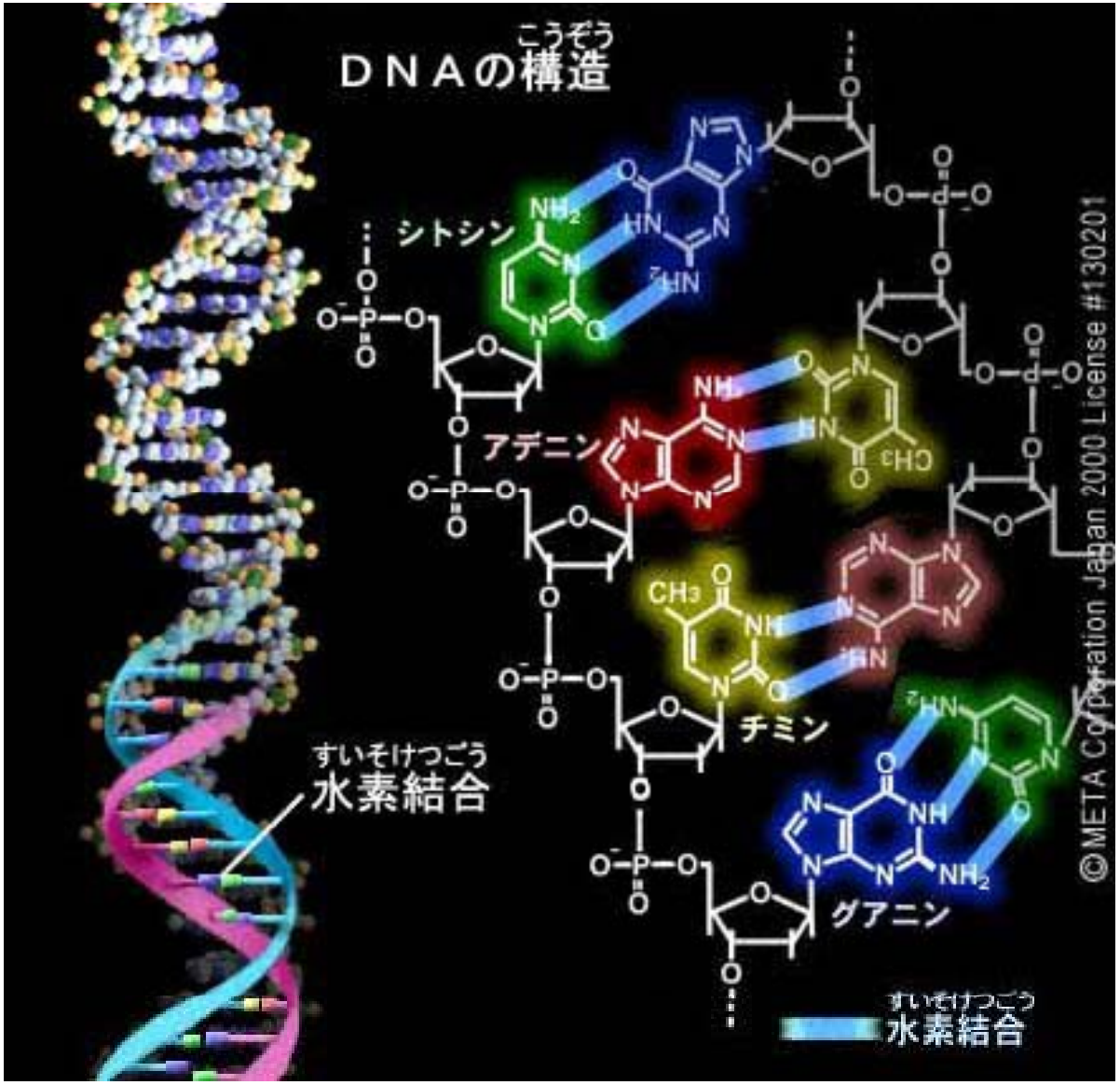
#9
Transcriptome analysis

- Transcription and Transcriptome
- DNA Microarray
  - Spotted microarray, Oligonucleotide microarray
- Data Normalization, Graphical Representation
  - MA-plot, Lowess correction, Volcano plot, Heat Map
- Clustering
  - Bottom up methods: Hierarchical clustering (UPGMA, etc.)
  - Top-down methods: k-mean clustering, SOM

Deoxyribonucleic Acid (DNA)

Review



four types of nucleobases

C: Cytosine（シトシン）

A: Adenine（アデニン）

T: Thymine（チミン）

G: Guanine（グアニン）

base complimentarity

A ･･･ T

G ･･･ C

From : IPA Educational Material Archive, http://www2.edu.ipa.go.jp/gz/

RNA (Ribonucleic Acid)

usually single stranded

possible base pairing:

