MicroArray Genome Imaging and Clustering Tool

MAGIC Tool
User Guide

MAGIC Tool v1.6

June 8, 2007

The Goal of MAGIC Tool

The purpose of MAGIC Tool is to allow the user to begin with DNA microarray tiff files and end with biologically meaningful information. Comparative hybridization data (glass chips) and Affymetrix data are compatible with MAGIC Tool. You can start with tiff files or expression files (spreadsheet of ratios or absolute expression levels).

MAGIC Tool was created with the novice in mind but it is not a “dumbed down” program. In fact, MAGIC Tool is designed to illuminate the algorithms being used rather than be a black box that produces results with little input from the user. MAGIC Tool allows the user to change parameters for clustering, data quantification etc. This User’s Guide will teach you how to use the software but leaves the theoretical explanations to the Instructor’s Guide.

Users are also encouraged to visit related sites:

MAGIC web site: http://www.bio.davidson.edu/MAGIC
Online MAGIC Tool lab: http://gcat.davidson.edu/GCAT/workshop2/derisi_lab.html
Tutorial for Clustering: http://gcat.davidson.edu/DGPB/clust/home.htm
GCAT: <www.bio.davidson.edu/GCAT>
Genomics Course: http://www.bio.davidson.edu/genomics

MAGIC Tool support is provided by the authors and student assistants (with NSF support). Email magictool.help@gmail.com or laheyer@davidson.edu for assistance. You can also email the GCAT listserve for help, as there are many MAGIC Tool users on this list.

Release Information

The following features were added in MAGIC Tool 1.6:

- Java TreeView has been incorporated into MAGIC Tool. JTreeView has more powerful and faster displays of hierarchical clustering (i.e. dendrograms). JTreeView also has a display that can help identify aneuploidy.
- Manual flagging has been implemented.
- Help inside of MAGIC Tool based on user manual.
  (Note: The user manual is currently being revised. A newer version of the user manual will be included in MAGIC Tool when it is completed.)
- Red vs. green dye swaps can be handled after expression data has been created.
- Dissimilarities using the method of 1 - |correlation| can now be computed, allowing genes that are oppositely expressed to be clustered together.

These features have not yet been documented. When the features are fully documented, a new version of the user’s guide will be released.

Installing MAGIC Tool

MAGIC Tool is distributed freely by Davidson College under the GNU public license. New versions of MAGIC Tool can be downloaded from the MAGIC Tool web page:

http://www.bio.davidson.edu/MAGIC
Beginning with version 1.5, the MAGIC Tool download consist of a single zip archive file, called MAGIC_Tool_1-x.zip, which you must decompress to see the MAGIC Tool folder, called MAGIC_Tool_1-x. The contents of the folder are described in the following table.

File Name	Description
Magic_launch.bat	Launcher for Windows (Executable)
MAGIC_launch	Launcher for Mac OS X (Executable)
MagicTool.jar	MAGIC Tool code (called by launcher)
MAGIC Users Guide v1-6.pdf	Users guide (this file)
Installation_guide.pdf	Detailed instructions for installing and running MAGIC Tool
MAGIC Instructor's Guide.pdf	Instructors guide with additional algorithmic details
Plugins	Necessary files for Java TreeView

After you unzip the downloaded file, navigate into the MAGIC_Tool_1-x folder and double click on the appropriate launcher file for your operating system. After a few seconds, the MAGIC Tool "splash screen" logo should appear, and in a few more seconds the program should be open. If the launcher does not properly start the MAGIC Tool program, see the MAGIC installation guide for detailed instructions.
Sample files and source code for MAGIC Tool are also available at the MAGIC Tool Website at http://www.bio.davidson.edu/magic/.

System Requirements

- Windows 2000 or later
- Mac OSX 10.3 or later
- Unix/Linux
- 512 MB RAM required for full size arrays; 1 GB of RAM recommended.
- Several hundred MB of hard drive space available, depending on the files you work with and what type of analyses you perform

Vocabulary

Addressing is the short process of telling MAGIC Tool the layout of the spots and grids in the tiff file as viewed within MAGIC.

Chip is a synonym for a microarray.

Feature is a synonym for a single spot on a microarray.

Flag is a verb that means you mark a particular spot to indicate its data are not reliable. This may be due to high background in the area, a dust bunny sitting on the spot, etc.

Grid is a compact arrangement of spots with even spacing.

Gridding is the process that MAGIC uses to find the spots on your tiff files

Metagrid is a higher order level of organization. A set of grids are organized into groups called metagrids. For a more complete description, see this web page <www.bio.davidson.edu/projects/GCAT/Griding.html>.

Segmentation is the process of finding the signal and distinguishing it from the background. There are three methods in MAGIC. Fixed circle is the fastest, and recommended for most purposes. Adaptive circle and seeded region growing are also provided.

Tiff files (e.g. file_name.tif) are the raw image data that are produced when a DNA microarray is scanned. One tiff file is produced for each color on each chip.

Getting Started

Overview of Steps

If you start with two tiff files, you will need to perform the following steps in order to produce clusters or explore your data.

1) Start a Project
2) Add files to project (recommended)
3) Load tiff files
4) Load gene list
5) Locate spots (Gridding and Addressing)
6) Distinguish signal from background and generate expression file (Segmentation)
7) Repeat steps 1-6 for all experimental conditions, appending to previous data and forming an expression file with several columns
8) Log-transform ratios
9) Add gene info to expression file (optional)
10) Explore data (recommended)
11) Filter data (recommended)

The following steps can only be performed if you have multiple columns in your expression file:

12) Calculate dissimilarity (e.g. 1 – correlation)
13) Cluster genes

(1) Start a Project

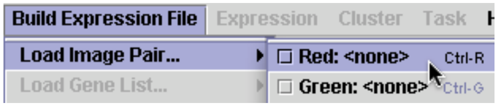
Under the Project menu, create a new Project. You can save this project in a convenient location on your hard drive. We recommend that you NOT use the MAGIC Tool software folder, since you may want to open this project with a newer version of MAGIC Tool in the future. Project files are automatically given a name that ends with the suffix ".gprj" and stored in a folder by the same name, automatically created by MAGIC Tool.

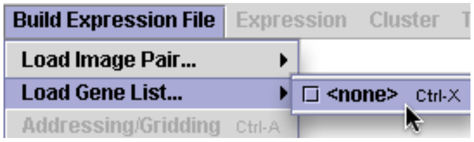
(2) Add Files to Project

We recommend that you copy files into your project, either through the Project menu options, or by dragging the files into the project folder and then selecting "Update Project" under the Project menu. Adding files to your project organizes your files for you into default folders, and simplifies future steps in the analysis.

(3) Load Tiff Files (Control R and Control G)

Under the Build Expression File menu, load the red and green tiff image pairs. Remember that red is a longer wavelength than green, so if your files are identified by the wavelengths, you should still be able to determine which color is which.





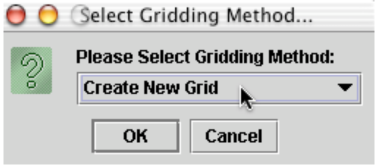
(4) Load Gene List (Control X)

Load the gene list, also under the Build Expression File menu. This should be a text file with suffix of ".txt" and be in MAGIC Tool format. (See full instructions below.)

(5) Locate Spots (Control A)

Under the Build Expression File, select Addressing/Gridding option. There are several distinct steps in Addressing and Gridding, which we will walk through one by one in the following paragraphs (a) – (i).

(a) Decide whether you want to create a new grid or load a saved grid.



Please Select Gridding Method: Create New Grid

Unless you have done this before, you will need to create a new grid.

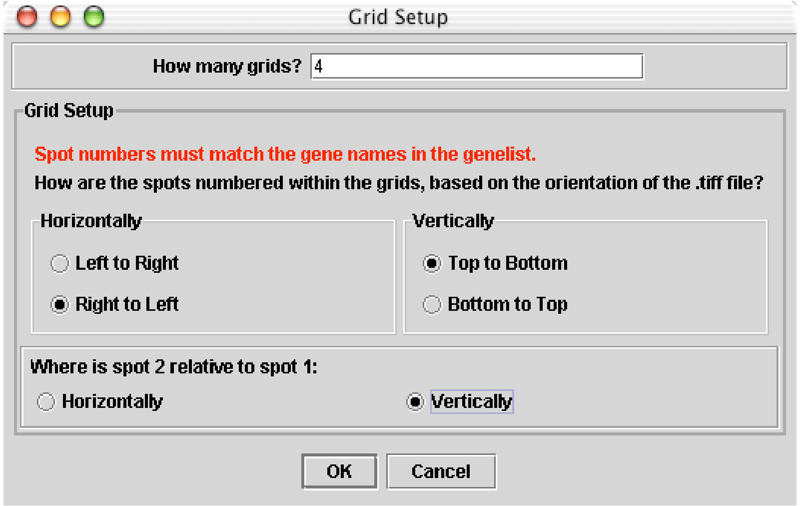
When you create a new grid, you will get a warning window that is normal and intentional. The warning is a reminder that you MUST understand how your spots are arranged on your microarray. For more information about this step, consult http://gcat.davidson.edu/GCAT/workshop2/addressing_MT.html

You must understand your genelist configuration before you start gridding.
First grid you create must contain spot number 1, corresponding to the first gene in the genelist.

Do not proceed any further if you do not understand the organization of your microarray.

Failure to perform Addressing and Gridding correctly will result in features being incorrectly identified.

You should see two windows. One will show your merged tiff files and the other will permit you to address the tiff file. The smaller (moveable) window will ask you information about how your microarray is organized; this is called addressing.


Grid Setup

How many grids? 4

Grid Setup

Spot numbers must match the gene names in the genelist.

How are the spots numbered within the grids, based on the orientation of the .tiff file?

Horizontally
○ Left to Right
● Right to Left

Vertically
● Top to Bottom
○ Bottom to Top

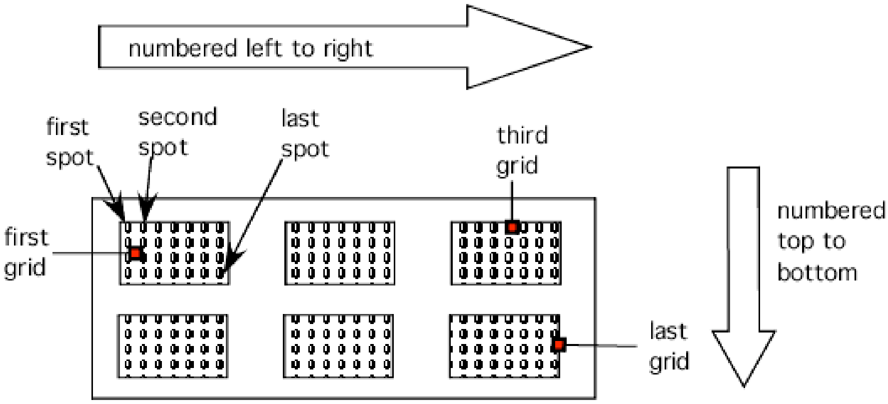
Where is spot 2 relative to spot 1:
○ Horizontally
● Vertically

***(b)* Answer the four questions in the Grid Setup window.**

First, enter the total number of grids on the tiff file.

Answering the remaining three questions is the easiest step to make a disastrous mistake. Answer the three questions based on **the way you are seeing your microarray at this time**. Here is an example to illustrate the point. Suppose the image has been rotated 90 degrees clockwise compared to the way you normally think about your chip, but your gene list is not altered to account for the rotation. Then the way you are seeing your tiff file will not match what you think of as your microarray organization. The following two images show the layout of the microarray before and after rotation.

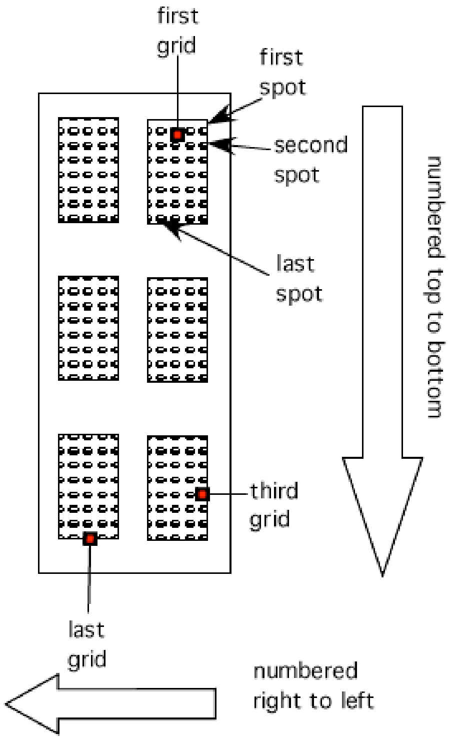
Before rotation, the spots would be described as being numbered from top to bottom and from left to right, with the second spot horizontal of the first spot (just like you would read a book). These are the default options. However, it is important that you keep track of the spots if the chip is rotated.

