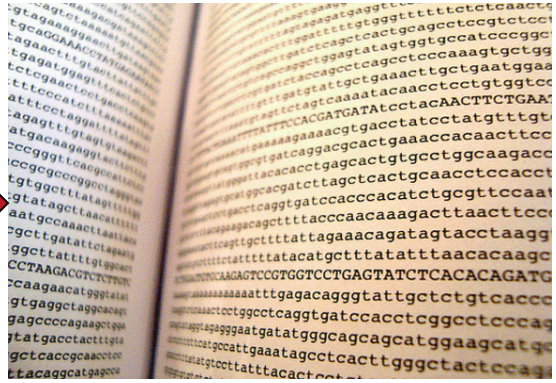
Clustering

With application to gene-expression profiling technology

Thanks to Kevin Wayne, Matt Hibbs, & SMD for a few of the slides

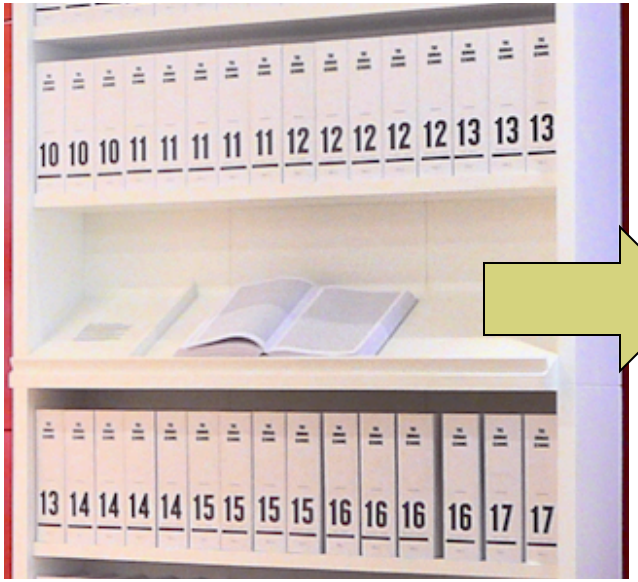
Why is expression important?



Understanding cellular and human biology

Understanding civil life and sociology

Why is expression important?



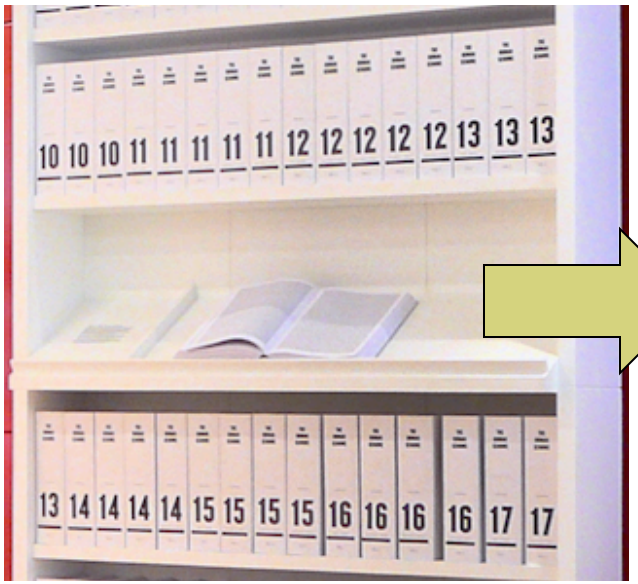
?

Understanding cellular and human biology

?

Understanding civil life and sociology

Why is expression important?



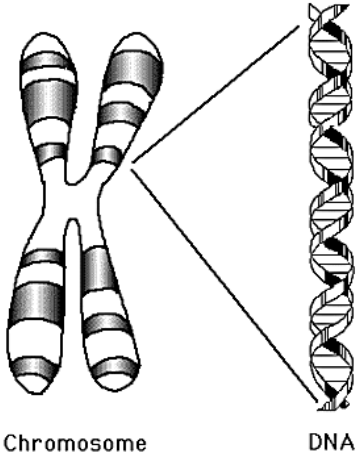
Measure the activity of genes in various cellular conditions

Understanding cellular and human biology

Measure the activity of people in various societal conditions

Understanding civil life and sociology

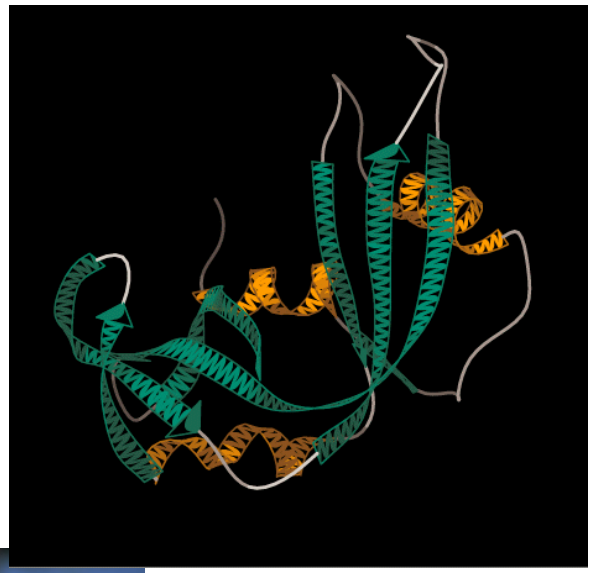
Why is expression important?



DNA

Gene Expression

Proteins



Phenotypes

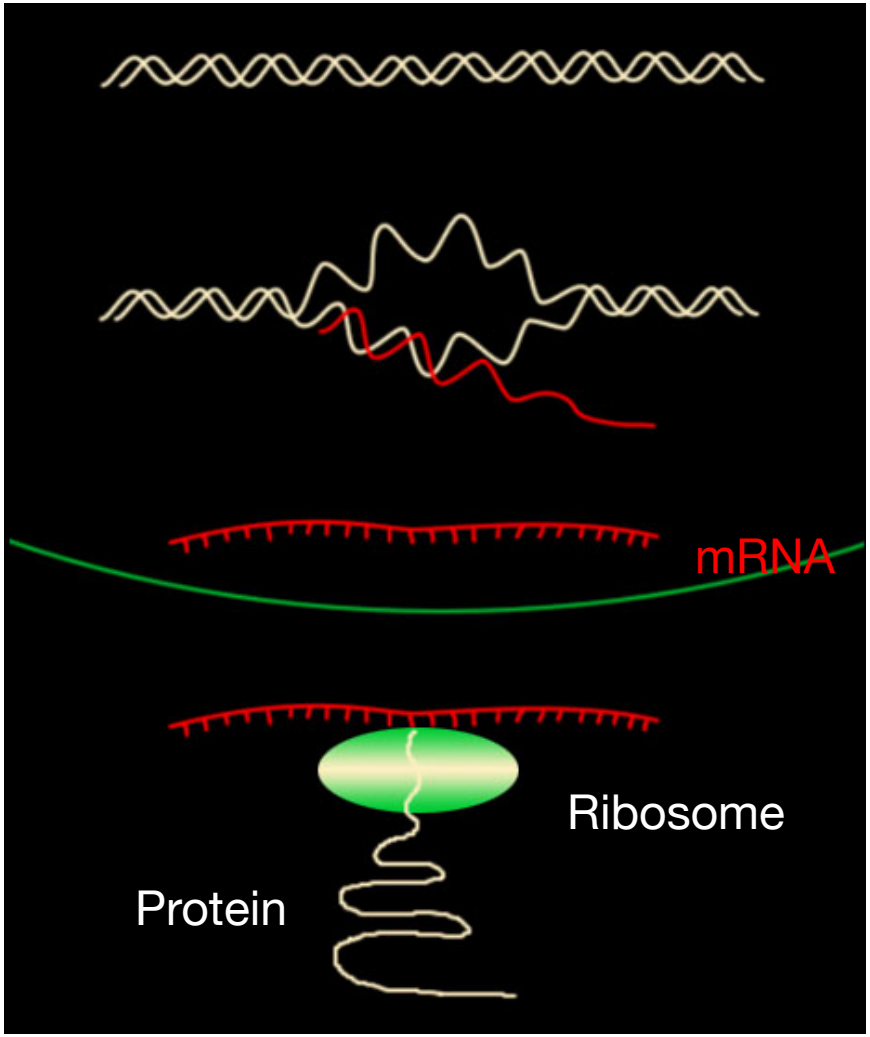
From Genes to Proteins

Transcription:

DNA to mRNA

Translation:

mRNA to Proteins



Proteins

Proteins are the "workhorses" of cells

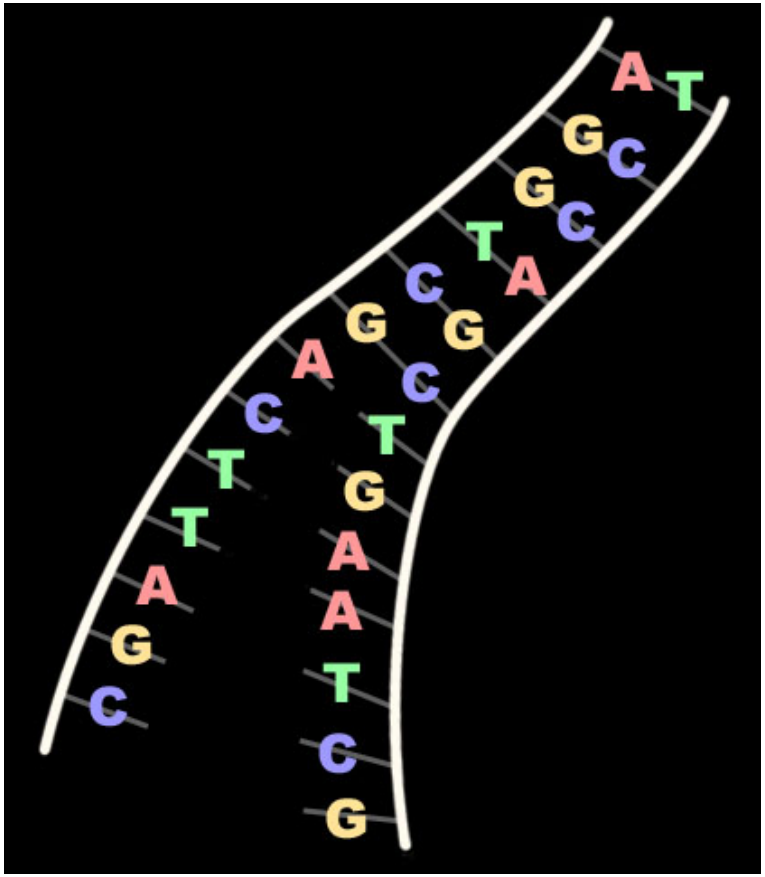
- To understand how cells work is to understand proteins

Understanding proteins and cells is key for finding disease treatments and cures

- Modern drug development is centered on affecting proteins (receptors, hormones, etc.)

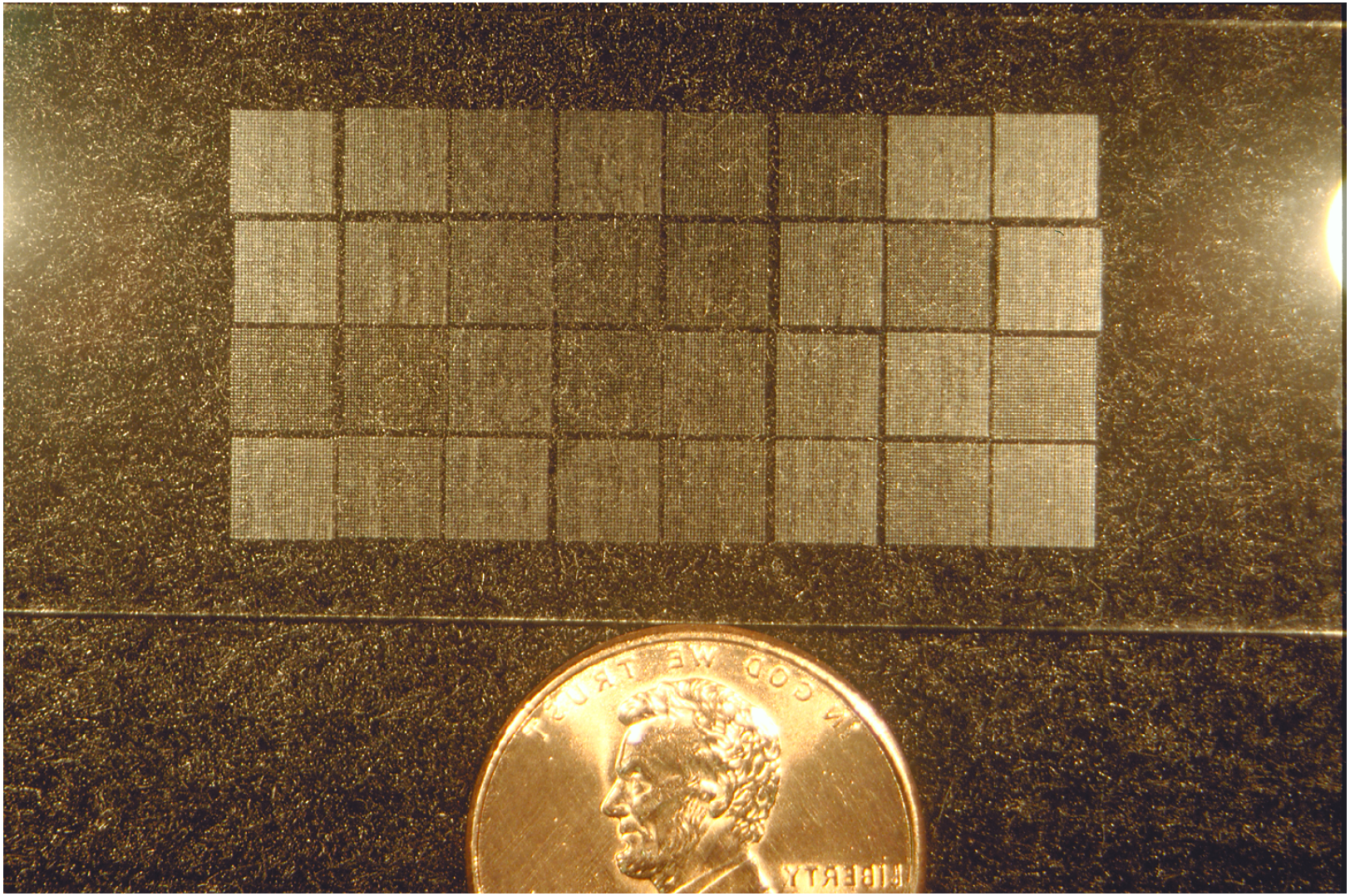
But… Proteins are hard to study directly, so microarrays look at the mRNA instead.

