

Structured
Genetic
Association
Mapping

GenAMap

An Integrated analytic and visualization platform for eQTL and GWA analysis

Ross E Curtis, Eric P Xing

Additional Algorithmic and Programming Support: Seyoung Kim, Michael Zuromskis, Kriti Puniyani, Sharath Babu, James Moffatt, and Kelly Chan.

User's Manual

Brought to you by the SAILING Laboratory at Carnegie Mellon University



And the Carnegie Mellon-University of Pittsburgh Joint PhD Program in Computational Biology



Also supported by the Lane Center for Computational Biology

RAY AND STEPHANIE LANE
Center for Computational Biology

Contents

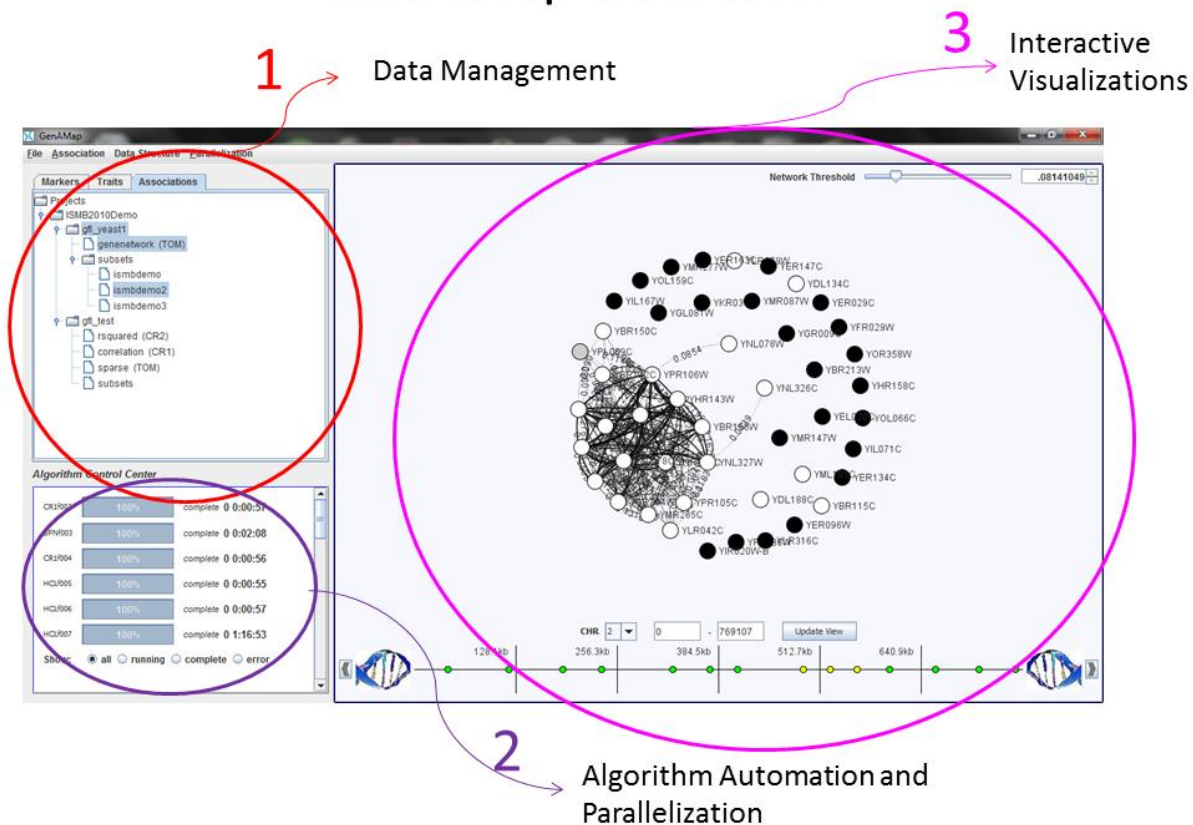
Starting the Application	4
Loading in Data	5
Data Organization	5
Loading in Data	5
Marker Tab-delimited format	5
PLINK style format.....	6
Trait Import	7
Trait Subsets.....	7
Removing or Renaming Data	8
The Algorithm Control Center	8
Controls.....	8
Error Handling	9
Visualizations	9
Opening and Changing What You See	9
The Chromosome View	9
Network Creation.....	9
Opening the Dialog	10
Loading a Pre-constructed Network	10
Algorithms.....	11
Correlation	11
Correlation r^2	11
SFN: Scale Free Network.....	11
TOM: Topological Overlap Matrix.....	11
Network Visualization	11
Matrix View for an Overall Picture - Controls.....	11
Downloading a Large Network.....	11
Setting a clustering.....	12
Adjusting the Color Scale	12
Panning and Zooming	13
JUNG View – for exploring data structure – controls	13
Association Set Creation	14

