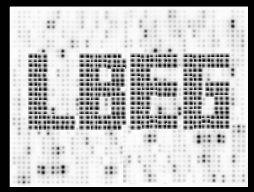


Laboratorio de Bioinformática y
Expresión Génica
INTA – Universidad de Chile

Curso BT51B – Facultad de Ciencias Físicas y Matemáticas:

Estrategias de análisis en experimentos de expresión génica

Ing. Chritian Hödar Q.

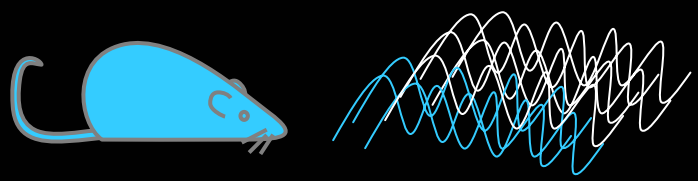
Experimentos de expresión génica

¿Qué medimos?

¿Cómo lo medimos?

Condición A

Condición B



-Northern Blot

-PCR tiempo real

-SAGE
(Serial Analisys of Gene expression)

-Microarrays

Cambios en los niveles de mRNA transcrito

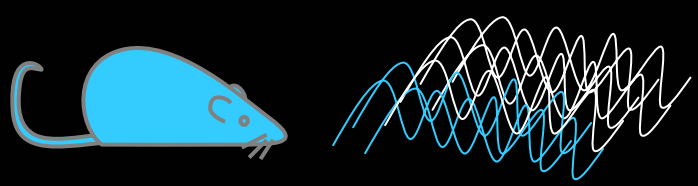
Experimentos de expresión génica

¿Qué medimos?

¿Cómo lo medimos?

Condición A

Condición B



-Northern Blot

-PCR tiempo real

-SAGE
(Serial Analisys of Gene expression)

-Microarrays

Cambios en los niveles de mRNA transcrito

Medición de la expresión génica a través de microarrays

Algunas ventajas

-Permite monitorear gran número de genes en simultáneo

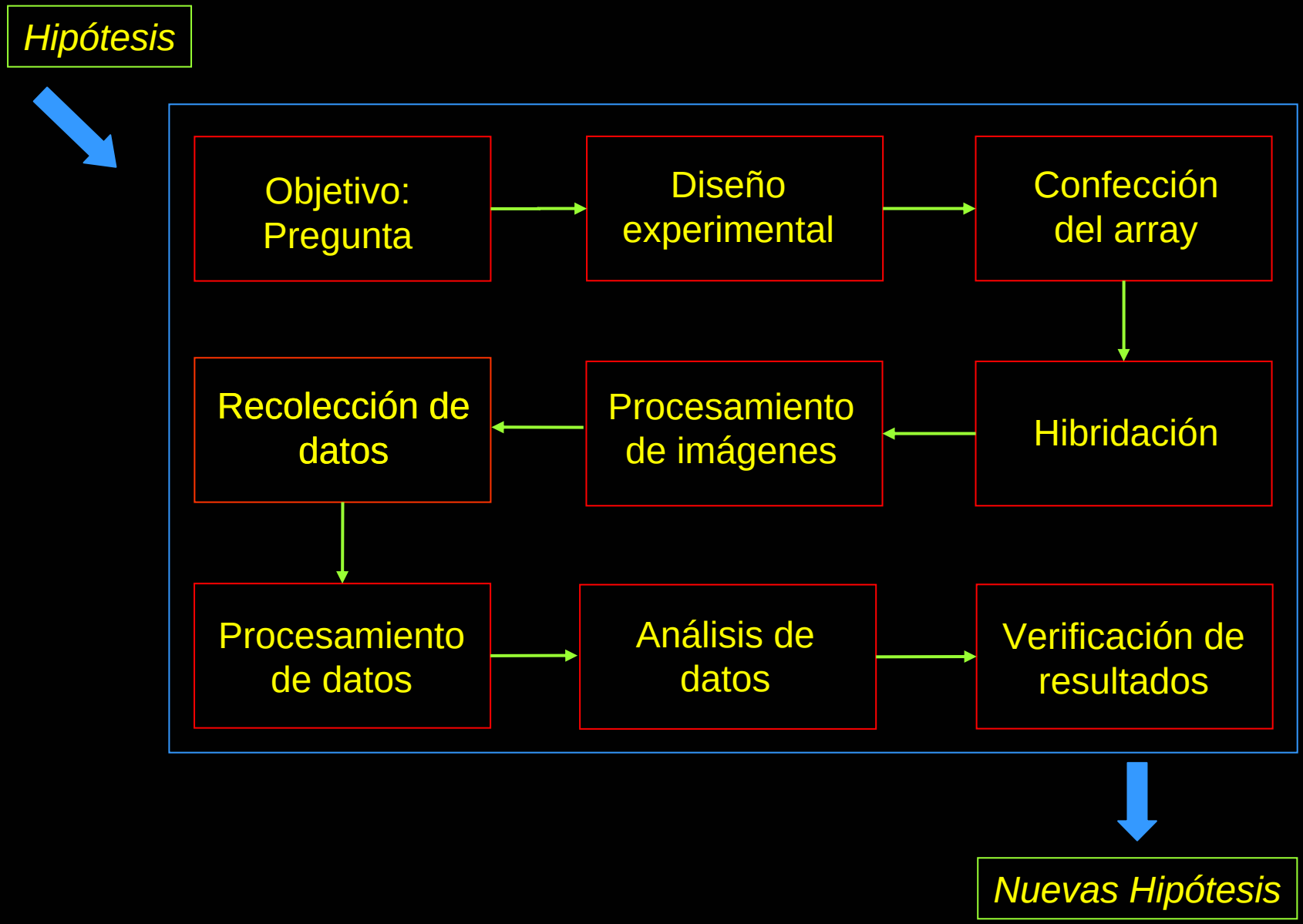
-Permite comparar varias condiciones experimentales en simultáneo

