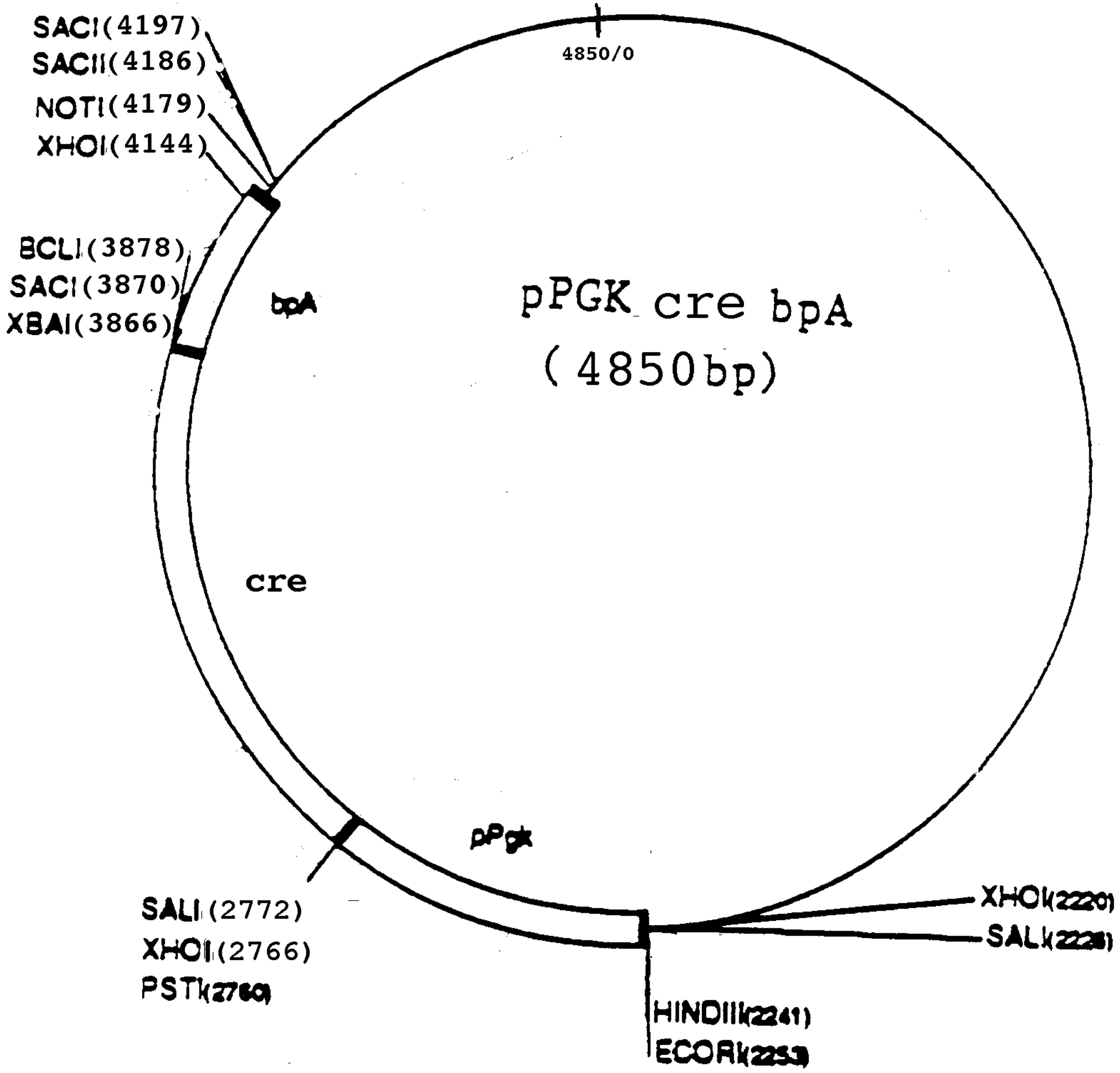
1. Why high-throughput (HT) data?

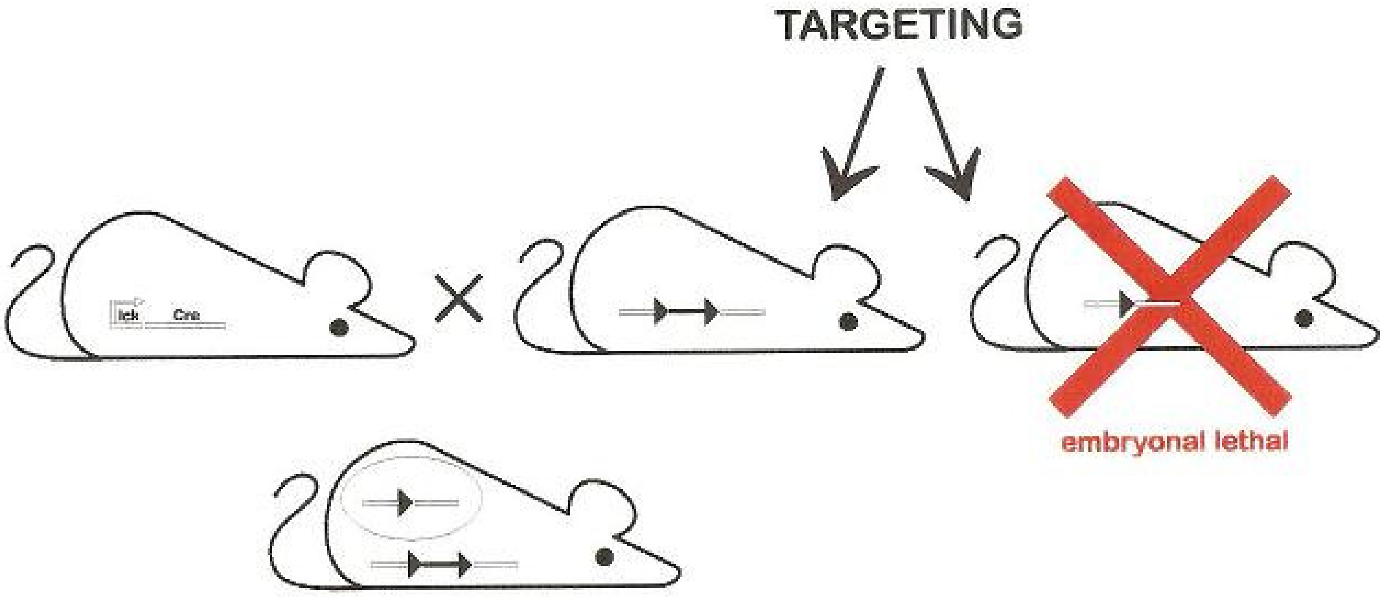
What is it that we cannot achieve by looking at only a few genes at a time?

Why high-throughput (HT) data?

- 1st example:
  Studying e.g. immunological coherences
  by knocking out genes

cre for eukaryotic expression





TARGETING

embryonal lethal

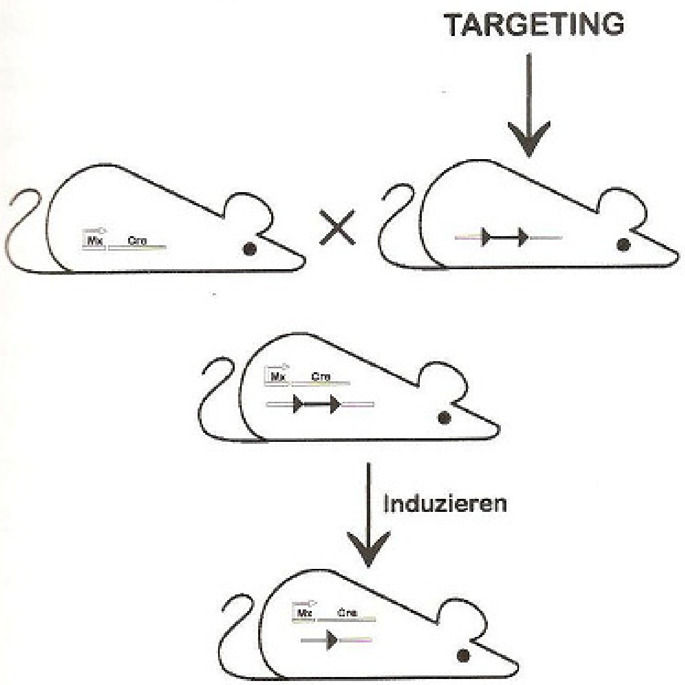


Abb. 1–3: Mx-cre System (Kuhn et al., 1995).

Der interferonabhängige Mx-Promotor dient hier zur induzierbaren Ausprägung von Cre (Kuhn et al., 1995). Bei anderen auf Transkriptionsebene kontrollierenden Systemen kann Cre sowohl induzierbar als auch zelltypspezifisch ausgeprägt werden.

Often:

✓ make construct ("flox" gene)
✓ transfect ES cells
✓ transfer to blastocyst
✓ transfer to foster mother
✓ get germline transmission
✓ cross with cre mice
✓ induce
✓ get considerable knockout rate

} at least
½ year
of work

☹ observing no phenotype

That sucks!

Why is that?

- Function of important genes is backuped by other pathways!

- In order to understand how that works

  (→ systems biology)

  one needs to know

  the status of **<u>many</u>** genes

Why high-throughput (HT)?

- 2nd example:
  cancer
- same problem:

About Cancer

- Cancer is a genetic disease
  - Not monogenic like MD or CF, but multigenic
- Cancer is caused by mutations in somatic cells
- Cancer can be caused by mutagens, chemicals that damage DNA, or viruses
- Cancer is caused by an accumulation of mutations in different genes in a single cell
- Cancer is caused by altered expression of genes or by accumulation of mutations in a single cell

Five Major Pathways: Cancer Cells

- There are five major pathways that must be activated or inactivated in a cell for the cell to become a cancer cell
  - Growth stimulatory signals
  - Growth inhibitory signals
  - Apoptosis resistance
  - Infinite proliferative capacity
  - Angiogenic potential

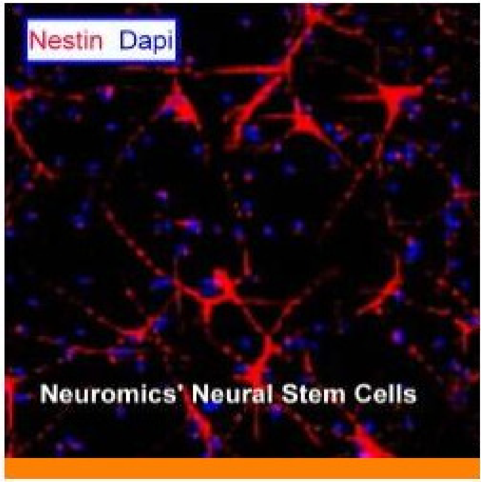
Why is that? Again:

- Function of important genes is backuped by other pathways!