Hybridization

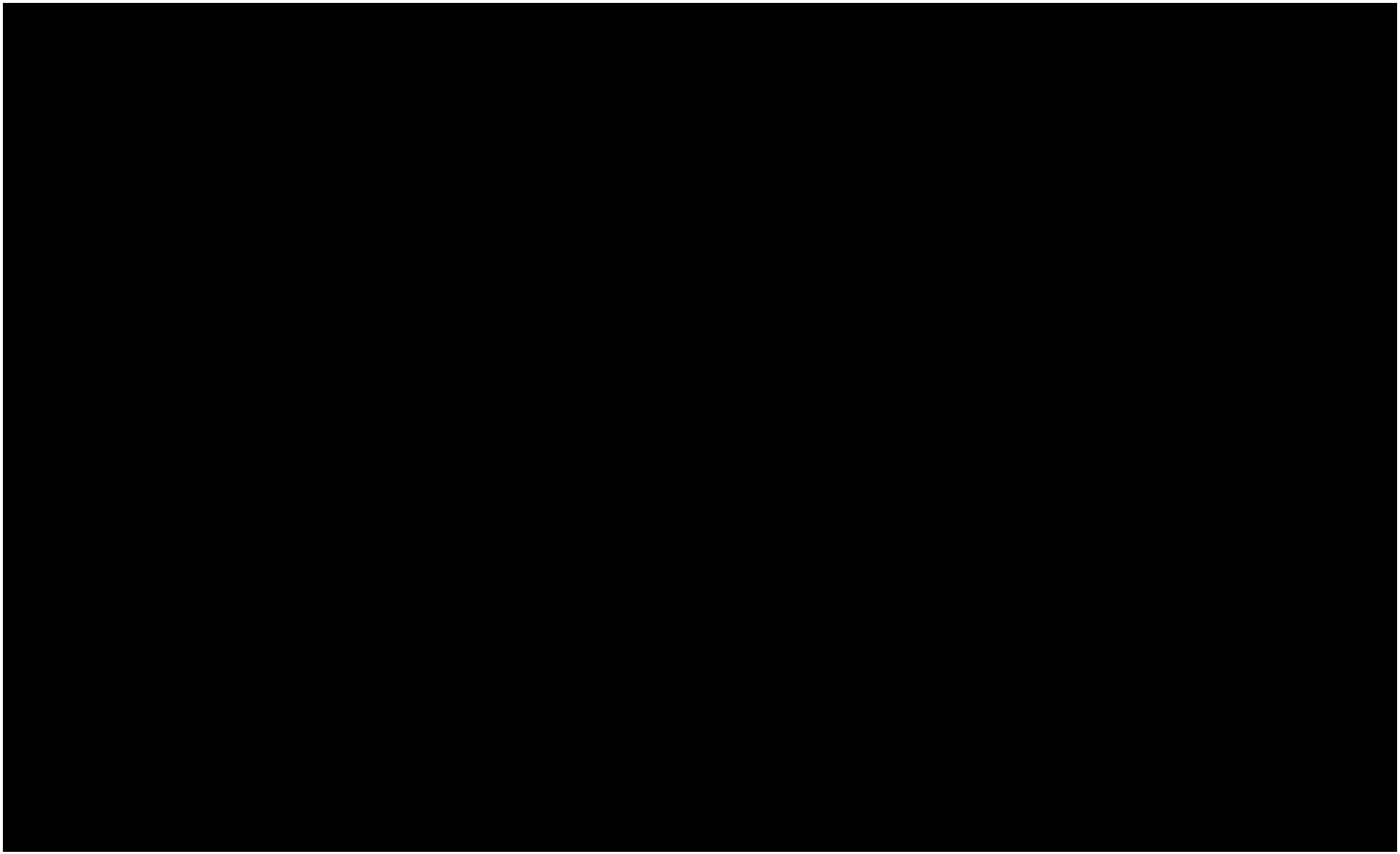


Expression microarrays use the fact that complementary strands will hybridize (attach) to each other

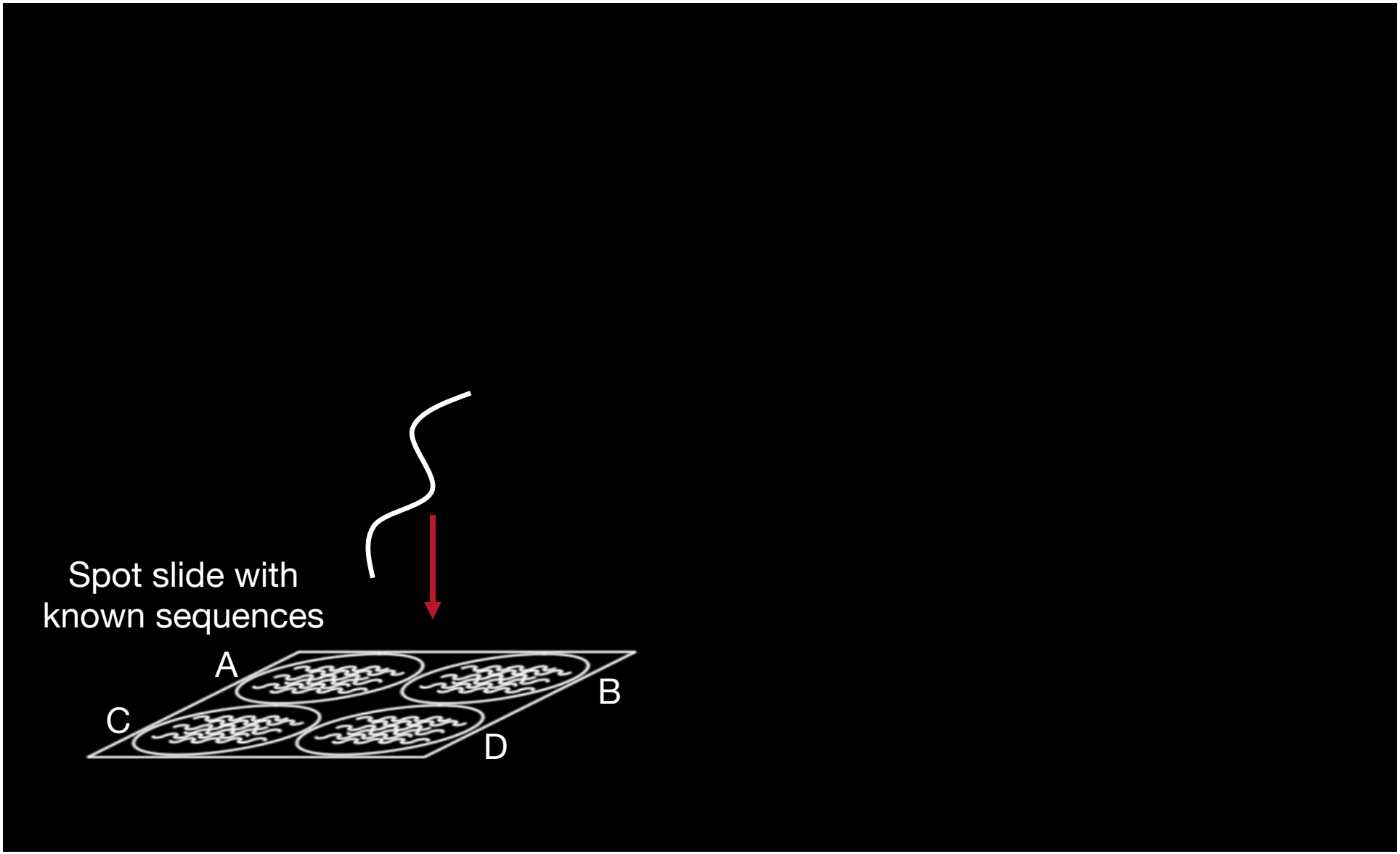
Early cDNA microarray (18,000 clones)



