Comprehensive functional analysis of large lists of genes and proteins

Bernhard Mlecnik, Jérôme Galon, Gabriela Bindea

Comprehensive functional analysis of large lists of genes and proteins

Bernhard Mlecnik[1,2,3,4], Jérôme Galon[1,2,3], Gabriela Bindea[1,2,3,#]

[1] INSERM, UMRS1138, Laboratory of Integrative Cancer Immunology, F-75006, Paris, France. [2] Université Paris Descartes, Sorbonne Paris Cité, UMRS1138, F-75006, Paris, France. [3] Sorbonne Universités, UPMC Univ Paris 06, UMRS1138, Centre de Recherche des Cordeliers, F-75006, Paris, France. [4] Inovarion, 75013 Paris, France.

[#] Correspondence should be addressed to GB (gabriela.bindea@crc.jussieu.fr)

Abstract

The interpretation of high dimensional datasets resulting from genomic and proteomic experiments in a timely and efficient manner is challenging. ClueGO software is a Cytoscape App that extracts representative functional biological information for large lists of genes or proteins. The functional enrichment analysis is based on the latest publicly available data from multiple annotation and ontology resources that can be automatically accessed through ClueGO. Predefined settings for the selection of the terms are provided to facilitate the analysis. Results are visualized as networks in which Gene Ontology (GO) terms and pathways are grouped based on their biological role. Many species are now supported by ClueGO and additional organisms are added on demand. ClueGO can be used together with the CluePedia App to enable the visualization of protein-protein interactions within or between pathways.

Introduction

With today's high-throughput technologies a continuous large amount of data is generated. The rapid advancement of DNA and RNA next generation sequencing, microarrays as well as mass spectrometry based proteomic technologies [1] facilitates the investigation of global aspects of health and disease. Studies of gene expression levels reveal expression patterns in particular cell types and how these change at particular stages of development or in the context of disease [2]. On the other hand, proteomic techniques interrogate the entire repertoire of proteins of an organism and can underline interactions of proteins involved in diverse cellular functions [3]. Biological processes are in fact multidimensional, with multiple informational levels including genes that code for several proteins. These proteins are involved in multiple pathways and can be further post-translationally modified in complex ways. The interpretation of such high dimensional datasets

in a timely and efficient manner is challenging. Novel computational methods and software allow to organize, integrate and analyze such data. Combining transcriptomic and proteomic data sets can reveal interesting new facets of biological and cellular functions [4], that can lead to the discovery of new therapeutic agents [5] and predictive biomarkers for an improved disease classification [6] [7] [8]. These kind of approaches can thus advance personalized healthcare.

High dimensional data bring both, opportunities and new challenges [9], hence statistical analysis steps are increasingly important. Data generated by researches can be used in combination with publicly available data [10]. Recommendations are provided on how to select candidate markers from large scale data [11] and how to find their biological role. This article presents a typical analysis workflow of a large dataset in which functional analysis were performed with ClueGO [12] software (Fig. S1A). Answers to questions addressed by ClueGO users were included in the manuscript.

How is biological data stored and organized?

More and more biological information becomes available for scientists through an increasing number of pan genomic projects and major public repositories where experimental data as well as biological knowledge are stored and organized.

Array and sequence based data as well as other large scale datasets are hosted in Gene Expression Omnibus (GEO) [13] and ArrayExpress [14]. This data represent original research results submitted upon publication to facilitate an independent evaluation and reanalysis of these results. A broad range of biological themes include disease, development, evolution, immunity, ecology, toxicology, metabolism, and other topics [15]. Additionally, genome methylation, chromatin structure, genome copy number variations, and genome–protein interactions can be investigated.

On the other hand, structured, controlled vocabularies and classifications for several biological domains are provided within the Gene Ontology (GO) collaborative project [16]. GO vocabularies are continuously updated and expanded by expert curators [17]. GO is structured in three parts, biological process, cellular component and molecular function, in which definitions called terms are included in a hierarchical graph (Fig. 1A). This graph illustrates terms as nodes and the relations between the terms as edges (links). Similar to the terms, each relation between GO terms is categorized and well defined. Commonly used relationships are: is_a, part_of, has_part, regulates or negatively/positively regulates (http://www.geneontology.org/page/ontology-relations). Based on the distance to the ontology's root term (e.g. Biological Process), terms can be considered to be placed in different levels. Terms closer located to the root have more general definitions with a high number of genes associated, whereas terms deeper down located in the tree describe more specific biological processes with less genes associated [16]. Each association of a gene to a term is qualified with an evidence code that reflects how the association was assessed (Fig. 1B). Most relevant GO annotations reflect a curatorial review of published literature and are supported by experimental evidence. More than a quarter of the evidence codes are experimentally derived for

all the three parts of GO (Fig. 1B).

Pathway databases instead present biological processes as a succession of reactions among molecules that lead to a certain product or a functional change in the cell. Important manually curated resources for pathways are the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome. KEGG [18] is a knowledge base for systematic analyses of gene functions that link genomic information with higher order functional information. Reactions, pathways and biological processes can be investigated using Reactome [19]. Other resources are WikiPathways [20], a community-curated database, and BioCyc [21], where metabolic pathways and enzymes can be investigated.

High quality biomedical, genomic, proteomic and functional information for many species are provided by The National Center for Biotechnology Information (NCBI) [22], The Universal Protein Resource (UniProt) [23] and Ensembl [24]. CORUM database [25] stores manually curated protein complexes. Protein sequence analysis is enabled in InterPro [26] while the online database resource Search Tool for the Retrieval of Interacting Genes (STRING) [27] provides experimental as well as predicted protein interaction information. Other databases are centered on one model organism, like dictyBase for the amoeba Dictyostelium discoideum [28]. Curated reference genome databases for cyanobacteria and rhizobia provide an easy way of accessing their sequences and annotation data [29]. Reference ontologies for plants [30], as well as specific annotations for species like rice [31] or gramene [32] are available.

These resources can be directly accessed at their respective web sites or within Cytoscape [33], a major platform for biological network analysis. A variety of Apps [34] provide remote access and analysis of data. Alternatively, markers can be mapped simultaneously on multiple resources using Apps like ClueGO [12] and CluePedia [35].

