

# PAC2 Desenvolupament del treball - Fase 1

Vasyl Druchkiv

Estudiant del Màster de Bioestadística i Bioinformàtica

15 d'Abril 2019

## Índice

<b>1</b>	<b>Descripció de l'avenç del projecte</b>	<b>2</b>
<b>2</b>	<b>Grau de compliment dels objectius</b>	<b>6</b>
<b>3</b>	<b>L'anàlisi comuna de GO, KEGG i Reactome</b>	<b>9</b>
3.1	ORA . . . . .	9
3.1.1	GO . . . . .	9
3.1.2	KEGG . . . . .	11
3.1.3	Reactome . . . . .	12
3.2	GSEA . . . . .	12
3.2.1	GO . . . . .	12
3.2.2	KEGG . . . . .	14
3.2.3	Reactome . . . . .	14
3.3	Bar-Plots . . . . .	15
3.4	Dot-Plots . . . . .	16
3.5	Enrichment Plots . . . . .	18
3.6	Category-Gene-Network Plot . . . . .	19
3.7	GSEA Plot . . . . .	19
<b>4</b>	<b>L'anàlisi específic de GO, KEGG i Reactome</b>	<b>21</b>
4.1	GO Plot . . . . .	21
4.2	KEGG Pathway . . . . .	22
4.3	Reactome Pathway . . . . .	22
<b>5</b>	<b>Validació dels resultats</b>	<b>23</b>
5.1	Exemple d'anàlisi 1. GEO: GSE100924 . . . . .	23
<b>6</b>	<b>Activitats no previstes</b>	<b>30</b>
	<b>Biblilografia</b>	<b>31</b>

---

## 1 Descripció de l'avenç del projecte

A data d'avui he desenvolupat l'aplicació d'anàlisi de les rutes. 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'anàlisi 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 LogRatios provinents de l'estudi de microarrays o NGS.

Figure 1: Pàgina d'entrada

The screenshot shows the 'Pathway analysis' web application. On the left is a dark sidebar with a 'Data input' icon and three menu items: 'GO Analysis', 'KEGG Analysis', and 'Reactome Analysis', each with a left-pointing arrow. The main content area has a light blue background and is divided into three horizontal sections. The top section is titled 'GO' in large blue letters, followed by 'Select Specie:' and a dropdown menu showing 'Homo Sapiens'. The middle section is titled 'Reactome' in large blue letters, followed by 'Select Specie:' and a dropdown menu showing 'Homo Sapiens'. The bottom section is titled 'KEGG' in large blue letters, followed by 'Enter Search Term for Specie' and a text input field containing 'homo'. Below this is 'Select KEGG Specie' and a dropdown menu showing 'Homo sapiens'. At the bottom of the main area, there are two sections: 'File with all genes' and 'File with selected genes'. Each section has a 'Browse...' button and a 'No file selected' status. A note at the very bottom states: 'Note: The file should contain two columns, one for gene ID (no duplicated ID allowed) and another one for fold change.'

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.

Figure 2: El resum de les dades selecció del *cut off* per a l'anàlisi ORA

**File with all genes**

Browse... Dose\_geneList.csv  
Upload complete

**File with selected genes**

Browse... Dose\_selectedGenes.csv  
Upload complete

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

---

You uploaded: 12495 genes  
First 10 entries

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

You selected: 207 genes  
First 10 entries

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

L'aplicació està dividida doncs en 4 parts substancials:

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

L'aplicació ofereix dos mètodes d'anàlisi: d'una banda es pot fer ORA (Over-Representation Analysis) i d'altra banda l'anàlisi GSEA (Gene Set Enrichment Analysis). 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 suggerides 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 escapen. 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

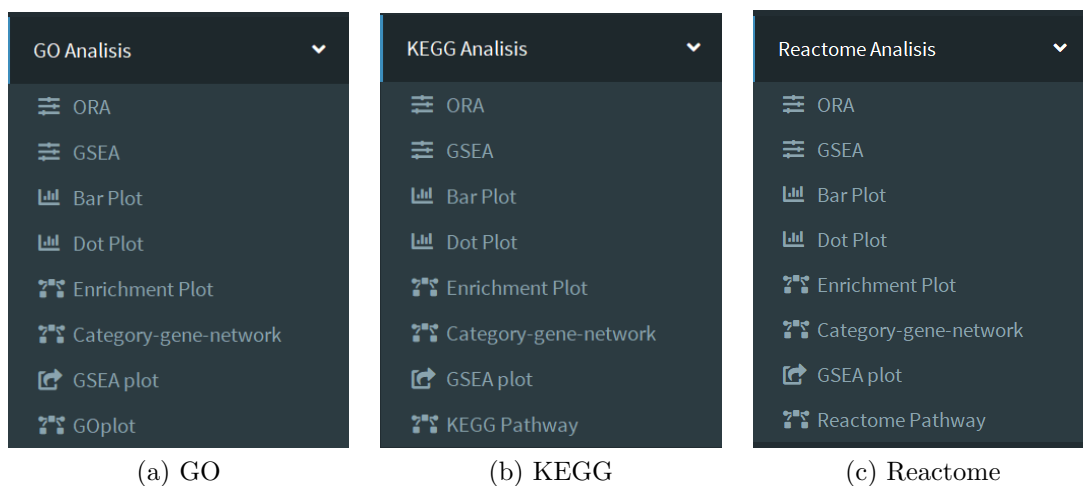


Figure 3: Els elements de les seccions d'anàlisi

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

Table 1: Resum de les anàlisis disponibles i recursos de Bioconductor R

El llistat de tots els paquets i funcions utilitzats en l'aplicació es troba a apèndix A.

## 2 Grau de compliment dels objectius

Recordem el calendari definit al pla de treball.

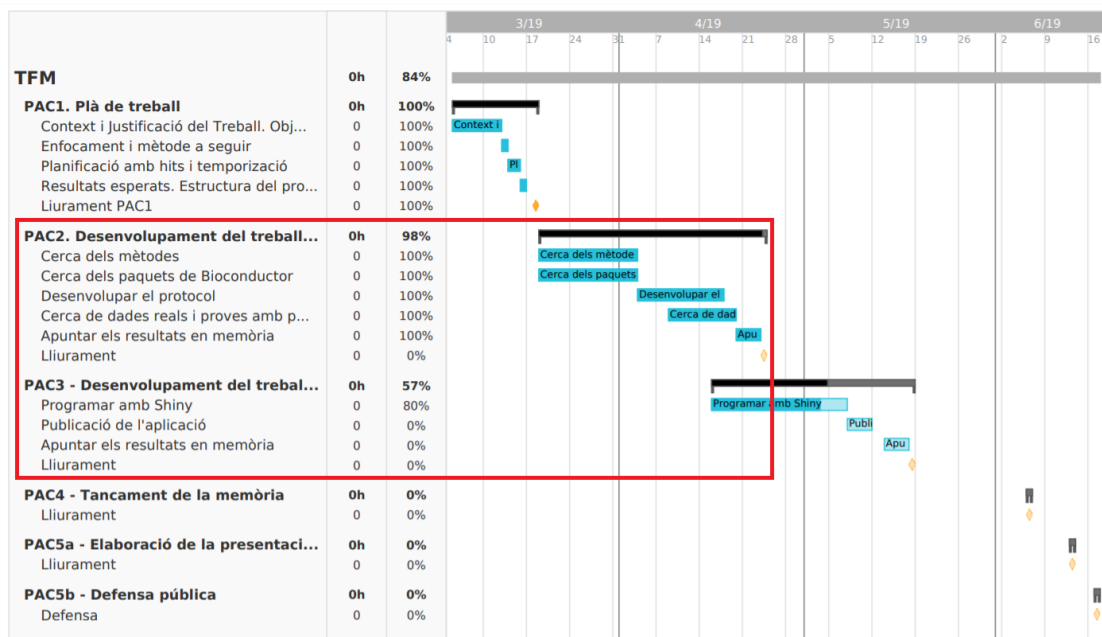


Figure 4: Diagrama de Gantt

Les tasques per a aquesta PAC eren:

### 1. Cerca dels mètodes

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 deixar-ho.

