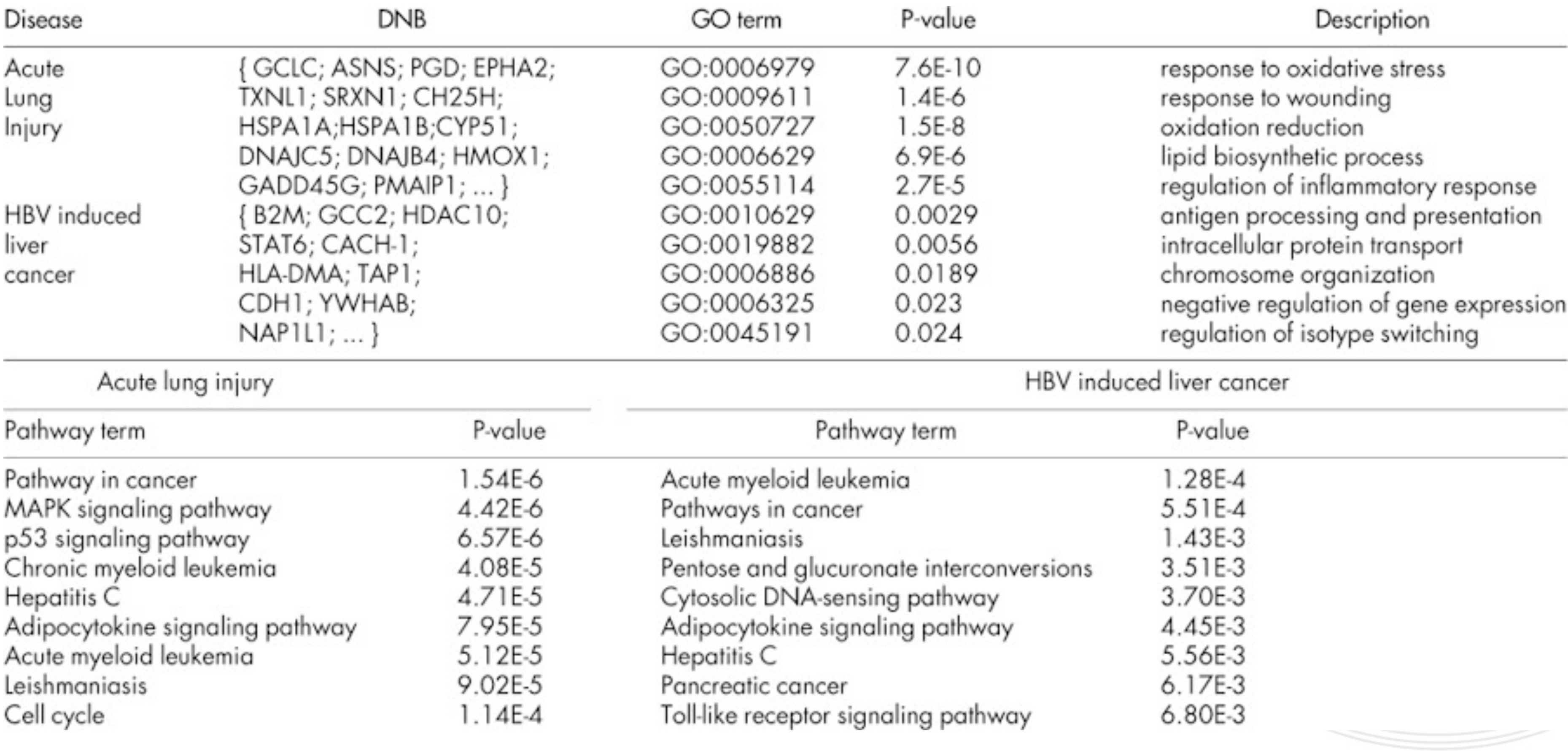
Ontology Enrichment and Analysis

Explotación semántica de datos
Máster Universitario en Bioinformática

Functional enrichment of GO biological processes and KEGG pathway enrichment analysis in the identified Dynamic Network Biomarkers (DNB) of two diseases

