



Universitat Oberta
de Catalunya

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

Vasyl Druchkiv

Programa de Master de Bioinformàtica i Bioestadística
Àrea del treball final: Estadística i Bioinformàtica

Consultor: Alex Sánchez-Pla

Professor responsable de l'asignatura: Alex Sánchez-Pla

Data d'entrega: 05/06/2017

Copyright © 2019 Vasyl Druchkiv.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

A copy of the license is included in the section entitled "GNU Free Documentation License".

FITXA DEL TREBALL FINAL

Títol del treball:	Implementació d'una eina en R/Shiny per a l'anàlisi de significació biològica utilitzant l'anàlisi de les rutes
Nom d'autor:	Vasyl Druchkiv
Nom del consultor:	Alex Sánchez-Pla
Nom del PRA:	Alex Sánchez-Pla
Data d'entrega (mm/aaaa):	06/2019
Titulació:	<i>Bioinformàtica I bioestadística</i>
Àrea del Treball Final:	<i>Anàlisi de dades òmiques</i>
Idioma del treball:	Català
Paraules clave	<i>Pathway analysis, R, Shiny</i>
Resum del Treball (màxim 250 palaules): Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball.	
L'objectiu de teball és trobar paquets de Bioconductor i desenvolupar una aplicació Shiny per dur a terme l'anàlisi de les rutes (<i>Pathway analysis</i>). Un <i>Pathway</i> és el conjunt de gens relacionats amb una funció biològica i descriu la relació entre els gens. Primer s'identifiquen els mètodes teòrics actualment presents per trobar i visualitzar les rutes diferencialment expressades. També s'identifiquen les bases de dades per anotar les rutes i els paquets de Bioconductor específics per fer l'anàlisi. D'aquesta manera es crea una aplicació Shiny que ofereix l'anàlisis ORA (Over-representation Analysis), GSEA (Gene Set Enrichment Analysis) i l'anàlisi de la topografia de les rutes. Els paquet elegits per a anàlisi de les rutes de Bioconductor són: <i>clusterProfiler</i> , <i>ReactomePA</i> i <i>patview</i> . L'anotació de gens es fa finalment via tres bases de dades: GO (Gene Ontology), KEGG (Kyoto Encyclopedia of Genes and Genomes) i Reactome.	
L'usuari pot seleccionar les ontologies de GO, especificar el nivell de significació i el mètode d'ajustament. S'obté per cada base de dades una taula per a l'anàlisi ORA i l'altra per GSEA. Els resultats es visualitzen via Bar plot, Dot plot, enrichment map, gene-concept network, GO plot, KEGG pathway i Reactome pathway. Es fa possible la descàrrega de les taules en format de .csv i la de les imatges en format .png. L'usuari pot especificar la resolució i altres atributs de les imatges.	
La creació de l'aplicació ve seguida per la seva validació, on es presenta la seva funcionalitat i es compara el resultat amb l'estudi original.	

Abstract (in English, 250 words or less):

Objective of the thesis is to find Bioconductor packages for pathway analysis and implement them using Shiny. A pathway is a set of genes related to a specific biological function and describes a relation between those genes. First, I identify theoretical methods available now for finding and visualising the differentially expressed pathways. Furthermore, I identify data bases for gene annotation and specific Bioconductor packages for performing analysis. Second, I develop Shiny application which offers ORA (Over-representation Analysis), GSEA (Gene Set Enrichment Analysis) and topologic analysis of the pathways. The packages selected for the analysis are *clusterProfiler*, *ReactomePA* i *patview*. Gene annotation is done using three data bases: GO (Gene Ontology), KEGG (Kyoto Encyclopaedia of Genes and Genomes) and Reactome.

User can select GO ontologies, specify significance level and adjustment method. For each annotation data base a table with results is generated. Additionally, results are visualized with bar plot, dot plot, enrichment map, gene-concept network, GO plot, KEGG pathway and Reactome pathway. It is possible to download results as .csv files and images as .png files. User can customize image resolution and other image attributes.

Development of the application is followed up with its validation, where its functionality is shown, and its results are compared to original study.

Agraïment

M'agradaria donar les gràcies a la meva dona i a la meva filla per acompañar-me en aquesta aventura que és el Màster. A més a més agraeixo especialment a la meva dona, Rosa de les Neus, l'assessorament lingüístic. Gràcies a tu el meu Català s'apropa a l'excellència.

I també agraeixo al meu consultor de la UOC, Alex Sánchez-Pla, la seva disponibilitat i acurats consells.

Contingut

1	Introducció	1
1.1	Context i justificació del treball	2
1.2	Objectius	2
1.2.1	Objectius generals	2
1.2.2	Objectius específics	3
1.3	Enfocament i mètode a seguir	3
1.4	Planificació del treball	4
1.5	Breu sumari dels productes obtinguts	5
1.6	Breu descripció dels capítols del treball	6
2	Marc teòric	8
2.1	Dades d'expressió genètica	8
2.2	Annotació dels gens	12
2.2.1	Gene ontology	12
2.2.2	KEGG	13
2.2.3	Reactome	14
2.3	ORA	14
2.4	GSEA	16
2.5	Analisi topològic de les rutes	18
2.5.1	El mapa d'enriquement	18
2.5.2	Gene-Concept-Network	19
2.5.3	GO-Plot	19
2.5.4	KEGG Pathway	20
2.5.5	Reactome Pathway	21
2.6	Desenvolupament del protocol	22
3	Tractament bioinformàtic	25
3.1	Cerca dels paquets de Bioconductor	25
3.2	Instal·lació de l'aplicació	26

Contingut

4 L'aplicació	29
4.1 ORA	33
4.1.1 GO	33
4.1.2 KEGG	34
4.1.3 Reactome	36
4.2 GSEA	37
4.2.1 GO	37
4.2.2 KEGG	38
4.2.3 Reactome	39
4.3 Visualització i l'anàlisi topològic	40
4.3.1 Bar-Plots	40
4.3.2 Dot-Plots	41
4.3.3 Enrichment Maps	43
4.3.4 Category-Gene-Network Plot	44
4.3.5 GSEA Plot	44
4.3.6 GO Plot	46
4.3.7 KEGG Pathway	47
4.3.8 Reactome Pathway	48
4.4 Manual i ajudes del programa	48
5 Validació dels resultats	54
5.1 Exemple d'anàlisi 1. GEO: GSE100924	55
6 Conclusions	67
Glossari	69
Bibliografia	71

Llista de les imatges

1.1	Gantt Plot	5
2.1	El procés d'anàlisi de microarrays.	9
2.2	El procés d'anàlisi de les rutes.	10
2.3	El mètode GSEA	17
2.4	L'anotació de les relacions dins de les rutes KEGG	21
2.5	Lucidchart per a l'aplicació	23
4.1	Pàgina d'entrada	30
4.2	Resum de les dades pujades	31
4.3	Els elements de les seccions d'anàlisi	33
4.4	Especificació d'ORA dels termes GO	34
4.5	El resultat d'anàlisi ORA. GO.	35
4.6	Configuració d'anàlisi KEGG	35
4.7	El resultat de l'anàlisi ORA. KEGG.	36
4.8	El resultat d'anàlisi ORA. Reactome.	37
4.9	El resultat de l'anàlisi GSEA. GO.	38
4.10	El resultat de l'anàlisi GSEA. KEGG.	39
4.11	El resultat d'anàlisi GSEA. Reactome.	40
4.12	Bar-Plot. GO.	41
4.13	Dot-Plot. GO.	42
4.14	Enrichment Map. GO.	43
4.15	Category-Gene-Network Plot. GO.	44
4.16	GSEA Plot. GO.	45
4.17	GO Plot	46
4.18	KEGG pathway	47
4.19	Reactome pathway	48
4.20	Manual per a aplicació	49
4.21	Manual per a l'anàlisi ORA amb l'anotació KEGG	49

Llista de les imatges

4.22 Senyals d'ajuda	50
4.23 Ajuda per a l'elecció de l'espècie	51
4.24 Ajuda per pujar les dades	51
4.25 Infromació per la interpretació d'anàlisi ORA	52
4.26 Ajuda per la selecció del mètode d'ajustament	52
4.27 Ajuda per la interpretació de GSEA	53
5.1 Selecció d'espècie	56
5.2 Breu resum de les dades	57
5.3 Resultat d'anàlisi ORA de Reactome	58
5.4 Gràfic de barres	59
5.5 Gràfic de punts	60
5.6 Mapa d'enriquiment	61
5.7 Red de les categories i gens	62
5.8 Rutes Reactome	63
5.9 Anàlisi GSEA	64
5.10 Gràfic GSEA	64
5.11 Anàlisi ORA de KEGG	65
5.12 Gràfic de les rutes KEGG	65
5.13 L'anàlisi ORA de GO	66

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 procés 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 proves no a nivell dels gens particulars sinó a 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 hi ha 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.

1 Introducció

1.1 Context i justificació del treball

La justificació d'aquest tema ve de dues fonts diferents: d'una banda tinc un interès personal en aquest tema, i d'altra banda entenc la potencial 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 entenent que aquesta eina és útil per al meu desenvolupament professional he buscat el tema que en requeria l'ús. Encara que hi ha algunes aplicacions de Shiny relacionades amb l'anàlisi de les rutes¹ i també en altres plataformes [Reimand et al., 2019], elles no representen tota la diversitat dels paquets disponibles a Bioconductor. L'ús d'aquests paquets queda restringit per a experts en informàtica i estadística i per tant són difícilment accessibles per la gran part de la comunitat científica, de manera que seria convenient donar-hi més accessibilitat via una aplicació amb interfície visual.

1.2 Objectius

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

1.2.1 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)

¹Algunes aplicacions existents de Shiny són: iDINGO [Class et al., 2017], ShinyGO [Ge and Jung, 2018], PAEA [Clark et al., 2015]

1.2.2 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 l'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

1.3 Enfocament i mètode a seguir

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 al mateix moment de l'anàlisi dels conceptes teòrics. D'aquesta manera he seguit aquests passos:

1. Trobar un mètode teòric que proporcioni un resultat interessant;
2. Buscar en Bioconductor aquest mètode;

1 Introducció

3. Repetir 1 i 2 fins que el conjunt dels mètodes facin l'anàlisi de les rutes completa.
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'utilitzessin 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.

1.4 Planificació del treball

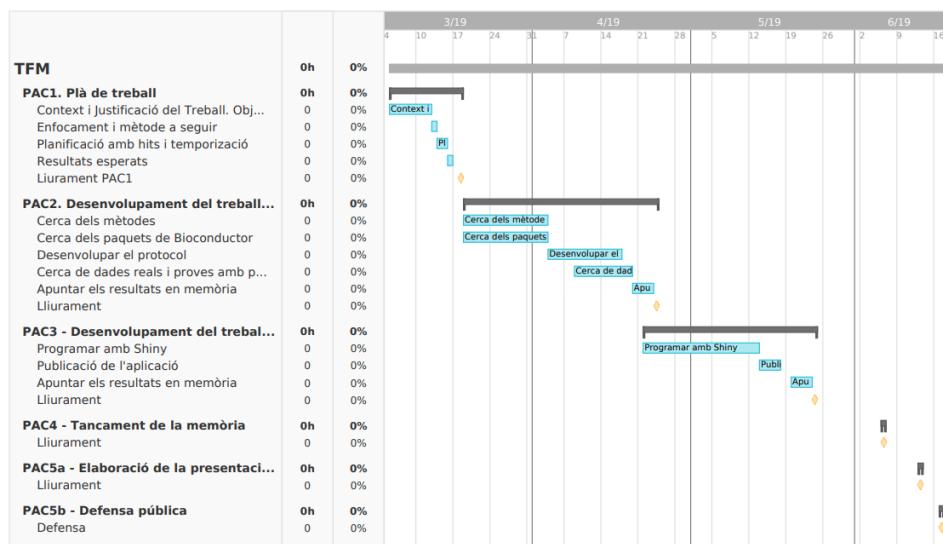
Es van definir les tasques següents per aconseguir els objectius:

1. Cerca de la literatura sobre els mètodes de l'anàlisi de les rutes;
2. Relacionar els mètodes trobats en 1 amb els paquets actuals de Bio-conductor;
3. Decidir sobre quins resultats són més interessants per a una aplicació Shiny i desenvolupar un protocol de l'anàlisi (*pipeline*) que formarà la base de l'aplicació. Documentar el protocol;
4. Buscar 3-5 exemples de dades i fer proves aplicant el protocol i comparant els resultats amb els resultats publicats sobre aquestes dades (si n'hi ha);
5. Fer els últims canvis en el protocol;
6. Dissenyar i programar l'aplicació de Shiny;
7. Publicar l'aplicació en web;
8. Tancar la memòria i fer la presentació per a la defensa.

Aquestes tasques les he distribuït de manera següent:

1 Introducció

Imatge 1.1: Gantt Plot



Els treballs previstos pel pla docent i realitzats i entregats a temps eren els següents:

Activitat	Nom d'activitat	Data d'inici	Data d'entrega
PAC0	Definició dels continguts del treball	20/02/19	04/03/2019
PAC1	Pla de treball	05/03/19	18/03/19
PAC2	Desenvolupament del treball - Fase 1	19/03/19	4/04/19
PAC3	Desenvolupament del treball - Fase 2	25/04/19	20/05/19
PAC4	Tancament de la memòria	21/05/19	05/06/19

1.5 Breu sumari dels productes obtinguts

El producte central obtingut és l'aplicació Shiny que actualment es pot descarregar del meu repositori a Github (veure l'apartat instal·lació de l'aplicació). Al mateix repositori es pot trobar tot el material relacionat amb la creació de l'aplicació: els arxius de latex per a les proves d'avaluació

1 Introducció

continuada, les captures de pantalla utilitzades en aquestes PACs, també el paquet modificat (pathview) que es diu pathviewPatched i que permet guardar les imatges de les rutes al directori especificat.

1.6 Breu descripció dels capítols del treball

Al primer capítol més teòric presentaré els mètodes més rellevants per a l'anàlisi de les rutes. Començaré però amb el context d'investigació de l'expressió genètica i més específicament amb l'anàlisi de microarrays per poder situar millor l'anàlisi biològica de les rutes. Donades les dades de l'experiment explicaré breument quins resultats habitualment se'n deriven. Veurem que aquests resultats són la llista de gens diferencialment expressats i la magnitud d'expressió diferencial mesurada via *logRatio* per tots els gens.

Seguidament investigaré quines estratègies existeixen per dur a terme l'anàlisi de les rutes. Aquí identificaré tres estratègies generals: l'anàlisi ORA (Over-Representation Analysis), FCS (Functional Class Scoring) i l'anàlisi topològica de les rutes. Veurem que aquests análisis accepten com a dades d'entrada els resultats obtinguts via *microarrays* descrits anteriorment. Per poder fer ús d'aquestes dades però és necessari anotar els gens de l'experiment. Per aquest motiu presentaré les tres bases de dades més rellevants per anotar els gens amb l'objectiu de fer l'anàlisi de les rutes: GO, KEGG i Reactome.

