Bon dia José Luis,

Aniré per punts:

1) Qui ets,  
2) que va motivar que decidissis estudiar aquest màster a la UOC, i  
3) quin és el teu objectiu professional un cop hagis finalitzat els estudis.

Vaig estudiar Ciències Polítiques i vaig fer posteriorment el màster en Politics, Economics and Philosophy a la universitat d’Hamburg a Alemanya. Ja com a estudiant vaig començar a treballar a l’hospital universitari d’Hamburg a la unitat d’oftalmologia on vaig fer anàlisi estadística per als estudis científics en oftalmologia. Aquest gir és degut al fet que dins dels meus estudis em vaig començar a interessar pels mètodes estadístics i la programació. Així a poc a poc el meu interès es va girar cap a biologia i medicina. Treballava com a investigador a la Universitat d’Hamburg fins l’any 2015 quan em vaig casar amb una noia de Tarragona i em vaig traslladar a València on des de llavors estic treballant full time en una I+D d’una clínica privada fent pràcticament el mateix: l’anàlisi estadística en oftalmologia. A part d’això continuo col·laborant amb la universitat d’Hamburg. M’agrada estudiar i així vaig decidir a adquirir més coneixement de biologia estudiant el màster a la UOC. Per què la UOC? Doncs perquè volia aprofitar i millorar el meu català. A part d’estadística, informàtica i biologia m’agraden també llengües i en domino moltes. El meu objectiu professional està connectat als meus objectius familiars: la meva dona treballa a Barcelona i jo a València i estem separats bona part de la setmana. Desprès del màster intentaré trobar una posició interessant a Barcelona i d’aquesta manera ser més a prop de la meva dona i filla. A part d’aquest objectiu professional tinc també l’objectiu acadèmic: vull fer un doctorat per continuar estimular-me intel·lectualment.

Dit això, passo fer-te alguns comentaris/suggeriments i plantejar-te pel mig una sèrie preguntes:

1) En la línia del que apunta l’Alex, és important que l’usuari disposi d’un exemple d’execució i d’unes dades que pugui usar per a testar l’aplicatiu. Què pots utilitzar? Doncs precisament pots aprofitar l’exemple que presentes a la “Validació dels resultats”.

Com observes correctament al punt següent no he creat cap de les funcions que generen els resultats. La meva feina consistia en donar-les vida en una aplicació amb interfície amigable. Crec que no seria possible crear-les en el marc temporal definit per TFM. Les funcions provenen del paquet clusterProfiler en conjunt amb els paquets ReactomePa i pathview de Bioconductor. Per aquest motiu penso que per mostrar que l’aplicació funciona seria millor utilitzar les mateixes dades que s’utilitzen en el vignette de clusterProfiler (i el paquet relacionat DOSE).

<https://yulab-smu.github.io/clusterProfiler-book/chapter3.html#input-data>

D’aquesta manera l’usuari pot ser confident que les funcions que jeien a base de l’aplicació són les mateixes funcions creades per als autors de clusterProfiler, perquè els resultats serien idèntics. De fet ho vaig fer a l’apartat “L’aplicació” de la memòria. Aquestes dades són guardades al meu Github i estan disponibles via enllaç que proporciono a d’alt. Podria modificar el codi fàcilment per incorporar aquestes dades en l’aplicació però al moment ho puc fer actualització del paquet. L’aplicació online és al servidor de l’Alex i hi tinc accés.

2) Lligant amb el punt anterior, és evident que tu el que has fet és implementar un aplicatiu que es basa en una sèrie de mètodes que ja estan validats. És a dir, entenc que no has modificat les funcions que afecten al còmput dels mètodes estadístics implementats pels diferents autors. Per tant, en el fons una validació com a tal, no és necessària. Però és obvi que el que vols és comprovar/mostrar que recuperes les mateixes vies i/o xarxes que els autors originals. En aquest sentit, jo el que et suggeriria és que utilitzessis un altre set de dades. Per què? Doncs perquè Li et al. 2017 utilitzen DAVID, que no és ni un mètode ORA ni tampoc un mètode FCS, si no un mètode que va a cavall entre aquests dos i que inicialment es va classificar com un mètode Modular Enrichment Analysis (MEA). Llavors, la meva pregunta és, què tan difícil seria extendre la implementació d’aquesta altra via en el teu aplicatiu?