-Presenta economía de escala en el diseño experimental

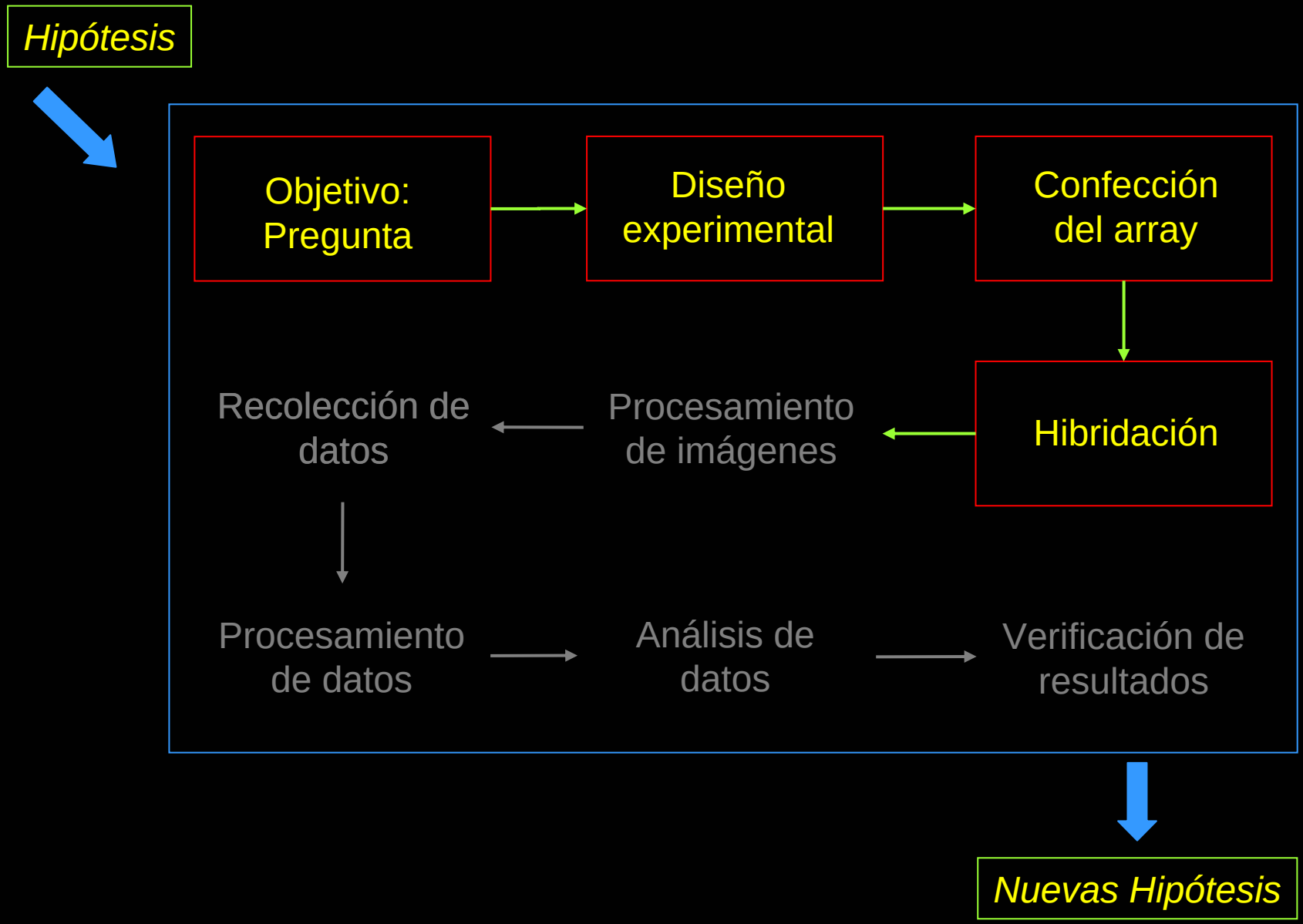
-Sistemas automatizados

-Permite inferir comportamientos globales de los genes

Medición de la expresión génica a través de microarrays



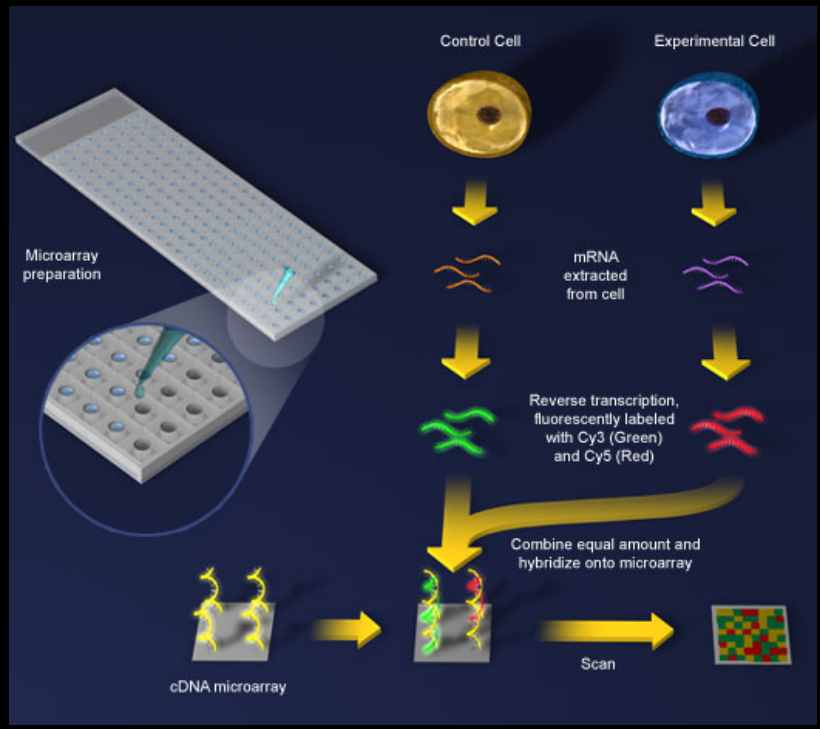
Medición de la expresión génica a través de microarrays



