Análisis de datos de microarray

Rodrigo Santamaría

Análisis de datos de microarray

Introducción

Tecnología de microarrays

Esquema de análisis

Fabricantes y BBDDs

Preprocesamiento

Análisis

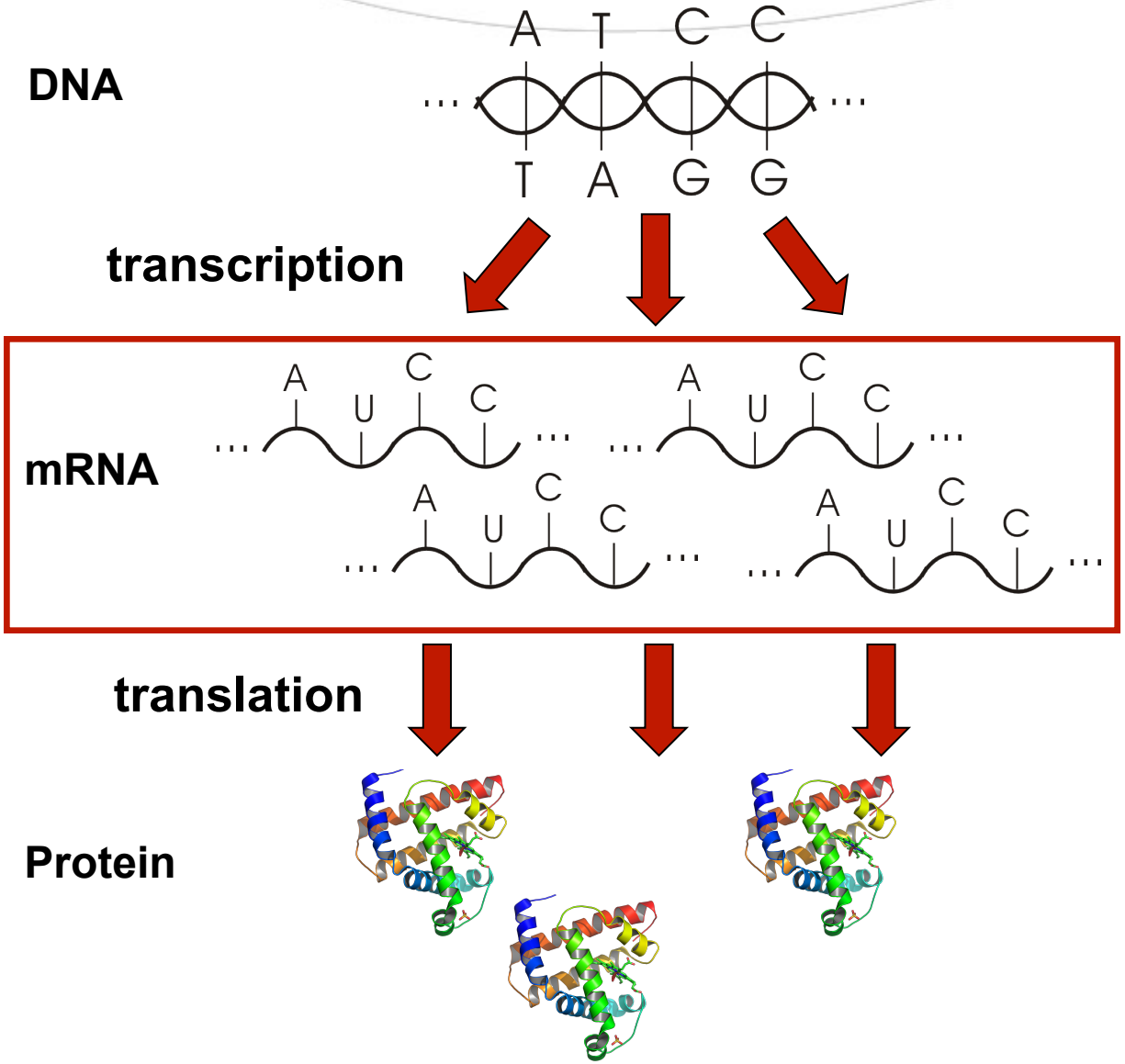
Validación

Otras tecnologías

Introducción

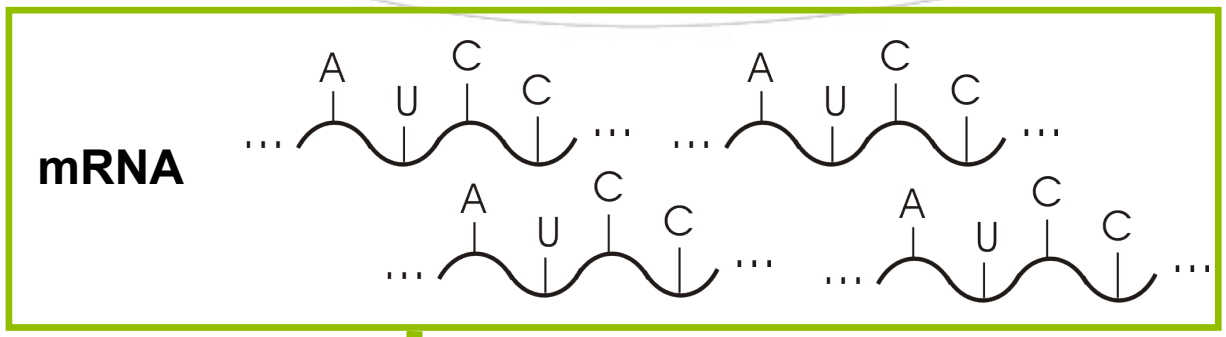
- Los experimentos de microarray son la herramienta más popular para realizar análisis de la expresión genética a gran escala
- A un nivel bioinformático, recibimos una matriz de intensidad con la que
  - Calculamos una matriz numérica normalizada (preprocesamiento)
  - Determinamos qué genes están más expresados (estadística inferencial)
  - Determinamos patrones de comportamiento comunes a distintos genes (estadística exploratoria y métodos de clasificación)

Expresión genética



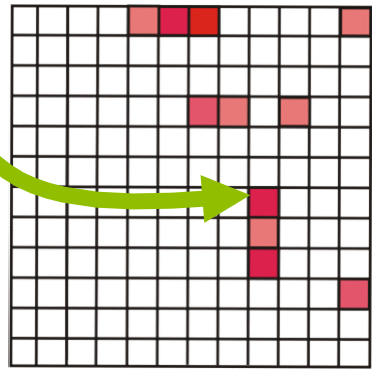
- La tecnología de microarrays se basa en el **dogma central** de la biología molecular
- Mide **los niveles de transcripción** de mRNA en un determinado caso de estudio
- El nivel de transcripción **NO** es equivalente al nivel de expresión, aunque se suele hacer dicha asunción
- La degradación del mRNA y otros procesos pos-transcripcionales influyen en la expresión y no son cuantificables con microarrays

Tecnología de microarray



Con la tecnología de microarray se "imprimen" las secuencias biológicas en un chip

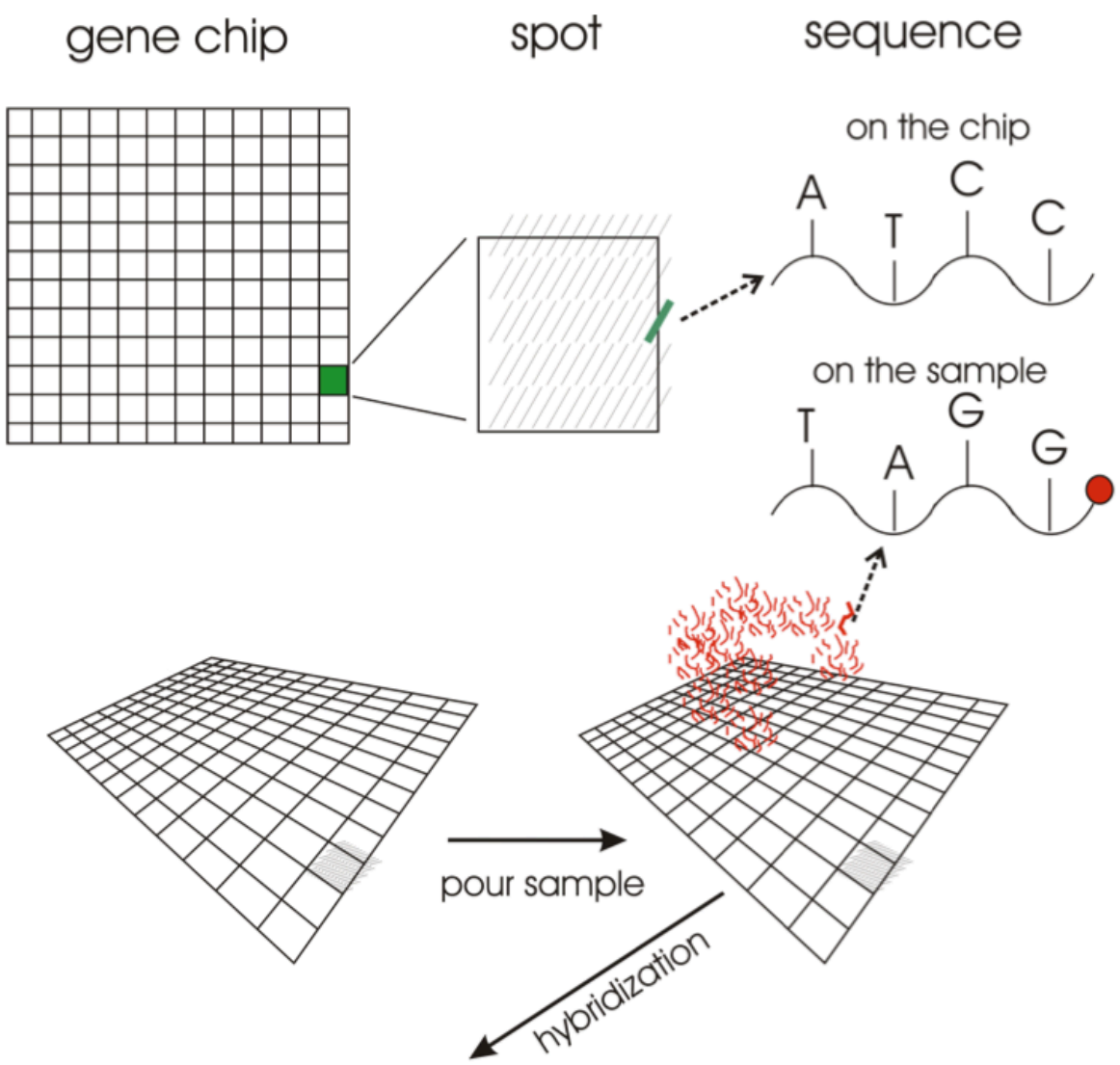
microarray chip



~

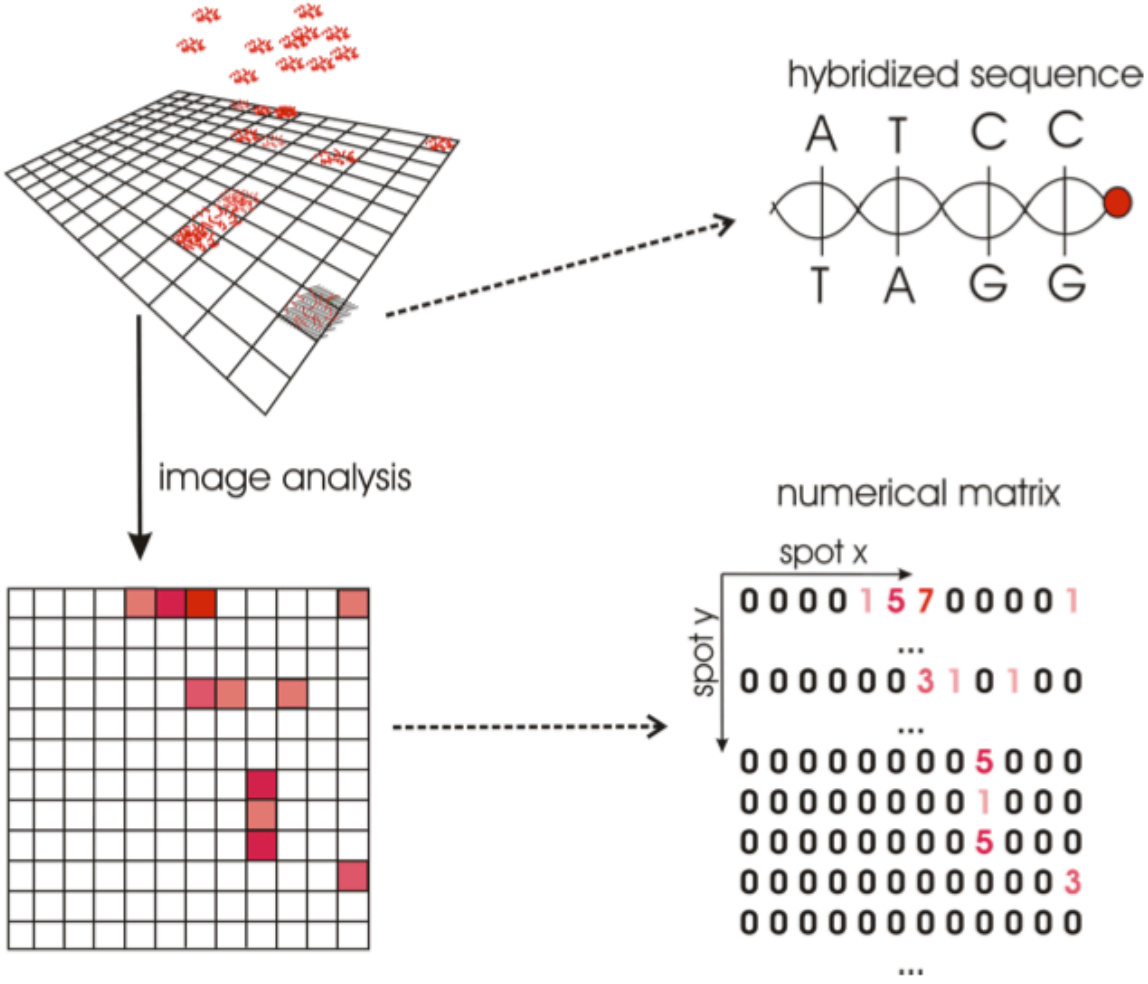
De manera que podemos cuantificar el nivel de transcripción en una matriz numérica

Microarray



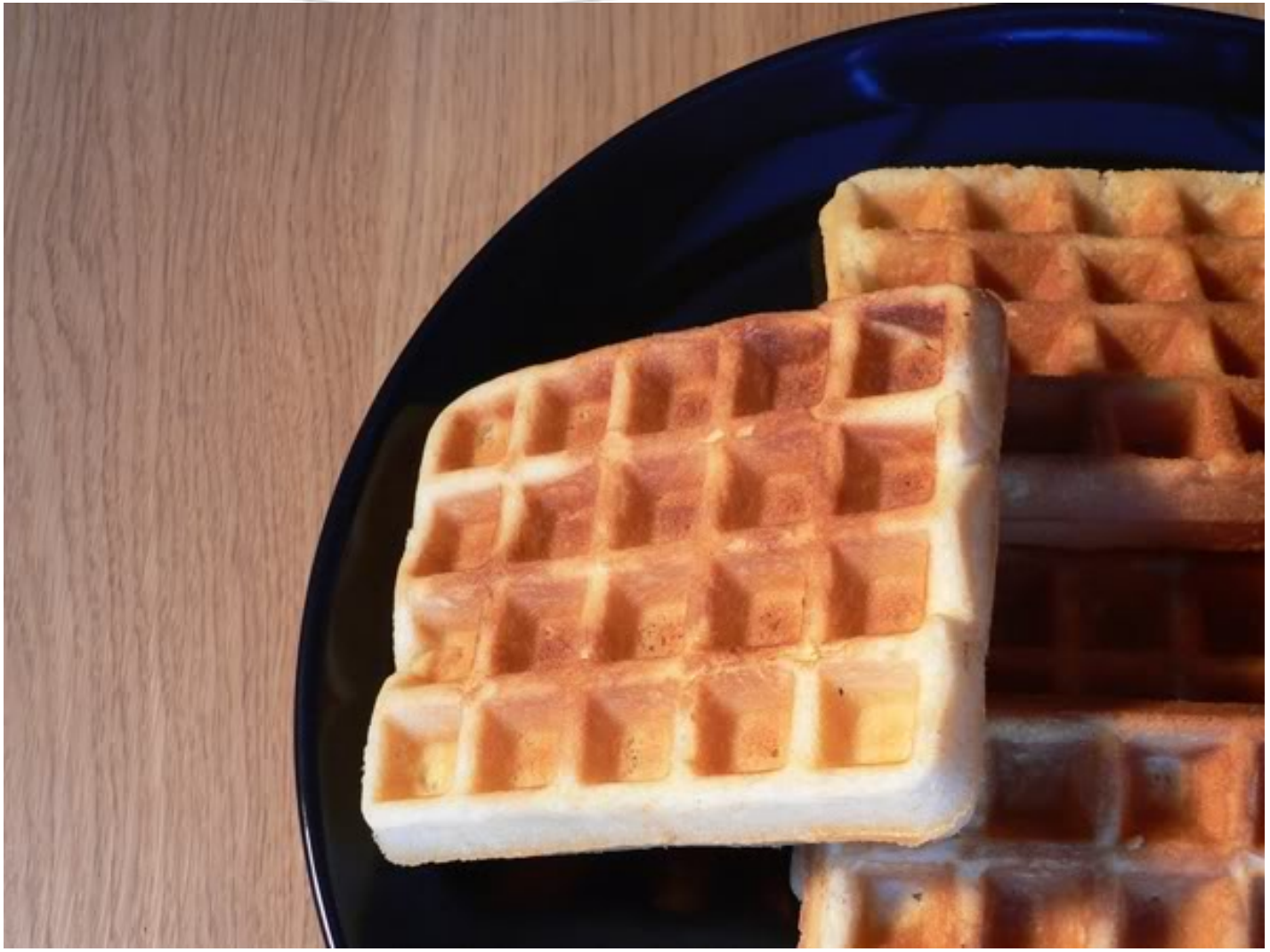
- En cada celda de un chip se "pegan" miles de copias de un segmento de mRNA (sonda)
- Celdas distintas contienen secuencias de mRNA distintas
- Todas las celdas contienen el mismo nº de segmentos

