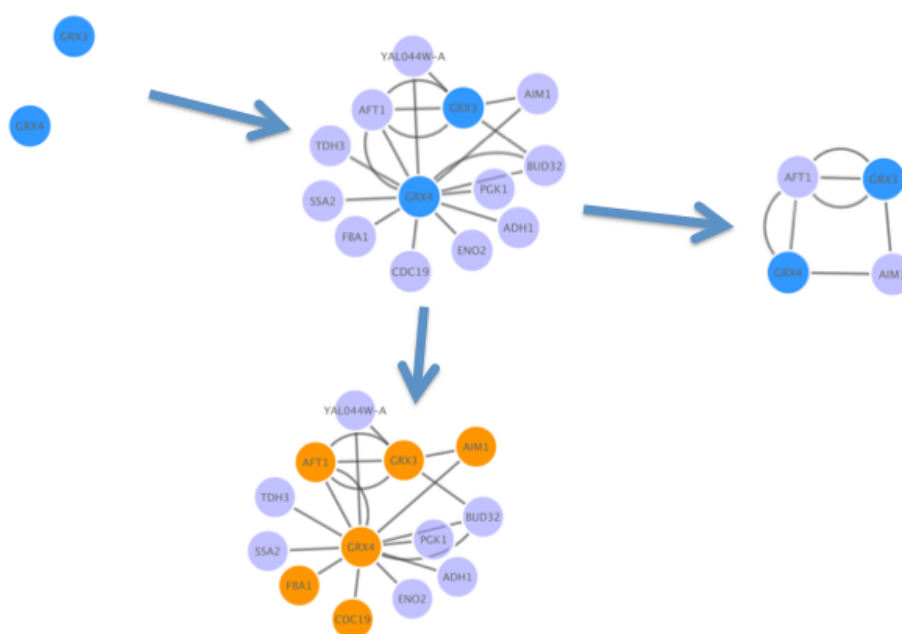


(v18, 1/12/15)

Network generation and analysis through Cytoscape and PSICQUIC



Author: Pablo Porras Millán

IntAct Scientific Database Curator

Contents

Summary.....	3
Objectives	3
Software requirements.....	3
Introduction to Cytoscape.....	3
Tutorial.....	4
Dataset description	4
Representing an interaction network using Cytoscape	7
Navigating through different networks.....	8
Selecting with edge and node columns.....	8
Integrating quantitative proteomics data: Loading columns from a user-generated table	10
Using the visual representation features of Cytoscape	12
Network clustering: finding topological clusters with clusterMaker2	13
Analysing network annotations: using BiNGO for functional annotation	16
Additional information	21
Installing apps in Cytoscape 3.3.....	21
Obtaining a GOSlim and species-specific annotation files.....	21
a) Generate a species-specific annotation file with QuickGO.....	22
b) Obtain a GO slim annotation file	22
Further reading.....	23
Links to useful resources	24
Contact details.....	24

Summary

The study of the interactome –the totality of the protein-protein interactions taking place in a cell– has experienced an enormous growth in the last few years. Biological networks representation and analysis has become an everyday tool for many biologists and bioinformatics, as these interaction graphs allow us to map and characterize signalling pathways and predict the function of unknown proteins. However, given the size and complexity of interactome datasets, extracting meaningful information from interaction networks can be a daunting task. Many different tools and approaches can be used to build, represent and analyse biological networks. In this tutorial, we will use a practical example to guide novice users through this process, making use of the popular open source tool Cytoscape and of other resources such as the PSICQUIC client to access several protein interaction repositories at the same time, the clusterMaker2 plugin to find topological clusters within the resulting network and the BiNGO app to perform GO enrichment analysis of the network as a whole or in its clusters as found by clusterMaker2.

Objectives

With the present tutorial you will learn the following skills and concepts:

- To build a molecular interaction network by fetching interaction information from a public database using the PSICQUIC client built in in the open source software tool Cytoscape.
- To load and represent that interaction network in Cytoscape.
- The basic concepts underlying network analysis and representation in Cytoscape: the use of visual styles, columns, filters and plugins.
- To integrate and make use of quantitative proteomics data in the network.
- To find highly interconnected groups of node, named clusters, using the clusterMaker2 Cytoscape plugin.
- To add Gene Ontology annotation to a protein interaction network.
- To use the BiNGO Cytoscape plugin to identify representative elements of GO annotation and to combine this approach with quantitative proteomics data to learn more about the biology represented in the network.

Software requirements

Cytoscape version 3.3.0 (downloadable from www.cytoscape.org) including the BiNGO 3.0.3 (www.psb.ugent.be/cbd/papers/BiNGO) and the clusterMaker2 0.9.5 apps (www.cgl.ucsf.edu/cytoscape/clusterMaker2/). See ‘Additional information’ for installation instructions.

Introduction to Cytoscape

Cytoscape 3 is an open source, publicly available network visualization and analysis tool (www.cytoscape.org) [1]. It is written in Java and will work on any machine running a Java Virtual Machine, including Windows, Mac OSX and Linux. The version we will use in this tutorial is 3.3.0, but you can have multiple versions installed in your computer if required. This version of Cytoscape requires having Java version 1.8 or newer installed to work. Version number in Cytoscape is very relevant depending on the analysis you want to perform, since some old apps (called plugins in the past) only work on the 2.x series. Updated versions are made available for the current 3.x series regularly.

Cytoscape is widely used in biological network analysis and it supports many use cases in molecular and systems biology, genomics and proteomics:

- It can import and load molecular and genetic interaction datasets in several formats.
 - ✓ In this tutorial, we will import a molecular interaction network fetching data from IMEx-complying databases, such as IntAct or MINT, using the Cytoscape built-in PSICQUIC client.

- It can make effective use of several visual features that can effectively highlight key aspects of the elements of the network. This can be saved in the form of visual styles, exported and imported for re-use.
 - ✓ We will use node and edge tables to represent quantitative proteomics data and interaction features.
- It can project and integrate global datasets and functional annotations.
 - ✓ We will make use of resources such as the Gene Ontology to annotate the interacting partners in our network.
- It has a wide variety of advanced analysis and modelling tools in the form of apps that can be easily installed and applied to different approaches.
 - ✓ The BiNGO app will be used to perform GO enrichment analysis and the clusterMaker2 app will identify topological clusters, so we will use them to try to identify the functional modules underlying our network.
- It allows visualization and analysis of human-curated pathway datasets such as Reactome or KEGG.

Tutorial

Dataset description

In order to easily illustrate the concepts discussed in this tutorial, we are going to follow a guided analysis example using a dataset from a work published by Ju *et al.* in 2014 [2]. Our working dataset is going to be a list of proteins coming from a proteomic analysis of A549 human lung adenocarcinoma cells treated with Multi-Walled Carbon Nanotubes (MWCNTs). The authors want to see if this material, increasingly used in the nanomaterials field, could have deleterious effects on exposed cells. One of the approaches they take to test this is treating the cells with increasing concentrations of MWCNTs in suspension and checking the effects this has on the proteome by monitoring expression over time using a combination of 2D-electrophoresis and MS identification of differentially expressed proteins. They provide a table (table 2) in which a list of 48 proteins whose concentration is significantly altered in one of the several conditions tested. We are going to use a modified version of this table to integrate this information into a molecular interactions network.

Generating an interaction network using the PSICQUIC client of Cytoscape

We are going to generate a protein interaction network that will help us identify the biological functions associated with those kinases identified in both regulatory and effector T cells. To do this, we will find out which proteins are interacting with the ones represented in the dataset as stored in some of the different molecular interaction databases that comply with the IMEx guidelines [3]. It is good practice to use IMEx-complying data if you aim to get only experimentally-derived data, since all databases that curate to these standards use the same type of identifiers and follow the same criteria when recording the data, so it is possible to merge datasets with minimum risk of redundancies or duplications. Here is a list of the databases that we will use:

- IntAct (www.ebi.ac.uk/intact): One of the largest available repositories for curated molecular interactions data, storing PPIs as well as interactions involving other molecules [4]. The European Bioinformatics Institute hosts it. IntAct has evolved into a multi-source curation platform and many other databases, such as MINT, I2D, InnateDB, UniProt or MatrixDB curate into IntAct and make their data available through it.
- MINT (mint.bio.uniroma2.it/mint): MINT (Molecular INteraction database) focuses on experimentally verified protein-protein interactions mined from the scientific literature by expert curators [5]. It is hosted in the University of Rome. MINT data has been recently integrated to the IntAct curation platform, so to access the most updated version of their data you need to check either the IntAct website or use PSICQUIC (see below).
- MatrixDB (matrixdb.ibcp.fr): Database focused on interactions of molecules in the extracellular matrix, particularly those established by extracellular proteins and

polysaccharides [6]. The data in MatrixDB comes from their own curation efforts, from other partners in the IMEx consortium and from the HPRD database. It also contains experimental data from the lab of professor Ricard-Blum in the Institut de Biologie et Chimie des Protéines in the University of Lyon, where it is hosted. Like MINT, it can be accessed through the IntAct website as well.

- DIP (dip.doe-mbi.ucla.edu/dip): DIP (Database of Interacting Proteins) is hosted in the University of California, Los Angeles and contains both curated data and computationally-predicted interactions [7].
- I2D (ophid.utoronto.ca/i2d): I2D (Interologous Interaction Database, formerly OPHID) integrates known, experimental (derived from curation) and predicted PPIs for five different model organisms and human [8]. It is hosted in the Ontario Cancer Institute in Toronto. Like MINT and MatrixDB, it curates into the IntAct curation platform and it can be queried from the IntAct website.
- UniProt (www.uniprot.org) and BHF-UCL (www.ucl.ac.uk/functional-gene-annotation/cardiovascular): although not interactions databases, UniProt and University College of London (UCL) curators do introduce interaction information into IntAct, and thus their data gets credited as a separate entities when you query it through PSICQUIC.
- InnateDB (www.innatedb.com/): Publicly available database of the genes, proteins, experimentally-verified interactions and signaling pathways involved in the innate immune response of humans, mice and bovines to microbial infection. The Brinkman and Hancock laboratories, at the Simon Fraser University and the University of British Columbia in Vancouver, host it jointly. It is another of the partners that use the IntAct curation platform, so their data can be accessed through IntAct as well.
- Molecular Connections (MolCon, www.molecularconnections.com): Private company specialized in the integration and analysis of scientific information. They do some curation work using the IntAct platform and make their molecular interaction data public through IntAct as well.

