ClueGO

Documentation



Integrative Cancer Immunology Team INSERM U872, Cordeliers Research Center, Paris, France

> Institute for Genomics and Bioinformatics Graz University of Technology, Graz, Austria

Contents

Installation
Documentation
Selection Panel
ClueGO Result
Functionally Grouped Annotation Network
Information table
Histogram with terms
Overview Chart with functional groups
Logging Information
Advanced Settings
Cluster Comparison
Result Folder
Statistics
Update
Include your organism of interest in ClueGO 17
Map other ID types than the IDs supported by ClueGO
Analysis from Cytoscape network
ClueGO and GOlorize
Sample data
Changing the network and charts
Running time
ClueGO examples
Bibliography

Installation

System Requirements:

- Windows, Linux, Unix or MacOS operating system.
- 512MB RAM needed, 1024 MB RAM recommended. Hard disk with at least 100MB free (for example files).
- Java 1.5+ needed, Java 1.6 recommended.
- Cytoscape 2.6.+ installed, with -Xmx1024m option in the start up file (.sh/.bat).

For downloading ClueGO, please visit: http://www.ici.upmc.fr/cluego/cluegoDownload.shtml. ClueGO is a Cytoscape plugin, thus it is necessary to copy the ClueGOPlugin_v1.x.jar in the plugin folder of Cytoscape.

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

If a new version of the ClueGO plugin is downloaded, please delete the old version from the Cytoscape plugin folder.

Documentation

Generally, 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.

The Gene Ontology (GO) projects [1] 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. The terms are structured in a hierarchical relationship (parent-child). Due to the complexity of the hierarchical structure (directed acyclic graph), the terms can be 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) [2] is a database of biological systems that integrates genomic, chemical and systemic functional information.

BioCarta (http://www.biocarta.com/) provides usefull pathway information.

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

To answer to this problematic, we developed ClueGO, an open-source Java tool that extracts the non-redundant biological information for large clusters of genes, using GO, KEGG and BioCarta. ClueGO is integrated in Cytoscape [3] as a plug-in and it is taking advantage of its complex visualization environment.

ClueGO features:

- ClueGO allows analysis of a single gene set (cluster) or cluster comparison.
- Different filter criteria can be applied to the terms.
- Fusion of the related terms that have similar associated genes.
- Functional grouping of the terms based on GO hierarchy or based on kappa score.
- Visualize the selected terms in a functionally grouped network.
- Charts presenting the specific terms and groups for the clusters compared.
- Statistical significance for the terms and for the groups.
- ClueGO can be used in combination with GOlorize.
- Easy updatable.
- Easy extendable.

Selection Panel

After starting Cytoscape, ClueGO can be found in the Plugins menu. Once selected, it will display the ClueGO selection panel on the left side in Cytoscape (see Figure 1):

Control Panel	2
🍃 Network \VizMapper™ \Filters \Editor \ ClueGO \	
ClueGO v1.1	
Analysis Type	
Single Cluster (1) O Compare Cluster	
Load Gene Cluster List(s)	
Homo Sapiens (2) - AccessionID (3) -	-
From File (4) From Network	
Bcell Healthy top200UpRegulated. Choose Cluster #1	
Select Ontology (♥) Evidence (*) GO MolecularFunction_2010.2008_171 All KEGG_14.10.2008 KEGG_14.10.2008 KEG	
Show Advanced Statistical Options (11)	
Set Network Specificity (12)	
Global Medium Detai	iled
Show Advanced Settings (13)	
ClueGO Functional Analysis	
Start (14	9

Figure 1: ClueGO selection Panel

- 1. Select the type of analysis. Besides analysis of a single gene set (cluster), ClueGO performs comparison of clusters, underlying the specificity and also the common aspects of their functionality.
- Select organism. For the moment, ClueGO supports Arabidopsis thaliana, Bos taurus, Caenorhabditis elegans, Danio rerio, Dictyostelium discoideum, Drosophila melanogaster, Escherichia coli, Galus galus, Homo sapiens, Magnaporthe grisea, Mus musculus, Oryza sativa, Rattus norvegicus, Saccharomyces cerevisiae.
- 3. Select identifiers type (AffymetrixID, AccessionID, SymbolID).
- 4. The identifiers can be uploaded from text files or interactively from a Cytoscape sub-network. ClueGO provides sample files containing specific genes (AccessionID) for Human B and NK cells. A Saccaromyces network sample (galFiltered.cys) can be found in the Cytoscape sampleData folder.
- 5. Choose the file with identifiers.

- Select the ontology/ontologies. To make ClueGO faster, we use precompiled files based on GO, KEGG and BioCarta. The precompiled files are included in the ClueGOPlugin_v1.x.jar. Perform point 8. before selecting the ontologies if necessary.
- 7. Select the evidence code, option available for the GO based files. For more details, see http://www.geneontology.org/GO.evidence.shtml.
- 8. Update the ontology (see ClueGO updates)
- 9. The significance of each term or group can be calculated with hypergeometric test (enrichment, depletion, two sided). Standard is hypergeometric test two sided.
- Several methods for PValue correction are proposed: Bonferroni, Bonferroni step-down and Benjamini-Hochberg.
- 11. Advanced statistical options
- Select network type. We predefined selection criteria leading to a "Global" network, a "Medium" network or a "Detailed" network.

The "Global" network displays GO terms found in the GO levels 1-3, with a high number of genes associated and a small percentage of uploaded genes found. Those terms provide a general biological information. At the opposite, the "Detailed" network shows very specific terms placed in GO levels 9-14, with only few associated genes but a high percentage of the uploaded genes found. Those terms are more specific and informative underlying particular aspects of the studied gene product.

The predefined settings provided by ClueGO should be combined with a customized selection for an analysis adapted to the biological question addressed (general/specific investigation) and to the number of genes involved.

- 13. Advanced Settings.
- 14. Start the analysis.

ClueGO result

A functionally grouped annotation network, a ClueGO information table, charts with terms and groups and the logging information are the result of ClueGO analysis. The created networks and analysis results can be saved in a specified project folder and used for further analysis. If the analysis is not needed anymore, it is recommended to close the project.



Figure 2: The biological role of B cell specific genes visualized with ClueGO.

Functionally Grouped Annotation Network

ClueGO visualizes the selected terms in a functionally grouped annotation network (see Figure 2) that reflects the relationships between the terms based on the similarity of their associated genes. The size of the nodes reflects the statistical significance of the terms (see Statistics). The degree of connectivity between terms (edges) is calculated using kappa statistics, in a similar way as described in [4]. The calculated kappa score is also used for defining functional groups. A term can be included in several groups. The reocurrence of the term is shown by adding "_n". The not grouped terms are shown in white color. Predefined, the group leading term is the most significant term of the group. The network

integrates only the positive kappa score term associations and is automatically laid out using Organic layout algorithm supported by Cytoscape.

