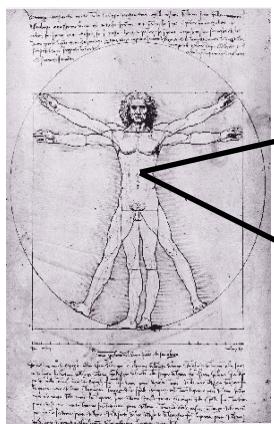


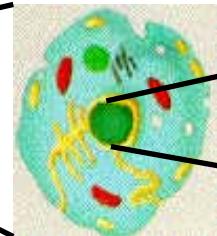
# Biología y Genómica Funcional

- Poseer información completa de todos los genes es muy importante, pero también lo es el estudio de la interacción entre estos genes en el organismo.
- Últimamente los esfuerzos se han concentrado en estudiar las interrelaciones de todos los genes simultáneamente.
- Para esto se requiere una tecnología de medición masiva de la expresión génica.

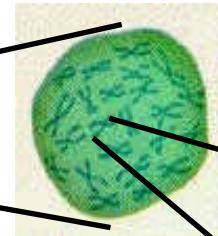
# Biología y Genómica Funcional (II)



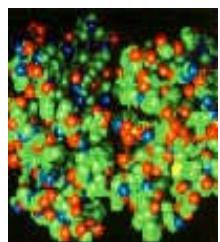
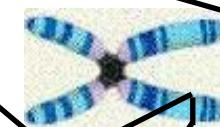
Cell



Nucleus

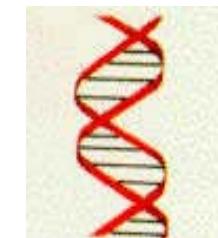
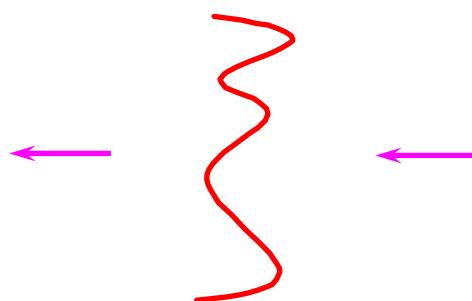


Chromosome



Protein

Gene (mRNA),  
single strand

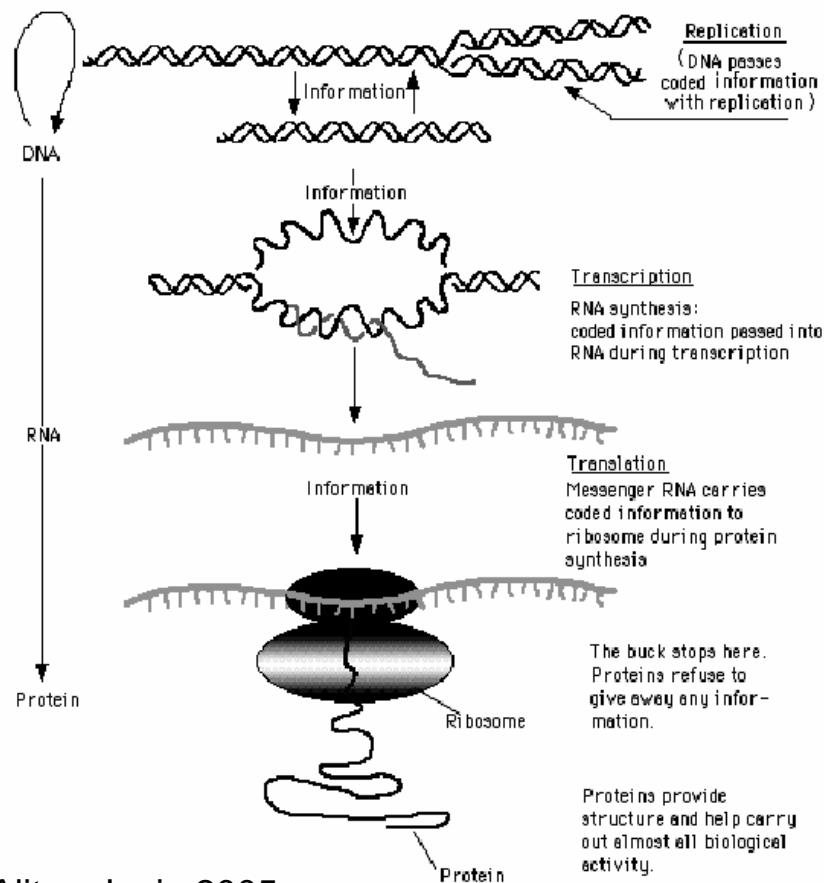


Gene (DNA)

Graphics courtesy of the National Human Genome Research Institute

# El dogma central de la Biología Molecular

## The Central Dogma of Molecular Biology

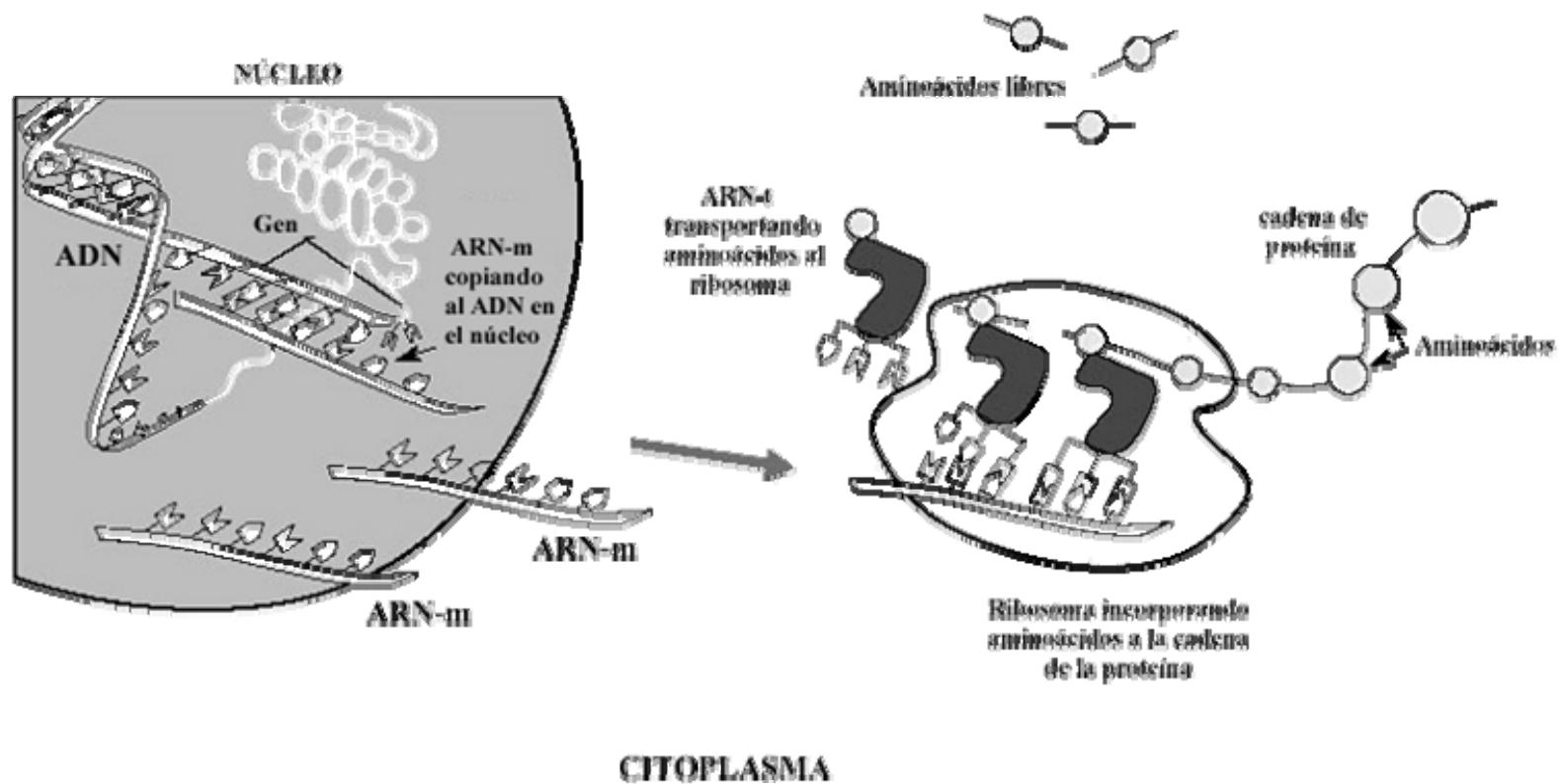


Aliter, Junio 2005.

## Transcripción del DNA al RNA a la proteína:

1. El DNA replica su información en un proceso llamado **replicación**
2. El DNA codifica para la producción del RNA mensajero (mRNA) durante la **transcripción**.
3. El RNA mensajero lleva información codificada a los ribosomas. El ribosoma "lee" esta información y la utiliza para la síntesis de proteínas. A este proceso se le llama **traducción**.

# Expresión génica



# ¿Por qué el estudio de la expresión génica?

- El **patrón de genes expresados** en una célula brinda información del estado actual de la misma.
- Existe una **correlación** entre el estado de la célula y los cambios en los niveles de mRNA de muchos genes.
- Los **patrones de expresion** de genes que nunca han sido caracterizados pueden proporcionar nuevas pistas sobre su posible función.

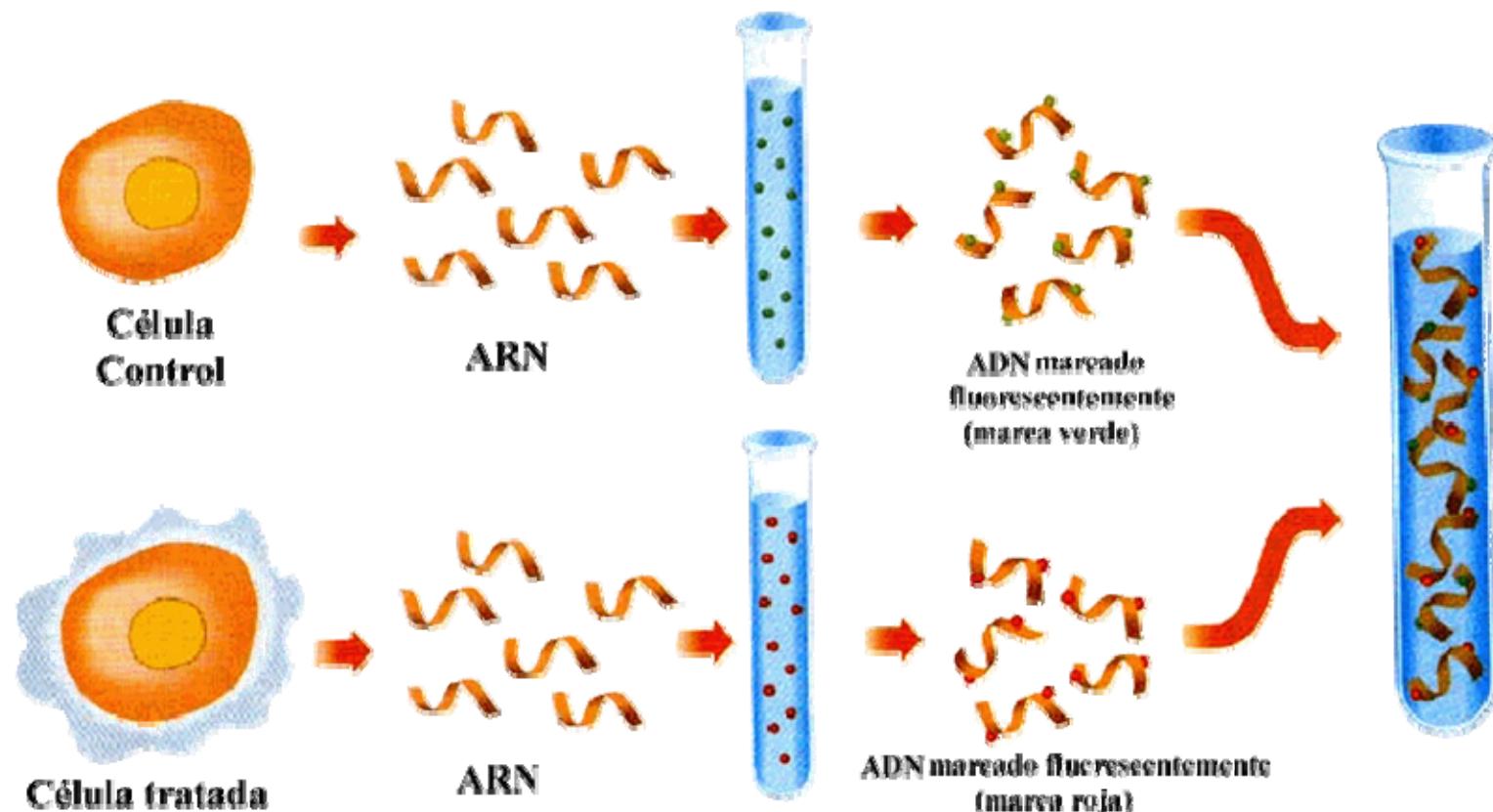
# La tecnología de los DNA microarray

- Es una tecnología concebida para detectar la **expresión de miles de genes simultáneamente**.
- Presenta multitud de aplicaciones potenciales:
  - Identificación de enfermedades genéticas complejas.
  - Estudio toxicológicos
  - Drug discovery
  - Múltiples estudios de expresión de los genes a través del tiempo (distintos tejidos, distintos estadíos de enfermedades...)