- Se prepara una muestra con mRNA de nuestro caso de estudio
- La muestras se marcan con una etiqueta fluorescente
- Nuestra muestra de estudio tendrá un número variable de mRNAs de cada segmento, según cuánto se transcriba

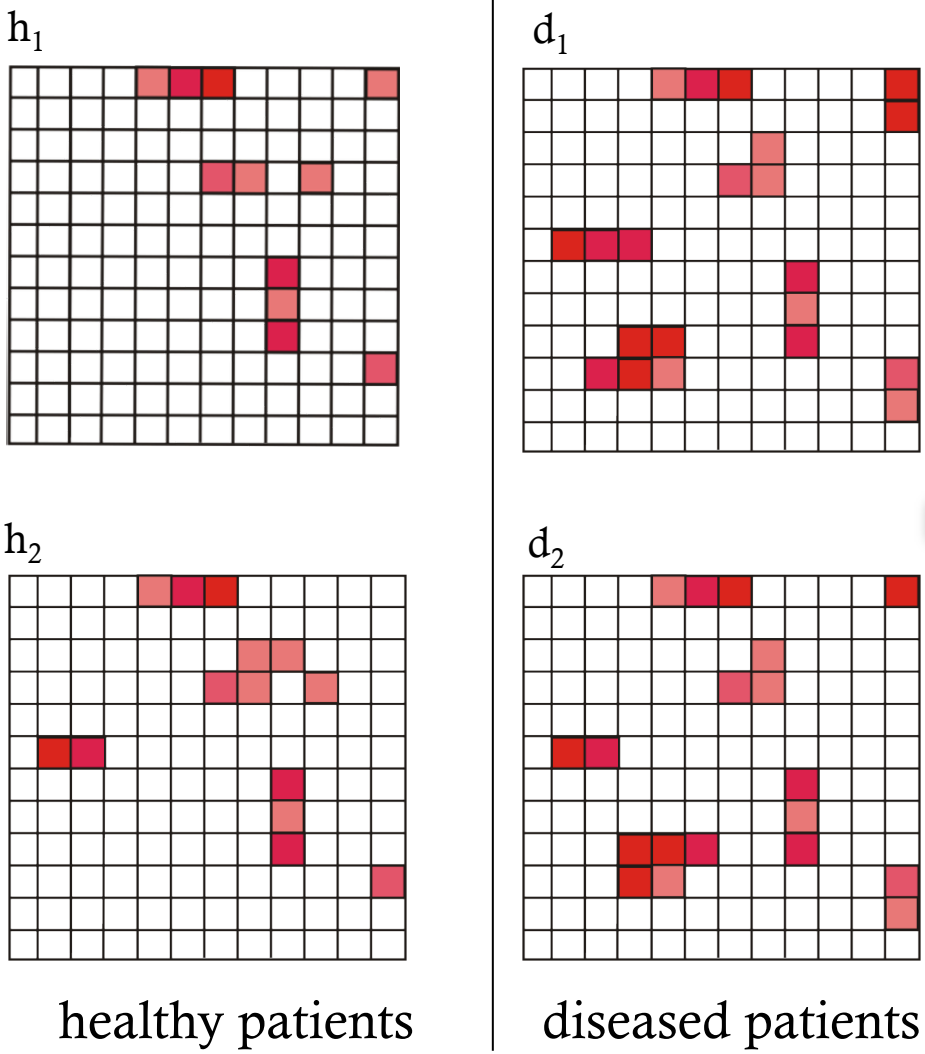


Hibridación: proceso por el que dos cadenas de RNA complementarias se combinan por atracción bioquímica

Análisis de la imagen: conversión de la cantidad de secuencias hibridadas (con etiqueta fluorescente) en una intensidad de luz (un número)

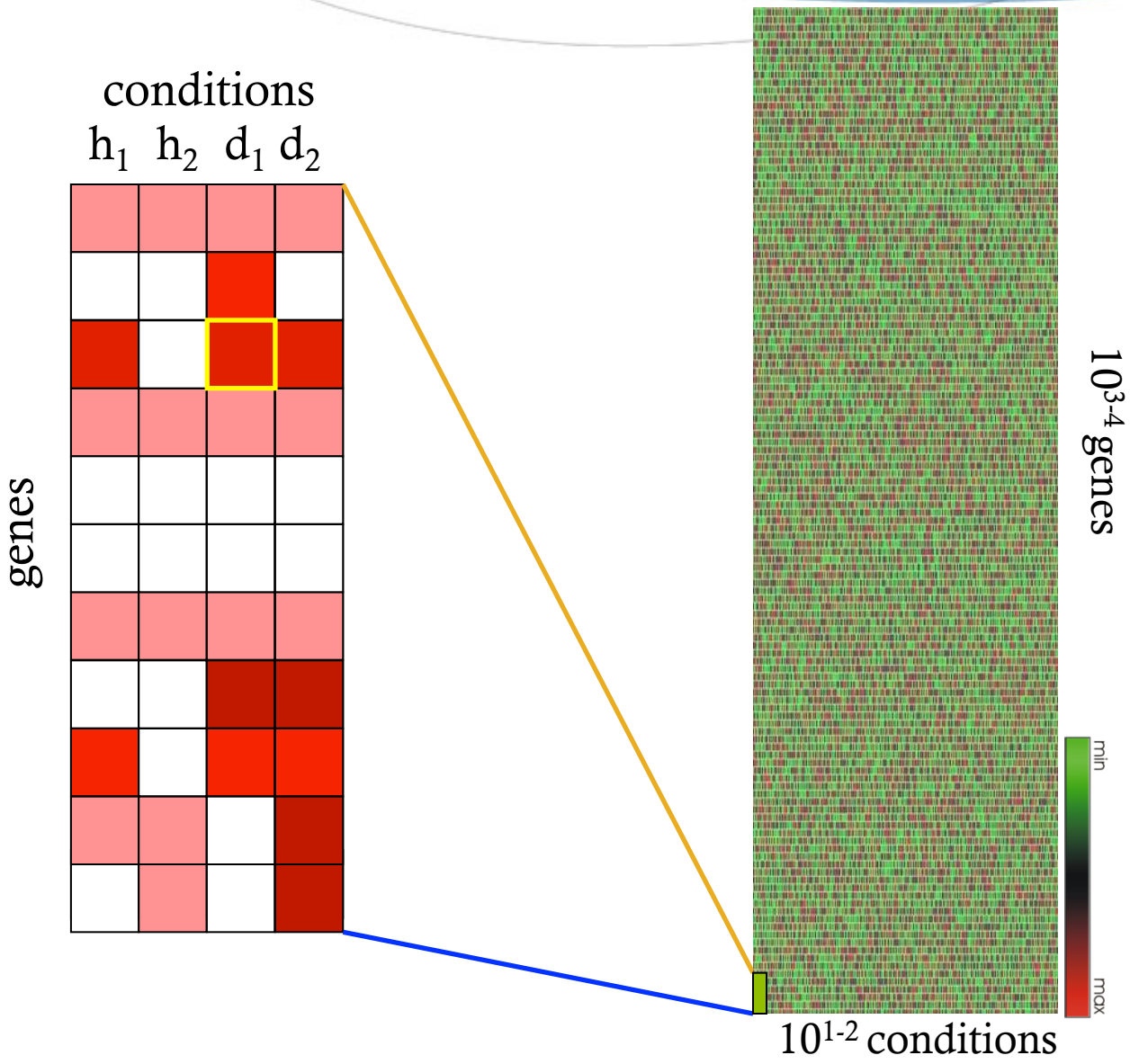


Matriz de expresión



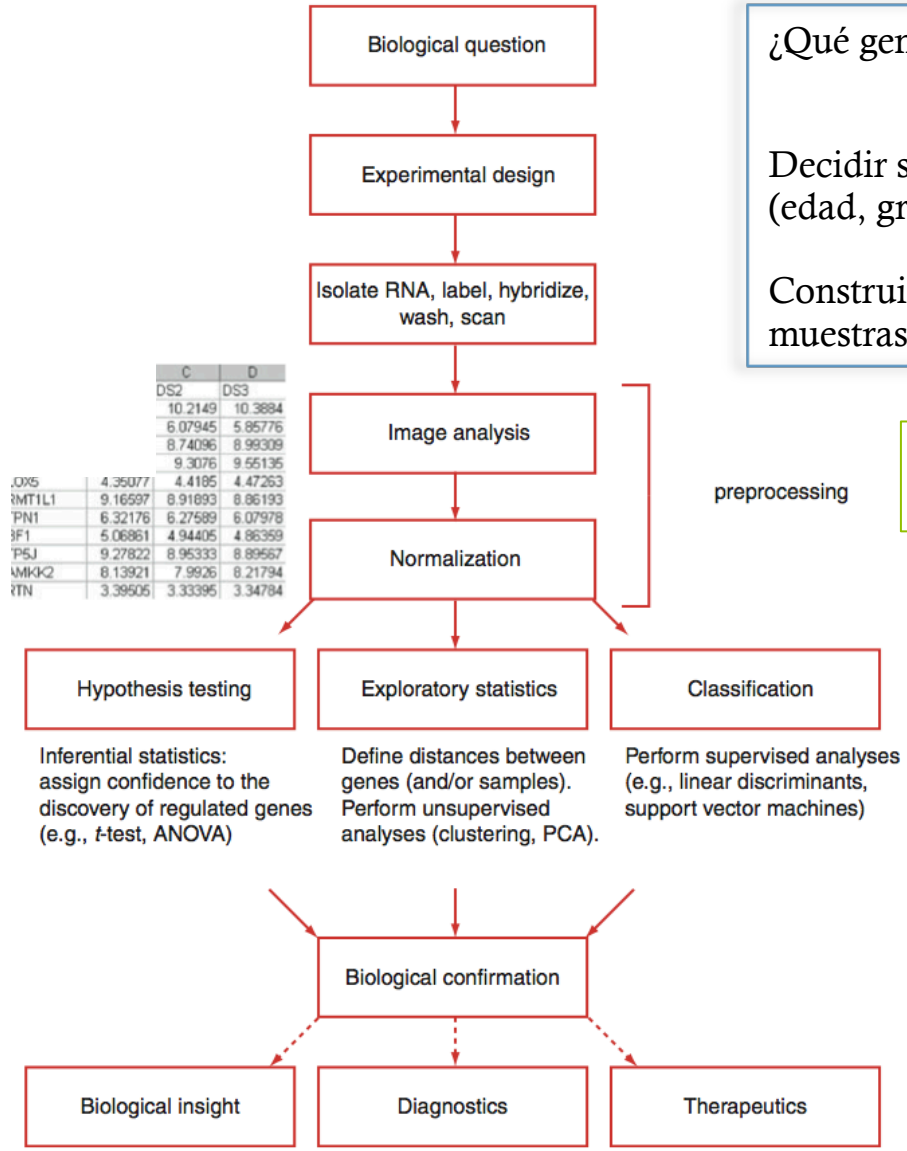
- **Matriz de expresión:** cada matriz de intensidad se "estira" a una columna de la matriz de expresión
- **Columna:** condiciones (arrays, ensayos, casos, muestras, factores experimentales)
- **Fila:** sondas o conjuntos de sondas. Se resumirán en genes en el preprocesamiento

Matriz de expresión



Una matriz de expresión tiene un tamaño considerable → necesidad de métodos informáticos para su análisis

Análisis de datos de microarray



¿Qué genes influyen en el cáncer de mama?

Decidir sobre los pacientes y los controles (edad, grado de la enfermedad, etc.)

Construir chips de microarray con las muestras recogidas

Convertir la información del microarray a una matriz numérica analizable

Realizar análisis inferenciales o exploratorios

Validar los resultados a partir del conocimiento existente y del diagnóstico en nuevos casos

Principales fabricantes

- Affymetrix
  - Es el principal fabricante y vendedor de chips
  - Muchos de los métodos de preprocesamiento han sido desarrollados por Affymetrix o para chips de Affymetrix
  - Los ficheros de intensidad de Affymetrix tienen extensión .cel
- Agilent
  - La segunda compañía en microarrays, heredera de HP
- Illumina
  - Introduce el concepto de microarray beads
  - Más centrada en la fabricación de secuenciadores (Solexa)

Terminología

- **Sobre-regulado** (upregulated): un transcrito con más intensidad que la intensidad base
  - También llamado activado, activo, sobreactivado
- **Infra-regulado** (downregulated): un transcrito con menor intensidad que la intensidad base
  - También llamado inhibido, inactivo o desactivado
- N-veces sobre/infra-regulado (n-fold up/down regulation): transcrito con intensidad N veces mayor/menor que una intensidad de referencia

Herramientas

- Existen dos BBDD principales que mantienen información sobre experimentos de microarray
  - GEO (NCBI) http://www.ncbi.nlm.nih.gov/geo/
  - ArrayExpress (EBI) http://www.ebi.ac.uk/arrayexpress/
- Existen distintas herramientas para adquisición de experimentos de microarray y su análisis
  - La más utilizada es el lenguaje de programación R junto con el paquete BioConductor

Preparación del entorno R

- Descargar R: http://www.r-project.org/
- Instalar Bioconductor: http://www.bioconductor.org/install/

```
source("http://bioconductor.org/biocLite.R")
biocLite() #Para instalar la base
biocLite("ArrayExpress") #Adquisición de microarrays
biocLite("GEOquery") #Adquisición de microarrays
biocLite("arrayQualityMetrics") #Análisis exploratorio
biocLite("affy") #Preprocesamiento (RMA)
biocLite("limma") #Paquetes adicionales microarrays
```

Preparación del entorno

- Para las explicaciones de esta sesión, usaremos dos experimentos de microarray:
  - GSE1397 (GEO): Experimento con muestras de cerebro de pacientes sanos y con síndrome de Down
    - Plataforma HG_U133A de Affymetrix
  - E-TABM-25 (ArrayExpress-AE): Experimento con muestras de distintas partes del cerebro de chimpancé a distintas edades
    - Plataforma HG_U95Av2 de Affymetrix
    - Los datos en crudo de intensidad están disponibles

Adquisición de datos

- Directamente a través de las páginas web de GEO y ArrayExpress
- Mediante BioConductor
  - bibliotecas GEOquery y ArrayExpress

```
library(GEOquery)
geo=getGEO("GSE1397”)

library(ArrayExpress)
ae=ArrayExpress("E-TABM-25”)
```

Análisis de datos de microarray

Introducción

Preprocesamiento

- Análisis exploratorio
- Corrección de fondo
- Normalización
- RMA

Análisis

Validación

Preprocesamiento

- La tecnología de microarrays y la preparación experimental pueden introducir ciertos artefactos en la medida de la expresión genética:
  - Artefactos debidos a la fluorescencia
    - Distinta eficiencia de las etiquetas fluorescentes
    - Variaciones en el rendimiento del escáner de fluorescencia
  - Artefactos debidos a la impresión
    - Variaciones en la densidad de impresión, superficies irregulares…
  - Artefactos debidos al experimento biológico
    - Diferencias en la pureza o calidad de las muestras biológicas
    - Diferencias en la manipulación de las muestras biológicas

Preprocesamiento

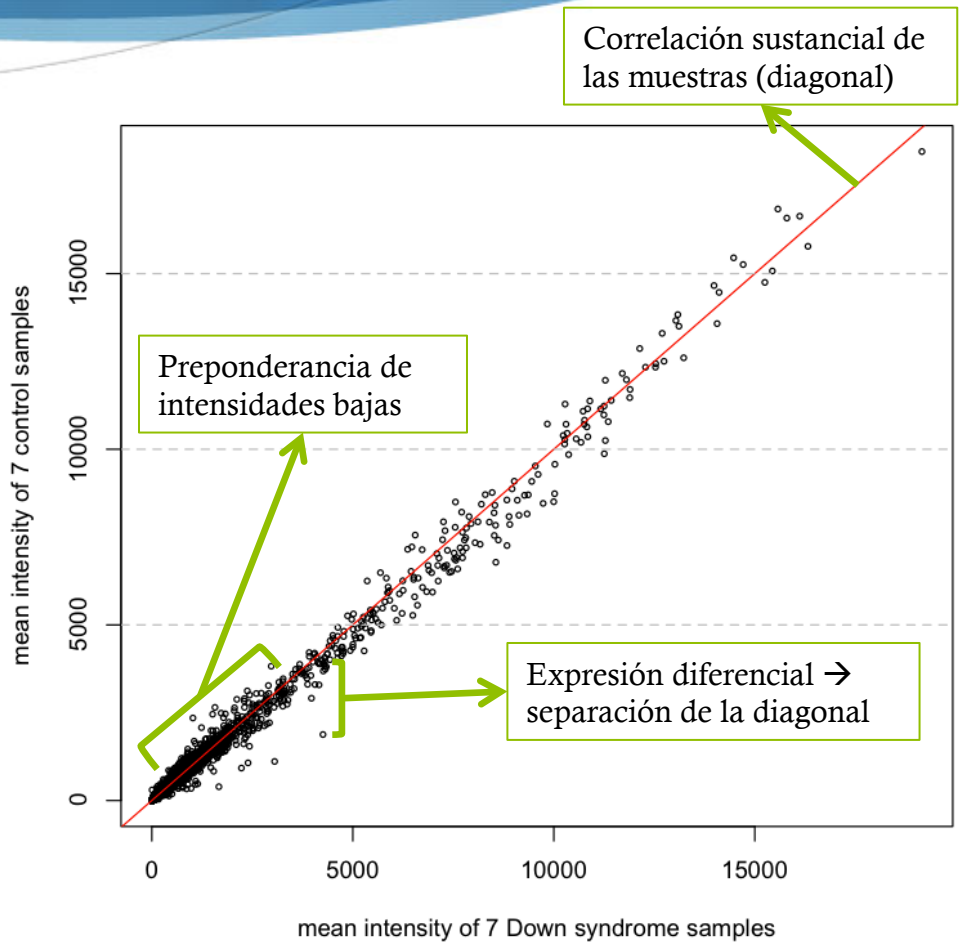
- El preprocesamiento trata de eliminar estas variaciones sistemáticas en los datos de la manera más completa posible, pero preservando la variación biológica real
- Hay cuatro pasos principales
  - Cuantificación de la imagen (no entraremos en detalle)
  - Exploración de los datos
  - Corrección del fondo, normalización y sumarización
  - Determinación de la calidad