- In order to understand how that works
  (→ systems biology)
  one needs to know
  the status of **<u>many</u>** genes

Medical Research News

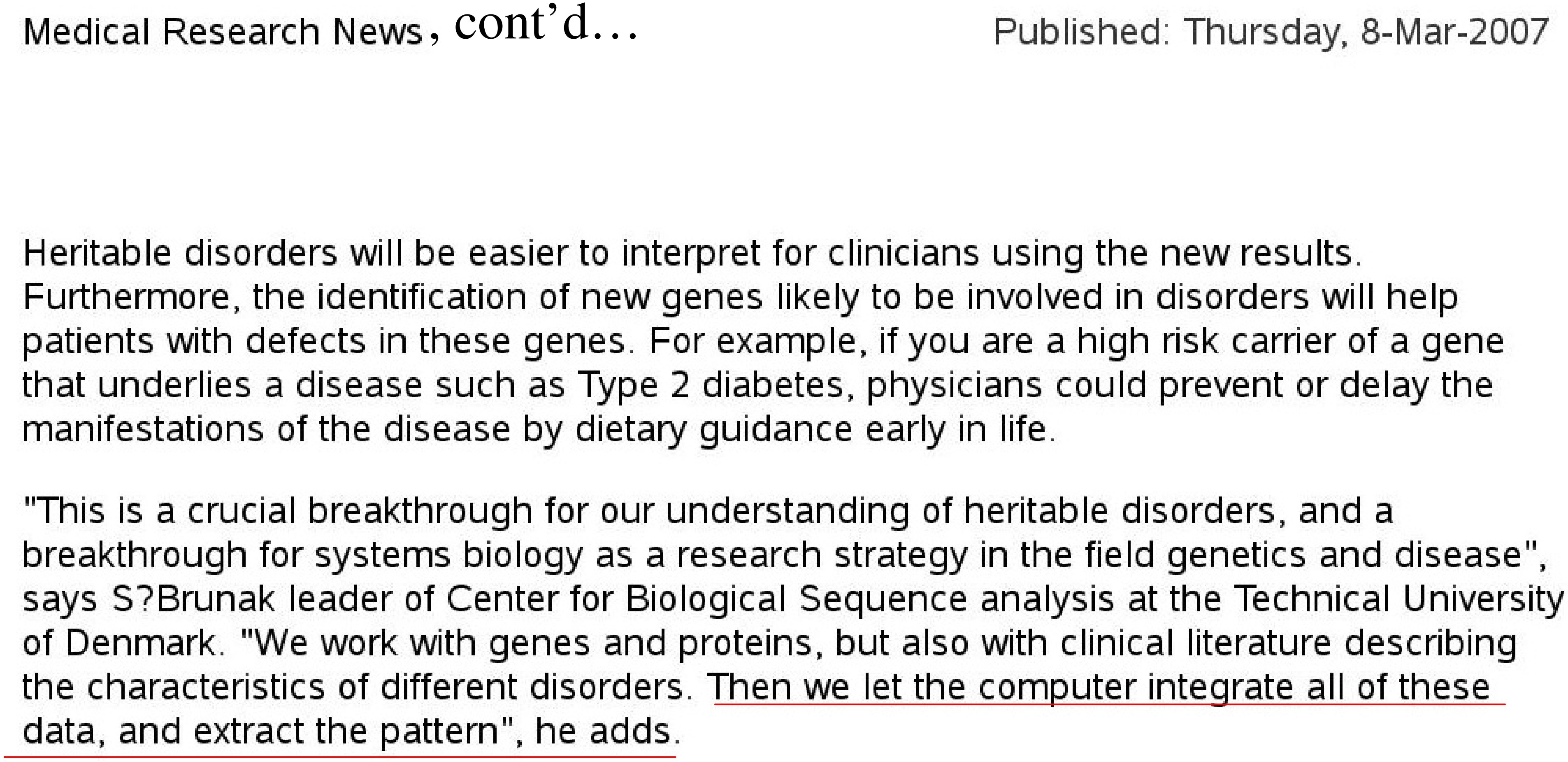
Published: Thursday, 8-Mar-2007



Danish and Belgian researchers have found a computer key that maps genes underlying heritable disorders, such as breast cancer, multiple sclerosis, and Alzheimer's disease. These results will possibly ease the discovery of new medicines and improve treatment in various disorders.

The results - which are published in the current issue of Nature Biotechnology - show that genes important for the development of diseases like Alzheimer's follow the same cellular rules as genes involved in fundamentally different disorders, such as heart disorders, multiple sclerosis, breast cancer, and Type 2 diabetes.

"Many disorders manifest themselves in fundamentally different ways, but the new surprising discovery is that the underlying genes play together after the same rules. Our results show that the genes that trigger diseases, regardless of the type of disease in question, are social team players who cooperate according to highly specific rules. These rules have now been mapped, and we have pointed at hundreds of new genes that are likely to be involved in disorders including multiple sclerosis, Parkinson, heart disorders, and diabetes", says Kasper Lage from Technical University of Denmark, who is the project coordinator on this work.



Medical Research News, cont’d…

Published: Thursday, 8-Mar-2007

Heritable disorders will be easier to interpret for clinicians using the new results. Furthermore, the identification of new genes likely to be involved in disorders will help patients with defects in these genes. For example, if you are a high risk carrier of a gene that underlies a disease such as Type 2 diabetes, physicians could prevent or delay the manifestations of the disease by dietary guidance early in life.

"This is a crucial breakthrough for our understanding of heritable disorders, and a breakthrough for systems biology as a research strategy in the field genetics and disease", says S?Brunak leader of Center for Biological Sequence analysis at the Technical University of Denmark. "We work with genes and proteins, but also with clinical literature describing the characteristics of different disorders. Then we let the computer integrate all of these data, and extract the pattern", he adds.

What HT data?

What is measured?

How?

- Genomics (DNA, → genome)

  (e.g. by sequencers, microarrays, …)
- Epigenetics (→ e.g. methylome)

  (e.g. by microarrays)
- Transcriptomics (→ transcriptome)

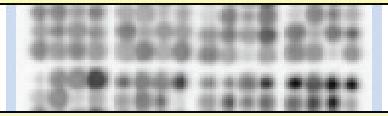
  (e.g. deep sequencing, microarrays, …)
- Proteomics (→ proteome)

  (e.g. 2D-gel electrophoresis, mass spectrometry, guess what – microarrays (AB or peptide chips)

First microarrays:

- cDNA
- on a nylon membrane
- prepared RNA reversely transcribed into cDNA (like today) ...
- ... using radioactively labelled nucleotides (today: mostly fluorescence labelling)

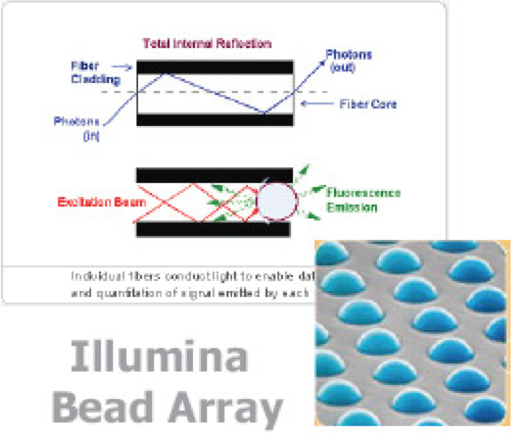
SAGE



1975: Southern Blotting Technology (Edward Southern)

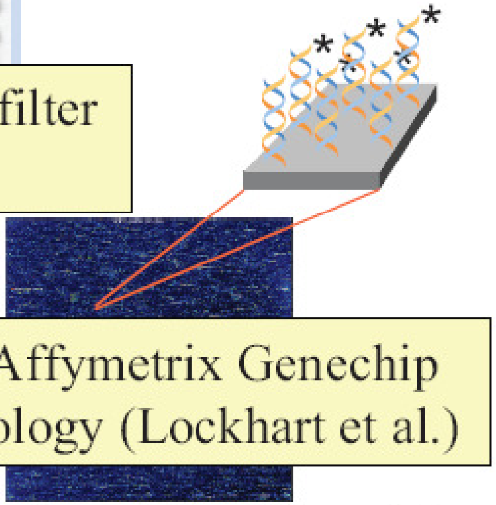
2003: Illumina Bead Arrays

1991: First high-density Nylon filter Arrays (Lennon, Lehrach)


Total Internal Reflection
Fiber Cladding
Photons (out)
Fiber Core
Photons (in)
Excitation Beam
Fluorescence Emission
Individual fibers conduct light to enable dat
and quantitation of signal emitted by each

Illumina Bead Array

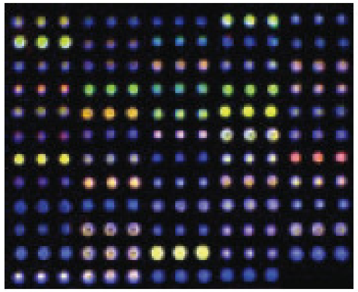
Different Technolo for Measuring Ge Expression



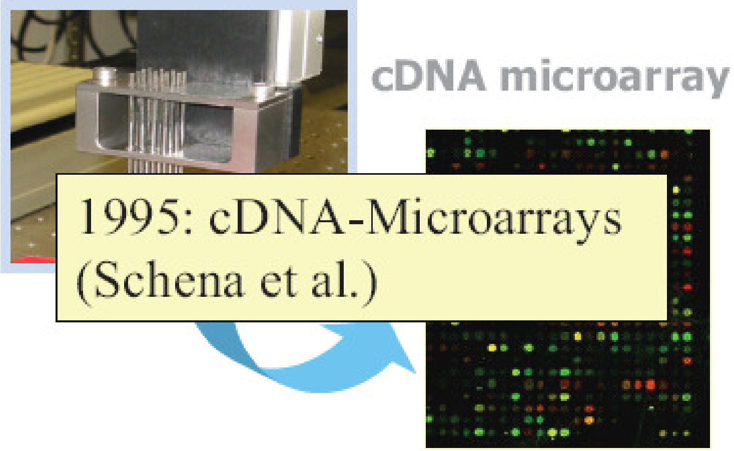
1996: Affymetrix Genechip Technology (Lockhart et al.)

