

Vasyl Druchkiv, MSc.

Implementació d'una eina en R/Shiny per a l'anàlisi de significació biològica utilitzant l'anàlisi de les rutes

Memoria del treball

per aconseguir el títol de

Master universitari en

Bioinformàtica i bioestadística

entregat a

Universitat oberta de Catalunya

Supervisat per Dr. Alexandre Sánchez Pla (UB)

València, Maig 2019

Abstract

This is a placeholder for the abstract. It summarizes the whole thesis to give a very short overview. Usually, this the abstract is written when the whole thesis text is finished.

Contingut

Abstract	iii
1 Introducció	1
1.1 Objectius	2
1.1.1 Objectius generals	2
1.1.2 Objectius específics	2
2 El marc teòric	5
2.1 Les dades d'expressió genètica	5
2.2 Annotació dels gens	9
2.2.1 Gene ontology	9
2.2.2 KEGG	10
2.2.3 Reactome	11
2.3 ORA	11
2.4 GSEA	12
2.5 L'anàlisi topològic de les rutes	14
2.5.1 El mapa d'enriquement	15
2.5.2 Gene-Concept-Network	15
2.5.3 GOpot	15
2.5.4 KEGG Pathway	16
2.5.5 Reactome Pathway	17
3 Tractament bioinformàtic	19
3.1 Cerca dels paquets de Bioconductor	19
3.2 Desenvolupament del protocol	19
3.3 Instal·lació de l'aplicació	21

Contingut

4 L'aplicació	25
4.1 ORA	25
4.1.1 GO	25
4.1.2 KEGG	26
4.2 GSEA	28
4.2.1 GO	28
4.2.2 KEGG	29
4.2.3 Reactome	30
4.2.4 Bar-Plots	30
4.2.5 Dot-Plots	31
4.2.6 Enrichment Plots	33
4.2.7 Category-Gene-Network Plot	34
4.2.8 GSEA Plot	34
4.3 L'anàlisi específic de GO, KEGG i Reactome	36
4.3.1 GO Plot	36
4.3.2 KEGG Pathway	37
4.3.3 Reactome Pathway	38
4.4 Manual i les ajudes del programa	38
5 Validació dels resultats	45
5.1 Exemple d'anàlisi 1. GEO: GSE100924	46
6 Discussió	57
Bibliography	61

Llista de les imatges

2.1	El procès d'anàlisi de microarrays	6
2.2	El procès d'anàlisi de les rutes	7
2.3	El mètode GSEA	13
2.4	L'anotació de les relacions dins de les rutes KEGG	16
3.1	Lucidchart per a l'aplicació	20
4.1	Especificació d'ORA dels termes GO	26
4.2	El resultat d'anàlisi ORA. GO.	26
4.3	Configuració d'anàlisi KEGG	27
4.4	El resultat de l'anàlisi ORA. KEGG.	27
4.5	El resultat d'anàlisi ORA. Reactome.	28
4.6	El resultat de l'anàlisi GSEA. GO.	29
4.7	El resultat de l'anàlisi GSEA. KEGG.	29
4.8	El resultat d'anàlisi GSEA. Reactome.	30
4.9	Bar-Plot. GO.	31
4.10	Bar-Plot. GO.	32
4.11	Bar-Plot. GO.	33
4.12	Category-Gene-Network Plot. GO.	34
4.13	GSEA Plot. GO.	35
4.14	GO Plot	36
4.15	KEGG pathway	37
4.16	Reactome pathway	38
4.17	Manual per a aplicació	39
4.18	Manual per a l'anàlisi ORA amb l'anotació KEGG	39
4.19	Senyals d'ajuda	40
4.20	Ajuda per a l'elecció de l'espècie	41
4.21	Ajuda per pujar les dades	41
4.22	Infromació per la interpretació d'anàlisi ORA	42

Llista de les imatges

4.23 L'ajuda per a la selecció del mètode d'ajustament	42
4.24 Ajuda per la interpretació de GSEA	43
5.1 Selecció d'espècie	46
5.2 Selecció d'espècie	47
5.3 Resultat d'anàlisi ORA de Reactome	47
5.4 Gràfic de barres	48
5.5 Gràfic de punts	49
5.6 Mapa d'enriquement	50
5.7 Red de les categories i gens	51
5.8 Rutes Reactome	52
5.9 Anàlisi GSEA	53
5.10 Gràfic GSEA	53
5.11 Anàlisi ORA de KEGG	54
5.12 Gràfic de les rutes KEGG	54
5.13 L'anàlisi ORA de GO	55

1 Introducció

El treball consistirà en el desenvolupament d'una aplicació per dur a terme l'anàlisi de les rutes (*Pathway analysis*). Amb les rutes entenem un conjunt de gens que actuen junts per dur a terme un procès biològic. Així doncs aquesta anàlisi permet donar més sentit a una expressió genètica diferencial entre les proves biològiques d'interès. Recordem que recents avenços tecnològics permeten mesurar els nivells d'expressió en una gran quantitat de gens, cosa que implica una gran quantitat de dades. Al nivell dels gens individuals es poden fer servir mètodes estadístics per comprovar si les diferències en les expressions entre els grups (provees biològiques) són estadísticament significatives. Per dotar encara de més sentit aquesta anàlisi és necessari agregar els resultats al nivell més raonable com ara al nivell de les rutes. Al final el que volem és comprovar si hi ha diferències estadísticament significatives entre les provees no al nivell dels gens particulars sinó al nivell de les rutes. Tan com en el cas dels gens particulars també en el nivell de les rutes s'han desenvolupat mètodes estadístics específics [Khatri et al., 2012]. En aquest treball vull analitzar quins mètodes són i quins tenen més avantatges que d'altres. A part d'aquest component més biològic i teòric del treball he buscat la possibilitat d'implementar aquests mètodes d'anàlisi en una aplicació intuitiva i d'un ús fàcil a la qual qualsevol científic que no disposi dels coneixements informàtics suficients per fer aquesta anàlisi podrà accedir gratuïtament. La plataforma que he utilitzat per crear l'aplicació és l'eina Shiny de Rstudio [Chang et al., 2018]. La feina ha consistit en la cerca dels paquets de Bioconductor que inclouen els mètodes per l'anàlisi de les rutes, selecció dels paquets més apropiats i la seva integració en una aplicació Shiny amb una interfície atractiva.

La justificació d'aquest tema ve de dues fonts diferents: d'una banda tinc un interès personal i d'altra banda entenc la importància de la meva aportació per a la comunitat científica. El meu interès personal és degut al fet que

1 Introducció

durant el màster he fet servir àmpliament el programa R però no he arribat a conèixer bé la creació d'una aplicació estadística amb Shiny. Per completar aquesta deficiència i entenent que aquesta eina és útil per al meu desenvolupament professional he buscat el tema que en requeria l'ús. Tot i la importància de l'anàlisi de les rutes, al meu saber encara no existeix cap aplicació Shiny que integri paquets diversos i molt efectius de Bioconductor. L'ús d'aquests paquets requereix coneixements informàtics i estadístics específics i per tant és difícilment accesible per la gran part de la comunitat científica. Encara que hi ha ja plataformes gratuïtes que ofereixen l'anàlisi de les rutes [Reimand et al., 2019] crec que val la pena desenvolupar una eina més que seria de codi obert.

1.1 Objectius

Entre els objectius del treball podem distingir els general i els més específics:

1.1.1 Objectius generals

1. Identificar els objectius i mètodes de l'anàlisi de les rutes (Bio/Stat)
2. Identificar els paquets de Bioconductor en R que s'aproximin als mètodes (Info)
3. Desenvolupar l'aplicació Shiny amb els paquets escollits per aproximar el resultat als objectius de l'anàlisi de les rutes (Info)

1.1.2 Objectius específics

1. Biologia/Estadística
 - a) Buscar literatura sobre l'anàlisi de rutes
 - Quins mètodes hi ha? Enumerar-los i explicar-los, especialment els tests estadístics.
 - Quines bases de dades es fan servir?

1.1 Objectius

- Determinar les opcions per visualitzar els resultats de l'anàlisi de les rutes.
 - b) Identificar les aplicacions existents i investigar què ofereixen
 - c) Analitzar els vignettes dels paquets de Bioconductor i provar-ne el seu ús localment amb R
2. Informàtics
- a) Crear i documentat un protocol (pipeline) de l'anàlisi utilitzant els paquets seleccionats.
 - b) Identificar les dades experimentals per passar-les pel pipeline creat
 - c) Fer proves amb les dades seleccionades
 - d) Fer canvis en el protocol si és necessari
 - e) Integrar el pipeline a l'aplicació Shiny

Com es pot entendre dels objectius la feina ha consistit d'una banda en l'anàlisi teòrica dels mètodes disponibles actualment per a l'anàlisi de rutes, i d'altra banda en el desenvolupament d'una aplicació que incorporarà aquests mètodes. El mètode triat per aconseguir aquests objectius era el mètode simultani on la programació es desenvolupava alhora de l'anàlisi dels conceptes teòrics. D'aquesta manera he seguit aquests pasos:

1. Trobar un mètode teòric que proporcioni un resultat interessant;
2. Buscar en Bioconductor aquest mètode;
3. Repetir 1 i 2 fins que el conjunt dels mètodes facin l'anàlisi de les rutes complet.
4. Quan tots els mètodes són triats dissenyar un protocol;
5. Aplicar el protocol a les dades independents;
6. Comparar els resultats amb els estudis d'on provenen les dades;
7. Ajustat últimament el protocol;
8. Desenvolupar l'aplicació

S'ha d'emfatitzar el punt 5 i 6. Era essencial trobar les dades que s'utilitzin per fer les proves durant la fase de desenvolupament de *pipeline*. Les dades havien de provenir d'uns resultats ja publicats per poder comparar-los amb els resultats obtinguts amb el programari elaborat.

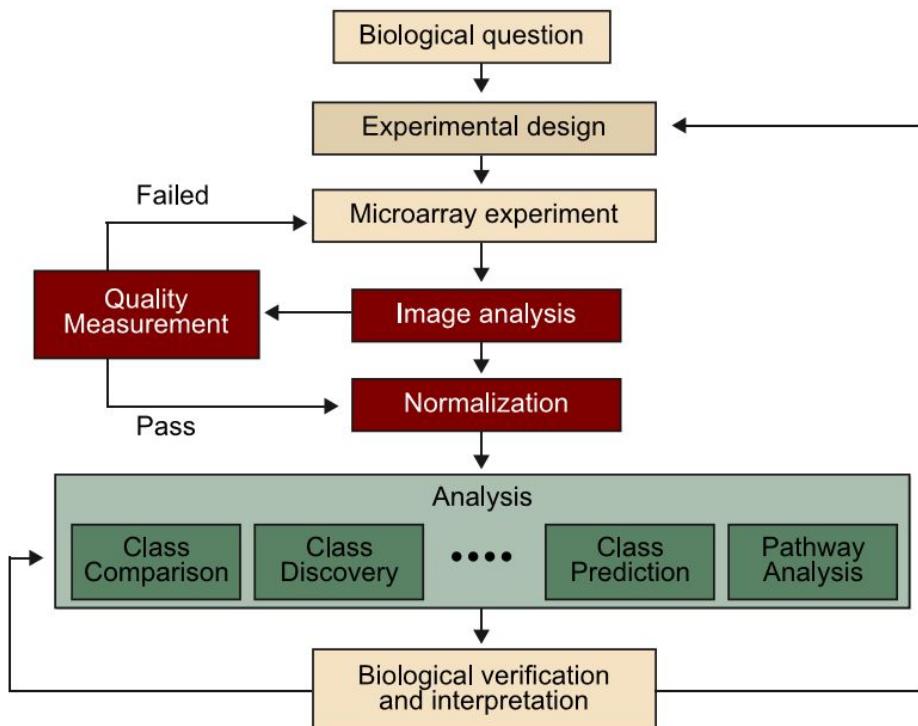
2 El marc teòric

2.1 Les dades d'expressió genètica

Les dades d'entrada per a l'anàlisi de les rutes provenen típicament de l'anàlisi de *microarrays* d'ADN, que produeix dades d'expressió de m gens (variables) per a n mARN mostres (observacions). Les dades com aquestes poden resultar d'un estudi d'investigació sobre efectes d'una proteïna com per exemple a l'estudi de [Li et al., 2017] on s'investiga la correlació entre la proteïna Zbtb7b (Zinc finger and BTB domain-containing protein 7B) i la formació de teixit adipós marró i beix i d'aquesta manera influex sobre fisiologia metabòlica. En aquest cas l'objectiu és comparar teixits de dos ratolins un de tipo salvatge i l'altre amb el gen ZBTB7B silenciat i investigar quins gens són diferencialment expressats entre aquests mostres biològiques.

Al gràfic següent veiem l'estructura habitual d'un experiment de *Microarray*.

2 El marc teòric



Imatge 2.1: El procès d'anàlisi de microarrays

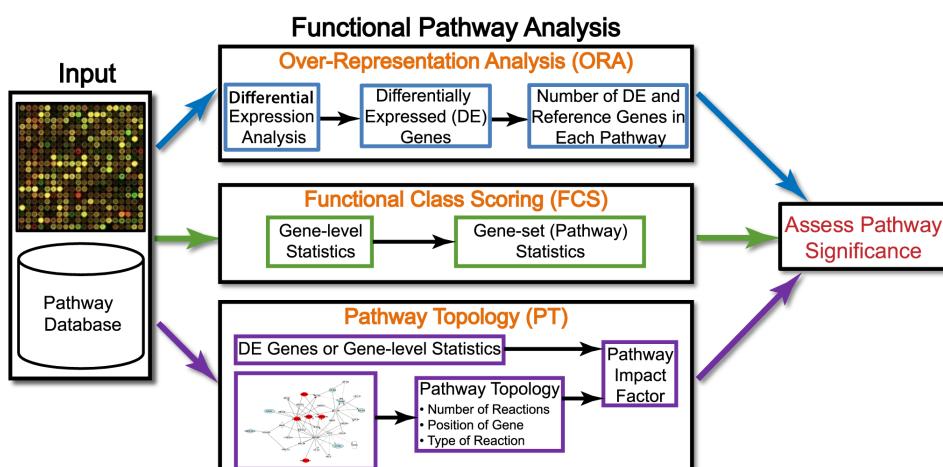
El *pipeline* de l'anàlisi consisteix doncs en plantejament d'una pregunta i un disseny experimental a partir del qual és fa l'experiment de *microarrays*. El producte d'experiment són bàsicament les imatges d'intensitats que es tradueixen als valors numèrics. Habitualment aquests valors són encara *raw values* i han de ser processats adequadament. Aquest processament inclou el control de qualitat d'imatges i la normalització dels valors d'intensitat per reduir la variabilitat tècnica. Finalment les dades normalitzades s'utilitzen per a l'anàlisi estadístic. Habitualment la mesura natural per comparar les mostres és el *log ratio* el qual podem denominar alternativament *logFC* on *FC* es refereix a *fold change*. Hi ha diversos tests estadístics per comprobar les diferències entre les mostres. En el cas de l'array d'un color podem fer servir tan els mètodes paramètrics com ara el test T o mètodes del modelatge lineal o bé mètodes nonparamètrics com ara la prova de Mann-Whitney.

