



NetAtlas: A Cytoscape Plugin Tool to Filter Cellular Signaling Networks Based on Tissue Gene Expression

Copyright © 2007 by The Hamner Institutes. All rights reserved.

'Getting Started' Tutorial

This tutorial was written to help users get started using the NetAtlas program.

I. General

1. [What is NetAtlas](#)
2. [NetAtlas License Agreement](#)
3. [SymAtlas Gene Expression Data](#)
4. [User Imported Gene Expression Data](#)
5. [Summary of Implementation](#)

II. How to Use the Application

1. [Starting Cytoscape and Loading a Cellular Signaling Network](#)
2. [Start NetAtlas Plugin](#)
3. [Cytoscape Network Options](#)
4. [Gene Atlas Options](#)
5. [User Input Options](#)
6. [Expression Data](#)
7. [Filter Options](#)
8. [View as New Network](#)
9. [Restore Nodes](#)

I. General

1. [What is the NetAtlas?](#)

NetAtlas is a Java plugin application designed to use tissue gene expression data to filter genes in a cellular signaling network. The NetAtlas plugin allows the creation of tissue-defined networks, identification of network components that are more highly expressed in specific tissues, and the identification of network components that show correlated expression across tissues. The default tissue gene expression data available in NetAtlas is from [SymAtlas](#) and contains human, mouse, and rat gene expression data for a wide range of tissues and has been previously published by the [Genomics Institute of the Novartis Research Foundation](#). The user is also allowed to import their own tissue gene expression data in text file format. The NetAtlas plugin has been developed to work inside [Cytoscape](#), a visualization platform for signaling networks.

2. [NetAtlas License Agreement](#)

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

3. [SymAtlas Gene Expression Data](#)

The default gene expression data used by NetAtlas are from the Genomic Institute of the Novartis Research Foundation's [SymAtlas](#). The [SymAtlas](#) is currently composed of 79 human tissues, 61 mouse tissues, and 44 combined tissues of 3 popular rat strains with 11 peripheral and 15 brain regions. Publications describing the gene expression dataset in detail can be found on the SymAtlas [portal page](#). The microarrays, tissue number, and data types currently available are as follows:

Array	Organism	Tissues	Data Type
-------	----------	---------	-----------

			MAS5	AP Call	RMA	gcRMA
HG-U133A	Human, <i>Homo sapiens</i>	79	Yes	Yes	Yes	Yes
gnGNF1Ba	Human, <i>Homo sapiens</i>	79	Yes	Yes	Yes	Yes
gnGNF1Musa	Mouse, <i>Mus musculus</i>	61	Yes	No	No	Yes
RG_U34A	Rat, <i>Rattus norvegicus</i>	44	Yes	No	No	Yes

4. [User Imported Gene Expression Data](#)

Users may import their own tissue gene expression data from local data files instead of using [SymAtlas](#). The data must have been collected using Affymetrix arrays and may be either expression values or Absent/Present (AP) calls. If expression values will be imported, the data must be normalized and log (base 2) transformed. The file format required by NetAtlas is described [below](#).

5. [Summary of Implementation](#)

The primary function of NetAtlas is to allow users to examine cellular signaling networks based on tissue gene expression patterns. The NetAtlas plugin allows the user to query the tissue gene expression data in ways that highlight different tissue-dependent properties of the signaling networks.

A cellular signaling network in Cytoscape is composed of nodes and edges. The nodes represent components of the network (e.g., proteins, complexes, mRNAs) and the edges represent processes associated with the nodes (e.g., phosphorylation, protein interaction). Within Cytoscape, the nodes and edges can have names and attributes associated with them. For example, the default label of nodes in Cytoscape network is 'canonicalName'. NetAtlas requires that one of the names or attributes be either the Entrez GeneID, Entrez Gene Symbol, GenBank Accession number, or Affymetrix probe identifier. NetAtlas uses the names or associated attributes of the nodes to assign a standardized gene identifier (i.e., Entrez GeneID). The standardized gene identifier is then used to match the specific node in the network to the various probe sets on the Affymetrix microarrays. The user is also asked select what type of organism the network is based upon (since many of the node attributes can be species specific) and whether they want NetAtlas to use the [NCBI HomoloGene](#) database to allow networks of one species to be matched to tissue gene expression data from another species.

