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

Fecha publicación del enunciado: 10-04-2022

Fecha límite de entrega de la solución: 23-04-2022

Presentación

Esta PEC plantea la realización de un análisis de microarrays a partir de datos de un estudio publicado con el 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 depositado en GEO con identificador "GSE177477" derivados de un estudio sobre los cambios en expresión génica asociados a la infección por SARS-COV2 en Pakistan.

Para su estudio, los investigadores trabajaron con muestras de sangre de humanos que clasificaron en tres grupos: HEALTHY, sin la enfermedad, ASYMPTOMATIC, enfermos sin síntomas, SYMPTOMATIC, enfermos con síntomas.

El objetivo de vuestro análisis será encontrar genes diferencialmente expresados entre los tres grupos y, a partir de éstos realizar un análisis de significación biológica que *de alguna forma* sirva para interpretar lñas diferencias observadas.

Para resolver la PEC deberéis:

i. Definir claramente los objetivos, es decir las comparaciones y preguntas que deseáis responder

ii. Realizar el proceso de obtención de datos y los preprocesados y los análisis adecuados.

iii. Elaborar un informe explicando problemas, métodos, resultados y discusión. Recordad que t**an 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.

Ahora bien, y como cosa importante, recordad que la PEC en si misma es un ejercicio de síntesis y aprendizaje en la que intenta valorar vuestra capacidad para resolver un problema muy parecido a los que se encuentra un/a bioinformática/a en su día a día. Esto quiere decir que para más de uno de los pasos que debéis realizar no hay una solución única. Plantead vuestra propia solución y explicad porqué creéis que es la adecuada. Entre otras cosas valoraremos:

- Capacidad de definir correctamente los objetivos a alcanzar
- Capacidad de organizar el análisis, obtención de los datos, preparación de los archivos etc.
- Dominio adecuado de las herramientas propias del tema (R, Rmarkdown, BioC)
- Capacidad de explicar qué y porqué se hace en cada paso.
- Capacidad de interpretar los resultados obtenidos.
- Capacidad de discutir las posibles limitaciones del estudio.
- Presentación del trabajo en un documento legible y bien organizado.

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 o .html (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 "Hadly Wickam", el fichero debe llamarse "wickam_hadly_ADO_PEC1.pdf").

¡¡¡Y no olvidéis de poner vuestro nombre y apellidos en el informe!!!

Indicaciones para la resolución

Datos para el análisis

Para esta PEC no os daremos los archivos .CEL ni el archivo “targets” como en los ejercicios, sino que la primera prueba que debéis hacer es localizarlos .CEL, descargarlos y crear el archivo .”targets”. Se puede hacer con facilidad desde la página del estudio en GEO:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE177477

Debéis descargaros el archivo “GSE1777477_RAW.tar” y tras descomprimirlo extraer de él los archivos .CEL

Selección de las muestras

Dado que todos los estudiantes vais a analizar el mismo estudio queremos potenciar que cada uno disponga de su propio "data set". Para ello debéis trabajar con un subconjunto de 15 muestras (de las 47 que hay en el estudio) para lo que *debéis escoger aleatoriamente cuatro muestras de cada grupo y formar con ellas vuestros "targets"*.

En el informé tenéis que indicar el código que habéis utilizado para la aleatorización.

Elaboración del informe

Lo natural hoy en día en bioinformática, como en otras disciplinas relacionadas (también en estadística, claro está) es realizar los informes utilizando un sistema que permita integrar código y explicaciones lo que facilitará la reproducibilidad del estudio y, sobre todo, simplificará vuestro trabajo a medida que avancéis. Dicho de otra forma, nuestro consejo es que elaboréis el informe directamente a partir del análisis utilizando Rmarkdown.
Ahora bien,

- Si para alguien esto resulta una dificultad tan importante que le impide centrarse en la PEC puede utilizar un enfoque más tradicional y separar análisis e informes.
- Recordad sin embargo que Rmarkdown genera fácilmente informes en HTML pero que si queréis generarlos en pdf tenéis que haber instalado el programa “TeX” en vuestro ordenador. Mi consejo es que no es necesario el pdf. Siempre podéis generar un documento HTML y, eventualmente imprimirlo a pdf.

“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. Exploración y 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 cómo 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 debe 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). 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 cuál ha sido el "input" suministrado al procedimiento y el "output" obtenido.
3. **Resultados**
    1. Que se obtiene como resultado del análisis
4. **Discusión**
    1. Que limitaciones consideramos que puede haber en el estudio (si consideramos que hay alguna…)
5. **Conclusión**: Vuestro "rol" aquí es técnico. Como bioinformátic@s 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.
6. **Apéndice**: Podéis poner el código de R que hayáis utilizado en un apéndice con comentarios si así lo creéis necesario

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
    - Referencias bibliográficas completas.

Una cosa importante: El informe NO DEBE SER una colección de salidas de R como en algunos de los scripts y Rmarkdown de ejemplo que os he ido facilitando. Podéis intercalar fragmentos de R en el texto si lo consideráis interesante, pero tenéis que separar el informe del código (poniendo el código en un apéndice sencillo de reproducir).

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 deberemos 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”.