Theoretical limitations of massively parallel biology

Genetic network analysis – gene and protein expression measurements

Zoltan Szallasi
Children's Hospital
Informatics Program
Harvard Medical School
www.chip.org

Vipul Periwal
Gene Network Sciences Inc.

Mattias Wahde
Chalmers University

John Hertz (Nordita)
Greg Klus (USUHS)

How much information is needed to solve a given problem ?

↕

How much information is (or will be) available ?

Conceptual limitations

Practical limitations

- **Finding transcription factor binding sites based on primary sequence information**

- **SNP < > disease association**

What are the problems we want to solve ?

So far the “DNA chip” revolution has been mainly technological:

The principles of measurements (e.g. complementary hybridization) have not changed.

It is not clear yet whether a conceptual revolution is approaching as well ?

potential breakthrough questions:

- **can we perform efficient, non-obvious reverse engineering ?**
- **can we identify non-dominant cooperating factors ?**
- **can we predict truly new subclasses of tumors based on gene expression patterns ?**
- **can we perform meaningful (non-obvious & predictive) forward modeling**

1. **Reverse engineering time series measurements**

2. **Identification of novel classes or separators in gene expression matrices in a statistically significant manner**

3. **Potential use of artificial neural nets (machine learning) in the analysis of gene expression matrices.**

Biological research has been based on the discovery of strong dominant factors.

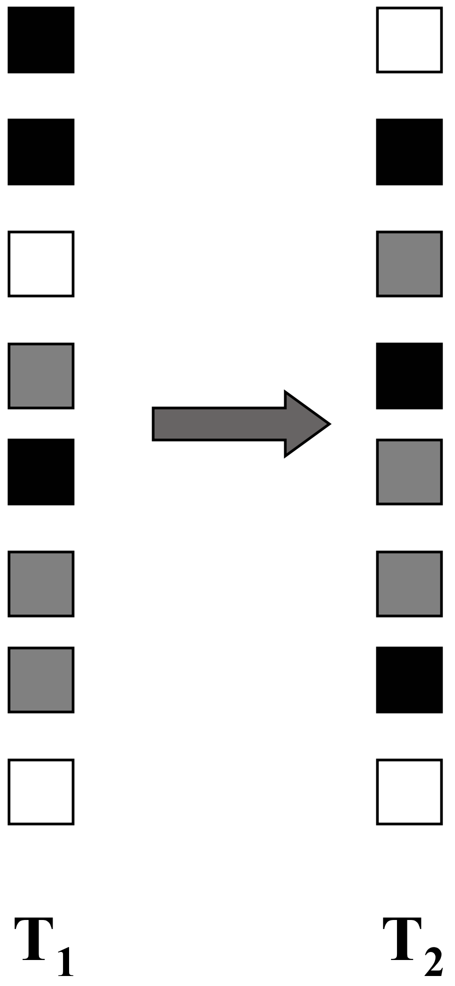
More than methodological issue ?

Robust network based on stochastic processes

↕

Strong dominant factors

The Principle of Reverse Engineering of Genetic Regulatory Networks from time series data:



Determine a set of regulatory rules that can produce the gene expression pattern at T_2 given the gene expression pattern at the previous time point T_1

Continuous modeling:

$$x_i(t+1) = g\,(b_i + \sum_j w_{ij} x_j(t))$$

(Mjolsness et al, 1991 - connectionist model;
Weaver et al., 1999, - weight matrix model;
D'Haeseleer et al., 1999, - linear model;
Wahde & Hertz, 1999 - coarse-grained reverse engineering)

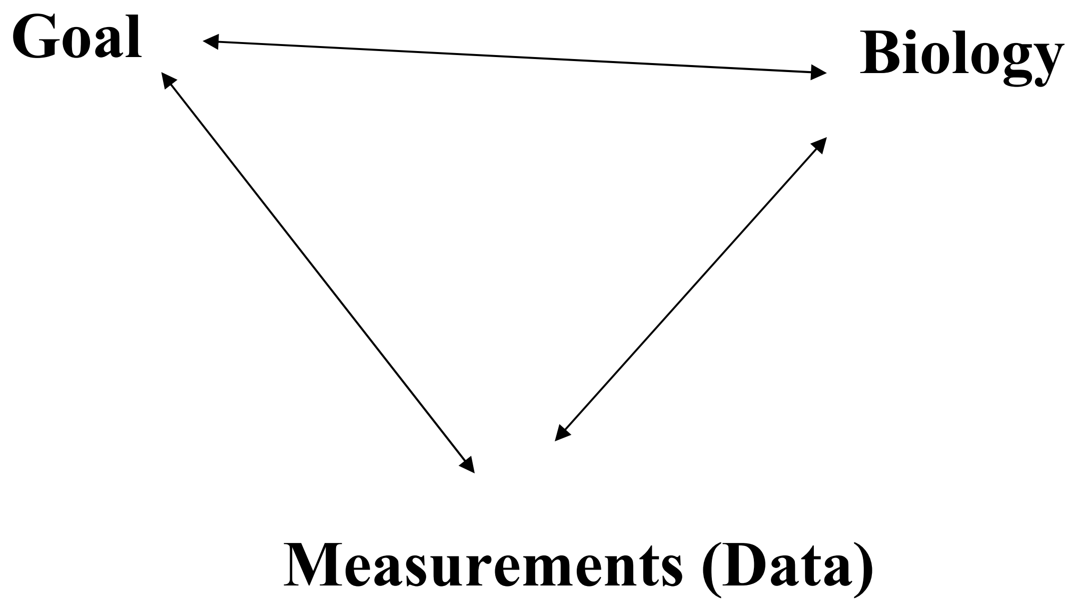
at least as many time points as genes: $T-1>N+2$
(Independently regulated entities)

For differential equations with r parameters 2r+1 experiments are enough for identification (E.D.Sontag, 2001)

How much information is needed for reverse engineering?

Boolean fully connected	2^N
Boolean, connectivity K	$K\ 2^K \log(N)$
Boolean, connectivity K, linearly separable rules	$K \log(N/K)$
Pairwise correlation	$\log (N)$

N = number of genes
K = average regulatory input/gene



Goal

Biology

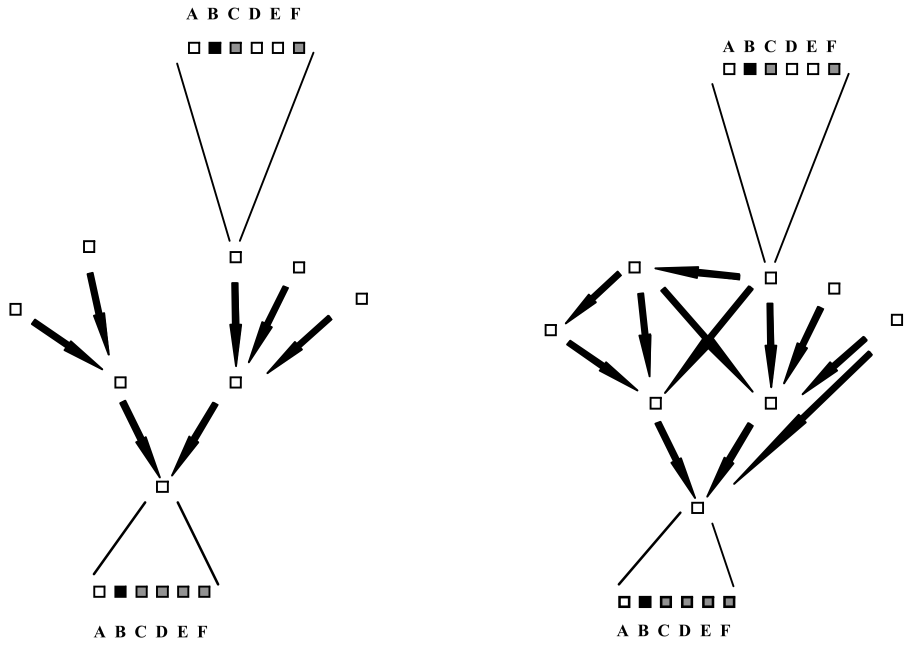
Measurements (Data)

Biological factors that will influence our ability to perform successful reverse engineering.

(1) the stochastic nature of genetic networks ,

(2) the effective size of genetic networks ,

(3) the compartmentalization of genetic networks,


A B C D E F
A B C D E F
A B C D E F
A B C D E F

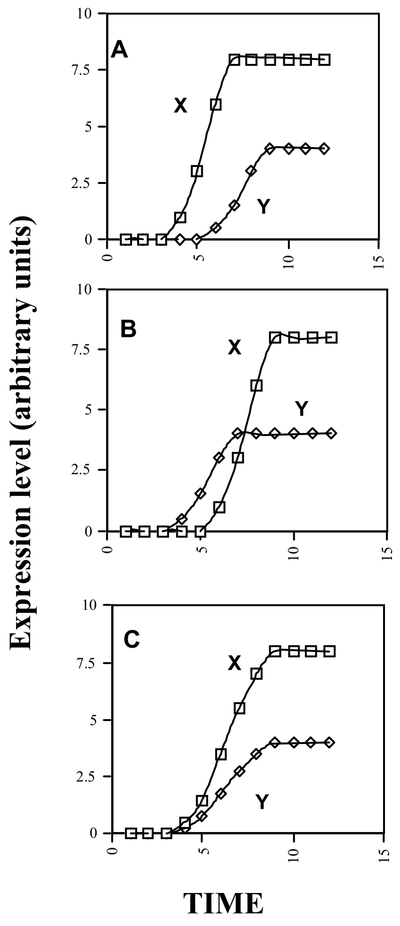
1. The prevailing nature of the genetic network

The effects of stochasticity:

1. It can conceal information (How much ?)

2. The lack of sharp switch on/off kinetics can reduce useful information of gene expression matrices.

(For practical purposes genetic networks might be considered as deterministic systems ?)



A
B
C
X
Y
0
2.5
5
7.5
10
0
5
10
15
Expression level (arbitrary units)
TIME

2. The effective size of the genetic network:

How large is our initial directed graph ?
(It is probably not that large.)

