ClueGO

Documentation

Gabriela Bindea and Bernhard Mlecnik
Laboratory of Integrative Cancer Immunology
UMRS1138 Cordeliers Research Center, Paris,
France

Contents

ClueGO Cytoscape App

Cytoscape is a major computational platform for the analysis and visualization of biological networks [1]. Apps with specific functionality extend Cytoscape and are available at the Cytoscape App Store [2].

ClueGO is a Cytoscape App that facilitates the biological interpretation of large lists of genes and proteins by selecting representative Gene Onology terms and pathways from multiple ontologies and visualizes them into functionally organized networks. ClueGO is extended by CluePedia, that allows a detailed pathways analysis.

ClueGO supports many organisms. Human and mouse data are included by default in the App, and more than 200 organisms are available for download.

ClueGO and CluePedia should be downloaded from the Cytoscape App Store through the Cytoscape App Manager (Cytoscape menu, Apps → ClueGO).

ClueGO v2.0.0 for Cytoscape 3.0.0 was released in 13th of December 2012. The functionality of ClueGO earlier described is valid in versions ported to Cytoscape 3.

Installation

System Requirements:

- Windows, Linux, Unix or MacOS operating system.
- 16 GB RAM recommended. Hard disk with at least 100 MB free (for example files).
- Java 1.8+ needed.
- Cytoscape 3.7.+ installed: https://cytoscape.org/

At the first startup, the ClueGOConfiguration folder containing precompiled ClueGO files and sample files will be created in the user home folder. If this folder is removed or the content damaged, it will be recreated automatically at the next startup.

License

The authors wish to make ClueGO available on a nonexclusive basis to interested parties for noncommercial internal research purposes.

ClueGO is available free of charge only to academic, government, and other nonprofit institutions for noncommercial, nonprofit internal research purposes. Note that the license terms specifically limit its use to such purposes. License applications are reviewed manually and an email including the license key will be sent to you as soon as the evaluation process has been completed.

To obtain a license key please visit: http://www.ici.upmc.fr/cluego/cluegoLicense.shtml.

Documentation

The biological interpretation of large gene clusters derived from high-throughput experiments is a real challenge. Many ontology sources exist in order to capture biological information in a meaningful way.

Annotation and ontology sources

The Gene Ontology (GO) project [3] aims to capture the increasing knowledge on gene function in a controlled vocabulary applicable to all the organisms. GO describes gene products in terms of their associated biological processes, cellular components and molecular functions, terms that have a hierarchical relationship (parent-child) (Figure 1).

Due to the complexity of hierarchy structure (directed acyclic graph), the terms can be found in several different levels. The specificity of the terms varies along the tree: from very general terms (in first levels of GO) to very specific ones.

Kyoto Encyclopedia of Genes and Genomes (KEGG) [4] is a database of biological systems that integrates genomic, chemical and systemic functional information in pathways.

Reactome is a free, open-source, curated and peer-reviewed pathway database that provides intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway

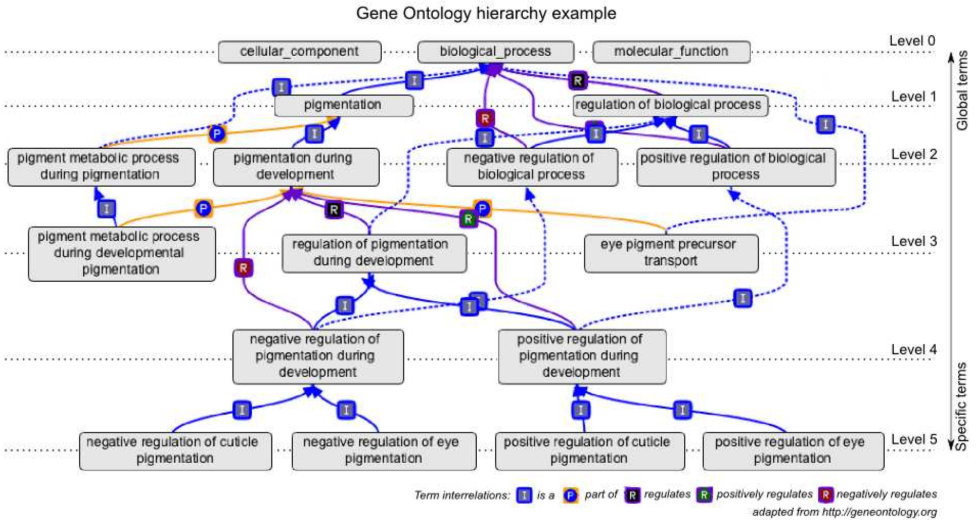

Gene Ontology hierarchy example
cellular_component
biological_process
molecular_function
Level 0
pigmentation
regulation of biological process
Level 1
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
Level 2
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
Level 3
negative regulation of pigmentation during development
positive regulation of pigmentation during development
Level 4
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation
Level 5
Global terms
Specific terms
Term interrelations: I is a P part of R regulates R positively regulates R negatively regulates
adapted from http://geneontology.org

Figure 1: GO hierarchy example

knowledge to support basic research, genome analysis, modeling, systems biology and education [5].

WikiPathways is an open, collaborative platform dedicated to the curation of biological pathways [6].

The Planteome is a resource for common reference ontologies and applications for plant biology [7].

Ensembl [8] is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

The National Center for Biotechnology Information [9] advances science and health by providing access to biomedical and genomic information.

InterPro [10] provides functional analysis of proteins by classifying them into families and predicting domains and important sites. Protein signatures from a number of member databases are combined into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

STRING [11] is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases.

The CORUM [12] database is a collection of experimentally verified mammalian protein complexes, key molecular entities that integrate multiple gene products to perform cellular functions.

IntAct [13] provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available.

BioCyc [14] is a collection of 14560 Pathway/Genome Databases (PGDBs), plus software tools for exploring them.

For a complete view on the studied process, several ontology sources should be consulted in order to integrate their complementary information. For each gene there is a large amount of information in each of these ontology sources. This makes the analysis of the relationship between genes and between terms very difficult to represent and comprehend. Also, for close related terms, a high degree of redundancy of their associated genes exists.

ClueGO, an open-source Java tool was design to extract the non-redundant biological information for large clusters of genes, using multiple ontologies. ClueGO is a Cytoscape [1] App and is taking advantage of its complex visualization environment.

ClueGO features

1. ClueGO allows the analysis of a single marker list (cluster) or compares multiple lists with markers.

2. ClueGO supports many organisms. Human and mouse are included in ClueGO by default and more than 200 other organisms are available for download.

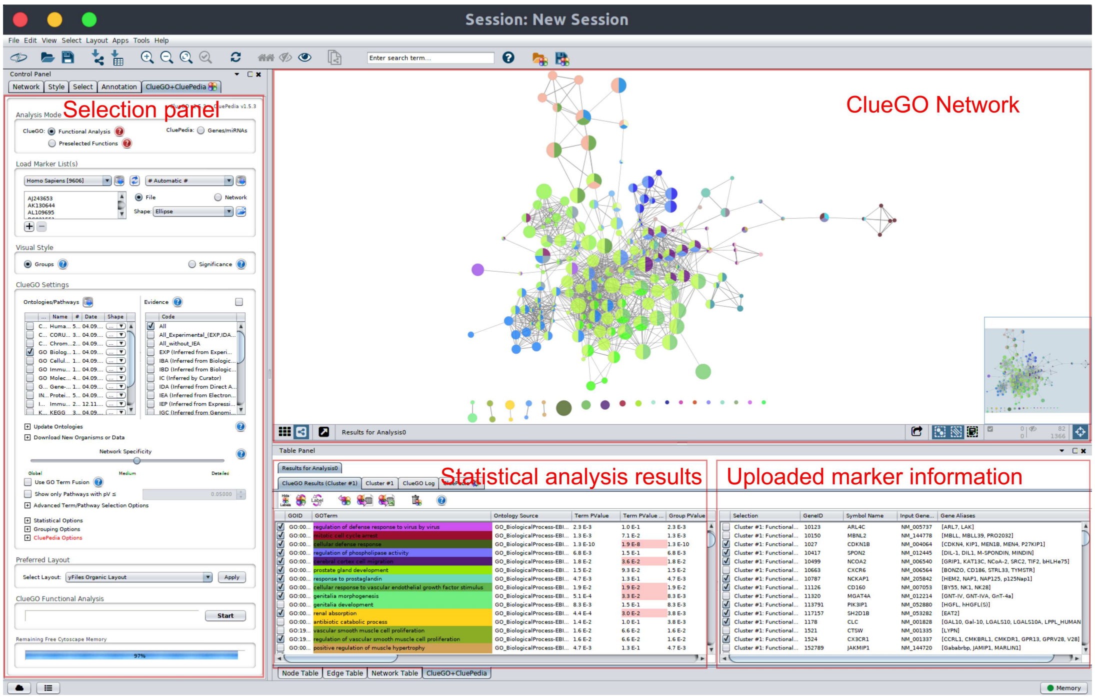
3. ClueGO recognizes automatically multiple gene and protein identifiers.

