Clustering and displaying microarray data

George Bell, Ph.D.

Bioinformatics and Research Computing

Hot Topics – March 2008

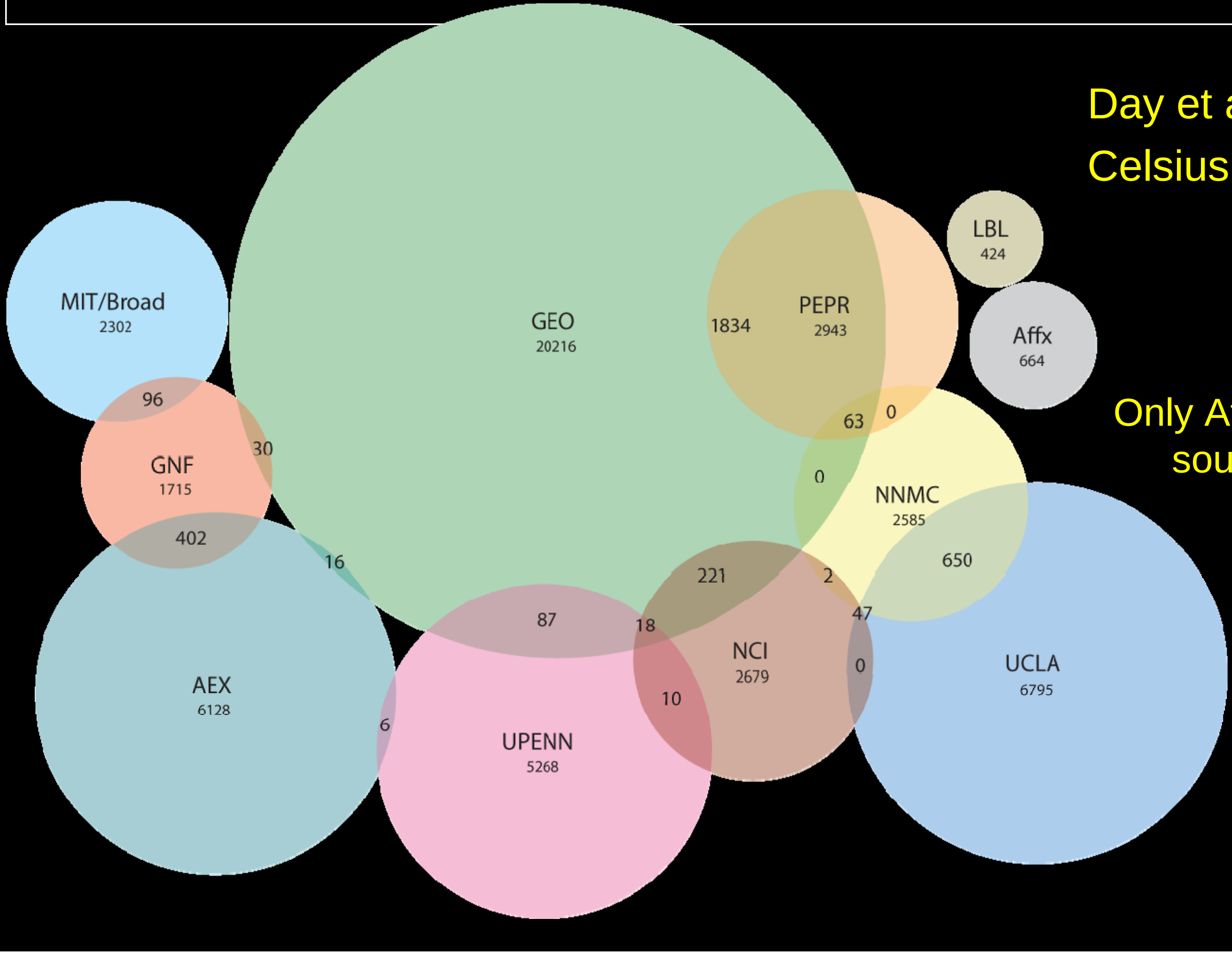
Why?

- Explore a large amount of expression or other data
- Get experiment-wide look at interesting subset of data
- Visually identify patterns for further analysis
- Order genes and/or experiments in a sensible way
- Split genes and/or experiments into a predefined number of groups

Why not?

- Clustering is not a substitute for rigorous statistics
- Clustering cannot identify
  - differentially expressed genes
  - profiles that are correlated with a reference profile
- Any data – even noise – can be clustered
- Clustering is not an essential step for most analyses

Where to get the data?



Day et al., 2007

Celsius project

Only Affymetrix sources shown

Types of data

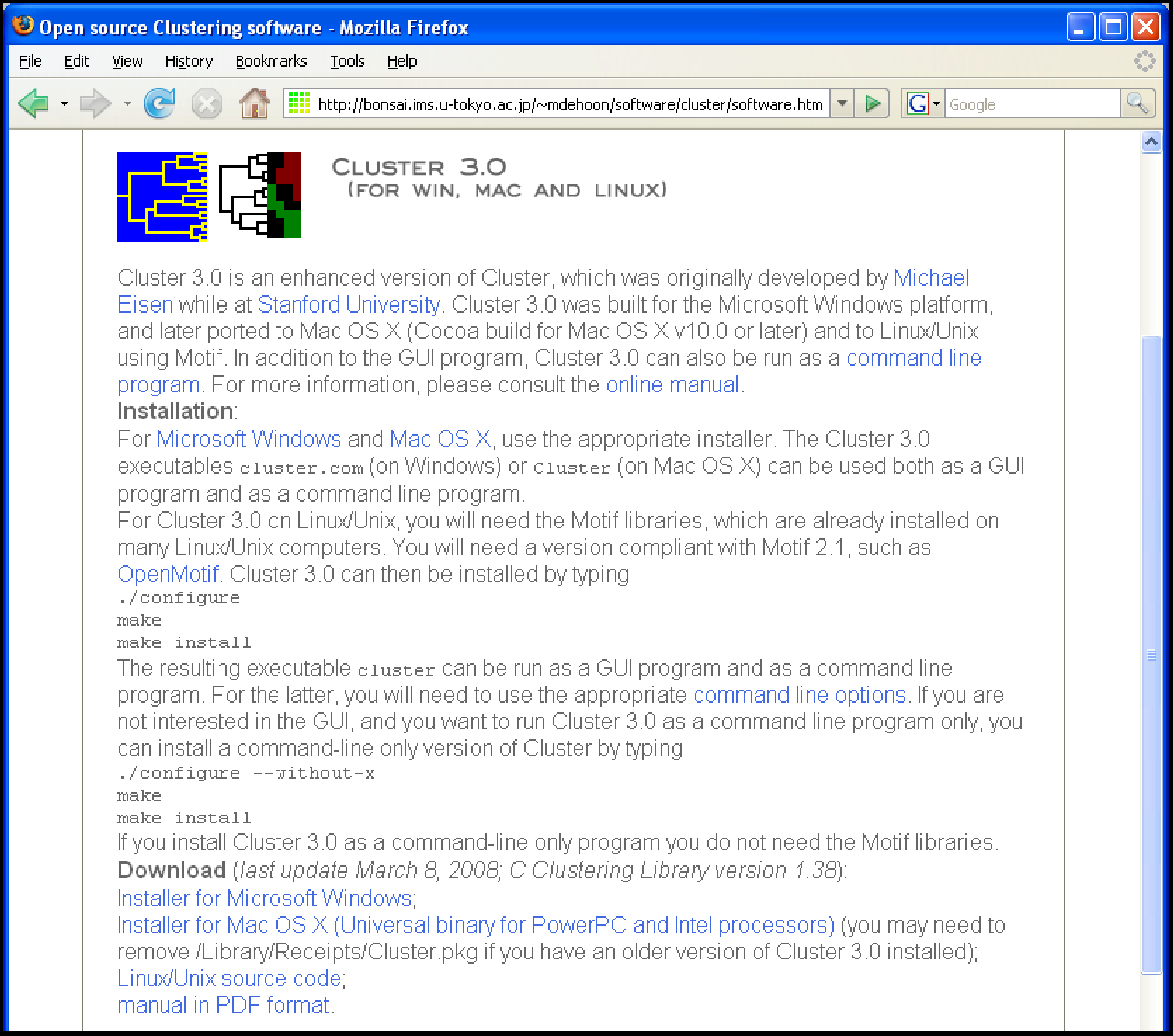
- Single-color arrays (mainly Affymetrix)
  - Data reported as expression values
  - Raw values or log2-transformed values (RMA; GCRMA)
- Two-color arrays
  - Data reported as expression ratios
  - Raw ratios or log2-transformed ratios