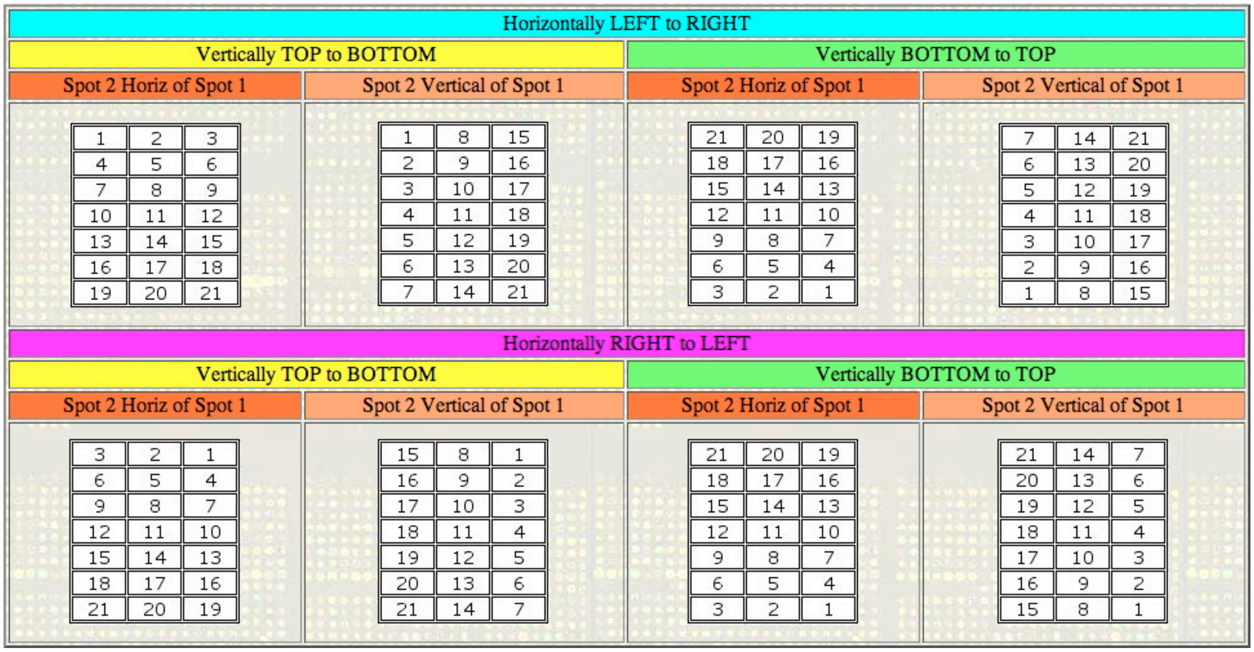


After rotation, the spots are numbered top to bottom, *right to left*, and the second spot is now *vertical* from (below) the first spot. Study the before and after rotation images, to understand how the spots have moved and why the new orientation resulted in the addressing provided in the figure above. Then study all the other options for numbering spots in the table below.

Use the pattern of missing spots and the comments in your gene list to help you become reoriented if necessary. The layout and number of grids is an easy way to orient yourself as well.

If you make a mistake, you can change your answers to these addressing problems by selecting “Grid properties…” under the file menu of the gridding window.





Horizontally LEFT to RIGHT

Vertically TOP to BOTTOM

Spot 2 Horiz of Spot 1

1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21

Spot 2 Vertical of Spot 1

1	8	15
2	9	16
3	10	17
4	11	18
5	12	19
6	13	20
7	14	21

Vertically BOTTOM to TOP

Spot 2 Horiz of Spot 1

21	20	19
18	17	16
15	14	13
12	11	10
9	8	7
6	5	4
3	2	1

Spot 2 Vertical of Spot 1

7	14	21
6	13	20
5	12	19
4	11	18
3	10	17
2	9	16
1	8	15

Horizontally RIGHT to LEFT

Vertically TOP to BOTTOM

Spot 2 Horiz of Spot 1

3	2	1
6	5	4
9	8	7
12	11	10
15	14	13
18	17	16
21	20	19

Spot 2 Vertical of Spot 1

15	8	1
16	9	2
17	10	3
18	11	4
19	12	5
20	13	6
21	14	7

Vertically BOTTOM to TOP

Spot 2 Horiz of Spot 1

21	20	19
18	17	16
15	14	13
12	11	10
9	8	7
6	5	4
3	2	1

Spot 2 Vertical of Spot 1

21	14	7
20	13	6
19	12	5
18	11	4
17	10	3
16	9	2
15	8	1

(c) Begin gridding.

The goal of gridding is to tell MAGIC where the spots within each grid are located. This feature is one of the best innovations in MAGIC Tool. Before you begin, you may want to adjust the contrast to help illuminate faint spots. To do this, slide the indicator that is currently pointing to 100% contrast near the top of this window. Adjusting contrast does NOT affect the raw data; it only allows you to see spots better for this step.

The number one tab should be selected as the default when you begin gridding. The tab numbers on the microarray correspond to the grid numbers. Selecting tab #1 indicates you are working with grid #1 (based on the gene list order). You may begin with a different grid if you wish, but be sure to keep straight where each grid is on the microarray. Again, if you do not follow this procedure of matching grid numbers with tab numbers, you will cause the features to be incorrectly identified. Grid #1 is the grid that contains spot #1, corresponding to gene #1 in the gene list.

(d) Center current grid in gridding window.

Scroll and zoom the image until you can see the first grid as defined by the gene list. To zoom in, click on the "Zoom In" button and then click on the grid where you want the zoom to center. Remember that spots and genes do not change their numbers with image rotation. In the example above where the image is rotated 90 degrees clockwise, the first grid would be the grid in the top right corner.

(e) Enter grid location information using "3-click" mouse method.

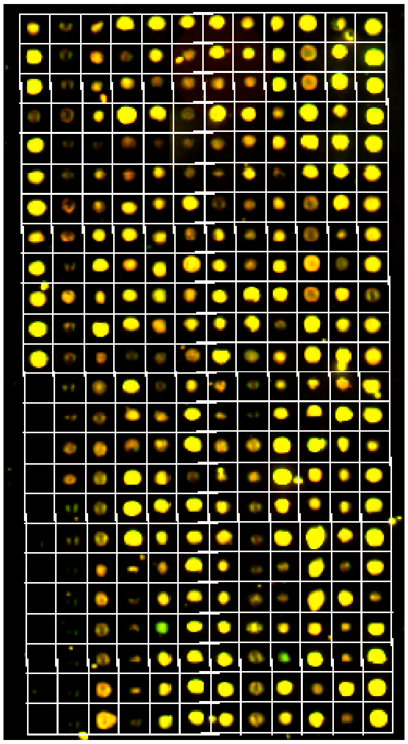
a. Click on the button that says "Set Top Left Spot" and then click on the center of the top left spot of the grid.

b. Click on the button that says "Set Top Right Spot" and then click on the center of the top right spot.

c. Click on the button that says "Set Bottom Row" and then click on the center of any spot in the bottom row. Choose a big round spot to make this step easier.

d. Enter the number of rows and columns. This is to be answered based on the way you are currently viewing the tiff file. In this example, there are 24 rows and 12 columns.

e. Click the "Update" button. At this time, you should see all the spots in the first grid surrounded by boxes as shown in the figure.



At any time in the gridding process, you can mouse over a spot and identify its location (x and y coordinates in pixels, row, column and spot number) as well as its identity from the gene list. This information is displayed in the bottom left corner and is especially useful for navigating during segmentation.

X:133 Y:353 Gene:YMR186W (Grid:1 Col:7 Row:18 Spot Number:162)

(f) Adjust the grid to center spots.

At this time, see if the spots look centered in the boxes. If not, then adjust the position of the boxes either by clicking on the appropriate button and then the correct spot, or by manually typing in numbers to adjust the boxes. Once you are familiar with the process, entering the numbers manually is much easier than using the mouse for fine-tuning. Note that the position of the mouse is displayed in the bottom left corner of the window so you can determine if the numbers should be bigger or smaller to shift the boxes in the correct direction. Gridding takes a bit of practice, but it is MUCH easier than most other methods for gridding.

(g) Define the next grid.

If you only have one grid, skip to step (i). If you have more than one grid, continue. Once the first grid is properly gridded (surrounded with boxes with the spots in the centers), it is time to repeat this process for grid #2. Be sure you know whether grid #2 is left, right, above or below grid #1.

a. Click on tab 2 at the top of the window.

b. Scroll and zoom to center grid #2 in the gridding window.

c. Click on the button that says “Set Top Left Spot” and then click on the center of the top left spot (or, if the top right spot is more visible, click the top right button and spot, instead).

d. As soon as you enter information for the top left or top right spot, you should see a box next to the phrase “Apply from grid ‘x’”. Click this box and then click on the “Update” button. The first time this option appears, ‘x’ should be 1, the only existing grid. Future grids can be applied from any existing grid, so you can select the previous grid that your current grid is most like. You can skip the other information (top right spot, bottom row, number of rows and columns). This should apply a set of boxes around the spots that nearly centers all spots in the boxes. Again, you can adjust to size and location of the boxes as needed by typing in X and Y values to resize and move the grid of boxes.

(h) Continue gridding.

Continue step 6 for each remaining grid on the microarray, so that all the grids on the microarray are boxed with the spots in the center of the boxes. At any time, you can change your answers to the four addressing problems by selecting “Grid properties…” under the file menu of the gridding window.

You may stop at any time and save your work so far, using the “Save Current Grid As…” under the file menu of the gridding window. Next time you begin Addressing/Gridding, you can simply open this saved grid file.

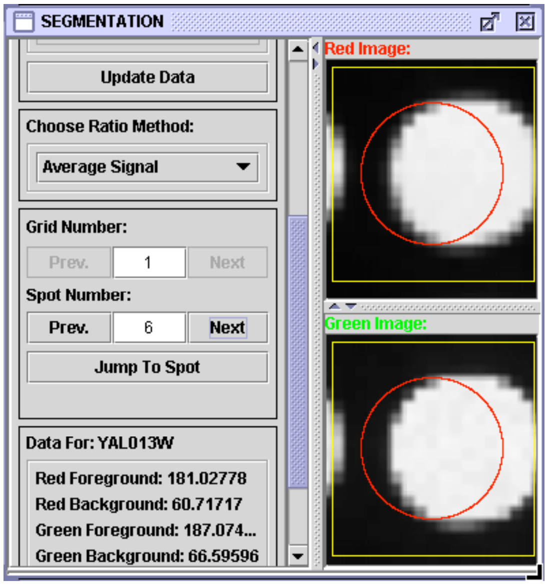
You can also save a snapshot of the combined tiff images at any time before or during the gridding process. You can save the image as tiff, jpg or gif. Tiff format works on all drawing and word processing programs so it is a universal format. Jpeg is good for images such as this that have many shades, like a photograph. Gif is the simplest format but may lose some of the subtlety of your original file. This saved merged image is useful if you want to take a picture of the overall grid and can be used for publishing or teaching.

(i) Complete the gridding process.

When you have finished gridding all your grids, click on the “Done!” button. If you have not already saved your grid, you will be prompted to do so before moving on to the next step. A grid file should be saved in your project folder and automatically given a suffix of “.grid” (so you do not need to type .grid yourself).

If the number of genes in your gene list and the number of spots you gridded do not match, you will get an error message. You must have exactly one grid square for each line (gene or gene replicate) in the gene list. If not, you probably will make an error identifying the spots later so you are required to fix this problem now. If your gene list and the number of gridded spots match, then you will be informed of the total number of spots and allowed to save the grid file for further use.

(6) Distinguish signal from background and generate expression file (Segmentation; Control S)



Choose Ratio Method: Average Signal

Grid Number: 1

Spot Number: 6

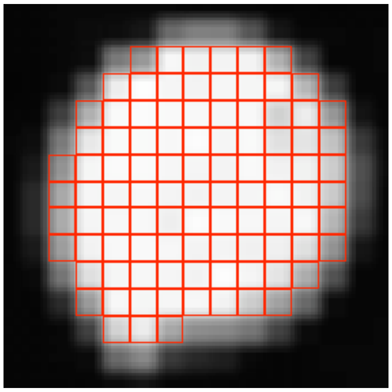
Data For: YAL013W

Red Foreground: 181.02778
Red Background: 60.71717
Green Foreground: 187.074...
Green Background: 66.59596

We will break this step into three parts, described in paragraphs (a) – (c).

(a) Select a method for distinguishing signal from background.

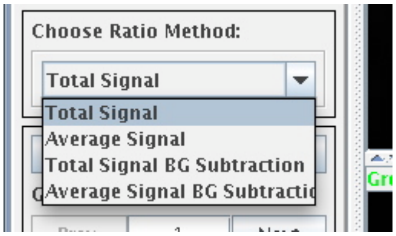
Fixed Circle: The most common way is to simply place a circle in the middle of the squares you drew for gridding. This is called *fixed circle*, though you can adjust the radius of this circle as shown in the figure to the right. Note that even if the circle is bigger than the box, only signal inside the box is used for measuring signal.