4. Identifiers can be pasted, uploaded from lists or from existing networks.

5. ClueGO allows the simultaneous analysis of multiple annotation and ontology sources.
6. Users of ClueGO can automatically update the ontology sources.
7. Predefined criteria for the selection of terms/pathways are provided. These criteria were defined using lists of 200 genes and can be customized.
8. Multiple statistical methods for the enrichment calculation are provided.
9. Results are illustrated as a functionally grouped network, as plots and tables.
10. Results are automatically mapped on the network in different visual styles.
11. The results and the entire project can be saved.
12. Selected terms and pathways are included in the network and are linked based on the kappa score that shows how similar are their associated genes.
13. Multiple visualization styles of the network are available.
14. To reduce the redundancy of GO terms can be applied the fusion of related terms with similar associated genes.
15. ClueGO can visualize also preselected GO terms and pathways.
16. ClueGO functionality is REST enabled, and can be accessed from R, Python, etc.
17. ClueGO can be used in combination with CluePedia to get detailed information of pathways.
18. Easy extendable.

ClueGO overview

The selection panel includes ClueGO features for enrichment analyses. The results of the enrichment analysis are shown as a functionally grouped network, as tables and graphs. A summary of uploaded markers can be found under the network.


Session: New Session
File Edit View Select Layout Apps Tools Help
Enter search term...
Control Panel
Network | Style | Select | Annotation | ClueGO+CluePedia
Selection panel
ClueGO Network
Statistical analysis results
Uploaded marker information
Results for Analysis0
Table Panel
Node Table | Edge Table | Network Table | ClueGO+CluePedia
Memory

ClueGO v2.5.3 + CluePedia v1.5.3

Analysis Mode
ClueGO: ● Functional Analysis
○ Preselected Functions
CluePedia: ○ Genes/miRNAs

Load Marker List(s)
Homo Sapiens [9606] | # Automatic #
AJ243653
AK130644
AL109695
● File ○ Network
Shape: Ellipse

Visual Style
● Groups ○ Significance

ClueGO Settings

Ontologies/Pathways

	...	Name	#	Date	Shape
□	C...	Huma...	5...	04.09....	...
□	C...	CORU...	3...	04.09....	...
□	C...	Chrom...	2...	04.09....	...
☑	GO	Biolog...	1...	04.09....	...
□	GO	Cellul...	1...	04.09....	...
□	GO	Immu...	1...	04.09....	...
□	GO	Molec...	4...	04.09....	...
□	G...	Gene-...	1...	04.09....	...
□	IN...	Protei...	5...	04.09....	...
□	I...	Immu...	2...	12.11....	...
□	K...	KEGG	3...	04.09....	...

Evidence □

	Code
☑	All
□	All_Experimental_(EXP,IDA...
□	All_without_IEA
□	EXP (Inferred from Experi...
□	IBA (Inferred from Biologic...
□	IBD (Inferred from Biologic...
□	IC (Inferred by Curator)
□	IDA (Inferred from Direct A...
□	IEA (Inferred from Electron...
□	IEP (Inferred from Expressi...
□	IGC (Inferred from Genomi...

⊞ Update Ontologies
⊞ Download New Organisms or Data

Network Specificity
Global — Medium — Detailed

□ Use GO Term Fusion
□ Show only Pathways with pV ≤ 0.05000
⊞ Advanced Term/Pathway Selection Options
⊞ Statistical Options
⊞ Grouping Options
⊞ CluePedia Options

Preferred Layout
Select Layout: yFiles Organic Layout | Apply

ClueGO Functional Analysis
Start

Remaining Free Cytoscape Memory
97%

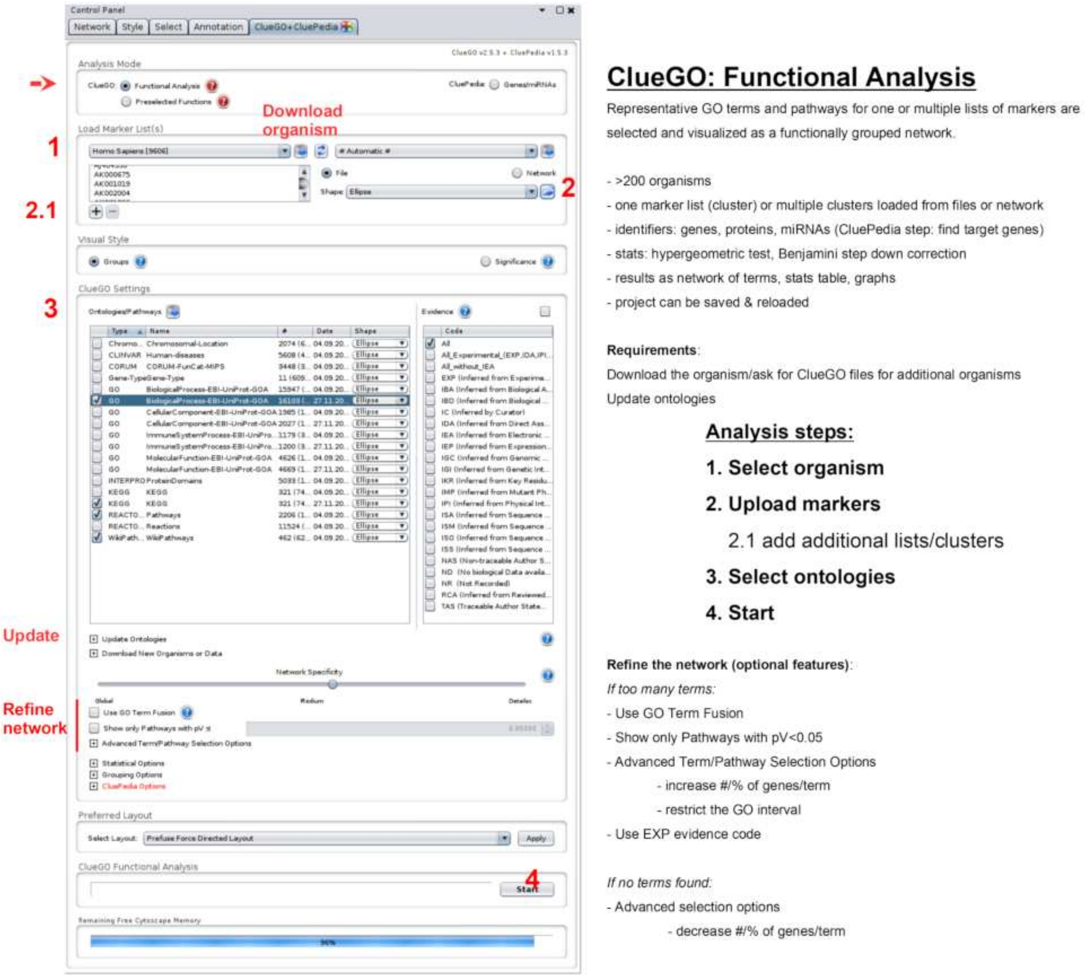
Results for Analysis0
ClueGO Results (Cluster #1) | Cluster #1 | ClueGO Log

	GOID	GOTerm	Ontology Source	Term PValue	Term PValue ...	Group PValue
☑	GO:00...	regulation of defense response to virus by virus	GO_BiologicalProcess-EBI...	2.3 E-3	1.0 E-1	2.3 E-3
☑	GO:00...	mitotic cell cycle arrest	GO_BiologicalProcess-EBI...	1.3 E-3	7.1 E-2	1.3 E-3
☑	GO:00...	cellular defense response	GO_BiologicalProcess-EBI...	1.3 E-10	1.9 E-8	1.3 E-10
☑	GO:00...	regulation of phospholipase activity	GO_BiologicalProcess-EBI...	6.8 E-3	1.5 E-1	6.8 E-3
☑	GO:00...	cerebral cortex cell migration	GO_BiologicalProcess-EBI...	1.8 E-2	3.6 E-2	1.8 E-2
☑	GO:00...	prostate gland development	GO_BiologicalProcess-EBI...	1.5 E-2	9.3 E-2	1.5 E-2
☑	GO:00...	response to prostaglandin	GO_BiologicalProcess-EBI...	4.7 E-3	1.3 E-1	4.7 E-3
☑	GO:00...	cellular response to vascular endothelial growth factor stimulus	GO_BiologicalProcess-EBI...	1.9 E-2	1.9 E-2	1.9 E-2
☑	GO:00...	genitalia morphogenesis	GO_BiologicalProcess-EBI...	5.1 E-4	3.3 E-2	8.3 E-3
□	GO:00...	genitalia development	GO_BiologicalProcess-EBI...	8.3 E-3	1.5 E-1	8.3 E-3
☑	GO:00...	renal absorption	GO_BiologicalProcess-EBI...	4.4 E-4	3.0 E-2	3.8 E-3
□	GO:00...	antibiotic catabolic process	GO_BiologicalProcess-EBI...	1.4 E-2	1.0 E-1	3.8 E-3
□	GO:19...	vascular smooth muscle cell proliferation	GO_BiologicalProcess-EBI...	1.6 E-2	6.6 E-2	1.6 E-2
☑	GO:19...	regulation of vascular smooth muscle cell proliferation	GO_BiologicalProcess-EBI...	1.6 E-2	6.6 E-2	1.6 E-2
□	GO:00...	positive regulation of muscle hypertrophy	GO_BiologicalProcess-EBI...	4.7 E-3	1.3 E-1	4.7 E-3

