

ANÁLISIS DE DATOS DE MICROARRAY 2ª PRÁCTICA OBLIGATORIA

OBJETIVO

Realizar un estudio de expresión génica relacionado con un tema de interés. Mediante esta práctica se evalúan las siguientes capacidades

- Búsqueda crítica de experimentos de microarray en BBDD
- Análisis crítico de microarrays
 - Análisis de la calidad del experimento
 - Uso de MAplots, Boxplots, histogramas, etc.
 - Preprocesamiento (si es necesario)
 - Análisis de la expresión, a nivel
 - Inferencial: lista de genes expresados diferencialmente de manera significativa
 - Descriptiva: detección de grupos de genes que comparten patrones. Manejo de clustering y PCA
 - Validación de los resultados
 - Explicación razonada de los genes/grupos encontrados
 - Enriquecimiento de anotaciones funcionales
 - Validación biológica (publicaciones, relaciones conocidas, etc.)

1.- Selección del tema

Elegir un tema biológico relacionado con la expresión génica que os interese, por ejemplo:

- ¿Cómo afecta la edad a la expresión génica del cerebro?
- ¿Cómo es la expresión génica en un cáncer de mama?

Este tema puede estar o no relacionado con el gen “nuevo” estudiado en la práctica de alineamientos, directamente o indirectamente con su familia asociada o con algún gen que nos haya llamado la atención sobre dicho estudio (por ejemplo BRCA1)

2.- Selección del experimento

Una vez seleccionado el tema, realizad una búsqueda en GEO y/o ArrayExpress (AE) sobre dicho tema. Por ejemplo:

- “brain aging homo sapiens”
- “BRCA1 breast cancer”

Veréis que aparecen muchos experimentos en cualquiera de las dos BBDD. Queremos elegir uno de ellos para aplicarle algunos de los análisis vistos en clase. Algunos criterios para elegirlo:

- No elijáis un experimento con más de 20-30 condiciones. Si no puede ser muy pesado de descargar y de analizar. Idealmente, 10-20 condiciones serían perfectas
- Elegid un experimento realizado en una plataforma Affymetrix. La mayoría de los métodos de análisis están diseñados para funcionar sobre todo con este tipo de plataformas
- Si es posible, elegid un experimento que tenga disponibles los datos en crudo (ficheros CEL) para descargar. Así podremos realizar un preprocesamiento si queremos.
- Elegid un experimento que sea sencillo y claro, especialmente en cuanto a factores experimentales. Si es posible, elegid un experimento que tenga su publicación asociada accesible.

Ejemplo de experimento demasiado complicado en cuanto a factores experimentales relacionado con el gen BRCA1: E-GEOD-25835. Dos factores experimentales, uno de ellos (genotipo) con 12 valores posibles, de difícil interpretación.

Experimental factors	Factor name	Factor values
	DISEASE STATE	Tumor tissue, uncultured epithelium
	GENOTYPE	BRCA1+/, BRCA1+/, mut p53, K-ras, myrPI3K,cyclinD1, BRCA1mut/+, BRCA1mut/+, mut p53, K-ras, myrPI3K,cyclinD1

Ejemplo de experimento más fácilmente interpretable: E-TABM-366. Un único factor, con dos valores fáciles de interpretar.

Experimental factors	Factor name	Factor values
	Genotype	BRCA1 mutant, BRCA1 wild type

3.- Análisis del experimento mediante R/Bioconductor

Intentad realizar un análisis completo del experimento mediante R/Bioconductor

1. Adquisición de los datos del experimento
2. Análisis de la calidad (y si es necesario preprocesamiento)
3. Análisis diferencial
4. Análisis de clustering
5. Otros análisis (GSEA, PCA...)

Se facilitan los archivos E-TABM-366.R y GSE1397.R con las distintas fases del análisis para dos ejemplos de ArrayExpress y GEO

En general, consultar los paquetes (funciones):

1. Adquisición: GEOquery (getGEO) y ArrayExpress (ArrayExpress)
2. Análisis de calidad: arrayQualityMetrics (arrayQualityMetrics)
3. Preprocesamiento:(justRMA), affyPLM
4. Análisis diferencial: limma
5. Clustering: (hclust), (kmeans), pvclust, cluster (diana), biclust
6. PCA: (princomp)

Se trata de un ejercicio explorativo para que os familiaricéis con BioConductor y sus métodos, y para que observéis cómo se realiza un análisis básico de microarrays.

Evaluación

Se tendrán en cuenta los siguientes aspectos

- *Comprensión*: de los métodos utilizados, explicación de por qué se han elegido dichos métodos (y sus parámetros), de la tecnología y análisis de microarrays, y del experimento seleccionado.
- *Análisis crítico*: por qué se ha seleccionado un determinado experimento biológico o sus muestras. Valoración de la calidad del experimento, valoración de la calidad de los análisis y sus resultados.
- *Concreción*: capacidad de síntesis en cuanto a la explicación del problema y del análisis, no repetir lo dicho por otros.
- *Claridad*: claridad en la redacción, corrección ortográfica, organización en apartados, la claridad del informe se considera un indicativo del grado de comprensión.
- *Compleción*: variedad de tipos de análisis y configuraciones de parámetros probados, número de herramientas utilizadas, número de experimentos explorados o al menos contemplados.
- *Ética*: rigor en las citas (publicaciones o páginas web), evitar copiar textos tal cual (entenderlos, sintetizarlos y explicarlos si es necesario, si no citarlos). La copia, total o en gran medida, del trabajo, es motivo suficiente para suspender de manera automática la práctica. Sé un buen profesional, no copies.

Entrega

Entregad un informe redactando los principales hallazgos, dificultades, etc. que habéis encontrado en cada paso. incluid el código R que hayáis desarrollado.

Especialmente, documentad:

- Experimento elegido: identificador, plataforma, organismo, número de muestras, descripción del experimento, factores experimentales, razones para su elección
- Análisis de calidad: distribución de la expresión, normalidad de las muestras, informe o imágenes de la calidad (MAplots, boxplots, etc.). Razonamiento de la calidad del experimento y de su re-preprocesamiento, si es necesario.
- Análisis de expresión diferencial: selección de muestras, explicación del proceso (método, parámetros, etc.), lista de genes expresados diferencialmente, con significatividad estadística.
- Análisis de expresión descriptiva: elección de muestras, filtrado de genes (si se hace), dendrogramas obtenidos (para muestras y/o genes), explicación de los dendrogramas. Análisis alternativos (biclustering, GSEA, etc.)
- Validación biológica: comparación de los resultados con la literatura existente. Enriquecimiento biológico de las anotaciones funcionales por parte de los grupos obtenidos o por la lista de genes expresados diferencialmente.

La entrega se hará a través de moodle, el nombre del fichero debe tener OBLIGATORIAMENTE el siguiente formato:

NombreApellido1Apellido2.pdf
NombreApellido1Apellido2.R

Si se desean incluir más ficheros distintos a los dos anteriores (con nombres arbitrarios), hacedlo en un fichero comprimido NombreApellido1Apellido2.zip

Los ficheros con nombres distintos a éstos serán ignorados