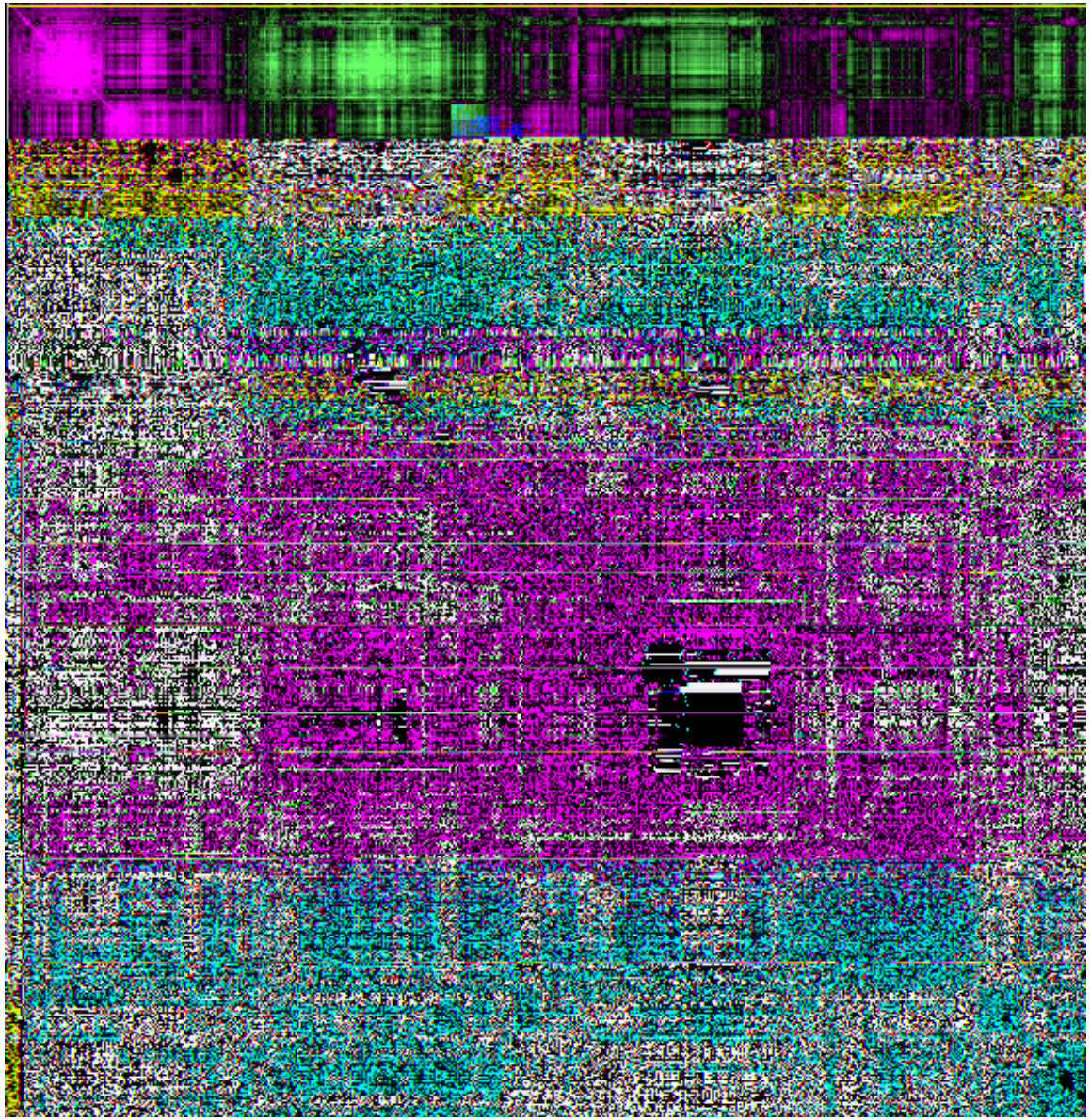
Practical Bioinformatics

Mark Voorhies

5/24/2013

Clustering exercises { Visualizing the distance matrix



Scripting Cluster

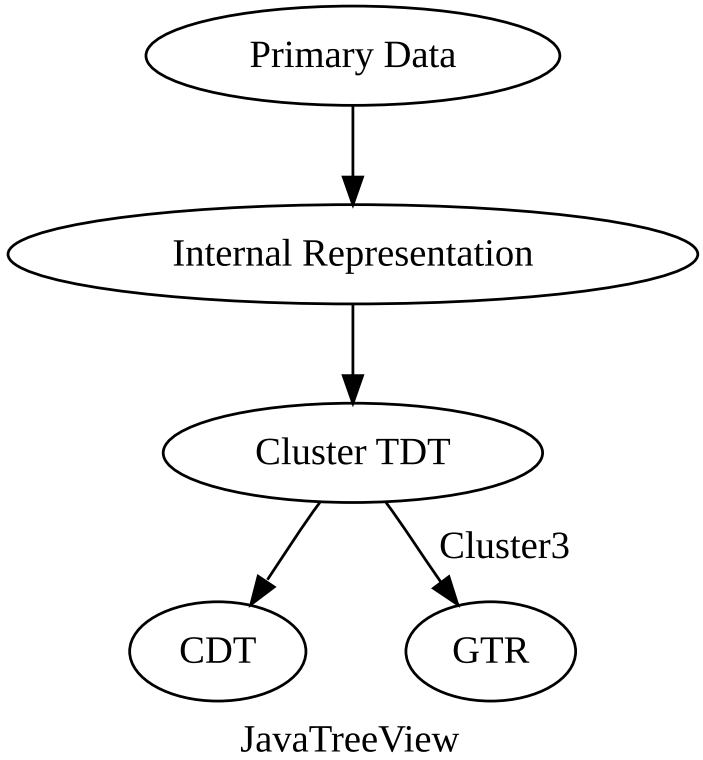
- Running Cluster3 from the command line
  - /Applications/Cluster.app/Contents/MacOS/Cluster
  - /Program Files/Stanford University/Cluster3/Cluster.com
- Command-line programs are like functions
- \man program" is like \help(function)"
- Use the `subprocess` module to run command-line programs from within Python.

USAGE: cluster [options]

-f filename	File loading
-u jobname	Allows you to specify a different name for the output files (default is derived from the input file name)
-g [0..8]	Specifies the distance measure for gene clustering 0: No gene clustering 1: Uncentered correlation 2: Pearson correlation 3: Uncentered correlation, absolute value 4: Pearson correlation, absolute value 5: Spearman's rank correlation 6: Kendall's tau 7: Euclidean distance 8: City-block distance (default: 0)
-m [msca]	Specifies which hierarchical clustering method to use m: Pairwise complete-linkage s: Pairwise single-linkage c: Pairwise centroid-linkage a: Pairwise average-linkage (default: m)

