CANB7640
Practical Workshop
Class 02

Aik Choon Tan, Ph.D.
Associate Professor of Bioinformatics
Division of Medical Oncology
Department of Medicine
aikchoon.tan@ucdenver.edu
9/10/2018
http://tanlab.ucdenver.edu/labHomePage/teaching/CANB7640/

Some thoughts - Assignment #1

- Applaud to those that have submitted on time

- Can a biologist do programming? – YES

- Is there a “right” or “wrong” program?

- Many ways to program - efficiency

@chomp

- New line can be a character (when you hit “Enter” in your keyboard, it generates a new line)
- You don’t see it, but it is there
- Cutting out new line

Reading in file

```
while (<>) # read in a file line by line
```

```
@array = (<>); # read in the file as any array, every line is an array element
```

Reading in file

`open (MYFILE, “file.txt”)` # open a file name `file.txt` and assign it to the variable `MYFILE`.

while (<MYFILE>) #read in the `file.txt` and read in line by line

`@array = (<MYFILE>);` # read in the `file.txt` as any array, every line is an array element

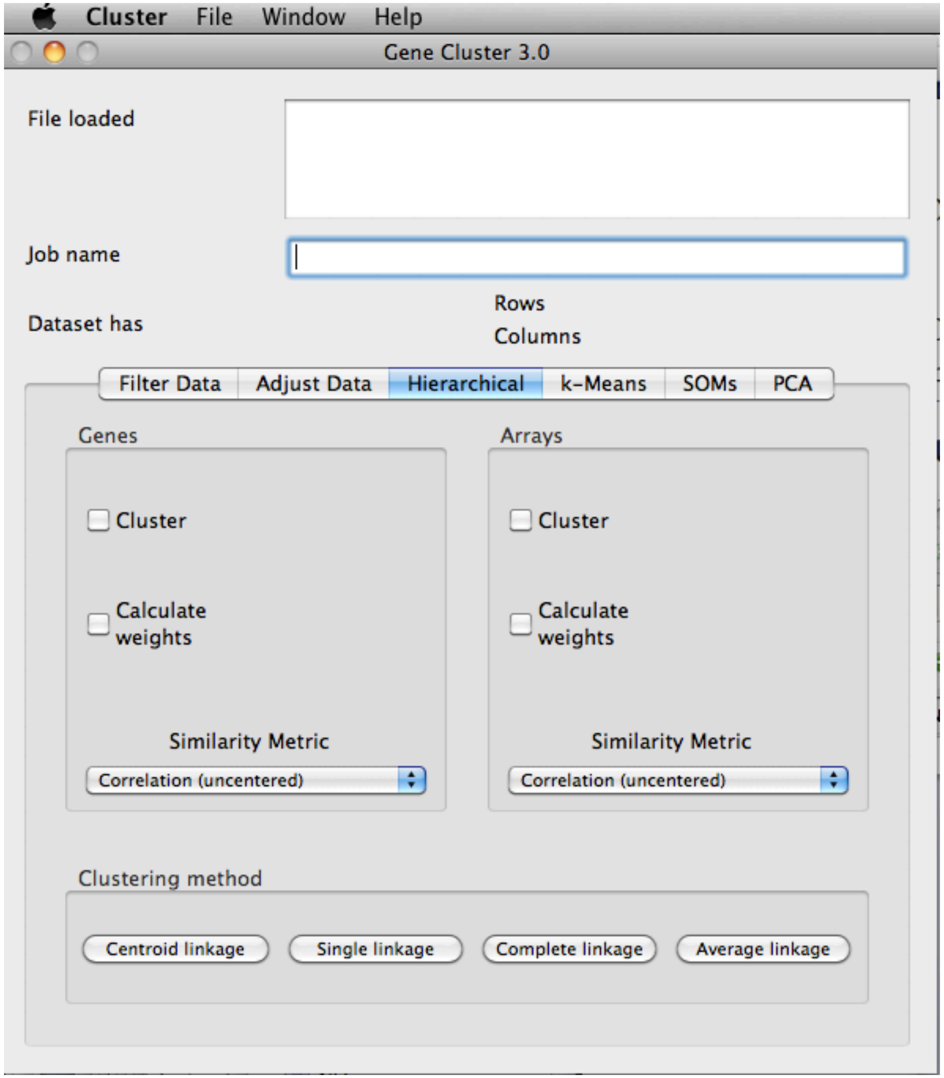
ReadFile_WHILE.pl

```
#! /usr/bin/perl -w

open (MYFILE, "file.txt");
open (OUTFILE, "> outfile_while.txt");
while (<MYFILE>)
{
        @tmp = split (/\t/);
        chomp @tmp;

        print OUTFILE "$tmp[0] || $tmp[1]\n";
}

close (MYFILE);
close (OUTFILE);
```

ReadFile_ARRAY.pl

```
#! /usr/bin/perl -w

open (MYFILE, "file.txt");
open (OUTFILE, "> outfile_array.txt");

@array = (<MYFILE>);
chomp @array;

for ($i = 0; $i <= $#array; $i++)
{
        print OUTFILE "$array[$i]\n";
}

close (MYFILE);
close (OUTFILE);
```

CLUSTER 3.0



File loaded

Job name

Dataset has
Rows
Columns

Genes

☐ Cluster

☐ Calculate weights

Similarity Metric: Correlation (uncentered)