Exploración de los datos

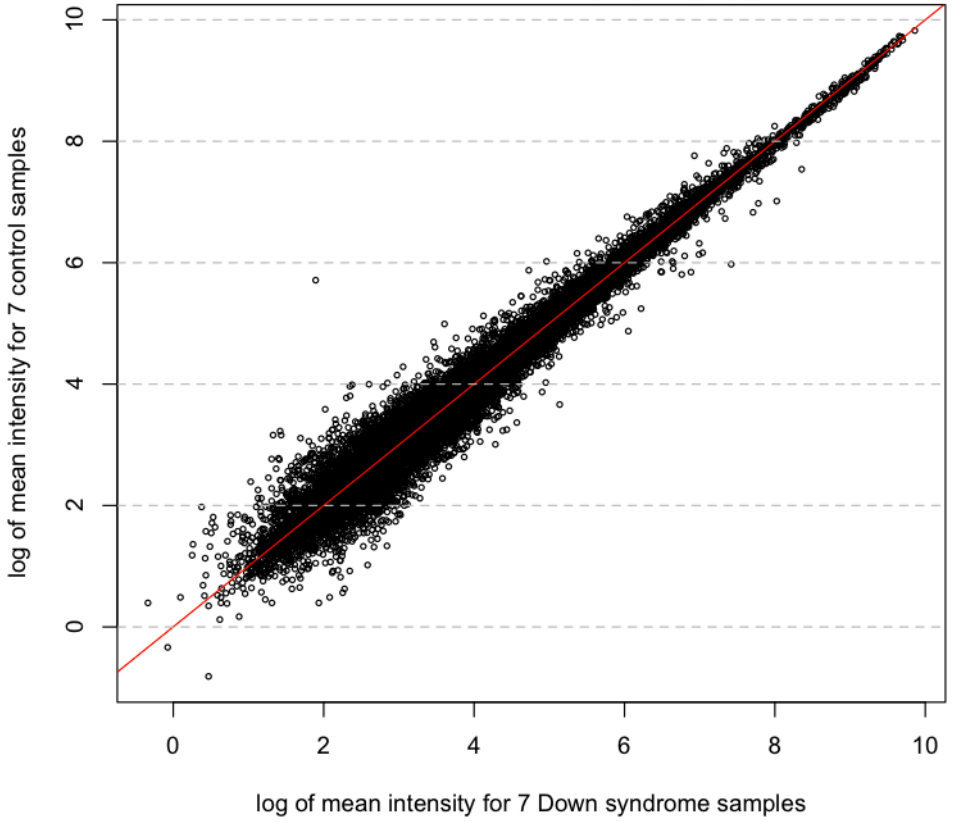
- Revisión inicial, mediante representaciones gráficas
  - **Scatterplots**: diagramas de dispersión que muestran la correlación de los niveles de expresión entre dos muestras
  - **MA plots**: evolución de los scatterplots que muestran los ratios de correlación
  - **Histogramas**: diagramas de distribución de los niveles de expresión en cada muestra del experimento
  - **Boxplots**: otro modo de mostrar la distribución de los niveles de expresión a lo largo de las muestras
- Su principal uso es detectar errores de bulto en el microarray

Exploración de los datos

- Scatter plots (diagramas de dispersión):
  - punto=transcrito
  - X=nivel de expresión para una condición o conjunto de condiciones
  - Y=nivel de expresión para otra condición o conjunto de condiciones
- Figura: caso típico de diagrama



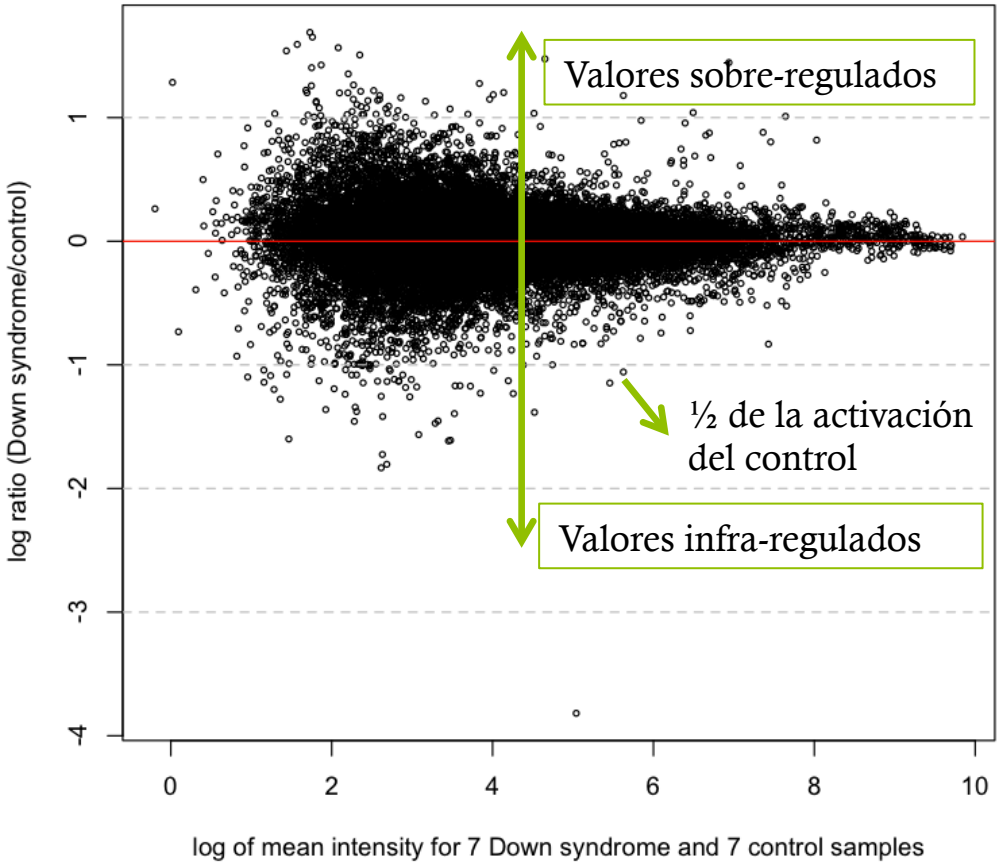
Diagramas de dispersión



- Se suele usar el logaritmo de las intensidades para que los datos estén distribuidos más uniformemente

MA plot

- Un diagrama de dispersión "girado" 45 grados
  - La intensidad no se suele medir contra la base del chip, si no con respecto a las muestras de control → ratios
- Punto: transcrito
- Y=log ratio de caso vs control (M)
- X=log de la media de la intensidad en todas las muestras (A)



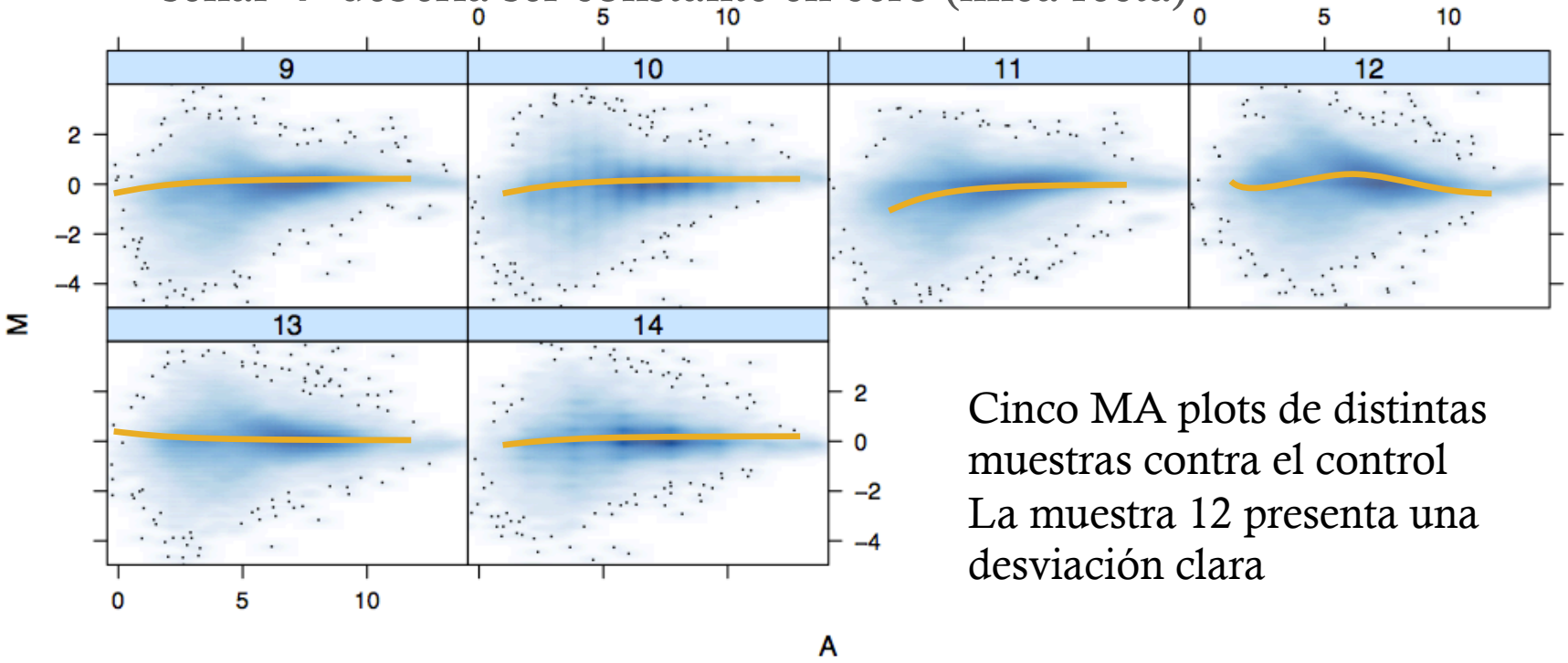
Ratios

Comportamiento	**Nivel Control (C)**	**Nivel Muestra (M)**	**Ratio (M/C)**	**Log Ratio $\log_2$(M/C)**
Nivel base de expresión	50	50	1.0	0.0
Sin cambio	50	50	1.0	0.0
Activación doble	50	100	2.0	1.0
Inhibición doble	50	25	0.5	-1.0

- Ejemplo de ratios para varios niveles de activación
- Los ratios tienen la interpretación más matemática
- Los los log ratios normalizan valores muy altos y tienen una interpretación simétrica y quizás más intuitiva

MA plots

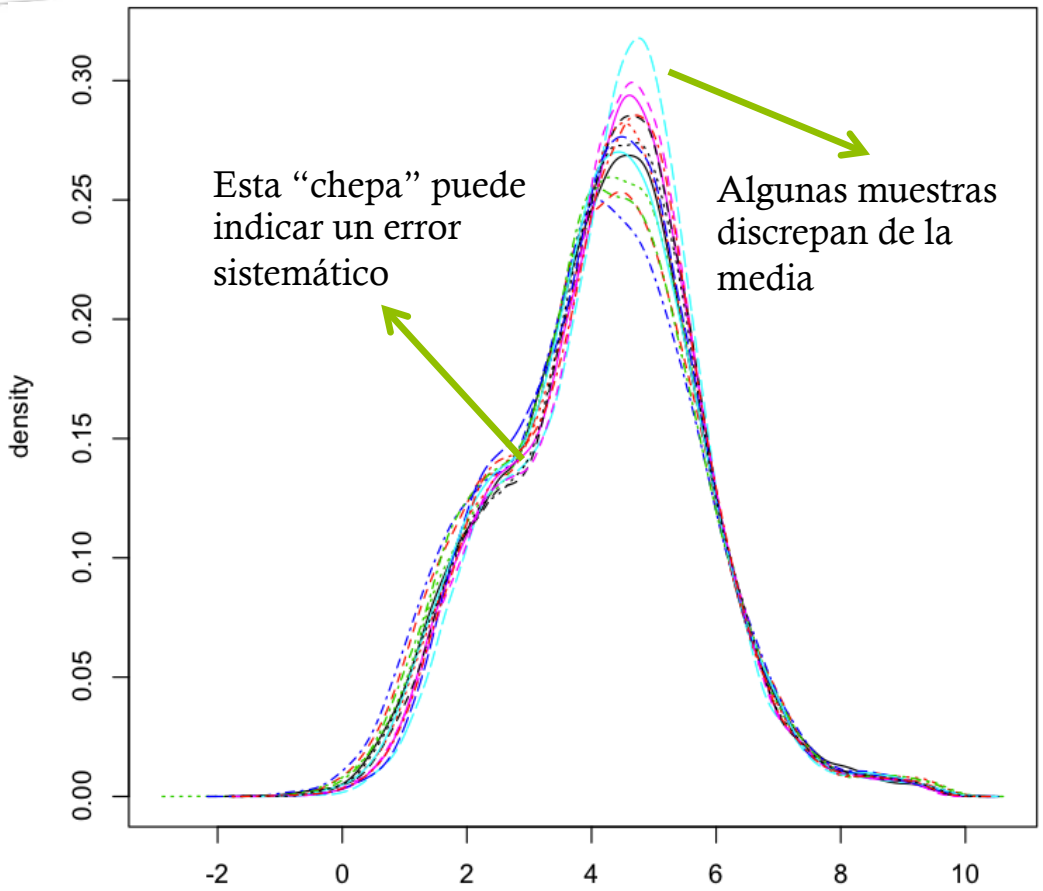
- Uno de los cometidos principales de los MA plots es conocer si los datos tienen una desviación dependiente de la intensidad de la señal → debería ser constante en cero (línea recta)



Cinco MA plots de distintas muestras contra el control
La muestra 12 presenta una desviación clara

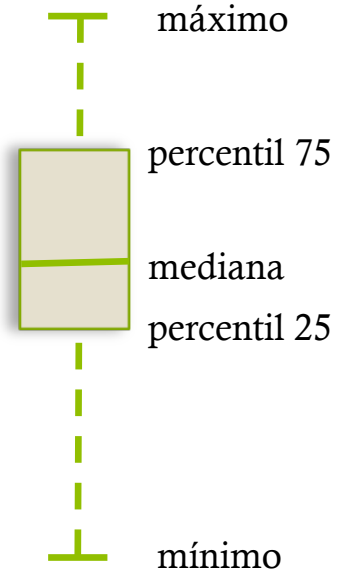
Histogramas

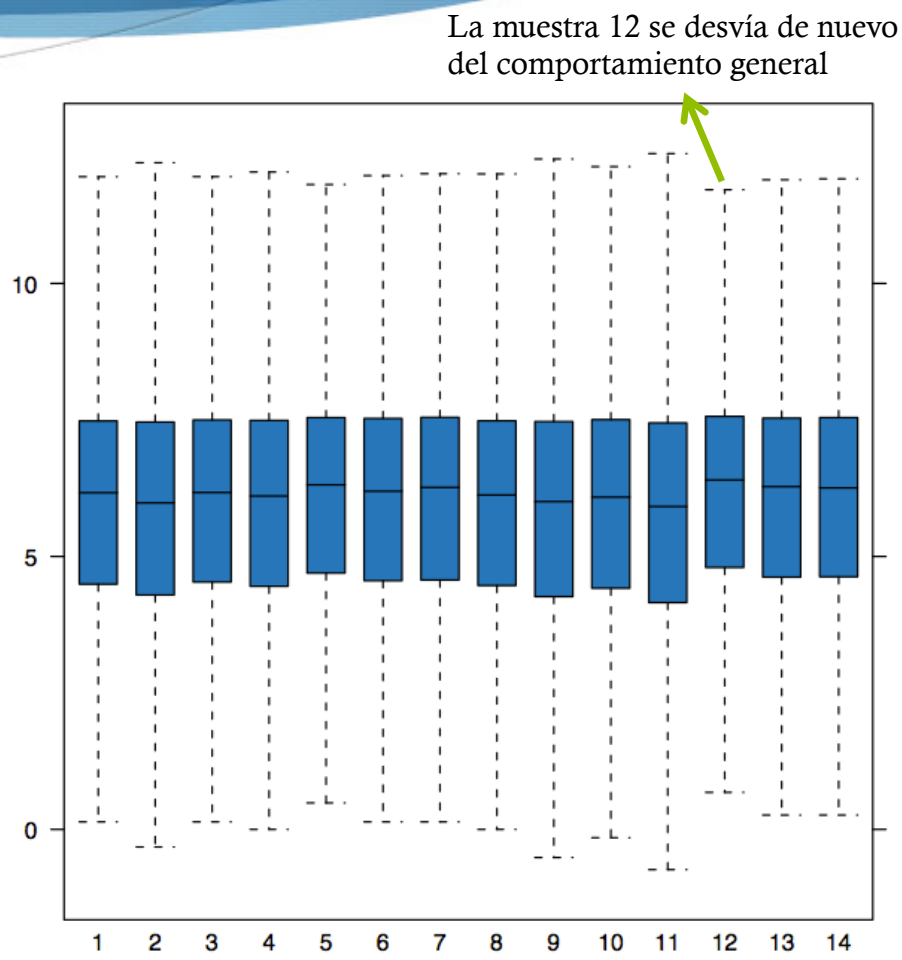
- Representación de la distribución de intensidades para cada muestra en el experimento
- Evaluación de la calidad de las muestras, se esperan
  - Formas similares
    - Alturas y anchuras
    - Posición
  - Distribución normal



Boxplots

- **Caja de Tukey**: resumen gráfico de algunos valores indicativos de la distribución