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

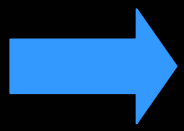
Estadístico

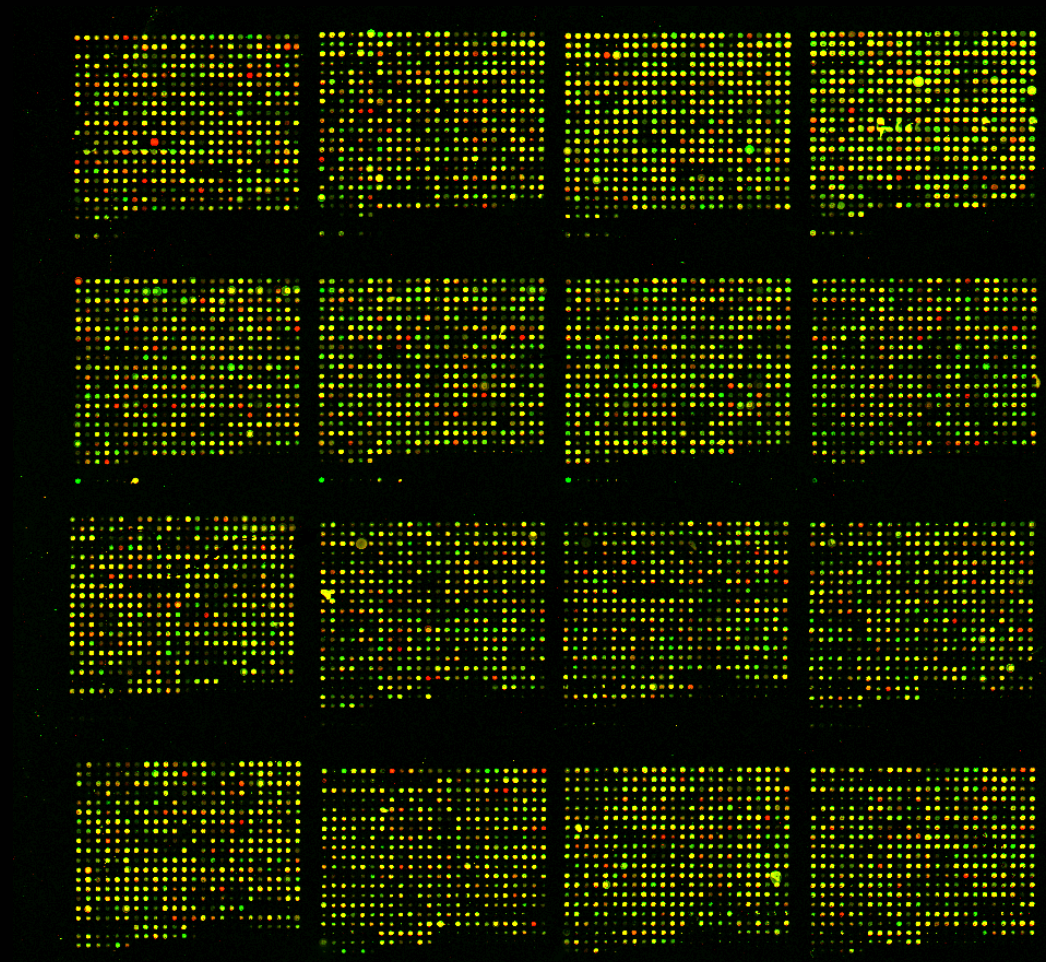
- número de muestras
- tipos de controles

Operacional

- sonda radioactiva
- sonda fluorescente
- número de sondas

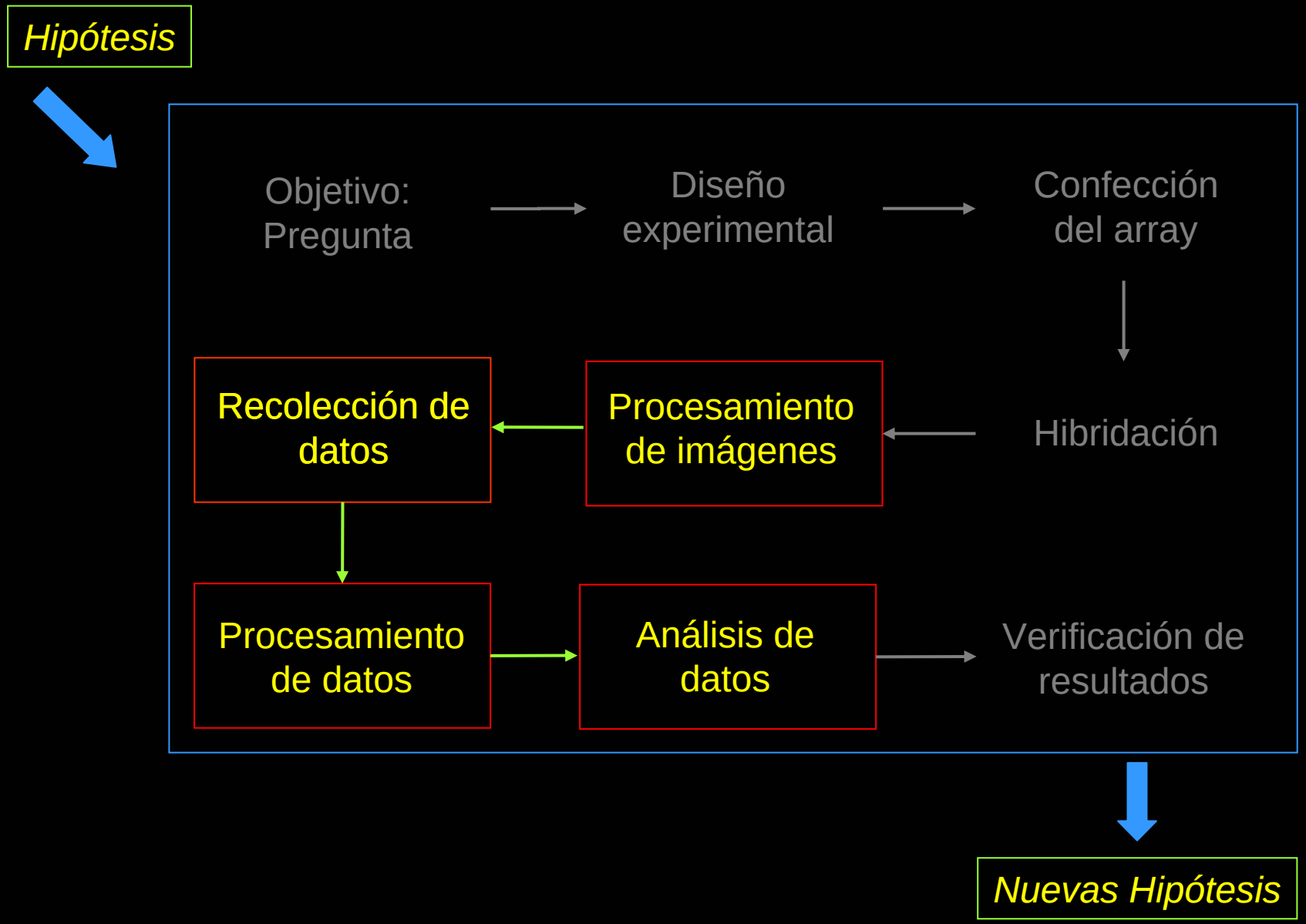
Medición de la expresión génica a través de microarrays





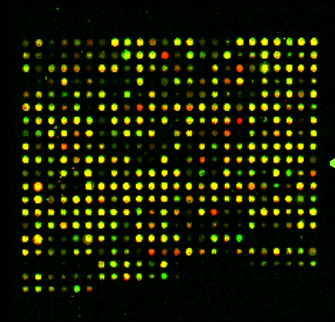
….y que hago con tanto punto de color??