2.1 Les dades d'expressió genètica

El test adequat dependrà bàsicament de la distribució de les dades. El resultat d'aquest ànalisi serveix com a base per a interpretació biològica dels resultats d'experiment. Per poder fer sentit de les dades d'expressió i de l'ànalisi estadístic al nivell de gens és imprescindible fer un ànalisi a nivell de les categories de gens o les rutes. Per aquest ànalisi es necessita, com ho veurem a l'apartat següent, una llista ordenada de les expressions relatives (*logFC*) i una subllista de gens que hem identificat mitjançant els tests estadístics com a diferencialment expressats.

En aquest treball m'ocupó de l'últim pas d'experiment d'escript, més específicament amb l'ànalisi de rutes (*Pathway analysis*).

La vista general de l'ànalisi de les rutes ofereix el gràfic següent:



Imatge 2.2: El procés d'anàlisi de les rutes

A part de les dades d'expressió, de les quals he parlat anteriorment, l'ànalisi requereix com a *input* també la base de dades de les rutes. De les dades que utilitzaré a la meva aplicació en parlaré a la secció següent. Per ara heis important entendre que els resultats d'expressió s'anoten a les bases de dades existents per comprobar si els gens sobre o sotaexpressats pertanyen a unes rutes específica. Per comprobar aquesta, o millor dit, aquestes hipòtesis (per que hi hauran hipòtesis múltiples) s'han establert tres grups dels mètodes:

2 El marc teòric

- **Over-Representation Analysis (ORA).** Aquest anàlisi necessita la preselecció dels gens diferencialment expressats (DE) i compara la freqüència dels gens de la ruta d'interès en la mostra dels gens diferencialment expressats i la freqüència dels gens de la ruta a la distribució de fons.
- **Functional Class Scoring (FCS)** Per a aquesta anàlisi no necessitem cap preselecció dels gens diferencialment expressats (DE) sinó ja basta amb tenir les estadístiques a nivell de gens, que al cas de l'aplicació és el *logFC*.
- **Pathway Topology (PT).** Aquest mètode enfoca la posició de gens diferencialment expressats en la ruta i d'aquesta manera utilitza el coneixement de les bases de dades més àmpliament. Per exemple, si una ruta està activada per un sol producte genètic o mitjançant un receptor i si aquesta proteïna particular no està produïda, la ruta estarà molt afectada, o fins i tot apagada. Més especificativament, si el receptor d'insulina no és en la ruta d'insulina (https://www.genome.jp/dbget-bin/www_bget?hsa04910) tota la ruta serà desactivada ([[Tarca et al., 2008](#)]). D'altra banda si un nombre de gens està involucrar en la ruta però apareixen riu abaux el seu efecte podria ser menys important. A més a més també el nombre de connexions amb altres gens a la ruta podria ser important [[Rahnenführer et al., 2004](#)]. O fins i tot les estadístiques que incorporen factors diferents com ara la posició, el tipus d'interacció etc. [[Draghici et al., 2007](#)]. Aquesta idea l'ha implementat en l'aplicació afegint les rutes dibuixades de KEGG i Reactome, on els gens estan emfatitzats d'acord amb els *logFCs* obtinguts mitjançant l'experiment.

Els mètodes interessants per a dur a terme l'anàlisi de les rutes són:

- ORA [[Boyle et al., 2004](#)]
- Mètodes GSA
 - Permutació de les mostres: GSEA [[Subramanian et al., 2005](#)], SAFE [[Dinu et al., 2007](#)] com els més representatius;
 - Permutació dels gens: PAGE [[Kim and Volsky, 2005](#)], T-Profiler [[Newton et al., 2007](#)] com els més representatius.
- GAGE (Generally Applicable Gene set Enrichment for pathway analysis) [[Luo et al., 2009](#)]

2.2 Anotació dels gens

Per fer l'aplicació més estructurada i menys complicada he elegit al final els dos mètodes: ORA [[Boyle et al., 2004](#)] i GSEA [[Subramanian et al., 2005](#)]. L'anàlisi GAGE seria un bon *Add-on* però per falta de temps per completar el TFM al final he decidit concentrar-me exclusivament en ORA i GSEA , els mètodes que descriure a les seccions següents.

2.2 Anotació dels gens

Com veurem més endavant per a analisi de les rutes és imprescindible tenir com a referència anotacions dels gens. Per a aplicació he utilitzat tres bases de dades: Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes i Reactome.

2.2.1 Gene ontology

Gene Ontology [[Consortium, 2004](#)] dona tan un vocabulari estructurat i controlat (ontologies) com la classificació que cobreix alguns dominis de la biologia molecular i cel·lular. És una base de dades gratuïta per a anotació de gens, el seu producte i les seqüències. El projecte GO proporciona ontologies per a descriure els atributs dels productes de gens als tres dominis separats de la biologia molecular:

1. **Molecular Function (MF).** Aquest domini descriu activitats al nivell molecular. És important entendré que el terme “molecular function” representa més les activitats i no pas les entitats (com per exemple molècules o complexos) que fan aquestes accions i a més a més no espeifíquen quan o a quin context l’acció té lloc. Un exemple podria ser *catalytic activity* o un terme més específic *adenylate cyclase activity*.
2. **Biological Process (BP).** Aquest domini descriu els objectius biològics aconseguits per una o conjunt de les funcions moleculars. Un exemple d'un procès biològic ampli podria ser *DNA repair*. Un exemple més específic podria ser *pyrimidine nucleobase biosynthetic process*.

2 El marc teòric

3. **Cellular Component (CC).** EL CC descriu l'emplaçaments al nivell d'estructures subcel·lulars (com *mitocondri*) i els complexos macromoleculars (com *ribosomes*) on el producte de gen fa la seva funció.

Dins de cada ontologia, els termens tenen tan una definició de text com un identificador únic. El vocabulari està estructurat en una classificació que manté les relacions “is-a” i “part-of” i “regulates”. Aquestes relacions les descriu amb més detall més endavant en la secció dedicada al gràfic acíclic de GO termes.

2.2.2 KEGG

La base de dades KEGG és la col·lecció del mapes dibuixades manualment que representen el coneixement sobre interacció molecular dividit en set dominis principals:

1. Metabolism
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases
7. Drug Development

Els mapes són dibuixades amb un software específí (KegSketch) que genera un arxiu KGML+. Aquest arxiu és un arxiu SVG que conté els objectes gràfics que són associats amb els objectes KEGG. Els objectes gràfics bàsics de les rutes KEGG són:

- caixes: gens o el seu producte
- cercles: altres molècules
- línies: reaccions

El significat més detallat d'aquests elements el presentaré a la secció dedicada a les rutes KEGG.

2.3 ORA

2.2.3 Reactome

Reactome és una base de dades gratuitament accèssible i manualment curada per a reaccions i rutes biològiques. Al centre de Reactome són reaccions que es definexen com qualsevol esdeveniment molecular com ara unió, fosforilització, catàlisi bioquèmic, transport molecular o esdeveniments moleculars espontàni. Aquestes reaccions involucren qualsevol molècula, però més típicament passen entre proteïnes i les molècules petites. Encara que els mapes de Reactome disponibles online contenen una relació entre les molècules més detallada el paquete de Bioconductor que utilitzaré per generar els mapes visualitza només la conecció bàsica entre els gens.

2.3 ORA

L'anàlisi de sobreexpressió és una tècnica d'identificació de les rutes significativament enriquitides en la mostra d'interès.

El paper original que se cita habitualment quan es parla d'anàlisi d'expressió genètica és de [Boyle et al., 2004]. 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 *log ratios* i/o en el valor de p provenint d'un test estadístic. *Log ratios* donen la magnitud amb la qual un gen és sobre o sotaexpressat. Les diferències entre els grups però són el resultat d'un procés estochàstic i per tant hem d'intentar de minimitzar el risc de prendre decisions falses. El valor de p representa la probabilitat d'aquest risc i per tant dona certa confiança sobre la significació de les diferències observades.

2. **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 en la distribució hipergeomètrica:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

2 El marc teòric

En aquesta equació N és el nombre total de gens en la distribució de fons, M és el nombre de gens dins d'aquesta distribució que són anotats a la ruta d'interès, n és el nombre total en la llista especificada de gens i k és el nombre de gens dins d'aquesta llista que són anotats a la ruta. La distribució de fons pot ser o bé tots els gens en la base de dades d'anotació o bé tots els gens de l'experiment.

El valor de P obtingut amb aquesta fórmula dona la probabilitat de veure el nombre x de gens de la llista relacionats amb la ruta específica en la llista del nombre total de gens n donat la proporció de gens relacionats amb aquesta ruta en la distibucó de fons.

L'aplicació utilitza aquesta idea i calcula una taula amb els camps següents:

- Description. El nom del terme GO;
- GeneRatio. El quocient $\frac{M}{N}$ on M és el nombre dels gens diferencialment expressats que pertanyen al conjunt de gens i N és el nombre total dels gens diferencialment expressats
- BgRatio. El quocient: $\frac{k}{n}$ on k és el nombre dels gens del conjunt d'interès en la distribució de fon i n és el nombre total dels gens en la distribució de fons;
- pvalue. Valor de p basat en la distribució hipergeomètrica descrita anteriorment.
- p.adjust. El valor de P ajustat. L'usuari pot seleccionar el mètode d'ajustament.

2.4 GSEA

Amb l'anàlisi GSEA 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 L el podem crear utilitzant els *logRatios*.

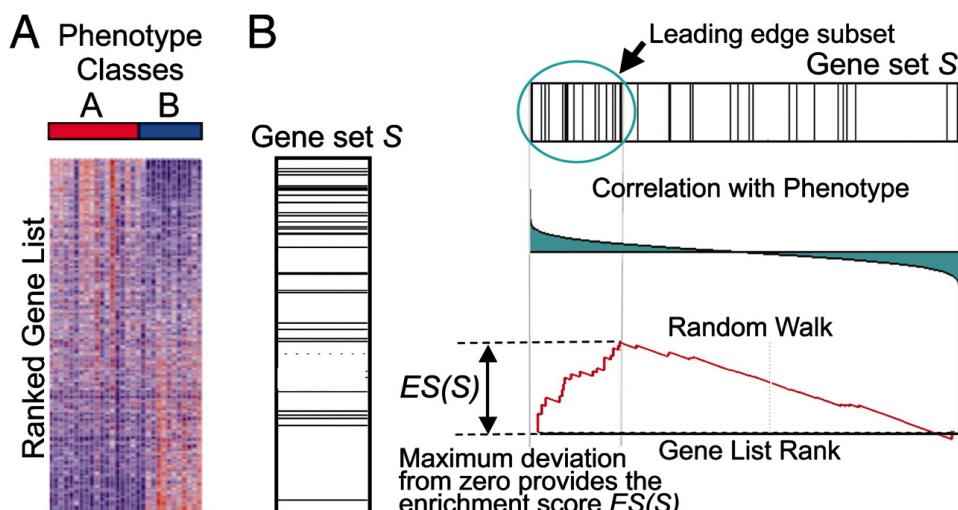
Donat el conjunt definit dels gens S , que pertanyen per exemple al mateix terme de Gene Ontology, l'objectiu de GSEA és determinar si els membres de S són distribuïts aleatoriament en el L o es troben més al cap o a la cua.

2.4 GSEA

S'esperaria que els gens relacionats amb la separació fenotípica mostraran aquesta última distribució.

L'anàlisi GSEA consisteix en tres passos:

1. Càlcul de la puntuació d'enriquiment (*ES: Enrichment Score*). La puntuació està calculada anant per la llista i augmentant la suma corrent sempre quan es troba un gen que pertany a S o, al contrari, restant-la quan el gen no forma part del conjunt S . La puntuació és la desviació màxima del zero observada en aquest camí. L'estadística obtinguda és l'estadística de Kolmogorov-Smirnov amb pesos.
2. Estimació del nivell de significació per a la puntuació *ES*. El valor de P nominal es pot obtenir mitjançant o bé la permutació de les classes o bé la permutació de gens, on l'estadística *ES* observada es compara amb la distribució obtinguda amb permutació. A l'aplicació es fa ús de l'última opció.
3. Càlcul del valor de P ajustat. El valor de P nominal s'ajusta per controlar l'error global que es produeix com a resultat de les comparacions múltiples.



Imatge 2.3: El mètode GSEA

L'aplicació que he desenvolupat agafa aquesta idea i calcula la taula que inclou les estadístiques següents:

2 El marc teòric

- enrichmentScore. Enrichment score per al conjunt dels gens. Amb altres paraules: el grau amb el qual el conjunt dels gens està sobre-expressat a dalt o a baix del llistat ordenat dels gens en les dades d'expressió.
- NES. Normalized enrichment score. La puntuació per al conjunt de gens després de ser normalitzat tenint en compte tots els conjunts de gens analitzats (la seva mida i la seva correlació amb les dades d'expressió). Aquesta puntuació ajuda a comparar els resultats entre els conjunts de gens.
- pvalue. El valor de p nominal.
- p.adjust. El valor de p ajustat.
- leading_edge
 - Tags. El percentatge de les ocurrències de gens del conjunt específic abans (per als ES positius) o després (per als ES negatius) del cim en la puntuació corrent d'enriquiment. Aquest valor indica el percentatge dels gens que contribueixen a la puntuació d'enriquiment.
 - List. El percentatge dels gens en el llistat ordenat de tots els gens abans o després del pic en la puntuació corrent d'enriquiment. Aquest valor ens indica on exactament el pic es produeix.
 - Signal. La fortalesa del senyal d'enriquiment que combina els dos valors anteriors.
- rank. La posició del pic en la llista ordenada dels gens. Els conjunts dels gens més interessants assoleixen el seu màxim o bé al principi o al final de la llista ordenada. Vol dir que tenen aquest valor o bé molt baix o bé molt alt.

