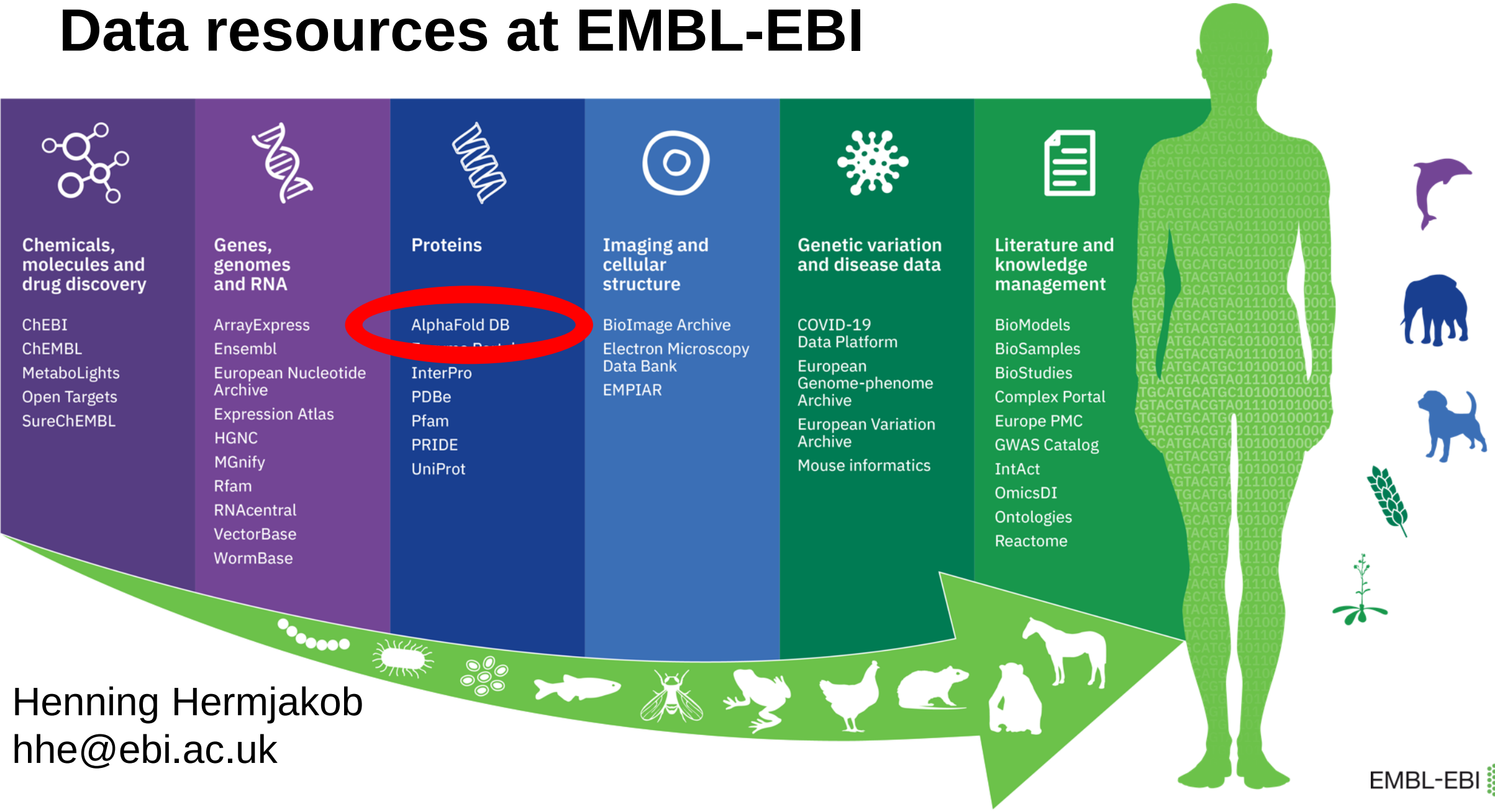
Data resources at EMBL-EBI



Henning Hermjakob
hhe@ebi.ac.uk

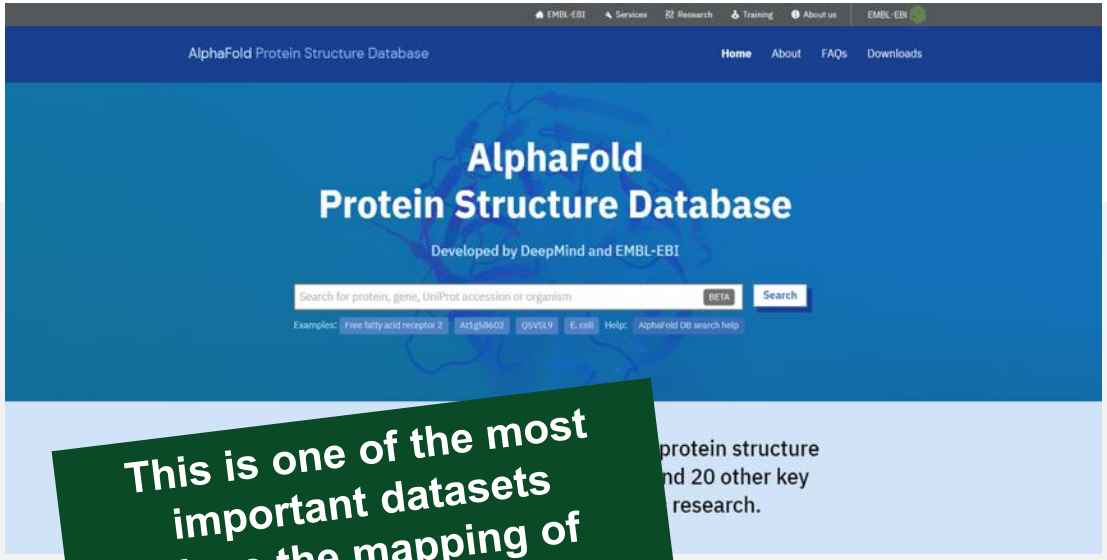
EMBL-EBI

AlphaFold Protein Structure Database

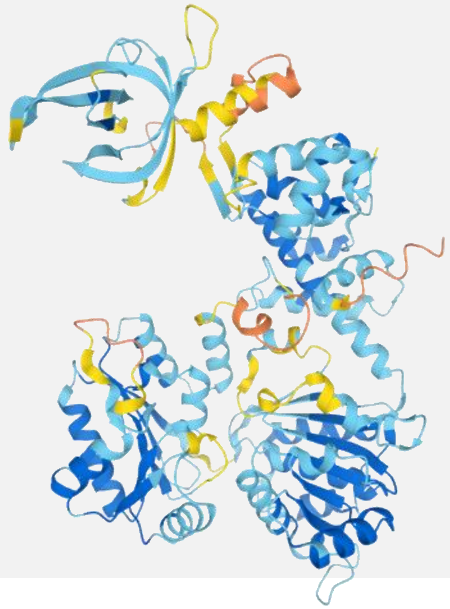
- Launched on 22nd July 2021
  - Open access (CC-BY-4.0 license) to all structure models

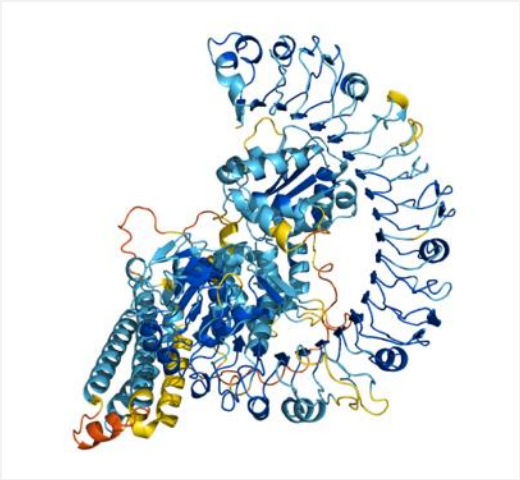
- V1: ~365,000 models (21 proteomes)
- V2: >800,000 models (SwissProt entries)
- V3: ~1,000,000 Structures (focus on pathogens and neglected diseases; 49 proteomes)
- Later in 2022: >100,000,000 models

NAR paper on AlphaFold DB: bit.ly/NAR-AFDB-2022



This is one of the most important datasets since the mapping of the Human Genome



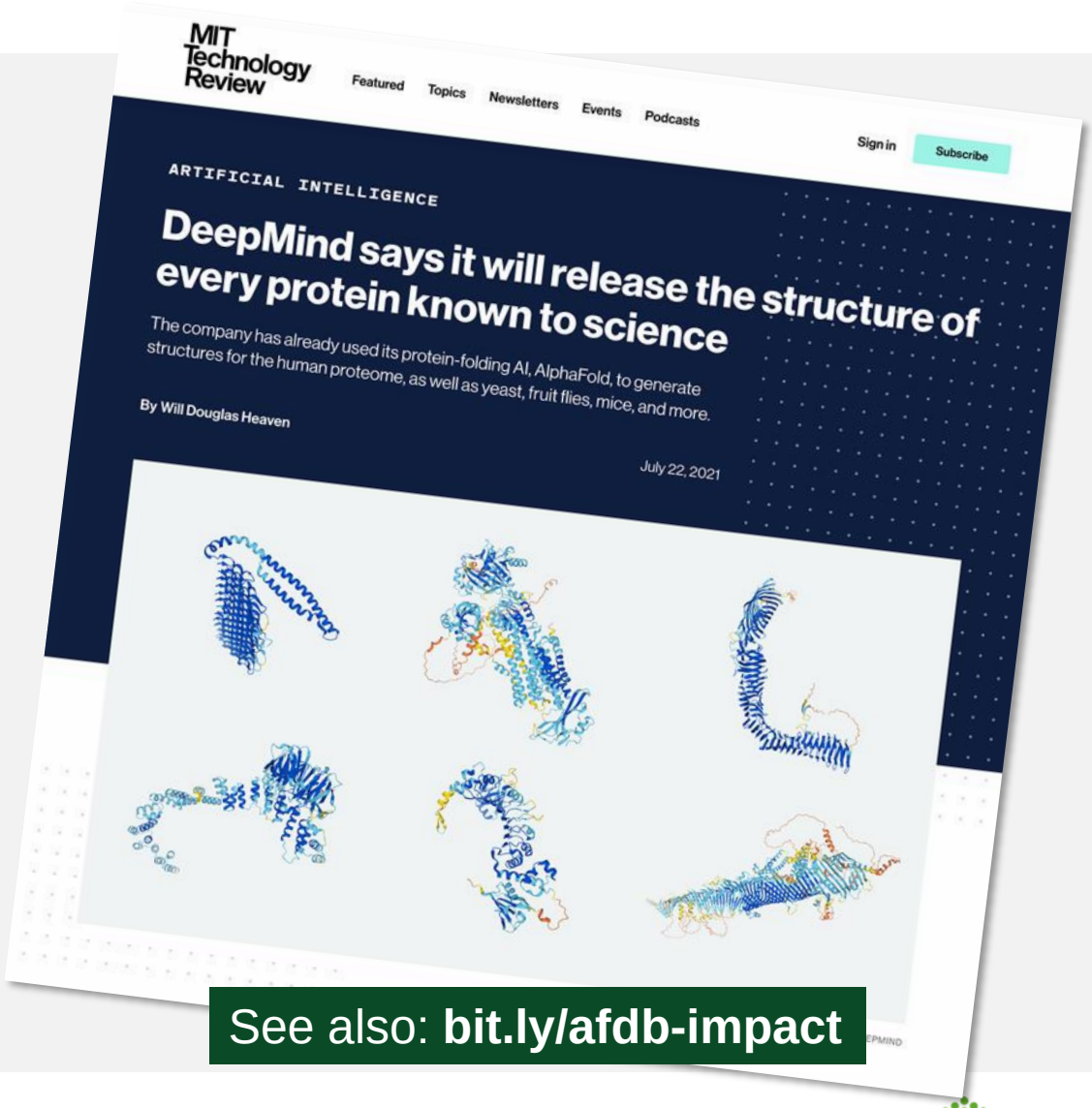


pdbe.org | pdbe-kb.org | alphafold.ebi.ac.uk | 3d-beacons.org

DeepMind EMBL-EBI

Impact of AlphaFold database on life science research

- Structure for every known protein in UniProt database will be available or can be modelled
- 3-D structures for virtually all (98.5%) of the human proteome
  - 92.5% residues are covered
  - 58% of all residues have very confident(36%) or confident (22%) predictions
  - ~30-40% are disordered and are not expected to have good prediction
- Protein Data Bank content
  - Only ~17% of the human proteome residues covered in the PDB

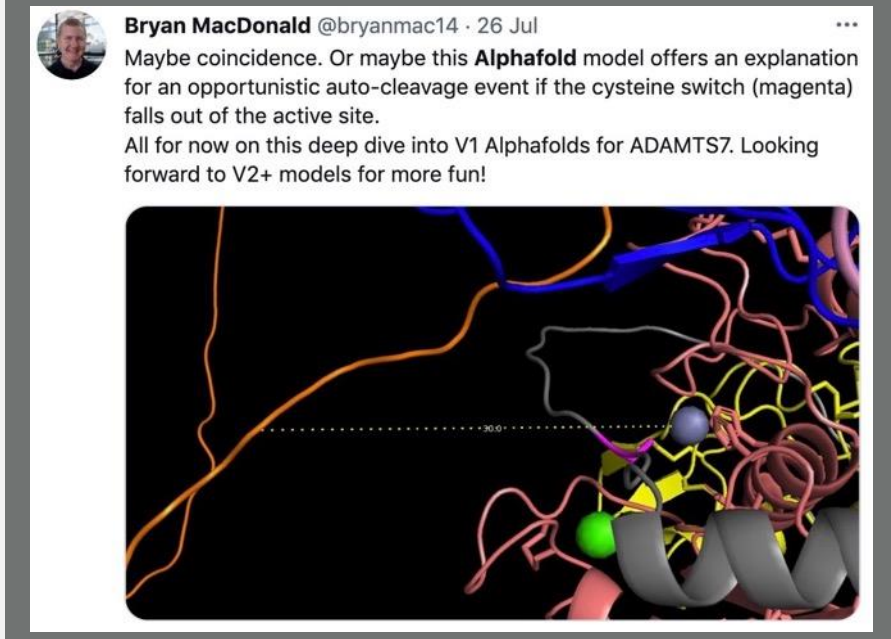


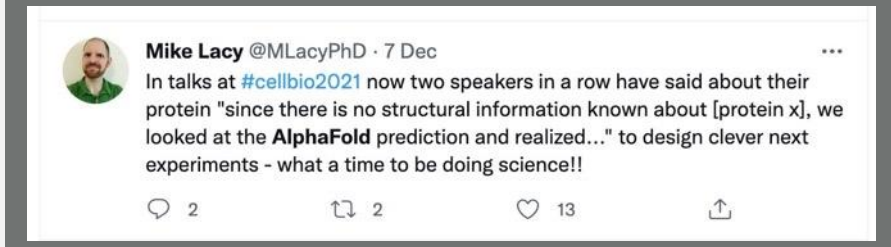
See also: **bit.ly/afdb-impact**

DeepMind EMBL-EBI

New normal for biology

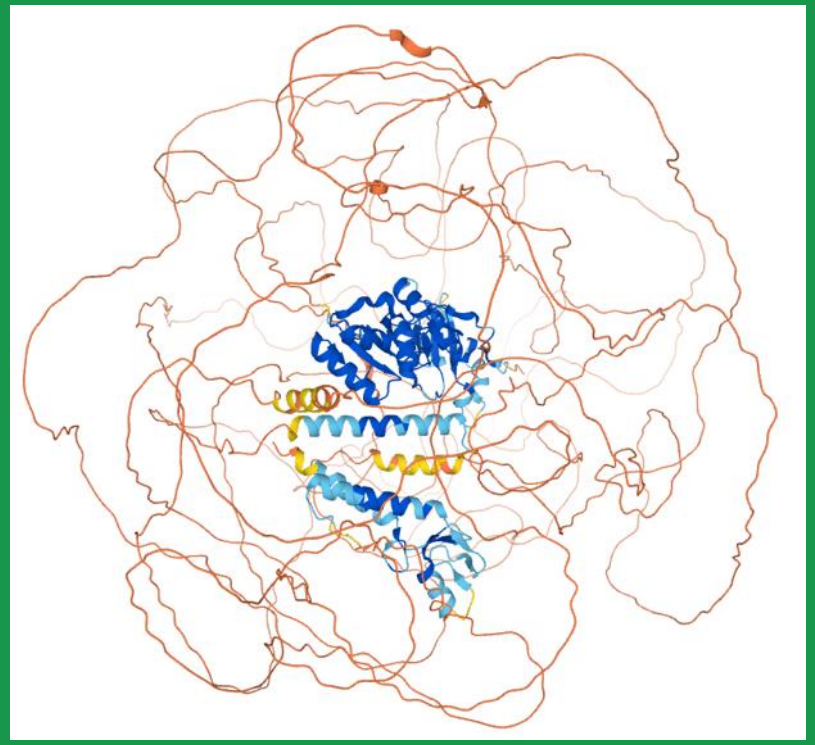
- Many studies of proteins can now start with a 3D model as the basis for hypotheses
- Does a model explain previously known data on function, binding, specificity, …?
- Design experiments based on the model (e.g., introduce mutations)





