Introduction, generalities and a brief summary on Spotted arrays

The questions: What is global expression data useful for?

Microarrays and other technologies allow us to simultaneously measure expression (i.e. a proxi for "activation level") for thousands of genes at a time: *Global* expression data.

Collecting these measurements on an appropriate set of conditions, one can address questions like:

- What genes are involved in the differentiation among certain cell types?

(e.g. normal vs cancerous cells, sub-classifications of cancerous cells, cells in different tissues or regions of an organ)

- Can we predict cell type on the basis of gene expression?
- Can we group cell types on the basis of expression similarity?

Here, the "conditions" on which expression is observed are cell types.

- What genes are involved in given biological processes?

(e.g. cell cycle, response to a stimulus or class of stimuli -- drug treatments, shocks to the cell)

- Can we identify characteristic expression patterns?
- Can we group relevant genes on the basis of expression similarity?

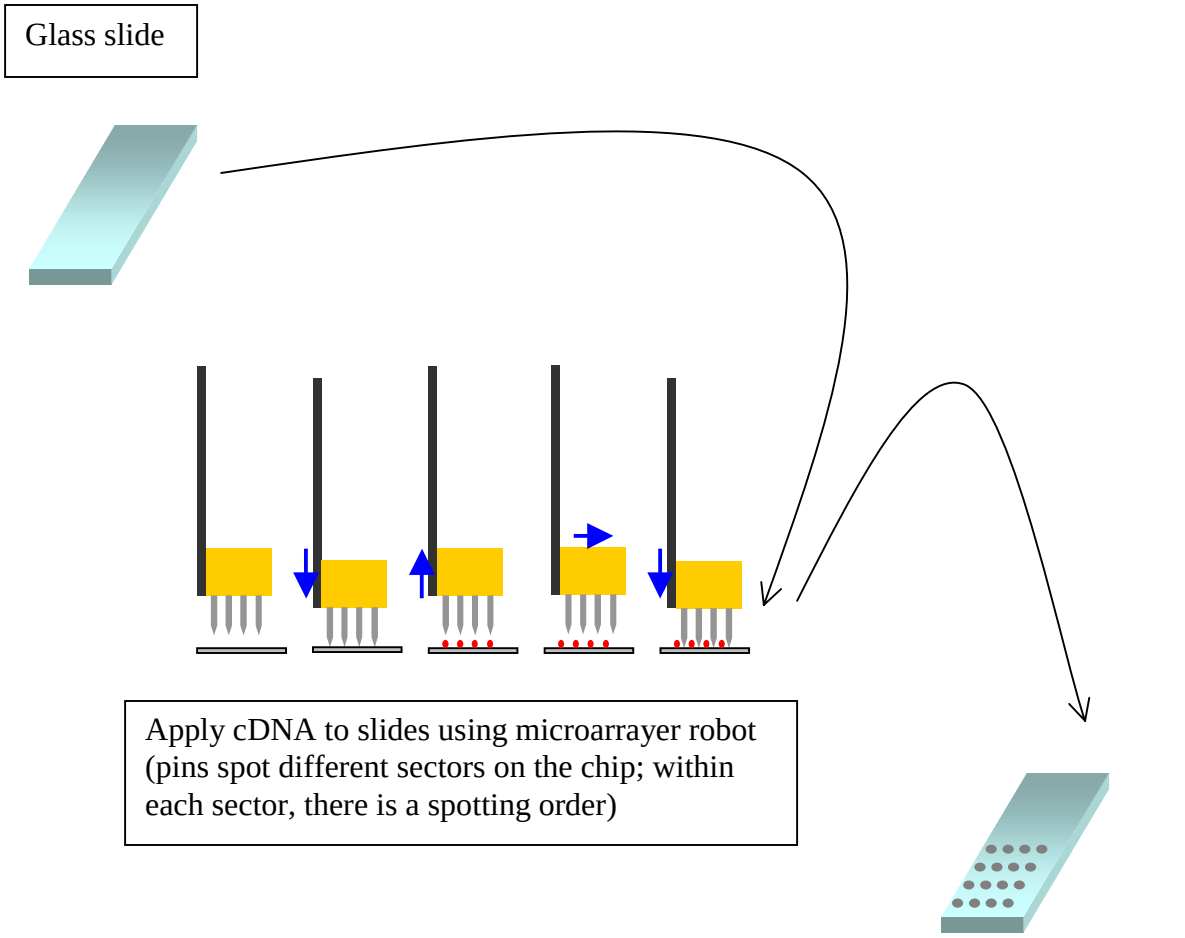
Here, "conditions" correspond to points in a time course.

