ULB
MLG

Veille Technologique en Informatique

Use of Machine Learning in Bioinformatics to Identify Molecular Signatures in Breast Cancer

Benjamin Haibe-Kains

bhaibeka@ulb.ac.be

Université Libre de Bruxelles

Institut Jules Bordet

Table of Contents

Breast Cancer

Breast Cancer

- The cancer is a **genetic disease** at the level of the cell.
- Several mutations are necessary in the genome of a normal cell to involve its transformation into malignant cell.
- The cancer can affect different types of organs as brain, bones, **breast**, etc.

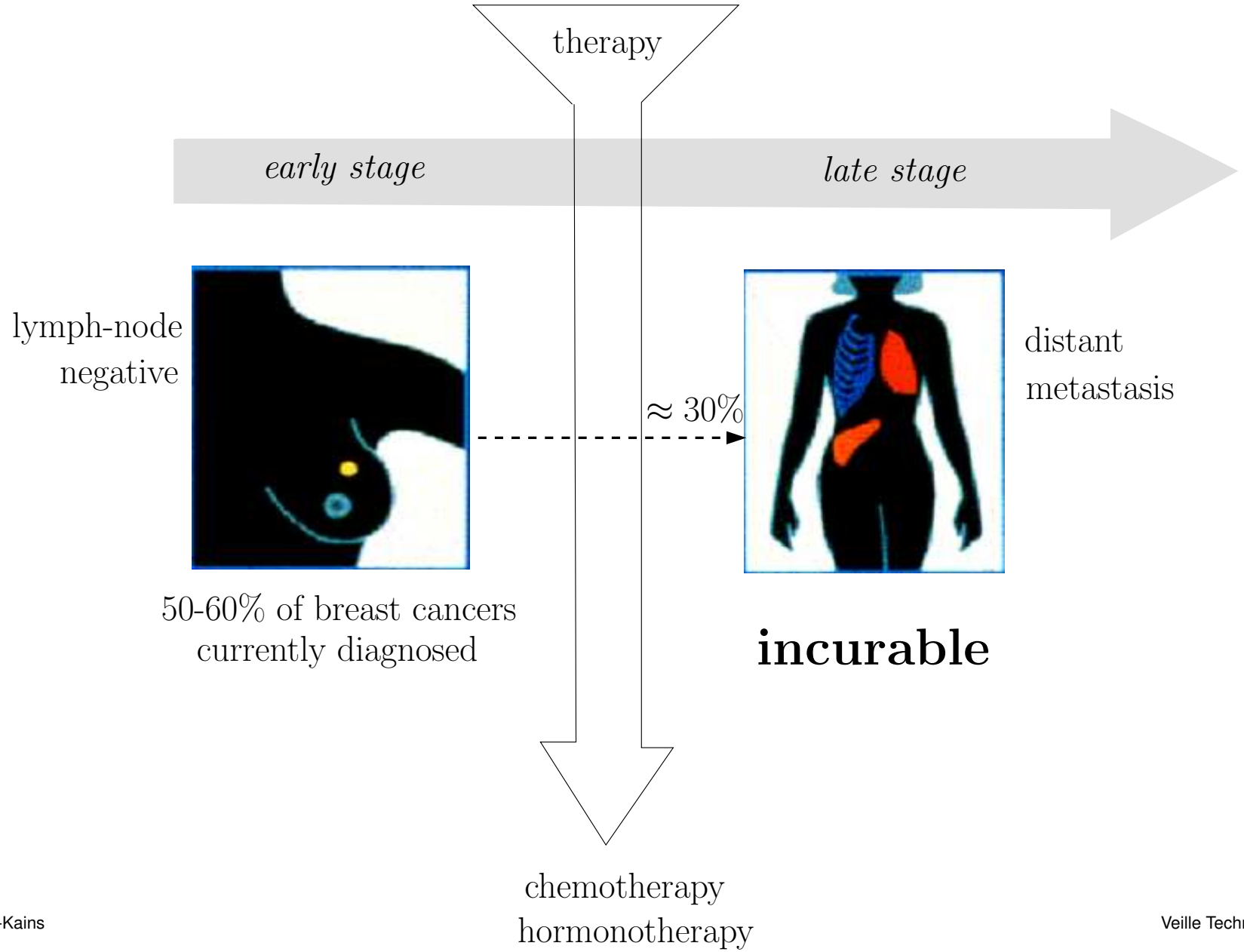
Breast Cancer(2)

- Breast cancer is one of the most common malignant tumors affecting women.
- Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome.
- Breast cancer classification has been based primarily on histological criteria. For instance :
  - invasive/non-invasive tumor
  - Number of involved lymph-nodes
  - tumor size
  - hormone receptor status (ER/PGR)

Breast Cancer(3)

- Unfortunately, current classifications have serious limitations. Tumors with similar histological criteria
  - can follow significantly different clinical courses (prognosis)
  - can show different responses to therapy (prediction).
- The strongest predictors for **metastasis** fail to classify accurately breast tumors according to their histological criteria.

Breast Cancer(4)



Breast Cancer(5)

- Chemotherapy or hormonotherapy reduces the risk of distant metastasis by $2-12\%$
- However
  - $\approx 70\%$ of patients receiving this treatment would have survived without it
  - these therapies frequently have **toxic side effects**.
- Classification of cancers must be accurate in order to give the correct treatment and so increase the chance of survival for the patient.

Breast Cancer(6)

- Such classification is difficult because of
  - the cellular and molecular heterogeneity of breast tumors
  - the large number of genes potentially involved in controlling cell growth, apoptosis and differentiation.
- These two characteristics emphasize the importance of studying **multiple genetic alterations** in cancer.

Research Groups

Microarray Unit

- Laboratory of the Institut Jules Bordet (IJB).
- 6 researchers (funds coming from IJB, Télévie, FNRS, etc.).
- Numerous running projects concerning the breast cancer :
  - Appearance of distant metastases
  - Response to therapies
  - Refinement of histological criteria by microarray, etc.
- Scientific collaborations :
  - Singapore Lab (Lence Miller and Edison)
  - Swiss Institute of Bioinformatics, etc.

Microarray Unit(2)

- Biological and medical facilities :
  - AFFYMETRIX© and cDNA microarray platforms
  - all the kits necessary to check the quality of the biological samples (AGILENT©) and to perform microarray experiments
  - access to the tumor bank of IJB.
- Computing facilities :
  - 2 workstations (P4 3.6 GHz, 4 Go RAM)
  - access to LIT5 cluster.
- Website : `http://www.bordet.be/servmed/array/`.

Machine Learning Group

- Research group of the Université Libre de Bruxelles (ULB).
- 7 researchers (funds coming from ULB, ARC, European Community, etc.).
- Research topics :
  - Local learning, Classification, Computational statistics, Data mining, Regression, Time series prediction, Sensor networks, Bioinformatics.
- Scientific collaborations in ULB :
  - IRIDIA (Sciences Appliquées), Physiologie Moléculaire de la Cellule (IBMM), Microarray Unit (IJB), Service d'Anesthésie (ERASME), etc.

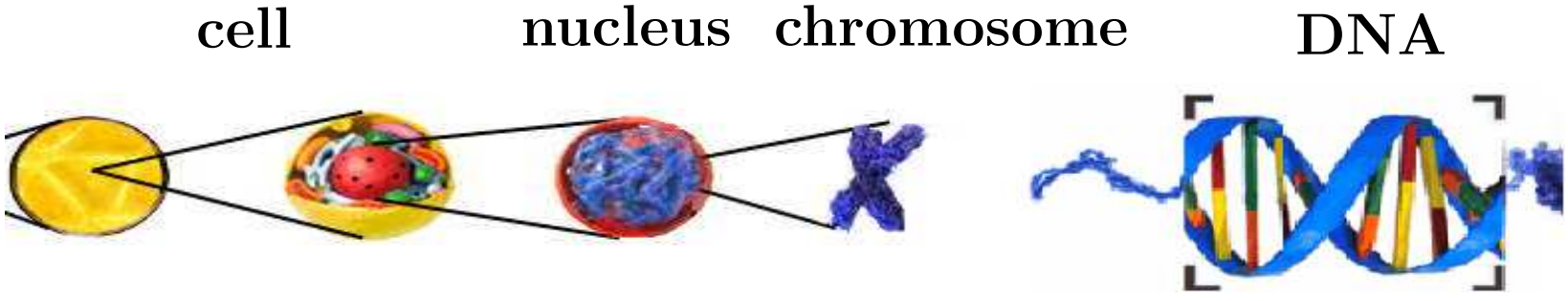
Machine Learning Group(2)

- Scientific collaborations outside ULB :
  - UCL Machine Learning Group (B), Politecnico di Milano (I), Università del Sannio (I), George Mason University (US), etc.
- Computing facilities :
  - LIT5 cluster (16 x P4 3.4 GHz, 16 x 2 Go RAM)
  - LEGO Robotics Lab.
- Website : `http://www.ulb.ac.be/di/mlg`.

