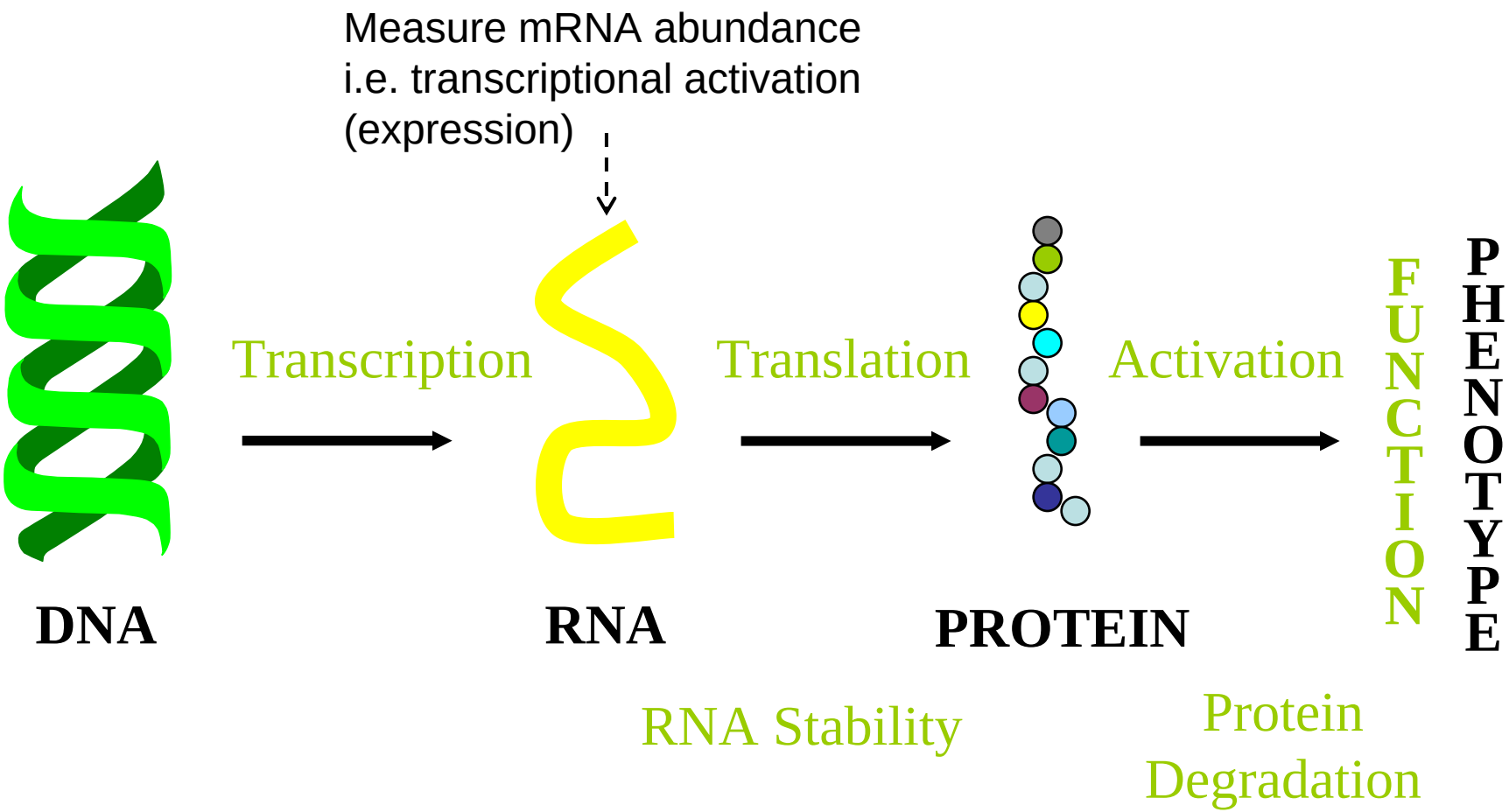
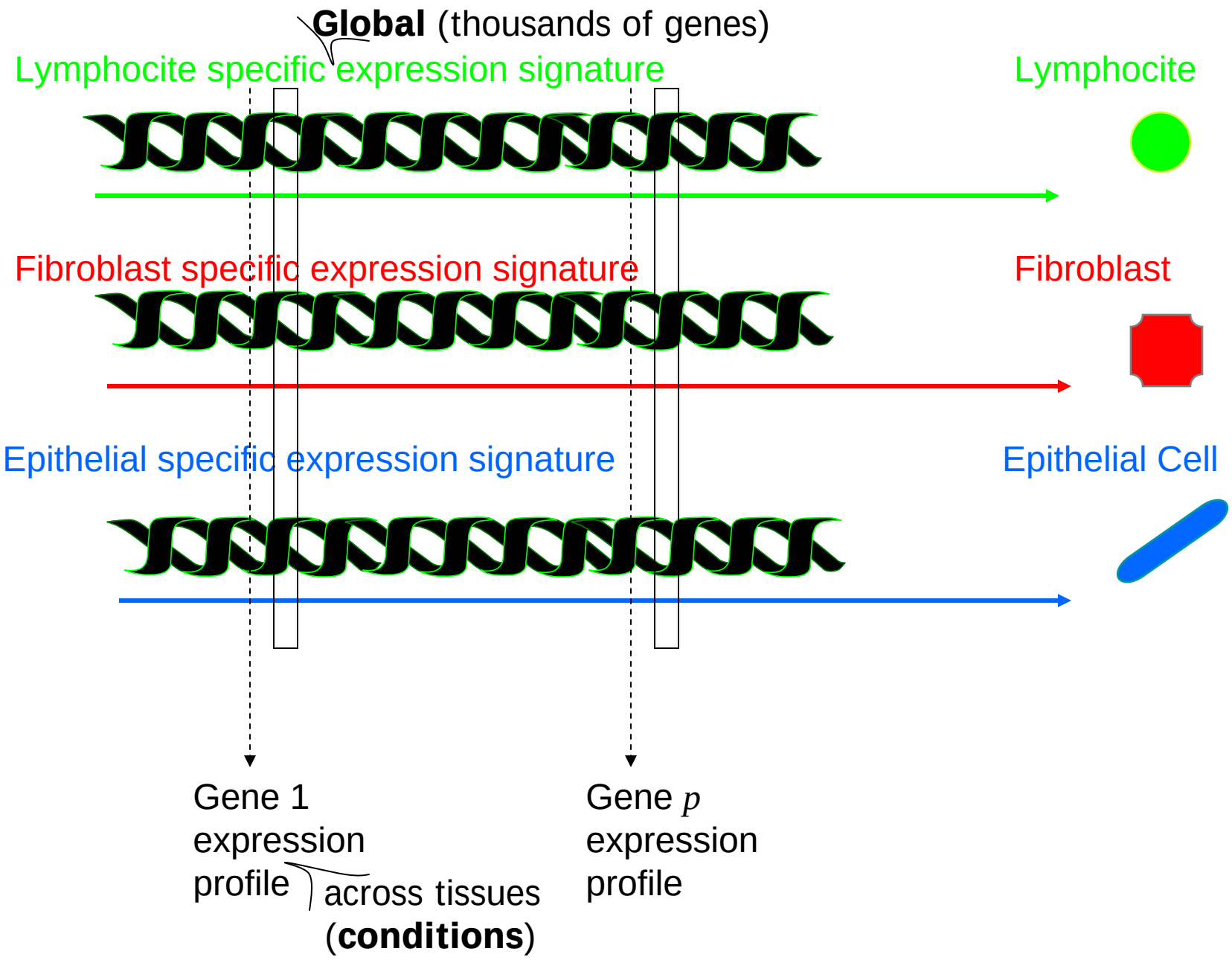
Introduction to Microarrays



