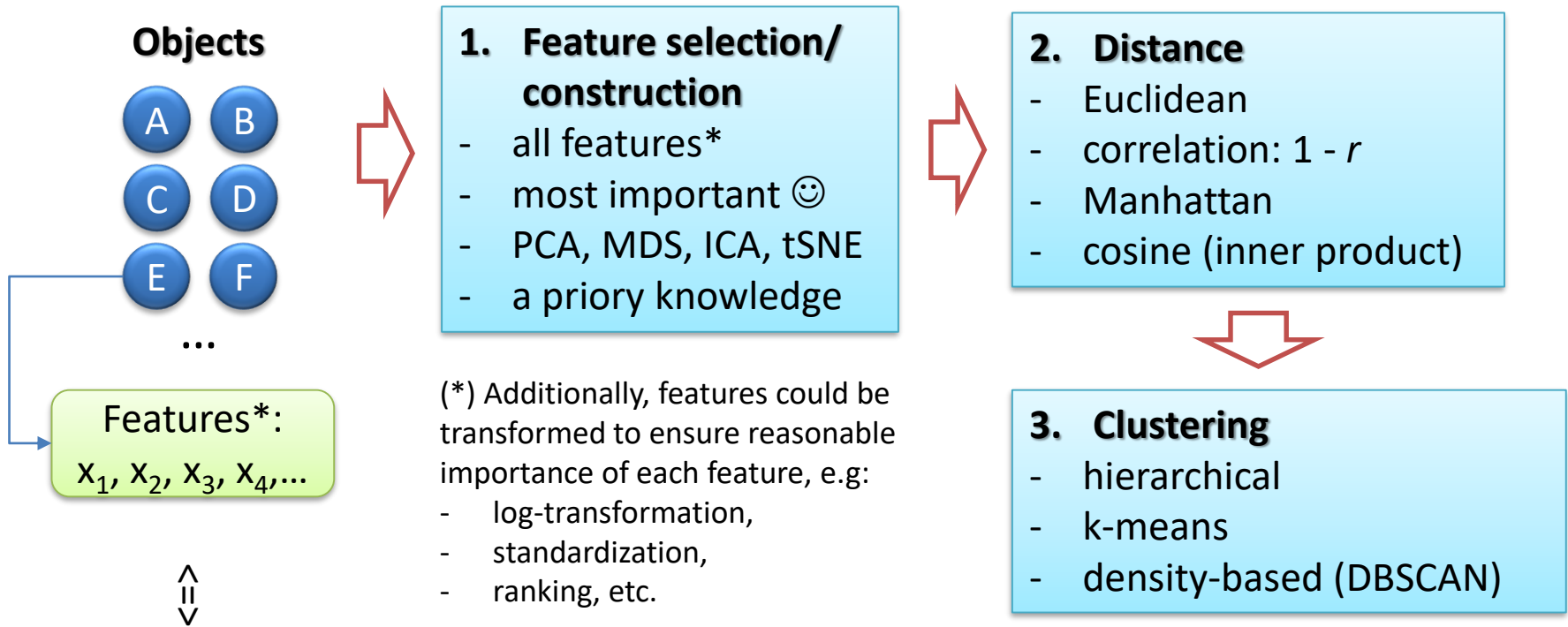
D Non-malignant cells



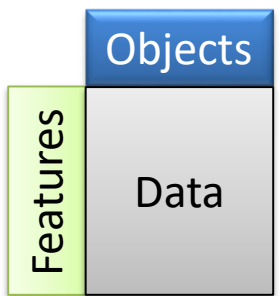
Take Home Messages

- ◆ Start your investigation with **PCA**, which will help
 - ◆ **Reduce dimensionality** and **help visualizing** your data
 - ◆ See which **factors** may play the **important role** in your data
 - ◆ **Find outlier** experiments
- ◆ **ICA** allows detecting **more biology-related** signals in the data, but requires large datasets to be applicable (~ 4 observations per component)
- ◆ **NMF** provides an **easy-to-interpret** deconvolution and dimensionality reduction, but have reduced robustness.
- ◆ **t-SNE** and **UMAP** can help visualizing large datasets, focusing on the **similarity between objects**: apply for single-cell studies, pan-cancer studies or visualizing features, such as genes and proteins

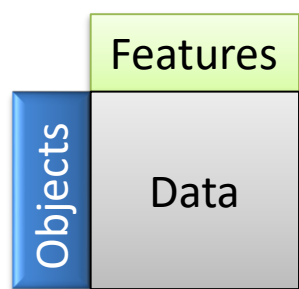
Clustering



In genomics:



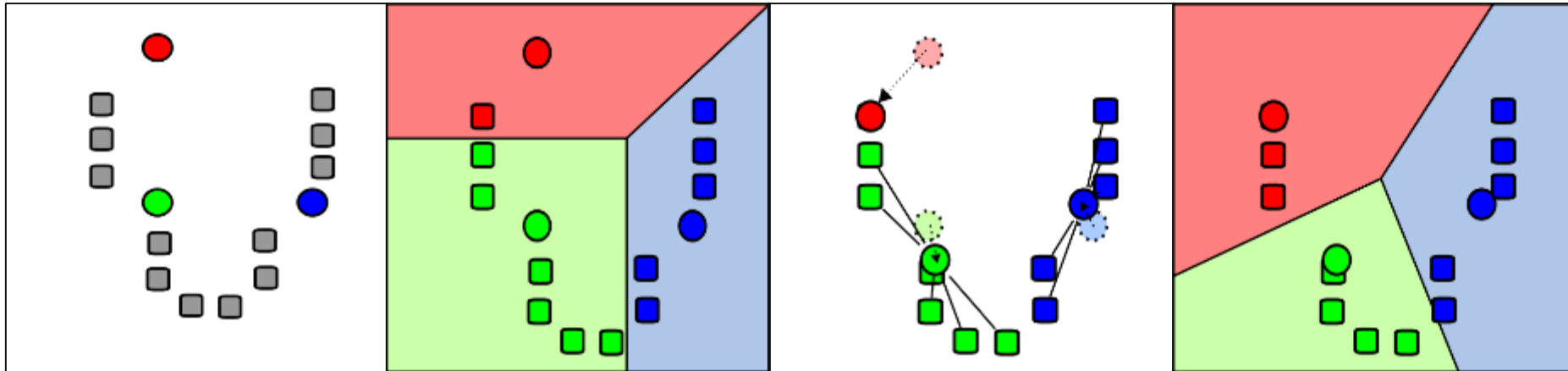
In machine learning:



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.

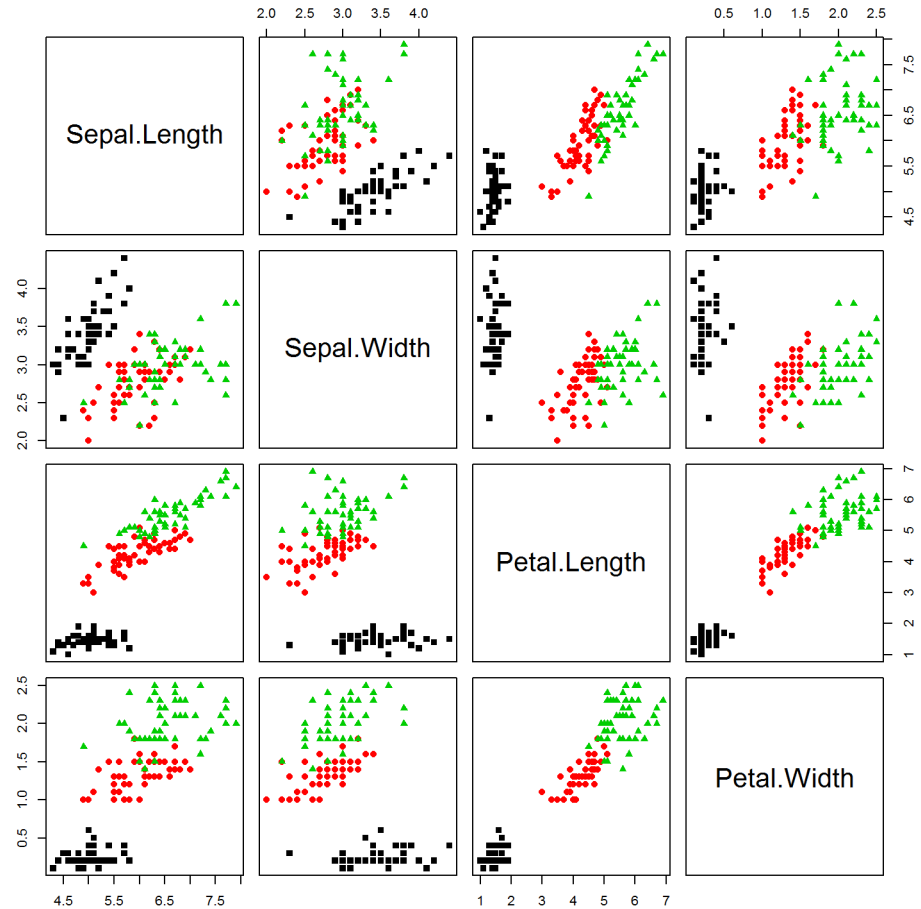
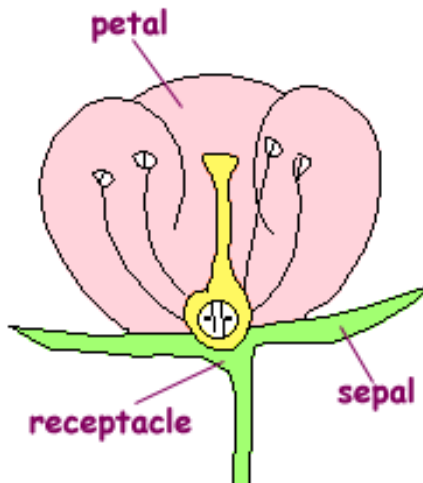
3) *The centroid of each of the k clusters becomes the new means.*

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

<http://wikipedia.org>

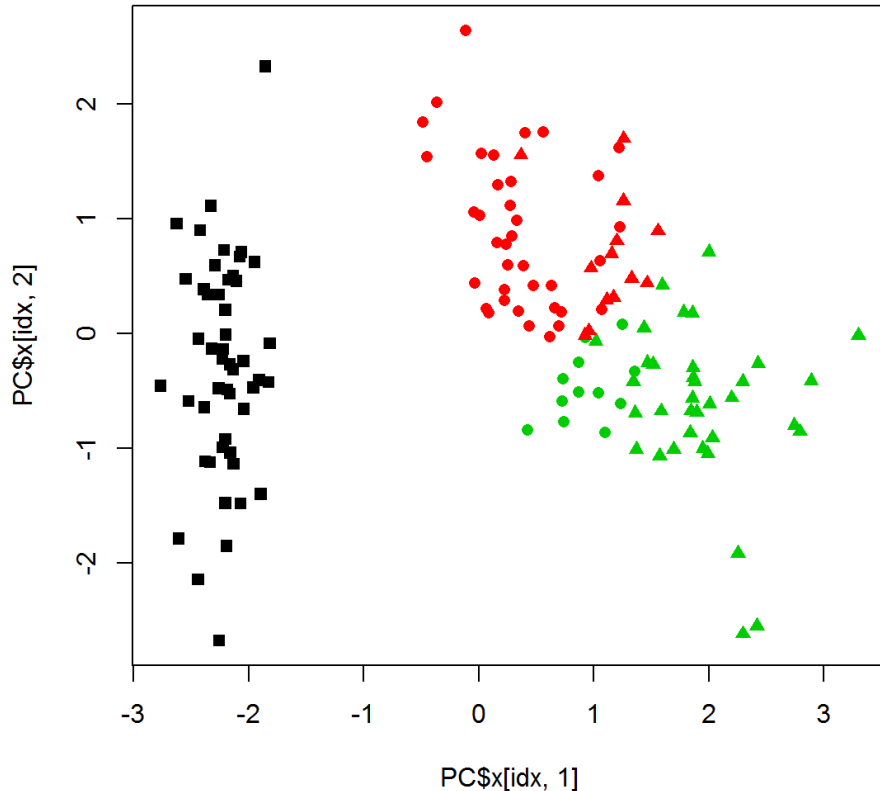
Iris Dataset (Fisher)

```
print(iris)
str(iris)
plot(iris[, -5],
     col=as.integer(iris[, 5]),
     pch=19)
```



How could we possibly represent these data on a single plot?

k-Means Clustering



Quite robust! But sensitive to outliers

```
cl = kmeans(X,
            centers=3,
            nstart=10)$cluster
```

