DNA Chip Technology

Benedikt Brors

Dept. Intelligent Bioinformatics Systems

German Cancer Research Center

Why DNA Chips ?

- Functional genomics: get information about genes that is unavailable from sequence

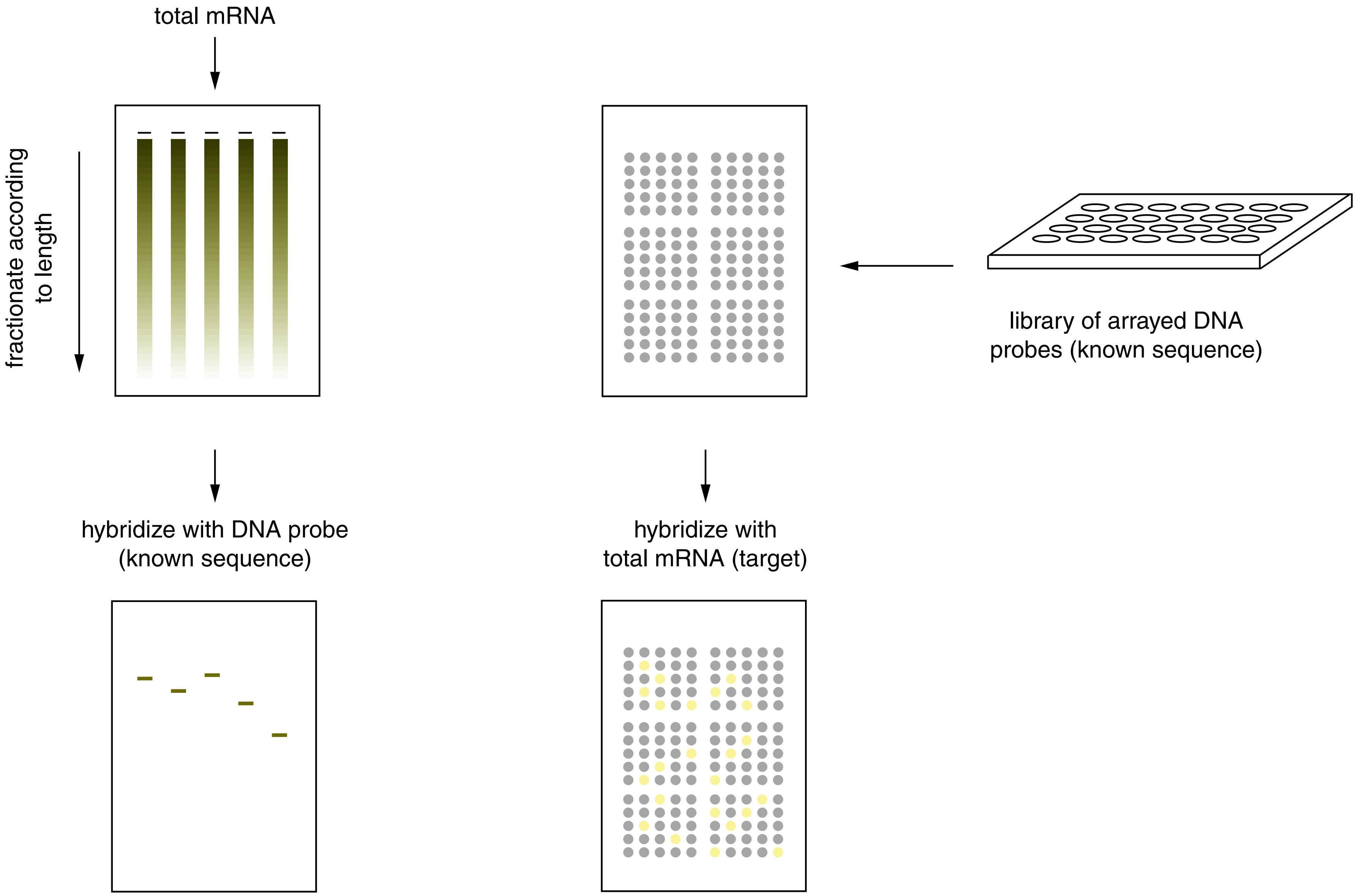
- Understand how cells/organisms react to external stimuli

- Understand gene regulation networks

- Determine what makes the difference between healthy and diseased tissue

- Simply do 15,000 Northern Blots at a time

Comparison Northern blot ↔ DNA array



total mRNA

fractionate according to length

library of arrayed DNA probes (known sequence)

hybridize with DNA probe (known sequence)

hybridize with total mRNA (target)

Northern Blot

DNA array

Functional Genomics

- There may be 100,000 different transcripts in human cells (± 50,000)

- We only have sound information on ≈ 12,000 genes

- All cells have the same genome, but there are more than 200 cell types in a single organism

- Gene expression determines the cell type (neuron, lymphocyte, fibroblast etc.) and directs development of an organism (by spatial/temporal patterns)

- DNA chip technology promises to solve such unanswered questions

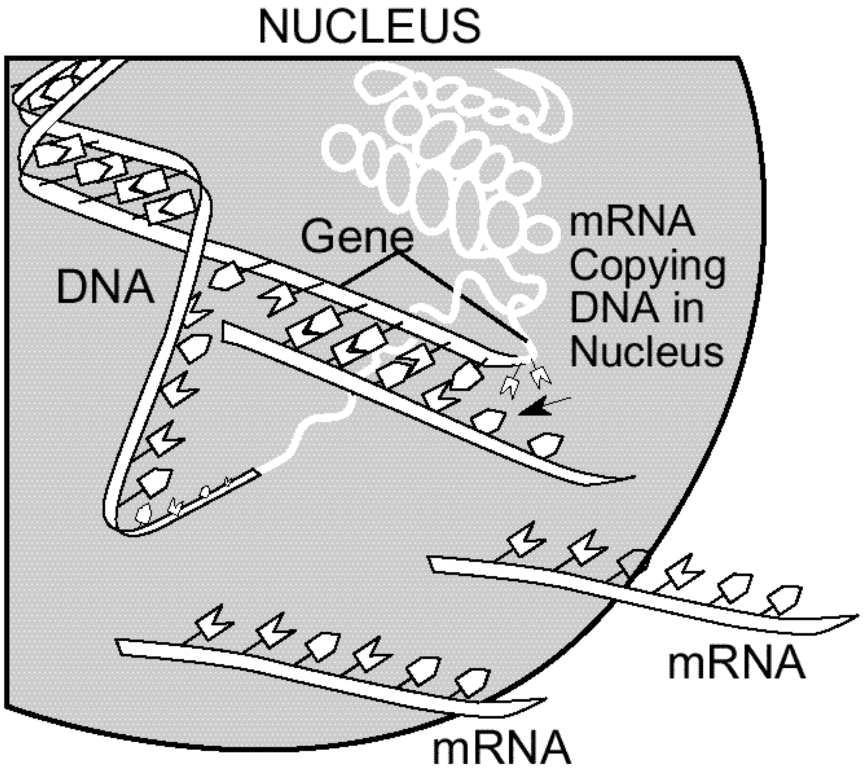
Basic Biology

- Genes contain construction information

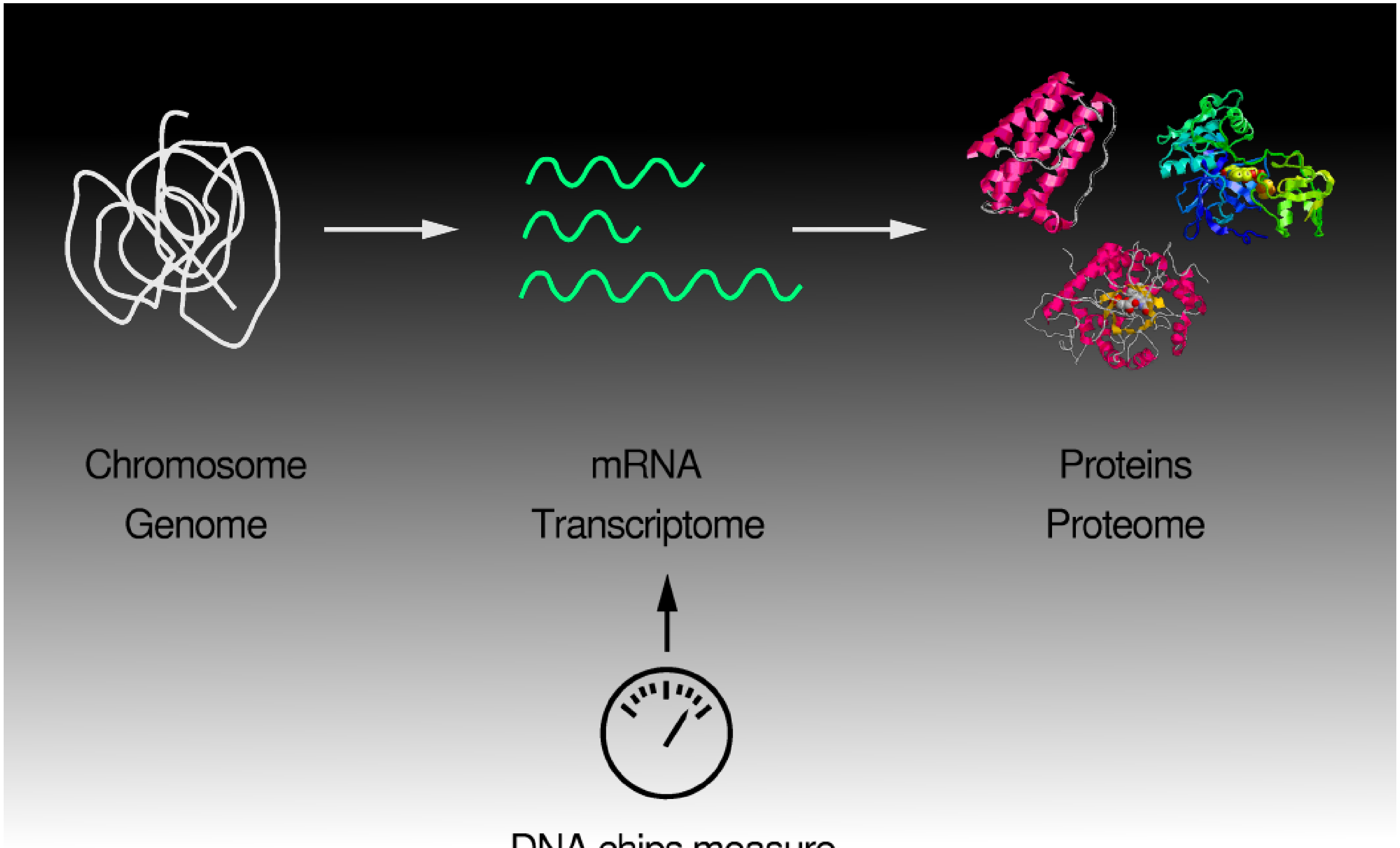
- All structure and function is made up by proteins

