ULB
MLG

Analysis of Genomic and Proteomic Data

AFFYMETRIX© Technology and Preprocessing Methods

Benjamin Haibe-Kains

bhaibeka@ulb.ac.be

Université Libre de Bruxelles

Institut Jules Bordet

Table of Contents

AFFYMETRIX© Technology

Microarray Technology

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

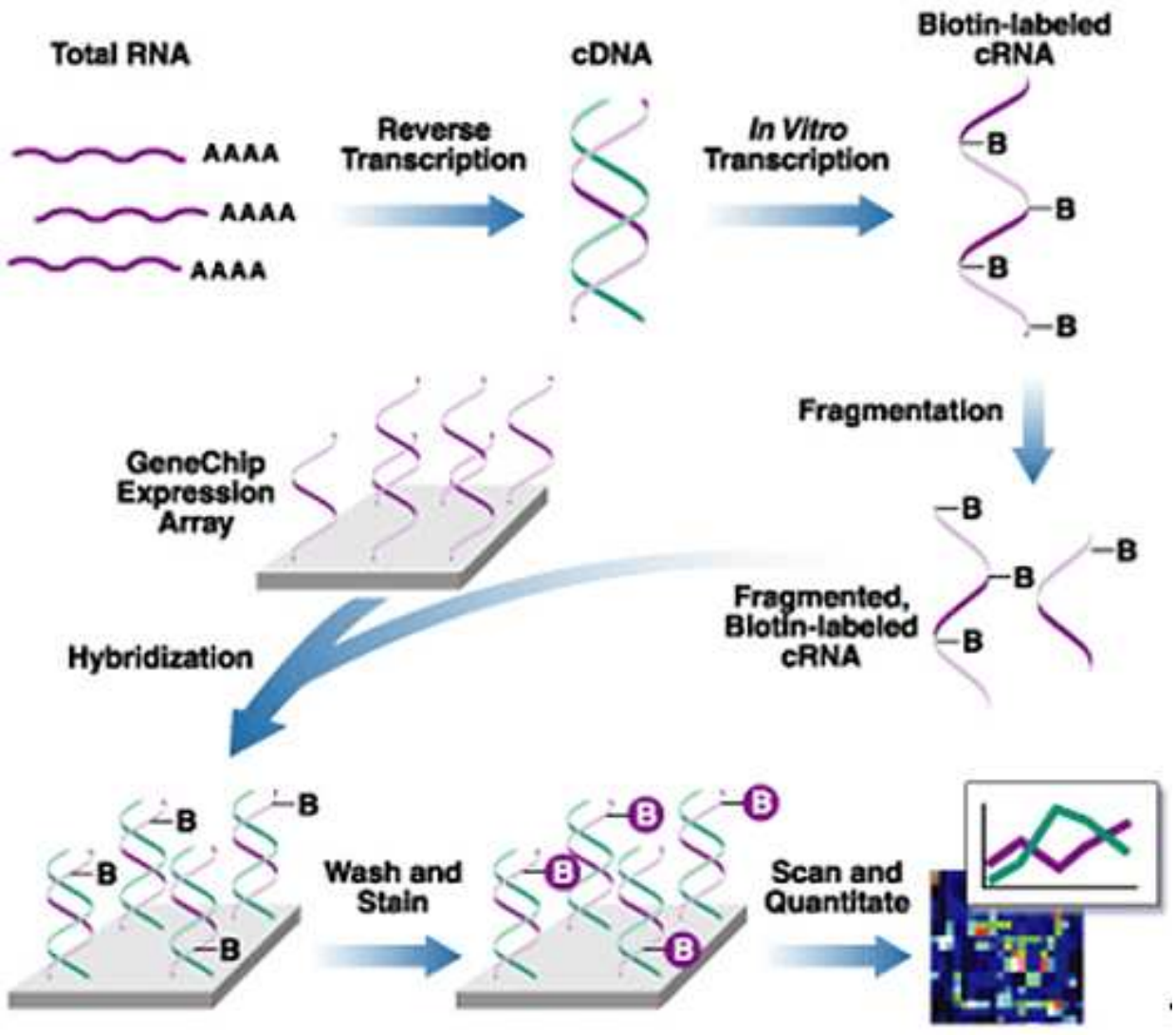
Microarray Technology(2)