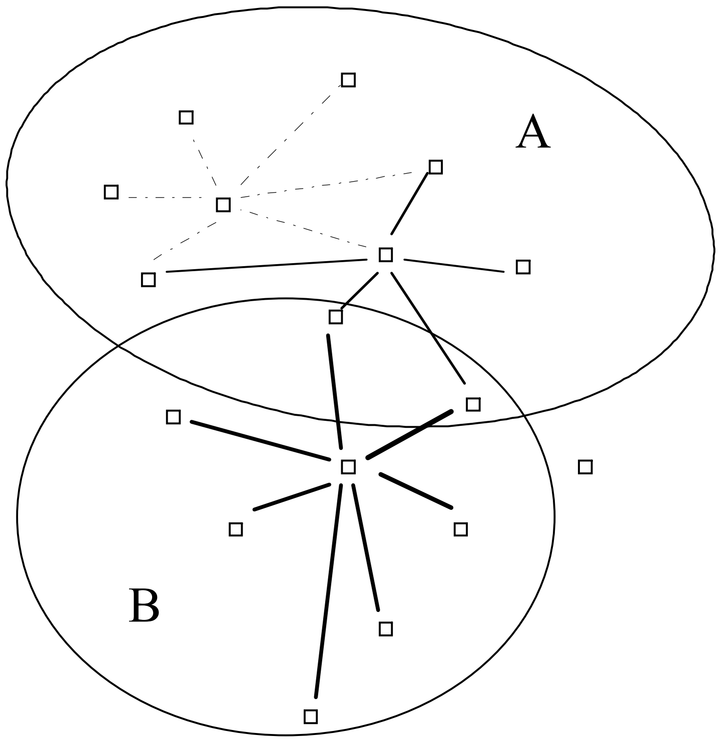
We might have a relatively well defined deterministic cellular network with not more than 10 times the number of total genes.

$$N_{bic} < 10 \text{ x } N_{gene}$$

10,000-20,000 active genes per cell
Splice variants <   > modules

3. The compartmentalization (modularity) of the genetic network:

The connectivity of the initial directed gene network graph
Low connectivity - better chance for computation.



Genetic networks exhibit:

Scale-free properties (Barabasi et al.)

Modularity

Flatness

(Useful) Information content of measurements
is influenced by the inherent nature of living systems

We can sample only a subspace of all gene expression patterns (gene expression space), because:

1. the system has to survive
(83% of the genes can be knocked out in S. cerevisiae)

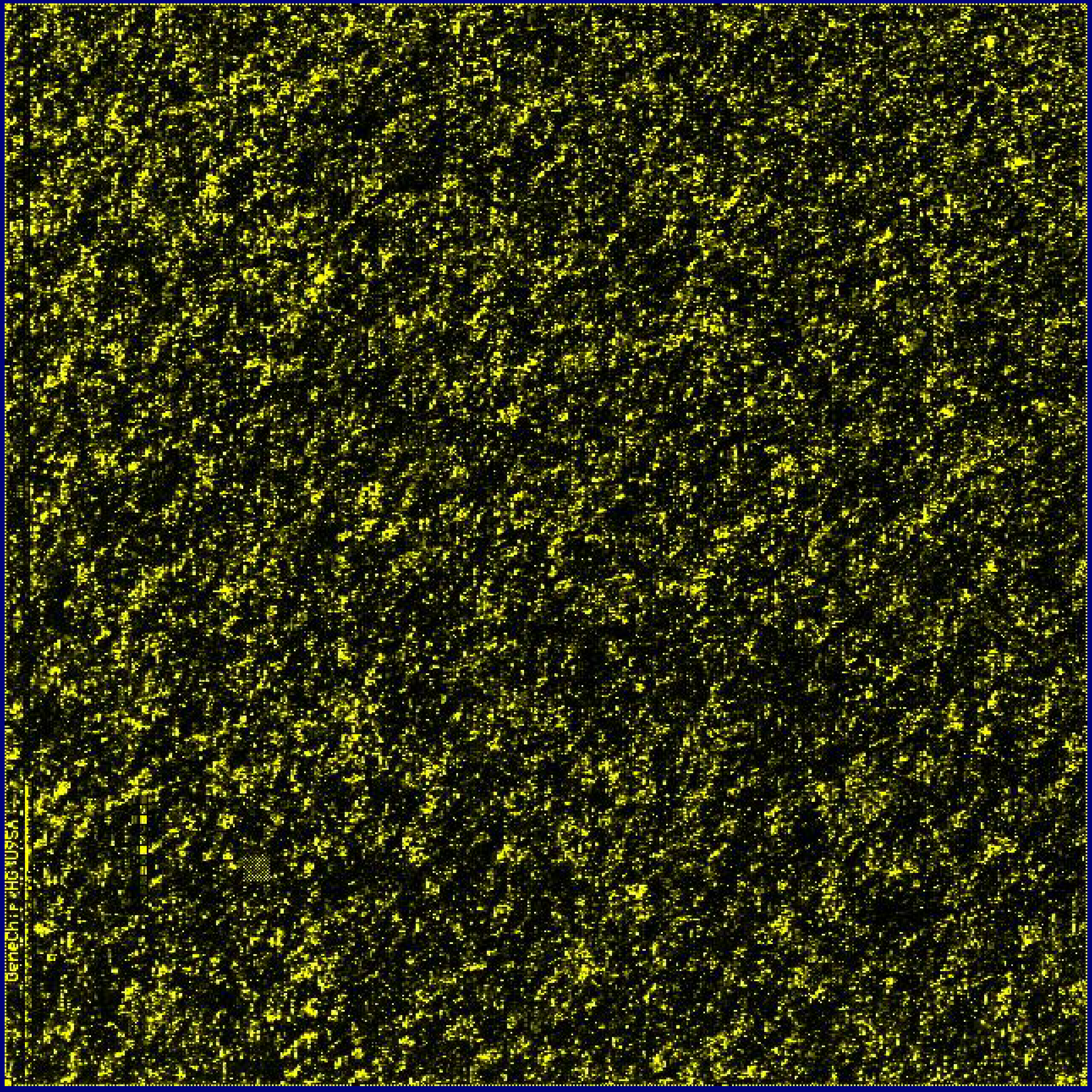
2. Gene-expression matrices (i.e. experiments) are coupled
Cell cycle of yeast under different conditions

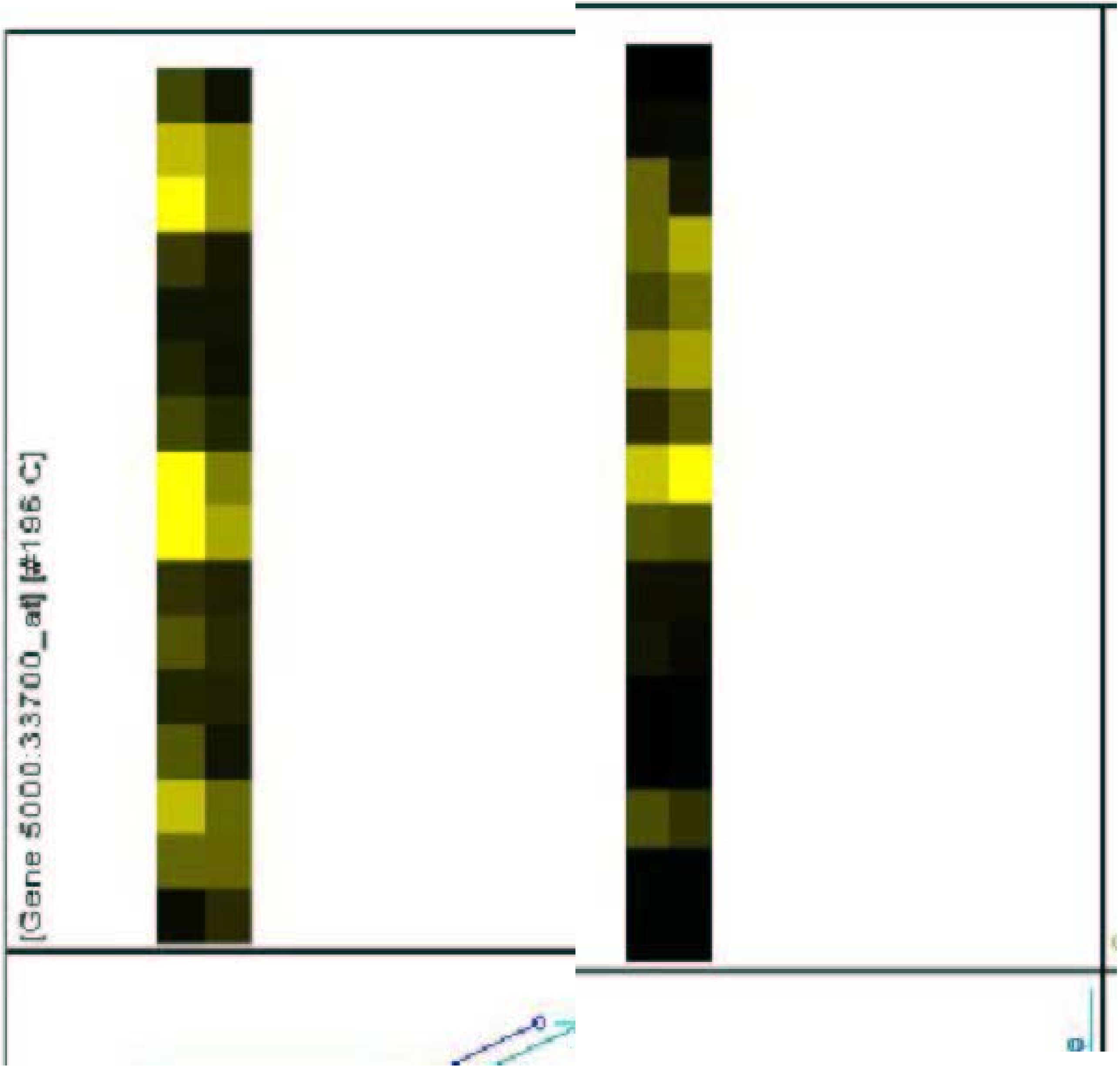
Data:

A reliable detection of 2-fold differences seems to be the practical limit of massively parallel quantitation.

