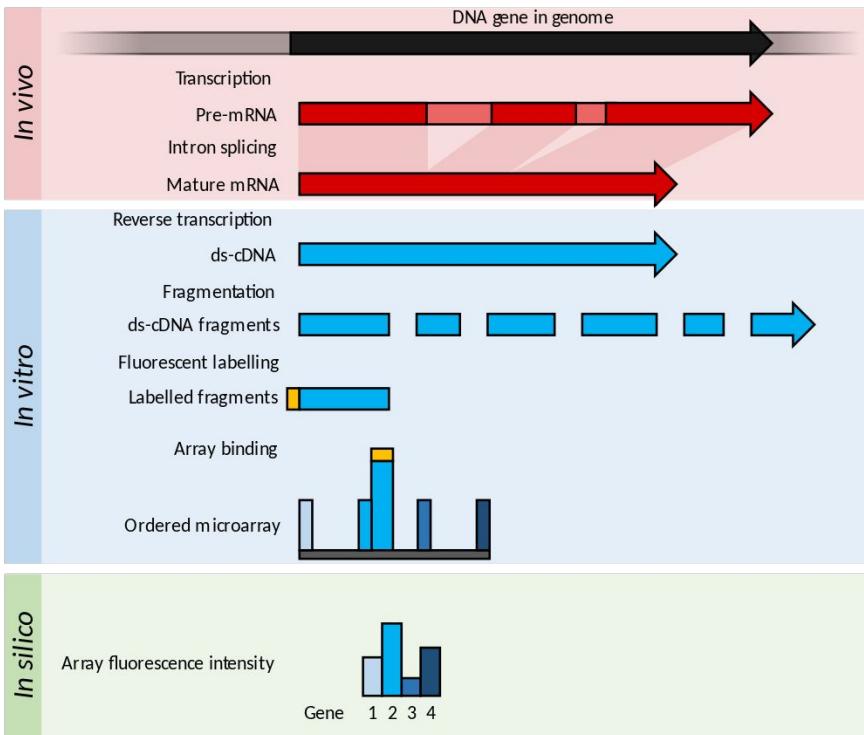


Functional genomics

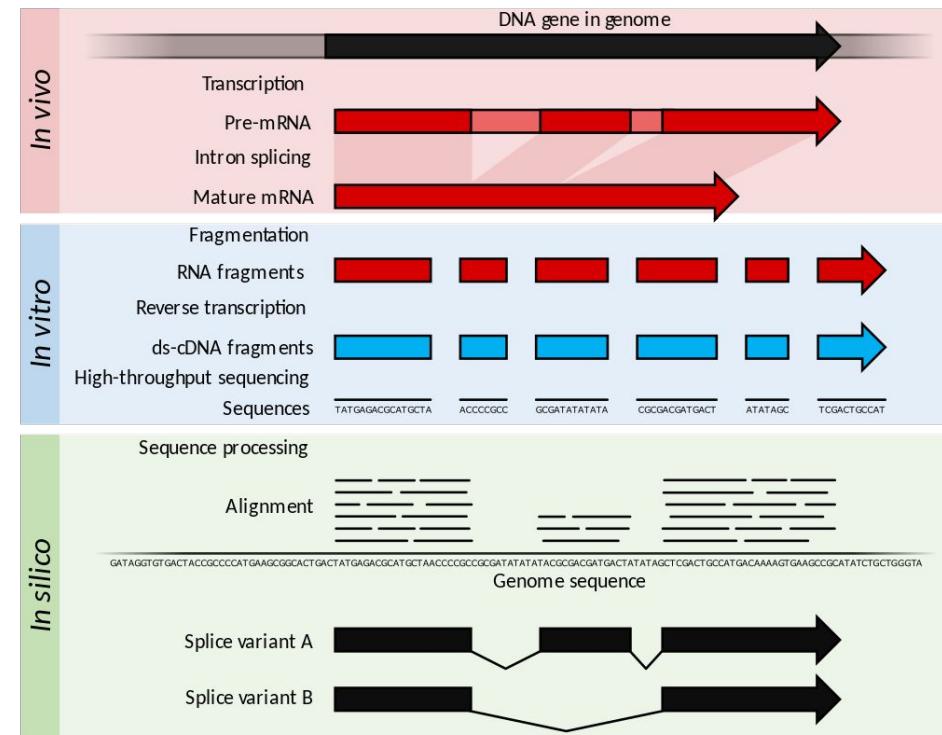
- Measuring gene-expression
- Distance/Similarity measures
- Clustering genes/samples
- Differential expression
- Functional enrichment

