Universidad Complutense de Madrid
ESCUELA DE VERANO 2007
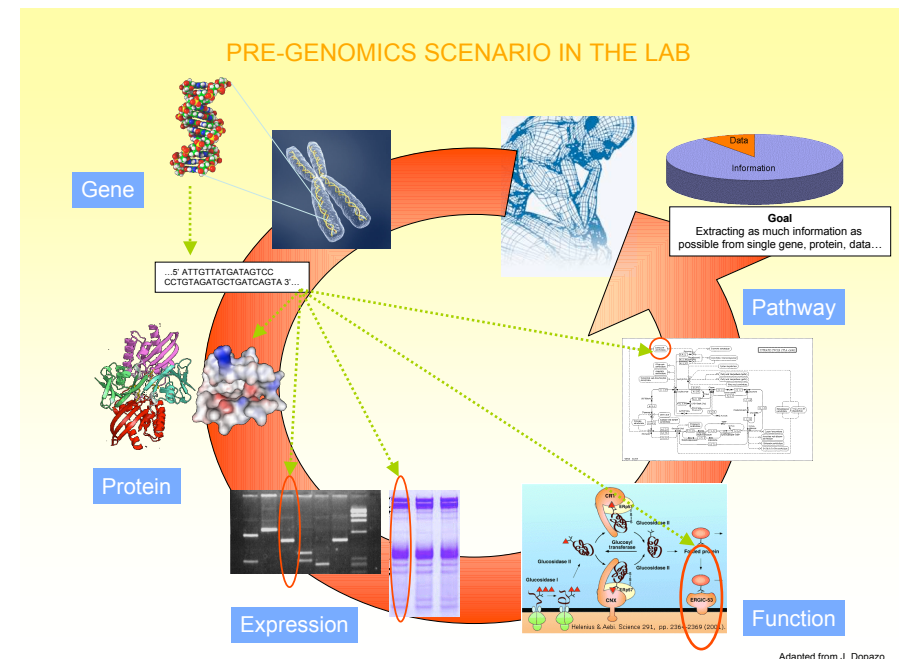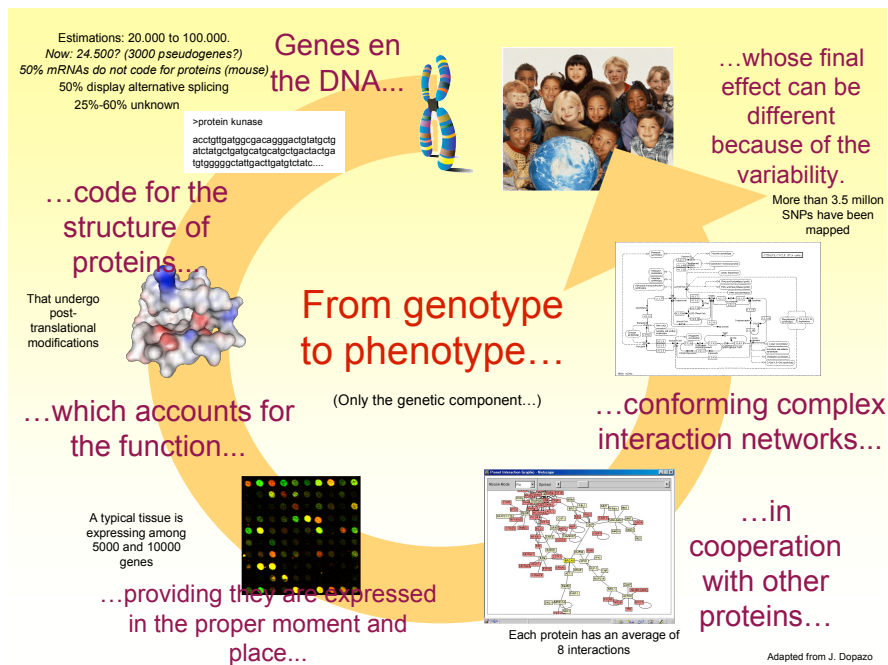
Introducción al análisis de datos en microarrays
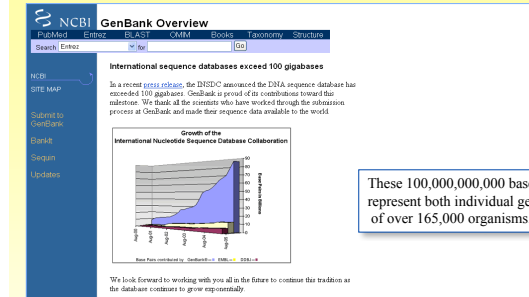
INB*
cnio

Gonzalo Gómez López
INB-CNIO
ggomez@cnio.es

Madrid, 26 de Julio, 2007

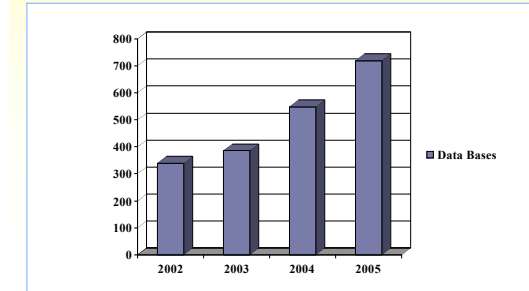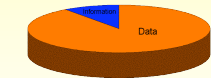---

Universidad Complutense de Madrid
ESCUELA DE VERANO
2007

1. Introducción
2. Microarrays (metodología, tipos, bases de datos, aplicaciones…)
3. JC Oliveros (Normalización, Análisis exploratorio, …)
4. Análisis de datos (Análisis supervisado y no supervisado, algoritmos…)
5. Practica

ggomez@cnio.es

Estimations: 20.000 to 100.000.
Now: 24.500? (3000 pseudogenes?)
50% mRNAs do not code for proteins (mouse)
50% display alternative splicing
25%-60% unknown

Genes en the DNA...

>protein kunase
acctgttgatggcgacagggactgtatgctg
atctatgctgatgcatgctgactactga
tgtgggggctattgacttgatgtctatc...

...whose final effect can be different because of the variability.

More than 3.5 millon SNPs have been mapped

...code for the structure of proteins...

That undergo post-translational modifications

From genotype to phenotype…

(Only the genetic component…)

...which accounts for the function...

...conforming complex interaction networks...

A typical tissue is expressing among 5000 and 10000 genes

...providing they are expressed in the proper moment and place...

Each protein has an average of 8 interactions

...in cooperation with other proteins…

Adapted from J. Dopazo

PRE-GENOMICS SCENARIO IN THE LAB

Gene

...5' ATTGTTATGATAGTCC CCTGTAGATGCTGATCAGTA 3'...

Data
Information

Goal
Extracting as much information as possible from single gene, protein, data…

Pathway

Protein

Expression

Function

Helenius & Aebi. Science 291, pp. 2364-2369 (2001)

Adapted from J. Dopazo

Crecimiento de GenBank, EMBL, DDBJ
(GenBank 5-10-2005)

GenBank Overview

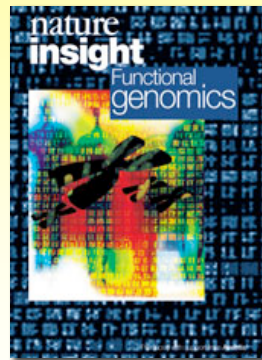International sequence databases exceed 100 gigabases