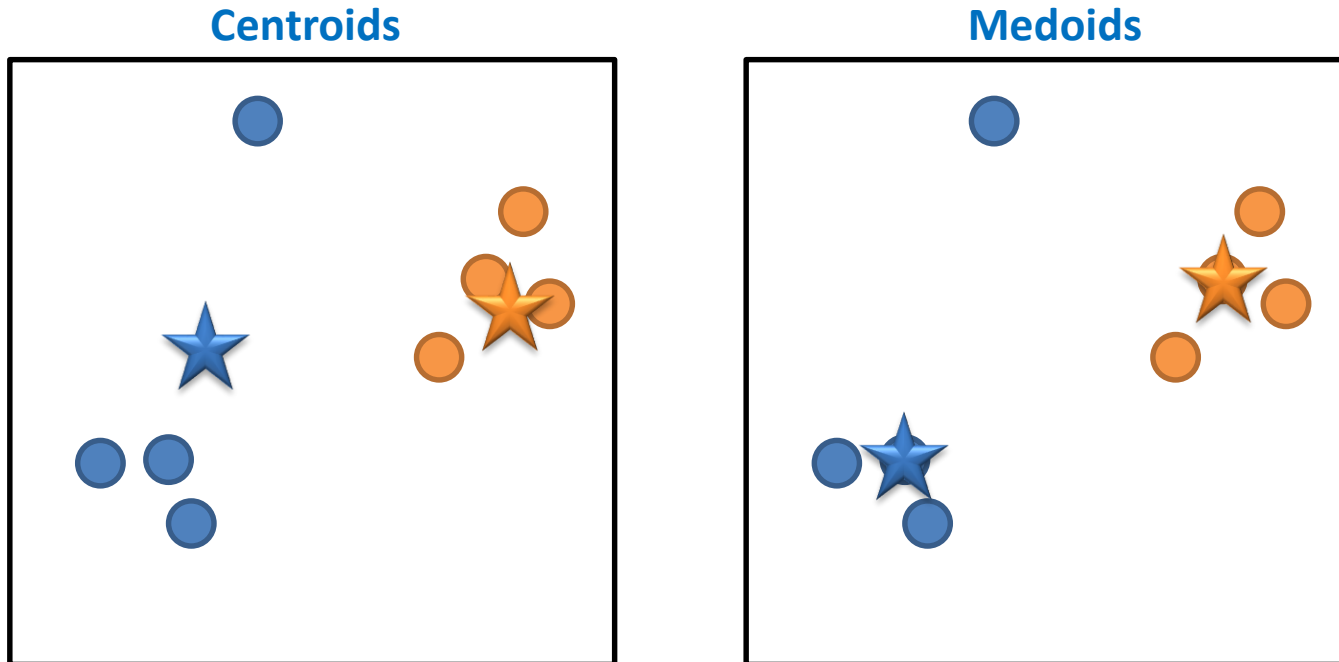
```
plot(PC$x[,1], PC$x[,2],
     col = cl, pch = point)
```

Let's remove 1 flower from each group
(-41,-98,-144) and repeat clustering:

	1	2	3
1	46	0	0
2	1	51	0
3	0	0	49

PAM: Partitioning Around Medoids (k-medoids)

PAM is a version of “k-means” that is more robust to outliers. Instead of calculated centroids, it uses medoids – representative objects of each class.



```
library(cluster)
cl = pam(X,k=3,nstart=10)$cluster
plot(PC$x[,1],PC$x[,2],col = cl, pch=point)
```

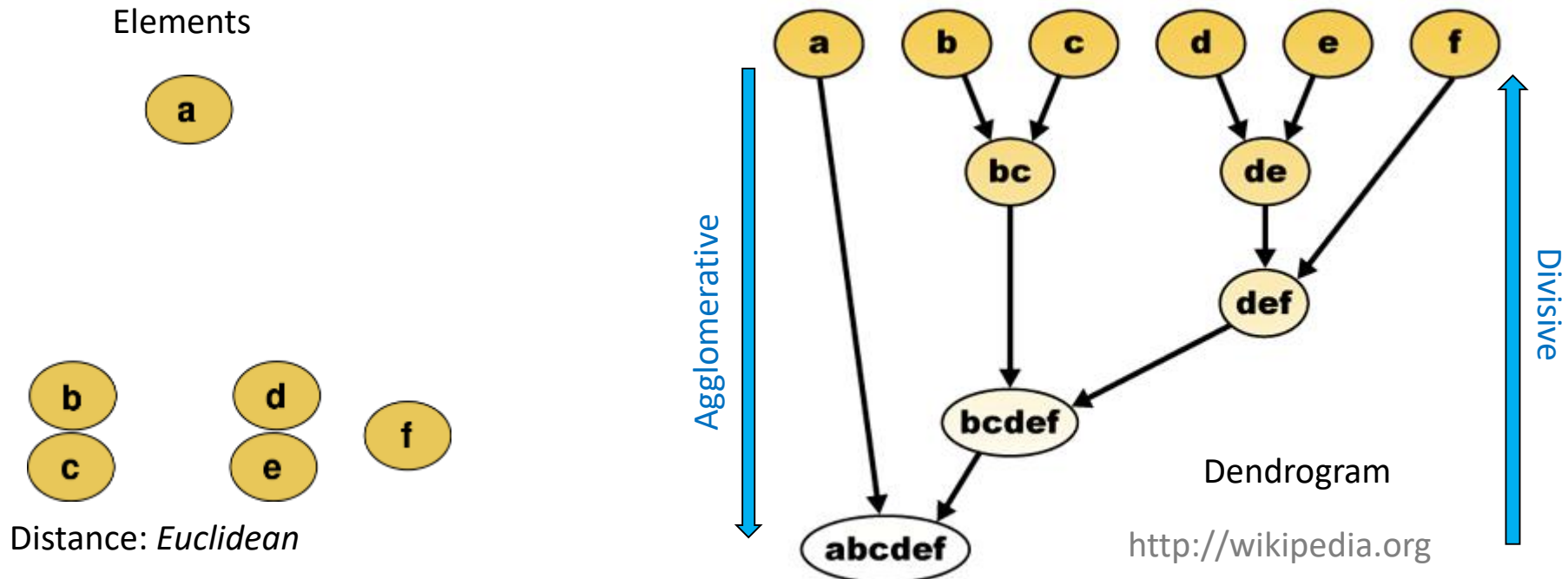
Clustering

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.

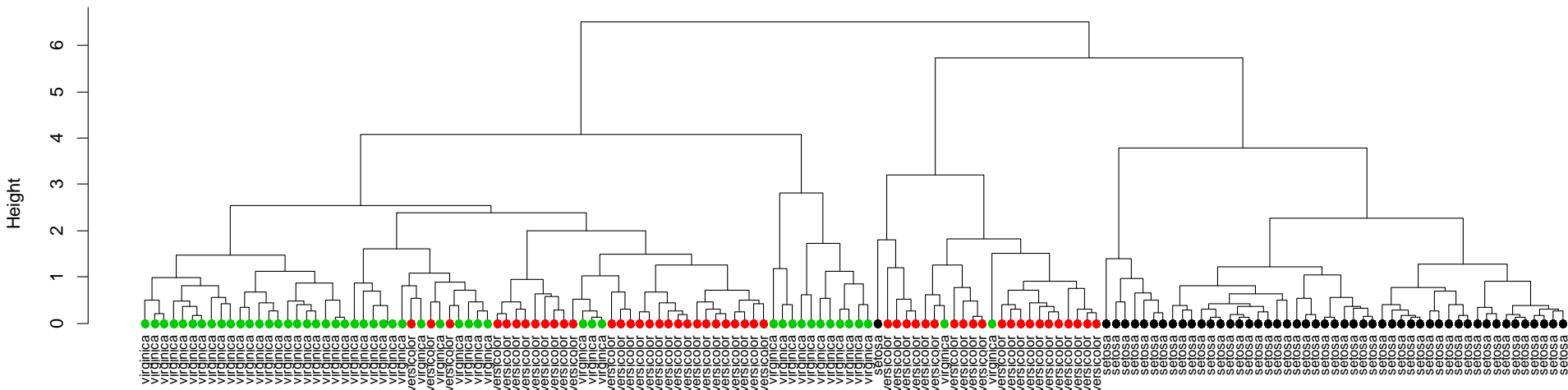


Iris dataset

```
H = hclust(dist(X))
plot(H,
      labels=iris$Species,
      hang=-1,
      cex=0.75)
```

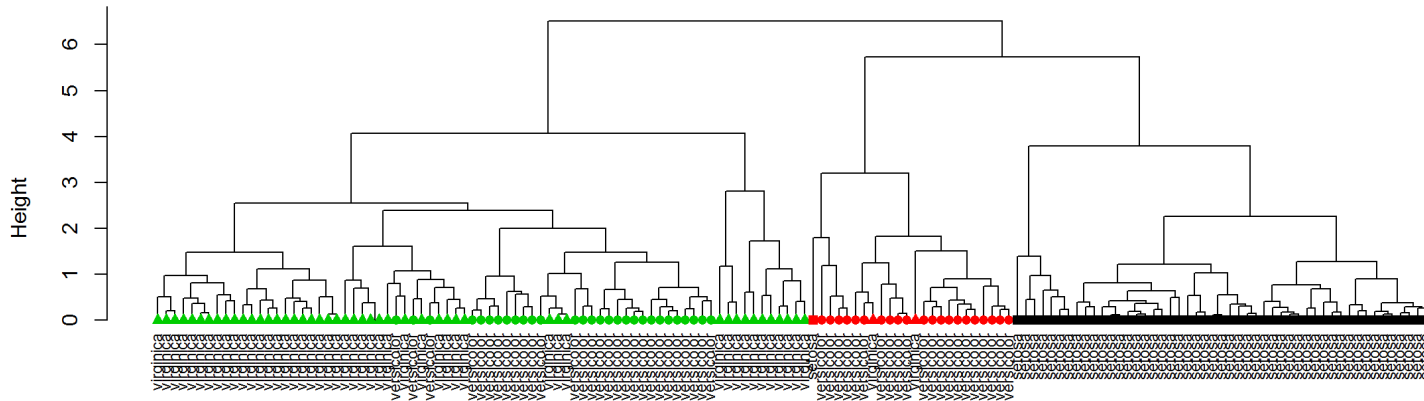
```
## optional - add points
points(x=1:nrow(X),
       y=rep(0,nrow(X)),
       pch=19,
       col=color[H$order])
```

Cluster Dendrogram



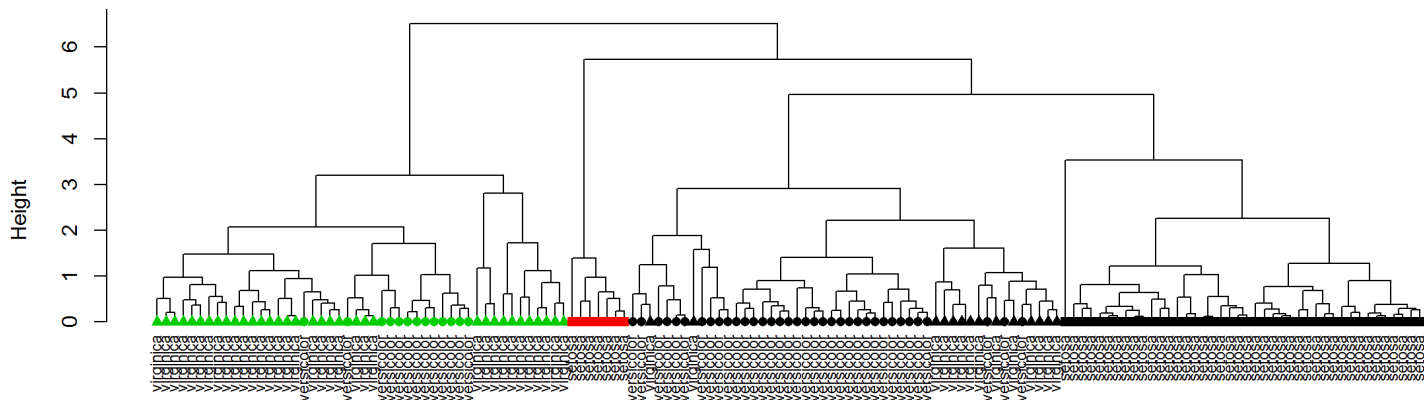
Problem: low stability

Cluster Dendrogram



Let's remove 1 flower from each group (-41,-98,-144) and repeat clustering:

Cluster Dendrogram after Removing

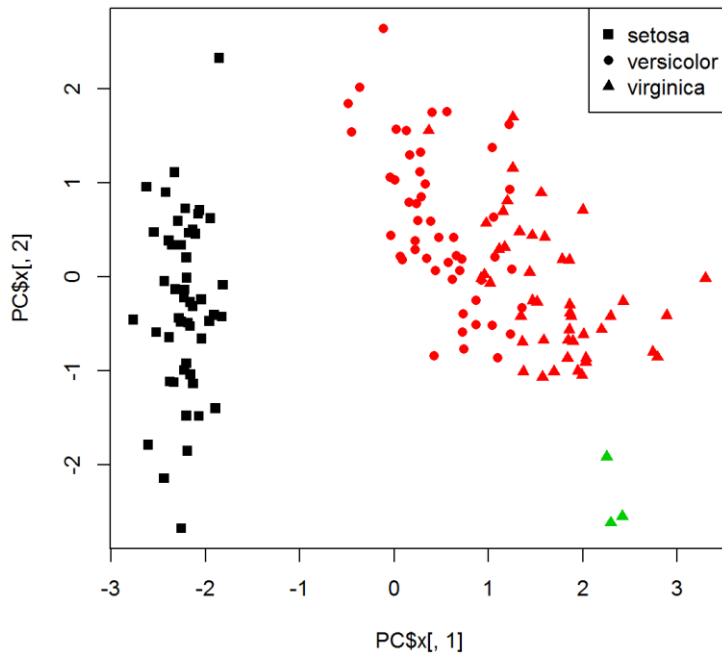


		Clustering 2		
		1	2	3
Clustering 1	1	41	7	0
	2	24	0	0
	3	27	0	48

Consensus Clustering

1. Resampling of the original set
2. Clustering
3. Summarizing the results

However, no guarantee that you will get what you expect... 3 group consensus HC:

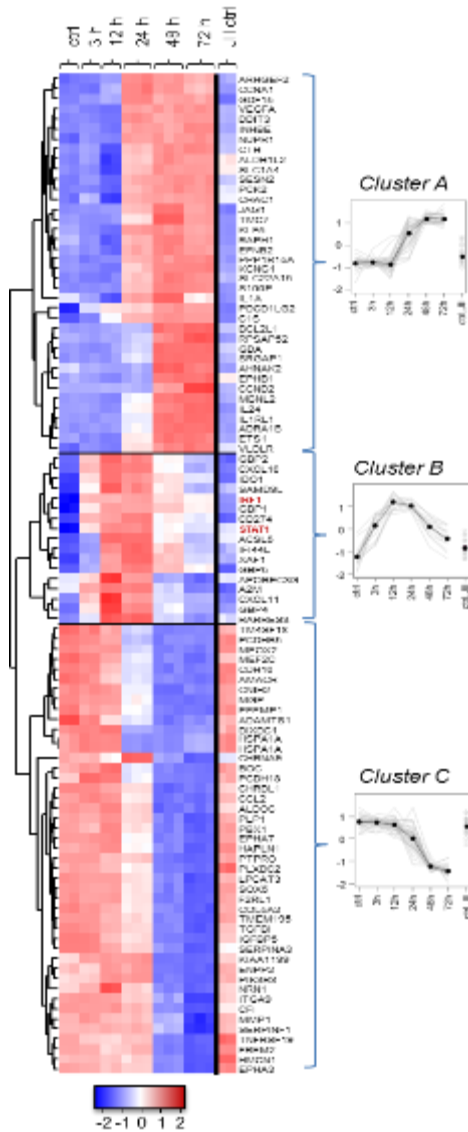


```
library(ConsensusClusterPlus)
```

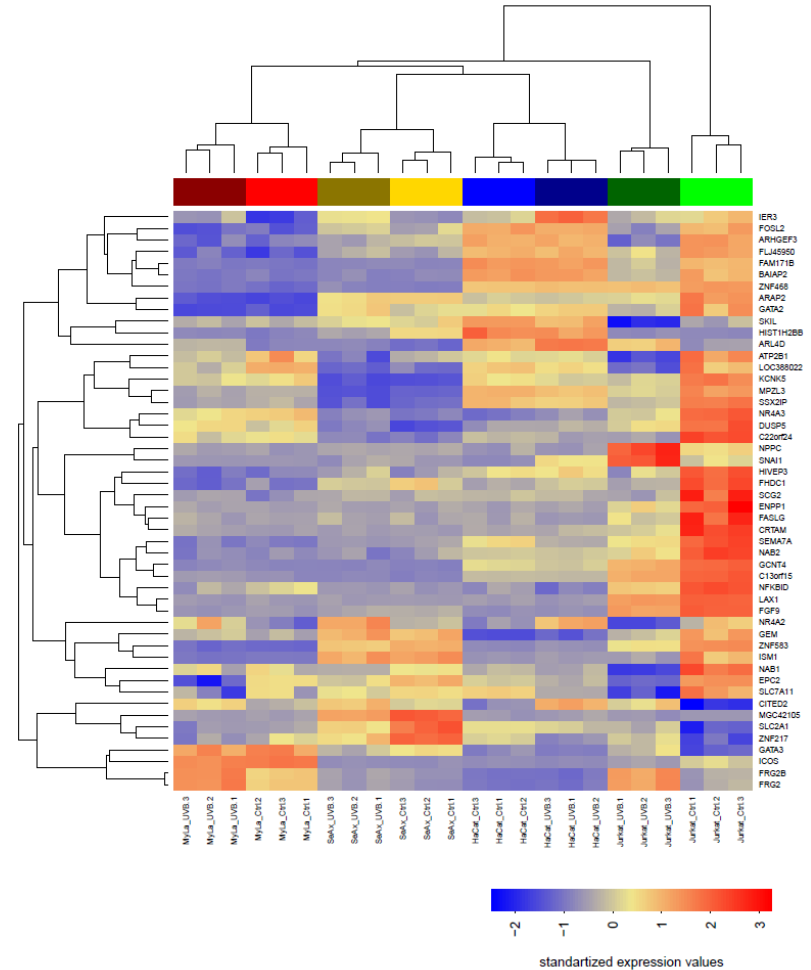
```
results = ConsensusClusterPlus(
  t(X), maxK=6, reps=50, pItem=0.8, pFeature=1,
  title="IRIS ConClust", clusterAlg="hc",
  distance="euclidean", seed=12345, plot="png")
```

```
plot(PC$x[, 1], PC$x[, 2],
  col=results[[3]]$consensusClass,
  pch=point)
```

Heatmaps



$$\text{Diff.SeAx.Jurkat} = (\text{SeAx,UVB} - \text{SeAx,Ctrl}) - (\text{Jurkat,UVB} - \text{Jurkat,Ctrl})$$



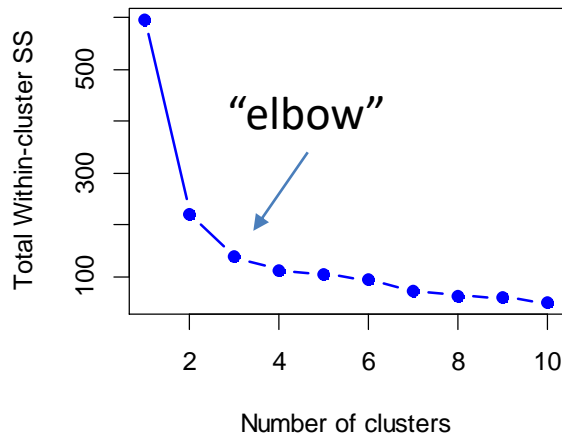
Number of Clusters

There is no universal (“magical”) solution, so aim at:

- method that gives most logical clustering (e.g. on “training” set)
- method that you could defend in your paper 😊

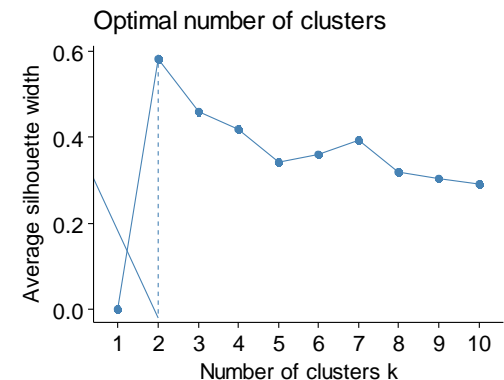
Elbow

Minimizes within-cluster sum of square (WSS)



Silhouette

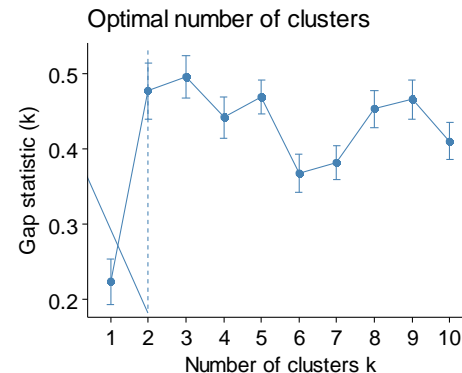
Silhouette – similarity to own cluster members compared to members of other clusters



Gap statistics

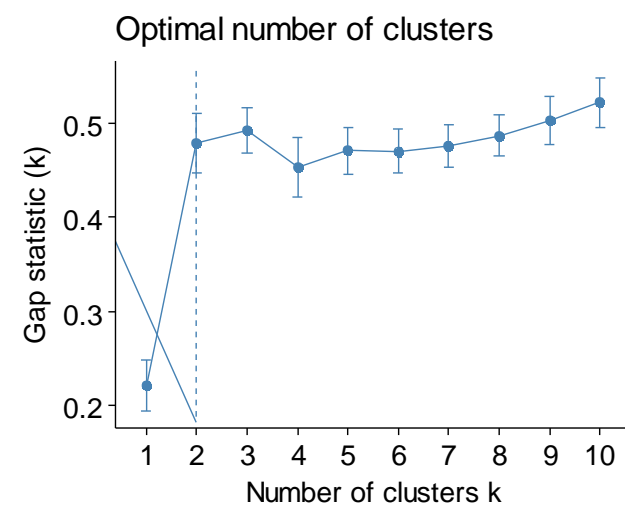
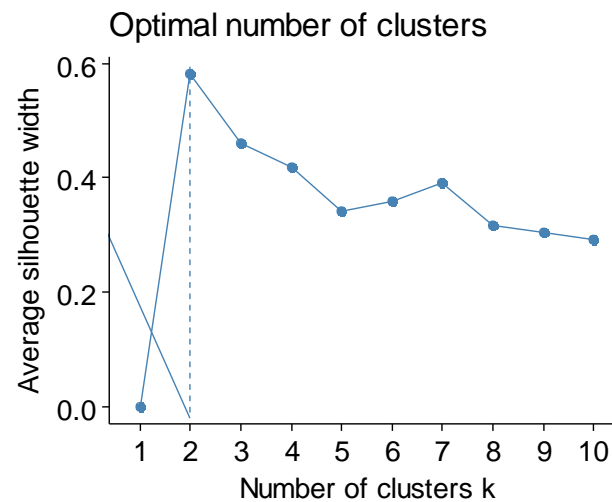
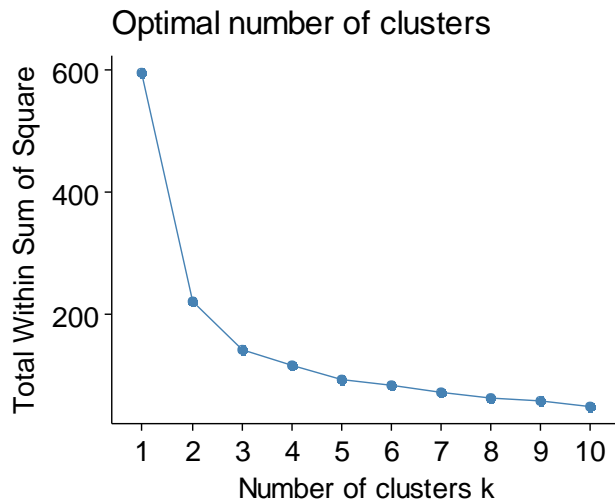
<http://web.stanford.edu/~hastie/Papers/gap.pdf>

Comparing intra-cluster variation to variation in random case

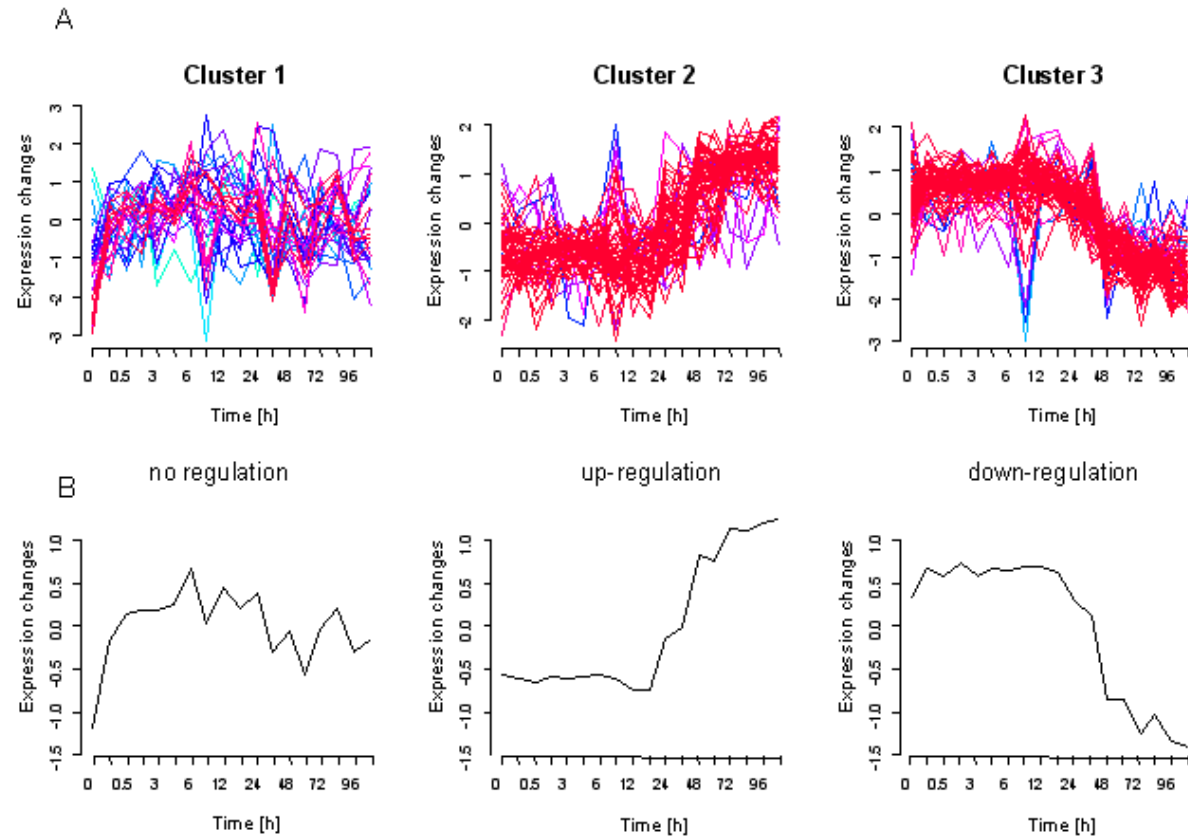
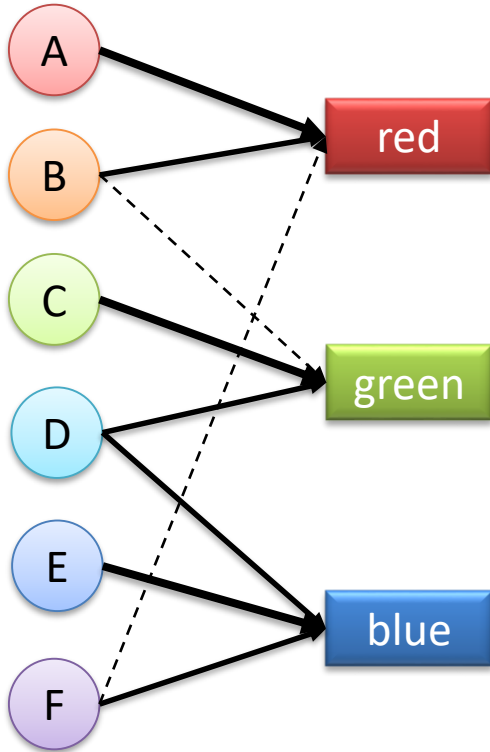