## 2. Cerca dels paquets de Bioconductor

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

## 3. Desenvolupar el protocol

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

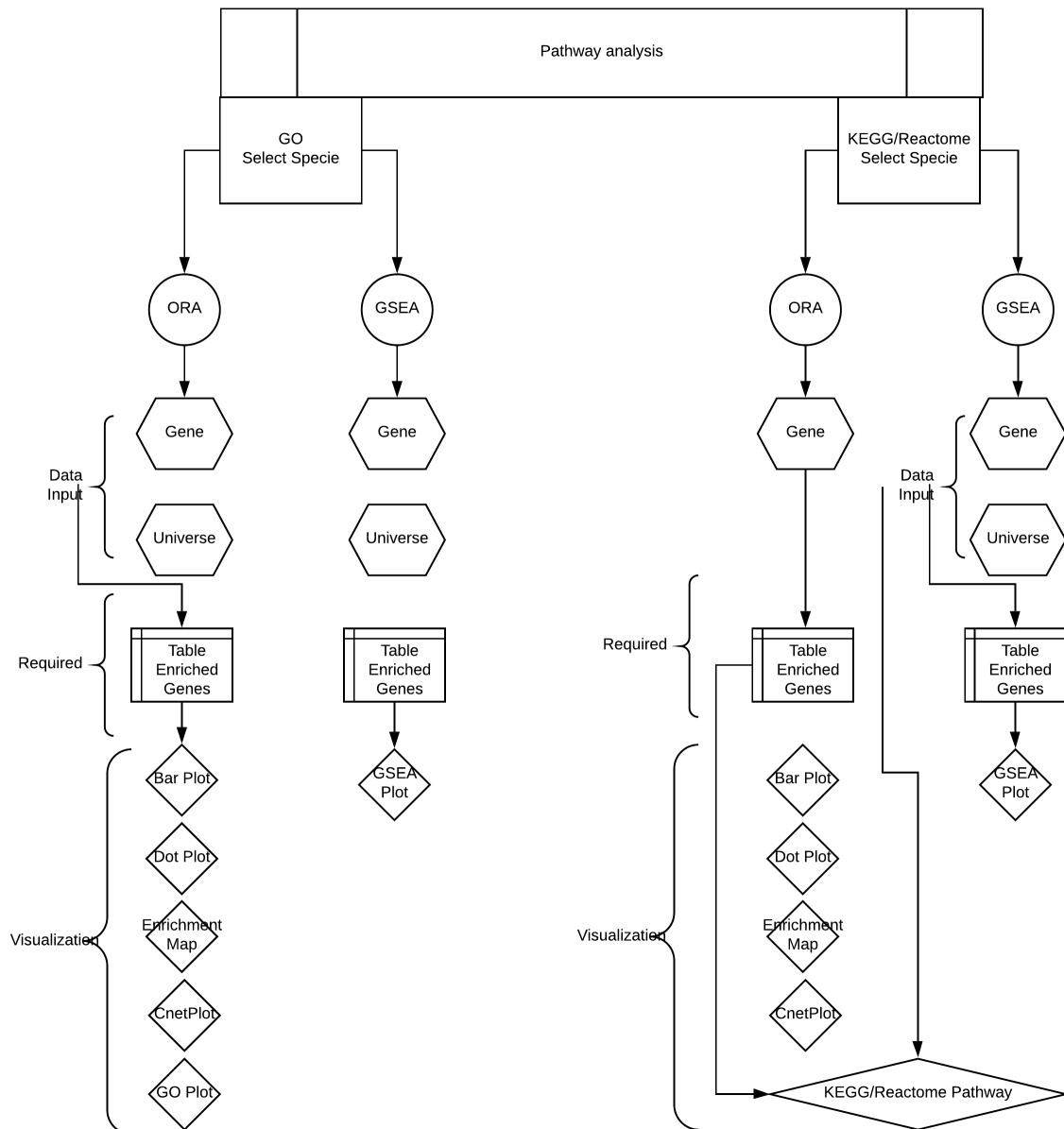


Figure 5: Lucidchart per a l'aplicació

D'aquí podem definir per exemple el protocol:

- Decidir quin anàlisi vol fer: GO, KEGG o Reactome
- Seleccionar l'espècie de referència
- Decidir quin mètode vol implementar: ORA o GSEA i respectivament pujar les dades necessàries.
  - Per a anàlisi GO tots dos arxius són necessaris: Gens Seleccionats (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.
- 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



- (e) 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.

#### 4. Cerca de dades i proves amb el protocol

Aquesta tasca ha resultat ser més difícil. Estic però satisfet que amb l'ajuda del professor he pogut trobar les dades i amb algunes d'elles ja fer les proves. Els detalls els explico en l'apartat Validació dels resultats.

#### 5. Programar amb shiny

Estic satisfet amb el grau del complement amb aquesta tasca. L'aplicació és ja funcional. D'altra banda noto que ara hi ha molt de codi repetitiu. Encara no he trobat la possibilitat de simplificar-lo. Estic buscant però les solucions.