(estimate: optimistic and not cross-platform)

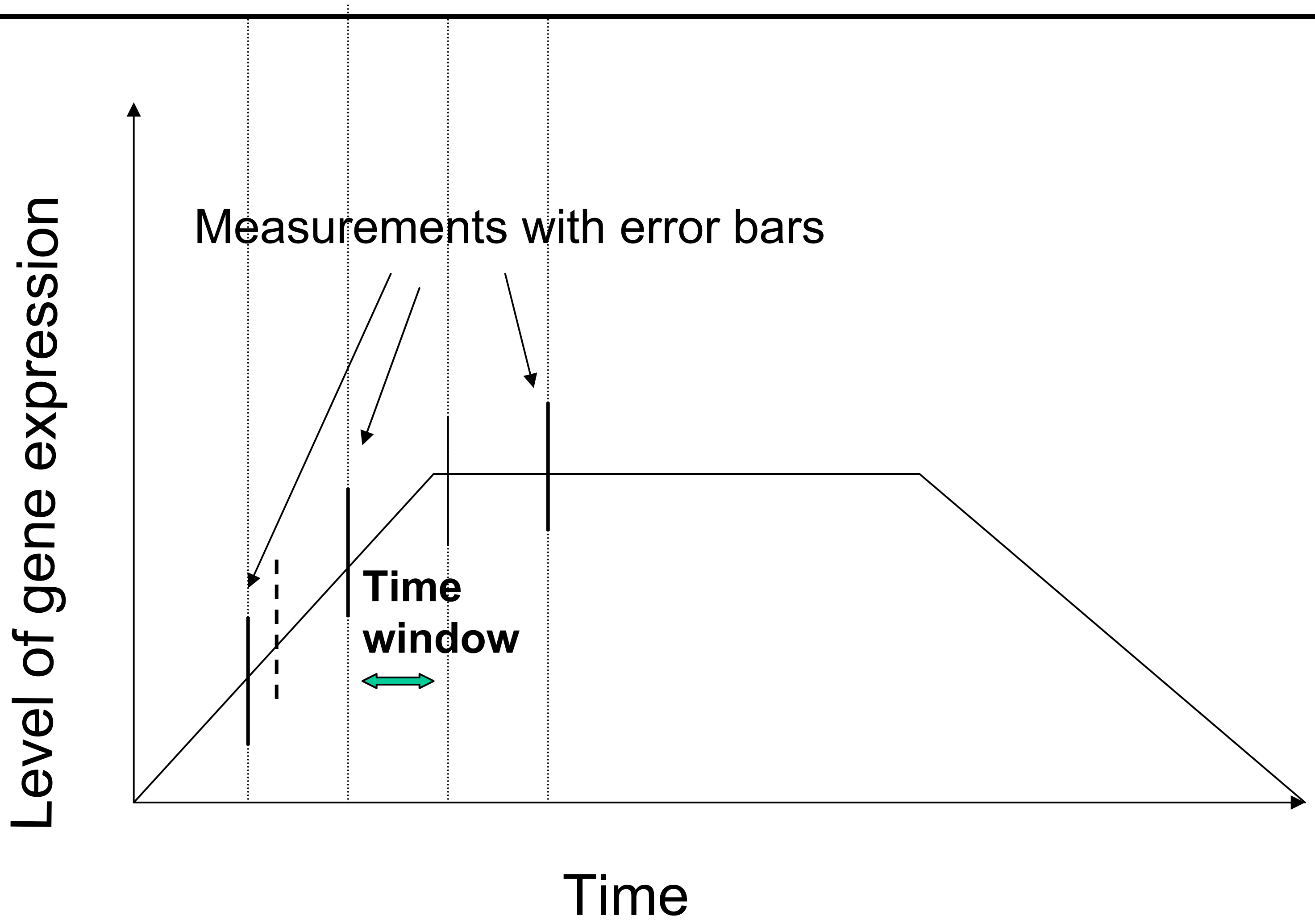
Population averaged measurements


GeneChip® HG U95A


[Gene 5000:33700_at] [#196 C]

The useful information content of time series measurements depend on:

1. Measurement error (conceptual and technical limitations, such as normalization)
2. Kinetics of gene expression level changes (lack of sharp switch on/off kinetics - stochasticity ?)
3. Number of genes changing their expression level.
4. The time frame of the experiment.



A rational experiment will sample gene-expression according to a time-series in which each consecutive time point is expected to produce at least as large expression level difference as the error of measurement: approximately 5 min intervals in yeast, 15-30 min intervals in mammalian cells.

$P = K \log(N/K)$ (John Hertz, Nordita)

P : gene expression states
N: size of network
K: average number of regulatory interactions

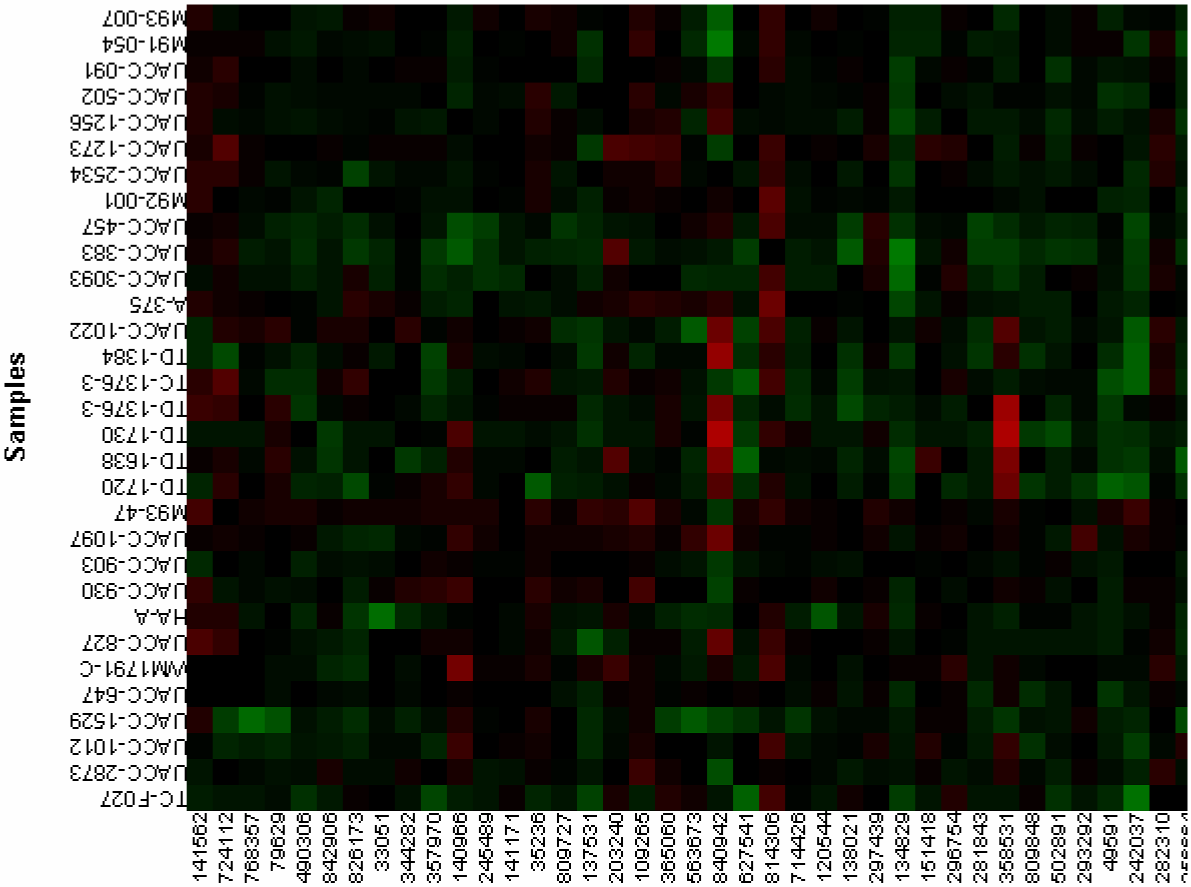
Applying all this to cell cycle dependent gene expression measurements by cDNA microarray one can obtain 1-2 orders of magnitude less information than expected in an ideal situation. (Szallasi, 1998)

