Bon dia,

**Diapositiva 1**

Em dic Vasyl Druchkiv i us vull presentar el resum i resultat del meu treball de màster. En el meu treball he tingut com a objectiu principal la implementació d’una eina shiny de R studio per dur a terme l’anàlisi de rutes. Aquesta feina l’he fet en el marc del màster de biosestadística i bioinformàtica de la universitat oberta de catalunya supervisat per professor Alex Sanchez-Plà.

**Diapositiva 2**

La presentació està estructurada de manera següent:

Primer explicaré breument que és l’analisi de rutes. Especialment el que volem aconseguir amb ella.

En segon lloc explicaré què és necessari per dur-la a terme. Aquí apuntaré dues línies importants: la procedència de les dades d’un experiment de microarrays i la seva anotació amb les bases de dades.

Tot seguit parlaré sobre els mètodes que he triat per fer aquesta anàlisi. Aquí destacaré els mètodes ORA i GSEA però també uns mètodes de visualització de les rutes.

A continuació us explicaré com instal·lar l’aplicació. Aquí explicaré que he habilitat un repositori Githib on he guardat, a part de l’aplicació empaquetada, tot el material relacionat amb el desenvolupament de l’aplicació.

Finalment presentaré un exemple d’ús de l’aplicació sobre les dades d’un experiment on s’investiga la relació del gen Zbtb7b amb producció dels greixos marrons.

**Diapositiva 3**

Un Pathway és un conjunt de gens que actuen junts per dur a terme un procès biològic. El que obtindrem però amb un experiment d’expressió genètica és una llista de gens indivuduals. D’aquesta manera agregant el resultat d’aquest experiment al nivell de les rutes podem reduir la informació i la complexitat. A més a més amb les rutes obtindrem una millor comprensió del funcionament mecànic de la biologia de la condició estudiada. Així podem agrupar els gens individuals als sets basant-nos en processos biològics, moleculars o els components cel·lulars. Les visualitzacions de les rutes són més intuïtives: aquí podem valorar la relació entre els gens i la seva importància relativa dins de la ruta. Des del punt de vista estadístic amb l’anàlisi de rutes podem aconseguir més potència estadística, perquè hi ha molts menys conceptes que els gens individuals, cosa que implica menys proves d’hipòtesis.

**Diapositiva 4**

Ja he començat a parlar sobre un experiment d’expressió genètica. Concretament m’agradaria apuntar l’experiment de microarrays. A la imatge podem apreciar el workflow d’aquesta anàlisi.

Comencem amb la formulació d'una pregunta científica i un disseny experimental. Els productes de l'experiment són bàsicament les imatges d'intensitats que es tradueixen en valors numèrics. Habitualment aquests valors han de ser processats adequadament. Aquest processament inclou el control de qualitat de les imatges i la normalització dels valors d'intensitat per reduir la variabilitat tècnica. Finalment les dades normalitzades s'utilitzen per a l'anàlisi estadística. El resultat d'aquesta anàlisi serveix com a base per a la interpretació biològica dels resultats de l'experiment. Per a l’anàlisi de les rutes es necessita una llista ordenada de les expressions relatives LogFc i una subllista de gens que hem identificat mitjançant els tests estadístics com a diferencialment expressats.

**Diapositiva 5**

Els gens individuals han de ser anotats als conceptes/rutes diferents. Aquesta anotació implica l’ús de les bases de dades que guarden aquest coneixement. Normalment l’assignació d’un gen a una o altra ruta es fa amb mètodes experimentals els resultats dels quals es publiquen. Les bases de dades grans com ara Gene Onthology, KEGG i Reactome reculen aquesta evidència sobre un gen i la integren dins de les seves categories o rutes.

El consortium Gene onthology inclou tres ontologies separades: 1) funció molecular; 2) funció biològica i components cel·lulars CLICK. És una base de dades jeràrquica i inclou connexions entre els conceptes com ara: is-a, part of o regulate. Els conceptes GO es representen habitualment en els gràfics acíclics on els conceptes estan ordenats jeràrquicament sense formar cercles. Notem però que la base de dades GO no representa rutes pròpiament dit. Per tant quan parlo sobre els termes GO m’hi refereixo com a conceptes.

Les bases de dades especialitzades en les rutes són KEGG i Reactome. Les rutes de KEGG estan organitzades en 7 categories (CLICK); cada una de les quals inclou rutes específiques.

La base de dades Reactome (CLICK) també està organitzada jeràrquicament amb top-level pathways representats com els nodes que es despleguen fins arribar a les reaccions específiques. Utilitzant el pathway browser podem trobar per exemple, la ruta [Synthesis of very long-chain fatty acyl-CoAs en Homo sapiens)](https://reactome.org/PathwayBrowser/#/R-HSA-75876). Aquesta ruta està involucrada en la producció dels greixos marrons.