No, no he modificat cap de les funcions generadores dels resultats ORA i GSEA. Però sí que volia validar aquestes funcions en sentit que tingués els resultats que caldria d’esperar d’un experiment fet d’una manera específica. Li et al. 2017 volien mostrar l’efecte de gen Zbtb7b al procés de producció dels greixos marrons. Per tan els ratolins amb aquest gen desactivat haurien de mostrar les rutes alteradament expressades relacionades amb quest procés. I com s’observa als resultats ho podem confirmar a partir de KEGG i Reactome. Les funcions de clusterprifiler no han trobat cap concepte enriquit per a les ontologies GO, cosa que em va preocupar bastant, perquè els autors que comentes tu sí que han trobat els conceptes significatius. Clar, com comentes aquí els Autors van utilitzar l’eina DAVID i el mètode que s’hi utilitza és diferent. Modular Enrichment Analysis és una extensió de l’anàlisi ORA molt útil tenint en compte que agrupa els conceptes de GO encara més dins d’un clúster dels conceptes relacionats i d’aquesta manera tracta millor la redundància que és deguda a l’estructura jeràrquica dels termes GO\*. Malauradament no veig cap paquet del Bioconductor que ofereix aquest mètode. Per tant no veig la incorporació d’aquest mètode a l’aplicació de manera fàcil. Seria però molt interessant desenvolupar un paquet que el podria oferir.

\* Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic acids research* 37.1 (2008): 1-13.

3) Hi han alguns conceptes del marc teòric que pot ser caldria donar-li alguna volta més, però no vull ser meticulós. En aquest cas no es tracta d’una tesi doctoral, i en el reduit temps que has tingut per a fer aquesta feinada ja és més que lloable el que presentes. Ara si, he notat que hi han alguns termes que són rellevants i m’agradaria que em comentessis, perquè crec que en el teu cas és més aviat una qüestió de “lost in translation”. Per tant, series tan amable d’explicar-me que entens per:  
(a) Anàlisi de sobreexpressió  
(b) Anàlisi de sobrerepresentació  
(c) Anàlisi de significació biològica  
(d) Anàlisi d’enriquiment

Tens raó. Hauria potser ser més detallat i exacte a l’hora d’utilitzar certes termes.

a+b) Els termes sobre i sotaexpresió he utilitzat a la memòria majoritàriament referint-me a l’experiment d’expressió genètica, com ara l’experiment de microarrays. Vol dir sobre o sotaexpressió d’un gen d’una mostra d’interès (com per exemple teixit de càncer) comparat a les mostres biològiques de referència (teixit normal). Veig també que al començament del paràgraf sobre ORA he escrit “L'anàlisi de sobreexpressió és una tècnica...”. Aquí sí que és un error de traducció. Hauria d’escriure “l’anàlisi de sobrerepresentació” referint-me a “over representation”.

c) Suposo que et refereixes a la frase al final:

“En tercer lloc, he descrit quins mètodes existeixen per calcular la significació biològica de les rutes.”

Aquí hauria d’escriure “significació estadística”. De totes maneres amb significació biològica entenc la fase d’interpretació dels resultats. Anàlisi de les rutes genera els resultats que poden ajudar a formular la significació biològica per a organisme.

c) Entenc que l’anàlisi d’enriquiment és el sinònim de l’anàlisi de les rutes. Vol dir tots els mètodes que he apuntat: ORA, FCS i PT (pathway topology). Per ser precís apunto però que els termes GO no són tècnicament rutes. Però entenia també l’enriquement d’aquests conceptes dins del marc de l’anàlisi d’enriquiment.

4) Tant a la memòria com a la presentació expliques que has basat el teu projecte en les bases de dades de GO, KEGG, i Reactome. No sé si tens clar que en realitat GO i KEGG no són bases de dades. De fet, GO és una ontologia i KEGG una col·lecció de bases de dades. Per contra si és una base de dades Reactome. Tanmateix, comentes que volies incorporar wikiPathways, però no ho has fet. M’agradaria saber doncs  
(a) Per què no has incorporat wikiPathways? Què en penses d’aquest projecte?  
(b) T’has plantejat incorporar altres recursos al teu aplicatiu? Si és que sí, quins? Si és que no, com procediries a fer una selecció representativa?

Sí. Tinc clar les diferències. El consortium go però es refereix a si mateix a moltes ocasions com a la base de dades. Per exemple llegim a un dels seus articles:

Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource." *Nucleic acids research* 32.suppl\_1 (2004): D258-D261.

Crec que es refereix al fet de com les anotacions són guardades: amb sql – la base de dades relacionals.