- mRNA is sort of ‘working copy’, containing design of one protein

- mRNA is transfered to cytoplasm where protein is made



More Schematically ...



Chromosome

Genome

mRNA

Transcriptome

Proteins

Proteome

DNA chips measure transcript levels

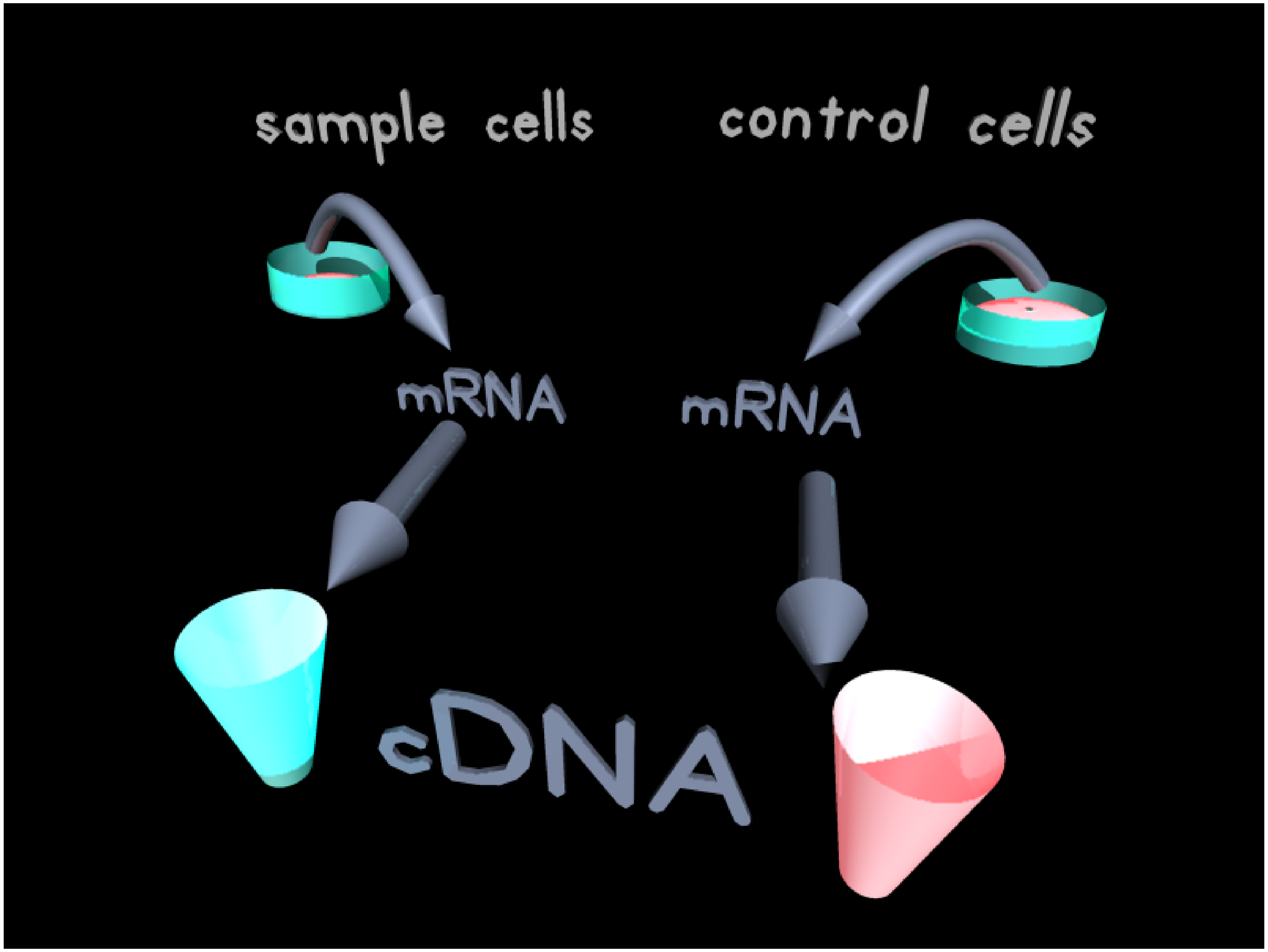
DNA Chip Technology 1

- Array: Small glass slide, contains 100s to 10,000s of DNA fragments (‘spots’) on few cm^2

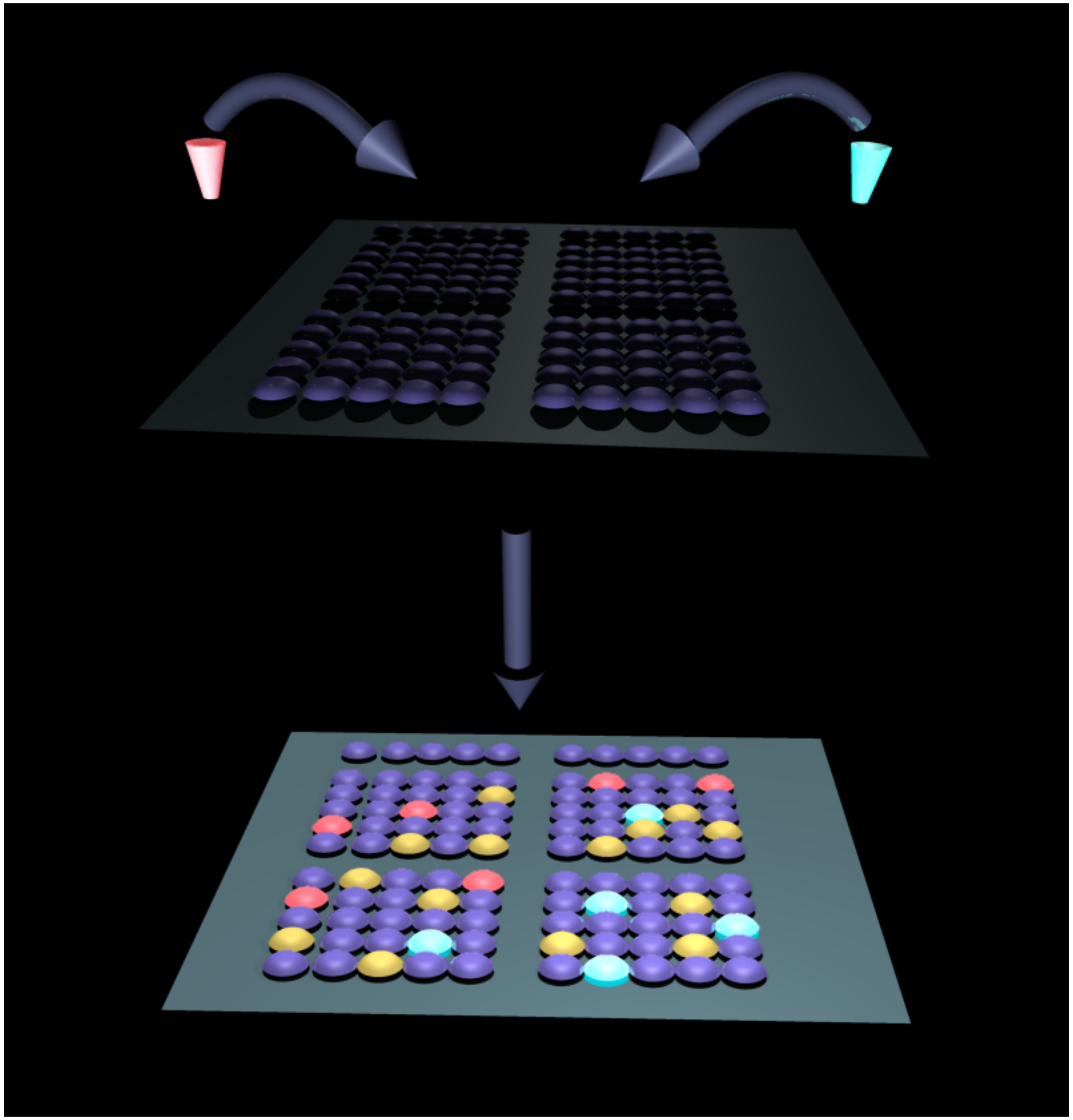
- Each DNA fragment will bind specifically a complementary DNA/RNA: ‘Hybridization’