Medición de la expresión génica a través de microarrays



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

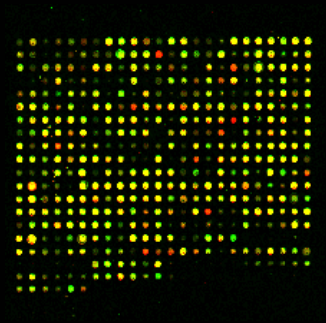
Cualquiera sea la pregunta, necesitamos pre-procesar los datos de expresión.

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



Señal Medida = Señal Real + Error

Error = Varianza + Tendencias

↓

Señal medida = Señal Real + Varianza + Tendencias

- Necesitamos disminuir la Varianza y las Tendencias

Varias réplicas → Reducción de la varianza

Normalización → Eliminación de tendencias y background

↓

Señal medida ~ Señal real

Medición de la expresión génica a través de microarrays

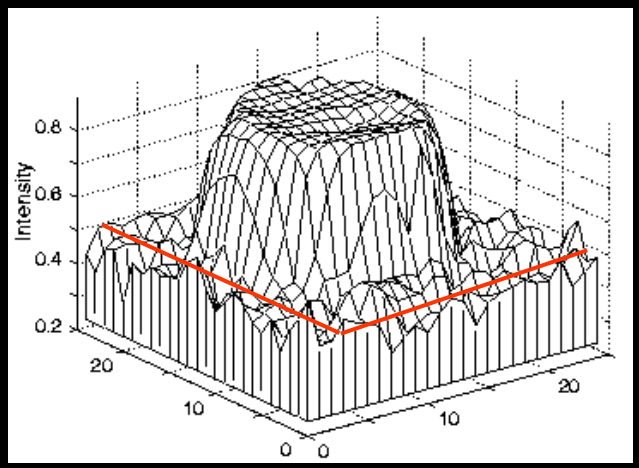
Varias réplicas → Reducción de la varianza

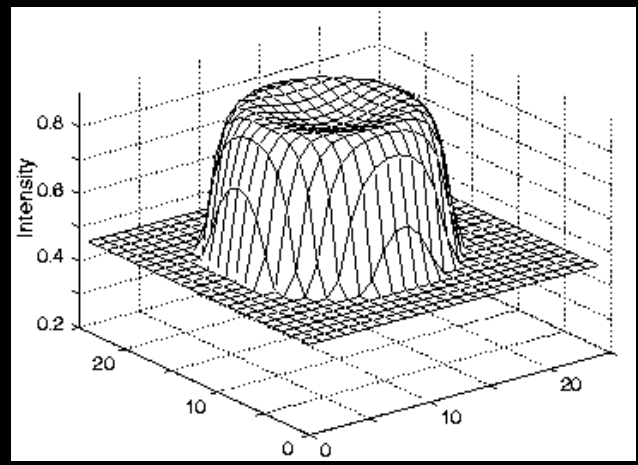
Se resuelve….haciendo más replicas…

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

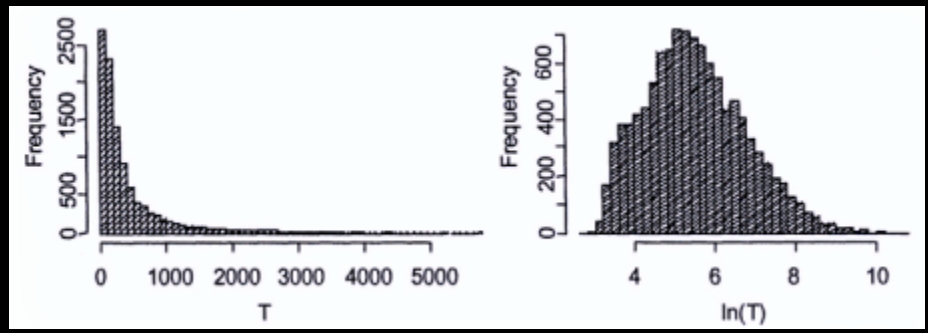




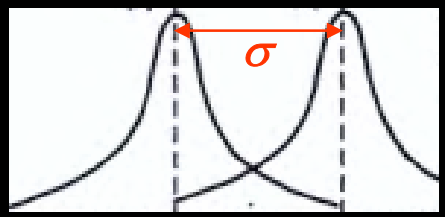
Medición de la expresión génica a través de microarrays

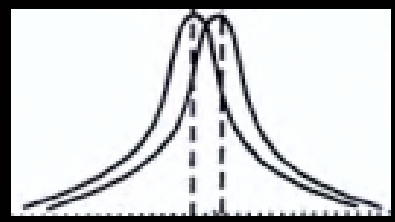
Normalización → Eliminación de tendencias

- *Transformación a logaritmo*



- *Un factor de corrección global σ.*

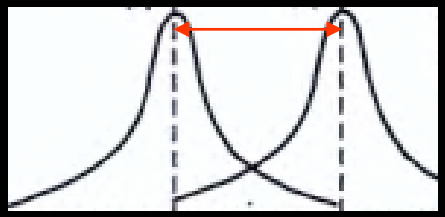


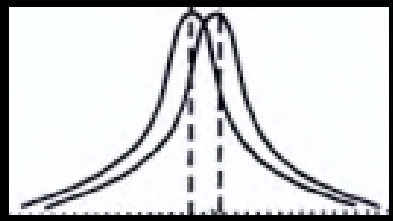


Medición de la expresión génica a través de microarrays

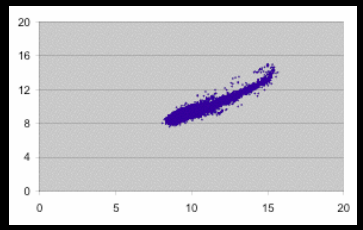
Normalización → Eliminación de tendencias

- *Uso de housekeepings o spikes, como referencia*



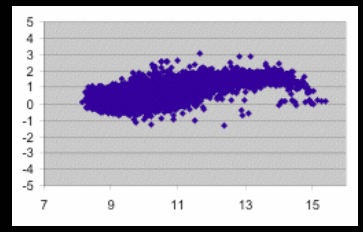


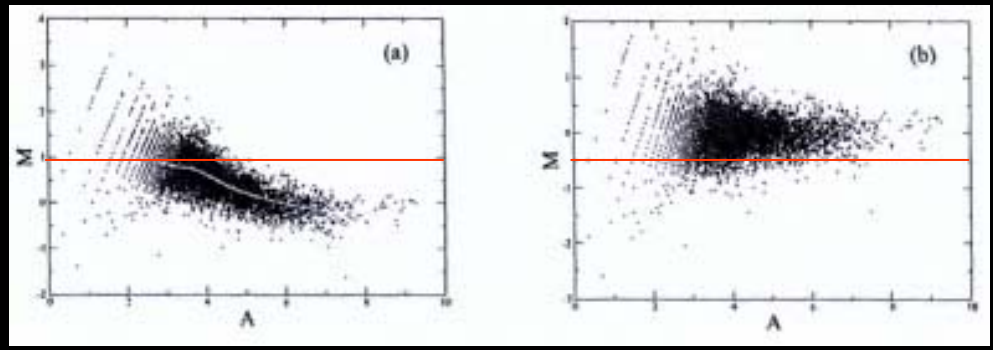
- *Métodos de regresión: lineales y no lineales*



M= log(R/G)

A= 0.5*(logR+logG)





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?