Adaptive Circle: The second method to choose from is the a*daptive circle*. The size and the location of the circle changes, depending of the size on the feature on the microarray. See the instructors guide for more details on this algorithm.

Seeded Region Growing: Seeded region growing is designed to find the signal for each spot based on the distribution of the signal. This method for segmentation looks for the brightest pixel near the center of the grid square, and then connects all pixels adjacent to this pixel and connects them into one shape. The algorithm simultaneously connects pixels to background and foreground regions, continuing until all pixels are in one of the regions. A user-specified threshold determines which pixels can be used to "seed" the regions. This is the slowest method since each pixel is processed individually. The bigger the threshold, typically the bigger the spot will be defined.

(b) Choose a Ratio Method



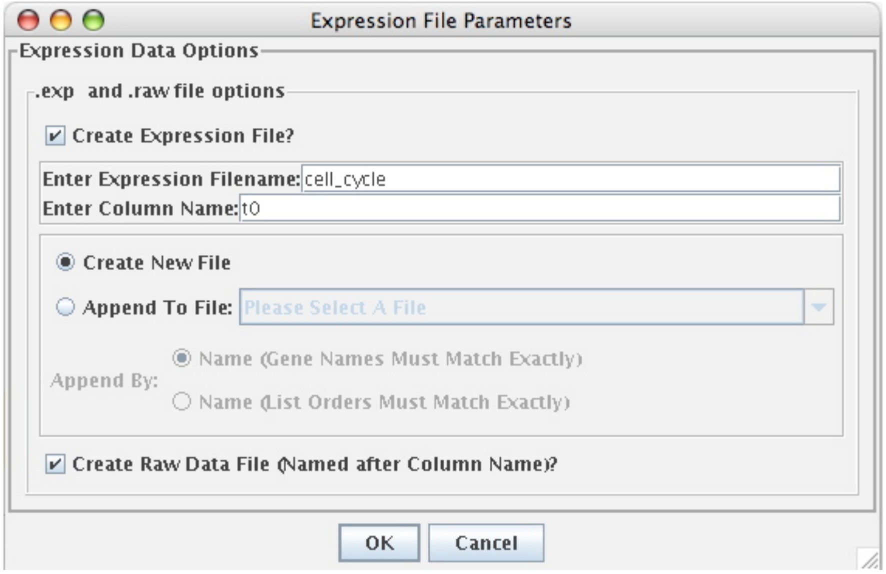
The final product of segmentation is a list of gene expression ratios. There are four choices for how to combine the four numerical values in segmentation (red foreground, red background, green foreground, green background) to determine a ratio for each feature on the microarray. Total signal adds the values in all the pixels designated as signal, and divides the red total by the green total. Average signal averages the values of signal pixels. The remaining two options subtract the background (total or average, respectively) before dividing the red by the green to get the ratio. Background subtraction introduces the possibility of a negative value (if background is greater than foreground). MAGIC Tool sets a negative value to 0. If background is greater than or equal to foreground in the green signal, this results in dividing by 0. In this case, MAGIC Tool resets the ratio to 998 or 999 (depending on whether the numerator of red foreground minus red background was also 0, or was greater than 0).

You can navigate around the spots, noting the summary of each spot's data below, to visually verify that the gridding and segmentation were performed adequately. This inspection gives you a chance to note any features you think should not be considered during subsequent data analysis.

(c) Generate expression file

Click on "Create Expression File" when you are satisfied with the segmentation process. This will generate an expression file, which was the goal of all the previous steps. An expression file contains the ratios for each spot (red ÷ green), according to the method chosen. MAGIC will ignore certain entries in the gene name column ("blank", "EMPTY", "missing" and "none"; case insensitive). The ratios will be used for all subsequent data analysis. You do not need the tiff files any more.

Unless you have already created an expression file for this microarray, you should check the box next to "Create Expression File?", and name the expression file and the column (e.g. time point, treatment, etc.). You can append this column to an existing file or create a new expression file consisting of this column only. MAGIC Tool will never erase one of your files, so if you append this column to an existing expression file, that file will remain as it was on your computer, and a new file will be created with the current column appended to the right of the columns in the existing file.



☑ Create Expression File?

Enter Expression Filename: cell_cycle

Enter Column Name: t0

● Create New File

○ Append To File: Please Select A File

Append By: ● Name (Gene Names Must Match Exactly) ○ Name (List Orders Must Match Exactly)

☑ Create Raw Data File (Named after Column Name)?

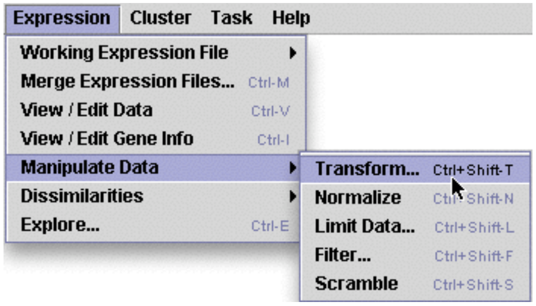
In the Expression File Parameters dialog box, you can also choose whether to save the "raw" data that was used to compute the expression ratios. If you check the box next to "Create Raw Data File," a tab-delimited text file will be created that contains 9 columns. The first column is the gene name. The next four columns are the pixel totals for red foreground, red background, green foreground, and green background. The final four columns are the pixel averages for these same four values. The raw data file will have the same name as your column label, with the extension ".raw". Your computer may think this is an image file, but it is just text. You can open the raw file from inside Excel (you may have to force it to look at files of all types for it to open). In future versions of MAGIC Tool, you will be able to use the raw data to filter your expression data, for example when signals are too weak to be reliable. In the meantime, this type of filtering must be done outside of MAGIC Tool.

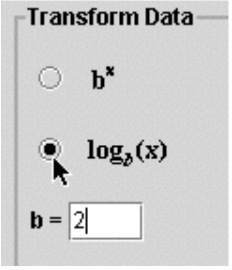
(7) Repeat steps 1-6 for all experimental conditions

If you have multiple time points or experimental conditions in your study, you should repeat steps 1-6 for each condition before continuing to the data manipulations of step 8. Once you have all data in one file, continue with the remaining steps. If you have only one condition, there will only be one column of data in your file, and you can do steps 8-11

(8) Manipulate Data

Although this step sounds like a point and click way to conduct scientific fraud, it is actually a beneficial step to consider (see Instructor’s Guide). You can: transform or normalize your data; temporarily restrict your data analysis to a subset of experimental conditions (e.g. certain time points, or dye reversals); filter out some features that don’t meet certain criteria; or generate a random set of data to use as a comparison.





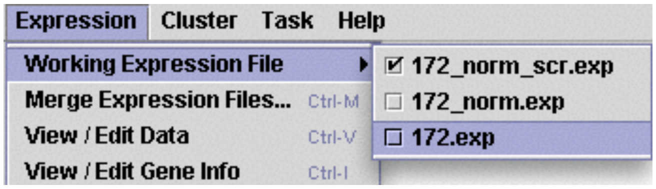
Transform Data

○ b^x

● $\log_b(x)$

b = 2

If you manipulate your data, you will generate a collection of new expression files with names that match the manipulation. MAGIC Tool will never erase your data, so the result of any of these data manipulations is stored in a new file, and the original file still exists as it was before the manipulation. Be sure to verify which expression file you are working with in subsequent steps. It is easy to get confused. The current file is checked on the list under “Working Expression File.”



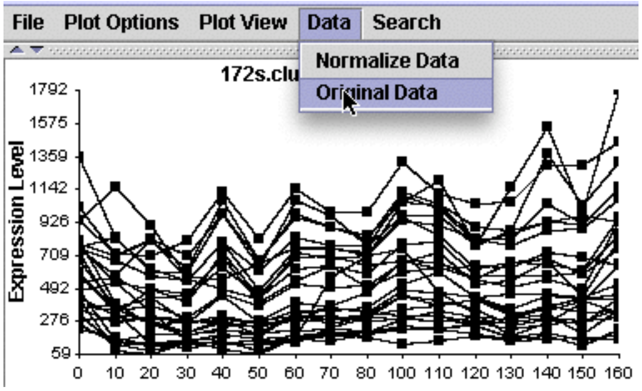
☑ 172_norm_scr.exp

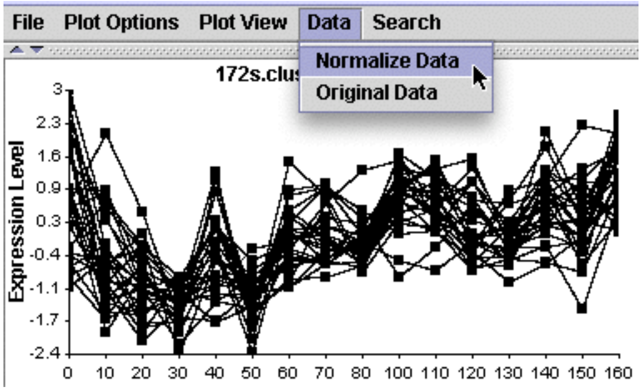
☐ 172_norm.exp

☐ 172.exp

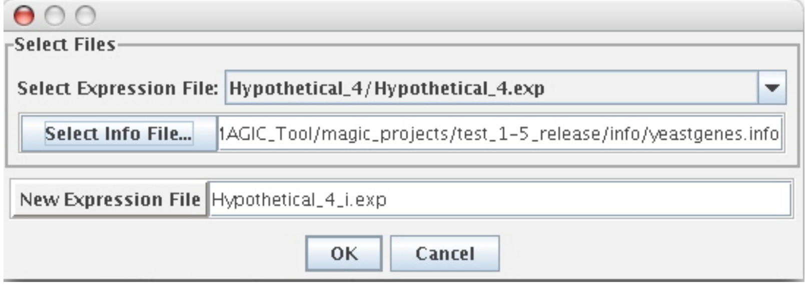
If you are working with ratio data, you should log transform your data. This will convert your ratios into values that are on the same numerical scale so that a gene that is 4 fold induced (+2) has the same numerical value as a gene that is 4 fold repressed (-2 instead of 0.25). Typically, this is done using a $\log_2$ transformation to indicate the number of two fold changes in gene expression (thus 4 fold changes resulted in numerical values of 2).

If you are working with absolute expression values (e.g. Affymetrix data) you may want to normalize your ratios. Normalizing in this case is on a gene-by-gene basis. For each gene, the mean value across the columns is subtracted from each value, resulting in an expression profile with a mean of 0. Then each value is divided by the standard deviation across the columns, resulting in an expression profile with a standard deviation of 1. This type of normalization is especially useful for viewing groups of genes on the same scale, so similarities are more easily seen when absolute expression levels vary greatly from gene to gene. Later, when you plot the various groups or clusters of genes, you can view the data in as normalized or original values, as shown in the following figure.





(9) Add gene info to expression file



Select Expression File: Hypothetical_4/Hypothetical_4.exp

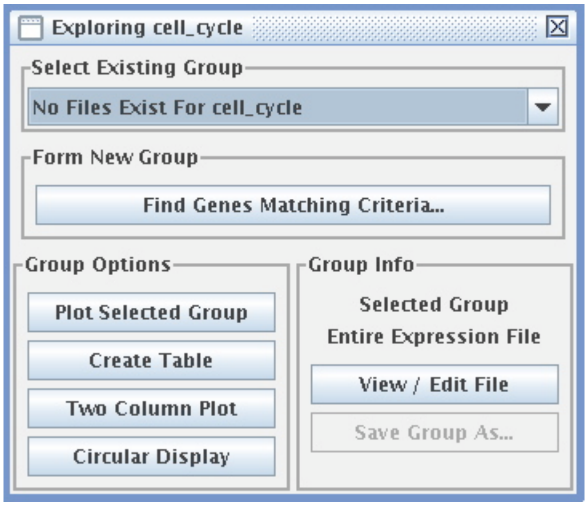
1AGIC_Tool/magic_projects/test_1-5_release/info/yeastgenes.info

Hypothetical_4_i.exp

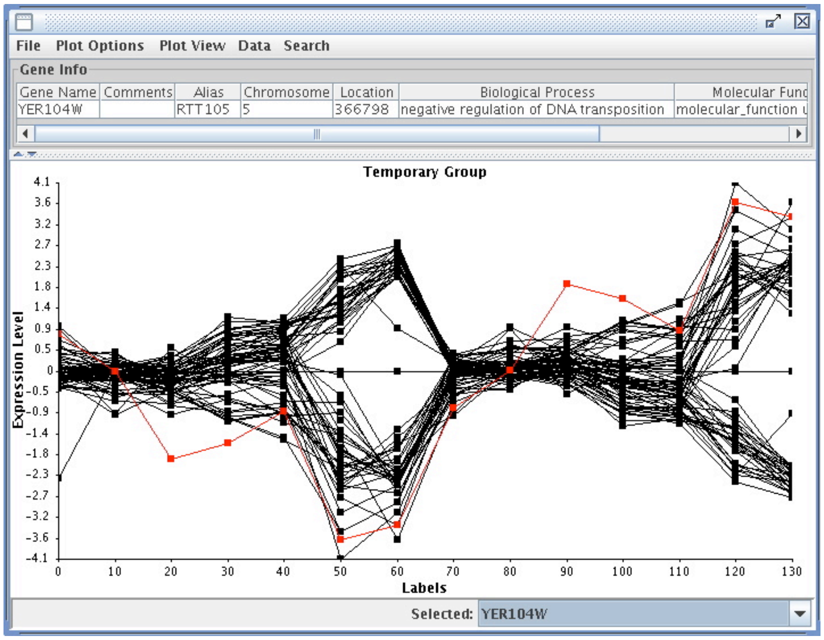
Now is the best time to add gene annotations to your expression file, so the annotations will be visible when you explore your data. Under the Expression menu, choose “Import Gene Info…” Select the expression file to which you wish to add annotations, and select the file containing the annotations. A file containing such annotations for yeast is included in the sample files. A similar file can be formed for any organism by creating a tab delimited text file with the appropriate columns (alias, chromosome, location on chromosome, biological process, molecular function, and cellular component).