GeneChip Affymetrix

Agilent: Long oligo Ink Jet



CGH

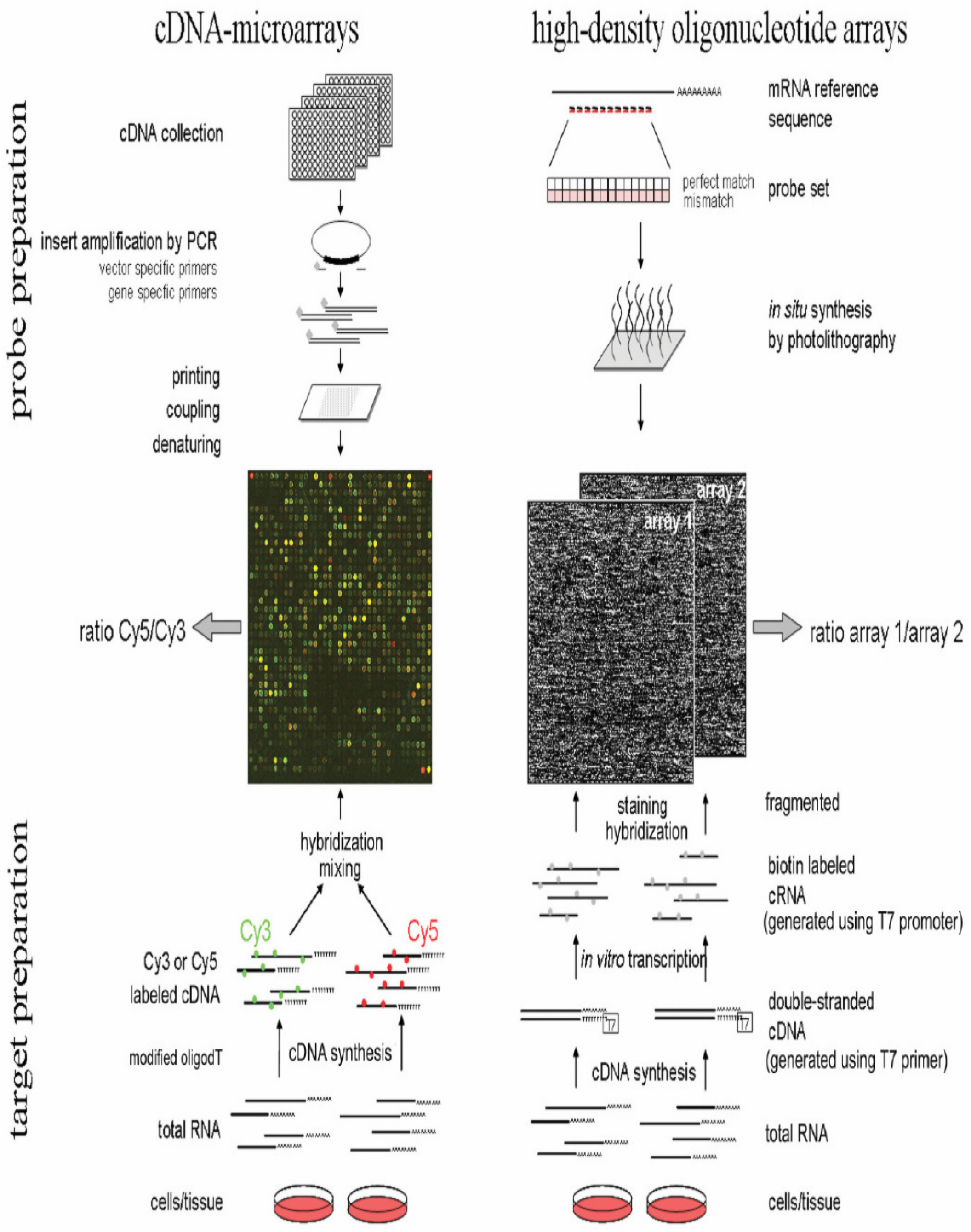


cDNA microarray

1995: cDNA-Microarrays (Schena et al.)

two-channel

single channel

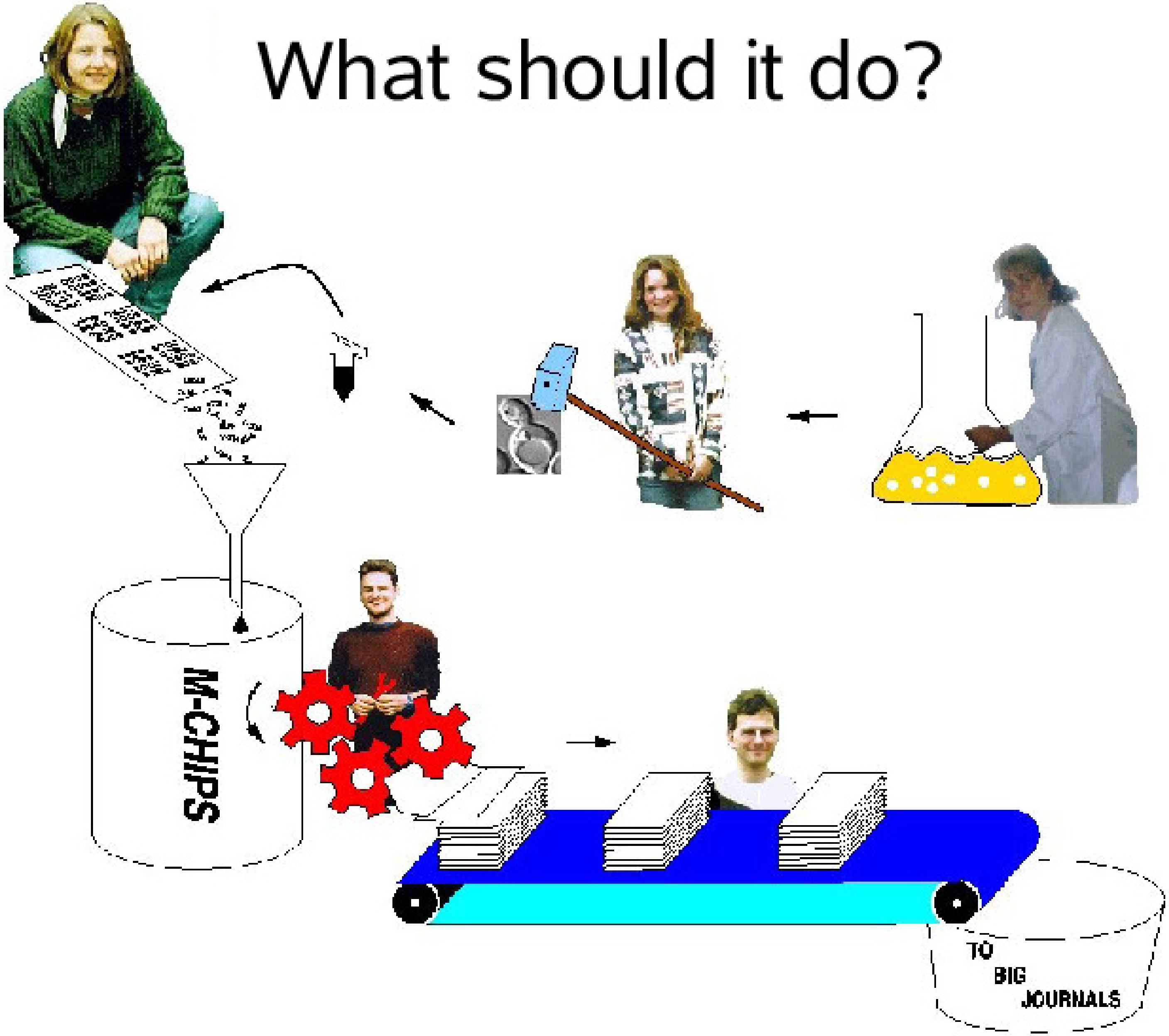


Microarray Experiment

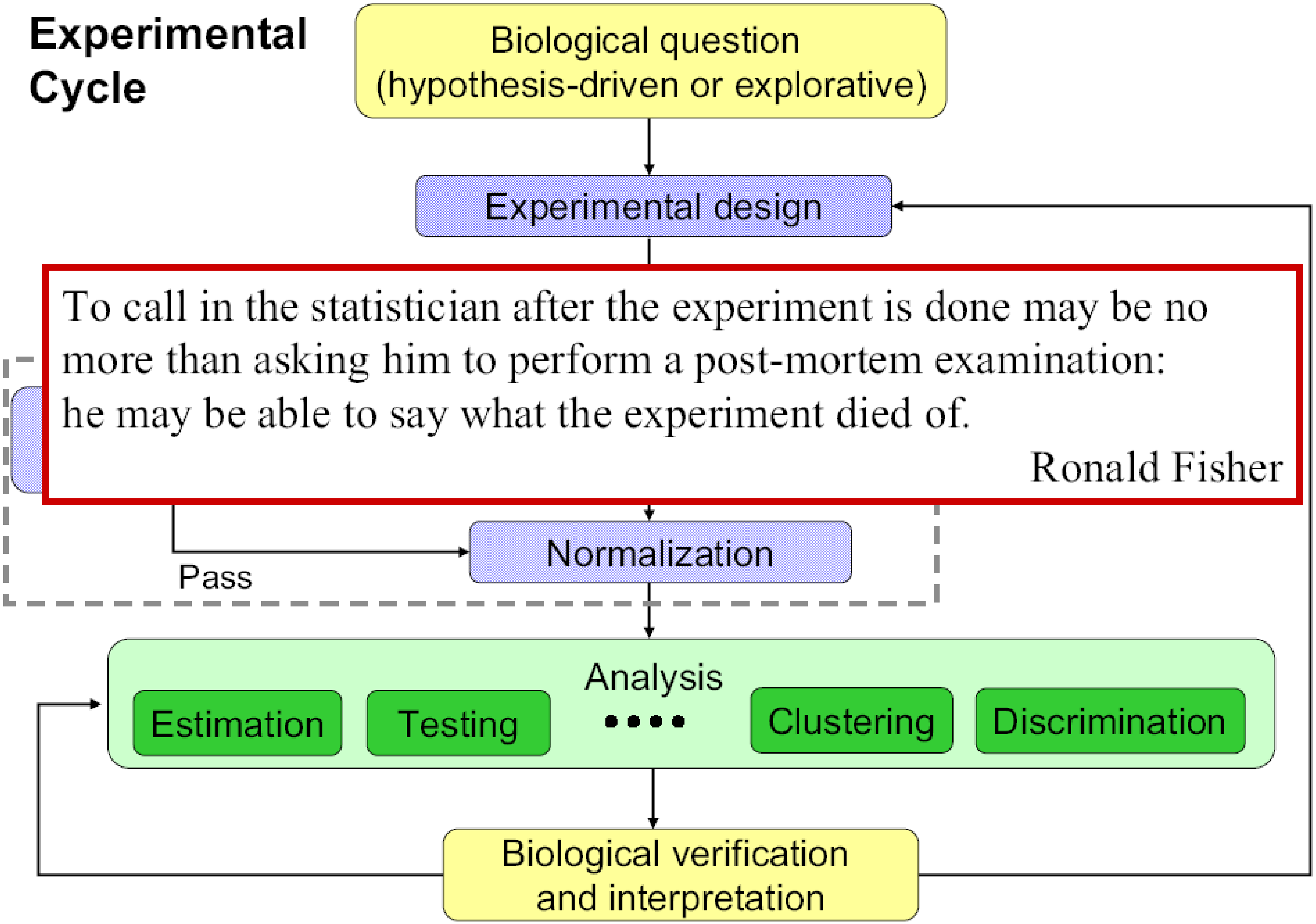
Animation:

http://www.bio.davidson.edu/Courses/genomics/chip/chip.html

What should it do?