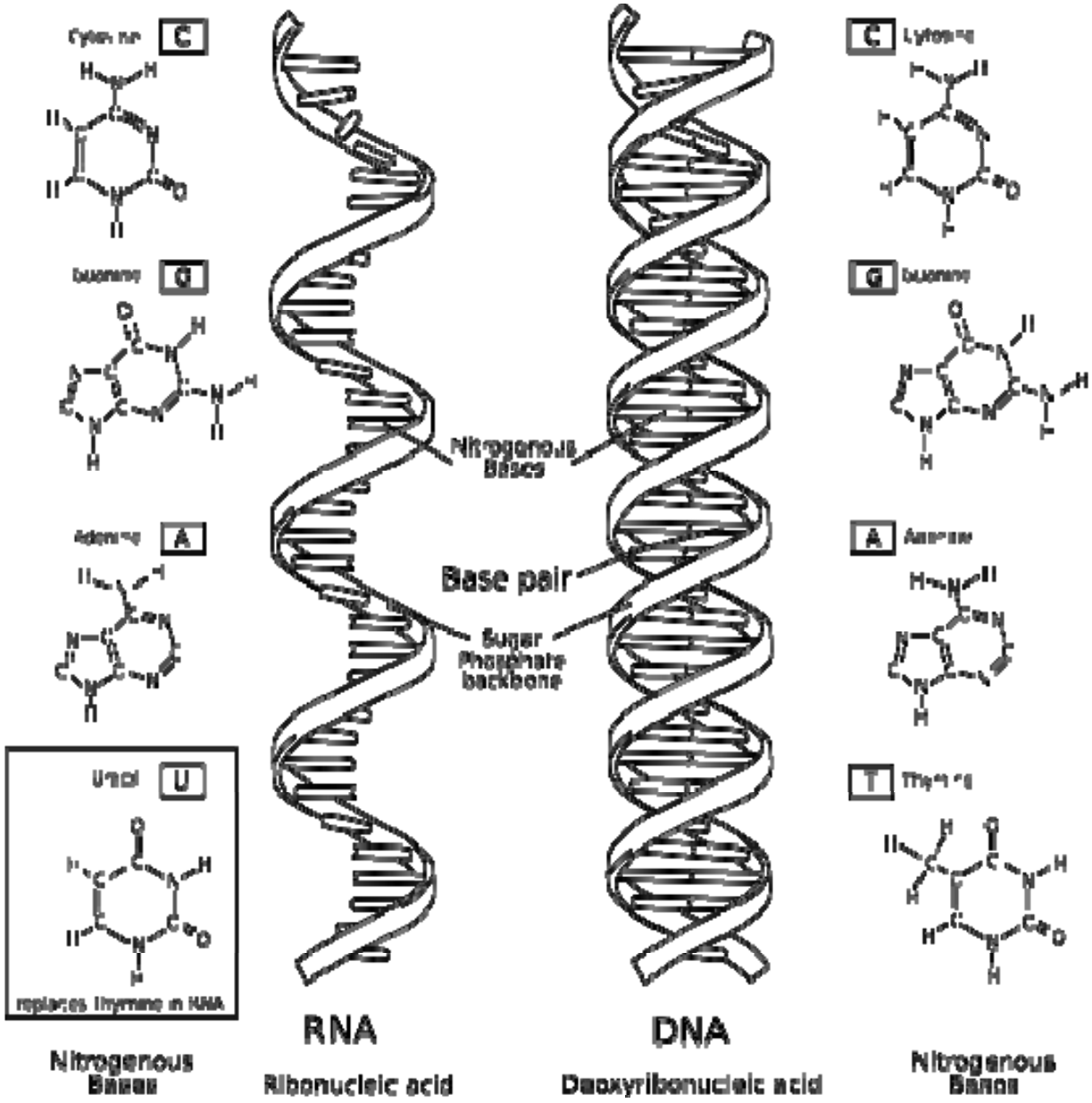
G - C

A - U

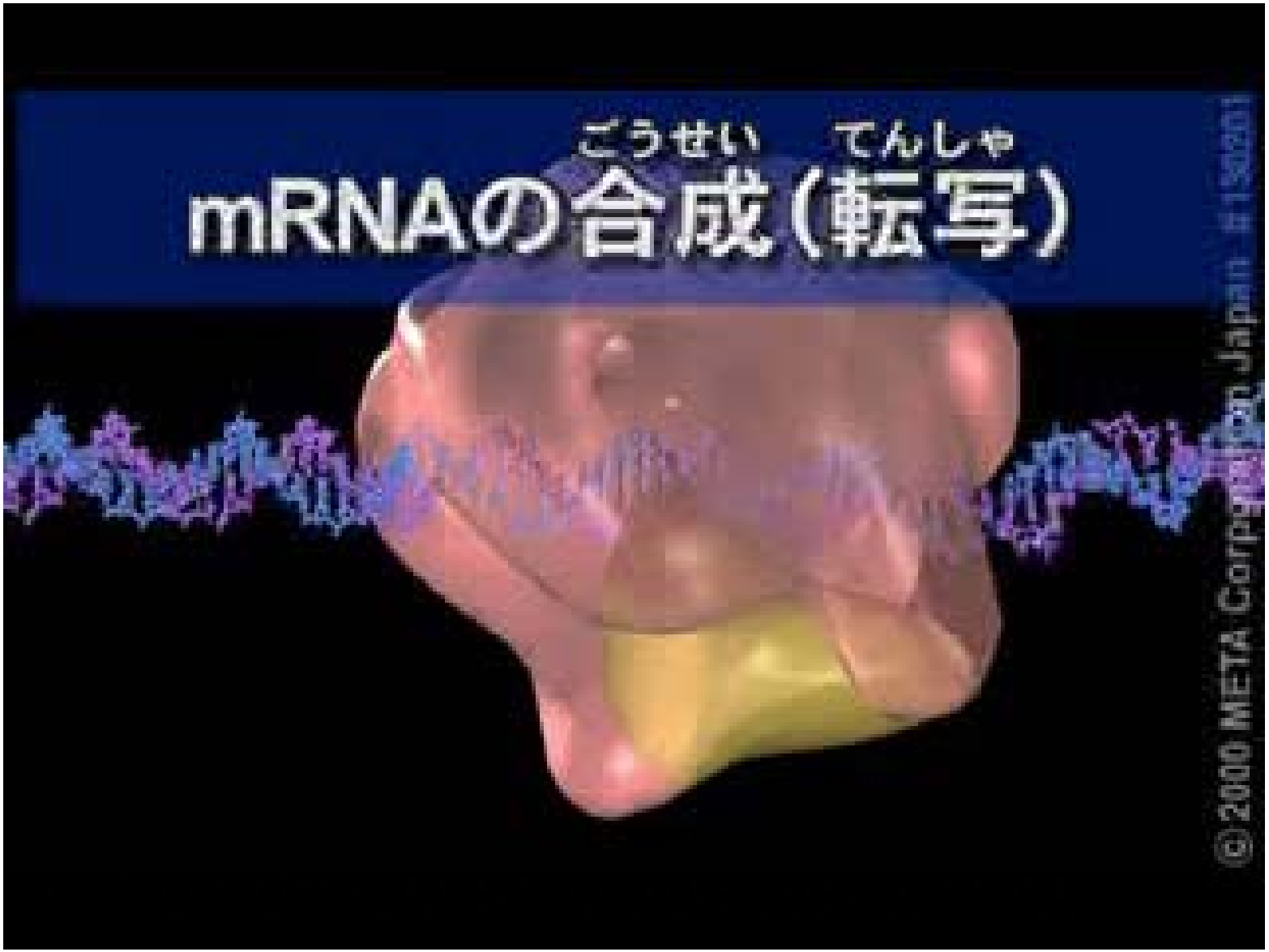
Watson-Crick pairs

(G - U), etc.

wobble pairs



mRNA synthesis (movie)

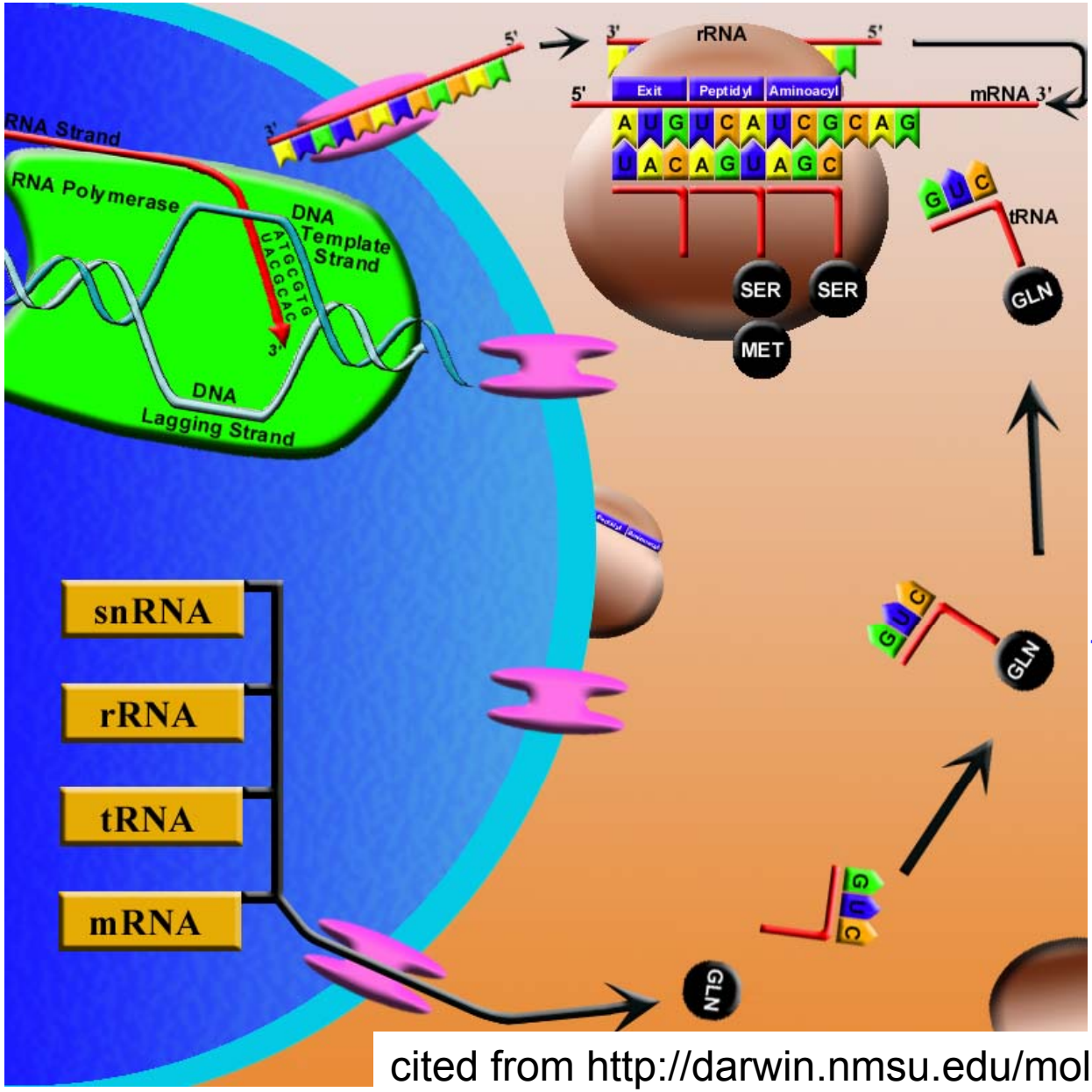


From : IPA Educational Material Archive, http://www2.edu.ipa.go.jp/gz/

From : IPA Educational Material Archive, http://www2.edu.ipa.go.jp/gz/

rRNA

mRNA



tRNA

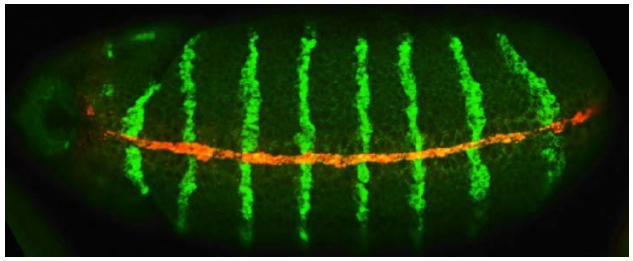
ncRNA

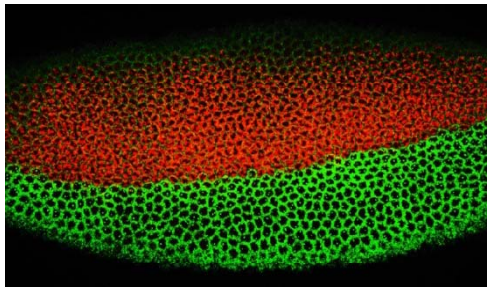
cited from http://darwin.nmsu.edu/molb_resources/

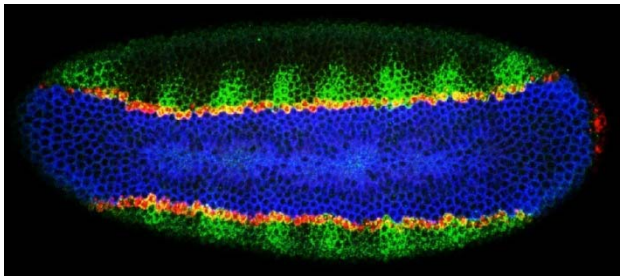
mRNA controls cell fate

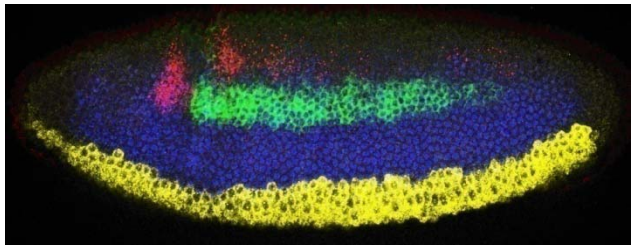
Multiplex fluorescent mRNA ***in situ*** **hybridization**