2.5 L'anàlisi topològic de les rutes

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 (GO, KEGG i Reactome) contenen però aquesta informació i l'aplicació, gràcies al paquet `clusterProfiler`, hi treu l'avantatge i visualitza aquestes relacions més detalladament.

2.5 L'anàlisi topològic de les rutes

2.5.1 El mapa d'enriquement

Navegant a la categoria **Enrichment plot** l'usuari obté el mapa d'enriquiment. El mapa organitza les rutes en una xarxa amb les línies que connecten les rutes amb els gens solpats. D'aquesta manera les rutes amb gens en comú s'agrupen més a prop l'una de l'altra.

2.5.2 Gene-Concept-Network

L'anàlisi ORA no visualitza per si sola els gens que contribueixen al fet que la ruta sigui diferencialment expressada. Amb la xarxa de gens-concepte es pretén visualitzar els gens al voltant dels conceptes on els gens poden ser connectats amb les rutes (conceptes) diferents. D'aquesta manera es fa possible identificar les associacions biològiques més complexes entre les rutes mitjançant els gens.

2.5.3 GOplot

El gràfic de GO està organitzat com direccional acíclic gràfic (Directed Acyclic Graph). Una manera útil de veure els resultats és mirar com els termes GO estan distribuïts per aquest gràfic. L'aplicació ensenya el gràfic GO induït pels gens més significatius. El gràfic mostra tres relacions possibles entre les rutes:

1. *is a*: Si dèiem que A *is a* B, volem dir que A és un subtíp de B. Per exemple el cicle mitòtic de la cèl·lula *is a* cicle de la cèl·lula.
2. *part of*: Aquesta relació s'utilitza per representar la relació entre una part i el tot. Aquesta relació entre A i B existeix nomès si B és necessàriament una part d'A: quan B existeix, ho fa només com una part de B i la presència de B implica la presència d'A.
3. *regulates*: La relació descriu el cas on un procès afecta directament la manifestació de l'altre procès.

Els conceptes al llarg del gràfic són marcats amb color depenent si són estadísticament significatius o no.

2 El marc teòric

2.5.4 KEGG Pathway

Aquest gràfic mostra les relacions entre els gens dins de la ruta específica. Els gens són remarcats amb el color dependent de l'expressió diferencial mesurada amb LogRatios. Per poder interpretar el gràfic és útil tenir present l'anotació següent:

Notation	Objects	Arrows
	gene product, mostly protein but including RNA	→ molecular interaction or relation
○	chemical compound, DNA and other molecule	→ link to/from another map
[]	map	---> indirect link or unknown reaction /→ missing interaction (eg., by mutation) → drug structure link or pointer used to add legend
	Protein-protein interactions	Gene expression relations
+p	phosphorylation	[] → ○ → [] expression
-p	dephosphorylation	[] → ○ ← [] repression
+u	ubiquitination	[] → c → [] expression
-u	deubiquitination	[] → c ← [] repression
+g	glycosylation	
+m	methylation	
	activation	
	inhibition	
---	indirect effect or state change	
	binding / association	
	dissociation	
	complex	
		Enzyme-enzyme relations
		[] → ○ → [] → two successive reaction steps

Imatge 2.4: L'anotació de les relacions dins de les rutes KEGG

2.5 L'anàlisi topològic de les rutes

2.5.5 Reactome Pathway

En canvi a Goplot i les rutes KEGG les relacions entre els gens dins les rutes Reactome són més senzilles. Aquí les relacions són mostrades només amb les línies, on es pot interpretar només la distància entre els gens.

3 Tractament bioinformàtic

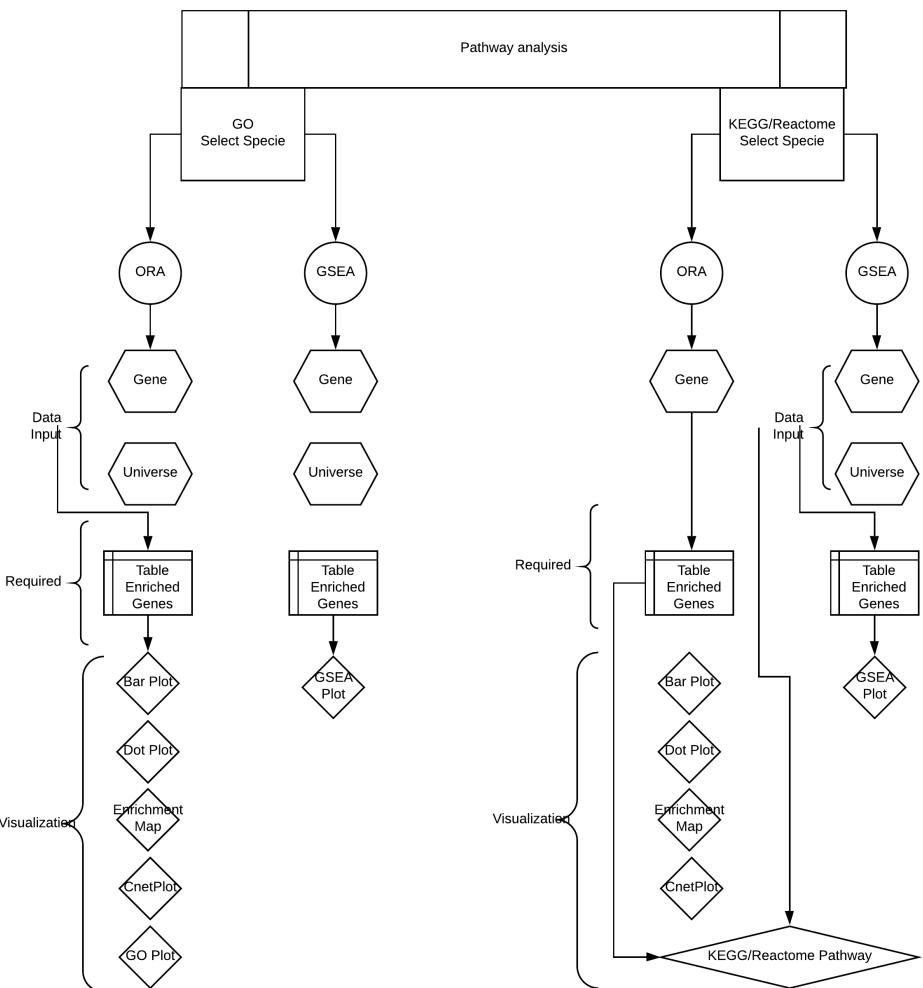
3.1 Cerca dels paquets de Bioconductor

El paquet clusterProfiler de Bioconductor integra els mètodes per dur a terme l'anàlisi de les rutes basant-se en les bases de dades GO, KEGG i Reactome. Els dos mètodes principals són ORA (Overrepresentation analysis) i GSEA (Gene set enrichment Analysis). També inclou les possibilitats de visualització dels resultats suficients per considerar l'anàlisi de les rutes complet. Notem però que el test de permutació a l'anàlisi GSEA implementat per clusterPrifiler es basa en la permutació dels gens i no de les mostres com originalment és proposat per [Subramanian et al., 2005].

3.2 Desenvolupament del protocol

Crec que l'aplicació és molt intuitiva i deixa entreveure l'esquema següent:

3 Tractament bioinformàtic



Imatge 3.1: Lucidchart per a l'aplicació

D'aquí podem definir per exemple el protocol:

1. Decidir quin ànalisi vol fer: GO, KEGG o Reactome
2. Seleccionar l'espècie de referència
3. Decidir quin mètode vol implementar: ORA o GSEA i respectivament pujar les dades necessàries.

3.3 Instal·lació de l'aplicació

- Per a l'anàlisi GO tots dos arxius són necessaris: Gens Seleccions (Gene) i Tots els gens (Universe).
 - Per a l'anàlisi KEGG o Reactome les dades necessàries varien: Pel mètode ORA l'arxiu amb els gens seleccionats és suficient. Dos arxius són necessaris pel mètode GSEA.
4. En el cas que volguem fer l'anàlisis ORA seleccionar a l'apartat corresponent (GO, KEGG o Reactome) la pestanya ORA i definir els criteris.
 - Els gràfics: Bar Plot, Dot Plot Enrichment Map, Cnet Plot, GO Plot (en cas d'anàlisi GO) i els gràfics de les rutes (KEGG/Reactome) es calculen automàticament
 5. En el cas que volguem fer l'anàlisi GSEA seleccionar a l'apartat corresponent (GO, KEGG o Reactome) la pestanya GSEA i definir els criteris.
 - El gràfic GSEA es genera automàticament. Es pot elegir la ruta mitjançant un menú desplegable.

3.3 Instal·lació de l'aplicació

La solució més plausible i ràpida era empaquetar tota l'aplicació dins d'un paquet R i fer-la disponible d'aquesta manera en el GitHub. Hi havia també dues opcions més:

- Publicar l'aplicació a CRAN
- Publicar l'aplicació en un servidor Shiny

La primera opció, publicació en CRAN, no l'he contemplat encara, perquè la solució no és immediata, sino que és un procès que no és fàcil i pot tardar fins que el paquet estigui publicat amb èxit. Com comenta [Wickham, 2015] “submitting to CRAN is a lot more work than just providing a version on github, but the vast majority of R users do not install packages from github, because CRAN provides discoverability, ease of installation and a stamp of authenticity. The CRAN submission process can be frustrating, but it's worthwhile...”. Normalment els paquets han d'estar en perfectes condicions abans d'entregar-los i seran revisats manualmet per un equip

3 Tractament bioinformàtic

dels voluntaris. D'aquesta manera l'aplicació no seria avaluable dins del marc temporal previst per al treball de màster. A més a més considero que podria millorar encara més l'aplicació abans d'entregar-lo.

La segona opció, publicació via Shiny Server, és molt interessant, però implicaria un treball considerable per configurar el servidor. Com que ho faria per primera vegada, no puc assegurar que tot estigui preparat a temps.

Per tant, el paquet PathwayApp es pot instal·lar del repositori GitHub seguint els passos següents:

1. Instal·lar, si encara no està fet, la versió actual de R;
2. Instal·lar, si encara no està fet, el Bioconductor;
3. Instal·lar, si encara no està fet, el paquet devtools

```
install.packages('devtools')
library(devtools)
```

4. Instal·lar el paquet PathwayApp

```
devtools::install_github("vdruchkiv/TFM/5_Packages/PathwayApp")
```

5. Iniciar l'aplicació

```
PathwayApp::runPathwayApp()
```

La funció runPathwayApp() iniciarà la comprovació dels paquets necessaris i començarà l'aplicació. Els paquets següents seran instal·lats, si no ho són encara:

3.3 Instal·lació de l'aplicació

Paquet	Font
clusterProfiler	Bioconductor
ReactomePA	Bioconductor
pathview	Bioconductor
pathviewPatched	GitHub vdruchkiv/TFM
dplyr	CRAN
ggplot2	CRAN
knitr	CRAN
kableExtra	CRAN
formattable	CRAN
shiny	CRAN
shinydashboard	CRAN
shinyhelper	CRAN
shinycssloaders	CRAN

4 L'aplicació

4.1 ORA

4.1.1 GO

Per realitzar l'anàlisi ORA per a termes GO s'utilitza la funció `enrichGO` del paquet `clusterProfiler`.

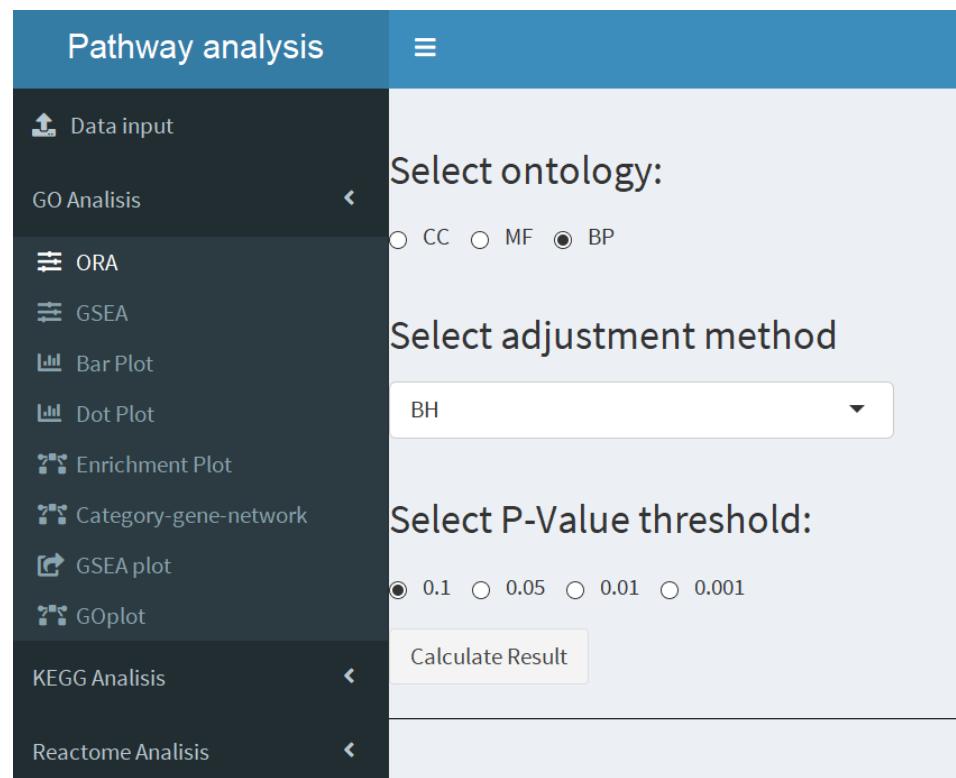
He implementat els valors per defecte amb la possibilitat per a l'usuari d'elegir entre:

- Ontologies GO
 - Molecular function, Biological proces, Cellular Components;
- Nivell de significació basant-se en els valors de P ajustats
 - 0.1, 0.05, 0.01, 0.001;
- Mètode d'ajustament
 - Holm; Hochberg; Hommel; Bonferroni; BH; BY; FDR; None.