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

Existen diferentes aproximaciones para responder esto, normalemente supervisadas

•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

$$\frac{\text{Control}}{\text{Tratado}} \quad \frac{30}{60} = 0.5 \qquad \frac{\text{Control}}{\text{Tratado}} \quad \frac{500}{1000} = 0.5$$

$$\downarrow$$

$$\frac{\text{Control}}{\text{Tratado}} \quad \frac{45}{45} = 1 \qquad \frac{\text{Control}}{\text{Tratado}} \quad \frac{515}{985} = 0.52$$

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

- Test de t

Para cualquier gen

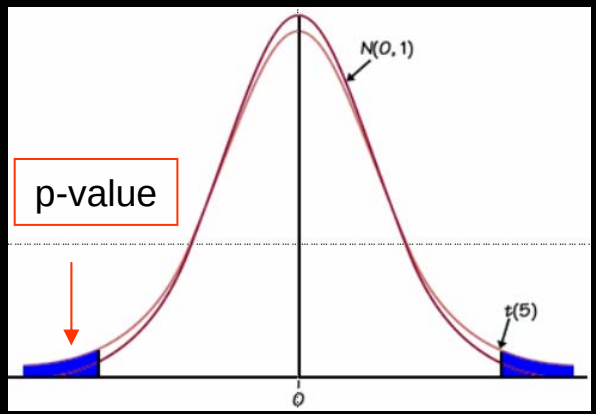
H_0: $\mu 1 = \mu 2$

H_1: $\mu 1 <> \mu 2$

μ 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

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

- Test de Welsh

Para cualquier gen

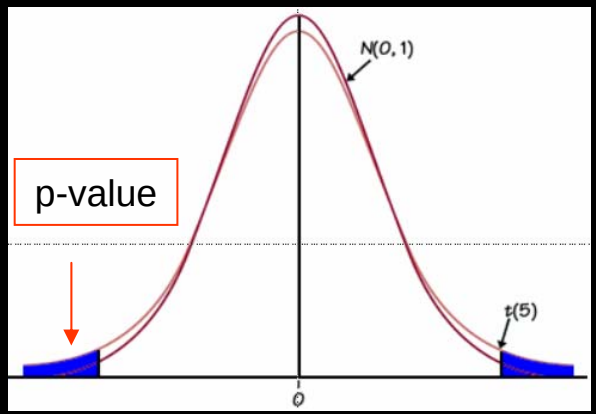
H_0: $\mu 1 = \mu 2$

H_1: $\mu 1 <> \mu 2$

μ 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

¿ 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)

