Introduction to Microarrays

Aproximations to study the activity of genes

Presentation

Goals

- Learn about microarray technology
- Understand its possibilities and limitations
- Get familiar with how to use microarrays in biological experiments
- Know where to go for more

Content

- Introduction
- Production and use of microarrays
- From Images to expression matrices
- Microarray bioinformatics
  - —Software for the analysis of microarray data
  - —Annotations and annotations databases
  - —More microarrays databases.

Introduction

Some history

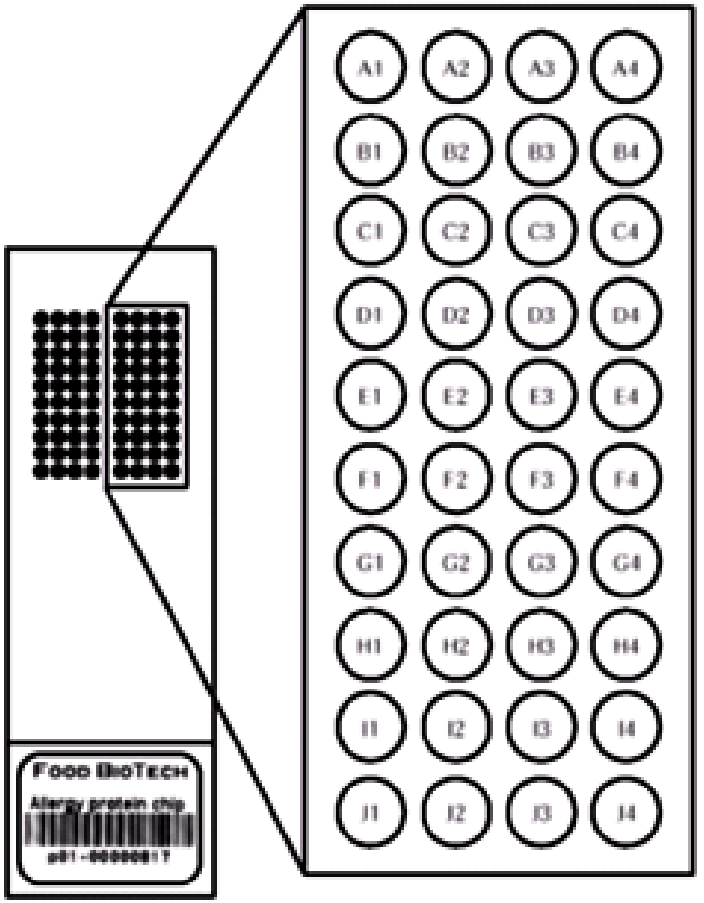
- Molecular biology has many techniques to measure RNA, DNA, proteins or metabolites.
  - Northern blot, differential display, SAGE
  - Southern blot: [similar to microarray]
- What characterizes the post genomic era is not what can be measured but the number of simultaneous measurements that can be performed.

The paradigm shift (J. Dopazo)

With the same resources
we obtain an image of lower resolution but wider scope

So, What is a microarray?

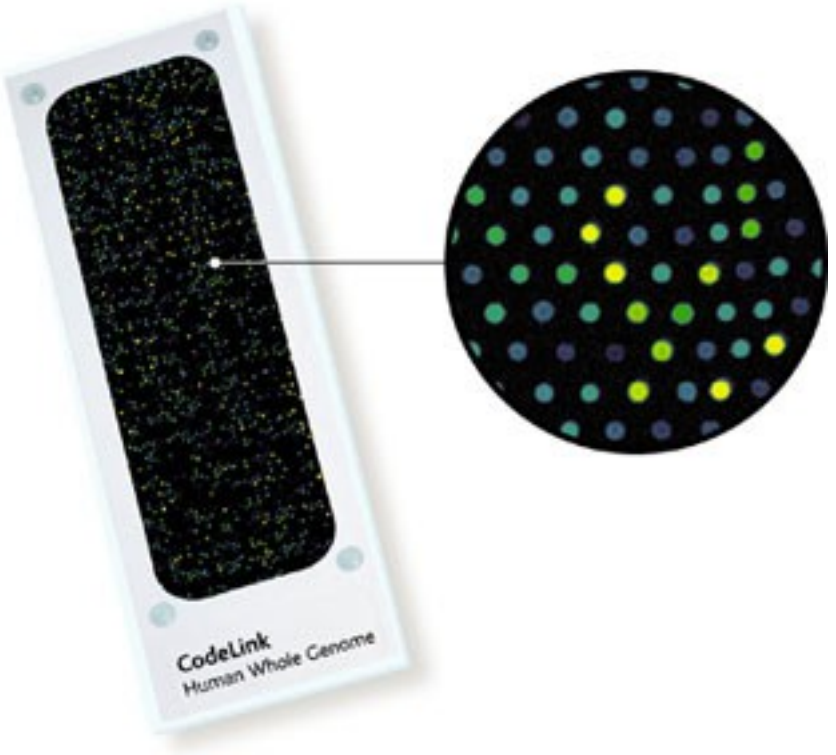
- An experimental format,
- based on the synthesis or attachment of probes, which represent genes (or proteins, or metabolites ...),
- on a solid substrate (glass, plastic, silica ...),
- Intended to be exposed to the target molecules (the sample).



How does it work?

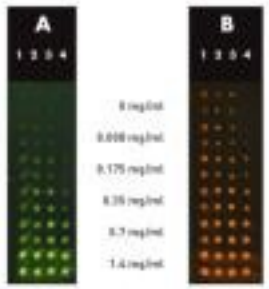
The level of hybridization between specific probes and target molecules is generally indicated by means of fluorescence and is measured by image analysis.

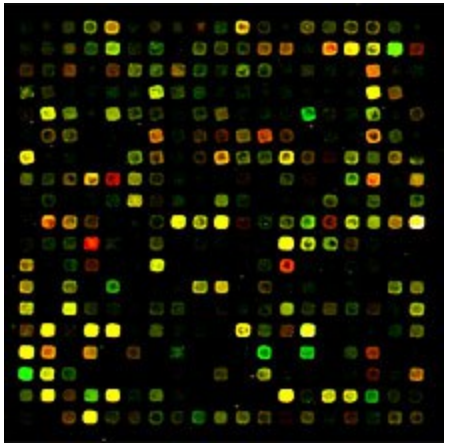
- The measure obtained indicates
  - the level of expression of the gene corresponding to the probe
  - in the test sample

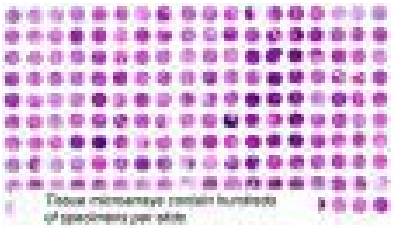


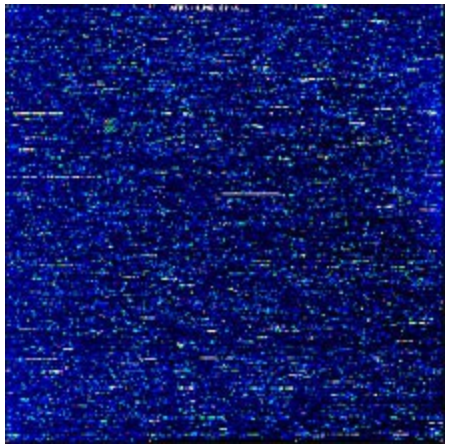
Types of microarrays

- Proteins
- Tissues
- DNA
  - CGH arrays
  - SNP arrays
- RNA (expression)
  - Two color (or cDNA)
  - e.g. Agilent
  - One color (or Affymetrix or oligonucleótidos)
  - GeneChip® Affymetrix
  - Illumina bead arrays









Microarray applications

- Study of genes that are differentially expressed between various conditions (Healthy / sick, mutant / wild, treated / untreated)
- Molecular classification of complex diseases
- Identify sets of genes characterizing a disease (signature)
- Predicting the response to treatment
- Detection of mutations and single polymorphisms (SNP)

But also

- Circadian clock analysis,
- Plant defence mechanisms,
- Environmental stress responses,
- Fruit ripening,