Loading in an Association Set.....	14
GFlasso	14
Wilcoxon Sum Rank Method.....	14
Association Visualization.....	14
Works Cited.....	15

Starting the Application

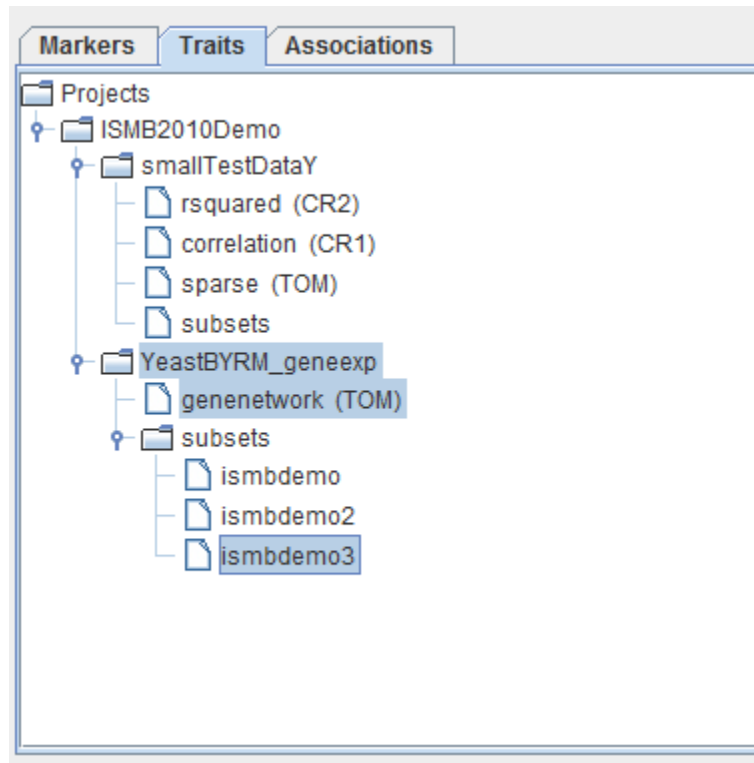
In order to run GenAMap, it must be connected to a cluster and a remote database. We have provided a web interface that GenAMap uses to access the SAILING Lab cluster. Upon login, you will need to provide a username and password into this system. Ross will give you this in a separate email. When you have downloaded the executable jar file, double click on it to run it and then enter in the logon information. This should allow the main GUI to display. If it does not, then the username and password are probably not correct.

GenAMap Overview



Once GenAMap is running, you will notice three parts of the application. The top left is the data control panel. Below it is where you will be managing your running algorithms. On the right is the data visualization panel. We will discuss each of these parts in this documentation.

Loading in Data



Data Organization

In the top left there is the data control panel. This is where we can browse through our stored data and select what we are going to be viewing or feeding to algorithms.

All data is organized into three tabs: Marker, Trait, and Association. For all data, it is organized into a project. The project stores all three types of data, which can be accessed from their appropriate tab.

In order to familiarize yourself with the GUI, right-click on the word “Projects” and select the “Create New Project” option. Type in a name and you will have a new project!

Loading in Data

GenAMap accepts data in two formats – the PLINK style ped/map format, or a tab-delimited format. All data imports can happen from the right-click menu from any project. Upon selection, an open file dialog will come up and the user can browse to the main file. Label files can later be selected from the dialog once a file is appropriately chosen.

Marker Tab-delimited format

SNP data can be imported into GenAMap from a delimited file. Samples can be labeled, or the ordering in the file can be an implicit identifier. If this is the case, ensure that the phenotype files are in the same order with the same number of samples.