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

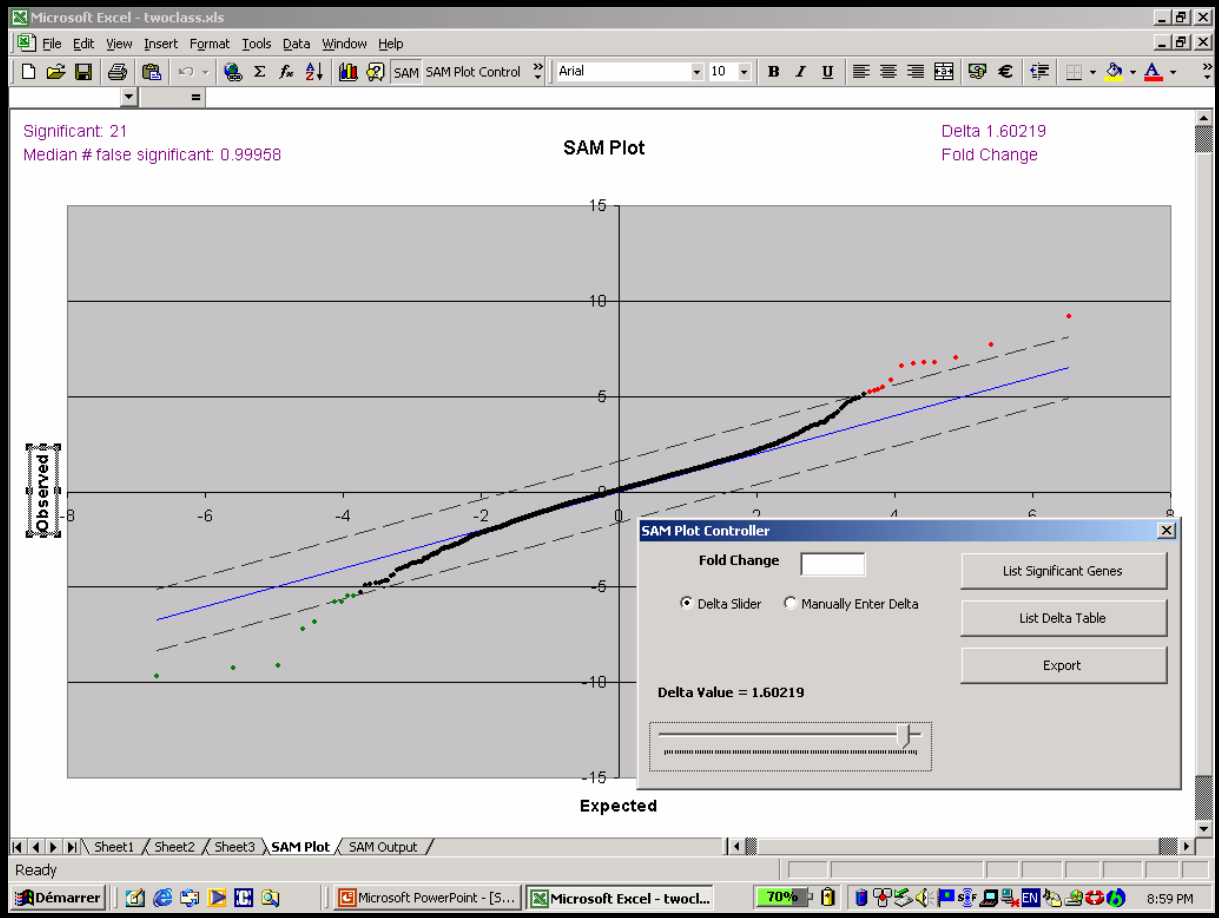
$$d_i = \frac{r_i}{s_i + s_0}$$

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

Medición de la expresión génica a través de microarrays

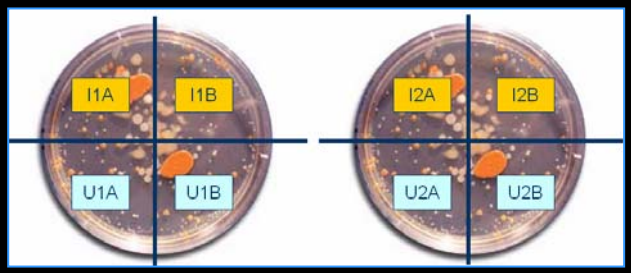
¿ Qué genes mostraron cambios significativos de expresión entre las condiciones estudiadas?

- Significance Analysis of Microarrays (SAM)

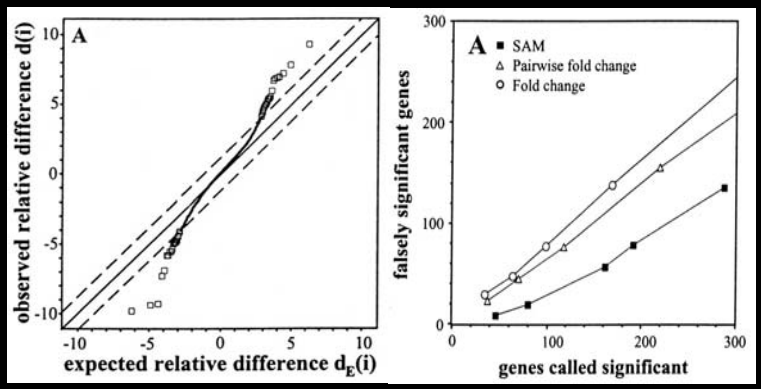


Medición de la expresión génica a través de microarrays

¿ 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

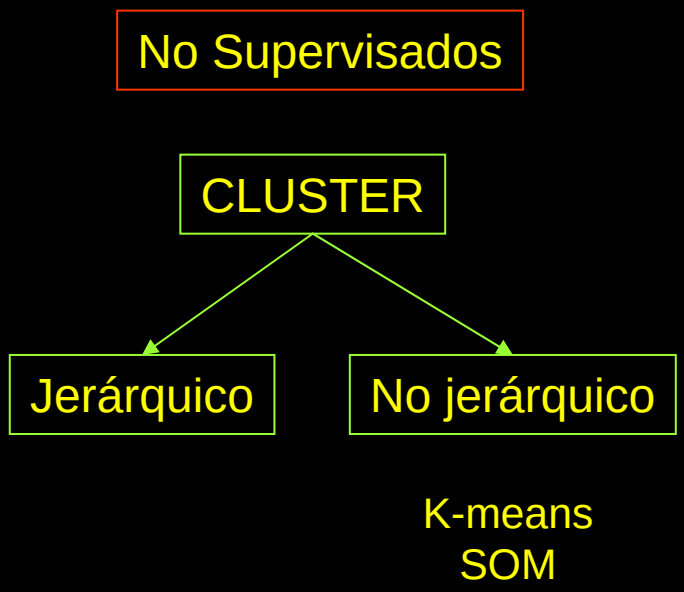


Los cambios de expresión significativos mas importantes:

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

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

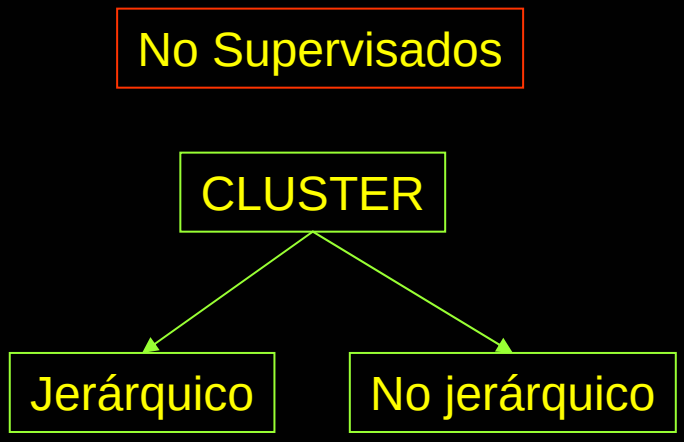
Existen diferentes métodos para responder esto…



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

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

Existen diferentes métodos para responder esto…



Supervisados

CLASIFICACIÓN

Análisis discriminante
Reconocimiento de patrones
Predicción de clases

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

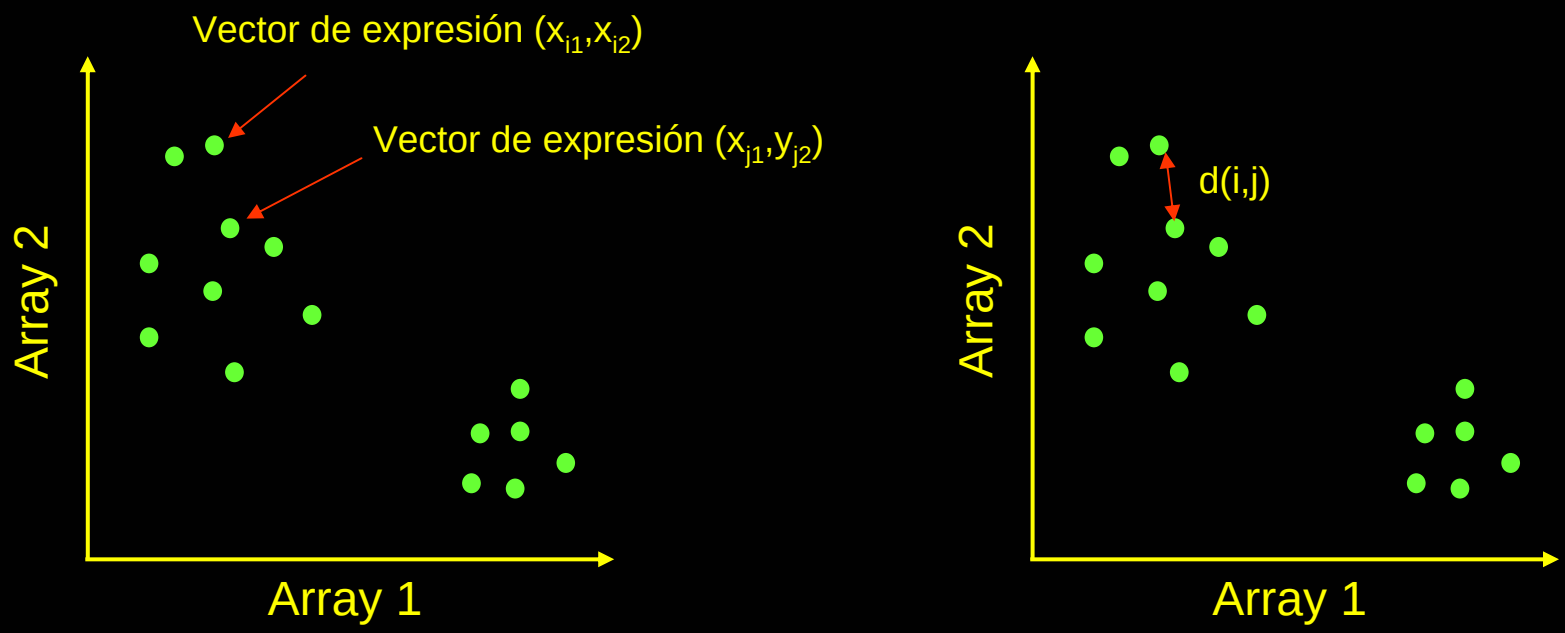
Medición de la expresión génica a través de microarrays