Number of Clusters

```
library(cluster)
library(factoextra)
library(NbClust)
fviz_nbclust(X, pam, method = "wss")
fviz_nbclust(X, pam, method = "silhouette")
fviz_nbclust(X, pam, method = "gap_stat")
```



Fuzzy Clustering: Mfuzz

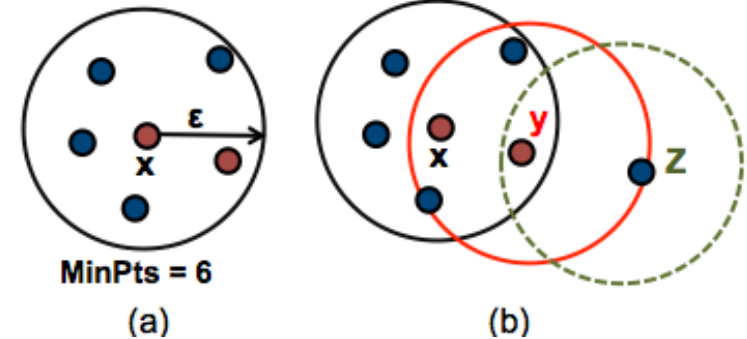


Density-based: DBSCAN

Important parameters:

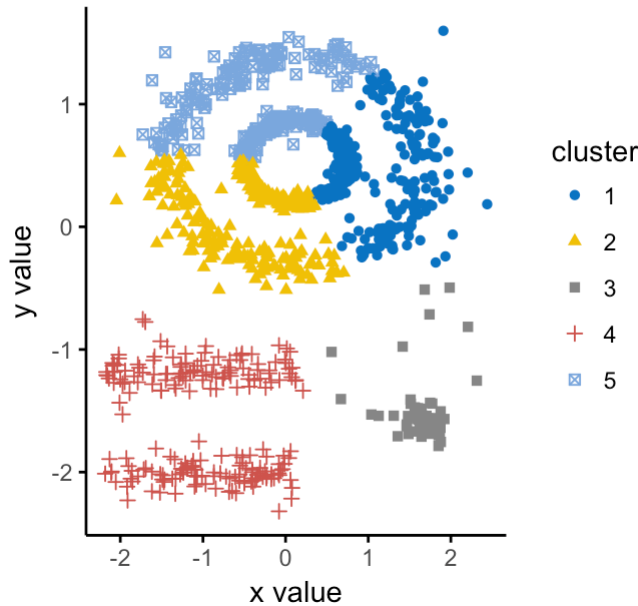
- Epsilon (ϵ)
- Minimal number of points (MinPts)

No need to define number of clusters!



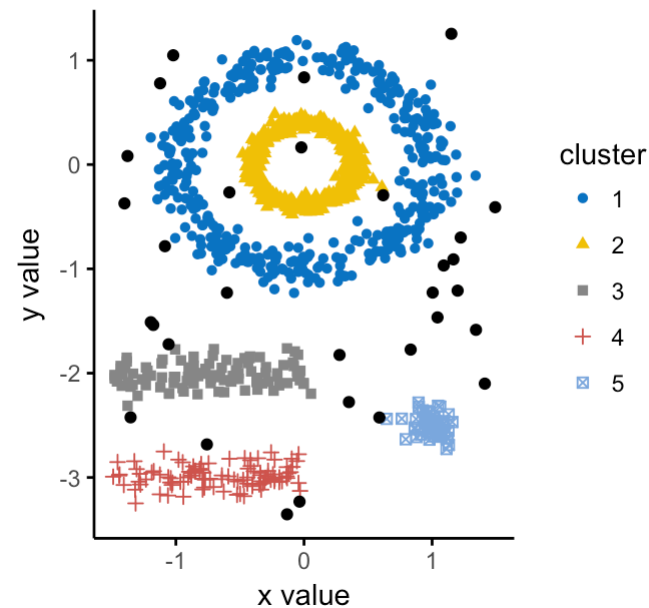
kmeans

Cluster plot



DBSCAN

Cluster plot



<http://www.sthda.com/english/articles/30-advanced-clustering/105-dbscan-density-based-clustering-essentials/>

DBSCAN

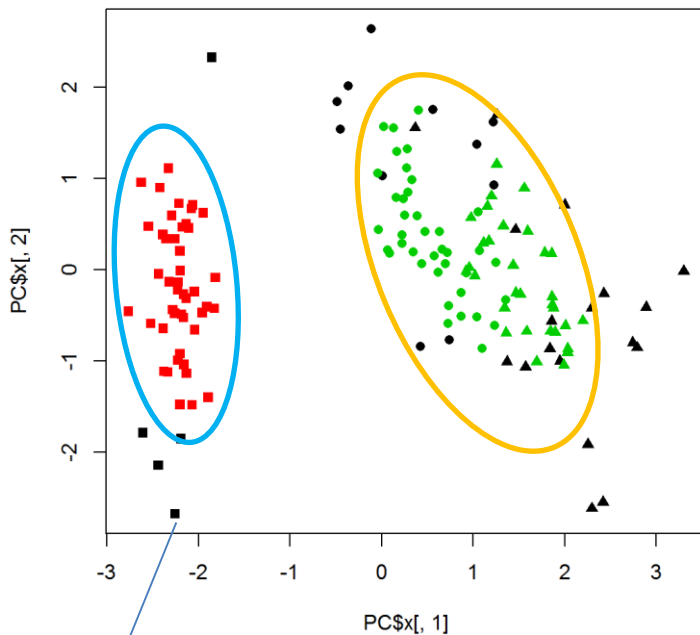
`library(dbscan)`

```
res = dbscan(X, eps=0.5, minPts=5)
```

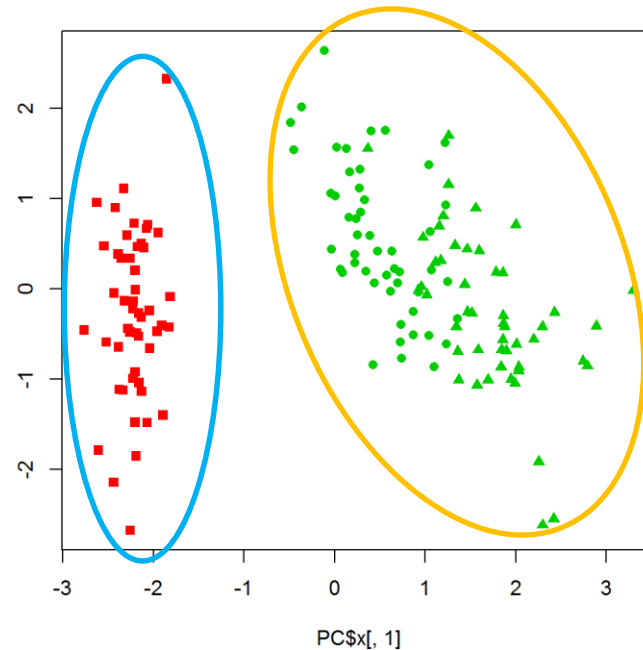
`res`

```
plot(PC$x[,1], PC$x[,2], col = 1+res$cluster, pch=point)
```

eps=0.5



eps=1



Issues:

- Number of clusters strongly depends on eps
- Method assumes similar density of points in clusters

unclustered outliers

Take Home Messages

- ◆ **Clustering** your data decide whether you would like to separate in a fixed number of groups and be **more robust (k-means)** or to a variable number of clusters and be **more flexible (hierarchical)**
- ◆ **Heatmap** allows you to visualize profiles of expression **among samples and among genes in one graph**
- ◆ For single-cell data, use **DBSCAN**
- ◆ There is no “magic bullet” to select number of clusters – try several approaches!

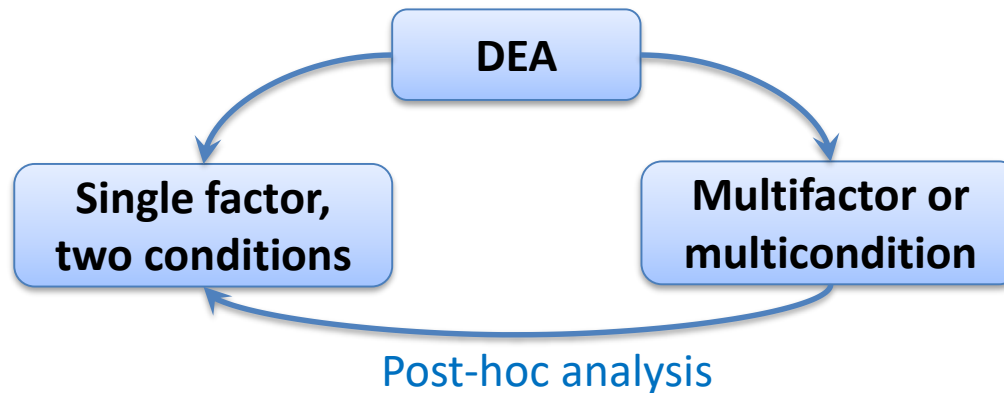
Differential Expression Analysis

Basics

Questions

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

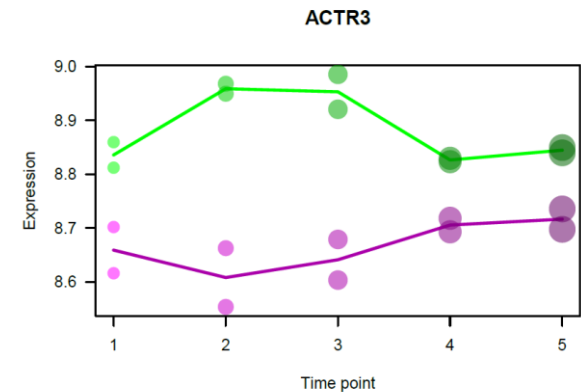
Similar to t-test with Student's statistics:
compare means



Similar to ANOVA with Fisher's statistics:
compare variances

And do not forget about multiple hypotheses testing

Example: 2 cell lines in time:



What is this 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

$$H_0: \mu \leq \mu_0$$

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

$$H_a: \mu > \mu_0$$

$$H_a: \mu < \mu_0$$

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.

$$\mu_0 = 3 \text{ lbm}$$

Suppose sample of $n=36$ coffee cans is selected. From the previous studies it's known that $\sigma = 0.18 \text{ lbm}$

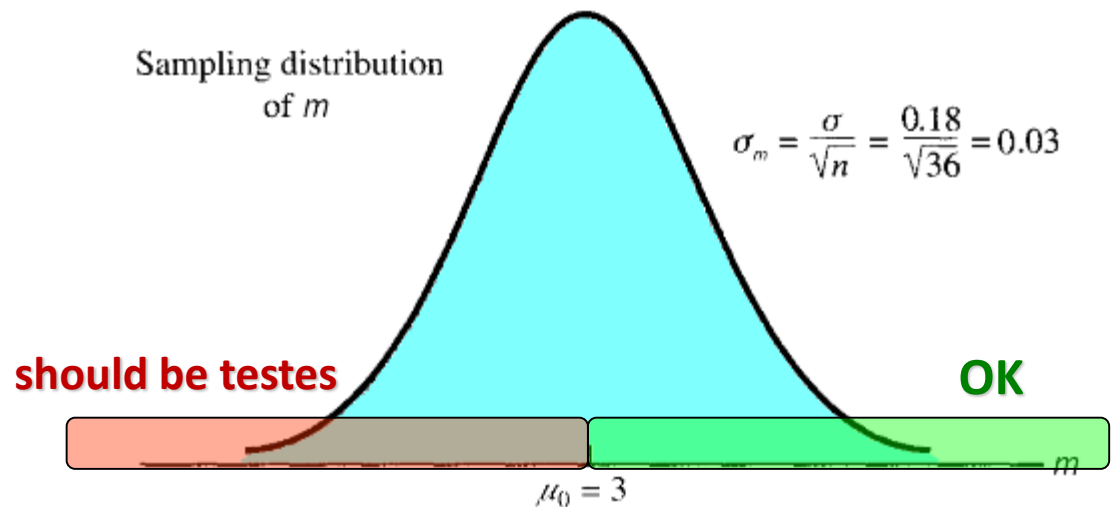
What is this p-value ?

