

Advances in Microbial Diagnostic Microarrays

Teramo, 15 - 16 Marzo 2010

Mathematical and statistical issues to identify differentially expressed genes from microarray experiments

Corrado Dimauro

Dipartimento di Scienze Zootecniche, Università di Sassari

Monitoring of gene expression patterns with a complementary DNA microarray

One gene at time
expression level



Northern blots

RT-PCR

SAGE

Simultaneous measurement of
the expression levels of tens of
thousand of genes



Microarray

Functional genomic

Microarray technology was first developed in human medicine

In the last years, microarrays have been applied also in animal field

Aims

- + Genetic control of physiological processes**
- + Genetic control of productive traits**
- + Candidate genes in QTL studies**
- + Diagnostic microarrays**

There are two main platforms for “expression chips” microarrays: cDNA and oligonucleotide microarrays

cDNA arrays are made with long double-stranded DNA molecules generated by enzymatic reactions such as PCR

Oligonucleotide arrays employ oligonucleotide probes spotted by either robotic deposition or *in situ* synthesis

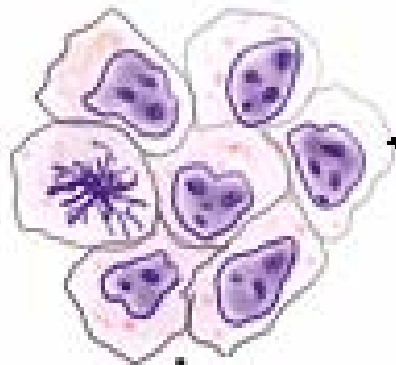
In cDNA arrays, two different samples are simultaneously hybridized onto the same array

In oligonucleotide arrays samples are hybridized separately onto different chips

cDNA microarray

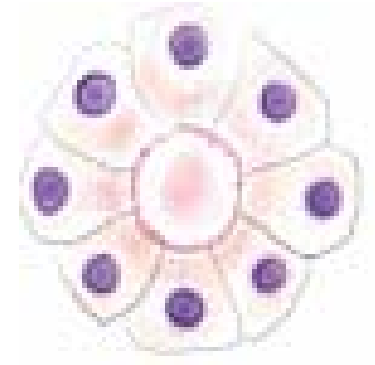
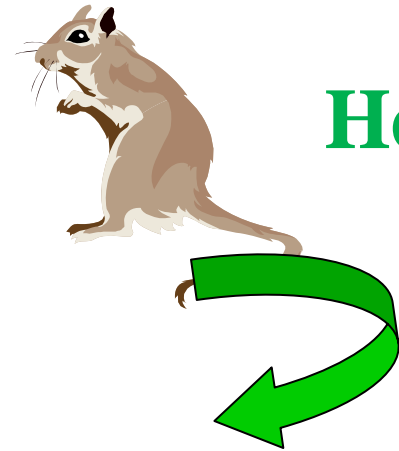
Goal of a microarray experiment is to identify a set of genes that are differentially expressed when, for example, sick and healthy samples are compared

Sick



RNA

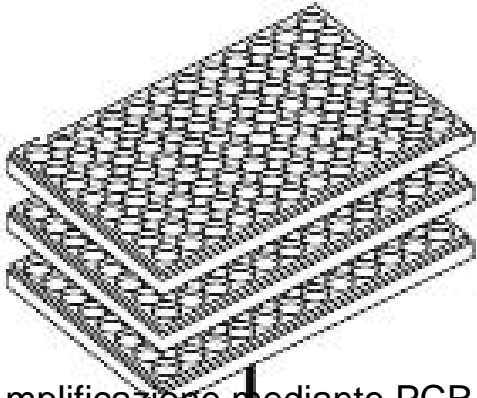
Healthy



RNA

cDNA microarray

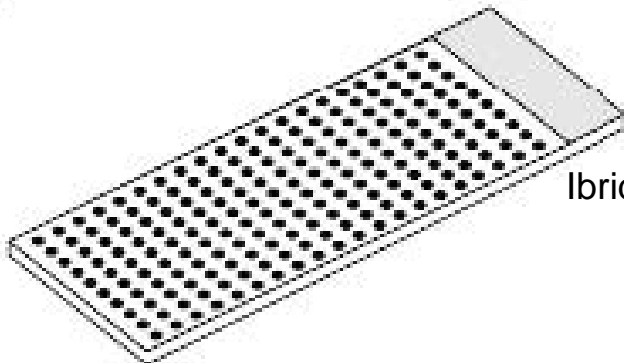
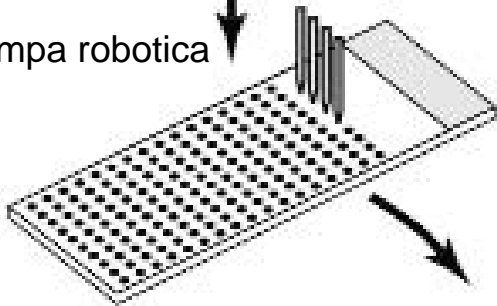
Cloni di DNA