Clustering with Cluster 3.0

- Based on original clustering program by Michael Eisen
- Code updated by Michiel de Hoon
- Runs on Windows, Mac, and Linux
- Free from

  http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm
- Hierarchical, k-means, SOMs
- Other option for large datasets:
  - XCluster, a command-line tool by Gavin Sherlock

Getting Cluster 3.0


Open source Clustering software - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm
Google

CLUSTER 3.0
(FOR WIN, MAC AND LINUX)

Cluster 3.0 is an enhanced version of Cluster, which was originally developed by Michael Eisen while at Stanford University. Cluster 3.0 was built for the Microsoft Windows platform, and later ported to Mac OS X (Cocoa build for Mac OS X v10.0 or later) and to Linux/Unix using Motif. In addition to the GUI program, Cluster 3.0 can also be run as a command line program. For more information, please consult the online manual.

Installation:

For Microsoft Windows and Mac OS X, use the appropriate installer. The Cluster 3.0 executables `cluster.com` (on Windows) or `cluster` (on Mac OS X) can be used both as a GUI program and as a command line program.

For Cluster 3.0 on Linux/Unix, you will need the Motif libraries, which are already installed on many Linux/Unix computers. You will need a version compliant with Motif 2.1, such as OpenMotif. Cluster 3.0 can then be installed by typing

```
./configure
make
make install
```

The resulting executable `cluster` can be run as a GUI program and as a command line program. For the latter, you will need to use the appropriate command line options. If you are not interested in the GUI, and you want to run Cluster 3.0 as a command line program only, you can install a command-line only version of Cluster by typing

```
./configure --without-x
make
make install
```

If you install Cluster 3.0 as a command-line only program you do not need the Motif libraries.

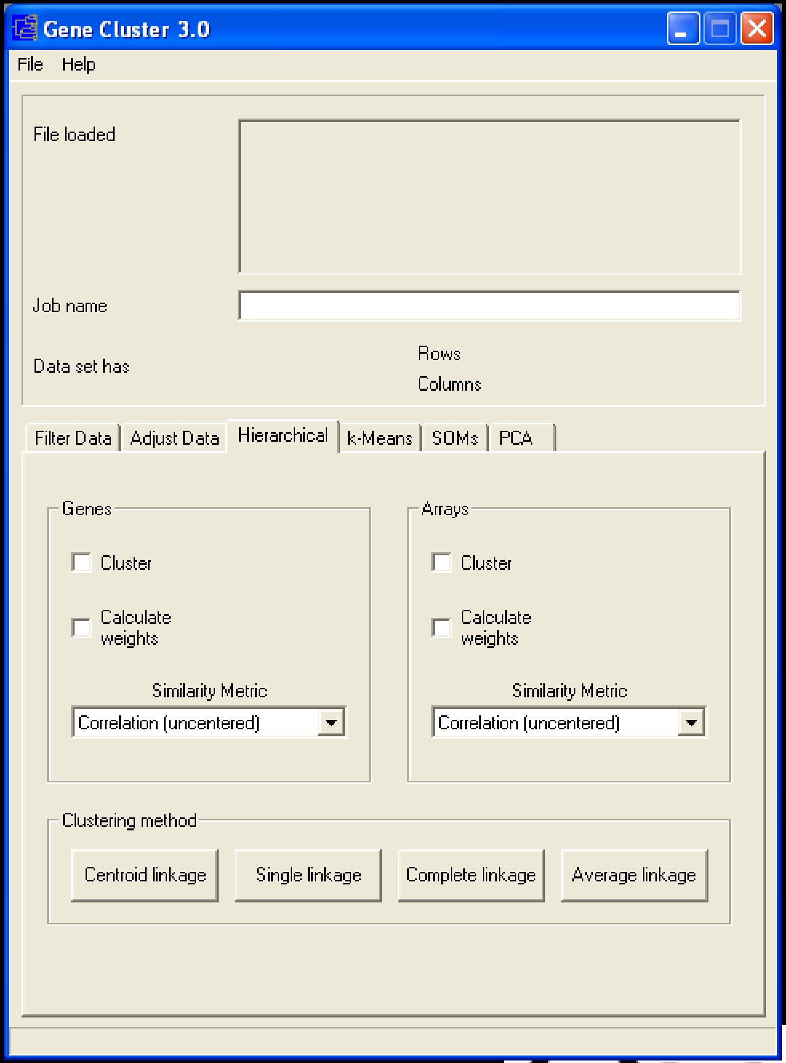
Download (*last update March 8, 2008; C Clustering Library version 1.38*):

Installer for Microsoft Windows;

Installer for Mac OS X (Universal binary for PowerPC and Intel processors) (you may need to remove /Library/Receipts/Cluster.pkg if you have an older version of Cluster 3.0 installed);

Linux/Unix source code;

manual in PDF format.