- ‘Active’ (transcribed) genes can be extracted from cells/tissues, labeled and hybridized to the array ⇒ ‘active’ genes will light up on the array

DNA Chip Technology 2



DNA Chip Technology 3



DNA Chip Technology 4

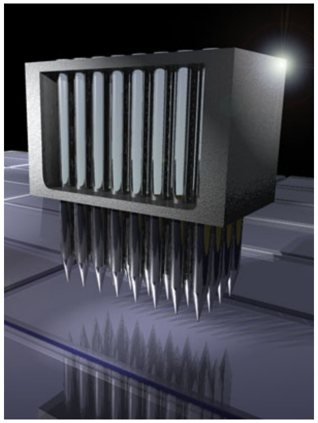
- Chip is read out by video camera
- Digitized image is analyzed by image analysis software
- Result: list of numbers

	R	G
spot1	1,346.2	1,575.8
spot2	100,326.1	30,872.0
spot3	987.1	177.2
spot4	(. . .)	(. . .)
(. . .)		

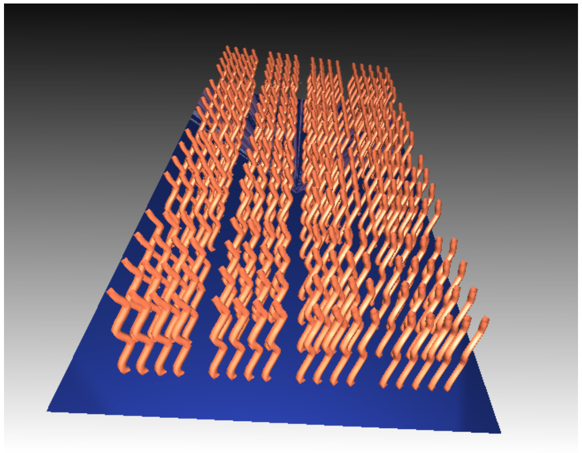
- N.B. the second column, ‘G’, is missing for one-color experiments

Competing Technologies

- Two systems: printed/‘spotted’ chips and on-chip synthesis



(A)



(B)

- For (A) mostly long DNA strands (500–3000 nt)
- For (B) only oligonucleotides ($\leq$ 25 nt)

Printed/Spotted Chips

- Any DNA fragment can be put on the chip. Often, cDNA from libraries is used; to get a suitable amount for spotting, mass PCR has to be used. However, though performed less frequently, synthetic oligonucleotides can be used as well.

- For cDNA chips, a new PCR reaction has to be performed for every batch of chips, these are quite different from each other with respect to the amount of DNA bound in a certain spot.

- DNA fragments are transferred to the chip either by a spotting robot that transfers nanoliter quantities of liquid, or by an ink-jet like device.

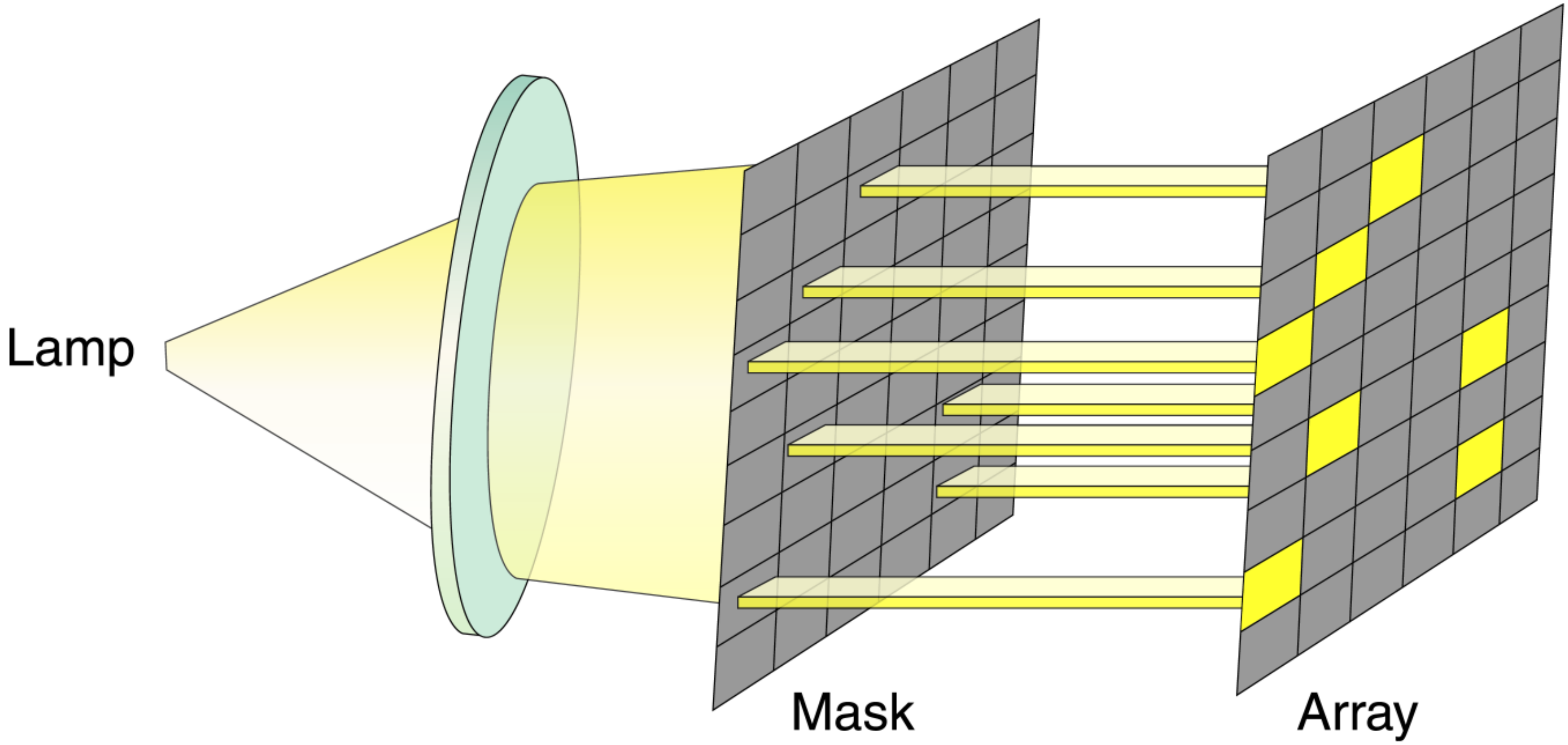
- Spotted or printed chips are usually hybridized with two differently labeled mRNA preparations (i.e. their cDNA representation) By competitively hybridizing with two targets, the DNA amount in a single spot becomes less important (but not irrelevant!!).