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

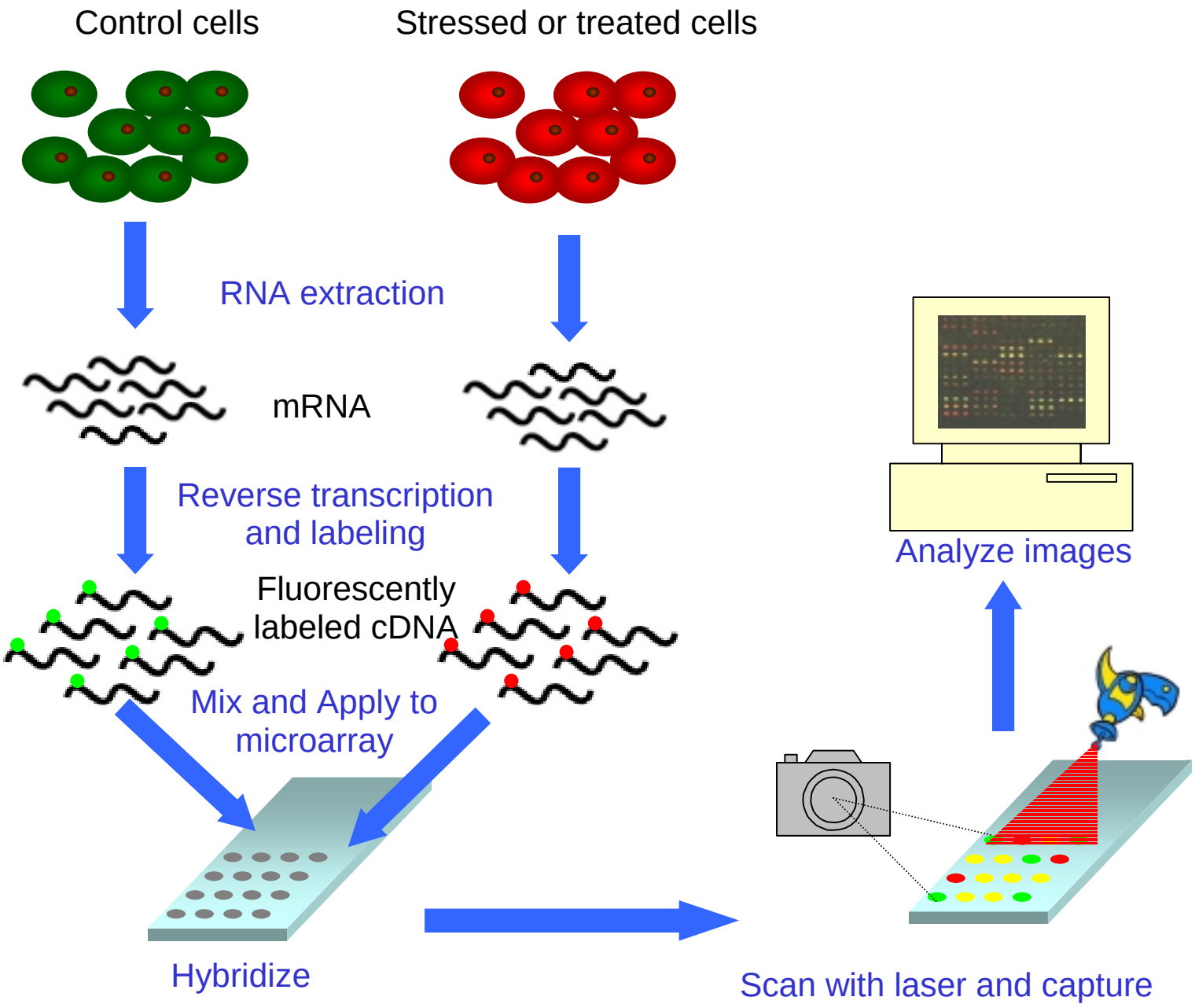
Summary description of spotted arrays technology:



Some spots contain "replicates" of the same gene. Some spots are "controls", i.e contain genes whose expression should not be subject to systematic variation:

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

Issue: where are replicate and control spots located?