De mateixa manera crec que també KEGG és una base de dades, o com a mínim o diu la mateixa KEGG:

Aquí: <https://www.genome.jp/kegg/pathway.html>

I aquí: <https://www.genome.jp/kegg/>

No he incorporat wikiPathways purament pels raons de temps. Volia fer realista i escollir com a molt tres recursos. Buscant a l’internet m’he adonat que actualment l’anàlisi de les rutes es fan majoritàriament basant-se en Reactome i KEGG. Volia però integrar també ontologies de gens (GO) per donar més variabilitat a l’anàlisi. A més a més veia la possibilitat de visualitzar les rutes KEGG i Reactome. En canvi les rutes de wikiPathways actualment no es pot visualitzar amb clusterProfiler. No vol dir però que no hi ha cap eina en Bioconductor per fer-ho. Aquesta visualització la podríem aconseguir per exemple amb el paquet rWikiPathways. En general penso que wikiPathways és una bona iniciativa especialment per fet de ser tan oberta.

5) He notat que en algun lloc parles de “topografia”. Novament, entenc que es tracta d’una qüestió lingüística. El terme correcte és el de “topologia” (que és la branca de les matemàtiques que s’encarrega d’estudiar les propietats de cossos geomètrics que es mantenen inalterables quan se’ls hi apliquen unes transformacions específiques). En el cas de l’estudi dels pathways i/o xarxes de característiques (e.g. gens, proteines, metobòlits,…) es parla de topologia de xarxa. Referent a aquest punt, el que has fet és implementar mètodes de visualització. Has explorat els mètodes/paquests de Bioconductor que fan anàlisi de xarxes? Segons la teva opinió valdria la pena incorporar-los en el “pipeline d’anàlisi” que li comentaves a l’Alex?

Sí, la paraula “topografia” no seria correcte. Utilitzo aquesta paraula freqüentment en context d’oftalmologia parlant sobre topografia corneal i s’ha encolat a la memòria. A l’aplicació, com comentes, he implementat la visualització de les rutes KEGG i Reactome on els gens que participen en la ruta són emfatitzats d’acord amb el grau de sobre o sota expressió (fold change). Clarament l’usuari ha de tenir un coneixement suficient per poder deduir com aquests gens influeixen desenvolupament normal de la funció de la ruta.

L’aplicació no ofereix però l’anàlisi topològic pròpiament dit. L’anàlisi topològic calcula la puntuació d’enriquement (enrichment score) no només utilitzant les diferències d’expressió dels gens individuals sinó també la topologia de la ruta. Així s’hi integra l’estructura de correlació entre els gens dins de la ruta.

En el Bioconductor trobem el paquet EnrichmentBrowser que integra altres paquets que ofereixen mètodes d’anàlisi topològic:

GGEA: Gene Graph Enrichment Analysis que evalua la consistència de les interaccions conegudes amb l’expressió observada en les dades.

SPIA: Signaling Pathway Impact Analysis que combina ORA amb la probabilitat que els canvis d’expressio es propaguen per la topologia de la ruta.

PathNet: Pathway Analysis using Network Information

DEGraph: Differential expression testing for gene graphs

TopologyGSA: Topology-based Gene Set Analysis

GANPA: Gene Association Network-based Pathway Analysis

CePa: Centrality-based Pathway enrichment

NetGSA: Network-based Gene Set Analysis

Veiem que hi ha moltes opcions i com ja he comentat a Alex en resposta al seu comentari, tanta varietat resulta a un factor de distracció dins del marc temporalment reduït del TFM. Clarament a mi m’agradaria incorporar l’anàlisi topològic a l’aplicació. Això però implicaria selecció de l’anàlisi adequat i implementació dins de l’estructura de l’aplicació. Tot això comportaria inversió de temps considerable, perquè tots aquests mètodes de topologia i la seva implementació via els paquets de Bioconductor son nous per a mi. En canvi tan ORA com GSEA vaig conèixer dins del currículum del màster i per tant la seva implementació a l’aplicació era mes o menys fàcil.

6) clusterProfiler és un paquet que ha anat evolucionant amb el temps, com la majoria de paquets de Bioconductor. De fet, en el darrer any va introduir canvis considerables en el seu codi que, fins i tot afectaven a la parametrització de funcions tan importants com les del enrichment map (i.e. emapplot). Suposa que tornen a introduir un gran canvi, i no tens l’aplicatiu instal·lat en un servidor. Afectaria molt a l’estructura interna del teu paquet? En aquest sentit, has pensat preparar una “solució dockeritzada”.

