

# Cluster Analysis of SAGE Data

What you need...

- SAGE Data in Cluster 3.0 format
- Obtain the CLUSTER and TREEVIEW programs
  - <http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv>
- Read the Cluster 3.0 manual to familiarize yourself with cluster analysis

# Step 1 - Get the SAGE data from GMOD

[Home](#) | [Gbrowse](#) | [Gblast](#) | [Assembly](#) | [ORFs](#) | [Domains](#) | [SAGE](#) | [Download](#) | [What's New](#) | [Getting Started](#) | [Login](#) |

Assembly Data: s\_mansoniest01

## SAGE Analysis

### Data Access

[Login](#) to access all SAGE libraries (SAGE Consortium members only). Data from all libraries will be publicly released upon publication.

### Select Libraries

- ☒ Adult male - single sex infection
- ☒ Adult female - single sex infection
- ☐ Adult male - bisex infection
- ☐ Adult female - bisex infection
- ☒ Sub-adult liver stage
- ☐ 3 hr NOS control
- ☐ 3 hr NOS exposed
- ☐ Miracidia
- ☐ 6d sporocysts (un-cond.)
- ☐ 6d sporocysts (cond.)
- ☐ 20d sporocysts (un-cond.)
- ☐ 20d sporocysts (cond.)

Start Search

### Minimum Tag Count

1

The Minimum Tag Count represents the minimum number of times a SAGE tag must be found in at least one of the selected libraries for inclusion in the analysis. Use this tool to filter out low frequency tags.

### Primary Tags Only

☐ True

Primary SAGE Tags are those generated by the most 3' [Nla III](#) restriction site on the theoretical transcript.

### R Value

4

greater than

The R-Value is the log-likelihood ratio statistic of [Stekel et al \(2000\)](#), which scores tags by their deviation from the null hypothesis of equal frequencies. Higher scores represent a greater deviation from the null hypothesis, while scores close to zero represent near constitutive expression.

Is Not Relevant

In

Adult male - single sex infection

Vs

All

### Regulation

This allows you to determine tags that are up or downregulated between libraries

### Perform Clustering

☐ True 10 Clusters

Analysis of SAGE data using this tool automatically includes clustering of gene expression profiles using log transformation, median centering, Pearson's correlation coefficient, and kmeans/median clustering.

### Shade Expression Levels Based on Median Centering

☐ True

Median centering forces shading of expression levels to ignore magnitude of expression, thus highlighting correlated patterns of expression among lowly and highly expressed genes in the search results.

### Sort By

Tag ID

descending

### View Data As

Cluster 3.0

For more complex analyses, output options include tab-delimited and the input formats for [Cluster 3.0](#), [wCLUTO](#), [TableView](#), and [IDEA](#).

Scientific enquiries should be sent to [nobody](#)

This database is hosted by us [HERE](#). Bug reports and technical problems should be reported to [nobody](#).