Experimental Cycle



Biological question
(hypothesis-driven or explorative)

Experimental design

Pass

Normalization

Analysis

Estimation

Testing

••••

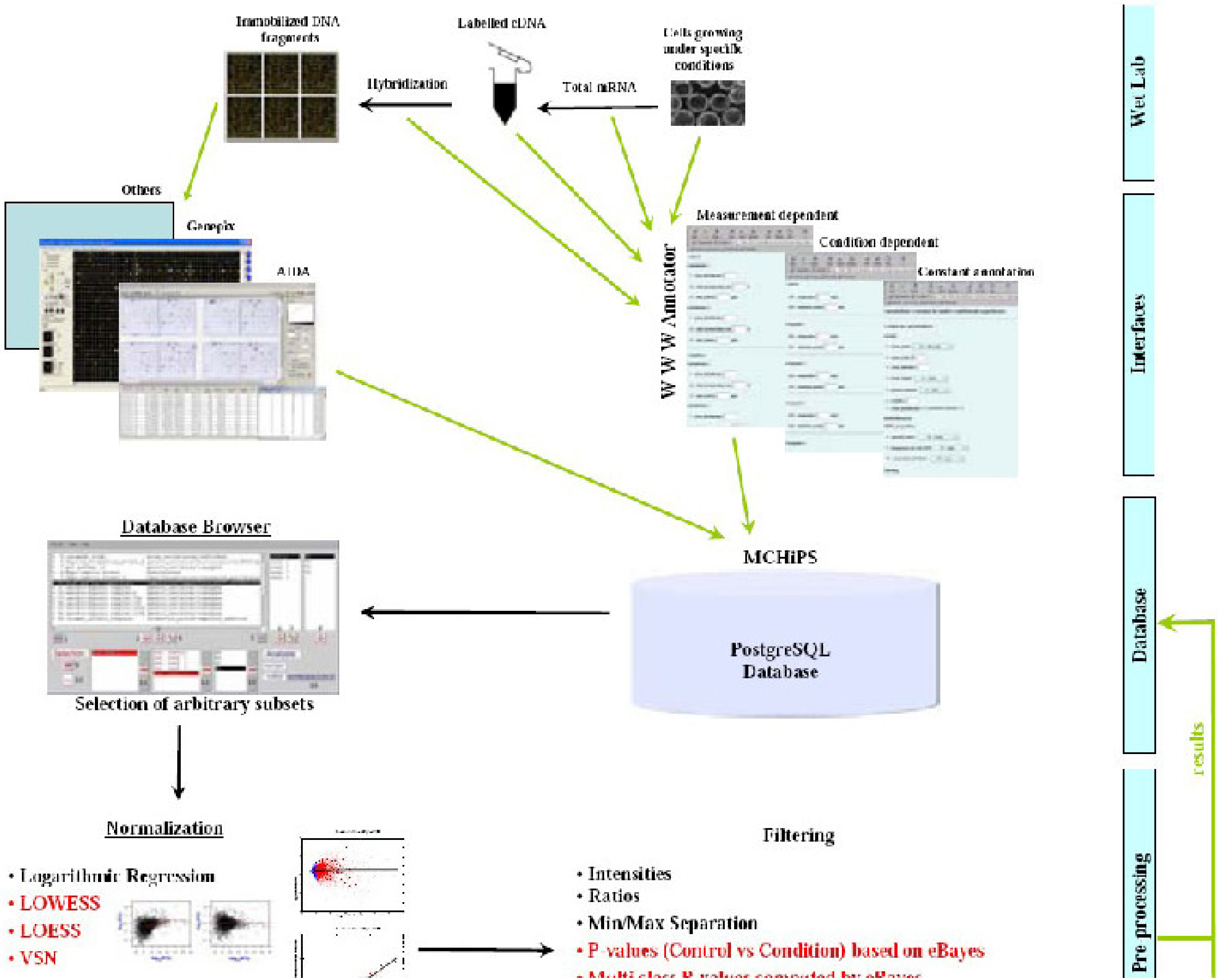
Clustering

Discrimination

Biological verification
and interpretation

> To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.
>
> Ronald Fisher

What is the biggest challenge?



Wet Lab

Interfaces

Database

Pre processing

results

Immobilized DNA fragments

Labelled cDNA

Cells growing under specific conditions

Hybridization

Total mRNA

Others

Genepix

AIDA

W W W Annotator

Measurement dependent

Condition dependent

Constant annotation

MCHiPS

PostgreSQL Database

Database Browser

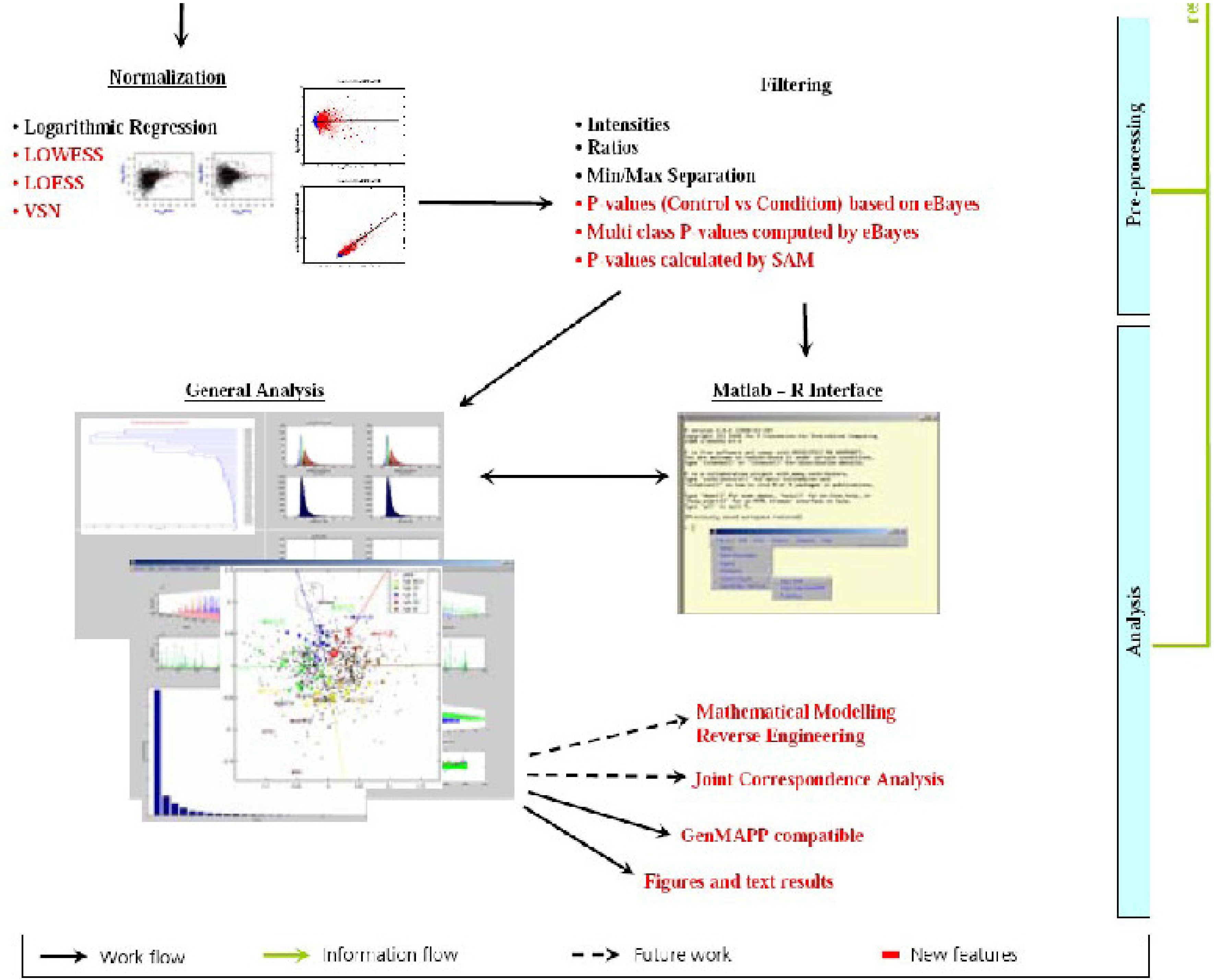
Selection of arbitrary subsets

Normalization

- Logarithmic Regression
- LOWESS
- LOESS
- VSN

Filtering

- Intensities
- Ratios
- Min/Max Separation
- P-values (Control vs Condition) based on eBayes
- Multi class P-values computed by eBayes



Normalization

- Logarithmic Regression
- LOWESS
- LOESS
- VSN

Filtering

- Intensities
- Ratios
- Min/Max Separation
- P-values (Control vs Condition) based on eBayes
- Multi class P-values computed by eBayes
- P-values calculated by SAM