Microarray Technology

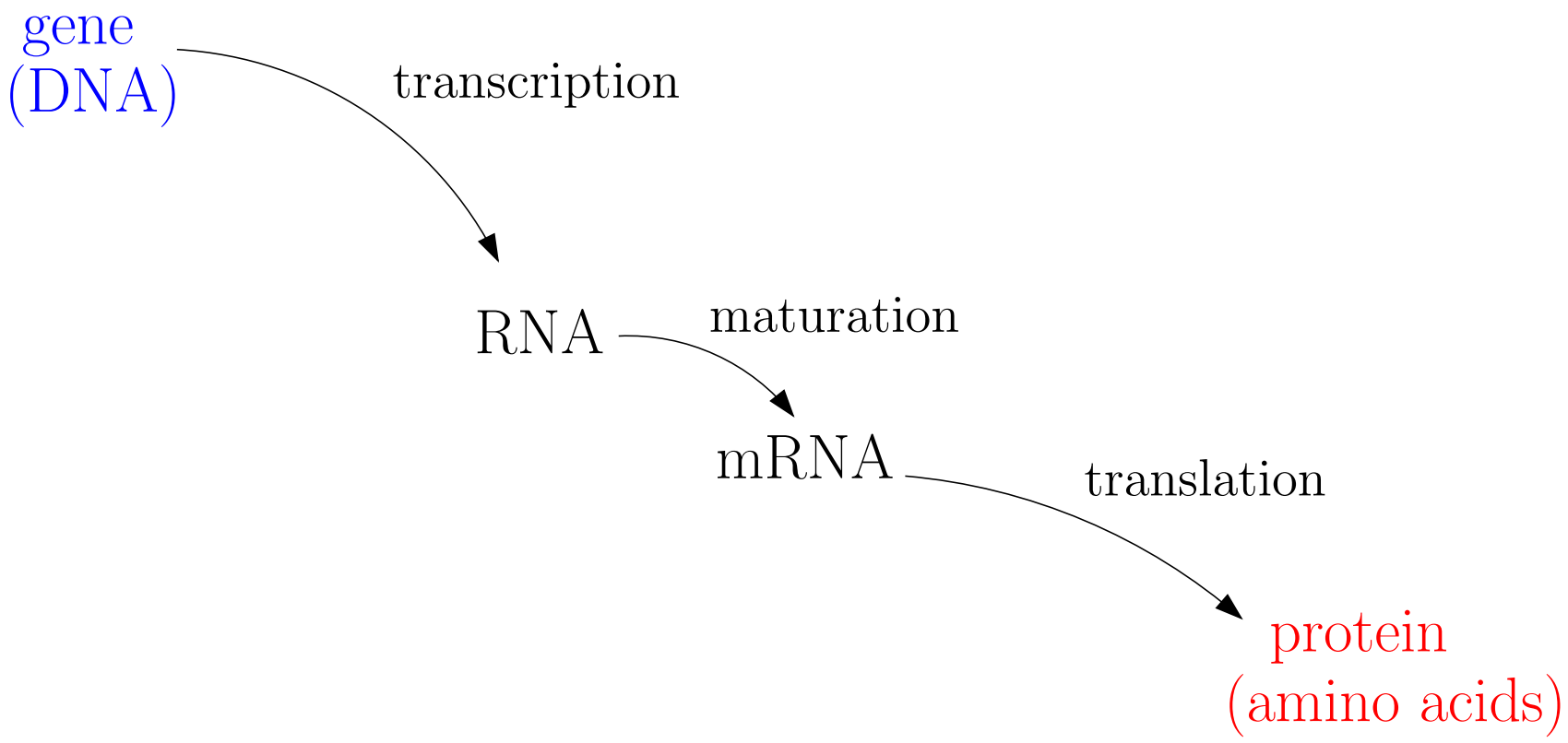
Genetics : basics

Cell to DNA :



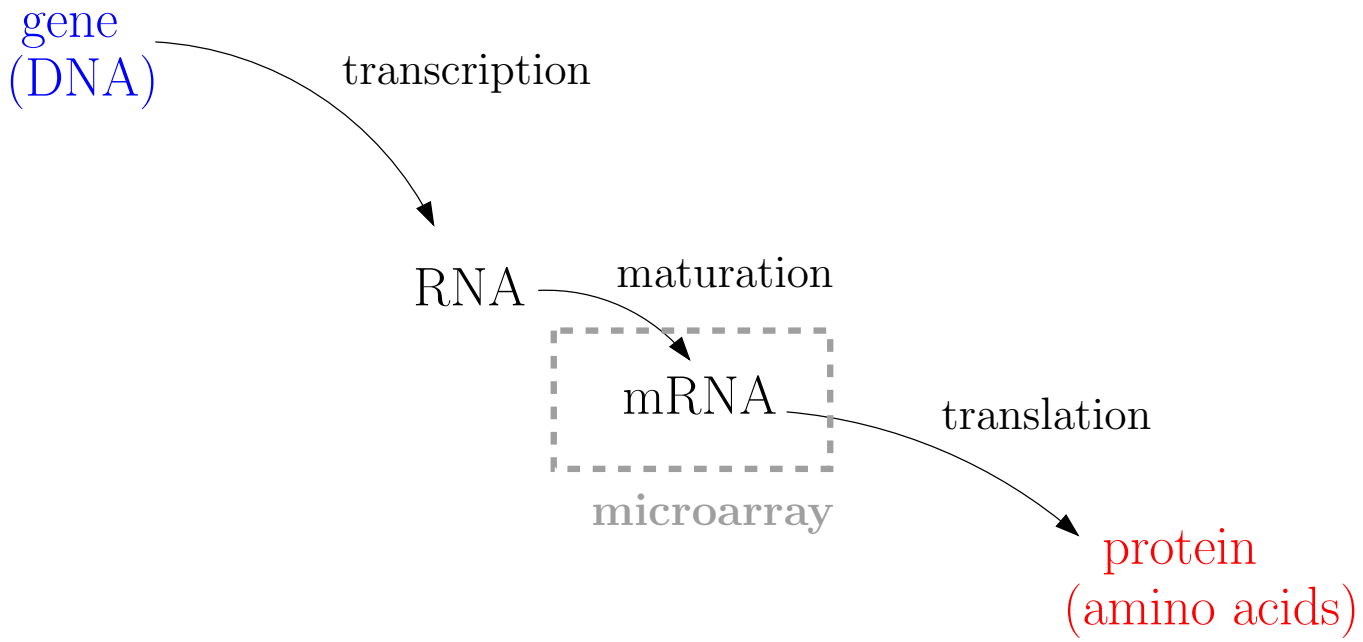
Genetics : basics (2)

DNA to protein :



Microarray Technology

- As we will see, the microarray technology allows us to study the **genetic profile** of breast tumors.
- Microarray works at the mRNA level :



Microarray Technology(2)

A *microarray* is composed of

- DNA fragments fixed on a solid support
- ordered position of probes
- principle of hybridization to a specific probe of complementary sequence
- radioactive labeling

➥ simultaneous detection of thousands of sequences in parallel

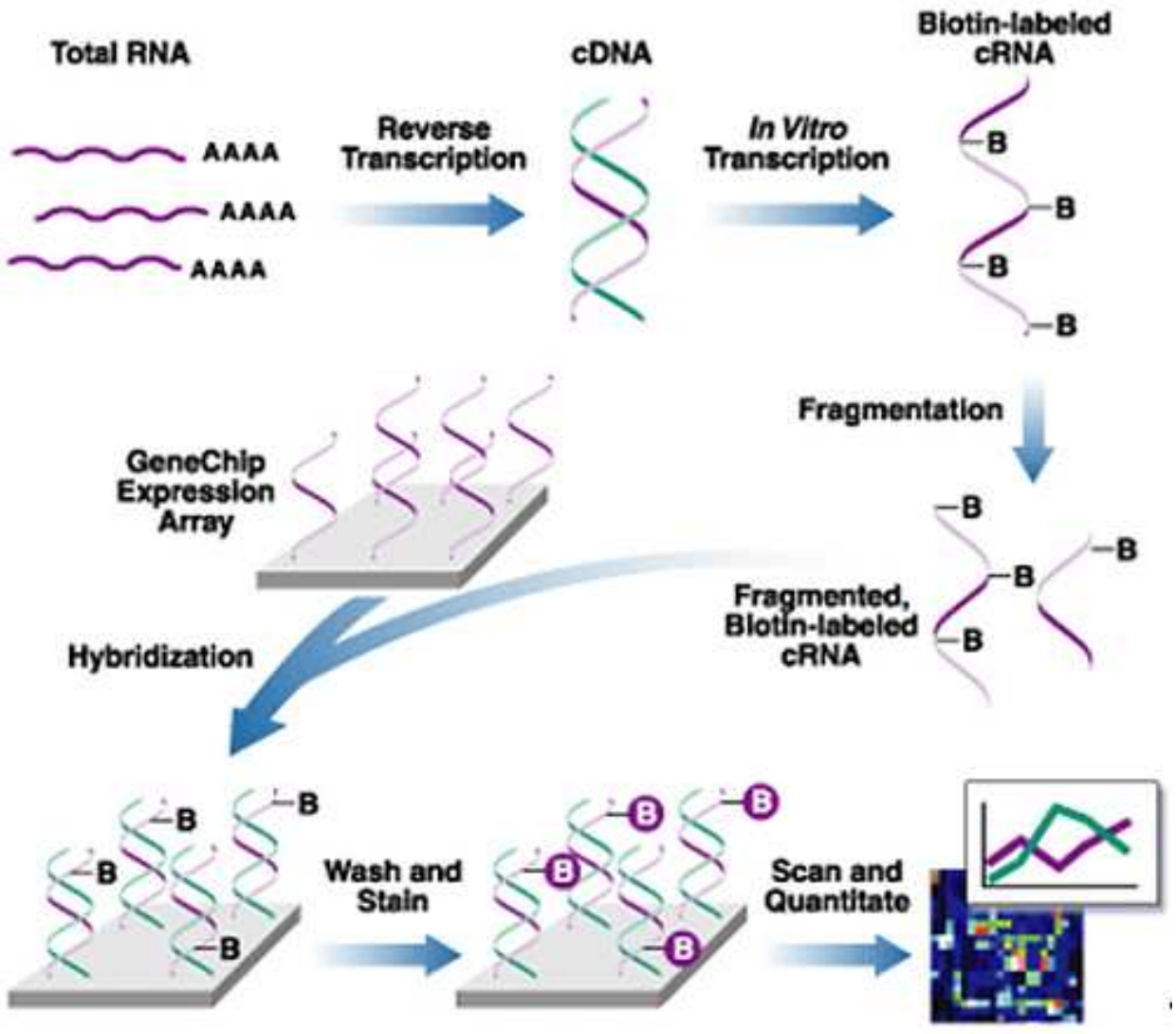
Microarray Technology(3)