Information table

ClueGO Information table provides information about the selected terms. From left to right:

SE688	7_Bcell_Healthy_top200UpRegula	ated.txt Cluster #1	Overview	Cluster #1	ClueGC	Log				
			Sa	ve ClueGO	Project	Close ClueGO Project				
GOID	GOTerm	Ontology Source	GOLevels	GOGro	% Assoc	Nr. Genes Term PVal	Term PValu	Group PVa	Group PValue	Associated Genes Found
15	Phosphoinositides and their do	REACTOME BioCart	[-1]	Group0	11.111	2 1.3 E-2	3.1 E-1	7.4 E-5	2.9 E-4	[BTK, PRKCE]
1619	Aspirin Blocks Signaling Pathwa	REACTOME BioCart	[-1]	Group0	10	2 1.6 E-2	3.8 E-1	7.4 E-5	2.9 E-4	[PTGS1, SRC]
46	B cell receptor signaling pathway	KEGG 14.10.2008	[-1]	Group0	6.25	4 3.7 E-3	8.6 E-2	7.4 E-5	2.9 E-4	[BTK, CD72, CD79B, FCG
72	-kappaB kinase/NF-kappaB cas	GO_BiologicalProc	[6, 7]	Group0	5.882	3 1.4 E-2	3.2 E-1	7.4 E-5	2.9 E-4	[BTK, TLR10, TLR7]
18	peptidyl-tyrosine modification	GO BiologicalProc	[7]	Group0	8.696	2 2.1 E-2	4.9 E-1	7.4 E-5	2.9 E-4	[BTK, SRC]
18	peptidyl-tyrosine phosphorylati	GO_BiologicalProc	[7,8]	Group0	9.524	2 1.8 E-2	4.1 E-1	7.4 E-5	2.9 E-4	[BTK, SRC]
380	Tryptophan metabolism	KEGG_14.10.2008	[-1]	Group1	6.897	4 2.6 E-3	6.0 E-2	2.1 E-3	8.4 E-3	[ALDH2, CYP1B1, INDO, K
65	tryptophan metabolic process	GO BiologicalProc	[5, 6, 7]	Group1	33.333	2 1.4 E-3	3.2 E-2	2.1 E-3	8.4 E-3	[INDO, KYNU]
90	aromatic amino acid family cata	GO_BiologicalProc	[5, 6, 7]	Group1	15.385	2 7.1 E-3	1.6 E-1	2.1 E-3	8.4 E-3	[INDO, KYNU]
19	pyridine nucleotide metabolic p	GO BiologicalProc	[5, 6, 7]	Group1	10.345	3 2.9 E-3	6.6 E-2	2.1 E-3	8.4 E-3	[INDO, KYNU, NMNAT1]
46	ndolalkylamine catabolic proce	GO_BiologicalProc	[6, 7]	Group1	50	2 5.8 E-4	1.3 E-2	2.1 E-3	8.4 E-3	[INDO, KYNU]
65	tryptophan catabolic process	GO BiologicalProc	[6, 7, 8]	Group1	50	2 5.8 E-4	1.3 E-2	2.1 E-3	8.4 E-3	[INDO, KYNU]
67	nicotinamide metabolic process	GO_BiologicalProc	[6, 7, 8]	Group1	11.538	3 2.1 E-3	4.8 E-2	2.1 E-3	8.4 E-3	[INDO, KYNU, NMNAT1]
19	pyridine nucleotide biosynthetic	GO BiologicalProc	[6, 7, 8]	Group1	15.385	2 7.1 E-3	1.6 E-1	2.1 E-3	8.4 E-3	[KYNU, NMNAT1]
19	NAD metabolic process	GO_BiologicalProc	[7, 8, 9]	Group1	23.077	3 2.5 E-4	5.8 E-3	2.1 E-3	8.4 E-3	[INDO, KYNU, NMNAT1]
94	NAD biosynthetic process	GO BiologicalProc	[7, 8, 9, 10]	Group1	28.571	2 1.9 E-3	4.5 E-2	2.1 E-3	8.4 E-3	[KYNU, NMNAT1]
220	Urea cycle and metabolism of a	KEGG_14.10.2008	[-1]	Group2	7.143	2 3.1 E-2	7.2 E-1	1.2 E-2	5.1 E-2	[ALDH2, GATM]
260	Glycine, serine and threonine m	KEGG_14.10.2008	[-1]	Group2	5	2 5.9 E-2	1.0000	1.2 E-2	5.1 E-2	[GATM, PSAT1]
330	Arginine and proline metabolism	KEGG_14.10.2008	[-1]	Group2	5.714	2 4.7 E-2	1.0000	1.2 E-2	5.1 E-2	[GATM, P4HA2]
64	mRNA catabolic process	GO_BiologicalProc	[6, 7]	Group 3	6.897	2 3.3 E-2	7.6 E-1	3.3 E-2	1.3 E-1	[SMG1, UPF3B]
956	nuclear-transcribed mRNA catab	GO_BiologicalProc	[7,8]	Group 3	8.333	2 2.3 E-2	5.3 E-1	3.3 E-2	1.3 E-1	[SMG1, UPF3B]
19/	nuclear-transcribed mRNA catab	GO RiologicalProc	[0 Q]	Group3	10	216 E.2	3.9 F.1	33 E.2	1 3 F.1	[SMG1_LIPE3R]

Figure 3: *ClueGO information table, detailed information about the selected terms and their associated genes*

- 1. Group leading term. There are several ways of defining the leading term of the group (see Advanced Settings). Predefined, the leading term has the highest significance in the group.
- 2. GOID.
- 3. GO Term.
- 4. Ontology Source.
- 5. GO levels. Due to the complex structure of GO tree (directed acyclic graph), the GO terms can be placed in several levels. In case of using sources without hierarchical structure (KEGG, BioCarta), the level it is assigned as -1.
- 6. The group or the groups that include the term.
- 7. The percentage of the genes from the uploaded cluster that were associated with the term, compared with all the genes associated with the term.

- 8. The number the genes from the uploaded cluster that were associated with the term.
- 9. Term significance (PValue).
- 10. Term significance (corrected PValue).
- 11. Group significance (PValue).
- 12. Group significance (corrected PValue).
- 13. The genes from the uploaded cluster that were associated with the term.

Histogram with terms

The chart presents the specific terms for the user genes and information related to their associated genes. The display can be customized (see Advanced Settings). Predefined, the bars represent the number of the genes from the analyzed cluster found to be associated with the term, and the label displayed on the bars is the percentage of found genes compared to all the genes associated with the term. Term significance information is included in the chart (see Statistics).



Figure 4: Functional terms in ClueGO Chart

Overview Chart with functional groups

The overview chart presents functional groups for the user genes. The name of the group is given by the group leading term (e.g. the most significant term in the group). The group sections correlate with the number of the terms included in group. The position of the sections can be changed (right click).



Figure 5: Functional groups in ClueGO Overview

Logging information

Complete information on the applied selection criteria for each analysis can be visualized in the logging information panel.



Node Attribute Browser / Edge Attribute Browser / Network Attribute Browser angle ClueGO /

Figure 6: ClueGO Logging Information

Advanced Settings

ClueGO advanced settings allow for a precise adjustment of the selection criteria.

 GO tree level. The identifiers can be mapped in the entire ontology tree or in a GO level interval (between Min and Max levels). If selected the first GO levels (1-3), the terms will be very general. Those terms have many associated genes, so, there will be a low percentage of found genes from the user genes. The very specific terms (e.g. level 12) have associated a small number of genes.

Set Network	Specificity		
Global	Med	ium	Detailed
GO Term Res	ranced Settings triction (GO Tree L	evels)	
(1) 3	 Min Level 	8	Max Level
GO Term Res	triction (#/% Gene	s)	
(2)4 -	Min #Genes	2 • 9	6Genes (3)
GO Term Fus	ion (Redundance F Term Fusion	Aeduction) 4) n (Kappa Score)	5)
Low	Medium	High	Score: 0.3
GO Term Gro Use GC Group col Fix	uping (Functional Term Grouping oring (7)	Grouping) (6) O Ra	andom (8)
Highest	Significance	(9) -
O Tree (10)	Kappa	Score (11)
3	, -	Initial Group	Size
50	•	% for Group	Merge

Figure 7: Advanced Settings

In this case, the percentage of found genes it is higher. Anyway, one has to consider the complex hierarchical GO structure that makes the terms to be in several levels. This feature is availabe while using a hierarchically structured ontology source (e.g. GO).

