廣州醫科大學
GUANGZHOU MEDICAL UNIVERSITY

Gene Set Analysis –Methods and Tools

Antonio Mora, Ph.D.

(Wechat: antoniocmora)

www.moralab.science

21.09.2020

MORA LAB
FUNCTIONAL BIOINFORMATICS

Contents

Contents

Why Databases?

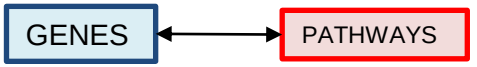
Databases are sources of Biological Annotation

What is Annotation?

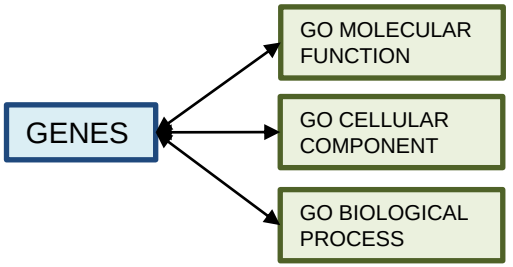
Annotation is structured information regarding a gene or other biomolecule

Three types of Databases are used in Gene Set Analysis

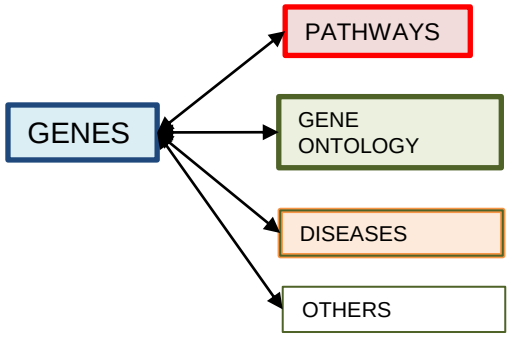
Pathway databases
(KEGG, Reactome, Wikipathways)



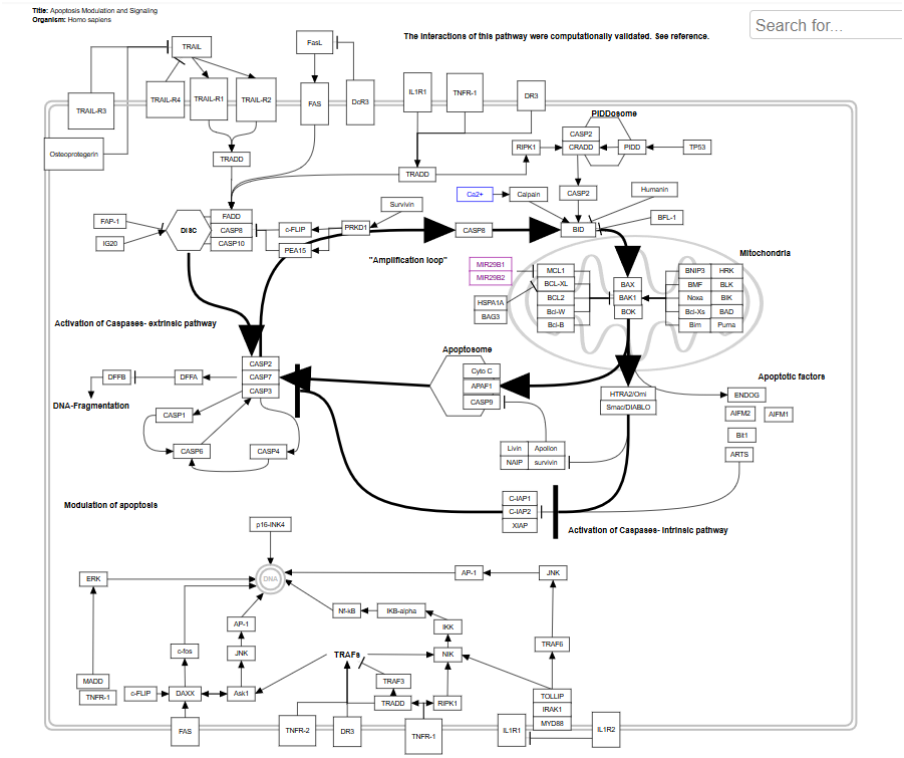
Ontology databases
(Gene Ontology)



Gene set databases
(GeneSetDB, MSigDB)



Pathways are chains of reactions



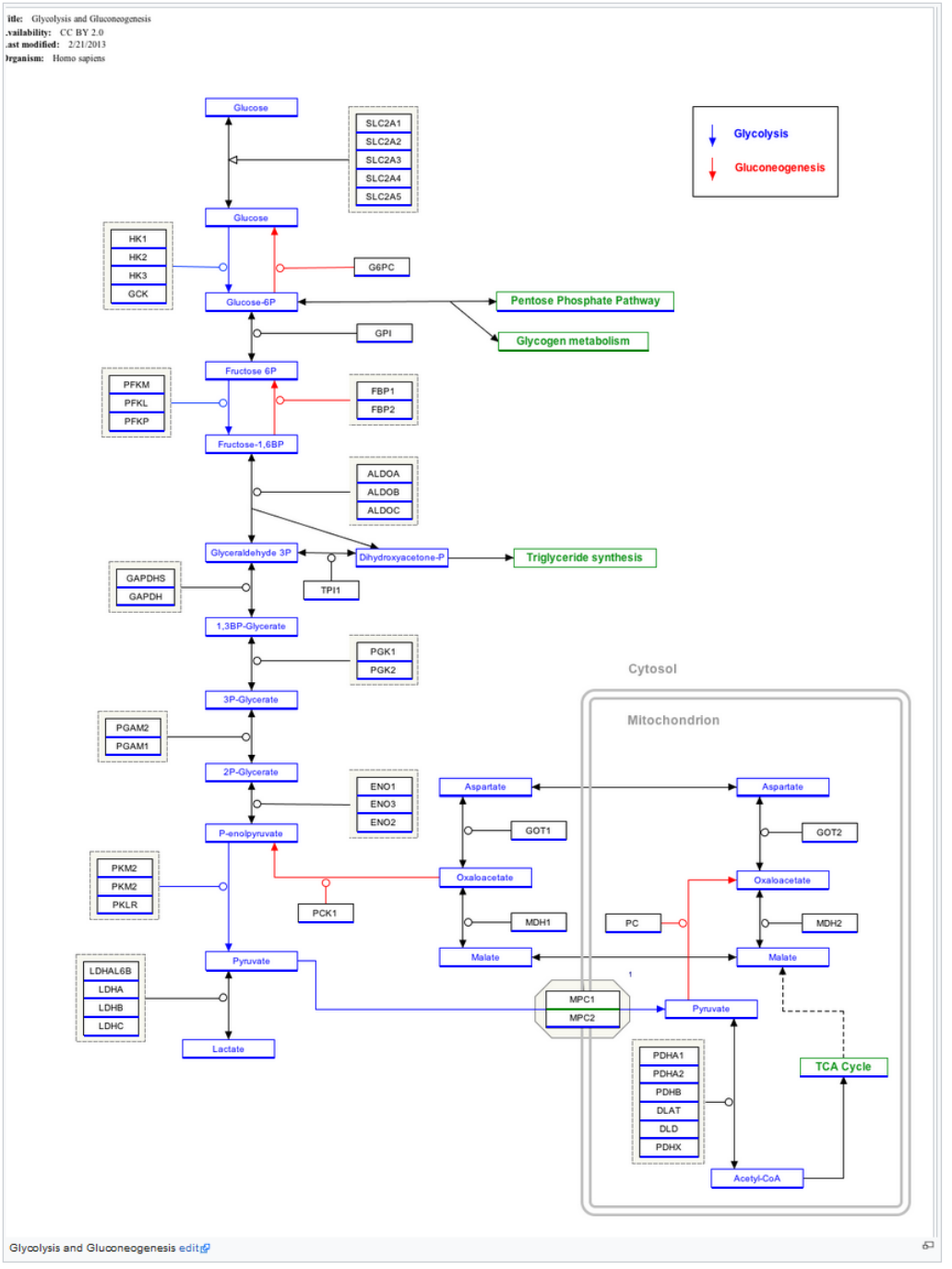
http://wikipathways.org/index.php/Pathway:WP1772

A ***biological pathway*** is a **chain of interactions or chemical reactions** among molecules that leads to one or more products.

Pathways are chains of reactions

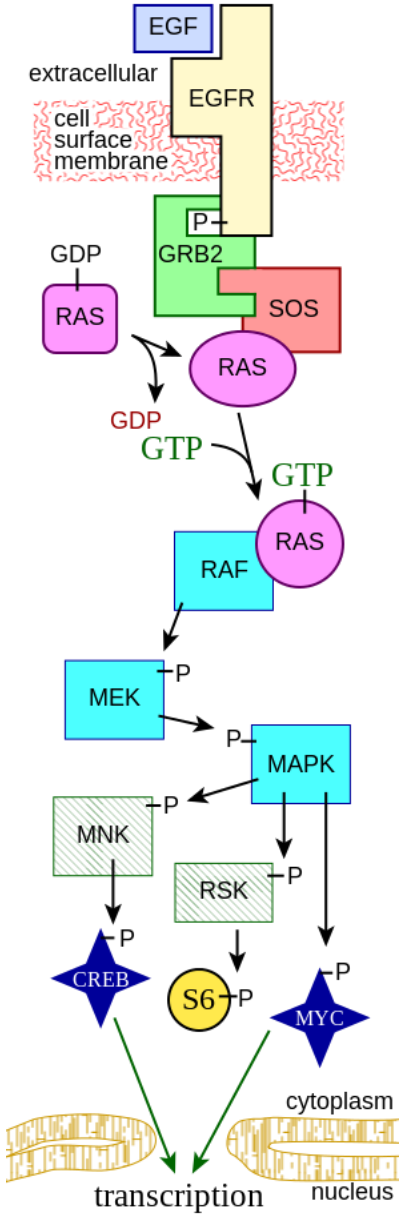
The most studied types of biological pathways are: Metabolic pathways, signal transduction pathways, and gene regulation pathways.

Metabolic pathways: Glycolysis

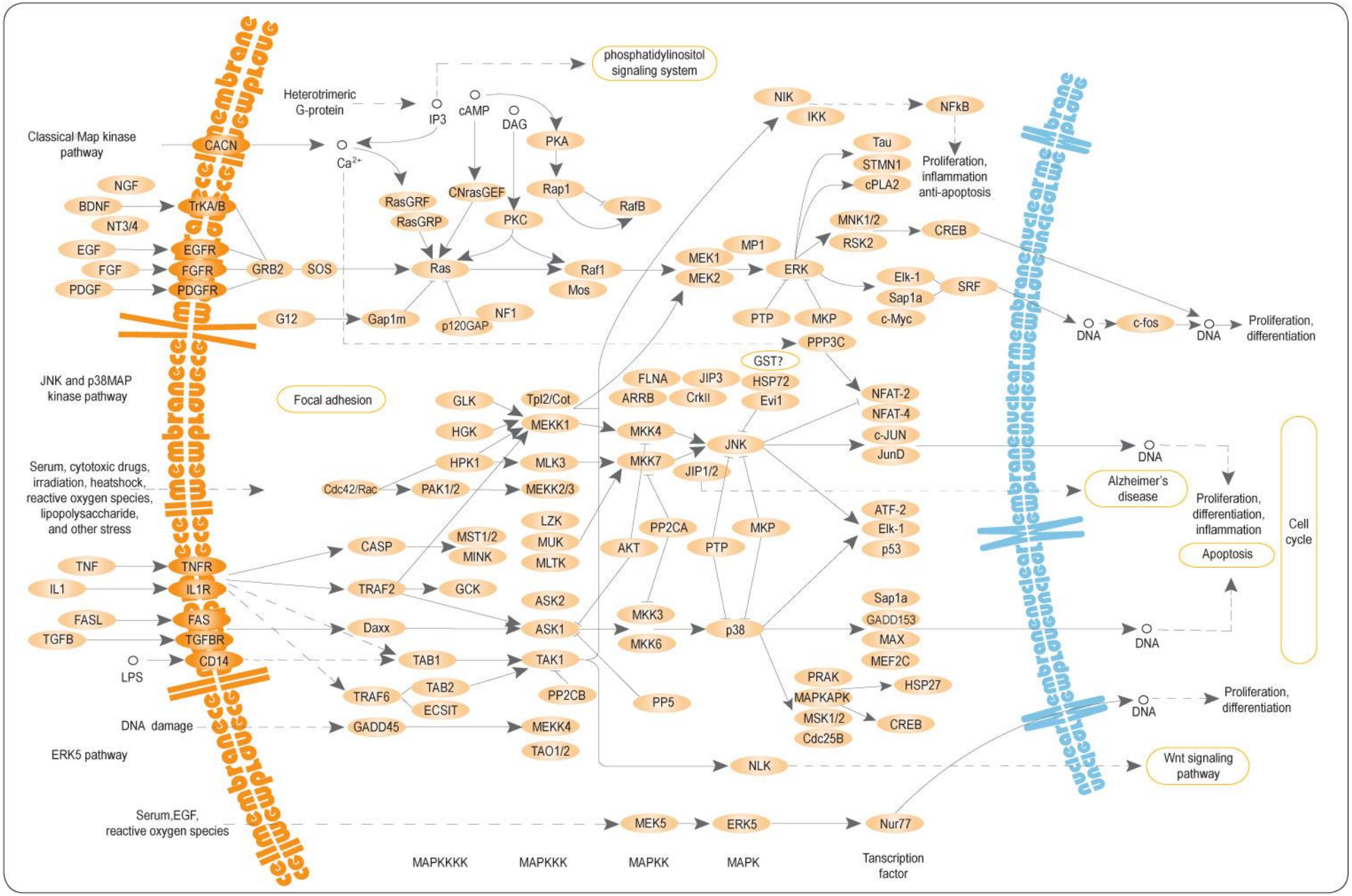


https://en.wikipedia.org/wiki/Glycolysis

Signaling pathways: MAPK/ERK pathway



https://en.wikipedia.org/wiki/MAPK/ERK_pathway



https://en.wikipedia.org/wiki/MAPK/ERK_pathway

Pathways are chains of reactions

A precise definition of a pathway?: The definition of a pathway is a little subjective. Three problems:

- Where to start and where to end
- Level of detail (intermediate reactions)
- Pathway cross-talk

Therefore, pathways may look slightly different according to the source.