General Analysis

Matlab – R Interface

Mathematical Modelling
Reverse Engineering

Joint Correspondence Analysis

GenMAPP compatible

Figures and text results

Pre-processing

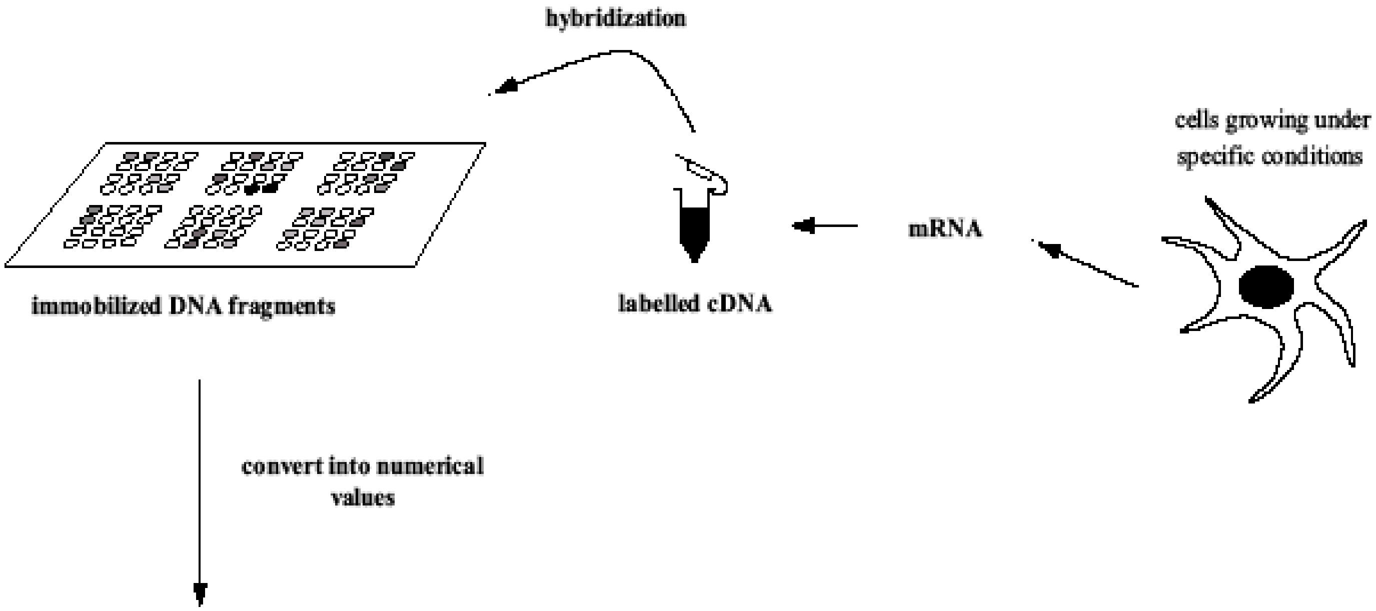
Analysis

Work flow

Information flow

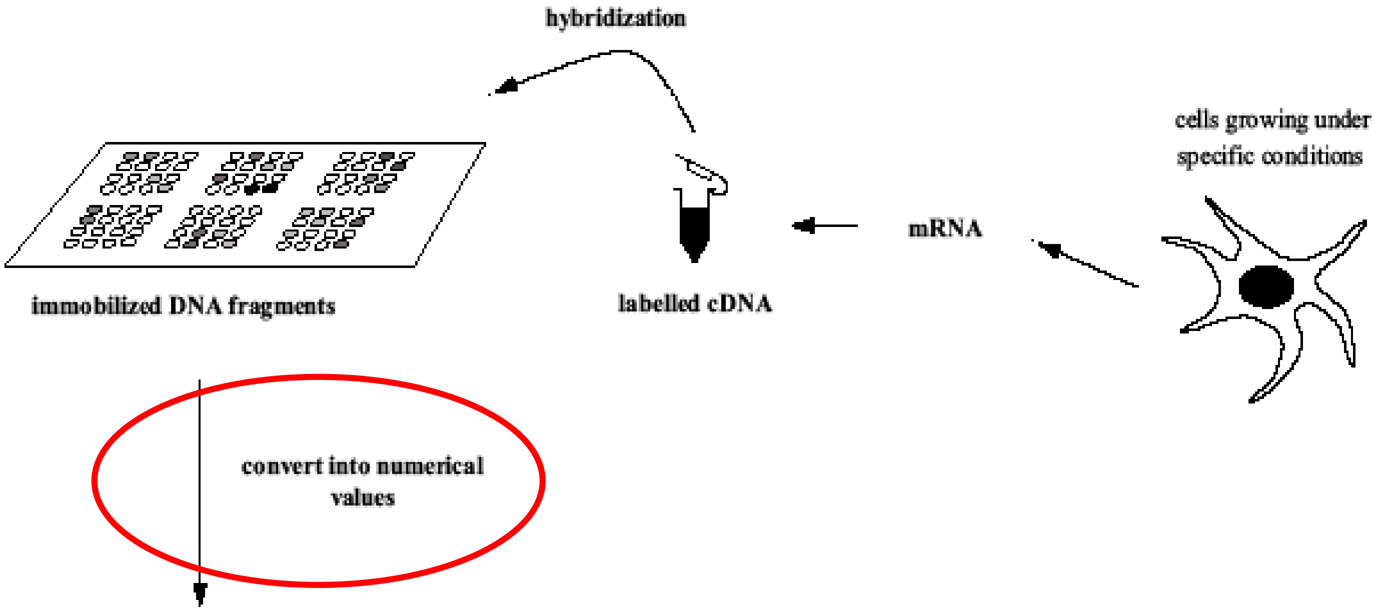
Future work

🟥 New features


hybridization
cells growing under specific conditions
mRNA
labelled cDNA
immobilized DNA fragments
convert into numerical values

Intensity Table

	control condition			condition 1			condition 2		
	hybr. 1	hybr. 2	hybr. 3	hybr. 4	hybr. 5	hybr. 6	hybr. 7	hybr. 8	hybr. 9
gene 1	14,243	12,154							
gene 2	5,323	27,152							
gene 3	10,300	1,407	...						
gene 4	1,007	3,101							
gene 5	100,232	120,993							
gene 6									
gene 7	⋮								
gene 8									
gene 9									
⋮									



hybridization

cells growing under specific conditions

immobilized DNA fragments

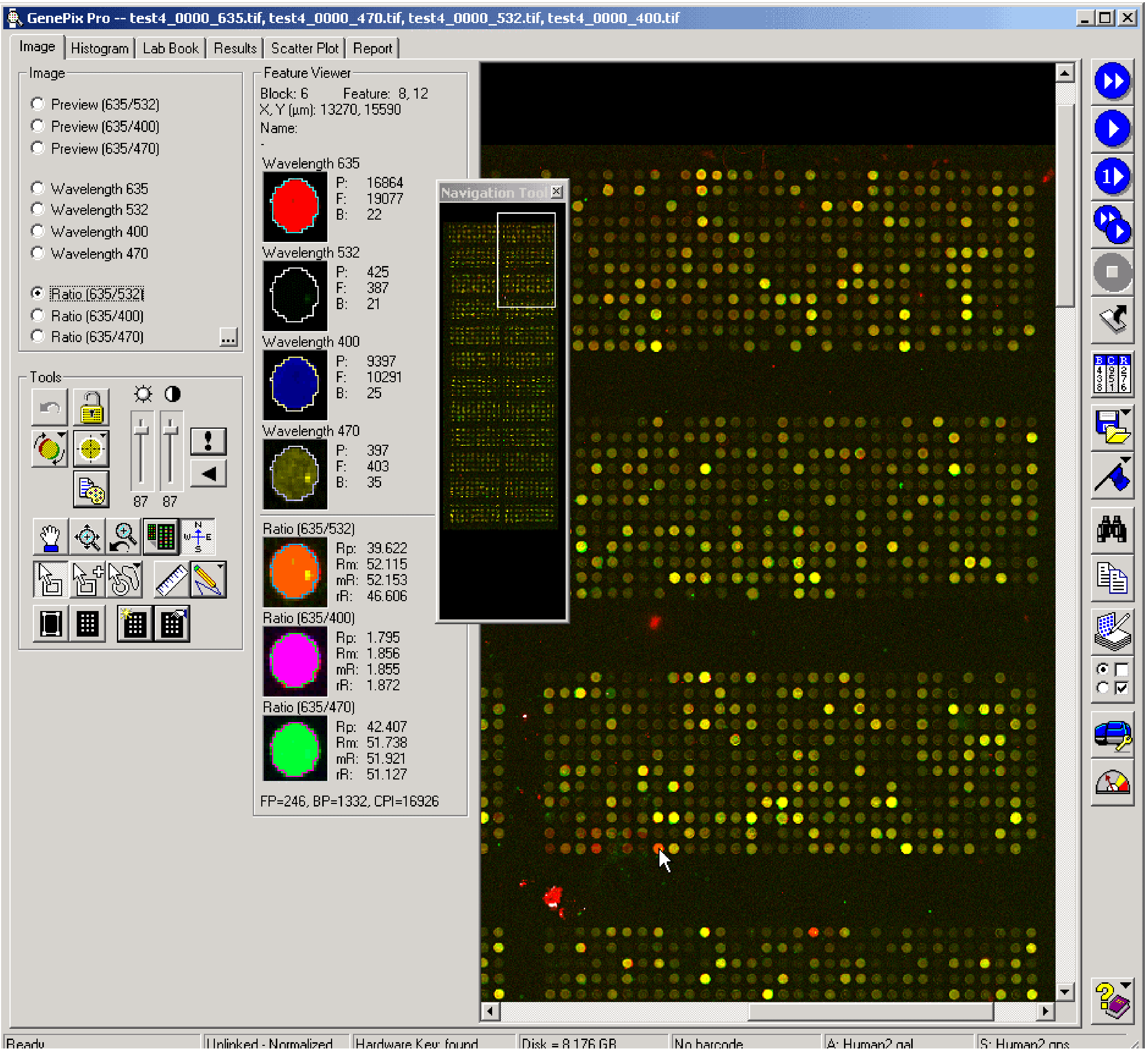
labelled cDNA

mRNA

convert into numerical values

Intensity Table

	control condition			condition 1			condition 2		
	hybr. 1	hybr. 2	hybr. 3	hybr. 4	hybr. 5	hybr. 6	hybr. 7	hybr. 8	hybr. 9
gene 1	14,243	12,154							
gene 2	5,323	27,152							
gene 3	10,300	1,407	...						
gene 4	1,007	3,101							
gene 5	100,232	120,993							
gene 6									
gene 7	⋮								
gene 8									
gene 9									
⋮									



GenePix Pro -- test4_0000_635.tif, test4_0000_470.tif, test4_0000_532.tif, test4_0000_400.tif
Image
Histogram
Lab Book
Results
Scatter Plot
Report
Tools
87
87
Navigation Tool
B C R

Image

○ Preview (635/532)
○ Preview (635/400)
○ Preview (635/470)

○ Wavelength 635
○ Wavelength 532
○ Wavelength 400
○ Wavelength 470

● Ratio (635/532)
○ Ratio (635/400)
○ Ratio (635/470)