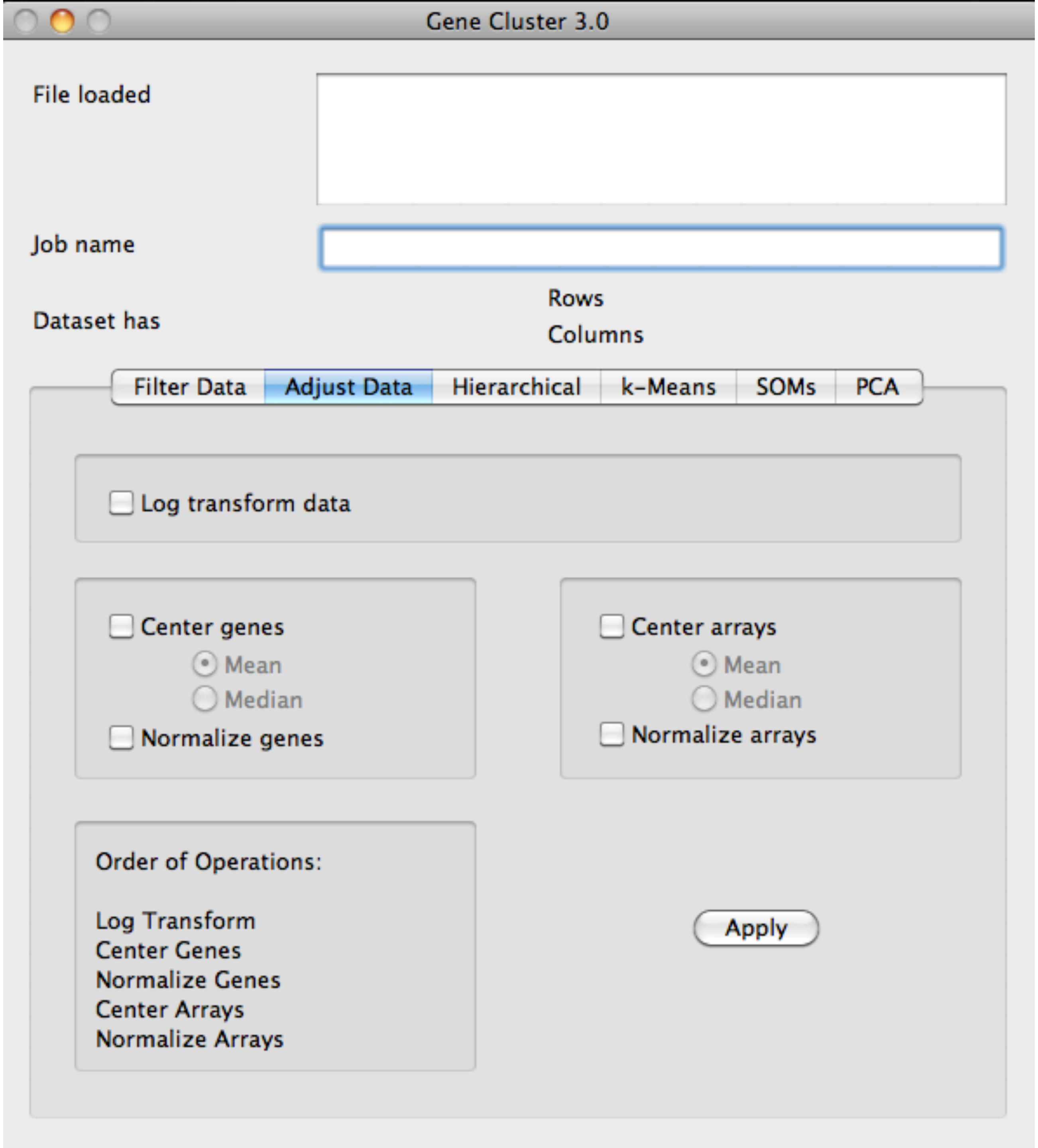
Arrays

☐ Cluster

☐ Calculate weights

Similarity Metric: Correlation (uncentered)

