



GeneSpring User Manual

version 6.1

November 14, 2003

Copyright 2003 Silicon Genetics. All rights reserved. GeneSpring, GeneSpider, GenEx, GeNet, MetaMine, ScriptEditor and MicroSift are trademarks of Silicon Genetics. All other products, including but not limited to Affymetrix GeneChip®, Affymetrix Global Scaling™, GenBank, Microsoft Excel®, Microsoft Notepad®, Pico™, SimpleText© and Adobe FrameMaker®, are the trademarks of their respective holders.

Table of Contents

1 Welcome to GeneSpring

Getting Started	1-2
Learning To Use GeneSpring	1-4
GeneSpring Basics	1-6
Data Loading	1-8
Loading Your Data.	1-8
Basic Actions	1-11
Commonly Used Functions	1-16
The Gene Inspector Window	1-16
Making Lists	1-16
Setting Preferences	1-18
Data Files	1-18
Database	1-18
Color.	1-18
Gene Labels	1-21
Browser	1-21
Firewall	1-21
System	1-22
GeNet.	1-22
Computation.	1-23
Miscellaneous.	1-23

2 Creating Genomes

The New Genome Installation Wizard	2-2
Creating a Genome from Experiment Data	2-9
Data Format	2-10
Layout Parameters	2-12
The .layout file	2-12
Examples of .layout files for Arrays	2-13

Renaming and Deleting Genomes	2-14
Renaming a Genome	2-14
Deleting a Genome.	2-14

3 Working With Experiments

Importing Experiment Data	3-2
Memory Use for Experiment Loading.	3-2
Loading an Experiment	3-3
Using the Column Editor	3-9
Default Column Assignments of Known Products	3-12
Creating New Experiments.	3-16
Copying and Pasting Experiments	3-17
Preparing to Paste.	3-17
Common Mistakes in Pasting	3-19
Pasting an Experiment into GeneSpring	3-20
Copying an Experiment or a List Out of GeneSpring	3-20
Default Normalizations.	3-21
The Sample Manager	3-23
Filtering Methods.	3-25
Experiment Parameters.	3-29
Parameters Displayed in the Navigator	3-29
A Note on Multiple Parameters	3-30
Parameter Display Options	3-30
Hidden Elements	3-31
Continuous Element.	3-31
Non-Continuous Element.	3-31
Color Code	3-31
Changing Experiment Parameters	3-32
Sample Attributes	3-35
Changing Experiment Attributes	3-35
Editing Standard Attributes	3-37
Experiment Interpretations	3-39
Vertical Axis Modes	3-40
Parameter Display Modes	3-43
Cross-gene Error Models	3-44
Using the Cross-gene Error Model	3-44
Technical Details	3-46

4 Viewing Data

Using the Genome Browser	4-2
Zooming In.	4-2
Panning	4-2
Modifying Display Options	4-2
Displaying a Gene List.	4-3
Finding and Selecting Genes	4-4
Performing a Simple Search	4-4

Performing an Advanced Search	4-4
Searching GeNet from GeneSpring	4-5
Selecting Genes	4-8
Inspectors	4-10
The Gene Inspector	4-10
The Sample Inspector.	4-13
The Experiment Inspector	4-16
The Condition Inspector.	4-18
The Gene List Inspector.	4-20
The Classification Inspector.	4-22
Display Options	4-25
Linked Windows	4-25
Split Windows	4-25
Bookmarks	4-26
The Vertical Axis.	4-27
Error Bars.	4-28
Legend	4-28
Color	4-30
Blocks View	4-36
Blocks View Display Options	4-36
Graph View.	4-37
Graph View Display Options.	4-37
Bar Graph View	4-39
Bar Graph View Display Options	4-39
Physical Position View.	4-41
Physical Position Display Options.	4-43
Scatter Plot View	4-45
Scatter Plot Display Options	4-45
3D Scatter Plot View	4-49
3D Scatter Plot Display Options	4-49
Tree View.	4-52
Viewing a Tree.	4-52
Selecting and Viewing Subtrees	4-52
Magnifying Trees.	4-55
Viewing Nodes.	4-55
Viewing Gene Names in Trees	4-56
Viewing Parameters in Trees.	4-56
Tree Display Options.	4-56
Ordered List View	4-58
Ordered List Display Options	4-58
Array Layout View.	4-60
Array Layout Display Options.	4-60
Pathway View.	4-62
Pathway Display Options.	4-62
Compare Genes to Genes	4-64
Graph by Genes View.	4-65

Graph by Genes Display Options	4-65
View as Spreadsheet	4-67
Condition Scatter Plot	4-68
Condition Scatter Plot Display Options	4-69
Showing/Hiding Window Display Elements	4-71

5 Normalizing Data

Experiment Normalizations	5-2
Using the Experiment Normalizations Window	5-2
Normalization Types	5-6
Start with Pre-Normalized Values	5-6
Data Transformation	5-6
Per Spot Normalization	5-7
Per Chip Normalizations	5-10
Per Gene Normalizations	5-13
Normalization Strategies for Specific Technologies	5-17
Normalization of Affymetrix Data	5-17
Normalization of Two-color Microarray Data	5-17
Region Normalization	5-17
Dealing with Repeated Measurements	5-18
Negative Control Strengths	5-20
References	5-21

6 Analyzing Data

Creating and Editing Gene Lists	6-2
Filtering Methods	6-3
Working with Gene Lists	6-6
The Find Similar Command	6-6
Making Lists by Applying Filters	6-11
Making Lists from Properties	6-12
Making Lists with the Venn Diagram	6-13
Making Lists from Classifications	6-14
Making Lists from Selected Genes	6-14
Creating Expression Profiles	6-15
Pathways	6-16
Regulatory Sequences	6-18
The Homology Tool	6-24
Annotation Tools	6-27
Updating your Master Gene Table with GeneSpider	6-27
Genome Databases	6-30
Building a Simplified Ontology	6-31
To Make Gene Lists From Properties	6-32
Building Homology Tables	6-32
Statistical Analysis (ANOVA)	6-33
1-Way ANOVA	6-34
Post-Hoc Tests	6-39

Viewing Post-Hoc Test Results	6-41
2-Way ANOVA	6-43
Details on 2-Way ANOVA	6-45
The Filtering Menu	6-51
The Basic Anatomy of a Filtering Window	6-51
Basic Filters	6-54
Filter on Expression Level	6-54
Filter on Fold Change	6-55
Filter on Error	6-58
Filter on Confidence	6-59
Filter on Flags	6-60
Filter on Gene List Numbers	6-64
Advanced Filtering	6-66
Creating Boolean Filters	6-67
Saving Filters	6-67
7 Clustering and Characterizing Data	
The Clustering Window	7-2
Using the Clustering Window	7-2
Clustering Methods	7-5
Gene Tree	7-5
Condition Tree	7-6
k-Means Clustering	7-8
Self-Organizing Maps	7-11
QT Clustering	7-14
Principal Components Analysis	7-16
PCA on Genes	7-16
PCA on Conditions	7-18
Interpreting your PCA Results	7-20
The Class Predictor	7-24
Interpreting the Results of a Prediction	7-25
Find Similar Samples	7-26
The Find Similar Samples Results Window	7-26
8 Scripts and External Programs	
Scripts	8-2
What is a Script?	8-2
The Run Script Window	8-3
The Script Inspector	8-5
Using the Remote Server	8-5
Using the ScriptEditor	8-7
ScriptEditor Concepts	8-7
The ScriptEditor Interface	8-7
The Icon Legend	8-11
The Properties Panel	8-12
Creating Scripts	8-13

Building Scripts	8-16
Saving Scripts	8-18
Warning Messages	8-18
Script Help	8-18
Script Building Blocks	8-20
Scripts to External Programs	8-32
Scripts and External Programs	8-34
External Programs	8-35
The New External Program Window	8-35
Examples	8-40
The External Program Inspector	8-42

9 Exporting GeneSpring Data

Saving Images	9-2
Saving Pictures and Printing	9-4
Exporting Gene Lists	9-5
Exporting MAGE-ML Data	9-8
Publishing Data to GeNet	9-11
Uploading Data Objects to GeNet	9-11
Deleting Data Objects from GeNet	9-12
Uploading Genomes to GeNet	9-12

A Installing from a Database

Custom Databases and GeneSpring	A-2
Databases	A-2
Open Database Connectivity	A-2
Structured Query Language	A-2
SQL Call Level Interfaces	A-3
The Genetic Analysis Technology Consortium	A-3
Databases and GeneSpring	A-3
Adding an Experiment from a Database	A-5
Connecting your Database to GeneSpring	A-6
Configuration File Reference	A-6
Tag Definitions	A-9
Entering your Database into GeneSpring	A-23
Prepared Databases	A-23
More Complicated Databases	A-23

B Equations for Correlations and other Similarity Measures

Measures of Similarity	B-2
Common Correlations	B-3
Standard Correlation	B-3
Pearson Correlation	B-3
Spearman Correlation	B-4
Spearman Confidence	B-4

Two-Sided Spearman Confidence	B-5
Distance	B-5
Smooth Correlation	B-6
Change Correlation	B-6
Upregulated Correlation.	B-7
Number of Samples Required to do Analyses	B-8

C Technical Details for the Predictor

Gene Selection	C-1
Classifying the Test Samples	C-1
Decision Threshold	C-1
References for the Predictor.	C-2

Welcome to GeneSpring

Congratulations on selecting the most advanced, flexible tool available for gene expression data analysis.

This manual is a guide to GeneSpring features. Chapter 1 covers installing GeneSpring, loading and setting up your data, and GeneSpring basics. The remaining chapters discuss loading, set-up and the various data analysis and visualization tools in detail.

Getting Started

Requirements

Windows:

- Windows 98/NT/2000
- Pentium II or better
- 256MB RAM (512MB recommended)
- 1024x768 display
- 40MB of free disk space

Macintosh:

- MacOS 9.1 or higher
- Power PC or better
- MRJ 2.2.5
- 256MB RAM (512MB recommended)
- 1024x768 display
- 40MB of free disk space

Unix:

- Most common Unix OSes (Linux and Solaris recommended)
- A JVM that supports JDK1.1 or later
- 256MB RAM (512MB recommended)
- 1024x768 display
- 40MB of free disk space

Installing from a CD

If you are installing GeneSpring from a CD, you have several options after you place your CD in the drive:

1. Select **Install GeneSpring Demo**. A splash screen and an InstallAnywhere© screen appears with a progress bar.
2. Follow the on-screen instructions. For more information see the ReadMe file included with the CD.

In Windows, you can also install the software by using the **Start > Run** command in the Start menu. Enter `D:\gspring.exe`, where D is the CD-ROM drive on your computer.

Installing from the Web

If you are reading this manual and do not have a copy of GeneSpring, you can download a copy by going to the following URL:

<http://www.sigenetics.com/cgi/SiG.cgi/Products/GeneSpring/download.smf>

Follow the on-screen directions and Silicon Genetics will send you a username, password and download link.

Starting GeneSpring

Once you have installed GeneSpring, the GeneSpring icon appears on your desktop.



Figure 1-1 The GeneSpring icon

To start GeneSpring, double-click the GeneSpring icon. Windows users can also start GeneSpring by selecting it from the **Start menu > Programs > GeneSpring** or navigating to **Program files > Silicon Genetics > GeneSpring** and double-clicking the GeneSpring icon.

Macintosh users can also start GeneSpring from the Applications folder/Silicon Genetics/GeneSpring.

A splash screen appears containing your GeneSpring version number, the expected expiration date and the JVM you are using. You will then see the GeneSpring main window. For further details, see “GeneSpring Basics” on page 1-6.

Obtaining a License Key

If you have already installed a demo copy of GeneSpring, your license key will expire within one month of the initial installation. Once you have purchased a full GeneSpring license, Silicon Genetics will send you a license key. Save this license key file in the Silicon Genetics/GeneSpring/Data folder. If you have kept the default settings of GeneSpring, on a Windows machine look in C:// Program Files, and on a Mac look in the Applications folder. When the key is about to expire, you will get a warning message 30 days in advance. If your license has expired, or is about to, contact Silicon Genetics at **1-866-SIGSOFT (744-7638)**.

Setting Memory Usage Options

Once GeneSpring is installed, make sure the default memory setting in GeneSpring preferences is half of your computer’s available memory (or more if you have a lot of RAM). To do this, select **Edit > Preferences**, choose **System** from the pull-down menu and enter the amount of memory in the **Desired Memory Use** field.

Configuring Virtual Memory

At least 150MB of virtual memory is required for optimal GeneSpring performance. To ensure that large files are not interfering with software performance, you may need to move some large files to a different hard disk.

If you continue to experience slow performance, check memory usage by selecting **Help > System Monitor** before invoking any functions. Make a record of the Total Mem-

ory and Free Memory listed in the System Monitor window and contact the Silicon Genetics Technical Services Department at **1-866-SIG-SOFT** or **support@sigenetics.com**.

Updating GeneSpring

To update an existing GeneSpring installation, select **Help > Update GeneSpring** and follow the on-screen instructions to obtain the current `GeneSpring.jar`.

Learning To Use GeneSpring

The Help Menu

The Help Menu is located on the right of the menu bar.

Tutorial

This command opens your default browser and takes you to the GeneSpring Basics Instructional Manual in PDF format. You can save this file to your local machine and print it. The tutorial covers many basic topics of GeneSpring.

User Manual

Select the User Manual command to open the manual installed on your hard drive during installation or updating. The *GeneSpring User Manual* is a PDF document you can save or print.

Version Notes

Select this option to view notes for your version of GeneSpring. These are located in `C:\Program Files\SiliconGenetics\GeneSpring\docs\Version-Notes.html`.

Update GeneSpring

Select this option to download the latest version of GeneSpring. You must have an active license key to update your software.

You can also automatically update the manuals that accompany GeneSpring. The manuals are published in HTML or PDF formats. It is important to download updated documentation when you update the GeneSpring software.

Technical Support

Select this option to contact Silicon Genetics technical support on the Web.

Silicon Genetics on the Web

Select this option to browse the Silicon Genetics website. This site contains a wealth of information including manuals and information on workshops designed to help you use GeneSpring more effectively.

GeNet Database

Select this option to browse the GeNet™ Web page. From here you can download a demo version of GeNet™ and upload or download additional information. See the *GeNet User Manual* for more information.

Support and Training Resources

Select this option to view the Silicon Genetics training page. Here you can take advantage of Silicon Genetic's many training options.

System Monitor

Select this option to view the Java system monitor to track free memory and view the processes running on your computer.

Test Database Connectivity

If you are using a database with GeneSpring, select this option to verify that GeneSpring and the database are communicating with each other.

Show Helpful Hints

Select this option to display a new helpful hint each time you start GeneSpring.

About

Select this option to view information about GeneSpring such as the version number and demo expiration date. If you contact Technical Support, they will ask you for the version of GeneSpring you are using.

GeneSpring Basics

GeneSpring is a powerful analysis tool. Like any professional level program, it can be intimidating to new users. The following section is a brief introduction to using GeneSpring and loading data designed to get you up and running in the shortest possible time. Figure 1-2 depicts the steps in a typical analysis session using GeneSpring. Note that this diagram represents what might occur in a typical data analysis session and does not include all of the types of analyses found in GeneSpring.

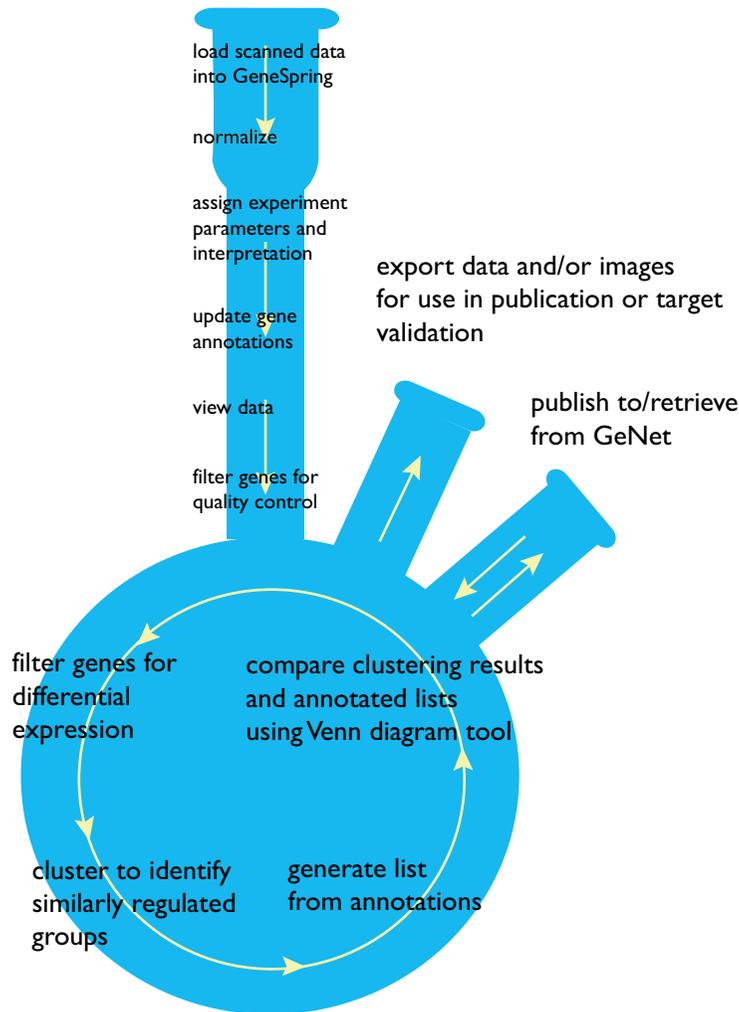


Figure 1-2 Typical GeneSpring workflow

In the process of loading your data, you will come across terms and concepts such as genome, parameter, parameter values, replicate, interpreted data, etc. Below are explanations of how these terms are used in GeneSpring.

What is a Genome?

In the context of GeneSpring, a genome contains information about all the genes in your chip or microarray setup. Note that a GeneSpring genome does not correspond exactly to

the biological definition of a genome. A genome in GeneSpring is composed of discrete genes as opposed to the full nucleotide sequence. This means that a GeneSpring genome can contain two genes representing alternately spliced variants of a single gene, whereas a true genome would include the DNA sequences for only one.

What is a Parameter?

Parameters are experiment variables, such as stage, time, concentration, etc.

Parameter values are values assigned to experiment parameters. For example Embryonic, Postnatal or Adult could be parameter values of the experiment parameter *stage*, while .01 ppm could be a parameter value of the experiment parameter *concentration*.

What are Replicates?

Replicates can be:

- multiple spots on the same array representing the same gene (also referred to as a copy)
- the same sample on more than one array
- a biological replicate, equivalent samples taken from more than one organism

Graphically, a parameter defined as a replicate is a hidden variable; no visual distinction is made based on this parameter or its parameter values.

What is a Region?

Regions divide your data into specific sections. This is important if you use multiple arrays and want to normalize sections of an array separately rather than normalizing across the entire data set.

What is Raw Data?

The analysis process begins by obtaining data in the form of flat files that were generated by your scanning software or other expression analysis technology. GeneSpring is capable of recognizing most commercially available formats and can be customized to work with other formats as necessary. Typically, the gene/spot/probe-set intensity values in these files are referred to as raw data.

What is Normalized Data?

If GeneSpring recognizes your file format, it applies a set of default normalizations appropriate for your expression analysis technology. The denominator used to normalize each measurement is referred to as the control strength.

What is Interpreted Data?

GeneSpring can interpret normalized data in many different ways. You can elect to have multiple samples treated as replicates and averaged, and indicate what assumptions you GeneSpring should make about the precision of these averaged values. You can display and perform analyses on normalized data using three modes: ratio (raw versus control strength), logarithm of ratio, or in terms of fold change (versus the control strength). *It is important to note that the graphical display of normalized values and the numbers used for all analyses (such as clustering) reflect the mode you have chosen. However, the num-*

bers displayed as text (as in the Gene Inspector window) and entered by the user as parameters for analyses (as in the Filter Genes tools) are always in ratio mode.

What are Flags?

Flags are additional measurement markers in your data set. They can be assigned as *present*, *marginal*, *unknown*, or *absent*.

Data Loading

The demonstration version of GeneSpring comes pre-loaded with sample yeast, rat and Affymetrix data. Many users benefit from performing trial analyses on these sample data sets. When you are ready to analyze your own data, you must load and set up the data for analysis. There are four steps to preparing data:

1. Loading gene information (optional)
2. Loading experiment information
3. Telling GeneSpring how to analyze and display the information by assigning normalizations, parameter values, and modes of display
4. Annotating/updating your genome

Loading Your Data

Step 1: Load gene information from your arrays (optional)

1. Start GeneSpring and select **File > New Genome Installation Wizard**.
2. Type the organism name (or the brand name of your array) and click **Next**.
3. Enter the information requested on each screen and click **Next** until you have completed the wizard. For details, see “The New Genome Installation Wizard” on page 2-2.

If you skip this step, GeneSpring can load gene information directly from your data files. However, to retrieve annotations for your genome using the GeneSpider (Step 4), you must enter the GenBank accession number of each gene in column 10 of the master gene table. Silicon Genetics can provide annotated genomes for many of the most commonly used arrays. Call **1-866-SIG-SOFT** or email support@sigenetics.com for details.

Step 2: Load an Experiment

1. Select **File > Import Data**.
2. Choose a file.
3. If GeneSpring recognizes the format of your data file, it asks you to name your genome. If the data format is unknown, you must set up columns using the column editor.

To set up columns, click each of the cells in the Function row and choose a data type from the pull-down menu. When you are done, click **Next**.

4. GeneSpring asks if there are more files to be loaded for this experiment. If there are additional files, select them from the menu and click **Add**. When you are done, click **Next**.

If there are no more files to load, click **Next**.

5. Enter required attributes, if any, and click **Next**.
6. Click **Yes** if you would like to create an experiment from the sample(s) you imported. If not, click **No**.
7. Enter an experiment name in the Choose Experiment Name window and click **Save**.

Step 3: Assigning Normalizations, Parameter Values, and Interpretations

1. Select **Experiments > Experiment Normalizations**. Choose the types of normalizations to apply. Four classes of normalizations are available:
 - background subtraction
 - per spot normalizations
 - per chip (global) normalizations
 - per gene normalizations.

Specify the desired normalizations and save. For information about normalizations and when to apply them, see Chapter 5, “Normalizing Data”.

2. Select **Experiments > Experiment Parameters**. Set parameter name, units, values, and value order, and add any missing parameters. For information about changing experiment parameters, see “Experiment Parameters” on page 3-29.
3. Select **Experiments > Experiment Interpretation**. Choose the following:

- mode of display
- lower and upper bounds of data
- flagged measurements to be included
- whether to use the Cross-gene Error Model
- whether the data should be continuous, non-continuous, viewed as a replicate, or color-coded

These assignments are an extremely important preparation for any type of data analysis. For information about changing experiment interpretations, see “Experiment Interpretations” on page 3-39.

Step 4: Annotate your genome (optional)

Most researchers will want to import the maximum amount of biological information available about each gene before beginning analyses. After collecting the data, it is a good idea to make lists of genes based on appropriate keywords.

1. Select **Annotations > GeneSpider**.
2. Select a database from which to update your annotations.

GeneSpring Basics

3. Select the column in your master gene table that contains the accession number (usually Column 10 for the GenBank locus). Make sure there are accession numbers in the column you select.
4. Click **Start** (the GeneSpider may continue gathering information for many hours).
5. Click **Save and close** when the GeneSpider is finished.

For details on the GeneSpider see “Annotation Tools” on page 6-27.

At this point your data are ready to work with.

Basic Actions

Once you have loaded your data, GeneSpring opens a window containing information from your new genome. Initially all the genes in your experiment are displayed. To see your new genome select **File > Open Genome or Array** and choose your genome from the pop-up list.

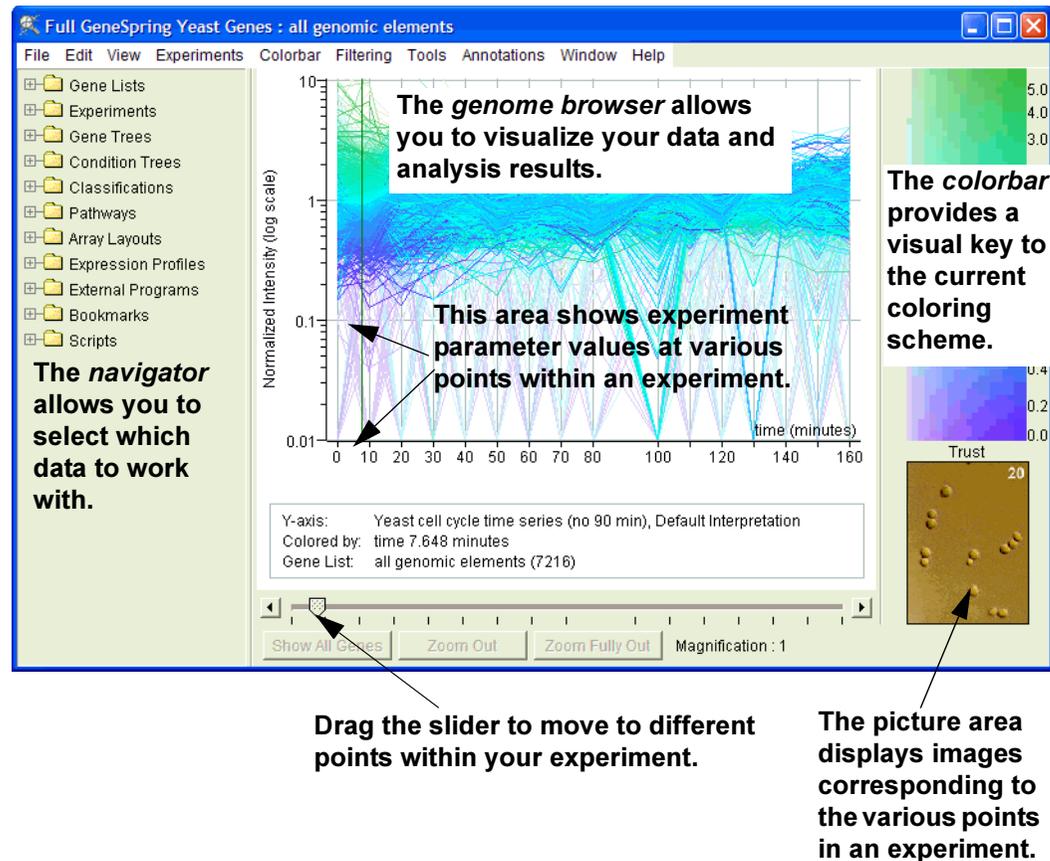


Figure 1-3 The main GeneSpring window

Below are some basic procedures for navigating the GeneSpring interface.

Changing the genes displayed:

Open the gene list folder in the navigator. GeneSpring initially displays the “all genes” list. You can change the genes shown in the display by choosing another list.

Views:

You can change the view in the genome browser using the View menu. GeneSpring initially displays the Classification view, in which genes are displayed according to pre-defined categories. However, you can also view displayed genes as a graph, a scatter plot, a bar graph, an ordered list, etc. Note that some views such as Tree, Pathway, and Array Layout require some preparation, such as creating a tree or adding a pathway or Array Layout image. For details on views, see Chapter 4, “Viewing Data”.

Zooming in:

Basic Actions

To zoom in on a region or gene, click on an area and drag your cursor diagonally. An expanding rectangle appears. Release the mouse and GeneSpring zooms in on the region enclosed by this rectangle.

Zooming out:

To zoom out, click **Zoom Out** or right-click (Control + click for Mac) and choose **Zoom Out** to go back one level or **Zoom Fully Out** to zoom out as far as possible.

Moving around the screen:

You can move around a zoomed-in screen by using Page Up, Page Down and the arrows keys.

Selecting a gene:

Click once on a single gene to select it.

Selecting multiple genes:

Hold down the Shift key and drag to select multiple genes. Or hold down the Shift key and click on individual genes to select them one by one.

Finding a specific gene:

Select **Edit > Find Gene** or **Ctrl+F**. Enter the gene name or keyword and click **OK**. GeneSpring selects and zooms in on the gene.

Inspecting genes:

You can view detailed information about a gene by double-clicking on it to bring up the Gene Inspector window. This is easier after zooming in on the gene. A shortcut to the Gene Inspector is **Ctrl + I**, or **⌘+I** for Mac users.

Undo:

You can undo your last action by selecting **Edit > Undo** or **Ctrl + Z** (**⌘ + Z** for Mac users).

Your First Gene Lists

To make lists from appropriate keywords:

1. Select **Annotations > Make Gene Lists from Properties**.
2. Choose the property you want to use for generating lists.
3. Click **OK**.

To make a list based on biological function:

1. Select **Annotations > Build Simplified Ontology**.
2. Name your new list.
3. Click **OK**.

To make lists from a group of selected genes:

1. Right-click over a highlighted group of genes.
2. Select **Make List from Selected Genes** from the pop-up menu.

Your new lists appear in the Gene Lists folder.

Tips for Macintosh Users

Except where otherwise noted, instructions in this manual describe GeneSpring usage on a PC. If you are a Macintosh user, you may find the following keystroke and mouse conversion information helpful:

- **Right-Click**—Hold Control and click. This most often activates a pop-up menu.
- **Ctrl = ⌘**—Substitute the ⌘ key wherever the manual mentions Ctrl. For example, if the manual says “press Ctrl + I to reach the Gene Inspector,” substitute the ⌘ (Apple) key for Ctrl.
- **Drawing genes on a pathway**—Hold down the Option key and drag your cursor diagonally to draw a gene on a pathway. See “Pathways” on page 6-16 for more information.

Note that on a Macintosh the menu bar is at the top of the screen, not on the individual GeneSpring windows as displayed in this manual.

The Navigator

GeneSpring organizes data elements relating to your genome into folders in the navigator. Each folder contains a specific type of information.

By default, folders in the navigator are closed, although on start-up GeneSpring displays an “all genes” or “all genomic elements” gene list. To change the default genome that GeneSpring initially open, select **Edit > Preferences** and click the **Data Files** tab. Enter a genome name in the Default Genome text field and click **OK**.

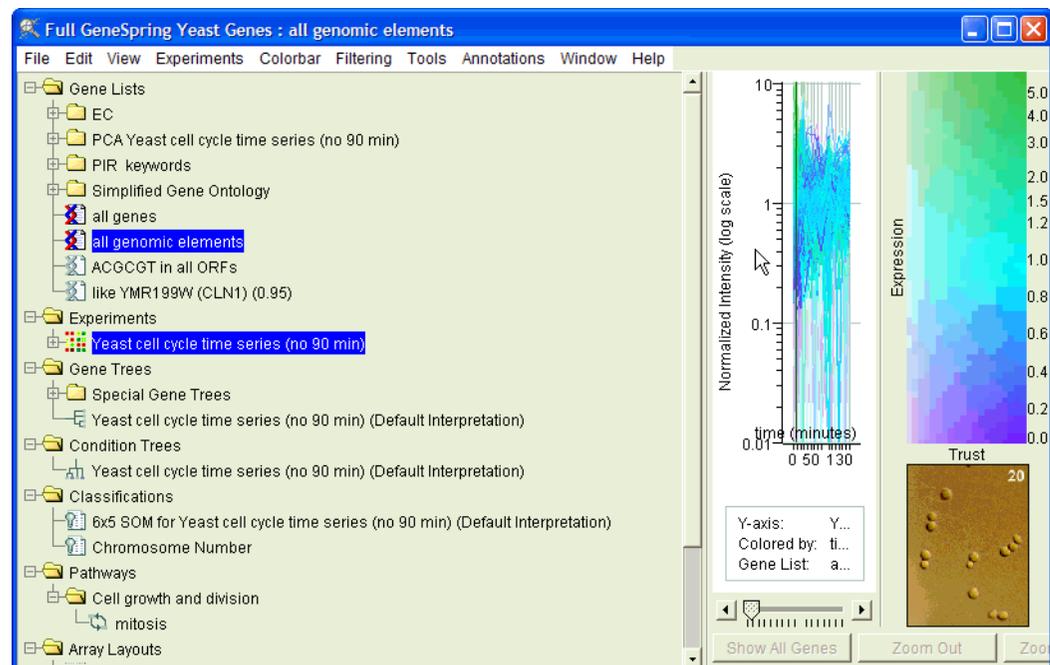


Figure 1-4 The GeneSpring Navigator

The Gene Lists Folder

During analysis, you will create and work with interesting collections of genes known as gene lists. These gene lists are stored in the Gene Lists folder. By default, GeneSpring makes and displays an “all genes” list containing all genes in the genome.

The Experiments Folder

The Experiments folder contains experiment information. Experiments are divided into *interpretations*. Experiment Interpretations tell GeneSpring how to treat and display your experiment variables, called experiment *parameters*.

Conditions are groupings of one or more samples. Each sample may be a condition, as in the “All Samples” interpretation or a condition may include multiple samples. For example, because the experiment above is organized according to the parameter values Embryonic, Postnatal and Adult, these can be called the *conditions* of the experiment. Within these conditions, the parameter *day* is being treated as a replicate and has been averaged for each condition, Embryonic, Postnatal and Adult, across all samples. Hence a condition can include data from more than one sample.

The Gene Trees Folder

Any gene trees created in GeneSpring are kept in the Gene Trees folder. Gene trees are dendrograms used as a method of showing relationships between the expression levels of genes over a series of conditions.

The Condition Trees Folder

Condition trees are like gene trees, except that instead of showing the relationships between genes, they show the relationships between the expression levels of samples. Condition trees are kept in the Condition Trees folder.

The Classifications Folder

The Classifications folder contains genes that have been grouped or *classified* to divisions defined by k-means or SOM clustering.

The Pathways Folder

Pathways are images of regulatory or metabolic pathways that can be imported into GeneSpring. Genes are overlaid on these images allowing you to observe their changing expression levels across experimental conditions. A feature called **Find Genes Which Could Fit Here** can be used as a tool to predict new pathway elements.

The Array Layouts Folder

The Array Layouts folder contains information about the arrangement of the spots on your array. These can be used to recreate an image of your arrays to check for regional abnormalities.

The Expression Profiles Folder

Expression profiles are lines representing gene profiles that you draw in the genome browser. You can then search for genes matching that profile. Any expression profiles you create are stored in the Expression Profiles folder.

The External Programs Folder

External programs are analysis programs outside GeneSpring that can be launched from within GeneSpring. Data from GeneSpring is sent to the program and output from the program is recognized by GeneSpring. These programs are kept in the External Programs folder.

The Bookmarks Folder

Bookmarks are saved display settings such as experiment, gene list, color scheme, selected genes, etc. You can always save your current display and return to it later by opening the Bookmarks folder and selecting a particular bookmark.

The Scripts Folder

Scripts are tools that save time by allowing a long series of data analysis steps to be performed at once. Scripts are re-usable and can be applied to any data set. You can create your own scripts using the Silicon Genetics ScriptEditor. All scripts, including complimentary scripts shipped with GeneSpring 4.2, are stored in the Scripts folder.

Commonly Used Functions

To open a different genome, choose **File > Open Genome or Array** and follow the submenus to your desired genome. To open another copy of the main window, choose **File > New Linked Window**. Each of these brings up a new main window similar to the one described in “GeneSpring Basics” on page 1-6.

To change preferences (colors, start up genome, etc.), choose **Edit > Preferences**. See “Setting Preferences” on page 1-18 for more details.

The Gene Inspector Window

Double-click a gene to bring up the Gene Inspector window. This window contains specific information about the selected gene. See “The Gene Inspector” on page 4-10 for details. Information presented in the Gene Inspector might include:

- knowledge you have about your selected gene (typically text).
- graphs of the selected gene’s expression profile from the current experiment.
- links to internet or intranet databases on the web for the selected gene.

Making Lists

There are many ways to create a list of genes, see Chapter 6, “Analyzing Data” for more details. From the Gene Inspector window you can do the following.

Making Lists with the Find Similar Command

The **Find Similar** button in the Gene Inspector allows you to create a list of genes having similar expression profiles to the gene being displayed. See “The Find Similar Command” on page 6-6 for more details.

Making Lists with the Complex Correlation Command

The Complex Correlation button in the Gene Inspector allows you to make a list of all the genes satisfying various conditions you define. See “The Find Similar Genes Window” on page 6-7 for more details.

Making Lists with the Venn Diagram

Select **Colorbar > Color by Venn Diagram** to begin. Right-clicking over lists in the navigator allows you to fill the diagram. This function allows you to make lists based on the membership of genes in a Venn Diagram. See “Making Lists with the Venn Diagram” on page 6-13 for more details.

Making Lists with the Filter Genes Command

Select **Tools > Filtering & Statistical Analysis**. It allows you to use expression level constraints and control strength restrictions to create a smaller gene list. See “The Filtering Menu” on page 6-51 for more details.

Making Lists from Selected Genes

You can make a list of all the genes you have selected in the genome browser by right-clicking and choosing **Make List from Selected Genes**. See the “Finding and Selecting Genes” on page 4-4 for how to select genes. See “Making Lists from Selected Genes” on page 6-14 for more details on this method of making a gene list.

Making Lists from Conjectured Regulatory Sequences

Once you have found possible regulatory sequences using the *Find Potential Regulatory Sequences* window (see “Regulatory Sequences” on page 6-18 for more details) and are inspecting one of the sequences in the *Conjectured Regulatory Sequence* window, you can make a list of all of the genes containing that sequence by selecting **List > Make Gene List**. See “Using the Conjectured Regulatory Sequence Window” on page 6-23 for more information.

Setting Preferences

The preferences screen allows you to change GeneSpring's global preferences. Note that some changes may not take effect in the currently open window or in your current GeneSpring session. Saved changes in the preferences window will not take effect until GeneSpring is restarted.

Select **Edit > Preferences**. To change any options in the Preferences window, click the appropriate tab to view the available settings.

Data Files

On this tab you can set the defaults of what you would like to see when GeneSpring opens. Set the defaults on this tab to have GeneSpring open directly to your chosen genome.

- **Data Directory**—The directory containing all GeneSpring data, including the genome that opens at startup. Use the browse button or the Navigator to choose the directory.
- **Load Sequence**—Load nucleic acid sequences with the genome data.
- **Suppress warnings about ambiguous gene identifiers when opening experiments**—Check this box to suppress the ambiguous gene identifier warning message (not recommended). For more information on ambiguous gene identifiers, see “Ambiguous Gene Identifiers” on page 2-9.
- **Default Genome**—The default genome to open when you start GeneSpring.
 - Select No default genome to be prompted for the genome to open each time you start GeneSpring.
 - Select Open the genome that was last used in the previous session to default to the last genome opened.
 - Select Open a specific genome to specify a default genome to open every time you start GeneSpring. To change this value, select the desired genome from the displayed directory. (On MacOSX, this menu is not displayed. To select a genome, click the browse button.)

Database

Use the pull-down menu to specify how GeneSpring assigns parameters for a series of numeric values in your database. You must also specify the fully qualified classname of the driver in the JDBC driver field.

Use the pull-down menu to specify how GeNetViewer assigns parameters for a series of numeric values in your database. You must also specify the fully qualified classname of the driver in the JDBC driver field.

Color

From this tab you can change the colors GeneSpring uses to represent different types of data and other screen elements. There are a variety of default color schemes available to choose from. The brightness of a color depends on the trust associated with it. For more information, see “Trust” on page 4-30.

Over- and under-expression color refers to the coloring of genes as shown in the genome browser and color bar. To change the definitions of overexpressed (upregulated) and underexpressed (downregulated) genes, right-click over the colorbar in the main genome browser. See “Changing the Colorbar Range” on page 4-32 for more details on this topic.

The colors you choose are blended to create a continuous spectrum from High to Normal to Low expression values.

There are two sections on this tab: Standard Colors and Group Colors.

Standard Colors

These are the colors applied to general display options.

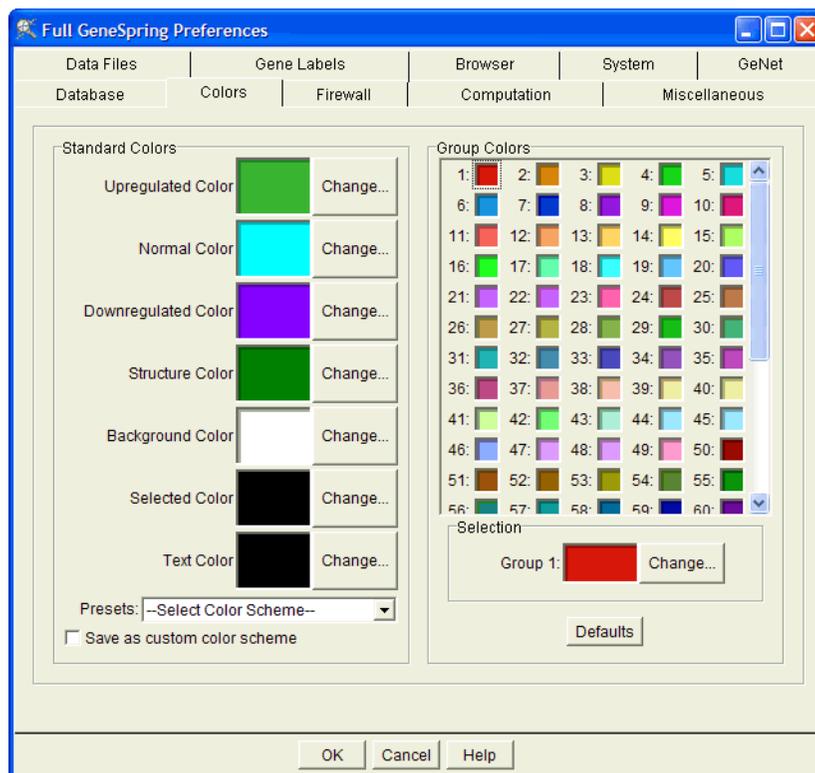


Figure 1-5 The Colors section of the Preferences window

- **Upregulated Color**—The Upregulated Color is the color used to display genes greater than or equal to the *High Expression* value selected for the current color bar.
- **Normal Color**—The Normal Color is the color used to represent genes having a normalized expression value of one. This is the only setting for which you can specify “no color”.
- **Downregulated Color**—The Downregulated Color is used to display genes less than or equal to the *Low Expression* value selected for the color bar.
- **Structure Color**—The Structure Color is used for the ConditionLine and for the lines between the genes in the Physical Position View, the Tree lines, the Ordered List lines, etc.

- **Background Color**—The Background Color defines the color behind the genes and other elements in the genome browser.
- **Selected Color**—The Selected Color is used for selected genes, gene names, and axes. For this, you will probably want the greatest contrast with the background color.
- **Text Color**—Defines the color of text displayed in the Genome Browser window.
- **Presets**—The Presets pull-down menu allows you to choose from a variety of pre-defined color schemes.

To create a custom color scheme, modify colors as desired and check the **Save as custom color scheme** box. This saves your current color scheme in the Presets menu under the name “Custom Color Scheme”. You can save only one custom color scheme at a time.

Group Colors

This section allows you to set colors used for the following:

- Classifications
- Parameters
- Gene Lists
- PCA
- Gene Inspector
- Find Similar Samples/Color by Attribute

Each box in the displayed grid indicates a color for that group.

Click on a box to select it. The Selection area at the bottom of the panel displays that box with the name and color of the selected group. Double-click the selected box or click **Change . . .** to view the Change Color window.

To restore the color defaults, click the **Defaults** button.

For more information on the various color options, see “Coloring” on page 4-47.

Specific Color Definition

You have the option to define your own colors to use in the genome browser. If your printer requires exact color definitions, specify them on this screen.

To change or adjust a color, select the **Change** button next to its element in the Preferences Color window.

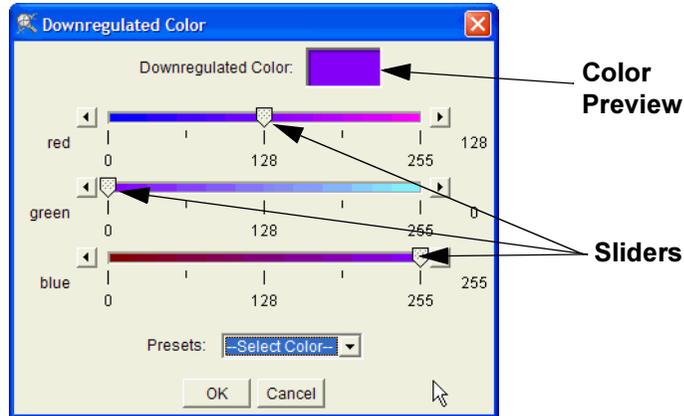


Figure 1-6 Color creation in the Preferences window

Click over any slider and move it horizontally to adjust the color. Watch the color preview box and stop moving the cursor when the desired color is reached. Click **OK** to accept the new color.

The checkbox **Specify no color** is only available for the “Normal Color” settings.

Gene Labels

On this tab you can specify how you would like to name your genes in the genome browser. The defaults are systematic name and common name. To change the defaults, select a new option from the drop-down menus. To restore the original defaults, click **Default Gene Labels**.

Browser

On this tab, specify default web browser settings if you want to use a particular browser for the GeneSpring application. You only need to set the **Arguments** option if you are using an obscure web browser that requires an argument.

Firewall

If your company has a firewall, you may need to specify settings to allow GeneSpring to access outside networks.

Click **Configure Automatically** to have GeneSpring attempt to automatically detect the appropriate settings. If the settings GeneSpring chooses do not allow you to reach the Internet, you may need to alter these settings.

The following settings are available:

- **Protocol**—The firewall protocol to use. Options are HTTP, SOCKS4, and SOCKS5.
- **Proxy Host Address**—The host address of the computer on which the firewall exists. This can be either a fully qualified hostname (i.e., *hostname.domainname.com*) or an IP number.
- **Proxy Port Number**—The port on which to connect to the firewall host.

- **Password Authenticate Connections**—Specify whether a password is required to connect to the firewall host.
- **Username**—The username used to connect to the firewall host (if required).
- **Password**—The password used to connect to the firewall host (if required).

If you are unsure how to proceed, contact your System Administrator for details about your firewall.

System

The System tab allows you to specify a number of different parameters regarding networking and memory usage.

- **Limit Filenames to less than 32 characters**—Allows you to limit the length of file names. This is a useful default setting for Macintosh users, since MacOS does not accept file names longer than 32 characters.
- **License Server**— Allows you to specify the IP address of the machine that dispenses concurrent licenses.
- **Desired Memory Use**— Allows you to set the amount of RAM GeneSpring attempts to use. If this value is set too high with respect to total available memory, unnecessary disk caching occurs and performance will be slow.
- **Disk Cache Size**—Specifies the amount of hard disk space GeneSpring uses to temporarily store HTML pages accessed by the GeneSpider or by other internet-based search functions.