Ontologies are a summary of all the concepts in a field and their relationships

Example: An ontology of the **finger**

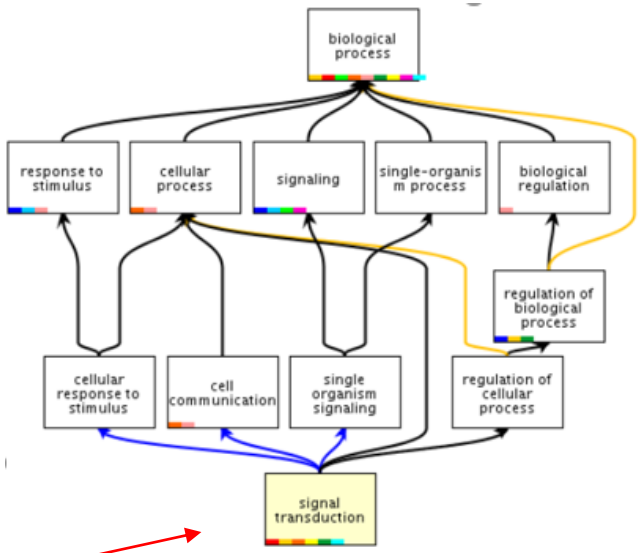
HUMAN BODY

- HEAD
- TRUNK
- EXTREMITIES
  - SUPERIOR
    - Hand
      - Palm
      - **Fingers**
    - Wrist
    - Forearm
    - Arm
    - Shoulder
  - INFERIOR

A finger IS PART OF a hand. A hand IS PART OF a superior extremity, which IS A extremity, which IS PART OF the human body

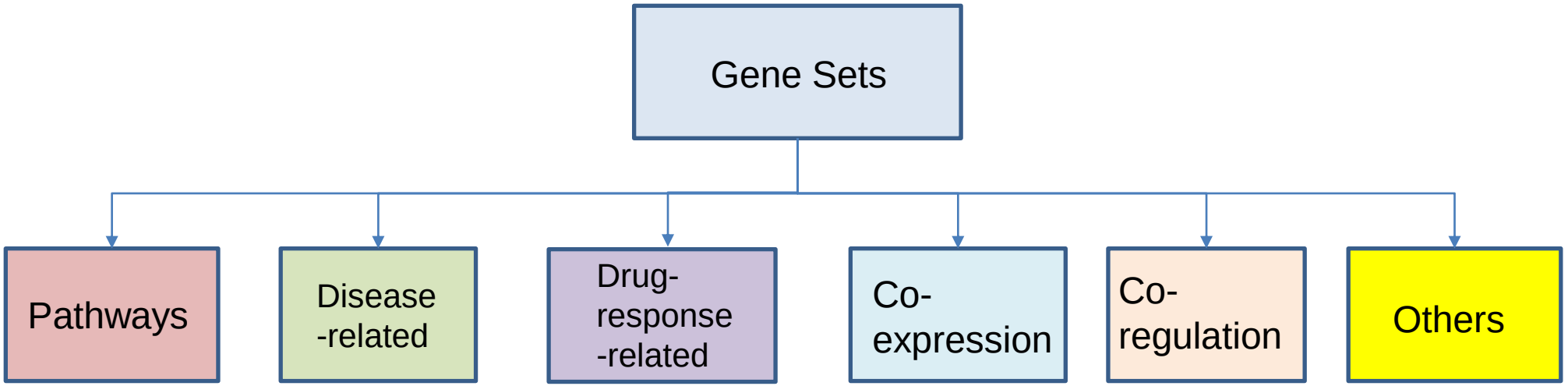
Ontologies are a summary of all the concepts in a field and their relationships

An ***ontology*** is a way of organizing the knowledge in a field. Knowledge is organized in terms of all of the concepts involved, and the relationships between them (IS A, IS PART OF, and others).



Example: The place of the concept «Signal transduction» inside an ontology graph and an ontology tree

Gene sets are... any set of genes



Essentially, any set of genes that can be grouped for some reason.

Contents

How many Pathway databases are out there...?

pathguide.org

Home | BioPAX | cBio | MSKCC

Pathguide » the pathway resource list

Navigation
- Protein-Protein Interactions
- Metabolic Pathways
- Signaling Pathways
- Pathway Diagrams
- Transcription Factors / Gene Regulatory Networks
- Protein-Compound Interactions
- Genetic Interaction Networks
- Protein Sequence Focused
- Other

Search
Organisms: All
Availability: All
Standards: All
Reset | Search

Analysis
- Statistics
- Database Interactions

Contact
Comments, Questions, Suggestions are Always Welcome!

Complete Listing of All Pathguide Resources

Pathguide contains information about **688** biological pathway related resources and molecular interaction related resources. Click on a link to go to the resource home page or 'Details' for a description page. Databases that are free and those supporting BioPAX, CellML, PSI-MI or SBML standards are respectively indicated.

If you know of a pathway resource that is not listed here, or have other questions or comments, please send us an e-mail.

News

Major new update of Pathguide August 2013
We now have information about ~550 resources!.

Visual navigation added
May 2010
Click the 'Database interactions' link on the left

Protein-Protein Interactions

Database Name (Order: alphabetically | by web popularity)	Full Record	Availability	Standards
2P2Idb - The Protein-Protein Interaction Inhibition Database	Details	Free	
3D-Interologs - 3D-Interologs	Details	Free	
3DID - 3D interacting domains	Details	Free	
ACSN - Atlas of Cancer Signalling Network	Details	Free	BioPAX
ADAN - Prediction of protein-protein interaction of modular domains	Details	X	
AHD2.0 - Arabidopsis Hormone Database 2.0	Details	Free	
AllFuse - Functional Associations of Proteins in Complete Genomes	Details	X	
aMAZE - Protein Function and Biochemical Pathways Project	Details	X	
ANAP - Arabidopsis Network Analysis Pipeline	Details	Free	
ANIA - ANnotation and Integrated Analysis of the 14-3-3 interactome	Details	Free	
AnimalTFDB - Animal Transcription Factor Database	Details	Free	
AntiJen - AntiJen a Kinetic, Thermodynamic and Cellular Database	Details	Free	
APID - Agile Protein Interactomes DataServer	Details	Free	PSI-MI
ARN - The Autophagy Regulatory Network	Details	X	PSI-MI BioPAX SBML
AS-ALPS - Alternative Splicing - induced ALteration of Protein Structure	Details	Free	
ASD - Allosteric Database	Details	Free	

