



Laboratorio de Bioinformática y  
Expresión Génica  
INTA – Universidad de Chile

**Curso BT51B – Facultad de Ciencias Físicas y Matemáticas:**

# Estrategias de análisis en experimentos de expresión génica

***Ing. Christian Hödar Q.***

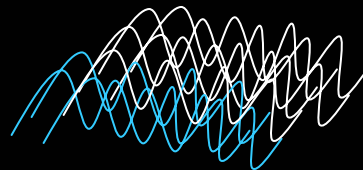
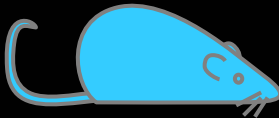
# Experimentos de expresión génica

¿Qué medimos?

Condición A



Condición B



Cambios en los  
niveles de mRNA  
transcrito

¿Cómo lo medimos?

-Northern Blot

-PCR tiempo real

-SAGE

(Serial Analysis of Gene expression)

**-Microarrays**

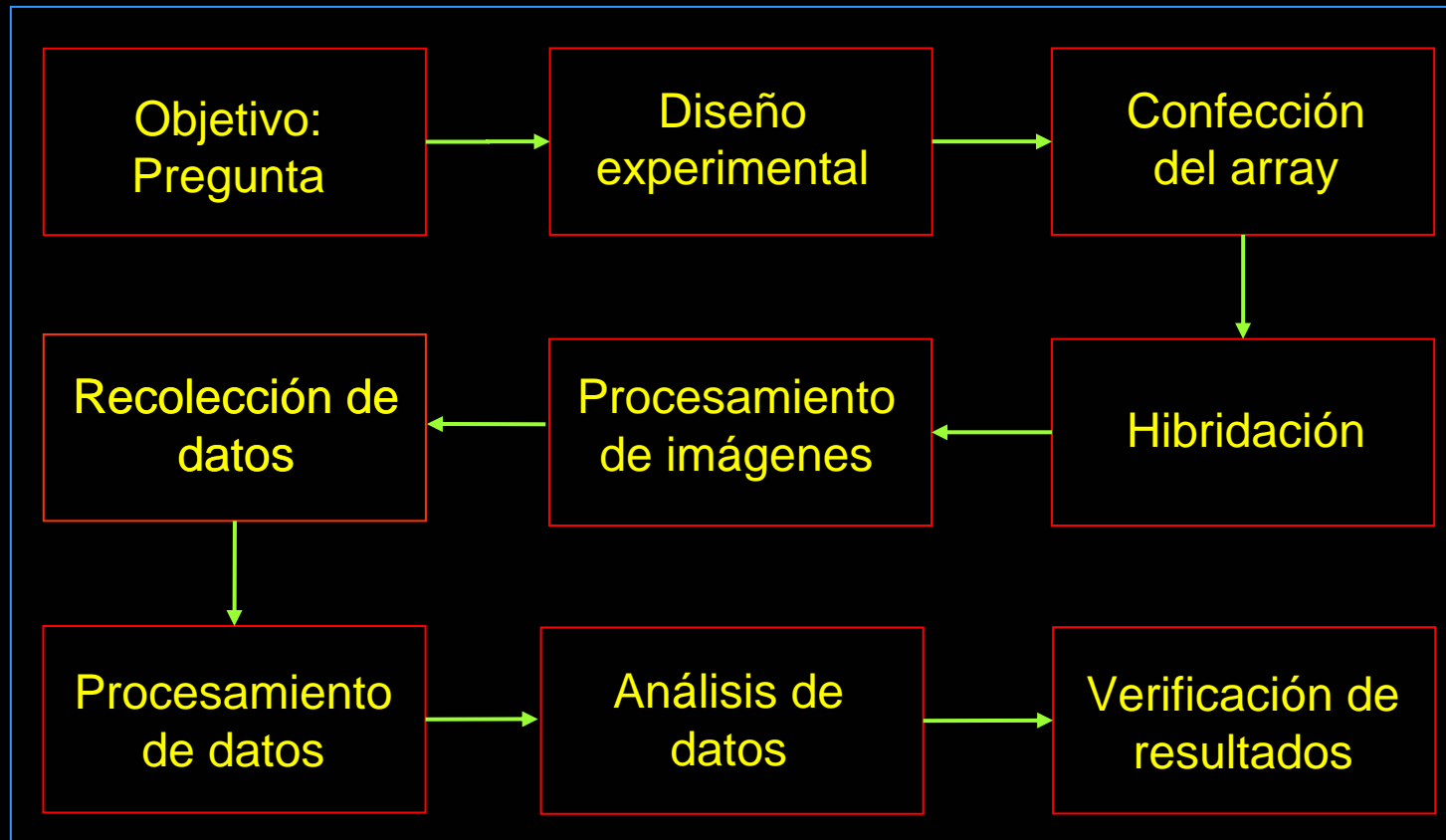
## Medición de la expresión génica a través de microarrays

### Algunas ventajas

- Permite monitorear gran número de genes en simultáneo
- Permite comparar varias condiciones experimentales en simultáneo
- Presenta economía de escala en el diseño experimental
- Sistemas automatizados
- Permite inferir comportamientos globales de los genes