Wow. Abans dels vostres comentaris ni he sentit parlar-ne... Estic molt content que puc aprendre fins i tot a l’última fase de TFM. Responent a la pregunta d’Alex sobre l’escalament de l’aplicació vaig comentar que es podria utilitzar el mètode de contenidors per crear instància de l’aplicació per cada usuari, i d’aquesta manera fugir de “single-threaded” naturalesa de R. Llegint més sobre dockering per poder respondre a la teva pregunta veig que es pot utilitzar els contenidors també amb l’objectiu de reproductibilitat de l’aplicació, com ho comentes tu, perquè el que fa el docker és no només empaquetar l’aplicació sinó totes les seves dependències. Clarament és un mètode molt útil que investigaré més en el futur.

7) A la diapositiva número 4 de la presentació, sobre el minut 2 aproximadament, parles de:

“Més potència estadístic via reducció del nombre d’hipòtesis múltiples”

No sé si acabo d’entendre el que vols transmetre. Series tant amable de desenvolupar-ho una miqueta més.

Sí. Aquest punt l’he vist també a la presentació següent a la diapositiva 4:

<https://reactome.org/docs/training/Pathways_&_Networks_Overview.pdf>

Em va semblar molt plausible, i el vaig utilitzar també en la meva presentació. Com ho entenc jo, en anàlisi estàndard d’expressió diferencial la significació estadística de cada gen està avaluada independent, i desprès s’utilitza un mètode de correcció del valor P per evitar la seva inflació com a resultat de les comprovacions de les hipòtesis múltiples. Aquests mètodes són normalment menys sensitius per detectat els gens amb les deferències petites en abundància d’ARNm i per tan poden tenir menys potència com a un li agradaria. Amb l’anàlisi de les rutes es redueixen les hipòtesis múltiples i d’aquesta manera s’augmenta la potència estadística.

Finalment, m’agradaria que m’expliquessis una sèrie de qüestions més generals, i són les següents:

1) Quina és la part que t’ha portat més dificultats en el desenvolupament de l’aplicatiu

A part de que el desenvolupament d’una aplicació Shiny era una cosa bastant nova per a mi, no em va resultat gens difícil programar perquè ja tenia experiència de amb R. Shiny té les seves peculiaritats però al final s’hi acostumes. Més problemes tenia amb implementació de les rutes KEGG via pathview, cosa que finalment vaig solucionar amb modificació de codi original de la funció pathview::pathview. Encara no tinc tots els conceptes clars de Shiny i per tant em sembla que el meu codi és molt repetitiu. Baix la pressió de temps no volia experimentar molt amb el codi, però estic segur que hi ha possibilitats de fer-lo més professional creant més funcions etc.

2) Quins aspectes positius destacaries del teu treball

Personalment he après moltes coses i per tant tot el projecte era molt positiu per a mi. L’aplicació va arribar a funcionar, i veure-la online és també enormement positiu. L’aplicació ofereix els mètodes bàsics, però què siguin ser actuals en l’anàlisi de les rutes. L’aplicació dona resultats numèrics i les visualitza, cosa que fa l’aplicació interessant i atractiu. Es dona possibilitat a valorar termes GO, Reactome i KEGG les “bases de dades” més conegudes actualment en quan l’anàlisi d’enriquement. L’aplicació és opensource i pot ser utilitzada i modificada per qualsevol que en tinc interès. És una base modesta, però estable per implementar noves funcionalitats.

3) Quin consell li donaries a un estudiant de la UOC que hagi de fer el seu TFM

Dividir tot el projecte en metes petites i intentar fer una cosa, la petit que sigui, cada dia. Cada pedra fa paret: una expressió que descriu molt bé aquest mètode de treball. Un projecte amb un objectiu molt llunya pot resultar frustrant al principi, però cal començar i continuar fent construint tot el projecte amb “pedres petites”. D’altra banda no cal obsessionar-se amb perfecció. Cal fer fins que funcioni, sí més no, amb un codi ineficient i anar cap a l’altre objectiu, Moltes vegades la solució per millorar el codi, o el text de memòria, apareix posteriorment, fins i tot, quan no hi penses. Això és inevitable amb una temporització estricta.

De nou felicitats pel teu treball, pot estar-ne molt orgullòs.

Resto doncs a l’espera de les teves respostes. Moltes gràcies.

Salutacions,