After identifying the attributes that describe the Cytoscape network, the user must select the tissue gene expression data that will be used to filter the network. The user may either utilize the default tissue gene expression data from [SymAtlas](#) or import their own gene expression data. With the default tissue gene expression data, the user can select the species, array, and data type used for the analysis. The species available are human, mouse, and rat and the arrays are dependent upon the species. The data types represent

various different preprocessing algorithms (MAS5, RMA, gcRMA) or the standard Affymetrix Absent/Present (AP) calls. When these are selected, NetAtlas retrieves the tissue gene expression data associated with the nodes in the network from a database that resides at The Hamner Institutes and displays it in spreadsheet format within the application. The tissues are on the X-axis and the genes are on the Y-axis. If there are multiple probe sets matching a gene or if there are replicates for a particular tissue, the expression values are averaged (signals) or concatenated (AP calls). The signal values are log (base 2) transformed. With imported tissue gene expression data, the user must select the file containing the data. It is helpful if the data file contains column headers that describe which tissue is contained in each column as the column headers can then be used by the software as labels. In the import process, the user is asked to select the microarray used to collect the data and the software will match the probe set identifiers to the node attributes described above. Similar to the default tissue gene expression data, the gene expression values or AP calls are displayed in a spreadsheet format within the application. The tissues/samples are on the X-axis and the genes are on the Y-axis and if there are multiple probe sets matching a gene, the expression values are averaged (signals) or concatenated (AP calls).

Once the tissue gene expression data is retrieved, NetAtlas provides three primary filtering options: (1) **Minimum Tissue/Sample Expression**; (2) **Selective Tissue/Sample Expression**; and (3) **Correlated Tissue/Sample Expression**. In the **Minimum Tissue/Sample Expression** option, a tissue or sample must be chosen and the nodes in the network are evaluated to see if they meet a minimum criteria based on its gene expression in that tissue in order to be called present. The minimum criteria can be based on a user defined value, confidence intervals calculated across tissues, or AP calls. This option is provided for the creation of tissue-defined cellular signaling networks by identifying which components of a network are likely to be expressed in a particular tissue. In the **Selective Tissue/Sample Expression** option, a tissue or sample must be chosen and the nodes in the network are evaluated to see if they show tissue-specific expression above a set criteria. The criteria is based on fold-change over or under the median value calculated across all tissues. This option is provided for identifying which components of a cellular signaling network show specific expression in a given tissue and may play a role in tissue-specific behavior of the network or may represent potential drug targets with fewer side effects in non-targeted tissues. In the **Correlated Tissue/Sample Expression** option, a gene must be chosen and all the nodes in the network are evaluated to see if their gene expression values are correlated across all tissues with the selected gene. This option is provided to identify components of a cellular signaling network that may have co-evolved due to dependent interactions (e.g., protein-protein interaction, enzyme/substrate interaction).

The nodes identified by each of the filtering criteria are flagged with a different color and can be used to create a new network graph with the original graph layout. The color of the filtered nodes can be selected by the user.

II. How to Use the Application

1. Starting Cytoscape and Loading a Cellular Signaling Network

If you haven't installed Cytoscape, please go to its web site at <http://www.cytoscape.org/> to download and install the program. The NetAtlas plugin requires Cytoscape version 2.4 or later.

With Cytoscape installed, add the NetAtlas.jar file to the 'plugins' folder under Cytoscape root directory. If you installed Cytoscape in the default mode, this will be C:\Program Files\Cytoscape_v2.X.X\plugins.