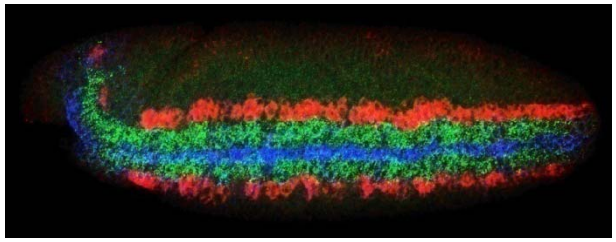
(Fruitfly: Drosophila)

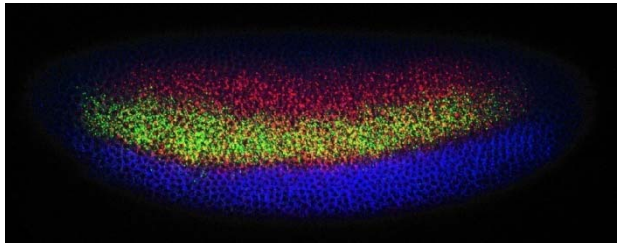








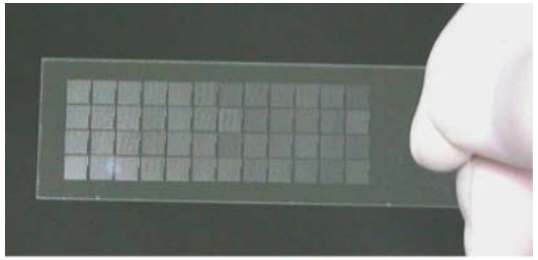




http://superfly.ucsd.edu/~davek/gallery.html

DNA Microarray

Many (10^4 - 10^5) kinds of DNA probe sequences are integrated on a chip.



大腸がんを診断するDNAチップ。小さなガラスに張り付けられている（DNAチップ研究所提供）

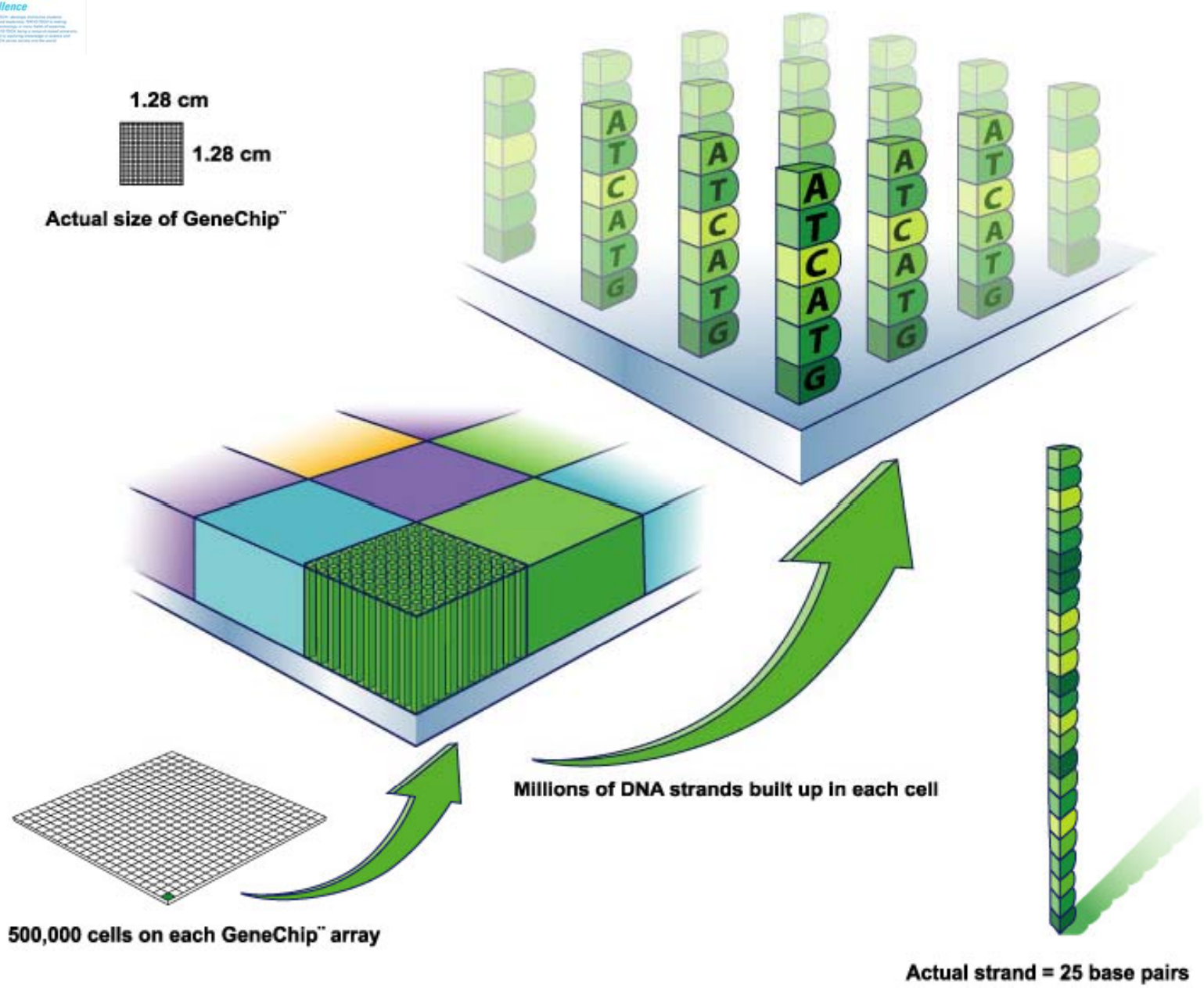
1) Spotted microarray (Stanford type)

Synthesized cDNAs are spotted on a glass by a spotting machine.
Two-color method.

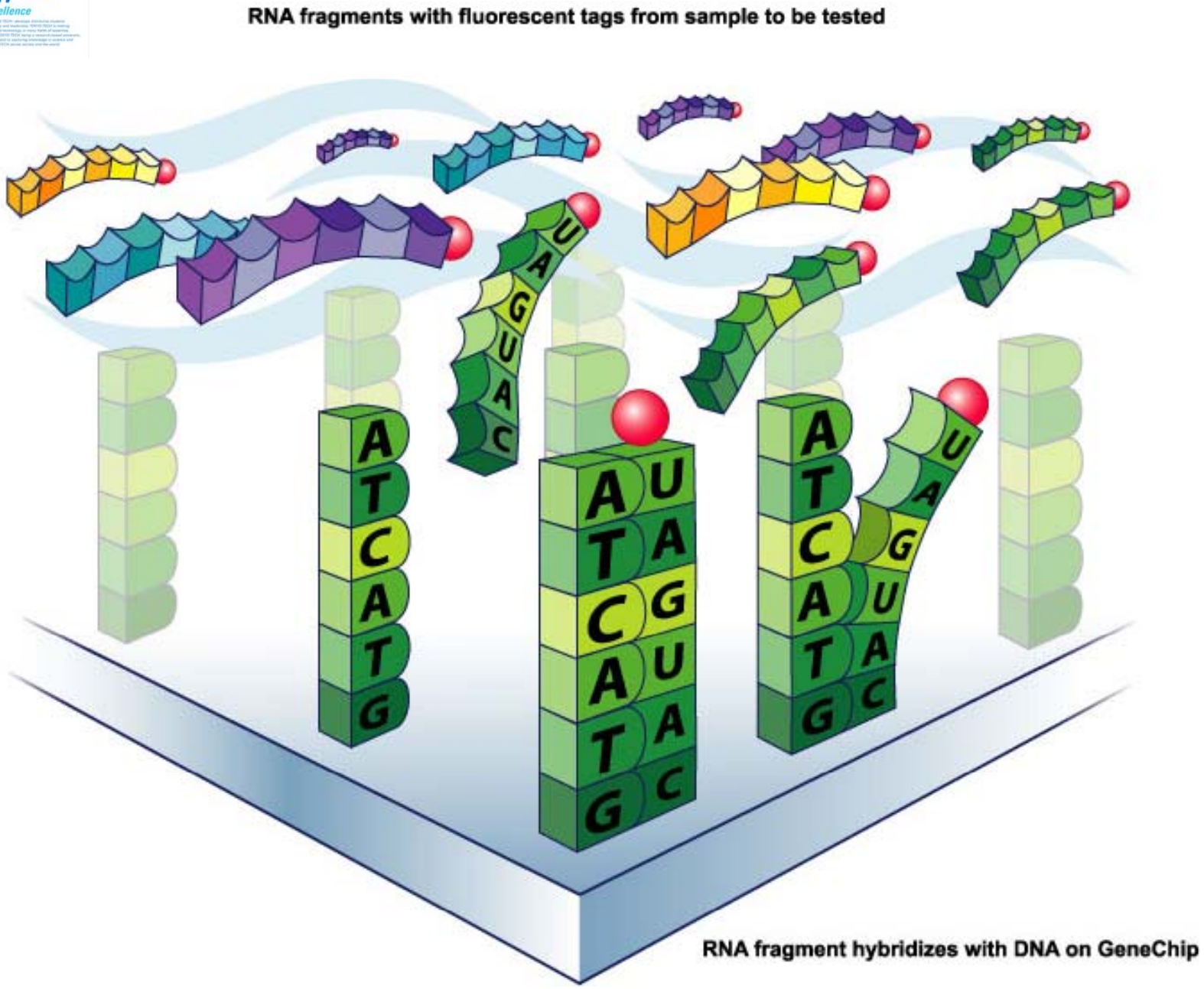
Affymetrix GeneChip

2) Oligonucleotide microarray

Short DNA probes are grown on a silicon wafer, base by base, using mask patterns.
One-color method.