How to select candidate genes/proteins?

After large scale data have been generated it is important to identify genes that are differentially expressed among the tested experimental conditions, and to investigate their biological role. These candidate genes can either support the experimental hypothesis, or can lead to new hypothesis generation. Genes representative for an experimental setting can be obtained by using software tools like R [36] or Apps like CluePedia, that can perform statistical tests through a user friendly interface. A common statistical approach is to calculate for each gene a p-value comparing two experimental conditions, adjust these p-values for multiple testing in case of large scale profiling, and create a list of candidate genes using an appropriate cut-off [11]. Other methods like the Gene Set Enrichment Analysis (GSEA) work with continuous data and search for genes enriched at the top or at the bottom of a ranked list containing all genes [37]. Another option is to test in large scale data a selection of markers a priori described to be involved in a biological process or pathology.

Candidate genes that might have a role in the tested hypothesis, should be further validated in independent datasets or by additional experimental techniques to confirm the discovery.

How to find representative biological functions for candidate genes

To investigate the biological role of candidate genes, ontologies and pathway databases can be used. This practical example illustrates the use of ClueGO for the functional analysis of these genes. The publicly available dataset GSE6887 [38] was downloaded from GEO. Gene expression data from blood of healthy donors was analyzed to select upregulated genes (n=200) in natural killer (NK) cells, important lymphocytes implicated in defense mechanisms. To investigate the most representative biological terms and pathways for these genes, ClueGO with predefined settings was used to perform gene enrichment analysis on GO (downloaded in 09.09.2016), KEGG and Reactome pathways.

Before performing the analysis, Cytoscape (http://cytoscape.org/) and ClueGO have to be installed. The Cytoscape App Manager (Fig. 2A) enables the automatic download of the latest version of ClueGO whereas older versions and their release history can be found in the Cytoscape App Store [39]. Users will be notified whenever a new version of ClueGO is available. It is recommended to always use the most recent version of ClueGO and Cytoscape, that contains latest bug fixes and new functionalities.

The ClueGO documentation (http://www.ici.upmc.fr/cluego/cluegoDocumentation.shtml) can be also consulted via the App Store, where users have the possibility to ask new questions or see FAQ concerning ClueGO's functionality.

The software is Java based and works on Windows, MacOS and Linux operating systems. ClueGO is freely available upon registration (http://www.ici.upmc.fr/cluego/cluegoLicense.shtml). At the first start up, the ClueGOConfiguration folder containing all ClueGO configurations, pre-compiled source files (ClueGOSourceFiles folder) and example gene lists (ClueGOExampleFiles folder) is created in the user's home directory.

Functional analysis with ClueGO is intuitive, and consists in 4 major steps. Additional features allow then to further customize the result output (step 5).

Step 1. Select the organism of interest and of the identifier type to analyze.

Once ClueGO has been started via the Cytoscape menu "Apps" (Fig. 2B), the control panel with all ClueGO parameters (Fig. 2C) appears on the left side of Cytoscape's main frame. After running the analysis, additional ClueGO features for customizing the network and saving results can be found in the ClueGO results panel, bellow the network (Fig. 2D). This example uses human data, so the standard settings can be kept ("Homo Sapiens", "#Automatic#").

ClueGO includes by default human and mouse data sources. More than 100 other organisms are available and prepared for download (Fig. 2C, *step 1.1*), including frequently used model

organisms. To see all supported organisms and to download additional ones, the “Download new organisms” button located next to the organism box has to be selected (Fig. 1E). Species that are already installed are marked in green. The organism of interest is then selected by clicking the “Download” button and the corresponding files will be added in the ClueGOConfiguration folder. The downloaded organism will be automatically included to the list of available organisms (*Step 1*). Additional organisms can be added to the repository upon request.

After downloading a new organism, the user should verify how recent the ClueGO data files are and if needed update the corresponding files with the latest information available from public data sources (Fig. 2C, *step 1.2*). After expanding the update feature (Fig. 2F), the ontology, pathway or annotation source can be selected and by clicking on the “Update” button the most recent data from the respective database is downloaded automatically, formatted and stored in the ClueGOConfiguration folder. Updates are important, to ensure the quality of gene annotations that directly impacts the results of the enrichment analysis [40]. Additionally, other customized gene annotation files can be used for the analysis.

Experimental data can be produced using a multitude of different techniques. For example, gene expression levels can be measured by Affymetrix microarray platforms or by quantitative real-time polymerase chain reaction (qPCR) techniques. More recently, the quantity of mRNA levels in biological samples can be assessed by next generation sequencing technologies (RNAseq). All these approaches rely on different probe, gene or protein identifiers that need to be recognized and link to their appropriate definition. To facilitate this issue, ClueGO automatically recognizes multiple identifier types and can be easily extended in case of unknown identifiers. Gene and protein information from NCBI, UniProt and Ensembl databases is extracted and various identifiers are converted from one type into another. Existing identifier conversion files are downloaded together with the organisms. To restrict the analysis to a certain identifier all available identifiers for a specie can be seen in the combo box menu next to the organism. ClueGO is centered on the NCBI EntrezGeneID for the majority of the organisms that are well annotated. Additional conversion files can be created and added to enable the analysis of data with other identifier types than yet supported by ClueGO.

Additionally ClueGO allows the analysis of metabolites annotated in KEGG and protein complexes from InterPro.

Step 2. Upload markers.

Marker identifier lists (Cluster) can be loaded from text files (first column of a tab delimited data file), as selected genes in an existing Cytoscape network (select “Network”) or they can be directly pasted into the text field. ClueGO example files or other lists with gene or protein identifiers can be accessed by clicking the button “Cluster #1”, like shown in Fig. 2C (s*tep 2*). The dialog that opens shows available example files and enables the navigation to other locations from where files with marker identifiers can be loaded (Fig. 2G). In this example the

"GSE6887_Nkcell_Healthy_top200UpRegulated.txt" list provided by ClueGO was used. Several marker lists (clusters) can be simultaneously analyzed to underline their common and specific biological functions. Additional marker lists can be added by selecting "+" below the cluster text field (Fig. 2C, *step 2.1*, Fig. 2H). Different colors and shapes are automatically attributed to each of the clusters, to facilitate the visualization of the origin of genes/proteins on the network. These parameters can be customized by clicking on the color button next to the cluster text field (Fig. 2H).