It exists several high-throughput methods to simultaneously measure the expression of a large number of genes :

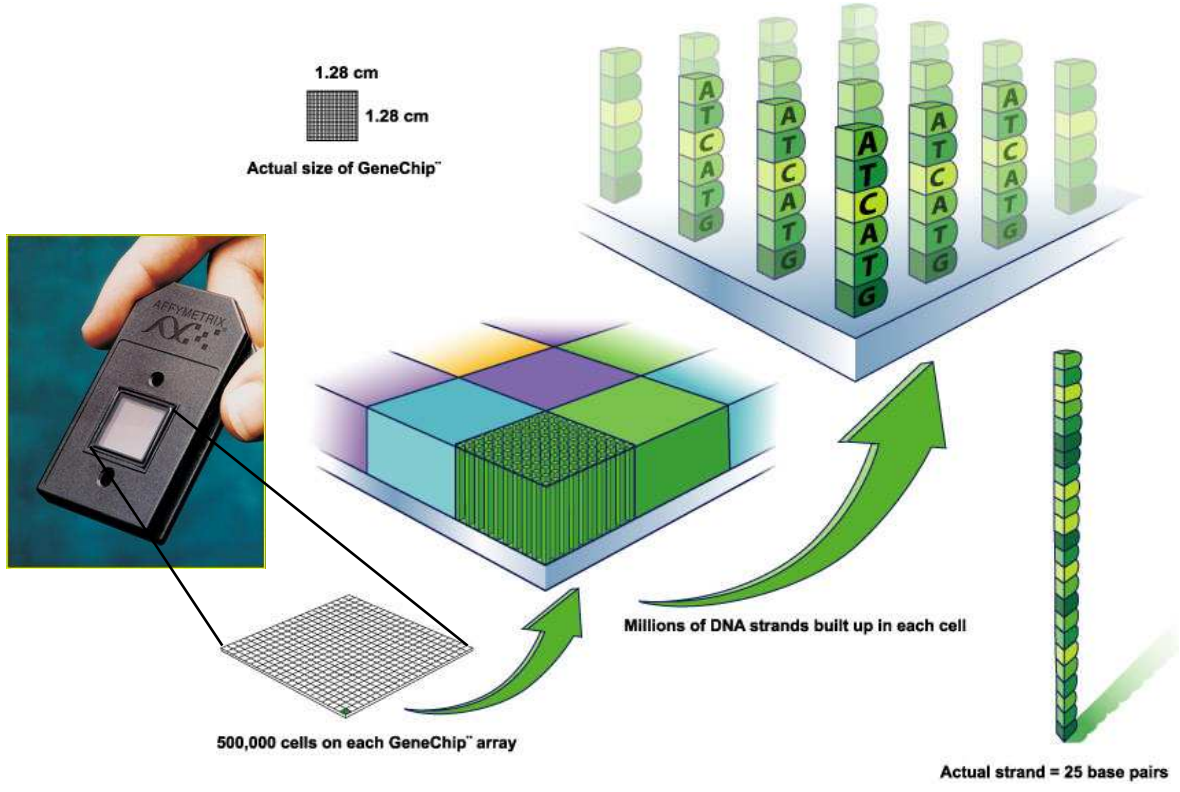
- cDNA microarray
- oligonucleotide microarray
  - short oligonucleotide (AFFYMETRIX©)
  - long oligonucleotide (AGILENT©, CODELINK©)
- multiplex quantitative RT-PCR

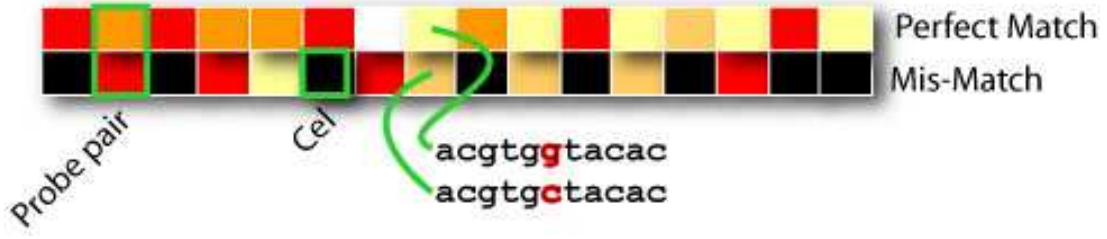
AFFYMETRIX© GeneChip

AFFYMETRIX© Design



AFFYMETRIX© GeneChip Structure



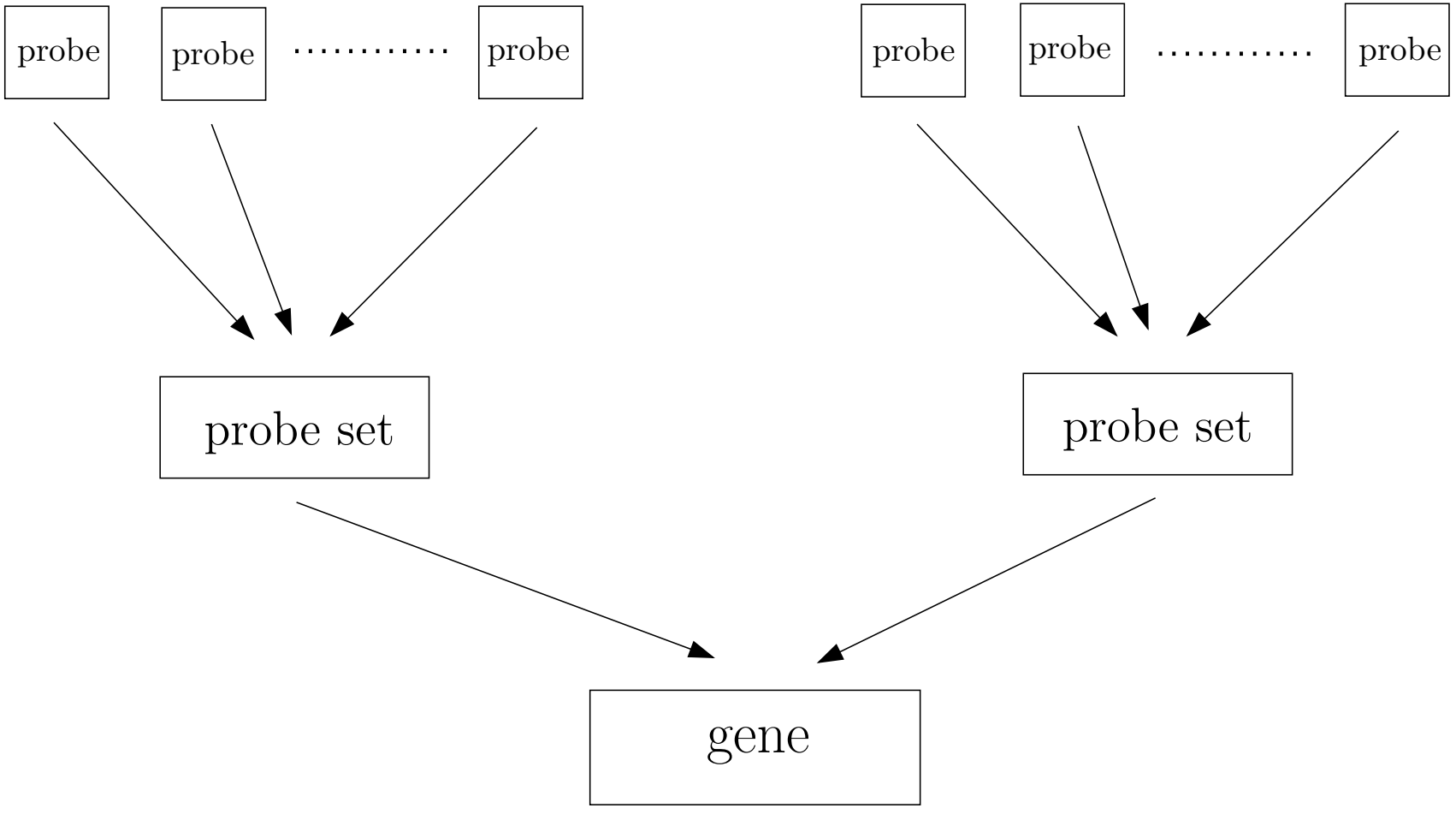


AFFY© GeneChip Structure(2)