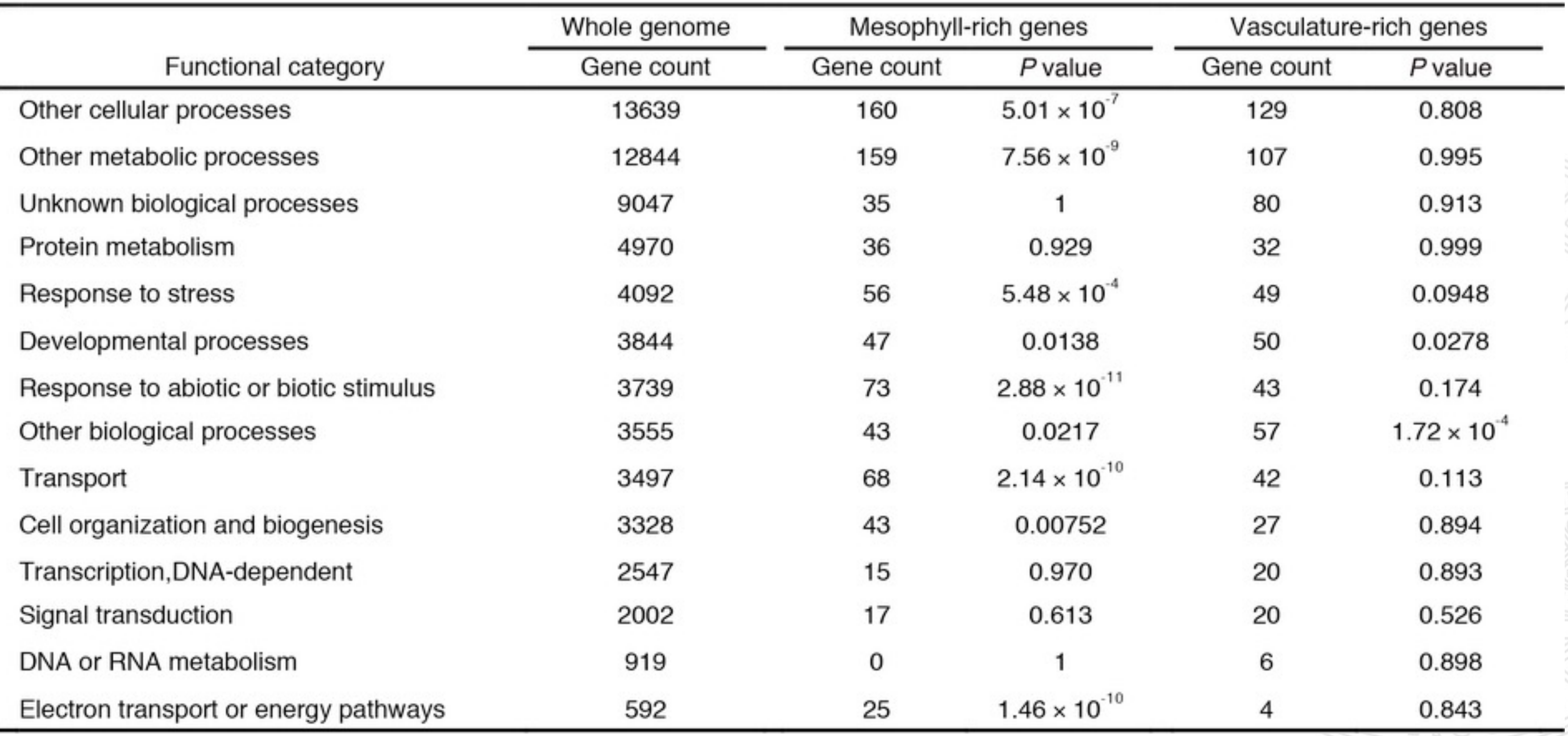


Disease	DNB	GO term	P-value	Description
Acute	{ GCLC; ASNS; PGD; EPHA2;	GO:0006979	7.6E-10	response to oxidative stress
Lung	TXNL1; SRXN1; CH25H;	GO:0009611	1.4E-6	response to wounding
Injury	HSPA1A;HSPA1B;CYP51;	GO:0050727	1.5E-8	oxidation reduction
	DNAJC5; DNAJB4; HMOX1;	GO:0006629	6.9E-6	lipid biosynthetic process
	GADD45G; PMAIP1; ... }	GO:0055114	2.7E-5	regulation of inflammatory response
HBV induced	{ B2M; GCC2; HDAC10;	GO:0010629	0.0029	antigen processing and presentation
liver	STAT6; CACH-1;	GO:0019882	0.0056	intracellular protein transport
cancer	HLA-DMA; TAP1;	GO:0006886	0.0189	chromosome organization
	CDH1; YWHAB;	GO:0006325	0.023	negative regulation of gene expression
	NAP1L1; ... }	GO:0045191	0.024	regulation of isotype switching

Acute lung injury		HBV induced liver cancer	
Pathway term	P-value	Pathway term	P-value
Pathway in cancer	1.54E-6	Acute myeloid leukemia	1.28E-4
MAPK signaling pathway	4.42E-6	Pathways in cancer	5.51E-4
p53 signaling pathway	6.57E-6	Leishmaniasis	1.43E-3
Chronic myeloid leukemia	4.08E-5	Pentose and glucuronate interconversions	3.51E-3
Hepatitis C	4.71E-5	Cytosolic DNA-sensing pathway	3.70E-3
Adipocytokine signaling pathway	7.95E-5	Adipocytokine signaling pathway	4.45E-3
Acute myeloid leukemia	5.12E-5	Hepatitis C	5.56E-3
Leishmaniasis	9.02E-5	Pancreatic cancer	6.17E-3
Cell cycle	1.14E-4	Toll-like receptor signaling pathway	6.80E-3