DeepMind EMBL-EBI

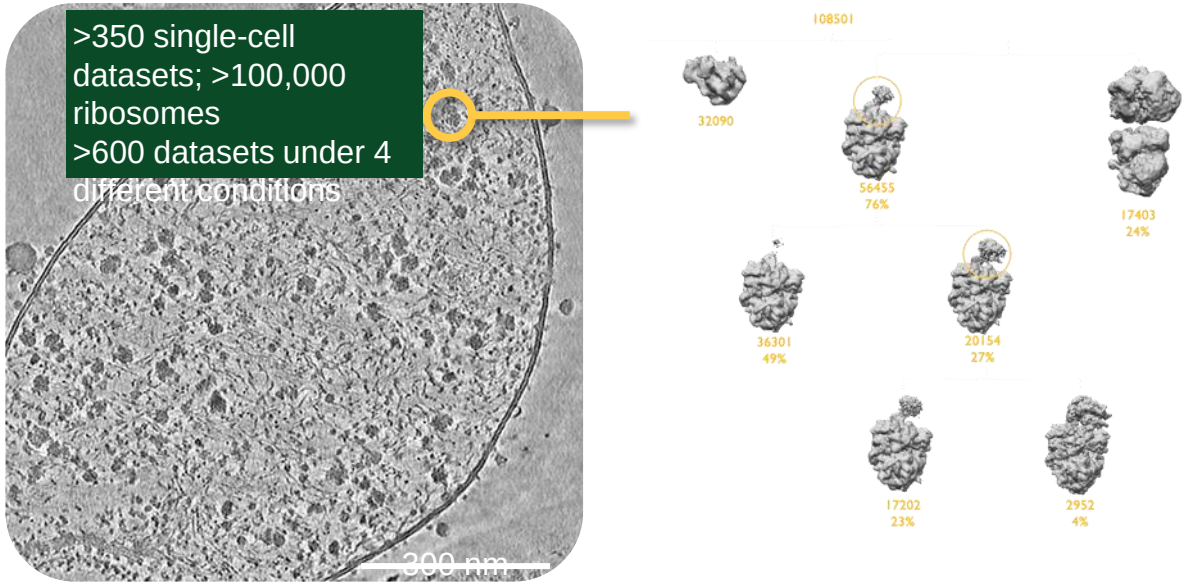
Impact of AlphaFold database on life science research



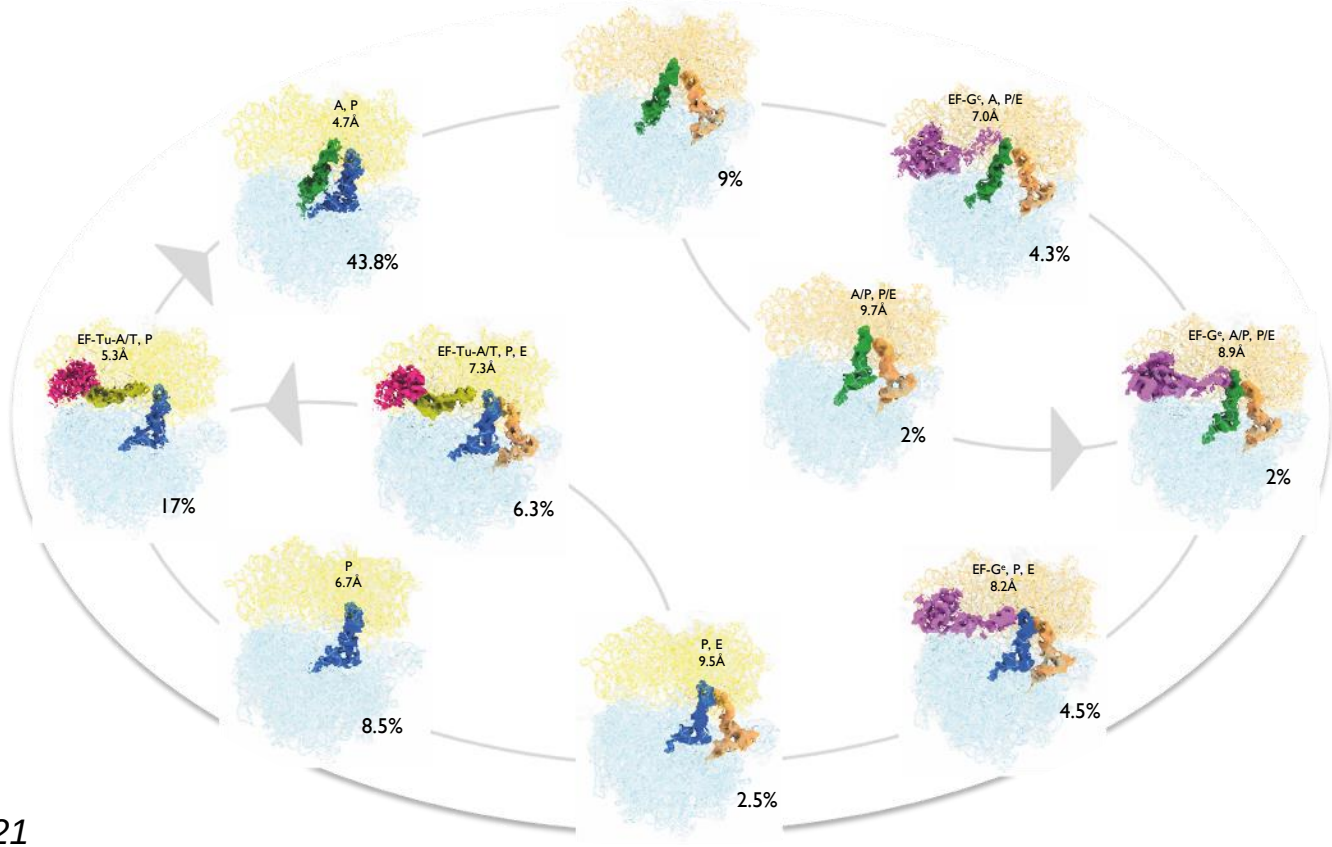
- “An unexpected effect of the AFDB is that it visually demonstrates the prevalence of intrinsically disordered regions (IDRs) in the human proteome”
  - Ried Alderson et al. doi: https://doi.org/10.1101/2022.02.18.481080
- “Proteins” are no longer seen as having a well defined and compact structure
- Inclusion of unstructured regions with pLDDT <70 in AFDB structures was a conscious decision

From Structural Inventories to Processes: Visual proteomics

Single-cell translation profiles
Molecular sociology of ribosomes
Translation landscapes reshaped by antibiotics



Tegunov, Xue, Dienemann, Cramer & Mahamid. *Nature Methods 2021*
O'Reilly*, Xue*, Graziadei *, Sinn, Lenz, Tegunov, Blötz, Hagen, Cramer, Stülke, Mahamid & Rappsilber. *Science 2020*

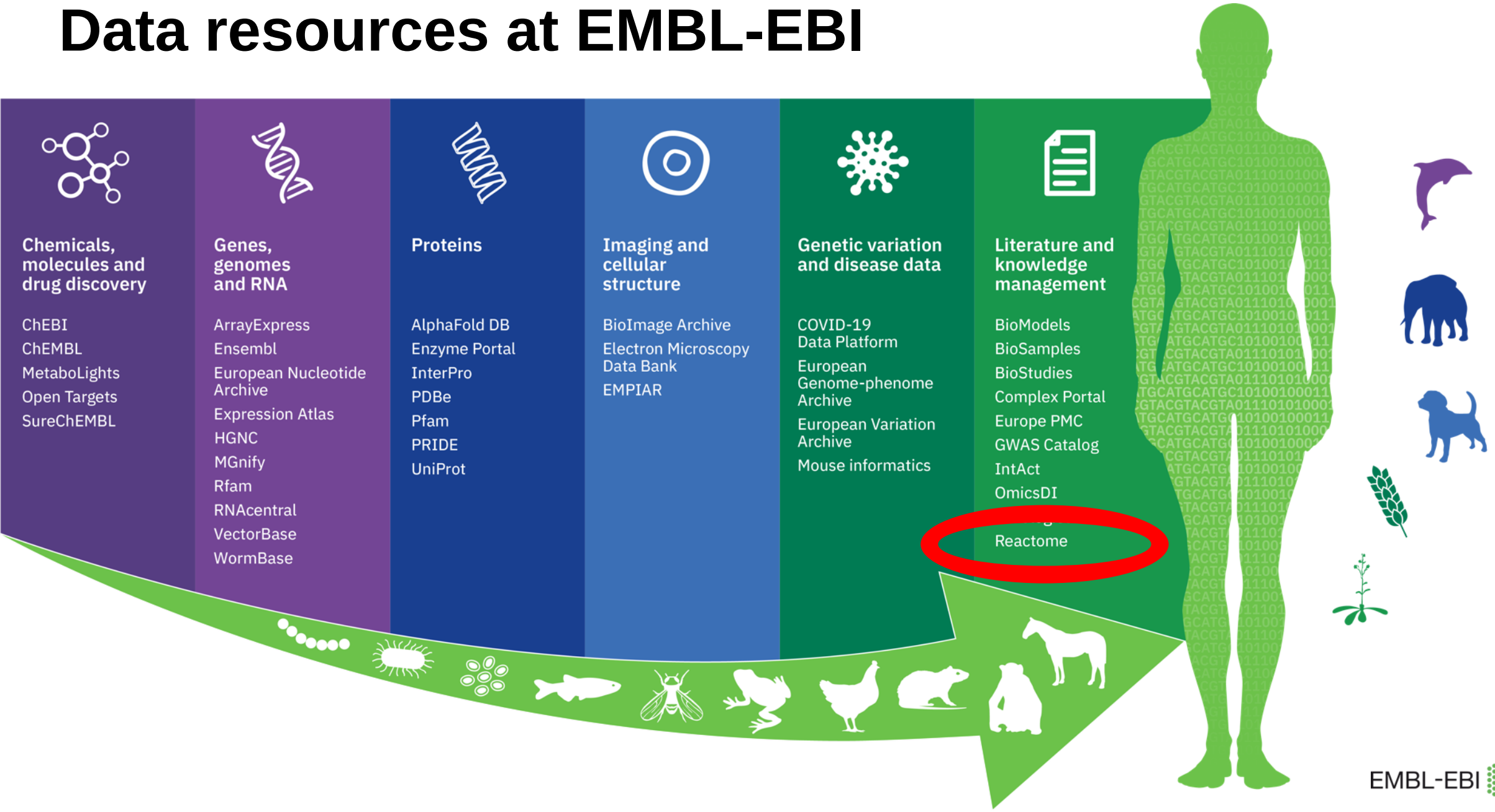


Xue, Lenz, Zimmermann-Kogadeeva, Tegunov, Cramer, Bork, Rappsilber & Mahamid. *Biorxiv 2021; under review*

pdbe.org | pdbe-kb.org | alphafold.ebi.ac.uk | 3d-beacons.org

DeepMind EMBL-EBI

Data resources at EMBL-EBI

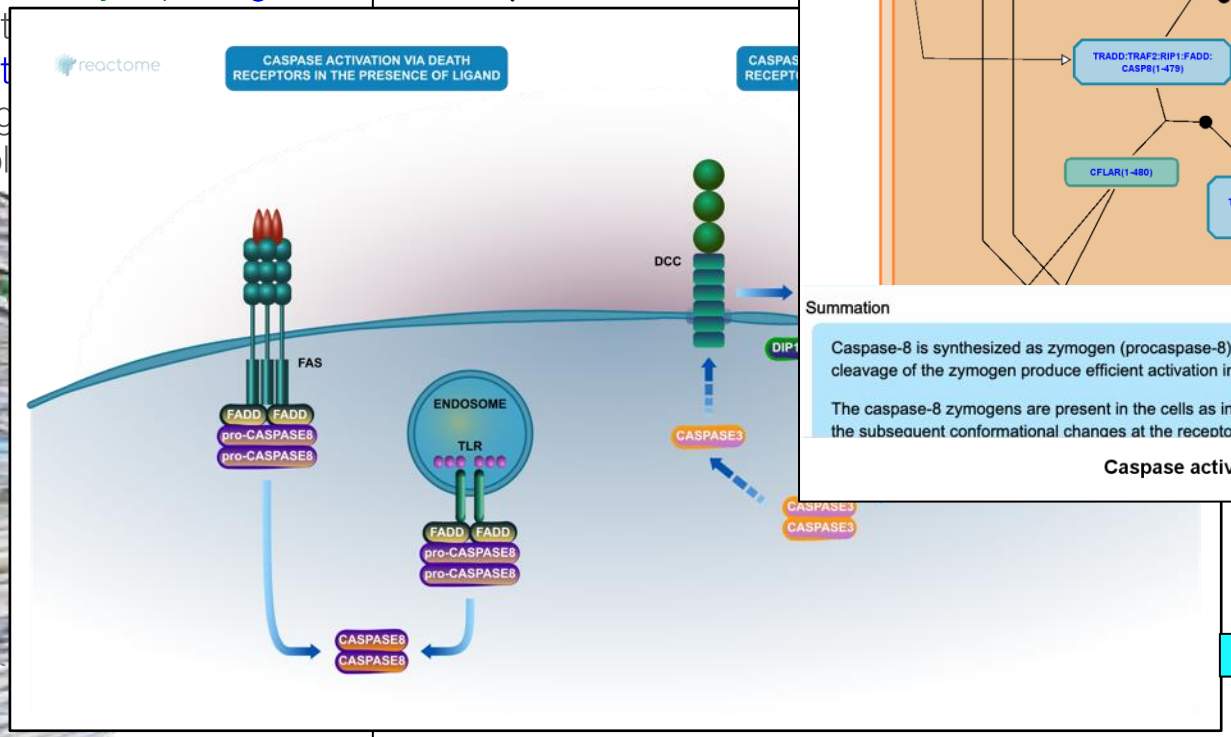


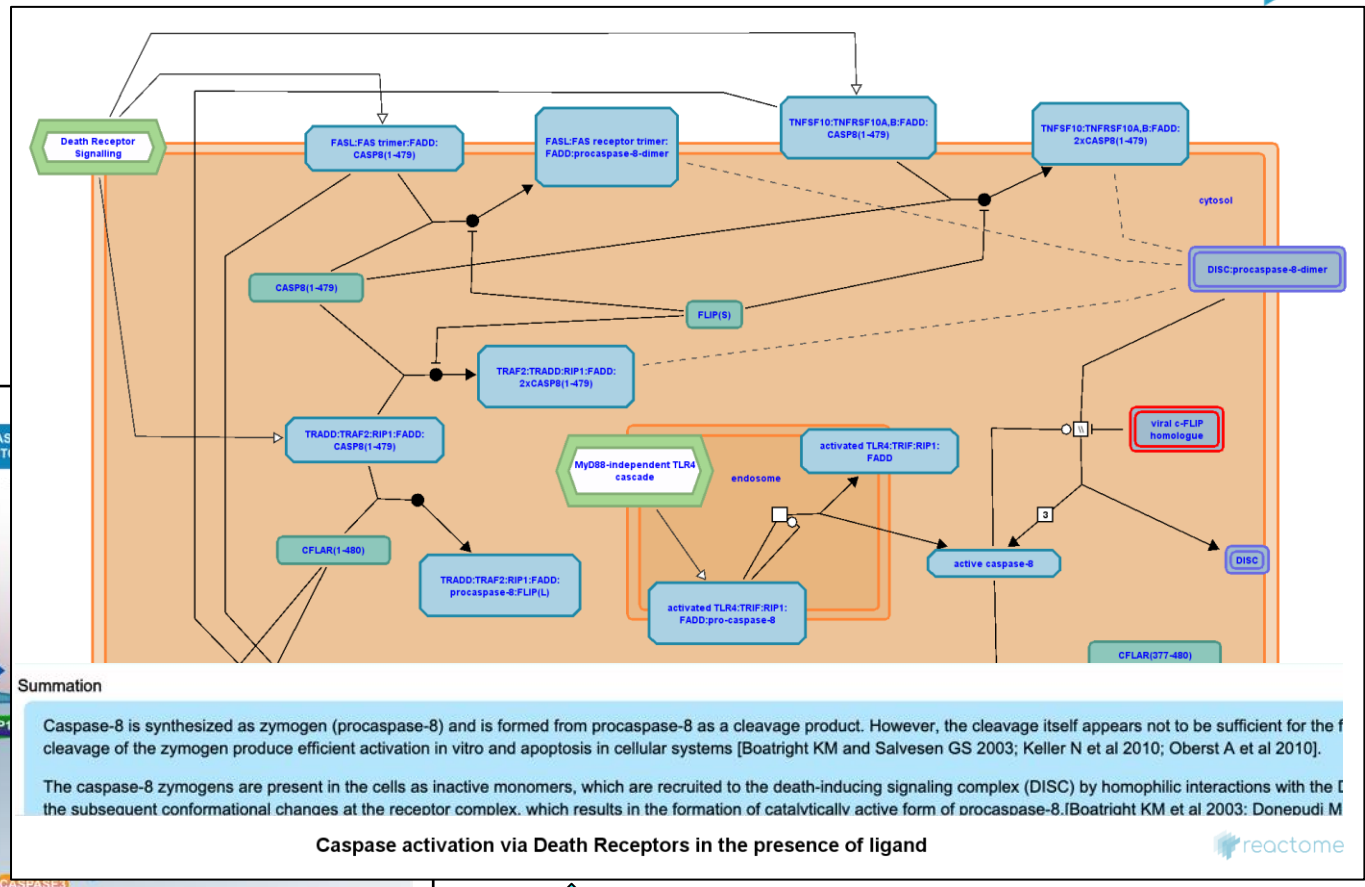
EMBL-EBI

Reactome: Computationally Accessible Pathway Reviews

Nature 407(6805):770-6.The Biochemistry of Apoptosis.