Silicon Genetics recommends that you set this value to 10% of your available disk space. In GeneSpring 6.0, experimental data is loaded into the disk cache instead of into system RAM. GeneSpring now loads only the data currently being used into memory. This enables GeneSpring to handle much larger experiments containing any number of samples.

- **Cached Internet Resources Expire After**—Specifies how long GeneSpring caches copies of Internet resources for quicker access.
- **Number of Processors**—Specify the number of processors in your computer. This allows several types of analysis (including k-means, build gene trees, promoter search, and predict parameter values) to be used most efficiently.

GeNet

On this tab you can specify the default GeNet server to connect to and enter the addresses of other GeNet servers to which you want GeneSpring to have access.

- To have GeneSpring connect automatically to the default GeNet server at startup, check the **Login to GeNet at Startup** box. Select the default GeNet server from the pull-down menu.
- To have GeneSpring invoke the Bulk Upload to GeNet window when you quit GeneSpring, check the **Remind me to upload new data to GeNet** box.

- When you create a new experiment in GeneSpring using samples from a GeNet server, GeneSpring saves local copies of those samples by default. This can cause slow performance when saving large experiments. To disable this feature, uncheck the **Save GeNet Samples Locally When Creating an Experiment** box.

To enter a new GeNet server, click **New . . .** and enter the following information:

- **GeNet Server Name**—The name of the server to connect to
- **GeNet Server Address**—The IP address of the GeNet server (Enter the numeric address only, i.e., 127.0.0.1. Do not enter “http://” before this address.)
- **Default Username**—The username with which to connect to GeNet
- **GeNet & GeneSpring on**—Specify whether the GeNet server is on the same side of your firewall as GeneSpring or not.
- **Use Secure Connection**—Specify whether communication between GeneSpring and GeNet should be secured or not. If it is secured, communication between GeNet and GeneSpring uses HTTPS (which uses the SSL library available in Java).

To edit an existing GeNet server, select it from the list and click **Edit . . .** To delete a GeNet Server, select it from the list and click **Delete . . .**

Computation

On this tab, specify settings for how to run scripts.

- **Default Computation**—Select Local to have scripts run on your local machine by default. Select Remote to run scripts on a remote execution server by default.
- **Local Computation Settings**—Check **Don't Show Script Result Summary Window** to skip the Script Result Summary when a script completes its execution. The **Current Scale Factor for Time Estimate** option allows you to modify the multiplier for GeneSpring's internal estimate of how long an analysis will take. This can be useful if you have significantly changed hardware or settings on your computer (added more RAM, etc.).
- **Remote Computation Settings**—Check **Automatically check for results** to have GeneSpring automatically check whether your script has finished running. Specify how often to check by entering a number of minutes in the **Delay between checks** box.

Miscellaneous

The Miscellaneous panel contains a variety of settings to customize your GeneSpring installation.

- **Default Minimum Correlation**—Specifies the default minimum correlation coefficient that appears near the Find Similar button in the Gene Inspector window.
- **Restrict Gene List Searches**—Allows you to limit the lists GeneSpring examines when searching for similar lists in the Gene Inspector window and during Tree building.

- **Search Gene Lists Stored**—Specify whether to search gene lists stored on your local machine, on a GeNet server, or both.
- **Use the Cross-Gene Error Model by Default in Experiment Interpretations**—Check this to use the Cross-Gene Error Model by default in experiment interpretations.
- **Font Name and Font Size**—Specify the style and point size of the default display font.
- **Reset Font**—Click this button to reset the display font to its default value.
- **Language**—Allow you to select from the available language choices for the GeneSpring interface. If your computer is set for a specific language, use the same setting here.
- **Your Name, Your Group Name, Your Email**—Specify the name, group name, and email address values contained in the HTML files that go into your data directories.
- **Entrez mirror**—Enter a web address.

Creating Genomes

In the context of GeneSpring, a genome contains information about all the genes in your chip or microarray setup. Note that a GeneSpring genome does not correspond exactly to the biological definition of a genome. A genome in GeneSpring is composed of discrete genes as opposed to the full nucleotide sequence. This means that a GeneSpring genome can contain two genes representing alternately spliced variants of a single gene, whereas a true genome would include the DNA sequences for only one.

Setting up a genome is usually the first step in the analysis workflow.

There are two ways to set up a genome:

- Request one from Silicon Genetics technical support
- Use the Genome Installation Wizard described in “The New Genome Installation Wizard” on page 2-2

Once you have set up a genome, you are ready to begin loading samples and building experiments. Key information that these Genome files include:

- a list of annotations (what the scientific community knows about each gene, including information used for building ontologies)
- a list of “gene hypertext links” (URLs from which you can find more information about each gene from public databases)
- mapping information about where each gene appears on a given chromosome
- identifiers (accession numbers) for the genes in various public databases

The New Genome Installation Wizard

The Genome Wizard guides you through the steps of creating a new GeneSpring genome. Most of these screens are self-explanatory. Which screens you see as you proceed through the Genome Wizard vary depending on the information you provide.

1. Select **File > New Genome Installation Wizard**. The New Genome Installation Wizard window appears.

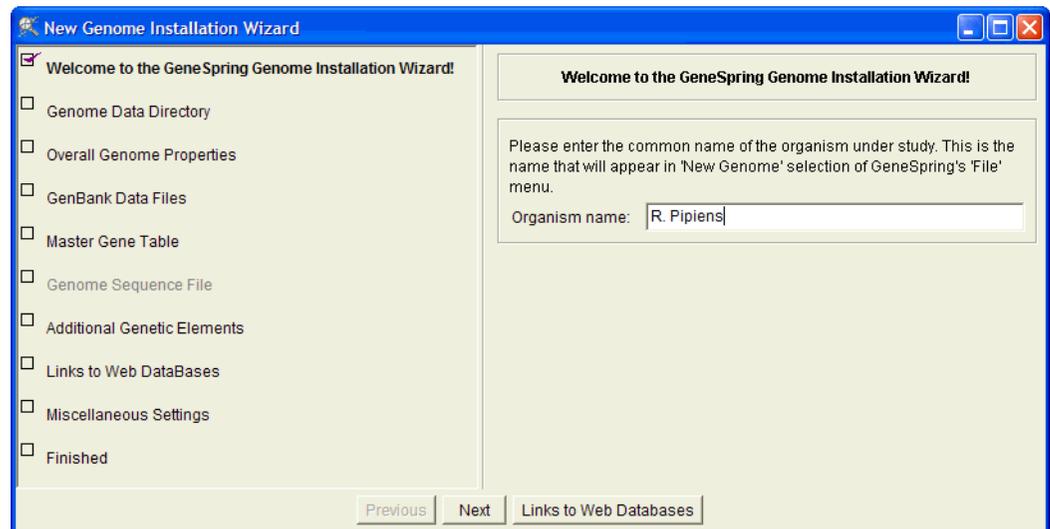


Figure 2-1 The New Genome Installation Wizard.

2. Enter a name for your new genome. Be sure the name you enter is descriptive. GeneSpring creates the genome using the capitalization and spelling you enter at this time.. Click **Next**. The Genome Data Directory window appears.
3. Select a directory in which to save your new genome, or create a new one. By default GeneSpring displays a new directory name in this field, using the same name you entered in the previous screen.
 - To accept the default directory, click **Next**.
 - To change the default name, enter a new directory location in the text box. If you enter a directory name that does not exist, GeneSpring creates it for you. If you leave the text box blank, GeneSpring saves your genome to the default location..
 - To select a different directory, click **Browse** and navigate to the desired directory.

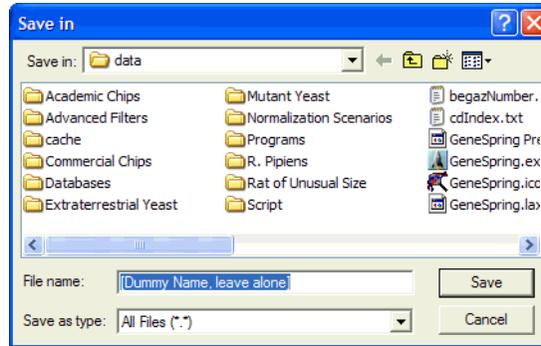


Figure 2-2 The Save In dialog

Note: When you click **Browse**, the words “Dummy Name, leave alone” appear in the File name field in the **Save in** window. This is expected behavior. Do not change this text. If you accidentally click on a filename in this window, the name of that file replaces the dummy text, and clicking **Save** generates an error message. If you get this error message, click **Yes**. *This does not replace the file you clicked.* It simply enters the correct directory name in the Specify Directory box.

When you click **Next**, the Overall Genome Properties window appears.

4. Specify genome properties:

- **If your organism has been sequenced and you have a file containing the full sequence**, select **Yes** in the first box. If not, leave **No** selected.
- **If your organism is a circular genome** (such as a bacterium, plasmid, or virus), select **Yes** in the second box. This tells GeneSpring to display your genome as a circle in the physical position display. If your organism does not have a circular genome, leave **No** selected.

When you have made your selections, click **Next**. The GenBank Data File screen appears.

5. Specify whether you are using a GenBank file as your data source, and, if so, the name of the file. If you are using an EMBL file, select **Yes** as you would for a GenBank file.

- **If you are using a GenBank or EMBL file**, select **Yes**. You are prompted to enter the filename. You cannot proceed until you have entered a filename. Type the complete path and filename, or click the **Browse** button to select it from a directory.
- **If you are not using a GenBank or EMBL file**, leave the **No** radio button selected.

Click **Next** to proceed to the next screen.

6. The Master Gene Table screen appears. (This screen appears only if you are not using a GenBank or EMBL file as your data source.)

On this screen, enter the name of your master gene table. Type its complete path and filename in the text box or click **Browse** to select it from a directory. You cannot proceed until you have entered the filename of a valid master gene table on this screen.

Note: Your master gene table must be in a name list, name function, SGD, or mapped format. For more information on these date formats, see “Data Format” on page 2-10.

Once you have entered the correct information, click **Next**.

7. The Genome Sequence File screen appears. (This screen appears only if you indicated on the Overall Genome Properties screen that your genome has been sequenced, and you are not using a GenBank or EMBL file.) On this screen, you specify where GeneSpring should look for the sequence data.

Place your cursor in the Enter Genome Sequence File Name box and type the complete file name and pathway, or click **Browse** to select the file from a directory. You cannot proceed to the next screen until you have entered a file name.

Once you have entered the correct information, click **Next**.

8. The Additional Genetic Elements screen appears. On this screen, specify whether you have a second table of genes. This is generally used to add genetic elements to a GenBank or EMBL-defined organism. In this case the supplementary table of genes probably contains alleles, centromeres, or genes from strains differing slightly from the sequenced strain.

- **If you do not have a separate table of genes**, leave **No** selected.
- **If you have a separate table of genes**, select **Yes**. You are prompted to enter a file name and select a file format. Enter the complete filename and path, or click **Browse** button to select a file. Then select the appropriate format from the **Select a file format** menu. For a description of the four format options, see “Data Format” on page 2-10.

When you are done, click **Next** to proceed to the next screen. The Links to Web Databases screen appears. From this screen, you can create a link to a web page or other online resource with relevance to your genome.

9. Click the Links to Web Databases button to view a table of commonly used links. You can copy these links and paste them into the table of links. This list is also available from http://www.silicongenetics.com/cgi/TNgen.cgi/GeneSpring/GSnotes/Notes/have_links.

If you do not want to include links to web databases, click **Next** to proceed to the next screen.

If you want to include links to web databases, select **Yes**.

- In the **Enter number of links** box, type the number of web databases to link to. When you enter a number in this box, the number of “Button” lines in the table below changes to match the number you entered. You can change this number at any time while you are on this screen.
- In the first column of this lower table (titled **Button label**) enter the name of the web database as you wish it to appear on a button within GeneSpring.
- In the right-hand column (titled **URL**), enter the URL of the database, with the systematic name of the gene replaced by a semicolon. If the semicolon representing the place the systematic name of the gene should go is at the end of the URL, it may be omitted.

You can also have links using names other than the systematic gene name. To use one of these, attach a special character before the link name (in the **Button label** column). Do not put a space or other character between the special character and the link name. To use the common name, use a dollar character (\$). To use

the GenBank Accession Number, use a percent sign (%). To use the systematic name, less anything after a dash, use the dash (-).

You can use any column in the Master Table of Genes in a link by entering `<name-of-column>` at the desired point in the url, where name-of-column is the name of the column you want to use.

When you right-click on this screen, there is no pop-up menu allowing you to cut and paste. However, you can still cut and paste URLs into the matrix fields by using the keyboard commands (for Windows this is Ctrl+C and Ctrl+V). Cutting and pasting is advised to ensure that URLs are properly entered.

Note: GeneSpring attempts to locate each URL you insert before it allows you to proceed to the next panel. This may be a problem if you are not connected to the internet when you are creating this genome. In this case you will have to skip this screen and add the web-links to the `.genomedef` file later. To add hyperlinks from GeneSpring, see Step on page 4.

GeneSpring cannot automatically locate the default web browser on NT or Macintosh systems. You must set the path manually. To set the path to the browser:

-Select **Edit > Preferences**.

-Click the **Browser** tab.

-In the **Browser path** box, either type the complete file name and pathway of the `.exe` file for your default browser, or click the **Browse** button to locate the proper executable, which is most likely located in the system directory. In a Windows NT environment your path may look something like this:

```
C:\Program Files\Plus!\Microsoft Internet\IEXPLORE.EXE
```

-Click **OK** to close the Preferences window.

When you are done with this screen, click **Next** to proceed to the Miscellaneous Settings screen.

10. From this screen, you can force all of the systematic gene names to appear in upper or lower case letters by selecting the appropriate checkbox. You do not have to select either of these options.

Click **Next**. The Finished screen appears.

11. Click the **Finish** button to save your genome.

Gene HyperText Links

The format for gene URLs in the `.genomedef` file has changed from earlier versions of GeneSpring. The new format is as follows:

```
GeneHypertextLinks: link:http://www.example.com&gene=<field1>&id=<field2>
```

where *link* is the name of the link (and must be followed by a colon (:), not a semicolon (;). Instances of `<field>` are replaced by the value of the specified parameter. The allowed parameters are:

- systematic
- common

- genbank,
- ec
- pubmed
- map
- chromosome
- synonyms
- description
- phenotype
- function
- product
- keywords
- dbid
- custom1
- custom2
- custom3

Links can be created using any column. Labeled format allows an unlimited number of columns.

A link is enabled for a particular gene only if all parameters mentioned in that URL are defined for that gene.

Experiment URLs work exactly the same way, except that they begin with `ExperimentHypertextLinks` instead of `GeneHypertextLinks` and the `<field>` variables contain names of parameters. A link is shown in the experiment inspector only if the experiment has parameters with names matching all fields in the URL.

In both cases, the parameter names are not case sensitive. Thus if an experiment has a parameter called Time, you can specify it as `<time>`, `<Time>`, or `<TIME>` in the URL.

In earlier versions of GeneSpring, URLs were specified in the Gene Inspector Window like this:

```
GeneHypertextLinks : #linkname;http://www.example.com&gene=;&org=Hs
```

where # is a symbol specifying the source of the gene annotation to query with: '%' to query with GenBank locus (column 10 of master gene table), '\$' to query with common name (col. 2), '-' to query with systematic name less anything after a dash.

If none of these symbols appears before the link name, the button automatically queries the designated database with the full systematic name. These names are added to the end of the specified URL. To drop the chosen gene identifier within the URL into a specific spot, mark the spot with a semicolon, ';'.

GenBank or EMBL Files

If you use a single GenBank file to describe a genome, you need not use a master gene table and therefore do not have to enter any of the information discussed in “Data Format” on page 2-10. You also do not need a separate file to contain the sequence data (the files for sequence data are described in “Sequence Data” on page 2-7).

The GenBank file can be downloaded directly from GenBank, if you open a web browser to the URL of the organism you are installing. For example, “*ecoli.gbk*” is a 9.5-MB file, from the URL:

```
ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/Ecoli/
```

This URL is usually the same for all of GenBank’s bacterial genomes, with the name of the organism you are installing in place of “Ecoli”. This URL may contain many file formats. Make certain to download the file with the suffix *.gbk*. An EMBL file may be used in place of a GenBank file.

Adding Extra Genes to a Genome Defined by a GenBank or EMBL file

You can use a GenBank or EMBL file to describe a genome and add extra genes. This is typically done to represent a strain slightly different from the sequenced strain. To do this you must create a separate master gene table containing all of the extra genes to add. Format these tables using one of the four table of genes formats discussed in “Data Format” on page 2-10. This file is parsed during genome creation but is not used again afterward. Contact Silicon Genetics technical support at 1-866-SIG-SOFT for help with this process.

If you are using an original *.gbk* file, you can simply go to their web site and update the entire file. Make sure you save it with the same name and to the same place as your current *.gbk* file.

Updating GenBank Information

After loading your data into GeneSpring you may want to update your annotations. For information on this procedure, see “Updating Annotations with GeneSpider” on page 6-27.

Sequence Data

GeneSpring loads in sequence data from a GenBank or EMBL file automatically. If you have sequence data that is not in a GenBank/EMBL file, place it in a separate file using the *.seq* format.

The Silicon Genetics *.seq* format is similar to the FASTA format, although there are some differences. The *.seq* format consists of one line of identifiers followed by lines of sequence. The identifier line consists of the "Greater than" sign (>) followed by the chromosome identifier, followed by a space which is followed by an optional description. An example is given here.

```
>CHR1 This is the description of Chromosome 1
GCTGACGGACTTTCTAGCGGTCTAGCAACTGAGCGGCGCGGGCATCGTA
CAGCAGCGAGCTACTATCTACGCGGCGGATATAAACTACAAAAAAA
```

Chromosomes in GeneSpring are given a number (1, 2, 3 etc.) and the number should be part of the Chromosome identifier. The Chromosome identifier can optionally contain the letters 'CHR' but is not required. The number used in the *.seq* format for the chromosome has to correspond to the number used in the Map position in the Master Table of Genes.

The *.seq* format is not the same as the FASTA format. There is an example of the FASTA format at <http://www.ncbi.nlm.nih.gov/BLAST/fasta.html>.

The New Genome Installation Wizard

A severely abridged example of the yeast .seq file might look like this:

```
>CHR1 Chromosome I data:
CCACACCACACCCACACACCCACACACCACCACCACACCACACCCACACACACA . . .
GTGGGTGTGGTGTGGTGTGTGGGTGTGGTGTGGGTGTGGTGTGTGTGGG
>CHR2 Complete DNA sequence of yeast chromosome II.
AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGA . . .
AGAATAGGGTACTGTTAGGATTGTGTTAGGGTGTGGGTGTGGTGTGTGTGGG
TGTGGTGTGTGGGTGTGT
>CHR3 LOCUS          SCCHRIII   315341 bp      DNA          PLN
25-NOV-1996
CCCACACACCACACCCACACCACACCCACACACCACACACACCACACACCCA . . .
AGTGTGTGGGTGTGGGTGTGTGGGTGTGGTGTGTGGGTGTGGTGTGTGTGGTGT
GTGGGTGTGGGTGTGTGGGTGTGGTGGGTGTGGTGTGTGTG
```

Name multiple chromosomes sequentially, i.e., CHR1, CHR2 and so on. If there is only one chromosome, name it CHR1.

Creating a Genome from Experiment Data

In GeneSpring, a genome includes all the genes on your chip. When you create a genome from experiment data, GeneSpring creates a genome on the fly based on genes in your experiment data files. This means that unlike a genome created in the New Genome Installation Wizard, this genome has no annotations and no means of obtaining annotations from public databases.

The genome consists of a master table of genes and a genome definition file. If you create a new genome after accepting a file format recognized by GeneSpring, anything not standard to that recognized format is not included in the master table of genes. (The master table of genes contains all the information associated with genes in a given genome.)

For example, if GeneSpring recognizes an Affymetrix file, but that file has GenBank accession numbers, the numbers are not loaded. You can add these numbers later to the GenBank column of the annotations file. (If your data files have a description column, GeneSpring includes it in the master gene table.) Clontech Atlas 2.0 and Incyte GEM Tools 2.4 have a GenBank Accession number column that is loaded into the master table of genes.

If you have difficulties creating a genome in this way, use the New Genome Installation Wizard. See “The New Genome Installation Wizard” on page 2-2.

Creating a New Genome

1. Select **File > Import Data**.
2. Choose the data file to load.
3. Specify the file format. For details, see “Importing Experiment Data” on page 3-2.
4. Select **Create a New Genome** and enter a name in the **Choose a Name** field.

You have the option to load additional files to the experiment. Choose the files to load. GeneSpring gives you the option of adding any new genes to the genome.

Ambiguous Gene Identifiers

When the gene identifier specified during data import is not unique to a single gene in the genome, GeneSpring can not determine which gene the measurement is for. In this case the identifier and all corresponding genes to are listed in the Ambiguous Gene Identifiers table and the measurement is not loaded.

To prevent this problem, edit the raw data file(s) to assign a gene identifier to each gene that is unique in the genome (the systematic gene name in the genome is always unique) and then re-import your data.

Contact Silicon Genetics Technical Support at 1-866-SIG-SOFT if you have further questions or experience difficulties.

Data Format

This section describes the format of the files created by GeneSpring during the genome creation process. This information may be helpful if you need to create or edit these files manually.

The Master Gene Table File

- To set up gene annotations, create a table with the following structure, using the tab-delimited text format:

Systematic Name	Common Name	Map	EC Number	Description	Product	Phenotype	Function	Keywords	GenBank Accession	Synonym	Sequence	PubMed ID	Custom 1	Custom 2	Custom 3	Type
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q

Note: There are five additional columns which are not displayed above due to space constraints. These are:

- DBid—18 (R)
- GO Biological Process—19 (S)
- GO Molecular Function—20 (T)
- GO Cellular Component—21 (U)
- RefSeq—22 (V)

For example, if you had Gene Identifiers (as the genes would be identified in a raw data file) and GenBank Accession number for all the probes on a chip, the table viewed in Excel might look something like this:

A1i																	L14754	
A1j																		X79882
A1k																		D78579
A1l																		M31630
A1m																		J04111
etc.																		etc.

Opened as a tab-delimited text file, the table might look like this:

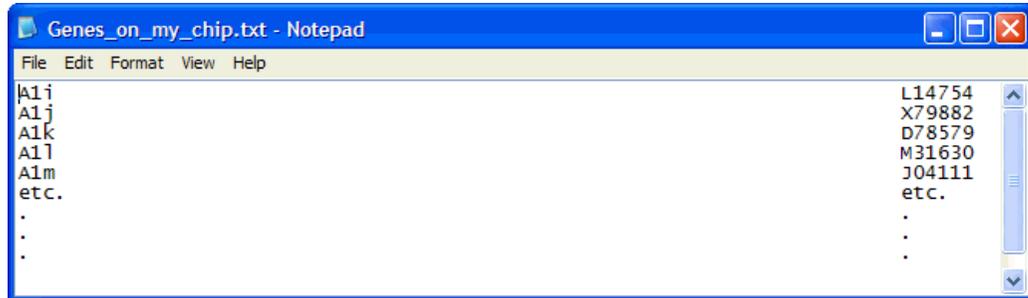


Figure 2-1 The Master GeneTable as tab-delimited text

2. Once the table is formatted, open to GeneSpring and select **File > New Genome Installation Wizard**.

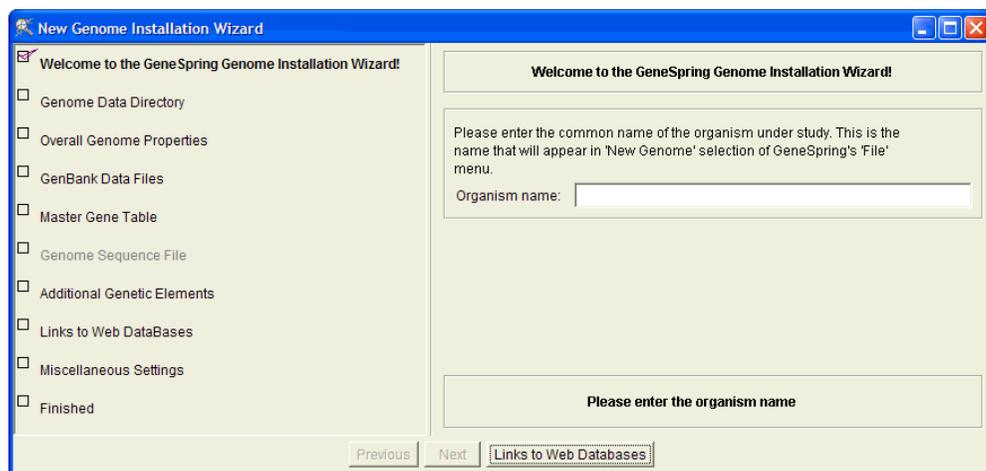


Figure 2-2 The New Genome Installation Wizard

3. Answer the series of questions, such as the genome name (i.e. 'Human custom'), whether the genome is circular or a series of linearized chromosomes, etc.
4. When you reach the 'Master Gene Table' screen, click **Browse** and select these tab-delimited text file you created in step 1.
5. Continue the genome installation process.

Once you are finished, the new genome can be selected using **File > New Genome or Array**. Click on any gene in the Genome Browser to invoke a Gene Inspector window from which resident gene annotation can be viewed.

If the annotations you imported include GenBank accession numbers, select **Annotations > GeneSpider** to import additional gene annotations available through NCBI.

Layout Parameters

The .layout file

To create an array layout file in GeneSpring, you need at least one file to tell GeneSpring general information about the array (size, shape, features, format, name, etc.). This file should end in the extension `.layout`. You usually need another file describing exactly which gene goes where.

The format of the `.layout` file is a series of lines. Order does not matter. Each line consists of a property, a colon, and a value.

For example, `property : value`. Blank lines and lines starting with a number sign (#) are ignored by GeneSpring. The following properties are allowed in the file. As always, GeneSpring is case-sensitive, so use the capitalizations as presented here:

- **Name**—The name of this layout, to appear in the navigator window of GeneSpring.
- **Icon**—(optional) The path of a 16 x 16 `.gif` file to appear next to the layout in the navigator window.
- **VerticalSubArrays**—(optional, default 1) The number of rows of sub-arrays.
- **HorizontalSubArrays**—(optional, default 1) The number of columns of sub-arrays.
- **HorizontalPerSubArray**—The number of columns of dots in a sub-array.
- **VerticalPerSubArray**—The number of rows of dots in a sub-array.
- **VerticalDuplication**—(optional, rarely used) When dots are duplicated vertically, the number of copies.
- **HorizontalDuplication**—(optional, rarely used) When dots are duplicated horizontally, the number of copies.
- **CommonArrayType**—The format of the array.
 - **Q-X-Y**—The data file contains two columns. The first is a list of genes, the second is a set of three numbers separated by commas or hyphens. The first is the “sub-array” number, the second is the X-coordinate, and the third is the Y-coordinate. All numbers start counting from 1. The subarrays are counted left to right, top to bottom. The second column can optionally be enclosed in quotation marks.
 - **Q-R-C**—Same as “Q-X-Y”, except the X and Y coordinates are swapped.
 - **CLONTECH LNL**—There is no datafile. All genes have systematic names of the form “B4c” indicating where they are in the array. The first (capital) letter indicates which sub-array; the number indicated which column, and the lower case letter indicates which row.
 - **CLONTECH LNNL**—Same as LNL, except there are two digits instead of one.
- **DataFileName**—The name of a datafile linking locations with gene names in format given by the CommonArrayType choice.

Once you have created the `.layout` file, save it in the ArrayLayouts folder of the genome folder for which the layout pertains.

For example, if you have not changed the defaults set-up of GeneSpring the path to the layout folder in the yeast genome is C:\Program Files\SiliconGenetics\GeneSpring\data\Demo Chips\yeast\ArrayLayouts.

Examples of *.layout* files for Arrays

Here is an example for Pat Brown's yeast layout. The following is from a file Pat.layout:

```
Name : Pat Brown's Yeast Layout
# Icon: XXX.gif
VerticalSubArrays: 2
HorizontalSubArrays: 2
HorizontalPerSubArray: 40
VerticalPerSubArray: 40
VerticalDuplication: 1
HorizontalDuplication: 1
CommonArrayType: Q-X-Y
DataFileName: PatLocationList.txt
```

Following are the first few lines of the file PatLocationList.txt:

```
YHR007C "1,13,1"
YBR218C "2,13,1"
YAL051W "1,14,1"
YAL053W "2,14,1"
YAL054C "1,15,1"
YAL055W "2,15,1"
YAL056W "1,16,1"
```

Here is an example for a CLONTECH Array, from a file Clontech.layout:

```
Name: Clontech 588
# Icon: XXX.gif
VerticalSubArrays: 2
HorizontalSubArrays: 3
HorizontalPerSubArray: 14
VerticalPerSubArray: 14
VerticalDuplication: 1
HorizontalDuplication: 2
CommonArrayType: Clontech
```

Making an array can be a complicated process. Contact Silicon Genetics Technical Support at **1-866-SIG-SOFT** or **support@silicongenetics.com** for more information on this topic.

Renaming and Deleting Genomes

GeneSpring saves information about each genome in the data subdirectory of the GeneSpring folder. For example, on a Windows system, this might be `C:\Program Files\SiliconGenetics\GeneSpring\data\`. This directory contains a folder for each genome contained in GeneSpring.

Renaming a Genome

To rename a genome in GeneSpring:

1. Using a text editor, open the `.genomedef` file for the desired genome. This file is located in the genome's folder in the GeneSpring data directory.

For example, the `.genomedef` file for the Extraterrestrial Yeast genome might be located in `C:\Program Files\SiliconGenetics\GeneSpring\data\Extraterrestrial Yeast\ExtraterrestrialYeast.genomedef`.

2. The genome name is set in the first line of this file. For the Extraterrestrial Yeast genome, it would look like this:

```
Name : Extraterrestrial Yeast
```

3. Delete the existing name and enter the new name, i.e., "Name: Martian Yeast".
4. Save your changes and exit. Your changes will appear the next time you start GeneSpring.

Deleting a Genome

Because each genome's folder contains all of the information associated with a genome in GeneSpring, including experimental data and annotations, do not delete a genome unless you are absolutely positive no one is using any of the data it contains.

You can remove a genome from GeneSpring without deleting its data by moving its folder to another location outside the GeneSpring data directory, such as a temporary directory or your own user directory. The genome will not appear the next time you start GeneSpring.

To restore a genome removed in this way, replace its folder in the GeneSpring data directory. The genome reappears the next time you start GeneSpring.

To permanently remove a genome, delete its folder from the GeneSpring directory.

Working With Experiments

Importing Experiment Data

GeneSpring can load data from nearly any expression analysis technology, provided the data are formatted as tab-delimited text. The following section describes methods of loading data that is automatically recognized by GeneSpring as well as for loading data from a custom source.

GeneSpring automatically recognizes the formats of the following products:

- Clontech AtlasImage 2.0
- Affymetrix Metrixs
- Affymetrix Pivot
- Affimetrix MAS 5.0
- Axon GenePix Pro 2
- Axon GenePix Pro 3
- BioDiscovery Imagene 4
- Incyte Internet
- Incyte GEM Tools 2.4
- Packard Biochip ScanArray/QuantArray
- Agilent Feature Extraction
- Amersham CodeLink
- dChip

If GeneSpring is unfamiliar with your file format, you can define a custom format to specify the type of data in each column. These specifications can be added to the list of known file types so that you can load subsequent experiments in batches.

Make sure you use the raw, tab-delimited files just as they come out of the scanner. GeneSpring uses this information in the column headers. If you have cut out header information, use your original tab-delimited data files.

Memory Use for Experiment Loading

In GeneSpring 6.0, experiments are loaded into the disk cache as well as into system memory (RAM). This requires some additional time when an experiment is first created. Once the experiment is loaded, it can then be reloaded in a fraction of the time. This change was made to accommodate the loading and creation of very large experiments, especially for systems with limited memory.

If you want to free some hard disk space or think that your cached data folder may be corrupted, delete the cache folder (**GeneSpring/data/cache**, where “GeneSpring” is the GeneSpring home directory on your machine). This forces GeneSpring to recreate the experimental data, which may solve the problem.

Loading an Experiment

1. Select **File > Import Data...** or type **Ctrl+O**. (On Windows and Unix systems, you can also drag and drop files into the main GeneSpring window directly from your desktop. This method is not supported for Macintosh.)
2. Choose the data file or folder to load. All files in a folder must have exactly the same format.

The Define File Format window appears.

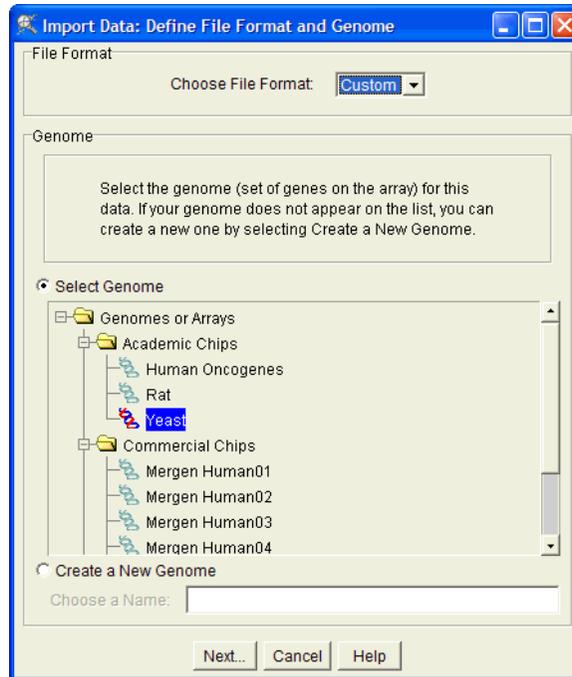


Figure 3-1 The Define File Format and Genome window

3. If the file format displayed in the **Choose File Format** box is correct, go to the next step. If not, select the correct file format from the pull-down menu.
4. From the Select Genome list, choose the genome in which to save the experiment data. To save the data in a new genome, select the **Create a New Genome** radio button and enter a name in the text box. For more information on creating a genome in this way, see “Creating a Genome from Experiment Data” on page 2-9.

If your data is in a known format, the Genome Browser window appears. For more information on the Genome Browser, see “Using the Genome Browser” on page 4-2.

If your data is in a custom format, the Column Editor appears. You must set up columns before continuing. See “Using the Column Editor” on page 3-9 for information on using the Column Editor.

5. Click **Next**. The Select Files window appears.

Importing Experiment Data

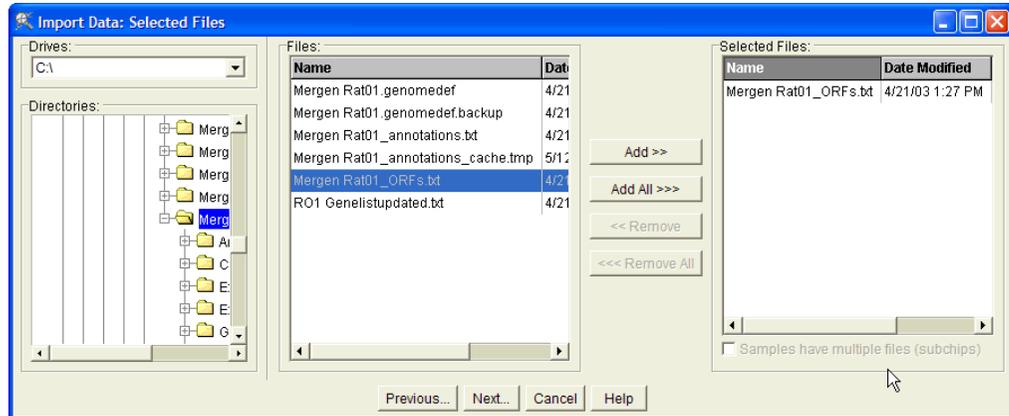


Figure 3-2 The Select Files window

From this screen you can select more files of the same type to add to your experiment.

To select files:

- Using the Drives and Directories menus, navigate to the folder containing the files you want to add. When you select a folder, the files it contains are listed in the Files section of the window.
- In the Files section, select the file or files to be added. To select multiple files, hold down the Ctrl key while clicking on the file names. You can use the same method to unselect one or more selected files.
- Click **Add**. The selected files are now listed in the Selected Files list. To add all files in the selected directory at once without selecting them individually, click **Add All**.

To remove a file or files from the Selected Files list, select them and click **Remove**. To remove all files from the list, click **Remove All**.

Note: These files must all be in the same format. GeneSpring verifies whether the format is correct, and if it is not, it does not add the files to your experiment.

- When you are done adding files, click **Next**.

If your signal and control files are in separate files, the Select Corresponding Files window appears. Proceed to Step 6.

If you have selected samples with multiple files (subchips) selected, the Merge Files window appears. Proceed to step 7.

If not, proceed to Step 8.

- In the Select Corresponding Files window, you can specify which signal file corresponds to which control file. (This step applies only to Imagen files.)

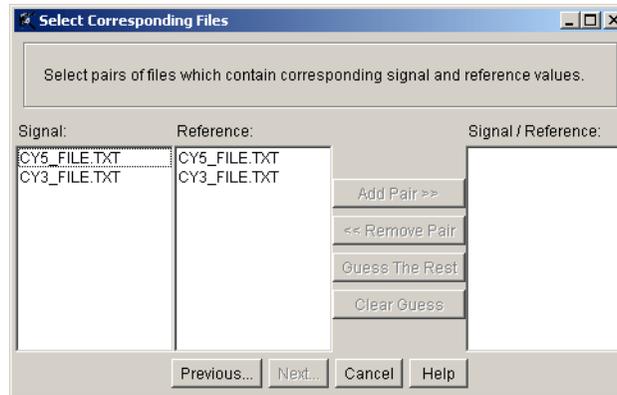


Figure 3-3 The Select Corresponding Values window

To do this:

- a. Select a filename in the left column.
 - b. Select the corresponding filename in the right column.
 - c. Click **Add Pair**. The pair you specified appears in the Signal/Reference list.
 - d. To have GeneSpring automatically select the rest of your files, click **Guess The Rest**. If GeneSpring guesses incorrectly, click **Clear Guesses**.
 - e. To remove a pair, select it in the **Signal/Reference** list and click **Remove Pair**.
 - f. When you are done, click **Next** and see Step 8.
7. **If you need multiple chips to cover one sample** (such as the Affymetrix Mu_u74 or Hu5 chip sets), the Merge Files window allows you to define all the files needed for each sample. This screen does not appear if it is not necessary.

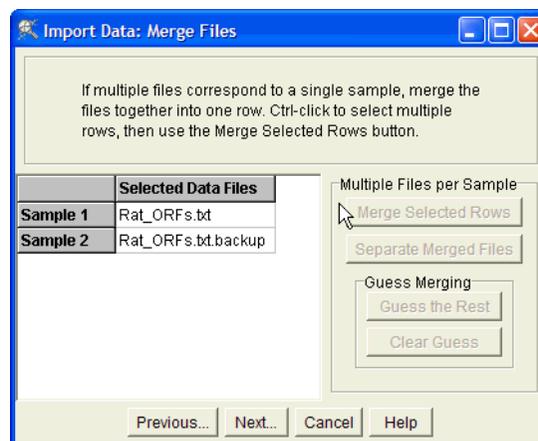


Figure 3-4 The Merge Files window

To merge files, select all the files that are on the same sample and click **Merge Selected Rows**. Use Ctrl+click to select multiple files in non-adjacent rows. You can also drag a file from one row and drop it in another to merge those two rows. To unmerge rows, click **Separate Merged Files**.

Importing Experiment Data

Click **Guess the Rest** to have GeneSpring try to match the pattern set by the names of the files you have already merged. If the guesses are incorrect, click **Clear Guesses**.

Click **Next** when you are done.

8. **If you imported genes that are not part of the selected genome**, the Extend Genome screen appears. If all the imported genes are part of the selected genome, this screen does not appear.

From this screen, you can specify whether or not to add those genes to the genome. To add the genes, click **Yes** and they are immediately added. This means that if you cancel the data loading process later, the genes are still part of the selected genome.

To skip the listed genes, click **No**.

9. **If you defined recommended or required attributes**, the Import Data: Sample Attributes screen appears. On this screen you must enter required attribute information before proceeding. You can also add recommended and optional attribute information.

You can also add new attributes at this time. For more information on sample attributes, see “Sample Attributes” on page 3-35.

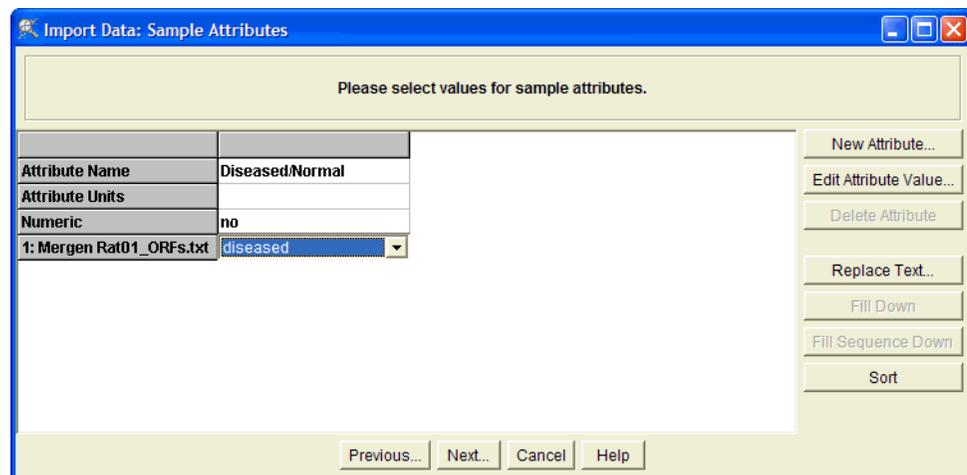


Figure 3-5 The Sample Attributes window

For attributes with standard values, a pull-down menu appears. Choose **Other** to turn the cell into a text field you can type in.

Click **Next** when you are done.

10. At this point, your new samples have been saved. You can either create an experiment using the new samples, or stop here.

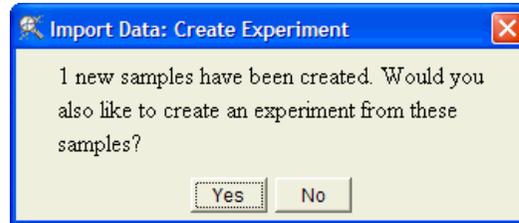


Figure 3-6 The Create Experiment dialog

To stop here, click **No**. The imported data are saved, but a new experiment file is not created. The data are saved as samples. The Sample Inspector displays each of the new samples. You can create experiments from these data later by selecting **Experiments > Create New Experiment**.

To create a new experiment, click **Yes**. You are prompted to enter a name and save the experiment.

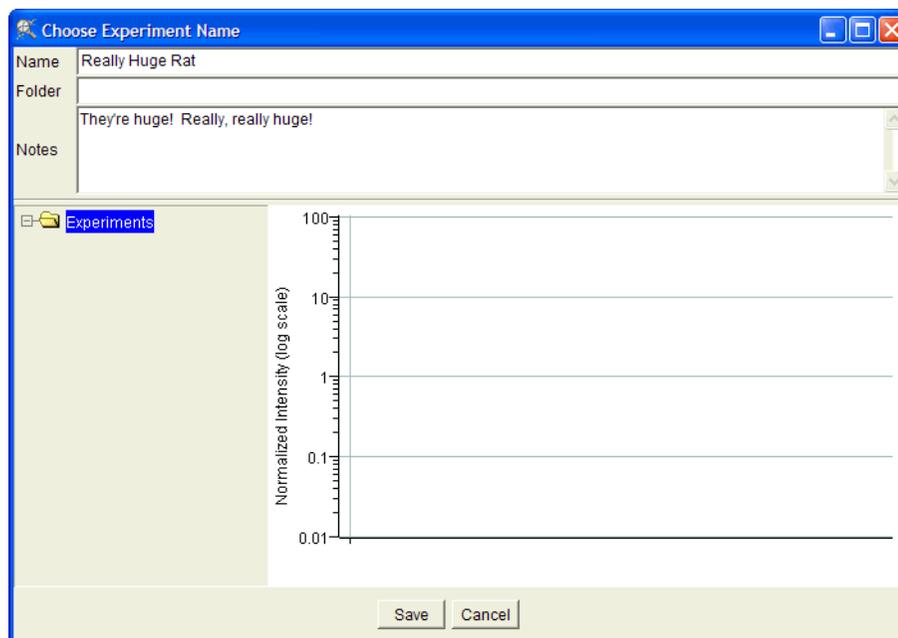


Figure 3-7 The Choose Experiment Name window

11. In the Choose Experiment Name window, do the following:

- Enter a name for the experiment in the **Name** field. Be sure to choose a descriptive name that you will remember later.
- To save the experiment in an existing folder, navigate to that folder in the directory browser in the lower left portion of the screen, and leave the **Folder** field blank. To save in a new subfolder, navigate to the desired parent folder and enter a name for the new folder in the **Folder** field.
- If desired, enter notes containing more descriptive information about the experiment.

When you are done, click **Save**. The New Experiment Checklist appears.