Select libraries for analysis

Select data filters

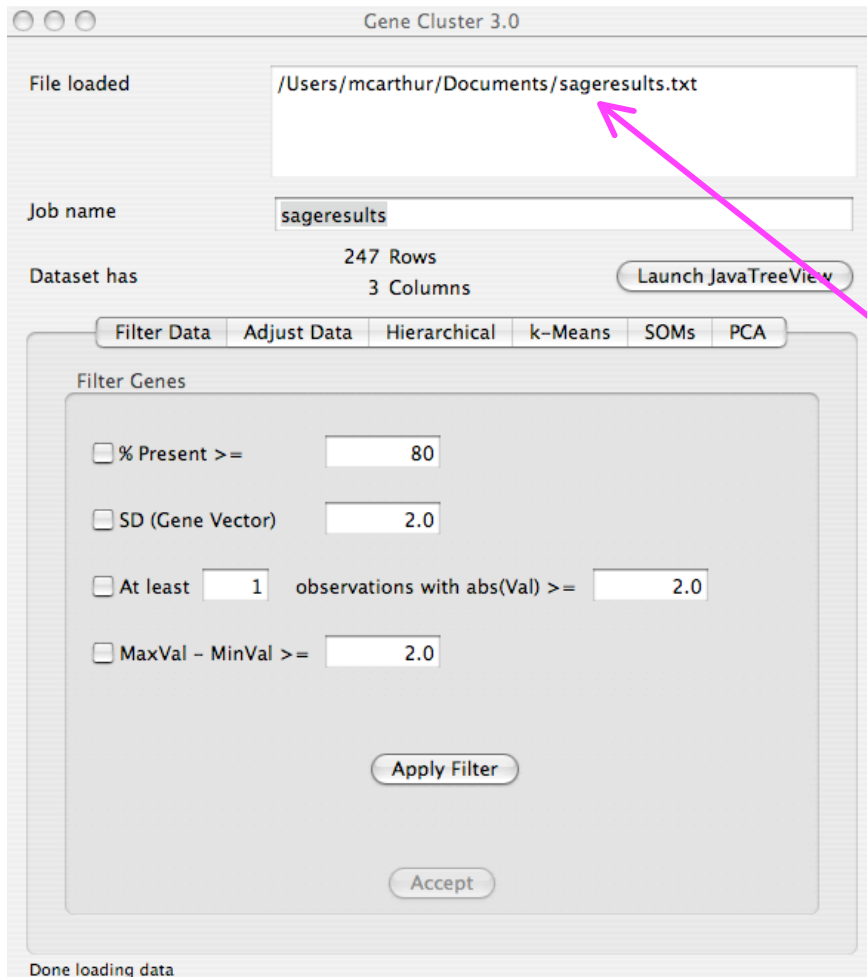
Save data in Cluster 3.0 format

# Step 1 - Get the SAGE data from GMOD

After you press submit, you should get a file called "sageresults.txt". This is the SAGE data in Cluster 3.0 format. You can rename the file if you like.

