Máster interuniversitario de Bioestadística y Bioinformática
Análisis de datos Ómicos (M0-157)
Primera prueba de evaluación continua.

Fecha publicación del enunciado: 6-04-2020

Fecha límite de entrega de la solución: 26-04-2020

Presentación

Esta PEC consta de ejercicios similares a los discutidos en los debates con los que podréis contrastar vuestra asimilación de los conceptos y métodos presentados en la primera parte del curso.

Objetivos

El objetivo de esta PEC es ilustrar el proceso de análisis de microarrays mediante la realización de un estudio, de principio a fin, tal como se llevará a cabo en una situación real.

Descripción de la PEC

La PEC se basará en los datos de un estudio **que debéis seleccionar de la base de datos GEO,** a partir del cual deberéis: (i) Plantear las cuestiones que deseáis responder (ii) Realizar los análisis necesarios y (iii) Elaborar un informe explicando problemas, métodos, resultados y discusión. Recordad que tan importante como el resultado es el razonamiento y el proceso que os leva a ello, es decir el consultor debe poder ver no tan sólo donde habéis llegado sino también como y porque habéis llegado hasta allí.

Recursos

Los recursos para la solución de la PEC son los que se han proporcionado en el aula para las primeras unidades, es decir los materiales del curso y casos de estudio.

Criterios de valoración

Tal como se indica en el plan docente la PEC vale el 40% de la nota.

Código de honor

Cuando presentáis ejercicios individuales os adherís al código de honor de la UOC, con el que os comprometéis a no compartir vuestro trabajo con otros compañeros o a solicitar de su parte que ellos lo hagan. Asimismo aceptáis que, de proceder así, es decir, en caso de copia probada, la calificación total de la PEC será de cero, independientemente del papel (copiado o copiador) o la cantidad (un ejercicio o todos) de copia detectada.

Formato

Para hacer la entrega se tiene que enviar un mensaje al buzón de entregas del aula. En este mensaje debéis adjuntar **únicamente** un fichero pdf (obtenido a partir de vuestro archivo Rmarkdown). El nombre del fichero debe ser la composición de vuestro apellido y vuestro nombre seguido de "_ADO_PEC1.doc" (por ejemplo: si vuestro nombre es "Jordi Pujol", el fichero debe llamarse "pujol_jordi_ADO_PEC1.pdf").

Además del archivo pdf podéis presentar vuestro estudio en formato reproducible mediante un repositorio de github cuya dirección deberá encontrarse en la primera página del informe.

No olvidéis de poner vuestro nombre y apellidos en el informe!!!

Enunciado

El objetivo de esta práctica es doble:

- Partiendo de un problema y unos datos públicos deberéis reanalizarlos siguiendo las pautas presentadas en los materiales y discutidas en los dos primeros segundo debate.
- Una vez obtenidos los resultados deberéis redactar un informe con la estructura tradicional de un informe científico técnico (ver "*Guías para el informe*").

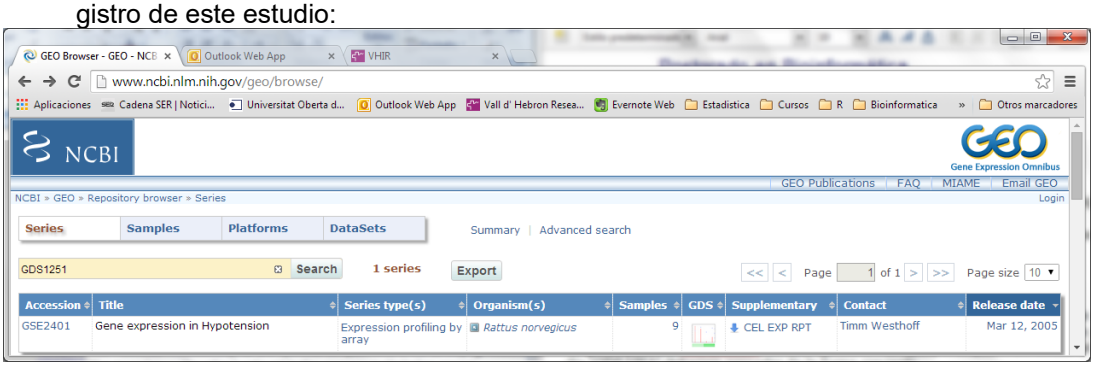
Selección de estudio/datos para el análisis

Lo primero que debéis hacer es escoger un estudio para re-analizar. *Si disponéis de uno podéis pasar a la sección siguiente*. Si no, seguid leyendo.

Podéis considerar reutilizar el estudio que seleccionasteis en el primer debate o bien escoger uno nuevo de la base de datos "Gene Expression Omnibus (GEO)". Por motivos obvios no podéis seleccionar ninguno de los estudios presentados en los debates o en los casos de estudio El estudio ideal debería tener pocos muestras (10-30) para que no tengáis problemas de memoria y preferiblemente 2-3 comparaciones aunque si tiene sólo una también servirá.