KEGG

Reactome

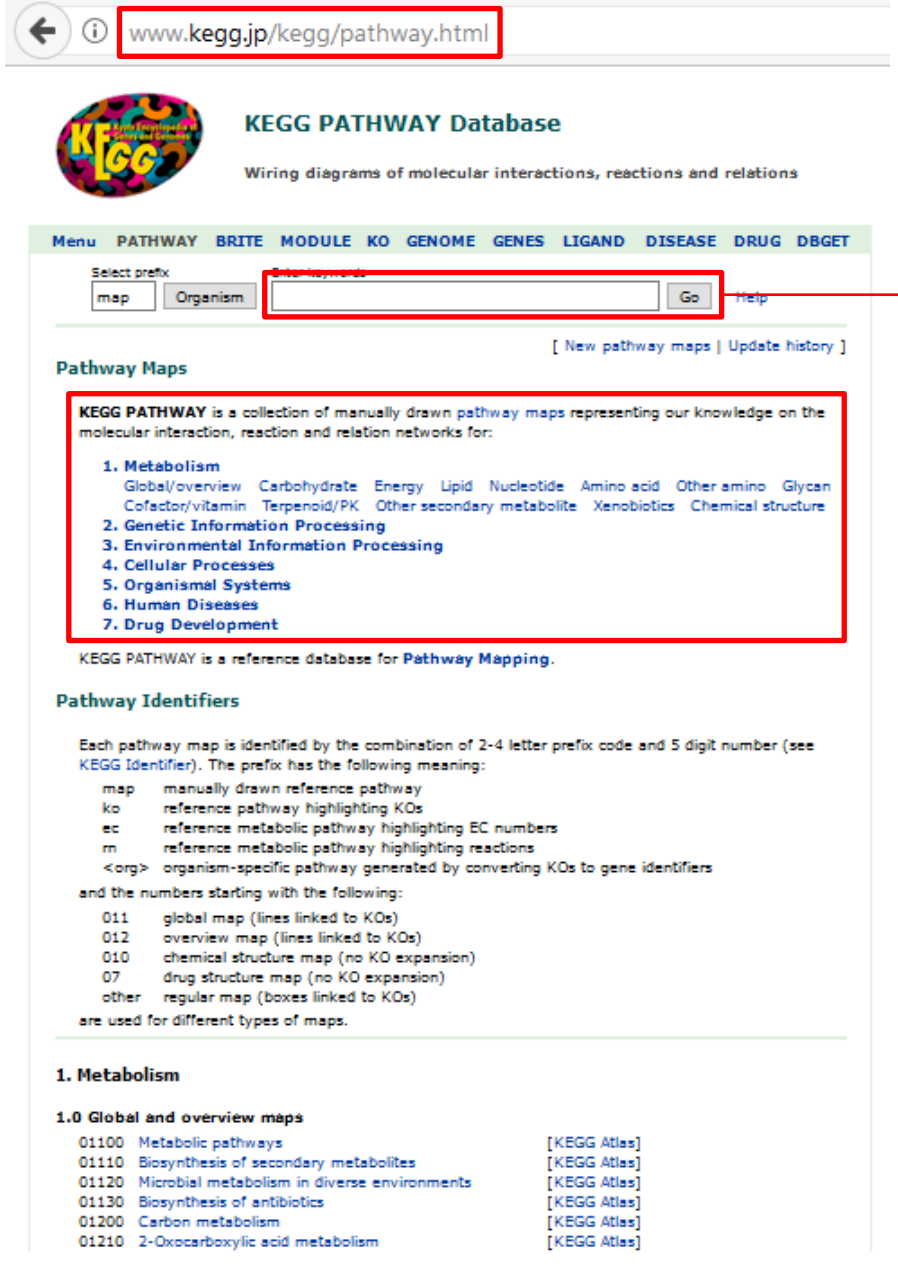
Wikipathways

Pathway Commons

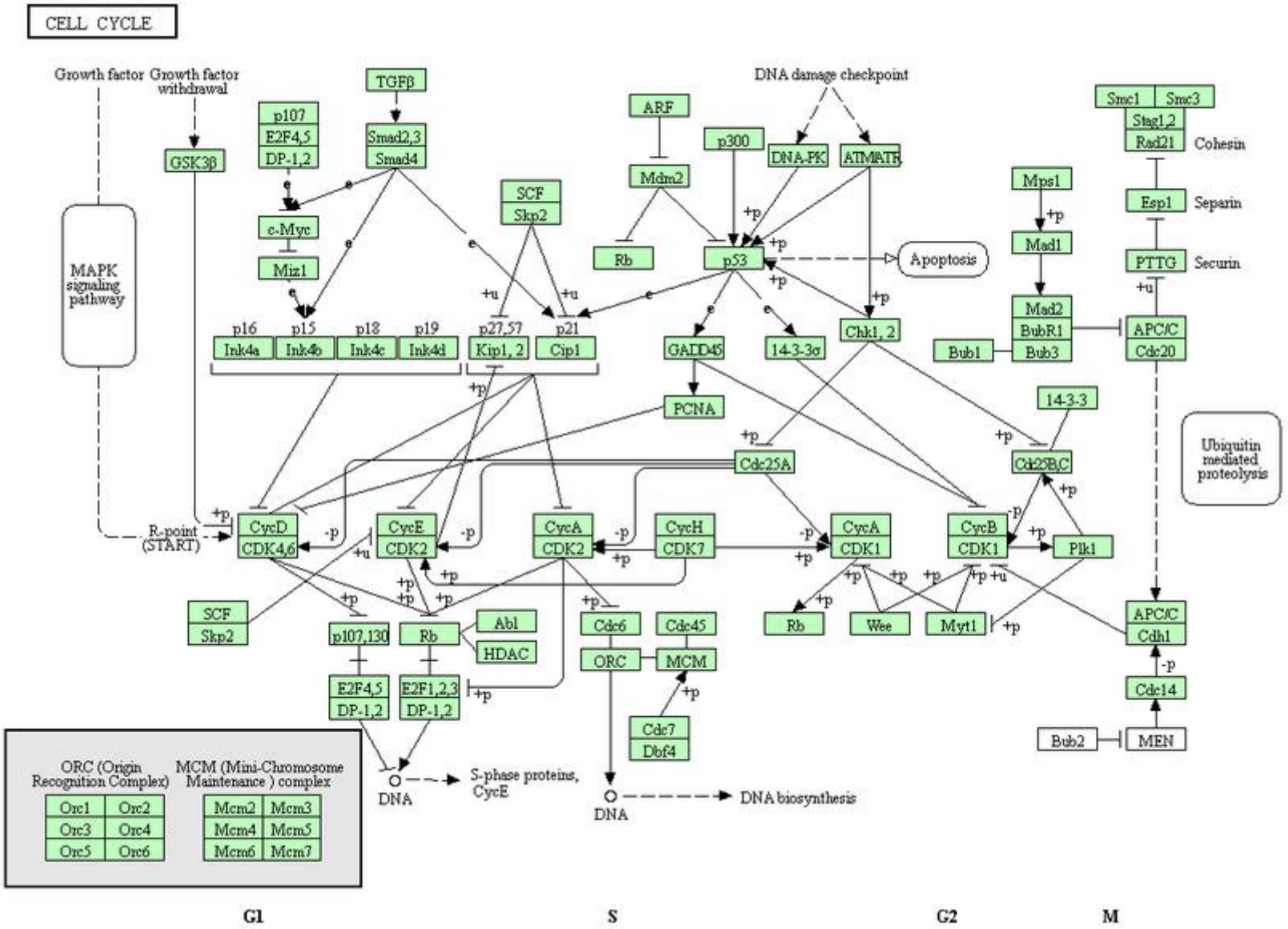
Biocyc

Panther Pathway

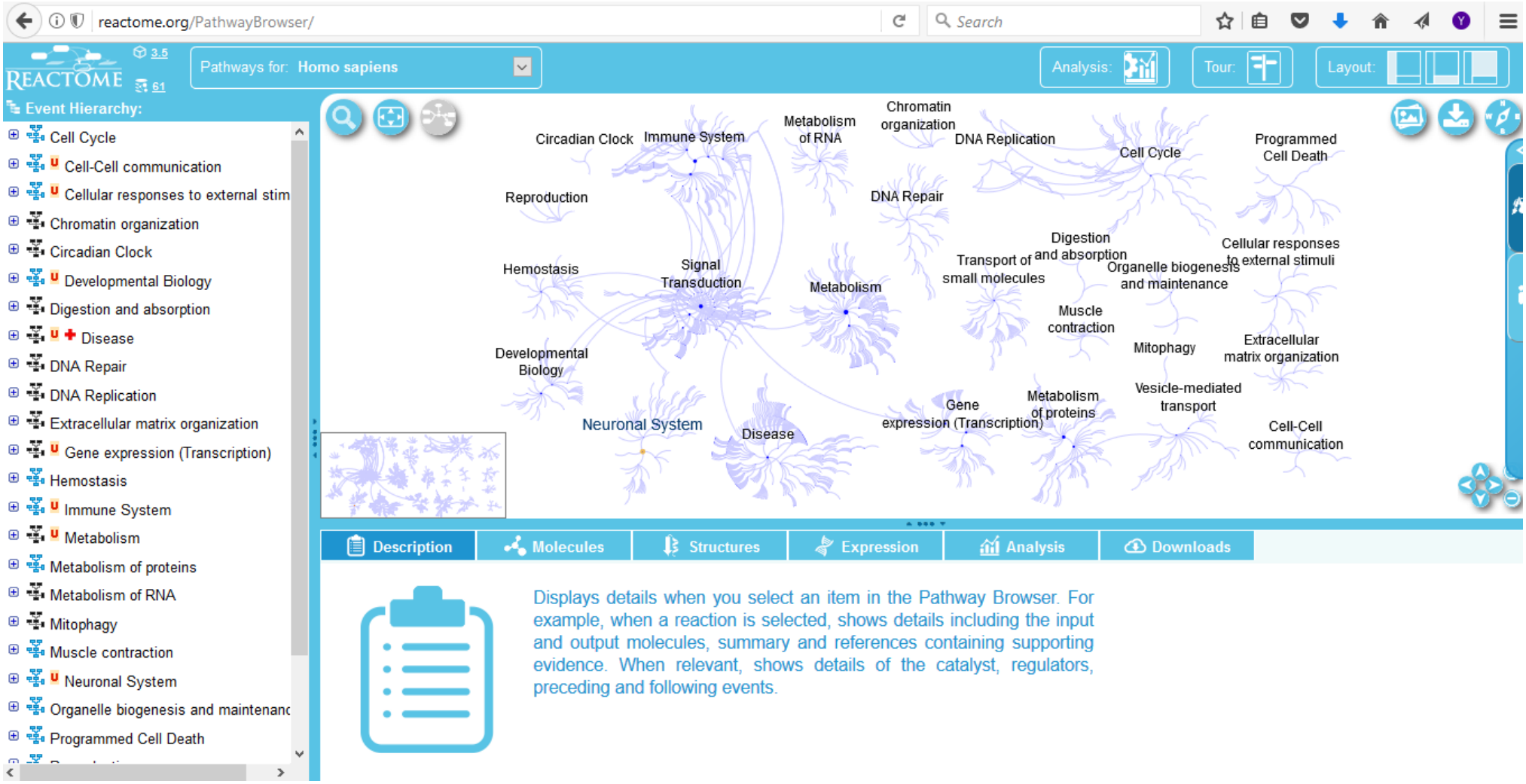
Pathway databases: KEGG



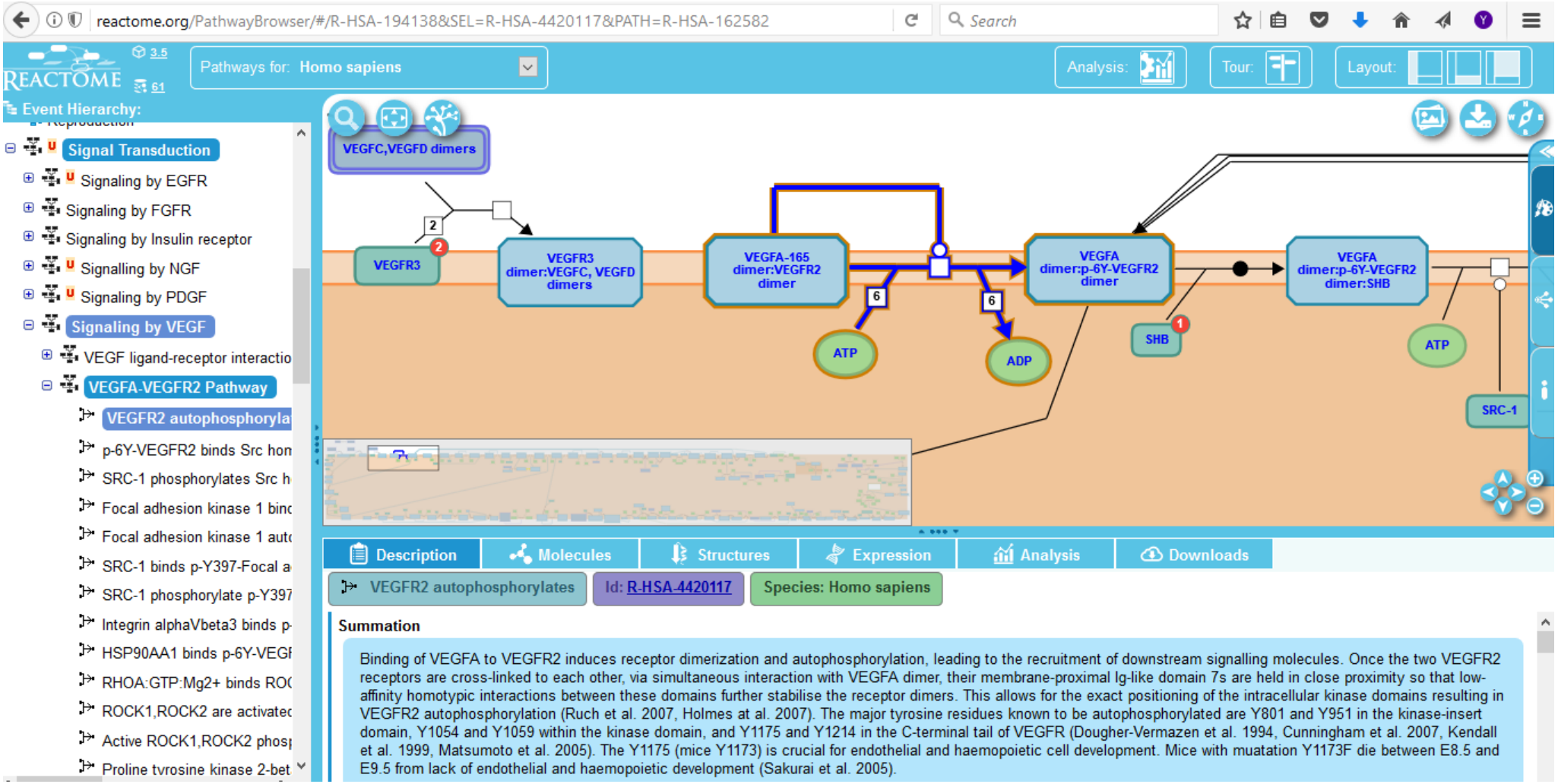
Pathway databases: KEGG



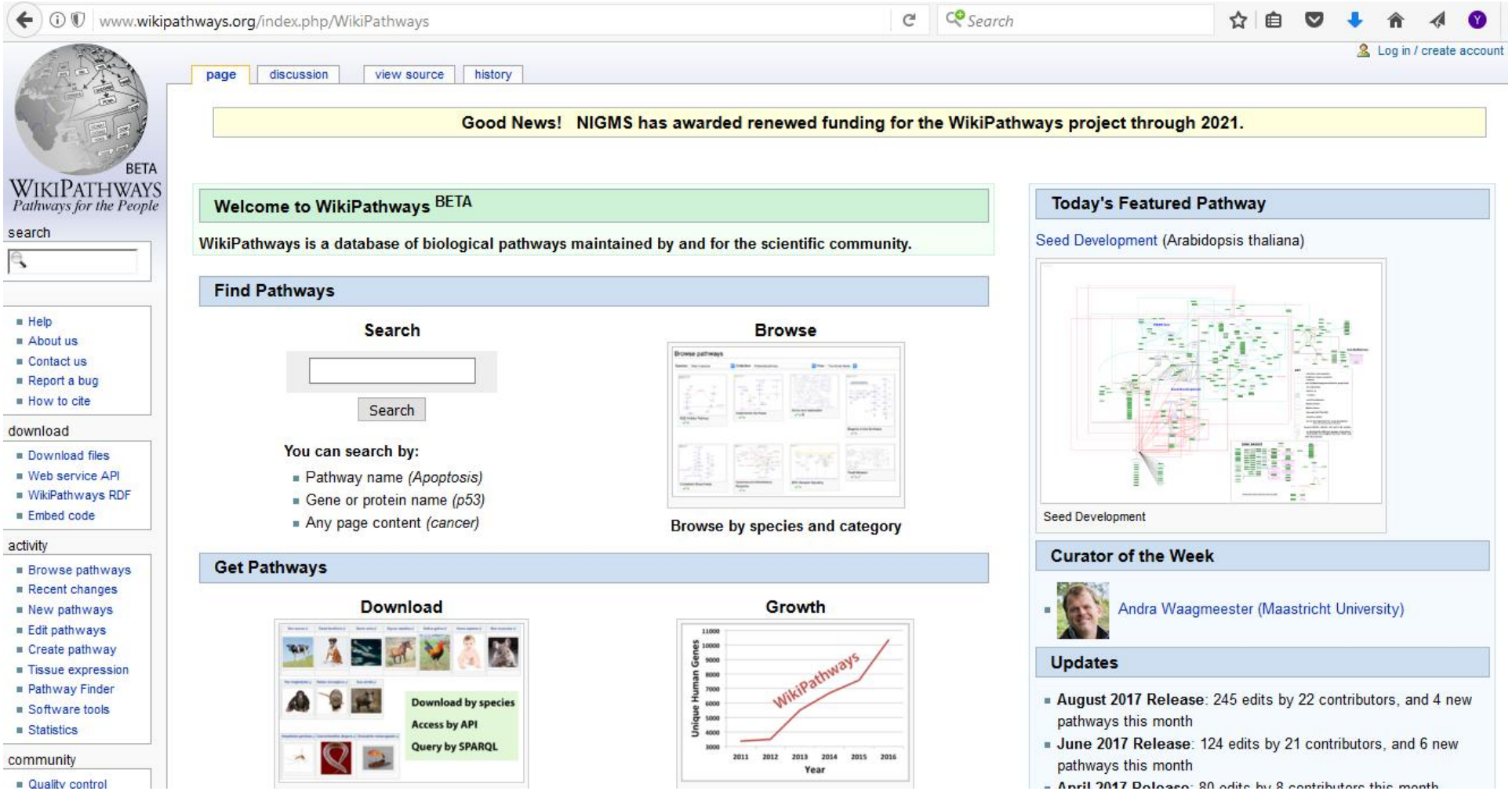
Pathway databases: Reactome



Pathway databases: Reactome



Pathway databases: Wikipathways



Pathway databases: Wikipathways

www.wikipathways.org/index.php/Pathway:WP1965 90% Search

BETA
WIKIPATHWAYS
Pathways for the People

search

- Help
- About us
- Contact us
- Report a bug
- How to cite

download