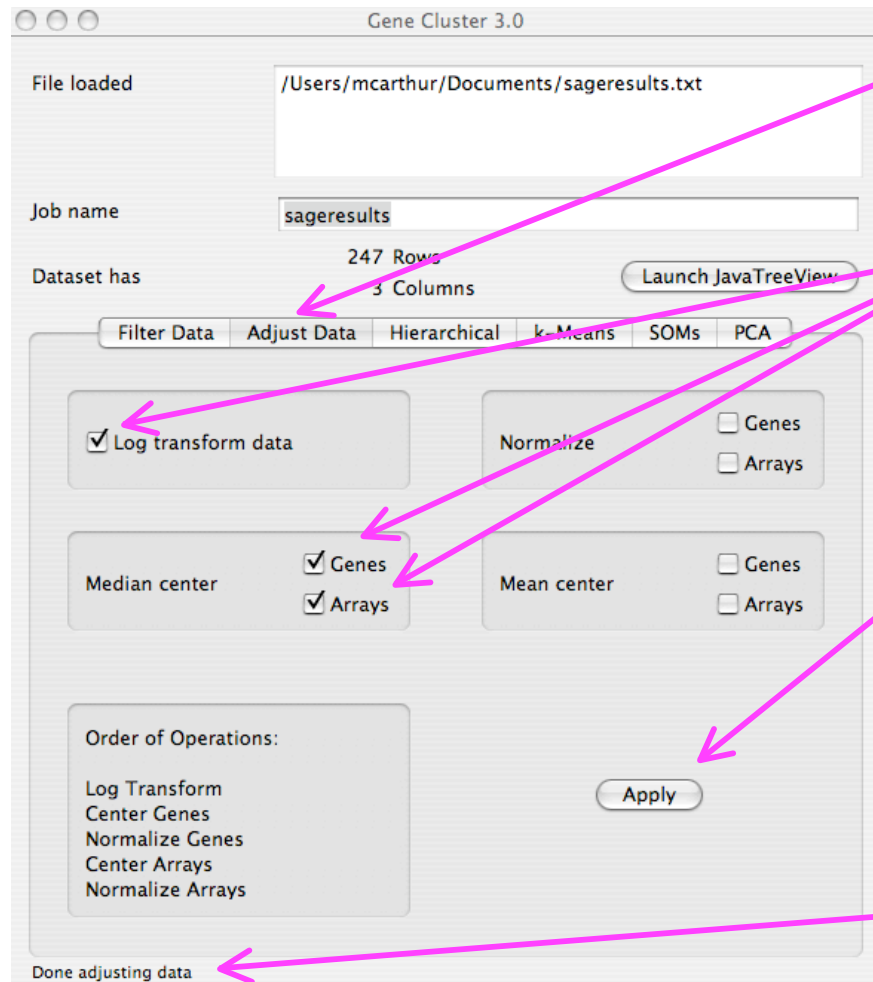
tagid	NAME	AM_SS	AF_SS	SA_LIVER
34563	tagtype:UK tagID:34563R-Value:6.2855 BLASTN of SAGETAG : gi 15778807 gb AC080889.5  Homo sapiens BAC clone RP11-785J10 from 4, complete sequence 0.836	1	1	19
26573	tagtype:UK tagID:26573R-Value:5.5081 BLASTN of SAGETAG : gi 25809521 emb AL845428.5  Zebrafish DNA sequence from clone CH211-133024 in linkage group 20, complete sequence 0.14	1	3	20
25464	tagtype:AA tagID:25464R-Value:4.0495 Orf:864 gi 22094807 gb AAM91993.1  egg secreted protein ESP15 [Schistosoma mansoni] 2e-39	1	13	1
25004	tagtype:PS tagID:25004R-Value:4.4212 Orf:5034 gi 19745168 ref NP_604448.1  lamin B receptor [Rattus norvegicus] gi 7513997 pir  JC5567 lamin B receptor - rat gi 2204062 dbj BAA20471.1  Rat NBP60 [Rattus norvegicus] 1e-32	1	4	18
23099	tagtype:UK tagID:23099R-Value:5.0395	1	1	16
22531	tagtype:PS tagID:22531R-Value:4.0679 Orf:8124 gi 56757568 gb AAW26946.1  unknown [Schistosoma japonicum] 3e-50	1	4	17
21385	tagtype:UK tagID:21385R-Value:5.2544 BLASTN of SAGETAG : gi 15420527 gb AF358445.1  Schistosoma mansoni glutamyl-tRNA synthetase mRNA, complete cds 0.009	1	9	23
21114	tagtype:UK tagID:21114R-Value:8.8321 BLASTN of SAGETAG : gi 33632062 emb BX569689.1  Synecococcus sp. WH8102 complete genome; segment 1/7 8.7	1	1	25
20795	tagtype:AS tagID:20795R-Value:4.8593 Orf:642 gi 66866675 gb AAV57921.1  22.6 kDa tegument antigen [Schistosoma mansoni] gi 135578 sp P14202 TEGU_SCHMA Tegument antigen (I(H)A) (Antigen SmA22.6) (A12) gi 161087 gb AAA29922.1  SM22.6 antigen (A12) gi 160933 gb AAA29856.1  antigen 1e-108	2	6	23
20113	tagtype:AS tagID:20113R-Value:5.1422 Orf:1959 gi 160955 gb AAC14467.1  Cu/Zn-superoxide dismutase [Schistosoma mansoni] 5e-86	2	3	21
17699	tagtype:UK tagID:17699R-Value:4.4279 BLASTN of SAGETAG : gi 68639430 emb CR339059.15  Zebrafish DNA sequence from clone CH211-121118 in linkage group 3, complete sequence 0.14	2	1	16
16710	tagtype:UK tagID:16710R-Value:4.4709 BLASTN of SAGETAG : gi 46240545 emb CR387786.1  Gallus gallus finished cDNA, clone CHEST533d6 8.7	1	14	1
16373	tagtype:AS tagID:16373R-Value:9.0481 Orf:3234 gi 66526630 ref XP_392104.2  PREDICTED: similar to CG31075-PA [Apis mellifera] 1e-166	1	29	4
15835	tagtype:UK tagID:15835R-Value:5.3058 BLASTN of SAGETAG : gi 18698807 gb AC090696.6  Homo sapiens chromosome 15, clone RP11-595N10, complete sequence 8.7	1	22	8