- 2. Minimum number of identifiers from the uploaded cluster found to be associated with a term
- 3. Minimum percentage of mapped identifiers present per term compared to the total of its associated genes
- Fusion. The terms in parent-child relationship that share similar genes (identical, or with one gene difference) are assessed and the most representative parent or child term is retained.
 Note: This is an optional feature. The fusion strongly reduces the redundancy.
- 5. Kappa Score shows the relationships between the terms based on their overlapping genes. It is used for creating the network and for creating the groups. Since the term-term matrix is of categorical origin, kappa statistic was found to be the most suitable method. Initially, a term-gene matrix containing the selected terms and their associated genes is created. Based on this, a term-term similarity matrix is calculated using kappa statistics to determine the association strength between the terms (kappa score). High score = visualize in the network only the connections between close related terms, with very similar associated genes (high stringency). Low score = allows to visualize in the network the connections between less related terms (low stringency).

6. Create functional groups. The terms can be associated in functional groups using the ontology's hierarchy or the kappa score.

When displaying the functional groups on the network, one has to consider that the network structure is calculated with kappa score. The kappa score grouping of the terms will correspond with the network structure that is calculated in the same manner. If the hierarchical grouping is used, it is possible to be a difference between the networks structure (based on the associated genes) and the functional groups defined using GO hierarchy.

- 7. Use fix coloring for the groups
- 8. Use random coloring for the groups
- 9. The most representative term in a group is giving the name of the group. It can be considered having:
 - a) the highest number of the found genes per term
 - b) the highest percentage of found genes per term
 - c) the highest percentage of found genes per total number of found genes
 - d) the most significant PValue (see Statistics)

This selection determines the ClueGO charts display: the number/percentage/significance of the terms. If the group leading term is the term with the higher number of genes or the most significant term from the group, a label with the percentage of the found genes compared with all the associated genes with the term is displayed on the bars of the chart. Reversely, if the group leading term is representing the percentage of genes (compared to all the genes associated with the term or with all the genes found from the uploaded cluster), on the label will be displayed the number of the genes found for this term.

- 10. The hierarchy based grouping consists in analyzing the selected terms through the perspective of their parent terms. The size of the groups can be defined using the number of common and different parents, that is depending on the level of the term in GO hierarchy. Since KEGG and BioCarta don't have a hierarchical structure, the terms selected from those souces will not be grouped.
- 11. For grouping the terms based on the kappa score, it is necessary to set the size of the initial group (e.g. 3) and the percentage (overlaping terms/group) for group merge (e.g. 50 percent).

For more details see [4]. All the possible initial groups with the terms showing a kappa score equal or above the predefined threshold (e.g. ≤ 0.3) are created. Further, the initial groups are iterratively compared and merged if they are overlapping in the defined percentage. A term can be part of several groups. The reocurrence of the term is shown by adding "_n", where n is the number off additional occurences. If the terms have very similar associated genes, it is possible that the iterative merging of the created groups to take a longer time in the attempt of merging all the groups. In this case, it is recommended to increase the kappa score level or to increase the percentage for group merge.

Advanced Statistical Options

- 1. Select mid-P-value for a less conservative hypergeometric test [5].
- 2. Select Doubling for a less conservative two tailed hypergeometric test [5].



Figure 8: Advanced Statistical Options

Cluster Comparison

Once selected the option: "Compare clusters", two file choosers are made available. The selection of the ontology terms is following the same steps as in the analysis for a single cluster. The difference refers to the number and the percentage of the genes per term.



Figure 9: ClueGO comparison, settings.

Those selection criteria can be flexible: Cluster1 OR Cluster2 options (1) or applied strictly for both clusters: Cluster1 AND Cluster2 options (2). It is recommended to have comparable number of identifiers in the analysed clusters. If one cluster is much larger, the selection criteria should be applied more strictly (e.g. increase the number of genes per term).

Further, we analyze the terms trough the perspective of their associated genes. It is possible that genes from both clusters to be associated with a term, but in different proportions. We defined a term as specific for one of the clusters if the percentage of associated genes from this cluster is higher than the selected threshold (e.g. 66 percent) (3). As result, charts with specific terms for each cluster are provided. The common terms are included in a separate chart.



Figure 10: Common terms for the analyzed clusters. A) The number of cluster1 genes associated with the term. B) The number of cluster2 genes associated with the term.

On the network, the different proportion of the genes from the analyzed clusters is represented with a color gradient from green, for the first cluster genes, to red for the second cluster. The visualization of the groups on the network can be switched with the one of the uploaded clusters distribution on the selected terms.



(a) Functionally grouped network

(b) Distribution of the two clusters on the network

Figure 11: The biological role (GO, KEGG, BioCarta) of NK cell up and down regulated genes visualized with ClueGO (kappa score: ≥ 0.3). (a) Groups network. Not grouped terms are shown in white. (b) Cluster distribution network. Terms with up/down regulated genes are shown in red/green, respectively. The color gradient shows the gene proportion of each cluster associated with the term. Equal proportions of the two clusters are represented in white.

Result folder

In the result folder are included (from up to down):

🗒 ClueGO Results for Analysis 0 All Genes Found For Selection.txt
E ClueGO Results for Analysis 0 BinaryGeneTermMatrix.txt
ClueGO Results for Analysis 0 ClueGOLog.txt
ClueGO Results for Analysis 0 Cluster #1.png
ClueGO Results for Analysis 0 EdgeAttributeTable.txt
🗒 ClueGO Results for Analysis 0 Functional Groups With Genes.txt
🔁 ClueGO Results for Analysis 0 Group NetworkView.pdf
ClueGO Results for Analysis 0 KappaScoreMatrix.txt
ClueGO Results for Analysis 0 NodeAttributeTable.txt
ClueGO Results for Analysis 0 Overview Cluster #1.png
ClueGO Results for Analysis 0 Project Networks.cys

Figure 12: ClueGO result folder

- 1. The list of genes that were found associated with the selected terms (under the selection criteria applied).
- 2. The binary gene-term matrix.

- 3. The logging information.
- 4. The terms (image, png).
- 5. Network edges information: the kappa score value for each connection.
- 6. The groups and the corresponding genes (txt).
- 7. The network (image, png).
- 8. The kappa score matrix (see Figure 17(a)).
- 9. The ClueGO browser information (GOID, GOTerm, Levels, Groups, Associated Genes, Significance).
- 10. The groups (image, png).
- 11. The network (.cys). It is used for recreating the network.

If a cluster comparison is performed, charts with specific and common terms and groups are provided.

Statistics

ClueGO calculates automatically the significance of the terms and groups.

The resulting PValues are included in the network visualization (the size of the nodes) and are also displayed on the charts. If a PValue correction method is selected in ClueGO selection panel, then on the network and on the charts the corrected PValue will be represented. The terms and groups significance can be found in the ClueGO browser.