We will use the Proteomics Standard Initiative Common QUery InterfaCe (PSICQUIC) importing client built into Cytoscape. PSICQUIC is an effort from the HUPO Proteomics Standard Initiative (HUPO-PSI, www.hupo.org/research/psi/) to standardise the access to molecular interaction databases programmatically, specifying a standard web service with a list of defined accessing methods and a common query language that can be used to search from data in many different databases. If you want to have more information about PSICQUIC, check their GitHub page at github.com/micommunity/psicquic or have a look at the Nature Methods publication where the client is described [9]. PSICQUIC allows you to access data from many different databases, like Reactome (www.reactome.org) [10], the pathways database hosted in the EBI; but we will limit our search to those resources that comply with the IMEx consortium curation rules (www.imexconsortium.org/curation) as listed before.



There are several ways to get molecular interaction data into Cytoscape apart from the one we present here. For example, from the IntAct web page, the user can generate files in tab-delimited or in Cytoscape-compatible XGMML formats that can be later imported into this software.

1. Open the file 'table2_formatted.xlsx'. This is an updated version of table 2 in Ju *et al.* where the relative quantity estimations have been formatted to avoid empty spaces and the different conditions have been separated in different columns. UniProt accessions are given in the table in the column "Swiss-prot ID"¹.

¹ UniProtKB identifiers are widely used among the different resources we are going to need along the tutorial, so it is highly recommended to use them when dealing with protein datasets. The advantages of using these ACs are that (i) they are stable (they are not changed or updated once assigned); (ii) they can reflect isoform information, if provided; and (iii) they are recognized by many interaction and annotation databases (in this instance, the two databases we will be using: IntAct and GO). To map other types of accessions to UniProt you can use the ID mapping tool they provide in their website (www.uniprot.org) or the PICR service

- Open Cytoscape and go to 'File' → 'Import' → 'Network' → 'Public Databases'. In the window that will appear, you will see as pre-selected the 'Interaction Database Universal Client' option from the 'Data source' drop-down menu. To search for the interactions in which the proteins from your list are involved, you just have to paste the list of the UniProt AC identifiers in the 'Enter Search Conditions' query box and click 'Search'².

A	B	C	D
Biological	Spot No	Swiss-prot ID	Protein
1			
2	Cell metal	1109 Q04760	Lactoylg
3	Cell metal	4209 P30084	Enoyl-C
4	Cell metal	4305 Q13011	Delta(3,
5	Cell metal	6204 P47985	Ubiquin
6	Cell metal	6205 P18669	Phospho
7	Cell metal	6301 P00491	Purine r
8	Cell metal	6304 P10768	S-formy
9	Cell metal	7407 (12 h	O96008
10	Cell signal	101 Q15185	Prostagl
11	Cell signal	2203 (12 h	P52565
12	Cell signal	2315 O00299	Chlorox
13	Cell signal	8106 P30086	Phospho
14	Cellular pi	1006 P09382	Galectin
15	Cellular pi	1201 P04632	Calpain
16	Cellular pi	2203 (24 h	P04792
17	Cellular pi	2303 (12 h	P08758
18	Cellular pi	306 P62258	14-3-3 p
19	Cellular pi	3204 P09211	Glutathi

copy & paste

- You will see that in the 'Select Database' box just below the numbers of interactions found by PSICQUIC among the different databases (or 'services') that the client can access are updated. You can then select the appropriate ones depending on your requirements.
- For the selection of the source of our interactions, we will stick to just IMEx-complying datasets. You should get interactions from IntAct, DIP, I2D-IMEx, MINT, BHF-UCL, MolCon, UniProt and MatrixDB, among other resources that store predicted interactions or pathways or are just not IMEx-compatible. We will ignore those to avoid problems while merging the data from the different repositories. Notice that some databases, such as I2D or InnateDB, identify a subset of their interactions as 'IMEx-complying'. The number of interactions found for each database changes with time, because they are constantly updated. Select just the IMEx-complying datasets we mentioned before in the 'Import' column and then click 'Import'.



The 'Database Type (Tags)' column in the interactor importer in Cytoscape gives a short description of the type of data you can find in each database accessed through PSICQUIC, indicating which of those host IMEx-complying data. The same information can be found in the PSICQUIC Registry:

www.ebi.ac.uk/Tools/webservices/psicquic/registry/registry?action=STATUS

- You will get yet another dialog box from which you will have a list of your databases of choice and the option to manually merge the results from them or just have them in separated networks. Click 'Yes' and the 'Advanced network merge' assistant will pop up (see screenshot next page).
- Now the 'Advanced Network Merge' assistant will open up. Select the networks you want to merge (in our case, all of them) and then click on the 'Advanced Network Merge' menu (on the little black triangle on its right side) to select the identifier you will use as a common ID for the merge. In our case, we are merging protein-protein interaction information and we will use UniProtKB accessions as our primary identifier. You will see

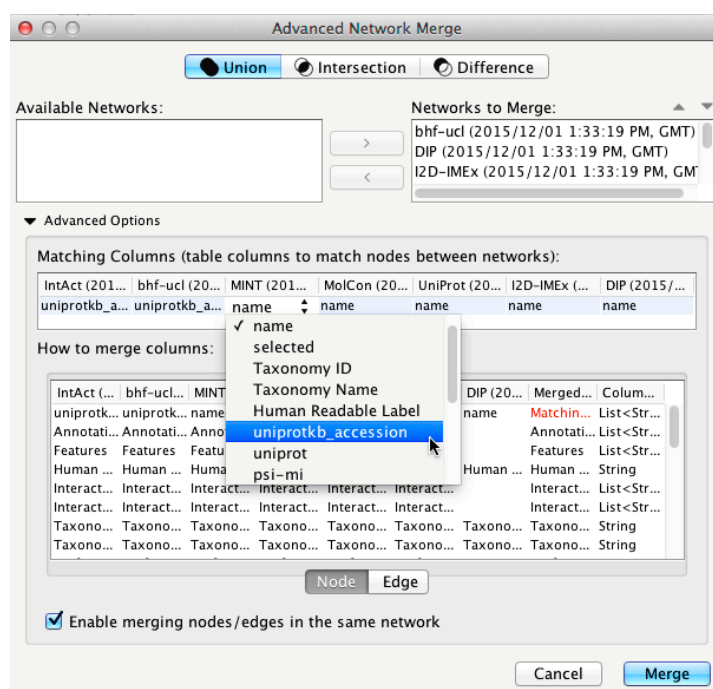
(Protein Identifier Cross-Reference Service) that can be accessed in www.ebi.ac.uk/Tools/picr.

² You can also perform queries using this tool by clicking on the 'Search mode' drop-down menu and selecting 'Search by Query Language (MIQL)'. Then you can search using TaxIDs, gene names or interaction detection methods and build complex queries with the MIQL syntax reference (check www.ebi.ac.uk/Tools/webservices/psicquic/view and click on the 'MIQL syntax reference' link you will find in the far-right upper corner by the search bar in that page).

a drop-down menu appearing for each network you select to be merged. In each drop-down menu you will find a list of the ‘columns’ that each node or edge of the network is assigned during the import. We will talk more about columns later, for now; just select the column ‘uniprotkb_accession’ in each menu. This column contains the UniProtKB AC for each node, so the merging can proceed properly.

- Finally, the interaction database universal client will create several networks. A different network will be created for each of the resources that were accessed by PSICQUIC and will be named accordingly. The final one will be called ‘Merged Network’ and is the one we will use for our analysis. The networks will look like a grid of squares (nodes) connected by many lines (edges). We will learn how to make sense of it in the following sections of the tutorial.
- Finally, since Cytoscape can be tricky (and buggy) and you don’t want your precious time to be wasted, **save your session** (go to ‘File’ → ‘Save’, click on the floppy disk icon up left or just press ‘Ctrl + s’ on a Linux or Windows machine or ‘Cmd + s’ on a Mac). A piece of advice: do this every time you want to try something new with Cytoscape, since going back to your initial file is sometimes not possible and you can waste a lot of time re-doing a lot of work!

NOTE: If you didn’t click ‘yes’ in step 5 you will not go through the merge networks routine. Don’t panic. You can just go to ‘Tools’ → ‘Merge’ → ‘Networks...’ and follow the instructions from step 6 onwards.



Representing an interaction network using Cytoscape

Finding a meaningful representation for your network can be more challenging than you might expect. Cytoscape provides a large number of options to customize the layout, colouring and other visual features of your network. This tutorial does not aim to be exhaustive in exploring the capabilities of Cytoscape; we just want to give you the basics. More detailed information and basic and advanced tutorials can be found in their documentation page: www.cytoscape.org/documentation/users.html.

