PAC2 Desenvolupament el treball - Fase 1

Vasyl Druchkiv

Estudiant de Màster de Bioestadística i Bioinformàtica

18 de Març 2019

Índice

1 Identificació del treball i data de l'informe

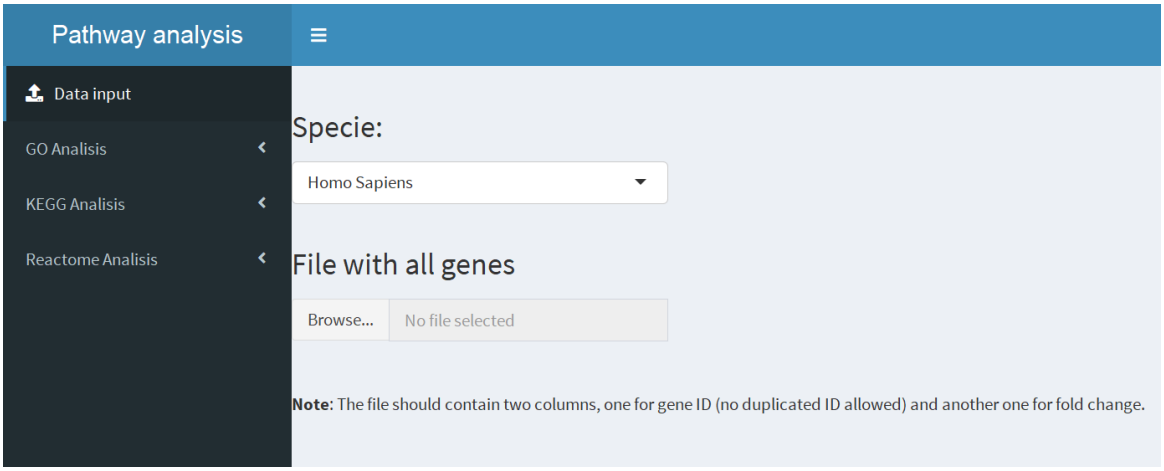
2 Descripció de l'avanç del projecte

A la data d'avui he desenvolpupat l'aplicació d'anàlisis de les rutes. L'aplicació es completament funcional localment i ofereix l'anàlisi a partir de les bases de dedes GO, KEGG i Reactome. Com estava previst, l'usuari indica l'especie, puja l'arxiu amb els gens i els LogRatios provenents d'estudi de microarrays o NGS.

L'aplicació està dividira doncs en 4 parts substancials:

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

Figure 1: Pàgina d'entrada



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 Analisis). Recordem que l'ORA consisteix en selecionar els gens diferencialment expressats i basant-se en GO, KEGG o Reactome comprobar si una de les agrupacións de gens suggerides per aquestes bases de dades està sobre o sotraexpressada en els gens selecionats. Per dur a terme l'ORA l'usuari té opció de definir un *cut-off* de Log-Ratio per formal el conjunt dels gens que s'hi utilitzara (*gene set*). ORA és una bona eina per veure els efectes grans però els effectes petits li escapen. Els efectes petits derivats dels gens individuals poden acommular-se en un efect conjunt substancial el qual ORA no serà capaç de detectar. És aquí on GSEA mostra la seva utilitat.

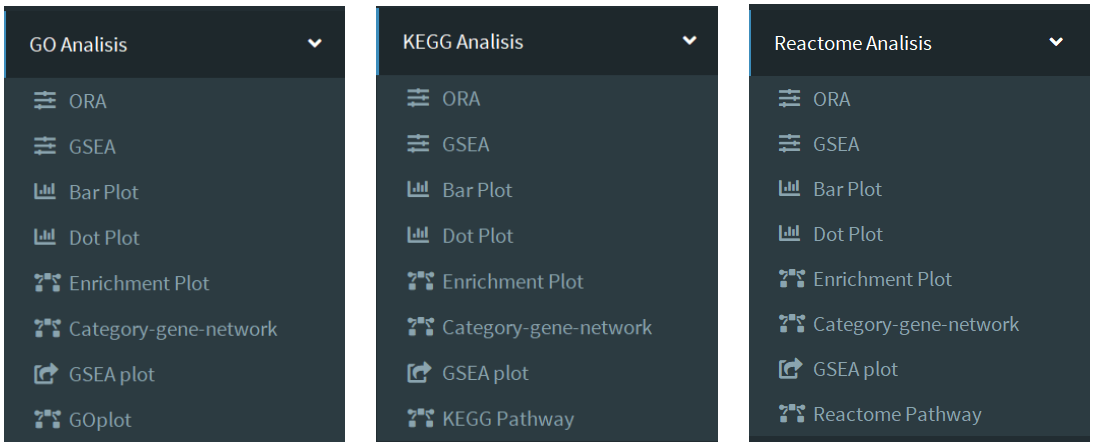
Els apartats d'anàlisi (GO, KEGG i Reactome) ofereixen tan representacions comunes com representacions específiques.

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 red dels gens en categories (Category-gene-network);
- El gràfic lde GSEA.

Els anàlisis específics són:

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



(a) GO (b) KEGG (c) Reactome

Figure 2: Els elements de les seccions d'anàlisi

3 L'anàlisi comú de GO, KEGG i Reactome

3.1 ORA

3.1.1 GO

Per realitzar l'anàlisi ORA per a termes GO s'utilitza la funciió 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'ellegir entre:

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

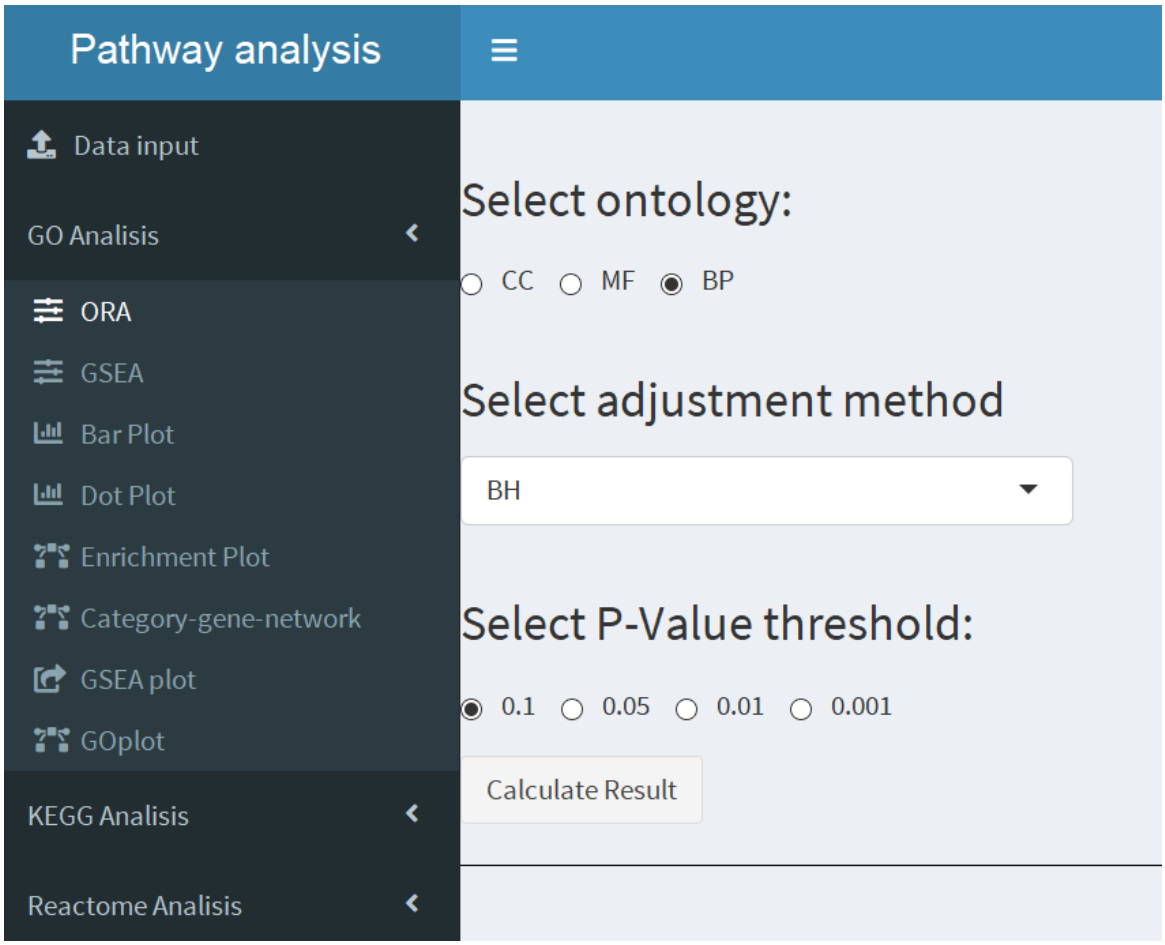


Figure 3: Especificació d'ORA dels termes GO

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

Apretant el botò apareix la taula i el botò nou mitjançant el qual l'usuari pot descargar els resultats en format .csv. He formateat la taula amb els paquets knitr, kableExtra, formattable i dplyr. Amb els dos últims he afegit les barres de color per el nombre dels gens diferencialment expressats del terme específic de GO i el gradient de color del verd fins vermell pels valors de més petits fins els més grans.

Els camps més interesants de la taula són:

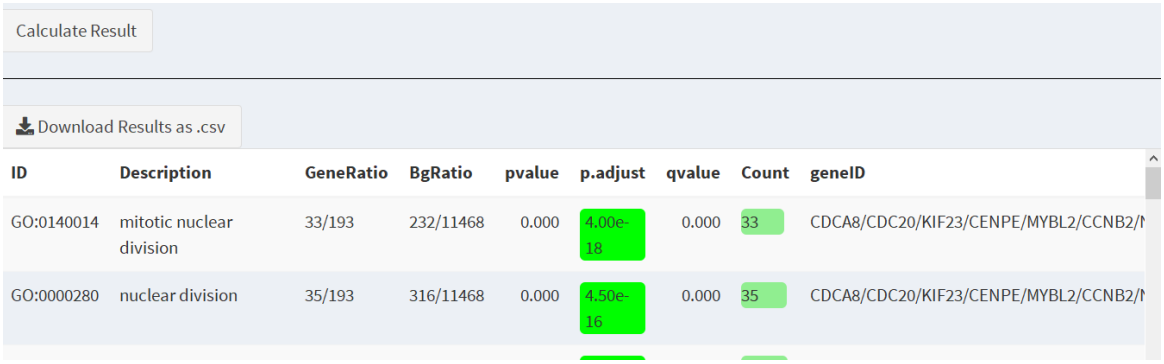


Figure 4: El resultat d'anàlisi ORA. GO.