In the charts, we mark the level of the significance for terms and groups using:

a) **: if the term/group is oversignificant, PValue<0.001

- b) *: if the term/group is significant, 0.001<=PValue<0.05
- c) . (dot): 0.05<=PValue<0.01

By default, the PValue is calculated with a two-sided minimal-likely-hood test on the hypergeometric distribution, equivalent to a classical Fisher's exact test [5]. Additional options provided are the left-sided (Enrichment) and right-sided (Depletion) hypergeometric tests. We also implemented the possibility to calculate mid-P-values and to apply doubling to two-sided hypergeometric tests. These features are a possibility to deal with the conservatism and discreteness of the hypergeometric distribution [5]. Several methods for PValue correction are proposed: Bonferroni, Bonferoni step-down and Benjamini-Hochberg.

We consider as reference the total number of the genes associated with all the terms included in the ontology source used. If several ontology sources are used, we consider all the unique genes found in all of the sources. The number of reference genes is mentioned in the logging information.

For the group significance we consider the number of unique genes found from the uploaded genes as associated with the terms included in the group and the total number of unique genes associated with those terms.

We provide a sample of PValue calculation for the term GO:0030837 (negative regulation of actin filament polymerization) using the classical Fisher Exact Test (equivalent to a two-sided hypergeometric test). The ontology used was: GO, Human, Biological Process (version from 26.09.2008). This ontology has 13592 associated genes. We did not apply any restriction regarding the evidence code of those genes. In the GO tree, genes with an evidence code inferred from electronic associations (IEA) are not displayed as associated with the terms in the tree.

The uploaded gene list had 300 of affymetrix identifiers (B cell sample file). 181 genes without any selection criteria of the term corresponding to these IDs were found in GO. From those genes, 2 genes were associated with the considered term.

	in term	NOT in term	Sum
user genes	2	179	181
reference genes	6	13586	13592
Sum	8	13765	13773

The calculated PValue was 4.6E-3, and corrected PValue 9.8E-2 (Bonferroni).

Table 1: Example of calculating the term significance

Update

ClueGO allows an easy integration of the most recent version of the GO and KEGG. The ontology and annotation source files are automatically downloaded and based on this, new ClueGO precompiled files are created.



Figure 13: ClueGO update

GO ontology (OBO file) and annotation files are directly downloaded from the FTP site of GO. The name of the GO based precompiled files contains the creation date of the ontology and is taken from the OBO file. The usage of multiple GO based precompiled files (e.g. BiologicalProcess, MolecularFunction, CellularComponent) is possible for files having the same data included in their name. GO has daily content changes.

ClueGO allows also KEGG updates via SOAP API. ClueGO KEGG based files will have in the name the date of the download. The KEGG update rate: http://www.genome.jp/en/info_dbget.html. Since Biocarta is not frequently changing and only supports human and mouse, updates will be supplied with each new release of ClueGO. BioCarta source files are taken from REACTOME via SOAP API. Source files and interfaces used:

- 1. GO Ontology, in OBO format: ftp://ftp.geneontology.org/pub/go/ontology/gene_ontology_edit.obo
- 2. GO Annotation files: ftp://ftp.geneontology.org/pub/go/gene-associations/
- 3. KEGG API: http://www.genome.jp/kegg/soap/
- 4. Biocarta API: http://www.reactome.org/download/index.html

ClueGO releases will be available for download at: http://www.ici.upmc.fr/cluego/cluegoDownload.shtml. Note that the update speed depends on the user's internet connection. The KEGG update can fail in case of high web traffic. The ClueGO precompiled files not needed anymore can be easely removed (right click). However, the initial ontology files will be recreated again at the next startup.

Include your organism of interest in ClueGO

ClueGO allows to add new organisms, not included in our initial selection.

For each new organism, a keyIdentifier file: organism.gene2accession.txt (e.g. Oryza Sativa.gene2accession) has to be created. In this file, the first column ("UniqueID#EntrezGeneID") will contain the keyID (e.g. EntrezGeneID) that is used for mapping. Next column, "SymbolID" contains the official gene symbol

or official symbol and synonyms separated by "|". Other types of identifiers can be included in the next columns.

The second file needed is organism.properties (e.g. Oryza Sativa.properties). This file contains the organism name, the keggID, the url of the GO association file, the taxonomy id. The same url can be used for other organisms after changing the name of the annotation file. In most of the cases, there are no special files required for the update (false).

Here is the content of the .properties file for Oryza sativa:

organism.name = Oryza sativa

organism.kegg.name = osa

organism.go.url = ftp://anonymous:anonymous@ftp.geneontology.org/pub/go/gene-associations/gene_

association.gramene_oryza.gz

organism.taxid = 39946

organism.update.needs.special.file = false

Those files have to be added in a new folder (e.g. Organism_Oryza sativa) created in UserHomeFolder, .cluegoplugin, currentVersion, ClueGOFiles, ClueGOSourceFiles.

Remark: At the addition of a new organism, make sure that the ids provided by KEGG are the same as the UniqueIDs from organism.gene2accession.txt.

Map other ID types than the IDs supported by ClueGO

Since ClueGO is supporting a limited number of identifiers type, we made available a feature that allows the mapping of other types of identifiers, not included in our selection.

For each supported organism, a keyIdentifier file: organism.gene2accession is provided (e.g. HomoSapiens.gene2accession). In this file, the first column ("UniqueID#EntrezGeneID") contains the keyID (e.g. EntrezGeneID) that is used for mapping. The type of keyID differs from one specie to another due to the most frequent usage of a certain type of ID for different species (e.g. for Drosophila is FBgnID). If the identifiers of interest are not supported, a new text file (tab delimited) containing the keyID (from the provided files) as first column and the corresponding new ids in the next colums can be created.

The name of the first column has to be "UniqueID#keyID". If several identifiers correspond to a keyID, they have to be separated with "|". This file is added in the UserHomeFolder, ClueGOFiles folder, ClueGOSourceFiles, OrganismOfInterest.

Analysis from Cytoscape network

In ClueGO, the identifiers can be uploaded interactively from a Cytoscape network.

If selected the option "From Network" (1) in the ClueGO Selection Panel, new features are made available. After selecting the genes from the network, their attributes have to be loaded (2: "Load Attributes" button). Then, from the available attributes (3) is selected the attribute to display (e.g. SGD ID). This attribute must correspond to the selected type of identifier (e.g. SGDID). The selected attribute list is uploaded in ClueGO (4: "Choose Cluster" button). The number of the selected nodes will be displayed in a text field (5). If the type of identifier is not the same with the selected attribute node, the analysis cannot be performed. In this case, after selecting a new type of attribute, make sure that the new selected attribute list is uploaded in ClueGO (4).

The comparison of two clusters is also available.

A sample Cytoscape network for Saccaromyces (galFiltered.cys) can be found in Cytoscape, sampleData folder. The node attributes can be visualised in Cytoscape in "Node Attribute Browser".



Figure 14: ClueGO analysis with genes uploaded interactively from Cytoscape network

ClueGO and GOlorize

ClueGO can be used in combination with GOlorize ([6]).

If both plugins are present in Cytoscape plugin folder, the ClueGO data panel is automatically transformed into ClueGOlorize panel. This panel combines the functionality of both, ClueGO and GOlorize panel.

From a Cytoscape network the genes of interest are selected. With ClueGO, biological functions are associated to those genes. In ClueGOlorize panel, the terms to be mapped on the network can be selected (1: "Select All"). Using GOlorize, the genes will be coloured on the network with "Validate"(2).



