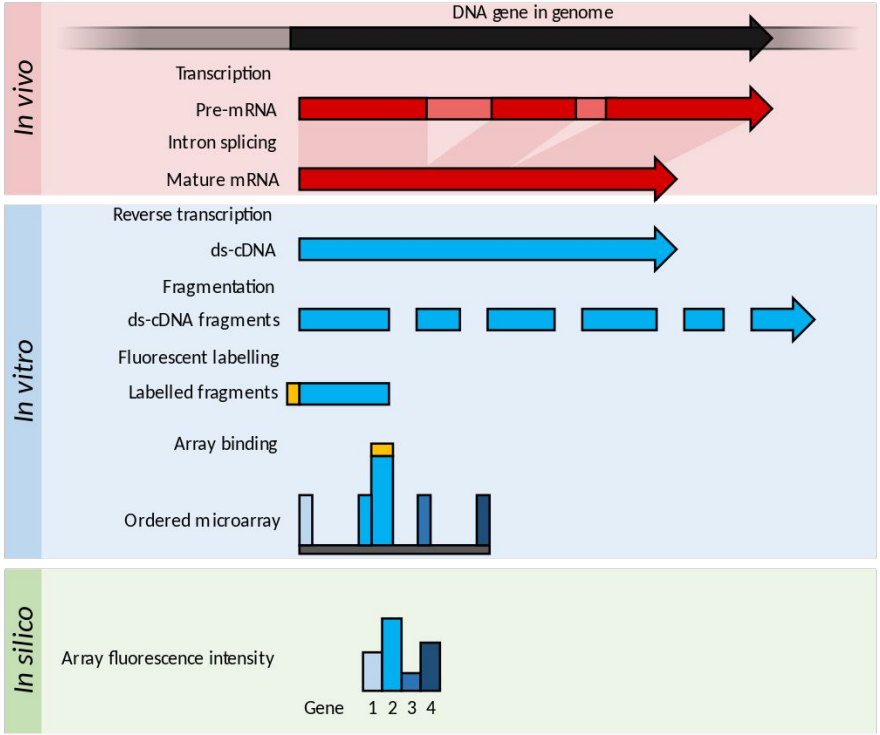
Functional genomics

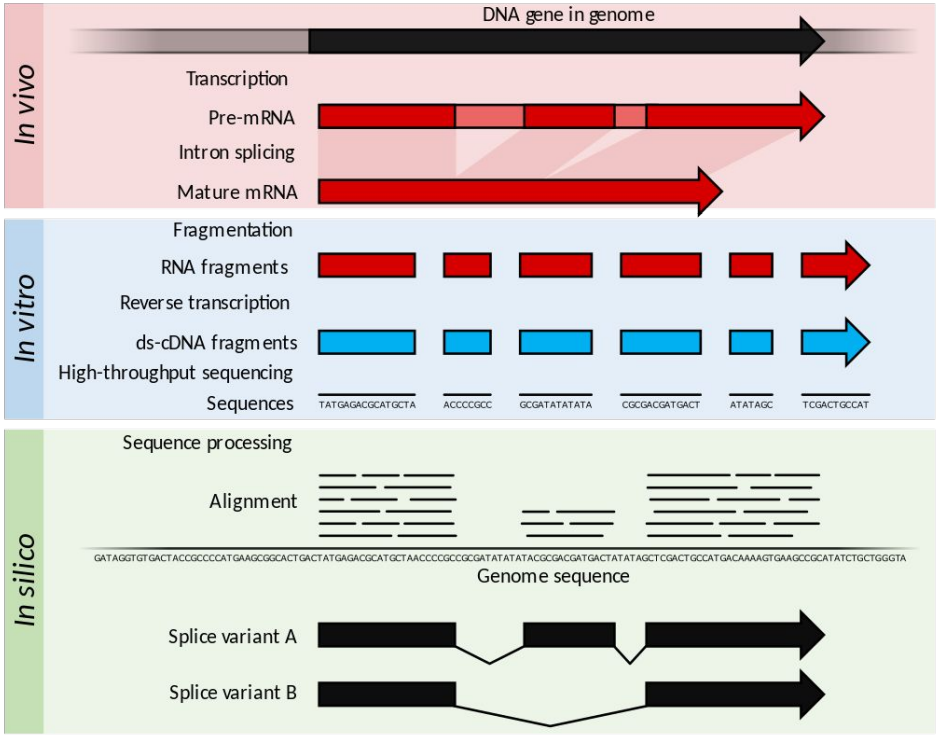
- Measuring gene-expression
- Distance/Similarity measures
- Clustering genes/samples
- Differential expression
- Functional enrichment

Measuring gene-expression on a large-scale

DNA microarrays



RNA-seq

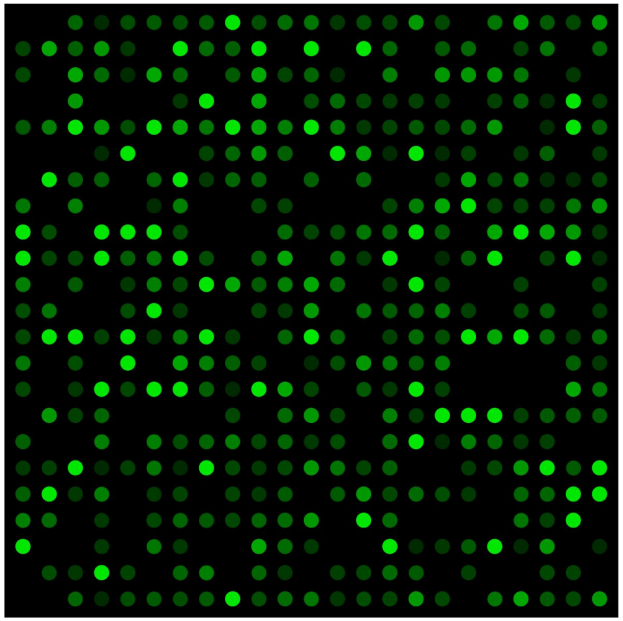


Lowe 2017 PLoS Comp. Biol.

Measuring gene-expression on a large-scale

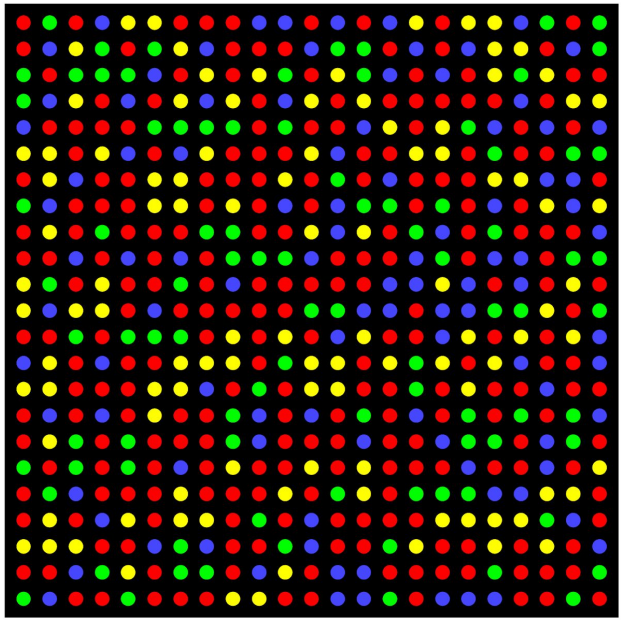
DNA microarrays

Microarray chip



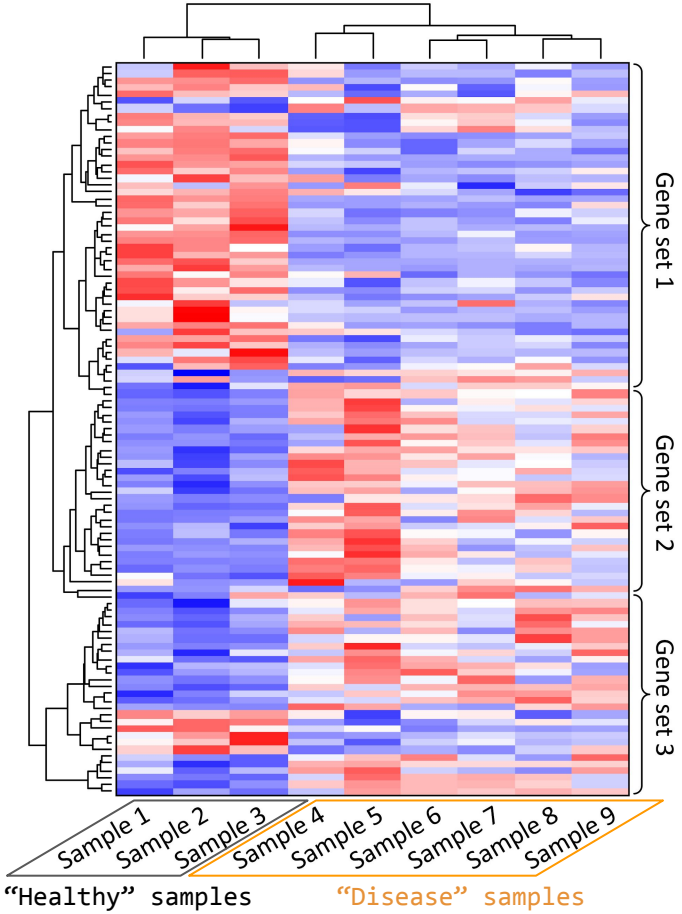
RNA-seq

Sequencing flow cell



Lowe 2017 PLoS Comp. Biol.

A gene expression dataset



Gene-level Qs:

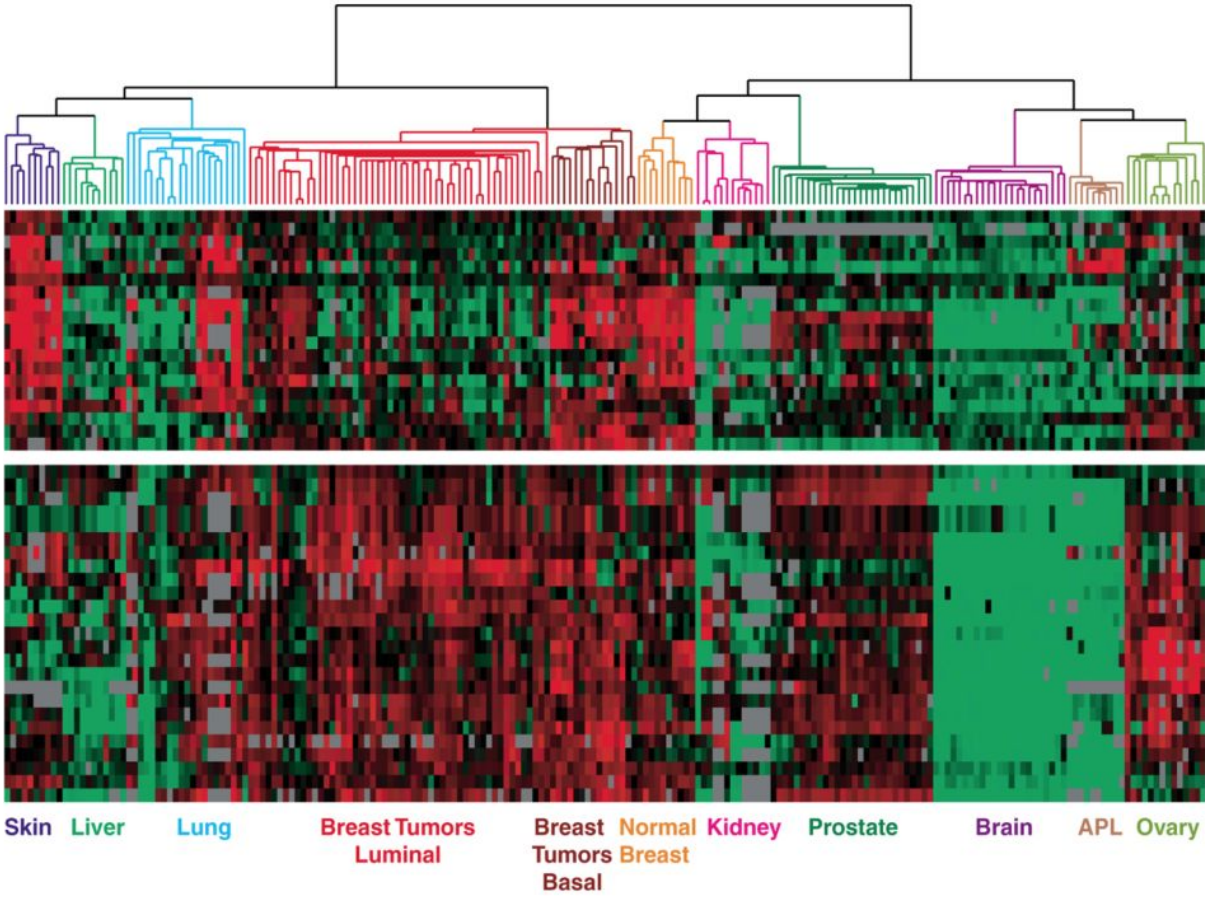
1. What's expressed (& by how much) in a given context/condition?
2. What's differentially expressed between two (or more) contexts/conditions?

Group-level Qs:

1. Are there groups of genes that respond similarly to changing contexts (across samples)?
2. Are there groups of samples that have very similar gene expression profiles?

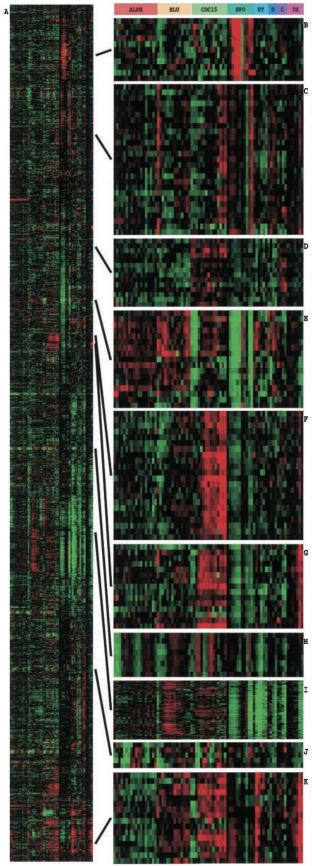
Lowe 2017 PLoS Comp. Biol.