Finalment parlaré sobre els mètodes ORA i GSEA més detalladament concentrant-me especialment en els resultat que proporcionen. També descriure breument les visualitzacions possibles de les rutes. Per acabar presentaré el protocol que utilitzaré per crear l'aplicació.

Tot seguit explicaré els paquets del Bioconductor per dur a terme l'anàlisi descrita al capítol anterior i la manera amb la qual he fet l'aplicació accessible. Veurem que al moment de redacció d'aquesta memòria l'aplicació pot ser descarregada de GitHub utilitzant el paquet *devtools* de R. S'intentarà però habilitar l'aplicació per a internet utilitzant el servidor de l'àrea d'estadística i bioinformàtica.

1 Introducció

A continuació presentaré l'aplicació. Aquí utilitzaré les dades del paquet `clusterProfiler` per mostrar el funcionament de l'aplicació: des de la pujada de les dades fins a l'obtenció de les representacions gràfiques.

Després intentaré validar l'aplicació. Parlaré dels problemes a l'hora de trobar les dades preprocessades i presentaré les dades amb les que en faig l'intent i em concentraré més en l'estudi de l'expressió gènica dels ratolins modificats genèticament de tal manera que el seu gen `Zbtb7b` està silenciat. Veurem que els resultats obtinguts amb l'aplicació s'assemblen als resultats descrits al *paper* original.

Per acabar redactaré les conclusions del meu treball.

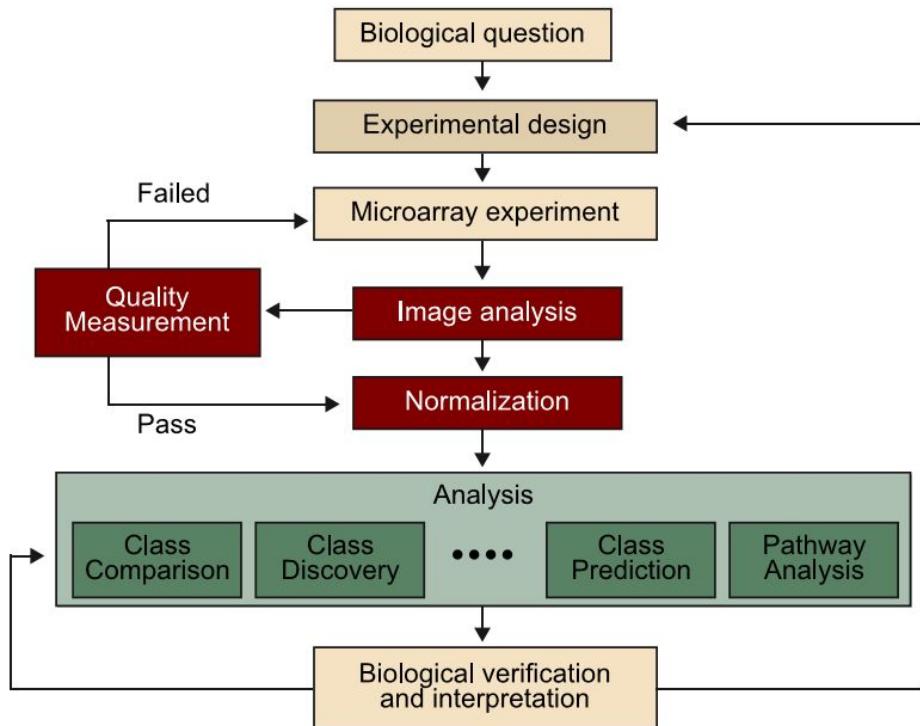
2 Marc teòric

2.1 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 influeix 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 silenciat i investigar quins gens són diferencialment expressats entre aquestes mostres biològiques.

Al gràfic següent veiem l'estructura habitual d'un experiment de *Microarray* [Ruíz de Villa and Sánchez-Pla, 2019]:

2 Marc teòric



imatge 2.1: El procès d'anàlisi de microarrays.

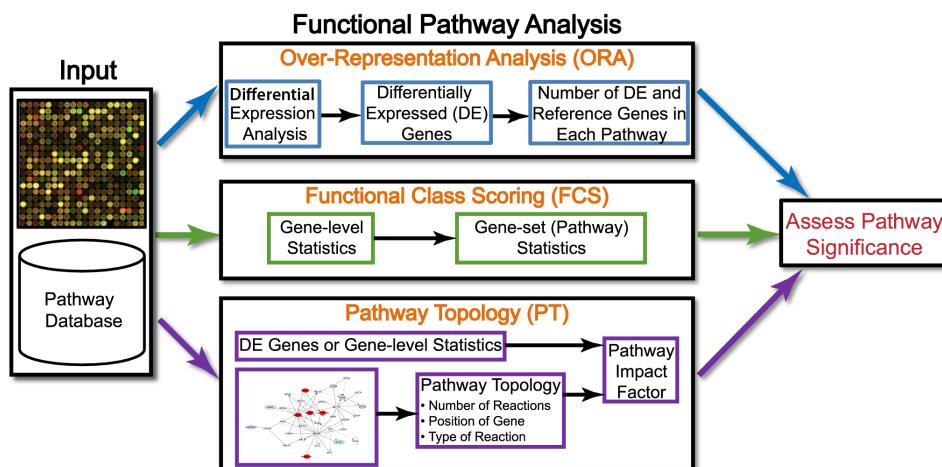
El *pipeline* de l'anàlisi consisteix doncs en el plantejament d'una pregunta i un disseny experimental a partir del qual es fa l'experiment de *microarrays*. Els productes de l'experiment són bàsicament les imatges d'intensitats que es tradueixen als valors numèrics. Habitualment aquests valors són encara *raw values* i han de ser processats adequadament. Aquest processament inclou el control de qualitat de les imatges i la normalització dels valors d'intensitat per reduir la variabilitat tècnica. Finalment les dades normalitzades s'utilitzen per a l'anàlisi estadística. Habitualment la mesura natural per comparar les mostres és el *logRatio* el qual podem denominar alternativament *logFC* on *FC* es refereix a *fold change*. Hi ha diversos tests estadístics per comprovar les diferències entre les mostres. En el cas de l'array d'un color podem fer servir tant els mètodes paramètrics -com ara el test T o els mètodes del modelatge lineal-, com els mètodes no-paramètrics -com

2 Marc teòric

ara la prova de Mann-Whitney. El test adequat depèndrà bàsicament de la distribució de les dades. El resultat d'aquest anàlisi serveix com a base per a la interpretació biològica dels resultats de l'experiment. Per poder donar sentit a les dades d'expressió i de l'anàlisi estadístic al nivell de gens és imprescindible fer una anàlisi a nivell de les categories de gens o les rutes. Per aquesta 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 de l'experiment descrit, més específicament de l'anàlisi de rutes (*Pathway analysis*).

La vista general de l'anàlisi de les rutes ofereix el gràfic següent [Khatri et al., 2012]:



Imatge 2.2: 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 comprovar si els gens sobre o sotaexpressats pertanyen a unes rutes específiques. Per comprovar aquesta, o millor dit, aquestes hipòtesis (per que hi haurà hipòtesis múltiples) s'han establert tres grups de mètodes:

2 Marc teòric

- **Over-Representation Analysis (ORA).** Aquesta 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 ([Boyle et al., 2004]).
- **Functional Class Scoring (FCS)** Per a aquesta anàlisi no necessitem cap preselecció dels gens diferencialment expressats (DE) sinó ja és suficient amb tenir les estadístiques a nivell de gens, que al cas de l'aplicació és el *logFC*. Hi ha diversos mètodes que generen una estadística per a tot el conjunt de gens d'una ruta i la comparen amb una distribució teòrica per a contrastar la hipòtesi nul·la. Els mètodes es diferencien bàsicament en el càlcul de la puntuació d'enriquiment que poden incloure l'estadística de Kolmogorov-Smornov, la suma, mitjana o mediana d'estadístiques al nivell de gens etc. [Khatri et al., 2012]. O bé es poden diferenciar en el càlcul de la distribució teòrica: aquí alguns mètodes utilitzen la permutació de les mostres o de gens, cosa que implica dues hipòtesis diferents.
- **Pathway Topology (PT).** Aquest mètode enfoca la posició dels 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 específicament, 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 ([Tarcia et al., 2008]). D'altra banda, si un nombre de gens està involucrat en la ruta però apareixen riu abaix 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'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.

En els capítols següents precisaré més formalment els mètodes d'ORA i GSEA i també descriuré algunes possibilitats per visualitzar les dependències de gens dins de les rutes específiques i també les relacions

entre les rutes diferencialment expressades.

2.2 Anotació dels gens

Com veurem més endavant per a l'anàlisi de les rutes és imprescindible tenir com a referència les anotacions dels gens. Per a l'aplicació he utilitzat tres bases de dades: Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) i Reactome. clusterProfiler també inclou WikiPathways però per raons de temps he decidit centrar-me només en les tres mencionades anteriorment. Aquestes bases de dades no són però úniques. N'hi ha també altres com per exemple *Pathway Studio pathways* o *IPA* amb l'inconvenient que són comercials i no gratuïtes.

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 les activitats a nivell molecular. És important entendre que el terme “molecular function” representa més les activitats que no pas les entitats (com per exemple molècules o complexos) que fan aquestes accions, i a més a més no especificuen 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 de les funcions moleculars o un conjunt d'elles. Un exemple d'un procés biològic ampli podria ser *DNA repair*. Un exemple més específic podria ser *pyrimidine nucleobase biosynthetic process*.

2 Marc teòric

3. **Cellular Component (CC).** EL CC descriu l'emplaçament 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 termes 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 dels mapes dibuixats 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 dibuixats 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.

2.2.3 Reactome

Reactome és una base de dades gratuïtament accessible i manualment curada per a reaccions i rutes biològiques. Al centre de Reactome hi ha reaccions que es defineixen com qualsevol esdeveniment molecular com ara unió, fosforització, catalisi bioquímic, transport molecular o esdeveniments moleculars espontanis. 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 paquet de Bioconductor que utilitzaré per generar els mapes visualitza només la connexió 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 *logRatios* i/o en el valor de p provenint d'un test estadístic. *logRatios* donen la magnitud amb la qual un gen és sobre o sotaexpressat. Les diferències entre els grups però són el resultat d'un procés estocàstic 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 Marc teòric

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 utilitzà 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 fons 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.

Debilitats d'aquest mètode són les següents:

- Les estadístiques utilitzades pel mètode ORA, com ara la distribució hipergeomètrica, són independents dels canvis mesurats. Això vol dir que aquests tests ignoren tots els valors associats amb ells com ara les intensitats (logRatios).
- Típicament ORA utilitzà només els gens més significatius i descarta els altres.
- ORA tracta cada gen igual i per tant assumeix que tots els gens són independents els uns dels altres.
- ORA assumeix que totes les rutes són independents l'una de l'altra.

2.4 GSEA

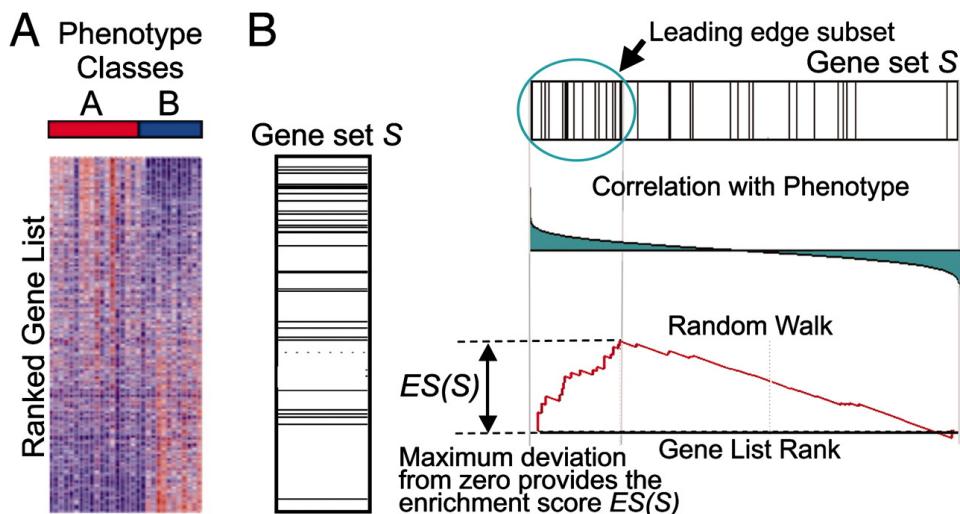
Amb l'anàlisi GSEA podem analitzar els resultats d'un experiment d'expressió per a dos grups. Aquí els gens són ordenats basant-se en la correlació entre la seva expressió i la separació entre les classes. Aquest llistat ordenat L el podem crear utilitzant els *logRatios*.

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

L'anàlisi GSEA consisteix en tres passos [Subramanian et al., 2005]:

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 aquet 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. Els dos modes de permutació comproven hipòtesis diferents. Mentre la permutació de gens comprova la hipòtesi que *els gens en la ruta com a màxim són diferencialment expressats com els gens fora de la ruta*, la permutació de les mostres implica la hipòtesi que cap de gens en la ruta són diferencialment expressats. Es diferencia doncs en el tractament dels gens fora de la ruta. A l'aplicació es fa ús de la permutació dels gens a causa bàsicament de la selecció dels paquets de Bioconductor per a l'aplicació.
3. Càlcul del valor de P ajustat. El valor de P nominal s'ajusta per controlar l'error global que es produueix com a resultat de les comparacions múltiples.

2 Marc teòric



Imatge 2.3: 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 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 ocurredades 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.

2 Marc teòric

- 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 es produeix exactament el pic.
- 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.

Encara que GSEA és una millora respecte a ORA, aquest mètode també té les seves debilitats:

- Tal com l'anàlisi ORA, també GSEA assumeix que les rutes són independents.
- GSEA utilitza els canvis d'expressió gènica dins de la ruta específica i descarta els canvis d'altres àrees. Per exemple, agafem dos gens A i B, que canvien dues vegades l'un (2-fold) i 20 vegades l'altre (20-fold) respectivament. Fins que aquests gens tinguin el mateix rang comparat amb altres gens de la ruta, GSEA els tractarà igual, encara que seria preferible donar més pes al gen B.

2.5 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.