L'execució de la funció és un procès temporalment costós. Per aquest motiu he afegit el botó d'acció, en lloc de deixar la funció reactiva. D'aquesta manera l'usuari ha de fer una decisió conscient de repetir l'anàlisi amb altres valors.

Prement el botó apareix la taula i el botó nou mitjançant el qual l'usuari pot descarregar els resultats en format .csv. He formatejat la taula amb els paquets `knitr`, `kableExtra`, `formattable` i `dplyr`. Amb els dos últims he afegit les barres de color pel nombre dels gens diferencialment expressats del terme específic de GO i la gradació de color del verd fins al vermell pels valors dels més petits fins els més grans.

4 L'aplicació



Imatge 4.1: Especificació d'ORA dels termes GO

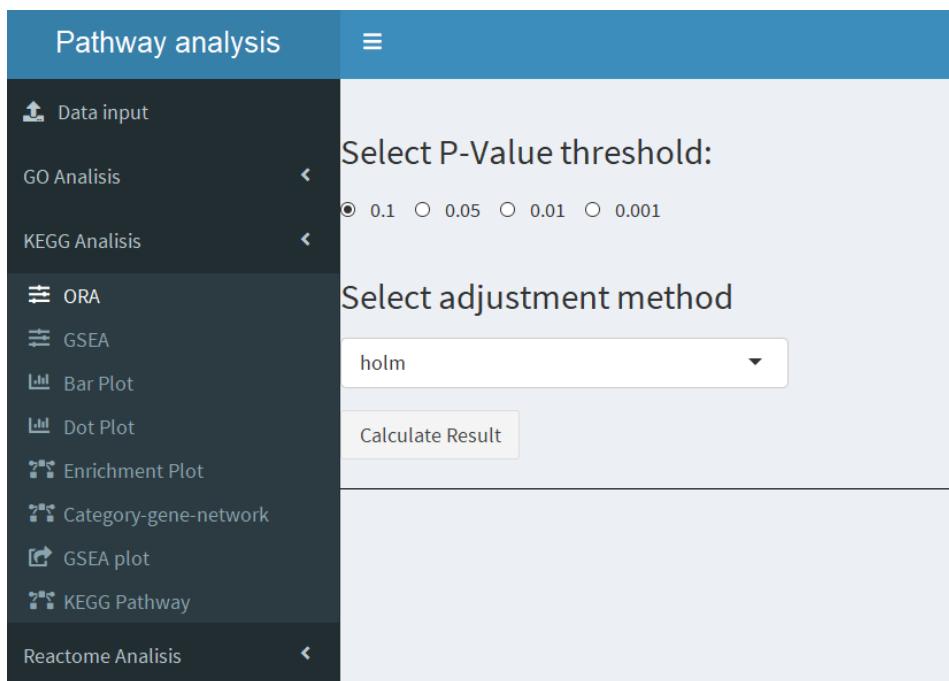
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	geneID
GO:0140014	mitotic nuclear division	33/193	232/11468	0.000	4.00e-18	0.000	33	CDCA8/CDC20/KIF23/CENPE/MYBL2/CCNB2/t
GO:00000280	nuclear division	35/193	316/11468	0.000	4.50e-16	0.000	35	CDCA8/CDC20/KIF23/CENPE/MYBL2/CCNB2/t

Imatge 4.2: El resultat d'anàlisi ORA. GO.

4.1.2 KEGG

Per l'ORA de base de dades KEGG he utilitzat la funció `enrichKEGG()` del paquet `clusterProfiler`.

4.1 ORA



Imatge 4.3: Configuració d'anàlisi KEGG

Una vegada introduïts els paràmetres i premut el botó **Calculate** apareix el botó **Download .csv** i la taula previsualitzada. Els camps de la taula són els mateixos com en l'anàlisi dels termes GO.

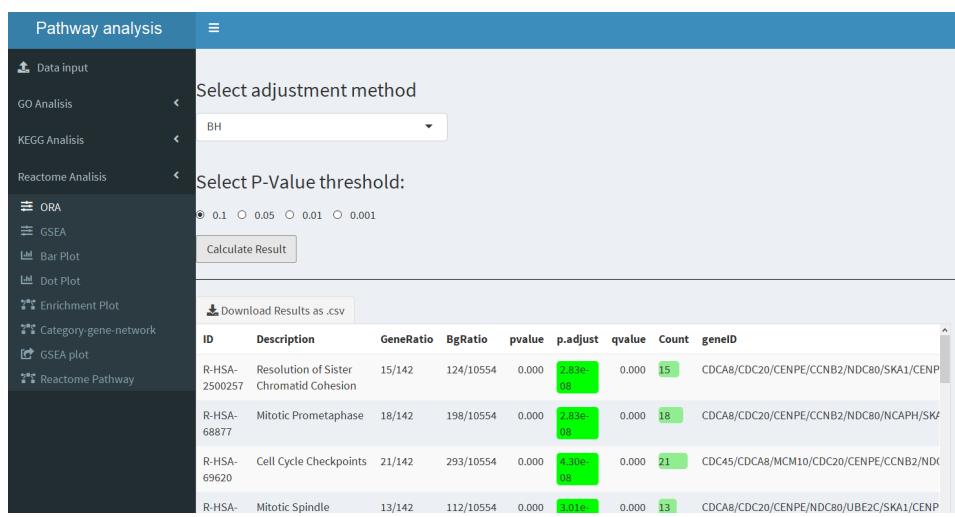
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	genelD
hsa04110	Cell cycle	11/92	124/7841	0.000	3.48e-05	0.000	11	8318/991/9133/890/983/4085/7272/1111/891/4174/9232
hsa04114	Oocyte meiosis	10/92	125/7841	0.000	1.70e-04	0.000	10	991/9133/983/4085/51806/6790/891/9232/3708/5241
hsa04218	Cellular senescence	10/92	160/7841	0.000	1.04e-03	0.001	10	2305/4605/9133/890/983/51806/1111/891/776/3708

Imatge 4.4: El resultat de l'anàlisi ORA. KEGG.

4 L'aplicació

Reactome

En el cas de Reactome el procediment és similar. La funció usada és `enrich-Pathway()` del paquet `ReactomePA`:



Imatge 4.5: El resultat d'anàlisi ORA. Reactome.

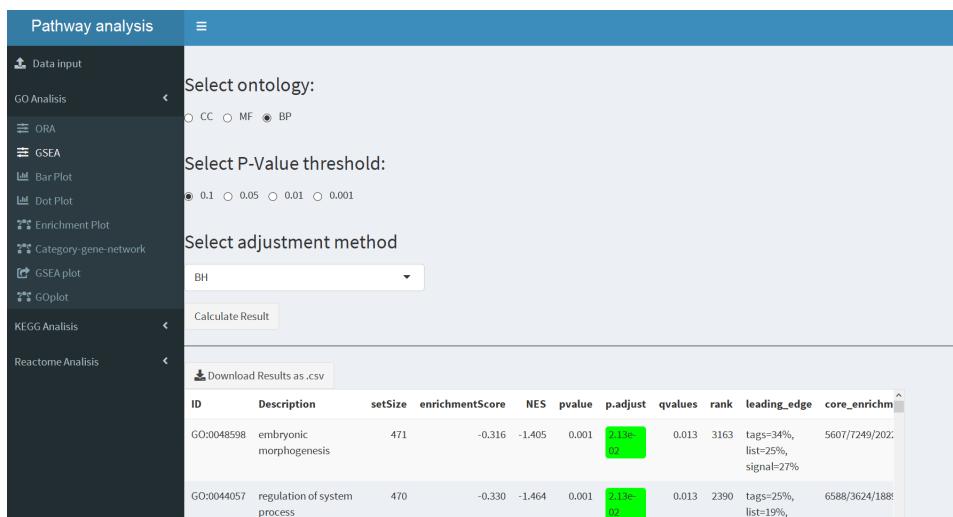
4.2 GSEA

4.2.1 GO

El mètode GSEA per a termes GO es calcula amb la funció `gseGO()` del paquet `clusterProfiler`.

L'usuari pot elegir l'ontologia GO, el *cut-off* del valor P i el mètode d'ajustament.

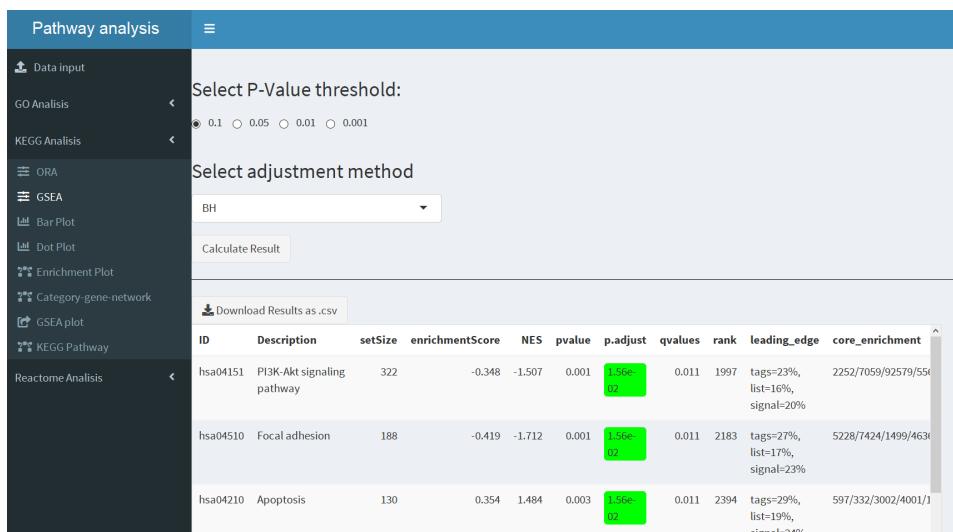
4.2 GSEA



Imatge 4.6: El resultat de l'anàlisi GSEA. GO.

4.2.2 KEGG

De la mateixa manera es calcula GSEA amb la funció `gseKEGG()` del paquet `clusterProfiler`:

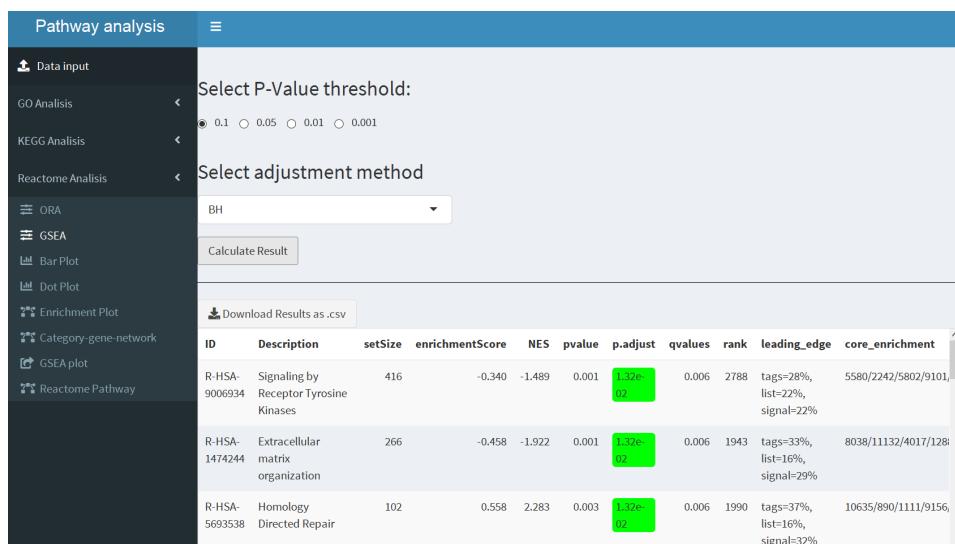


Imatge 4.7: El resultat de l'anàlisi GSEA. KEGG.

4 L'aplicació

4.2.3 Reactome

Per completar l'anàlisi l'usuari pot calcular GSEA per a base de dades Reactome. Com als altres casos utilitzo el paquet `clusterProfiler` i específicament la funció `gsePathway()`

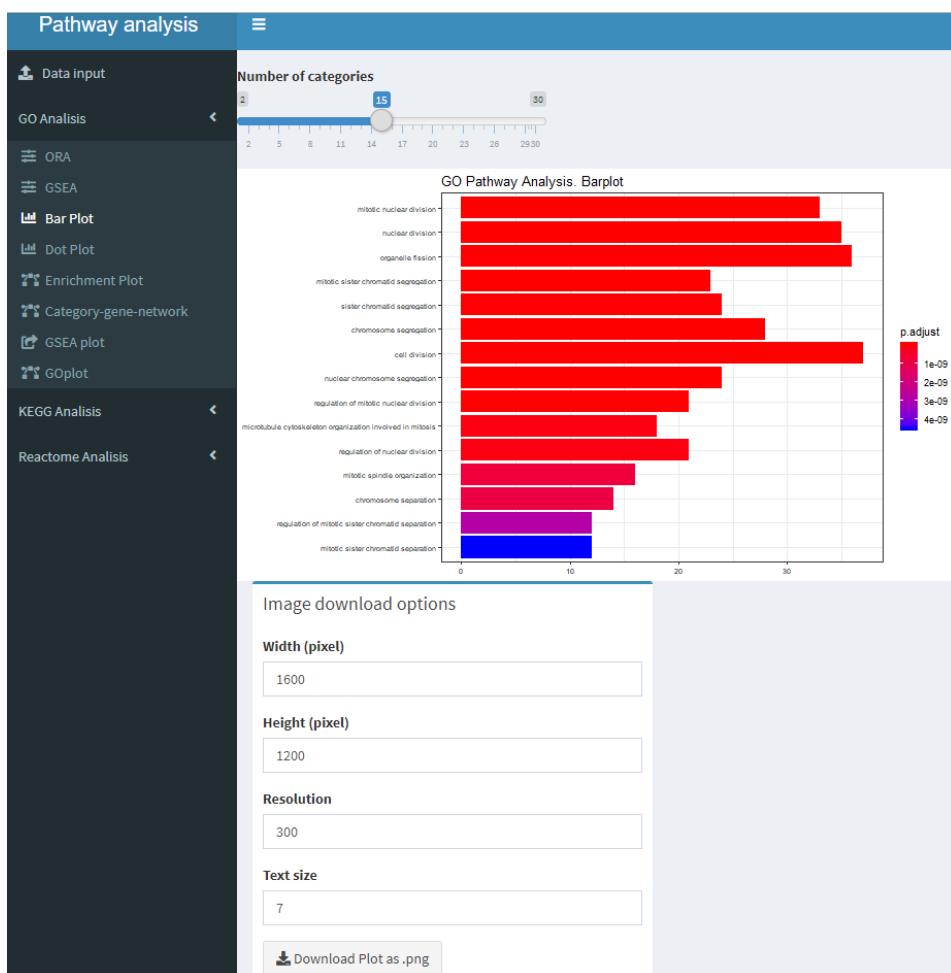


Imatge 4.8: El resultat d'anàlisi GSEA. Reactome.