On-chip synthesis of oligonucleotides

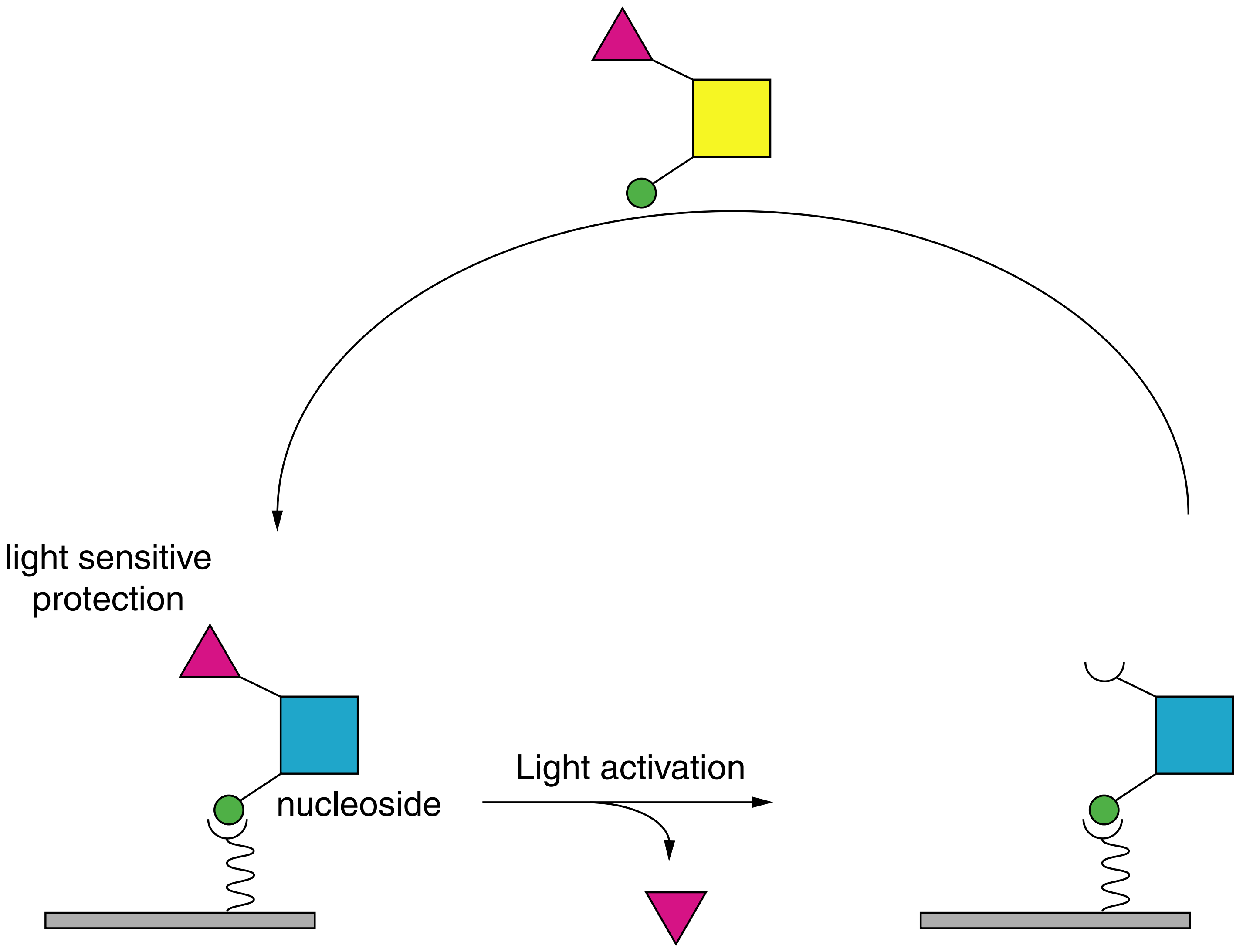
- Oligonucleotides are built up on the chip surface by sequentially elongating the growing chain with a single nucleotide. To determine the sequence of the final oligonucleotides on each position of the chip, a process called *photolithography* is used.

- As chemical yield of the stepwise elongation is limited, oligonucleotides can’t be grown to more than 25 nt length.

Photolithography



Light activated oligo synthesis



On-chip synthesis of oligonucleotides

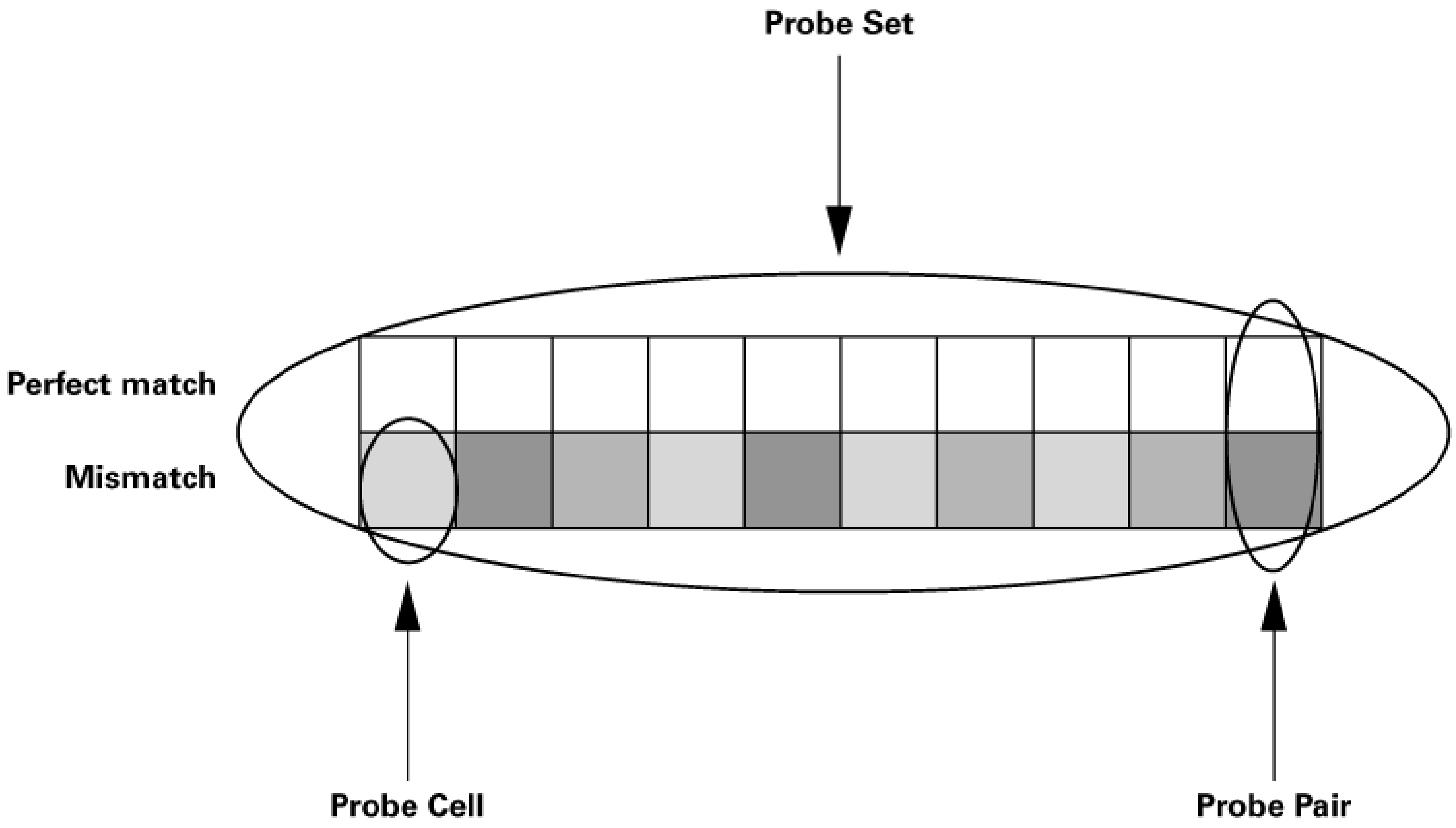
- Oligonucleotides are built up on the chip surface by sequentially elongating the growing chain with a single nucleotide. To determine the sequence of the final oligonucleotides on each position of the chip, a process called *photolithography* is used.

- As chemical yield of the stepwise elongation is limited, oligonucleotides can't be grown to more than 25 nt length.

- Hybridization to short oligos is quite unspecific, thus a number of them has to be used to probe for a single gene (usually 12–25).

- Frequently, cross-hybridization occurs. To eliminate this effect, hybridization is compared with that of an oligo that bears a single mismatch.

Affymetrix probe set



cDNA vs. Oligo Chips

- long DNA strands are more specific than oligos:
  - ★ cDNA chips: 1 (2,3 identical) spots per gene
  - ★ oligo chips: many oligos per gene

- Oligo chips by on-chip synthesis:
  Affymetrix GeneChip^{TM}:
  - ★ Single-color readout
  - ★ approx. 20 oligos per gene
  - ★ mismatched control for *every* oligo
  - ★ sophisticated weighting and averaging over 20 oligo pairs
  - ★ much of the information is proprietary

References

- J. Khan, M. Bittner, Y. Chen, P.S. Meltzer, and J. Trent (1999) DNA microarray technology: the anticipated impact on the study of human disease. *Biochim Biophys Acta* **1423**: M17–M28

- D.J. Duggan, M. Bittner, Y. Chen, and J.M. Trent (1999) Expression profiling using cDNA microarrays. *Nat Genet* **21** (Suppl): 10–14

- P. Hedge, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snesrud, N. Lee, and J. Quackenbush (2000) A concise guide to cDNA microarray analysis. *Biotechniques* **29**: 548–562

- A. Brazma and J. Vilo (2000) Gene expression data analysis. *FEBS Lett* **480**:17–24

- D.J. Lockhart and E.A: Winzeler (2000) Genomics, gene expression and DNA arrays. *Nature* **405**: 827–836

- S. Lampel and P. Lichter (2000) Nukleinsäurechips. *Medizinische Genetik* **12**: 287–289