Step 3. Select ontology and pathway sources.

In this example the GO (09.09.2016), KEGG and Reactome data resources were used. The download date when the ontology or pathway data was created, as well as the number of pathways and associated genes or proteins is shown in the selection table (Fig. 2C, s*tep 3*). ClueGO performs gene enrichment analysis on pathways from multiple sources, that can be further identified by their customizable node shape on the network (Fig. 2C, s*tep 3*).

Additionally, particular evidence codes for the gene-term associations can be selected (Fig. 2C, *step 3.1*). Besides individual evidence codes provided by GO (Fig. 1B), new categories were defined by grouping all experimental based evidences and all evidences except the ones inferred from electronic annotations (IEA), respectively. The "Update Ontologies" and "Download New Organisms or Data" features can be accessed bellow the the "Ontologies section/Pathways" section (Fig. 2C, s*tep 1.2*).

Step 4. Run the analysis. Select "Start".

All selected ontologies are enriched with the uploaded cluster(s) and the results are shown as a network of pathways and terms grouped based on their common biological role. Each of the genes can have several functional associations, and in total, the 200 NK genes were associated with 3.366 terms and pathways (Fig. 3A), where four of these terms were found significant after multiple testing correction (Fig. 3B). These pathways: "cellular defense response", "immune system process", "immune response" and "defense response" describe very general immune related processes.

The predefined ClueGO parameters enable the selection of representative terms, neither too general nor too detailed. These parameters were optimized for lists with around 200 genes and take into account the GO level, the minimum number of genes to be associated with a term, and the minimum percentage that these genes represent among the total number of associated genes (Fig. 4A). Global terms from the upper part of the GO tree are unspecific and have many associated genes where the investigated genes generally only represent a very low percentage. On the other end of the GO tree terms are much more specific, with only few associated genes that represent very detailed biological functions. Depending on the number of genes in the list to analyze and the purpose of the analysis, the user can choose to represent the enrichment results as a general, medium or very detailed network of functions. For a more fin-grained analysis the

user can customize the parameters for the GO term selection in the “Advanced Term/Pathway Selection Options” section. After applying predefined ClueGO parameters for upregulated genes in NK cells, terms related to “natural killer cell mediated immunity”, “regulatory T cell differentiation”, “cellular defense response” were kept in the network with pathways like “Natural killer mediated cytotoxicity” and “Chemokine receptors bind chemokines” (Fig. 4B). Such terms and pathways illustrate well the biological role of the NK cells, immune subpopulation with cytotoxic properties. Through the release of proteins such as perforin or proteases like granzymes the NK cells are killing virus infected cells and respond to tumor formation. The NK cells are activated by cytokines and chemokines released by cells upon viral infection.

In the network, the terms are connected based on kappa score, a measure that takes into account how many genes are shared among two terms. All terms are compared to each other, and functional groups are defined using the kappa score, in a similar way as described by Huang *et al*. [41]. For each group, the most significant term is highlighted by a large name label. Functional groups are represented in a piechart (Fig. 4C), where the proportion of each group is calculated based on the number of the terms included in the group. Another representation illustrates the percentage of found genes per term as a bar chart (Fig. 4D), as illustrated for the “natural killer mediated immunity” group. The statistical significance is calculated for both, terms and groups and shown as: ** (pValue < 0.001), * (0.001 < pValue < 0.05), . (0.05 < pValue < 0.1).

Visualizing and saving results

ClueGO provides several interchangeable visual styles that can be applied to represent the results on the network. By default, the “Groups” visual style (Fig. 2C) will be applied, that colors nodes based on their functional groups (Fig. 4B). The size of the nodes indicates their significance where the most significant pathways are illustrated with the largest node size. Beside this representation, the “Significance” style is available, that colors pathways based on their significance, with most significant terms shown in dark red, and with non significant terms in gray, like illustrated in Fig. 3A. With this visual style, the size of the nodes indicates the number of genes associated with this term, thus pathways with high number of gene associations will be shown as largest nodes on the network. In addition, if several lists of markers are investigated, the origin of these clusters of genes/proteins can be shown on the network by selecting the “Clusters” visual style. Pathways in which more than 60% of the genes are originating from one of the clusters, will be shown on the network with the predefined color of this cluster. The visual styles are extended to networks of pathways and genes, if CluePedia is installed.

If the list to analyze comprises many well annotated genes, the resulting network with a high number of terms might be difficult to interpret. To reduce the complexity of the network users can apply the fusion option (“Use GO Term Fusion”, Fig. 2C, *step 5*). This feature compares genes associated with GO terms found in a parent - child relation and if shared genes are similar, only the more informative of these two terms will be kept. Users can choose as well to visualize only

significant pathways ("Show only Pathways with pV≤", Fig. 2C, *step 5*). Further, a custom selection of the pathways is available using parameters defined in the "Advanced Term/Pathway Selection Options". GO terms found in a particular level on the ontology tree can be selected (Fig. 2C, *step 5*) and filters for pathways having a minimum number and/or percentage of genes from the user list can be defined. To get all GO terms, "All" has to be selected as maximum GO level, as well as in the minimum number of genes/term.

In the figure 4E, the upregulated NK genes were analyzed again, but this time additionally to the standard ClueGO term selection (Fig. 4B) the fusion was applied. The number of terms kept in the network was much reduced, and e.g. only 6 out of 17 terms from the "regulation of regulatory T cell differentiation" group were kept. In a simplified view of the network, only the name of the most significant term per group is shown using the "Show/Hide small labels" feature (Fig. 2D, Fig. 4F). The color of the groups can be adjusted ("Change group colors" feature, Fig. 2D). Additionally also other terms of interests can be selected in the ClueGO table and highlighted on the network using Update term labels feature (Fig. 2D), like illustrated with the "Natural killer cell mediated cytotoxicity" pathway (Fig. 4F).