2.5.1 El mapa d'enriquement

L'anàlisi ORA resulta en una llista de les rutes significativament enriquides. Els gens dins dels conjunts o les rutes poden encavalcar i descriuen gairebé

2 Marc teòric

conceptes biològics idèntics [Merico et al., 2010]. Aquest problema de redundància és més evident als conceptes que són organitzats jeràrquicament, com és el cas dels conceptes de la base da dades GO. El mapa d'enriquiment redueix la redundància als conjunts de gens. Els conjunt de gens estan representats com a nodes els radis dels quals són proporcionalment relacionats amb el nombre de gens que formen part d'aquests conjunts. Els cantells indiquen els nodes que tenen gens compartits, on el seu gruix depèn del nombre de gens compartits. A més a més, es pot utilitzar el color dels nodes per representar una altra dimensió com ara el nivell de significació expressat per el valor de P. Si no hi ha cap gen compartit entre els conceptes (o rutes) els nodes no són connectats via cantells. Aquest mètode de representació és molt útil per poder reduir/simplificar la informació obtinguda mitjançant els mètodes ORA o GSEA i per tant m'he posat l'objectiu també d'implementar-lo a l'aplicació.

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 rutes (conceptes) diferents. D'aquesta manera es fa possible identificar les associacions biològiques més complexes entre les rutes mitjançant els gens.

2.5.3 GO-Plot

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 subtip de B. Per exemple el cicle mitòtic de la cèl·lula *is a* cicle de la cèl·lula.

2 Marc teòric

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 procés afecta directament la manifestació de l'altre procés.

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

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:

2 Marc teòric

Notation	Objects	Arrows	
	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		[] → e → [] expression
	-U → [] deubiquitination		[] → e → [] 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

2.5.5 Reactome Pathway

Les rutes de Reactome són similares a les rutes de KEGG. La seva implementació amb Bioconductor, com ho veurem properament, no visualitza les rutes originals mostrades per [Pathway Browser](#) de Reactome. La visualització amb el paquet ReactomePA és més modesta i ofereix només les relacions nominals sense mostrar direccionalitat, com ho fa el gràfic de la ruta original de Reactome. Tot i així podem identificar quantes connexions amb altres gens de la ruta tenen els gens diferencialment expressats i

2 Marc teòric

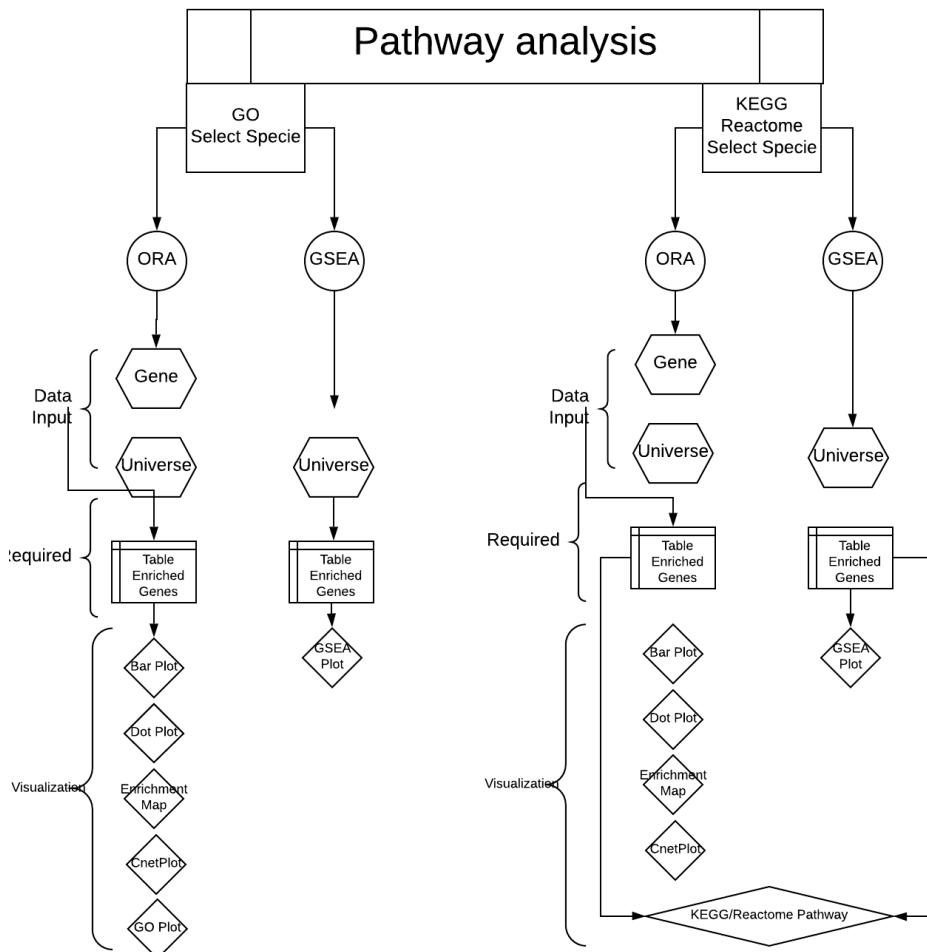
d'aquesta manera intuir la seva importància relativa.

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

2.6 Desenvolupament del protocol

Tenint en compte el marc teòric he intentat dissenyar l'aplicació de tal manera que ofereixi els mètodes ORA, GSEA i algunes visualitzacions de la topografia de les rutes. Tan se val amb quina base de dades l'usuari anoti les dades d'expressió, ja que l'usuari podrà decidir quin mètode vol aplicar. En el millor dels casos l'usuari faria els dos mètodes ORA i GSEA. Això implicaria la disponibilitat de dos arxius: l'arxiu amb tots els gens (Universe) i l'arxiu amb el grup de gens diferencialment expressats (Gene set). Si l'usuari vol fer només l'anàlisi GSEA haurà de pujar només l'arxiu amb tots els gens. Una vegada seleccionada l'estrategia l'usuari puja els arxius necessaris i genera el resultat de ORA i/o GSEA aplicant uns criteris com ara selecció d'ontologia en ORA, valor de P com al filtre de visualització de les rutes més significatives, i el mètode d'ajustament.

2 Marc teòric



imatge 2.5: 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 anàlisi GO tots dos arxius són necessaris: Gens Seleccions (Gene) i Tots els gens (Universe).

2 Marc teòric

- 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 vulguem fer l'anàlisi 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 vulguem 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 Tractament bioinformàtic

3.1 Cerca dels paquets de Bioconductor

El Bioconductor ofereix molts paquets per dur a terme l'anàlisi de les rutes implementant algoritmes diferents a l'hora de calcular les estadístiques de les anàlisis ORA i GSEA. La cerca s'ha reduït a tres paquets principals: `clusterProfiler`, `ReactomePA` i `pathview`. D'aquests tres paquets el paquet `clusterProfiler` és el més complet i 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 i GSEA. També inclou les possibilitats de visualització dels resultats suficients per considerar l'anàlisi de les rutes completa. 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].

Els paquets i les seves funcions per generar els resultats són els següents:

Base de dades	Mètode	Paquet Bioconductor	Funció	Observació
GO	ORA	<code>clusterProfiler</code>	<code>enrichGO()</code>	Només 7 espècies disponibles
GO	GSEA	<code>clusterProfiler</code>	<code>gseGO()</code>	Permutació de gens
GO	Bar-Plot	<code>enrichplot</code>	<code>barplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
GO	Enrichment Map	<code>enrichplot</code>	<code>emapplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
GO	Gene-Concept-Network	<code>enrichplot</code>	<code>cnetplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
GO	GO directed acyclic graph	<code>enrichplot</code>	<code>goplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
KEGG	ORA	<code>clusterProfiler</code>	<code>enrichKEGG()</code>	Totes les espècies de KEGG
KEGG	GSEA	<code>clusterProfiler</code>	<code>gseKEGG()</code>	Permutació de gens
KEGG	Bar-Plot	<code>enrichplot</code>	<code>barplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
KEGG	Enrichment Map	<code>enrichplot</code>	<code>emapplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
KEGG	Gene-Concept-Network	<code>enrichplot</code>	<code>cnetplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
KEGG	Pathway	<code>pathview</code>	<code>pathview()</code>	Cal modificar la funció per guardar els gràfics en el directori temporal
Reactome	ORA	<code>ReactomePA</code>	<code>enrichPathway()</code>	Totes les espècies de KEGG
Reactome	GSEA	<code>ReactomePA</code>	<code>gsePathway()</code>	Permutació de gens
Reactome	Bar-Plot	<code>enrichplot</code>	<code>barplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
Reactome	Enrichment Map	<code>enrichplot</code>	<code>emapplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
Reactome	Gene-Concept-Network	<code>enrichplot</code>	<code>cnetplot()</code>	Necessita l'objecte del class <code>enrichResult</code>
Reactome	Pathway	<code>ReactomePA</code>	<code>viewPathway()</code>	

Table 3.1: Resum de les anàlisis disponibles i recursos de Bioconductor R

3.2 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, sinó que és un procès que no és fàcil i pot tardar fins que el paquet estigui publicat amb èxit. Com comenta [[Wickham, 5 15](#)] “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 de 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 implica 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