Labels can be in a separate file, or they can be column and row headers in the file. In all files imported into GenAMap, missing values are not supported at this time.

If the marker file is to have labels in it on import, the first three columns should constitute the label. If not, a file with three columns must be prepared in the format NAME, CHR, and LOCUS. Name represents the name of the marker, chromosome represents the chromosome and locus represents the physical location on the chromosome. An example file may look like:

```
rs3345as23    4      1234567890
rs22390120    4      1234456780
rs23094580    5      124355670
```

...

PLINK style format

Data can be prepared similar to the PLINK style with PED and MAP files. The main difference between our support and PLINK's support is that the PED file can contain 0 to n phenotypes, where n is a user-specified term on import.

Import Data

PED/MAP Import

Project:

Dataset name:

PED file: ... ?

MAP file: ... ?

Labels: ... ?

File Format

☒ Phenotype/column headers in PED file

☐ Phenotype names in Label file

delimiter:

Marker type:

☒ Numeric

☐ Alphabetic

We use the PLINK file import as an example of what users will see when importing data. They can choose which project to create, name the dataset, and browse to the ped and map file. Other limited options help specify the exact format from the data.

Trait Import

Trait values can be imported as a delimited matrix file or in the phenotype PLINK-like style where the first two columns are fid and pid and then the remaining columns are the data. Labels can be in the file itself, or in an accompanying file. Sample labels, if omitted in a delimited file, will be generated as with marker data. It will therefore be imperative to have the samples in the same ordering between the phenotype file and the genotype files.

Trait Subsets

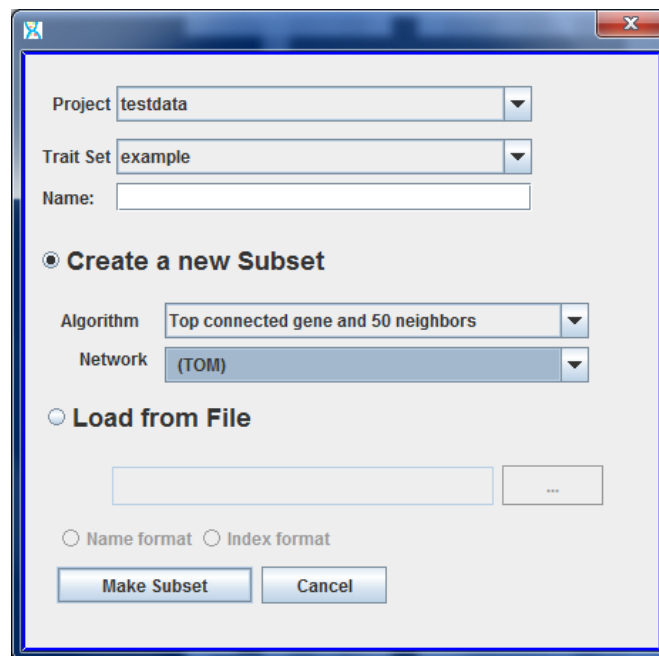
As a user explores through the trait data, they will have opportunities to create subsets of the traits. This will allow the user to store traits of interest, traits that are associated to particular chromosomal regions, or to save smaller sets for viewing.

Save traits as a subset

When using a visualization to view the traits, the user can select or zoom into a particular set of traits and then choose to save the selected or visible traits as a subset (which ever option is applicable in the view).

Add a Subset algorithmically or from a file

The user can right click on the subset node in the data tree and choose to add a subset. This will bring up the subset dialog.



From the subset dialog, the user can choose to load in a subset in either name format (one trait name per line), or in index format (where index is the index of the trait in the original file). This will create the subset from a file that the user selects. The user can also choose one of four subset creating algorithms to generate a subset of the most connected genes or its neighbors. These algorithms are different in that they are run locally.