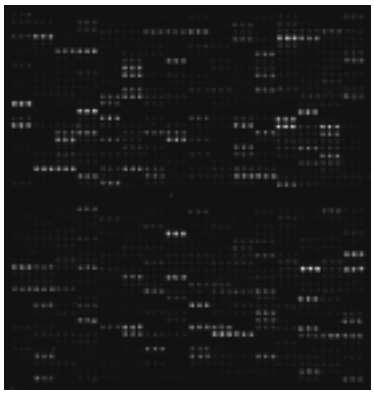


Issue: is there enough material in each spot for any amount of cDNA to hybridize?

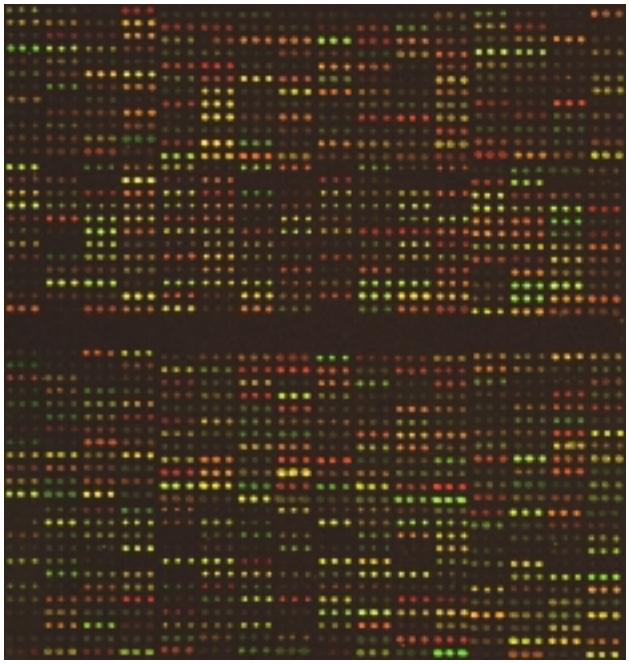
Issue: hybridization conditions, spatial homogeneity on chip?

Issue: cross-hybridization?

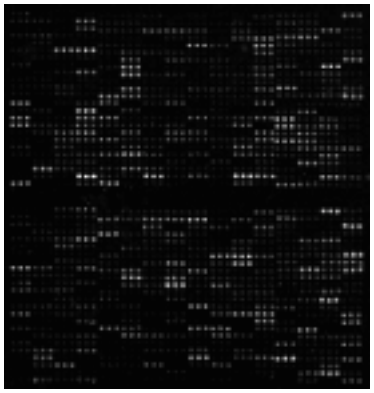
Cy3 channel (control)



False-color overlay



Cy5 channel (treatment)

