CHAT User Manual

Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks

Tanja Muetze[1*], Ivan H. Goenawan[1*], Heather L. Wiencko[2], Manuel Bernal-Llinares[1], Kenneth Bryan [1] and David J. Lynn [1,3,†]
[1]EMBL Australia Biomedical Informatics Group, Infection & Immunity Theme, South Australian Medical and Health Research Institute, North Terrace, Adelaide, SA 5000, Australia.
[2]Animal and Bioscience Research Department, Animal and Grassland Research and Innovation Centre, Teagasc, Grange, Co. Meath, Ireland.
[3]School of Medicine, Flinders University, Bedford Park, SA 5042, Australia.

* These authors contributed equally
† To whom correspondence should be addressed.
david.lynn@sahmri.com

About CHAT

Highly connected nodes (hubs) in biological networks are topologically important to the structure of the network and, have been shown to be preferentially associated with a range of phenotypes of interest. The relative importance of a hub node, however, can change depending on the biological context. Here, we report a Cytoscape[1] App, the Contextual Hub Analysis Tool (CHAT), which enables users to easily construct and visualize a network of interactions from a list of genes of interest, integrate contextual information, such as gene expression data, and identify nodes that are more highly connected to contextual nodes, e.g. those that are differentially expressed, than expected by chance. In a case study, we show that such *contextual hubs* are more relevant to the biology of the dataset being investigated than degree-based hubs.

Availability

CHAT can be downloaded from the Cytoscape App Store (http://apps.cytoscape.org/apps/chat). The latest source code is available at https://bitbucket.org/dynetteam/chat.

System/software requirements (sequence is important)

- Java 8 (http://www.oracle.com/technetwork/java/javase/downloads/index.html)
- Cytoscape 3.0 or higher (http://www.cytoscape.org/download.php)
- Installation: Open Cytoscape, click on "Apps" in the menu bar and select "App Manager". In the App Manager, on the "Install Apps" tab, type "CHAT" in the search bar, click on CHAT when it appears, and then click on "Install".

List of features

- Network construction.

    - Create a network from a user-supplied list of genes based on species-specific interaction data from a database that supports the PSICQUIC[2] interface, such as IntAct or BioGrid.
- Analysis
    - Unlike many existing tools, which simply identify whether a node is highly connected in a network, CHAT identifies hub nodes that interact with more "contextual" (e.g. differentially expressed) nodes than statistically expected.
    - CHAT allows users to compare the top contextual hubs to the top degree-based hubs.
    - Highlight contextual nodes vs non-contextual nodes.
    - Adjust the network to only show the top *n* hubs (either contextual or degree-based) and their interactors.
    - Customize colors.
    - Analyze individual nodes and their interactors.

Overview

The process of identifying top contextual hubs consists of three main steps:

1) input of a user-supplied gene list and contextual data,

2) network construction and statistical analysis to identify nodes that preferentially interact with contextual nodes,

3) and visualization of the top contextual hubs and their interactions, compared to the top degree-based hubs.

The Tool

1. Initialization dialog
    1.1. Network settings
        1.1.1. Input data (Identifier and attributes list)
        1.1.2. Select PSICQUIC database to query
        1.1.3. Restrict query to a specific interaction type or to specific interaction types
        1.1.4. Specify the species for the query
    1.2. Network name and layout
2. Results panel
    2.1. User-supplied settings
    2.2. Compare contextual with degree-based hubs
    2.3. Visualize the top n hubs
3. Node analyzer

1. Initial Dialog

After successfully installing CHAT via the App Manager, CHAT can be launched by clicking on the “Apps” menu in Cytoscape, and then choosing “CHAT” in the dropdown menu. This launches the initial dialog where the user defines the parameters to create the network (Figure 1).