$$\mu_0 = 3 \text{ lbm}$$

$$H_0: \mu \geq 3 \quad \text{no action}$$

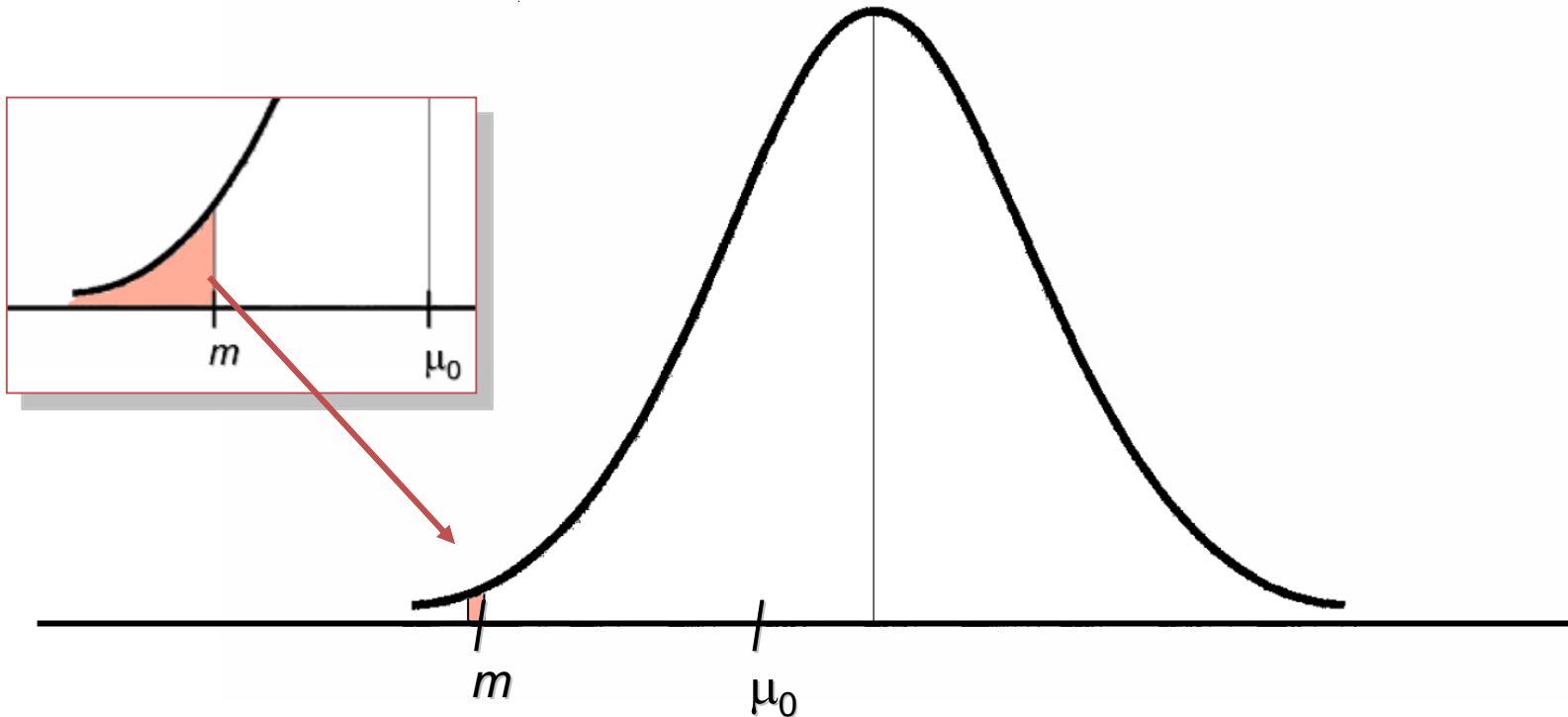
$$H_a: \mu < 3 \quad \text{legal action}$$

Let's say: **in the extreme case**, when $\mu=3$, 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 **$\alpha = 0.01$**

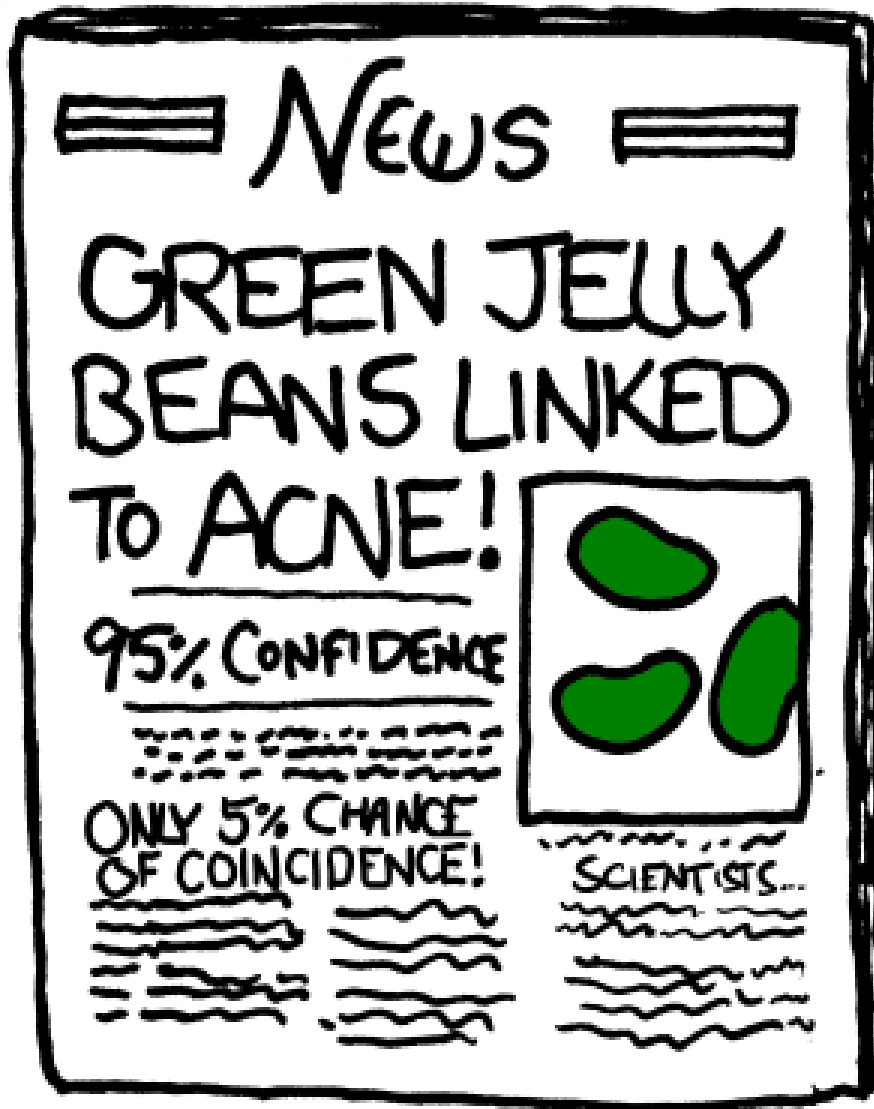


What is this p-value ?

Let's find the probability of observation m for all possible $\mu \geq 3$. We start from an extreme case ($\mu=3$) 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 $\mu \geq 3$?



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



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

44

Multiple Hypotheses

		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:

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

Differential Expression Analysis

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		Total
		H ₀ is TRUE	H ₀ is FALSE	
Conclusion	Accept H ₀ (non-significant)	<i>U</i>	<i>T</i>	$m - R$
	Reject H ₀ (significant)	<i>V</i>	<i>S</i>	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 = $\alpha = 0.05$**

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

Expected value for $FDR < \alpha$ if

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



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

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

Theoretically, the sign should be " \leq ".
But for practical reasons it is replaced by " $<$ "

Familywise Error Rate (FWER)

Bonferroni – simple, but too stringent, not recommended

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

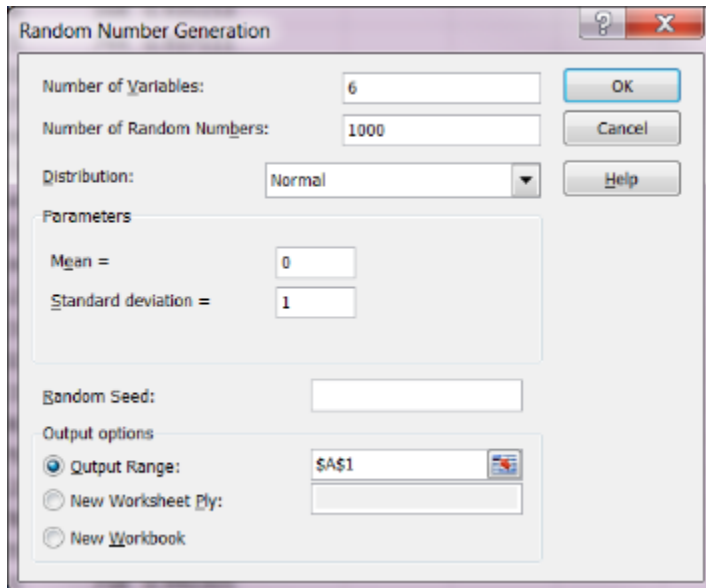
Holm-Bonferroni – a more powerful, less stringent but still universal FWER

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

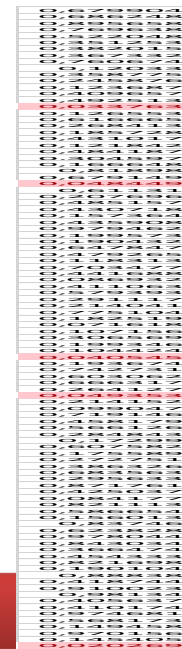
Why is it so important to correct p-values?..

Let's generate a completely random experiment (Excel)

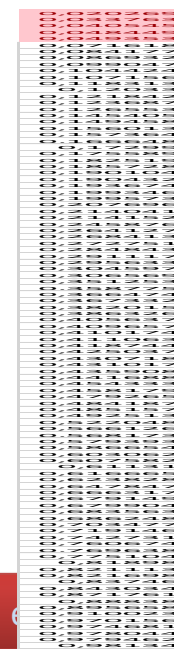
- ◆ Generate 6 columns of normal random variables (1000 points/candidates in each).
- ◆ Consider the first 3 columns as “treatment”, and the next 3 columns as “control”.
- ◆ Using t-test calculate p-values b/w “treatment” and “control” group. How many candidates have $p\text{-value} < 0.05$?
- ◆ Calculate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



Candidates.
5% are false



Same candidates.
Just sorted



Top 5%
selected
???

Linear Models

Many conditions

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

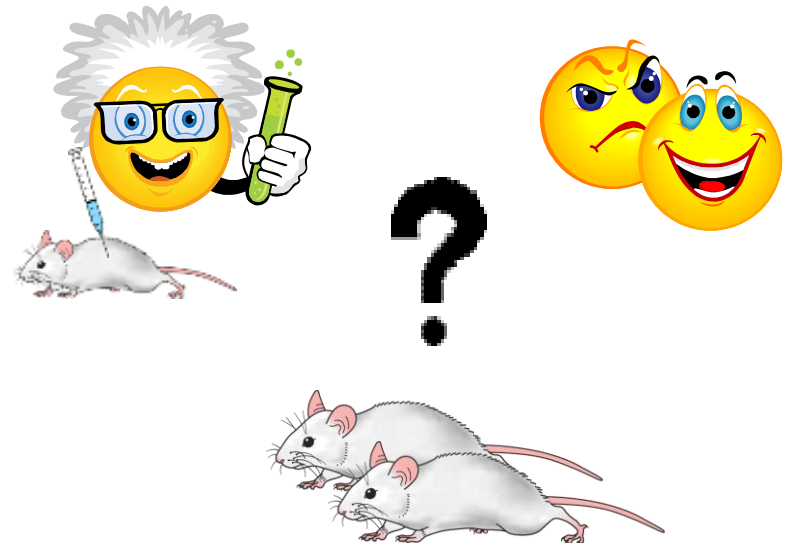
Many factors

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™

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?

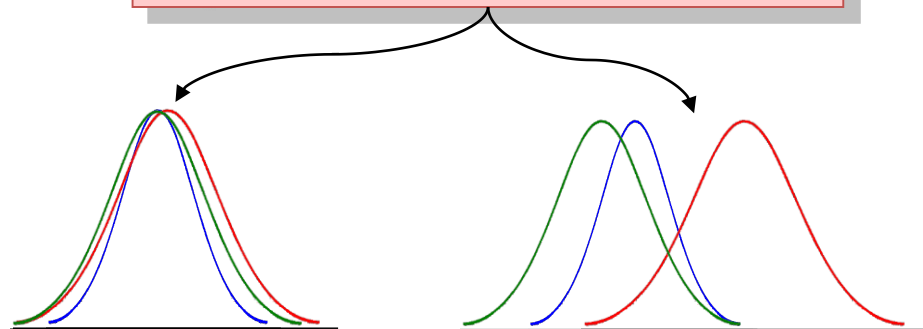
`depression.txt`

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
...

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

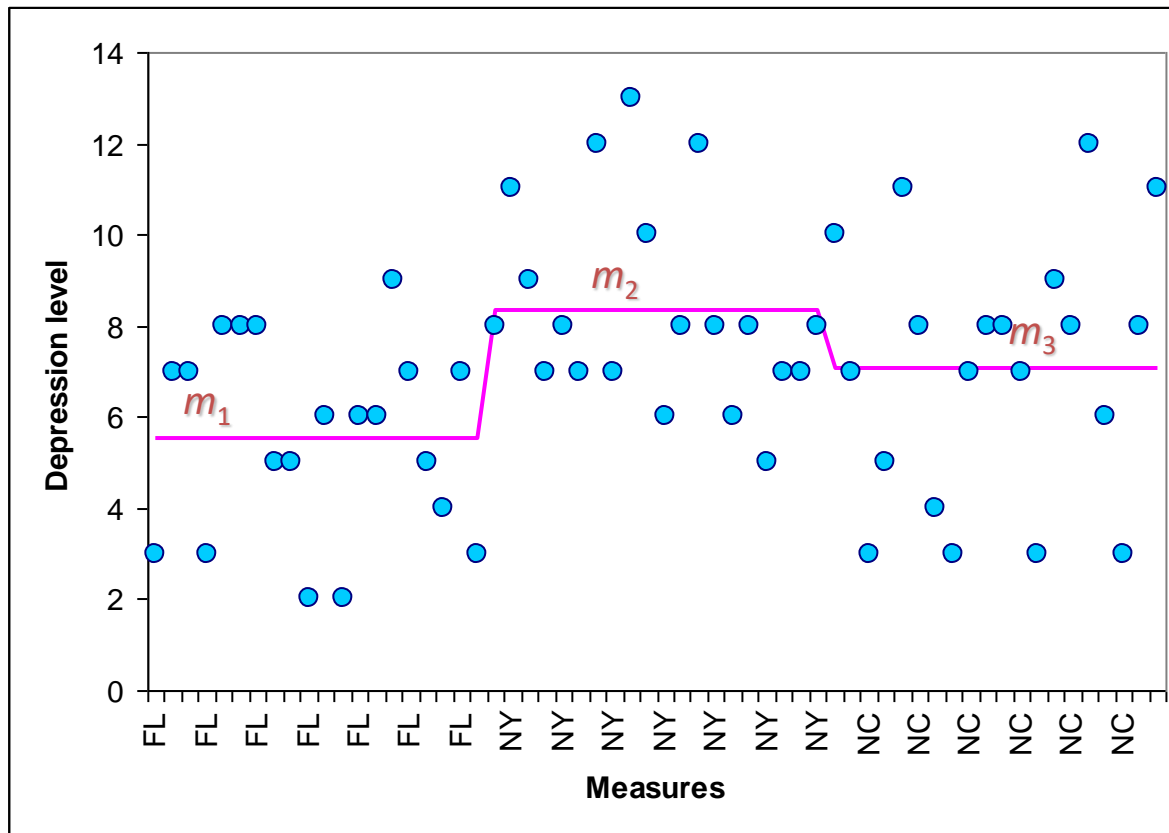
$$H_a: \text{not all 3 means are equal}$$