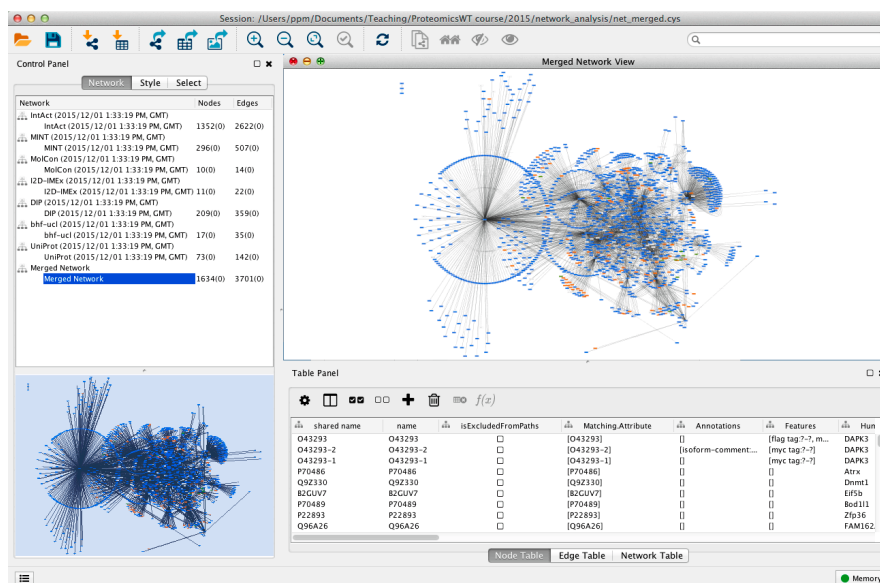
Now we will learn how to use the basic tools that Cytoscape provides to manage the appearance of your network and make the information that it provides easier to understand.

- If it is the first time you use Cytoscape, have a look at the user interface and get familiar with it. The main window displays the network (all the network manipulations and ‘working’ will be visualized in this window). The lower-right pane (the Data Panel)

contains three tabs that show tabulated information about node, edge and network tables. The left-hand pane (the Control Panel) is where navigation, visualization, editing and filtering options are displayed.

- By default, Cytoscape lays out all the nodes in a grid, so that is why your network is looking so ugly. You can change the layout going to the 'Layout' menu. There is a wide range of different layouts that will help displaying certain aspects of the network, like which proteins have a large number of interaction partners (the so called 'hubs'). Give some of them a try and stick to the one you prefer, like the 'organic' layout shown in the following screenshot.

Save your session




Navigating through different networks

A Cytoscape session, saved as a .cys file, can hold more than one network, as you have noticed after finishing the import. The 'Network' tab in the Control Panel allows us to navigate from one network to another and, by use of the right-click, to change the names or delete the networks we have stored in our session. One concept that was created in the 3.x versions of Cytoscape is that of "network collection", reflected in the 'Network' tab as a hierarchy where you can see different networks grouped under the same network collection (see screenshot on the right). Network collections help grouping together networks that share the same type of columns, for example, or networks that are "children" of another network. More about columns in the next section.


Control Panel		
Network Style Select		
Network	Nodes	Edges
IntAct (2015/12/01 1:33:19 PM, GMT)		
IntAct (2015/12/01 1:33:19 PM, GMT)	1352(0)	2622(0)
MINT (2015/12/01 1:33:19 PM, GMT)		
MINT (2015/12/01 1:33:19 PM, GMT)	296(0)	507(0)
MolCon (2015/12/01 1:33:19 PM, GMT)		
MolCon (2015/12/01 1:33:19 PM, GMT)	10(0)	14(0)
I2D-IMEx (2015/12/01 1:33:19 PM, GMT)		
I2D-IMEx (2015/12/01 1:33:19 PM, GMT)	11(0)	22(0)
DIP (2015/12/01 1:33:19 PM, GMT)		
DIP (2015/12/01 1:33:19 PM, GMT)	209(0)	359(0)
bhf-uct (2015/12/01 1:33:19 PM, GMT)		
bhf-uct (2015/12/01 1:33:19 PM, GMT)	17(0)	35(0)
UniProt (2015/12/01 1:33:19 PM, GMT)		
UniProt (2015/12/01 1:33:19 PM, GMT)	73(0)	142(0)
Merged Network		
Merged Network	1634(16...	3701(0)
over-12h30	8(8)	1(0)
Merged Network--clustered	1634(0)	2828(0)
wholenet_cc		
wholenet_cc	648(0)	1291(0)

Selecting with edge and node columns

In network graphs, interacting partners are represented as **nodes**, which are objects represented as circles, squares, plain text... that are connected by **edges**, the lines depicting the interactions. All information referred to an interacting partner or an interaction must then be loaded in Cytoscape as a node or an edge **column**. A column can be a string of text, a number (integer or floating point) or even a Boolean operator and can be used to load information and represent it as a visual feature of the network. For example, a confidence score for a given interaction between two participants represented as nodes can be represented as the thickness of the edge connecting those nodes.

Columns can be created and loaded directly in Cytoscape using the ‘Create New Column’ icon  and then values can be added individually (double-clicking on any cell), to a subset of selected nodes/edges or to the whole column (by right-clicking on a single value and then selecting ‘Apply to entire column’ or ‘Apply to selected nodes/edges’). The columns can also be imported from data tables defined by the user or from external resources, as we will see later, and directly imported with the network from different network formats, as we will see right now.

Because we have used the PSICQUIC interaction database universal client, the information we took from the different PPI databases will be represented complying with the PSI-MI-2.7 tabular format³, so the fields requested by the format will be loaded as columns and we can start making use of them right away.

1. Let’s have a look at the columns that have been loaded with our network. First, select all the nodes and edges of the network.
2. Have a look at the Data Panel below the main window. By default, you should be in the ‘Node Table’ tab. You can see a number of columns being listed there; some of them with obvious meaning and some others whose content may not be so clear to you. It might be interesting to clear this view a bit, so only meaningful information is shown.
3. Click on the ‘Show Column’ icon . All the columns that have been loaded from the XGMML file will now be visible as a selectable list. Choose the following node columns to be displayed and try to figure out their meaning:
 - name
 - Human Readable Label
 - Interactor Type
 - Interactor Type ID
 - Taxonomy name
 - Taxonomy ID
 - uniprotkb_accession
 - Features
 - Annotation
4. Now go to the ‘Edge Table’ tab and do the same with the following edge columns:
 - interaction
 - Annotation
 - Author
 - Complex Expansion
 - Confidence-Score-intact-miscore / -author-score
 - Detection Method
 - Host Organism Taxonomy
 - Interaction Type / Primary Interaction Type
 - Publication DB
 - Publication ID
 - Source / Target Biological Role
 - Source / Target Experimental Role
 - Source / Target Participant Detection Method
 - Xref
 - Xref ID
 - Parameters

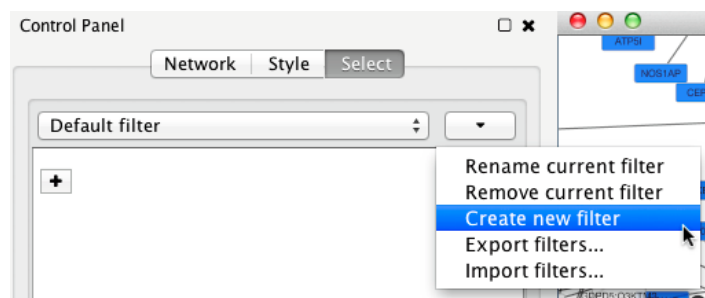
Save your session


Let’s make use of some of these columns. Sometimes, homolog proteins coming from different

³ The PSI-MI-TAB-2.7 format is part of the PSI-MI standard and it was originally derived from the tabular format that the BioGrid database used. You can learn more about the fields represented in the format checking their Google Code wiki at github.com/MICCommunity/psimi/blob/wiki/PsimiTab27Format.md.

species are used to perform interaction experiments. For this reason there is a number of ‘human-other species’ interactions in the databases. Now we will use the ‘Taxonomy’ node column to produce a human proteins-only network.

1. In the Control Panel, go to the ‘Select’ tab (see next screenshot).



2. Choose ‘Create new filter’ in the far-right drop-down menu and give your filter a name (e.g., ‘human only’).
3. Go to the ‘+’ icon and select to create a ‘Column Filter’. Choose the column you want to use for filtering. In this case, we will use the node column ‘Taxonomy ID’. Select it and you will get a search bar and two drop-down menus: one called with the name of the column you selected and the other in which you can select the operator you want to use for the search (‘contains’, ‘doesn’t contain’, ‘is’, ‘is not’ and ‘contains regex’).
4. The search bar can be used to type the value you want to select for. The ‘Taxonomy ID’ column stores NCBI taxonomy identifiers for the species origin of each protein in the network. The code for human is ‘9606’, write it down in the search bar and then click ‘Apply’.
5. The nodes that bear the ‘9606’ column will be then selected and highlighted in the network. Combinations of different columns can be applied by adding more selection criteria using the ‘+’ icon.
6. Now generate a new network containing only human proteins by going to ‘File’ → ‘New’ → ‘Network’ → ‘From Selected Nodes, All Edges’. Alternatively, you can click the quick ‘New Network From Selection’ button .

Save your session

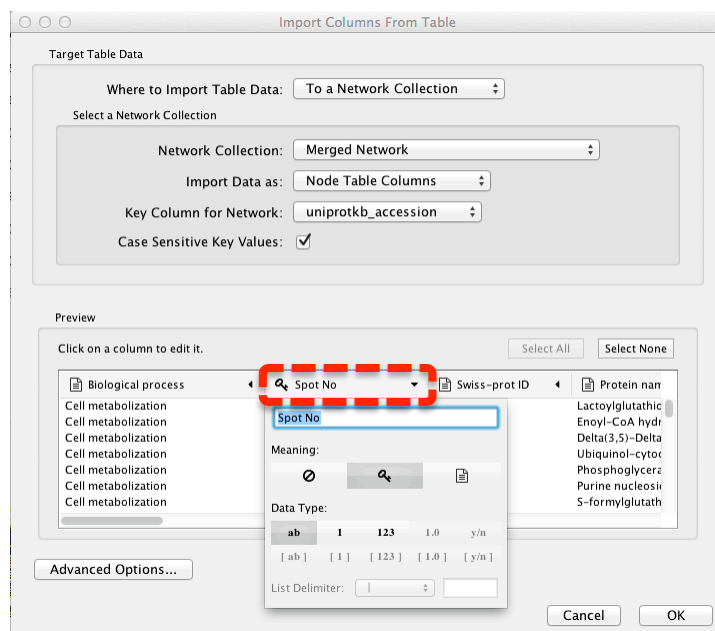



Multiple methodologies can be used for PPI detection, each method entailing its own strength and weaknesses and none of them being perfect, since every PPI detection approach must be considered artefactual to some degree (several reviews on the subject are recommended in the ‘Additional information’ section at the end of the tutorial). Nevertheless, sometimes you want to look at interactions found with a particular methodology. Use edge columns to create a network in which all the interactions have been found using the ‘two hybrid’ method.