Genes:
□ Cluster
□ Calculate weights
Similarity Metric: Correlation (uncentered)

Arrays:
□ Cluster
□ Calculate weights
Similarity Metric: Correlation (uncentered)

BaRC
Bioinformatics and Research Computing

Cluster data import

- Minimal matrix (text, not Excel format)

Probe	Amygdala	Heart	Kidney	Liver	Lung
1000_at	0.85	0.19	-0.92	-0.32	-0.27
1009_at	0.02	0.44	0.32	0.53	-0.80
1014_at	-0.25	0.17	-5.83	-5.83	0.93
1030_s_at	-0.25		0.13	-2.09	0.21
1031_at	-0.35	-0.19	-0.22	-5.00	

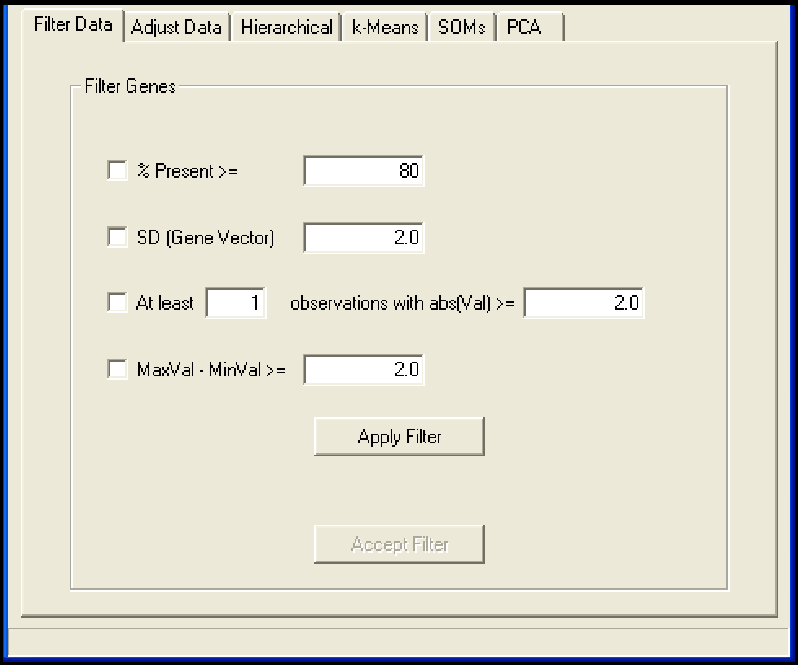
- Matrix with annotation and cluster weights

GeneID	NAME	GWEIGHT	Amygdala	Heart	Kidney	Liver	Lung
EWEIGHT			0	0	1	1	1
1000_at	MAPK3	1	0.85	0.19	-0.92	-0.32	-0.27
1009_at	HINT1	1	0.02	0.44	0.32	0.53	-0.80
1014_at	POLG	1	-0.25	0.17	-5.83	-5.83	0.93
1030_s_at	TOP1	0	-0.25		0.13	-2.09	0.21
1031_at	SRPK1	0	-0.35	-0.19	-0.22	-5.00	

BaRC
Bioinformatics and Research Computing

Data filtering

- Why filter?
  - Noise (unexpressed or uninteresting genes) can hide signal
  - A complete dataset is too much to visually process
- What are you looking for?
  - Differentially expressed genes
  - Most variable genes
  - Most interesting profile (expression pattern)
- Select list of genes of interest
- Select set of genes with GO annotation of interest
- Do in spreadsheet or Cluster ("Filter Data" tab)



Filter Genes

☐ % Present >= 80

☐ SD (Gene Vector) 2.0

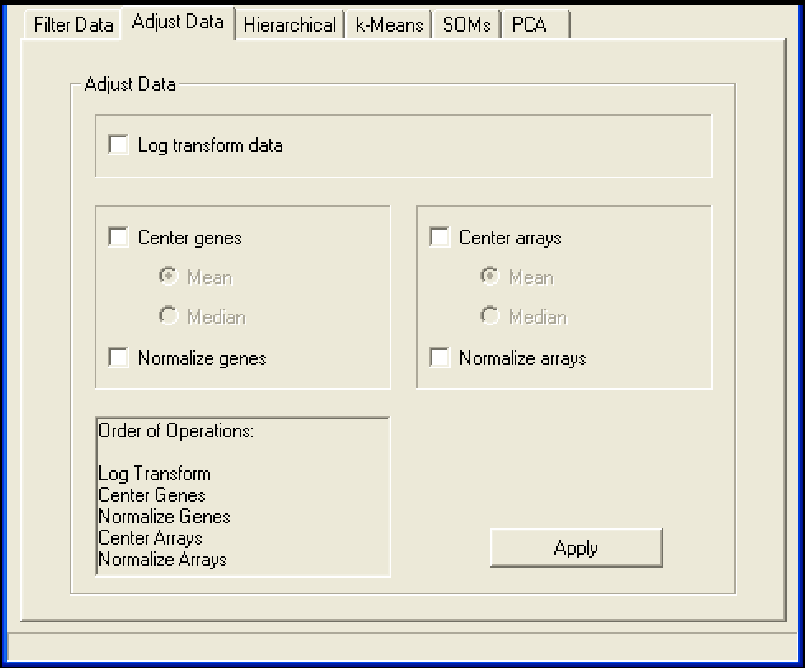
☐ At least 1 observations with abs(Val) >= 2.0

☐ MaxVal - MinVal >= 2.0

BaRC
Bioinformatics and Research Computing

Transforming data

- Do in spreadsheet or Cluster (“Adjust Data” tab)
- Common methods
  - Log-transformation
  - Converting values into ratios
  - Centering:
    - value – mean (row or column)
    - value – median (row or column)
  - Many normalization methods (from elsewhere)



Adjust Data

□ Log transform data

□ Center genes
● Mean
○ Median
□ Normalize genes

□ Center arrays
● Mean
○ Median
□ Normalize arrays

Order of Operations:

Log Transform
Center Genes
Normalize Genes
Center Arrays
Normalize Arrays

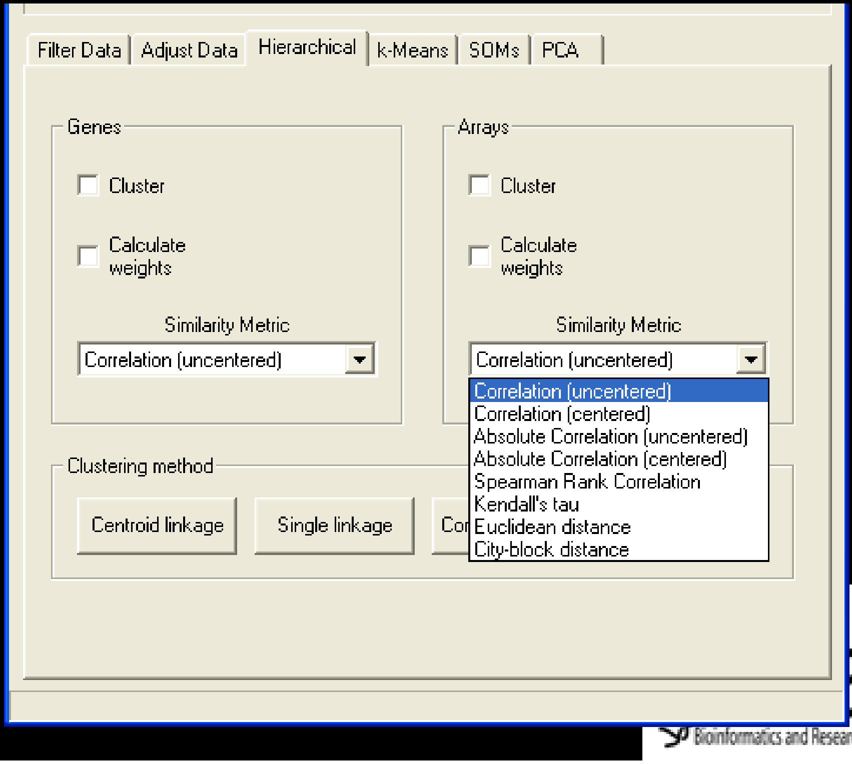
BaRC
Bioinformatics and Research Computing

Clustering goals and caveats

- Potential goal: organize a set of data to show relationships between data elements
- With microarray analysis: genes and/or chips
- Most data does not inherently exist in clusters
- Most effective with optimal quantity of data
- Interpretation of data in obvious clusters: is it filtered?
- Clustering vs segmenting

Hierarchical clustering

- Agglomerative, unsupervised analysis
- Steps
  1. Create an all vs. all distance matrix
  2. Fuse closest objects
  3. Compare fused object to all others
  4. Repeat steps 2-3 until one inclusive cluster is created
- Can be performed on genes and/or arrays
- Efficiency = $O(n^2m)$
- Need to select:
  – Similarity Metric
  – Clustering method



Genes

□ Cluster

□ Calculate weights

Similarity Metric: Correlation (uncentered)

Arrays

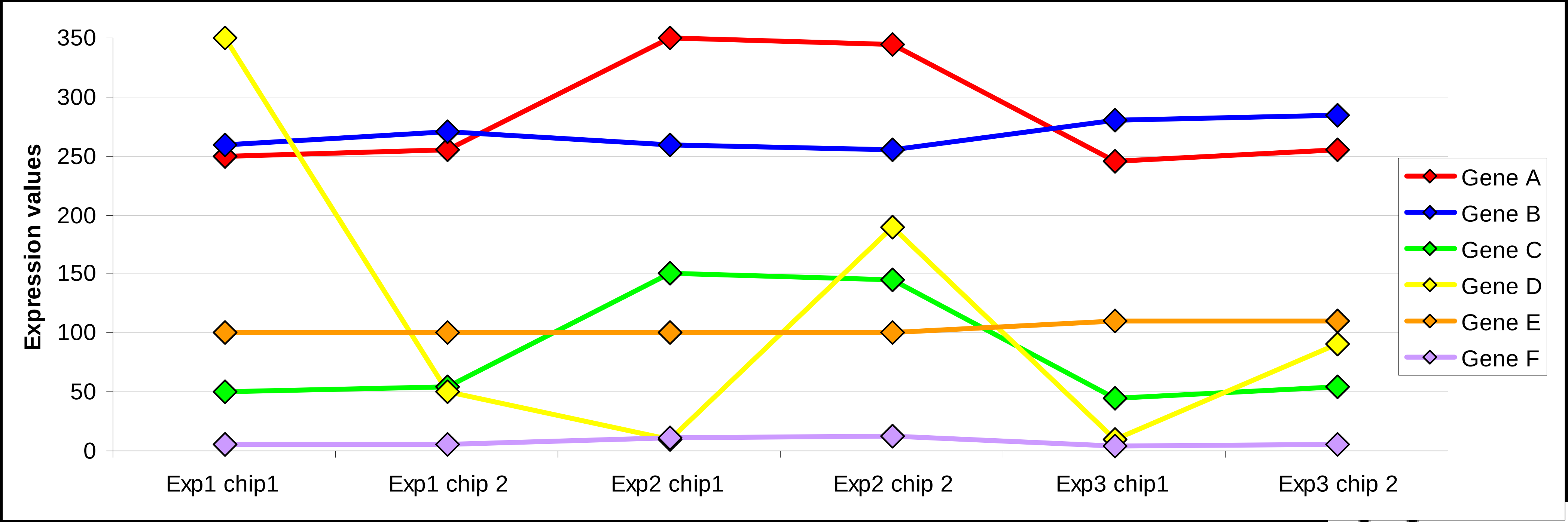
□ Cluster

□ Calculate weights

Similarity Metric: Correlation (uncentered)

Measuring similarity between profiles

- Similarity (distance) metric is an important choice when comparing genes and/or experiments
- What are you trying to group?



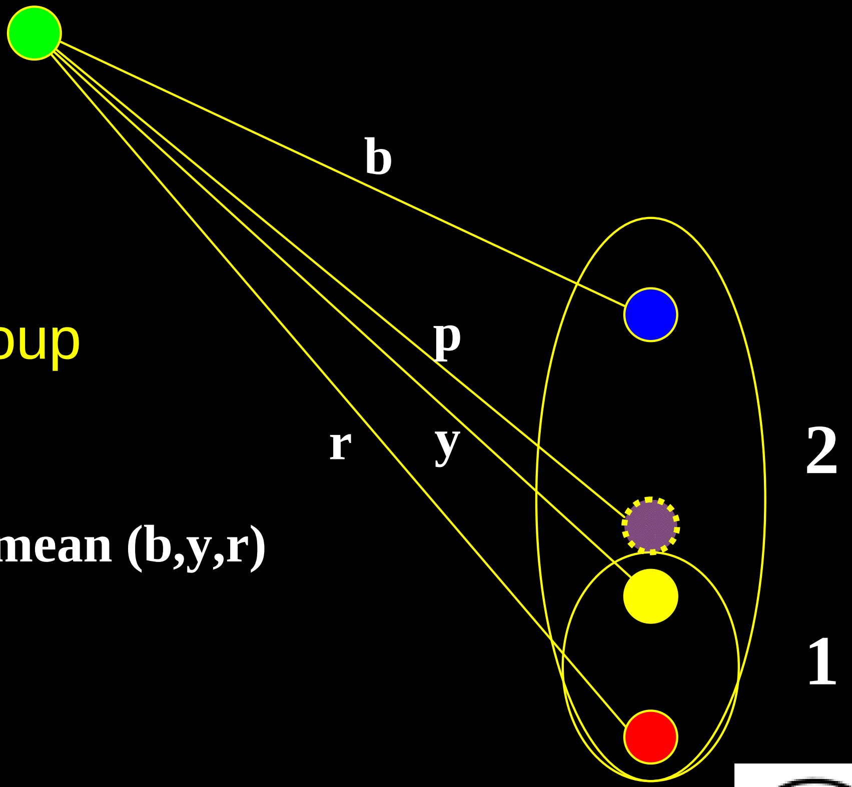
Common similarity metrics

- Pearson correlation
  - Measures the difference in the shape of two curves
  - modifications:
    - uncentered correlation: for offset profiles, coefficient < 1
    - absolute correlation: opposite profiles cluster together
- Euclidean distance: multidimensional Pythagorean Theorem
  - Measures the distance between two curves
- Nonparametric or Rank Correlation
  - Similar to the Pearson correlation but data values are replaced with their ranks
  - Ex: Spearman Rank, Kendall’s Tau
  - Good idea if distribution of data is not normal
  - More robust (against outliers) than other methods