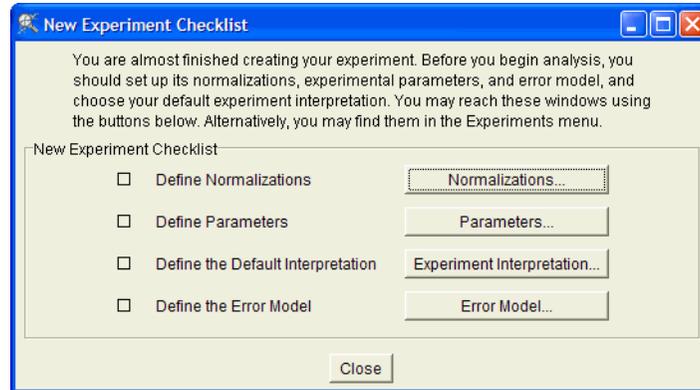


Figure 3-8 The New Experiment Checklist

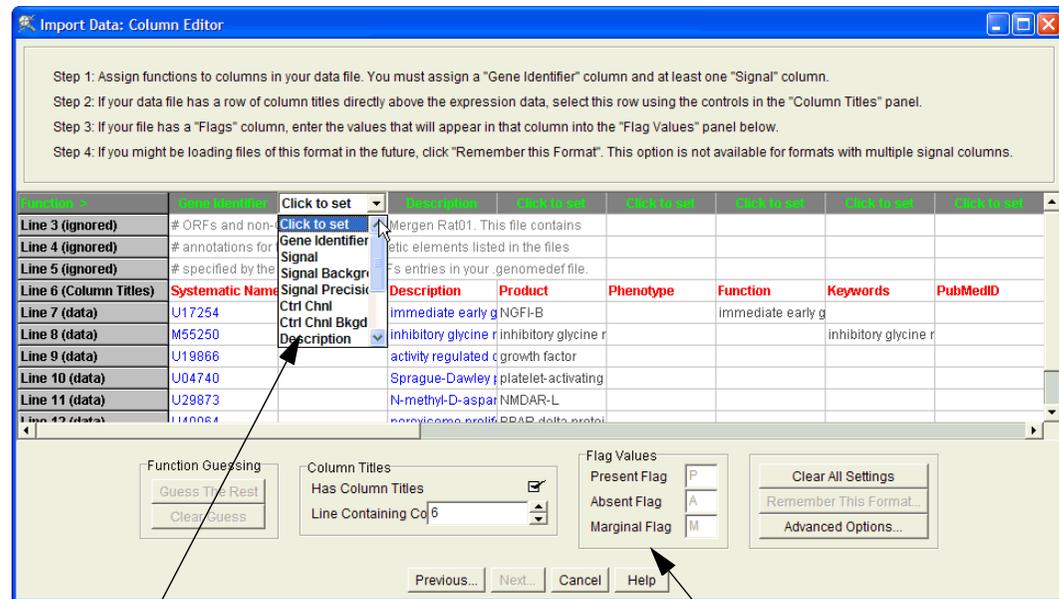
12. At this point, you can examine and change your normalizations, interpretations, and parameters.

- To define or edit normalizations, click **Normalizations...** For information on defining normalizations, see “Default Normalizations” on page 3-21.
- To define or edit parameters, click **Parameters...** For information on defining parameters, see “Experiment Parameters” on page 3-29.
- To define or edit default interpretations, click **Experiment Interpretation...** For information on defining default interpretations, see “Experiment Interpretations” on page 3-39.

If you prefer to make these changes later, click **Close**. You can load the experiment another time and select **Experiment Normalizations**, **Change Experiment Parameters**, or **Change Experiment Interpretation** from the Experiments menu in the main GeneSpring window.

Using the Column Editor

If GeneSpring does not recognize your file format, use the Column Editor to assign headings and functions to each column in your data file.



Function pull-down menu

Flag Translation Table

Figure 3-9 The Column Editor

When you first load a file, GeneSpring analyzes it to determine which row contains the column titles. If the row chosen is incorrect, use the **Line Containing Column Titles** field to adjust the number of rows.

If there are no column titles in your data file, uncheck the box marked **Has column titles**.

To set up columns:

1. Assign functions to each data column. Choose a function from the pull-down menu in each column. (See Figure 3-9 for an example.)

You must designate at least one Gene Name column and one Signal (raw data) column before the **Load Now** button becomes active.

Using the Column Editor

The available column assignments are listed below:

Name	Required?	# Allowed	Description
Unused	Optional	Any	These columns are not visible within GeneSpring, but can be used to filter data via the Filter Genes window. See "Filter on Data File" on page 6-61 for details.
Gene Identifier	Required	One	Gene identifiers must be unique to the genes in this genome. Duplicate genes are treated as replicates. It is recommended that the Gene Identifier in the raw data files be the gene's Systematic Name.
Signal	Required	One or more	You must have at least one Signal column.
Signal Background	Optional	Any	You can have as many Signal Background columns as you have Signal columns. If you are using Signal Background, you must have a Signal Background for each Signal column.
Signal Precision	Optional	Any	Used only when the scanner software used for your experiment produces an estimate of the precision of the value in the signal column. This information is merged with other information as part of the GeneSpring Cross-gene Error Model. These numbers are the standard deviation of the measured signal around the true expression level (signal) for that sample as expressed by the scanner software. See "Cross-gene Error Models" on page 3-44 for details.
Control Channel	Optional	Any	If you have control channels (i.e., a two-color experiment), you must have the same number of control channel columns as signal columns.
Control Channel Background	Optional	Any	If you are using control channel backgrounds, the number of columns must be the same as the number of Control Channel columns.
Description	Optional	One	A description of the gene, if known. This information is included in the new master table of genes, and is accessible with the Find Gene command and the Gene Inspector. This field applies only to new genomes created through the Column Editor.

Name	Required?	# Allowed	Description
GenBankID	Optional	One	The GenBank identifier for the gene, if known. If the GenBank identifiers for your genes are not used as their systematic or common names, including the GenBank accession number in this field allows you to update information about the gene directly from GenBank. See "Updating Annotations with GeneSpider" on page 6-27 for more information. This field is included in the new master table of genes, and applies only to new genomes created through the column editor.
Common Name	Optional	One	Adds a common name column to a genome if it is being newly created.
Flags	Optional	Any	Specifies the letter or number indicating Present, Absent, and Marginal calls. You can have as many Flag columns as you have Signal columns.
Region	Optional	One	If your experiment uses multiple arrays or sections of arrays that must be normalized separately, this column tells GeneSpring the region of the array and/or which array a particular gene reading came from.

2. Click **Guess the Rest**. GeneSpring attempts to label the remaining columns. If the labels are incorrect, click **Clear Guess** to remove the column labels and select them yourself.
3. Click **Advanced Options** if any of the following are true:
 - The gene identifiers in your experiment files have a prefix or suffix that must be stripped.
 - Your signal and control values are in separate files.
 - You want to apply a default normalization scheme to your experiment files.

From this screen, you can select the appropriate options.

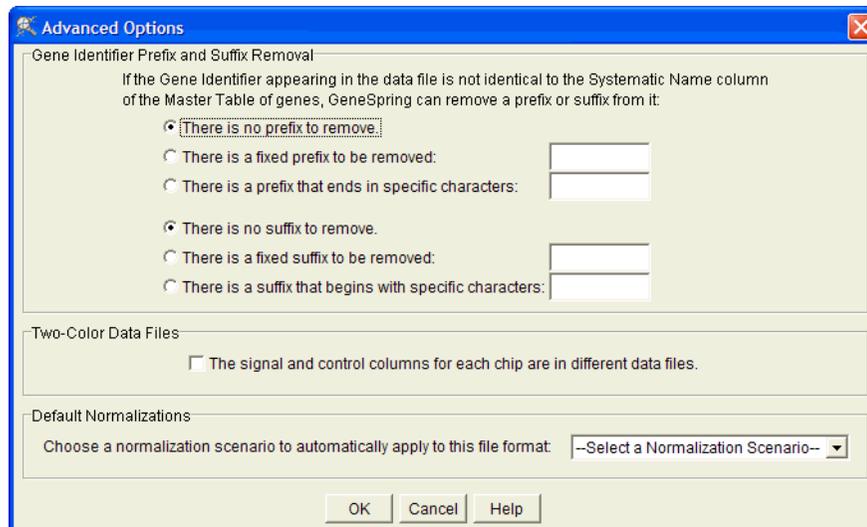


Figure 3-10 Advanced Column Editor Options

- a. **To strip a gene identifier prefix or suffix**—in the Gene Identifier Prefix and Suffix Removal section, select the appropriate radio button and enter the characters to be stripped in the text box next to your choice.
 - b. **To specify that your signal and control values are in separate files**—check the box in the Two-Color Data Files section.
 - c. **To apply a default normalization scenario to your experiment files**—select the appropriate scenario from the pull-down menu in the Default Normalizations section. For more information on the available default normalizations, see “Default Normalizations” on page 3-21.
4. To save this file format setup for future use, click **Remember This Format**. The format is added to the cache of recognized formats so that GeneSpring recognizes it. When prompted, enter a name for the new format.

Default Column Assignments of Known Products

GeneSpring recognizes column titles of various commercially available products and place them as described in the following lists.

Affymetrix

Pivot Table:

- **Column 1**—interpreted as Gene Name
- **Average Difference or Signal**—interpreted as Signal
- **Detection or Abs Call**—interpreted as Flags

Metrixs:

- **Gene Name or Probe Set or Probe Set Name**—interpreted as Gene Name
- **Signal or Average Difference**—interpreted as Signal
- **Detection or Abs Call**—interpreted as Flags

- **P, M, A**—interpreted as Flag Designators
- **Region**—interpreted as Experiment Name

d-Chip

- **Probe Set**—interpreted as Gene Name
- **Column to left of column that ends in “call”**—interpreted as Signal
- **Description**—interpreted as Description
- **Accession**—interpreted as GenBank ID
- **Column to the right of “call” that is to the right of a Signal column**—interpreted as Flags
- **P, M, A**—interpreted as Flag Designators

Agilent

- **ProbeName**—interpreted as Gene Name
- **rBGSubSignal**—interpreted as Signal
- **gBGSubSignal**—interpreted as Control
- **Description or GeneName**—interpreted as Description
- **GenBank**—interpreted as GenBank ID

Amersham

- **GeneID**—interpreted as Gene Name
- **Signal Mean**—interpreted as Signal
- **Background Mean**—interpreted as Signal Background
- **Flag**—interpreted as Flags
- **0=P, 2=A, 3=M**—interpreted as Flag Designators

Axon

GenePix Pro 2 & 3:

Note: The Ratio Formulation entry is used to determine which channel is Signal and which is Control.

- **ID**—interpreted as Gene Name
- **F635 Median or F532 Median**—interpreted as Signal
- **B635 Median or B532 Median**—interpreted as Signal Background
- **F635 Median or F532 Median**—interpreted as Control Channel
- **B635 Median or B532 Median**—interpreted as Control Channel Background
- **Name**—interpreted as Description
- **Flags**—interpreted as Flags

BioDiscovery

Imagene 4:

- **Gene ID**—interpreted as Gene Name

- **Signal Median**—interpreted as Signal
- **Background Median**—interpreted as Signal Background
- **Signal Median**—interpreted as Control Channel
- **Background Median**—interpreted as Control Channel Background
- **Flag**—interpreted as Flags

Incyte

GEMTools 2.4:

- **CloneID**—interpreted as Gene Name
- **P2 BalancedSignal or P2 Balanced**—interpreted as Signal
- **P1Signal or P1**—interpreted as Control Channel
- **Gene Name**—interpreted as Description
- **AccessionNum or Accession**—interpreted as GenBankID

Internet Download:

- **CloneID**—interpreted as Gene Name
- **Varies (format of PS# Cy5, determined in header)**—interpreted as Signal
- **Varies (format of PS# Cy3, determined in header)**—interpreted as Control Channel
- **Gene name**—interpreted as Description
- **PS# Absent/Present (where # is the sample name)**—interpreted as Flags
- **P, A**—interpreted as Flag Designators
- **GEM ID**—interpreted as Custom1
- **Gene ID**—interpreted as Custom2

Packard Biochip ScanArray/QuantArray

Note: Checks file header to determine which channel is Signal and which is Control

- **Name**—interpreted as Gene Name
- **ch1 Intensity or ch2 Intensity**—interpreted as Signal
- **ch1 Background or ch2 Background**—interpreted as Signal Background
- **ch1 Intensity or ch2 Intensity**—interpreted as Control Channel
- **ch1 Background or ch2 Background**—interpreted as Control Channel Background

Clontech Atlas Image 2-Color

- **Gene Code**—interpreted as Gene Name
- **Intensity_2**—interpreted as Signal
- **Background_2**—interpreted as Signal Background
- **Intensity_1**—interpreted as Control Channel
- **Background_1**—interpreted as Control Channel Background
- **Protein/gene**—interpreted as Description
- **Column 11**—interpreted as GenBankID

Clontech Atlas Image 1-Color

- **Gene Code**—interpreted as Gene Name
- **Intensity_2**—interpreted as Signal
- **Background_2**—interpreted as Signal Background
- **Protein/gene**—interpreted as Description
- **Column 11**—interpreted as GenBankID

Creating New Experiments

You can create a new experiment using existing data from either your local system, from a GeNet server, or both. GeneSpring provides a variety of filters to make it easy to select the appropriate samples for your experiment.

The following sections describe the basic process of creating a new experiment, followed by more detailed information on each screen.

To create a new experiment:

1. From the main GeneSpring window, select **Experiments > Create New Experiment**. The Select Samples screen appears. For a detailed description of this screen, see “The Sample Manager” on page 3-23.
2. Select the samples to include in your experiment. To add a sample, select it in the Filter Results List and click **Add**. The sample appears in the Samples for New Experiment list.

To select multiple samples, hold down the Ctrl key while clicking the desired samples. To add all the samples in the list to your experiment, click **Add All**.

To view detailed information on a sample, click **Inspect** to invoke the Sample Inspector. For more information on the Sample Inspector, see “The Sample Inspector” on page 4-13.

3. If you need to edit parameters or normalizations for this experiment, click **Next**. For detailed information on the Edit Parameters screen, see “Experiment Parameters” on page 3-29. For detailed information on normalizations, see , “Normalizing Data”. Once you are finished with parameters and normalizations, click **Finish**. The Choose Experiment Name screen appears.

To accept the default parameters and normalizations, click **Finish**. The Choose Experiment Name screen appears.

4. In the Choose Experiment Name window, do the following:
 - Enter a name for the experiment in the **Name** field. Be sure to choose a descriptive name that you will remember later.
 - To save the experiment in an existing folder, navigate to that folder in the directory browser in the lower left portion of the screen, and leave the **Folder** field blank. To save in a new subfolder, navigate to the desired parent folder and enter a name for the new folder in the **Folder** field.
 - If desired, enter notes containing more descriptive information about the experiment.

When you are done, click **Save**. Your new experiment has been created.

Copying and Pasting Experiments

You can use the copy (Ctrl+C) and paste (Ctrl+V) functions to insert a new experiment or lists from the clipboard into GeneSpring.

Preparing to Paste

You should have normalized data in an Excel file or saved as tab-delimited text. (Figure 3-12). You must have all of the following three parts to your data. Your data *must* be in the following format to correctly paste into GeneSpring:

- Name
- Parameters
- Data

The seven parameters for this experiment

Parameter values for third sample

	A	B	C	D	E
1	Multiple Disease Example				
2	Sick ()*	no	y	y	y
3	Disease ()*	no	hepatitis	hepatitis	syphilis
4	Infectious Disease ()*	n	y	y	y
5	Hepatitis ()*	n	y	y	n
6	Type Hepatitis ()*	n	a	b	n
7	Cancer ()*	n	n	n	n
8	Type Cancer ()*	n	n	n	n
9	Time (minutes)	0	10	20	30
10	YAL001C	0.941667	0.575	0.95	0.925
11	YAL002W	1.738318	0.971963	0.570093	0.635514
12	YAL003W	0.710966	0.68773	0.964863	0.679229

Figure 3-11 Example of parameter arrangements and values

Name

The first line must be the unique name of the experiment.

Parameters

The second line must be the first parameter. You can have an unlimited number of parameters.

- The first column must contain the parameter name.
- Subsequent columns contain values for the parameter in that sample.

Each parameter must have units in parentheses in the same column as the name. For example, the parameter “time” should be immediately followed by (minutes). If your

parameters have no units you must follow the name with an empty set of parentheses, or GeneSpring does not recognize it as a parameter.

By default, GeneSpring assumes that the parametric values to follow are numeric and to be displayed in numeric order. If the parametric values for a parameter are non-numeric, enter an asterisk immediately after the unit-indicating parentheses (empty if no units). There must be a space between the right parenthesis and the asterisk. This tells GeneSpring to expect non-numeric parametric values and treat the data appropriately.

The default setting for interpretation of parameters is as a continuous element. See “Continuous Element” on page 3-31 for details. To have the parameters treated differently, enter the following codes just after the parentheses:

- **S** — means the data is interpreted as a non-continuous element, also known as a discrete element. See “Non-Continuous Element” on page 3-31 for details.
- **C** — data is colored by the different parametric values assigned automatically by GeneSpring. In Figure 3-12 each column would get a different color as time values 0-160. See “Color Code” on page 3-31 for details.
- **R** — data is interpreted as a replicate (not shown). See “Hidden Elements” on page 3-31 for details.

You can enter all parameters with the default (with no code after the parentheses) and change the interpretation later from within GeneSpring. See “Experiment Interpretations” on page 3-39.

For example, for the parameter `tissue type`, a non-continuous non-numeric parameter, the first column might look like this:

```
tissue type() *S.
```

If you have no parameters, enter arbitrary (but meaningful) names so that you can distinguish each sample from those in other columns.

Data

- There can be only one gene per line.
- The name of the gene must be in the first column.
- The following columns are data points for each sample.

Experiment Name	First Parameter Name with units	Parameter Values	Normalized Data
Multiple·Disease·Example			
Sick·()	*no	y	y
Disease·()	*no	hepatitis	hepatitis
Infectious·Disease·()	*n	y	y
Hepatitis·()	*n	y	y
Type·Hepatitis·()	*n	a	b
Cancer·()	*n	n	n
Type·Cancer·()	*n	n	n
YAL001C			0.941666722+0.575000048+0.950000048+0.925000072+1.166666746

Figure 3-12 Example of a correctly formatted tab-delineated file

Common Mistakes in Pasting

- forgetting the title
- not using parentheses
- not having parameters
- using non-normalized data (data can be normalized within GeneSpring)
- having extraneous columns
- forgetting to indicate parameters having non-numeric parametric values with an asterisk (*)
- using more than one type of decimal marker, or the wrong type for your computer's settings.

Pasting an Experiment into GeneSpring

1. If you have not done so already, give your experiment a unique name. If the name is already in use, GeneSpring appends a number to distinguish it from other experiments of the same name.

Select all or part of a properly formatted Excel or tab-delineated file and click **Copy** or press **Ctrl-C**.

Note: Some computers have a limit on the amount of data you can place on the clipboard. If you are consistently crashing at the point, you may need a JVM update.

2. In the main GeneSpring window, select **Edit > Paste > Paste Experiment**. GeneSpring automatically updates the window, regardless of the current display settings. Larger files may take longer to paste, depending on your system.

When the paste is complete, a new Choose Experiment Name box appears with the current name of the experiment already in the Name text box.

When you return to the main window, your new experiment is displayed automatically. From here, you can alter the normalizations with **Experiment > Experiment Normalizations** command or the interpretation with the **Experiment > Experiment Interpretation** command.

Copying an Experiment or a List Out of GeneSpring

When you copy an experiment, only data for the currently selected gene list is copied. To copy data for all the genes in the current experiment, right-click over the “All genes” list and select **Display List** before you begin to copy.

When you paste, the gene list is sorted into the order presented in the Ordered List view.

1. Choose an experiment or a gene list from the navigator.
2. In the main GeneSpring window, select **Edit > Copy > Copy Experiment**. Your data is saved to the clipboard.
3. Paste your experiment or gene list into Microsoft Excel or a text editor such as Microsoft Notepad or Microsoft Word.

Default Normalizations

GeneSpring normalizes your new files based on the technology used to create the original data files. For more information on normalizations, see , “Normalizing Data”.

One-Color Experiments

GeneSpring counts only samples of the same data type which have both the data transform and the normalize to median steps applied. One-Color normalizations automatically display all measurements:

- **Data transformation**—Set measurements less than 0.01 to 0.01
- **Per Chip**—Normalize to 50th percentile.
- **Options**—Use all flags, never apply background correction
- **Per Gene**—Normalize to median, cutoff=10 in raw data (if 3+ samples)

Two-Color Experiments

Two-color experiments are automatically normalized to a signal ratio. Two-color normalizations automatically display all measurements:

- **Per Spot**—Intensity dependent (Lowess) if more than 100 genes per region, divide by control channel if fewer than 100 genes per region. Cutoff = 10 in raw data, 20% of data used for smoothing
- **Per Chip**—Intensity dependent (Lowess) if more than 1000 genes per chip, divide by control channel if fewer than 1000 genes per chip
- **Options**—Use background correction if necessary, anything but absent. Cutoff = 10 in raw data

Pre-Normalized Data

This applies only to samples uploaded to GeNet before version 3.0, created from experiments using the Merge/Split functionality in an earlier version of GeneSpring, or samples imported using the Paste Experiment function.

- **Start with pre-normalized data**
- **Per spot**—Reserve control channel

Replicates

If you have multiple measurements for the same gene in the same sample, GeneSpring takes the average of the measurements. In addition, GeneSpring saves the minimum and maximum values and display them in the Gene Inspector. See “Dealing with Repeated Measurements” on page 5-18 for a mathematical explanation of this process.

Remembered Formats

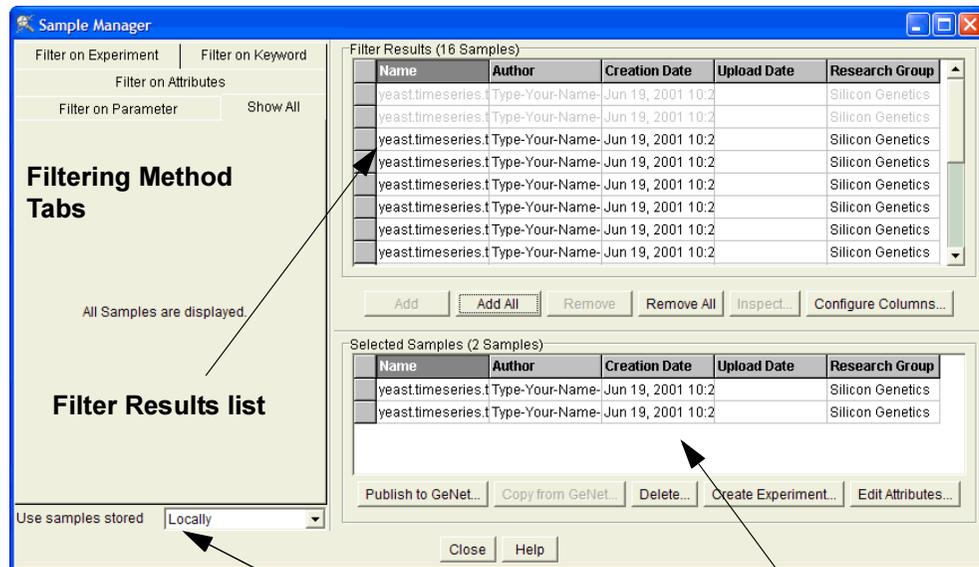
While you cannot edit remembered formats, you can share them. (If you must change a remembered format, you must build a new one.) To share remembered format files, use your favorite browser or file management program to copy the file from:

```
YourLocalDrive:\Program Files\SiliconGenetics\  
GeneSpring\data\Experiment Formats\name.expformat
```

The above path must be typed all on one line. You can then paste the file into a shared drive.

The Sample Manager

The Sample Manager is an important part of the experiment creation process, but it can also be used on its own. From the main GeneSpring window, select **Experiments > Sample Manager...**



Display samples from the local machine, GeNet, or both

Samples to include in the new experiment

Figure 3-13 The Create New Experiment window

From this screen you can select samples to add to your experiment, or choose a filtering method to narrow the samples.

The left side of the screen contains a tab for each filtering method. The available methods are:

- Show All—Display all available samples without applying a filter
- Filter on Experiment—Display samples associated with a particular experiment
- Filter on Attributes—Display samples based on selected attributes
- Filter on Keyword—Display samples containing a keyword
- Filter on Parameter—Display samples based on the parameter values of experiments containing them

Click on the appropriate tab to view the options for that filtering method. For detailed information on these filters, see “Filtering Methods” on page 3-25.

The right portion of the screen contains two sample lists. The upper list contains all of the samples resulting from the current filtering method. The lower list contains the samples you have selected.

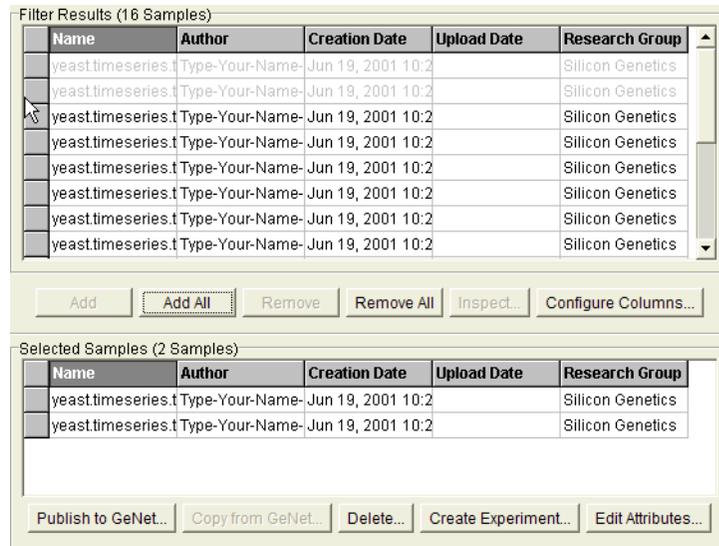


Figure 3-14 Sample Lists

Below the Filter Results list are six buttons:

- **Add**—Add a selected sample in the Filter Results list to the Selected Samples list.
- **Add All**—Add all samples in the Filter Results list to the Selected Samples list.
- **Remove**—Remove a selected sample from the Selected Samples list.
- **Remove All**—Remove all samples from the Selected Samples list.
- **Inspect**—View the selected sample in the Sample Inspector. For more information on the Sample Inspector, see “The Sample Inspector” on page 4-13.
- **Configure Columns**—Select which columns to display in the sample lists. The available choices are:
 - Sample Name
 - Identifier
 - Authors
 - Creation Date
 - Upload Date
 - Research Group
 - Organization
 - Application
 - Sample Attributes
 - Experiment Parameters

Below the Selected Samples list are five buttons:

- **Publish to GeNet...**—Publish all samples in the Selected Samples list to GeNet.
- **Copy from GeNet...**—Make a local copy of all GeNet samples in the Selected Samples list.
- **Delete...**—Delete the samples in the Selected Samples list.

- **Create Experiment...**—Create a new experiment from the samples in the Selected Samples list.
- **Edit Attributes...**—Edit the attributes of the highlighted samples in the Selected Samples list.

Filtering Methods

Filter on Experiment

This method allows you to filter based on samples associated with a selected experiment.

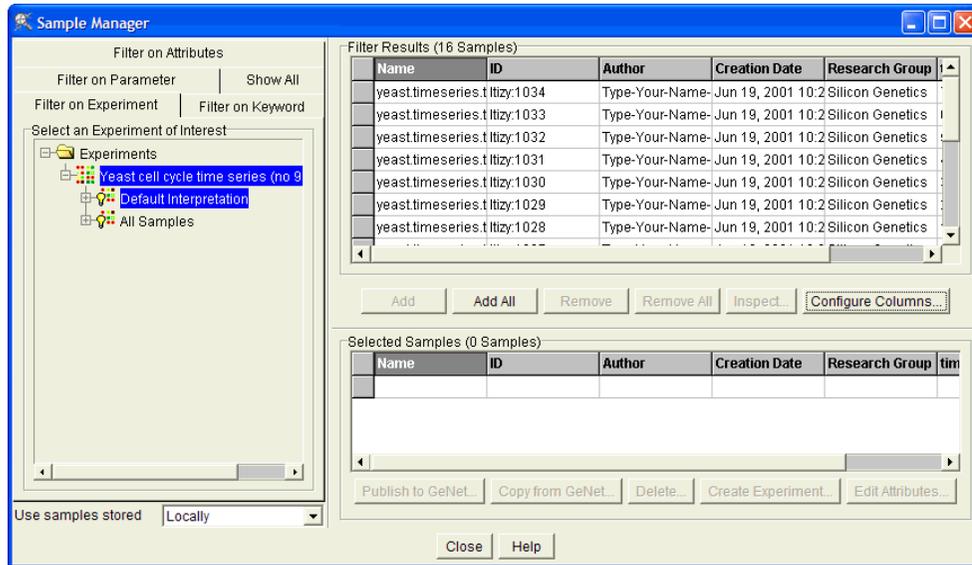


Figure 3-15 The Filter on Experiment tab

To filter by experiment, use the GeneSpring navigator to locate the desired experiment and select it in the list. All samples associated with that experiment appear in the Filter Results list.

Filter on Parameter

This method filters samples based on a parameter and value range. Parameters are associated with experiments, not individual samples. The samples listed are those contained in any experiment matching the selected parameter and value(s).

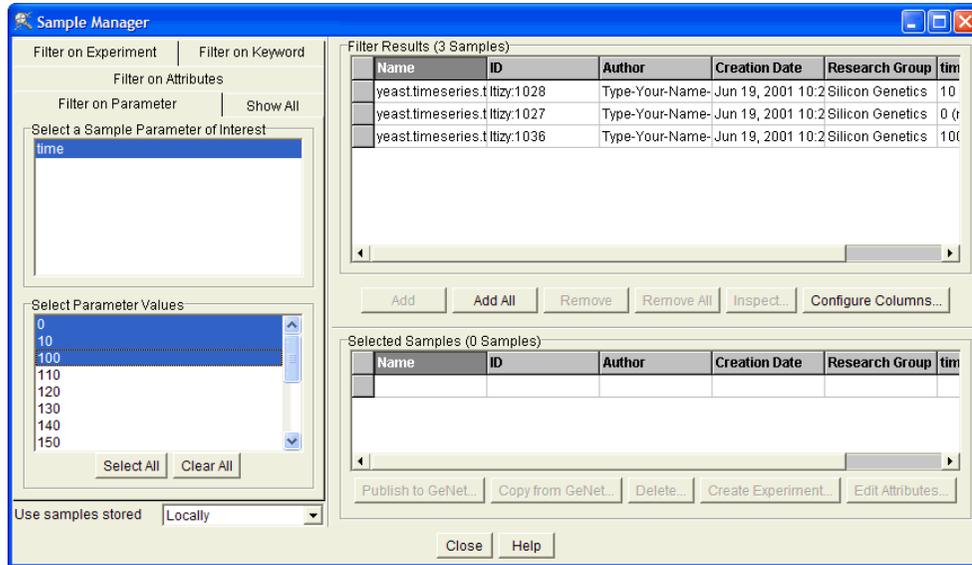


Figure 3-16 The Filter on Parameter tab

To select a parameter, click its name in the Select a Sample Parameter of Interest list.

To select parameter values, click the desired values in the Select Parameter Values list. To select all values in the list, click **Select All**. To clear your selections click **Clear All**.

The Filter Results list is updated dynamically as you make your selections.

Filter on Attributes

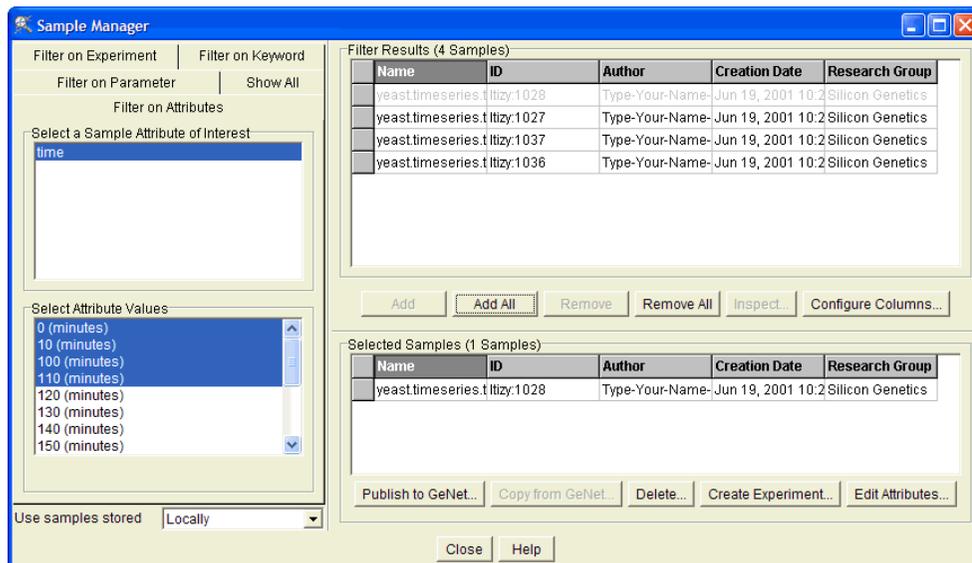


Figure 3-17 The Filter on Attributes tab

This method allows you to filter samples based on their attributes. Attributes are very similar to parameters, but are associated with individual samples rather than entire experiments. Attributes can also be paragraphs long, since they do not appear as labels on a

graph. They contain sample-specific information that would not be used as a basis for analysis. For example, the following might be attributes of a sample:

- Patient Biography
- Lab Technician
- Date
- Ambient Temperature

Attributes may also contain the same information as parameters, and can be imported as parameters when creating an experiment.

To select an attribute, click its name in the Select a Sample Attribute of Interest list.

To select attribute values, click the desired values in the Select Attribute Values list. To select all values in the list, click **Select All**. To clear your selections click **Clear All**.

The **Filter Results** list is updated dynamically as you make your selections.

Filter on Keyword

This method allows you to filter based on whether a particular keyword appears in any of the specified fields. This is useful in cases where a given string (such as “cancer”) is sometimes the parameter, sometimes the parameter value, and sometimes is part of the experiment name.

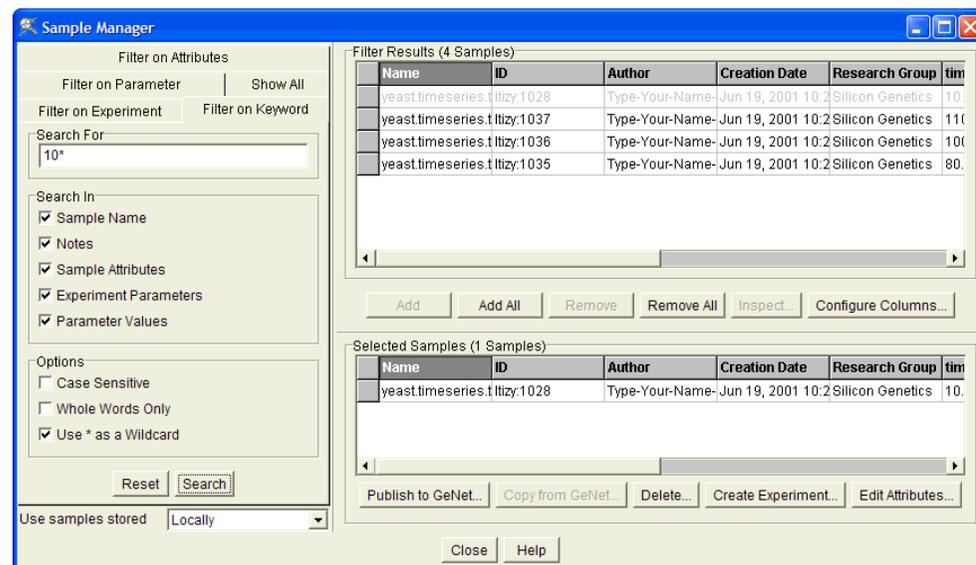


Figure 3-18 The Filter by Keyword tab

To filter by keyword:

1. Enter the desired keyword in the **Search For** box. This keyword can be a word or a number. It can also contain an asterisk (*) as a wildcard character.
2. Select the fields to search. To select a file, check the box next to the desired field in the **Search In** panel. To unselect it, click in the box to remove the check. You can select as many or as few fields as you like. The available choices are:

- Sample Name
 - Notes
 - Sample Attributes
 - Experiment Parameters
 - Parameter Values
3. In the **Options** panel, select any additional search features. You can choose from the following:
 - **Case Sensitive**—search only for words using the specific capitalization you entered.
 - **Whole Words Only**—search only for whole words matching your keyword. For example, if you search for the string “statistic”, the search results will contain samples that contain the word “statistic” in the specified fields, but ignore those containing “statistic” as part of a larger word, such as “statistician”.
 - **Use * as a Wildcard**—search for samples containing the specified keyword plus any other characters. For example, you might want to look for samples that are named using a prefix with sequential numbers appended to the end. To do this, you would enter the prefix followed by an asterisk, e.g., “ex*”.
 4. Click **Search**. Any samples matching your search term appear in the Filter Results list.

Experiment Parameters

Parameters are the variables you use to describe your experiment.

Experiment parameters

These are variables that can incorporate many sample values. Generally speaking, when the term parameter is used, it means an experimental parameter. As an example, experiment parameters could be:

- Drug Concentration
- Strain of Yeast
- Infection
- Replicate Number

Parameter Values

The values of the parameters in the previous list could be:

- Drug Concentration in ppm, 0, 10, 20, 30, 40
- Strain of Yeast, A or B
- Infection, Healthy or Infected
- Replicate Number, 1 or 2

Parameters Displayed in the Navigator

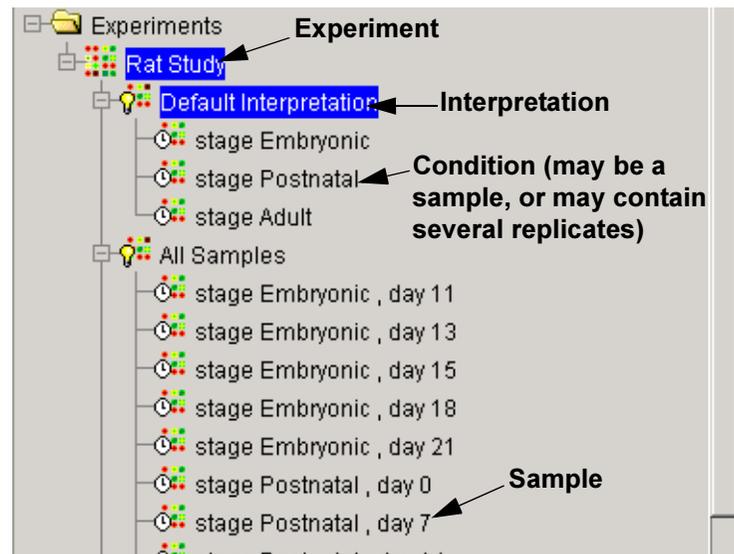


Figure 3-19 Data objects in the navigator

- **Sample**—The data generated from a biological object placed onto an array or set of arrays. Sample data is visible in the GeneSpring navigator, under the All Samples icon.

- **Condition**—A unique combination of parameters as applied to your sample. Each condition may be a single sample or a group of replicate samples combined based upon the parameter values defined for each sample. The easiest way to think of this is as the parameters under which the sample(s) was observed. If you have no replicates, condition and sample can be considered synonymous. In Figure 3-19 the conditions are Embryonic, Postnatal and Adult.
- **Interpretation**—A description of how GeneSpring displays the data for you to view. It would include a definition of applicable parameters and how to treat the normalized numbers. This is the way a set of conditions is grouped. In Figure 3-19 the interpretation is the Default Interpretation.
- **Experiment**—a set of samples, generally designed to answer specific types of questions. The data are usually (but not always) manipulated in a normalized form. In Figure 3-19, the experiment is the Rat Study.

A Note on Multiple Parameters

The more experiment parameters you have, the more options you have for visually querying your data. If you have samples of tissues with different diseases such as breast cancer, kidney cancer, liver cancer, brain cancer, hepatitis A, hepatitis B, osteoporosis, arthritis, syphilis, and no disease, you might want to use several parameters for this experiment. Using multiple parameters (even if they all refer to the same information) allows you to group the data in many different ways which may give you different insights into your data set.

As another example, these could be given parameters of Cancer, Pathogen and Genetic Disorder. Parameter values can also be assigned in different ways:

- Cancer: Malignant, Benign or Normal
- Cancer: Cancerous or Normal
- Cancer: Tumor, Metastatic or Normal

If all of these parameter values are used they should be assigned to multiple parameters, such as Cancer type, Cancer Presence and Cancer Stage. Additional parameters could be Age, Gender and Treatment.

Parameter Display Options

GeneSpring offers four ways of visually displaying a parameter: a continuous element, a non-continuous element, a replicate (or hidden) element, or a color code. When you create a new experiment, your chosen display option becomes the default for that parameter. If you simply paste in a new experiment, all parameters are assigned in the Paste Experiment format.

Regardless of how a parameter was created in GeneSpring, you can use the **Experiment > Change Experiment Interpretation** command to change how it is displayed. For more details on this, see “Experiment Interpretations” on page 3-39.

Hidden Elements

Parameters defined as replicates are averaged together and appear as a single parameter. A parameter defined as a replicate is graphically a hidden variable. Defining a parameter as a replicate is the easiest way to deal with repeated samples inside GeneSpring.

The equation used for averaging repeated samples is exactly the same one used to average repeated measurements in a raw data file. See “Dealing with Repeated Measurements” on page 5-18 for more information. The only difference is that averaging of repeated parameters is done after the raw data has been normalized.

Continuous Element

In a continuous variable, each parameter value exists in series on a continuum with the other values in that parameter, rather than as discrete points. Each value is related to the values on either side of it and adjacent data points are connected together by lines. Typically, continuous variables are numeric. This requires that the values be in a particular order. GeneSpring automatically orders numerical parameters from highest to lowest and non-numerical parameters in alphabetical order.

When graphing by a continuous parameter each value is placed on the X-axis in order from left to right. You can change this default order. See “Set Value Order” on page 3-34 for more details.

Non-Continuous Element

In a non-continuous (or set) variable, each parameter value exists independent of the others, as a discrete point. When a non-continuous element is graphed, each parameter value is placed on the horizontal axis, in order from left to right. GeneSpring automatically orders numerical parameters from highest to lowest. Non-numerical parameters are in alphabetical order. See “Set Value Order” on page 3-34 if you need non-numerical parameter values to be graphed in a particular non-alphabetical order.

When displaying data from a non-continuous parameter, data points are graphed in histograms, as discrete points. A gene deletion is a simple example of a non-continuous element, but it is by no means the only possible non-continuous parameter. A non-continuous parameter is occasionally referred to as a set when there are other parameter display options employed (especially when a continuous parameter is used) because the non-continuous parameter separates the data into a series of discrete graphs viewed next to each other on the same screen. When a continuous parameter is used in conjunction with a non-continuous parameter each discrete graph contains all of the values of the continuous parameter, making each of the separate graphs look like a set of parameter values.

Color Code

A color code is used for experimental parameters whose parameter values exist independently of one another but are not unrelated. When the genome browser is colored by parameter, GeneSpring orders the parameter values from top to bottom in the colorbar. See “Color by Parameter” on page 4-33 for details. Values are listed in alphabetic or numerical order.

Each color represents a category (or set of categories). When coloring the browser display by parameter, each value defined as a condition is assigned a color and every data point described by that parameter is drawn in that parameter's color. This can be referred to as *Color by Parameter*. This display option shows the same gene multiple times. The number of times a single gene is drawn is equal to the number of values defined as conditions.

When the browser display is colored using a color option other than Color by Parameter, it is impossible to visually distinguish which value a particular gene line or gene point represents, although separate gene lines for each value defined as a condition are still drawn. See "Set Value Order" on page 3-34 for details on how to change that order.

Individual patients, or strain types, are variables commonly defined as color codes (conditions) because, although they are different values, it is interesting to see them visually compared to one another. It is likely the expression patterns of individual patients with the same disease will react in a similar way under similar conditions. Often it is when the expression patterns are not similar that the results are interesting. This is where graphs of parameter-values defined as color-coded conditions are useful as they allow you to easily compare varying conditions of the same gene.

Changing Experiment Parameters

Use the Experiment Parameters window to assign parameter names and units (e.g., time and minutes) to your data.

You can also use this window to add and delete parameters and rearrange the order of non-numeric parameter values on the horizontal axis. If you set up your file names as described below, your parameter assigning process is automated.

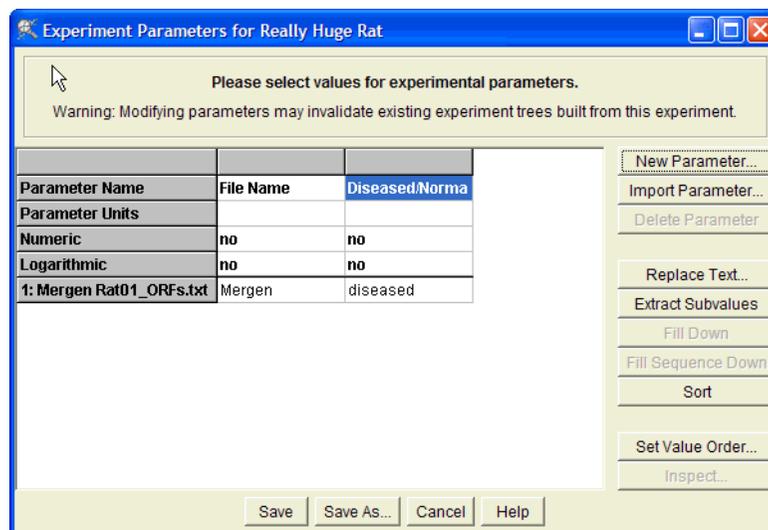


Figure 3-20 The Experiment Parameters window

Import Parameter

You can import a parameter from another experiment or from a list of sample attributes defined in any of the samples in the current experiment. To import a parameter:

1. Click **Import Parameter**. The Import Parameters window appears.

2. Select the parameter or parameters to import.

To select a sample attribute to import as a parameter, click its name in the list at the top of the screen. To select all attributes in this list, click **Select All**. To clear your selections, click **Clear All**.

To import parameters from another experiment, find the desired experiment in the navigator and select it. The parameters associated with that experiment appear in the Parameters from Selected Experiment list. Select the desired parameters from the list.

3. When you are done, click **OK**. You are returned to the Experiment Parameters screen. A new column appears for each parameter you imported.

New Parameter

To create a new parameter:

1. Click **New Parameter**. A dialog appears.
2. To add a standard parameter, select it from the pull-down menu and click **OK**. To add a custom parameter, select the Custom radio button and click **OK**.

A new column appears in the Experiment Parameters window. If you want to accept the default values for a standard parameter, simply click **Save**.

3. Fill in the Parameter Name and Parameter Units in the new column (the latter only if applicable).
4. In the Numeric and Logarithmic rows, select **Yes** or **No** from the pull-down menus. (Click in a cell in either row to make the pull-down menu appear.) You can also paste data in the **Sample** cells.
5. Click **Save** to change the parameters in your current experiment or **Save As** to save this parameter set-up as a new experiment.

You can paste in columns of information by clicking the cells of the Sample section. For example, if you had an Excel spreadsheet of data and wanted to copy and paste a column from it, you could copy a large section of column and paste it into the new column. You can also copy information out. You can only add columns (parameters and parameter values), you cannot add rows (samples) into this table.

Delete Parameter

To delete a parameter, click the gray bar above the column you want to delete and click **Delete Parameter**.