Amplificazione mediante PCR



Stampa robotica



Ibridizzazione



635 = Cy5 = **Red**

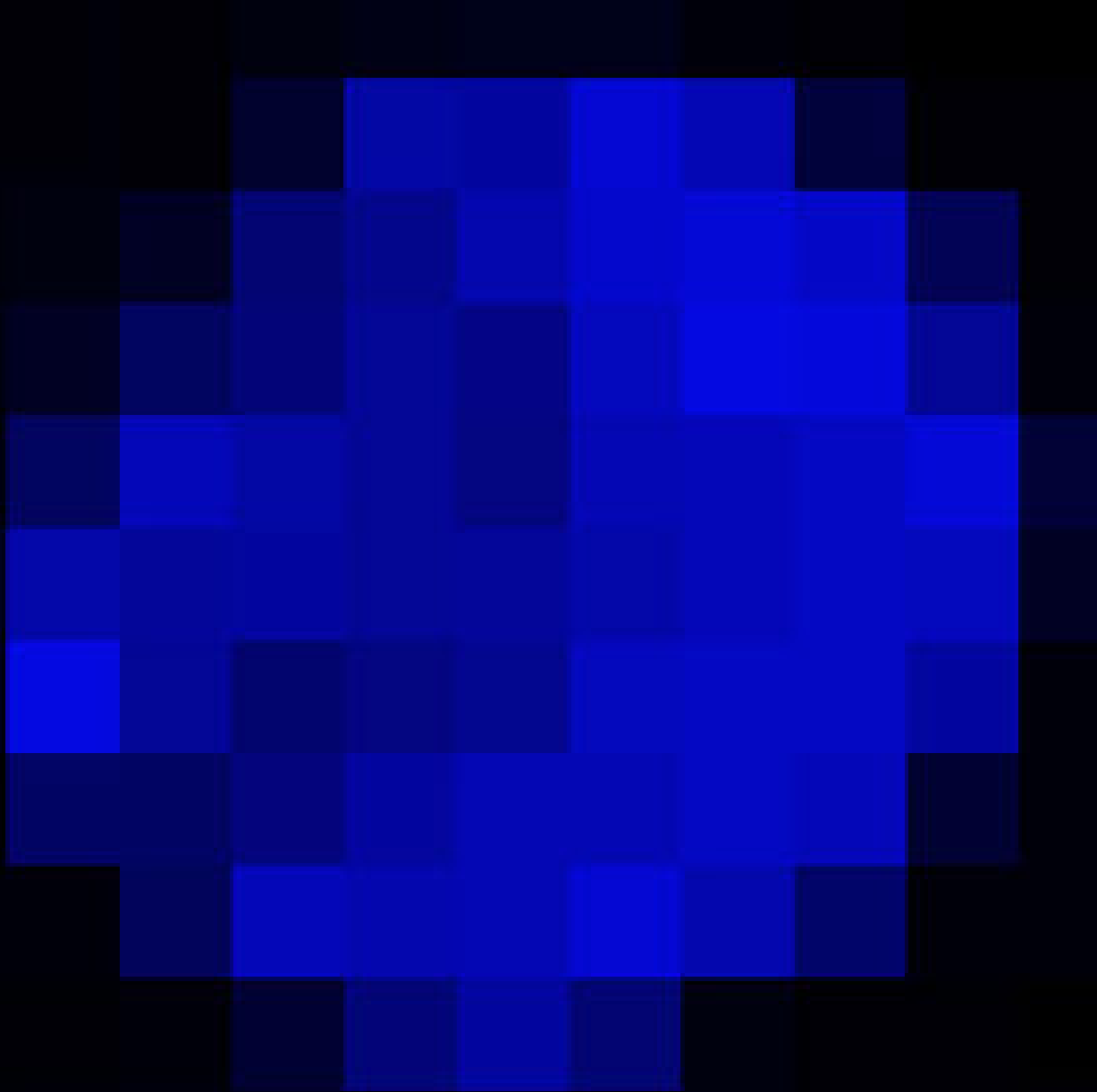
Output GenePix

523 = Cy3 = **Green**

Block	Column	Row	Name	ID	X	Y	Dia.	F635 Med	F635 Mea	F635 SD	F635 CV	B635	B635 Med	B635 Mea	B635 SD
1	1	1	01A12	Lambda Q	1400	12320	70	194	206	66	32	117	117	120	27
1	2	1	1,00E+12	Negative	1600	12310	100	114	117	29	24	114	114	116	24
1	3	1	01I12	GAPDH	1810	12320	110	649	679	185	27	113	113	116	25
1	4	1	01M12	Beta-actin	2020	12330	120	329	832	958	115	112	112	114	25
1	5	1	01A24	Negative	2230	12310	100	120	124	34	27	112	112	113	23
1	6	1	1,00E+24	GAPDH	2440	12320	120	711	702	186	26	112	112	114	24
1	7	1	01I24	Beta-actin	2650	12320	120	1243	1337	542	40	115	115	116	28
1	8	1	01M24	Lambda Q	2850	12310	50	238	231	33	14	129	129	134	35
1	9	1	02A12	NBFGC_B	3060	12290	60	542	575	196	34	126	126	129	29
1	10	1	2,00E+12	NBFGC_B	3280	12320	120	435	512	216	42	115	115	116	23
1	11	1	02I12	NBFGC_B	3490	12320	120	172	223	112	50	116	116	118	27
1	12	1	02M12	NBFGC_B	3690	12280	50	450	469	140	29	123	123	123	27
1	13	1	02A24	NBFGC_B	3910	12310	120	948	1145	640	55	114	114	117	27
1	14	1	2,00E+24	NBFGC_B	4120	12310	120	285	297	87	29	115	115	117	26
1	15	1	02I24	NBFGC_B	4330	12310	120	661	684	212	30	114	114	116	22
1	16	1	02M24	NBFGC_B	4540	12310	120	754	856	342	39	117	117	119	44
1	17	1	03A12	NBFGC_B	4750	12310	120	386	382	101	26	115	115	116	24
1	18	1	3,00E+12	NBFGC_B	4960	12310	120	231	293	148	50	115	115	116	24
1	19	1	03I12	NBFGC_B	5170	12300	120	1704	1895	1010	53	116	116	117	24
1	20	1	03M12	NBFGC_B	5370	12300	110	515	598	220	36	115	115	117	23
1	1	2	03A24	NBFGC_B	1400	12530	120	272	308	141	45	116	116	120	29
1	2	2	3,00E+24	NBFGC_B	1610	12530	120	448	457	126	27	117	117	119	24
1	3	2	03I24	NBFGC_B	1820	12530	120	216	262	122	46	116	116	117	25
1	4	2	03M24	NBFGC_B	2020	12530	120	188	230	116	50	113	113	114	23
1	5	2	04A12	NBFGC_B	2240	12520	100	280	281	86	30	111	111	112	23
1	6	2	4,00E+12	NBFGC_B	2430	12500	60	268	267	73	27	110	110	113	24
1	7	2	04I12	NBFGC_B	2660	12550	110	348	377	198	52	115	115	118	30
1	8	2	04M12	NBFGC_B	2890	12550	60	1586	1702	1003	58	120	120	127	37

Data analysis

One single spot



**In each channel
intensity**

=

**Mean of pixel
intensities**

OR

**Median of pixel
intensities**

635 = Cy5 = **Red**

523 = Cy3 = **Green**

Gene	F635 median	F635 mean	F523 median	F523 mean	B635 median	B635 mean	B523 median	B523 mean
BE483318	195	206	325	387	46	62	87	102

An amount of statistical techniques are required to analyze raw data

No standardization in data analysis procedures

Low reliability and repeatability of results



No routine application

two main problems are involved



Statistical



Technical

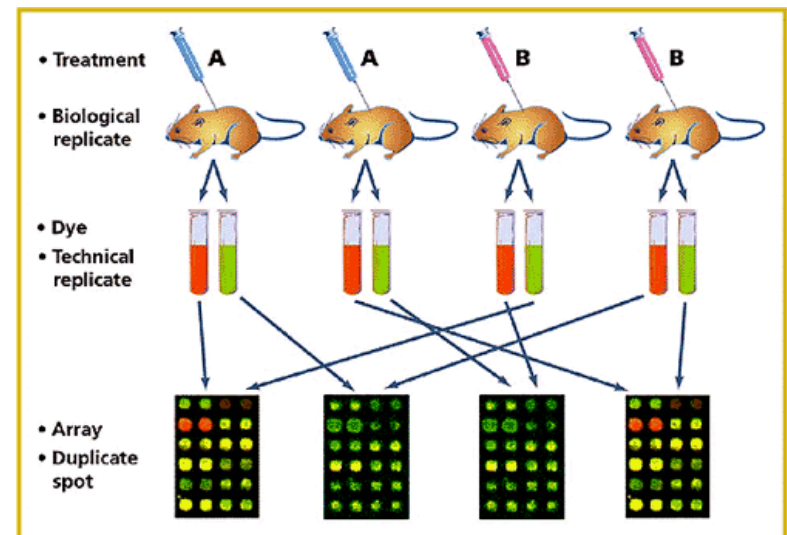
Statistical drawbacks

Goal of the microarray analysis is to identify a set of genes that are differentially expressed when, for example, sick and healthy samples are compared

SAMPLE INFERENCE P-VALUE POPULATION

**Biological replicates
are required**

**Different comparisons
are required**



Technical replicates improve the precision of the measurement

Biological replicates are required

This is expensive and, sometimes,

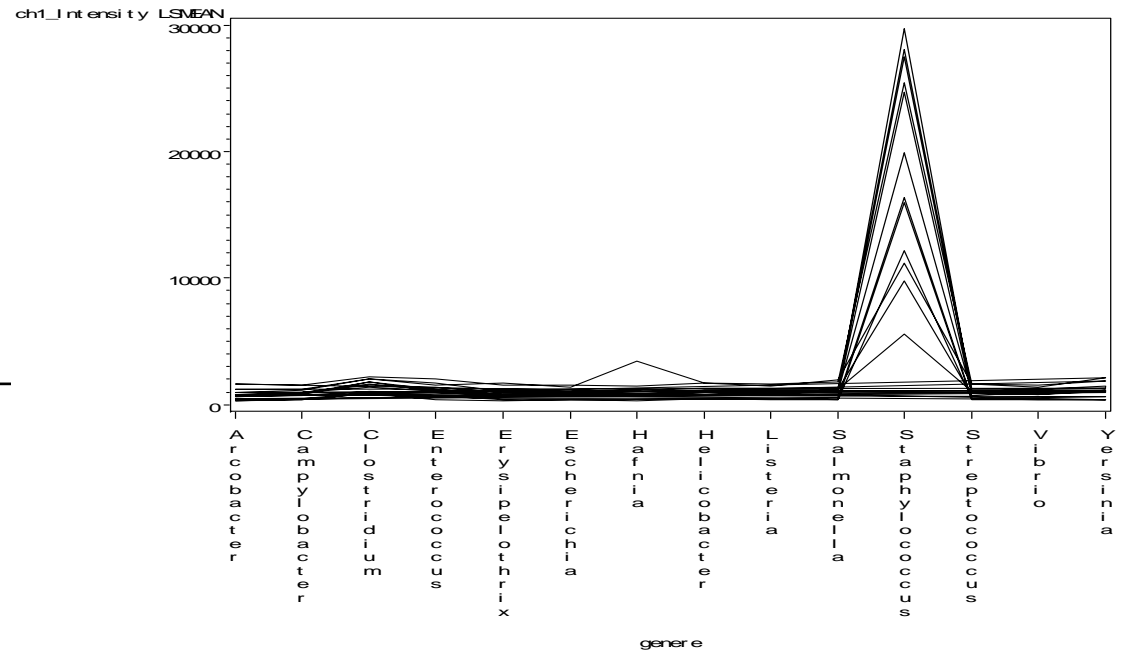
impossible

Each probe represents a single information

There is a great difference in the analysis of data from diagnostic and expression microarray

Genere	Numero geni (tre repliche)	Numero specie
Arcobacter	2	1
Campylobacter	473	19
Clostridium	113	30
Enterococcus	75	26
Erysipelothrix	10	1
Escherichia	6	1
HK	119	4
Hafnia	1	1
Helicobacter	16	10
Listeria	116	6
Salmonella	620	5
Staphylococcus	372	37
Streptococcus	220	42
Vibrio	5	1
Yersinia	316	14

Regarding to staphylococcus, and some of its species, it is highly probable that they are detected, having 372 genes



Classic statistic

vs.

non classic data



Severe problems to assess the FDR levels

Technical drawbacks

Several sources of error

■ **Dusty signal**

■ **Labelling**

■ **Signal intensity**

■ **RNA concentration**

Fundamental hypothesis

Goal of the microarray analysis is to identify a subset of genes that are differentially expressed between the sick and healthy samples.

The null hypothesis being tested is that there is no difference in expression between the conditions

As a consequence, only “few” genes should be differentially expressed

Measured intensities for each gene represent its relative expression level

If **G** is the green channel and **R** the red channel

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

If **T > 1** then the gene is up-regulated

If **T < 1** then the gene is down-regulated

Genes **up-regulated by a factor 2** have **T=2**

Genes **down-regulated by a factor 2** have **T=0.5**

Generally, the log-base 2 of T is considered

$$T_i = \log_2 \frac{R_i}{G_i}$$

if $\frac{R_i}{G_i} = 1$ then $\log_2 \frac{R_i}{G_i} = 0$ therefore $T_i = 0$

if $\frac{R_i}{G_i} = 2$ then $\log_2 \frac{R_i}{G_i} = 1$ therefore $T_i = 1$

if $\frac{R_i}{G_i} = \frac{1}{2}$ then $\log_2 \frac{R_i}{G_i} = -1$ therefore $T_i = -1$

if $\frac{R_i}{G_i} = 4$ then $\log_2 \frac{R_i}{G_i} = 2$ therefore $T_i = 2$

if $\frac{R_i}{G_i} = \frac{1}{4}$ then $\log_2 \frac{R_i}{G_i} = -2$ therefore $T_i = -2$