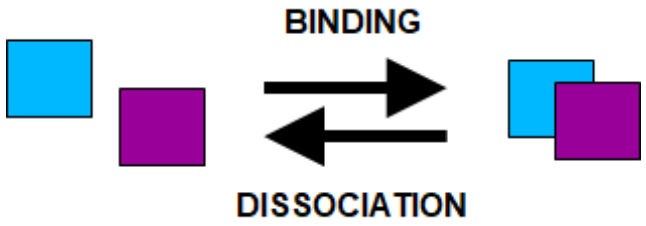
"Caspase-8 is the key initiator caspase in the death-receptor pathway. Upon ligand binding, death recept (Apo-1/Fas) aggregat membrane-bound sig (Box 3). These compl throug of pro conc proxi crowd protea sufficie molec each o action activati includi caspas

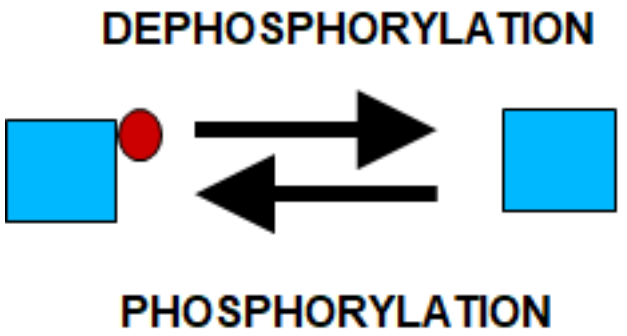


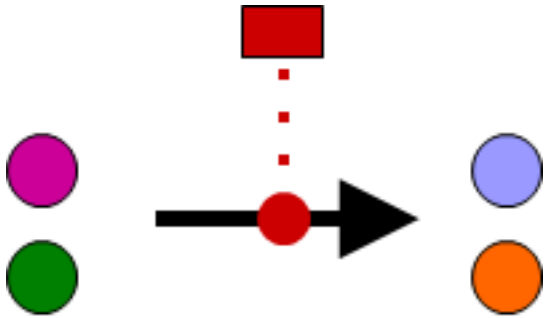


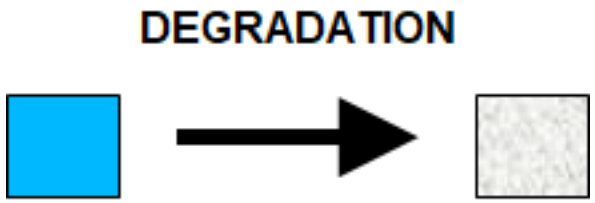
http://www.reactome.org

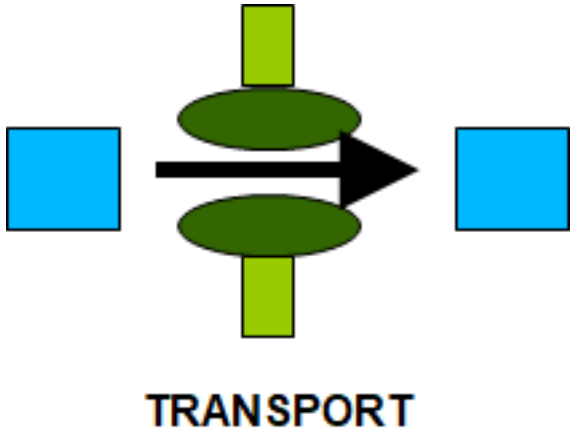
Reactions as Building Blocks



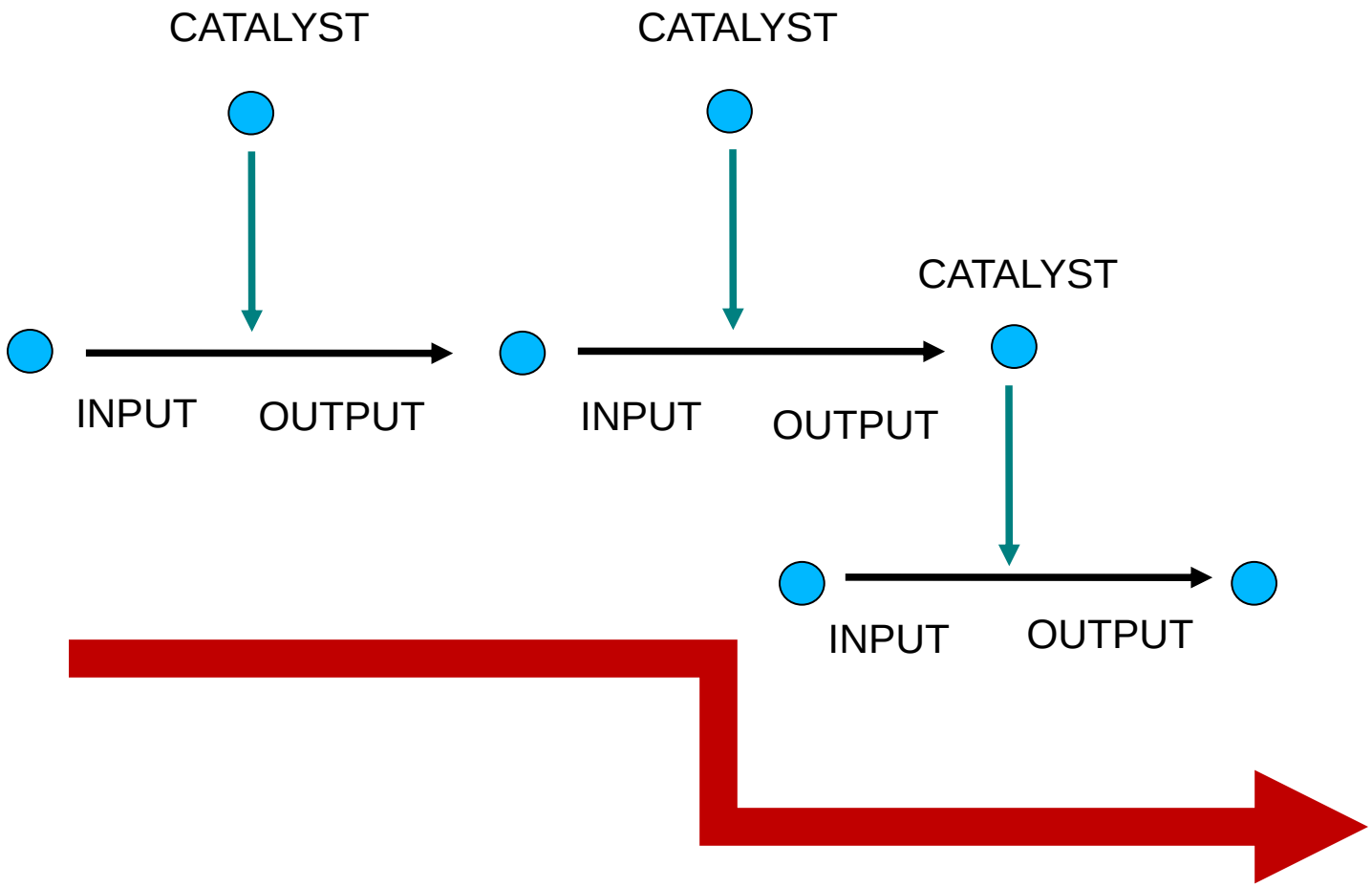




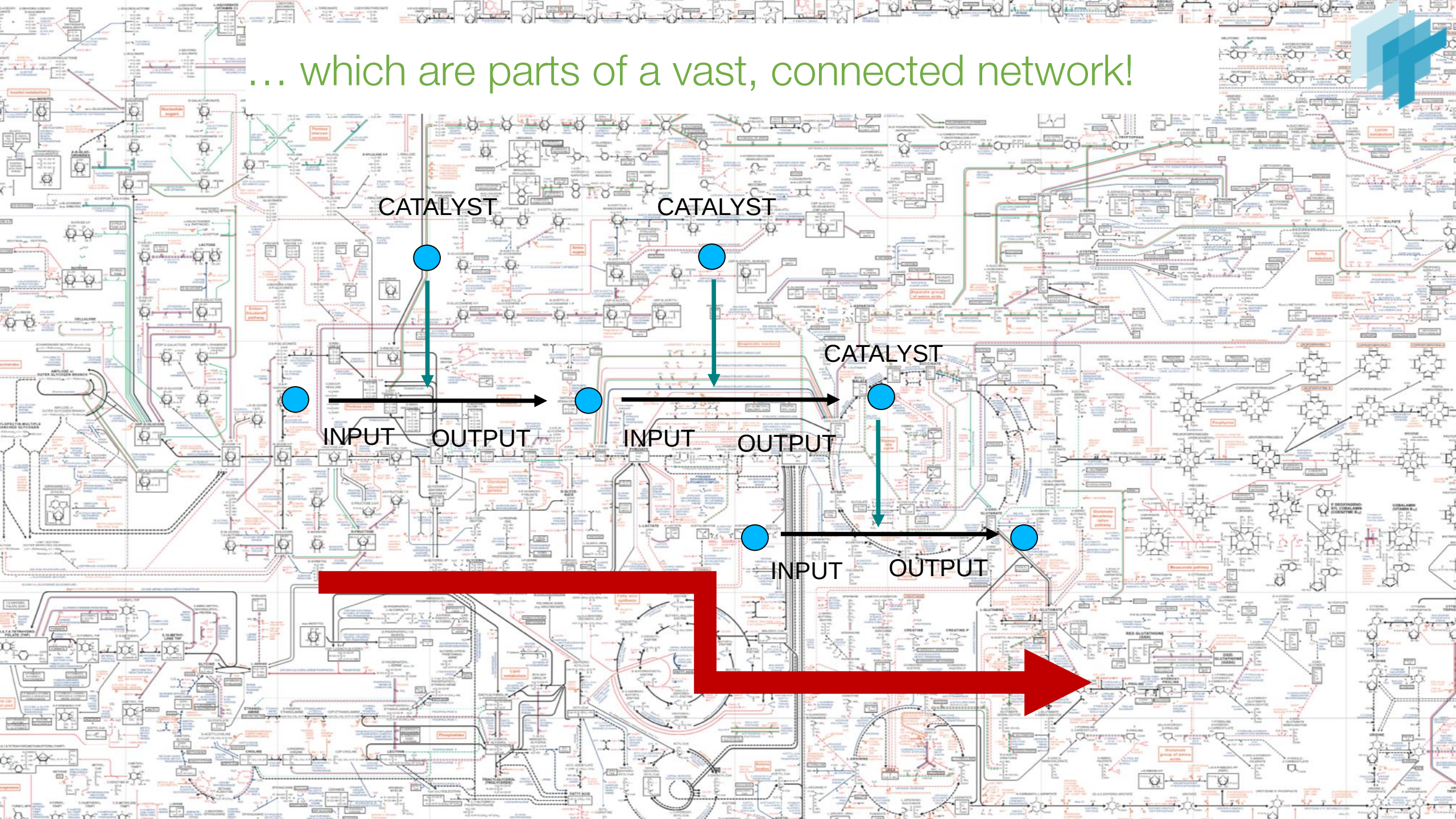




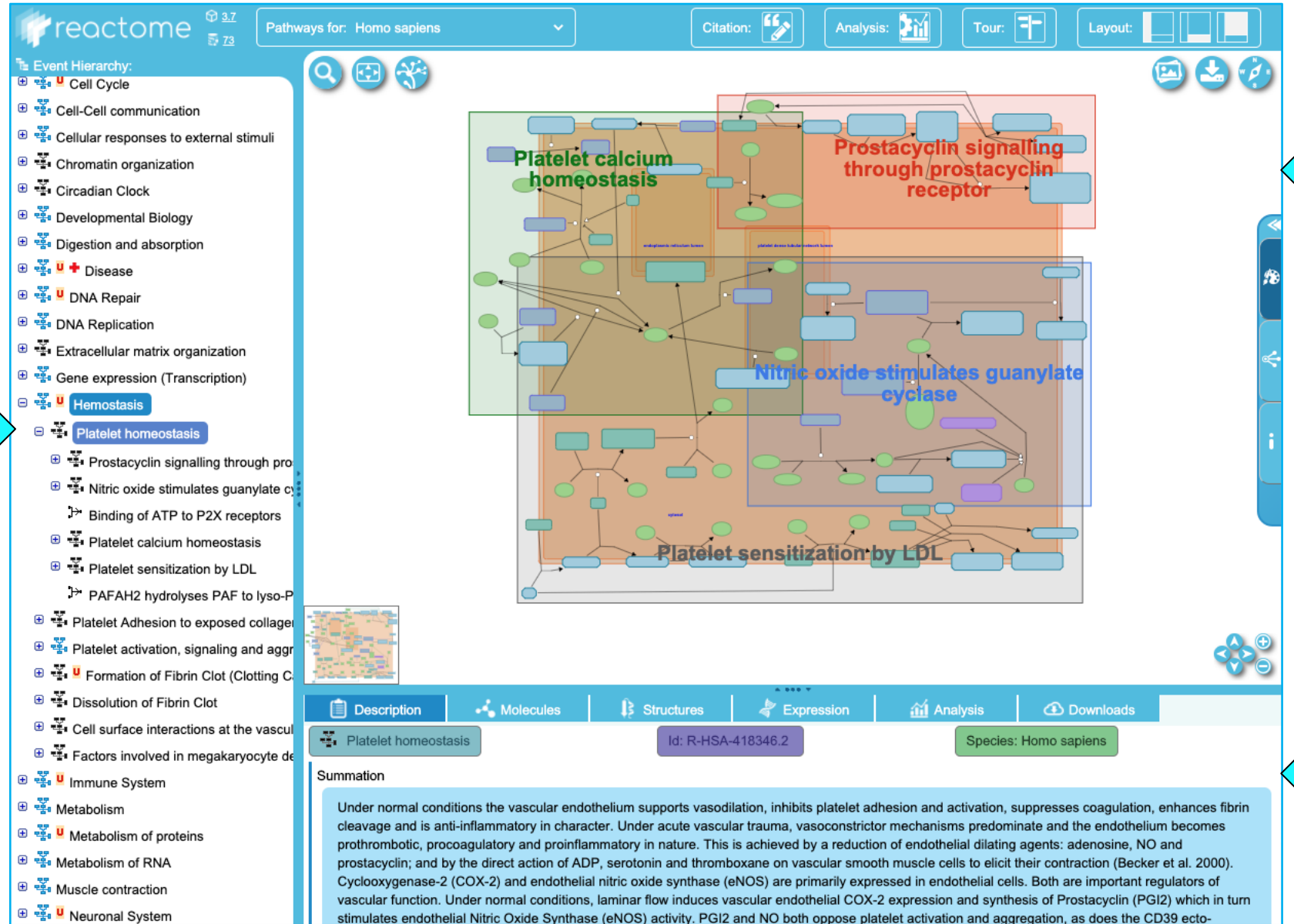
Reactions Connect into Pathways



… which are parts of a vast, connected network!