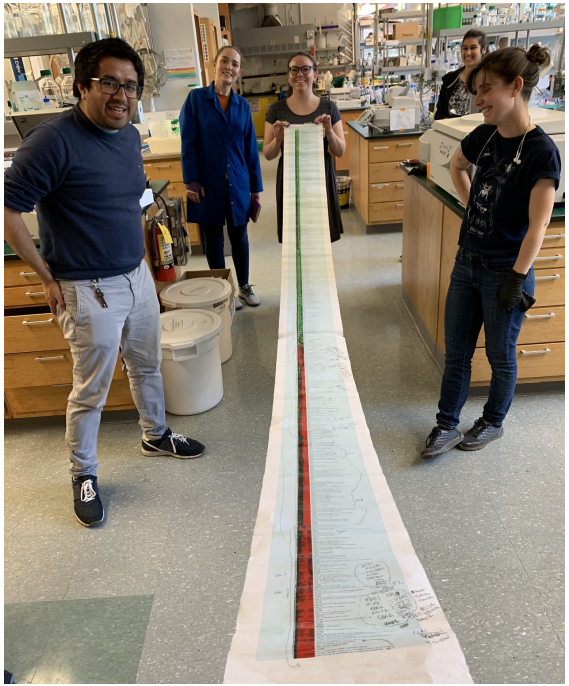
Example dataset and group-level analysis



Wikipedia

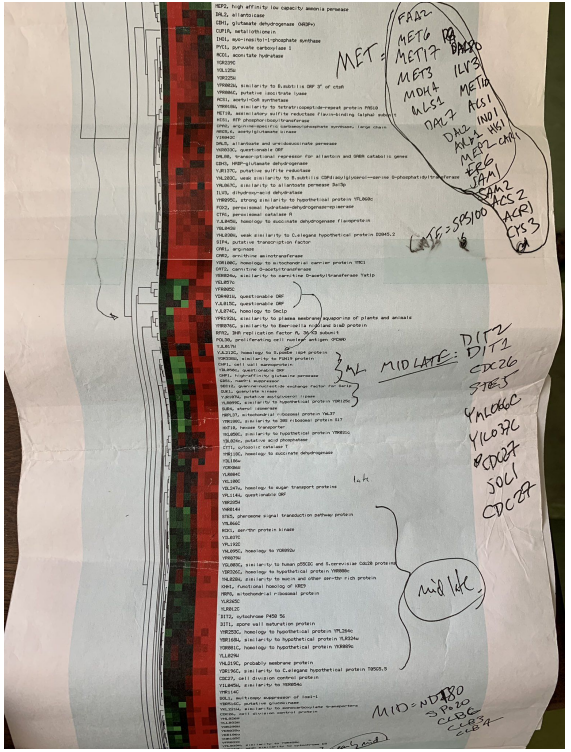
One of the first large-scale gene expression analyses (yeast)





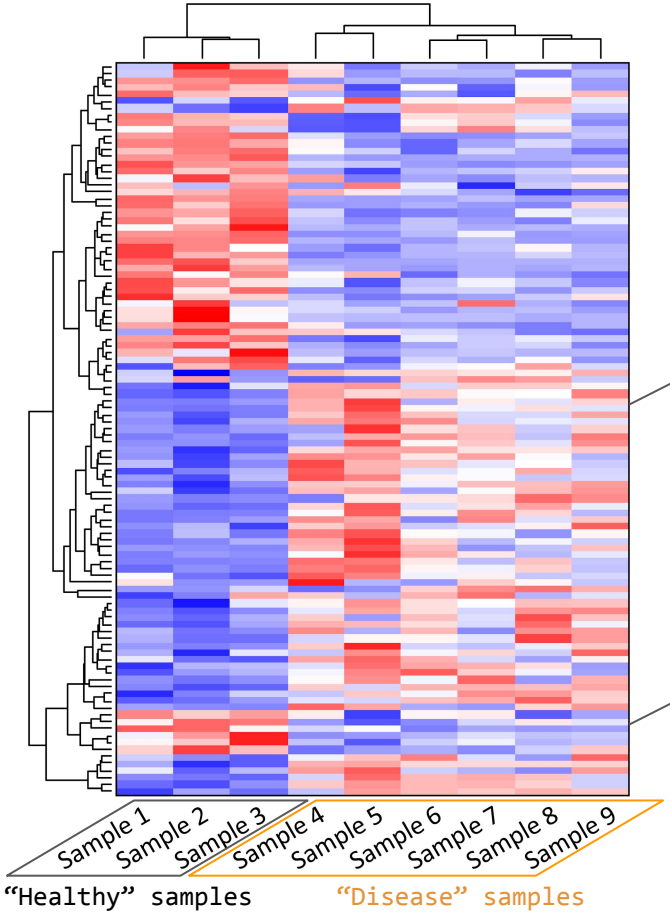
Michael Eisen 虫めづるマイケル
@mbeisen

Inspired by @UCSDCooperLab's question about origins of the red/green color scheme in microarray clustering, I present THE FIRST dna microarray cluster analysis made by me in 1997 for ncbi.nlm.nih.gov/m/pubmed/97841... w/handwritten notes from Pat Brown and the late Ira Herskowitz.



Eisen 1998 PNAS

A gene expression dataset



Gene_X	10.1	12.8	13.4	9.05	10.2	5.47	6.09	4.56	11.3

Gene_Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.8

Lowe 2017 PLoS Comp. Biol.

Many distance measures

Euclidean Distance

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Pearson Correlation Coefficient

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

Spearman Rank Correlation

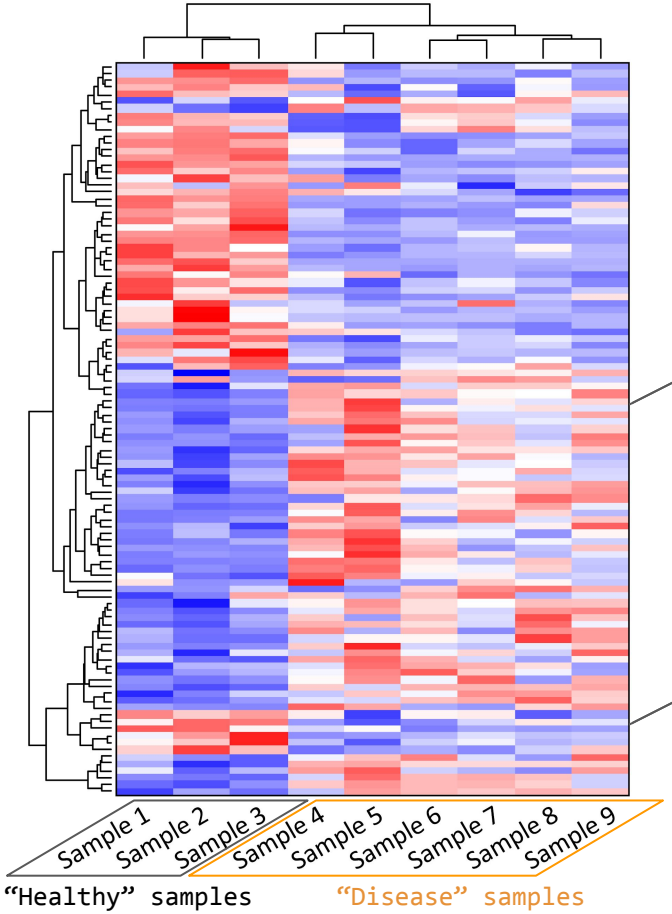
$$\rho = 1 - \frac{6 \sum_{i=1}^{n} [rank(x_i) - rank(y_i)]}{n(n^2 - 1)}$$

Mutual Information

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{(X,Y)}(x,y) \log \left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)} \right)$$

Wikipedia

Gene-level analysis



	“Healthy” samples			“Disease” samples					
Gene_X	10.1	12.8	13.4	9.05	10.2	5.47	6.09	4.56	11.3
Gene_Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.8

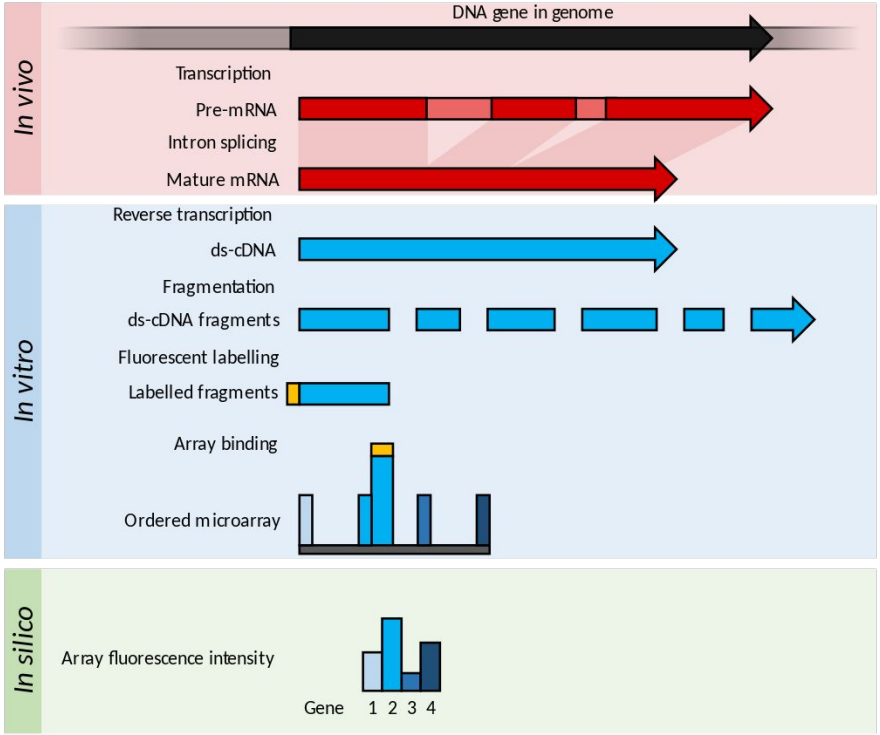
Lowe 2017 PLoS Comp. Biol.

Functional genomics

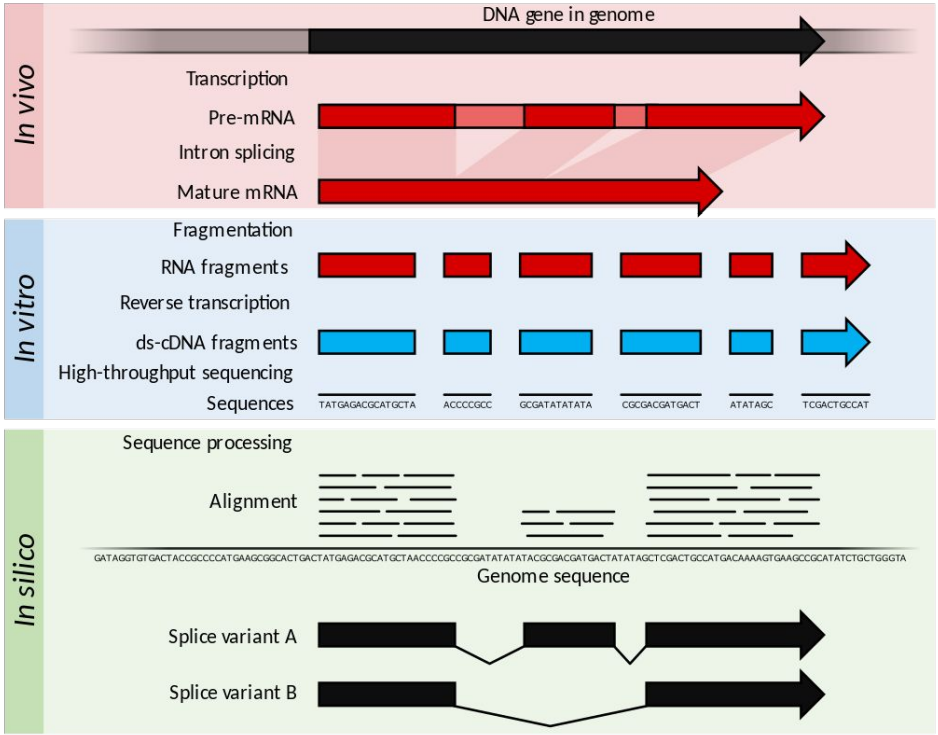
- Measuring gene-expression
- Distance/Similarity measures
- Clustering genes/samples
- Differential expression
- Functional enrichment

Measuring gene-expression on a large-scale

DNA microarrays



RNA-seq

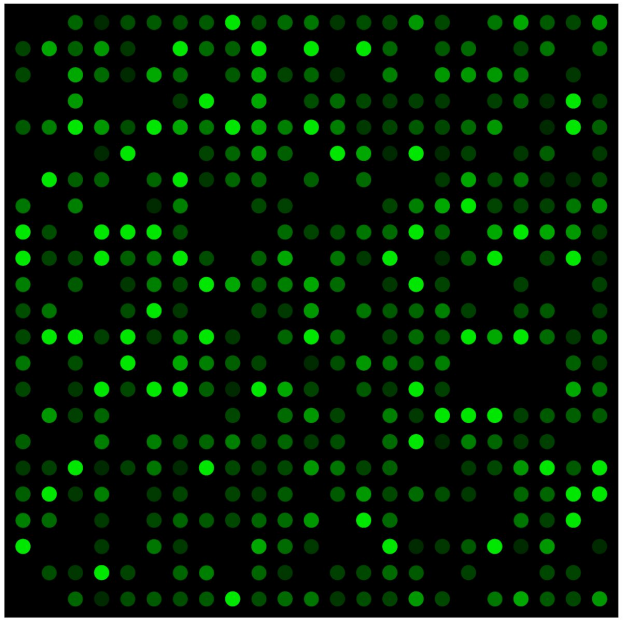


Lowe 2017 PLoS Comp. Biol.

Measuring gene-expression on a large-scale

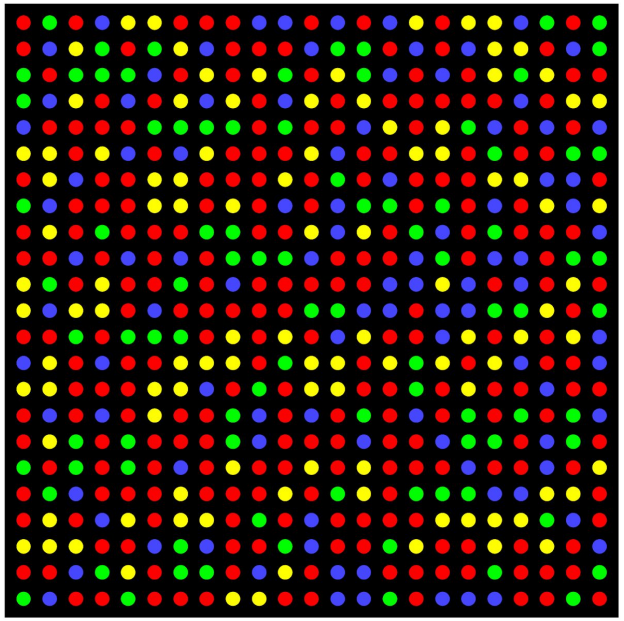
DNA microarrays

Microarray chip



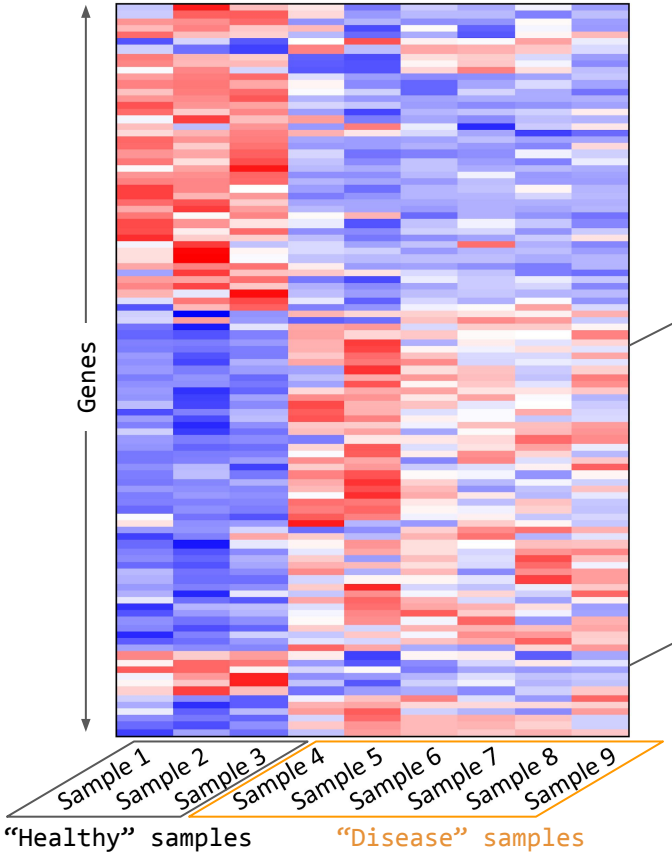
RNA-seq

Sequencing flow cell



Lowe 2017 PLoS Comp. Biol.