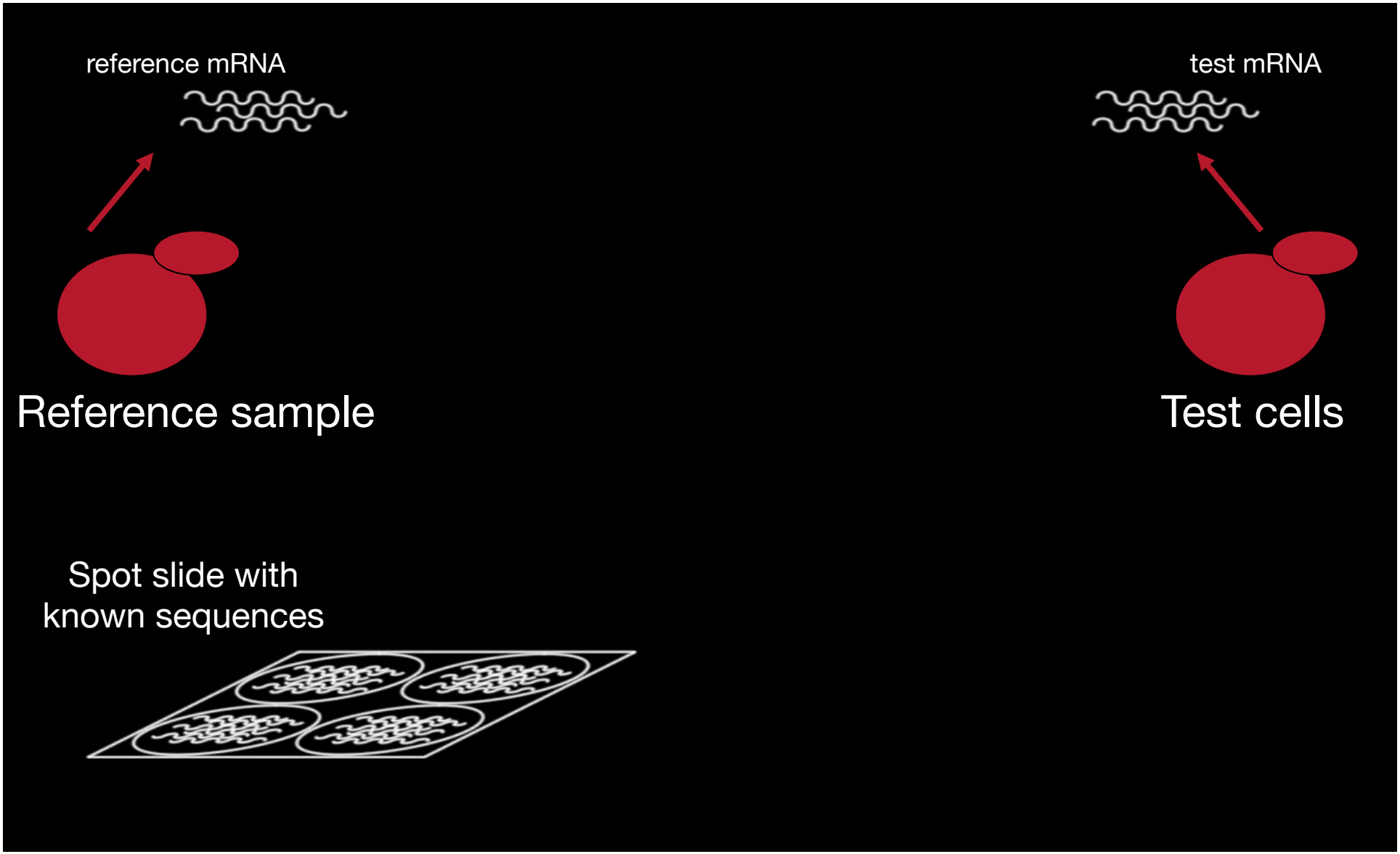
Microarray Methodology



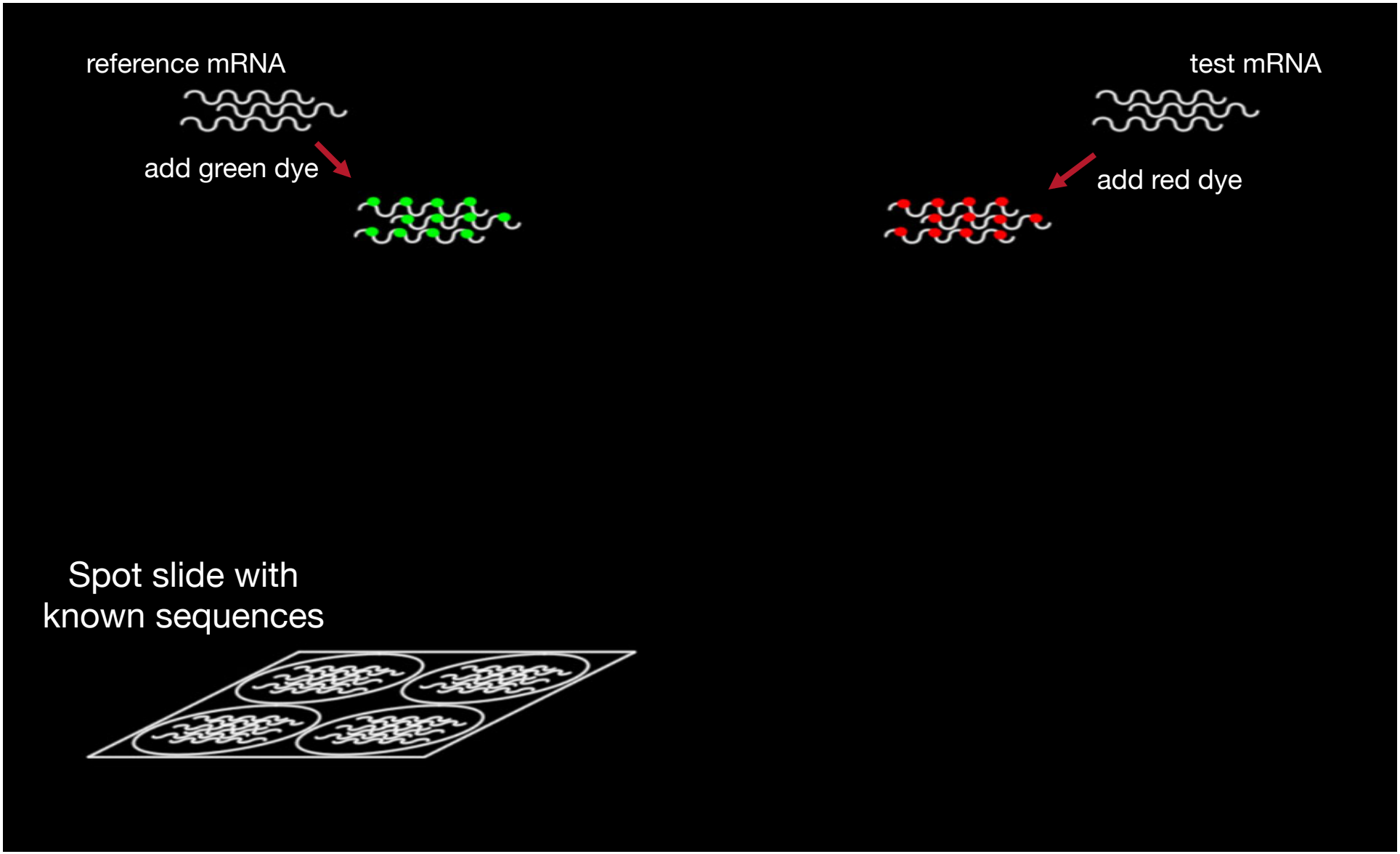
Microarray Methodology



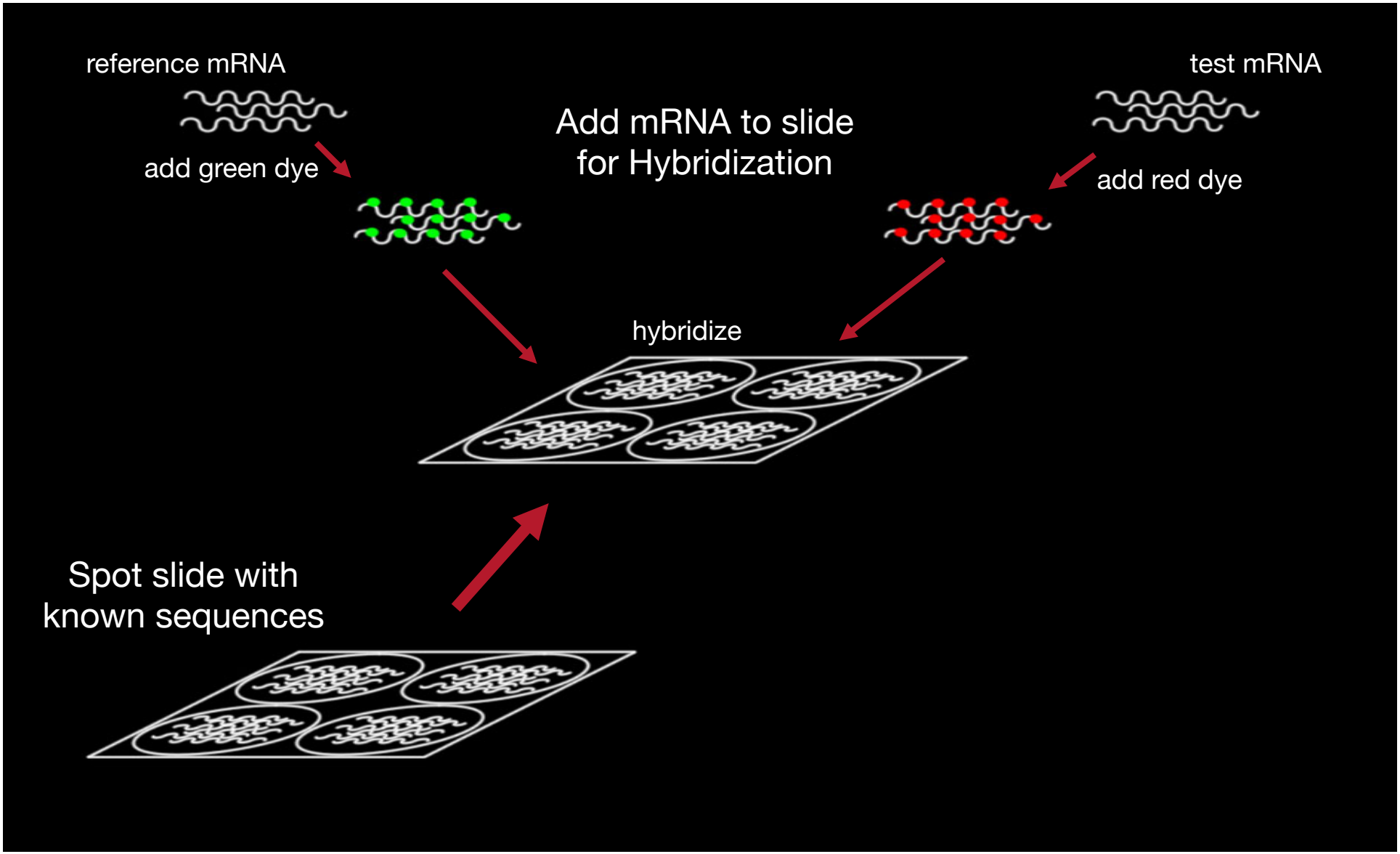
Microarray Methodology



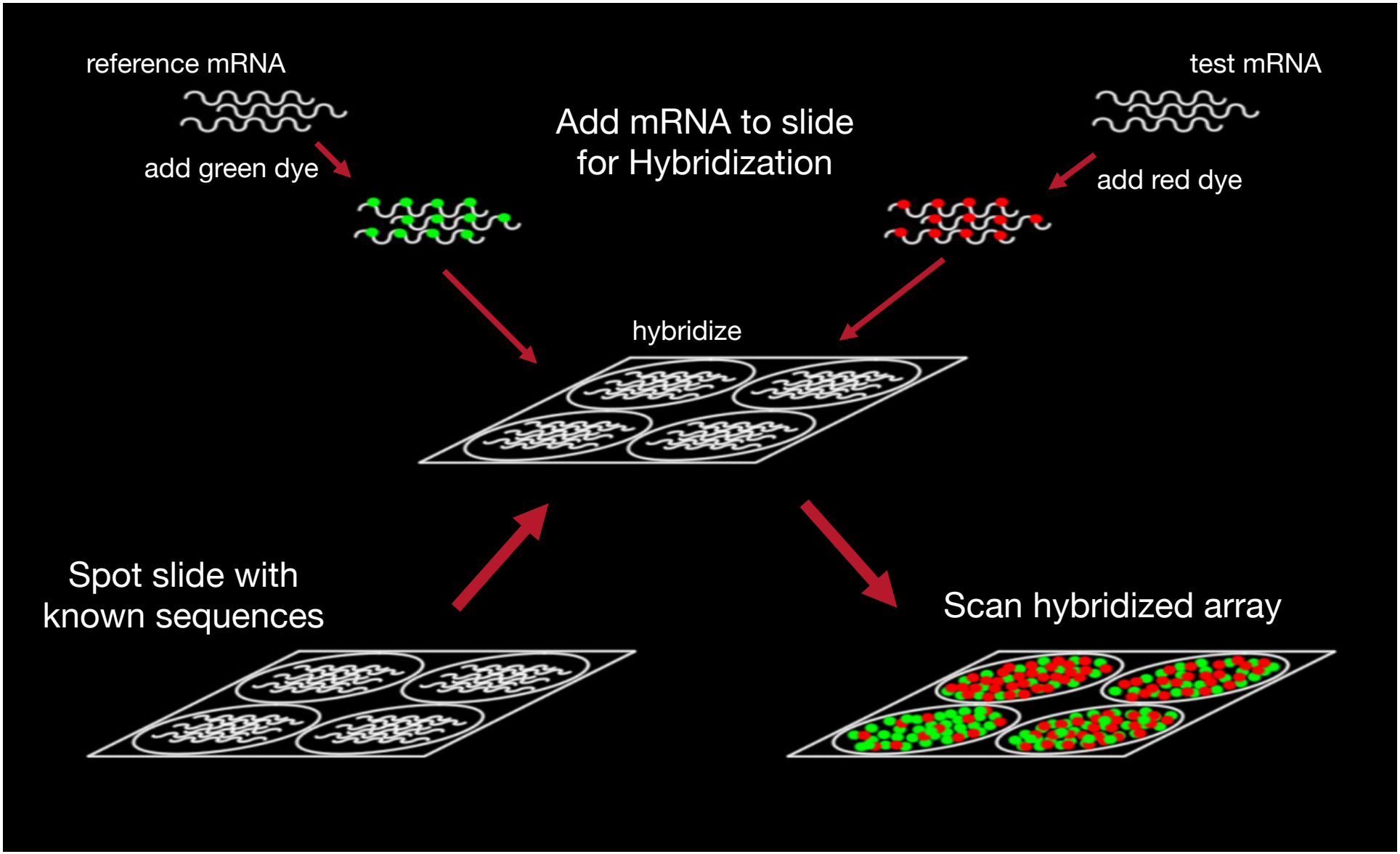
Microarray Methodology



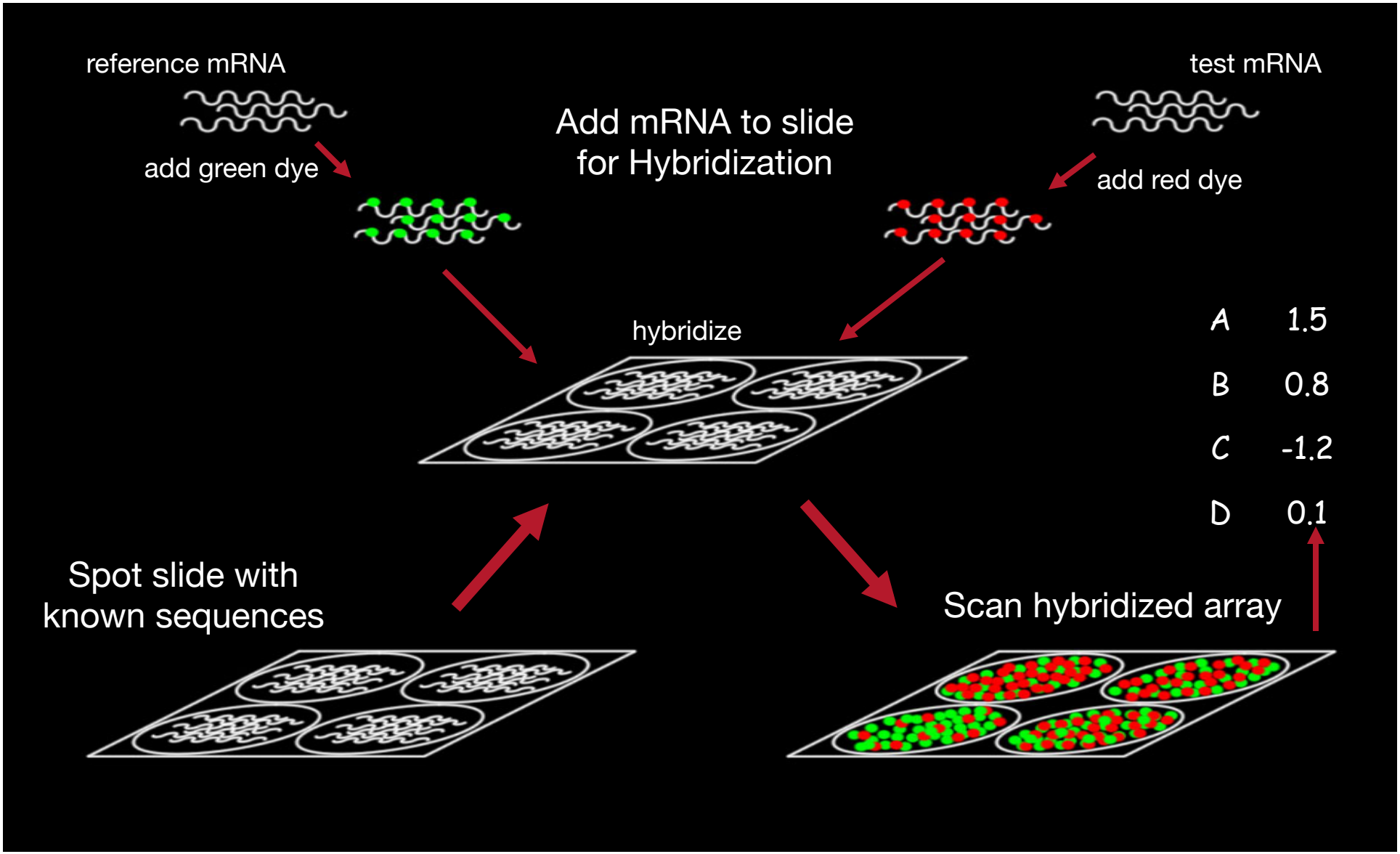
Microarray Methodology



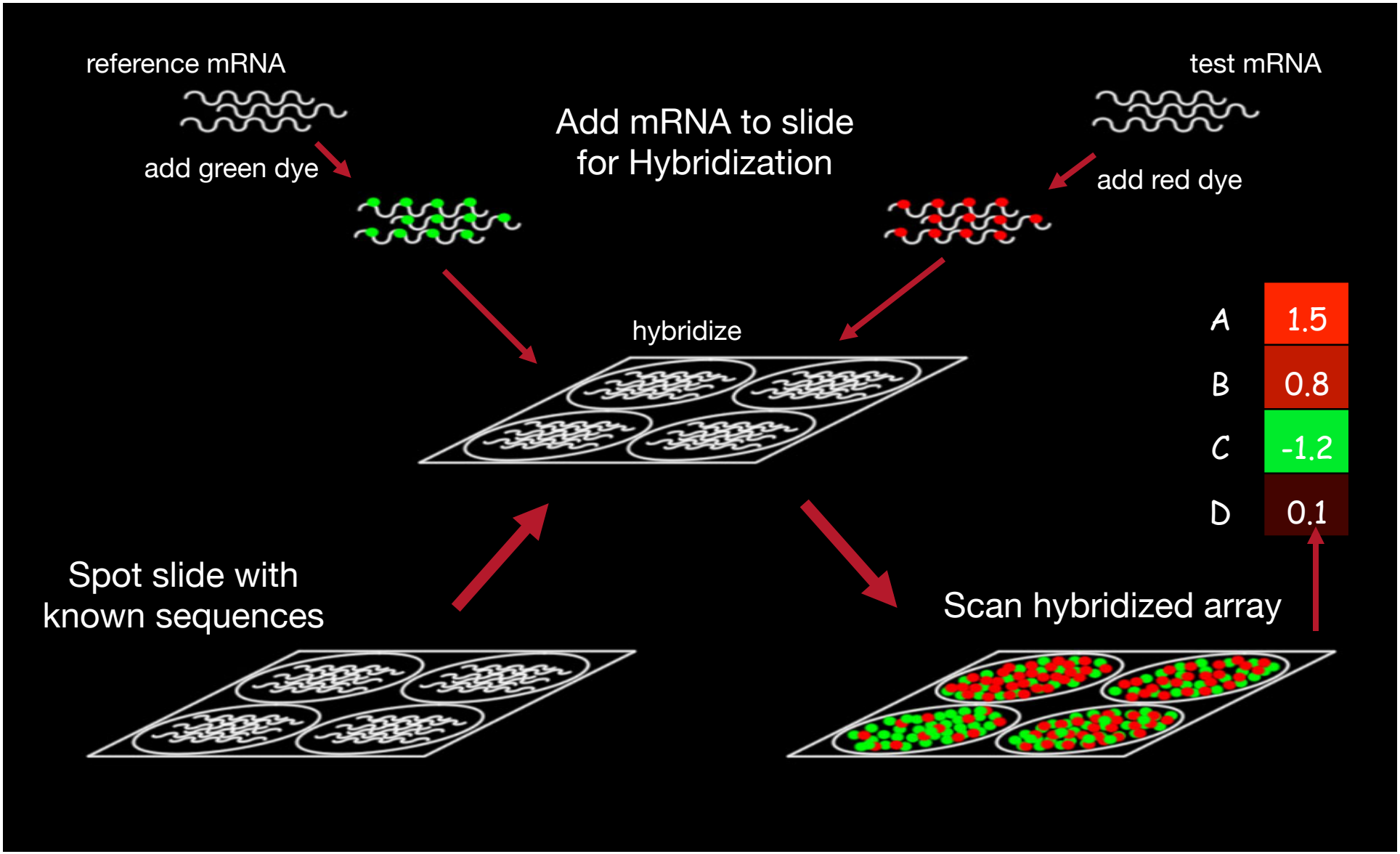
Microarray Methodology



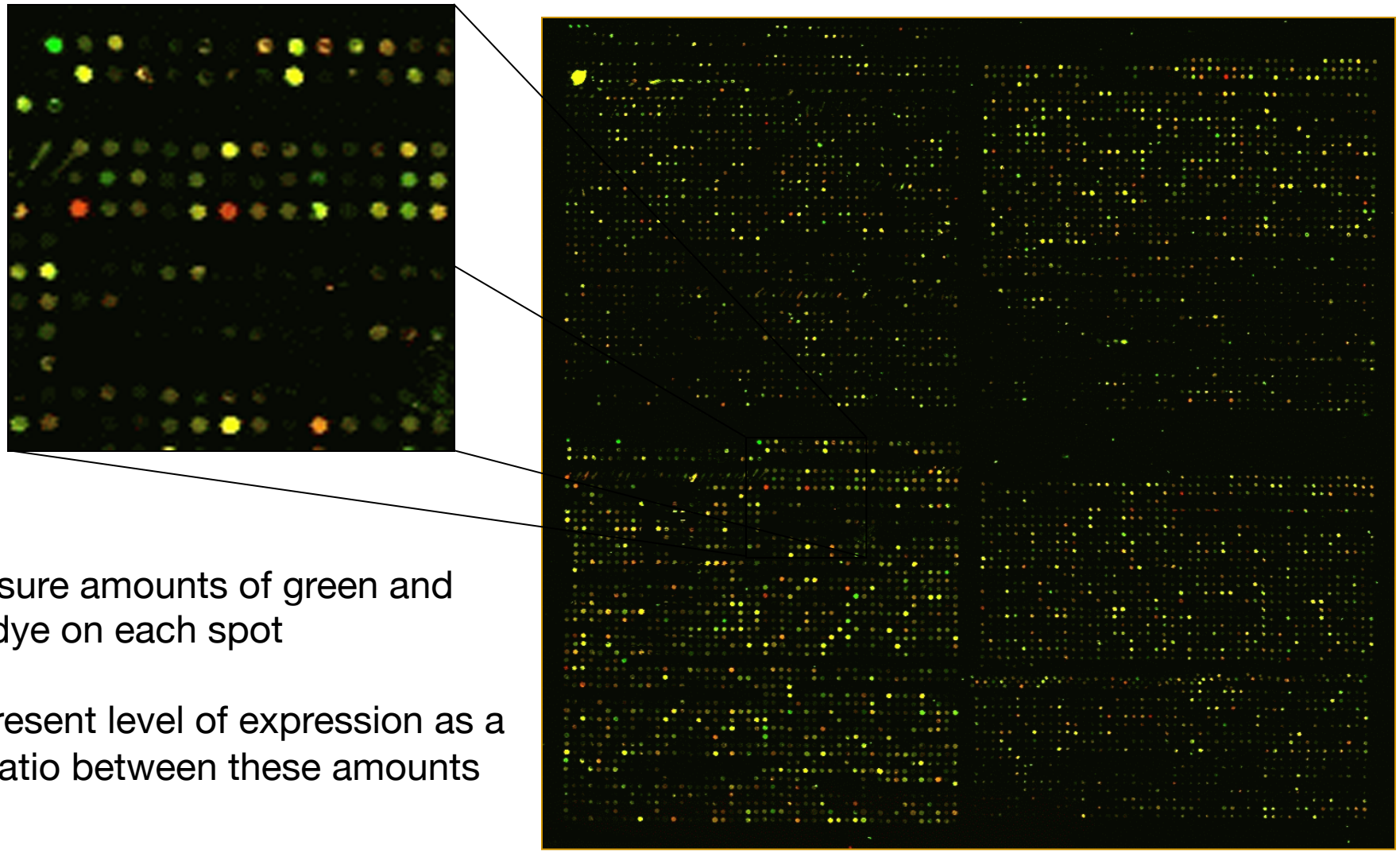