## 3 L'anàlisi comuna de GO, KEGG i Reactome

### 3.1 ORA

#### 3.1.1 GO

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

```
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.

L'execució de la funció és un procés 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

Figure 6: Especificació d'ORA dels termes GO

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.

Els camps més interessants de la taula són:

- 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:  $\frac{\text{Nombre dels gens del conjunt d'interès en tota la mostra}}{\text{Nombre total dels gens en la mostra}} = \frac{k}{n}$ ;

- pvalue. Valor de p basat en la distribució hipergeomètrica:  $p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$
- p.adjust. El valor de P ajustat.

### 3.1.2 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	geneID
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.

### 3.1.3 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								
Data input								
GO Analysis								
KEGG Analysis								
Reactome Analysis								
ORA								
GSEA								
Bar Plot								
Dot Plot								
Enrichment Plot								
Category-gene-network								
GSEA plot								
Reactome Pathway								
Select adjustment method								
BH								
Select P-Value threshold:								
0.1 0.05 0.01 0.001								
Calculate Result								
Download Results as .csv								
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	geneID
R-HSA-2500257	Resolution of Sister Chromatid Cohesion	15/142	124/10554	0.000	2.35e-08	0.000	15	CDCA8/CDC20/CENPE/CCNB2/NDCC80/SKA1/CENP
R-HSA-68877	Mitotic Prometaphase	18/142	198/10554	0.000	2.83e-08	0.000	18	CDCA8/CDC20/CENPE/CCNB2/NDCC80/NCAPH/SK
R-HSA-69620	Cell Cycle Checkpoints	21/142	293/10554	0.000	4.30e-08	0.000	21	CDC45/CDCA8/MCM10/CDC20/CENPE/CCNB2/ND
R-HSA-	Mitotic Spindle	13/142	112/10554	0.000	3.01e-08	0.000	13	CDCA8/CDC20/CENPE/NDCC80/UBE2C/SKA1/CENP