Feature Viewer

Block: 6 Feature: 8, 12
X, Y (µm): 13270, 15590
Name:
-

Wavelength 635
P: 16864
F: 19077
B: 22

Wavelength 532
P: 425
F: 387
B: 21

Wavelength 400
P: 9397
F: 10291
B: 25

Wavelength 470
P: 397
F: 403
B: 35

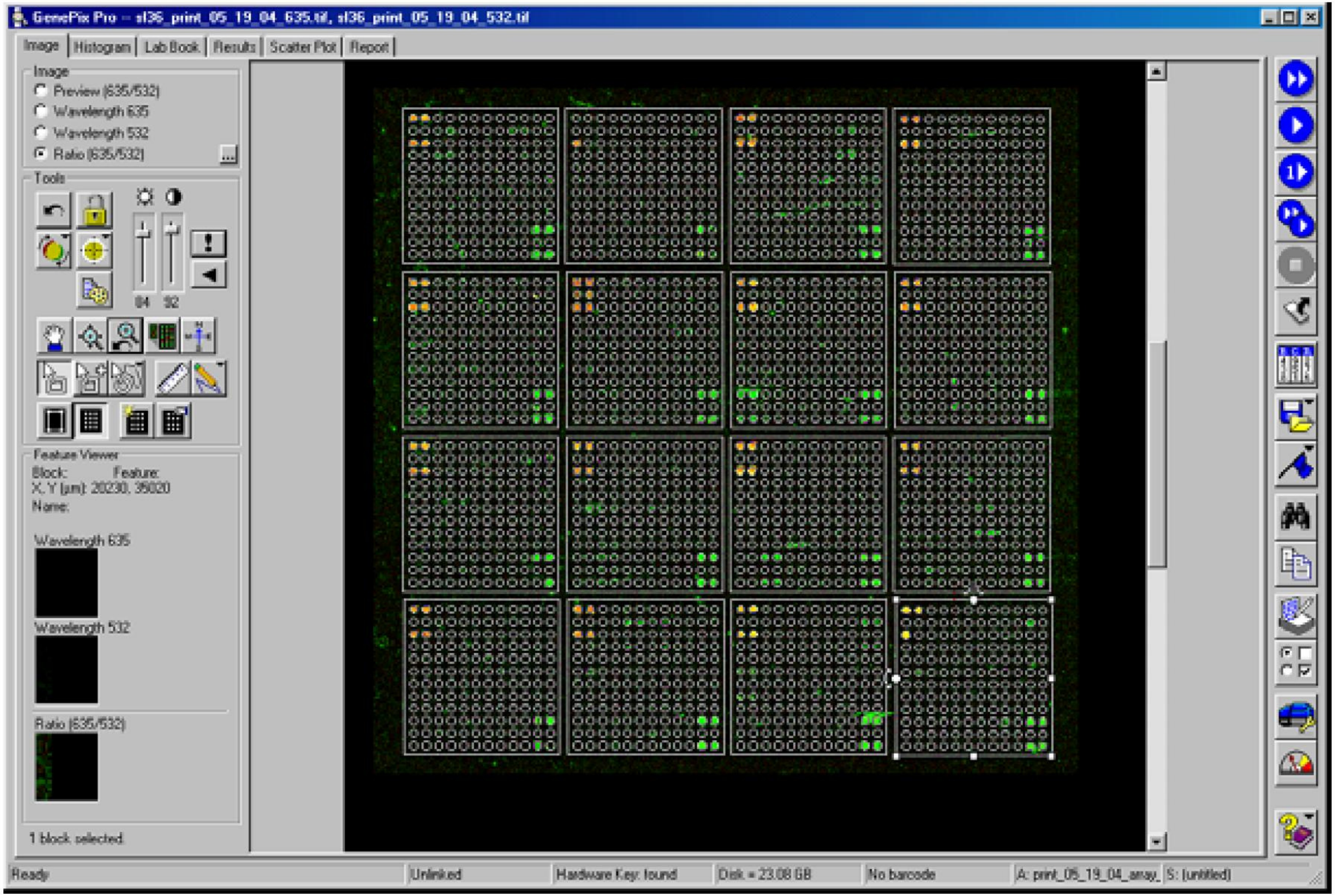
Ratio (635/532)
Rp: 39.622
Rm: 52.115
mR: 52.153
rR: 46.606

Ratio (635/400)
Rp: 1.795
Rm: 1.856
mR: 1.855
rR: 1.872

Ratio (635/470)
Rp: 42.407
Rm: 51.738
mR: 51.921
rR: 51.127

FP=246, BP=1332, CPI=16926

Ready | Unlinked - Normalized | Hardware Key: found | Disk = 8.176 GB | No barcode | A: Human2.gal | S: Human2.gps



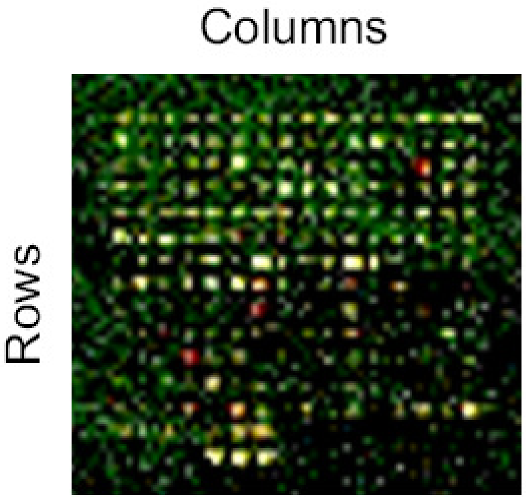
Block: Feature:
X, Y (µm): 20230, 36020
Name:

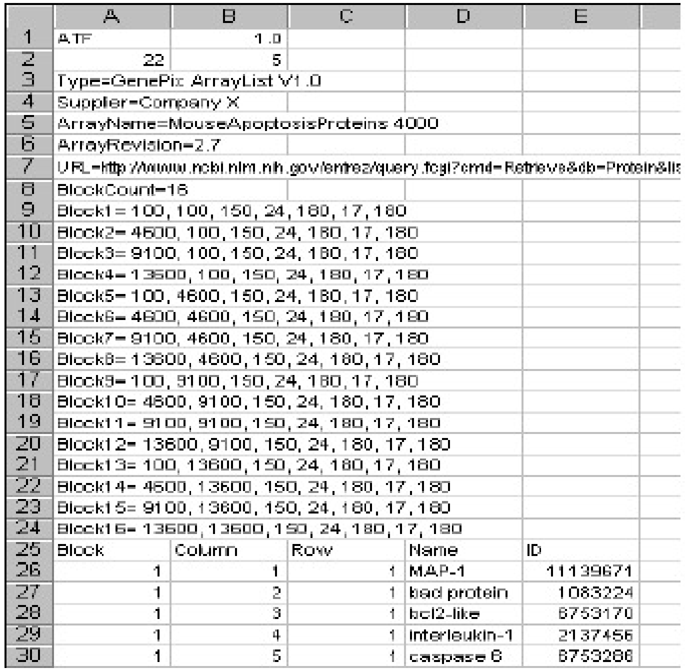
1 block selected

Ready | Unlinked | Hardware Key: found | Disk = 23.08 GB | No barcode | A: print_05_19_04_array_ | S: (untitled)

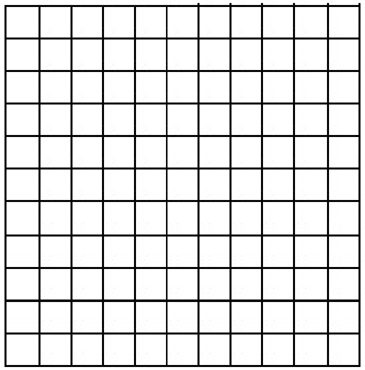
Image Analysis – Spot Identification

- The grid structure is provided by the manufacturer or generated individually for custom-made microarrays (e.g. GAL-files)
- The grid is overlaid by hand or automatically onto the image (beware of column/row displacement errors!)

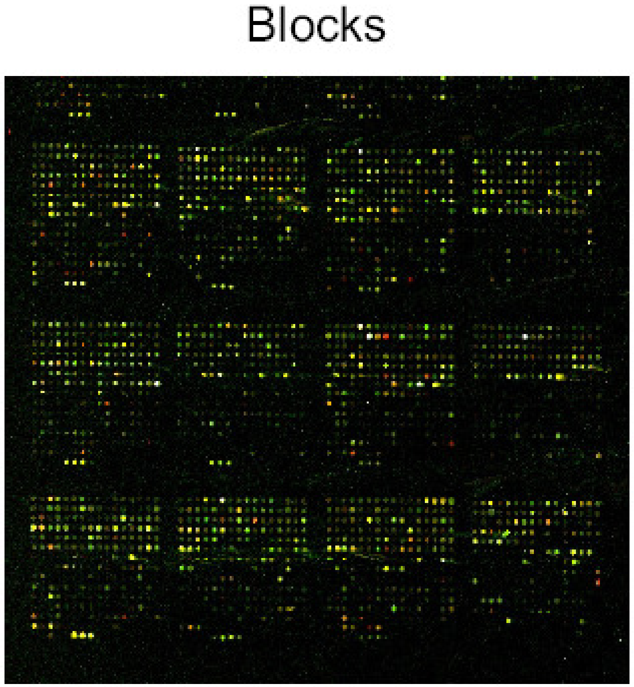




	A	B	C	D	E
1	ATF	1.0			
2	22	5			
3	Type=GenePix ArrayList V1.0				
4	Supplier=Company X				
5	ArrayName=MouseApoptosisProteins 4000				
6	ArrayRevision=2.7				
7	URL=http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Protein&lis				
8	BlockCount=16				
9	Block1= 100, 100, 150, 24, 180, 17, 180				
10	Block2= 4600, 100, 150, 24, 180, 17, 180				
11	Block3= 9100, 100, 150, 24, 180, 17, 180				
12	Block4= 13600, 100, 150, 24, 180, 17, 180				
13	Block5= 100, 4600, 150, 24, 180, 17, 180				
14	Block6= 4600, 4600, 150, 24, 180, 17, 180				
15	Block7= 9100, 4600, 150, 24, 180, 17, 180				
16	Block8= 13600, 4600, 150, 24, 180, 17, 180				
17	Block9= 100, 9100, 150, 24, 180, 17, 180				
18	Block10= 4600, 9100, 150, 24, 180, 17, 180				
19	Block11= 9100, 9100, 150, 24, 180, 17, 180				
20	Block12= 13600, 9100, 150, 24, 180, 17, 180				
21	Block13= 100, 13600, 150, 24, 180, 17, 180				
22	Block14= 4600, 13600, 150, 24, 180, 17, 180				
23	Block15= 9100, 13600, 150, 24, 180, 17, 180				
24	Block16= 13600, 13600, 150, 24, 180, 17, 180				
25	Block	Column	Row	Name	ID
26	1	1	1	MAP-1	11139671
27	1	2	1	bad protein	1083224
28	1	3	1	bcl2-like	6753170
29	1	4	1	interleukin-1	2137456
30	1	5	1	caspase 6	6753286