Fabrication and use

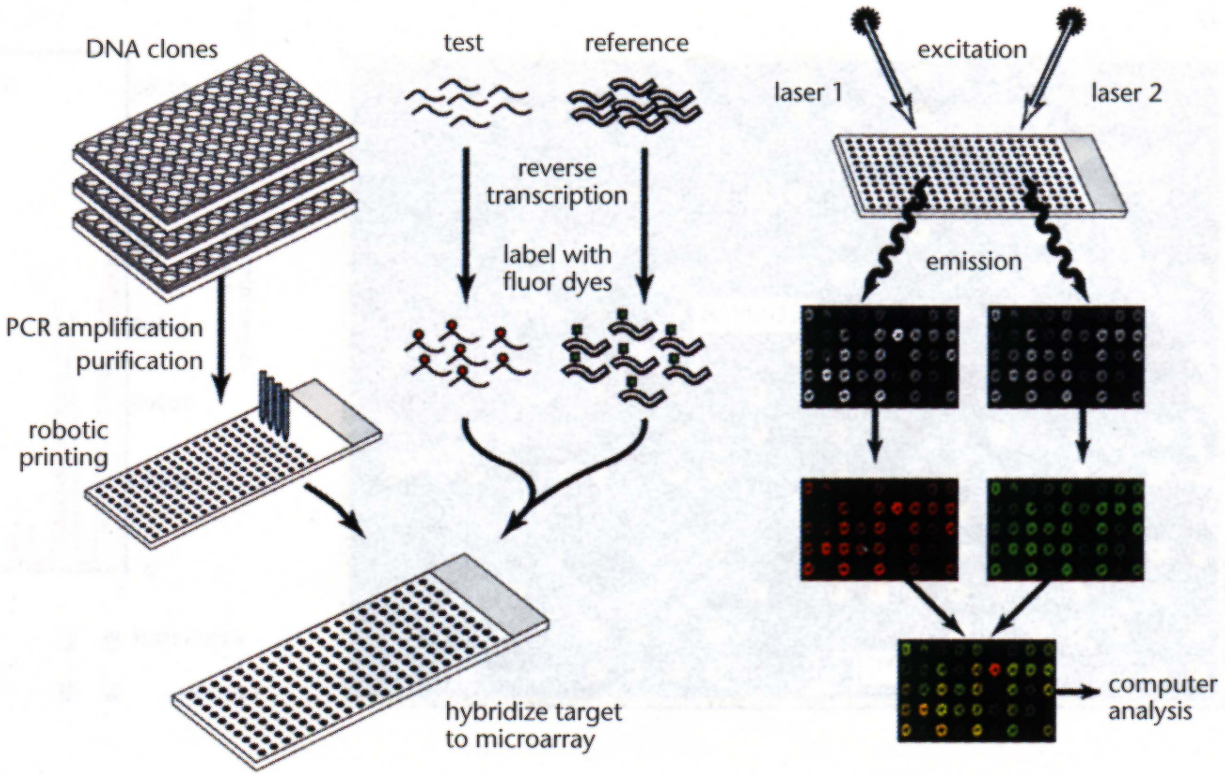
Expression microarrays

- There are many types of microarrays
- They rely on similar principles but the details of its operation change from one to another case
- Here we focus on expression arrays
  - 2-color arrays (spotted)
  - Oligonucleotide arrays (in situ synthesized).

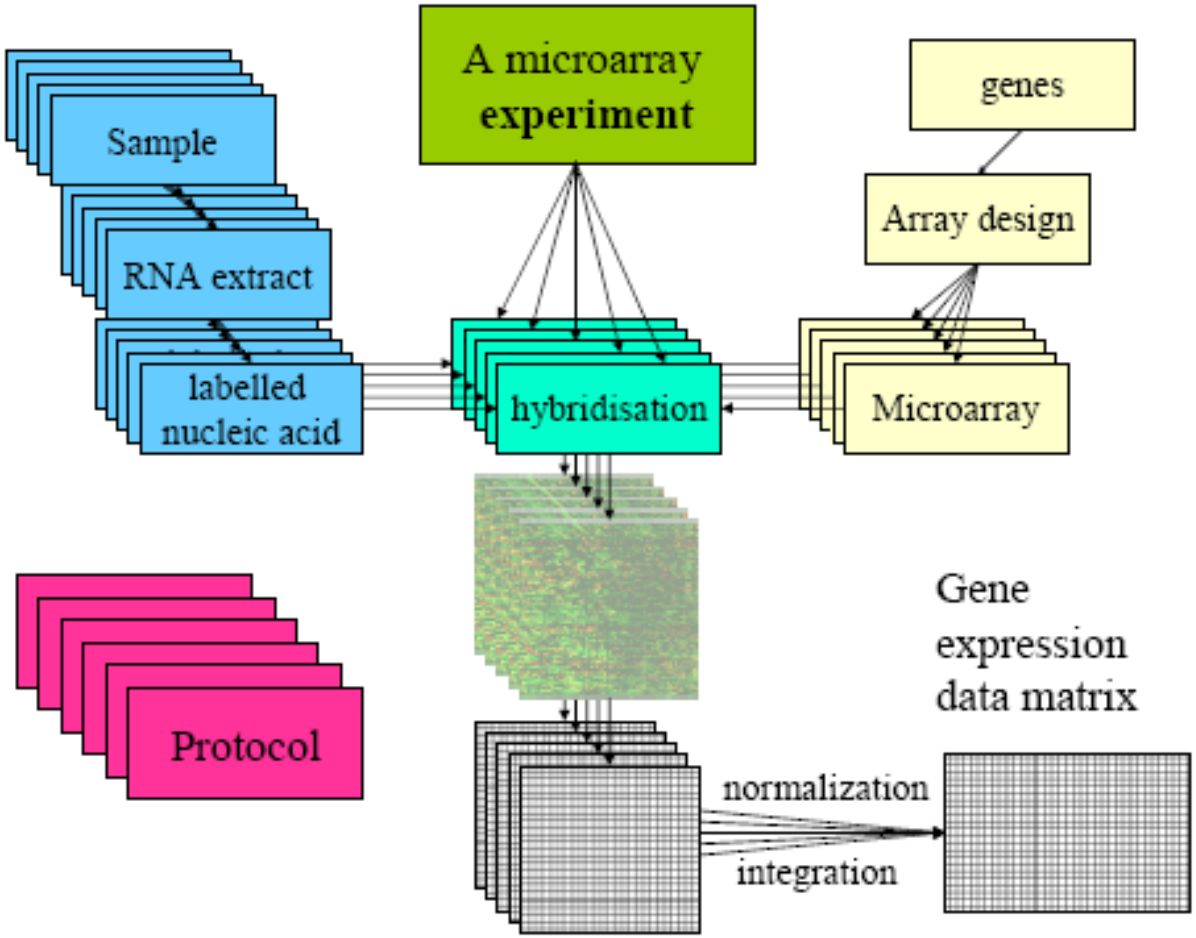
Two color microarrays (spotted)

- Chip design and production
- Sample Preparation
- Hybridization
- Scanning the chip
- Image analysis

General overview of the process



To visualize an animation go to:
http://www.bio.davidson.edu/courses/genomics/chip/chip.html


A microarray experiment
genes
Array design
Sample
RNA extract
labelled nucleic acid
hybridisation
Microarray
Protocol
Gene expression data matrix
normalization
integration

Oligo microarrays sinthesized *in situ*

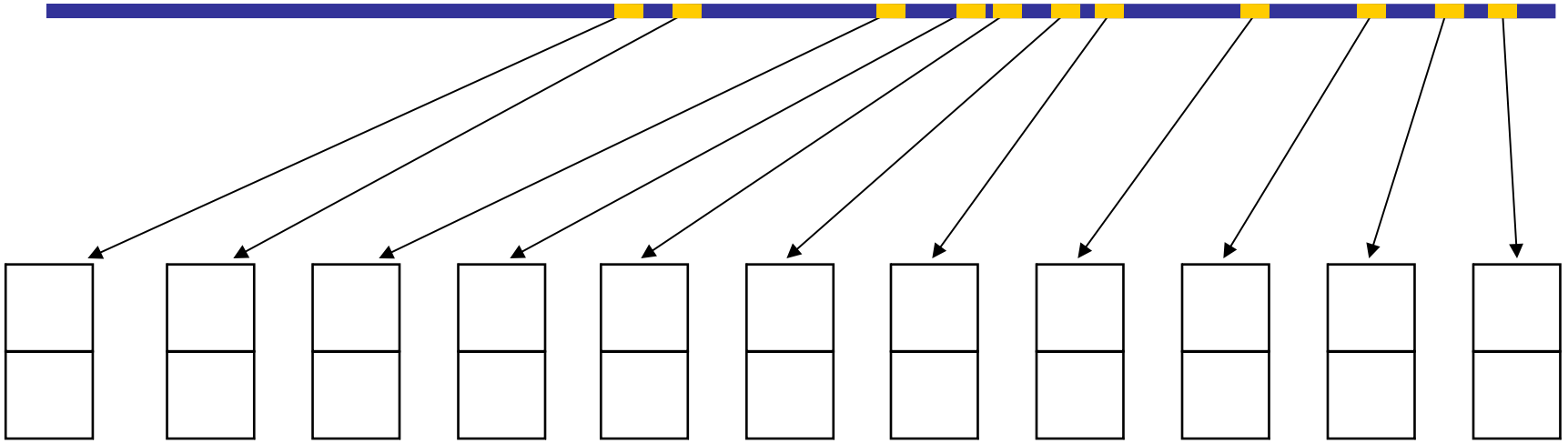
- More advanced design than 2 colors
- Rely on technologies developed for microelectronics
- Some distinctive features
  - —Not based on competitive hybridization: each chip containing samples from a single type
  - —Probes are synthesized directly on the chip instead of in vitro synthesized and then attached to slides
  - —Each gene is represented by a group of short probes rather than a single long probe