	Selection	GeneID	Symbol Name	Input Gene...	Gene Aliases
□	Cluster #1: Functional...	10123	ARL4C	NM_005737	[ARL7, LAK]
□	Cluster #1: Functional...	10150	MBNL2	NM_144778	[MBLL, MBLL39, PRO2032]
☑	Cluster #1: Functional...	1027	CDKN1B	NM_004064	[CDKN4, KIP1, MEN1B, MEN4, P27KIP1]
☑	Cluster #1: Functional...	10417	SPON2	NM_012445	[DIL-1, DIL1, M-SPONDIN, MINDIN]
☑	Cluster #1: Functional...	10499	NCOA2	NM_006540	[GRIP1, KAT13C, NCoA-2, SRC2, TIF2, bHLHe75]
□	Cluster #1: Functional...	10663	CXCR6	NM_006564	[BONZO, CD186, STRL33, TYMSTR]
☑	Cluster #1: Functional...	10787	NCKAP1	NM_205842	[HEM2, NAP1, NAP125, p125Nap1]
☑	Cluster #1: Functional...	11126	CD160	NM_007053	[BY55, NK1, NK28]
□	Cluster #1: Functional...	11320	MGAT4A	NM_012214	[GNT-IV, GNT-IVA, GnT-4a]
☑	Cluster #1: Functional...	113791	PIK3IP1	NM_052880	[HGFL, hHGFL(S)]
☑	Cluster #1: Functional...	117157	SH2D1B	NM_053282	[EAT2]
☑	Cluster #1: Functional...	1178	CLC	NM_001828	[GAL10, Gal-10, LGALS10, LGALS10A, LPPL_HUMAN
□	Cluster #1: Functional...	1521	CTSW	NM_001335	[LYPN]
☑	Cluster #1: Functional...	1524	CX3CR1	NM_001337	[CCRL1, CMKBRL1, CMKDR1, GPR13, GPRV28, V28]
□	Cluster #1: Functional...	152789	JAKMIP1	NM_144720	[Gababrbp, JAMIP1, MARLIN1]

Figure 2: *ClueGO overview*

ClueGO analysis steps



ClueGO: Functional Analysis

Representative GO terms and pathways for one or multiple lists of markers are selected and visualized as a functionally grouped network.

- >200 organisms
- one marker list (cluster) or multiple clusters loaded from files or network
- identifiers: genes, proteins, miRNAs (CluePedia step: find target genes)
- stats: hypergeometric test, Benjamini step down correction
- results as network of terms, stats table, graphs
- project can be saved & reloaded

Requirements:

Download the organism/ask for ClueGO files for additional organisms

Update ontologies

Analysis steps:

1. Select organism

2. Upload markers

2.1 add additional lists/clusters

3. Select ontologies

4. Start

Refine the network (optional features):

If too many terms:

- Use GO Term Fusion
- Show only Pathways with pV<0.05
- Advanced Term/Pathway Selection Options
  - increase #/% of genes/term
  - restrict the GO interval
- Use EXP evidence code

If no terms found:

- Advanced selection options
  - decrease #/% of genes/term

Figure 3: *ClueGO analysis steps*

Requirements:

Download the organism or ask for additional organisms. If needed, update ontologies.

Analysis steps:

1. Select the organism

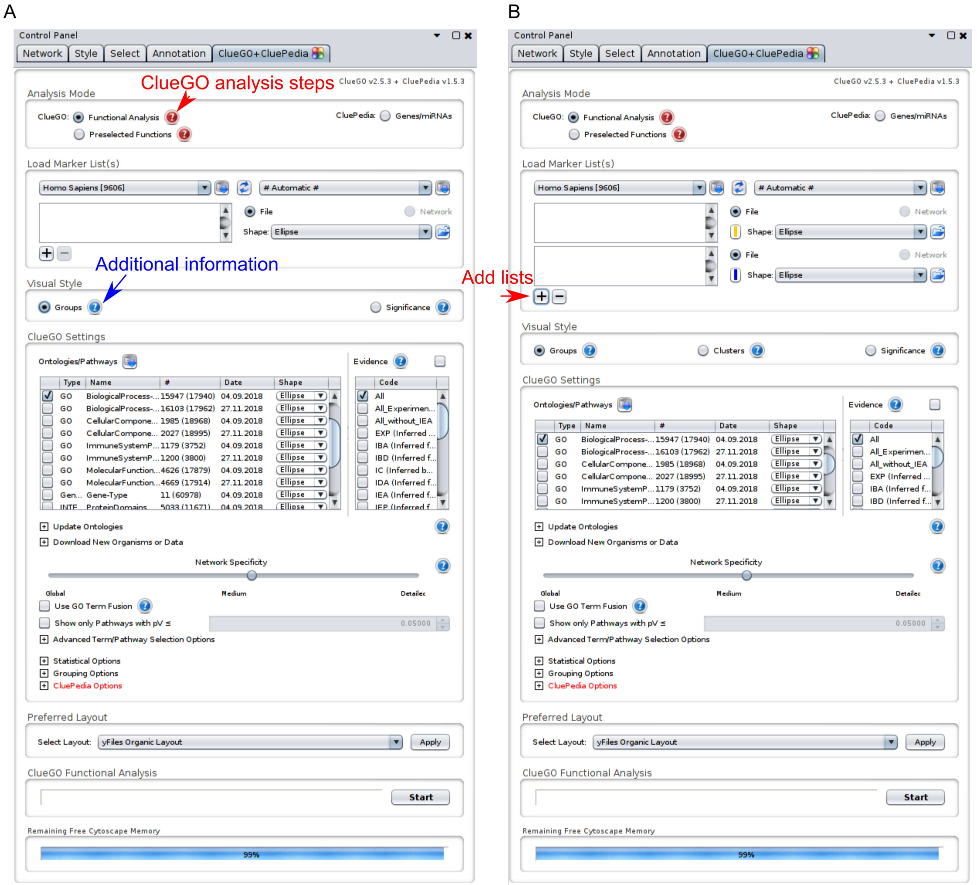
2. Upload markers. Optional (2.1) Add additional lists

3. Select ontologies

4. Start the analysis

ClueGO analysis with one or multiple lists with markers

After starting Cytoscape, ClueGO can be found in the Cytoscape menu/Apps. Once selected, it will display the ClueGO selection panel options (Figure 4). ClueGO provides short explanations for each feature (click on help icons).



ClueGO: ● Functional Analysis ○ Preselected Functions

CluePedia: ○ Genes/miRNAs

Homo Sapiens [9606] · # Automatic #

● File ○ Network · Shape: Ellipse

Visual Style (A): ● Groups ○ Significance

Visual Style (B): ● Groups ○ Clusters ○ Significance

	Type	Name	#	Date	Shape
☑	GO	BiologicalProcess-...	15947 (17940)	04.09.2018	Ellipse
□	GO	BiologicalProcess-...	16103 (17962)	27.11.2018	Ellipse
□	GO	CellularCompone...	1985 (18968)	04.09.2018	Ellipse
□	GO	CellularCompone...	2027 (18995)	27.11.2018	Ellipse
□	GO	ImmuneSystemP...	1179 (3752)	04.09.2018	Ellipse
□	GO	ImmuneSystemP...	1200 (3800)	27.11.2018	Ellipse
□	GO	MolecularFunction...	4626 (17879)	04.09.2018	Ellipse
□	GO	MolecularFunction...	4669 (17914)	27.11.2018	Ellipse
□	Gen...	Gene-Type	11 (60978)	04.09.2018	Ellipse
□	INTE	ProteinDomains	5033 (11671)	04.09.2018	Ellipse

	Code
☑	All
□	All_Experimen...
□	All_without_IEA
□	EXP (Inferred ...
□	IBA (Inferred f...
□	IBD (Inferred f...
□	IC (Inferred b...
□	IDA (Inferred f...
□	IEA (Inferred f...
□	IEP (Inferred f