Start the Cytoscape program and load your cellular signaling network from an existing file with correct format supported by Cytoscape. You can also construct your network by adding new nodes and edges in Cytoscape. In order to use NetAtlas, the node labels (canonicalName) or one of the node attributes must contain its Entrez GeneID, Entrez Gene Symbol, GenBank Accession number, or Affymetrix probe identifier. If the nodes of your network are not labeled with any of these identifiers, there are additional [plugins](#) for Cytoscape that can expand a node's annotation or you can create your own annotation file and import it as node attributes.

If you have multiple networks already displayed in Cytoscape, please select the network of interest as the current view.

2. Start NetAtlas Plugin

From the Cytoscape pulldown menus, select 'Plugins->NetAtlas' to start NetAtlas user interface. The NetAtlas program will display a welcome screen as it initializes and creates the required database connections. If the NetAtlas program successfully loads, the following screen will appear.

NetAtlas Gene Expression Data Filter

File Help

Cytoscape Network Options

Node Attribute: Organism:

Identifier Type: Match to Orthologs on Array

SymAtlas Tissue Gene Expression Options

Organism: Array Type: Data Type:

User Input Options

Import Affymetrix Tissue Gene Expression Data From Local File (tab delimited text)

Expression Data

Filter Options

Tissue/Sample Selected: None. Gene Selected: None.

Minimum Tissue/Sample Expression:

Minimum Maximum Filter Value

Confidence Interval: Mean Median AP_Call

Selective Tissue/Sample Expression: Fold Change versus Median: Over Under

Correlated Tissue/Sample Expression: Correlation Cutoff:

Genes without Expression Data: Hide Fill Color Genes Selected by Filter:

3. Cytoscape Network Options

There are four parameters provided as Cytoscape network options:

Node Attribute: Upon startup, the NetAtlas program will automatically retrieve the attributes (if any) of nodes in the current network loaded in Cytoscape. These attributes will be displayed in the 'Node Attribute' pulldown box. The default label of a node is 'canonicalName'. The user must select the 'Node Attribute' that matches the identifier listed under 'Identifier Type'.

Identifier Type: Four types of identifiers are supported by NetAtlas. These include the Entrez GeneID, Entrez Gene Symbol, GenBank Accession number, and Affymetrix Probe identifier. The user must select the 'Identifier Type' provided in the 'Node Attribute' above.

Organism: The user must select what type of organism the cellular signaling network is based upon (since many of the node attributes can be species specific). The pulldown menu lists all the species included in the [NCBI HomoloGene](#) database. NetAtlas assumes that the model organism represented by the cellular signaling network is one of the species on this list.

Match to Orthologs on Array: If this option is selected, genes from the Cytoscape network of one model organism are matched to orthologs in the tissue gene expression data from another species using the [NCBI HomoloGene](#) database. For example, if the Cytoscape network contains nodes labeled with Entrez GeneIDs from mouse genes and the user wants to filter this network based on human tissue gene expression values, this box must be checked. If the box is not checked, NetAtlas assumes that the model organism of the Cytoscape network matches the species selected for the tissue gene expression data.

4. [Gene Atlas Options](#)

There are three parameters provided as Gene Atlas Options:

Organism: The pulldown menu lists the three species (human, mouse, and rat) for which tissue gene expression data are currently available.

Array Type: List of the Affymetrix microarrays available for the species selected.

Data Type: The data types available for each microarray and species represent various different preprocessing algorithms (MAS5, RMA, gcRMA) or the standard Affymetrix Absent/Present (AP) calls. Not all data types are currently available for each microarray. The currently available data types are listed [here](#).