The ontologies and the selection criteria used are summarized in a log file (Fig. S1B). The hypergeometric test (Fisher's Exact Test), a simple and efficient method [11], is used to calculate the statistical significance of the overrepresentation of enriched genes in pathways and functional groups, taking as reference all the genes from selected ontologies (Fig. S1C). The Bonferroni step-down method corrects by default for multiple testing. Other statistical methods are available in the "Statistical Options" (Fig. 2C).

All information about terms and groups including their statistical significance is included in the ClueGO result table (Fig. 4G). This table can be saved together with other figures and tables that summarize functional groups and their associated genes, or enriched genes and their biological role as a list or as a gene-term matrix (Fig. 2D, S1D). Interrelations calculated based on shared genes between all term-term pairs are saved in a kappa score matrix. The entire ClueGO project can be saved (Cytoscape, File, "Save ClueGO Session As" feature) and reopened to continue the analysis. The network and the other graphical representations can be saved as scalable vector graphics, and modified with vector graphic editor software like Inkscape (https://inkscape.org).

Limitations

Gene annotation data is stored at many locations in online data repositories, and is based on vocabularies and definitions that are often specific for individual resources. The usage of multiple data sources increases the completeness of available information, a feature that is provided by ClueGO. Annotations are continuously improved by the curators, an effort that takes a lot of time because of the manual curation process. With the multitude of publications, most recent findings may not have yet been included into the annotation sources. Scientists should contribute to the effort of systematically organizing the data by submitting to GO newly refined annotations [42].

Most of the databases are developed as independent software that, in the absence of established naming conventions, use custom generated identifiers for genes/proteins. These identifiers are continuously updated in different resources and the conversion from one type to another can be challenging. In addition, due to the constant development of these data sources and repositories, the web addresses where the different data sources can be accessed are sometimes changing. Since such changes directly impact the automatic update of source files, recent versions of ClueGO use QuickGO [43] to access GO data.

Conclusions

Publicly available data are now widely used in research, often integrated with new data generated by researchers. Software that perform such type of analyses to select candidate markers are more and more needed. ClueGO performs up to date functional analyses for large lists of genes and proteins revealing new biological insights that add to previous knowledge. Compared to many other tools that represent the results as long lists of terms, ClueGO visualizes GO terms and pathways as networks, and groups them based on their biological role. Several lists of markers can be simultaneously analyzed to underline their common or specific functions. To facilitate the analysis, ClueGO provides predefined settings for the selection of the terms. In addition, pathways and terms representative for the investigated genes/proteins, their significance and interrelations in functional groups are directly visualized on the network. Several ClueGO visual styles can be used to highlight different important information on nodes and edges. The network and other figures can be exported and saved as high quality images. Data sources, including the date when the ontology files were created, the version of the software used as well as the parameters applied to select pathways should be mentioned to ensure the reproducibility of the results. Many species are now supported by ClueGO, and additional organisms will be included on demand. ClueGO is extended by CluePedia, that enables the analysis of experimental data and the visualization of protein-protein interrelations within a pathway. Based on experimental derived or *in silico* scores, new genes/proteins potentially associated with a pathway can be found. New versions of the software will continue to support the community of ClueGO and Cytoscape with new features and visualizations.

Acknowledgements

The authors would like to thank ClueGO users for their feedback and valuable suggestions that essentially contribute to the development of the software.

Funding

This work was supported by grants from INSERM, the Cancer research for personalized medicine (CARPEM), the LabEx Immuno-oncology.

References

[1] D. Penque, T. Simões and F. Amado. Proteomics advances in the last decade: What is next?, J Proteomics 75(1) (2011) : 1-3.

[2] G. Bindea, B. Mlecnik, M. Tosolini, A. Kirilovsky, M. Waldner, A. Obenauf, H. Angell, T. Fredriksen, L. Lafontaine, A. Berger, P. Bruneval, W. Fridman, C. Becker, F. Pagès, M. Speicher, Z. Trajanoski and J. Galon. Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer , Immunity 39 (2013) : 782 - 795.

[3] P. de Sousa-Pereira, M. Cova, J. Abrantes, R. Ferreira, F. Trindade, A. Barros, P. Gomes, B. Colaço, F. Amado, P. Esteves and R. Vitorino. Cross-species comparison of mammalian saliva using an LC-MALDI based proteomic approach, Proteomics 15(9) (2015) : 1598-607.

[4] P. Bastos, J. da Costa and R. Vitorino. A glimpse into the modulation of post-translational modifications of human-colonizing bacteria., J Proteomics 152 (2017) : 254-275.

[5] J. da Costa, V. Carvalhais, F. Amado, A. Silva, R. Nogueira-Ferreira, R. Ferreira, L. Helguero and R. Vitorino. Anti-tumoral activity of human salivary peptides, Peptides 71 (2015) : 170-8.

[6] B. Mlecnik, G. Bindea, H. Angell, M. Sasso, A. Obenauf, T. Fredriksen, L. Lafontaine, A. Bilocq, A. Kirilovsky, M. Tosolini, M. Waldner, A. Berger, W. Fridman, A. Rafii, V. Valge-Archer, F. Pagès, M. Speicher and J. Galon. Functional network pipeline reveals genetic determinants associated with in situ lymphocyte proliferation and survival of cancer patients, Sci Transl Med 6(228) (2014) : 228ra37.

[7] B. Mlecnik, G. Bindea, A. Kirilovsky, H. Angell, A. Obenauf, M. Tosolini, S. Church, P. Maby, A. Vasaturo, M. Angelova, T. Fredriksen, S. Mauger, M. Waldner, A. Berger, M. Speicher, F. Pagès, V. Valge-Archer and J. Galon. The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis, Sci Transl Med 8(327) (2016) : 327ra26.

[8] B. Mlecnik, G. Bindea, H. Angell, P. Maby, M. Angelova, D. Tougeron, S. Church, L. Lafontaine, M. Fischer, T. Fredriksen, M. Sasso, A. Bilocq, A. Kirilovsky, A. Obenauf, M. Hamieh, A. Berger, P. Bruneval, J. Tuech, J. Sabourin, F. Le Pessot, J. Mauillon, A. Rafii, P. Laurent-Puig, M. Speicher, Z. Trajanoski, P. Michel, R. Sesboüe, T. Frebourg, F. Pagès, V. Valge-Archer, J. Latouche and J. Galon. Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability, Immunity 44(3) (2016) : 698-711.