3 Tractament bioinformàtic

```
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 estan ja:

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

He decidit no forçar la instal·lació de totes les bases d'anotacions per GO i Reactome. Al servidor sí que ho faria, però per a la instal·lació local podria resultar ser una experiència massa ferragosa, perquè encara que l'usuari necessités per a la seva anàlisi només un genoma anotat específic, hauria d'instal·lar tots els altres innecessàriament; i per tant dedicar més temps a la instal·lació que iniciaria amb la primera crida de la funció; `runPathwayApp` i a més a més també ocuparia espai al seu disc dur. Recordem que cada base d'anotació té un pes important. La base de dades `org.Mm.eg.db` per a ratolí, per exemple, ocuparà aprox. 275 megabytes al disc dur. Si anteriorment

3 Tractament bioinformàtic

les anotacions no són descarades per a l'espècie que l'usuari vol analitzar, l'usuari rebrà un error: **ERROR: object 'org.Mm.eg.db' is not found.**

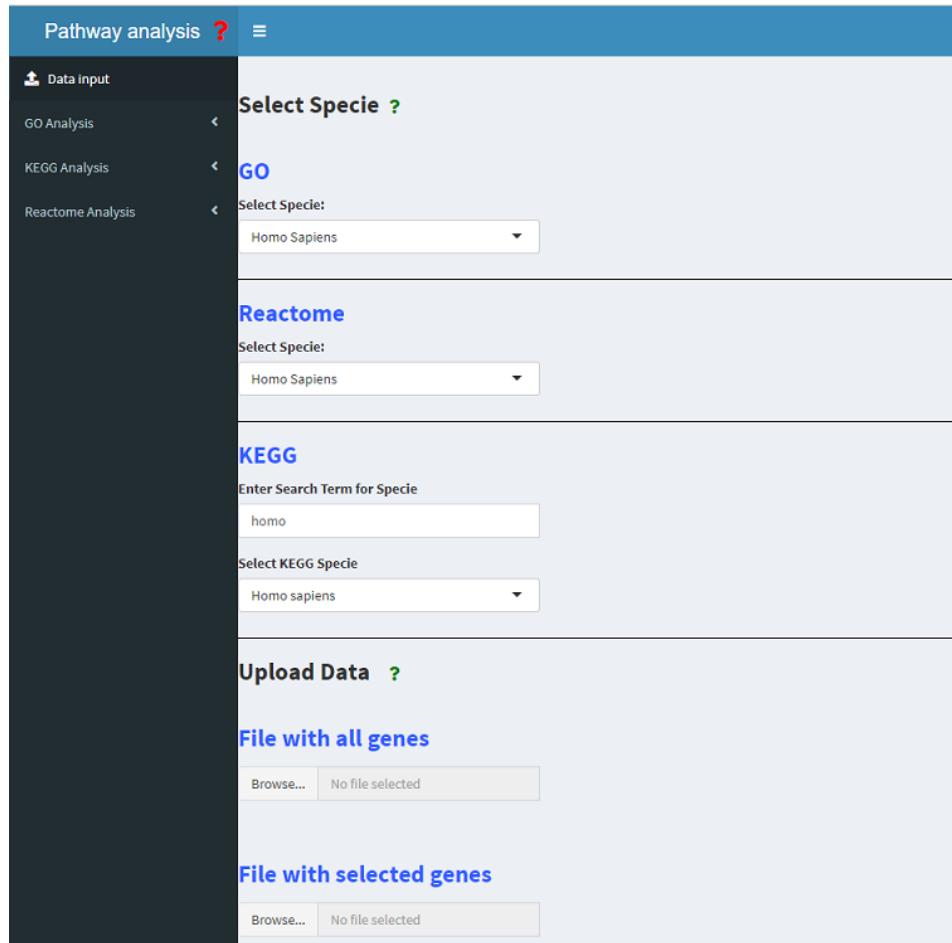
4 L'aplicació

Després de la instal·lació l'aplicació és completament funcional localment i ofereix l'anàlisi a partir de les bases de dades GO, KEGG i Reactome. A l'apartat **Input data** l'usuari primer ha d'indicar l'espècie per a totes tres bases de dades. Per les bases de dades de Reactome l'usuari pot elegir entre Homo Sapiens, Rat, Mouse, Celegans, Yeast, Zebrafish, Fly. Per a l'anàlisi GO, a més de les anteriors, hi ha disponibles aquestes espècies addicionals: Arabidopsis, Bovine, Chicken, Canine, Pig, Rhesus, E coli strain K12, Xenopus, Anopheles, Chimp, Malaria, E coli strain Sakai. Hi ha més espècies disponibles per a l'anàlisis KEGG, perquè la funció de `culsterProfiler enrichKEGG()` descarrega les últimes anotacions directament de la base de dades KEGG. Es poden trobar totes les espècies [aquí](#). També l'usuari pot buscar l'espècie introduint els termes de cerca. Finalment l'usuari puja l'arxiu amb els gens i els logRatio provinents de l'estudi de *microarrays*.

A la presentació següent faré ús de les dades disponibles al paquet DOSE de Bioconductor que provenen de l'estudi de schmidt2008humoral on es comparen les mostres del càncer de mama del grau III vs. el grau I.

4 L'aplicació

Imatge 4.1: Pàgina d'entrada



L'usuari té la possibilitat d'introduir l'arxiu amb tots els gens i els gens seleccionats. Un cop introduïdes les dades es mostra un petit resum del contingut dels arxius.

4 L'aplicació

imatge 4.2: Resum de les dades pujades

The screenshot shows a user interface for uploading gene lists. At the top, there are two sections: "File with all genes" and "File with selected genes". Both sections have a "Browse..." button and a file input field containing "Dose_geneList.csv" and "Dose_selectedGenes.csv" respectively. Below each section is a blue "Upload complete" button.

Under "File with all genes", it says "You uploaded: 12495 genes". Below that, "First 10 entries" are listed in a table:

Entrez ID	FoldChange
4312	4.573
8318	4.515
10874	4.418
55143	4.144
55388	3.876
991	3.678
6280	3.502
2305	3.292
9493	3.286
1062	3.220

Under "File with selected genes", it says "You selected: 207 genes". Below that, "First 10 entries" are listed in a table:

Entrez ID	FoldChange
4312	4.573

L'aplicació està dividida doncs en 4 parts substancials:

1. Entrada de les dades;
2. Anàlisi GO;
3. Anàlisi KEGG;
4. Anàlisi Reactome.

4 L'aplicació

L'aplicació ofereix dos mètodes d'anàlisi: d'una banda es pot fer ORA i d'altra banda l'anàlisi GSEA. Recordem que l'ORA consisteix a seleccionar els gens diferencialment expressats i basant-se en GO, KEGG o Reactome comprovar si una de les agrupacions de gens sugerides per aquestes bases de dades està sobre o sotraexpressada en els gens seleccionats. Per dur a terme l'ORA l'usuari té l'opció de definir un *cut-off* de Log-Ratio per formar el conjunt dels gens que s'hi utilitzarà (*gene set*). ORA és una bona eina per veure els efectes grans però els efectes petits se li escapan. Els efectes petits derivats dels gens individuals poden acumular-se en un efecte conjunt substancial el qual ORA no serà capaç de detectar. És aquí on GSEA mostra la seva utilitat.

Els apartats d'anàlisi (GO, KEGG i Reactome) ofereixen tan representacions comunes com representacions específiques.

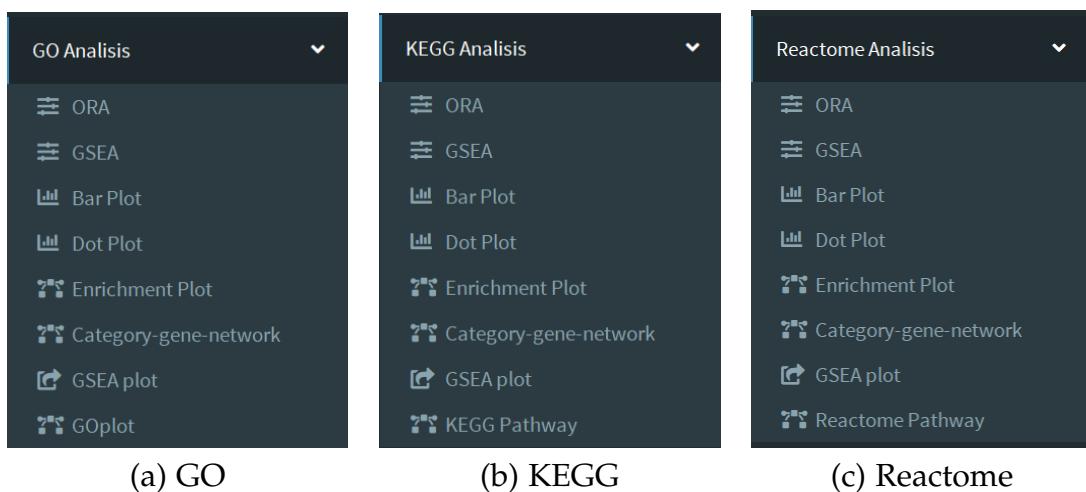
Els anàlisis i representacions en comú són:

- Taula dels resultats ORA;
- Taula dels resultats GSEA;
- Gràfic de barres del resultat ORA;
- Gràfic de punts del resultat ORA;
- El mapa d'enriquement (Enrichment Map);
- La xarxa dels gens en categories (Category-gene-network);
- El gràfic de GSEA.

Les anàlisis específics són:

- GO → Gràfic GO
- KEGG → Rutes de la base de dades KEGG
- Reactome → Rutes de la base de dades Reactome

4 L'aplicació



imatge 4.3: Els elements de les seccions d'anàlisi

4.1 ORA

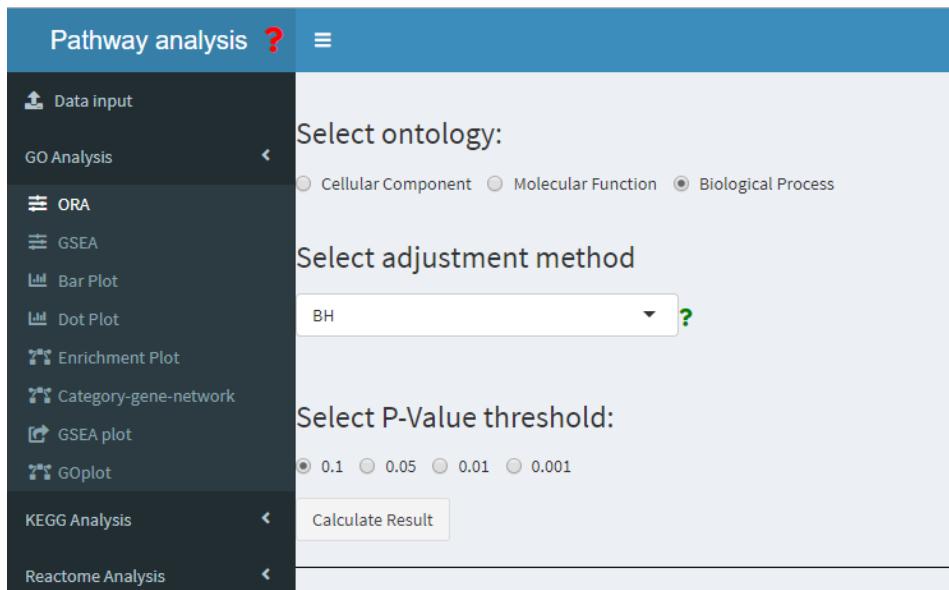
4.1.1 GO