□ Use GO Term Fusion

□ Show only Pathways with pV ≤ 0.05000

Select Layout: yFiles Organic Layout

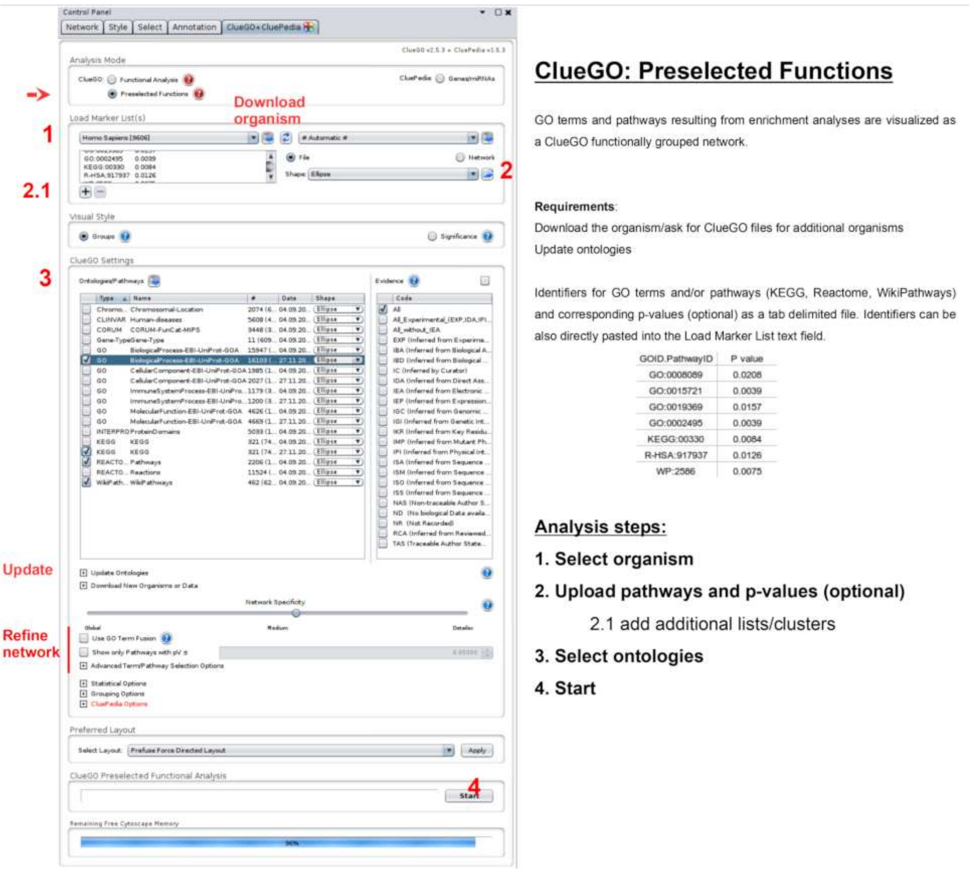
Figure 4: *ClueGO selection Panel with one (A) or multiple (B) lists*

The user can choose between a ClueGO functional analysis, a ClueGO like visualization of preselected terms or to run CluePedia for a detailed analysis of pathways (Analysis Mode section). ClueGO analysis steps are summarized (click on the help icon).

Additional lists can be added by selecting on ”+” in Load Marker Lists(s) section. A color is automatically attributed to each uploaded list.

ClueGO visualization of preselected GO terms and pathways

GO terms and pathways resulted from enrichments can be visualized with ClueGO. Upload your file with pathways and Pvalues obtained from the enrichment analysis, mapp the terms on ontologies and visualized them into a ClueGO funtionally grouped network.



ClueGO: Preselected Functions

GO terms and pathways resulting from enrichment analyses are visualized as a ClueGO functionally grouped network.

Requirements:
Download the organism/ask for ClueGO files for additional organisms
Update ontologies

Identifiers for GO terms and/or pathways (KEGG, Reactome, WikiPathways) and corresponding p-values (optional) as a tab delimited file. Identifiers can be also directly pasted into the Load Marker List text field.

GOID.PathwayID	P value
GO:0008089	0.0208
GO:0015721	0.0039
GO:0019369	0.0157
GO:0002495	0.0039
KEGG:00330	0.0084
R-HSA:917937	0.0126
WP:2586	0.0075

Analysis steps:

1. Select organism

2. Upload pathways and p-values (optional)

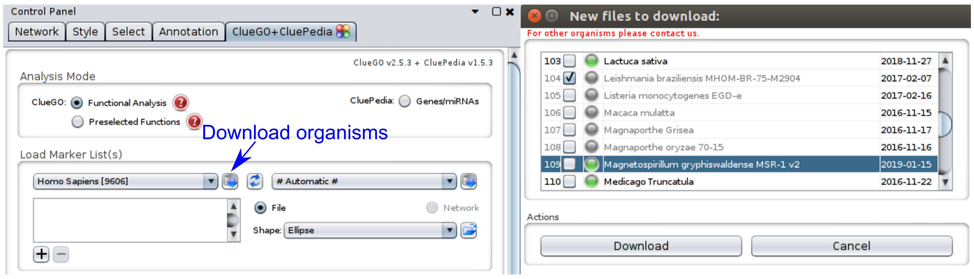
2.1 add additional lists/clusters

3. Select ontologies

4. Start

Figure 5: *ClueGO pre selected pathways*

ClueGO supported organisms



ClueGO: ● Functional Analysis ○ Preselected Functions

CluePedia: ○ Genes/miRNAs

Homo Sapiens [9606]

Automatic

● File ○ Network

Shape: Ellipse

For other organisms please contact us.

103	□	●	Lactuca sativa	2018-11-27
104	☑	●	Leishmania braziliensis MHOM-BR-75-M2904	2017-02-07
105	□	●	Listeria monocytogenes EGD-e	2017-02-16
106	□	●	Macaca mulatta	2016-11-15
107	□	●	Magnaporthe Grisea	2016-11-17
108	□	●	Magnaporthe oryzae 70-15	2016-11-16
109	□	●	Magnetospirillum gryphiswaldense MSR-1 v2	2019-01-15
110	□	●	Medicago Truncatula	2016-11-22

Figure 6: *ClueGO supported organisms*

Human and mouse data are included in ClueGO. More than 200 other organisms can be downloaded (Figure 6). Downloaded organisms are shown in green. Select a new organism and click Download to install it.

Additional organisms can be added upon request (contact the authors). The taxonomy identifier and functional annotations: genes/proteins and their corresponding GO terms are needed. This data should be GAF or tab delimited.

Functional annotations for non-model organism sequences can be obtained using annotation software like: EggNOG mapper [15].

ClueGO predefined selection criteria of GO terms and pathways

ClueGO provides predefined selection criteria of representative pathways defined using lists of around 200 genes. These criteria include the 3-8 GO tree interval, a minimum of 3 genes from the uploaded list found to be associated to a term, and that these genes represent at least 4% from the total number of associated genes.