Pathway Browser

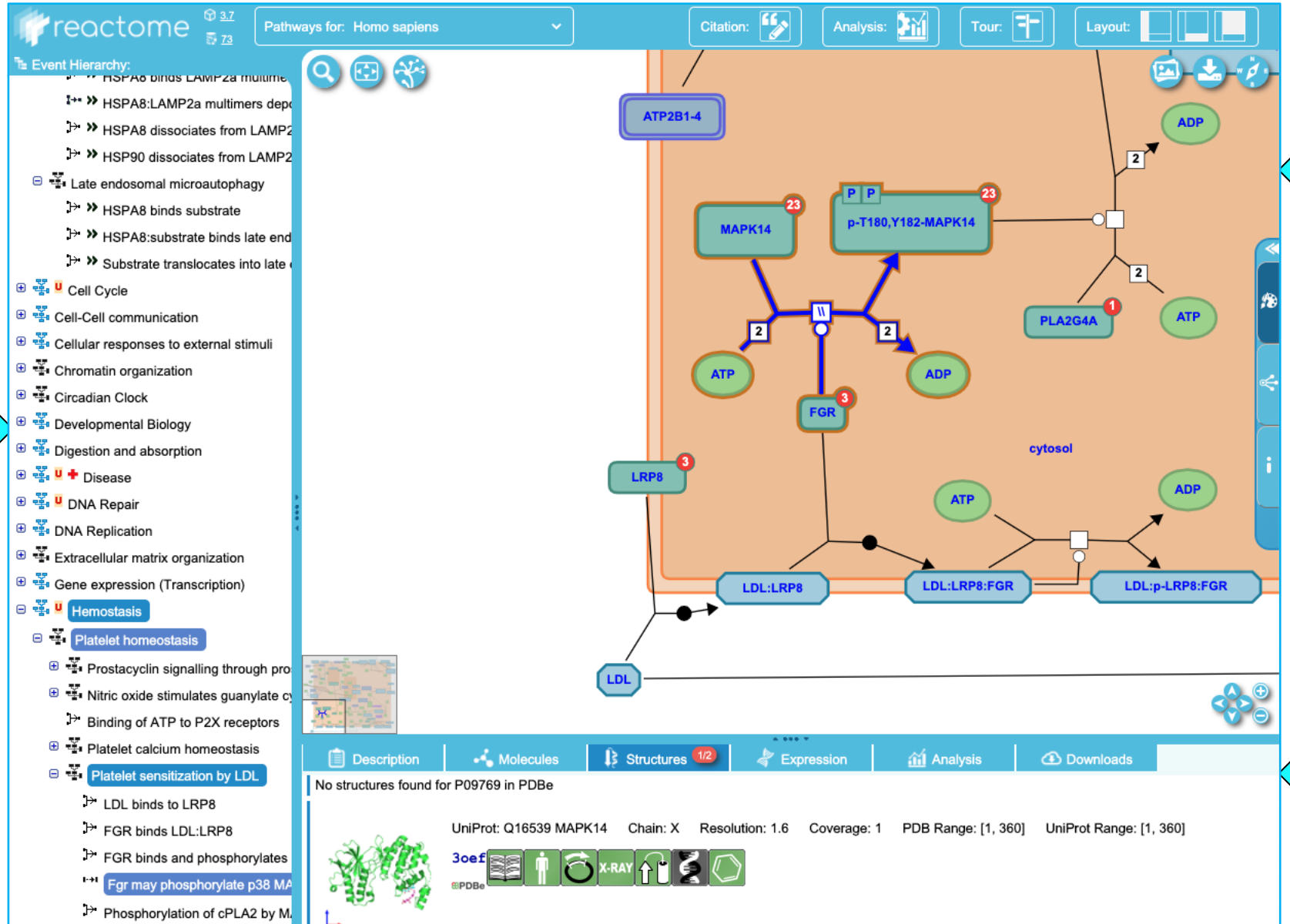


Hierarchy

Mol. Map

Details

Pathway Browser

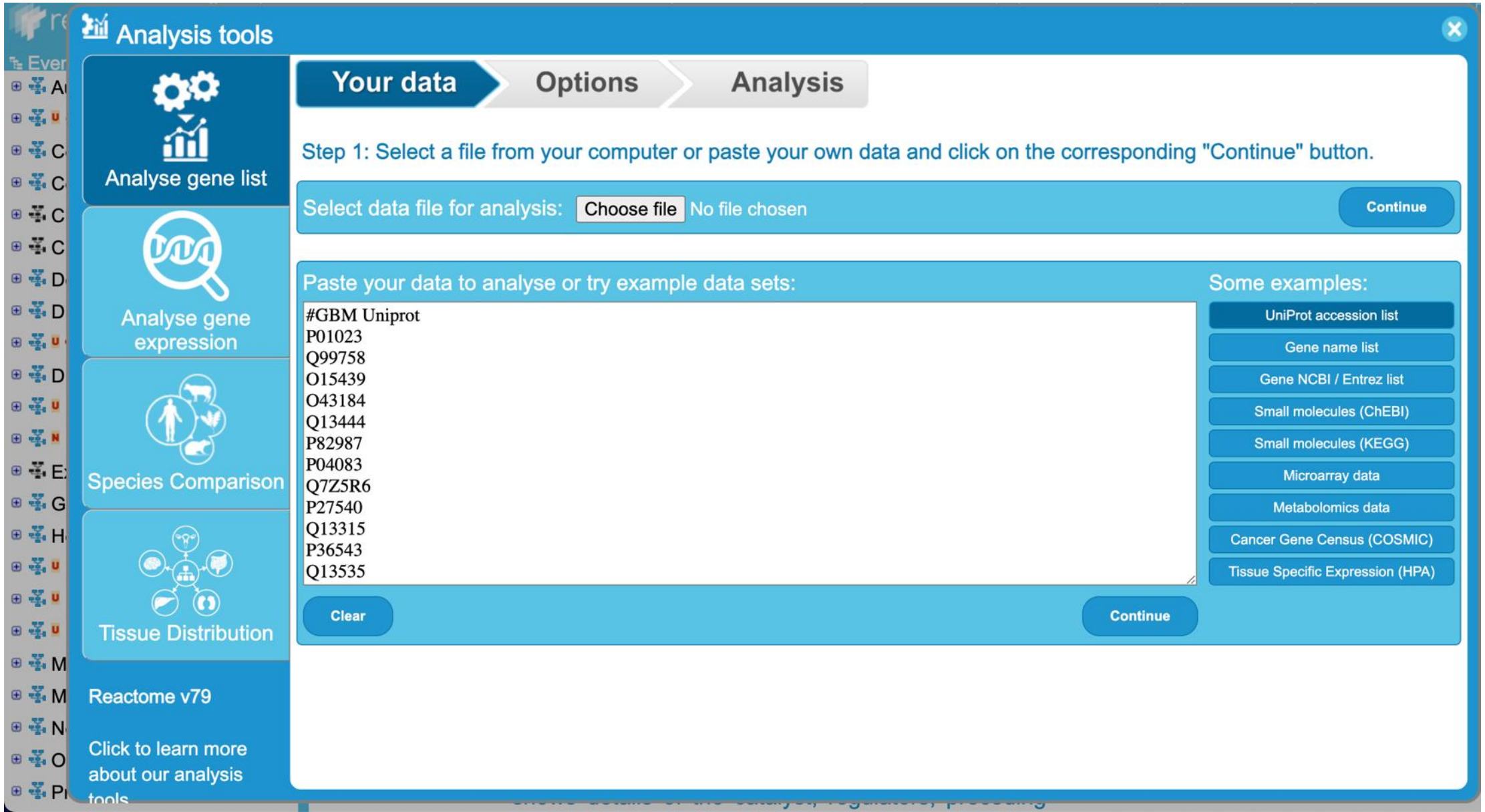


Hierarchy

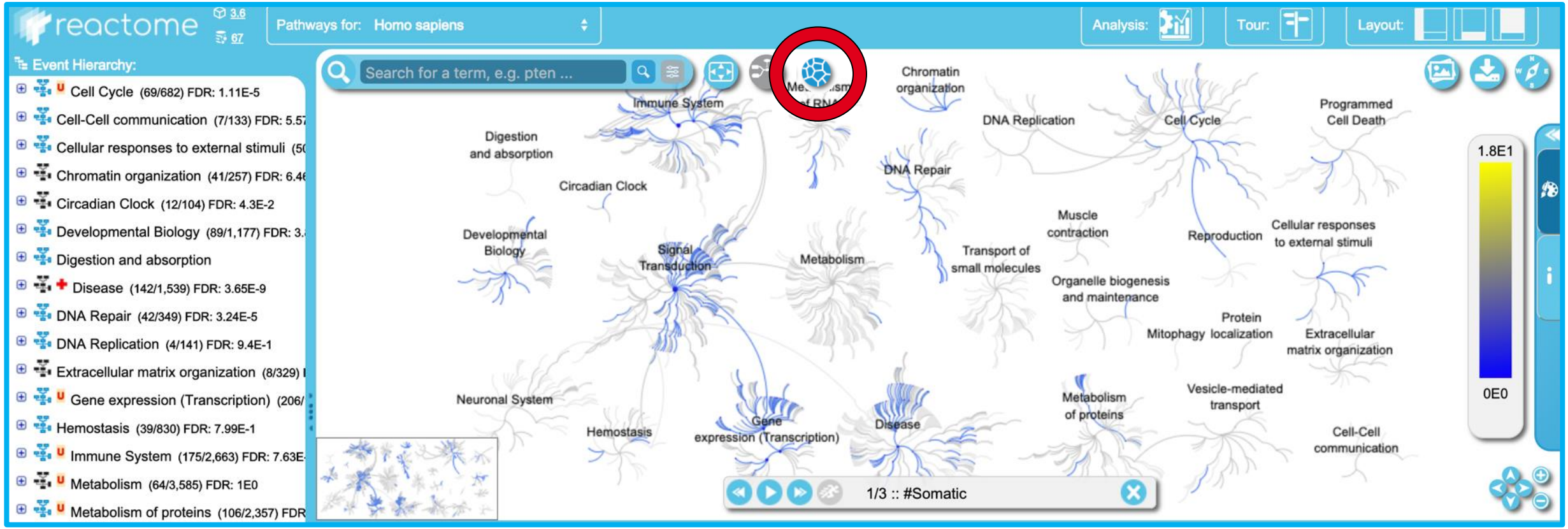
Mol. Map

Details

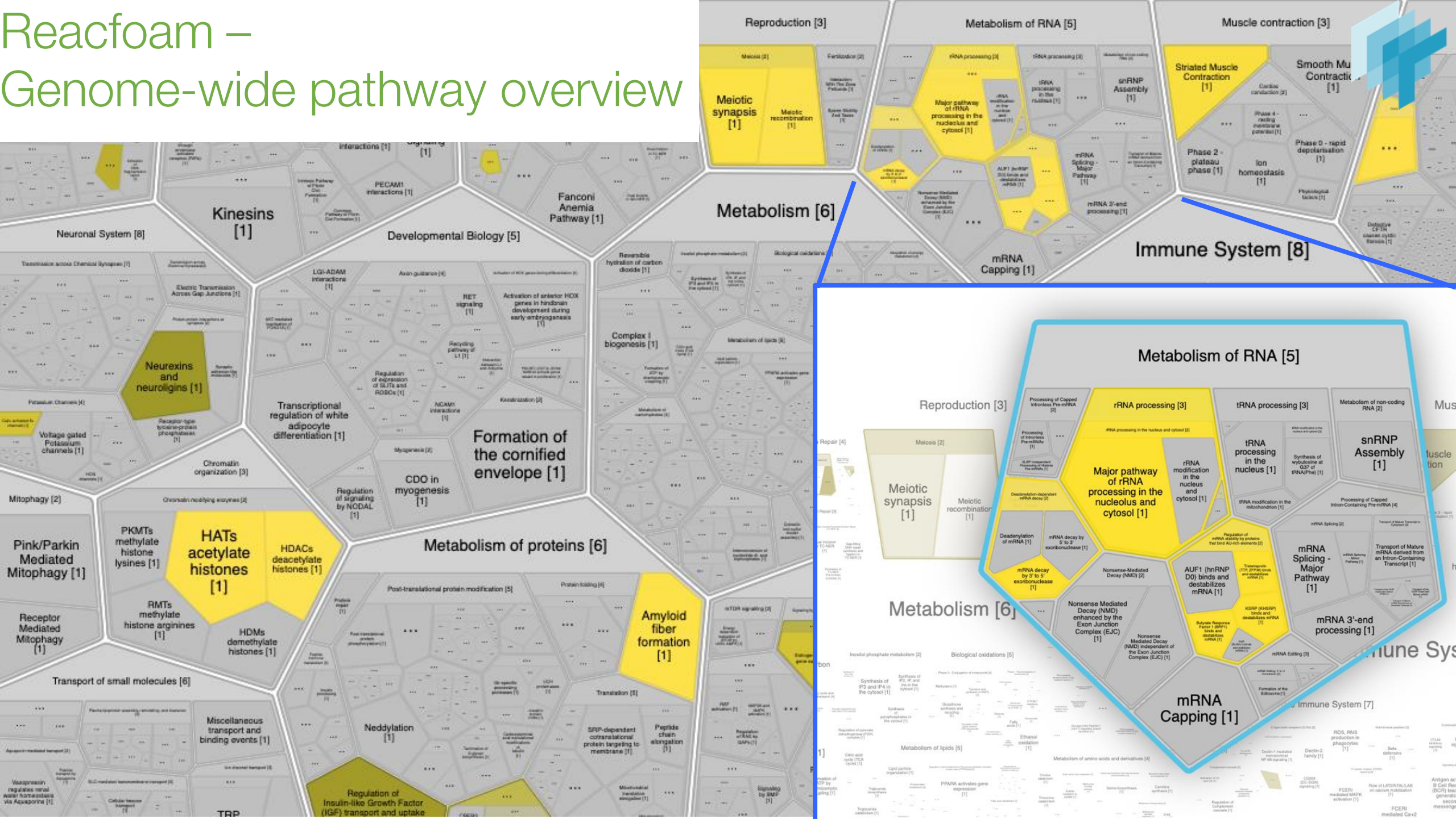
Pathway Analysis



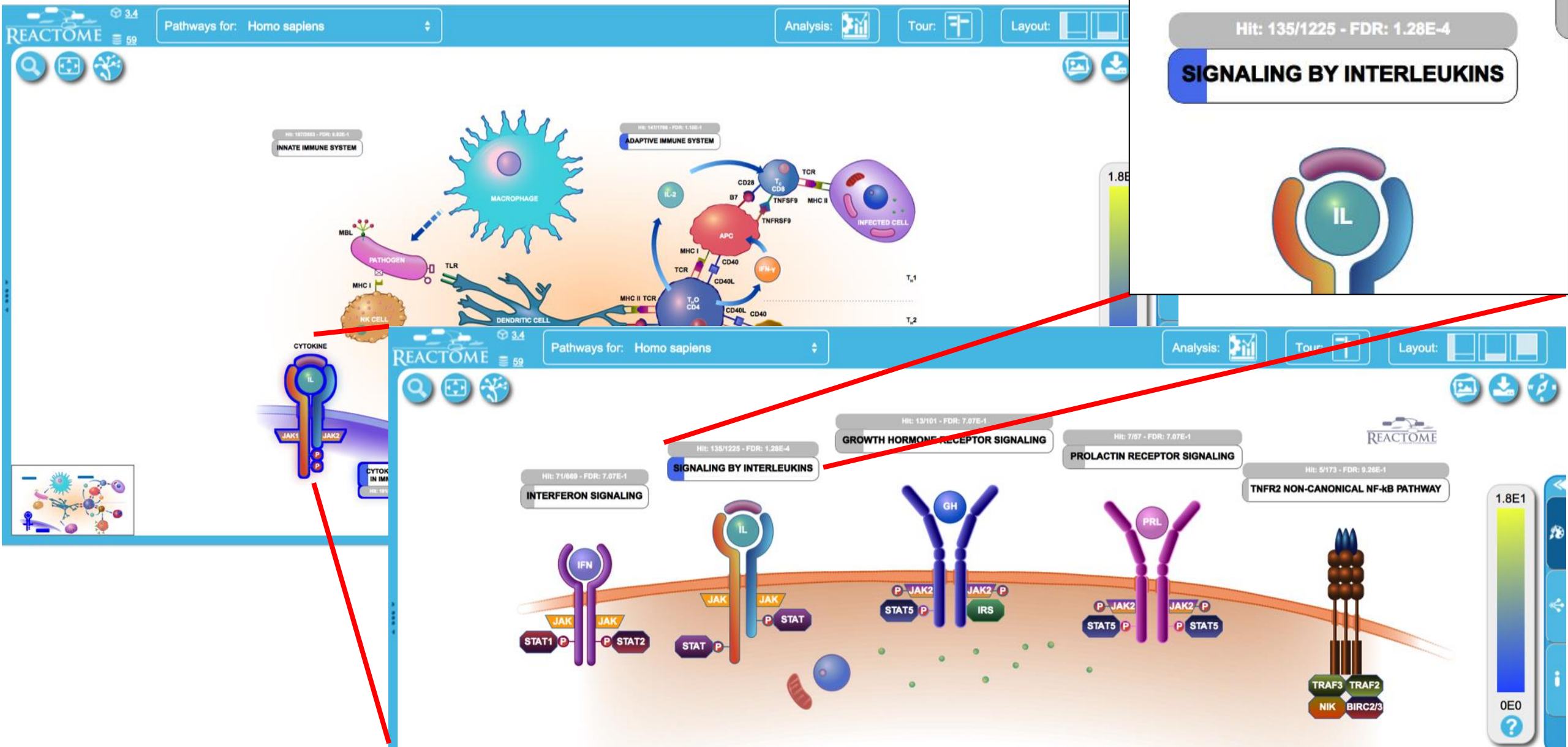
Pathway Analysis: Results overview - Fireworks