Integrating quantitative proteomics data: Loading columns from a user-generated table

In order to load large amounts of information associated with the proteins in our network, it is often useful to import user-defined tables containing external data that can complement the network analysis. In our particular case, we will make use of the relative expression values that are given in table 2 of our selected publication in order to highlight the proteins that are enriched over exposure to different concentrations of MWCNTs. Since no interaction information was extracted from the original article, the information we put in will be exclusively node-centric (no edge annotations) and can be loaded in the form of a user-produced table.

1. Open the 'table2_formatted.xlsx' file. This is an adaptation of the table 2 in the original article. Have a look at the different fields and figure out what is represented in each column.
2. In Cytoscape, go to 'File' → 'Import' → 'Table' → 'File...'. Select the 'table2_formatted.xlsx' file and the 'Import Column From Table' wizard will pop up (next screenshot).



3. First, have a look at the 'Target Table Data' header section. There you can select to which network collection do you wish to apply the imported columns or if you prefer to restrict them to specific networks. This becomes of practical importance particularly when you run different types of analysis and you want to integrate different types of columns to different networks or collections. Select the 'Merged Network' collection in this case.
4. In the same menu you have to select also the column already existing in the network that will be used to map the values from the file you are importing. In the 'Key Column for Network' drop-down menu, select 'uniprotkb_accession' as the correct column to do the mapping.
5. Now check the 'Advanced Options...' menu. It allows you to import the first line of a text file as column names, to choose the separator used in delimited files or to skip commented lines at the beginning of the file. In our case we do not need to worry about these issues, since we are using an excel file with no comment lines.
6. In order to choose the primary key from the file that will map with the key column in the network, we need to click on the column name. This opens a small menu (see previous screenshot) in which you can define which is the column to be used as key. Once selected, it will get a small key icon to identify it. Notice that this menu also allows you to discard columns for import and to select the type of data that each column holds (text, integer, double or Boolean, plus lists of multiple values of those types).
7. Click 'Import' to finish the process.
8. Finally, the new node columns should be already visible in the 'Table Panel'. You can select which you prefer to use with the 'Show Columns' icon . Notice that only the proteins that were part of the original proteomics dataset from the paper have values in the newly imported columns.

Save your session



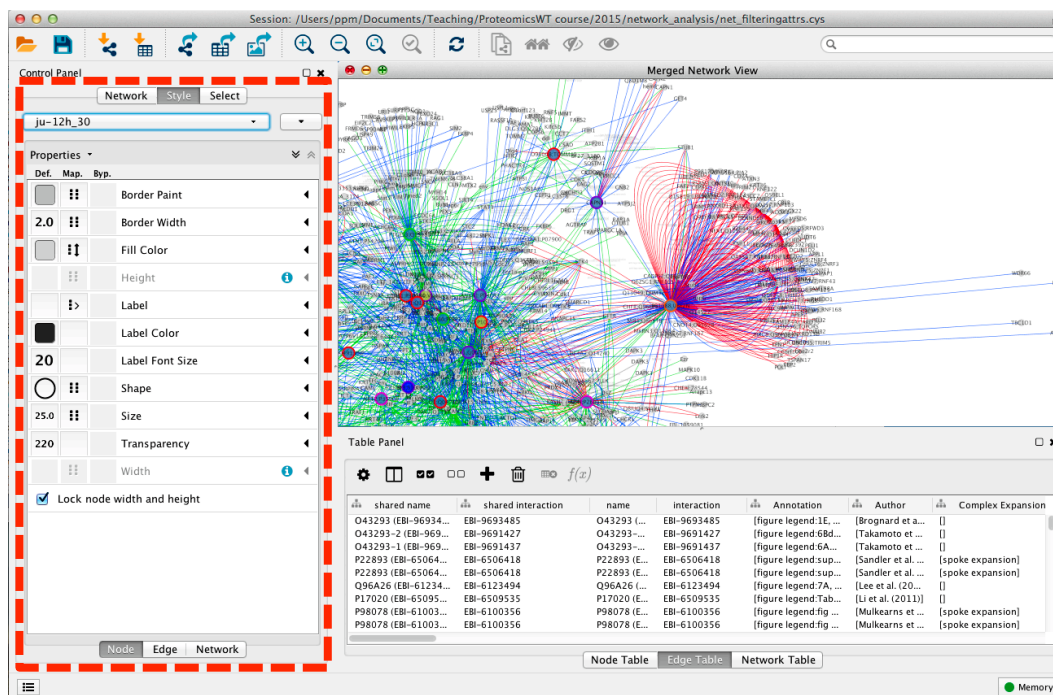
Try to create a sub-network to see how the proteins that are over-represented after 12 hours of treatment with 30 µg/ml of MWCNTs are connected. Make use of filters and the ‘Create new networks for selected nodes, all edges’ function. Try first picking only those protein for which you can find values and then try again creating a network including their 1st-level interacting partners.

Using the visual representation features of Cytoscape

After having integrated the quantitative proteomics information from the publication in the form of node columns, we can use the visual style editor of Cytoscape to represent this information in our network in a meaningful way. The ‘Style’ tab in the Control Panel controls all the visual features of a network, features that are saved in the form of ‘styles’. In a style, the default visual features of the network, such as the size of the nodes or the colour of the edges, are defined and columns can be used to define specific characteristics for specific column values. For example, the thickness of the edges in a PPI network can depend on a confidence score for the interaction it represents. Visual styles can be saved and re-used if it is necessary. We are going to import a pre-created style to visualize the new columns that we imported to our network.

1. Go to the ‘Style’ tab in the Control Panel and check the drop-down menu on the top of the tab. Here you can select different visual styles to apply to your network. Have a play with some of the default types and see how the properties listed below also change depending on the style you apply.
2. Now we are going to import a new visual styles file, one that includes a style specifically developed with this network in mind. Go to ‘File’ → ‘Import’ → ‘Style...’. Select the file ‘custom_style.xml’.
3. In the styles tab, select the ‘ju-12h_30’ style from the drop-down menu. The representation of the network will then change.
4. The changes of the visual features of the network are controlled through the ‘Properties’ menu, where properties can be chosen and columns loaded to be used for differential display of each one of them. Notice that there are separate tabs to control properties referred to nodes and edges. You can take some time to check which properties have been used to highlight certain aspects of the network and which columns were mapped to them.
5. Now you have a representation in which we can easily differentiate between the original protein dataset, in which quantitative proteomics data has been integrated and represented, and its interactome context as given by PSICQUIC (see next screenshot).

Save your session



Try creating a new network in which you represent the relative quantification of the proteins affected by the exposure to 30 $\mu\text{g/ml}$ of MWCNTs after 24 hours.

Network clustering: finding topological clusters with clusterMaker2

The study of the protein interactome is essentially the study of how proteins work together. The strategies that aim to interpret PPINs generally try to find common attributes within members of the network. Nodes may be grouped on the basis of network topology: groups of highly interconnected nodes may form clusters. Although clusters are identified solely on the basis of the topology, the assumption underlying this approach is that clusters will identify groups of proteins that share a similar function.

clusterMaker2 is a Cytoscape app developed at UCSF that allows the user to easily create a visualize topological clusters using a great variety of methodologies⁴. Their website (www.rbvi.ucsf.edu/cytoscape/clusterMaker2) provides extensive documentation about each method and some useful guided examples that illustrate how this app can be used.

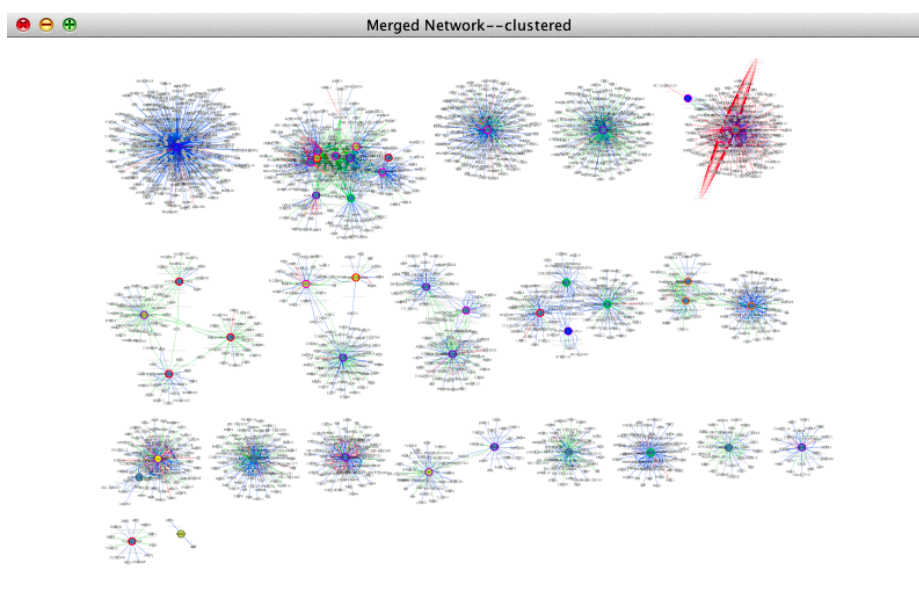
We are going to first use the GLayer Community Clustering algorithm due to its ease of use. Community clustering analysis was originally developed for the study of social networks. The algorithm begins by simplifying the network to give it a community-like structure by removing duplicate edges (you count each friend only once) and self-looping (you cannot be friends with yourself). The clusterMaker2 app has incorporated the GLayer implementation of the Newman-Girvan fast greedy algorithm [13,14]. The algorithm identifies clusters by iteratively removing edges from the network and then checking to see which groups of nodes are still connected, aiming to find heavily connected sub-networks.