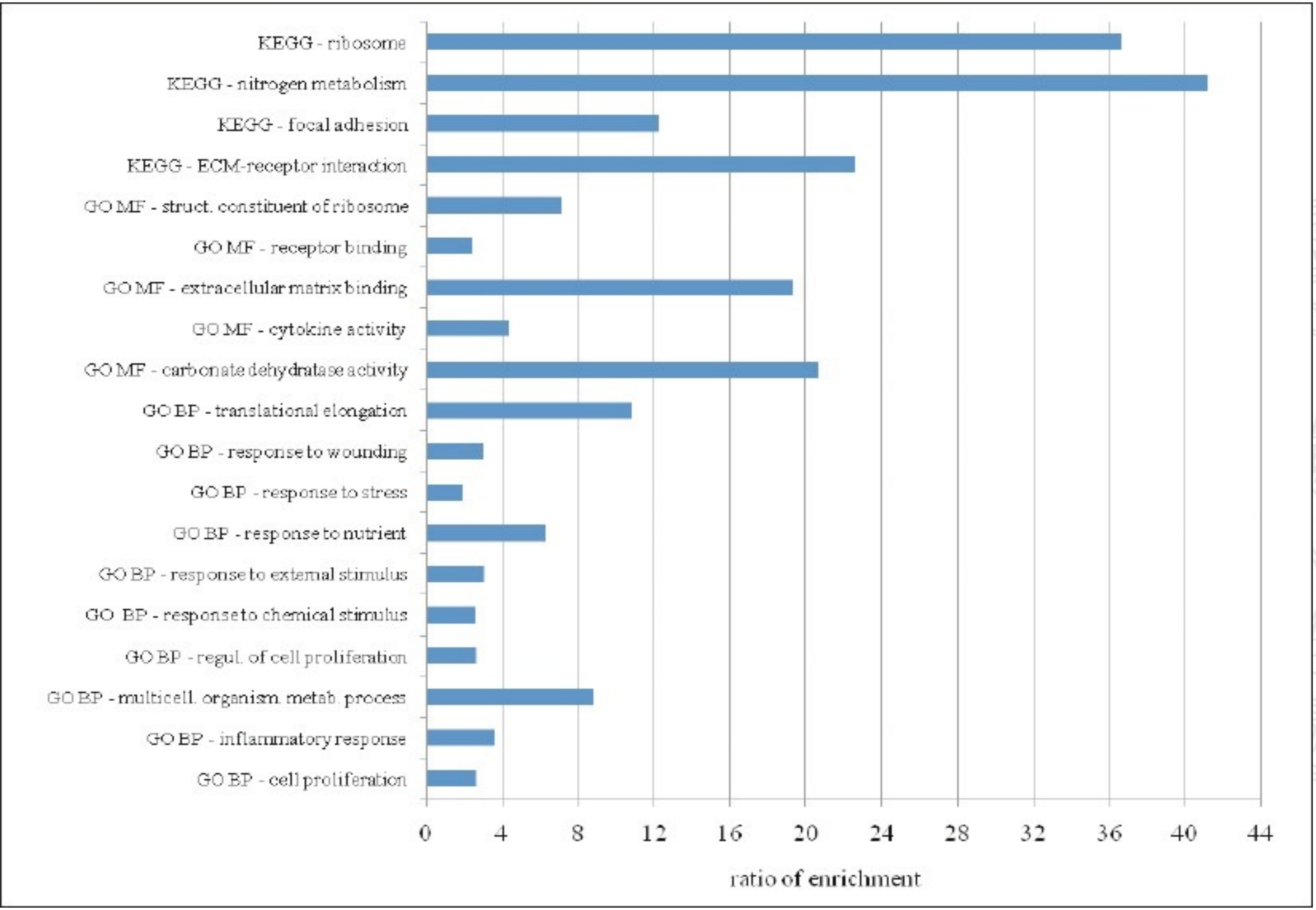
Tissue-specific clocks in Arabidopsis show asymmetric coupling Gene ontology slim term enrichment analysis

UNIVERSIDAD DE
MURCIA



Functional category	Whole genome	Mesophyll-rich genes		Vasculature-rich genes	
	Gene count	Gene count	*P* value	Gene count	*P* value
Other cellular processes	13639	160	5.01×10^{-7}	129	0.808
Other metabolic processes	12844	159	7.56×10^{-9}	107	0.995
Unknown biological processes	9047	35	1	80	0.913
Protein metabolism	4970	36	0.929	32	0.999
Response to stress	4092	56	5.48×10^{-4}	49	0.0948
Developmental processes	3844	47	0.0138	50	0.0278
Response to abiotic or biotic stimulus	3739	73	2.88×10^{-11}	43	0.174
Other biological processes	3555	43	0.0217	57	1.72×10^{-4}
Transport	3497	68	2.14×10^{-10}	42	0.113
Cell organization and biogenesis	3328	43	0.00752	27	0.894
Transcription,DNA-dependent	2547	15	0.970	20	0.893
Signal transduction	2002	17	0.613	20	0.526
DNA or RNA metabolism	919	0	1	6	0.898
Electron transport or energy pathways	592	25	1.46×10^{-10}	4	0.843

Systematic enrichment analysis of gene expression profiling studies identifies consensus pathways implicated in colorectal cancer development



Ontology Enrichment Analysis

- Given a set of genes of interest (with some particular properties), find which ontology terms are overrepresented using the annotations for the global set of genes.