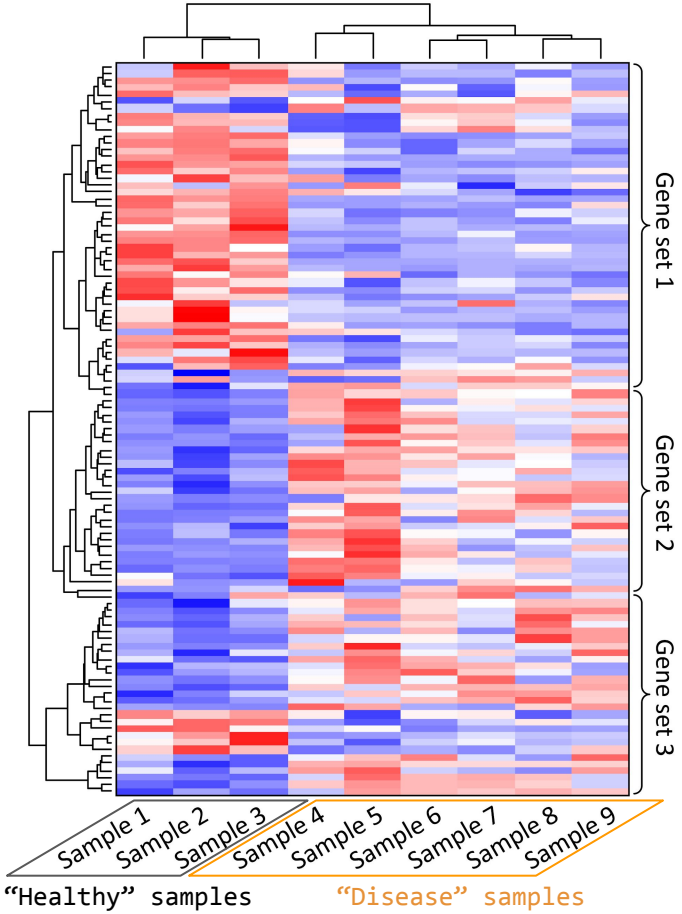
A gene expression dataset



	“Healthy” samples			“Disease” samples					
Gene_X	10.1	12.8	13.4	9.05	10.2	5.47	6.09	4.56	11.3
Gene_Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.8

Lowe 2017 PLoS Comp. Biol.

A gene expression dataset



Gene-level Qs:

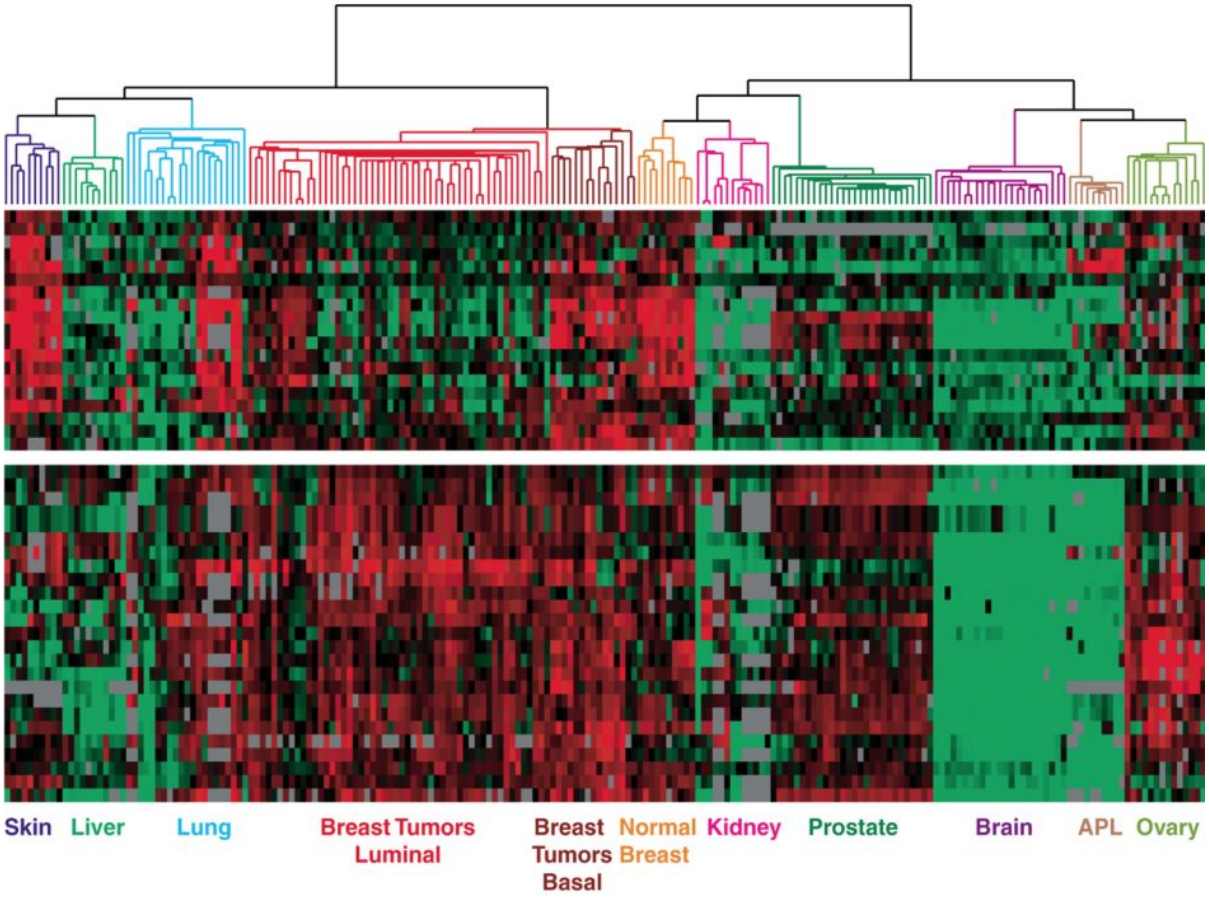
1. What's expressed (& by how much) in a given context/condition?
2. What's differentially expressed between two (or more) contexts/conditions?

Group-level Qs:

1. Are there groups of genes that respond similarly to changing contexts (across samples)?
2. Are there groups of samples that have very similar gene expression profiles?

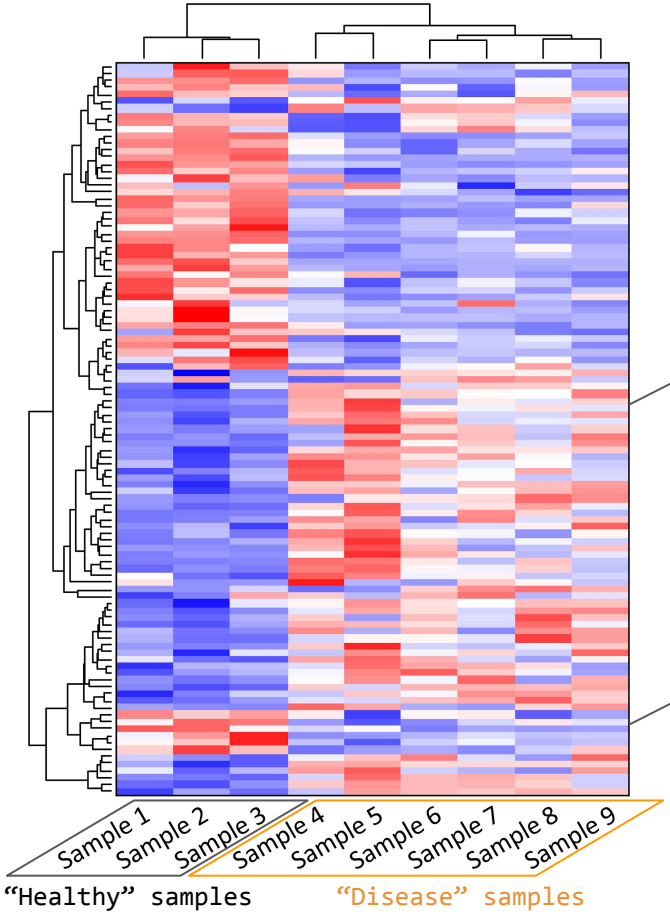
Lowe 2017 PLoS Comp. Biol.

One of the first large-scale gene expression analyses



Wikipedia

A gene expression dataset



Gene_X	10.1	12.8	13.4	9.05	10.2	5.47	6.09	4.56	11.3

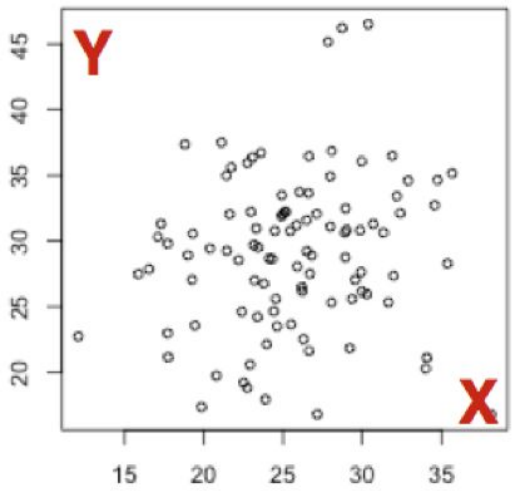
Gene_Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.8

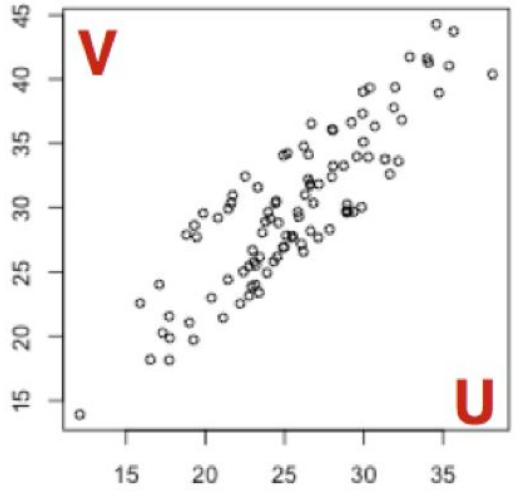
Lowe 2017 PLoS Comp. Biol.

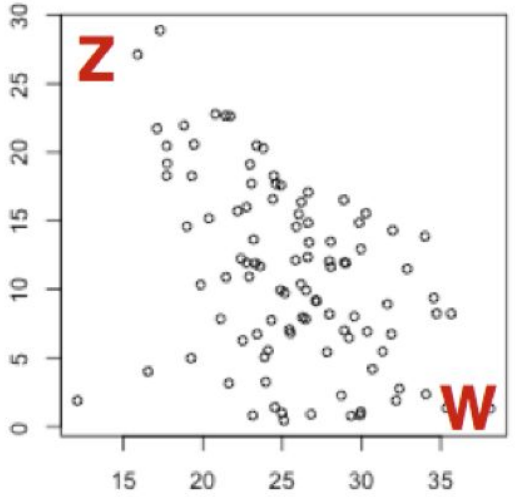
Calculating “distance”/“similarity” between genes or samples

Gene_X	10.1	12.8	13.4	9.05	10.2	5.47	6.09	4.56	11.3

Gene_Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.8







Correlation coefficient

Pearson Correlation Coefficient

- Measures linear relationship between variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- n is the sample size
- x_i, y_i are the single samples indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

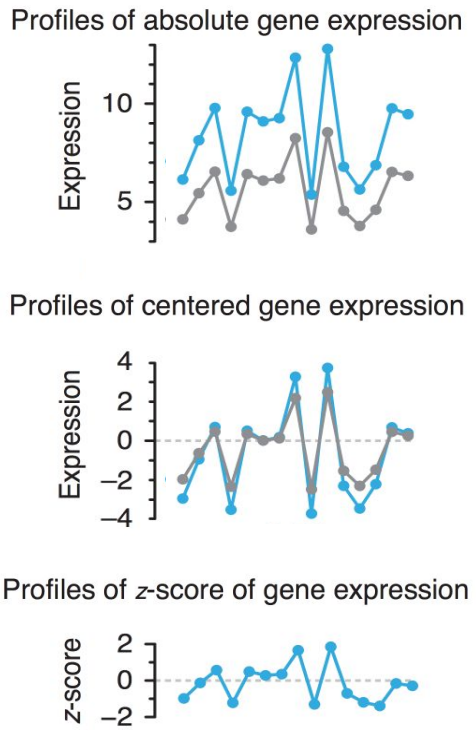
Wikipedia

Correlation coefficient

Pearson Correlation Coefficient

- Captures the relationship between 2 vectors after centering each vector by its mean and scaling by its standard deviation.
- The final quantities for each vector are called z-scores.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$



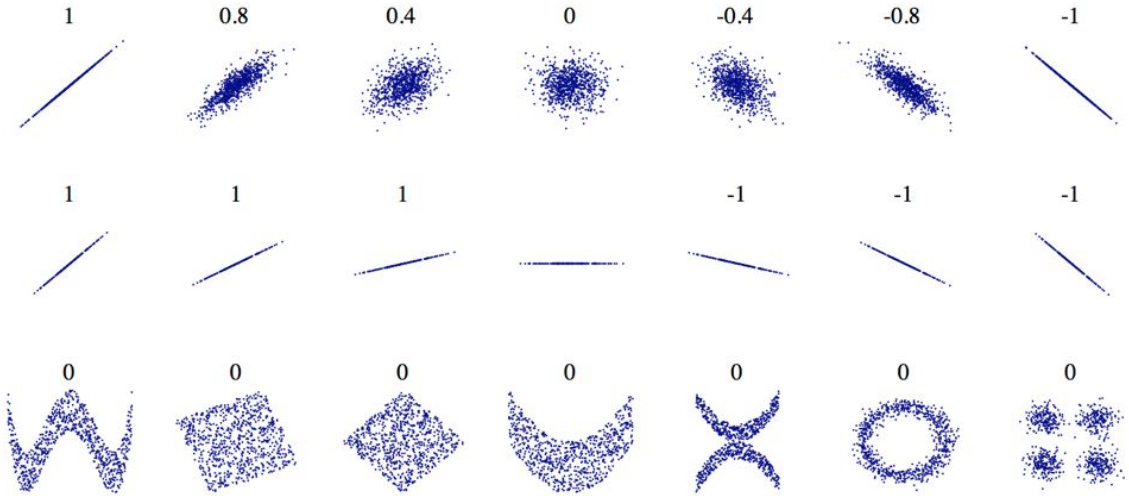
https://www.nature.com/articles/nmeth.4299

Correlation coefficient

Pearson Correlation Coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Measures linear relationship between variables.