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

## 3.2 GSEA

### 3.2.1 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.

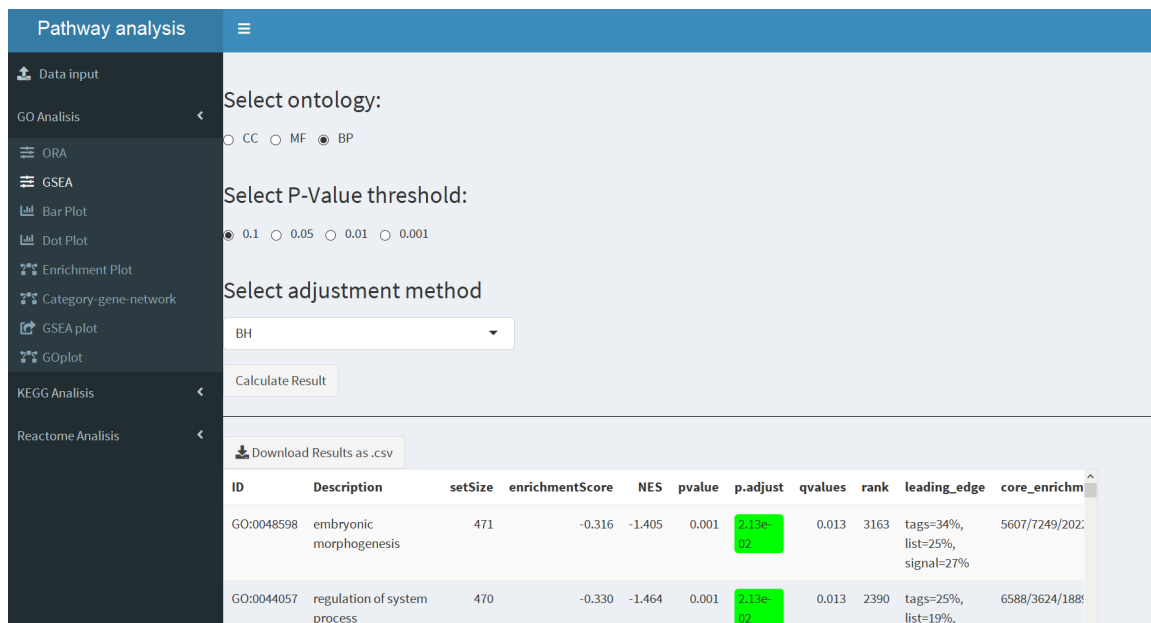


Figure 11: El resultat de l'anàlisi GSEA. GO.

Per entendre l'anàlisi:

- 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 ocurrencies de gens del conjunt específic abans (per als ES positius) o després (per als ES negatius) del cim en la puntuació corrent d'enriquiment. Aquest valor indica el percentatge dels gens que contribueixen a la puntuació d'enriquiment.
  - List. El percentatge dels gens en el llistat ordenat de tots els gens abans o després del pic en la puntuació corrent d'enriquiment. Aquest valor ens indica on exactament el pic es produeix.
  - Signal. La fortalesa del senyal d'enriquiment que combina els dos valors anteriors.

- **rank**. La posició del pic en la llista ordenada dels gens. Els conjunts dels gens més interessants assoleixen el seu màxim o bé al principi o al final de la llista ordenada. Vol dir que tenen aquest valor o bé molt baix o bé molt alt.

### 3.2.2 KEGG

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

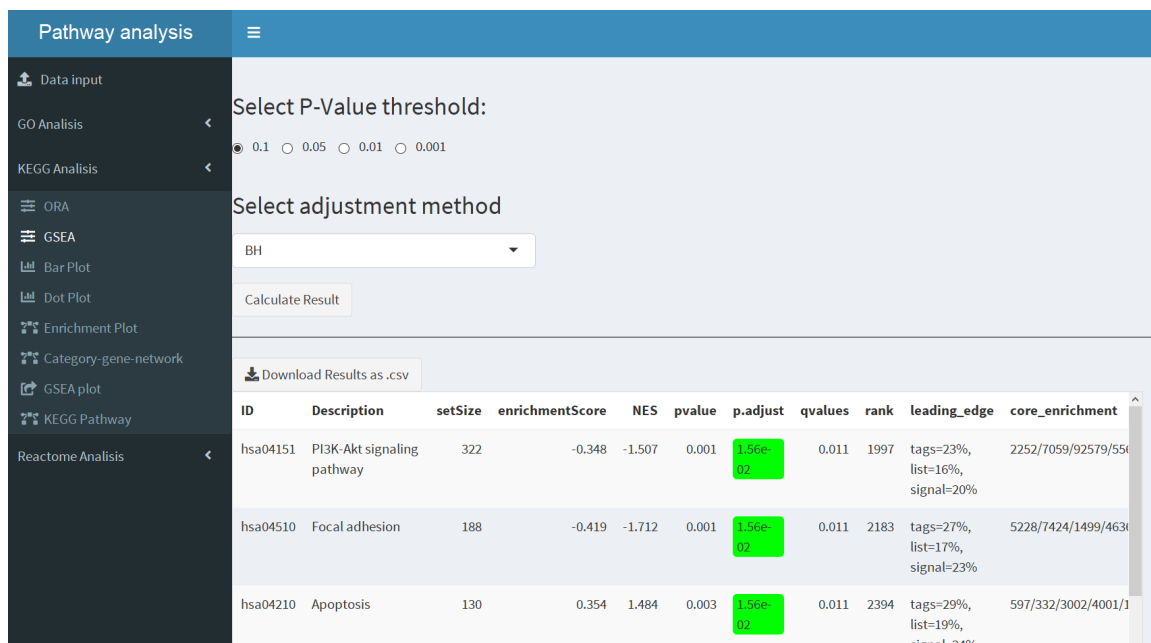


Figure 12: El resultat de l'anàlisi GSEA. KEGG.

### 3.2.3 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()`

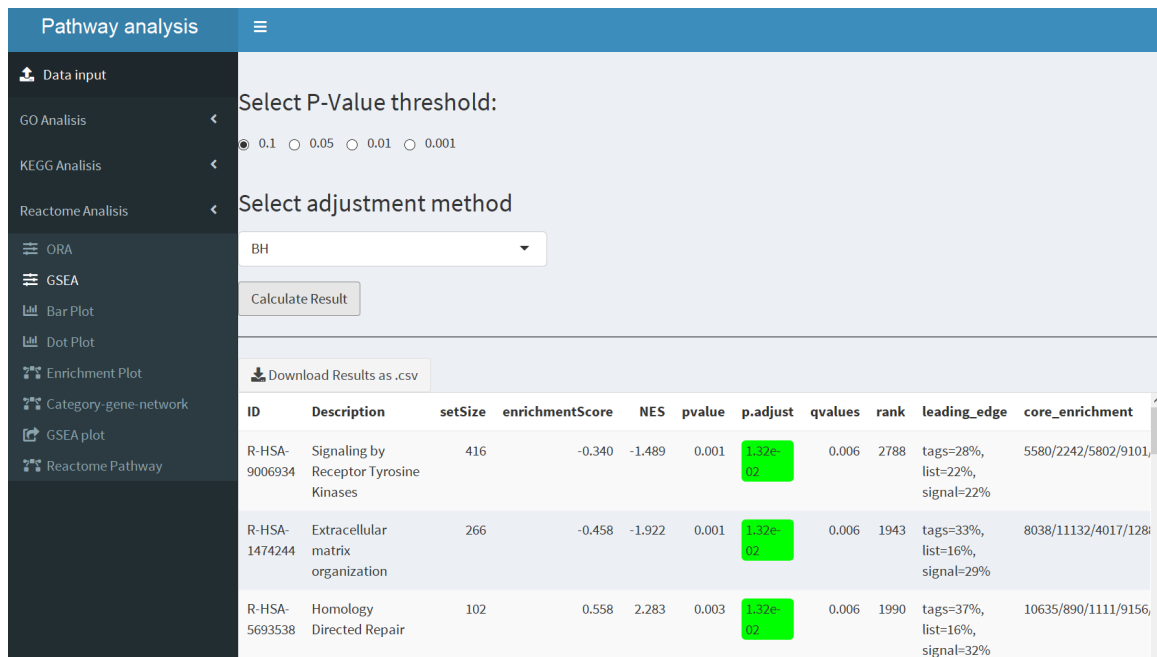


Figure 13: El resultat d'anàlisi GSEA. Reactome.

### 3.3 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.

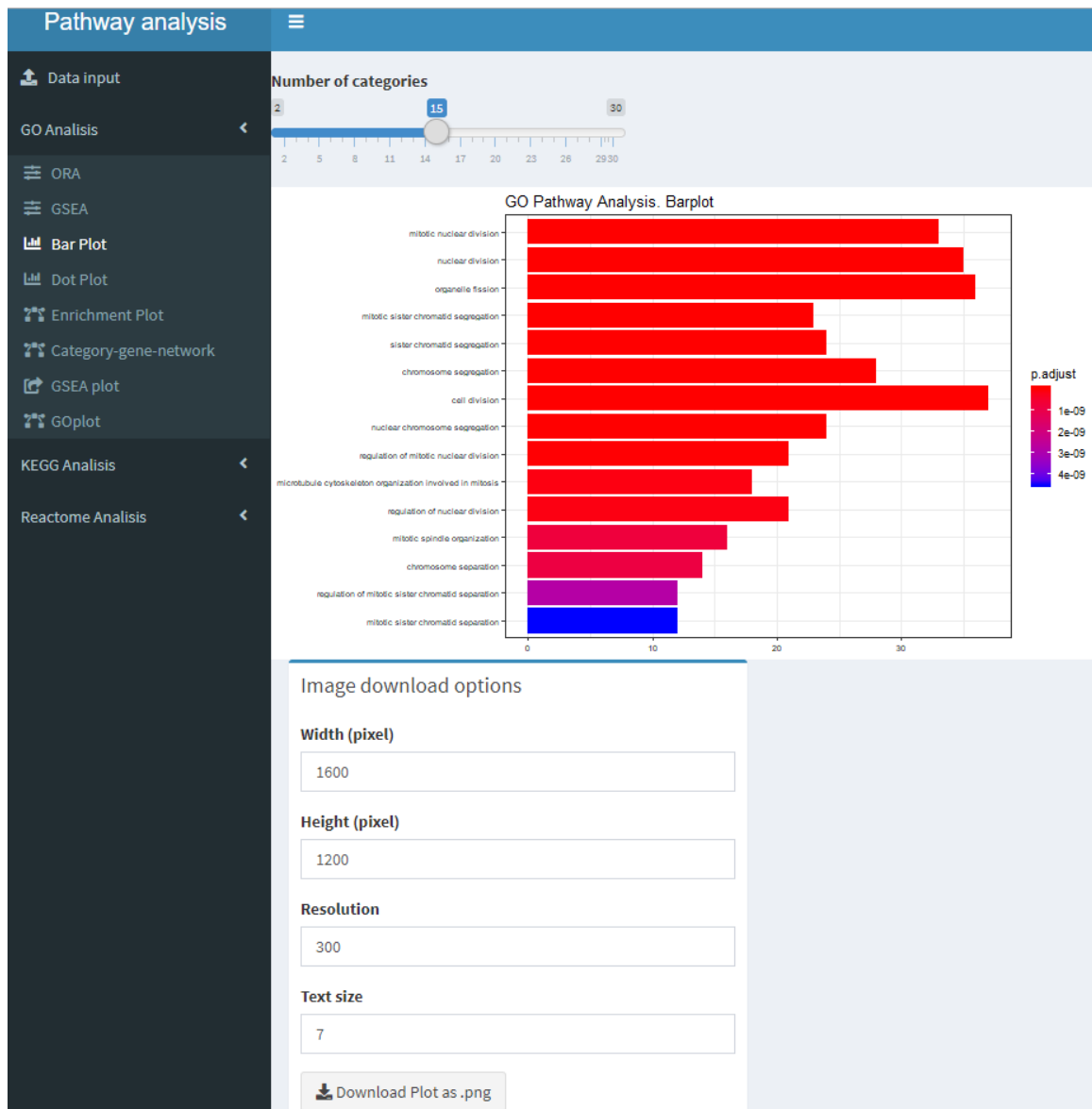


Figure 14: Bar-Plot. GO.

### 3.4 Dot-Plots

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



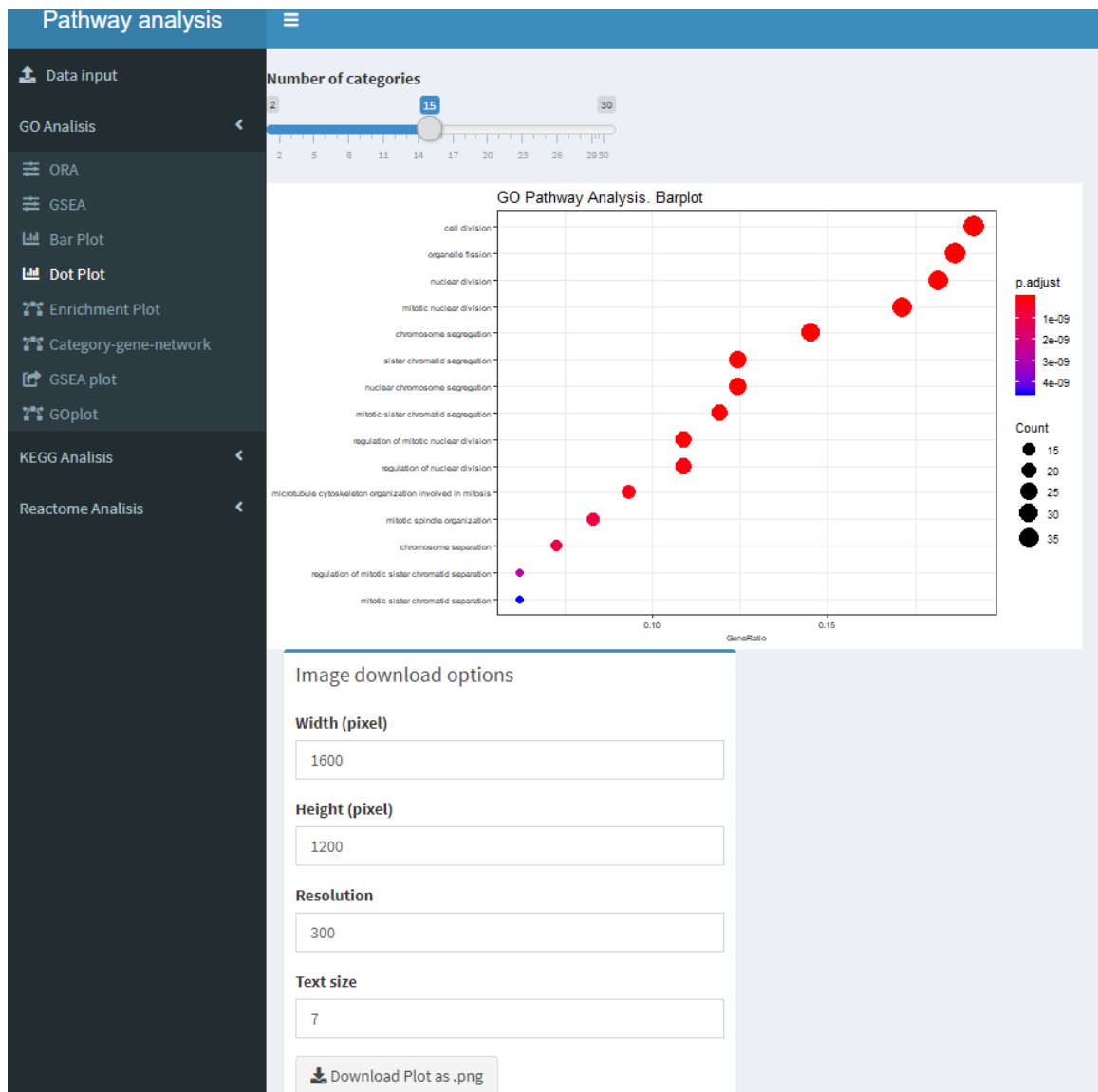


Figure 15: Bar-Plot. GO.

### 3.5 Enrichment Plots

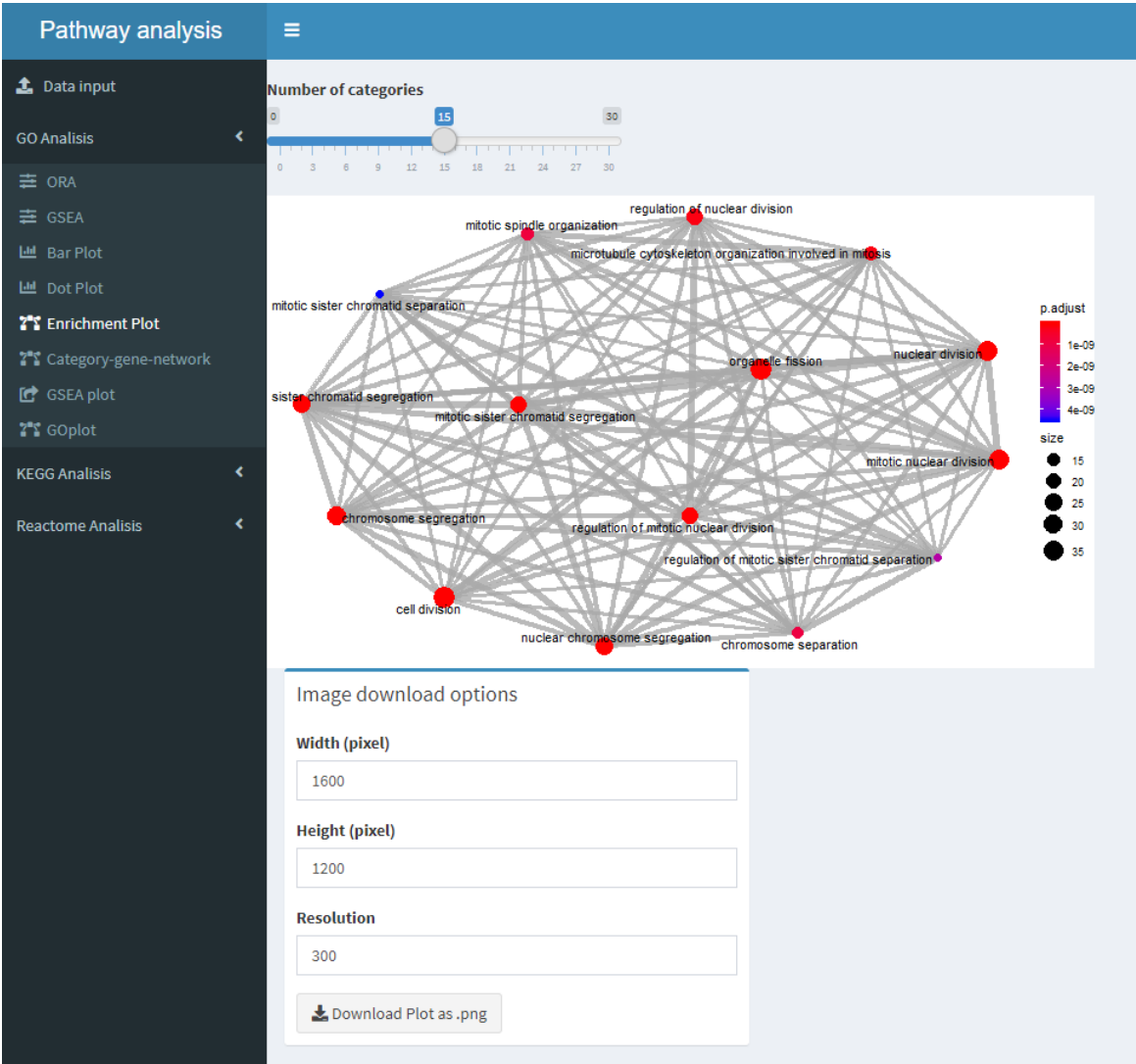


Figure 16: Bar-Plot. GO.

### 3.6 Category-Gene-Network Plot

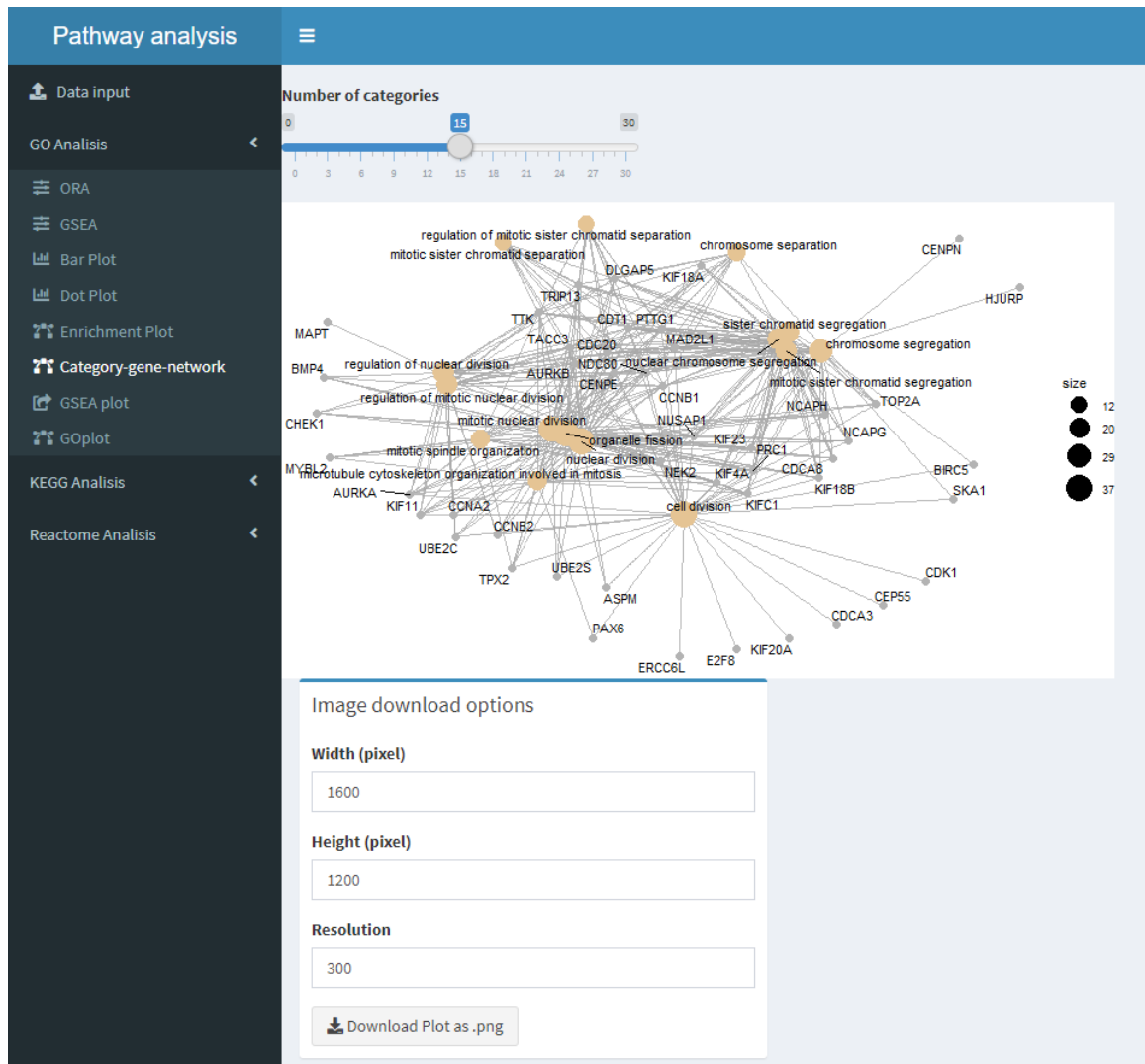