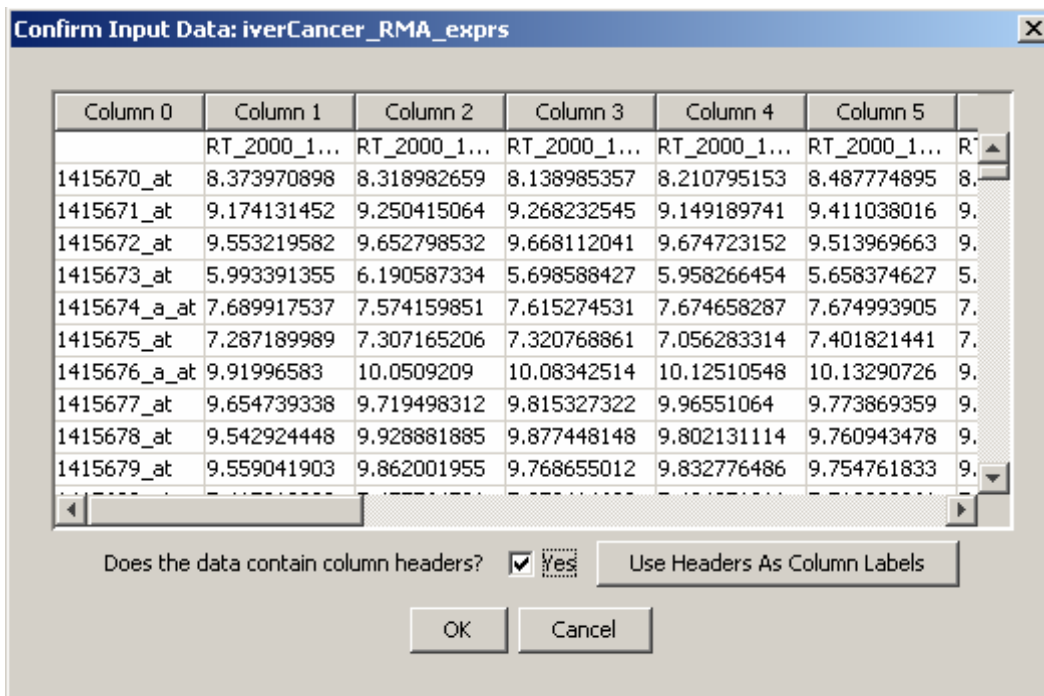
Following selection of the three Gene Atlas Options, the user must press the 'Retrieve' button to load the gene expression data from the database.

NOTE: The tissue gene expression data are retrieved from a MySQL database that resides at The Hamner Institutes. If your program cannot connect to the database, it may

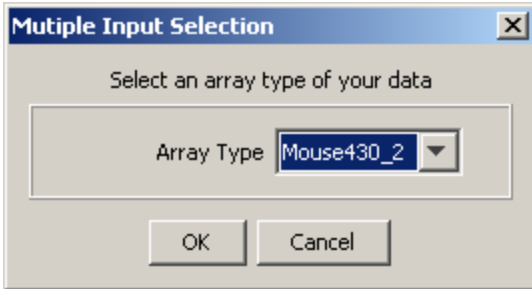
be due to limitations from your firewall. Please check with your IT department if this occurs.

5. User Input Options

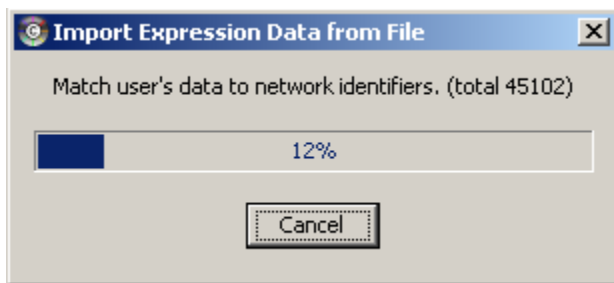
NetAtlas allows users to import their own tissue gene expression data from local files instead of using the [SymAtlas](#) database. The gene expression data must have been acquired using Affymetrix arrays and must be either expression values or Absent/Present (AP) calls. If the user is importing expression values, the data must be normalized and log (base 2) transformed prior to importing into NetAtlas. The input file must be in tab-delimited format with the first column containing the Affymetrix probe identifier and each subsequent column containing a single sample or tissue. Each row of the data must be a separate probe set. It is also advisable for the data file to contain column headers that describe which tissue is contained in each column. To import gene expression data, the user must select the 'Browse' button under the 'User Input Options' section and locate the input file. After NetAtlas reads the file, the data will be displayed in spreadsheet format. If the data contain column headers, the user must select the 'Yes' checkbox and press the 'Remove Headers' button. The column headers will then be recognized by the software as column labels.