Microarray Methodology



Microarray Methodology



Microarray Outputs

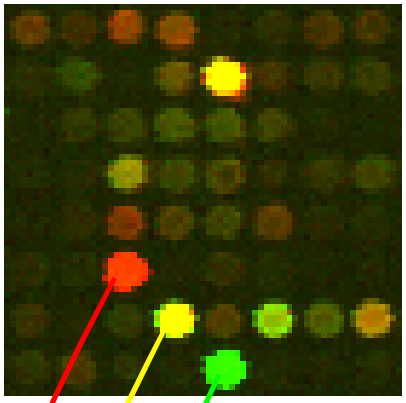


Measure amounts of green and red dye on each spot

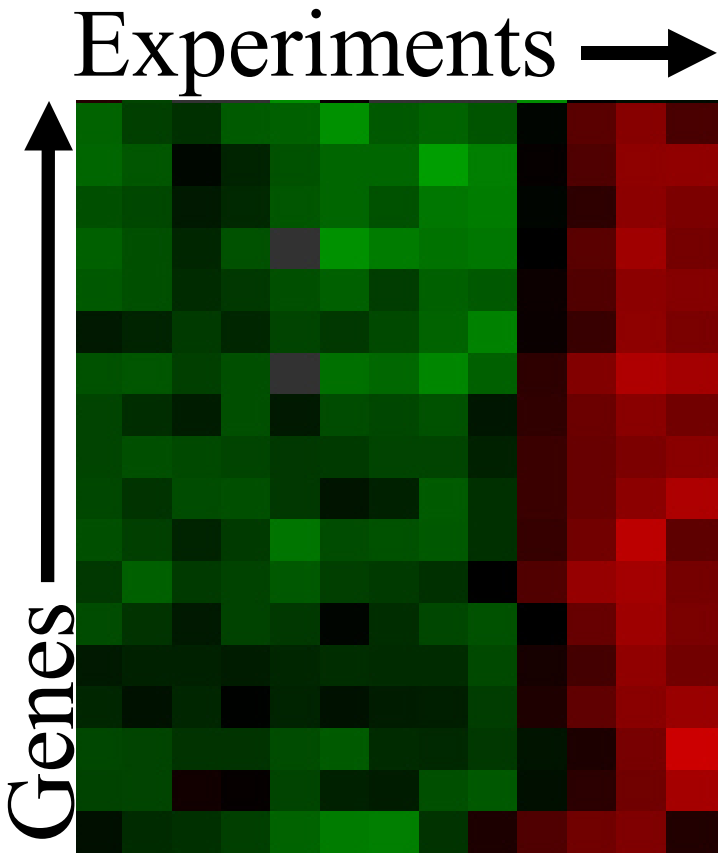
Represent level of expression as a log ratio between these amounts

Raw Image from Spellman et al., 98

Extracting Data



Cy3	Cy5	$\frac{Cy5}{Cy3}$	$\log_2\left(\frac{Cy5}{Cy3}\right)$
200	10000	50.00	5.64
4800	4800	1.00	0.00
9000	300	0.03	-4.91