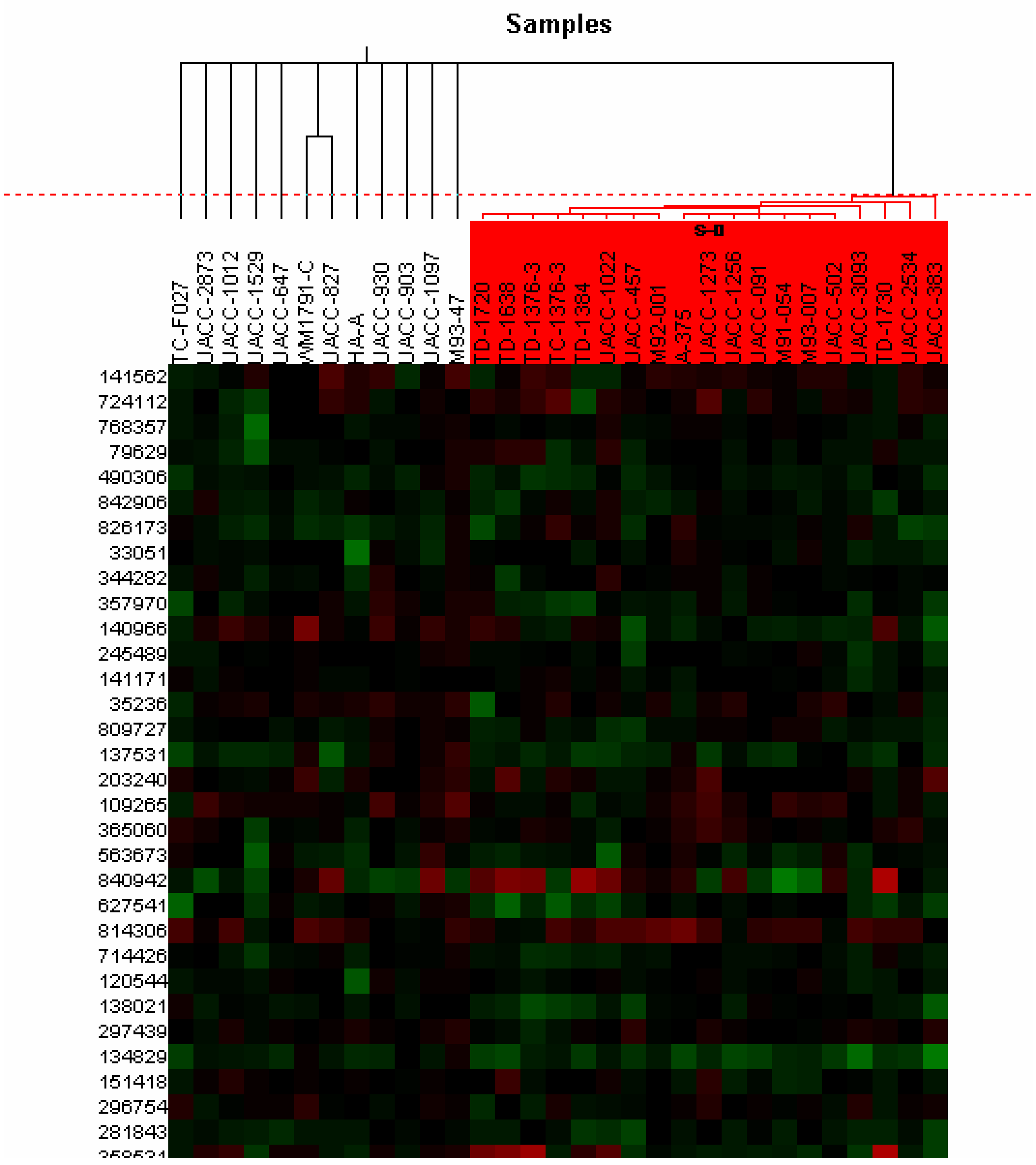
Can we identify non-dominant cooperating factors ?

Can we predict truly new subclasses of tumors based on gene expression patterns ?

How much data is needed ?

How much data will be available ?


Samples
M93-007
M91-054
UACC-091
UACC-502
UACC-1256
UACC-1273
UACC-2534
M92-001
UACC-457
UACC-383
UACC-3093
A-375
UACC-1022
TD-1384
TC-1376-3
TD-1376-3
TD-1730
TD-1638
TD-1720
M93-47
UACC-1097
UACC-903
UACC-930
HA-A
UACC-827
WM1791-C
UACC-647
UACC-1529
UACC-1012
UACC-2873
TC-F027
141562
724112
768357
79629
490306
842906
826173
33051
344282
357970
140966
245489
141171
35236
809727
137531
203240
109265
365060
563673
840942
627541
814306
714426
120544
138021
297439
134829
151418
296754
281843
358531
809848
502891
293292
49591
242037
282310


Samples
S-0
TC-F027
UACC-2873
UACC-1012
UACC-1529
UACC-647
WM1791-C
UACC-827
HA-A
UACC-930
UACC-903
UACC-1097
M93-47
TD-1720
TD-1638
TD-1376-3
TC-1376-3
TD-1384
UACC-1022
UACC-457
M92-001
A-375
UACC-1273
UACC-1256
UACC-091
M91-054
M93-007
UACC-502
UACC-3093
TD-1730
UACC-2534
UACC-383
141562
724112
768357
79629
490306
842906
826173
33051
344282
357970
140966
245489
141171
35236
809727
137531
203240
109265
365060
563673
840942
627541
814306
714426
120544
138021
297439
134829
151418
296754
281843

<u>Analysis of massively parallel data sets</u>

Unsupervised - avoiding artifacts in random data sets

avoiding artifacts in data sets retaining the internal data structure

Supervised

INFORMATION REQUIREMENT

Consistently mis-regulated genes in random matrices

“E” different samples

“N”-gene microarray

M_i genes mis-regulated in the “i”-th sample,

K consistently mis-regulated across all E samples.

What is the probability that (at least) K genes were mis-regulated by chance ?

$$\mathbf{P(E, k \geq K) = 1 - \sum_{i=0}^{K-1} P(E, k)}$$

Where P(E,k) is the probability that exactly k genes are consistently mis-regulated

$$\mathbf{P(E, k) = \sum_{j=k}^{M} \frac{\binom{j}{k} * \binom{N-j}{M-k} * P(E-1, j)}{\binom{N}{M}}}$$

$$\mathbf{P(2, k) = \frac{\binom{M}{k} * \binom{N-M}{M-k}}{\binom{N}{M}}}$$

If N>>M, then

$$P(E, k \geq K) \approx \frac{\binom{N}{K} * \binom{N-K}{M-K}^{E}}{\binom{N}{M}^{E}}$$

or

$$P(E, k \geq K) \approx \binom{N}{k} * (q^{E})^{k} * (1 - q^{E})^{N-k}$$

For a K gene separator:

$$n_K = \binom{N}{K} * \left(1 - q^K\right)^E$$

N	M	E	K	n_K simulated	n_K calculated
500	100	4	3	1172455±123637	1174430
500	100	8	3	69630 ± 17487	66605
300	50	15	3	760 ± 579	785
200	40	20	4	2032 ± 1639	1713

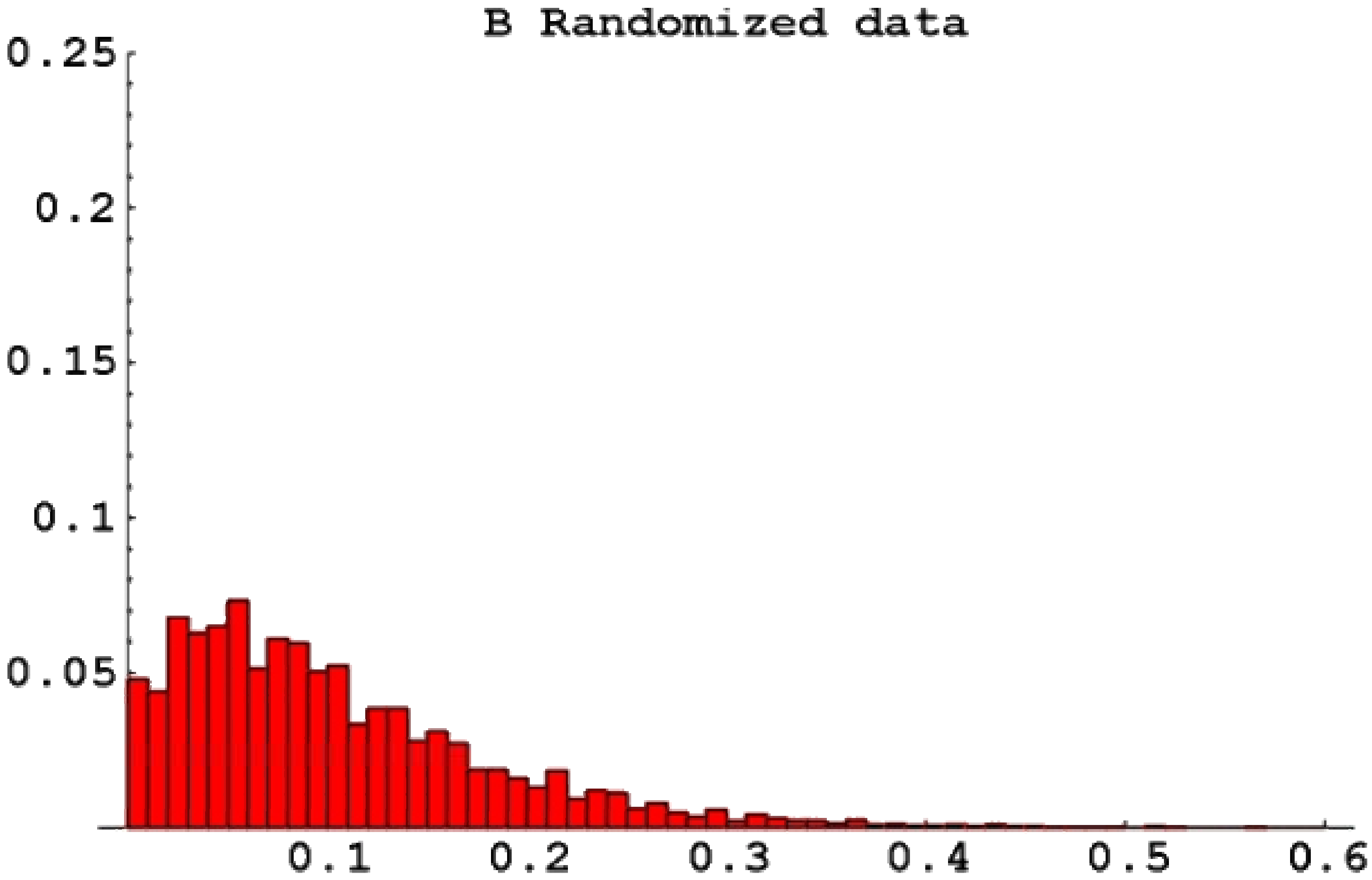
how many cell lines do we need in order to avoid accidental separators ?