15604	tagtype:UK tagID:15604R-Value:4.2835	1	15	2
15598	tagtype:AS tagID:15598R-Value:6.3453 Orf:7727 gi 56753475 gb AAW24941.1  unknown [Schistosoma japonicum] 1e-25	1	20	2
15463	tagtype:PS tagID:15463R-Value:8.5004 Orf:4630 gi 67463829 pdb 1TD1 C Chain C, Crystal Structure Of The Purine Nucleoside Phosphorylase From Schistosoma Mansoni In Complex With Acetate gi 67463828 pdb 1TD1 B Chain B, Crystal Structure Of The Purine Nucleoside Phosphorylase From Schistosoma Mansoni In Complex With Acetate gi 1e-162	1	33	12
15405	tagtype:UK tagID:15405R-Value:7.0170 BLASTN of SAGETAG : gi 26103723 dbj AK086729.1  Mus musculus 15 days embryo head cDNA, RIKEN full-length enriched library, clone:D930048E06 product:unknown EST, full insert sequence 0.14	1	24	4
15402	tagtype:UK tagID:15402R-Value:5.7568 BLASTN of SAGETAG : gi 55467283 emb BX510316.11  Zebrafish DNA sequence from clone CH211-194J3 in linkage group 12, complete sequence 0.14	1	17	1
15359	tagtype:UK tagID:15359R-Value:5.7568 BLASTN of SAGETAG : gi 20197018 gb AC003096.3  Arabidopsis thaliana chromosome 2 clone T29F13 map ve016, complete sequence 0.14	1	17	1
13441	tagtype:PS tagID:13441R-Value:5.8825 Orf:1095 gi 60692116 gb AAV30610.1  unknown [Schistosoma japonicum] 8e-40	3	20	1
13148	tagtype:UK tagID:13148R-Value:5.1224 BLASTN of SAGETAG : gi 9581783 emb AL117374.39 HSDJ47A22 Human DNA sequence from clone RP1-47A22 on chromosome 20q12 Contains STSs and GSSs, complete sequence 0.14	2	17	1
12165	tagtype:PS tagID:12165R-Value:5.0856 Orf:3629 gi 7494503 pir  T30855 multidrug resistance protein 2 - fluke (Schistosoma mansoni) gi 425476 gb AAA66477.1  SMDR2 7e-61	3	18	1
11265	tagtype:UK tagID:11265R-Value:5.5802	4	3	25
11156	tagtype:PS tagID:11156R-Value:4.0269 Orf:1997 gi 30794206 ref NP_084385.1  splicing factor 3b, subunit 2 [Mus musculus] gi 29144992 gb AAH49118.1  Splicing factor 3b, subunit 2 [Mus musculus] 1e-126	1	18	8
11086	tagtype:UK tagID:11086R-Value:4.3637 BLASTN of SAGETAG : gi 26984795 emb AL845171.5  Mouse DNA sequence from clone RP23-38N8 on chromosome 4, complete sequence 8.7	3	18	2
10809	tagtype:UK tagID:10809R-Value:38.2557 BLASTN of SAGETAG : gi 3063366 dbj AB003713.1  Protula magnifica mRNA for elongation factor-1alpha, partial cds 0.14	1	1	91
10802	tagtype:UK tagID:10802R-Value:11.4264 BLASTN of SAGETAG : gi 1791220 gb U82283.1 SMU82283 Schistosoma malayensis 18S ribosomal RNA gene, partial sequence, ITS1, complete sequence and 5.8S ribosomal RNA gene, partial sequence 0.836	1	1	31