(10) Explore data

Data exploration is a way to get familiar with your data, and find important functional relationships that may not be apparent from clustering. For example, you can find all genes that were upregulated after a certain time point, or all genes that increased their fold repression four times or greater at any time point. Once you have identified such genes, you can display them in a number of dynamic ways and save these images for publishing or teaching.



The top option in the Exploring window is to Select Existing Group, but if you have not explored the file before and saved group files, there will not be any existing files. You can create a temporary group by clicking "Find Genes Matching Criteria…" and filling out the form to find the genes and expression patterns you are interested in. If you want a group to be available the next time you explore your data, and even the next time you open this project, you need to save the group file (which will automatically be given an extension of ".grp". A group file is just a text file that lists the names of the genes in the group.



Gene Name	Comments	Alias	Chromosome	Location	Biological Process	Molecular Func
YER104W		RTT105	5	366798	negative regulation of DNA transposition	molecular_function

Selected: YER104W

Each of the displays on the left hand side of the Exploring window gives you a different visualization of your data. The "Plot Selected Group" display is shown here, with gene YER104W highlighted. Note that the annotations of this gene can be revealed above the plot of the group. This group was formed by finding all genes whose minimum value was less than -2. Interestingly, the group seems to consist of two distinct sub-groups: genes that are upregulated early and downregulated later, and genes that have the opposite profile.

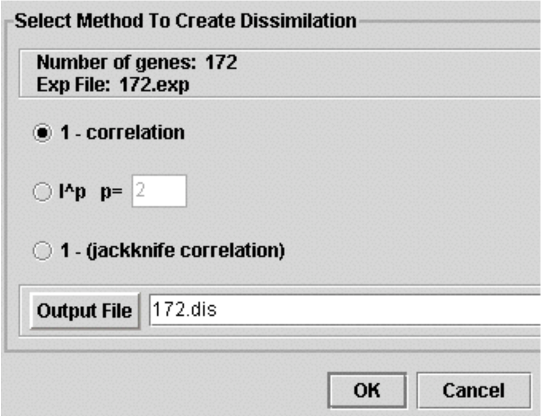
(11) Filter data

You should filter your data to remove uninteresting genes before proceeding to the next steps. For example, you might keep only those genes whose expression pattern has a sufficiently large standard deviation across the columns (in other words, whose expression is not constant). Or you might remove genes with unreliable ratios (including those involving a division by 0). It is important that your expression file be as small as possible, without losing important information, before beginning the clustering process.

(12) Calculate correlation coefficients

To form clusters of similar genes you need a way to compare the expression profiles of different genes. In this step, you will generate a huge table of "dissimilarities," measuring the difference between every pair of gene expression patterns. This step can take a very long time for a large number of genes. Be sure you have filtered your data suitably, and that you know you will learn something from the clustering process before you begin this step.

Under the Expression menu, choose “Dissimilarities” and then “compute”. When you do this, a window will appear where you have to choose from three choices. This is another decision that will affect the data analysis.


Select Method To Create Dissimilation

Number of genes: 172
Exp File: 172.exp

● 1 - correlation
○ l^p p= 2
○ 1 - (jackknife correlation)

Output File: 172.dis

The most common method is the default, which is 1 – correlation. The other two methods are described in the Instructor’s Guide. When this step is complete, MAGIC Tool generates a dissimilarity file, which you can name in the output file box. The file will automatically be given the suffix “.dis”. Click on OK to begin the computation process. The progress is monitored in a popup scale bar (not shown here). You can calculate dissimilarities on any expression file (.exp) but you should use your transformed ratios rather than non-transformed ratios. You can also use transformed and normalized expression files containing absolute expression values.

(13) Cluster genes

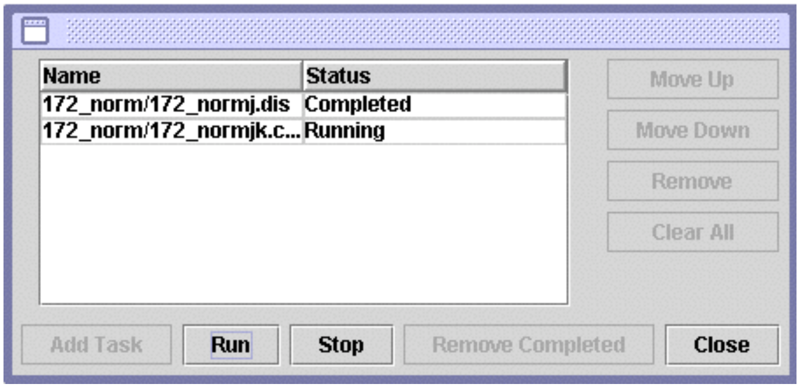
At this point, you can generate a series of clusters using four different methods. Clustering is a very popular process for DNA microarrays, so we will describe this first, but remember that exploration is equally valid and may tell you more about your genes and experimental conditions than clustering can. Exploring your data can be performed any time after segmentation. All you need to explore are expression files (*.exp).

With MAGIC Tool, there are four ways to cluster genes. You can cluster from any dissimilarity file. First you have to calculate the clusters and then you can display them in a variety of ways. The most common way to cluster is called hierarchical clustering, which you can do with MAGIC. However, we prefer Q-T clustering (see Instructor’s Guide for details). You can also cluster by k-means or supervised clustering.

Once you have clustered the genes, you can display the results in several ways. MAGIC allows you to view these clusters in a variety of dynamic displays. Each display can be saved as an image file for publishing or teaching. Display options are addressed in more detail later in this manual.

Automating Tasks

As your datasets get bigger, the time it will take to make all the necessary calculations will increase rapidly. Therefore, MAGIC Tool allows you to establish a list of tasks to be performed in sequence. You can tell MAGIC Tool to begin a series of steps and then walk away from your computer. MAGIC Tool will perform this sequence of tasks while you do other things. For example, you can establish a list of tasks to perform and go home for the night. When you return the next morning, MAGIC will have completed the series of tasks.



Name	Status
172_norm/172_normj.dis	Completed
172_norm/172_normjk.c...	Running

Closing Comments

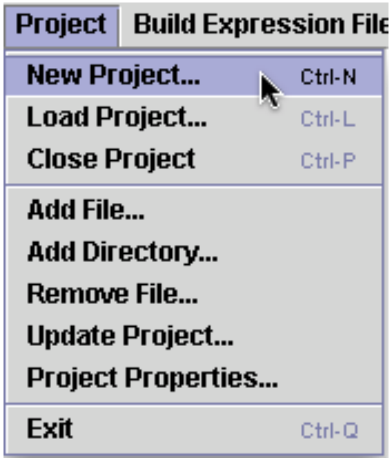
This section was intended as a way to get you launched into the MAGIC Tool way of working with DNA microarrays. MAGIC allows you to compare the consequences of different choices for quantifying, comparing and clustering the same raw dataset. This capacity to compare methods is a powerful way to understand better the assumptions and implications inherent in data analysis as published each week. MAGIC allows you to explore data and data analysis during the early days of DNA microarrays when the research community has not settled upon standards for comparing results. MAGIC was designed to empower the user and make DNA microarrays more approachable for a wider audience. In the following section, every option available in MAGIC Tool will be spelled out so you can utilize the full potential of MAGIC Tool.

Complete List of MAGIC Tool Options

Project Menu

New Project (Control N)

This begins a new project. A project is a way of organizing all related MAGIC Tool work in a folder. The name you give to the project is the name of the folder, and the folder is automatically created by MAGIC Tool. Each project name should be unique and descriptive. Within the folder created by MAGIC Tool will be a file that ends with the suffix “.gprj”. All subsequent steps and files will be stored automatically in this project folder, until you start another new project. The .gprj file is a text file that is essentially a table of contents of your project.



Load Project (Control L)

This allows you to reopen a previous project. Navigate to the location of the project on your hard drive, and select the .gprj file within the project folder.

Close Project (Control P)

Allows you to stop a project without quitting MAGIC Tool completely. You can also stop a project by opening a new project and confirming that you wish to close the currently open project.

Add File….

Allows you to add files (e.g. tiff files, gene lists, info files, existing expression files) from other projects to your current project. You will be directed to a window from which you can click your way through the hard drive in search of the files you want to add. Holding down Shift and clicking allows you to select a consecutive range of files. (On Windows, you can hold down the control key and click on multiple files to select them.)

Add Directory…..

Alows you to add all files in the selected folder to your current project.

Remove File….

Lets you remove unwanted files from your current project folder. You can also delete files by writing over the older version (you will be prompted to verify you want to write over the existing file with the same name).

Update Project….
Allows you to drag files into existing folders and then update the currently active project. This allows the user to quickly move tiff, grid, expression, dissimilarity, and cluster files around and then utilize them in different projects.

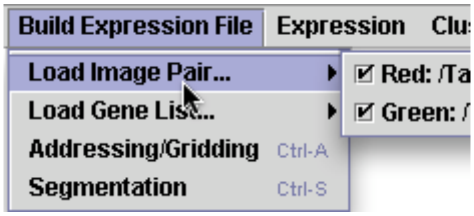
Project Properties…
Allows you to modify the default properties and configure the behavior of MAGIC Tool. There are three tabs, each containing properties of different types.

> **Data Handling:** Currently the only data handling option is how to handle missing data. You can choose to *remove* or *ignore* any genes in your current project that are missing data. When a DNA microarray is printed, some features will be missing and therefore you cannot collect data for this gene. If you choose to *remove* all genes missing data, then genes missing any data from one or more columns will not be used for calculating dissimilarities. If you choose to *ignore*, you will be prompted for what percentage of possible data (in percent) must be available for a gene to be included in your data analysis. This allows you to work with genes that are missing data from less than that percentage of columns. Genes missing more than the input threshold percentage of columns will be removed.
>
> **Image Saving:** Controls maximum image size saved from MAGIC Tool
>
> **Group Files:** There are two options under this tab. The first, "New Expression Files Carry Group Files When:" controls how group files go along with expression files. This option comes into play whenever you create a new expression file from an existing expression file, for example by log-transforming, adding information, filtering, normalizing or limiting data. Since a group file is simply a list of genes, you may wish groups that you selected based on values in an earlier version of the expression data to be accessible after you do one of the above processes to create a new expression file. The default setting is Always, meaning all group files are copied to the folder containing the new expression file. You can also choose to never copy group files, or to only copy the group files when the expression data itself was not changed (e.g. when adding info to the expression file).

Exit (Control Q)
This quits MAGIC Tool. All completed steps and files will be saved in your project folder. Steps only partially completed will be lost. Open tiff files will not be reopened when the project is opened next.

Build Expression File Menu

Load Image Pair…. (Control R and Control G)



This allows you to browse your hard drive to find the tiff files for the two colors. You can load the two tiff files in either order. If you have added files to the project, or moved files into the project folder and updated the project, all tiff files will be located in the Images folder of the project. Otherwise, you can navigate to the location of the files on your hard drive. Just be sure to match the colors and the files. Remember that red is a longer wavelength than green.

Load Gene List… (Control X)

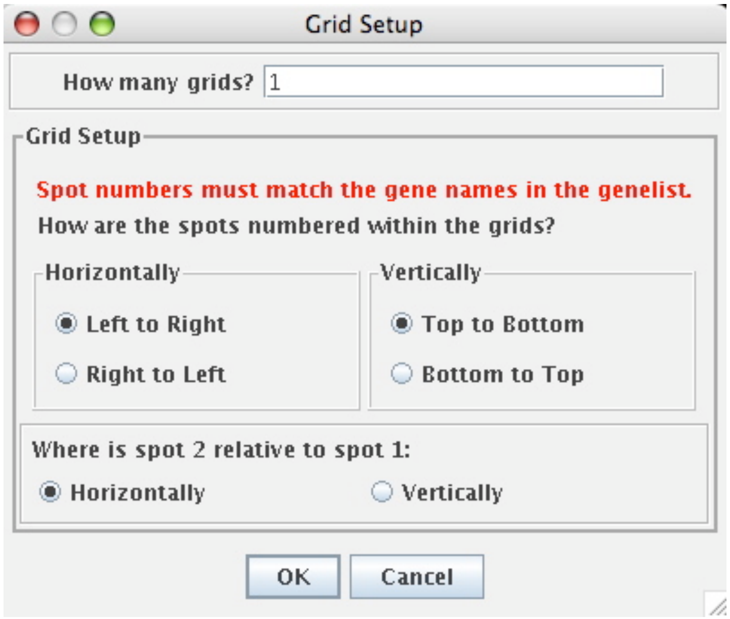
Reads a file that associates each feature on the microarray with a gene name. MAGIC Tool requires you to have this file, called the gene list, in a particular format. Gene lists in MAGIC Tool format are available for downloading from the GCAT and MAGIC Tool web sites, and are included in the Sample_Files folder of the MAGIC_Tool_1-x folder.