- Description. El nom del terme GO;
- GeneRatio. El quotient: $\frac{\text{Nombre dels gens diferencialment expressats}}{\text{Nombre total dels gens en la mostra}}$;
- BgRatio. El quotient: $\frac{\text{Nombre dels gens de la ruta}}{\text{Nombre total dels gens en la base de dades GO}}$;
- 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)
```

Com en el cas de l'anàlisi dels termes GO també aquí l'usuari té la llibertat d'elegir l'organisme, el *cut-off* del valor de P i el mètode d'ajustament. Perquè la funció necessita el codi kegg', 'ncbi-geneid', 'ncib-proteinid' o 'uniprot' l'usuari ha d'especificar altra vegada l'especie. La llista de les especies disponibles per a anàlisi de KEGG és molt llarga. Per aquest motiu he abilitat l'eina de cerca d'espècie per a usuari.

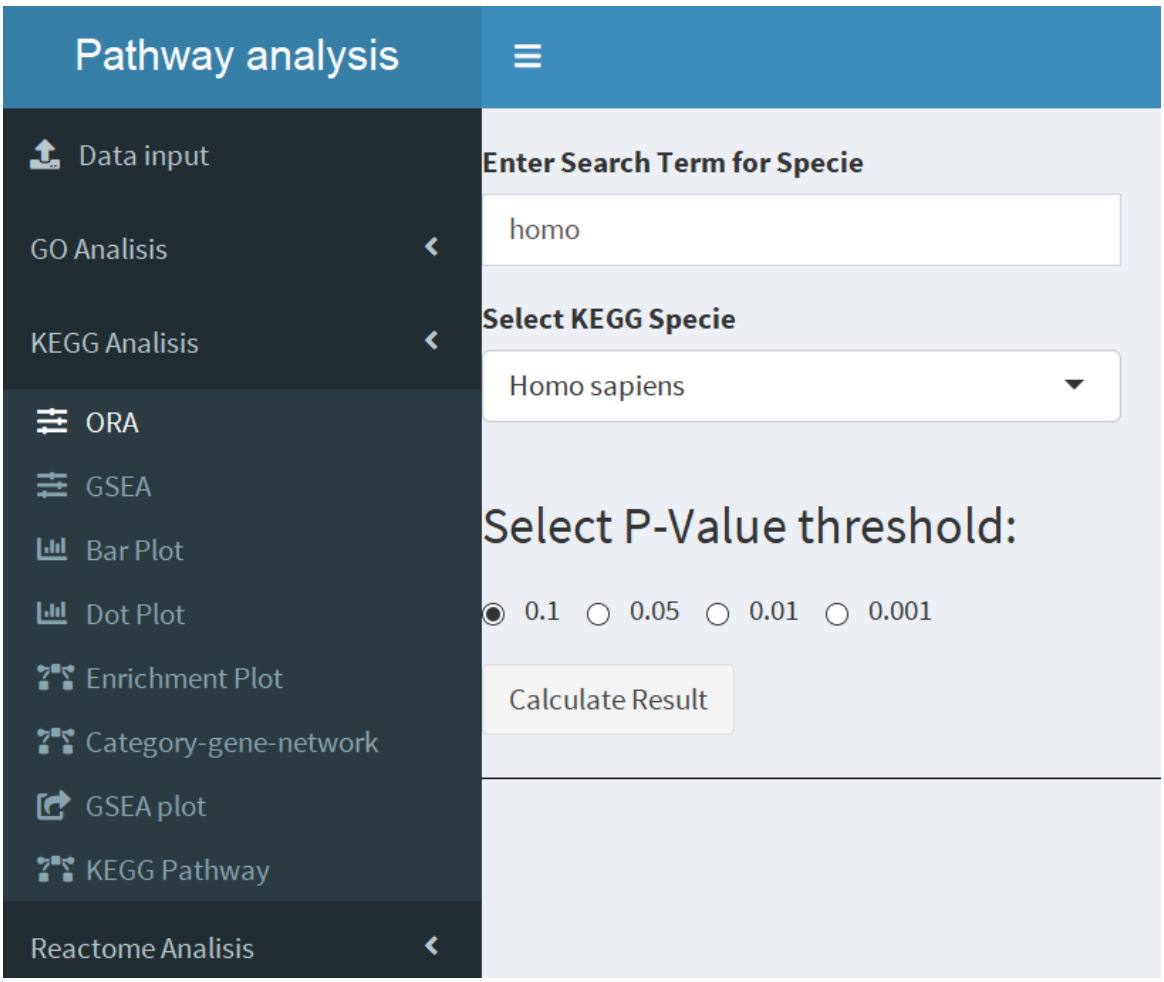


Figure 5: El resultat d'anàlisi ORA. KEGG.

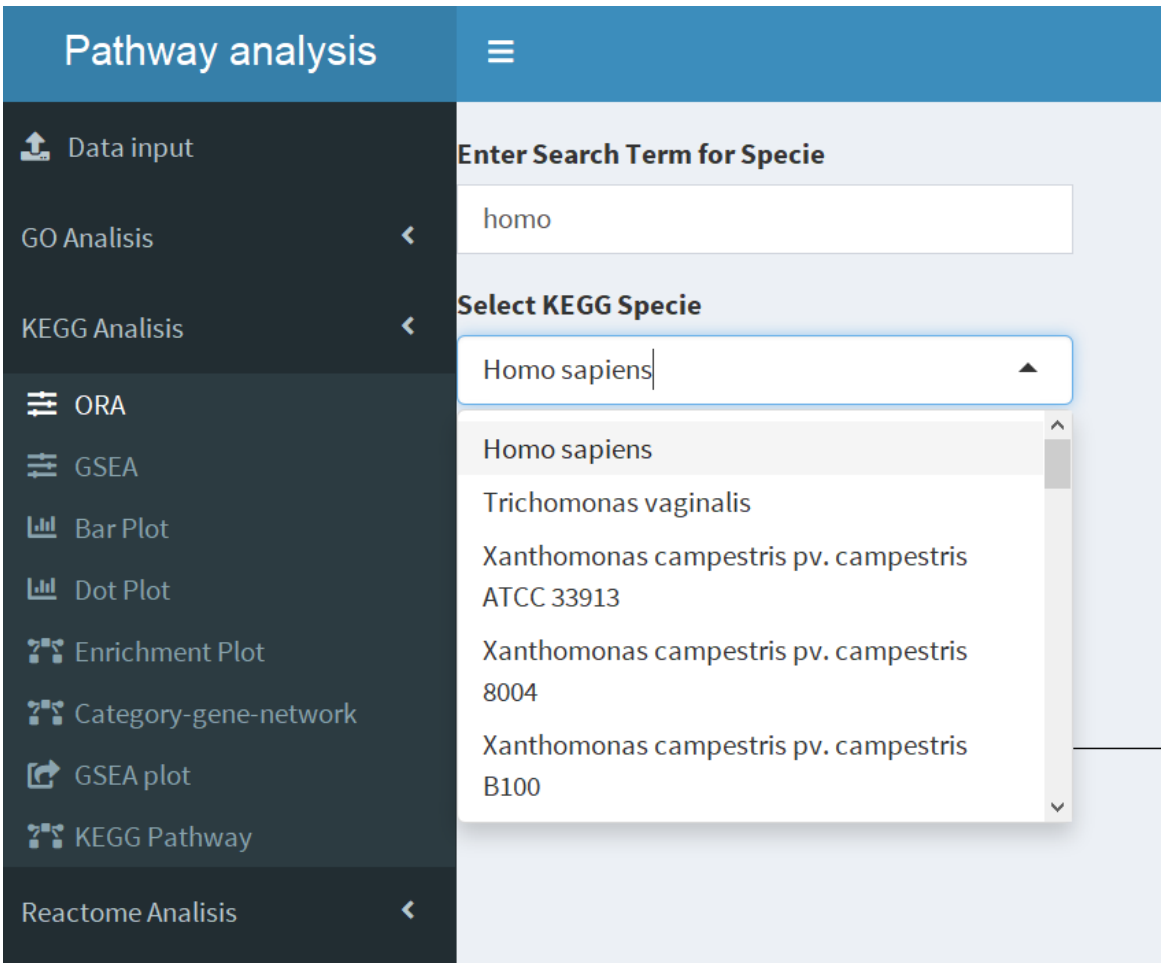
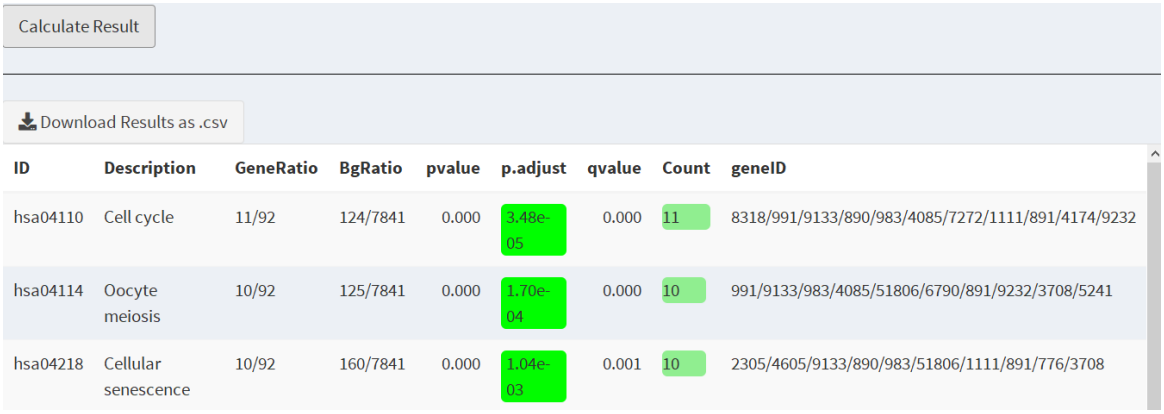


Figure 6: L'eina de cerca d'espècie. KEGG.

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



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 d'anàlisi ORA. KEGG.

3.1.3 Reactome

Al 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)
```

Aquí l'usuari ha de seleccionar l'altra vegada l'organisme. Les opcions disponibles són: "human", "rat", "mouse", "celegans", "yeast", "zebrafish", "fly".