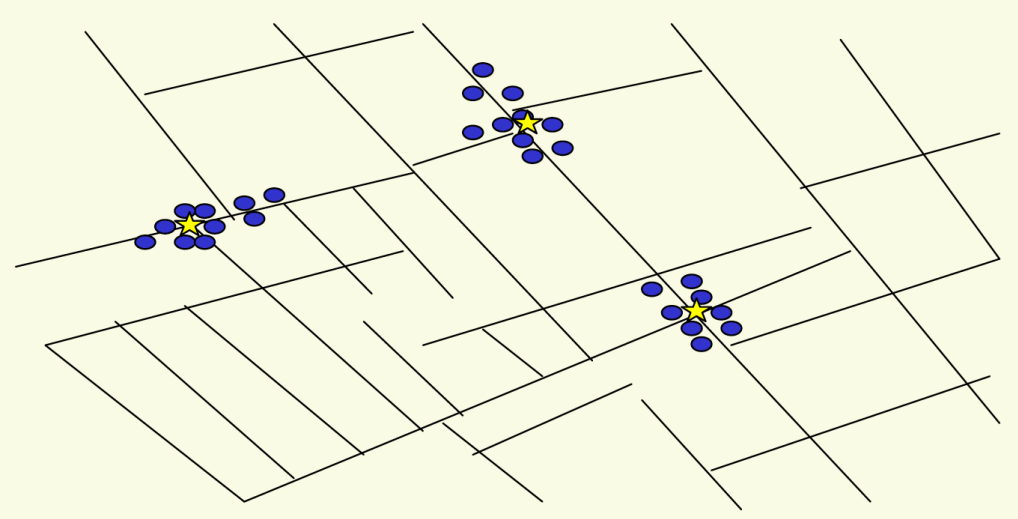


Some questions you can tackle with high-throughput gene-expression

Large-scale study of biological processes

- What is going on in the cell at a certain point in time?
  - what genes/pathways are active?
- On a genomic level, what accounts for differences between phenotypes?
  - which genes/pathways are activated in stress response?

Clustering

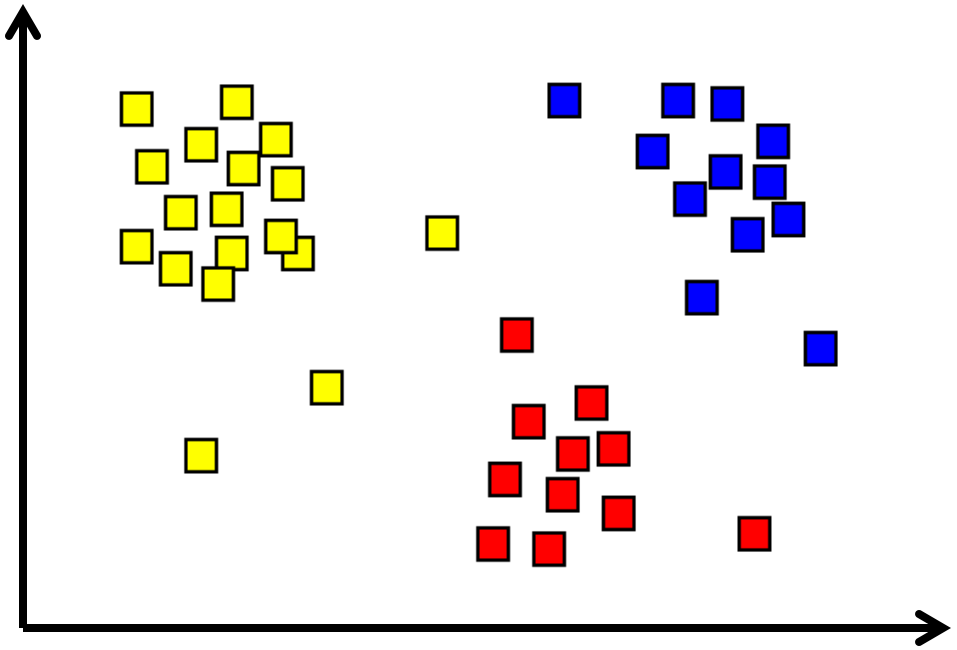


Outbreak of cholera deaths on map in 1850s.
Reference: Nina Mishra, HP Labs

History: London physicist John Snow plotted outbreak of cholera deaths on map in 1850s. Location indicated that clusters were around certain intersections with polluted wells; this exposed the problem and solution!

What is clustering?

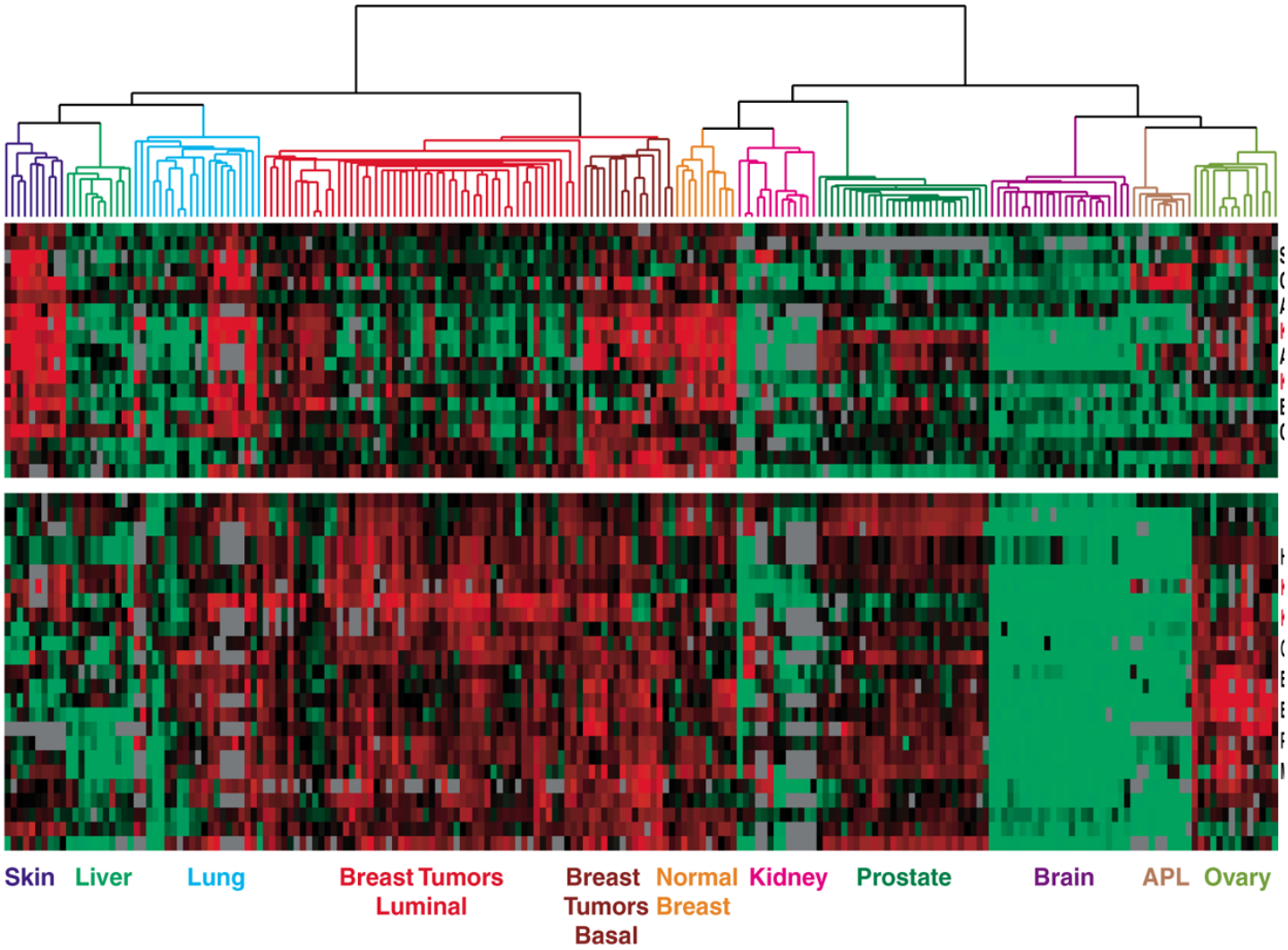
Reordering of vectors in a dataset so that similar patterns are next to each other



Why cluster microarray data?

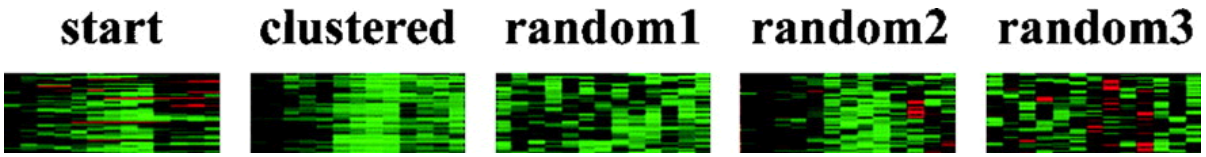
- Guilt-by-association: if unknown gene ***i*** is similar in expression to known gene ***j***, maybe they are involved in the same/related pathway

- Dimensionality reduction: datasets are too big to be able to get information out without reorganizing the data

