

# PAC2 Desenvolupament del treball - Fase 1

Vasyl Druchkiv

Estudiant del Màster de Bioestadística i Bioinformàtica

15 d'Abril 2019

## Índice

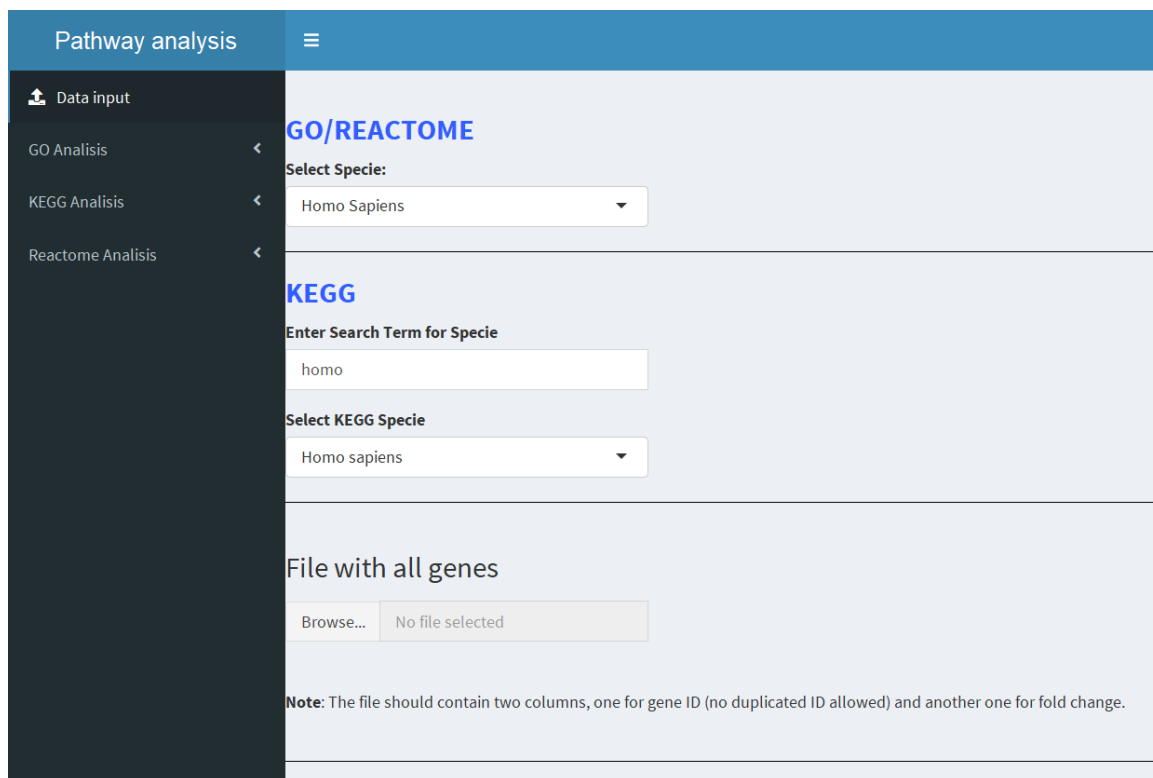
<b>1</b>	<b>Descripció de l'avenç del projecte</b>	<b>1</b>
<b>2</b>	<b>L'anàlisi comuna de GO, KEGG i Reactome</b>	<b>4</b>
2.1	ORA . . . . .	4
2.1.1	GO . . . . .	4
2.1.2	KEGG . . . . .	6
2.1.3	Reactome . . . . .	7
2.2	GSEA . . . . .	8
2.2.1	GO . . . . .	8
2.2.2	KEGG . . . . .	9
2.2.3	Reactome . . . . .	10
2.3	Bar-Plots . . . . .	10
2.4	Dot-Plots . . . . .	11
2.5	Enrichment Plots . . . . .	12
2.6	Category-Gene-Network Plot . . . . .	12
2.7	GSEA Plot . . . . .	13
<b>3</b>	<b>L'anàlisi específic de GO, KEGG i Reactome</b>	<b>14</b>
3.1	GO Plot . . . . .	14
3.2	KEGG Pathway . . . . .	15
3.3	Reactome Pathway . . . . .	15
<b>4</b>	<b>Validació dels resultats</b>	<b>16</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 GO i Reactome l'usuari pot elegir entre "human", "rat", "mouse", "celegans", "yeast", "zebrafish", "fly". 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