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



Distancia Euclidiana

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

Distancia Manhattan

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

Distancia Minkowski

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

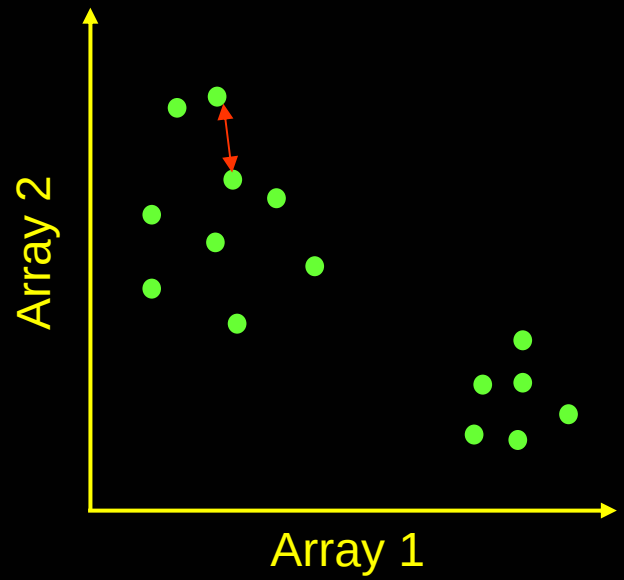
Distancia pesada

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

Correlacion de Pearson

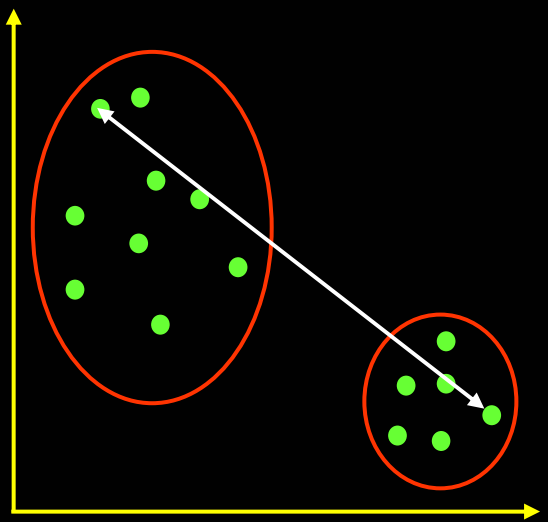
$$r = \frac{1}{N} \sum_{i=1,N} \left(\frac{X_i - \overline{X}}{\sigma_X} \right) \left(\frac{Y_i - \overline{Y}}{\sigma_Y} \right)$$

Medición de la expresión génica a través de microarrays

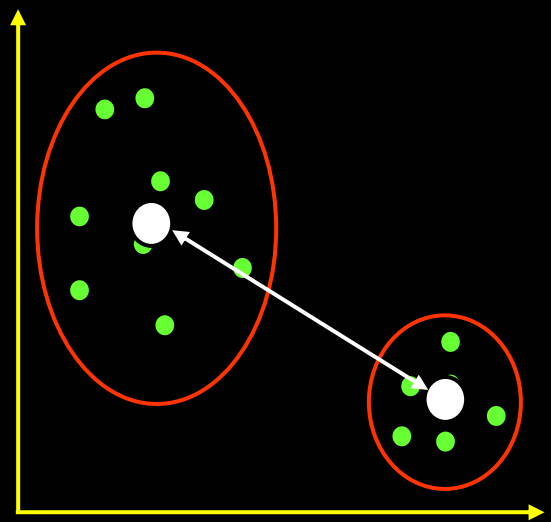


-d(i,j) constituyen una nueva matriz
- ¿Cómo agrupo los más similares?