Clustering method



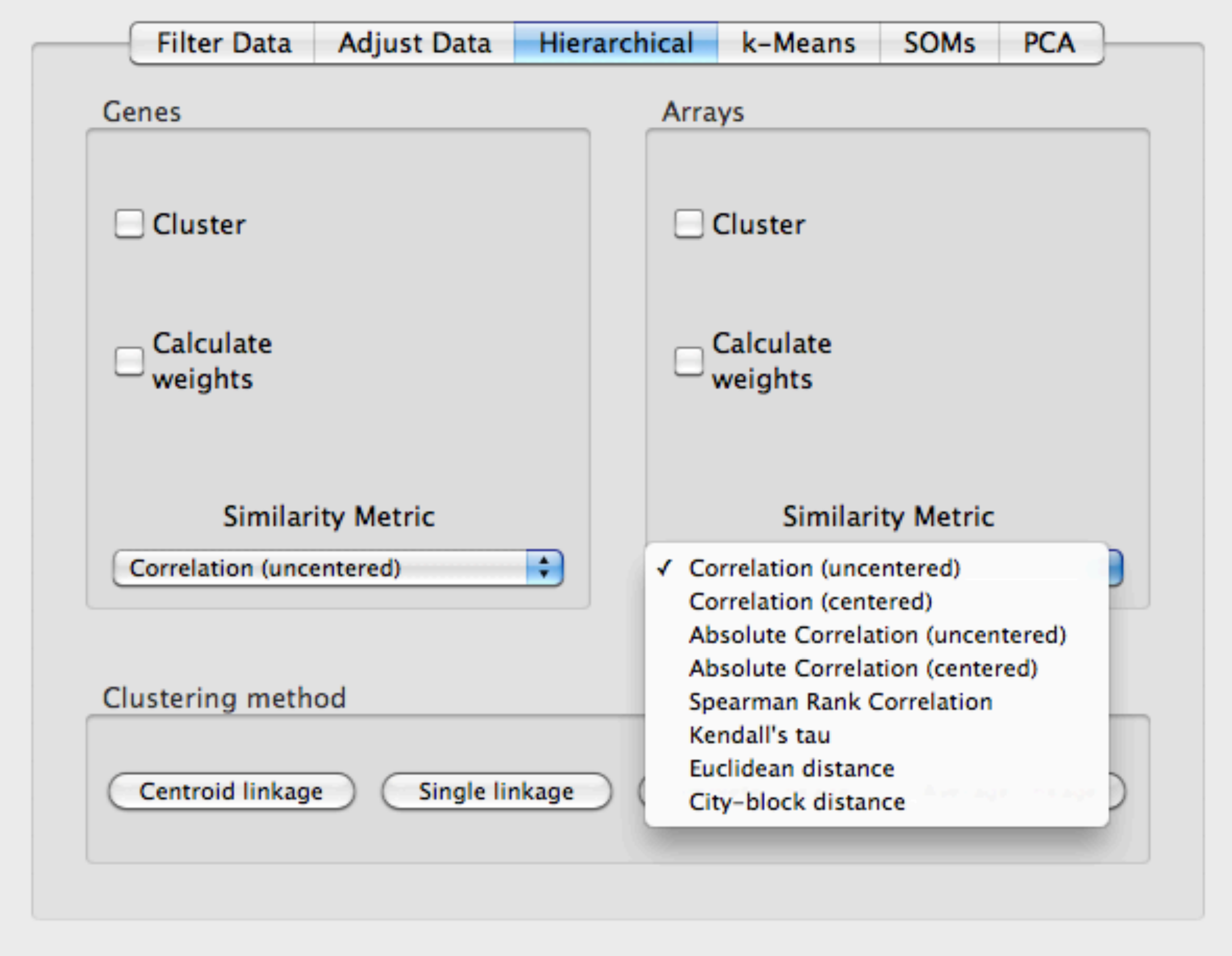
☐ Log transform data

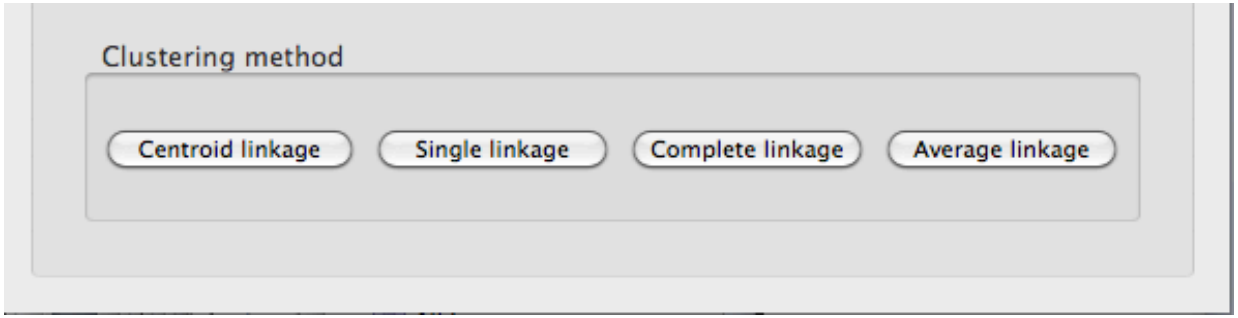
☐ Center genes
● Mean
○ Median
☐ Normalize genes

☐ Center arrays
● Mean
○ Median
☐ Normalize arrays

Order of Operations:

Log Transform
Center Genes
Normalize Genes
Center Arrays
Normalize Arrays

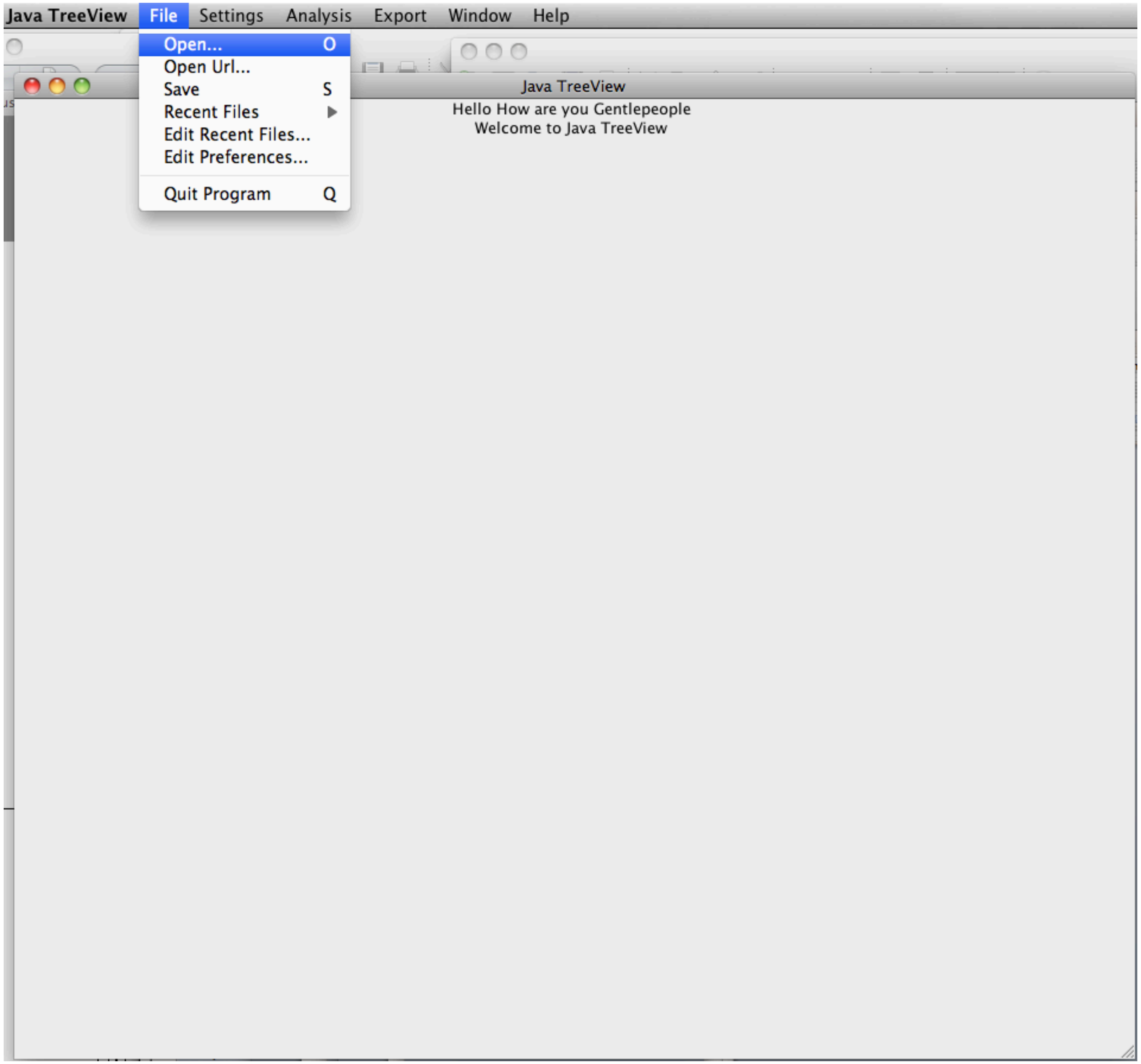

Filter Data
Adjust Data
Hierarchical
k-Means
SOMs
PCA
Genes
☐ Cluster
☐ Calculate weights
Similarity Metric
Correlation (uncentered)
Arrays
☐ Cluster
☐ Calculate weights
Similarity Metric
✓ Correlation (uncentered)
Correlation (centered)
Absolute Correlation (uncentered)
Absolute Correlation (centered)
Spearman Rank Correlation
Kendall's tau
Euclidean distance
City-block distance
Clustering method
Centroid linkage
Single linkage


Clustering method
Centroid linkage
Single linkage
Complete linkage
Average linkage

Clustering Results

- *.cdt – clustered data table
- *.atr – array tree (samples)
- *.gtr – gene tree (genes)