The software program will use the Affymetrix probe identifiers in the first column to automatically determine the possible types of microarrays used to collect the data. The user must select the array type from the drop down box. If your Affymetrix microarray is not supported by NetAtlas, please [inform](#) us so that the can add it to the database.



Finally, the program will match the Affymetrix probe identifiers in the imported data to network identifiers. If there are multiple probe sets that match the sample identifier, the data displayed will be an average value (signal) or concatenated (AP call). A progress bar is provided by NetAtlas to allow the user to monitor this process.



6. Expression Data

The tissue gene expression data are displayed in spreadsheet format. If the default [SymAtlas](#) tissue gene expression data are being used, column labels are named with tissues/organs used in the microarray experiments. If imported data are being used, the column labels will be those supplied by the user. The spreadsheet rows are genes named with node attributes from the Cytoscape network. The gene expression values in the spreadsheet are letters if the data type selected is AP calls. Otherwise, numeric values will be displayed. If there are multiple probe sets matching a specific gene or if there are replicates for a particular tissue, the expression values are averaged (signals) or concatenated (AP calls).

NetAtlas Gene Expression Data Filter - NFKB_network

File Help

Cytoscape Network Options

Node Attribute: Organism:

Identifier Type: Match to Orthologs on Array

SymAtlas Tissue Gene Expression Options

Organism: Array Type: Data Type:

User Input Options

Import Affymetrix Tissue Gene Expression Data From Local File (tab delimited text)

Expression Data

canonicalNa...	721 B-lymp...	Adipocyte	Adrenal cor...	Adrenal gland	Amygdala	Appendix	Atrioventric...
CRK	3.30029243...	4.56269747...	3.39323502...	3.67857754...	3.99953252...	3.40438753...	3.28830999...
TBL1X	3.37544295...	2.44027097...	2.51560800...	2.47058701...	2.46353797...	2.62634997...	2.50047903...
LTA4H	7.94465517...	6.27298998...	7.21611499...	6.56308507...	5.37041497...	3.92925989...	4.00024986...
CCDC71	4.91197013...	5.93399000...	6.17075490...	6.05860996...	4.99679493...	5.96094012...	7.10278487...
CYBB	2.58392175...	2.36121670...	2.51166331...	2.42804658...	2.32236162...	2.58485829...	2.55726166...
RASL11B	2.50510001...	3.11776995...	2.74515998...	2.69183993...	3.31848502...	2.69087004...	2.72438502...

Filter Options

Tissue/Sample Selected: Amygdala Gene Selected: None

Minimum Tissue/Sample Expression:

Minimum Maximum Filter Value

Confidence Interval: Mean Median AP_Call

Selective Tissue/Sample Expression: Fold Change versus Median: Over Under

Correlated Tissue/Sample Expression: Correlation Cutoff:

Genes without Expression Data: Hide Fill Color Genes Selected by Filter:

7. Filter Options

There are three primary gene expression filtering options provided by NetAtlas. Applying the gene expression filter is either tissue-based (Minimum Tissue/Sample Expression or Selective Tissue/Sample Expression) or gene-based (Correlated Tissue/Sample

Expression) and either a tissue or gene needs to be chosen before the filter option can be used. A tissue/sample is chosen by clicking on the desired column header in the spreadsheet and a gene is chosen by clicking on the row of data representing the desired gene.

The features of the three filtering options are as follows:

Minimum Tissue/Sample Expression: For this option, a tissue must be chosen and the nodes in the network are evaluated to see if they meet a minimum criteria based on its gene expression in that tissue in order to be called present. This option is provided for the creation of tissue-defined cellular signaling networks by identifying which components of a network are likely to be expressed in a particular tissue.