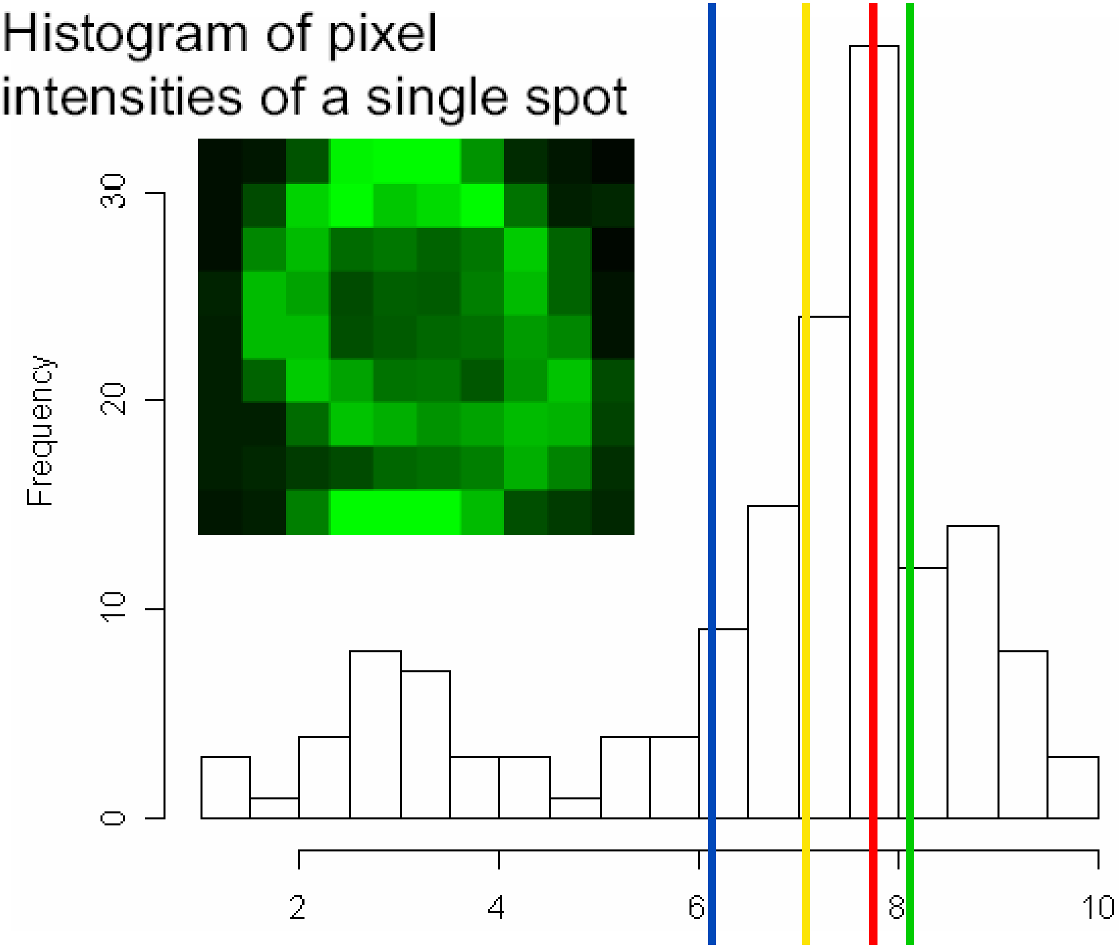
GAL-file contains Clone-IDs and defines their position on the grid

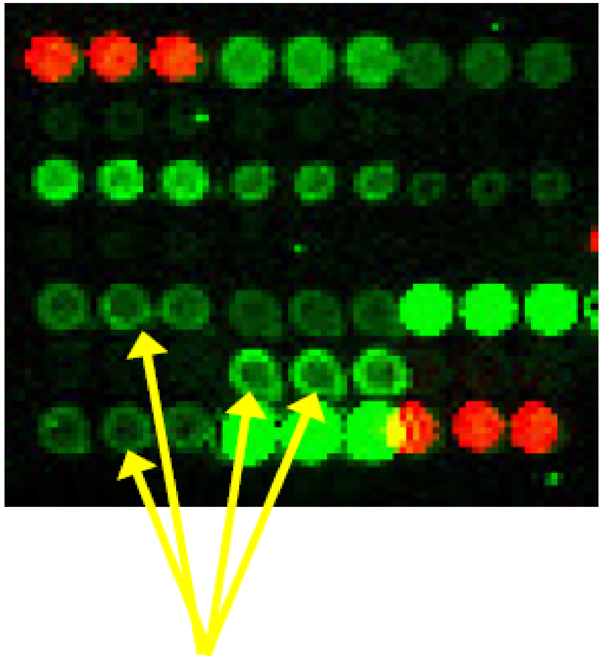


Spot identification

- The signal of the spots is quantified.

Histogram of pixel intensities of a single spot





„Donuts“

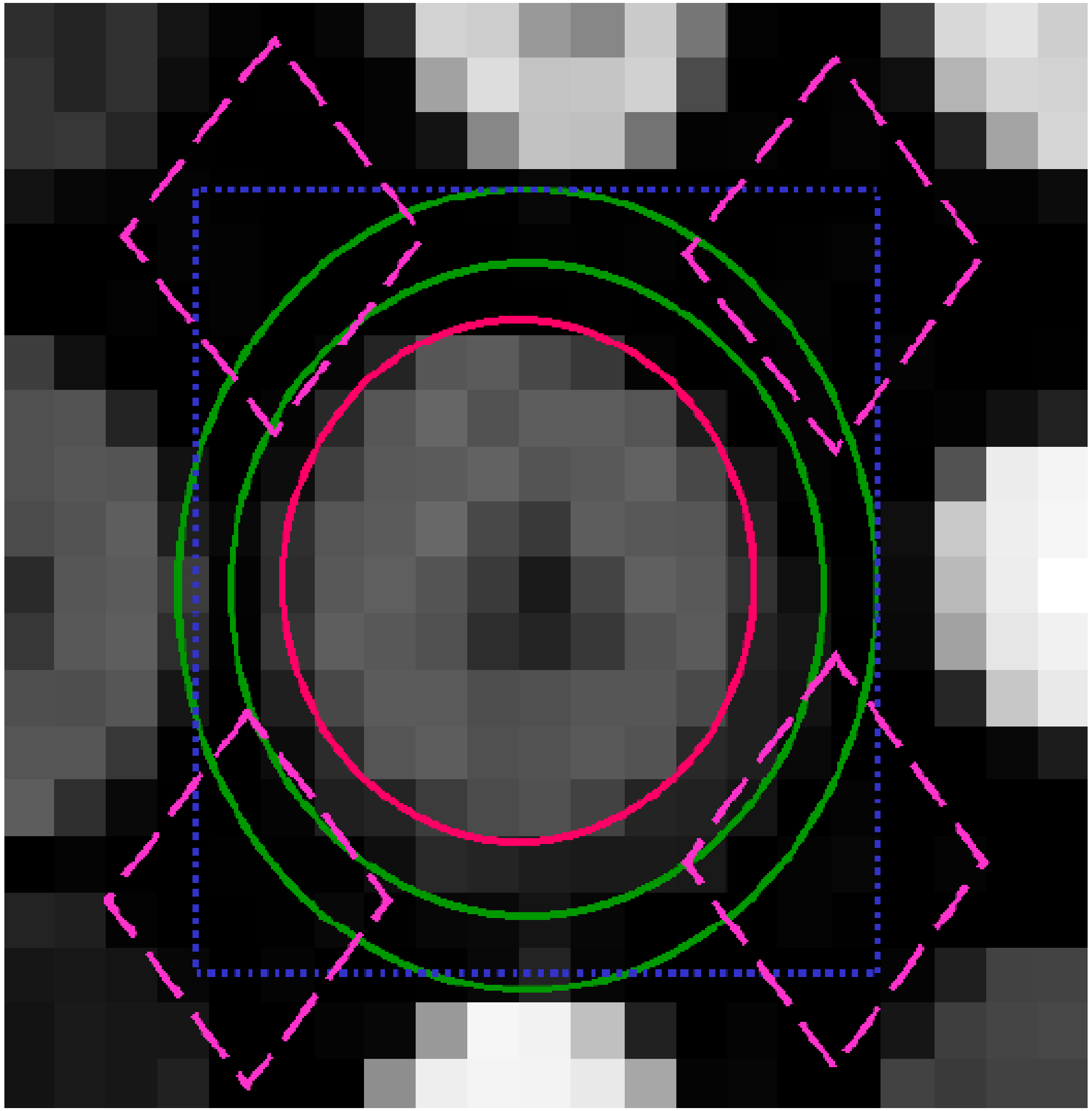
Mean / Median / Mode / 75% quantile

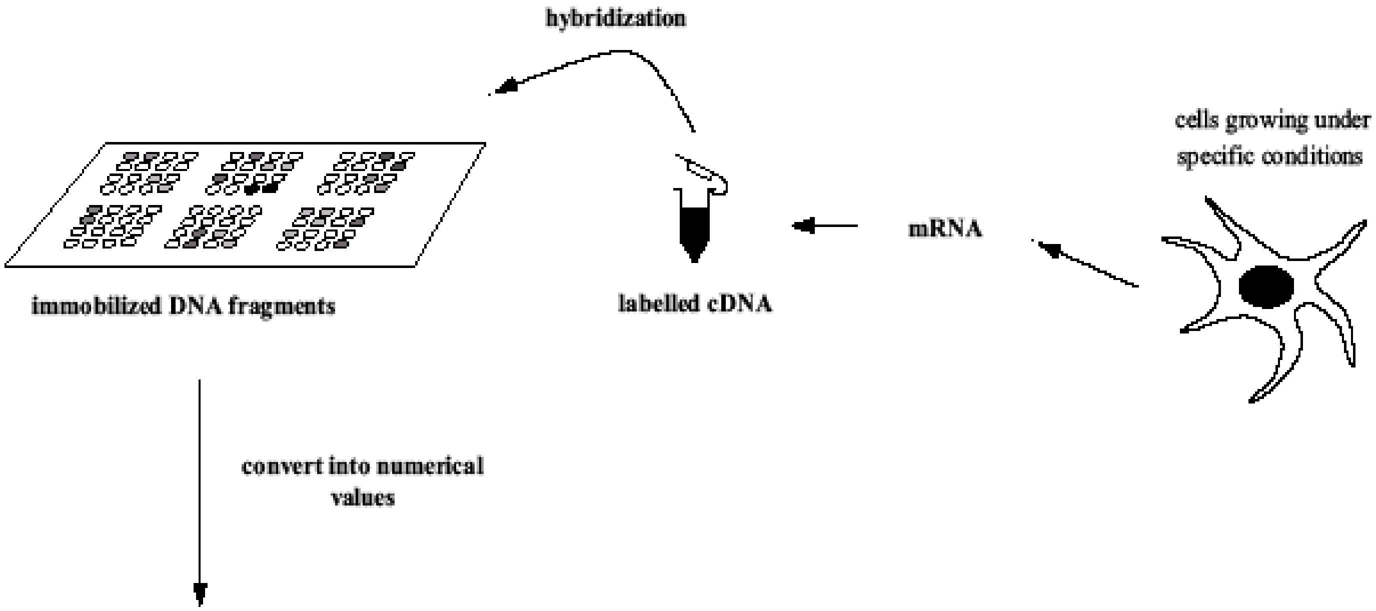
Different Regions around the spot are quantified to measure local background.