**Diapositiva 6.**

Ara tenim d’una banda resultats d’un experiment de microarrays, i d’altra banda les bases d’anotacions. El que ens falta són els mètodes estadístics que identifiquen les rutes o categories diferencialment anotades. Hi ha molts mètodes que podem agrupar dins de les tres categories:

1. Over-Representation analysis
2. Functional Class Scoring
3. Pathway topology

Per dur a terme ORA necessitem una llista de gens diferencialment expressats i el nombre de gens de referència dins de les rutes. Aquest nombre de referència pot ser obtingut de les bases de dades d’anotació o de tots els gens analitzats via l’experiment de microaarays. El problema amb ORA és que ella utilitza només els gens més significatius i descarta els altres. Amb els mètodes de funcional class scoring s’intenta corregir aquesta deficiència.

Per als mètodes de funcional class scoring no necessitem la llista dels gens diferencialment expressats, sinó la llista de tots els gens d’experiment amb la seva expressió relativa, on l’expression relativa pot ser donada amb els log Fold changes.

Tan ORA com GSEA no visualitzen les relacions entre les rutes i entre els gens dins de les rutes. Els avenços en anotació manual de les bases de dades disponibles contenen però aquesta informació i l'aplicació, que he desenvolupat, gràcies al paquet \helvetica{clusterProfiler}, hi treu l'avantatge i visualitza aquestes relacions més detalladament.

**Diapositiva 7**

El mètode que utilitzo en l’aplicació és el del paper de Boyle. El mètode estadístic descrit consisteix bàsicament en els passos següents:

1. De tots els gens de la mostra seleccionar un grup de gens que es considera que són significativament expressats.

Els criteris de selecció poden basar-se en logRatios i/o en el valor de p provinent d'un test estadístic

1. Determinar si algunes rutes anoten la llista especificada de gens amb la freqüència més alta que la que s’esperaria per casualitat.

El test estadístic es basa, com ho veiem a la diapositiva, en la distribució hipergeomètrica. La idea és simple (Click) i la podem expressar en una taula creuada. Imaginem nos, que tenim 263 gens diferencialment expressats i 25 d’ells formen part d’una ruta específica. Ho comparem amb la distribució de fons on 177 de gens pertanyen a la ruta d’interès. L’objectiu és determinar sí el nombre de gens de la llista de gens diferencialment expressats és significativament enriquida comparant amb la distribució de fons. Aixó implica el test de Fisher que ens dona el valor p molt baix per poder rebutjar la hipòtesi nulla amb molta confiança.

1. Els valors de p han de ser ajustats per tenir en compte les proves d’hipòtesis múltiples.

**Diapositiva 8**

Amb Gen Set Enrichment Anàlisis podem analitzar els resultats d'un experiment d'expressió per a dos grups. Aquí els gens són ordenats basant-se en la correlació entre la seva expressió i la separació entre les classes. Aquest llistat ordenat el podem crear utilitzant els logRatios.

Donat el conjunt definit dels gens, que pertanyen per exemple al mateix terme de Gene Ontology, l'objectiu de GSEA és determinar si els membres d’aquest conjunt són distribuïts aleatòriament en el llistat o es troben més al cap o a la cua. S'esperaria que els gens relacionats amb la separació fenotípica mostrin aquesta última distribució.

L'anàlisi GSEA consisteix en tres passos:

Primer hem de calcular la puntuació d'enriquiment, on la puntuació està calculada anant per la llista i augmentant la suma corrent sempre quan es troba un gen que pertany a la ruta específica, al contrari, restant-la quan el gen no hi forma part. La puntuació és la desviació màxima del zero observada en aquet camí. L'estadística obtinguda se sembla a l’estadística de Kolmogorov-Smirnov.

Segon hem d’estimar el nivell de significació per a la puntuació obtinguda anteriorment. El valor de P nominal el podem obtenir mitjançant la permutació de gens.

Finalment ajustem el valor de pa nominal per acomodar el problema de les proves múltiples.

**Diapositiva 9**

Els anàlisis ORA i GSEA retornen unes taules amb les categories i es convenient resumir les dades visualment. Els gràfics de barres i de punts visualitzen el nombre de gens diferencialment expressats en les rutes identificades per ORA. Aquí com ho veurem més endavant es pot relacionar el nombre de gens de la ruta específica en la llista de gens diferencialment expressats, el nombre total de gens diferencialment expressats i la significació estadística al nivell de la ruta.

El gràfic GSEA ens mostrarà la distribució de gens de la ruta al llarg del llistat ordenat de tots els gens d’experiment.

Els gràfics de Kegg i Reactome són més específics i detallats per a cada ruta significativament enriquida. Aquí es pot observar les relacions entre els gens de ruta i les reaccions en les quals són implicats.

