

Memòria del treball de final de màster

Vasyl Druchkiv

Estudiant del Màster de Bioestadística i Bioinformàtica

20 de Maig 2019

Índice

1	Introducció	1
2	Objectius	1
2.1	Objectius generals	1
2.2	Objectius específics	2
3	El marc teòric	3
3.1	Les dades d'expressió genètica	3
3.2	Annotació dels gens	5
3.2.1	Gene ontology	5
3.2.2	KEGG	6
3.2.3	Reactome	7
3.3	ORA	7
3.4	GSEA	8
3.5	L'anàlisi topològic de les rutes	10
3.5.1	El mapa d'enriquement	10
3.5.2	Gene-Concept-Network	10
3.5.3	GOplot	10
3.5.4	KEGG Pathway	10
3.5.5	Reactome Pathway	11
4	Tractament bioinformàtic	11
4.1	Cerca dels paquets de Bioconductor	11
4.2	Desenvolupar el protocol	12
5	Instal·lació de l'aplicació	13
6	L'anàlisi comuna de GO, KEGG i Reactome	14
6.1	ORA	14
6.1.1	GO	14
6.1.2	KEGG	16
6.1.3	Reactome	17
6.2	GSEA	17
6.2.1	GO	17
6.2.2	KEGG	18
6.2.3	Reactome	19
6.3	Bar-Plots	19
6.4	Dot-Plots	20
6.5	Enrichment Plots	22

6.6	Category-Gene-Network Plot	23
6.7	GSEA Plot	23
7	L'anàlisi específic de GO, KEGG i Reactome	25
7.1	GO Plot	25
7.2	KEGG Pathway	26
7.3	Reactome Pathway	26
8	Manual i les ajudes del programa	27
9	Validació dels resultats	31
9.1	Exemple d'anàlisi 1. GEO: GSE100924	32
10	Discussió	39
	Bibliografia	39

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 intuïtiva 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 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 entenenent 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 accessible 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.

2 Objectius

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

2.1 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)

2.2 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?
 - 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 probes 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.

3 El marc teòric

3.1 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 tipus salvatge i l'altre amb el gen ZBTB7B silenciado i investigar quins gens són diferencialment expressats entre aquests mostres biològiques.

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

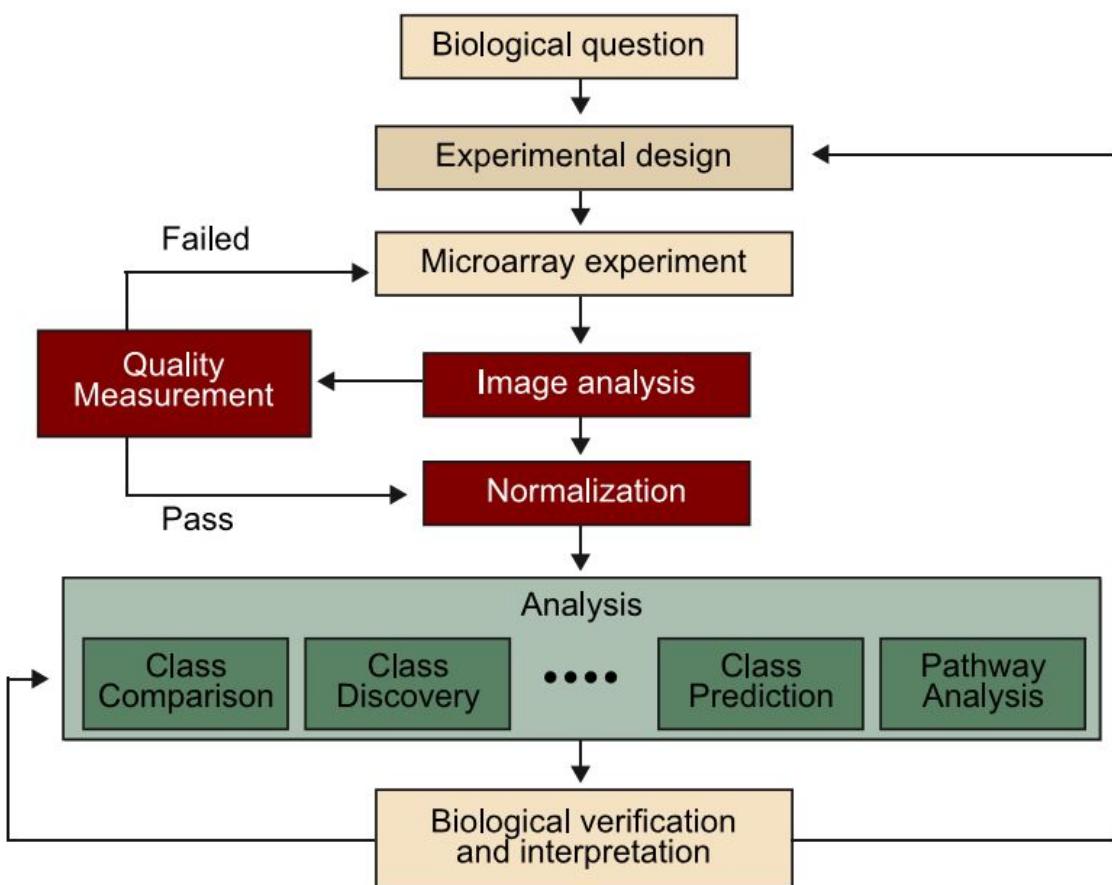


Figure 1: El procès d'anàlisi de microarrays

El *pipeline* de l'anàlisi consisteix doncs en plantejament d'una pregunta i un disseny experimental a partir del qual es 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. El test adequat dependrà bàsicament de la distribució de les dades. El resultat d'aquesta anàlisi serveix com a base per a interpretació biològica dels resultats d'experiment. Per poder fer sentit de les dades d'expressió i de l'anàlisi estadístic al nivell de gens és imprescindible fer un anàlisi a nivell de les categories de gens o les rutes. Per aquest anàlisi 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'anàlisi de rutes (*Pathway analysis*).

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

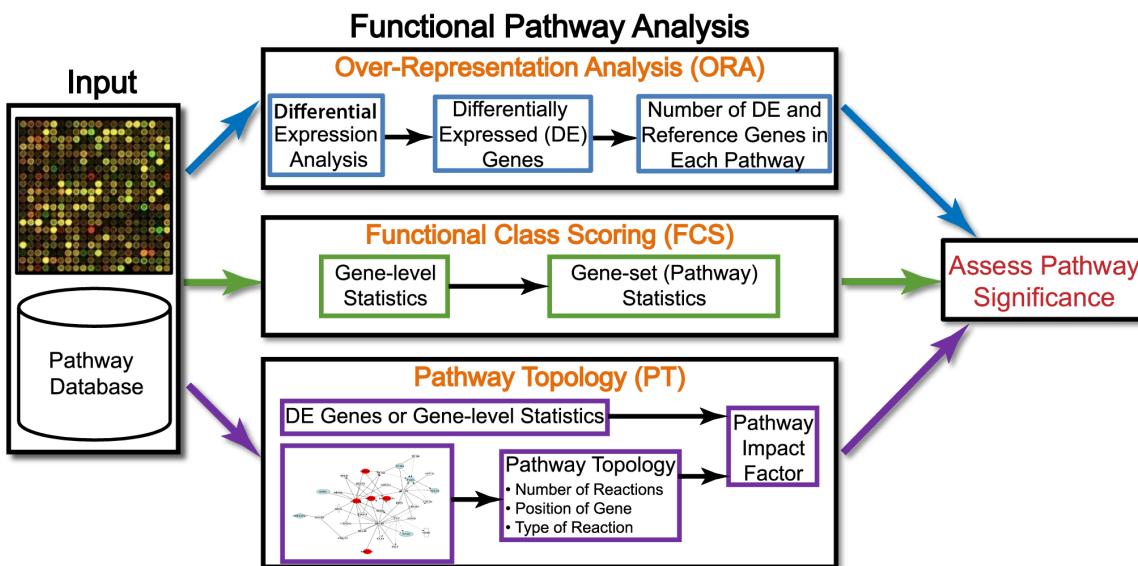


Figure 2: El procés d'anàlisi de les rutes

A part de les dades d'expressió, de les quals he parlat anteriorment, l'anàlisi 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 és 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:

- **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 ([Tarcà 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 coneccions amb altres gens a la ruta podria ser important [Rahnenführer et al., 2004]. O fins i tot leas estadístiques que incorporen factors diferents com ara la posició, el tipo d'interacció etc. [Draghici et al., 2007]. Aquesta idea l'he 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]

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.

3.2 Anotació dels gens

Com veurem més endavant per a anàlisi 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.

3.2.1 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 entendre 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 espeifiquen 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*.
3. **Cellular Component (CC)**. El CC descriu l’emplaçaments al nivell d’estructures subcel·lulars (com *mitochondri*) 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.

3.2.2 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 específic (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.

3.2.3 Reactome

Reactome és una base de dades gratuitament accessible 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, catalisi 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.

3.3 ORA

L'anàlisi de sobreexpressió és una tècnica d'identificació de les rutes significativament enriquides 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 procès 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:

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

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 distribució de fons.

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

- Description. El nom del terme GO;
- GeneRatio. El quocient: $\frac{\text{Nombre dels gens diferencialment expressats que pertanyen al conjunt de gens}}{\text{Nombre total dels gens diferencialment expressats}} = \frac{M}{N};$
- BgRatio. El quocient: $\frac{\text{Nombre dels gens del conjunt d'interès en la distribució de fons}}{\text{Nombre total dels gens en la distribució de fons}} = \frac{k}{n};$
- 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.

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

L'anàlisi GSEA consisteix en tres passos:

1. Càcul de la puntuació d'enriquement (*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. Càcul 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.

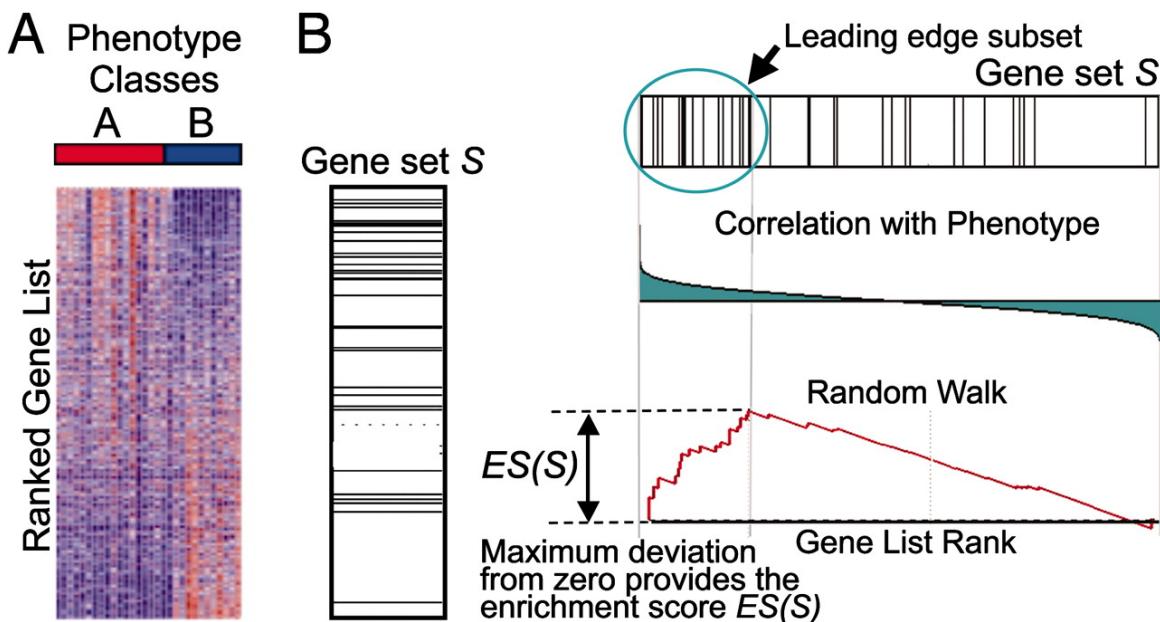


Figure 3: El mètode GSEA

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

- enrichmentScore. Enrichment score per al conjunt dels gens. Amb altres paraules: el grau amb el qual el conjunt dels gens està sobreexpressat 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'enriquement.
 - 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.