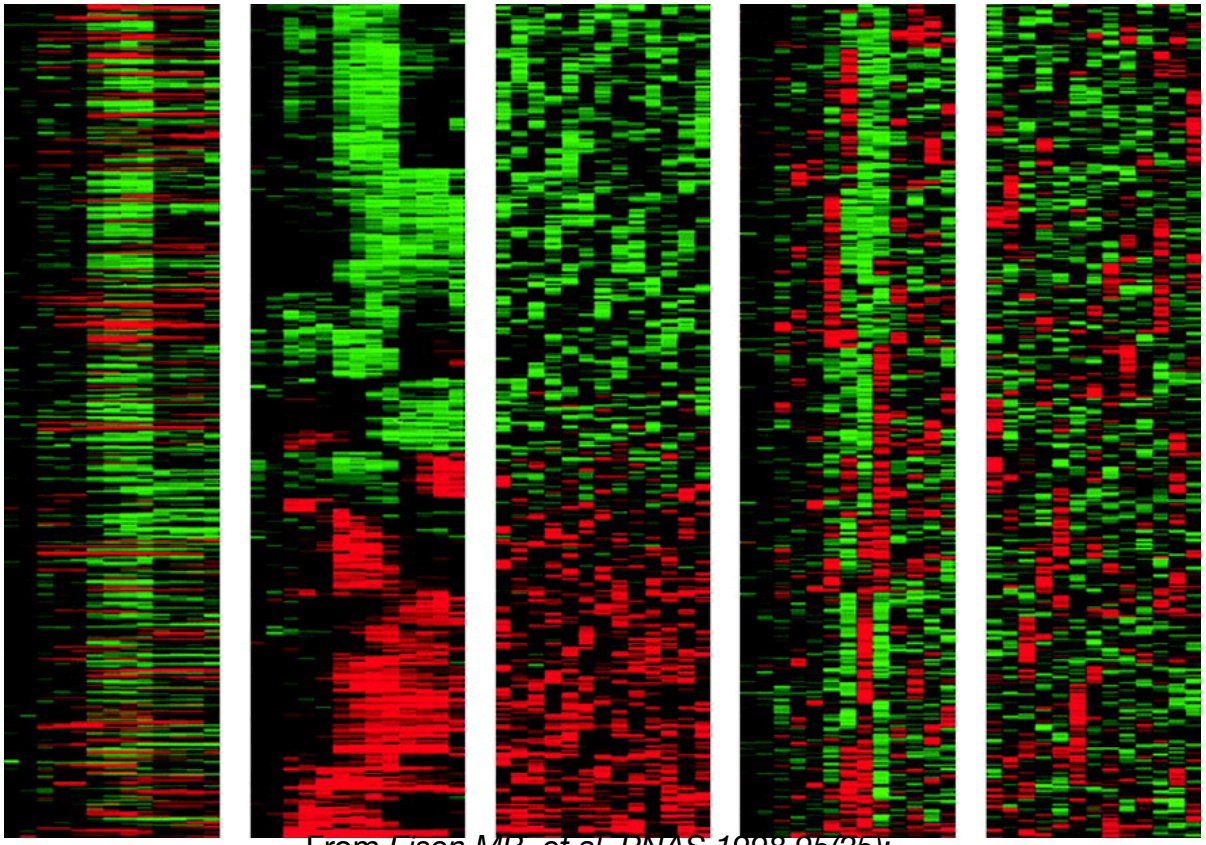


Botstein & Brown group

Clustering Random vs Biological Data

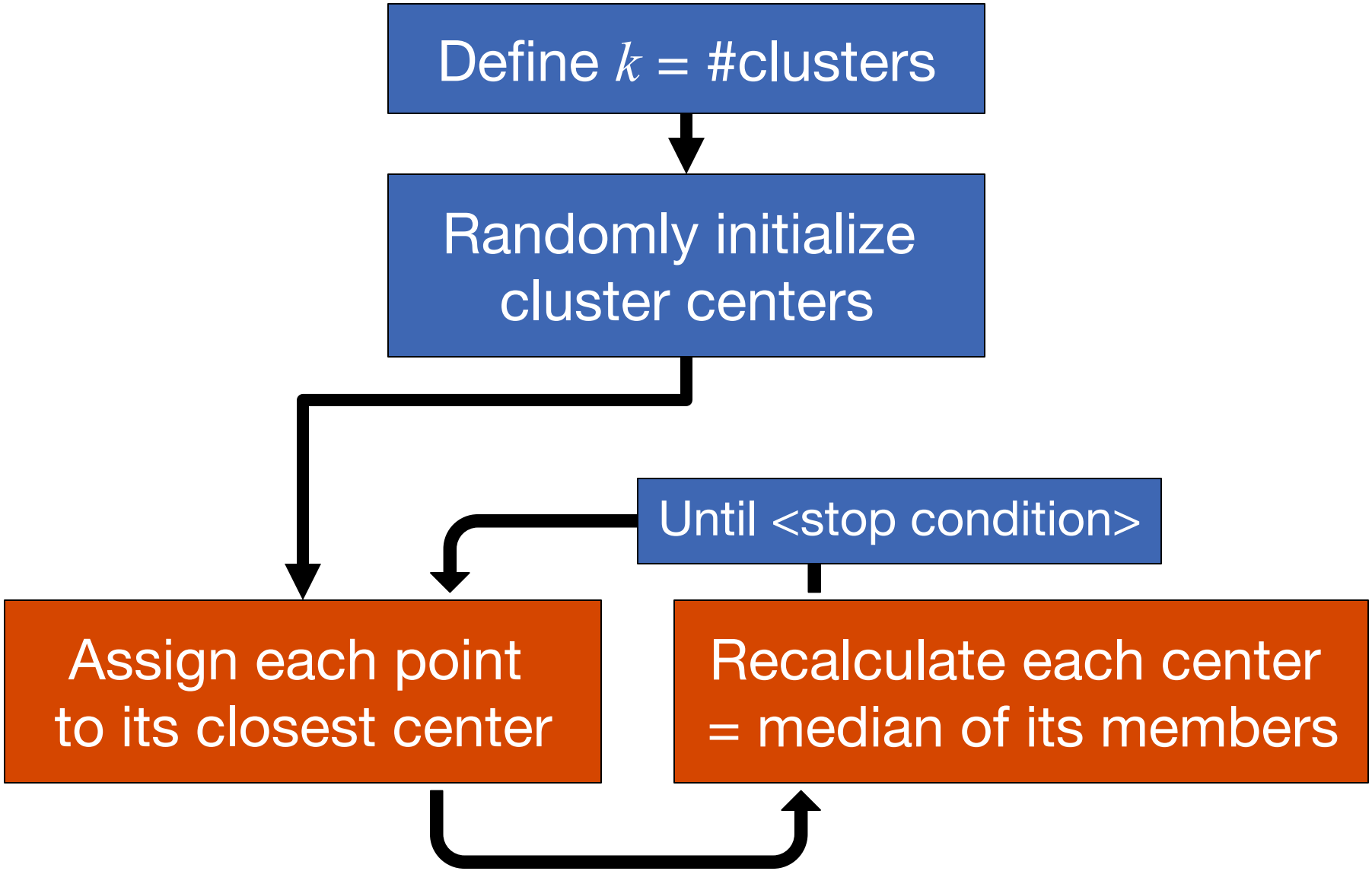


Challenge – when is clustering “real”?



From *Eisen MB, et al, PNAS 1998 95(25): 14863-8*

K-means clustering



K-means clustering

DEMO

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

K-means clustering

Conceptually similar to Expectation-Maximization

EM iteration alternates 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, and

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 twice in a row
- Until some minimal number of genes (e.g. 90%) get assigned to the same partition twice in a row

K-means clustering

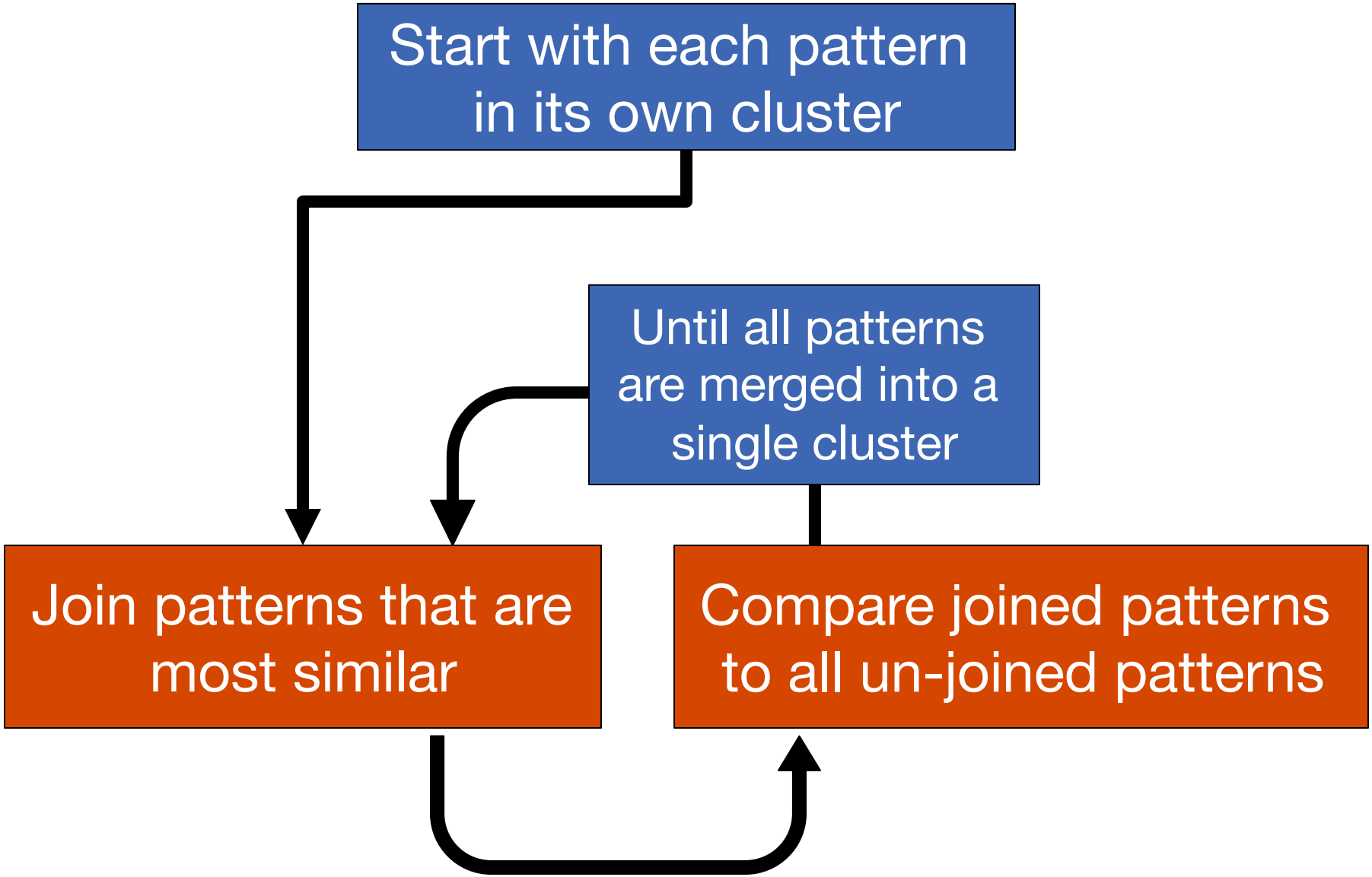
Some issues

- Have to set ***k*** ahead of time
- Prefers clusters of approx. similar sizes
- Each gene only belongs to 1 cluster
- Genes assigned to clusters on the basis of all experiments