10787	tagtype:UK tagID:10787R-Value:4.2234 BLASTN of SAGETAG : gi 30349104 gb AC124499.3  Mus musculus BAC clone RP24-68G23 from chromosome 12, complete sequence 0.14	1	1	14
10596	tagtype:PS tagID:10596R-Value:10.6517 Orf:702 gi 56753077 gb AAW24748.1  unknown [Schistosoma japonicum] 4e-70	1	41	17
10542	tagtype:AS tagID:10542R-Value:11.5700 Orf:3275 gi 15986447 gb AAL11633.1  putative histamine-releasing factor [Schistosoma mansoni] gi 20140691 sp Q95WA2 TCTP_SCHMA Translationally controlled tumor protein homolog (TCTP) (Histamine-releasing factor) 2e-92	2	32	1
10467	tagtype:UK tagID:10467R-Value:26.0714 BLASTN of SAGETAG : gi 161027 gb J04017.1 SCMHSP86 S.mansoni heat shock protein 86 mRNA, complete cds 0.14	1	1	64
10260	tagtype:UK tagID:10260R-Value:4.3064 BLASTN of SAGETAG : gi 25168716 emb AL928696.6  Mouse DNA sequence from clone RP23-387G11 on chromosome 2, complete sequence 8.7	3	16	1
7449	tagtype:UK tagID:7449R-Value:4.0537 BLASTN of SAGETAG : gi 13157532 emb AL159154.16  Human DNA sequence from clone RP11-428G23 on chromosome 13 Contains part of a novel gene (KIAA0916), complete sequence 0.009	2	16	2
7020	tagtype:UK tagID:7020R-Value:4.9445 BLASTN of SAGETAG : gi 11095132 gb AC084686.1 CBRMM39H11 Caenorhabditis briggsae cosmid NM39H11, complete sequence 0.14	3	21	3
6957	tagtype:UK tagID:6957R-Value:7.4940 BLASTN of SAGETAG : gi 42569710 ref NM_129333.3  Arabidopsis thaliana aldo/keto reductase family protein (At2g37770) mRNA, complete cds 0.036	12	32	
6907	tagtype:AS tagID:6907R-Value:9.6772 Orf:4528 gi 60600029 gb AAV26558.1  unknown [Schistosoma japonicum] 1e-44	13	1	38
6767	tagtype:UK tagID:6767R-Value:53.3903 BLASTN of SAGETAG : gi 22094806 gb AF527012.1  Schistosoma mansoni egg secreted protein ESP15 (ESP15) mRNA, partial cds 2.2	3	125	1
6728	tagtype:PS tagID:6728R-Value:14.7359 Orf:1683 gi 605647 gb AAA57567.1  fructose 1,6 biphosphate aldolase [Schistosoma mansoni] gi 1703248 sp P53442 ALF_SCHMA Fructose-bisphosphate aldolase gi 2598926 gb AAB04014.1  fructose bisphosphate aldolase [Schistosoma mansoni] 0	3	58	15
6440	tagtype:UK tagID:6440R-Value:4.6639 BLASTN of SAGETAG : gi 12123261 emb U169298.8  Mouse DNA sequence from clone RP23-340H11 on chromosome 4 Contains the 3' end of the Col27a1 gene for procollagen			