Clustering methods

How can groups of objects be represented?
How is distance measured to a cluster of objects?

- Single linkage (b)
  - minimum distance
- Complete linkage (r)
  - maximum distance
- Centroid linkage (p)
  - distance to "centroid" of group
- Average linkage (x)
  - average distance



x = mean (b,y,r)

Weighting?

- GWEIGHT, EWEIGHT

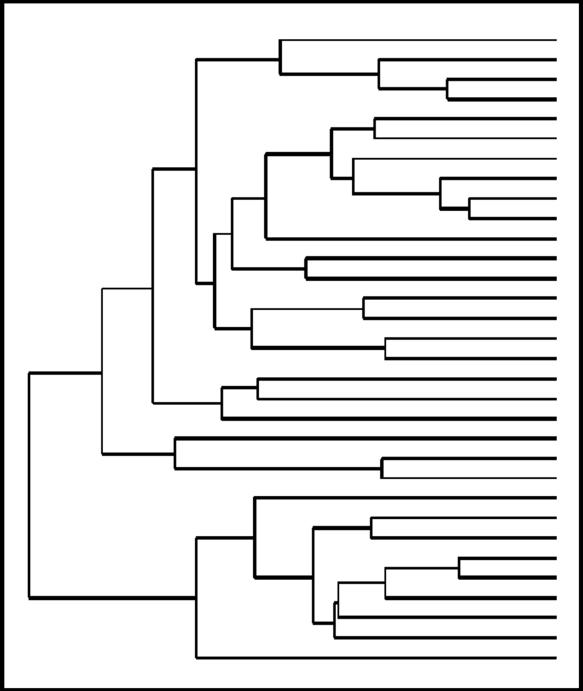
Cluster data output

- For hierarchical clustering by genes and arrays, 3 output files are created:
  - .cdt (“clustered data table”)
  - .gtr (“gene tree”)
  - .atr (“array tree”)
- All are tab-delimited text and can be opened as a spreadsheet
- Create your own ‘cdt’ file and bypass Cluster 3.0:
  - Tab-delimited text
  - First 2 columns are gene identifiers

Gene ID	Symbol	Amygdala	Heart	Kidney	Liver	Lung
1000_at	MAPK3	0.85	0.19	-0.92	-0.32	-0.27
1009_at	HINT1	0.02	0.44	0.32	0.53	-0.80

Representation of clustered data

- Hierarchical clustering produces a dendrogram(s) showing relationships between objects
- Order of leaves: 2^{N-1} choices
- How can objects be partitioned into groups?
  - k-means clustering
  - self-organizing maps
  - How many clusters (k)?



- Are the data really hierarchical?
- Original distance matrix may be informative

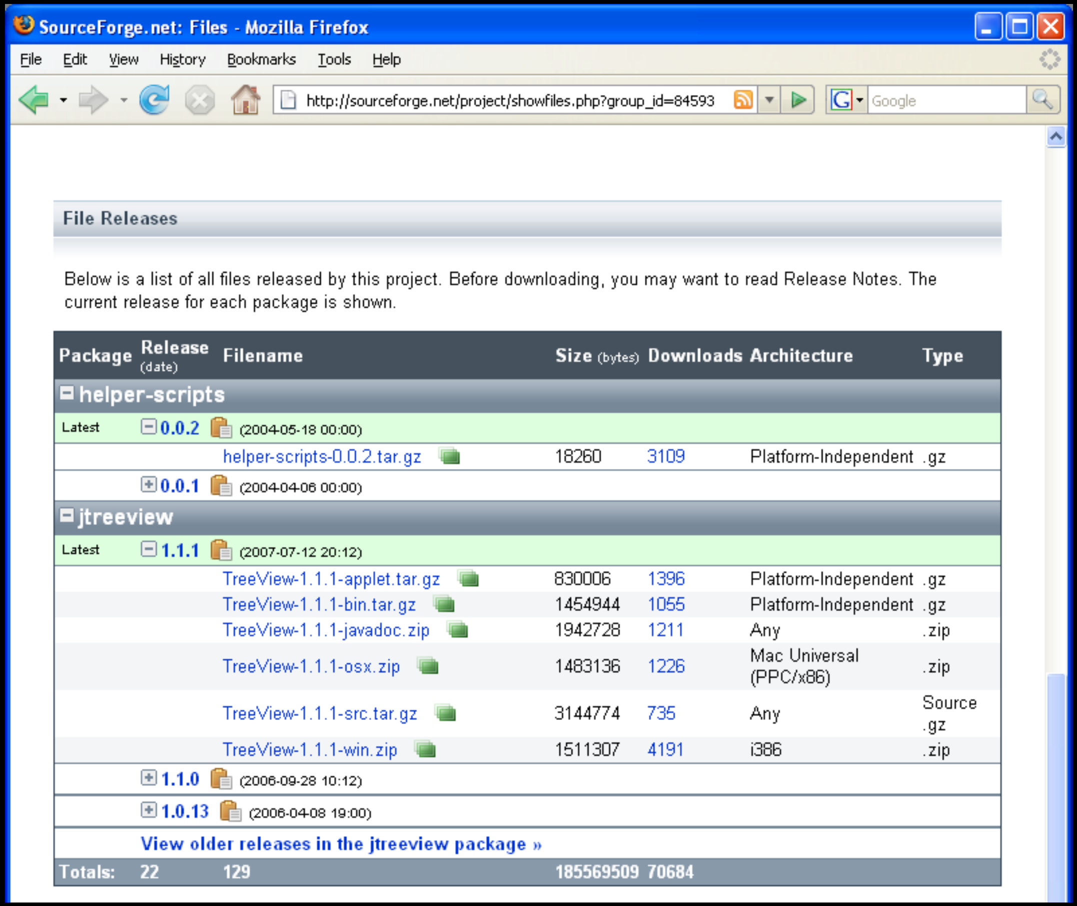
Visualizing clustered data with Java TreeView