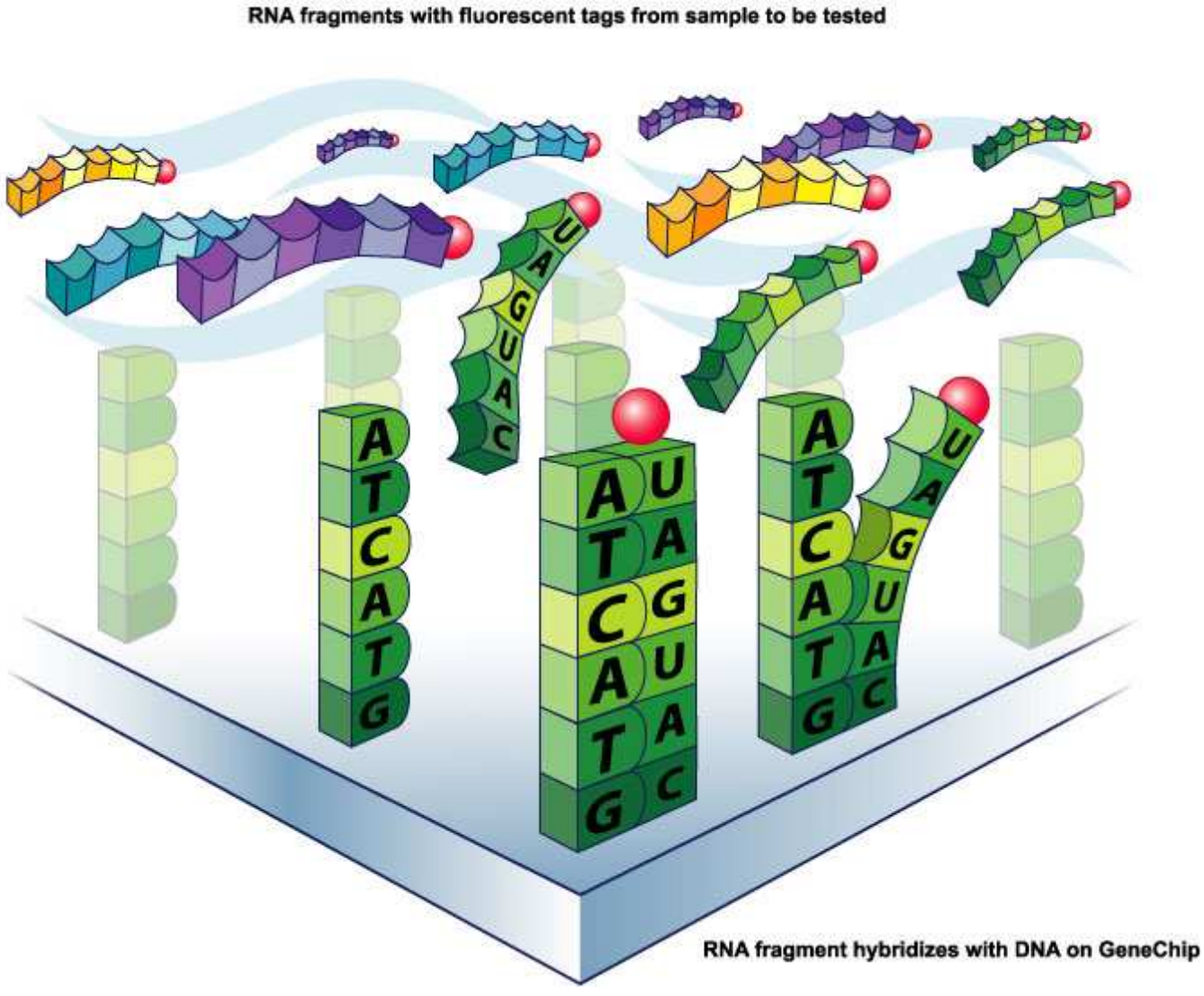
- 1 probe pair includes a Perfect Match (PM) value and a Mis-Match value (MM).
- 1 probe set includes 11 to 20 probe pairs.
- 1 gene is represented by 1 or more probe sets.

To facilitate the explanation, we assume that 1 gene is represented by 1 probe set including 20 probe pairs (PM and MM)

AFFY© GeneChip Structure(3)