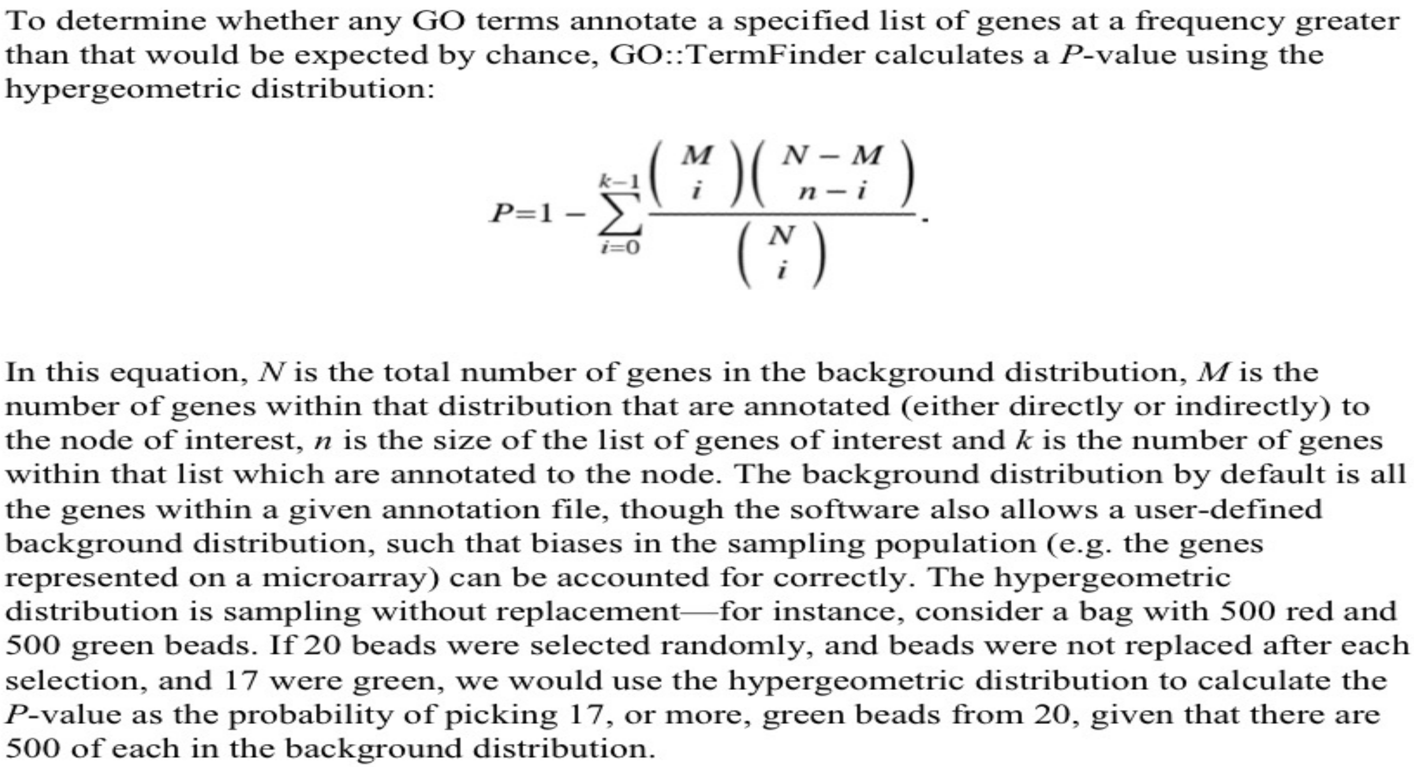
- The hierarchical structure of the ontologies permit to perform such studies at different levels

- Not only GO can be used for enrichment analysis

- KEGG: which pathways are overrepresented

Gene Enrichment Analysis

- Use of a subset of genes of interest



To determine whether any GO terms annotate a specified list of genes at a frequency greater than that would be expected by chance, GO::TermFinder calculates a *P*-value using the hypergeometric distribution:

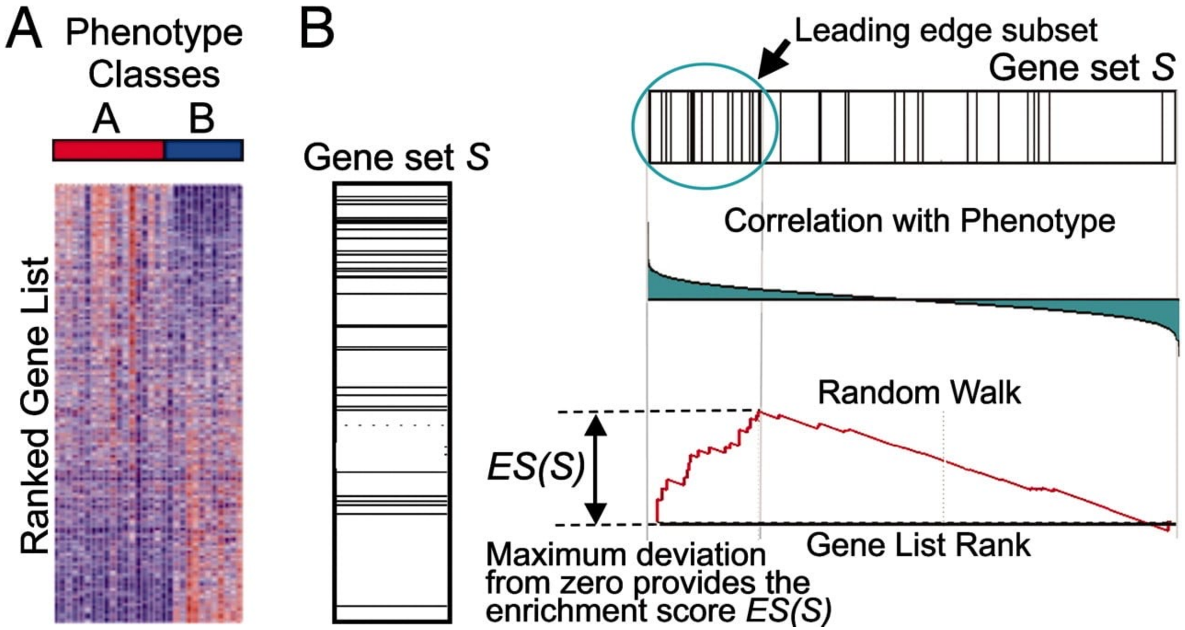
$$P=1-\sum_{i=0}^{k-1}\frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}.$$

In this equation, *N* is the total number of genes in the background distribution, *M* is the number of genes within that distribution that are annotated (either directly or indirectly) to the node of interest, *n* is the size of the list of genes of interest and *k* is the number of genes within that list which are annotated to the node. The background distribution by default is all the genes within a given annotation file, though the software also allows a user-defined background distribution, such that biases in the sampling population (e.g. the genes represented on a microarray) can be accounted for correctly. The hypergeometric distribution is sampling without replacement—for instance, consider a bag with 500 red and 500 green beads. If 20 beads were selected randomly, and beads were not replaced after each selection, and 17 were green, we would use the hypergeometric distribution to calculate the *P*-value as the probability of picking 17, or more, green beads from 20, given that there are 500 of each in the background distribution.

http://geneontology.org/docs/go-enrichment-analysis/

Gene Set Enrichment Analysis

- Use of all the genes associated with the functional concept, not only a subset of interest



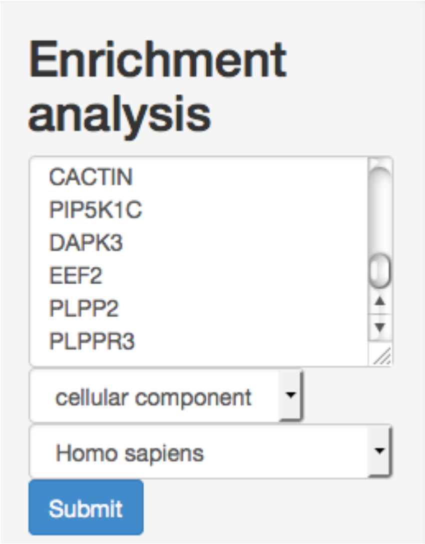
(*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the “gene tags,” i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Ontology Enrichment

- Online enrichment
  - Gene Ontology
  - Reactome

- R libraries for ontology enrichment
  - CLUSTERPROFILER
    - https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
  - GPROFILER
    - https://cran.r-project.org/web/packages/gProfileR/index.html

