Inferencia Estadística (M0.155)

Segunda Prueba de Evaluación Contínua

Fecha de publicación del enunciado: 9/12/2024

Fecha límite para presentar la PEC: 22/12/2024[1]

Presentación y objetivos

En esta PEC, una vez familiarizados con los datos de expresión, y los métodos y herramientas para la selección de genes y el análisis de la significación biológica, procedemos a la realización de un análisis de datos, que nos permitirán mejorar nuestra comprensión de un problema biológico mediante métodos y herramientas estadísticas y bioinformáticas.

El análisis es parecido, aunque no necesariamente coincidente, con algunos de los casos resueltos que os hemos proporcionado, por lo que podéis inspiraros en ellos pero, sobretodo, debéis entender cada paso que hagáis.

Descripción de la PEC

La PEC se basará en los datos de un estudio que, utilizando un modelo murino (de ratón) investigo la utilidad de los antibióticos LINEZOLID y VANCOMICINA para inmunomodulación durante infecciones por Staphylococcus aureus resistente a meticilina (MRSA).

sample	infection	time	agent
GSM944831	uninfected	hour 0	untreated
GSM944838	uninfected	hour 0	untreated
GSM944845	uninfected	hour 0	untreated
GSM944852	uninfected	hour 0	untreated
GSM944859	uninfected	hour 0	untreated
GSM944833	uninfected	hour 0	linezolid
GSM944840	uninfected	hour 0	linezolid

[1]La fecha de entrega es la que se indica en el enunciado de la PEC. En caso de no coincidir con la indicada en el aula, ésta (la del enunciado) será la que predomine.

sample	infection	time	agent
GSM944847	uninfected	hour 0	linezolid
GSM944854	uninfected	hour 0	linezolid
GSM944861	uninfected	hour 0	linezolid
GSM944834	uninfected	hour 0	vancomycin
GSM944841	uninfected	hour 0	vancomycin
GSM944848	uninfected	hour 0	vancomycin
GSM944855	uninfected	hour 0	vancomycin
GSM944862	uninfected	hour 0	vancomycin
GSM944832	S. aureus USA300	hour 2	untreated
GSM944839	S. aureus USA300	hour 2	untreated
GSM944846	S. aureus USA300	hour 2	untreated
GSM944853	S. aureus USA300	hour 2	untreated
GSM944860	S. aureus USA300	hour 2	untreated
GSM944835	S. aureus USA300	hour 24	untreated
GSM944842	S. aureus USA300	hour 24	untreated
GSM944849	S. aureus USA300	hour 24	untreated
GSM944856	S. aureus USA300	hour 24	untreated
GSM944863	S. aureus USA300	hour 24	untreated
GSM944836	S. aureus USA300	hour 24	linezolid
GSM944843	S. aureus USA300	hour 24	linezolid
GSM944850	S. aureus USA300	hour 24	linezolid
GSM944857	S. aureus USA300	hour 24	linezolid
GSM944864	S. aureus USA300	hour 24	linezolid
GSM944837	S. aureus USA300	hour 24	vancomycin
GSM944844	S. aureus USA300	hour 24	vancomycin
GSM944851	S. aureus USA300	hour 24	vancomycin
GSM944858	S. aureus USA300	hour 24	vancomycin
GSM944865	S. aureus USA300	hour 24	vancomycin

Como puede verse en la tabla 1, el dataset consta de 35 muestras, 15 tomadas antes de la infección y 20 después, 5, que eliminaremos, a las 2 horas de la misma y 15 a las 24 horas.

Preguntas

Nuestro objetivo sera intentar caracterizar, a través del cambio en la expresión génica, el efecto de la infección y del tratamiento con antibióticos así como comparar los efectos de éstos.

Es decir deberéis hacer las comparaciones siguientes:

- Infectados vs no infectados sin tratamiento
- Infectados vs no infectados tratados con LINEZOLID
- Infectados vs no infectados tratados con VANCOMICINA

Esto generará tres listas de genes que deberéis, por un lado caracterizar mediante análisis de significación biológica y, por otro lado, comparar entre ellas.

Preparación de los datos

Podéis descargaros los datos crudos del sitio de GEO https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38531 donde también encontraréis información sobre el estudio original.

Con el fin de simplificar el análisis y dificultar un mínimo el intercambio no permitido de información deberéis eliminar algunas muestras: - Por un lado prescindiremos de las cinco muestras tomadas a las dos horas - Por otro lado sortearemos las muestras restantes de forma que hayáis de conservar tan sólo cuatro muestras de cada grupo.