Java Tree

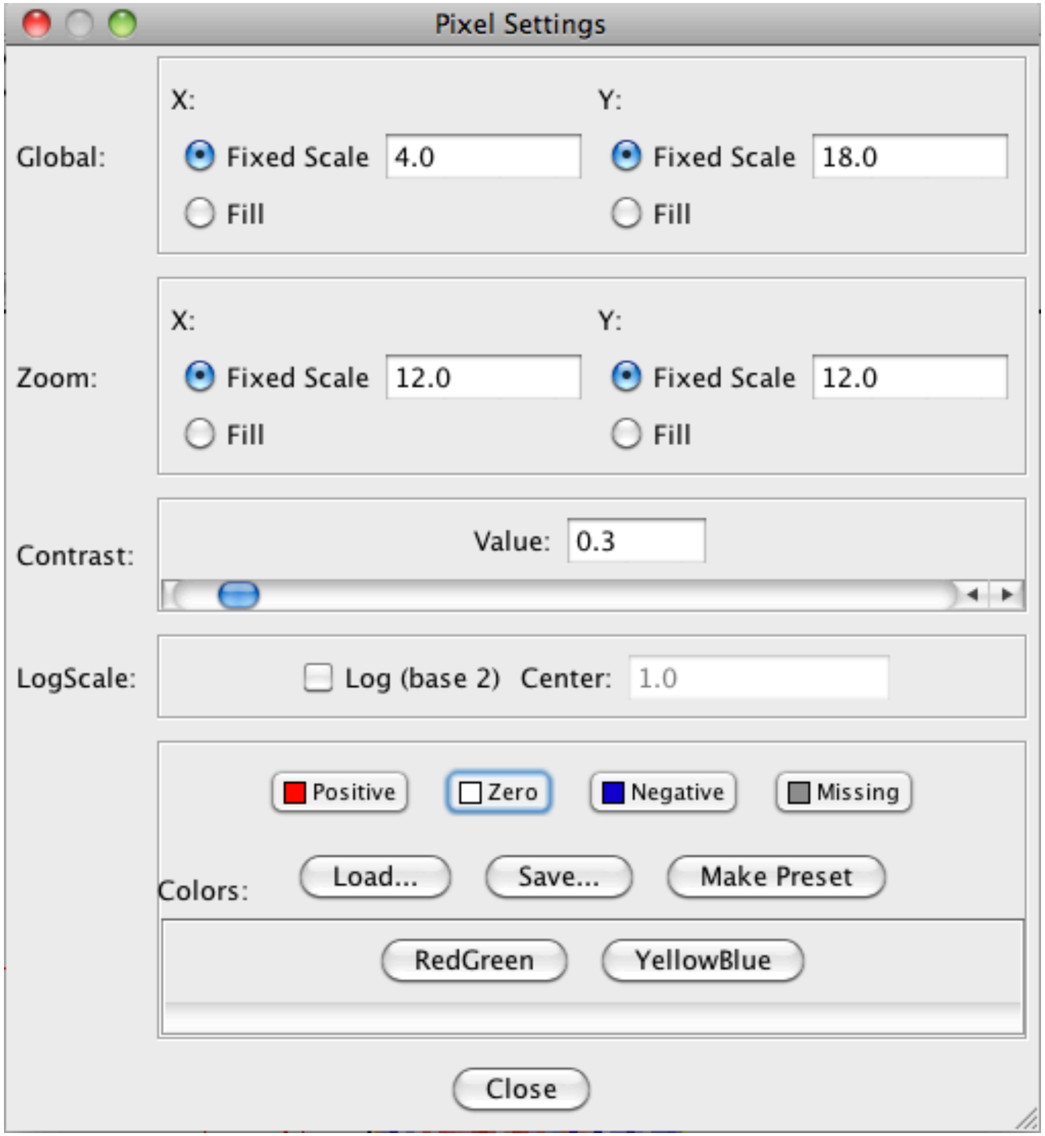


Hello How are you Gentlepeople
Welcome to Java TreeView

View Status
No status info for ArrayName

Usage Hints
Click and drag to scroll

MB-1 gene (U05259_rna1_at)
HKR-T1 (S50223_at)
Inducible protein mRNA (L47738_at)
CCND3 Cyclin D3 (M92287_at)
Transcriptional activator hSNF2b (D26156_s_at)
PROTEASOME IOTA CHAIN (X59417_at)
MYL1 Myosin light chain (alkali) (M31211_s_at)
C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete
Adenosine triphosphatase, calcium (Z69881_at)
TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) (M3
TOP2B Topoisomerase (DNA) II beta (180kD) (Z15115_at)
RETINOBLASTOMA BINDING PROTEIN P48 (X74262_at)
Cytoplasmic dynein light chain 1 (hdlc1) mRNA (U32944_at)
ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain (M91432_at)
IRF2 Interferon regulatory factor 2 (X15949_at)
MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3 (D38073_at)
GTF2E2 General transcription factor TFIIE beta subunit, 34 kD (X63469_at)
DHPS Deoxyhypusine synthase (U26266_s_at)
Transcriptional activator hSNF2b (U29175_at)
SRP9 Signal recognition particle 9 kD protein (U20998_at)
Oncoprotein 18 (Op18) gene (M31303_rna1_at)
ADA Adenosine deaminase (M13792_at)
RABAPTIN-5 protein (Y08612_at)
Heterochromatin protein p25 mRNA (U35451_at)
IL7R Interleukin 7 receptor (M29696_at)
DF D component of complement (adipsin) (M84526_at)
Homeodomain protein HoxA9 mRNA (U82759_at)
Azurocidin gene (M96326_rna1_at)
PRG1 Proteoglycan 1, secretory granule (X17042_at)
LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog (M16038_at)
CD33 CD33 antigen (differentiation antigen) (M23197_at)
Zyxin (X95735_at)
FAH Fumarylacetoacetate (M55150_at)
Leukotriene C4 synthase (LTC4S) gene (U50136_rna1_at)
CTSD Cathepsin D (lysosomal aspartyl protease) (M63138_at)
ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide) (M81695_s_at)
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage) (M27891_at)
LYZ Lysozyme (M19045_f_at)
Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA (U46751_at)
INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1 (L08246_at)
INTERLEUKIN-8 PRECURSOR (Y00787_s_at)
Interleukin 8 (IL8) gene (M28130_rna1_s_at)
MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3 (M69043_at)
ATP6C Vacuolar H+ ATPase proton channel subunit (M62762_at)
PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (M80254_at)
LEPR Leptin receptor (Y12670_at)
Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joi
Epb72 gene exon 1 (X85116_rna1_s_at)
LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of sy
PFC Properdin P factor, complement (M83652_s_at)



Pixel Settings

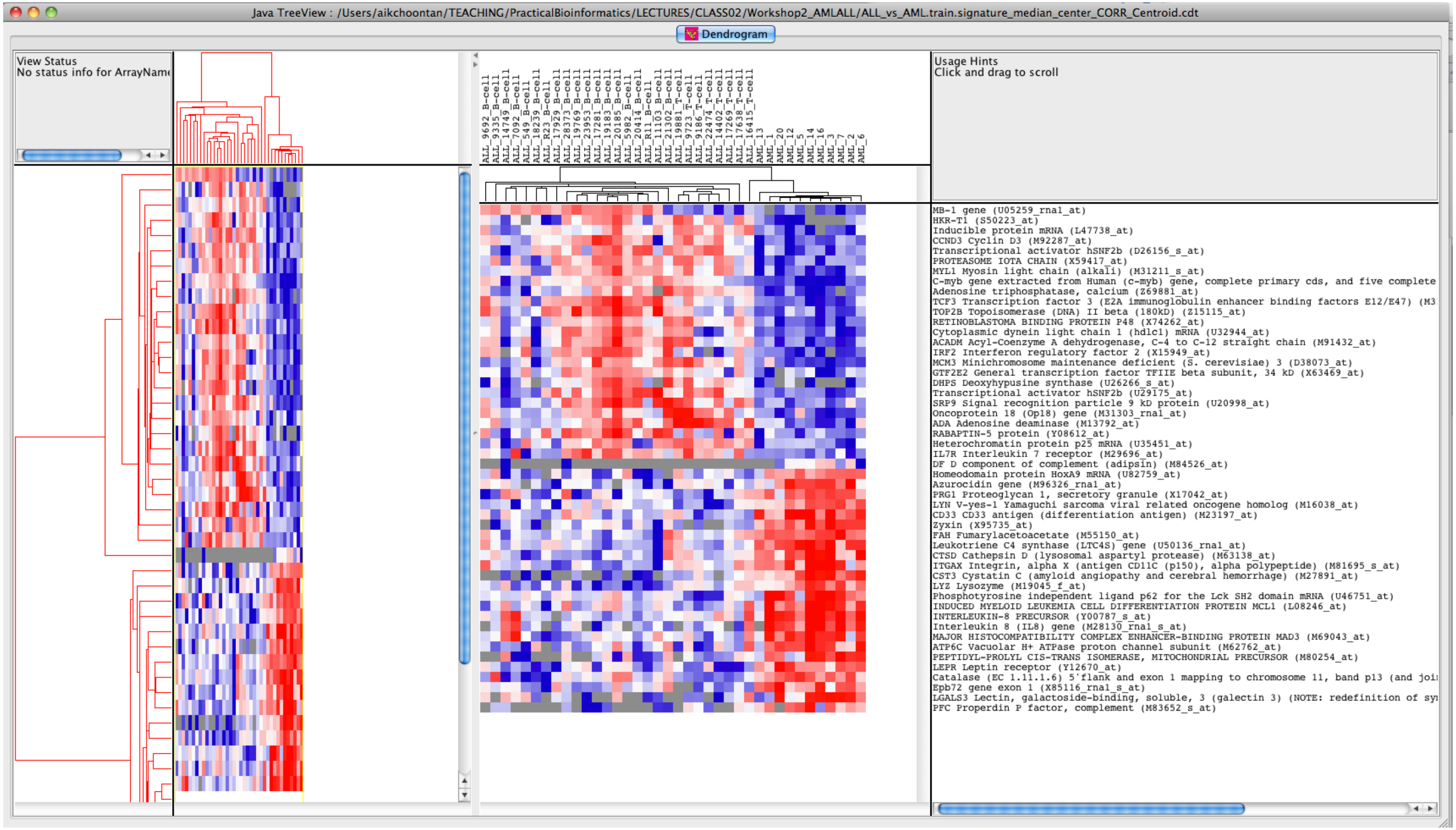
Global:
X: ● Fixed Scale 4.0 ○ Fill
Y: ● Fixed Scale 18.0 ○ Fill