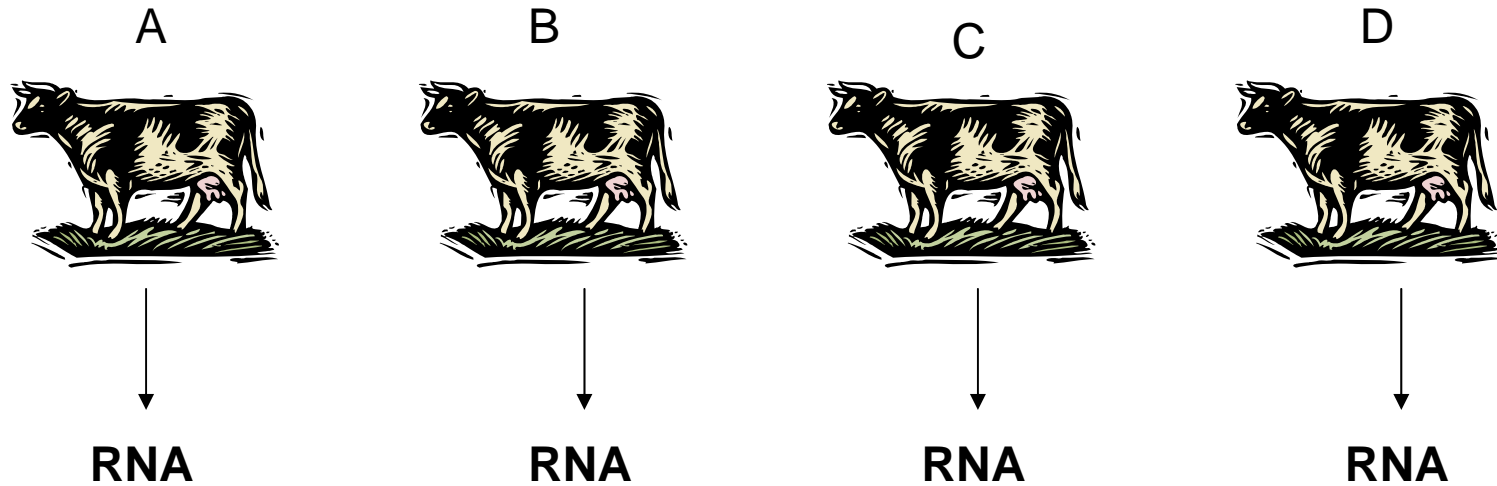
} symmetry

} symmetry

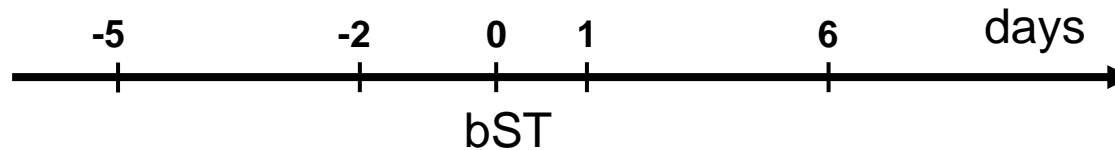
A REAL EXPERIMENT

SOMATOTROPIN (bST) ADMINISTRATION TO LACTATING COWS

WHAT ABOUT THE GENETIC CONTROL?

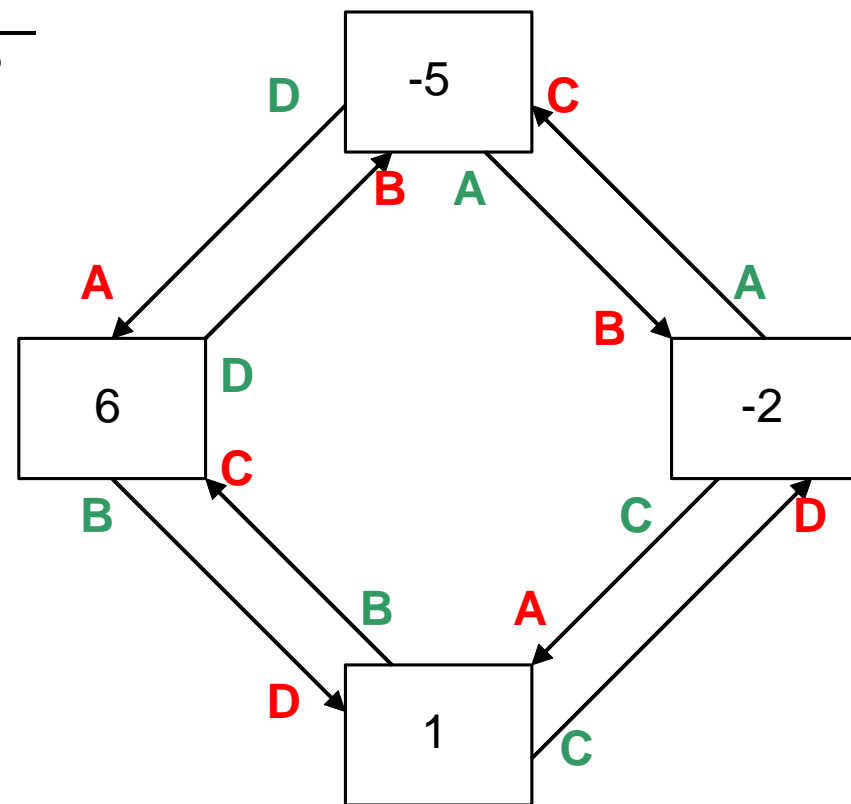


4 time points



The experimental plan: closed loop design with dye swap

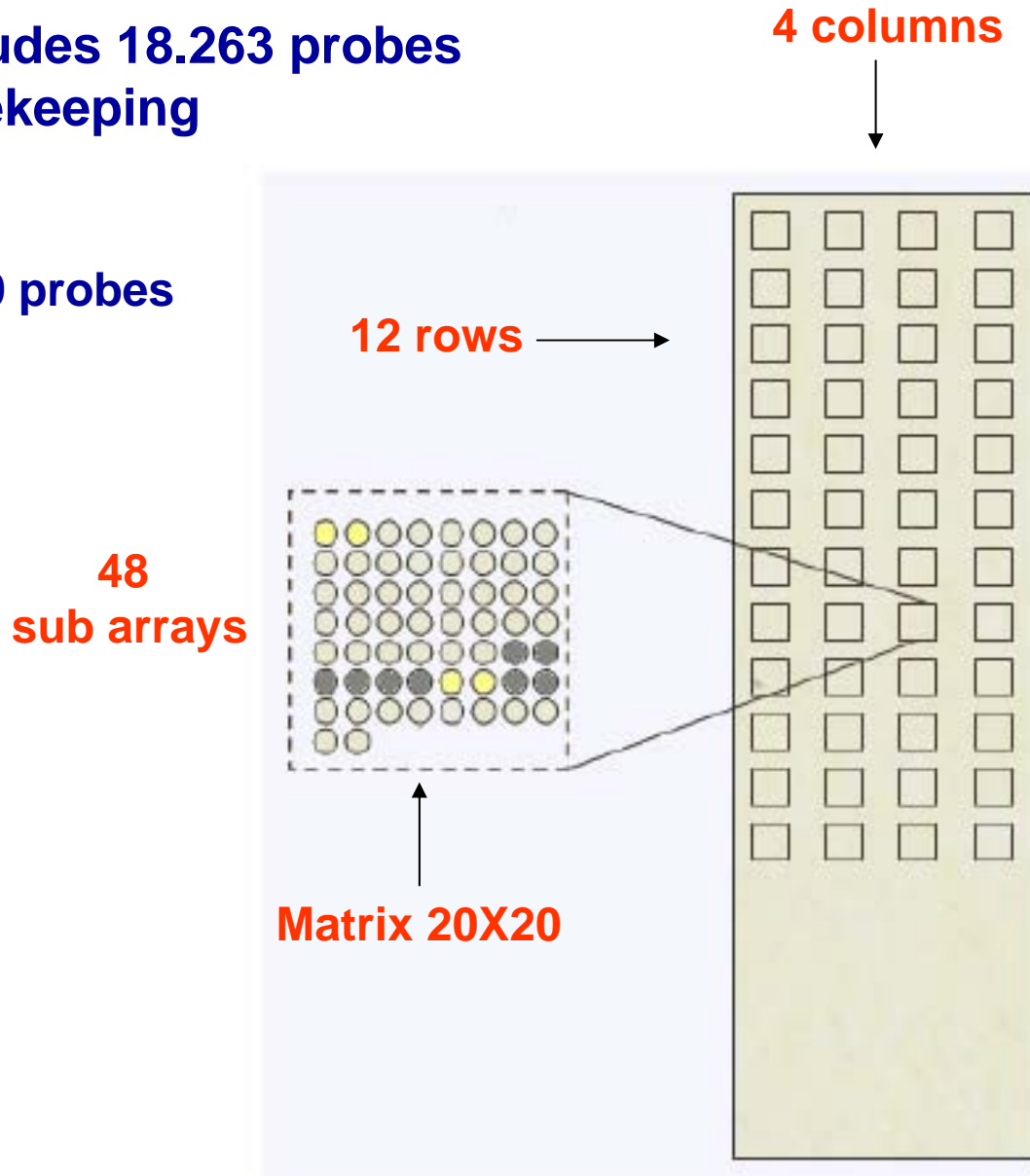
Piano 1					
Array	Rosso		vs.	Verde	
	Vacca	Giorno		Vacca	Giorno
1	A	-5	vs.	B	-2
2	A	-2	vs.	C	-5
3	D	-5	vs.	A	6
4	D	6	vs.	B	-5
5	C	-2	vs.	A	1
6	C	1	vs.	D	-2
7	B	1	vs.	C	6
8	B	6	vs.	D	1
Piano 2					
9	D	-5	vs.	C	-2
10	D	-2	vs.	A	-5
11	B	-5	vs.	D	6
12	B	6	vs.	C	-5
13	A	-2	vs.	D	1
14	A	1	vs.	B	-2
15	C	1	vs.	A	6
16	C	6	vs.	B	1