Often, non-MAGIC Tool formatted gene lists have additional information such as which features did not print, alternative names for the gene, etc. You can open your gene list to see what information it contains. If it contains information about the plates and wells for each gene, this is not useful information for MAGIC but was used to help the people who printed the chips to keep track of what they were doing during the manufacturing of the chips.

If you have a gene list that is not in MAGIC Tool format, you can use these instructions, and examples at http://www.bio.davidson.edu/people/macampbell/ACS_MAGIC/genelists.html to create a gene list with the proper format. First, open your gene list from inside Excel. Find the column that contains ORF names such as YBL023c or YAR002W, etc. Copy this ORF column and paste it in the first column (you may have to create a new column to hold this information). Next, remove all header rows, so that the first row in your file is the first gene in the list. Save the modified file as a tab-delimited text file, with a new name that ends with the suffix ".txt". This file is now a valid MAGIC Tool gene list. Although it takes a bit of manual labor to create this MAGIC gene list, it allows the user to quickly adapt to different microarray production styles. Later, you will learn how to import additional information about genes from commonly studied organisms.

Addressing/Gridding (Control A)

When you begin the addressing and gridding process, you should first see a merged image of your red and green tiff files, and where red and green are superimposed, you should see a shade of yellow. Then you will be asked four questions that tell MAGIC Tool how the spots are numbered, shown in the snapshot below. This step, called *Addressing,* is the easiest one to make a mistake on, so be very careful when answering the four questions as they appear in the window. It is vital you understand how your spots are organized on the microarray and in the gene list. All questions should be answered according to the way you see the merged image of your microarray in the viewing window. Are the genes printed in duplicate? If so, are the duplicate spots horizontal or vertical? You will need to know how many grids there are as well as the order of the spots in your gene list compared to the image in MAGIC Tool. The default answers to the Grid Setup questions correspond to the way you would read a book: left to right, top to bottom, with the second spot horizontal of the first one. It cannot be overemphasized how critical this step is. If you get this part wrong, you will not know the correct identity of any of the spots. Once you press OK, you have finished the Addressing step, but you can always choose File, Grid Properties in the Gridding window to get another chance to answer the four questions.



How many grids? 1

Grid Setup

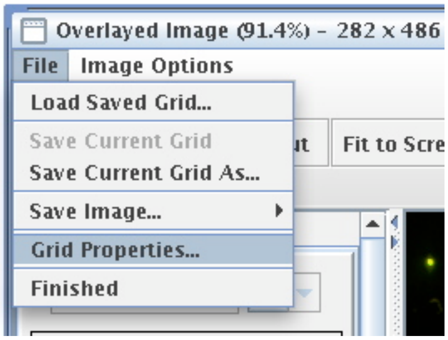
Spot numbers must match the gene names in the genelist.

How are the spots numbered within the grids?

Horizontally: ● Left to Right ○ Right to Left

Vertically: ● Top to Bottom ○ Bottom to Top

Where is spot 2 relative to spot 1: ● Horizontally ○ Vertically



Gridding is much easier. The purpose of gridding is to draw little boxes around each feature so the spots are in the center of the boxes. You may find it helpful to zoom in on the first grid of spots. To zoom in, click on the “Zoom In” button and then click where you want the zoom to center. The number one tab should be selected as the default.

Navigate the image until you can see the first grid as the one you know to be the first grid in the original layout of your microarray. If you want, you can adjust the contrast to help illuminate faint spots. To do this, slide the indicator that is currently pointing to 100% contrast near the top of this window. If the maximum value of the slider is still not enough contrast, you can adjust further by typing the percentage contrast you want in the box next to the slider. Adjusting contrast does NOT affect the raw data; it only allows you to see spots better for this step.

To grid, you simply click on three spots. First, click on the button that says “Set Top Left Spot” and then click on the center of the top left spot. Second, click on the button that says “Set Top Right Spot” and then click on the center of the top right spot. Third, click on the button that says “Set Bottom Row” and then click on the center of any spot in the bottom row. Choose a good spot to make this step easier. Enter the information for the number of rows and columns. Rows and columns are defined based on the way you are currently viewing the tiff file. To finish this grid, click on “Update” button. At this time, you should see all the spots in the first grid surrounded by boxes as shown to the right. (You may need to zoom out to see the full grid.)

At this time, see if the spots look centered in the boxes. If not, then adjust the position of the boxes either by clicking on the appropriate button and then the correct spot, or by manually typing in numbers to adjust the boxes. Note that the position of the mouse is displayed in the bottom left corner of the window so you can determine if the numbers should be bigger or smaller to shift the boxes in the correct direction. This step takes a bit of practice, but it is WAY easier than most other methods for manual gridding, gives you more control and understanding of the process.

Once the first grid is properly gridded, it is time to repeat this process for grid number two. Click on tab 2 at the top of the window. You may need to zoom in again, and scroll over or down to center the second grid in the screen. Click on the button that says “Set Top Left Spot” and then click on the center of the top left spot. At this time, you should see a box next to the phrase (apply from grid 1). Select this box and then click on the “Update” button. This should apply a set of boxes around the spots that is close to what you wanted to do, essentially a shifted copy of the first grid. Again, you can adjust this second grid of boxes as needed. Continue this process until all the grids are surrounded with the boxes.

You can save your current grid at any time, using File, Save Current Grid (or Save Current Grid As… to save under a different name). Grid files are automatically given a suffix of “.grid”. You can close the gridding window without saving, and the current grid will automatically be restored the next time you open the gridding window (without asking the four questions again). If you close the project, however, you must save your grid before you close the project, and choose the option Load Saved Grid when you begin gridding next time. This lets you pick back up where you left off with gridding.

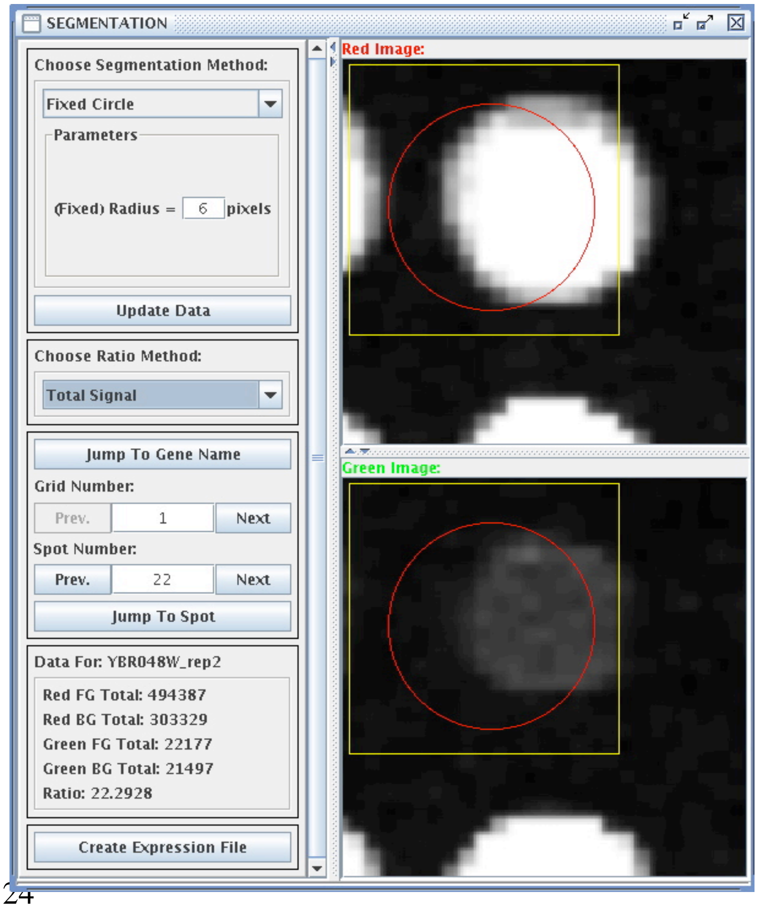
When you have finished gridding all the grids on the microarray, click on the “Done!” button. If you have not already saved your grid, you will be prompted to do so before moving on to the next step. If the number of genes in your gene list and the number of spots you gridded do not match, you will get an error message. You must have exactly one grid square for each line (gene or gene replicate) in the gene list. If not, you probably will make an error identifying the spots later so you are required to fix this problem now. If your gene list and the number of gridded spots match, then you will be informed of the total number of spots and allowed to save the grid file for further use.

You can take a snapshot of the combined tiff images, before, after, or during the gridding process. You can save an image of whatever is currently showing inside the gridding window, in tiff, jpg or gif format. (Tiff format works on all drawing and word processing programs so it is a universal format. Jpeg is good for images such as this that have many shades, like a photograph. Gif is the simplest format but may lose some of the subtlety of your original file.) This saved merged image is useful if you want to document your gridding process and can be used for publishing or teaching.

Segmentation (Control S)

Segmentation is the process of distinguishing signal from background. There are three methods available for this process. During segmentation, you will have the opportunity to view each feature on the entire microarray, one at a time. In this step, the two tiff files are separated again, with the red image on top and the green image on bottom. There are three algorithms available in MAGIC Tool for finding the foreground (signal) and background (noise) in each channel (red and green) separately. In addition, there are four choices for how to combine these four numerical values to determine the ratio.

You might want to experiment with the different algorithms and choices before settling on the best method. By browsing from spot to spot, or jumping to potential problem spots you noticed while gridding, you can see how these choices will affect the final answer. When you are satisfied with your choices, hit the “Create Expression File” button, and you will be prompted for a file in which MAGIC Tool will save all the ratios, one for each feature on the microarray. When you save the whole list, all values are recomputed, so it does not matter if you have browsed two spots or two hundred. In addition to saving the list of ratios, you will be given the opportunity to save “raw data,” i.e. all foreground and background values in the red and green channels.



SEGMENTATION

Choose Segmentation Method: Fixed Circle

(Fixed) Radius = 6 pixels

Choose Ratio Method: Total Signal

Grid Number: 1

Spot Number: 22

Data For: YBR048W_rep2

Red FG Total: 494387
Red BG Total: 303329
Green FG Total: 22177
Green BG Total: 21497
Ratio: 22.2928

Fixed Circle

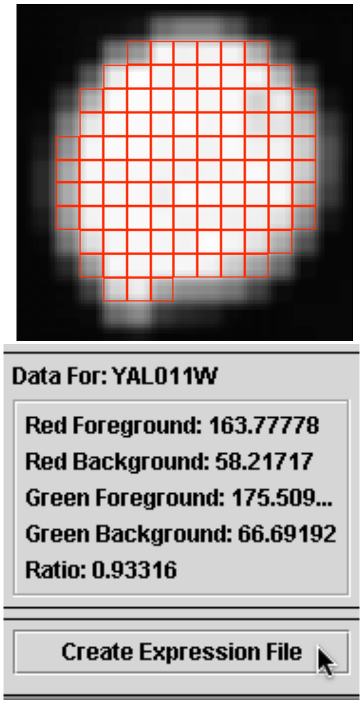
Fixed circle simply places a circle in the middle of the box. All pixels inside the circle (that are also inside the box) will be considered signal and pixels outside the circle (but still inside the box) will be background. You can set the radius of the circle in pixel units. In the above figure, you can see the features are in the box, but they are not centered. The foreground and background values of spots that are off center and spots that are bigger or smaller than the selected fixed radius will not be exactly right. However, the ratio between the red and green values should still be fairly accurate. Fixed circle is the most common method for segmentation, and is the fastest of the three segmentation methods.

Adaptive Circle

This method changes the center and radius of the circle to fit the size and location of each feature. The algorithm considers all pixels above a user-specified threshold to be "on," and finds the circle with the highest percentage of pixels that are on. The radius can range between a user-specified lower and upper bound; the center can be anywhere inside the grid box. This method is slower than Fixed Circle, but generally covers the actual spot better.

Seeded Region Growing

This method for segmentation is designed to find the signal for each spot based on the distribution of the signal. Seeded region growing looks for the brightest pixel and then connects all pixels adjacent to this pixel into one shape. The algorithm simultaneously connects pixels to background and foreground regions, continuing until all pixels are in one of the regions. A user-specified threshold determines which pixels can be used to "seed" the regions. This is the slowest method since each pixel is processed individually.