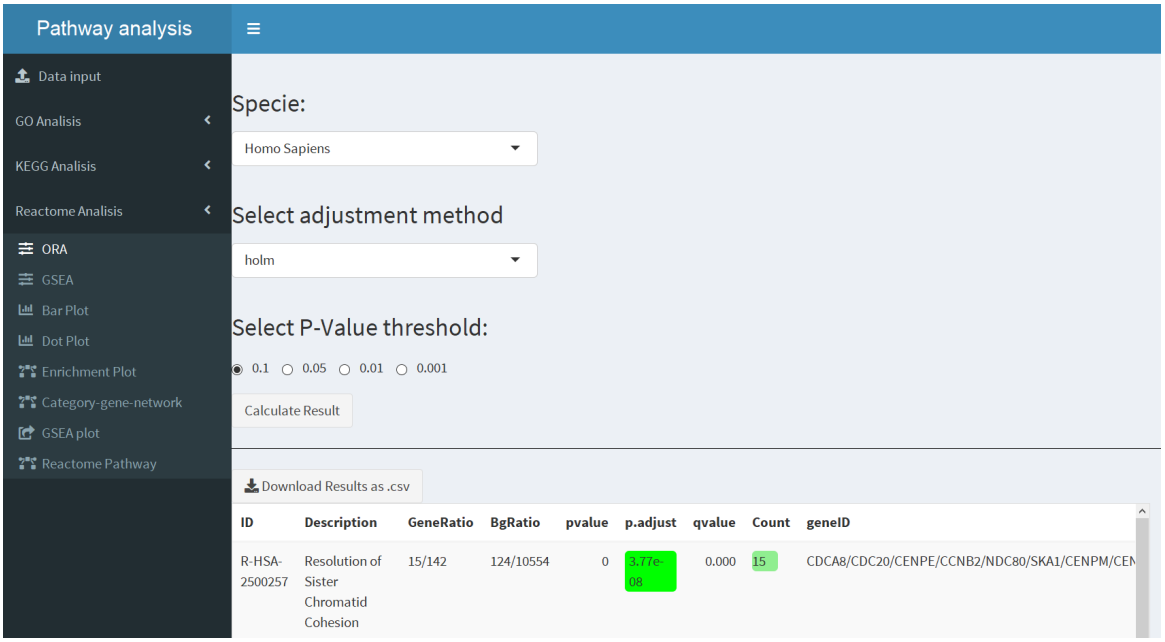


Figure 8: El resultat d'anàlisi ORA. Reactome.

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.

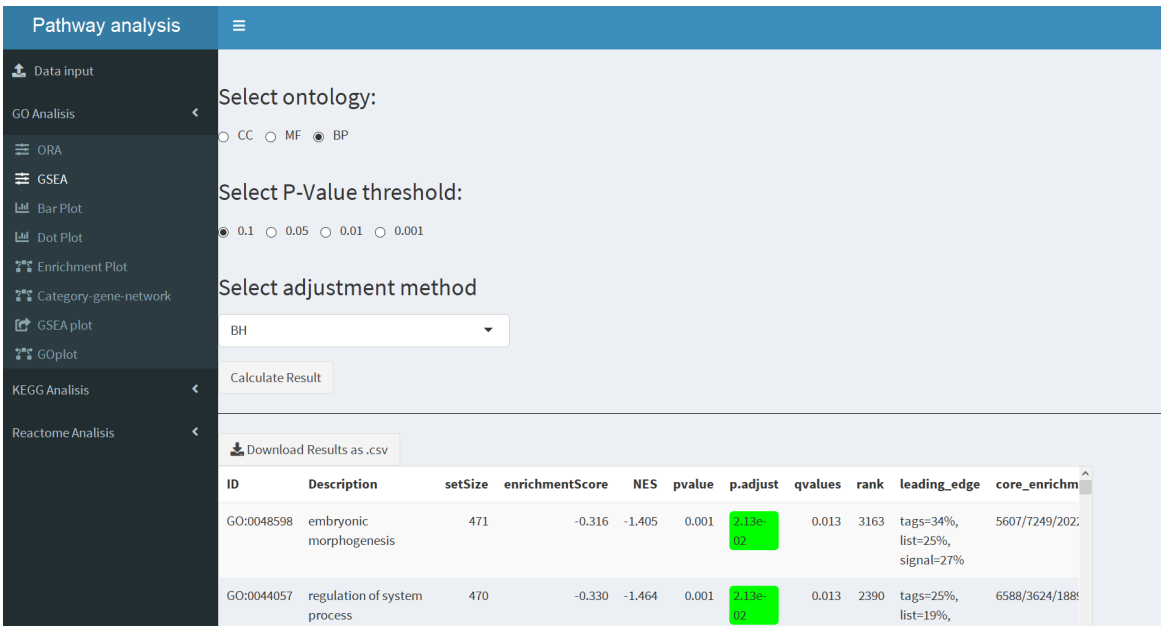


Figure 9: El resultat d'anàlisi GSEA. GO.