16 high-density microarrays from NBFGC consortium

The array includes 18.263 probes
and 937 housekeeping

A total of 19.000 probes



Output GenePix

Block	Column	Row	Name	ID	X	Y	Dia.	F635 Med	F635 Mean	F635 SD	F635 CV	B635	B635 Med	B635 Mean	B635 SD
1	1	1	01A12	Lambda Q	1400	12320	70	194	206	66	32	117	117	120	27
1	2	1	1,00E+12	Negative	1600	12310	100	114	117	29	24	114	114	116	24
1	3	1	01I12	GAPDH	1810	12320	110	649	679	185	27	113	113	116	25
1	4	1	01M12	Beta-actin	2020	12330	120	329	832	958	115	112	112	114	25
1	5	1	01A24	Negative	2230	12310	100	120	124	34	27	112	112	113	23
1	6	1	1,00E+24	GAPDH	2440	12320	120	711	702	186	26	112	112	114	24
1	7	1	01I24	Beta-actin	2650	12320	120	1243	1337	542	40	115	115	116	28
1	8	1	01M24	Lambda Q	2850	12310	50	238	231	33	14	129	129	134	35
1	9	1	02A12	NBFGC_B	3060	12290	60	542	575	196	34	126	126	129	29
1	10	1	2,00E+12	NBFGC_B	3280	12320	120	435	512	216	42	115	115	116	23
1	11	1	02I12	NBFGC_B	3490	12320	120	172	223	112	50	116	116	118	27
1	12	1	02M12	NBFGC_B	3690	12280	50	450	469	140	29	123	123	123	27
1	13	1	02A24	NBFGC_B	3910	12310	120	948	1145	640	55	114	114	117	27
1	14	1	2,00E+24	NBFGC_B	4120	12310	120	285	297	87	29	115	115	117	26
1	15	1	02I24	NBFGC_B	4330	12310	120	661	684	212	30	114	114	116	22
1	16	1	02M24	NBFGC_B	4540	12310	120	754	856	342	39	117	117	119	44
1	17	1	03A12	NBFGC_B	4750	12310	120	386	382	101	26	115	115	116	24
1	18	1	3,00E+12	NBFGC_B	4960	12310	120	231	293	148	50	115	115	116	24
1	19	1	03I12	NBFGC_B	5170	12300	120	1704	1895	1010	53	116	116	117	24
1	20	1	03M12	NBFGC_B	5370	12300	110	515	598	220	36	115	115	117	23
1	1	2	03A24	NBFGC_B	1400	12530	120	272	308	141	45	116	116	120	29
1	2	2	3,00E+24	NBFGC_B	1610	12530	120	448	457	126	27	117	117	119	24
1	3	2	03I24	NBFGC_B	1820	12530	120	216	262	122	46	116	116	117	25
1	4	2	03M24	NBFGC_B	2020	12530	120	188	230	116	50	113	113	114	23
1	5	2	04A12	NBFGC_B	2240	12520	100	280	281	86	30	111	111	112	23
1	6	2	4,00E+12	NBFGC_B	2430	12500	60	268	267	73	27	110	110	113	24
1	7	2	04I12	NBFGC_B	2660	12550	110	348	377	198	52	115	115	118	30
1	8	2	04M12	NBFGC_B	2890	12550	60	1586	1702	1003	58	120	120	127	37

F635 = Cy5 = Rosso

F523 = Cy3 = Verde

Data analysis

Three fundamental steps

1. Data normalization
2. Detection of differentially expressed genes
3. Clusters

Errors in data acquisition due to

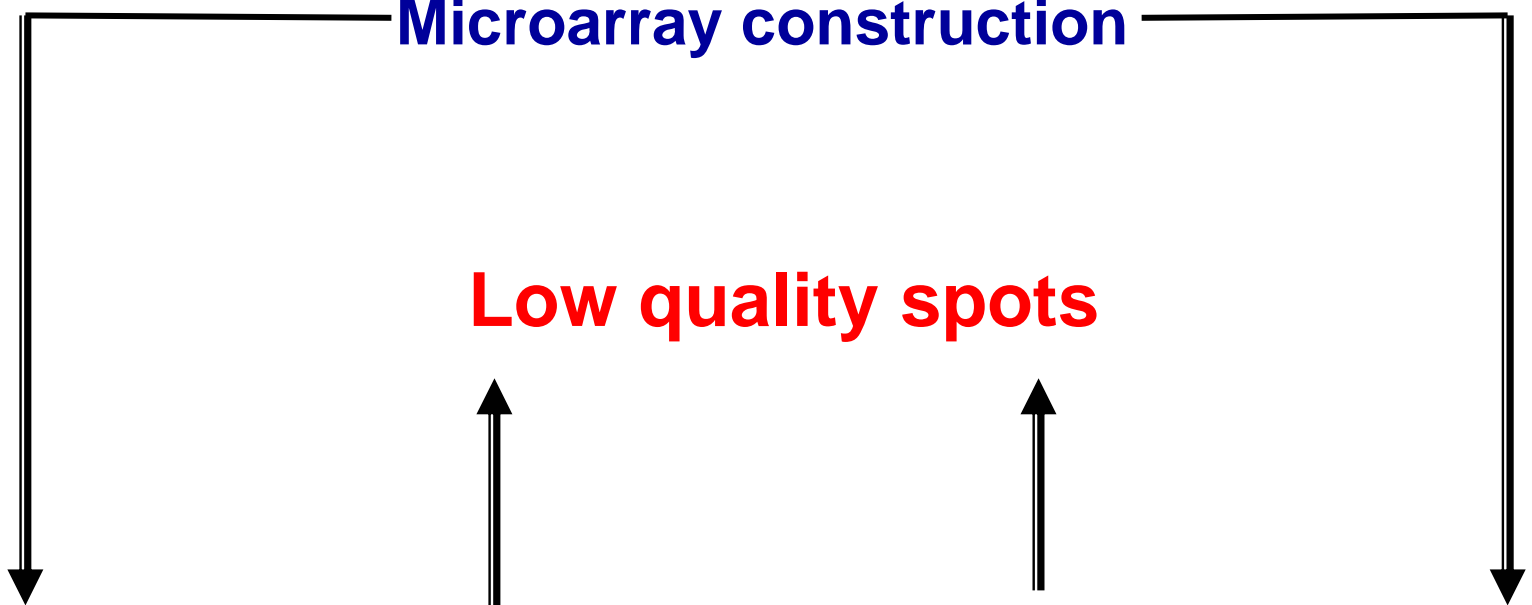


Microarray construction

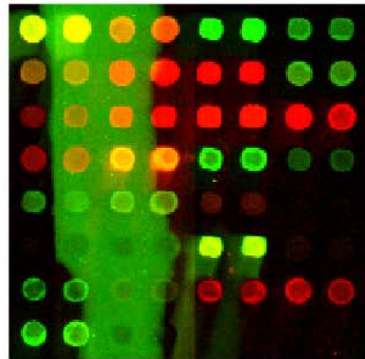
Low quality spots

**Abnormal shape
and dimension**

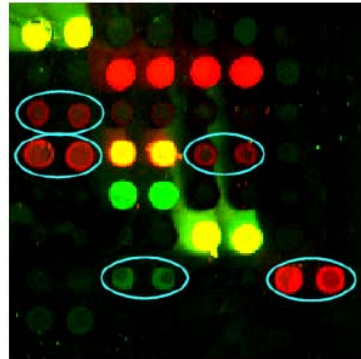
**Dust and dirt on the
microarray**



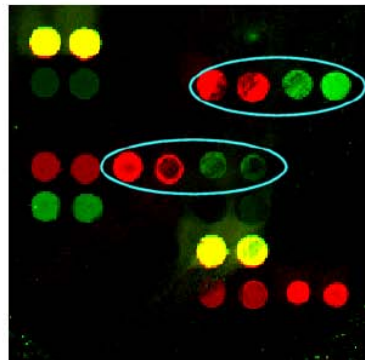
Data normalization: spot analysis



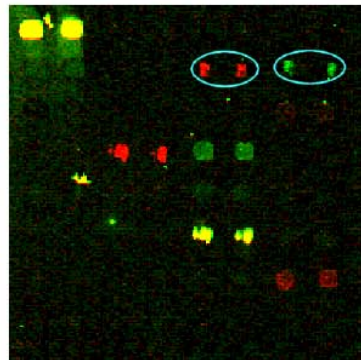
(a) Striscia di sporco netta



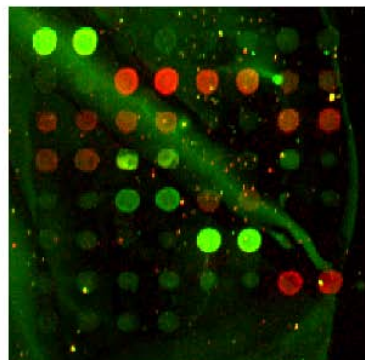
(b) Spot a "ciambella"



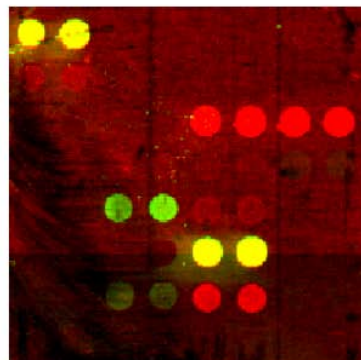
(c) Spot rovinati



(d) Spot molto piccoli



(e) Spot con sfondo scuro o basso segnale



(f) Spot con sfondo scuro o basso segnale

Data normalization: spot analysis

gene	f1mediana	b1mediana	f1media	b1media	f2mediana	b2mediana	f2media	b2media
Lambda Q	150	134	150	135	130	145	131	145
Negative	141	136	145	140	147	146	146	146
GAPDH	790	137	795	142	1154	149	1119	150
Beta-actin	1050	135	1045	136	3776	147	3726	149
Negative	140	135	142	135	156	144	162	147
GAPDH	950	135	944	137	1225	145	1210	146
Beta-actin	1418	136	1407	139	4810	151	4773	152
Lambda Q	156	138	160	147	166	148	173	152
BE483243	182	139	198	145	210	152	242	154
BE483318	327	137	325	140	295	153	296	155
BE483151	303	140	307	142	298	154	299	156
BE483227	232	140	237	142	194	154	194	155
BE479768	600	138	640	141	397	153	408	156
BE479860	225	138	228	139	262	152	265	153
BE479916	247	138	246	140	253	155	255	156
BE479712	1046	139	997	144	3045	158	2893	159