Figure 17: Category-Gene-Network Plot. GO.

### 3.7 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*

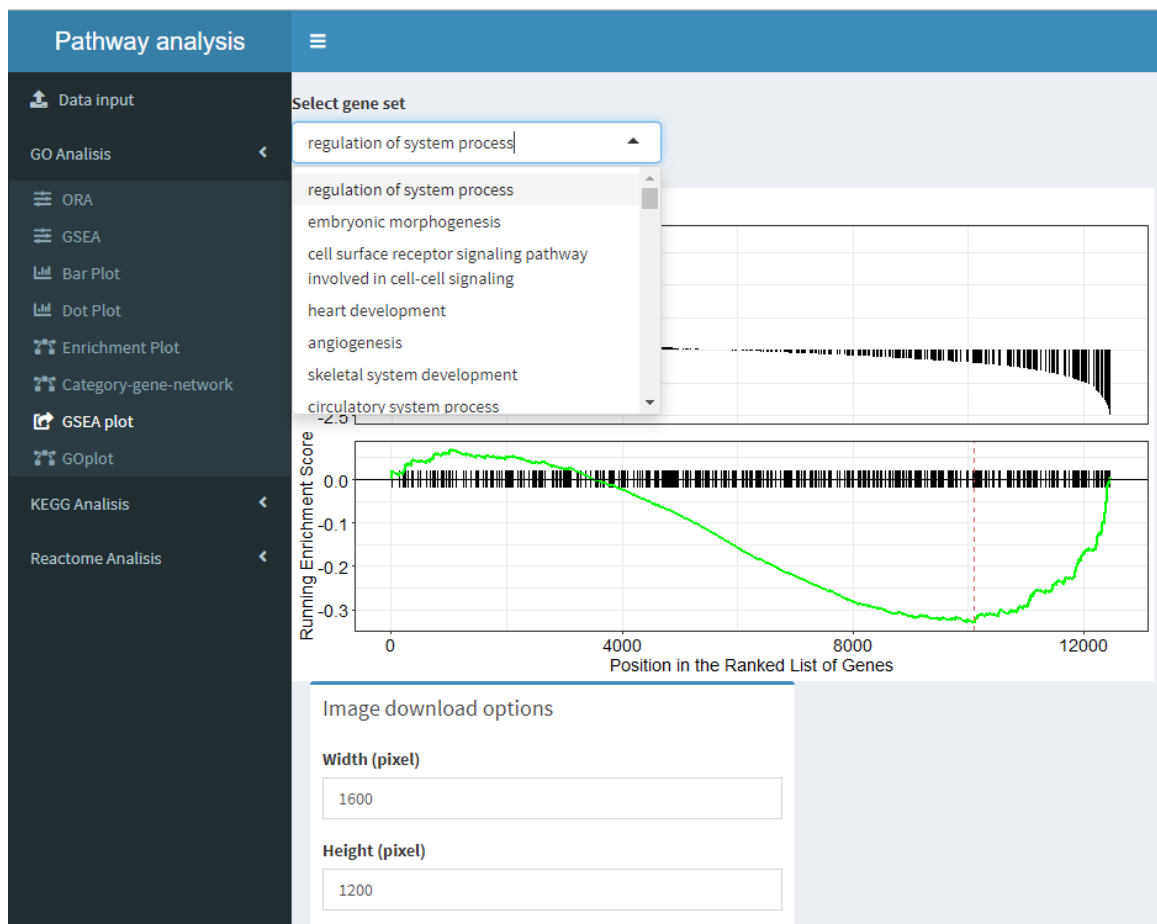


Figure 18: GSEA Plot. GO.



## 4.2 KEGG Pathway

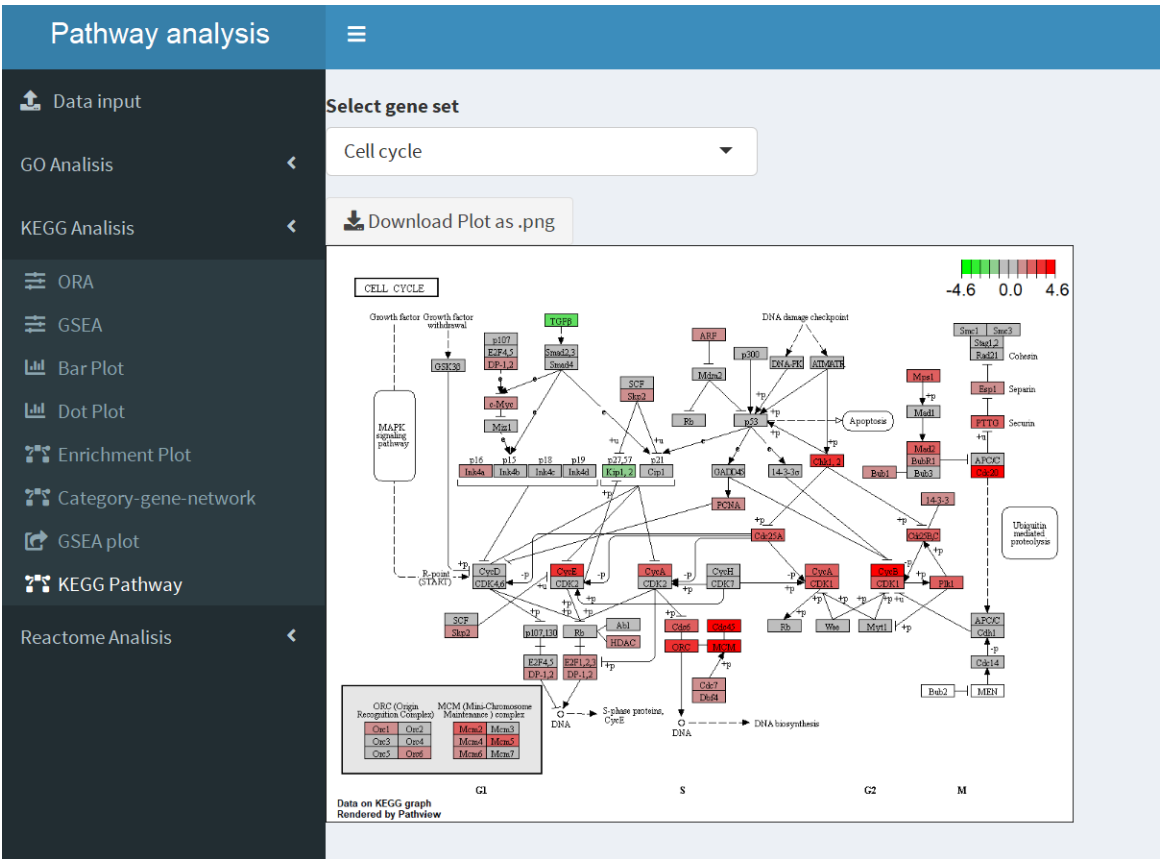


Figure 20: KEGG pathway

## 4.3 Reactome Pathway

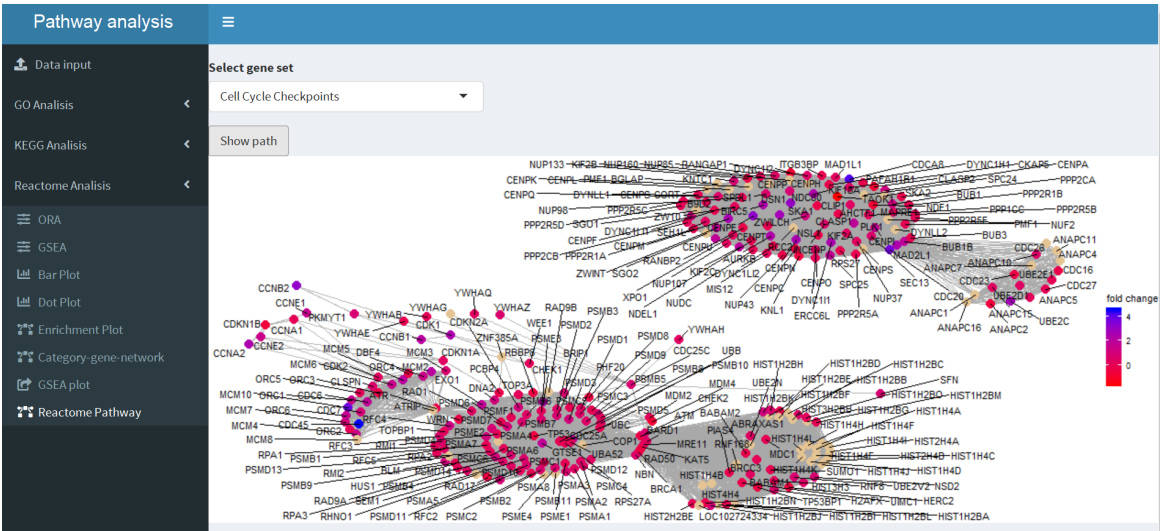


Figure 21: Reactome pathway

## 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, càlcul de les LogRatios, ajustament dels gens repetits a l'array, selecció dels gens diferencialment expressats, etc.). Les dades de GEO (Gene Expression Omnibus) estan però disponibles com a màxim en format normalitzat. Caldria doncs fer una anàlisi per arribar a un llistat de gens diferencialment expressats amb les logRatios per tots els gens de la mostra. Fer això no seria cap problema i de fet ho he fet per altres estudis. El problema és que arribo a resultats diferents dels resultats dels estudis d'on provenen les dades (i no parlo de l'anàlisi de les rutes sinó ja del 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	Espècie	Tipo d'experiment	Font
[Schmidt et al., 2008]	GSE11121	Homo sapiens	Microarrays	Paquet <b>DOSE</b> 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] .

### 5.1 Exemple d'anàlisi 1. GEO: GSE100924

Les dades d'estudi [Li et al., 2017] són ja preprocessades per Ricardo Gonzalo Sanz i Sanchez Pla i estan disponibles a github. De la carpeta *results* he agafat la taula *topAnnotated\_KOvsWT\_COLD.csv*. Sanz i Pla utilitzen el paquet **ReactomePA** per a l'anàlisi d'enriquiment. Repeteixo doncs el seu anàlisi utilitzant l'aplicació.

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

**Pathway analysis**

**Data input**

GO Analysis <

KEGG Analysis <

Reactome Analysis <

**GO**

Select Species:

Mouse

**Reactome**

Select Species:

Mouse

**KEGG**

Enter Search Term for Specie

mus

Select KEGG Specie

Mus musculus