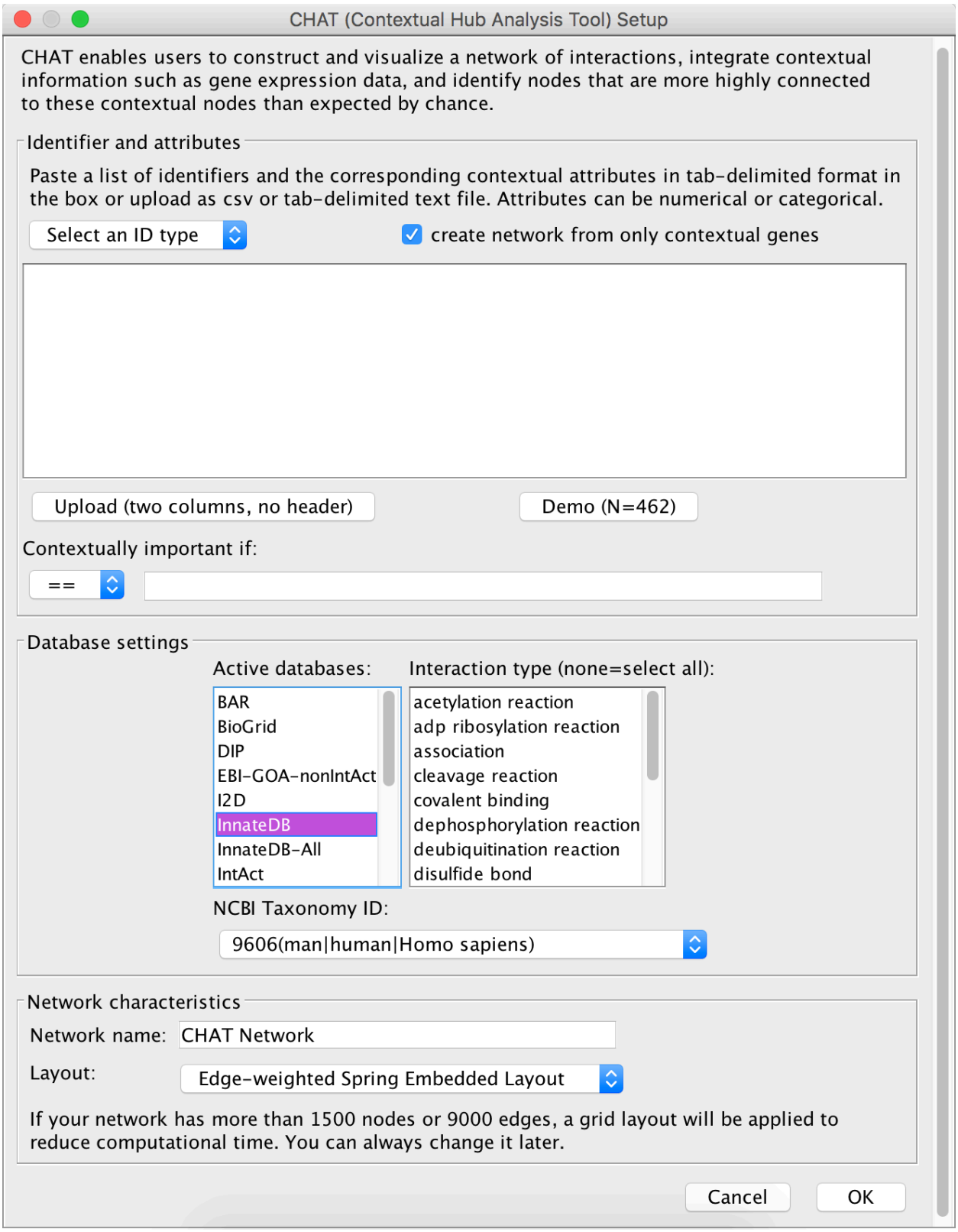

Figure 1. Initial Dialog.

1.1. Network settings

1.1.1. Input Data (Identifier and attributes list)

To build a network using CHAT, the user must provide a list of gene identifiers, and associated numerical or categorical attributes, in tab-delimited format. This data can be entered into the text box provided, e.g. by copy and pasting from Excel, or it can be uploaded as a csv or tab-delimited file via the upload button.

The first column in the uploaded data must contain the gene identifiers, and the second column the attributes. The attributes can be either numerical (p-values, fold changes, or any other quantitative data) or categorical.

The uploaded or pasted data must not have a header or additional columns.

Incorrectly formatted files or input will result in errors and warning messages.

If the user has not already pre-processed the input data, they can then specify which genes in the uploaded list should be considered as contextually important, based on the user-provided numerical or categorical contextual attributes of those genes (e.g. genes with > 2 fold-change in expression). If the user provides numerical attributes, they can choose among the six options (==,!=,<,<=,> and >=). If the attributes are categorical, the user chooses among “equal to” (==) and a “not equal to” (!=). The user can decide if they want the network to include all uploaded genes (including the non-contextual ones) or only genes that meet the threshold criteria by ticking a check box.

Next, from the dropdown menu, select a gene identifier type to indicate which ID types have been provided to CHAT, e.g. UniProt or Ensembl. The IDs which are supported will depend on the selected PSICQUIC[2] database.

1.1.2. Select PSICQUIC database to query

Choose one of the available PSICQUIC databases which will be used as the source of molecular interaction data. CHAT provides access to any PSICQUIC service that has at least 10,000 interactions and interactions between genes of the same species (unlike ChEMBL). Note that at a given time the number of available databases might vary if a database is temporarily down on the registry at the time of query. The selected database will be used to identify interactors of the uploaded or contextual genes dependent on the user’s choice and will be used as the background universe for computing the hypergeometric tests. Hovering over a database will display a tooltip that indicates the number of genes for the selected species, ID type and interaction type, and the provider’s (database) URL, where the user can find out more information (Figure 2).