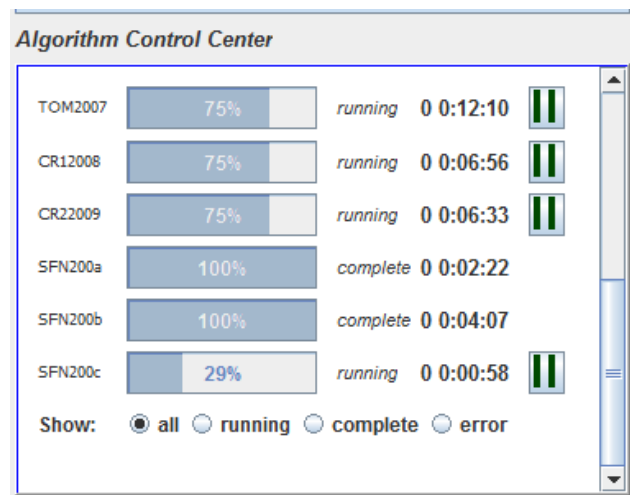
Subset Tree

Once a subset is created, it is added to the tree under the subset nodes. When visualizing a network the user can select the subset to view only the nodes contained in that subset.

Removing or Renaming Data

Data can be renamed or removed from the right-click menu. Deleting a set of data will cause a recursive deletion of all other associations sets, networks, etc, that reference it.

The Algorithm Control Center



The algorithm control center allows the user to have a glimpse and some control into what is happening as algorithms are run on the cluster.

Controls

The algorithms are each assigned an id when they are initially created. This allows them to be tracked on the database. The id as shown in the algorithm control center is what gives the user some control over the algorithm. When the user right-clicks on the id, they can see error information or delete an algorithm (stopping all processing on the cluster). The id has a three letter abbreviation of the algorithm name, a user identifier, and an algorithm number (hex format).

Additionally, the user can use the algorithm control center to track an algorithm's progress. Each algorithm will have a variable number of steps, and each step will spawn a variable number of processes. As these processes happen, the percent complete will update and the status will reflect what is happening on the cluster.

If at any time an algorithm needs to be stopped, the user can push the pause button. This will not stop the algorithm immediately, but will allow current processes to reach their stopping point before moving to the next step. This, of course, could be instantaneous, or could take a week depending on the length of the current step. ☺ To start the algorithm, the user just repushes the button.

The user can also sort the algorithms that they have tracking using the filter controls at the bottom.

Error Handling

At this point in time, GenAMap has very limited error handling capability. The user can restart an algorithm in error and it will start from the step before the error occurred. Other errors should contact Ross for further investigation.

Visualizations

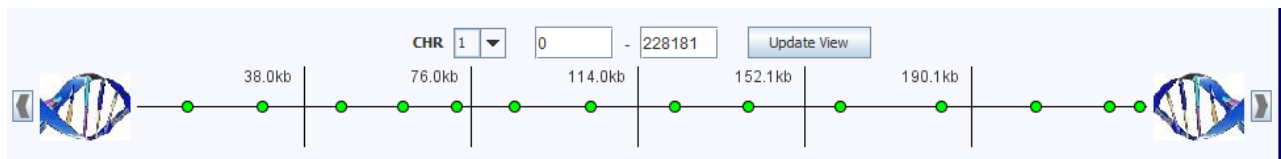
Opening and Changing What You See

As you switch between tabs and select datasets in the appropriate tabs, the visualization for that tab will open. For example, if I want to open up a chromosome view for a particular dataset, I can browse to the marker tab and select that dataset. The chromosome view will open.

The trait tab allows the user to select the type of structure they want to view the traits in. Without some kind of structure, no traits can be viewed. However, once traits have structure the user has a variety of controls from a few different views and management of subsets.

The Chromosome View

Once a user has loaded in marker data, they can instantly view it using the chromosome view.



The chromosome view allows the user to explore the SNP markers, and where they fall on the chromosome. The user can change between chromosomes using the drop down menu. The user can also manually set which base pairs they want to visualize.

The user can also zoom into certain regions by selecting and then right-clicking to select the zoom selection option. The user can also use the scroll wheel. Once the user has zoomed in, they can pan around by dragging the SNP markers (circles), or using the arrows on the sides.

Further functionality will eventually be added, but this basic visualization gives the user a feel for their overall data.