Measures of expression

- For cDNA chips, mostly the *ratio of expressions* is used:

$$\text{ratio}_i = \frac{R_i}{G_i}$$

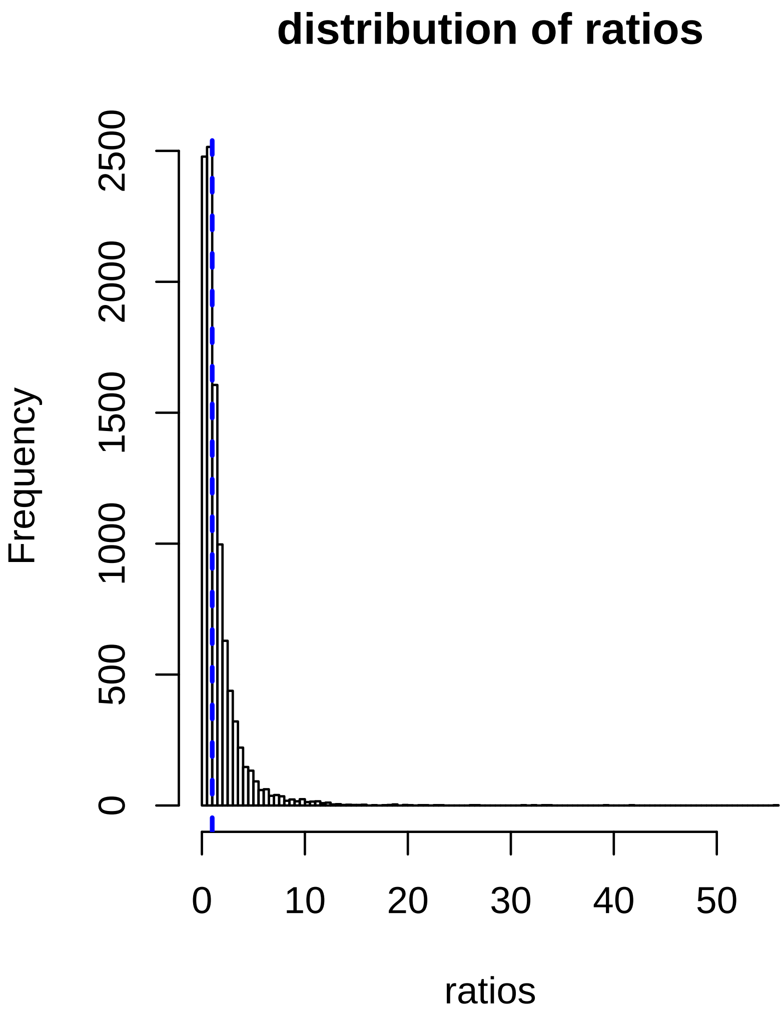
- The logarithm of the ratios is symmetric around ratio=1 (no change with respect to control condition):

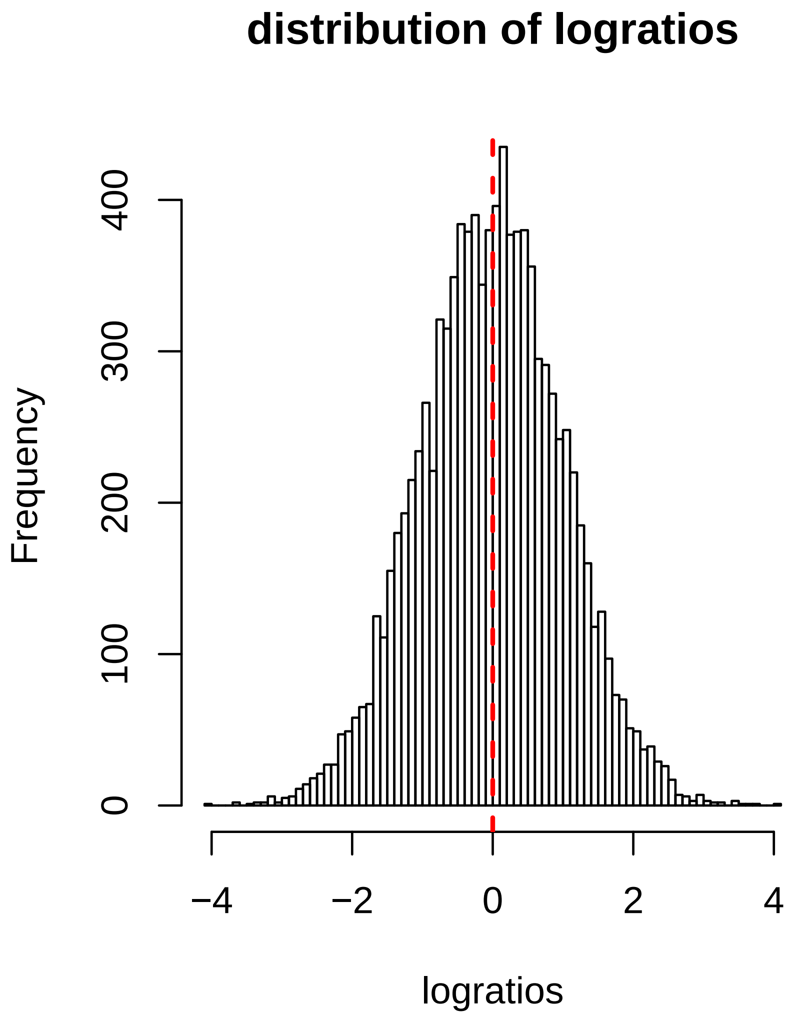
$$\text{logratio}_i = \log \frac{R_i}{G_i} = \log R_i - \log G_i$$

- Logratios to different bases of the logarithm (2, e, 10) are identical up to a constant factor:

$$\log_2(x) = \log_{10}(x) \cdot \log_2(10)$$

Distribution of ratios/logratios


distribution of ratios


distribution of logratios

Measures of expression 2

- Ratios are independent of absolute signal intensity, i.e. $R_i/G_i = 20/10 = 2$ will give the same ratio as $R_i/G_i = 20,000/10,000$. Sometimes, values of M and A are used:

$$\begin{aligned} M &= \log \frac{R_i}{G_i} && \text{(logratio)} \\ A &= 1/2 \log(R_i \cdot G_i) && \text{(average expression)} \end{aligned}$$

- For Affymetrix-type arrays, the signal intensities of the whole probe set have to be aggregated first. Affymetrix software (MAS) uses trimmed means:

$$\text{AvgDiff} = \frac{1}{|A|} \sum_{j \in A} (PM_j - MM_j)\,, \quad A \subset N$$

Questions asked to microarray data: three case studies

Alizadeh et al.: Lymphoma

- Study was published in *Nature* **403**:503–511 (2000)

- Gene expression profiling of Diffuse Large B-Cell Lymphoma (DLBCL)

- Lymphoma is a blood cancer where *peripheral* blood cells degenerate and divide without control

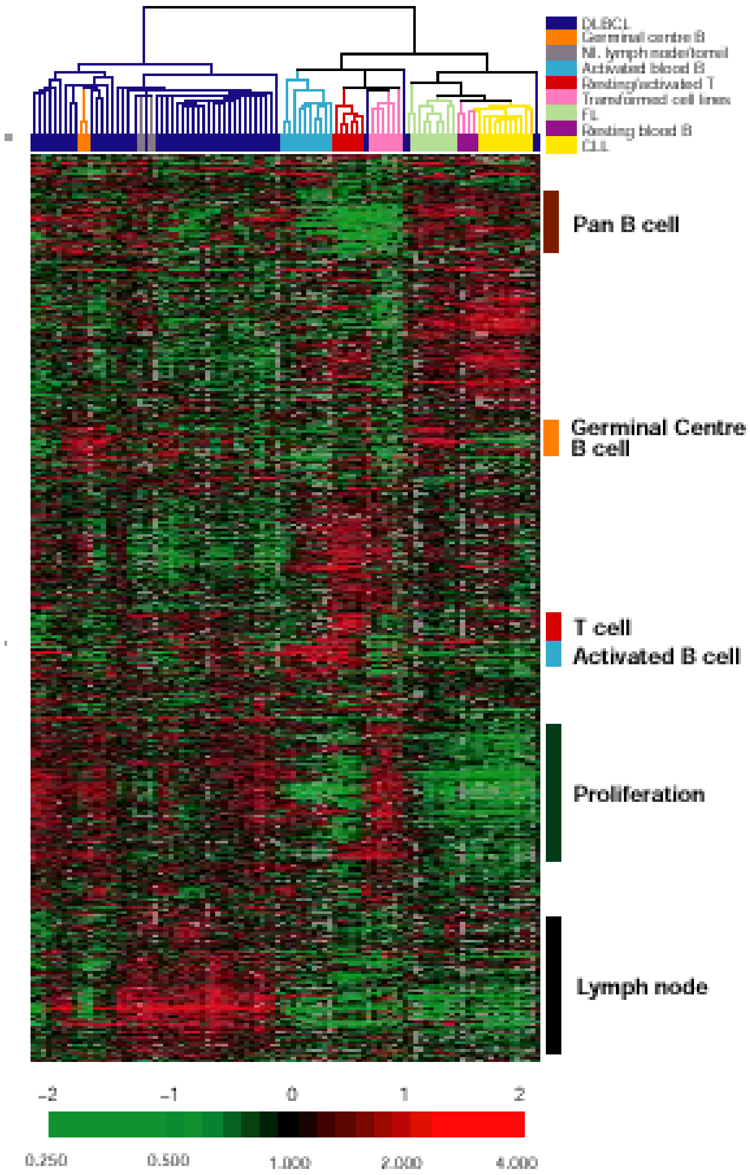
- DLBCL is an aggresive form of this disease, originating from B-lymphocytes. Overall 5-year survival is about 40%.