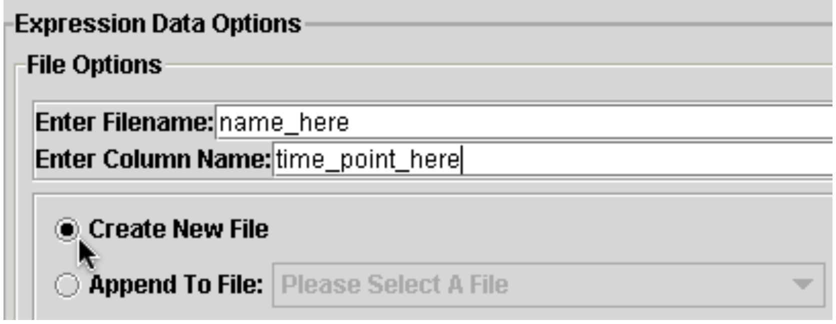


Data For: YAL011W

Red Foreground: 163.77778
Red Background: 58.21717
Green Foreground: 175.509...
Green Background: 66.69192
Ratio: 0.93316

Regardless which method you choose, you can visually inspect the features to verify the gridding and segmentation were performed adequately. This inspection gives you a chance to flag any features you think should not be considered during subsequent data analysis.

When you complete segmentation, you will produce an expression file. Click on "Create Expression File" when you are satisfied with the segmentation process. This will generate an expression file, which was the goal of the first half of MAGIC Tool. An expression file contains



Enter Filename: name_here
Enter Column Name: time_point_here

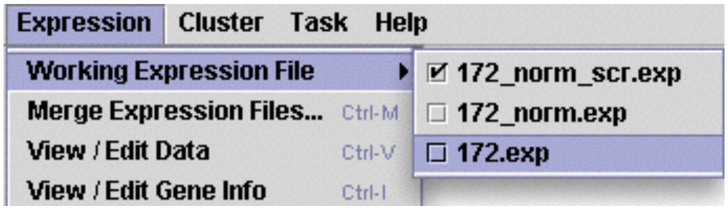
● Create New File
○ Append To File: Please Select A File

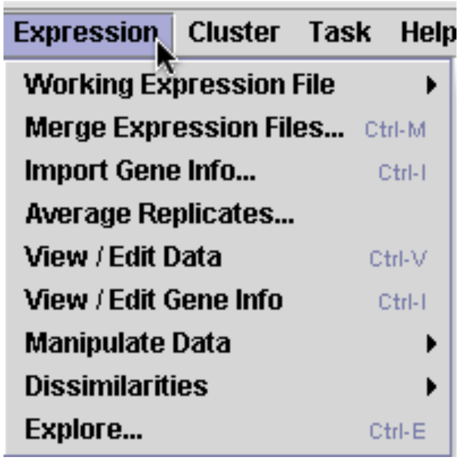
the ratios for each spot (red ÷ green). MAGIC will ignore certain entries in the gene name column (“blank”, “EMPTY”, “missing” and “none”; case insensitive). The ratios will be used for all subsequent data analysis. You do not need the tiff files any more.

You will need to name the expression file and the column (e.g. time point, treatment, etc.). You can append this to an existing file or create a new one. You can also save raw signal and background intensity levels.

Expression Menu

Working Expression File

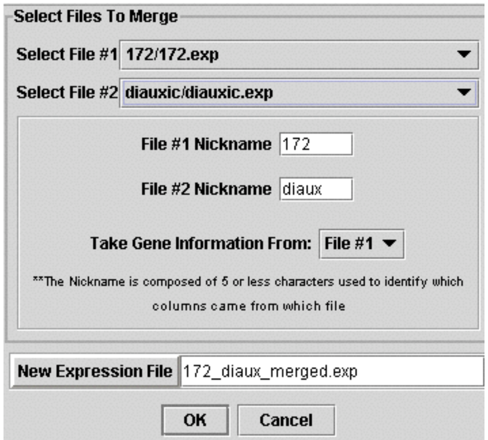




This option allows you to choose from a range of expression files within a single project. As you can see from the image on the left, you can choose which one is active simply by clicking on it.

Merge Expression Files… (Control M)

Merging expression files allows you to combine data from multiple chips so you can evaluate time course data, or other related data sets. You merge files one at a time and provide nicknames to assist MAGIC in keeping track of the soon to be combined data. Also, you can select one set of gene annotations as the one that is retained with the merged data set. A new file will be created, so your two original files are not lost.


Select Files To Merge

Select File #1 172/172.exp

Select File #2 diauxic/diauxic.exp

File #1 Nickname 172

File #2 Nickname diaux

Take Gene Information From: File #1

**The Nickname is composed of 5 or less characters used to identify which columns came from which file

New Expression File 172_diaux_merged.exp

Import Gene Info… (Control I)

This allows you to compile more complete information about your ORFs. For example, we have created a text file that describes the chromosomal location, the three categories of gene ontology annotation, and synonym for all yeast genes. This permits you to search by each of these fields to help detect trends and meaningful information.

Average Replicates

MAGIC Tool treats every spot as a unique feature and does not average for replicate genes automatically. This preserves all your original ratio data. If a set of feature names are identical in the gene list, MAGIC Tool will give each replicate a unique name by appending _rep1, _rep2, etc. After you have created expression files, you may choose to average replicate spots as defined by ORF name. When you average replicates, all features with identical names (disregarding _rep#) then the data will be averaged.

View/Edit Data (Control V)

After an expression file is created or merged, you can view and edit the data. This option should not be used often, but we did want you to have access to the ratio data if you deem it necessary. It is helpful if you want to verify steps or pick up a project after an extended period of time.

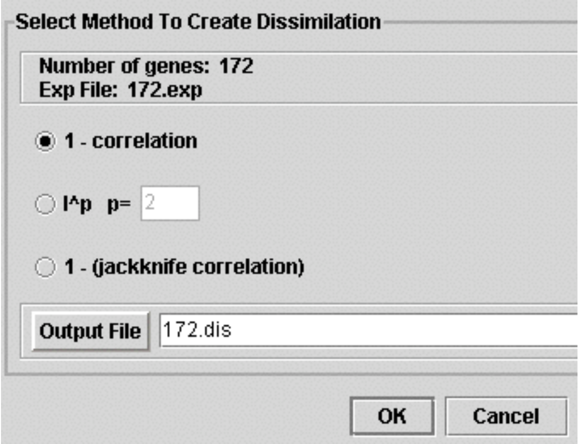
View/Edit Gene Info (Control I)

This option allows you to view and modify the gene annotations. Of course, you can view and edit this file outside MAGIC Tool, but this option provides you an opportunity to do so within MAGIC. Perhaps you will want to perform a search on the gene function. Viewing the list can allow you to select appropriate terms for searching.

Dissimilarities (Control D)

Calculating dissimilarities allows you to compare different genes to one another. When you do this, a window will appear where you have to choose from three options.

The most common method is the default 1 – correlation (see Instructor's Guide for a detailed explanation of this and the other two methods). When this step is complete, MAGIC generates a dissimilarity file which you can name in the output file box, automatically given the extension ".dis". Click on OK to begin this process. The progress is monitored in a popup scale bar (not shown here).



Select Method To Create Dissimilation

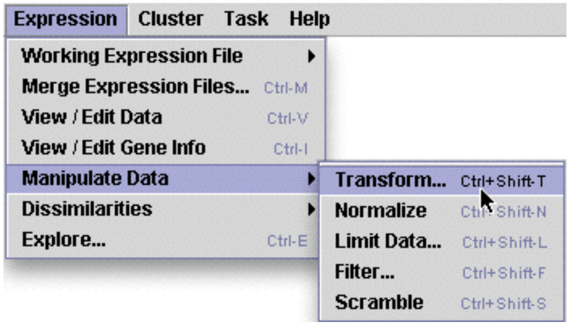
Number of genes: 172
Exp File: 172.exp

● 1 - correlation
○ l^p p= 2
○ 1 - (jackknife correlation)

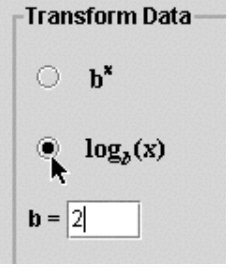
Output File 172.dis

Manipulate Data

Manipulating data is not as bad as it sounds. This option allows you to choose from five options. These options do NOT alter your original data, they simply allow you to process the data further prior to clustering or exploring your data.



Transform (Control Shift T)



○ b^x

● $\log_b(x)$

b = 2

A standard process you should perform is transforming your data before performing any analysis (exploring or calculating dissimilarities and clustering). You want to log-transform your ratios so you eliminate any fractions. It is important to get all ratios on the same scale of magnitude. For example, if a gene is repressed 16 fold, the ratio will be 0.0625 while a gene that is induced 16 fold will have a ratio of 16.0. Before analyzing your data, you should log-transform your data. After transformation (typically $\log_2$), the two genes would be altered (-4 vs. +4) with equal magnitude but in opposite directions. See Instructor's Guide for more information. You should explore after transforming, but may or many not want to normalize before exploring (see below). If you want to "un-transform" your transformed data, you can use the exponent function b^x.

Normalize (Control Shift N)

This process takes your (transformed) ratios and corrects for the magnitude of a gene's ratios and the variation among each gene's ratios. Normalization is not appropriate for ratio data, but is useful for absolute expression values. See Instructor's Guide for more details.

Limit Data (Control Shift L)

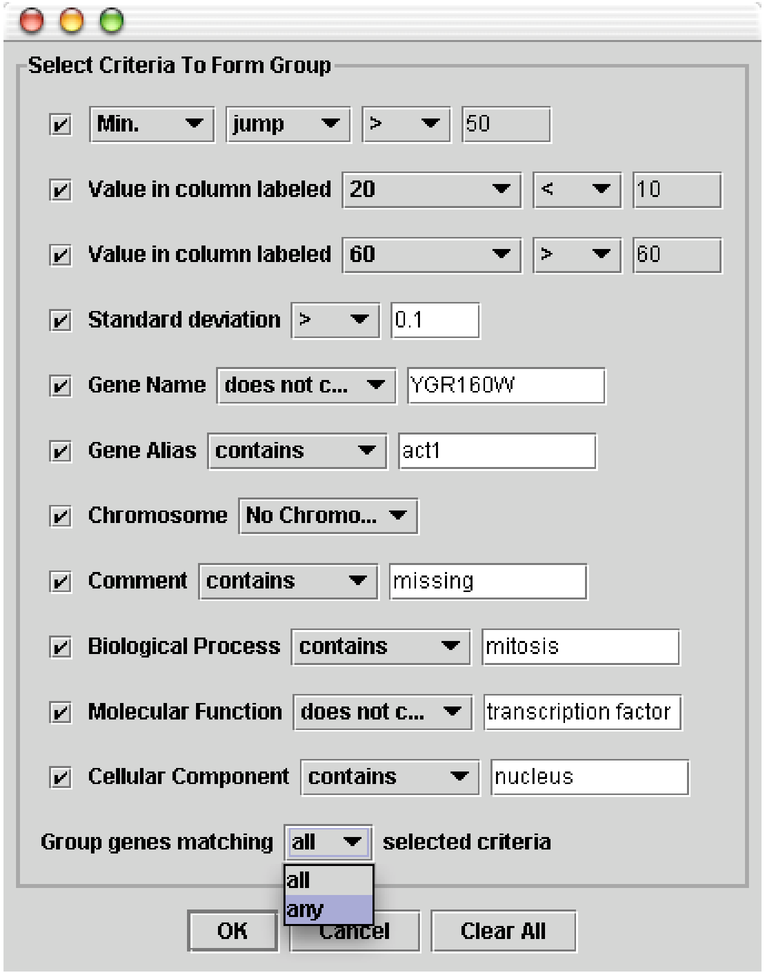
If you have merged data from many microarrays (e.g. a time course experiment), you may want to study only certain portions of your merged data independently. Limiting data allows you to select column headings and retain these selected data for analysis in a "limited data set". Your original merged file is left unaltered and a new file is created. The new expression file will terminate with the name "x_limted.exp" where x would be the original expression file name.

Filter (Control Shift F)

Filtering allows you to remove from further consideration genes that do or do not meet user-defined criteria. The image below shows you the types of filters that can be applied.

Scramble (Control Shift S)

Gives three different methods for creating a gene expression file with the same exact numbers as your current file, but in random order. Randomization can help indicate whether the patterns found through exploration and clustering are real effects of the experimental conditions.



Select Criteria To Form Group

☑ Min. jump > 50

☑ Value in column labeled 20 < 10

☑ Value in column labeled 60 > 60

☑ Standard deviation > 0.1

☑ Gene Name does not c... YGR160W

☑ Gene Alias contains act1

☑ Chromosome No Chromo...