3.2.2 KEGG

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

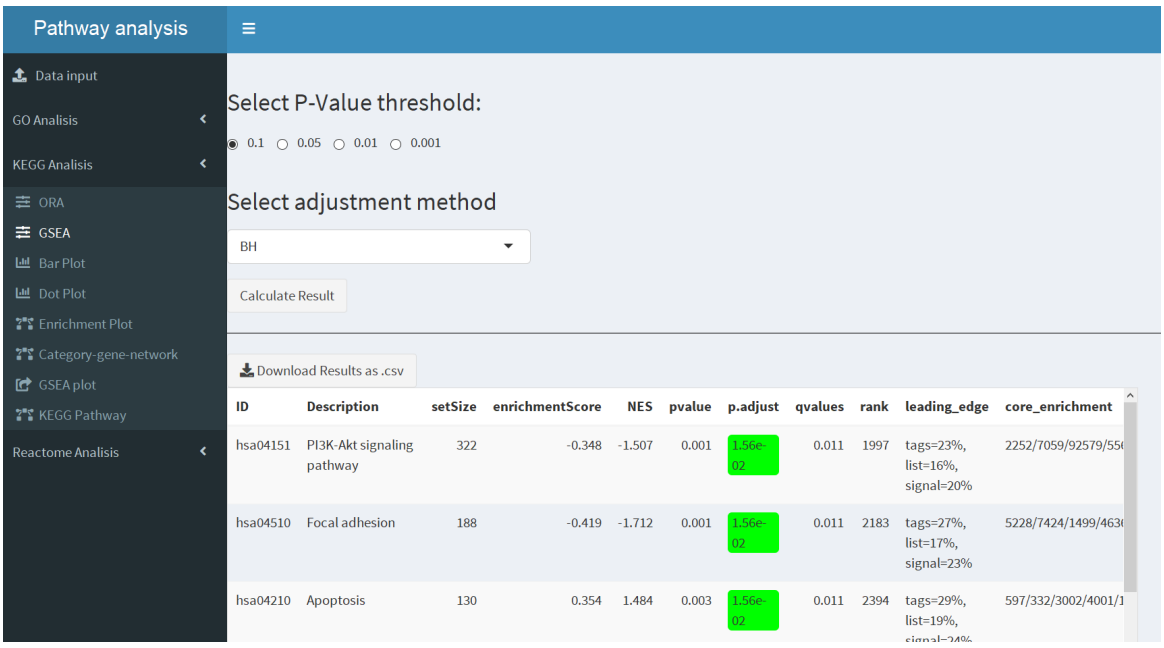
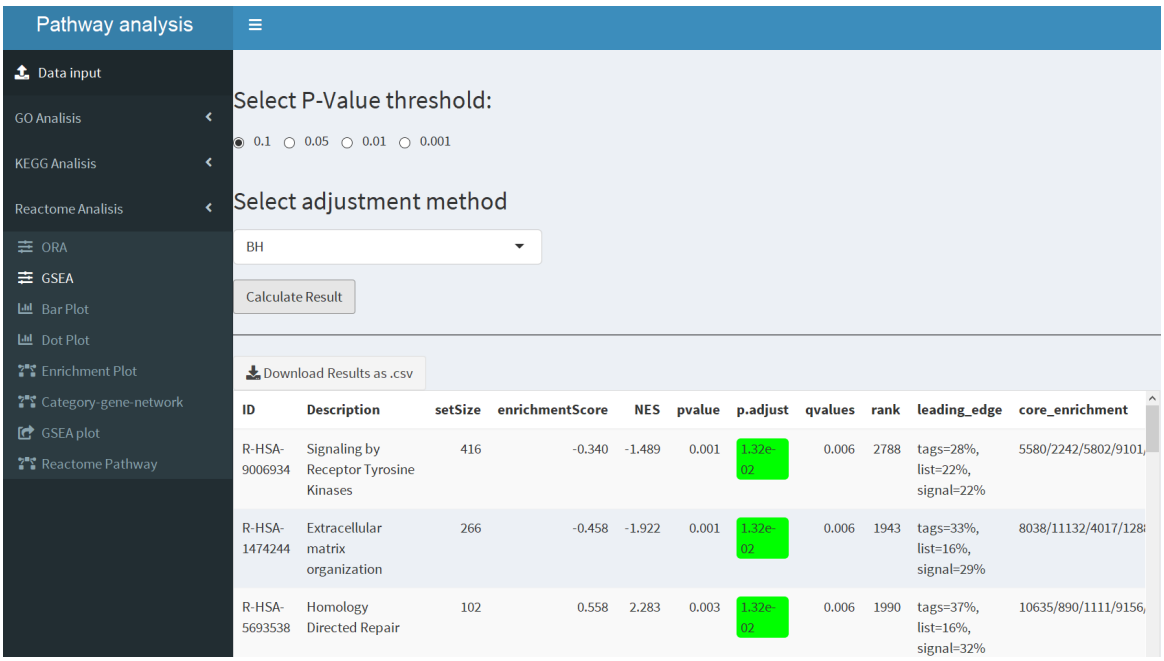


Figure 10: El resultat d'anàlisi GSEA. KEGG.

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()



ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrichment
R-HSA-9006934	Signaling by Receptor Tyrosine Kinases	416	-0.340	-1.489	0.001	1.32e-02	0.006	2788	tags=28%, list=22%, signal=22%	5580/2242/5802/9101/
R-HSA-1474244	Extracellular matrix organization	266	-0.458	-1.922	0.001	1.32e-02	0.006	1943	tags=33%, list=16%, signal=29%	8038/11132/4017/128
R-HSA-5693538	Homology Directed Repair	102	0.558	2.283	0.003	1.32e-02	0.006	1990	tags=37%, list=16%, signal=32%	10635/890/1111/9156/

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

3.3 Bar-Plots

Els resultats de enrichGO, enrichKEGG i enrichPathway es pot 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 descargar el gràfic en format .png.

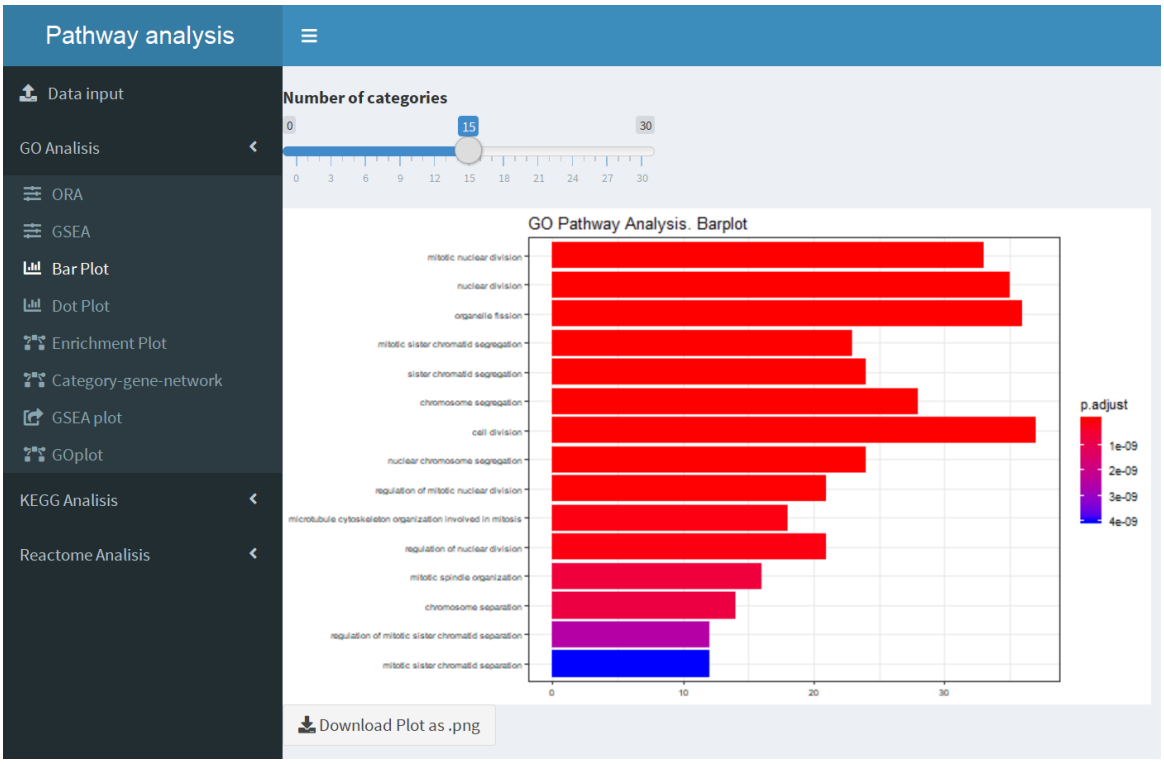


Figure 12: Bar-Plot. GO.