Problem solution



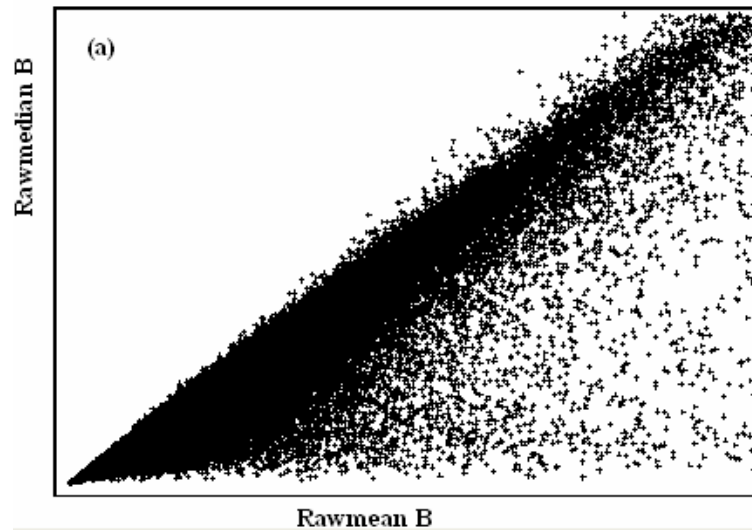
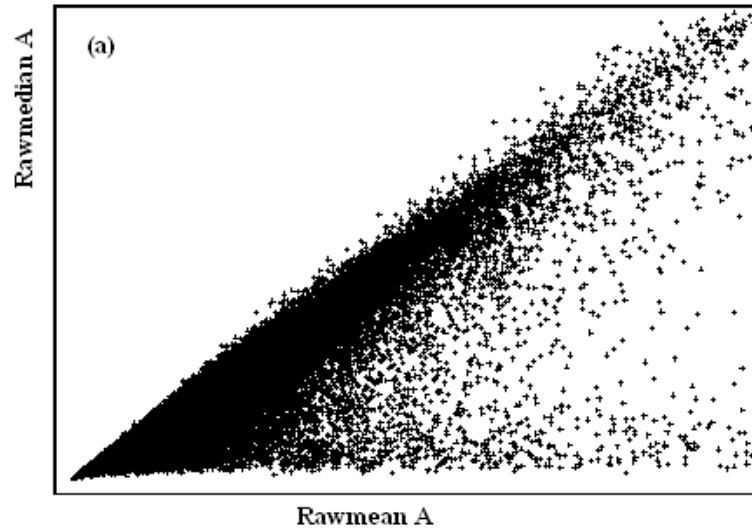
Good Spot **⇒** **median intensity ~ mean intensity**

Test: median-mean correlation ~ 99%

$$R = \frac{\text{mean intensity}}{\text{median intensity}}$$

If $R < 0.80$ the spot is deleted

Data normalization: spot analysis



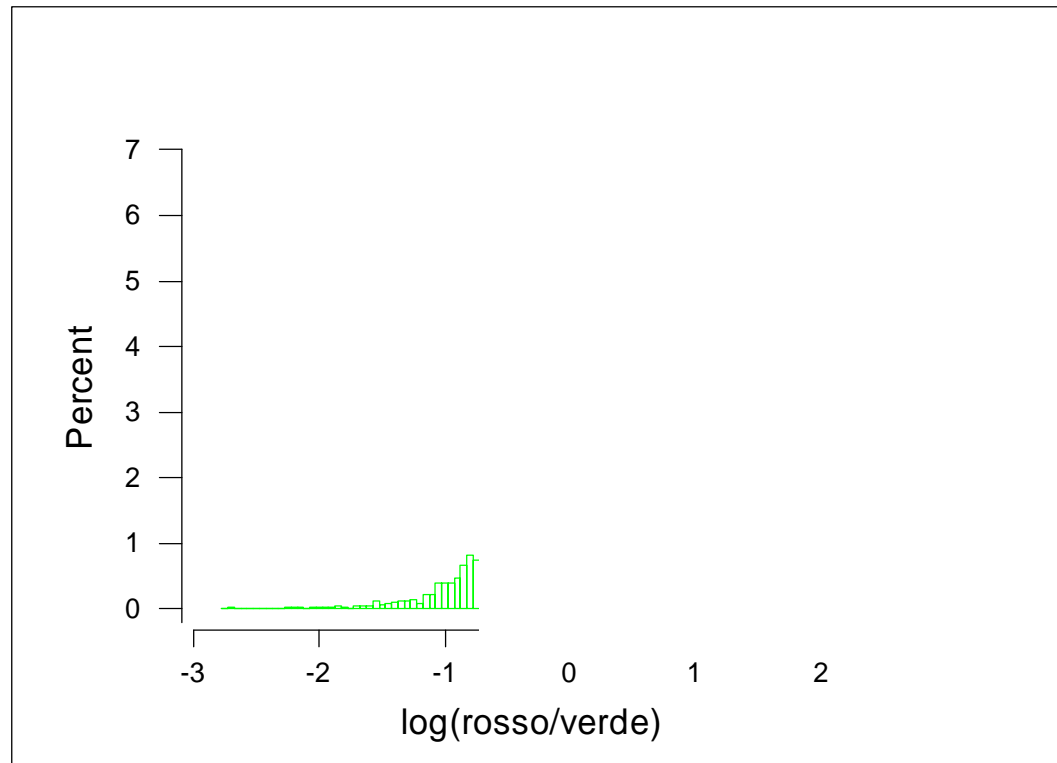
~ 20% of spots were deleted in the bST experiment

Data normalization: logarithmic transformation

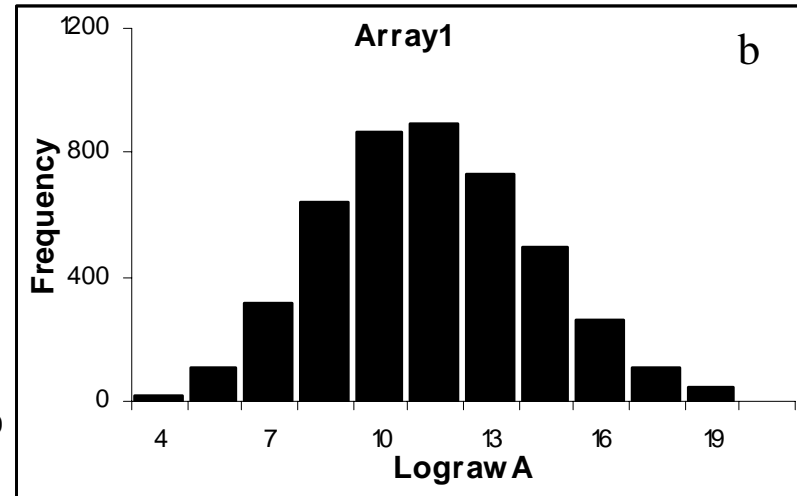
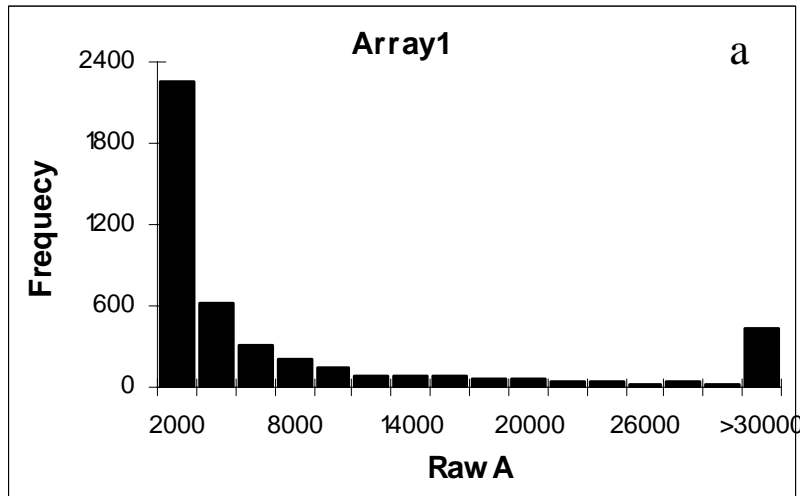
Source of error \implies **Data are not normally distributed**

Problem solution

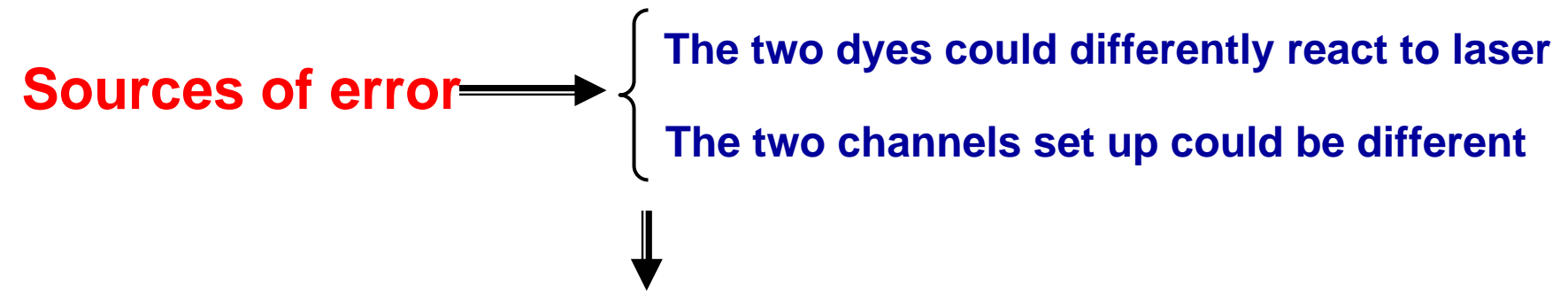
Log-base 2 transformation of data



In the bST experiment



Data normalization: intensity correction



Bias due to the fluorescence intensity (M-A plot)