Network Creation

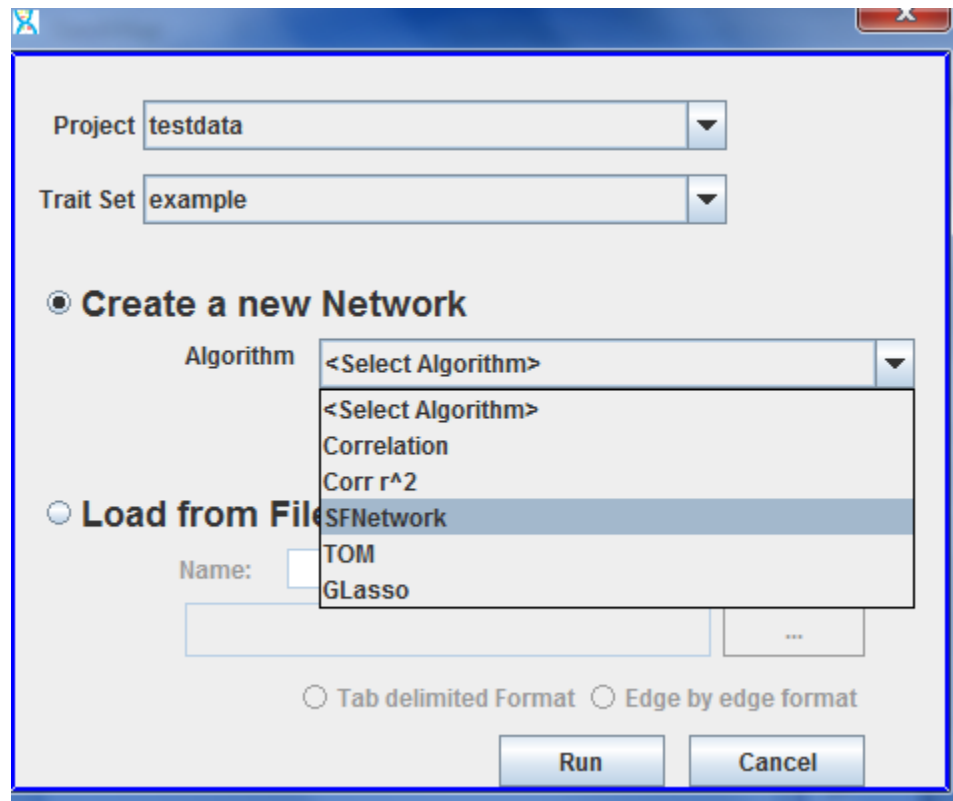
In order to run an algorithm like the GFlasso, networks will need to be created or loaded in. These networks can be created through a few algorithms, which are described here. Current visualization tools will also be discussed.

Opening the Dialog

In order to run an algorithm to create a structure, like a network, for a trait:

- Browse to the trait tab
- Right click on the traitset that you will be adding structure to
- Click on Add Network Data

At this point you will be given a dialog which can be used to change the project and traitset, if desired. It can also be used to select an algorithm that you would like to run. At this time, GenAMap supports just four different algorithms. GenAMap also supports loading in a network from a data file in two formats.



Loading a Pre-constructed Network

When a network is loaded in, the user will need to give it a name and browse to its location in the file system. Two different formats are supported, and the user must select which format they are using.

The tab delimited format takes a $K \times K$ matrix of weights, where K is the number of traits in the trait set. When no edge is present, the value should be set to exactly 0. The weights can include both positive and negative numbers – negative numbers will be treated as strongly as positive numbers in the visualization algorithms. (We take the absolute value in all visualizations).

The edge by edge format is also a delimited file. However, each edge has its own row. The first two columns are the names of the traits involved in the edge, and the third column is the weight between the traits. These names must match exactly the trait labels imported into the database.

T1	T5	1.73
T1	T7	1.72
T1	T8	1.01
T1	T9	-1.73
T1	T15	0.73

Algorithms

Select which algorithm you would like to run, and then press run. The algorithm will be kicked out to the cluster and will go through a series of steps. The job can be monitored through the Algorithm Control Center.

Correlation

This algorithm will calculate Pearson's correlation coefficient between each set of nodes and construct a network based on this value. The algorithm currently has an automatic cutoff where it determines that the correlation is too small to be an edge. This will be adjustable in future versions.

Correlation r^2

This algorithm calculates the r^2 value between all nodes in the network and adds an edge if it is over the applied threshold.

