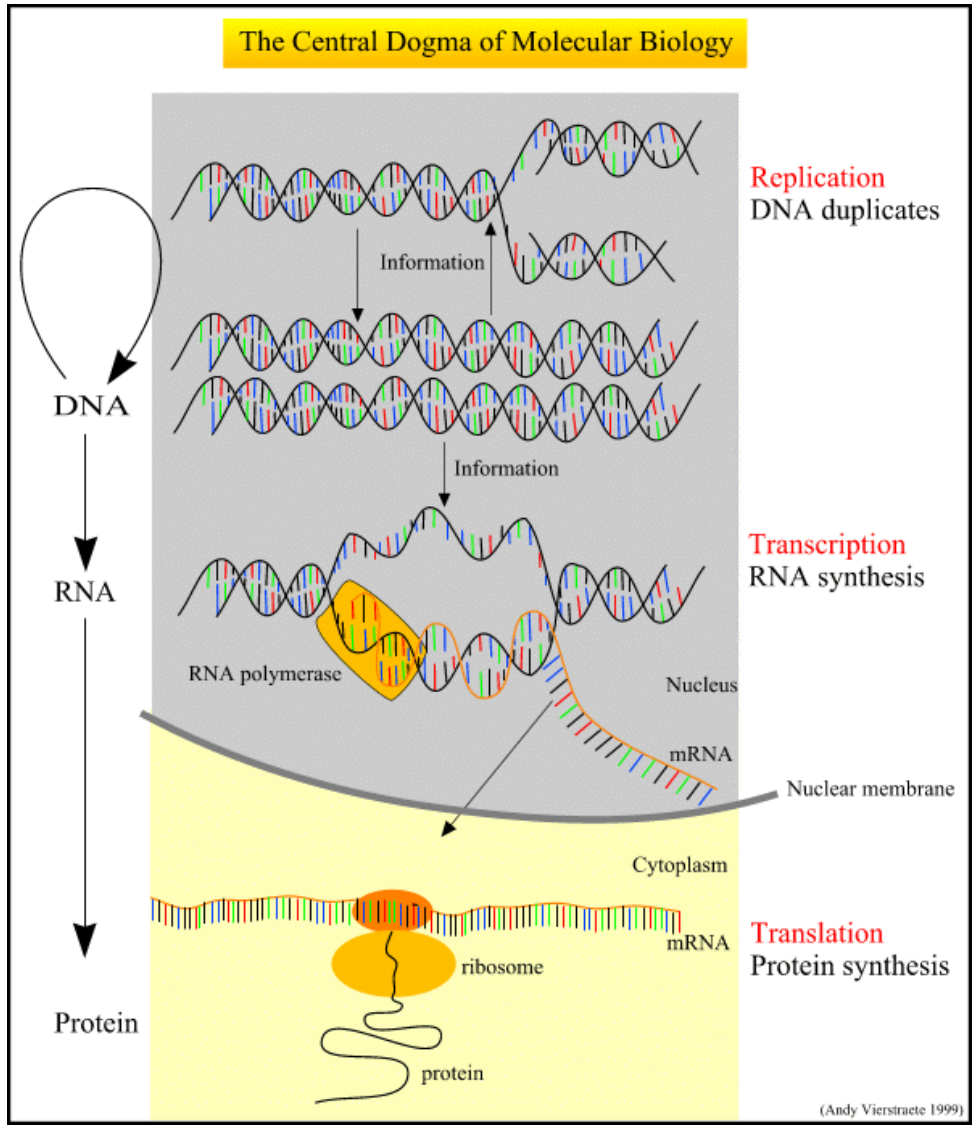
EECS730: Introduction to Bioinformatics

Lecture 14: Microarray

Some slides were adapted from Dr. Luke Huan (University of Kansas), Dr. Shaojie Zhang (University of Central Florida), and Dr. Dong Xu and Trupti Joshi (University of Missouri Columbia)

