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

Fecha publicación del enunciado: 27-12-2021

Fecha límite de entrega de la solución: 09-01-2022

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 segunda parte del curso.

Objetivos

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

Descripción de la PEC

A partir de los datos proporcionados y de la información sobre el problema 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 unidades de la segunda parte del curso, es decir los vídeos, artículos y, sobre todo, los 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 o html obtenido a partir de vuestro análisis. El nombre del fichero debe ser la composición de vuestro apellido y vuestro nombre seguido de "_ADO_PEC2" (por ejemplo: si vuestro nombre es "Jordi Pujol", el fichero debe llamarse "pujol_jordi_ADO_PEC2.html" o "pujol_jordi_ADO_PEC2.pdf").

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

Análisis de datos de RNA-Seq

El objetivo de esta práctica es doble:

- Partiendo de un problema y unos datos seleccionados como se indica a continuación deberéis
    - Decidir un pipeline de análisis apropiado, con los paquetes de (R/Bioconductor) que consideréis más adecuados.
    - Realizar el análisis siguiendo las pautas presentadas en los materiales.
- 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 los datos

El archivo RawCounts.csv contiene la información de las muestras de un estudio obtenido del repositorio Gene Expression Omnibus pertenecientes a un estudio que, entre otros resultados investigó los cambios en expresión génica asociados a la respuesta inmune a la infección con SARS-COV-2.

Como muestra una rápida inspección del archivo éste contiene 34 muestras, 17 pertenecientes a pacientes afectados de COVID y 17 de controles sanos.

En este ejercicio no os pedimos que busquéis un estudio para analizar sino que ya os proporcionamos los datos preprocesados en una tabla de contajes y os pedimos que seleccionéis **20 muestras** aleatoriamente, 10 de cada grupo. Para ello deberéis preparar un fragmento de código que selecciones10 muestras del grupo "COVID", y 10 del grupo "SANO" y, con la información obtenida generar vuestra propia matriz de contajes con sólo 20 columnas.

NOTA: Podéis hacer esta extracción aleatoria en la forma que prefiráis -y que garantice la aleatoriedad. Únicamente aseguraros de explicarlo con detalle en el informe.

Una vez hayáis preparado los datos podéis proceder a realizar un análisis de expresión diferencial.

"Pipeline" de análisis

El pipeline de análisis de datos de RNA-seq es parecido al de microarrays salvo algunos pasos, que no llevaremos a cabo completamente, aunque si que lo haréis de forma simplificada.

Los pasos básicos son los habituales:

1. Definición de los datos tal como se ha descrito en el párrafo anterior
2. Preprocesado de los datos: filtraje y normalización, estudio de posibles efectos batch.
    - Aquí no filtraremos por variabilidad sino que eliminaremos los tránscritos con demasiados ceros. Es decir: *eliminad todos aquellos transcritos que no tengan como mínimo tres contajes distintos de cero en* ***cada grupo****.*
3. Identificación de genes diferencialmente expresados
    - Haremos lo que se sugirió en el debate: Probad como mínimo dos métodos de entre *limma-voom*, *DESEQ2* y *edgeR* y ***comparad los resultados*** **que obtenéis**.
4. Anotación de los resultados.
5. Busca de patrones de expresión y visualización de los mismos.
6. Análisis de significación biológica ("Gene Enrichment Analysis")
    - Volviendo a la idea del debate: Investigad el uso del paquete *GOSeq* y comparadlo con algún otro que vayáis a utilizar, ya sea *GOstats o clusterProfiler*

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,
    2. Métodos y **herramientas** que habéis utilizado en el análisis:
        1. Procedimiento general de análisis (pasos, "workflow" o "pipeline" que habéis seguido)
        2. Software que habéis utilizado
    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**: Poned el código de R que hayáis utilizado en un apéndice con comentarios (no se trata de que expliquéis el código. Simplemente dejadlo como una sección más que permitiría, si así lo deseáramos, copiar y re-ejecutar un fragmento).

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.

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".