- Current clinical risk factors are not sufficient.

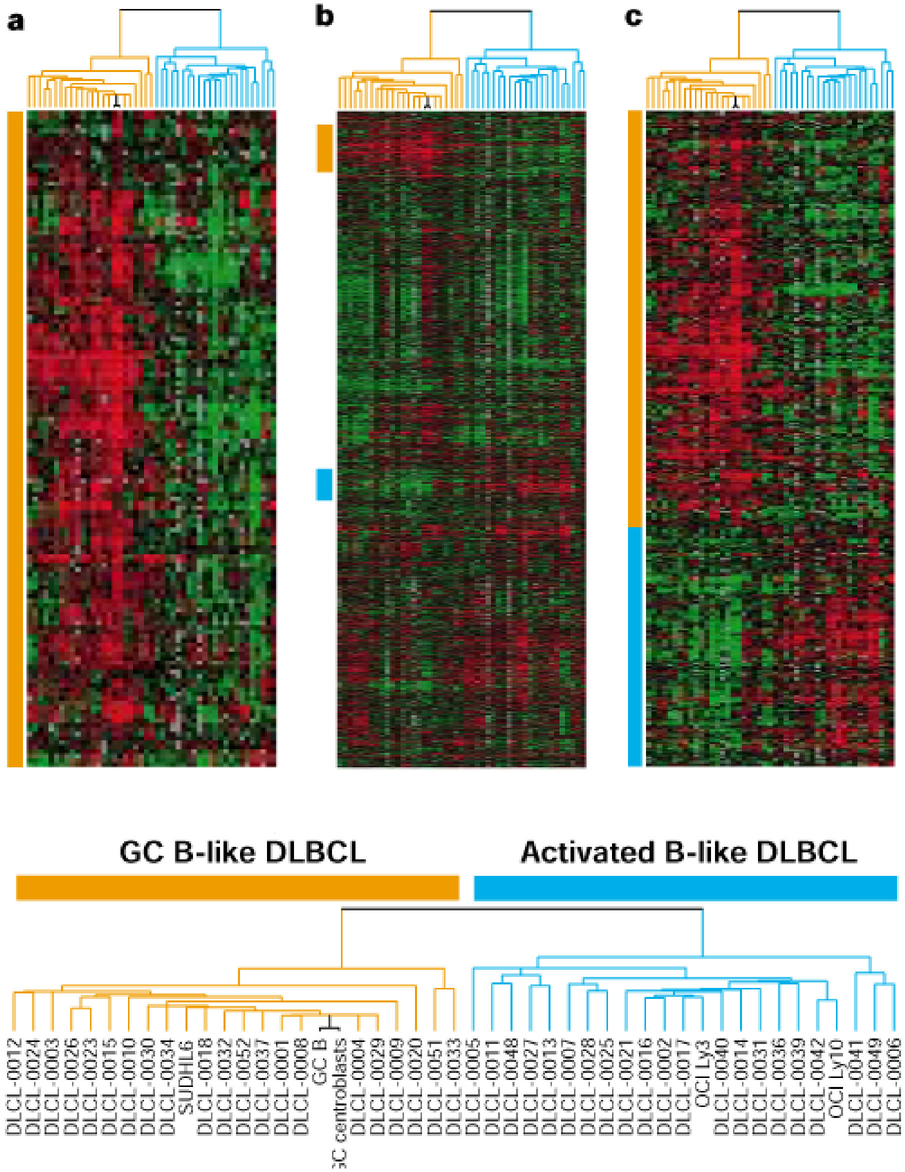
Alizadeh et al.: Methods

- A special cDNA chip was used, the *Lymphochip*
- spotted cDNA array of approximately 17,000 clones related to Lymphocytes
- 42 samples of DLBCL were analyzed, plus additional samples of normal B cells and of related diseases
- mRNA from these samples was competitively hybridized against control mRNA, stemming from a pool of lymphoma cell line mRNA preparations
- Data were analyzed by clustering

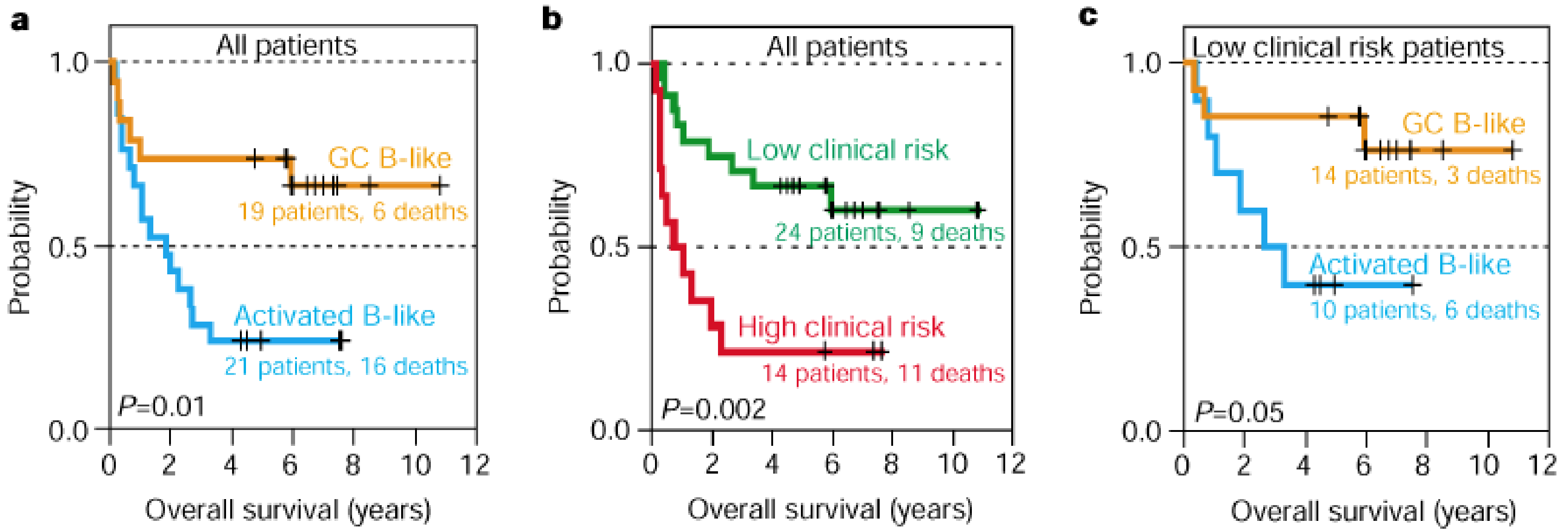
Alizadeh et al.: Results 1



Alizadeh et al.: Results 2



Alizadeh et al.: Results 3



Van’t Veer et al.: Breast cancer

- published in *Nature* **415**:530–536 (2002)

- looks for prognostic markers in breast cancer

- two classes of patients: those with distant metastasis (other than in breast) within 5 years, and those without (also had negative lymph node status)

- In statistical thinking, this is a *classification* problem: given a set of *variables*, can we train a *classifier* such that it predicts for any new sample the *class* as correctly as possible?

Van’t Veer et al.: Methods

- A custom-made 25,000-clone chip was used; each feature contained a unique 60-mer oligonucleotide. This oligo was transferred to the chip by ink-jet likr printing.