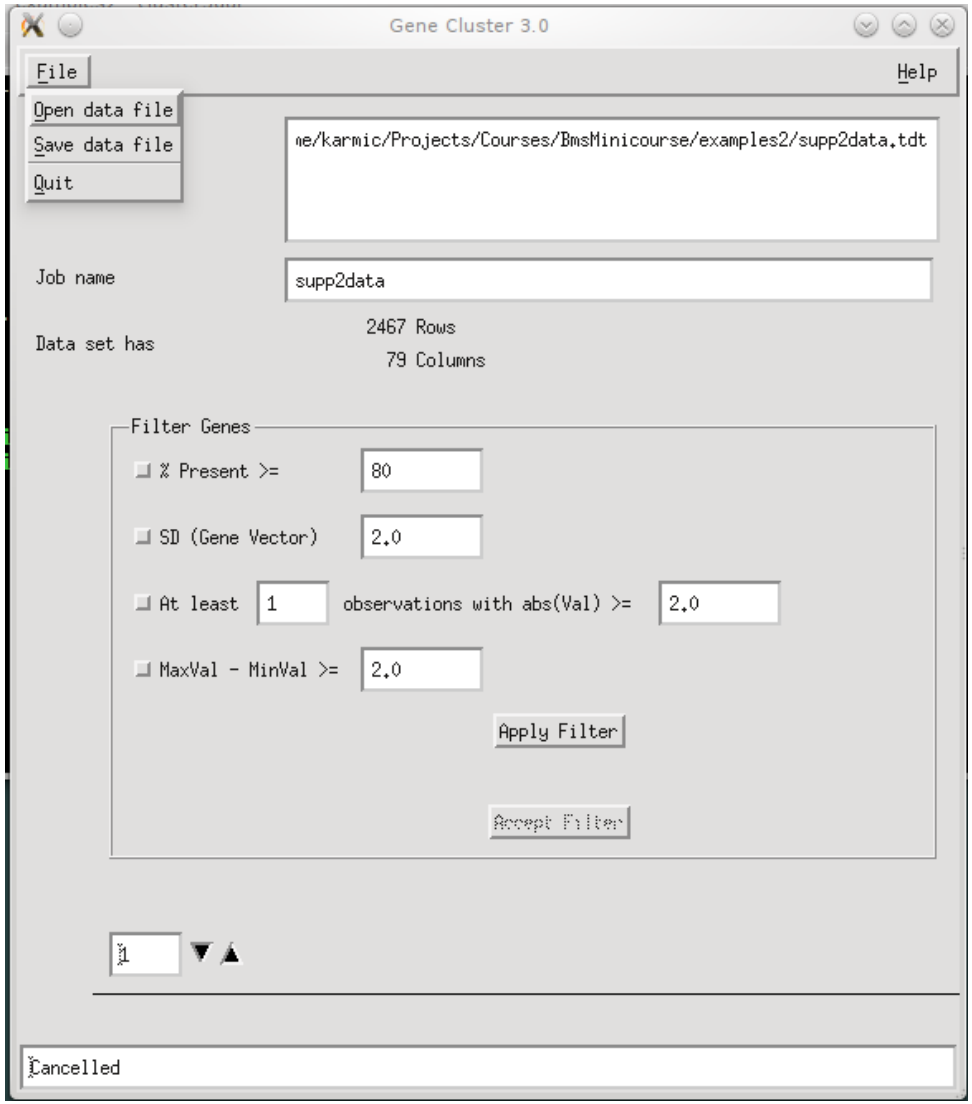
Scripting the Protocol

```
from subprocess import check_call
check_call(
    # Which program to run
    ("cluster",
     # Input file
     "-f","supp2data.tdt",
     # Output prefix
     "-u","supp2data.Uncentered.Complete",
     # Clustering method: complete linkage
     "-m","m",
     # Distance function: uncentered Pearson
     "-g","1"))
```