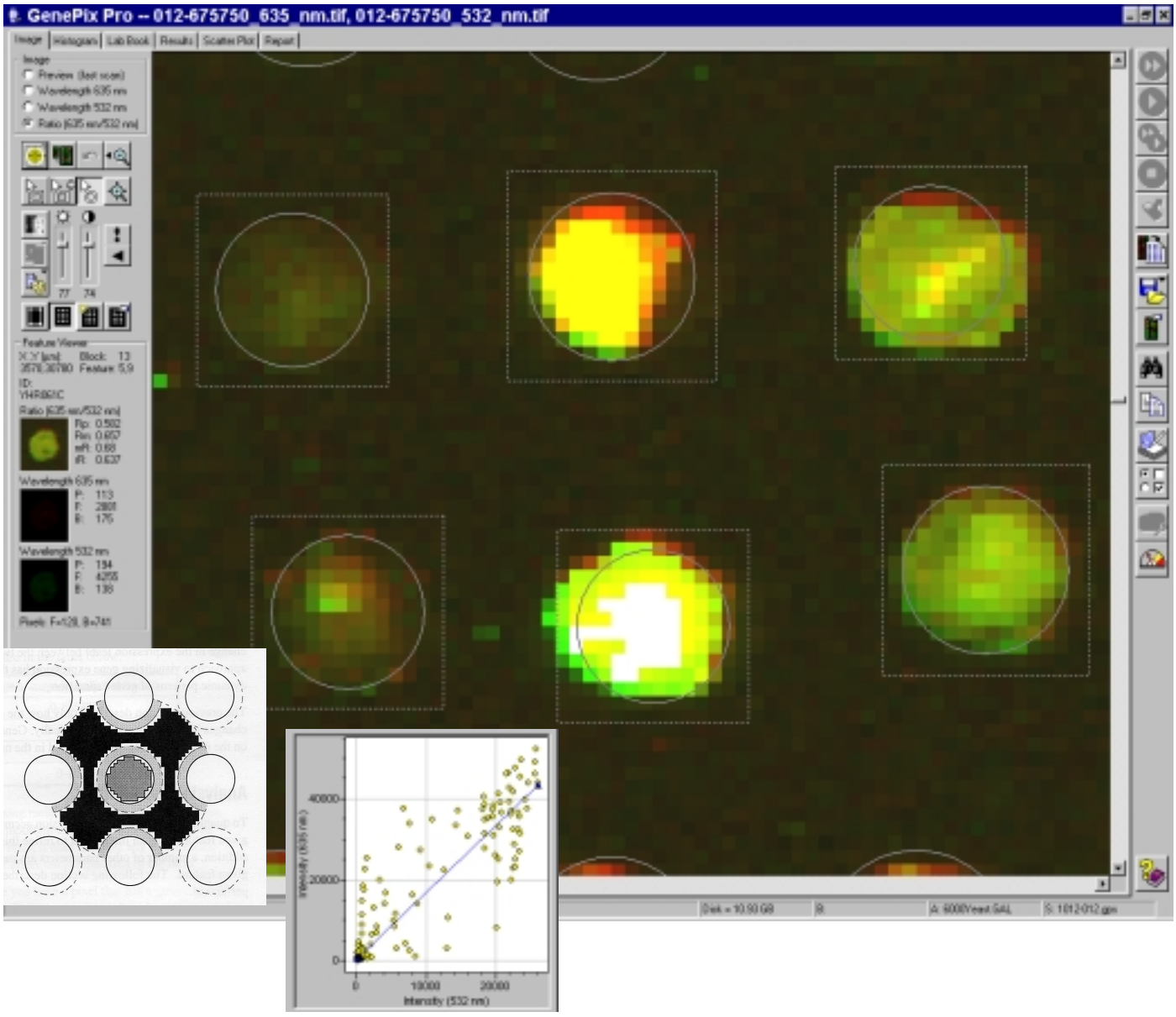


Output of the scanner: two monochromatic images of the array, one for each channel.

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).



For instance, summarize through two numbers per spot:
Red = Ave. Foregr. – Ave. Backgr. pixel intensities on Red
Green = Ave. Foregr. – Ave. Backgr. pixel intensities on Green