M = Comparison of the red and green intensities in each spot



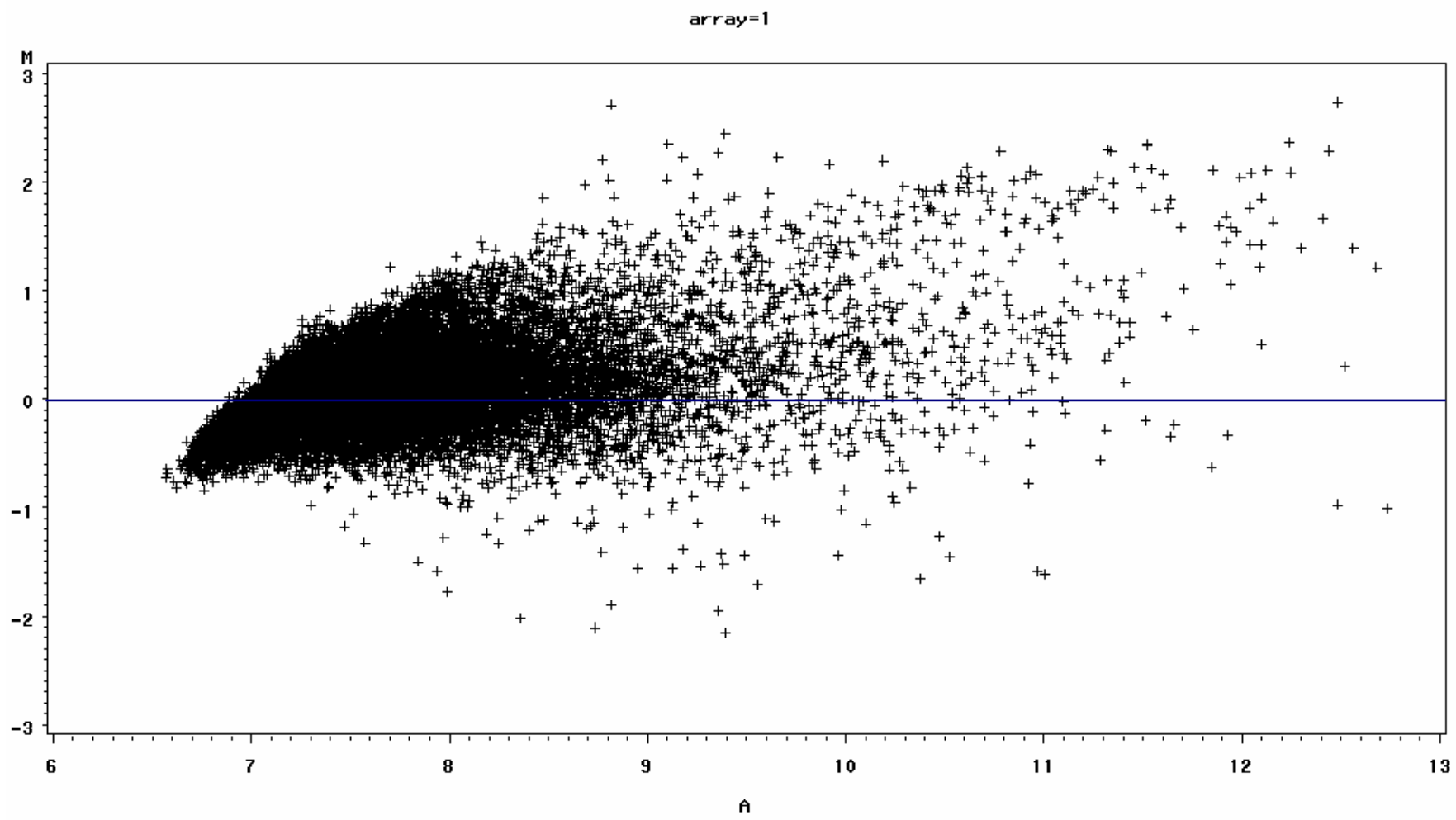
$$M = \log_2 R / G = \log_2 R - \log_2 G$$

A = Mean intensity of each spot



$$A = \log_2 \sqrt{RG} = \frac{1}{2} (\log_2 R + \log_2 G)$$

Data normalization: intensity correction



Problem solution



LOcally WEighted Scatterplot Smoothing regression (LOWESS)

Regression M vs A locally weighted $\Longrightarrow M'$

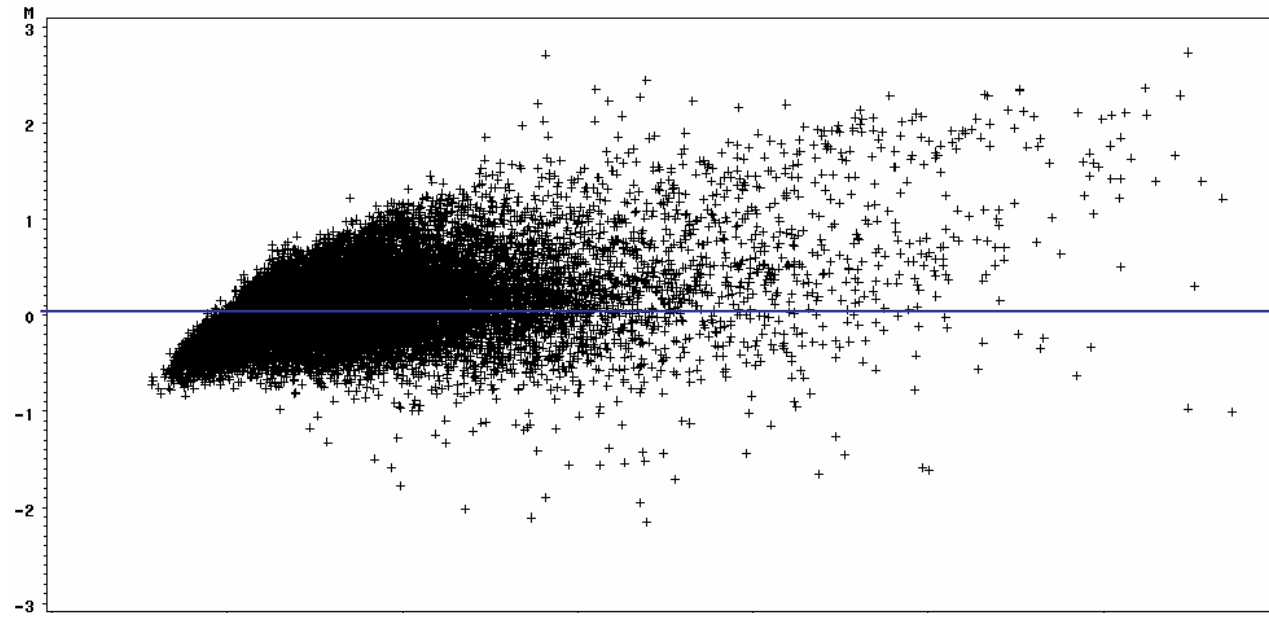
The normalized intensity $\Longrightarrow M^* = M - M'$