☑ Comment contains missing

☑ Biological Process contains mitosis

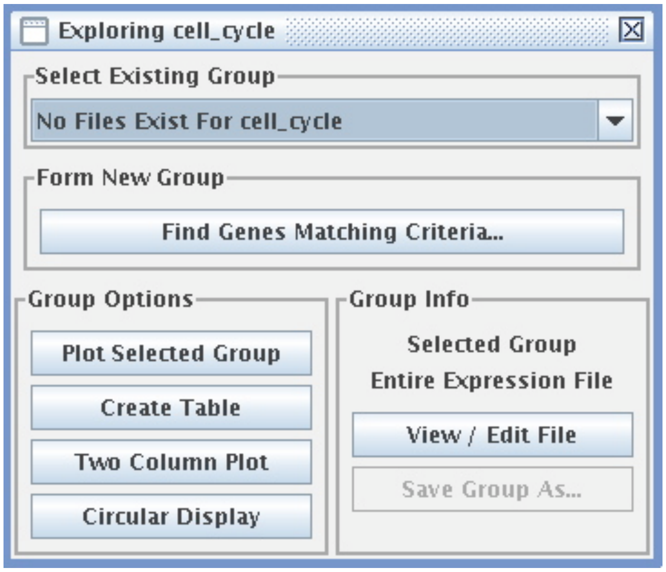
☑ Molecular Function does not c... transcription factor

☑ Cellular Component contains nucleus

Group genes matching all selected criteria

Explore (Control E)

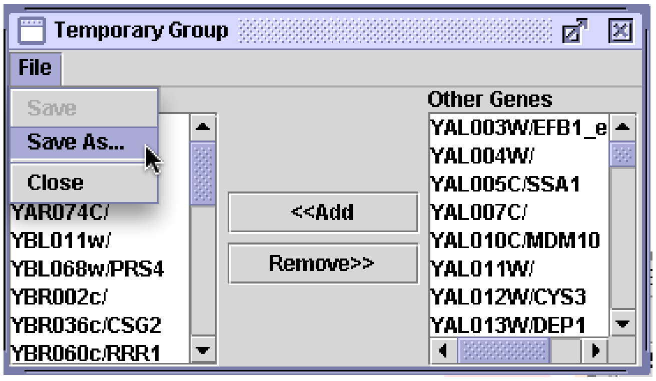
After you have transformed your data, you can explore it in a number of ways. The default group of genes is the full set. You can select a subset of genes via the Form New Group button called “Find Genes Matching Criteria…” You can search for criteria similar to those shown for the filter set on the previous page. When you have identified genes of interest, the window changes as shown to the right in red text. To save this new group of genes, click on the “View/Edit file” button just below the red text.



Exploring cell_cycle

Select Existing Group: No Files Exist For cell_cycle

A new window will appear that lets you view the list of genes in your newly formed group. You can modify this group if you want, or you can “save as” under the file menu. You can create many subgroups of genes and explore them individually using the “select Existing Group” pull down menu. Once you have subsets of genes to explore, you can perform visualize them in a number of ways:

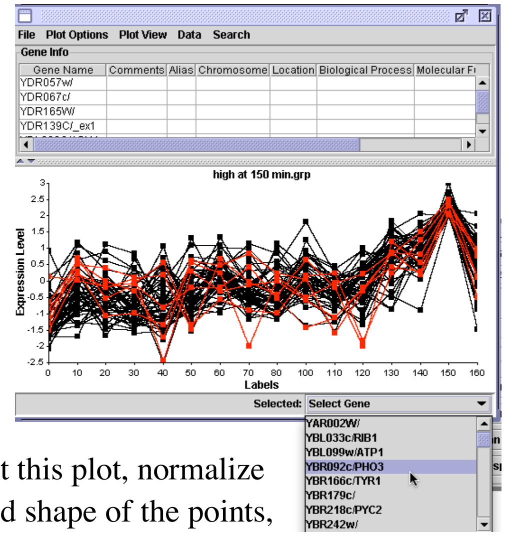


YAR074C/
YBL011w/
YBL068w/PRS4
YBR002c/
YBR036c/CSG2
YBR060c/RRR1

YAL003W/EFB1_e
YAL004W/
YAL005C/SSA1
YAL007C/
YAL010C/MDM10
YAL011W/
YAL012W/CYS3
YAL013W/DEP1

Plot Selected Group

You can have the ratios plotted graphically. You can select one gene using the pull down menu in the bottom right corner. Or, as shown here, you can click on one node at a time and hold down the shift key to select multiple genes (in this case, those with the lowest ratios in the group). These selected genes are listed in the top window (which you can pull down to see) as well as any other information about these genes in your gene list. You can adjust the size of the plot, as well as zoom in on a section. For example, this group of genes was selected by having a ratio of 2 or more at 150 minutes. To untangle the crowded lines, you can zoom in on any region of interest. To do this, hold down the control button then click and drag a box around the crowded area to zoom in. You can unzoom using the Plot View menu at the top of the window.



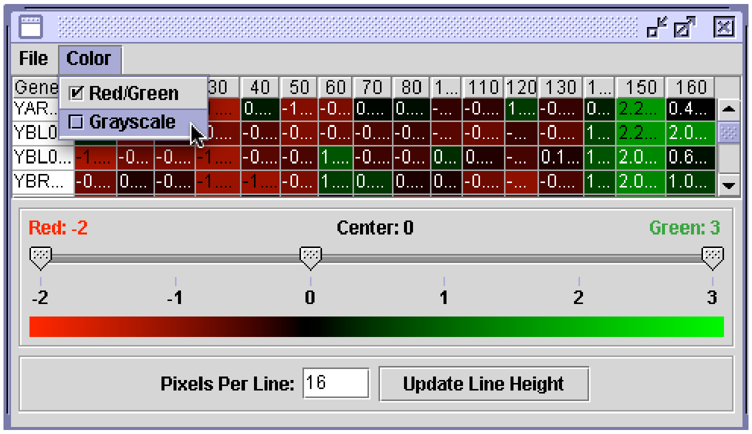
Gene Name	Comments	Alias	Chromosome	Location	Biological Process	Molecular Fı
YDR057w/						
YDR067c/						
YDR165W/						
YDR139C/_ex1						

YAR002W/
YBL033c/RIB1
YBL099w/ATP1
YBR092c/PHO3
YBR166c/TYR1
YBR179c/
YBR218c/PYC2
YBR242w/

In addition, you can label the axes, save this as a file, print this plot, normalize the data (if you have not already done so), change the size and shape of the points, and search for certain terms for the genes based on the gene list from which these genes are derived.

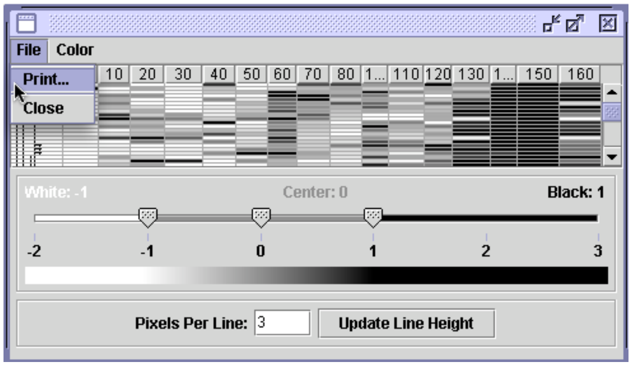
Create Table

This feature is unique to MAGIC Tool and creates a dynamic table. The default is a grayscale table, but you can change this to a red-green scale if you prefer. The most interesting feature of this interactive table is the scale bar and the three sliding tabs. Imagine a gene set that has one gene with a very high ratio (e.g. +16) and one gene with a very low ratio (-16) but with most genes having ratios between +3 and –3. Because of these two extreme genes, the color differences in the remaining genes would be lost. However, if you adjust the tabs, you can compress the color scale on the extreme ends and bring more color variation to the middle of the range of ratios, where most of your genes are located. In the image below, the scale has been adjusted and converted to grayscale to illustrate the potential views in the Create Table option. In this view, the gene lines have been reduced from 16 pixels high to 3 pixels high, the color scale changed to grayscale and the range reduced to –1 to +1. This reduction makes all high and low values either white or black, but allows the intermediate values to be on the grayscale.



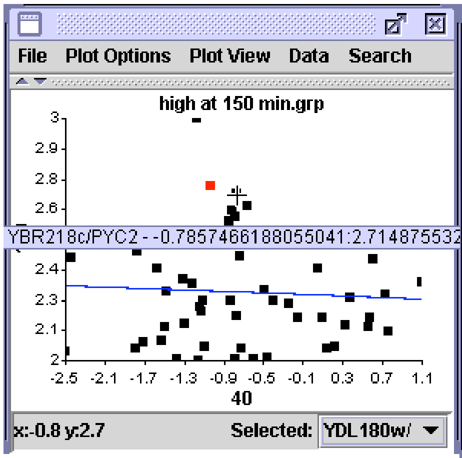
Red: -2 Center: 0 Green: 3

Pixels Per Line: 16



White: -1 Center: 0 Black: 1

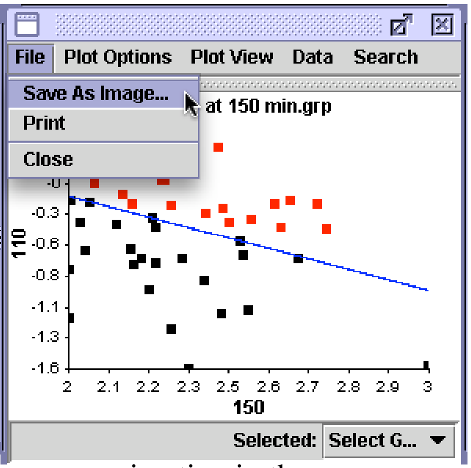
Pixels Per Line: 3



YBR218c/PYC2 - -0.7857466188055041:2.714875532

x:-0.8 y:2.7

Selected: YDL180w/



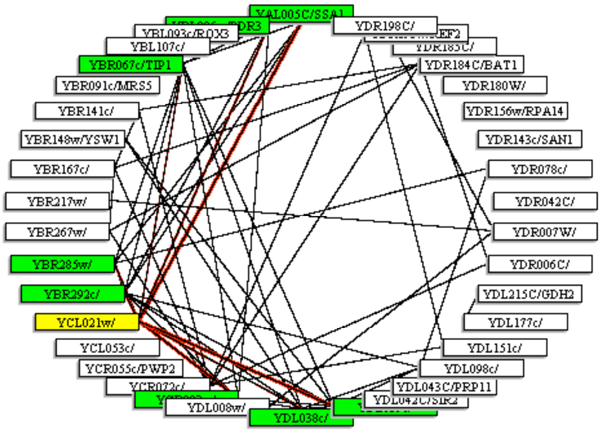
Selected: Select G...

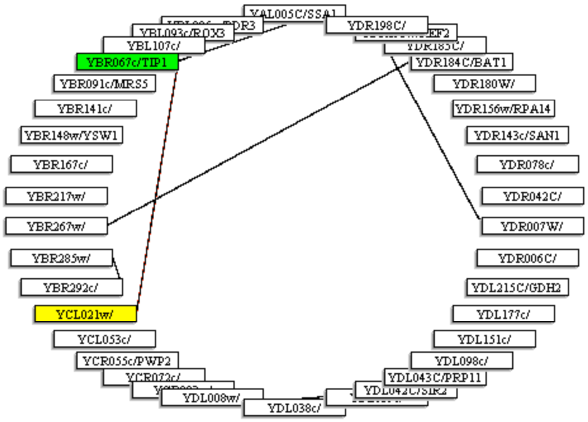
Two Column Plot

This plot allows you to select two columns of data and compare their ratios. As you can see, some comparisons are more similar than others. In this plot, you can select a single gene (left) or many genes (right; hold down the shift key while clicking). If you mouse over a gene, the display will tell you the two ratios for the two time points. You can also see an approximation in the bottom left corner.

Circular Display

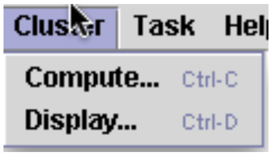
Another unique MAGIC Tool display is the circular one. Let’s imagine you have created a group of genes and you want to know how correlation coefficient for these genes, and to which genes the correlation exists. The default setting is correlation coefficient of 0.8 which is shown on the left. Using the display menu, you can change the radius of the circle and the threshold for reporting correlations. Change the threshold to 0.1 (correlation of 0.9) and you see fewer lines connecting the genes (right). In this case, the same gene was clicked on (yellow) and the genes which met the threshold are colored green with the lines colored red.