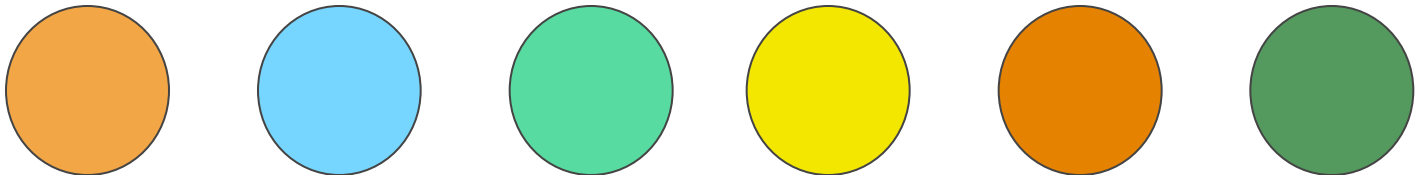
Hierarchical clustering

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

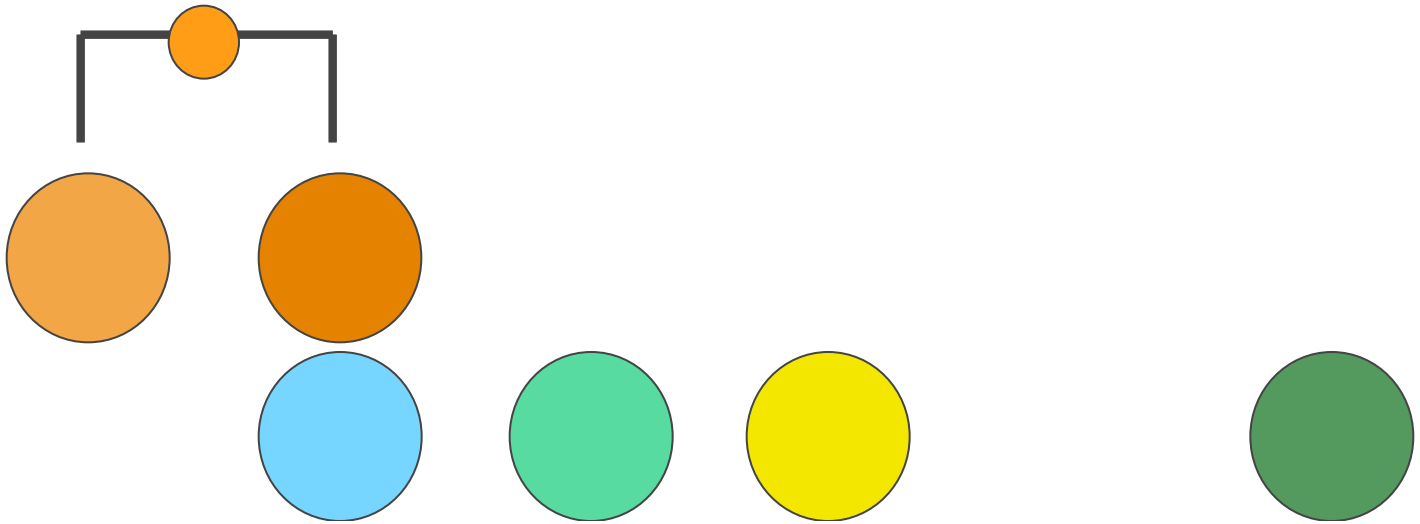
Hierarchical clustering



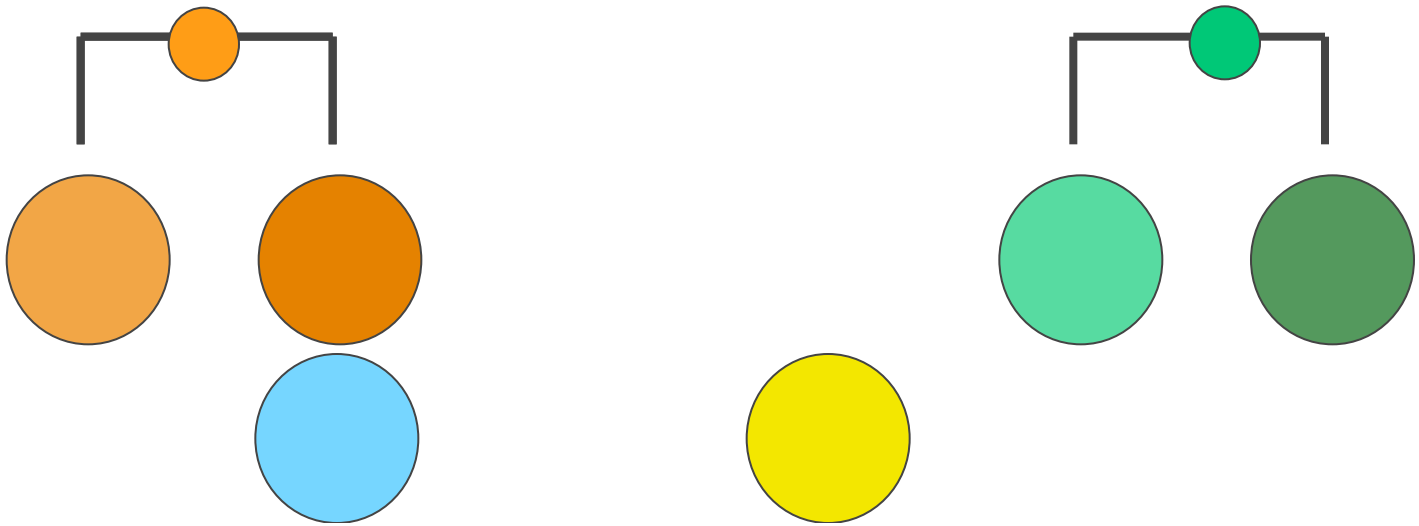
Hierarchical clustering



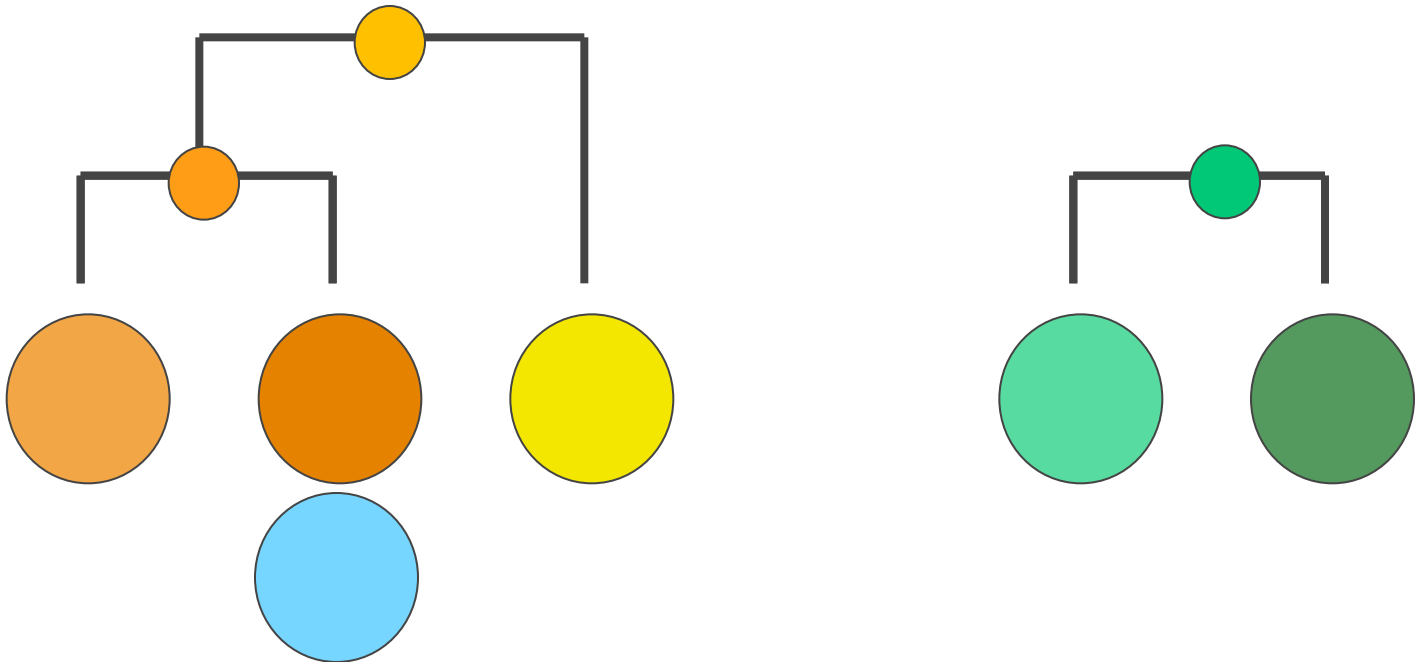
Hierarchical clustering



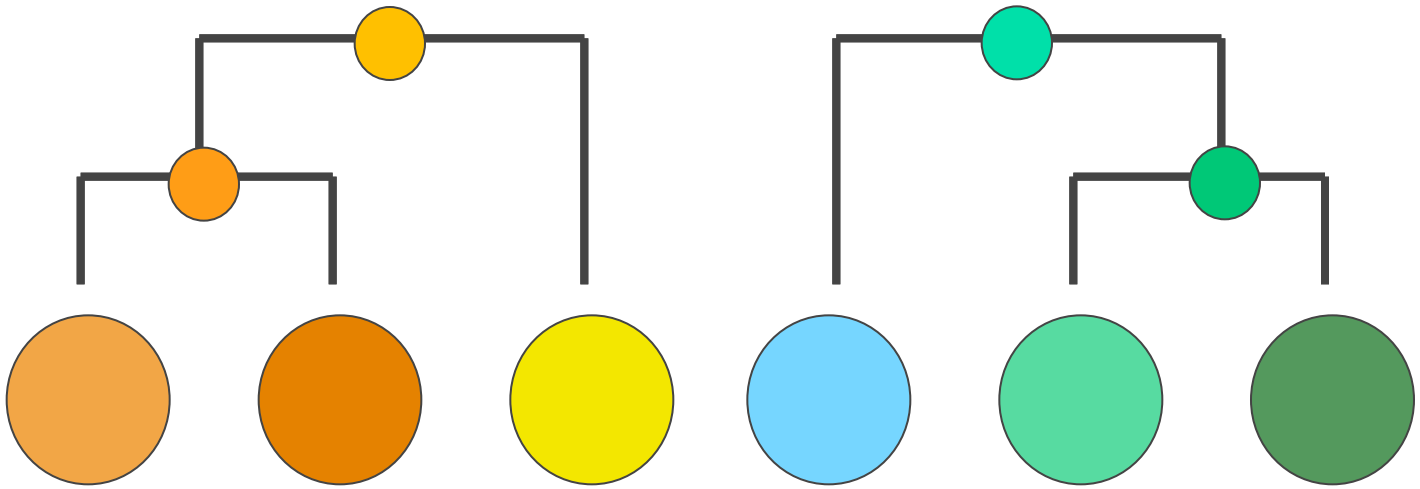
Hierarchical clustering



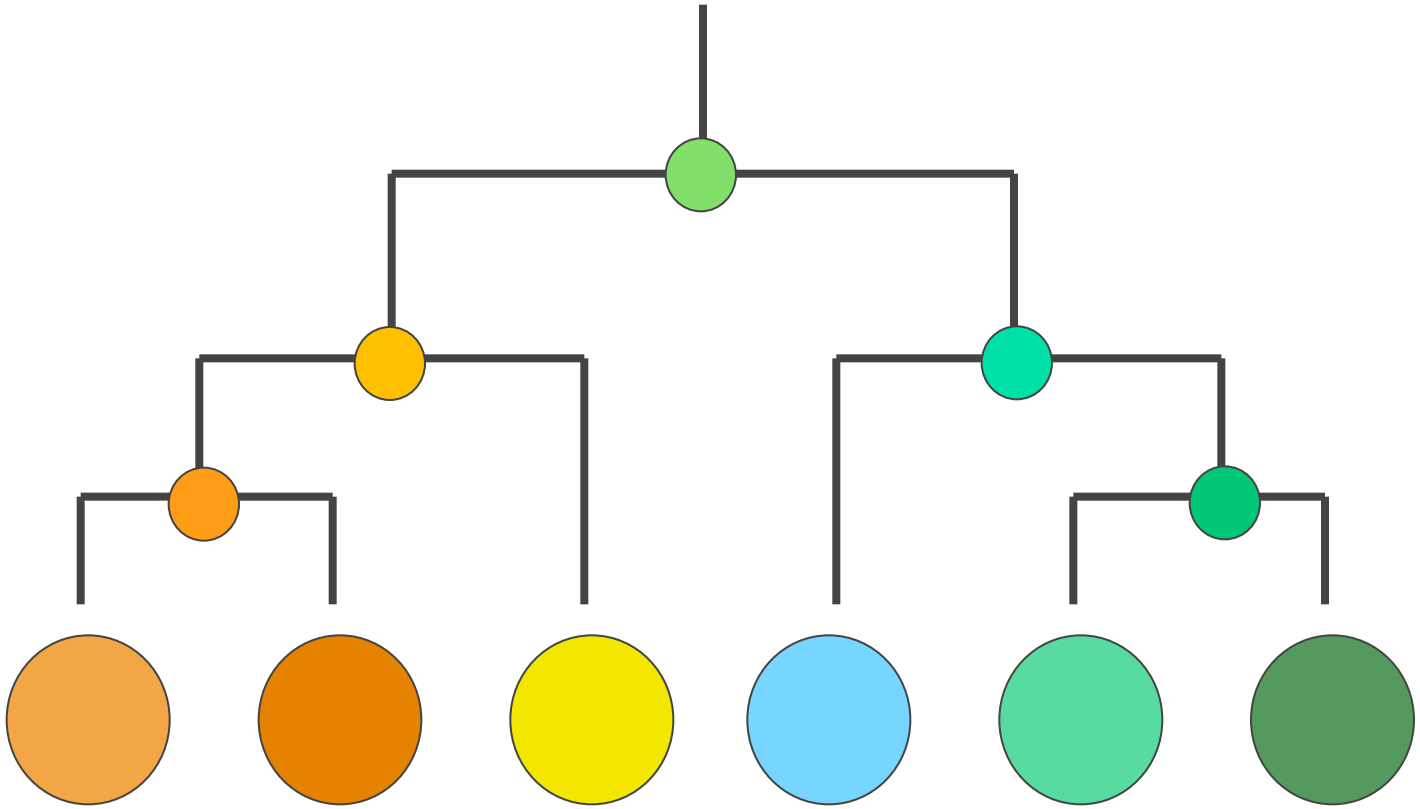
Hierarchical clustering



Hierarchical clustering

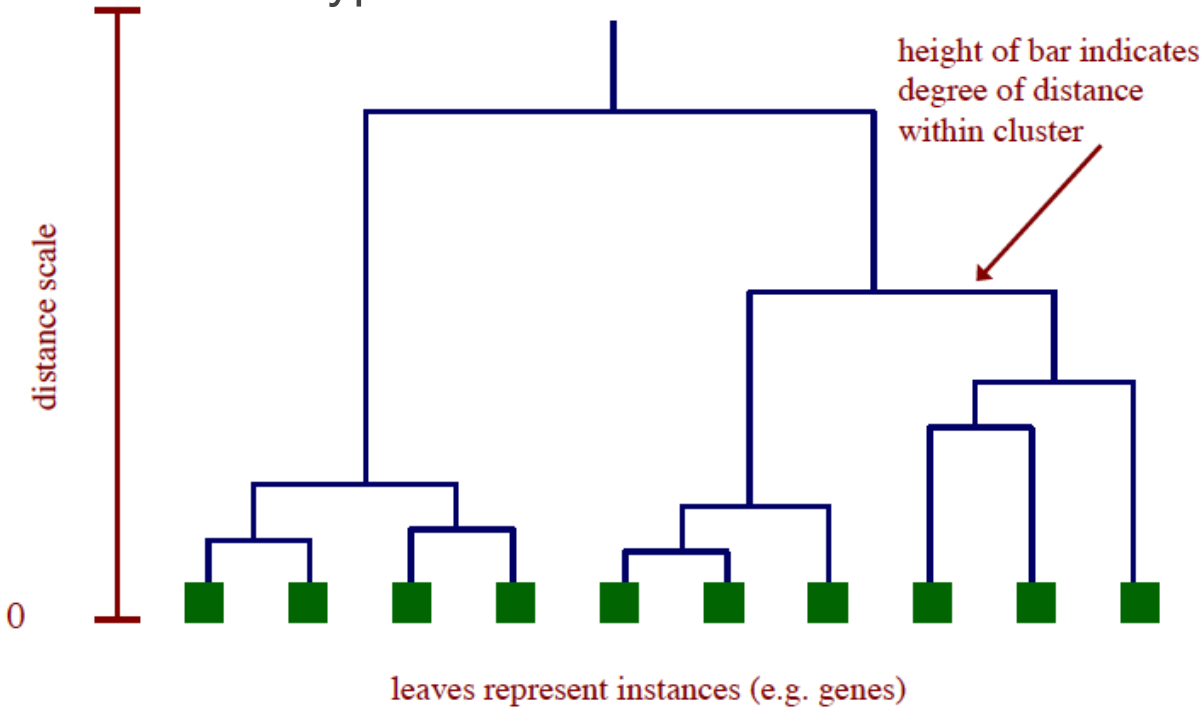


Hierarchical clustering



Dendrogram

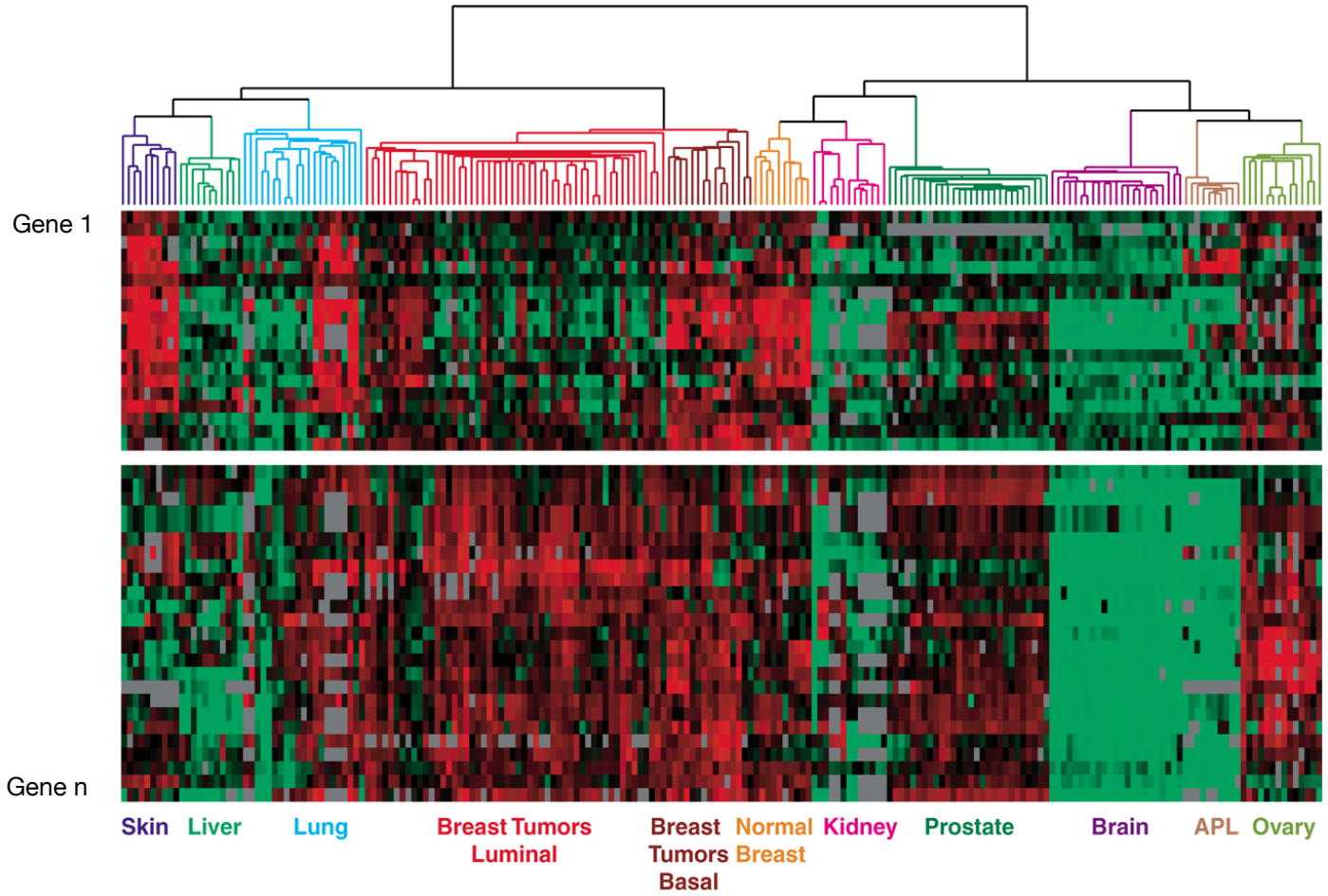
- Dendrogram. Scientific visualization of hypothetical sequence of evolutionary events.
  - Leaves = genes.
  - Internal nodes = hypothetical ancestors.



Reference: http://www.biostat.wisc.edu/bmi576/fall-2003/lecture13.pdf

Dendrogram of Human tumors

Tumors in similar tissues cluster together.



Reference: Botstein & Brown group

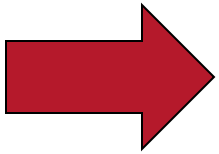
Hierarchical clustering: problems

- Hard to define distinct clusters
- Genes assigned to clusters on the basis of all experiments
- Optimizing node ordering hard (finding the optimal solution is NP-hard)
- Can be influenced by one strong cluster – a problem for gene expression b/c data in row space is often highly correlated

Distance Metrics

- Choice of distance measure is important for most clustering techniques