Review of Central Dogma



Gene Expression

mRNA level

Protein level

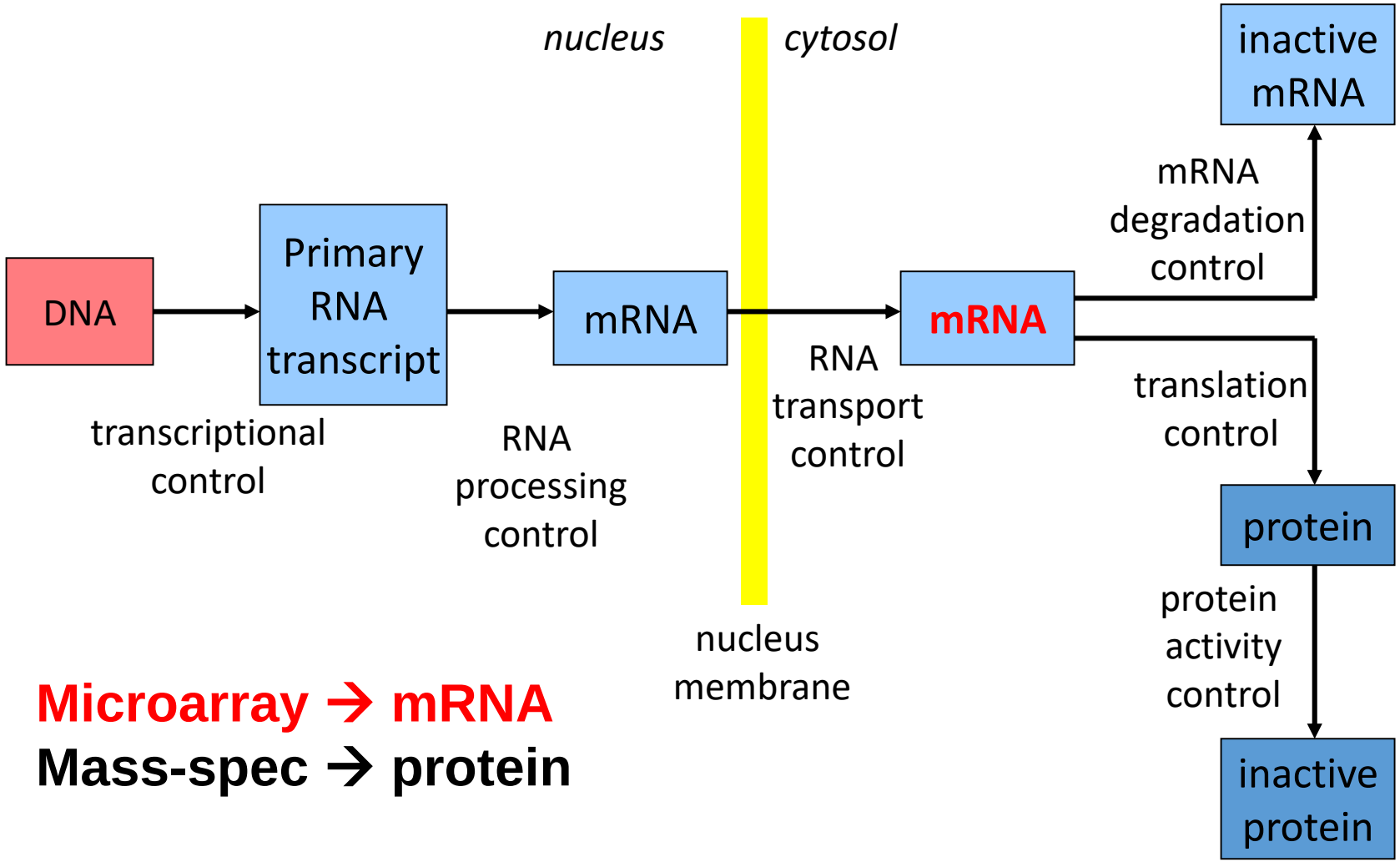
Gene expression profile

- Gene expression profile represents a specific “state” of the cell

Microarray

- Profile the DNA transcription level
- Microarrays measure the activity (expression level) of the genes under varying conditions/time points
- Expression level is estimated by measuring the amount of mRNA for that particular gene
  - A gene is active if it is being transcribed
  - More mRNA usually indicates more gene activity

Information we can measure



Microarray → mRNA

Mass-spec → protein

Applications

- **Gene discovery**
- **Biological mechanisms** (gene regulatory network, etc.)
- **Disease diagnosis** (cancer, infectious disease, etc.)
- **Drug discovery: *Pharmacogenomics***
- **Toxicological research: *Toxicogenomics***
- **Microbial diversity in the environment**
- **…**