## Medición de la expresión génica a través de microarrays

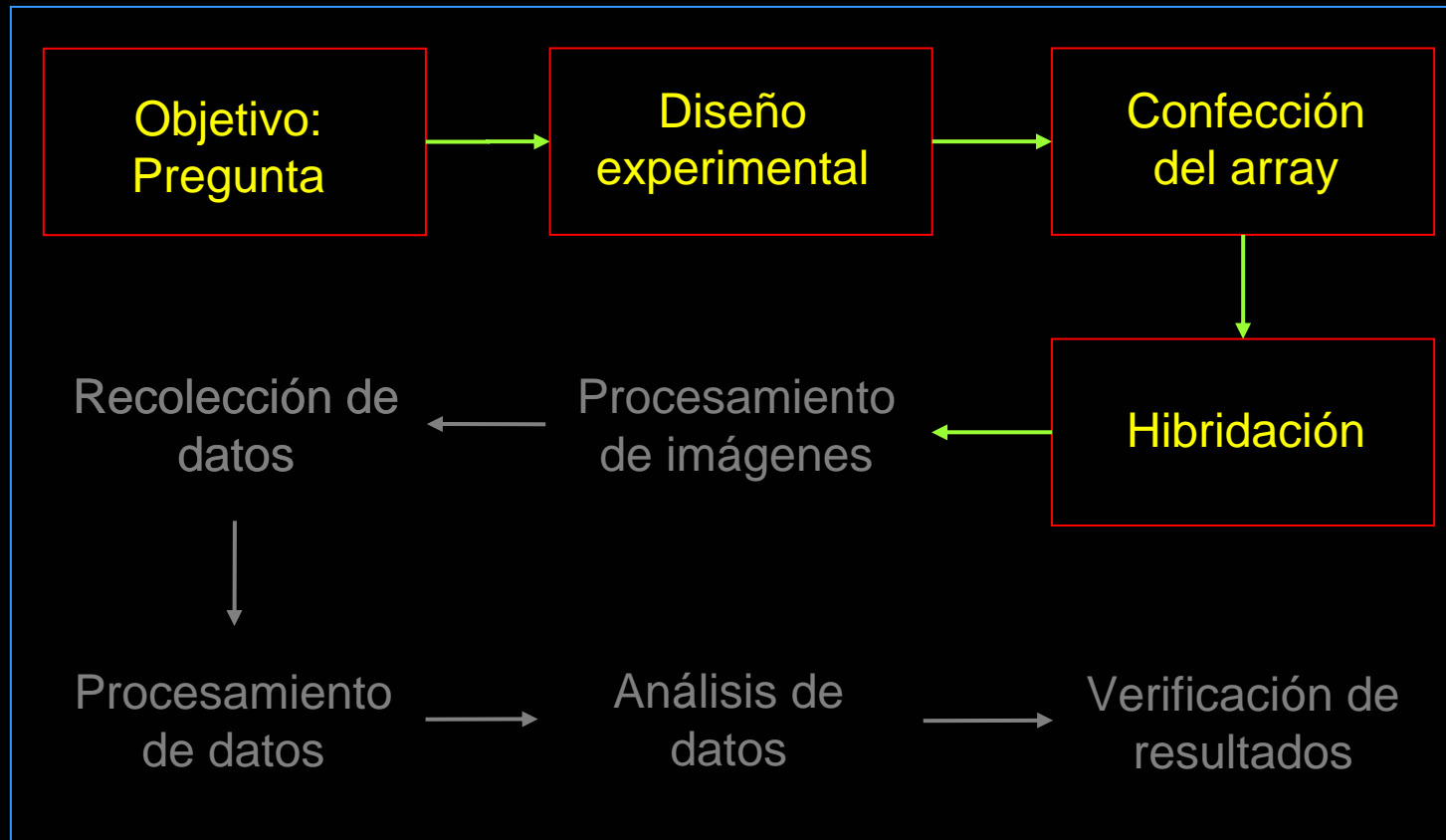
*Hipótesis*



*Nuevas Hipótesis*

## Medición de la expresión génica a través de microarrays

*Hipótesis*

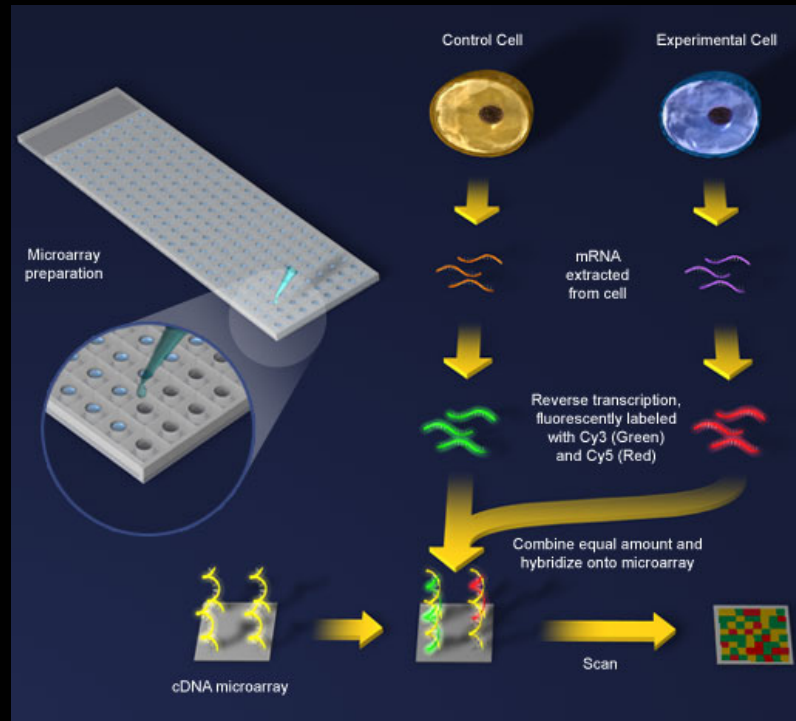


*Nuevas Hipótesis*

# Medición de la expresión génica a través de microarrays

## Confección del array

- cDNA
- oligonucleótidos
- trozos cromosomas



## Hibridación

- cantidad de sonda
- temperatura
- tiempo

## Diseño experimental

### Biológico

- tipos de muestras
- condiciones experimentales

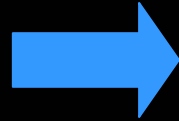
### Estadístico

- número de muestras
- tipos de controles

### Operacional

- sonda radioactiva
- sonda fluorescente
- número de sondas

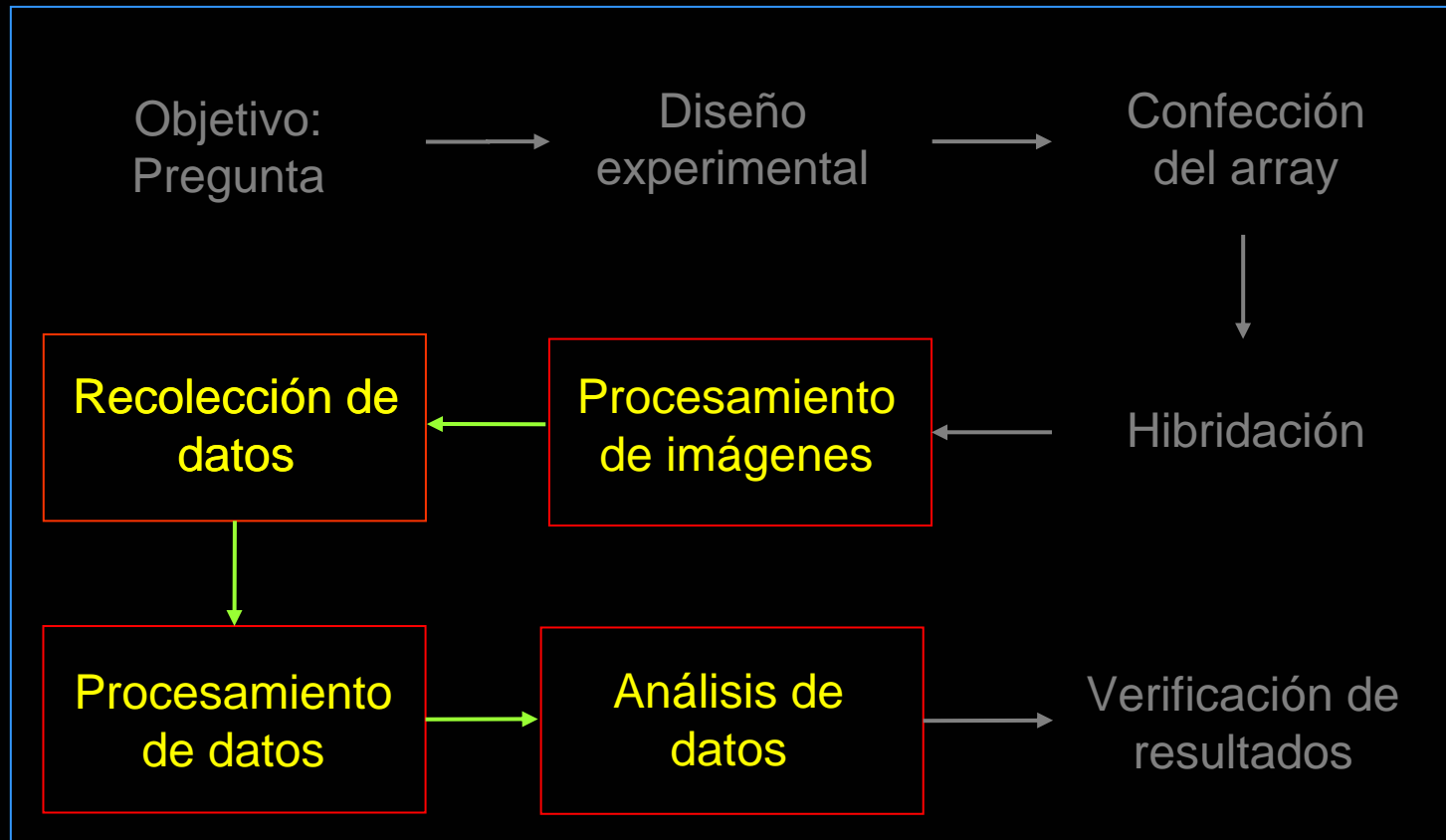
## Medición de la expresión génica a través de microarrays



....y que hago con tanto punto de color??