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

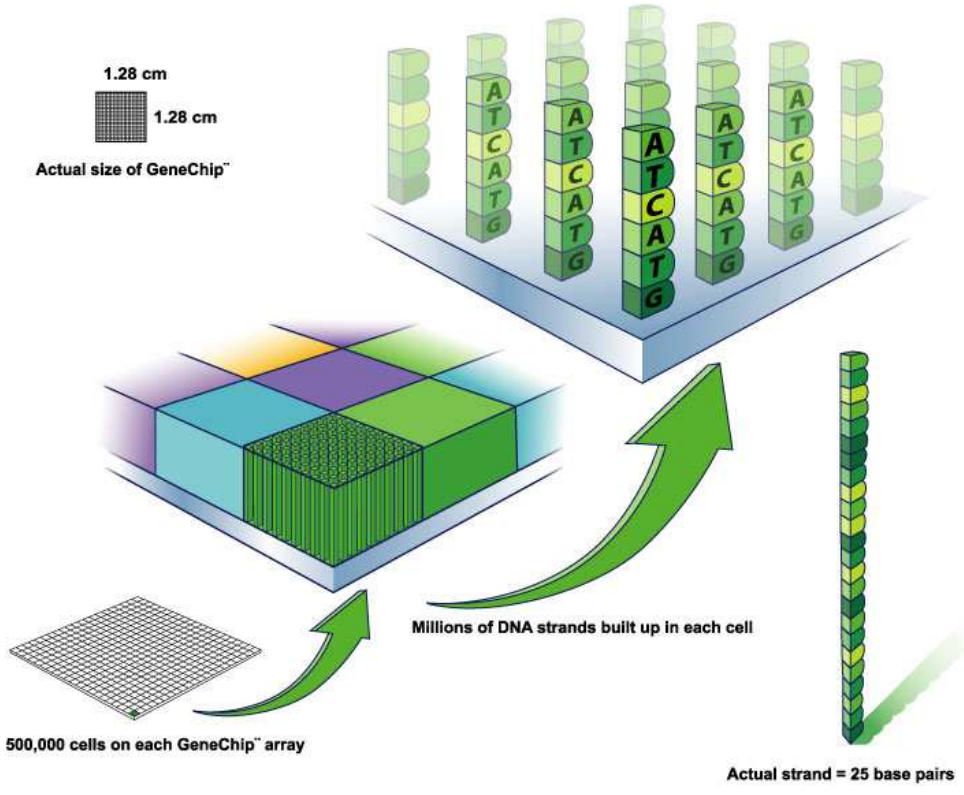


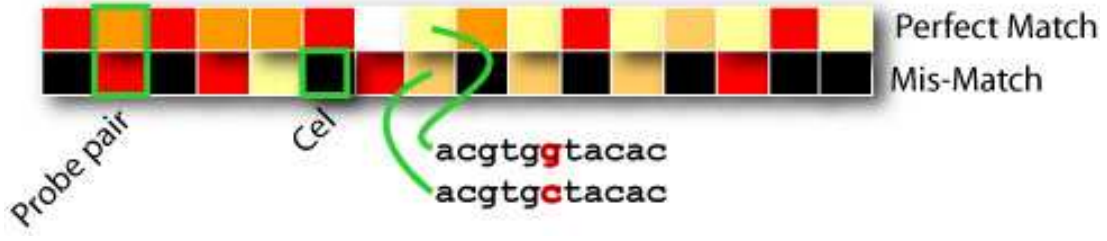
AFFY© GeneChip Structure

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

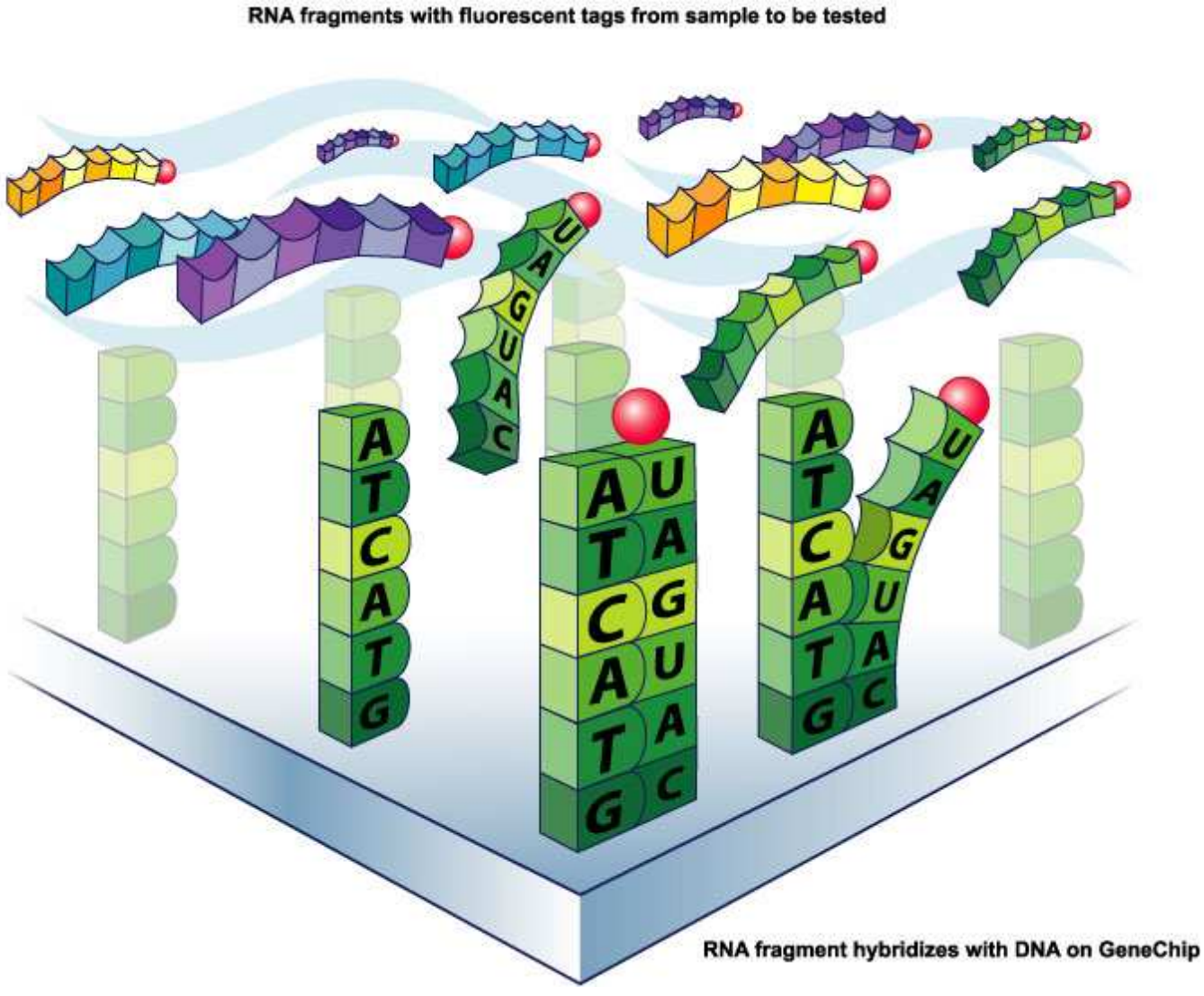
To facilitate the explanation, we assume that 1 gene is represented by 1 probe set including 20 probe pairs (PM and MM)