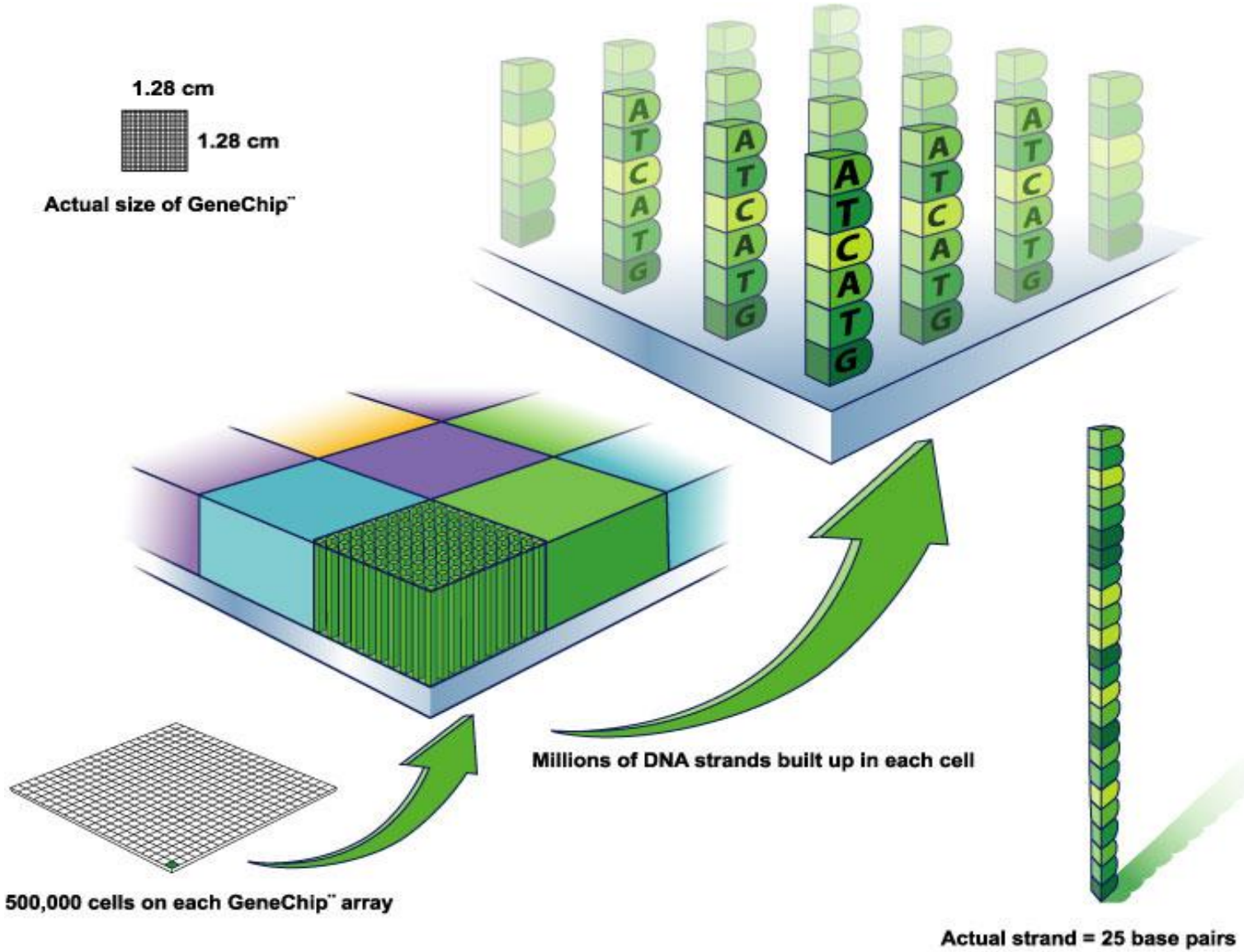
Caveat

- mRNA levels and protein levels are not always directly correlated.
- Translational control
- But we roughly get ~50-70% correlation
- Measuring mRNA is much cheaper!!!

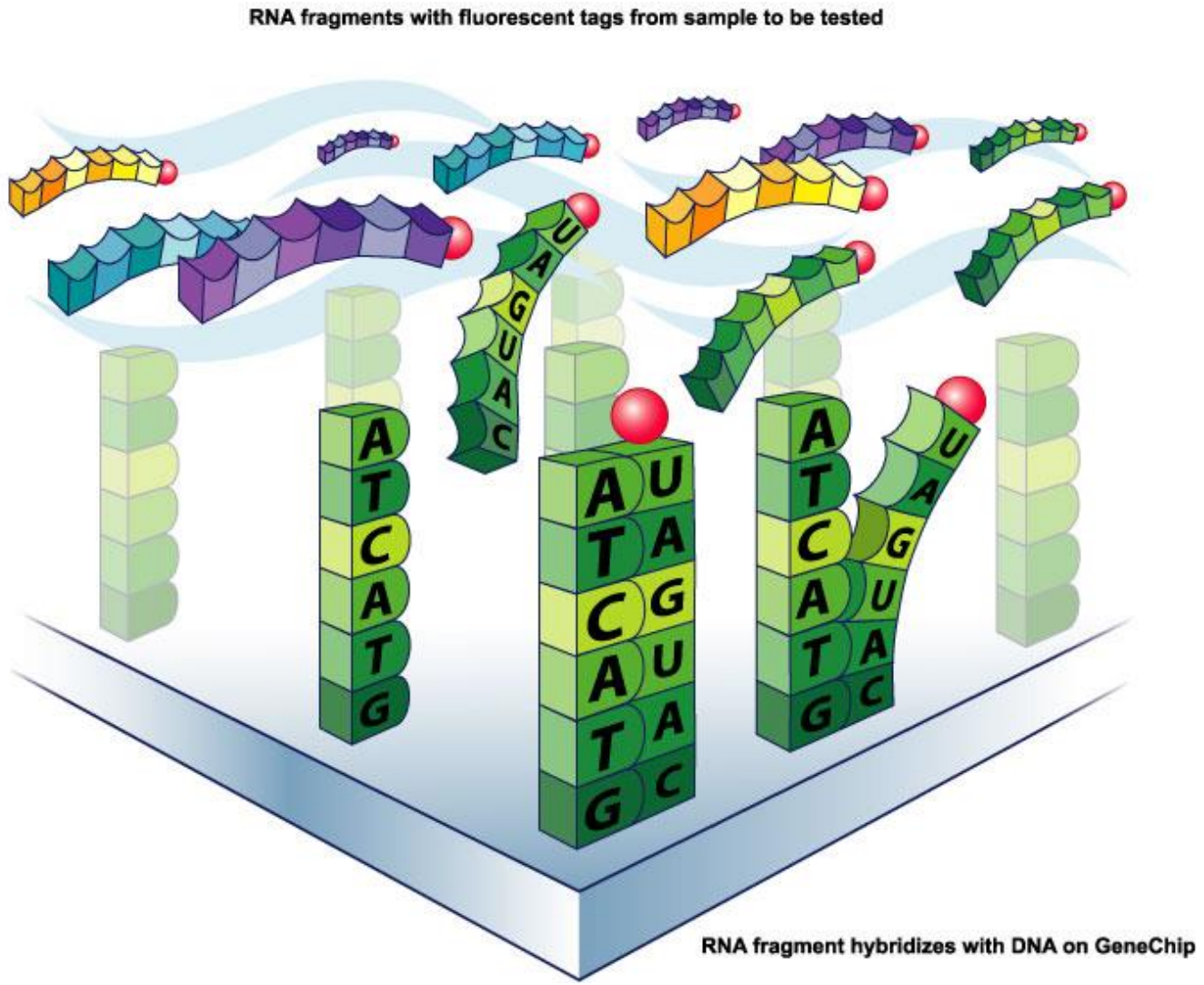
Microarray technology

- Typically a glass slide with cDNA or oligo
- Printed by robot or synthesized by photo-lithography.
- Typical arrays are 25x75 mm. Contains up to 500,000 probed gene fragments.