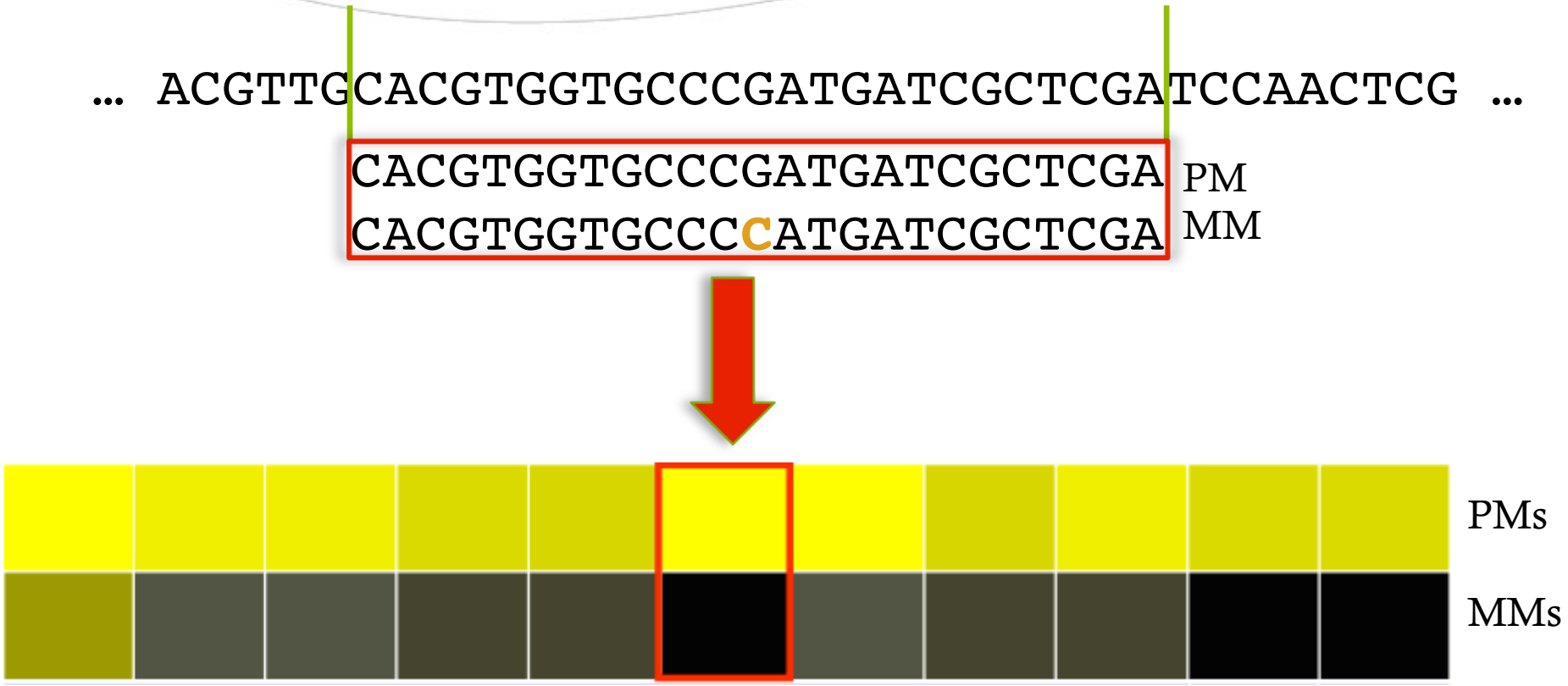
Normalización

- Corrección de dos o más muestras antes de comparar sus valores de expresión
- Suele constar de tres pasos
  - Corrección del fondo (background)
    - Estimar y eliminar la intensidad de ruido de fondo
  - Normalización global o local
    - Asegurar que la mayoría de las sondas varíen igual
  - Sumarización
    - Conversión de sondas o conjuntos de sondas a transcritos o genes

Corrección de fondo

- Affymetrix
  - En las celdas del chip, por cada sonda con la secuencia deseada, de 25 nucleótidos normalmente (PM), se coloca también una sonda con esa misma secuencia, pero con el nucleótido 13 cambiado al complementario (MM)
    - PM: Perfect Match, secuencia exacta
    - MM: MisMatch, secuencia cambiada
  - MM permite medir la hibridación no específica
    - Aquéllas sondas que se "pegan" sin tener la secuencia buscada
    - Es decir, mide el fondo debido a esta causa

Corrección de fondo



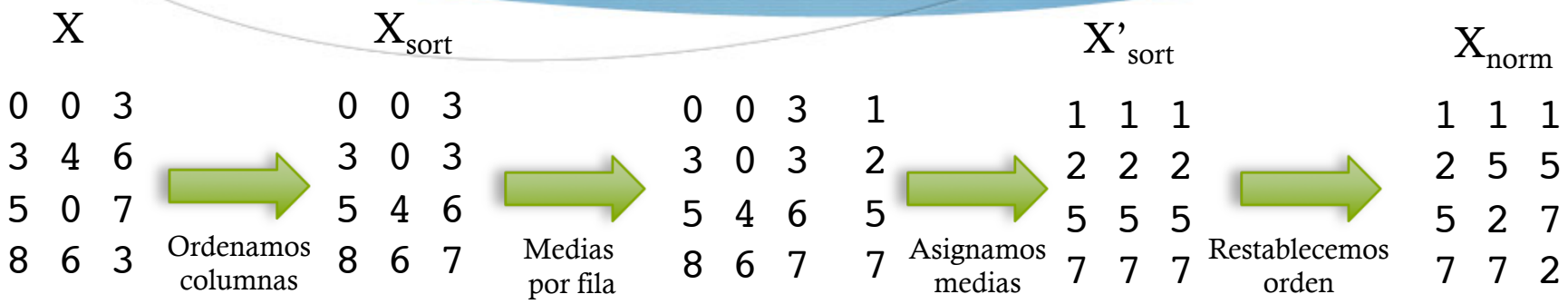
Normalización

- Parte de la hipótesis de que la mayoría de los genes en un microarray no varían su valor bajo diferentes condiciones experimentales
  - Su expresión media es cero (o su ratio medio es uno)
- Normalización paramétrica
  - Asume que los datos se parecen a una distribución normal
  - ANOVA y t-test son normalizaciones paramétricas muy usadas
- Normalización no paramétrica
  - No se asume ninguna distribución por defecto
  - La normalización por cuantiles es muy utilizada en microarrays

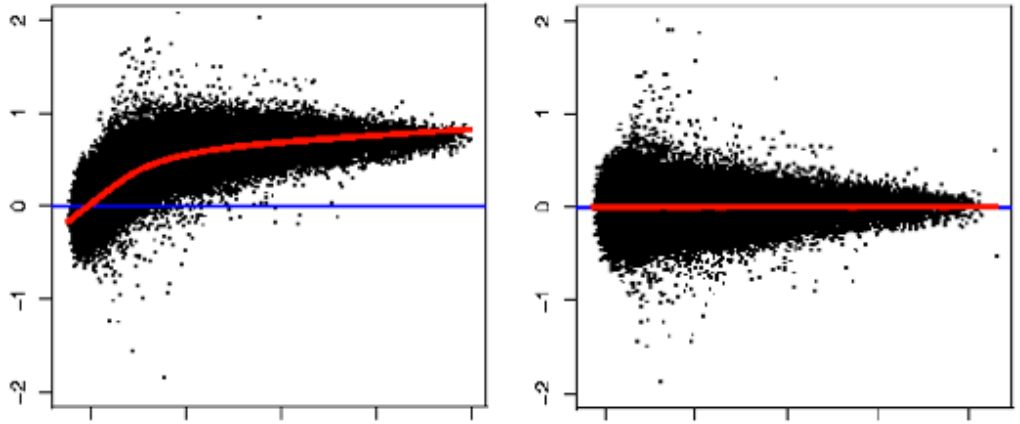
Normalización por cuantiles

- Asume que todos los arrays de nuestro experimento tienen la misma distribución (aunque no asume ninguna en particular)
- Método
  - Ordenar las columnas de la matriz de intensidad X → X_{sort}
  - Calcular la media de las filas de X_{sort}, y aplicar dichos valores a cada elemento → X'_{sort}
  - Restaurar el orden original de X → X_{norm}

Normalización por cuantiles



- Es la normalización más usada
  - Utilizada por Affymetrix
  - Incorporada por RMA
  - Artículo original: Bolstad et al. 2003
    - http://bmbolstad.com/misc/normalize/normalize.html



MAplot antes y después de la normalización por cuantiles (tomado de las figuras de Bolstad et al. 2003)

Housekeeping genes

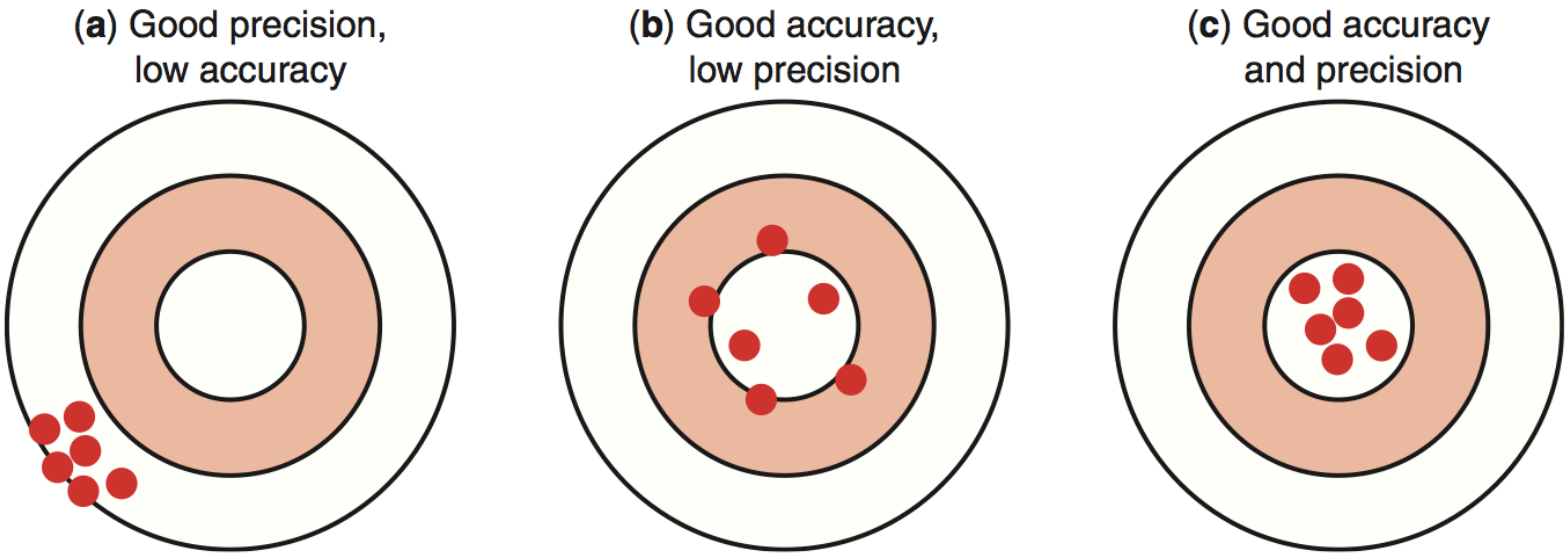
- Genes que tienen una intensidad constante independientemente de las condiciones experimentales
  - P. ej. debido a que son fundamentales para que la célula viva
- Se pueden utilizar para normalizar
  - Dividiendo todas las intensidades por la expresión media de los housekeeping genes
  - Problema: la asunción de que no varían no siempre es cierta
  - HuGE: proyecto que analiza 7000 genes en 19 tejidos para determinar una lista de 451 genes que nunca varían

Sumarización

- Para cada transcrito (o gen) tenemos varias sondas que miden su intensidad
  - Por ejemplo, en chips *affy* suele haber 11 sondas para cada transcrito (o conjunto de sondas – *probeset*)
- La sumarización es el proceso por el que determinamos la intensidad del transcrito a partir de la intensidad de sus sondas
  - Suele ser un proceso simple (calcular una media)

Precisión y Puntería

- El preprocesamiento busca mejorar la puntería (disminuir las tendencias erróneas) y la precisión (disminuir la varianza)

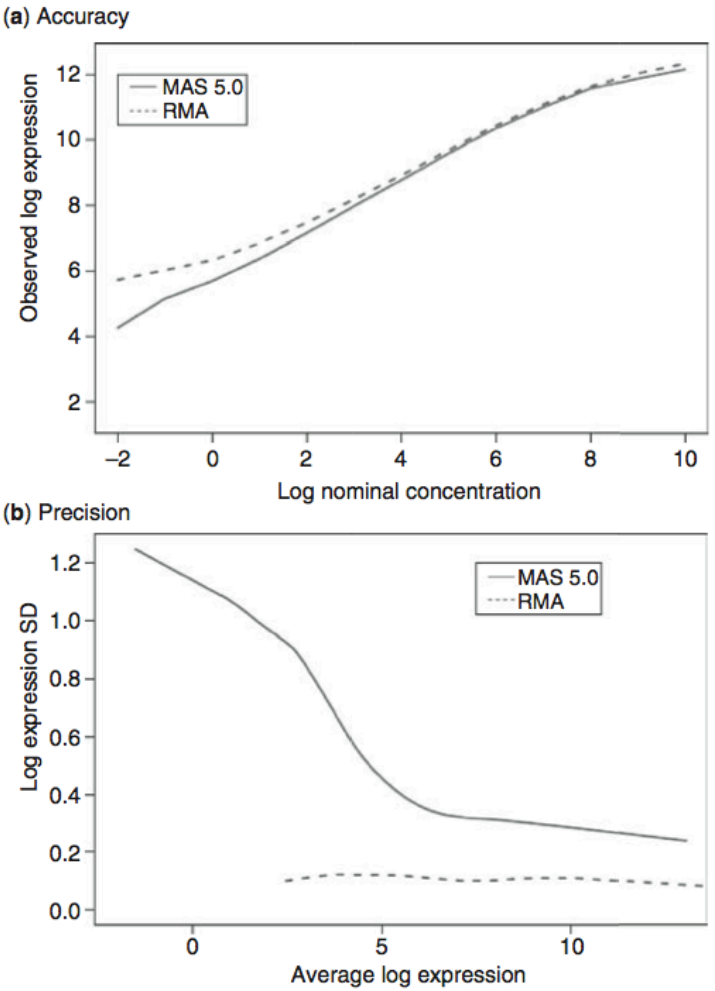


Precisión y Puntería

- Precisión
  - Una buena precisión se caracteriza por resultados reproducibles
  - Se comprueba mediante medidas repetidas de la misma muestra (réplicas técnicas)
- Puntería (accuracy)
  - Una buena puntería se comprueba mediante resultados bien conocidos, independientes de nuestro experimento
    - Por ejemplo, el uso de spike-in ARN, medidas conocidas de las concentraciones de unas determinadas sondas

Robust Multiarray Analysis (RMA)

- Método para realizar la corrección de fondo, normalización y sumarización en chips de Affymetrix
- Tiene una precisión mucho mayor que MAS 5.0 (el método de Affymetrix para preprocesar sus chips)
  - Corrección de fondo sin contar MM
  - Normalización por cuantiles
  - Estimación por median polish



RMA

- Corrección de fondo
  - RMA estima que MM contiene hibridación específica e inespecífica y no es por tanto útil para corregir el fondo
    - **Los MMs se descartan**
    - Sea *n* la sonda, *j* el conjunto de sondas al que pertenece e *i* el array
      - Se estima que $PM_{ijn} = bg_{ijn} + s_{ijn}$
        - bg_{ijn} es el fondo, tanto debido a hibridación inespecífica como a errores en el reconocimiento óptico, igual para todas las sondas de un mismo array *i*
        - s_{ijn} es la señal biológica que nos interesa extraer
      - Se utiliza un modelo de convolución para separar bg_{ijn} de s_{ijn}

RMA

- Normalización por cuantiles de los PMs
- Cálculo del $\log_2$ de los niveles de expresión
- Ajuste por median polish, considerando $\mu = 0$

$$x_{ijn} = \alpha_{in} + \beta_{jn} + r_{ijn}$$

- Artículo original (difícil de leer)
  - Irizarry et al. *Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data*. 2003

Median polish

- Método de ajuste lineal para matrices introducido por John Tukey (1977)
  - Utilizado en RMA por Irizarry et al. (2003) para la estimación de los valores de expresión
- La idea es que un valor de intensidad se puede ajustar como la suma de una constante de fondo, constantes por filas y columnas y residuos para cada valor

$$x_{ij} = \mu + \alpha_i + \beta_j + r_{ij}$$

Median polish

	X	
0	0	3
3	4	6
5	0	7
8	6	3

Median polish

	0	0	2
-4.5	0	0	1
-0.5	-1	0	0
0.5	0	-5	0
1.5	2	0	-5

β_j

α_i

$\mu = 4.5$

residuos (r_{ij})

RMA

- GCRMA
  - Ajuste de RMA para la extracción de hibridación no específica
  - Mejora la puntería de RMA, manteniendo su precisión
- fRMA (frozen RMA)
  - Modificación del método para el análisis de varios experimentos de microarray
    - No es lo mismo analizar varias muestras del mismo experimento que analizar varias muestras de distintos experimentos
      - Factores de variación introducidos por los laboratorios, preparación de las muestras, etc.