Assumption: changes in expression relate to changes in function (phenotype)


Global (thousands of genes)
Lymphocite specific expression signature
Lymphocite
Fibroblast specific expression signature
Fibroblast
Epithelial specific expression signature
Epithelial Cell
Gene 1 expression profile
Gene p expression profile
across tissues (**conditions**)

Many technologies allow to measure transcriptional activation …

… The Microarray revolution: Global gene expression

Although “noisy” in many and complicated ways microarrays allow us to measure transcriptional activation for **thousands** of genes simultaneously.

By doing so on appropriate sets of conditions, e.g.

- different cancer types
- different tissues of an organism
- points in a time course

we can address complex questions such as:

- What are the global expression signatures of cancers? Can they inform cancer taxonomies? Help in diagnosing?
- What are the global expression signatures of tissues, and can they help in understanding differentiation or failures thereof?
- What genes are involved in certain processes (e.g. cell cycle) or the reactions to certain stimuli/offenses, and what are their typical expression patterns?

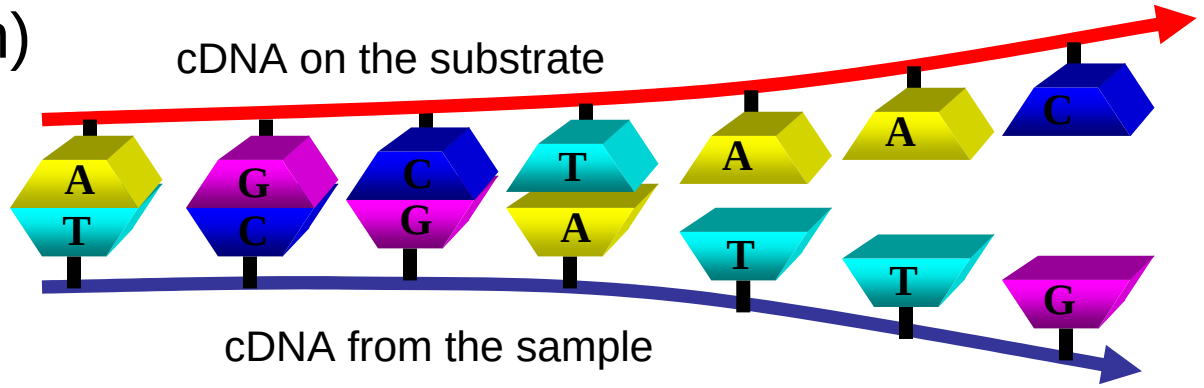
Even more complex questions:

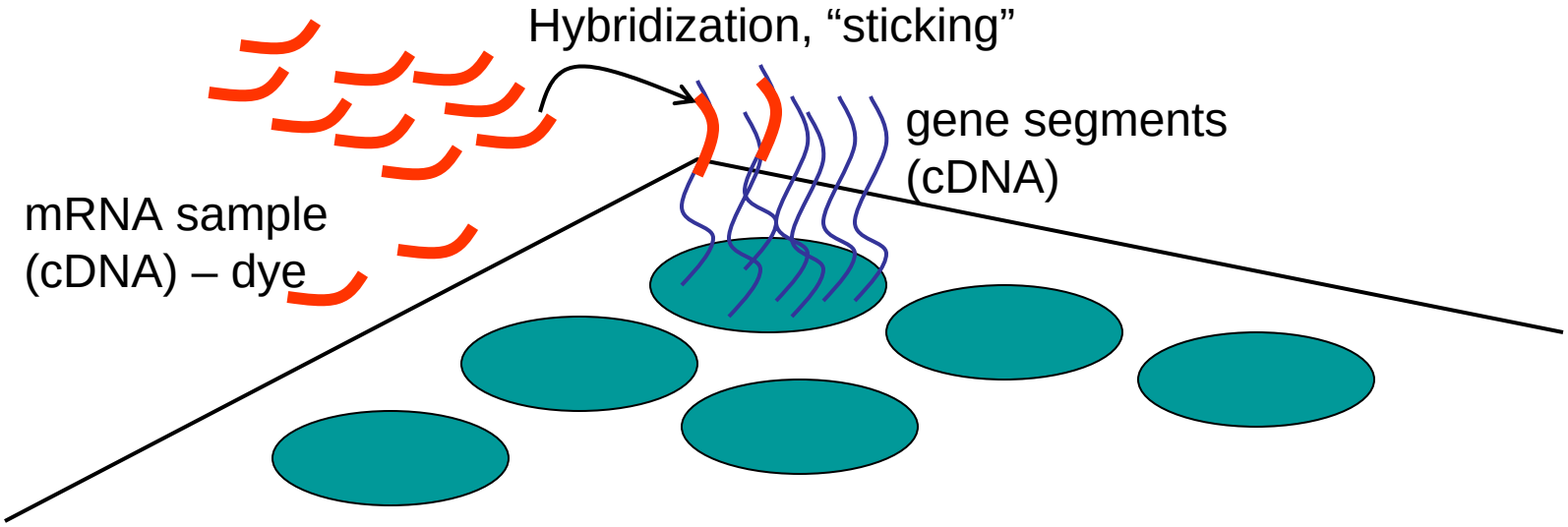
- What are the relationships among genes?
- Can we use expression profiles to identify genes whose products perform similar or related functions?
- Can we use expression profiles to identify genes that are co-regulated?
- Can we use expression profiles to infer regulatory relationships? (genes acting upon one another through their products)

More generally, reconstructing modular networks among genes (the genome-level “contraction” of complex pathways involving extra-nuclear and/or extra-cellular signaling and interactions).

For these questions, in particular for those relative to networks, expression information needs to be merged with product, sequence, conservation, and other types of information (appropriately “mined” and formatted).