## Medición de la expresión génica a través de microarrays

*Hipótesis*



*Nuevas Hipótesis*

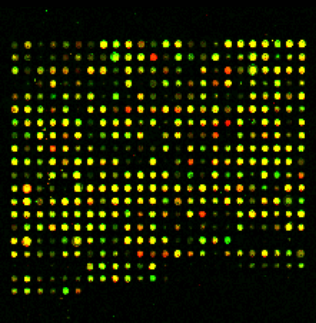


## Medición de la expresión génica a través de microarrays

Procesamiento  
de imágenes



Recolección de  
datos



Asignación posición de cada spot

Identificación de zonas conflicto

Cuantificación de spots



Construcción de bases de datos con  
matrices de expresión génica

Cualquiera sea la pregunta, necesitamos pre-procesar los datos de expresión.

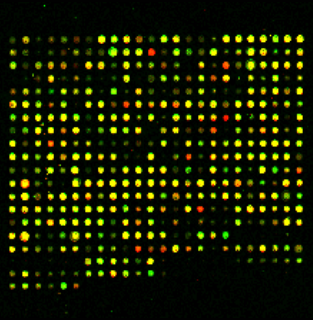
Artefactos que pueden darse:

- variaciones en las cantidades de mRNA iniciales
- diferente incorporación de marca en la sonda
- spot irregulares por efecto de la siembra
- variaciones en la calidad de los lavados post-hibridación
- variaciones en la calidad del escaneo



La señal que debemos analizar debe ser sólo producto de la condición biológica estudiada.

## Medición de la expresión génica a través de microarrays



$$\text{Señal Medida} = \text{Señal Real} + \text{Error}$$

$$\text{Error} = \text{Varianza} + \text{Tendencias}$$



$$\text{Señal medida} = \text{Señal Real} + \text{Varianza} + \text{Tendencias}$$

- Necesitamos disminuir la Varianza y las Tendencias

Varias réplicas  Reducción de la varianza

Normalización  Eliminación de tendencias y background



$$\text{Señal medida} \sim \text{Señal real}$$

## Medición de la expresión génica a través de microarrays

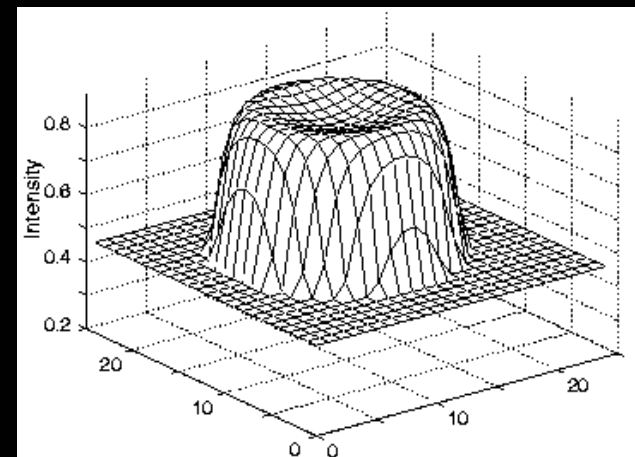
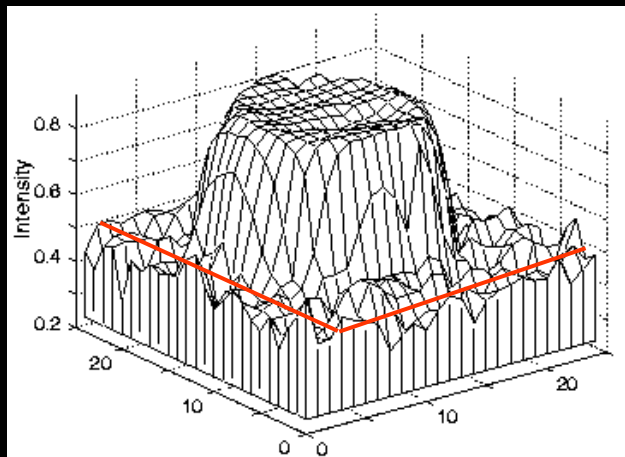
Varias réplicas → Reducción de la varianza

Se resuelve....haciendo más réplicas...

*Réplicas biológicas:* mRNA obtenidos por separado bajo el mismo estímulo  
*Réplicas experimentales:* diferentes eventos de marcaje para un mismo pool de mRNA

Background → Eliminación de señales inespecífica

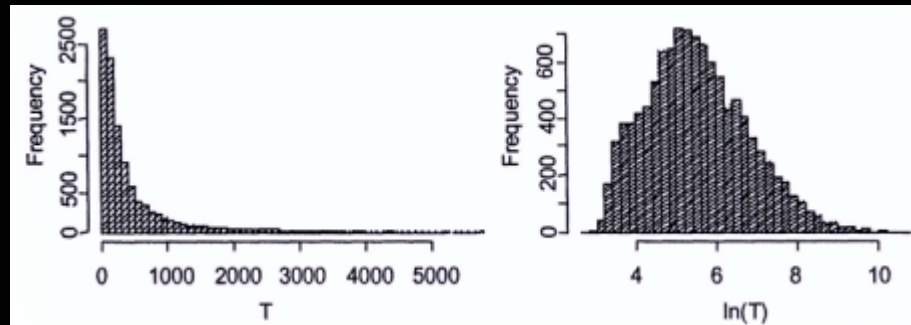
*Corrección por señal local o global*



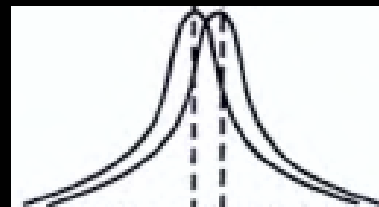
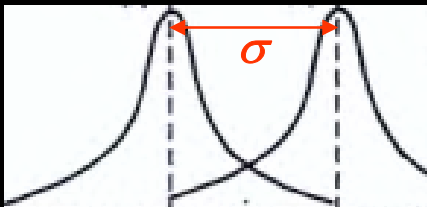
# Medición de la expresión génica a través de microarrays

Normalización → Eliminación de tendencias

- *Transformación a logaritmo*



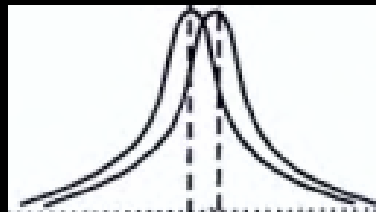
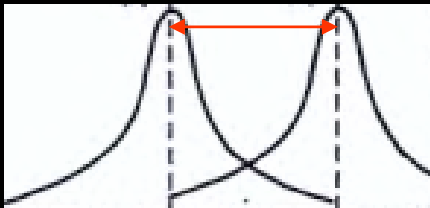
- *Un factor de corrección global  $\sigma$ .*



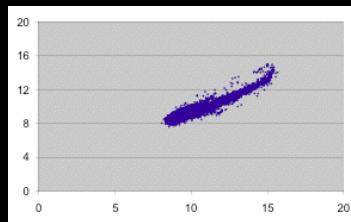
# Medición de la expresión génica a través de microarrays

Normalización → Eliminación de tendencias

- *Uso de housekeepings o spikes, como referencia*

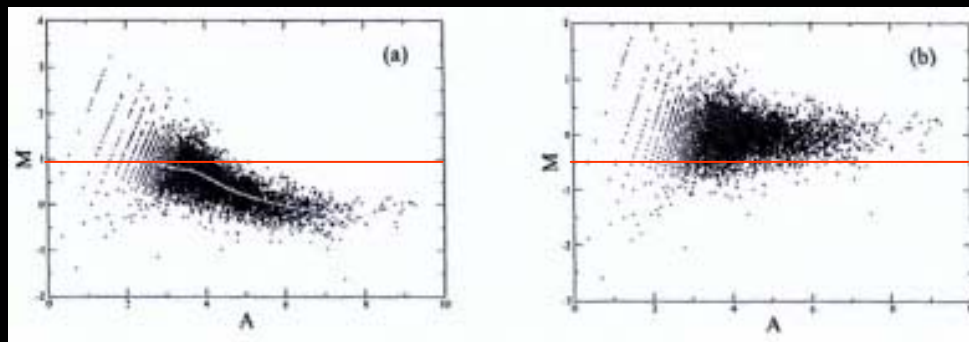
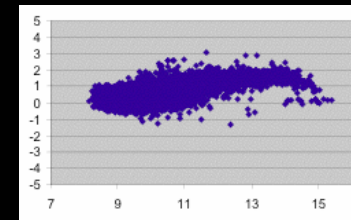


- *Métodos de regresión: lineales y no lineales*



$$M = \log(R/G)$$

$$A = 0.5 * (\log R + \log G)$$



*lowess*

Ya tenemos entonces...