Figure 15: ClueGO used in combination with GOlorize

Sample data

From Gene Expression Omnibus we downloaded the dataset GSE6887, a gene expression profile of peripheral blood lymphocytes of melanoma patients and healthy controls. From this dataset we selected the expression profiles NK cells in healthy controls and calculated the mean over all replicates. The genes were sorted based on their expression and the top 200 most expressed genes were included in the list with up regulated genes and the top 200 less expressed genes in the list with down regulated genes. The same steps we applied to the B cell data. The reference used in this dataset was a pool of all immune cell types.

We use as ClueGO sample files (Human, AccessionID) the lists with the 200 up and the 200 down regulated genes for B cells and NK cells. The sample lists can be found in .cluegoplugin folder, in the user's home folder.

Changing the ClueGO network and charts

• In ClueGO, the colors attributed to the groups are created automatically each time the analysis is performed.

One can change the color of the groups using the Cytoscape feature: "Node color", available under "VizMapper". If a comparison analysis is performed, in order to change the colors of the groups make sure that the groups are displayed on the network (option "Show Groups"). If the "Show Difference" option is active in ClueGO, under "Node color" feature will be displayed the color gradient used for showing the percentage of the genes found from the analysed clusters.

- The two colors defining the gradient displayed under "Show Difference" can be changed by using the color choosers available in the ClueGO selection panel, next to the file choosers.
- Cytoscape Layout features allow modifications of the network (e.g. rescaling, rotating, apply different layout patterns).
- The size of the displayed text on the network can be modified in Cytoscape VizMapper under "Node Font size". By default, the font size of the term names is 12 and of the group names 20. If the font size is set to a small value (e.g. 0.1), the names will not be anymore visible on the network.

• One can sort the data from ClueGO browser and select the terms of interest (e.g. the terms significant after PValue correction). Then, in Cytoscape, with File, New, Network, From selected nodes (Ctrl N) a new network is created. For the display, we use the Organic layout.

Running time

The running time depends on the number of genes uploaded but also on the selection criteria.

A longer time will take the analysis of a large list of genes mapped in the entire GO with a small minimum number of genes found per term, small percentage of those genes and small kappa score. If the percentage for group to merge is increased, there will be a high number of groups in the end. The status of the analysis can be followed in ClueGO selection panel, under Current Action. If the number of the groups visualized for different iterations of an analysis is very high, it is recommended to review the selection criteria:

- increase the minimum number of the genes found from the list for a term to be selected
- increase the minimum percentage of those genes for a term
- increase the kappa score
- decrease percentage for group to merge
- restrict the number of the terms based on GO levels

Also, if a comparison of cluster is performed, it is recommended to have a similar number of genes in each cluster. If the cluster have different number of genes, for the one with more genes more strict selection criteria should be applied.

To note is that certain parts of GO are more developed than others, due to an increased interest from the scientific community in that specific area. This could also impair the analysis. The number of iteration and the number of groups of the current iteration can be seen above the progression panel in the left side of Cytoskape. The number of iterations depends on the group-merging options, and on the fact that the algorithm converges. If the number of groups per iteration exceeds 500, the algorithm stops and applies the grouping with minimal number of groups found. An iteration with around 300 groups can take around 30 - 50 seconds on a 2.4GHz computer.

ClueGO examples

Fusion example



Figure 16: Fusion example. A) The test gene list and the total number of corresponding GO terms. Selection criteria applied in ClueGO. B) GO terms selected under the specified selection criteria. C) Fusion criteria. It applies to GO terms in parent-child relation, with similar associated genes. The most representative parent or child term is preserved. The terms are organized based on their mean level specificity (up: general terms with low specificity, down: specific terms). D) GO terms mentained after fusion (colored in blue). Colored arrows show the fusion criteria used (concordance with C).

Kappa statistics in ClueGO

Steps in creating functional groups and the network.







Figure 17: Gene-term matrix

2. Based on the gene-term matrix (step 1) and using kappa statistics, a term-term similarity matrix is calculated. Since the term-term matrix is of categorical origin, kappa statistics was found to be the most suitable method [7]. The method used for defining functional groups is similar with the one described in [4].



Figure 18: Kappa score matrix

3. Functional groups are created.

Initial groups are defined as having a minimum number of terms (e.g 3 terms/group) with a similarity of associated genes (kappa score) bigger than a predefined threshold (e.g. 0.03).

In the next step, the initial groups are iteratively merged if they overlap with a defined percentage (e.g 60). The iteration ends when the merging is not possible anymore.



Figure 19: Kappa score matrix (visualized in Genesis)

4. The network with terms (nodes) linked based on the predefined kappa score level is created. The edges with a kappa score smaller than 0.03 are not displayed on the network. E.g. the link between regulation of cell size and regulation of leukocyte activation with kappa 0.029. The groups are visualized with different colors on the network.



Figure 20: ClueGO network

Walk-through example

ClueGO settings in Figure 1 (paper)

- set the type of analysis: Compare Cluster
- select the organism: Homo Sapiens and the the type of ids used: AccessionID
- load sample gene lists:

choose Cluster #1: select GSE6887_NKcell_Healthy_top200UpRegulated.txt choose Cluster #2: GSE6887_NKcell_Healthy_top200DownRegulated.txt

- select the Ontologies: GO_BiologicalProcess_10.10.2008 (All Evidence codes), KEGG_Pathways_14.10.2008 and REACTOME_BioCarta_10.10.2008
- select the statistical test: Enrichment/Deplection (Two sided hypergeometric test), FisherExactTest
- select the correction method: Bonferroni
- click Show Advanced Settings
- set GO Tree Level: Min 4 and Max 5
- set the selection criteria for the terms that have associated genes from cluster 1: min 2 genes/term and minimum 4% from all the Genes associated with the term
- select OR (e.g. min 2 genes from cluster #1 or min 2 genes from cluster #2)
- set is specific to 66% (if 66% or the genes associated with the term are from cluster #1, the term is considered specific for this cluster)
- set the selection criteria for the terms that have associated genes from cluster 2: min 2 genes/term and minimum 4% from all the Genes associated with the term
- select Use GO Term Grouping
- select Fix Group coloring

- select Leading Group Term based on Highest Significance
- select Kappa Score grouping with 3 terms in initial group and 50% overlap for groups to merge
- select ShowDifference
- Start