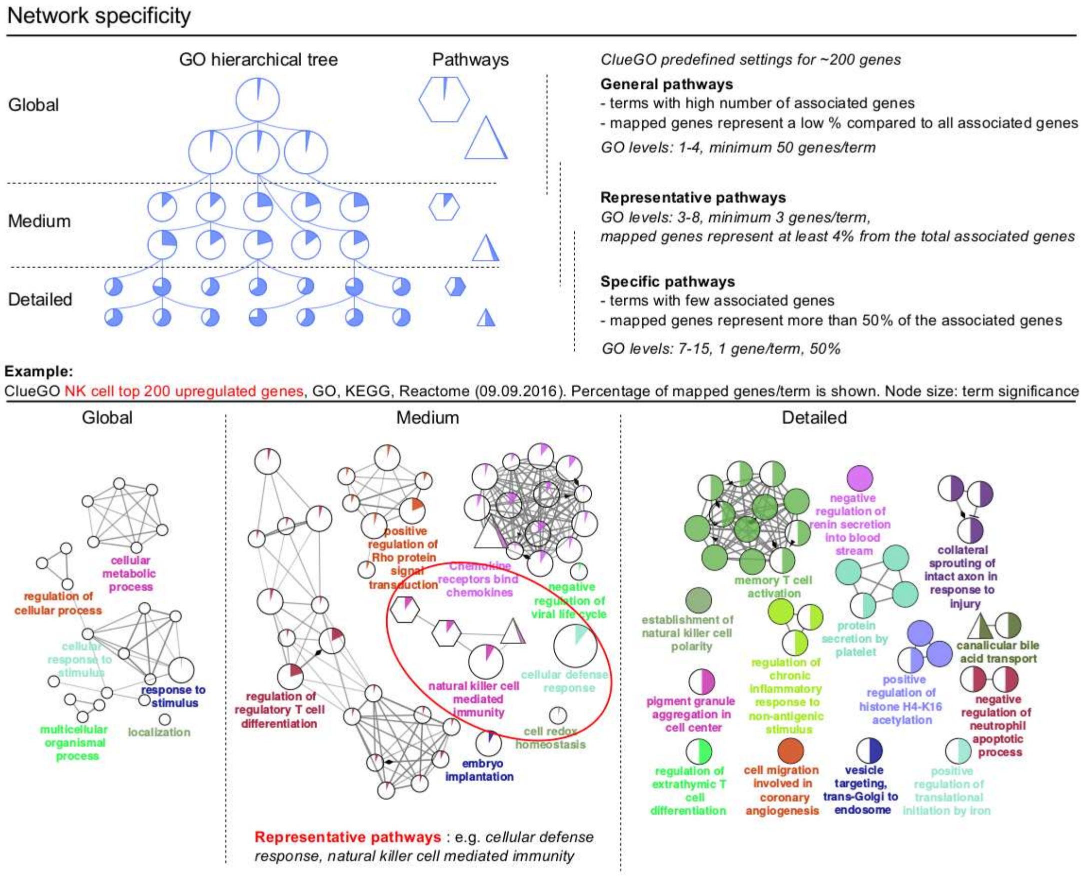
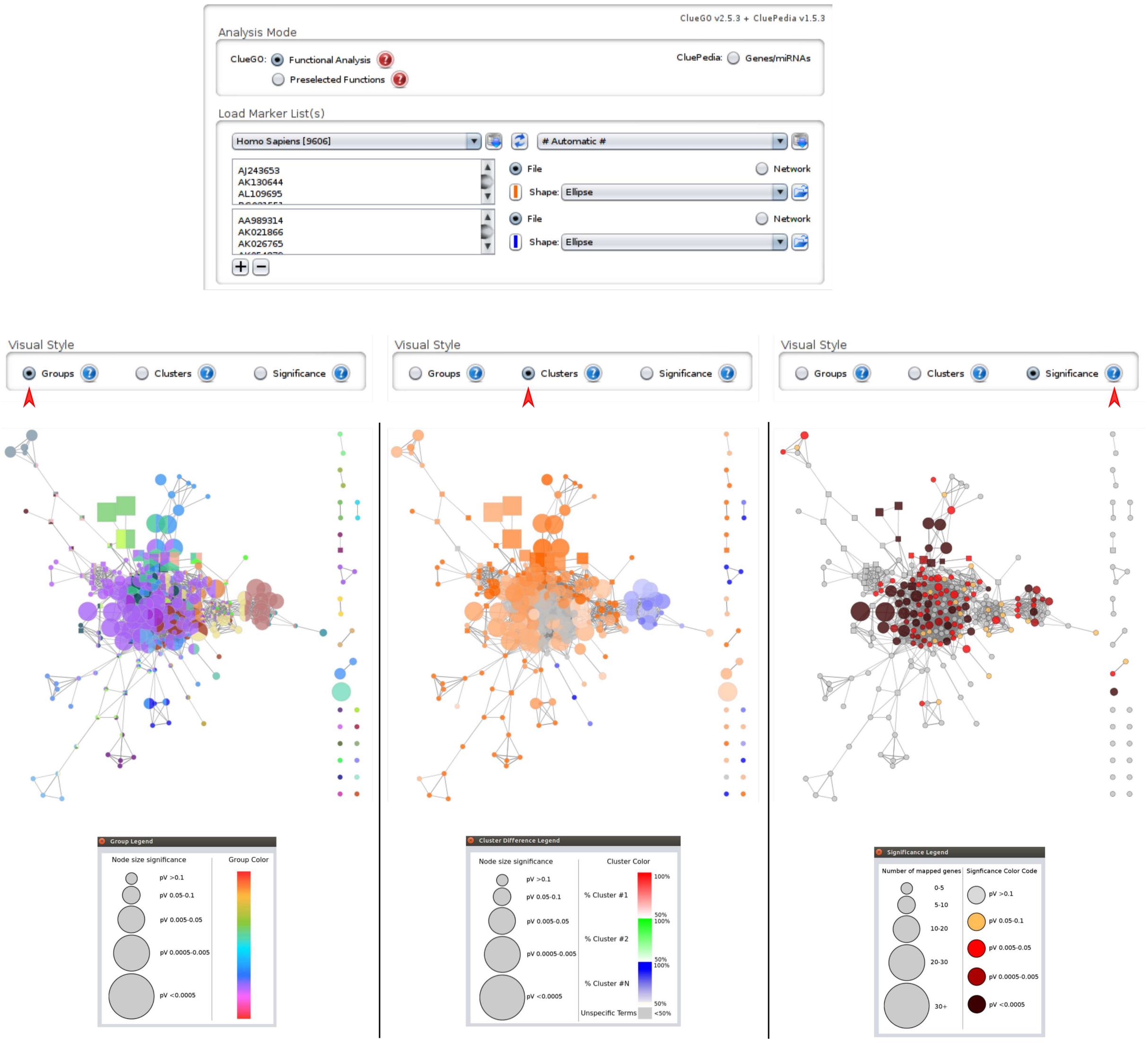

Network specificity
GO hierarchical tree
Pathways
Global
Medium
Detailed
ClueGO predefined settings for ~200 genes
General pathways
- terms with high number of associated genes
- mapped genes represent a low % compared to all associated genes
GO levels: 1-4, minimum 50 genes/term
Representative pathways
GO levels: 3-8, minimum 3 genes/term,
mapped genes represent at least 4% from the total associated genes
Specific pathways
- terms with few associated genes
- mapped genes represent more than 50% of the associated genes
GO levels: 7-15, 1 gene/term, 50%
Example:
ClueGO NK cell top 200 upregulated genes, GO, KEGG, Reactome (09.09.2016). Percentage of mapped genes/term is shown. Node size: term significance
Global
Medium
Detailed
cellular metabolic process
regulation of cellular process
cellular response to stimulus
response to stimulus
multicellular organismal process
localization
positive regulation of Rho protein signal transduction
Chemokine receptors bind chemokines
negative regulation of viral life cycle
regulation of regulatory T cell differentiation
natural killer cell mediated immunity
cellular defense response
cell redox homeostasis
embryo implantation
Representative pathways : e.g. cellular defense response, natural killer cell mediated immunity
negative regulation of renin secretion into blood stream
collateral sprouting of intact axon in response to injury
memory T cell activation
establishment of natural killer cell polarity
protein secretion by platelet
canalicular bile acid transport
regulation of chronic inflammatory response to non-antigenic stimulus
positive regulation of histone H4-K16 acetylation
pigment granule aggregation in cell center
negative regulation of neutrophil apoptotic process
regulation of extrathymic T cell differentiation
cell migration involved in coronary angiogenesis
vesicle targeting, trans-Golgi to endosome
positive regulation of translational initiation by iron

Figure 7: *ClueGO predefined selection criteria*

ClueGO visual styles, multiple marker lists



ClueGO v2.5.3 + CluePedia v1.5.3

Analysis Mode

ClueGO: ● Functional Analysis
○ Preselected Functions

CluePedia: ○ Genes/miRNAs

Load Marker List(s)

Homo Sapiens [9606]

Automatic

AJ243653
AK130644
AL109695

● File ○ Network
Shape: Ellipse

AA989314
AK021866
AK026765

● File ○ Network
Shape: Ellipse

Visual Style: ● Groups ○ Clusters ○ Significance

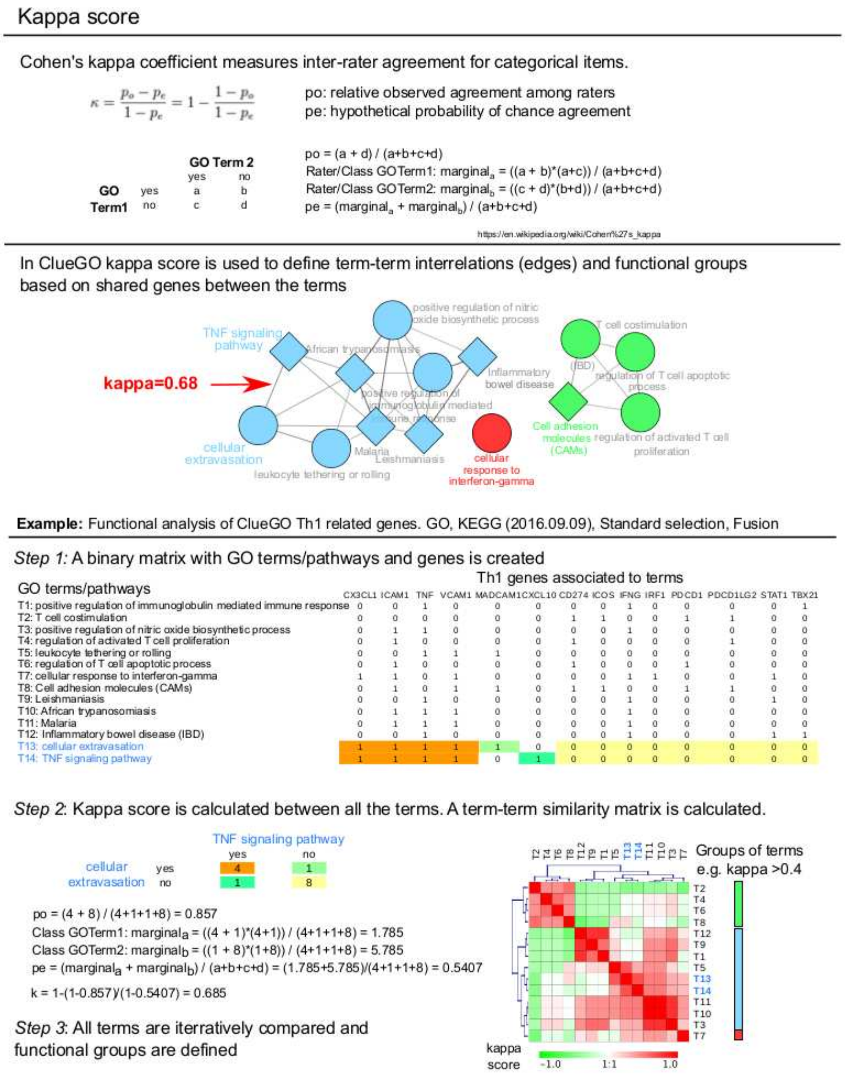
Visual Style: ○ Groups ● Clusters ○ Significance