3.5 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.

3.5.1 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 soplats. D'aquesta manera les rutes amb gens en comú s'agrupen més a prop l'una de l'altra.

3.5.2 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.

3.5.3 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 els gens més significatius. El gràfic mostra tres relacions possibles entre les rutes:

1. *is a*: Si diem que A *is a* B, volem dir que A és un subtip 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 només 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.

3.5.4 KEGG Pathway

Aquest gràfic mostra les relacions entre els gens dins de la ruta específica. Els gens són remarcats amb el color depenent 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
			chemical compound, DNA and other molecule
			map
			molecular interaction or relation
			link to/from another 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	
			phosphorylation
			dephosphorylation
			ubiquitination
			deubiquitination
			glycosylation
			methylation
			activation
			inhibition
			indirect effect or state change
			binding / association
			dissociation
			complex
Enzyme-enzyme relations			two successive reaction steps

Figure 4: L'anotació de les relacions dins de les rutes KEGG

3.5.5 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.

4 Tractament bioinformàtic

4.1 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

propost per [Subramanian et al., 2005].

4.2 Desenvolupar el protocol

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

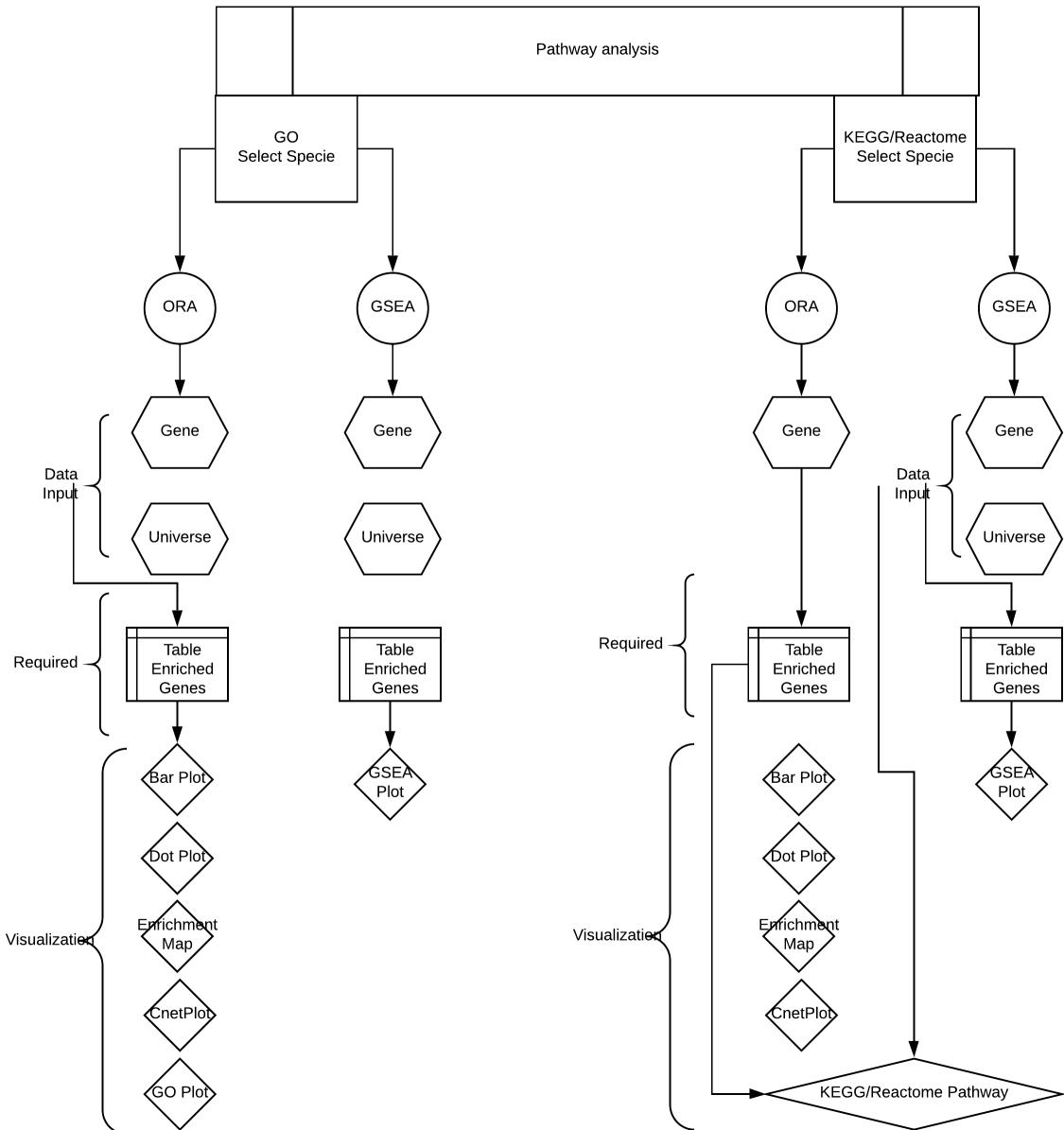


Figure 5: Lucidchart per a l'aplicació

D'aquí podem definir per exemple el protocol:

1. Decidir quin anàlisi 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.

- Per a l'anàlisi GO tots dos arxius són necessaris: Gens Selecciónats (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.

5 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 dels voluntaris. D'aquesta manera l'aplicació no seria available 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/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:

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

6 L'anàlisi comuna de GO, KEGG i Reactome

6.1 ORA

6.1.1 GO

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

```
enrichGO(gene, OrgDb, keyType = "ENTREZID", ont = "MF", pvalueCutoff = 0.05,
pAdjustMethod = "BH", universe, qvalueCutoff = 0.2, minGSSize = 10, maxGSSize = 500,
readable = FALSE, pool = FALSE)
```

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.

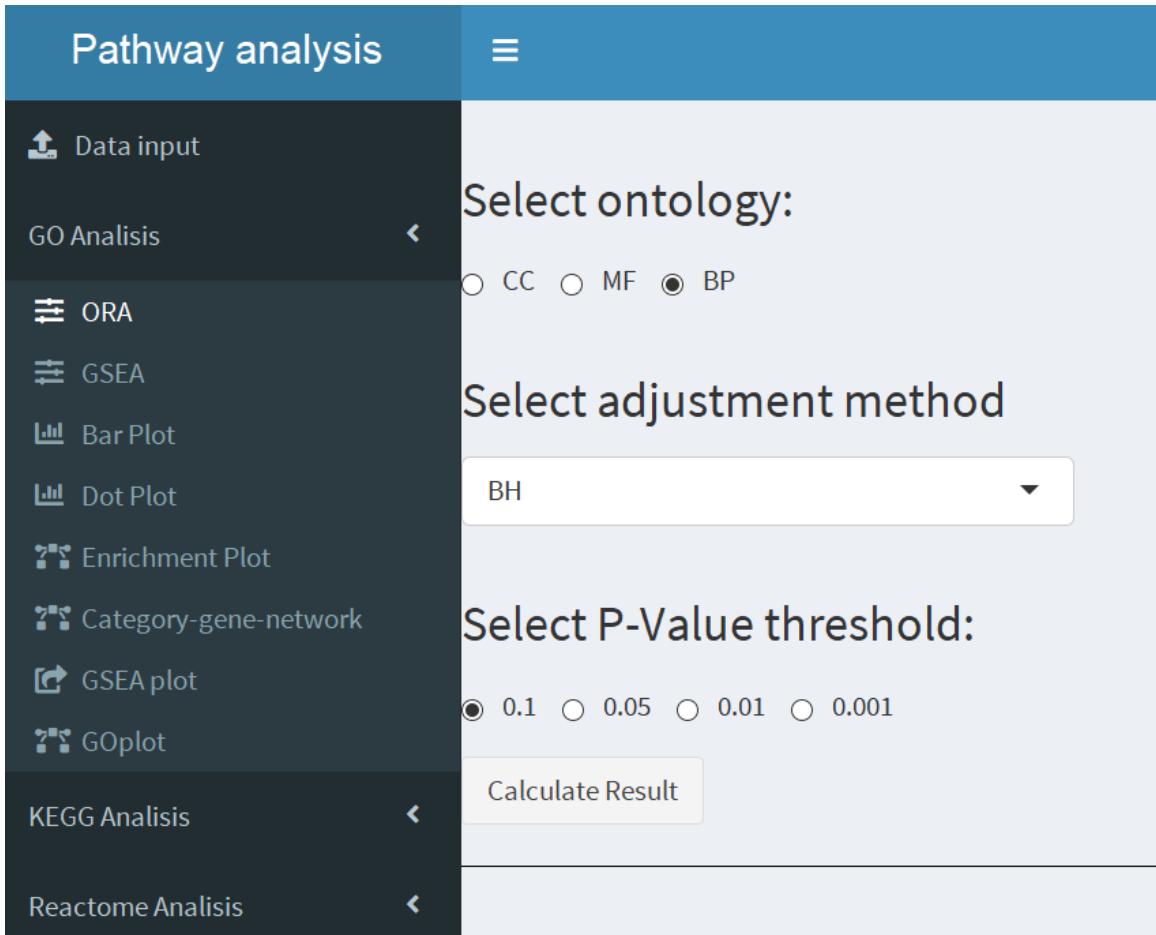


Figure 6: Especificació d'ORA dels termes GO

L'execució de la funció és un procès temporalment costós. Per aquest motiu he afegit el botó d'accio, 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.

Calculate Result								
 Download Results as .csv								
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/
GO:0000280	nuclear division	35/193	316/11468	0.000	4.50e-16	0.000	35	CDCA8/CDC20/KIF23/CENPE/MYBL2/CCNB2/

Figure 7: El resultat d'anàlisi ORA. GO.

6.1.2 KEGG

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

```
enrichKEGG(gene , organism = "hsa" , keyType = "kegg" , pvalueCutoff = 0.05 ,
pAdjustMethod = "BH" , universe , minGSSize = 10 , maxGSSize = 500 ,
qvalueCutoff = 0.2 , use_internal_data = FALSE)
```

Figure 8: 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.

Calculate Result								
 Download Results as .csv								
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	genelID
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

Figure 9: El resultat de l'anàlisi ORA. KEGG.

6.1.3 Reactome

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

```
enrichPathway(gene, organism = "human", pvalueCutoff = 0.05,
pAdjustMethod = "BH", qvalueCutoff = 0.2, universe, minGSSize = 10,
maxGSSize = 500, readable = FALSE)
```

Pathway analysis								
 Download Results as .csv								
Select adjustment method								
BH								
Select P-Value threshold:								
<input checked="" type="radio"/> 0.1 <input type="radio"/> 0.05 <input type="radio"/> 0.01 <input type="radio"/> 0.001								
Calculate Result								
 Download Results as .csv								
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	genelID
R-HSA-2500257	Resolution of Sister Chromatid Cohesion	15/142	124/10554	0.000	2.83e-08	0.000	15	CDCA8/CDC20/CENPE/CCNB2/NDC80/SKA1/CENP
R-HSA-68877	Mitotic Prometaphase	18/142	198/10554	0.000	2.83e-08	0.000	18	CDCA8/CDC20/CENPE/CCNB2/NDC80/NCAPH/SKA1/CENP
R-HSA-69620	Cell Cycle Checkpoints	21/142	293/10554	0.000	2.30e-08	0.000	21	CDC45/CDCA8/MCM10/CDC20/CENPE/CCNB2/NDC80/NCAPH/SKA1/CENP
R-HSA-	Mitotic Spindle	13/142	112/10554	0.000	3.61e-08	0.000	13	CDCA8/CDC20/CENPE/NDC80/UBE2C/SKA1/CENP

Figure 10: El resultat d'anàlisi ORA. Reactome.