These 100,000,000,000 bases, or "letters" of the genetic code, represent both individual genes and partial and complete genomes of over 165,000 organisms.
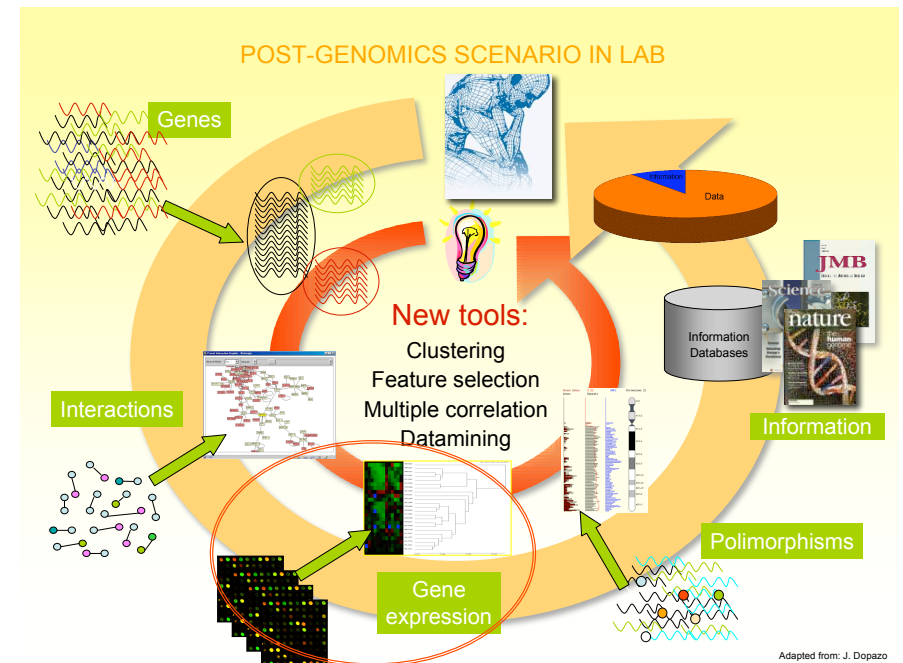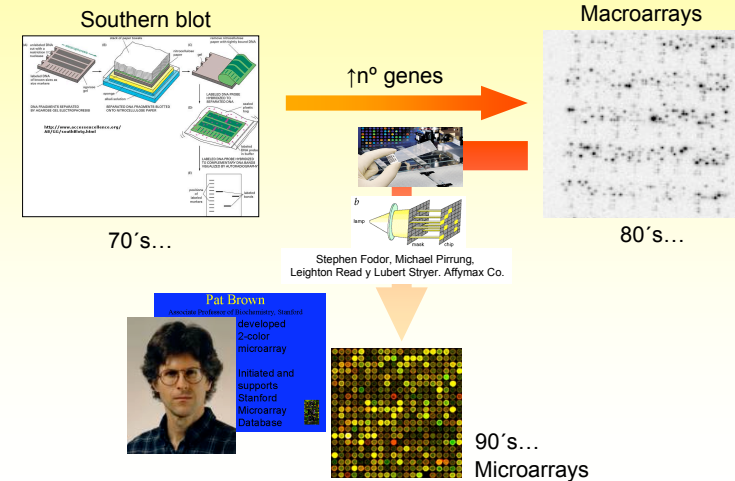
2002	2003	2004	2005

■ Data Bases

Bases de datos públicas en Biomedicina y Biología Molecular
Nucleic Acids Research, 2005, Vol. 33.

Era Pregenómica y Postgenómica

"Mientras que hace unos años los científicos se centraban en el estudio particular de determinadas proteínas de la célula, hoy catálogos completos de genes y proteínas de muchos organismos están sobre la mesa esperando resolvamos el puzzle de cómo funcionan esas células. Es necesario seguir con ambas estrategias (se complementan)"

POST-GENOMICS SCENARIO IN LAB

Genes

Interactions

New tools:
Clustering
Feature selection
Multiple correlation
Datamining

Information Databases

Information

Polimorphisms

Gene expression

Adapted from: J. Dopazo

Slide 1

Universidad Complutense de Madrid
ESCUELA DE VERANO
2007

1. Introducción
2. Microarrays (metodología, tipos, bases de datos, aplicaciones…)
3. JC Oliveros (Normalización, análisis exploratorio…)
4. Análisis de datos (Análisis supervisado y no supervisado, algoritmos…)
5. Práctica

ggomez@cnio.es

# Slide 2

¿Qué es un microarray?

UN POCO DE HISTORIA…

Southern blot

↑nº genes

Macroarrays

70´s…

Stephen Fodor, Michael Pirrung,
Leighton Read y Lubert Stryer. Affymax Co.

80´s…

Pat Brown
Associate Professor of Biochemistry, Stanford
developed
2-color
microarray

Initiated and
supports
Stanford
Microarray
Database

90´s…
Microarrays

Slide 3

¿Qué es un microarray?

Sondas (impresas en el chip):
ss cDNA
ss oligos
proteínas
anticuerpos
cromosomas,
etc.
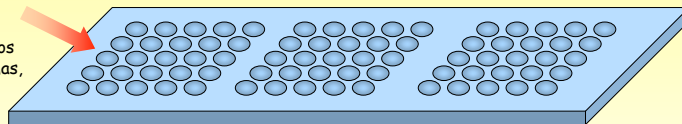
Muestras dianas marcadas
(extraída en el laboratorio):
proteínas
ss ac.nucleicos
etc

Slide 4
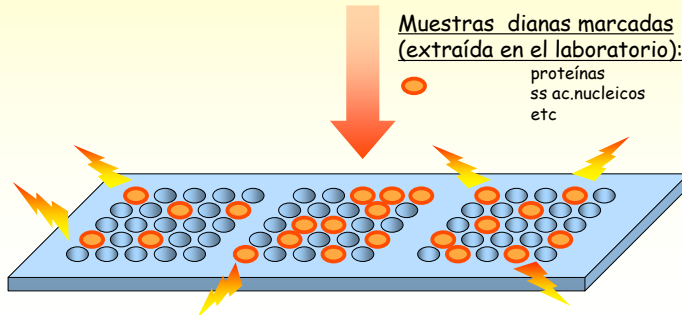
EL PRIMER DNA MICROARRAY

45 GENES DE ARABIDOPSIS
3 GENES CONTROL.
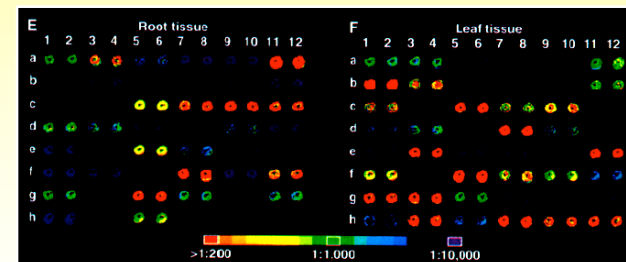
Schena *et al.* (1995) Science, 270, 467-470
Quantitative monitoring of gene expression patterns with a complementary DNA microarray.

Provided by JC Oliveros

Publicaciones basadas en microarrays
Pub Med
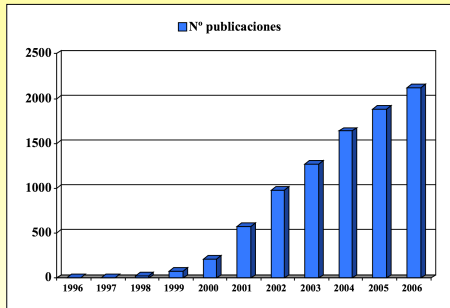
Microarrays en Internet (5-10-05)
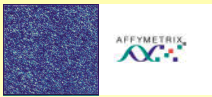Google
YAHOO!
alltheweb
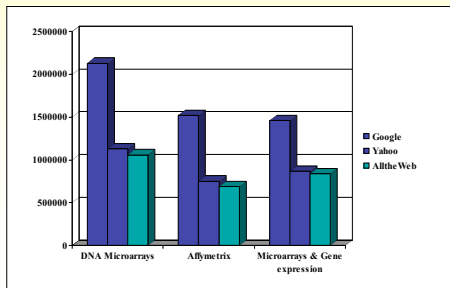
GEO submissions
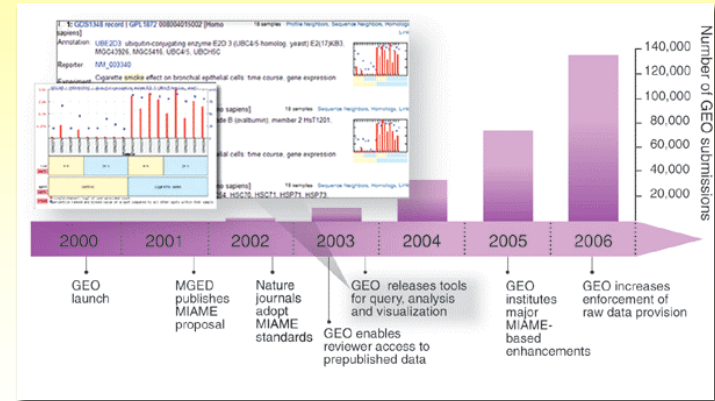
Tipos de microarrays