General idea (cartoon)





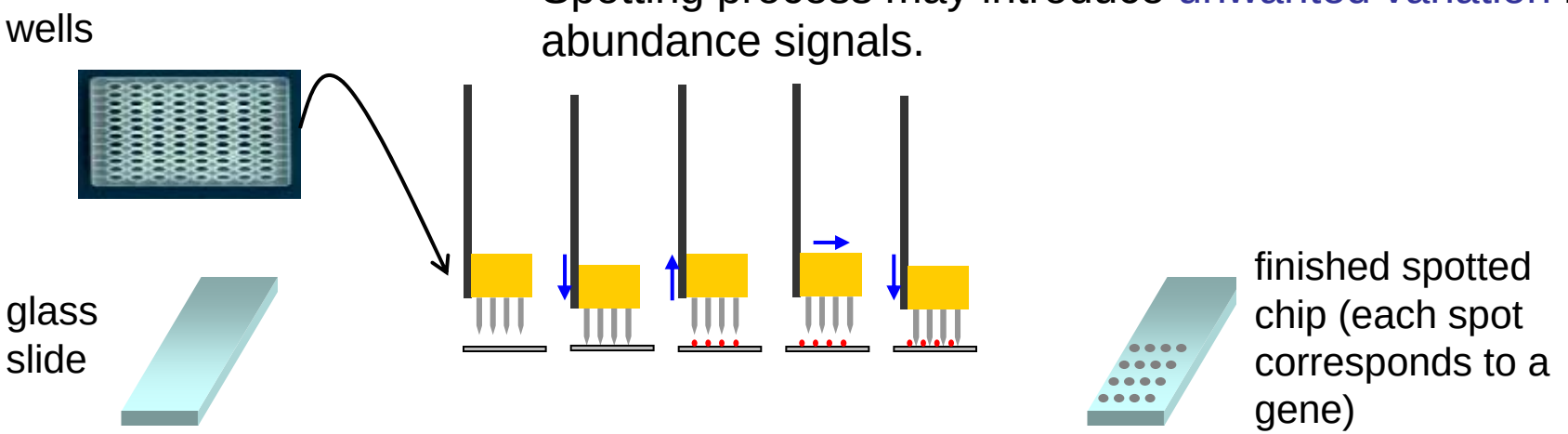
a physical substrate with thousands of “spots”/ “cells”

scan and process image(s) to obtain abundance signals

cDNA "spotted" 2-color arrays
(in house)

Adhere micro-quantities of selected cDNA to a substrate, using microarrayer robot (pins spot different sectors on the chip; within each sector, there is a spotting order)

Spotting process may introduce unwanted variation in abundance signals.

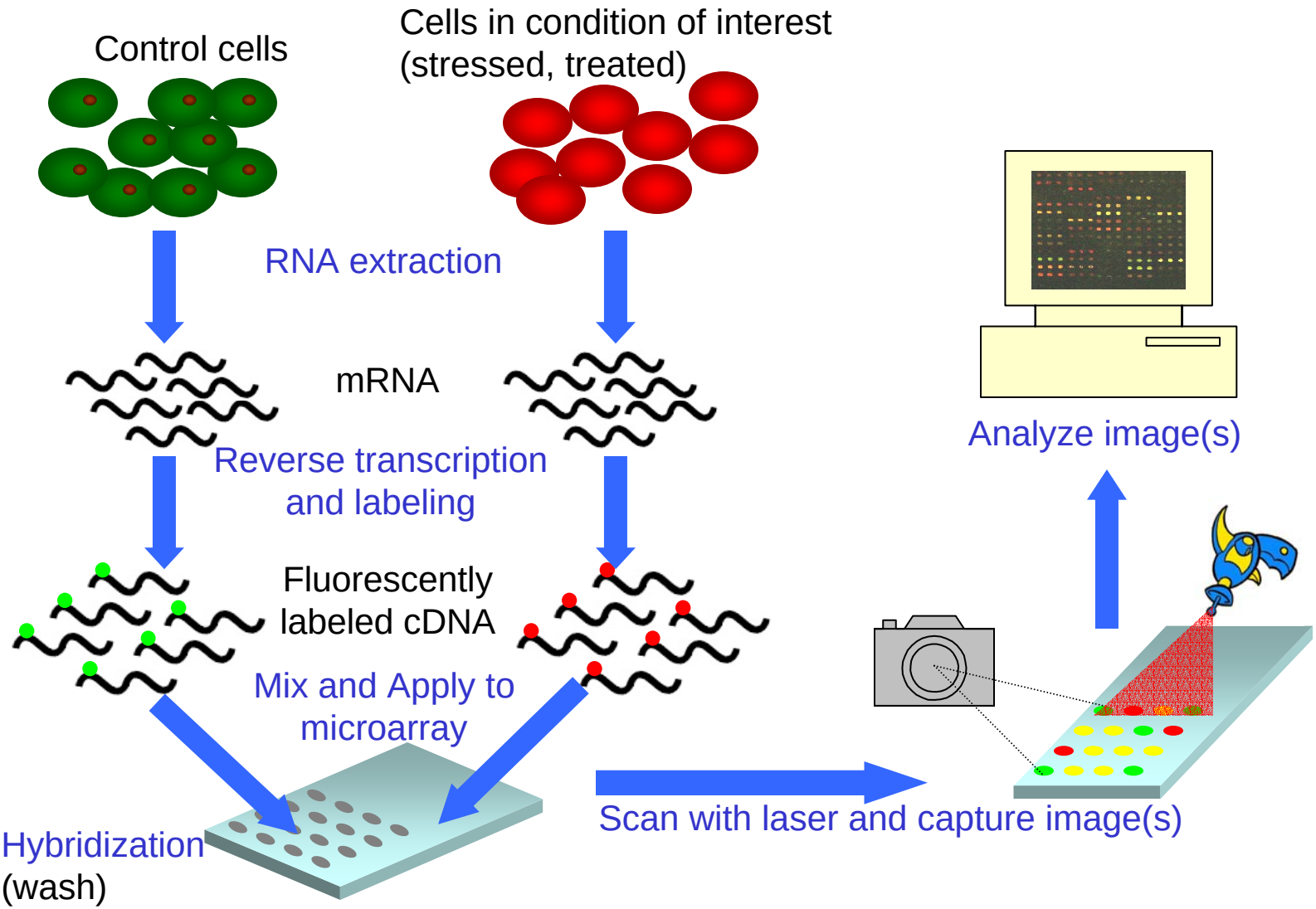


Some spots are replicates (cDNA from same gene).

Some spots are controls, i.e contain cDNAs from genes whose transcript abundance should not be subject to systematic variation across the conditions of interest:

- Housekeeping genes
- Genes from other organisms, or synthetic (spiked controls)

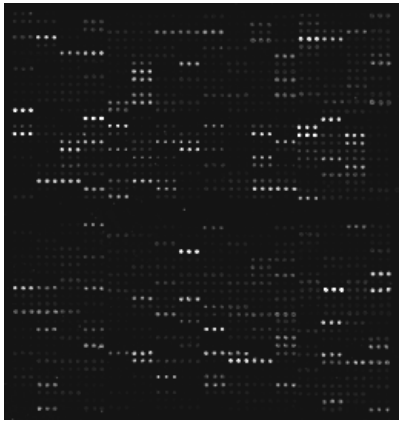
Issue: where are replicate and control spots located?