**Señal medida ~ Señal real**

Y ahora??



Las preguntas más tradicionales:

¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?

¿ Existen genes co-regulados que establecen grupos de patrones de expresión particulares en las condiciones estudiadas?

## Medición de la expresión génica a través de microarrays

¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?

Existen diferentes aproximaciones para responder esto:

### •Veces de Cambio

- Sólo identifica los que más cambian.
- Alta tasa de falsos positivos
- La razón de cambio es sensible cuando el denominador es pequeño

$$\begin{array}{ccc} \text{Control} & \frac{30}{60} & = 0.5 \\ \text{Tratado} & & \\ \text{Control} & \frac{500}{1000} & = 0.5 \\ \text{Tratado} & & \\ \downarrow & & \\ \text{Control} & \frac{45}{45} & = 1 \\ \text{Tratado} & & \\ \text{Control} & \frac{515}{985} & = 0.52 \\ \text{Tratado} & & \end{array}$$



## Medición de la expresión génica a través de microarrays

¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?

### • Test de t

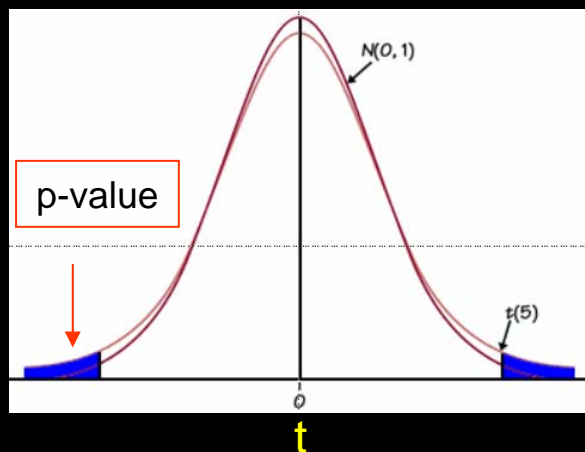
Para cualquier gen

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

$\mu$  representa su expresión

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{v}{n_1} + \frac{v}{n_2}}}$$

Permite determinar la probabilidad de que la diferencia observada sea dada por azar



- Requiere alto numero de réplicas para aproximar normalidad
- Alta tasa de falsos positivos
- Requiere que las varianzas de ambos genes a lo largo de las muestras sea igual

## Medición de la expresión génica a través de microarrays

¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?

### • Test de Welsh

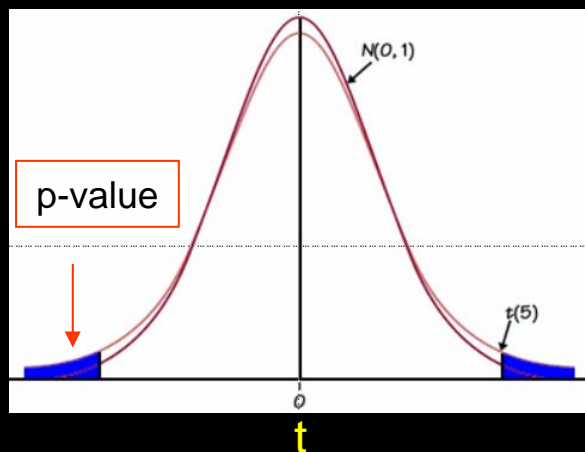
Para cualquier gen

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

$\mu$  representa su expresión

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{v_1}{n_1} + \frac{v_2}{n_2}}}$$

Permite determinar la probabilidad de que la diferencia observada sea dada por azar



- Requiere alto numero de réplicas para aproximar normalidad
- Alta tasa de falsos positivos
- Mayor dificultad en el cálculo de los grados de libertad

## Medición de la expresión génica a través de microarrays

¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?

- Significance Analysis of Microarrays (SAM)

- Simula el número de réplicas a través de permutaciones de los mismos datos
- Permite controlar la tasa de falsos positivos
- No necesita asumir distribuciones de los datos
- Utiliza estadísticao similar al del test de t (o al de Welsh)

$$d_i = \frac{r_i}{s_i + s_0}$$

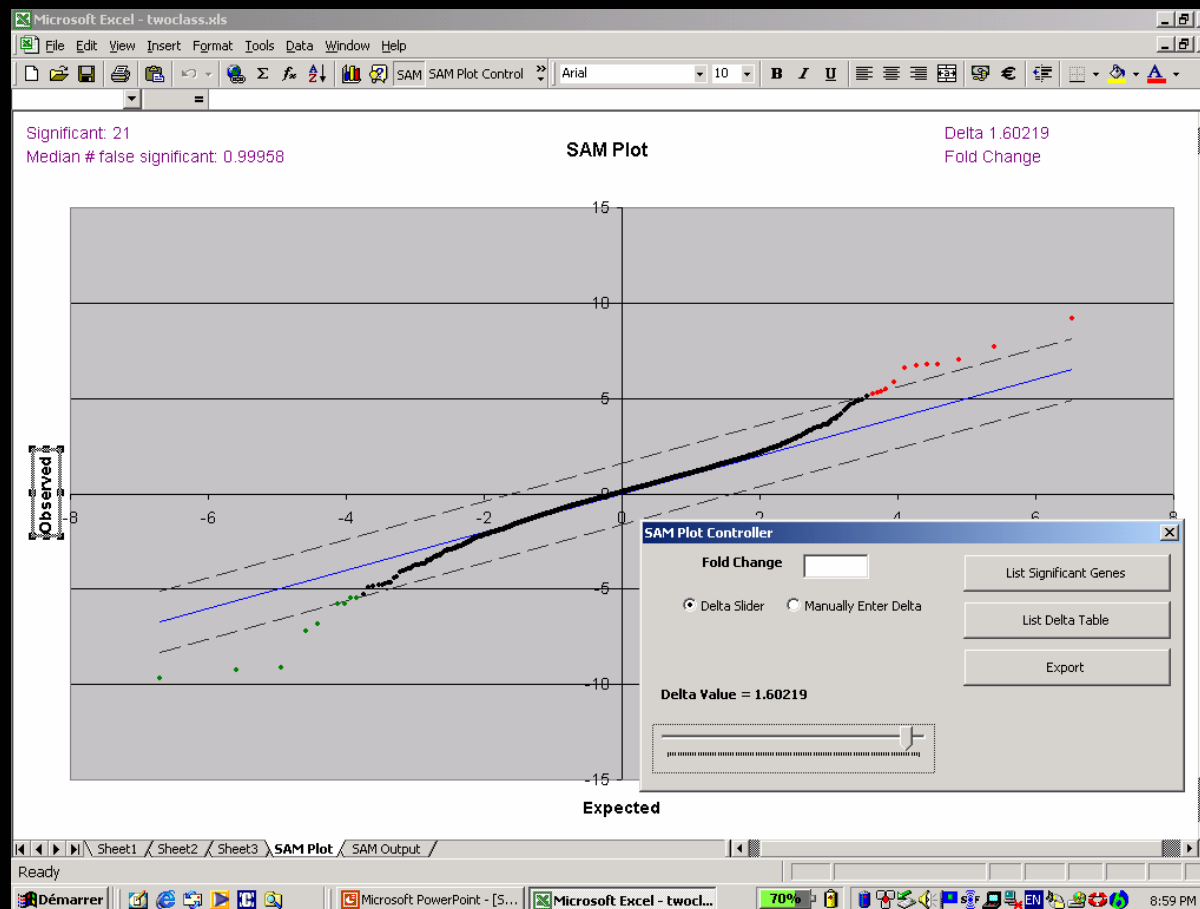
$$r_i = \bar{x}_{i1} - \bar{x}_{i2}$$

$$s_i = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{\sum_{j \in C1} (x_{ij} - \bar{x}_{i1})^2 + \sum_{j \in C2} (x_{ij} - \bar{x}_{i2})^2}{n_1 + n_2 - 2}}$$

# Medición de la expresión génica a través de microarrays

¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?