**File with all genes**

Browse... topAnnotated\_KOvsWT\_COLD\_ur

Upload complete

**File with selected genes**

Browse... topAnnotated\_KOvsWT\_COLD\_ge

Upload complete

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

Figure 22: 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 23: 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*



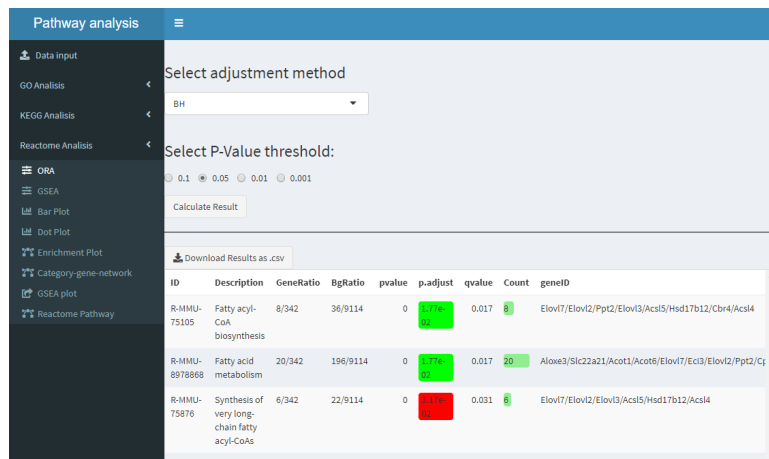


Figure 24: Resultat d'anàlisi ORA de Reactome

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

### 3. Visualització del resultat ORA

- Selecciono *Reactome Analysis* → *Bar Plot*

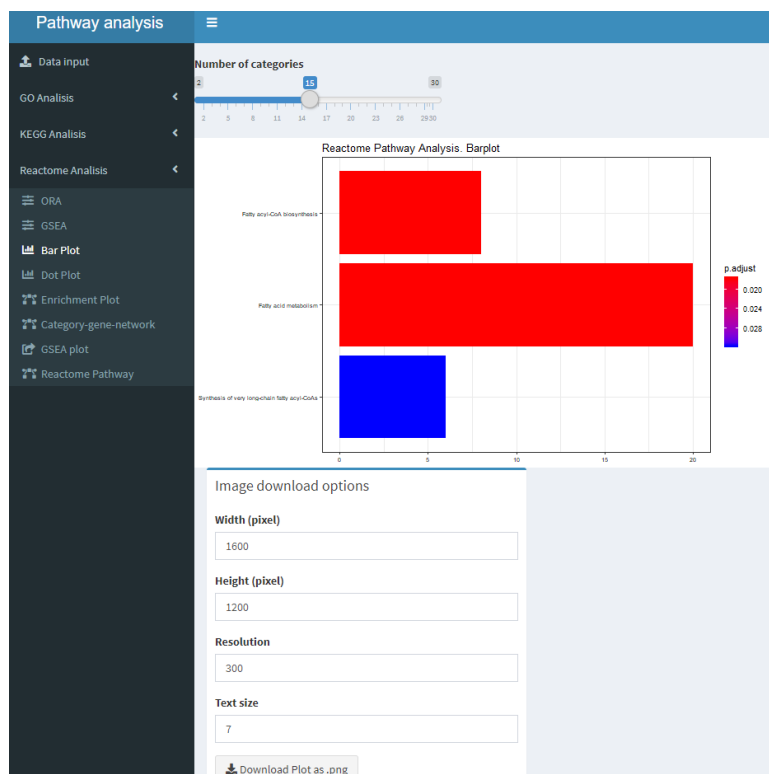


Figure 25: Gràfic de barres

- Selecciono *Reactome Analysis* → *Dot Plot*

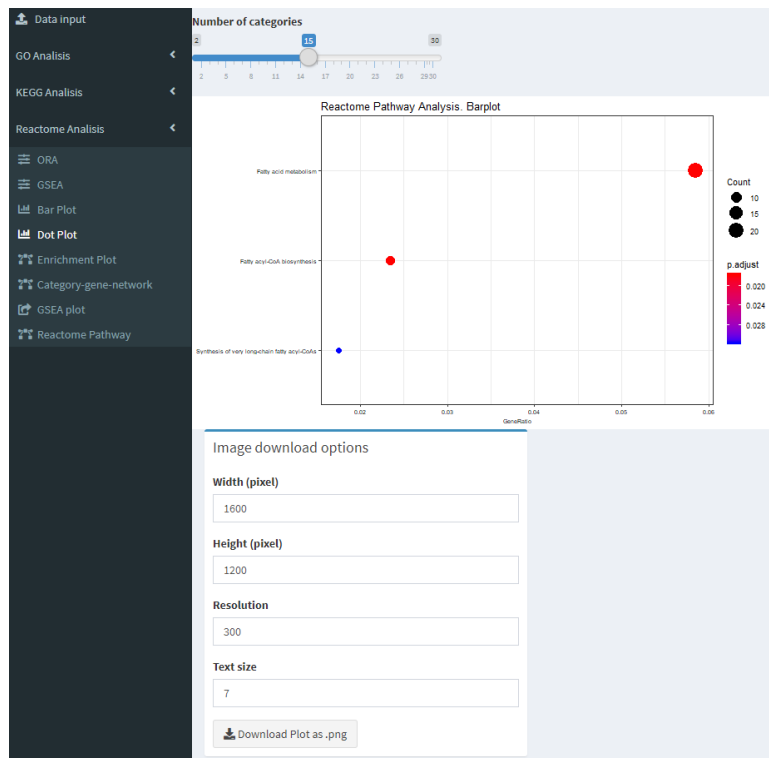


Figure 26: Gràfic de punts

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

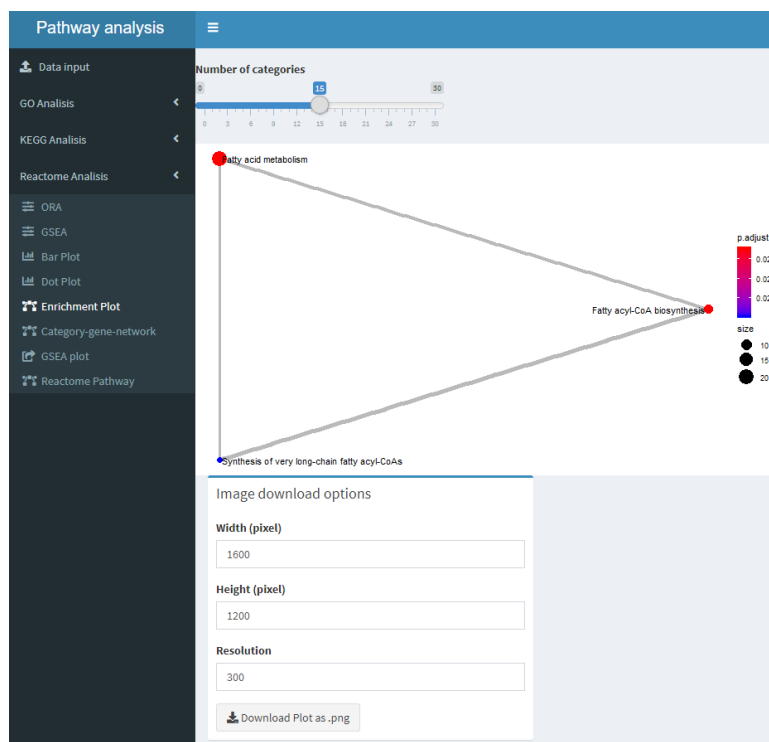


Figure 27: Mapa d'enriquement

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

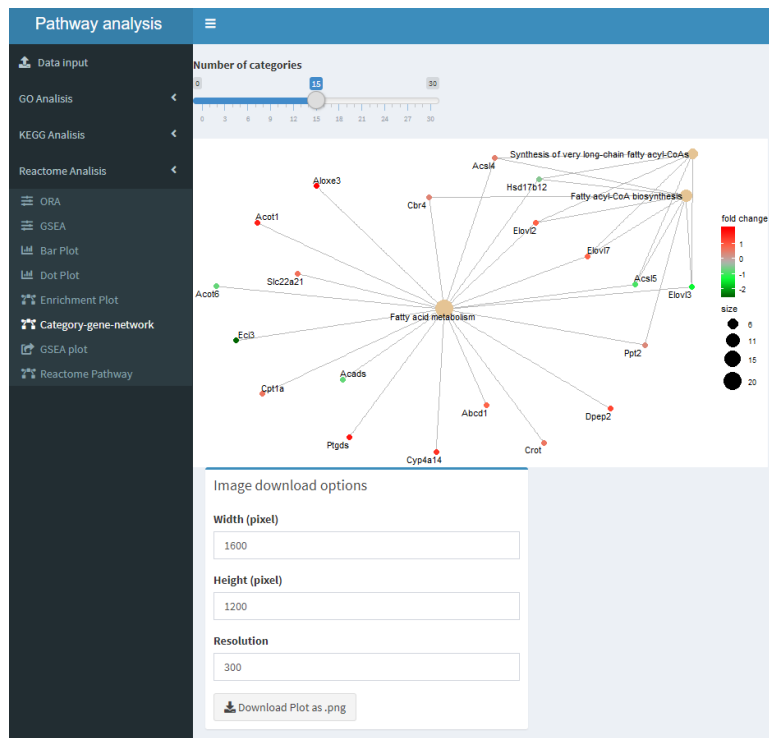


Figure 28: Red de les categories i gens

- Selecciono *Reactome Analysis*→*Reactome Pathway*

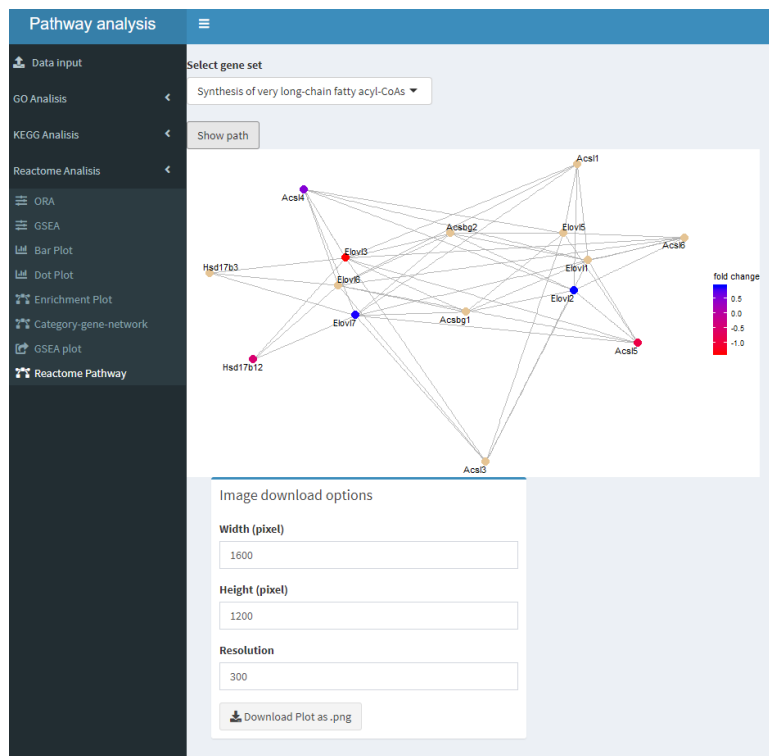


Figure 29: Rutes Reactome

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.

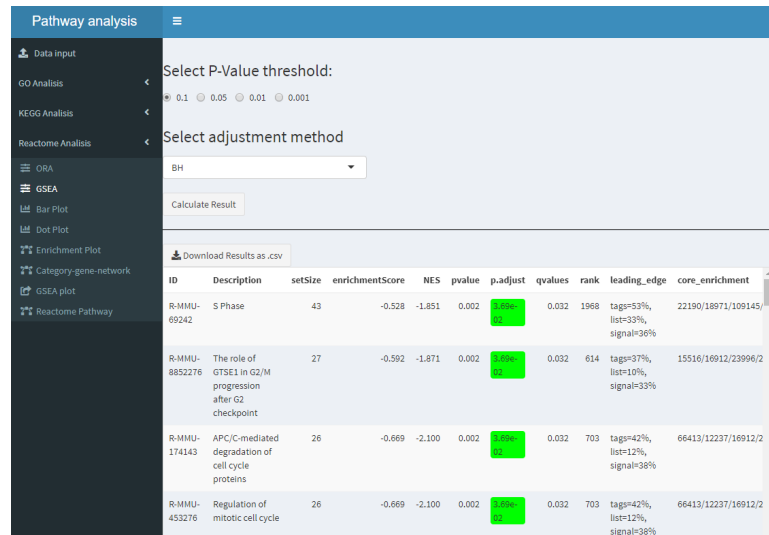


Figure 30: Anàlisi GSEA

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

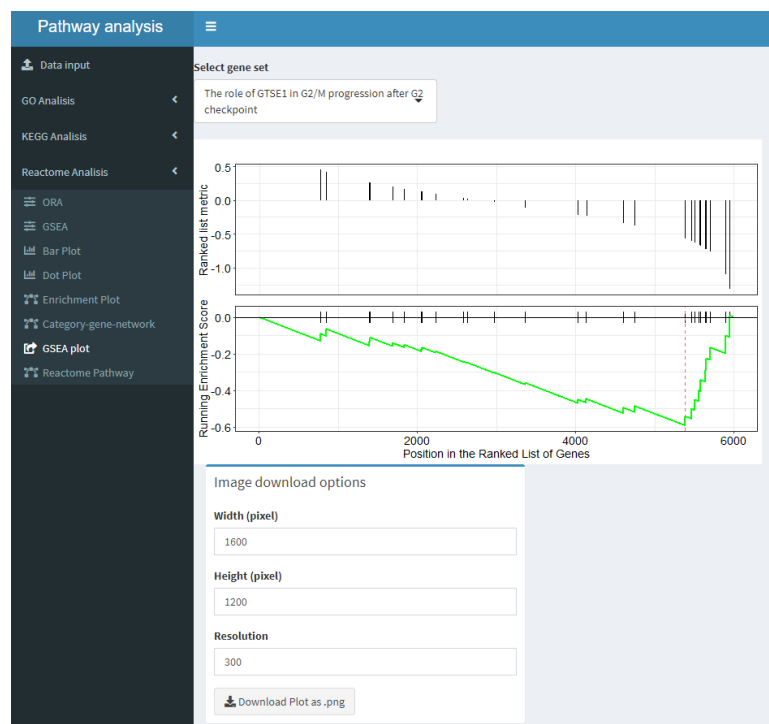


Figure 31: Gràfic GSEA

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*

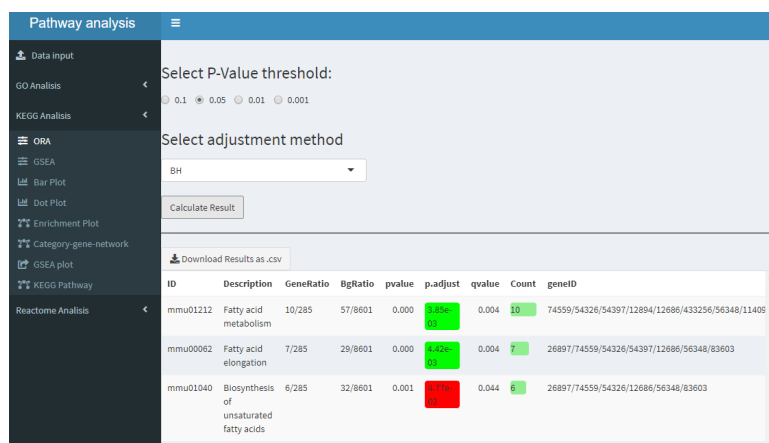


Figure 32: Anàlisi ORA de KEGG

2. Anem a *KEGG*→*KEGG Pathway*

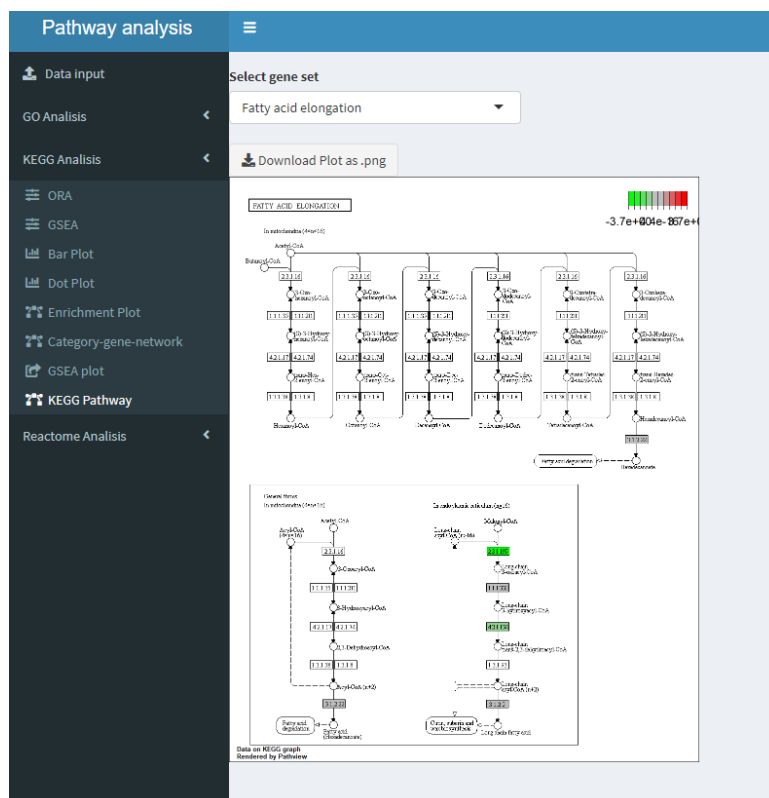


Figure 33: Gràfic de les rutes KEGG

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*

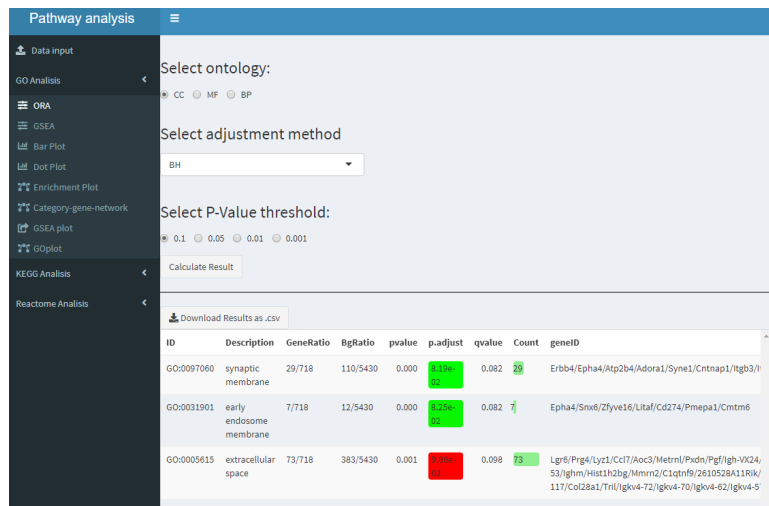


Figure 34: L'anàlisi ORA de GO