3.4 Dot-Plots

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

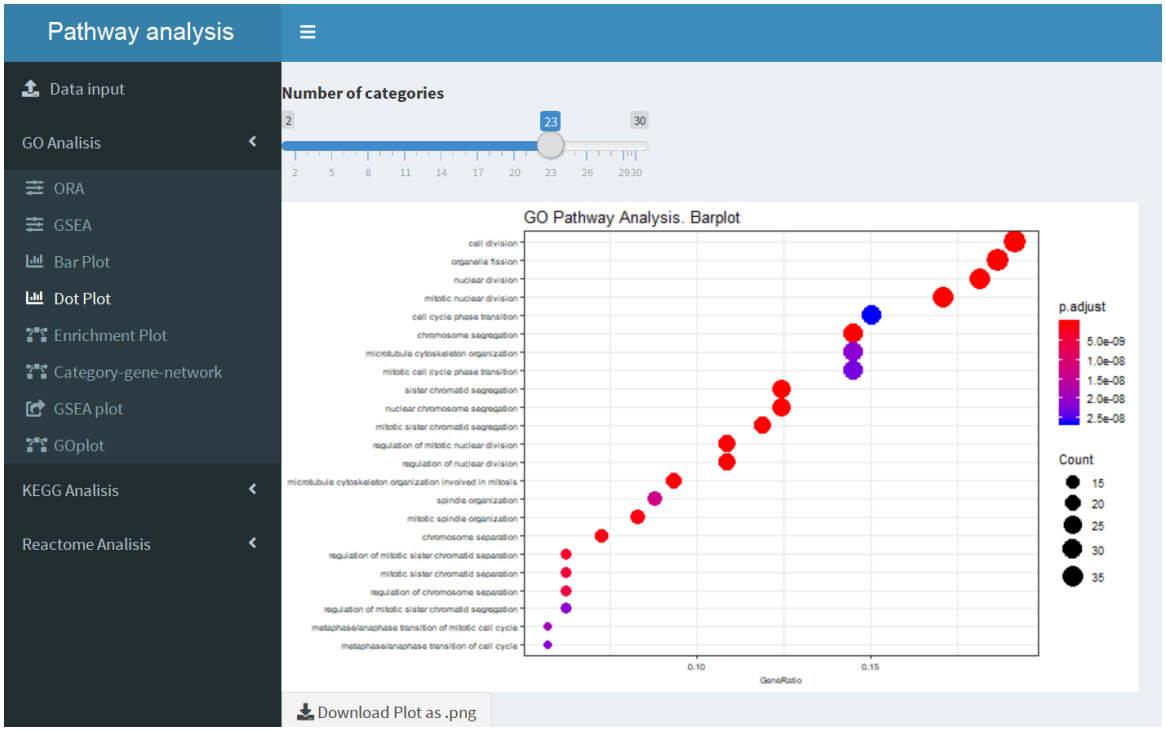


Figure 13: Bar-Plot. GO.

3.5 Enrichment Plots

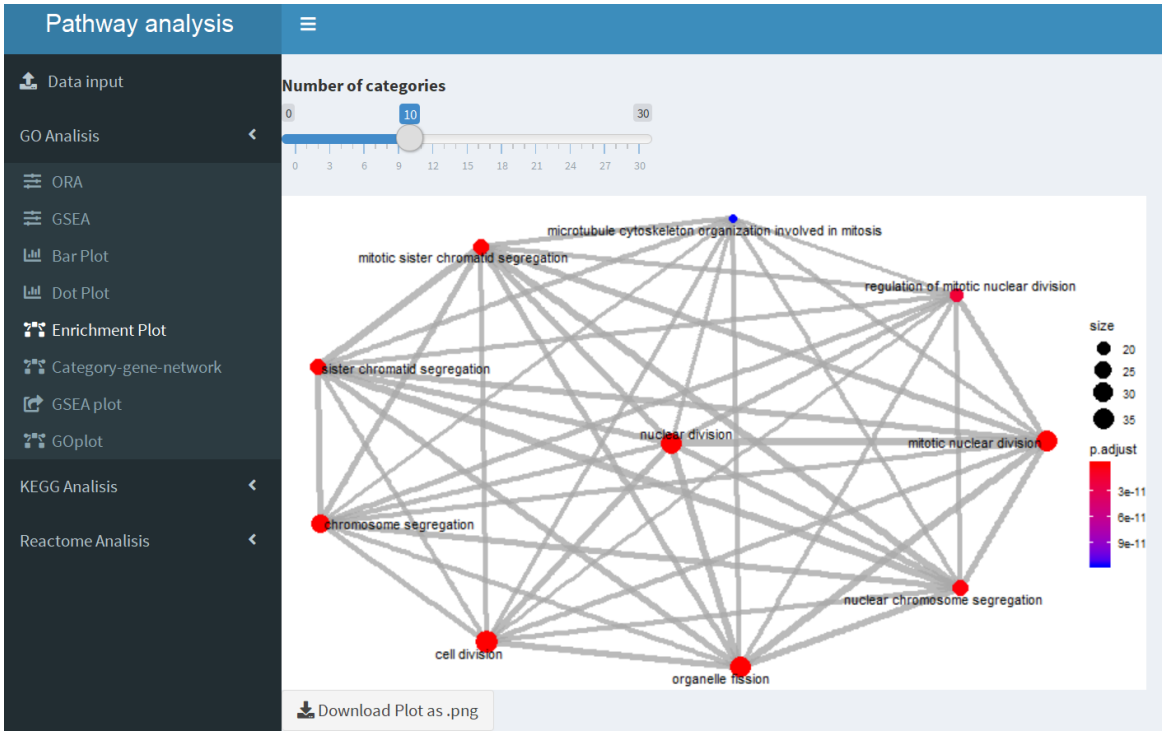


Figure 14: Bar-Plot. GO.