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

Contrast: Value: 0.3

LogScale: ☐ Log (base 2) Center: 1.0

Colors:

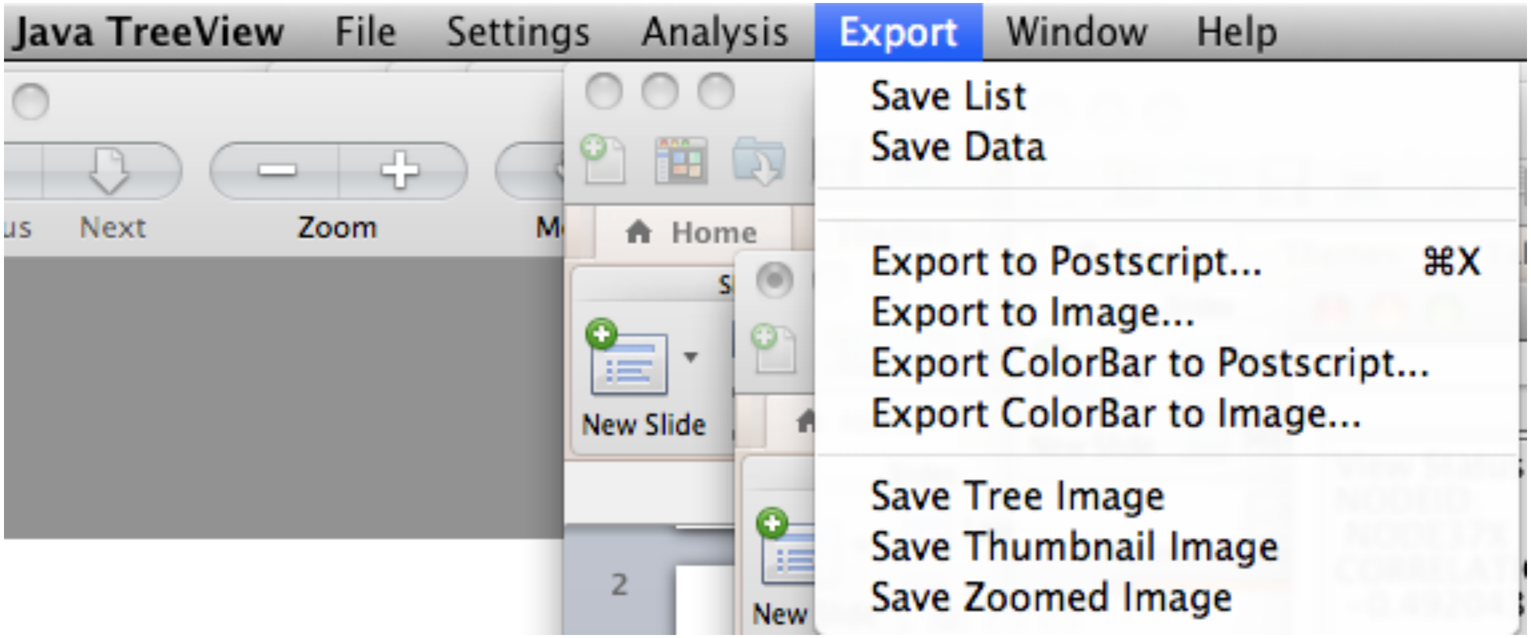


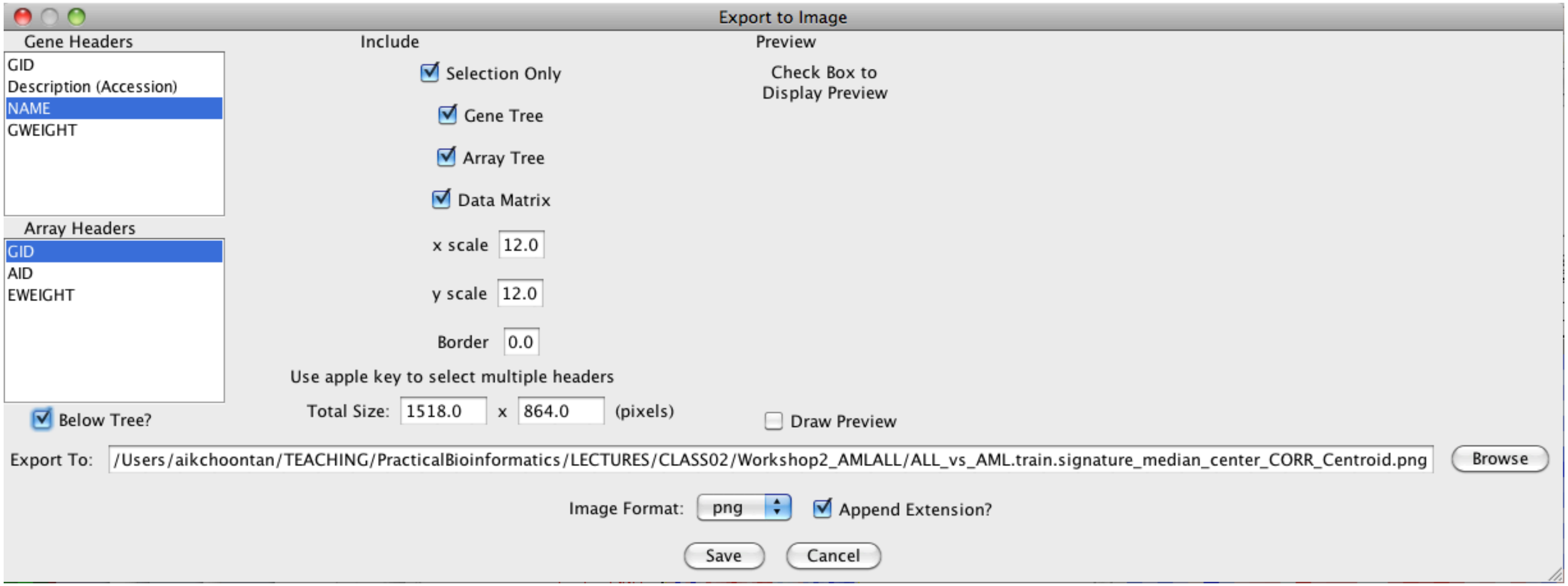
Java TreeView : /Users/aikchoontan/TEACHING/PracticalBioinformatics/LECTURES/CLASS02/Workshop2_AMLALL/ALL_vs_AML.train.signature_median_center_CORR_Centroid.cdt