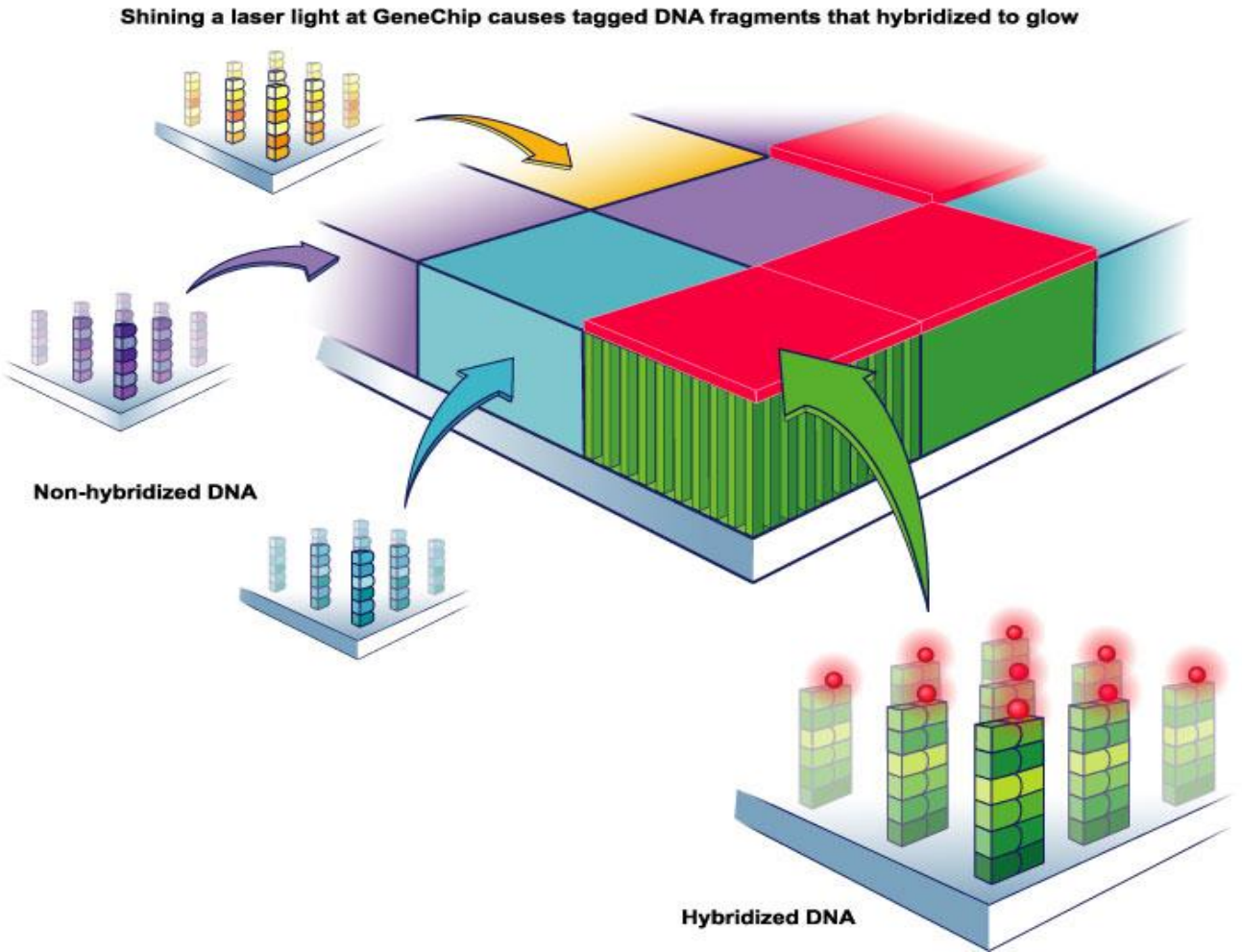
Microarray technology

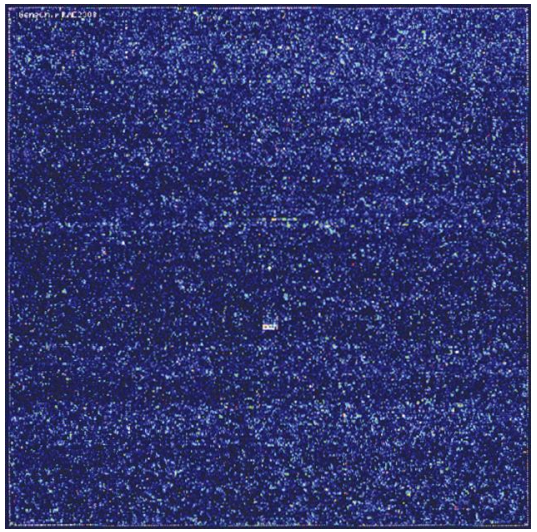


Hybridization



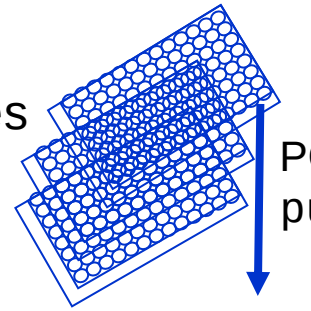
The Chip is scanned





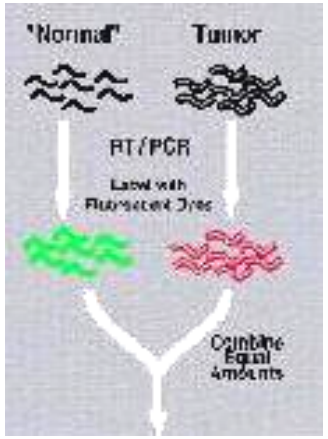
Summary

cDNA clones
(probes)