- Significance Analysis of Microarrays (SAM)

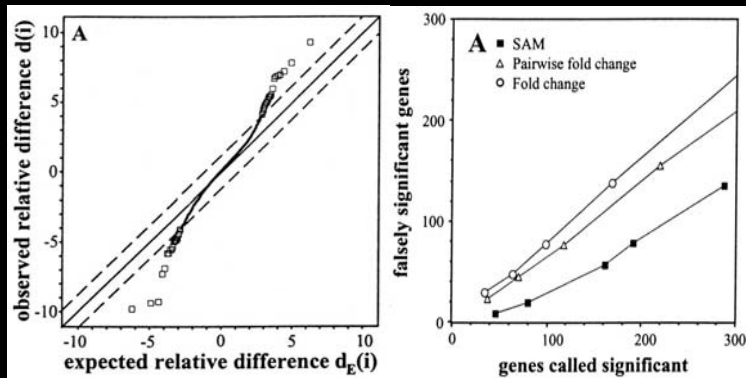


## Medición de la expresión génica a través de microarrays

¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?



**Significance analysis of microarrays applied to the ionizing radiation response.**  
*Tusher et al. 2001. PNAS 98:5116*



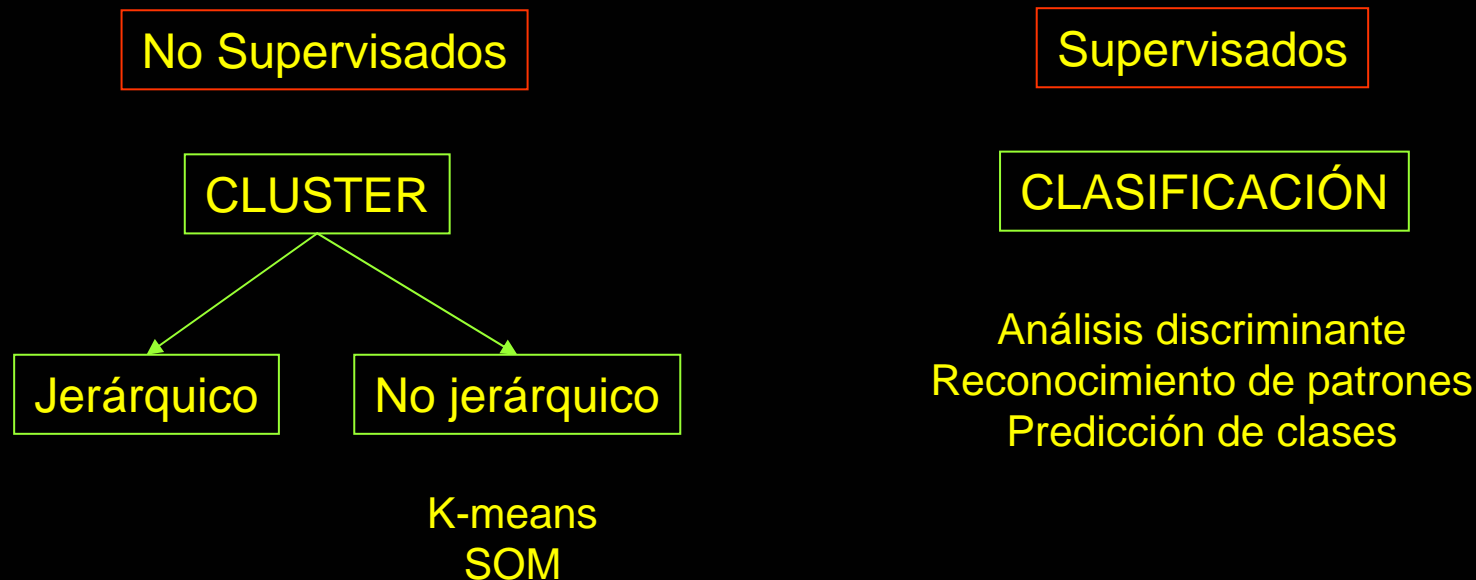
Los cambios de expresión significativos mas importantes:

- Genes que participan en apoptosis
- Genes que participan en ciclo celular
- Genes que participan en reparación del daño al DNA

## Medición de la expresión génica a través de microarrays

¿ Existen genes co-regulados que establecen grupos de patrones de expresión particulares en las condiciones estudiadas?

Existen diferentes métodos para responder esto...

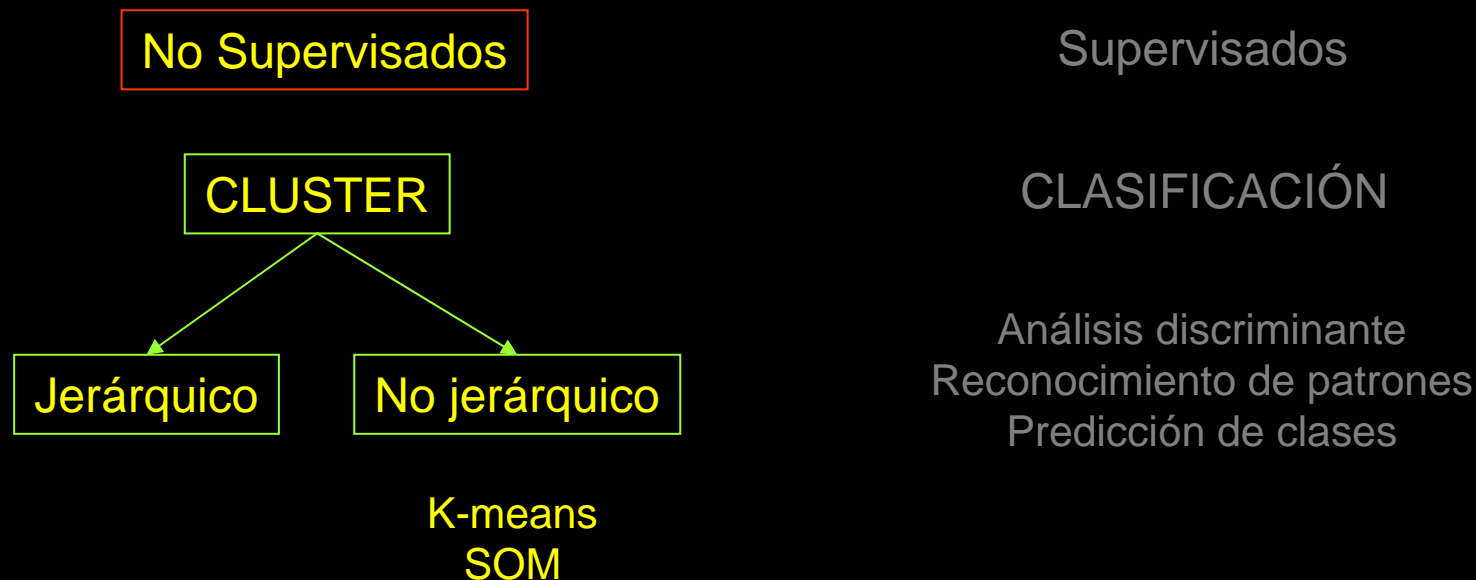


...pero la base para la respuesta es la misma: genes co-regulados deben poseer patrones de expresión similares.