- Download files
- Web service API
- WikiPathways RDF
- Embed code

activity

- Browse pathways
- Recent changes
- New pathways
- Edit pathways
- Create pathway
- Tissue expression
- Pathway Finder
- Software tools
- Statistics

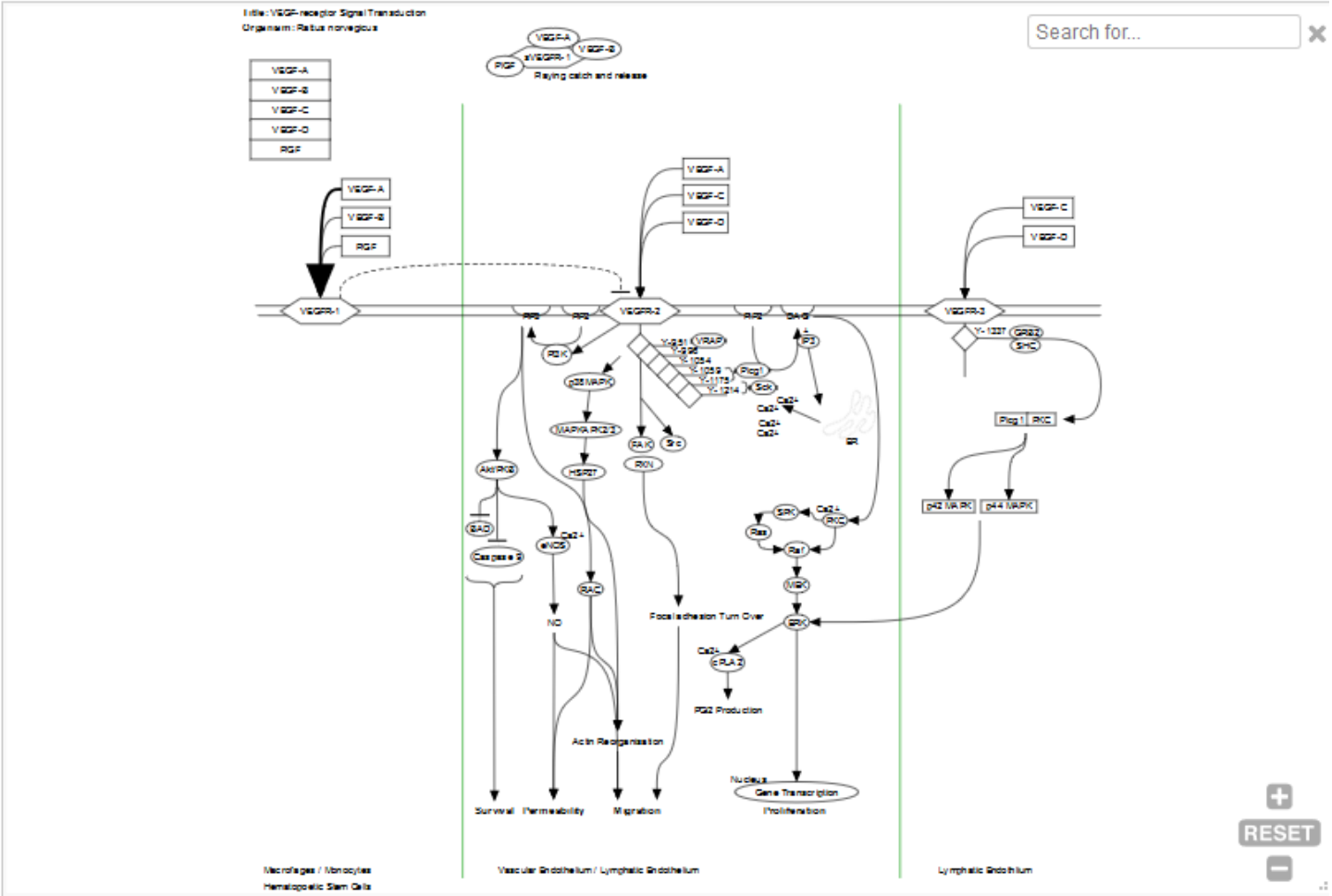
community

- Quality control
- Development
- WikiPathways Blog
- AOP portal
- CIRM portal

pathway | discussion | view source

VEGF-receptor Signal Transduction (Rattus norvegicus)

camiel hoogendoorn, Martina Kutmon, Daniela Digles, et al.



Pathway databases: Reactome vs Wikipathways

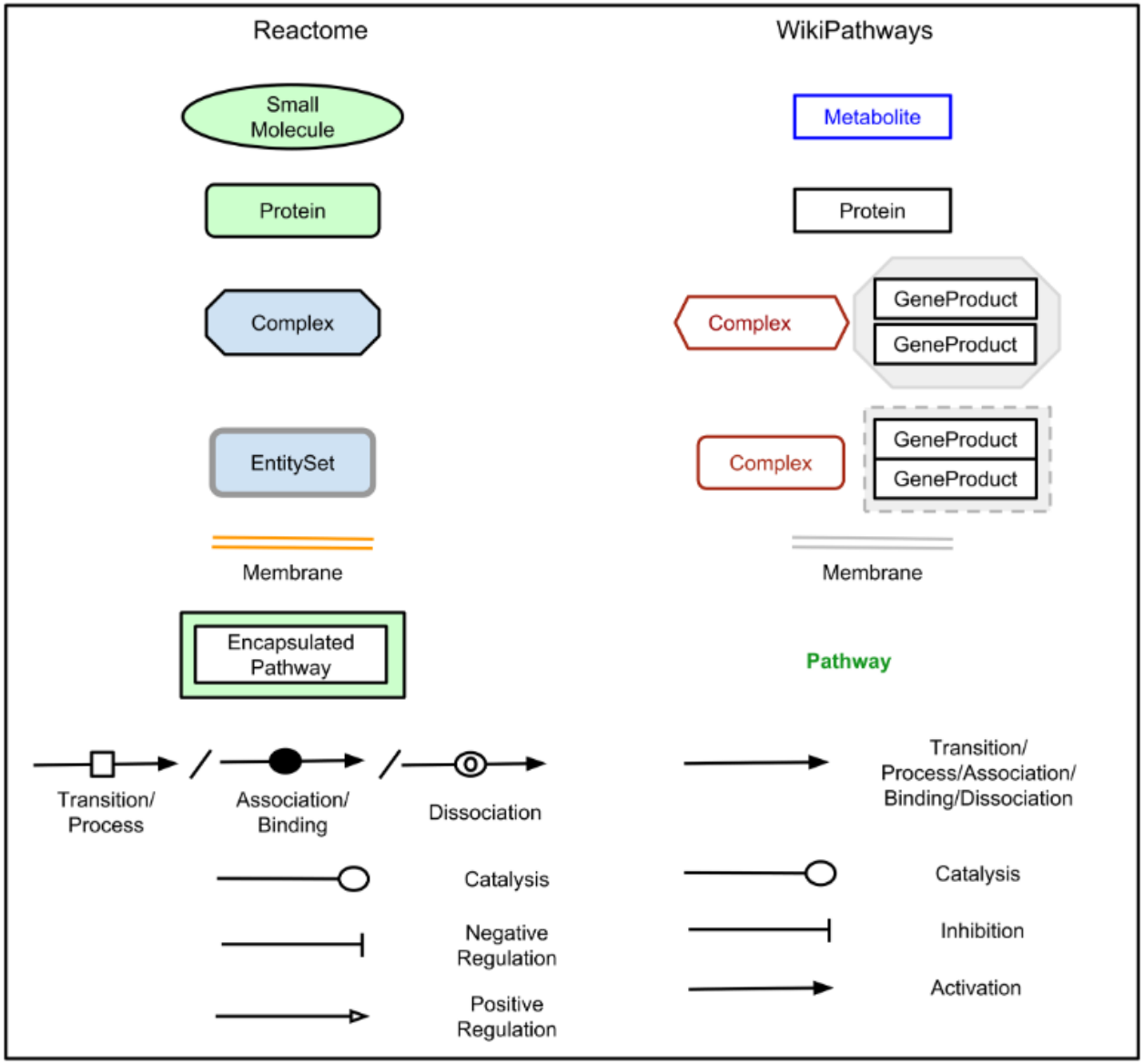
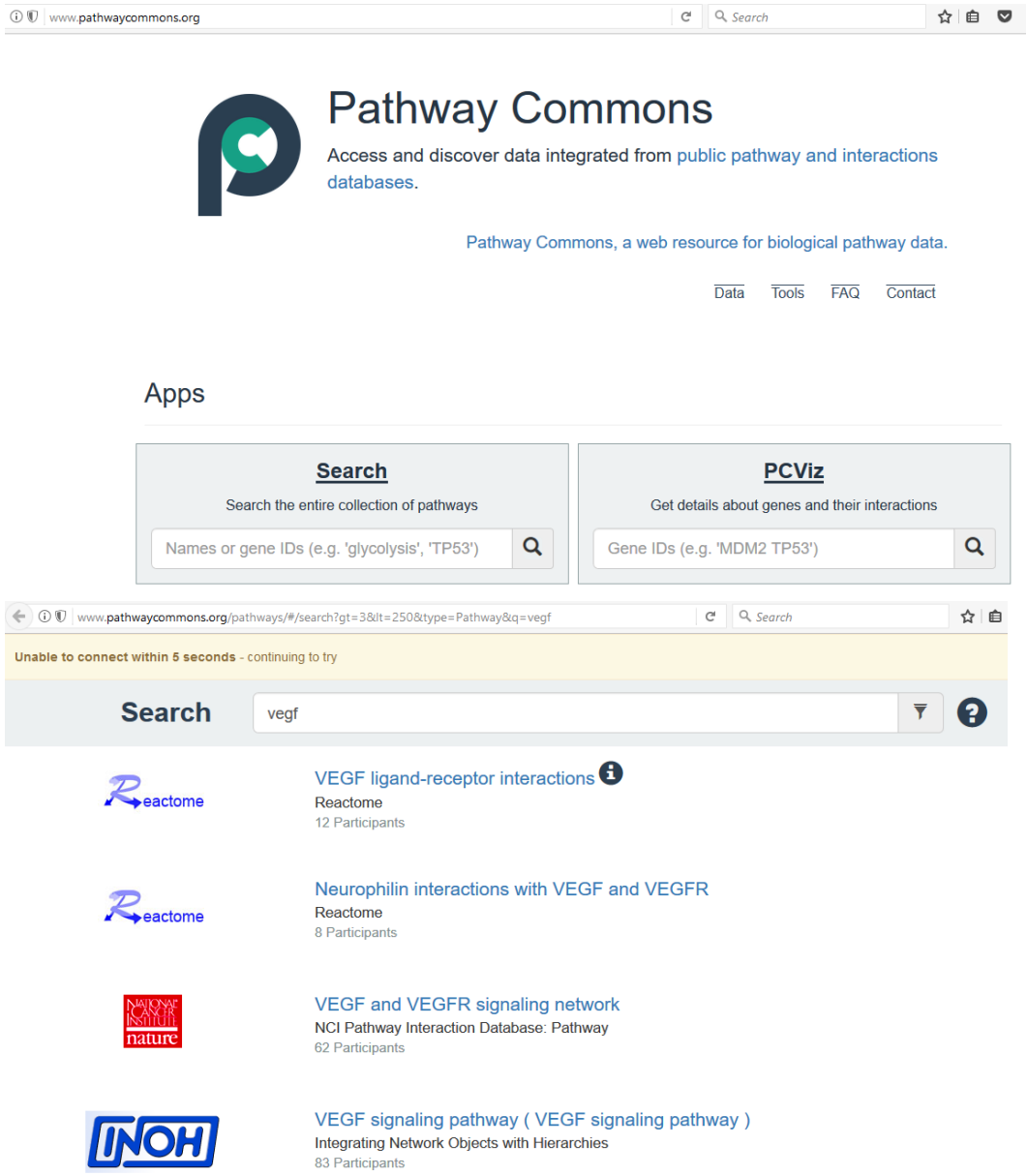