[9] J. Fan, F. Han and H. Liu. Challenges of Big Data Analysis, National science review 1(2) (2014) : 293-314.

[10] J. Rung and A. Brazma. Reuse of public genome-wide gene expression data, Nat Rev Genet 14(2) (2013) : 89-99.

[11] R. A. Irizarry, C. Wang, Y. Zhou and T. P. Speed. Gene set enrichment analysis made simple, Statistical Methods in Medical Research 18 (2009) : 565-575.

[12] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, WH Fridman, F. Pages, Z. Trajanoski and J. Galon. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, Bioinformatics 25(8) (2009) : 1091-1093.

[13] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A.

Soboleva, M. Tomashevsky and R. Edgar. NCBI GEO: mining tens of millions of expression profiles—database and tools update, Nucleic Acids Research 35 (2007) : D760-D765.

[14] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar and A. Brazma. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments., Nucleic acids research 39 (2011) : D1002-D1004.

[15] T. Barrett. Gene Expression Omnibus (GEO), The NCBI Handbook[Internet]. 2nd edition .

[16] 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 and G. Sherlock. Gene ontology: tool for the unification of biology The Gene Ontology Consortium, Nat Genet 25 (2000) : 25-29.

[17] R. Huntley, T. Sawford, P. Mutowo-Muellenet, A. Shypitsyna, C. Bonilla, M. Martin and C. O'Donovan. The GOA database: Gene Ontology annotation updates for 2015, Nucleic Acids Research 43(Database issue):D1057-63.

[18] M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya. The KEGG databases at GenomeNet, Nucleic Acids Res 30 (2002) : 42-46.

[19] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio and L. Stein. Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Research 39 (2011) : D691-D697.

[20] A. Pico, T. Kelder, M. van Iersel, K. Hanspers, B. Conklin and C. Evelo. WikiPathways: pathway editing for the people, PLoS Biol 6(7):e184.

[21] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases, Nucleic Acids Research 42 (2014) : D459-D471.

[22] N. R. Coordinators. Database resources of the National Center for Biotechnology Information, Nucleic Acids Research 44(Database issue), D7–D19.

[23] UniProtConsortium. The Universal Protein Resource (UniProt), Nucleic Acids Research 35 (2007) : D193-D197.

[24] 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. Girón, 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 and P. Flicek. Ensembl 2017, Nucleic Acids Research .

[25] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone and H.-W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes—2009, Nucleic Acids Research 38 (2010) : D497-D501.

[26] A. Mitchell, H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas and R. D. Finn. The InterPro protein families database: the classification resource after 15 years, Nucleic Acids Research 43 (2015) : D213-D221.

[27] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. v. Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, Nucleic Acids Research 39 (2011) : D561-D568.

[28] S. Basu, P. Fey, Y. Pandit, R. Dodson, W. A. Kibbe and R. L. Chisholm. dictyBase 2013: integrating multiple Dictyostelid species, Nucleic Acids Research 41 (2013) : D676-D683.

[29] T. Fujisawa, S. Okamoto, T. Katayama, M. Nakao, H. Yoshimura, H. Kajiya-Kanegae, S. Yamamoto, C. Yano, Y. Yanaka, H. Maita, T. Kaneko, S. Tabata and Y. Nakamura. CyanoBase and RhizoBase: databases of manually curated annotations for cyanobacterial and rhizobial genomes, Nucleic Acids Research 42 (2014) : D666-D670.

[30] A. Deans, S. Lewis, E. Huala, S. Anzaldo, M. Ashburner, J. Balhoff, D. Blackburn, J. Blake, J. Burleigh, B. Chanet, L. Cooper, M. Courtot, S. Csosz, H. Cui, W. Dahdul, S. Das, T. Dececchi, A. Dettai, R. Diogo, R. Druzinsky, M. Dumontier, N. Franz, F. Friedrich, G. Gkoutos, M. Haendel, L. Harmon, T. Hayamizu, Y. He, H. Hines, N. Ibrahim, L. Jackson, P. Jaiswal, C. James-Zorn, S. Kohler, G. Lecointre, H. Lapp, C. Lawrence, N. Le Novere, J. Lundberg, J. Macklin, A. Mast, P. Midford, I. Miko, C. Mungall, A. Oellrich, D. Osumi-Sutherland, H. Parkinson, M. Ramírez, S. Richter, P. Robinson, A. Ruttenberg, K. Schulz, E. Segerdell, K. Seltmann, M. Sharkey, A. Smith, B. Smith, C. Specht, R. Squires, R. Thacker, A. Thessen, J. Fernandez-Triana, M. Vihinen, P. Vize, L. Vogt, C. Wall, R. Walls, M. Westerfeld, R. Wharton, C. Wirkner, J. Woolley, M. Yoder, A. Zorn and P. Mabee. Finding Our Way through Phenotypes, PLoS Biology 13(1).

[31] H. Sakai, S. S. Lee, T. Tanaka, H. Numa, J. Kim, Y. Kawahara, H. Wakimoto, C.-c. Yang, M. Iwamoto, T. Abe, Y. Yamada, A. Muto, H. Inokuchi, T. Ikemura, T. Matsumoto, T. Sasaki and T. Itoh. Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics, Plant and Cell Physiology 54 (2013) : e6.

[32] M. K. Tello-Ruiz, J. Stein, S. Wei, J. Preece, A. Olson, S. Naithani, V. Amarasinghe, P. Dharmawardhana, Y. Jiao, J. Mulvaney, S. Kumari, K. Chougule, J. Elser, B. Wang, J. Thomason, D. M. Bolser, A. Kerhornou, B. Walts, N. A. Fonseca, L. Huerta, M. Keays, Y. A. Tang, H. Parkinson, A. Fabregat, S. McKay, J. Weiser, P. D'Eustachio, L. Stein, R. Petryszak, P. J. Kersey, P. Jaiswal and D. Ware. Gramene 2016: comparative plant genomics and pathway resources, Nucleic Acids Research .

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

[34] R. Saito, M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, S. Lotia, A. R. Pico, Bader Gary D and T. Ideker. A travel guide to Cytoscape plugins, Nat Methods 9 (2012) : 1069-1076.