Análisis de datos de microarray

Introducción

Preprocesamiento

Análisis

Estadística inferencial

Test-t, ANOVA, limma

Estadística descriptiva

Clustering, PCA

Validación

Análisis

- Una vez preprocesados, tenemos dos tipos de análisis sobre datos de microarrays
  - **Estadística inferencial**: determinar qué genes están expresados diferencialmente (DEGs) y si dicha expresión es significativa
  - **Estadística descriptiva**: determinar grupos de genes que presentan patrones similares
    - **Análisis no-supervisado**: sin tener información de la estructura de los datos en el microarray
    - **Análisis supervisado**: contando con información de la estructura

Estadística inferencial

- Umbrales de expresión
  - La forma más evidente y sencilla de determinar DEGs es realizar el ratio de expresión entre una condición experimental y el control, y tomar los genes con un ratio mayor (o menor) que un umbral
  - Es una forma rápida de determinar los genes muy expresados diferencialmente, pero
    - Los umbrales sólo se pueden establecer de una manera arbitraria
    - No podemos determinar la significatividad estadística de su expresión diferencial

Test-t

- Testeo de hipótesis
  - **Hipótesis nula H_0**: no hay diferencia de señal entre las condiciones que estamos testeando
  - **Estadístico**: figura matemática que caracteriza los datos de expresión y en función suya rechacemos o aceptemos H_0
  - **Nivel de significación (α)**: probabilidad de rechazar H_0 cuando es verdadera (~ probabilidad de un falso positivo)
    - Típicamente α <0.05 (ver temas anteriores sobre la significatividad estadística y la probabilidad de rechazo)

Test-t

- **Estadístico-*t***
  - Es el más usado para probar H_0 en microarrays
  - Sea una matriz de expresión con n condiciones experimentales de tipo A y m condiciones de tipo B
    - A y B pueden significar, por ejemplo, "enfermedad" y "control"
    - Para un determinado gen, tenemos los niveles de expresión:
      - $(x_{a1}, \ldots, x_{an})$ y $(x_{b1}, \ldots, x_{bm})$
      - Con media y varianza (para la condición A, análogo para B):

$$\bar{x}_A = \frac{1}{n}\sum_{i=a_1}^{a_n} x_i \qquad s_A^2 = \frac{1}{n-1}\sum_{i=a_1}^{a_n}(x_i - \bar{x}_A)^2$$

Test-t

- El estadístico-*t* mide la expresión diferencial teniendo en cuenta el ratio entre señal (~media) y ruido (~varianza) en el experimento:

$$estadístico - t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n} + \frac{s_B^2}{m}}}$$

- A partir del estadístico-t podemos calcular un *p*-valor que comparar con el nivel de significatividad para aceptar o rechazar H_0
  - Un $p = 0.01$ quiere decir que de cada 100 tests uno será un falso positivo

Test-t

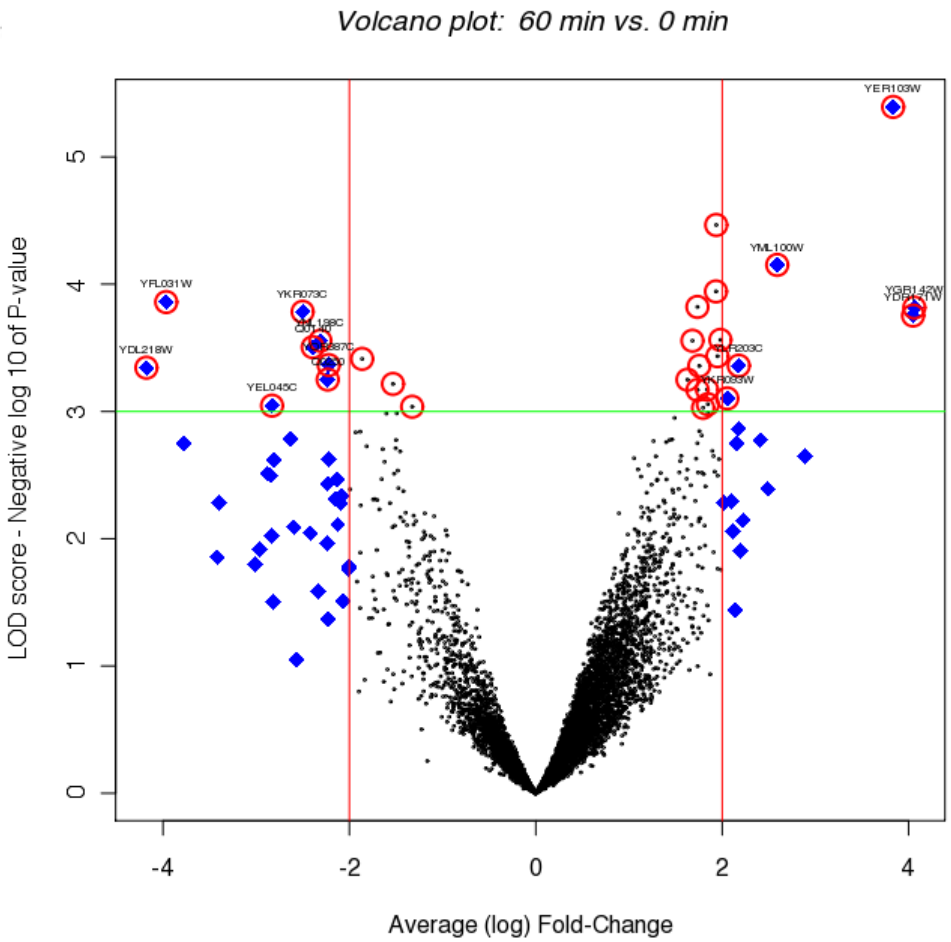
- Un test-t asume que los valores de expresión siguen una distribución normal (tests paramétricos)
  - Los tests no paramétricos (como el de Wilcoxon) no asumen una distribución normal, pero no se suelen usar en microarrays
- Es importante tener en cuenta el diseño experimental cuando se aplica un test-t
  - Cada condición puede tener uno o más factores experimentales
    - Por ejemplo: edad (31), estado (leucemia A), sexo (M)
  - Podemos comparar factores siempre que sea razonable, lo que generalmente implica tener intersecciones nulas
    - Por ejemplo, pacientes masculinos con leucemia A vs sanos
    - Probablemente no tendrá sentido comparar pacientes masculinos con pacientes con leucemia B

Múltiples comparaciones

- Un $p < 0.01$ para un test nos dice que hay un 1% de obtener un falso positivo
  - Si tenemos 10000 tests, quiere decir que tendremos ~100 falsos positivos!
- Necesitamos redefinir los límites cuando hacemos múltiples comparaciones para evitar
  - Errores tipo I (falsos positivos)
  - Errores tipo II (falsos negativos)
- Corrección de Bonferroni, FDR, FWER → ver temas anteriores

Volcano plot

- Representación de genes según su expresión diferencial y significatividad estadística
  - Punto = gen
  - X = expresión diferencial
  - Y = significatividad estadística
- Figura: volcano plot para genes de levadura a los 0 minutos y a los 60 minutos de aplicarle calor
  - Eje Y: $-\log_{10}$(p-valor)
  - Eje X: $\log_2$(ratio)



ANOVA

- El análisis de varianza (ANalysis Of VAriance) es un método apropiado si queremos comparar más de dos condiciones
  - Por ejemplo, múltiples puntos temporales, o control contra dos tipos de tratamiento o de enfermedad
- ANOVA es un modelo que toma la siguiente forma:

$$Y = \mu + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

  - Y es una función de X bajo distintas condiciones ($x_1 \ldots x_n$)
    - $\beta_1 \ldots \beta_n$ son los pesos dados a estas condiciones
  - ε es el error o residuo, no explicable por el modelo

limma

- El modelo lineal de análisis (*limma*) es un modelo de ANOVA, pero diseñado para cada gen por separado, en vez de hacer un único modelo general y aplicárselo a cada gen
  - Es muy utilizado en análisis de datos de microarray

- Todos estos modelos, al igual que el test-t, nos darán un p-valor para la significatividad de la expresión genética en cada contraste de condiciones

Estadística descriptiva

- "La maldición de la dimensionalidad"
  - Cada una de nuestras muestras tiene tantas dimensiones como genes (para humano, aprox. 20000)
    - Cada condición se puede ver como un punto de 20000 dimensiones
      - Es imposible imaginar un espacio de 20000 dimensiones
      - Comparar dos puntos de 20000 dimensiones generalmente dará distancias muy largas y aproximadamente iguales
    - Análogamente, cada gen tiene tantas dimensiones como condiciones
  - Necesitamos métodos matemáticos para explorar estos datos que se encuentran en un espacio altamente dimensional

Estadística descriptiva

- Hay muchas técnicas en estadística descriptiva, nos centraremos en las dos más utilizadas
  - Clustering
  - Análisis de Componentes Principales (PCA)
- En ambos casos, tratamos de reducir la dimensionalidad del problema para extraer conclusiones sobre el comportamiento de los genes en nuestros experimentos.
  - En ambos casos, necesitaremos definir alguna medida de similitud entre los datos

Clustering

- Es probablemente la técnica más utilizada para encontrar grupos de genes o condiciones en microarrays
  - Eisen et. al (1998) popularizaron su uso en microarrays
- El clustering es la agrupación (cluster significa literalmente “racimo”, se traduce por grupo) de elementos según las distancias entre ellos
  - Típicamente, se usa la distancia euclídea
- El resultado de estas agrupaciones se representa mediante
  - Dendrogramas (árboles de similitud)
  - Diagramas de dispersión
- El clustering jerárquico tiene dos fases
  - Cálculo de distancias entre genes o condiciones (euclídea, pearson, etc.)
  - Construcción del árbol a partir de las distancias (aglomerativo o divisivo)

Clustering jerárquico

- Cálculo de distancias
  - Hay varias, las más usadas son
    - Distancia Euclídea
    - Coeficiente de correlación de Pearson
  - Otras distancias son
    - Manhattan
    - Canberra
    - Binaria
    - Minkowski
    - Información mutua

- Sean dos puntos n-dimensionales $x = (x_1, \ldots, x_n)$ e $y = (y_1, \ldots, y_n)$

$$d_{euclidea} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$r_{pearson} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

$$d_{manhattan} = \sum_{i=1}^{n} \left(|x_i - y_i| \right)$$

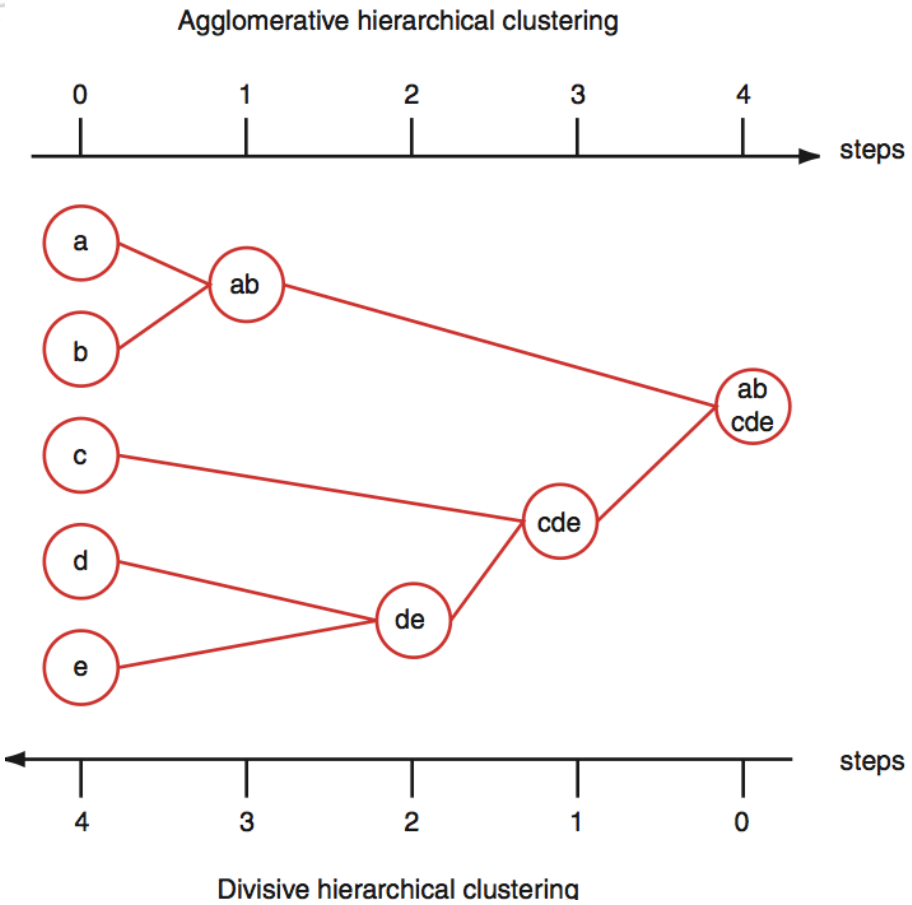
Coeficiente de correlación de Pearson

- Si dos variables X e Y varían juntas, se dice que correlacionan
- El coeficiente de correlación de Pearson (r) varía entre -1 (perfecta correlación negativa) a 1 (perfecta correlación positiva), pasando por 0 (no correlación o independientes)
- Para dos series de valores $X=\{X_1, \ldots, X_n\}$ e $Y=\{Y_1, \ldots, Y_n\}$, con medias $\overline{X}$ e $\overline{Y}$ y desviaciones típicas σ_x, σ_y:

$$r = \frac{\sum_{i=1}^{N}\left[\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]}{(N-1)\sigma_x\sigma_y}$$

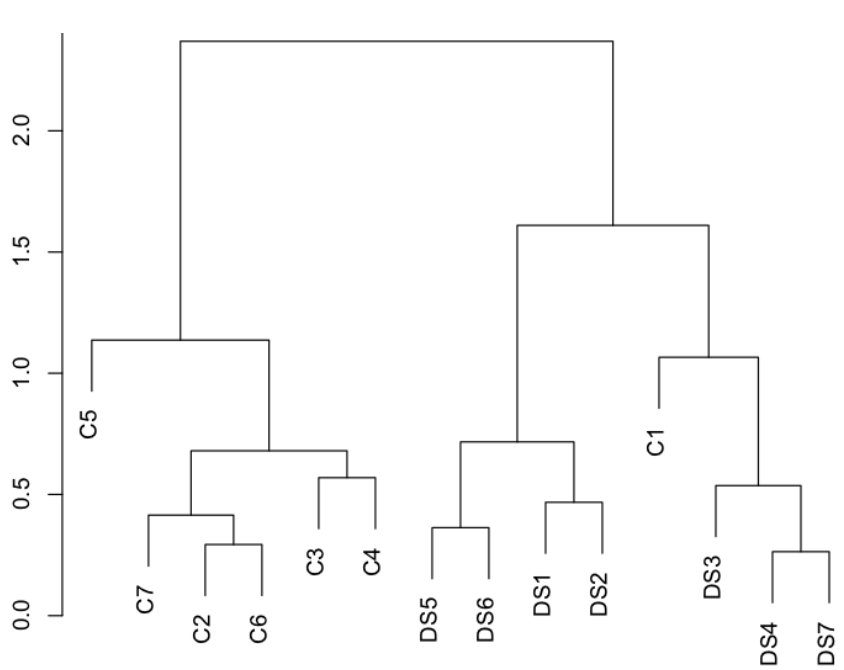
Clustering jerárquico

- Construcción del árbol
  - **Técnica aglomerativa**: se considera cada elemento por separado y se van uniendo los que tienen distancias más pequeñas
  - **Técnica divisiva**: técnica inversa, se considera todo el conjunto de elementos y se van separando los que tienen distancias más grandes

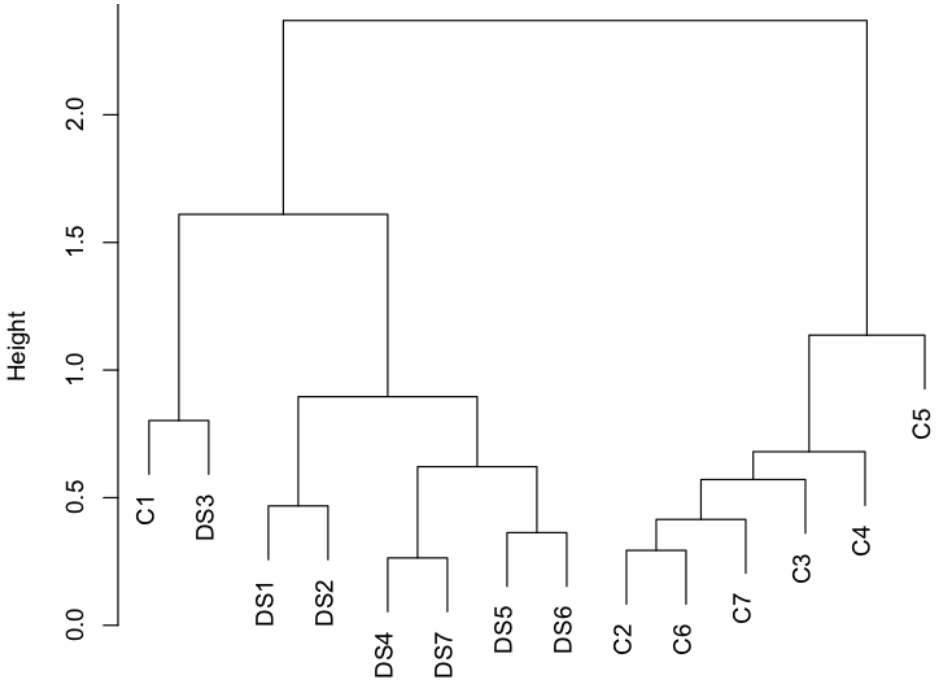


Clustering jerárquico

- Aunque muy similares, el clustering aglomerativo suele dar más precisión en la raíz del árbol y el divisivo en las hojas