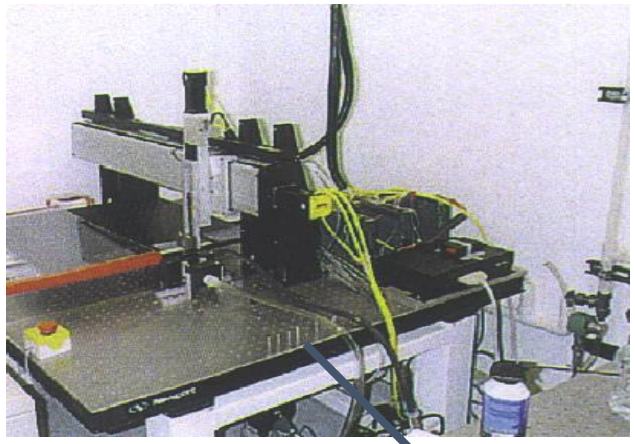


PCR product amplification
purification

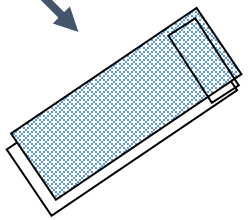
mRNA target)



printing



microarray

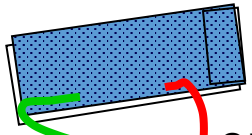


Hybridise
target to
microarray

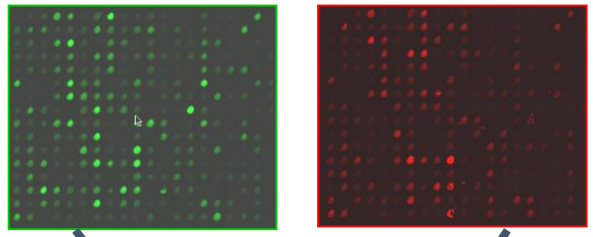
scanning

laser 2

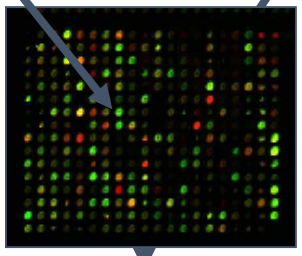
laser 1



emission

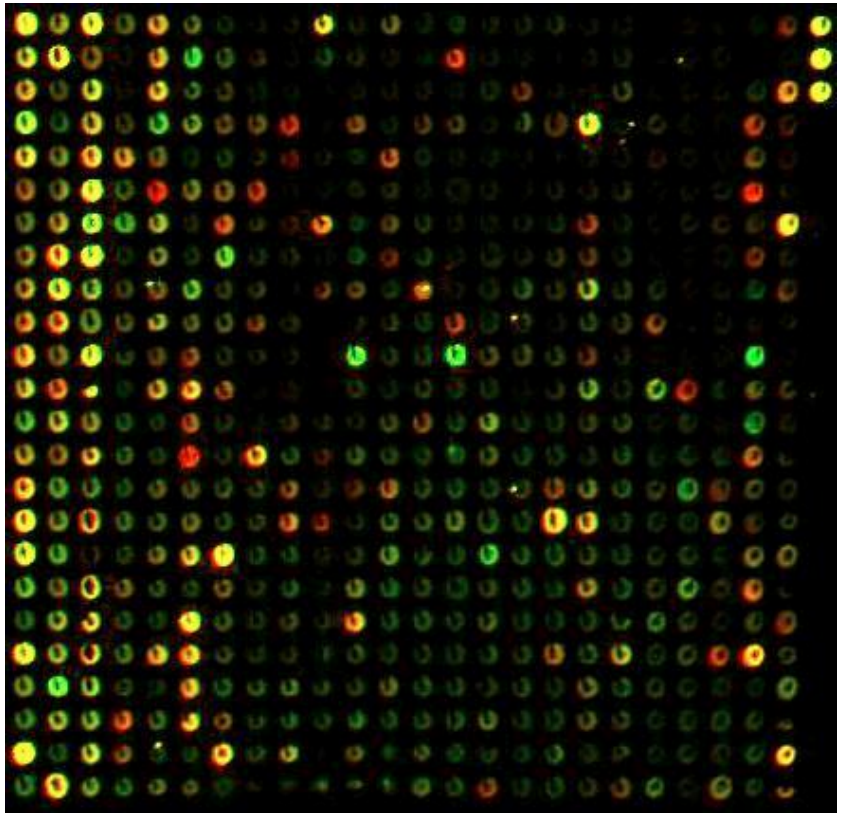


overlay images and normalize



analysis

What does the data look like



- **Green**: expressed only from control
- **Red**: expressed only from experimental cell
- **Yellow**: equally expressed in both samples
- **Black**: NOT expressed in either control or experimental cells