Fig 1. Mapping Reactome pathways elements to WikiPathways pathway elements. This diagram shows the symbols used to represent different biological entities in Reactome and the corresponding symbol used to represent the same biological entity in WikiPathways.

doi:10.1371/journal.pcbi.1004941.g001

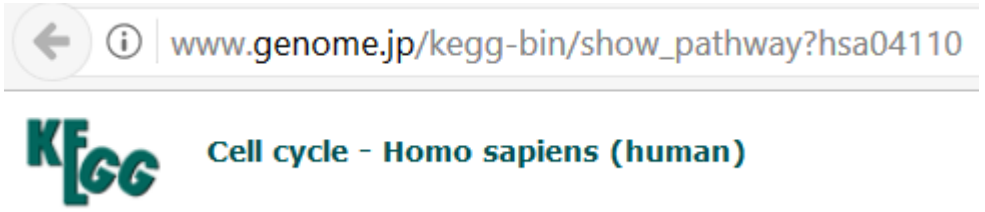
Pathway databases: Pathway Commons

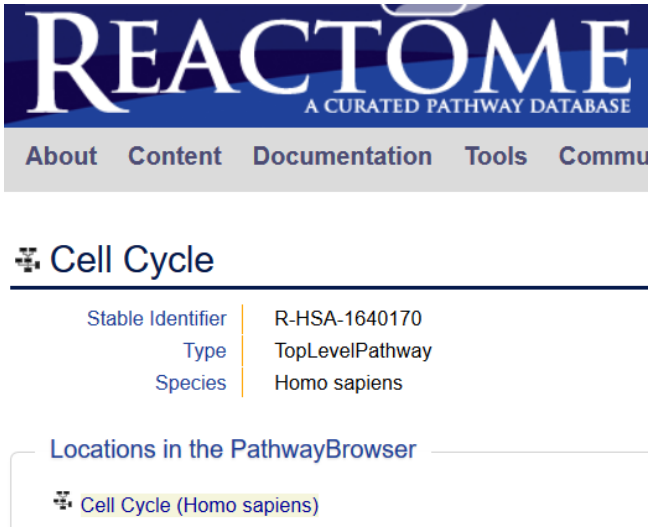
Search for pathways in multiple pathway databases

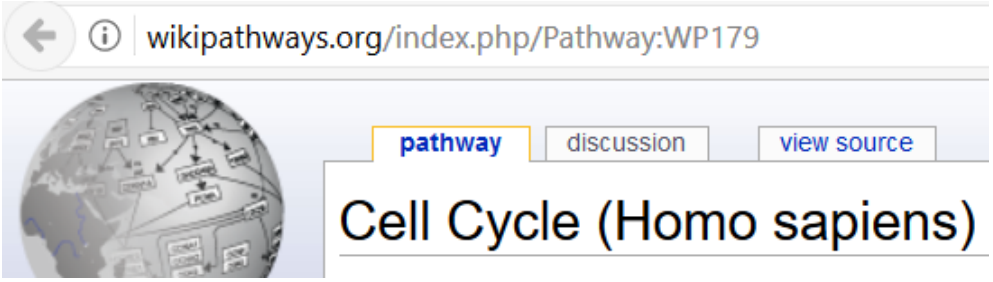


Be aware of... Pathway database identifiers

Identifiers (IDs) are ideally unique, stable names or numbers that help track database records. For example, your wechat ID, Entrez Gene ID 41232, etc. Each DB has its own type of identifier.







Be aware of... Pathway file formats

- Simple graphical file (png, jpeg, etc)
- SBML (Systems Biology Markup Language): Popular in Systems Biology (mathematical models of pathways). Databases of models such as “BioModels”.
- BioPax (Biological Pathway Exchange).

You will need tools that can read the pathway format you choose. Many graphical tools can read SBML and BioPax files.

Databases with BioPAX Export [edit]

Online databases offering BioPAX export include:

- Signaling Gateway Molecule Pages (SGMP)
- Reactome
- BioCyc
- INOH
- BioModels
- Nature/NCI Pathway Interaction Database
- Cancer Cell Map
- Pathway Commons
- Netpath - A curated resource of signal transduction pathways in humans
- ConsensusPathDB - A database integrating human functional interaction networks
- PANTHER (List of Pathways)
- WikiPathways
- PharmGKB/PharmGKB*

Software [edit]

Software supporting BioPAX include:

- Paxtools, a Java API for handling BioPAX files
- Systems Biology Linker (Sybil), an application for visualizing BioPAX and converting BioPAX to SBML, as part of the Virtual Cell.
- ChiBE (Chisio BioPAX Editor),[2] an application for visualizing and editing BioPAX.
- BioPAX Validator - syntax and semantic rules and best practices (project wiki)
- Cytoscape includes a BioPAX reader and other extensions, such as PathwayCommons plugin and CyPath2 app.
- BiNoM, a cytoscape plugin for network analysis, with functions to import and export BioPAX level 3 files.
- BioPAX-pattern, a Java API for defining and searching graph patterns in BioPAX files.

Contents

Pathway visualization: PathVisio

PathVisio

a tool to edit and analyze biological pathways

Home | Getting Started | Support/Help | Downloads | Plugins | Cite Us | About

What is PathVisio?

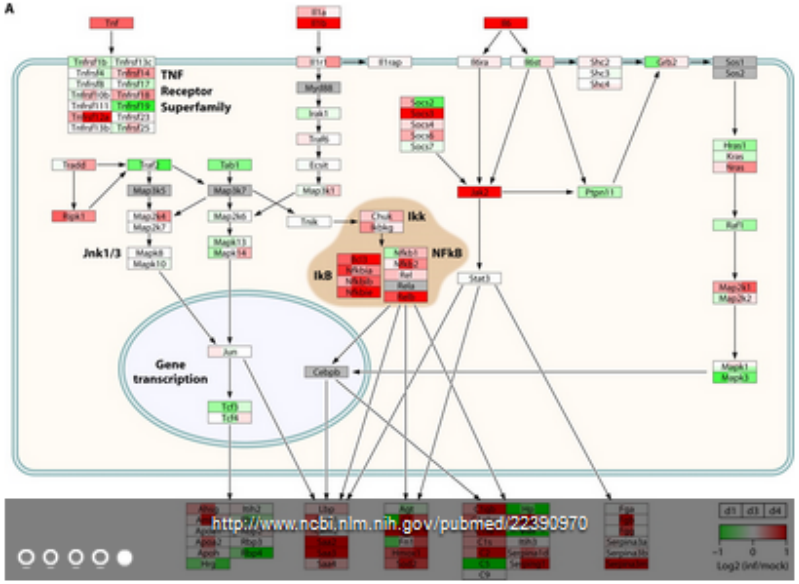
PathVisio is a free open-source biological pathway analysis software that allows you to draw, edit and analyze biological pathways. Learn more.

How to use PathVisio?

Learn how to download PathVisio and go through the tutorials to find out how to do pathway analysis and how to visualize and analyze your data. Get started.

PathVisio Plugins

Plugins are extensions that provide advanced analysis methods, visualization options or additional import/export functionality. Find out more.