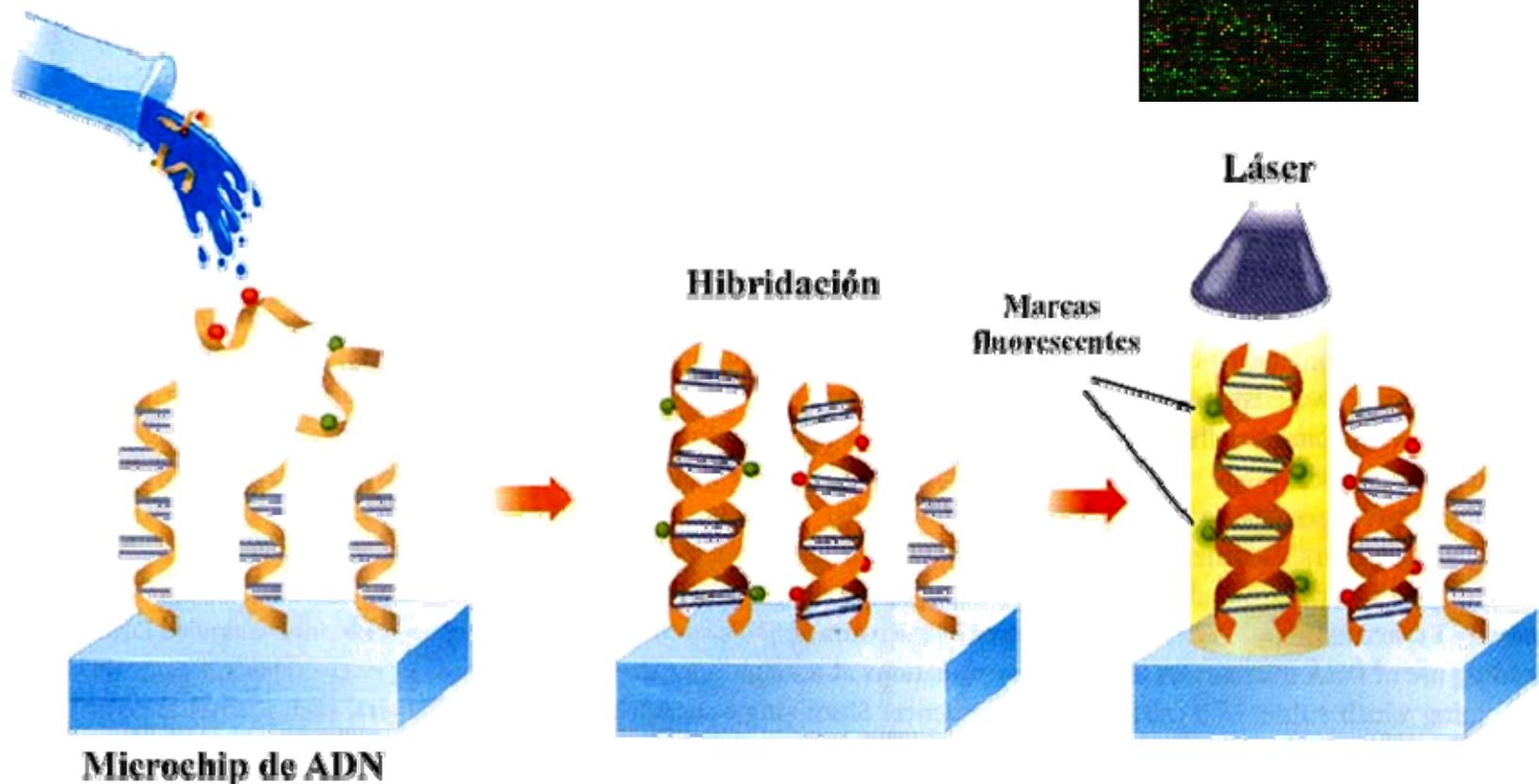
# Principios de Microarrays de ADN

- El ADN complementario a los genes de interés es generado y depositado en posiciones específicas en placas de vidrio.
- El ADN de las muestras a analizar es marcado con sustancias fluorescentes y vertido sobre la superficie de estas placas. Como el el DNA complementario tiende a unirse (hibridación), aquellos genes que se han expresado en la célula se fijarán a su copia en la placa.
- La presencia del DNA expresado se detecta por fluorescencia al excitarse con láser.

## Proceso (I)

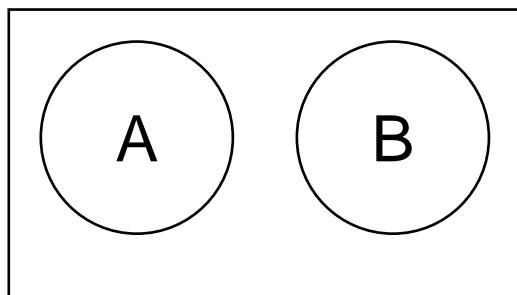


## Proceso (II)



## Proceso (III)

On the surface

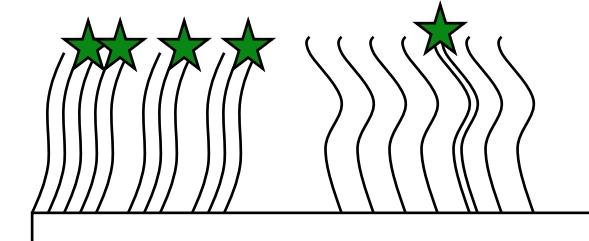
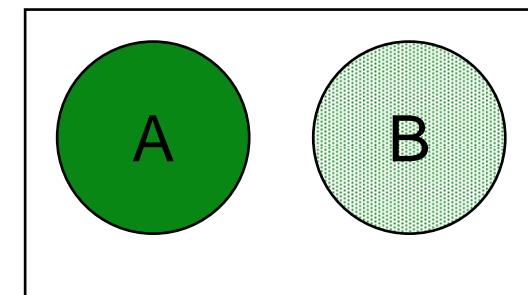