GO Enrichment Analysis

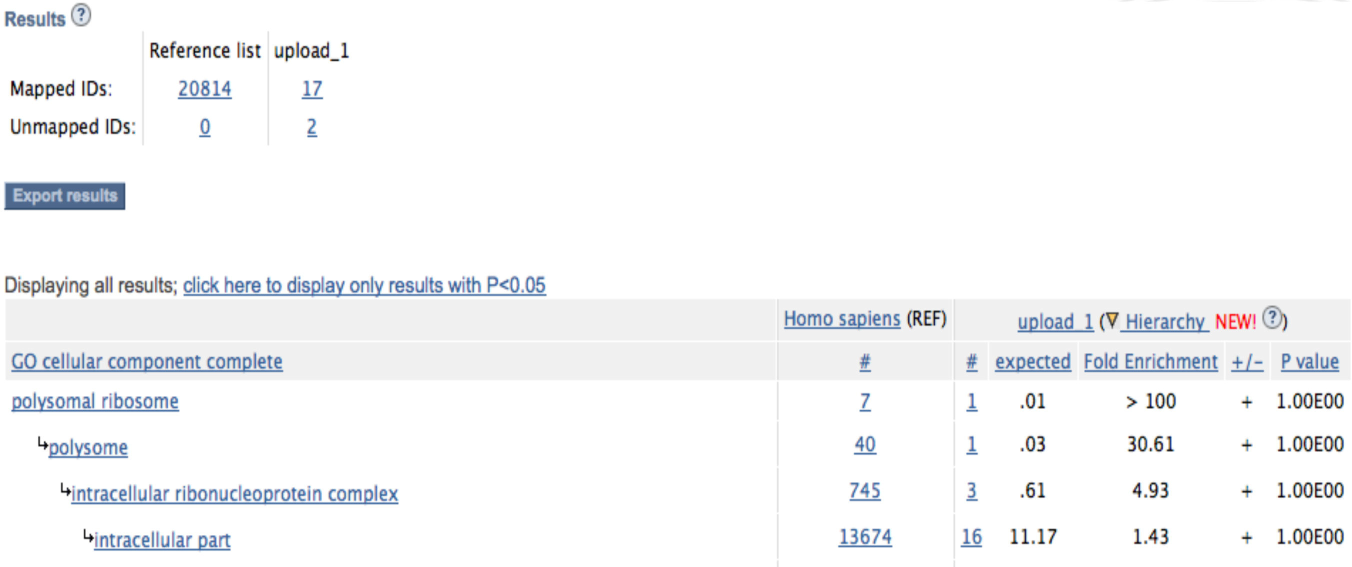


Enrichment analysis

CACTIN
PIP5K1C
DAPK3
EEF2
PLPP2
PLPPR3

cellular component

Homo sapiens



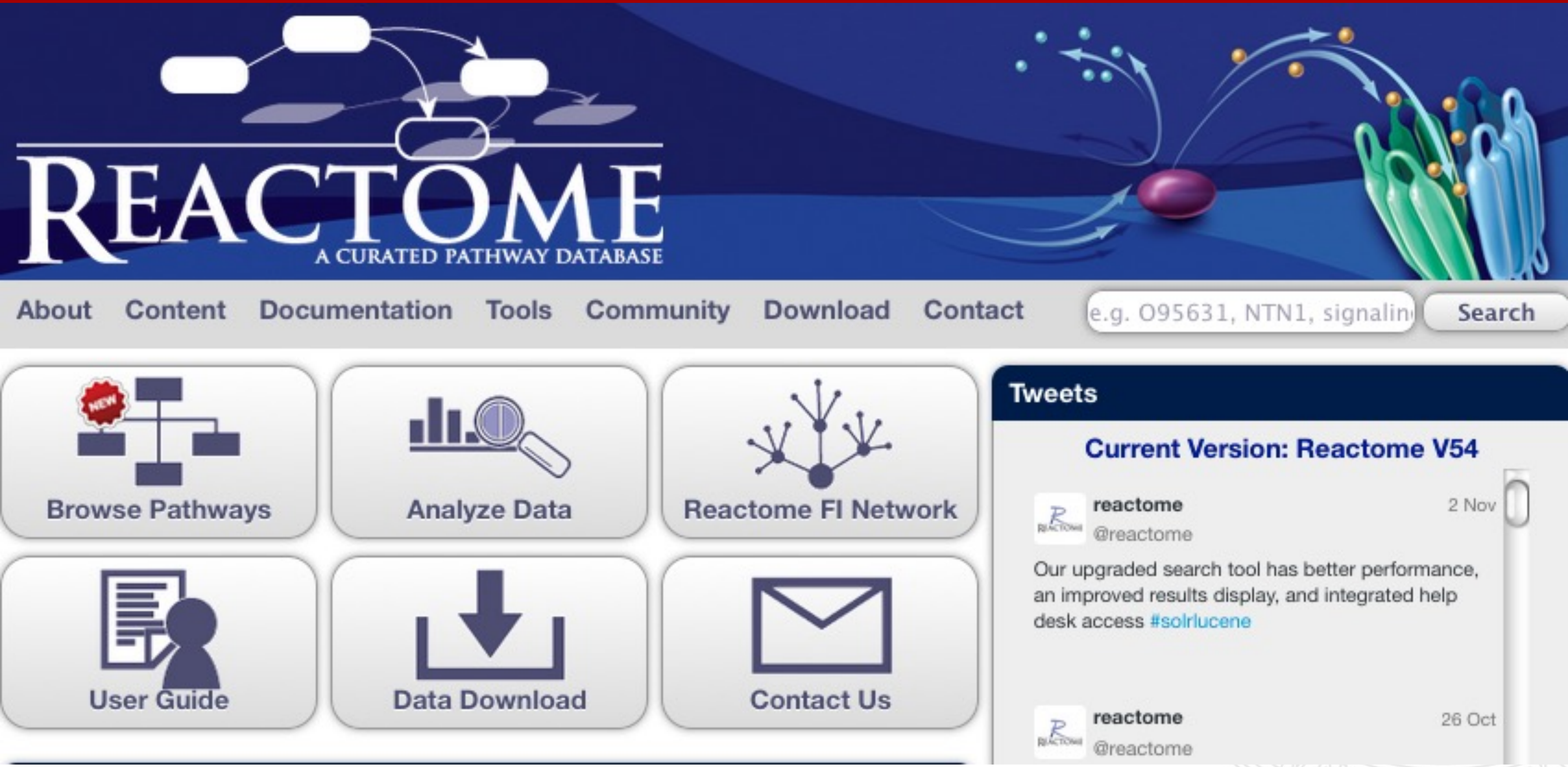
Results ?