Supongamos **por ejemplo** que queráis utilizar el primer estudio "GDS1251" para hacer la PEC. Deberíais proceder de la forma siguiente

- Acceder a: http://www.ncbi.nlm.nih.gov/geo/browse/?view=series
- Realizar una búsqueda con el código proporcionado (GDS1251). Ésta os llevara al registro de este estudio:



- Para simplificar el análisis os sugiero que **sólo consideréis estudios realizados con microarrays de marca Afymetrix**, es decir que utilizan archivos .CEL
- Si los archivos .CEL están disponibles podéis descargarlos desde esta misma página. (Haciendo clic donde dice "CEL" o en la flecha que aparece a su lado).
- Alternativamente haciendo clic en el enlace correspondiente al "Accession" (GSE2401) accederéis a la página del estudio.
    - En la parte inferior encontrareis el enlace a los archivos suplementarios desde donde podéis descargar los archivos CEL, *siempre que estén disponibles. OJO, porque podrían no estarlo. Si es así cambiad de estudio.*
    - Para saber a que grupo pertenece cada muestra (En este caso "Case" o "Control") no tenéis más remedio que hacer clic en el enlace de la muestra y mirar a que grupo pertenece.
        - Esto lo encontrareis en el campo "Description" de la ficha de la muestra. Por ejemplo la muestra GSM45184 pertenece al grupo "Control"
    - Una vez sepáis a que grupo pertenece cada muestra podéis preparar el archivo "targets" y utilizarlo para crear la matriz del diseño y analizar los datos del ejemplo.
    - Si el estudio se ha publicado podéis acceder a él desde el apartado "citations". El abstract debería bastar para haceros una idea sobre de que trata el estudio y probablemente que comparaciones realizar.
- En vez de descargar los archivos .CEL y construir el archivo "targets" podéis descargar los datos con el paquete geoQuery lo que generará de forma automática el objeto ExpressionSet que necesitáis para el análisis.

Unos últimos comentarios sobre el formato de los datos:

- Para realizar las anotaciones debéis saber con que modelo de array trabajáis. Esta información la encontrareis en el apartado "plattform" en la ficha del estudio Gene Expression Omnibus.

"Pipeline" de análisis

Tal como habéis aprendido en esta unidad un análisis de microarrays suele seguir una serie de pasos ordenados. El proceso estándar consistirá en:

1. Identificar que grupos hay y a qué grupo pertenece cada muestra.
2. Control de calidad de los datos crudos
3. Normalización
4. [Control de calidad de los datos normalizados] (opcional)
5. Filtraje no específico [opcional]
6. Identificación de genes diferencialmente expresados
7. Anotación de los resultados
8. Comparación entre distintas comparaciones (si hay más de una comparación, ver que genes han sido seleccionados en más de una comparación)
9. Análisis de significación biológica ("Gene Enrichment Analysis")

Los capítulos 4 a 7 de los materiales muestran como llevar a cabo cada paso utilizando bioconductor.

Informe del análisis

Una vez realizado el análisis debéis redactar un informe exponiendo qué habéis hecho, como lo habéis hecho y qué resultados habéis obtenido.
Como cualquier informe científico-técnico vuestro informe tiene que tener las partes siguientes:

1. **Abstract**, con un resumen breve de no más de cinco líneas.
2. **Objetivos**: Que se pretende con este estudio
3. **Materiales y Métodos**
    1. Naturaleza de los datos, tipo de experimento, diseño experimental, tipo de microarrays utilizados,…
    2. Métodos que habéis utilizado en el análisis:
        1. Procedimiento general de análisis (pasos, "workflow" o "pipeline" que habéis seguido)
    3. Que habéis hecho en cada paso (NO ES PRECISO entrar en el detalle de los métodos, más bien hacer una descripción cualitativa indicando porque se ha llevado a cabo cada paso, y cual ha sido el "input" suministrado al procedimiento y el "output" obtenido.
4. **Resultados**
    1. Que se obtiene como resultado del análisis
5. **Discusión**
    1. Que limitaciones consideramos que pueden haber en el estudio (si consideramos que hay alguna…)
6. **Conclusión**: NO HACE FALTA. Vuestro "rol" aquí es técnico. Como bioinformáticos se os presupondrá la capacidad de manejar la información biológica mediante los programas adecuados, pero ello no implica que debáis tener los conocimientos específicos que puede requerir la interpretación biológica de los resultados.
7. **Apéndice**: Podéis poner el código de R que hayáis utilizado en un apéndice con comentarios.

Algunos comentarios sobre el formato de entrega

- La estructura indicada no es más que una propuesta. Podéis modificarla o adaptarla según vuestro propio criterio.
- Procurad facilitar la revisión
    - Tabla de contenidos
    - Secciones y subsecciones bien organizadas.
    - Gráficos bien centrados, preferiblemente con número y pie
    - Código o salida en formato courier y bien justificado
    - Páginas numeradas
    - Referencia bibliográficas completas.

Una cosa importante: El informe NO DEBE SER una colección de salidas de R como en algunos de los scripts y markdown de ejemplo que os he ido facilitando. Podéis poner algún fragmento de R si lo consideráis interesante pero tenéis que separar el informemdel código.
Si, como es de esperar, trabajáis con Rmarkdown os será muy sencillo ocultar las salidas de código que no deseáis mostrar utilizando las opciones del paquete knitr.

Observad especialmente que el objetivo de la práctica no es que generéis un "tocho" con un montón de información cogida de todas partes (que luego yo deberé leer) sino que realicéis un trabajo de síntesis que ilustre, de forma general, el proceso que va desde que el investigador se presenta delante vuestro diciendo "tengo unos datos que me gustaría que analicéis" hasta que le presentáis un informe con un "esto es lo que ha salido".

Reproducibilidad del estudio

Una habilidad que debéis adquirir como bioinformatic@s es aseguraros de que vuestro trabajo sea reproducible. Una forma de conseguirlo es crear un proyecto de Rstudio y ponerlo bajo control de versión en github, tal como hemos discutido en los debates.

Cread un repositorio en github y poned en él vuestro proyecto de forma que se pueda clonar en otro ordenador y reproducir vuestro trabajo. Debéis indicar la url de vuestro repositorio en el informe.