GenePattern

HierarchicalClusteringViewer Documentation

Description: Visualizes and manipulates hierarchical clustering results.

**Author:**   Joshua Gould (Broad Institute), gp-help@broadinstitute.org

Introduction

This visualizer provides a tree view (dendrogram) of the results of the HierarchicalClustering module. The dendrogram provides a graphical view of the distance between clusters identified by the algorithm, as well as a heat map of the associated data. This module requires the following files, which are output from the HierarchicalClustering module:

- A CDT file, which contains the original data, sorted into clusters
- Either a GTR file, which contains the distance measures between gene (row) clusters in the dataset, or an ATR file, which contains the distance measures between sample (column) clusters. When both files are provided, a 2-dimensional "biclustering" of the data is shown

Values are displayed in a heat map format, representing higher values in more intense red colors and lower values in more intense blue colors.

The following figure shows an example of this view.
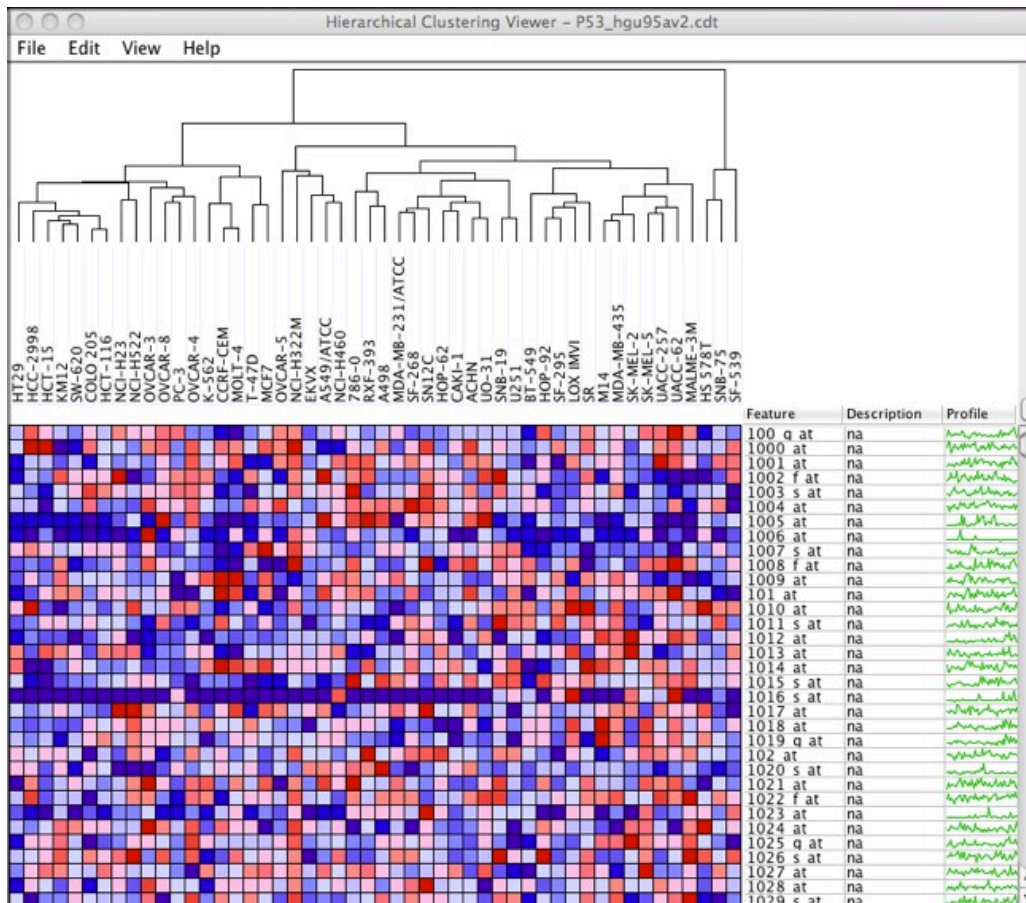
Available Options

Display options

There are a number of options controlling the appearance of the heat map that can be set by selecting *View>View Options* from the menu:

- You can choose between radio buttons for *Relative* and *Global* color schemes. When you select *Relative*, the heat map converts values to colors based on the mean, minimum, and maximum values in each row. When you select *Global*, the heat map converts values to colors based on the mean, minimum, and maximum values in the entire dataset. To display the color legend, select *View>Color Scheme Legend*.
- You can choose between the radio buttons:
    - *Use Color Gradient*: The heat map uses a linear color gradient to convert values to colors
    - *Use Discrete Colors* : The heat map uses a discrete color map to convert values to colors
- You can optionally load a color map by clicking *Load Color Map*. This color map is a file containing three colors, with one color listed per line:
    - First color = the color for the minimum values
    - Second color = the color for the mean value
    - Third color = the color for the maximum values
    The colors can be specified as a decimal, octal, hexadecimal integer, or r:g:b triplet (e.g., #9900FF, 0:0:0, 171). (For more information on numerical color specification, see http://en.wikipedia.org/wiki/RGB_color_model#Numeric_representations)
- The *Show Grid* checkbox controls whether a grid is drawn around each element.
- The *Show Profile* checkbox controls whether the profile is shown in the table for each feature.
- The row size and column size sliders control the size of each element in the heat map. When the *Maintain Square Aspect* checkbox is selected, the row and column sizes are kept in synch.
- You can choose to hide/show feature descriptions, feature names, and sample names by clearing/selecting the corresponding checkbox.
- The *Feature Label Width* slider controls the size of feature (gene) labels. Similarly, the *Sample Label Height* slider controls the size of sample labels. These controls do not appear unless you have features and/or samples loaded. To load feature labels, select *File>Label Features.* To load sample labels, select *File>Label Samples*.
- You can control the sample or feature (gene) dendrogram height by editing the corresponding text box and pressing RETURN (or ENTER) to apply your changes.
- You can set the sample or feature (gene) dendrogram line thickness by editing the corresponding text box and pressing RETURN (or ENTER) to apply your changes.

Saving images

Users can save the heat map image to a file by selecting *File>Save Image* from the menu. Supported image formats are BMP, EPS, JPEG, PNG, and TIFF.

Saving datasets

You can use the viewer to create a new dataset.

1. Select *File>Save Dataset*. A window appears.
2. Choose the features and samples to include in the dataset by selecting the features and sample names in the viewer.
3. Choose a location and name for the new dataset.
4. Click *Save* to save the new dataset.

Dendrogram color

Note: Feature (gene) dendrograms are displayed when a GTR file is provided and sample dendrograms are displayed when an ATR file is provided.

You can set the color of a selected branch in the sample or feature dendrogram branch color by selecting *Edit>Sample Dendrogram Branch Color* or *Edit>Feature Dendrogram Branch Color*.

You can also set the color of a dendrogram node by editing the ATR or GTR file. Add a column to the ATR or GTR file that contains the node color. For example (added column shown in bold):

NODE1X	GENE78X	GENE77X	0.964702	**#9900FF**
NODE2X	GENE25X	GENE8X	0.963298	**#990000**

Dendrogram rotation

You can flip a selected branch in the sample or feature dendrogram by selecting *View>Flip Sample Dendrogram* or *View>Flip Feature Dendrogram*.

Feature labels

Feature (gene) labels use color to annotate features in the heat map.

To use feature labels:

1. Create a GRP, GMX, or GMT file. (Or download from a site like the Molecular Signatures Database [MSigDB] or from supplementary files for journal articles of interest.)
2. Select *File>Label Features* to open your feature (gene) set file. You have the option to restrict the view to features in your feature set if you did not cluster the features in your dataset. A color bar appears next to each feature in the feature set in the table.
3. Select E*dit>Feature Labels* to edit the color or close the feature set.
4. In the Feature Labels window, select your feature set from the drop-down list. The color assigned to that feature list appears in the box to the right.

   - To change the color, click the box and select a new color.
   - To delete the feature set and remove the color bars from the table, click *Delete*.

The feature annotations legend can be saved to a file by selecting *File>Save Feature Labels Legend*.

Sample labels

Sample labels use color to annotate samples in the heat map.

To use sample labels:

1. Create a [sample info file](#).
2. Select *File>Label Samples* to open your sample info file. A color bar appears below each sample name.
3. Select *Edit>Sample Labels* to edit the color or close the sample info file.
4. In the Sample Labels window, select your sample class from the drop-down list. The color assigned to that sample class appears in the box to the right.

   - To change the color, click the box and select a new color.
   - To delete the sample class and remove the color bars from the table, click *Delete*.

The sample annotations legend can be saved to a file by selecting *File>Save Sample Labels Legend*.

Loading descriptions

GeneCruiser retrieves information about Affymetrix probe identifiers and adds the information to the feature table.