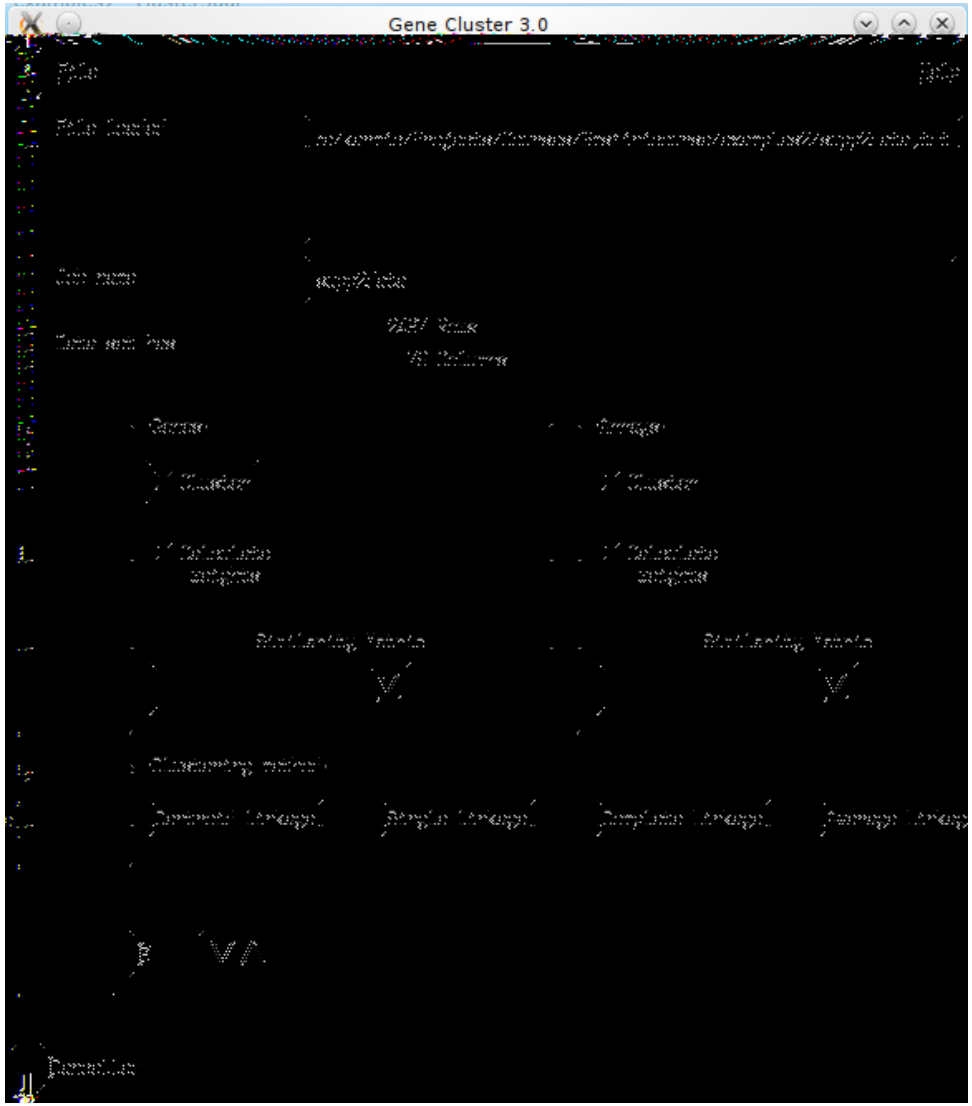


Using the Cluster3 GUI

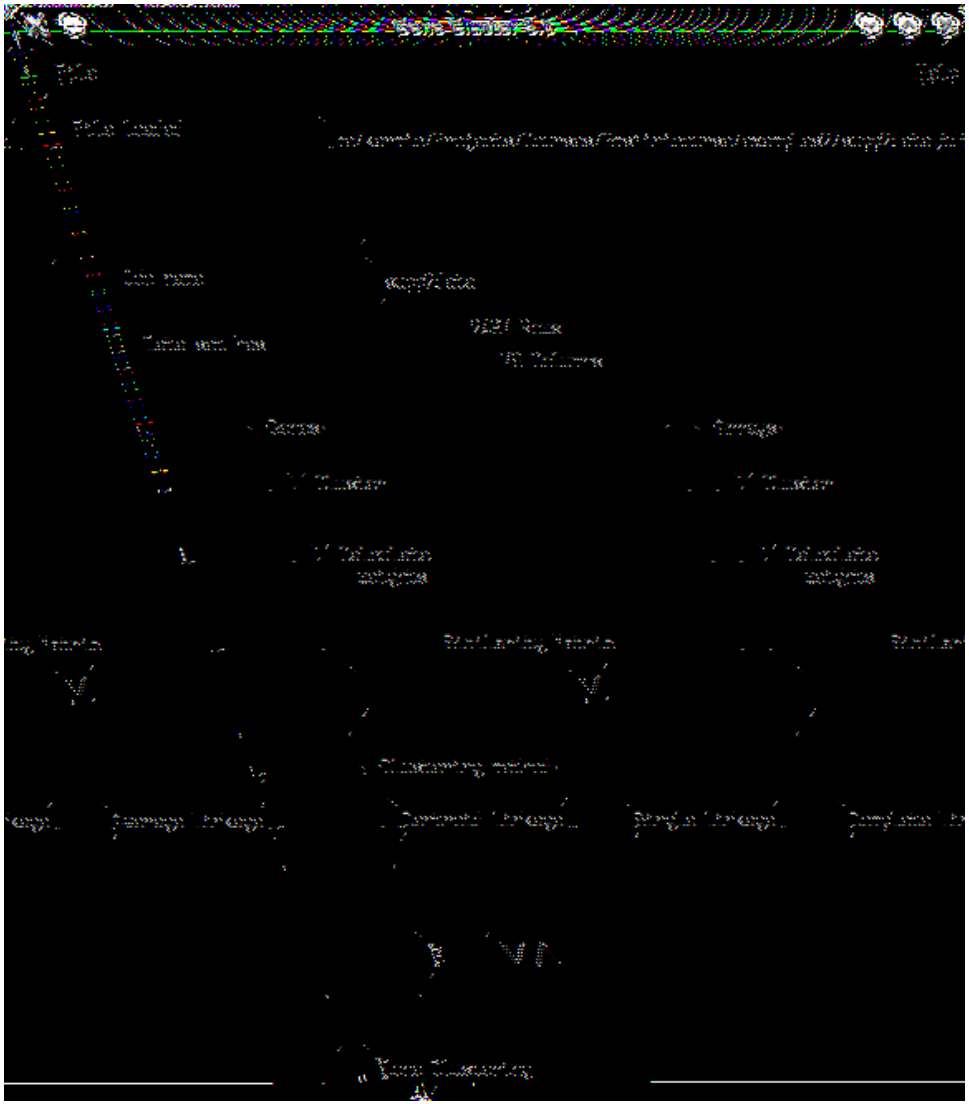
Load your data



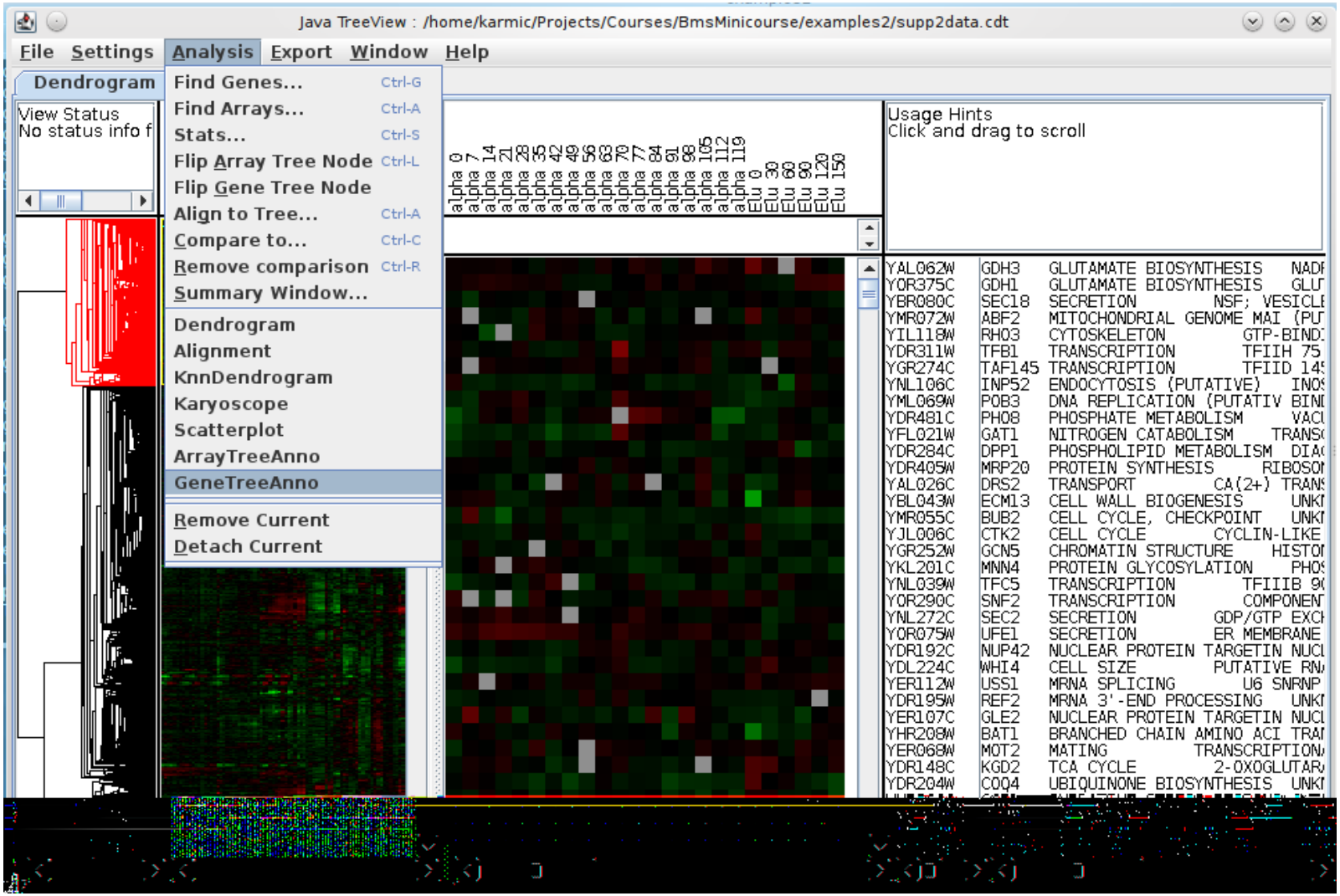
Choose distance function



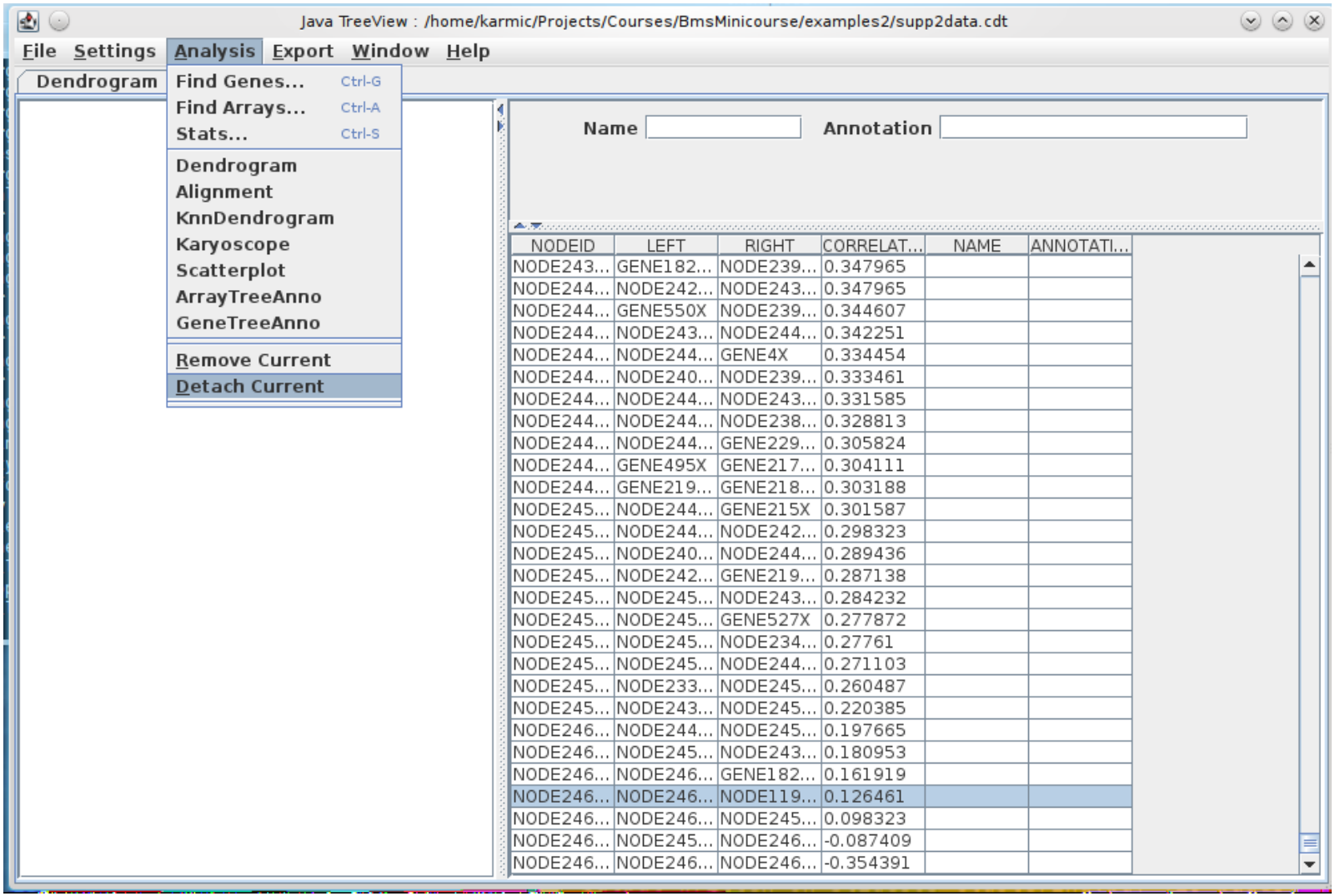
Choose linking method



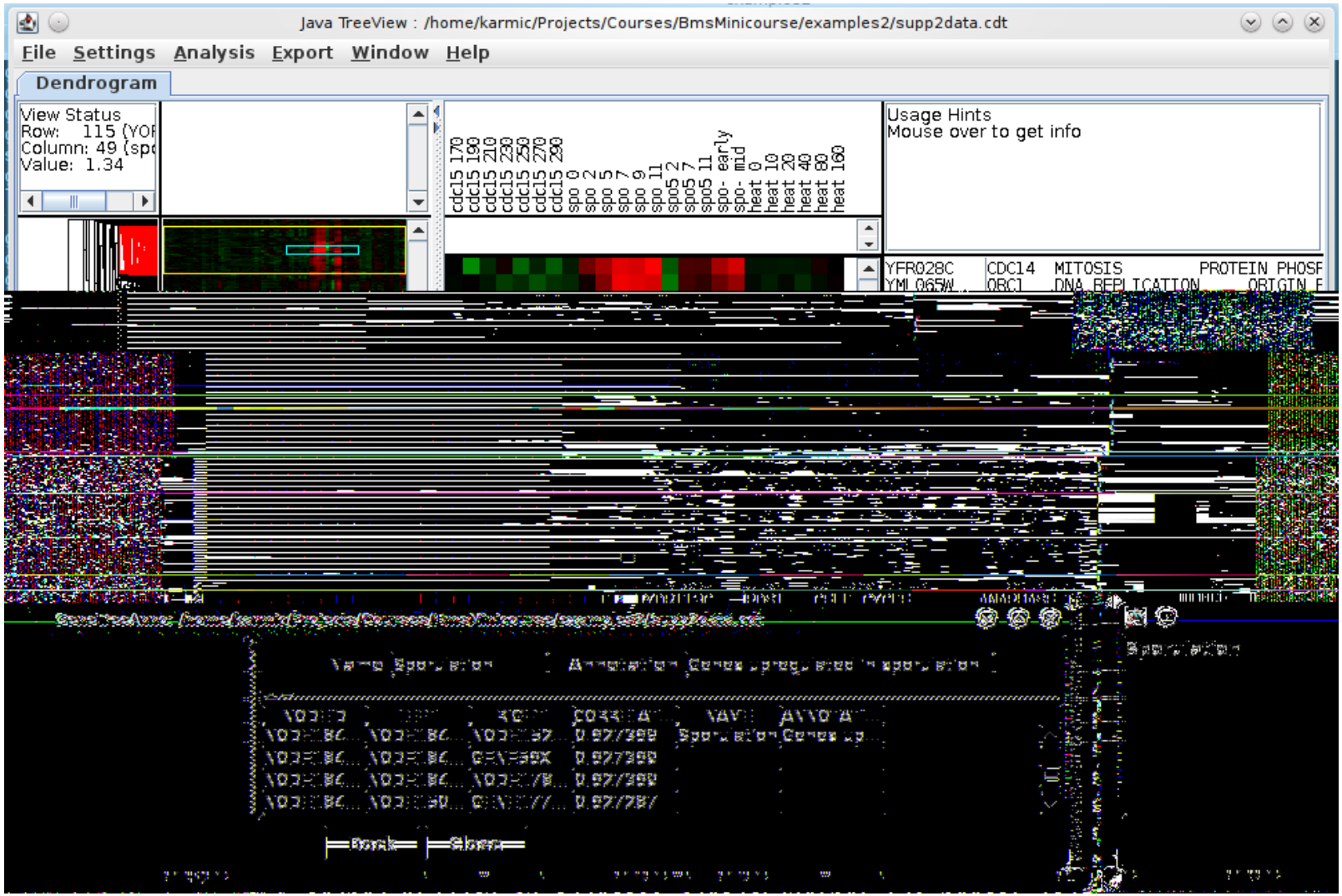
Activate and detach annotation window



Activate and detach annotation window



Activate and detach annotation window



Clustering exercises { Negative controls