Probesets, probes, PM & MM

- A set of probes is used to measure mRNA level of a single gene.
- Each group (probeset) consists of multiple pairs of cells (probe cells)
  —with millions of copies of a 25bp oligo.
- Pairs consist of
  —a Perfect Match (PM) which coincides exactly with a portion of the gene
  —A Mismatch (MM) identical to PM except in the central nucleotide replaced with its complementary

1gene=1 probeset

Gene sequence



Distinct "Probe pairs" represent different parts of same gene

Probes are selected to be specific to the gene they represent and to have good hybridization properties

Oligo synthesis on-the-chip

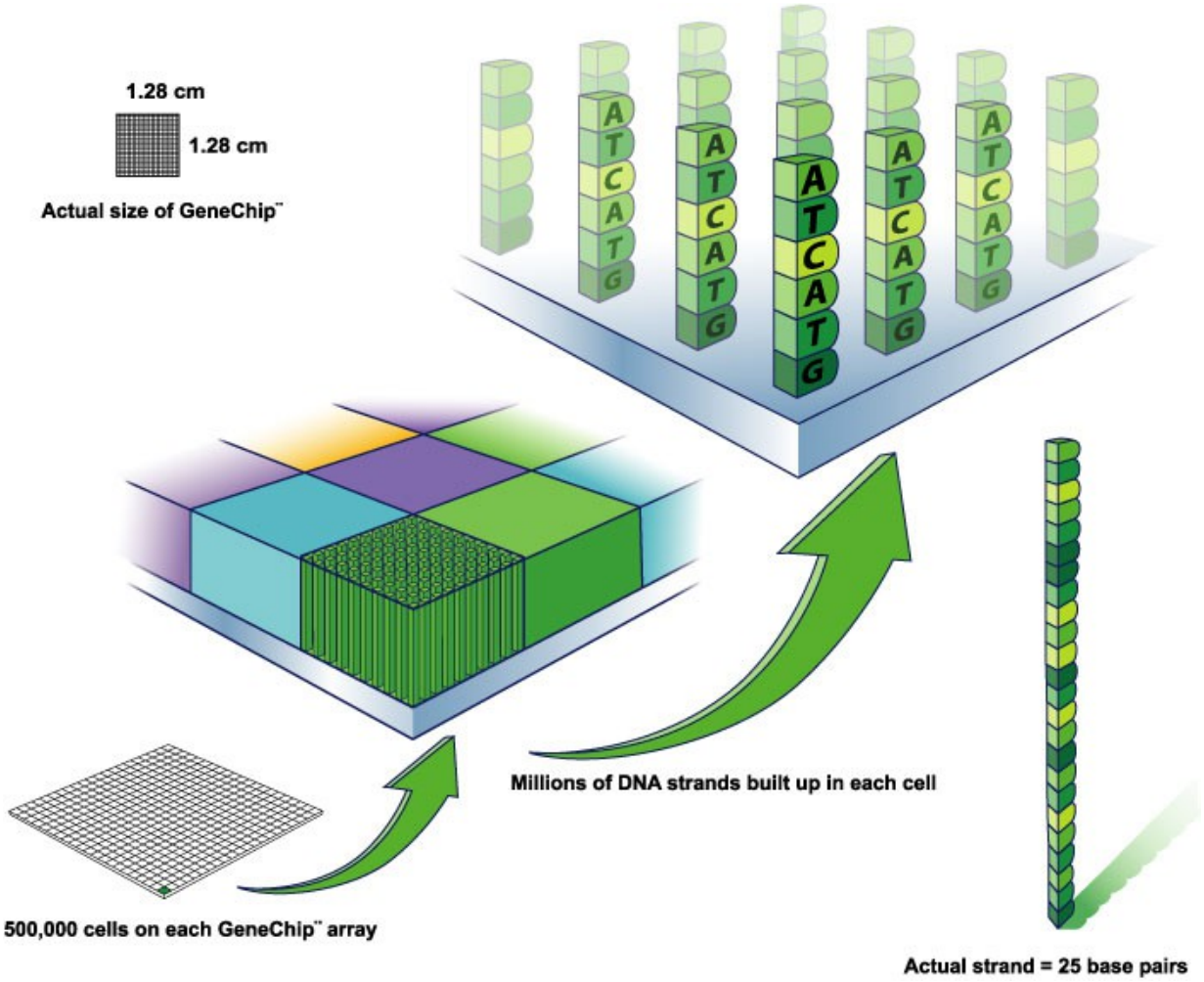


Image courtesy of Affymetrix.