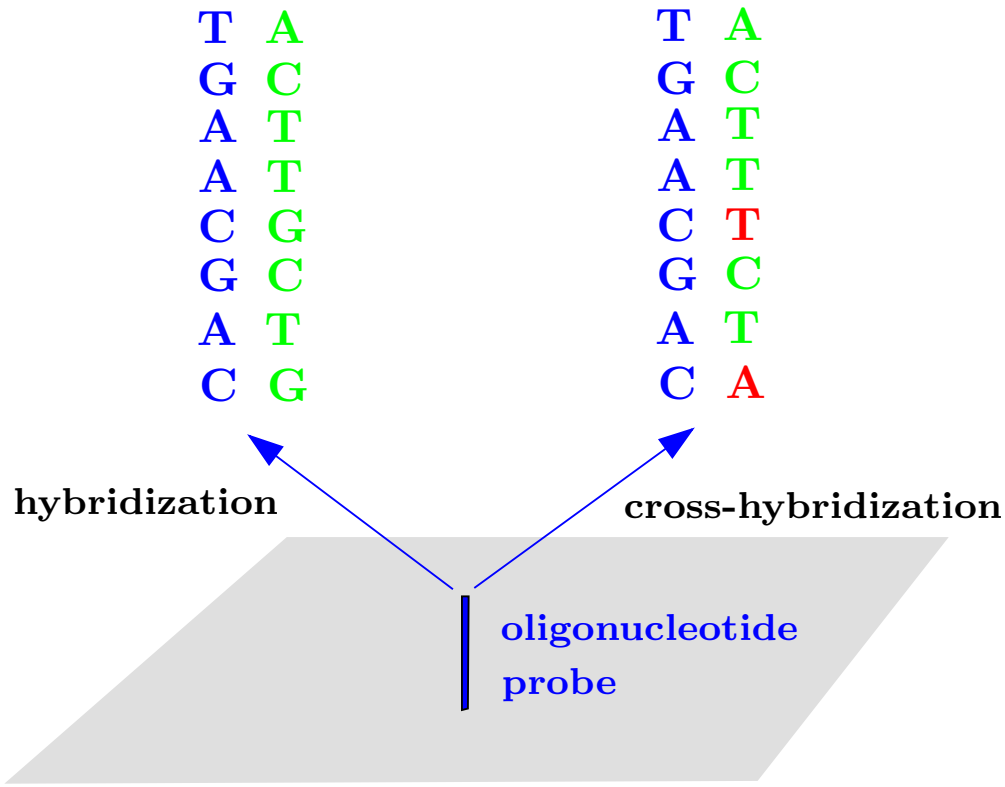


AFFYMETRIX© Hybridization

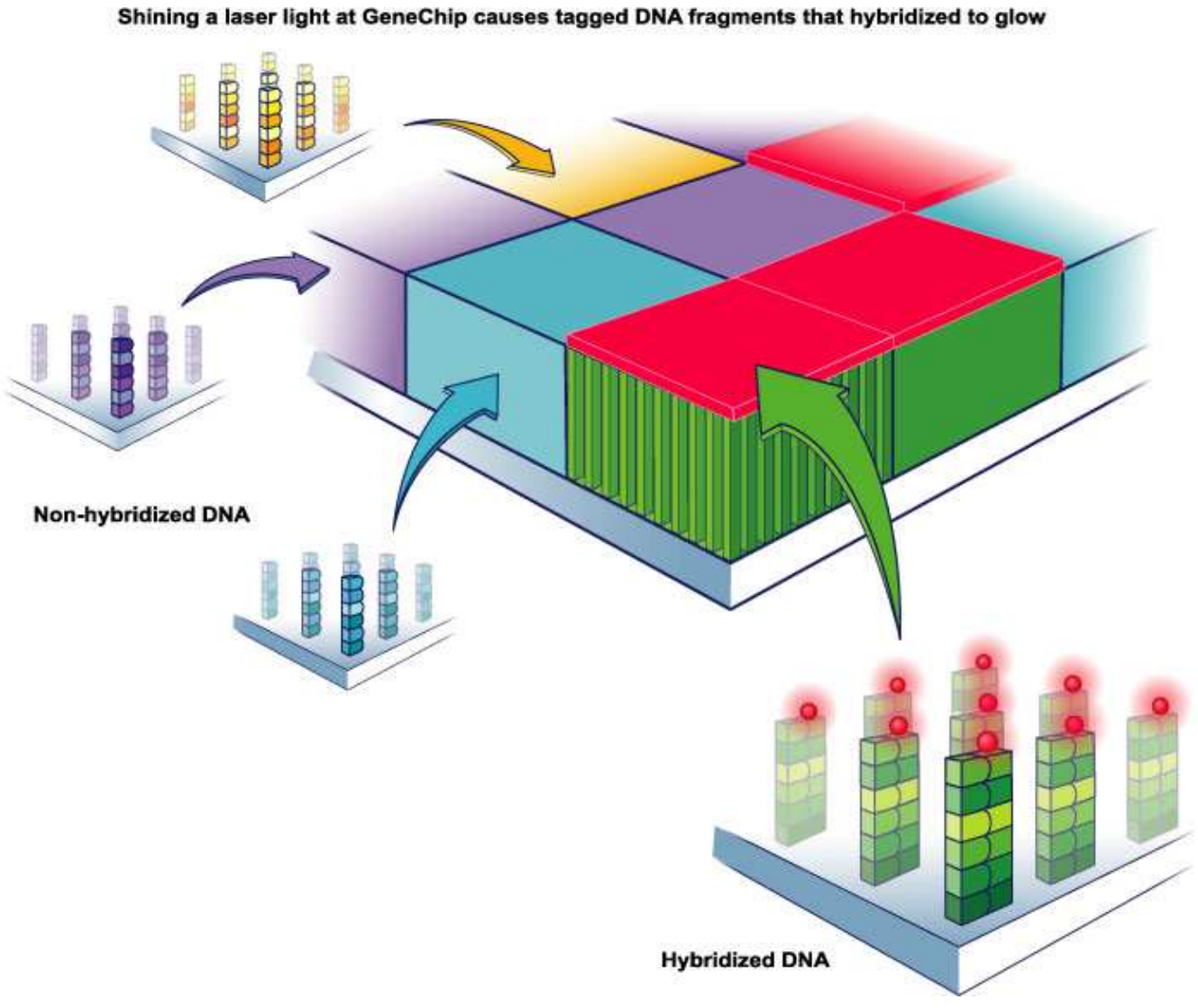


AFFYMETRIX© Hybridization(2)

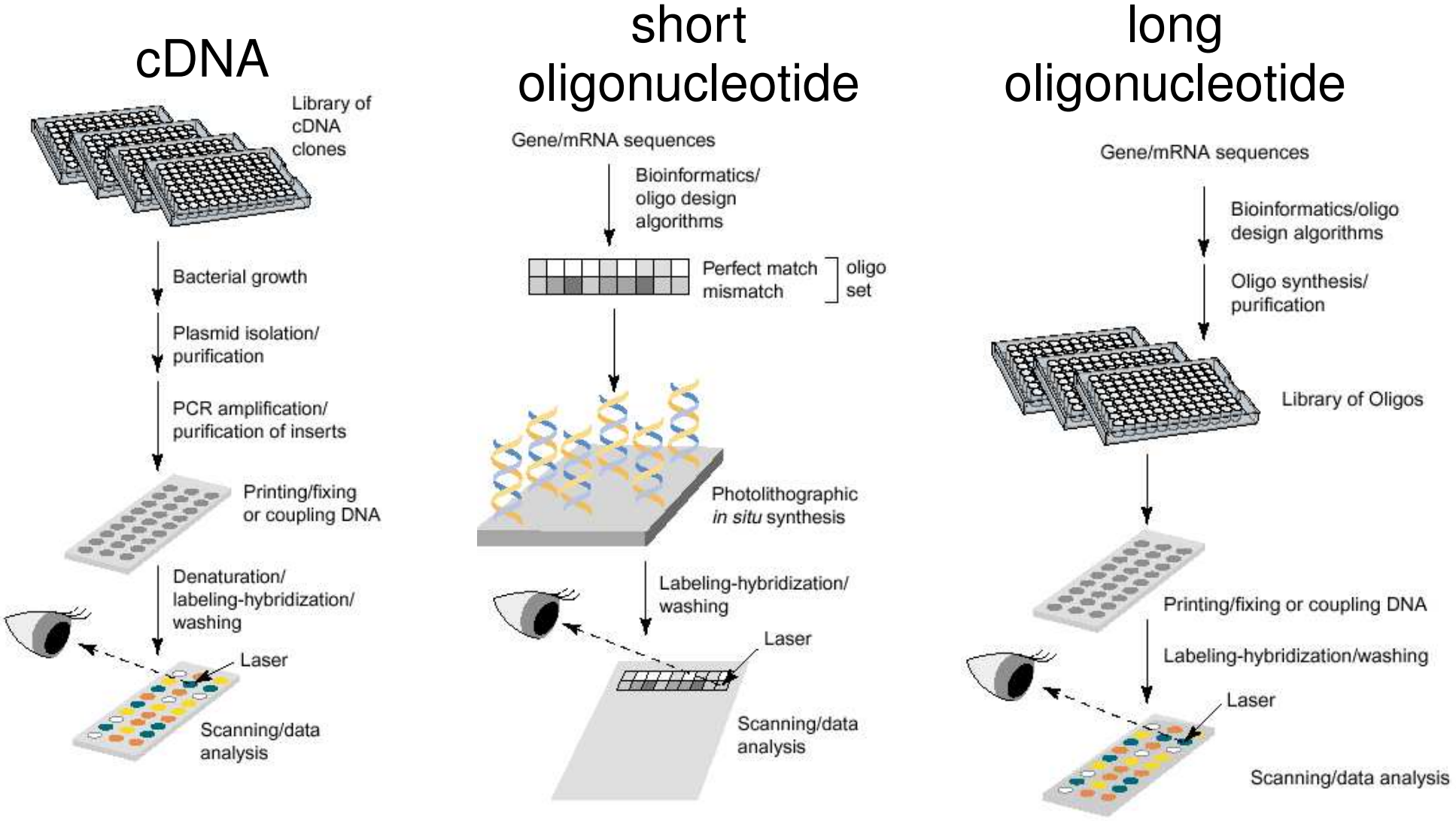
- The process of 2 complementary DNA strands binding is called *hybridization*.
- Ideally, an oligonucleotide probe will only bind to the DNA sequence for which it was designed and to which it is complementary.
- However, many DNA sequences are similar to one another and can bind to other probes on the array.
- This phenomenon is called *cross-hybridization*.



AFFYMETRIX© Detection



Microarray Comparison



Microarray Comparison(2)

AFFYMETRIX© advantages :

- commercially available for several years (strong manufacturing)
- large number of published studies (generally accepted method)
- no reference sample $\rightarrow$ possible comparison between studies

Microarray Comparison(3)

AFFYMETRIX© disadvantages :

- cost of the devices and the chips (but easy use)
- changes in probe design is hard (but new program permits to create his own design)
- short oligos $\longrightarrow$ several oligos per gene, specificity/sensitivity trade-off (complex methods to get gene expression)

Breast Cancer and Microarray

- The microarray technology provides the opportunity of correlating genome-wide expressions with the cancer evolution with/without therapies.
- Systematic investigation of expression patterns of thousands of genes in tumors and their correlation to specific features of phenotypic variation might provide the basis for an improved taxonomy of cancer.
- It is expected that variations in gene expression patterns in different tumors could provide a “molecular portrait” of each tumor, and that the tumors could be classified into subtypes based solely on the difference of expression patterns.

Breast Cancer and Microarray(2)

- In [van't Veer et al., 2002], classification techniques have been applied to identify a gene expression signature strongly predictive of a short interval (5 years) to distant metastases.
- In [jansen et al., 2005], survival analysis methods have been applied to identify a gene expression signature strongly predictive to the response of hormonotherapy (TAMOXIFEN©).
- In this context the number n of features equals the number of genes (ranging from 6,000 to 30,000) and the number N of samples is the number of patients under examinations (about hundreds).