SFN: Scale Free Network

Studies have shown that biological networks are often scale-free and modular. GenAMap has its own implementation of the method used by Zhang and Horvath to generate a network with a soft threshold (Zhang B 2005).

TOM: Topological Overlap Matrix

Once the Scale-free network has been constructed, GenAMap can use its own implementation to calculate the TOM as described in (Zhang B 2005).

Network Visualization

Networks can be visualized using two different views – the matrix view and the JUNG view.

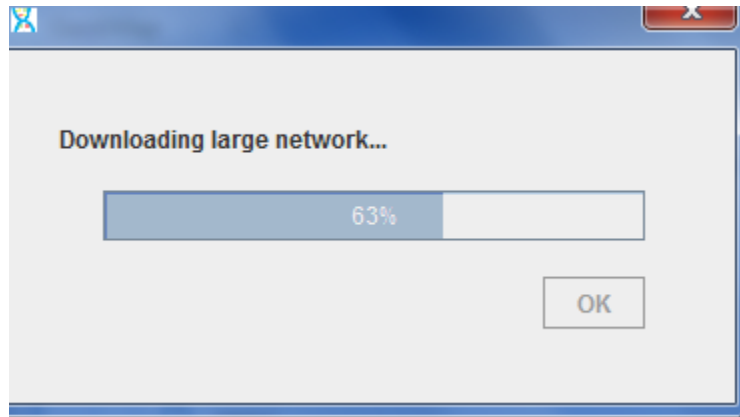
Matrix View for an Overall Picture - Controls

Any time that there are more than 200 genes selected to view, GenAMap will default to a matrix view, or a heatmap. At this time, we support black and white images to display the values, where white is high and black is low.

Downloading a Large Network

When a new network has been newly created, it will exist on the database in its entirety. However, in order for us to provide the user with a good experience zooming in and out, we create several resolutions. This has to happen once per network per clustering.

Thus, the first time a network is requested to be viewed, the user will see the Downloading Large Network dialog and will have to wait until this is finished to continue using the network. This will only happen the first time. It will store some data on the user's local machine in the drive that they place their executable jar file.



Setting a clustering

Genes are often not ordered in the best format for viewing when the user has loaded them into the network. In order to enhance viewing, GenAMap supports running an agglomerate hierarchical clustering on the network to make things more viewable. These clusterings are assigned to a trait set and are stored on the database, and are only applicable to the Matrix View.

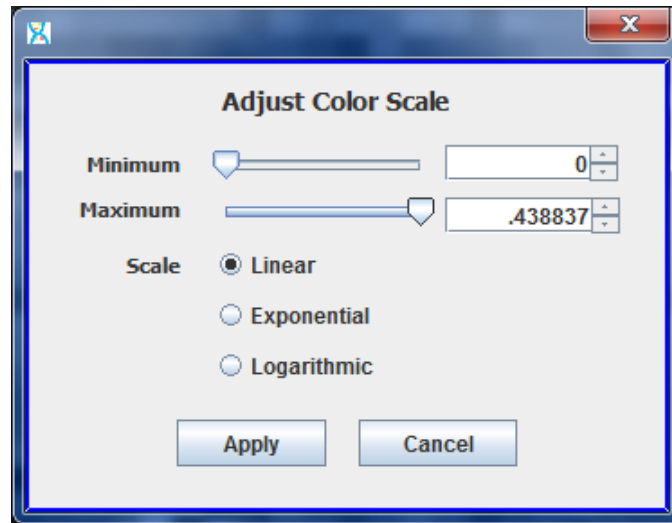
In order to create a clustering, the user must right click on the Matrix View and select Add a Cluster. They will be presented with a dialog to give the new clustering a name and assign a network from the trait set to use in the clustering. This will be kicked out to the cluster and when finished, it will be available to the user. The user can select the clustering they would like to see from the right-click menu.

In order to set a cluster that has already finished running, right click on the Matrix View and select Select a Cluster and choose the desired cluster according to its name.

A cluster can also be loaded. Each line is a single number, and represents the index of the trait in the original file.

Adjusting the Color Scale

The user has quite a bit of control over the coloring of the image. Once a user right clicks on the image, he can select to adjust the color scale, and retrieve the Adjust Color Scale dialog.