Replace

To replace many entries at once, select the entries to change and click **Replace**. Enter the appropriate text in the dialog box that appears. To replace all instances of an entry, choose **Replace** and then uncheck the **Replace in selected cells only** box before clicking **OK**.

Extract Sub-Values

This feature automates parameter assignment. To use it, create file names based on your parameter values (e.g., Rlr001a.txt, where “Rlr0” is an experiment and “01” is your sample number and “a” is the region designator).

When you use the Extract Sub-values feature, file names are broken down into sub-values. GeneSpring is programmed to first look for alternating constant fields and variable fields and to make parameters out of the variable fields. Next it divides the variable fields into groups consisting of uninterrupted stretches of either numbers, letters, or non-alphanumeric characters and makes parameters out of each of these groups.

Fill Down

To replace entries using the top selected cell, click on the cell you want to use as the replacement and then, holding down the Shift key, click on the cells underneath whose values you would like replaced with the original cell. Then click **Fill Down**.

Fill Sequence Down

This allows you to fill down as described above, but automatically continue a simple numeric or alphabetic sequence.

Set Value Order

To change the order of your parameters as they are displayed along the X-axis in the main GeneSpring window, select an entire column or part of a column and click **Set Value Order**.

For example, to show the numeric, continuous parameter “Kryptonite Concentration” in reverse order (40, 30, 20, 10, 0) of the normal arrangement (0, 10, 20, 30, 40) you first must change the setting to a non-numeric parameter and select the column by clicking on the gray bar at the very top. You cannot change the order of a parameter defined as numeric.

To select part of a column, highlight it in the normal fashion. Click in the topmost cell you want to select while holding down the **Shift** key. GeneSpring selects down the column for you. Click **Set Value Order**.

To use the Sort Ascending or Sort Descending buttons, select all the values to be ordered. The main GeneSpring window sorts your parameters according to the new system.

You can also sort manually by selecting one parameter value and use the move up/move down buttons to arrange the order to your liking.

Sample Attributes

Attributes are values associated with a sample, such as time, drug concentration, etc. You may have many attributes applying to a single sample. These attributes are selected and associated with a sample in the Sample Manager. For more information, see “Attributes and Parameters” on page 4-14.

You can add any number of additional attributes you like using the Sample Manager or the Standard Attribute Editor. Additional sets of attributes are available for download from the Silicon Genetics website at <http://www.silicongenetics.com/cgi/SiG.cgi/support/stdatt.smf>. These include sets of attributes tailored for various applications, including MIAME standards.

Changing Experiment Attributes

You can add, edit, or delete sample attributes through the Edit Attributes window.

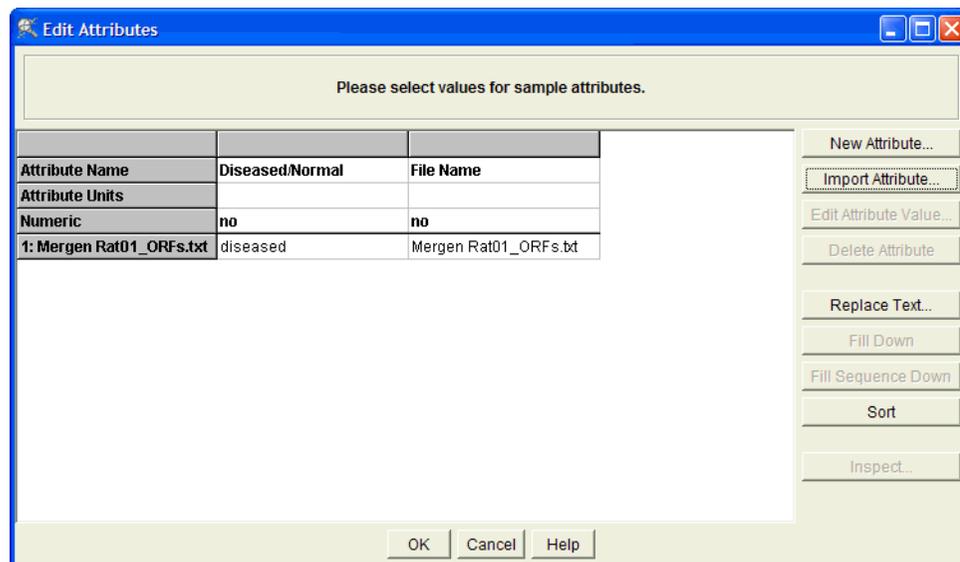


Figure 3-21 The Experiment Attributes window

Import Attribute

You can import a attribute from another experiment or from a list of sample attributes available for any of the samples in the current experiment. You can also convert a parameter into a sample attribute.

To import a attribute:

1. Click **Import Attribute**. The Import Attributes window appears.
2. Select the attribute or attributes to import.

To select a sample attribute to import as a attribute, click its name in the list at the top of the screen. To select all attributes in this list, click **Select All**. To clear your selections, click **Clear All**.

To import attributes from another experiment, find the desired experiment in the navigator and select it. The attributes associated with that experiment appear in the Attributes from Selected Experiment list. Select the desired attributes from the list.

3. When you are done, click **OK**. You are returned to the Experiment Attributes screen. A new column appears for each attribute you imported.

New Attribute

To create a new attribute:

1. Click **New Attribute**. A dialog appears.
2. To add a standard attribute, select it from the pull-down menu and click **OK**. To add a custom attribute, select the Custom radio button and click **OK**.

A new column appears in the Experiment Attributes window. If you want to accept the default values for a standard attribute, simply click **Save**.

3. Fill in the Attribute Name and Attribute Units in the new column (the latter only if applicable).
4. In the Numeric and Logarithmic rows, select **Yes** or **No** from the pull-down menus. (Click in a cell in either row to make the pull-down menu appear.) You can also paste data in the **Sample** cells.
5. Click **Save** to change the attributes in your current experiment or **Save As** to save this attribute set-up as a new experiment.

You can paste in columns of information by clicking the cells of the Sample section. For example, if you had an Excel spreadsheet of data and wanted to copy and paste a column from it, you could copy a large section of column and paste it into the new column. You can also copy information out. You can only add columns (attributes and attribute values), you cannot add rows (samples) into this table.

Delete Attribute

To delete a attribute, click the gray bar above the column you want to delete and click **Delete Attribute**.

Replace Text

To replace many entries at once, select the entries to change and click **Replace Text**. Enter the appropriate text in the dialog box that appears. To replace all instances of an entry, choose **Replace Text** and then uncheck the **Replace in selected cells only** box before clicking **OK**.

Fill Down

To replace entries using the top selected cell, click on the cell you want to use as the replacement and then, holding down the Shift key, click on the cells underneath whose values you would like replaced with the original cell. Then click **Fill Down**.

Fill Sequence Down

This allows you to fill down as described above, but automatically continue a simple numeric or alphabetic sequence.

Editing Standard Attributes

GeneSpring comes with a set of “standard attributes”. These attributes are available for use in any experiment in GeneSpring. You can add as many additional attributes as you like, or edit existing attributes, using the standard attribute editor.

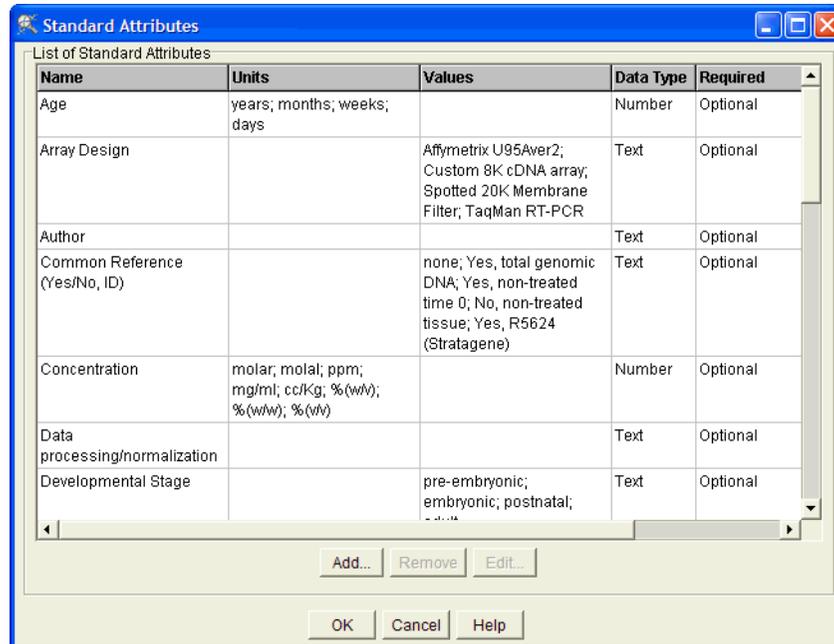


Figure 3-22 The Standard Attributes Editor

To add a standard attribute:

1. Select **Edit > Standard Attributes**. The Standard Attributes window appears.
2. Click **Add**. The Add Attribute window appears.

The screenshot shows the 'Add Attribute' dialog box with the following fields and tables:

- Attribute Name:** Planet of Origin
- Suggested Attribute Units (optional):** An empty table with a 'Remove Row' button.
- Suggested Attribute Values (optional):** A table with the following values: Earth, Mars, Venus, Mercury, Saturn. It has a 'Remove Row' button.
- Data Type:** Text
- This Attribute Is:** Recommended

Buttons at the bottom: OK, Cancel, Help.

Figure 3-23 The Add Attribute window

3. Enter a name for the new attribute.
4. To add a suggested unit of measurement for the attribute, such as “minutes” or “ppm”, click in a row of the Suggested Units table and enter a unit type. You can add as many suggested units as you like. These units will be selectable by users when they assign this attribute to a sample.
5. To add a suggested value for the attribute, such as “Control Group” or “Martian”, click in a row of the Suggested Values table and enter a value. You can add as many suggested values as you like. These values will be selectable by users when they assign the attribute to a sample
6. Select the data type (Text or Numeric) from the Data Type pull-down menu.
7. Specify whether the attribute is Required, Recommended, or Optional from the This Attribute Is pull-down menu.
8. Click **OK**. You are returned to the Standard Attributes window.

To edit a standard attribute, double-click it in the Standard Attributes window, or select it and click **Edit**. The Edit Attribute window appears. This window looks exactly like the Add Attribute window. Make any desired changes and click **OK**.

To delete a standard attribute, select it in the Standard Attributes window and click **Remove**.

Experiment Interpretations

The Experiment Interpretation window allows you to determine how an experiment is to be displayed. You can change the upper and lower bounds of the vertical axis of your graph, the mode used to represent your data, whether to turn on the cross-gene error model, how you want to view each parameter, and which flagged measurements to be displayed.

Changing an experiment interpretation is useful not only for customizing initial display settings, but also because statistical analysis techniques in GeneSpring are carried out based on how your data is characterized in the interpretation. Because of this, it can be valuable to set up more than one experiment interpretation, then perform analyses on each one to compare the results of statistical testing on data that has been grouped and characterized in different ways.

When you load your experiment, GeneSpring automatically creates a Default Interpretation and an All Samples interpretation. The Default Interpretation is the first item listed under the experiment in the navigator. It may be most convenient to set up your most frequently used interpretation as your Default Interpretation. You can rename the Default Interpretation, but you cannot delete it. The All Samples interpretation makes all parameters non-continuous, so that each parameter is viewed and analyzed individually. The All Samples interpretation cannot be changed, renamed or deleted.

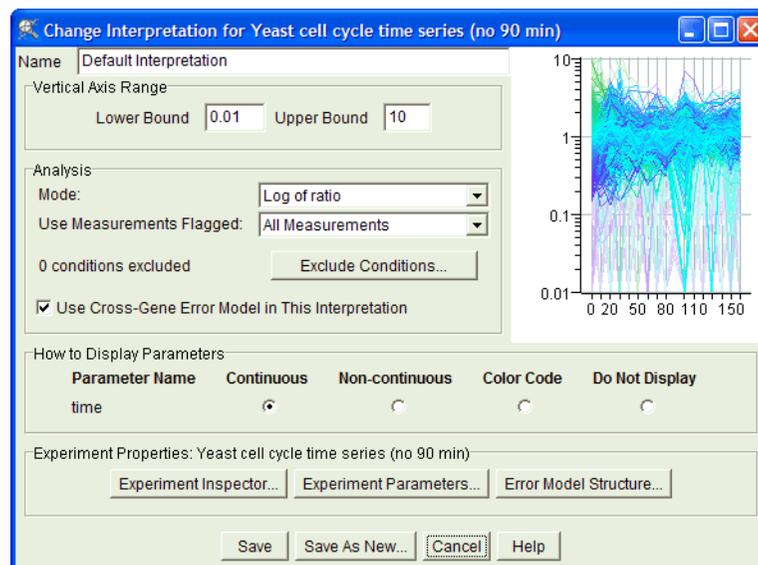


Figure 3-24 The Change Experiment Interpretation window

Changing Experiment Interpretation

1. Select **Experiments > Experiment Interpretation**. The Experiment Interpretation window appears. (You can also right-click the genome browser in graph view and select **Options > Experiment Interpretation**.)
2. From the **Mode** pull-down menu, choose a data display mode for the vertical axis. You have the following choices:
 - **Ratio (signal/control)**

- **Log of ratio**
- **Fold Change**

The mode you choose is used in such statistical procedures as Statistical Group Comparison, k-means Clustering, Self-organizing Maps, and Principal Components Analysis. See below for details on these modes. Choose the lower and upper bounds of the vertical axis in the fields provided.

3. Depending on your instrumentation, you may have flags indicating the degree to which your data is reliable. If you have flags, choose from the **Use Measurements Flagged** pull-down menu to limit data based on these flags.
4. (optional) To use the cross-gene error model, check the **Use Cross-Gene Error Model** box. Cross-gene error models, if used, are assigned an equal number of degrees of freedom as the direct variability estimates for that gene.
5. Choose a mode for each parameter: **Continuous Element**, **Non-continuous**, **Color Code**, or **Do Not Display**. Note that if you choose Color Code, you must also select **Colorbar > Color by Parameter**. See below for details on these modes.
6. To make any additional desired changes in your experiment before you continue, click any of the buttons in the Experiment Properties panel. Your options are:
 - Experiment Inspector—Opens the Experiment Inspector window. For a detailed description of this window, see “The Experiment Inspector” on page 4-16.
 - Experiment Parameters—Opens the Experiment Parameters window. For a detailed description of this window, see “Changing Experiment Parameters” on page 3-32.
 - Error Model Structure—Opens the Cross-gene Error Model window. For a detailed description of this window, see “Cross-gene Error Models” on page 3-44.
7. When you are done, name your interpretation and click **Save** to update your current interpretation or **Save As** to create a new interpretation.

Find saved interpretations by clicking on the relevant experiment in the Experiments folder of the navigator. You can delete an interpretation you have created by right-clicking over it in the navigator and selecting **Delete** from the pop-up menu.

Vertical Axis Modes

The default display is Ratio, where normalized intensity values are graphed on the vertical axis. In this mode, values range from negative infinity to infinity.

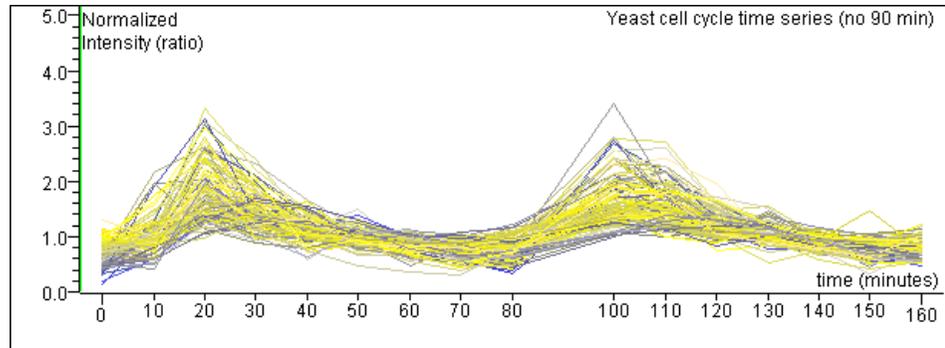


Figure 3-25 The gene list “like *CLN1*” graphed using the [signal/control] formula. The Y-axis is graphed from 0 to 5.

The Ratio is determined by dividing the signal (raw data) by the control strength. (In a one-color experiment the control strength refers to the denominator used to normalize the raw data. In a two-color experiment control strength refers to the control channel.) When data is reported as the signal divided by the control, it is assumed that all expression values are positive. The number 1 is considered normal expression; any expression value above one is overexpressed, and all underexpressed data is less than one, but greater than zero.

This means that all underexpressed data appears flattened because it has to graphically fit between zero and 1, whereas overexpressed data takes up a much larger percentage of the graph (from 1 to positive infinity). Raw signal values that are negative (which is commonly the case in Affymetrix data) produce normalized values that are negative. (To deal with these negative values, see “The Affine Background Correction” on page 5-13.)

Log of Ratio

The Log of Ratio mode graphs normalized values (i.e., the ratio of the signal to the control, not their logs), but spaces them logarithmically. The normal expression is 1. The Log of Ratio interpretation solves the problem mentioned above under “Ratio”, where all underexpressed data appears flattened because it has to graphically fit between zero and 1. In this mode underexpressed genes take up as much space visually as overexpressed genes. Logarithms of the expression ratios are used as the basis for statistical analysis.

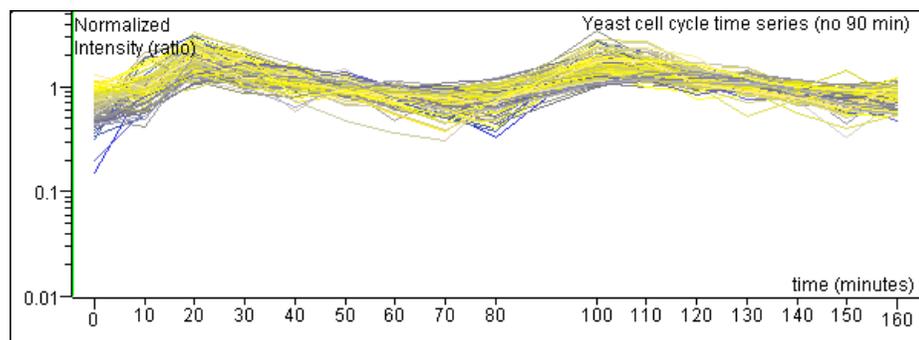


Figure 3-26 The gene list “like *CLN1*” graphed using the log ratio formula

Note that in Log of Ratio interpretation, the lower limit of the vertical axis is 0.01. Any expression values below 0.01 are plotted as 0.01. Note also that when you export your

data, GeneSpring reinterprets the data as the ratio. Measurements below .01 are exported as .01

Fold Change

Fold Change mode creates a more balanced visual representation between over- and underexpressed genes than Ratio mode and emphasizes the increase and decrease of expression levels. For example, x1 would refer to normal expression, x2 to an expression level twice normal, and /2 to an expression level half normal. When using the upper or lower bound fields to change the vertical axis range enter either the ratio values in integers, or the fold change value (i.e., x4 or /4). Any integers you enter are converted.

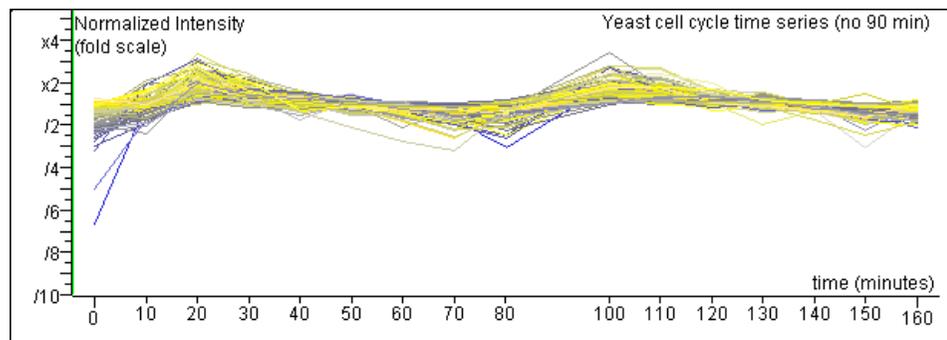


Figure 3-27 New Fold Change Image

Note that in Fold Change interpretation, the lowest measured value is 0.01. Any values below 0.01 are calculated as 0.01. The minimum display value is /100. Note also that when you export your data, GeneSpring reinterprets the data as the ratio. Measurements below .01 are exported as .01.

Ratio Numbers	Display
-5	/110
0	/110
.01	/100 (this is the lower cutoff)
.25	/4
.33	/3
.5	/2
1.5	x1.5
3	x3
5	x5

Note: In Fold Change mode, the values on the vertical axis are not the same as those used for subsequent analyses or those that are exported using the Copy Annotated Gene List function. In fold change mode, the values are stored as 1-N for values greater than 1 and 1-1/N for values less than 1 (where N is the normalized signal). These values can be thought of as representing the distance away from normality (i.e., the

point that represents neither over-expression nor under-expression). These values may not be particularly useful for most users.

Parameter Display Modes

Continuous Element

Applicable only to Graph view, the Continuous Element mode shows parameter values existing on a continuum, where each point is connected with a line. GeneSpring automatically orders numerical parameters from highest to lowest and non-numerical parameters in alphabetical order. See “Parameter Display Options” on page 3-30 for details.

Non-Continuous

Applicable only to Graph view, Non-continuous mode shows parameter values existing independently of one another, where each value is represented as a discrete point. GeneSpring automatically orders numerical parameters from highest to lowest and non-numerical parameters in alphabetical order. See “Parameter Display Options” on page 3-30 for details.

Do Not Display (Replicate)

Select this mode when the parameter does not differentiate between samples. For example, if you have multiple patient samples with different cancer types, you could select Do Not Display for the parameter Age and Continuous for Cancer Type. This would group all the samples by the parameter values for Cancer Type and ignore Age when grouping samples into conditions:

Note that when the same gene occurs twice in the course of an experimental set, it is called a “repeat” and the measurements are averaged together. This cannot be changed.

Color Code

The Color Code mode colors genes by parameter. the number of times a single gene is drawn is equal to the number of parameter-values defined as conditions allowing you to easily compare varying conditions of the same gene. By default, parameter values are listed in alphabetic or numerical order. See “Parameter Display Options” on page 3-30 for details.

Cross-gene Error Models

Using the Cross-gene Error Model

The ability to estimate measurement and sample-to-sample variation in microarray-based experiments is often compromised by the fact that the cost (in both time and materials) of performing large numbers of replicate samples is quite high. If the cross-gene error model is turned on, GeneSpring accounts for error instead by assuming that the amount of variability is a function of the control strength within all the measurements for a single experimental condition. The advantage of making this assumption is that the number of measurements used to estimate the global error is equal to the total number of genes on any given chip.

In addition, measurement precision information supplied by the scanner software or independently by the user can be loaded into GeneSpring via the “Signal Precision” column type in the column editor. The value given in this column is interpreted as the standard deviation of the raw measured value.

The sample-to-sample variability includes the effect of both types of variation, and the statistical separation of these effects is called *variance components analysis*. The GeneSpring Cross-gene Error Model performs this variance components analysis, and uses the estimates of these two components of variation to accurately estimate standard errors and compare mean expression levels between experimental conditions.

Separate estimates of two different kinds of random variation are used to estimate the variability in gene expression measurements:

- **Measurement variation**—This comprises the lowest level of variation, corresponding to the variation of the measurement of a gene on a single chip around the true value that would be achieved by a perfect measurement of the expression level of the gene for that sample.
- **Sample-to-sample variation**—This is the variation between samples in the same condition. This represents biological or sampling variability, such as variability between multiple subjects in a condition, between multiple physical samples for an experimental subject or patient, or between multiple hybridizations of a physical sample. GeneSpring can represent any one of these kinds of variability, depending on the types of replicate samples you have specified in your interpretation and in the error model dialog. GeneSpring assumes all replicate samples in the same condition correspond to one kind of variability.

When you turn the Cross-gene Error Model on the Error Model is used as the basis for:

- standard deviation, representing the variability of individual population members.
- standard error, representing the precision of the mean of the gene expression measurements in the condition with respect to the true condition mean.
- error bars corresponding to standard deviation or standard error in the Graph view and Gene Inspector.
- t-test p-value, representing the statistical test of differential expression for a specific condition.

- color by significance, coloring according to the t-value from the t-test of differential expression.
- finding differentially expressed genes using the Statistical Group Comparison, if the error model option is chosen

To Turn On the Cross-Gene Error Model

1. Select **Experiments > Cross-Gene Error Model**. The Cross-Gene Error Model window appears.

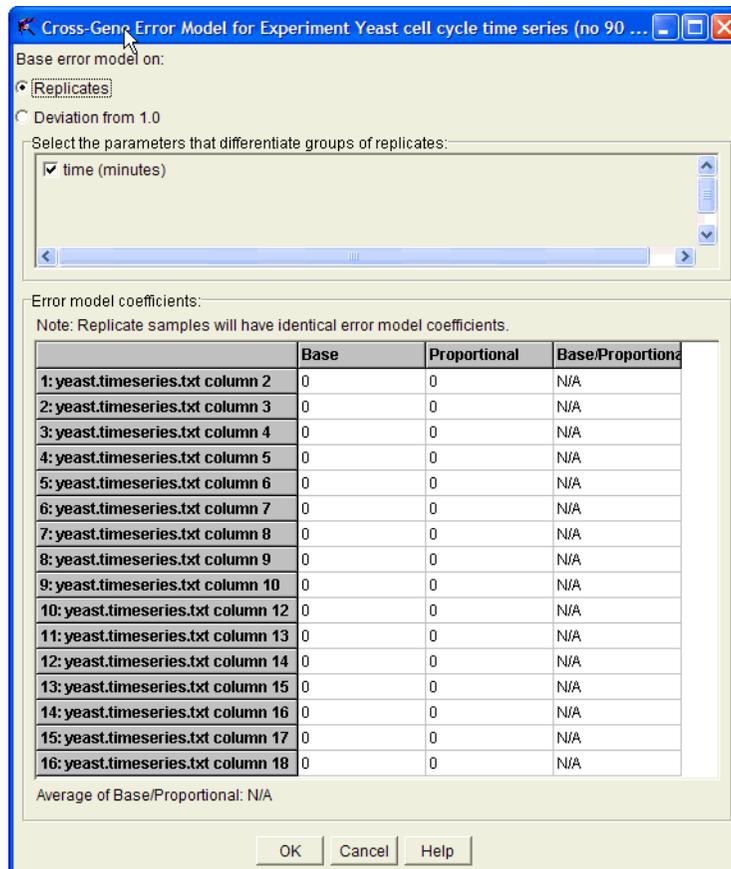


Figure 3-28 The Cross-Gene Error Model window

2. **If you have replicates for each condition**, select the Replicates radio button and select parameters to treat as replicates. Click **OK**.

If you do not have replicates for each condition, select the Deviation from 1.0 radio button and click **OK**.

Note: Double-click on a row to view that sample in the Sample Inspector.

3. Select **Experiments > Experiment Interpretation**. The Change Interpretation window appears.
4. Click the box marked **Use Cross-Gene Error Model**.

- Click **Save** to save as part of your current interpretation or **Save As** to create a new interpretation.

Technical Details

The two-component model for estimating variation from control strength is known as the Rocke-Lorenzato model. The two components are an absolute error component that dominates at low measurement levels, and a relative error component that dominates at high measurement levels. The formula for the error model for raw (pre-normalization) expression levels can be written as:

$$\sigma_{RAW} = \sqrt{a^2 + b^2 S^2}$$

where σ_{RAW} is the measurement standard error of the raw expression data, S is the measurement level (control strength), and a and b are the fitted coefficients of the model.

Expressed in terms of the normalized expression levels, which are the result of dividing raw expression levels by control strength, the standard errors can be written as:

$$\sigma_{NORM} = \sqrt{\frac{a^2}{S^2} + b^2}$$

Before fitting the error model, the genes are ordered by their control strengths. A median variance and median control strength is calculated for each non-overlapping set of eleven genes. If replicates are used, this variance is the standard error of the samples in the current condition. If the “deviation from 1” option is selected, error is approximated by using the median deviation from 1.0. The goal in this step is to remove outliers (when replicates are being used) and to disregard genes whose high or low expression level is the result of biological activity. In the absence of replicates the working assumption is that the vast majority of the genes do not change over the conditions in the experiment, and thus deviation from one represents error in a gene whose expression level changes little over the course of the experiment. Then an iteratively reweighted linear regression of variation or squared deviation versus squared control strength is fitted to estimate the parameters.

Estimation of the 2-level variance components model is done by the method of moments. In order to eliminate negative estimates of variance components, within-sample variation is taken as a lower bound on total between-sample variation. Different sources of information in the analysis are weighted by their appropriate statistical degrees of freedom. Precision estimates based on replicate genes or samples are assigned degrees of freedom equal to the number of replicates minus one. User-supplied precision values, if available, are assigned 1 degree of freedom. Cross-gene error models, if used, are assigned an equal number of degrees of freedom as the direct variability estimates for that gene. Between-sample analyses are done according to the interpretation mode (ratio, log, fold). Within-sample variability is calculated in terms of normalized ratio expression, and translated as necessary to the interpretation mode by use of the delta method.

Results of the variance components analysis are used to estimate standard deviations and standard errors, according to the grouping of samples into conditions as specified by the experiment interpretation. Two different types of interpretation affect the assumed context of the calculation:

- **Single-sample interpretation**—If all conditions contain only one sample (for instance the “All Samples” interpretation), precision calculations are based solely on the estimated within-sample measurement variation. The error bars, standard deviations, and standard errors represent the variability of all possible measurements on this specific sample.
- **Multi-sample interpretation**—If at least one condition contains multiple samples, precision calculations for all samples are based on the combined within-sample and between-sample variation, and error bars, standard deviations, and standard errors represent the variation of measurements of samples representing the population of all possible samples in the condition.

In a multi-sample interpretation, if no replicate samples are available for a specific condition, then no error calculations are made and no error bars are shown, since there is no information available on the variability of that condition.

References

- Milliken, G. A. and Johnson D, E. (1984) *Analysis of Messy Data*, Volume 1: Designed Experiments. Wadsworth, Inc. Belmont, California.
- Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978) *Statistics for Experimenters*, John Wiley and Sons, New York.
- Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2:110-14.

Viewing Data

Using the Genome Browser

The large panel in the center of the GeneSpring window is the genome browser, which graphically displays information about the genes in the selected gene list. The genome browser often presents so much information that individual genes and gene names are not visible. To look more closely at fewer genes you can zoom in and pan around.

Zooming In

You can enlarge a region of the screen by “zooming in”.

Click and drag a rectangle across the region to enlarge. Release the cursor. Repeat steps 1 and 2 until you reach the desired magnification level.

To undo a zoom, type **Ctrl+[,** or click **Zoom Out**.

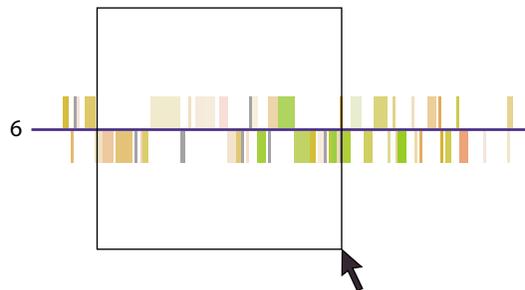


Figure 4-1 Zooming

To return directly to the unmagnified state, do one of the following:

- Select the **View > Zoom Fully Out** option.
- Click **Zoom Fully Out**.
- Type **Ctrl + Home**.

Panning

If you have zoomed in and must view genes that are not visible in the genome browser but are nearby, you can pan in any direction.

To pan, do one of the following:

- Use the arrow keys to move in the desired direction.
- Use the Page Up or Page Down keys to travel one screen’s distance up or down.

Modifying Display Options

Each of the displays in the **View** menu provide multiple options for modifying the way your data are represented. To see what display options are available in the current view, select **View > Display Options....**

For more information on display options, see “Display Options” on page 4-25.

Displaying a Gene List

Displaying a Gene List

To display a gene list, select a list with an ordinary mouse-click.

Alternately:

1. Right-click on the gene list to view in the Gene List folder in the navigator. A submenu appears.
2. Select **Display List**.

Displaying a Gene List as a Secondary List

1. Display a gene list as outlined above, then right-click above the gene list to view as your secondary gene list. A submenu appears.
2. Select **Display As a Second List**.

To remove the secondary gene list, go to the **View** menu and select **Remove Secondary Gene List**.

Show All Genes

At any time in any display mode, you can click the **Show All Genes** button to revert to a display of all genomic elements.

Finding and Selecting Genes

The Find Gene and Advanced Find Gene functions allow you to quickly find one or more genes. This is especially useful when there are too many genes in the genome browser to easily identify individual genes.

Performing a Simple Search

1. Select **Edit > Find Gene**. The Find Gene in View window appears.
2. Type a synonym, systematic name or common name of a particular gene in the Find Gene window text box.
3. Check any or all of the checkboxes for the fields to search.
4. Click **Find** or press the **Enter** key.

In some views, the genome browser zooms in on the “found” gene. This gene is automatically selected. If more than one gene is found that matches your search criteria, the total number of genes found are listed at the bottom of the Find Gene Window along with the number of genes found that are visible in the current gene list. At this point, click **Find Next** to show the next gene that matches your search criteria.

Performing an Advanced Search

1. Select **Edit > Advanced Find Gene...** The Advanced Find Genes window appears.

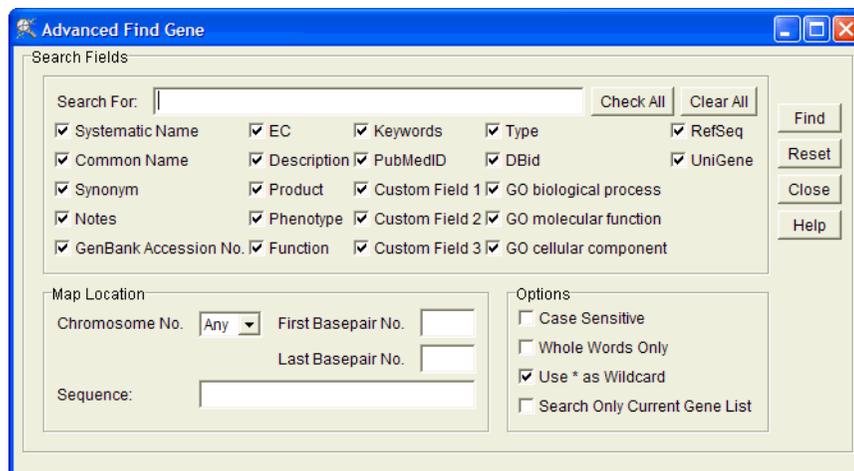


Figure 4-2 The Advanced Search Window

2. Check or uncheck the fields to search. For more information about the contents of these fields see “Annotation Options” on page 9-6.
3. Enter a search term in the **Search For** field.
 - Enter multiple terms separated by the words “and”, “or”, or “and not” in this field.
 - **AND**—match any gene containing both of these terms, even if they do not appear in the same field

- **OR**—match any gene containing either of these terms, even if they do not appear in the same field
 - **AND NOT**—match any gene containing the first term, but not containing the second
 - Enter an asterisk ‘*’ in this field to match any gene with an entry in any of the fields you specified in step 2.
 - Restrict your search to specific regions within the genome using the Map Location fields. You can specify the chromosome number or search for a particular sequence for any organism that has mapping information.
 - For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
 - You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.
4. Click **Find**. A list of the genes that match the search criteria is displayed at the base of the window. The field that matched the search criteria is colored red.
 5. Identify the genes from the list to learn more about by clicking on the appropriate row. To identify more than one gene, hold down the ctrl key while clicking multiple rows. Once you have highlighted the gene(s) of interest, click one of the five buttons on the right:
 - **Select** - to highlight and select the selected gene in the genome browser
 - **Select All** - to select and highlight all of the genes identified by the search
 - **Zoom & Select** - to select and zoom-in on a gene in the genome browser
 - **Inspect** - to bring up the Gene Inspector window for the selected gene
 - **Make Gene List** - to make a gene list from all of the genes identified by the search

Searching GeNet from GeneSpring

Many researchers use multiple types of arrays to study the same class of disease. These arrays may even come from diverse organisms, representing animal models for human diseases, etc. In this instance it may be helpful to search a pool of data much larger than that which GeneSpring defines as a “genome”.

To search for data on GeNet from within GeneSpring, select **Edit > Search GeNet**. The Search GeNet screen appears. From this screen you can search for either genes or associated data objects. However, searches will only return data you have permission to view.

If you are logged into more than one GeNet server, a dialog appears that prompts you to select the GeNet server on which to search. You cannot search multiple GeNet servers at the same time.

Search for Genes

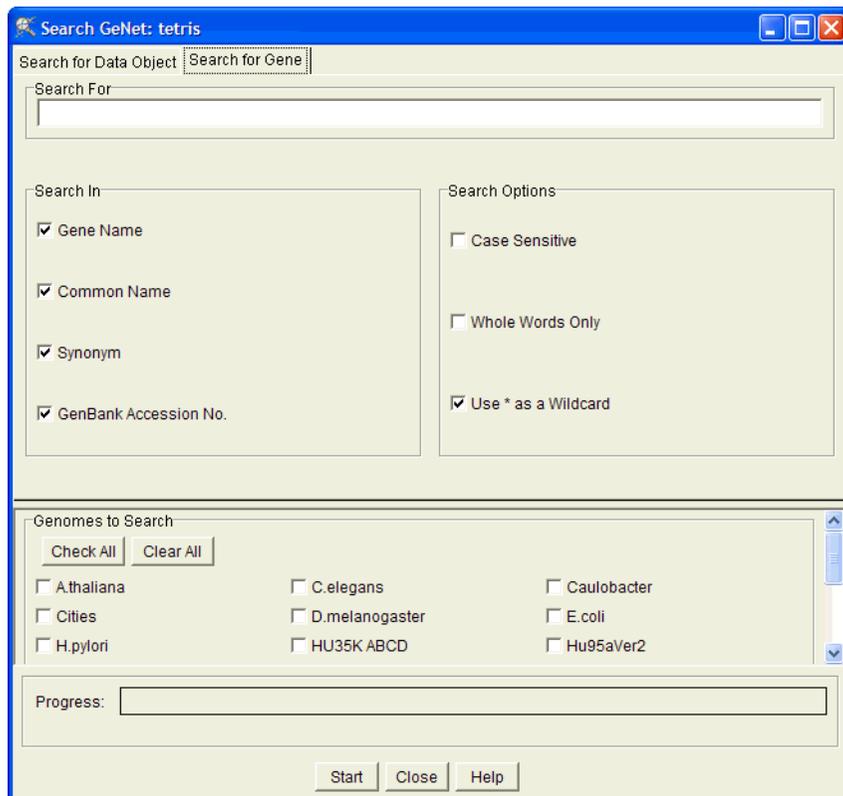


Figure 4-3 The Search GeNet for Genes window

To search for genes:

1. Enter a search term in the **Search For** field.
 - Enter multiple terms separated by the word “or” in this field to match any gene containing either of these terms, even if they do not appear in the same field.
 - Enter an asterisk “*” in this field to match any gene with an entry in any of the fields you specify in step 2.
 - For organisms that are not completely sequenced, you can search for cytogenetic band markers. For organisms that are completely sequenced you can restrict your search to regions between specified bases. Only the fields appropriate for the given genome appear. Any gene that falls even partially within the specified region is identified by the search.
 - You can restrict your search to genes containing specific sequences. These sequences can include the IUPAC-IUB ambiguity codes, A, K, Y, W etc. Note that the symbol, X, is not allowed and users who want to specify a single wildcard-base should use N instead. Searching for NNNN therefore identifies all the genes in the genome and may result in an out of memory error.
2. Check or uncheck the annotation fields to search. Searches are limited to the annotations listed on the screen.

3. Specify the genomes in which to search. The more genome names you check, the longer the search process will take.
4. Click **Start**. When the search is complete, a list of the genes that match the search criteria appears.

Search for Data Objects

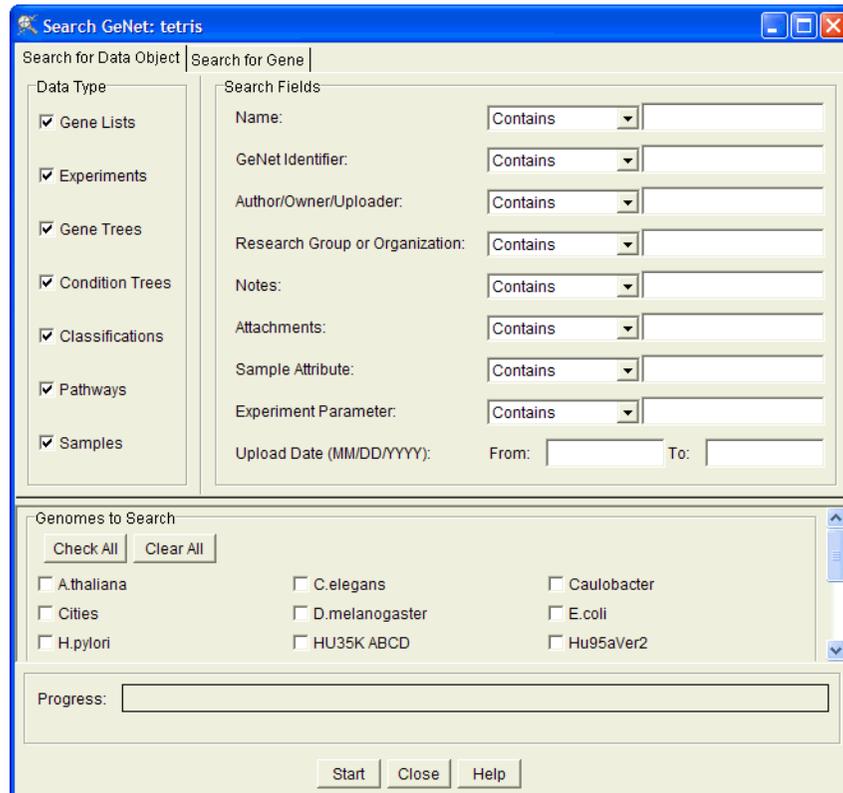


Figure 4-4 The Search GeNet for Data Objects window

To search for data objects:

1. In the Data Type section, specify the type of data to search for.
2. Enter search terms in the appropriate fields in the Search Fields section. Use the pull-down menus to specify how the search term appears in that field. Available options are:
 - Contains
 - Equals
 - Starts with
 - Ends with
 - Does not contain

All entries in the Search Fields section are considered “ands”; i.e., the search will look for data objects that match all of the terms entered on this screen.

3. Specify the genome or genomes in which to search. The more genome names you check, the longer the search process will take.

Selecting Multiple Genes

- Click once on any line or square representing a gene. Hold down Shift to add more genes. (Clicking a selected gene while holding **Shift** deselects that particular gene.)

Or...

- **Shift** and drag your mouse across genes you want to select. A box appears as you drag. When you release the mouse, the selected genes are highlighted.

When several genes are selected, the number of genes selected appears in the genome browser.

If some selected genes do not appear in the current displayed gene list, the legend displays the message “ x genes selected, y genes not in list” where x is the number of selected genes and y is the number of genes not in the list.

Click anywhere in the browser to unselect a the selected genes.

List Inspector

Right-click over a list icon in the navigator and select **Inspect**. A Gene List Inspector window appears, displaying the common and systematic names of all the genes in the gene list currently being displayed in the genome browser. You can select one of the listed genes (by double-clicking) for closer inspection. For more information on this window, see “The Gene List Inspector” on page 4-20.

Inspectors

The Inspect windows allow you to view the current defaults and available details of any gene, condition, classification or experiment.

The Gene Inspector

The Gene Inspector window allows you to look at all the data associated with a particular gene, see the lists that include your gene, make correlations, and link directly to Internet databases.

In the upper left corner of the Gene Inspector window is the name of the gene and an area for notes. The table in the upper right corner displays the normalized, control, and raw values, as well as the t-test p-value and flag for each measurement. In the center of the window is a browser showing a graph of the gene across all conditions. At the bottom of the window, from left to right, are correlation functions, lists containing your gene, and links to databases.

Accessing the Gene Inspector Window

There are three ways to access this window:

- Double-click a gene (this may be easier when you zoom in)
- Select **Edit > Find Gene** and enter the name of your gene
- Press **Ctrl+I** (when one or more genes are selected)

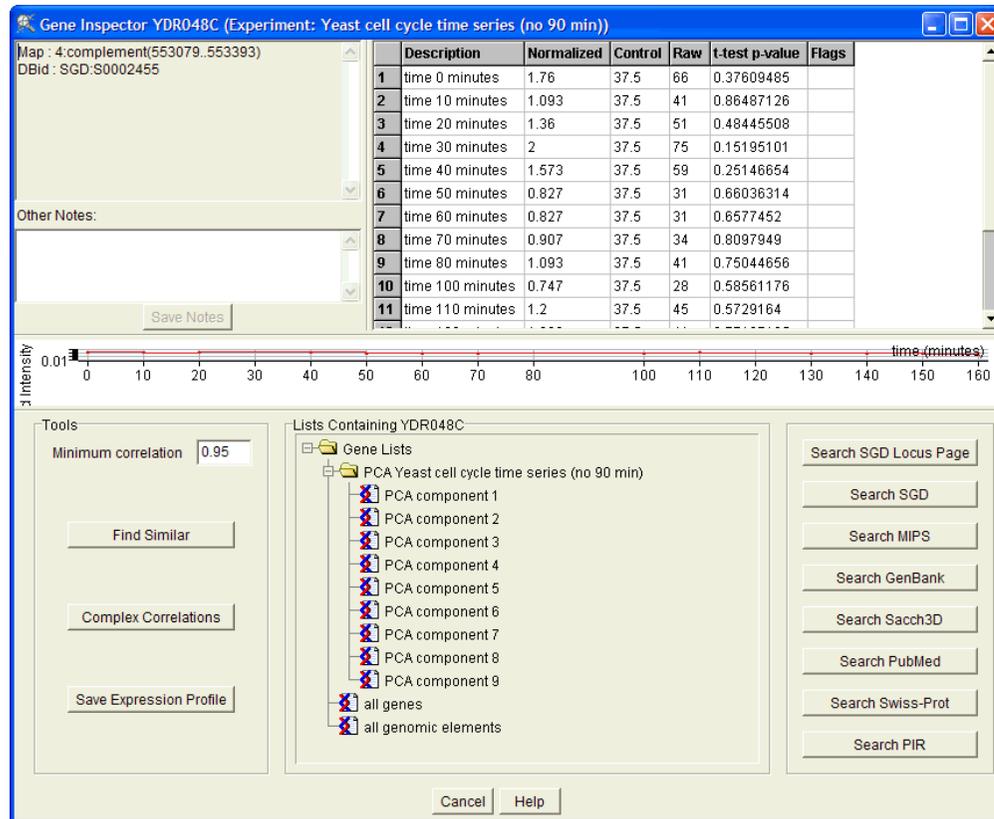


Figure 4-6 Gene Inspector window for gene RPS3

Gene Identification Section

Information on the selected gene from the master gene table is displayed in the upper left corner of the Gene Inspector window in the Gene Identification section.