View Status
No status info for ArrayName

Usage Hints
Click and drag to scroll

```
MB-1 gene (U05259_rna1_at)
HKR-T1 (S50223_at)
Inducible protein mRNA (L47738_at)
CCND3 Cyclin D3 (M92287_at)
Transcriptional activator hSNF2b (D26156_s_at)
PROTEASOME IOTA CHAIN (X59417_at)
MYL1 Myosin light chain (alkali) (M31211_s_at)
C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete
Adenosine triphosphatase, calcium (Z69881_at)
TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) (M3
TOP2B Topoisomerase (DNA) II beta (180kD) (Z15115_at)
RETINOBLASTOMA BINDING PROTEIN P48 (X74262_at)
Cytoplasmic dynein light chain 1 (hdlc1) mRNA (U32944_at)
ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain (M91432_at)
IRF2 Interferon regulatory factor 2 (X15949_at)
MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3 (D38073_at)
GTF2E2 General transcription factor TFIIE beta subunit, 34 kD (X63469_at)
DHPS Deoxyhypusine synthase (U26266_s_at)
Transcriptional activator hSNF2b (U29175_at)
SRP9 Signal recognition particle 9 kD protein (U20998_at)
Oncoprotein 18 (Op18) gene (M31303_rna1_at)
ADA Adenosine deaminase (M13792_at)
RABAPTIN-5 protein (Y08612_at)
Heterochromatin protein p25 mRNA (U35451_at)
IL7R Interleukin 7 receptor (M29696_at)
DF D component of complement (adipsin) (M84526_at)
Homeodomain protein HoxA9 mRNA (U82759_at)
Azurocidin gene (M96326_rna1_at)
PRG1 Proteoglycan 1, secretory granule (X17042_at)
LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog (M16038_at)
CD33 CD33 antigen (differentiation antigen) (M23197_at)
Zyxin (X95735_at)
FAH Fumarylacetoacetate (M55150_at)
Leukotriene C4 synthase (LTC4S) gene (U50136_rna1_at)
CTSD Cathepsin D (lysosomal aspartyl protease) (M63138_at)
ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide) (M81695_s_at)
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage) (M27891_at)
LYZ Lysozyme (M19045_f_at)
Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA (U46751_at)
INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1 (L08246_at)
INTERLEUKIN-8 PRECURSOR (Y00787_s_at)
Interleukin 8 (IL8) gene (M28130_rna1_s_at)
MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3 (M69043_at)
ATP6C Vacuolar H+ ATPase proton channel subunit (M62762_at)
PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (M80254_at)
LEPR Leptin receptor (Y12670_at)
Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joi
Epb72 gene exon 1 (X85116_rna1_s_at)
LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of sy
PFC Properdin P factor, complement (M83652_s_at)
```


Java TreeView
File
Settings
Analysis
Export
Window
Help
Save List
Save Data
Export to Postscript... ⌘X
Export to Image...
Export ColorBar to Postscript...
Export ColorBar to Image...
Save Tree Image
Save Thumbnail Image
Save Zoomed Image
Next
Zoom
Home
New Slide
2
New



Export to Image

Gene Headers

GID
Description (Accession)
NAME
GWEIGHT

Array Headers

GID
AID
EWEIGHT

☑ Below Tree?

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

Total Size: 1518.0 x 864.0 (pixels)

