Microarray Data Analysis

C. Ambroise

Laboratoire Statistique et Génome
UMR CNRS 8071

Autumn 2010

Plan

Plan

Textbooks

- McLachlan, G.J , Do, K.A. , Ambroise, C. Analyzing Microarray Gene Expression Data, Wiley Series in Probability and Statistics, 2004.
- Speed T (ed). Statistical Analysis of Gene Expression Microarray Data. Chapman and Hall, New York, 2003.
- Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (eds). The Analysis of Gene Expression Data. Springer-Verlag, New York, 2003.
- Dalgaard P. Introductory Statistics with R. Springer-Verlag, New York, 2002.

Course Outline

Lecture 1 Preprocessing

- Introduction to microarray technologies
- Visualization: Descriptive Statistics, Principal Component Analysis, Multidimensional Scaling

Lecture 2 Testing

- Methods for selecting differentially expressed genes (Hypothesis testing)
- Multiple Testing

Lecture 3 Interpreting

- Clustering microarray data (Hierarchical, kmeans, mixture)
- Enrichment of the cluster via public database

Course objectives

Microarrays are important for the study of gene expression. This technology changes the way biologists approach problems and introduces new challenges for statisticians. The literature now contains more than 10,000 papers using microarrays; biologists should understand how the data is processed in order to evaluate these publications. Statisticians need to understand where the data comes from, in order analyze it appropriately.

- Understand how microarrays work and how they are analyzed.
- Evaluate the analysis of microarray data in a published paper.
- Perform some basic analyses of microarrays.

Evaluation

1. Home Assignement (reading papers and performing a small analysis in R)
2. Evaluation in January ()Multiple Choice, summarizing a reseach paper).

Plan

Principle of microarrays

Technology

A DNA microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides of a specific DNA sequence, known as probes (or reporters) that are used to hybridize a cDNA or cRNA sample (called target)

Different types of microarrays

genome	SNP, CGH
transcriptome	Gene expression pofiling, Tiling array ...
proteome	CHIP-Chip,..
epigenome	...

Regulation networks

Gene expression is regulated (up or down)

- By region (e.g., liver vs brain)
- By development (e.g., fetal vs. adult)
- In dynamic response to environmental signals
- In disease states
- By gene activity (e.g. mutant vs. wild)

Regulation networks

Gene expression is regulated (up or down)

- By region (e.g., liver vs brain)
- By development (e.g., fetal vs. adult)
- In dynamic response to environmental signals
- In disease states
- By gene activity (e.g. mutant vs. wild)

What is it useful for ?

Principle

- Microarrays allow to have a picture of the cell functionning at a given time
- The problem is the fuzzyness of the picture.....

It is useful for

- Functional genomic
- Medical and clinical diagnosis
- Drug discovery, targeting and monitoring
- ....

What is it useful for ?

Principle

- Microarrays allow to have a picture of the cell functionning at a given time
- The problem is the fuzzyness of the picture.....

It is useful for

- Functional genomic
- Medical and clinical diagnosis
- Drug discovery, targeting and monitoring
- ....

Why do statisticians are interested in?

Specificities

- Numerous applications
- Difficult problems:
  1. Many features,
  2. few observations,
  3. very noisy data