for $N=10000$ $M=1000$ for $p<0.001$

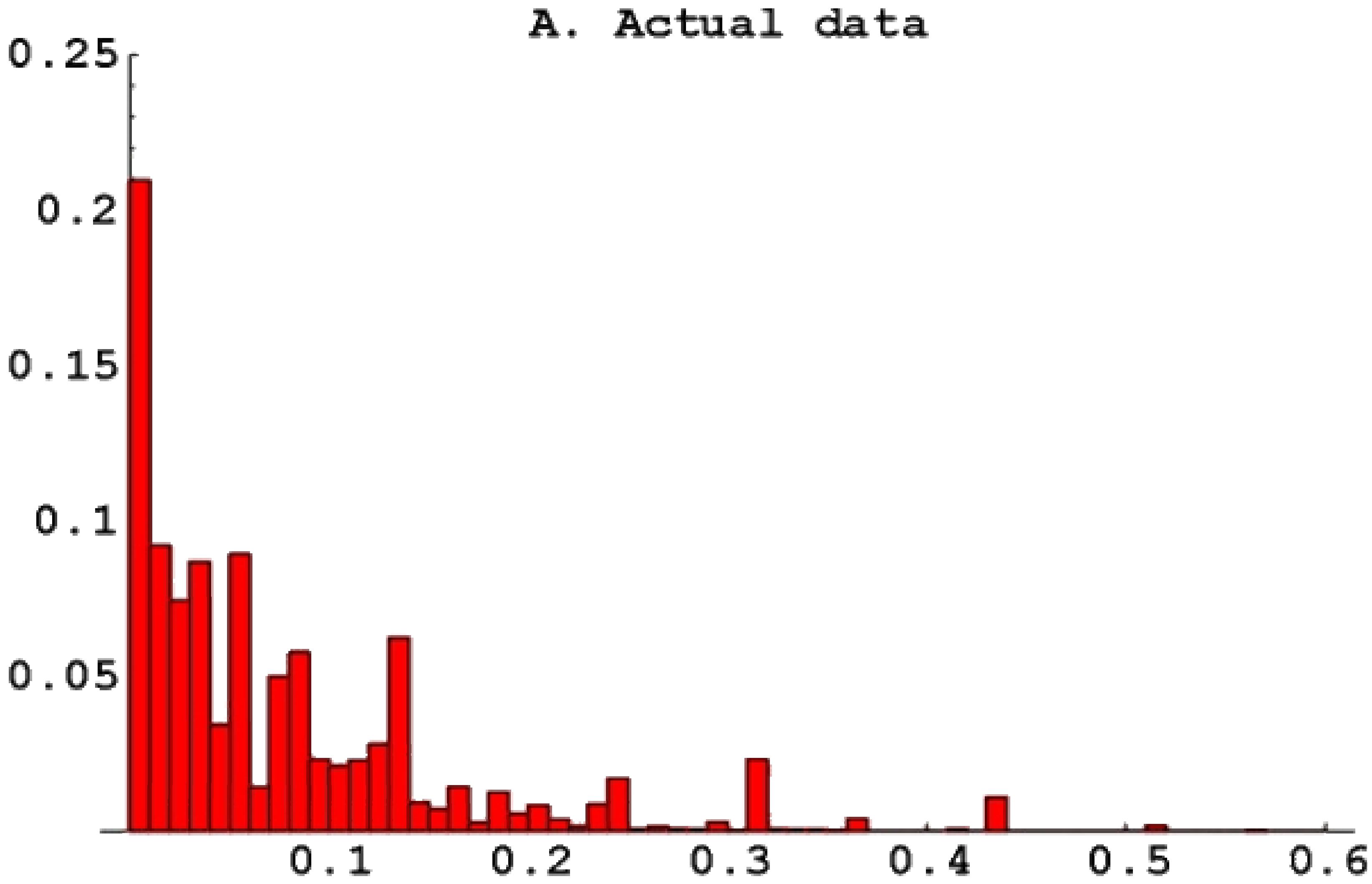
K=1	E=7

Higher order separator

K=2	E=15
K=3	E=25
K=4	E=38
K=5	E=54
K=6	E=73


B Randomized data
0.25
0.2
0.15
0.1
0.05
0.1
0.2
0.3
0.4
0.5
0.6

Genes are not independently regulated



Generative models (gene expression operator) will simulate <u>realistic looking</u> gene expression matrices ?

- the number of genes that can be mis-regulated
- the independence of gene mis-regulation.

	N_1	N_2	N_3	. . . .	N_i	T_1	T_2	T_3	. . . .	T_i
$gene_1$	0	0	1		0	1	1	1	…………………………...	0
$gene_2$	0	0	0		0	0	1	0	……………………….. ..	1
$gene_3$	0	0	0		0	1	0	1	…………………………...	0
$gene_4$	0	1	1		0	0	0	1	……………………………	1

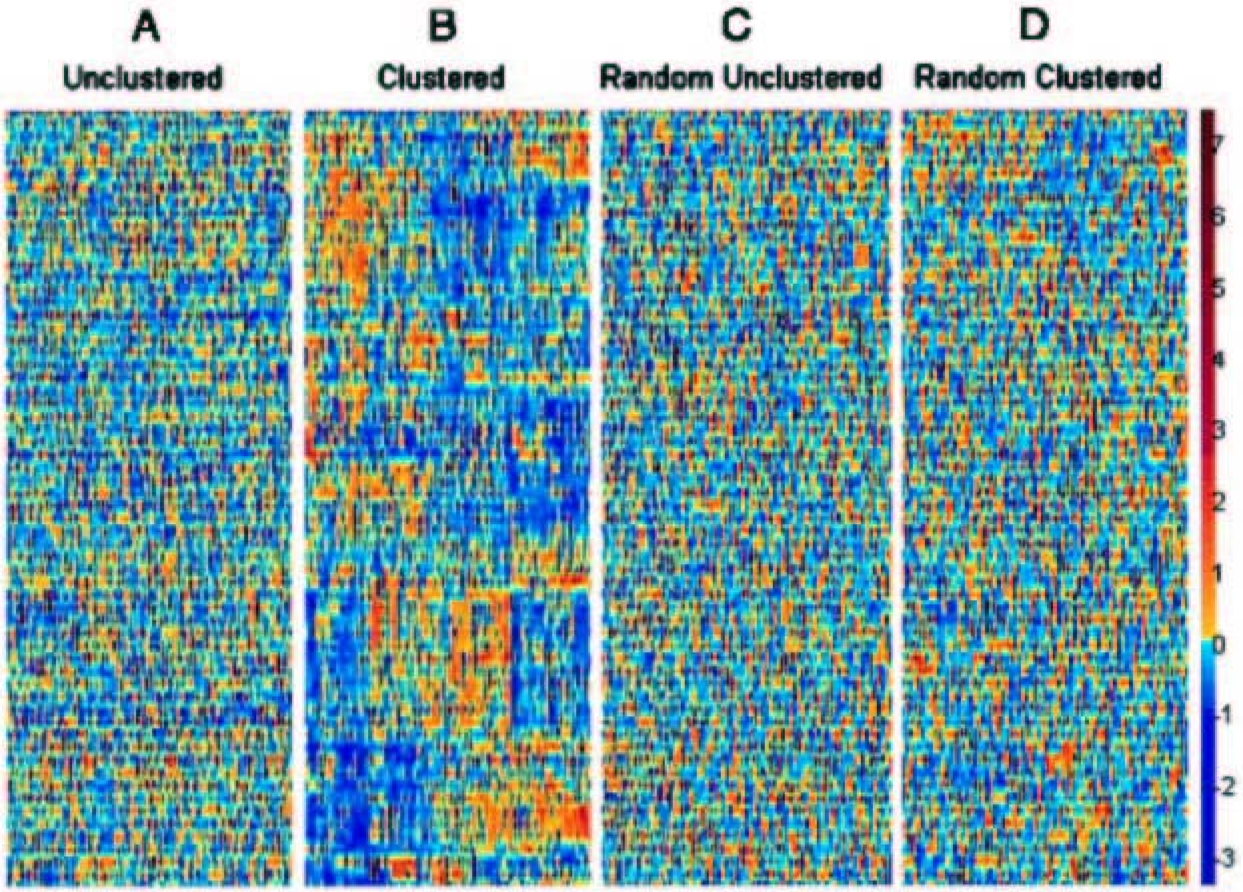
<u>Algorithm</u> to extract Boolean separators from a gene expression matrix.

U. Alon data set (colon tumors) : N=2000, $M_{average}$=180   K=2

E	Alon data	calc.	Num. sim.
10	708	131	130
11	120	~ 1	1
12	45	8.6×10^{-3}	8.6×10^{-3}
13	3	7.0×10^{-5}	-
14	3	5.6×10^{-7}	-
15	1	4.6×10^{-9}	-
16	1	3.7×10^{-11}	-

Generative model: 4+/-2 separators

features, while preserving step-like changes in intensity. The features were arranged in the order they appear in the EST sequence, the PM-MM intensities in a moving window of five features were sorted, and the filtered intensity was given by the mean of the middle three sorted intensities. The total intensity



coefficient of x a

The binary tre
to reorganize the
this end, we incl
in a determinist

Fi
norn
are t
the s
the h
so its
is 1.
scale
62 ti
the c
on tl
(C)
data
in tl
rand
algo
are
colo

Pearson-disproportion of an array:

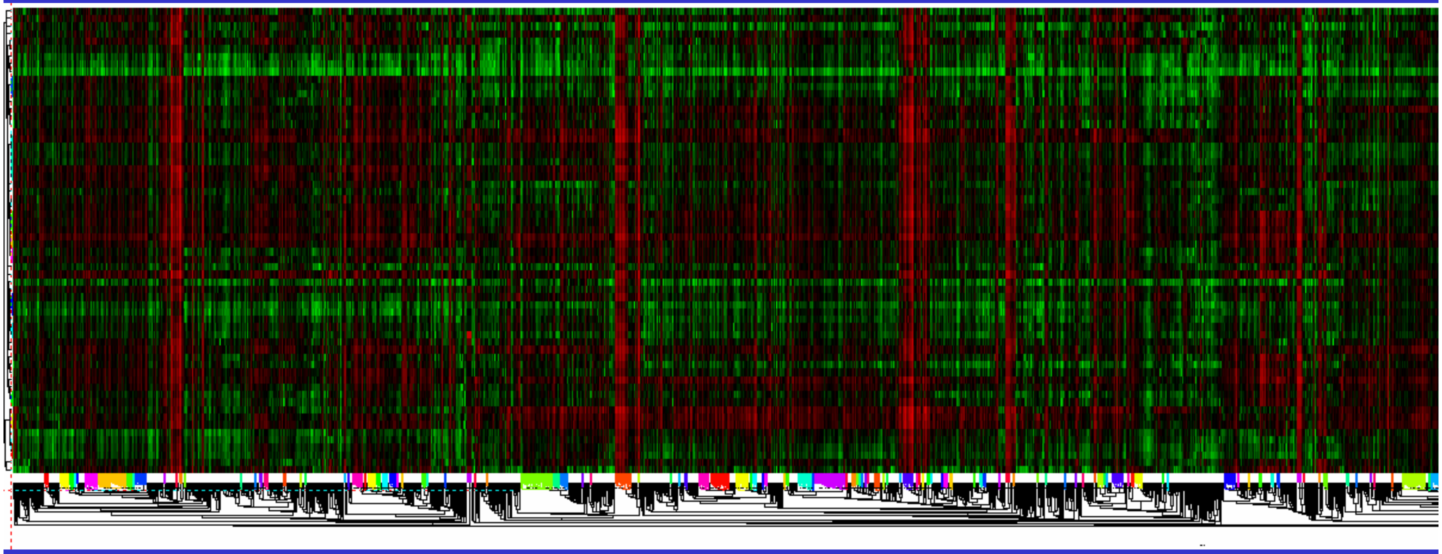
$$PE(\mathrm{y}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(y_{ij} - \frac{m_i n_j}{N})^2}{\frac{m_i n_j}{N}}$$