[35] G. Bindea, J. Galon and B. Mlecnik. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data, Bioinformatics 29(5) (2013) : 661-3.

[36] R Development Core Team. R: A Language and Environment for Statistical Computing, .

[37] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirova. Gene Expression Omnibus (GEO), Proc Natl Acad Sci U S A 102(43) (2005) : 15545–15550.

[38] R. J. Critchley-Thorne, N. Yan, S. Nacu, J. Weber, S. P. Holmes and P. P. Lee. Down-Regulation of the Interferon Signaling Pathway in T Lymphocytes from Patients with Metastatic Melanoma, PLoS Medicine 4(5).

[39] S. Lotia, J. Montojo, Y. Dong, G. D. Bader and A. R. Pico. Cytoscape App Store, Bioinformatics 29(10) (2013) : 1350–1351.

[40] L. Wadi, M. Meyer, J. Weiser, L. Stein and J. Reimand. Impact of outdated gene annotations on pathway enrichment analysis, Nature Methods 13 (2016) : 705–706.

[41] d. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane and R. A. Lempicki. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists, Genome Biol 8 (2007) : R183-R183.

[42] R. C. Lovering. How Does the Scientific Community Contribute to Gene Ontology?, Methods Mol Biol 1446 (2017) : 85-93.

[43] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan and R. Apweiler. QuickGO: a web-based tool for Gene Ontology searching, Bioinformatics 25(22):3045-6.