1. Select all the nodes in the network, excluding the orphan interactions.

⁴ There are many different strategies to find topological clusters. For example, Brohée and van Helden evaluated four methods for the detection of previously annotated complexes [11]. Whichever clustering strategy is chosen, the question of greatest interest to the biologist is how well a particular algorithm identifies biologically meaningful protein complexes. Cluster-detection algorithms remain an active area of research and the interested reader is referred to a review by Wang et al. [12].

2. Now start clusterMaker2. Go to 'Apps' → 'clusterMaker'. Have a look at all the different methods that are listed in the menu.
3. We can then choose our favourite algorithm (Community clustering in this case). Choose 'Apps' → 'clusterMaker' → 'Community cluster (GLay)'.
4. A pop-up menu appears with some check boxes where you can select whether to use the whole network as searching space or just a subset of selected nodes and to allow the algorithm use directionality in its calculations or not. This last feature is not of interest for us, since our edges are undirected. Click on the 'Cytoscape Advanced Settings' to have a look at the options there.
5. The only thing that concerns us under 'Cytoscape Advanced Settings' is the 'Cluster column' section. There you have a name for a column that will identify the cluster to which each node belongs.
6. Now check the 'Visualization Options' box. You can create a new network that will show the newly found clusters grouped together if you click on the 'Create new clustered network' tick-box. If you want to keep the interactions connecting these clusters, you can do so by clicking the 'Restore inter-cluster edges after layout' tick-box.
7. Now click 'Ok' to start the algorithm and wait a few seconds until you get another pop-up giving you the results.
8. If you did not click the 'Create new clustered network' option, it would seem not much happened. To create a view, just go to 'clusterMaker Visualizations' → 'Create new Networks from Clusters'.
9. As seen in the next screenshot, several clusters are produced and laid out in a new window, arranged in decreasing order of size. The '___glayCluster' node column created by clusterMaker2 can be used to identify the clusters. This information can then be exported (see the 'File' → 'Export' → 'Table' menu) for further analysis in other tools or to be used to highlight clusters in the full network by defining a new visual style and colour them accordingly in this view or in the big network.

Save your session



Now if you want to have more control over the way the clusters are found, we recommend to use the Molecular Complex DETection (MCODE) algorithm [15]. This fast and versatile tool uses a three-stage process to find highly connected complexes in a network. Its default setting is much

more conservative than the GLay Community Clustering algorithm and will find less clusters and with a lower number of nodes. The process works as follows:

- a) **Weighting:** the algorithm gives a higher score to those nodes whose neighbours are more interconnected.
- b) **Molecular complex prediction:** starting with the highest-weighted node (seed), the algorithm recursively moves out and adds nodes to the complex that are above a given threshold. This threshold value is calculated by multiplying a user-defined cut-off by the seed node score. This way, the bigger the cut-off, the bigger the clusters you will find.
- c) **Post-processing,** which applies filters to improve the cluster quality. It goes through two optional processes: haircutting and fluffing. The haircut option drops all nodes from the cluster if they only have a single connection to it. The fluffing option expands the clusters by one step if the nodes have a score greater than the node score cut-off.

These are the advanced tuning options that the MCODE implementation in clusterMaker2 has, as can be found in the clusterMaker2 documentation webpage (see www.rbvi.ucsf.edu/cytoscape/clusterMaker2/#mcode):

- **Network Scoring**
 - a. **Include loops:** If checked, loops (self-edges) are included in the calculation for the vertex weighting. This shouldn't have much impact.
 - b. **Degree Cutoff:** This value controls the minimum degree necessary for a node to be scored. Nodes with less than this number of connections will be excluded.
- **Cluster Finding**
 - a. **Haircut:** If checked, drops all of nodes from a cluster if they only have a single connection to the cluster.
 - b. **Fluff:** If checked, after haircutting (if checked) all of the cluster cores are expanded by one step and added to the cluster if the score is greater than the Node Score Cutoff.
 - c. **K-Core:** Filters out clusters that do not contain a maximally interconnected sub-cluster of at least k degrees.
 - d. **Max Depth:** Controls how far out from the seed node the algorithm will search in the molecular complex prediction step.



Try to repeat the clustering search using MCODE this time. What is the main difference to GLay?

Try different values for the advanced parameters in MCODE and see how that affects your results.

Analysing network annotations: using BiNGO for functional annotation

Protein interaction networks can be used as backbones in which to set up the elements of new pathways or functions; but in order to be able to do that, we need to have access to information about the elements of the network. We can make use of the functional annotation that is associated to genes and proteins to enrich our network with such information. One of the most important resources that annotate genes and proteins is the Gene Ontology (GO) project [16], which provides structured vocabulary terms for describing gene product characteristics⁵. However, just incorporating raw GO annotation to a relatively large list of proteins will tell us very little, since the amount of information we integrate is just too much to handle manually. Some of the terms will be redundant as well, and distributed through many of the proteins represented in our list or network. GO enrichment analysis aims to figure out which terms are over- or under-represented in the population, thus extracting the most important biological features that can be learned from that particular set of proteins.

For starters, you need to have solid knowledge about the biological and experimental background of the data you are analysing to draw meaningful conclusions. For example, if you analyse a list of genes that are overexpressed in a lab cell line, you have to be aware that cell lines are essentially cancer cells that have adapted to live in Petri dishes. You will find a lot of terms related to negative regulation of apoptosis, cell adhesion or cell cycle control; but that just reflects the genetic background your cells have.

It is also important to take into account that certain areas of the gene ontology are more thoroughly annotated than others, just because there is more research done in some particular fields of biology than in others, so you have to be cautious when drawing conclusions. GO terms are assigned either by a human curator that performs manual, careful annotation or by computational approaches that use the basis of manual annotation to infer which terms would properly describe uncharted gene products. They use a number of different criteria always referred to annotated gene products, such as sequence or structural similarity or phylogenetic closeness. The importance of the computationally derived annotations is quite significant, since they account for roughly 99% of the annotations that can be found in GO. If you do not want to use computationally inferred annotations in your analysis, they can be filtered out by excluding those terms assigned with the evidence code 'IEA' (Inferred from Electronic Annotation). Most analysis tools support this feature. Since there are significantly less manual than automatic annotations, this also tends to simplify the output of your analysis, which can be very useful, as we will see later.

Finally, another factor that will make the analysis of GO annotation challenging is the level of detail and complexity you can reach when annotating large datasets. GO terms can describe very specific processes or functions -what is called 'granularity'- and it is often the case that even the result of a GO enrichment analysis is way too complex to understand due to the large number of granular terms that come up. In order to solve this problem, specific sets of GO annotation that are trimmed down in order to reduce the level of detail and the complexity in the annotation are provided by GO or can be created by a user in need of a specific region of the ontology to be 'slimmed'. Check www.geneontology.org/GO.slims.shtml and the 'Additional Information' section to learn more about GO slim. Apart from that, some tools, such as ClueGO [17], give the option to cluster together related terms of the ontology, highlighting groups of related, granular terms together.

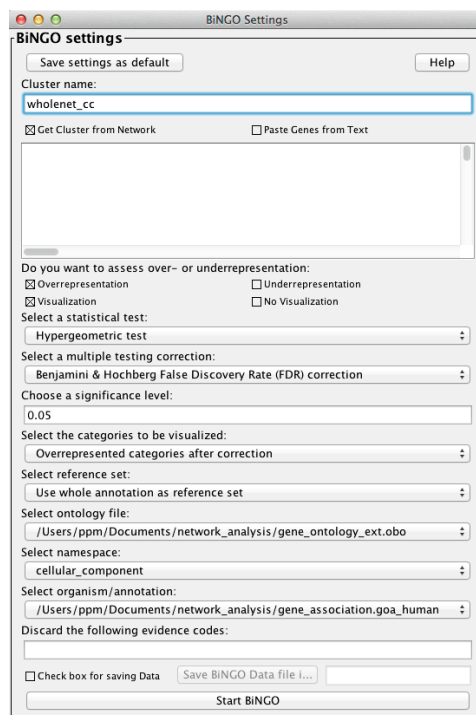
In order to perform network-scale ontology analysis, we are going to use the BiNGO tool (www.psb.ugent.be/cbd/papers/BiNGO), a Cytoscape app that annotates proteins (nodes) with gene ontology (GO) terms and then performs an enrichment analysis in order to figure out which terms are over- or under-represented in the population [18]. BiNGO will help us by providing an answer to this basic question:

⁵ The GO project is an international initiative that aims to provide consistent descriptions of gene products (i.e., proteins). These descriptions are taken from controlled, hierarchically organized vocabularies called 'ontologies'. GO uses three ontologies covering three biological domains. These are (1) Cellular Component, or the location of the protein within the cell (e.g., cytosol or mitochondrion); (2) Biological Process, or a series of events accomplished by one or more ordered assemblies of molecular functions (e.g., glycolysis or apoptosis); and (3) Molecular Function, which is the activity proteins possess at a molecular level (e.g., catalytic activity or trans-membrane transporter activity). More information can be found in their website, geneontology.org.

'When sampling X proteins (test set) out of N proteins (reference set; graph or annotation), what is the probability that x or more of these proteins belong to a functional category C shared by n of the N proteins in the reference set.'