GenePix

QuantArray

ScanAlyse




hybridization
cells growing under specific conditions
mRNA
labelled cDNA
immobilized DNA fragments
convert into numerical values

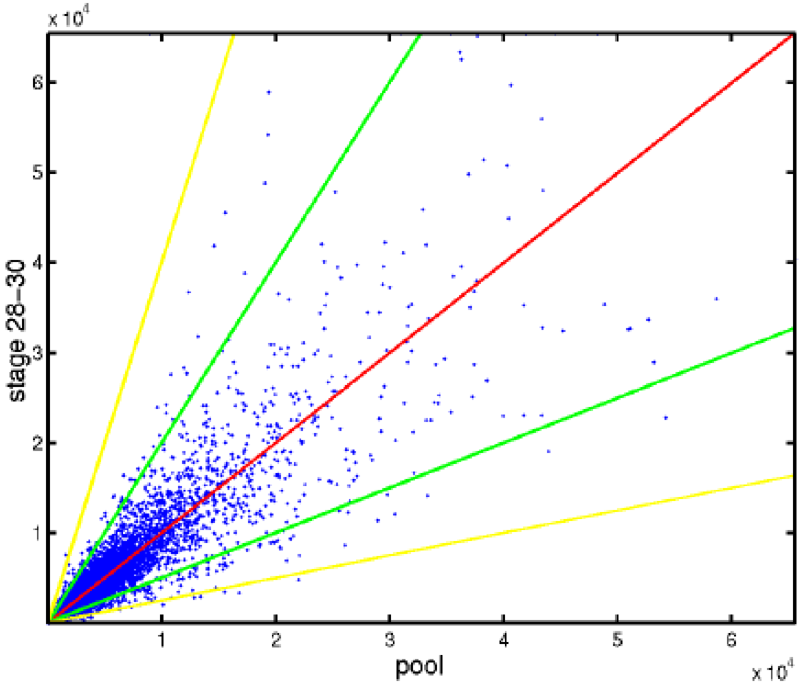
Intensity Table

	control condition			condition 1			condition 2		
	hybr. 1	hybr. 2	hybr. 3	hybr. 4	hybr. 5	hybr. 6	hybr. 7	hybr. 8	hybr. 9
gene 1	14,243	12,154							
gene 2	5,323	27,152							
gene 3	10,300	1,407	...						
gene 4	1,007	3,101							
gene 5	100,232	120,993							
gene 6									
gene 7	⋮								
gene 8									
gene 9									
⋮									

Scatterplot

Data

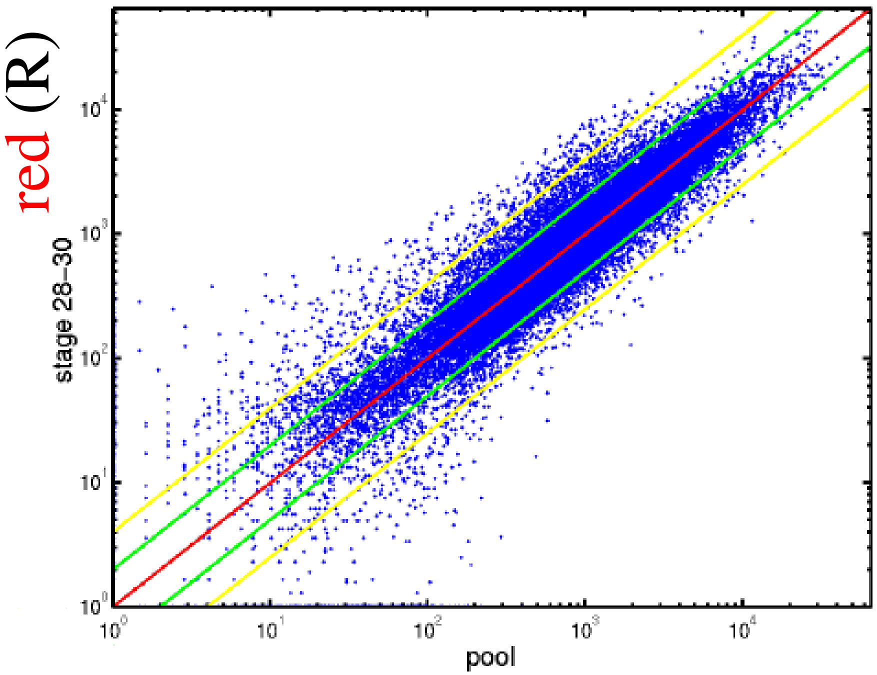
red (R)



green (G)

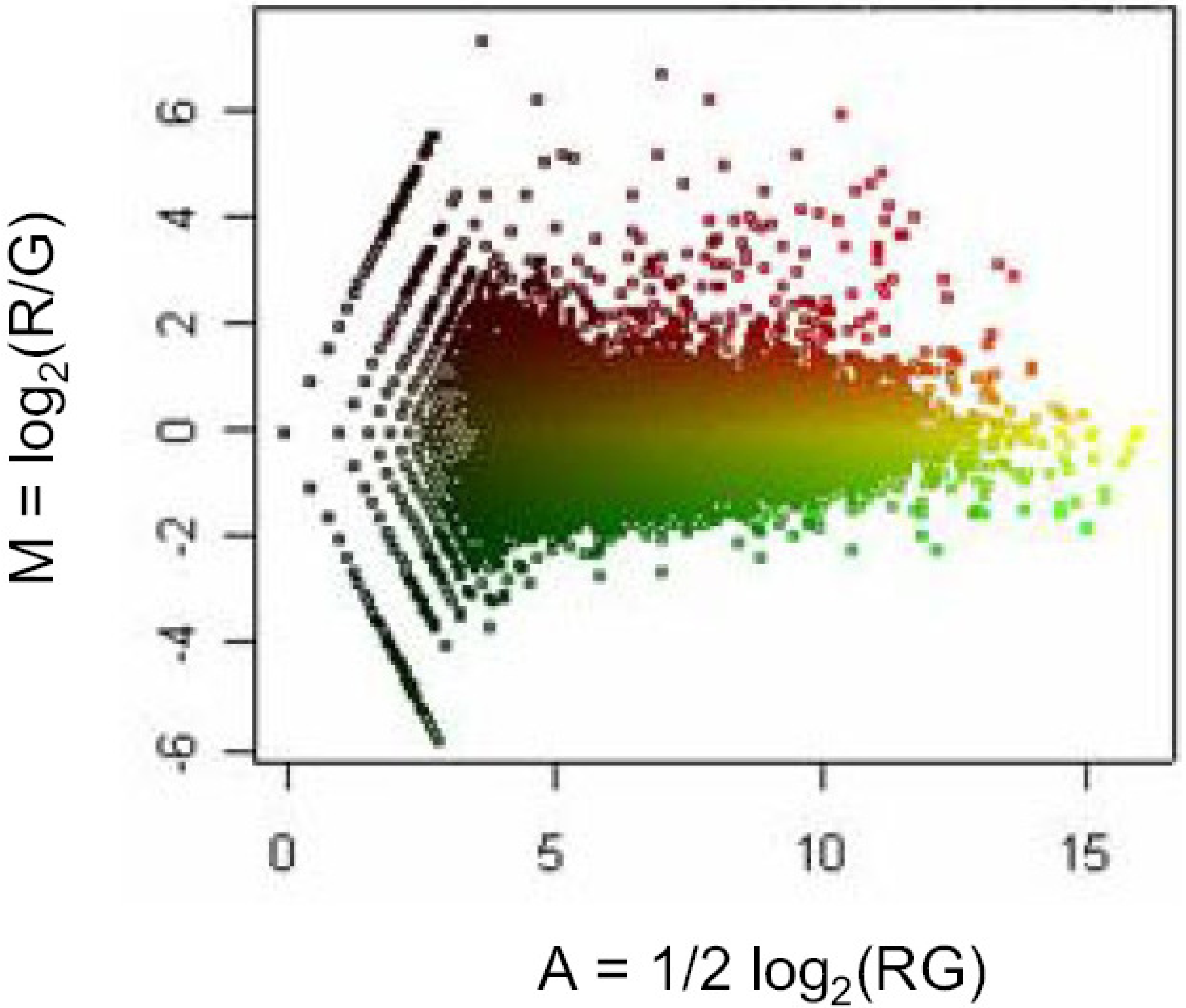
Data (log scale)

red (R)



green (G)

MA Plot



Computing

- Microarray data analysis does not need much processor time (interactive instead of batch processing)

- However, it needs considerable amounts of memory (RAM)

Computing, cont.

- Imaging or scatterplots

  comprise one hybridization at a time

  → often done on PCs

  → mostly running Windows

ADVENT
lore

Computing, cont.

- High level analyses

  (classification, clustering, projection, modelling, ...)

  may comprise hundreds of hybridizations

  → often done on Servers

  → mostly running Unix

lima
paris
theran
rome
PRIMERGY
FUJITSU SIEMENS
RX300 S4

fez
osaka
lima
paris
theron
SWN1-GQ8
10.40.128.1
SWN2-GQ8
10.40.128.2
ProCurve Networking