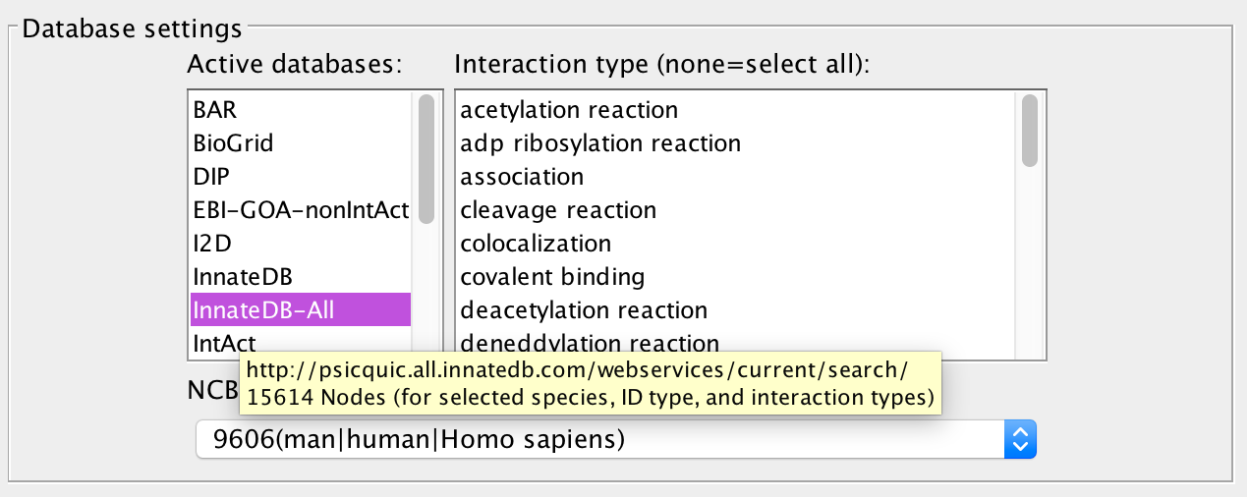

Figure 2. Tooltip for the URL of the selected database as well as the number of nodes matching the currently selected species, ID type and interaction type(s).

Not all ID types are covered by all the molecular interaction databases, thus, the user available ID type choices will change once a particular PSICQUIC provider has been selected.

1.1.3. Restrict query to a specific interaction type or to specific interaction types

By default, all interaction types are used to generate the network, however, the user can restrict the query to a specific interaction type by selecting it in the "Interaction type" menu. Multiple selection is allowed by holding down the Ctrl (Windows) or Command (OS X) key while clicking on the desired interaction types, or by using the shift key to highlight a section of interaction types. The background universe for the hypergeometric tests will be adjusted accordingly.

1.1.4. Specify the species for the query

The default species for queries is the species with the most number of interactors in the selected database based on the chosen ID type, however, the user can choose to query data from any species that the chosen database provides.

1.2. Network name and layout

For clarity reasons, and to ensure unique Cytoscape windows and networks (e.g. for computational access to the network), CHAT suggests a unique network name. The user is, however, free to rename the network. The name they enter will be displayed as the network name in the control panel, the window name and the name of the tab in the result panel.

Any Cytoscape network layout algorithm can be used, however, if the number of nodes exceeds 1500, or the number of edges exceeds 9000, a grid layout is automatically applied to reduce computational time, but it can still be changed afterwards.

2. Results panel

Once the user hits “OK”, CHAT retrieves the network data via the selected PSICQUIC service, creates the network, applies the network layout algorithm and adjusts the node sizes and colors based on their p-value. The network consists of the contextual or uploaded genes and their interactors. Self-interactions are omitted. Edges are only created between all uploaded or contextual genes (as chosen by the user) and their interactors, but not between interactors, to keep the visualization simple and not overloaded as Cytoscape slows down the larger the network becomes. Once the network creation completes, a CHAT tab is created on the result panel, which is the starting point for further analysis (Figure 3).