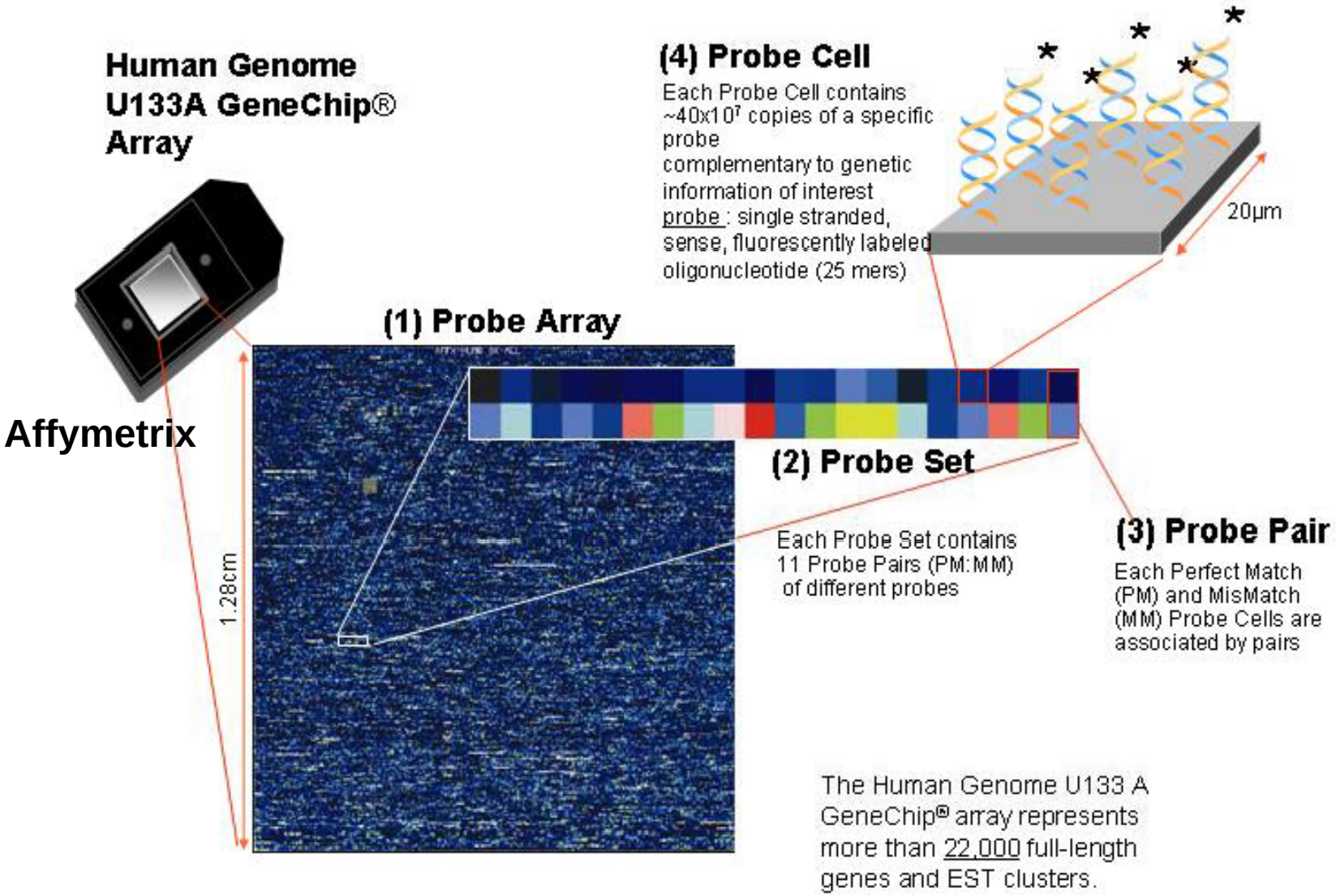


http://www.well.ox.ac.uk/genomics/facilitites/Microarray/



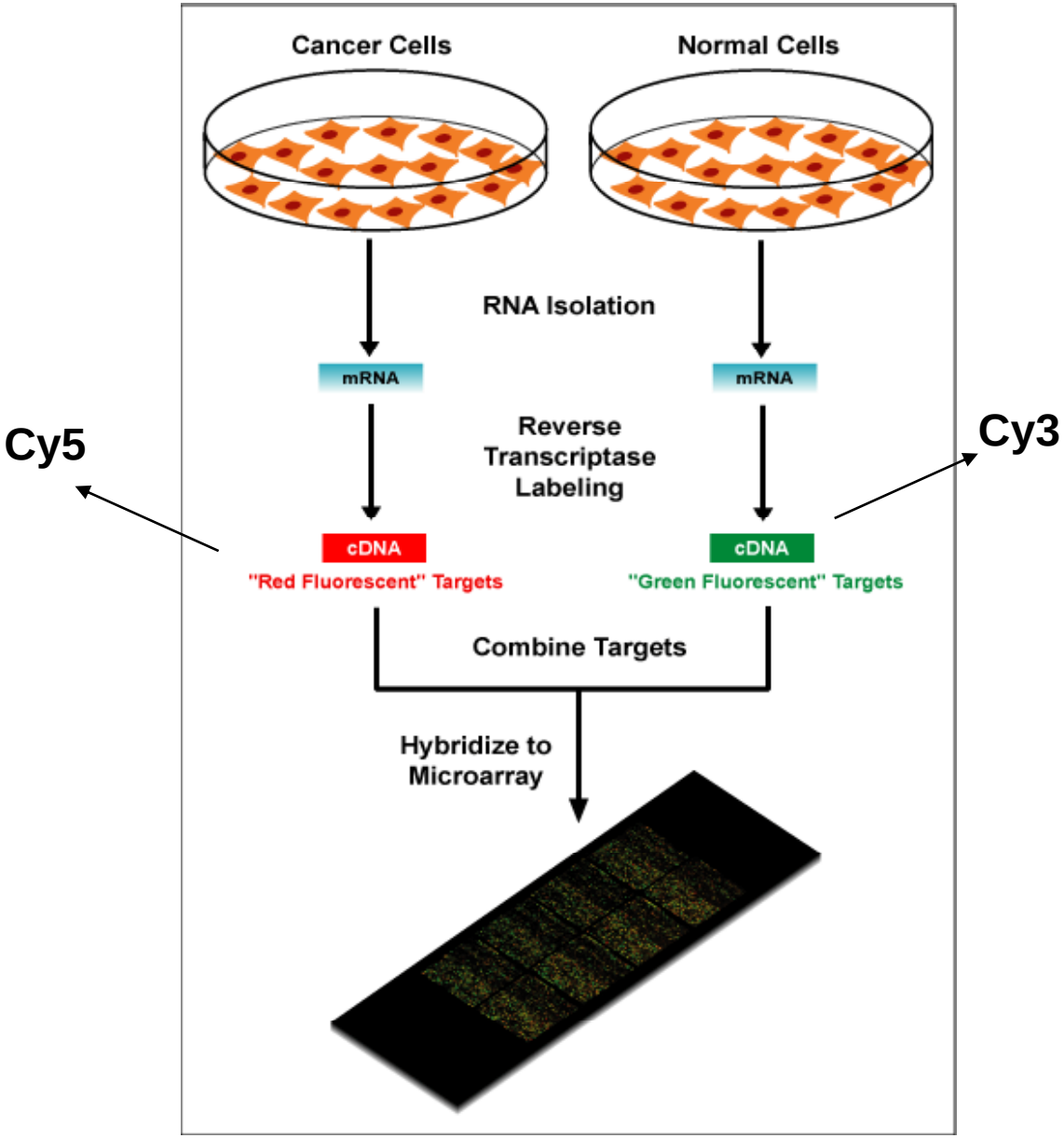
http://www.well.ox.ac.uk/genomics/facilitites/Microarray/