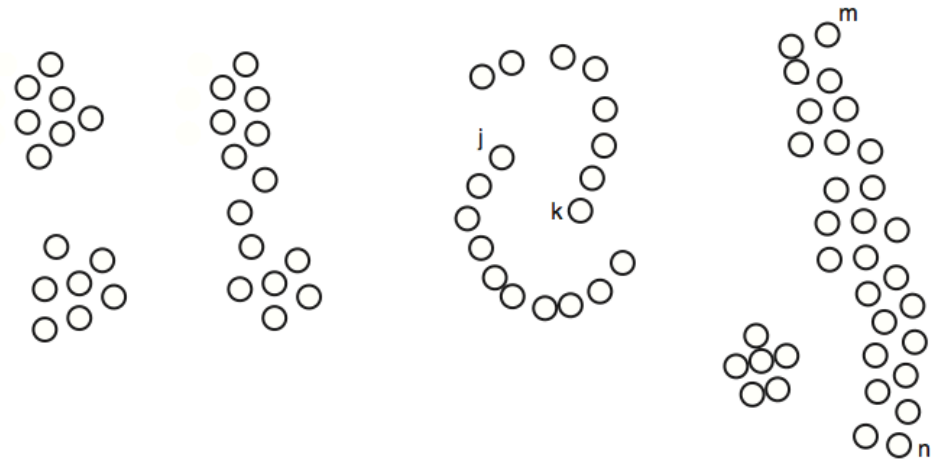
aglomerativo

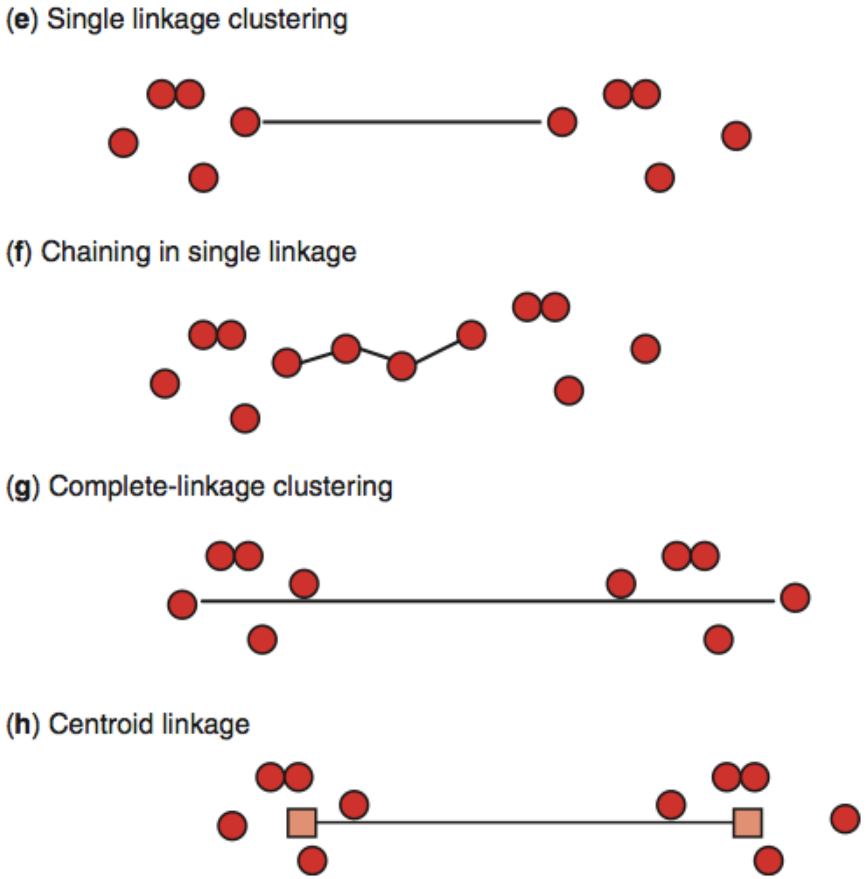


divisivo

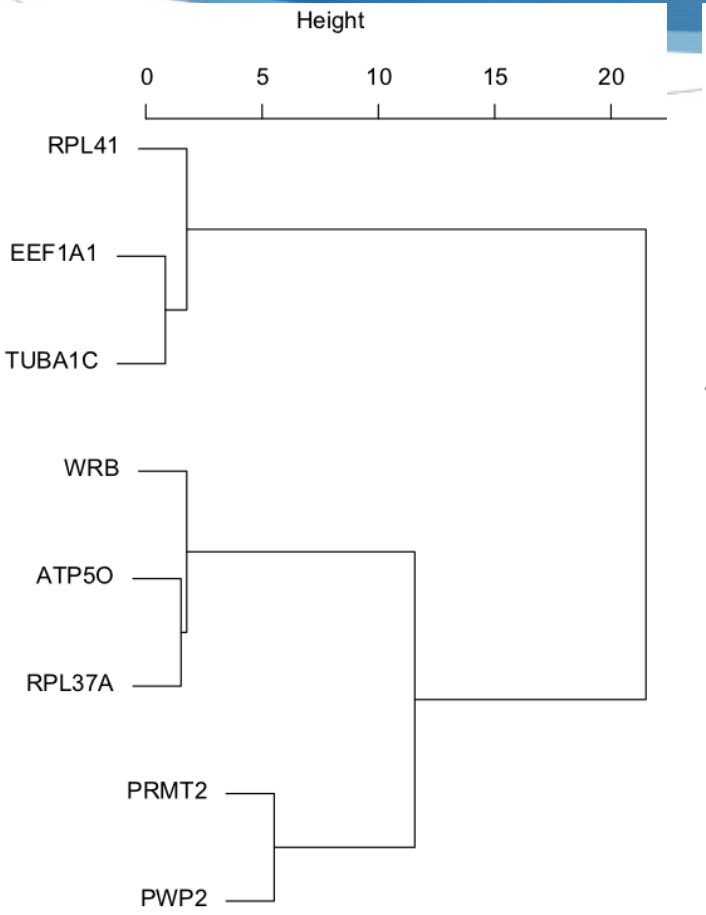
Clustering jerárquico

- Linkage: modo de definir la distancia entre dos clusters
  - **Single**: dada por los elementos más cercanos
    - Problema: fenómeno de la cadena
  - **Complete**: dada por los más lejanos
  - **Centroid**: dada por el centroide del cluster

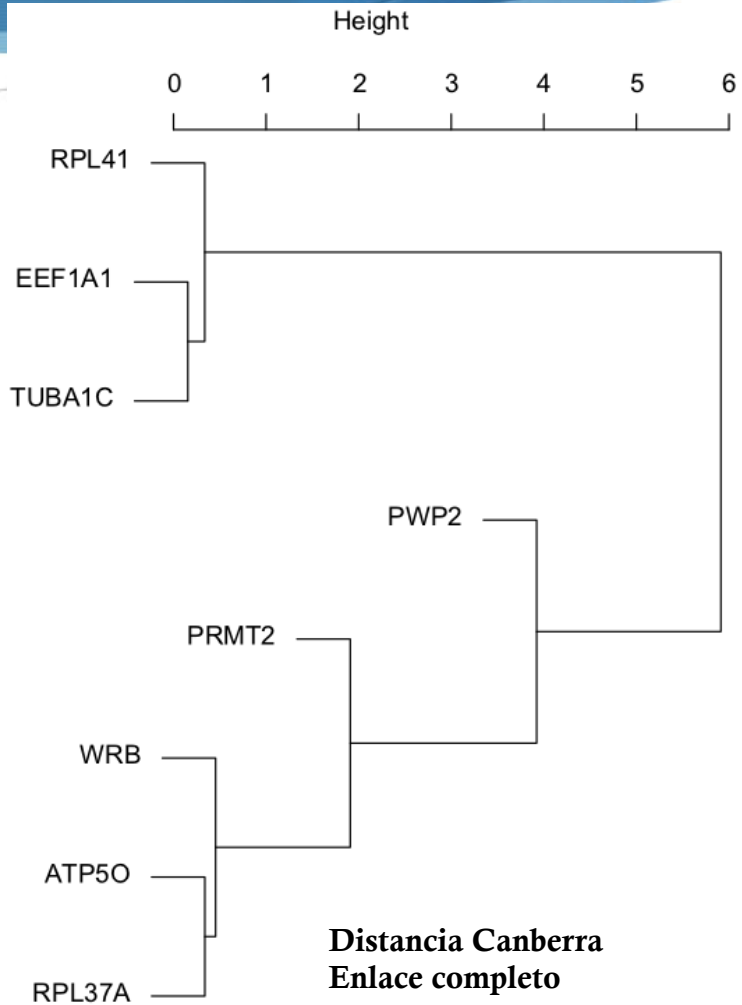




Clustering jerárquico

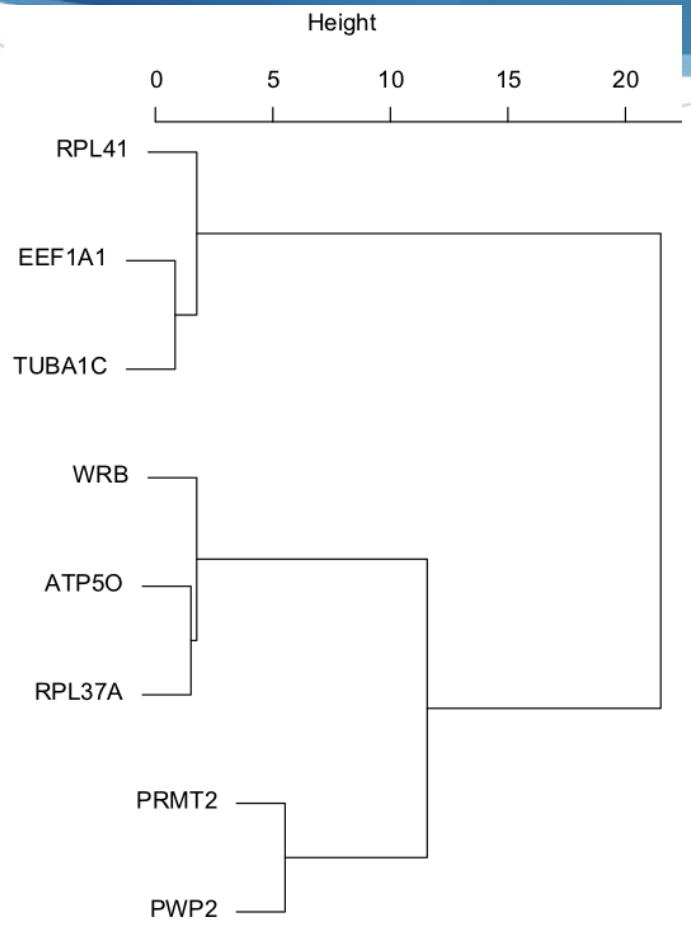


Distancia Euclídea
Enlace completo

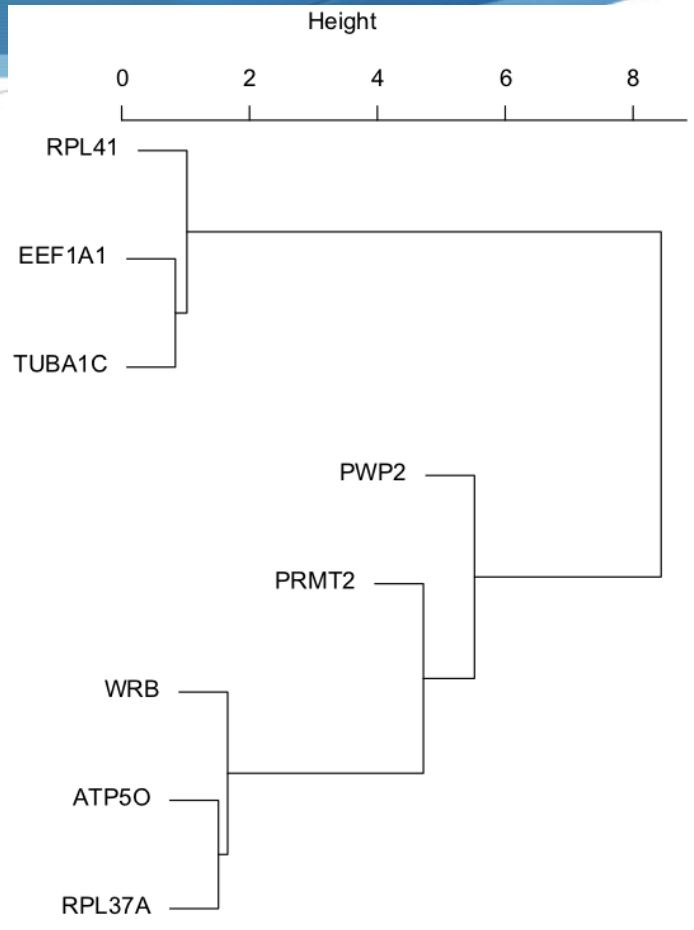


Distancia Canberra
Enlace completo

Clustering jerárquico



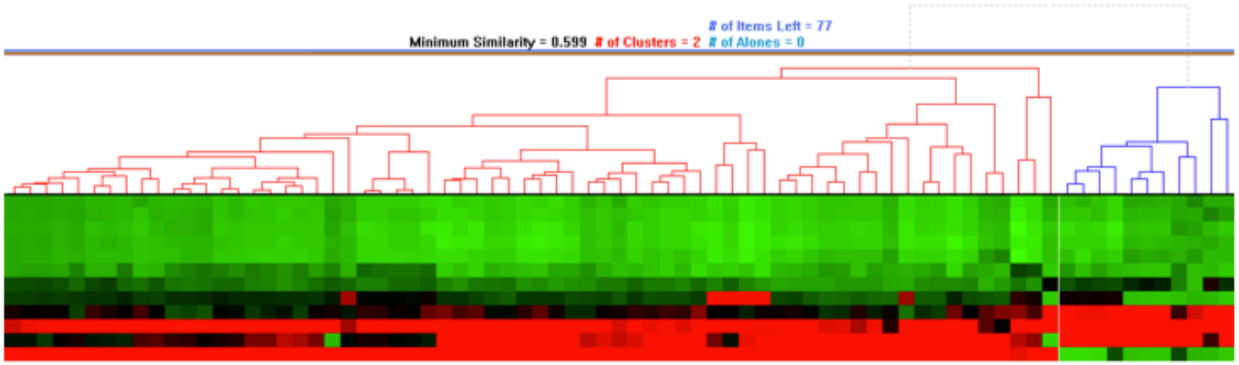
Distancia Euclídea
Enlace completo



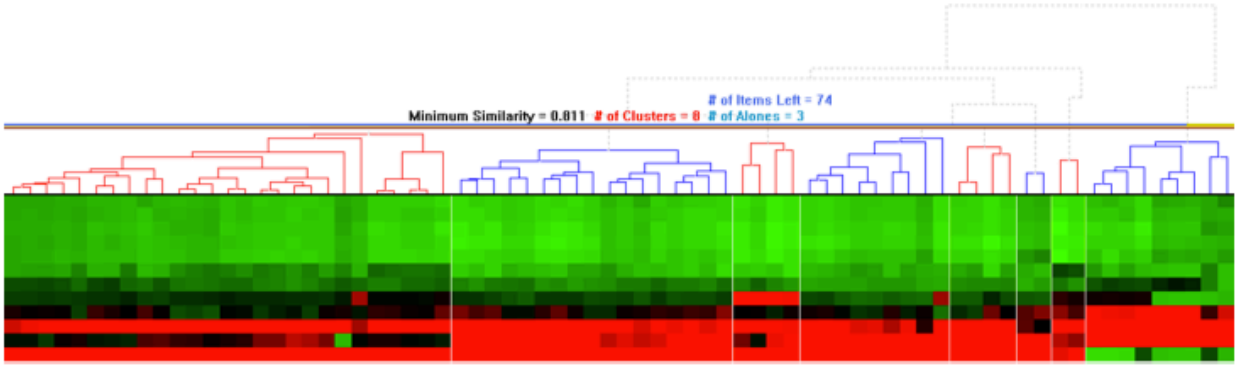
Distancia Euclídea
Enlace simple

Clustering jerárquico

- A partir del árbol, debemos determinar qué grupos hay en nuestros datos
  - Necesidad de establecer un umbral
  - No es trivial decidir cuál es el mejor "corte"
  - Uso de índices para determinar
    - La cohesión interna de los clusters
    - La separación entre clusters

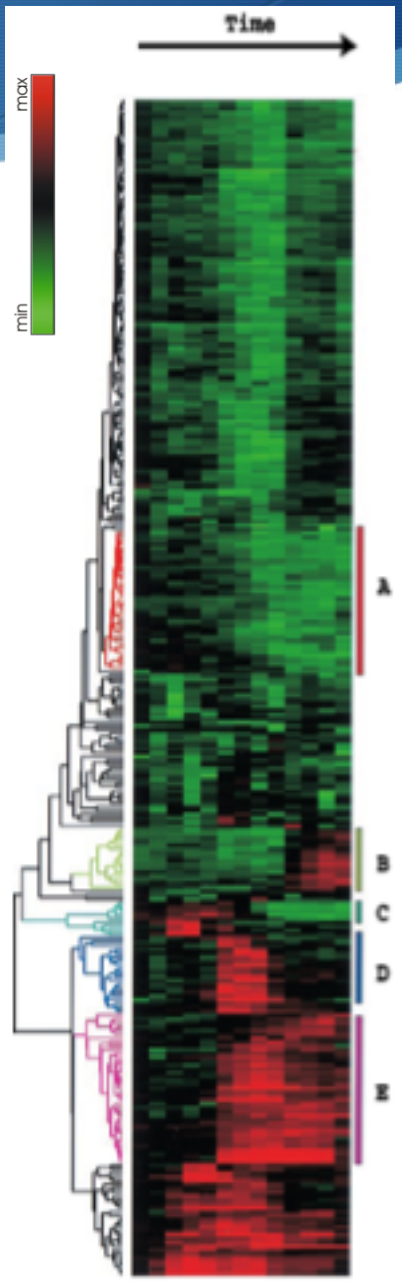


(a) Hierarchical clustering cut with two clusters



(b) Two clusters (red and blue) obtained by hierarchical clustering

- Eisen et al. 1998
  - Levadura bajo estrés en distintos momentos
  - Se identifican grupos de genes con patrones similares de activación (A-E)
    - A través de un clustering jerárquico



Clustering de k-medias

- Generación de clusters si sabemos exactamente el número de grupos (k) en que están divididos nuestros datos
  - Por ejemplo, si tenemos muestras de dos tipos de enfermedades y de control, tendremos k=3
- No se genera una jerarquía, simplemente los k clusters
- Es un método iterativo
  - Se asigna aleatoriamente cada elemento a un grupo
  - En cada iteración, se reasignan los grupos intentando minimizar la distancia media entre elementos de un grupo