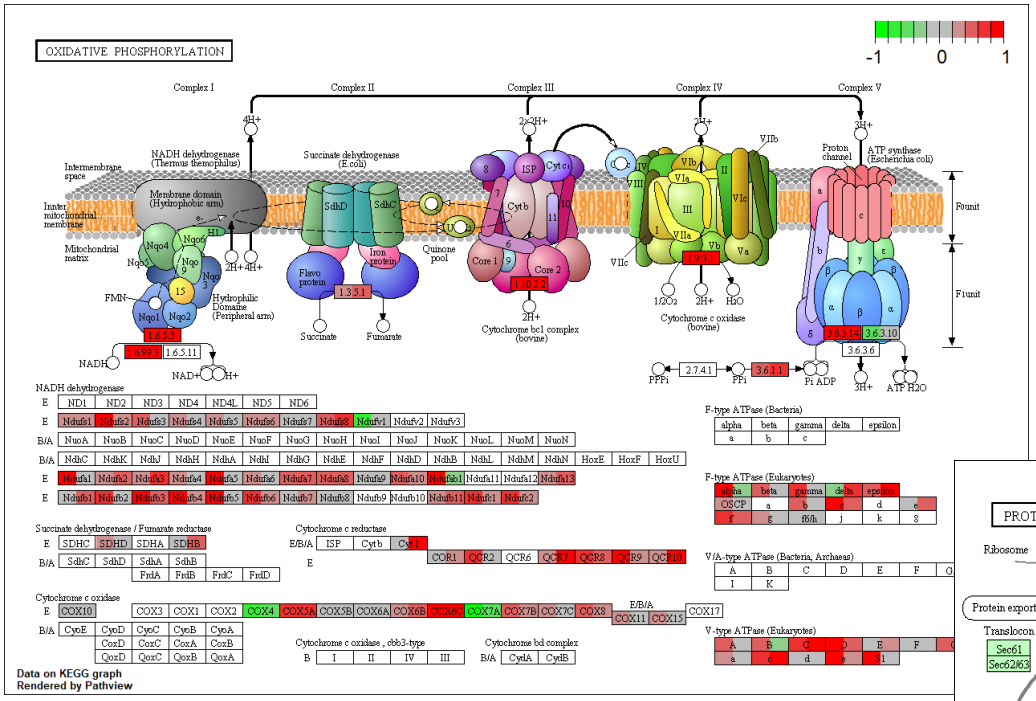
News

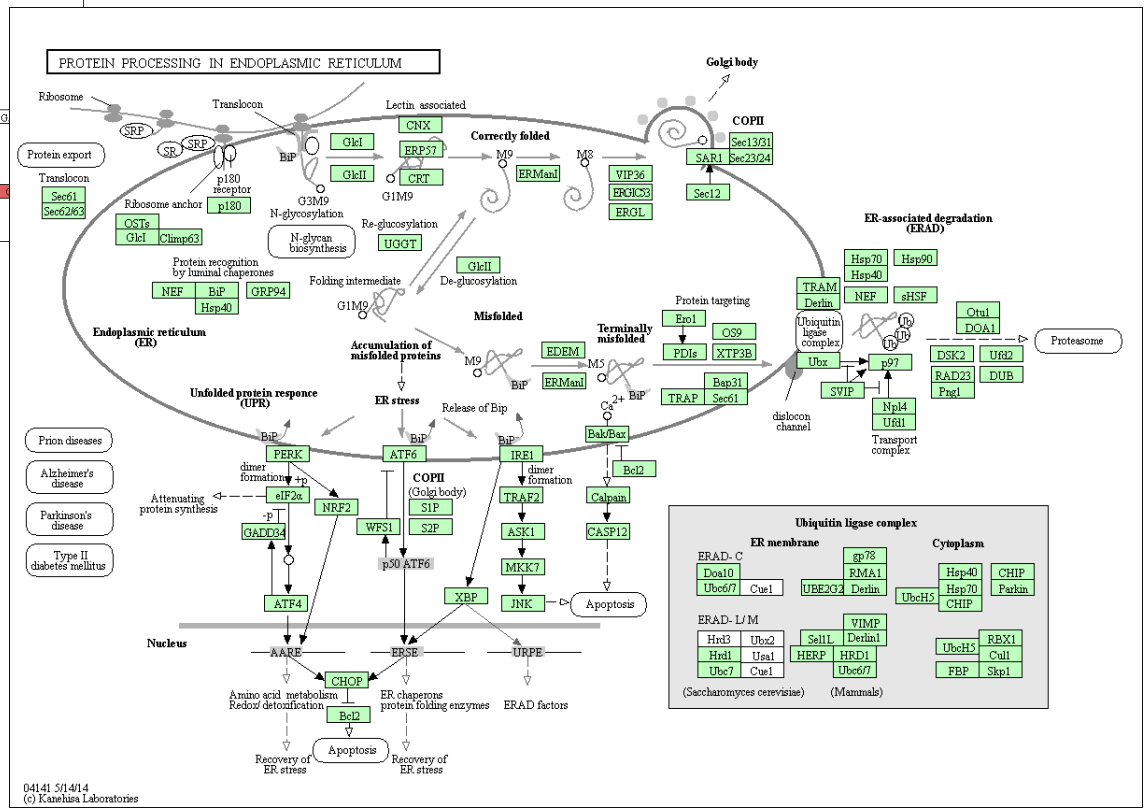
- **WikiPathways survey:** Are you using WikiPathways in your research? Let us know in order to improve the odds of us working on what matters to you: https://www.surveymonkey.com/r/wikipathways
- **Check out the new WikiPathways paper!** "WikiPathways: capturing the full diversity of pathway knowledge." Nucl. Acids Res. first published online October 19, 2015. doi: 10.1093/nar/gkv1024
- **New PathVisio 3 paper!!** "PathVisio 3: An Extendable Pathway Analysis Toolbox." PLoS Comput Biol. 2015 Feb 23;11(2):e1004085. doi: 10.1371/journal.pcbi.1004085
- **Developer website:** The previous website will stay available on developers.pathvisio.org and will be used as a developers website.
- **Plugin repository:** Find plugins in the plugin repository and install them through the plugin manager.

Releases

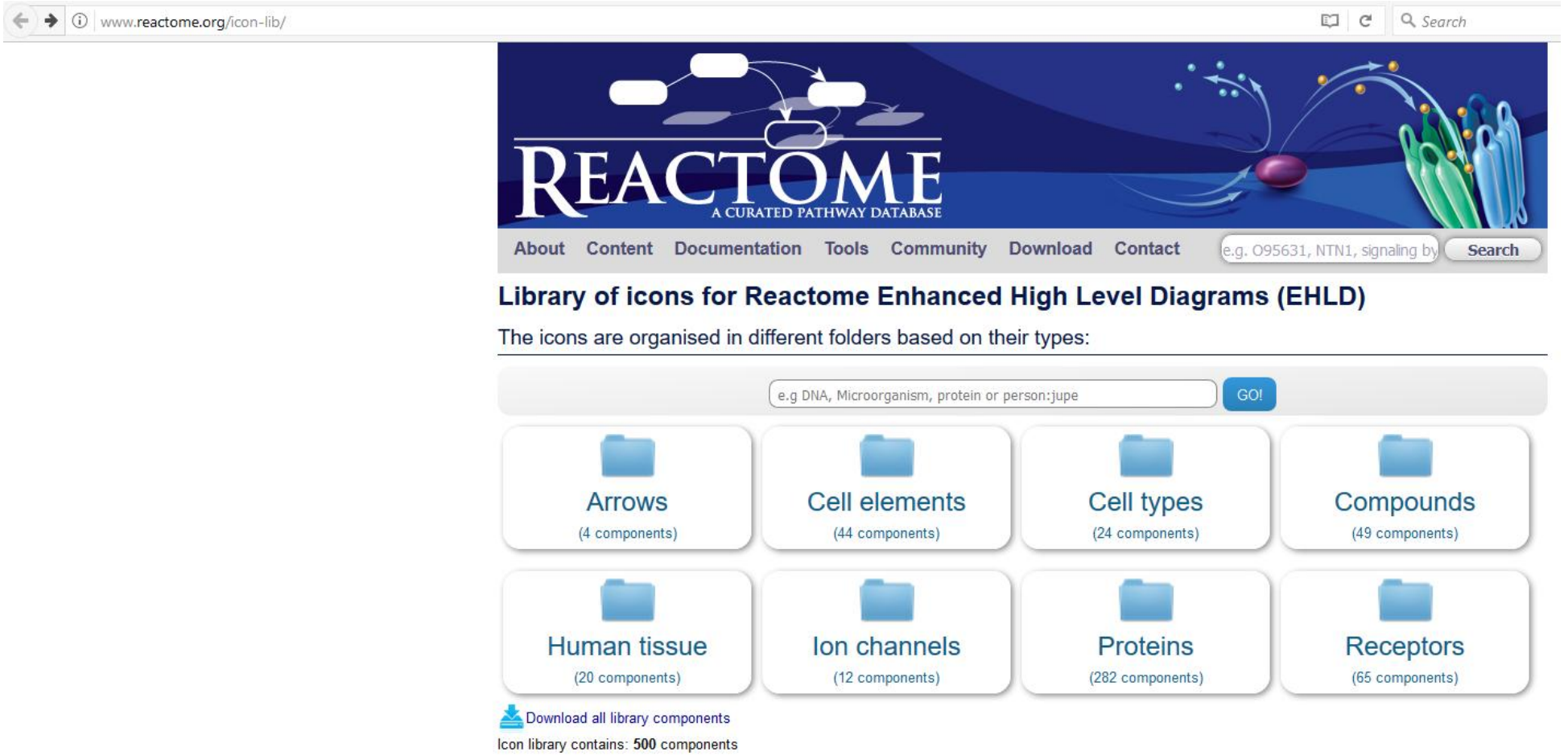
- **Nov 24 2016 PathVisio 3.2.4 has been released!**
- **Sep 20 2016 PathVisio 3.2.3 has been released!**
- **May 03 2016 PathVisio 3.2.2 has been released!**
- **Oct 07 2015 PathVisio 3.2.1 has been released!**
- **Feb 28 2015 PathVisio 3.2.0 has been released!**

Pathway visualization: R / pathview

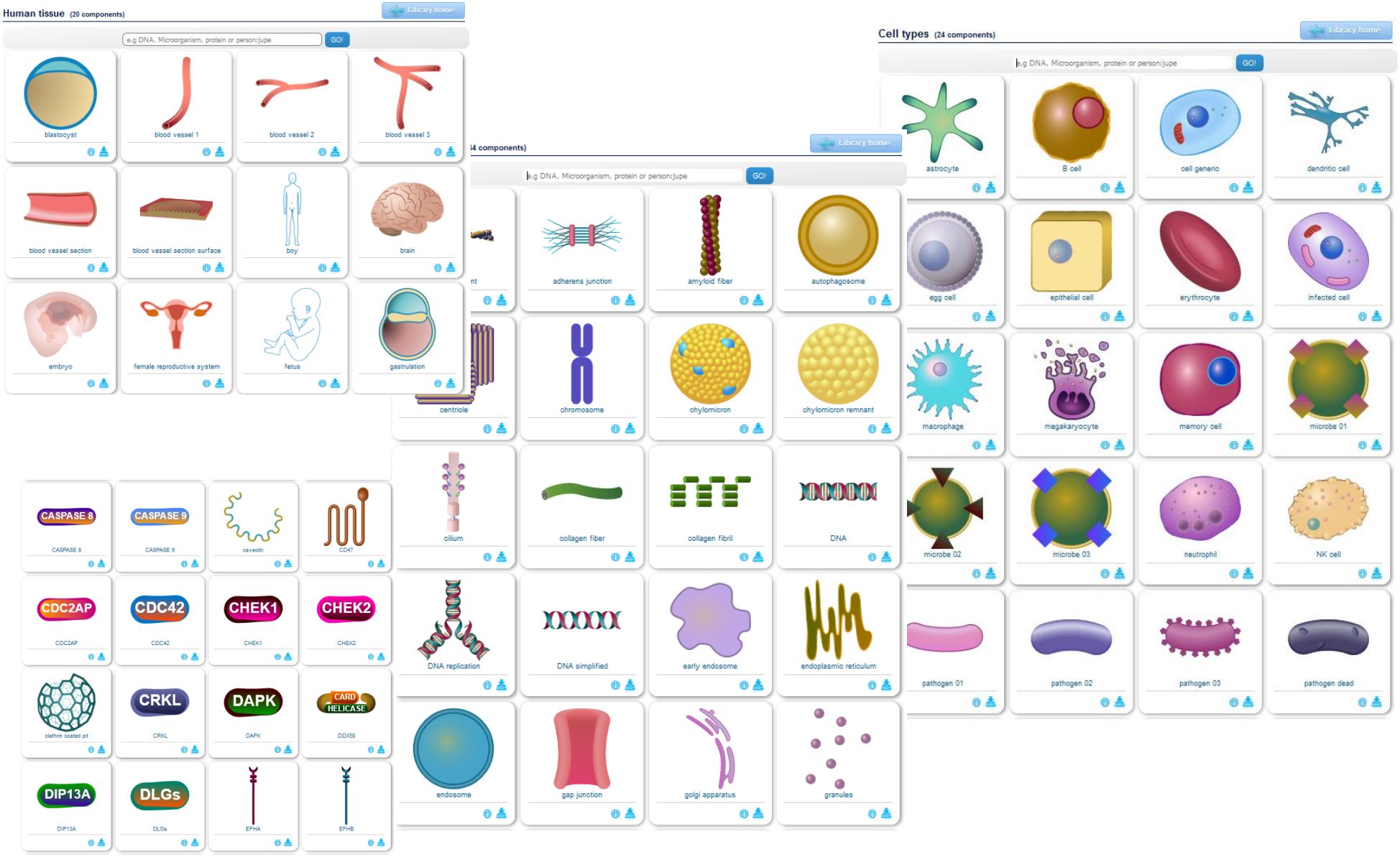




Pathway visualization: Reactome Library of Icons



Pathway visualization: Reactome Library of Icons



Contents

What is the Gene Ontology (GO)?

CAN YOU PLEASE SUMMARIZE ALL CONCEPTS IN THIS MOLECULAR BIOLOGY BOOK AND TELL ME HOW THEY RELATE TO EACH OTHER?

SURE! BASICALLY, THERE ARE 3 MAJOR CATEGORIES IN BIOLOGY: BIOLOGICAL PROCESSES, MOLECULAR FUNCTIONS, AND CELLULAR COMPONENTS. NOW, I WILL SHOW THE SUBDIVISION OF THE THREE CATEGORIES…

WOW! HE ORGANIZED ALL BIOLOGICAL KNOWLEDGE IN AN **ONTOLOGY!!**

https://www.freepik.com/

What is the Gene Ontology (GO)?

ALSO… CAN YOU FIND A WAY TO TELL ME ALL THE BIOLOGY CONCEPTS RELATED TO A GIVEN GENE?