Measuring gene-expression on a large-scale

DNA microarrays

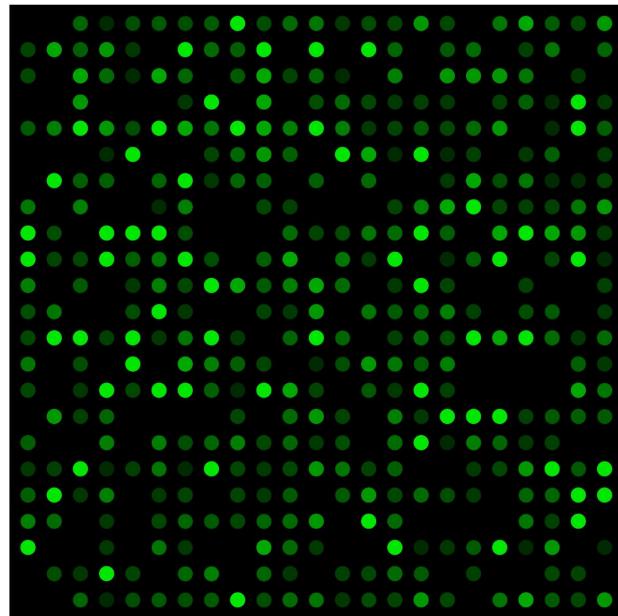


RNA-seq

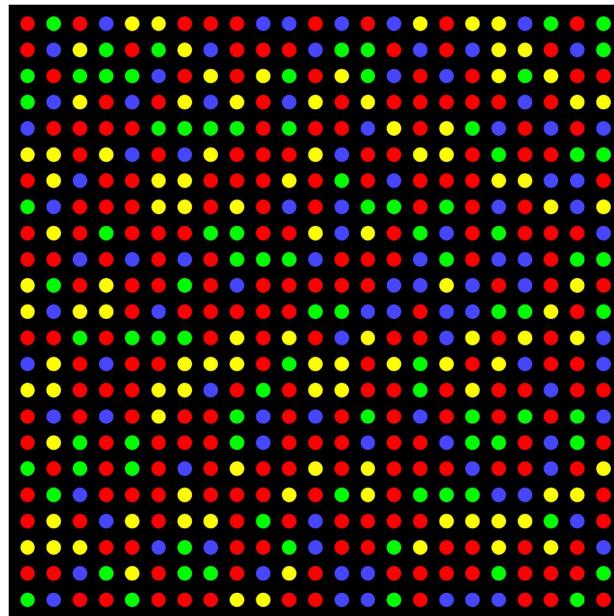


Measuring gene-expression on a large-scale

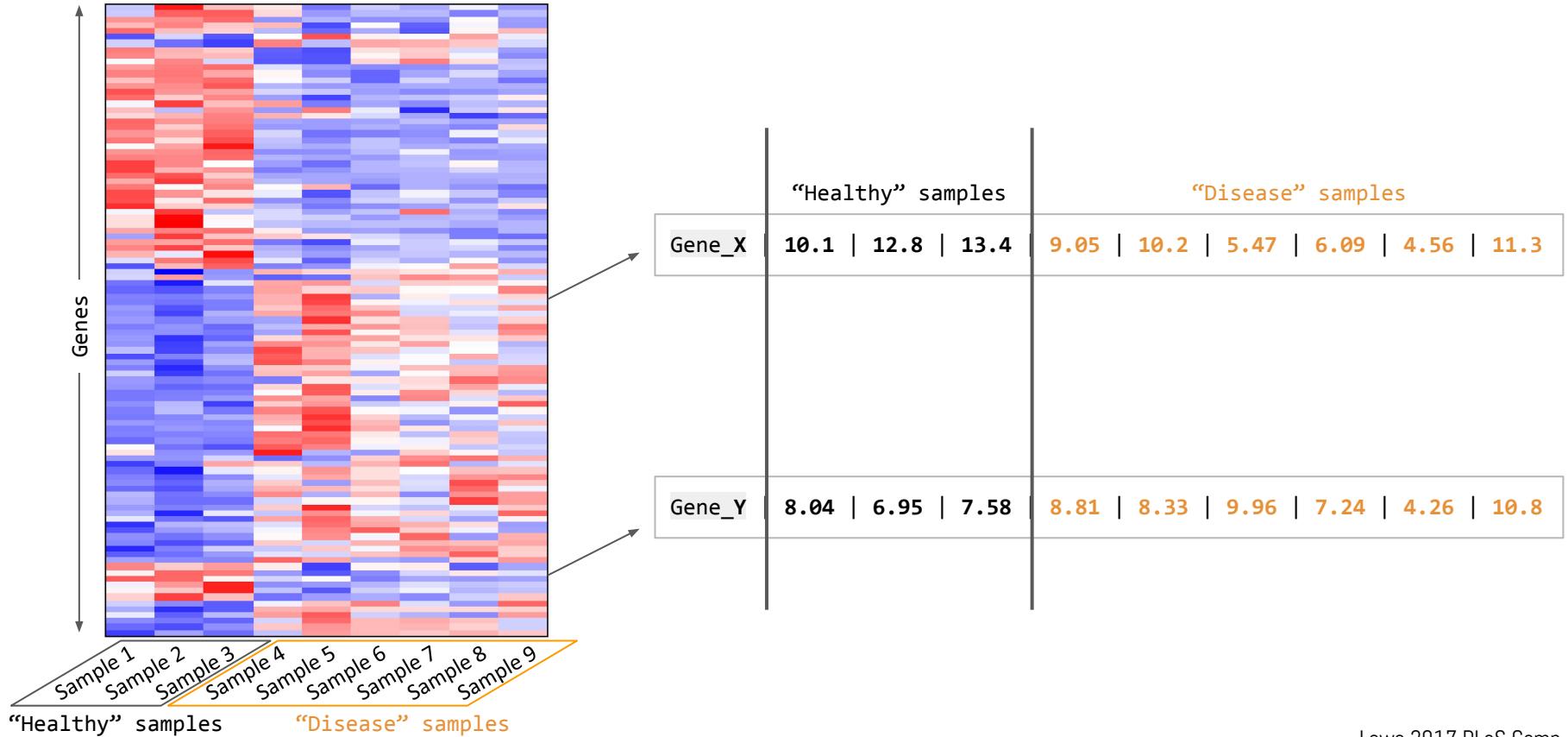
DNA microarrays



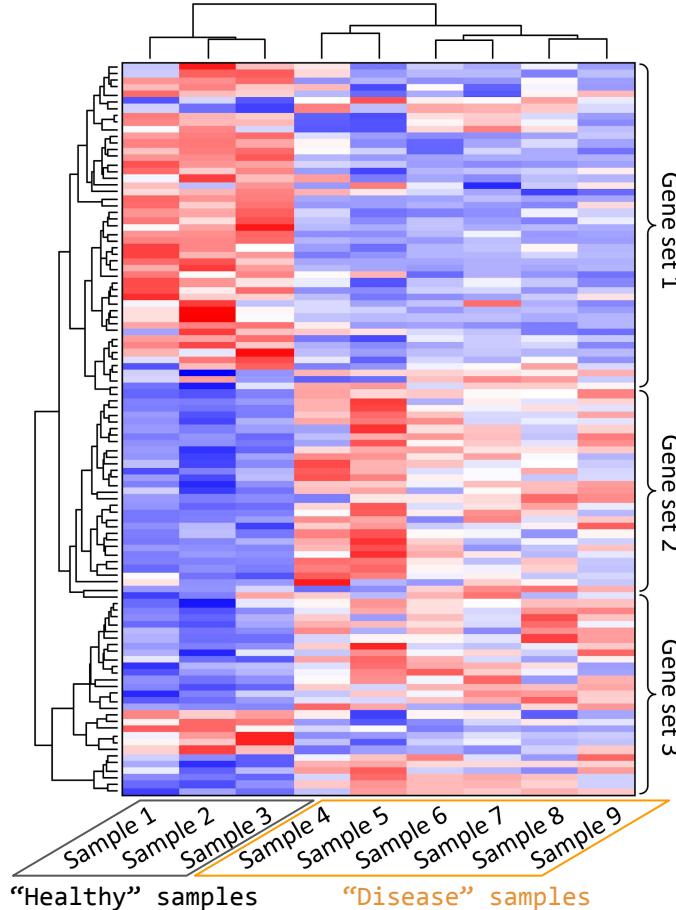
RNA-seq



A gene expression dataset



A gene expression dataset



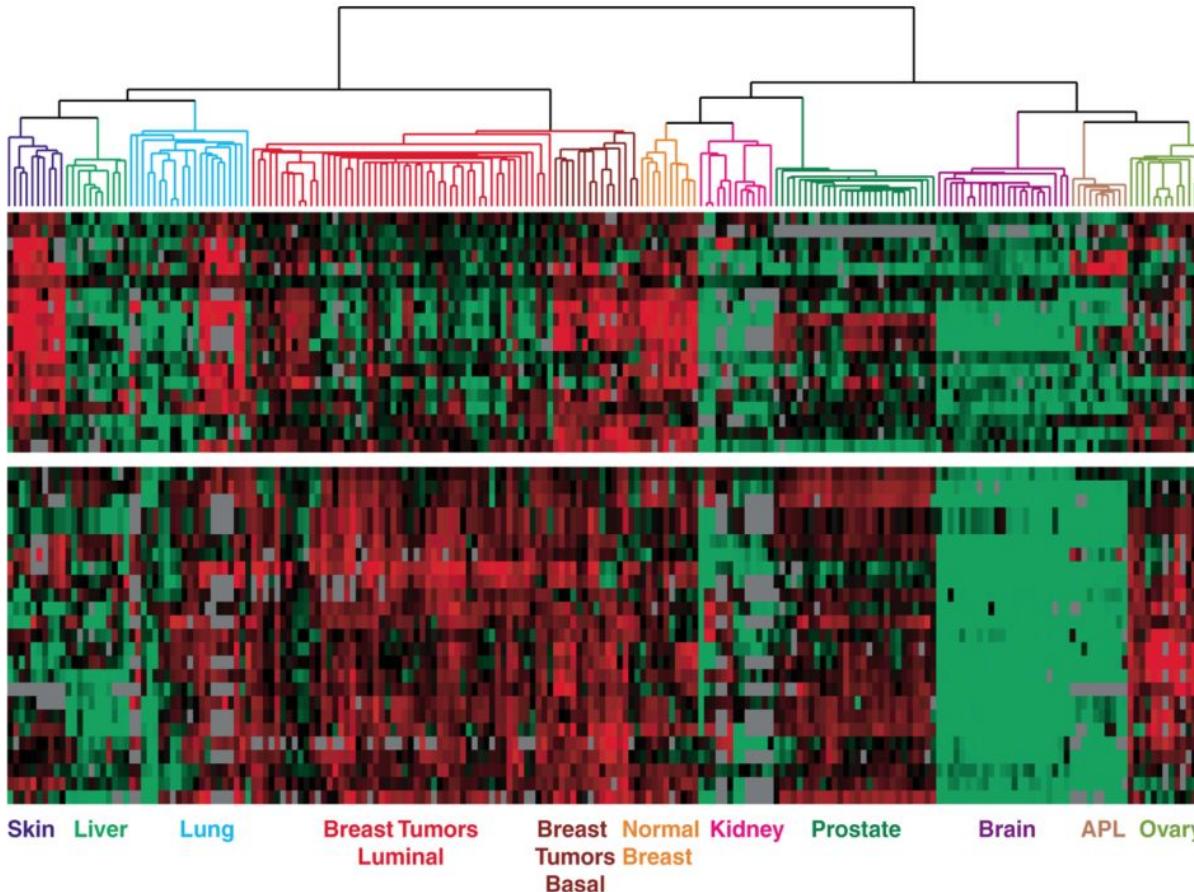
Gene-level Qs:

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

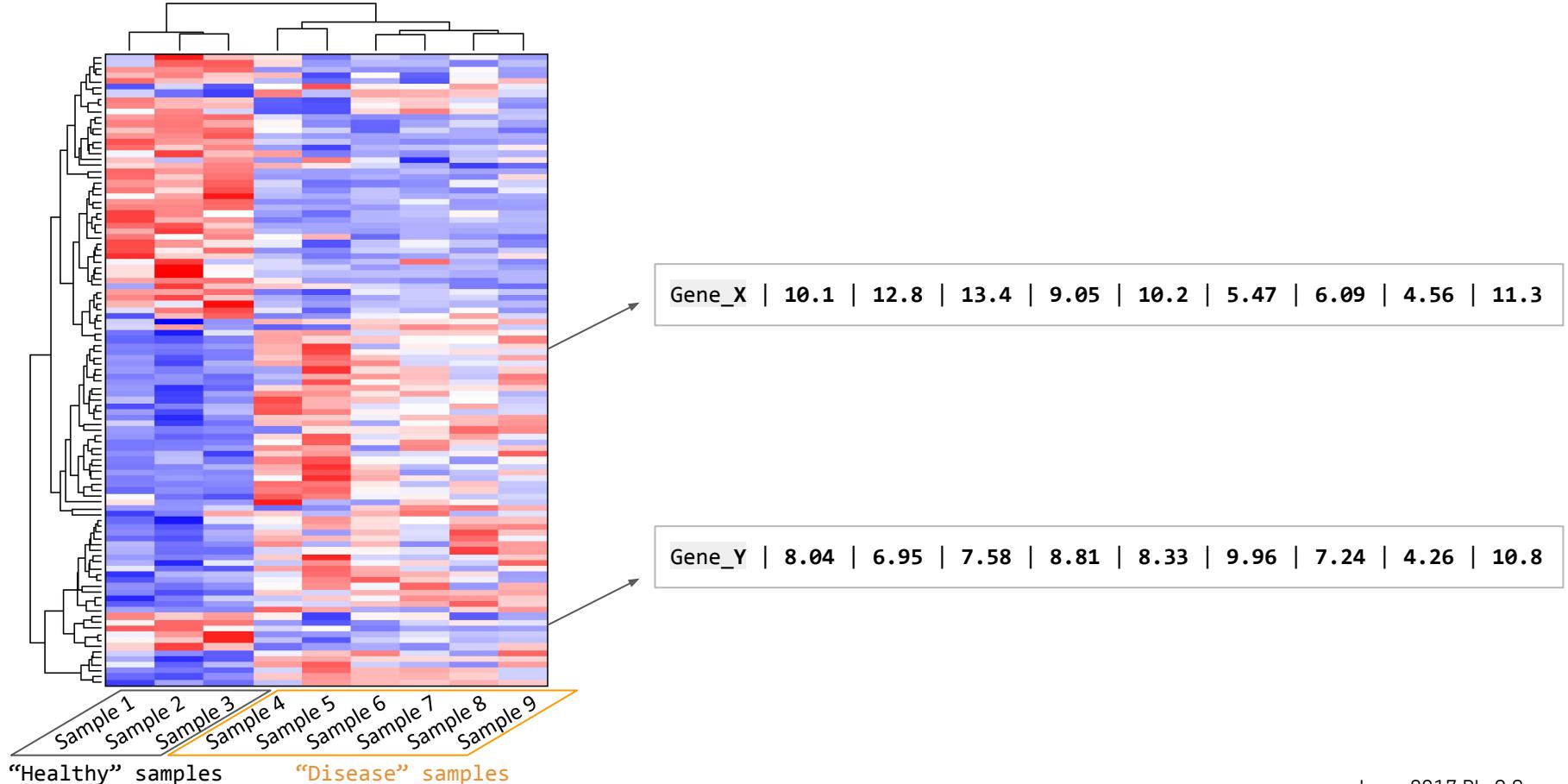
Group-level Qs:

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?

One of the first large-scale gene expression analyses



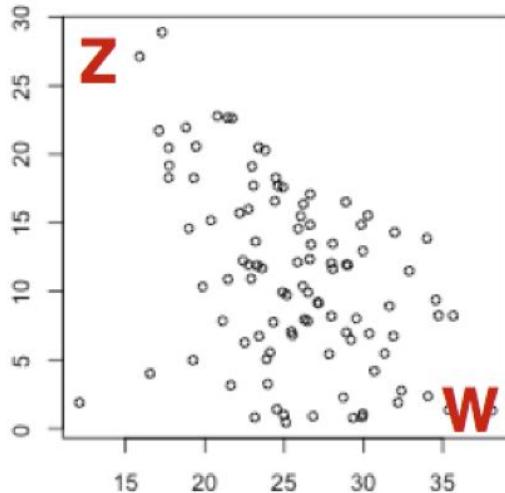
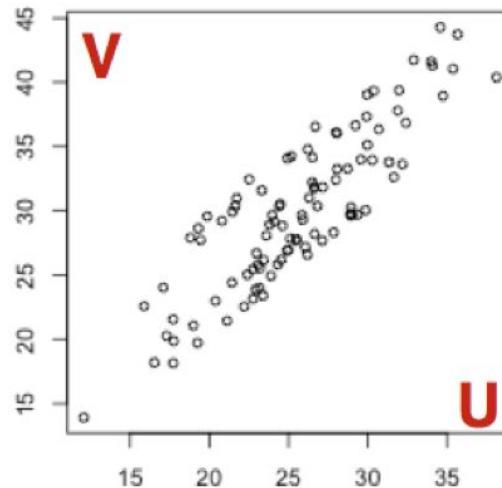
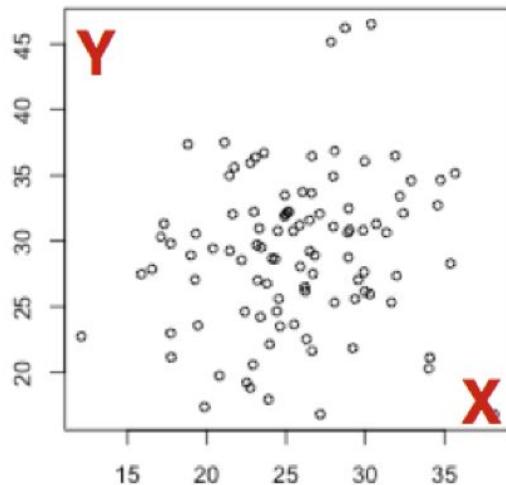
A gene expression dataset



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}

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

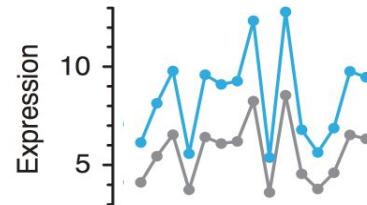
Correlation coefficient

Pearson Correlation Coefficient

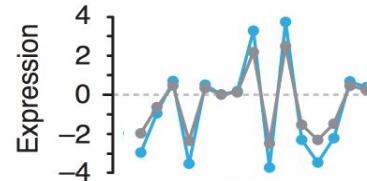
- Captures the relationship between 2 vectors after centering each vector by its mean and scaling by its standard deviation.
- The final quantities for each vector are called z-scores.

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

Profiles of absolute gene expression



Profiles of centered gene expression



Profiles of z-score of gene expression

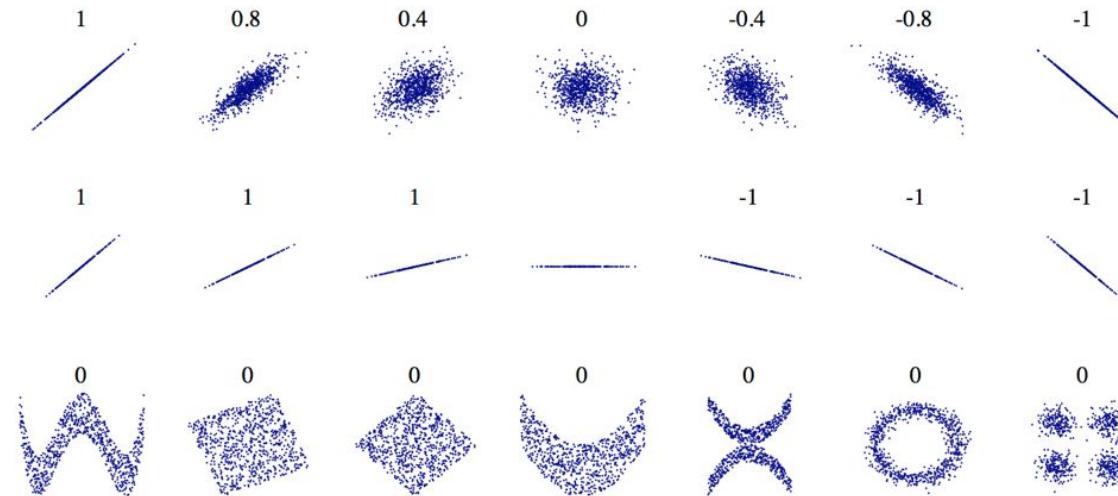


Correlation coefficient

Pearson Correlation Coefficient

- Measures linear relationship between variables.