The user is allowed to set the minimum criteria using:

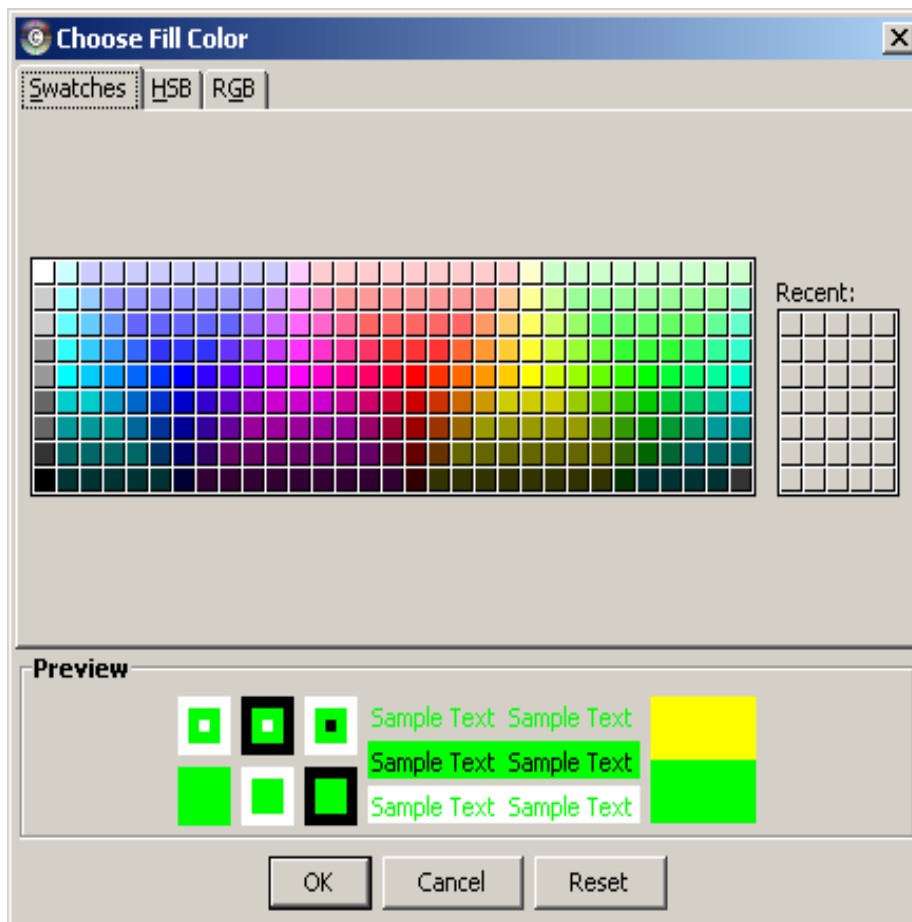
1. **Filter Value:** Based on the tissue/sample selected, the minimum and maximum gene expression values are shown in the respective boxes. The user can input any cutoff between the minimum and maximum values using the slider or directly typing the number into the box. Genes in the network with an expression value in the chosen tissue/sample equal to or above this cutoff will be highlighted by the filter.
2. **Confidence Interval of the Mean:** The mean expression value across all tissues/samples and associated confidence interval is calculated for each gene in the network. The user can select between a 95% or 99% confidence interval. Genes in the network with an expression value in the chosen tissue/sample equal to or above the lower confidence limit will be highlighted by the filter.
3. **Confidence Interval of the Median:** The median expression value across all tissues/samples and associated confidence interval is calculated for each gene in the network. The user can select between a 95% or 99% confidence interval. Genes in the network with an expression value in the chosen tissue/sample equal to or above the lower confidence limit will be highlighted by the filter. The lower median confidence limit is calculated as follows: $(N * 0.5 - Z * \text{Sqrt}(N * 0.5 * (1 - 0.5)))$ where N is the number of tissues and Z is 2.58 for 99% confidence or 1.96 for 95% confidence.
4. **AP Call:** Based on the tissue/sample selected, each gene in the network will have one or more Affymetrix Absent/Present (AP) calls associated with it. The AP call can be either 'P' for present, 'A' for absent, or 'M' for marginal. The genes in the network with the selected AP call or better in the chosen tissue/sample will be highlighted by the filter (e.g., if the selected AP call is 'M' then genes with a 'P' or 'M' will be highlighted; if the selected AP call is 'P', then only genes with an AP call of 'P' will be highlighted). If the gene has more than one AP call, the final value will be based on a majority vote.