The main advantages of BiNGO with respect to other enrichment analysis tools is that it is very easy to use and it can be complemented with the basic network manipulation and analysis tools that Cytoscape offers. It also can provide its results in the form of a network that can be further manipulated in Cytoscape, a feature that eases the analysis, and it can be used in combination with its sister tool PiNGO [19], which can be used to find candidate genes for a specific GO term in interaction networks. On top of that, it is relatively lightweight when it comes to usage of computer resources and it can be run with reasonable speed in any desktop computer. On the negative side, it is not as customizable and does not offer as many visualization options as the more advanced tool ClueGO, for example, which is displacing BiNGO in terms of popularity and is also improving dramatically its performance. We feel ClueGO is a bit too complex to give an introduction in a course of this nature, so we decided to go for BiNGO to show you the basics of GO enrichment analysis. Whichever the plugin you finally decide to use, GO enrichment analysis is a useful tool that can become even more powerful when combined with other types of analysis, such as the cluster analysis we performed before using clusterMaker2.

1. As a starting point, we will apply the BiNGO analysis to the whole dataset, in order to see an overview of all the processes overrepresented in this network. Subsequent analyses may then focus on sub-sets of the network, using a view suitable to pick out functional modules. Select all the nodes in the network.
2. To start BiNGO, go to 'Apps' → 'BiNGO'. Do this only once: Cytoscape will not stop you from opening multiple copies of the BiNGO setup menu (which will lead to confusion and chaos!).



3. The BiNGO setup screen will now appear (previous screenshot). There are several operations you need to perform in this screen:
 - a. Name the fraction of the network you are going to analyse in the text box 'Cluster name'.
 - b. We will take the standard significance level and statistical analysis options for this exercise. For a detailed comment on these options, you might want to have a look at the BiNGO User Guide that can be found in their website: www.psb.ugent.be/cbd/papers/BiNGO/User_Guide.html.

- c. We want to know which terms are over-represented in the network with respect to the whole annotation, so we leave the corresponding categories as they are.
 - d. Under 'Select ontology file' choose the Gene Ontology file 'gene_ontology_ext.obo' using the 'custom' option in the drop-down menu⁶.
 - e. Under Select namespace select 'Cellular Component'.
 - f. Under Select organism/annotation choose the 'gene_association.goa_human' file using the 'custom' option in the drop-down menu⁶.
 - g. The 'Discard the following evidence codes' box allows you to limit the analysis discarding annotations that are given based on a particular evidence code⁷.
 - h. If you want to save the results of the analysis, mark the check box and choose a path to save your files.
 - i. Finally, press the 'Start BiNGO' button.
- You will receive a warning saying, "Some category labels in the annotation file are not defined in the ontology". The warning refers to identifiers that are not properly mapped in the GO reference file by BiNGO. There might often be a small discrepancy between the identifiers provided in the interaction network and those found in the GO reference file (when using isoforms, for example). Ignore this warning and click OK.
 - The GO terms found are displayed in two ways. The first is a table of GO terms found, as seen in the next screenshot; the second is a directed acyclic network in which nodes are the GO terms found and directed edges link parent terms to child terms.

The screenshot shows the BiNGO output window with a table of GO terms and their associated genes. The table has columns for GO ID, Description, p-val, corr p-val, cluster freq, total freq, and genes. The first few rows are:

GO ID	Description	p-val	corr p-val	cluster freq	total freq	genes
44424	intracellular part	0.0000E-1	0.0000E-100	1150/1213 94.8%	39003/88970 43.8%	P30405 A6ND89 Q12824 Q9BRZ2 Q9C035 Q95793 Q8IZ07 Q70EL4 P84022 Q9BYC4 Q70EL2 Q68E01 Q92973...
5829	cytosol	0.0000E-1	0.0000E-100	524/1213 43.1%	3927/88970 4.4%	Q9UB84 Q9BRZ2 Q9C035 Q60566 P84022 Q00273 Q9BYC4 Q92973 P49137 P27348 Q92734 P52565 Q969K3...
5737	cytoplasm	3.1111E-310	0.0000E-100	1007/1213 83.0%	27715/88970 31.1%	P30405 A6ND89 Q9BRZ2 Q9C035 Q95793 Q8IZ07 P84022 Q9BYC4 Q92973 Q92731 Q92734 Q00839 Q9Y27...
5622	intracellular	4.9904E-298	0.0000E-100	1162/1213 95.7%	42729/88970 48.0%	P30405 A6ND89 Q12824 Q9BRZ2 Q9C035 Q95793 Q8IZ07 Q70EL4 P84022 Q9BYC4 Q70EL2 Q68E01 Q92973...
43226	organelle	8.0828E-283	0.0000E-100	1073/1213 88.4%	34977/88970 39.3%	P30405 Q12824 Q9C035 Q95793 Q8IZ07 Q70EL4 P84022 Q70EL2 Q68E01 Q92973 Q92731 Q92734 Q00839...
43227	membrane-bounded organelle	1.9016E-268	0.0000E-100	1013/1213 83.5%	31214/88970 35.0%	P30405 Q12824 Q9C035 Q95793 Q8IZ07 Q70EL4 P84022 Q70EL2 Q68E01 Q92973 Q92731 Q92734 Q00839...
43229	intracellular organelle	1.4185E-260	0.0000E-100	1036/1213 85.4%	33664/88970 37.8%	P30405 Q12824 Q9C035 Q95793 Q8IZ07 Q70EL4 P84022 Q70EL2 Q68E01 Q92973 Q92731 Q92734 Q00839...
5634	nucleus	2.1555E-258	0.0000E-100	725/1213 59.7%	14649/88970 16.4%	Q9H845 P67809 Q12948 Q12824 Q9C035 Q00267 Q60566 Q70EL4 P84022 Q00273 Q70EL2 Q68E01 Q9297...
31982	vesicle	1.2992E-244	0.0000E-100	422/1213 34.7%	4224/88970 4.7%	P67809 Q95793 Q92973 P49137 P27348 Q92734 Q9H2H9 P04004 P52565 Q9Y277 P08727 P27361 P16104 P...
44422	organelle part	2.9324E-239	0.0000E-100	829/1213 68.3%	21195/88970 23.8%	P30405 Q12824 Q95793 Q70EL4 P84022 Q70EL2 Q68E01 Q92973 Q92731 Q92734 Q00839 Q9Y277 P08727...
31974	membrane-enclosed lumen	3.1629E-237	0.0000E-100	510/1213 42.0%	7025/88970 7.8%	P30405 P67809 Q12948 Q12824 Q00267 Q60566 Q70EL4 P84022 Q00273 Q70EL2 Q68E01 Q92973 Q92731...
43233	organelle lumen	2.4834E-236	0.0000E-100	507/1213 41.7%	6957/88970 7.8%	P30405 P67809 Q12948 Q12824 Q00267 Q60566 Q70EL4 P84022 Q00273 Q70EL2 Q68E01 Q92973 Q92731...
44446	intracellular organelle part	5.0013E-235	0.0000E-100	818/1213 67.4%	20851/88970 23.4%	P30405 Q12824 Q95793 Q70EL4 P84022 Q70EL2 Q68E01 Q92973 Q92731 Q92734 Q00839 Q9Y277 P08727...
31988	membrane-bounded vesicle	5.5385E-232	0.0000E-100	404/1213 33.3%	4066/88970 4.5%	P67809 Q95793 Q92973 P49137 P27348 Q92734 Q9H2H9 P04004 P52565 Q9Y277 P08727 P27361 P16104 P...
1903	extracellular vesicle	2.7947E-231	0.0000E-100	350/1213 28.8%	2807/88970 3.1%	P67809 P13747 P30086 Q04760 P10599 Q95793 Q9UB83 Q92973 P46783 Q8WUM4 P49137 P62913 P27348...
43230	extracellular organelle	3.1607E-231	0.0000E-100	350/1213 28.8%	2808/88970 3.1%	P67809 P13747 P30086 Q04760 P10599 Q95793 Q9UB83 Q92973 P46783 Q8WUM4 P49137 P62913 P27348...

- The table displays the most over-represented terms sorted in with the smallest p-values on top. In this table we see a list of GO terms (with their names and GO-IDs) and the uncorrected p-value and corrected p-value. Apart from that, total frequency values and a list of corresponding proteins (listed under the title 'genes') are listed for each term. You can visualize which nodes have been significantly annotated under the listed terms by selecting the terms and then using the 'Select nodes' button. Since the list is sorted just by p-value, many general terms, (less descriptive terms) rise to the top of the table, making it difficult to see the more specific terms that are more useful. If you clicked the 'save' option in the BiNGO setup window, then this table is already saved to file. If not, then you will need to copy and paste these results into an Excel file (or similar). The data in

⁶ The Gene Ontology is updated continuously and the ontologies and annotations that are loaded by default in BiNGO are out of date. The files that we provide for this tutorial were freshly downloaded for this course from the following links. There are also specific instructions for downloading custom annotation files in the 'Additional information' section.

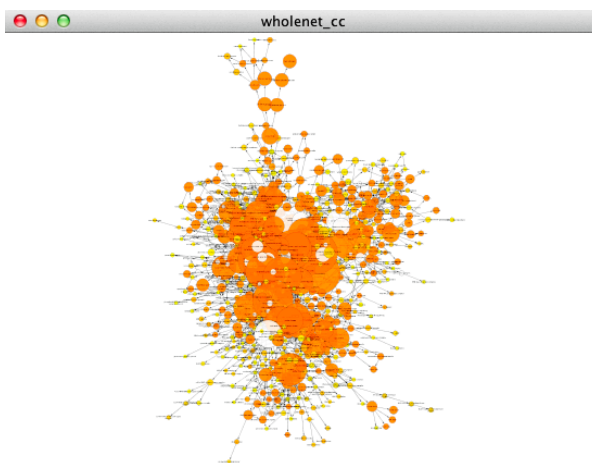
Ontology file (.obo extension): geneontology.org/page/download-ontology

Annotation file: geneontology.org/page/download-annotations

⁷ Every GO annotation is associated to a specific reference that describes the work or analysis supporting it. The evidence codes indicate how that annotation is supported by the reference. For example, annotations supported by the study of mutant varieties or knock-down experiments on specific genes are identified with the IMP (Inferred from Mutant Phenotype) code. All the annotations are assigned by curators with the exception of those with the IEA code (Inferred from Electronic Annotation), which are assigned automatically based in sequence similarity comparisons. See geneontology.org/page/guide-go-evidence-codes for more information about evidence codes.

this table is not saved as part of a Cytoscape session file and you will lose this data if you do not save it separately.