4.2.4 Bar-Plots

Els resultats de `enrichGO`, `enrichKEGG` i `enrichPathway` es poden visualitzar amb el gràfic de barres. L'usuari pot elegir el nombre de les categories visualitzades entre 2 i 30. Es dona l'opció per descarregar el gràfic en format `.png`.

4.2 GSEA

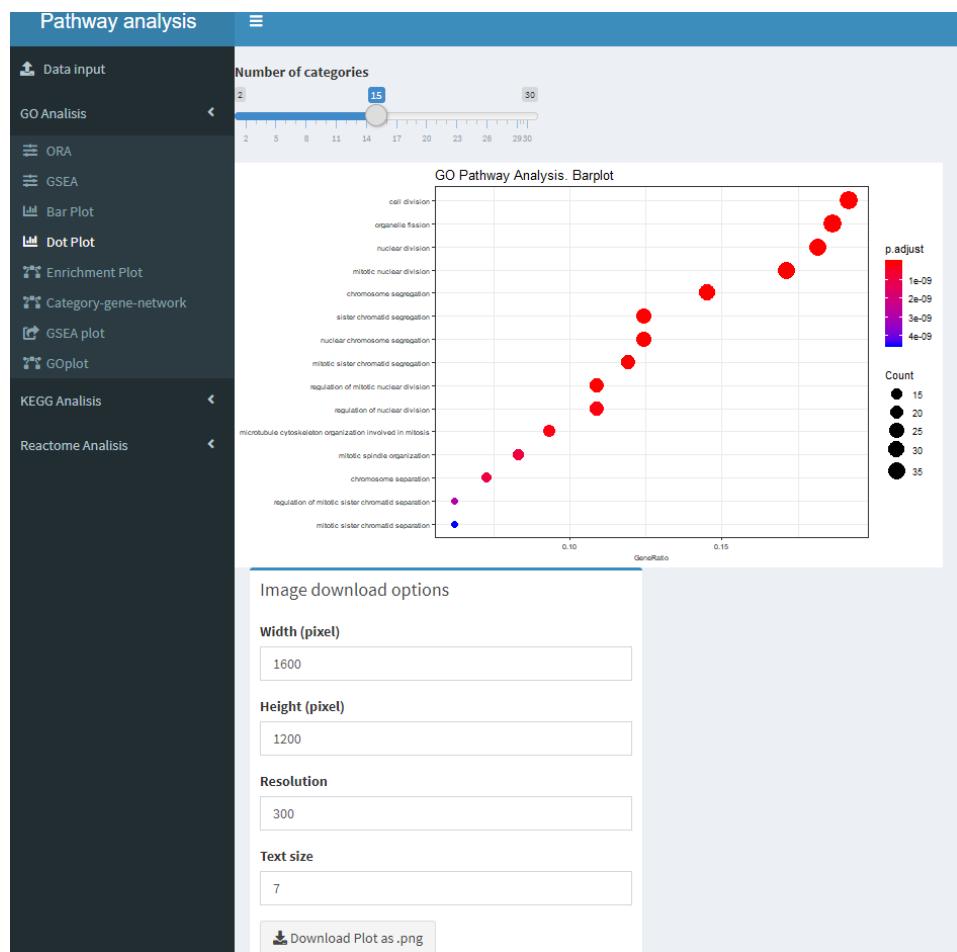


Imatge 4.9: Bar-Plot. GO.

4.2.5 Dot-Plots

El *dot plot* visualitza addicionalment el *gen ratio*. També aquí l'usuari pot seleccionar el nombre de categories.

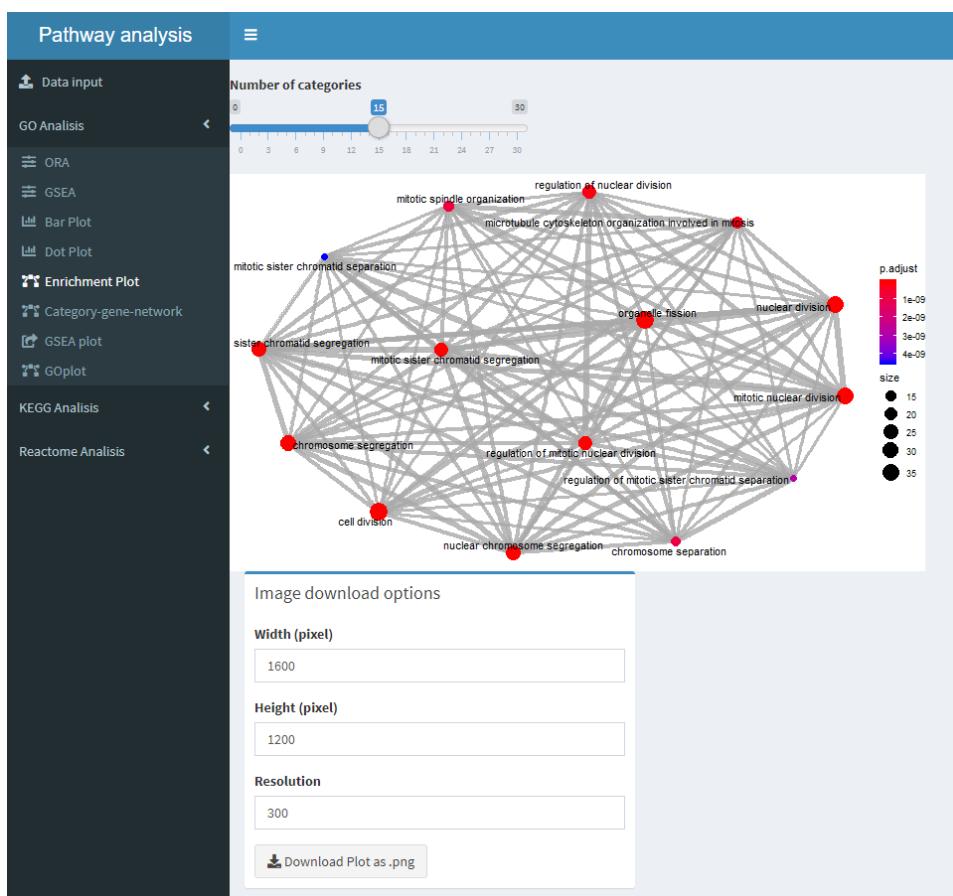
4 L'aplicació



Imatge 4.10: Bar-Plot. GO.

4.2 GSEA

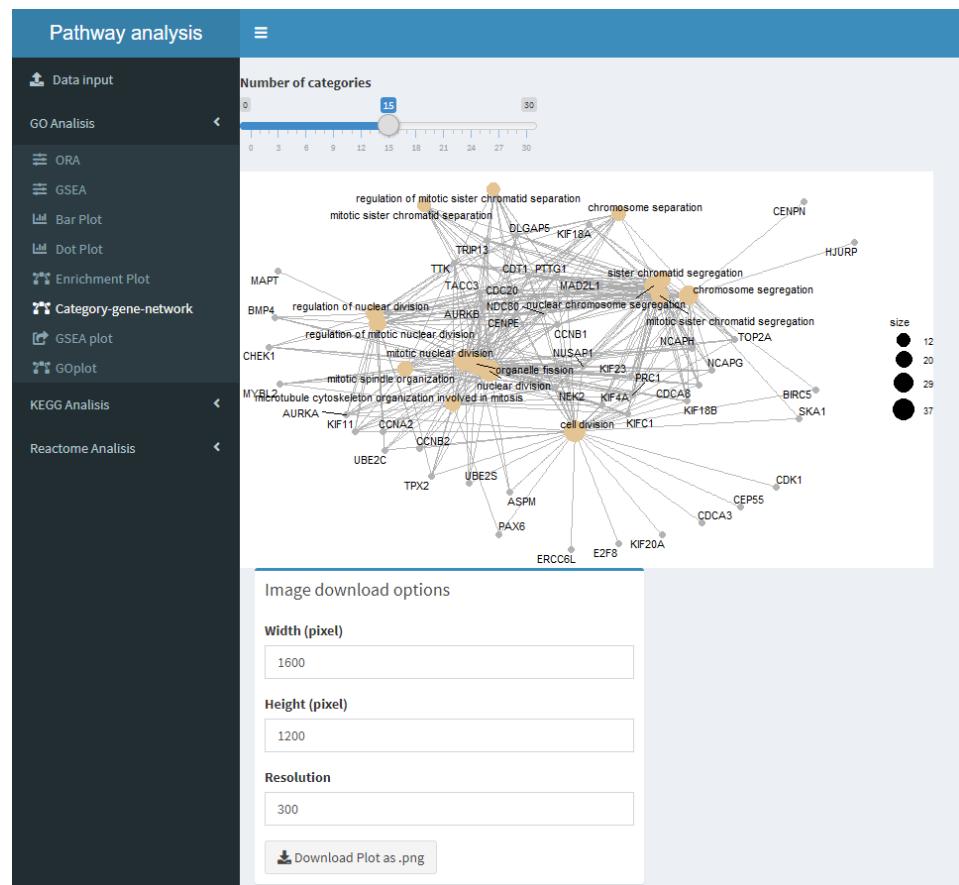
4.2.6 Enrichment Plots



Imatge 4.11: Bar-Plot. GO.

4 L'aplicació

4.2.7 Category-Gene-Network Plot

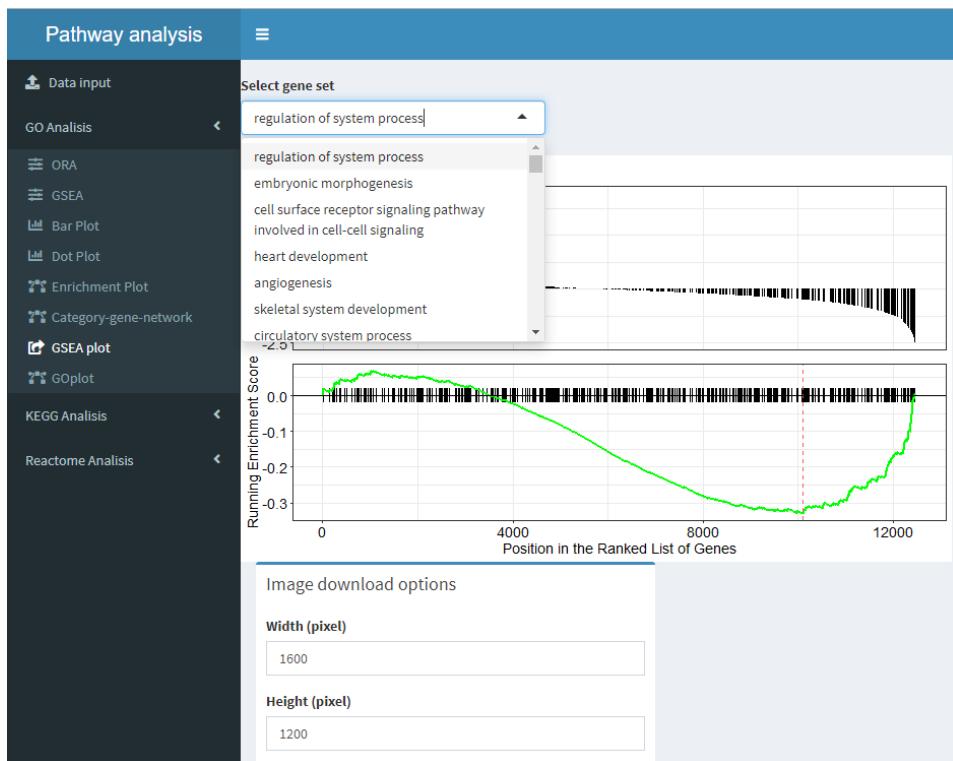


Imatge 4.12: Category-Gene-Network Plot. GO.