To use GeneCruiser annotations:

1. Select *File>GeneCruiser*.
2. Select the features that you want to retrieve annotations for in the table.
3. Choose which fields to retrieve from GeneCruiser in the GeneCruiser dialog.
4. The annotations appear in additional columns in the table.

You can also load descriptions from a comma separated values (CSV) file. To load descriptions from a CSV file, select *File>Load Descriptions From Csv File.*

Finding features

To find a feature (gene) in the heat map, select *Edit>Find*. You can choose whether to match the case of the text you're searching for by selecting the *Match case* checkbox at the bottom of the find dialog. You can also choose which column to search from by selecting the desired column name in the dropdown box.

Selecting samples and features by name

To select a sample or feature by name:

1. Select *View>View Options*.
2. Ensure that the corresponding *Show* (*Feature* or *Sample*) *Name* checkbox is selected.
3. Increase the *Column Size* or *Row Size* to a value large enough to make the name visible.
4. Select the first name by clicking it; add to the selection by control-clicking (PC) or command-clicking (Mac) more names.

Profile plot

The expression profile for a feature plots expression value per sample.

To display an expression profile:

1. Select one or more features and optionally select one or more samples.
2. Click *View>Profile*. Alternatively, right-click and select *Profile* from the context menu or click the profile plot column in the feature table.

Centroid plot

The centroid plot shows the mean expression value for each sample. The error bars represent the standard deviation.

To display a centroid plot:

1. Select two or more features (genes) and optionally select one or more samples.
2. Click *View>Centroid Plot.* Alternatively, right-click and select *Centroid Plot* from the context menu.

Histogram

The histogram plot shows the distributions of expression values. The vertical axis represents number of occurrences, and the horizontal axis represents the binned expression values.

To display a histogram:

1. Select one or more features (genes) and optionally select one or more samples.
2. Click *View>Histogram.* Alternatively, right-click and select *Histogram* from the context menu.

Nearest neighbors

The nearest neighbors item shows other features (genes) whose expression values follow similar trends to the selected feature. There are four choices for the distance metric (which defines the method used to determine mathematical distance between features):

- Pearson (default): Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. It is a measure for how well a straight line can be fitted to a scatter plot of x and y. If all the points in the scatter plot lie on a straight line, the Pearson correlation coefficient is either +1 or -1, depending on whether the slope of line is positive or negative. If it is equal to zero, there is no correlation between x and y.
- Cosine: A measure of similarity between two features based on the cosine of the angle between the vectors (consisting of the expression values for each) in mathematical space.
- Euclidean: Corresponds to the length of the shortest path between two points. Takes into account the difference between two features directly, based on the magnitude of changes in the sample levels. This distance type is usually used for data sets that are normalized or without any special distribution problem.
- Manhattan: Also known as the city-block or taxi cab distance; the Manhattan distance is the sum of distances along each mathematical dimension between two points.

To display the nearest neighbors:

1. Select one feature.
2. Click *View>Nearest Neighbors.* Alternatively, right-click and select *Nearest Neighbors* from the context menu.

Scatter plot

The scatter plot lets you compare expression values in two samples.

To display a scatter plot:

1. Select two samples.
2. Optionally select one or more features.
3. Click *View>Scatter Plot.* Alternatively, right-click and select *Scatter Plot* from the context menu.

Keyboard Shortcuts

You can use your keyboard to quickly accomplish many tasks. To find the shortcuts for common commands, look in the menus or select *Help>Keyboard Shortcuts* to see a list of available shortcuts.

Parameters

Name	Description
cdt file (required)	A clustered data table (CDT) file contains the original data, but reordered according to clusters.
gtr file	The Gene Tree file (GTR) records the order in which genes (rows) were joined during clustering. *Either a GTR or an ATR file is required.
atr file	The Array Tree file (ATR) records the order in which samples (columns) were joined. *Either a GTR or an ATR file is required.

Input Files

1. CDT file

   Contains the original data reordered according to clusters.

2. GTR file *or* ATR file

   Records the order in which genes/rows (GTR) or samples/columns (ATR) were joined during clustering

Example Files

To see an example sample information file for sample labeling, see:
ftp://ftp.broadinstitute.org/pub/genepattern/example_files/HierarchicalClusteringViewer/ALLAML
Test_SampleInfo.txt

Platform Dependencies

Module type:	Visualizer
CPU type:	any
OS:	any
Language:	Java

GenePattern Module Version Notes

Version	Release Date	Description
10	3/19/2010	
11	4/10/2013	Fixed to work with Java 7.