In solution

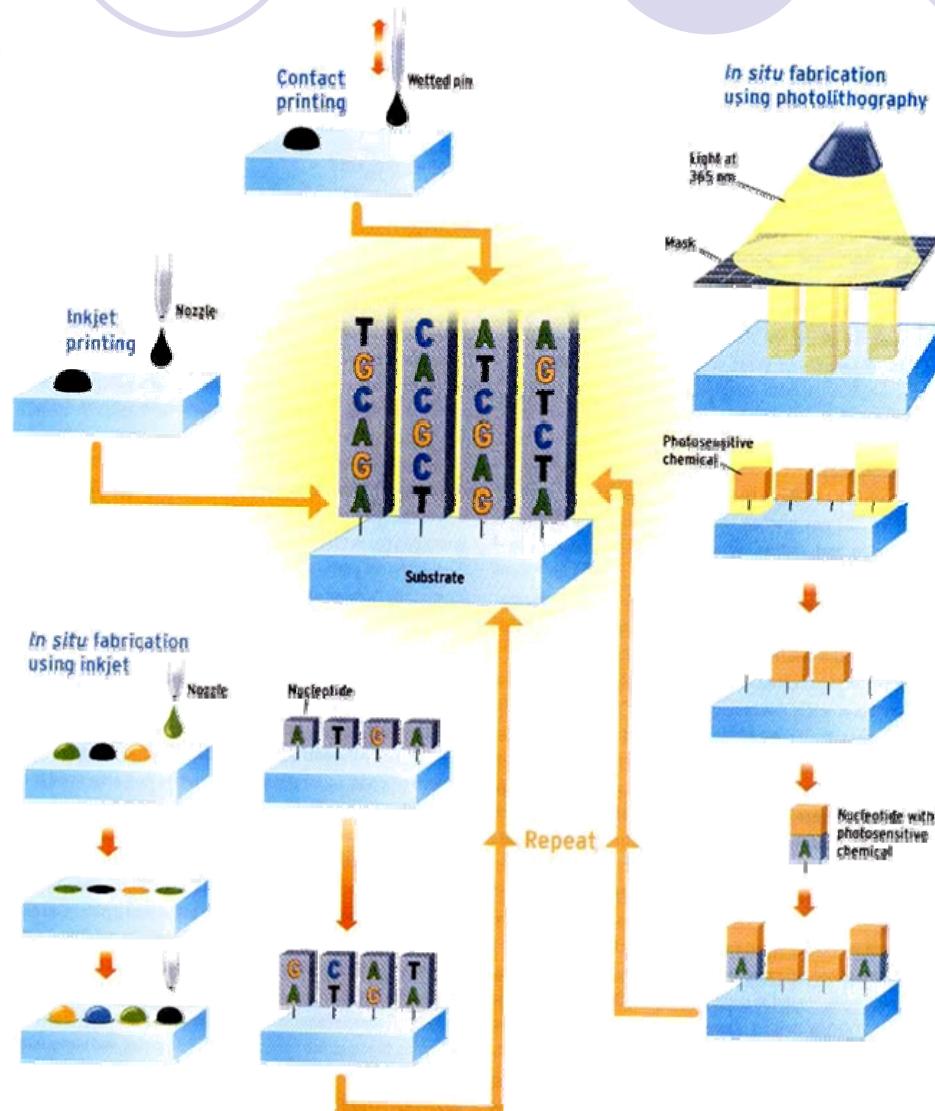
4 copies of gene A,  
1 copy of gene B



After Hybridization



# ¿Cómo se hace el microarray?



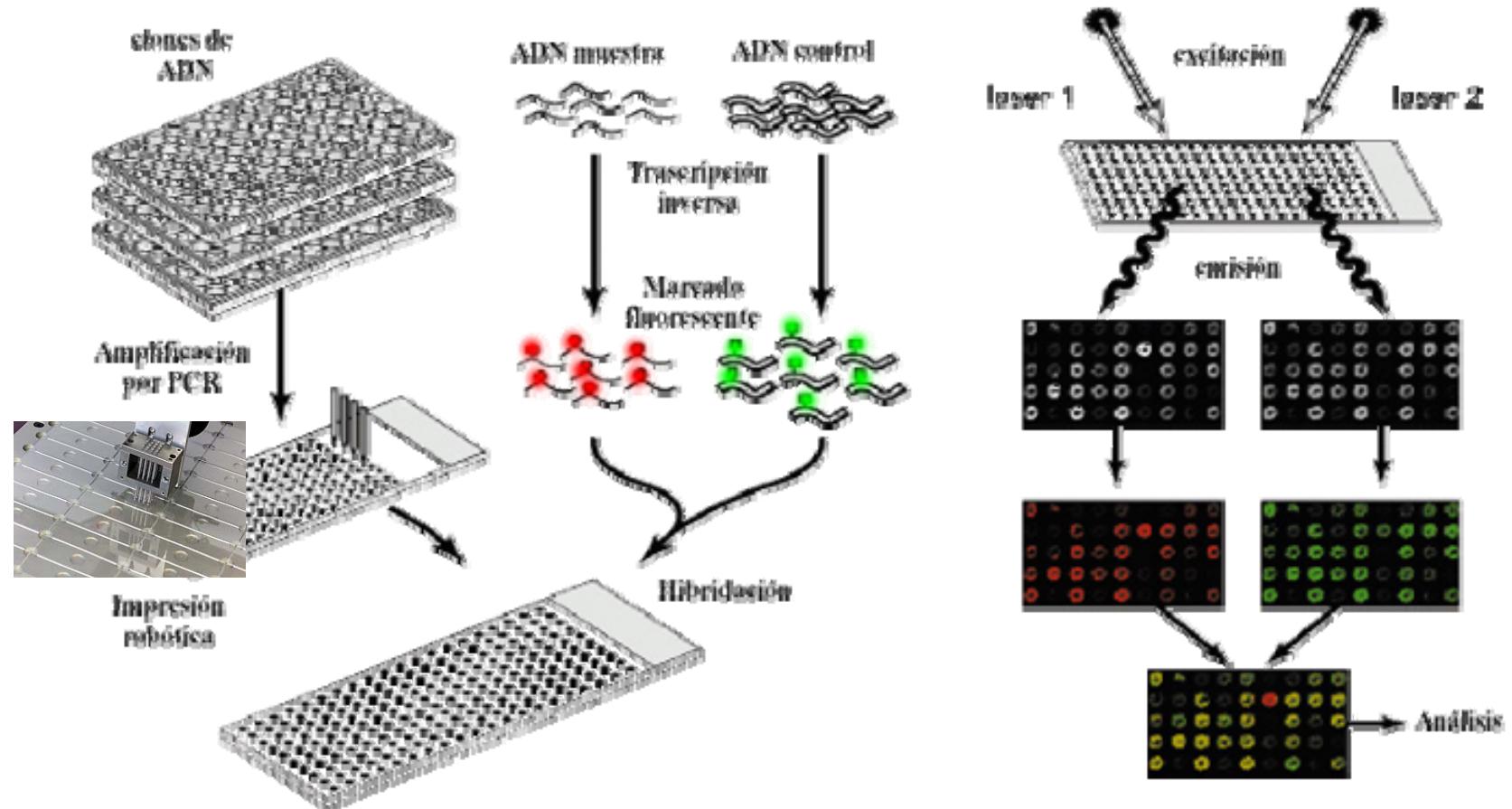
## Tecnologías de Microarrays:

- Short oligonucleotide arrays (**Affymetrix**)
- cDNA or spotted arrays (**Brown/Botstein**).
- Long oligonucleotide arrays (**Agilent Inkjet**)
- Fiber-optic arrays

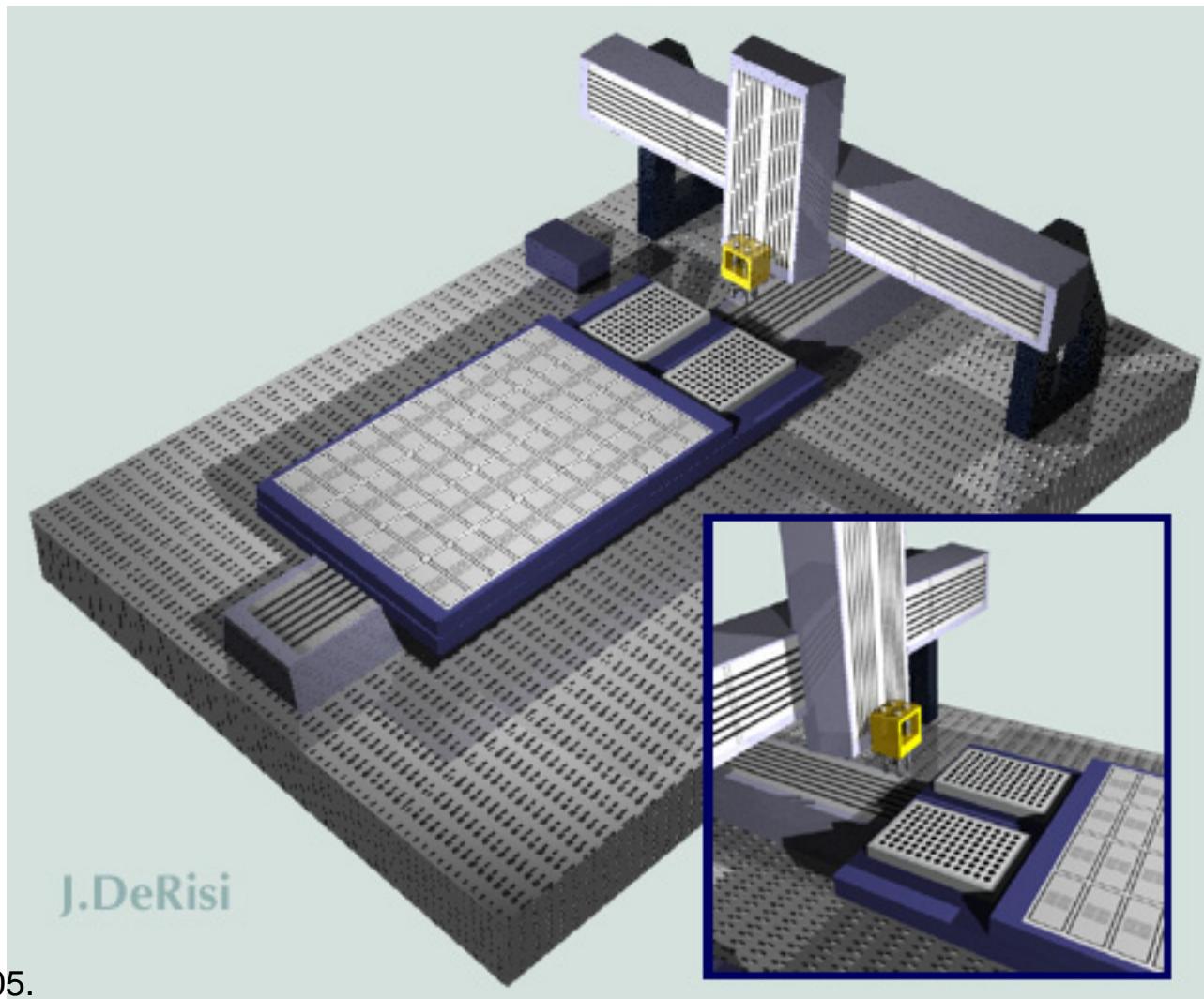
Las tecnologías difieren fundamentalmente en:

- La forma en que el **DNA es depositado en el sustrato** (spotting, lithography, Inkjet printing,...).
- **Longitud de la secuencia** del DNA que es depositada (secuencia completa o fragmentos del gen).
- El tipo de **señal que se mide de cada spot** (e.g. fluorescencia)

# La tecnología del cDNA microarray

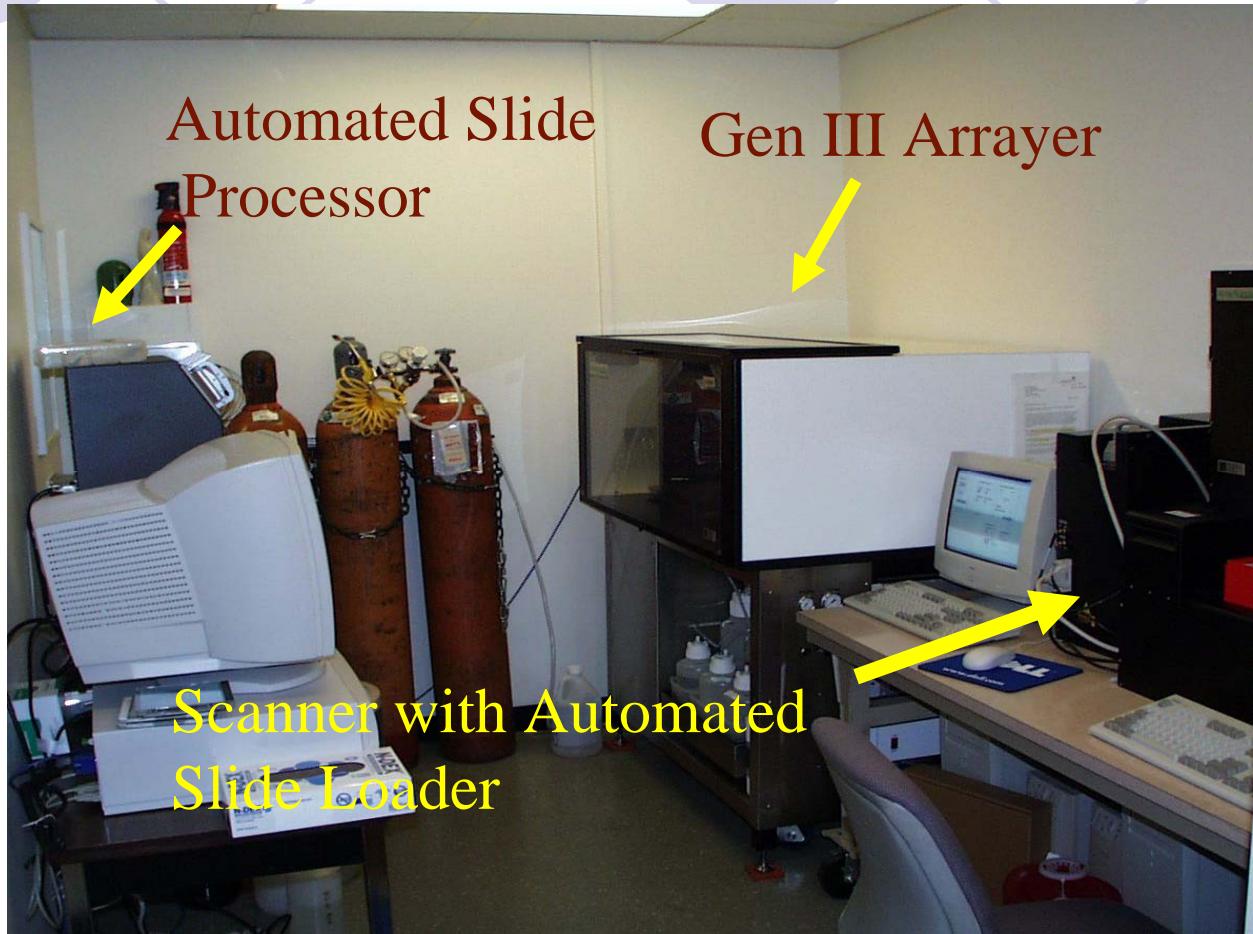


# Arrayer (Robot):



Aliter, Junio 2005.

# Laboratorio de Microarrays

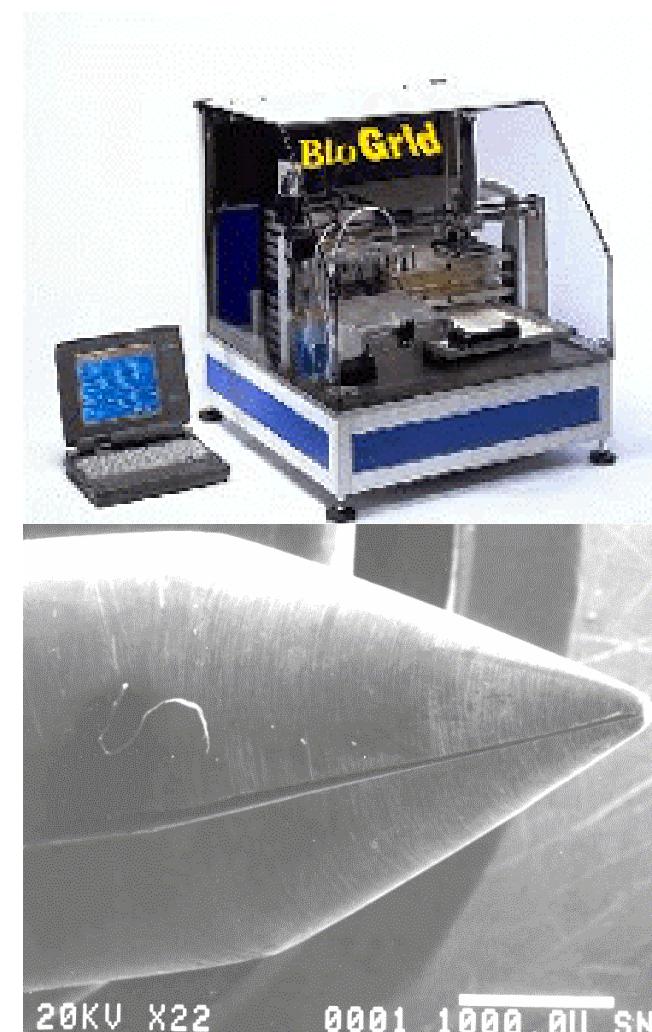


Aliter, Junio 2005.