Linear Models

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

H_a : not all 3 means are equal



LIMMA & EdgeR : Linear Models for Microarrays

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

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 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 RNA-Seq, 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

Take Home Messages

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

- ◆ DEA provides the genes which have variability in **mean** gene expression between condition
 - ◆ => more data you have, smaller differences you will be able to see

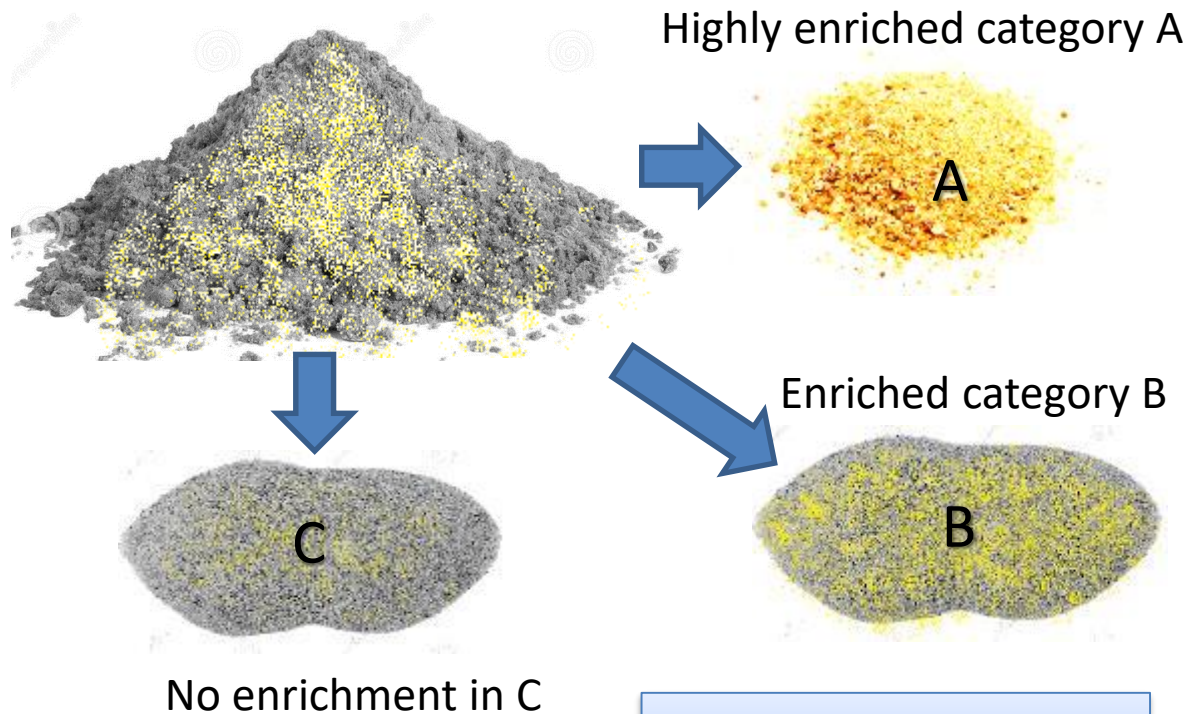
- ◆ Several factors can be taken into account in ANOVA approach. This will give you insight into significance of each experimental factor but at the same time will correct batch effects and allow answering complex questions (remember shoes affecting ladies...).

Enrichment Analysis

Enrichment Analysis

1. Category Enrichment Analysis

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



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

How surprised will you be if
he grabbed

●●●●●●●●●●●●●●●●●●●●●●
(17 red , 3 green)

Method of the analysis:
Fisher’s exact test

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

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

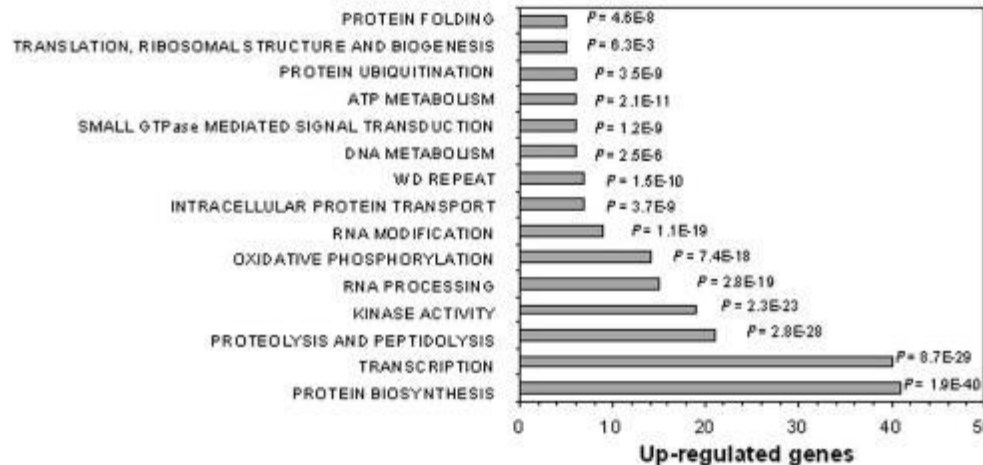
N: total number of genes

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

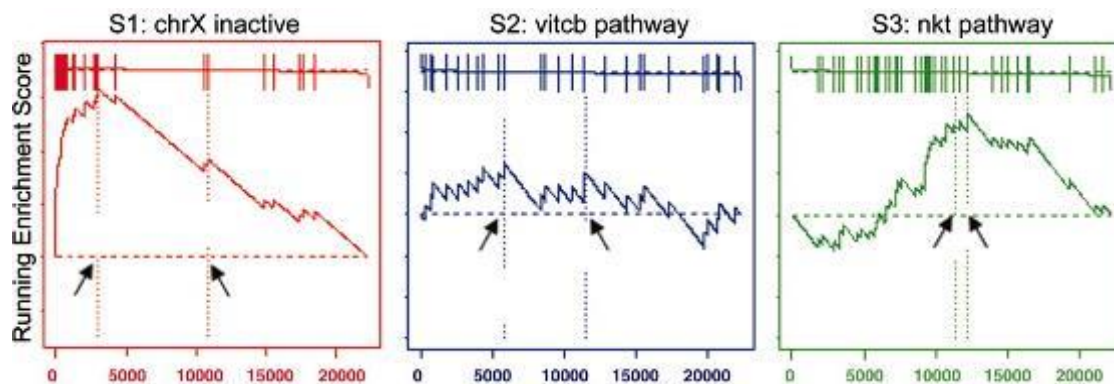
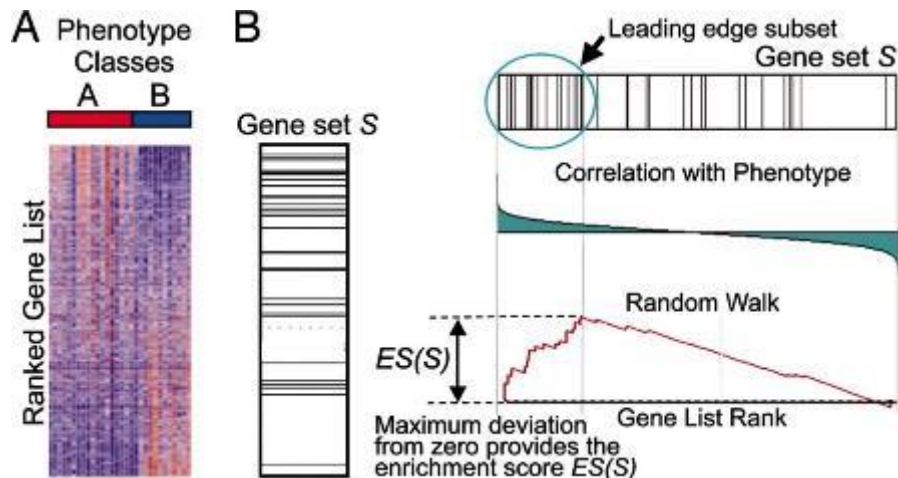
$$C_k^n = C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$



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

2. Gene Set Enrichment Analysis (GSEA)

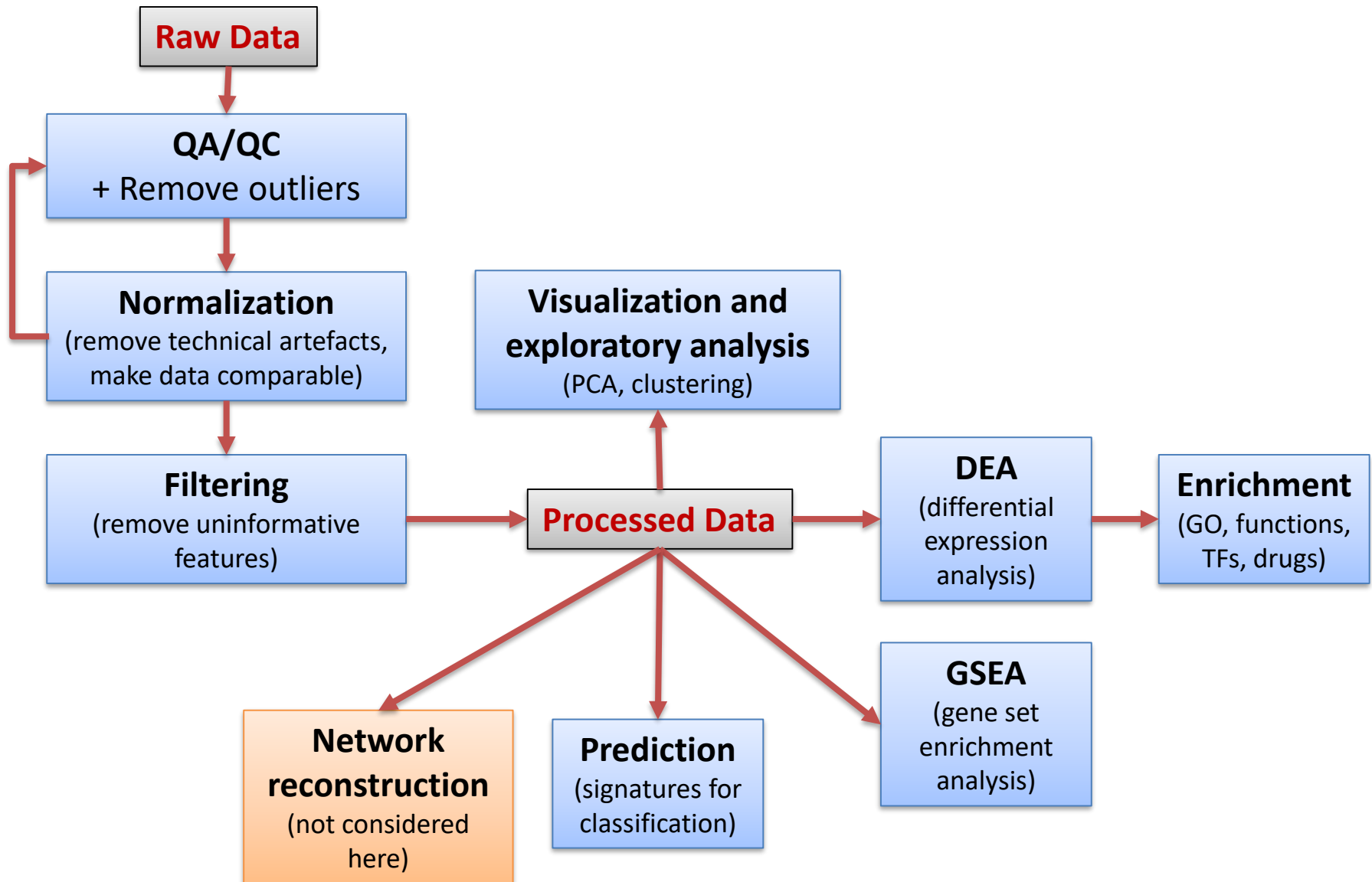
Is direction of genes in a category random?



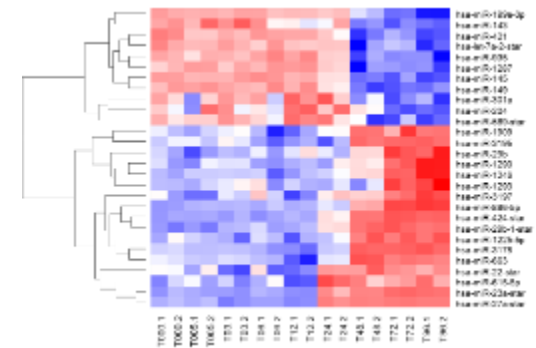
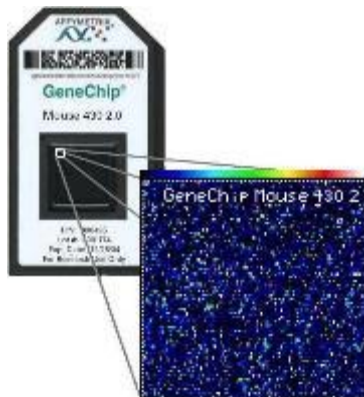
A. Subramanian et al. PNAS 2005,102,43

Take Home Messages

- ◆ To find biological meaning of the significantly regulated genes use enrichment analysis methods linking known groups of genes to DEA results
- ◆ Enriched categories are usually more robust than individual genes



Thank you for your attention !