Write functions to reproduce the shu ing controls in gure 3 of the Eisen paper (removing correlations among genes and/or arrays).

Clustering exercises { Negative controls

Write functions to reproduce the shu ing controls in gure 3 of the Eisen paper (removing correlations among genes and/or arrays).

```
def shuffleRows(self, seed = None):
    """Permute ratio values within rows."""
    import random
    if(seed != None):
        random.seed(seed)
    for i in self.num:
        random.shuffle(i)
```

Clustering exercises { Scripting Cluster

Modify the clustering protocol script to run Cluster3 multiple times on the same input, varying distance metric and/or clustering method. Be sure to give the output files distinct names.

Clustering exercises { Scripting Cluster

Modify the clustering protocol script to run Cluster3 multiple times on the same input, varying distance metric and/or clustering method. Be sure to give the output files distinct names.

```
metrics = ("None",
           "Uncentered",
           "Pearson",
           "UncenteredAbs",
           "PearsonAbs",
           "Spearman",
           "Kendall",
           "Euclidean",
           "City")
linkage = (("Complete","m"),
           ("Single","s"),
           ("Centroid","c"),
           ("Average","a"))

# Loop over all 32 possible methods
print "Starting hierarchical clustering runs..."
from subprocess import check_call
for metric in xrange(1,len(metrics)):
    print "    ",metrics[metric],"..."
    for (linkname, link) in linkage:
        print "        ",linkname
        check_call(("cluster","-f","shuffled.txt",
                    "-u",".".join(("shuffled",
                                   metrics[metric],
                                   linkname)),
                    "-m",link,"-g",str(metric)))
```