4.2.8 GSEA Plot

L'usuari pot visualitzar una de les categories disponibles via *dropdown list*. El llistat inclou totes les rutes generades durant l'anàlisi GSEA en els apartats *Go Analysis*→*GSEA*; *KEGG*→*GSEA*

4.2 GSEA

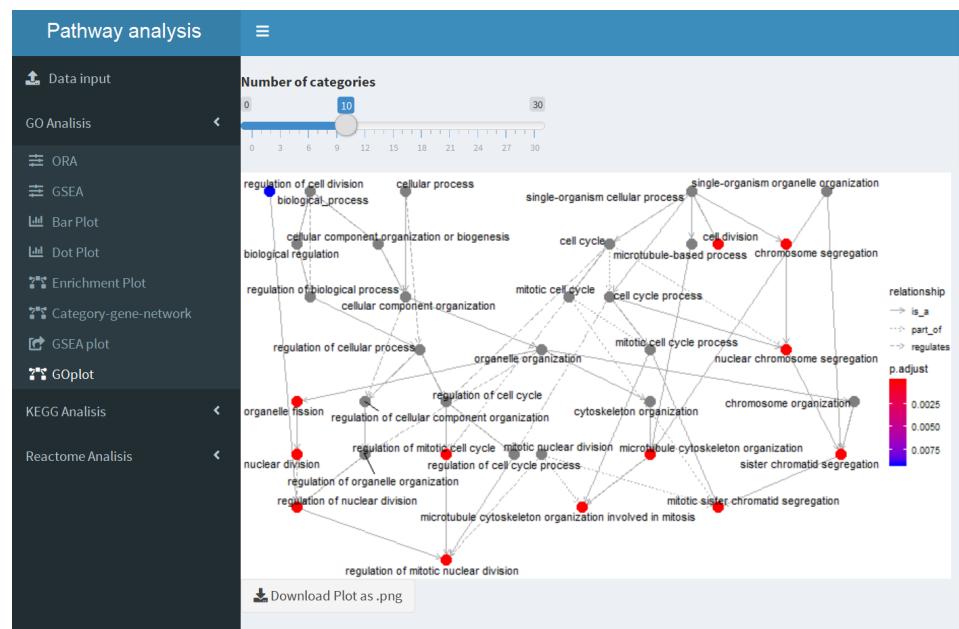


imatge 4.13: GSEA Plot. GO.

4 L'aplicació

4.3 L'anàlisi específic de GO, KEGG i Reactome

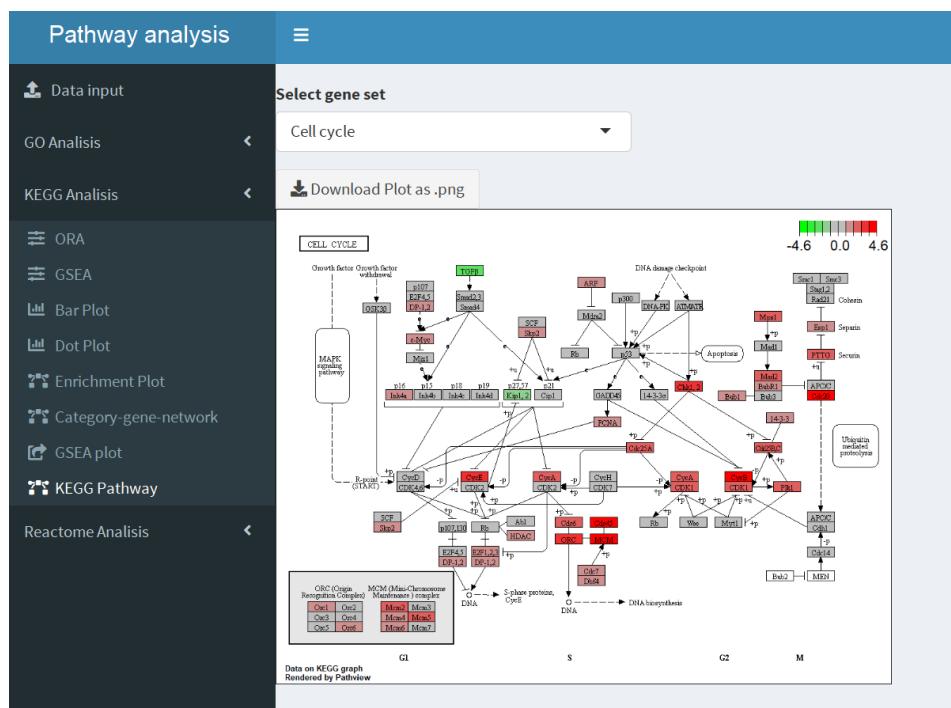
4.3.1 GO Plot



imatge 4.14: GO Plot

4.3 L'anàlisi específic de GO, KEGG i Reactome

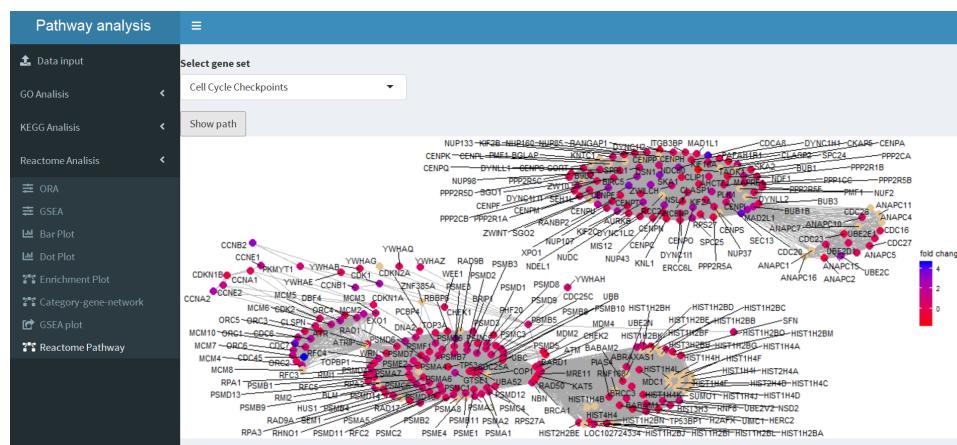
4.3.2 KEGG Pathway



imatge 4.15: KEGG pathway

4 L'aplicació

4.3.3 Reactome Pathway

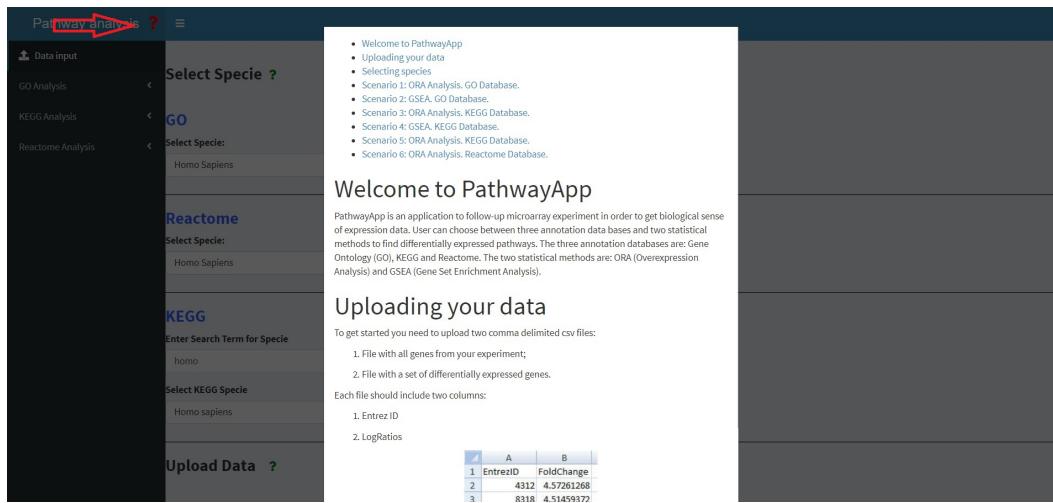


imatge 4.16: Reactome pathway

4.4 Manual i les ajudes del programa

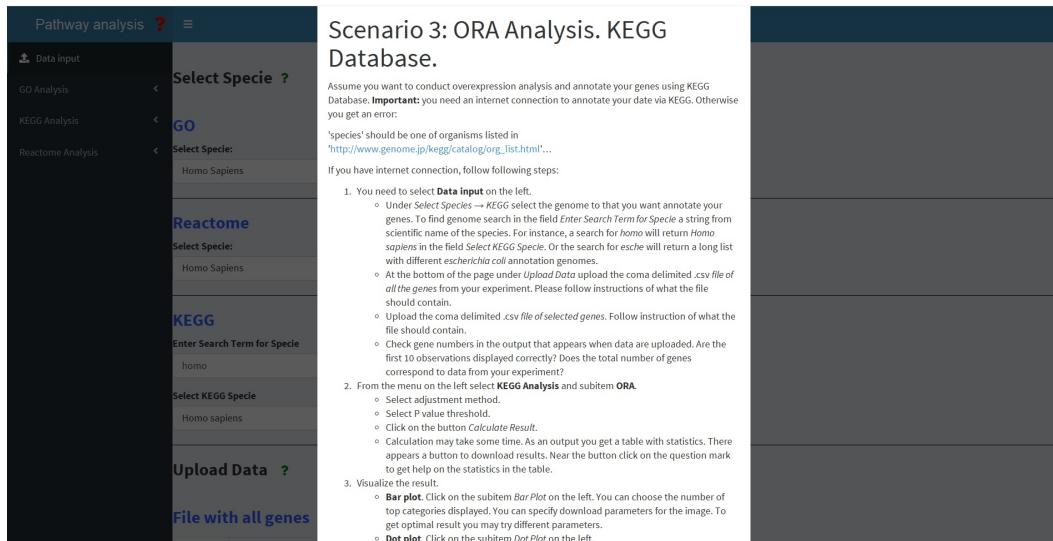
Per facilitar l'ús de l'aplicació he pensat com es podria fer de manera més intuïtiva possible. Primer cal destacar que com a llengua de manual he elegit l'anglès per poder fer l'ús de l'aplicació el més inclusiu possible. Segon, l'usuari pot accedir tant al manual com a l'ajuda, que es guarden en arxius .Md separats. Per accedir al manual l'usuari ha de clicar al símbol d'interrogació a prop del títol **Pathway analysis**:

4.4 Manual i les ajudes del programa



Imatge 4.17: Manual per a aplicació

Com es veu hi ha apartats diferents. Depenen dels objectius de l'usuari, aquest pot seleccionar l'apartat que més l'interessi. Així, si l'usuari vol fer l'anàlisi ORA amb l'anotació KEGG pot navegar en la secció —textbf{Scenario 3: ORA Analysis. KEGG Database}.

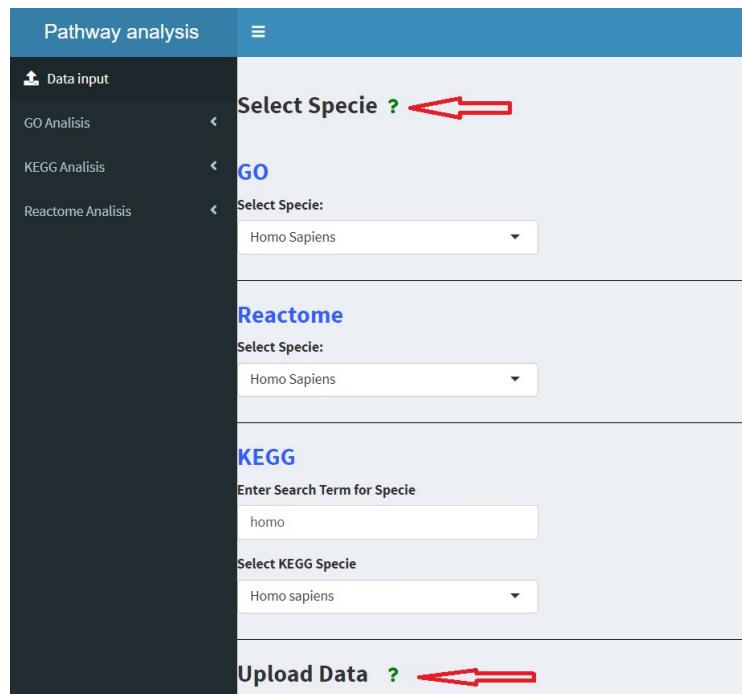


Imatge 4.18: Manual per a l'anàlisi ORA amb l'anotació KEGG

4 L'aplicació

També, l'usuari pot accedir a l'ajuda clicant als símbols d'interrogació distribuïts per l'aplicació en els llocs que penso que poden generar dubtes.

Per fer-ho possible s'utilitza el paquet `shinyhelper` que s'instal·la en executar la funció `runPathwayApp()`.

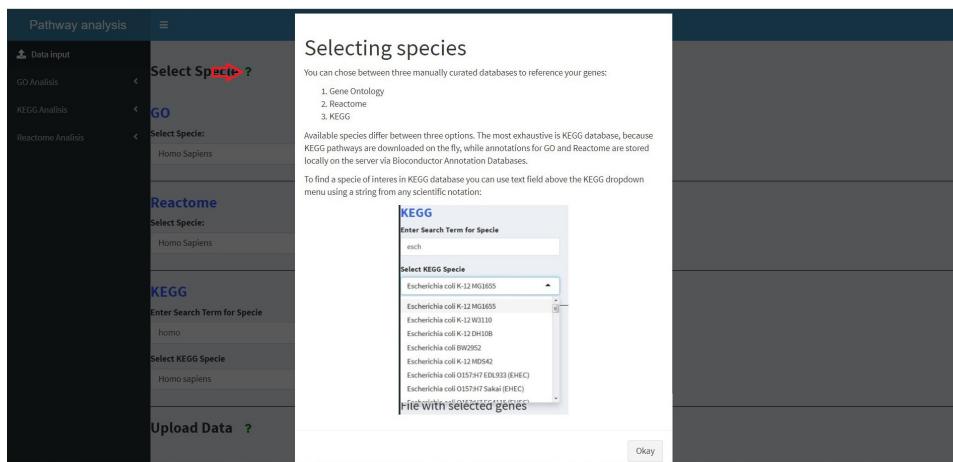


imatge 4.19: Senyals d'ajuda

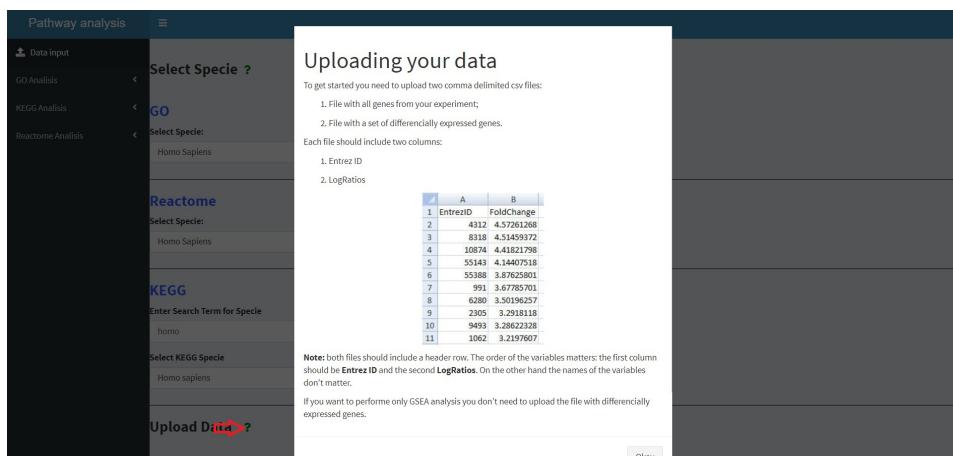
El clic en aquests senyals fa que aparegui una finestreta amb la informació d'ajuda.