The Data Table

The table in the upper right corner is the Data Table. It contains the following information:

- **Description**—The condition under which the measurement was taken.
- **Normalized**—The normalized data value. For details on normalization, see , “Normalizing Data”.
- **Control**—The control strength for the gene. For information about control strengths. See “Per Gene Normalizations” on page 5-13.
- **Raw**—The raw value of the data, just as it came off the chip or out of the scanner.
- **t-test p-value**—This is a measure of the likelihood of this gene’s expression value being different from one, assuming the data is centered around one. The t-test p-value is applicable only to replicated data.
- **Flags**—Flags indicate whether or not your data are reliable. Whether or not you have flags depends on your instrumentation and what you have entered into your master gene table. See “Measurement Flags” on page 5-19.

The T-test P-value

In cases where there is replicate data, a one-sample Student's t-test is calculated to test whether the mean normalized expression level for the gene is statistically different from 1.0. The t-statistic is calculated as:

$$t = \frac{\bar{X} - 1}{S_x / (\sqrt{n})}$$

where $\bar{X} = \sum_{i=1}^n X_i$ is the sample average of the n normalized expression levels X_1, \dots, X_n ,

$$\text{and } S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

is the sample standard deviation of the replicates. The value of t is compared with a table of the distribution of Student's t-distribution with $n - 1$ degrees of freedom to yield the significance level (or *p-value*) for a two-sided test that the mean gene intensity differs significantly from 1.0.

The Browser Display

Inspecting a gene shows you the gene's expression over the experimental parameter, time (minutes). The browser image reflects the experiment interpretation in the main browser window. The only view option available in the Gene Inspector window is the Graph view.

Right-click on the browser to use error bars in the browser display or create a resizable picture of the browser. Right-click and select **Options** to change the vertical axis range, show or hide many of the browser elements, and switch your view from normalized to raw data. For details on the latter options, see "Using the Genome Browser" on page 4-2. For information about error bars, see "Cross-gene Error Models" on page 3-44. For information about creating a resizable picture, see "Saving Pictures and Printing" on page 9-4. For information on bookmarks, see "Bookmarks" on page 4-26.

Gene Inspection Tools

The box in the bottom left corner of the Gene Inspector window contains tools allowing you to search for genes having similar expression profiles to the gene currently displayed.

- **Find Similar**—Allows you to search for genes with similar expression profiles to the gene being inspected. Each gene expression profile must have the required minimum correlation to be considered similar. The higher the minimum correlation (maximum 1), the closer the gene expression profiles must be. Enter this number in the Minimum correlation box above the **Find Similar** button. For information on using the Find Similar function, see "The Find Similar Command" on page 6-6.
- **Complex Correlation**—Allows you to make a gene list comparing the gene being inspected to genes having similar expression profiles in multiple experiments, with

more complex parameters than the Find Similar tool allows. For information on using the Complex Correlation function, see “The Find Similar Genes Window” on page 6-7.

- **Save As Expression Profile**—Allows you to save your gene expression profile as an expression profile, which you can use to make lists. For information on making lists from expression profiles, see “Creating Expression Profiles” on page 6-15.

Lists Containing Your Gene

In the bottom center of the Gene Inspector window is a navigator for the lists containing your gene. Select a list to view the Inspect List window. For information about this window, see “The Gene List Inspector” on page 4-20.

Searching Internet Databases

In the Windows version of GeneSpring, you can set up the Gene Inspector window to search public databases. See “The New Genome Installation Wizard” on page 2-2.

To configure a web browser with a Macintosh, go to **Edit > Preferences > Browser** and enter the appropriate pathway.

Notes Section

In the upper left corner of the Gene Inspector window, under the Gene Identification Section, is an area where you can make notes. To save these notes, click **Save Notes**.

The Sample Inspector

The Sample Inspector allows you to view and edit detailed data on a particular sample. It can be accessed by clicking the **Inspect** button in the Sample Manager, or by right-clicking on a sample in the navigator and selecting **Inspect**.

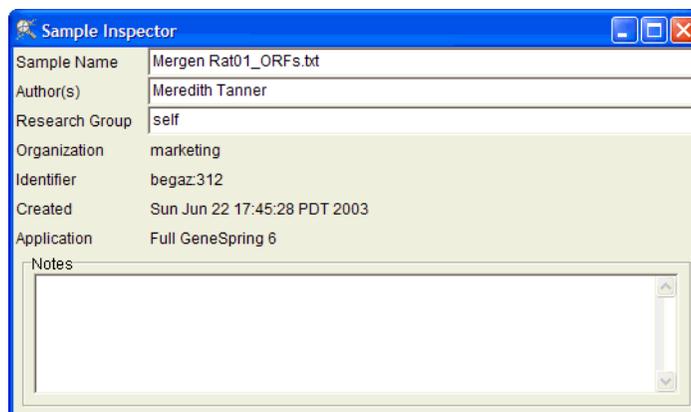


Figure 4-7 The upper half of the Sample Inspector Window

The Sample Inspector screen has two sections. The upper section contains basic information about a sample

Weblinks buttons appear in the upper right portion of the window only if there are web links associated with the genome containing the sample. These link to LIMS-type data-

bases. For more information on weblinks in genomes, see page 4 in the Genome Installation Wizard section.

The lower section contains four tabbed menus from which you can view or edit a variety of information about the sample being inspected. Click a tab to view the available options.

Attributes and Parameters

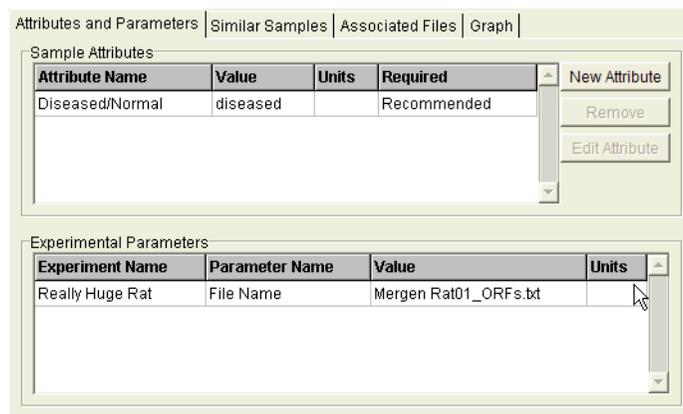


Figure 4-8 The Attributes and Parameters tab

There are two lists visible on this screen: the Sample Attributes list and the Experimental Parameters list.

In the Sample Attributes list, you can view the attributes of the sample being inspected. You can also add, remove, or edit any of these attributes.

The Experimental Parameters list displays parameters assigned to this sample in any experiment. Since a sample may be part of several experiments, you cannot edit experiment parameters from this screen. For information on how to edit experiment parameters, see “Experiment Parameters” on page 3-29.

To add a new sample attribute, click **New Attribute**. The Sample Attribute screen appears.

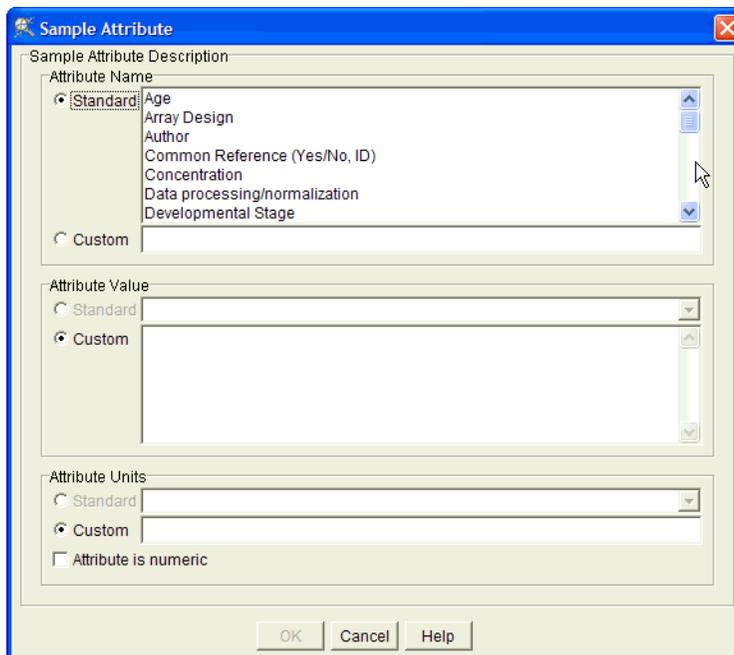


Figure 4-9 The Sample Attribute window

You can specify the following on this screen:

- **Attribute Name**—The name of the attribute being added. You can select a standard attribute name from the scrolling list, or select the **Custom** radio button and enter a new attribute name.
- **Attribute Value**—The value of the attribute being added. For many standard attributes, there is a default value. You can accept the default or select the **Custom** radio button and enter a new value.
- **Attribute Units**—The units in which the attribute is measured. For many standard attributes, there is a default unit. You can accept the default or select the **Custom** radio button and enter a new unit. If the unit is numeric, check the **Attribute is numeric** box.

When you are done, click **OK** to save your new attribute and return to the Sample Inspector.

To remove an attribute, select it in the **Sample Attributes** list and click **Remove**.

To edit an attribute, select it in the **Sample Attributes** list and click **Edit Attribute**. The Sample Attribute screen appears. Proceed as you would if you were adding a new attribute.

The Similar Samples Tab

This tab lists the correlation between the current sample and the samples in the currently selected experiment. This list appears only when an experiment containing the sample is selected.

You can click any of the samples in this list and view them in another Sample Inspector window by clicking **View Sample**.

The Associated Files Tab

This tab lists any files that may be associated with the sample, including data files, array images, sample images, etc. If there is a sample image included among the associated files, it is displayed in the **Sample Image** panel to the right of the list.

You can re-order the files in this list by clicking the property to sort by in the list headers. For example, to sort by file type rather than file name, click the **File Type** column header.

From this screen you can do the following:

- **Add File**—To add a file, click **Add File** and select the desired file from the Browse menu that appears. You can also drag and drop a file directly from the desktop into the Associated Files list.
- **Extract File**—To save (extract) a file in the list to another location, select the desired file from the list and click **Extract File**. Choose a location from the Browse menu and click **Save**. This does not remove the file from your list. It simply places a copy of the file in a new location.
- **Delete File**—To remove an associated file, select it in the list and click **Delete**.
- **View File**—Select a file name in the list and click **View File** to view the contents of the file in an external program. The appropriate program to display is selected automatically the file if the file type is known.
- **View Data File Format**—Click to display the column assignments for the selected file as it was loaded into GeneSpring.

The Graph Tab

This view is available only if an experiment containing the current sample is selected. It displays a graph of the raw sample data.

The Experiment Inspector

Just as you can inspect a gene with the Gene Inspector window, you can inspect experiments with the Experiment Inspector window.

Accessing the Experiment Inspector

1. Right-click over the name of any experiment in the navigator.
2. Select **Inspect** from the pop-up menu.

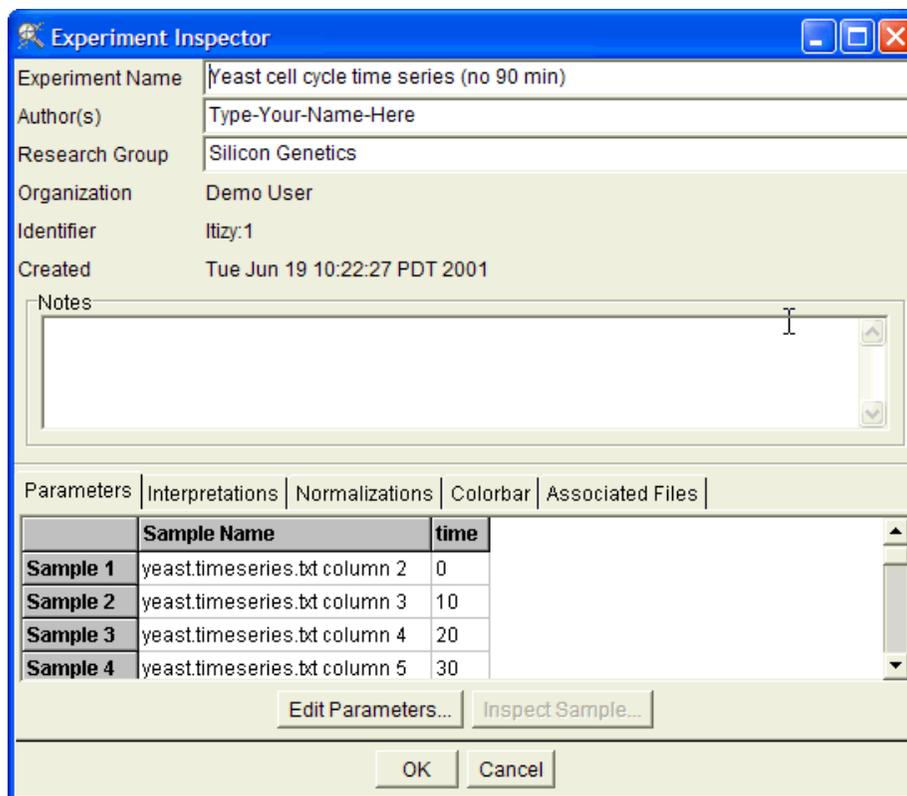


Figure 4-10 The Experiment Inspector window

The upper section of the Experiment Inspector window contains the experiment information. You can edit the text in the white boxes as desired. Below this are five tabs. Click a tab to view its contents.

There are two buttons at the bottom of the window, regardless of which tab is active. Click the **OK** button to save your data and exit the window. Click **Cancel** to close the Experiment Inspector without saving any changes.

The Parameters Tab

On the Parameters tab you can view the experiment parameters and their possible values.

Click **Edit Parameters** to view the Change Parameters window. See “Experiment Parameters” on page 3-29 for details on this window. When you click **OK**, any changes you make are saved and applied to your experiment.

To view details on a particular sample, select it in the list and click **Inspect Sample**. This invokes the Sample Inspector window. For more information about the Sample Inspector, see “The Sample Inspector” on page 4-13.

The Interpretations Tab

This tab allows you to view all the interpretations associated with the selected experiment. Click on an interpretation in the list to select it. To edit an interpretation, double-click it or select it in the list and click **Edit Interpretation**. The Change Interpretation win-

dow appears. When you click **OK**, any changes you make are saved and applied to your experiment.

The Normalizations Tab

This tab allows you to view the normalizations currently being used in your experiment. To edit normalizations, click **Edit Normalizations**. The Experiment Normalizations window appears. See “Experiment Normalizations” on page 5-2 for details on this window. Click **OK** to save your changes.

To view a more detailed description of a particular normalization, select its name in the list on the Normalizations tab and click **View Text Description**. A dialog appears with a description of the selected normalization. You can copy the text in this dialog to the keyboard by clicking Copy to Clipboard. You can then paste the text into a text editor.

The Colorbar Tab

Use this tab to edit the default range of expression in your experiment’s coloration scheme. You can also specify whether or not to show trust on the colorbar by clicking the radio button next to your preferred choice. Click **OK** to save your changes or **Cancel** to exit without saving.

Use this tab to view the default range of expression in your experiment’s coloration scheme. You can also see whether or not trust is shown on the colorbar by clicking the radio button next to your preferred choice.

For more information about the range of expression and how it affects the coloration of your experiment, see “Changing the Colorbar Range” on page 4-32

The Associated Files Tab

This screen lists any files that may be associated with the experiment, including data files, array images, sample images, etc.

From this screen you can do the following:

- **Add File**—To add a file, click **Add File** and select the desired file from the Browse menu that appears. You can also drag and drop a file directly from the desktop into the Associated Files list.
- **Extract File**—To save (extract) a file in the list to another location, select the desired file from the list and click **Extract File**. Choose a location from the Browse menu and click **Save**. This does not remove the file from your list. It simply places a copy of the file in a new location.
- **Delete File**—To remove an associated file, select it in the list and click **Delete**.
- **View File**—Select a file name in the list and click **View File** to view the contents of the file in an external program. GeneSpring automatically selects the appropriate program to display the file if the file type is known.

The Condition Inspector

A condition is a unique combination of parameters as applied to your sample. Each condition may be a single sample or a group of replicate samples combined based upon the

parameter values defined for each sample. The easiest way to think of this is as the parameters under which the sample(s) was observed. If you have no replicates, condition and sample can be considered synonymous.

1. Open the **Experiment** folder in the navigator by clicking on its icon.
2. Click the + sign next to the experiment icon.
3. Click the + sign next to the interpretation icon.
4. Right-click over a condition.
5. Select **Inspect** from the pop-up menu.

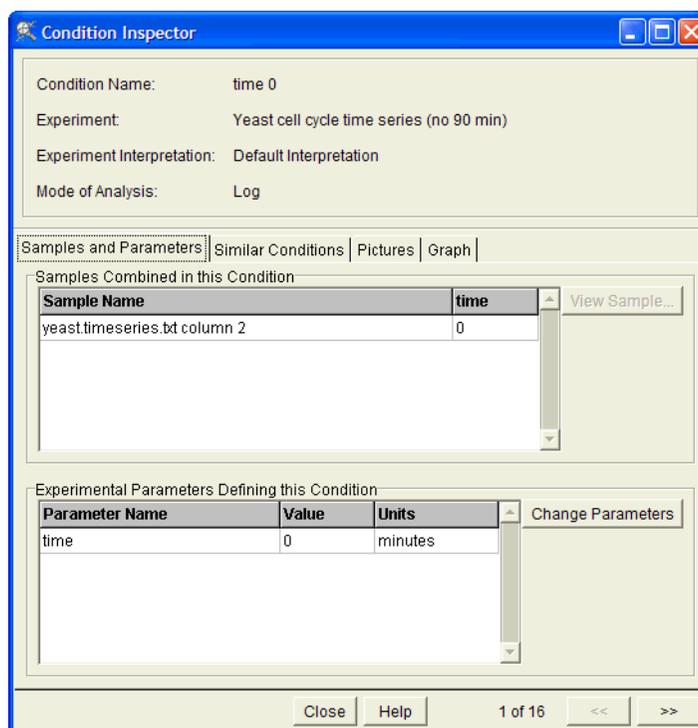


Figure 4-11 The Condition Inspector window

The Samples and Parameters Tab

This tab contains two sections:

- **Samples Combined in this Condition**—Lists the samples in the selected condition. Select a sample and click **View Sample** to invoke the Sample Inspector for that sample. See “The Sample Inspector” on page 4-13 for more information.
- **Experimental Parameters Defining this Condition**—Lists the parameters associated with this condition. To edit parameters, click **Change Parameters**. For more information on the Change Parameters window, see “Experiment Parameters” on page 3-29.

The Similar Conditions Tab

This tab contains a list of similar conditions in the experiment with columns corresponding to their associated values.

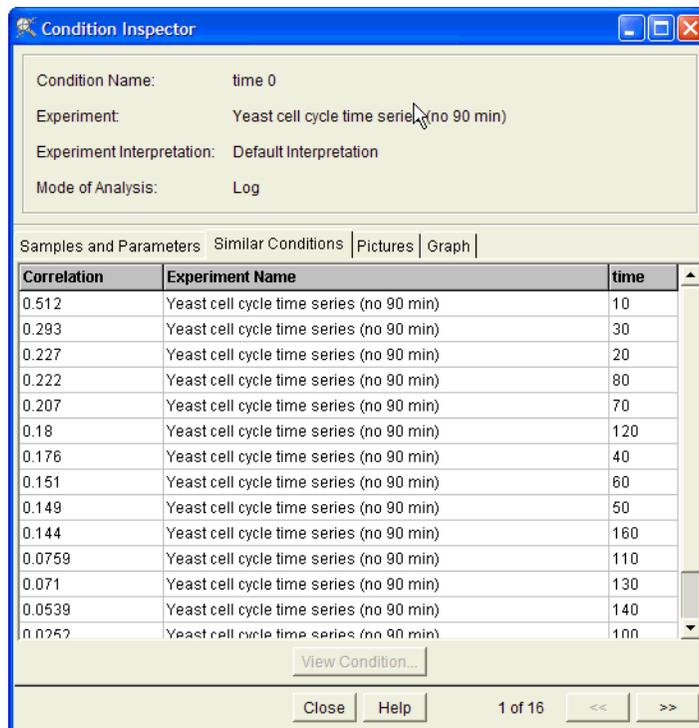


Figure 4-12 The Similar Conditions tab

The Correlation column lists how closely correlated the other conditions in the experiment are to the one under inspection. The conditions are listed from most closely correlated to least correlated. This feature uses “standard correlation” to measure the similarity of the selected condition and all others in the experiment. This cannot be changed. To use another metric, you must create a script using the “Condition Correlation” building block. For details on creating scripts, see “Creating Scripts” on page 8-13.

The Pictures Tab

This tab displays the sample images, if any, associated with this condition. Double-click an image to view it at its full size.

The Graph Tab

Displays the condition in graph form.

The Gene List Inspector

You can view the contents of a gene list and its creation method using the Gene List Inspector window. This window is especially useful for learning about lists identified using the Similar List function.

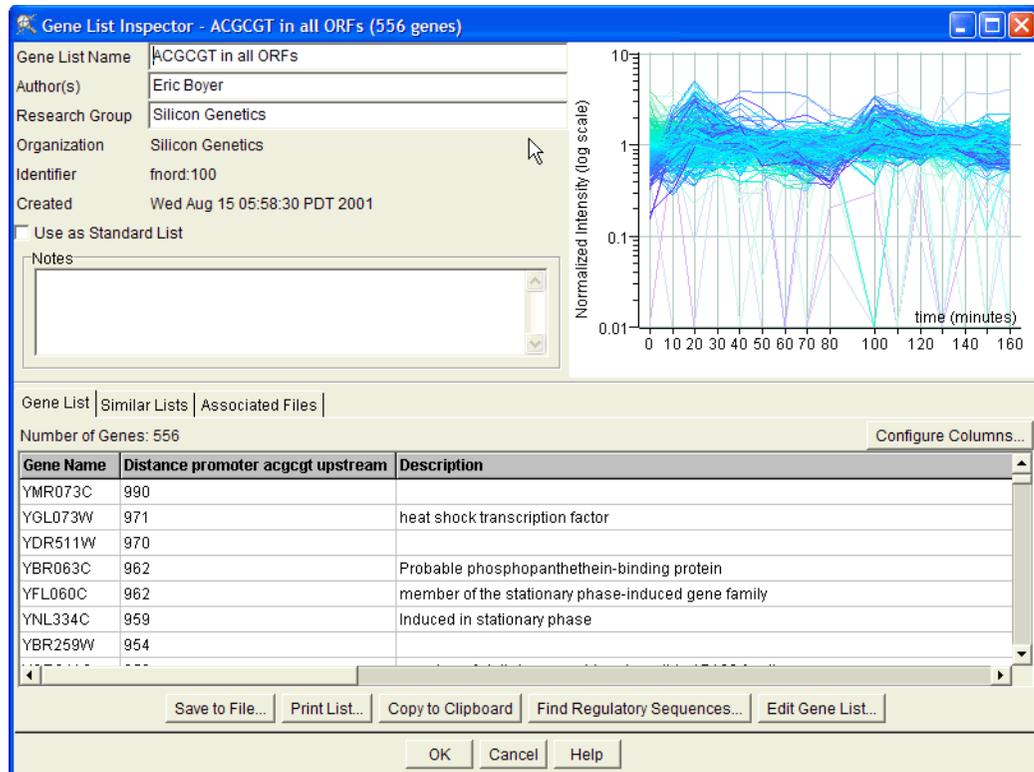


Figure 4-13 The List Inspector window

The history of the selected gene list is displayed in the upper left corner of the window. You can edit the information in these fields.

In the upper right corner of the window is a browser graphing your list. Right-click on the graph for a menu of options. See “Using the Genome Browser” on page 4-2 for information on browser options.

Below this section there are three tabs containing a variety of options. Click a tab to view its contents. At the bottom of the screen, regardless of which tab is active, are three buttons. Click **OK** to save your changes and exit. Click **Cancel** to exit without saving. Click **Help** to view available documentation for the Gene List Inspector.

The Gene Lists Tab

This tab displays a table of all the genes included in the selected list. Double-click a gene or cell in this table to view a Gene Inspector window for the selected gene. See “The Gene Inspector” on page 4-10 for information on the Gene Inspector window. Click on any column header in the displayed gene list to sort the table by the values in that column.

From this tab, you have the following options:

- **Configure Columns**—Allows you to select which columns to display on the Gene Lists tab. You can choose from any of the columns in your Master Table of Genes except the Sequence column.
- **Save to File**—Allows you to save the entire gene list as a plain text file.
- **Print List**—Prints the selected gene list.

- **Copy to Clipboard**—Copies the contents of the gene list to the clipboard. You can then paste the list into another application, such as a text editor.
- **Find Regulatory Sequences**—Opens the Find Potential Regulatory Sequences window with the current gene list pre-selected. This button is available only if the genome is fully sequenced. For more information on this window, see “Regulatory Sequences” on page 6-18.
- **Edit Gene List**—Opens the Gene List Editor window. For more information, see “Creating and Editing Gene Lists” on page 6-2.

The Similar Lists Tab

This tab displays names of lists resembling the selected list, or containing a statistically significant number of overlapping genes.

There are two ways to view these lists:

- **List View**—Displays a simple two-column list. In this view, statistical significance is listed as the p-value for each of the similar lists.
- **Navigator View**—Displays a navigator-style listing.

Right-click a list to print or copy. Double-click a list to view a Gene List Inspector window for that list.

The Associated Files Tab

This screen lists any files that may be associated with the selected gene list, including data files, array images, sample images, etc.

From this screen you can do the following:

- **Add File**—To add a file, click **Add File** and select the desired file from the Browse menu that appears. You can also drag and drop a file directly from the desktop into the Associated Files list.
- **Extract File**—To save (extract) a file in the list to another location, select the desired file from the list and click **Extract File**. Choose a location from the Browse menu and click **Save**. This does not remove the file from your list. It simply places a copy of the file in a new location.
- **Delete File**—To remove an associated file, select it in the list and click **Delete**.
- **View File**—Select a file name in the list and click **View File** to view the contents of the file in an external program. The appropriate program is automatically selected if the file type is known.

The Classification Inspector

The Classification Inspector allows you to learn about the method used to construct a classification or to learn more about the variability explained by each class within a classification. To use the Classification Inspector, right-click a classification in the navigator panel and select the **Inspect** option.

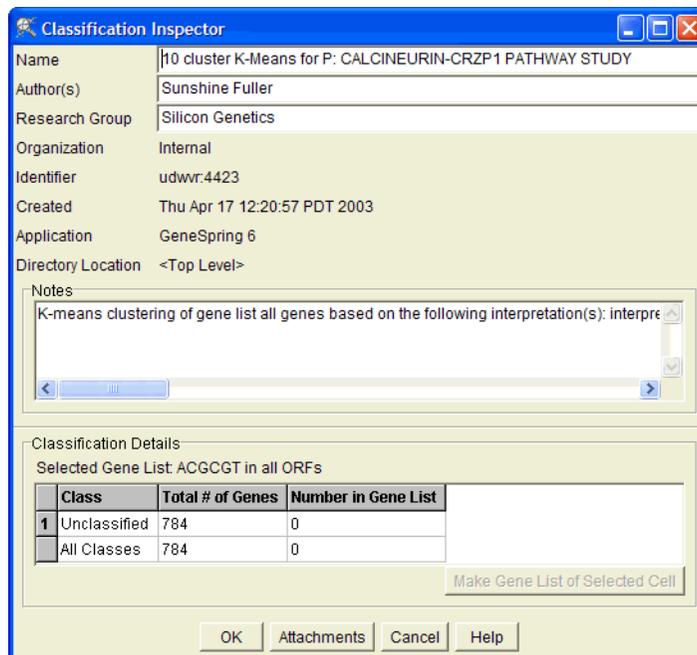


Figure 4-14 Classification Inspector for a *k*-means clustering with 10 groups

In Figure 4-14, the notes field contains information about the method used to make the classification. If the classification is the result of clustering, this field displays information such as the type of clustering, the distance metric, and the number of iterations that were used to perform the clustering. You can save your own comments about the classification here for future reference. The bottom half of the Classification Inspector contains a table with three columns:

- **Class**—the name given to each class
- **Genes**—the number of genes in each class
- **Average Radius**—the root mean square of the Euclidean distances between each gene and the centroid of each class. Classes with large radii are spread out and classes with small radii are tightly grouped.

Percent Explained Variability

In the Classification Inspector, there is an improved formula for calculating percentage explained variability. This new formula properly weights classes by the number of genes in each class.

Let:

c be the number of classes (including “unclassified”, but not “no data”).

n be the total number of genes with data.

n_i be the number of genes with data in class i .

D_i be the distance of each class centroid from the overall data centroid.

d_{ij} be the distance of each gene from the centroid of its class.

Calculate:

$$B = \sum_{i=1}^c n_i D_i^2$$

$$W = \sum_{i=1}^c \sum_{j=1}^{n_i} d_{ij}^2$$

$$E = 100 \times \max\left(1 - \frac{W/(n-c)}{(B+W)/(n-1)}, 0\right) \quad (\text{If } c \geq n \text{ then } E=0)$$

E is the percent variability explained.

Note that the percent explained variability depends on the selected experiment, and the selected gene list. It is calculated using Euclidean distance of the gene expression profiles of the conditions interpreted in the interpretation made (ratio, log, fold). Details about the number of genes in each class matching the selected gene list, and the number of those with available data, are shown in the class table. Any of these lists of genes can be examined by selecting a table cell with a Gene count and clicking on Make Gene List of Selected Cell.

References for the Classification Inspector

Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1-27.

Gordon, A. D. *Classification*, 2nd Ed. Monographs on Statistics and Applied Probability 82. Chapman & Hall/CRC, Boca Raton (1999).

Display Options

Linked Windows

Allows you to select one gene or gene list in two windows simultaneously. Simply select a gene or gene list in one window and the same gene or gene list is automatically selected in the other window.

To create a linked window, go to the **File** menu and select **New Linked Window**.

Split Windows

Another interesting way to view classifications is with the Split windows function. The Split windows feature allows you to see multiple sets simultaneously in the main GeneSpring screen.

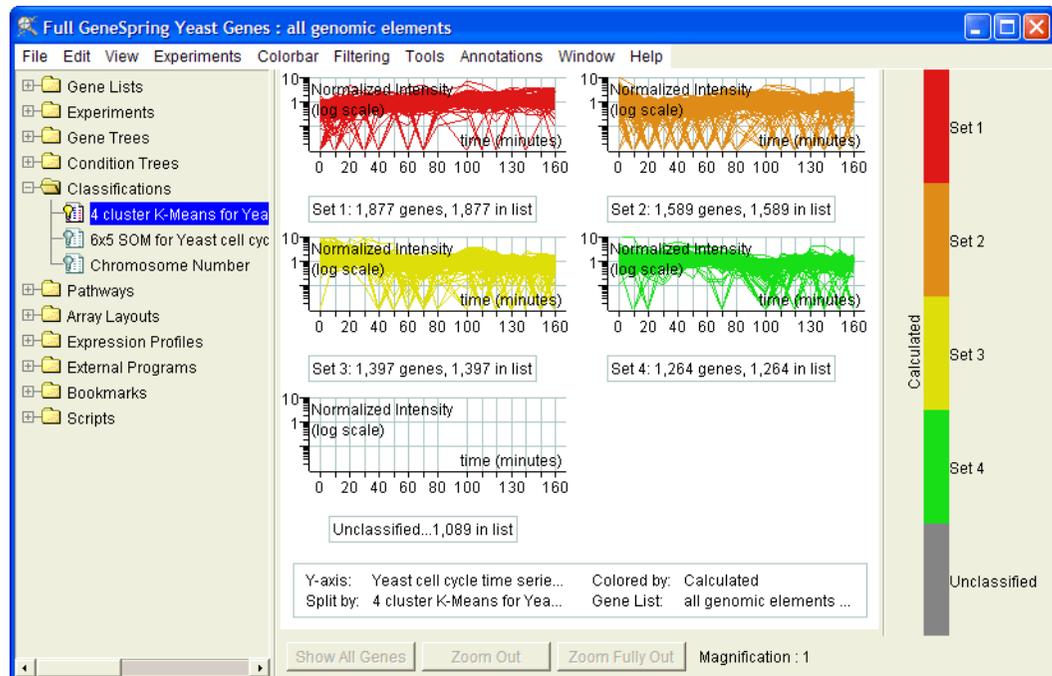


Figure 4-15 Example of a k-means clustering

Display Options

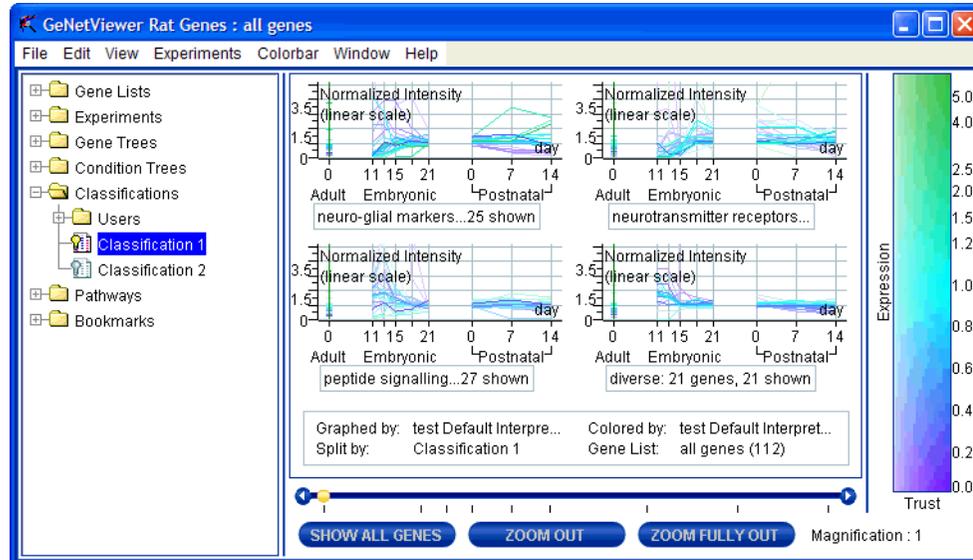


Figure 4-16 A split window

In Figure 4-15, the example represents a k-means clustering, colored by expression values. Note the list name and number of genes shown in the upper right corner of each small screen. In this instance, the names are set numbers from the original k-means clustering.

To reach the split windows command, right-click over any item in the classification folder or any folder of classifications and move the cursor down to **Split window**. A small pop-up menu appears.

Select one of the options. The main screen of the genome browser splits into several small screens. Notice the number of genes beneath each small screen. In addition, clicking any classification automatically splits the window in both directions.

Note: In the Eisen-like subtree view, the thumbnail of the full tree remains in its usual position, but no marquee is shown specifying the subtree that has been zoomed in on. Each classification shows the same subtree.

To unsplit the screen, select **View > Unsplit window** or right-click over the original data object and select **Split > Neither**.

You can also hide the labels appearing in the main window.

All of the Hide and Show commands are simple toggle switches. Re-select that option to show what has been hidden. You may have to enlarge your screen before you can see all the labels.

Bookmarks

If you ever need to pause in the midst of your analysis, you can create a Bookmark to hold your place. The Bookmark saves all your current display settings, including experiment, gene list, coloration, and selected genes.

Creating a Bookmark

1. Go to the **File** menu and select **Save Bookmark**. The Save Bookmark dialog box appears.
2. Name your bookmark.
3. Click **Save**.

Accessing an Existing Bookmark

1. Click the Bookmarks folder in the navigator.
2. Click the name of any bookmark to open.
Or:
 1. Go to **File** and select **Load Bookmark File**. The Load Bookmark dialog box appears.
 2. Select your bookmark.
 3. Click **Open**.

The Vertical Axis

In Graph, Graph by Genes, Bar Graph, Scatter Plot, and 3D Scatter Plot modes, the display options window contains a panel for modifying the presentation of data on the vertical axis. The following section describes the vertical axis options for each of these views except the scatter plot views (see “Scatter Plot Display Options” on page 4-45 and “3D Scatter Plot Display Options” on page 4-49).

The Display Options window allows you to select the experiment interpretation that is graphed. To change the current interpretation:

1. Select **View > Display Options...**
2. Click the **Vertical Axis** tab.
3. Select an interpretation from the Display Options navigator panel.
4. Click **Graph Experiment>>**.
5. Click **Apply**.

The vertical axis display format can be altered by modifying the experiment interpretation or by using the display options window. If the vertical axis format is modified in the display options window, this new format overrides the display settings within each interpretation. To modify the vertical axis format using the display options window:

1. Select **View > Display Options...**
2. Click the **Vertical Axis** tab.
3. Uncheck the **Lock Vertical Axis Format to the Interpretation** checkbox.
4. Select a value to graph (Normalized, Control, Raw, Average of Raw and Control, or Max of Raw and Control).

Display Options

5. Select a graph mode (Linear, Logarithmic, or Fold Change)
6. To adjust the vertical axis so that all measurements are visible, check the **Scale Vertical Axis to Show all Values** box. The upper and lower bounds are adjusted automatically. Alternatively you can manually set the upper and lower bounds to values of your choosing.
7. Click **Apply**.

In addition to the vertical axis format, you can also modify the tick spacing on the vertical axis. Unlike the vertical axis format, the tick spacing is not specified in the interpretation. To manually adjust the tick spacing:

1. Select **View > Display Options**
2. Click the **Vertical Axis** tab.
3. Uncheck the **Automatic Tick Spacing on Vertical Axis** box.
4. Enter the distance between major ticks in the **Major Tick Interval** field.
5. Enter the number of divisions between major ticks in the **Minor Ticks per Major Tick** field. Note that the number of *visible* tick-marks between major ticks is one less than the number you enter.
6. Click **Apply**.

Error Bars

You have the option of using error bars in the Graph and Scatter Plot views. To turn the error bars on, right-click in the genome browser and select **Display Options**. Click the Error Bars tab. The error bars are visible in the Gene Inspector as well as in the main GeneSpring window.

You can choose one of the following three kinds of error bars:

- Standard Error
- Standard Deviation
- Minimum/Maximum Value of Each Gene

In Figure 4-15, the example represents a k-means clustering, colored by expression values. Note the list name and number of genes shown beneath each small screen. In this instance, the names are set numbers from the original k-means clustering.

Legend

You can specify what information to display in most views using the Legend tab on the Display Options screen.

1. Select **View > Display Options**
2. Click the **Legend** tab.
3. Check the **Show Legend** box to display the information you specify. Uncheck the box to show no text information.

4. Check or uncheck these options as desired.
5. Click **Apply**. Your changes are applied to the display in the main GeneSpring window. The available options depend on whether they are applicable to the current view.

Legend Options

Selected Object in Navigator

Displays the following:

- In Tree view, the name of the selected gene tree and condition tree
- In Array Layout view, the name of the array layout
- In Pathway view, the name of the pathway

Experiment(s) Plotted on Axis

Displays the following:

- The name of the experiment and interpretation being displayed on the Y Axis. This is available in the following views: Graph, Graph by Genes, Bar Graph, Scatter Plot, and 3D Scatter Plot.
- The name of the experiment and interpretation (or gene list and type of associated values) being displayed on the X Axis. This is available in the Scatter Plot and 3D Scatter Plot views.
- The name of the experiment and interpretation (or gene list and the type of associated values) being displayed on the Z Axis. This is available in the Scatter Plot and 3D Scatter Plot views.

Split Window Information

Displays the name of the Classification or Gene List folder used to split the window.

Coloring Information

Displays different information depending on the coloring scheme selected:

- **Color by Expression**—The name of the experiment, interpretation, and condition used for coloring. For experiments with continuous numeric parameters, the “condition” may actually be an interpolation between two measured conditions. In Scatter Plot and 3D Scatter Plot views, the parameter value is also displayed since it affects where the genes are graphed.
- **Color by Significance**—The name of the experiment, interpretation, and condition used for coloring. For experiments with continuous numeric parameters, the “condition” may actually be an interpolation between two measured conditions. In Scatter Plot and 3D Scatter Plot views, the parameter value is also displayed since it affects where the genes are graphed.
- **Venn Diagram**— displays “Venn Diagram”
- **Color by Parameter**—If the experiment has parameters designated as color codes, displays the name of the experiment, interpretation, and parameter(s) used for coloring. In Scatter Plot and 3D Scatter Plot views, the parameter value is also displayed since it affects where the genes are graphed.

Display Options

- **Color by Classification**—The name of the Classification or Gene List Folder used for coloring.

Equation for Line of Best Fit

(Scatter Plot view only) If Line of Best Fit is selected, displays the equation for the line of best fit (data-dependent).

Error Bar Information

Displays whether the error bar is based on Standard Error, Standard Deviation or the minimum/maximum data values, and whether the error/deviation information is based on within-sample information or between-sample information. This is available in the following views: Graph, Graph by Genes, Bar Graph, Scatter Plot, and 3D Scatter Plot.

Gene List and Information on Selected Genes

Displays the name of the selected gene list, the number of genes in the gene list, and the name of the selected gene (if only one is selected) or the number of selected genes (if multiple are selected).

Secondary Gene List Name

If a secondary gene list is being displayed, this displays the name of the secondary gene list and the number of genes in this list

Condition or Gene List Sorted By

In the Graph by Genes View, displays the name of the gene list or the name of the condition, experiment and interpretation used to sort the genes on the X Axis.

Color

Color by Expression

This option colors genes according to their normalized expression values and trustworthiness. To color your genes by expression, select **Colorbar > Color by Expression** or select the **Expression** option from the Coloring tab in the Display Options dialog.

Expression

The vertical axis of the colorbar represents expression levels on a continuous scale. Using the default colors, red indicates overexpression, yellow indicates average expression, and blue indicates underexpression. Genes are colored by their expression level in the selected condition as indicated by the condition line. If you have specified the parameter on the horizontal axis to be continuous, expression levels in between conditions are interpolated.

Trust

The horizontal axis of the colorbar indicates the degree to which you can trust your data, where dark or unsaturated colors represent low trust, and bright, saturated colors represent high trust. GeneSpring uses the following guidelines to automatically create trust values:

- In **two-color experiments**, the trust value is usually the control channel (typically Cy5), unless you do a per chip normalization in which case it is:

(the control channel) x (the median of the control channel) x
(the median of the signal channel)

- For **Affymetrix and other one-color experiments**, the trust value is constructed based on the normalizations you have chosen. If you accept the default normalizations for Affymetrix data (use distribution of all genes using the 50th percentile and normalize to the median for each gene), then trust is:

(the median value of the chip) x (the median value of the gene)

- If you choose to use distribution of all genes using the 50th percentile and normalize to sample(s), trust is calculated as follows:

(the median value of the chip)
x
(the average of the gene's measurement in control samples)

Changing Colorbar Settings

To set the trust interpretation:

1. Right-click the colorbar.
2. Click **Set Coloring**. The Display Options screen appears, with the Coloring tab pre-selected.
3. Click **Set Colorbar Range**. This button is active only when coloring by expression.

The Colorbar Range dialog appears.

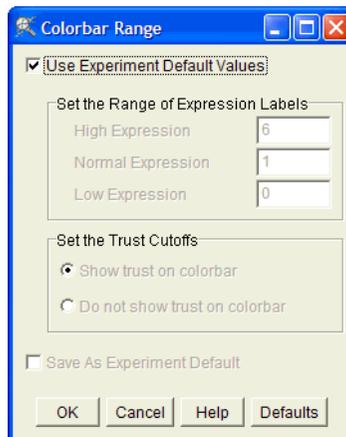


Figure 4-17 The Colorbar Range dialog

4. To enable custom settings, uncheck the **Use Experiment Default Values** box.
5. Enter values for High Control Strength, Medium Control Strength, and Low Control Strength.
6. (optional) By default, trust is shown on the colorbar. To disable this, select the **Do not show trust on colorbar** radio button.
7. To save these settings as the default for this experiment, check the **Save As Experiment Default** box. If you leave this box unchecked, your changes will affect the display options only in your current session.

8. Click **OK**.

Changing the Colorbar Range

1. Right-click over the colorbar and select **Set Coloring** from the pop-up menu.
2. Select **Color by Expression** from the pull-down menu.
3. Reset the values determining the intensity of the colors used by the genome browser.
4. Click **OK**.

There are six categories you can change:

- **High Expression**—High expression refers to the normalized expression of your genes, it is the vertical axis of the color bar. The default for this is 6.0.
- **Normal Expression**—For most normalization procedures the data are normalized to 1.0. The default for this is 1.0.
- **Low Expression**—For most normalization procedures the data do not have negative numbers. The default for this is 0.0.

For example, you could change the usual range of an experiment to high = 10, normal = 5 and low = -2 resulting in a very different color scheme once you click **OK**.

There is no **Edit > Undo (Ctrl+Z)** function for this type of change. To return to your previous coloration scheme, you must re-open the Experiment Data Range pop-up window and enter your old values.

For more details on trust, see “Trust” on page 4-30. For more details on normalization, see , “Normalizing Data”.

Color by Significance

Data are colored based on how far the gene is over- or underexpressed (relative to a normalized expression level of 1), in terms of the standard error of the measurement. The standard colorbar is replaced with a colorbar ranging from $+3\sigma$ to -3σ . The standard error model is based on the Cross-gene Error Model, if the Cross-gene Error Model is turned on. (For more information about the Cross-gene Error Model, see “Cross-gene Error Models” on page 3-44.) Otherwise the standard error is based on the standard deviation of the replicate data for a particular gene and condition (for information about the calculation of this error, see “The Gene Inspector” on page 4-10).

To color your genes by significance, select **Colorbar > Color by Significance** or select the **Significance** option from the pull-down menu on the Coloring tab in the Display Options window.

Color by Venn Diagram

This option colors genes based on their membership in one or more gene lists in a Venn diagram. To assign a gene list to the Venn diagram:

1. Select **Colorbar > Color by Venn Diagram** or select the **Venn Diagram** option from the pull-down menu on the Coloring tab in the Display Options window.
2. Drag the list from the navigator to the appropriate section of the venn diagram at the right side of the window.