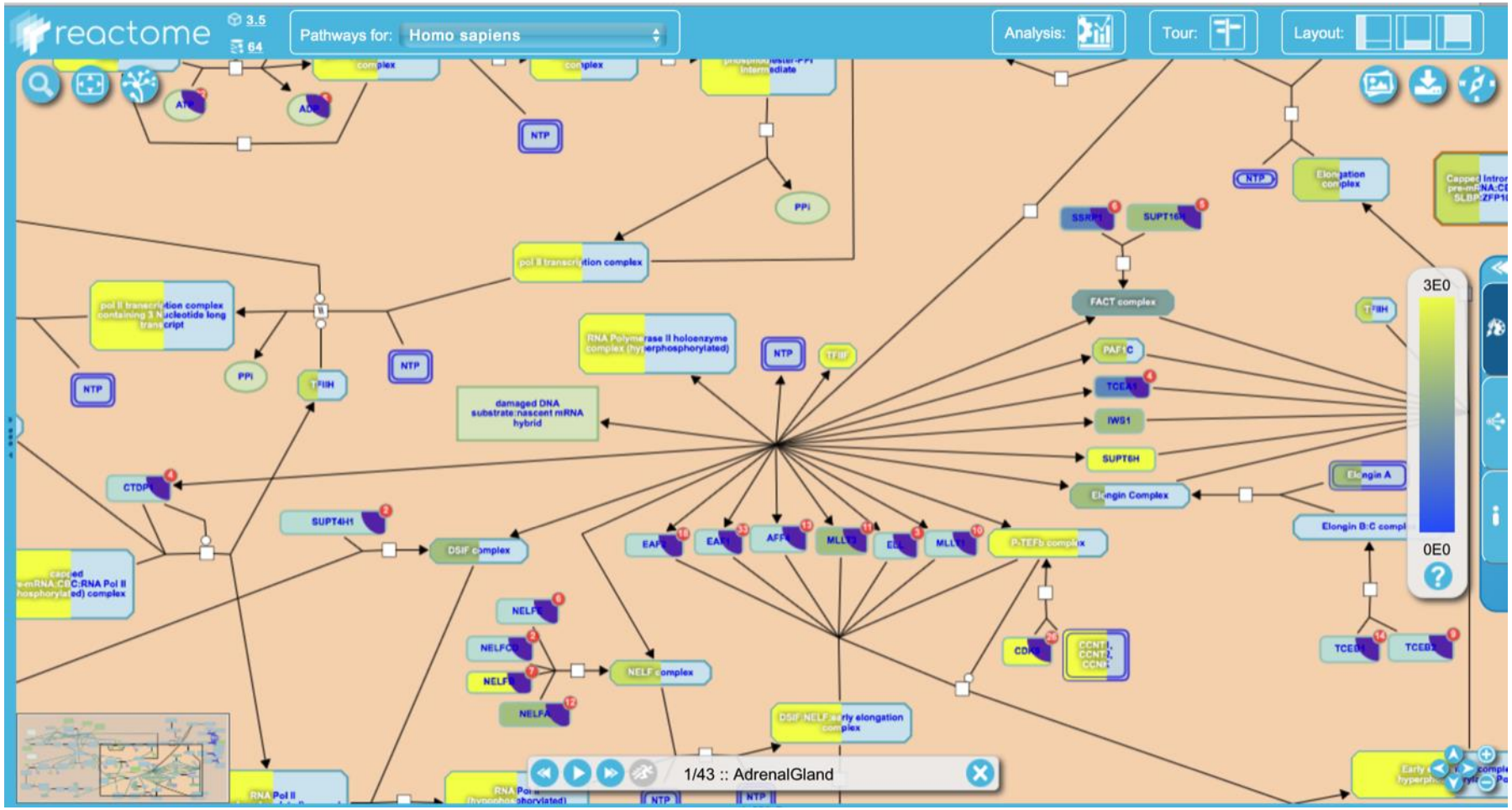
Reacfoam – Genome-wide pathway overview



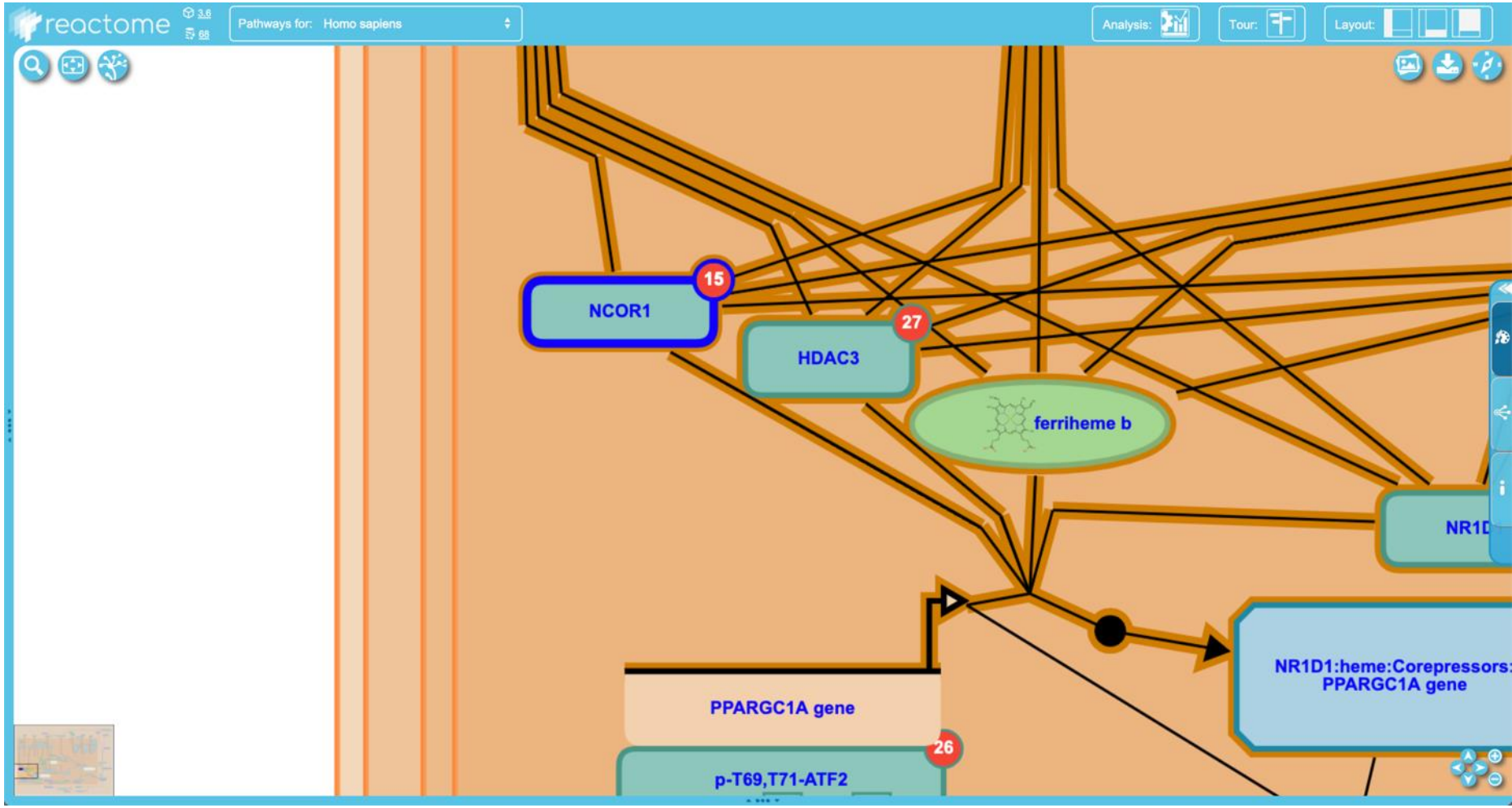
Textbook-style high level diagrams



Pathway Analysis: Detail view

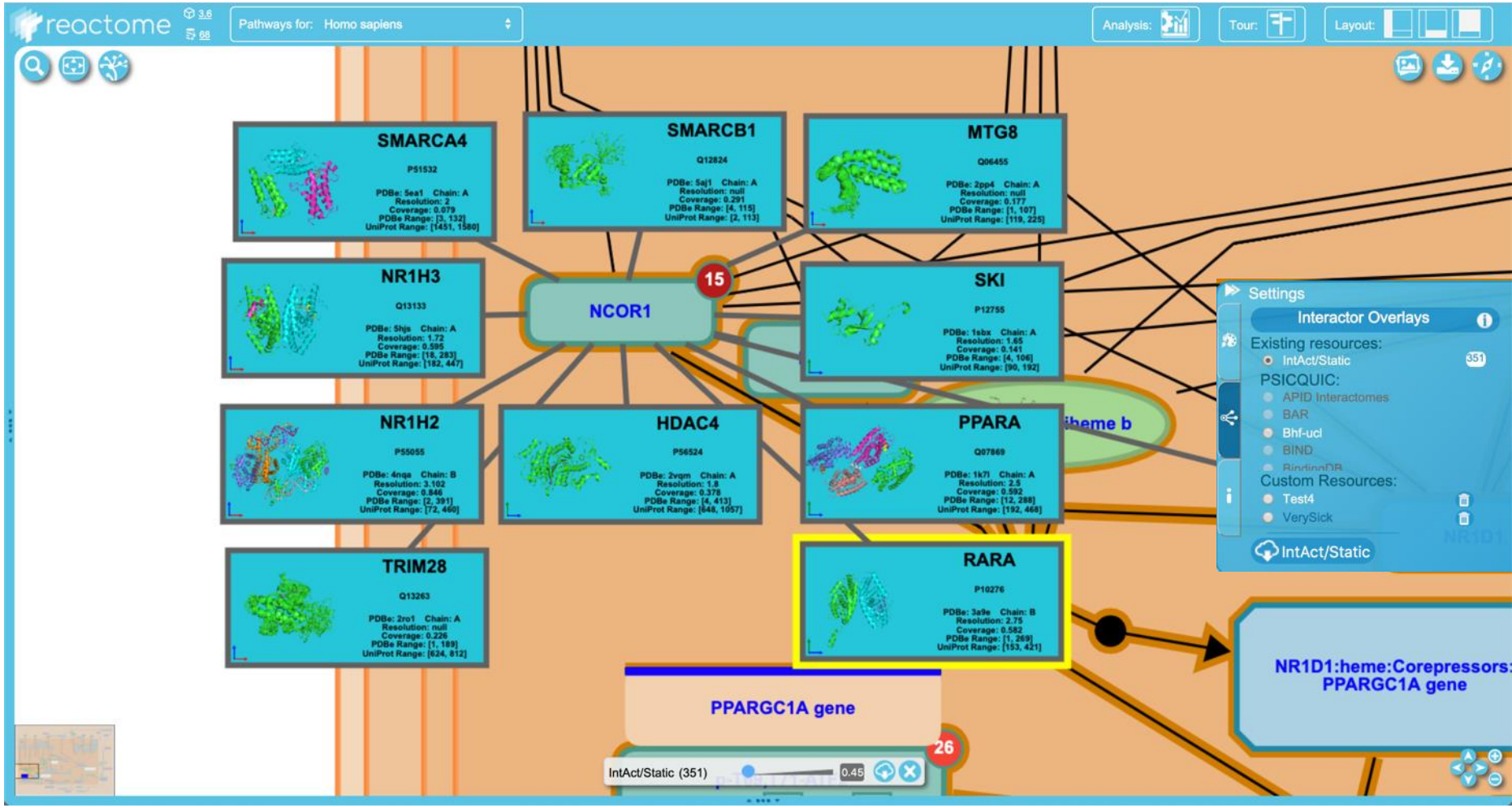


Pathway Analysis: Detail view

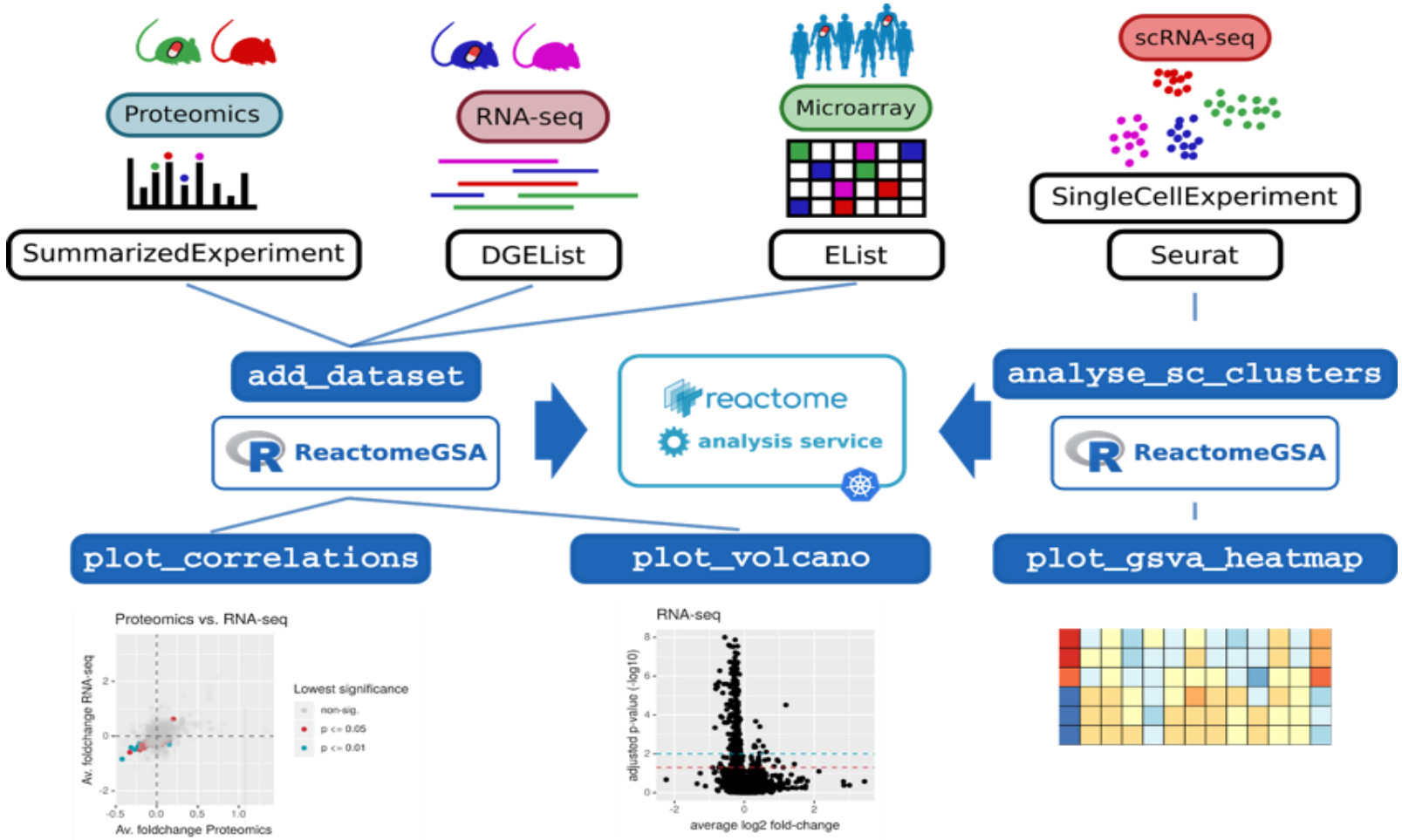


Inclusion of External Data: IntAct

- Medium to high confidence interactors from IntAct are included in the Reactome pathway database
- Where available, PDB structure pictograms are included in high zoom levels
- Many other resources available through PSICQUIC web services



ReactomeGSA - Quantitative Multi-Omics Analysis



Quantitative Analysis

ReactomeGSA

```r
# S4 request class
my_request <- new("ReactomeAnalysisRequest", method="Camera")

# directly import limma dataset
data("griss_melanoma_proteomics")
my_request <- add_dataset(request = my_request,
                          expression_values = griss_melanoma_proteomics,
                          name = "Proteomics",
                          type = "proteomics-int",
                          comparison_factor = "condition",
                          comparison_group_1 = "MOCK",
                          comparison_group_2 = "MCM",
                          additional_factors = c("cell.type", "patient.id"))

# directly import edgeR dataset
data("griss_melanoma_rnaseq")
my_request <- add_dataset(request = my_request,
                          expression_values = griss_melanoma_rnaseq,
                          name = "RNA-seq",
                          type = "rnaseq",
                          comparison_factor = "treatment",
                          comparison_group_1 = "MOCK",
                          comparison_group_2 = "MCM",
                          additional_factors = c("cell_type", "patient"))
```

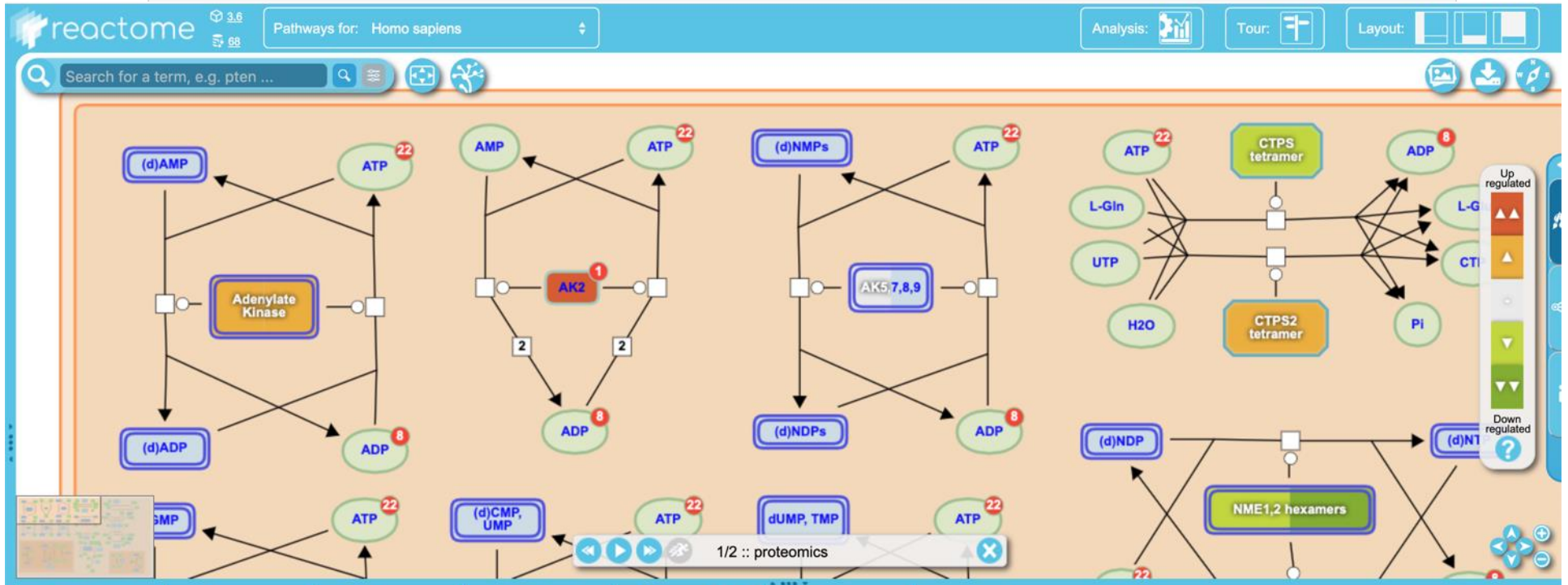
https://bioconductor.org/packages/release/bioc/html/ReactomeGSA.html

Combined Visualization

```
# perform the analysis
analysis_result <- perform_reactome_analysis(my_request)

# get the pathway statistics
pathway_result <- pathways(analysis_result)

# open the visualization
open_reactome(analysis_result)
```



Gsa Regulation analysis results for TOTAL [Data: Multi-sample analysis]

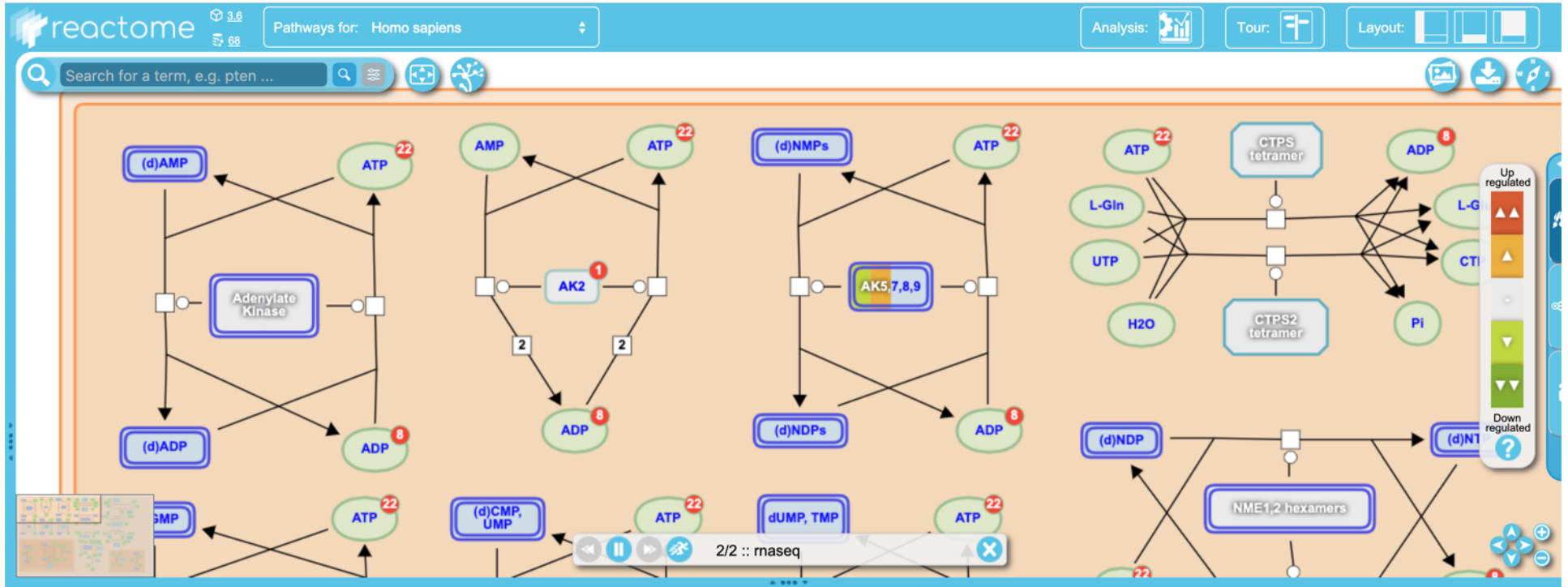
Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	proteomics	rnaseq	Species name
Propionyl-CoA catabolism	5	14	0.001	1.48E-2	7.4E-2	3	3	0	▲▲	▲	Homo sapiens
Signaling by FGFR	68	106	0.007	1.49E-2	7.44E-2	139	142	0.012	▼	▼	Homo sapiens
Mitochondrial protein import	57	69	0.005	1.49E-2	7.44E-2	14	14	0.001	▲▲	▼	Homo sapiens
Interferon gamma signaling	179	250	0.018	1.52E-2	7.53E-2	15	15	0.001	▲	▲	Homo sapiens
Intracellular metabolism of fatty acids regulates insulin secretion	3	9	0.001	1.6E-2	7.92E-2	2	2	0	▲	▼	Homo sapiens