- Based on original clustering program by Michael Eisen
- Code updated by Alok Saldanha
- Runs on Windows, Mac, and Linux
- Free from

  http://sourceforge.net/project/showfiles.php?group_id=84593

Getting Java TreeView

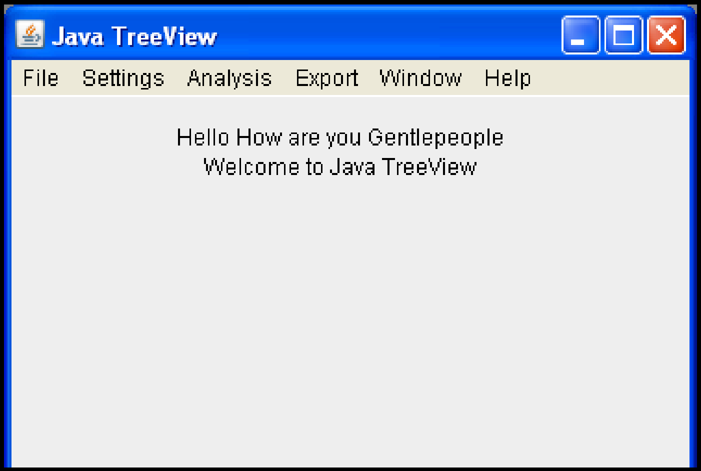
- http://sourceforge.net/project/showfiles.php?group_id=84593


SourceForge.net: Files - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://sourceforge.net/project/showfiles.php?group_id=84593
Google

File Releases

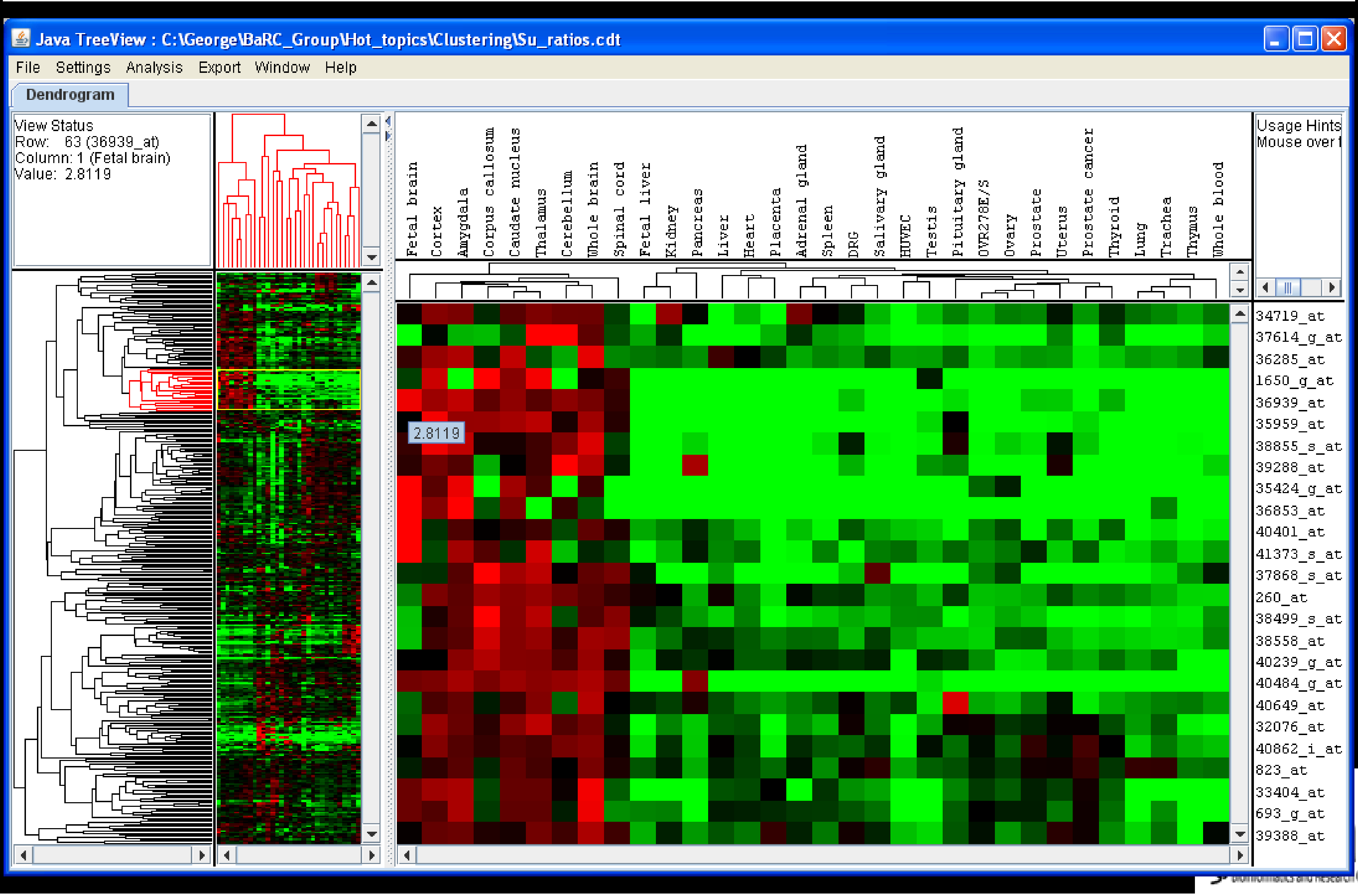
Below is a list of all files released by this project. Before downloading, you may want to read Release Notes. The current release for each package is shown.

Package	Release (date)	Filename	Size (bytes)	Downloads	Architecture	Type
⊟ helper-scripts						
Latest	⊟ 0.0.2 (2004-05-18 00:00)					
		helper-scripts-0.0.2.tar.gz	18260	3109	Platform-Independent	.gz
	⊞ 0.0.1 (2004-04-06 00:00)					
⊟ jtreeview						
Latest	⊟ 1.1.1 (2007-07-12 20:12)					
		TreeView-1.1.1-applet.tar.gz	830006	1396	Platform-Independent	.gz
		TreeView-1.1.1-bin.tar.gz	1454944	1055	Platform-Independent	.gz
		TreeView-1.1.1-javadoc.zip	1942728	1211	Any	.zip
		TreeView-1.1.1-osx.zip	1483136	1226	Mac Universal (PPC/x86)	.zip
		TreeView-1.1.1-src.tar.gz	3144774	735	Any	Source .gz
		TreeView-1.1.1-win.zip	1511307	4191	i386	.zip
	⊞ 1.1.0 (2006-09-28 10:12)					
	⊞ 1.0.13 (2006-04-08 19:00)					
	View older releases in the jtreeview package »					
Totals:	22	129	185569509	70684		