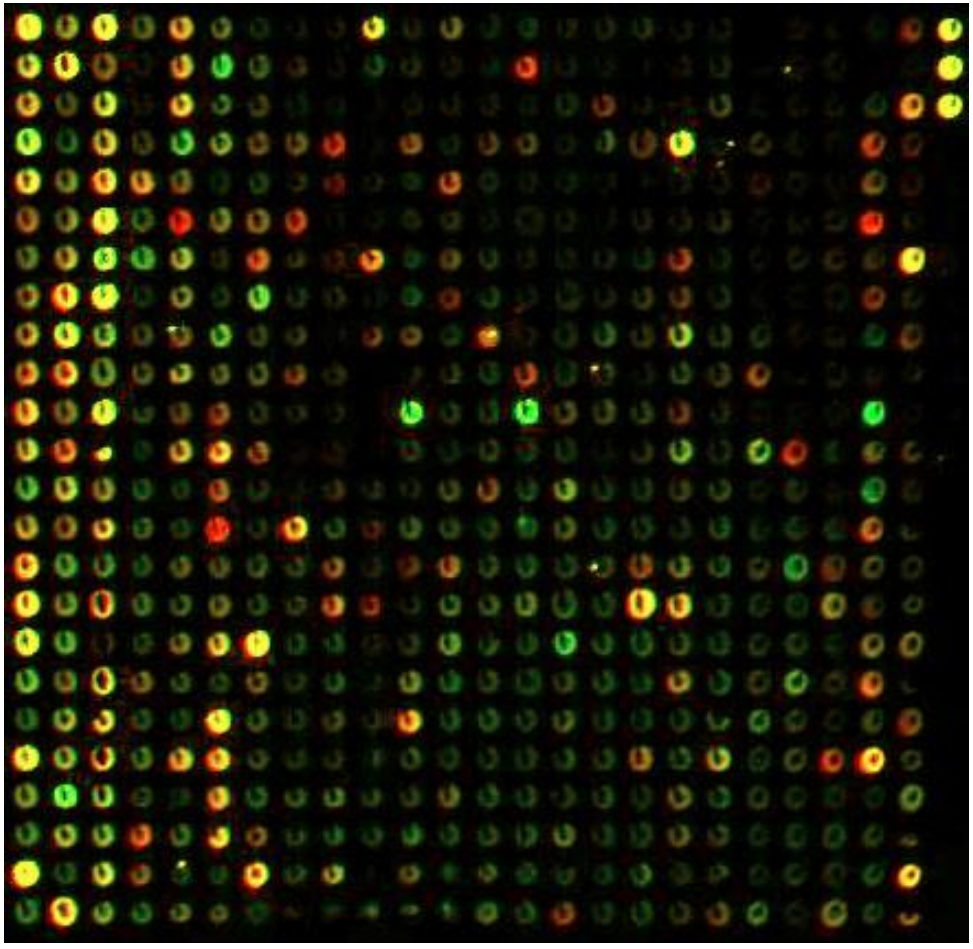
History

Gene expression Microarray platforms

- Spotted cDNA on nylon membranes (obsolete)
  - Commerically produced: Research Genetics, Clontech
  - Radioactive labeling, single channel
- Multiple synthesized short oligos (25-mers) on silicon
  - Commerically produced: Affymetrix
  - Single channel fluorescent labeling
  - Between 11 and 20 probes per gene target
- Spotted cDNA or long oligos (60- or 70-mers) on glass slides
  - Home-grown or commercial Two-channel: simultaneous
  - Co-hybridization of two samples
  - Two-color fluorescent labeling