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

Example 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

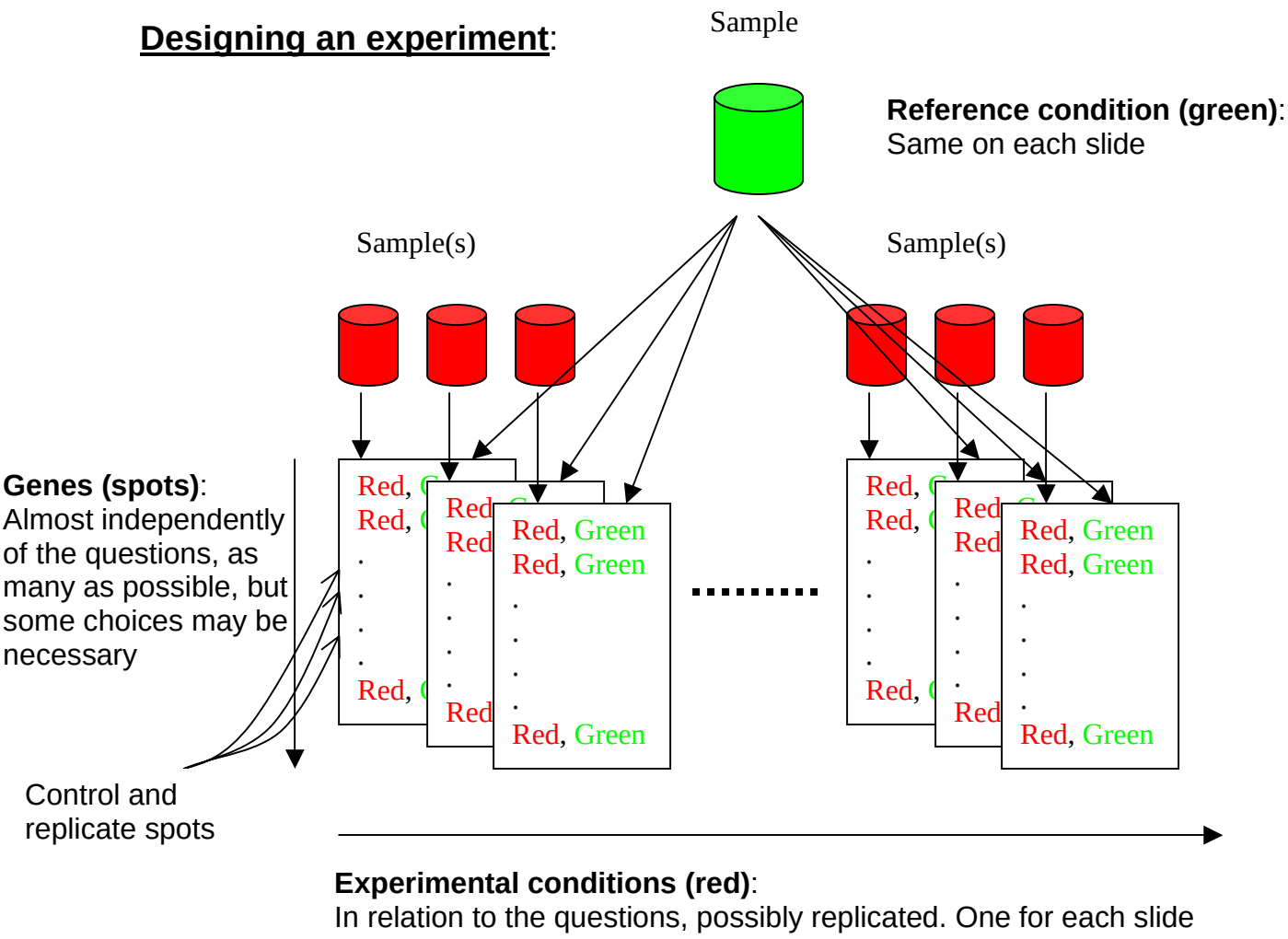
% > B635+1SD	% > B635+2SD	F635 % Sat.	F532 Median	F532 SD	B532 Median	B532 SD	% > B532+1SD
50	25	0	626	177	520	113	48
36	23	0	724	164	691	156	30
85	79	0	15921	11517	688	631	85
77	69	0	7613	6790	608	357	78