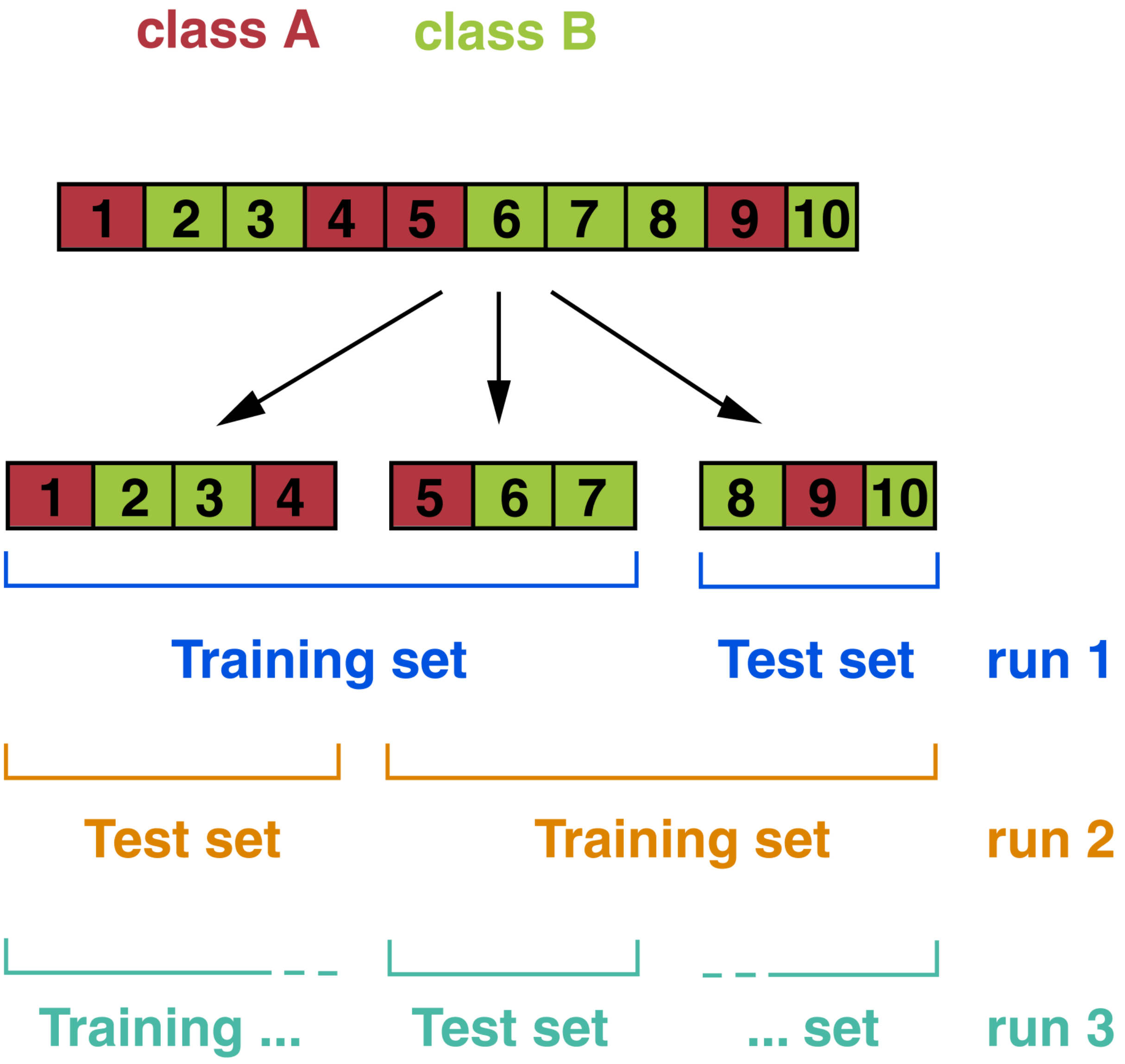
- The chips were hybridized competitively; the reference mRNA was obtained from a pool of patient mRNA (98 patients in total).

- Only data from certain genes (231) were used; finding out informative genes is called *feature selection* in machine learning.

- A home-made *ad hoc* classification method was used (no details given here). You can do better with established classification methods (tought later in this course).

- The model was validated by cross validation and by an independent test set.