6.2 GSEA

6.2.1 GO

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

```
gseGO(geneList, ont = "BP", OrgDb, keyType = "ENTREZID",
```

```

exponent = 1, nPerm = 1000, minGSSize = 10, maxGSSize = 500,
pvalueCutoff = 0.05, pAdjustMethod = "BH", verbose = TRUE,
seed = FALSE, by = "fgsea")

```

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

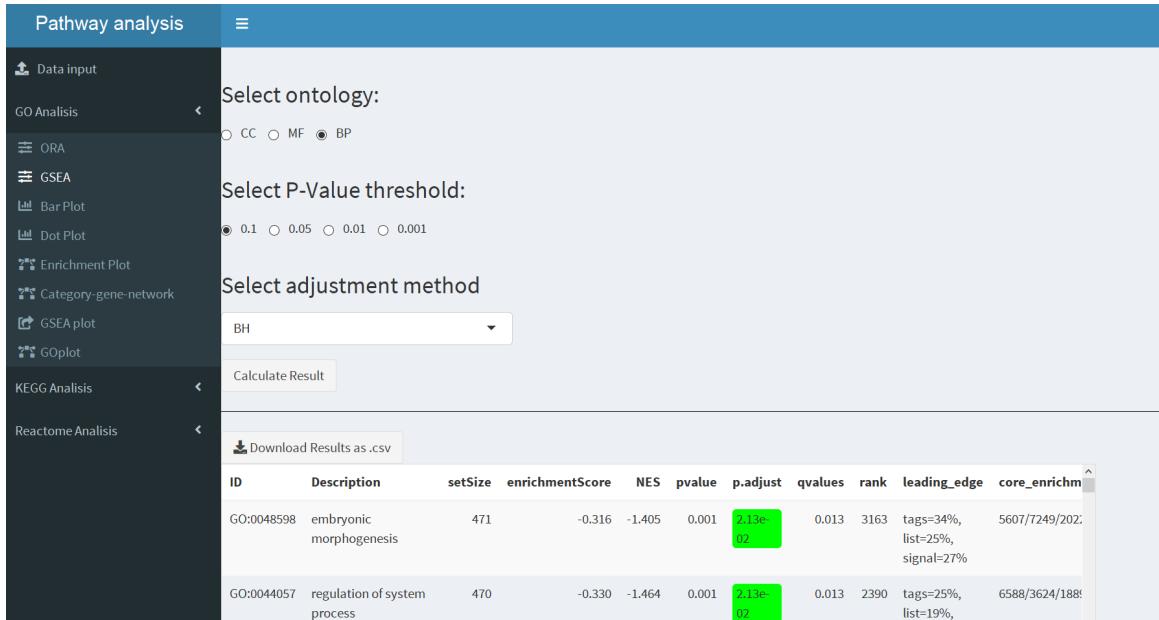


Figure 11: El resultat de l'anàlisi GSEA. GO.

6.2.2 KEGG

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

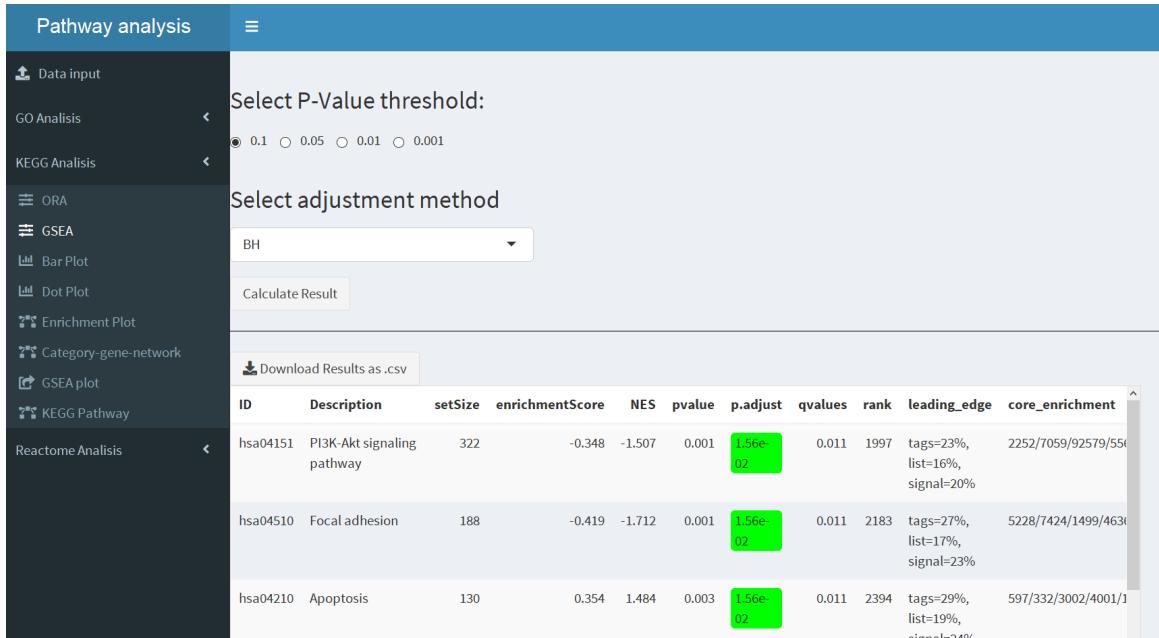


Figure 12: El resultat de l'anàlisi GSEA. KEGG.

6.2.3 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()**

The screenshot shows the 'Pathway analysis' interface for Reactome. On the left, a sidebar lists analysis types: Data input (GO Analysis, KEGG Analysis, Reactome Analysis), visualization methods (ORA, GSEA, Bar Plot, Dot Plot, Enrichment Plot, Category-gene-network, GSEA plot, Reactome Pathway), and a download option for CSV results. The main panel is titled 'Select P-Value threshold:' with radio buttons for 0.1, 0.05, 0.01, and 0.001 (0.1 is selected). Below that is 'Select adjustment method' with a dropdown set to 'BH'. A 'Calculate Result' button is present. The results table displays three rows of pathway data:

ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrichment
R-HSA-9006934	Signalling by Receptor Tyrosine Kinases	416	-0.340	-1.489	0.001	1.32e-02	0.006	2788	tags=28%, list=22%, signal=22%	5580/2242/5802/9101,
R-HSA-1474244	Extracellular matrix organization	266	-0.458	-1.922	0.001	1.32e-02	0.006	1943	tags=33%, list=16%, signal=29%	8038/11132/4017/1281
R-HSA-5693538	Homology Directed Repair	102	0.558	2.283	0.003	1.32e-02	0.006	1990	tags=37%, list=16%, signal=32%	10635/890/1111/9156,

Figure 13: El resultat d'anàlisi GSEA. Reactome.

6.3 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.

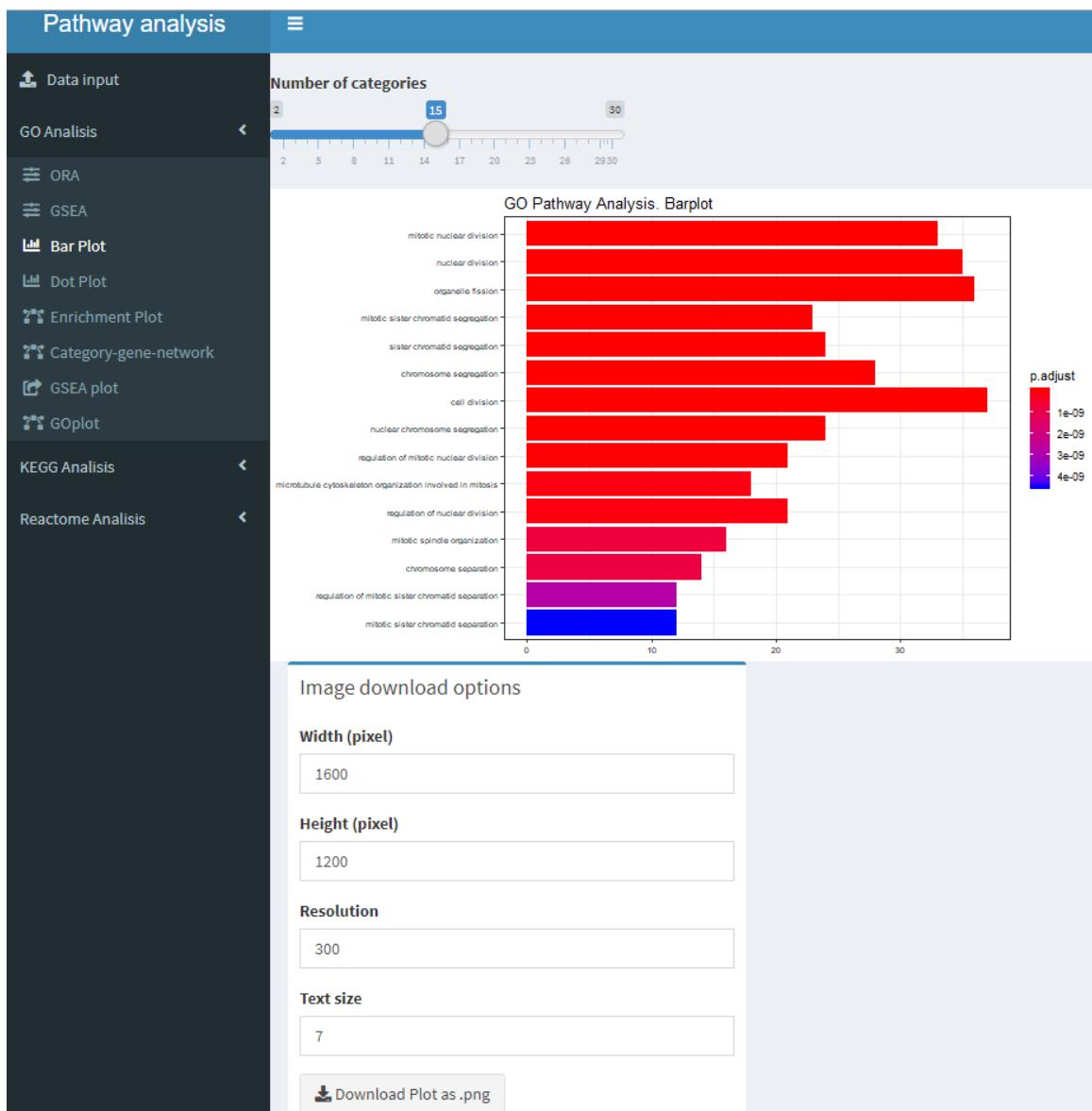


Figure 14: Bar-Plot. GO.

6.4 Dot-Plots

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

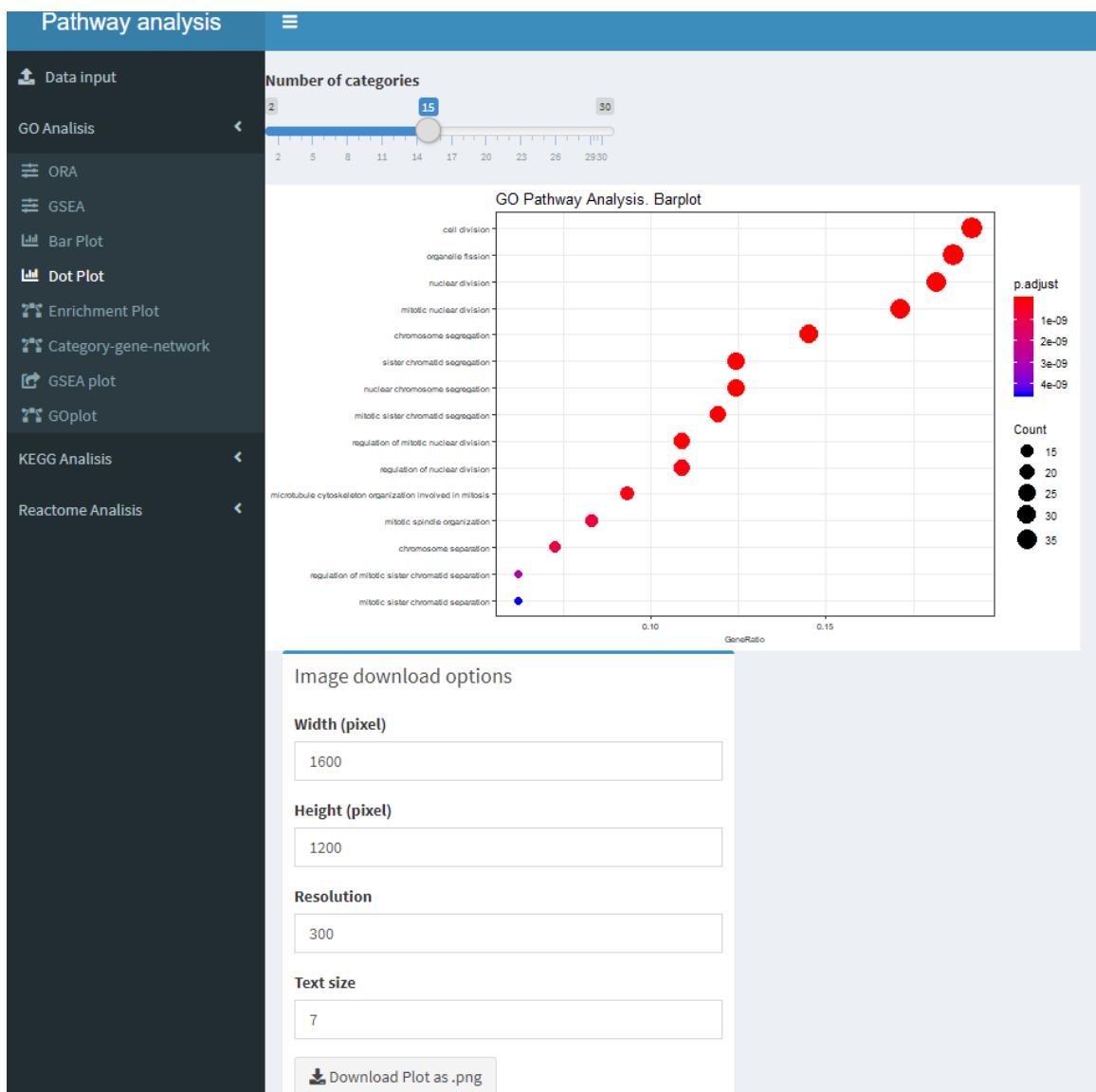


Figure 15: Bar-Plot. GO.