7. The other representation of the results is a graphical depiction of the enriched GO terms in the form of a network. Each node is a GO term, and GO terms are linked by directed edges representing parent-to-child relationships. Nodes are coloured by p-value (a small window depicting the legend is also produced) and the size of each node is proportional to the number of proteins annotated with that term. The default layout is less easy to read if the list of terms is long (as it is in our case), but we may take advantage of one of Cytoscape's tools to provide an alternative representation.

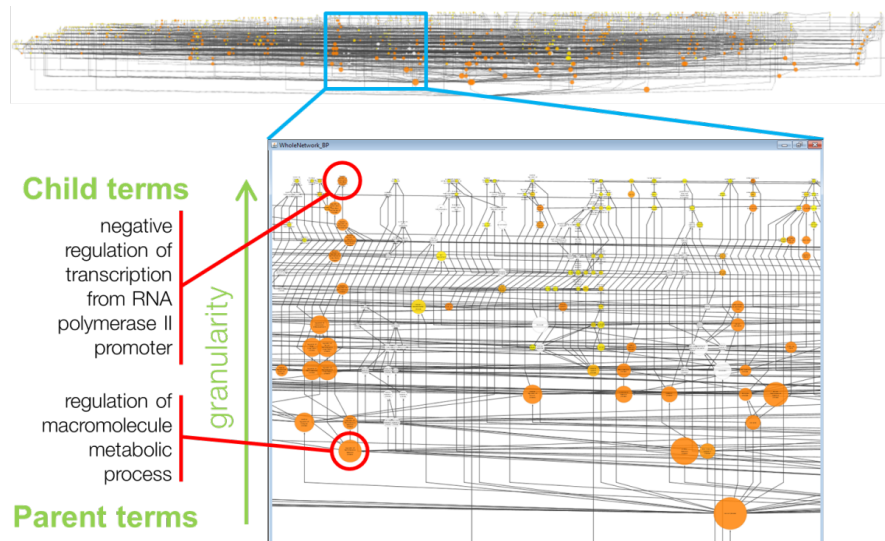


8. Make sure the graphical representation of the BiNGO results is selected. Choose 'Layouts' → 'Cytoscape Layouts' → 'Hierarchical layout'. Gene ontologies are a directed acyclic graph: Cytoscape utilizes this topology to organize the BiNGO results graph so that more specific and informative terms float to the top, while general, less informative terms sink to the bottom. You want to focus on orange-coloured terms that branch-up the graph to find significantly enriched functions, as shown in the figure in the next page. Navigating through this view provides a more useful impression of what biological processes are present in this network. When you find a term of interest, you may look it up in the table to see what proteins in the network were annotated with that term.

Save your session



The graphical representation of your BiNGO results is just another network that can be modified and analysed in Cytoscape by making further use of analysis plugins.



A final set of exercises

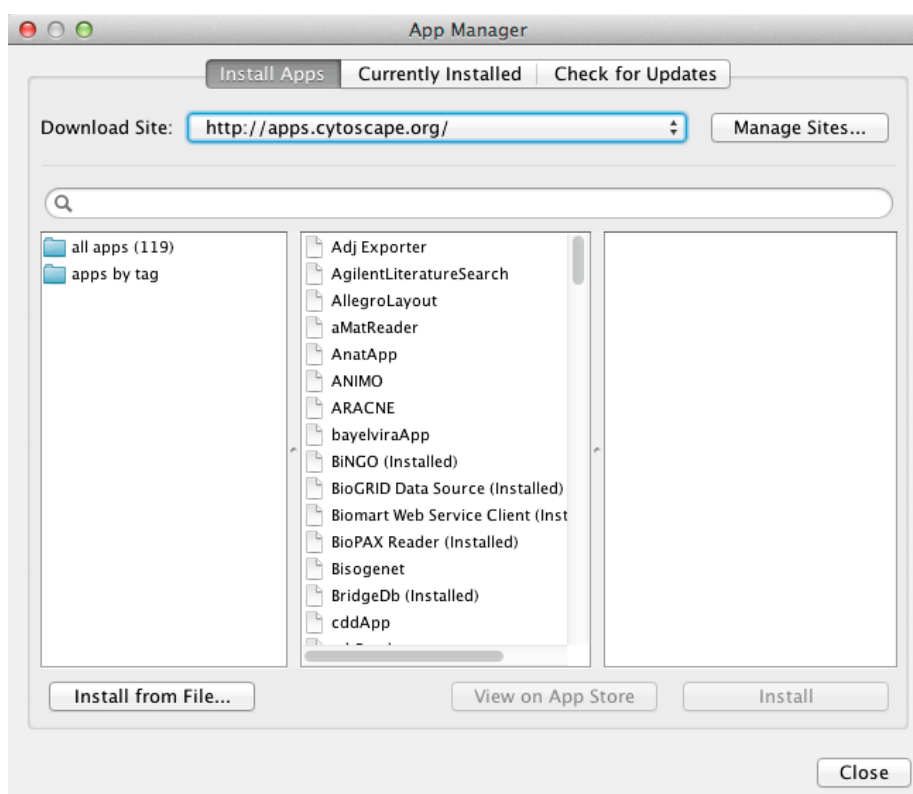
- The output of the analysis as we have performed it is terribly complicated. Try to repeat it using the generic GOSlim ontology and annotation that are provided in the files 'goslim_generic.obo' and 'gene_association_human_goslim.goa'. How does the analysis look like now?
- Which processes are specifically over-represented in those proteins over-represented proteins in cells exposed to 30 µg/mL of MWCNTs in the 12 hours time point?
 - Repeat the BiNGO analysis and find out which processes are involving proteins connected to over-represented proteins in cells exposed to 30 µg/mL of MWCNTs in the 12 hours time point.
 - 💡 What is the difference between the results of these two analysis?
- Combine the ontology enrichment and the cluster analysis to try to figure out the functionality behind the different topological clusters found by clusterMaker2. Check for clusters where over/under-represented proteins are especially prominent to find functional modules potentially different cells exposed to different doses of MWCNTs at different time points.

Additional information

Installing apps in Cytoscape 3.3

This set of instructions is specific for the BiNGO app as an example, but it can be used for any other plugin you might need to install using the plugins manager in Cytoscape 3.3, such as clusterMaker2.

1. In Cytoscape, go to 'Apps' → 'App Manager' (see next screenshot).
2. Look for BiNGO using the search box or browsing through the 'apps by tag' folder and the 'enrichment analysis' subfolder.
3. Press 'Install'
4. Check that the app was installed; it should be visible in your 'Apps' menu. You might need to re-start Cytoscape if it is not there.



You can also install the app if you have Cytoscape opened and just go to the apps store in the Cytoscape site (<http://apps.cytoscape.org/>). Look for the app you need and you will find a Cytoscape 3 'Install' button on the right hand-side of the screen. You can use this and the app will be immediately installed in Cytoscape.

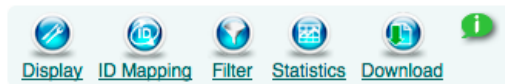
Obtaining a GOSlim and species-specific annotation files

The Gene Ontology consortium makes available their ontology and generic annotation files at <http://geneontology.org/>. While it is not problematic to obtain the ontology files for the full GO along with a set of more or less generic GOSlims in their website, the annotation files can be a bit more problematic if you want something specific. A generic GO slim annotation file is not available there, for example, and species-specific annotations can be problematic if they are not mapped to UniProt accessions (as happens in MGI-produced annotation files for mouse).

In order to help you solve this issue, here is a short set of instructions on how to use the QuickGO ontology browser (www.ebi.ac.uk/QuickGO/) to produce a) species-specific annotation files and b) GO slim annotation files.

a) Generate a species-specific annotation file with QuickGO

1. Open QuickGO in www.ebi.ac.uk/QuickGO/ and click the 'Search and Filter GO annotation sets'.
2. You will get a big table with all the GO annotation available. Above the table you get the total number of annotations available in GO (over 200 million annotations to over 30 million proteins at the moment of writing this tutorial). Notice the menu on the upper-right corner and click on the 'Filter' icon (see screenshot).



3. This opens a window with several tabs. Each tab refers to a different level for filtering. We are now interested in the 'Taxon' tab, so click on it.
4. Select the species that you need if it is on the pre-defined list or enter the taxon ID in the box. You can select more than one taxon if you want. Select or enter "9606" to download all human annotations. Then click 'Submit'.
5. Now you are back on the table view and notice the 'Displaying annotations...' text referring to a lower number of annotations (about one million annotations to 100000 human proteins).
6. To download the annotations, just click on the 'Download' button on the upper-right menu. Select the GAF format if you want to use the file in BiNGO. Be sure to set a high limit, so all the annotations are downloaded.

b) Obtain a GO slim annotation file

1. Open QuickGO in www.ebi.ac.uk/QuickGO/ and click the 'Investigate GO slims'.
2. You get to a window with several tabs. The first tab is a set of instructions that can help you during the process. Click on the 'Choose Terms' tab.
3. You get a menu where you can select a pre-defined GO slim or add a list of GO identifiers representing a slim you have created externally. For this example, we will select the pre-defined GO slim. Click on the green cross by the 'goslim_generic' box.
4. You automatically get to the 'Refine Selection' tab. Here you could remove specific terms or branches if you wish. We can ignore this process for this example, along with the 'Comparison Chart' tab. Just click on 'Find annotations' to get your results.
5. Once you get the results, you can apply further filters to, for example, reduce the number of annotations and limit them to one species, as we proposed in the previous section. Just click download as before to save your GO slim.