I) Acidos nucleicos
(expresión, SNPs, miRNAs…)

Oligos

Fotolitografía (25 b.)
(vidrio)
AFFYMETRIX

Impresión
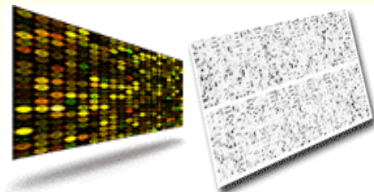(25-60 b.) (vidrio, plástico)

cDNAs

Impresión
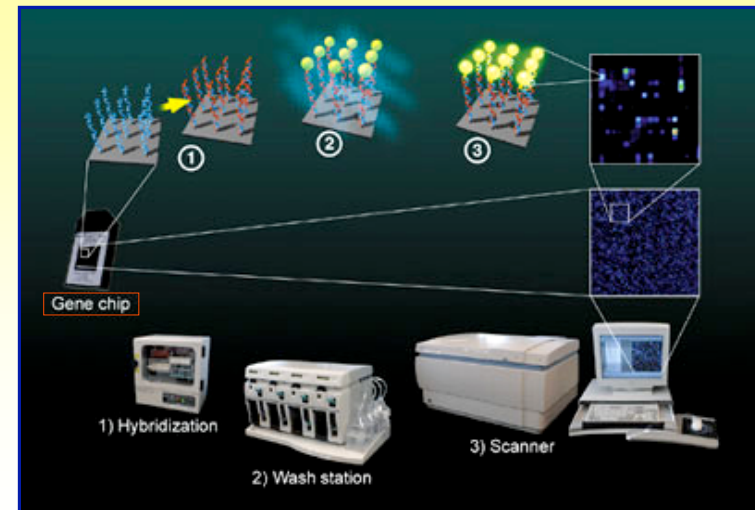(vidrio, plástico, membrana)
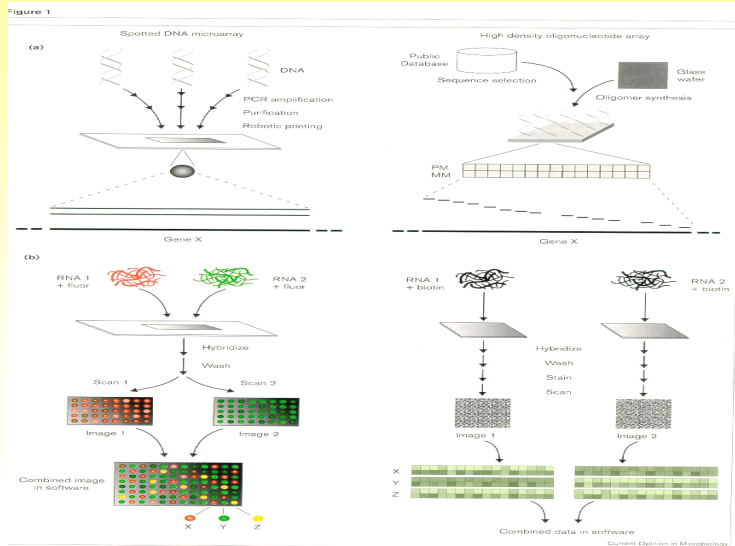
II) Proteínas
III) Tejidos
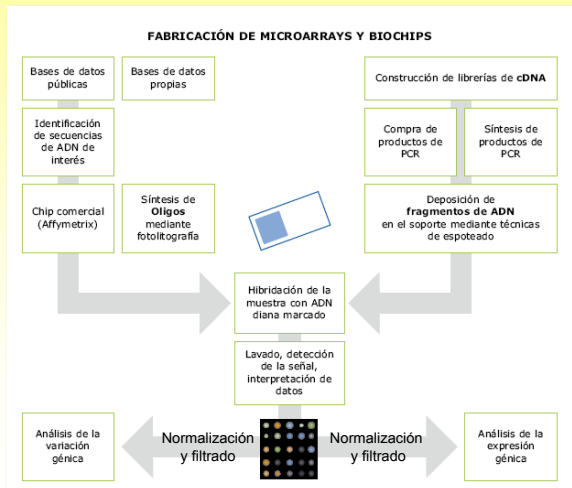IV) Cultivos celulares
V) CGH

I) Microarrays de ac. nucleicos

Gene chip

1) Hybridization

2) Wash station

3) Scanner

www.mskcc.org

Spotted microarrays and Affymetrix technologies

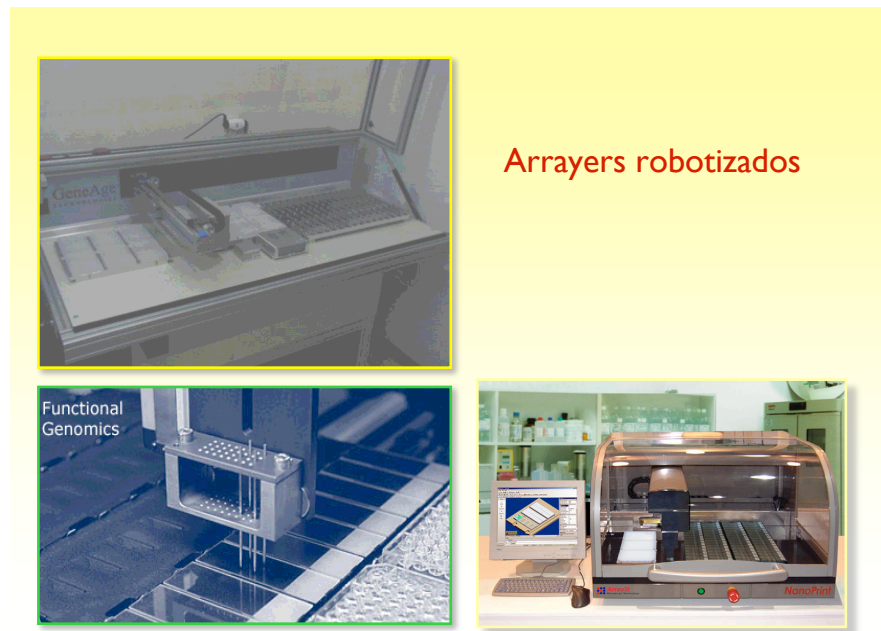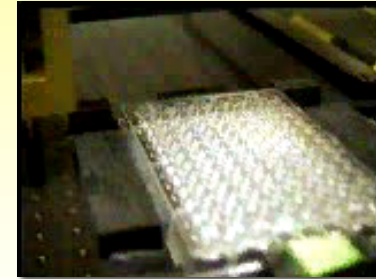## EJEMPLO
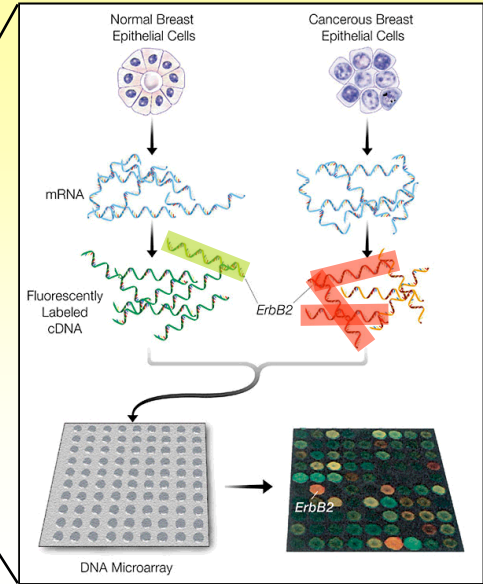
Haciendo microarrays…