What does the data look like

- begin with a data matrix (gene expression values versus samples)
- Typically, there are many genes (> 10,000) and few samples (~ 10)
- The log2 ratio

$$T_i = \frac{R_i}{G_i}$$

		1	2	3
		log2.t0	log2.t0.5	log2.t2
1		-0.40	-0.91	-1.60
2		-0.99	-0.07	-0.83
3		-0.22	-0.49	-0.28
4		-0.31	-0.01	-0.09
5		-0.48	1.31	0.36
6		-0.38	0.35	0.60
7		-0.41	-0.49	-0.54
8		-0.46	-2.72	-3.16
9		-0.15	0.06	0.13
10		0.12	-0.67	-0.77
11		-0.03	-1.87	-2.58
12		0.31	0.02	-1.64
13		-0.06	-0.22	0.17
14		-0.03	-0.23	0.02
15		-0.12	0.11	-0.01
16		-0.21	-0.66	-0.30
17		-0.40	1.66	1.13
18		-0.58	0.25	0.72
19		-0.77	-0.05	1.11
20		-0.28	0.43	-0.57

log2 transformation

- Logarithm base 2 transformation, has the advantage of producing a continuous spectrum of values and treating up and down regulated genes in a similar fashion.

- The logarithms of the expression ratios are also treated symmetrically, such that
  - genes up regulated by a factor of 2 has a log2(ratio) of 1,
  - gene down regulated by a factor of 2 has a log2(ratio) of −1,
  - gene expressed at a constant level (ratio of 1) has a log2(ratio) equal to zero.

The data needs to be normalized

- Unequal quantities of starting RNA
- Differences in labeling
- Differences in detecting efficiencies between the fluorescent dyes
- Scanning saturation
- Systematic biases in the measured expression levels

Normalization by total intensity

$$N_{total} = \frac{\sum_{i=1}^{Narray} R_i}{\sum_{i=1}^{Narray} G_i}$$

$$G_k' = N_{total} G_k \text{ and } R_k' = R_k$$

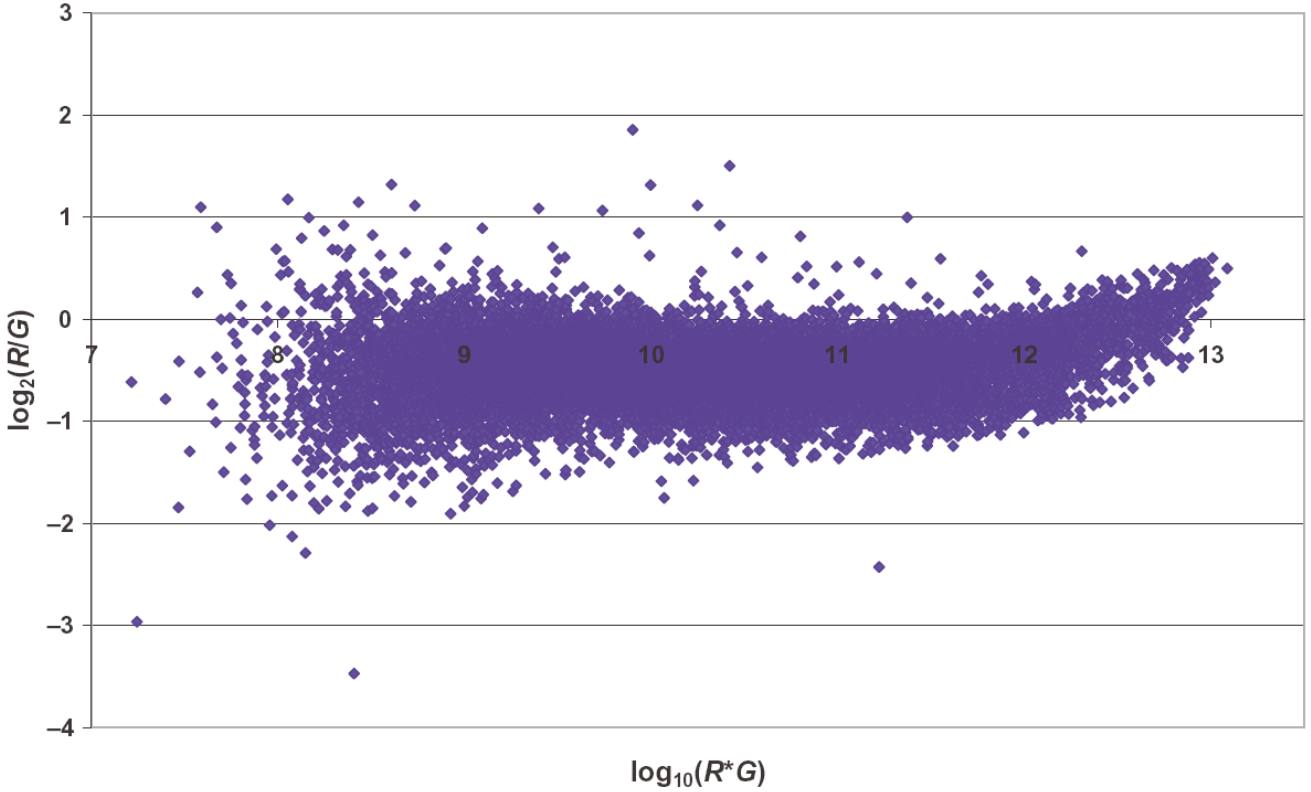
$$T_i = \frac{R_i}{G_i} = \frac{1}{N_{total}} \frac{R_i}{G_i}$$

$$\log_2(T_i') = \log_2(T_i) - \log_2(N_{total})$$

- G_i and R_i are the measured intensities for the *ith* array element
- *Log2(T_i'*) is the normalized value

Normalizing the quenching effect

- quenching (a phenomenon where dye molecules in close proximity, re-absorb light from each other, thus diminishing the signal)
- The log ratio is also dependent of the absolute values of the intensities



Normalize the quenching effect

- You can view the intensity of a given gene is a linear combination of the quenching effect, its true expression change, and measurement error.