# Microarray Gridder

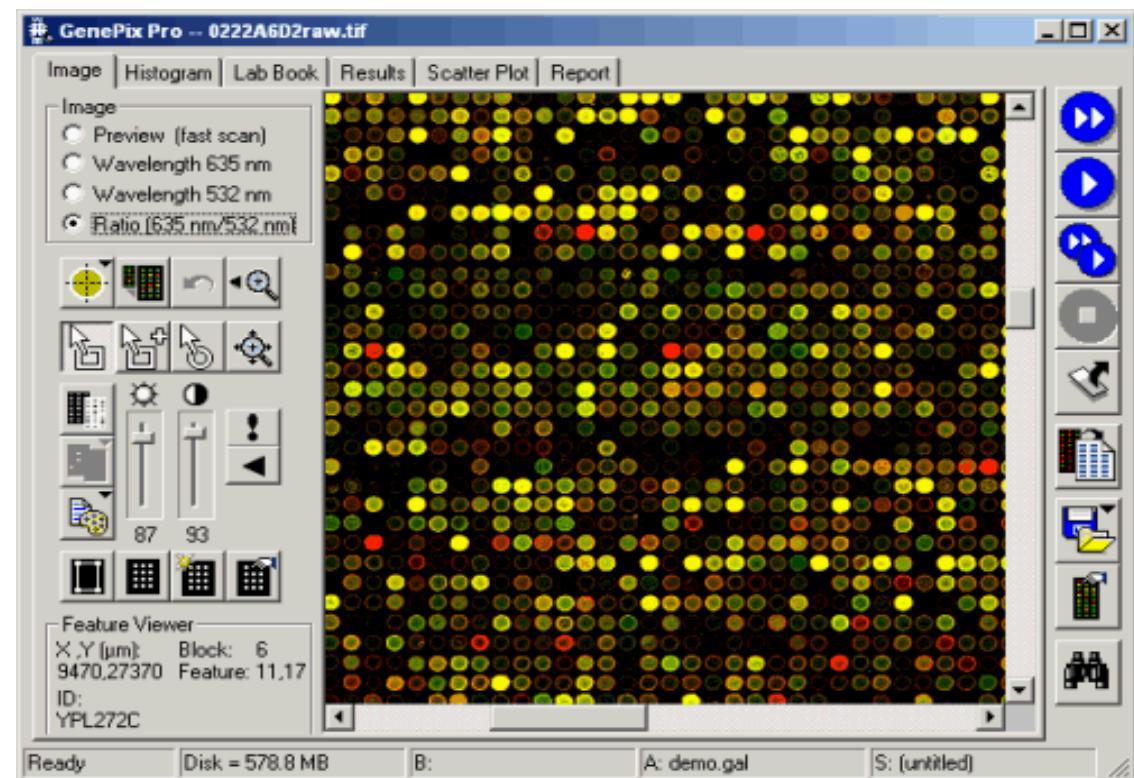


Aliter, Junio 2005.



20KV X22 0001 1000.00 nm

# Scanner

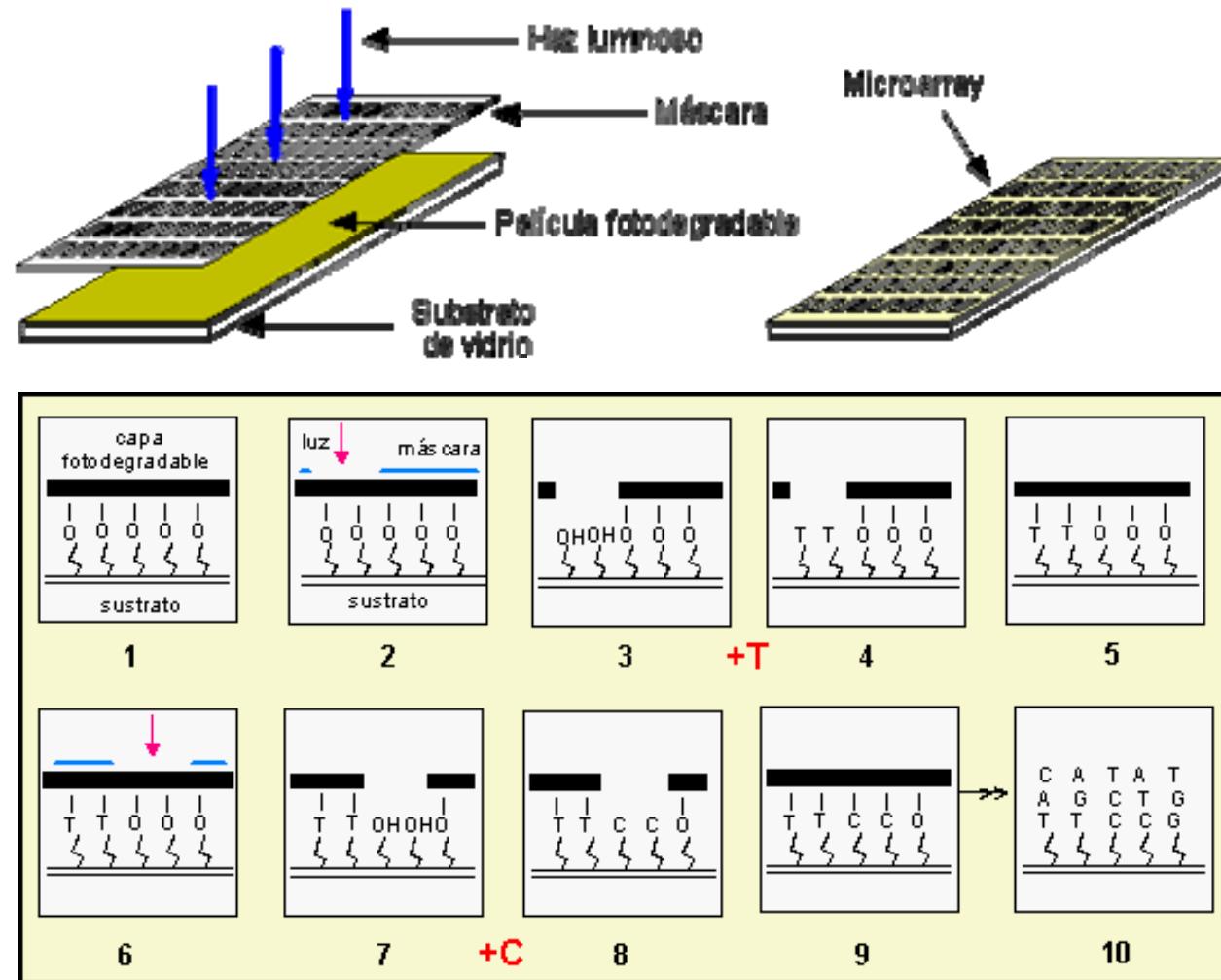


# Tecnología Affymetrix



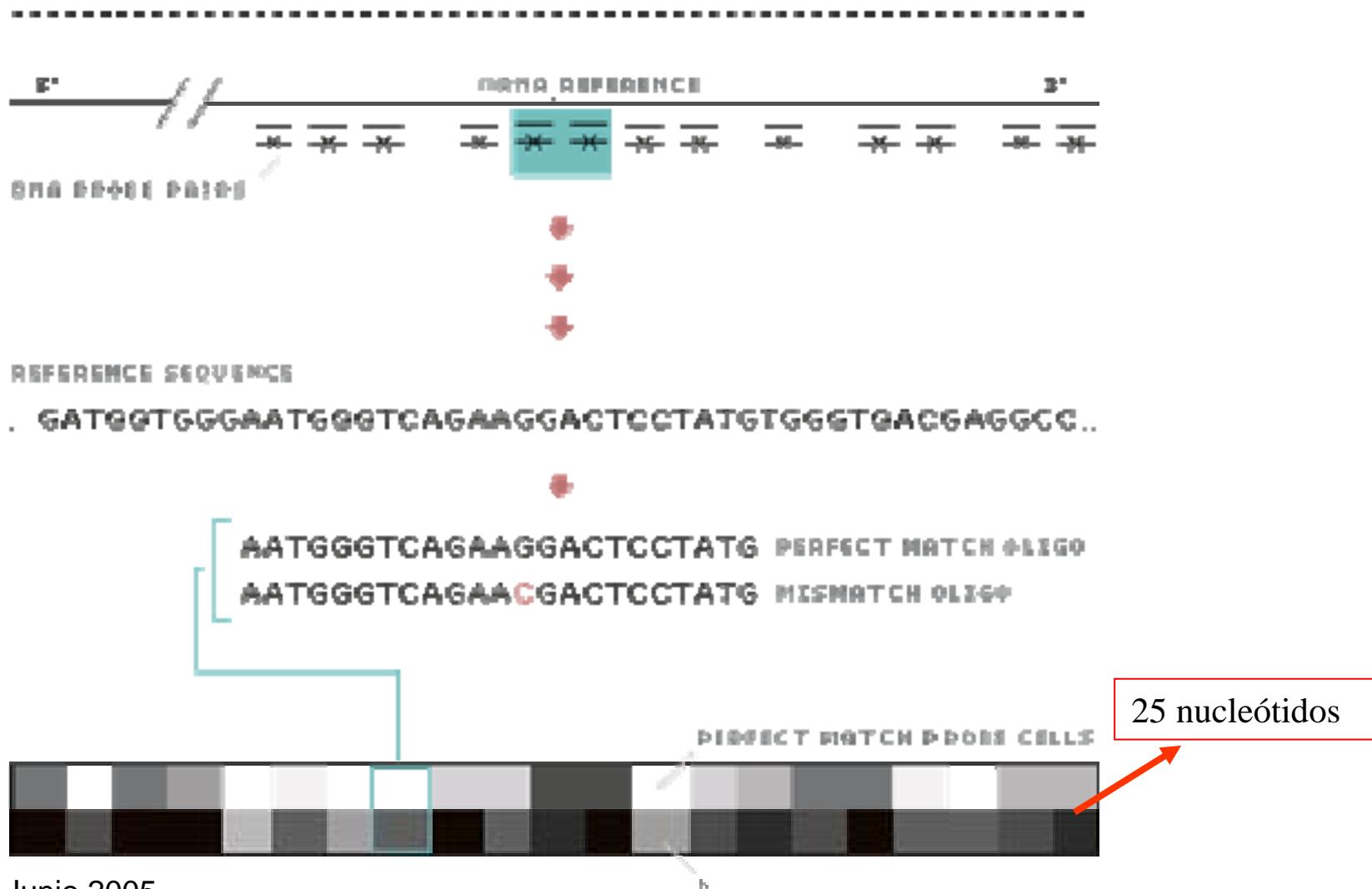
Aliter, Junio 2005.

# Light directed oligonucleotide synthesis:



Aliter, Junio 2005.