Bioinformatics

Bioinformatics Definition

Bioinformatics *or computational biology is the use of techniques from applied mathematics, informatics, statistics, and computer science to solve biological problems. Research in computational biology often overlaps with systems biology. Major research efforts in the field include sequence alignment, protein structure prediction,* ***prediction of gene expression*** *and protein-protein interactions, and the modeling of evolution ... A common thread in projects in bioinformatics and computational biology is the use of mathematical tools to extract useful information from noisy data produced by* ***high-throughput biological techniques****. (The field of data mining overlaps with computational biology in this regard.)*

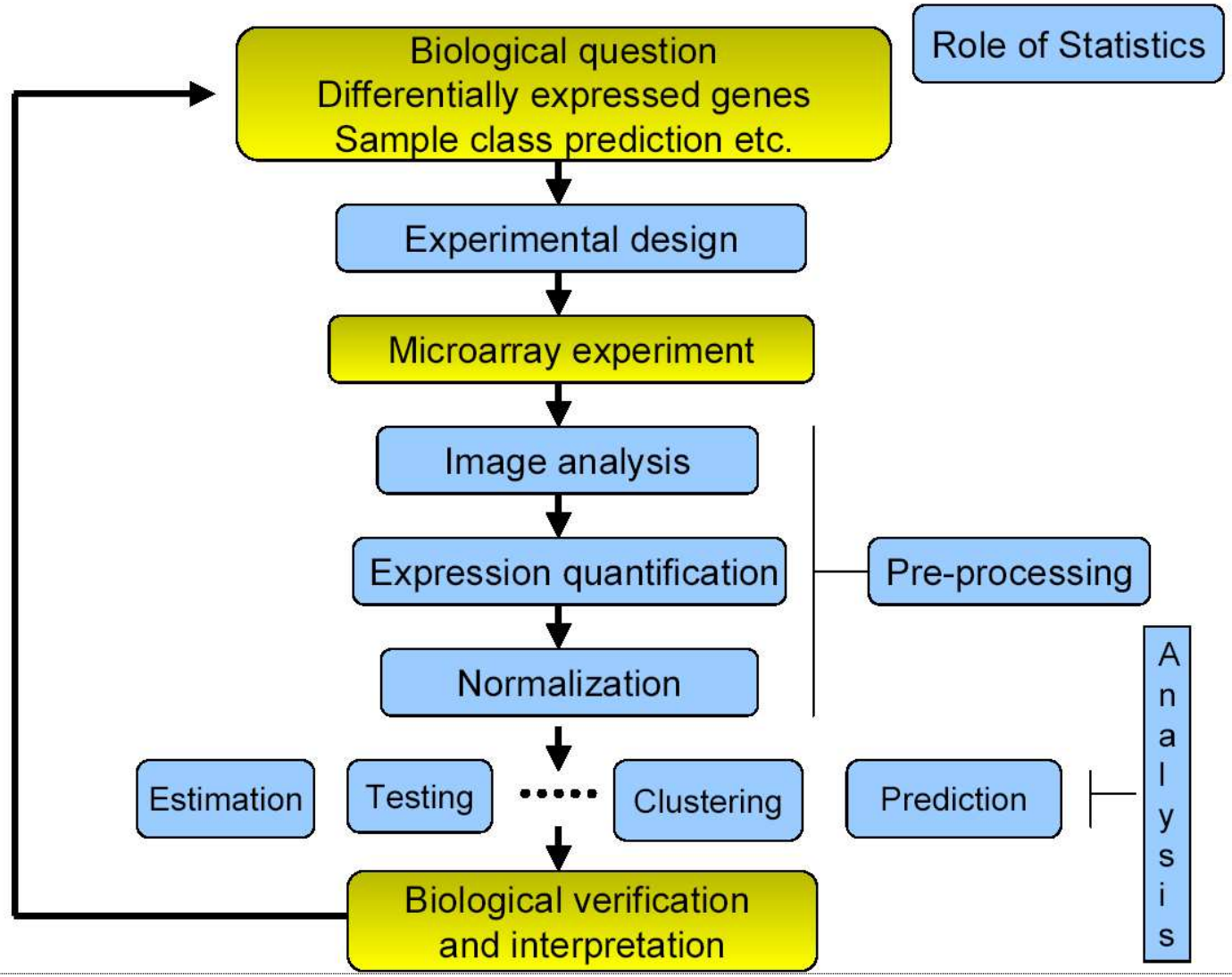
The definition comes from `http://en.wikipedia.org`

Machine Learning Definition

Machine learning *is an area of artificial intelligence concerned with the development of techniques which allow computers to "learn"Ṁore specifically, machine learning is a method for creating computer programs by the analysis of data sets. Machine learning overlaps heavily with statistics, since both fields study the analysis of data, but unlike statistics, machine learning is concerned with the algorithmic complexity of computational implementations. Many inference problems turn out to be NP-hard so part of machine learning research is the development of tractable approximate inference algorithms.*