$$\log_2^* G = A + M^*/2$$

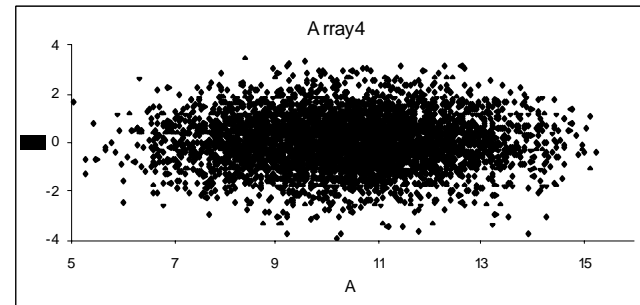
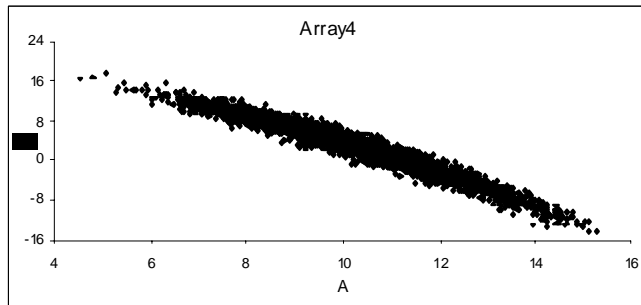
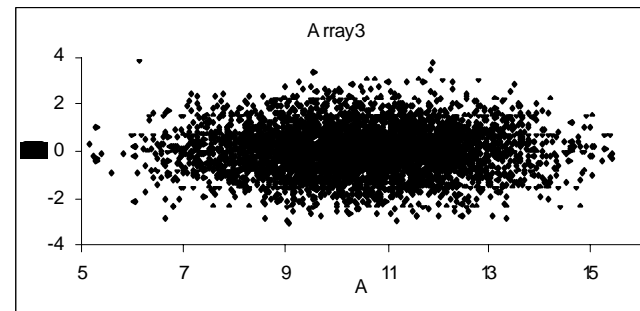
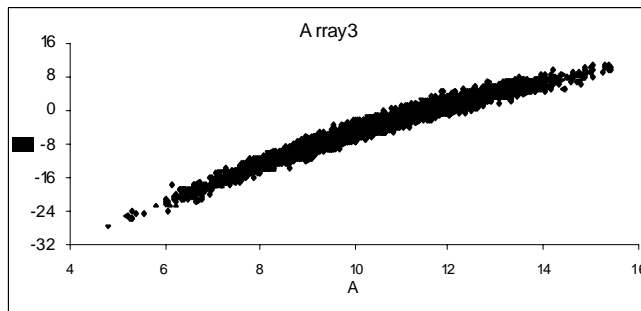
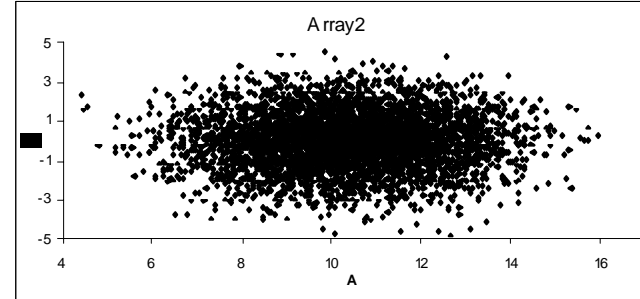
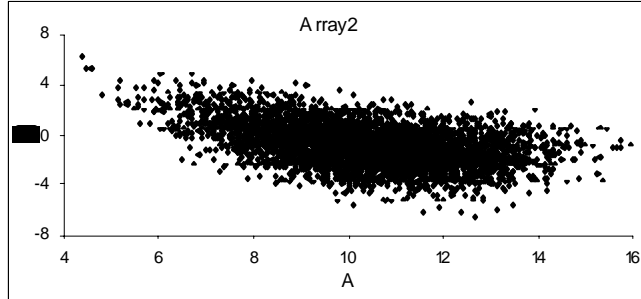
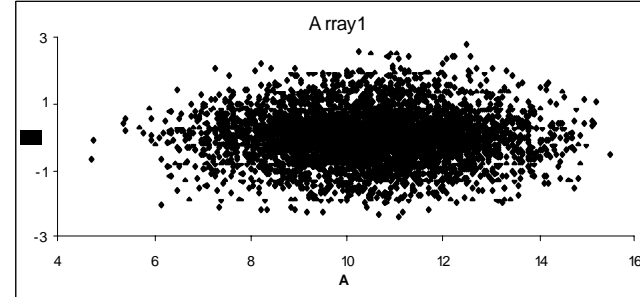
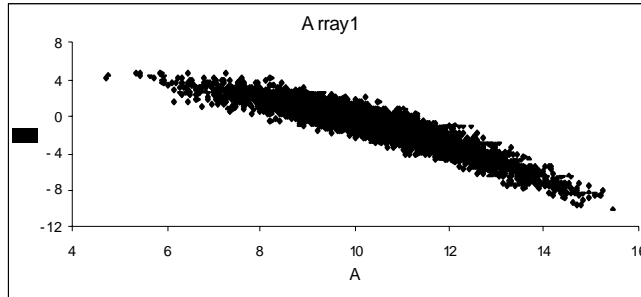
$$\log_2^* R = A - M^*/2$$

Data normalization: intensity correction

array=1



In the bST experiment

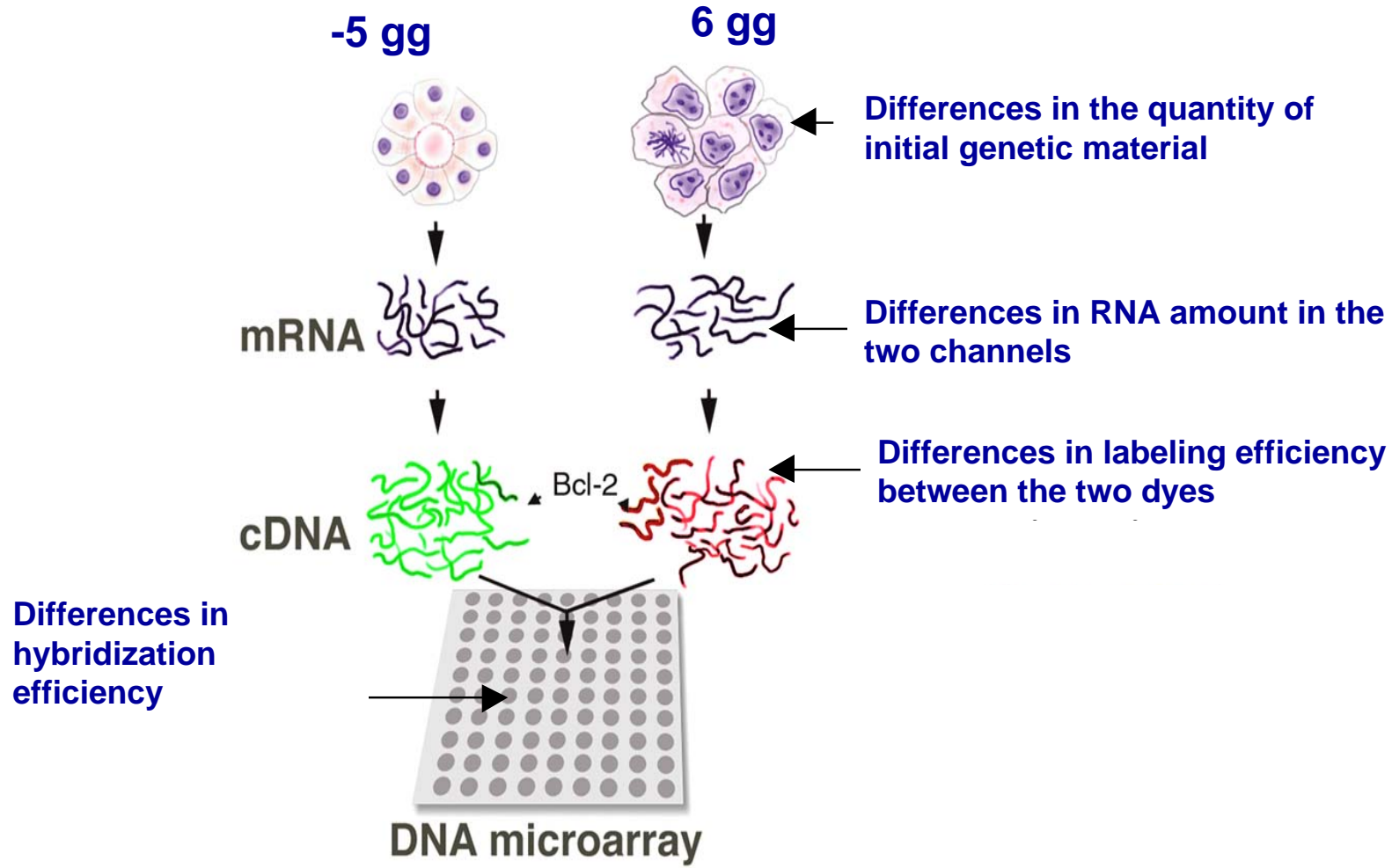


a

b

Data normalization: global dye correction

Sources of error



$$\frac{\sum R}{\sum G} = 1$$

Problem solution



ANOVA mixed model



$$y_{ijk} = \mu + D_j + A_k + (DA)_{jk} + \varepsilon_{ijk}$$

D= dye (fixed effect)

A= array (random effect)

DA= Dye*Array interaction (random effect)

$Y - Y_{\text{pred}} = r = \text{residuals} = \text{normalized data}$

Results: differentially expressed genes

OBJECTIVE



Detection of differentially expressed genes

Problem solution



GENE SPECIFIC mixed model

$$r_{ijk} = \mu + T_i + D_j + A_k + \gamma_{ijk}$$

r = normalized data (residuals of the former mixed model)

T= fixed effect of time points: -5, -2, 1, 6

D= fixed effect of dye, specific for each gene

A= random effect of array

Results: differentially expressed genes

Source of error



Multiple Testing Error Rate

Problem solution



PERMUTATION TEST

Data are randomly assigned to groups. The statistic test is developed and the F (for example) è annotated

$$p - \text{value} = \frac{\text{numero di } F - \text{value} > F - \text{value del data set originale}}{\text{numero di permutazioni}}$$

VOLCANO PLOT

Volcano plot combines the relative intensity (fold-change, x axis) and the results of statistical tests (P-values, y axis)