	Reference list	upload_1
Mapped IDs:	20814	17
Unmapped IDs:	0	2

Displaying all results; click here to display only results with P<0.05

	Homo sapiens (REF)	upload_1 (∇ Hierarchy NEW! ?)				
GO cellular component complete	#	#	expected	Fold Enrichment	+/-	P value
polysomal ribosome	7	1	.01	> 100	+	1.00E00
↳polysome	40	1	.03	30.61	+	1.00E00
↳intracellular ribonucleoprotein complex	745	3	.61	4.93	+	1.00E00
↳intracellular part	13674	16	11.17	1.43	+	1.00E00

Enrichment Analysis in Reactome



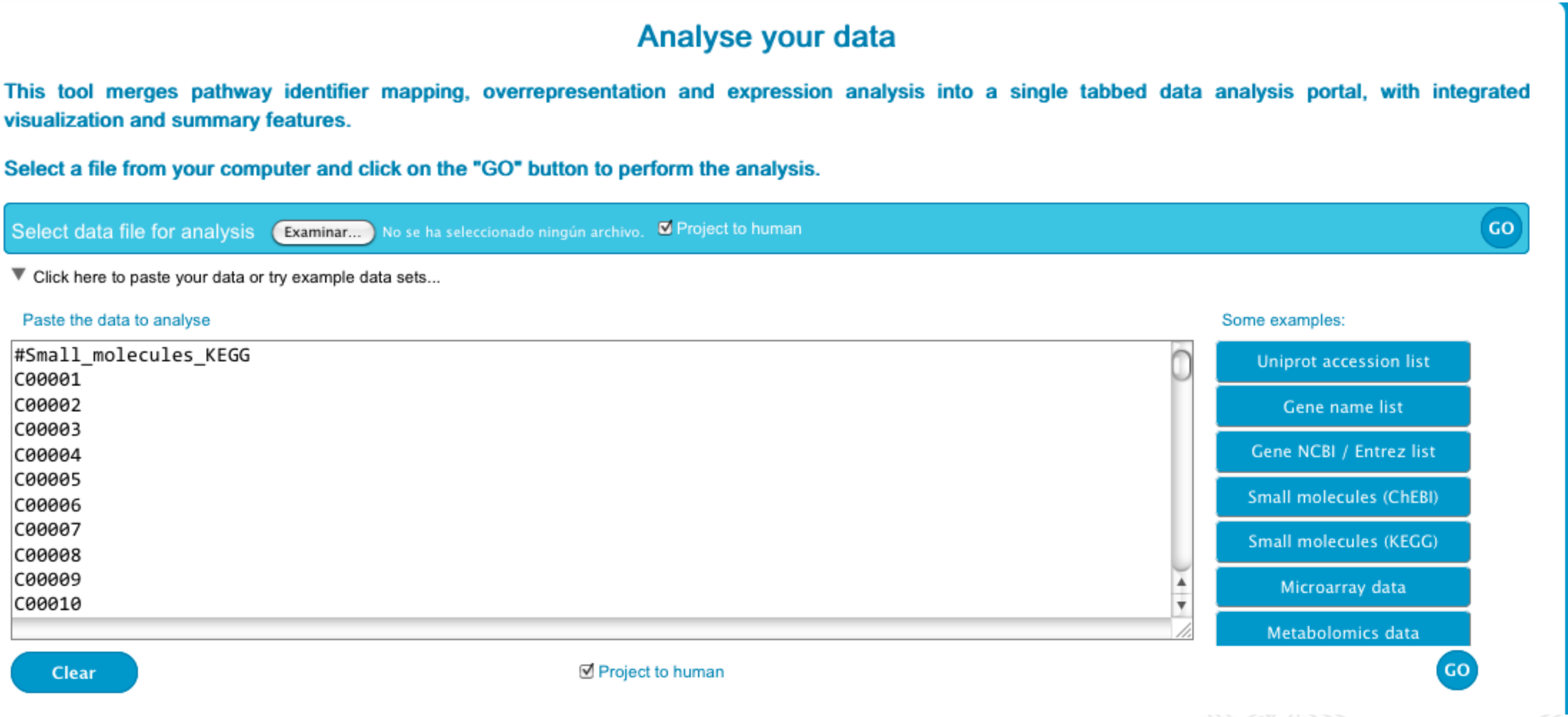
Current Version: Reactome V54

reactome @reactome 2 Nov

Our upgraded search tool has better performance, an improved results display, and integrated help desk access #solrlucene

reactome @reactome 26 Oct

Enrichment Analysis in Reactome



Analyse your data

This tool merges pathway identifier mapping, overrepresentation and expression analysis into a single tabbed data analysis portal, with integrated visualization and summary features.

Select a file from your computer and click on the "GO" button to perform the analysis.

Select data file for analysis Examinar... No se ha seleccionado ningún archivo. ☑ Project to human GO

▼ Click here to paste your data or try example data sets...