Visual Style: ○ Groups ○ Clusters ● Significance

Figure 8: *ClueGO visual styles*

Kappa score

In ClueGO the Kappa score is used to define term-term interractions shown as edges on the network.



Kappa score

Cohen's kappa coefficient measures inter-rater agreement for categorical items.

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

po: relative observed agreement among raters
pe: hypothetical probability of chance agreement

		GO Term 2 yes	GO Term 2 no
GO Term1	yes	a	b
	no	c	d

po = (a + d) / (a+b+c+d)
Rater/Class GOTerm1: marginal_a = ((a + b)*(a+c)) / (a+b+c+d)
Rater/Class GOTerm2: marginal_b = ((c + d)*(b+d)) / (a+b+c+d)
pe = (marginal_a + marginal_b) / (a+b+c+d)

https://en.wikipedia.org/wiki/Cohen%27s_kappa

In ClueGO kappa score is used to define term-term interrelations (edges) and functional groups based on shared genes between the terms

TNF signaling pathway
African trypanosomiasis
positive regulation of nitric oxide biosynthetic process
kappa=0.68
Inflammatory bowel disease
positive regulation of immunoglobulin mediated immune response
cellular extravasation
Malaria
Leishmaniasis
leukocyte tethering or rolling
cellular response to interferon-gamma
T cell costimulation
(IBD)
regulation of T cell apoptotic process
Cell adhesion molecules (CAMs)
regulation of activated T cell proliferation

Example: Functional analysis of ClueGO Th1 related genes. GO, KEGG (2016.09.09), Standard selection, Fusion

Step 1: A binary matrix with GO terms/pathways and genes is created

GO terms/pathways	CX3CL1	ICAM1	TNF	VCAM1	MADCAM1	CXCL10	CD274	ICOS	IFNG	IRF1	PDCD1	PDCD1LG2	STAT1	TBX21
T1: positive regulation of immunoglobulin mediated immune response	0	0	1	0	0	0	0	0	1	0	0	0	0	1
T2: T cell costimulation	0	0	0	0	0	0	1	1	0	0	1	1	0	0
T3: positive regulation of nitric oxide biosynthetic process	0	1	1	0	0	0	0	0	1	0	0	0	0	0
T4: regulation of activated T cell proliferation	0	1	0	0	0	0	1	0	0	0	0	1	0	0
T5: leukocyte tethering or rolling	0	0	1	1	1	0	0	0	0	0	0	0	0	0
T6: regulation of T cell apoptotic process	0	1	0	0	0	0	1	0	0	0	1	0	0	0
T7: cellular response to interferon-gamma	1	1	0	1	0	0	0	0	1	1	0	0	1	0
T8: Cell adhesion molecules (CAMs)	0	1	0	1	1	0	1	1	0	0	1	1	0	0
T9: Leishmaniasis	0	0	1	0	0	0	0	0	1	0	0	0	1	0
T10: African trypanosomiasis	0	1	1	1	0	0	0	0	1	0	0	0	0	0
T11: Malaria	0	1	1	1	0	0	0	0	1	0	0	0	0	0
T12: Inflammatory bowel disease (IBD)	0	0	1	0	0	0	0	0	1	0	0	0	1	1
T13: cellular extravasation	1	1	1	1	1	0	0	0	0	0	0	0	0	0
T14: TNF signaling pathway	1	1	1	1	0	1	0	0	0	0	0	0	0	0

Step 2: Kappa score is calculated between all the terms. A term-term similarity matrix is calculated.

		TNF signaling pathway yes	TNF signaling pathway no
cellular extravasation	yes	4	1
	no	1	8

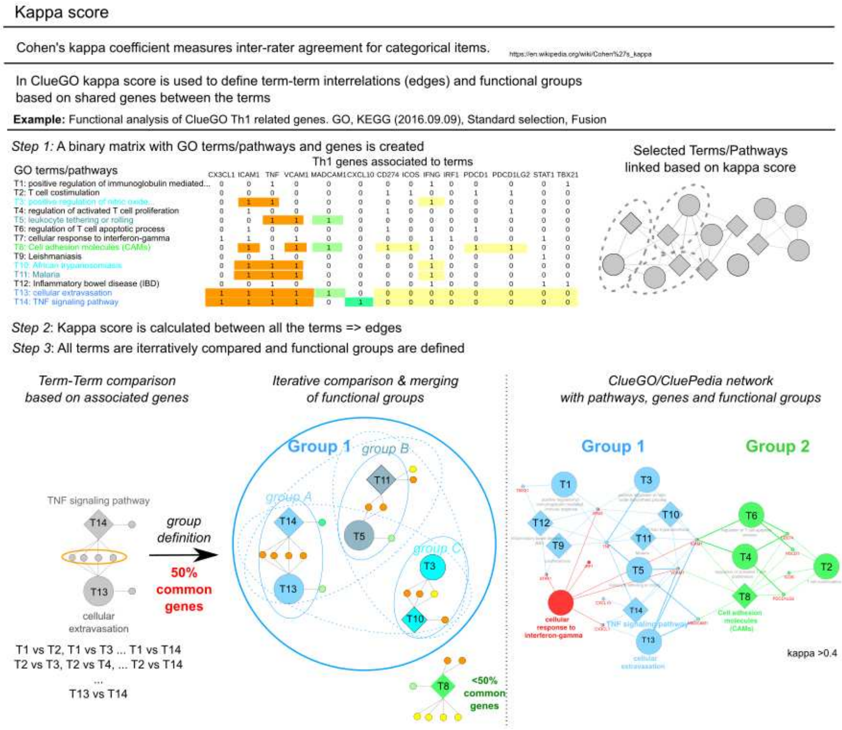
po = (4 + 8) / (4+1+1+8) = 0.857
Class GOTerm1: marginal_a = ((4 + 1)*(4+1)) / (4+1+1+8) = 1.785
Class GOTerm2: marginal_b = ((1 + 8)*(1+8)) / (4+1+1+8) = 5.785
pe = (marginal_a + marginal_b) / (a+b+c+d) = (1.785+5.785)/(4+1+1+8) = 0.5407

k = 1-(1-0.857)/(1-0.5407) = 0.685

Step 3: All terms are iterratively compared and functional groups are defined

Figure 9: *Kappa score*

Kappa score is used also to associate terms and pathways into functional groups based on shared genes.