Tratamiento de la muestra

?
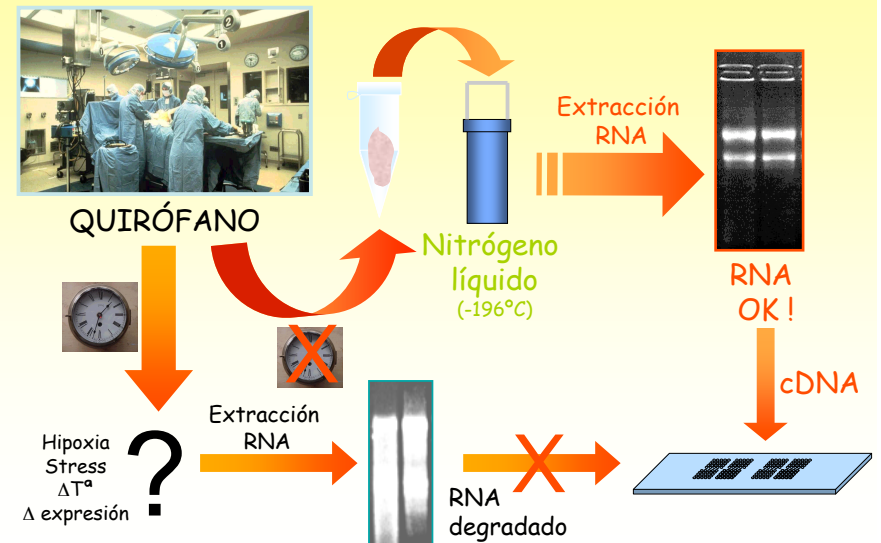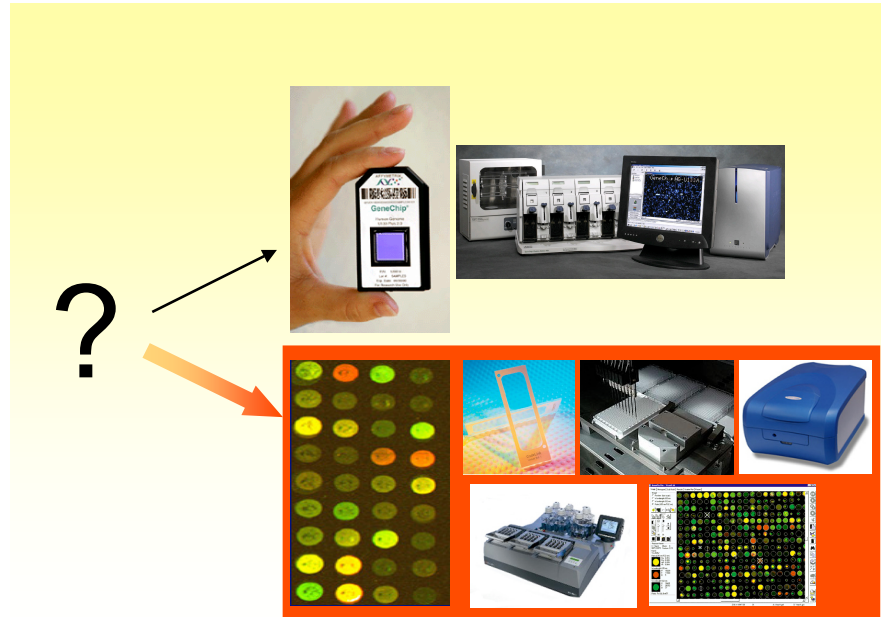
Imprimiendo un microarray

Arrayers robotizados

Functional Genomics
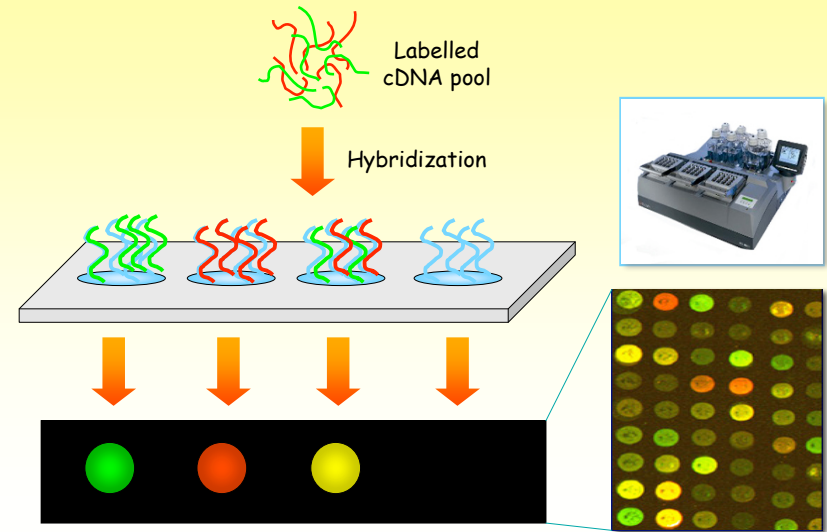

Ink Jet Tech

2-color standart protocol (spotted arrays)

Cloning

Growing

×2^n

Printing

Control RNA

Tumoral RNA

(RNA T7amplification)

RT reaction (Cy3)

RT reaction (Cy5)

Labelled cDNAs

Hibridization

Washes

Scanning and Image analysis
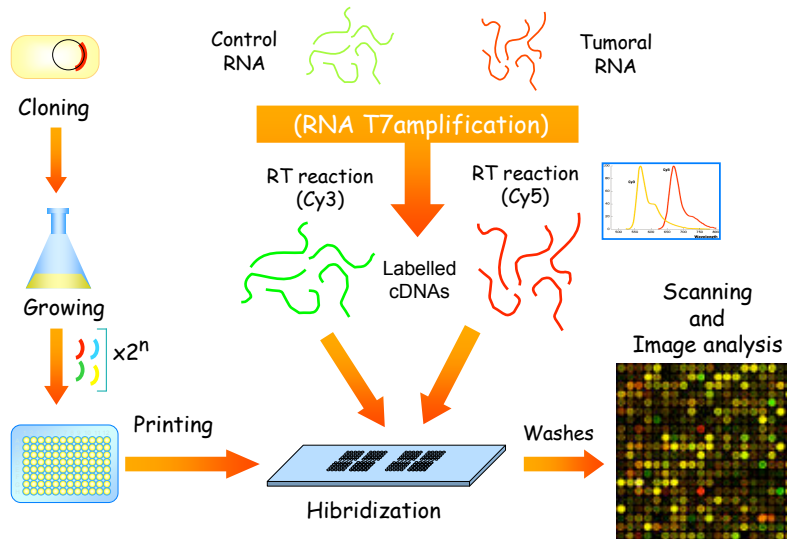
Microarray hybridization

Labelled cDNA pool

Hybridization

Scanning

Ch 1 / Ctrl - E:\imgsamps\BMS\MDYN\MD...

Ch 2 / Data - E:\imgsamps\BMS\MDYN\M...