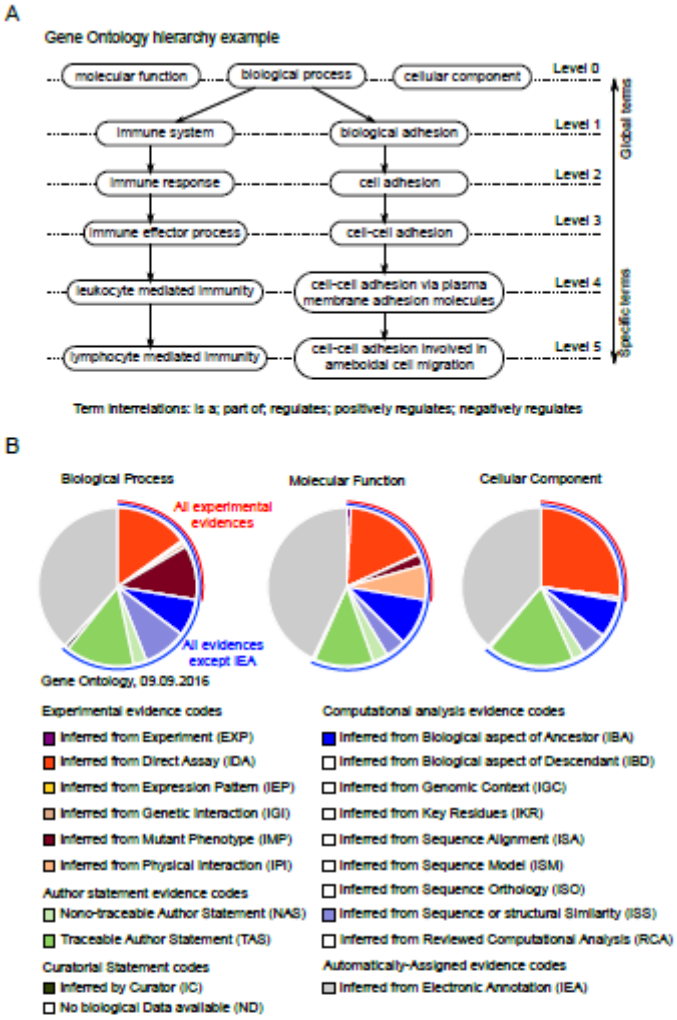


Figure 1

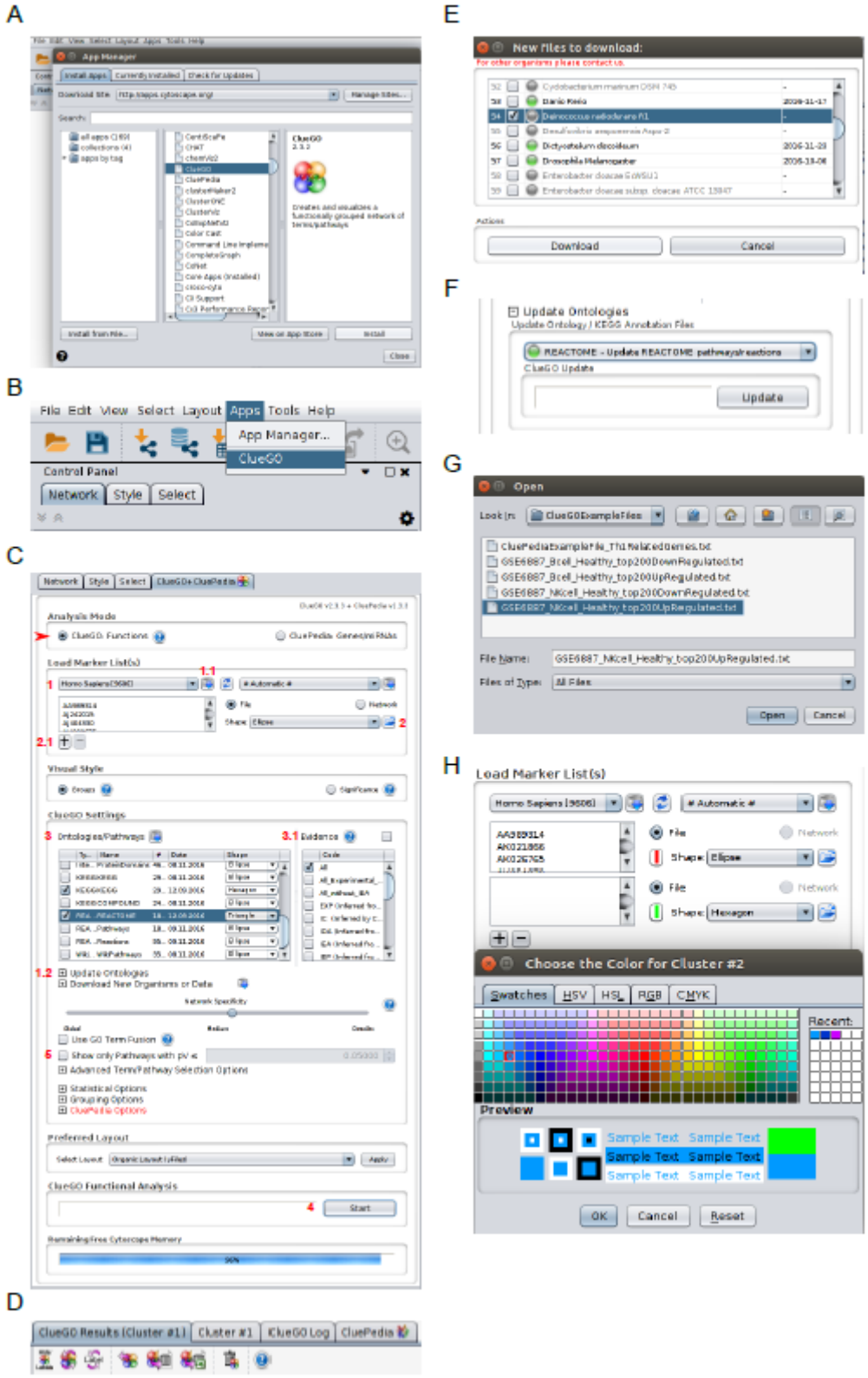
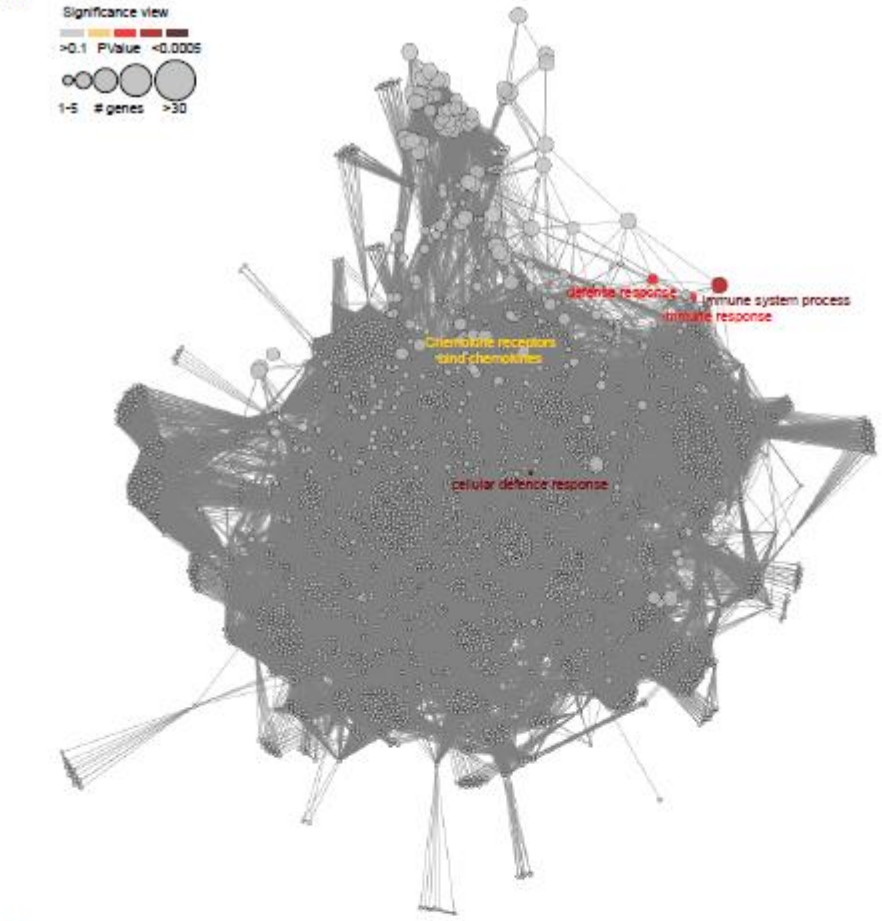