Hybridization process

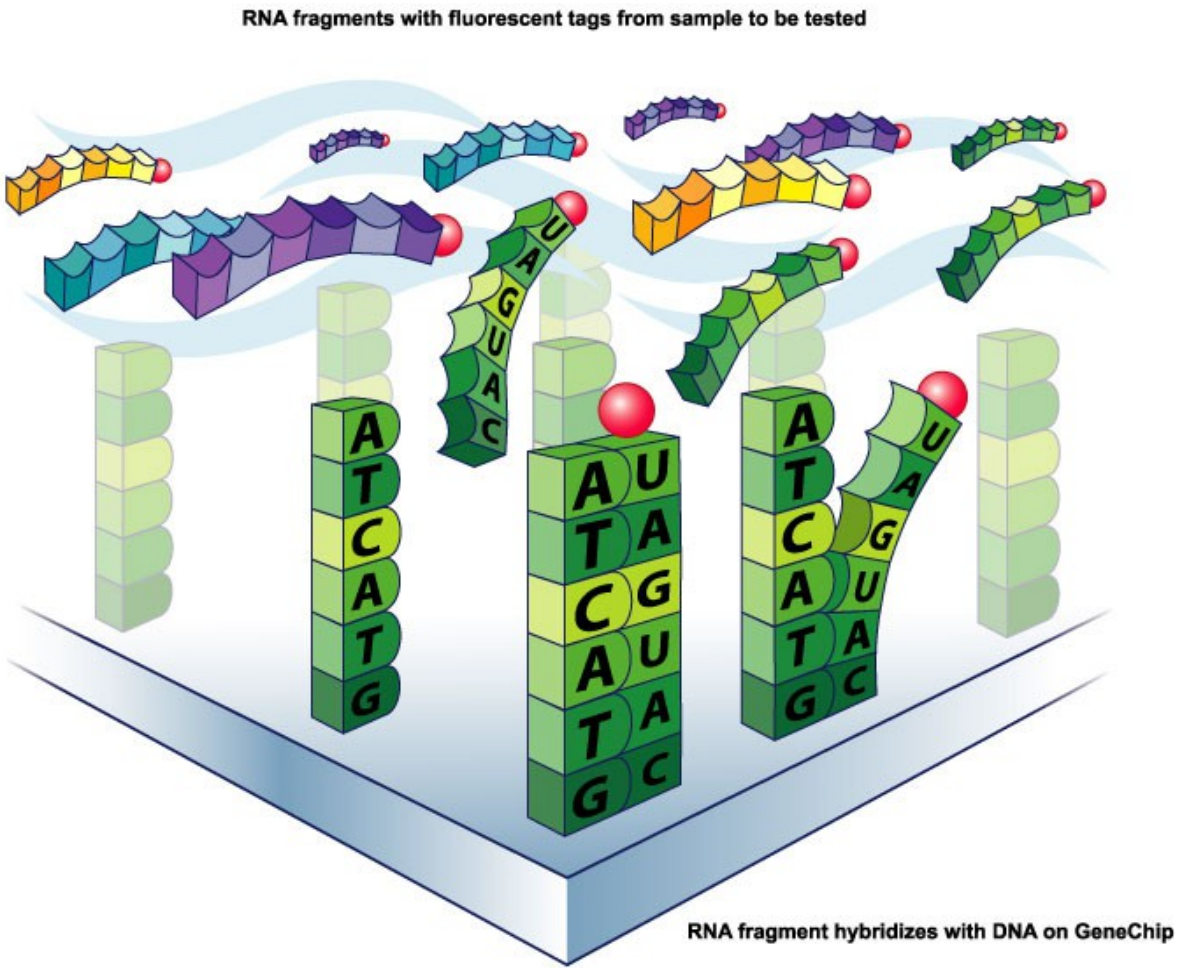
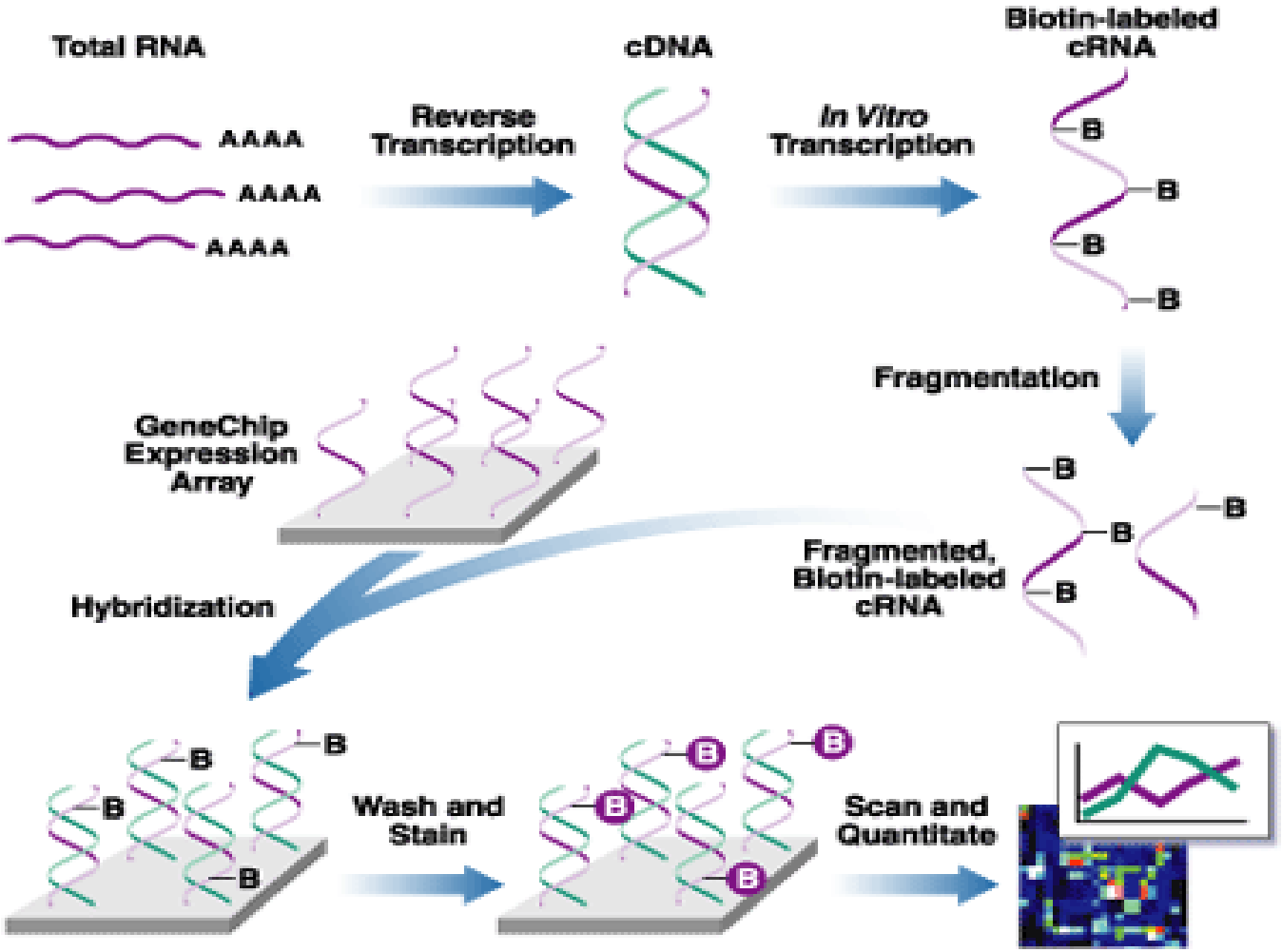


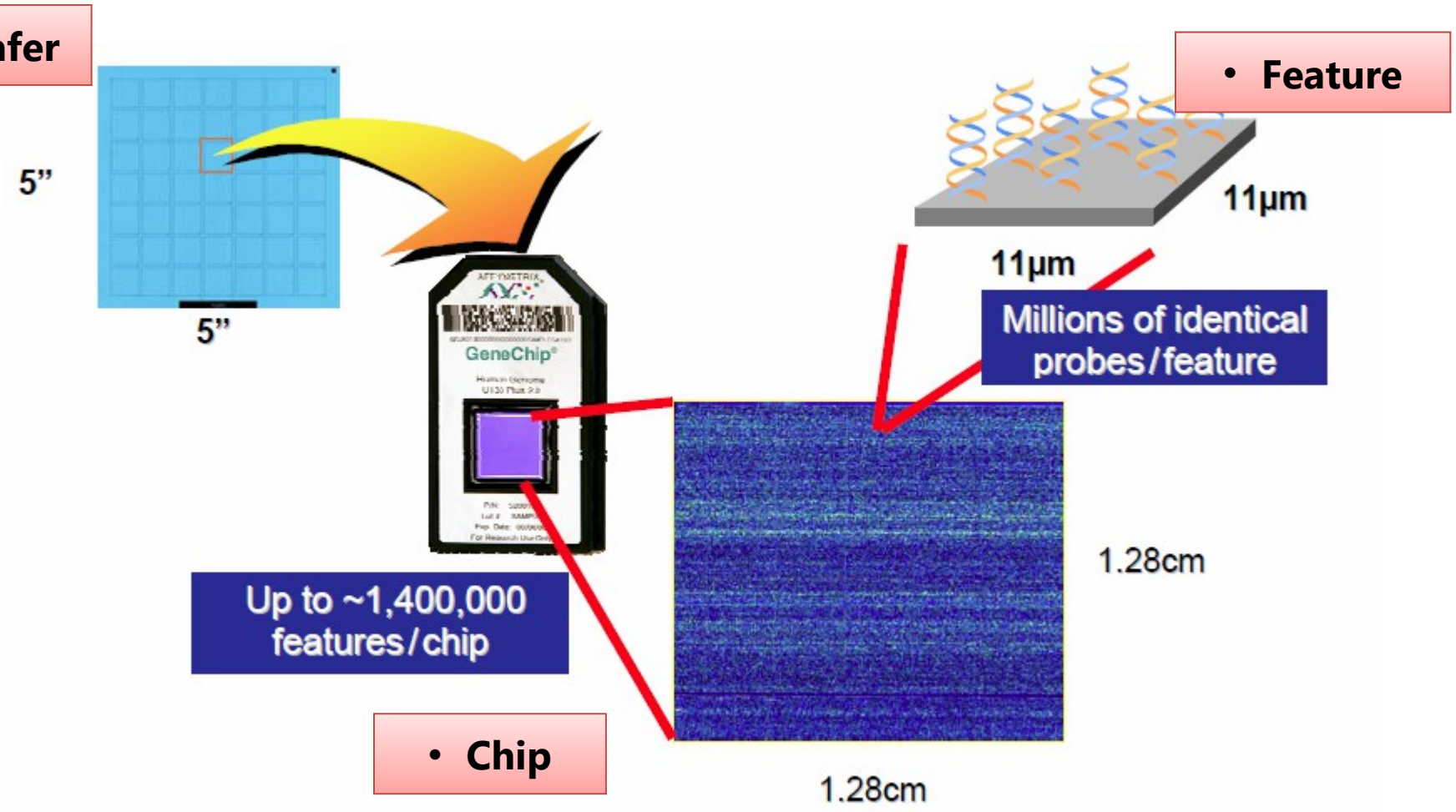
Image courtesy of Affymetrix.