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



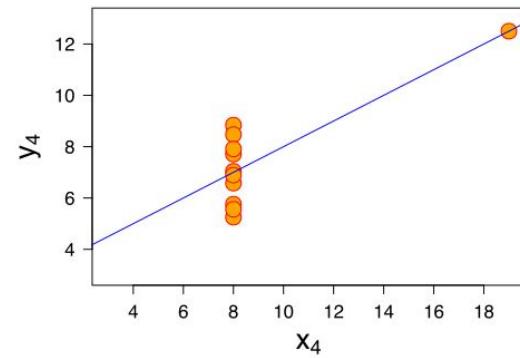
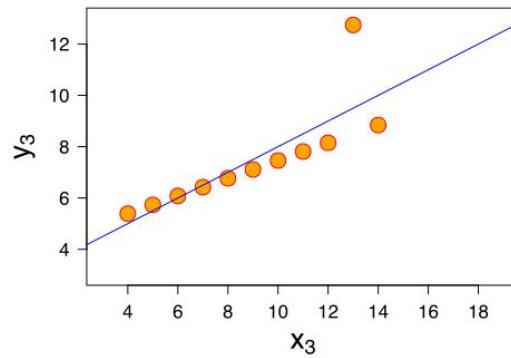
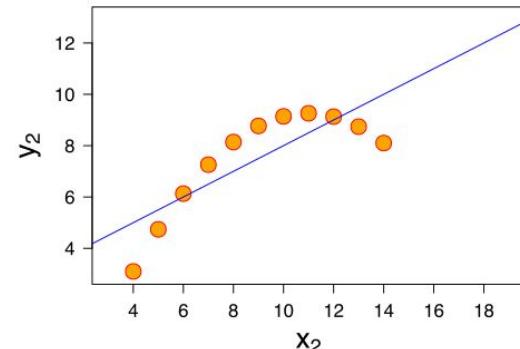
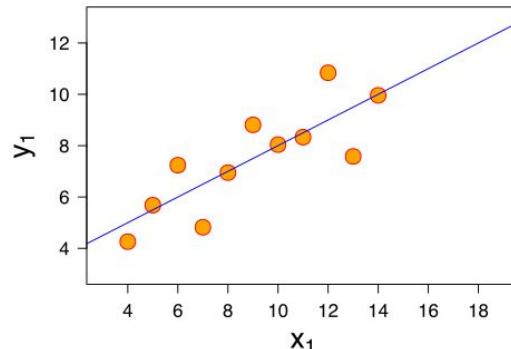
$$-1 \leq r \leq +1$$

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

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

11 data points

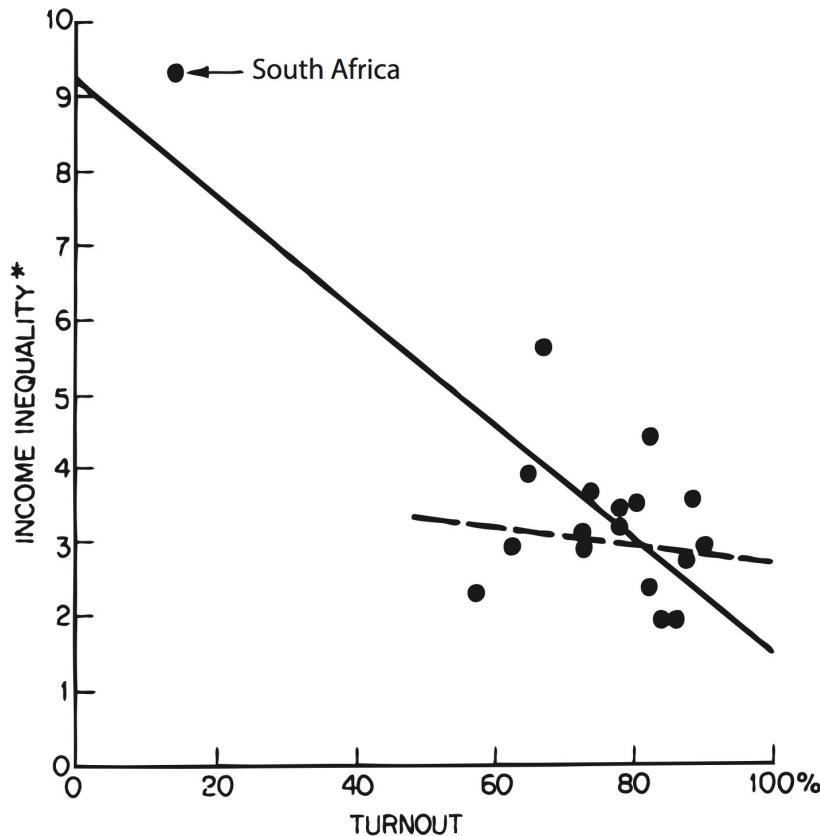
- Mean (x) = 9
- Var (x) = 11
- Mean (y) = 7.50
- Var (y) ~ 4.12
- Cor (x, y) = 0.816
- Linear regression line:
 - $y = 3.00 + 0.500x$



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

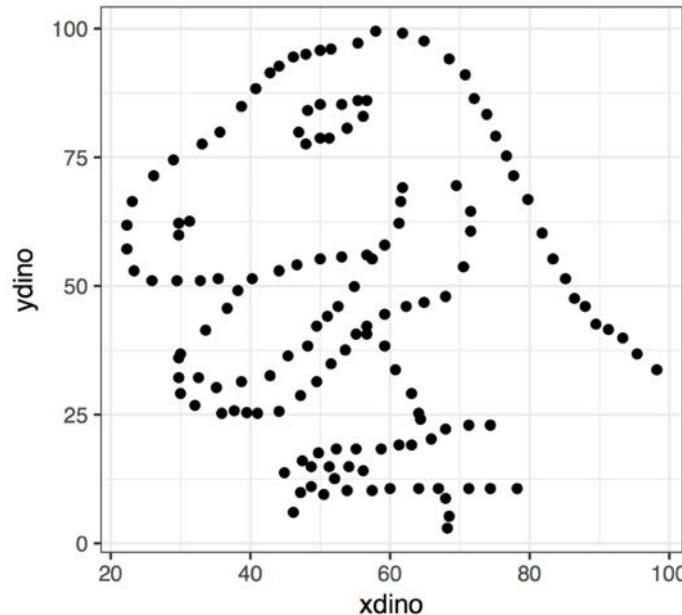
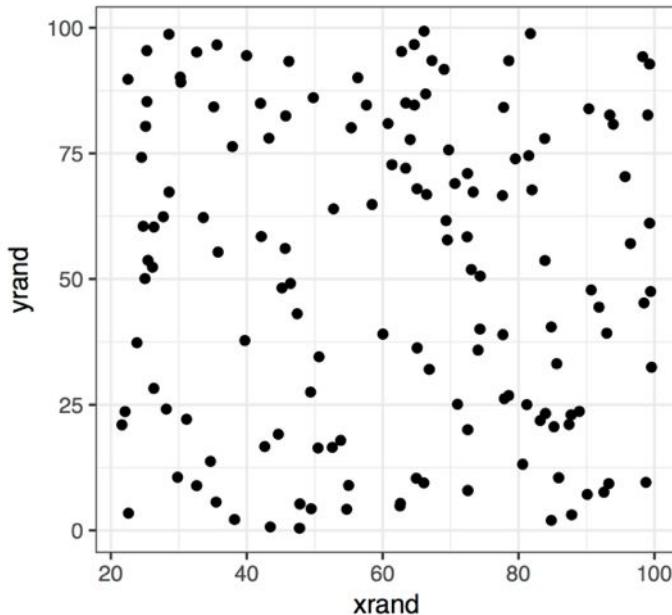
Wikipedia

What does a correlation coefficient tell you about the data?



What does a correlation coefficient tell you about the data?

Correlation = -0.06



What does a correlation coefficient tell you about the data?