Paste the data to analyse

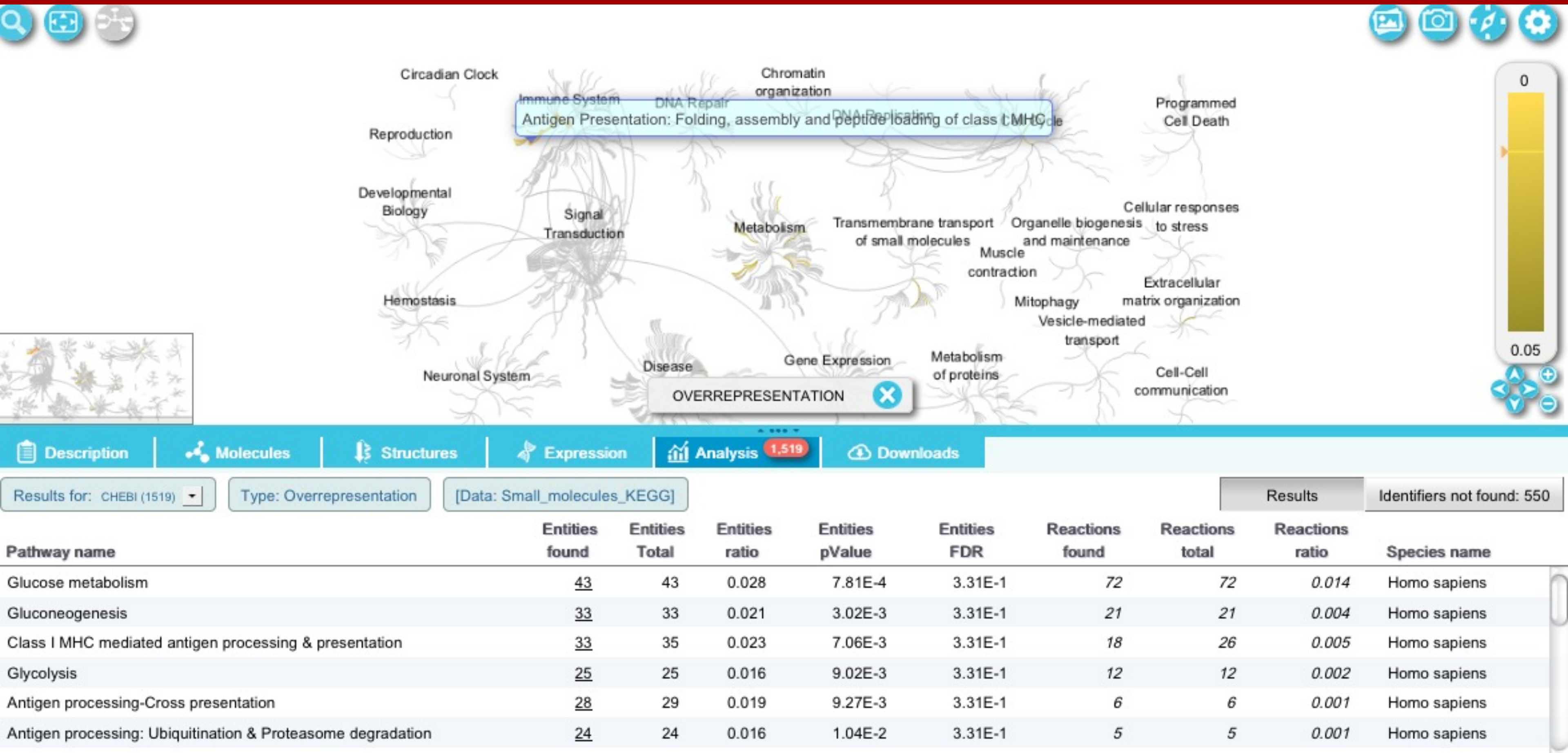
```
#Small_molecules_KEGG
C00001
C00002
C00003
C00004
C00005
C00006
C00007
C00008
C00009
C00010
```

Some examples:

☑ Project to human

Enrichment Analysis in Reactome

UNIVERSIDAD DE
MURCIA



Results for: CHEBI (1519)

Type: Overrepresentation

[Data: Small_molecules_KEGG]

Results

Identifiers not found: 550

Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	Species name
Glucose metabolism	43	43	0.028	7.81E-4	3.31E-1	72	72	0.014	Homo sapiens
Gluconeogenesis	33	33	0.021	3.02E-3	3.31E-1	21	21	0.004	Homo sapiens
Class I MHC mediated antigen processing & presentation	33	35	0.023	7.06E-3	3.31E-1	18	26	0.005	Homo sapiens
Glycolysis	25	25	0.016	9.02E-3	3.31E-1	12	12	0.002	Homo sapiens
Antigen processing-Cross presentation	28	29	0.019	9.27E-3	3.31E-1	6	6	0.001	Homo sapiens
Antigen processing: Ubiquitination & Proteasome degradation	24	24	0.016	1.04E-2	3.31E-1	5	5	0.001	Homo sapiens

Exercises –
Online Enrichment Analysis

- AulaVirtual (Recursos ➔ prácticas ➔enrichment ➔ ejercicio1-GOReactomeOnlineEnrichment.pdf)

Enrichment Analysis in R

- AulaVirtual (Recursos → prácticas→enrichment → ejercicio2-EnrichmentR.pdf)

- Upload to Rstudio: Recursos → prácticas→enrichment → files-enrichment.zip

Questions, comments…