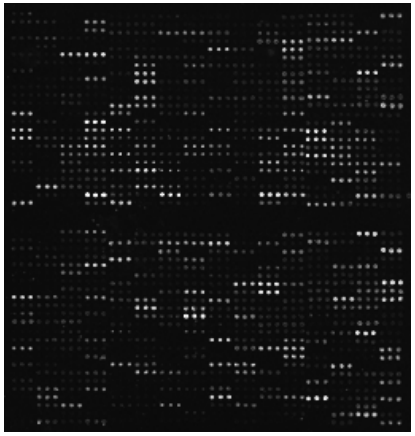
- is there enough material in each spot for any amount of cDNA to hybridize?
- cross or spurious hybridization?

- hybridization conditions, spatial in-homogeneity on chip
- scanning process
- different chemical and optical properties of dyes may introduce unwanted variation in abundance signals.

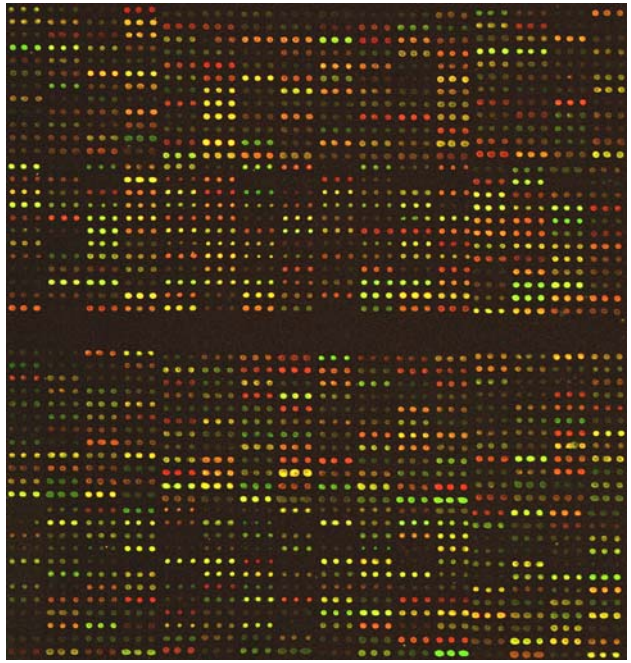
Cy3 channel (control)



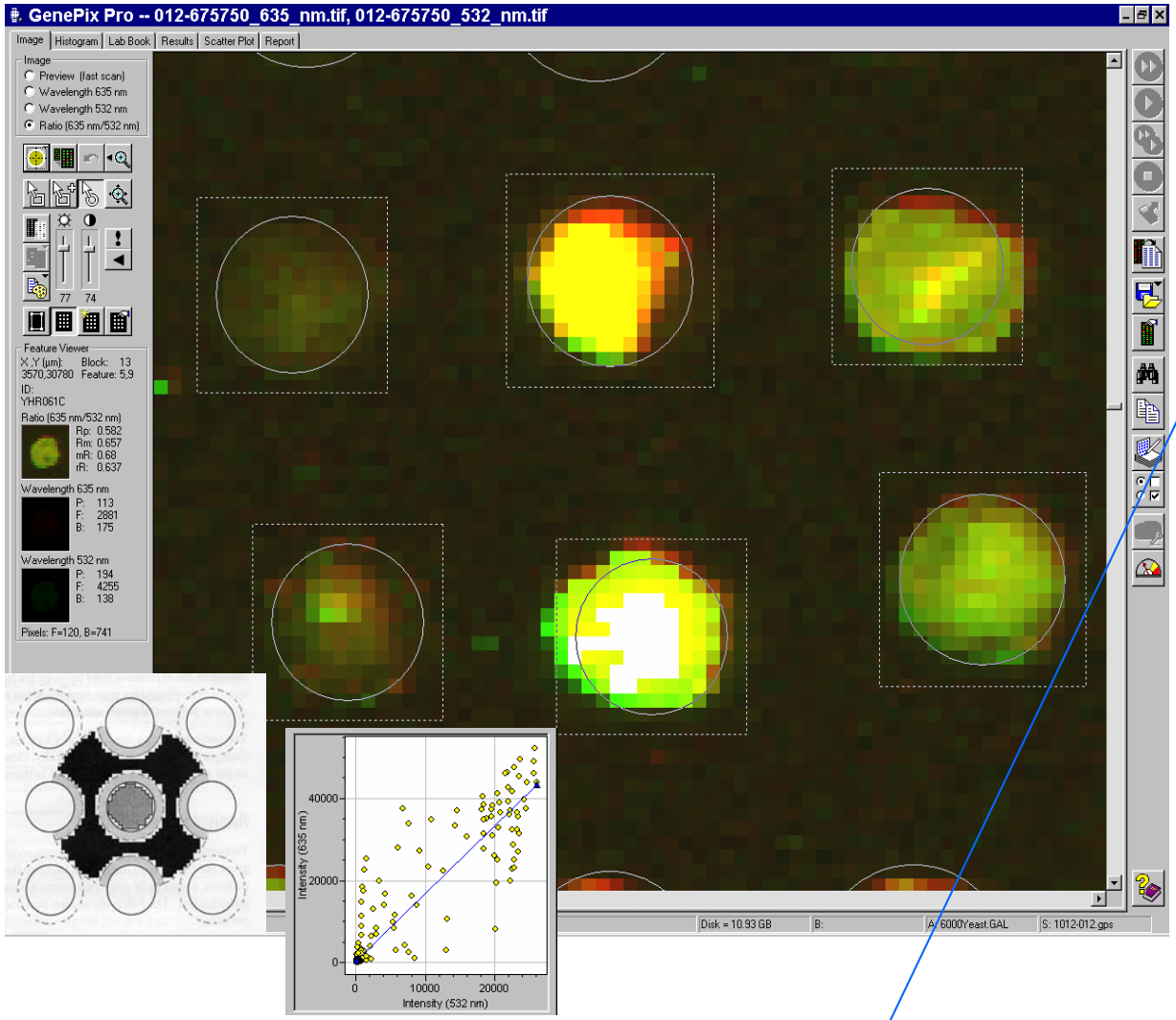
Cy5 channel (treatment)



False-color overlay



- Output of scanning: two monochromatic images of the array, one for each channel.
- Relative gene expression can be “visualized” by coloring the two images (green and red) and overlaying them.



The basic numbers: produced through special software (image analysis and synthesis).