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.

4 L'aplicació



imatge 4.4: Especificació d'ORA dels termes GO

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.

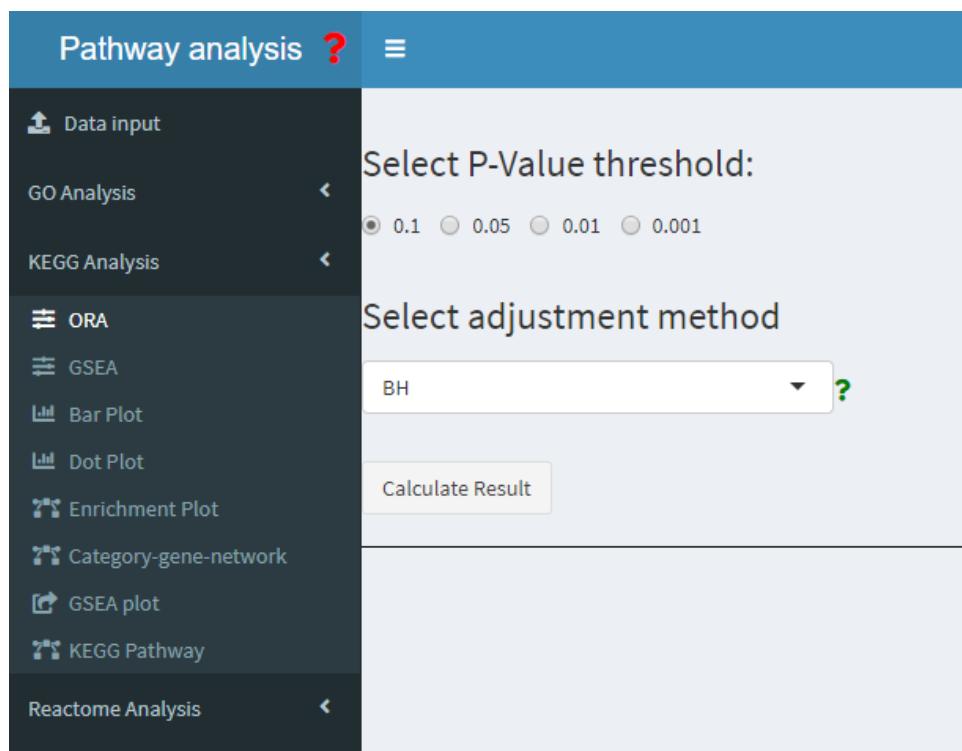
4.1.2 KEGG

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

4 L'aplicació

ID	Description	GeneRatio	BgRatio	pvalue	p-adjust	qvalue	Count	geneID
GO:0140014	mitotic nuclear division	29/189	201/11248	0.000	9.51e-16	0.000	29	CDCA8/CDC20/KIF23/CENPE/MYBL2/NDC80
GO:0000280	nuclear division	31/189	274/11248	0.000	3.38e-14	0.000	31	CDCA8/CDC20/KIF23/CENPE/MYBL2/NDC80
GO:0048285	organelle fission	32/189	303/11248	0.000	5.00e-14	0.000	32	CDCA8/CDC20/KIF23/CENPE/MYBL2/NDC80
GO:0000070	mitotic sister chromatid	21/189	109/11248	0.000	6.28e-13	0.000	21	CDCA8/CDC20/KIF23/CENPE/NDC80/NCAP

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



Imatge 4.6: 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.

4 L'aplicació

The screenshot shows a table titled 'ORA KEGG Results' with the following columns: ID, Description, GeneRatio, BgRatio, pvalue, p.adjust, qvalue, Count, and geneID. The table lists several biological pathways with their respective enrichment statistics. The 'p.adjust' and 'qvalue' columns are highlighted in green, indicating statistical significance.

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	geneID
hsa04110	Cell cycle	11/93	124/7860	0.000	3.82e-05	0.000	11	8318/991/9133/890/983/4085/7272/1111/891/4174/92
hsa04114	Oocyte meiosis	10/93	125/7860	0.000	1.85e-04	0.000	10	991/9133/983/4085/51806/6790/891/9232/3708/5241
hsa04218	Cellular senescence	10/93	160/7860	0.000	9.47e-04	0.001	10	2305/4605/9133/890/983/51806/1111/891/776/3708
hsa04061	Viral protein interaction with cytokine and cytokine receptor	8/93	100/7860	0.000	9.47e-04	0.001	8	3627/10563/6373/4283/6362/6355/9547/1524
hsa03320	PPAR signalling pathway	7/93	74/7860	0.000	6.47e-04	0.001	7	4312/9415/9370/5105/2167/3158/5346
hsa04914	Progesterone-mediated oocyte maturation	7/93	99/7860	0.000	5.14e-03	0.005	7	9133/890/983/4085/6790/891/5241
hsa04115	cAMP signalling	5/93	72/7860	0.003	2.29e-02	0.043	2	8122/6241/892/1111/901

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

4.1.3 Reactome

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

4 L'aplicació

The screenshot shows the 'Pathway analysis' section of the Reactome application. On the left, a sidebar lists various analysis options: Data input, GO Analysis, KEGG Analysis, Reactome Analysis, ORA, GSEA, Bar Plot, Dot Plot, Enrichment Plot, Category-gene-network, GSEA plot, and Reactome Pathway. The 'ORA' option is selected. The main panel is titled 'Select adjustment method' with 'BH' chosen, and 'Select P-Value threshold:' with 0.1 selected. A 'Calculate Result' button is present. Below this is a table titled 'Download Results as .csv'. The table has columns: ID, Description, GeneRatio, BgRatio, pvalue, p.adjust, qvalue, Count, and geneID. Four rows of data are shown:

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	geneID
R-HSA-2500257	Resolution of Sister-Chromatid Cohesion	15/142	126/10619	0.000	3.25e-09	0.000	15	CDCA8/CDC20/CENPE/CCNB2/NDC80/SKA1/C
R-HSA-68877	Mitotic Prometaphase	18/142	200/10619	0.000	3.25e-09	0.000	18	CDCA8/CDC20/CENPE/CCNB2/NDC80/NCAPH,
R-HSA-69620	Cell Cycle Checkpoints	21/142	293/10619	0.000	4.13e-09	0.000	21	CDC45/CDCA8/MCM10/CDC20/CENPE/CCNB2,
R-HSA-20510	Mitotic Spindle	13/142	112/10619	0.000	3.00e-09	0.000	13	CDCA8/CDC20/CENPE/NDC80/UBE2C/SKA1/C

imatge 4.8: El resultat d'anàlisi ORA. Reactome.

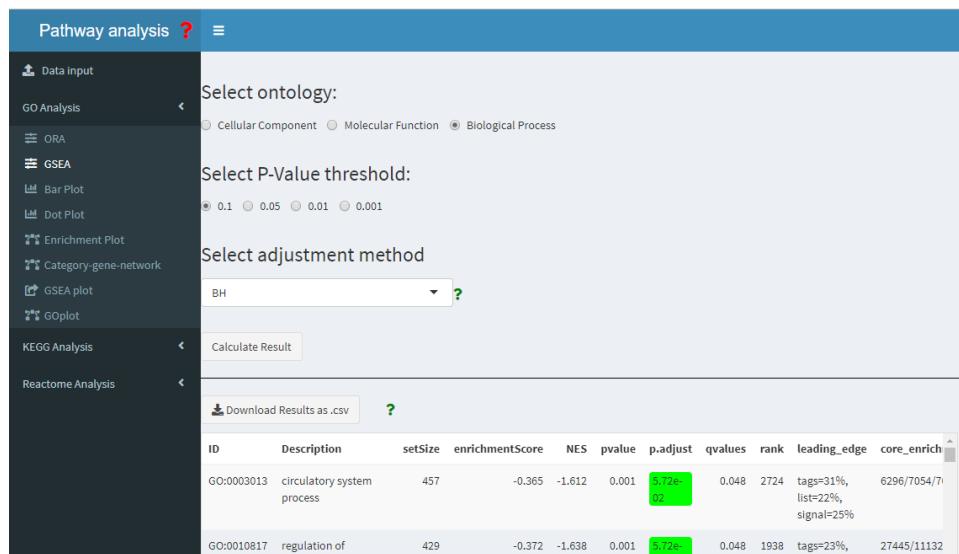
4.2 GSEA

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 L'aplicació

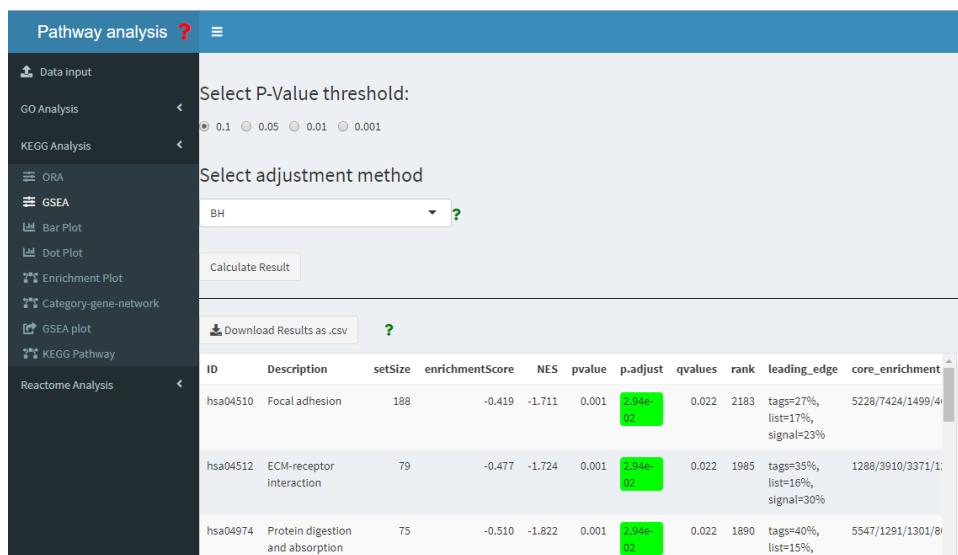


Imatge 4.9: El resultat de l'anàlisi GSEA. GO.

4.2.2 KEGG

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

4 L'aplicació



Imatge 4.10: El resultat de l'anàlisi GSEA. KEGG.

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()`

4 L'aplicació

The screenshot shows the 'Pathway analysis' interface. On the left, a sidebar lists various analysis types: Data input, GO Analysis, KEGG Analysis, Reactome Analysis, ORA, GSEA, Bar Plot, Dot Plot, Enrichment Plot, Category-gene-network, GSEA plot, and Reactome Pathway. The 'Reactome Pathway' option is selected. The main panel has two sections: 'Select P-Value threshold:' with radio buttons for 0.1, 0.05, 0.01, and 0.001 (0.01 is selected), and 'Select adjustment method:' with a dropdown menu showing 'BH' (selected) and 'holm'. Below these are 'Calculate Result' and 'Download Results as .csv' buttons. The results table displays three rows of data:

ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrich
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	418	-0.338	-1.490	0.001	2.57e-02	0.020	2788	tags=27%, list=22%, signal=21%	26052/534/2
R-HSA-1474244	Extracellular matrix organization	266	-0.458	-1.937	0.001	2.37e-02	0.020	1943	tags=33%, list=16%, signal=29%	8038/11132/-