Preview

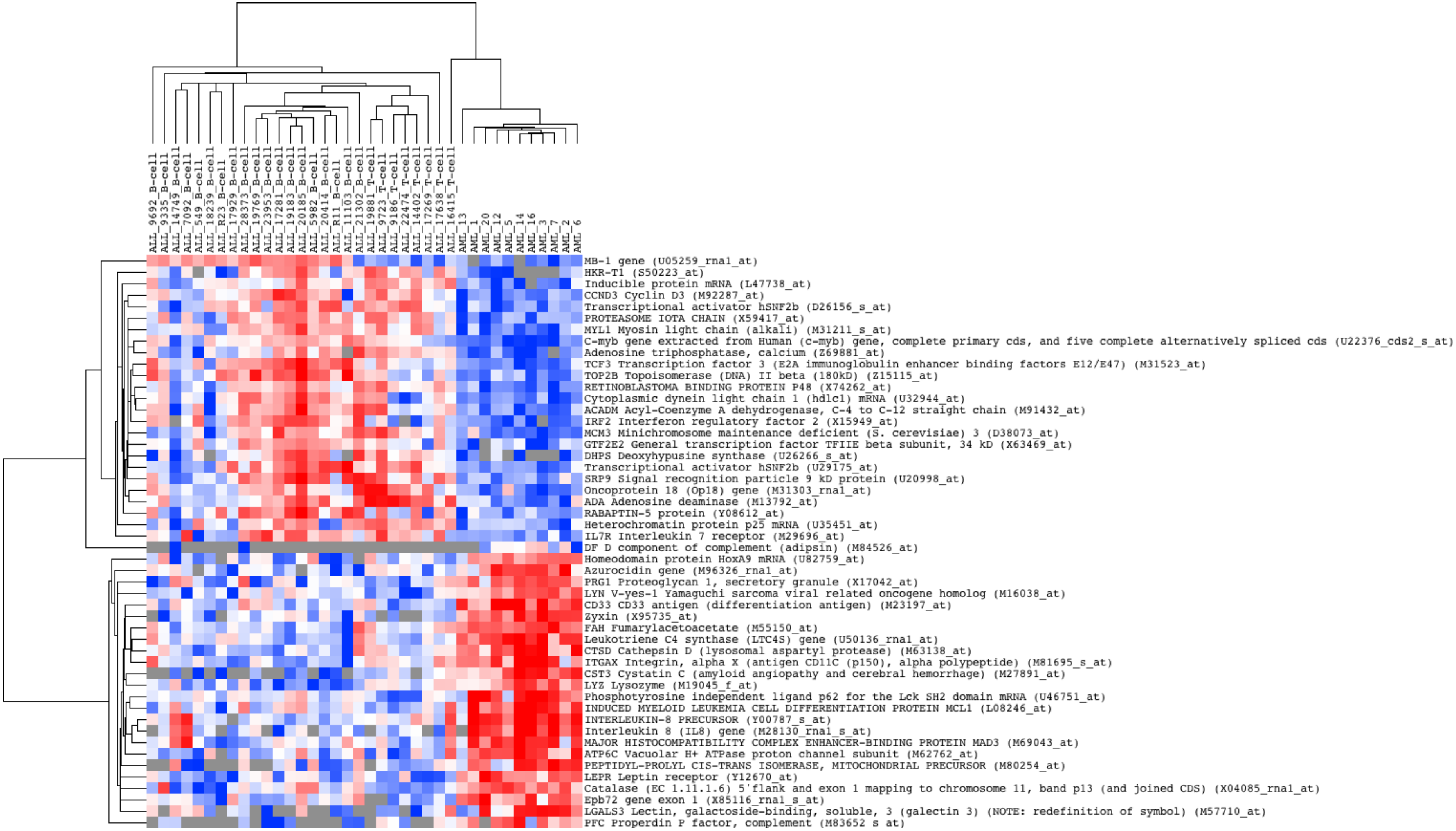
Check Box to Display Preview

☐ Draw Preview

Export To: /Users/aikchoontan/TEACHING/PracticalBioinformatics/LECTURES/CLASS02/Workshop2_AMLALL/ALL_vs_AML.train.signature_median_center_CORR_Centroid.png

Image Format: png

☑ Append Extension?


ALL_9692_B-cell
ALL_9335_B-cell
ALL_14749_B-cell
ALL_7092_B-cell
ALL_549_B-cell
ALL_18239_B-cell
ALL_R23_B-cell
ALL_17929_B-cell
ALL_28373_B-cell
ALL_19769_B-cell
ALL_23953_B-cell
ALL_17281_B-cell
ALL_19183_B-cell
ALL_20185_B-cell
ALL_5982_B-cell
ALL_20414_B-cell
ALL_R11_B-cell
ALL_11103_B-cell
ALL_21302_B-cell
ALL_19881_T-cell
ALL_9723_T-cell
ALL_9186_T-cell
ALL_22474_T-cell
ALL_14402_T-cell
ALL_17269_T-cell
ALL_17638_T-cell
ALL_16415_T-cell
AML_13
AML_1
AML_20
AML_12
AML_5
AML_14
AML_16
AML_3
AML_7
AML_2
AML_6
MB-1 gene (U05259_rna1_at)
HKR-T1 (S50223_at)
Inducible protein mRNA (L47738_at)
CCND3 Cyclin D3 (M92287_at)
Transcriptional activator hSNF2b (D26156_s_at)
PROTEASOME IOTA CHAIN (X59417_at)
MYL1 Myosin light chain (alkali) (M31211_s_at)
C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds (U22376_cds2_s_at)
Adenosine triphosphatase, calcium (Z69881_at)
TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) (M31523_at)
TOP2B Topoisomerase (DNA) II beta (180kD) (Z15115_at)
RETINOBLASTOMA BINDING PROTEIN P48 (X74262_at)
Cytoplasmic dynein light chain 1 (hdlc1) mRNA (U32944_at)
ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain (M91432_at)
IRF2 Interferon regulatory factor 2 (X15949_at)
MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3 (D38073_at)
GTF2E2 General transcription factor TFIIE beta subunit, 34 kD (X63469_at)
DHPS Deoxyhypusine synthase (U26266_s_at)
Transcriptional activator hSNF2b (U29175_at)
SRP9 Signal recognition particle 9 kD protein (U20998_at)
Oncoprotein 18 (Op18) gene (M31303_rna1_at)
ADA Adenosine deaminase (M13792_at)
RABAPTIN-5 protein (Y08612_at)
Heterochromatin protein p25 mRNA (U35451_at)
IL7R Interleukin 7 receptor (M29696_at)
DF D component of complement (adipsin) (M84526_at)
Homeodomain protein HoxA9 mRNA (U82759_at)
Azurocidin gene (M96326_rna1_at)
PRG1 Proteoglycan 1, secretory granule (X17042_at)
LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog (M16038_at)
CD33 CD33 antigen (differentiation antigen) (M23197_at)
Zyxin (X95735_at)
FAH Fumarylacetoacetate (M55150_at)
Leukotriene C4 synthase (LTC4S) gene (U50136_rna1_at)
CTSD Cathepsin D (lysosomal aspartyl protease) (M63138_at)
ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide) (M81695_s_at)
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage) (M27891_at)
LYZ Lysozyme (M19045_f_at)
Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA (U46751_at)
INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1 (L08246_at)
INTERLEUKIN-8 PRECURSOR (Y00787_s_at)
Interleukin 8 (IL8) gene (M28130_rna1_s_at)
MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3 (M69043_at)
ATP6C Vacuolar H+ ATPase proton channel subunit (M62762_at)
PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (M80254_at)
LEPR Leptin receptor (Y12670_at)
Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS) (X04085_rna1_at)
Epb72 gene exon 1 (X85116_rna1_s_at)
LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol) (M57710_at)
PFC Properdin P factor, complement (M83652_s_at)