- **Abundance signals**
- Variability measurements and quality indicators (flags)

For instance, summarize through two numbers per spot:

Red = Average Foreground – Average Background pixel intensities on Red

Green = Average Foreground – Average Background pixel intensities on Green

- defining foreground and background areas.
- what pixel “summary” measurements to use

Example of GenePix output

Block	Column	Row	ID	F635 Mean	B635 Mean	F532 Mean	B532 Mean
8	9	20	3xSSC	707	626	641	527
15	10	20	3xSSC	941	851	755	703
2	5	1	B.subtilis_Dap	14388	978	15163	876
2	4	1	B.subtilis_Trp	6363	784	8212	678
4	8	1	BAR1	1786	557	1761	493
2	7	1	BL21_1	32035	872	40969	739
2	9	1	BL21_2	2945	572	4089	490
2	11	1	BL21_3	2645	603	3425	504
2	14	1	Cy3_YDR363W	20535	716	27725	597
12	14	20	Cy3_YDR363W	28488	9479	35953	8893
8	14	20	Cy3_YDR363W	25492	4466	27598	4601
2	6	1	DH5a_1	53830	1038	50606	927
2	8	1	DH5a_2	5016	690	6058	568
2	10	1	DH5a_3	2673	596	3361	502
16	16	20	empty	798	792	645	642

Filter	X	Y	Dia.	F635 Median	F635 SD	B635 Median	B635 SD
	10820	15850	180	24673	11055	678	196
sat. 532nm	19570	29300	160	27768	21977	888	19816
	9930	15850	140	1895	1933	574	118
	19940	33790	160	789	113	778	130

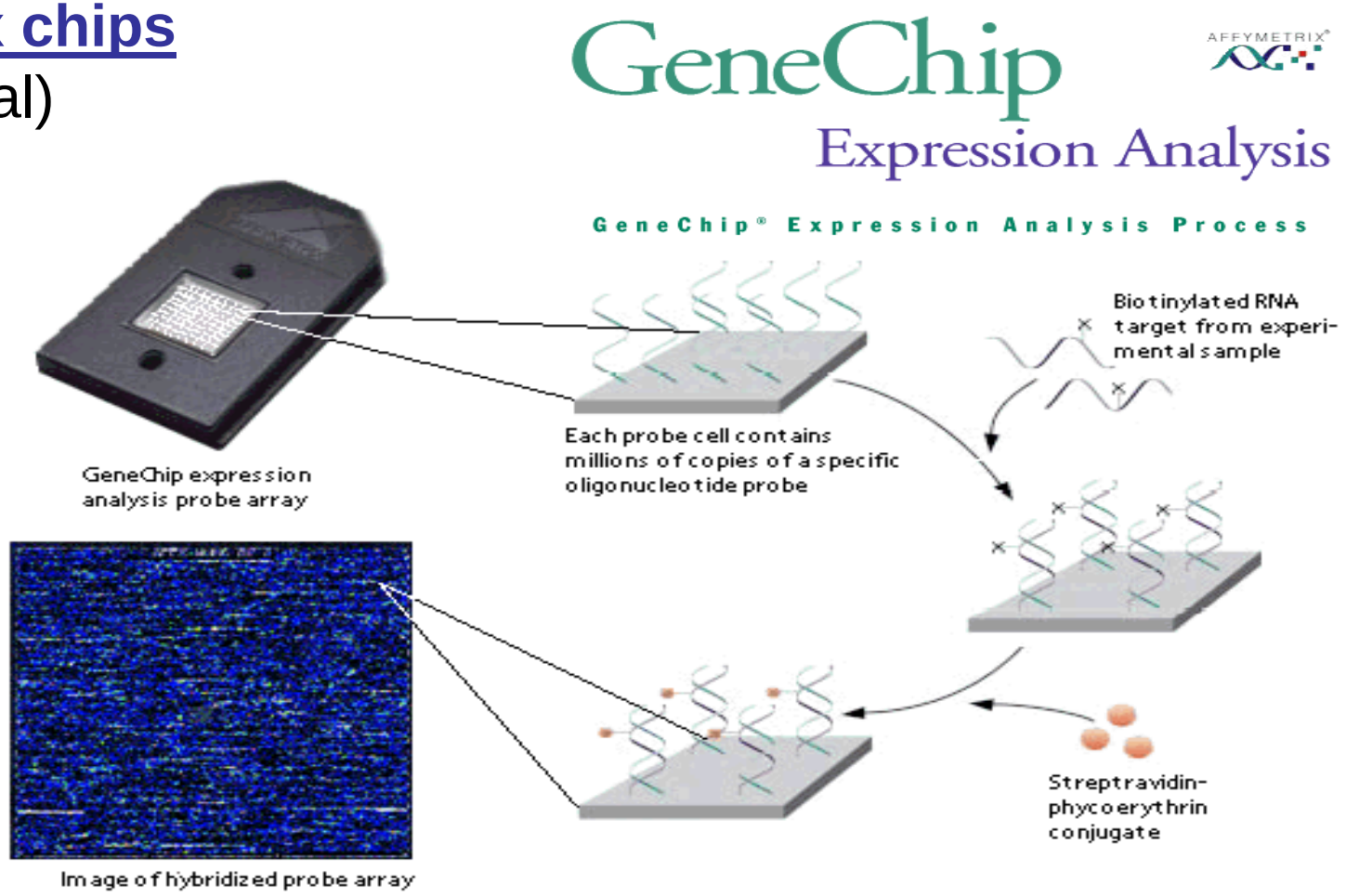
% > B635+1SD	% > B635+2SD	F635 % Sat.	F532 Median	F532 SD	B532 Median	B532 SD	% > B532+1SD
50	25	0	626	177	520	113	48
36	23	0	724	164	691	156	30
85	79	0	15921	11517	688	631	85
77	69	0	7613	6790	608	357	78

% > B532+2SD	F532 % Sat.	Ratio of Medians	Ratio of Means	Median of Ratios	Mean of Ratios	Ratios SD	Rgn Ratio
23	0	0.821	0.744	0.734	1.18	1.434	0.169
7	0	2.03	1.578	1.452	3.677	6.068	0.2
80	0	0.89	0.939	0.935	1.345	1.728	0.901
67	0	0.736	0.741	0.779	1.48	3.302	0.695