Proces overview (Affy)



@Affymetrix

Final Chip


• Wafer
5”
5”
GeneChip®
Up to ~1,400,000 features/chip
• Chip
1.28cm
1.28cm
• Feature
11µm
11µm
Millions of identical probes/feature

Comparison between two types

cDNA Microarrays

ADVANTAGES

Cheaper (not anymore)
Flexibility in experimental design
High signal intensity (long secs)

DISADVANTAGES

Low reproducibility
Cross-hybridization (low specificity)
Need more manual handling (possibility of contamination)

Oligonucleotide Microarrays

ADVANTAGES

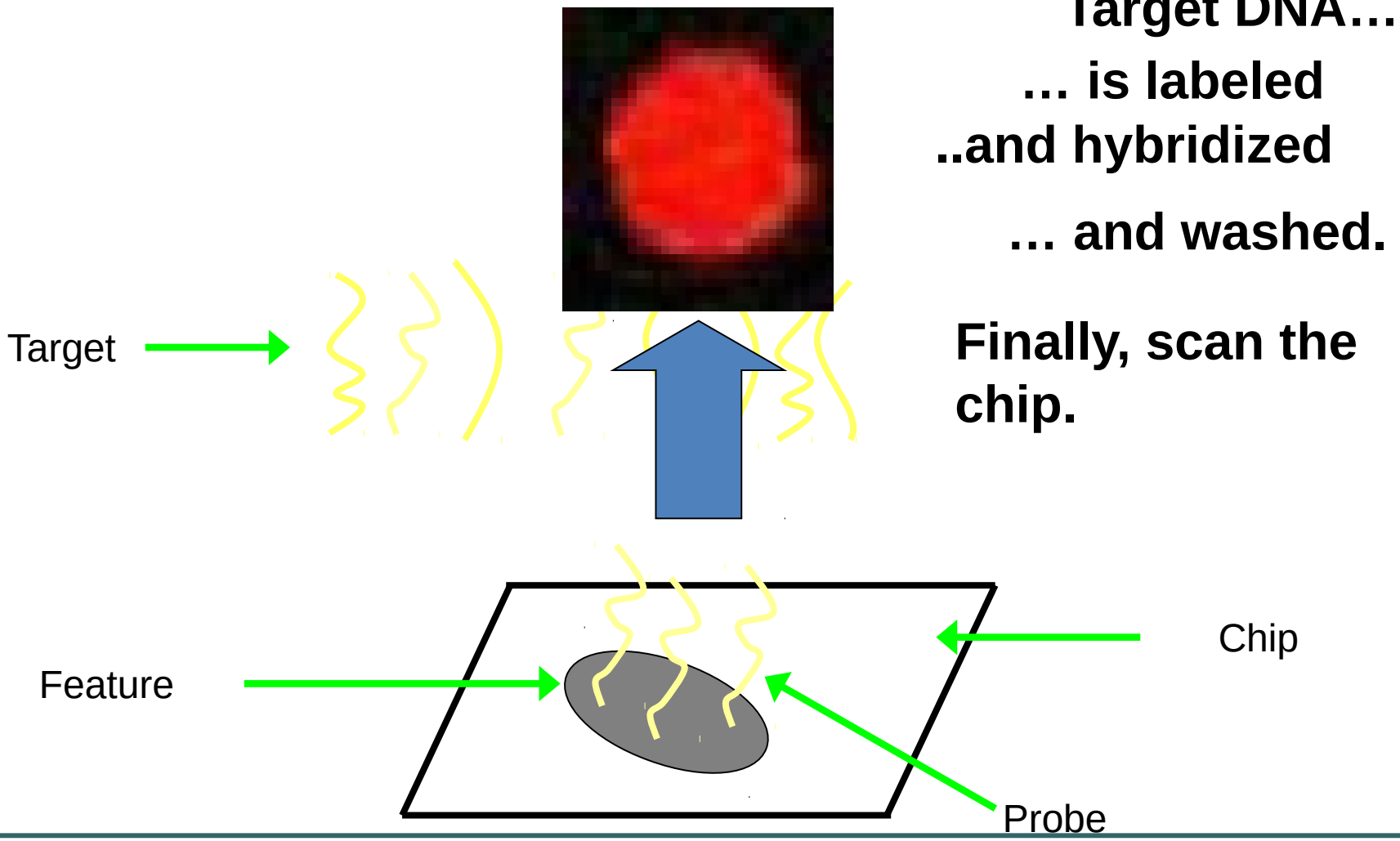
Quick and robotic manufacturing
High Reproducibility
High specificity (short sequences)
Use many probes / gene

DISADVANTAGES

Requires more specialized equipment
Expensive
Less flexible (genes on the chip cannot be selected)

Microarray Basics

Imagine a one-spot microarray…



Target DNA…

… is labeled

..and hybridized

… and washed.

Finally, scan the chip.

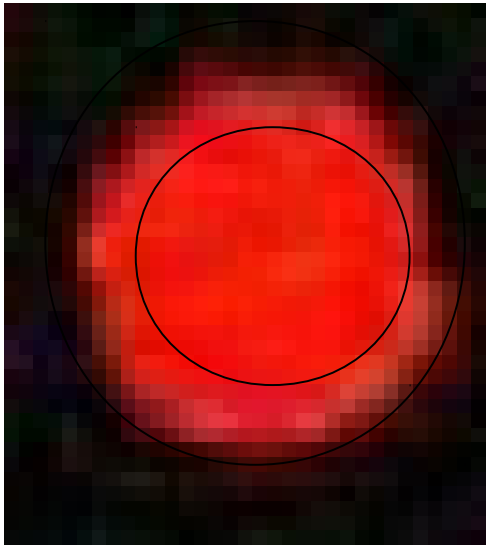
Spot-Weighting and Background Correction

Background correction: To remove stray signal using a Model-based method

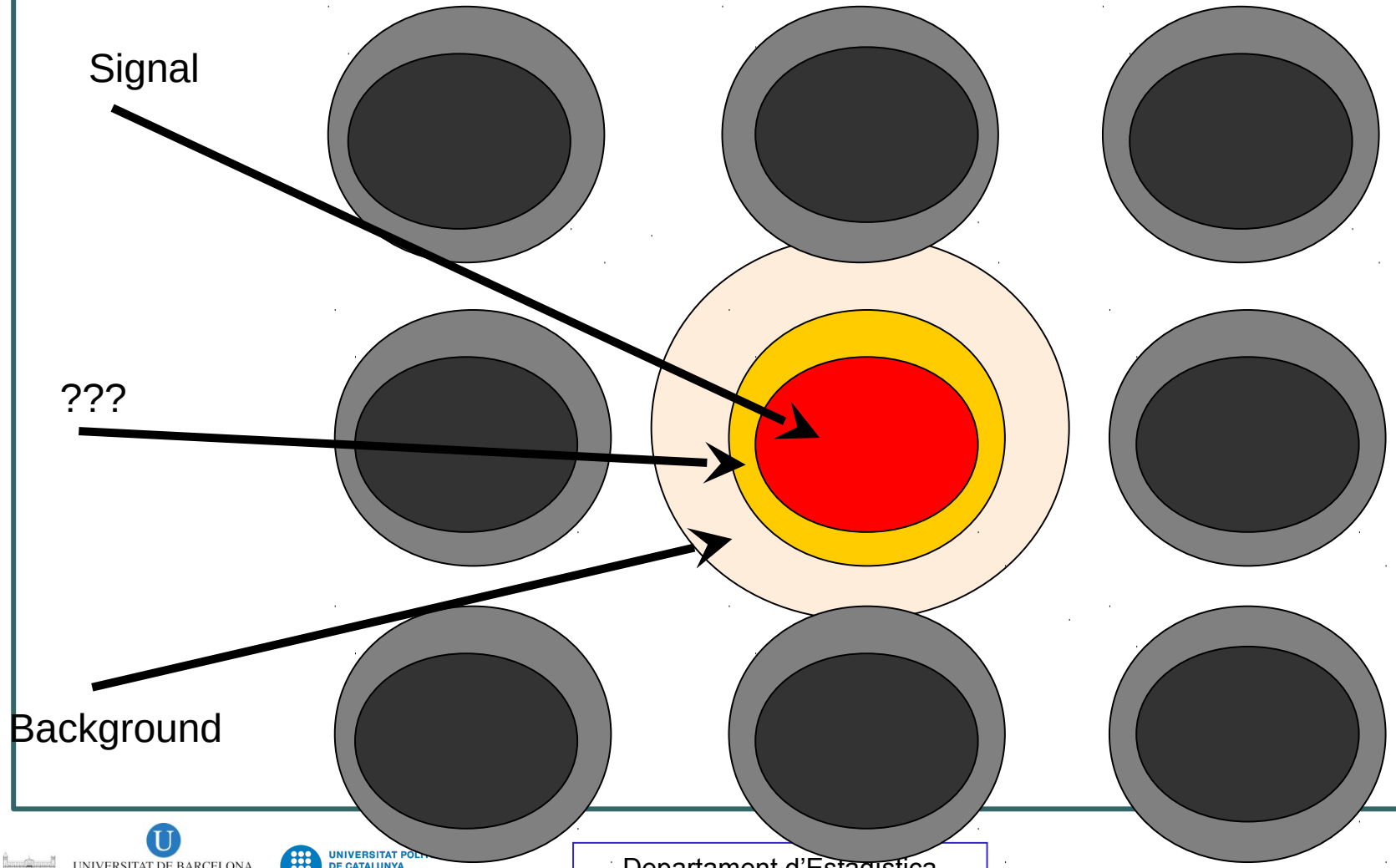
Spot-Weighting:

A “perfect” spot is used normally in analysis : Weight = 1

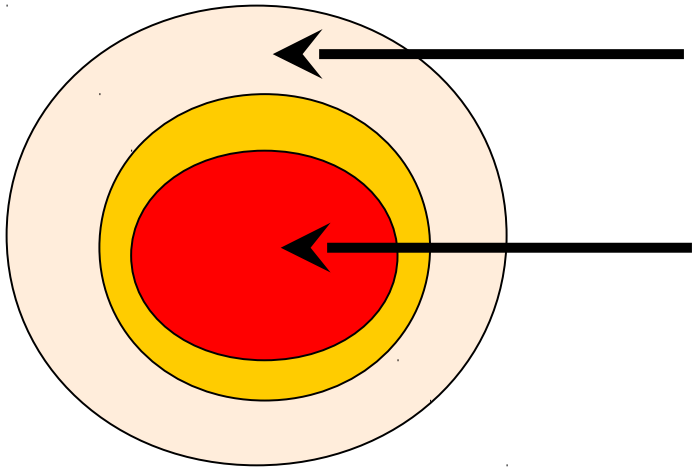
A “poor” spot is given less consideration : 0 < Weight < 1



Spot Segmentation



So what do we get?



Background Intensity: BG

Foreground Intensity: FG

Possibility:
Signal = FG - BG

NO!

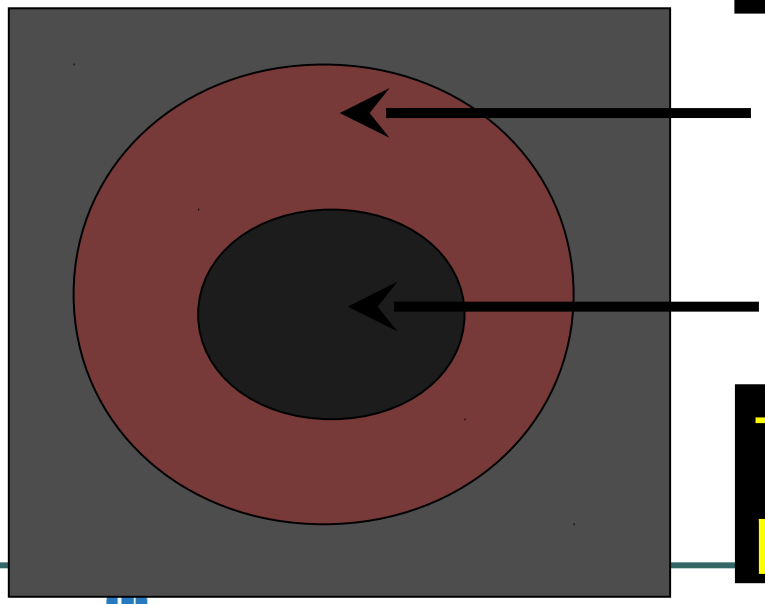
If BG > FG
Then negative Signal

0.1-2% of spots

Why Might This happen?

In 2001 two papers showed that empty spots have less signal than background

Unbound spots correspond to low-expression genes

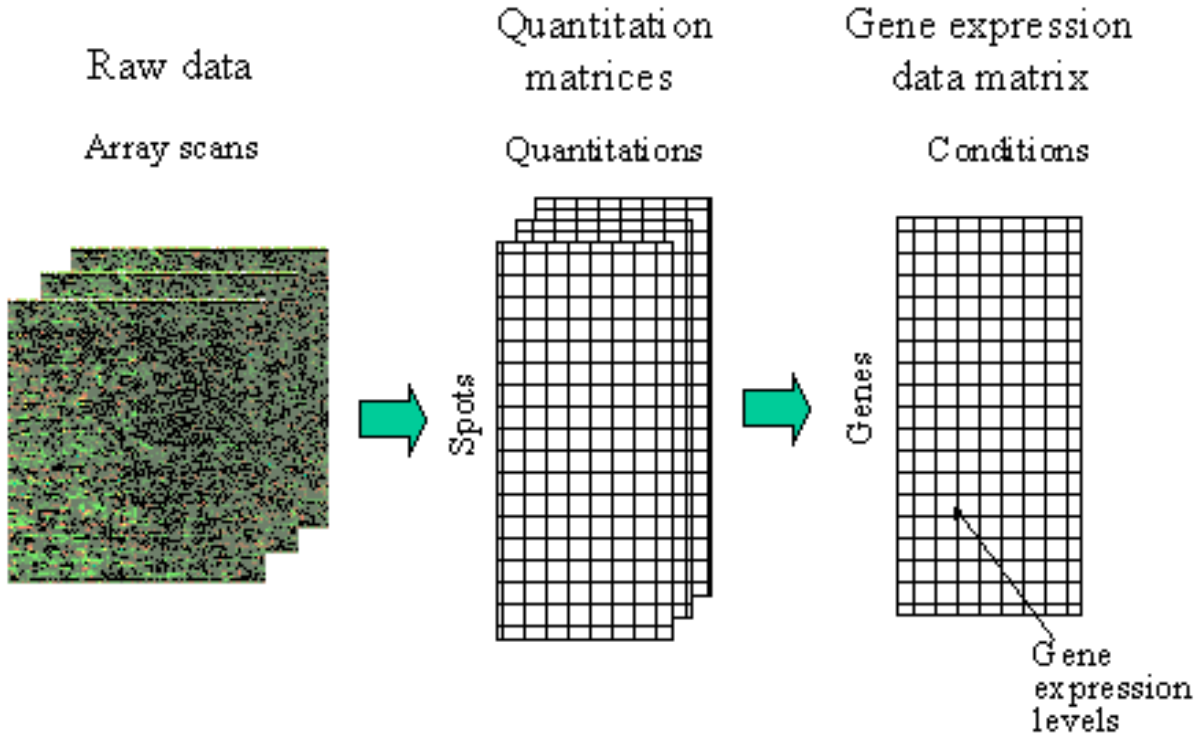


Background Intensity: BG

Foreground Intensity: FG

Thus unbound spots are particularly prone to problems

From image to expression matrix



Expression measures (cDNA)

Gene expression is measured from intensity measures as the (corrected) relative intensity of one dye vs. the (corrected) relative intensity of the other.

Background correction may be needed, or not, according to the array quality.

Example: gene expression data

Gene expression data on *6348* genes for *16* samples.

mRNA samples

Genes	T1	C1	T2	C2	T3	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene *i* in mRNA sample *j*

M = { Log(TSRBI / C*) ; Log(CFVB / C*)

C* = Control sample pools

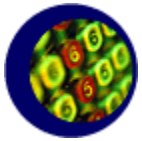
Expression measures (affy)

- Obtaining expression measures for affymetrix arrays is less straightforward.
  - —Background correction and normalization is required.
  - —There are multiple *PM* and *MM* values which must be integrated into one single expression value.
- The resulting summarized values are absolute expression measures which are more difficult to interpret than relative expression values.

Example: absolute expression values

	C01-001.CEL	C02-001.CEL	C03-001.CEL
1415670_at	8.954387	9.088924	8.833863
1415671_at	10.700876	10.639307	10.610953
1415672_at	10.377266	10.510106	10.461701
1415673_at	7.320335	7.252635	7.112313
1415674_a_at	8.381129	8.332256	8.393718
1415675_at	8.120937	8.082713	8.051514
1415676_a_at	10.322229	10.287371	10.282812
1415677_at	9.038344	8.979641	8.905711

Software for microarray data analysis



BRB-ArrayTools

Which software for the analysis?