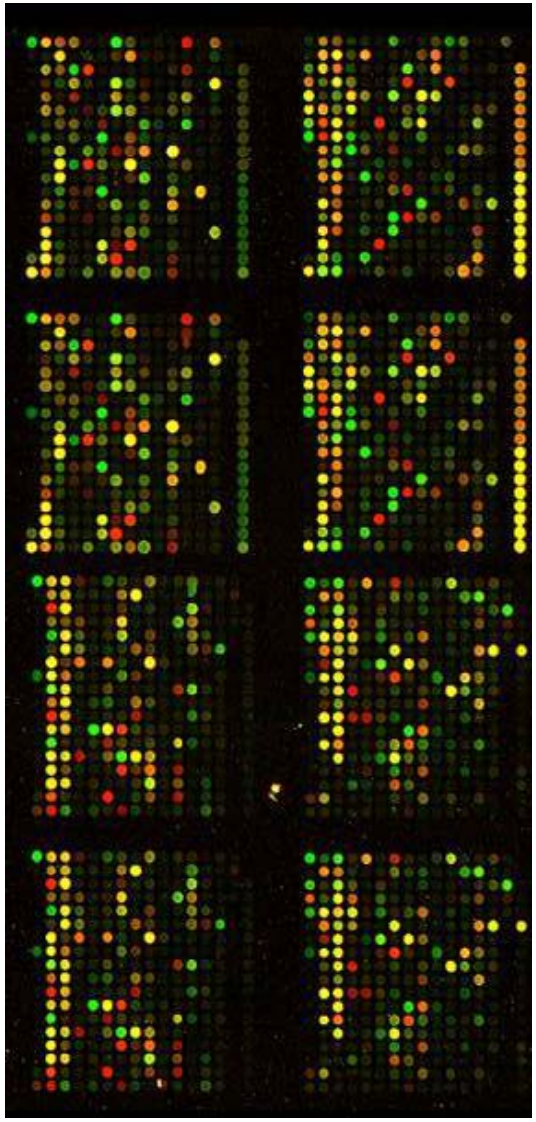
Probe Architecture on a Chip



http://www.weizmann.ac.il/home/ligivol/pictures/system.jpg

Two-color (dual-color) method

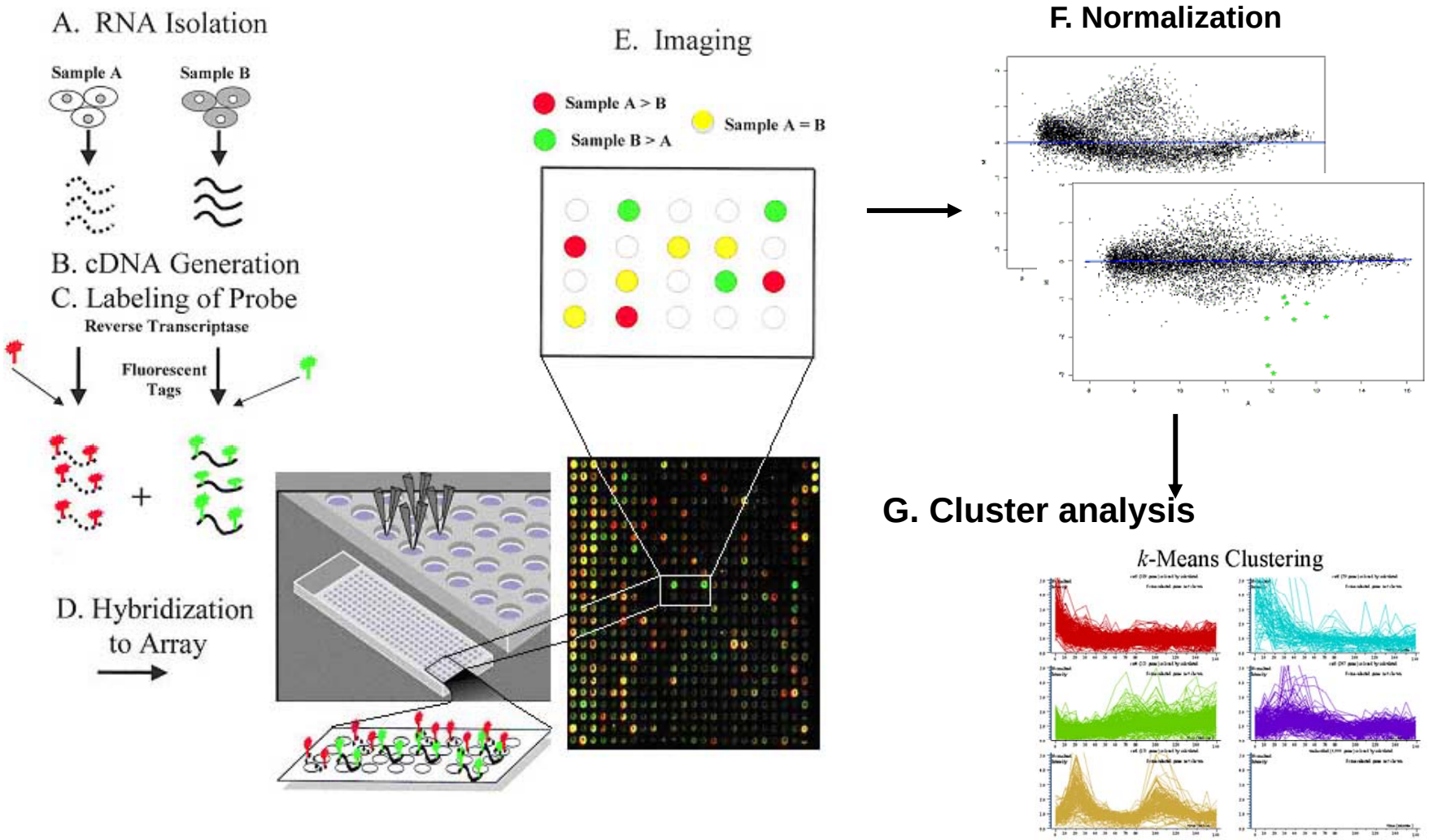




modified from http://upload.wikimedia.org/wikipedia/en/3/32/Microarray-schema.gif

cDNA Microarray Analysis Flow

cDNA: Complimentary DNA (made from RNA sample)



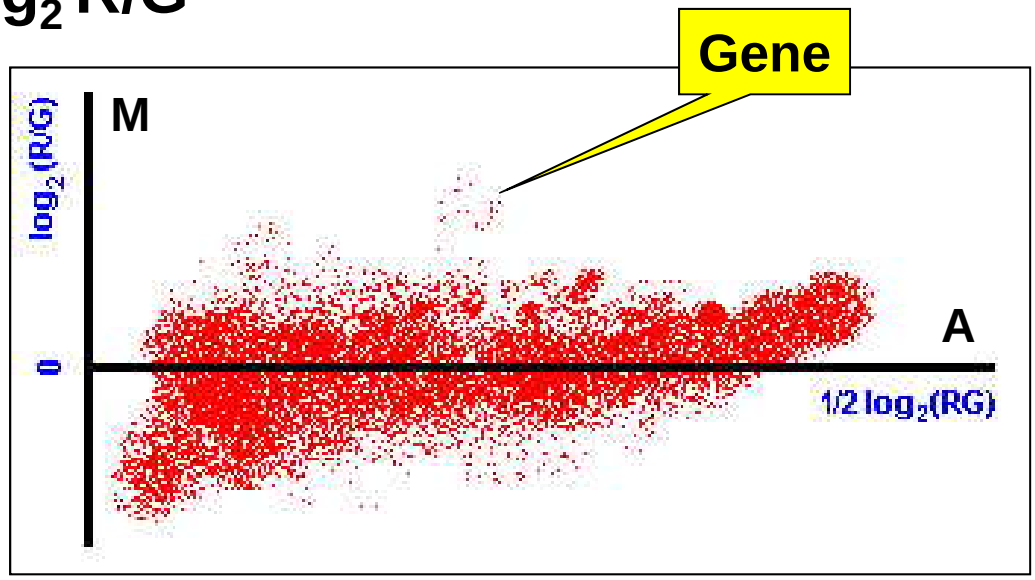
Original slide: courtesy of Dr. W. Fujibuchi (CBRC)

MA-plot

Used for data normalization in two-color method.

$\log_2 R/G$

Ratio of two colors



$1/2\log_2 RG$

Multiplication of two

$$M = \log_2 R/G$$
$$= \log_2 R - \log_2 G$$

(M = Minus)

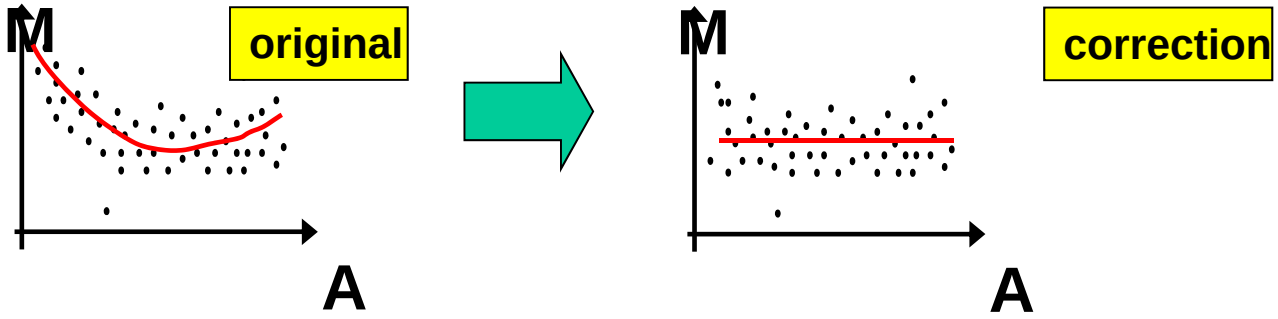
$$A = 1/2\log_2 RG$$
$$= 1/2[\log_2 R + \log_2 G]$$