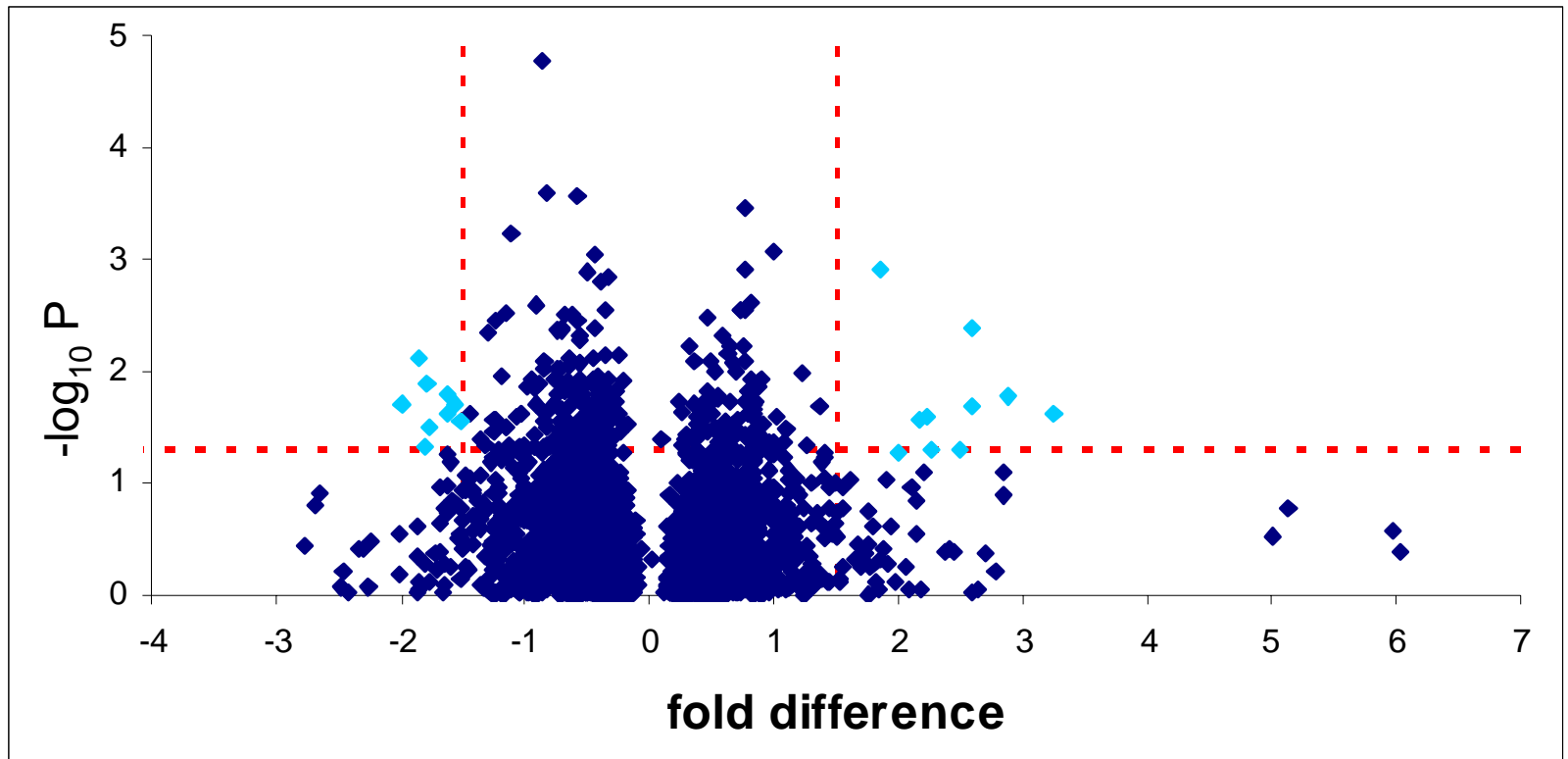
T	$\text{Log}_2 T$	Fold-change
$\frac{1}{1} = 1$	0	0
$\frac{2}{1} = 2$	1	1
$\frac{1}{2} = 0.5$	-1	-1
$\frac{4}{1} = 4$	2	2
$\frac{1}{4} = 0.25$	-2	-2

p-value	Log_{10}	$-\text{Log}_{10}$
0,2	-0,7	0,7
0,05	-1,3	1,3
0,01	-2,0	2,0
0,001	-3,0	3,0
0,0001	-4,0	4,0

Results: differentially expressed genes

VOLCANO PLOT

time points -5 e 6



Results: clustering of significant genes

OBJECTIVE

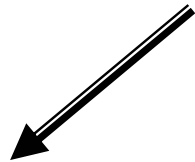


Clustering of differentially expressed genes

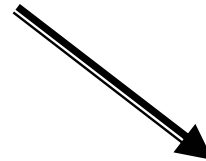
Solution



Cluster analysis



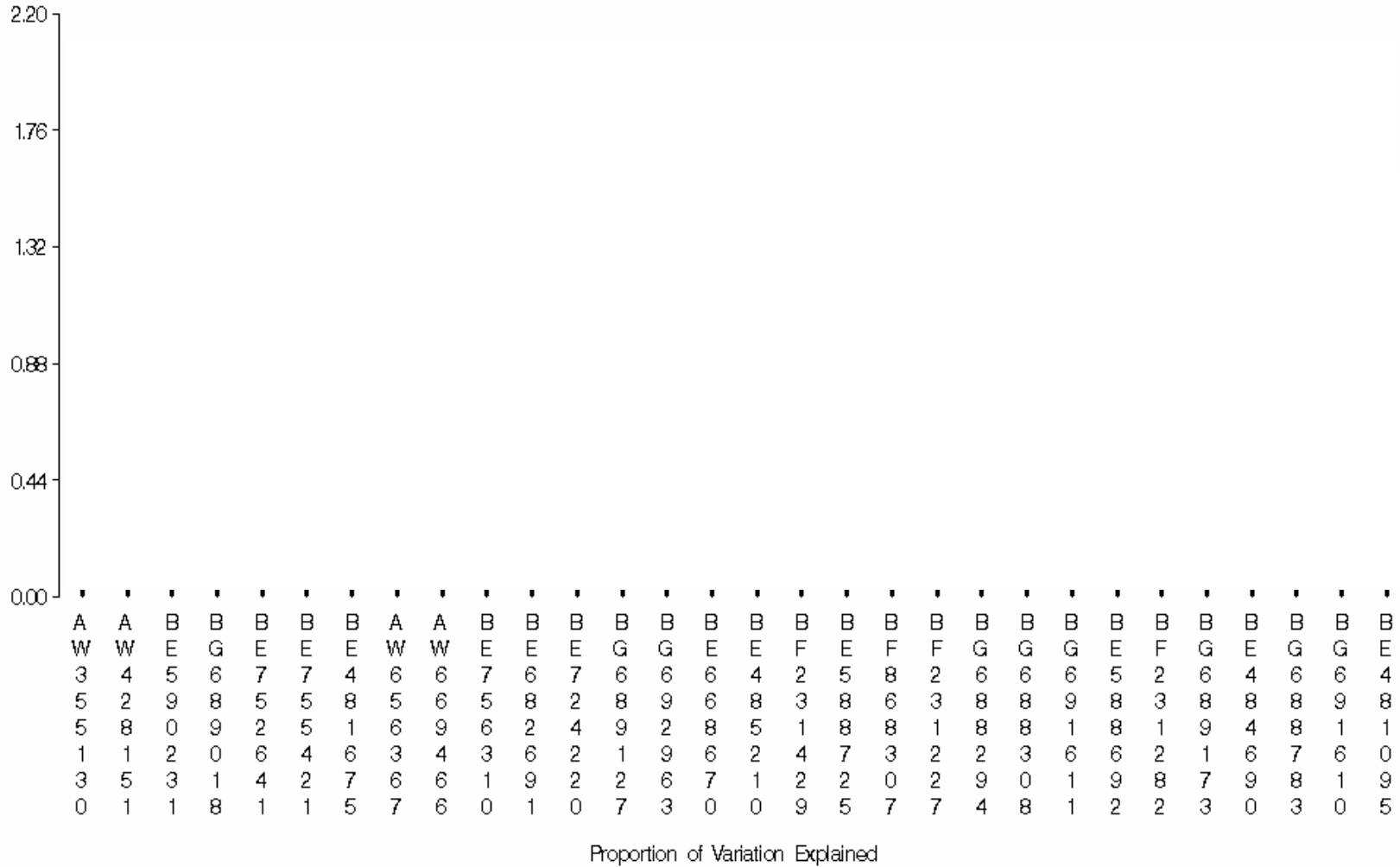
Hierarchical clustering



Trajectory clustering

Results: clustering of significant genes

Hierarchical clustering



Problems

- ❑ Hypothesis test
- ❑ Clusters cut-off
- ❑ Several clustering methods

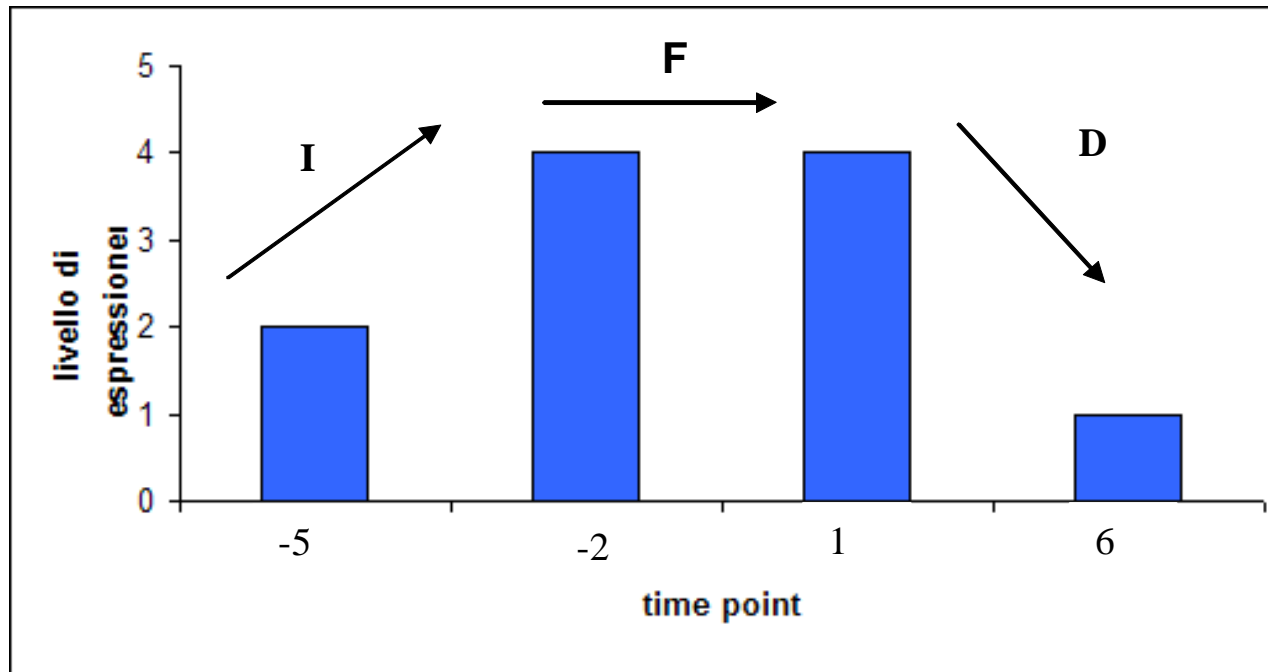
Problem solution



Combine different clustering methods

Trajectory clustering

Genes are grouped on the basis of the direction of the expression change between adjacent time points



$$3^{T-1} = 3^{4-1} = 3^3 = 27$$

possible trajectories

Results: clustering of significant genes