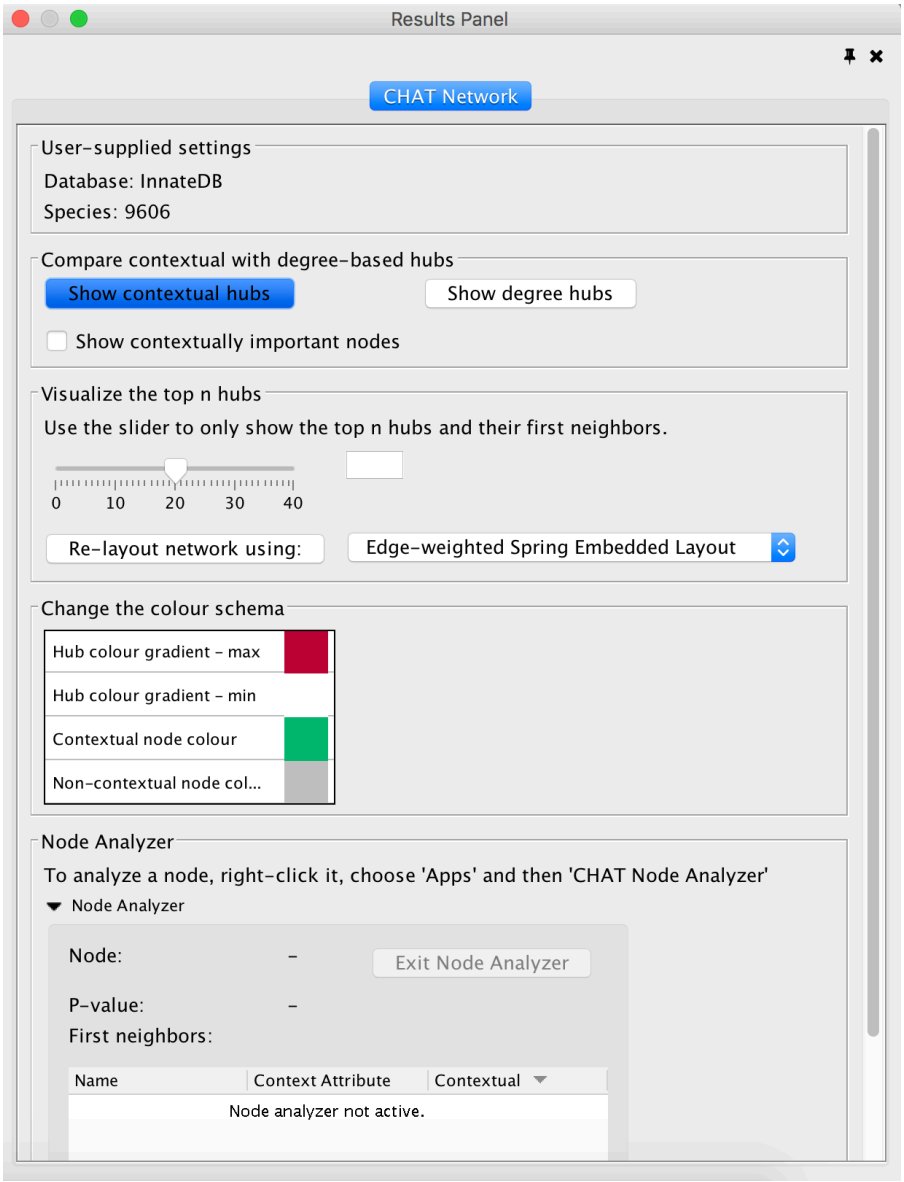

Figure 3. Results panel.

2.1. User-supplied settings

The top part of the results panel displays the selected database, species, ID type and interaction type(s) chosen by the user to create the network.

2.2. Compare contextual with degree-based hubs

By default, CHAT displays the top 20 contextual hubs and their direct interactors based on the results of the hypergeometric distribution test, which determines whether each node in the network interacts with more contextual nodes than expected by chance. In the CHAT network visualization, node size and node color are proportional to the node's p-value in this test, such that the smaller the p-value (i.e. more statistically significant), the larger the node size and the darker the red coloring of the node. The color scheme is adjustable via the "Change the color schema" menu. Selecting the "show contextually important nodes" box highlights the nodes in the network that were specified by the user as being contextually important. By default, these are colored green when this box is checked. Non-contextual nodes are shown in light-grey.

The user can quickly and easily compare the CHAT results to the results they would have obtained based on node degree (connectivity) by clicking the "show degree hubs" button (see Figure 4). This will change the network visualization such that the node size and node color are now proportional to the node's degree.

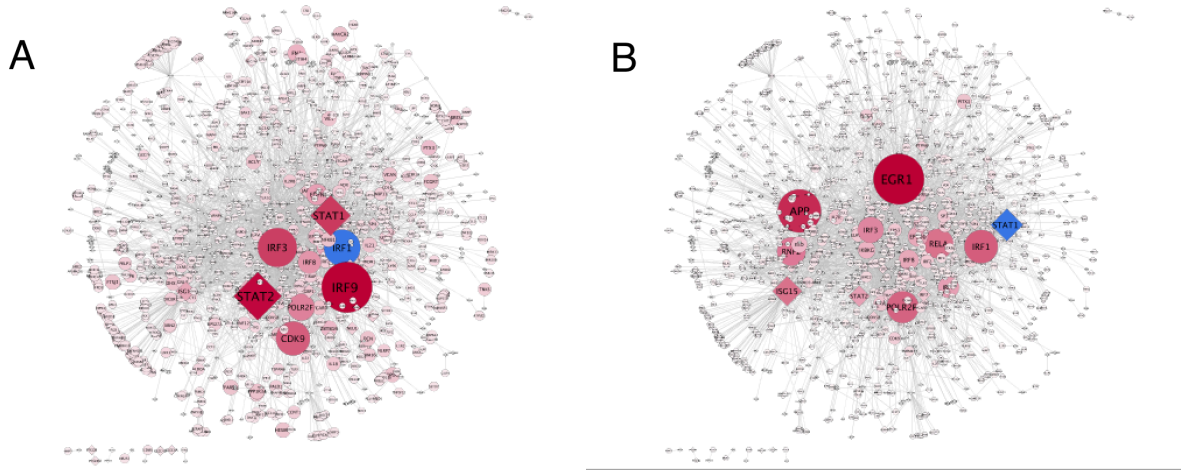

Figure 4. Comparison between context-specific hubs (A) and high degree hubs (B) in the Dengue fever gene expression dataset (see case study below).