$-1 \leq r \leq +1$

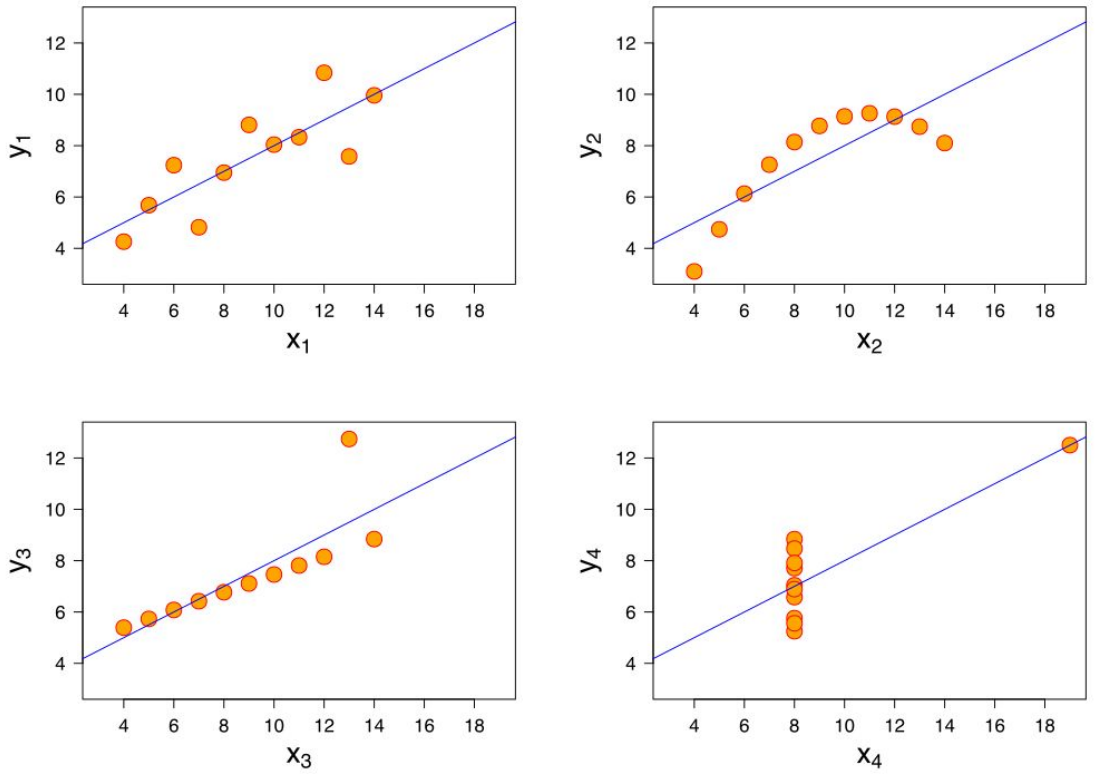
−1 is total −ve correlation | 0 is no correlation | +1 is total +ve correlation

Wikipedia

Anscombe's quartet: "calculation are exact; graphs are rough!"

11 data points

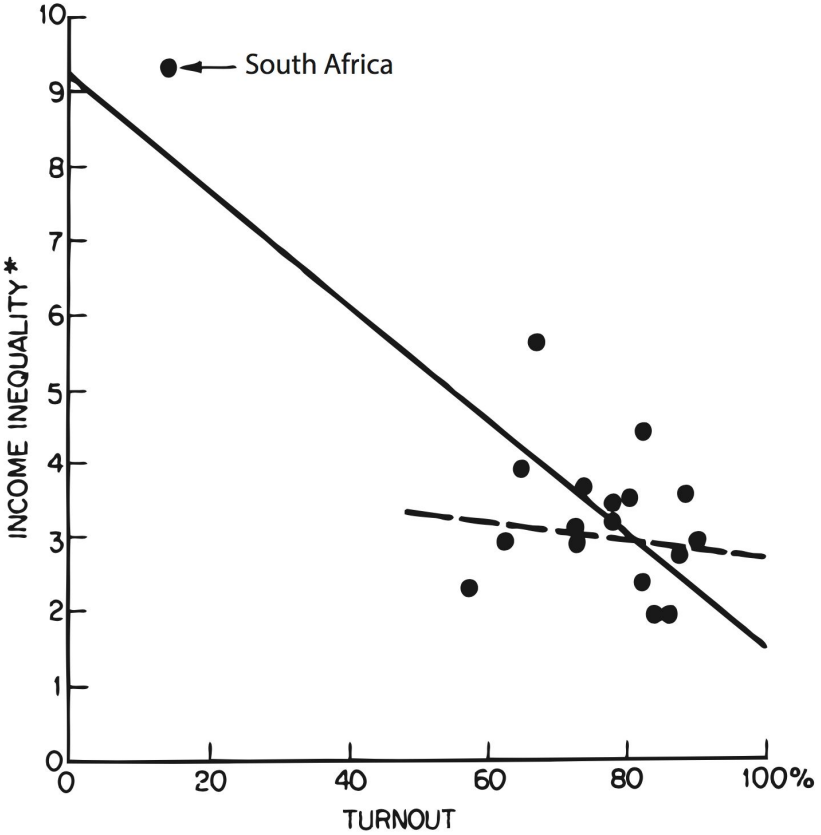
- Mean (x) = 9
- Var (x) = 11
- Mean (y) = 7.50
- Var (y) ~ 4.12
- Cor (x, y) = 0.816
- Linear regression line:
  - y = 3.00 + 0.500x



Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician 27 (1): 17–21.

Wikipedia

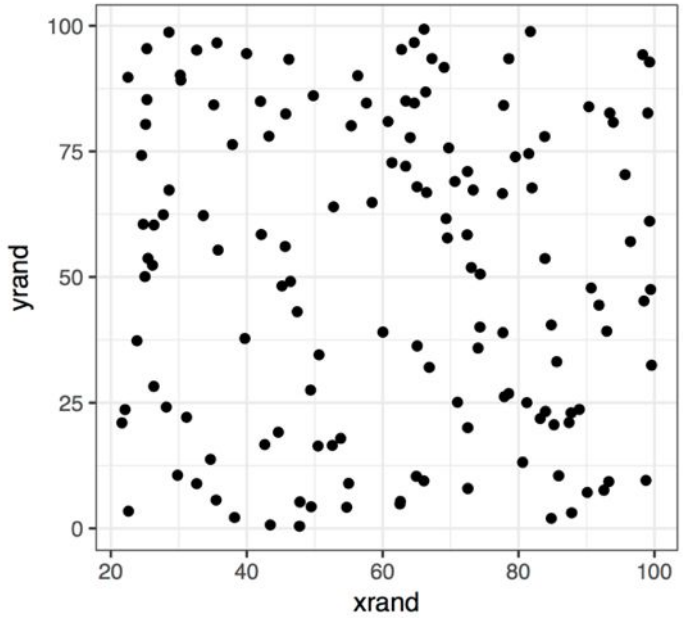
What does a correlation coefficient tell you about the data?

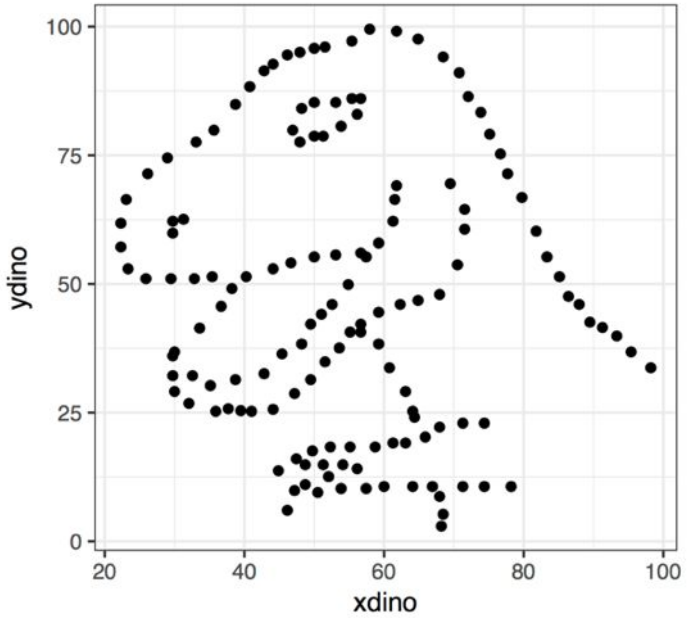


http://socviz.co/lookatdata.html

What does a correlation coefficient tell you about the data?

Correlation = -0.06

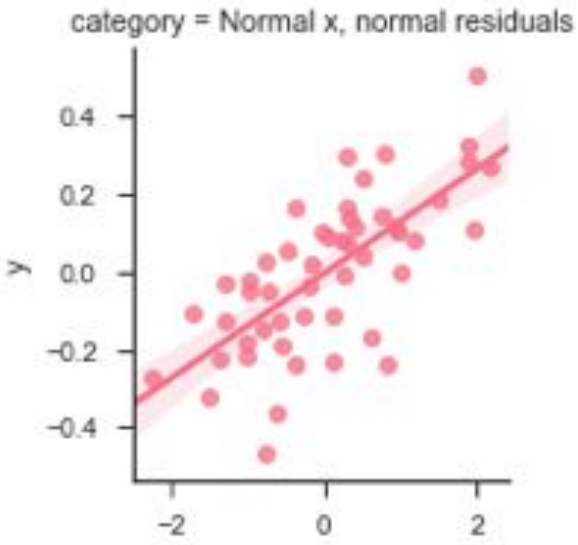




https://www.autodeskresearch.com/publications/samestats

What does a correlation coefficient tell you about the data?

Correlation = 0.7



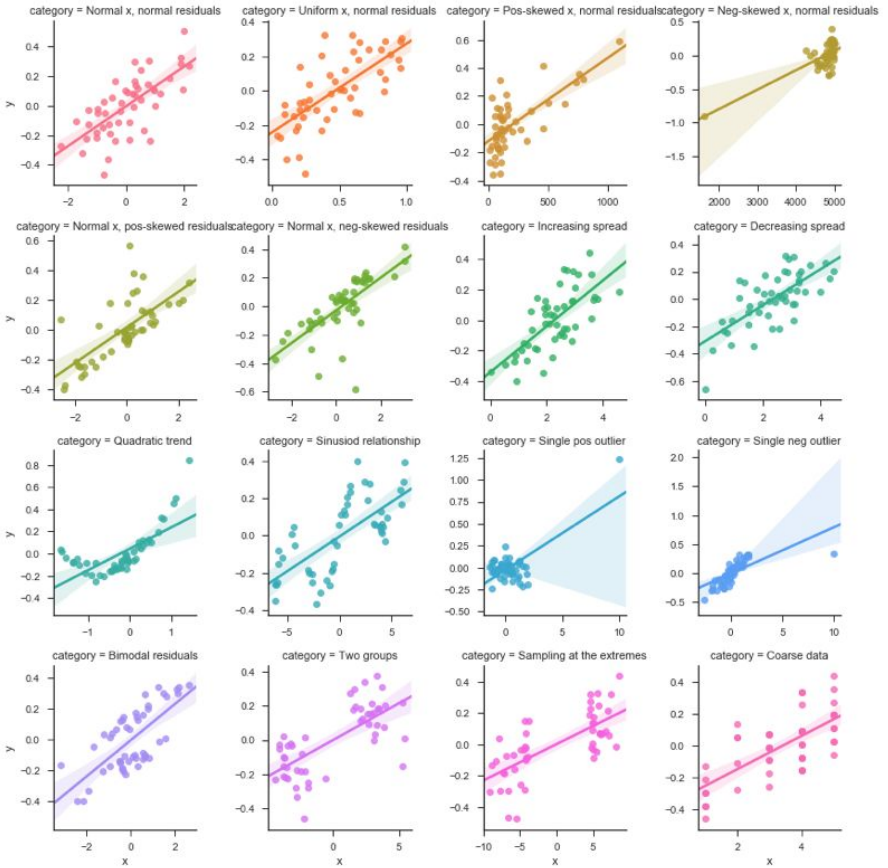
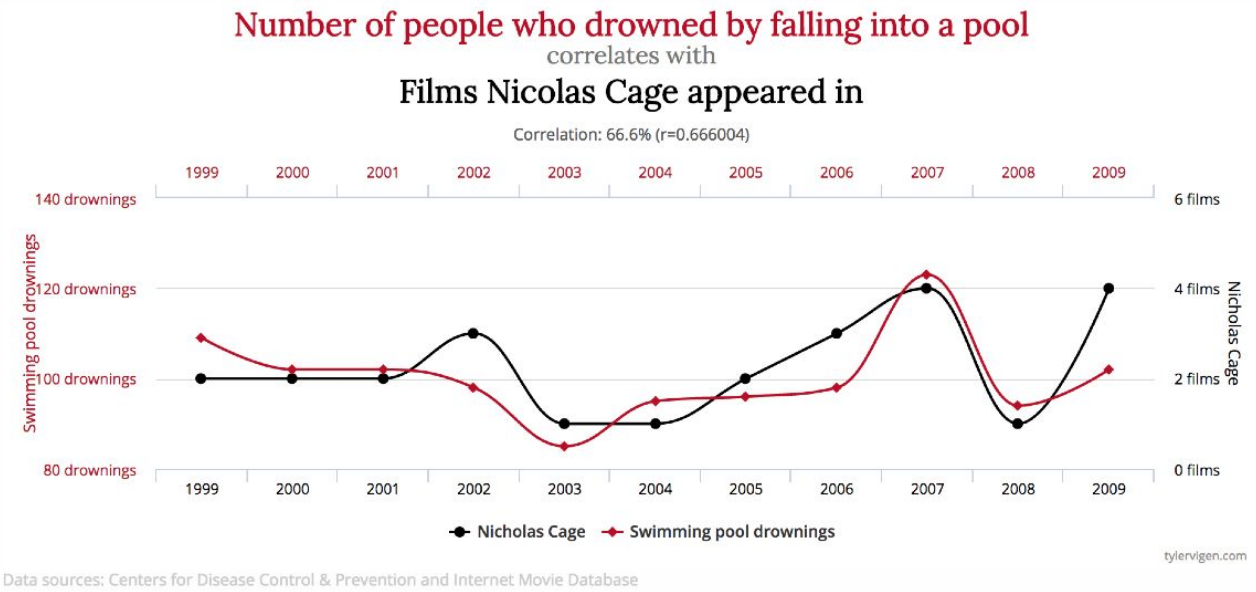


Figure adapted from code by Jan Vanhove: https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like

Spurious correlations

What does Nicholas Cage have to do with people drowning in swimming pools?

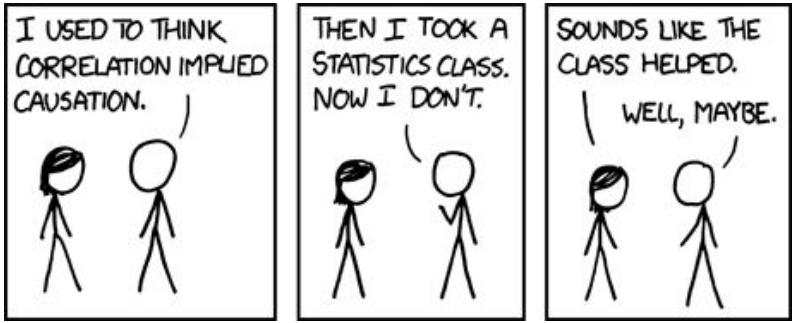


Checkout https://www.google.com/trends/correlate

Correlation does not imply causation

There is a significant correlation between annual chocolate consumption and number of Nobel laureates for different countries ($r(20)=.79$; $p<0.001$) → chocolate intake provides nutritional ground for sprouting Nobel laureates.