$\mathrm{y}_{\mathrm{ij}} =$ gene expression level in the *i*th row and *j*th column

$\mathrm{m}_{\mathrm{i}} = \sum_{\mathrm{j=1}}^{\mathrm{c}} \mathrm{y}_{\mathrm{ij}}$ $\qquad$ $\mathrm{n}_{\mathrm{j}} = \sum_{\mathrm{i=1}}^{\mathrm{r}} \mathrm{y}_{\mathrm{ij}}$

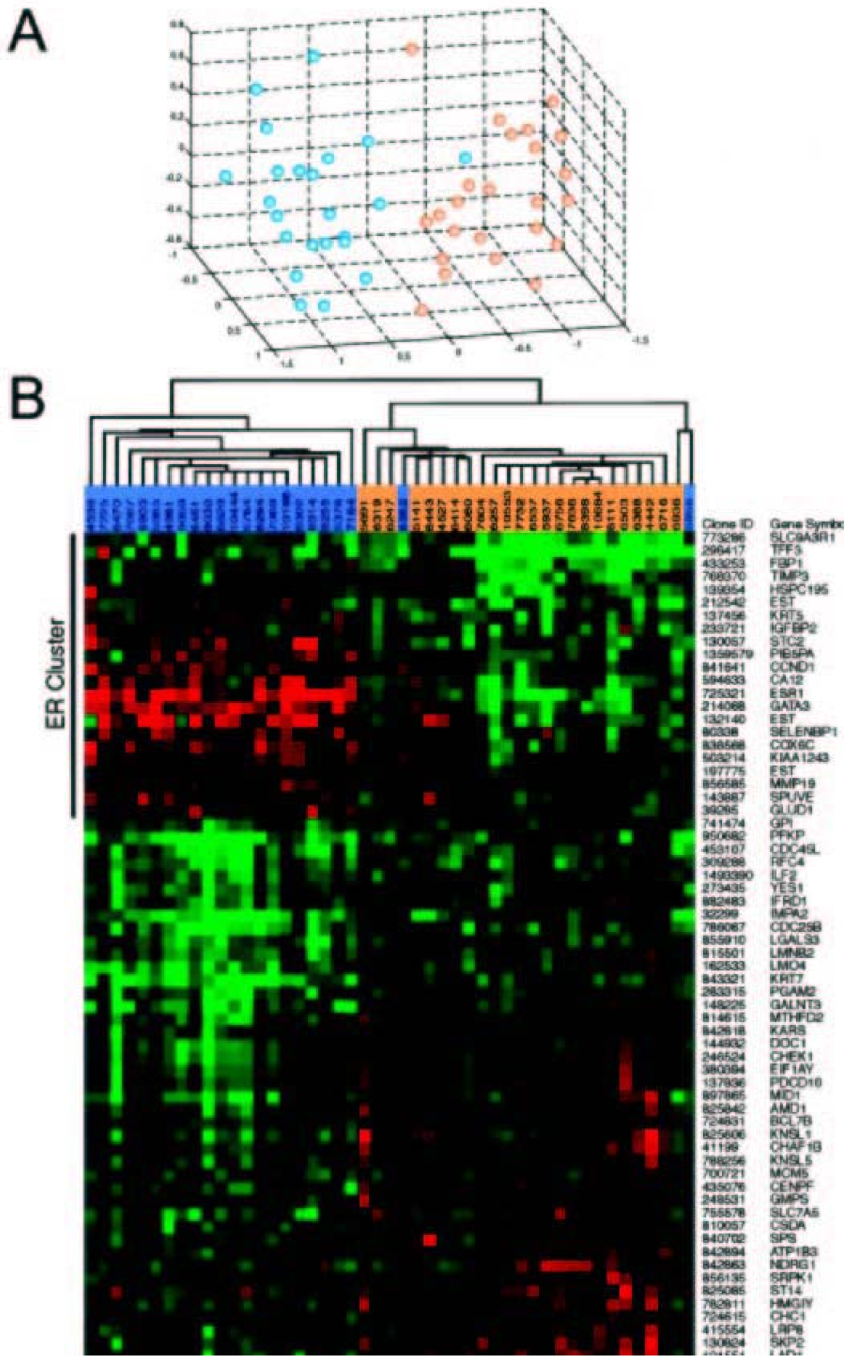
$\mathrm{N} = \sum_{\mathrm{i}} \mathrm{m}_{\mathrm{i}} = \sum_{\mathrm{j}} \mathrm{n}_{\mathrm{j}}$

Random matrices with the same intensity distribution and same (or larger) disproportion measure as the original matrix (Monte Carlo simulations)



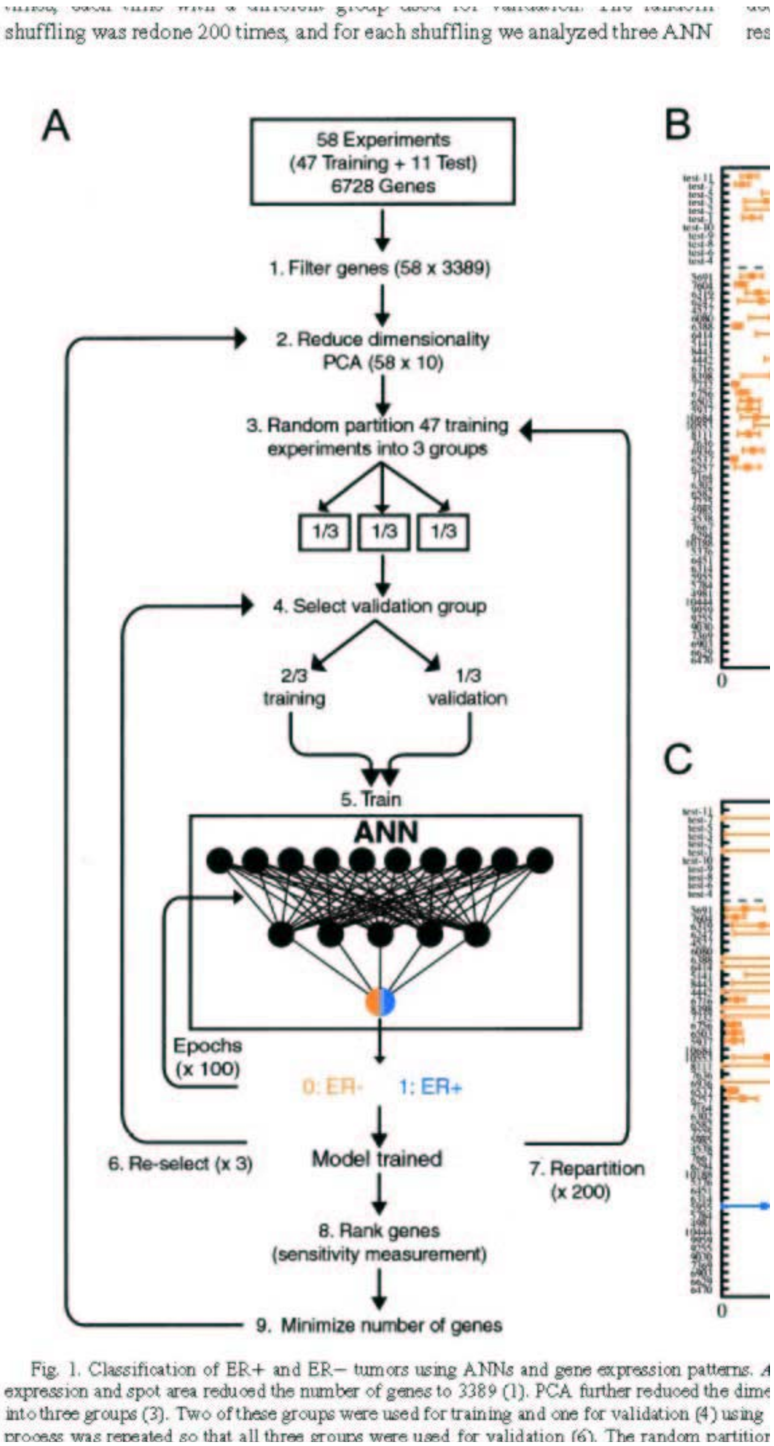
Generative models (random matrices retaining internal data structure) will help to determine the required sample number for statistically meaningful identification of classes and separators.

Machine learning – Artificial Neural Nets in the analysis Cancer associated gene expression matrices



he
is.
er-
ed
ns
ng
ed
nt
ay
ls.
us

as
nd
er-
ed
at
es
es
he
he
or
to
he



times, each time with a different group used for validation. The random
shuffling was redone 200 times, and for each shuffling we analyzed three ANN

Fig. 1. Classification of ER+ and ER− tumors using ANNs and gene expression patterns.
expression and spot area reduced the number of genes to 3389 (1). PCA further reduced the dime
into three groups (3). Two of these groups were used for training and one for validation (4) using
process was repeated so that all three groups were used for validation (6). The random partition

P. Meltzer,
J. Trent
M. Bittner

ANN (artificial neural nets) work well when a large number of samples is available relative to the number of variables

(e.g. for the pattern recognition of hand written digits one can create a huge number of sufficiently different samples).