Ch 3 / Output - Processed Image

Cy3/Cy5 labelling

Cy3

Cy5

Wavelength

Scanners

Image analysis

Image analysis
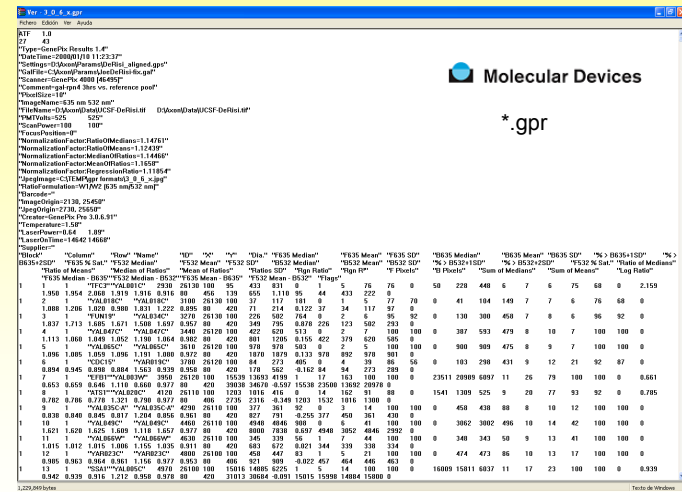
Microarrays flash movie…

http://www.bio.davidson.edu/Courses/genomics/chip/chip.html

Síntesis *in situ* por fotolitografia
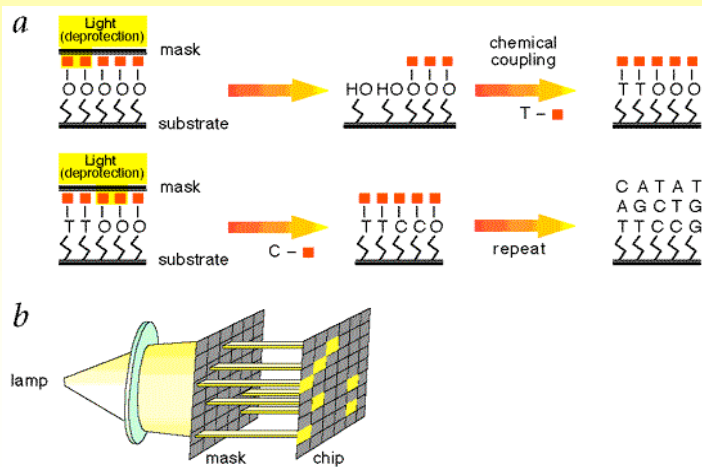
RNA EXTRACTION

1- color standart protocol (Affymetrix)

sscDNA

dscDNA

Wash columns

dscDNA

cRNA
Biot- labelling

Wash columns
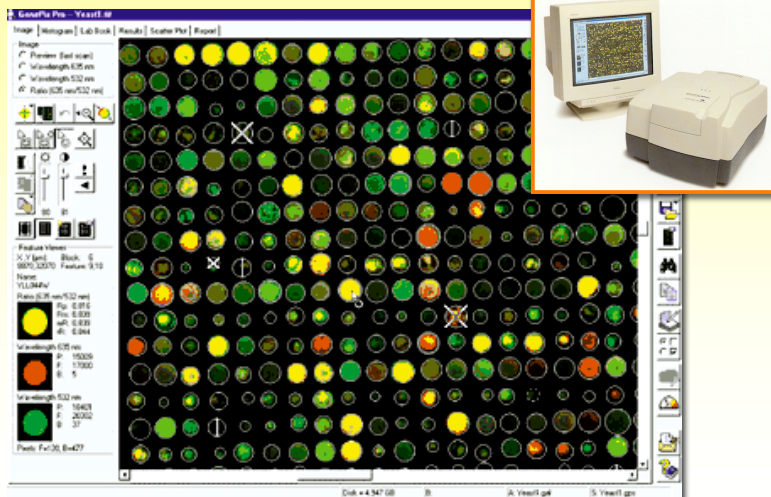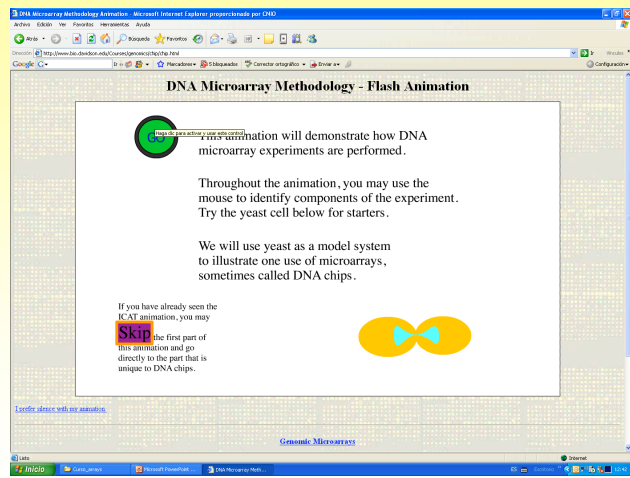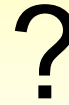
FRAGMENTATION

MICROARRAYS

1 color arrays

2 color arrays

Ventajas

Protocolos estandarizados
Procesos automatizados
Reproducibilidad
Ampliamente aceptado
Sistema cerrado

Limitaciones

Flexibilidad
Sistema cerrado
Precio

Ventajas

Precio
Flexibilidad
Producción propia
Sistema abierto
Procesos automatizados
(en progresión)

Limitaciones

Estandarización en proceso
Reproducibilidad (en progresión)
Sistema abierto

- AB Human Genome Survey v2.0 (29.098 genes)
- AB Mouse Genome Survey (32.000 genes)
- AB Rat Genome Survey (27.000 genes)

GE Healthcare
Life Sciences

Amersham
Biosciences

GE-CodeLink

Oligos

HUMAN
CodeLink™ Human Whole Genome Bioarray (~55 000 human genes)
CodeLink UniSet™ Human 20K I Bioarray (~20 000 human genes)
CodeLink UniSet Human I Bioarray (~10 000 human genes)
CodeLink UniSet Human II Bioarray (~10 000 human genes)

RAT
CodeLink™ Rat Whole Genome Bioarray (~34 000 rat gene targets)
CodeLink UniSet™ Rat I Bioarray (~10 000 rat gene targets)
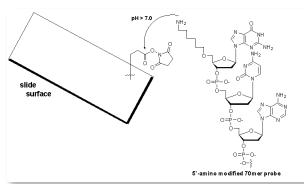
MOUSE
CodeLink™ Mouse Whole Genome Bioarray (~36 000 mouse gene targets)
CodeLink UniSet™ Mouse 20K I Bioarray (~20 000 mouse gene targets)
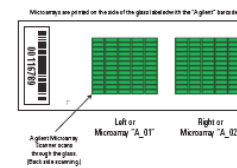CodeLink UniSet Mouse I Bioarray (~10 000 mouse gene targets)

SNPs

CodeLink slides contain a gel-like polymer surface
that is derivatized with an N-hydroxysuccinimide ester.

Agilent Technologies

60mer - Oligos

cDNA

50-mer probes

Nanochip™
Lab-on-a-chip

Spectrochip
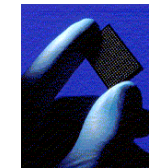
Etc, etc, etc…..

- ASOCIACIÓN GÉNIKA
- CENTRO NACIONAL DE BIOTECNOLOGÍA (CNB-CSIC)
- FUNDACIÓN DE INVESTIGACIÓN DEL CÁNCER
- HOSPITAL UNIVERSITARIO CENTRAL DE ASTURIAS (HUCA)
-INSTITUT D´INVESTIGACIONS BIOMÉDIQUES AUGUST PI I SUNYER (IDIBAPS)
- INSTITUT DE RECERCA DE L'HOSPITAL UNIVERSITARI VALL D'HEBRON
- INSTITUT DE RECERCA HOSPITAL DE LA SANTA CREU I SANT PAU (IRHSCSP)
- PARQUE CIENTÍFICO DE MADRID (PCM)
-UNIVERSITAT DE VALENCIA (UVEG)
-FUNDACIÓN BIOMÉDICA (FICHUVI)

CENTRO NACIONAL DE INVESTIGACIONES ONCOLÓGICAS CARLOS III (CNIO)
DOMINION PHARMAKINE S.L. (DPK)
FUNDACIÓN LEIA, CENTRO DE DESARROLLO TECNOLÓGICO (LEIA, C.D.T.)
HOSPITAL CLÍNICO SAN CARLOS (HCSC)
UNIVERSIDAD DE CÓRDOBA (UCO)
CENTRO NACIONAL DE INVESTIGACIONES CARDIVASCULARES (CNIC)

GE Healthcare Life Sciences
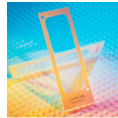
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS (IIB-CSIC)
INSTITUTO VALENCIANO DE INFERTILIDAD I+D, VALENCIA (IVI I+D)
PARQUE CIENTÍFICO DE MADRID (PCM)
FUNDACIÓN BIOMÉDICA (FICHUVI)