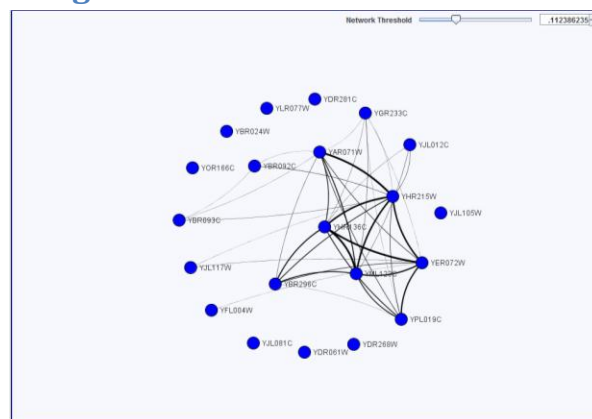


In the adjust color scale dialog, the user can set the minimum (black) value and the maximum (white) value. The user can also determine what kind of scale to use between the minimum and maximum values. A linear scale gives equal weight to all values, while Exponential will emphasize the difference between larger values. A logarithmic scale will emphasize the difference between smaller values. The changes will be applied when the dialog closes.

Panning and Zooming

A user can zoom in and out using the mouse wheel. Additionally, the user can hold down the left mouse button and select an area of the image. The program will then zoom in on this particular area. Once zoomed in, the user can hold down the control key and then drag the mouse to pan around the image.

JUNG View – for exploring data structure – controls



In the JUNG view, named after the open source library it was generated from (Madahain JO 2005), the user can experience the gene-gene interactions up close and personal with the visualization tools. The network is shown using a ball and stick model. The user has the ability to turn on and off labels, weights, and change the layout from the right-click menu.

Additionally, from the right-click menu the user can switch back to the Matrix View, reset the view, and toggle between a transformation mode and a picking mode. In the picking mode, the user can move certain vertices around, while in the transformation mode they have control over the entire graph.

The user does not have to always show all edges in the database at once, but can play with the network threshold in the top right to add or remove edges as they see fit.

Association Set Creation

Although the data exploration tools in GenAMap are limited, they provide the user with an overall feel for the structure of the data. These tools are especially more powerful when combined with an association set and the user can browse through the structure of the phenome while visualizing the structure of the genome.

In order to visualize this connection, an association set must be created. The prerequisites for creating an association set are to have a trait set and a marker set with the same set of samples loaded in and ready. This will allow the user to open up the association creation dialog and select to load or run an algorithm. At this time only the GFlasso and Wilcoxon Sum Rank test methods are implemented.

Right click on a project and select to add an association set to get started. Once a trait set is selected, only valid marker sets will be available.

Loading in an Association Set

An association set can be loaded in as a matrix in a tab delimited format. The traits and markers are not labeled and must be in the same order as they were upon import. The values in the matrix can be either p-values or association strength value. In the case of p-values, 1 means no association while a 0 means no association in the other case.

GFlasso

The GFlasso is currently implemented in a 17 step process which runs the algorithm on subnetworks of the data after a clustering step. A faster implementation will be available soon. This implementation gets the most promising 1000 markers and then runs that set on each subset of the network.

For details on a run or help to interpret the results, please talk to Ross.

Wilcoxon Sum Rank Method

The nonparametric WSR method with FDR control is implemented as described in (Zhu J 2008).

Association Visualization

Once the results from an association run are complete, the user is able to view them much the same way that they would visualize a network. This is done with a matrix view at the start, along with a zoomed in JUNG view, if they switch to that view. Clusterings are also available.

Inside the JUNG view, the user has the opportunity to explore both the genome space and the network space with the same controls as before. At this point, the user can also color the nodes based on the values of the association strength. To do this, the user would select a genome region, right click, and then highlight associated traits. This will continue to show the structure while allowing the user to see the association strengths.

Works Cited

Madahain JO, Fisher D, Smyth P, White S, Boey YB. "Analysis and Visualization of Network Data using JUNG." (Journal of Statistical Software) VV, no. II (2005).

Zhang B, Horvath S. "A General Framework for Weighted Gene Co-Expression Network Analysis." *Statistical Applications in Genetics and Molecular Biology* 4(1), 2005.

Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks." (Nature Genetics) 40, no. 7 (2008): 854-861. .