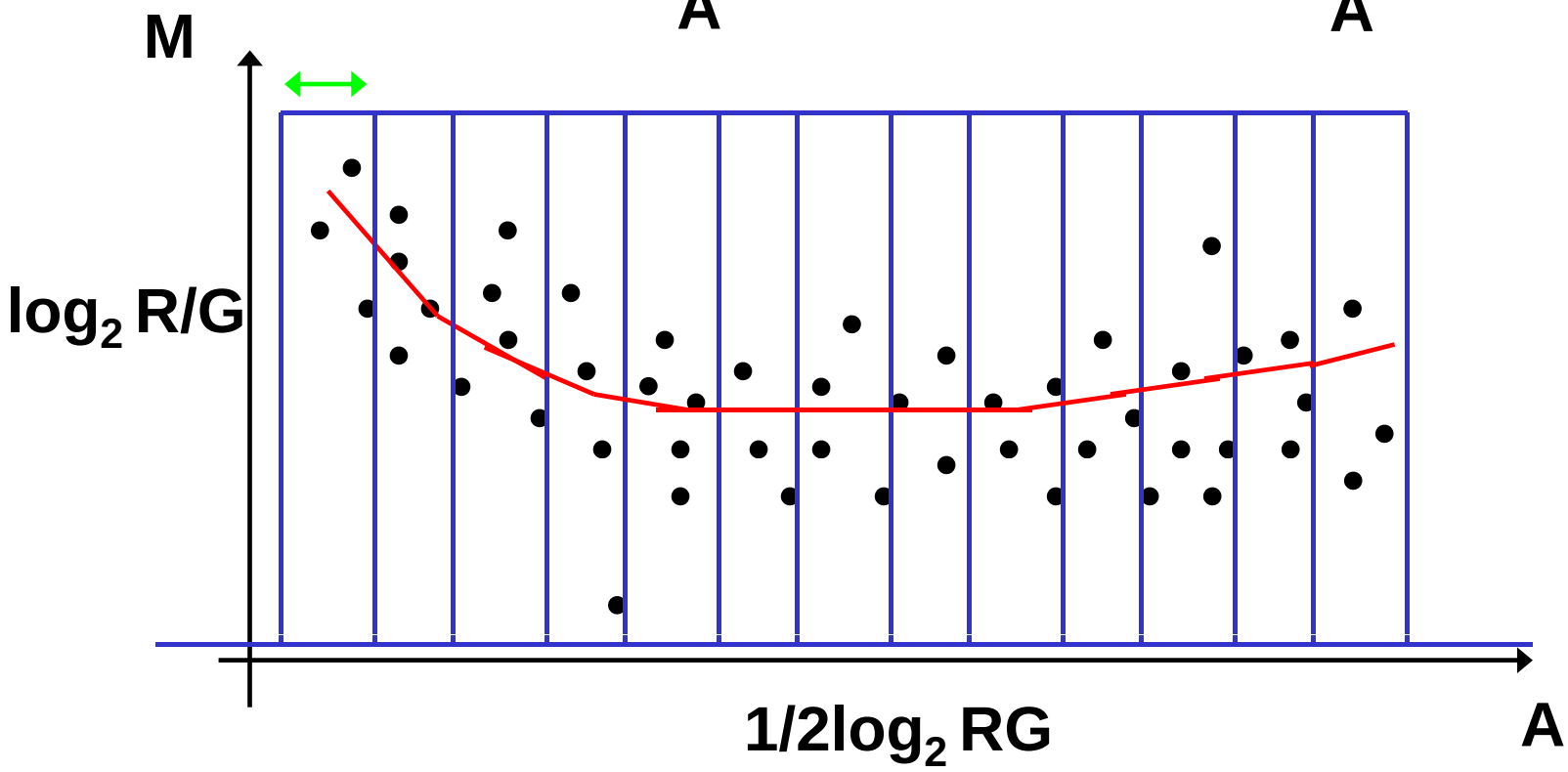
(A = Add)

Original slide: courtesy of Dr. W. Fujibuchi (CBRC)

Lowess correction

Lowess (Local weighted regression)

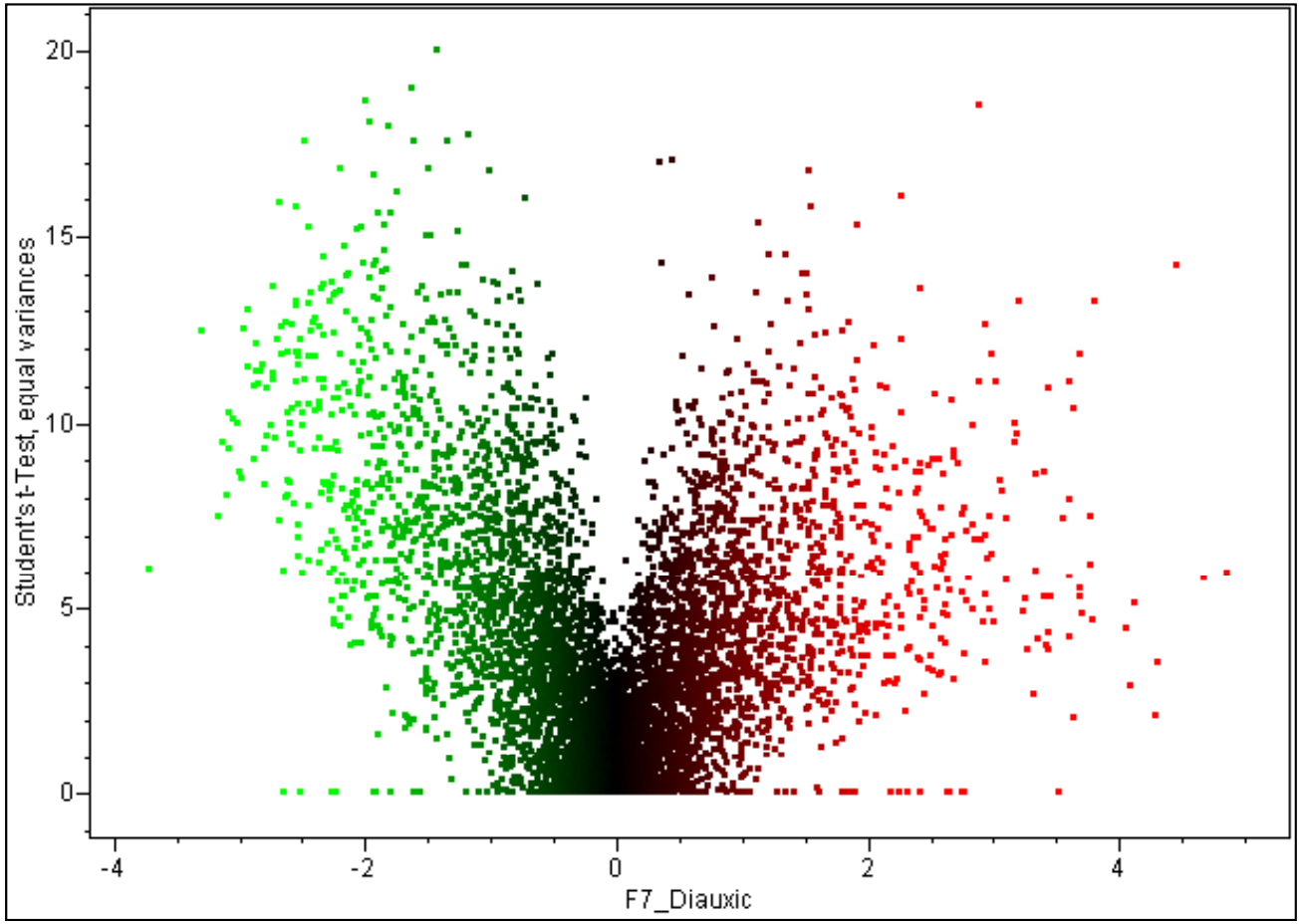




Volcano Plot

T-value (statistical significance)