## Medición de la expresión génica a través de microarrays

¿ Existen genes co-regulados que establecen grupos con patrones de expresión particulares en las condiciones estudiadas?

Existen diferentes métodos para responder esto...



...pero la base para la respuesta es la misma suposición: genes co-regulados deben poseer patrones de expresión similares.

## Medición de la expresión génica a través de microarrays

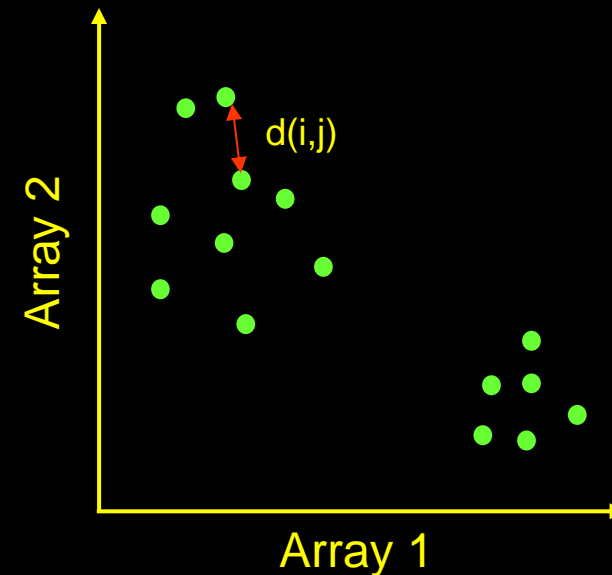
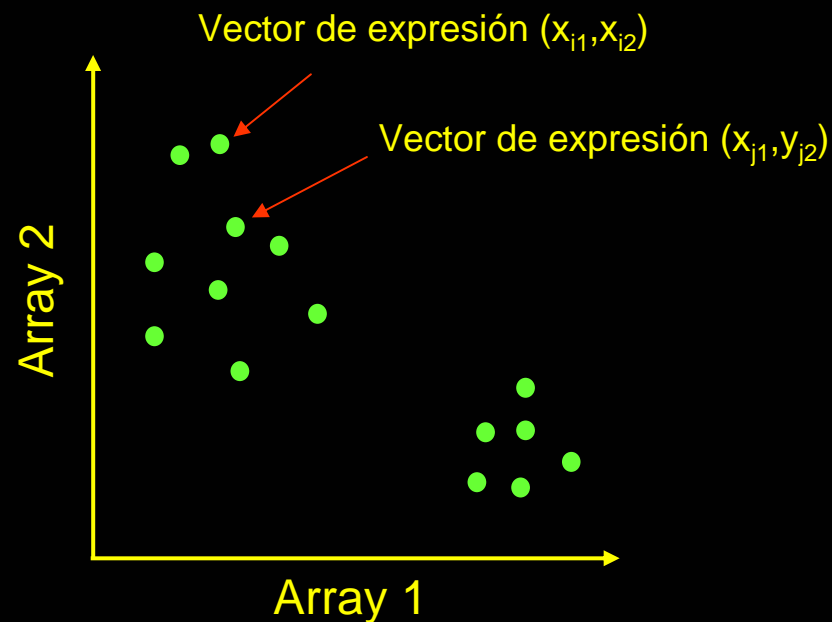
### CLUSTER

- Jerárquico: Aglomerativos y divisivos
- Analisis microarrays: aglomerativos
- Agrupar datos de expresión en base a similitud

Matriz de datos  
de expresión



Cálculo de distancia (o similitud)  
a través de alguna métrica





## Medición de la expresión génica a través de microarrays

### Distancia Euclidiana

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{in} - x_{jn}|^2}$$

### Distancia Manhattan

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{in} - x_{jn}|$$

### Distancia Minkowski

$$d(i, j) = (|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{in} - x_{jn}|^q)^{1/q}$$

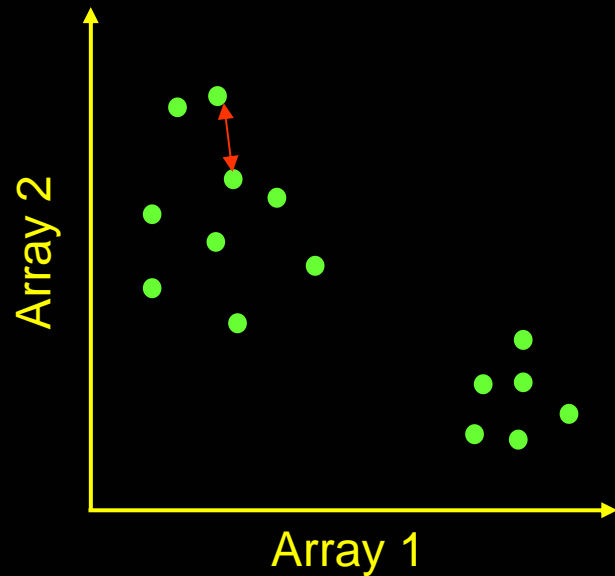
### Distancia pesada

$$d(i, j) = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \dots + w_n|x_{in} - x_{jn}|^2}$$

### Correlacion de Pearson

$$r = \frac{1}{N} \sum_{i=1, N} \left( \frac{X_i - \bar{X}}{\sigma_X} \right) \left( \frac{Y_i - \bar{Y}}{\sigma_Y} \right)$$

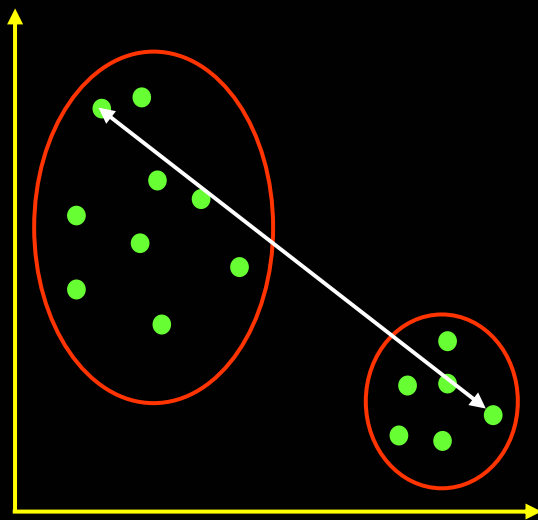
## Medición de la expresión génica a través de microarrays



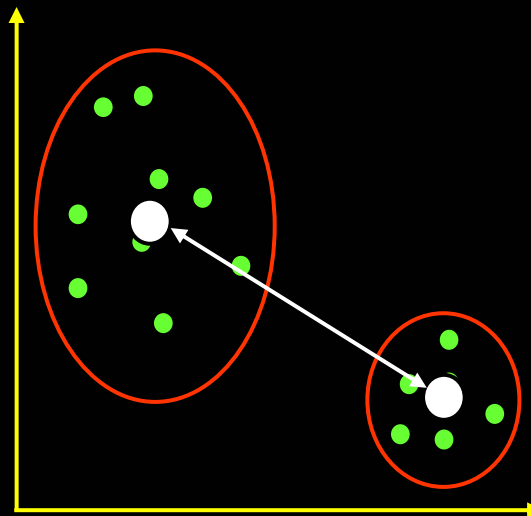
- $d(i,j)$  u  $r$  constituyen una nueva matriz
- ¿Cómo agrupo los más similares?