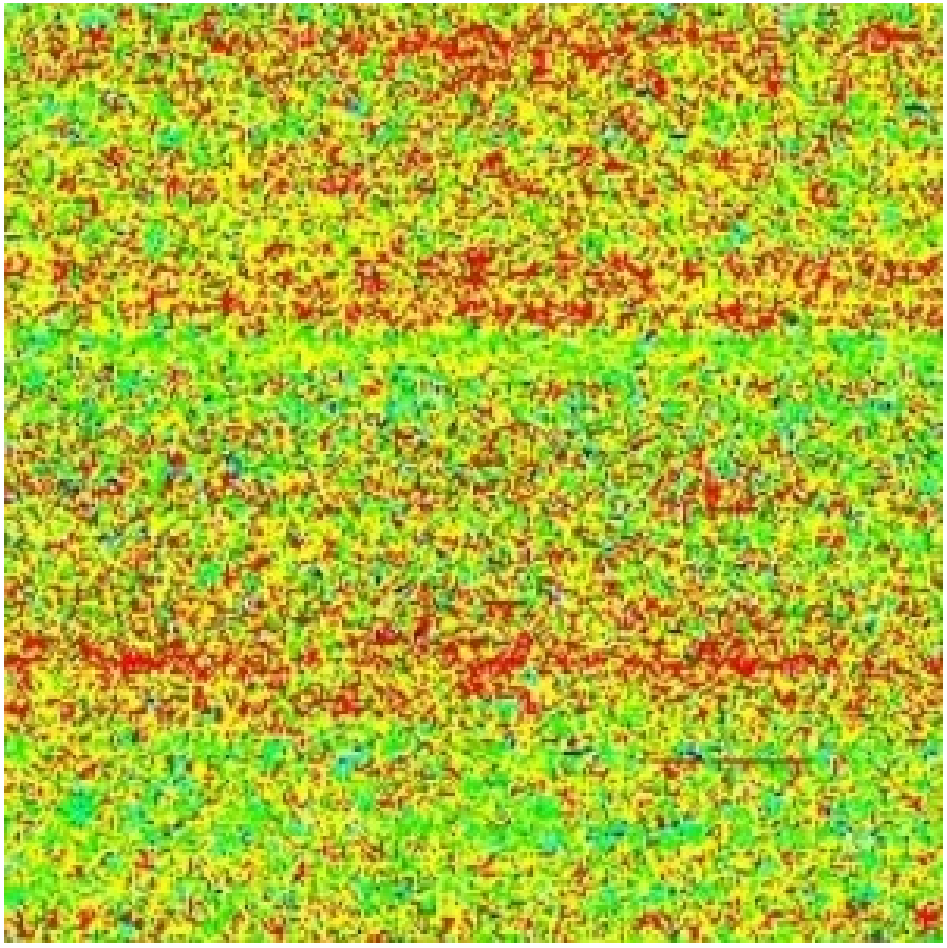
Glass slides

Glass slides (microarrays): fluorescent labeling comparing two conditions

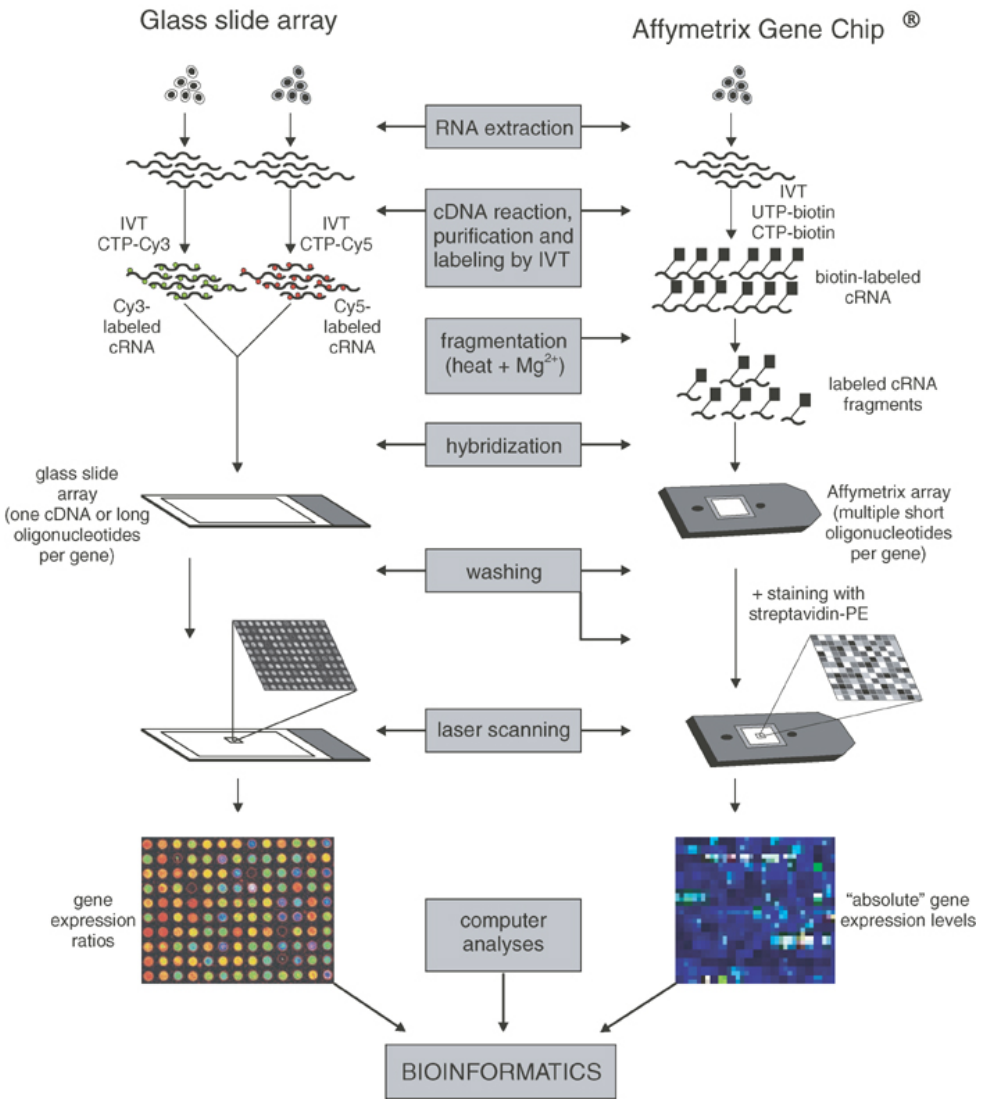


Affimetrix

oligo- nucleotides
20 couples
perfect match /
mismatch for
each gene
mismatch =
perfect match
with one mutated
nucleotide

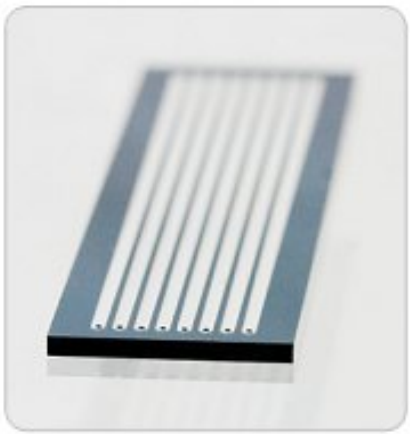


Two color Glass versus Affymetrix microarrays



Is sequencing the end of microarrays ?

FIGURE 1: ILLUMINA GENOME ANALYZER FLOW CELL



Up to eight samples can be loaded onto the flow cell for simultaneous analysis on the Illumina Genome Analyzer.

Illumina Genome Analyzer Commonly referred to as 'the Solexa', can be used to analyse gene expression:

- sequence expression data
- align sequences on a reference genome