YAL005C/SSA1
YBL093c/ROX3
YBL107c/
YBR067c/TIP1
YBR091c/MRS5
YBR141c/
YBR148w/YSW1
YBR167c/
YBR217w/
YBR267w/
YBR285w/
YBR292c/
YCL021w/
YCL053c/
YCR055c/PWP2
YCR072c/
YDL008w/
YDL038c/
YDL042C/SIR2
YDL043C/PRP11
YDL098c/
YDL151c/
YDL177c/
YDL215C/GDH2
YDR006C/
YDR007W/
YDR042C/
YDR078c/
YDR143c/SAN1
YDR156w/RPA14
YDR180W/
YDR184C/BAT1
YDR185C/
YDR198C/


YAL005C/SSA1
YBL093c/ROX3
YBL107c/
YBR067c/TIP1
YBR091c/MRS5
YBR141c/
YBR148w/YSW1
YBR167c/
YBR217w/
YBR267w/
YBR285w/
YBR292c/
YCL021w/
YCL053c/
YCR055c/PWP2
YCR072c/
YDL008w/
YDL038c/
YDL042C/SIR2
YDL043C/PRP11
YDL098c/
YDL151c/
YDL177c/
YDL215C/GDH2
YDR006C/
YDR007W/
YDR042C/
YDR078c/
YDR143c/SAN1
YDR156w/RPA14
YDR180W/
YDR184C/BAT1
YDR185C/
YDR198C/

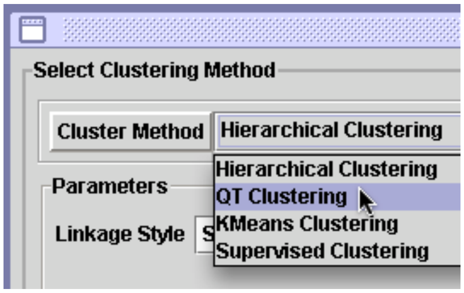
Cluster Menu

Compute… (Control C)
Once you have created dissimilarity file, you may cluster your data. To do this you must computer the cluster using one of four methods. Details for these four methods can be found in the Instructor's Guide.



Hierarchical Clustering

Hierarchical clustering produces a tree-like structure (a *dendrogram*) by connecting genes according to the similarity of their expression data. When a gene joins with another gene or group of genes in the tree, the entire collection of genes is represented as a single pseudo-gene. The similarity between a given gene and the gene (or pseudo-gene) to which it is connected, is indicated by the horizontal length of the branches joining them. At each stage in the algorithm, the two most similar genes or pseudo-genes are joined together. The process continues until all genes have joined the tree.



QT Clustering

QT Cluster takes every gene under consideration and one at a time, builds a temporary cluster for each gene with a user-defined cutoff value for similarity. Whichever gene garnered the most genes in its cluster is used to create permanent cluster and all the genes associated in this cluster are removed from the list of genes for the next round of creating permanent clusters. QT Cluster repeats the process of creating temporary clusters, one gene at a time, and then forms the second permanent cluster using the largest temporary cluster. This process is repeated until all the genes are in clusters, or the remaining genes form clusters smaller than a user-defined size. These remaining genes (called *singletons*) are not presented in the clustering displays unless the user defined 1 as the minimal size for a permanent cluster.

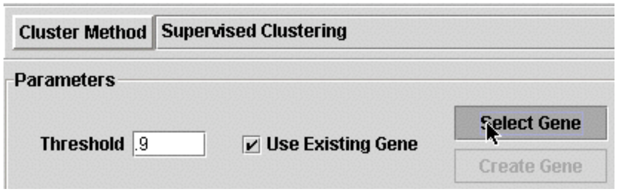
When you use QT Cluster, you should adjust the threshold value. The default of 0.9 means correlation coefficients of +0.1 through +1.0. If you change the threshold setting to 0.2, you will cluster genes only if their correlation coefficients are +0.8 through +1.0. The range of settings for threshold is from 0 (correlation of +1.0) through 1 (correlation of 0, i.e. not similar at all) to 2 (correlation of –1.0; track opposite each other). Therefore, by setting the threshold at 2, you would get every single gene placed in one cluster.

K-Means Clustering

In this method, you determine *a priori* how many clusters there will be (K = the number of clusters) and MAGIC tool will make sure all genes fit into this number of clusters. This is the first step in Self Organized Maps but both methods begin with the investigator determining how many clusters to generate.

Supervised Clustering

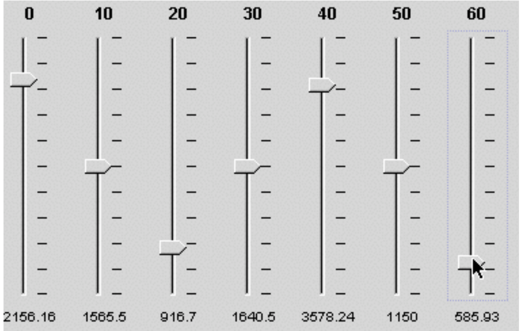
This method performs a QT cluster but you can define the threshold and choose one gene around which you want your cluster built. This allows you to focus your research on your favorite gene. On the left, you see that "Use Existing Gene" is selected. Click on the "Select Gene" button and then choose form the genes in your gene list of the currently active expression file.


Cluster Method: Supervised Clustering

Parameters

Threshold: .9

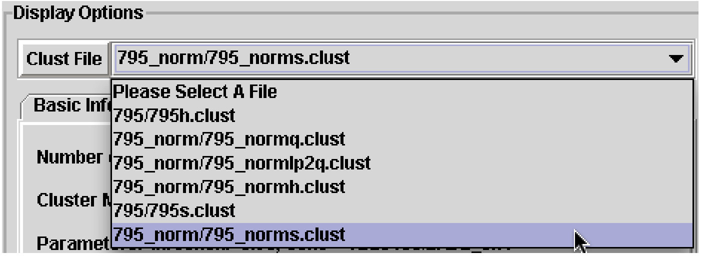
☑ Use Existing Gene



Alternatively, you can deselect the "Use Existing Gene" option and then click on "Create Gene". This produces a window that allows you to manipulate the sliders to create an expression profile for which you want to find genes with similar profiles (based on the threshold you choose). This is a quick way to find complex patters of interest to you.

Display…

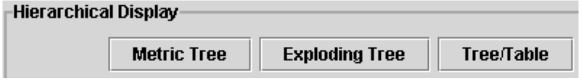
Once you have create a cluster or two, you can display them. First, choose the cluster file you want to display. Each type of cluster has its own display options.


Display Options

Clust File: 795_norm/795_norms.clust

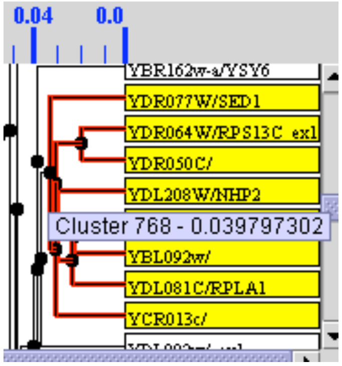
- Please Select A File
- 795/795h.clust
- 795_norm/795_normq.clust
- 795_norm/795_normlp2q.clust
- 795_norm/795_normh.clust
- 795/795s.clust
- 795_norm/795_norms.clust

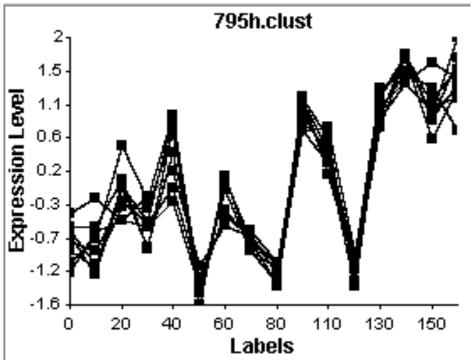
Hierarchical Cluster Display



You have three options for display, each of which has its own options. <u>Metric Tree</u> is unique to hierarchical clustering. It produces a dendrogram with nodes plotted at indicated thresholds. The smaller the threshold number, the higher the correlation coefficient.

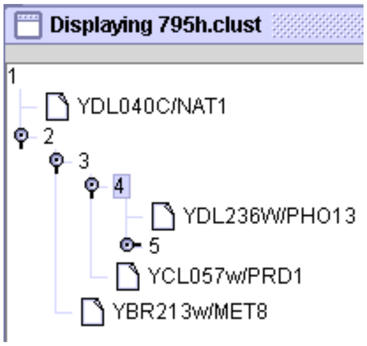
You can click on a branch point and highlight all the genes within this cluster as shown. If you mouse over the branch point, you can see the exact threshold which is 1 miuns the correlation coefficient (~0.96). You can plot this cluster and as you would image with this high a correlation coefficient, the normalized data plot as a very tight group.





<u>Exploding Tree</u> is an efficient way to show clusters and gradually expand the contents of each node. In this example, there is one gene and then all other genes are within node number 2. As you click on the nodes, they expand

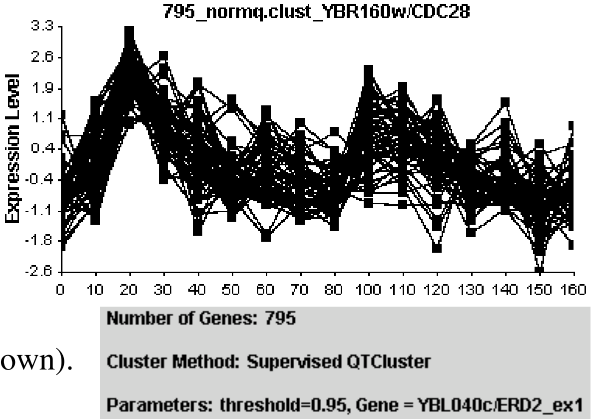
and if you click a second time, they collapse. You can explode the node completely by highlighting the number and clicking on the explode button, or explode it one at a time by clicking on the node directly. You can also plot any cluster within a node by clicking on the "Plot Node As Group" button.



Tree/Table is a way to combine the Table view and the dendrogram. The dendrogram is on the far left and the colored table (the majority of the window) is displayed on the right (view not shown).

QT Cluster Display

QT cluster also allows Exploding tree and Tree/Table, but it has replaced the metric tree with List. List allows you to see the name of the root gene for each cluster. If you click on the root gene, then all the genes within this cluster are displayed. You can plot this cluster as shown here.



Number of Genes: 795

Cluster Method: Supervised QTCluster

Parameters: threshold=0.95, Gene = YBL040c/ERD2_ex1

Supervised (QT) Cluster Display

Supervised Cluster hast the same display options as regular QT Cluster. However, when you are choosing your display, you should note the box that indicates what threshold was used and which gene was used as the root. In this case, ERD2, the KDEL receptor exon 1 was used as the root for this cluster with a correlation coefficient of 0.95 (plot not shown).

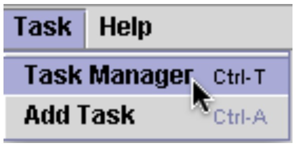
K-means Cluster Display

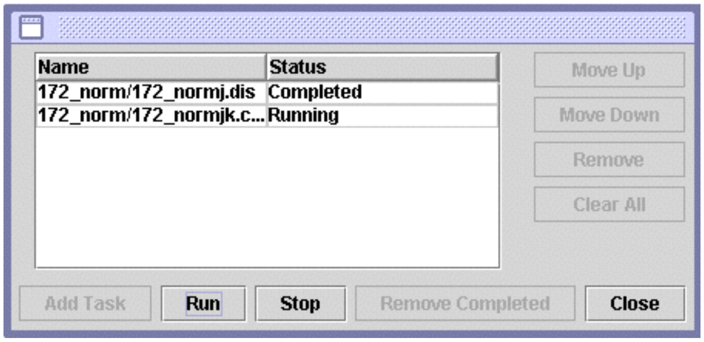
The three displays possible for K-means cluster display are described above.

Export…

When gene lists get longer than about 5000 genes, displaying clusters becomes slow in MAGIC Tool. One way to handle this is to export a cluster computed by MAGIC Tool for viewing in other software. We export files that are readable by the open source software JTreeView. Currently, you must exit MAGIC Tool and open JTreeView to begin exploring your very large clusters using this more advanced software. In the future, we hope to integrate the two programs more tightly.

Task Menu





Name	Status
172_norm/172_normj.dis	Completed
172_norm/172_normjk.c...	Running

As your datasets get bigger, the time it will take to make all the necessary calculations will increase rapidly. Therefore, MAGIC allows you to establish a list of tasks to be performed in sequence. You can tell MAGIC to begin a series of steps and then walk away from your computer. MAGIC will perform this sequence of tasks while you do other things. For example, you can establish a list of tasks to perform and go home for the night. When you return the next morning, MAGIC will have completed the series of tasks. At this time, the only tasks that can be performed are calculating dissimilarities and clusters.

Task Manager (Control Shift M)
The window above is the task manager. It allows you to add or remove a task, change the order of a task as well as various housekeeping chores.

Add Task (Control T)
This option allows you to add a task without going through the task manager.

Help (Control H)
Currently, electronic help is not available. We hope to get this working soon. The contents will be very similar to this User's Guide, but be available real time inside of MAGIC Tool.

Credits

MAGIC Tool version 1.0 was written in JAVA by Adam Abele, Brian Akin, Danielle Choi, and Parul Karnik, David Moskowitz. Contributors to subsequent versions are Mackenzie Cowell, Gavin Taylor, and Bill Hatfield. Laurie J. Heyer and A. Malcolm Campbell are advisors to the code-writing team. MAGIC Tool was developed at Davidson College and supported by the NSF, Duke Endowment, and Davidson College.

We would like to thank Wolfgang Christian and Mario Belloni for sharing their knowledge and resources with us.