Hello How are you Gentlepeople
Welcome to Java TreeView

Java TreeView main view

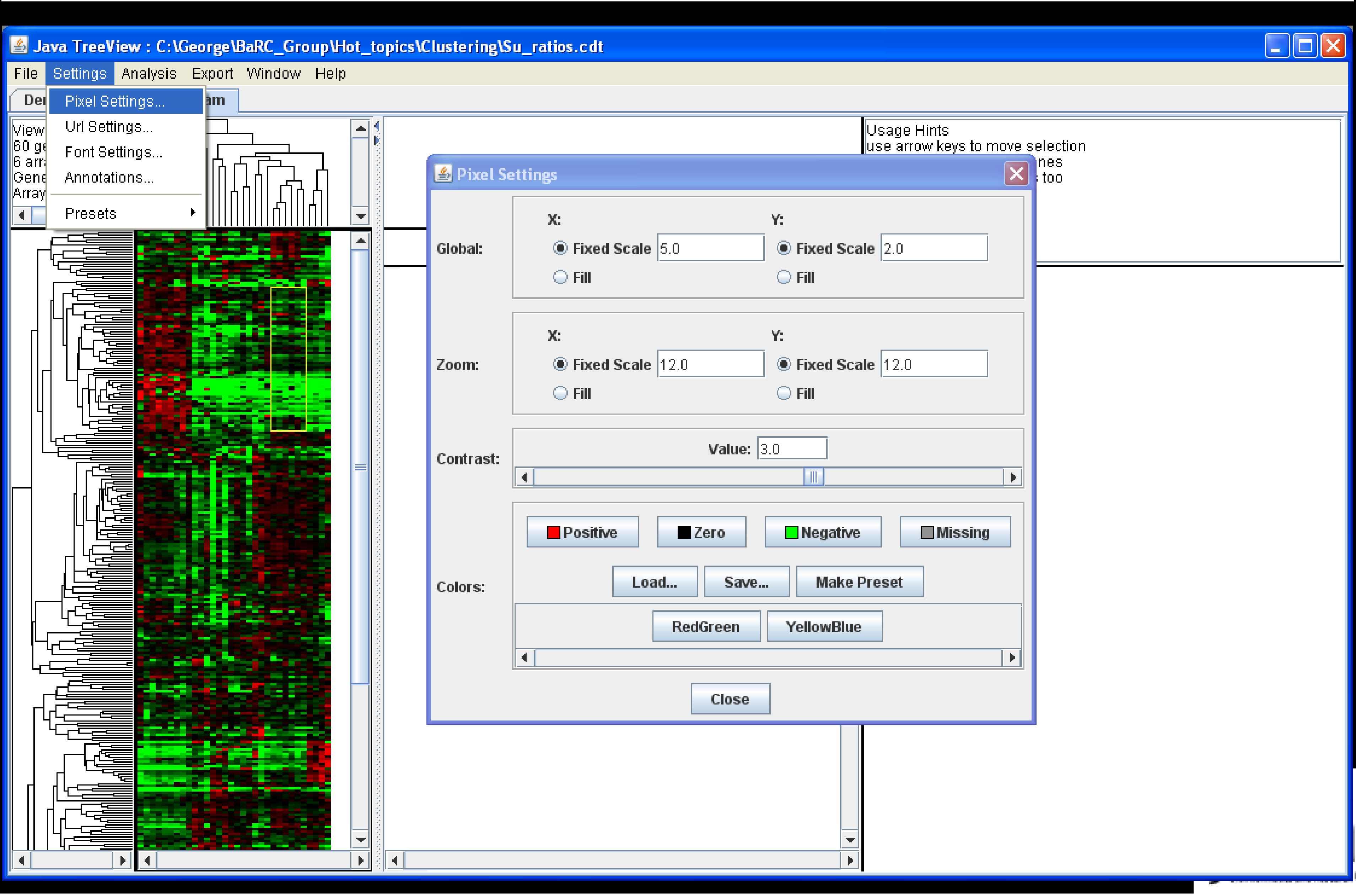


Java TreeView : C:\George\BaRC_Group\Hot_topics\Clustering\Su_ratios.cdt

View Status
Row: 63 (36939_at)
Column: 1 (Fetal brain)
Value: 2.8119

Usage Hints
Mouse over t

Java TreeView: settings



Java TreeView : C:\George\BaRC_Group\Hot_topics\Clustering\Su_ratios.cdt

View
60 ge
6 arr
Gene
Array

Usage Hints
use arrow keys to move selection
nes
too

Pixel Settings

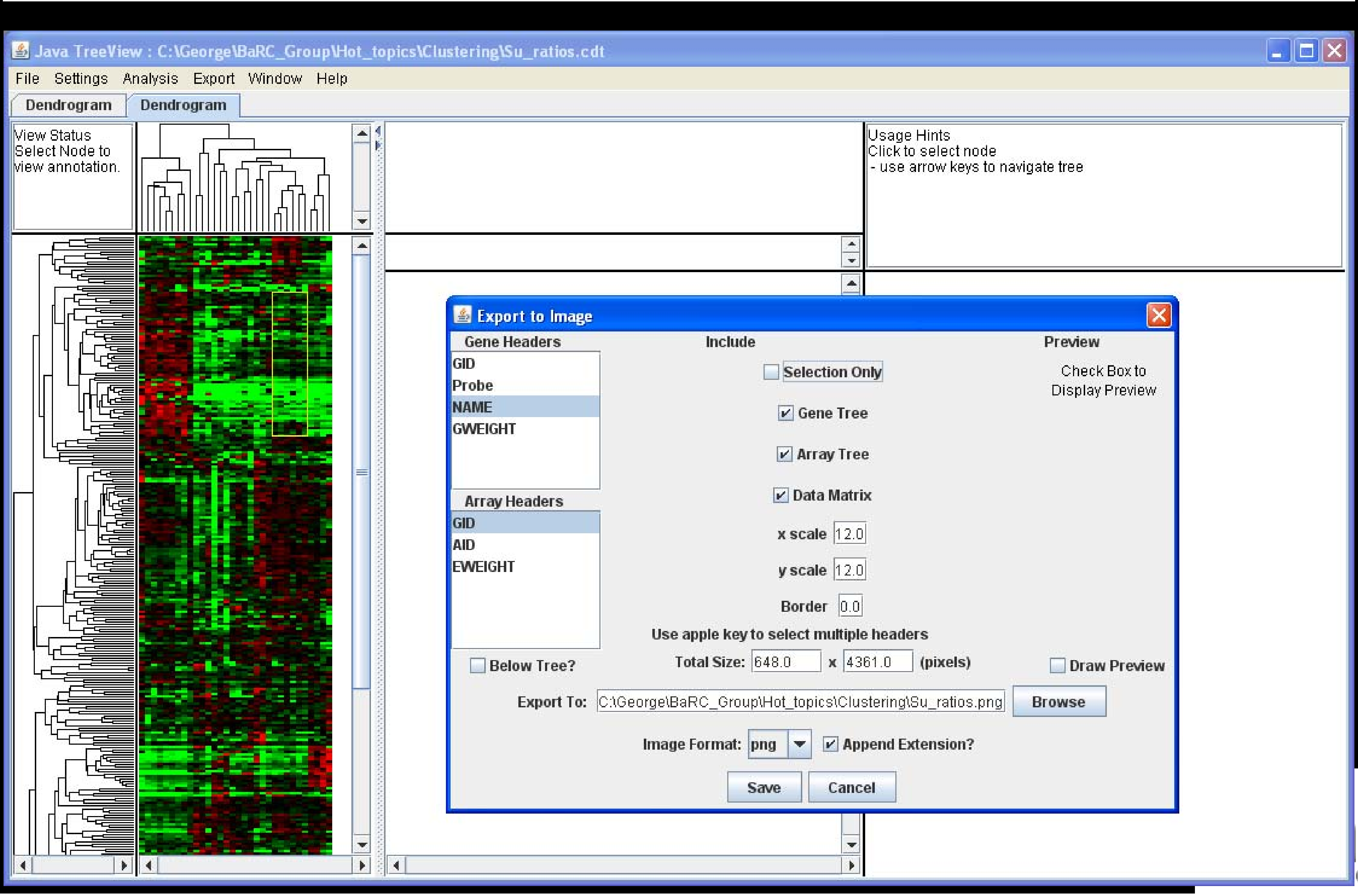
Global:
X: ● Fixed Scale 5.0 ○ Fill
Y: ● Fixed Scale 2.0 ○ Fill