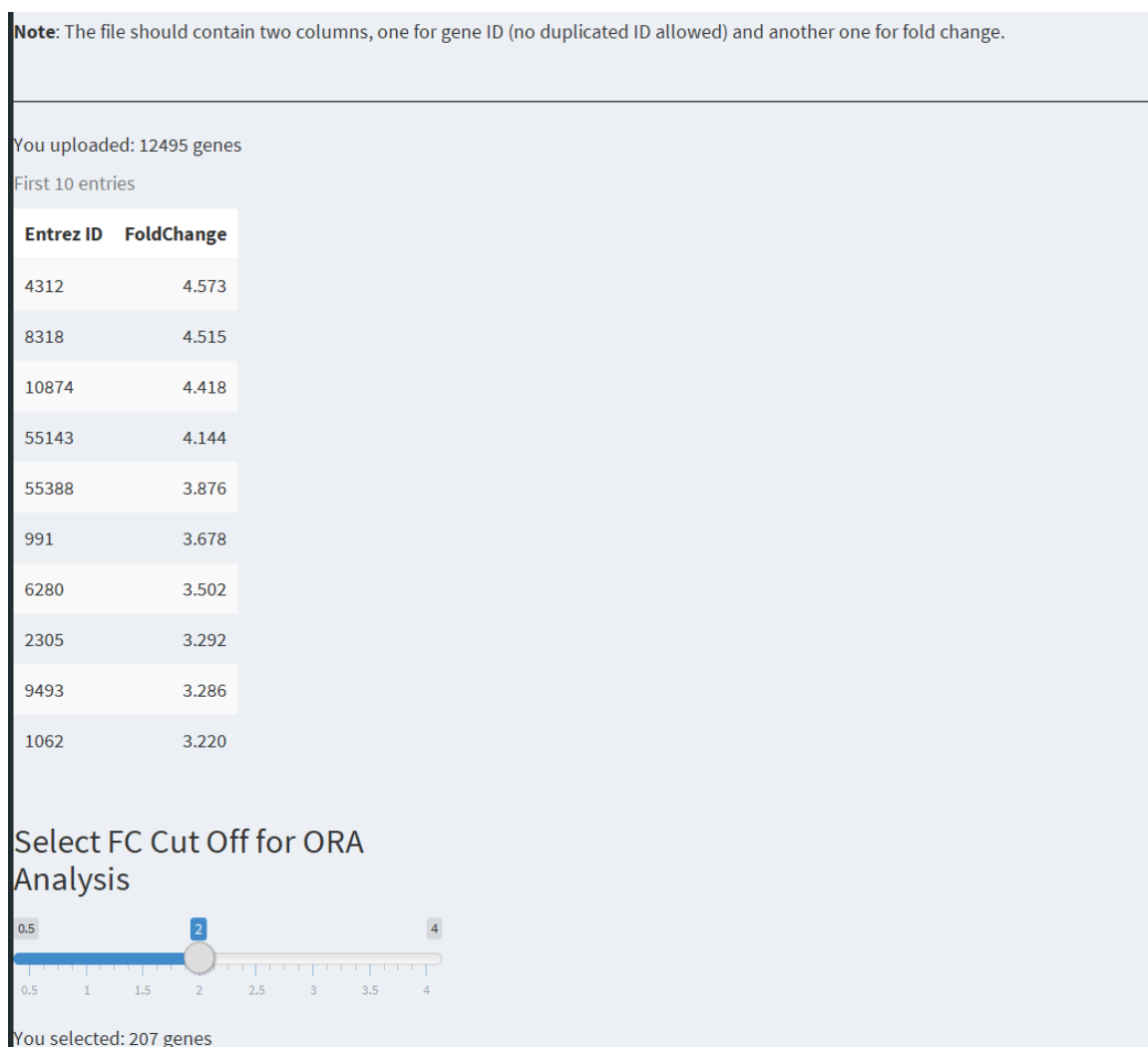
Un cop introduïdes les dades es mostra un petit resum del contingut i es dona la possibilitat d'elegir el *cut-off* de FoldChange per a l'anàlisi ORA. Per defecte s'agafa el valor de FoldChange=2. En funció del *cut-off* elegit es mostra el nombre dels gens en aquest grup (gens sobre o sotaexpressats).

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 sotaexpressada en els gens seleccionats. Per dur

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



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

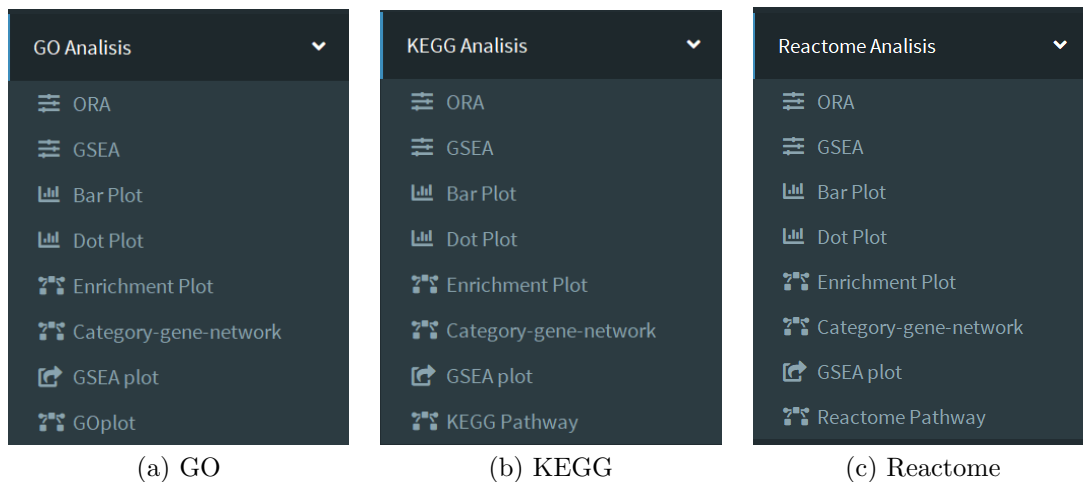


Figure 3: Els elements de les seccions d'anàlisi

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()	Nececita l'objecte del class enrichResult
GO	Enrichment Map	enrichplot	emapplot()	Nececita l'objecte del class enrichResult
GO	Gene-Concept Network	enrichplot	cnetplot()	Nececita l'objecte del class enrichResult
GO	GO directed acyclic graph	enrichplot	goplot()	Nececita 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()	Nececita l'objecte del class enrichResult
KEGG	Enrichment Map	enrichplot	emapplot()	Nececita l'objecte del class enrichResult
KEGG	Gene-Concept Network	enrichplot	cnetplot()	Nececita 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()	Nececita l'objecte del class enrichResult
Reactome	Enrichment Map	enrichplot	emapplot()	Nececita l'objecte del class enrichResult
Reactome	Gene-Concept Network	enrichplot	cnetplot()	Nececita l'objecte del class enrichResult
Reactome	Pathway	ReactomePA	viewPathway()	Molt lent. Per aquest motiu no he afegit botó de descarga.

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

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

### 2.1 ORA

#### 2.1.1 GO

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

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

He implementat els valors per defecte amb la possibilitat per a 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.