Combined Visualization

```
# perform the analysis
analysis_result <- perform_reactome_analysis(my_request)

# get the pathway statistics
pathway_result <- pathways(analysis_result)

# open the visualization
open_reactome(analysis_result)
```

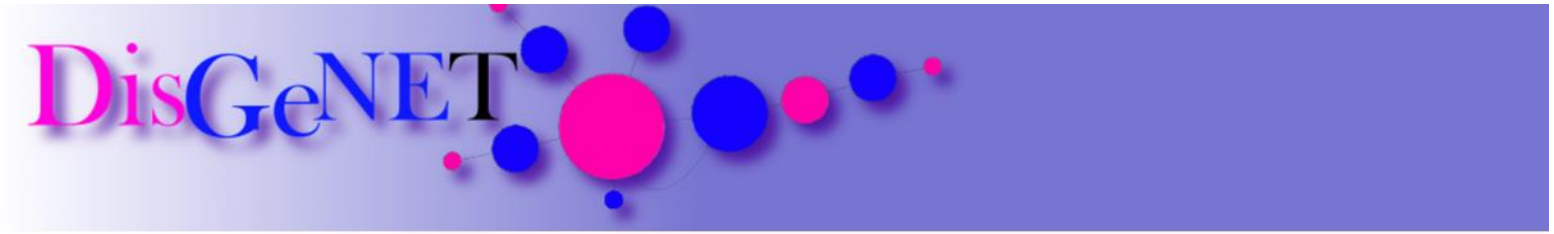


Gsa Regulation analysis results for TOTAL [Data: Multi-sample analysis]

Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total	Reactions ratio	proteomics	rnaseq	Species name
Propionyl-CoA catabolism	5	14	0.001	1.48E-2	7.4E-2	3	3	0	▲▲	▲	Homo sapiens
Signaling by FGFR	68	106	0.007	1.49E-2	7.44E-2	139	142	0.012	▼	▼	Homo sapiens
Mitochondrial protein import	57	69	0.005	1.49E-2	7.44E-2	14	14	0.001	▲▲	▼	Homo sapiens
Interferon gamma signaling	179	250	0.018	1.52E-2	7.53E-2	15	15	0.001	▲	▲	Homo sapiens
Intracellular metabolism of fatty acids regulates insulin secretion	3	9	0.001	1.6E-2	7.92E-2	2	2	0	▲	▼	Homo sapiens

External Binary Data Overlay: DisGeNET

- Many types of binary association data:
  - Disease-Gene, Drug-Target, Antibody-Antigen, …
- Example: DisGeNet



DisGeNET is a discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases (Piñero *et al.*, 2016; Piñero *et al.*, 2015). DisGeNET integrates data from expert curated repositories, GWAS catalogues, animal models and the scientific literature. DisGeNET data are homogeneously annotated with controlled vocabularies and community-driven ontologies. Additionally, several original metrics are provided to assist the prioritization of genotype–phenotype relationships.

	A	B	C	D	E	F	G	H	I
1	geneId	geneSymbol	DSI	DPI	diseaseId	diseaseName	diseaseType	diseaseClass	diseaseSemanticType
1421	3949	LDLR	0.475	0.828	C0519066	Acute Q fever	disease	C01	Disease or Syndrome
1422	4047	LSS	0.676	0.448	C0519066	Acute Q fever	disease	C01	Disease or Syndrome
1423	10211	FLOT1	0.727	0.414	C0519066	Acute Q fever	disease	C01	Disease or Syndrome
1424	1152	CKB	0.636	0.483	C0857501	Acute schizophrenia	disease		Mental or Behavioral Dysfunc
1425	3458	IFNG	0.323	0.897	C0857501	Acute schizophrenia	disease		Mental or Behavioral Dysfunc
1426	7167	TPI1	0.564	0.655	C0857501	Acute schizophrenia	disease		Mental or Behavioral Dysfunc
1427	4297	KMT2A	0.46	0.793	C0280141	Acute Undifferentiated Leukemia	disease		Neoplastic Process
1428	283	ANG	0.547	0.69	C0001349	Acute-Phase Reaction	phenotype	C23	Pathologic Function
1429	28976	ACAD9	0.727	0.345	C1970173	Acyl-CoA Dehydrogenase Family, Member 9, Deficien	disease	C05;C10;C14;C1	Disease or Syndrome
1430	673	BRAF	0.352	0.793	C0431129	Adamantinous Craniopharyngioma	disease	C04	Neoplastic Process

DisGeNET overlay of gene-disease associations

DisGeNET (Pinero J, et al, Nucleic Acids Res. 2019) is a database of gene-disease associations. We have pre-processed DisGeNET curated gene-disease associations (Release v7.0) for overlay onto Reactome. For each disease, clicking on the "Analysis" button will show the results of Reactome pathway analysis with the set of genes associated with that disease. If you are interested in overlaying other data sources onto Reactome in a similar manner, please contact us.

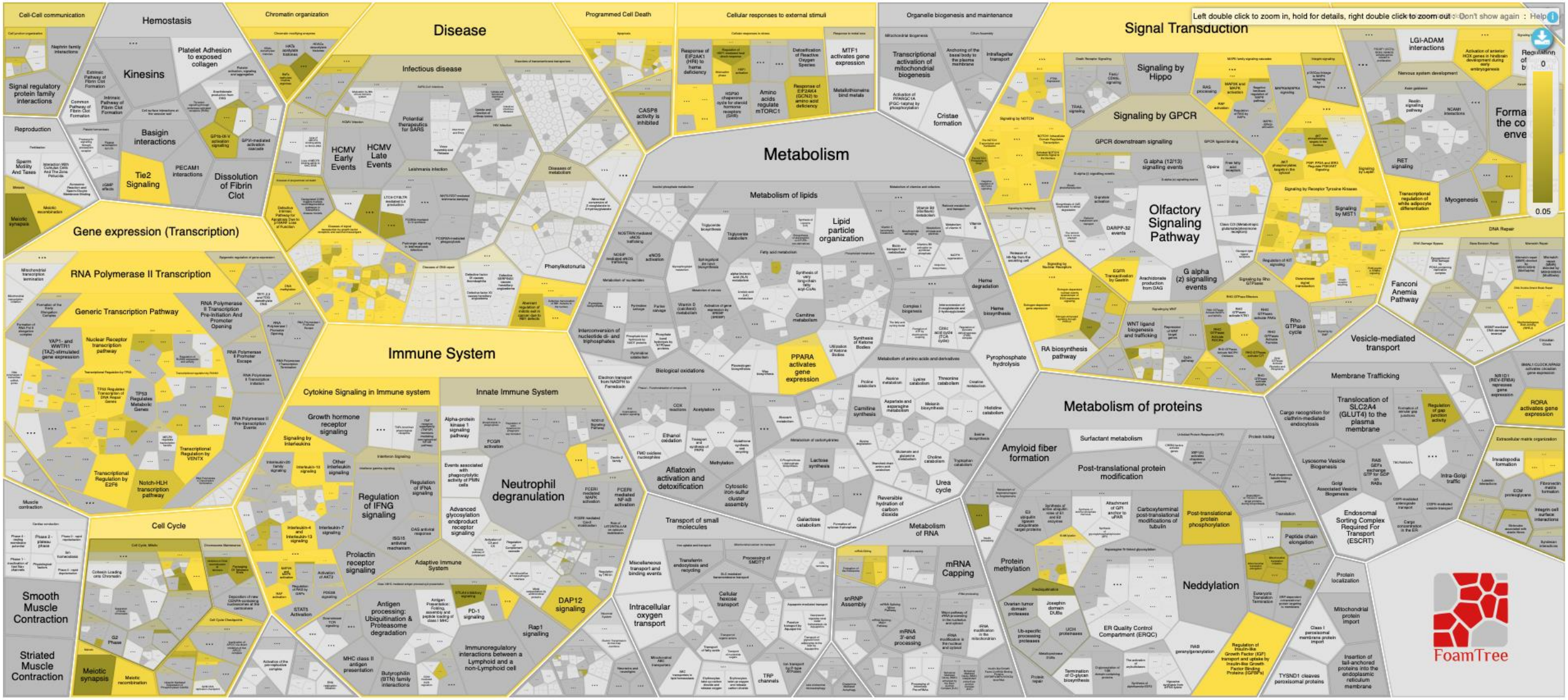
Parameters

Parameter	Option	Description
Minimum number of genes per disease	10	The Reactome pathway analysis presented here only makes sense for multiple genes. Diseases with less gene associations than this parameter are not displayed in the table. For small numbers of genes, it might be helpful to switch from "pValue" to "coverage" at the bottom of the visualisation.
Score	◉ Low (no filter) ○ Medium (> 0.33) ○ High (> 0.66)	Score of the gene-disease association as provided by DisGeNET.
Include interactors	☐	Include high confidence interactors from IntAct in Reactome.
Default result view	◉ ReacFoam ○ Fireworks	Reactome provides two different options for the first view of the analysis results. Choose your preference.
Reset the filters	Reset	Clear disease name filter, and reset to default values.

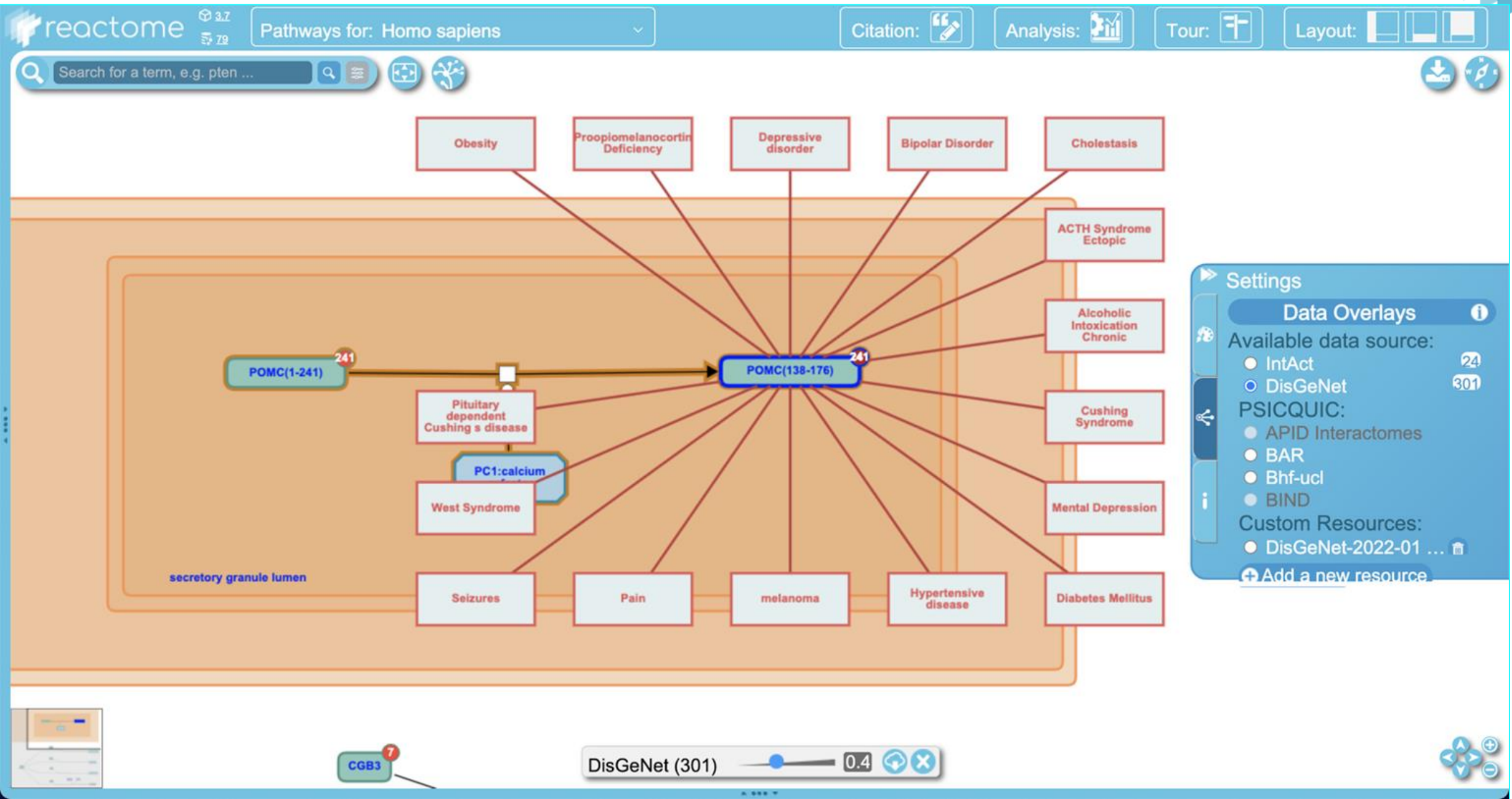
Overlay Table

Check in Pathway Browser	Disease name (Disease name filter)	Number of genes	Gene list	Disease id
Analysis	Malignant_Neoplasm_Of_Breast	1074	ABCA3, ABCA4, ABCB1, ABCB10, ABCB6, ABCB8, ABCC1, ABCG2, ABHD12B, ABL1, ABRAXAS1, ACADM, ACAP1, ACCS, ACHE, ACO2, ACTA2, ACVR1, ACY1, ADAM10, ADAM12, ADAM33, ADAMTS1, ADAMTS19, ADAR, ADAT3, ADGRF4,	C0006142
Analysis	Schizophrenia	883	A1BG, ABCA1, ABCA13, ABCB1, ACE, ACHE, ACOT6, ACP1, ACSL6, ACSM1, ACTB, ACTR2, ADAM12, ADAMTS12, ADAMTS3, ADAMTSL3, ADARB1, ADCY7, ADCYAP1, ADCYAP1R1, ADGRF4, ADK, ADM, ADNP, ADNP2, ADORA1,	C0036341
Analysis	Liver_Cirrhosis_Experimental	774	A2M, AADAC, AADAT, ABAT, ABCB1, ABCC2, ABCC5, ABCG1, ABCG5, ABCG8, ABLIM3, ACAT1, ACBD4, ACHE, ACKR3, ACOT9, ACOX2, ACP5, ACSL1, ACTA2, ACTN1, ADA, ADAM17, ADD3, ADGRE1, ADH4, ADHFE1, ADIPOQ, ADM,	C0023893
Analysis	Colorectal_Carcinoma	702	ABCA1, ABCA10, ABCA12, ABCA13, ABCA3, ABCA4, ABCA5, ABCA6, ABCA8, ABCA9, ABCB1, ABCB10, ABCB11, ABCB4, ABCB5, ABCB6, ABCC1, ABCC13, ABCC2, ABCC3, ABCC4, ABCC5, ABCC6, ABCC8, ABCD2, ABCD3, ABCD4,	C0009402
Analysis	Prostatic_Neoplasms	616	AAAS, ABCC4, ABCG5, ABO, ABR, ACE, ACHE, ACRBP, ACSL4, ACSM3, ADAM28, ADAM9, ADAMTS8, ADI1, ADRB2, AGR2, AHCYL2, AHR, AKAP13, AKR1C3, AKT1, ALAD, ALDH1A2, ALOX12B, ALOX5, ALOXE3, AMACR, ANTXR2, ANXA1, ANXA3,	C0033578
Analysis	Malignant_Neoplasm_Of_Prostate	616	AAAS, ABCC4, ABCG5, ABO, ABR, ACE, ACHE, ACRBP, ACSL4, ACSM3, ADAM28, ADAM9, ADAMTS8, ADI1, ADRB2, AGR2, AHCYL2, AHR, AKAP13, AKR1C3, AKT1, ALAD, ALDH1A2, ALOX12B, ALOX5, ALOXE3, AMACR, ANTXR2, ANXA1, ANXA3,	C0376358
Analysis	Breast_Carcinoma	538	ABCB1, ABCC1, ABCG2, ABL1, ACHE, ACTA2, ACVR1, ADAM10, ADAM33, ADAMTS1, ADAR, AFP, AGR2, AHR, AKAP12, AKT1, AKT2, ALDOA, ALK, ALKBH8, ANGPTL4, ANKRD34A, APC2, APOBEC3A, APOBEC3B, APRT, AR, ARAF, AREG,	C0678222
Analysis	Mammary_Neoplasms	527	ABCB1, ABCC1, ABCG2, ABL1, ACHE, ACTA2, ACVR1, ADAM10, ADAM33, ADAMTS1, ADAR, AFP, AGR2, AHR, AKAP12, AKT1, AKT2, ALDOA, ALK, ALKBH8, ANGPTL4, ANKRD34A, APC2, APOBEC3A, APOBEC3B, APRT, AR, ARAF, AREG,	C1458155
Analysis	Mammary_Neoplasms_Human	525	ABCB1, ABCC1, ABCG2, ABL1, ACHE, ACTA2, ACVR1, ADAM10, ADAM33, ADAMTS1, ADAR, AFP, AGR2, AHR, AKAP12, AKT1, AKT2, ALDOA, ALK, ALKBH8,	C1257931

