ENTREGA DE TAREA: CASO PRÁCTICO

INSTRUCCIONES

- **Cada estudiante deberá entregar:**
  1. **Un documento en formato PDF con su respuesta a las preguntas del caso práctico. Las figuras deberán formar parte del documento (esto es, no se entregarán como archivos separados). Si para responder una pregunta se emplea código de R, dicho código debe incluirse en la respuesta.**
  2. **Un archivo de texto, de extensión `.R`, conteniendo todo el código de R empleado.**
- **Cada pregunta se puntúa sobre 1 punto. Si una pregunta tiene varios apartados, cada apartado se puntúa proporcionalmente (por ejemplo, si una pregunta tiene 4 apartados, cada uno puntúa sobre 0.25 puntos). Las preguntas no contestadas no puntúan.**
- **Entrega: por el Campus Virtual, en el bloque 'Prueba Práctica', 'Entrega de Tarea', desde el 2 de abril hasta el 30 de abril de 2022 inclusive.**

Descargar de la base de datos GEO del NCBI (`https://www.ncbi.nlm.nih.gov/gds` el conjunto de datos de microarrays de expresión génica con código de acceso GSE132575.

Asimismo, descargar de la Biblioteca de la UDC el artículo asociado a los datos:

Kovi RC, Bhusari S, Mav D, Shah RR, Ton TV, Hoenerhoff MJ, Sills RC, Pandiri AR (2019). Genome-wide promoter DNA methylation profiling of hepatocellular carcinomas arising either spontaneously or due to chronic exposure to Ginkgo biloba extract (GBE) in B6C3F1/N mice. *Archives of Toxicology*, 93 (8): 2219–2235. https://doi.org/10.1007/s00204-019-02505-7.

Debe descargarse también el 'Supplementary material' del artículo. La descarga de artículo y datos suplementarios puede hacerse accediendo a la Biblioteca de la UDC *desde fuera de la UDC* (`https://www.udc.es/es/biblioteca/recursos_informacion/acceso_fora_udc`), proceso que requiere *Autentificarse*. Una vez en la web de la revista, se recomienda usar el doi, 10.1007/s00204-019-02505-7, como término de búsqueda.

Responder a las preguntas siguientes.

1. En relación con el experimento al que está asociado el conjunto de datos:
   a) Resumir su objetivo.
   (INDICACIÓN: Responder en 10-15 líneas, redactadas en lenguaje científico, pero no excesivamente técnico.)
   b) ¿Qué tipo de microarray se utiliza?
2. Importar los datos con el paquete `affy` de Bioconductor y guardarlos en un objeto de nombre `gse132575`. Responder a las siguientes cuestiones:
   a) ¿De qué clase es el objeto `gse132575`? ¿Es una clase de tipo S3 o S4? ¿Cuántos *slots* tiene la clase?
   b) Dar el ejemplo de un *slot* del objeto `gse132575` que esté vacío (esto es, que no contenga ninguna información).
   c) Mostrar el contenido del *slot* `phenoData` empleando diversas funciones de acceso. En concreto, mostrar los nombres de los microarrays con `samplenames()` y, si son excesivamente largos (por ejemplo, si son los nombres de los archivos CEL.gz originales), cambiarlos por otros más cortos.
   d) ¿Podría estar vacío el *slot* `assayData`? ¿Por qué? ¿Qué función se utiliza para acceder al contenido del *slot* `assayData`? Empleando dicha función, mostrar *parte* del contenido del *slot*.

e) ¿Se pueden aplicar métodos de la clase `eSet` al objeto `gse132575`? ¿Por qué? En caso afirmativo, poner un ejemplo.

3. En relación con el objeto `gse132575` de la Pregunta 2 y empleando las funciones de acceso adecuadas, responder las siguientes cuestiones:

   a) ¿Cuántas medidas de intensidad hay en cada uno de los microarrays?

   b) ¿Cuántos pares de sondas tiene cada microarray? ¿Cuántas sondas?

   c) ¿Coincide el número de sondas con el número de medidas de intensidad? ¿Por qué?

   d) ¿Cuántos conjuntos de pares de sondas tiene cada microarray?

4. Escribir una función de R, de nombre `dePosicionATipoDeSonda`, que permita saber si una posición sobre el microarray corresponde a una sonda PM o MM. En concreto, la función tendrá como *argumentos* un objeto de clase AffyBatch y el número correspondiente a una posición sobre el microarray, y la *salida* será la cadena `"PM"`, la cadena `"MM"` o la cadena `"Ninguno"` según que en esa posición haya, respectivamente, una sonda PM, una sonda MM o no haya ningún tipo de sonda.

   Ejemplo de ejecución:

   ```
   > dePosicionATipoDeSonda(gse132575, 471259)
   [1] "MM"
   ```

5. Crear el objeto de R `dni` como se indica en el archivo auxiliar `generarConDNI.R`, que contiene el código de R necesario (este archivo se descarga del Campus Virtual, bloque 'Prueba Práctica'). El objeto `dni` contendrá los dígitos del DNI sin letra del estudiante.