The screenshot shows a web application titled "Pathway analysis". On the left is a dark sidebar with a menu containing: "Data input", "GO Analysis", "ORA", "GSEA", "Bar Plot", "Dot Plot", "Enrichment Plot", "Category-gene-network", "GSEA plot", "GOplot", "KEGG Analysis", and "Reactome Analysis". The "GO Analysis" option is selected, indicated by a chevron. The main content area is light blue and contains three sections: "Select ontology:" with radio buttons for "CC", "MF", and "BP" (where "BP" is selected); "Select adjustment method" with a dropdown menu showing "BH"; and "Select P-Value threshold:" with radio buttons for "0.1", "0.05", "0.01", and "0.001" (where "0.1" is selected). A "Calculate Result" button is located below the threshold selection.

Figure 4: Especificació d'ORA dels termes GO

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 consient de repetir l'anàlisi amb altres valors.

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

Calculate Result								
Download Results as .csv								
ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	geneID
GO:0140014	mitotic nuclear division	33/193	232/11468	0.000	4.00e-18	0.000	33	CDCA8/CDC20/KIF23/CENPE/MYBL2/CCNB2/...
GO:0000280	nuclear division	35/193	316/11468	0.000	4.50e-16	0.000	35	CDCA8/CDC20/KIF23/CENPE/MYBL2/CCNB2/...

Figure 5: 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.

### 2.1.2 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)
```

**Pathway analysis**

- Data input
- GO Analysis
- KEGG Analysis
- ORA
- GSEA
- Bar Plot
- Dot Plot
- Enrichment Plot
- Category-gene-network
- GSEA plot
- KEGG Pathway
- Reactome Analysis

Select P-Value threshold:

☒ 0.1 ☐ 0.05 ☐ 0.01 ☐ 0.001

Select adjustment method

holm

Calculate Result

Figure 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.

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 7: El resultat de l'anàlisi ORA. KEGG.

### 2.1.3 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)
```

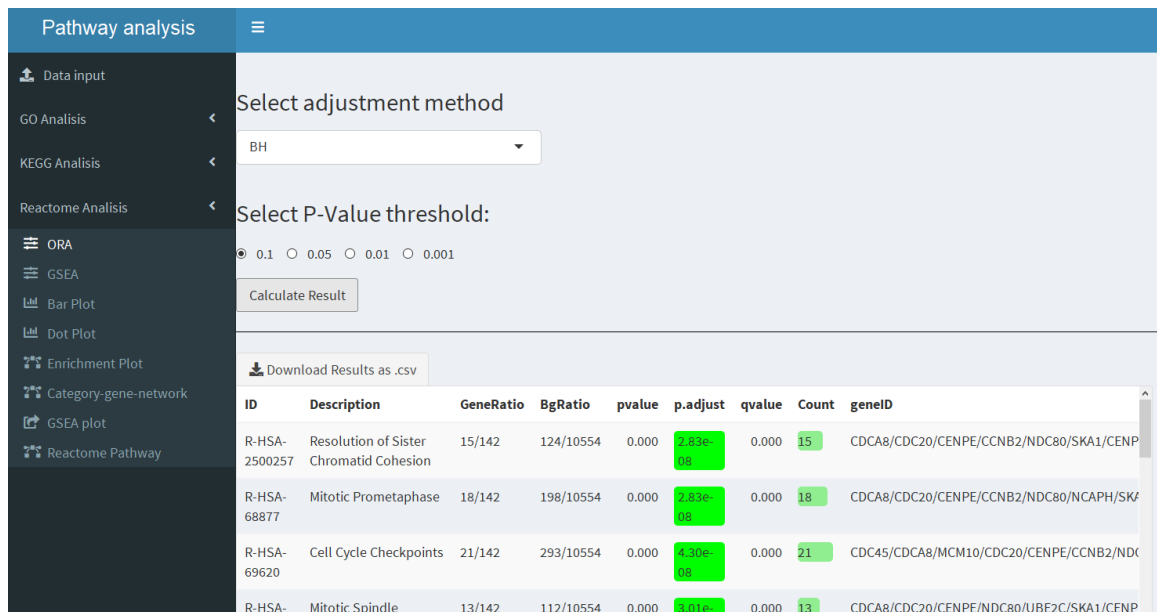


Figure 8: El resultat d'anàlisi ORA. Reactome.

## 2.2 GSEA

### 2.2.1 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.



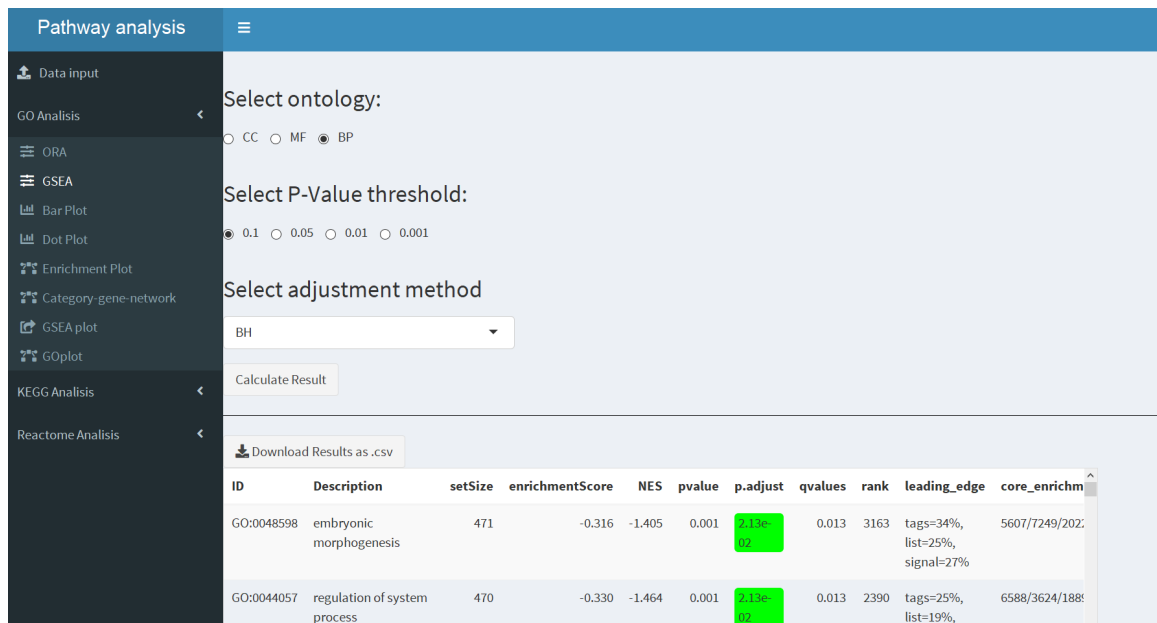


Figure 9: El resultat de l'anàlisi GSEA. GO.