- Correlation can occur by random chance.
- Confounding variables could lead to correlation.
- Even when there is causation, there might not be obvious correlation.



xkcd.com/552

Many distance measures

Euclidean Distance

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Pearson Correlation Coefficient

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

Spearman Rank Correlation

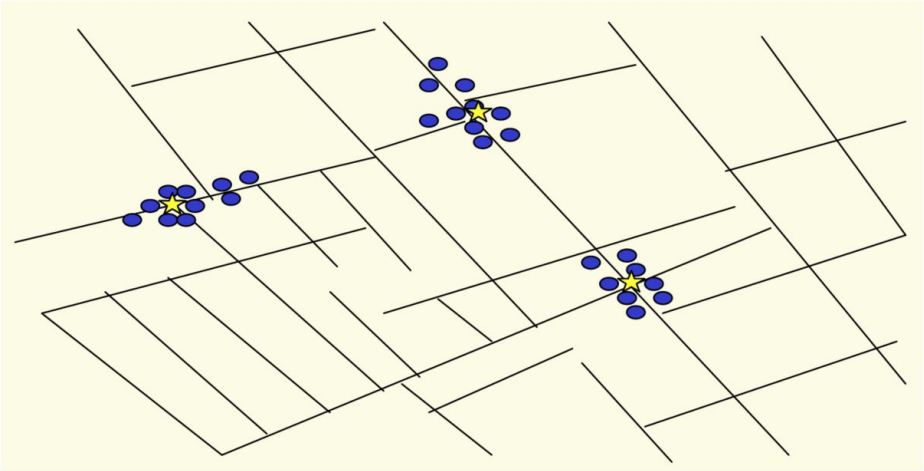
$$\rho = 1 - \frac{6 \sum_{i=1}^{n} [rank(x_i) - rank(y_i)]}{n(n^2 - 1)}$$

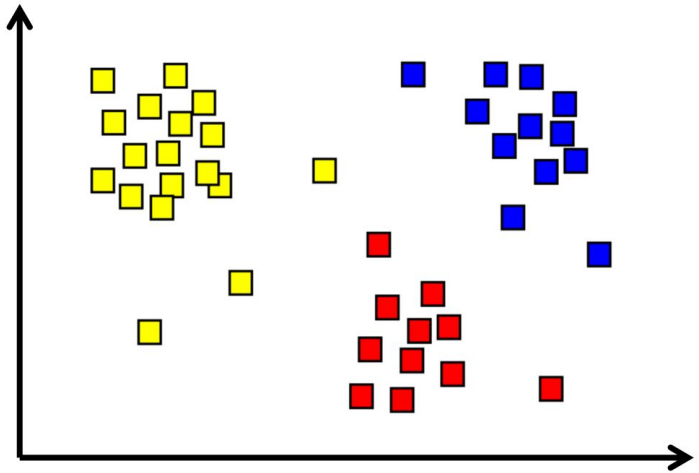
Mutual Information

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{(X,Y)}(x,y) \log \left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)} \right)$$

Wikipedia

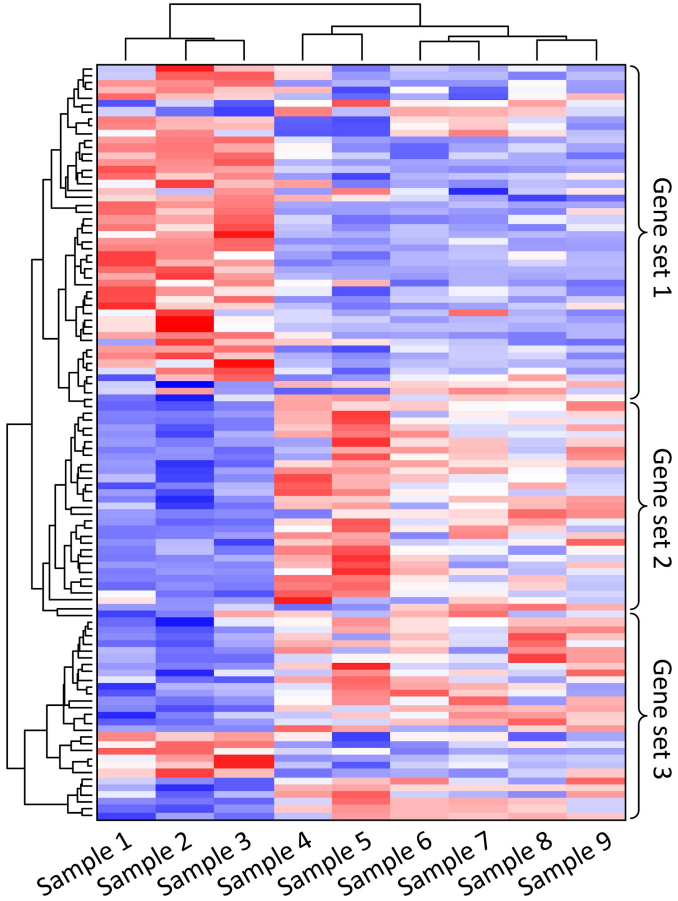
Clustering





Wikipedia

Clustering gene-expression profiles

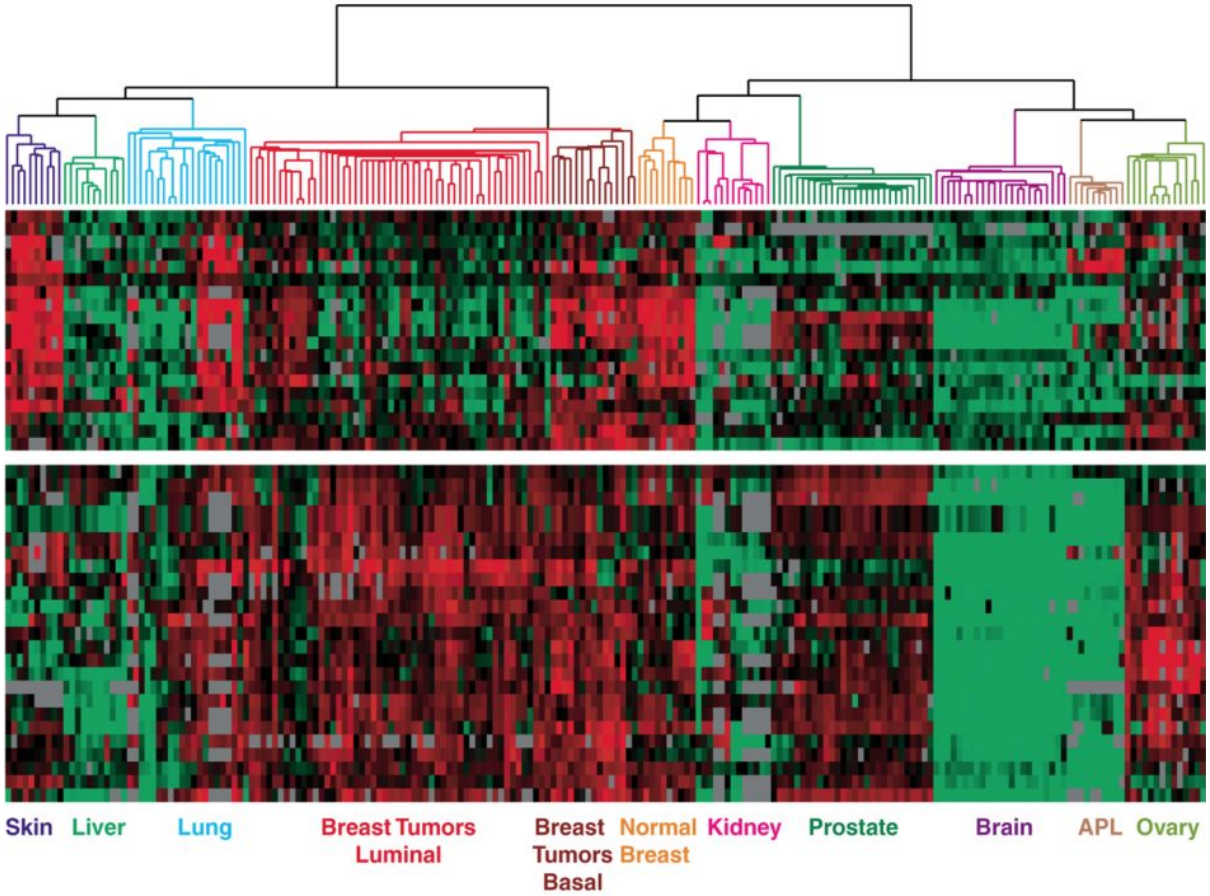


Group-level Qs:

1. Are there groups of genes that respond similarly to changing contexts (across samples)?
2. Are there groups of samples that have very similar gene expression profiles?

Lowe 2017 PLoS Comp. Biol.

Clustering gene-expression profiles

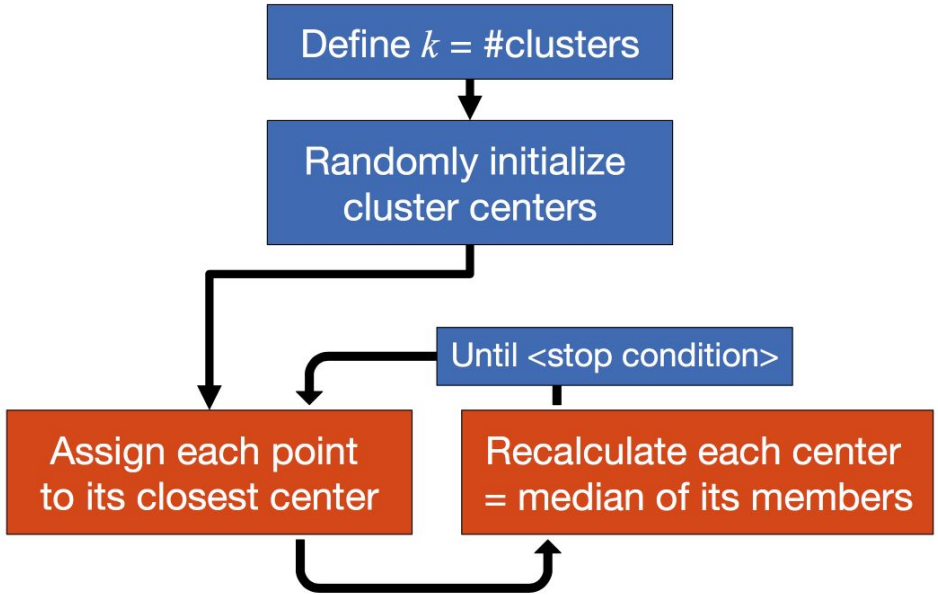


Group-level Qs:

1. Are there groups of genes that respond similarly to changing contexts (across samples)?
2. Are there groups of samples that have very similar gene expression profiles?

Wikipedia

K-means clustering



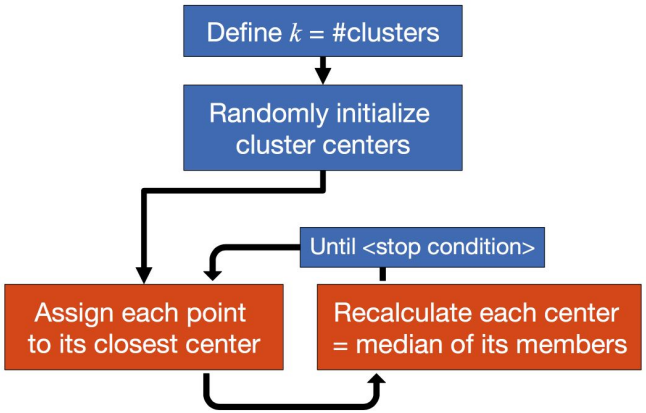
Conceptually similar to Expectation-Maximization, alternating between 2 two steps:

1. **E step:** Creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters.
2. **M step:** Computes parameters maximizing the expected log- likelihood found on the E step.

These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Checkout http://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-means clustering



Stopping condition

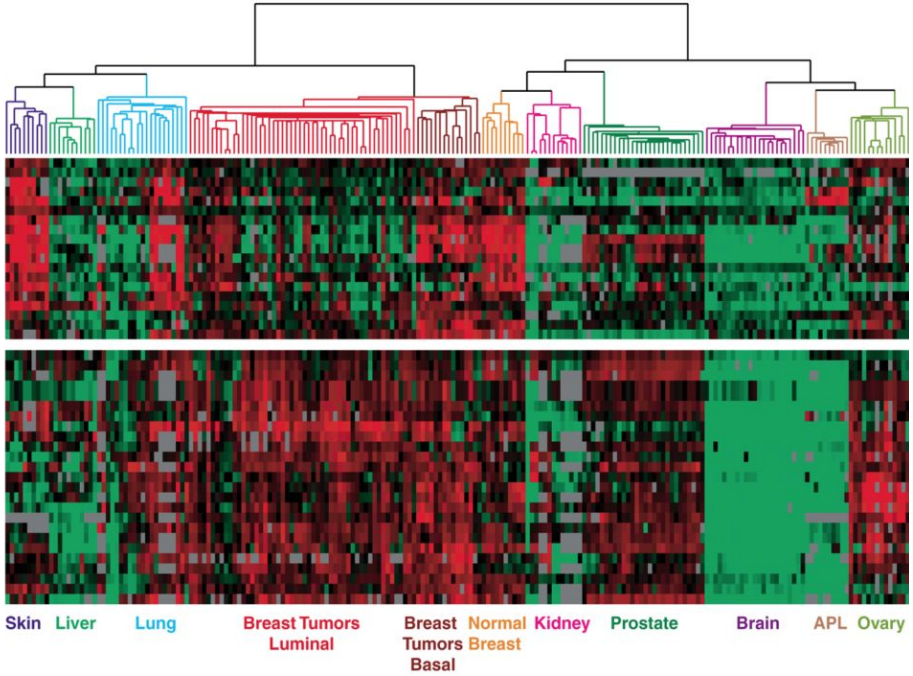
- Until the change in centers is less than <constant>.
- Until all genes get assigned to the same partition.
- Until some minimal number of genes (e.g. 90%) get assigned to the same partition twice in a row.

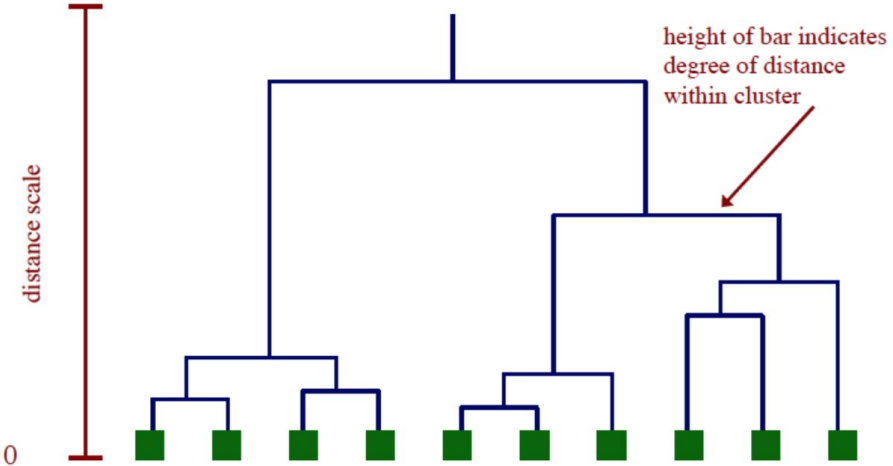
Some issues

- Have to set **k** ahead of time.
- Works well if clusters of approx. similar sizes.
- Each gene only belongs to 1 cluster.

- Genes assigned to clusters on the basis of all experiments; Experiments assigned to clusters based on all genes.