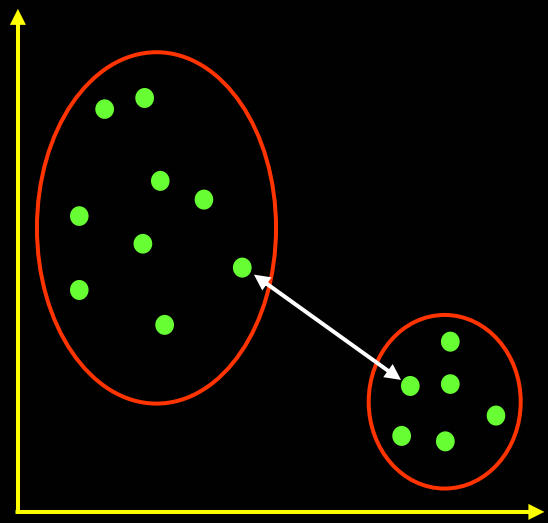
Linkage



Complete Linkage

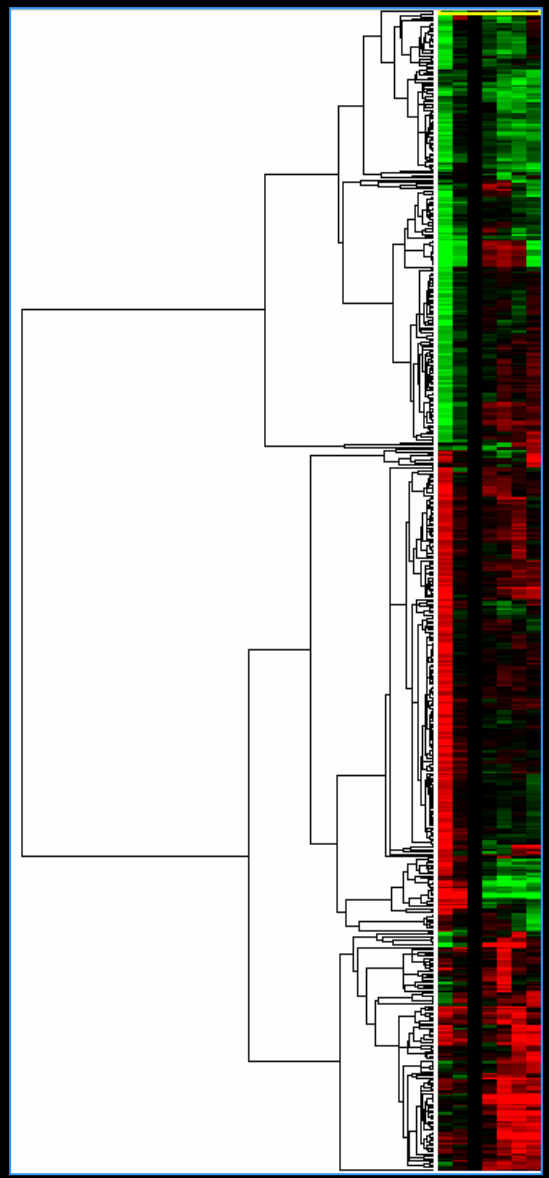


Average Linkage



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 confidencia 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 asiganado a un cluster en una sóla 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 redimensionalizar la data (2D o 3D)
-Permite tener una observación gráfica

Una breve nota sobre métodos supervisados

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

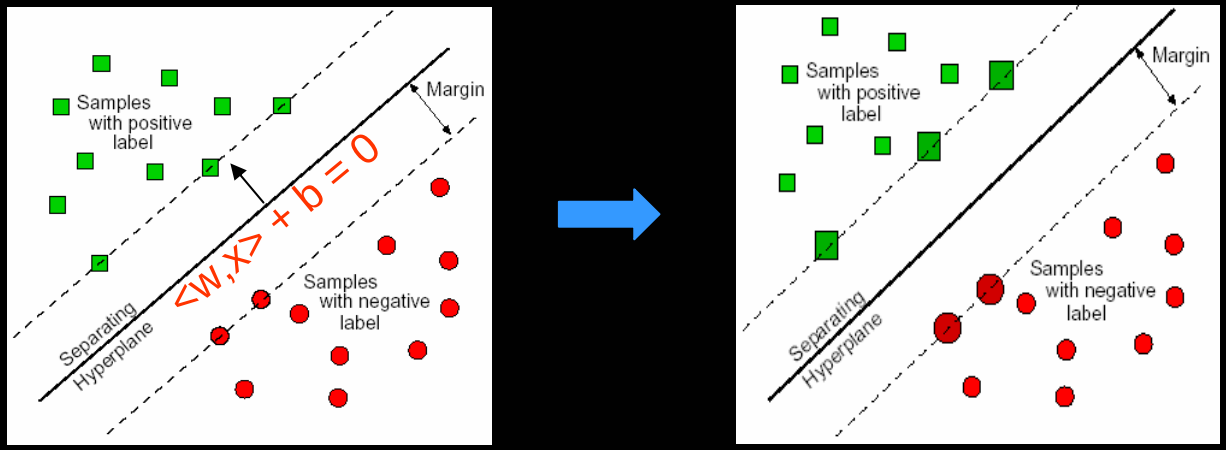
Características de estos algoritmos:

-Perfiles de expresión de genes conocidos en términos de su función (o ubicación)

-Una regla de desición que pueda explicar el set de entrenamiento

-El desafío: encontrar una regla de desición que permita generalizar hacia el resto de los datos

Ej. Support Vector Machine (SVM)

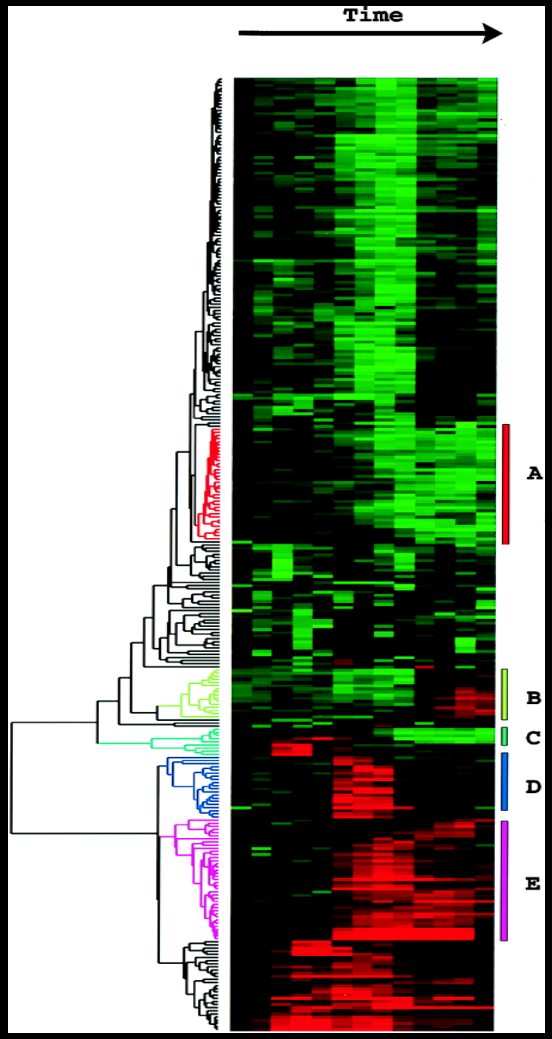


Los elementos más cercanos al hiperplano conforman el “Support Vector”

W y b se determinan en el entrenamiento

Medición de la expresión génica a través de microarrays

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



Transcriptional Program in the Response of Human Fibroblasts to Serum. *Iyer et al 1999. Science 283, 83-7.*

- Privación de suero 48hrs.

tpo 0: Estimulación con suero
Monitoreo cada 15min x 24hrs

•Disminución temporal de expresión de genes involucrados en biosintesis de colesterol (Cluster A)

•Aumento temporal de expresión de Genes involucrados en el ciclo celular (Cluster B)

•Aumento temprano de expresión en genes relacionados con remodelamianto de tejidos y angiogenesis (Cluster D y E)

Medición de la expresión génica a través de microarrays