R-HSA-.....	Cardiac conduction	117	-0.402	-1.545	0.001	2.57e-02	0.020	2829	tags=36%, list=18%, signal=29%	6543/3750/1

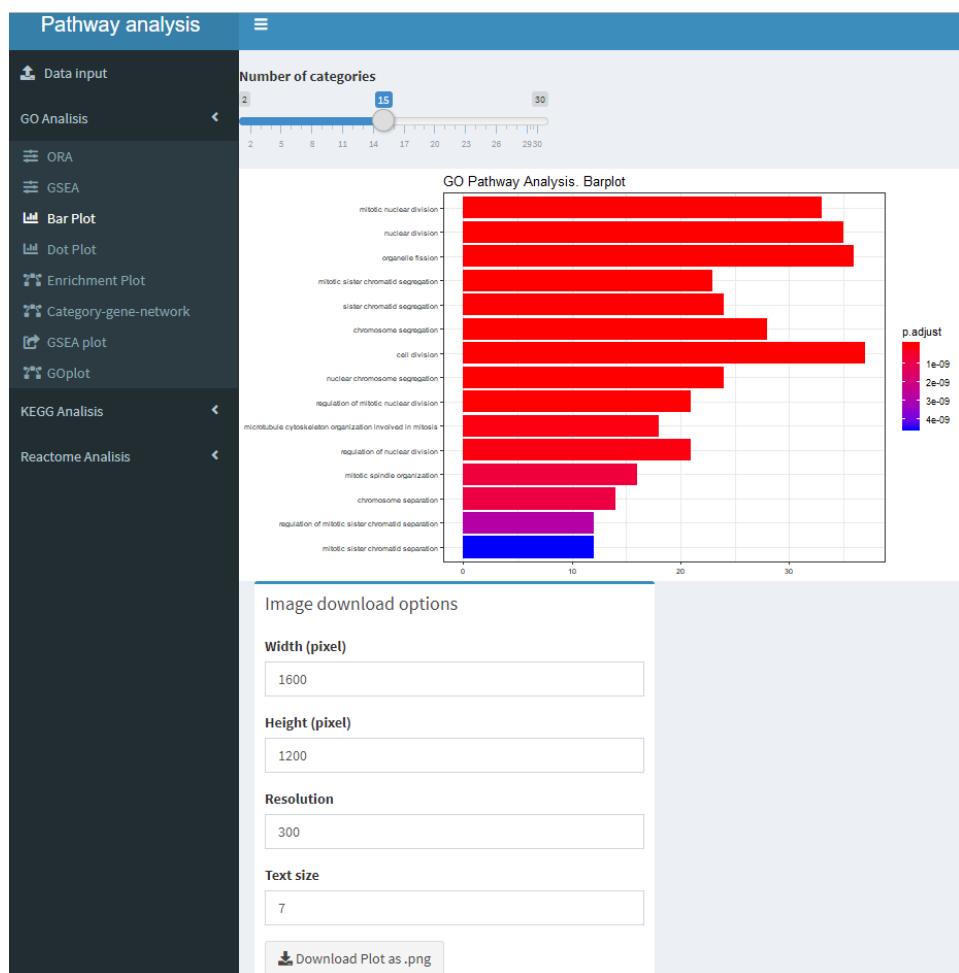
Imatge 4.11: El resultat d'anàlisi GSEA. Reactome.

4.3 Visualització i l'anàlisi topològic

4.3.1 Bar-Plots

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

4 L'aplicació

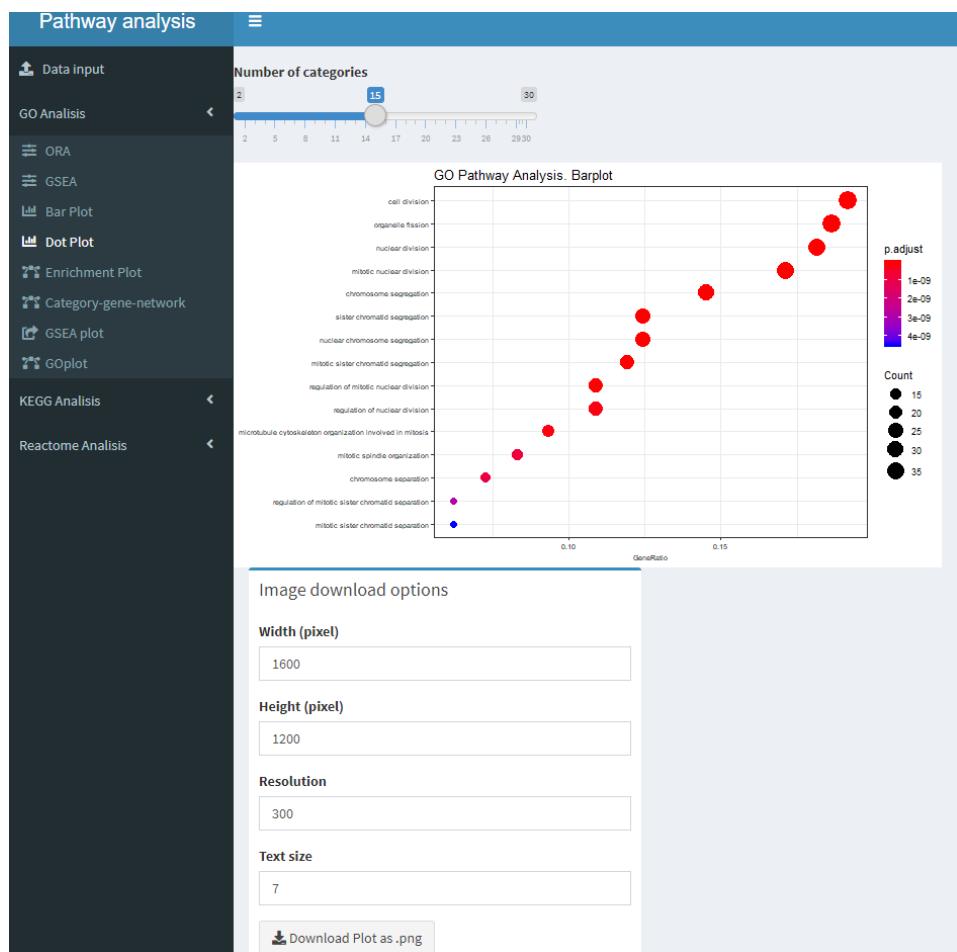


Imatge 4.12: Bar-Plot. GO.

4.3.2 Dot-Plots

El *Dot-Plot* visualitza addicionalment el *gen ratio*. També aquí l'usuari pot seleccionar el nombre de categories.

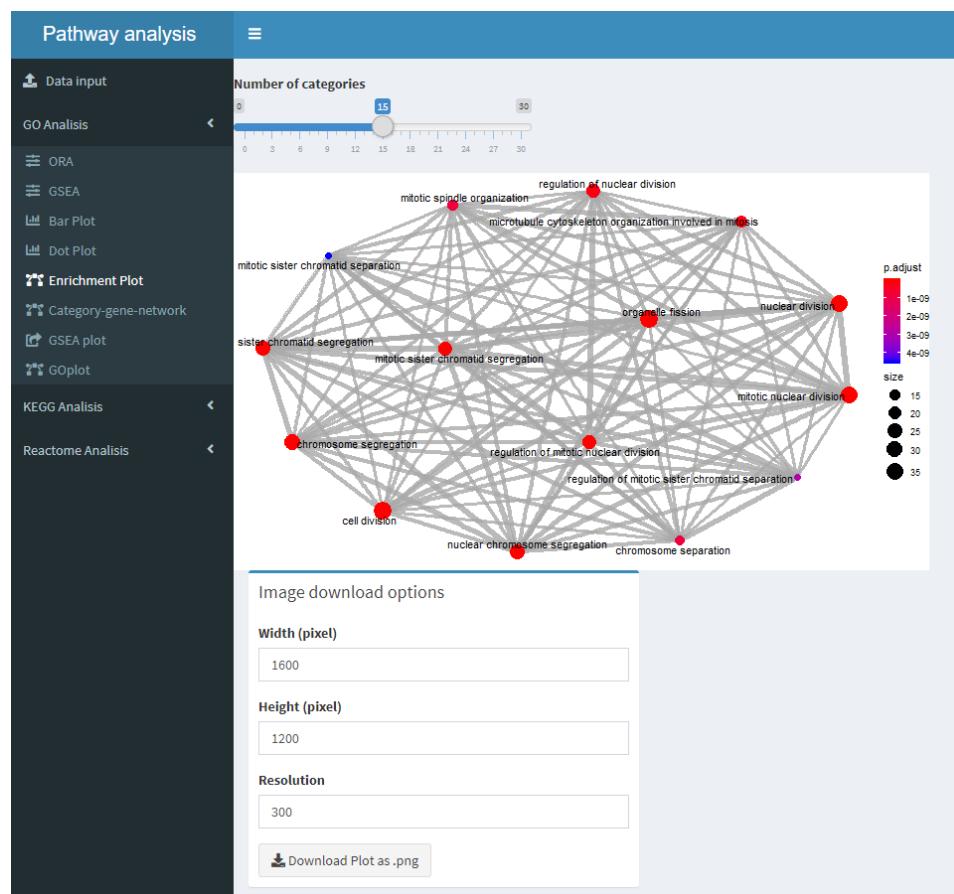
4 L'aplicació



imatge 4.13: Dot-Plot. GO.

4 L'aplicació

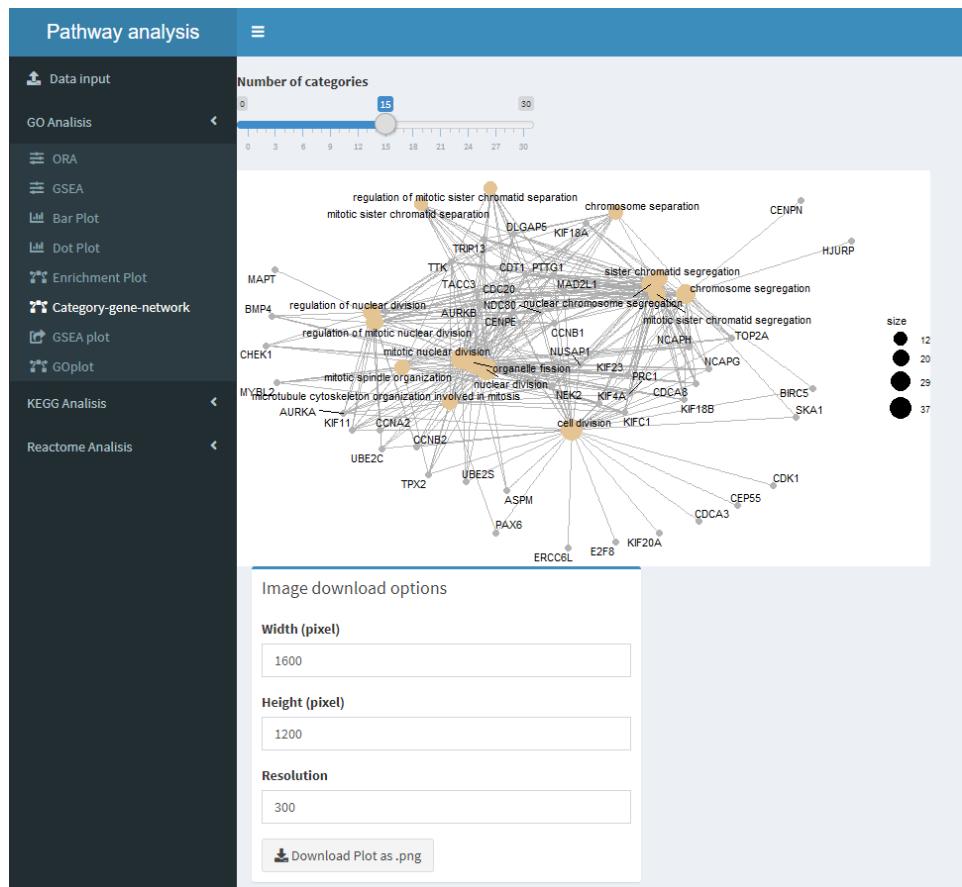
4.3.3 Enrichment Maps



Imatge 4.14: Enrichment Map. GO.

4 L'aplicació

4.3.4 Category-Gene-Network Plot

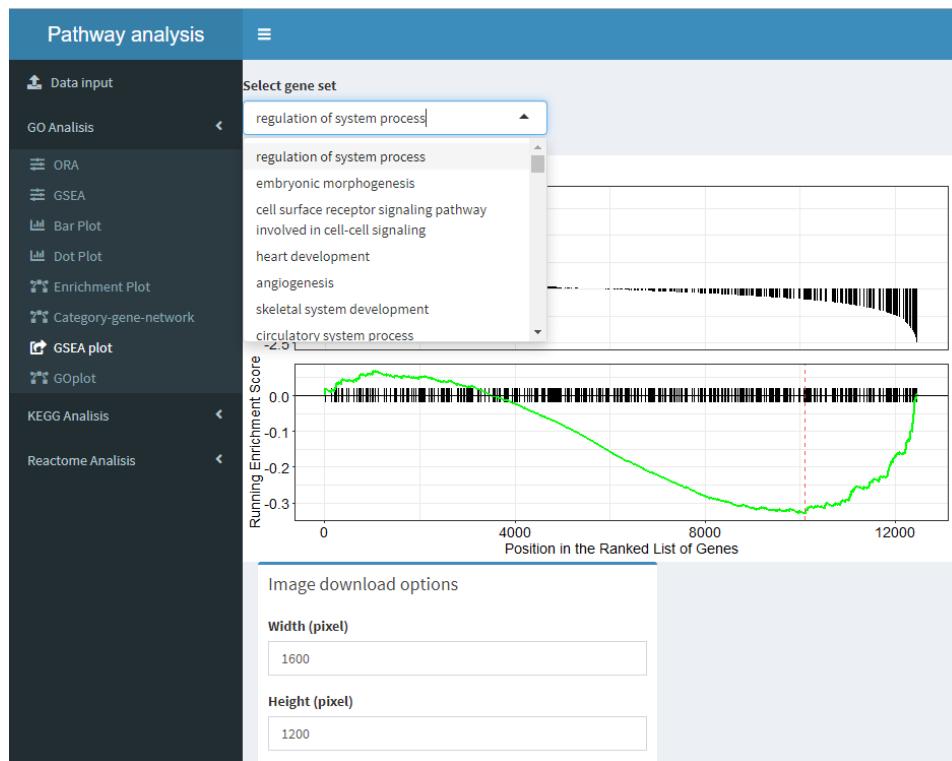


imatge 4.15: Category-Gene-Network Plot. GO.

4.3.5 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*

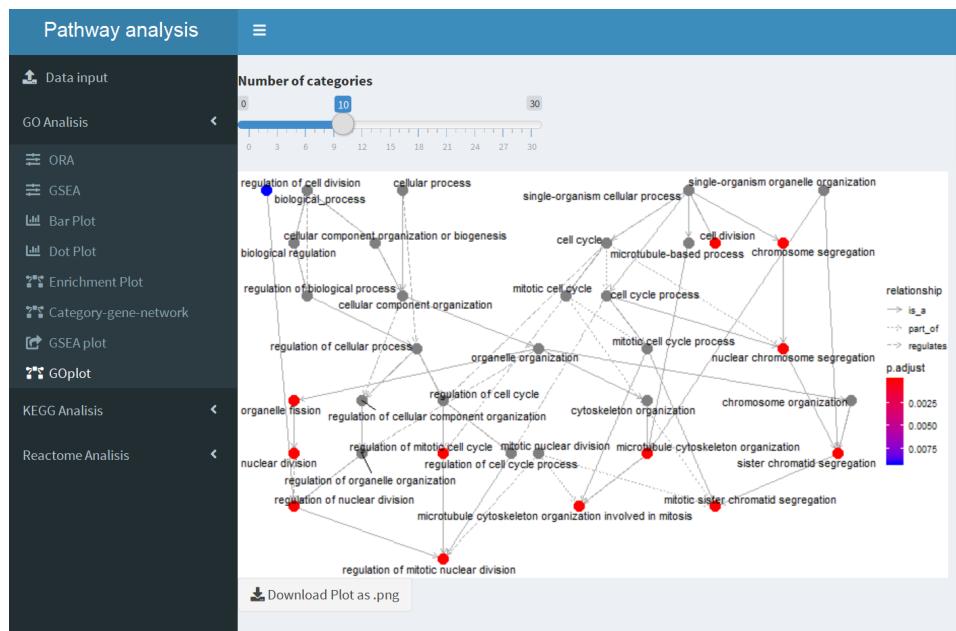
4 L'aplicació



imatge 4.16: GSEA Plot. GO.

4 L'aplicació

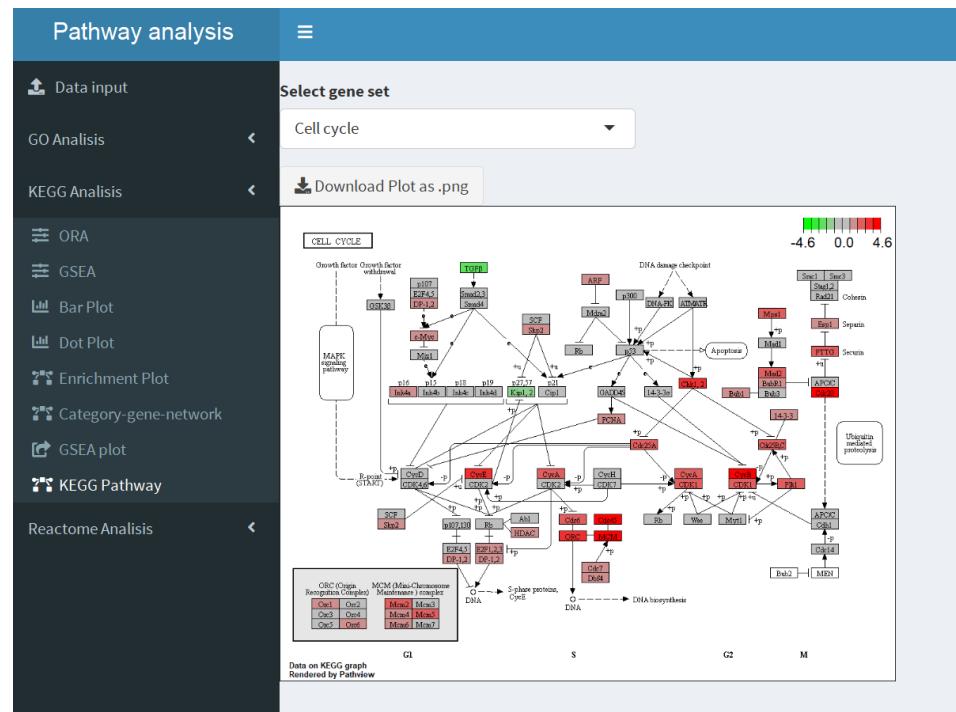
4.3.6 GO Plot



imatge 4.17: GO Plot

4 L'aplicació

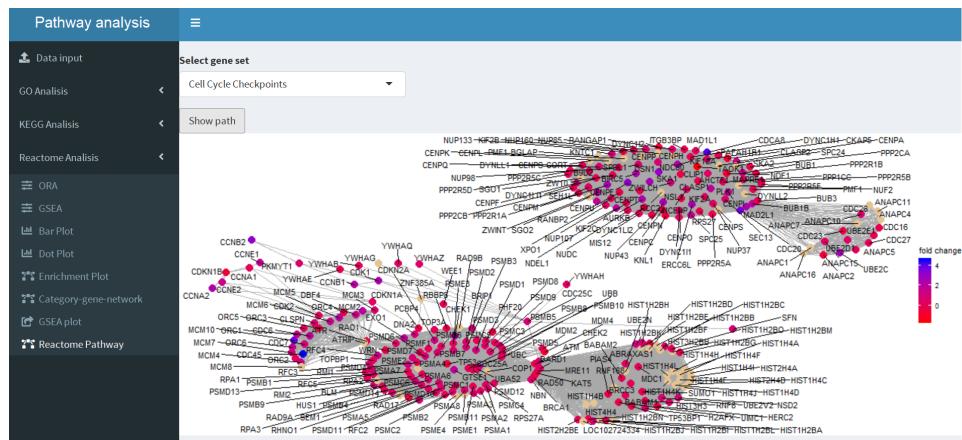
4.3.7 KEGG Pathway



imatge 4.18: KEGG pathway

4 L'aplicació

4.3.8 Reactome Pathway

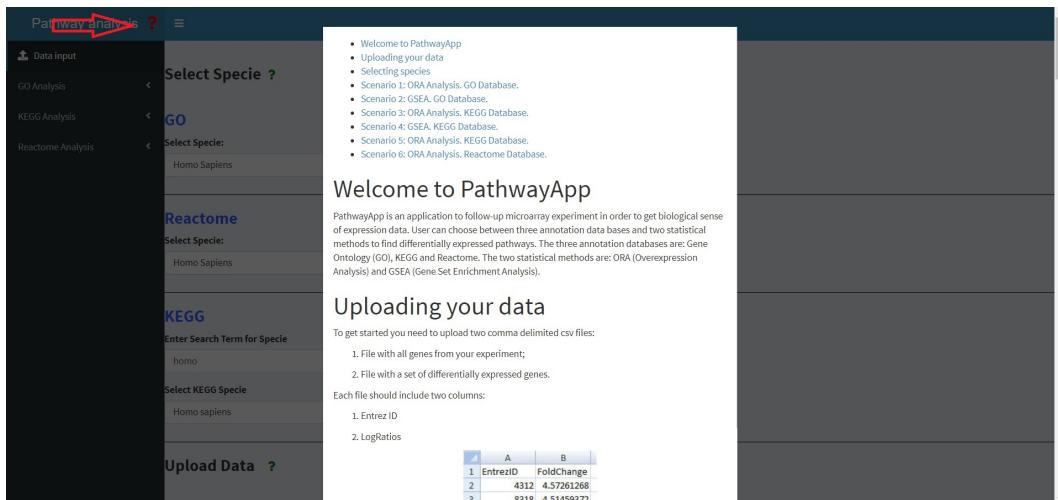


Imatge 4.19: Reactome pathway

4.4 Manual i ajudes del programa

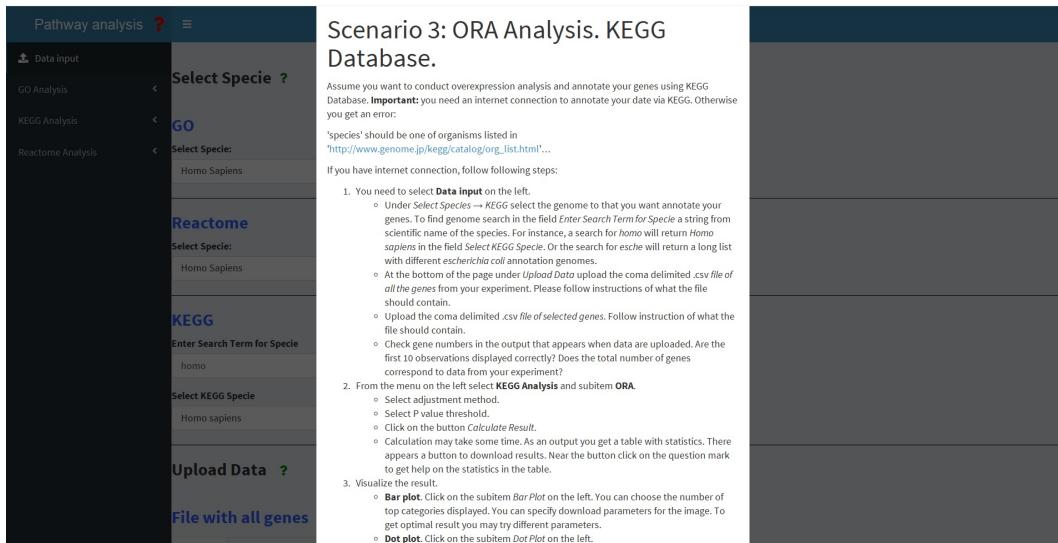
Per facilitar l'ús de l'aplicació he pensat com es podria fer de manera el més intuïtiva possible. Primer cal destacar que com a llengua de manual he elegit l'anglès per poder fer l'ús de l'aplicació el més inclusiu possible. Segon, l'usuari pot accedir tant al manual com a l'ajuda, que es guarden en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

4 L'aplicació



Imatge 4.20: Manual per a aplicació

Com es veu hi ha apartats diferents. Dependent dels objectius de l'usuari, aquest pot seleccionar l'apartat que més li interessa. Així, si l'usuari vol fer l'anàlisi ORA amb l'anotació KEGG pot navegar en la secció **Scenario 3: ORA Analysis.KEGG Database**.

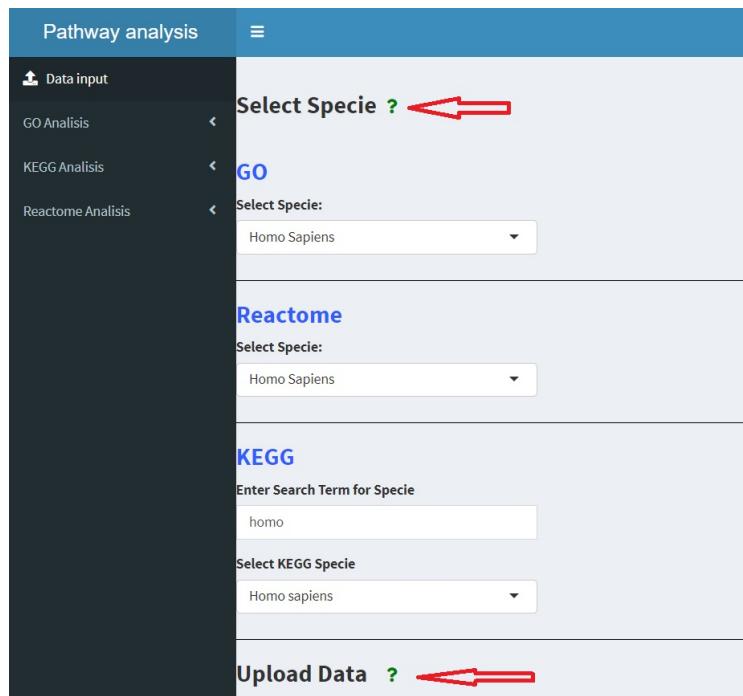


Imatge 4.21: Manual per a l'anàlisi ORA amb l'anotació KEGG

4 L'aplicació

També, l'usuari pot accedir a l'ajuda clicant els 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()`.

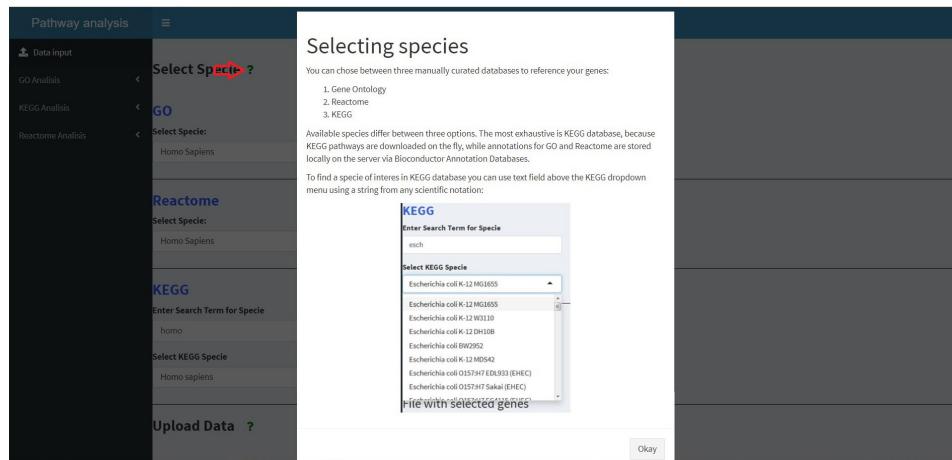


imatge 4.22: Senyals d'ajuda

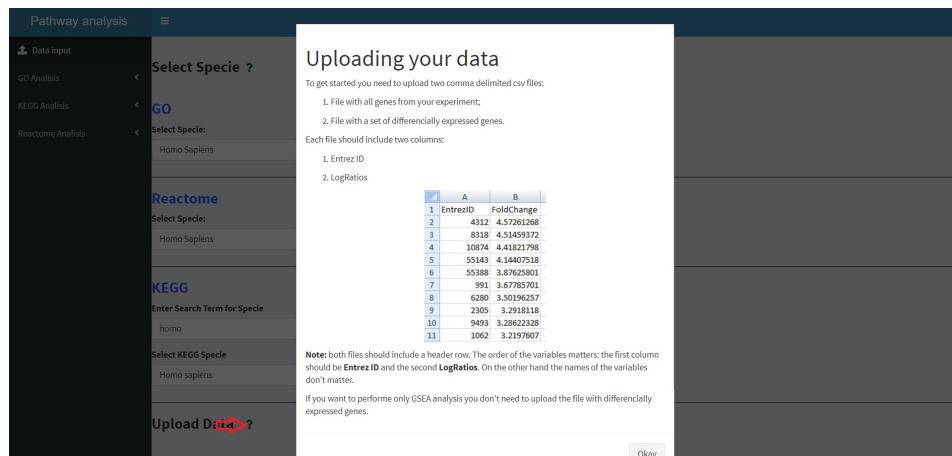
El clic en aquests senyals fa que aparegui una finestreta amb la informació d'ajuda.

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

4 L'aplicació



Imatge 4.23: Ajuda per a l'elecció de l'espècie



Imatge 4.24: Ajuda per pujar les dades

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

4 L'aplicació

The screenshot shows the RStudio Browser interface with the 'Pathway analysis' tab selected. On the left, there's a sidebar with various analysis options like GO Analysis, GSEA, and KEGG Analysis. The main panel displays the results of an ORA analysis. A modal window titled 'Understanding ORA output' is open, explaining the statistical test based on the hypogeometrical distribution. It includes a formula for calculating the p-value:

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

The modal also provides definitions for terms like 'Description', 'GeneRatio', 'BgRatio', 'pvalue', and 'p.adjust'. Below the modal, the analysis results table shows three pathways: spindle, microtubule, and condensed chromosome, centromeric.

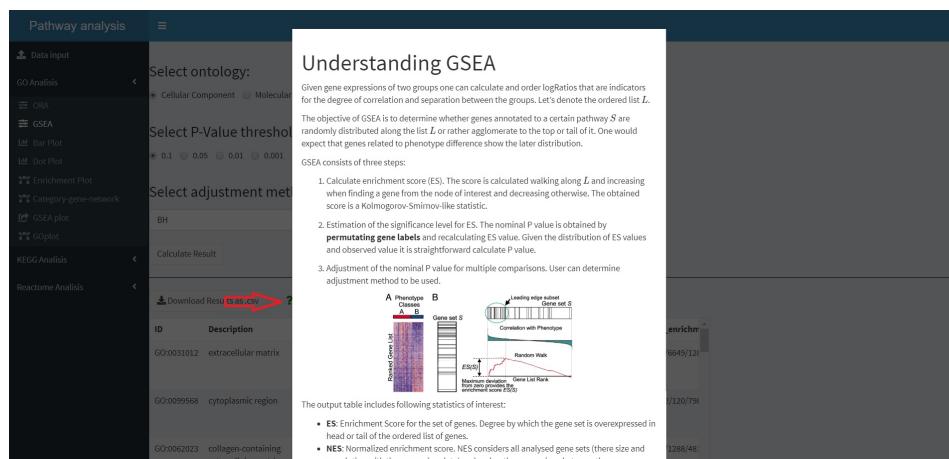
Imatge 4.25: Infromació per la interpretació d'anàlisi ORA

Aquí cal destacar que les fórmules, 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 fórmules. Pot ser un problema amb Java, que s'ha d'actualitzar? Ho estic investigant.

The screenshot shows the RStudio Browser with the 'Pathway analysis' tab selected. A modal window titled 'From R manual p.adjust{stats}' is open, providing detailed information about the adjustment methods used in the p.adjust function. It explains the Bonferroni correction, Holm's method, Hochberg's method, Hommel's method, Benjamini and Hochberg's method, Benjamini, Yekutieli, and others' method, and the BH method. The modal also includes a 'References' section with citations for Holm, Hochberg, and Benjamini et al.

Imatge 4.26: Ajuda per la selecció del mètode d'ajustament

4 L'aplicació



Imatge 4.27: 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. 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. 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 a [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] .

5 Validació dels resultats

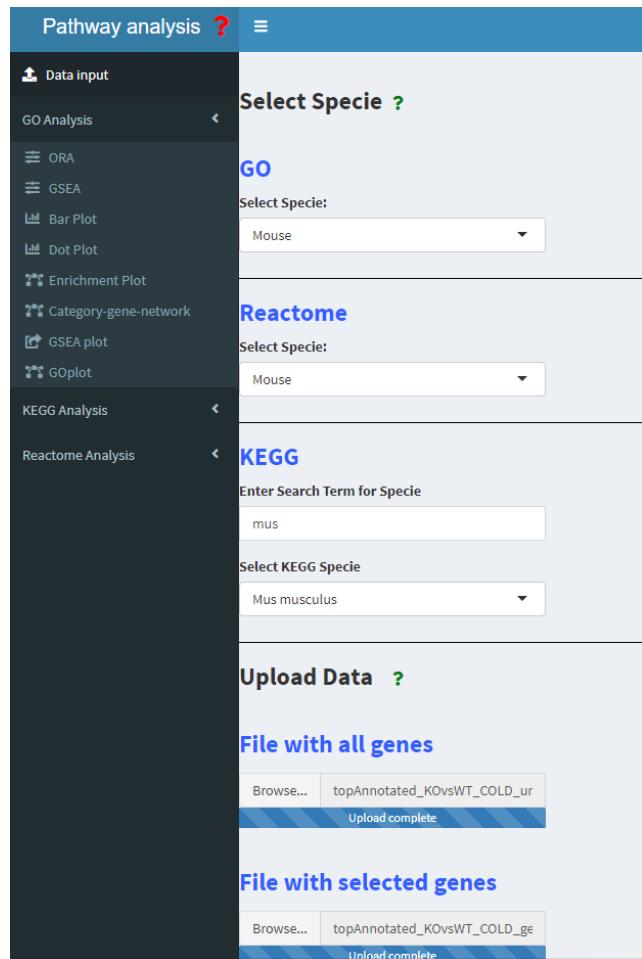
5.1 Exemple d'anàlisi 1. GEO: GSE100924

[Li et al., 2017] analitzen l'associació del gen *Zbtb7b* amb la producció dels greixos marrons que al seu torn influeixen en termogènesis i processos metabòlics diferents. D'aquesta manera els greixos marrons són importants per al tractament dels desordres metabòlics.