2.3. Visualize the top n hubs

Given that protein-protein interaction networks can be large, by default, CHAT displays the top 20 most significant nodes (which is determined by the p-value from the hypergeometric test or the number of their first neighbors, depending on the option selected) and their interactors. This default can be adjusted by sliding the slider, or by entering a positive number into the text box next to the slider, and hitting enter. A warning will be displayed if the value entered is not valid.

3. Node analyzer

If the user right clicks on a node and then chooses “Apps” and then “CHAT Node Analyzer”, all nodes except the selected node and its interactors (first neighbors) are hidden. The table at the bottom of the results panel will also display attributes for the selected node and its interactors: the selected node’s name, p-value and information about its interactors, their names, their input contextual attributes and whether they are considered contextual (Figure 5 and Figure 6). This allows the user to closely analyze the selected node. The user can further click on another node and launch the node analyzer for that node. To return to a full view of the network, click on “Exit Node Analyzer”.

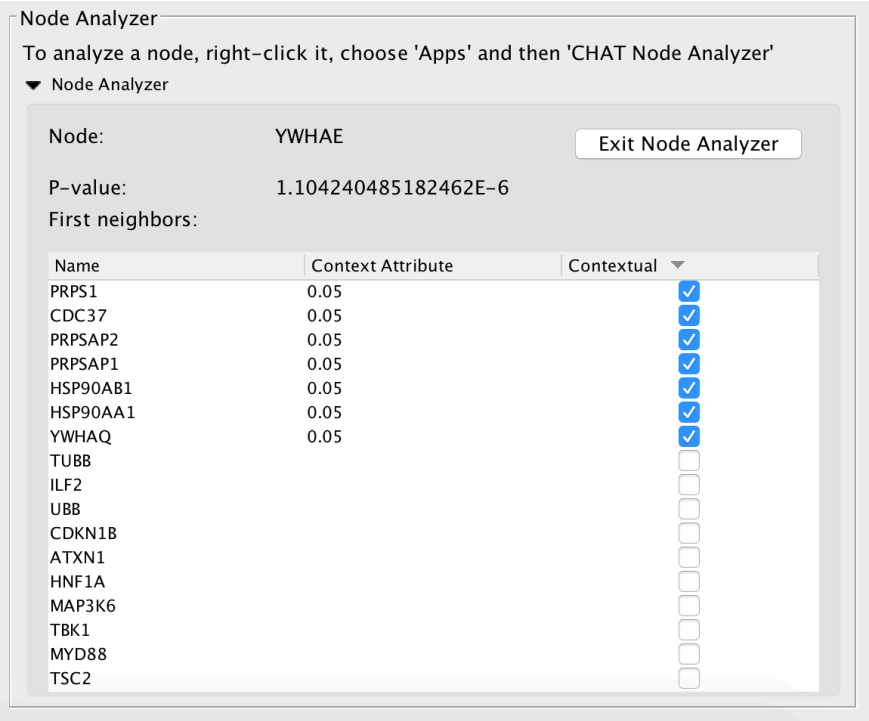

Figure 5. Node Analyzer table.