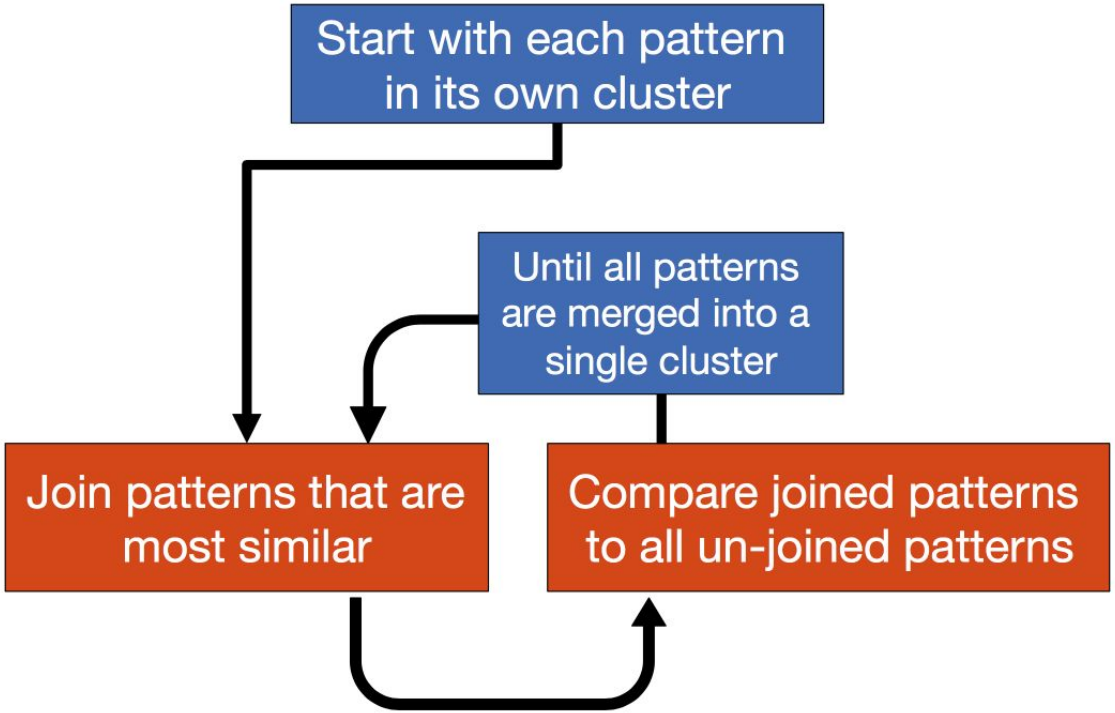
Hierarchical clustering



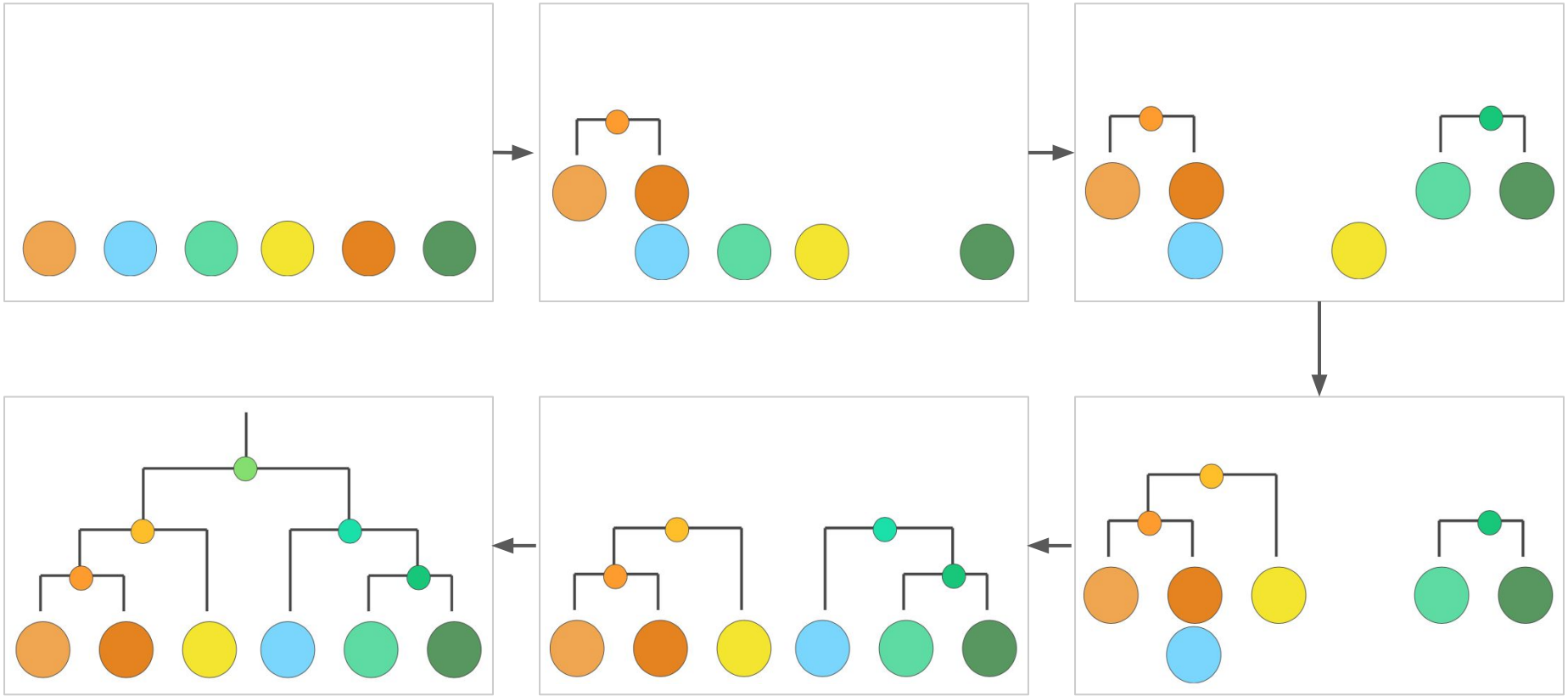


- Imposes hierarchical structure on all of the data.
- Easy visualization of similarities and differences between genes (experiments) and clusters of genes (experiments).

Hierarchical clustering



Hierarchical clustering



Hierarchical clustering

Linkage criteria:

- Single/Minimum linkage (nearest neighbors)
- Complete/Maximum linkage (farthest neighbors)
- Average linkage (average of all pairs)

Complete linkage clustering of 5 objects

1

	A	B	C	D	E
A	0				
B	5	0			
C	10	3	0		
D	15	6	7	0	
E	20	8	2	11	0

2

	A	B	CE	D
A	0			
B	5	0		
CE	20	8	0	
D	15	6	11	0

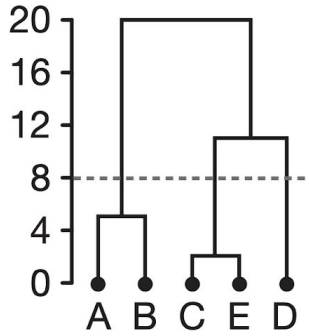
3

	AB	CE	D
AB	0		
CE	20	0	
D	15	11	0

4

	AB	CED
AB	0	
CED	20	0

Dendrogram

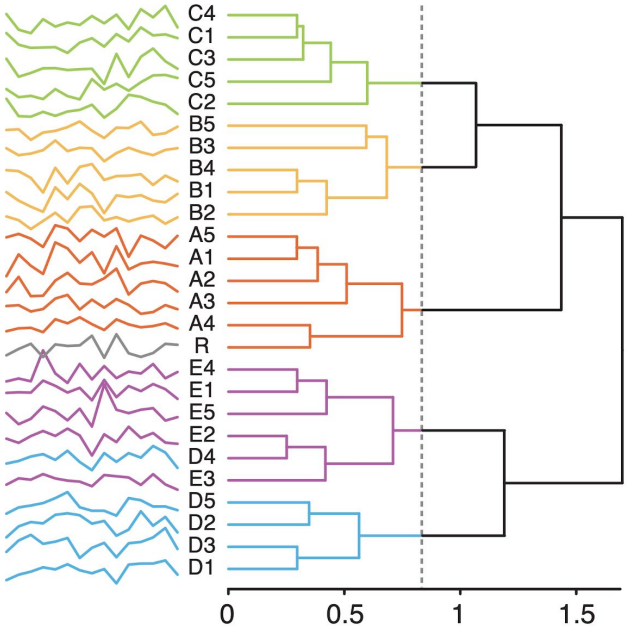


Altman & Krzywinski (2017) Nat. Meth.

Hierarchical clustering

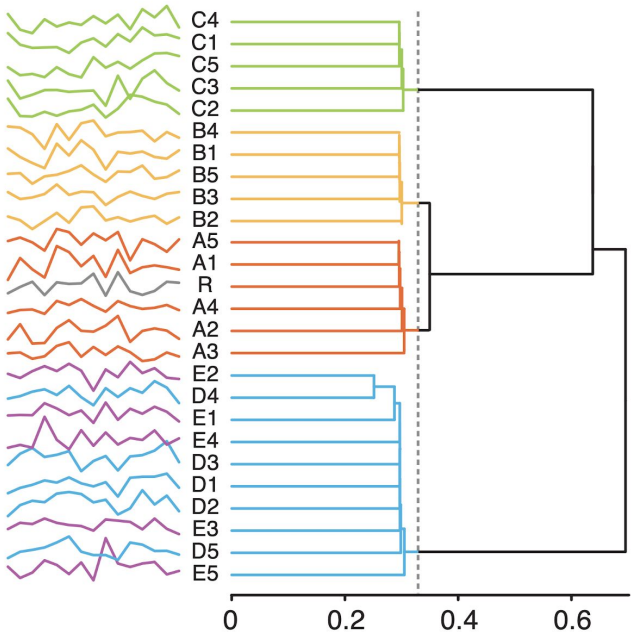
Complete linkage clustering

Tends to create balanced dendrograms by first clustering objects into small nodes and then clustering the nodes



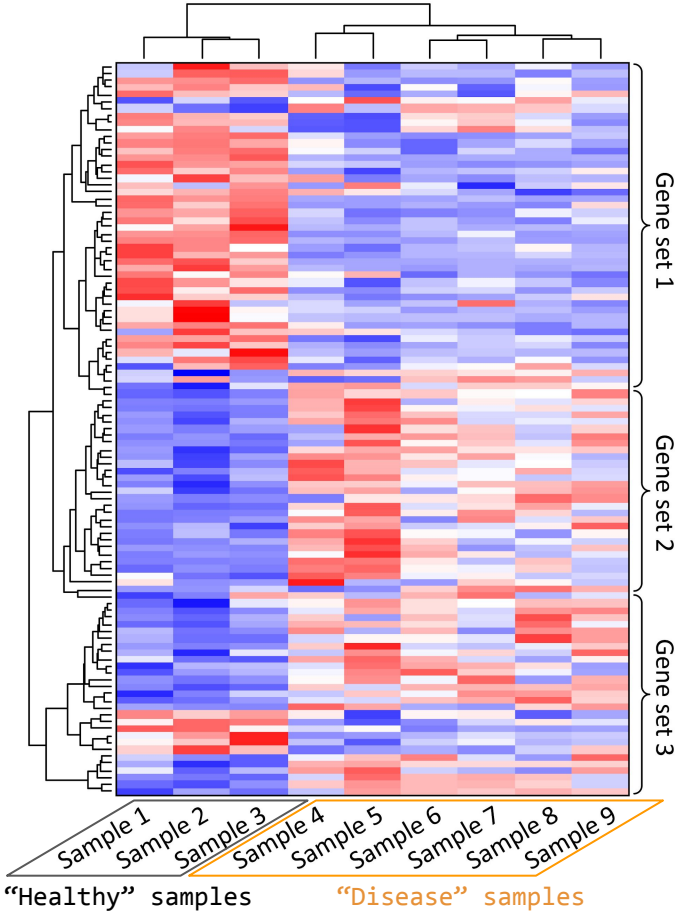
Single linkage clustering

Tends to create stringy dendrograms by first creating a few nodes and then adding objects to them one at a time



Altman & Krzywinski (2017) Nat. Meth.

A gene expression dataset



Gene-level Qs:

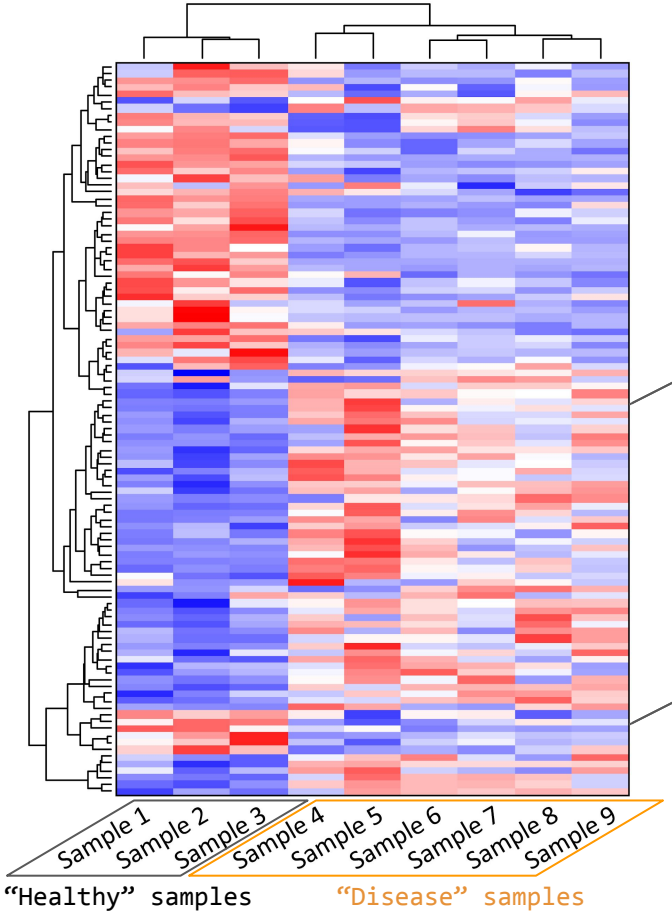
1. What's expressed (& by how much) in a given context/condition?
2. What's differentially expressed between two (or more) contexts/conditions?

Group-level Qs:

1. Are there groups of genes that respond similarly to changing contexts (across samples)?
2. Are there groups of samples that have very similar gene expression profiles?

Lowe 2017 PLoS Comp. Biol.

Gene-level analysis



	“Healthy” samples			“Disease” samples					
Gene_X	10.1	12.8	13.4	9.05	10.2	5.47	6.09	4.56	11.3
Gene_Y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.8

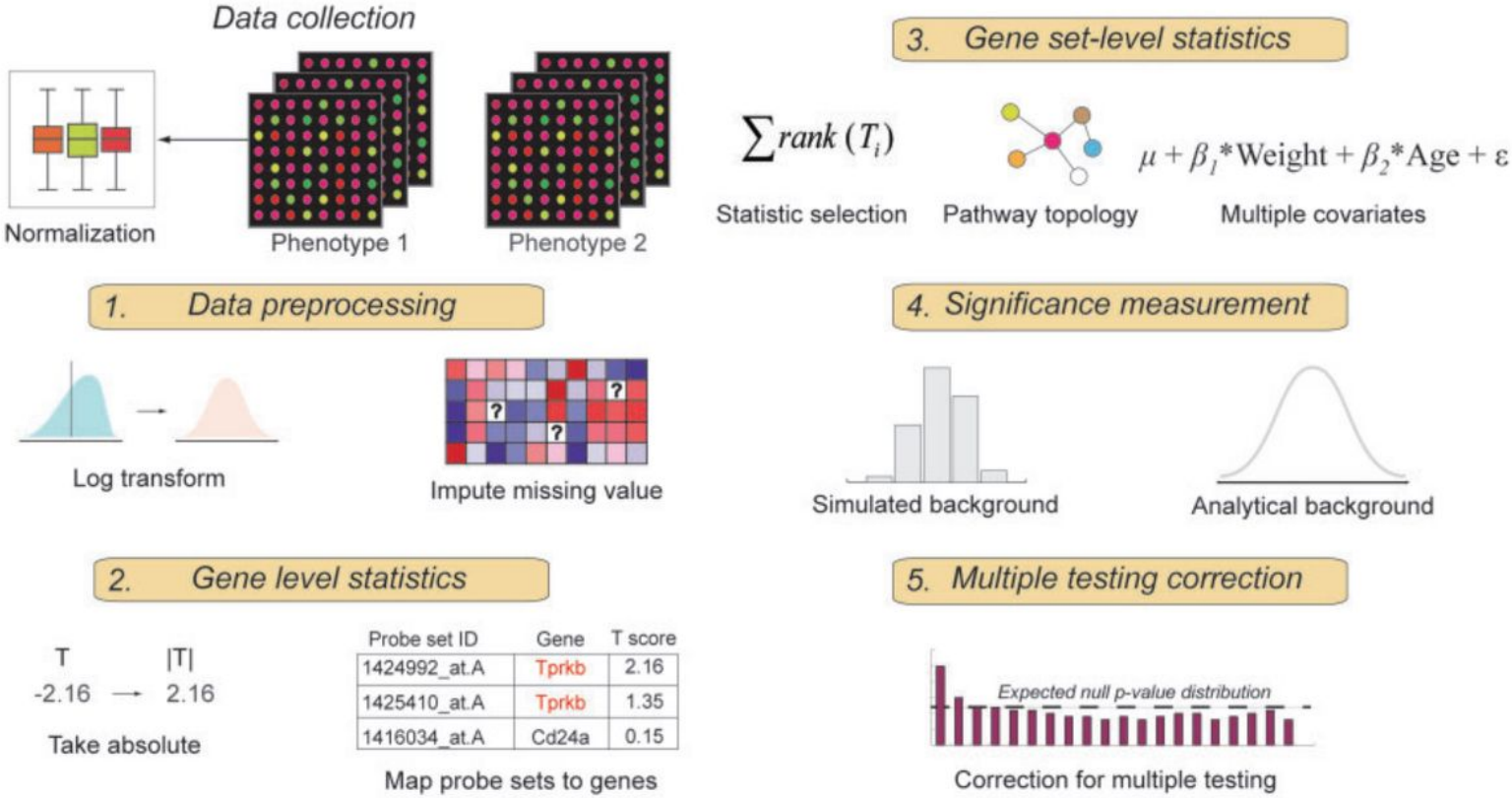
Lowe 2017 PLoS Comp. Biol.