Single Cluster Compare Cluster Single Cluster Compare Cluster Single Cluster Compare Cluster Compare Cluster Compare Cluster Form File Form File Choose Cluster #2 Leader Statistics Select Choose Cluster #2 Leader Statistics Compare Cluster #2 Leader Statistics Compare Cluster #2 Leader Statistics Compare Cluster #2 Leader Market Cluster #2 Leader Statistical Test Enrichment/Depletion (Two-sided hypergeometric test) Softwork Specificity Show Advanced Statistical Options Set Network Specificity Compare Cluster #2 Show Advanced Statistical Options Set Network Specificity Compare Cluster #2 Show Advanced Statistical Options Set Network Specificity Compare Cluster #3 Compare Cluster #4 Win #Genes Compare Cluster #3 Compare Cluster #4 Win #Genes Compare Cluster #4 Specific (Cluster #2 Min #Genes Go Term Flusion Co Term Flusion Co Term Flusion Co Term Connection Reduction Co Term Grouping Content Cluster Biologic (Cluster #3 Co Term Grouping Fire Compare Cluster Biologic (Cluster #3 Show Morupe Term Based on Highs Store) Co Term Grouping Fire Compare Statistical on Cluster #3 Show Morupe Term Based on Highs Store) Show Morupe Term Based on Highs Store Show Moru	eGO v1.1 nalvsis Type				
	nalysis lyna				
ad Gene Cluster List(s) Homo Sapiens AccessionID From File File From File File From File File File From File File File File File File File File	 Single Clust 	er		Compa	re Cluster
Homo Sapiens AccessionD From Network From Network From Network Select Drobody Select Drobody Select Statistical Test Enchment/Depletion (Two-sided hypergeometric test) Benferroni Activation (Select Select Select Select Select Select Select Select Select Test Enchment/Depletion (Two-sided hypergeometric test) Benferroni Protocol (Select Select Sele	oad Gene Cli	ister List(s)			
From File Choose Cluster #1 Choose Cluster #2 Healthy top2000_PMRegulated Choose Cluster #2 Healthy top2000_PMRegulated From File From Fi	Homo Sapiens		Acces	sionID	-
ell Healthy top200UpRequiated Healthy top200UpRequiated Choose Cluster #1 Healthy top200DownRequiated Choose Cluster #2 Choose Cluster #2 Choose Cluster #2 Choose Cluster #2 Choose Cluster #2 Choose Cluster #2 Choose Cluster #2 All who LeA REACTONE BioCarta 10 10 2008 17 # All REACTONE BioCarta 10 10 2008 17 # All REACTONE BioCarta 10 10 2008 17 # All REACTONE BioCarta 10 10 2008 17 # All Extended Statistical Options Select Statistical Test EnrichmentDepletion (Two-sided hypergeometric test) Bonferroni PV Correction Show Advanced Statistical Options Sel Network Specificity Sol Medium Detz Show Advanced Statistical Options Set Network Specificity Cluter #1 Cluter #1 Min #Genes 4 % Genes GO Term Restriction (#0 Grees) Cluter #2 Min #Genes 4 % Genes GO Term Fusion (Redundance Reduction) Cluter #2 Min #Genes 4 % Genes GO Term Fusion (Redundance Reduction) Cluter #2 Min #Genes A % Genes GO Term Grouping (functional Grouping) Clute GO Term Grouping Go Term Grouping (functional Group Firm Based on Highest Significance Kappa Score) Free	From File				n Network
Control (M2) CODOWNRequirated Choose Cluster #2. Headity (to 2000) winkequirated Choose Cluster #2. Select Ontology Choose Cluster #2. MacGO Settings Evidence Select Ontology File All winker (EA) Select Ontology File All winker (EA) Show Antology Update Options Evidence Select Satistical Test Enrichment/Depletion (Two-sided hypergeometric test) Bonferroni V Correction Show Advanced Statistical Options Select Satistical Options Select Satistical Test Max Lew Correct Restriction (BO Tree Levis) Max Lew Correct Restriction (BO Tree Levis) Max Lew Correct Restriction (BO Tree Levis) Max Lew Correct Restriction (Res Genes) Correct Restriction (Res Genes) Correct Restriction (Res Genes) Min #Genes Correct Restriction (Res Genes) Go Term Fusion Correct Correct Restriction (Restriction (Kappa Score)) Use GO Term Fusion Correct Guine (Restriction (Respa Score)) Go Term Grouping Group Correct Statistical Group Size Show Moroups (winkal Group Size Go Term Grouping Guine Cleand Group Merge Fread Fusion Clean Size <td>ell Healthy to</td> <td>n 2001 In Regula</td> <td>ted</td> <td>Choose</td> <td>Cluster #1</td>	ell Healthy to	n 2001 In Regula	ted	Choose	Cluster #1
	Healthy top2	00DownRegula	ted:	Choose	Cluster #2
UseGO Settings Select Othology Select Othology Select Othology Select Othology Select Othology All whoull EA REACTONE BioCarta 10 10 2008 All whoull EA REACTONE BioCarta 10 10 2008 Select Statistical Test Enrichment/Depletion (Two sided hypergeometric test) Bond Advanced Statistical Options Select Statistical Test Enrichment/Depletion (Two sided hypergeometric test) Bond Advanced Statistical Options Select Statistical (For the Levels) Or Term Restriction (#0 free Levels) Of Term Carticon (
Select Ontology Evidence Go Molecularitorio 20 10 200 11 M All KEGC 1410 2005 All All All All All All All All All Al	lueGO Settin	igs			
Gol Malecular-unclon 2010 2008 All without IEA REACTONE BIOCATA ID 10.2008 EAC Inferred Initiation of the set of t	Select Ontolog	y		Evide	nce
BEACTORE BIOCATE 10 10 2008 ■ EXP (Interred Inf Show Ontology Update Options Select Statistical Test Enrichment/Depletion (Two-sided hypergeometric test) Bionferron	GO_Molecular	unction_20.10.	2008_1		ithout IEA
Show Ontology Update Options Select Statistical Text Enrichment/Depletion (Two-sided hypergeometric test) Bonnernai PV Correctice Show Advanced Statistical Options Set Network Specificity Jobal Medium Deta Of Term Retroit on (G) Tree Levels) Max Leve GO Term Retroit on (G) Tree Levels) Max Leve CO Term Retroit on (G) Tere Levels) Max Leve CO Term Retroit on (G) Tere Levels) Max Leve CO Term Retroit on (G) Tere Levels) Max Leve CO Term Retroit on (G) Tere Levels) Max Leve CO Term Retroit (G) Tere Levels) Max Leve CO Term Subin Retruition (Rapa Score) Vere Min #Genes CO Term Louin Retruitance Reduction) O Use GO Term Fusion CO Term Grouping Go Term Grouping G' Use GO Term Grouping Graph Score: 0.3 GO Term Grouping Graph Score: 0.3 GO Term Grouping Final Group Score: 3 Go Term Grouping Wind Score: Score: 3 Go Term Grouping Wind Score: Score: 3 Soon % for Group Merge Staco Comparis	REACTOME_Big	Carta_10.10.20	800	EXP (Inferred fr 💌
Select Statistical Test Enrichment/Depicetion (Two-sided hypergeometric test) Bonferroni V Corrector Show Advanced Statistical Options Set Network Specificity Set Network Specificity Solution (Go Tree Levels) Go Term Restriction (60 Tree Levels) Go Term Restriction (60 Tree Levels) Go Term Restriction (67% Genes) Cutter 21 Win #Genes Go Term Restriction (67% Genes) Cutter 22 Win #Genes Go Term Fusion Go Term Fusion Go Term Fusion Go Term Grouping Cutter 20 Go Term Grouping Cutter 30 Go Term Grouping Cutter 30 Go Term Grouping Cutter 30 Go Term Grouping Cutter 30 Solution Corporation Networks Solution Corporation Show Orcups Show Orcups Show Orcups Show Ofference TereCine Manager Show Orcupa Show Orcupa Show Orcupa Show Ofference Show	 Show Ontol 	ogy Update Op	otions		
Enrichment/Depletion (Two-sided hypergeometric test) Bonferroni V Corrector Ston Advanced Statistical Options Set Network Specificity Solution (Solution) Set Network Specificity Show Advanced Settings Show Advanced Settings Clutter #1 Clutter #1 Clutter #2 Clutter #2 Clutter #2 Clutter #2 Clutter #2 Correm Fusion (Redundance Reduction) Clutter #2 Clutter #2 Store Correm Fusion Correm Fusion (Redundance Reduction) Clutter Correm Clution Correm Connection Reduction Correm Connection Correm Connectio	Select Statistic	al Test			
loomeroni PV Corrector loomeroni PV Corrector Show Advanced Statistical Options Set Network Specificity Sibobal Medium Deta Show Advanced Settings Oo Term Restriction (60 Tree Levels) Max Level O Term Restriction (60 Tree Levels) Max Level Max Level O Term Restriction (60 Tree Levels) Max Level Max Level Co Term Restriction (67 Tree Levels) Max Level Set NeGenes Co Term Fusion 66 Tree Vision & Wis Specific Deta Co Term Connection Restriction (Rappa Score) Set O Term Fusion Go Term Grouping Go Term Grouping Grappa Score: Set O Term Grouping Final Group Stree Use GO Term Grouping Final Group Stree Statis Significance Wishal Group Stree Sool Wis Groups Term Based on Hingla Group Stree Statis Significance Show Ofference Use GO Term Grouping Final Group Stree Statis Significance Show Ofference UseOC Comparison View Settings Show Ofference Show Ofference Show Ofference	Enrichment/De	pletion (Two-s	ided hy	pergeome	tric test) 🔻
Set Network Specificity Set Network Specificity Solution Show Advanced Settings O Term Restriction (60 Tree Levels) O Term Restriction (74 Genes) Uter 3. O Term Restriction (74 Genes) Uter 4. O Term Cancel Restriction Solo Morcups Solo Morcups O Term Cancel Restriction Solo Morcups O Term Cancel Restriction Solo Morcups O Term Cancel Restriction Solo Morcup Solo Morc	Bonterroni	and constant and	Online	• P	V Correction
Set Network Specificity Set Network Specificity Show Advanced Settings Corem Restriction (60 Tree Levels) Counter Restriction (7% Genes) Counter Restriction (7% Genes) Counter Restriction (7% Genes) Counter Restriction (7% Genes) Counter Restriction (Respective) Counter Restriction (Restriction (Respective) Counter Restriction (Restriction (Respective) Counter Restriction (Restriction (Restricti		iced Stausucal	opuon	>	