Zoom:
X: ● Fixed Scale 12.0 ○ Fill
Y: ● Fixed Scale 12.0 ○ Fill

Contrast:
Value: 3.0

Colors:

Java TreeView: exporting images


Java TreeView : C:\George\BaRC_Group\Hot_topics\Clustering\Su_ratios.cdt
File Settings Analysis Export Window Help
Dendrogram
Dendrogram
Save
Cancel
Browse

View Status
Select Node to view annotation.

Usage Hints
Click to select node
- use arrow keys to navigate tree

Export to Image

Gene Headers: GID, Probe, NAME, GWEIGHT

Array Headers: GID, AID, EWEIGHT

Include

□ Selection Only

☑ Gene Tree

☑ Array Tree

☑ Data Matrix

x scale 12.0

y scale 12.0

Border 0.0

Use apple key to select multiple headers

□ Below Tree?

Total Size: 648.0 x 4361.0 (pixels)

Preview

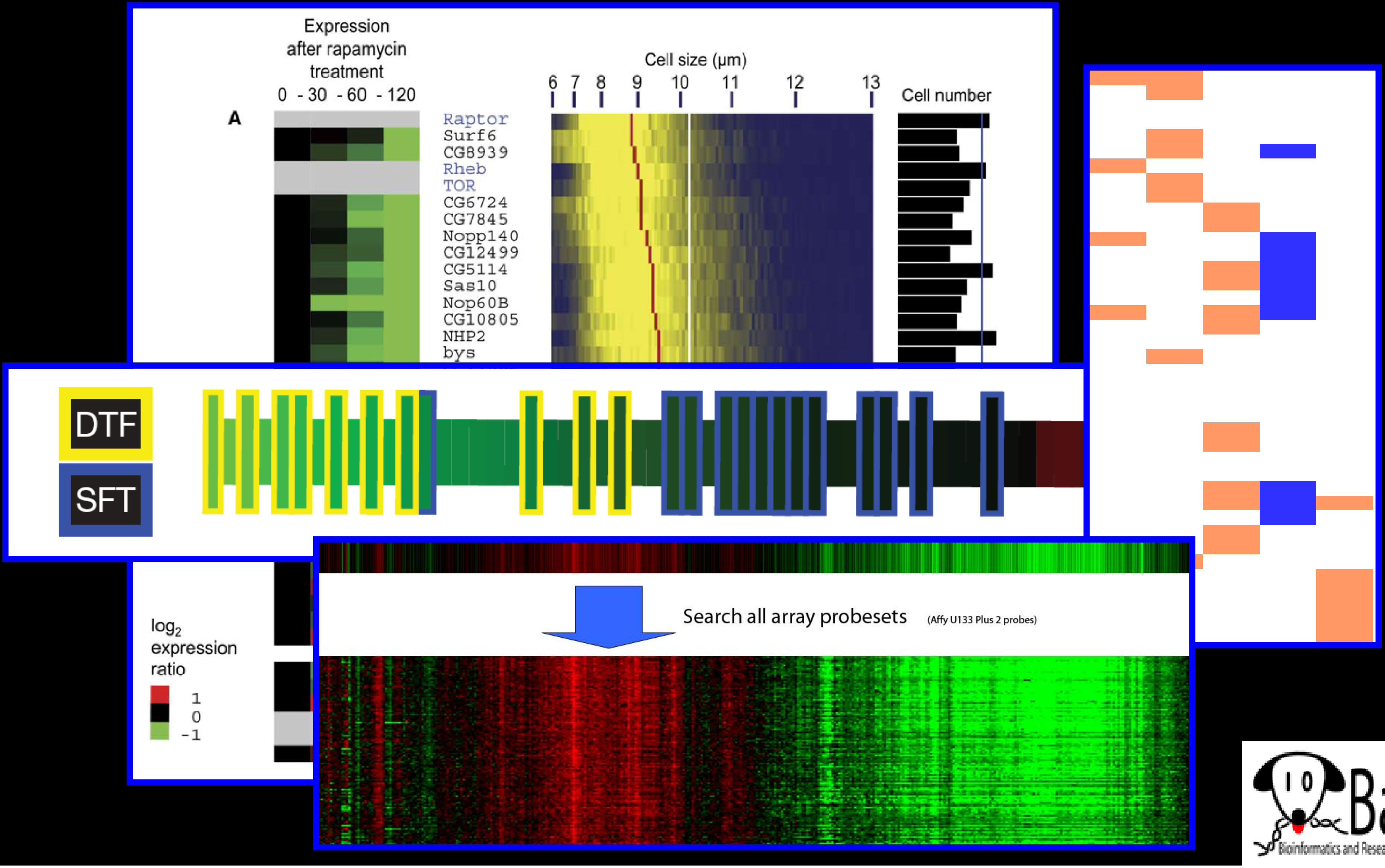
Check Box to Display Preview

□ Draw Preview

Export To: C:\George\BaRC_Group\Hot_topics\Clustering\Su_ratios.png

Image Format: png ☑ Append Extension?

Displaying other types of data



Demo ?