AFFY© GeneChip Structure(2)

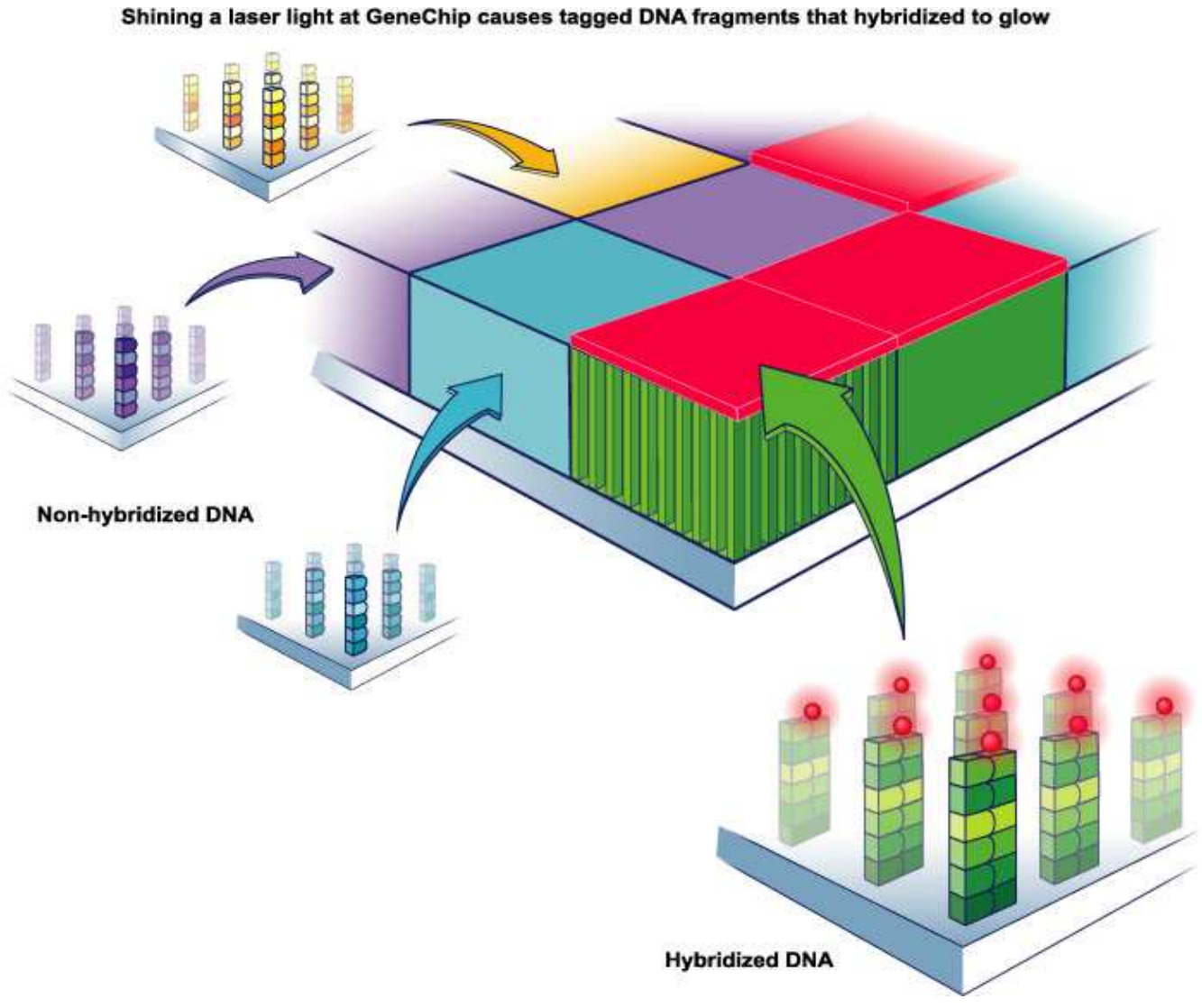




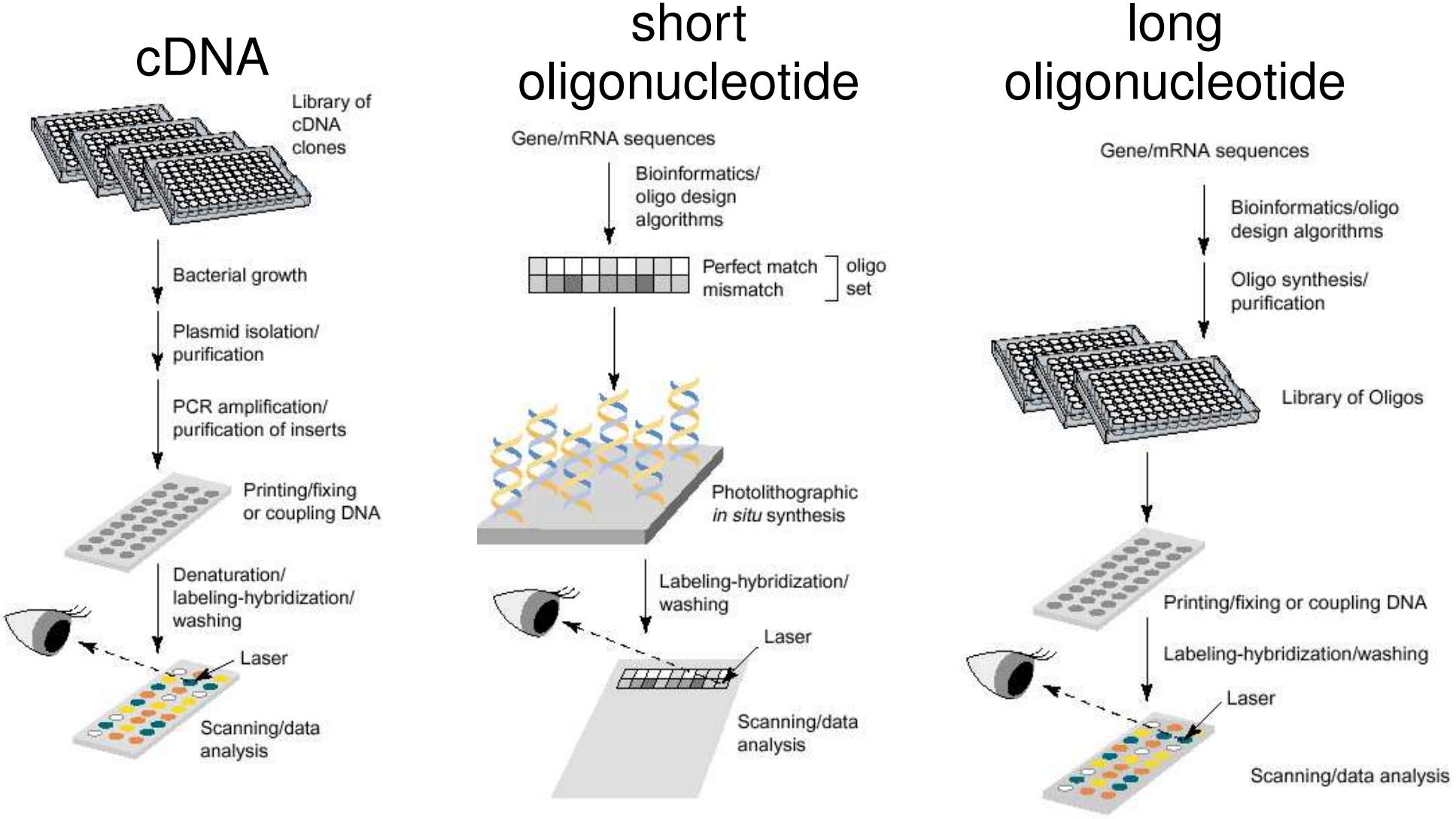
AFFYMETRIX© Hybridization



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 :

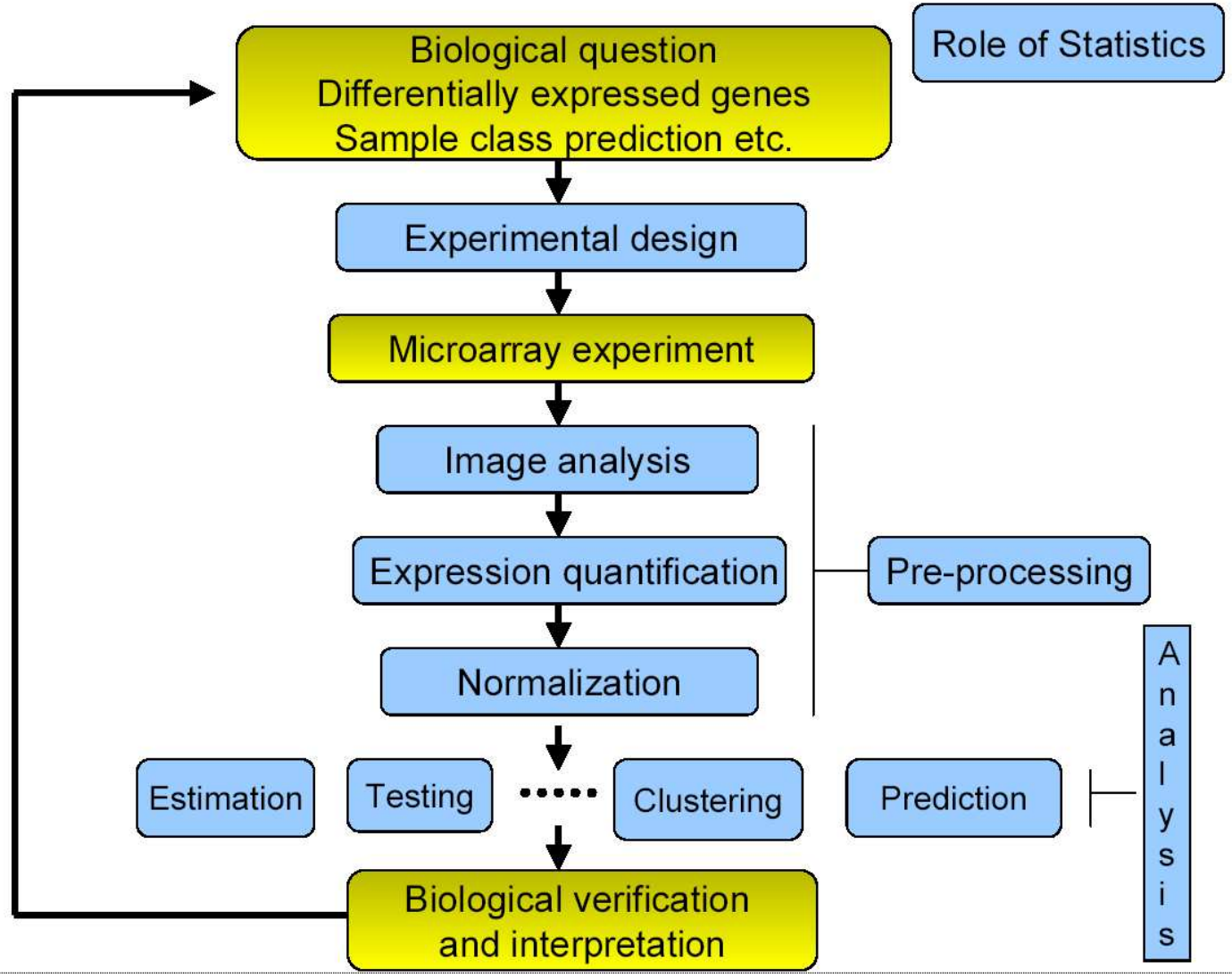
- cost of the devices and the chips (but easy use)
- changes in probe design is hard (but new program permits to create his own design)
- short oligos $\longrightarrow$ several oligos per gene, specificity/sensitivity trade-off (complex methods to get gene expression)