Figure 2

A



B

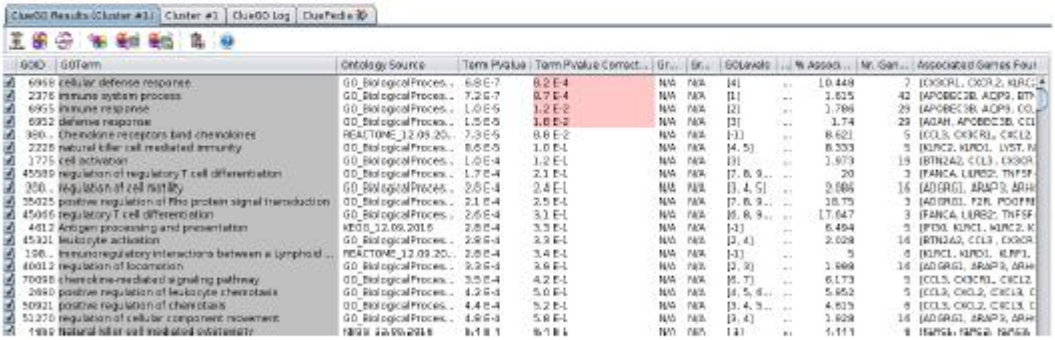


Figure 3

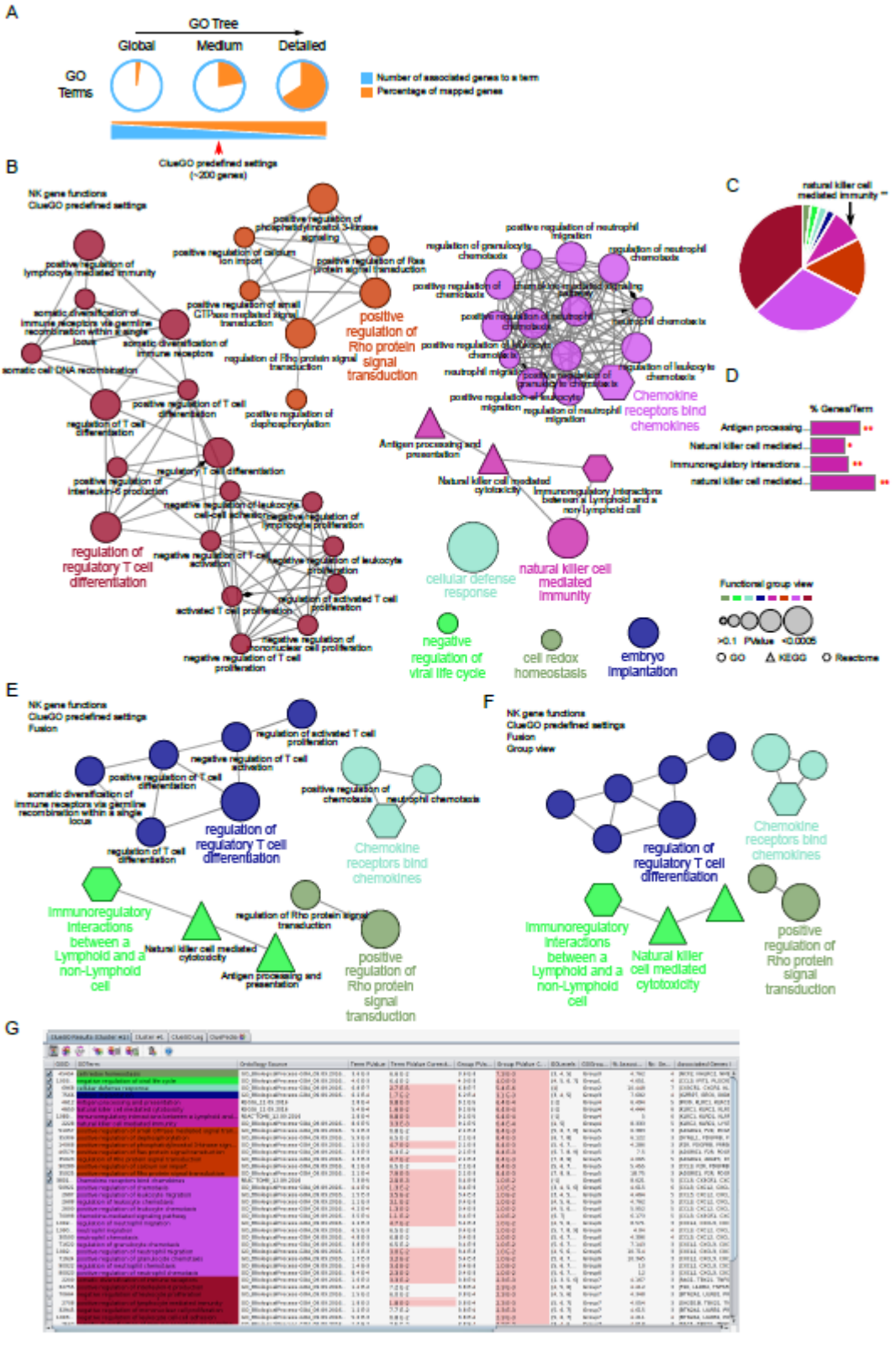
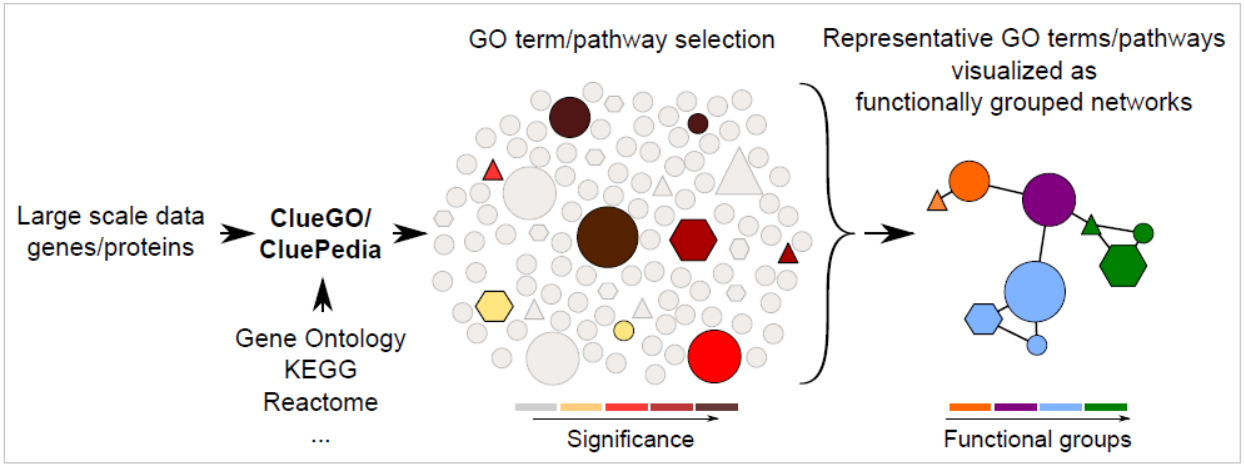


Figure 4



Graphical Abstract

Comprehensive functional analysis of large lists of genes and proteins

Bernhard Mlecnik[1,2,3,4], Jérôme Galon[1,2,3], Gabriela Bindea[1,2,3,#]

[1] INSERM, UMRS1138, Laboratory of Integrative Cancer Immunology, F-75006, Paris, France. [2] Université Paris Descartes, Sorbonne Paris Cité, UMRS1138, F-75006, Paris, France. [3] Sorbonne Universités, UPMC Univ Paris 06, UMRS1138, Centre de Recherche des Cordeliers, F-75006, Paris, France. [4] Inovarion, 75013 Paris, France.

[#] Correspondence should be addressed to GB (gabriela.bindea@crc.jussieu.fr)

Conflict of interests

The authors declare that they have no competing interests.

Highlights

- Omics data analysis and interpretation is challenging
- Public repositories and projects store and organize experimental data and biological knowledge
- ClueGO extracts representative biological information for large lists of genes or proteins and visualizes pathways in functional grouped networks
- Intuitive visualization facilitates the interpretation of complex biological data