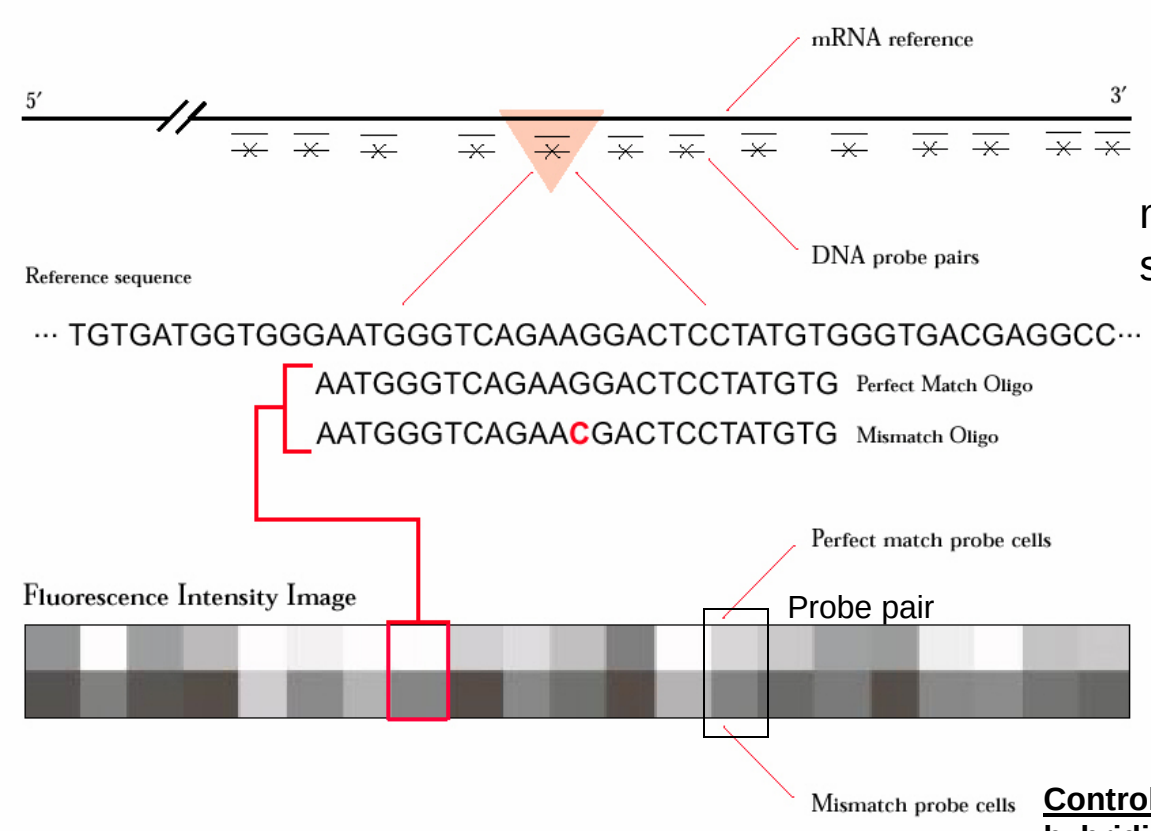
Rgn R^2	F Pixels	B Pixels	Sum of Medians	Sum of Means	Log Ratio	F635 Median - B635	F532 Median - B532
0.05	52	340	193	211	-0.285	87	106
0.034	52	324	100	165	1.022	67	33
0.876	256	1049	28784	28064	-0.169	13551	15233
0.917	316	1132	12158	13238	-0.443	5153	7005

F635 Mean - B635	**F532 Mean - B532**	Flags
90	121	0
101	64	0
13589	14475	0
5634	7604	0

Affymetrix chips
(commercial)



Similar process, different substrate and **one** sample on each chip…

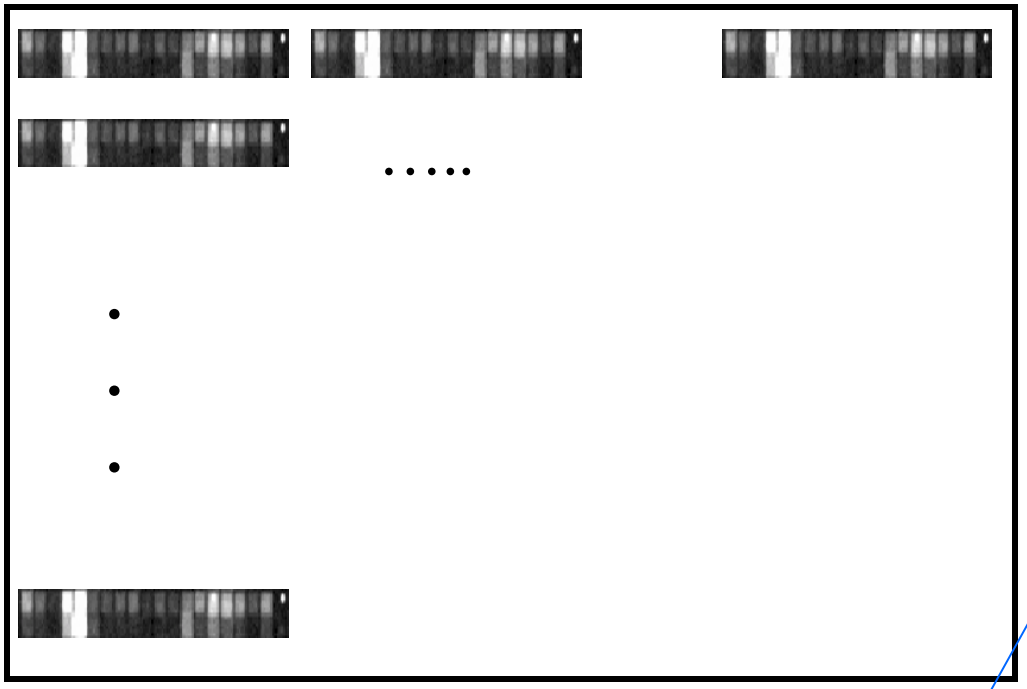


multiple (16, 20)
short segments (25-meres)

Probe set (gene)

Control for spurious hybridization

(again, replicate and control probe sets on chip)

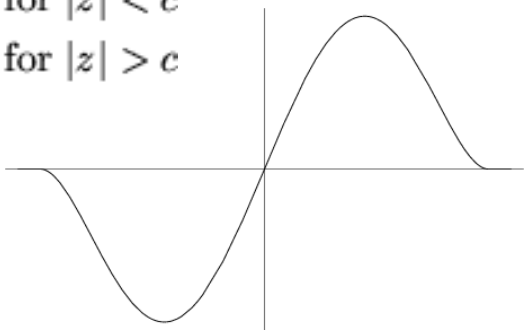


The basic numbers: produced through special software (image analysis and synthesis).

- **Abundance signals**
- Variability measurements
- "Present/Absent" calls

- Summarize pixel intensities per cell
- Correct for background
- Summarize cell measurements in set. For instance, Tukey bi-weight averaging of PM-MM, robust

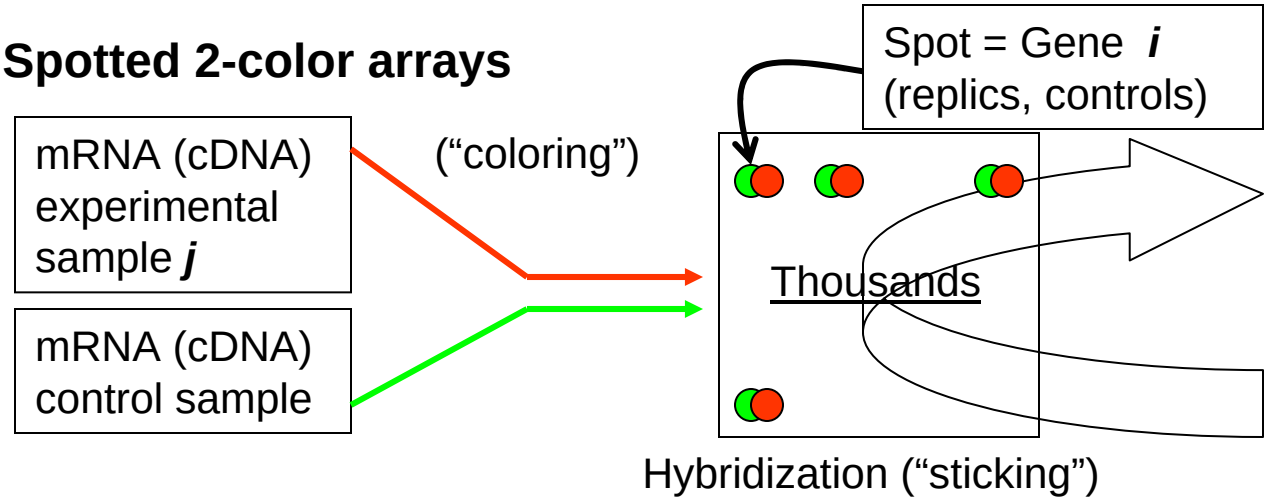
$$\psi(z) = \begin{cases} z\left(1 - \frac{z^2}{c^2}\right)^2 & \text{for } |z| < c \\ 0 & \text{for } |z| > c \end{cases}$$