Esto lo podéis hacer con la función `selectSamples` que encontraréis en el archivo `selectSamples.R` y que os permitirá extraer 24 muestras distintas a cada uno con tan solo llamarla usando como semilla (argumento "seed") vuestro DNI (sin la letra) o, preferiblemente vuestro identificador de la UOC.

Tras aplicar la función `selectSamples` obtendréis un nuevo objeto `targets` que os permitirá crear un nuevo `ExpressionSet` leyendo únicamente aquellos archivos .CEL que hayáis seleccionado.

Observad que la tabla no contiene los nombres exactos de los archivos.CEL por lo que *deberéis encargaros vosotros de adaptar lo qe creáis necesario para poder leerlos.*

A partir de este `ExpressionSet` personalizado, con 24 muestras muestras deberéis realizar vuestro análisis que consistirá en lo siguiente:

Análisis exploratorio y control de calidad

Empezad con la exploración habitual que os permita decidir si los datos necesitan alguna transformación, si presentan algún problema y si los grupos que deseáis comparar se separan mínimamente

Podéis complementar vuestra exploración con un control de calidad con el paquete `arrayqualitymetrics`,

Tened en cuenta que vais a empezar trabajando con datos crudos que convertiréis en una matriz de expresión después de normalizar dichos datos usando, por ejemplo, el algoritmo RMA.

Filtrado de los datos

Aunque, como sabéis, el filtraje es algo discutido, podéis por ejemplo eliminar las sondas menos variables y quedaros con el 10% de sondas que presenten mayor variabilidad.

Construcción de las matrices de diseño y de contrastes

Para realizar el análisis debéis crear las matrices de diseño y de contrastes y utilizarlas para llevar a cabo las comparaciones propuestas. Recordad que debéis hacer tres comparaciones lo que equivale a tres contrastes.

Obtención de las listas de genes diferencialmente expresados *para cada comparación*

Utilizad `limma` para obtener una lista de genes diferencialmente expresados, siguiendo los ejemplos presentados en las notas y los casos resueltos.

Las comparaciones entre las listas de genes la podéis hacer gráficamente o usando la función `decideTests` de `limma`.

Anotacion de los genes

El análisis con `limma` nos arroja listad de identificadores basados en los identificadores originales. Con estas listas debéis anotarlos, es decir asociarles algún identificador como "Symbol", "EntrezID" o "EnsemblID"

Análisis de la significación biológica

Una vez anotados los genes podemos intentar interpretar los resultados intentando determinar si las listas se encuentran enriquecidas en algunas categorías biológicas

Para ello podéis llevar a cabo un análisis de sobre-representación o un Gene Set Enrichment Analysis. Podéis utilizar para ello el paquete `clusterProfiler` que os permite hacer ambos análisis de forma muy muy similar.

También permite, de forma muy sencilla, visualizar los resultados del análisis de significación biológica, lo que ayuda a comprender *y comparar* los resultados.

Informe del análisis

Finalmente, y como de costumbre, debéis elaborar un informe de vuestro trabajo usando Rmarkdown. Aquí debéis tener en cuenta el contenido y la construcción.

- En cuanto al contenido el informe debe tener la estructura habitual de cualquier trabajo: (i) Tabla de contenidos, (ii) Introducción y Objetivos, (iii) Métodos, (iv) Resultado (v) Discusión (vi) Referencias y (vii) Apéndices. En el apéndice podéis poner el código R que habréis utilizado para vuestro trabajo y así será un único documento.

- En cuanto a la construcción debéis preparar documento en Rmarkdown que genere el informe en HTML y que debéis imprimir a pdf para entregarlo. Si tenéis instalado alguna versión de LateX es probable que podáis generar el archivo .pdf directamente. El cómo generéis el pdf queda a vuestra elección.

Tenéis que entregar **un único archivo** en formato pdf con la estructura anterior. El archivo debe ser legible por lo que no debe contener listas inmensas ni largos fragmentos de código, que tenéis que colocar en apéndices del documento.

Recursos

Los recursos para la resolución de la PEC son los que se han proporcionado en el aula hasta el momento, es decir, los materiales del curso y casos de estudio.

Criterios de valoración

Tal como se indica en el plan docente, esta PEC vale el 40% de la nota dado que a este segundo reto se dedica este porcentaje de las horas del curso.

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, BioConductor)
- 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.

Adicionalmente os recuerdo que el uso de programas de IA, como ChatGPT, Perplexity o similares, no está permitido. Obviamente esto es muy difícil de controlar, pero, además de vuestro compromiso, mediante el código de honor en que aceptáis seguir la normativa de las universidades, no debéis olvidar que estos programas tienden a alucinar, por lo que son más sencillos de identificar de lo que uno cree.