Differential expression & Enrichment analysis

- Quantile normalization
- Differential expression using Empirical Bayes
  - Moderated t-test
  - Permutation test
- Enrichment analysis
  - Hypergeometric test
  - PAGE

Altman & Krzywinski (2017) Nat. Meth.

Enrichment analysis

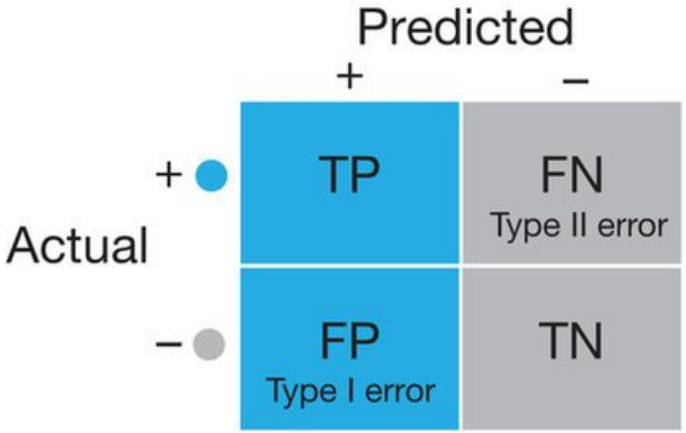


Hung (2012) Brief. Bioinfo.

Enrichment analysis

Hung (2012) Brief. Bioinfo.

Evaluating a clustering with external/prior knowledge



Sensitivity (recall) TP/●

False negative rate FN/●

False positive rate FP/●

Specificity TN/●

Precision TP/■

FDR FP/■

Accuracy

$(TP + TN)/(● + ●)$

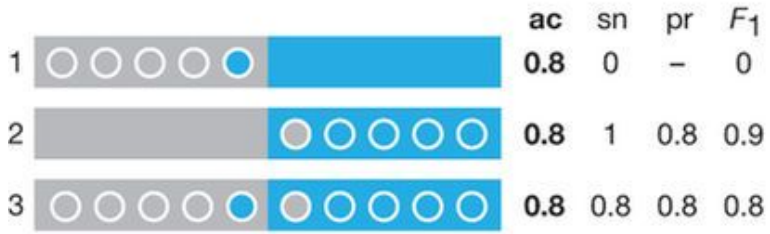
F_1 score

$2TP/(2TP + FP + FN)$

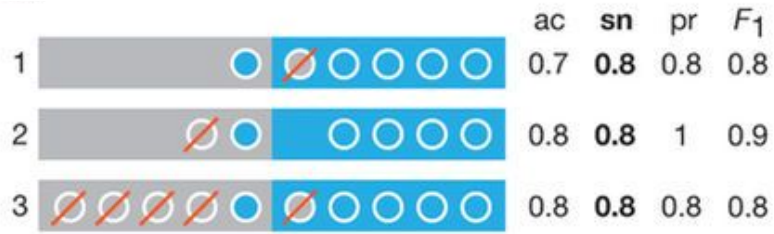
Lever, Krzywinski, & Altman (2016) Nat. Meth.

Evaluating a clustering with external/prior knowledge

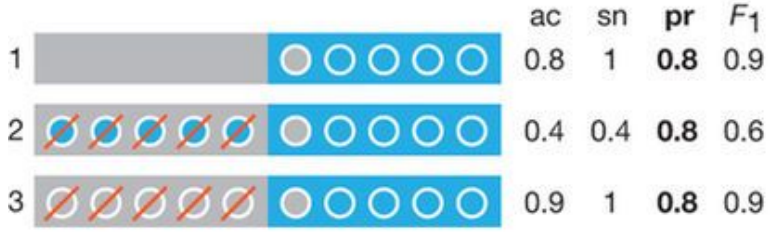
Four groups of 3 different classification scenarios that have the same value (0.8) for the metric in bold



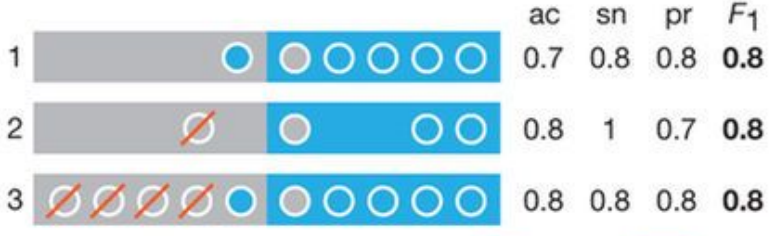
	ac	sn	pr	F_1
1	**0.8**	0	–	0
2	**0.8**	1	0.8	0.9
3	**0.8**	0.8	0.8	0.8



	ac	sn	pr	F_1
1	0.7	**0.8**	0.8	0.8
2	0.8	**0.8**	1	0.9
3	0.8	**0.8**	0.8	0.8



	ac	sn	pr	F_1
1	0.8	1	**0.8**	0.9
2	0.4	0.4	**0.8**	0.6
3	0.9	1	**0.8**	0.9

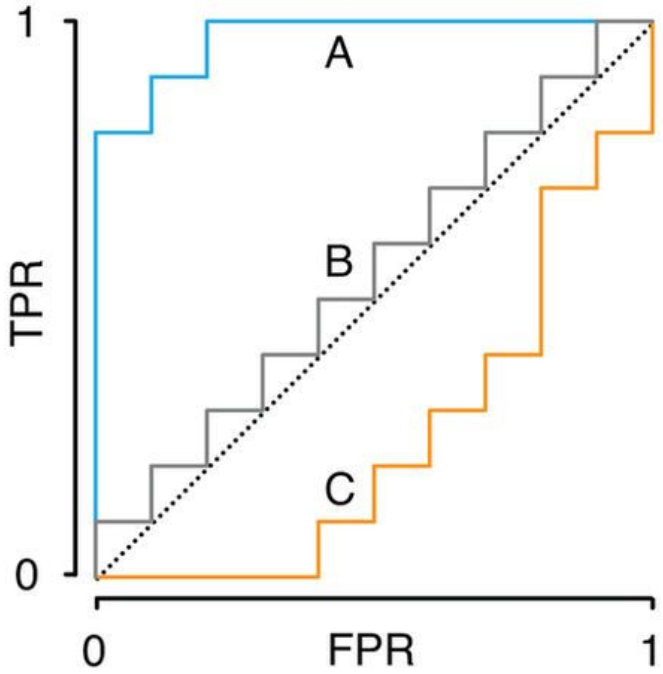


	ac	sn	pr	F_1
1	0.7	0.8	0.8	**0.8**
2	0.8	1	0.7	**0.8**
3	0.8	0.8	0.8	**0.8**

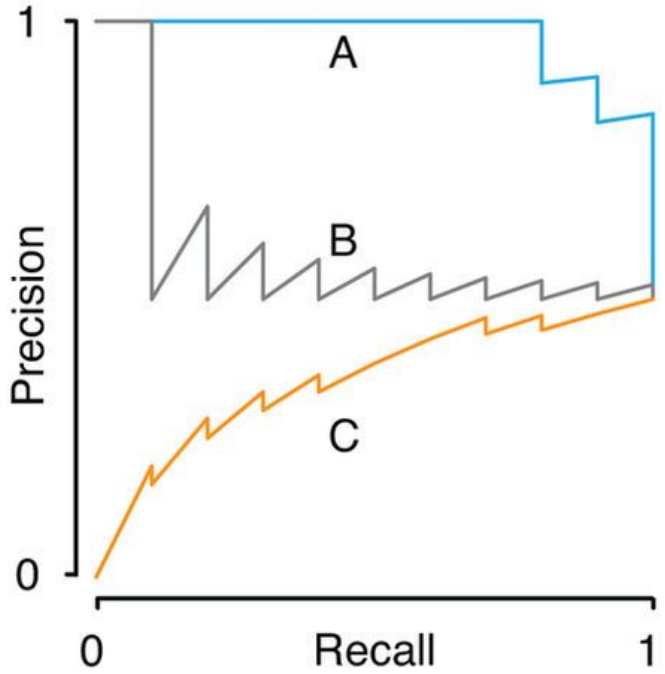
Lever, Krzywinski, & Altman (2016) Nat. Meth.

Evaluating a clustering with external/prior knowledge

ROC curve



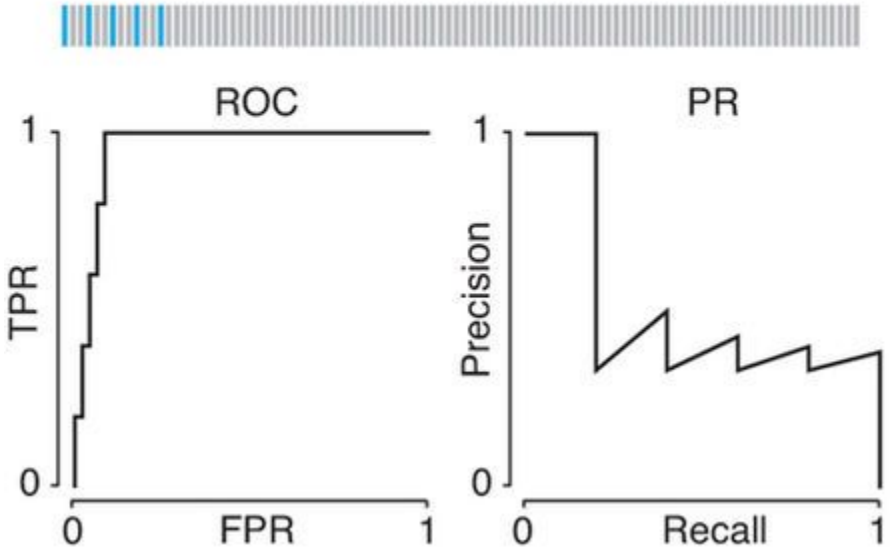
Precision recall curve



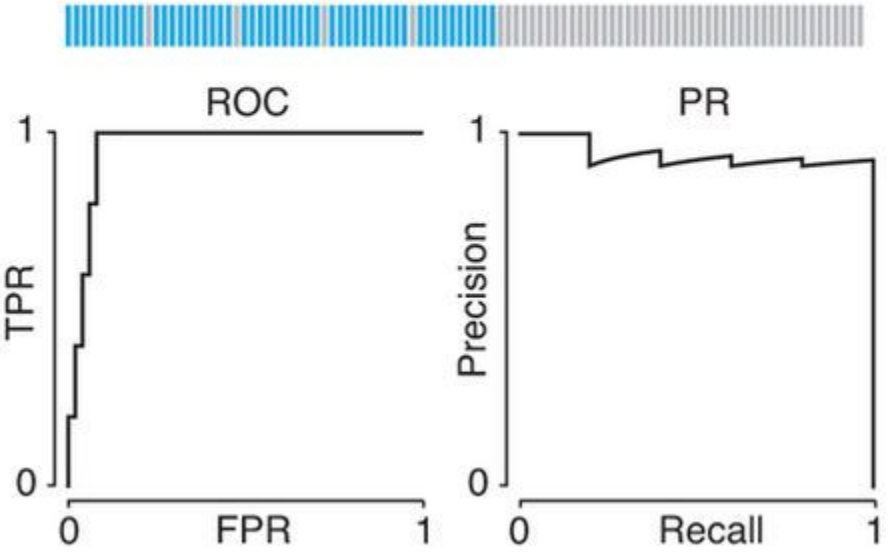
Lever, Krzywinski, & Altman (2016) Nat. Meth.

Evaluating a clustering with external/prior knowledge

5% positive



50% positive

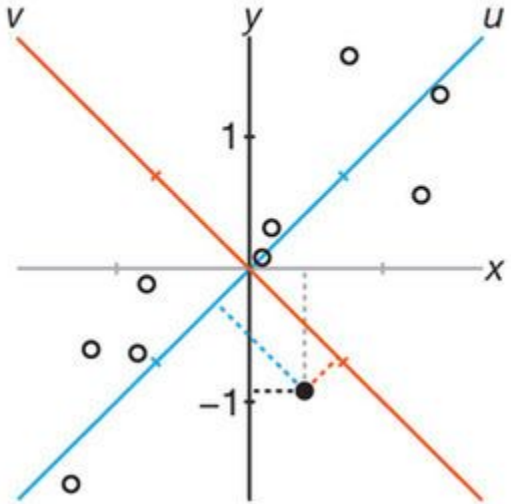


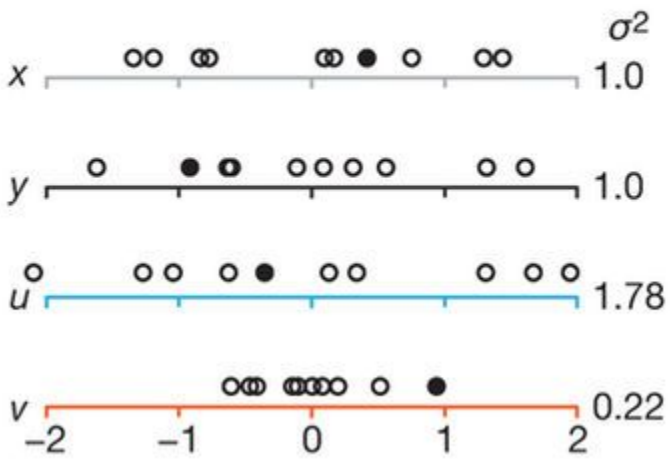
Lever, Krzywinski, & Altman (2016) Nat. Meth.