Altres gràfics interessants que integro en l’aplicació són:

1. GO-Plots
2. Enrichment Maps
3. Gene-Concept Networks.

El gràfic GO està organitzat jeràrquicament com direccional acíclic gràfic (Directed Acyclic Graph) i mostra tres relacions: is-a, part of regulates,

Enrichment Map és un mètode de representació molt útil per poder reduir/simplificar la informació obtinguda mitjançant els mètodes. El mapa d’enriquiment redueix la redundància als conjunts de gens, que poden incloure els mateixos gens. Això és encara més significatiu per als conceptes GO, que són estructurats jeràrquicament. Els conjunt de gens estan representats com a nodes els radis dels quals són proporcionalment relacionats amb el nombre de gens que formen part d’aquests conjunts. Els cantells indiquen els nodes que tenen gens compartits, on el seu gruix depèn del nombre de gens compartits.

Amb Gene-Concept Networks es pretén visualitzar els gens al voltant dels conceptes on els gens poden ser connectats amb rutes (conceptes) diferents. D'aquesta manera es fa possible identificar les associacions biològiques més complexes entre les rutes mitjançant els gens.

**Diapositiva 10**

He contemplat opcions diferents de publicar l’aplicació. La més preferible seria publicar-la en un servidor WEB. Ho podria fer al meu ordenador de casa configurant el servidor Shiny. Clar com un exercici seria bé però no seria accessible de facto: el meu ordenador de casa no és pas un servidor i l’apago de tant en tan. Per poder fer-ho factible hauria de llogar un servidor i configurar-lo. Això implica un factor econòmic i temporal que no estic preparat d’assumir. El meu professor Alex Sanches-Pla, em volia donar accés al servidor de seu grup de recerca però en al moment actual, no hi tinc accés. Per tant que he fet és empaquetar l’aplicació i fer la disponible via GitHub. Però, tenint servidor shiny configurat es podrà depositar-hi l’aplicació relativament fàcil i ràpid.

L’usuari ha de tenir Bioconductor i el Paquet devtools instal·lats prèviament al seu R Studio. Quan l’usuari ho ha fet, ell pot instal·lar l’aplicació via funció install\_github. Quan l’aplicació està instal·lada l’usuari por accedir a l’aplicació via la funció runPathwayApp(). Si alguns dels paquets no són instal·lats localment l’aplicació primer procedirà instal·lar-los. He decidit no forçar la instal·lació de totes les bases d’anotacions per GO i Reactome, perquè això implicaria més temps d’instal·lació i requeriria més espai al disc dur localment, cosa que resultaria a una experiència ferragosa.

**Diapositiva 11**

Li i coautors analitzen l'associació del gen Zbtb7b amb la producció dels greixos marrons que al seu torn influeixen en termogènesis i processos metabòlics diferents. D'aquesta manera els greixos marrons són importants per al tractament dels desordres metabòlics.

Les dades d'estudi són ja preprocessades per Ricardo Gonzalo Sanz i Sanchez Pla i estan disponibles a Github. De la carpeta results he agafat la taula topAnnotated\\_KOvsWT\_COLD.csv. Sanz i Pla utilitzen el paquet ReactomePA per a l'anàlisi d'enriquiment. Repeteixo doncs el seu anàlisi utilitzant l'aplicació.

RStudio

1. Primer instal·lo l’aplicació. Ja tinc tant el Bioconductor com devtools preinstal·lats.
2. Comprobo si l’anotacions per al ratolí ja són instal·lades.
3. Crido la funció runPathwayApp. L’execució no triga gaire perquè tots els paquets necessaris els tinc ja instal·lats
4. Així arribem a interfície de l’aplicació.

L’aplicació està dividida en quatre parts: Selecció de dades, Anàlisi dels conceptes  GO, l’anàlisi de les rutes KEGG i l’anàlisi de les rutes Reactome. Per cada base de dades podem generar resultats ORA, GSEA i visualitzar-los via Bar-Plots , Dot-Plots, Enrichment-Plot, Category-Gene-Networ i les visualitzacions específics com ara Go-Plot, rutes KEGG i rutes Reactome.

Observem que el manual i ajudes estan indicats amb els símbols interrogatius als llocs oportuns.

Anem primer a la secció de les dades.

Seleccionem l’espècie d’interès, que en el nostre cas és Mus Musculus.

Pugem l’arxiu amb tots els gens i les gens diferencialment expressats.

Anem a l’apartat de Reactome i fem primer l’anàlisi ORA. Per això selecciono el nivell de significació de 0.05 i el mètode d’ajustament  Benjamini and Hochberg.

D’aquesta manera obtindrem tres categories. Les podem visualitzar amb gràfic de barres i de punts. Mirem també el mapa d’entiquement. Viem que totes les categories són conectades, cosa que implilca que comparten els gens. Amb el category-gene network podem identificar els gens compartits. Aquí veiem la importància d’aquesters representacions, perquè  sembla que aquí hi ha una estructura jeràrquica amb la ruta fatty acid metabolism com a top-level.

Visualitzem la ruta synthesis of very-long-chain fatty-acid acyl-coas.

Ara podem fer gene set enrichment analysis.

Totes les imatgen, les podem guardar especificant els paràmetres d’imatge.