Note: use of MM's highly controversial

Summary

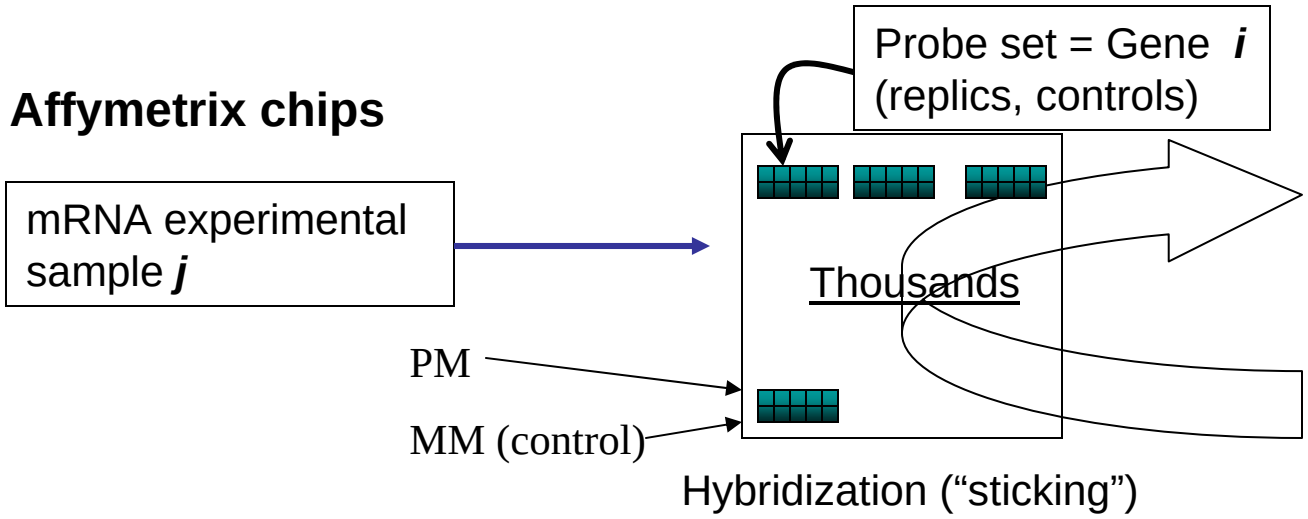
Spotted 2-color arrays



Scan, process image (summarize pixel intensities, correct for background, etc) on two channels

R_{ij} , G_{ij}

Affymetrix chips



Scan, process image (summarize pixel intensities in cell and cell readings in set, correct for background, etc)

B_{ij}

j ranges over conditions (e.g. time points, treatments, types) and possibly replicates for each

Questions

Experimental design

(plus knowledge of sources of unwanted variation)

Experiment

Image analysis

Basic numbers

R_{ij} , G_{ij} or B_{ij}

Dots: Statistics!

Preprocessing

Data matrix: rows=genes, columns=chips

$$\begin{bmatrix} X_{1,1(1)} & \cdots & X_{1,1(R1)} \\ \vdots & \ddots & \vdots \\ X_{p,1(1)} & \cdots & X_{p,1(R1)} \end{bmatrix} \cdots \cdots \begin{bmatrix} X_{1,T(1)} & \cdots & X_{1,T(RT)} \\ \vdots & \ddots & \vdots \\ X_{p,T(1)} & \cdots & X_{p,T(RT)} \end{bmatrix}$$

$i = 1 \ldots p$ (*genes*), $j = 1 \ldots T$ (*conditions*), $r = 1 \ldots Rj$ (*replicates*)

(N chips)

Analysis

Answers

Comparable numbers, on a scale and format appropriate to the questions and the data analysis tools one intends to use. Possibly no missing values, possibly a reduced set of genes.

Preprocessing the data:

- Normalization

  Ensure comparability of different sets of expression measurements, and mitigate effects of unwanted variation. Sources, control with Design.

- Imputing missing values

- Other preprocessing steps

  1. Further improve comparability of measurements across experimental conditions and/or across genes, e.g. centering and standardization.
  2. Further decrease the effect of unwanted sources of variation by reducing noise and systematic errors, e.g. low-dimensional reconstructions.
  3. Eliminate “inert bulk” that can affect detection of interesting signals by standard methods; filtering out genes with negligible expression in all conditions (e.g. absent calls in affy), or negligible variation across them.

Analyzing the data:

- Identifying differentially expressed genes
  Computational approaches to significance assessment
  Multiple comparisons

- Investigating aggregate features of global expression data (constitutive or typical expression patterns, groups of genes with similar profiles, groups of conditions with similar signatures)
  Dimension Reduction
  Clustering (unsupervised classification)

- Investigating expression correlates of sample classifications (cancer types, tissues), or quantitative responses (relapse times, levels of chemicals)
  Classification (supervised) & Regression