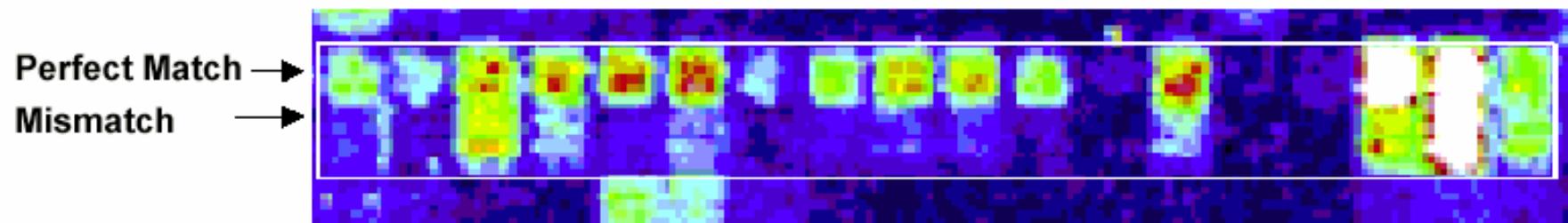
## Multiples “probe pairs” por gen:



## Características del proceso de Affymetrix

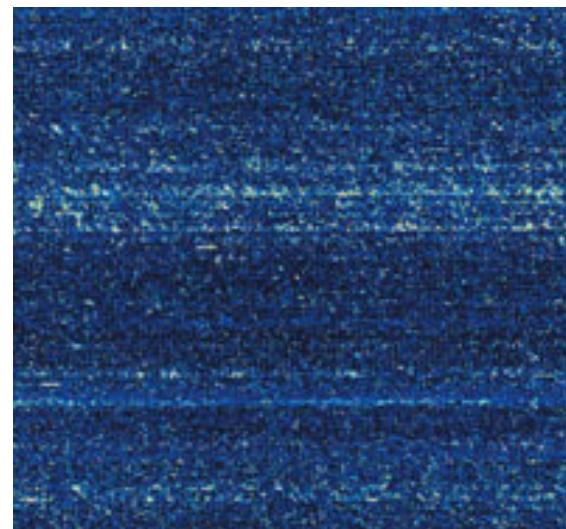
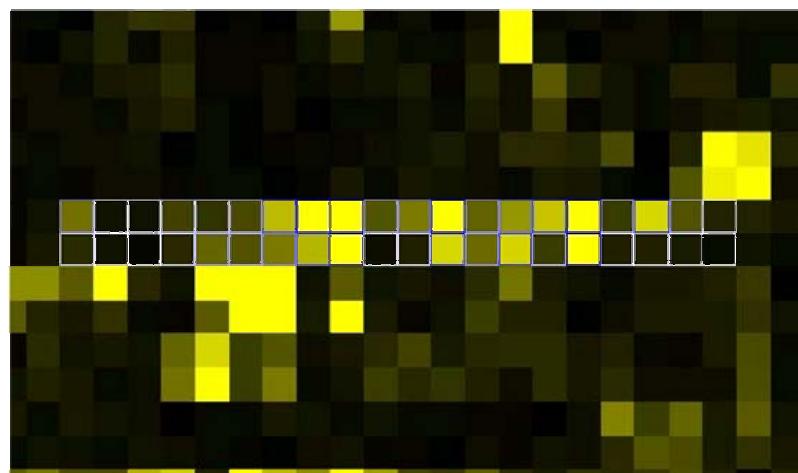
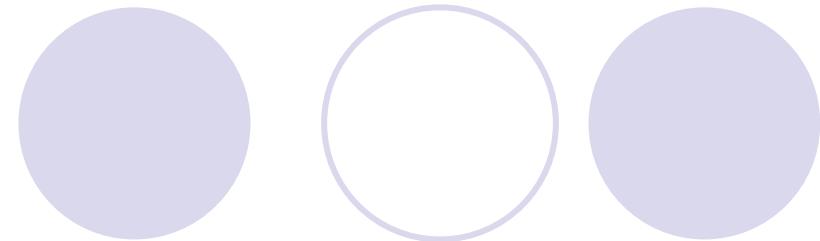
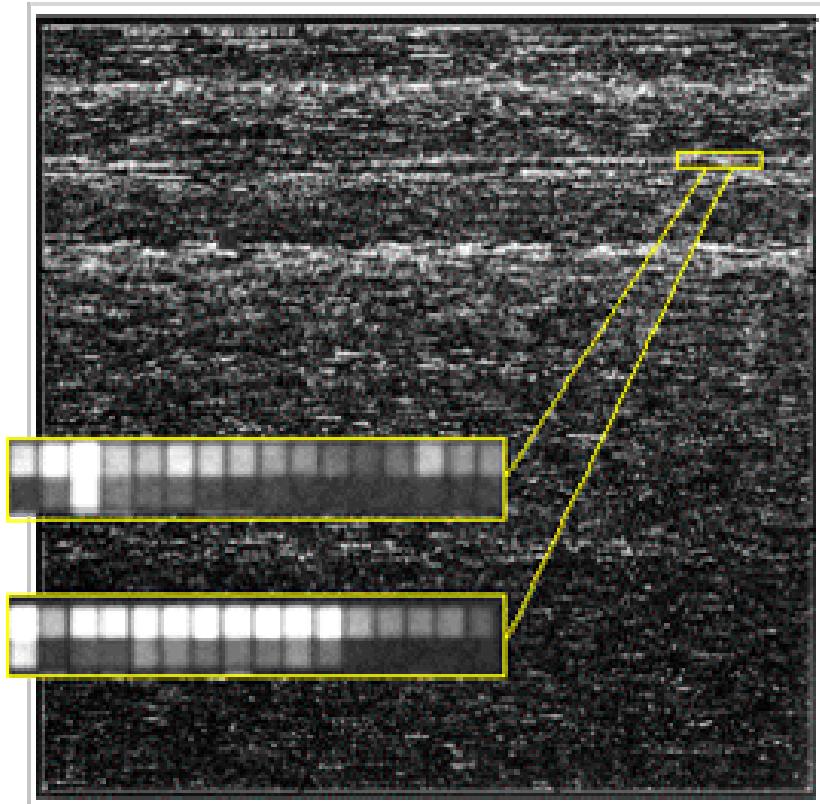
- Cada “probe pair” contiene millones de copias de una secuencia de oligonucleotidos.
- Existe una reducción significativa del número de errores que ocurren por hibridación cruzada.
- Es un método mas “cuantitativo”.
- Coste muy alto

# A Probe Set (DNA Chip)



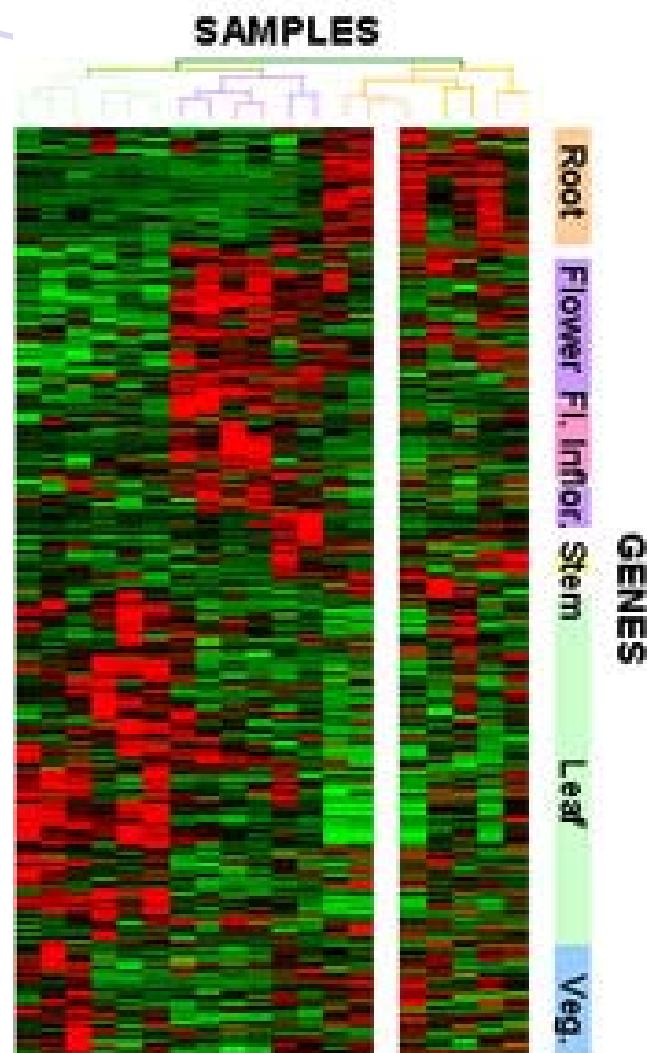
Perfect Match      AGGCTATCGCACTCCAGTGG  
                        AGGCTATCGTACTCCAGTGG  
                        |

## Imagen producida



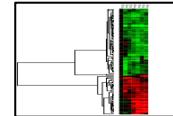
Aliter, Junio 2005.

# Procesamiento de Datos



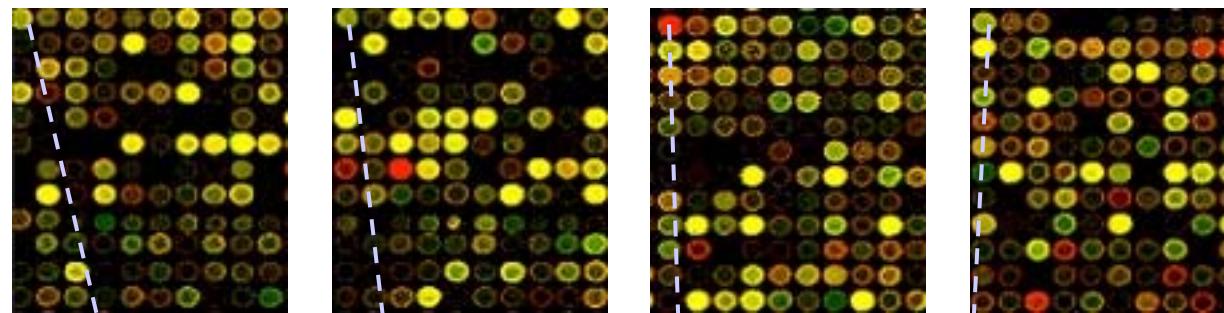
Aliter, Junio 2005.

# Flujo de procesamiento

- Adquisición de datos 
- Almacenamiento 
- Preprocesamiento (Normalización, Duplicados, etc)
- Filtrado
- **Análisis (*agrupamiento, clasificación, predicción, etc*)** 
- Visualización 
- Interpretación/Anotación
- Publicación en repositorio público 