Correlation = 0.7

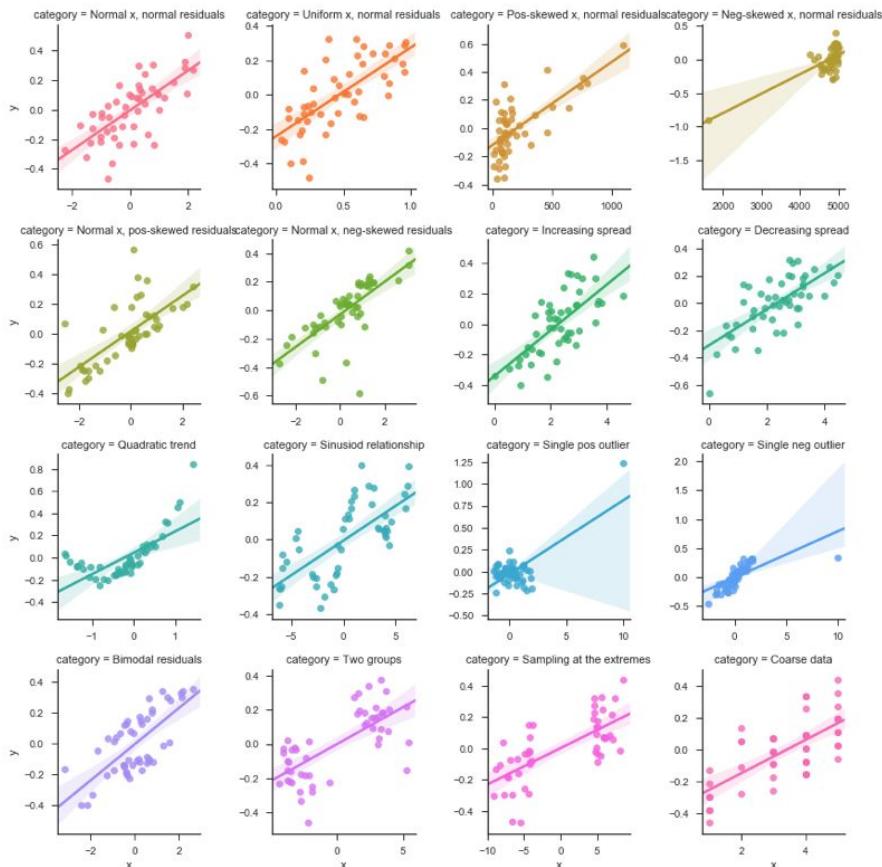
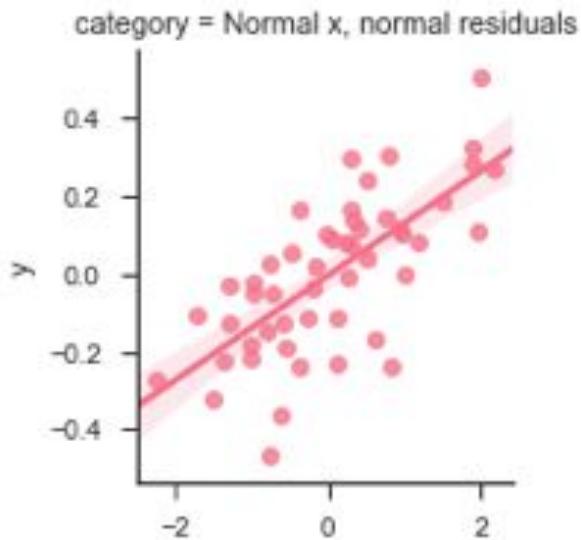
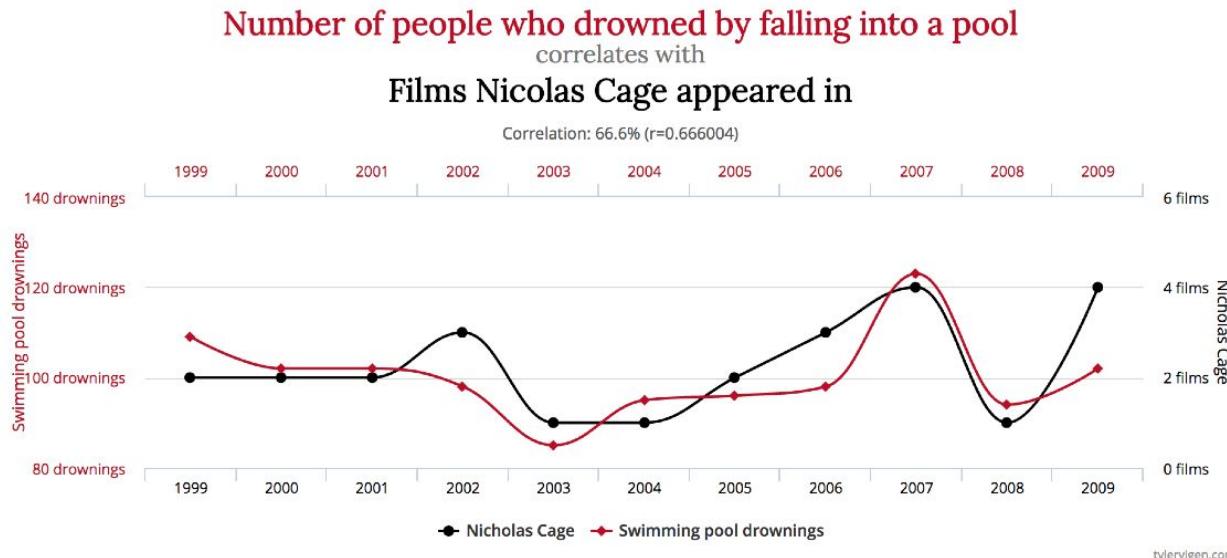


Figure adapted from code by Jan Vanhove: <https://ianhove.github.io/teaching/2016/11/21/what-correlations-look-like>

Spurious correlations

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



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

tylervigen.com

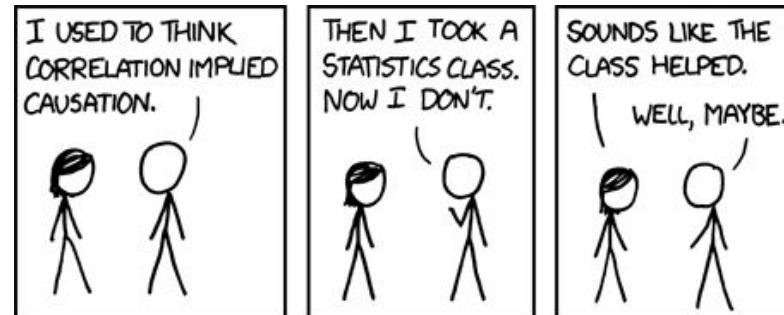
Checkout <https://www.google.com/trends/correlate>

tylervigen.com

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.



Many distance measures

Euclidean Distance

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

Pearson Correlation Coefficient

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

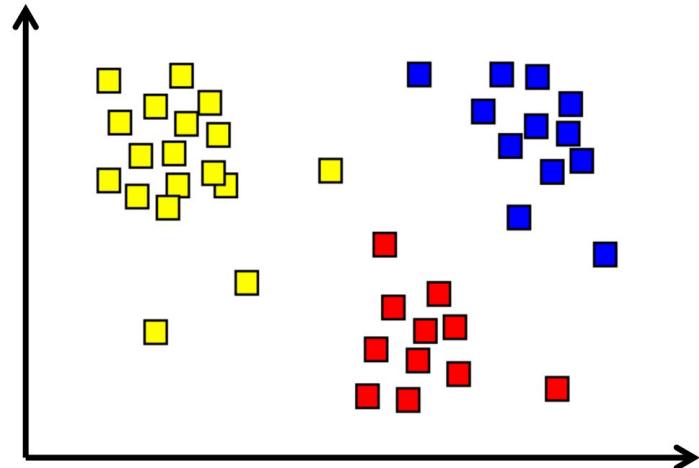
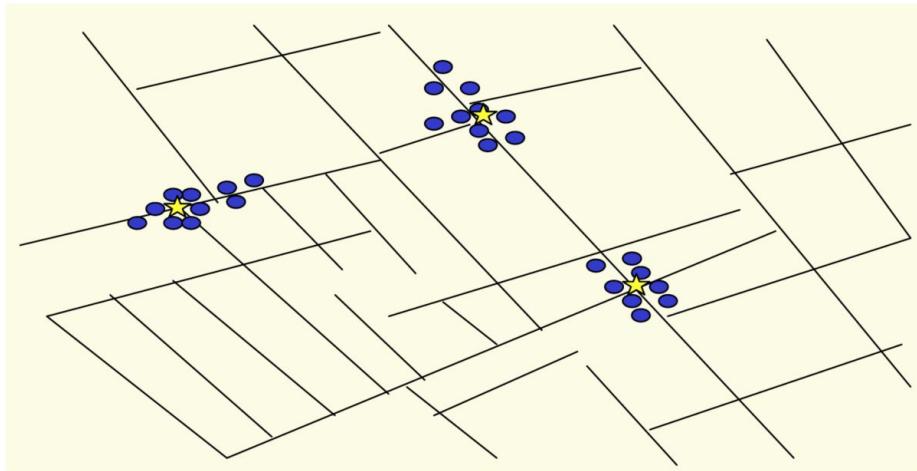
Spearman Rank Correlation