- Genes that are adjacent to each other should have similar strength of quenching effect, but we can assume that they have independent expression change and measurement error.

- So we can perform linear regression on a set of adjacent genes in the graph, depict the potential quenching effect, and normalize it.

Normalizing the quenching effect

- LOWESS (locally weighted scatterplot smoothing) regression
- Normalize the value point by point

$$\text{set } x_i = \log_{10}(R_i{}^{*}G_i) \text{ and } y_i = \log_2(R_i/G_i)$$

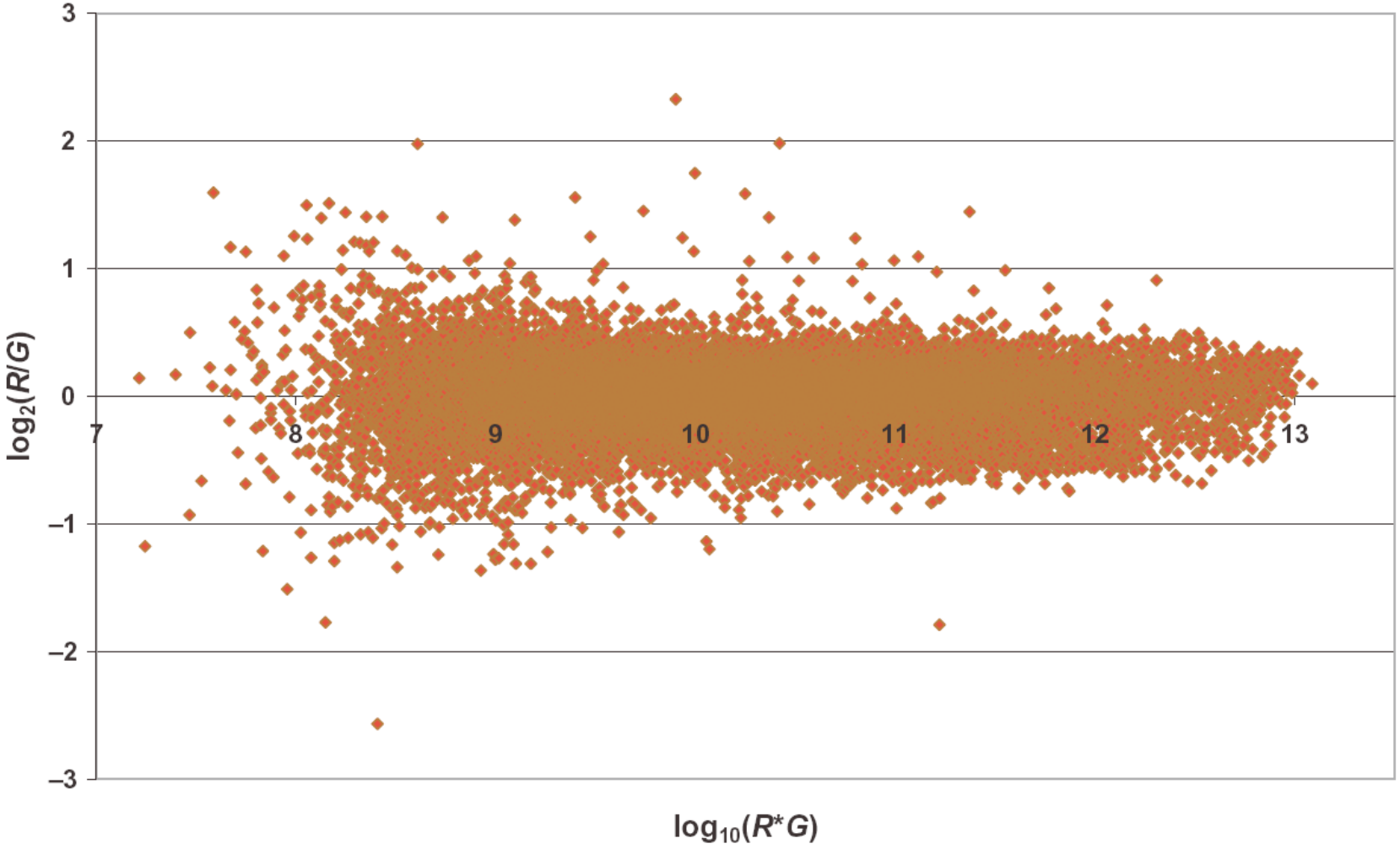
$$\log_2(T_i')=\log_2(T_i)-\mathrm{y}(x_i)=\log_2(T_i)-\log_2(2^{y(x_i)}),$$

or equivalently,

$$\log_2(T_i)=\log_2\left(T_i * \frac{1}{2^{y(xi)}}\right)=\log_2\left(\frac{R_i}{G_i} * \frac{1}{2^{y(xi)}}\right)$$

$$G_i'= G_i * 2^{y(x_i)} \text{ and } \mathrm{R}_i'= \mathrm{R}_i.$$

Normalizing the quenching effect



Statistical analysis of significance

- What can we tell if we find a gene whose expression is upregulated by two fold between two samples?
- Unfortunately… nothing (at least this is what the statistician would argue)
- Biological variation / measure errors…

Is the gene significantly differentially expressed???