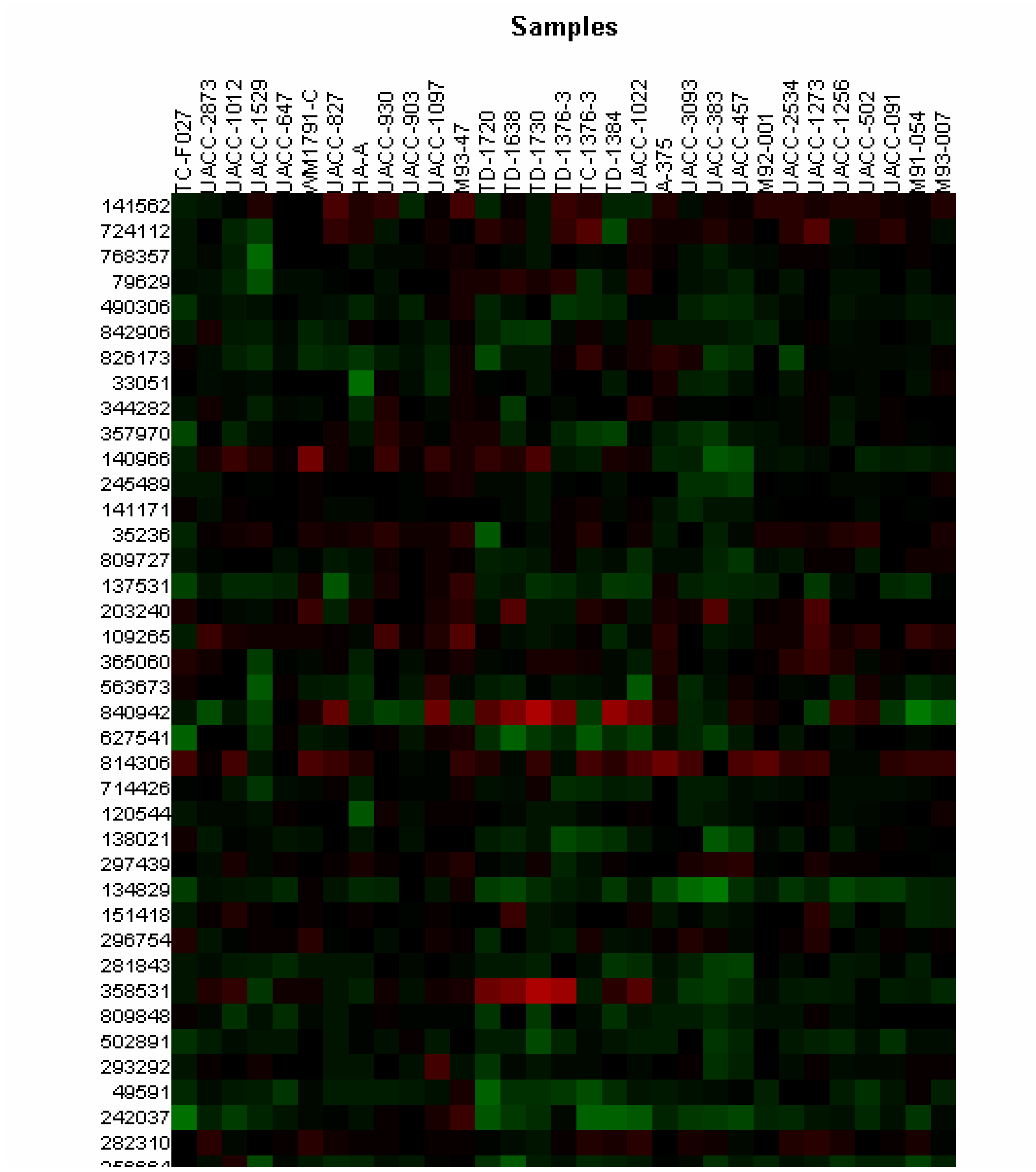
In biology there might be two limitations:

1. the number of samples might be quite limited, at least relative to the complexity of the problems (The cell has to survive)

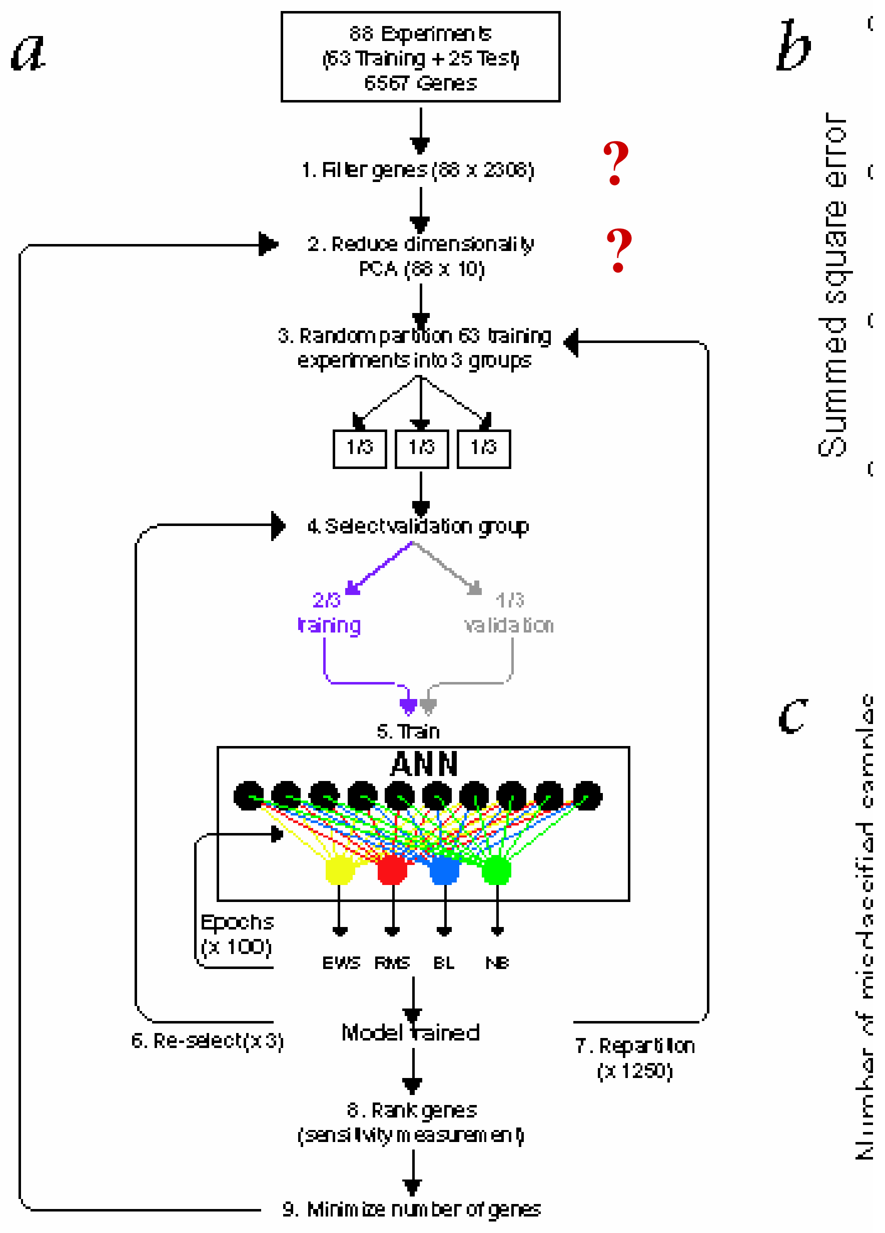
2. There might be a practical limit to collecting certain types of samples

< 100 samples

> 1000

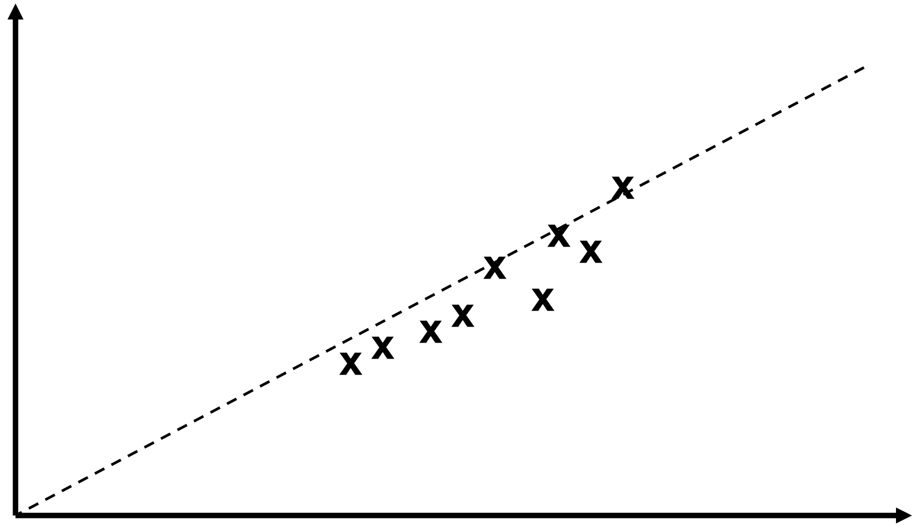


tic
:a-
ed
'e-
to
:x-
its
its
ne
nd
IN
n-
at-
ne
es
3
in
ng
ili-
ng
d-
ne
si-
ed
:x-
all
a-
or