# Step 2 - CLUSTER Analysis



- Start the CLUSTER software
- From the *File* menu, open the "sageresults.txt" file
- If you have succeeded, the file should now be listed

# Step 2 - CLUSTER Analysis



Change to the *Adjust Data* section

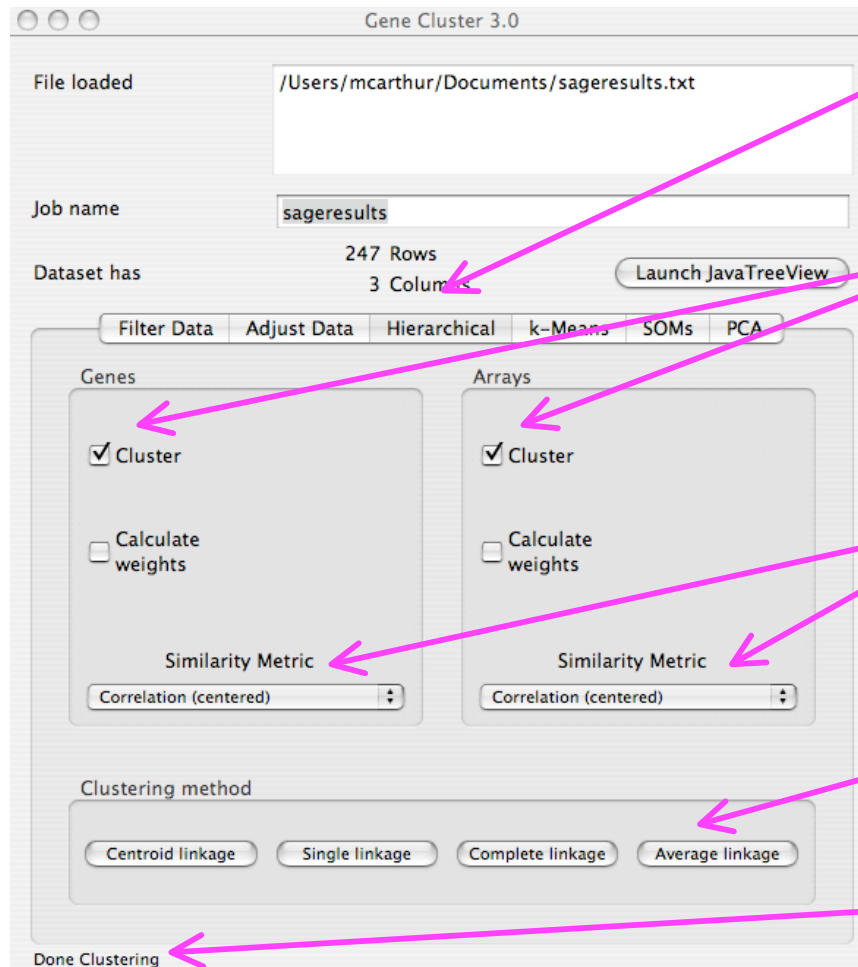
Select *Log transform data* and *Median center of Genes and Arrays*

Press *Apply*

If you have succeeded, you should see this message

A log transformation is used to reduce the power relationship that exists with SAGE data (i.e. the standard deviation of tag frequencies is not independent of the mean of tag frequencies). Note that this is a different reason for log transforming than for microarray data (see manual). Median centering adjusts tag frequencies to a range of -1 to +1, with the median being zero. This reduces the influence of tag magnitude - lowly expressed genes are treated the same way as highly expressed genes such that only that pattern of change counts, not the magnitude. Normalization is not selected here as the distance metric we use does not require it (see manual).

# Step 2 - CLUSTER Analysis



Change to the *Hierarchical* section

Select *Cluster* for both *Genes* (i.e. tags) and *Arrays* (i.e. SAGE libraries)

