Actividad 1. Search GEO datasets (GEO web query)

Participamos en un estudio sobre **artritis reumatoide**, cuyos objetivos son conocer mejor los mecanismos moleculares de esta enfermedad y desarrollar herramientas que mejoren su diagnóstico y tratamiento.

Queremos conocer qué estudios se han publicado referente a esta enfermedad y qué datos están disponibles en GEO, porque nos gustaría realizar una revisión de estos resultados.

1. Determina el número de ítems totales. Detalla cuántos datasets, plataformas, muestras y series aparecen vinculados a "rheumatoid arthritis".
2. ¿Cuántos de ellos son específicamente para humano?
3. ¿Cuántos de estos estudios publicaron sus datos en GEO, durante el último año?
4. Guarda la relación de esta selección progresiva de estudios en un fichero de texto.
5. Reproduce la estrategia de búsqueda desde la opción "Advanced".

Actividad 2. Search GEO datasets (GEO web query)

Hemos leído este paper "Molecular signatures and new candidates to target the pathogenesis of rheumatoid arthritis" y nos ha gustado. En él se indica que los datos de microarrays de expresión utilizados en el experimento están en GEO, en la serie **GSE1919**.

1. ¿Qué plataforma comercial se ha utilizado para estos arrays? ¿ Agilent, Affymetrix o Illumina? Indica el tipo de array empleado.
2. ¿Cuántas muestras se incluyeron en este estudio?
3. ¿A qué organismo se hace referencia?
4. ¿Qué estatus tienen estos datos?
5. ¿Cuándo se subieron estas muestras a GEO y quién las incorporó en el respositorio? Queremos contactar a los responsables de estos datos. ¿Sería posible?
6. ¿Qué diseño experimental se utilizó?

Actividad 3. Analyze a study with GEO2R

Desde la misma web de GEO y siguiendo con el estudio anterior, analizaremos los datos para detectar los cambios de expresión significativos entre los enfermos de "osteoartritis" y "controles". Para ello selecciona la opción "Analyze with GEO2R":

1. Define dos grupos: controles ("normal donor") y enfermos ("osteoarthritis"). Asigna las muestras a cada sugbrupo.
2. ¿Qué 250 genes son los que muestran mayores diferencias entre ambos grupos? ¿Conoces algunos de los indicadores que aparecen en la cabecera de estos resultados?
3. Visualizar la información de los 250 genes con mayor diferencia de expresión está bien, pero nos gustaría disponer de un archivo con la información referente a la expresión diferencial para todos los genes incluidos en el array. ¿Te lo podrías

descargar en un fichero de texto?

4. Ahora nos centramos sólo en un gen de interés. ¿Qué información conocemos del gen **MMP3** en este análisis?
5. Muestra el perfil gráfico de expresión de este gen para todas las muestras. ¿Dónde está más expresado en controles o en enfermos? ¿Qué valores de expresión hay en cada una de las 10 muestras para este gen?
6. Revisa el script de R que se utilizó para este análisis.
7. ¿Qué información te puedes descargar de este estudio? ¿Qué diferencias hay entre las siguientes opciones? (Descarga la información en cada opción para comprobar las diferencias).
   - SOFT formatted family file(s)
   - MINiML formatted family file(s)
   - Series Matrix File(s)

Actividad 4. . FTP files

Explora las opciones que nos proporciona GEO para la descarga de datos por FTP: ftp://ftp.ncbi.nlm.nih.gov/geo/

Determina cómo descargar los datos correspondientes al estudio GSE1919 utilizando FTP.

Actividad 5. More GEO web queries

En nuestro departamento nos gustaría profundizar en el conocimiento de los genes de que cuyas variaciones transcriptómicas pueden ser causantes de la obsesidad. Previamente queremos conocer que estudios hay disponibles en GEO en está línea de trabajo:

1. Selecciona los estudios de obesidad. ¿Cuántos hay?
2. Inicialmente nos centraremos en mus musculus. ¿Qué número de estudios disponemos?
3. El primero de ellos nos gusta. Necesitamos saber: (empezamos seleccionando "Analyze DataSet")
   - ¿Qué perfil de expresión tiene el gen FTO? (Pista: "Find genes")
   - ¿Podrías obtener una representación gráfica global de todos los genes y muestras del estudio? (Pista: "heatmap").
   - ¿Y un gráfico con las distribuciones de expresión para cada muestra) (Pista: "Experimental design and value distribution")

Actividad 6. Detección de perfiles de expresión de un gen a lo largo de un grupo de estudios seleccionados.

1. Realizamos nuevamente una búsqueda en estudios de "rheumatoid arthritis" en humano (GeoDatasets).
2. A continuación queremos conocer los perfiles de expresión del gen HLA-DRB1, ya descrito como un biomarcador de esta enfermedad, en cada uno de los estudios seleccionados.
3. Además nos descargaremos los perfiles de expresión en un fichero txt.

Actividad 7. Descarga y extracción de los datos de estudios en GEO mediante el paquete GEOquery de Bioconductor

1. Ejecuta el script con los ejemplos que se incluyen.
2. Reproduce la descarga y extracción de datos para el estudio que utilizamos en actividades anteriores: **GSE1919.**

Actividad 8. Descarga y extracción de los datos de estudios en GEO mediante el paquete GEOquery de Bioconductor

GEOmetadb tiene como objetivo facilitar el acceso a los metadatos asociados con muestras, plataformas y conjuntos de datos. Esto se logra al analizar todos los metadatos de NCBI GEO en una base de datos SQLite que se puede almacenar y consultar localmente.

GEOmetadb es simplemente una envoltura alrededor de la base de datos SQLite junto con la documentación asociada. Finalmente, la base de datos SQLite se actualiza regularmente a medida que se agregan nuevos datos a GEO y se pueden descargar a voluntad para obtener los metadatos más actualizados.

1. Ejecuta el script “geo_geometadb.r” y comprueba la funcionalidad con los estudios de “**melanoma**” que sugieren.
2. Repite el proceso con los estudios de “**rheumatoid arthritis**”