6.5 Enrichment Plots

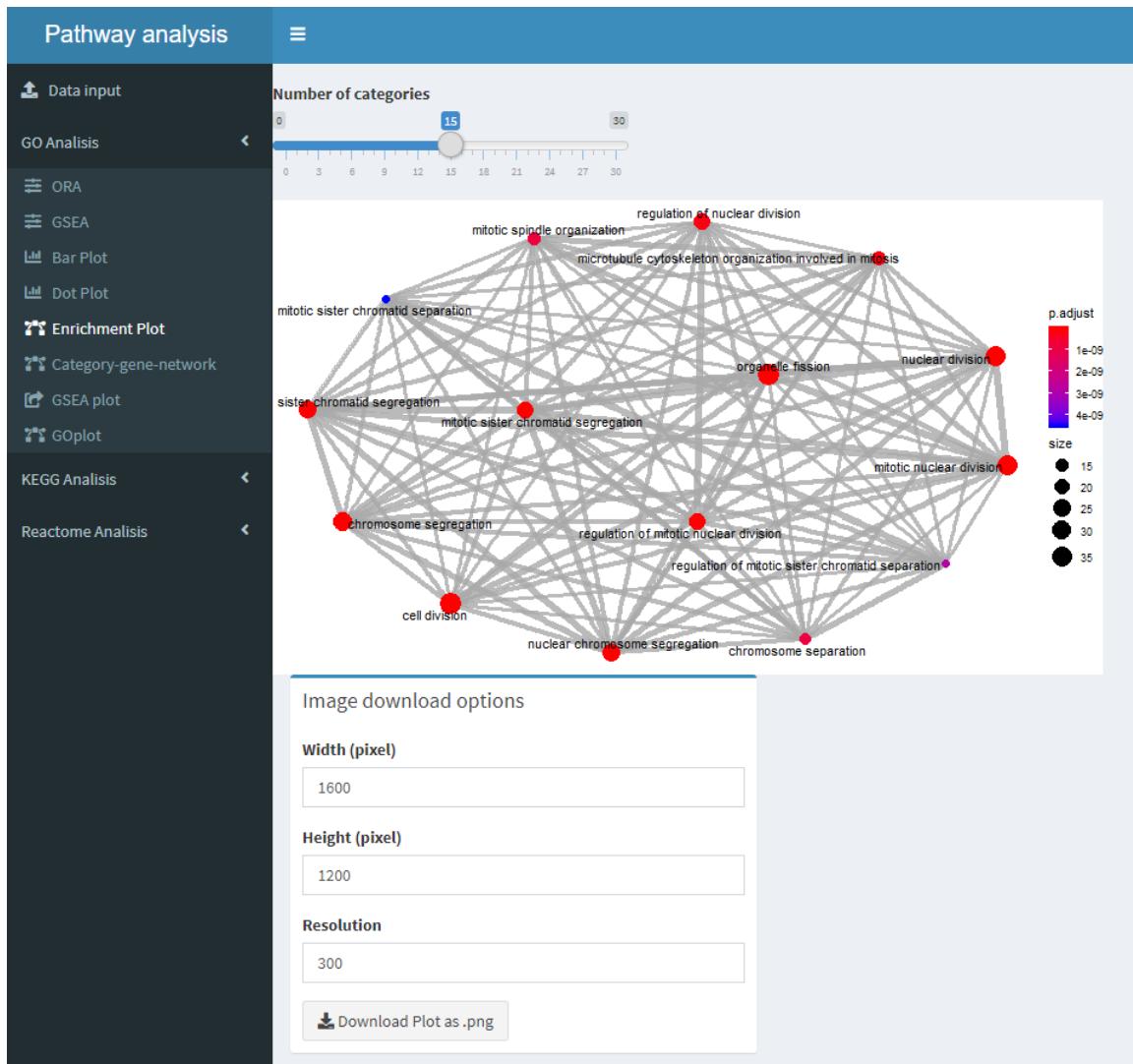


Figure 16: Bar-Plot. GO.

6.6 Category-Gene-Network Plot

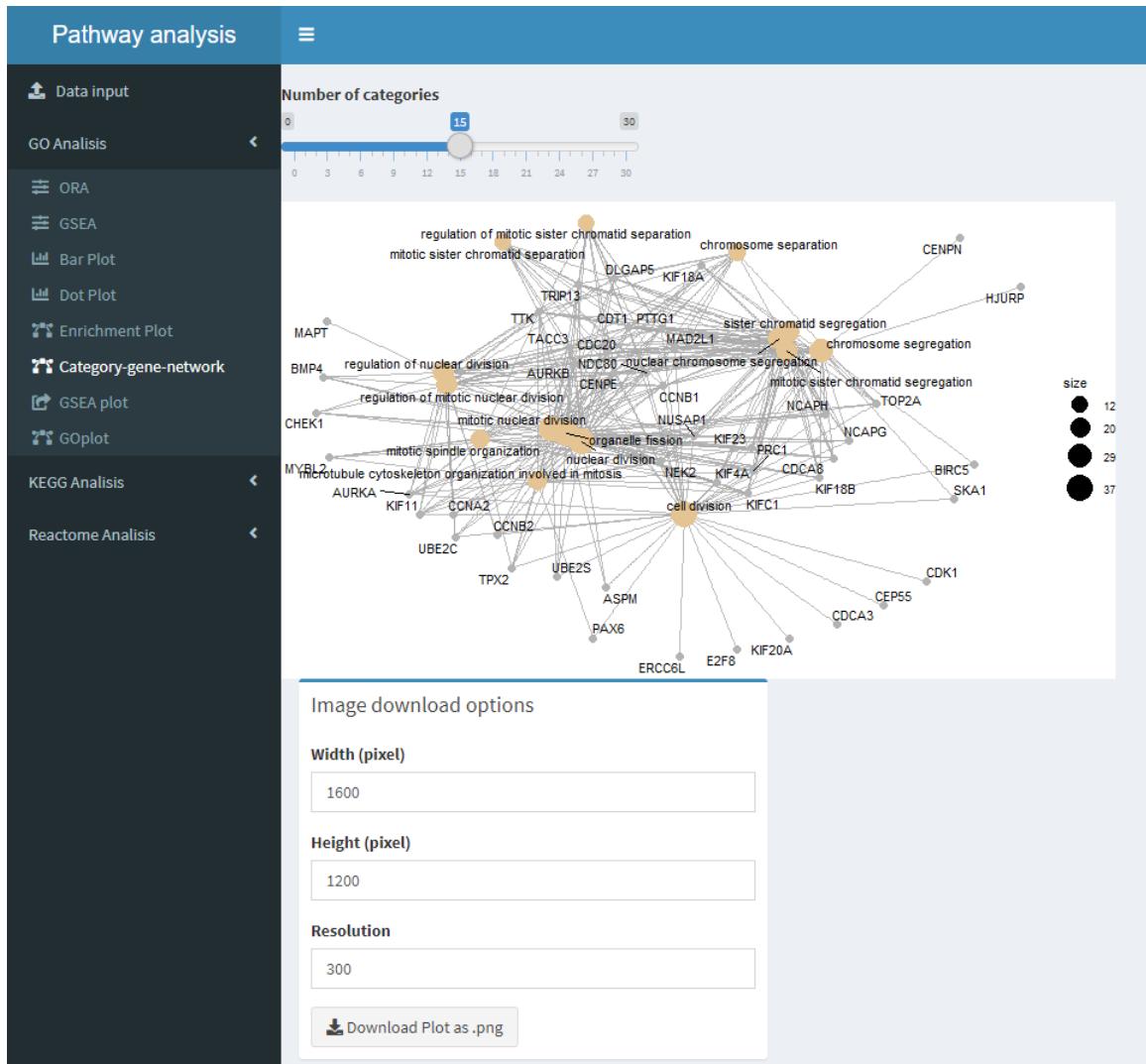


Figure 17: Category-Gene-Network Plot. GO.

6.7 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*

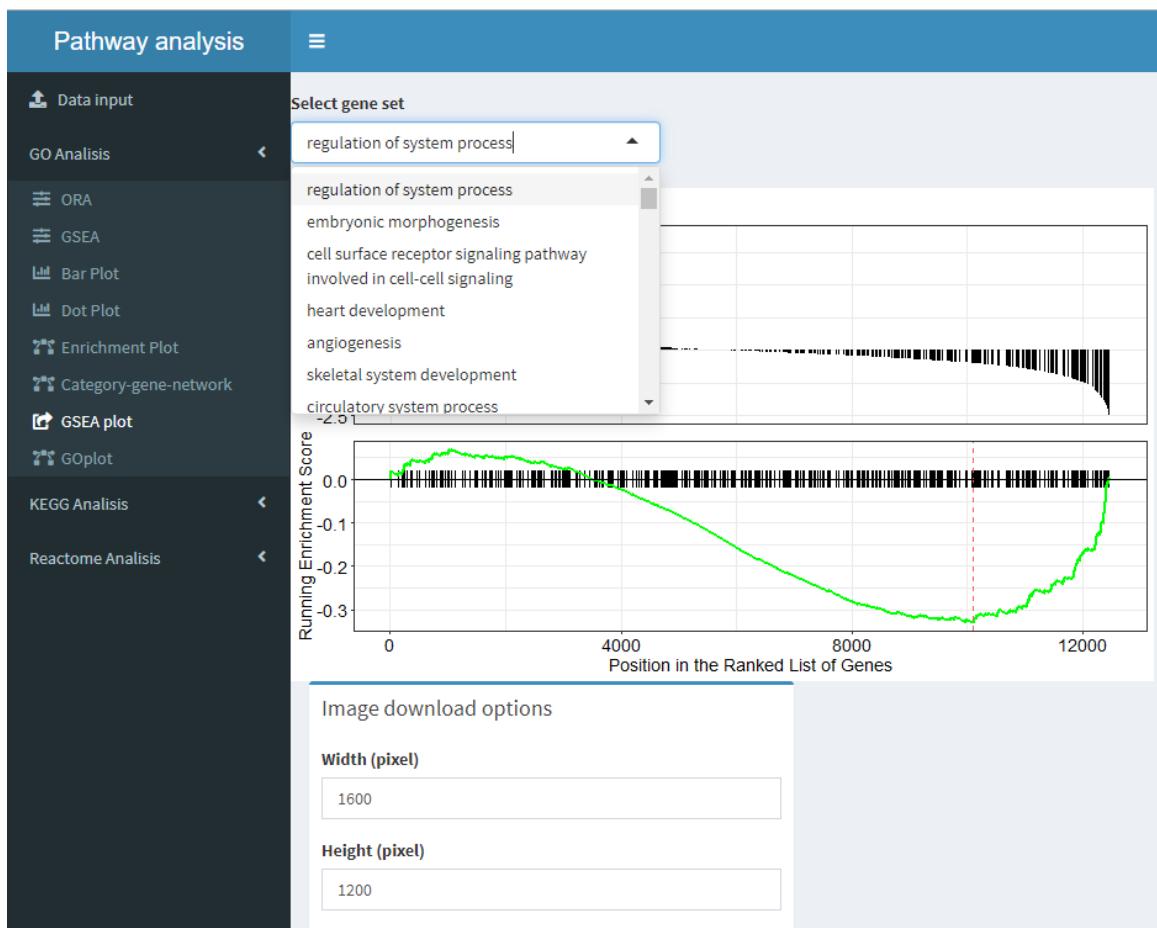


Figure 18: GSEA Plot. GO.