Microarray experiments generate huge quantities of data which have to be:

Stored, managed, visualized, processed …

Many options available.

However…No tool satisfies all user's needs.

A tool must be:

- Powerful but user friendly.
- Complete but without too many options,
- Flexible but easy to start with and go further.
- Available, to date, well documented but affordable.

R and Microarrays

R is a popular tool between statisticians. Once they started to work with microarrays they continued using it.

- To perform the analysis.
- To implement new tools.

This gave rise very fast to lots of free R-based software to analyze microarrays.

The Bioconductor project groups many of these (but not all) developments.

The Bioconductor project

http://www.bioconductor.org

Open source and open development software project for the analysis and comprehension of genomic data.

Most early developments as R packages.

Extensive documentation and training material from short courses.

Has reached some stability but still evolving !!!

→ *what is now a standard may not be so in a future.*

Some pro's & con's

- Powerful,
- Used by statisticians
- Easy to extend
  - —Creating add-on **packages**
  - —Many already available
- Freely available
- Unix, windows & Mac
- Lot of documentation

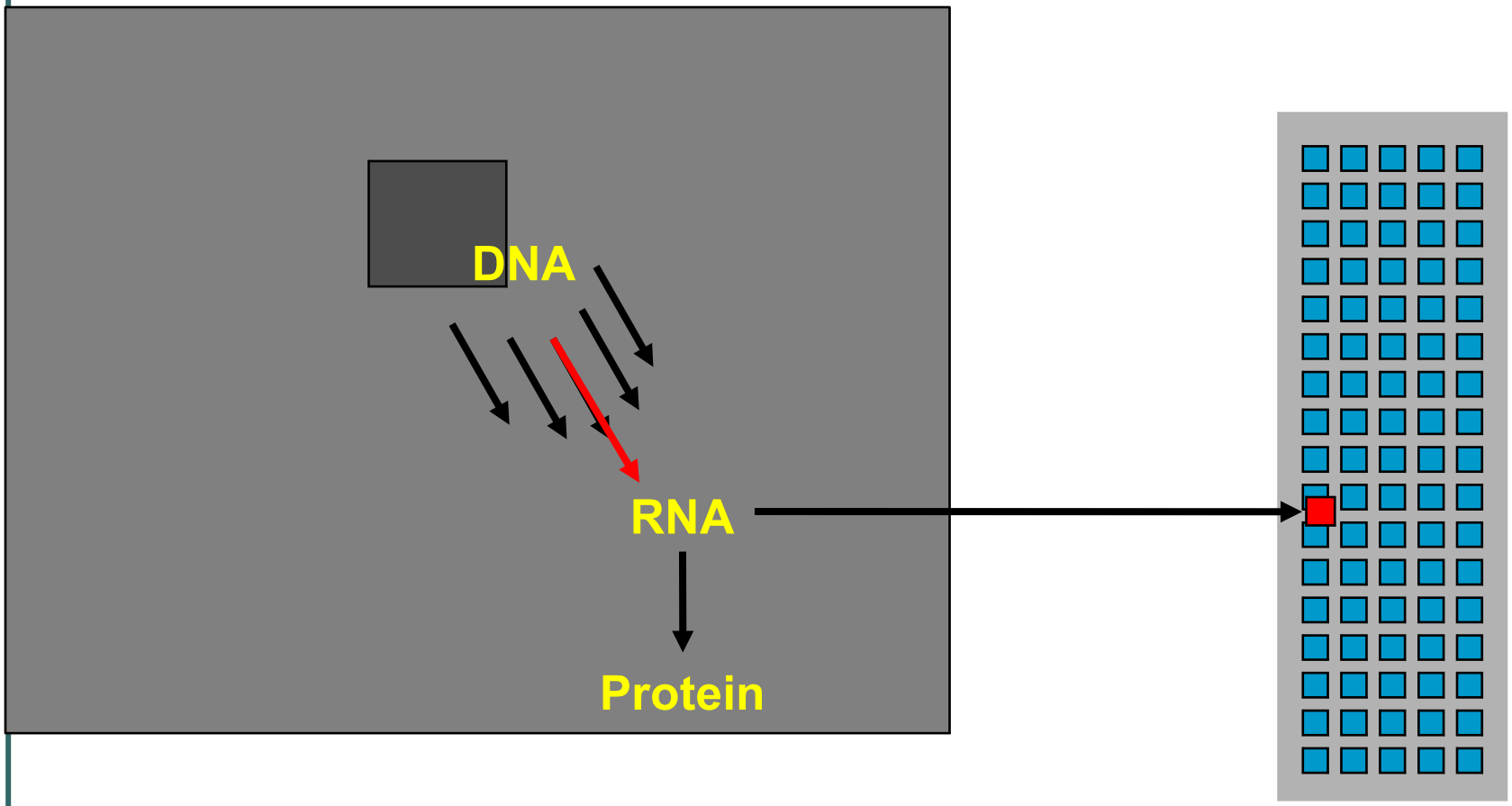
- Not very easy to learn
- Command-based
- Documentation sometimes cryptic
- Memory intensive
  - Worst in windows
- Slow at times

Gene Annotations in Genomics Experiments

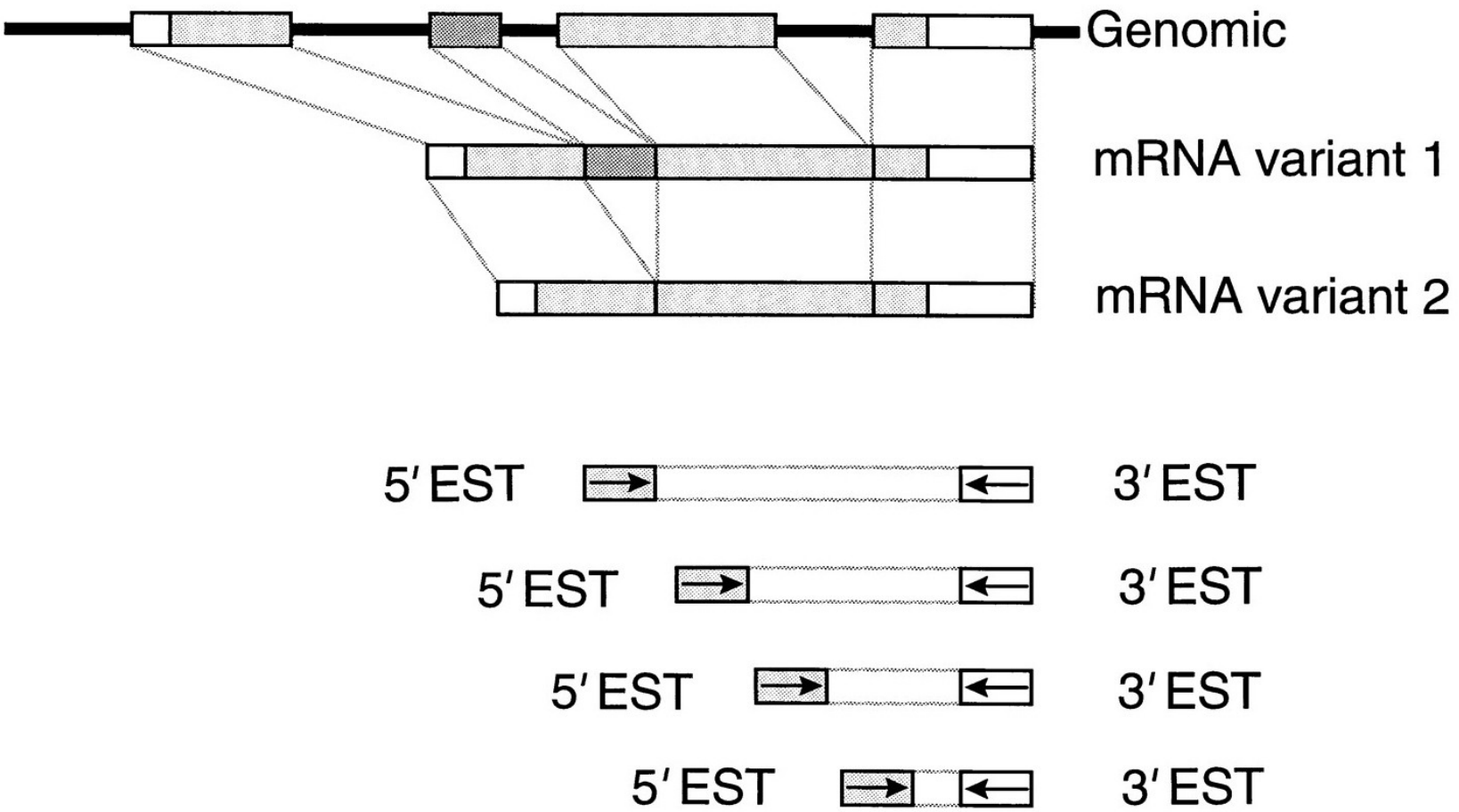
Biological preliminaries

- Every cell in the human body contains the entire human genome: 3.3 Gb or ~30K genes.
- The investigation of gene expression is meaningful because different cells, in different environments, doing different jobs express different genes.
- To-do list to *create platforms* for gene expression analysis:
  - Define what a gene is.
  - Identify genes in a sea of genomic DNA where <3% of DNA is contained in genes.
  - Design and implement probes that will effectively assay expression of ALL (most? many?) genes simultaneously.
  - Cross-reference these probes.