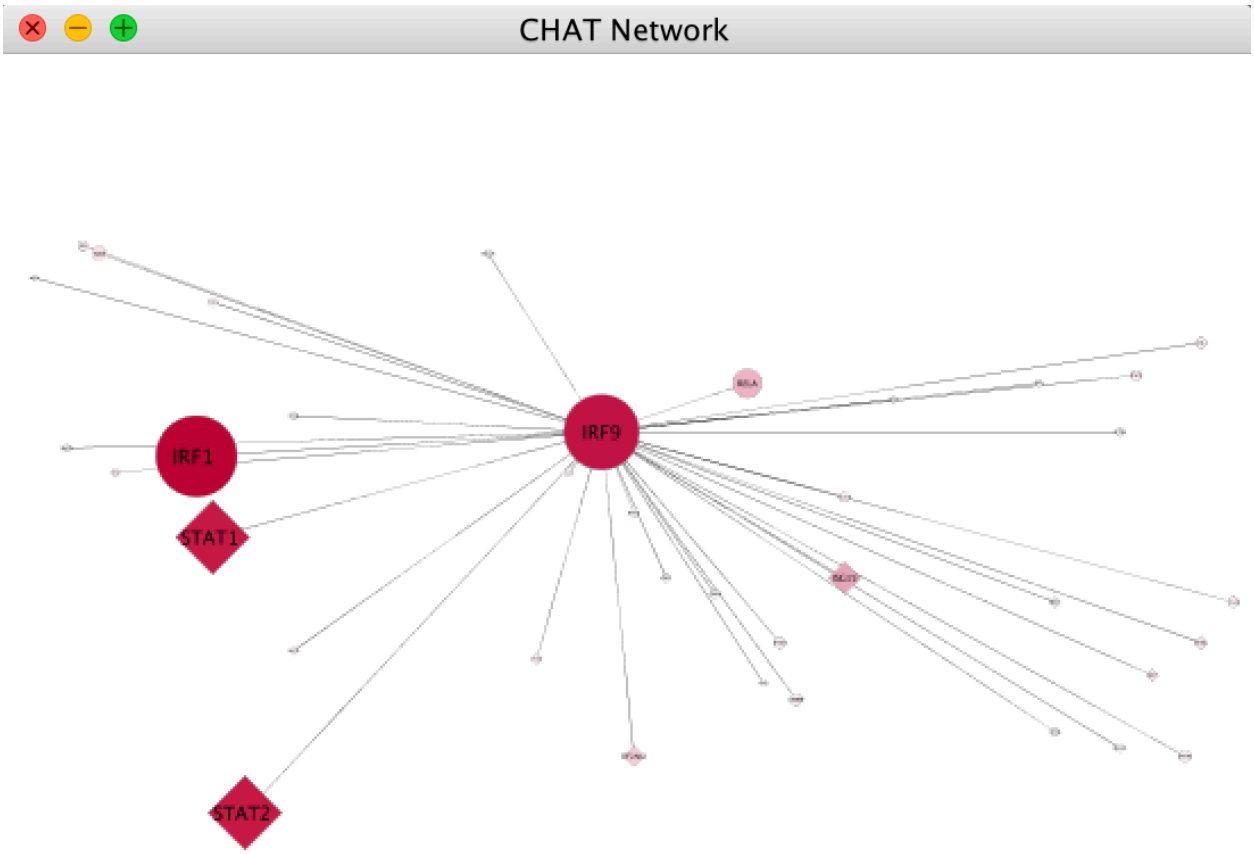

Figure 6. Node Analyzer view.

The network window

Hovering over nodes and edges reveals their unique ID in the database (Figure 7). You can further perform other modifications to analyze the network, such as applying different layouts or styles or running other apps.

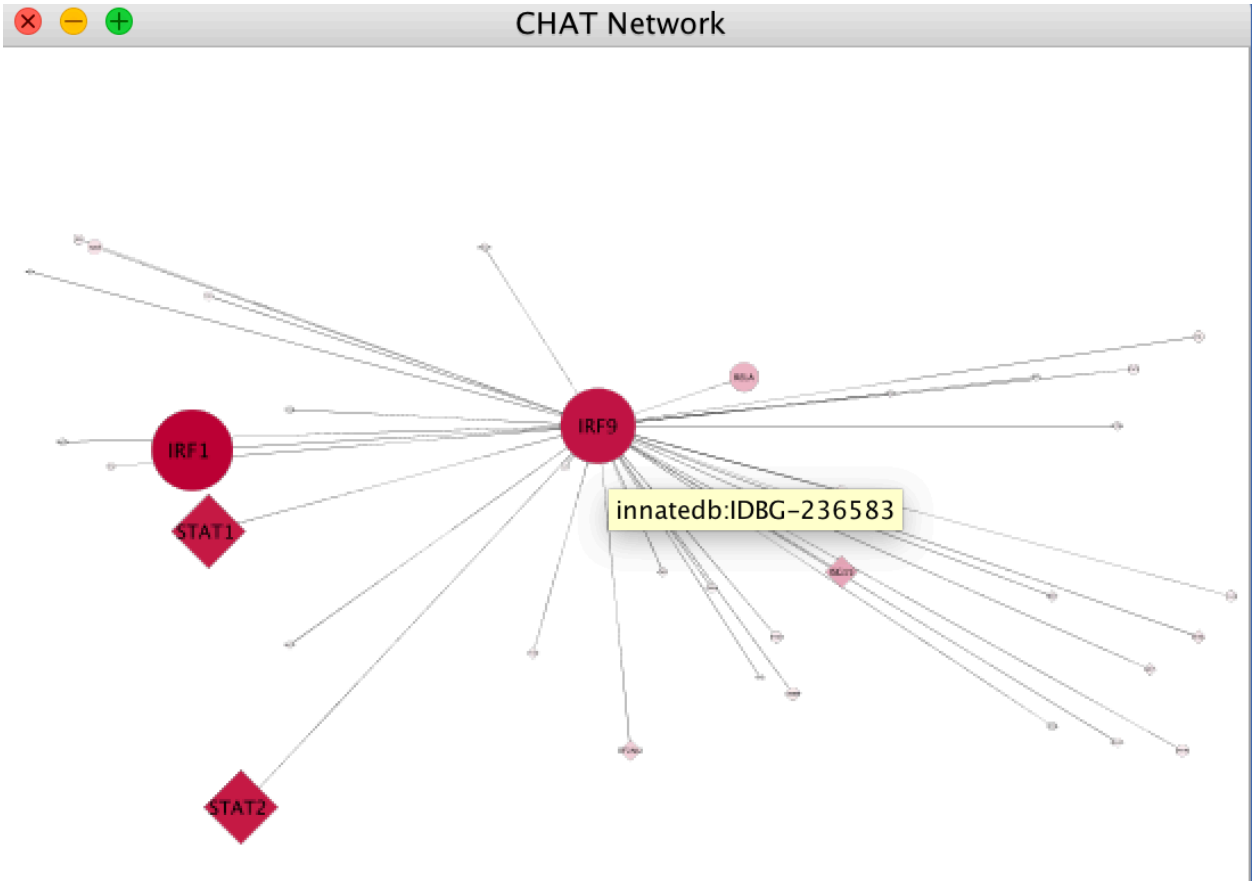

Figure 7. Node and edge tooltips.