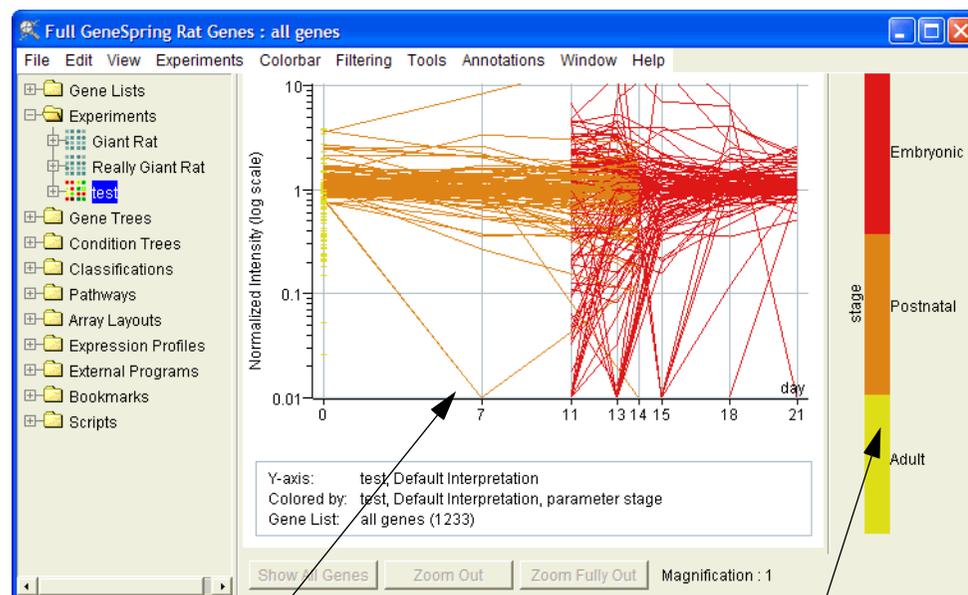
You can also assign the circles in a Venn diagram by right-clicking on a gene list and selecting the **Venn Diagram** option. For more information about creating Venn diagrams and using them for analysis, see “Making Lists with the Venn Diagram” on page 6-13.

Color by Parameter

This option colors genes based on the value of parameters. This coloring scheme is best suited for use with Graph view and Bar Graph view when different conditions are indicated with discrete symbols.

To color by parameter:

1. Select **Experiments > Change Experiment Interpretation**.
2. Choose the parameter(s) to color by and click **Color Code** for that parameter. Click **Save** to create a new interpretation.
3. Select **Colorbar > Color by Parameter**.



The conditions in the selected interpretation

Parameter values in alphabetic order

Figure 4-18 An experiment colored by parameter

You can also choose to color by parameter from the Display Options menu:

1. Select **View > Display Options** and click the Coloring tab.
2. Select an experiment from the navigator on the left side of the Display Options menu.
3. Click the **Set Experiment** button.
4. Click **OK**.

No Color

This option allows you to view genes with no coloration, showing all genes in gray. To implement this option, select **Colorbar > No Color**. You can also select a single color in which to display genes by selecting the **Solid Color** option from the pull-down menu on the Coloring tab in the Display Options menu.

Color by Classification

This coloring scheme allows you to color-code the genes by some previously defined knowledge about them. You can use a folder of lists to color by classification or a classification method such as k-means or SOM.

To color by a previously saved classification:

1. Open the **Classifications** folder by clicking its icon.
2. Select a classification by right-clicking over the name.
3. Select **Use Coloring** from the pop-up menu and GeneSpring automatically updates to reflect the new coloring scheme.

The colorbar shows the names of the sets present in the chosen classification.

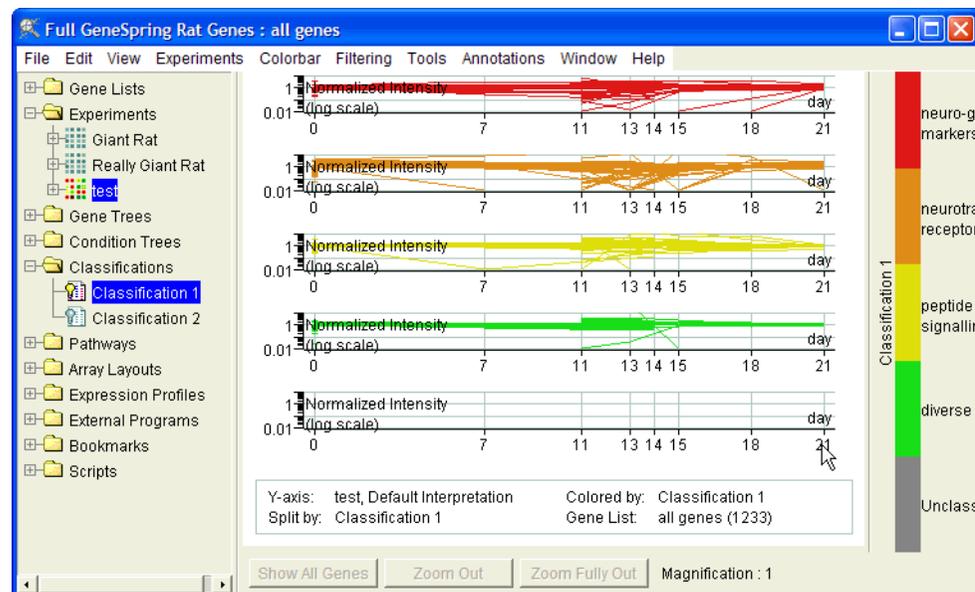


Figure 4-19 A Split Window, colored by Classification

You can also color by classification using the Display Options window:

1. Select **View > Display Options** and click the Coloring tab.
2. Select a classification from the navigator on the left side of the Display Options window.
3. Click **Set Experiment**.
4. Click **OK**.

Split Window and Color by Classification

You can also use the Split Window feature with the Color by Classification scheme.

1. Select a gene list to view.
2. Right-click over a folder or a previously saved classification and select **Use as Classification**.
3. Right-click over that folder again and select **Split Window > Both**.

Color by Secondary Experiment

The Graph and Scatter Plot displays lend themselves to being colored in many different ways because the display presents expression levels of the genes through the entire experiment. These are the only views in which you may choose to color the genes by a secondary experiment. This means the color of each gene line graphed correlates to the expression level of that gene in a different experiment, at the point in the second experiment marked by the secondary scroll bar.

1. From the navigator, open the **Experiments** folder by clicking on its icon.
2. Position your cursor over an experiment (not the one currently displayed) you want to use for coloration.
3. Right-click and select **Set Secondary Experiment** from the pop-up menu.

The coloring scheme of the genome browser is shown in the colorbar on the right. There are two versions of the animation controls in the Experiment Specification Area.

Changing the Default Colors

You can change the colors used to display the genes. This does not affect interpretation of your data, but it can help you to make genes more visible on-screen or make it easier to print screen shots.

1. Select **Edit > Preferences** and click the **Color** tab, or click the **Change Colors...** button on the **Display Options > Color** tab.
2. Select the type of information whose color you want to change and click **Change**.
3. Adjust the sliders until the color you want is displayed in the preview window at the top of the Structure Color window.
4. Click **OK**.

For more details about the other options in the Preferences window, see “Setting Preferences” on page 1-18.

Blocks View

This view displays a rectangle for every gene in the active genome, ordered by trust. To choose the Blocks view, select **View > Blocks**.

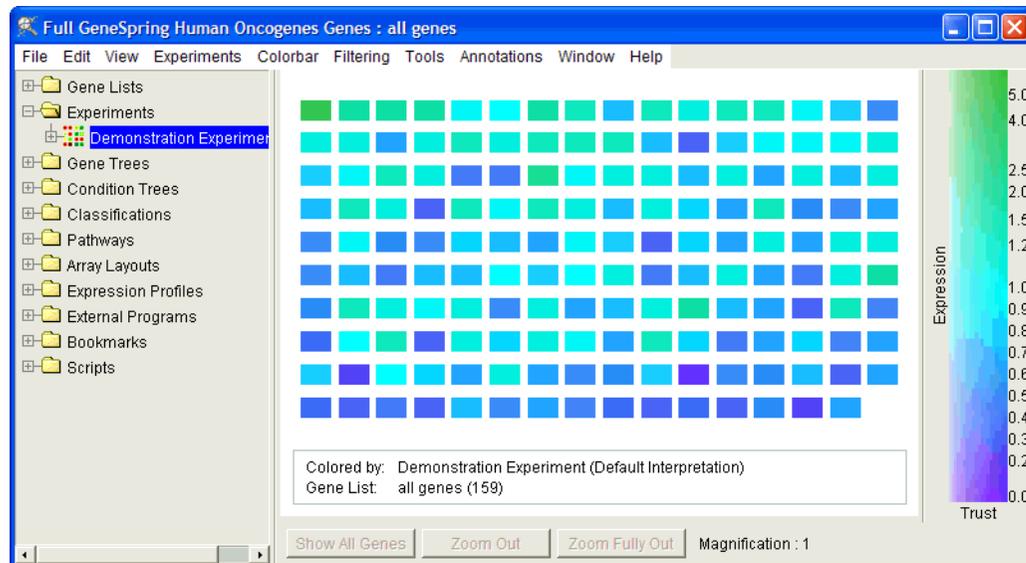


Figure 4-20 The Blocks View

Blocks View Display Options

The following display options are available in blocks view:

- **Features**—The available options for this view are listed below.
- **Coloring**—See “Color” on page 4-30.
- **Legend**—See “Legend” on page 4-28

Features

The Features panel of the display options window contains a column of check-boxes that allow you to toggle on or off certain items in the genome browser.

- **Color by all conditions** – Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the genome browser.
- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Graph View

The Graph view allows you to visualize one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a line. To choose the Graph view, select **View > Graph**.

Note: Genes with no data cannot be displayed in this view.

The Graph option consists of two views: the continuous graph view and the histogram view, which appears if the experiment being displayed contains any non-continuous parameters.

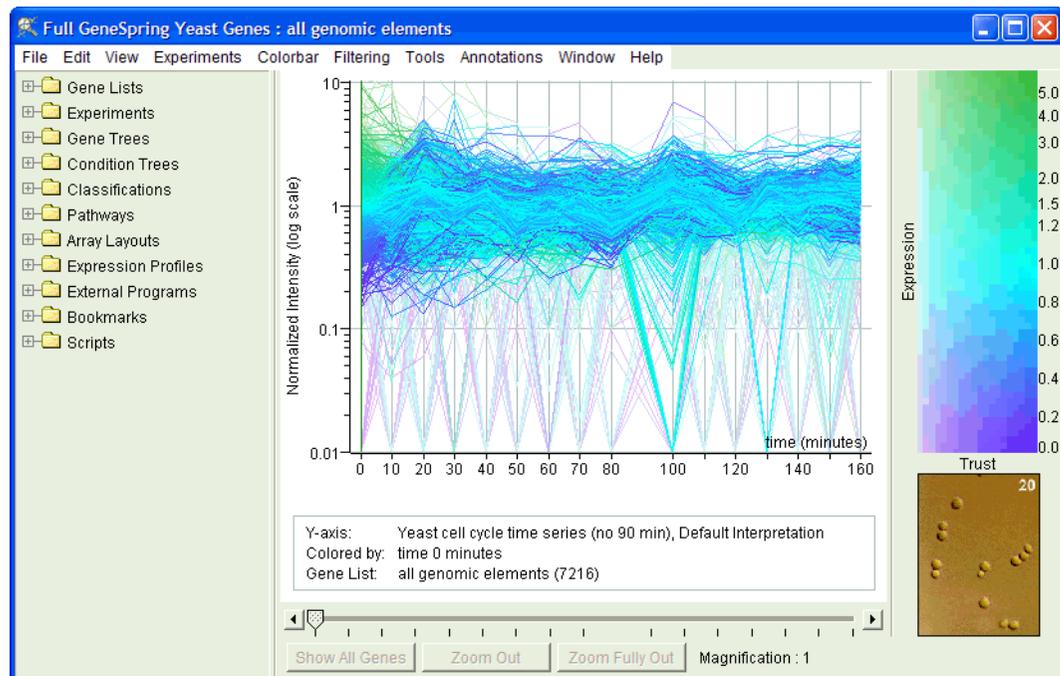


Figure 4-21 The Graph view

The figure above shows the genes in the “all genes” list in Graph view. The gene in white has been selected; its name appears in the legend, after the name of the gene list.

Graph View Display Options

The following display options are available in graph view:

- **Vertical Axis**—See “The Vertical Axis” on page 4-27.
- **Features**—The available options for this view are listed below.
- **Lines to Graph**—The available options for this view are listed below.
- **Coloring**—See “Color” on page 4-30.
- **Error Bars**—See “Error Bars” on page 4-28
- **Legend**—See “Legend” on page 4-28

Lines to Graph

You have the option to draw grid lines to help distinguish distinct groups of data points. To modify the use of lines:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**.
2. Click the **Lines to Graph** tab.
3. To see a grid inside the plot area, you can have lines drawn at the major and minor tick intervals of each axis. Check any of the the **Major Intervals/Minor Intervals** boxes and click **Apply** to view your data with grid lines.

Features

The Features panel of the display options window contains a column of check-boxes that allow you to toggle on or off certain items in the genome browser.

- **Show Experiment Name**—Displays the name of the current experiment in the upper right-hand corner of the genome browser.
- **Show Horizontal Axis Label**—Displays the parameter that is graphed on the horizontal axis.
- **Show Vertical Axis Label**—Displays the parameter that is graphed on the vertical axis.
- **Label Vertical Axis on Side**—Displays the vertical axis label vertically. If this is unchecked the vertical axis label sits to the right of the top of the vertical axis.
- **Show Condition Line**—Displays the vertical bar that can be moved with the condition slider.
- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Bar Graph View

The Bar Graph view allows you to visualize one experiment or a set of experiments by plotting the relative expression of each gene against experimental parameters, such as time or drug concentration. Each gene is represented as a vertical bar. To switch to Bar Graph view, select **View > Bar Graph**.

Note: Genes with no data cannot be displayed in this view.

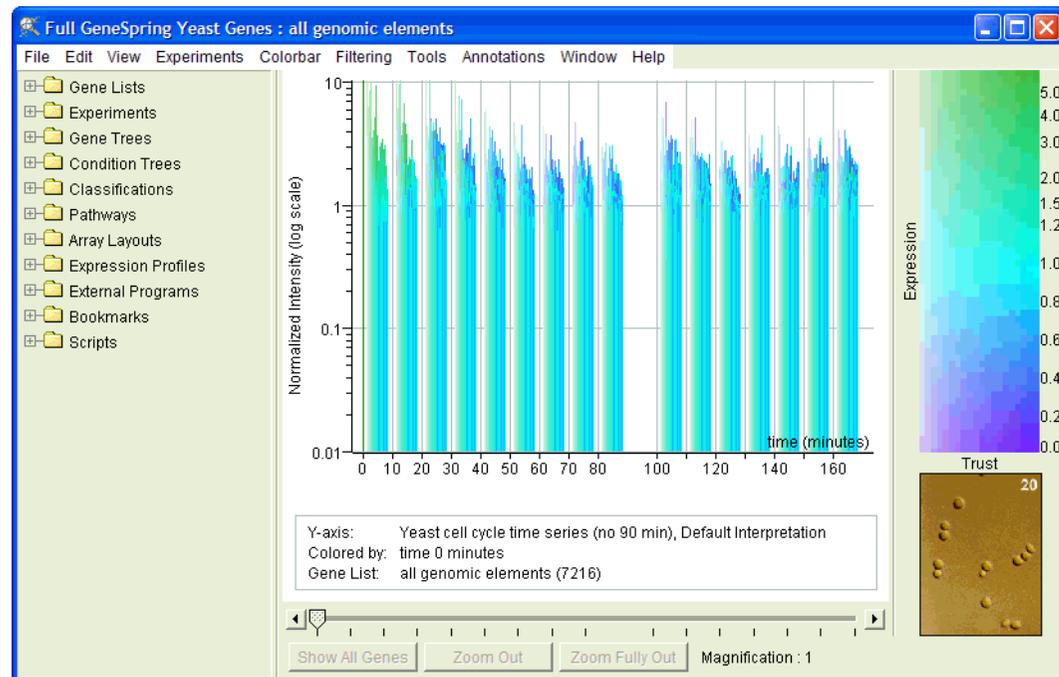


Figure 4-22 The Bar Graph view

The figure above shows a Yeast cell cycle time series in Bar Graph view.

Bar Graph View Display Options

In bar graph view, the following display options are available:

- **Vertical Axis**—See “The Vertical Axis” on page 4-27.
- **Features**—The available options for this view are listed below.
- **Lines to Graph**—The available options for this view are listed below.
- **Coloring**—See “Color” on page 4-30.
- **Legend**—See “Legend” on page 4-28

Lines to Graph

GeneSpring provides the option to draw grid lines to help distinguish distinct groups of data points. To modify the use of lines:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**.

2. Click the **Lines to Graph** tab.
3. To see a grid inside the plot area, you can have lines drawn at the major and minor tick intervals of each axis. Check any of the the **Major Intervals/Minor Intervals** boxes and click **Apply** to view your data with grid lines.

Features

The Features panel of the display options window contains a column of check-boxes that allow you to toggle on or off certain items in the genome browser.

- **Show Horizontal Axis Label**—Displays the parameter that is graphed on the horizontal axis.
- **Show Vertical Axis Label**—Displays the parameter that is graphed on the vertical axis.
- **Label Vertical Axis on Side**—Displays the vertical axis label vertically. If this is unchecked the vertical axis label sits to the right of the top of the vertical axis.
- **Show Condition Line**—Displays the vertical bar that can be moved with the condition slider.
- **3D Look**—Places the bars on a diagonal line so as to imply that genes in each condition are stacked in rows perpendicular to the horizontal axis.
- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Physical Position View

The Physical Position display allows you to see an experiment or a set of experiments by organizing the genes according to their physical position (when the gene loci are known and loaded into GeneSpring) within the DNA sequence of the organism. Select **View > Physical Position**. The Physical Position view works for any organism whose mapping data is at least partially available. An illustration of what Physical Position View looks like for humans is given in Figure 4-24. For organisms already sequenced, the physical position views looks more like yeast (illustrated in Figure 4-23).

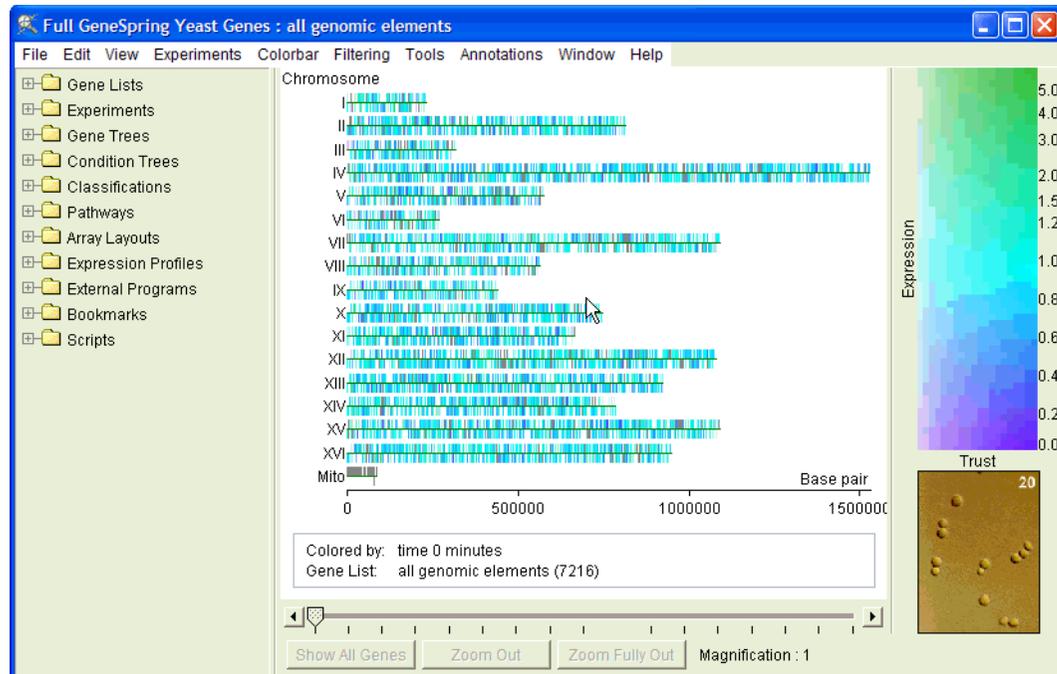


Figure 4-23 The Physical Position view

At greater magnification, you can see the base pairs.

Physical Position View

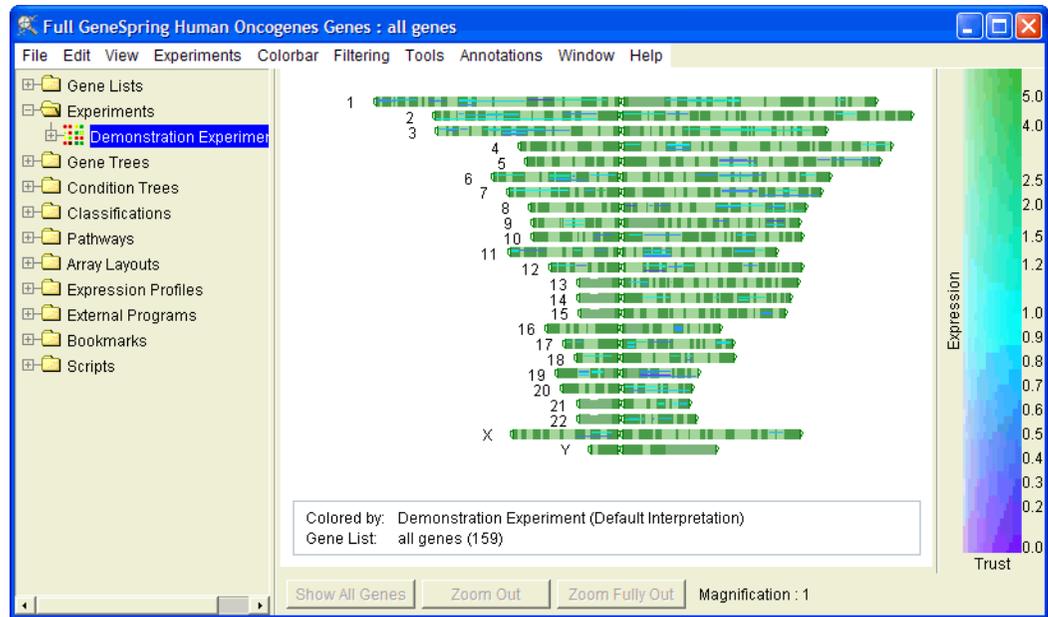


Figure 4-24 Physical position view for a human genome

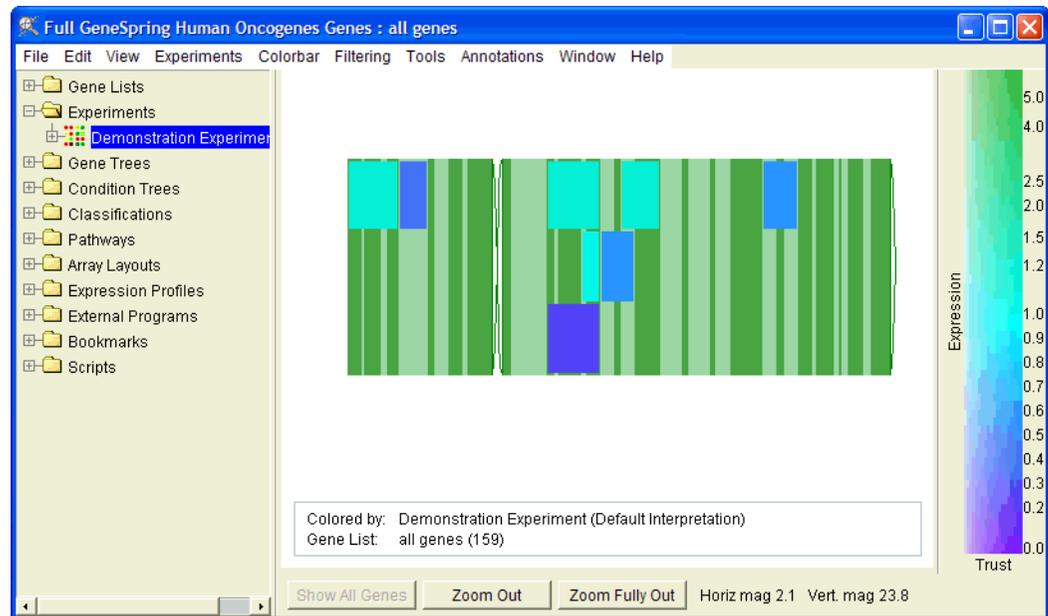


Figure 4-25 Zooming in for a closer look at chromosome 12

At high magnifications the labels associated with the chromosome's cytogenetic bands are visible.

The Load Sequence command

In GeneSpring versions 4.0 and later, sequence information is loaded by default if it is available. If you have an old version of GeneSpring and cannot update it (see "Updating GeneSpring" on page 1-4), follow these directions.

The Load Sequence command is applicable only for sequenced organisms. Load the nucleic acid sequence to magnify a section of the physical position view until the nucleic acid sequence is displayed. Loading the sequence also allows you to take advantage of GeneSpring's other sequence-based features such as **Tools > Find Potential Regulatory Sequences**.

You can load the nucleic acid sequence in a number of ways.

Method 1 (takes immediate effect)

1. Right-click while the cursor is in the black genome browser. A menu appears.
2. Select **Options > Load Sequence**.

A window saying *Please wait while nucleic acid sequence is loaded* appears. After the loading is complete it is possible to zoom in and see the nucleic acid sequence of a particular gene.

The sequence is shown in the magnified genes. However, this information is not saved, so when you exit and re-open GeneSpring you must reload the nucleic acid sequence.

If you would like the sequences to always be readily available, you must change the defaults through the Preferences window. You may choose to make the load sequence feature automatically load with the program. Again, note that this applies to version 4.0 and earlier.

Method 2 (takes effect in your next GeneSpring session)

1. Select **Edit > Preferences**. The GeneSpring Preferences window appears.
2. Select **Data Files** from the pull-down at the top of the window.
3. Select the **Load Sequence** checkbox.
4. Click **OK** at the bottom of the window.
5. Close and restart GeneSpring. (Or, you can select **File > New Window**.)

Changing the defaults in the Preferences window does not initiate the load sequence feature in your current session, but it does change future initial loading practices. The nucleic acid sequence can also be loaded as a side effect of using **Tools > Find Regulatory Sequences**. For more information on this particular feature, see "Regulatory Sequences" on page 6-18.

Physical Position Display Options

In the Physical Position view, the following display options are available:

- **Features**—The available options for this view are listed below.
- **Coloring**—See "Color" on page 4-30.
- **Legend**—See "Legend" on page 4-28

The Features panel of the display options window contains a column of check-boxes that allow you to toggle on or off certain items in the genome browser.

Physical Position View

- **Show Chromosome Label**—Displays the word “Chromosome” next to the chromosome names or numbers.
- **Show Chromosome Label on Side**—Displays the word “chromosome” vertically beside the chromosome names or numbers.
- **Show Base Pair Label**—Displays the words “Base Pair” next to the axis representing the sequence location.
- **Show ORF direction**—Places genes above or below the chromosomes depending on the direction they are transcribed. Genes on the top of the line are transcribed from left to right. Leaving this option unchecked places all of the genes on top of the chromosome lines.
- **Show Just One Strand of Bases**—Displays only the bases on the Watson strand (when the genome browser is zoomed-in enough to display them).
- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Scatter Plot View

The Scatter Plot view is useful for examining the expression levels of genes in two distinct conditions, samples, or normalization schemes. For instance, you can use the scatter plot to identify genes that are differentially expressed in one sample versus another. A scatter plot can also be used to compare two values associated with genes in two gene lists. Such associated values might include the relative contribution of principal components as determined from principal components analysis, or two similarity scores from the Find Similar function in the Gene Inspector.

Note: Genes with no data cannot be displayed in this view.

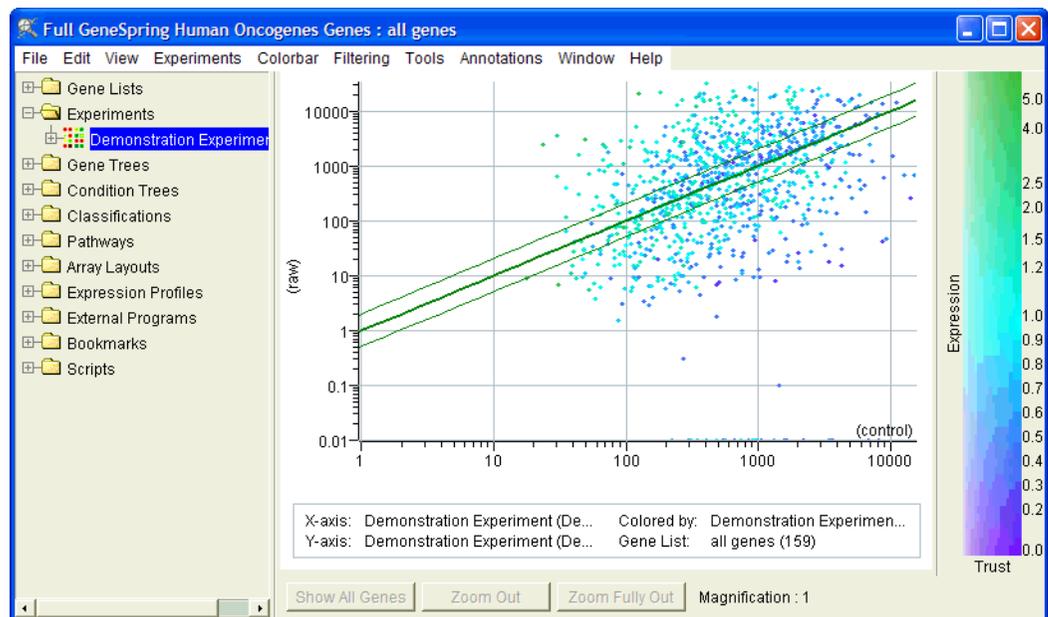


Figure 4-26 The Scatter Plot view

In the scatter plot above, each '+' symbol represents a gene. The vertical position of each gene represents its expression level in the current condition, and the horizontal position represents its control strength (in this case, the median expression level of this gene in all conditions). Genes that fall above the diagonal are overexpressed and genes that fall below the diagonal are underexpressed as compared to their median expression level over the course of the experiment.

Viewing a Scatter Plot

To view a scatter plot select the **View > Scatter Plot** option. The scatter plot view is the most flexible in its ability to customize the way that data are displayed.

Scatter Plot Display Options

The following display options are available for this view:

- **Vertical Axis**—The available options are listed below.
- **Horizontal Axis**—The available options are listed below.

- **Features**—The available options are listed below.
- **Lines to Graph**—The available options are listed below.
- **Coloring**—The available options are listed below.
- **Error Bars**—See “Error Bars” on page 4-28
- **Legend**—See “Legend” on page 4-28

Vertical/Horizontal Axes

The most critical option to set is the type of data that is displayed on the two axes. To modify the function, as well as the appearance of the axes:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**. The Scatter Plot Display Options window appears.
2. Click either the Horizontal Axis or Vertical Axis tab.
3. In the Display Options navigator select the gene list, experiment, interpretation, or condition to use on the selected axis.
4. Click the **Horizontal/Vertical Axis Value** pull-down menu. The list of options includes only those that are appropriate for the type of data object you selected.
5. Choose a graph mode for the specified axis. The three options are linear, logarithmic, and fold change. Note that the fold change option is only available if you are looking at normalized data from an interpretation or a condition.
6. To adjust the vertical axes so that all measurements are visible, check the **Scale Axis to Show all Values** box. The upper and lower bounds are adjusted automatically. Alternatively you can manually set the upper and lower bounds to values of your choosing.
7. To automatically choose tick spacings, check the **Automatic Tick Spacing on Axis** box. To set the tick spacings manually, leave this box unchecked and enter the major tick interval as well as the number of minor ticks. For more information about setting tick spacings, see “The Vertical Axis” on page 4-27.

Adding Lines

You have the option to draw lines that help distinguish distinct groups of data points. Although these lines can represent many types of data thresholds, they are generically called fold change lines. These fold lines are valuable because you can select points that lie above or below them by right clicking in the appropriate position in the genome browser. In addition to fold lines, you can add lines to the origin of each axis as well as draw a line of best fit. To modify the use of lines:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**.
2. Click the Lines to graph tab.
 - To use fold change lines click the **Fold Change Lines** box. If you only want one pair of fold change lines, select the **Set Lines At** radio button and enter a number in the **fold** box. If you would like more than one pair of lines, check the

Set Lines at Multiple Intervals box and list the “fold-values” to view, separated by commas.

- To show a trend in your data check the **Line of Best Fit** box. Note that the regression is performed on the transformed data, and this line is always linear regardless of how the axes are chosen.
- To make the origin of each axis more visible, check the **Lines Through Origin** option.
- To see a grid inside the plot area, you can have lines drawn at the major and minor tick intervals of each axis. Check the **Horizontal/Vertical Grid Lines** checkboxes that to see. The color of these grid lines is represented in the Grid Color box at the bottom of the window. To modify the grid color, click **Change . . .**

Changing Labels and Features

The scatter plot view also allows you to change the appearance of data points and data labels. To modify these features:

1. Select **View > Display Options . . .** or right click anywhere in the genome browser and select **Display Options**.
2. Click the Features tab.
3. To modify the size and shape of the points choose from among the options in the **Style** and **Size** pull-down menus.
4. There are five options for labeling the plot:
 - **Show Gene Names**—Displays the name of each gene to the lower right of each point. These names become unreadable if more than ~100 genes are visible in the current gene list and magnification.
 - **Show Horizontal Axis Label**—Displays the parameter that is graphed on the horizontal axis.
 - **Show Vertical Axis Label**—Displays the parameter that is graphed on the vertical axis.
 - **Label Vertical Axis on Side**—Displays the vertical axis label vertically. If this is unchecked the vertical axis label sits to the right of the top of the vertical axis.
 - **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Coloring

Coloring in the scatter plot view is more complicated than in other views because the color of each gene can be derived from the data in either axis. In other views, the color of the gene is usually linked to the data plotted on the vertical axis. In addition, the scatter plot allows you to color genes based on a third experiment or condition that is not plotted on *either* axis. To modify the way data points are colored:

1. Select **View > Display Options . . .** or right click anywhere in the genome browser and select **Display Options**.

Scatter Plot View

2. Click the Coloring tab.
3. Select the type of data that is to be used for coloring from the **Color data points by** pull-down menu. For more information about the types of data that are available for coloring, see “Color” on page 4-30.
4. Choose from the **Use the expression levels in the following experiment or condition** radio-buttons to select the axis to be used for coloring. Note that only axes which represent experiments, interpretations or conditions are available for coloring. To color genes by an experiment that is not represented by either axis click **Other Experiment**, select an experiment in the navigator, and click **Set Experiment>>**.

3D Scatter Plot View

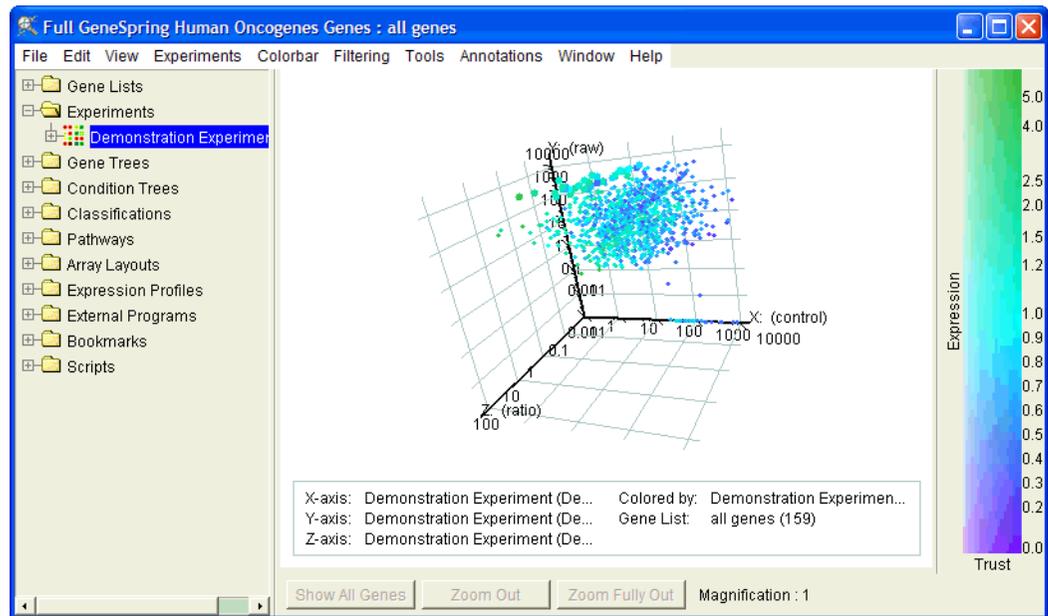


Figure 4-27 The 3-D Scatter Plot View

In the 3D scatter plot above, each dot represents a gene. The vertical position of each gene represents its expression level in the current condition, and the horizontal position represents its control strength (in this case, the median expression level of this gene in all conditions).

Pressing the **x**, **y**, or **z** keys rotates the graph on the specified axis. Hold down the **Shift** key to speed this rotation. Hold down the **Alt** key to reverse the direction of rotation.

Note: Genes with no data cannot be displayed in this view.

3D Scatter Plot Display Options

The following display options are available for this view:

- **X Axis**—See “X, Y, and Z Axes” on page 4-50
- **Y Axis**—See “X, Y, and Z Axes” on page 4-50
- **Z Axis**—See “X, Y, and Z Axes” on page 4-50
- **Features**—See “Changing Labels and Features” on page 4-47
- **Lines to Graph**—See “Adding Lines” on page 4-46
- **Coloring**—See “Coloring” on page 4-47.
- **Error Bars**—See “Error Bars” on page 4-28
- **Legend**—See “Legend” on page 4-28

X, Y, and Z Axes

The most critical option to set is the type of data that is displayed on the three axes. To modify the function, as well as the appearance of the axes:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**. The 3D Scatter Plot Display Options window appears.
2. Click the **X Axis**, **Y Axis**, or **Z Axis** tab.
3. In the Display Options navigator select the gene list, experiment, interpretation, or condition to use on the selected axis.
4. Click the **X/Y/Z Axis Value** pull-down menu. The list of options includes only those that are appropriate for the type of data object you selected.
5. Choose a graph mode for the specified axis. The three options are linear, logarithmic, and fold change. Note that the fold change option is only available if you are looking at normalized data from an interpretation or a condition.
6. To adjust the axes so that all measurements are visible, check the **Scale Axis to Show all Values** box. The upper and lower bounds are adjusted automatically. Alternatively you can manually set the upper and lower bounds to values of your choosing.
7. To automatically choose tick spacings, check the **Automatic Tick Spacing on Axis** box. To set the tick spacings manually, leave this box unchecked and enter the major tick interval as well as the number of minor ticks. For more information about setting tick spacings, see “The Vertical Axis” on page 4-27.

Adding Lines

You have the option to draw lines that help distinguish distinct groups of data points. Although these lines can represent many types of data thresholds, they are generically called fold change lines. These fold lines are valuable because you can select points that lie above or below them by right clicking in the appropriate position in the genome browser. In addition to fold lines, you can add lines to the origin of each axis as well as draw a line of best fit. To modify the use of lines:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**.
2. Click the Lines to Graph tab.
3. To see a grid inside the plot area, you can have lines drawn at the major and minor tick intervals of each axis. Check the **X/Y/Z Axis Grid Lines** checkboxes that to see. The color of these grid lines is represented in the Grid Color box at the bottom of the window. To modify the grid color, click **Change...**

Changing Labels and Features

The scatter plot view also allows you to change the appearance of data points and data labels. To modify these features:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**.
2. Click the Features tab.
3. To modify the size and shape of the points choose from among the options in the **Style** and **Size** pull-down menus.
4. There are five options for labeling the plot:
 - **Show Gene Names**—Displays the name of each gene to the lower right of each point. These names become unreadable if more than ~100 genes are visible in the current gene list and magnification.
 - **Show X Axis Label**—Displays the parameter that is graphed on the X axis.
 - **Show Y Axis Label**—Displays the parameter that is graphed on the Y axis.
 - **Show Z Axis Label**—Displays the parameter that is graphed on the Z axis.
 - **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Coloring

Coloring in the scatter plot view is more complicated than in other views because the color of each gene can be derived from the data in any axis. In other views, the color of the gene is usually linked to the data plotted on the vertical axis. In addition, the scatter plot allows you to color genes based on a fourth experiment or condition that is not plotted on *either* axis. To modify the way data points are colored:

1. Select **View > Display Options...** or right click anywhere in the genome browser and select **Display Options**.
2. Click the Coloring tab.
3. Select the type of data that is to be used for coloring from the **Color data points by** pull-down menu. For more information about the types of data that are available for coloring, see “Color” on page 4-30.
4. Choose from the **Use the expression levels in the following experiment or condition** radio-buttons to select the axis to be used for coloring. Note that only axes which represent experiments, interpretations or conditions are available for coloring. To color genes by an experiment that is not represented by either axis click **Other Experiment**, select an experiment in the navigator, and click **Set Experiment>>**.

Tree View

The Tree view allows you to view the results of hierarchical clustering in the form of a mock phylogenetic tree, or dendrogram. In such a tree, genes having similar expression patterns are clustered together.

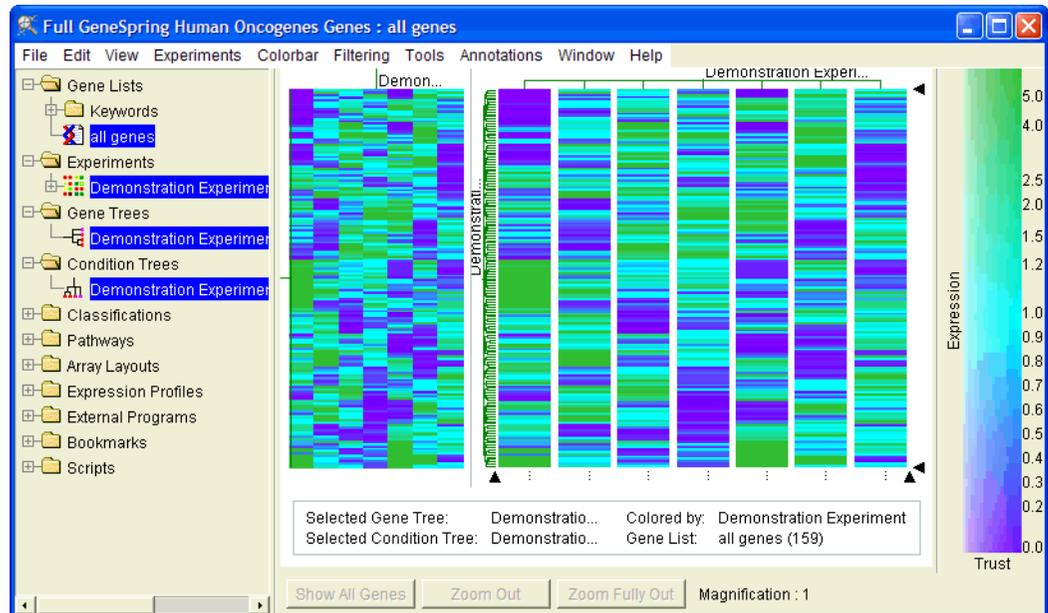


Figure 4-28 Tree View

The genome browser above is displaying a gene tree. The genes are the columns of colored rectangles to the right of the tree structure, displayed in green. Similarly colored genes tend to be clustered together.

Viewing a Tree

1. From the navigator, open the Gene Trees or the Condition Trees folder.
2. Select a tree. If there are no trees available for viewing you must create one. See “Gene Tree Clustering Options” on page 7-5.

Selecting and Viewing Subtrees

A single green line ending in a gene is a branch of the gene tree. Each bar crossing a set of branches forms a node of the intersecting branches. The distance from gene X to the node connecting it to gene Y indicates how closely the genes X and Y are correlated. The shorter the distance, the higher the correlation is. Select any node by clicking over its intersection with your cursor. All the genes associated with that node changes to your *selected* color.

- To create a new tree from a node of a larger tree, select a node as described above, then right-click in the genome browser and select **Make Subtree** from the pop-up menu.

- To make a gene list from a subtree, select a node as described above, then right-click in the genome browser and select **Make List from Subtree**.

Eisen-Like Tree View

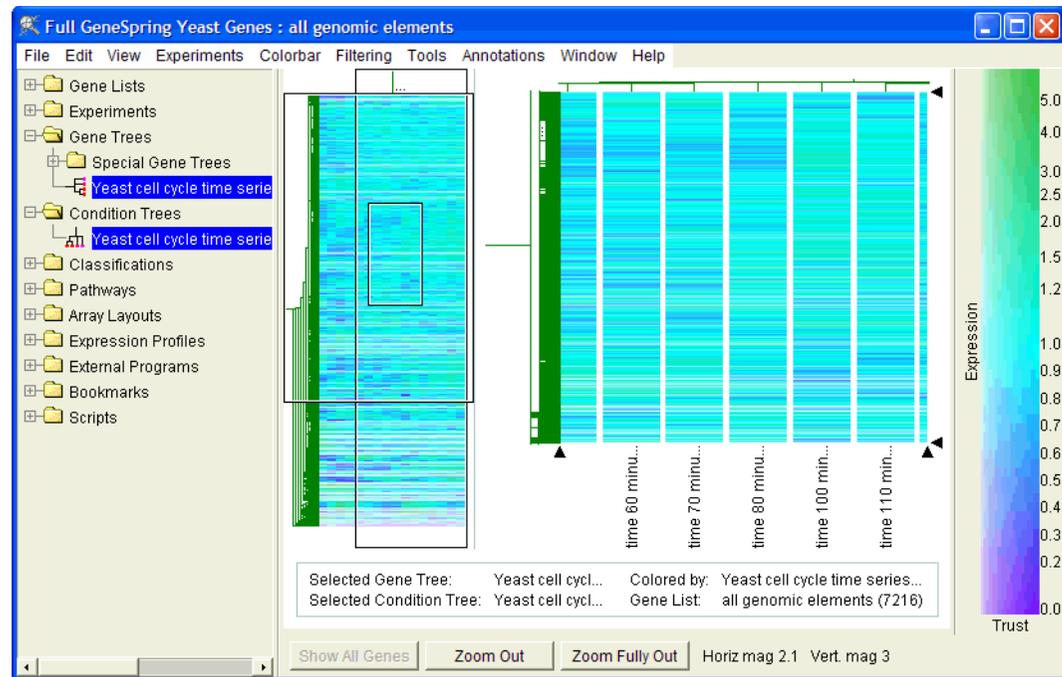


Figure 4-29 Eisen-like Tree View

In GeneSpring 6.0, subtrees can be viewed in an Eisen-like format. There are two ways to select a subtree for this view:

In GeNetViewer, subtrees can be viewed in an Eisen-like format. There are two ways to select a subtree for this view:

- Double-click on the node defining the desired subtree
- Right-click on the node defining the desired subtree and select **Display Sub-tree**

Double-clicking a node changes the selected subtree. You can double-click on nodes both in the thumbnail and in the main part of the window.