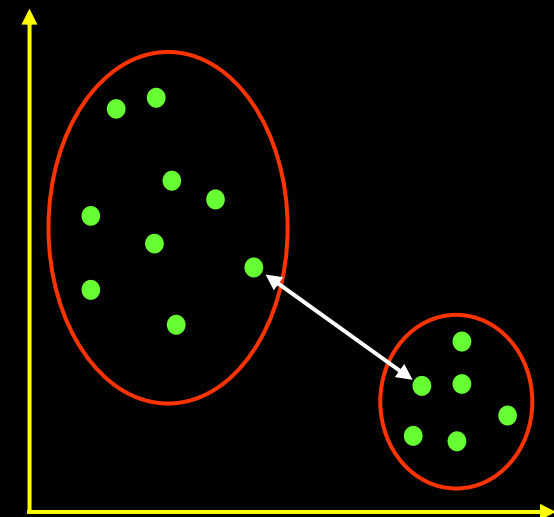
Linkage



Complete Linkage

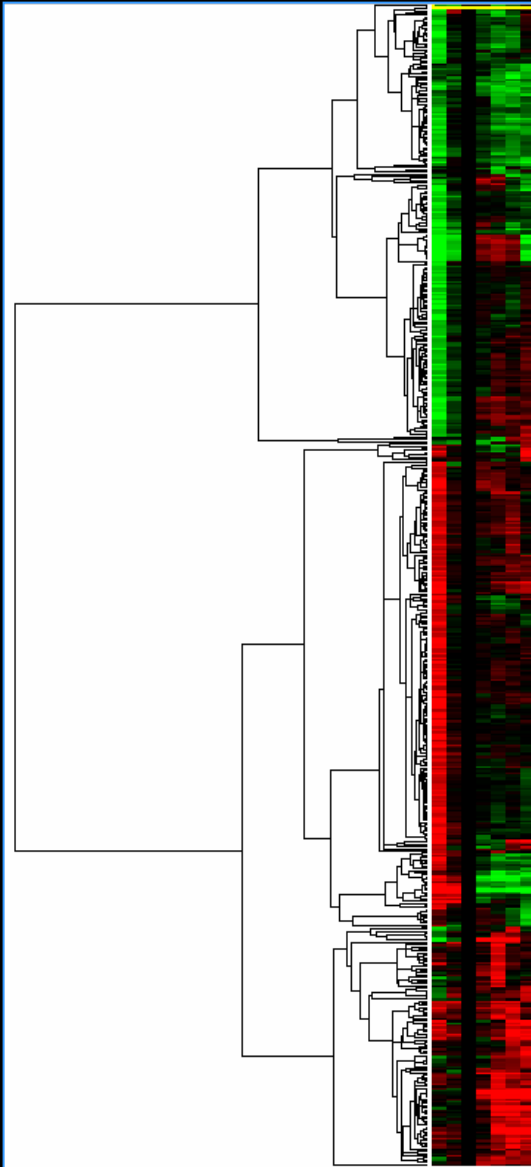


Average Linkage



Single Linkage

## Medición de la expresión génica a través de microarrays



### Ventajas de la técnica:

- Es exploratoria, por lo que no necesita conocimiento a priori
- Existen numerosas implementaciones para algoritmos de cluster
- Permite una mirada “visual” de la distribución de la información

### Desventajas de la técnica:

- Determinación del número de grupos
- Resultados varían al variar la métrica o el linkage
- No tiene confianza estadística

## Medición de la expresión génica a través de microarrays

### CLUSTER

- No Jerárquico: particionales

### K-means

- Requiere a priori el número de grupos a encontrar
- Cada objeto (valor de expresión) es asignado a un cluster en una sólo etapa
- Itera hasta que los objetos dentro del cluster sean lo mas similares posibles
- El resultado del cluster depende de la partición inicial

### Self organizing maps (SOM)

- Es similar a K-means en el core de la operación
- Permite redimensionar la data (2D o 3D)
- Permite tener una observación gráfica

# Medición de la expresión génica a través de microarrays

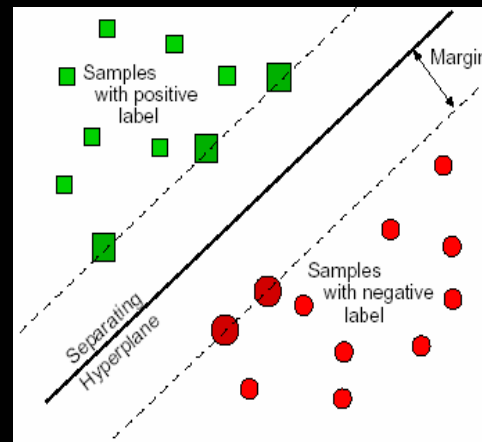
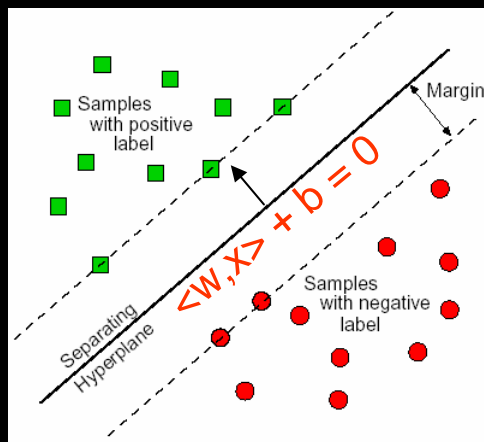
## Una breve nota sobre métodos supervisados

En general se usan en clasificación basándose en la idea de que para un número de muestras biológicas ya existe información preliminar, que puede utilizarse para agrupación de nuevos datos en clusters.

Características de estos algoritmos:

- Perfiles de expresión de genes conocidos en términos de su función (o ubicación)
- Una regla de decisión que pueda explicar el set de entrenamiento
- El desafío: encontrar una regla de decisión que permita generalizar hacia el resto de los datos

## Ej. Support Vector Machine (SVM)



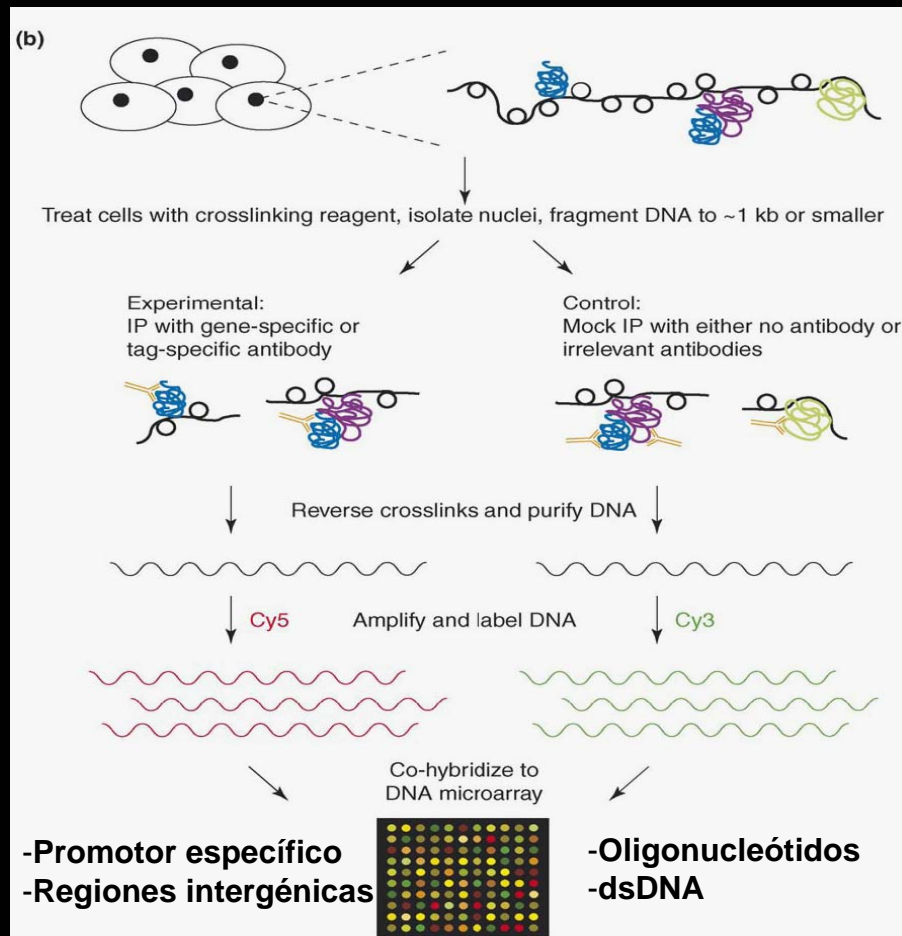
Los elementos más cercanos al hiperplano conforman el "Support Vector"