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 ocurrences de gens del conjunt específic abans (per als ES positius) o després (per als ES negatius) del cim en la puntuació corrent d'enriquement. Aquest valor indica el percentatge dels gens que contribueixen a la puntuació d'enriquement.
  - List. El percentatge dels gens en el llistat ordenat de tots els gens abans o després del pic en la puntuació corrent d'enriquement. Aquest valor ens indica on exactament el pic es produeix.
  - Signal. La fortalesa del senyal d'enriquement que combina els dos valors anteriors.
- rank. La posició del pic en la llista ordenada dels gens. Els conjunts dels gens més interessants assoleixen el seu màxim o bé al principi o al final de la llista ordenada. Vol dir que tenen aquest valor o bé molt baix o bé molt alt.

### 2.2.2 KEGG

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

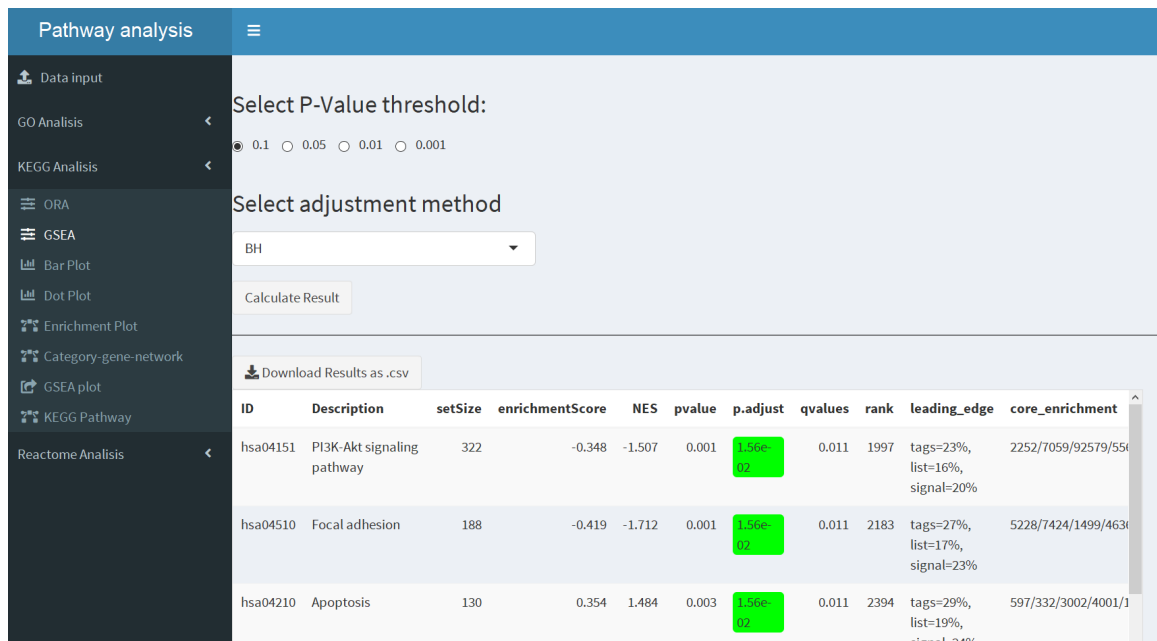


Figure 10: El resultat de l'anàlisi GSEA. KEGG.

### 2.2.3 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()`

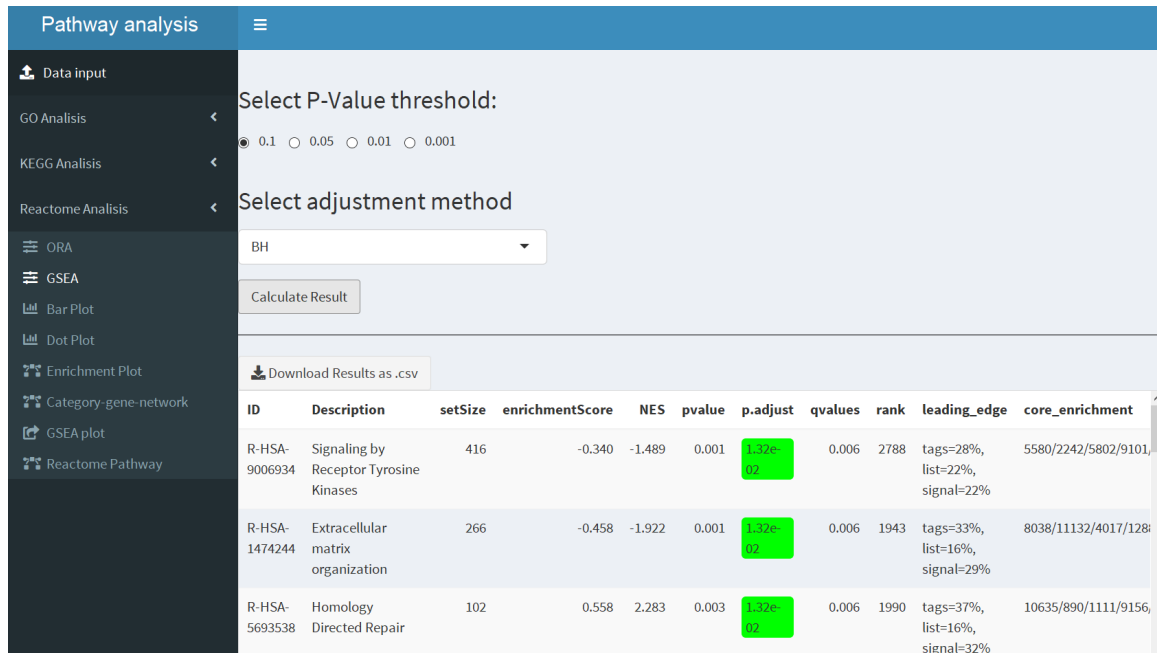


Figure 11: El resultat d'anàlisi GSEA. Reactome.