Les dades d'estudi són ja preprocessades per Ricardo Gonzalo Sanz i Sanchez Pla i estan disponibles a [github](#). De la carpeta *results* he agafat la taula *topAnnotated_KOvsWT_COLD.csv*. Sanz i Pla utilitzen el paquet ReactomePA per a l'anàlisi d'enriquiment. Repeteixo doncs el seu anàlisi utilitzant l'aplicació.

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

5 Validació dels resultats



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.

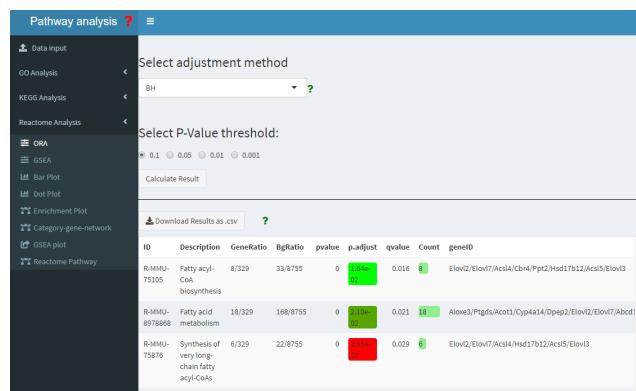
5 Validació dels resultats

File with selected genes	
Browse...	topAnnotated_KOvsWT_COLD_genes.xls
Upload complete	
<hr/>	
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
<hr/>	
You selected: 769 genes	
First 10 entries	
Entrez ID	FoldChange
320492	-1.454
50785	0.743
12859	-0.822
98404	-0.805

Imatge 5.2: Breu resum de les dades

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

5 Validació dels resultats



Imatge 5.3: Resultat d'anàlisi ORA de Reactome

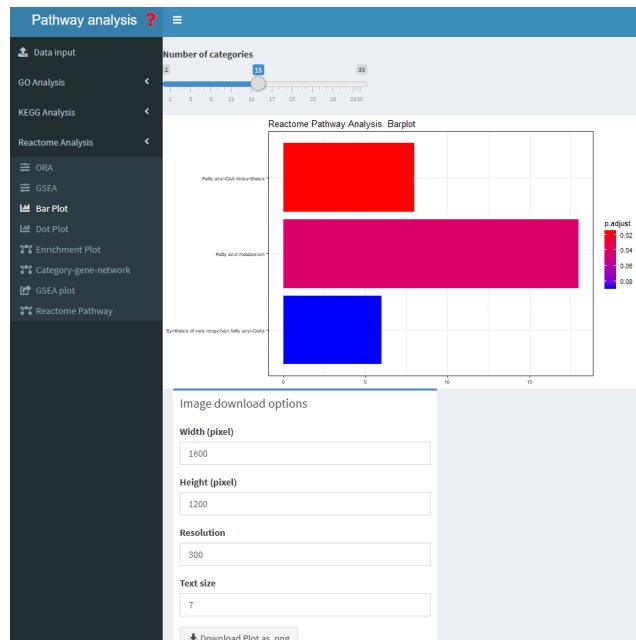
Observem que les rutes mostrades són les mateixes esmentades per Sanz i Pla. També destaquem que el resultat coincideix amb les trobades a l'estudi de [Li et al., 2017]. S'observa la pertorbació de les rutes relacionades amb els greixos marrons **Fatty acyl-CoA biosynthesis i Fatty-acid metabolism**.

3. Visualització del resultat ORA

L'aplicació permet visualitzar els resultats obtinguts amb la ORA.

- Seleccions *Reactome Analysis→Bar Plot*

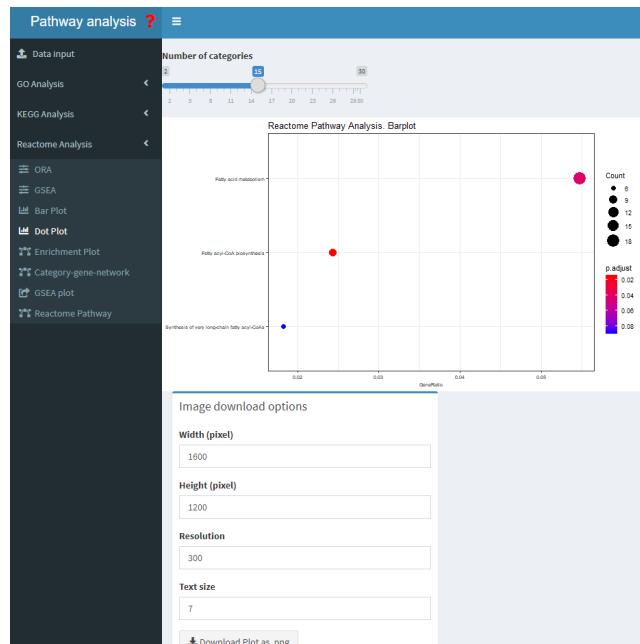
5 Validació dels resultats



Imatge 5.4: Gràfic de barres

- Seleccionso *Reactome Analysis*→*Dot-Plot*

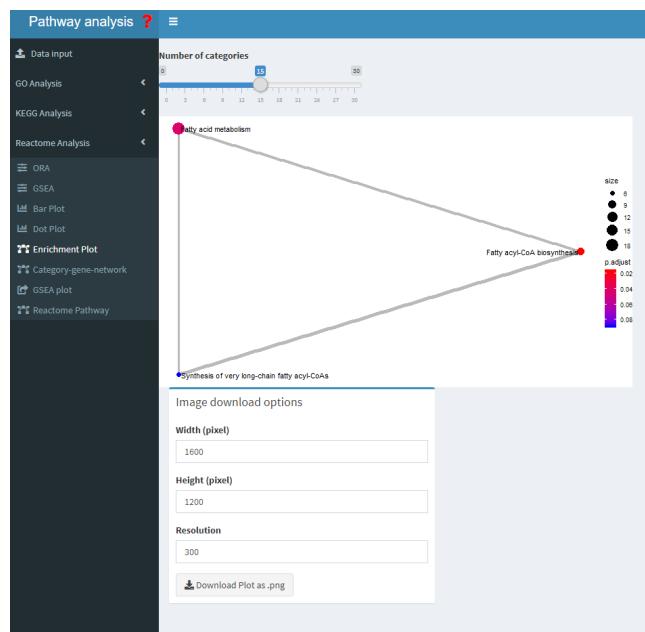
5 Validació dels resultats



Imatge 5.5: Gràfic de punts

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

5 Validació dels resultats

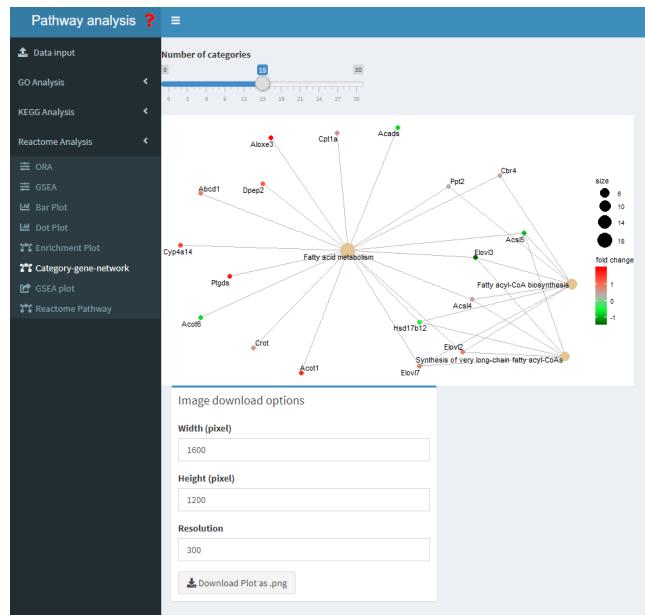


Imatge 5.6: Mapa d'enriquiment

Observem que totes les rutes identificades comparteixen els gens.

- Selecciono *Reactome Analysis* → *Gene-Concept-Network* Aquesta visualització incorpora els gens individuals de la ruta i la magnitud de la seva expressió diferencial. Observem que el gen Elov13 està sotaexpressat, tal com es comenta al paper de [Li et al., 2017]. El gen Elov13 és el component important per a reclutament dels lípids al teixit adipós marró [Westerberg et al., 2006]

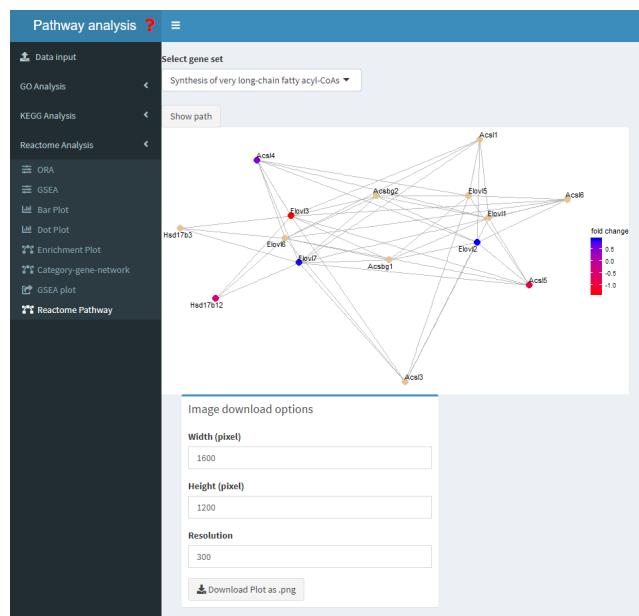
5 Validació dels resultats



Imatge 5.7: Red de les categories i gens

- Selecciono *Reactome Analysis*→*Reactome Pathway*

5 Validació dels resultats

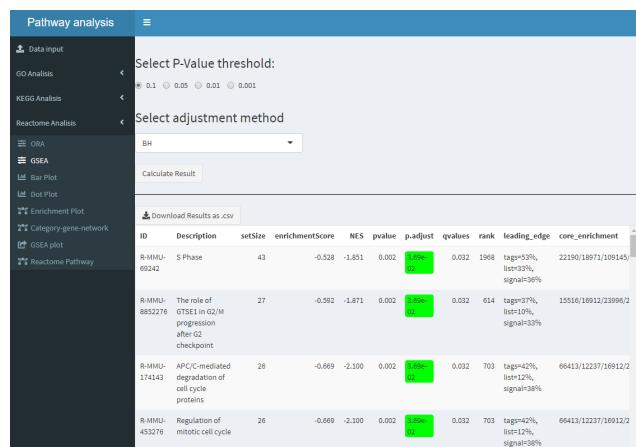


Imatge 5.8: Rutes Reactome

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

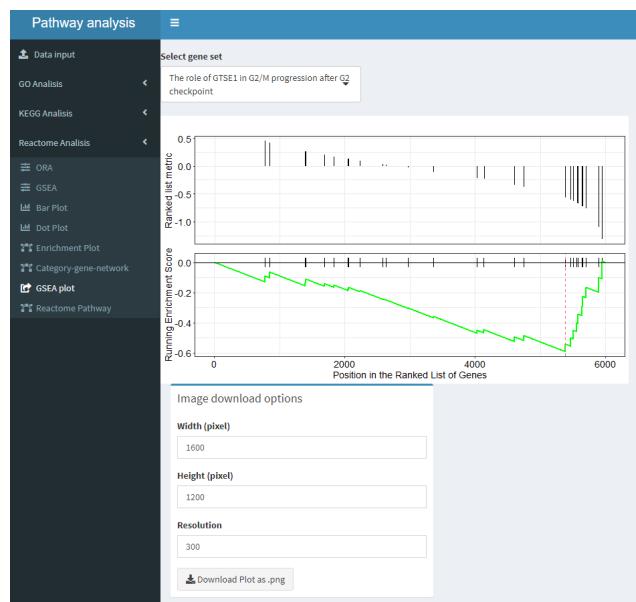
1. Clico en l'apartat *Reactome Analysis* → *GSEA*. Selecciono com a mètode d'ajustament *BH* i el cut-off del valor de P ajustat 0.05. Clico a *Calculate results*
Amb el valor de P de 0.05 l'anàlisi no troba cap ruta enriquida.
2. Augmento el Cut-Off del valor de P a 0.1
Amb el Cut-Off més alt l'aplicació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

5 Validació dels resultats



Imatge 5.9: Anàlisi GSEA

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

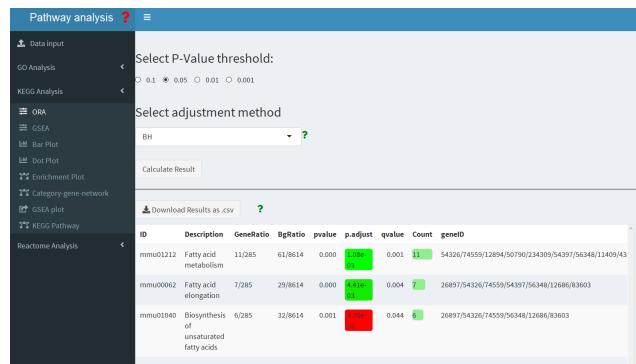


Imatge 5.10: Gràfic GSEA

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:

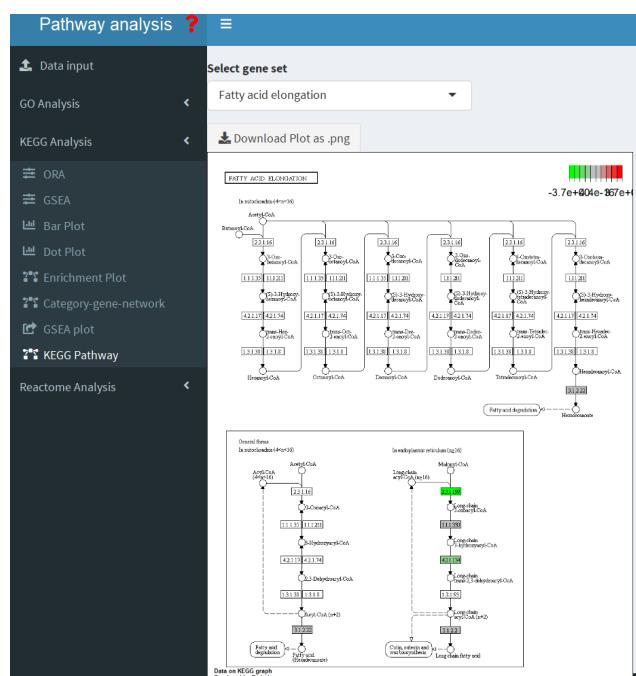
5 Validació dels resultats

- 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

- Anem a *KEGG* → *KEGG Pathway*

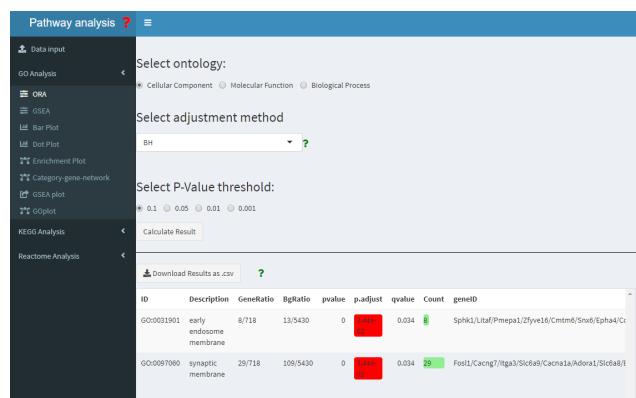


imatge 5.12: Gràfic de les rutes KEGG

5 Validació dels resultats

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 Conclusions

Al treball final he tingut com a objectiu la creació d'una aplicació per a l'anàlisi de les rutes. Primer he determinat què és l'anàlisi de les rutes i on se situa en l'anàlisi global d'expressió genètica. He explicat que l'anàlisi de les rutes és l'últim pas de l'anàlisi d'expressió gènica on es pretén donar sentit biològic a les dades d'expressió. Per dur a terme aquesta anàlisi s'utilitza una llista dels gens diferencialment expressats i els logRatios de tots els gens de l'experiment. Així, he explicat que es necessari anotar aquests gens a les bases existents per poder agrupar-los en conceptes (o rutes).

En segon lloc, he triat les bases de dades més conegudes i que són utilitzades pels paquets de Bioconductor. Aquestes bases de dades són: GO, KEGG i Reactome. En tercer lloc, he descrit quins mètodes existeixen per calcular la significació biològica de les rutes. Entre altres he identificat tres estratègies: ORA, FCS i l'anàlisi topològic de les rutes. En quart lloc, he descrit amb més detall els mètodes ORA i GSEA i també he identificat les possibilitats per visualitzar les rutes. En cinquè lloc, després de formular un protocol d'anàlisi, he buscat i identificat els paquets de Bioconductor que permeten aplicar els mètodes descrits. Entre altres he triat clusterProfiler, ReactomePA i pathview. He posat aquests paquets com a base per la futura aplicació.