Three marquees may be shown on the thumbnail:

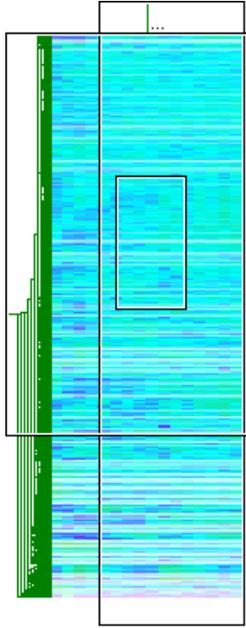


Figure 4-30 Marquees in the Tree View

The area displayed within all of these marquees is displayed to the right of the thumbnail. To change the size of the thumbnail, use the drag arrows below and to the right of the tree. To enable these drag arrows:

1. Right-click in the browser and select **Display Options**.
2. Click the Features tab.
3. Check the **Display Drag Arrows** box.
4. Click **OK** to return to the tree view.

Navigating Subtrees

- To navigate through subtrees, right-click on any node and select **Display Sub-tree**.
- To view the tree immediately above the one selected, right-click anywhere and choose **Display Parent of Sub-tree**.
- To return to the top and view the entire tree, right-click anywhere and select **Display Entire Tree**. This returns you to the default GeneSpring tree view.
- Double-clicking a terminal branch (a line indicating only one condition or gene) invokes either the Condition Inspector or the Gene Inspector, depending on the branch.

Keyboard commands for the right-hand or top tree (usually the Condition tree)

- Alt+left arrow—Jump to the sibling to the left of the selected node
- Alt+right arrow—Jump to the sibling to the right of the selected node
- Alt+up arrow—Jump to the parent of the selected node

- Alt+down arrow—Jump to the first child of the selected node (counting from left to right)

Keyboard commands for the left-hand tree (usually the Gene Tree)

- Ctrl+left arrow—Jump to the parent of the selected node
- Ctrl+right arrow—Jump to the first child of the selected node (counting from top to bottom)
- Ctrl+up arrow—Jump to the sibling directly above the selected node
- Ctrl+down arrow—Jump to the sibling directly below the selected node

Magnifying Trees

Magnification in the Tree View is not quite the same as in the other views due to the need to keep the genes in the view along with the immediate tree branches. Zooming in by dragging a rectangle with the cursor usually produces a magnified view that contains more elements than were in the selected area. The amount of magnification is visible in the parameter specification area just below the genome browser.

Use arrow keys to pan the screen while zoomed. Panning never takes you outside the bounds of the selected subtree (if any).

When a subtree is selected, clicking **Zoom Fully Out** displays the entire subtree, not the entire tree. To return to the top level, right-click anywhere and select **View Entire Tree**.

You cannot zoom in on the thumbnail in the Eisen-like view.

Viewing Nodes

After clustering the genes according to their expression patterns, all known lists are checked against all subtrees of the new gene tree, to assign names to the tree nodes where possible. These labels are taken from the gene lists in the standard lists.

Place your cursor as close as possible to a label or intersection to view the text. When the cursor pauses over an intersection, a label appears. It disappears when the cursor is moved.

All of the branches intersecting to form a node constitute the subtree defined by that node. A label such as “ribosome [15.1]” means the subtree from that node has a lot in common with the genes in the “ribosome” list. The numbers in square brackets are a measure of statistical significance. The higher the value, the more significant the comparison is.

The comparisons between the lists and the subtrees are not looking for exact matches, but rather statistically significant overlaps, which may include subsets and supersets. When there is enough space on the screen, a label, if one exists, is displayed along the top (horizontal bar) of the subtree. Otherwise, when there is space, a “...” is displayed. An “&” symbol after a list name indicates the subtree is statistically similar to more than one list, all of whom, when there is enough room, are displayed as labels along the top of the subtree.

To take a screen shot that includes the label, hover your cursor over the node, take the screen shot when the label appears. For most Windows applications, the cursor is not visi-

ble, just the label. For more information about screen shots, see “Saving Pictures and Printing” on page 9-4.

Viewing Gene Names in Trees

You can magnify the tree until the names are visible along the edge of the genes.

1. Place your cursor anywhere over the group of genes to view the gene name. When the cursor pauses over a gene, a label appears. It disappears when the cursor is moved.
2. Click once and that gene becomes the selected gene. The name of the selected gene appears in the upper right corner of the genome browser.

Viewing Parameters in Trees

For most experiments, each measurement was taken under certain conditions. These conditions are listed in the far right side of the tree view. If one of the parameters has been designated as a continuous parameter, it is shown directly beneath the genome browser.

Tree Display Options

The following display options are available in tree view:

- **Gene Tree/Condition Tree**—The available options for this view are listed below.
- **Features**—The available options for this view are listed below.
- **Coloring**—See “Color” on page 4-30.
- **Legend**—See “Legend” on page 4-28

To modify the appearance of your tree, select **View > Display Options....**

Gene Tree/Condition Tree Tab

The Display Options window includes a Gene Tree or Condition Tree tab, depending on whether you have selected a gene tree, condition tree, or both. This tab contains the following options:

- **Draw Genes Horizontally** – Orients your tree so that the genes appear as horizontal bars on the right extending from tree branches on the left.
- **Show Tree Structure**—Specifies whether to show or hide the tree structure.
- **Show Gene Name Labels**—If genes are displayed vertically, shows the name of each gene to its right if there is space. You must be at a very high magnification for these labels to be visible. This option is available only if a gene tree is selected.
- **Show Tree Annotation Labels**—Displays annotations for tree nodes if they are available and space permits. This option is available only if a gene tree is selected.
- **Show Experiment Condition Labels**—Displays experiment condition labels if they are available and space permits. This option is available only if an experiment or condition tree is selected.
- **Color Branches by Classification**—Color the tree branches based on classification. This option is available only if a gene tree is selected. To color by classification:

- a. Check the **Color Branches by Classification** box.
- b. Select a classification from the Display Options minibrowser.
- c. Click **Set Classification>>**.
- d. Click **Apply**.

Unclassified genes are displayed using the background color.

- **Color Branches by Experiment Parameter**—Color the tree branches based on experiment parameters. This option is available only if a condition or condition tree is selected. To color by experiment parameters:
 - a. Check the **Color Branches by Experiment Parameter** box.
 - b. Select an experiment from the Display Options minibrowser.
 - c. Click **Set Experiment>>**.
 - d. Select a parameter from the pull-down menu.

To display a row of blocks at the bottom of the condition tree indicating their classification, select the **Show Coloring Blocks** radio button.
 - e. Click **Apply**.

Features Tab

The Display Options window also includes a Features tab with the following options for reorganizing your tree:

- **Color by all conditions**—Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the genome browser.
- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.
- **Place gaps between Heatmap Tiles**—Uncheck this option to remove gaps between tiles.
- **Display Navigational Tree**—Specifies whether to display the navigational tree for the Eisen-like subtree view on the left or the top of the viewing area.
- **Use Custom Heatmap Borders**—Allows you to customize the amount of screen space dedicated to tree branches and labels. When this option and the Show Drag Arrows option are selected, use the drag arrows in the genome browser window to make adjustments.
- **Show Drag Arrows**—Displays arrows used for changing the size of the area dedicated to tree branches. This affects both the thumbnail and the displayed subtree in the Eisen-like view.

Ordered List View

Allows you to view a gene list in the order of its associated values. Values are listed in descending order. If you do not have associated values, genes are ordered according to the way they are listed in the master gene table. Vertical lines representing genes are proportional to the gene's associated number.

To view genes in an ordered list, go to **View > Ordered List**. Your list appears in its order.

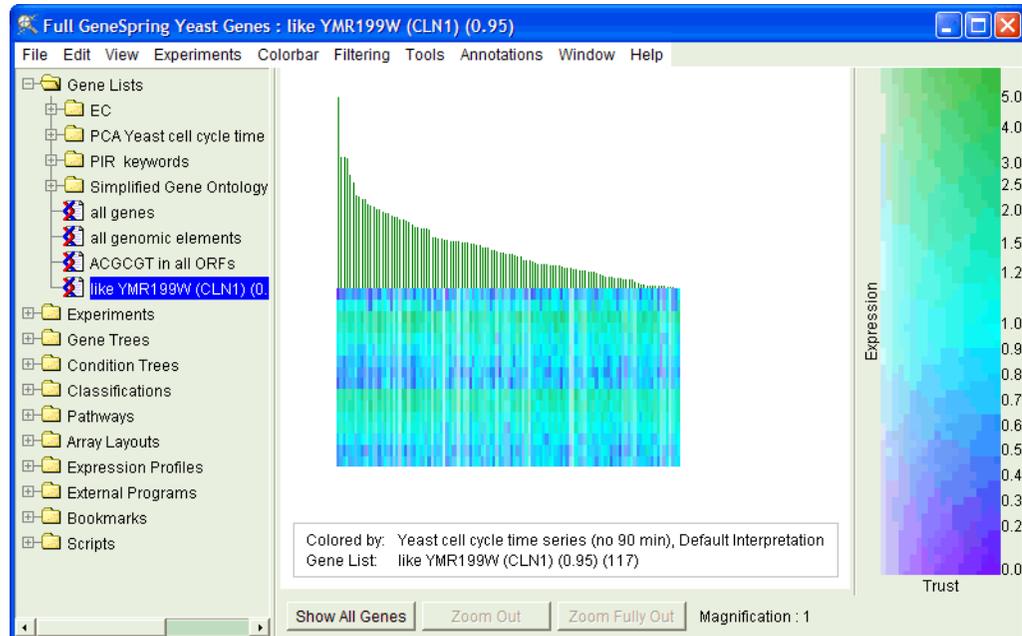


Figure 4-31 Ordered List View

Ordered List Display Options

The following display options are available in ordered list view:

- **Features**—The available options for this view are listed below.
- **Coloring**—See “Color” on page 4-30.
- **Legend**—See “Legend” on page 4-28

To modify the appearance of your tree, select **View > Display Options...** or right-click anywhere in the genome browser and select **Display Options...**. The Display Options window includes a Features panel with the following options:

- **Show Associated Value** – When the view is zoomed, so as to enlarge the tops of the lines, selecting this options displays the numerical value associated with each line.
- **Color by all conditions** – Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the genome browser.

- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

The Coloring panel allows you to modify the way color is used to represent different types of data. For more information see “Color” on page 4-30.

Array Layout View

The Array Layout view produces a synthetic picture of the arrays used in the current experiment. This view is useful in identifying arrays that display local shifts in intensity due to problems in probe deposition, hybridization, washing, or blocking. To use this view you must first create an array layout file (see “Layout Parameters” on page 2-12).

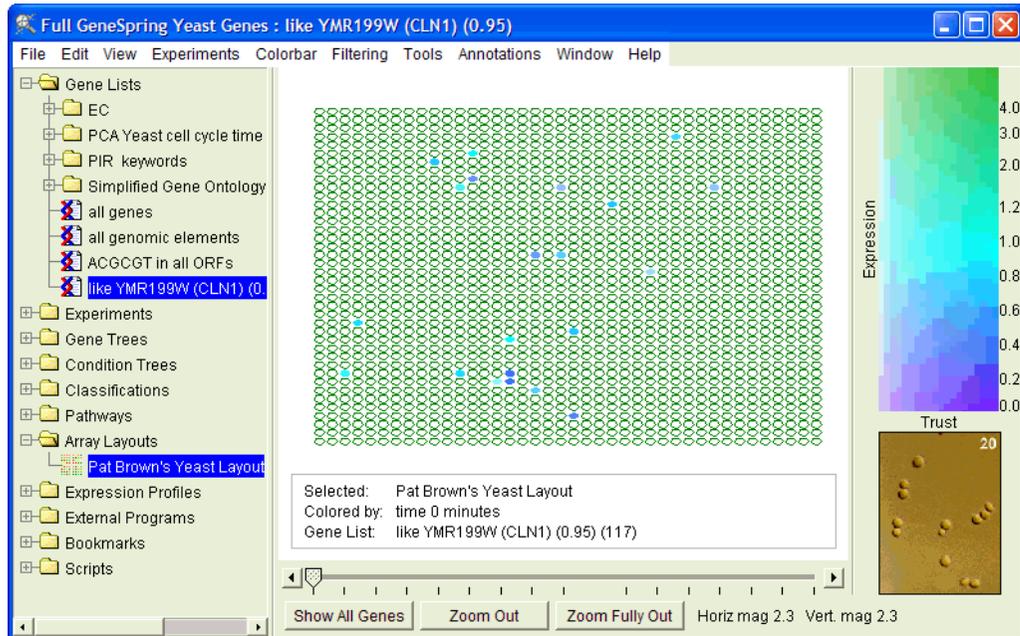


Figure 4-32 The Array view

In Figure 4-32, each solid circle represents an oligonucleotide on the array. If you zoom in, the gene names become visible. Circles are numbered from left to right and top to bottom. For example, a 3X3 array is:

```

1 2 3
4 5 6
7 8 9

```

Viewing an Array Layout

1. Select the **View > Array Layout** option.
2. Select an array from the navigator.

Array Layout Display Options

The following display options are available in graph view:

- **Features**—The available options for this view are listed below.
- **Coloring**—See “Color” on page 4-30.
- **Legend**—See “Legend” on page 4-28

The only feature that can be changed is the **Show unclassified Group When Splitting the Window** option within the features panel. When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Pathway View

The Pathway view lets you display and place genes on an imported .gif or .jpeg image. For information on downloading and importing pathways, see “Pathways” on page 6-16.

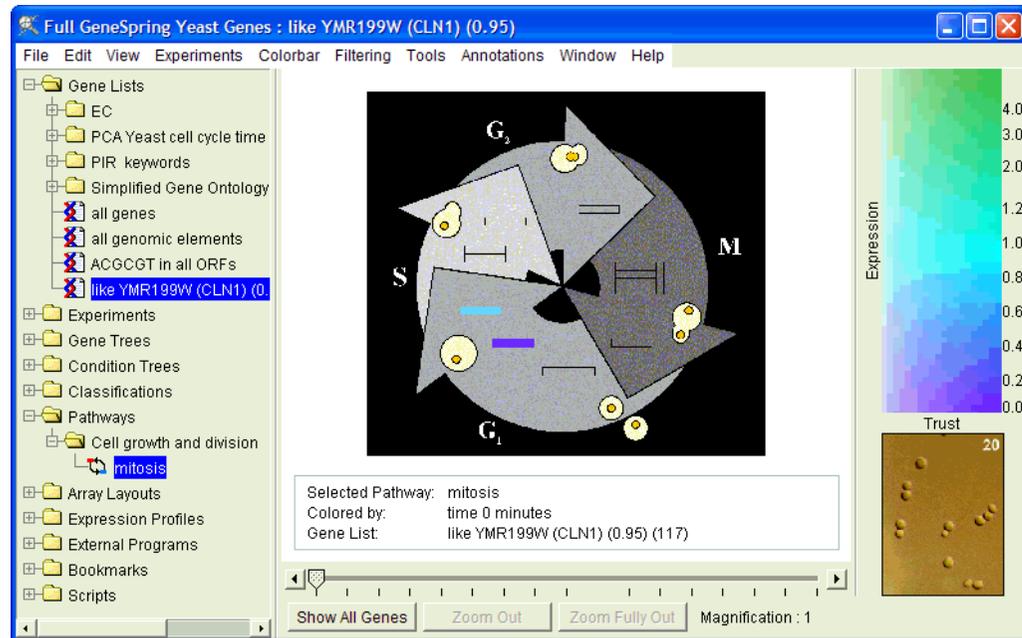


Figure 4-33 The Pathway view

Viewing a Pathway

1. Select a pathway from the Pathways folder in the navigator. (You must have already created a pathway. See “Pathways” on page 6-16.)
2. Select a gene list. If a pathway contains a gene on a selected gene list, then the gene is colored according to its expression level in the selected experiment.

See the example of the mitosis pathway in Figure 4-33.

- To add a gene to the pathway, hold **Ctrl** and drag mouse over the desired placement area. Type a gene name or keyword. If a keyword is used, select the gene from the resulting list.
- To delete a gene from the pathway, right-click over the gene and select **Delete Pathway Element**.

Zooming, coloration, movement and the **Find Genes Which Could Fit Here** features work in this view. **Find Genes Which Could Fit Here** suggests genes that might be appropriate in certain areas of the picture. See “Pathways” on page 6-16 for more details.

Pathway Display Options

The following display options are available in graph view:

- **Features**—The available options for this view are listed below.

- **Coloring**—See “Color” on page 4-30.
- **Legend**—See “Legend” on page 4-28

To modify the appearance of your pathway, select **View > Display Options . . .**. The Display Options window includes a Features panel with the following options:

- **Color by all conditions** – Divides the genes into sections representing multiple conditions, so that all conditions in the selected interpretation can be viewed simultaneously. Using this feature disables the condition slider at the bottom of the genome browser.
- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Compare Genes to Genes

The Compare Genes to Genes view allows you to observe the similarity between the expression profiles of two genes in one list or in two separate lists. Genes being compared are listed along respective graph axes. The correlation between any two genes is shown by a colored square at their point of intersection. Strong correlations in expression level are shown by a higher intensity color, weak correlations by a lower intensity color.

Associated values for gene lists are shown as lines extending perpendicularly from each axis. The length of the line represents the magnitude of the associated value. You can view these associated values by zooming in on the ends of the lines.

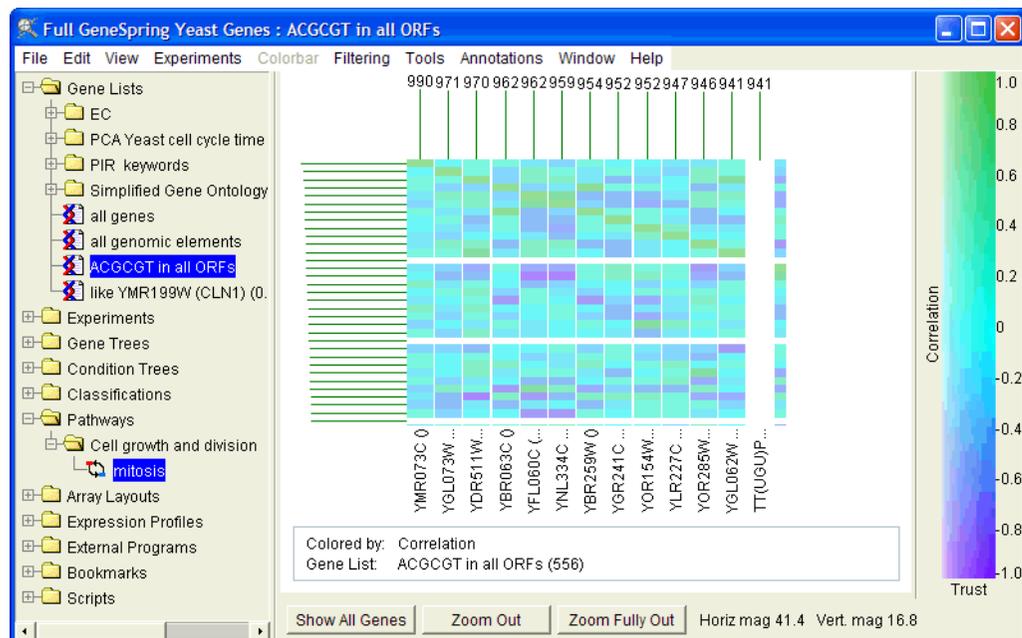


Figure 4-34 Compare Genes to Genes

In the Compare Genes to Genes view, GeneSpring employs a Pearson correlation to measure the pair-wise similarities (see “Pearson Correlation” on page B-3). Note that if you place the same list on both axes, a line of perfect correlation values descends diagonally across the grid. There are no display options in this view.

Viewing Compare Genes to Genes

1. Click the first gene list to compare in the navigator. (Do this before you switch the view type, as large gene lists take a very long time to compare.)
2. Select the **View > Compare Genes to Genes** option. The default display places the selected gene list on both axes.
3. If desired, select a second gene list from the navigator by right-clicking on a gene list and selecting the **Display as Second List** option. To remove this second list, select the **View > Remove Secondary Gene List**.

Graph by Genes View

The Graph by Genes view allows you to visualize an experiment as one line, where each point on the line represents the relative expression of one gene. Select **View > Graph by Genes**.

Note: Genes with no data cannot be displayed in this view.

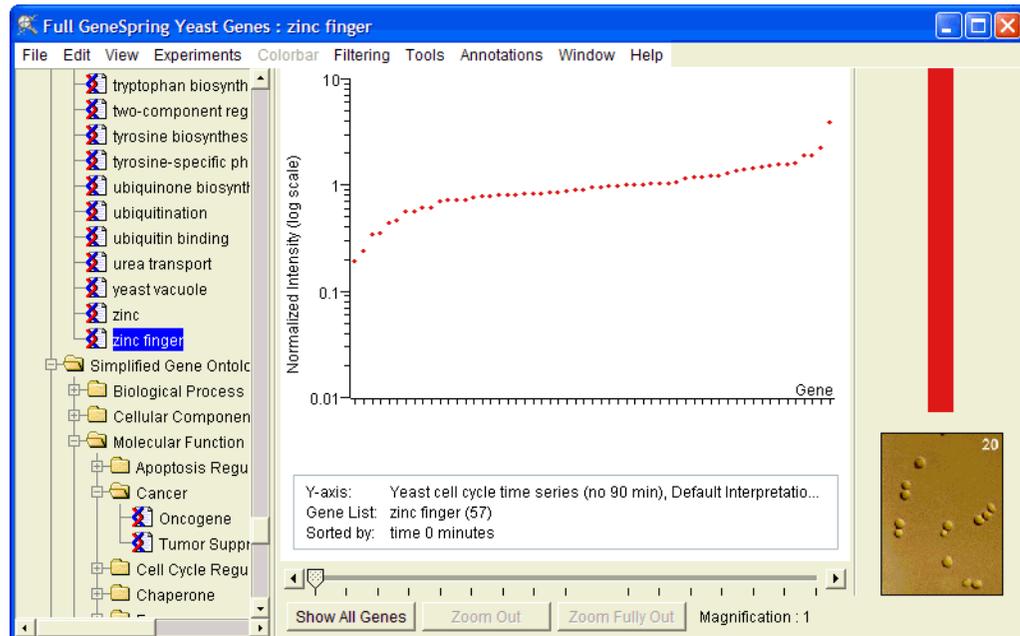


Figure 4-35 The Graph by Genes view, limited to the “Zinc Finger” list

Genes at the top of the selected gene list are displayed at the left end of the experiment line and genes at the bottom of the gene list are displayed at the right end of the experiment line. Generally, your gene lists are ordered so that the associated values appear in descending order. If you do not have associated values, your genes appears in the same order as in the master gene table.

Graph by Genes Display Options

The following display options are available in graph by genes view:

- **Horizontal Axis**—The available options for this view are listed below.
- **Vertical Axis**—See “The Vertical Axis” on page 4-27.
- **Features**—The available options for this view are listed below.
- **Coloring**—“Color” on page 4-30.
- **Legend**—See “Legend” on page 4-28

The Horizontal Axis Tab

- **Sort by Gene List**—Sorts genes in their order in the gene list (by their associated numbers, if they exist; otherwise by their order in the Master Table of Genes).

- **Set Gene List**—Specify the gene list by which to sort genes. This button is only active if the **Sort by Gene List** radio button is selected.
- **Sort by Condition (Normalized Data)**—Sort genes in the order of their normalized values within the selected condition.
- **Sort by Condition (Raw Data)**—Sort genes in the order of their raw values within the selected condition.
- **Sort by Condition (Control Data)**—Sort genes in the order of their control data within the selected condition.
- **Set Condition**—Specify the condition by which to sort genes. This button is only active if one of the **Sort by Condition** radio buttons is selected.

The Features Tab

The Features tab of the display options window contains a column of check-boxes that allow you to toggle on or off certain items in the genome browser. To use them, select **View > Display Options...** or right-click anywhere in the genome browser and select **Display Options...** Click the Features panel and chose from any of the following:

- **Plot Symbol**—Using the **Style** and **Size** pull-down menus, specify the symbol with which to display each gene. If the **Line** option is selected, individual genes cannot be selected in the Genome Browser window.
- **Show Horizontal Axis Label**—Displays the parameter that is graphed on the horizontal axis.
- **Show Vertical Axis Label**—Displays the parameter that is graphed on the vertical axis.
- **Label Vertical Axis on Side**—Displays the vertical axis label vertically. If this is unchecked the vertical axis label sits to the right of the top of the vertical axis.
- **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

View as Spreadsheet

This option allows you to view your data as a spreadsheet. The spreadsheet color scheme and gene list reflect what is showing in the genome browser at the time you activate the new window. The order of the genes is the same as in your master table of genes.

	stage Postnatal, day 0 normalized	stage Adult, day 0 normalized	stage Postnatal, day 7 normalized	stage Embryonic, day 11 normalized	stage Embryonic, day 13 normalized
keratin	1	0.01*	0.725	6.744	1.316
cellubrevin	1	0.87	1.608	4.15	3.107
nestin	0.746	0.238	0.369	3.079	3.886
MAP2	0.973	1	0.976	0.0475	0.594
GAP43	1.177	0.994	1	0.816	1.365
L1	1.68	0.6	1	0.213	0.557
NFL	1.376	0.371	1.596	0.0896	1.019
NFM	1.17	3.731	1.163	0.173	1
NFH	1.071	2.01	1.296	0.189	0.161
synaptophysin	1.13	0.8	1.406	0.206	0.634
нено	0.99	0.78	1.088	0.314	0.797
S100beta	0.957	1.24	1.06	0.0645	0.0127
GFAP	1.381	1.961	3.37	0.01*	0.01*
MOG	33	127.5	132.8	0.01*	0.01*
GAD65	1	0.682	1.001	0.229	0.709
pre-GAD67	1.138	0.655	1	0.105	0.265
GAD67	1.396	0.736	1.194	0.232	0.235
G67180/86	0.752	0.182	0.519	1.043	1.85
G67186	1	0.01*	0.265	0.635	1.779
GAT1	1.031	0.83	1.117	0.332	0.418

Mode: log * Value adjusted due to log interpretation

Figure 4-36 Spreadsheet View

To Copy a Row for Pasting into another Document

1. Click on the row to copy.
2. Right-click on the row and select **Copy**.

To copy the entire spreadsheet, click **Copy All**. Note that if you have any rows selected, you must first click **Clear Selection**.

Locating a Particular Gene

1. Type **Ctrl+F**.
2. Enter the gene name.
3. Click **OK**.

Inspect Found Gene

To bring up the Gene Inspector for your found gene, type **Ctrl+I**.

Condition Scatter Plot

The condition scatter plot displays a fundamentally different type of information than any other view with the possible exception of the condition tree. Unlike other GeneSpring views, each colored point (dot, circle, square, etc.) represents a condition, not a gene.

This view is the most common way to visualize the results of principal components analysis performed on conditions. It is also useful for presenting complex multidimensional data in the context of conditions. For example, a 3D condition scatter plot can be configured to display a principal component score on one axis, a parameter value on a second axis, and the normalized expression level of a given gene on the third axis. A simpler possibility is to plot the expression values for two genes on two axes. Such a plot is useful for demonstrating whether the expression pattern of the genes is correlated or anti-correlated.

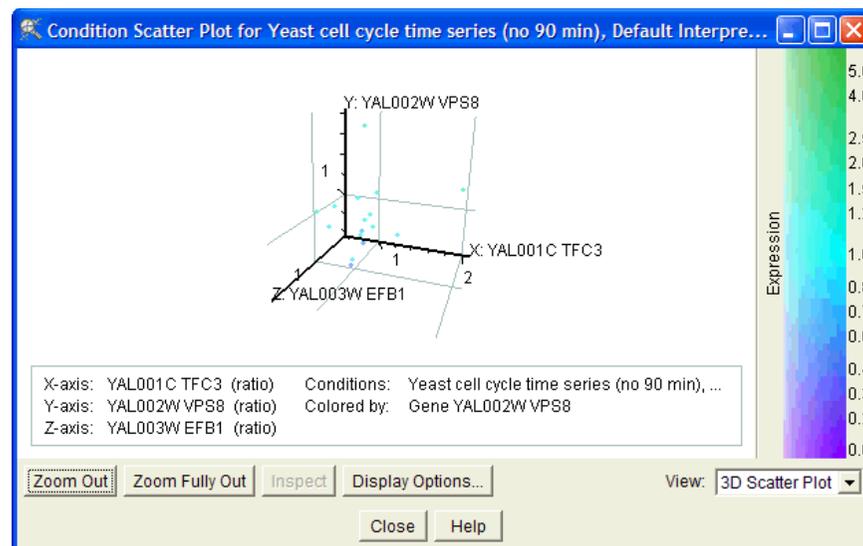


Figure 4-37 The Condition Scatter Plot view

To view the condition scatter plot, select **View > Condition Scatter Plot** in the main GeneSpring window. Unlike most views, the condition scatter plot is displayed in a separate window, which appears when the option is selected. This window also appears when you run a PCA on Conditions analysis. For details, see “PCA on Conditions” on page 7-18. For a 2D view of this plot, select **2D Scatter Plot** from the View menu in the lower right portion of the window.

In the example above, each dot represents a condition. When this window is opened from the main GeneSpring window, the first three parameters (if available) are selected for the axes. If only two parameters are available, the plot is displayed in 2D format. If there are fewer than two parameters, a 3D plot is displayed using the first three genes from the selected gene list.

You cannot select the experiment to be displayed from within this view. To change the experiment being viewed, exit this window, select the desired experiment, and choose **View > Condition Scatter Plot** in the main GeneSpring window.

Pressing the **x**, **y**, or **z** keys rotates the graph on the specified axis. Hold down the **Shift** key to speed this rotation. Hold down the **Alt** key to reverse the direction of rotation.

Condition Scatter Plot Display Options

The following display options are available:

- **X-axis**—See “X, Y, and Z Axes” on page 4-69
- **Y-axis**—See “X, Y, and Z Axes” on page 4-69
- **Z-axis**—See “X, Y, and Z Axes” on page 4-69
- **Lines to Graph**—See “Adding Lines” on page 4-69
- **Features**—See “Changing Labels and Features” on page 4-70
- **Coloring**—See “Coloring” on page 4-70
- **Error Bars**—See “Error Bars” on page 4-28
- **Legend**—See “Legend Options” on page 4-29

X, Y, and Z Axes

The most critical option to set is the type of data that is displayed on the three axes. To modify the function, as well as the appearance of the axes:

1. Click **Display Options . . .** or right click anywhere in the condition scatter plot window and select **Display Options**. The Display Options window appears.
2. Select the **X Axis**, **Y Axis**, or **Z Axis** tab.
3. Specify the type of data to display on the selected axis from the pull-down menu. The available options are Gene, Expression Profile, or Experimental Parameter.
4. Specify the gene from which to use data in the plot. To choose a gene other than the one selected by default, click **Choose Gene . . .** and use the search screen to locate the desired gene.

Note: You can select genes only from the currently active experiment. To work with data from a different experiment, you must exit this screen and select that experiment in the main GeneSpring window before re-opening the Condition Scatter Plot window.

5. Specify the data type. Available options are Control, Raw, or Normalized.
6. Choose a graph mode for the specified axis. Available options are linear, logarithmic, and fold change. Note that the fold change option is only available if you are looking at normalized data from an interpretation or a condition.

Adding Lines

You have the option to draw lines that help distinguish distinct groups of data points. Although these lines can represent many types of data thresholds, they are generically called fold change lines. These fold lines are valuable because you can select points that lie above or below them by right clicking in the appropriate position in the genome browser. In addition to fold lines, you can add lines to the origin of each axis as well as draw a line of best fit. To modify the use of lines:

1. Click **Display Options . . .** or right click anywhere in the condition scatter plot window and select **Display Options**.

2. Click the Lines to Graph tab.
3. To see a grid inside the plot area, you can have lines drawn at the major and minor tick intervals of each axis. Check the **X/Y/Z Axis Grid Lines** checkboxes to display. The color of these grid lines is represented in the Grid Color box at the bottom of the window. To modify the grid color, click **Change . . .**

Changing Labels and Features

The scatter plot view also allows you to change the appearance of data points and data labels. To modify these features:

1. Click **Display Options . . .** or right click anywhere in the condition scatter plot window and select **Display Options**.
2. Click the Features tab.
3. To modify the size and shape of the points choose from among the options in the **Style** and **Size** pull-down menus.
4. There are five options for labeling the plot:
 - **Show Condition Names**—Displays the name of each gene to the lower right of each point. These names become unreadable if more than ~100 genes are visible in the current gene list and magnification.
 - **Show X Axis Label**—Displays the parameter that is graphed on the X axis.
 - **Show Y Axis Label**—Displays the parameter that is graphed on the Y axis.
 - **Show Z Axis Label**—Displays the parameter that is graphed on the Z axis.
 - **Show unclassified Group When Splitting the Window**—When the window is split, this option displays the genes that were not put into any classification into their own section of the genome browser.

Coloring

Coloring in the scatter plot view is more complicated than in other views because the color of each gene can be derived from the data in any axis. In other views, the color of the gene is usually linked to the data plotted on the vertical axis. In addition, the scatter plot allows you to color genes based on a fourth experiment or condition that is not plotted on *either* axis. To modify the way data points are colored:

1. Select **View > Display Options . . .** or right click anywhere in the genome browser and select **Display Options**.
2. Click the Coloring tab.
3. Specify whether to color conditions by gene, parameter, or attribute.
4. If you are coloring by parameter or attribute, select the appropriate parameter or attribute from the pull-down menu.
5. To change the default color, click the Change... button and choose the desired color.

Showing/Hiding Window Display Elements

You have the option of showing or hiding many of the elements in the GeneSpring window. To change the visibility of these elements, select **View > Visible** and choose one of the following options:

- **Picture**—Shows or hides the optional picture at the bottom right corner of the window
- **Animation Controls**—Shows or hides the slider and the Animate check box at the bottom of the window (hiding this check box does not disable the Animation feature)
- **Magnification**—Shows or hides the Magnification feature and the **Zoom Out** button at the bottom of the window (hiding the **Zoom Out** button does not disable the **Zoom Out** menu option)
- **Secondary Picture**—Shows or hides your secondary picture when you are viewing two gene lists or experiments simultaneously in the genome browser
- **Secondary Animation Controls**—Shows or hides the secondary Animation Controls check box and slider when you are viewing two gene lists or experiments simultaneously
- **Navigator**—Shows or hides the navigator panel
- **Hide All**—Hides everything in the window except the genome browser
- **Show All**—Shows all elements
- **Hide All in All Windows**—Hides everything in all windows except the genome browser
- **Show All in All Windows**—Shows all elements in all windows

Normalizing Data

Experiment Normalizations

Experiment normalizations are used to standardize your microarray data to enable differentiation between real (biological) variations in gene expression levels and variations due to the measurement process. Normalizing also scales your data so that you can compare relative gene expression levels.

GeneSpring assumes the data you have entered is raw data and must be normalized. If your data has been pre-normalized around a median other than 1, it may not be accurately interpreted during analysis. If your data are pre-normalized this way, see “Normalize to a Constant Value” on page 5-12.

There are several ways to normalize your data in GeneSpring. Typically, you will want to do either one per chip normalization together with one per gene normalization or one per spot normalization with one per chip normalization. There are important exceptions to this, which are discussed below under the relevant normalization.

Most normalizations can be applied in any order, and different samples in the same experiment can be normalized in different ways. You have the option of applying most normalization step only to specific samples in your experiment. To do this:

1. Check the **Apply Only to Specific Samples** box. A list of samples in the current experiment appears. If these samples are named, the names appear as sample identifiers. If they are not named, by default each sample is named for the file it is from, possibly including the column name if there is more than one sample in a file.

Samples that cannot be normalized, i.e., samples with no normalized column, appear grayed out in the list, and cannot be selected.

2. Check the box of any samples to which you want to apply this step.
3. Click **OK** to add this step to your normalization scenario, or **Cancel** to quit without adding this step.

Using the Experiment Normalizations Window

To access the Experiment Normalizations window, select **Experiments > Experiment Normalizations**.

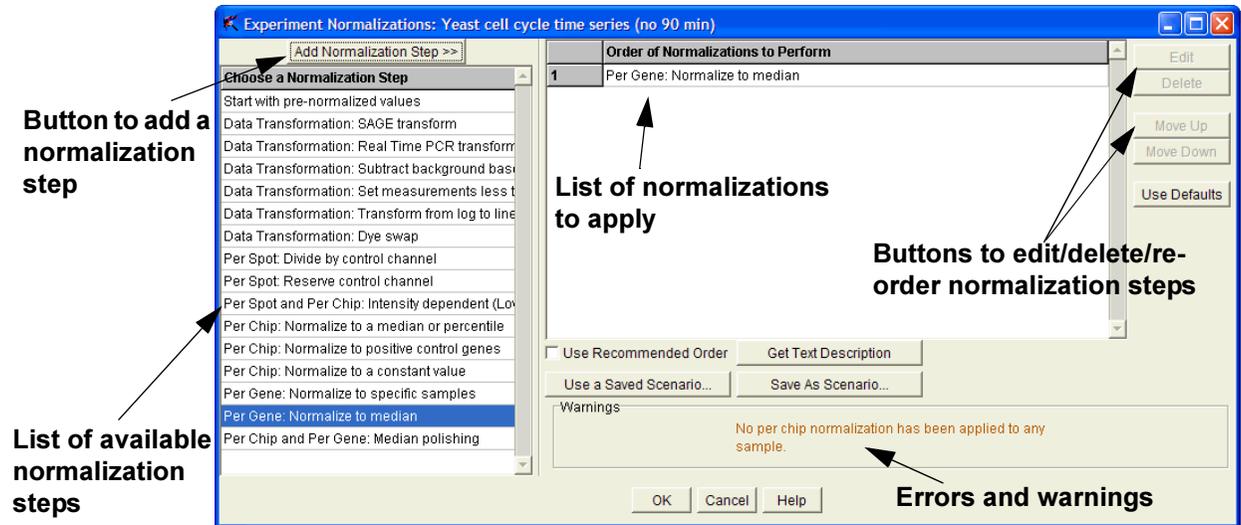


Figure 5-1 The Experiment Normalizations window

The Experiment Normalizations window lists the normalizations currently being applied to your experiment and allows you to add, edit, delete, or re-order normalization steps. You can save the current normalization steps as a scenario for future use, or load a previously saved scenario. The Warnings panel displays information about unmet requirements or other potential problems with the currently specified normalizations.

Adding a Normalization Step

To add a normalization step, select the desired normalization from the list on the left side of the screen and click **Add Normalization Step**. For detailed information on the available normalization types, see the following sections:

- “Per Spot Normalization” on page 5-7
- “Per Chip Normalizations” on page 5-10
- “Per Gene Normalizations” on page 5-13
- “Normalization Strategies for Specific Technologies” on page 5-17

Editing a Normalization Step

1. Select the desired step in the list of normalizations to be performed and click **Edit**, or double-click the step number of the selected normalization.
2. The configuration screen for the selected normalization appears.
3. Make any desired changes to the normalization settings.
4. Click **OK** to save your changes, or **Cancel** to exit without saving.

Removing Normalization Steps

To remove normalization steps, select one or more steps in the list of normalizations to be performed and click **Delete**. Be certain you have selected the correct step, since no confirmation dialog appears.

Re-Ordering Normalization Steps

To move a normalization step, select its name in the list and click **Move Up** or **Move Down**. Continue until the steps are in your desired order.

Applying Default Normalizations

When an experiment is created during the sample import process, normalizations are applied before you reach the Experiment Normalization window. These normalizations are determined from the data format of the samples in the experiment. For information on the default normalizations used during sample import, see “Default Normalizations” on page 3-21.

When you create an experiment from samples that have already been imported, the default normalizations are the Generic One-Color and Generic Two-Color scenarios. These normalizations can be applied when you create an experiment using the Create New Experiment menu, or by clicking the **Use Defaults** button in the Experiment Normalizations window.

To remove any changes you have made and apply only the default normalizations for your data type, click **Use Defaults**. A confirmation dialog appears. Click **OK** to continue or **Cancel** to quit and return to the main Normalizations screen.

Viewing Text Descriptions

To view a more detailed description of a particular normalization, select its name in the list and click **Get Text Description**. A dialog appears with a description of the selected normalization. You can copy the text in this dialog to the keyboard by clicking **Copy to Clipboard**. You can then paste the text into a text editor.

Saving a Normalization Scenario

Click **Save As Scenario...** to save the current normalization sequence for use in other experiments. A dialog appears prompting you to enter a name for the sequence to be saved. Once you save a normalization scenario, it is available for use in all genomes.

A saved scenario records whether each step was applied to all samples or a limited number of samples, but *not* which samples the steps were applied to. It also does not record a list of positive or negative controls or a list of control samples (as in the Normalize to Specific Samples option).

Working with Saved Scenarios

To load a saved normalization scenario, click **Use a Saved Scenario...** The Select a Normalization Scenario screen appears. From this screen you can do the following:

- **Load Scenario**—Select a scenario from the list and click **Load Scenario** to load it for use in the current experiment.
- **Delete Scenario**—Select a scenario from the list and click **Delete Scenario**. The scenario is removed from the list.

- **Rename Scenario**—Select a scenario from the list and click **Rename Scenario**. A dialog appears prompting you to enter a new name for the saved scenario.
- **Close**—Return to the previous screen without making any changes.

Normalization Warnings

Warnings occur under the following circumstances:

- A normalization step is missing
- A normalization step is inappropriate (i.e., there are too few genes or samples)
- Normalizations are applied to only some of the samples

Warnings appear in orange. You can proceed with an active warning, but the results may not be what was intended. Fatal errors appear in red. A fatal error means that the current normalization steps will not produce a usable result. In this case, the OK button is disabled until the problem is solved.

When a warning or error applies to a specific normalization step, that step is displayed in the list in the appropriate color for the warning or error.

Normalization Types

Start with Pre-Normalized Values

This option is provided for backwards compatibility, and allows you to maintain normalizations from a previous experiment. It can be applied only to samples that were created by GeneSpring's Merge/Split window or uploaded before GeNet 3.0. If you select this option when none of the data in your experiment have been previously normalized, a message alerting you to do this is displayed and the **OK** button is disabled.

Data Transformation

SAGE Transform

This method is recommended only for SAGE data. It fills in zeroes for all genes not mentioned in your data file.

Real Time PCR Transform

In this method, doubling measurements are converted into measurements of mRNA concentration using the equation 2^{-n} .

Subtract background based on negative controls

In this method, the median value of the gene list is subtracted from the raw values for each gene. The gene list used can be typed in or loaded from a file.

To type in a gene list, simply enter a gene in each line of the text box provided. Right-click in this box to use the Copy and Paste options.

To load a gene list from a file, click **Load From File** and select the gene list from the browse window that appears. If there are already genes listed when you click **Load From File**, the genes from the list you select are added to the existing list of genes. Any genes you have already entered are not overwritten.

Note: This text box can contain no more than 32,000 text characters (including carriage returns).

The list of negative control genes should be intersected with the regions, if there are any. Negative controls should be averaged within each region. If any region does not have any negative controls, an error message appears alerting you that the normalization cannot be performed.

Set measurements less than 0.01 to 0.01

This option sets any measurements less than a specified cutoff value to the cutoff value. By default, this value is 0.0. To enter another value, click in the Cutoff text box and enter a new value. This step can be applied before or after other normalizations.

Transform from log to linear values

This option transforms logarithmic data into linear expression values. This is required if your raw data are reported in log values, since GeneSpring requires data to be linear. To view your data on a logarithmic scale, use the experiment interpretation. For more information on experiment interpretations, see “Experiment Interpretations” on page 3-39.

To apply this normalization, specify the base of the original measurements by selecting the appropriate radio button. The available options are:

- Base 2
- Natural Log (e)
- Base 10
- Other— Enter a base value in the provided text box

Dye Swap

This option swaps your Control channel and Signal channel in order to do dye comparisons. Dye swap is available only for two-color experiments.

This step must be the first normalization step that changes the units of signal and control. For example, a log transform can come before this step, but not a per chip normalization. This is because the Signal and Control must be in the same units.

Per Spot Normalization

Per Spot normalizations are commonly used for two-color experiments. The formula for this normalization is:

$$\frac{\text{(signal strength of gene A in sample X)}}{\text{(control channel value for gene A in sample X)}}$$

Divide by control channel

This option divides the measured intensity of each gene by the value of its Control channel. This is recommended for two-color experiments if you do not use intensity-dependent normalization. If the Control channel value is very low, a cutoff value is used instead. By default, this value is 10.0. To change this value, click in the Cutoff text box and enter the desired cutoff value.

Note: The cutoff value cannot be lower than 0.000001.

This normalization works as follows:

	Signal >= Cutoff	Signal < Cutoff
Control >= Cutoff	Signal/Control	Signal/Control
Control < Cutoff	Signal/Cutoff	No Data

Reserve Control Channel

This option was previously known as Use Control Channel for Trust. This option tells GeneSpring to use the control channel to determine the saturation of the color of your genes. This is recommended when you imported the Signal to Control Ratio and the Control Channel.

Enter the value below which you do not trust the control signal in the Cutoff text box. By default, this value is 10.0.

Intensity Dependent Normalization

Intensity dependent normalization (often called non-linear or LOWESS normalization) is recommended for use in most two-color experiments. This step can be applied only to chips with more than 100 genes. LOWESS normalization uses region designators in the same way that other per-chip normalization methods do. For details, see “Region Normalization” on page 5-12.