# Generación de los datos Patrón de expresión

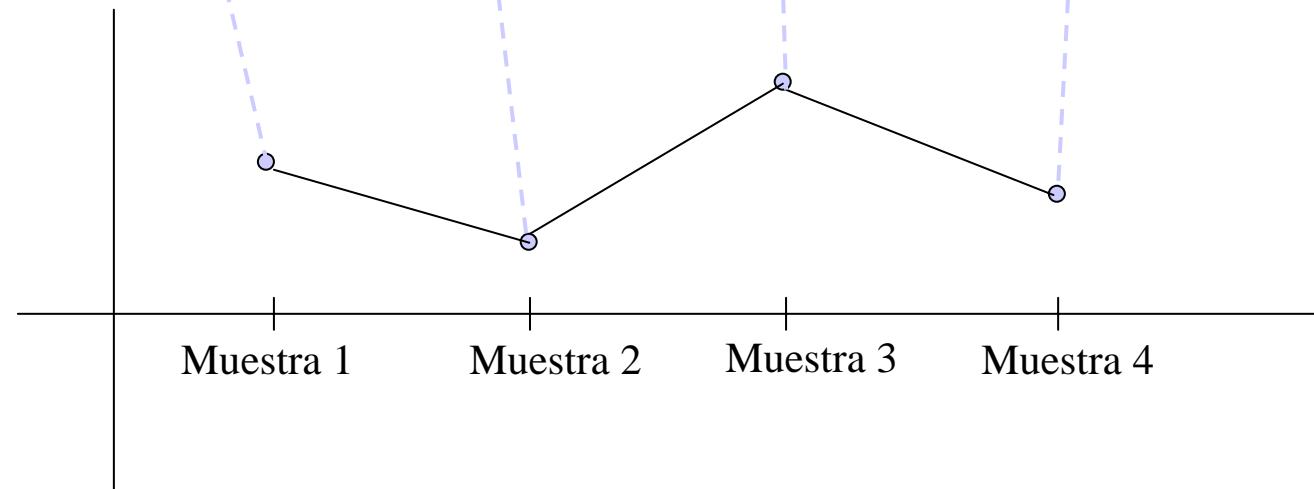


Muestra 1

Muestra 2

Muestra 3

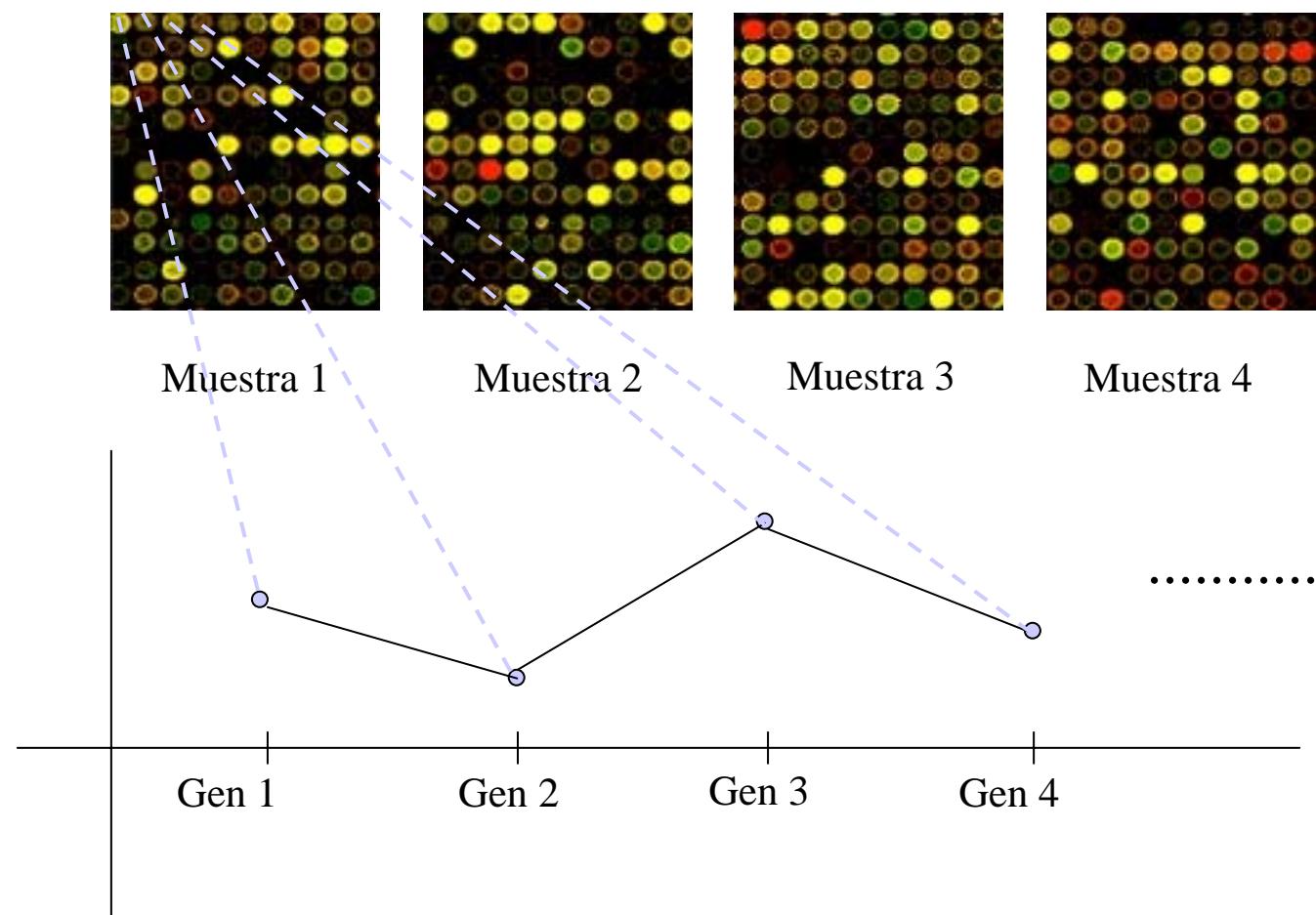
Muestra 4



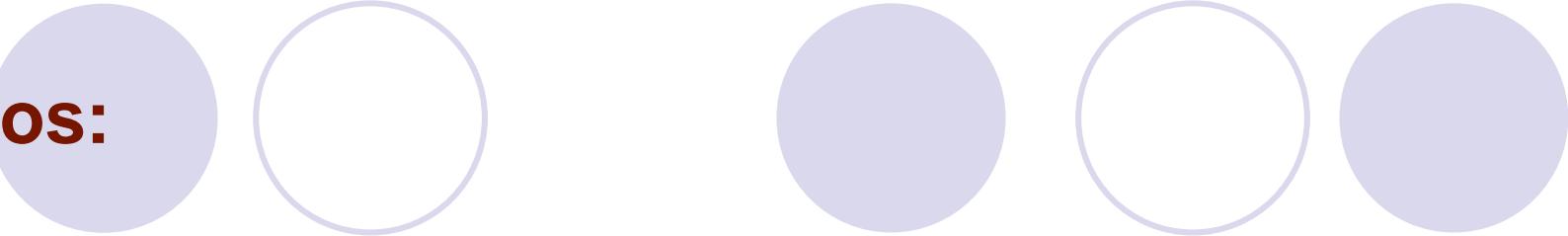
Para cada gen tenemos un perfil de expresión formado por los  $n$  experimentos.

Aliter, Junio 2005.

# Generación de los datos: Por genes (Análisis fenotípico)



Para cada experimento (muestra) tenemos un perfil de expresión (huella molecular)  
Aliter, Junio 2005. formado por  $p$  genes.



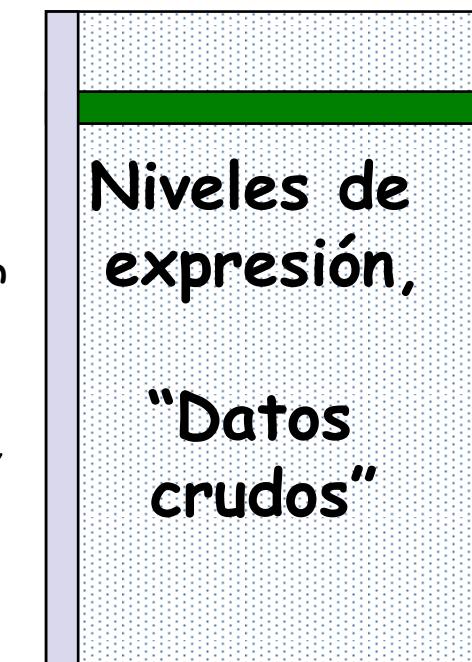
## Datos:

- **Análisis de expresión:**  
 $n$  vectores de  $p$  variables  
 $n$ : número de genes (puntos en el chip)  
 $p$ : número de muestras (número de chips)
- **Análisis fenotípico:**  
 $n$  vectores de  $p$  variables  
 $n$ : número de muestras (número de chips)  
 $p$ : número de genes (puntos en el chip)

## Matriz de datos:

Elementos de la matriz de datos:

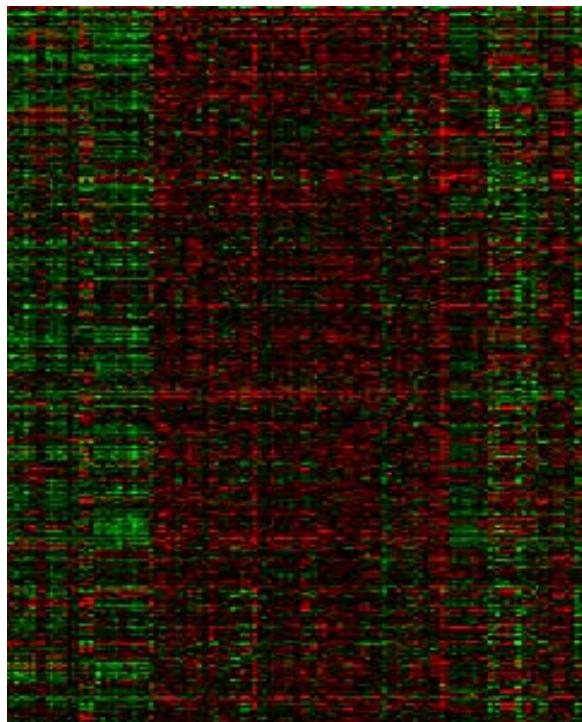
- valores relativos de expresión (ratios)      condiciones →
- valores absolutos
- Distribuciones...
- Fila = patrón de expresión/  
vector huella de un gen
- Columna = perfil condición /  
tejido/ chip



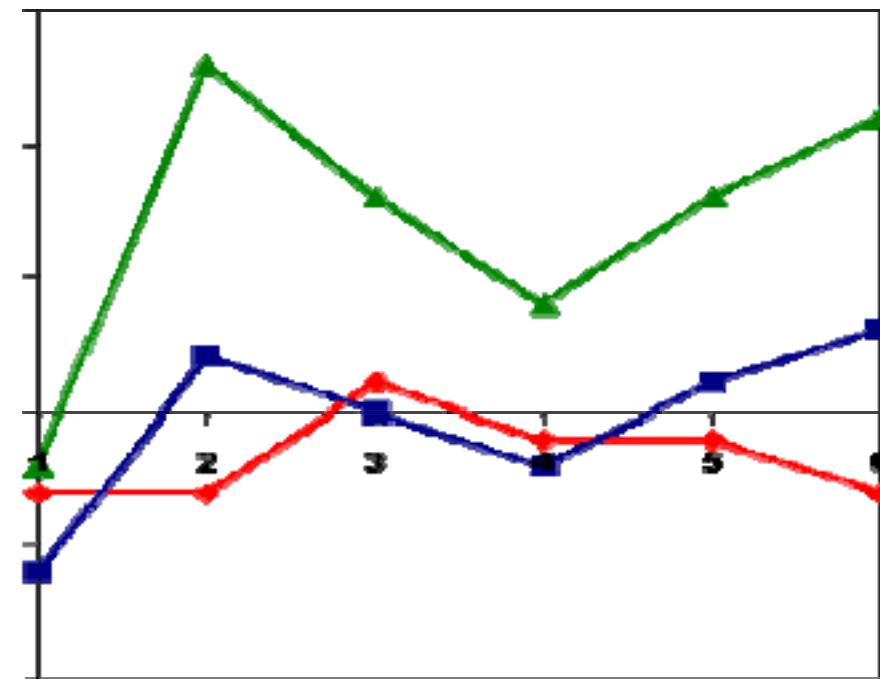
## Matriz de datos:

	Gene Attribute	Exp 1	Exp 2	Exp 3	Exp4	Exp5
Exp Attributes		type I	type II	type III	type II	type III
Gene 1	foo	0.51	0.70	0.88	0.21	0.83
Gene 2	bar	0.35	0.87	0.96	0.22	0.97
Gene 3	blee	0.20	0.06	0.72	0.50	0.99
Gene 4	bas	0.06	0.17	0.37	0.16	0.42
Gene 5	groo	0.54	0.70	0.41	0.86	0.50
Gene 6	gar	0.57	0.28	0.58	0.61	0.58
Gene 7	glee	0.57	0.20	0.45	0.11	0.51
Gene 8	glas	0.52	0.68	0.21	0.43	0.08
Gene 9	gree	0.35	0.91	0.25	0.72	0.67
Gene 10	goe	0.68	0.35	0.25	0.53	0.18

## Visualización:



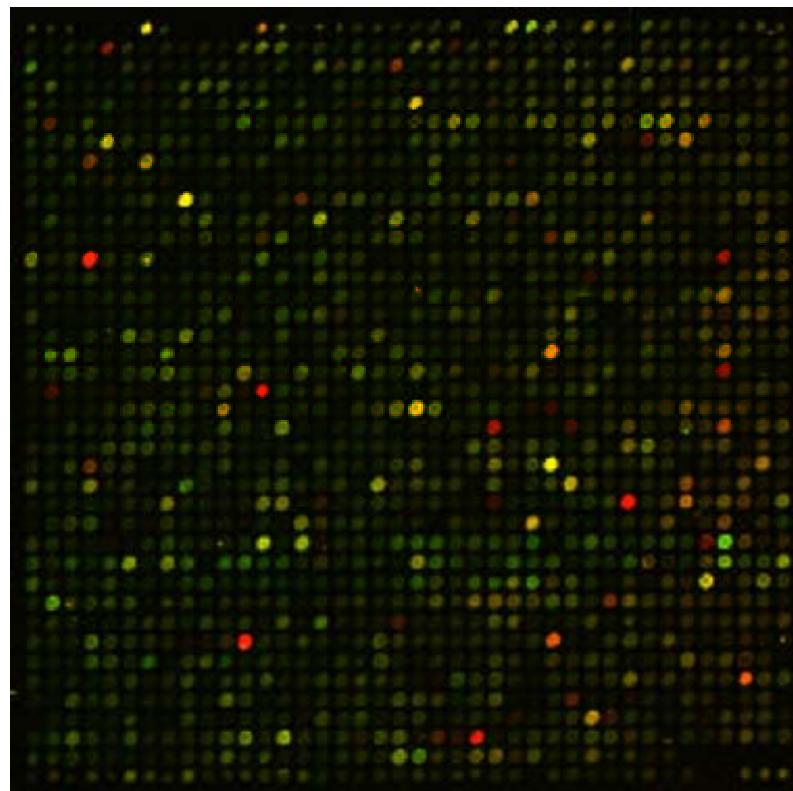
**Filas:** genes  
**Columnas:** muestras  
**Color:** Nivel de expresión