## 2.3 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.

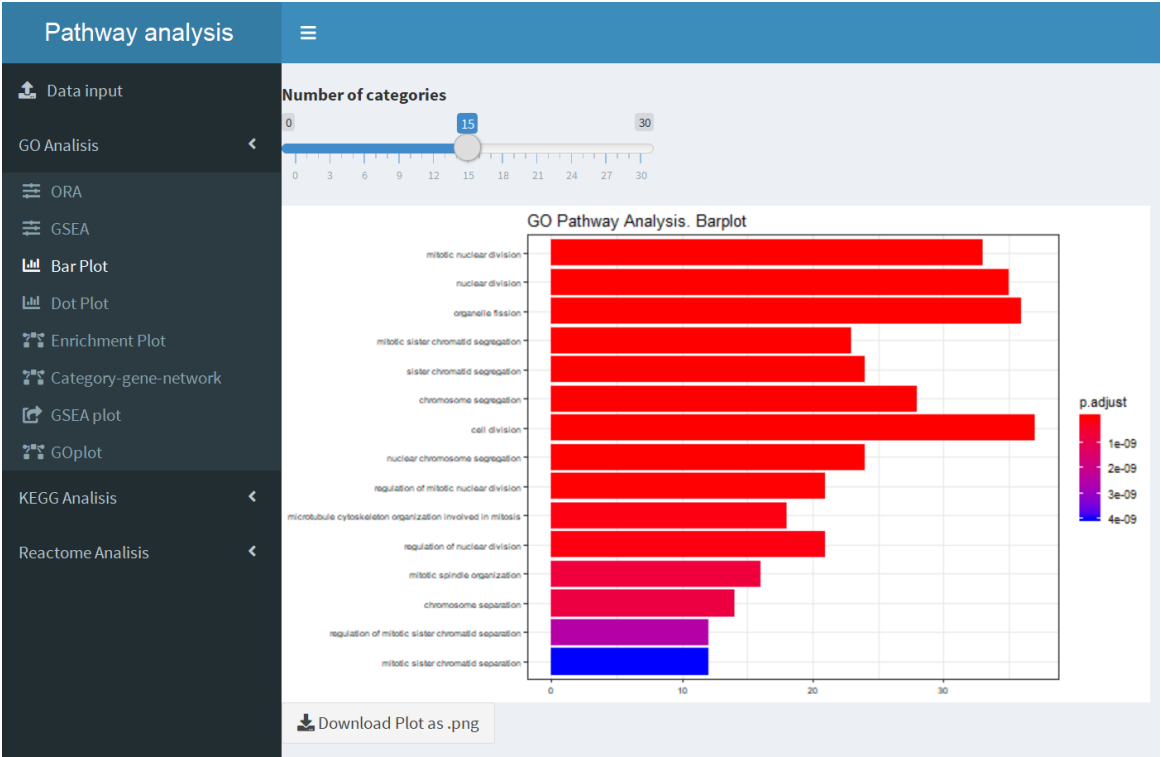


Figure 12: Bar-Plot. GO.

## 2.4 Dot-Plots

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

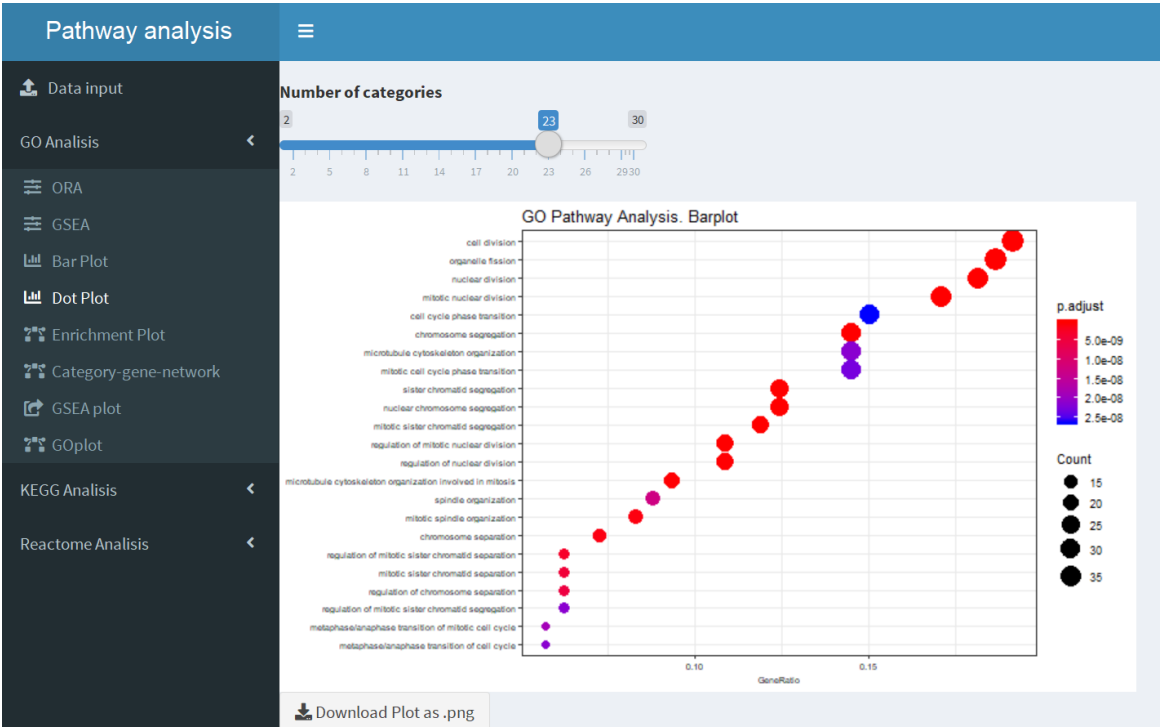


Figure 13: Bar-Plot. GO.