3.6 Category-Gene-Network Plot

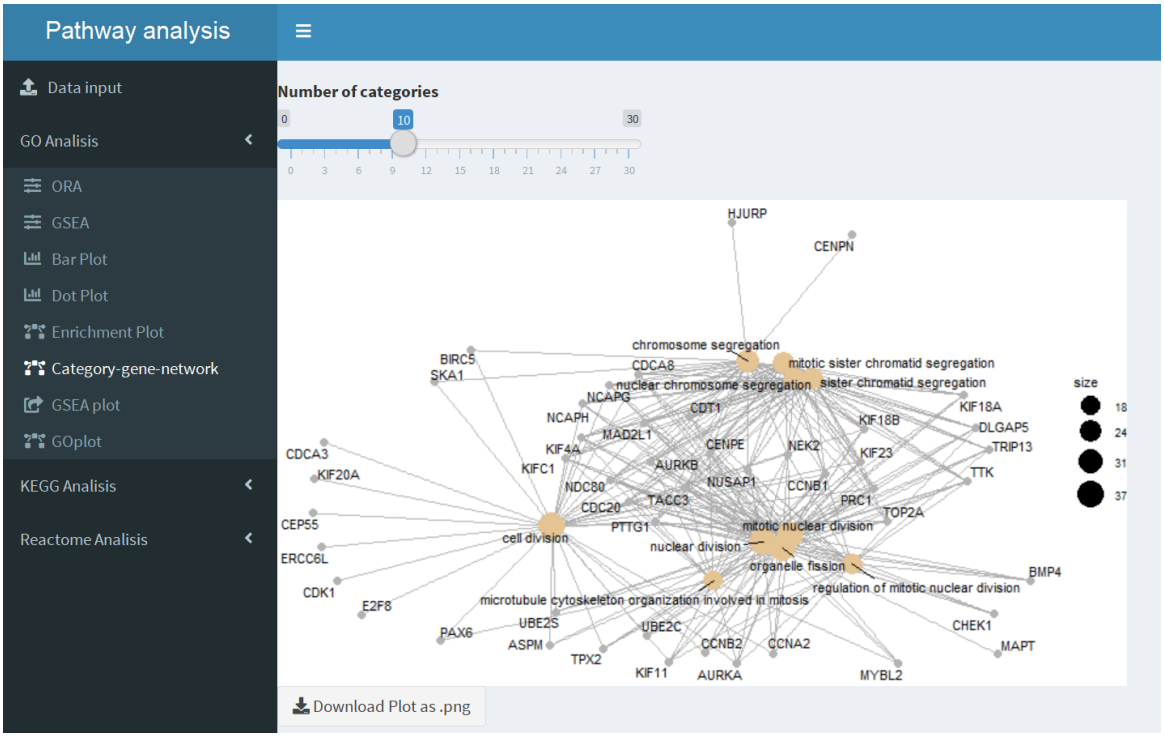


Figure 15: Category-Gene-Network Plot. GO.

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*

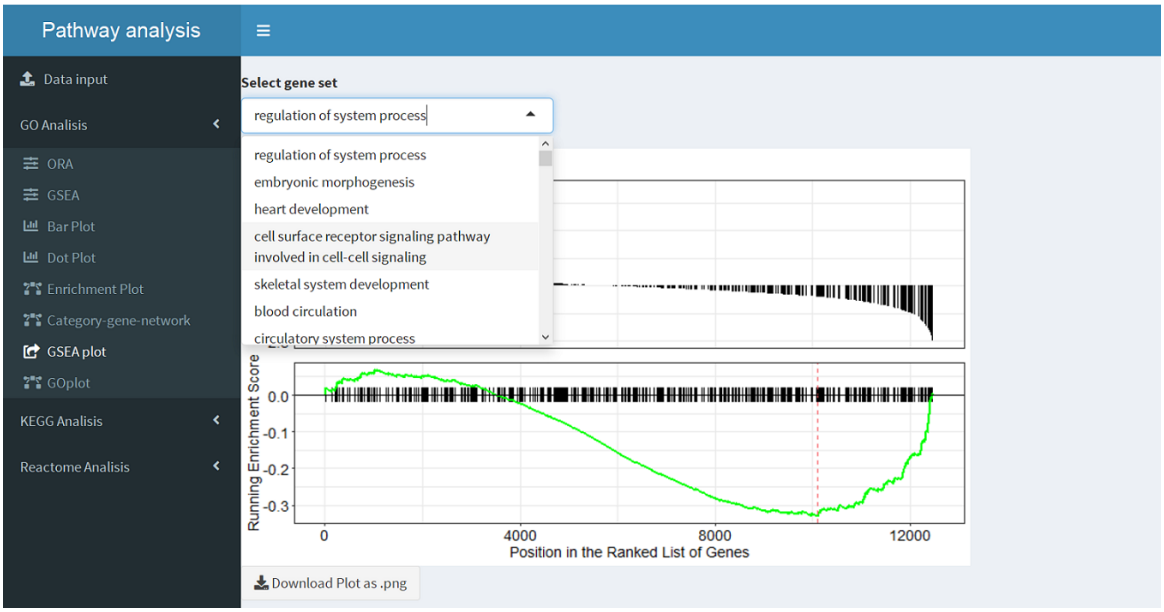


Figure 16: GSEA Plot. GO.