7 L'anàlisi específic de GO, KEGG i Reactome

7.1 GO Plot

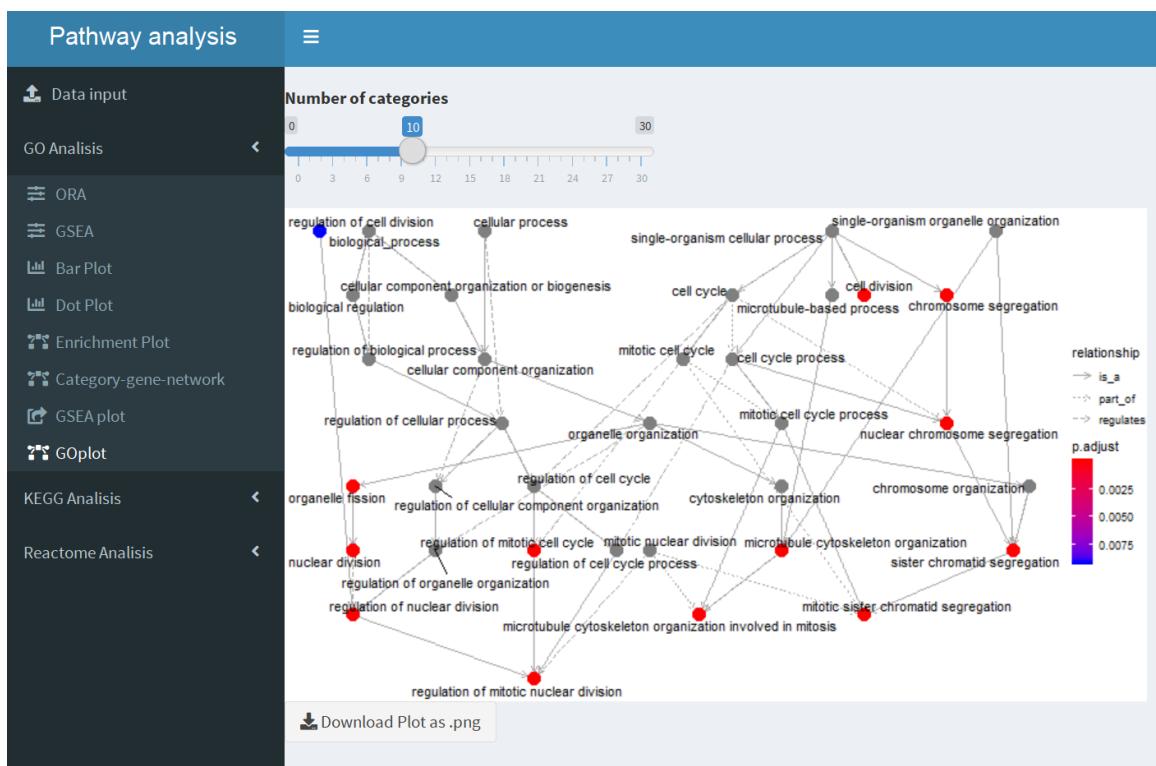
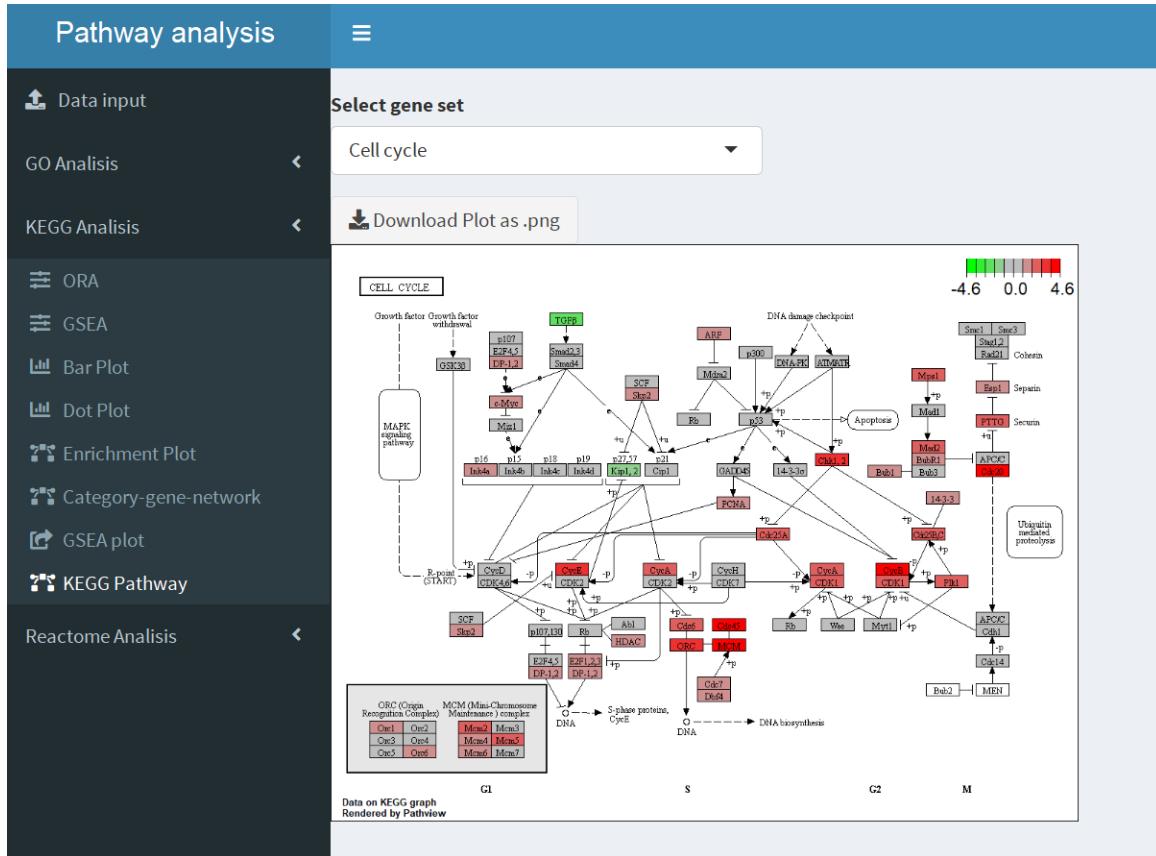


Figure 19: GO Plot

7.2 KEGG Pathway



8 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**:

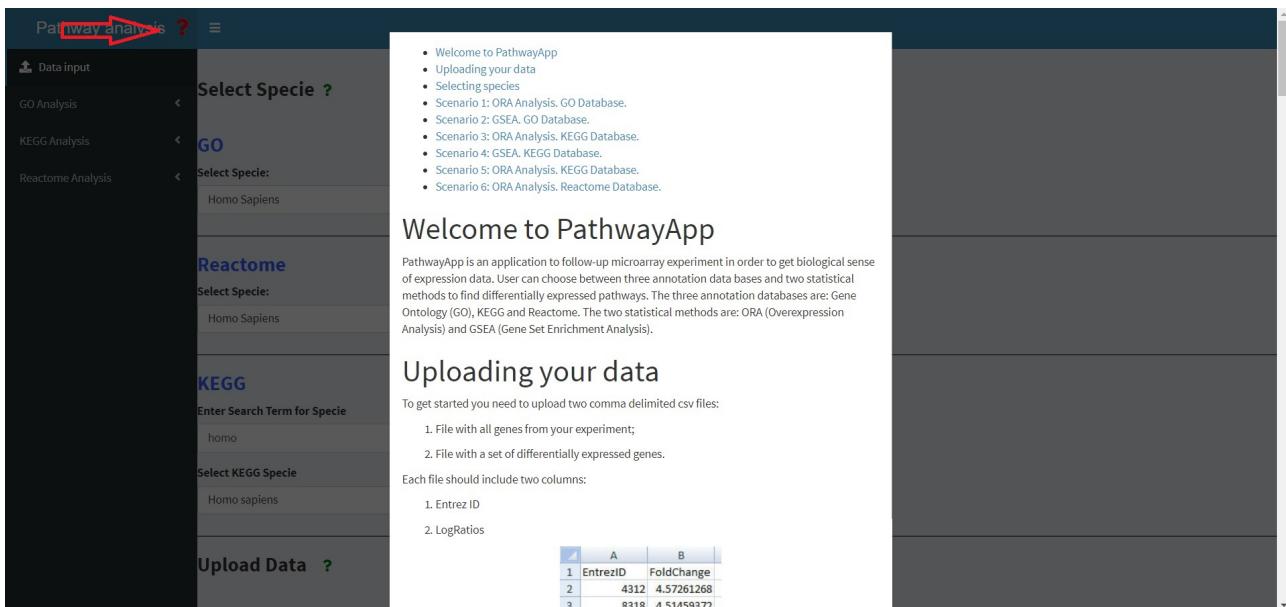


Figure 22: Manual per a aplicació

Com es veu hi ha apartats diferents. Dependent 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}.

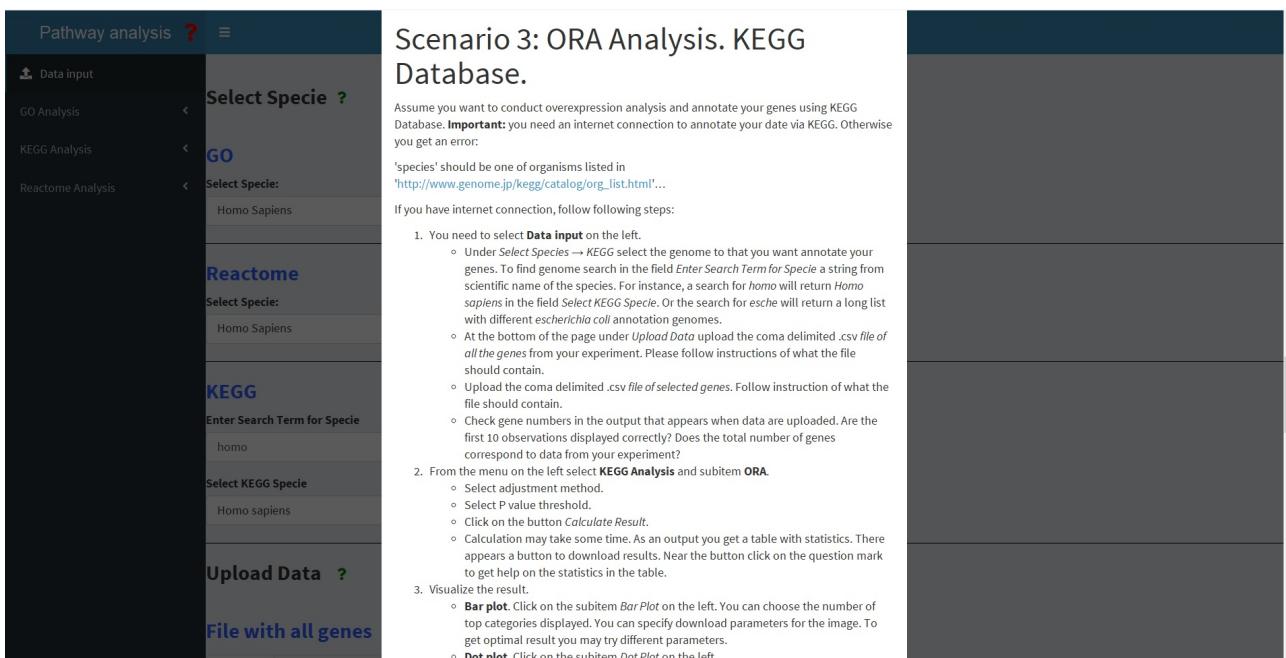


Figure 23: Manual per a l'anàlisi ORA amb l'anotació KEGG

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()`.