Preprocessing Methods

R and Bioconductor

- **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`

Microarray Analysis Design



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

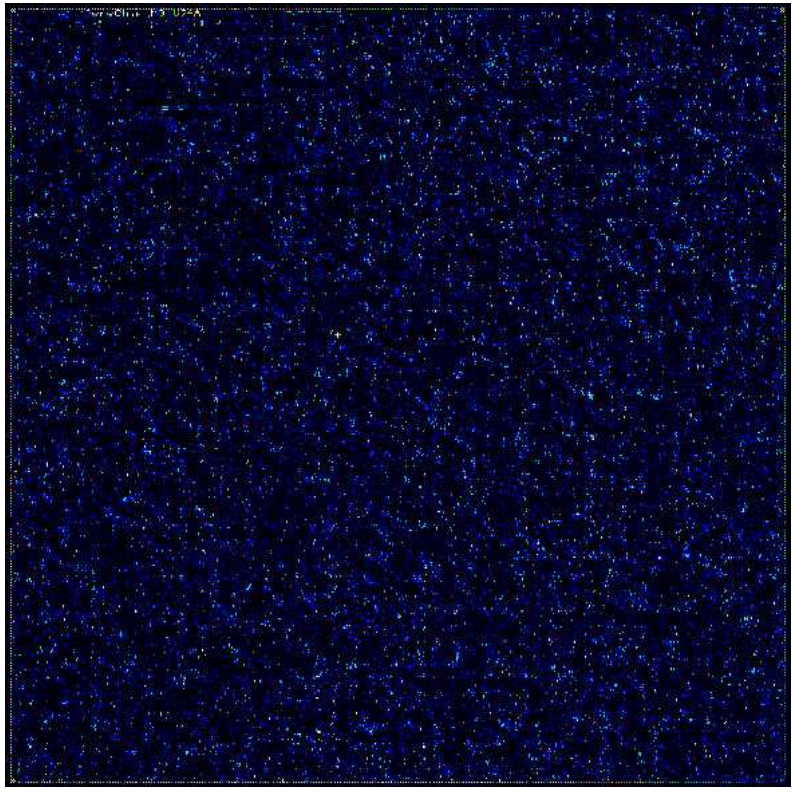
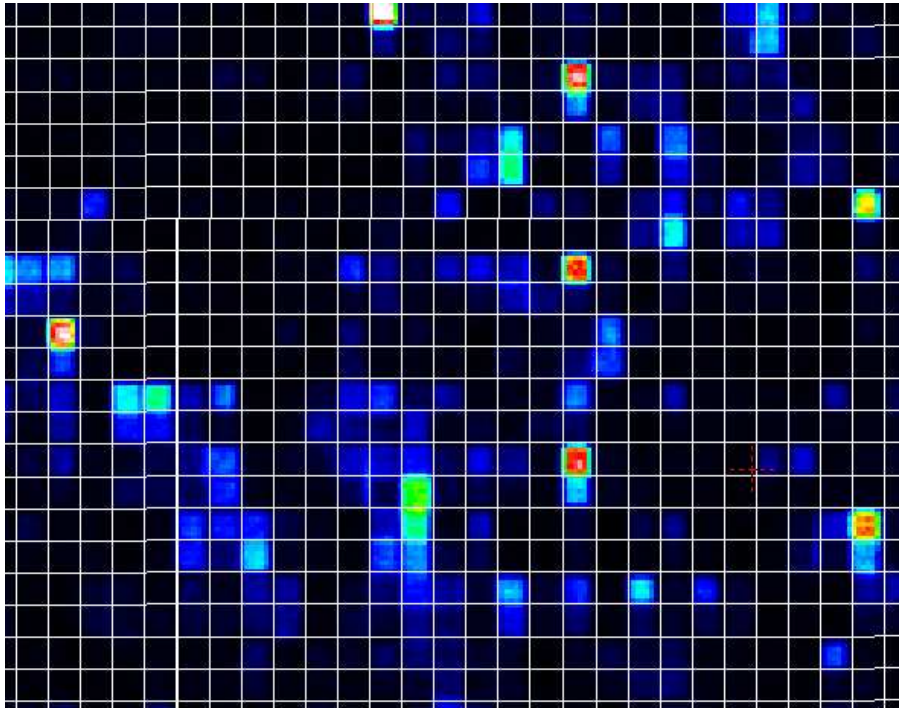


Image Analysis(2)



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 CELI 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