$W$  y  $b$  se determinan en el entrenamiento

# Medición de la expresión génica a través de microarrays

## Nuevos Usos de la tecnología de microarrays

### ChIP on Chip Arrays (Chromatin Immunoprecipitation microarrays)



#### Ventajas:

- Provee una “instantánea *in vivo*” de la célula bajo las condiciones determinadas
- Puede determinar sitios de interacción en el DNA que actúan directa o indirectamente con la proteína (a través de otras proteínas)

#### Desventajas:

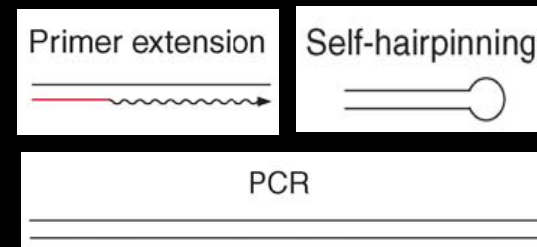
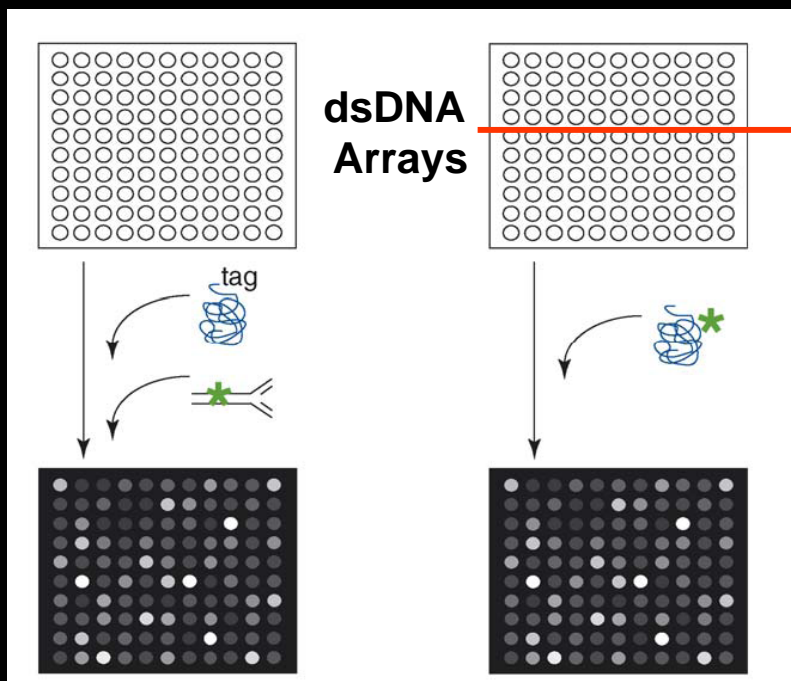
- Anticuerpos específicos
- Momento de interacción DNA-Prot

#### Usos:

- Modificaciones de la cromatina (Nucleosoma/Histonas)
- Identificación de sitios de inicio transcripcionales
- Determinación de sitios de unión de TFs

# Medición de la expresión génica a través de microarrays

## -PBM (Protein Binding Microarrays)



### Ventajas:

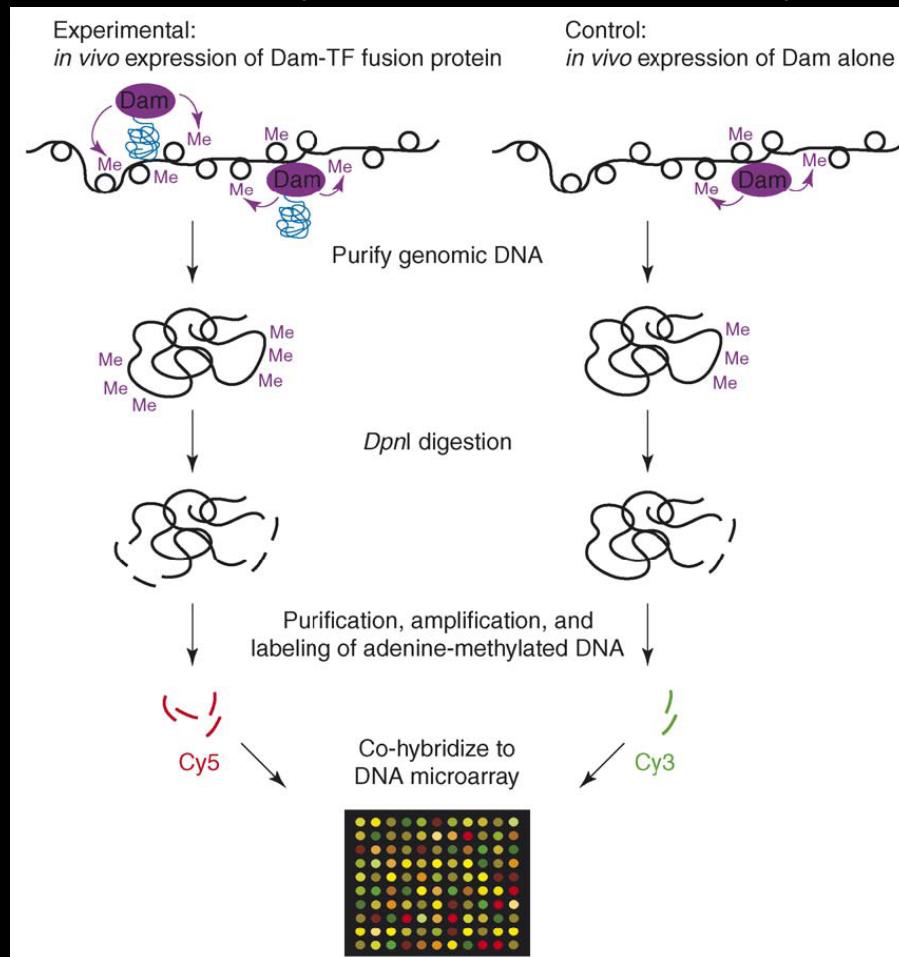
- No precisa anticuerpos específicos
- No requiere conocimiento del momento de la interacción DNA-Proteína (*in vitro*)
- Puede entregar datos de unión de proteínas para variantes de secuencias de DNA
- Rapidez

### Desventajas:

- Necesita de mucho análisis previo de las secuencias de DNA a utilizar (expresión, filogenia, etc.)
- La unión puede no corresponder a una unión que se da en forma endógena.

# Medición de la expresión génica a través de microarrays

## -DamID Arrays (DNA adenine methyltransferase Microarrays)



### Ventajas:

- Provee una “instantánea *in vivo*” de la célula bajo las condiciones determinadas
- No requiere un anticuerpo específico contra la proteína.

### Desventajas:

- La proteína de fusión no tiene la misma dinámica que la endógena.
- Requiere de una leve sobreexpresión de la proteína de fusión
- Dificulta la identificación de regiones de regulación debido a la extensión de las metilaciones.



GRACIAS