Anàlisi de dades Òmiques (M0-157)

Segona prova d' avaluació contínua

1. Presentació i objectius 2
   1.1 Descripció de la PEC 2
2. Preguntes 3
   2.1 Preparació de les dades 3
   2.2 Anàlisi exploratòria i control de qualitat 4
   2.3 Filtratge de les dades 4
   2.4 Construcció de les matrius de disseny i de contrastos 4
   2.5 Obtenció de les llistes de gens diferencialment expressats *per a cada comparació* 5
   2.6 Anotació dels gens 5
   2.7 Anàlisi de la significació biològica 5
   2.8 Informe de l' anàlisi 5
3. Recursos 6
4. Criteris de valoració 6
   4.1 Codi d' honor 6

Data de publicació de l'enunciat: 9/12/2024

Data límit per presentar la PEC: 22/12/2024[1]

[1] La data de lliurament és la que s' indica en l' enunciat de la PEC. En cas de no coincidir amb la indicada a l'aula, aquesta (la de l'enunciat) serà la que predomini.

1. Presentació i objectius

En aquesta PEC, un cop familiaritzades amb les dades d' expressió, i els mètodes i eines per a la selecció de gens i l' anàlisi de la significació biològica, procedim a la realització d' una anàlisi de dades, que ens permetran millorar la nostra comprensió d' un problema biològic mitjançant mètodes i eines estadístiques i bioinformàtiques.

L'anàlisi és semblant, tot i que no necessàriament coincident, amb alguns dels casos resolts que us hem proporcionat, per la qual cosa podeu inspirar-vos-hi però, sobretot, deveu entendre cada pas que feu.

1.1 Descripció de la PEC

La PEC es basarà en les dades d'un estudi que, utilitzant un model murí (de ratolí) investigo la utilitat dels antibiòtics LINEZOLID i VANCOMICINA per a immunomodulació durant infeccions per Staphylococcus aureus resistent 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
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

sample	infection	time	agent
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

Com es pot veure a la taula 1, el dataset consta de 35 mostres, 15 preses abans de la infecció i 20 després, 5, que eliminarem, a les 2 hores de la mateixa i 15 a les 24 hores.

2. Preguntes

El nostre objectiu serà intentar caracteritzar, a través del canvi en l'expressió gènica, l'efecte de la infecció i del tractament amb antibiòtics així com comparar els efectes d'aquests.

És a dir haureu de fer les comparacions següents:

- Infectats vs no infectats sense tractament
- Infectats vs no infectats tractats amb LINEZOLID
- Infectats vs no infectats tractats amb VANCOMICINA

Això generarà tres llistes de gens que haureu, d'una banda caracteritzar mitjançant anàlisi de significació biològica i, d'altra banda, comparar entre elles.

2.1 Preparació de les dades

Podeu descarregar-vos les dades crues del lloc de GEO https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38531 on també trobareu informació sobre l'estudi original.

Per tal de simplificar l'anàlisi i dificultar un mínim l'intercanvi no permès d'informació haureu d'eliminar algunes mostres: - D'una banda prescindirem de les cinc mostres preses a les dues hores - D'altra banda sortejarem les mostres restants de manera que cal conservar tan sols quatre mostres de cada grup.

Això ho podeu fer amb la funció `selectSamples` que trobareu a l'arxiu `selectSamples.R` i que us permetrà extreure 24 mostres diferents a cadascú amb tan sols anomenar-la usant com a llavor (argument "seed") el vostre DNI (sense la lletra) o, preferiblement el vostre identificador de la UOC.

Després d'aplicar la funció `selectSamples` obtindreu un nou objecte que us permetrà crear un nou `ExpressionSet` llegint únicament aquells arxius . CEL que hi ha.

Observeu que la taula no conté els noms exactes dels arxius. CEL pel que *haureu d'encarregar-vos vosaltres d'adaptar el que creieu necessari per poder llegir-los*.

A partir d'aquest `ExpressionSet` personalitzat, amb 24 mostres haureu de realitzar la vostra anàlisi que consistirà en el següent:

2.2 Anàlisi exploratòria i control de qualitat

Comenceu amb l'exploració habitual que us permeti decidir si les dades necessiten alguna transformació, si presenten algun problema i si els grups que voleu comparar se separen mínimament

Podeu complementar la vostra exploració amb un control de qualitat amb el paquet `arrayqualitymetrics`,

Teniu en compte que vaig començar treballant amb dades crues que convertireu en una matriu d'expressió després de normalitzar aquestes dades fent servir, per exemple, l'algoritme RMA.

2.3 Filtratge de les dades

Tot i que, com sabeu, el filtratge és una cosa discutida, podeu per exemple eliminar les sondes menys variables i quedar-vos amb el 10% de sondes que presentin més variabilitat.