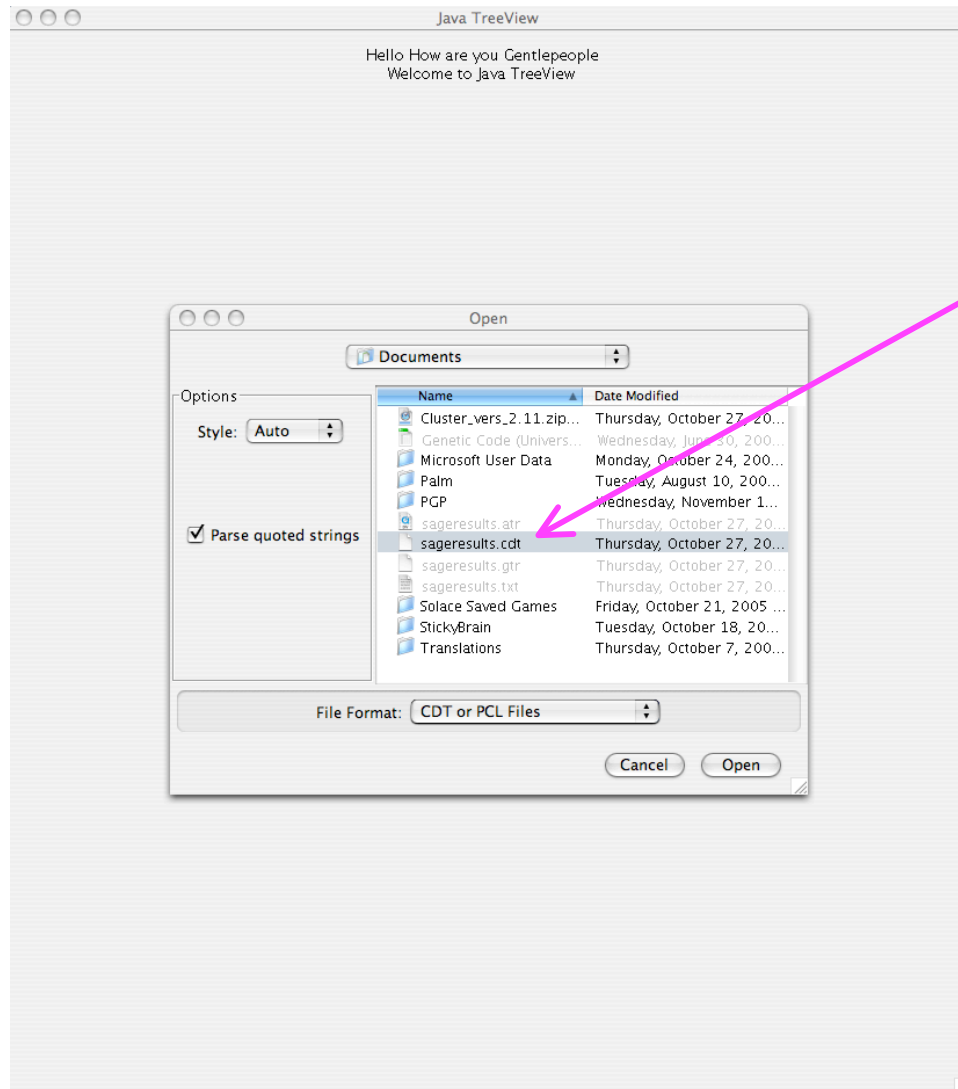
Select *Correlation (centered)* for both *Genes* and *Arrays*

Press *Average Linkage*

If you have succeeded, you should see this message

We are clustering both SAGE tags (*Genes*) to find tags with correlated expression profiles and SAGE libraries (*Arrays*) to discern correlated global gene expression among libraries. Centered correlation (i.e. Pearson's correlation) is a reasonable metric for SAGE data (Poisson would be better!). Note that this metric results in an analysis that is insensitive to magnitude of expression. This means our cluster analysis is focused upon correlated patterns of expression, whether genes are lowly or highly expressed. This can be particularly important for highlighting tags from less abundant transcripts (which otherwise do not have striking *R*-values). Other metrics in CLUSTER may be more robust to outliers, but are sensitive to magnitude. See the manual for an explanation of different clustering methods - we like to focus on average differences between tags, but each method has its strengths and weaknesses.

# Step 3 - View the Results



- Start the TREEVIEW software

From the *File* menu, open the “sageresults.cdt” file

# Step 3 - View the Results

This cluster has 55.4% correlated expression

Select a cluster to obtain details

SAGE libraries, with global correlation results

Tags and their annotations