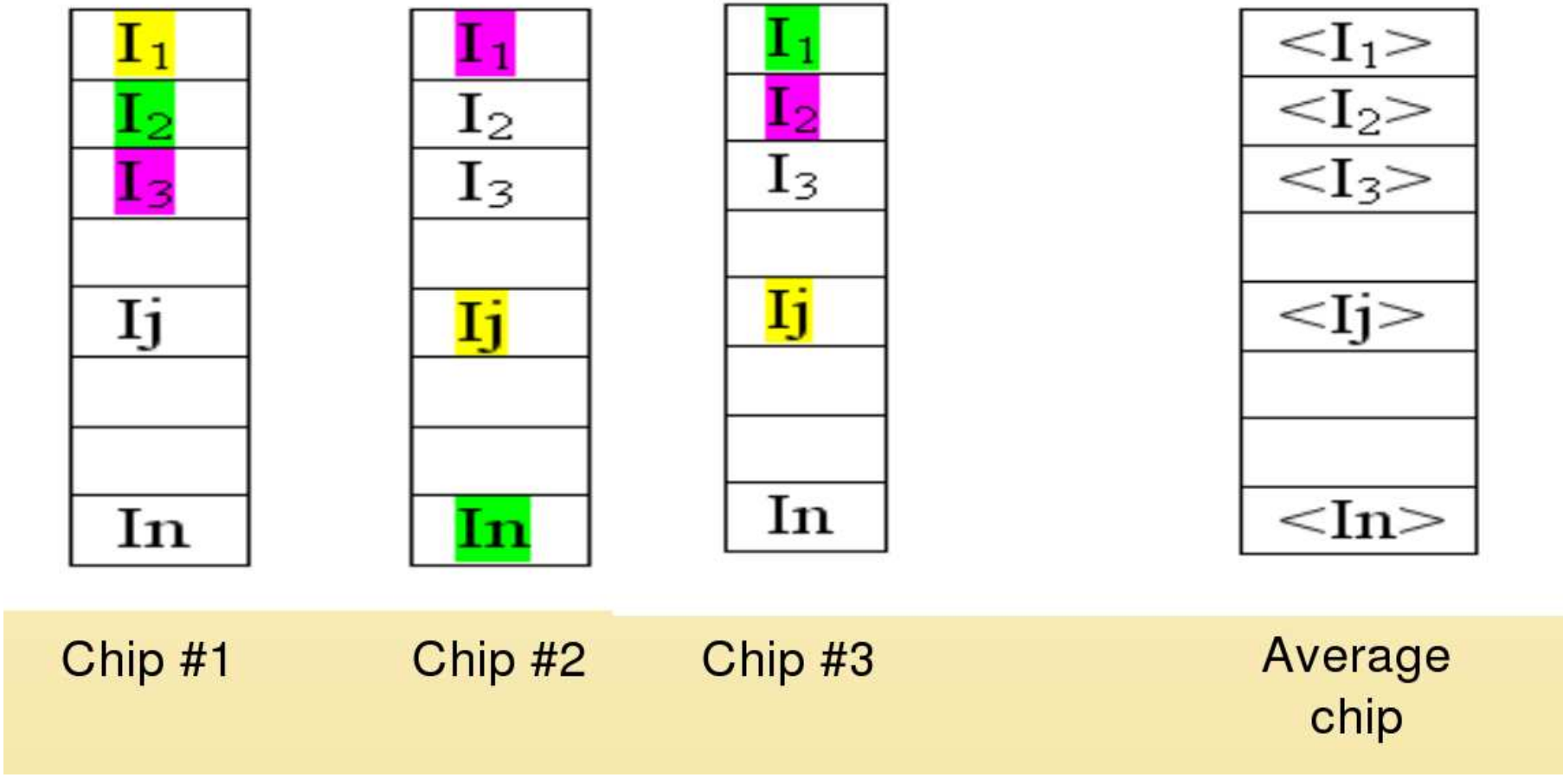
Expression Quantification

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

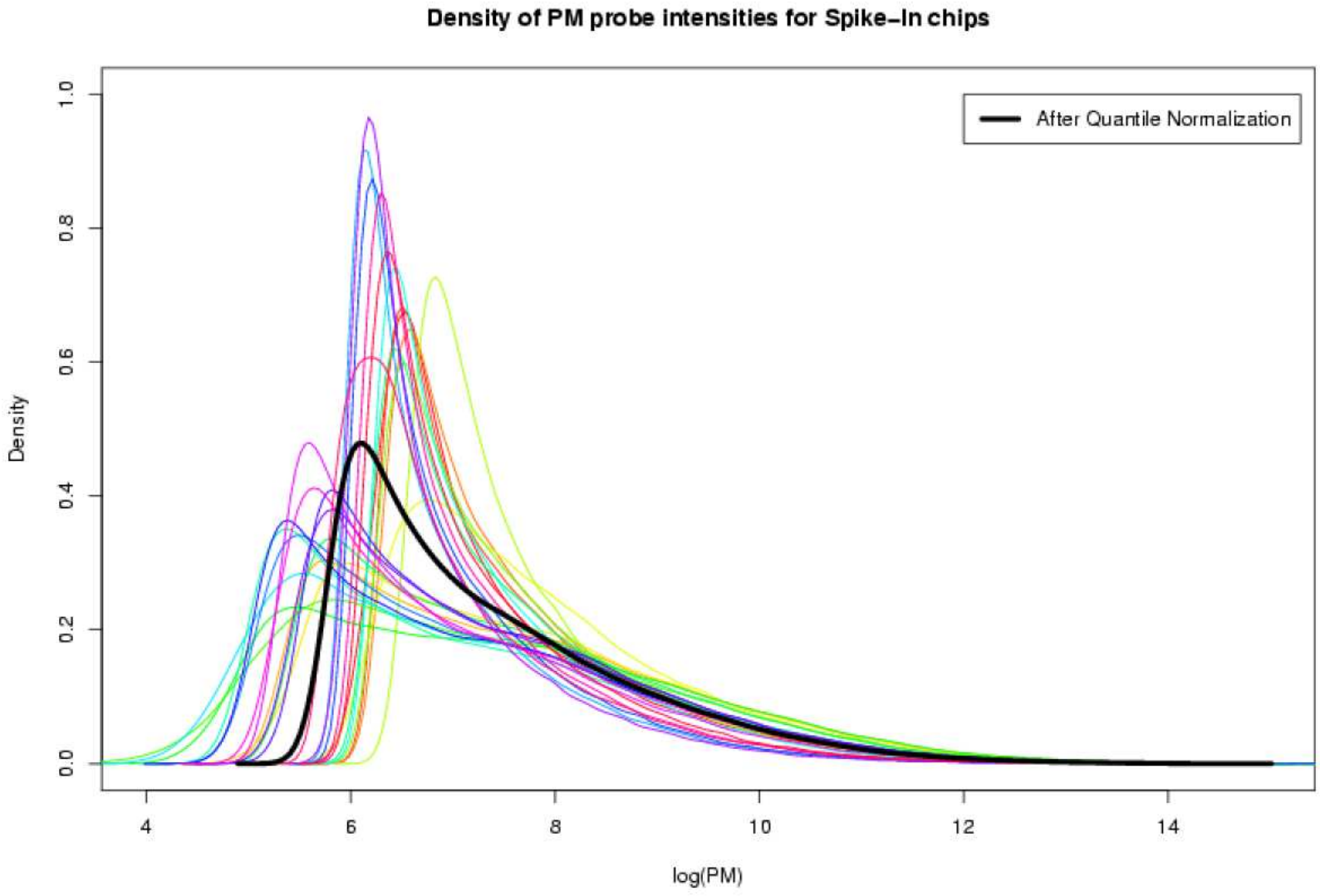
RMA procedure

- 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 $log_2 n(PM - \hat{BG})$

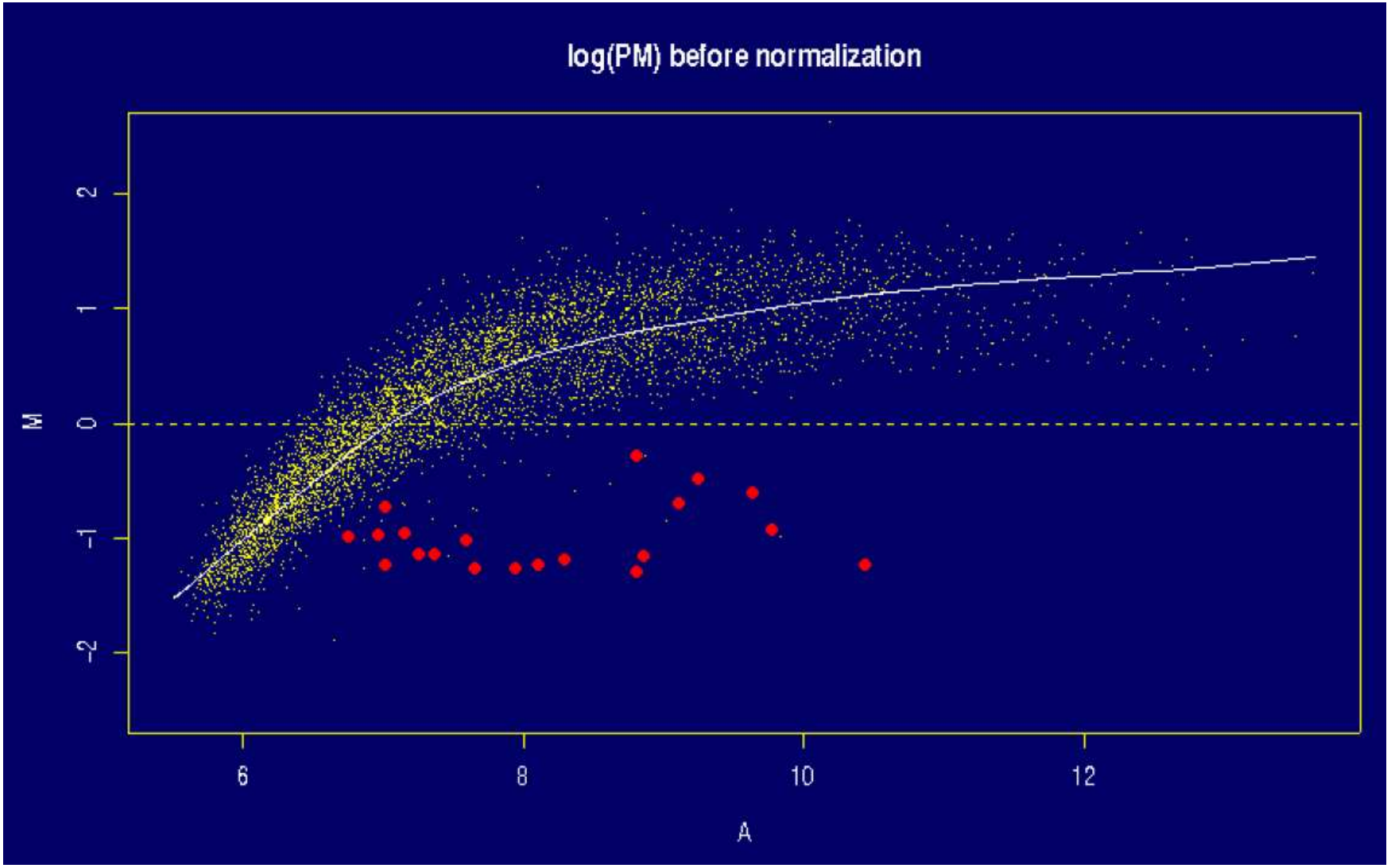
Quantile Normalization



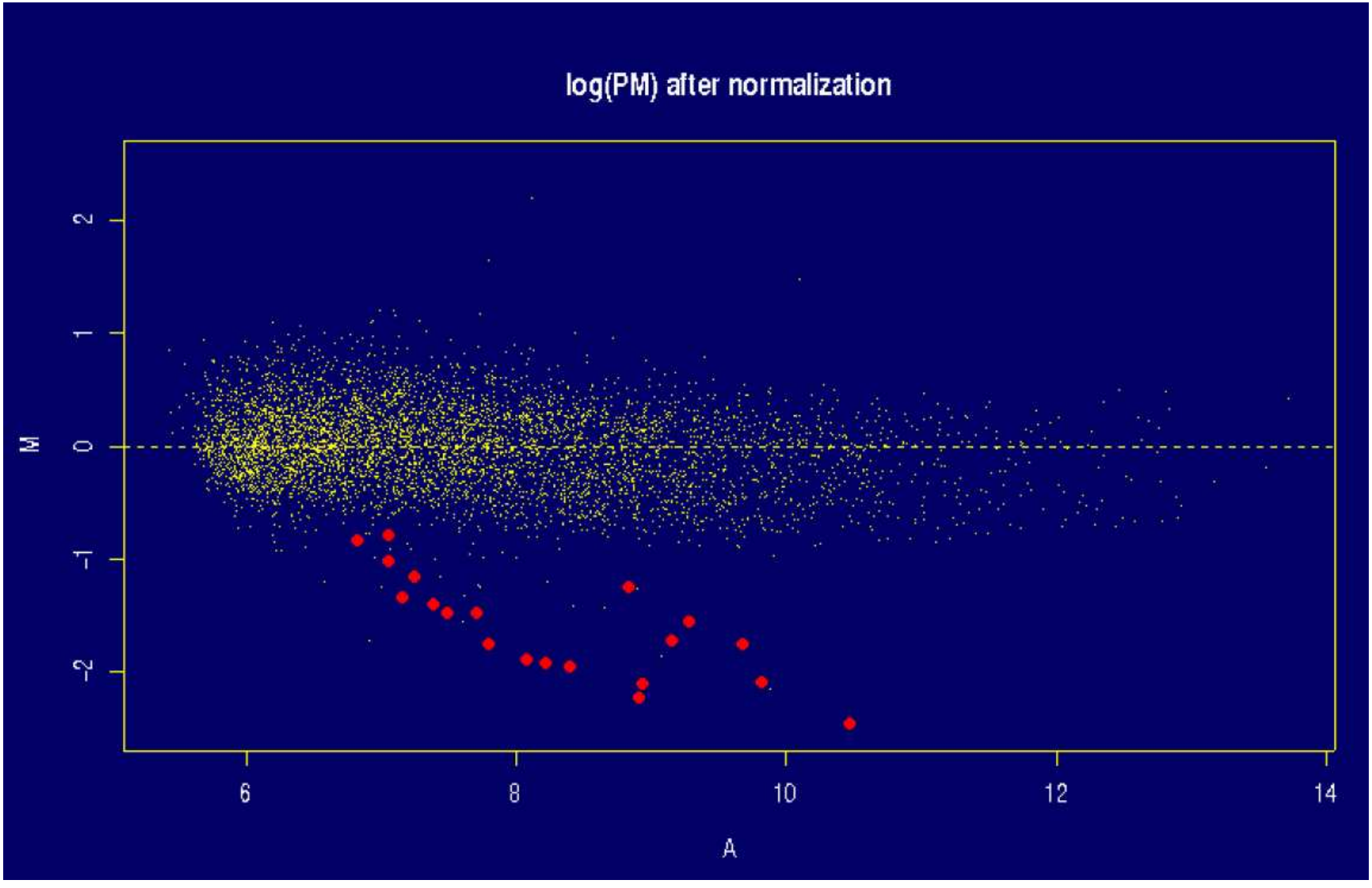
Quantile Normalization(2)



Quantile Normalization(3)



Quantile Normalization(4)



Bioconductor Functions

Most of the functions are in the **affy** package

```
> library(affy) #load the library
> library(help=affy) #help about the library
contents
```

All the CEL files have to be read in an **AffyBatch** object

Bioconductor Functions(2)

AffyBatch structure (see `?AffyBatch`)

- cdfName : object of class *character* representing the name of CDF file associated with the arrays in the *AffyBatch* (e.g. hgu133plus2)
- exprs : object of class *matrix* inherited from *exprSet*. The matrix contains one probe per row and one chip per column
- phenoData : object of class *phenoData* inherited from *exprSet*