2.4 Construcció de les matrius de disseny i de contrastos

Per realitzar l'anàlisi haureu de crear les matrius de disseny i de contrastos i utilitzar-les per dur a terme les comparacions proposades. Recordeu que deveu fer tres comparacions el que equival a tres contrastos.

2.5 Obtenció de les llistes de gens diferencialment expressats *per a cada comparació*

Utilitzeu `limma` per obtenir una llista de gens diferencialment expressats, seguint els exemples presentats en les notes i els casos resolts.

Les comparacions entre les llistes de gens la podeu fer gràficament o usant la funció de `limma`.

2.6 Anotació dels gens

L'anàlisi amb `limma` ens arreplega llistat d'identificadors basats en els identificadors originals. Amb aquestes llistes haureu d'anotar-los, és a dir associar-los algun identificador com "Symbol", "EntrezID" o "EnsemblID"

2.7 Anàlisi de la significació biològica

Un cop anotats els gens podem intentar interpretar els resultats intentant determinar si les llistes es troben enriquides en algunes categories biològiques

Per això podeu dur a terme una anàlisi de sobre-representació o un Gene Set Enrichment Analysis. Podeu utilitzar per a això el paquet `clusterProfiler` que us permet fer totes dues anàlisis de forma molt molt similar.

També permet, de forma molt senzilla, visualitzar els resultats de l' anàlisi de significació biològica, la qual cosa ajuda a comprendre *i comparar* els resultats.

2.8 Informe de l' anàlisi

Finalment, i com de costum, haureu d'elaborar un informe del vostre treball fent servir Rmarkdown. Aquí deveu tenir en compte el contingut i la construcció.

- Pel que fa al contingut l'informe ha de tenir l'estructura habitual de qualsevol treball: (i) Taula de continguts, (ii) Introducció i Objectius, (iii) Mètodes, (iv) Resultat (v) Discussió (vi) Referències i (vii) Apèndixs. A l'apèndix podeu posar el codi R que heu utilitzat per a la vostra feina i així serà un únic document.

- Pel que fa a la construcció deveu preparar document a Rmarkdown que generi l'informe a HTML i que haureu d'imprimir a pdf per lliurar-lo. Si teniu instal·lat alguna versió de LateX és probable que podeu generar l'arxiu .pdf directament. El com genereu el pdf queda a la vostra elecció.

Haureu de lliurar **un únic arxiu** en format pdf amb l'estructura anterior. L'arxiu ha de ser llegible per la qual cosa no ha de contenir llistes immenses ni llargs fragments de codi, que haureu de col·locar en apèndixs del document.

3. Recursos

Els recursos per a la resolució de la PEC són els que s' han proporcionat a l' aula fins al moment, és a dir, els materials del curs i casos d' estudi.

4. Criteris de valoració

Tal com s' indica en el pla docent, aquesta PEC val el 40% de la nota atès que a aquest segon repte es dedica aquest percentatge de les hores del curs.

Ara bé, i com a cosa important, recordeu que la PEC en si mateixa és un exercici de síntesi i aprenentatge en què intenta valorar la vostra capacitat per resoldre un problema molt semblant als que es troba un/a bioinformàtica/a en el seu dia a dia. Això vol dir que per a més d'un dels passos que haureu de fer no hi ha una solució única. Plantegeu la vostra pròpia solució i expliqueu perquè creieu que és l'adequada. Entre altres coses valorarem:

- Capacitat de definir correctament els objectius a assolir
- Capacitat d' organitzar l' anàlisi, obtenció de les dades, preparació dels arxius etc.
- Domini adequat de les eines pròpies del tema (R, Rmarkdown, BioConductor)
- Capacitat d' explicar què i perquè es fa en cada pas.
- Capacitat d' interpretar els resultats obtinguts.
- Capacitat de discutir les possibles limitacions de l' estudi.
- Presentació del treball en un document llegible i ben organitzat.

4.1 Codi d' honor

Quan presenteu exercicis individuals us adheriu al codi d'honor de la UOC, amb el qual us comprometeu a no compartir la vostra feina amb altres companys o a demanar de la seva part que ells ho facin. Així mateix, accepteu que, de procedir així, és a dir, en cas de còpia provada, la qualificació total de la PEC serà de zero, independentment del paper (copiat o copiador) o la quantitat (un exercici o tots) de còpia detectada.

Addicionalment us recordo que l'ús de programes d'IA, com ChatGPT, Perplexity o similars, no està permès. Òbviament això és molt difícil de controlar, però, a més del vostre compromís, mitjançant el codi d'honor en què accepteu seguir la normativa de les universitats, no haureu d'oblidar que aquests programes tendeixen a al·lucinar, per la qual cosa són més senzills d'identificar del que un creu.