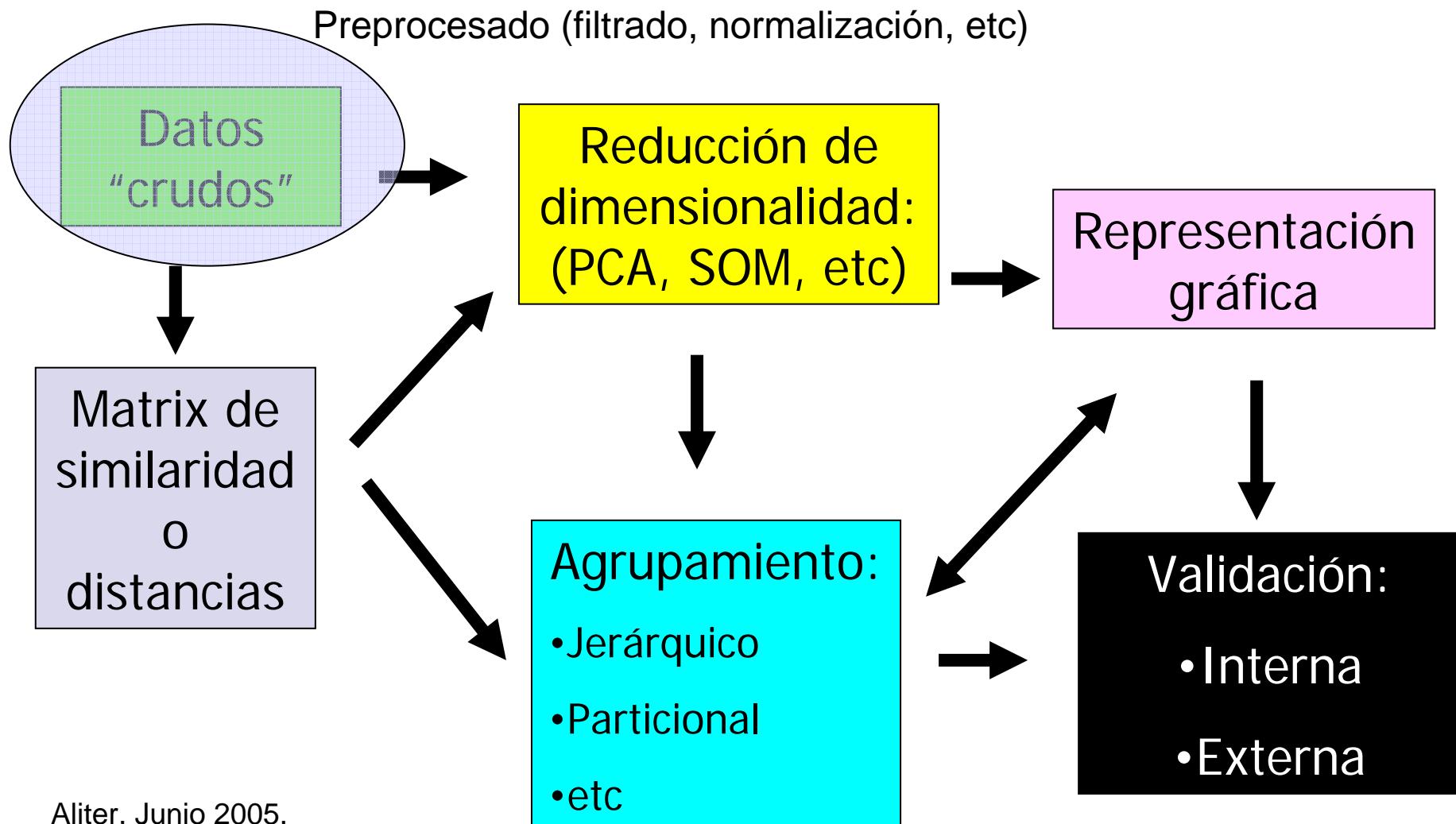
**Eje X:** muestras  
**Eje Y:** nivel de expresión  
**Cada gráfica:** Un gen

## Interpretación de la imagen:

- = more abundant in cell type A
- = more abundant in cell type B
- = equally abundant in both cell types

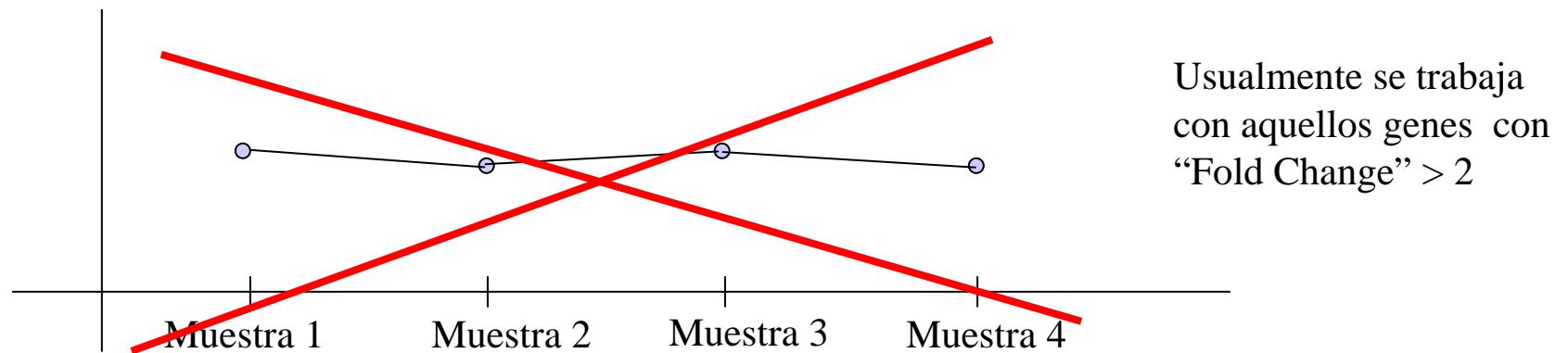


# Aprendizaje no supervisado

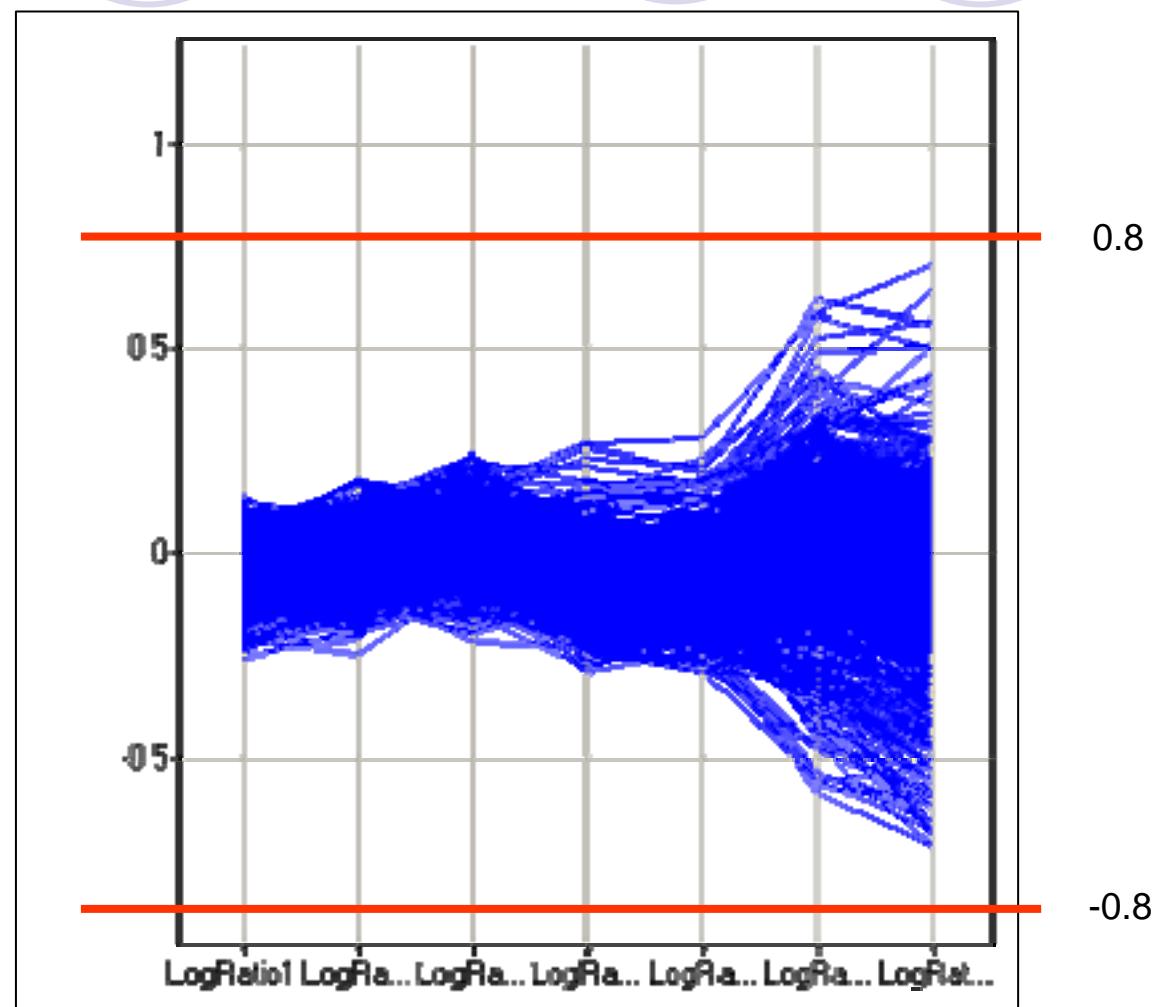


## Filtrado:

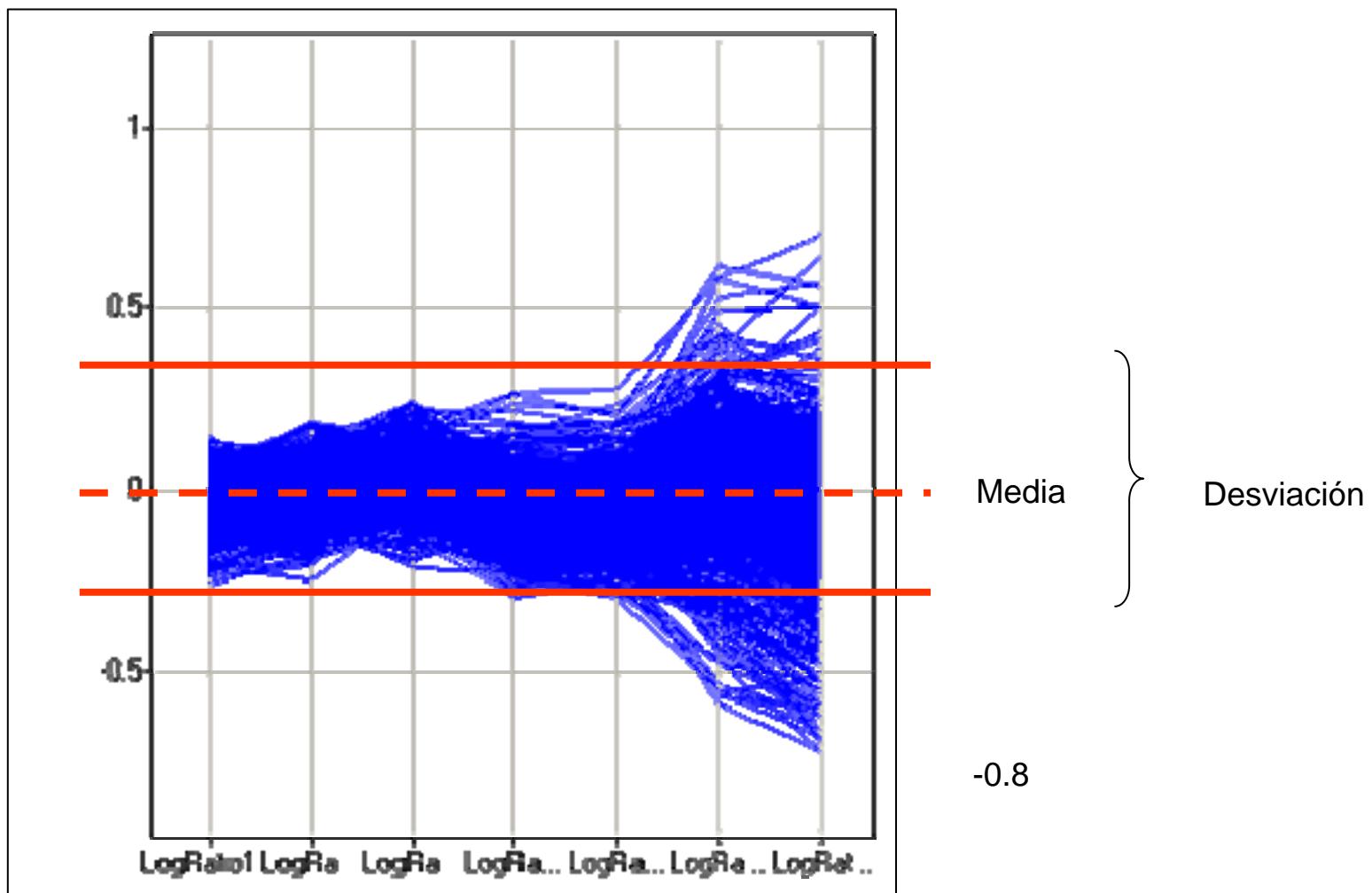
- Los datos de DNA microarray generalmente hay que pre-procesarlos antes de trabajar con ellos:
  - No todos los genes en un chip nos interesan, solo aquellos que hayan variado al menos en una condición experimental.