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

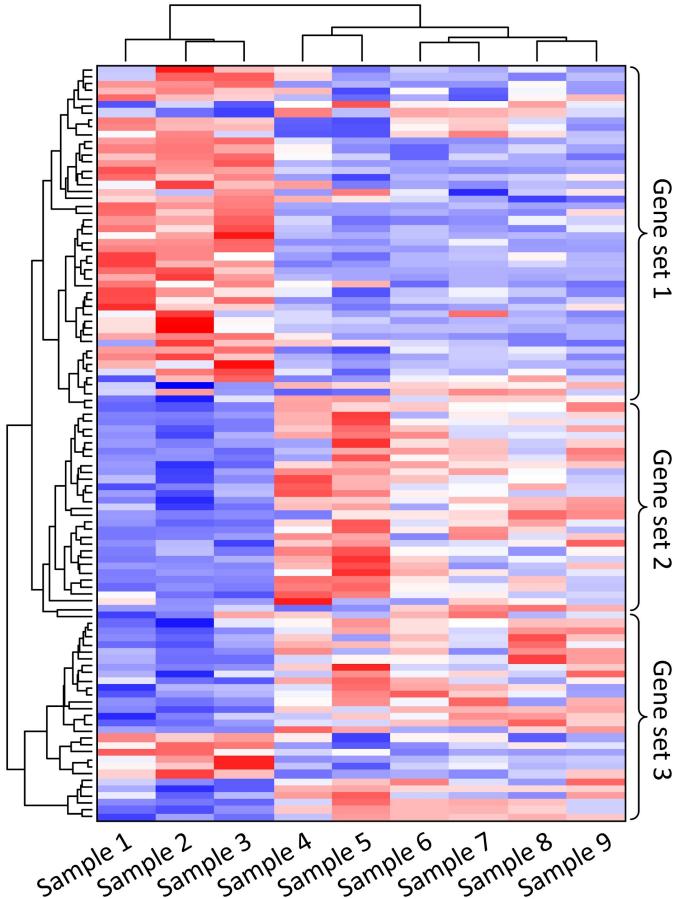
Mutual Information

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

Clustering



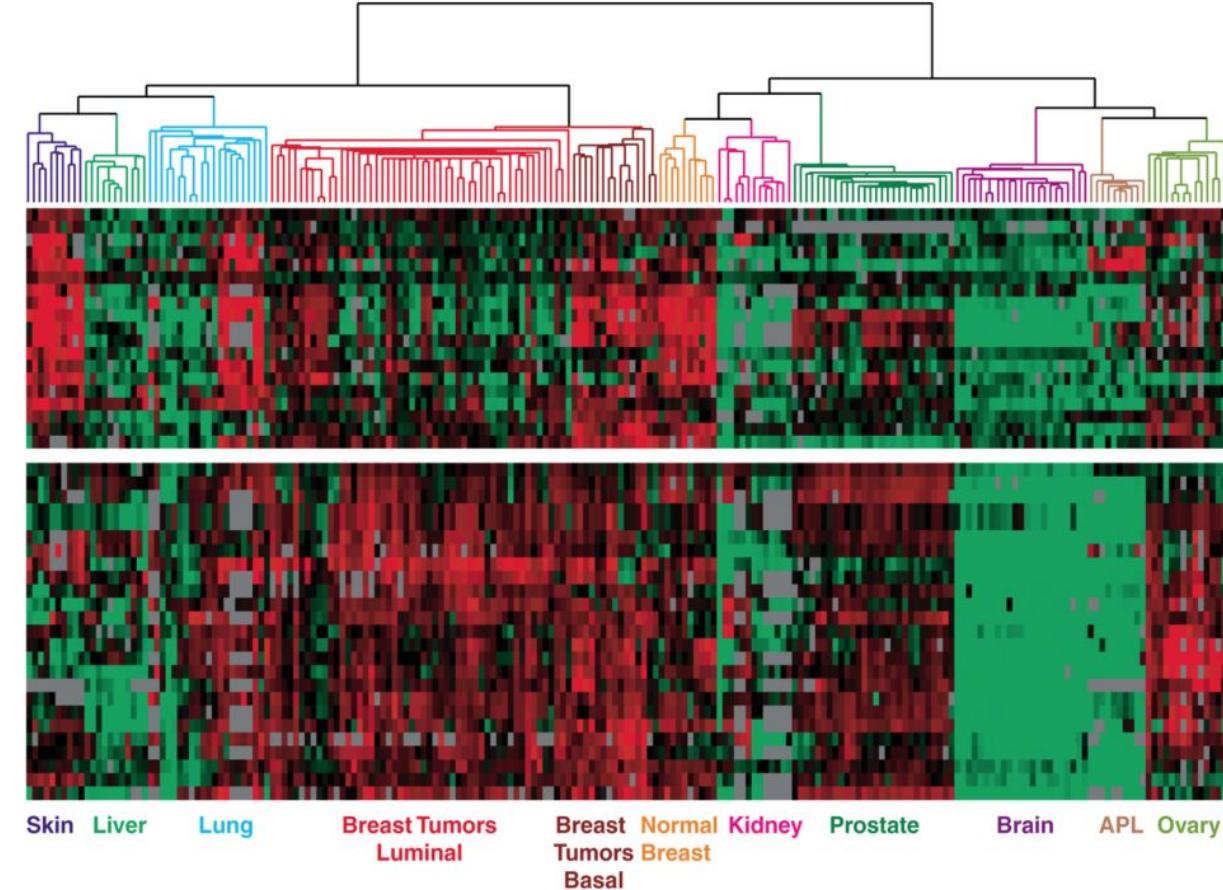
Clustering gene-expression profiles



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?

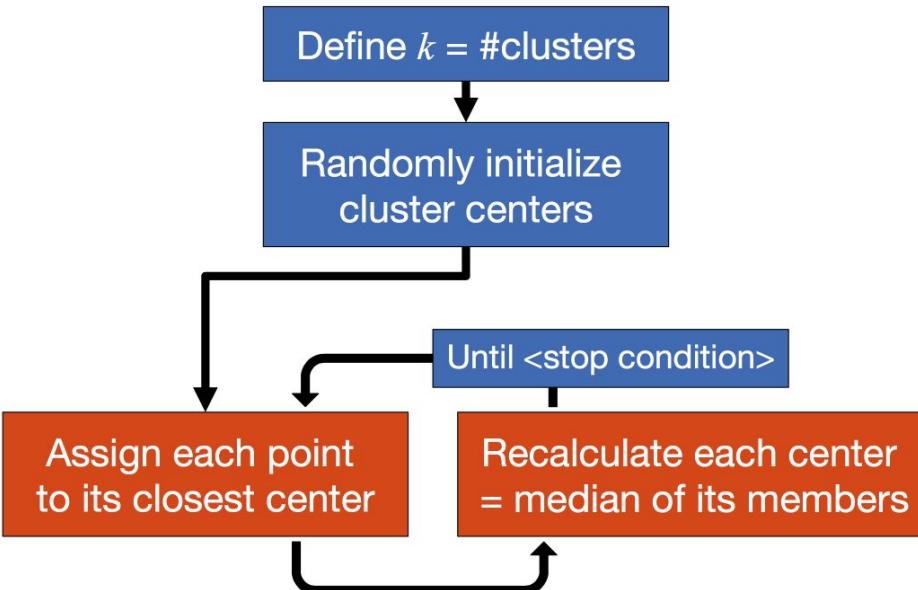
Clustering gene-expression profiles



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?

K-means clustering

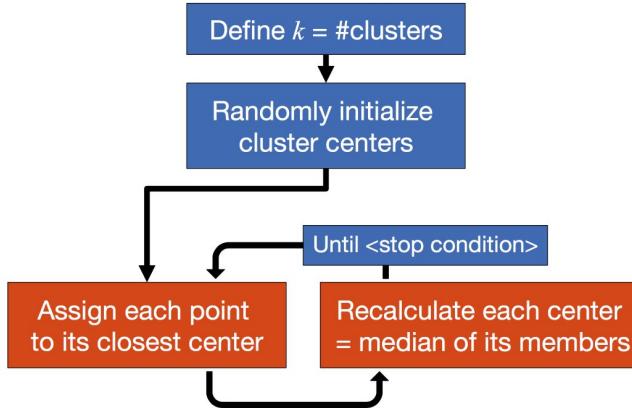


Conceptually similar to Expectation-Maximization, alternating between 2 two steps:

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.