- Rank results by confidence with significance metrics (e.g. *p*-value)
- Estimate the false positive (Type I errors) and false negatives (Type II errors)
- Achieve the desired balance of sensitivity and specificity
- Result in a certain amount of flexibility (and arbitrariness) when interpreting significance metrics generated by a test

T-test

- Paired t test:
  - the size of two groups should be same
  - Comparison for organism before or after treatment (before and after heat shock)
- Unpaired t test:
  - the size of two groups do not need to be same
  - Comparison between organisms with treatment or non-treatment
  - Assume equal variance (otherwise use Welch's test)

T-test

Paired T-test

$$T = \frac{X_1 - X_3}{\sqrt{\frac{\sum (d_1 - d_2)^2}{n-1}}}$$

$X_1 - X_3$ = difference between means

$\sqrt{\frac{\sum (d_1 - d_2)^2}{n-1}}$ = standard error

$\sum (d_1 - d_2)^2$ = the variance of the diffence scores for each individual

$n - 1$ = the sample number minus 1

Un-Paired T-test

$$T = \frac{X_1 - X_2}{\sqrt{\frac{S^2 p}{N_1} + \frac{S^2 p}{N_2}}}$$

$X_1 - X_2$ = difference between means

$\sqrt{\frac{S^2 p}{N_1} + \frac{S^2 p}{N_2}}$ = standard error

$S^2 p$ = pooled variance

N_1 = population # of group 1

N_2 = population # of group 2

Example

	Control Group	Experimental Group
Signal R1	3700	4900
Signal R2	4000	5200
Signal R3	4200	4900
Signal R4	3900	5000
Signal R5	4100	4800
Signal R6	4000	4750

Paired T test

Mean1 = 3983 v1 = 5

Mean2 = 4925 v2 = 5

$$\frac{SE(d_1 - d_2)}{5} = \sqrt{228.065} = 45.61$$

$$t = \frac{941.67}{45.61} = 20.65$$

$t_{0.05,10} = 2.228$ as $20.65 > 2.228$ then reject Ho :

P < 0.0001. The differences between the means is greater than 0.

Unpaired T test

Mean1 = 3983 Sum of Squares 1 = 148334 v1 = 5

Mean2 = 4925 Sum of Squares 2 = 128750 v2 = 5

$$S^2 p = \frac{148334 + 128750}{5 + 5} = 29666.8$$

$$S_{mean1-mean2} = \sqrt{\frac{29666.8}{6}} + \sqrt{\frac{29666.8}{6}} = 140.63$$

$$t = \frac{3983 - 4925}{140.63} = -6.70$$

$t_{0.05,10} = 2.228$ as $6.7 > 2.228$ then reject Ho :

P < 0.0001. The two means are not the same.

Wilcoxon Signed-Rank Test

- Use if sample is not distributed normally
- Similar to paired T test but non-parametric
- Rank the absolute difference between arrays, i.e. $|x_{2,i} - x_{1,i}|$.
- If the difference between two pairs is 0, the value is not used
- If the difference is identical between 2 pairs, the average rank of the two groups is used
- Compute W value using $W = \sum_{i=1}^{N_r} [\operatorname{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]$, the sum of the signed ranks
- Look up Wilcoxon Table for significance value

Mann-Whitney Test

- Use if sample is not distributed normally
- Similar to non-paired T test but non-parametric
- Use the rankings of the numerical values instead of variance
- Take the less U value and look up table for significance

$$U = \frac{n_1(n_1+1)}{2} - R_1$$

$$n_1 = \#\text{ of individual in group 1}$$

$$R_1 = \text{sum of the ranks for group 1}$$

Mann-Whitney Test

	Control Group	Experiment Group	Control Rank	Experiment Rank
Signal R1	4500	3700	7	9
Signal R2	5200	3300	2	11
Signal R3	4700	4600	4	6
Signal R4	5500	3500	1	10
Signal R5	5000	3900	3	8
Signal R6	4650		5	

n1 = 6; n2 = 5; N = 11; R1 = 22; R2 = 44

Ranks of N are assigned in either lowest to highest or vice versa.

$$U = \frac{(6)(5)+(6)(7)}{2} - 22 = 29$$

$$U' = \frac{(6)(5)+(5)(6)}{2} - 44 = 1$$

Measure of performance of prediction

	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_A)	Total
Test is declared significant	V	S	R
Test is declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m is the total number hypotheses tested
- m_0 is the number of true null hypotheses, an unknown parameter
- $m - m_0$ is the number of true alternative hypotheses
- V is the number of false positives (Type I error) (also called "false discoveries")
- S is the number of true positives (also called "true discoveries")
- T is the number of false negatives (Type II error)
- U is the number of true negatives
- $R = V + S$ is the number of rejected null hypotheses (also called "discoveries", either true or false)

In m hypothesis tests of which m_0 are true null hypotheses, R is an observable random variable, and S, T, U, and V are unobservable random variables.

-wikipedia

Multiple testing

- Statistical hypothesis testing is based on rejecting the null hypothesis if the likelihood of the observed data under the null hypotheses is low.
- If multiple comparisons are done or multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a Type I error) increases.
- The development of "high-throughput" sciences, such as genomics, allowed for rapid data acquisition.

-wikipedia

Correction

- Bonferroni Correction:
  - Reject the null hypothesis if $p_i \leq \frac{\alpha}{m}$

- False discovery rate:
  - Benjamini–Hochberg procedure

1. For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha.$
2. Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k.$

  - And many more other correction methods…

-wikipedia