Applied Biosystems

CENTRO NACIONAL DE INVESTIGACIONES CARDIOVASCULARES (CNIC)

...estandarización en progresión...

www.mged.org

MIAME

OWG
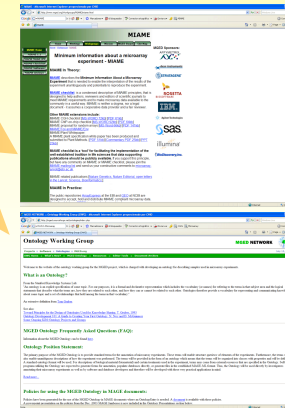
II) Microarrays de proteínas

II) Microarrays de proteínas
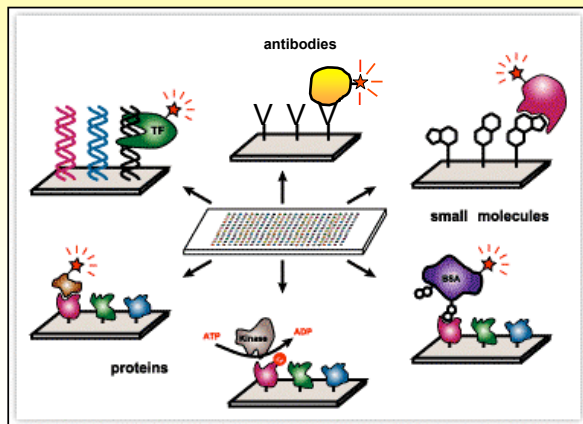
PROTEIN MICROARRAY SYSTEM
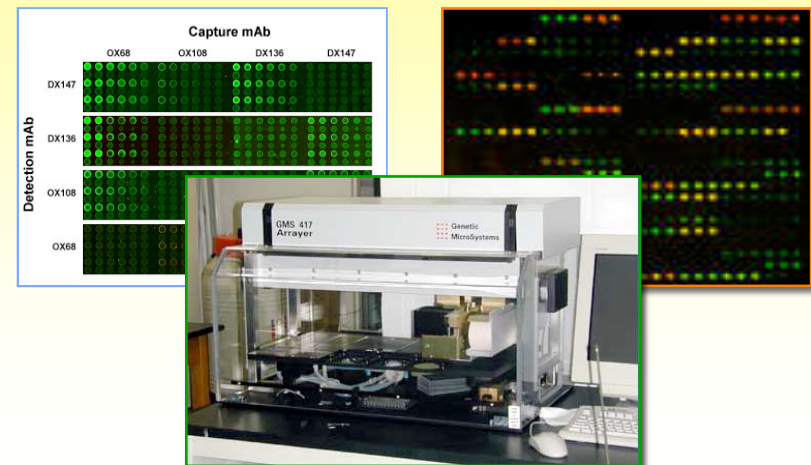
EJEMPLO
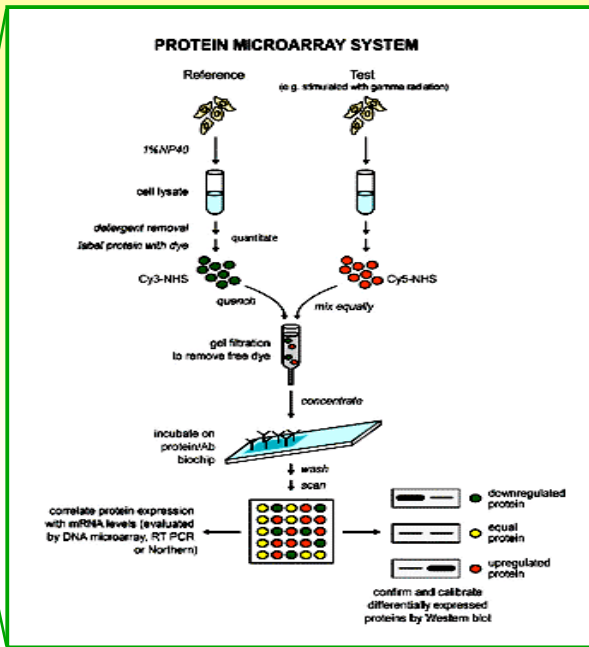
Antibody microarray approaches
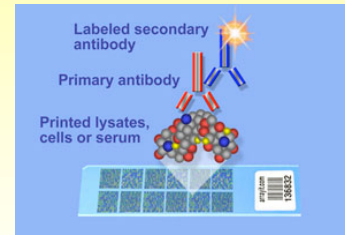
Reverse phase protein arrays
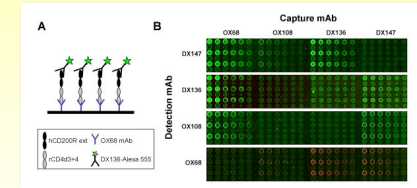
Forward phase protein arrays

III) Microarrays de tejido (*tissue chips*)
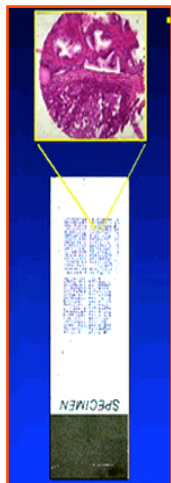
- Mas de 1000 muestras de tejido por array.
- Análisis *in situ* de DNA, RNA y proteínas.