K-means clustering



Stopping condition

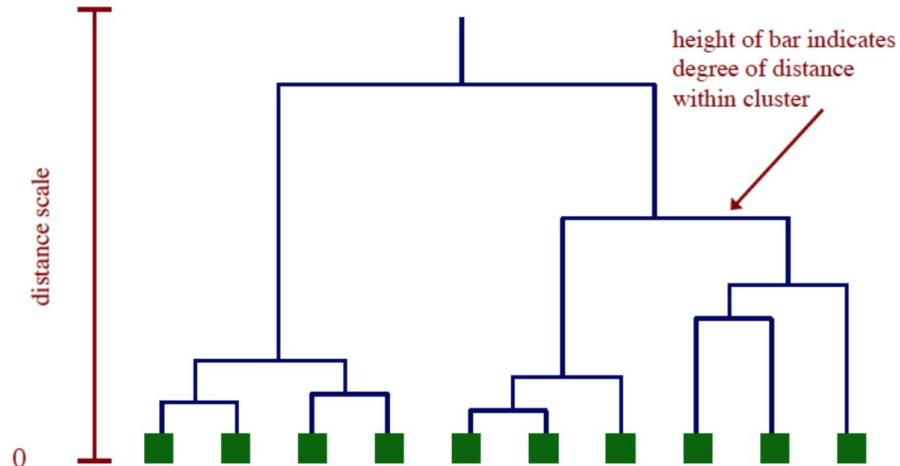
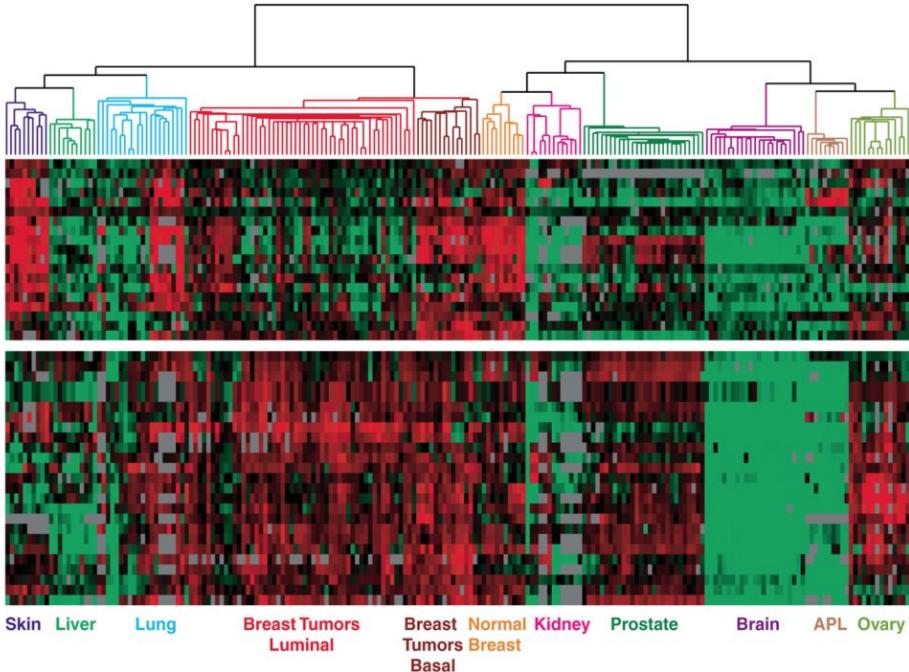
- Until the change in centers is less than <constant>.
- Until all genes get assigned to the same partition.
- Until some minimal number of genes (e.g. 90%) get assigned to the same partition twice in a row.

Some issues

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

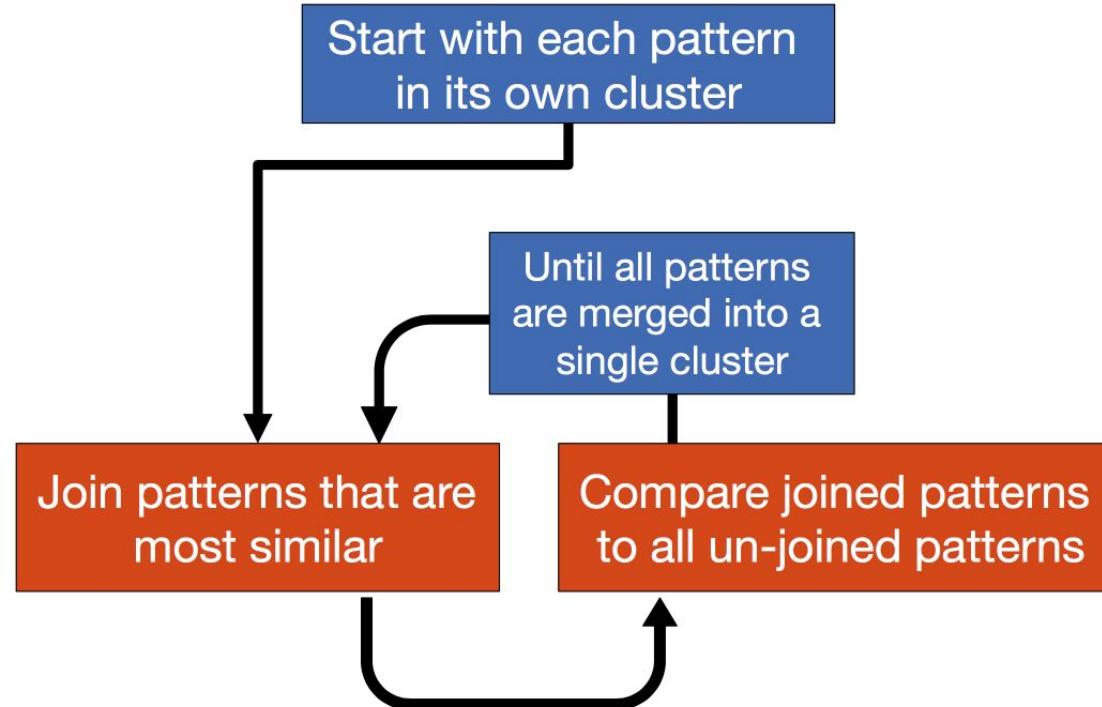
- Genes assigned to clusters on the basis of all experiments;
Experiments assigned to clusters based on all genes.

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

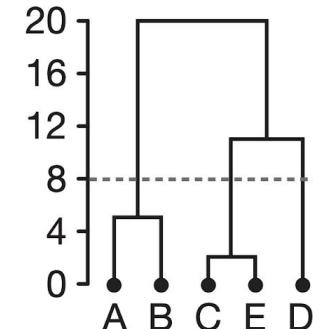
Linkage criteria:

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

Complete linkage clustering of 5 objects

		1					2					3					4				
		A	B	C	D	E	A		B	CE	D	AB		CE	D	AB		CED	AB		CED
A	0						A	0				AB	0			AB	0		AB	0	
B	5	0					B	5	0			CE	20	0		CE	20	0	CED	20	0
C	10	3	0				CE	20	8	0		D	15	11	0	D	15	11	0		
D	15	6	7	0			D	15	6	11	0										
E	20	8	2	11	0																