Kappa score

Cohen's kappa coefficient measures inter-rater agreement for categorical items. https://en.wikipedia.org/wiki/Cohen%27s_kappa

In ClueGO kappa score is used to define term-term interrelations (edges) and functional groups based on shared genes between the terms

Example: Functional analysis of ClueGO Th1 related genes. GO, KEGG (2016.09.09), Standard selection, Fusion

Step 1: A binary matrix with GO terms/pathways and genes is created

Th1 genes associated to terms

GO terms/pathways	CX3CL1	ICAM1	TNF	VCAM1	MADCAM1	CXCL10	CD274	ICOS	IFNG	IRF1	PDCD1	PDCD1LG2	STAT1	TBX21
T1: positive regulation of immunoglobulin mediated...	0	0	1	0	0	0	0	0	1	0	0	0	0	1
T2: T cell costimulation	0	0	0	0	0	0	1	1	0	0	1	1	0	0
T3: positive regulation of nitric oxide...	0	1	1	0	0	0	0	0	1	0	0	0	0	0
T4: regulation of activated T cell proliferation	0	1	0	0	0	0	1	0	0	0	0	1	0	0
T5: leukocyte tethering or rolling	0	0	1	1	1	0	0	0	0	0	0	0	0	0
T6: regulation of T cell apoptotic process	0	1	0	0	0	0	1	0	0	0	1	0	0	0
T7: cellular response to interferon-gamma	1	1	0	1	0	0	0	0	1	1	0	0	1	0
T8: Cell adhesion molecules (CAMs)	0	1	0	1	1	0	1	1	0	0	1	1	0	0
T9: Leishmaniasis	0	0	1	0	0	0	0	0	1	0	0	0	1	0
T10: African trypanosomiasis	0	1	1	1	0	0	0	0	1	0	0	0	0	0
T11: Malaria	0	1	1	1	0	0	0	0	1	0	0	0	0	0
T12: Inflammatory bowel disease (IBD)	0	0	1	0	0	0	0	0	1	0	0	0	1	1
T13: cellular extravasation	1	1	1	1	1	0	0	0	0	0	0	0	0	0
T14: TNF signaling pathway	1	1	1	1	0	1	0	0	0	0	0	0	0	0

Selected Terms/Pathways linked based on kappa score

Step 2: Kappa score is calculated between all the terms => edges

Step 3: All terms are iterratively compared and functional groups are defined

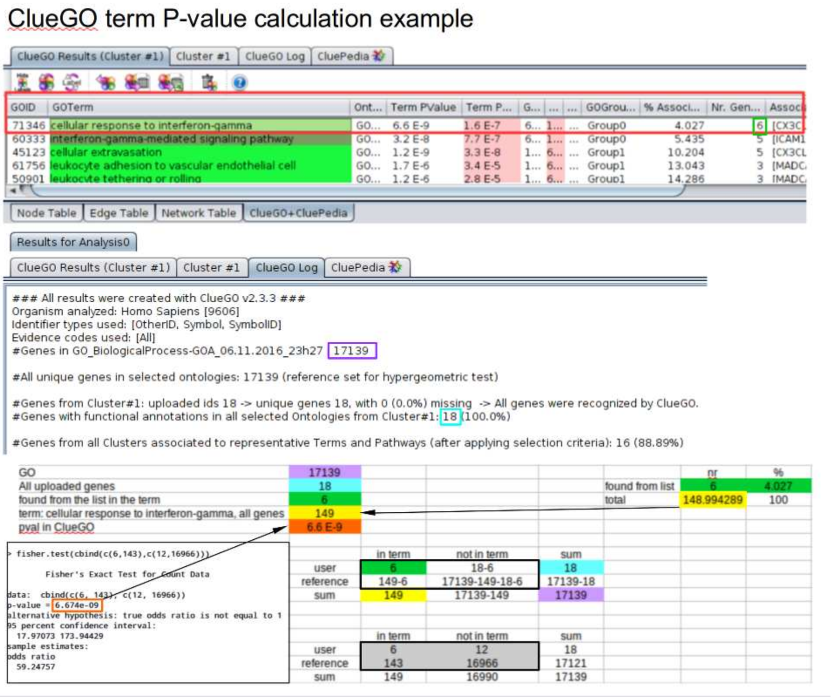
Term-Term comparison based on associated genes

Iterative comparison & merging of functional groups

ClueGO/CluePedia network with pathways, genes and functional groups

Figure 10: *Functional groups*

Statistics



ClueGO term P-value calculation example

GOID	GOTerm	Ont...	Term PValue	Term P...	G...	...	...	GOGrou...	% Associ...	Nr. Gen...	Assoc
71346	cellular response to interferon-gamma	GO...	6.6 E-9	1.6 E-7	6...	1...	...	Group0	4.027	6	[CX3C
60333	interferon-gamma-mediated signaling pathway	GO...	3.2 E-8	7.7 E-7	6...	1...	...	Group0	5.435	5	[ICAM1
45123	cellular extravasation	GO...	1.2 E-9	3.3 E-8	1...	6...	...	Group1	10.204	5	[CX3CL
61756	leukocyte adhesion to vascular endothelial cell	GO...	1.7 E-6	3.4 E-5	1...	6...	...	Group1	13.043	3	[MADC
50901	leukocyte tethering or rolling	GO...	1.2 E-6	2.8 E-5	1...	6...	...	Group1	14.286	3	[MADC

```
### All results were created with ClueGO v2.3.3 ###
Organism analyzed: Homo Sapiens [9606]
Identifier types used: [OtherID, Symbol, SymbolID]
Evidence codes used: [All]
#Genes in GO_BiologicalProcess-GOA_06.11.2016_23h27 17139

#All unique genes in selected ontologies: 17139 (reference set for hypergeometric test)

#Genes from Cluster#1: uploaded ids 18 -> unique genes 18, with 0 (0.0%) missing  -> All genes were recognized by ClueGO.
#Genes with functional annotations in all selected Ontologies from Cluster#1: 18 (100.0%)

#Genes from all Clusters associated to representative Terms and Pathways (after applying selection criteria): 16 (88.89%)
```

GO	17139					nr	%
All uploaded genes	18				found from list	6	4.027
found from the list in the term	6				total	148.994289	100
term: cellular response to interferon-gamma, all genes	149						
pval in ClueGO	6.6 E-9						

```
> fisher.test(cbind(c(6,143),c(12,16966)))

        Fisher's Exact Test for Count Data

data:  cbind(c(6, 143), c(12, 16966))
p-value = 6.674e-09
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  17.97073 173.94429
sample estimates:
odds ratio
  59.24757
```

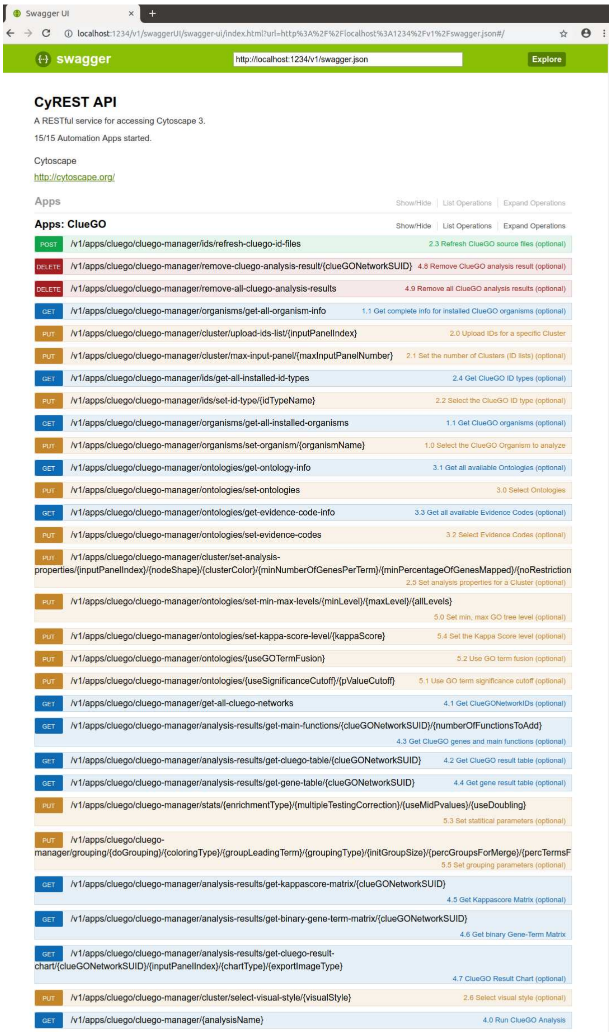
	in term	not in term	sum
user	6	18-6	18
reference	149-6	17139-149-18-6	17139-18
sum	149	17139-149	17139

	in term	not in term	sum
user	6	12	18
reference	143	16966	17121
sum	149	16990	17139

Figure 11: *Stats example*

ClueGO automation

Cytoscape Automation enables scientific workflows written in many languages through the CyREST [16]. ClueGO (v2.5.0+) implements the cyREST core plugin API and provides programmatical access to its functionality [17]. ClueGO features REST enabled can be explored in the cyREST API Swagger (Figure 12) (Cytoscape menu, Help/Automation/CyREST API).



localhost:1234/v1/swaggerUI/swagger-ui/index.html?url=http%3A%2F%2Flocalhost%3A1234%2Fv1%2Fswagger.json#/

http://localhost:1234/v1/swagger.json

CyREST API

A RESTful service for accessing Cytoscape 3.

15/15 Automation Apps started.