- annotation : object of class *character* identifying the annotation that may be used for the chips
- description : object of class *MIAME* (Minimal Information About Microarry Experiment)

Bioconductor Functions(3)

AffyBatch creation (see `?read.affybatch`)

```
> abatch <- read.affybatch(filenames, phenoData,
description, verbose=TRUE)
```

Remarks

- the AffyBatch class is an extension of the exprSet class
- *filenames* is an object of class *character* containing the whole paths to CEL files
- all the CEL files have to come from the same chip (e.g. hgu133plus2)

Bioconductor Functions(4)

RMA function (see `?rma`)

```
> eset <- rma(abatch, verbose=TRUE,
normalize=TRUE, background=TRUE)
```

Remarks

- *rma* function for background correction
- *quantile* function for normalization
- *pmonly* function for probe specific correction
- *medianpolish* function for summarization

Bioconductor Functions(5)

Due to memory limitations, **just.rma** function can be used without create an AffyBatch object (see `?just.rma`)

```
> eset <- just.rma(filenames, phenoData,
description, verbose=TRUE, background=TRUE,
normalize=TRUE)
```

Conclusion

- AFFYMETRIX© is a widespread technology and used in a large number of studies
- R and Bioconductor are a gold mine to analyze AFFYMETRIX© data
- Preprocessing methods have an important impact on the high-level analyses
- RMA methods seem very efficient (see bias/variance studies with spike-in and dilution experiments [Bolstad et al., 2003])

Conclusion(2)

- Take care about memory consumption ... Need to rewrite some methods (C language, parallel programming) ?
  - I propose a project about the parallelization of the RMA methods in order to manage large datasets (current limit is about 250 chips)
  - the project description is available from the course web page or my homepage
- The slides of this presentation are available from the course web page or my homepage
- A practical course will be given february 4, 2005 at 10 to 12h

Links

- Course web page :
  `http://www.ulb.ac.be/di/map/gbonte/DEA_appli.html`

- Personal homepage :
  `http://www.ulb.ac.be/di/map/bhaibeka/`

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.