Jabal Medium Detal Show Advanced Settings Co Term Restriction (60 Tree Levels) Min Level S Max Level Co Term Restriction (#% Genes) Coltater #J Second Min Level Co Term Restriction (#% Genes) Coltater #J Second Second Coltater #J Min #Genes % Genes Second Coltater #J Min #Genes % Second Second Coltater #J Min #Genes Second Second Coltater #J Min #Genes Second Second Coltater #J Min #Genes Second Second Coltater #J Genes Coltater #J Second Coltater #J Second Random -Leading Soughtframe Coltater #J Min #Genes % for Group Merge Show Groups Show Difference Show Difference MacO Show Difference Show Difference	Set Network Sp	pecificity			
Jobal Medium Deta Show Advanced Settings (O Term Restriction (40 Free Levels) (a medium) (a medium) Min Level (b medium) (b Term Restriction (47% Genes) (b medium) (b Level 3) (b medium) (b Min #Genes 4) % Genes (c Term Restriction (47% Genes) (b medium) (b Min #Genes 4) % Genes (c Term Gruping) (b medium) (c Term Gruping) (c medium) (c term) (c medium) </td <td></td> <td></td> <td></td> <td></td> <td></td>					
Show Advanced Settings Of Term Restruction (160 Tree Levels) Go Term Restruction (160 Tree Levels) Go Term Restruction (17% Genes) Cutater #1 2 Min #Genes Go Term Set Settings Cutater #2 Min #Genes Go Term Fusion Go Term Connection Reduction Use Go Term Fusion Go Term Connection Reduction Go Term Connection Go Term Conn	Slobal	Me	dium		Detail
GO Term Restriction (#0) Free Levels) Image: Min Level S GO Term Restriction (#7% Genes) Cluster #1 Image: Min #Genes Image: Min	Show Advar	nced Settings			
A Min Evers Min A Levre Col Term Restriction (#% Genes) Culture #1 Culture #1 Go % Sis Specific Culture #2 Min #Genes AND Go %% is Specific Culture #2 Min #Genes Col Term Fusion Go Term Fusion Go Term Fusion Go Term Grouping Go Term Grouping Group Grouping) Go Term Grouping Fix Go Term Grouping Random Leading Group Term Based on Highest Significance Inside Group Size So So % for Group Merge	GO Term Restri	ction (GO Tree L	evels)		
CO Term Restriction (#%) Genes Q Min #Genes Q Term Grouping S	4	 Min Level 	5	-	Max Level
Cluter #1 2 Min #Genes 4 %Genes 0 R AND 56 %S is Specific Cluter #2 2 Min #Genes 4 %Genes 0 Term Fusion (Redundance Reduction) 0 Go Term Fusion 60 Term Connection Restriction (Kappa Score) 0 Term Grouping 0 Terus Connection Restriction (Kappa Score) 0 Tere Restriction Restriction (Kappa Score) 0 Show Ordons 0 Show Ofference 1 TereConnection Restriction	GO Term Restri	ction (#/% Gene	s)		
Vini #Genes AND Go So AND Go Vis Specific Cluster #2 Vis AND Go Vis Specific Cluster #2 Vis AND Go Vis Co Co Ten Co	Cluster #1				
AND 66 % is Specific Cuter #2 Min #Genes 4 %Genes GO Term Fusion (Redundance Reduction) Use GO Term Fusion Go Term Grouping O Term Grouping	2 🔻	Min #Genes	4	• %(Genes
Comparison View Settings Show Ordups Sore Show Ordups	• OR		66	▼% is 5	pecific
Otem Fusion Redundance Reduction Otem Fusion Redundance Reduction Use GO Term Fusion GO Term Fusion GO Term Grouping Term Grouping Grad Crem Grad Crem Grad Crem Grad Crem Grad Crem Grad Crem Grad Crem Grad Crem Grad Crem Grad Crem Grad Crem Grad Cre	and the second se				
o Term Fusion (Redundance Reduction) Use GO Term Connection Restriction (Kappa Score) Cor Term Connection Restriction (Kappa Score) Cor Term Grouping (functional Grouping) ✓ Use GO Term Grouping ✓ Group Connection Restriction (Kappa Score) Group Connection Restriction (Kappa Score) Solo Terme Solo Terme Solo Terme Solo Terme Standard Score) Solo Terme Show Orcups Show Difference Inscol Restrictional Analysis	2 V	Min #Genes	4		Senes
Use GO Term Fusion GO Term Croueding Restrictions (Kappa Score) Low Medium High Score: 0.3 GO Term Grouping Ortuna Crouping Ortuna Ortuna Crouping Ortuna	2 V	Min #Genes	4	▼ %(Genes
CO Term Connection Restriction (Kappa Score) Low Medium High Score: 0.3 CO Term Grouping (Functional Grouping) ☑ Use GO Term Grouping ◎ Fix R Random Highest Significance So Tere & Kappa Score 3 ♥ Initial Group Size 50 ♥ % for Group Merge UseGO Comparison View Settings Show Groups @ Show Difference HueGO Exercitional Analysis	GO Term Fusio	Min #Genes	4 Reduction	▼ %(Genes
Low Medium High Score: 0.3 Co Tem Grouping (functional Grouping) C Use Co Tem Grouping Orarup colony Orarup colony P Kix Random Leading forup Term based on Highest Significance So So So So So So So So So So	GO Term Fusio	Min #Genes n (Redundance F erm Fusion	4 Reduction	• %(Genes
Low Medium High Score: 0.3 GO Term Grouping (Functional Grouping) Group colong First Colong First Colong First Colong Highest Significance So So So So So So So So So So	GO Term Fusio	Min #Genes n (Redundance F erm Fusion ection Restriction	4 Reduction n (Kappa	• %(Genes
CO Term Grouping (functional Grouping)	GO Term Fusio	Min #Genes n (Redundance F erm Fusion ection Restriction	4 Reduction n (Kappa	▼ %(Genes
Constructional Analysis Show Difference Show Difference	GO Term Fusio Use GO T GO Term Conn	Min #Genes n (Redundance F erm Fusion ection Restriction Medium	4 Reduction	Score) High S	Genes
Group coloning ●Fix ●Fix Cading Group Term based on Highest Significance Tree ● Kappa Score initial Group Size ⑤ for Group Merge UeGO Comparison View Settings ⑤ show Oriforence "ueGO Exectional Apalvis" ● Show Difference	GO Term Fusio Use GO T Low GO Term Group	Min #Genes n (Redundance F erm Fusion ection Restriction Medium ping (Functional	4 Reduction n (Kappa Grouping	> %() Score) High S	Senes
Fix Readom Fix Readom Fix Readom Fighest Significance So Tree So Fix Readom Fighest Significance So Fix Readom Fighest Significance So Fix Readom Fix Reado	GO Term Fusio GO Term Fusio GO Term Conn GO Term Conn Cow GO Term Group V Use GO 1	Min #Genes n (Redundance F erm Fusion ection Restriction Medium ing (Functional ferm Grouping	4 Reduction n (Kappa Grouping	> %() Score) High S	Genes
Leading Group Term Based on Highest Significance Kappa Score 3 While Group Size 50 Start Significance Start 50 Start Significance Start 50 Show Ordpus Start Show Difference TuesC0 Exerctional Apalysis	GO Term Fusio Use GO 1 COVERT Conn COVERT CONN CONN COVERT CONN CONN CONN CONN CONN CONN CONN CONN	Min #Genes n (Redundance F erm Fusion ection Restriction Medium bing (Functional erm Grouping ing	4 Reduction n (Kappa Grouping	• %(Score) High S	Genes
Highest Significance Kappa Score Tree Kappa Score 3 Initial Group Size 50 % for Group Merge	GO Term Fusio Use GO 1 GO Term Conn GO Term Group GO Term Group GO Term Group Group color Fix	Min #Genes n (Redundance F erm Fusion ection Restriction Medium bing (Functional ferm Grouping ing	4 Reduction n (Kappa Grouping		Genes
Tree ● Rappa Score 3 Imital Group Size 50 * % for Group Merge clueGO Comparison View Settings Show Groups ● Show Difference "UseGO Excitional Analysis	GO Term Fusio Use GO T GO Term Conn GO Term Conn GO Term Group Use GO T Group color Fix Leading Gro	Min #Genes n (Redundance F Ferm Fusion ection Restriction Medium ning (Functional Ferm Grouping Ing	4 Reduction n (Kappa Grouping	> %6 Score) High S () Rar	Genes
3 Initial Group Size 50 Initial Group Size % for Group Merge % for Group Merge ClueGO Comparison View Settings Initial Group Size Show Groups Is Show Difference UneGO Exectional Analysis Is Show Difference	GO Term Fusio GO Term Fusio GO Term Conn GO Term Conn GO Term Group Use GO 1 Group color Fix Leading Group Highest Sig	Min #Genes n (Redundance F erm Fusion ection Restriction Medium bing (Functional erm Grouping ing up Term based ognificance	4 Reduction n (Kappa Grouping	> %6 Score) High S () Rat	Senes
100 % for Group Merge	GO Term Fusio GO Term Fusio GO Term Conn GO Term Conn GO Term Group Use GO 1 Go Term Group GO Term Group GO Term Group GO Term Group GO Term Group GO Term Group Tuse GO 1 GO Term Group GO Term GO T GO Term GO TER GO TER	Min #Genes n (Redundance F Term Fusion ection Restriction Medium oing (Functional Term Grouping Ing up Term based i gnificance	4 Reduction n (Kappa Grouping	Wight Store (Control of the second	Senes
LlueGO Comparison View Settings Show Groups Show Difference	GO Term Fusio GO Term Fusio GO Term Conn GO Term Conn GO Term Conn GO Term Group Use GO 1 Group color Fix Leading Gro Highest Sid Tree 3	Min #Genes n (Redundance F ferm Fusion ection Restriction Medium bing (Functional ferm Grouping ing up Term based gnificance	4 Reduction n (Kappa Grouping on Init	Kappa	senes
ClueGO Comparison View Settings Show Groups Show Difference Uw GO Functional Analysis	GO Term Fusio GO Term Fusio GO Term Conn GO Term Group GO Term Group GO Term Group GO Term Group Fusion Highest Sig Tree 3 50	Min #Genes In (Redundance F erm Fusion Medium Joing (Functional erm Grouping Ing up Term based aprificance	4 Reduction n (Kappa Grouping on Init	Kappa ial Group	senes
Show Groups Show Difference	Cluxef #2 2 3 - 60 Term Fusio - 0 Use 60 Term Conn - 60 Term Group - 70 Term - 70 Te	Min #Genes n (Redundance IF term Fusion ection Restriction Medium ing (Functional term Grouping ing up Term based phificance	4 Reduction n (Kappa Grouping on - - - - - - - - - - - - - - - - - -	Score) High S	Senes
JueGO Eunctional Analysis	CO Term Fuel CO Term Fuel Use GO Term Conn CO Term Fuel Cov GO Term Conn Cov GO Term Conn Cov GO Term Group Fuel Fuel Fix -Leading Gro Highest Six 50	Min #Genes n (Redundance fe ern Fusion ection Restriction Medium bing (Functional errm Grouping ing up Term based anificance	4 Reduction n (Kappa Grouping on Init Setting	%64 %	Senes
aucoo ranctionar Analysis	Custer #2/ 2 3 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	Min #Genes n (Redundance F erm Fusion ection Restriction Medium Medium ing (Functional erm Grouping ing - up Term based - gnificance arison View S is	4 Reduction n (Kappa Grouping on Init v % 1	Score) High S Kappa Iaial Group S for Group Sow E Show E	Senes