Gene Expression and Microarray analysis



From Genomic DNA to mRNA Transcripts



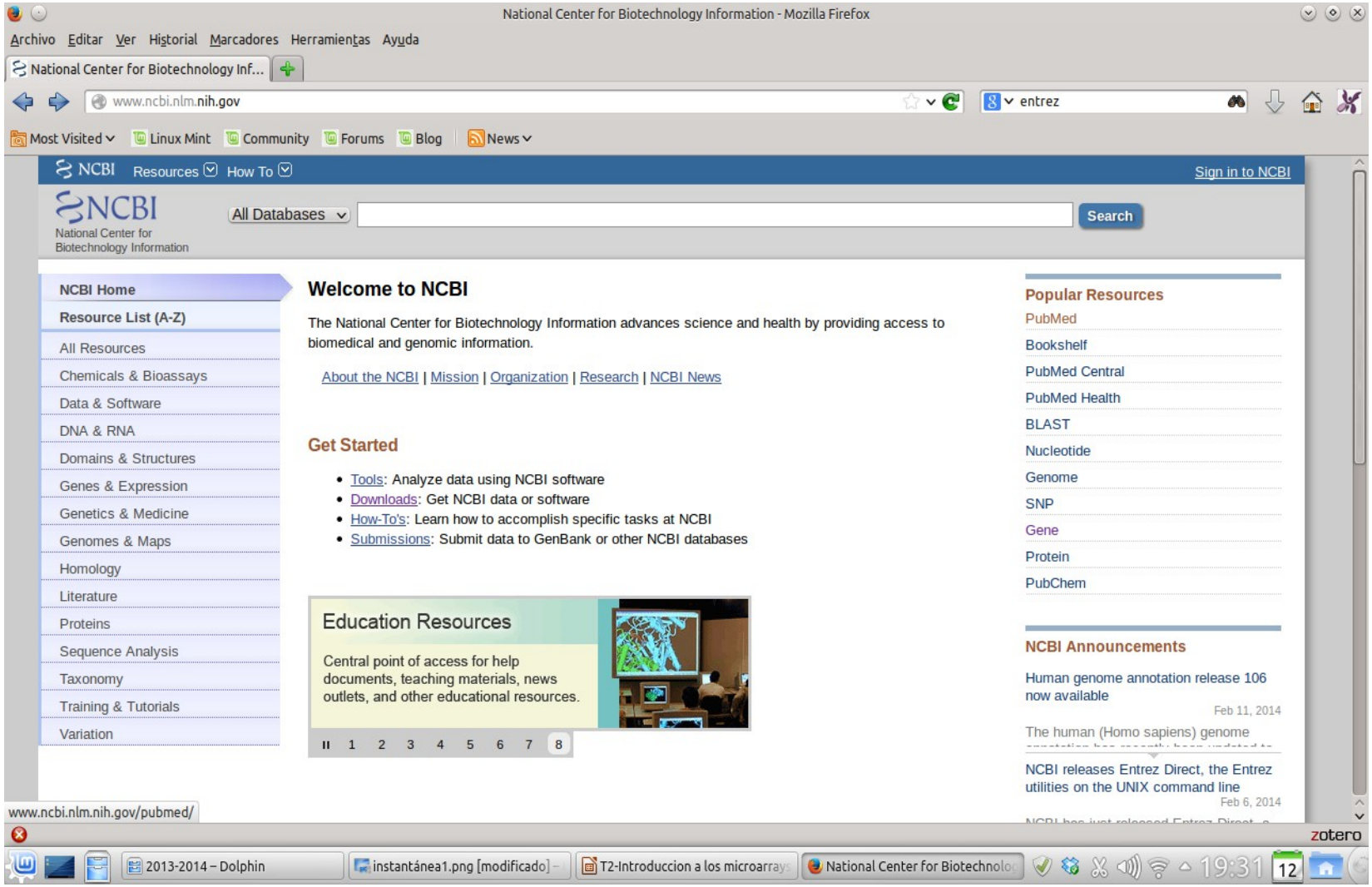
Sequence and Gene databases

- Probes have to be mapped to databases.
- These may be either gene or sequence databases

- Sequence databases:
  - From which sequence has the probe been synthesized?

- Gene databases
  - Which gene is the probe intended to interrogate

NCBI


National Center for Biotechnology Information - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
National Center for Biotechnology Inf...
www.ncbi.nlm.nih.gov
entrez
Most Visited Linux Mint Community Forums Blog News
NCBI Resources How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | Research | NCBI News
Get Started
Tools: Analyze data using NCBI software
Downloads: Get NCBI data or software
How-To's: Learn how to accomplish specific tasks at NCBI
Submissions: Submit data to GenBank or other NCBI databases
Education Resources
Central point of access for help documents, teaching materials, news outlets, and other educational resources.
1 2 3 4 5 6 7 8
Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI Announcements
Human genome annotation release 106 now available
Feb 11, 2014
The human (Homo sapiens) genome
NCBI releases Entrez Direct, the Entrez utilities on the UNIX command line
Feb 6, 2014
www.ncbi.nlm.nih.gov/pubmed/
zotero
2013-2014 – Dolphin
instantánea1.png [modificado]
T2-Introduccion a los microarrays
National Center for Biotechnolog
19:31

RefSeq

- The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.
- Similar to a review article, a RefSeq is a synthesis of information integrated across multiple sources at a given time. RefSeqs provide a foundation for uniting sequence data with genetic and functional information.
- They are generated to provide reference standards for multiple purposes ranging from genome annotation to reporting locations of sequence variation in medical records.

http://www.ncbi.nlm.nih.gov/RefSeq/

Unigene

- UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters.
- Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.
- In addition to sequences of well-characterized genes, hundreds of thousands novel expressed sequence tag (EST) sequences have been included. Consequently, the collection may be of use to the community as a resource for gene discovery.

http://www.bioinfo.org.cn/relative/NCBI-UniGene.htm

Databases for microarrays

Microarray Data in a Nutshell

Lots of data to be managed before and after the experiment:

Data to be stored before the experiment .

- Description of the *array* and the *sample*.
- Direct access to all the cDNA and gene sequences, annotations, and physical DNA resources.

Data to be stored after the experiment

- Raw Data - scanned images.
- Gene Expression Matrix - Relative expression levels observed on various sites on the array.

Hence we can see that ***database software capable of dealing with larger volumes of numeric and image data is required.***

Why Databases?

Tailored to datatype
Tailored to the Scientists

Intuitive ways to query the data

- Diagrams, forms, point and click, text etc.

Support for efficient answering of queries.

- Query optimisation, indexes, compact physical storage.

Gene Expression Databases:Integration

- There are many different types of data presenting numerous relationships.
- There are a number of Databases with lots of information.
- Experiments need to be compared because the experiments are very difficult to perform and very expensive.
- Solution: Make all the databases talk the same language.
- XML was the choice of data interchange format.

Gene Express Omnibus

The Gene Expression Omnibus ia a gene expression database hosted at the National library of Medicine
It supports four basic data elements
Platform (the physical reagents used to generate the data)
Sample (information about the mRNA being used)
Submitter (the person and organisation submitting the data)
Series (the relationship among the samples).
It allows download of entire datasets, it has not ability to query the relationships
Data are entered as tab delimited ASCII records,with a number of columns that depend on the kind of array selected.
Supports Serial Analysis of Gene Expression (SAGE) data.

http://www.ncbi.nlm.nih.gov/geo/