Cytoscape

http://cytoscape.org/

Apps — Show/Hide | List Operations | Expand Operations

Apps: ClueGO — Show/Hide | List Operations | Expand Operations

```
POST    /v1/apps/cluego/cluego-manager/ids/refresh-cluego-id-files    2.3 Refresh ClueGO source files (optional)
DELETE  /v1/apps/cluego/cluego-manager/remove-cluego-analysis-result/{clueGONetworkSUID}    4.8 Remove ClueGO analysis result (optional)
DELETE  /v1/apps/cluego/cluego-manager/remove-all-cluego-analysis-results    4.9 Remove all ClueGO analysis results (optional)
GET     /v1/apps/cluego/cluego-manager/organisms/get-all-organism-info    1.1 Get complete info for installed ClueGO organisms (optional)
PUT     /v1/apps/cluego/cluego-manager/cluster/upload-ids-list/{inputPanelIndex}    2.0 Upload IDs for a specific Cluster
PUT     /v1/apps/cluego/cluego-manager/cluster/max-input-panel/{maxInputPanelNumber}    2.1 Set the number of Clusters (ID lists) (optional)
GET     /v1/apps/cluego/cluego-manager/ids/get-all-installed-id-types    2.4 Get ClueGO ID types (optional)
PUT     /v1/apps/cluego/cluego-manager/ids/set-id-type/{idTypeName}    2.2 Select the ClueGO ID type (optional)
GET     /v1/apps/cluego/cluego-manager/organisms/get-all-installed-organisms    1.1 Get ClueGO organisms (optional)
PUT     /v1/apps/cluego/cluego-manager/organisms/set-organism/{organismName}    1.0 Select the ClueGO Organism to analyze
GET     /v1/apps/cluego/cluego-manager/ontologies/get-ontology-info    3.1 Get all available Ontologies (optional)
PUT     /v1/apps/cluego/cluego-manager/ontologies/set-ontologies    3.0 Select Ontologies
GET     /v1/apps/cluego/cluego-manager/ontologies/get-evidence-code-info    3.3 Get all available Evidence Codes (optional)
PUT     /v1/apps/cluego/cluego-manager/ontologies/set-evidence-codes    3.2 Select Evidence Codes (optional)
PUT     /v1/apps/cluego/cluego-manager/cluster/set-analysis-properties/{inputPanelIndex}/{nodeShape}/{clusterColor}/{minNumberOfGenesPerTerm}/{minPercentageOfGenesMapped}/{noRestriction
        2.5 Set analysis properties for a Cluster (optional)
PUT     /v1/apps/cluego/cluego-manager/ontologies/set-min-max-levels/{minLevel}/{maxLevel}/{allLevels}
        5.0 Set min, max GO tree level (optional)
PUT     /v1/apps/cluego/cluego-manager/ontologies/set-kappa-score-level/{kappaScore}    5.4 Set the Kappa Score level (optional)
PUT     /v1/apps/cluego/cluego-manager/ontologies/{useGOTermFusion}    5.2 Use GO term fusion (optional)
PUT     /v1/apps/cluego/cluego-manager/ontologies/{useSignificanceCutoff}/{pValueCutoff}    5.1 Use GO term significance cutoff (optional)
GET     /v1/apps/cluego/cluego-manager/get-all-cluego-networks    4.1 Get ClueGONetworkIDs (optional)
GET     /v1/apps/cluego/cluego-manager/analysis-results/get-main-functions/{clueGONetworkSUID}/{numberOfFunctionsToAdd}
        4.3 Get ClueGO genes and main functions (optional)
GET     /v1/apps/cluego/cluego-manager/analysis-results/get-cluego-table/{clueGONetworkSUID}    4.2 Get ClueGO result table (optional)
GET     /v1/apps/cluego/cluego-manager/analysis-results/get-gene-table/{clueGONetworkSUID}    4.4 Get gene result table (optional)
PUT     /v1/apps/cluego/cluego-manager/stats/{enrichmentType}/{multipleTestingCorrection}/{useMidPvalues}/{useDoubling}
        5.3 Set statitical parameters (optional)
PUT     /v1/apps/cluego/cluego-manager/grouping/{doGrouping}/{coloringType}/{groupLeadingTerm}/{groupingType}/{initGroupSize}/{percGroupsForMerge}/{percTermsF
        5.5 Set grouping parameters (optional)
GET     /v1/apps/cluego/cluego-manager/analysis-results/get-kappascore-matrix/{clueGONetworkSUID}
        4.5 Get Kappascore Matrix (optional)
GET     /v1/apps/cluego/cluego-manager/analysis-results/get-binary-gene-term-matrix/{clueGONetworkSUID}
        4.6 Get binary Gene-Term Matrix
GET     /v1/apps/cluego/cluego-manager/analysis-results/get-cluego-result-chart/{clueGONetworkSUID}/{inputPanelIndex}/{chartType}/{exportImageType}
        4.7 ClueGO Result Chart (optional)
PUT     /v1/apps/cluego/cluego-manager/cluster/select-visual-style/{visualStyle}    2.6 Select visual style (optional)
GET     /v1/apps/cluego/cluego-manager/{analysisName}    4.0 Run ClueGO Analysis
```

Figure 12: ClueGO in the Cytoscape API Swagger

Examples: http://www.ici.upmc.fr/cluego/ClueGOcyRESTExample.R and http://www.ici.upmc.fr/cluego/ClueGOcyRESTExample.py.

References

[1] P Shannon, A Markiel, O Ozier, N S Baliga, J T Wang, D Ramage, N Amin, B Schwikowski, T Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13:2498–2504, 2003.

[2] S. Lotia, J. Montojo, Y. Dong, G. D. Bader, A. R. Pico. Cytoscape app store. *Bioinformatics*, 29(10):1350–1351, May 2013.

[3] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, G Sherlock. Gene ontology: tool for the unification of biology The Gene Ontology Consortium. *Nat Genet*, 25:25–29, 2000.

[4] M Kanehisa, S Goto, S Kawashima, A Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30:42–46, 2002.

[5] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, 46(D1):D649–D655, Jan 2018.

[6] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, A. R. Pico. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.*, 40(Database issue):D1301–1307, Jan 2012.

[7] L. Cooper, P. Jaiswal. The Plant Ontology: A Tool for Plant Genomics. *Methods Mol. Biol.*, 1374:89–114, 2016.

[8] B. L. Aken, P. Achuthan, W. Akanni, M. R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, T. Juettemann, S. Keenan, M. R. Laird, I. Lavidas, T. Maurel, W. McLaren, B. Moore, D. N. Murphy, R. Nag, V. Newman, M. Nuhn, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, S. P. Wilder, A. Zadissa, M. Kostadima, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, F. Cunningham, A. Yates, D. R. Zerbino, P. Flicek. Ensembl 2017. *Nucleic Acids Res.*, 45(D1):D635–D642, 01 2017.

[9] R. Agarwala, T. Barrett, J. Beck, D. A. Benson, C. Bollin, E. Bolton, D. Bourexis, J. R. Brister, S. H. Bryant, K. Canese, M. Cavanaugh, C. Charowhas, K. Clark, I. Dondoshansky, M. Feolo, L. Fitzpatrick, K. Funk, L. Y. Geer, V. Gorelenkov, A. Graeff, W. Hlavina, B. Holmes, M. Johnson, B. Kattman, V. Khotomlianski, A. Kimchi, M. Kimelman, M. Kimura, P. Kitts, W. Klimke, A. Kotliarov, S. Krasnov, A. Kuznetsov, M. J. Landrum, D. Landsman, S. Lathrop, J. M. Lee, C. Leubsdorf, Z. Lu, T. L. Madden, A. Marchler-Bauer, A. Malheiro, P. Meric, I. Karsch-Mizrachi, A. Mnev, T. Murphy, R. Orris, J. Ostell, C. O'Sullivan, V. Palanigobu, A. R. Panchenko, L. Phan, B. Pierov, K. D. Pruitt, K. Rodarmer, E. W. Sayers, V. Schneider, C. L. Schoch, G. D. Schuler, S. T. Sherry, K. Siyan, A. Soboleva, V. Soussov, G. Starchenko, T. A. Tatusova, F. Thibaud-Nissen, K. Todorov, B. W. Trawick, D. Vakatov, M. Ward, E. Yaschenko, A. Zasypkin, K. Zbicz. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 46(D1):D8–D13, Jan 2018.

[10] P. Jones, D. Binns, H. Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Y. Yong, R. Lopez, S. Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, May 2014.

[11] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, C. von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(Database issue):D447–452, Jan 2015.

[12] M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, A. Ruepp. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.*, 47(D1):D559–D563, Jan 2019.

[13] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni, H. Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, 42(Database issue):D358–363, Jan 2014.

[14] P. D. Karp, R. Billington, R. Caspi, C. A. Fulcher, M. Latendresse, A. Kothari, I. M. Keseler, M. Krummenacker, P. E. Midford, Q. Ong, W. K. Ong, S. M. Paley, P. Subhraveti. The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinformatics*, Aug 2017.

[15] J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, P. Bork. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.*, 34(8):2115–2122, 08 2017.

[16] K. Ono, T. Muetze, G. Kolishovski, P. Shannon, B. Demchak. CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. *F1000Res*, 4:478, 2015.

[17] B. Mlecnik, J. Galon, G. Bindea. Automated exploration of Gene Ontology term and pathway networks with ClueGO-REST. *Bioinformatics*, Mar 2019.