Mathematical foundations: The hypergeometric test

In order to calculate the contextual importance of a node or a hub, CHAT performs a hypergeometric test on each node in the network. This test assesses the cumulative probability that a node interacts with contextual nodes more than expected by chance. The p-value is calculated using the equation:

$$p(X \geq k) = \sum_{x=k}^{n} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

Where *N* is the number of genes with at least one non-self-interaction in the database queried in which both interactors have the specified species and ID type, and *K* is the number of contextually relevant nodes provided by the user (with at least one interaction in the database queried), *n* is the number of nodes connected to the hub, and *k* is the number of contextually relevant nodes connected to the hub.

Case study: Dengue fever gene expression dataset

Start CHAT via the "Apps" menu. On the initial dialog, select Ensembl as ID type, leave the check box checked, upload the file Dengue_Ensembl_up_reg_IDs_ready_for_upload_to_CHAT.csv, set

the comparator to “>=” and the attribute value field to 2.0 (Figure 8). Select InnatedDB-All as database and do not click on an interaction type. Change the network name to “Dengue fever gene expression set – CHAT analysis”, leave the default layout, the edge-weighted spring embedded layout, and hit ok. Once the network is created, we can toggle between viewing the nodes in the network sized and colored by the p-value from the hypergeometric tests and the node degree. The network structure is very different for the contextual hubs and high degree hubs (Figure 4). Only 4 hubs of the top 20 hubs are the same among these two networks (Figure 4). By default, only the top 20 hubs are shown but we can adjust this via the slider. We can zoom into specific nodes using the Node Analyzer (Figure 6 and Figure 9).