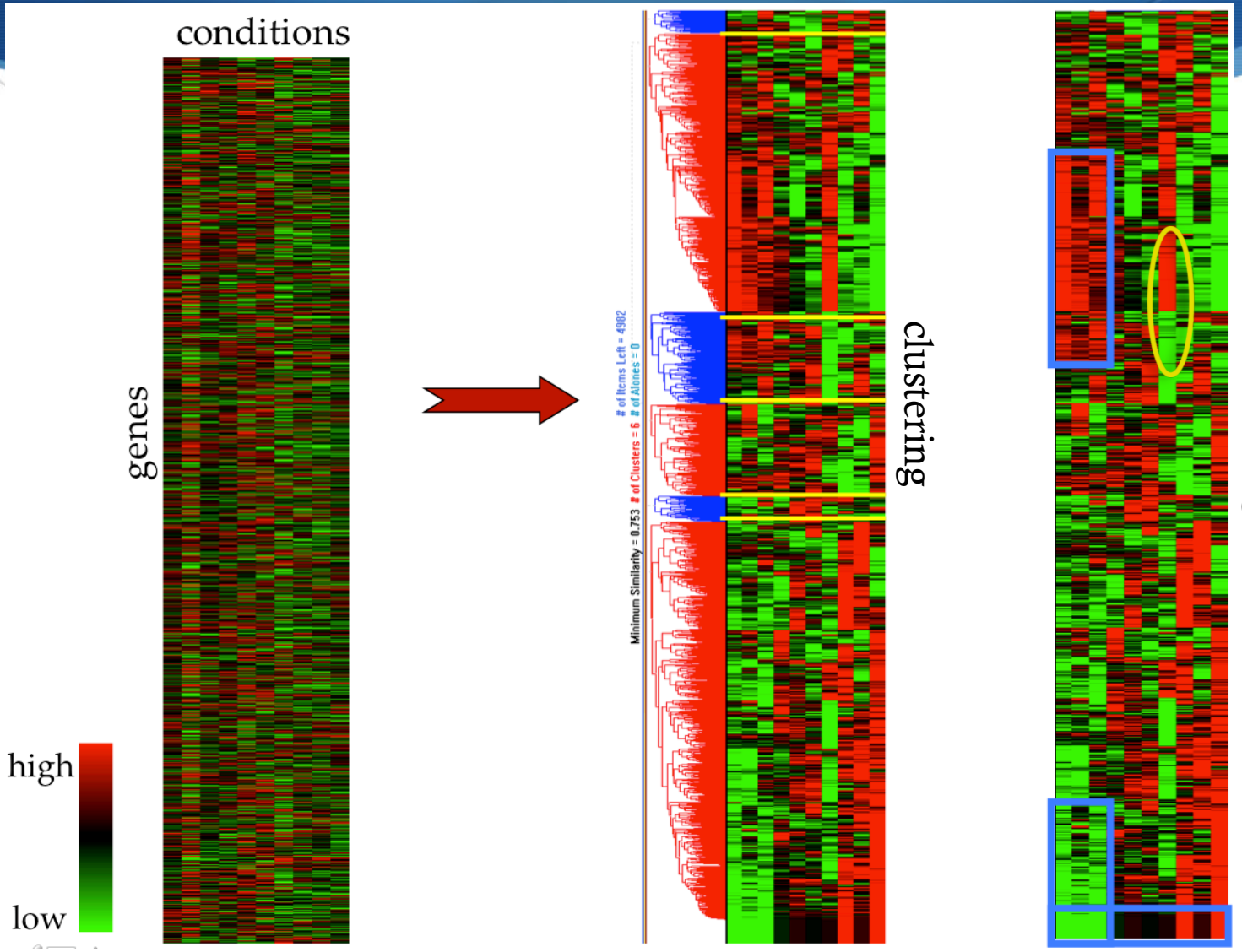
Clustering jerárquico

- Software
  - Hierarchical Clustering Explorer (HCE)
    - http://www.cs.umd.edu/hcil/hce/ (sólo para Windows)
  - TreeView
    - http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
- Métodos y paquetes R
  - Método "hclust" para clustering jerárquico aglomerativo
  - Biblioteca "cluster" y método "diana" para clustering jerárquico divisivo
  - Método "kmeans" para clustering de k-medias
  - Biblioteca "pvclust" para significatividad estadística

Biclustering

- En clustering, buscamos grupos de genes con similar expresión bajo *todas* las condiciones del experimento (análogo para condiciones)
- En biclustering, se buscan grupos *solapables* de genes con similar expresión bajo *algunas* condiciones
  - La activación de un gen puede deberse a varias funciones
  - Los genes "colaboran" bajo unas condiciones pero no bajo otras
- Técnica nueva y todavía no demasiado implantada
  - Mejora los resultados de clustering en precisión y puntería
  - Aún no existen benchmarks y es difícil determinar su calidad

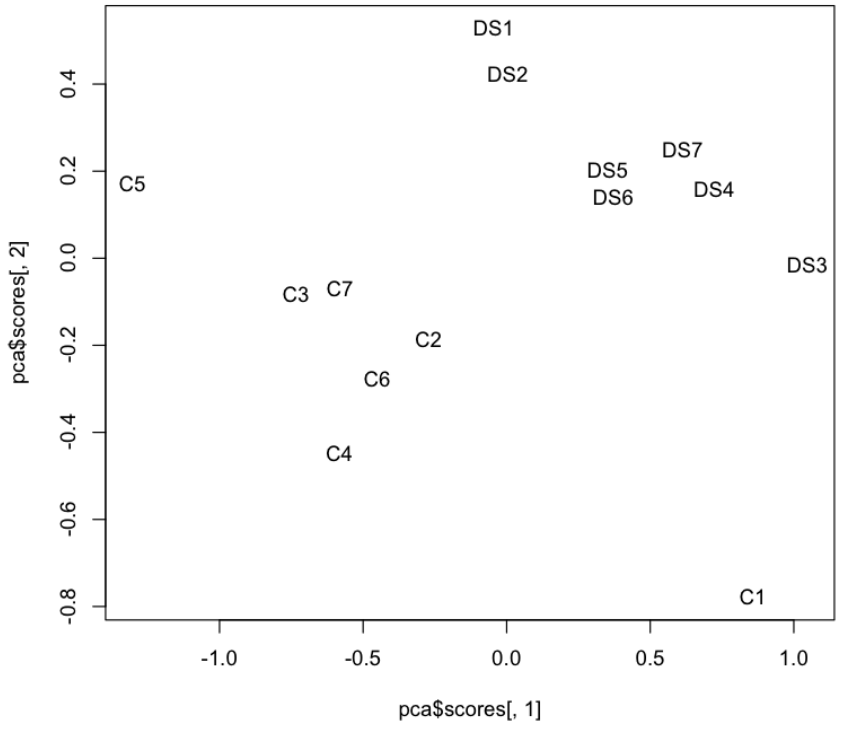
Biclustering



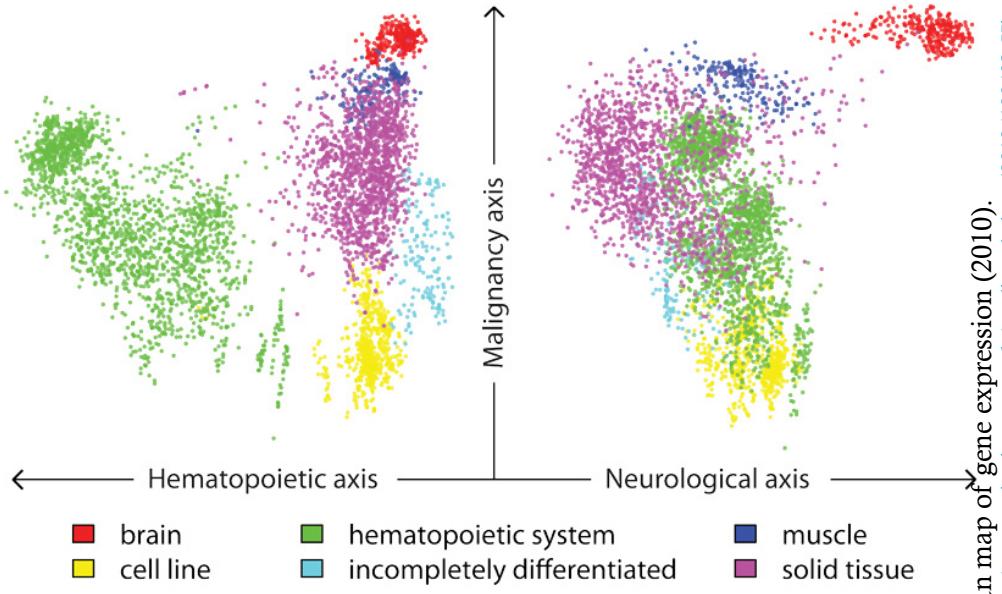
PCA

- Análisis de Componentes Principales
  - Reduce la dimensionalidad del problema de *n* a 2 ó 3 dimensiones
    - A cada gen/condición se le asigna un punto en una representación bi/tri-dimensional.
  - Para ello, se extraen los "componentes principales" de los puntos *n*-dimensionales
    - Las características más relevantes de su expresión génica
    - Generalmente las 2-3 primeras componentes caracterizan la mayoría del comportamiento

PCA



Las 2 primeras componentes para la expresión génica de 8 genes en pacientes de control (Cx) y pacientes con síndrome de Down (DSx)



Representación de tres componentes (asociadas a neurología, hematopoiesis y malignidad) para 5372 muestras de tejido humano. Se identifican 6 grupos

Brazma et al. A human map of gene expression (2010). http://www.embl.de/aboutus/communication_outreach/media_relations/2010/100408_Hinxton/press08apr10.pdf

Análisis de datos de microarray

Introducción

Preprocesamiento

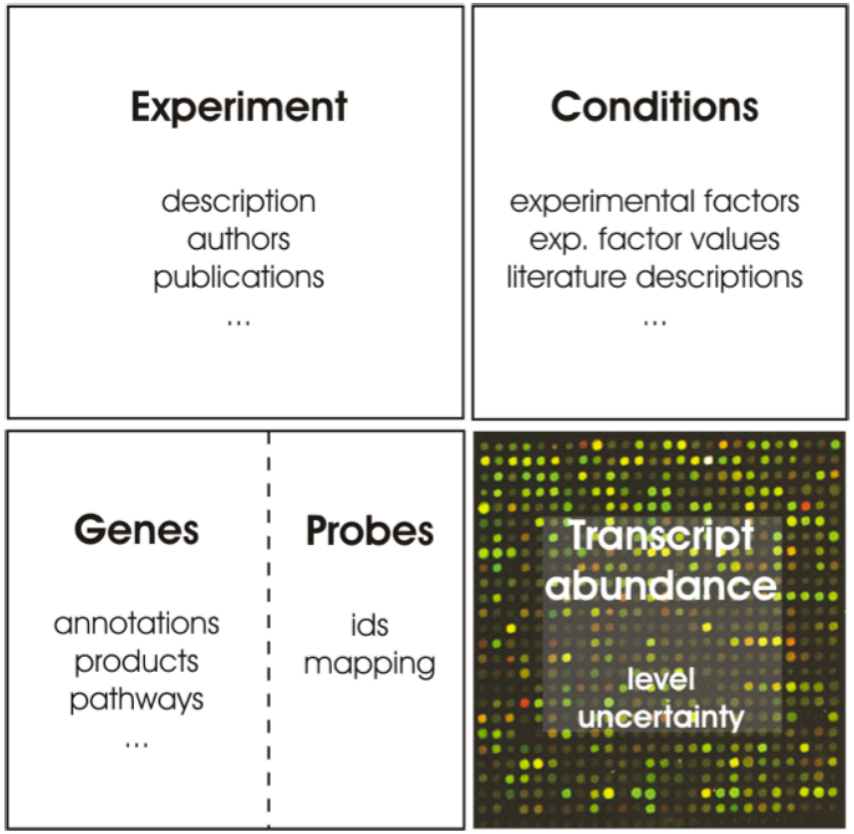
Análisis

Validación

- Anotación biológica
- Validación estructural
- Validación biológica

Anotación biológica

- Contamos con mucha más información de nuestro experimento que los niveles de transcripción
  - Condiciones
    - Factores experimentales (edad, sexo, estado, tiempo, localización…) y sus valores (31 años, mujer, cáncer, 3 meses, pecho…)
  - Genes
    - Funciones conocidas (anotaciones)
    - Relaciones conocidas (redes)
- Veremos cómo utilizar esta información para realizar y validar análisis de expresión



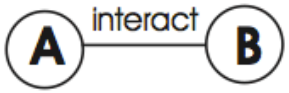
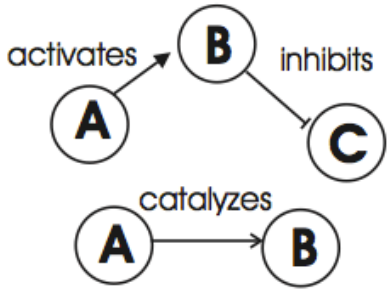
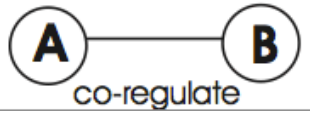
Anotación funcional

- Gene Ontology (GO)
  - Vocabulario controlado de información acerca de los genes
    - Función molecular (MF)
      - Qué hace su producto a nivel molecular: catálisis, enlazado del DNA
    - Proceso biológico (BP)
      - Qué hace su producto a nivel biológico: respuesta a estrés, biosíntesis de grasas
    - Componente celular (CC)
      - Dónde se localiza su producto: mitocondria, pared celular
- Podemos utilizar esta anotación (términos GO) para
  - Evaluar la coherencia biológica de los grupos encontrados (p.ej. "todos los genes del grupo se relacionan con la respuesta a estrés")
  - Guiar la búsqueda de grupos

Anotación relacional

- Muchos repositorios mantienen información sobre relaciones entre genes a varios niveles, en forma de redes
  - Topológico:
    - Interacciones a nivel físico o de co-locación.
    - Típicamente a nivel de proteínas (protein-protein interaction – PPI)
  - Cualitativo:
    - Existencia de una relación causal de algún tipo
      - P.ej. regulación (activación o inhibición) o colaboración en una red metabólica (KEGG pathways)
  - Cuantitativo:
    - Correlación entre valores, por ejemplo, de expresión
    - Correlación NO implica por sí sola una relación causal

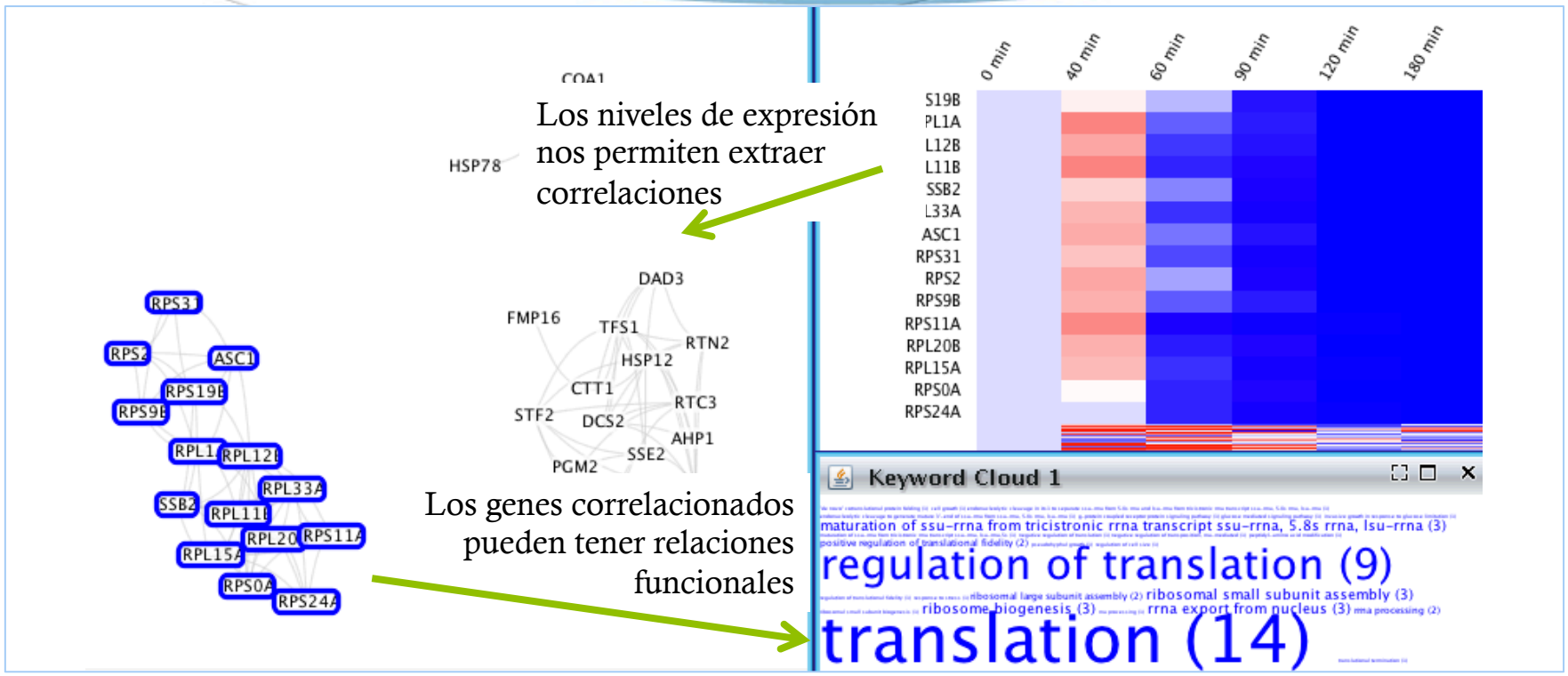
Anotación relacional

Category	Network examples	Relationship examples
Topological	Protein-Protein interactions Protein-DNA interactions Protein-Chemical interactions	
Qualitative	Regulatory networks Metabolic networks	
Quantitative	Co-regulation networks	

number of entities

R Santamaria et al. (2011) Systems biology of infectious diseases: a focus on fungal infections Immunobiology 216: 11. 1212-1227 11

Anotación biológica



Red de correlación (izquierda) determinada a partir de niveles de expresión (derecha, arriba). Un grupo de 15 genes bien relacionados en la red (en azul) muestran una clara relación funcional de traducción (14 de ellos anotados con “translation”, 9 con “regulation of translation”)

R. Santamaría, R. Therón, L. Quintales (2008) BicOverlapper : A tool for bicluster visualization Bioinformatics 24: 9. 1212-1213 05

Anotación biológica

- Ventaja
  - Dota a nuestros resultados de un valor biológico añadido
- Desventaja
  - Si se usa para guiar el análisis, puede desviar los resultados hacia el conocimiento biológico ya conocido
- Si un grupo no tiene sentido biológico según lo conocido…
  - ¿Es resultado de un mal análisis…
  - … o hemos encontrado información nueva?