III y IV) Microarrays de tejidos y cultivos celulares

Tissue array

Cell culture array

J. Ziauddin & D. Sabatini Nature, vol 411, May 2001

ChIP on chip

Binding site identification   **Array data analysis**   **Genomic Arrays**

Chromosome CGH   Microarray CGH

Normal DNA   Tumor DNA

Amplification detected at
2p24
2p15

5'SAGE-5'-end serial analysis of gene expression-http://5sage.gi.k.u-tokyo.ac.jp/
ArrayExpress-Public collection of microarray gene expression data-http://www.ebi.ac.uk/arrayexpress
Axeldb-Gene expression in *Xenopus laevis*-http://www.dkfz-heidelberg.de/abt0135/axeldb.htm
BodyMap-Human and mouse gene expression data-http://bodymap.ims.u-tokyo.ac.jp/
BodyMap-Xs-A database for cross-species comparison of vertebrate gene expression-http://bodymap.jp/
BGED-Brain gene expression database-http://genome.mc.pref.osaka.jp/BGED/
CAGE-CAGE tags for cap-analysis of gene expression -http://fantom31p.gsc.riken.jp/cage/mm5/
CleanEx-Expression reference database, linking heterogeneous expression data to facilitate cross-dataset comparisons-http://www.cleanex.isb-sib.ch/
dbERGEII-Database of experimental results on gene expression: Genomic alignment, annotation and experimental data-http://dberge.cse.psu.edu/menu.html
Deniz-Beta-thalassemia allele frequencies in world populations-http://biobase.fatih.edu.tr/deniz1.html
EHCO-Encyclopedia of hepatocellular carcinoma genes online-http://ehco.nchc.org.tw
EICO DB-Expression-based imprint candidate organiser: a database for discovery of novel imprinted genes-http://fantom2.gsc.riken.jp/EICODB/
emap Atlas -Edinburgh mouse atlas: a digital atlas of mouse embryo development and spatially-mapped gene expression -http://genex.hgu.mrc.ac.uk/
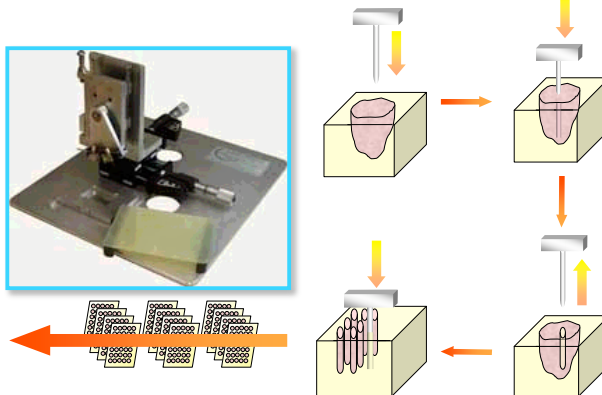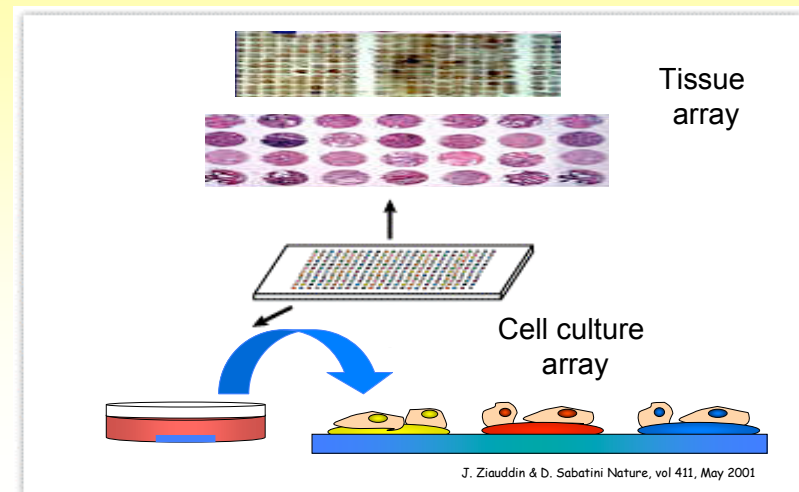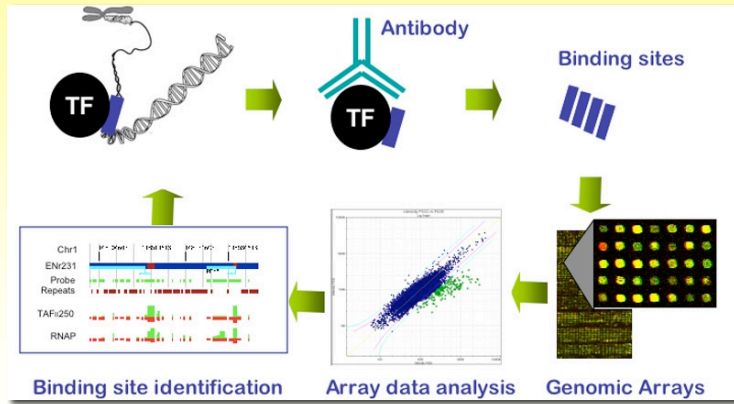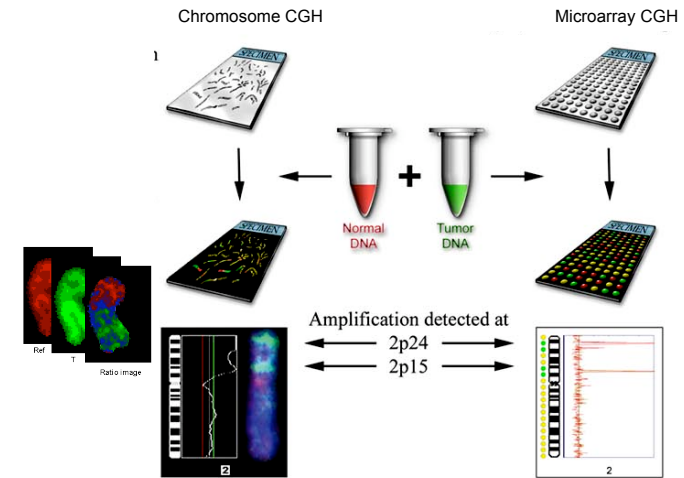**EMAGE**-Edinburgh mouse atlas gene expression database-http://genex.hgu.mrc.ac.uk/Emage/database/emageIntro.html
EPConDB-Endocrine pancreas consortium database-http://www.cbil.upenn.edu/EPConDB
EpoDB-Genes expressed during human erythropoiesis-http://www.cbil.upenn.edu/EpoDB/
FlyView-Drosophila development and genetics-http://pbio07.uni-muenster.de/
GeneAnnot-Revised annotation of Affymetrix human gene probe sets -http://genecards.weizmann.ac.il/geneannot/
GeneNote-Human genes expression profiles in healthy tissues -http://genecards.weizmann.ac.il/genenote/
GenePaint-Gene expression patterns in the mouse-http://www.genepaint.org/Frameset.html
GeneTide-A transcriptome-focused member of the GeneCards suite-http://genecards.weizmann.ac.il/genetide/
GeneTrap-Expression patterns in an embryonic stem library of gene trap insertions-http://www.cmhd.ca/genetrap/
GENSAT-Gene expression nervous system atlas: a map of gene expression in the central nervous system of the mouse-http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gensat
GEO-Gene expression omnibus: Gene expression profiles-http://www.ncbi.nlm.nih.gov/geo/
GermOnline-Gene expression in mitotic and meiotic cell cycle -http://www.germonline.org/
GXD-Mouse gene expression database-http://www.informatics.jax.org/menus/expression_menu.shtml
H-ANGEL-Human anatomic gene expression library-http://www.jbirc.aist.go.jp/hinv/index.jsp
HemBase-Genes expressed in differentiating human erythroid cells -http://hembase.niddk.nih.gov/
HugeIndex-Expression levels of human genes in normal tissues-http://zlab.bu.edu/HugeSearch
IGTC-International mouse Gene Trap Consortium data-http://www.genetrap.org
Kidney Development Database-Kidney development and gene expression-http://golgi.ana.ed.ac.uk/kidhome.html
LOLA-List of lists annotated: a comparison of gene sets identified in different microarray experiments-http://www.lola.gwu.edu/
MAGEST-Ascidian (*Halocynthia roretzi*) gene expression patterns-http://www.genome.ad.jp/magest
MAMEP-Molecular anatomy of the mouse embryo project: Gene expression data on mouse embryos-http://mamep.molgen.mpg.de/
MEPD-Medaka (freshwater fish *Oryzias latipes*) gene expression pattern database-http://www.embl.de/mepd/
MethDB-DNA methylation data, patterns and profiles-http://www.methdb.de/
Mouse SAGE-SAGE libraries from various mouse tissues and cell lines -http://mouse.biomed.cas.cz/sage
NASCarrays-Nottingham *Arabidopsis* Stock Centre microarray database-http://affymetrix.arabidopsis.info
NetAffx-Public Affymetrix probesets and annotations-http://www.affymetrix.com/
OncoMine-Cancer microarray data by gene or cancer type-http://www.oncomine.org/
Osteo-Promoter Database-Genes in osteogenic proliferation and differentiation-http://www.opd.tau.ac.il
PEDB-Prostate expression database: ESTs from prostate tissue and cell type-specific cDNA libraries-http://www.pedb.org/
PEPR-Public expression profiling resource: Expression profiles in a variety of diseases and conditions -http://pepr.cnmcresearch.org
RECODE-Genes using programmed translational recoding in their expression-http://recode.genetics.utah.edu/
RefExA-Reference database for human gene expression analysis-http://www.lsbm.org/db/index_e.html
rOGED-Rat ovarian gene expression database-http://app.mc.uky.edu/kolab/rogedendo.aspx
SAGEmap-NCBI's resource for SAGE data from various organisms-http://www.ncbi.nlm.nih.gov/SAGE
SIEGE-Smoking Induced Epithelial Gene Expression-http://pulm.bumc.bu.edu/siegeDB
Stanford Microarray Database-Raw and normalized data from microarray experiments-http://genome-www.stanford.edu/microarray
TmaDB-Tissue microarray database-http://www.bioinformatics.leeds.ac.uk/tmadb/
Tooth Development Database-Gene expression in dental tissue-http://bite-it.helsinki.fi/