4 L'anàlisi específic de GO, KEGG i Reactome

4.1 GO Plot

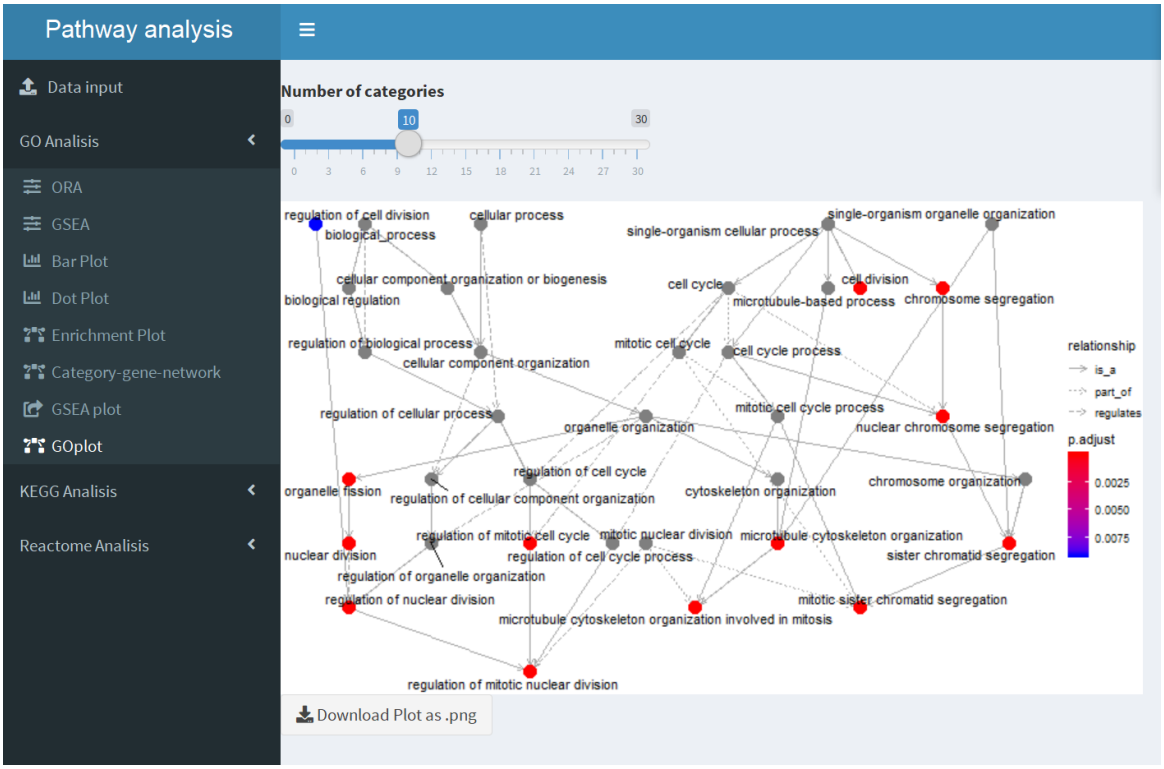


Figure 17: GO Plot

4.2 KEGG Pathway

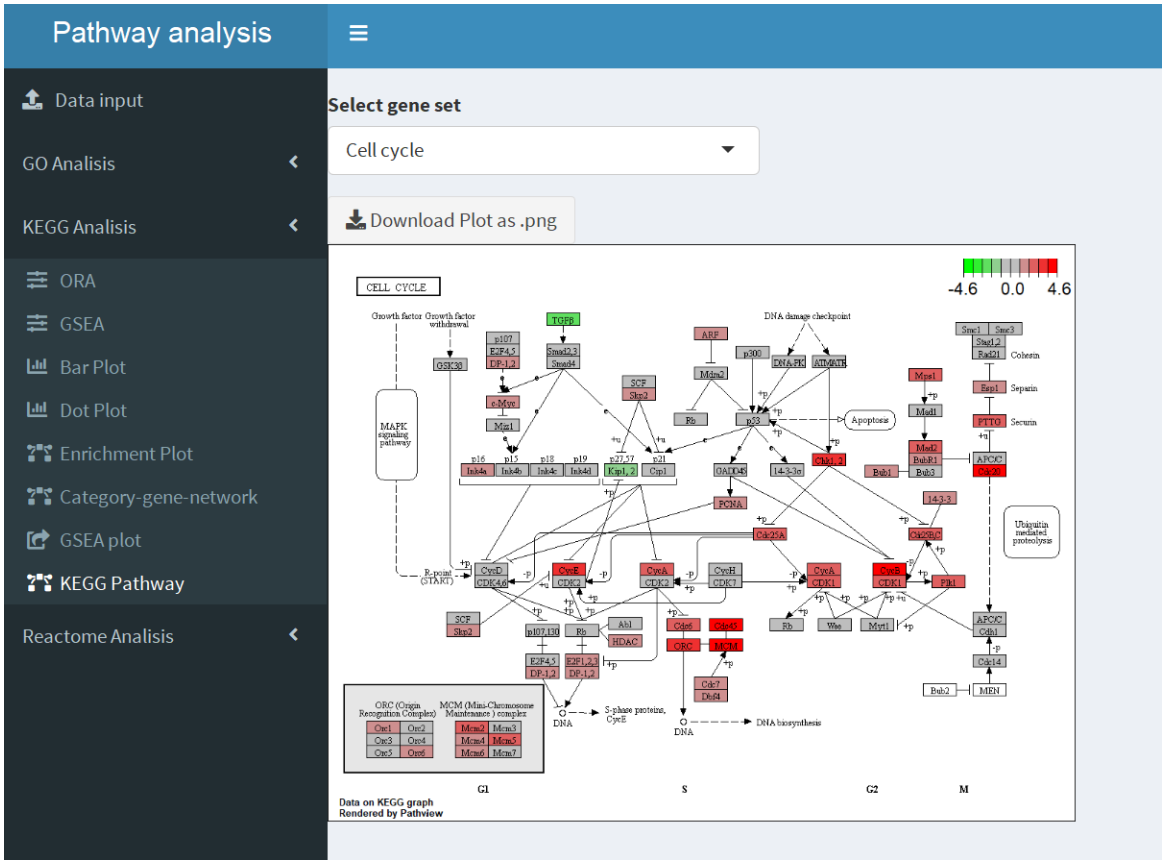


Figure 18: KEGG pathway

4.3 Reactome Pathway

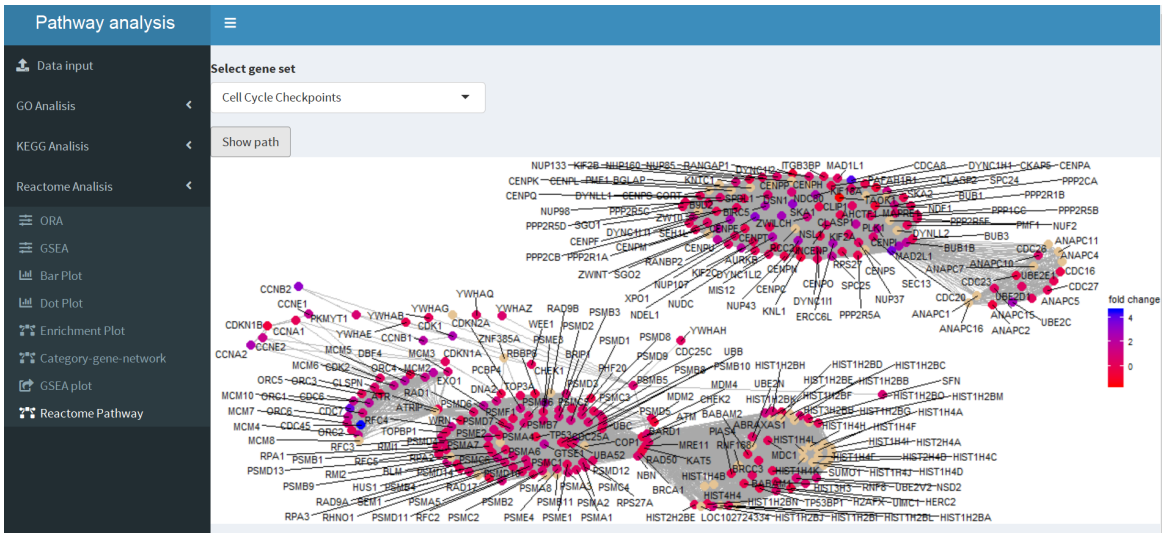


Figure 19: Reactome pathway