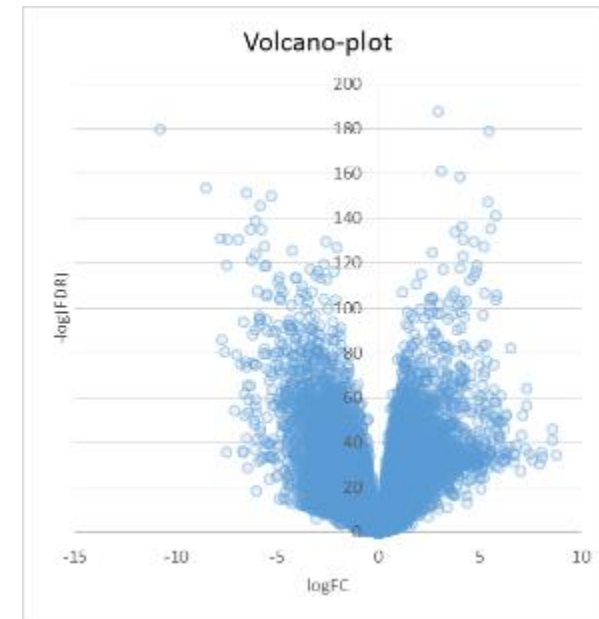


Practice

Example – LUSC data from TCGA

<http://edu.sablab.net/transcript/lusc20.xlsx>

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



Task 1. 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 \uparrow and \downarrow functions

Enrichr

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

BioCompendium

<http://biocompendium.embl.de/>

Strategy 2:

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

- Can be biased (gene can be $\uparrow\downarrow$)
- Assume \uparrow gene \Rightarrow \uparrow function
- Can distinguish \uparrow and \downarrow functions

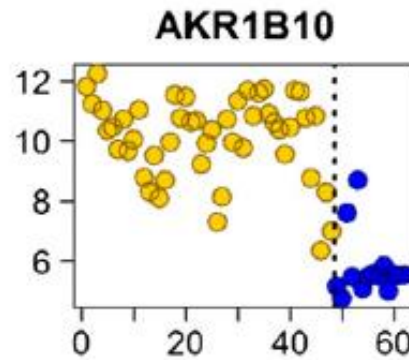
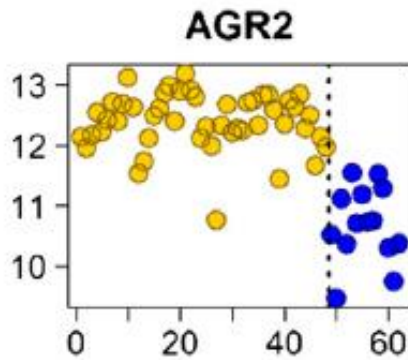
Classification

Gene Markers

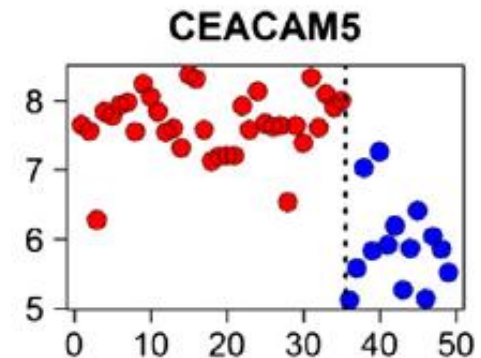
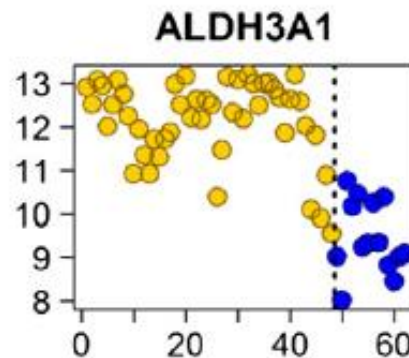
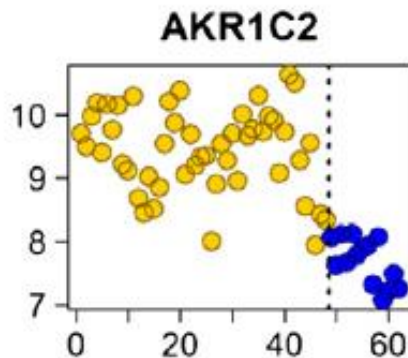
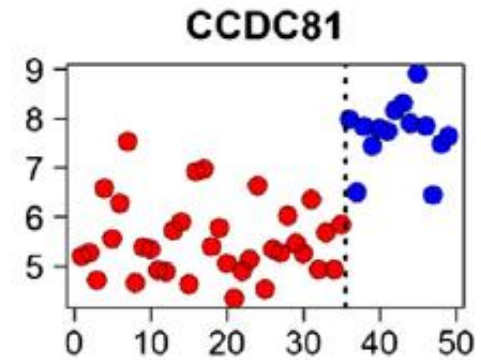
Questions

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

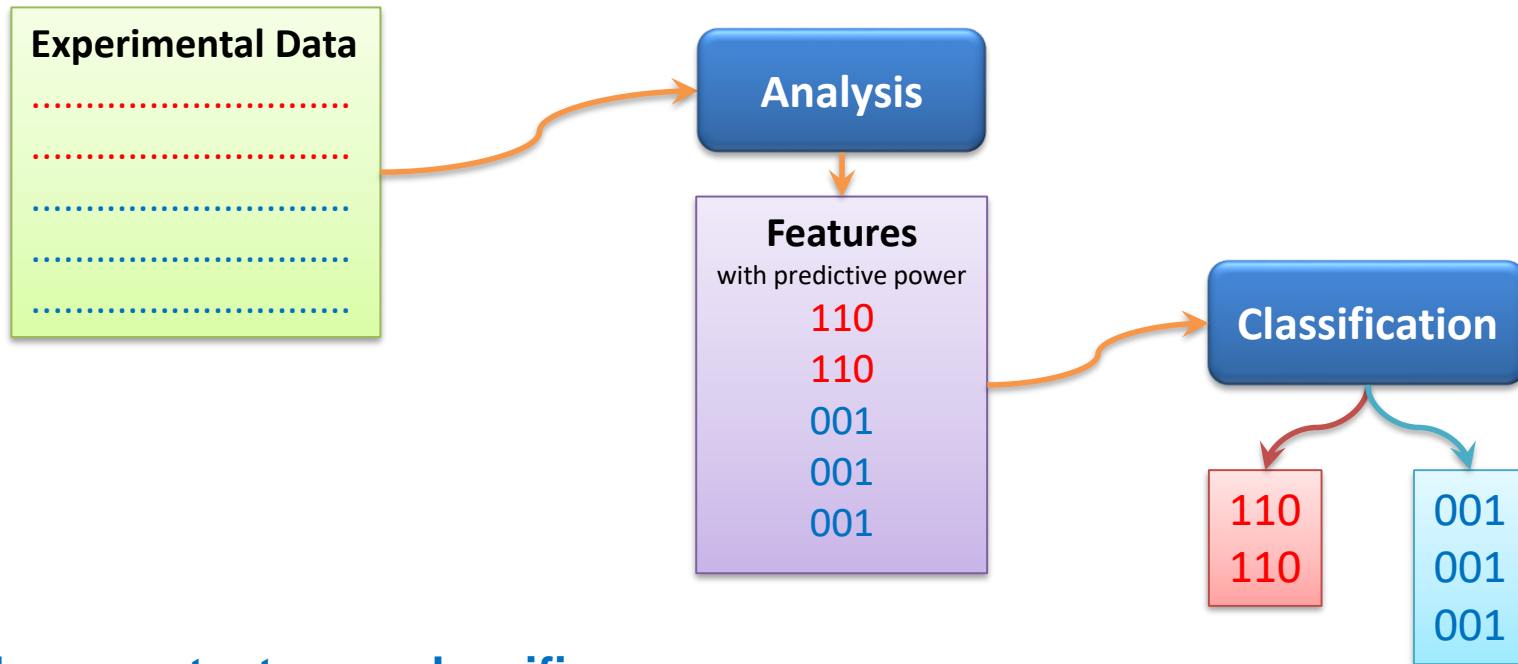
A SNC vs NS



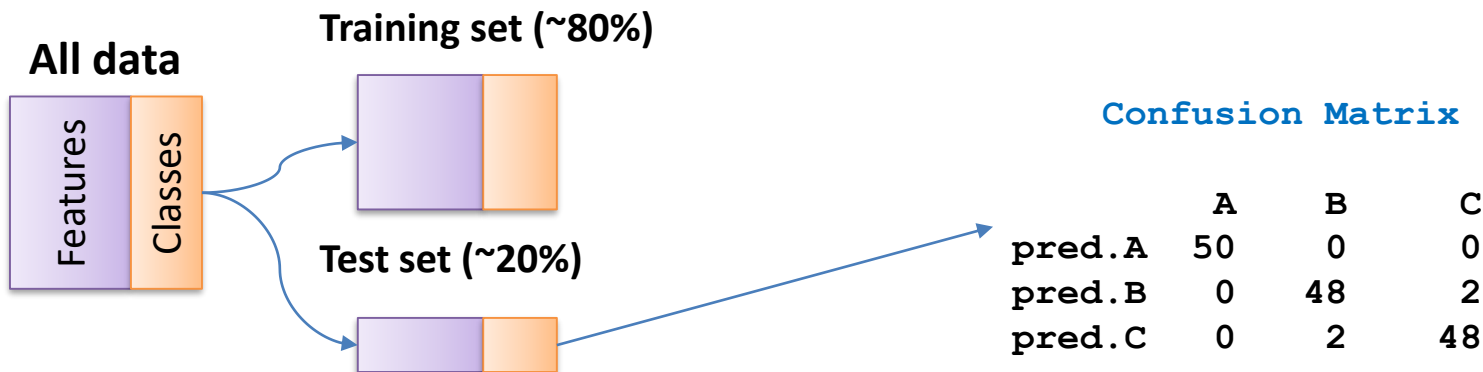
B SC vs NS



General Scheme



When you test your classifier:



Selection of Features: ROC and AUC

ROC curve

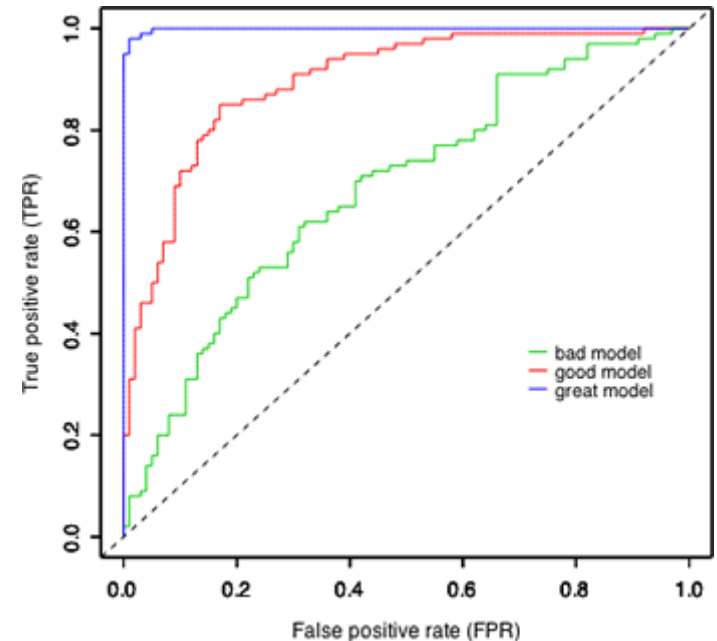
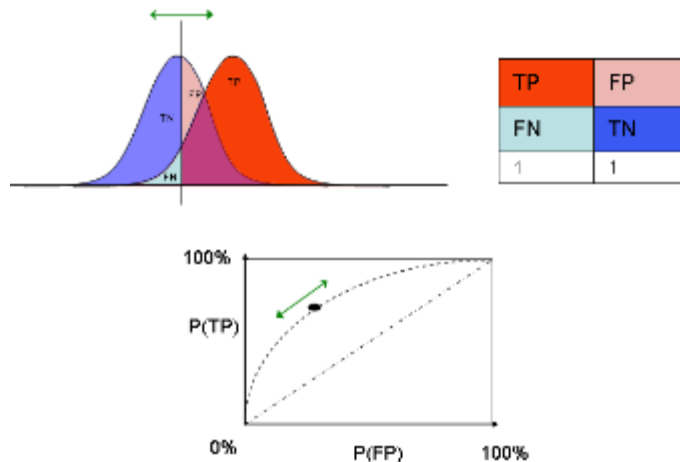
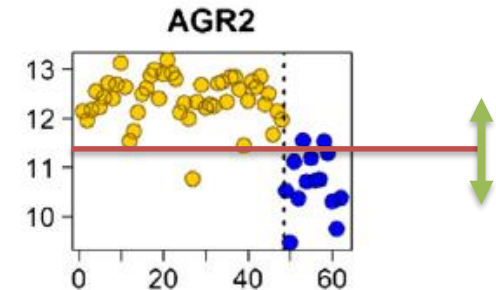
(receiver operating characteristic) is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate (1-specificity or false positive rate)

AUC

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

ROC is introduced for 2 classes.