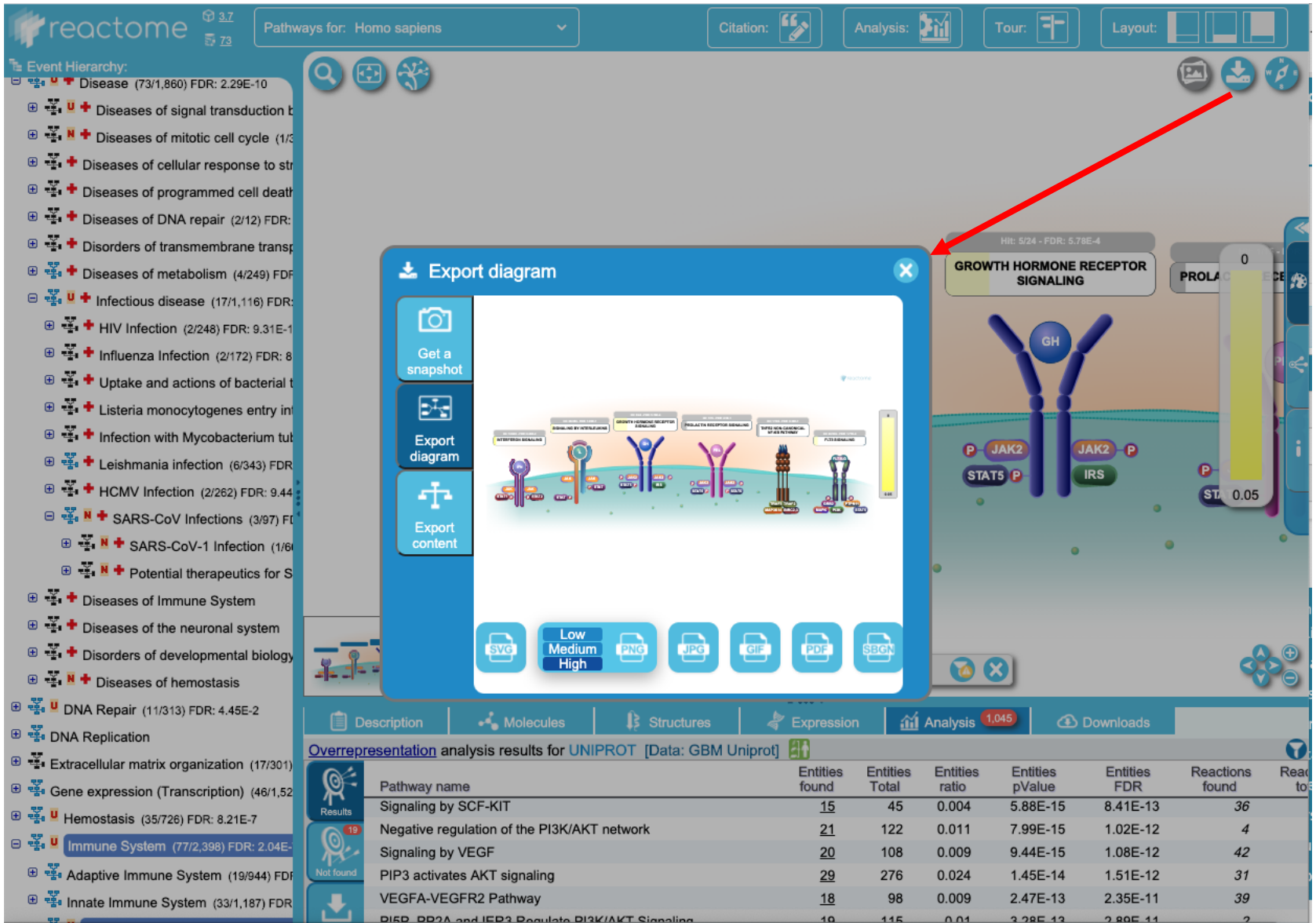
Enrichment of DisGeNET “Breast Carcinoma” associated genes



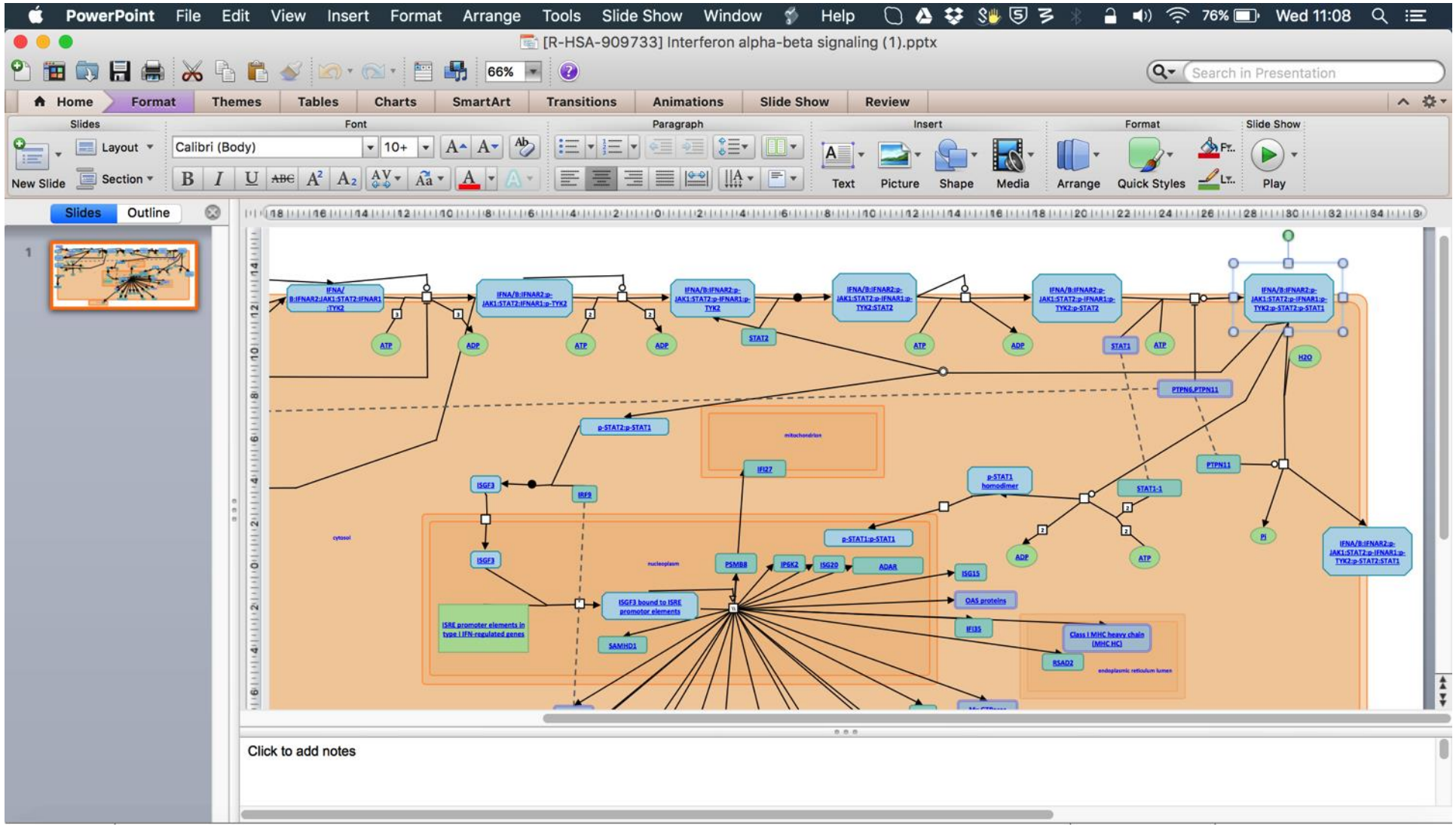
Binary Disease Relation Overlay



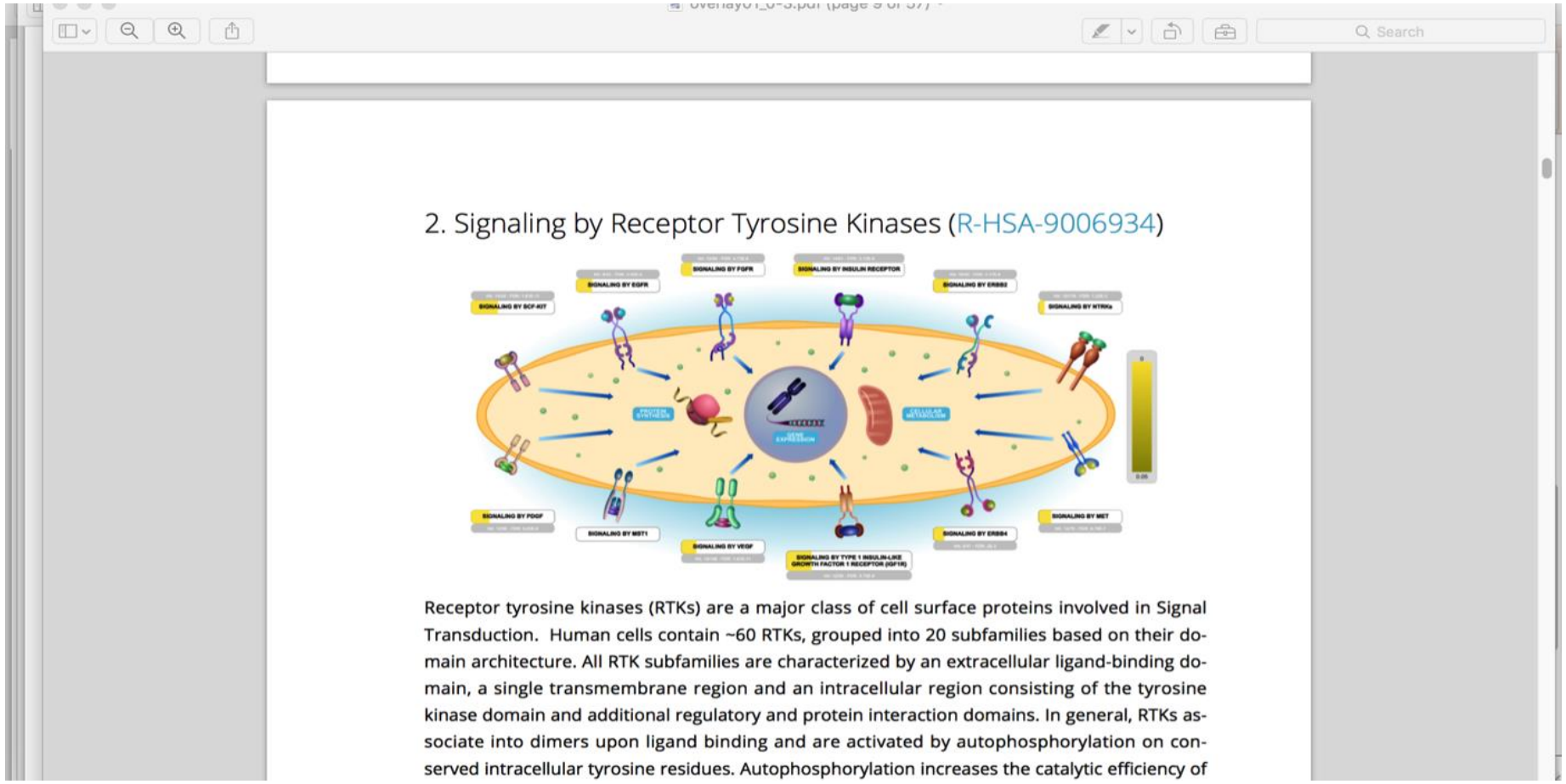
Download as PNG or SVG



Download as pptx



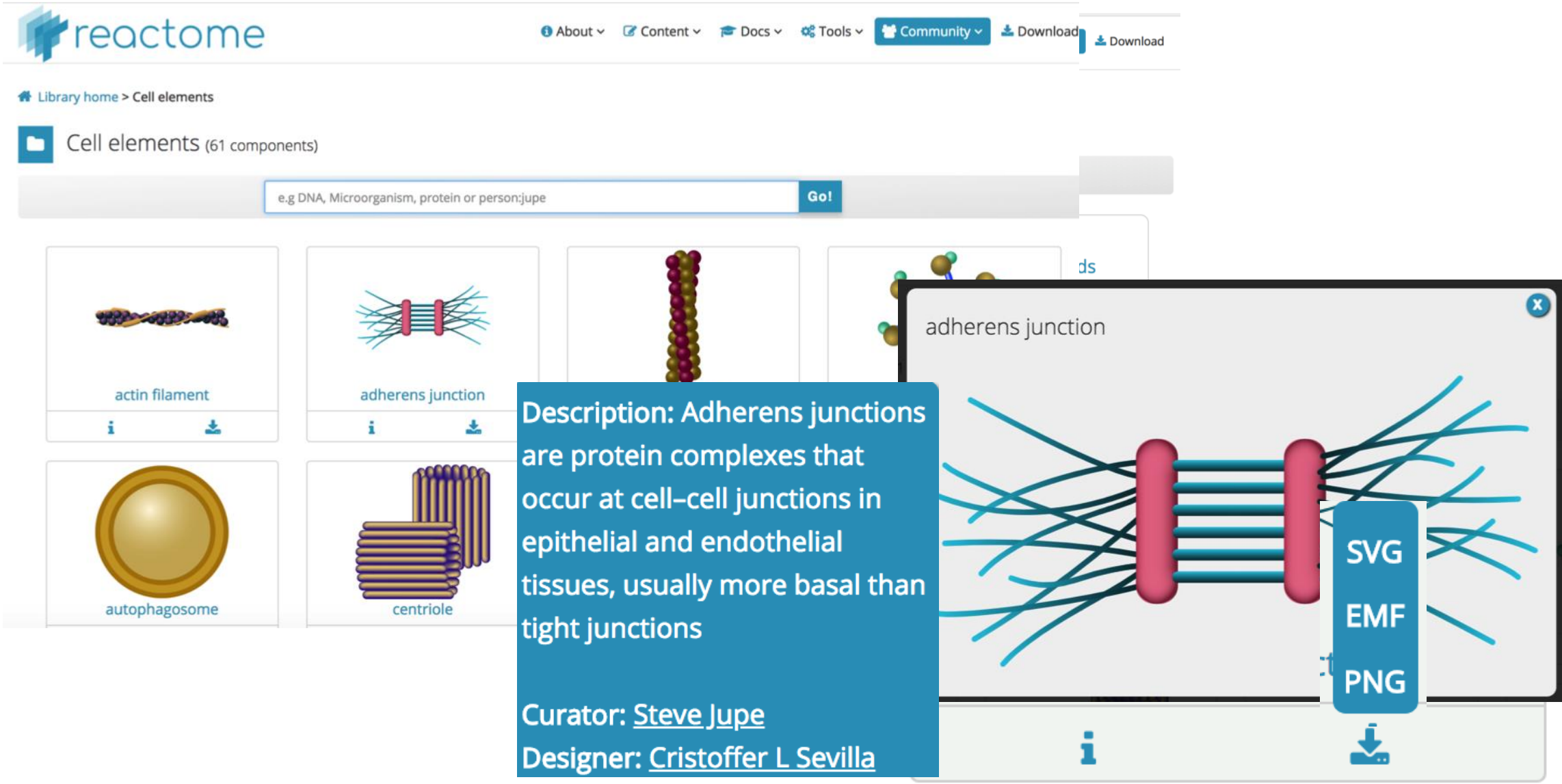
Pathway Analysis: PDF Export



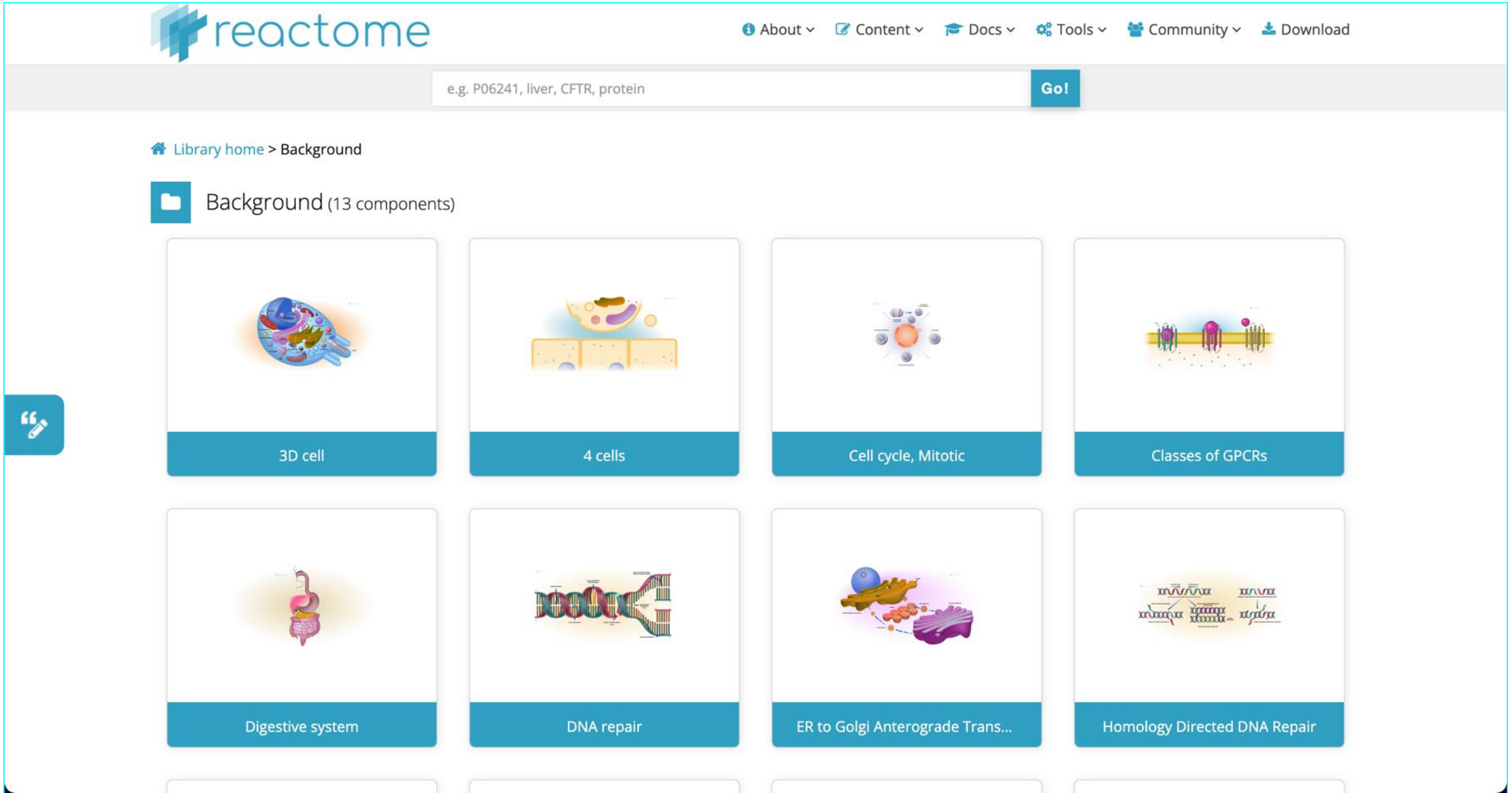
2. Signaling by Receptor Tyrosine Kinases (R-HSA-9006934)