## 6 Activitats no previstes

Treballant en el projecte he notat la necessitat de fer tot el procés més segur. El moment clau era quan no he pogut trobar l'USB on he guardat el meu projecte. Ho tenia en un USB perquè hi treballava ded de molts ordinadors diferents: de casa, de la feina, en un portàtil quan era de viatge. Per tan he decidit guardar tot el projecte en github.com. He creat un repositori al qual puc accedir des d'ordinadors diferents.

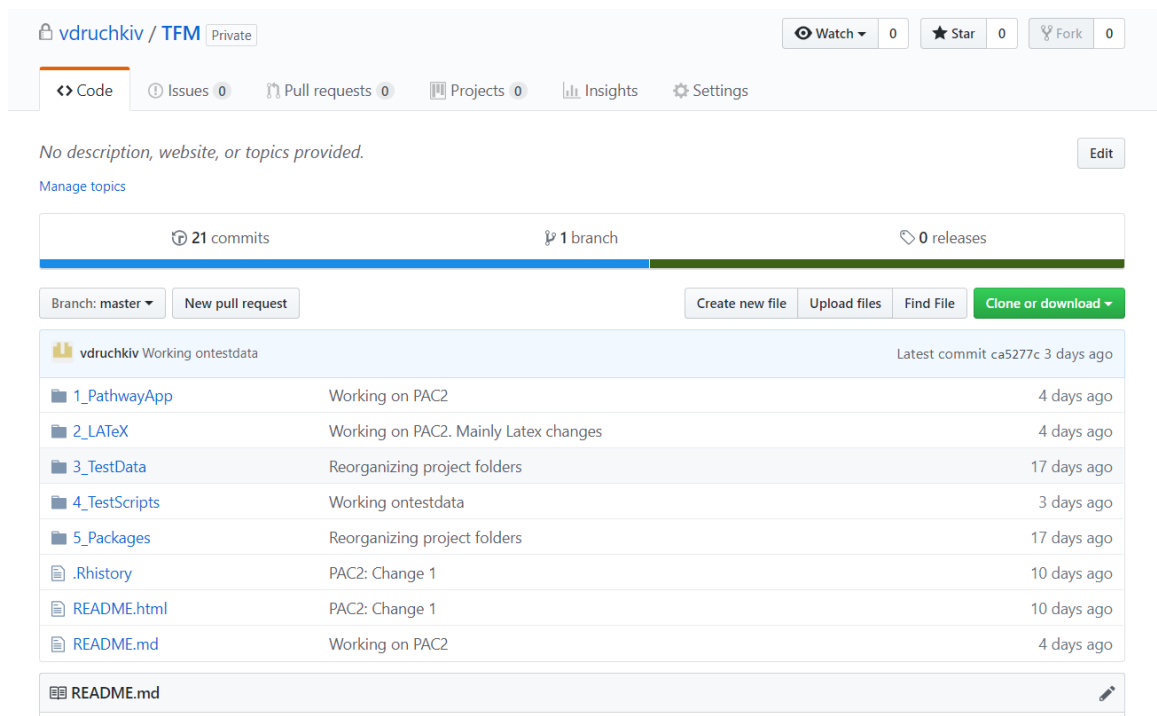


Figure 35: Github repositori del TFM

Utilitzant el GitBash es pot documentar els canvis (**commit**) i pujar (**push**) o baixar (**pull**) els arxius. Així el treball en el projecte és més segur i pràctic.

# Bibliografia

## References

- [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.
- [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.
- [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.
- [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.
- [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 Gehrman, 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.

## Apèndix

### A Els paquets i funcions utilitzats en el app

Package	Funcion
c("package:clusterProfiler", "package:ReactomePA")	cnetplot
c("package:clusterProfiler", "package:ReactomePA")	dotplot
c("package:clusterProfiler", "package:ReactomePA")	emapplot
c("package:clusterProfiler", "package:ReactomePA")	gseaplot
c("package:dplyr", "package:stats")	filter
c("package:kableExtra", "package:knitr")	kable
c("package:shinydashboard", "package:graphics")	box
character(0)	enrichrekegg
character(0)	enrichresgo
character(0)	enrichresgo_gsea
character(0)	enrichreskegg
character(0)	enrichreskegg_gsea
character(0)	enrichresRA
character(0)	enrichresRA_gsea
character(0)	geneList
character(0)	genes
character(0)	kegg_organism1
character(0)	kegg_organism2
character(0)	PathPlotRA
character(0)	pathview
package:base	abs
package:base	as.character
package:base	as.numeric
package:base	c
package:base	file.copy