% > 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

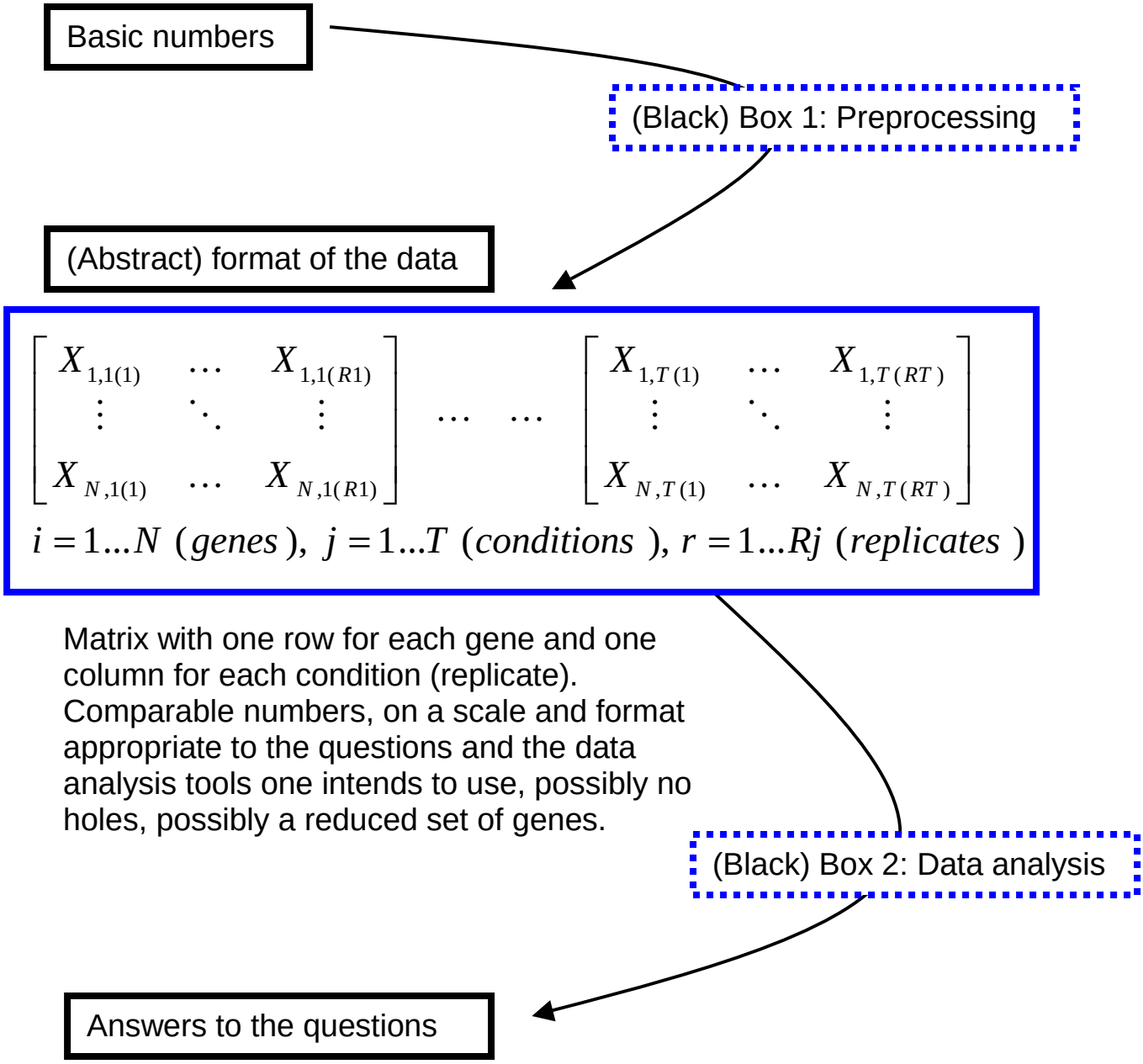
F635 Mean - B635	**F532 Mean - B532**	Flags
90	121	0
101	64	0
13589	14475	0
5634	7604	0

Designing an experiment:



Controls: important to compute measurements that are comparable across experimental conditions, and corrected from gross unwanted sources of variability.

Replicates: important to evaluate reproducibility, unwanted sources of variability, and (non-systematic) error variance.


Basic numbers
(Black) Box 1: Preprocessing
(Abstract) format of the data
$\begin{bmatrix} X_{1,1(1)} & \cdots & X_{1,1(R1)} \\ \vdots & \ddots & \vdots \\ X_{N,1(1)} & \cdots & X_{N,1(R1)} \end{bmatrix} \cdots \cdots \begin{bmatrix} X_{1,T(1)} & \cdots & X_{1,T(RT)} \\ \vdots & \ddots & \vdots \\ X_{N,T(1)} & \cdots & X_{N,T(RT)} \end{bmatrix}$
$i = 1...N\ (genes),\ j = 1...T\ (conditions),\ r = 1...Rj\ (replicates)$
Matrix with one row for each gene and one column for each condition (replicate). Comparable numbers, on a scale and format appropriate to the questions and the data analysis tools one intends to use, possibly no holes, possibly a reduced set of genes.
(Black) Box 2: Data analysis
Answers to the questions