Receptor tyrosine kinases (RTKs) are a major class of cell surface proteins involved in Signal Transduction. Human cells contain ~60 RTKs, grouped into 20 subfamilies based on their domain architecture. All RTK subfamilies are characterized by an extracellular ligand-binding domain, a single transmembrane region and an intracellular region consisting of the tyrosine kinase domain and additional regulatory and protein interaction domains. In general, RTKs associate into dimers upon ligand binding and are activated by autophosphorylation on conserved intracellular tyrosine residues. Autophosphorylation increases the catalytic efficiency of

Reactome Icon Library



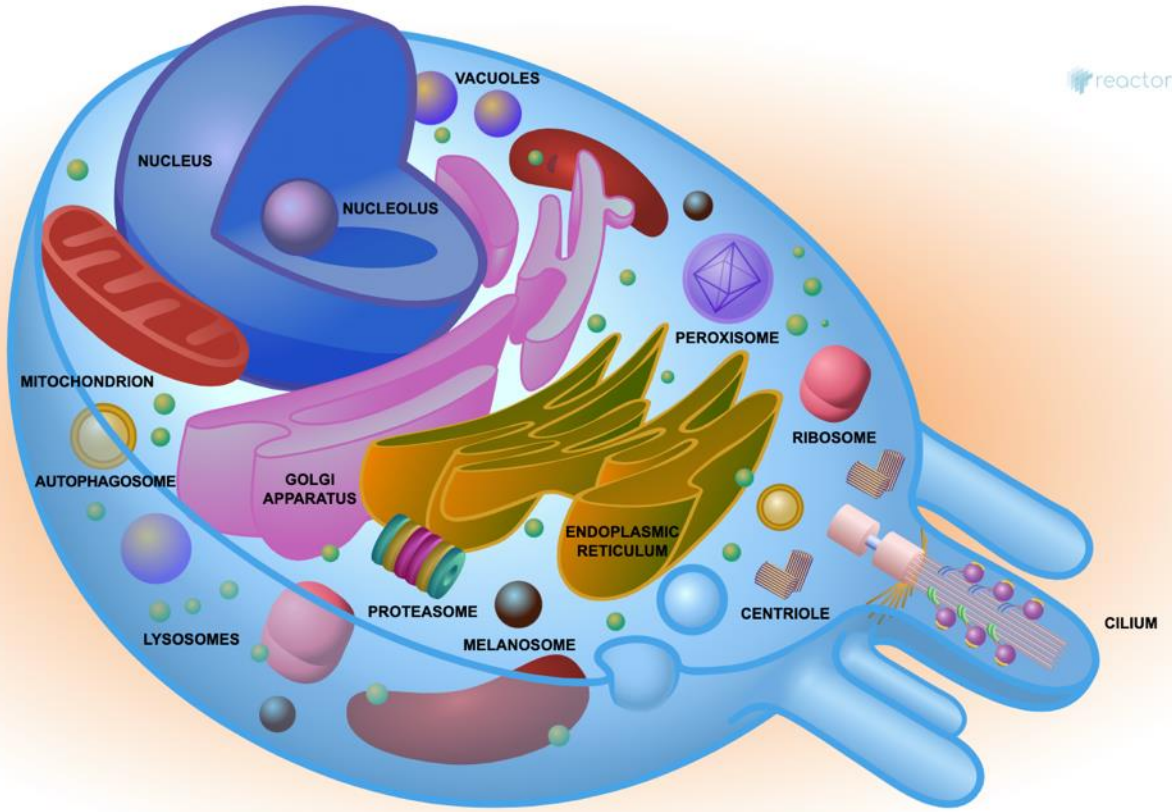
New: Backgrounds



3D cell

Categories	Background
Curator	Karen Rothfels
Designer	Cristoffer Sevilla

Icon preview



SVG

PNG

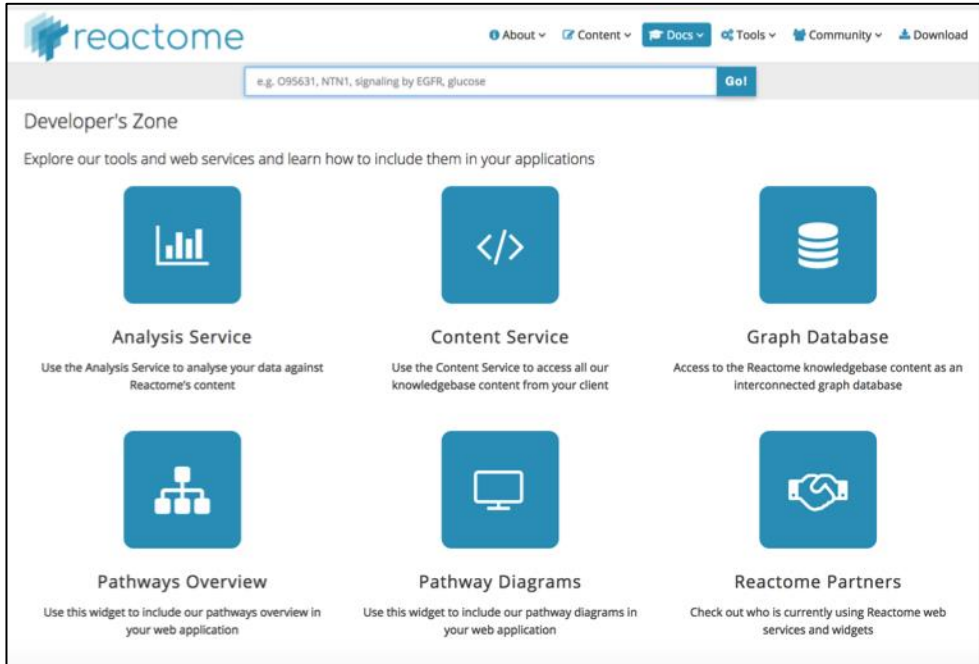
EMF

Reactome is Open and Documented

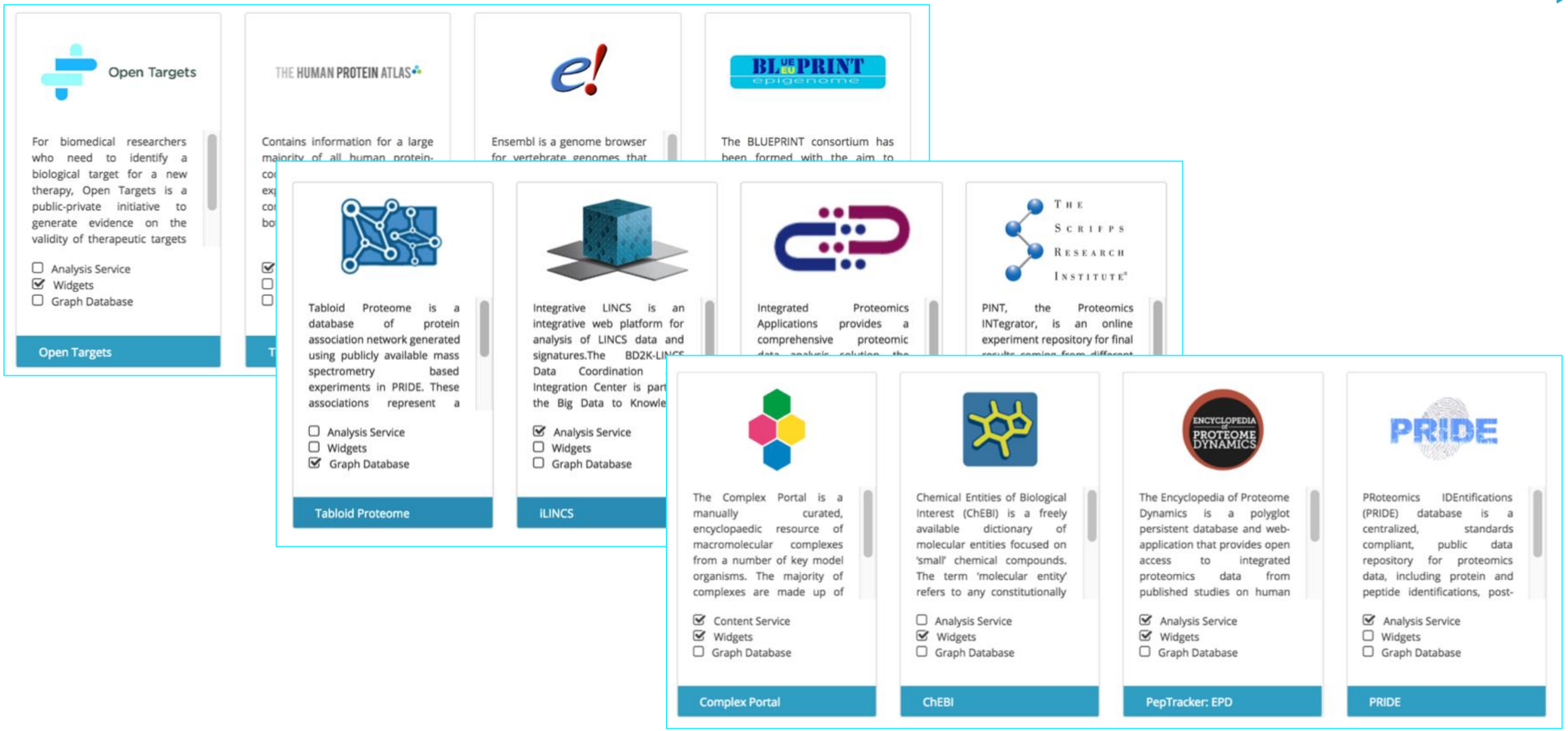
- Reactome is open source, open data, “open graphics”
- Most components CC-BY
- Recently released content under CC-0 for Wikidata integration

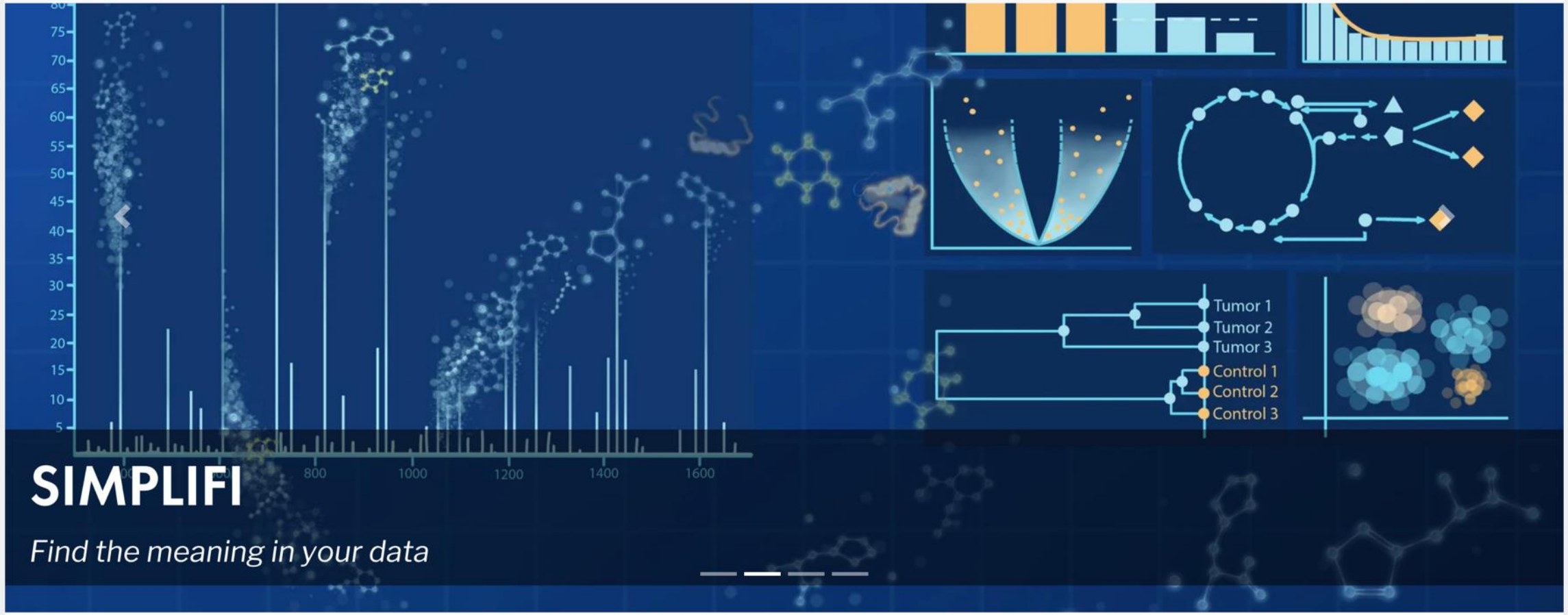
Reactome is Open and Documented

- Reactome is open source, open data, “open graphics”
- Most components CC-BY
- Recently released content under CC-0 for Wikidata integration
- Third parties can integrate Reactome components:
  - http://www.reactome.org/pages/documentation/developer-guide/



Use of Reactome software by 3rd parties





What is ProtiFi?

ProtiFi, LLC was founded to understand life beyond genes. In contrast to static DNA, which is the same in every organ, the structures and machinery of life are dynamic and change with age or in conditions of health and disease. ProtiFi develops and provides innovative solutions to study proteins, metabolites, lipids and glycans. From sample

SimpliFi

p-value < 0.05

log2 fold change > 0 up/down

≤ 25 % missing values

sensitivity ≥ 0 %

specificity ≥ 0 %

Text search

Only include favorite analytes

Found 782 of 5411 analytes.

Sort by p-value

Sort order Normal

Fold change format log2

Analytes | Pathways | GO Terms | Cell Markers | Volcano Plot | Venn Diagram

Filters | p-value < 0.05 | Text search | Only show disease-associated pathwa

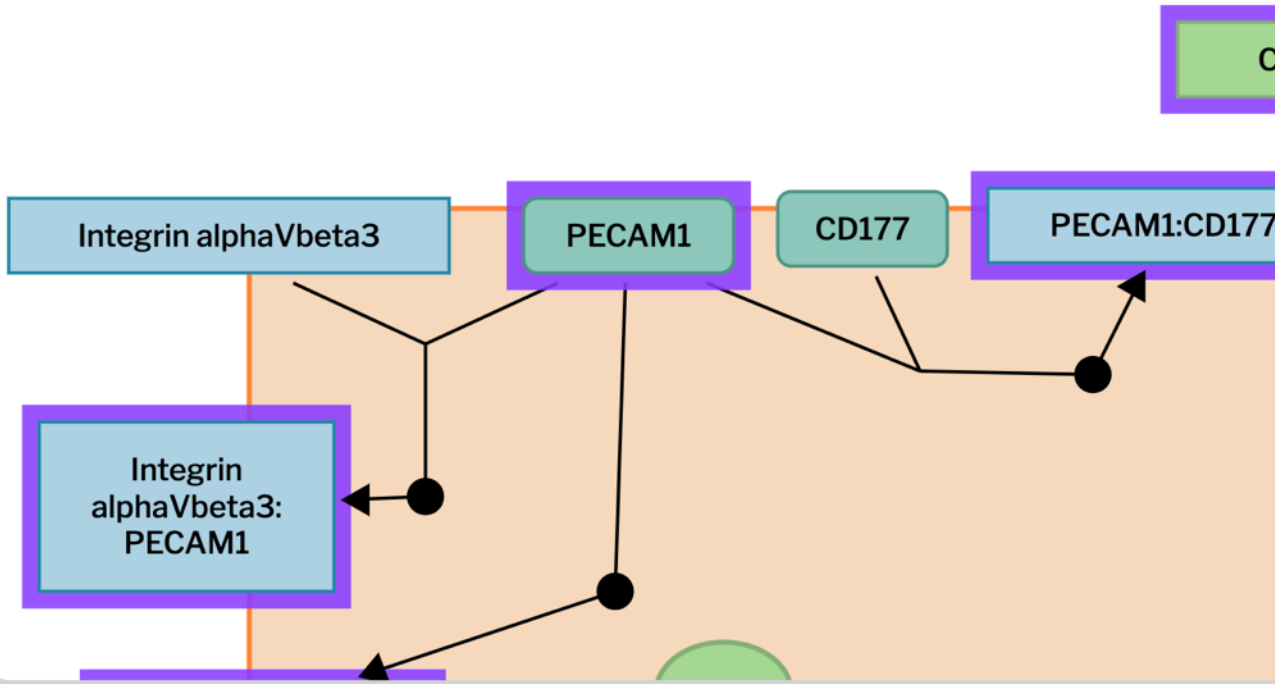
R-HSA-6798695 Neutrophil degranulation	2.1602e-69 53 out of 480
R-HSA-6791226 Major pathway of rRNA processing in the nucleolus and cytosol	4.7494e-47 31 out of 183
R-HSA-381426 Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	1.9134e-28 19 out of 124
R-HSA-202733 Cell surface interactions at the vascular wall	3.5695e-26 20 out of 196
R-HSA-1268020 Mitochondrial protein import	6.92e-26 16 out of 83
R-HSA-8957275 Post-translational protein phosphorylation	5.4973e-24 16 out of 107
R-HSA-72764 Eukaryotic Translation Termination	8.351e-23 15 out of 97
R-HSA-198933 Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	1.6406e-22 20 out of 297
R-HSA-6790901 rRNA modification in the nucleus and cytosol	5.2134e-22 13 out of 60

R-HSA-202733

Cell surface interactions at the vascular wall

Found 20 of 196 analytes.
Hypergeometric p-value: 3.56

GSEA p=0 Score



New Bucket | Contains 0 analytes. | Filter OFF

Acknowledgements

Alphafold DB

- Sameer Velankar, EMBL-EBI
- PDBe team
- AlphaFold team

Reactome

- Lincoln Stein, OICR
- Peter D'Eustachio, NYU
- Guanming Wu, OHSU
- Reactome team

- US NIH P41 HG003751
- US NIH BD2K 1U54GM114833-01
- European Molecular Biology Laboratory
- EU Horizon 2020 grant 788042
- FWF-Austrian Science Fund (P30325-B28)
- Open Targets
- DeepMind