package:base	formatC
package:base	length
package:base	library
package:base	list
package:base	max
package:base	names
package:base	ncol
package:base	nrow
package:base	paste0
package:base	print
package:base	return
package:base	sort
package:base	tempdir
package:base	tempfile
package:clusterProfiler	enrichGO
package:clusterProfiler	enrichKEGG
package:clusterProfiler	goplot



package:clusterProfiler	gseGO
package:clusterProfiler	gseKEGG
package:clusterProfiler	search_kegg_organism
package:dplyr	everything
package:dplyr	mutate
package:dplyr	select
package:formattable	color_bar
package:formattable	color_tile
package:graphics	barplot
package:grDevices	dev.off
package:grDevices	png
package:kableExtra	kable_styling
package:kableExtra	scroll_box
package:png	readPNG
package:ReactomePA	enrichPathway
package:ReactomePA	gsePathway
package:ReactomePA	viewPathway
package:shiny	actionButton
package:shiny	downloadButton
package:shiny	downloadHandler
package:shiny	eventReactive
package:shiny	fileInput
package:shiny	fluidRow
package:shiny	h3
package:shiny	hr
package:shiny	HTML
package:shiny	htmlOutput
package:shiny	icon
package:shiny	imageOutput
package:shiny	numericInput
package:shiny	radioButtons
package:shiny	reactive
package:shiny	renderImage
package:shiny	renderText
package:shiny	renderUI
package:shiny	req
package:shiny	selectInput
package:shiny	shinyApp
package:shiny	sliderInput
package:shiny	strong
package:shiny	tableOutput
package:shiny	textInput
package:shiny	textOutput
package:shiny	uiOutput
package:shinycssloaders	withSpinner
package:shinydashboard	dashboardBody
package:shinydashboard	dashboardHeader
package:shinydashboard	dashboardPage
package:shinydashboard	dashboardSidebar

package:shinydashboard	menuItem
package:shinydashboard	sidebarMenu
package:shinydashboard	tabItem
package:shinydashboard	tabItems
package:utils	head
package:utils	read.csv
package:utils	write.csv

---