## 2.5 Enrichment Plots

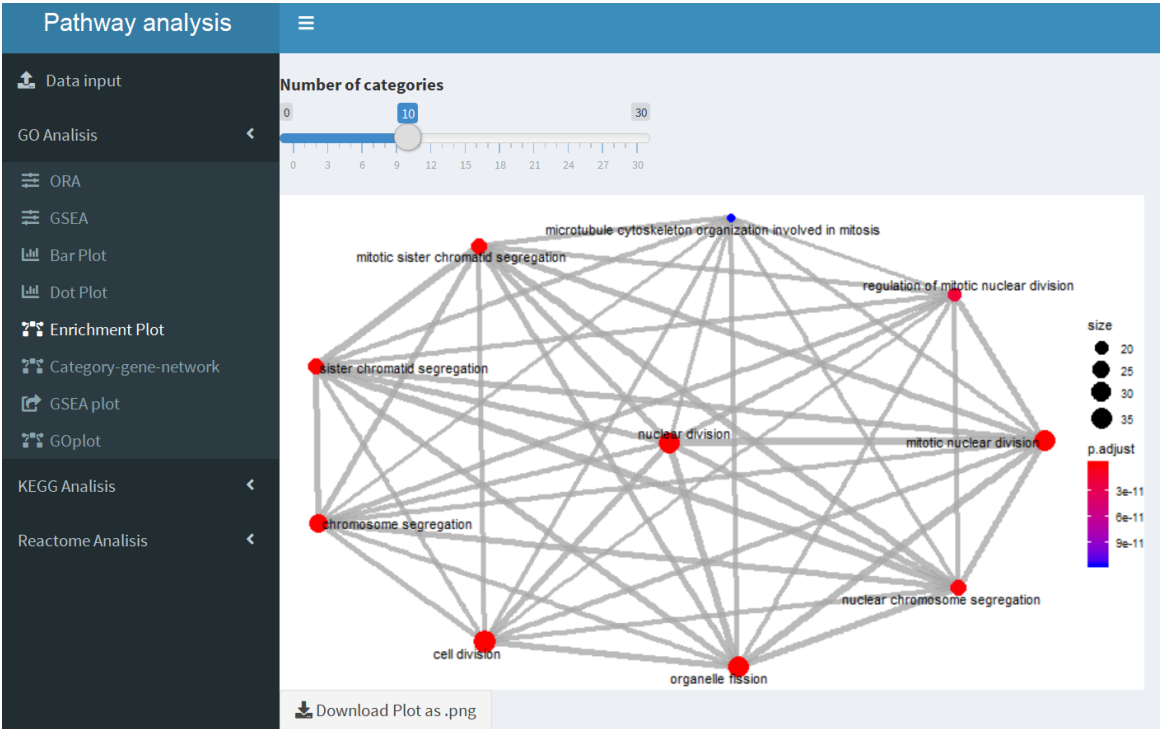


Figure 14: Bar-Plot. GO.

## 2.6 Category-Gene-Network Plot

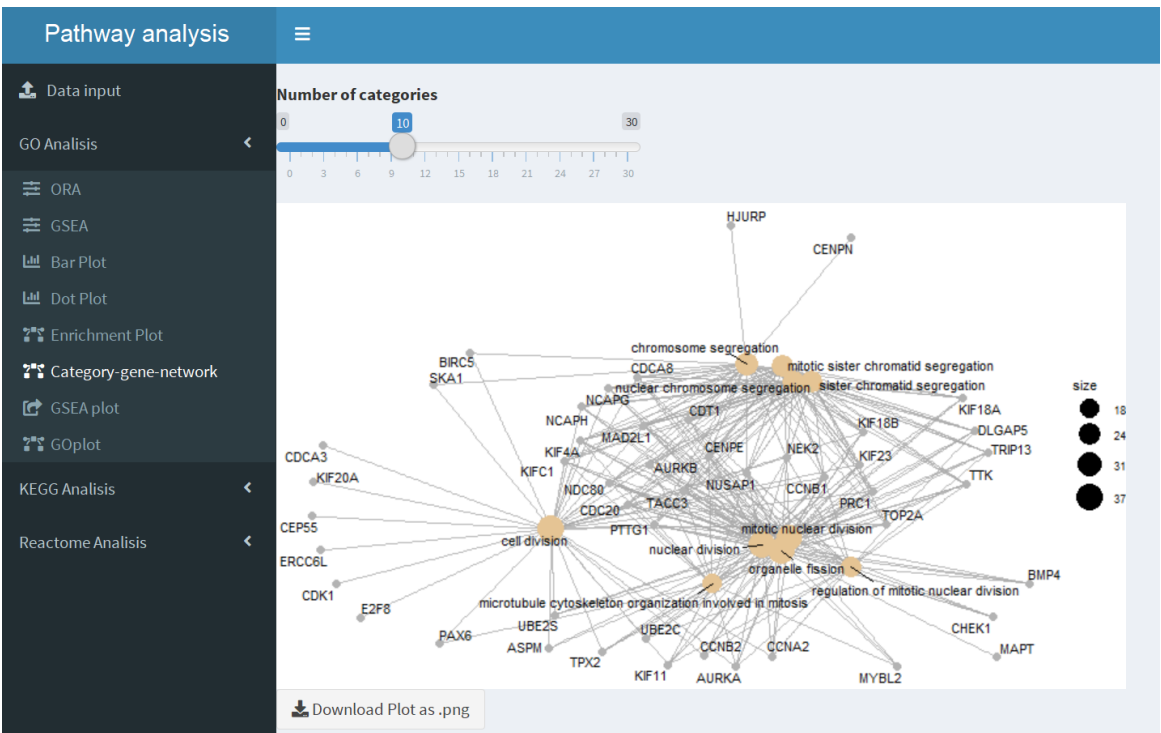


Figure 15: Category-Gene-Network Plot. GO.