p-value in Student's T-test



M=log(R/G) **(= Fold Change)**

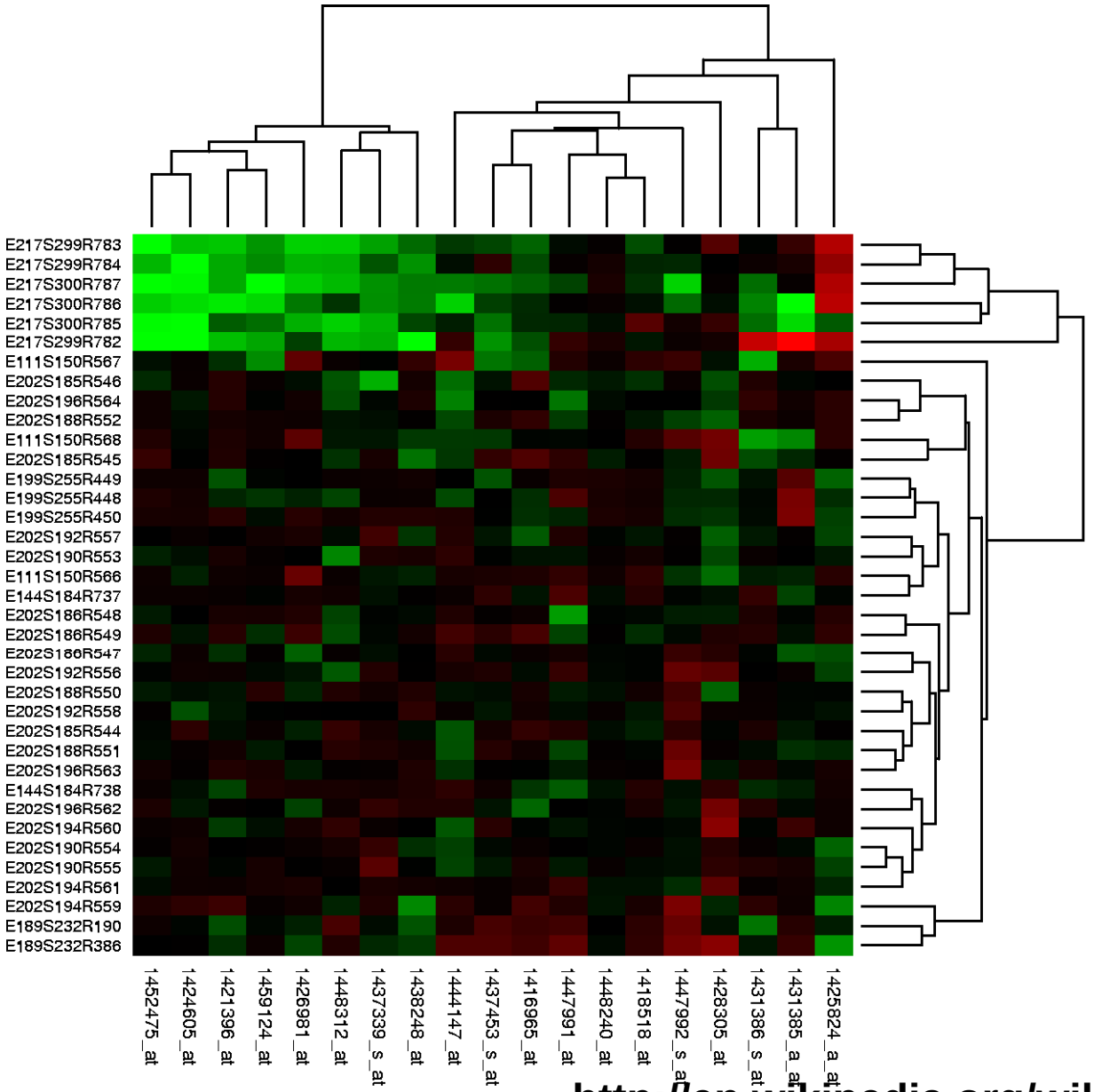
www.dcs.gla.ac.uk/~rb106x/talks/Analysis_of_Gene_Expression_Data2005.ppt

Gene Expression Data

	Exp.1	Exp.2		Exp.*n*
Gene 1	x_{11}	x_{12}	$\cdots$	x_{1n}
Gene 2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Gene *m*	x_{m1}	x_{m2}	$\cdots$	x_{mn}

Expression profile of i-th Gene $\mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{in}\right)$

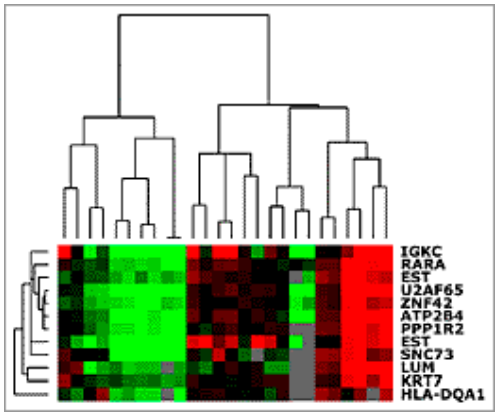
Heat Map



http://en.wikipedia.org/wiki/Heat_map

Clustering of related genes

- Bottom-up methods
  - Hierarchical Clustering
    - Deterministic
    - Progressive
    - Tree representation obtained



http://www.oncology.cam.ac.uk/Brenton.html

- Top-down methods
  - K-means clustering
  - SOM (Self Organization Map)
    - Group number must be determined *a priori*
    - Non-deterministic

Distance Definition

Gene expression profiles: g1 (x1, x2, …, xn), g2 (y1, y2, …, yn)

- Distance measures
  - Euclidian distance

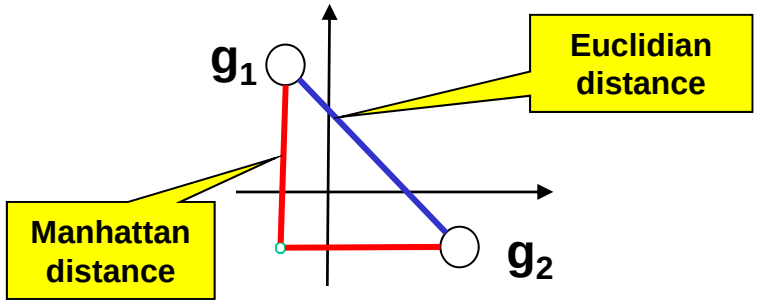
$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

    - Measuring absolute value. Normalization is important. D≧0

  - Pearson's correlation coefficient
    - -1 < Corr < +1 (D = 1 - |Corr|)

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

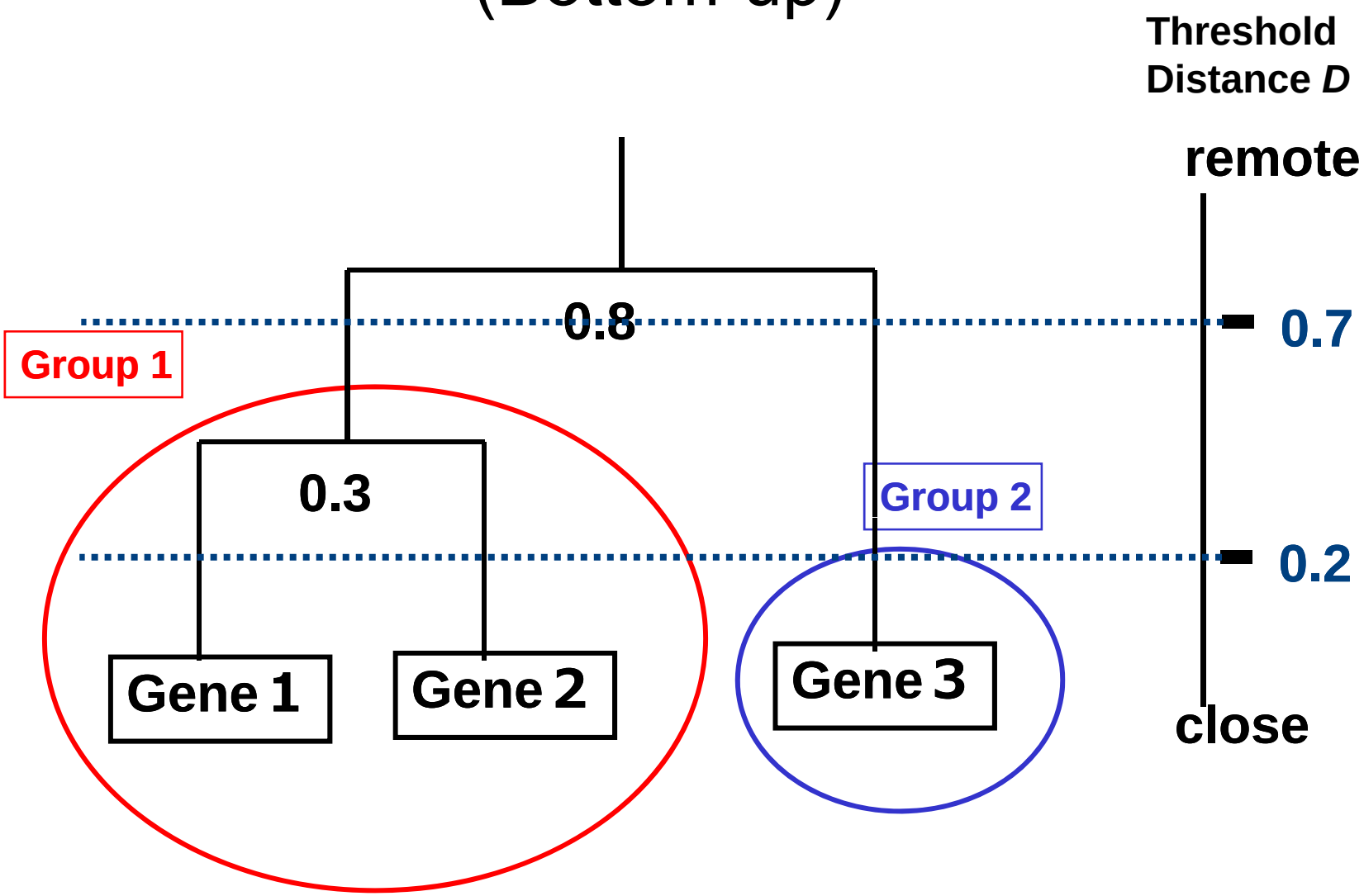
  - Manhattan distance (city-block distance)



$$d = \sum_{i=1}^{n} |x_i - y_i|$$

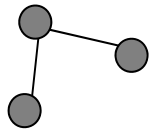
Hierarchical Clustering

(Bottom-up)