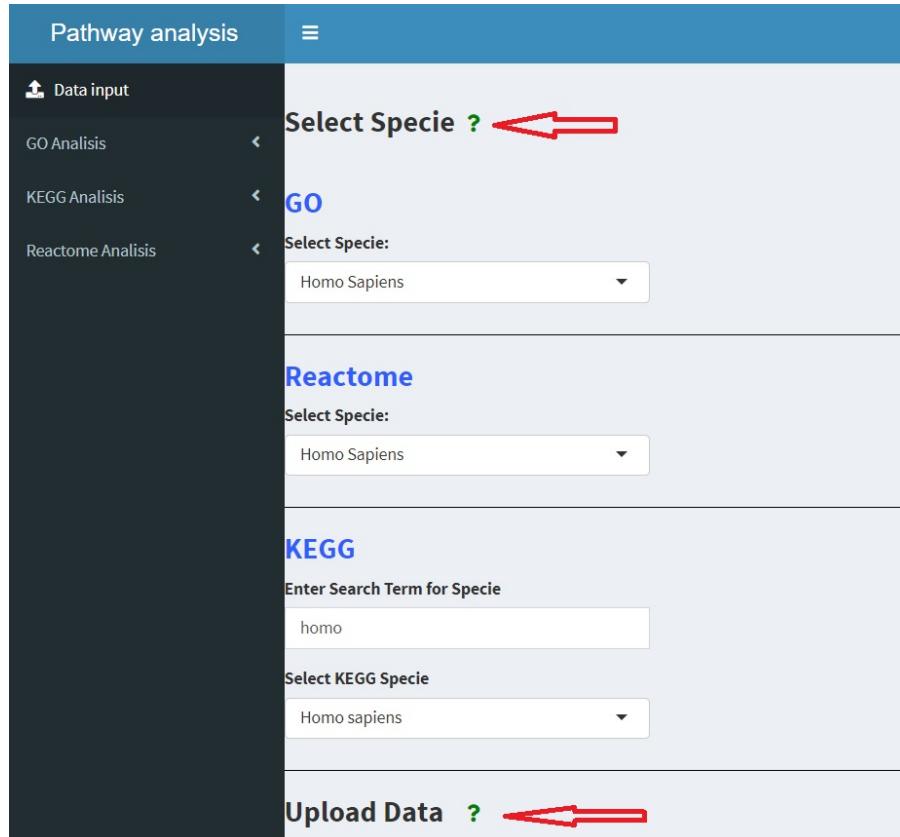


Figure 24: Senyals d'ajuda

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

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

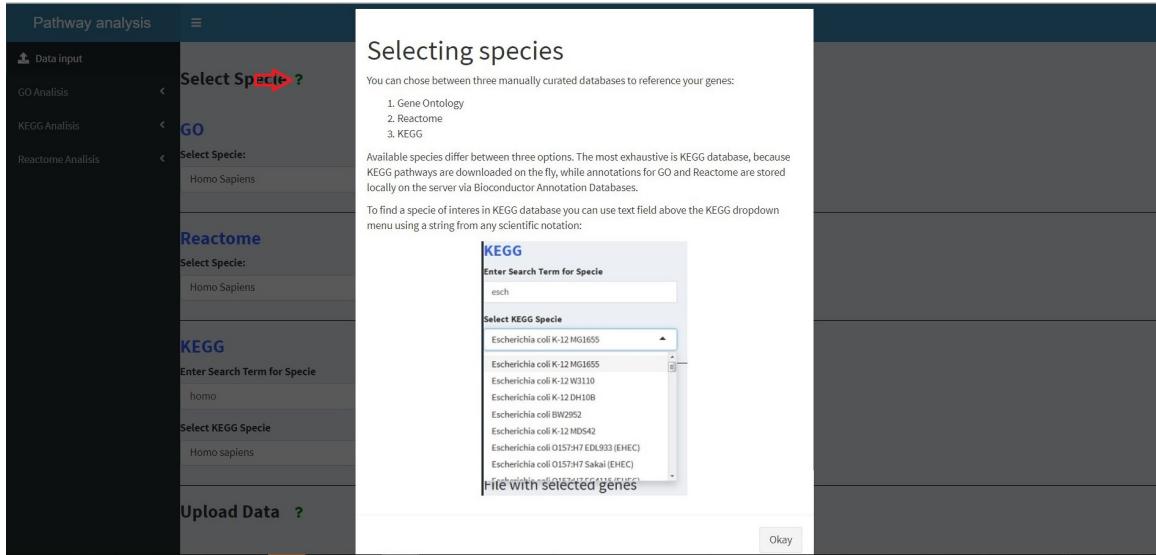


Figure 25: Ajuda per a l'elecció de l'espècie

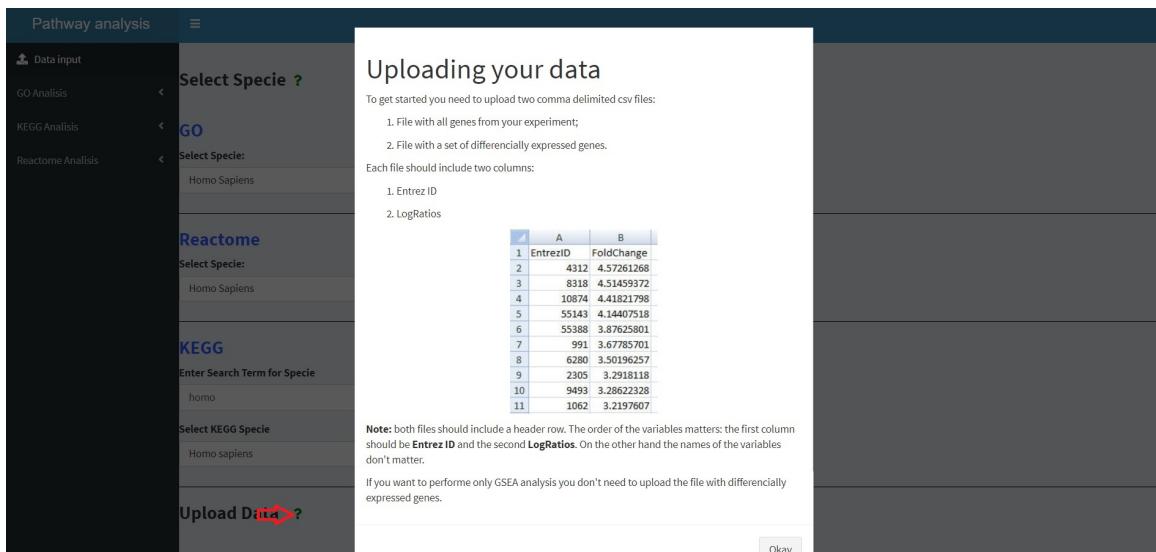


Figure 26: Ajuda per pujar les dades

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

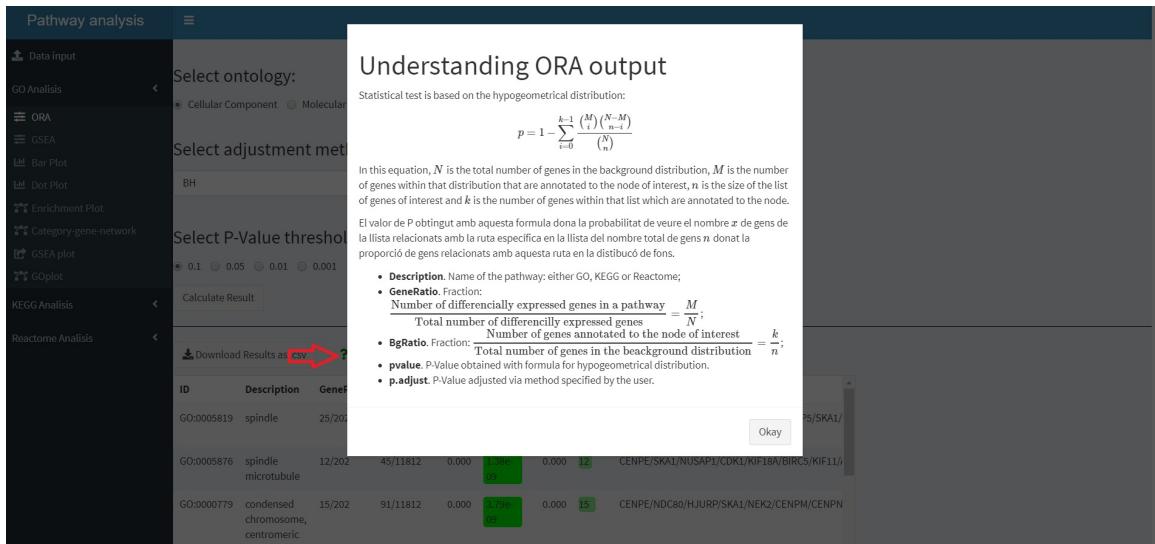


Figure 27: Infromació per la interpretació d'anàlisi ORA

Aquí cal destacar que les fòrmules, depenen 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 fórmula 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.

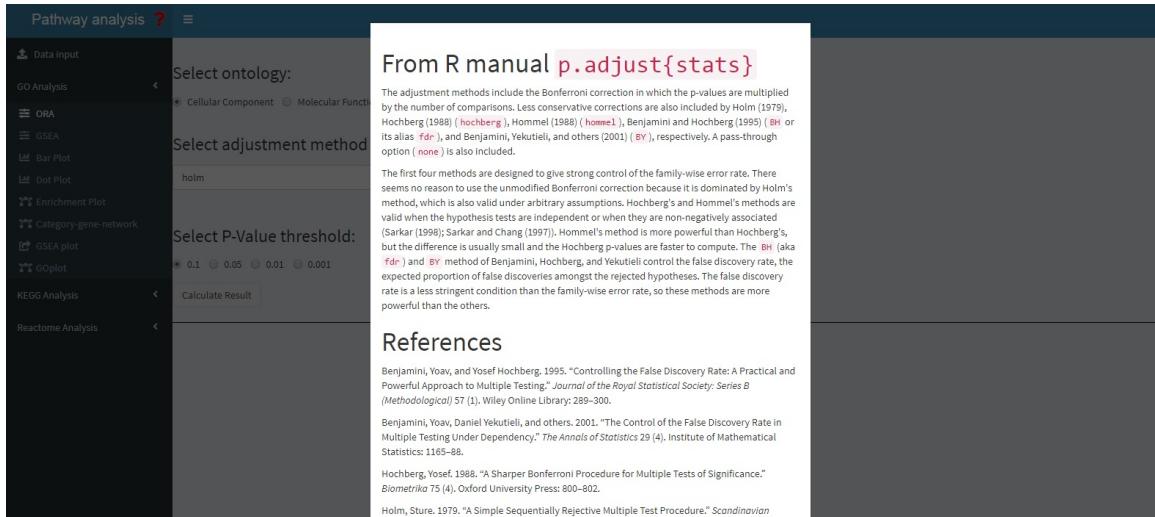


Figure 28: L'ajuda per a la selecció del mètode d'ajustament

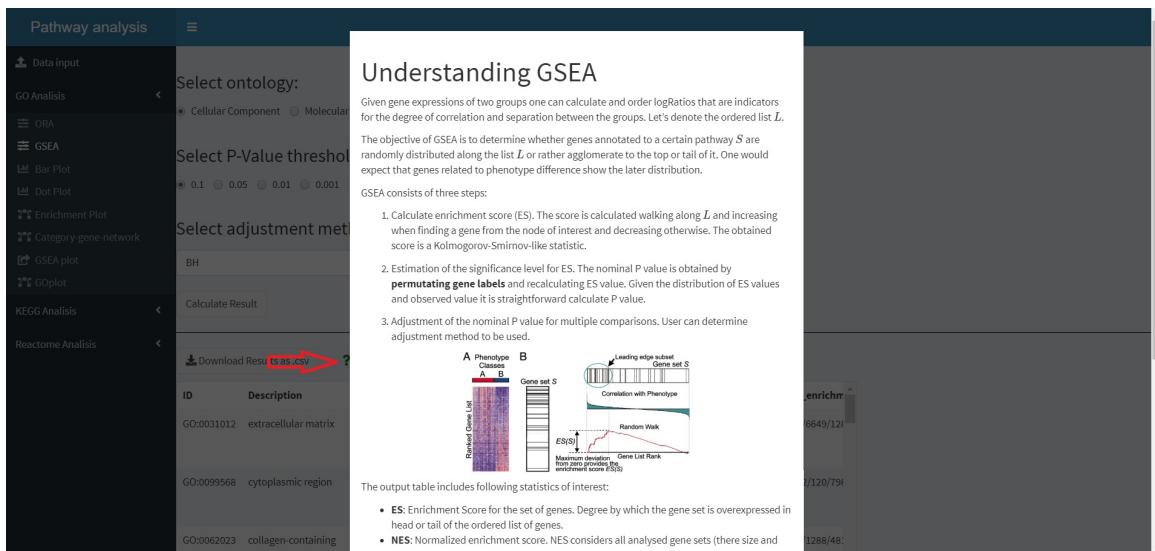


Figure 29: Ajuda per la interpretació de GSEA

9 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 arriba a resultats diferents dels resultats dels estudis d'on provenen les dades (i no parlo de l'anàlisi de les rutes sinó ja del càlcul 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	Especie	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 als resultats en els vignettes mencionats. Procediré doncs amb l'exemple basat en les dades de [Li et al., 2017] .