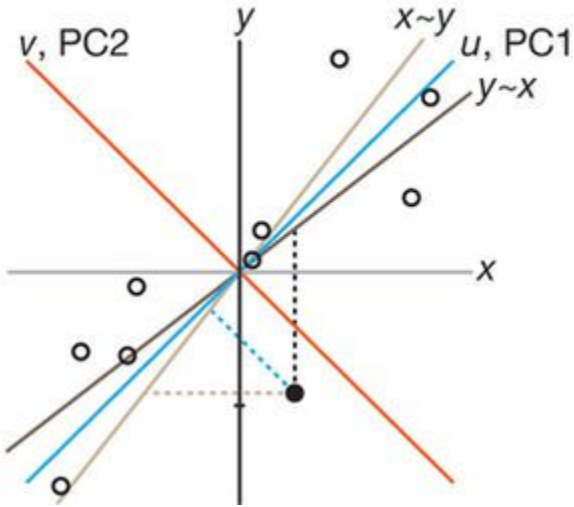
Dimensionality reduction by PCA

PCA geometrically projects data onto a lower-dimensional space

- Each lower dimension is a 'linear' combination of correlated original dimensions.
- The principal components (PCs) represent the directions of maximum variation.

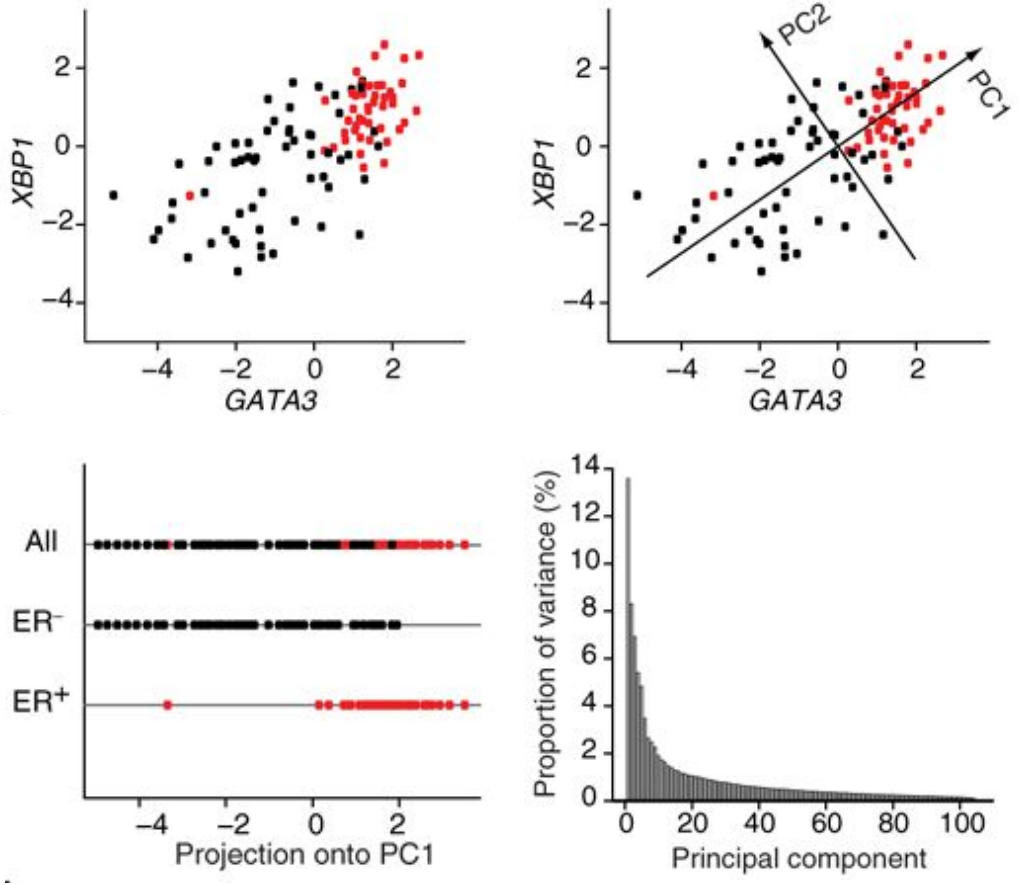






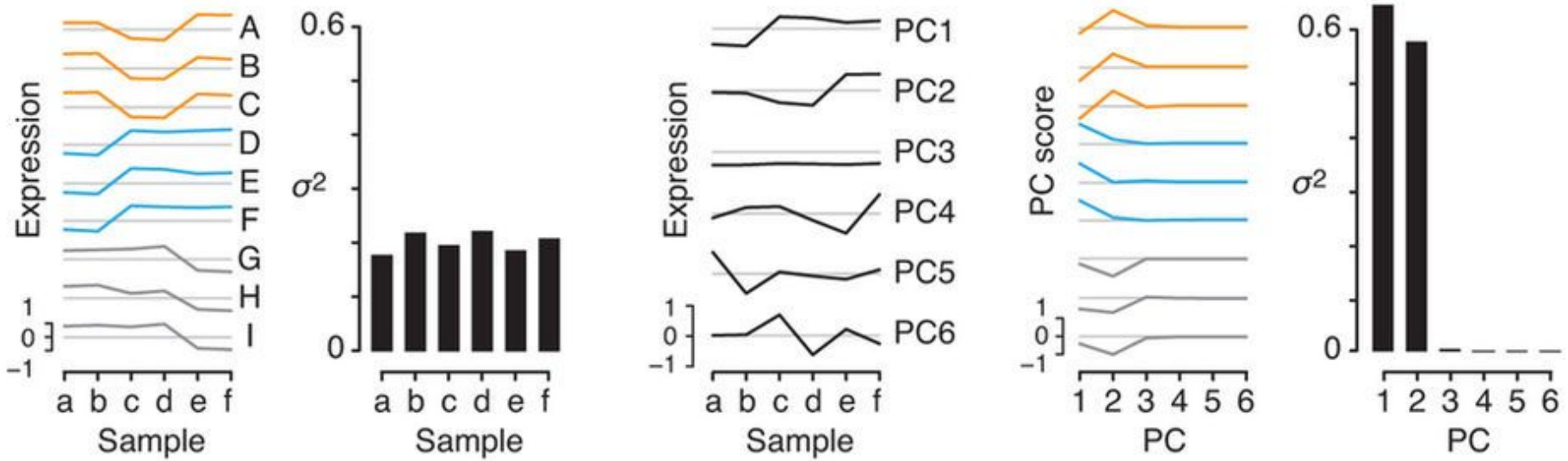
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



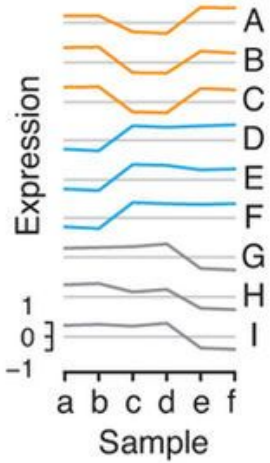
Ringnér (2008) Nat. Biotech

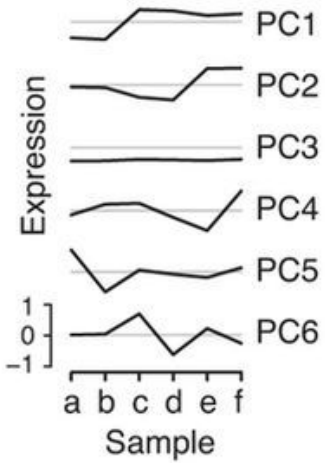
Dimensionality reduction by PCA

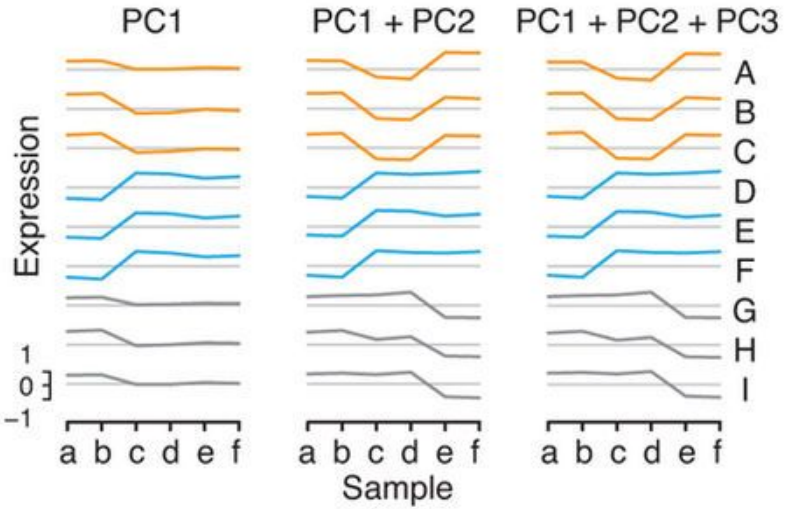


Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA

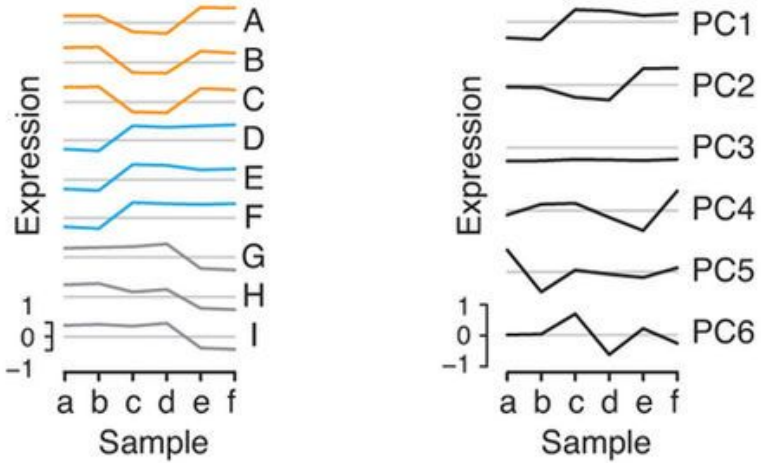


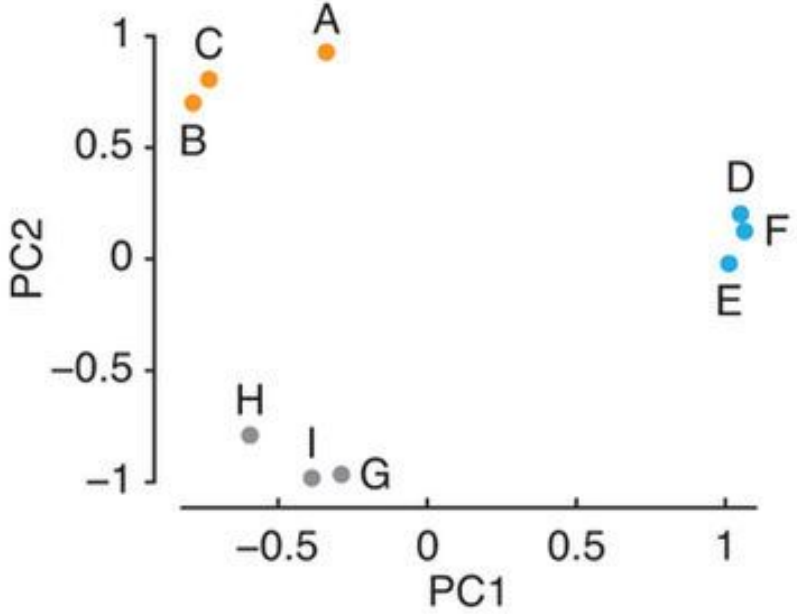




Lever, Krzywinski, & Altman (2017) Nat. Meth.

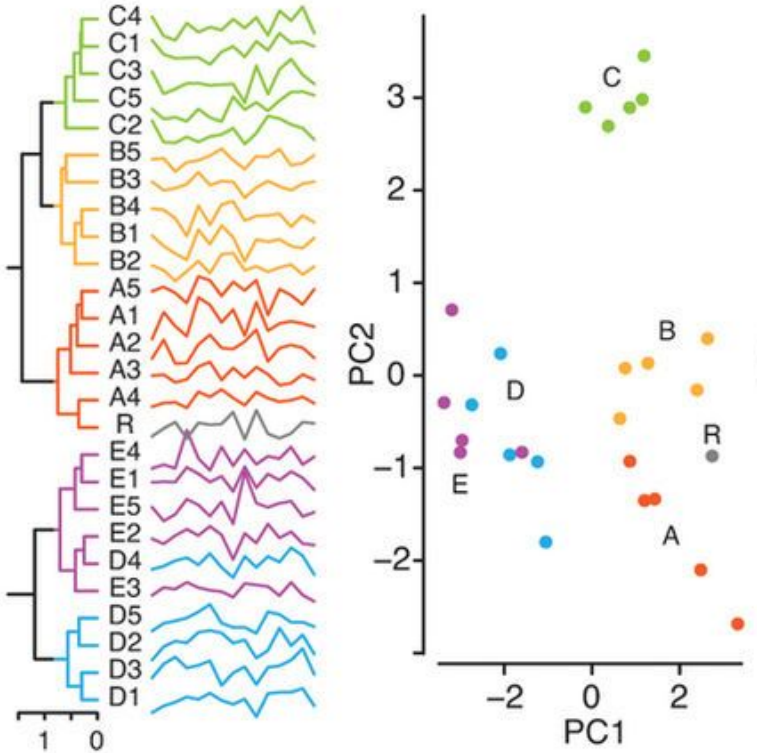
Dimensionality reduction by PCA





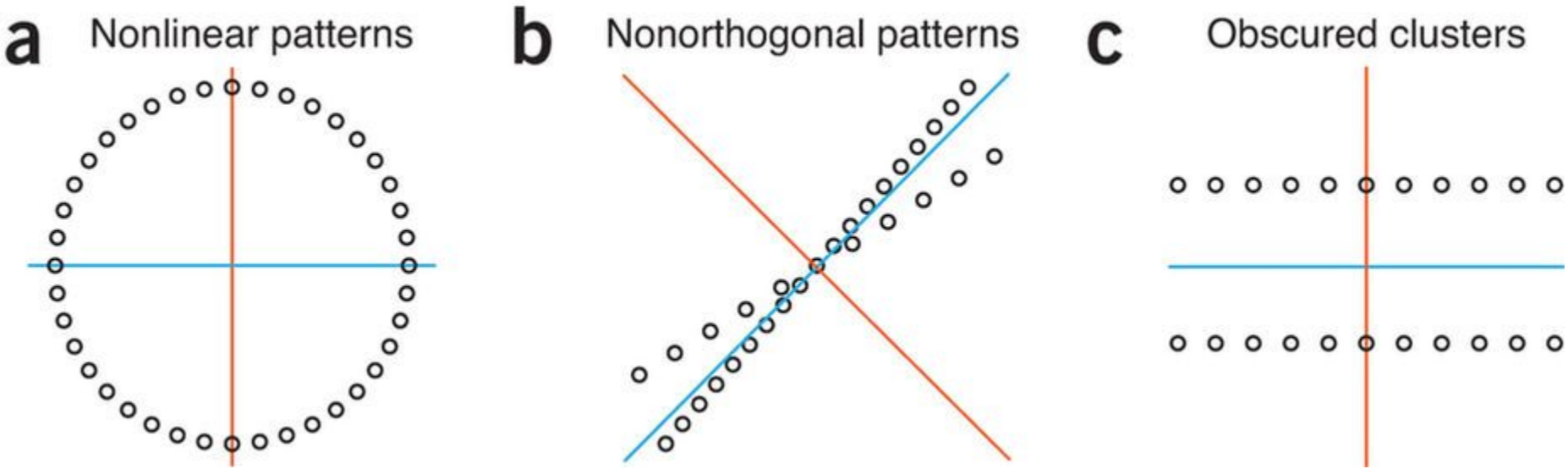
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.