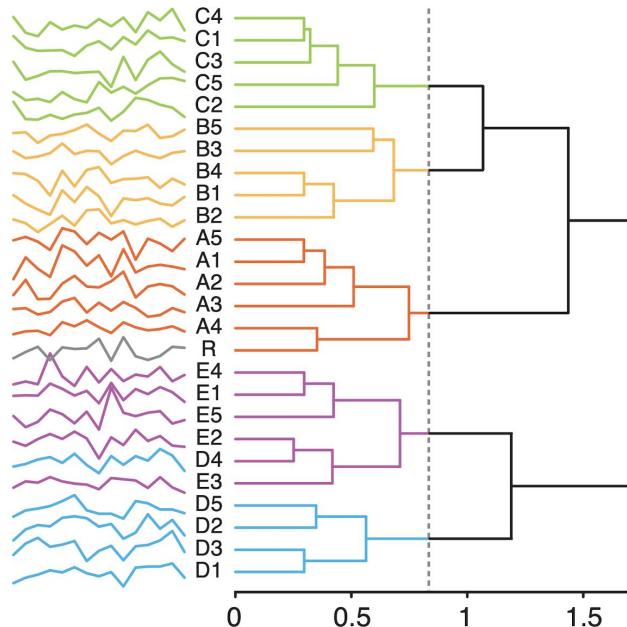
Dendrogram



Hierarchical clustering

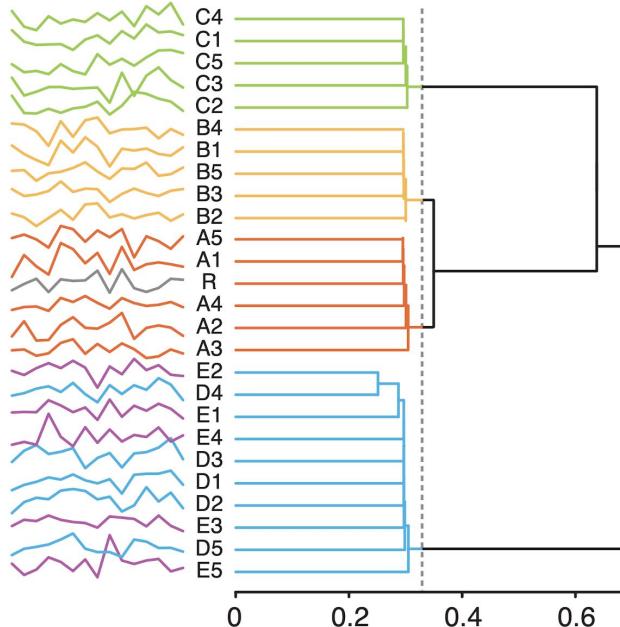
Complete linkage clustering

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

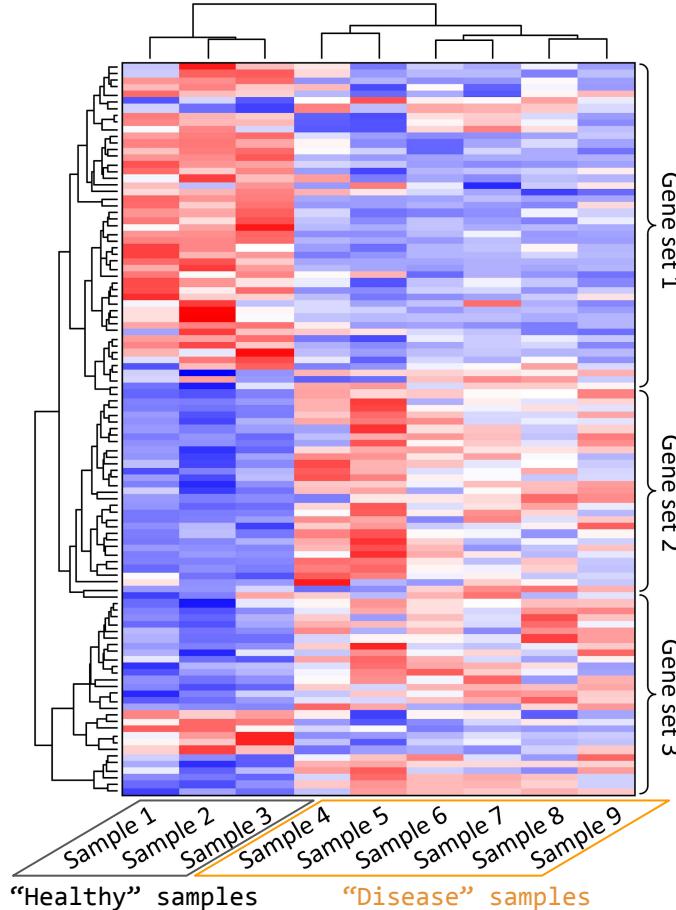


Single linkage clustering

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



A gene expression dataset



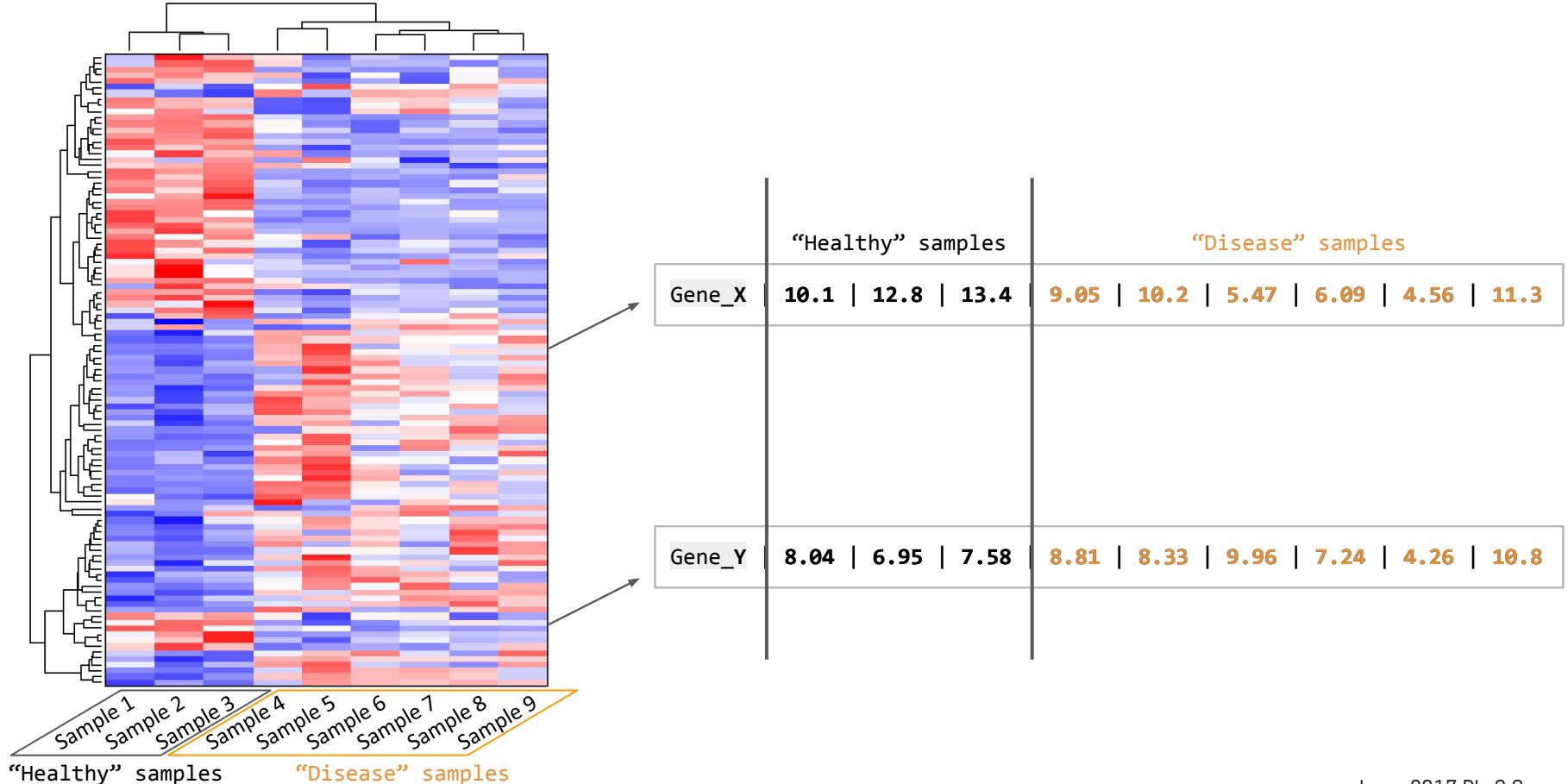
Gene-level Qs:

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?

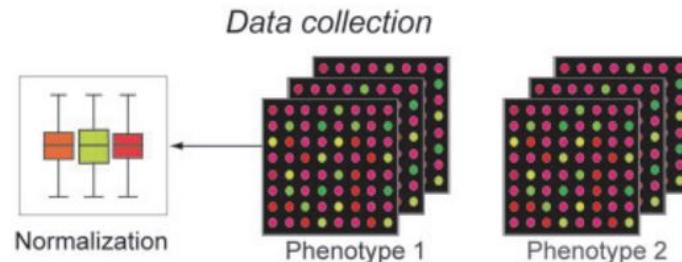
Gene-level analysis



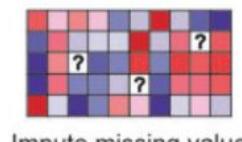
Differential expression & Enrichment analysis

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

Enrichment analysis



1. Data preprocessing



2. Gene level statistics

$T \rightarrow |T|$
-2.16 → 2.16
Take absolute

Probe set ID	Gene	T score
1424992_at.A	Tprkb	2.16
1425410_at.A	Tprkb	1.35
1416034_at.A	Cd24a	0.15

Map probe sets to genes

3. Gene set-level statistics

$$\sum rank(T_i)$$

Statistic selection

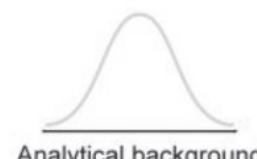
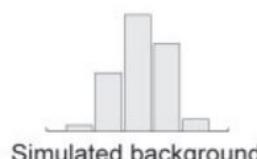


Pathway topology

$$\mu + \beta_1 * \text{Weight} + \beta_2 * \text{Age} + \epsilon$$

Multiple covariates

4. Significance measurement



5. Multiple testing correction



Correction for multiple testing