9.1 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.

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

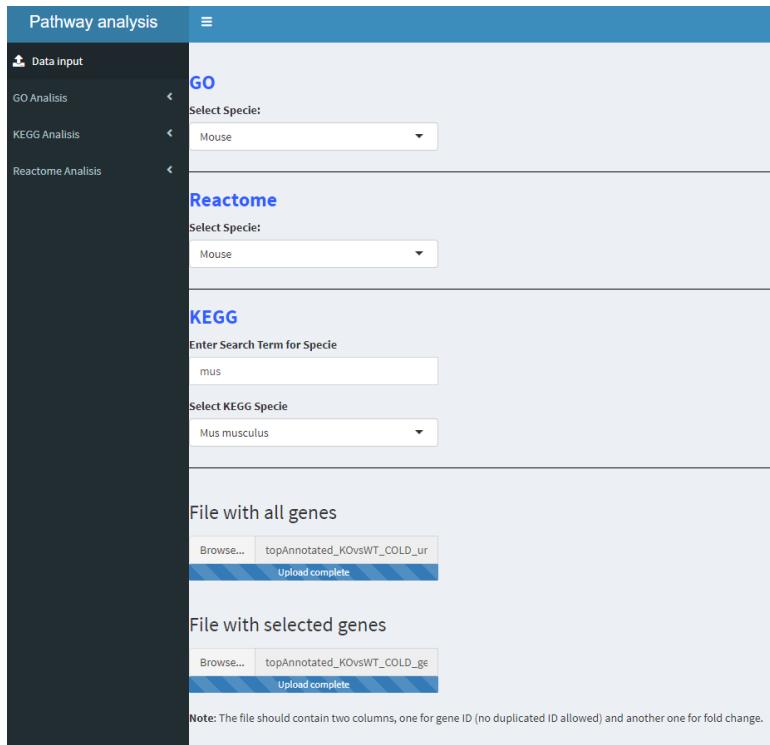


Figure 30: Selecció d'espècie

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.

Note: The file should contain two columns, one for gene ID (no duplicated ID allowed) and another one for fold change.

You uploaded: 5995 genes

First 10 entries

Entrez ID	FoldChange
108664	-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

Figure 31: Selecció d'espècie

2. 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*

ID	Description	GeneRatio	BgRatio	pvalue	p-adjust	qvalue	Count	geneID
R-MMU-75105	Fatty acyl-CoA biosynthesis	8/342	36/9114	0	0.017	0.017	8	Elov17/Elov12/Ppt2/Elov3/Acsf5/Hsd17b12/Cbr4/Acsf4
R-MMU-897886	Fatty acid metabolism	20/342	198/9114	0	0.017	0.017	20	Alox5/Slc22a21/Acot11/Acot6/Elov17/Eci3/Elov12/Ppt2/Cf
R-MMU-75876	Synthesis of very long-chain fatty acyl-CoAs	6/342	22/9114	0	0.031	0.031	6	Elov17/Elov12/Elov3/Acsf5/Hsd17b12/Acsf4

Figure 32: Resultat d'anàlisi ORA de Reactome

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

3. Visualització del resultat ORA

- Selecciono *Reactome Analysis*→*Bar Plot*

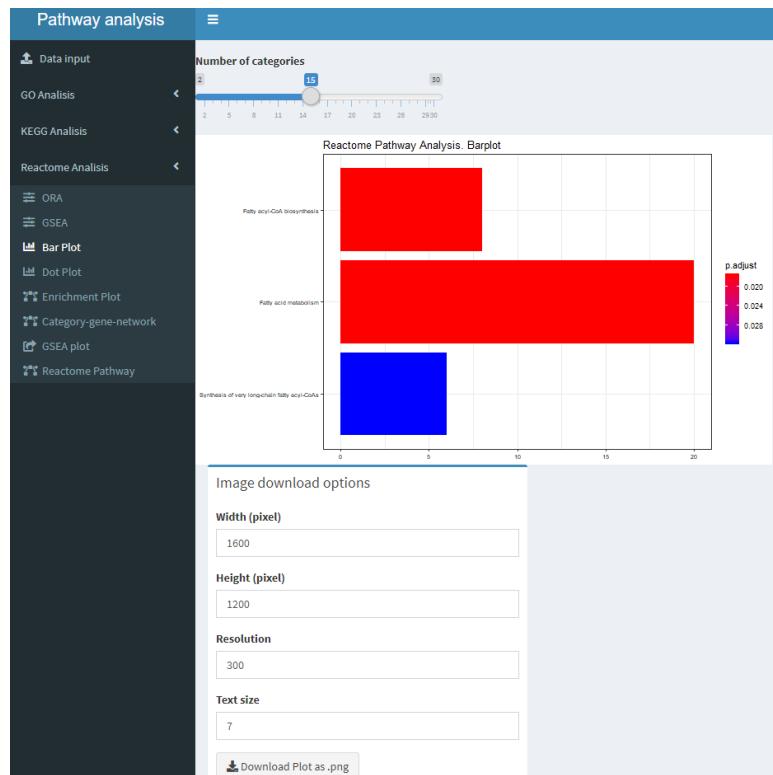


Figure 33: Gràfic de barres

- Selecciono *Reactome Analysis*→*Dot Plot*

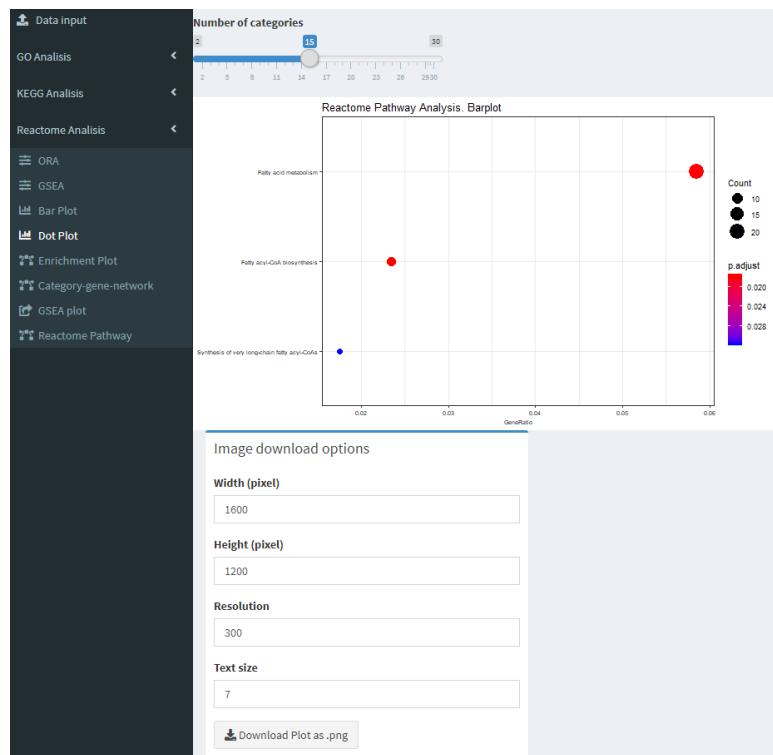


Figure 34: Gràfic de punts

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

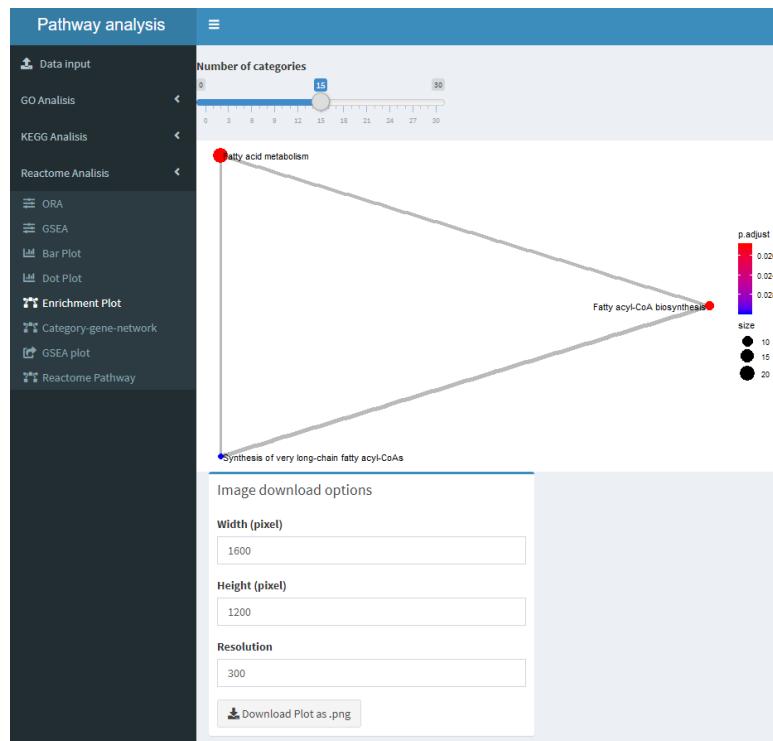


Figure 35: Mapa d'enriquement

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

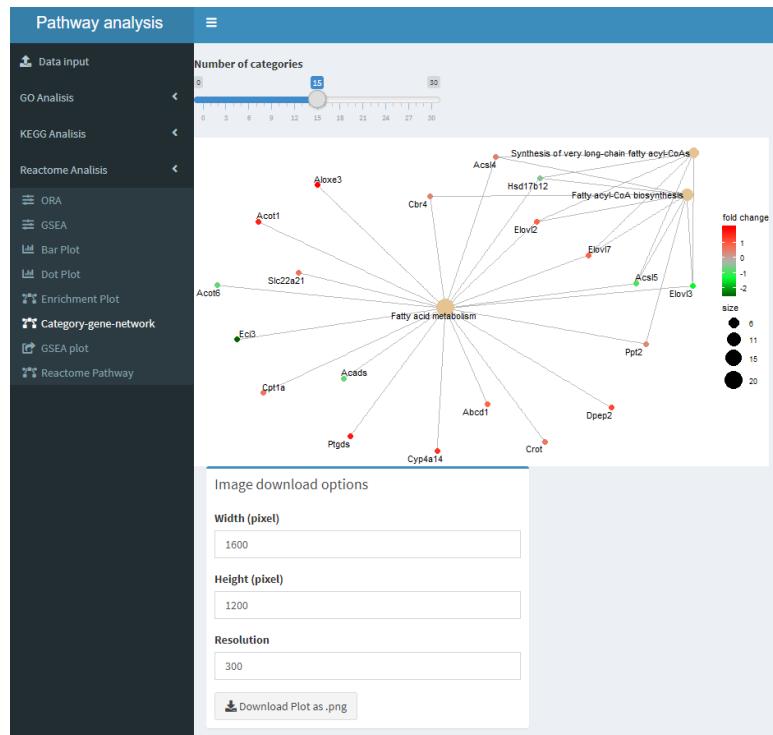


Figure 36: Red de les categories i gens

- Selecciono *Reactome Analysis*→*Reactome Pathway*

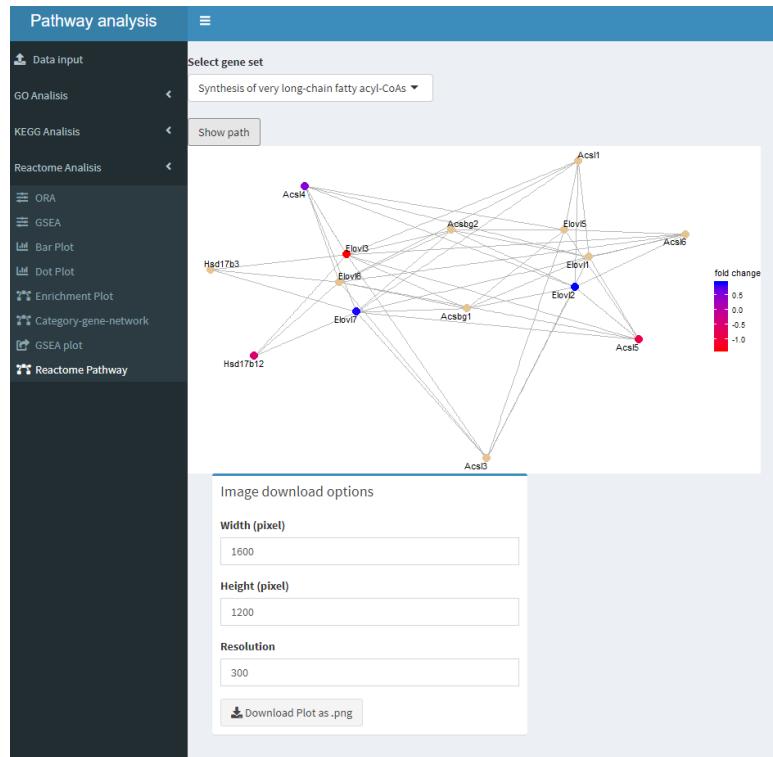


Figure 37: Rutes Reactome

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.

