Anàlisi de dades Òmiques (M0-157)
Segona prova d'avaluació contínua.

Contingut

Data de publicació de l'enunciat: 7/12/2023

Data límit per presentar la PEC: 21/12/2023[1]

1. Presentació i objectius

En aquesta PEC, un cop familiaritzades amb les dades de microarrays, 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 hagen de seleccionar de la llista de "datasets" de GEO proporcionada en l'arxiu "GEOdatasets_Enhanced.xlsx", el contingut

[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.

del qual apareix també en l'apèndix. Un cop seleccionat l' estudi, amb les dades d' aquest, deveu:

1. Plantejar les qüestions que desiebreu respondre (*normalment seleccionar gens diferencialment expressats entre dos o més grups*)
2. Realitzar les anàlisis necessàries (*preprocessat, selecció de gens, anàlisi de significació ORA i GSEA*) i
3. Elaborar un informe explicant objectius (problemes a resoldre), mètodes, resultats i discussió.

Recordeu que tan important com el resultat és el raonament i el procés que us hi aixequi, és a dir el consultor ha de poder veure no tan sols on heu arribat sinó també com i perquè heu arribat fins allà.

A la pràctica això significa que deveu:

1. Seleccionar el "dataset" amb el qual treballareu de la llista que us proporcionem. -El meu consell és que no treballeu amb cap "dataset" amb un nombre excessiu d'arrelaments
2. Per simplificar la PEC, **no caldrà que treballeu amb les dades crues (arxius .CEL)** per la qual cosa podeu utilitzar directament les dades que obteniu en descarregar l'estudi de GEO amb el paquet `geoQuery` (tal com vam veure a l'activitat 2). Això us proporcionarà un objecte `amb` dades normalitzades i llestes per a l'anàlisi[2].
3. En aquest cas el control de qualitat no analitzarà les sondes sinó directament els valors normalitzats. Val la pena fer alguns boxplots i PCA per comprovar que, efectivament es disposa de dades normalitzades. Podeu fer-ho amb el paquet `arrayQualityMetrics` com en els casos d'exemple, o bé directament fent servir codi R "ad-hoc".
4. Per realitzar l'anàlisi heu de crear les matrius de disseny i de contrastos i utilitzar-les per dur a terme les comparacions proposades (per vosaltres).

- [2] Per a l'extracció de dades de GEO usant GEOQuery us serà d'especial utilitat revisar els vídeos i el cas d'estudi de Bioconductor que us proporcioneu per a l'activitat 3.
    - BioC-3-Trabajando con ExpressionSets ***(Nuevo)***
    - BioC-4-Descarga de datos con GEOquery ***(Nuevo)***
    - Caso de estudio de introducción a Bioconductor ***(Actualizado)***

- En la majoria dels estudis la vostra anàlisi es basarà en un model d'un factor, però si creieu que deveu elaborar un model més complex, no us en priveu.
- Si l'estudi s'hi ha publicat podeu accedir-hi des de l'apartat "citations". L'abstractament n'hi hauria prou per fer-vos una idea sobre què tracta l'estudi i per tant *que hauríeu de fer*.

5. Amb la llista de gens resultants de l'anàlisi deveu
   - Anotar-los, és a dir associar-los algun identificador com "Symbol", "EntrezID" o "EnsemblID"
   - Un cop anotats, realitzar un estudi de significació biològica que us serveixi per aventurar el significat dels resultats que obteniu.
   - El meu consell és que, en comptes d'utilitzar el paquet `GOstats` utilitzeu `clusterProfiler` que us permet fer Anàlisi d'enriquiment i GSEA amb un codi molt similar[3]. També permet, de forma molt senzilla, visualitzar els resultats de l' anàlisi de significació biològica, la qual cosa ajuda a comprendre els resultats.
6. 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.

---

[3] Per a l'anàlisi de significació biològica podeu basar-vos en els materials d'estudi però també pot ser-vos d'utilitat el material complementari: que us proporcioneu en l'activitat 3:

https://aspteaching.github.io/An-Introduction-to-Pathway-Analysis-with-R-and-Bioconductor/

7. Haureu de lliurar **un únic arxiu** en format pdf amb l'estructura anterior i el nom del qual sigui la concatenació dels vostres cognoms, el nom i la paraula PEC1, **sense accents ni espais en blanc**.
    - Per exemple, en el meu cas l'arxiu que lliuraria es denominaria "Sanchez_Pla_Alex-PEC2.pdf"

1.2 Recursos

Els recursos per a la solució 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. En les notes al peu de la secció anterior s' indiquen també els materials complementaris d' interès per a la PEC.

1.3 Criteris de valoració

Tal com s' indica en el pla docent, cada PEC val el 30% de la nota.

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.

1.4 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

2. Apèndice: *Conjunts de dades*

A l'arxiu "GEOdatasets_Enhanced.xlsx" trobareu informació sobre cada "dataset". La informació inclou el nombre de mostres, el tipus de microarray i els grups experimentals. Alguns estudis tenen dos o més factors i aquests apareixen a la columna "Description" perquè no s'ha elaborat suficientment la preparació de la taula.