   A continuación, ejecutar el siguiente código de R (que también se encuentra en el archivo auxiliar `generarConDNI.R`), para obtener el nombre de un conjunto de pares de sondas:

   ```
   > set.seed(dni)
   > id5 <- sample(length(featureNames(gse132575)), size = 1)
   > featureNames(gse132575)[id5]
   ```

   En relación con este conjunto de pares de sondas:

   a) ¿Cuántas sondas PM lo componen?

   b) Obtener las medidas de intensidad de las sondas PM del microarray de la muestra con identificador GSM3876426 (archivo GSM3876426_8353_473_CTL_Rep5_Mouse430_2_.CEL.gz).

   c) Obtener las medidas de intensidad de las sondas MM del microarray del apartado *b*).

   d) Con un gráfico adecuado, mostrar la relación numérica entre las medidas de intensidad de los apartados *b*) y *c*). ¿Son siempre mayores las medidas de intensidad de las sondas PM que las de las correspondientes sondas MM?

   e) ¿Todos los conjuntos de pares de sondas del microarray tienen el mismo número de pares de sondas que el conjunto de pares de sondas estudiado? Justificar la respuesta.

6. Con las etiquetas `"Espontáneo"` y `"Tratamiento"` se hará referencia a las muestras del experimento correspondientes a ratones que desarrollaron carcinoma hepatocelular espontáneamente o tras el tratamiento con extracto de *Ginkgo biloba*, respectivamente. Las demás muestras del experimento son controles. Ejecutando el siguiente código de R (contenido en el archivo auxiliar `generarConDNI.R`)), se guardará en el objeto `id6` una de las cadenas `"Espontáneo"` o `"Tratamiento"` elegida al azar:

   ```
   > set.seed(dni)
   > id6 <- sample(c("Espontáneo", "Tratamiento"), size = 1)
   > id6
   ```

   En el código anterior, el objeto `dni` es el mismo de la Pregunta 5.

   Responder los apartados siguientes:

a) Representar un diagrama MA que compare las medidas de intensidad de las sondas PM de todos los pares de microarrays correspondientes al valor de `id6`.

b) Examinando visualmente el diagrama del apartado *a*), ¿a qué par de microarrays le corresponde el mayor valor absoluto de la mediana de los valores de M? ¿Cuál es el valor de dicha mediana?

c) En vez de responder a las preguntas del apartado *b*) mediante un examen visual del diagrama del apartado *a*), escribir código de R que haga los cálculos automáticamente.

7. Ejecutando el siguiente código de R (código que se encuentra en el archivo auxiliar `generarConDNI.R`), se guardará en el objeto `id7` el identificador de un conjunto de sondas. En dicho código, el objeto `dni` es el mismo de las Preguntas 5 y 6, y el símbolo '+' indica que la línea esta cortada para que entre en el ancho de página.

```
> set.seed(dni)
> nombres.id <- c("1415789_a_at", "1415861_at", "1416240_at", "1417612_at",
+   "1421320_a_at", "1421899_a_at", "1423997_at", "1425128_at","1428553_at",
+   "1428689_at", "1429840_at", "1430649_at", "1434268_at", "1435718_at",
+   "1437060_at", "1439922_at", "1440278_at", "1440867_at", "1441604_at",
+   "1441876_x_at", "1442171_at", "1452055_at", "1455720_at", "1456500_at",
+   "1456753_at", "1457629_at", "1459034_at", "1459825_x_at")
> id7 <- sample(nombres.id, size = 1)
> id7
```

Responder los apartados siguientes:

a) Usando los recursos de anotación de Bioconductor y detallando el código de R utilizado, averiguar la correspondencia del identificador guardado en `id7` con los genes de las bases de datos Entrez Gene y GenBank del NCBI (`https://www.ncbi.nlm.nih.gov/`).

b) Usando los recursos del NCBI, responder a las siguientes cuestiones: ¿En qué cromosoma está localizado el gen? ¿En qué hebra del cromosoma? ¿Qué coordenadas tiene?. Finalmente, resumir la función de la proteína codificada por el gen.

NOTA IMPORTANTE: Las preguntas 8, 9 y 10 se refieren al subconjunto de microarrays formado por las muestras control y por las muestras correspondientes al valor de `id6` de la Pregunta 6.

8. Con el subconjunto de microarrays al que se alude en el recuadro al final de la pregunta 7, es decir, el subconjunto formado por las muestras control y las correspondientes al valor de `id6`, crear el objeto de nombre `gse132575.sub`. Partiendo de dicho objeto:

a) Hacer una corrección de fondo, considerando sólo las sondas PM.

b) Tras haber realizado la corrección de fondo, llevar a cabo una normalización de cuantiles, considerando sólo las sondas PM.

c) Representar gráficamente estimaciones de la densidad de las intensidades brutas, con corrección de fondo y normalizadas. Interpretar los gráficos.

9. Partiendo del objeto `gse132575.sub`, preprocesar los microarrays en un solo paso empleando la función `rma()` del paquete `affy`. ¿Cómo se efectuaría exactamente el mismo preprocesamiento por el método RMA con la función `expresso()`? ¿A qué se deben las diferencias en eficiencia computacional de las dos funciones?

10. Partiendo del objeto `gse132575.sub` ya preprocesado creado en la Pregunta 9, realizar un estudio de la expresión génica diferencial entre ratones de las muestras control y de las muestras correspondientes al valor de `id6` (ver Pregunta 6). Con tal fin, utilizar los procedimientos implementados en los paquetes `multtest` y `limma`, entre otros. Comparar los resultados obtenidos con los del artículo de Kovi *et al* (2019) asociado al experimento.