El resultat ha estat l'aplicació dividida en quatre apartats: 1) Entrada de les dades; 2) Anàlisi amb la base d'anotació GO; 3) Anàlisi amb la base d'anotació KEGG i anàlisi amb la base d'anotació Reactome. Per a cada base de dades l'usuari pot fer ORA, GSEA i obtenir la visualització de la topologia de les rutes. L'aplicació permet descarregar els resultats com a arxiu .csv i imatges .png, els quals l'usuari pot personalitzar. Pel que fa a la validació dels resultats, s'ha mostrat una bona coincidència amb l'estudi triat per validar les dades. Aquest estudi investiga l'impacte del gen *Zbtb7b* sobre la producció dels greixos marrons. Justament les rutes relacionades

6 Conclusions

amb la síntesi de greixos marrons estaven identificades amb l'aplicació. En darrer lloc, l'aplicació pot ser descarregada i instal·lada lliurement del repositori GitHub i la via per fer-ho s'explica a la memòria. Actualment es busca el mètode per publicar l'aplicació a internet per fer-la encara més accessible.

Acrònims

ES Enrichment Score. 16

FCS Functional Class Scoring. 6, 11, 67

GO Gene Ontology. v, vi, vii, 6, 12, 13, 15, 18, 19, 23, 24, 25, 29, 31, 32, 33, 34, 35, 37, 38, 41, 45, 46, 55, 66, 67

GSEA Gene Set Enrichment Analysis. iv, v, vi, vii, 6, 11, 16, 17, 18, 19, 22, 23, 24, 25, 32, 37, 38, 39, 40, 44, 45, 53, 63, 64, 67

KEGG Kyoto Encyclopedia of Genes and Genomes. iv, v, vi, vii, 6, 11, 13, 18, 21, 22, 23, 24, 25, 29, 31, 32, 34, 35, 36, 38, 39, 47, 49, 55, 65, 67

logFC logarítmic de Fold Change. 9, 10, 11

logRatio El mateix que logFC. 6, 14, 16, 29, 67

ORA Over-Representation Analysis. iv, v, vi, vii, 6, 11, 14, 18, 19, 22, 23, 24, 25, 32, 33, 34, 35, 36, 37, 49, 51, 52, 58, 63, 65, 66, 67

PT Pathway Topology. 11

Glossari

Bar-Plot La visualització de les rutes enriquides mitjançant ORA amb el nombre de gens diferencialment expressats dins de la ruta i el valor de P. v, vi, 24, 25, 40, 41

Bioconductor Projecte lliure i del codi obert per a anàlisi de les dades genòmiques a base de la llengua de programació R. iv, 1, 2, 3, 4, 6, 14, 16, 21, 25, 26, 27, 29, 54, 67

Dot-Plot La visualització de les rutes enriquides mitjançant ORA amb el nombre de gens diferencialment expressats dins de la ruta i el valor de P a relació amb GeneRatio. v, vi, 24, 41, 42, 59

Enrichment Map La visualització per reduir la redundància als conjunts de gens obtinguts amb ORA. Els cantells indiquen els nodes que tenen gens compartits, on el seu gruix depèn del nombre de gens compartits. v, vi, 24, 25, 32, 43, 60

Gene-Concept-Network La visualització de gens al voltant dels conceptes on els gens poden ser connectats amb rutes diferents. D'aquesta manera es pot veure quins gens contribueixen a què les rutes siguin diferencialment expressades. iv, 19, 25, 61

GO-Plot Directed Acyclic Graph que visualitza els conceptes jeràrquicament mostrant les relacions is-a, part of, i regulates. iv, 19

KEGG Pathway Les rutes descarregades de la base de dades KEGG on els gens enriquits són emfatitzats. iv, 20, 65

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.
- [Clark et al., 2015] Clark, N. R., Szymkiewicz, M., Wang, Z., Monteiro, C. D., Jones, M. R., and Ma'ayan, A. (2015). Principle angle enrichment analysis (paea): Dimensionally reduced multivariate gene set enrichment analysis tool. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 256–262. IEEE.
- [Class et al., 2017] Class, C. A., Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2017). idingo—integrative differential network analysis in genomics with shiny application. *Bioinformatics*, 34(7):1243–1245.
- [Consortium, 2004] Consortium, G. O. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261.
- [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.

Bibliografia

- [Ge and Jung, 2018] Ge, S. and Jung, D. (2018). Shinygo: a graphical enrichment tool for animals and plants. *bioRxiv*, page 315150.
- [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.
- [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.
- [Merico et al., 2010] Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G. D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one*, 5(11):e13984.
- [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.
- [Ruíz de Villa and Sánchez-Pla, 2019] Ruíz de Villa, M. C. and Sánchez-Pla, A. (accessed may 2019). Pid_00192743.
- [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.

Bibliografia

- [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.
- [Westerberg et al., 2006] Westerberg, R., Måansson, J.-E., Golozoubova, V., Shabalina, I. G., Backlund, E. C., Tvrđik, P., Retterstøl, K., Capecchi, M. R., and Jacobsson, A. (2006). Elovl3 is an important component for early onset of lipid recruitment in brown adipose tissue. *Journal of Biological Chemistry*, 281(8):4958–4968.
- [Wickham, 5 15] Wickham, H. (2015 (accessed 2019-05-15)). R packages. <http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm>.