Aquí hi ha informació de l'apartat **Data Input**:

4.4 Manual i les ajudes del programa



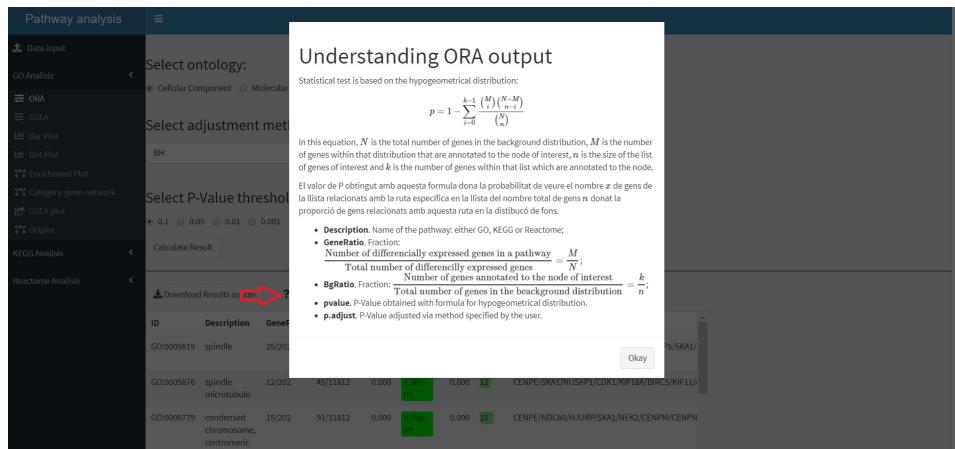
Imatge 4.20: Ajuda per a l'elecció de l'espècie



Imatge 4.21: Ajuda per pujar les dades

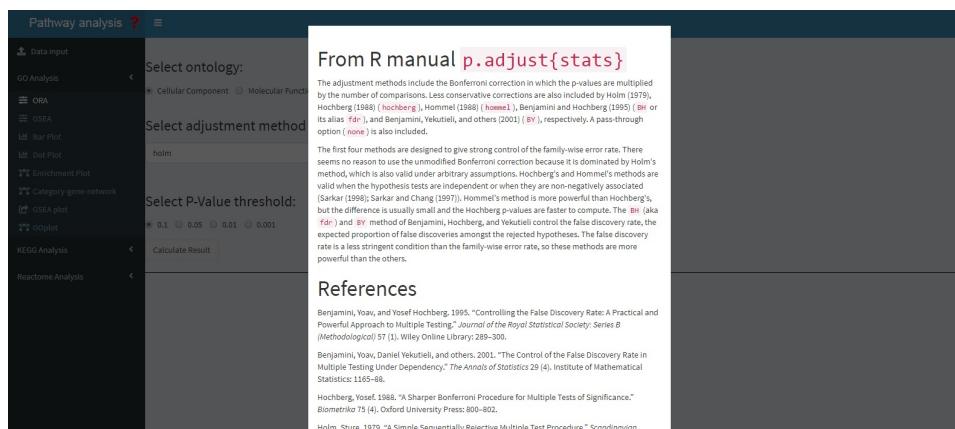
Les informacions per a l'apartat ORA són les següents:

4 L'aplicació



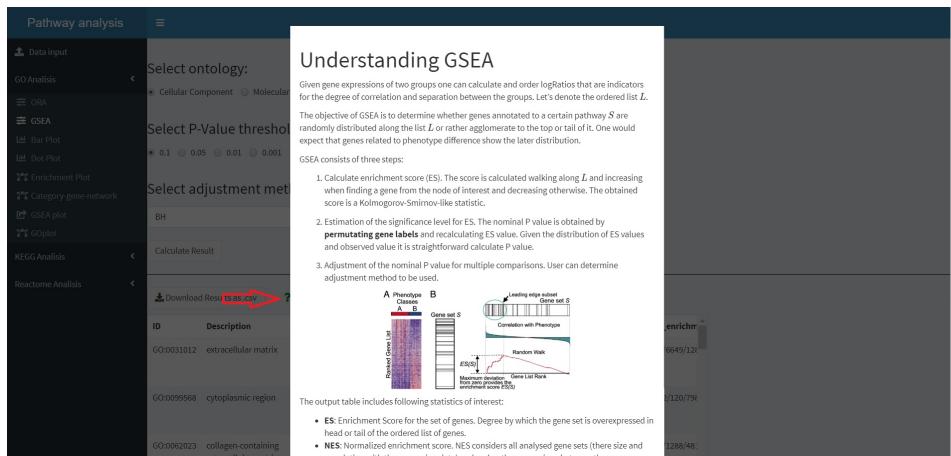
Imatge 4.22: Infromació per la interpretació d'anàlisi ORA

Aquí cal destacar que les fòrmules, depenent de l'ordinador, no apareixen degudament en el RStudio Browser. Sí que apareixen bé quan l'aplicació s'obre via l'internet browser. L'usuari ha de tenir connexió amb internet perquè l'aplicació pugui descodificar la fòrmula via MathJax. Encara no he trobat la causa per la qual el Rstudio Browser en alguns ordinadors no visualitza bé les fòrmules. Pot ser un problema amb Java, que s'ha d'actualitzar? Ho estic investigant.



Imatge 4.23: L'ajuda per a la selecció del mètode d'ajustament

4.4 Manual i les ajudes del programa



Imatge 4.24: Ajuda per la interpretació de GSEA

5 Validació dels resultats

L'anàlisi de les rutes representa l'últim pas de l'anàlisi d'expressions. Per dur a terme l'anàlisi de rutes és necessari tenir unes dades que ja estiguin processades prèviament (normalització, càcul de les LogRatios, ajustament dels gens repetits a l'array, selecció dels gens diferencialment expressats, etc.). Les dades de [GEO \(Gene Expression Omnibus\)](#) estan però disponibles com a màxim en format normalitzat. Caldria doncs fer una anàlisi per arribar a un llistat de gens diferencialment expressats amb les logRatios per tots els gens de la mostra. Fer això no seria cap problema i de fet ho he fet per altres estudis. El problema és que arribo a resultats diferents dels resultats dels estudis d'on provenen les dades (i no parlo de l'anàlisi de les rutes sinó ja del càcul de les logRatios). Per tant les dades que entraria a l'aplicació serien diferents de les dades de l'estudi i lògicament amb aquesta comprovació no comprovo el que realment m'interessa. Podria, doncs, dedicar-me a trobar el motiu pel qual els resultats són diferents, però fer totes aquestes comprovacions prèvies no té a veure amb l'objecte del meu treball de màster, l'anàlisi de les rutes. Per tant he procedit a contactar el meu professor per si tindria (o coneixeria) dades preprocessades fins a un llistat de gens amb logRatios i amb el set de gens diferencialment expressats, per tal que les pugui utilitzar en la meva aplicació. El meu professor m'ha redirigit, entre altres enllaços molt útils, al seu repositori en [github.com](#).

Estudi	GEO ID	Espècie	Tipo d'experiment	Font
[Schmidt et al., 2008]	GSE11121	Homo sapiens	Microarrays	Paquet DOSE de Bioconductor
[Li et al., 2017]	GSE100924	Mus musculus	Microarrays	Github Sanchez Pla
[Farmer et al., 2005]	GSE1561	Homo sapiens	Microarrays	Github Sanchez Pla
[Hengel et al., 2003]	DAVID Demo List 1	Homo sapiens	Microarrays	DAVID

Les dades de [\[Schmidt et al., 2008\]](#), que s'utilitzen en els vignettes de clusterProfiler i ReactomePA, ja les he mostrat en gran part a dalt quan explicava el contingut de l'aplicació. Els resultats obtinguts amb l'aplicació són iguals

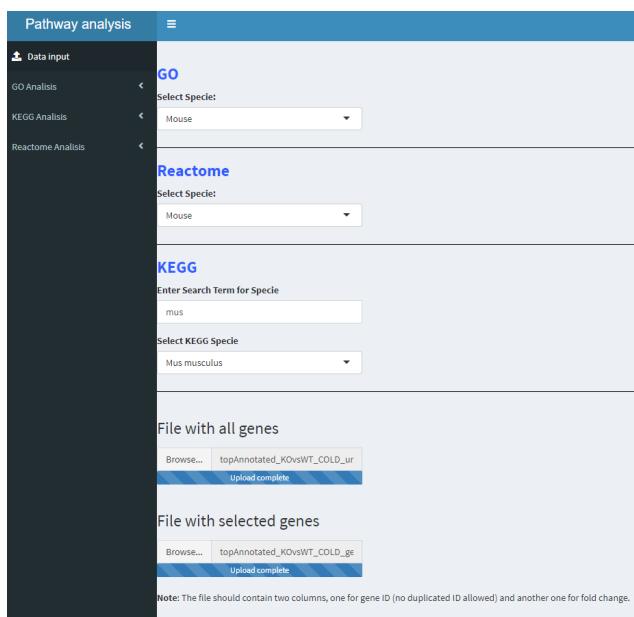
5 Validació dels resultats

als resultats en els vignettes mencionats. Procediré doncs amb l'exemple basat en les dades de [Li et al., 2017] .

5.1 Exemple d'anàlisi 1. GEO: GSE100924

Les dades d'estudi [Li et al., 2017] 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.

1. Ellegeixo l'espècie *Mus musculus* per a GO, KEGG i Reactome.



imatge 5.1: Selecció d'espècie

L'output a baix indica que s'ha pujat el total de 5995 gens. Per a l'arxiu dels gens seleccionats l'aplicació diu que s'han pujat 769 gens.

5.1 Exemple d'anàlisi 1. GEO: GSE100924

You uploaded: 5995 genes
First 10 entries

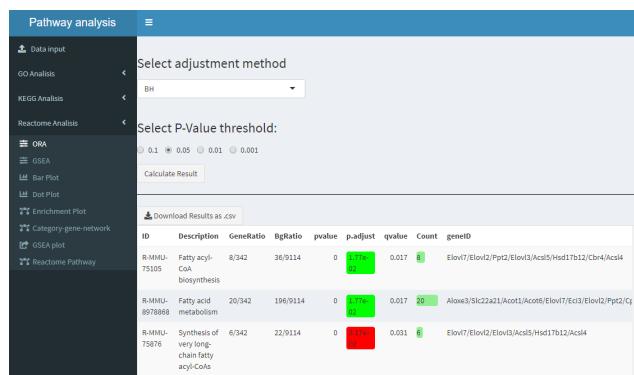
Entrez ID	FoldChange
108864	-0.420
319263	0.049
59014	-0.143
109294	0.114
320492	-1.454
98711	0.072
17087	-0.653
75712	-0.384
14859	0.378
27993	-0.113

You selected: 769 genes
First 10 entries

Entrez ID	FoldChange
320492	-1.454
50785	0.743

Imatge 5.2: Selecció d'espècie

- Clico en l'apartat *Reactome Analysis*→*ORA*. Selecciono com a mètode d'ajustament *BH* i el cut-off del valor de P ajustat 0.05. Clico a *Calculate results*



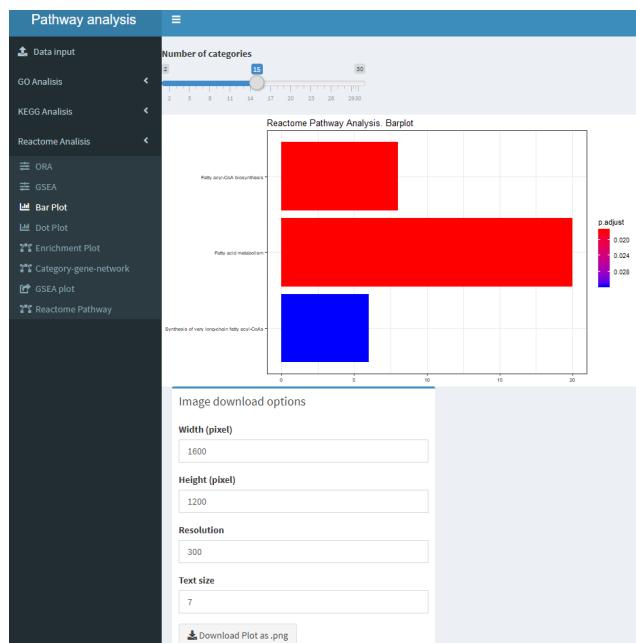
Imatge 5.3: Resultat d'anàlisi ORA de Reactome

Observem que els gens mostrats són els mateixos esmentats per Sanz i Pla.