## 2.7 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*

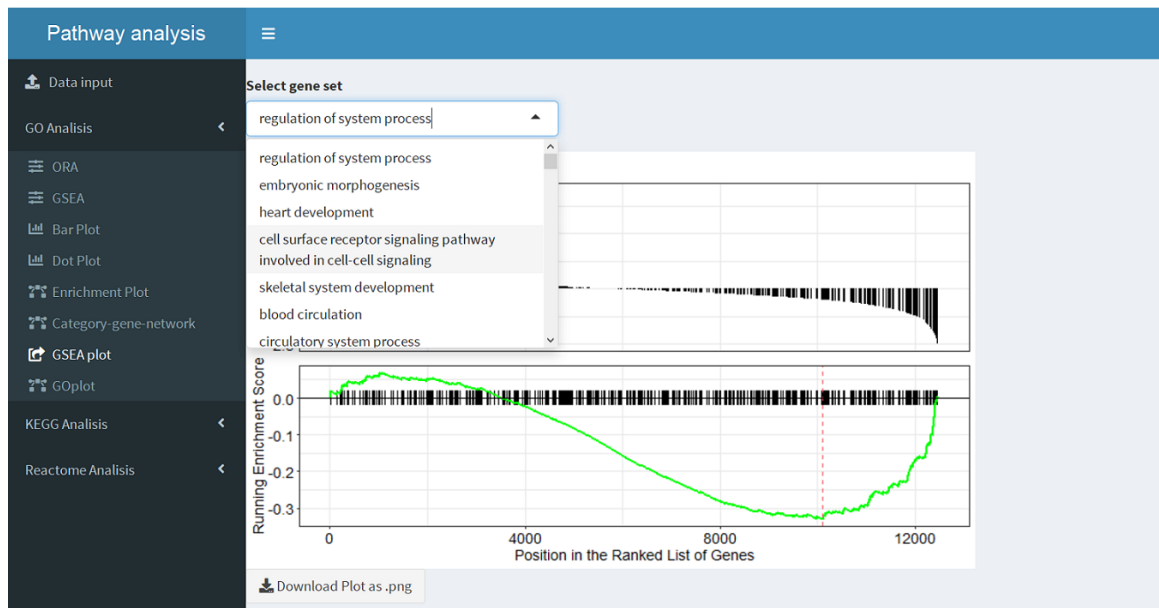


Figure 16: GSEA Plot. GO.