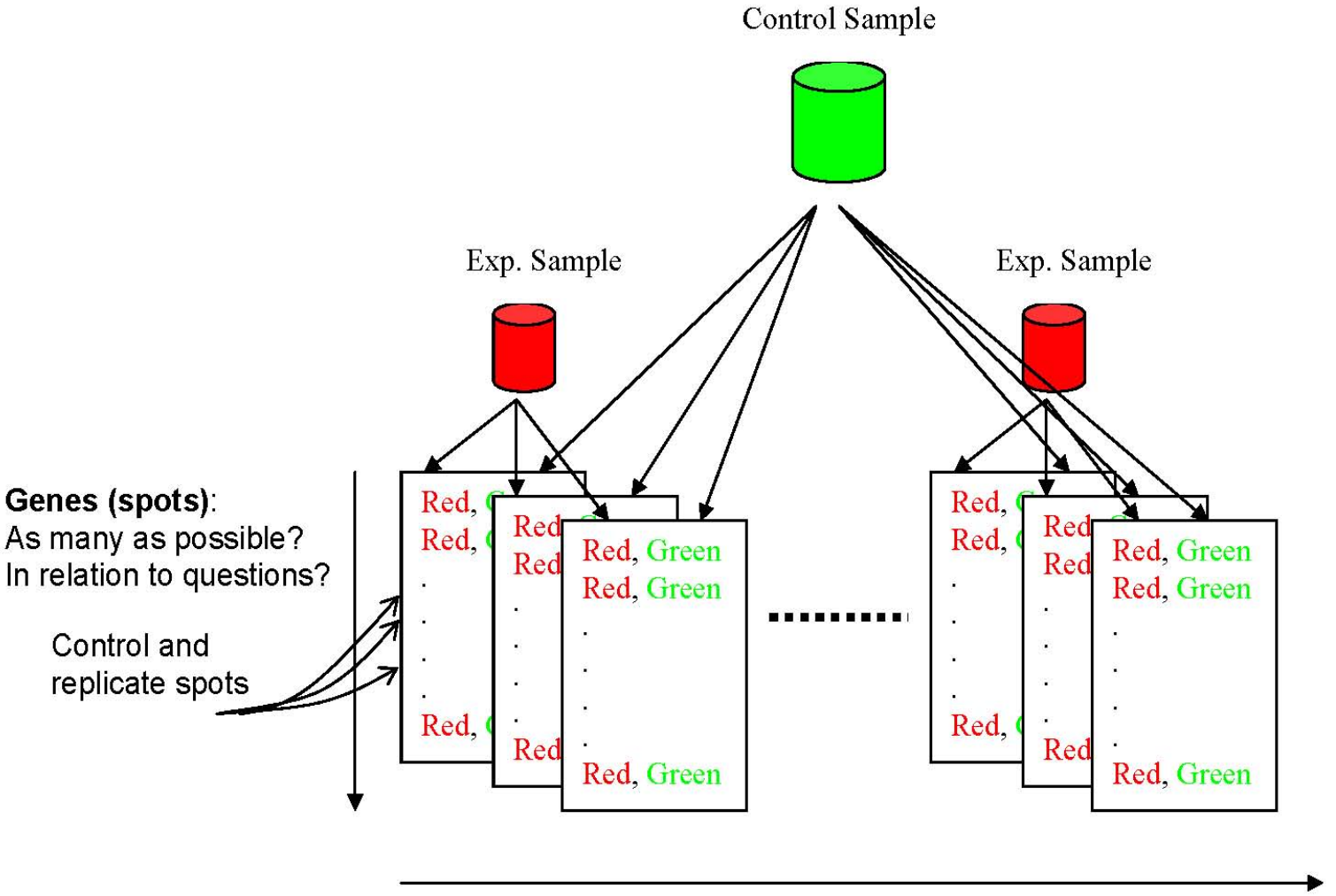
- Integrating expression with genomic sequence, alignment (conservation), protein and/or other information

- Investigating gene networks

(many others)

Designing an experiment…

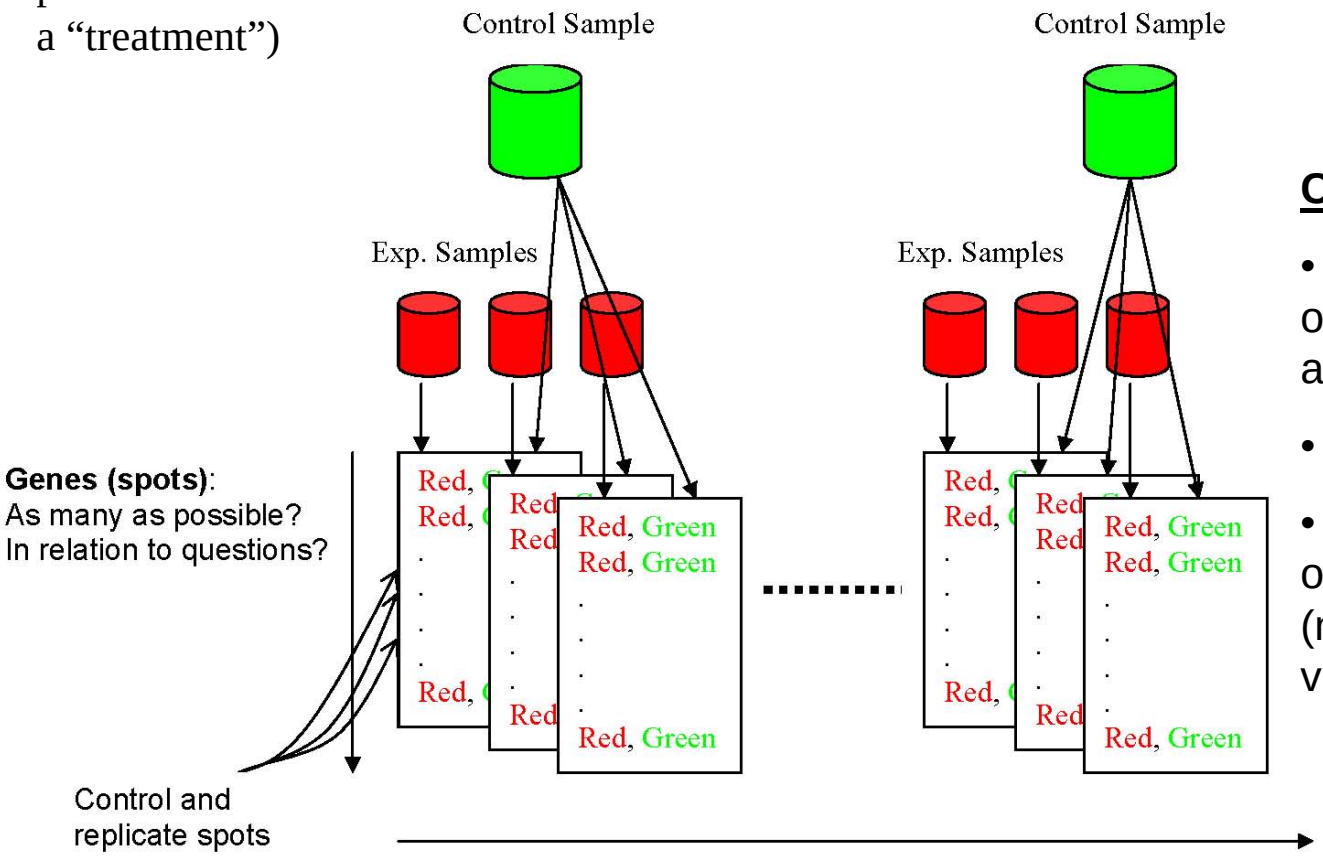
Reference condition (green): technical replicates on each slide.



Experimental conditions (red): In relation to the questions. Technical replicates on each group of slides.

(e.g. observing a process in time w/out a “treatment”)

Reference conditions (green): To match experimental. Technical replicates on each group of slides



Experimental conditions (red): In relation to the questions. Biological replicates on each group of slides.

Controls and Replicates:

- obtain measurements that ought to be comparable across experimental conditions
- assess reproducibility, and
- evaluate unwanted sources of (systematic) variability and (non-systematic) error variance; **NORMALIZATION**.

What sources can be evaluated, and how effectively replicates are used to do so, depends on **EXPERIMENTAL DESIGN**.