The definition comes from `http://en.wikipedia.org`

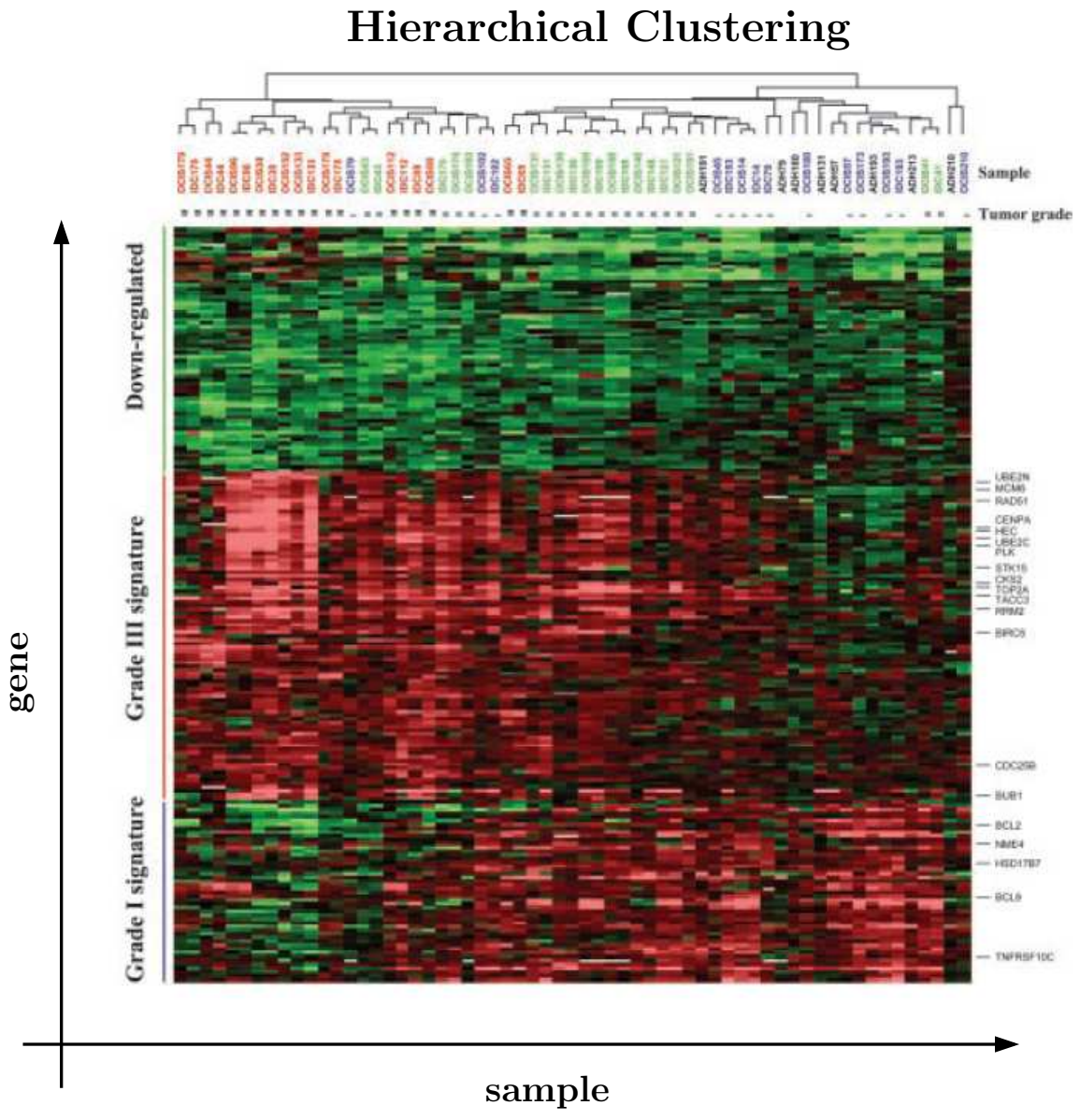
Microarray Analysis Design



Microarray Analysis

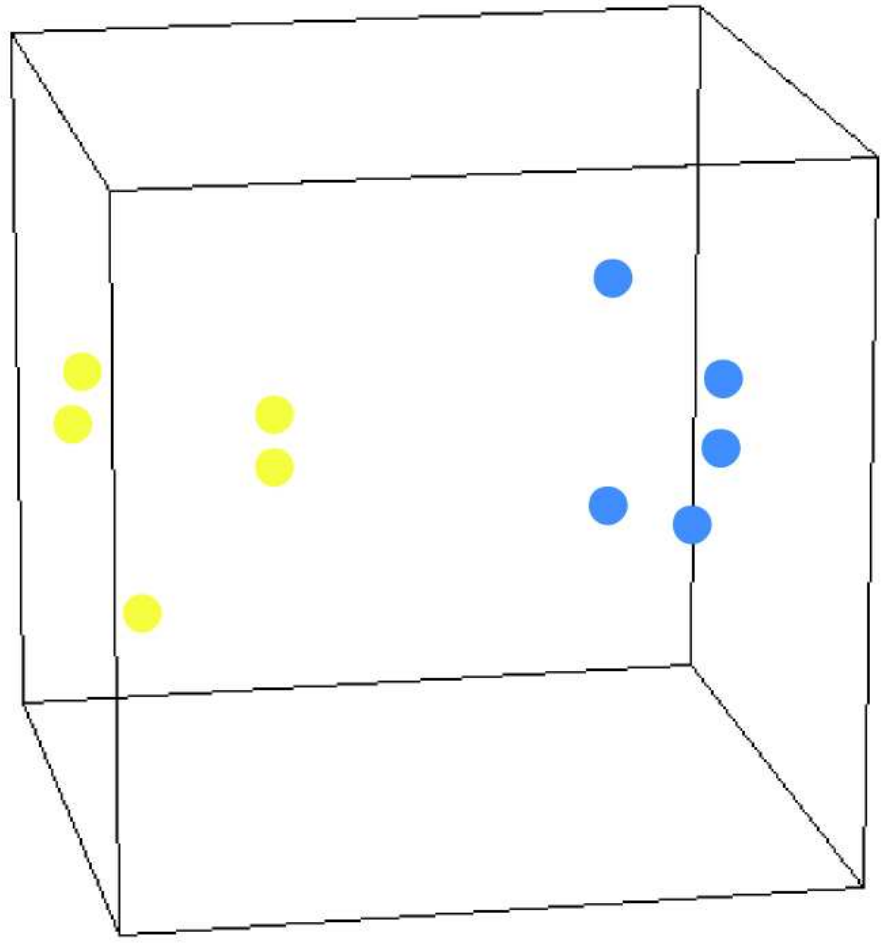
- Microarray analysis uses several machine learning methods as
  - feature selection : how to select the relevant genes ?
  - class-discovery methods : can we highlight unknown biological classes ?
  - supervised classification : can we predict known classes of new samples ?
  - ...
- These methods have to deal with noisy data, high feature/sample ratio and highly correlated variables.

Microarray Analysis : Example



Microarray Analysis : Example(2)

Dimension Reduction and Classification



bad responder

good responder

Bioinformatics Software

- **R** is a widely used open source language and environment for statistical computing and graphics
  - Software and documentation are available from `http://www.r-project.org`
- **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data
  - Software and documentation are available from `http://www.bioconductor.org`

Conclusion

Conclusion

- The microarray technology is very promising in breast cancer.
- Unfortunately, literature have been inflated by over-optimistic results.
- Bioinformatics analysis have to be performed very carefully.
- Microarray Technology :
  - AFFYMETRIX© is a widespread technology and used in a large number of studies.
- Bioinformatics software :
  - R and Bioconductor are a gold mine to analyze AFFYMETRIX© data.

Links

- Course web page :
  `http://www.ulb.ac.be/di/map/gbonte/Info079.html`.
- Personal homepage :
  `http://www.ulb.ac.be/di/map/bhaibeka/`.
- Microarray Unit :
  `http://www.bordet.be/servmed/array/`.
- Machine Learning Group :
  `http://www.ulb.ac.be/di/mlg`.
- DEA/DES in Bioinformatics :
  `http://www.bioinfomaster.ulb.ac.be/`.

Thanks for your attention

Benjamin Haibe-Kains

Appendix

Microarray Unit

Project - TransBIG

- Joint project with Microarray Unit headed by Dr. Sotiriou.
- Motivations :
  - current risk evaluation of early breast tumors (see St Galln, NIH and NPI) fails to classify correctly the tumors. It results :
    - unnecessary therapies
    - toxic side effects
    - waste of money
  - the goal of the data mining analysis is to identify those patients at higher risk of distant metastases appearance on the basis of their genetic profile.

Project - TAMOXIFEN©

- Joint project with Microarray Unit headed by Dr. Sotiriou.
- Motivations :
  - majority of early-stage breast cancers express estrogen receptors (ER) and receive TAMOXIFEN© in the adjuvant setting.
  - 40% of these patients will relapse and develop incurable metastatic disease.
  - the goal of the data mining analysis is to identify those patients at higher risk of TAMOXIFEN© resistance on the basis of their genetic profile.

Members of Microarray Unit

Preprocessing Methods

We will focus on **preprocessing** methods for AFFYMETRIX© data

- image analysis : get raw probe intensities from chip image
- expression quantification : get gene expressions from raw probe intensities
- normalization : remove systematic bias to compare gene expressions

It exists several methods but we will focus on **Robust Multi-array Analysis** (RMA) methods [Irizarry et al., 2003]

Image Analysis

Main software from AFFYMETRIX© was MicroArray Suite 5 (MAS5), now called GeneChip Operating Software 1.2 (GCOS)

- **DAT** file is the image file

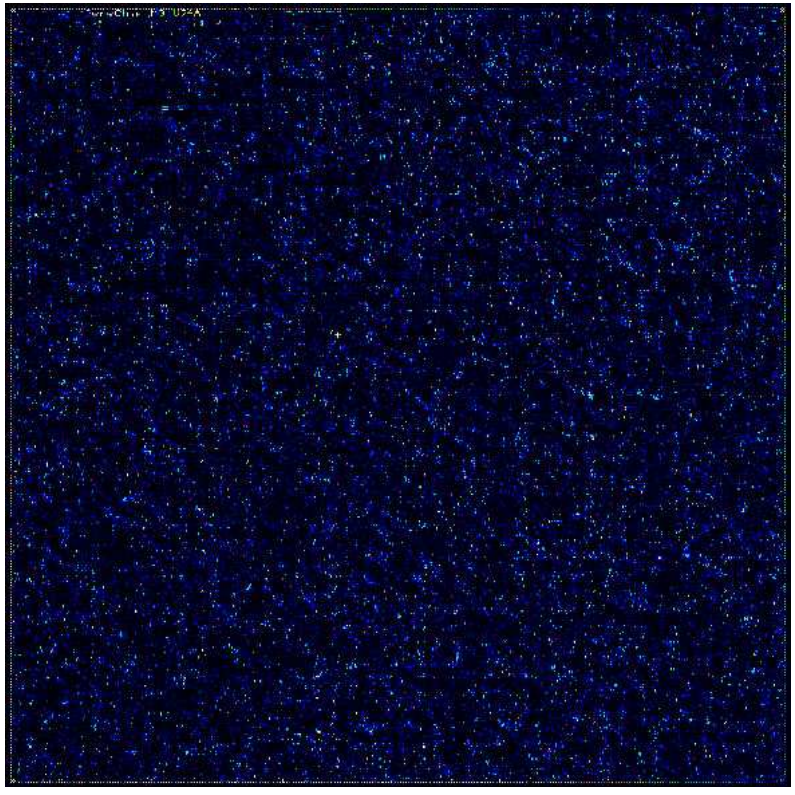
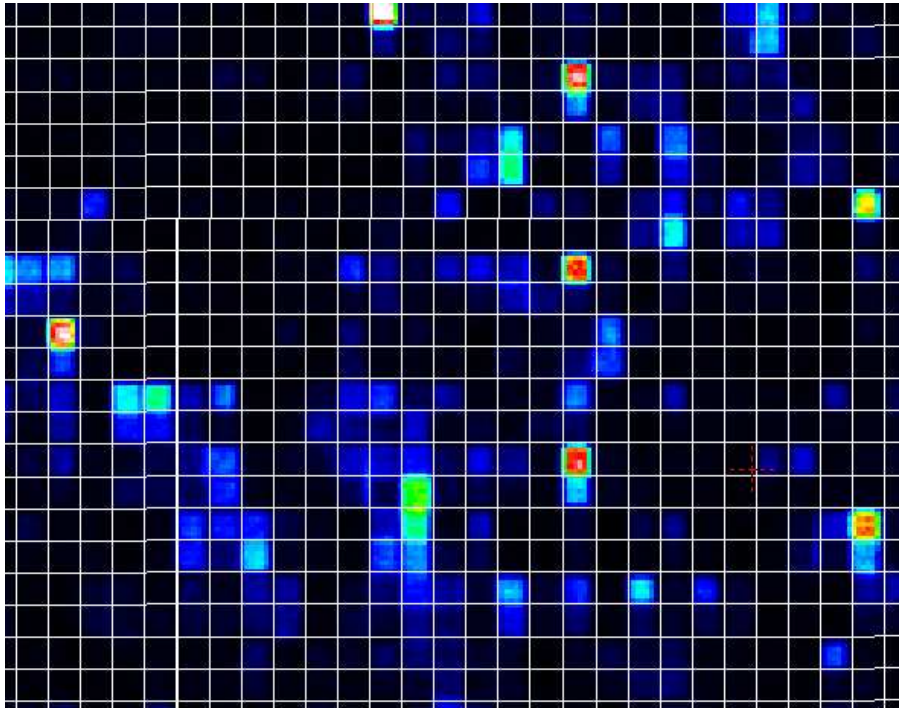


Image Analysis(2)



Each probe cell is composed by 10x10 pixels

- remove outer 36 pixels $\rightarrow$ 8x8 pixels
- probe cell signal, PM or MM, is the 75th percentile of the 8x8 pixel values.