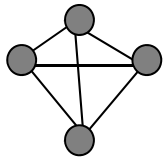


Linkage methods

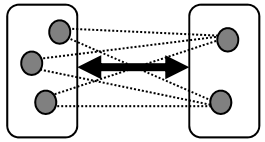
- Single linkage clustering
  - （min. distance）d1 ≤ Threshold *D* → Linkage



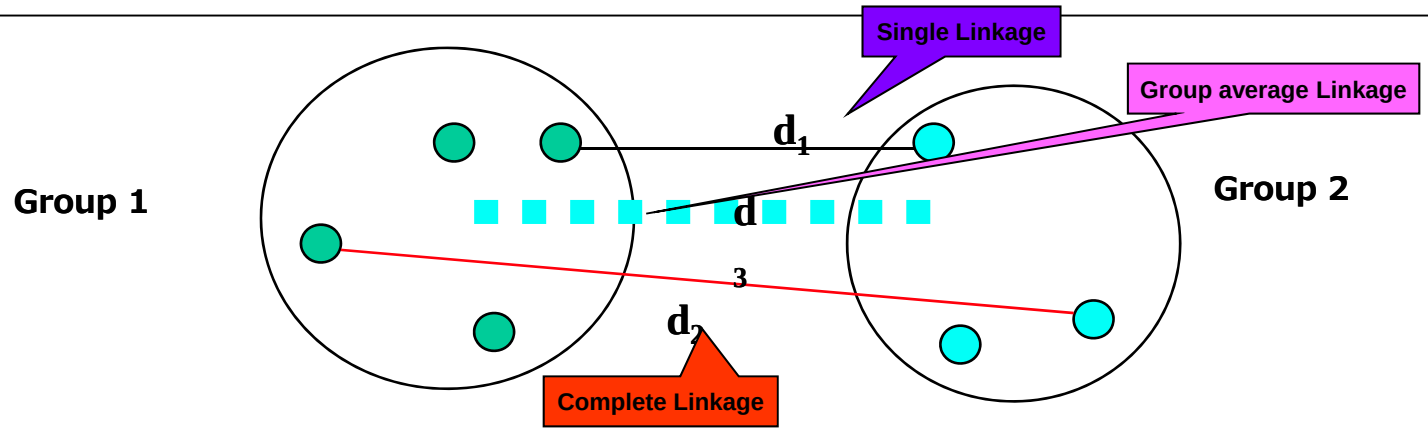
- Complete linkage clustering
  - （max. distance）d2 ≤ Threshold *D* → Linkage



- Group distance linkage clustering
  - （group's distance）d3 ≤ Threshold *D* → Linkage
  - **UPGMA**:
    - Unweighted pair-group method using average
    - Average distance of all possible pairs between two groups



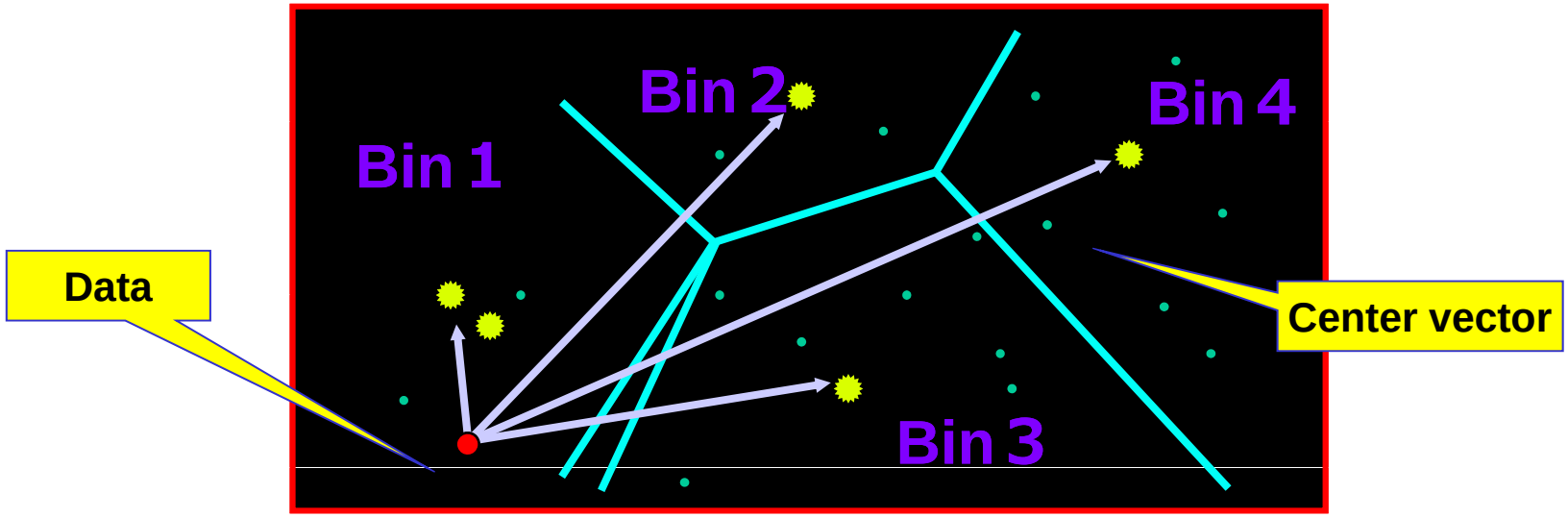
  - UPGMC:
    - **U**nweighted **p**air-**g**roup **m**ethod using **c**entroid
    - Distance between centroids of each groups



K-means clustering

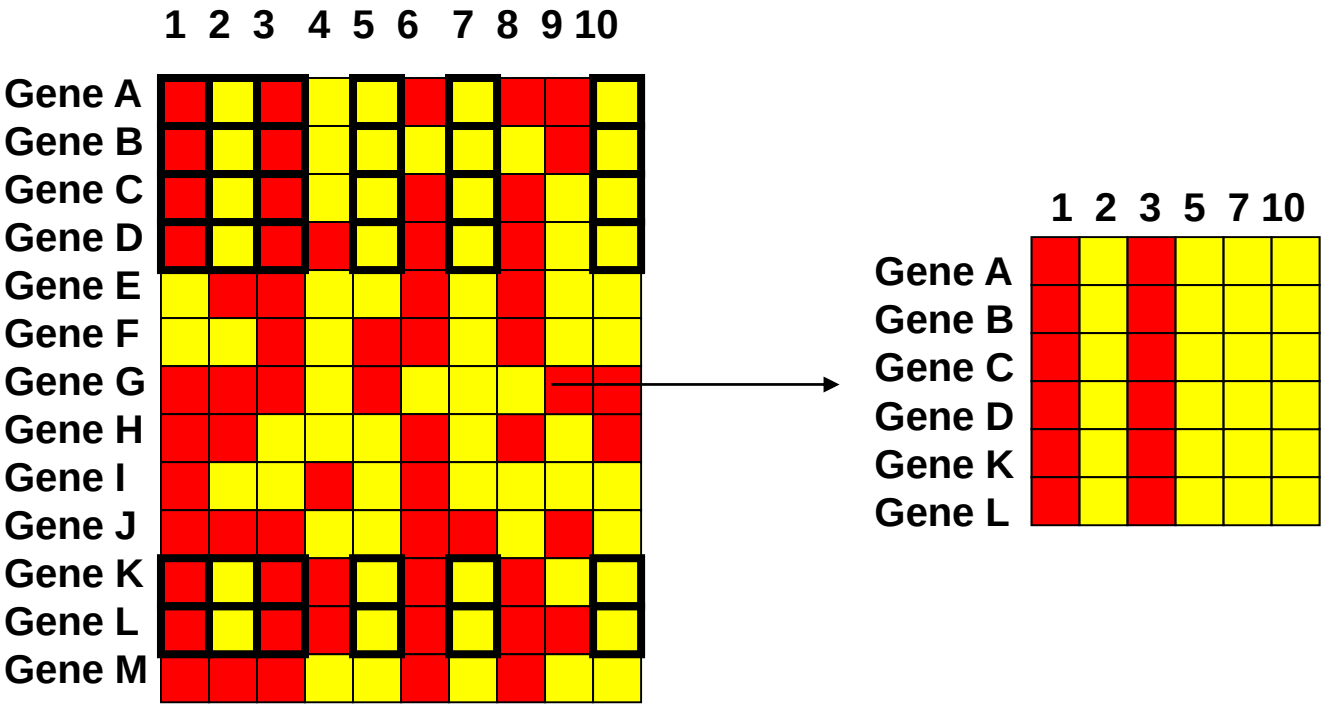
(Top-down)

- Prepare K boxes (bins) and its centroid vector
- Repeat the following procedure
  - 1) New data are assigned to a bin with the closest centroid vector
  - 2) Center vectors are recalculated
  - 3) Take one vector from field, and go back to 1
- Until clustering is converged and no change on centroids.



Bi-directional Clustering

- Simultaneous clustering for “genes” and “samples”



https://www.cs.tcd.ie/research_groups/mlg/kdp/presentations/Bryan_PH