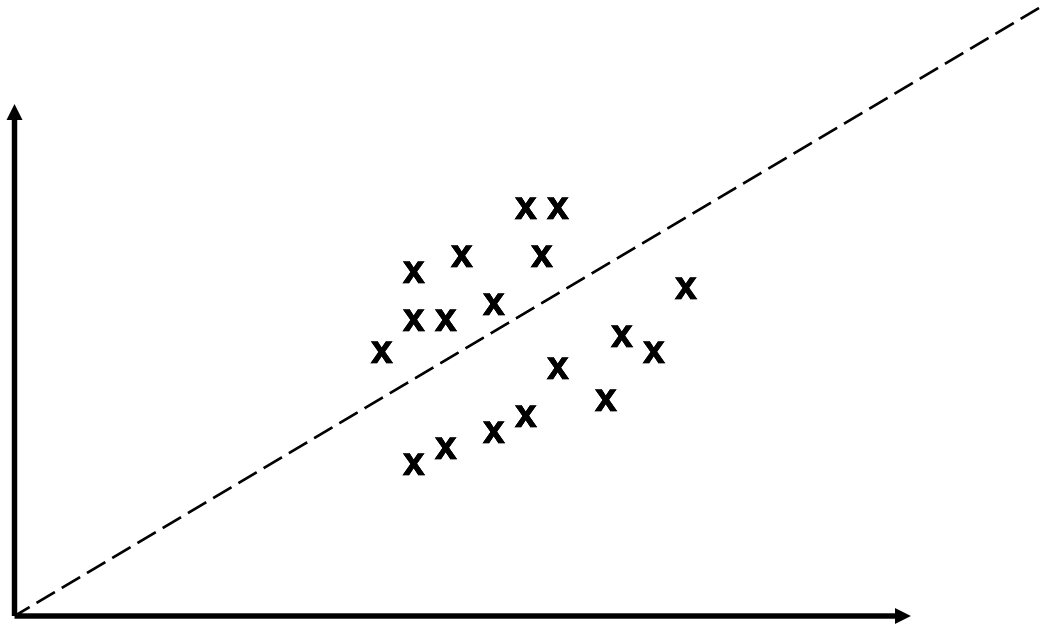


Reducing dimensionality

Principal component analysis
retain variance

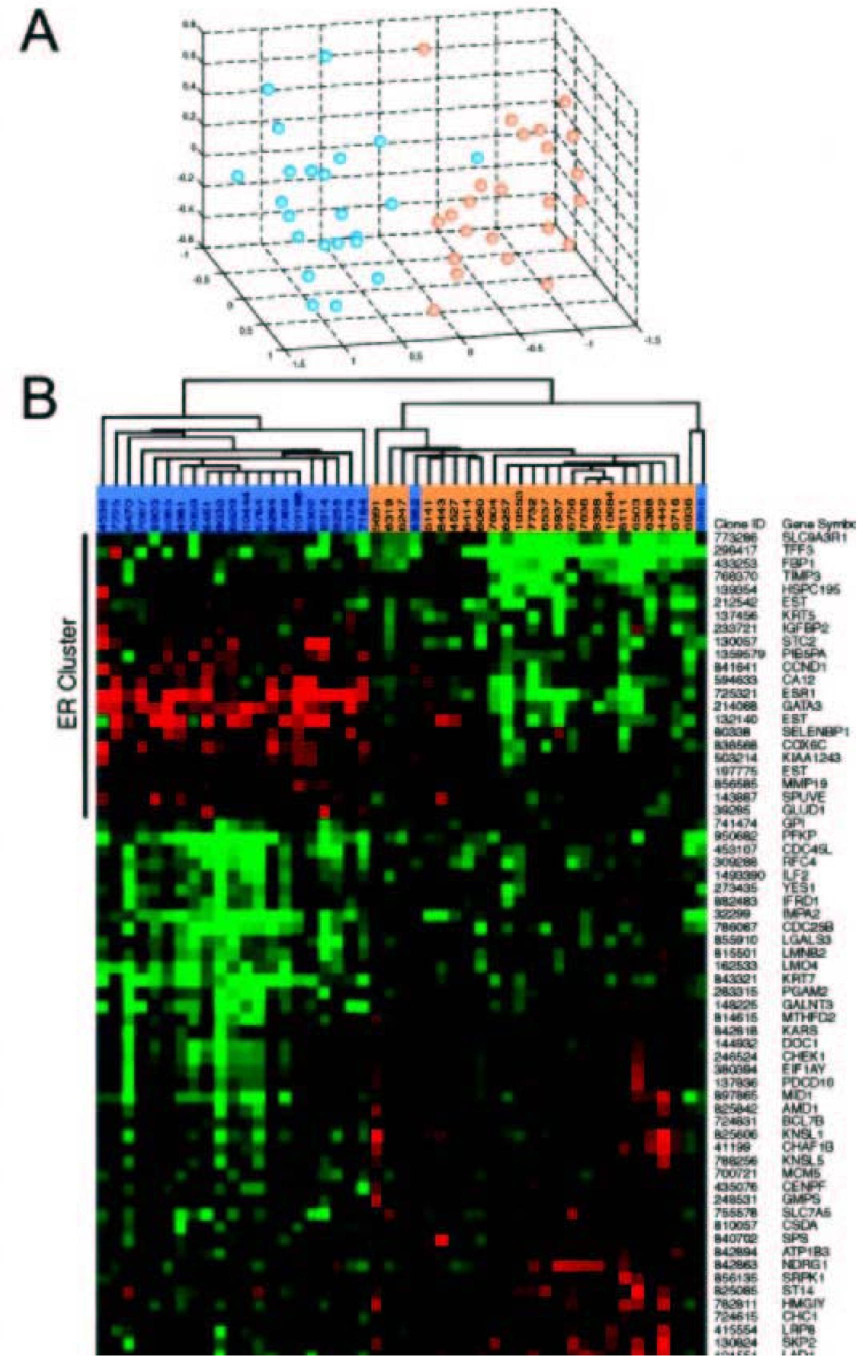


The risk of reducing dimensionality by PCA



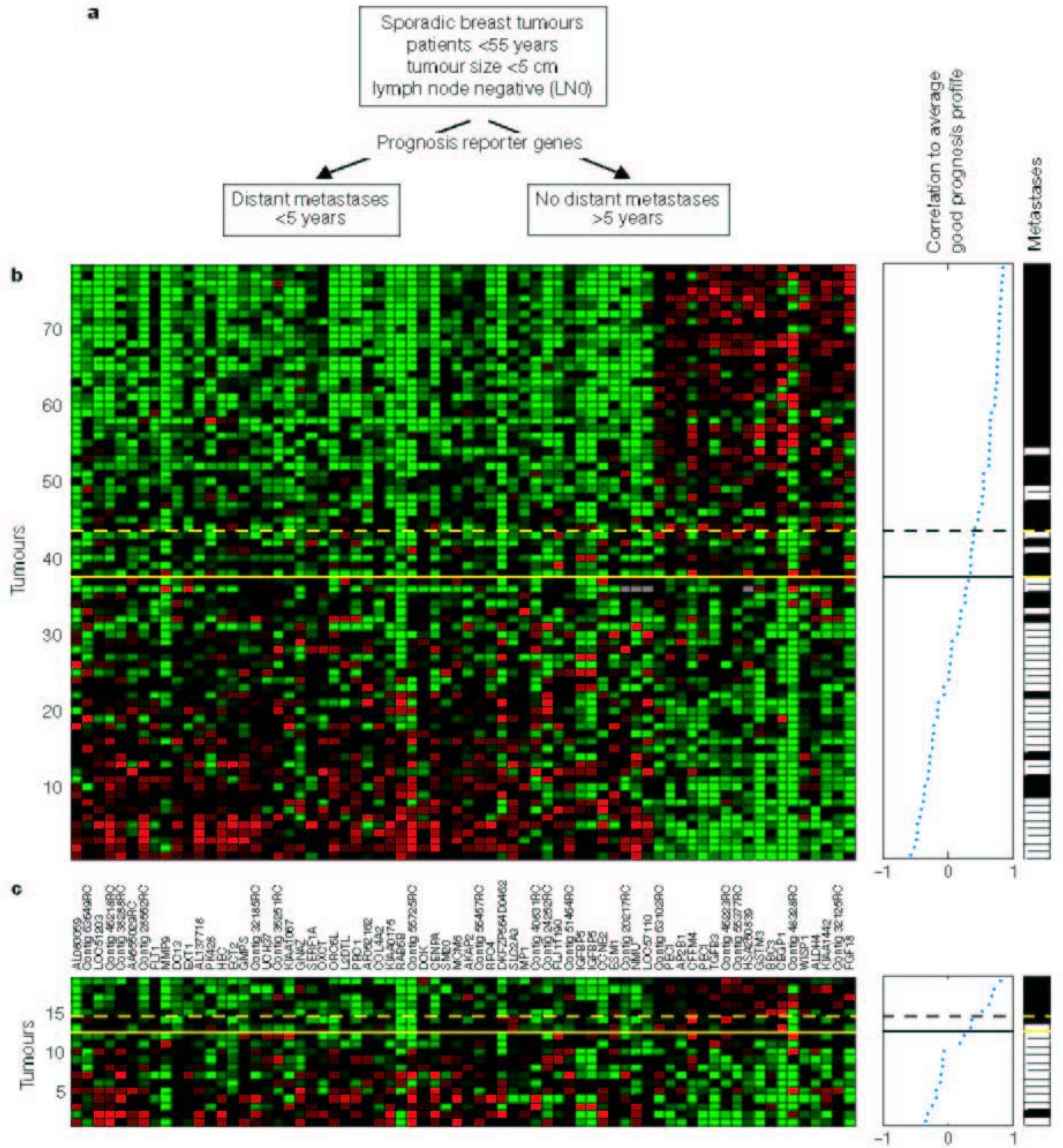
he
is.
er-
ed
ns
ng
ed
nt
ay
ls.
us

as
nd
er-
ed
at
es
es
he
he
or
to
he



(Rosetta)
83% accuracy
with 70 genes

Simple genetic
algorithm by us:
93% with 3 genes



gure 2 Supervised classification on prognosis signatures. **a**, Use of prognostic reporter
nes to identify optimally two types of disease outcome from 78 sporadic breast tumours

prognostic classifier with optimal accuracy; dashed line, with optimized sensitivity. Above
the dashed line patients have a good prognosis signature, below the dashed line the