Figure 21: ClueGO settings example

Customize the network using Cytoscape features

- select VizMapper (Cytoscape Control Panel)
- select Node Font Size
- set the value of FALSE (size for the name of the terms) to 0.001 and press Enter

- set the value of TRUE (size for the name of groups) to 26 and press Enter
- select Layout (Cytoscape menu bar)
- set scale to 1/3
- zoom the image
- change the position of the leading terms to make visible the name of the group
- change the possition of notgrouped terms if due to rescaling they are too close to a group
- for visualizing groups press Show Groups



Figure 22: ClueGO customized figure

Bibliography

- [1] 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.
- [2] M Kanehisa, S Goto, S Kawashima, A Nakaya. The KEGG databases at GenomeNet. Nucleic Acids Res, 30:42–46, 2002.
- [3] 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.
- [4] da W Huang, B T Sherman, Q Tan, J R Collins, W G Alvord, J Roayaei, R Stephens, M W Baseler, H C Lane, R A Lempicki. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 8:R183–R183, 2007.
- [5] I Rivals, L Personnaz, L Taing, MC Potier. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23:401–7, 2007.
- [6] O Garcia, C Saveanu, M Cline, M Fromont-Racine, A Jacquier, B Schwikowski, T Aittokallio. GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. *Bioinformatics*, 23:394–396, 2007.
- [7] J Cohen. A coefficient of agreement for nominal scales. Educ Psychol Meas, 20:37-46, 1960.