Image Analysis(3)

- **CEL** file is the CELl intensity file including probe level PM and MM values

The last file provided by AFFYMETRIX© concerns the annotations of the chip

- **CDF** file is the Chip Description File describing which probes go in which probe sets (genes, gene fragments, ESTs)

The bioconductor functions (especially from **affy** package) use the CEL files

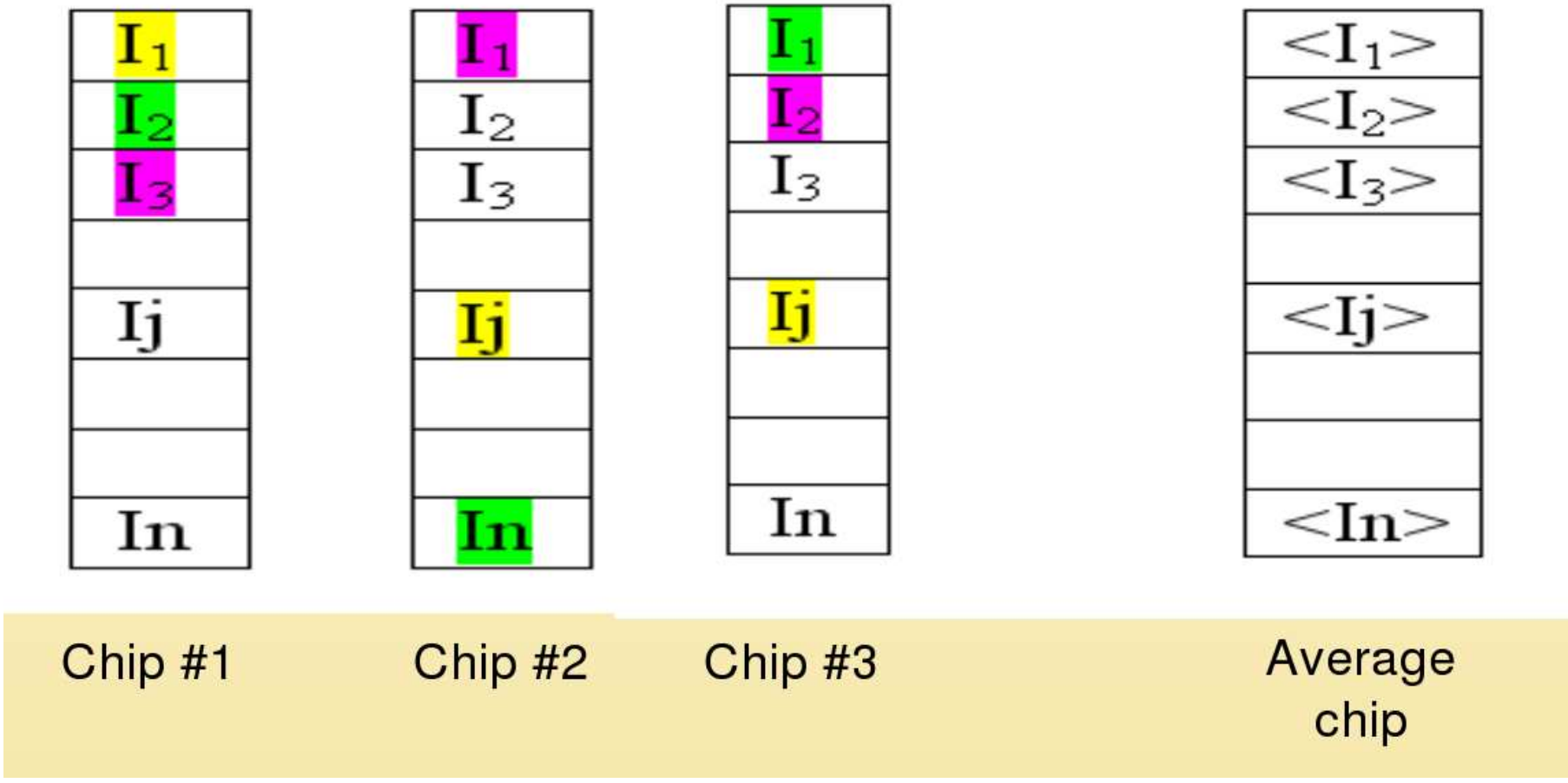
Expression Quantification

For each probe set, **summarization** of the probe level data (11-20 PM and MM pairs) into a single expression measure

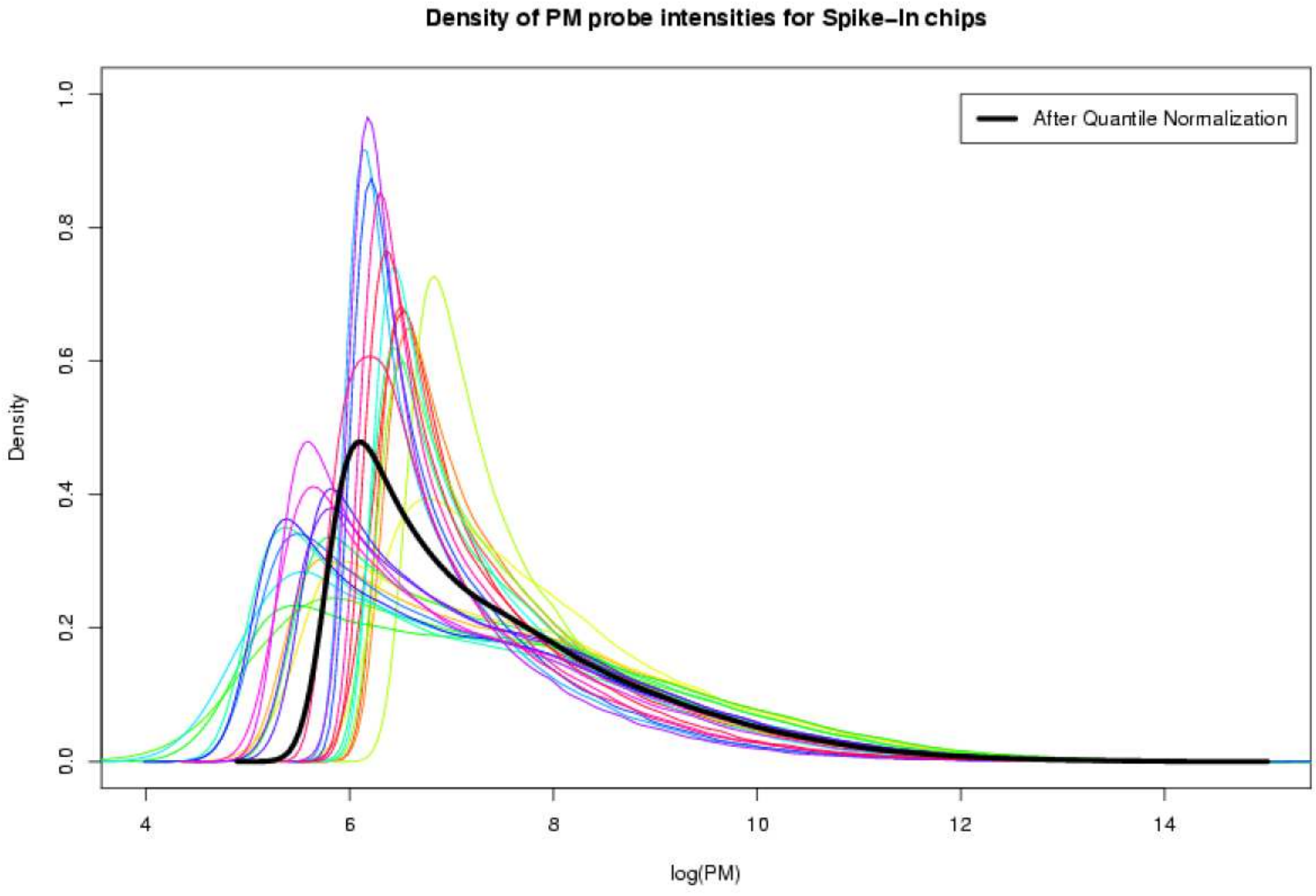
RMA procedure

- use only PM and ignore MM
- adjust for background on the raw intensity scale
- carry out **quantile** normalization [Bolstad et al., 2003] of $PM - \hat{BG}$ and call the result $n(PM - \hat{BG})$
- Take log2 of normalized background adjusted PM
- Carry out a **medianpolish** of the quantities $log_2 n(PM - \hat{BG})$

Quantile Normalization



Quantile Normalization(2)



References

[Bolstad et al., 2003] Bolstad, B. M., Irizarry, R. A., Astrand, M., and TP, T. S. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

[Irizarry et al., 2003] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15.

[jansen et al., 2005] jansen, M., Foekens, J. A., van Staveren, I. L., Dirkzwager-Kiel, M. M., Ritstier, K., Look, M. P., van Gelder, M. E. M., Sieuwerts, A. M., Portengen, H., Dorssers, L. C., Jlijn, J., and Berns, M. (2005). Molecular clasification of tamoxifen-resistant breast carcinomas by gene expression profiling. *Journal of Clinical Oncology*, 23(4):732–740.

[van't Veer et al., 2002] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhiven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536.