Gene Expression Omnibus (NCBI)

http://www.ncbi.nlm.nih.gov/geo/
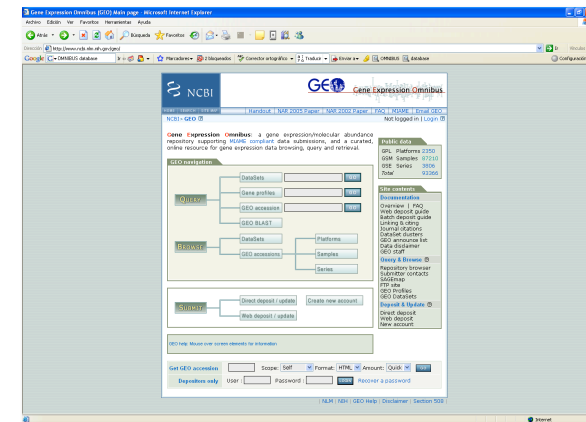
Array Express (EMBL-EBI)
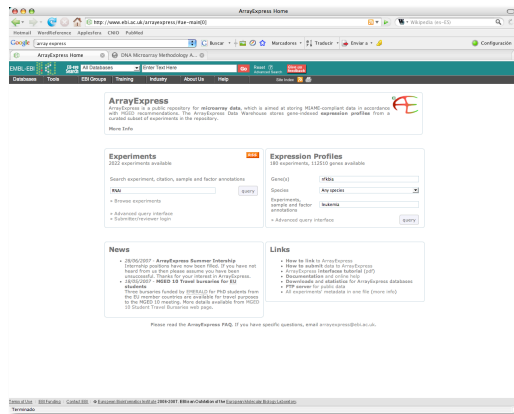
http://www.ebi.ac.uk/arrayexpress/#ae-main[0]

VENTAJAS

1. 1 ensayo ⇒ hasta 30.000 genes
2. Poco tiempo
3. Gran volumen de información
4. Formato
5. "Precio"

LIMITACIONES

1. "Precio"
2. Tratamiento de la muestra
3. Análisis de datos
4. Estandarización
5. Cantidades grandes de mRNA

APLICACIONES

- Búsqueda de biomarcadores y vias celulares
- Rastreo de mutaciones, SNPs...
- Complemento diagnóstico
- Interacciones proteicas
- Perfiles moleculares
- Transfección celular
- Farmacogenómica
- Union Prot-DNA, etc.
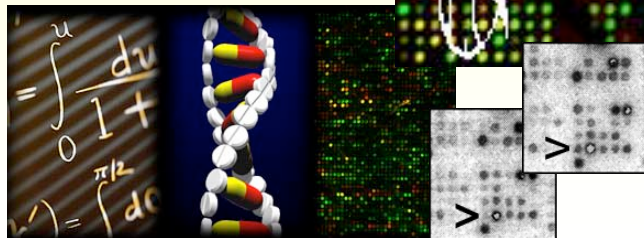
APLICACIONES

¿Cómo están las cosas a día de hoy?

Table 1	Status of microarray-based processes	
Process	Status*
Transcriptional profiling	Mature, but still to be improved
Genotyping	Mature, but still to be improved
Splice-variant analysis	In progress
Identification of unknown exons	Early stages
DNA-structure analysis	Pilot phase
ChIP-on-chip	In progress
Protein binding	Under development
Protein–RNA interaction	Idea
Chip-based CGH	In progress
Epigenetic studies	Under development
DNA mapping	Mature
Resequencing	In progress
Large-scale sequencing	Under development
Gene/genome synthesis	Early stages
RNA/RNAi synthesis	Pilot phase
Protein–DNA interaction	Under development
On-chip translation	Under development
Universal microarray	Under development

*From most to least developed: mature, in progress, under development, early stages, pilot phase, idea. CGH, comparative genomic hybridization; ChIP-on-chip, on-chip chromatin immunoprecipitation.

Nature Reviews, vol 7, March 2006

1) Distinción entre diferentes fenotipos moleculares
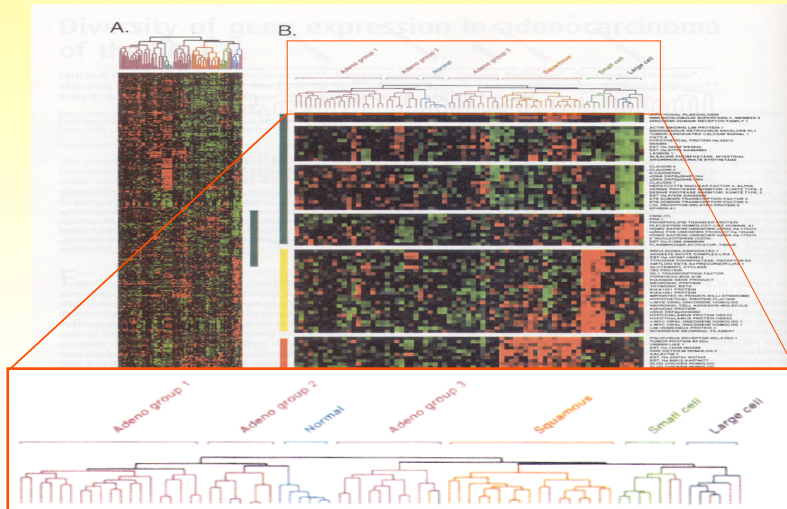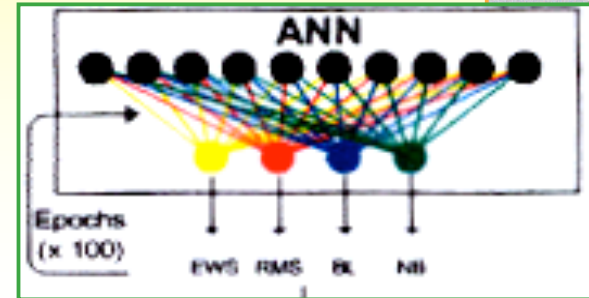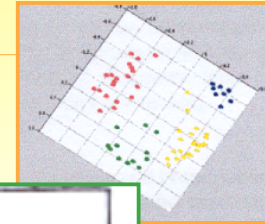

Garber M. E. *et al*. PNAS, vol.98, nº24; 2001.

---

2) Complemento diagnóstico → predicción del tipo tumoral

- 6567 initial genes to 96 finally genes.
- 63 training experiments
- 25 test samples
- ✓ 100% of SRBCTs (5 samples, $p < 0.05$)
- ✓ 100% of non SRBCTs recognized

SRBCTs
Ewing family tumours
Rabdomyosarcoma
Hodgkin lymphoma
Neuroblastoma

Khan et al. NATURE MEDICINE. VOLUME 7 .NUMBER 6 .JUNE 2001

Mas información...

Genoma España	http://www.gen-es.org/
Applied Biosystems	http://www.appliedbiosystems.com/
Affymetrix	http://www.affymetrix.com
Molecular Devices	http://www.moleculardevices.com/
Agilent	www.agilent.com/
CodeLink	http://www4.amershambiosciences.com/aptrix/upp01077.nsf/Content/codelink_activated_slides
Nanogen	http://www.nanogen.com/
Incyte	http://www.incyte.com
Sequenom	http://www.sequenom.com/
Illumina	www.illumina.com/
Lab on a chip	http://www.lab-on-a-chip.com
	http://www.micronit.com/en/about_microfluidics.php
Pat Brown´s Lab	http://cmgm.stanford.edu/pbrown/
DNA microarrays	http://www.dna-microarrays.org/
	http://affymetrix.com/corporate/outreach/educator.affx
	http://www.ebi.ac.uk/microarray/
Protein arrays	http://www.lcsciences.com/custom_peptide_chips.html?gclid=CLqYi9O_kYYCFRBKVgodD3cuug
	http://www.functionalgenomics.org.uk/sections/resources/protein_arrays.htm
	http://genome-www.stanford.edu/proteinarrays/
	http://proteomics.cancer.gov/technology/protein_microarrays.asp
Tissue arrays	http://www.ihcworld.com/tissuearray.htm
	http://www.foliobio.com/
CGH	http://amba.charite.de/cgh/
Artículos	http://www.ii.uib.no/~inge/micro/
	http://ihome.cuhk.edu.hk/~b400559/journalwatch.html
Links	http://www.deathstarinc.com/science/biology/chips.html
	http://www.biocompare.com/jump/192/Microarrays.html
	http://www.gene-chips.com/
	http://www.statsci.org/micrarra/
	http://biowww.net/

Gracias

Gonzalo Gómez López
ggomez@cnio.es