### 3.2 KEGG Pathway

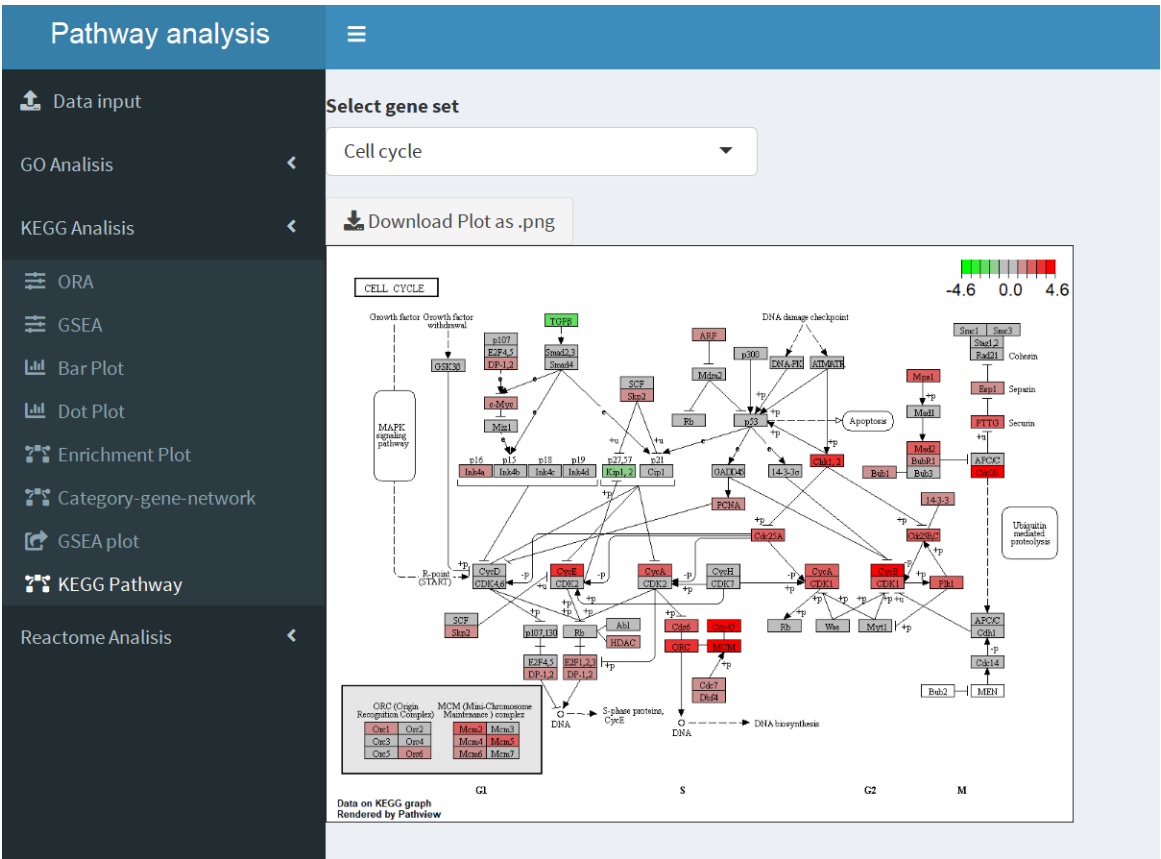


Figure 18: KEGG pathway

### 3.3 Reactome Pathway

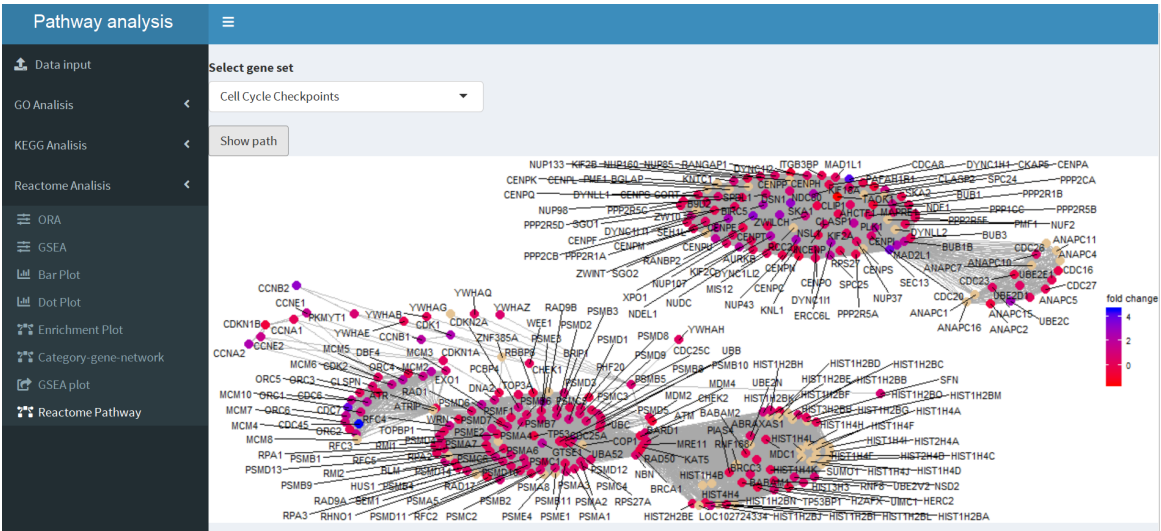


Figure 19: Reactome pathway

## 4 Validació dels resultats

Per validar l'aplicació he intentat trobar estudis que investiguin les espècies diferents i utilitzin com el tipus d'experiment o bé *Microarrays* o bé *NGS (New Generation Sequencing)*. En la base de dades GEO (Gene Expression Omnibus) he elegit els estudis següents:

Estudi	GEO ID	Espècie	Tipo d'experiment
[Li et al., 2017]	GSE100924	Mus musculus	Microarrays

L'aplicació necessita un arxiu amb les IDs de Entrez i els LogChanges. Malauradament GEO no disposa d'aquestes dades. Només té les dades en format .cel. Per tant, primer s'hauria fer els passos d'anàlisi de dades d'expressió per derivar les dades que m'interessen. Aquests passos inclouen: normalització d'una banda i calculació del model per derivar els LogChanges d'altra banda.

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 USB perquè en treballava dels molts ordinadors diferents: de casa, de feina, en 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.

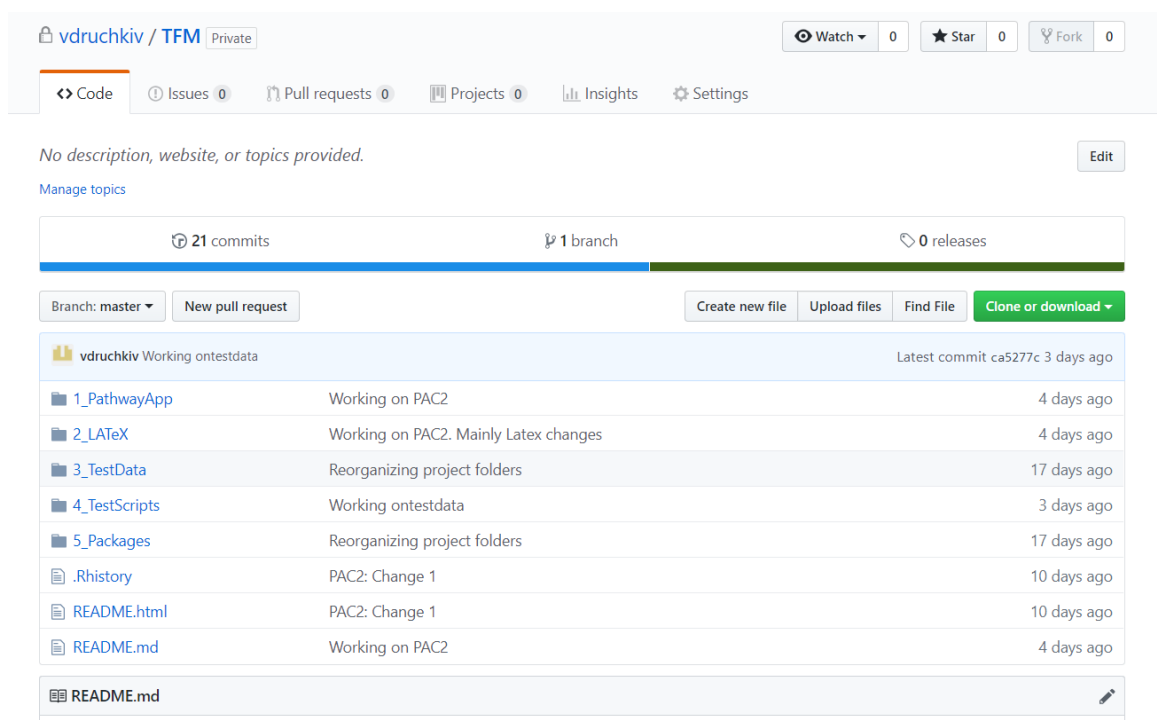


Figure 20: Github repositori del TFM

## References

[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.