Pathway analysis																																																																		
Data input		Select gene set																																																																
GO Analysis		Synthesis of very long-chain fatty acyl-CoAs																																																																
KEGG Analysis		Show path																																																																
Reactome Analysis		ORA GSEA Bar Plot Dot Plot Enrichment Plot Category-gene-network GSEA plot Reactome Pathway																																																																
		Select P-Value threshold: <input checked="" type="radio"/> 0.1 <input type="radio"/> 0.05 <input type="radio"/> 0.01 <input type="radio"/> 0.001																																																																
		Select adjustment method: BH																																																																
		Calculate Result																																																																
		Download Results as .csv																																																																
		<table border="1"> <thead> <tr> <th>ID</th><th>Description</th><th>setSize</th><th>enrichmentScore</th><th>NES</th><th>pvalue</th><th>p.adjust</th><th>qvalues</th><th>rank</th><th>leading_edge</th><th>core_enrichment</th></tr> </thead> <tbody> <tr> <td>R-MMU-69242</td><td>S Phase</td><td>43</td><td>-0.528</td><td>-1.851</td><td>0.002</td><td>0.000</td><td>0.032</td><td>1968</td><td>tags=53%, list=33%, signal=36%</td><td>22190/18971/109145</td></tr> <tr> <td>R-MMU-8852276</td><td>The role of GTF2E1 in G2/M progression after G2 checkpoint</td><td>27</td><td>-0.592</td><td>-1.871</td><td>0.002</td><td>0.000</td><td>0.032</td><td>614</td><td>tags=27%, list=10%, signal=33%</td><td>15516/16912/23996/2</td></tr> <tr> <td>R-MMU-174143</td><td>APC/C-mediated degradation of cell cycle proteins</td><td>26</td><td>-0.669</td><td>-2.100</td><td>0.002</td><td>0.000</td><td>0.032</td><td>703</td><td>tags=42%, list=12%, signal=38%</td><td>66413/12237/16912/2</td></tr> <tr> <td>R-MMU-453276</td><td>Regulation of mitotic cell cycle</td><td>26</td><td>-0.669</td><td>-2.100</td><td>0.002</td><td>0.000</td><td>0.032</td><td>703</td><td>tags=42%, list=12%, signal=38%</td><td>66413/12237/16912/2</td></tr> </tbody> </table>										ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrichment	R-MMU-69242	S Phase	43	-0.528	-1.851	0.002	0.000	0.032	1968	tags=53%, list=33%, signal=36%	22190/18971/109145	R-MMU-8852276	The role of GTF2E1 in G2/M progression after G2 checkpoint	27	-0.592	-1.871	0.002	0.000	0.032	614	tags=27%, list=10%, signal=33%	15516/16912/23996/2	R-MMU-174143	APC/C-mediated degradation of cell cycle proteins	26	-0.669	-2.100	0.002	0.000	0.032	703	tags=42%, list=12%, signal=38%	66413/12237/16912/2	R-MMU-453276	Regulation of mitotic cell cycle	26	-0.669	-2.100	0.002	0.000	0.032	703	tags=42%, list=12%, signal=38%	66413/12237/16912/2
ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrichment																																																								
R-MMU-69242	S Phase	43	-0.528	-1.851	0.002	0.000	0.032	1968	tags=53%, list=33%, signal=36%	22190/18971/109145																																																								
R-MMU-8852276	The role of GTF2E1 in G2/M progression after G2 checkpoint	27	-0.592	-1.871	0.002	0.000	0.032	614	tags=27%, list=10%, signal=33%	15516/16912/23996/2																																																								
R-MMU-174143	APC/C-mediated degradation of cell cycle proteins	26	-0.669	-2.100	0.002	0.000	0.032	703	tags=42%, list=12%, signal=38%	66413/12237/16912/2																																																								
R-MMU-453276	Regulation of mitotic cell cycle	26	-0.669	-2.100	0.002	0.000	0.032	703	tags=42%, list=12%, signal=38%	66413/12237/16912/2																																																								

Figure 38: Anàlisi GSEA

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

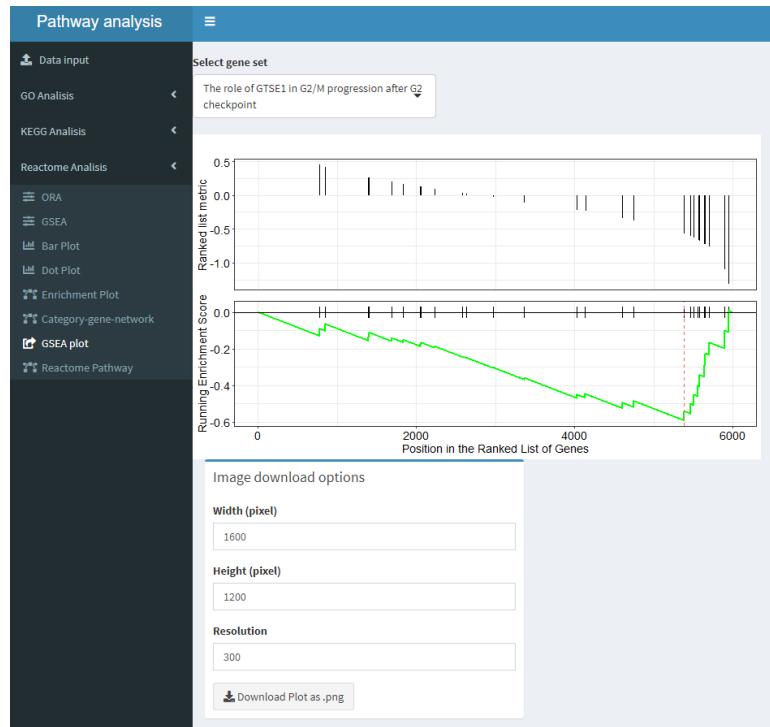


Figure 39: Gràfic GSEA

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:

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*

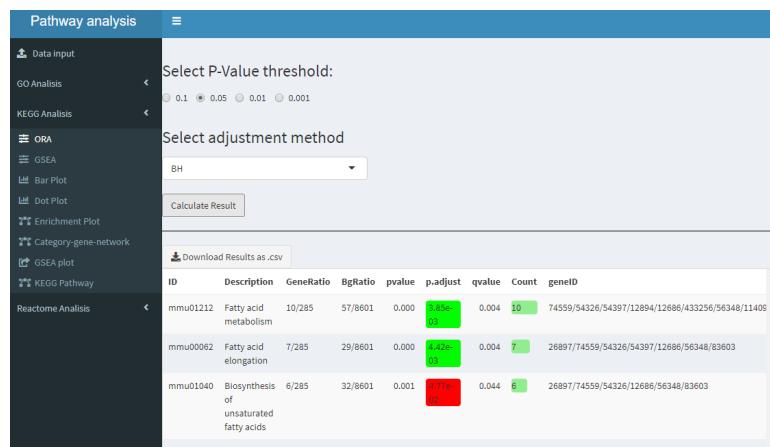


Figure 40: Anàlisi ORA de KEGG

2. Anem a *KEGG*→*KEGG Pathway*

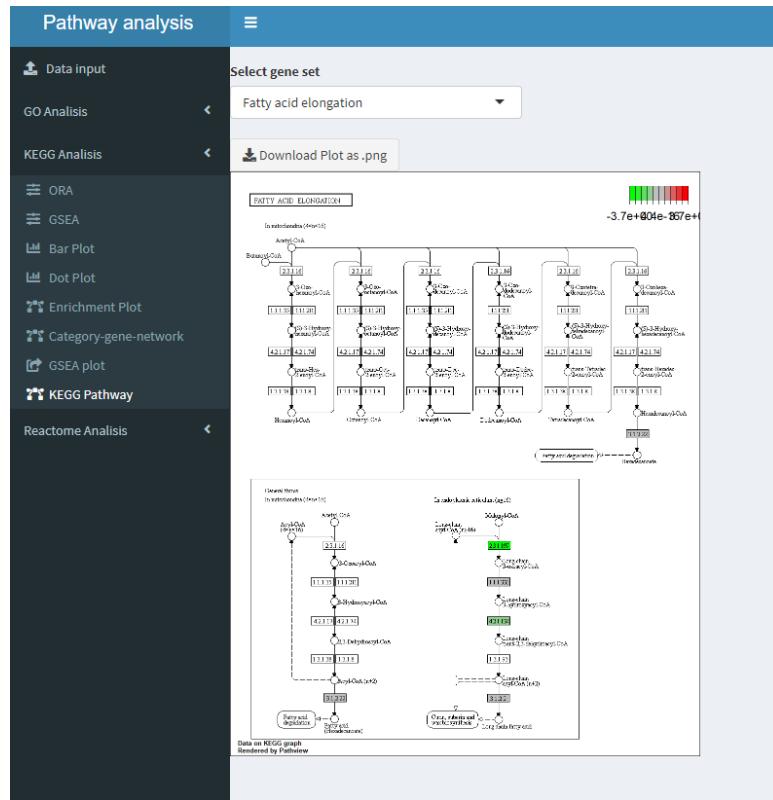


Figure 41: Gràfic de les rutes KEGG

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*

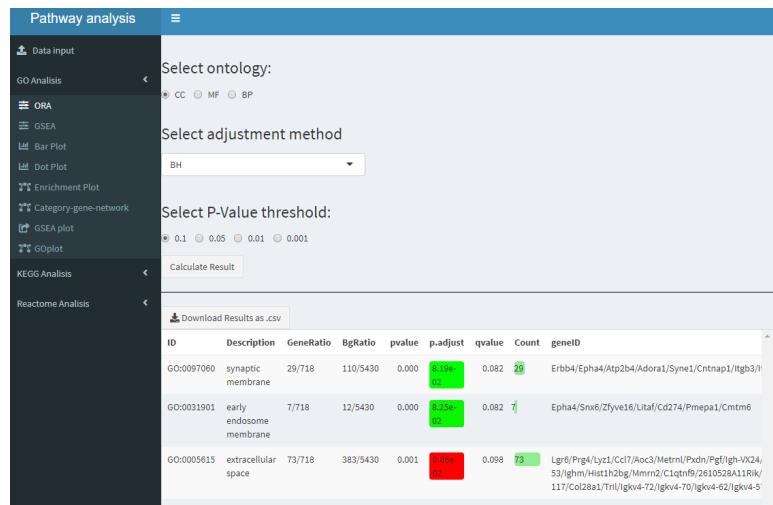


Figure 42: L'anàlisi ORA de GO

10 Discussió

Bibliografia

- [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.
- [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 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.