SURE! WE BUILT THIS DATABASE CALLED “GO” WHERE EVERY GENE IS RELATED TO EVERY CONCEPT IN OUR ONTOLOGY

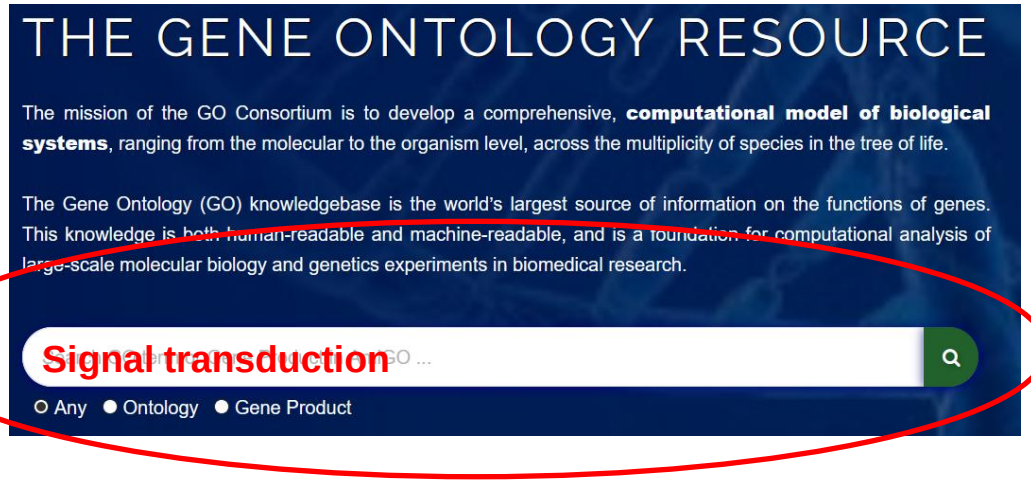
WOW! HIS ONTOLOGY IS **ANNOTATED!!**

https://www.freepik.com/

What is the Gene Ontology (GO)?

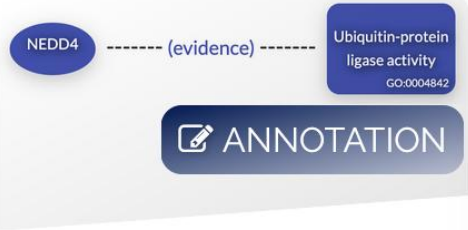
- The GO is a set of words or phrases (called GO terms) which are related to genes. For example: “protein kinase” (a molecular function), “glycolysis” (a biological process), or “nucleus” (a cellular component).
- It is a ***Dictionary***: Term definitions
- It is an ***Ontology***: A formal system for describing knowledge
- It is ***Annotated***: Genes linked to GO terms

The Gene Ontology (GO)



The network of biological classes describing the current best representation of the "universe" of biology. The molecular functions, cellular locations, and processes gene products may carry out.

GO Ontology Overview
Browse in AmiGO
Download



Statements, based on specific, traceable scientific evidence, asserting that a specific gene product is a real exemplar of a particular GO class.

GO Annotations Overview
View in AmiGO
Download

Gene Ontology Consortium

The Gene Ontology (GO)

Description (Name, Ontology, GO Term, Synonym, Definitions):

signal transduction

Term Information

Data health

Accession	GO:0007165
Name	signal transduction
Ontology	biological_process
Synonyms	signaling pathway, signalling pathway, signaling cascade, signalling cascade
Alternate IDs	GO:0023033
Definition	The cellular process in which a signal is conveyed to trigger a change in the activity or state of a cell. Signal transduction begins with reception of a signal (e.g. a ligand binding to a receptor or receptor activation by a stimulus such as light), or for signal transduction in the absence of ligand, signal-withdrawal or the activity of a constitutively active receptor. Signal transduction ends with regulation of a downstream cellular process, e.g. regulation of transcription or regulation of a metabolic process. Signal transduction covers signaling from receptors located on the surface of the cell and signaling via molecules located within the cell. For signaling between cells, signal transduction is restricted to events at and within the receiving cell. *Source:* GOC:go_curators, GOC:mtg_signaling_feb11
Comment	Note that signal transduction is defined broadly to include a ligand interacting with a receptor, downstream signaling steps and a response being triggered. A change in form of the signal in every step is not necessary. Note that in many cases the end of this process is regulation of the initiation of transcription. Note that specific transcription factors may be annotated to this term, but core/general transcription machinery such as RNA polymerase should not.
History	See term history for GO:0007165 at QuickGO
Subset	goslim_metagenomics goslim_aspergillus goslim_chembl goslim_plant goslim_generic gosubset_prok goslim_candida
Related	Link to all **genes and gene products** annotated to signal transduction. Link to all direct and indirect **annotations** to signal transduction. Link to all direct and indirect **annotations download** (limited to first 10,000) for signal transduction.

Gene Ontology Consortium

The Gene Ontology (GO)

Annotations:

Annotations | Graph Views | Inferred Tree View | Neighborhood | Mappings

Filter results

Total annotations: 16657

User filters

+ taxon_subset_closure_label: Homo sapiens
+ aspect: P

Your search is pinned to these filters

− document_category: annotation
− regulates_closure: GO:0007165

Ontology (aspect)

Organism

Nothing to filter.

Evidence

GO class

GO class (direct)

Annotation qualifier

Annotation extension

Contributor

PANTHER family

Total annotations: 16657; showing: 1-10
Results count 10

«First <Prev Next> Last» Download

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference	Date
MSX2	Homeobox protein MSX-2		signal transduction involved in regulation of gene expression		Ensembl	Homo sapiens	IEA	UniProtKB:Q03358 ensembl:ENSMUSP00000021922	family not named pthr24338		GO_REF:0000107	20170826
MSX2	Homeobox protein MSX-2		positive regulation of BMP signaling pathway		Ensembl	Homo sapiens	IEA	UniProtKB:Q03358 ensembl:ENSMUSP00000021922	family not named pthr24338		GO_REF:0000107	20170826
MSX2	Homeobox protein MSX-2		BMP signaling pathway involved in heart development		Ensembl	Homo sapiens	IEA	UniProtKB:Q03358 ensembl:ENSMUSP00000021922	family not named pthr24338		GO_REF:0000107	20170826
MAPK8IP3	C-Jun-amino-terminal kinase-interacting protein 3		activation of JUN kinase activity		GO_Central	Homo sapiens	IBA	PANTHER:PTN000356517	jnk/sapk-associated protein pthr13886		GO_REF:0000033	20141001
MAPK8IP3	C-Jun-amino-terminal kinase-interacting protein 3		regulation of JNK cascade		UniProt	Homo sapiens	ISS	UniProtKB:Q9ESN9	jnk/sapk-associated protein pthr13886		GO_REF:0000024	20041006
EREG	Proepiregulin		MAPK cascade		Reactome	Homo sapiens	TAS		epiregulin pthr22610		Reactome:R-HSA-5673001	20170526
EREG	Proepiregulin		epidermal growth factor receptor signaling pathway		UniProt	Homo sapiens	ISS	UniProtKB:Q61521	epiregulin pthr22610		GO_REF:0000024	20060119
EREG	Proepiregulin		epidermal growth factor receptor signaling pathway		GO_Central	Homo sapiens	IBA	PANTHER:PTN001098750	epiregulin pthr22610		GO_REF:0000033	20140922
EREG	Proepiregulin		regulation of phosphatidylinositol 3-kinase signaling		Reactome	Homo sapiens	TAS		epiregulin pthr22610		Reactome:R-HSA-6811558	20170526
EREG	Proepiregulin		cytokine-mediated signaling pathway		UniProt	Homo sapiens	IDA		epiregulin pthr22610		PMID:9419975	20060120

Ontology tree:

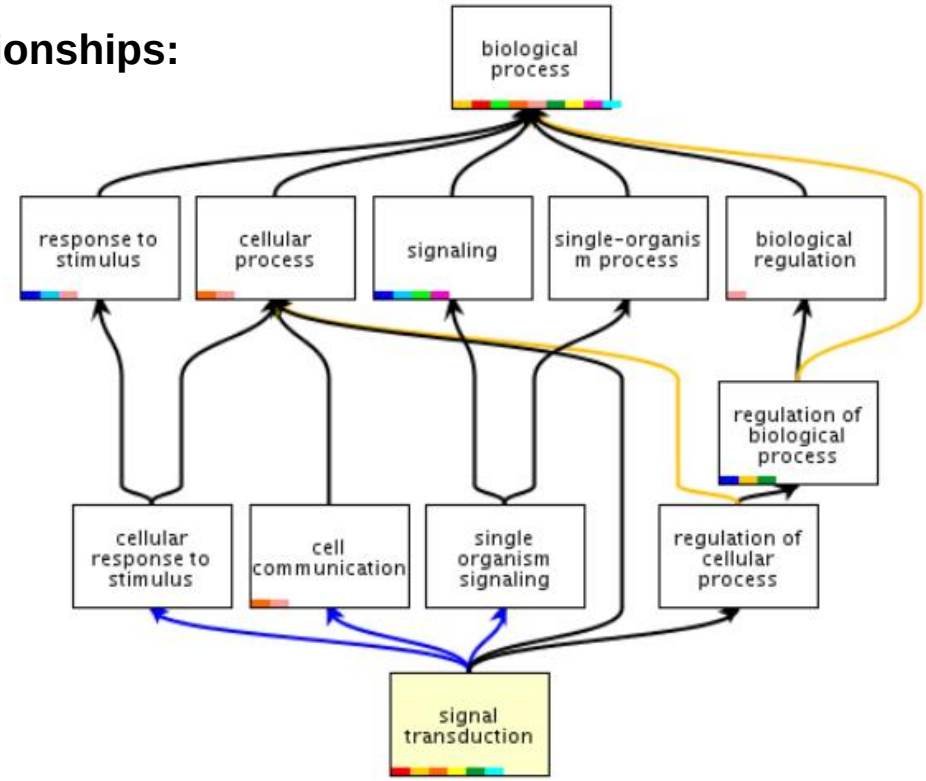
Parents and children

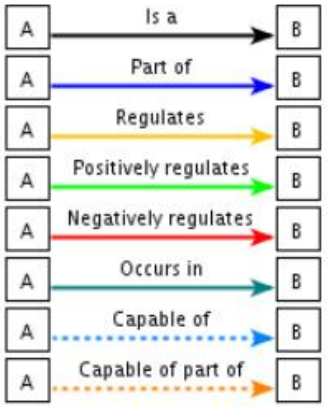
amigo.geneontology.org/amigo/term/GO:0007165

Annotations | Graph Views | Inferred Tree View | Neighborhood | Mappings

- GO:0008150 biological_process
  - GO:0065007 biological regulation
    - GO:0009987 cellular process
    - GO:0050789 regulation of biological process
    - GO:0050896 response to stimulus
      - GO:0007154 cell communication
      - GO:0051716 cellular response to stimulus
      - GO:0050794 regulation of cellular process
      - GO:0023052 signaling
        - **GO:0007165 signal transduction**
          - GO:0095500 acetylcholine receptor signaling pathway
          - GO:0007196 adenylate cyclase-inhibiting G-protein coupled glutamate receptor signaling pathway
          - GO:0007198 adenylate cyclase-inhibiting serotonin receptor signaling pathway
          - GO:0071875 adrenergic receptor signaling pathway
          - GO:0098990 AMPA selective glutamate receptor signaling pathway
          - GO:0097190 apoptotic signaling pathway
          - GO:0038183 bile acid signaling pathway
          - GO:0099004 calmodulin dependent kinase signaling pathway
          - GO:0038171 cannabinoid signaling pathway
          - GO:0009756 carbohydrate mediated signaling
          - GO:0007166 cell surface receptor signaling pathway
          - GO:0010019 chloroplast-nucleus signaling pathway
          - GO:0009870 defense response signaling pathway, resistance gene-dependent
          - GO:0010204 defense response signaling pathway, resistance gene-independent
          - GO:0030968 endoplasmic reticulum unfolded protein response
          - GO:2000803 endosomal signal transduction
          - GO:0006984 ER-nucleus signaling pathway
          - GO:0007213 G-protein coupled acetylcholine receptor signaling pathway
          - GO:0007216 G-protein coupled glutamate receptor signaling pathway
          - GO:0007186 G-protein coupled receptor signaling pathway
          - GO:0098664 G-protein coupled serotonin receptor signaling pathway
          - GO:0007215 glutamate receptor signaling pathway
          - GO:0009755 hormone-mediated signaling pathway
          - GO:0071588 hydrogen peroxide mediated signaling pathway
          - GO:0097411 hypoxia-inducible factor-1alpha signaling pathway
          - GO:0002764 immune response-regulating signaling pathway
          - GO:0030522 intracellular receptor signaling pathway
          - GO:0035556 intracellular signal transduction
          - GO:0035235 ionotropic glutamate receptor signaling pathway
          - GO:0098991 kainate selective glutamate receptor signaling pathway
          - GO:0055095 lipoprotein particle mediated signaling
          - GO:0031930 mitochondria-nucleus signaling pathway
          - GO:0097527 necroptotic signaling pathway

Ontology relationships:





- Terms are related within a hierarchy
- Terms can have more than one parent or child

QuickGO - http://www.ebi.ac.uk/QuickGO

Reactome
REACT_89740
REACT_100624
REACT_112549
REACT_102354
REACT_114820
REACT_114657
REACT_113601
REACT_113964
REACT_12478
REACT_114910
REACT_114690
REACT_93680
REACT_98872
REACT_113151
REACT_78535
REACT_112130
REACT_115037
REACT_115147
REACT_31232

Wikipedia Signal_transduction

GO Terms and GO Annotations

- GO terms are added by editors at EBI
- Some terms may be added by request

- Genes are associated with GO terms either by trained curators or created automatically (without human review)
- Multiple annotations per gene
- Manual annotation is considered of higher quality but it is time-consuming.
- Electronic annotation may have variable quality.