Further reading

Apart from the references given throughout the text, here you have a couple of suggestions that I hope you might find useful:

Nice review on PPI network generation and their use to study genetic disease. Provides a very nice overview of how consensus and paradigms within the field have evolved and how our confidence on PPI data and its coverage has evolved over the years: Lage, 2014 [20].

General review on the utility of molecular interactions to study disease, with special emphasis on the strengths and limitations of the field, quite useful for the newcomer: Schramm *et al.*, 2013 [21].

Entry-point review about the basic concepts required to understand protein-protein interactions: De Las Rivas *et al.*, 2010 [22].

Excellent recent overview on the strategies and tools used for network and pathway analysis, with a focus on cancer: Creixell *et al.*, 2015 [23].

More on human diseases and network biology in this review in which the author provides a clear explanation of how topological characteristics of the networks can be used to learn new things about disease pathogenesis: Furlong, 2013 [24].

A review about differential network biology, the study of the differences between particular biological contexts in contrast with the static interactome: Ideker & Krogan, 2012 [25].

The assessment of confidence values to molecular interactions requires the use of several, complementary approaches. In this study, the performance of different protein interaction detection methods with respect to a golden standard set is evaluated: Braun *et al.*, 2008 [26].

Our group produced a tutorial in the HUPO discussing the importance of molecular interactions network analysis and applying a similar approach to the one presented here, using BiNGO in combination with clusterMaker. It is a bit old now, but the basics still apply. See Koh *et al.*, 2012 [27].

A good example of network analysis using data coming from literature-curated databases can be found in this paper in Nature Biotechnology: Wang *et al.*, 2012 [28]. They construct a network with high-quality binary protein-protein interactions where there is information about the interaction interfaces at atomic resolution and integrate disease-related mutation information, finding out an enrichment of disease-causing mutations in interacting interfaces.

A very nice network analysis paper in which the authors outline the full power of integrating different sorts of data to analyse the immensely complex human interactome and derive context-filtered networks that can help to drive experimental research: Schaefer *et al.*, 2012 [29].

This very recent large-scale resource publication providing over 14000 interactions between human proteins by use of the yeast two-hybrid method is particularly interesting for its detailed analysis on issues such as popularity bias in PPI database information and description of general properties of the human interactome. It also has a detailed methods section in which the process of constructing a network from multiple databases is dealt with extensively. Check Rolland *et al.*, 2014 [30].

If you want to learn more about the interaction confidence score used in IntAct, MINT, MatrixDB and other IMEx-complying databases, check this publication in Database [31]. It describes the algorithms used to deal with redundant interactions (MImerge) and to score these interactions using the experimental evidence given for each interacting pair (MIscore).

Finally, a visualization-based review highlighting the possibilities that representing deep-curated interaction datasets can offer, focused in LRRK2, a kinase linked to familial forms of Parkinson's disease [32].

Links to useful resources

First, some useful repositories, databases and ontologies:

- The Universal Protein Resource, UniProt : www.uniprot.org
- The Gene Ontology: geneontology.org
- QuickGO, a GO browser: www.ebi.ac.uk/QuickGO
- The IntAct molecular interactions database: www.ebi.ac.uk/intact
- Lots of other IMEx-complying interaction databases in the IMEx website: www.imexconsortium.org/about-imex

And some useful tools:

- How do I get interaction data from most of the interaction databases that are out there? Easy answer: use the Proteomics Standard Initiative Common Query Interface (PSICQUIC). You can learn more about it here github.com/micommunity/psicquic and here you have a link to its search interface, PSICQUIC View: www.ebi.ac.uk/Tools/webservices/psicquic/view
- To learn more about Cytoscape or to get access to documentation and tutorials, go to its website: www.cytoscape.org. You can see a list of version 2.8.3 plugins here: chianti.ucsd.edu/cyto_web/plugins. For plugins (apps) in the 3.x series, visit their app store: apps.cytoscape.org. Last, but not least, an introductory article about Cytoscape plugins for newcomers: Saito *et al.*, 2012 [33].
- More about the BiNGO plugin in their website, with a nice tutorial and useful documentation: www.psb.ugent.be/cbd/papers/BiNGO
- A more advanced enrichment analysis tool, with lots of useful features, especially for different cluster comparison. Check their website www.ici.upmc.fr/cluego.
- To find functional circuits in large networks, try clusterMaker2, a Cytoscape plugin for topological cluster analysis. Lots of documentation and useful tutorials in their website: www.rbvi.ucsf.edu/cytoscape/clusterMaker2
- A very clear and useful video explaining in detail how the MCODE algorithm works: <http://www.youtube.com/watch?v=7wA4ZEoFGI8>
- A repository to share networks produced in Cytoscape as webpages that allow interactive functionalities such as zooming, moving nodes and edges, etc... Here is a link to the tool and some basic documentation: <http://idekerlab.github.io/cy-net-share/>. Some more detailed instructions and other possibilities for publishing Cytoscape-produced networks can be found here: http://wiki.cytoscape.org/Cytoscape_3/UserManual/Publish

Contact details

Don't hesitate to write if you have any questions, comments or random thoughts.

Pablo Porras Millán, PhD
EMBL-EBI
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SD, U.K.
Tel: +44 1223 494482
email: pporras@ebi.ac.uk

References

- [1] Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T., Cytoscape 2.8: new features for data integration and network visualization. *Bioinforma. Oxf. Engl.* 2011, 27, 431–432.
- [2] Ju, L., Zhang, G., Zhang, X., Jia, Z., et al., Proteomic analysis of cellular response induced by multi-walled carbon nanotubes exposure in a549 cells. *PLoS One* 2014, 9, e84974.
- [3] Orchard, S., Kerrien, S., Abbani, S., Aranda, B., et al., Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* 2012, 9, 345–350.
- [4] Orchard, S., Ammari, M., Aranda, B., Breuza, L., et al., The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 2014, 42, D358–63.
- [5] Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., et al., MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* 2010, 38, D532–539.
- [6] Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N., Ricard-Blum, S., MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* 2011, 39, D235–240.
- [7] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., et al., The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 2004, 32, D449–451.
- [8] Brown, K.R., Jurisica, I., Online predicted human interaction database. *Bioinforma. Oxf. Engl.* 2005, 21, 2076–82.
- [9] Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S.L., et al., PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 2011, 8, 528–529.
- [10] Croft, D., O’Kelly, G., Wu, G., Haw, R., et al., Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 2011, 39, D691–697.
- [11] Brohée, S., van Helden, J., Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006, 7, 488.
- [12] Wang, J., Li, M., Deng, Y., Pan, Y., Recent advances in clustering methods for protein interaction networks. *BMC Genomics* 2010, 11 Suppl 3, S10.
- [13] Su, G., Kuchinsky, A., Morris, J.H., States, D.J., Meng, F., GLay: community structure analysis of biological networks. *Bioinforma. Oxf. Engl.* 2010, 26, 3135–7.
- [14] Newman, M.E.J., Girvan, M., Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 2004, 69, 026113.
- [15] Bader, G.D., Hogue, C.W. V., An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, 4, 2.
- [16] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25, 25–29.
- [17] Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., et al., ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 2009, 25, 1091–1093.
- [18] Maere, S., Heymans, K., Kuiper, M., BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks. *Bioinformatics* 2005, 21, 3448–3449.
- [19] Smoot, M., Ono, K., Ideker, T., Maere, S., PiNGO: a Cytoscape plugin to find candidate genes in biological networks. *Bioinforma. Oxf. Engl.* 2011, 27, 1030–1031.
- [20] Lage, K., Protein–protein interactions and genetic diseases: The interactome. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* n.d.
- [21] Schramm, S.-J., Jayaswal, V., Goel, A., Li, S.S., et al., Molecular interaction networks for the analysis of human disease: utility, limitations, and considerations. *Proteomics* 2013, 13, 3393–405.
- [22] De Las Rivas, J., Fontanillo, C., Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. *PLoS Comput Biol* 2010, 6, e1000807.
- [23] the Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium, Pathway and network analysis of cancer genomes. *Nat. Methods* 2015, 12, 615–621.
- [24] Furlong, L.I., Human diseases through the lens of network biology. *Trends Genet. TIG* 2013, 29, 150–159.
- [25] Ideker, T., Krogan, N.J., Differential network biology. *Mol. Syst. Biol.* 2012, 8, 565.
- [26] Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., et al., An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* 2008, 6, 91–97.
- [27] Koh, G.C.K.W., Porras, P., Aranda, B., Hermjakob, H., Orchard, S.E., Analyzing Protein-Protein Interaction Networks (†). *J. Proteome Res.* 2012.
- [28] Wang, X., Wei, X., Thijssen, B., Das, J., et al., Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 2012, 30, 159–164.
- [29] Schaefer, M.H., Lopes, T.J.S., Mah, N., Shoemaker, J.E., et al., Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.* 2013, 9, e1002860.
- [30] Rolland, T., Tasan, M., Charlotiaux, B., Pevzner, S.J., et al., A Proteome-Scale Map of the Human Interactome Network. *Cell* 2014, 159, 1212–1226.
- [31] Villaveces, J.M., Jiménez, R.C., Porras, P., Del-Toro, N., et al., Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database J. Biol. Databases Curation* 2015, 2015.
- [32] Porras, P., Duesbury, M., Fabregat, A., Ueffing, M., et al., A visual review of the interactome of

- LRRK2: Using deep-curated molecular interactions data to represent biology. *Proteomics* 2015.
- [33] Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., et al., A travel guide to Cytoscape plugins. *Nat. Methods* 2012, 9, 1069–1076.