If we have more than 2 classes – create several ROC curves (1 per class)



<http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture22.htm>

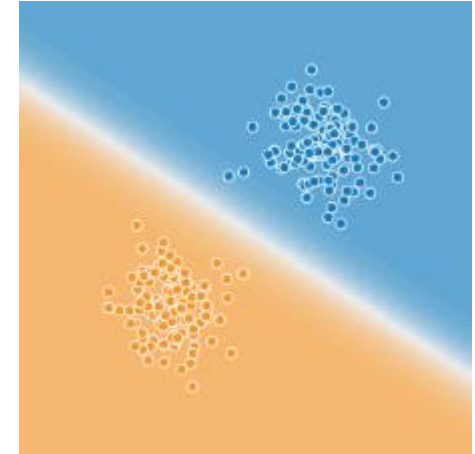
https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Simple Classifier: Logistic Regression

Logistic regression

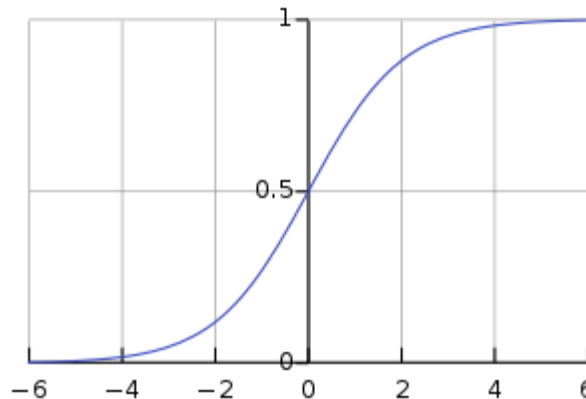
Linearly combines the features and calculates

- 1) will divide you data to 2 groups, and
- 2) has the optimal distance from the closest elements of the groups



Logistic regression: sigmoid function upon linear regression:

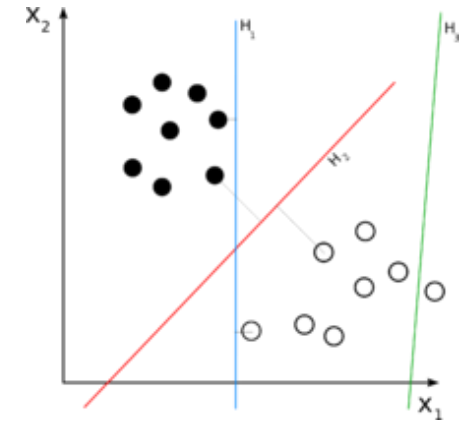
$$F(z) = \frac{1}{1 + e^{-(b_1x_1 + b_2x_2 + \dots + b_0)}}$$



More Advanced Classification Methods

Support vector machine (SVM)

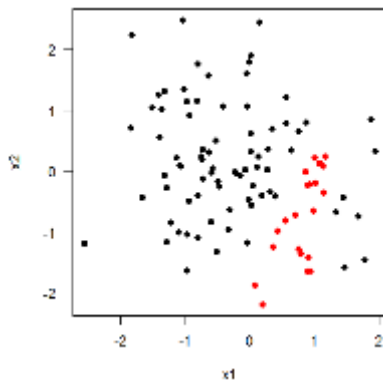
System tries to find a line (hyper plane) which
 1) will divide you data to 2 groups, and
 2) has the optimal distance from the closest elements of the groups



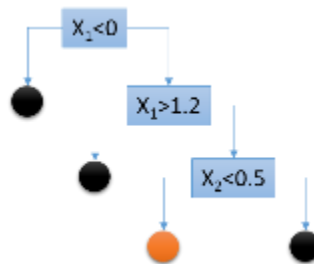
Random Forest (RF)

Makes a set of decision trees (if value x is less than x_0 then we choose class A), which is called "forest". Then vote among the trees.

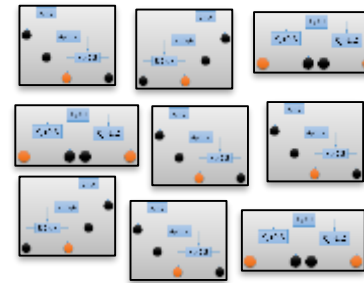
space of features



tree



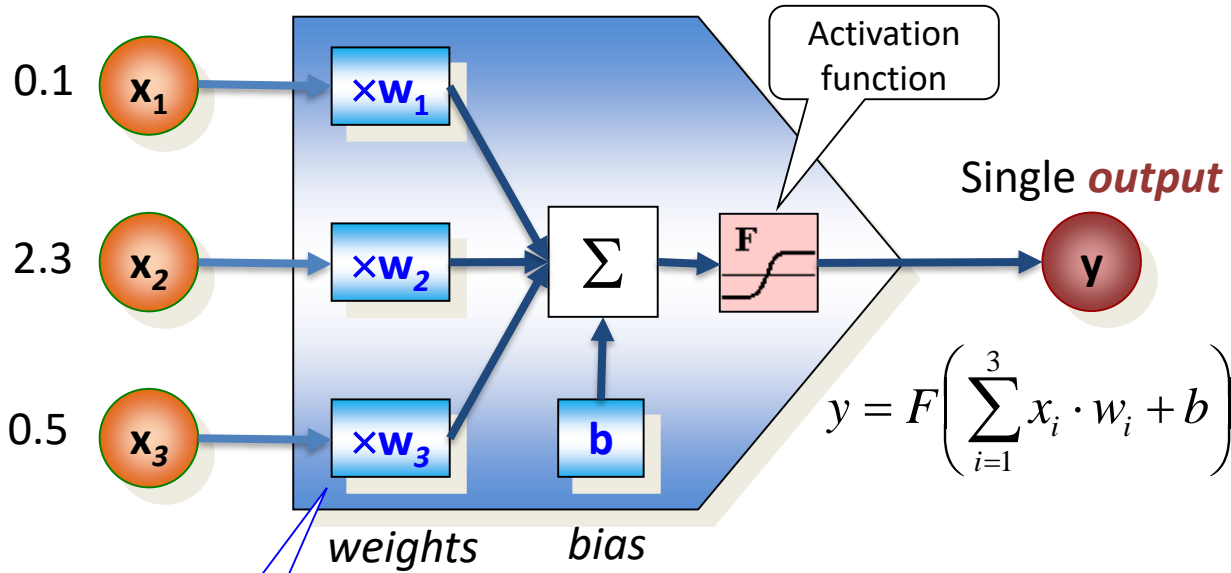
forest



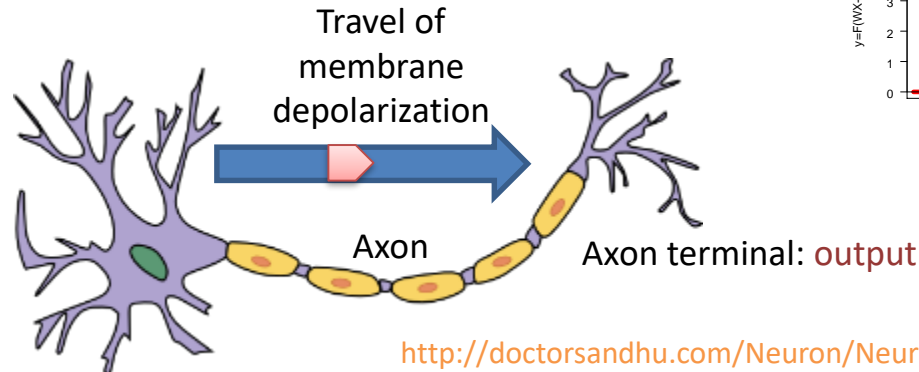
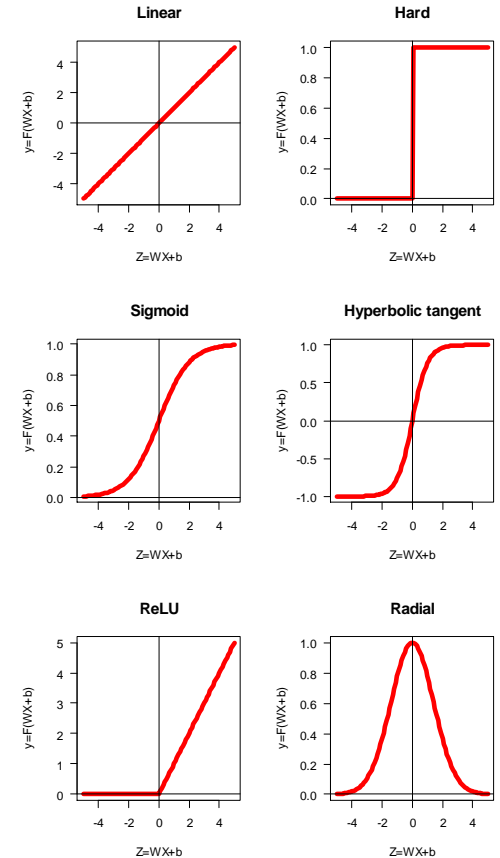
Classification and Marker Genes

Artificial Neuron – a Simple Processing Unit (~ logistic regression)

Multiple *inputs*



Adjusted coefficients

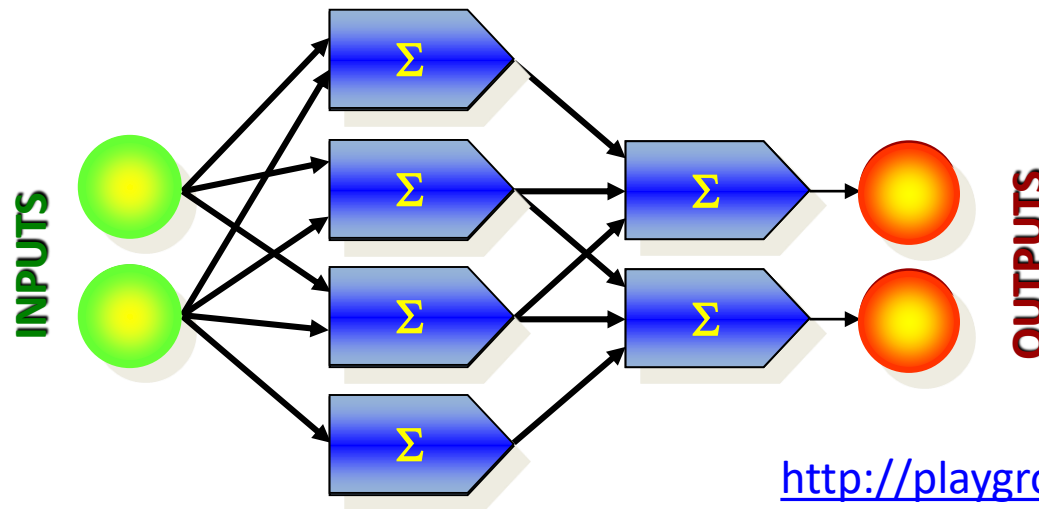


<http://doctorsandhu.com/Neuron/Neuron.shtml>

Classification and Marker Genes

Feed Forward Network (FFN), a.k.a. Multi-layer Perceptron (MLP)

Forward propagation of information



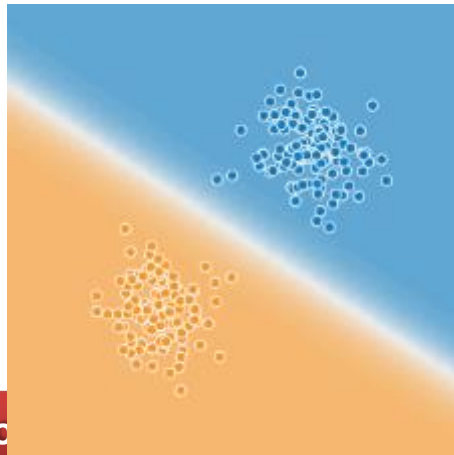
Normalized data: raw, features, variables etc.

In classification the output is considered as probability of a class (with *softmax*)

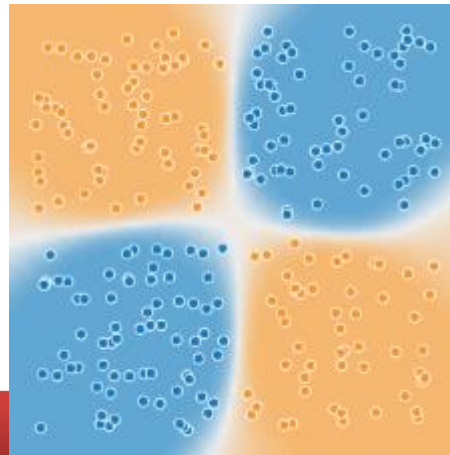
$$p(y_i|X) = \frac{y_i}{\sum y_j}$$

<http://playground.tensorflow.org/>

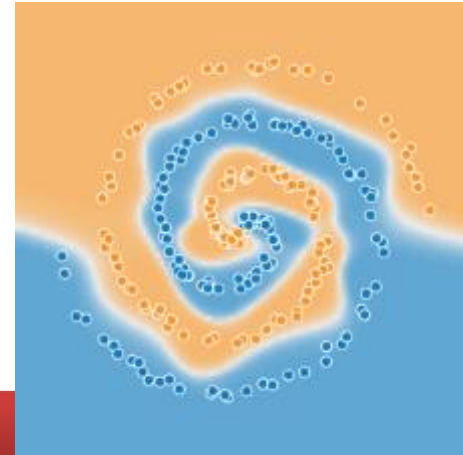
1 layer



2 layers



4 layers



Take Home Messages

- ◆ Diagnostics & prediction include 3 main steps:
 - ◆ 1. Data analysis – transforms data into set of features
 - ◆ 2. Select features with predictive properties
 - ◆ 3. Use a classification algorithm

- ◆ AUC is one of the measures to select genes with strong predictive properties. Ideal AUC = 1, minimal AUC (worst situation) = 0.5

- ◆ Classifiers: logistic regression, SVM, RF, neural networks

- ◆ When doing classification for a real application - always divide your data in two groups: training and testing subsets to avoid overtraining