**Umbral**



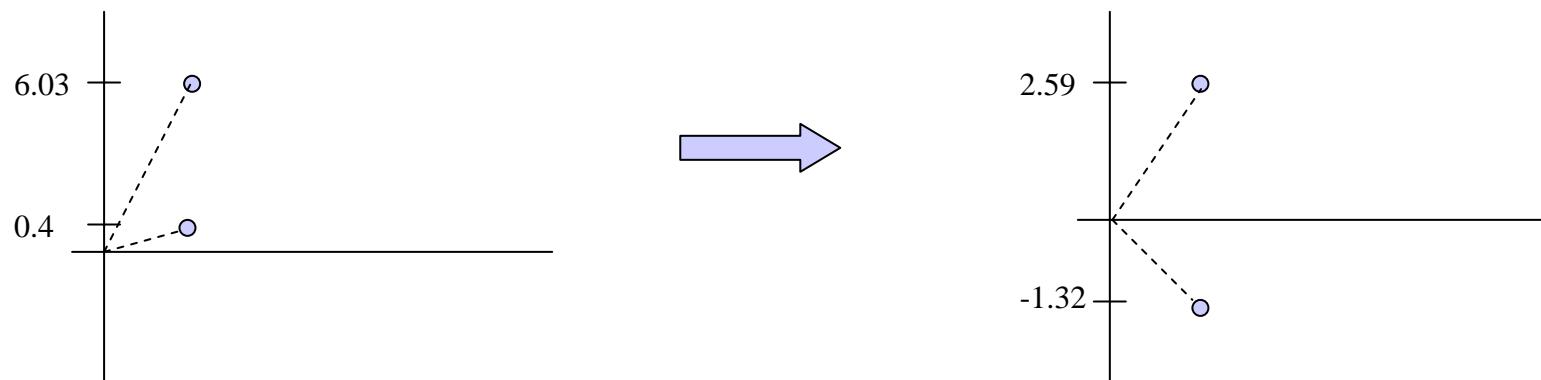
# Desviación



## Transformación logarítmica:

Los datos de expresión generalmente muestran distribuciones asimétricas respecto a la expresión o inhibición, lo cual dificulta el uso de medidas de distancias para establecer diferencias entre ellos. Para compensar estas diferencias, se utiliza generalmente la transformación logarítmica.

Por ejemplo, en cDNA, genes expresados ocupan la escala de 1 a infinito (o al menos 1000-fold), pero los genes inhibidos ocupan solamente la escala de 0 a 1. La transformación logarítmica pone la escala simétrica alrededor del cero.



## Transformación logarítmica:

Muestra/Control:

$$100/1 = 100$$

$$10/1 = 10$$

$$1/1 = 1$$

$$1/10 = 0.1$$

$$1/100 = 0.01$$

Logaritmo:

2

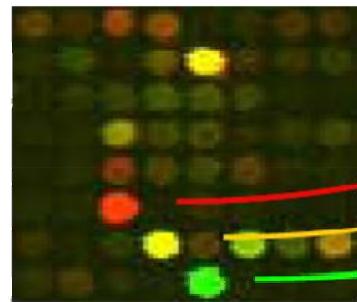
1

0

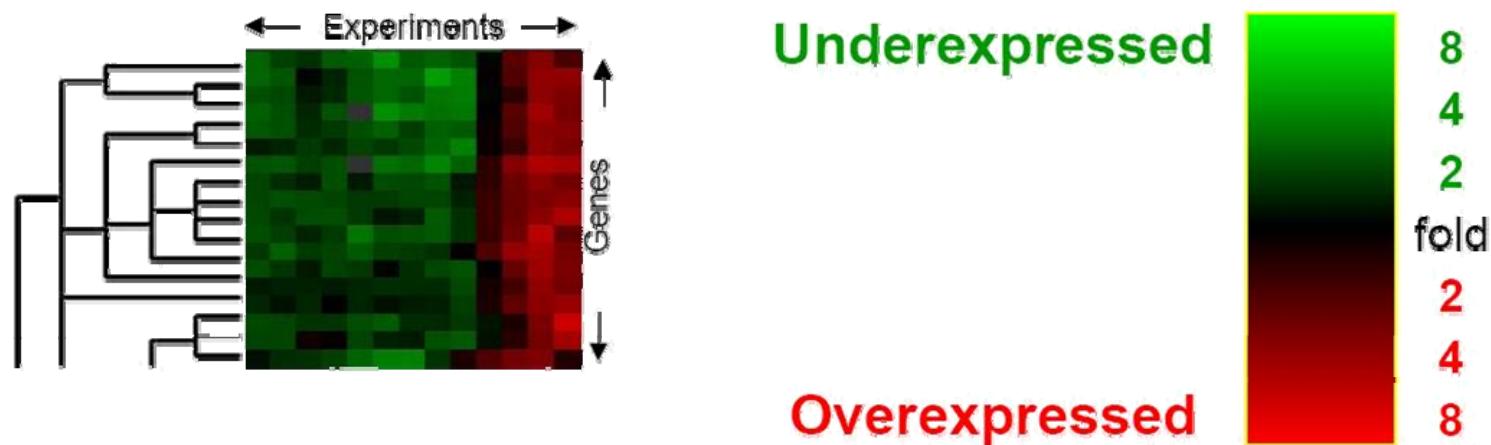
-1

-2

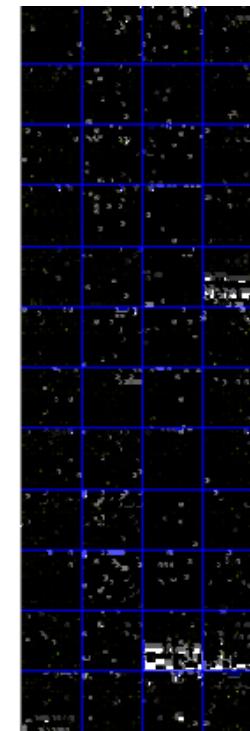
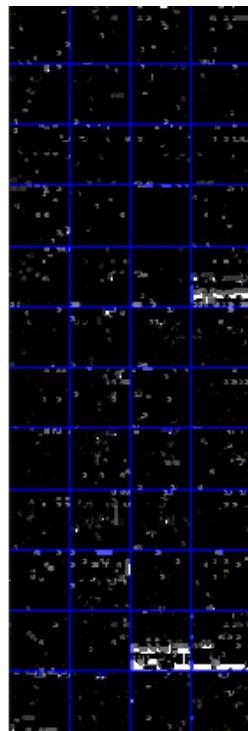
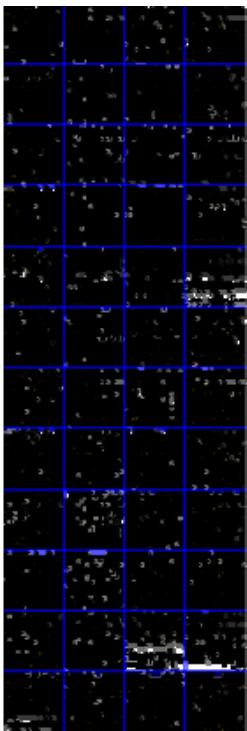
# Representación



Cy3	Cy5	$\frac{\text{Cy5}}{\text{Cy3}}$	$\log_2 \left( \frac{\text{Cy5}}{\text{Cy3}} \right)$	
200	10000	50.00	5.64	Red
4600	4800	1.00	0.00	Black
9000	300	0.03	-4.91	Green



# Valores incompletos (Missing values)



# Tabla

Index	Name	ID	Log2Replicate1	Log2Replicate2	Log2Replicate3
13902	0610005A07Rik   BG072517	mAB0628	0.332	0.447	0.477
5971	0610006A03Rik   BG084074	mAA9100	0.175	0.021	0.095
10740	0610006H08Rik   BG074701	mAB3178			
7981	0610006H10Rik   BG064737	mAA2144	-0.188	-0.039	-0.385
13179	0610006H10Rik   BG064737	mAA2144	-0.12	-0.02	-0.072
4961	0610006I08Rik   BG076213	mAB4974	0.303	0.53	0.358
7442	0610006K04Rik   BG064645	mAA2043	0.307	0.399	0.672
13718	0610006K04Rik   BG064645	mAA2043	-0.123	0.112	
7225	0610006O17Rik   BG086123	mAB1625	-0.004	0.228	0.541
10892	0610007A03Rik   AU023429	mAA6410	0.545	0.508	0.323
16983	0610007H07Rik   BG085143	mAB0381	0.286	0.282	
6906	0610007L03Rik   BG069702	mAA7510			
12509	0610007L05Rik   BG087267	mAB3112	0.215	0.192	0.318
8006	0610007N03Rik   BG077453	mAA1400	-0.053	-0.042	-0.157
13206	0610007N03Rik   BG077453	mAA1400	0.043	0.034	0.021
7235	0610007N19Rik   BG073658	mAB2057	0.432	0.339	1.176
5641	0610007T007Rik   BG079781	mAA4228	0.381	0.497	-0.519
6608	0610007P06Rik   BG063045	mAA0046	0.453	0.475	0.224
14640	0610007P06Rik   BG063045	mAA0046			
20029	0610007P06Rik   BG086880	mAB2552	0.314	0.349	0.495
8593	0610008C08Rik   BG071900	mAA9933			
4034	0610008F14Rik   BG076975	mAA0805	0.445	0.614	0.041
17162	0610008F14Rik   BG076975	mAA0805	0.46	0.672	-0.003
18305	0610008K04Rik   BG085113	mAB0350	0.04	0.014	0.276
6801	0610008N23Rik   BG086860	mAB2526	0.087	0.153	-0.007
5974	0610009C03Rik   BG071176	mAA9082	0.368	0.198	0.188
9383	0610009D07Rik   BG076503	mAB5243	0.392	0.363	0.036
11483	0610009D07Rik   BG076503	mAB5243	0.292	0.313	0.098
2643	0610009D10Rik   BG063439	mAA0605		0.479	
18621	0610009D10Rik   BG063439	mAA0605	0.259	0.372	-0.099
2588	0610009D10Rik   BG077769	mAA1787	-0.223	-0.157	-0.496
18598	0610009D10Rik   BG077769	mAA1787	-0.203	-0.032	-0.667
9878	0610009D16Rik   BG086417	mAB1991	0.217	0.065	
7466	0610009E20Rik   BG077106	mAA0945	0.539	0.59	0.41
13772	0610009E20Rik   BG077106	mAA0945	0.25	0.362	0.194
9084	0610009H04Rik   BG083210	mAA8139	0.178	0.246	0.193
8568	0610009H04Rik   BG087149	mAB2975	0.126	0.187	
13812	0610009J22Rik   BG075746	mAB4420	-0.138	-0.029	0.079
1063	0610009M14Rik   BG074040	mAB2401			
8888	0610009M14Rik   BG077537	mAA1508	-0.062	0.083	-0.608
12324	0610009M14Rik   BG077537	mAA1508	-0.128	-0.074	0.138
4911	0610009N12Rik   BG065259	mAA2695	0.199	0.251	-0.047
16239	0610009N12Rik   BG065259	mAA2695	0.258	0.445	0.149