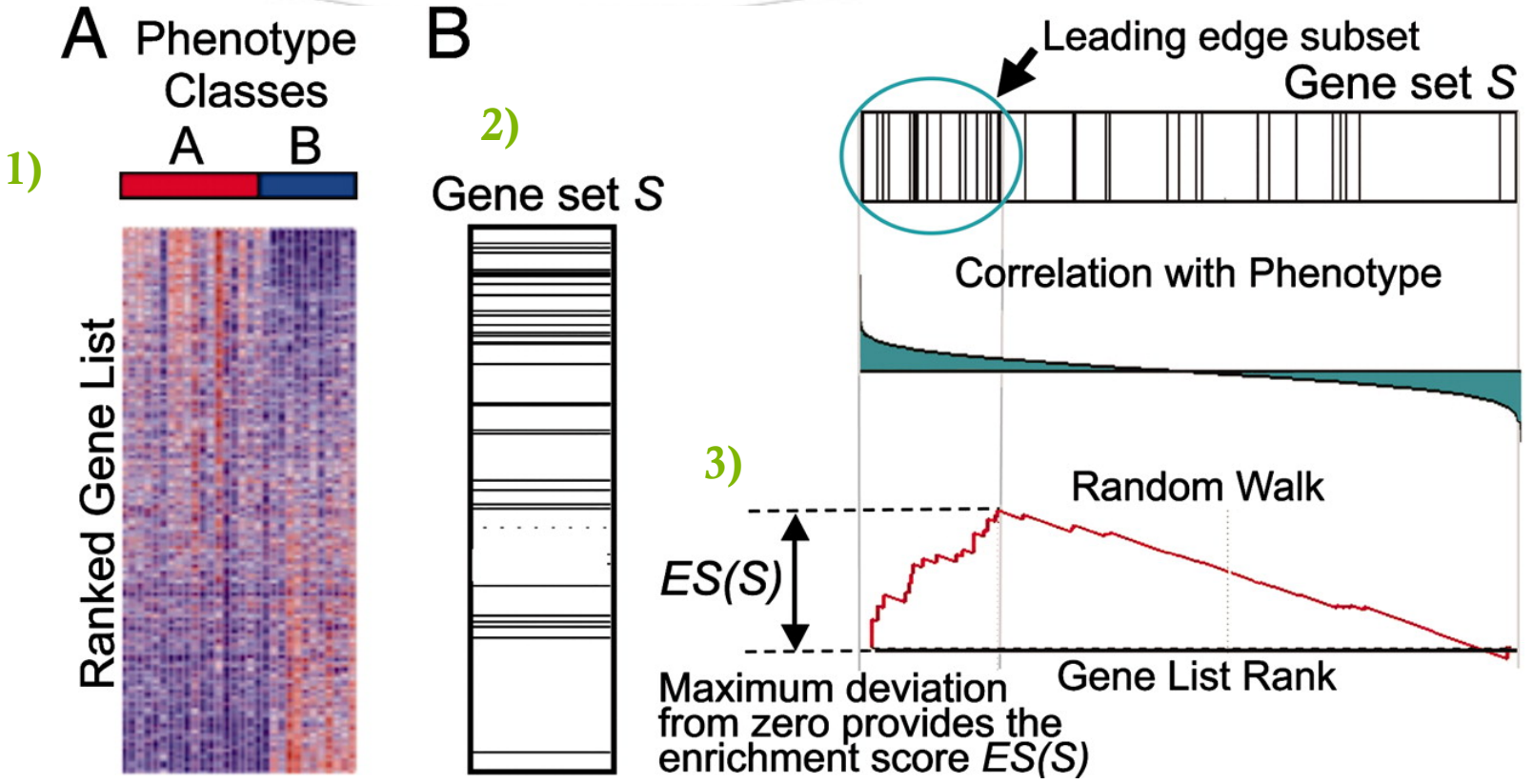
Validación basada en anotaciones

- Determinar el grado de significatividad estadística de alguna anotación biológica en un grupo encontrado
  - Si el grupo hace significativa dicha anotación, se dice que el grupo "enriquece" la anotación.
- Básicamente, se trata de ejecutar un test estadístico
  - Sea S una anotación biológica, con n_e genes anotados con S en nuestro experimento, y n_g genes anotados con S en nuestro grupo
  - Tratamos de determinar si n_g es un número significativamente alto de anotaciones respecto a n_e, o si seguramente se debe al azar

Análisis basado en anotaciones

- Gene Set Enrichment Analysis (GSEA)
  1. Se escogen dos conjuntos de muestras *A* y *B*, y se calcula su expresión diferencial para todos los genes
     - Se ordenan los genes según su nivel de expresión diferencial
  2. Se escoge una anotación funcional *S*
     - P.ej. genes anotados con el término GO "respuesta a estrés"
  3. Se calcula el "valor de enriquecimiento" *ES* de la anotación *S* entre los genes ordenados
     - Se suma un valor por cada gen anotado y se resta otro por cada uno no anotado
     - Se toma como *ES* el valor máximo de dicha función
  4. Se repiten los pasos 2 y 3 para muchas anotaciones distintas S_i, calculando sus ES_i, y se lleva a cabo un test estadístico para determinar la significatividad estadística de cada ES_i
  5. Se reportan las S_i con p-valor menor que el nivel de significatividad fijado

GSEA



Subramanian et al. *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. 2005

Análisis de datos de microarray

Introducción

Preprocesamiento

Análisis

Validación

Otras tecnologías

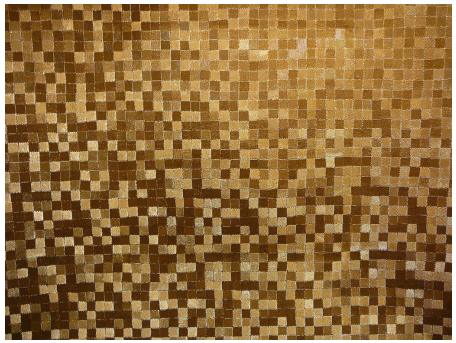
Tiling arrays y RNA-seq

ChIP-on-chip

Otras tecnologías

- Con un microarray medimos el nivel de expresión de unas determinadas secuencias de DNA
- Existen tecnologías derivadas de un modo u otro de este concepto
  - Incremento de la densidad de las sondas en el array → tiling
  - Técnicas de captura de secuencias de ADN → ChIP-on-chip
- Además, como veremos en el tema siguiente, las nuevas técnicas de secuenciación pueden aplicarse a la medición de niveles de transcripción
  - RNA-Seq, ChIP-Seq

Tiling arrays



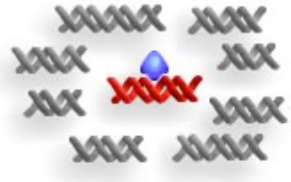
- **tile** 1 /taɪl/ sustantivo
  - (for floor) baldosa f, losa f;
  - (for wall) azulejo m
- Llevan la hibridación al máximo de las capacidades técnicas
  - Usamos como sondas un conjunto de secuencias cortas de nucleótidos (~25 bases) solapadas (~5 bases)
  - No sólo secuencian el transcriptoma, si no el genoma completo a una resolución muy alta
- Técnica intermedia que no se ha explotado demasiado debido a la aparición de la secuenciación de alto rendimiento

ChIP-on-chip

- Chromatin ImmunoPrecipitation (ChIP)
  - Técnica experimental para investigar la interacción entre el ADN y proteínas que se asocian con regiones del ADN
    - Son típicamente factores de transcripción que se unen a regiones promotoras o sitios de enlace (DNA binding sites)
  - La técnica se puede combinar con otras
    - ChIP-on-chip: combinación de ChIP y microarrays
    - ChIP-seq: combinación de ChIP y secuenciación

ChIP

1. Se mezclan las proteínas de interés (POI, en azul) con el ADN para que se acoplen a su sitio de enlace (rojo)



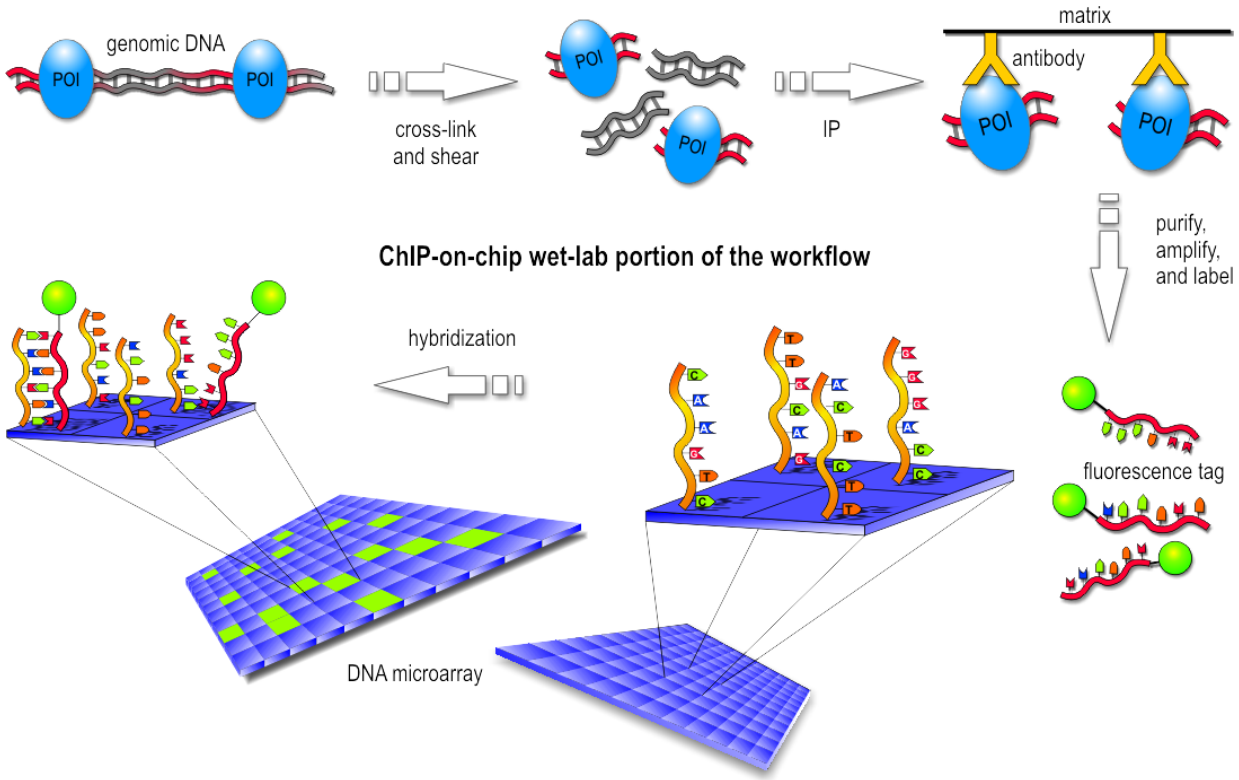
2. Se rompe la cadena de ADN, separando así la cromatina (combinación de proteína y ADN)

3. Se diseña un anticuerpo que ligue la proteína en cuestión y recoja sólo los fragmentos de cromatina (immunoprecipitación)

4. Estos fragmentos se amplifican mediante PCR para su medida, secuenciación, etc.

ChIP-on-chip

- Las secuencias extraídas mediante ChIP pueden lanzarse sobre un microarray para medir el nivel de enlace de la POI a las distintas sondas



Preguntas a debate

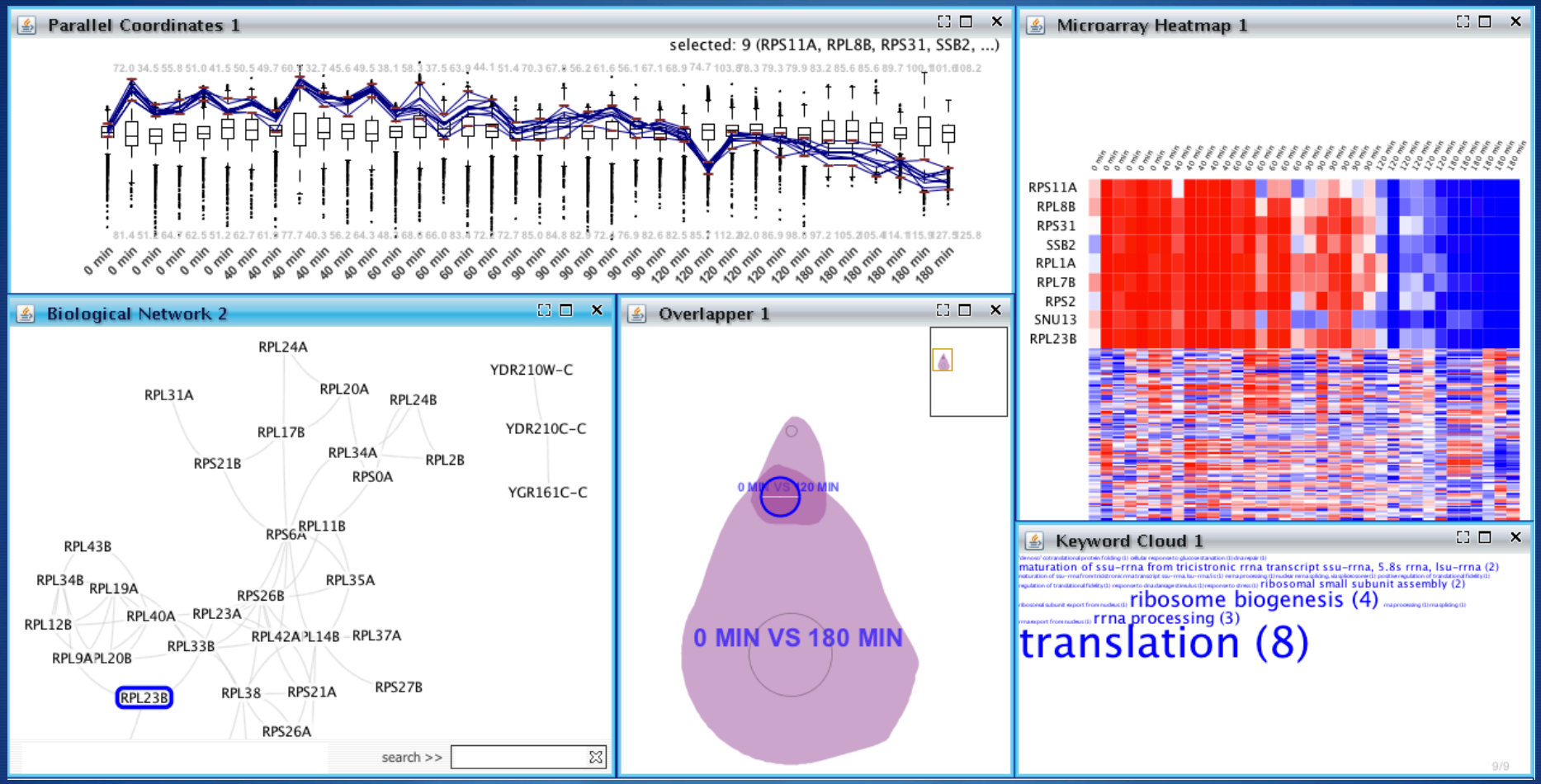
- ¿Se te ocurre una situación en que un ratio de expresión sea significativo estadísticamente, pero probablemente irrelevante biológicamente?
- ¿Considerarías el mismo umbral de p-valor en un análisis de expresión de una levadura y en un análisis de humano?
  - Si aplicas una corrección conservadora, y el resultado es que ningún gen cambia significativamente, ¿es esto un resultado posible biológicamente?
- ¿Qué opinas del uso de anotaciones funcionales para guiar y validar el análisis de expresión? ¿Es útil? ¿Es correcto?
  - ¿Qué otros modos hay de determinar la "bondad" de un grupo o de un conjunto de genes expresados diferencialmente? ¿Se pueden aplicar los conceptos de precisión y puntería?

Resumen

- La tecnología de microarrays mide el nivel de expresión (transcripción), resultando en una matriz de expresión de genes (filas) bajo distintas condiciones (columnas)
- El análisis de expresión génica tiene como objetivo determinar qué genes se encuentran diferencialmente expresados entre dos condiciones (estadística inferencial), y qué grupos de genes/condiciones tienen un patrón de expresión similar (estadística descriptiva)
- Existen una gran cantidad de métodos de análisis. Casi todos devuelven una salida. Lo difícil es confirmar que esa salida es válida desde un punto de vista biológico y estructural
- Hay que tener en cuenta que la expresión depende en gran medida del tipo de tejido y del momento celular, y que el nivel de transcripción no es directamente el nivel de expresión
- Muchos análisis caen en errores a la hora de identificar grupos, siendo los más corrientes 1) no hacer correcciones para contrastes de múltiples hipótesis, 2) no hacer una normalización adecuada, ni chequeos de la calidad de los arrays, 3) exceso de libertad paramétrica en el análisis de los datos y 4) exceso de limitaciones biológicas en el análisis de los datos
- Las correlaciones entre genes a nivel de expresión que queramos concluir como causales deben acompañarse de experimentos de laboratorio que aseguren que la relación a nivel transcriptómico se mantiene a niveles superiores (qRT-PCR, chIP-on-chip, etc.)

Lecturas adicionales

- Pevsner, 2009: Ch 9 *Gene Expression: Microarray Data Analysis*
- Eisen et al. *Cluster analysis and display of genome-wide expression patterns.* PNAS 95(25): 14863-14868. 1998
- Ashburner et al. *Gene Ontology: tool for the unification of biology*. Nat Genet 25(1): 25-29. 2000. PMCID: PMC3037419
- Subramanian et al. *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. PNAS 102 (43): 15545-15550. 2005



BicOverlapper es una herramienta desarrollada en la Universidad de Salamanca para la integración y análisis de datos de expresión génica, desde redes de correlación y ontología GO hasta expresión diferencial, GSEA y biclustering

http://carpex.usal.es/~visusal/bicoverlapper/