Evidence Codes

GO terms include information regarding the type of evidence. For example:

EXP: Inferred from Experiment

IEA: Inferred from electronic annotation

IC: Inferred by curator

Many others (see: http://geneontology.org/docs/guide-go-evidence-codes/)

Key point: Be aware of annotation origin

Guide to GO evidence codes

A GO annotation is a statement about the function of a particular gene. Each annotation includes an evidence code

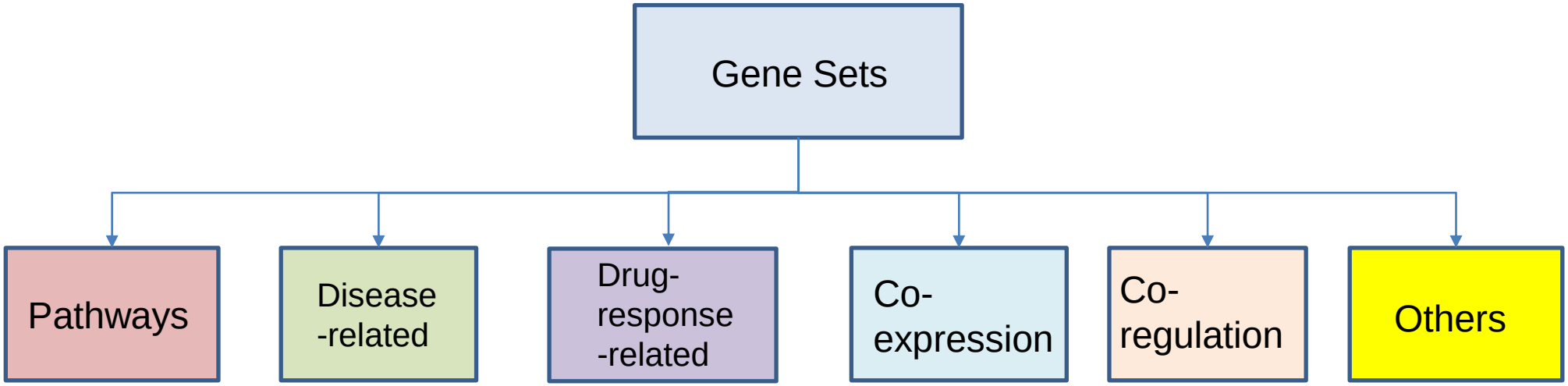
Evidence codes fall into six general categories:

- experimental evidence
- phylogenetic evidence
- computational evidence
- author statements
- curatorial statements
- automatically generated annotations

http://geneontology.org/docs/guide-go-evidence-codes/

Contents

From pathways to «gene sets»



From pathway databases to “gene set” databases, such as **GeneSetDB** (Araki, 2012) and **MSigDB** (Broad Institute), which include pathways, phenotypes, GO, and others.

GeneSetDB

Table 1
Sources databases included in GeneSetDB.

Subclass Name	Sources database	Reference/URL
Pathway	Biocarta	http://www.biocarta.com
	EHMN	[15]
	HumanCyc	[16]
	INOH	[17]
	NetPath	[18]
	PID	[19]
	Reactome	[20]
	SMPDB	[21]
	Wikipathways	[22]
Disease/Phenotype	CancerGenes	[23]
	HPO	[24]
	KEGG Disease	[25]
	MethCancerDB	[26]
	MethyCancer	[27]
	MPO	[28]
	SIDER	[29]
Drug/Chemical	CTD	[30]
	DrugBank	[31]
	MATADOR	[32]
	STITCH	[33]
	T3DB	[34]
Gene Regulation	MicroCosm Targets	[35]
	miRTarBase	[36]
	Rel/NF-κB target genes	http://bioinfo.lifl.fr/NF-KB
	TFactS	[37]
GO	Gene Ontology	[8]

MSigDB

MSigDB
Molecular Signatures Database

MSigDB Collections

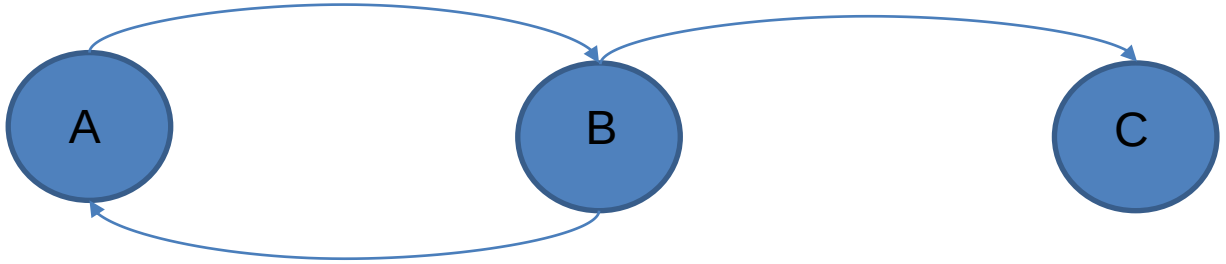
The 17779 gene sets in the Molecular Signatures Database (MSigDB) are divided into 8 major collections, and several sub-collections. See the table below for a brief description of each, and the MSigDB Collections: Details and Acknowledgments page for more detailed descriptions. See also the MSigDB Statistics and the MSigDB Release Notes.

Click on the "browse gene sets" links in the table below to view the gene sets in a collection. Or download the gene sets in a collection by clicking on the links below the "Download GMT Files" headings. For a description of the GMT file format see the Data Formats in the Documentation section. The gene sets can be downloaded as Entrez Gene Identifiers or HUGO Gene Symbols. An XML file containing all the MSigDB gene sets is available on the Downloads page.

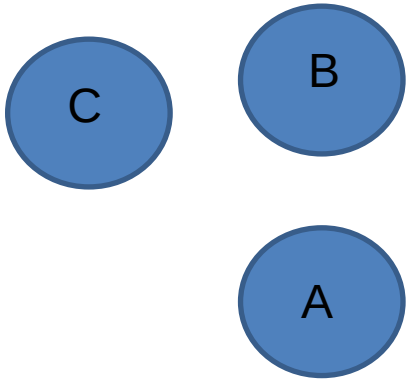
H: hallmark gene sets (browse 50 gene sets)	Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression. details	Download GMT Files gene symbols entrez genes ids
C1: positional gene sets (browse 326 gene sets)	Gene sets corresponding to each human chromosome and each cytogenetic band that has at least one gene. details	Download GMT Files gene symbols entrez genes ids
C2: curated gene sets (browse 4731 gene sets)	Gene sets curated from various sources such as online pathway databases, the biomedical literature, and knowledge of domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into two sub-collections: CGP and CP. details	Download GMT Files gene symbols entrez genes ids
CGP: chemical and genetic perturbations (browse 3402 gene sets)	Gene sets represent expression signatures of genetic and chemical perturbations. A number of these gene sets come in pairs: xxx_UP (and xxx_DN) gene set representing genes induced (and repressed) by the perturbation.	Download GMT Files gene symbols entrez genes ids
CP: Canonical pathways (browse 1329 gene sets)	Gene sets from pathway databases. Usually, these gene sets are canonical representations of a biological process compiled by domain experts.	Download GMT Files gene symbols entrez genes ids
CP:BIOCARTA: BioCarta gene sets (browse 217 gene sets)	Gene sets derived from the BioCarta pathway database.	Download GMT Files gene symbols entrez genes ids
CP:KEGG: KEGG gene sets (browse 186 gene sets)	Gene sets derived from the KEGG pathway database.	Download GMT Files gene symbols entrez genes ids
CP:REACTOME: Reactome gene sets (browse 674 gene sets)	Gene sets derived from the Reactome pathway database.	Download GMT Files gene symbols entrez genes ids
C3: motif gene sets (browse 836 gene sets)	Gene sets representing potential targets of regulation by transcription factors or microRNAs. The sets consist of genes grouped by short sequence motifs they share in their non-protein coding regions. The motifs represent known or likely cis-regulatory elements in promoters and 3'-UTRs. The C3 collection is divided into two sub-collections: MIR and TFT details	Download GMT Files gene symbols entrez genes ids
MIR: microRNA targets (browse 221 gene sets)	Gene sets that contain genes sharing putative target sites (seed matches) of human mature miRNA in their 3'-UTRs.	Download GMT Files gene symbols entrez genes ids
TFT: transcription factor targets (browse 615 gene sets)	Gene sets that share upstream *cis*-regulatory motifs which can function as potential transcription factor binding sites. Based on work by Xie et al. 2005	Download GMT Files gene symbols entrez genes ids
C4: computational gene sets (browse 858 gene sets)	Computational gene sets defined by mining large collections of cancer-oriented microarray data. The C4 collection is divided into two sub-collections: CGN and CM. details	Download GMT Files gene symbols entrez genes ids
CGN: cancer gene neighborhoods (browse 427 gene sets)	Gene sets defined by expression neighborhoods centered on 380 cancer-associated genes. This collection is described in Subramanian, Tamayo et al. 2005	Download GMT Files gene symbols entrez genes ids
CM: cancer modules (browse 431 gene sets)	Gene sets defined by Segal et al. 2004. Briefly, the authors compiled gene sets ('modules') from a variety of resources such as KEGG, GO, and others. By mining a large compendium of cancer-related microarray data, they identified 456 such modules as significantly changed in a variety of cancer conditions.	Download GMT Files gene symbols entrez genes ids
C5: GO gene sets (browse 5917 gene sets)	Gene sets that contain genes annotated by the same GO term. The C5 collection is divided into three sub-collections based on GO ontologies: BP, CC, and MF. details	Download GMT Files gene symbols entrez genes ids
BP: GO biological process (browse 4436 gene sets)	Gene sets derived from the GO Biological Process Ontology.	Download GMT Files gene symbols entrez genes ids
CC: GO cellular component (browse 580 gene sets)	Gene sets derived from the GO Cellular Component Ontology.	Download GMT Files gene symbols entrez genes ids
MF: GO molecular function (browse 901 gene sets)	Gene sets derived from the GO Molecular Function Ontology.	Download GMT Files gene symbols entrez genes ids
C6: oncogenic signatures (browse 189 gene sets)	Gene sets that represent signatures of cellular pathways which are often dis-regulated in cancer. The majority of signatures were generated directly from microarray data from NCBI GEO or from internal unpublished profiling experiments involving perturbation of known cancer genes. details	Download GMT Files gene symbols entrez genes ids
C7: immunologic signatures (browse 4872 gene sets)	Gene sets that represent cell states and perturbations within the immune system. The signatures were generated by manual curation of published studies in human and mouse immunology. details	Download GMT Files gene symbols entrez genes ids

But pathways in gene set databases are gene-sets

A pathway



A gene set



What have we learned today?

What are biological pathways
Where and how to find biological pathways
Pathway database formats and identifiers
How to use the Gene Ontology
What are the main Gene Set databases


MORA LAB
FUNCTIONAL BIOINFORMATICS

EXTRA. Automatic reconstruction of pathways

Final remark: Automatic reconstruction of pathways

Pathway databases follow two main strategies: Either a curator team, such as in KEGG or Reactome, or open to public submission, such as in Wikipathways.

However, there are huge amounts of pathway information in the scientific literature that would take many years to human beings to process it. Therefore, we need **text mining** methodologies to automatically extract pathway knowledge from the literature.

Final remark: Automatic reconstruction of pathways

One example of this is **MELODI**, a text mining tool that extracts mechanisms of disease based on subject-predicate-object triples from **SemMedDB** (Semantic Medline Database).

For example, the sentence "*We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia*" produces the following four triples:

- Hemofiltration-TREATS-Patients
- Digoxin overdose-PROCESS_OF-Patients
- Hyperkalemia-COMPLICATES-Digoxin overdose
- INFERENCE: Hemofiltration-TREATS-Digoxin overdose

Final remark: Automatic reconstruction of pathways

Building a database of triples for all PubMed, we can let computers link information from different papers and reconstruct the pathway for us!

Final remark: Automatic reconstruction of pathways