Selective Tissue/Sample Expression: For this option, a tissue/sample must be chosen and the nodes in the network are evaluated to see if they show tissue- or sample-specific expression above a set criteria. The criteria is based on fold-change over or under the

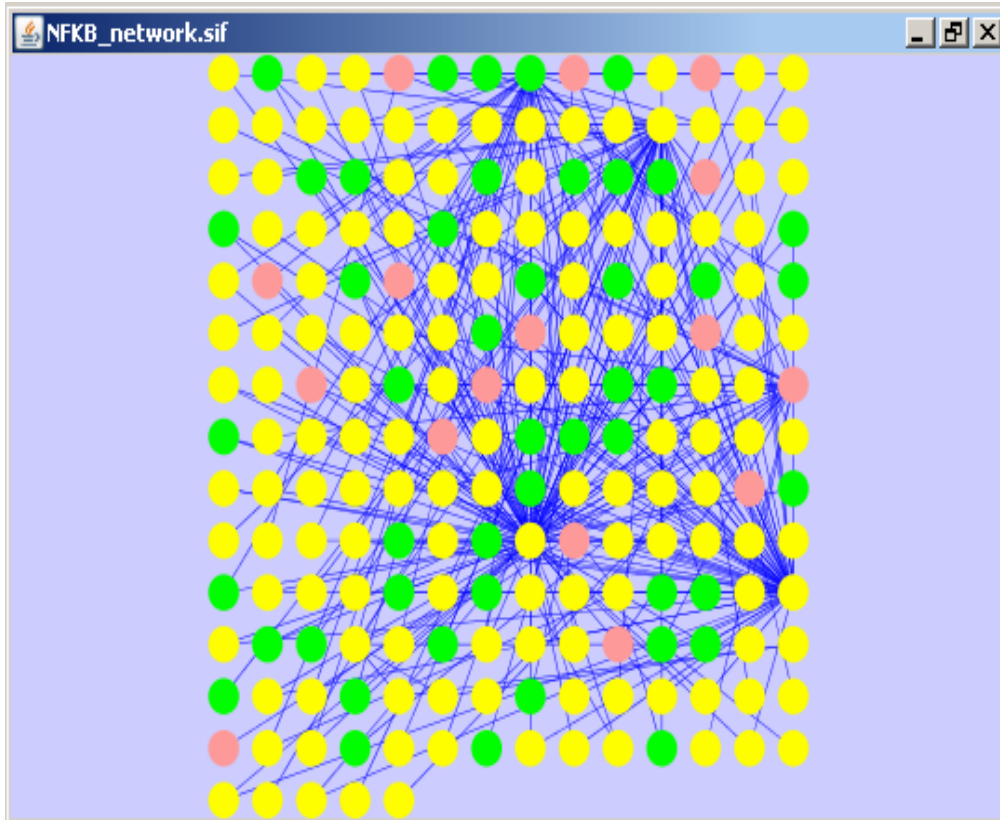
median value calculated across all tissues/samples. The user is allowed to input the fold-change amount and select whether the fold-change is over or under the median value. This option is provided for identifying which components of a cellular signaling network show specific expression in a given tissue/sample and may play a role in tissue-specific behavior of the network or may represent potential drug targets with fewer side effects in non-targeted tissues.

Correlated Tissue/Sample Expression: For this option, a gene must be chosen and all the nodes in the network are evaluated to see if their gene expression values are correlated across all tissues/samples with the selected gene. This option is provided to identify components of a cellular signaling network that may have co-evolved due to dependent interactions (e.g., protein-protein interaction, enzyme/substrate interaction).