- Visualització del resultat ORA

- Seleccio *Reactome Analysis*→*Bar Plot*

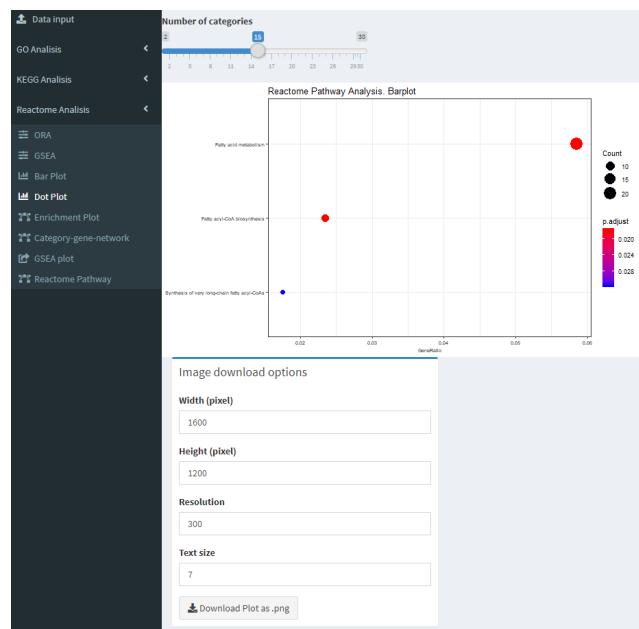
5 Validació dels resultats



Imatge 5.4: Gràfic de barres

- Selecciono *Reactome Analysis* → *Dot Plot*

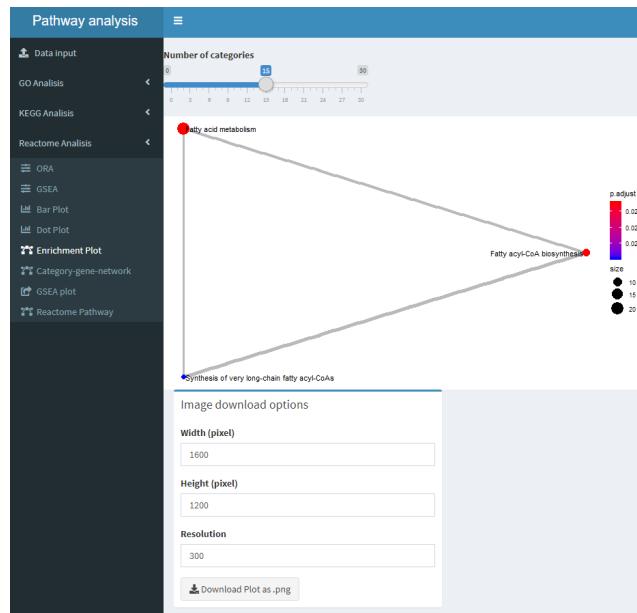
5.1 Exemple d'anàlisi 1. GEO: GSE100924



Imatge 5.5: Gràfic de punts

- Seleccions *Reactome Analysis* → *Enrichment Map Plot*

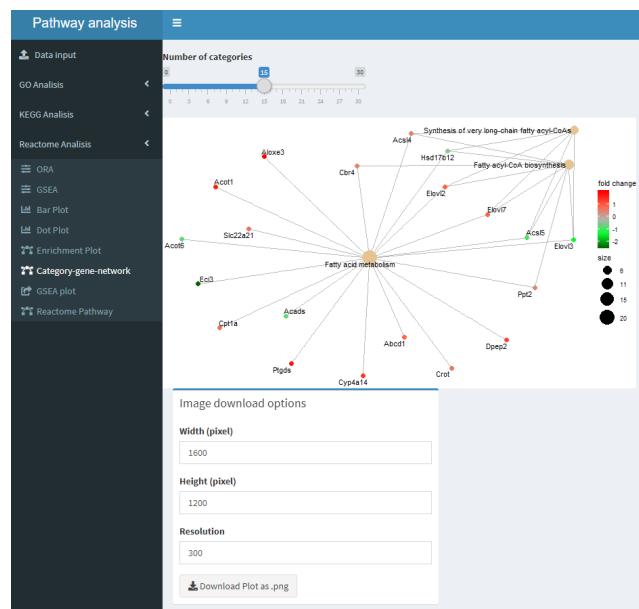
5 Validació dels resultats



Imatge 5.6: Mapa d'enriquement

- Selecciono *Reactome Analysis* → *Category Gene Network*

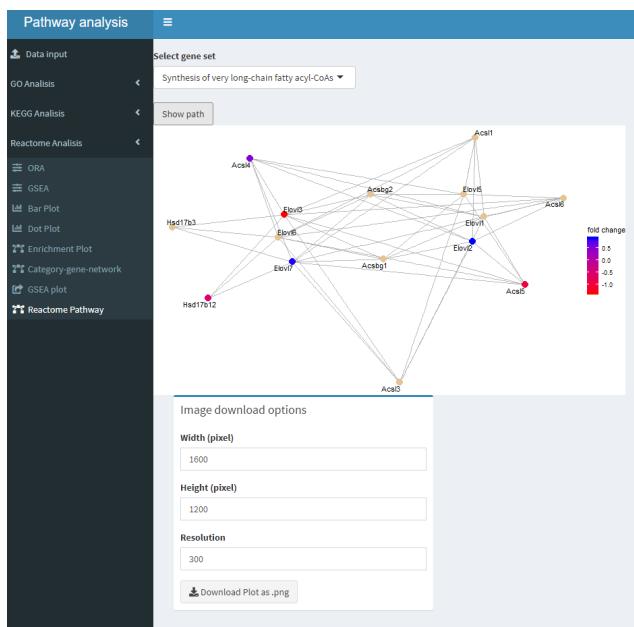
5.1 Exemple d'anàlisi 1. GEO: GSE100924



Imatge 5.7: Red de les categories i gens

- Seleccions *Reactome Analysis* → *Reactome Pathway*

5 Validació dels resultats

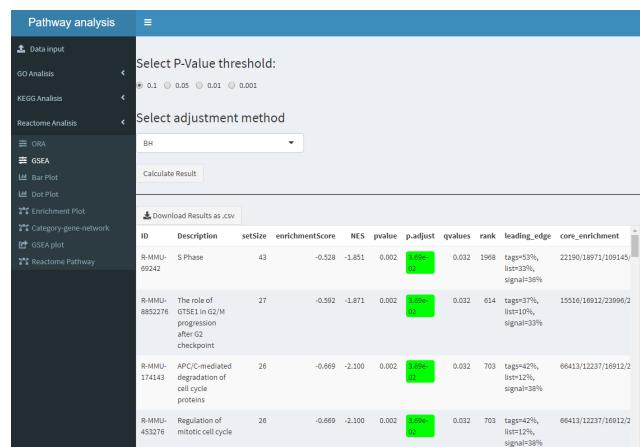


Imatge 5.8: Rutes Reactome

Addicionalment a l'anàlisi ORA podem fer, mitjançant l'aplicació, l'anàlisi GSEA per les rutes de Reactome. Per fer-ho:

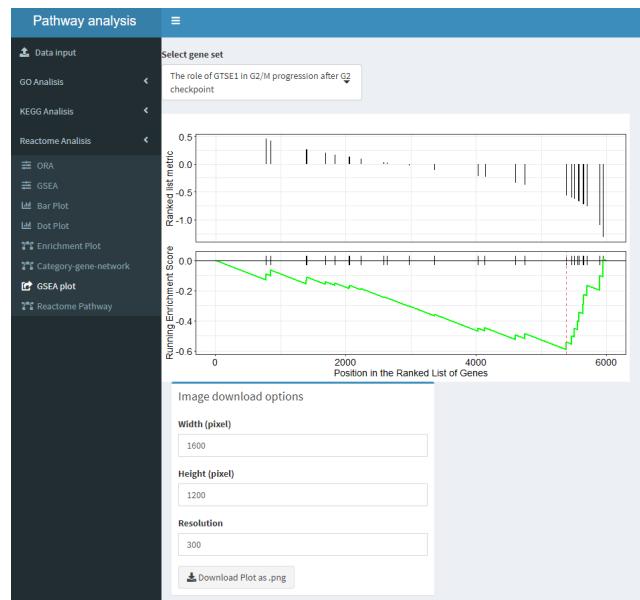
1. Clico en l'apartat *Reactome Analysis* → *GSEA*. Selecciono com a mètode d'ajustament *BH* i el cut-off del valor de P ajustat 0.05. Clico a *Calculate results*
Amb el valor de P de 0.05 l'anàlisi no troba cap ruta enriquida.
2. Augmento el Cut-Off del valor de P a 0.1
Amb el Cut-Off més alt l'aplicació retorna un llistat de gens.

5.1 Exemple d'anàlisi 1. GEO: GSE100924



imatge 5.9: Anàlisi GSEA

3. Per obtenir els gràfics GSEA anem a *Reactome Analysis*→*GSEA plot*

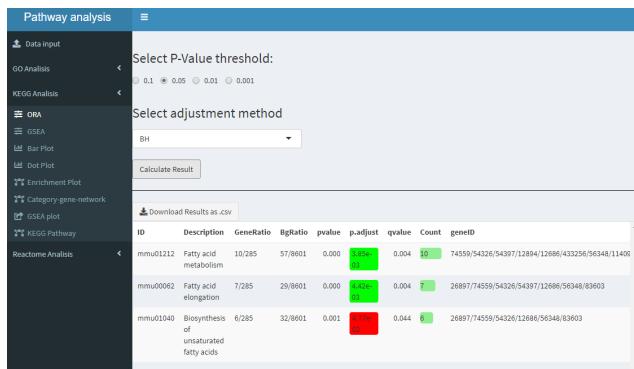


imatge 5.10: Gràfic GSEA

També podem fer l'anàlisi de KEGG. El resultat de KEGG és similar a l'anàlisi de Reactome. L'aplicació permet però generar les rutes KEGG. Per obtenir-les:

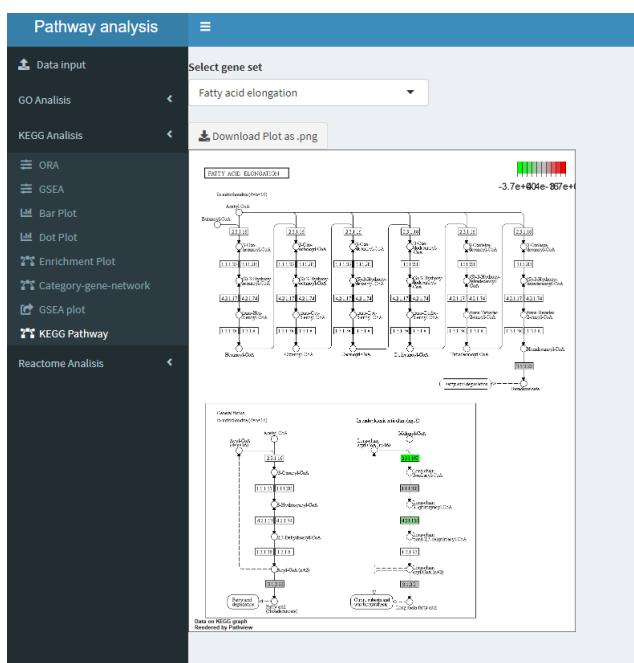
5 Validació dels resultats

1. Clico en l'apartat *KEGG Analysis*→*ORA*. Selecciono com a mètode d'ajustament *BH* i el cut-off del valor de P ajustat 0.05. Clico a *Calculate results*



Imatge 5.11: Anàlisi ORA de KEGG

2. Anem a *KEGG*→*KEGG Pathway*

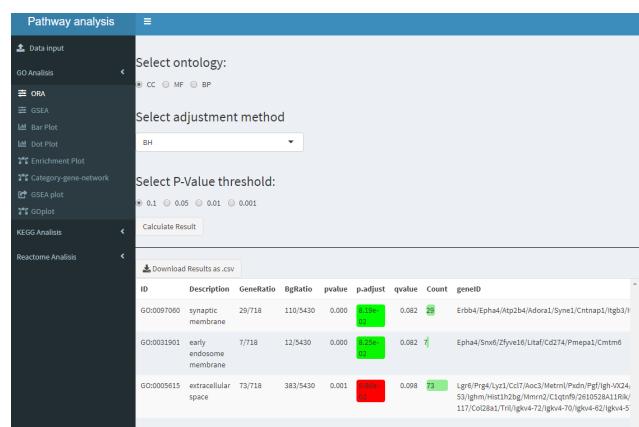


Imatge 5.12: Gràfic de les rutes KEGG

5.1 Exemple d'anàlisi 1. GEO: GSE100924

L'anàlisi GO no retorna cap terme GO amb el nivell de significació de 0.05. Pujant el nivell de significació fins 0.1 retorna un llistat dels termes enriquits per als components cel·lulars.

Clico en l'apartat *GO Analysis→ORA*. Selecciono com a mètode d'ajustament *BH* i el cut-off del valor de P ajustat 0.1. Selecciono també CC. Clico a *Calculate results*



Imatge 5.13: L'anàlisi ORA de GO

6 Discussió

Appendix

Bibliography

- [Boyle et al., 2004] Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004). Go:: Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715.
- [Chang et al., 2018] Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.2.0.
- [Consortium, 2004] Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261.
- [Dinu et al., 2007] Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by sam-gs. *BMC bioinformatics*, 8(1):242.
- [Draghici et al., 2007] Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome research*, 17(10):1537–1545.
- [Farmer et al., 2005] Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., MacGrogan, G., Bergh, J., Cameron, D., Goldstein, D., et al. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Breast Cancer Research*, 7(2):P2–11.
- [Hengel et al., 2003] Hengel, R. L., Thaker, V., Pavlick, M. V., Metcalf, J. A., Dennis, G., Yang, J., Lempicki, R. A., Sereti, I., and Lane, H. C. (2003). Cutting edge: L-selectin (cd62l) expression distinguishes small resting memory cd4+ t cells that preferentially respond to recall antigen. *The Journal of Immunology*, 170(1):28–32.

Bibliography

- [Khatri et al., 2012] Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375.
- [Kim and Volsky, 2005] Kim, S.-Y. and Volsky, D. J. (2005). Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, 6(1):144.
- [Li et al., 2017] Li, S., Mi, L., Yu, L., Yu, Q., Liu, T., Wang, G.-X., Zhao, X.-Y., Wu, J., and Lin, J. D. (2017). Zbtb7b engages the long noncoding rna blnc1 to drive brown and beige fat development and thermogenesis. *Proceedings of the National Academy of Sciences*, 114(34):E7111–E7120.
- [Luo et al., 2009] Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, 10(1):161.
- [Newton et al., 2007] Newton, M. A., Quintana, F. A., Den Boon, J. A., Sengupta, S., Ahlquist, P., et al. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, 1(1):85–106.
- [Rahnenführer et al., 2004] Rahnenführer, J., Domingues, F. S., Maydt, J., and Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical applications in genetics and molecular biology*, 3(1):1–29.
- [Reimand et al., 2019] Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, page 1.
- [Schmidt et al., 2008] Schmidt, M., Böhm, D., Von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kölbl, H., and Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research*, 68(13):5405–5413.
- [Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a

Bibliography

- knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.
- [Tarca et al., 2008] Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2008). A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82.
- [Wickham, 2015] Wickham, H. (2015). R packages.