Cross-validation



class A class B

1 2 3 4 5 6 7 8 9 10

1 2 3 4

5 6 7

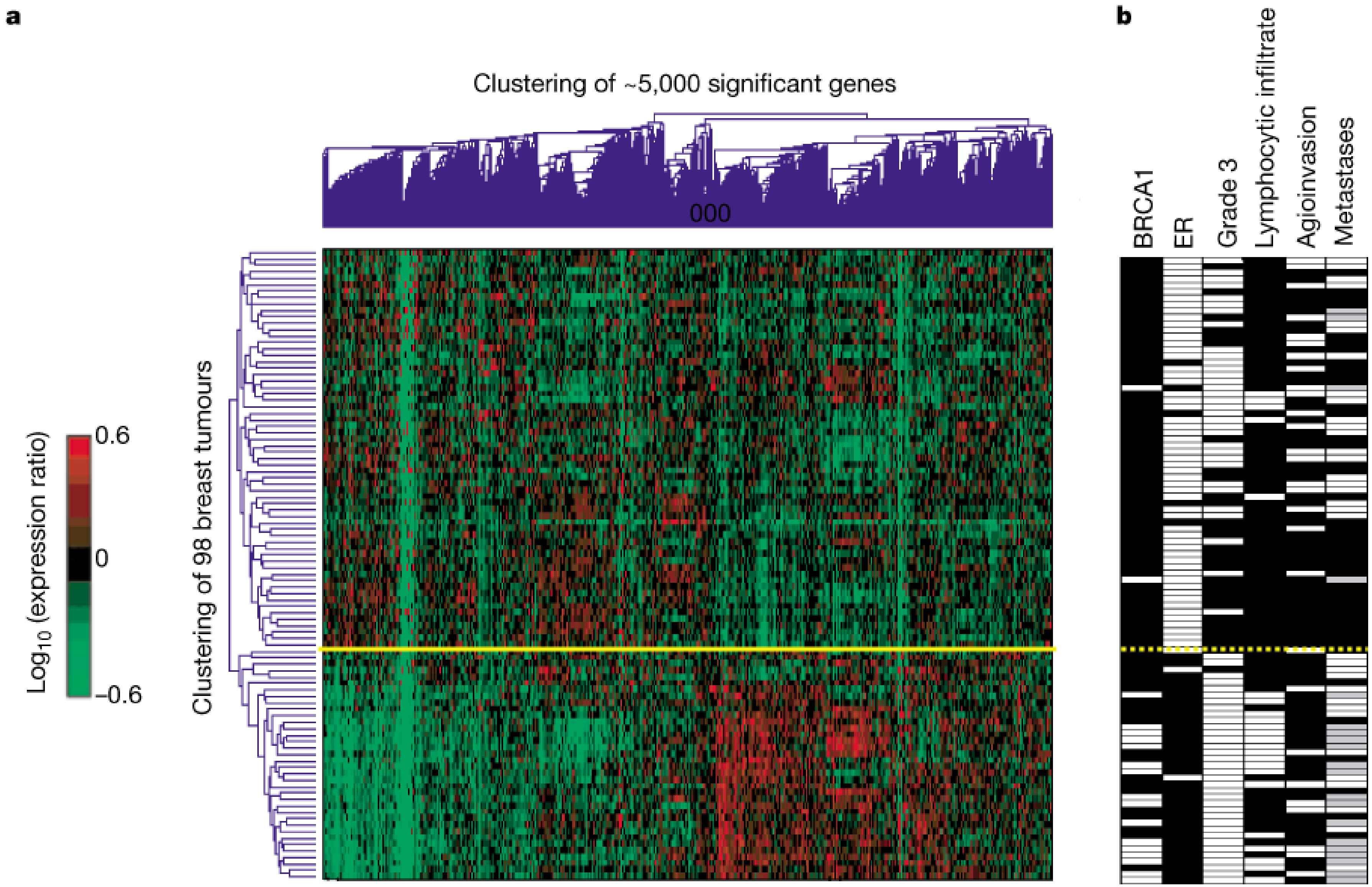
8 9 10

Training set Test set run 1

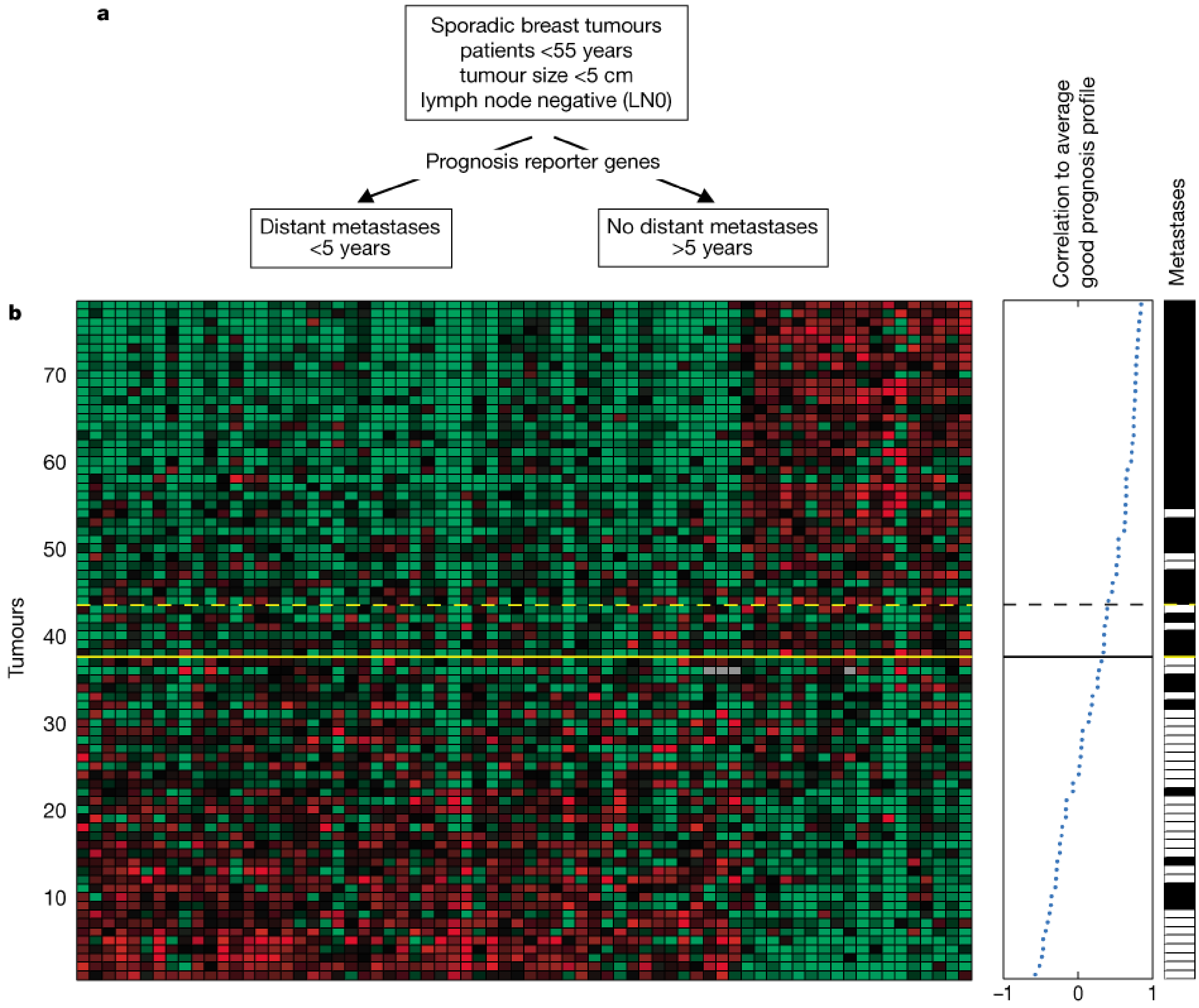
Test set Training set run 2

Training ... Test set ... set run 3

Van’t Veer at al.: Results 1



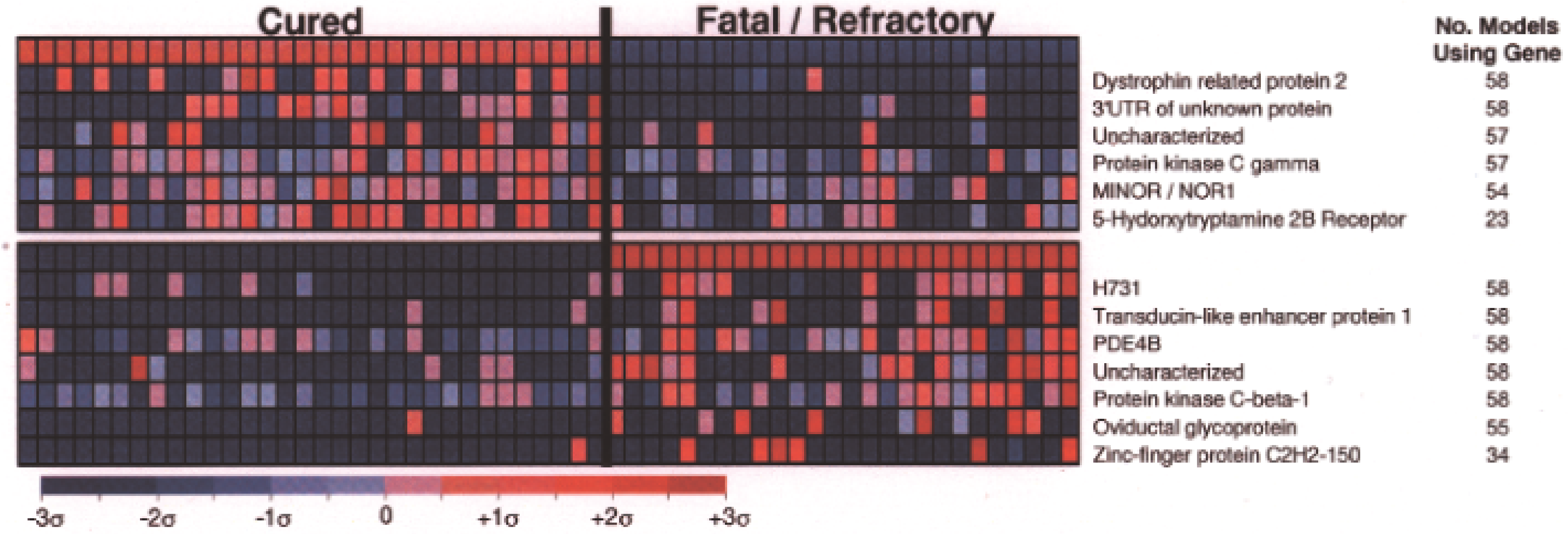
Van’t Veer et al.: Results 2



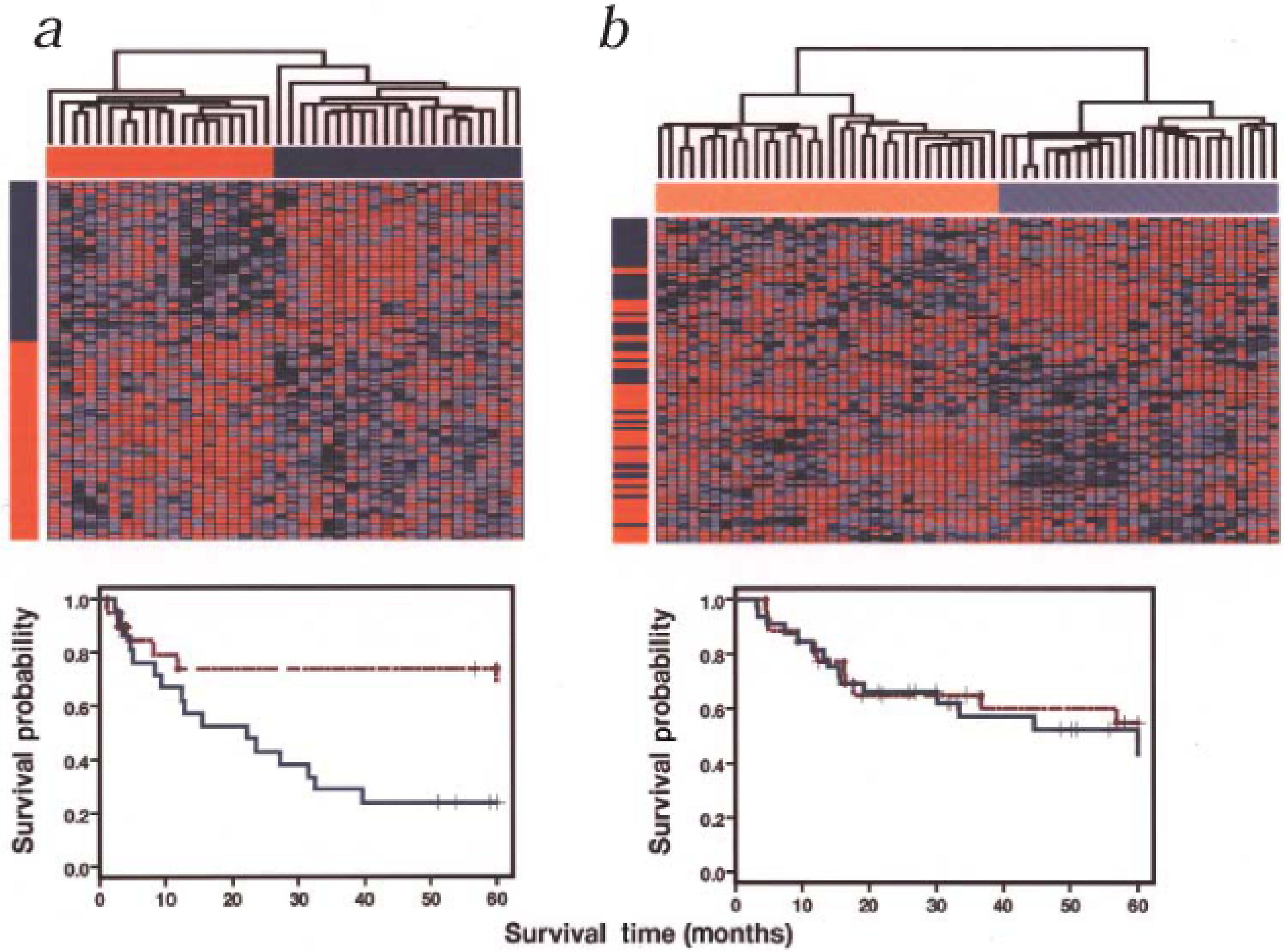
Shipp et al.: Lymphoma again

- published in *Nat. Med.* **8**:68–74
- Same lymphoma (DLBCL) as in the study of Alizadeh et al. was investigated
- Samples from 58 patients with DLBCL were subjected to gene expression analysis
- Affymetrix chip was used (6,800 probe sets)
- A classification (supervised) approach was taken
- Results were compared with those of Alizadeh et al.

Shipp et al.: Results 1



Shipp et al.: Results2



Possible extension: Regression

- This was treated as a classification problem, i.e. there were distinct *classes* (cured vs. fatal) as *response variables*

- One could also use a *continuous* response variable: e.g. survival time, or the probability of being cured

- Fitting a model that predicts a continuous response is called *regression* in statistics (methods to be discussed later)