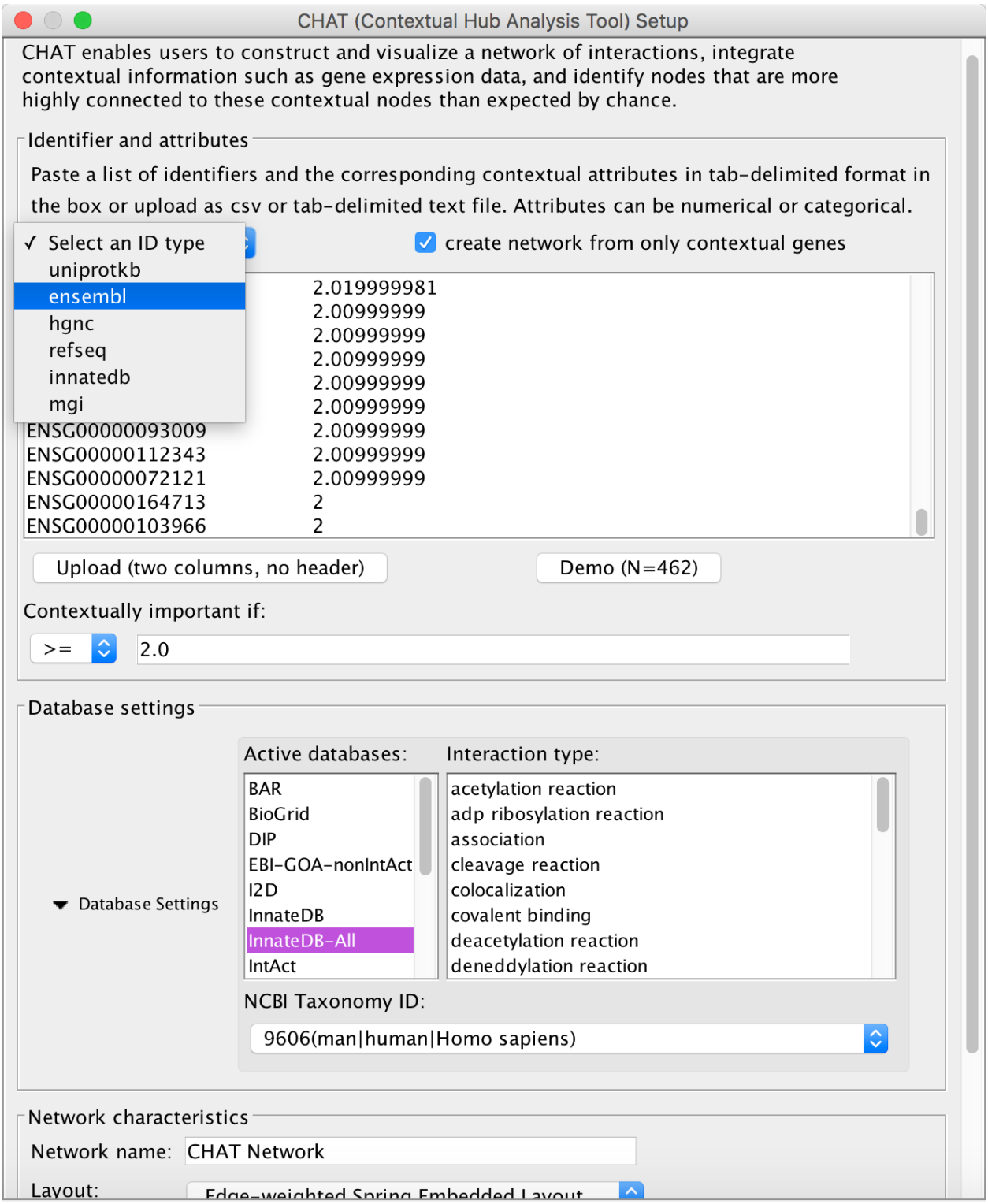

Figure 8. Initial dialog filled with data for the analysis of the example Dengue fever gene expression dataset

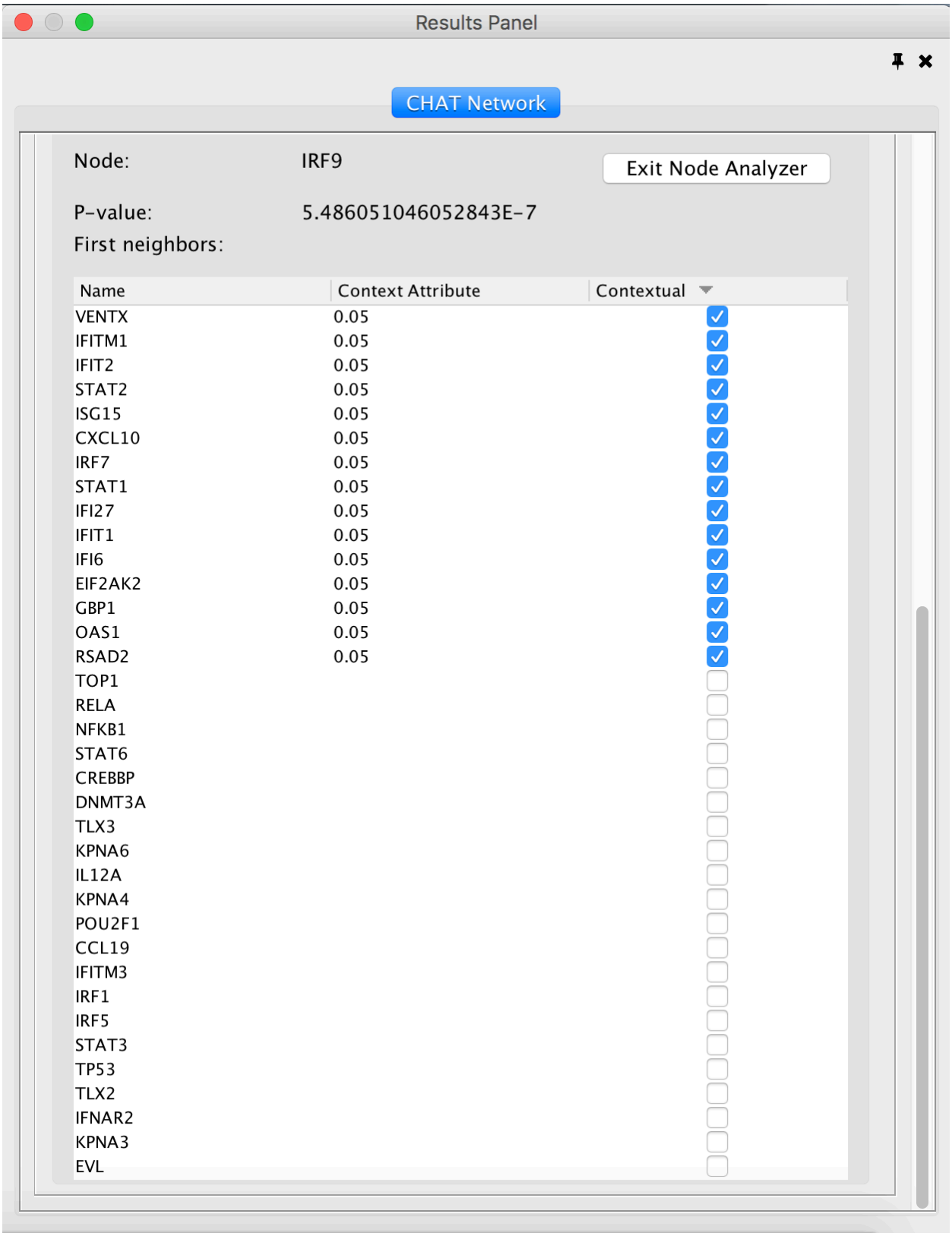

Figure 9. Zooming into specific nodes using the Node Analyzer

References

1. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
2. Aranda, B. *et al.* PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* **8,** 528–529 (2011).