- Linear measures: Euclidean distance, Pearson correlation
- Non-parametric: Spearman correlation, Kendall's tau

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

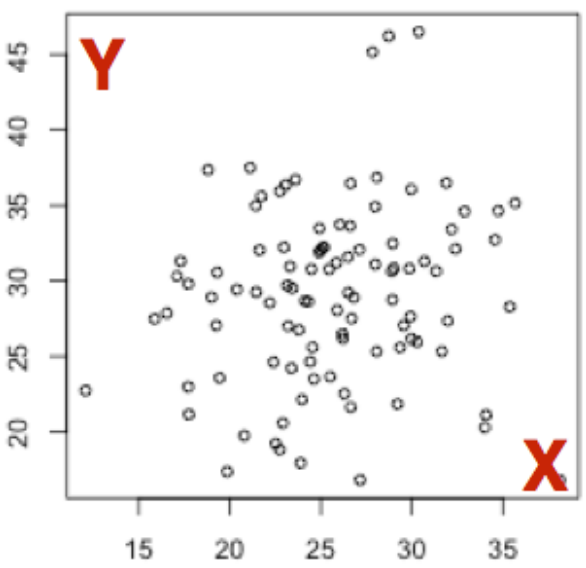


$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

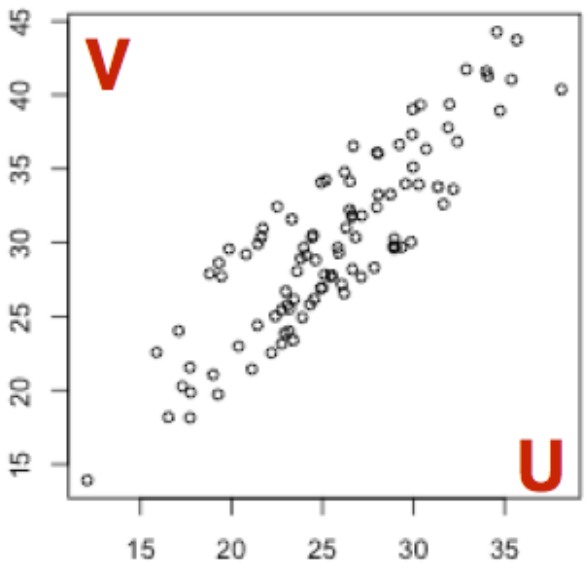
$$\rho = 1 - \frac{6\sum_{i=1}^{n}[rank(x_i) - rank(y_i)]}{n(n^2 - 1)}$$

Distance Metrics

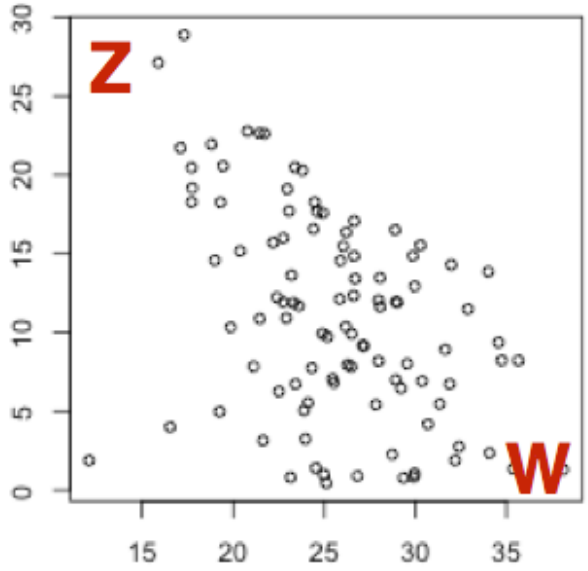
Consider the following plot of 3 pairs of genes



No correlation



Positive correlation

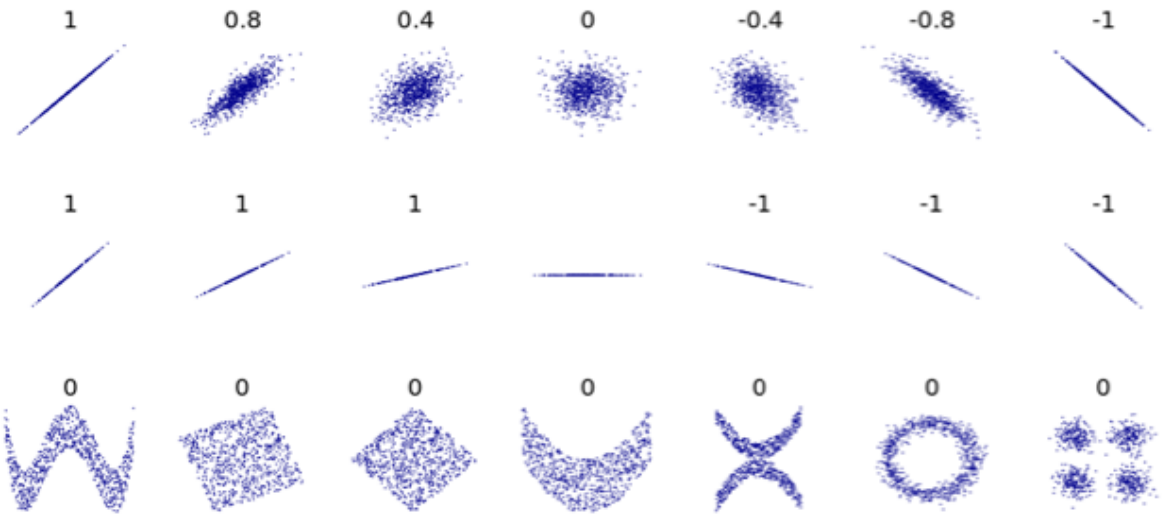


Negative correlation

Distance Metrics

Pearson correlation (r) is a measure of the linear correlation (dependence) between two variables X and Y.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$



+1 ≤ r ≤ –1

+1 is total positive correlation

0 is no correlation

–1 is total negative correlation.

Distance Metrics

11 datapoints

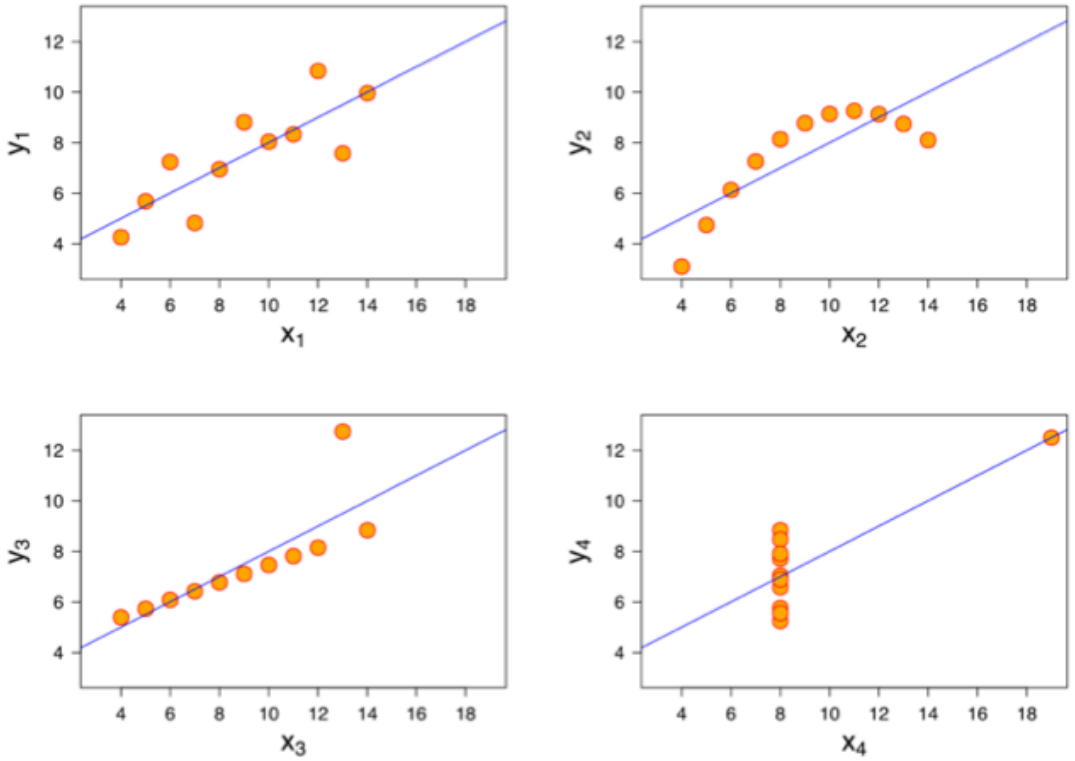
Mean (x) = 9
Var (x) = 11

Mean (y) = 7.50
Var (y) ~ 4.12

Cor (x, y) = 0.816

Linear regression line:
y = 3.00 + 0.500x

Anscombe's quartet



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

Distance Metrics

- Choose your distance measure carefully after:
  - Exploring your data using sanity-checks
  - **Looking** at your data. There is no substitute for this.

- Linear measures: Euclidean distance, Pearson correlation
- Non-parametric: Spearman correlation, Kendall's tau