After defining the filter option, the user is allowed to select what to do with nodes that are not represented on the associated Affymetrix array. The user may either 'Hide' the nodes or select a different 'Fill Color' so they are flagged with a specified color. The 'Fill Color' is selected by clicking on the color box and selecting the desired color. The fill color for 'Genes Selected by Filter' can be modified in the same way.

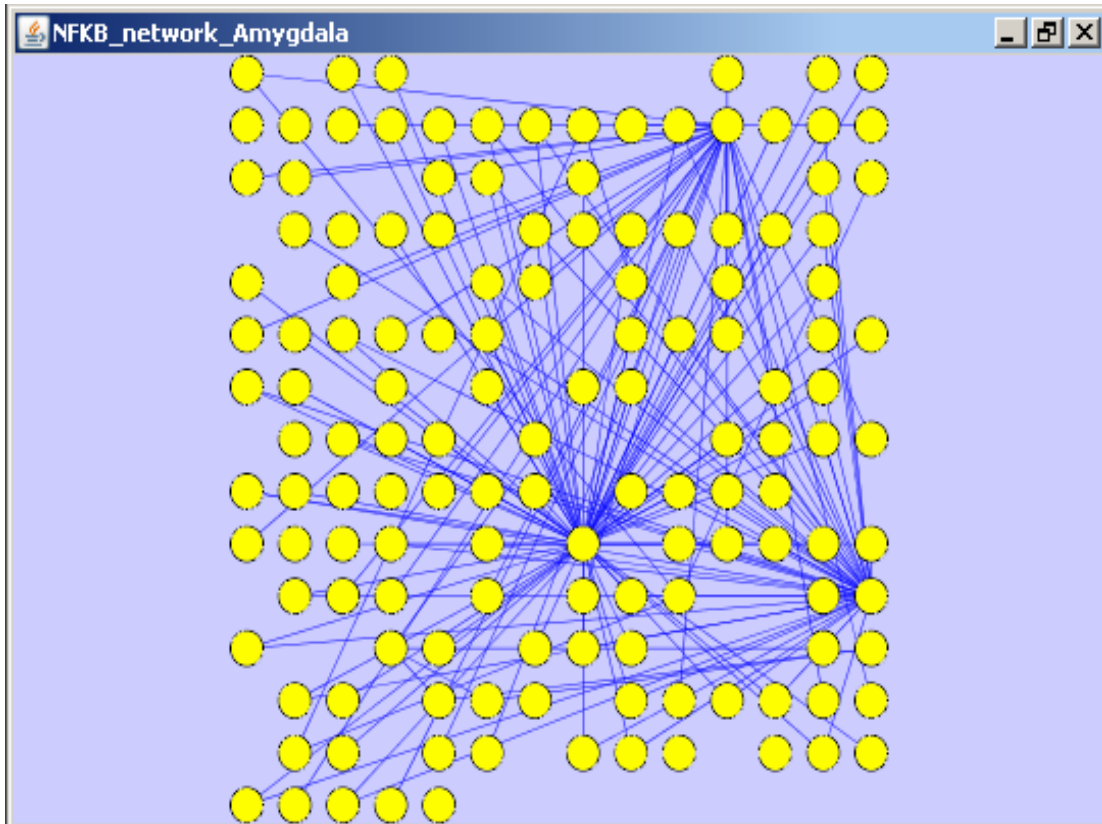


By clicking 'Apply Filter' button, nodes without data will be hidden or filled with the chosen color (green in the example below); nodes selected by the filter will be flagged with the chosen color (yellow in the example below); the remaining nodes will stay unchanged (red in the example below).



8. [View as New Network](#)

By selecting 'New Network', NetAtlas will extract all the nodes selected by the filter to form a new network view. The example below shows the creation of a new network with only those nodes selected by the filter (i.e., all nodes without gene expression data and those not selected by the filter are not included).



9. Restore Nodes

By clicking 'Restore Nodes', nodes without gene expression data that were hidden from the network will be brought back into the current network view.

More Questions?

If you have further questions, concerns, comments or suggestions, please contact us at:
rthomas@thehamner.org
lyang@thehamner.org.