Intensity dependent normalization is a technique that is used to eliminate dye-related artifacts in two-color experiments that cause the Cy5/Cy3 ratio to be affected by the total intensity of the spot. This normalization process attempts to correct for artifacts caused by non-linear rates of dye incorporation as well as inconsistencies in the relative fluorescence intensity between some red and green dyes. Such artifacts often result in a curve in the graph of raw versus control signal (see panel A in Figure 5-2).

In the absence of bias, one would expect there to be no dependence of raw signal on control signal and thus the data points would be scattered symmetrically around the 45° line.

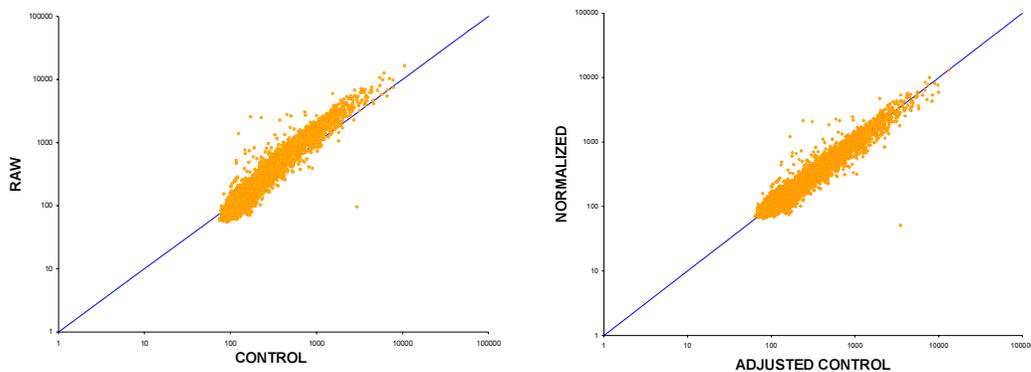


Figure 5-2 The Effect of Intensity Dependent Normalization

GeneSpring’s intensity dependent normalization feature fits a curve through the data and uses this curve to adjust the control value for each measurement. When the resulting normalized data are graphed versus the adjusted control value, the points are distributed more symmetrically around the 45° line (see Figure 5-2, panel B). You can specify the percent of data to be used for smoothing. By default, this value is 20.0%.

To counter the problem of taking logarithms of negative values (in subsequent steps), and to discount outliers, the raw and control values for each spot (R, G) are shifted by a constant:

$$shift = \max\left(\frac{\text{median}(R)}{100} - \text{min}(R), \frac{\text{median}(G)}{100} - \text{min}(G), 0\right)$$

The data goes through the following transformations:

$$\begin{aligned} (G, R) &\longrightarrow (G - shift, R - shift) \\ &\longrightarrow (\log(G - shift), \log(R - shift)) \\ &\xrightarrow{L} (A, M) \\ &\longrightarrow (A, M - f(A)) \\ &\xrightarrow{L'} (\log(G'), \log(R')) \\ &\longrightarrow (G', R') \\ &\longrightarrow (G' + shift, R' + shift) \\ &\longrightarrow \left(\frac{R(G' + shift)}{R' + shift}, R\right) \end{aligned}$$

...where $f(A)$ is the fitted function of the transformed data, and the coordinates are transformed by the operation:

$$L = \begin{bmatrix} 1/2 & 1/2 \\ -1 & 1 \end{bmatrix}$$

The last step of the transformation is performed because normalizations in GeneSpring are accomplished by adjusting the control value and leaving the raw value unchanged. The fit of the data, $f(A)$ is made using the LOWESS algorithm where the value $f = 0.2$ is used for the fraction of the total data points used for smoothing at each point (see “References” on page 5-21 for more information on the LOWESS algorithm). The degree of the polynomial fitted is 1. For efficiency, the regression is not calculated at each data point, but at a progressively fitted mesh that adjusts to the sparsity of the data.

If you attempt to re-normalize an experiment that has been constructed using the Merge-Split Experiment tools, you will be unable to apply intensity dependent normalization. The control values in merged experiments have already been adjusted and thus do not reflect the intensity of the reference dye.

If you perform an intensity dependent normalization, it is usually not necessary to perform a per chip normalization. Like normalizing to the distribution of all genes, *intensity dependent normalization should not be used on specialized arrays that contain a small number of genes, or on arrays where a majority of the genes may react similarly to experimental conditions.*

Per Chip Normalizations

Per Chip normalizations control for chip-wide variations in intensity. Such variations may be due to inconsistent washing, inconsistent sample preparation, or other microarray production or microfluidics imperfections. GeneSpring does not allow you to perform more than one per chip normalization, as they all address the same issue.

There is no dedicated option for region normalization. However, if you have region designators, all per-chip normalizations are performed on each region independently.

Normalize to a median or percentile

This option allows you to divide all of the measurements on each chip by a specified percentile value. By default, this value is 50.0%. To change this value, enter a new one in the text box. You do not have to restrict the measurements used in the calculation of the percentile. You can limit measurements based on a specified cutoff or by flag values.

If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction. To limit measurements by flag values, check the **Use only measurements flagged** box and select the appropriate option from the pull-down menu. The available options are:

- Present Only
- Present or Marginal
- Anything but Absent

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

Partially normalized means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

To limit by a cutoff:

1. Check the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the pull-down menu.
3. Enter the cutoff figure in the text box. The default value is 10.0.

You can choose to apply additional background correction in this step. To apply background correction, check the appropriate box in the Background Correction section of the screen. You have the following options:

- **Never apply extra background correction**
- **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes

- **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Note: Global Per Chip normalization is not recommended in any experiment where more than 50% of the genes on the chip are likely to be affected similarly by the experimental conditions. For example, if a chip containing only known growth factors were used to study differential expression in malignant and benign tumors, you might expect a majority of the genes to be differentially expressed. In this case, applying a per chip normalization would mask the changes in expression.

Normalize to Positive Control Genes

Some chips come with positive controls (mRNA from another genome or *housekeeping genes*), which are used to control for differences in the amount of exposure between samples. The formula for this difference is:

$$\frac{\text{(signal strength of gene A in sample X)}}{\text{(median signal of the positive controls in sample X)}}$$

(median signal of the positive controls in sample X)

To normalize to positive control genes, first enter a list of genes. This gene list can be typed in or loaded from a file.

To type in a gene list, simply enter a gene in each line of the text box provided. Right-click in this box to use the Copy and Paste options.

To load a gene list from a file, click **Load From File** and select the gene list from the browse window that appears. If there are already genes listed when you click **Load From File**, the genes from the list you select are added to the existing list of genes. Any genes you have already entered are not overwritten.

Note: This text box can contain no more than 32,000 text characters (including carriage returns).

Select the percentile of the positive controls by which to divide each sample. By default, this value is 50.0%.

You can limit measurements based on a specified cutoff or by flag values.

If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction. To limit measurements by flag values, check the **Use only measurements flagged** box and select the appropriate option from the pull-down menu. The available options are:

- Present Only
- Present or Marginal
- Anything but Absent

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. Negative numbers are allowed. These measurements are back-calculated

Normalization Types

based on the previous normalization steps. Rounding errors may be introduced in this process.

Partially normalized means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

To limit by a cutoff:

1. Check the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the pull-down menu.
3. Enter the cutoff figure in the text box. The default value is 10.0.

You can choose to apply additional background correction in this step. To apply background correction, check the appropriate box in the Background Correction section of the screen. You have the following options:

- **Never apply extra background correction**
- **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes
- **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Normalize to a Constant Value

If you are using a technology that calculates its own number for normalization, you will want to use constant values. For instance, Affymetrix's Global Scaling™ centers your data around 2500; in this case you would need to normalize your data to 2500 to center it around 1.

(signal strength of gene A in sample X)

(hard number in sample X)

To normalize to a constant value, simply enter the desired value in the **Per Chip: Normalize to a constant value** text box. By default, this value is set to 1.0.

Region Normalization

Regions are assigned in the column editor during the data loading process. For more information, see “Using the Column Editor” on page 3-9. If you have defined regions in your data, all normalization steps are applied on a per-region basis.

There are three ways of designating regions:

- Each data file for a sample is assumed to be a separate region
- Each distinct value in the Region Column is designated as a region
- A specified list of region codes (which may or may not be suffixes in the region column). This option is included for backwards compatibility.

The Affine Background Correction

If negative values form a large fraction of your data set, GeneSpring may automatically do what is known as the affine background correction. If a large percentage of your data are negative, normalization can be a problem. For instance, the median, which GeneSpring divides your data by in Use Distribution of All Genes, can be very small or even negative.

In such cases, GeneSpring readjusts the background level for your data by adding a constant to all raw control strengths such that the 10th percentile is set equal to 0. The affine background correction is applied only when the 10th percentile is more negative than the median of the data are positive. If the correction is applied, a warning message appears during data loading. Also, in the Gene Inspector, control strengths adjusted by this correction are flagged with asterisks.

You can choose to apply additional background correction in this step. To apply background correction, check the appropriate box in the Background Correction section of the screen. You have the following options:

- **Never apply extra background correction**
- **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes
- **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Per Gene Normalizations

Divide by Specific Samples

In this normalization, each gene is divided by the intensity of that gene in a specific control sample or by the average intensity in several control samples. The formula for this is:

$$\frac{\text{(signal strength of gene A in sample X)}}{\text{(signal strength of gene A in the control sample[s])}}$$

Or,

$$\frac{\text{(signal strength of gene A in sample X)}}{\text{(average signal strength of gene A in several control samples)}}$$

Normalization Types

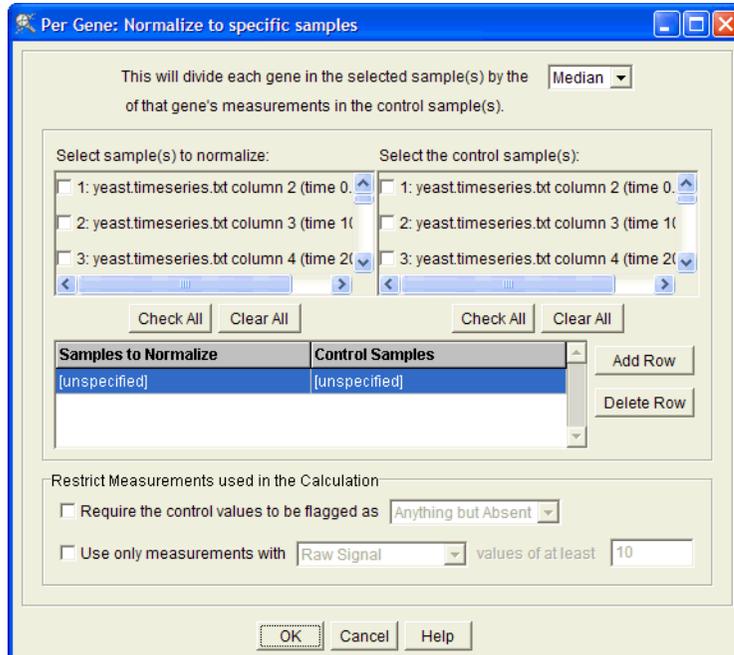


Figure 5-3 per gene: Normalize to specific samples window

Specify whether to divide by the mean or median by selecting from the pull-down menu at the top of the screen.

Choose a pair of samples and their control samples by checking the boxes next to the desired samples or typing them into the table below. Copy and Paste functions will also work in this table. Use commas or semicolons to separate samples, or dashes to indicate a range of samples. To select all samples in a column, click **Check All**. To unselect all samples in a column, click **Clear All**.

To define a new pair, use the **Add Row** button. To remove a pair, use the **Delete Row** button.

You can limit measurements based on a specified cutoff or by flag values.

If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction. To limit measurements by flag values, check the **Use only measurements flagged** box and select the appropriate option from the pull-down menu. The available options are:

- Present Only
- Present or Marginal
- Anything but Absent

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

Partially normalized means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

To limit by a cutoff:

1. Check the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the pull-down menu.
3. Enter the cutoff figure in the text box. The default value is 10.0.

Note: You cannot perform this normalization and normalize to the median of each gene, because they address the same issue.

Normalize to Median

This per gene normalization accounts for the difference in detection efficiency between spots. It also allows you to compare the relative change in gene expression levels, as well as display these levels in a similar scale on the same graph. GeneSpring uses the following formula to normalize to the median for each gene:

$$\frac{\text{(signal strength of gene A in sample X)}}{\text{(median of every measurement taken for gene A throughout your experiment)}}$$

(median of every measurement taken for gene A throughout your experiment)

If the median of the gene's measurements is below the specified cutoff value, the cutoff is used instead. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

Partially normalized means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

GeneSpring does not allow you to perform this normalization and normalize to sample(s), as they address the same issue.

Median Polishing

Median polishing means that each chip is normalized to its median and each gene is normalized to its median. These normalizations are repeated until the medians converge, up to a maximum of five iterations. This limit prevents endless looping if the normalization coefficients do not converge.

If measurements are limited by flag values, the percentile is calculated using only the genes that pass the flag restriction. To limit measurements by flag values, check the **Use only measurements flagged** box and select the appropriate option from the pull-down menu. The available options are:

- Present Only
- Present or Marginal
- Anything but Absent

Normalization Types

If measurements are limited by a cutoff, the percentile is calculated from all measurements above the cutoff. This cutoff can be in either raw or partially normalized units.

The Raw Signal option means that the cutoff is applied to the raw measurements in the original data file. These measurements are back-calculated based on the previous normalization steps. Rounding errors may be introduced in this process.

Partially normalized means that the cutoff is applied to the gene values resulting from the previous normalization steps (which may or may not be equivalent to the raw measurements).

To limit by a cutoff:

1. Check the **Use only measurements with** box.
2. Select whether to limit by raw signal or current normalized values from the pull-down menu.
3. Enter the cutoff figure in the text box. The default value is 10.0.

You can choose to apply additional background correction in this step. To apply background correction, check the appropriate box in the Background Correction section of the screen. You have the following options:

- **Never apply extra background correction**
- **Always apply extra background correction**—Prior to taking the specified percentile, the bottom tenth percentile is used as a background correction and subtracted from all genes
- **If needed apply extra background correction**—For samples in which the bottom tenth percentile is less than the negative of the specified percentile, the tenth percentile is used as a background correction and subtracted from all genes before the specified percentile is taken.

Normalization Strategies for Specific Technologies

Normalization of Affymetrix Data

Often data in affymetrix .chp files are either pre-scaled or are pre-normalized. While Affymetrix's scaling and normalizations are designed to meet the same needs as GeneSpring's, they are not equivalent.

The Affymetrix global scaling procedure, which is comparable to GeneSpring's per-chip normalization, scales the data of each chip to a user-defined target intensity. However, GeneSpring's per chip normalization option, Use Distribution of All Genes, divides each intensity value by the median of all of values on the chip. The resulting expression levels on each chip are centered around 1.

For pre-scaled Affymetrix data, we recommend applying per chip normalizations using the distribution of all genes. In addition we recommend applying a per gene normalization using the median of each gene. The greatest benefit of performing these normalizations is that each gene intensity is centered artificially around 1. Several GeneSpring functions depend on scaling around 1, especially the cross-gene error model.

Normalization of Two-color Microarray Data

Like most other technologies, two-color experiment data should be normalized at the gene-level (to standardize expression levels between genes) and at the chip level (to standardize expression values between arrays). Two-color experiments are designed to provide an internal standard at the spot level. This per spot normalization often provides the same scaling that would be provided by a per gene normalization, and thus per gene normalization is often unnecessary.

Per Chip normalizations are useful in two-color experiments to standardize the global intensities across multiple arrays. Even after applying a per spot normalization, global variability between chips often remains due to differences in the total amount of the dyes added to the sample and reference samples from one chip to another. However, the intensity dependent normalization option (which is actually a per gene normalization) succeeds in centering all of the values on each array around 1. In addition it provides protection from dye incorporation artifacts that lead to unwanted relationships between signal intensity and normalized expression values (see "Intensity Dependent Normalization" on page 5-8).

It is generally recommended to apply a per spot normalization using the **Divide by control channel** option, followed by selecting the **Per Chip Normalization** step. Both of these normalization options can be accessed by selecting **Experiments > Experiment Normalizations...**

Region Normalization

This normalization option allows you to normalize sections of a sample rather than normalizing over the entire sample. This is especially important if you used multiple arrays for each experimental point or if there is some reason you must normalize sections of an array separately from one another. Region normalization is not a separate mathematical

Normalization Strategies for Specific Technologies

formula the way the previous normalizations discussed in this chapter are. Using this normalization means if you normalize to negative controls, to positive controls or normalize each sample to itself you do not actually normalize over each sample, but rather perform the normalization over each region. Hence the formulas for these normalization options become:

Normalizing to Negative Controls for a Region:

$$\frac{\text{(the control strength of gene A in region Y of sample X)}}{\text{(the median signal of the negative controls in region Y of sample X)}}$$

Normalizing to Positive Controls for a Region:

$$\frac{\text{(the control strength of gene A in region Y of sample X)}}{\text{(the median signal of the positive controls in region Y of sample X)}}$$

Normalizing Each Region to Itself:

$$\frac{\text{(the signal for gene A in region Y of sample X)}}{\text{(the median signal of the genes region Y of sample X)}}$$

Dealing with Repeated Measurements

Single Data File

Occasionally the raw experimental data in the data file for your sample has more than one line devoted to a particular gene. This may be because you did the sample twice or because you did the sample once but took the measurements twice. If the same gene name is reported multiple times on different horizontal lines in your data file, GeneSpring automatically considers the measurements repeats and averages the signal strengths together.

GeneSpring reports the average and keeps track of the minimum and maximum values for each gene, but it cannot access the particular values falling between the minimum and maximum values. The formula for averaging a repeated gene is:

$$\frac{\text{[(the signal strength of gene A1) + (the signal strength of gene A2) + ... + (the signal strength of gene An)]}}{N}$$

This process is repeated for each gene repeated in a data file *before* any other normalizations are applied to the raw values.

Frequently samples are repeated with exactly the same parameters, but are reported in different data files. If this is the case, the fact the samples are repeats is represented via parameter. The same normalization is employed when dealing with an experimental parameter considered to be a repeat, but in that case the averaging takes place after the raw data for each gene has been normalized. See “Experiment Parameters” on page 3-29 for more information about repeats reported in separate data files.

Mathematical Illustration of the Dealing with Repeated Measurements in a Single Data File Method

Given this raw data, with four repeats of YMR199W (marked with the arrows):

➤	YMR199W	1117
	YMR200W	1384
	YMR201C	1101
	YMR202W	1357
	YMR203W	1162
	YMR204C	1464
	YMR206W	978
	YMR199W	973
	YMR207C	1618
	YMR208W	1374
	YMR209C	1432
	YMR210W	1068
	YMR211W	1568
➤	YMR199W	1313
	YMR212C	1638
	YMR213W	1648
	YMR214W	1282
➤	YMR199W	1218
➤	YMR199W	1496

GeneSpring averages all of the measurements of YMR199W to get a signal strength of 1286. GeneSpring notices the maximum control strength for YMR199W in this sample is 1496 and the minimum is 1117. These values are the end points of YMR199W’s error bar which GeneSpring plots when you choose to display error bars in either the graph or the scatter plot displays.

Measurement Flags

Measurement flags are markers in your data set, and data can be assigned as one of four flags:

- Passed (or OK)
- Marginal
- Absent
- Failed
- Unknown

Flags assigned by you when the experiment is entered into GeneSpring:

- **Good Data**—data are present and reliable. Marked with a “P” for passed or “O” for ok.
- **Marginal Data**—data are present, but of unknown or dubious quality. Marked with an “M” for marginal.
- **Absent Data**—There is no data available, and there should have been. Marked with an “A” for absent or “F” for failed.

Flags assigned by GeneSpring:

- **Unavailable Data**—If there is no flag in the column, GeneSpring assigns that measurement a “U”.

Only measurements at the highest available flag level are combined and treated as replicates. The order of flag precedence is **P M U A**. If one or more Ps are present, only Ps are used. If no Ps are present and one or more Ms are present, only Ms are used, etc. Summary statistics are collected over these cases and stored with the corresponding flag. All other flag data are discarded for the gene. This is done when the experiment is loaded into GeneSpring and is not affected in any way by later choices about which codes are to be used or displayed. The only way to avoid this is to not declare a flag column during data load, in which case the flags are not available for other uses.

For information about measurement flags and how to load them into your experiment, see “Using the Column Editor” on page 3-9.

Negative Control Strengths

Some types of microarray technology report negative control strengths. This is usually the result of subtracting estimated background levels that are larger than the raw signal. This can happen in situations where the expression levels of the gene are low compared to the measurement error. It can also happen when there is background subtraction or when a mismatched probe set has higher intensity levels than the perfect match probe sets.

If negative signal levels occur in a large fraction of the data used for normalization, there can be problems with the normalization, as the median across the normalization set can be very small or even negative. This leads to unreasonable results of normalization. In such cases, which only occur in a few situations, GeneSpring does an extra step in the normalization, where it readjusts the background level for that data by adding a constant to all the raw control strengths in such a way that the 10th percentile of the signal is set equal to 0, before proceeding with the median normalization. This correction, called the affine background correction, is applied only when the 10th percentile of the data is more negative than the median of the data is positive. A warning message appears when you first load your data into GeneSpring if this background correction has been applied.

Whether or not the above correction is applied, negative signal levels may still be present for a few measurements. GeneSpring offers the option as the last step of normalization to set these values to zero. Also, when interpreting data in logarithm or fold interpretations, GeneSpring treats all normalized ratio values less than 0.01 (including 0 and negative values) as if they had a ratio of 0.01 to prevent transformation problems.

References

Cleveland, W. S., and S. J. Devlin. (1988). Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* 83, 596-610.

Yang, Y.H., Dudoit, S., Luu, P., and T.P. Speed. (2001) Normalization for cDNA Microarray Data. *SPIE BiOS 2001*, San Jose, California, January 2001.

References

6

Analyzing Data

Creating and Editing Gene Lists

You can create a gene list by selecting **Edit Gene List...** from the Edit menu. The Gene List Editor screen appears. From this screen you can create a new gene list based on a variety of selection criteria, or add/remove genes from existing gene lists. This screen is very similar to the Sample Manager.

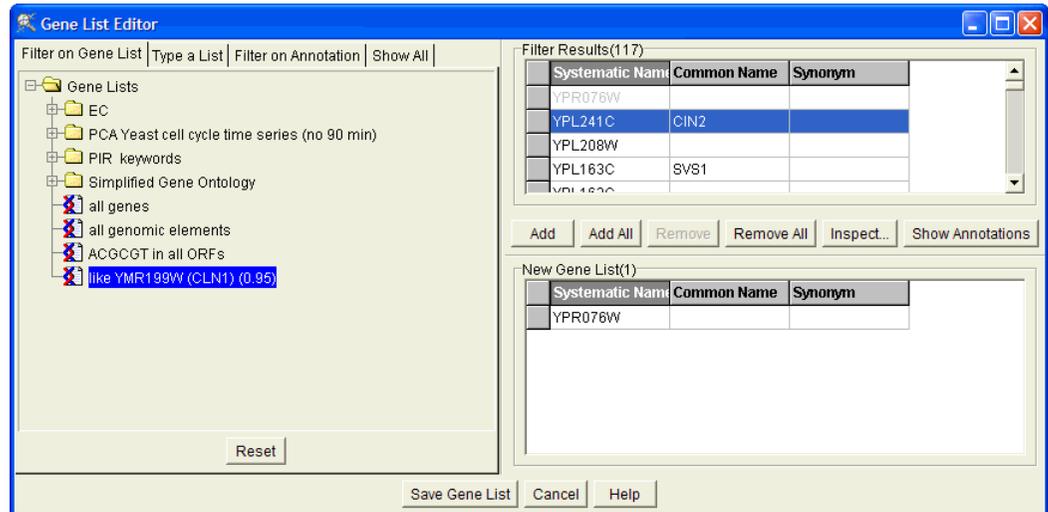


Figure 6-1 The Gene List Editor window

The left side of the screen contains tabs for each filtering method. Click a tab to view options for that method. These methods are:

- **Show All**—Display all available genes without applying a filter
- **Filter on Annotation**—Display genes based on a specified annotation
- **Type a List**—Manually enter a list of genes
- **Filter on Gene List**—Display genes from a selected gene list

The right portion of the screen contains two tables. The upper table contains all of the genes resulting from the current filtering method. The lower table contains the genes you have selected to add to your list. Between the two tables are six buttons.

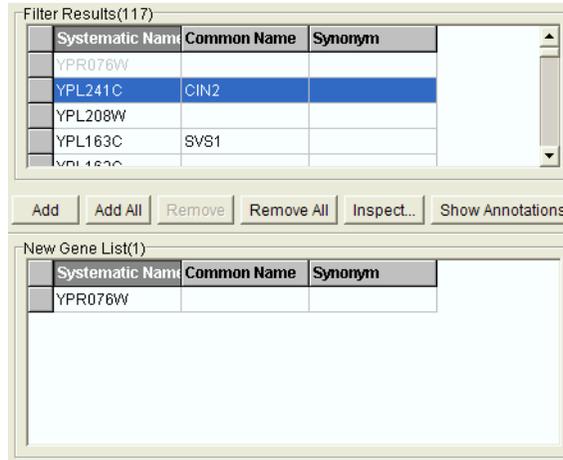


Figure 6-2 Gene Tables

The buttons are as follows:

- **Add**—Add a selected gene in the Filter Results table to the New Gene List table.
- **Add All**—Add all genes in the Filter Results table to the New Gene List table.
- **Remove**—Remove a selected gene from the New Gene List table.
- **Remove All**—Remove all genes from the New Gene List table.
- **Inspect**—View the selected gene in the Gene Inspector. For more information on the Gene Inspector, see “The Gene Inspector” on page 4-10.
- **Show Annotations**—Show or hide annotations from a gene list.

Filtering Methods

Filter on Gene List

To filter on gene list, use the GeneSpring navigator to locate the desired gene list and select it in the list. All genes in that list appear in the Filter Results table.

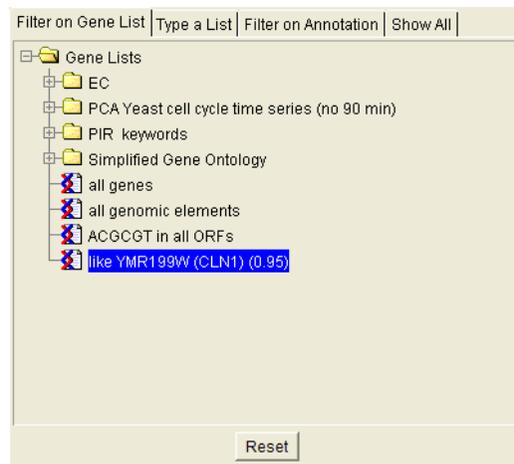


Figure 6-3 The Filter on Gene List tab

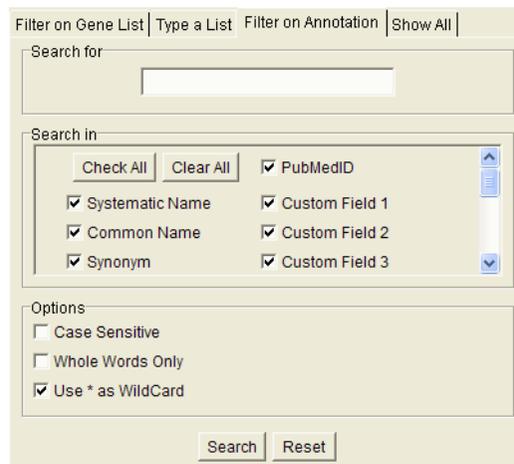
Type a List

This method allows you to manually enter a list of genes. To enter genes, simply click in the gene list box, type a gene's Common Name, Systematic Name, Synonym, or GenBank Accession Number and press Enter. You can also use copy and paste to enter one or more genes to this list.

If the gene is found, it appears in the Filter Results table. If the gene is not found, it is colored in red. To clear your entries, click Reset. This removes all entered genes from both the list of genes you have typed in and the Filter Results table.

Filter on Annotation

This method allows you to filter genes based on text or values in a specified annotation.



The screenshot shows a web-based interface for filtering genes. At the top, there are four tabs: 'Filter on Gene List', 'Type a List', 'Filter on Annotation' (which is selected), and 'Show All'. Below the tabs is a 'Search for' text input field. Underneath that is a 'Search in' section containing a 'Check All' button, a 'Clear All' button, and a list of checkboxes for search criteria: 'PubMedID', 'Systematic Name', 'Common Name', 'Synonym', 'Custom Field 1', 'Custom Field 2', and 'Custom Field 3'. Below the search criteria is an 'Options' section with three checkboxes: 'Case Sensitive', 'Whole Words Only', and 'Use * as WildCard'. At the bottom of the form are 'Search' and 'Reset' buttons.

Figure 6-4 The Filter on Annotation tab

To filter on annotation:

1. Enter a search term in the Search For text box.
2. Select the annotation fields to search in. You have the following options:
 - Systematic Name
 - Common Name
 - Synonym
 - Notes
 - GenBank Accession Number
 - EC
 - Description
 - Product
 - Phenotype
 - Function
 - Keywords
 - PubMed ID
 - Custom Fields 1-3

- Type
- DBid
- GO Biological Process
- GO Molecular Function
- GO Cellular Component
- RefSeq
- UniGene
- Any custom annotation fields you may have added

Some fields may not be visible due to window size. To view all fields, scroll down in the Search In panel. To check all boxes, click **Check All**. To uncheck all boxes, click **Clear All**.

3. Select any desired search options. The available choices are:
 - Case Sensitive
 - Use Whole Words Only
 - Use * as WildCard
4. Click **Search**. Genes matching the specified search parameters appear in the Filter Results table.

Working with Gene Lists

The Find Similar Command

Similar lists in the Gene List Inspector window are gene lists that contain a significant number of overlapping genes with the one selected. The p-value is calculated using the hypergeometric probability. This equation calculates the probability of overlap corresponding to k or more genes between a gene list of n genes compared against a gene list of m genes when randomly sampled from a universe of u genes:

$$\frac{1}{\binom{u}{m}} \sum_{i=k}^n \binom{m}{i} \binom{u-m}{n-i}$$

The “standard list” checkbox in the Gene List Inspector window allows you to define a newly created list as “standard list”. If a list is defined as standard, the list is included in the search for similar lists. Some lists, such as those created using the Simplified Gene Ontology tool, are automatically defined as “standard lists”.

You can also change the GeneSpring Preferences to allow the program to search through all lists in your genome for similar lists. Open **Edit > Preferences**, select the Miscellaneous tab and change the settings under **Restrict Gene List Searches**.

Each gene expression profile must contain the set minimum correlation to be considered similar. The higher you set the minimum correlation (maximum 1), the closer the gene expression profiles must be.

To make a list with the Find Similar command:

1. Double-click a gene to bring up the Gene Inspector.
2. Click **Find Similar**. The New Gene List window appears, which includes the genes in that list, as well as lists that are similar to your new gene list.

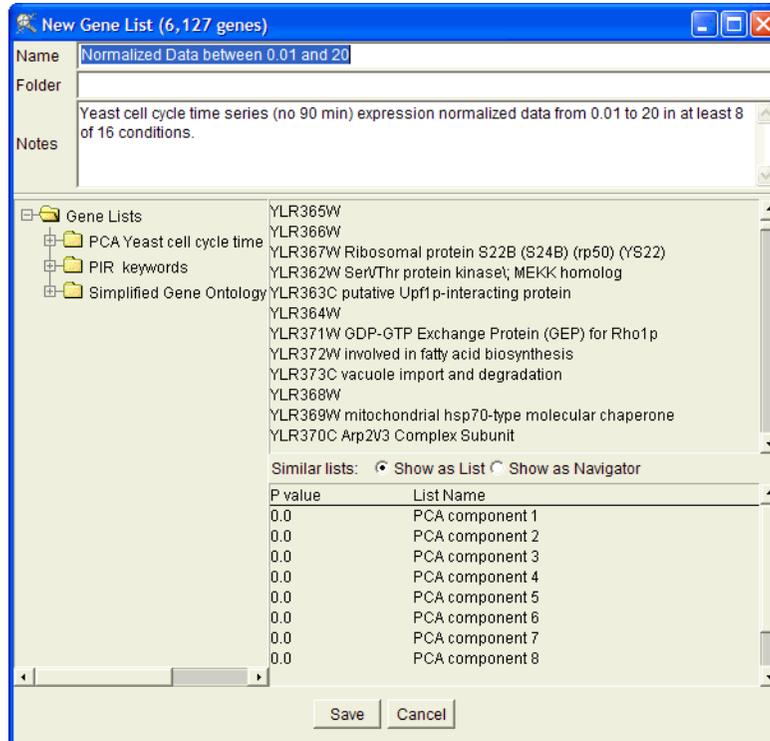


Figure 6-5 The New Gene List window

3. Enter a name for the new gene list, or accept the default and click **Save**.

The Find Similar Genes Window

This command allows you to set up complex correlations against the inspected gene. These correlations may involve more than one experiment or condition or extra restrictions on experiments.

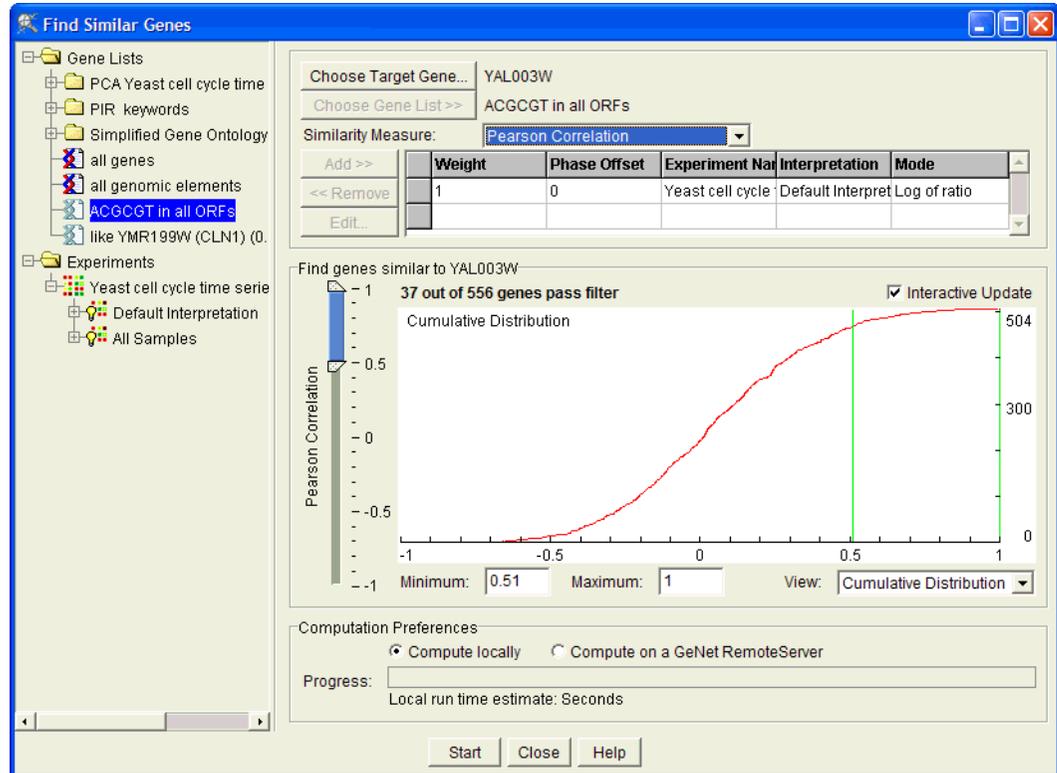


Figure 6-6 The Find Similar Genes window

You can view the preview pane as a cumulative distribution, a graph, or by linking the preview to the main GeneSpring window. When you are working with large experiments, you may want to uncheck the Interactive Update box to the upper right of the preview pane to avoid slowing the analysis.

1. Double-click a gene to bring up the Gene Inspector.
2. Click **Complex Correlations**. The Find Similar Genes window appears.

This window can also be accessed by selecting **Tools > Find Similar Genes** from the GeneSpring main menu. In this case you must also click **Choose Target Gene** and enter a gene name on the Find Target Gene screen. For more information on this screen, see “Performing an Advanced Search” on page 4-4.

3. Select a gene list from the navigator and click **Choose Gene List**.
4. Select an option from the **Similarity Measure** menu. Available options are:
 - Standard Correlation
 - Smooth Correlation
 - Change Correlation
 - Upregulated Correlation
 - Pearson Correlation
 - Spearman Correlation
 - Spearman Confidence
 - Two Sided Spearman Confidence

- Distance

You can specify minimum and maximum settings for the similarity measure by moving the sliders or entering values in the appropriate fields.

5. Select an experiment or condition in the navigator and click **Add**. The New Correlation window appears.

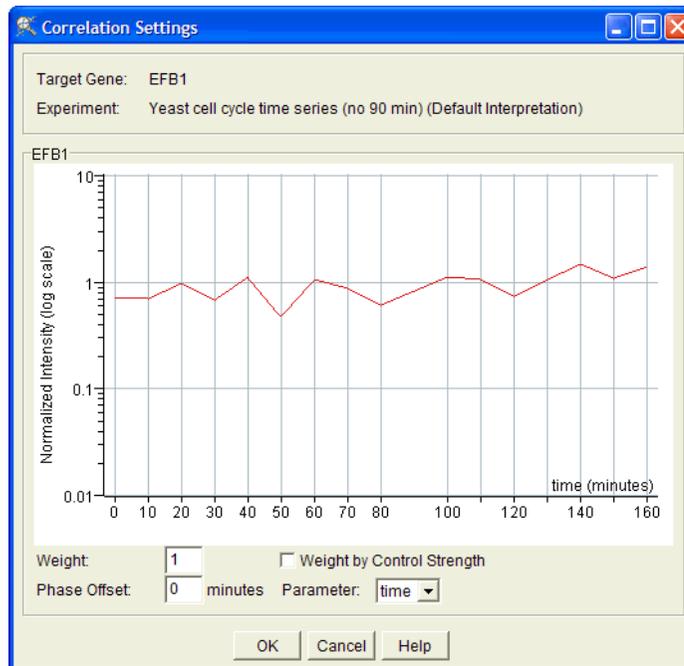


Figure 6-7 The Correlation Settings window

A cumulative distribution graph of gene correlations appears in the center of the New Correlation window. The horizontal axis shows the correlation from zero to 1. The vertical axis represents the number of genes. The green lines are your specified maximum and minimum values. If you change these values the green lines move accordingly.

6. If desired, enter new values for Phase Offset and Weight. You can select a parameter from the pull-down menu in the phase offset section.
7. Click **OK**. You are returned to the Find Similar Genes window. You can now add additional experiments or conditions if you wish. For more information on this portion of the screen, see “The Correlations table” on page 6-10.

To remove an experiment or condition, click on the name of the experiment or condition in the white center box and click **Remove**. To change the settings, click the experiment’s name to select its row and click **Edit**.

8. Specify whether to run the search locally or on a GeNet Remote Server.
9. When you are done, click **OK**. The New Gene List window appears.
10. Name your gene list and click **Save**. The list appears in the Gene Lists folder of the main navigator.

The Correlations table

The Correlations table lists the experiments chosen to correlate against the specified gene. The experiments selected may be weighted, making one more important than another. If both experiments chosen are given a weight of 1, they are averaged equally. To modify an experiment's weight, click in the Weight column to the left of its name and enter a new weight.

The equation used to determine the overall correlation is:

$$X = \frac{(Aa + Bb + Cc + \dots)}{(a + b + c + \dots)}$$

A	The correlation coefficient between the gene in question in experiment 1 and the selected gene, also from experiment 1.
a	The weight specified for experiment 1.
B	The correlation coefficient of the gene in question in experiment 2, to the selected gene, also from experiment 2.
b	b is the weight associated with experiment 2.
C	The correlation coefficient of the gene in question in experiment 3 to the selected gene, also from experiment 3.
c	The weight associated with experiment 3.

...and so on.

Experiments 1, 2, 3, and so forth, are all of the experiments selected in the white Correlations table. If **X** is between the minimum and maximum correlations specified in the Find Similar Genes window, the gene in question passes the correlations.

Standard Correlation	Measures the angular separation of expression vectors for Genes A and B around zero. <i>Result = $\mathbf{a \cdot b} / (\mathbf{a} \mathbf{b})$</i>
Smooth Correlation	Make a new vector A from a by interpolating the average of each consecutive pair of elements of a . Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph screen. Do the same to make a vector B from b . <i>Result = $\mathbf{A \cdot B} / (\mathbf{A} \mathbf{B})$</i>
Change Correlation	Make a new vector A from a by looking at the change between each pair of elements of a . Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\text{atan}(a_{i+1}/a_i) - \pi/4$. Do the same to make a vector B from b . <i>Result = $\mathbf{A \cdot B} / (\mathbf{A} \mathbf{B})$</i>

Upregulated Correlation	<p>Make a new vector A from a by looking at the change between each pair of elements of a. Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\max(\text{atan}(a_{i+1}/a_i)-\pi/4, 0)$. Do the same to make a vector B from b.</p> <p><i>Result = $\mathbf{A} \cdot \mathbf{B} / (\mathbf{A} \mathbf{B})$</i></p>
Pearson Correlation	<p>Calculate the mean of all elements in vector a. Then subtract that value from each element in a. Call the resulting vector A. Do the same for b to make a vector B.</p> <p><i>Result = $\mathbf{A} \cdot \mathbf{B} / (\mathbf{A} \mathbf{B})$</i></p>
Distance	<p>Distance is not a correlation at all, but a measurement of dissimilarity. Distance is the measurement of Euclidian distance between the expression profile for gene A (defined by its expression values for each point in N-dimensional space, where N is the number of conditions with data in your experiment) and the expression profile for gene B.</p> <p><i>Result = $\mathbf{a}-\mathbf{b}$ divided by the square root of the number of conditions with data</i></p>
Spearman Correlation	<p>Order all the elements of vector a. Use this order to assign a rank to each element of a. Make a new vector a' where the i^{th} element in a' is the rank of a_i in a. Now make a vector A from a' in the same way as A was made from a in the Pearson Correlation. Similarly, make a vector B from b.</p> <p><i>Result = $\mathbf{A} \cdot \mathbf{B} / (\mathbf{A} \mathbf{B})$</i></p>
Spearman Confidence	<p>Compute a value r of the spearman correlation as described above.</p> <p><i>Result = $1 - (\text{probability you would get a value of } r \text{ or higher by chance.})$</i></p>
Two sided Spearman Confidence	<p>Compute a value r of the spearman correlation as described above.</p> <p><i>Result = $1 - (\text{probability you would get a value of } r \text{ or higher, or } - r \text{ or lower, by chance.})$</i></p>

Making Lists by Applying Filters

To make a list from filter data, open any filter window, set the desired options, and click **Save . . .** The New Gene List window appears.

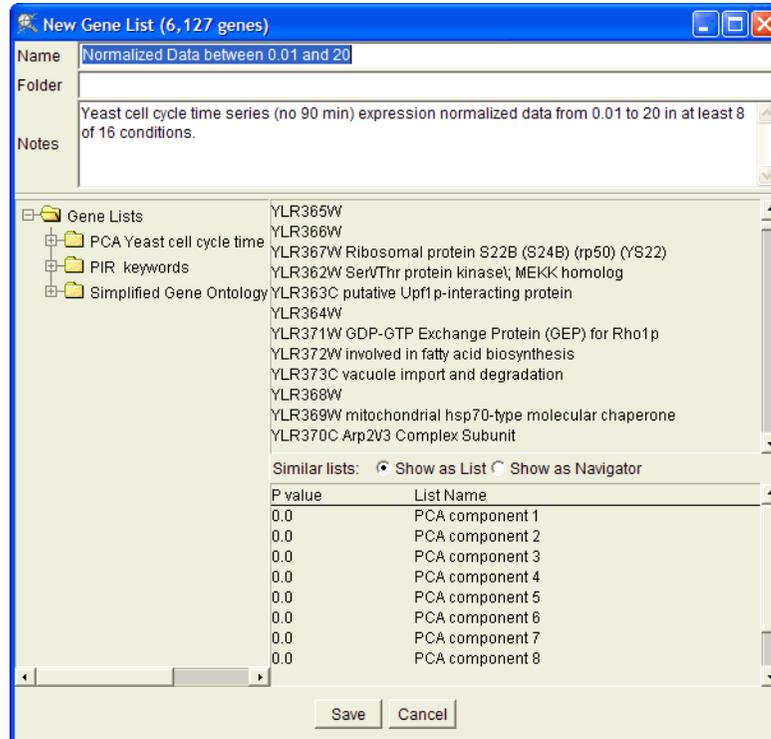


Figure 6-8 The New Gene List window

Enter a name and destination folder for the new gene list, or accept the default and click **Save**. To specify the location for the destination folder, select the desired parent folder from the navigator.

Making Lists from Properties

You can make gene lists based on the properties (annotations) contained in your master gene table. These lists are not ordered.

To make a list from properties:

1. Select **Annotations > Make Gene List from Properties** (pre-4.1 users select **Tools > Make Gene List from Properties**).
2. Choose a property on which to base your list from the pull-down menu.
3. Uncheck the Divide by semicolons box if you do not want your data separated by semicolons.
4. You can specify to include a list only if it has a certain number of members, or you can include all lists.

By default, GeneSpring removes gene lists with one or fewer members. Change this number in the text box provided, or include everything by unchecking the Remove classifications with 1 or fewer box.

5. Enter a name for your gene list folder.

6. Click **OK**. A new folder with the gene list you created appears in your Gene Lists folder.

Making Lists with the Venn Diagram

A Venn Diagram allows you to quickly visualize genes common to more than one gene list. You can also find genes present only in a particular list. The gray area behind the circles represents the Venn Diagram “universe” (the selected gene list). Genes in the selected list that are common to gene lists represented by the Venn diagram circles appear as numbers in those circles. For information about creating and filling Venn Diagrams, see “Color by Venn Diagram” on page 4-32.

To make a list with a Venn Diagram:

1. Right-click the area of the Venn Diagram in which you want to make a list.
2. Select an option from the pop-up menu. A New Gene List window appears.

If you click in an area where two circles overlap, you have the following options:

- **Make list of these genes**—list genes in the immediate geometric area.
- **Make list of genes in both lists**—list genes common to the two circles, i.e. the intersection.
- **Make list of genes in either list**—list all genes in the two circles, i.e. the union.

If you click in an area where three circles overlap, you have the following options—

- **Make list of genes in all lists**—list genes common to the three circles, i.e. the intersection.
- **Make list of genes in any list**—list all genes in the three circles, i.e. the union.

If you click a non-overlapping (gray) area, you can make a list of genes in that section only.

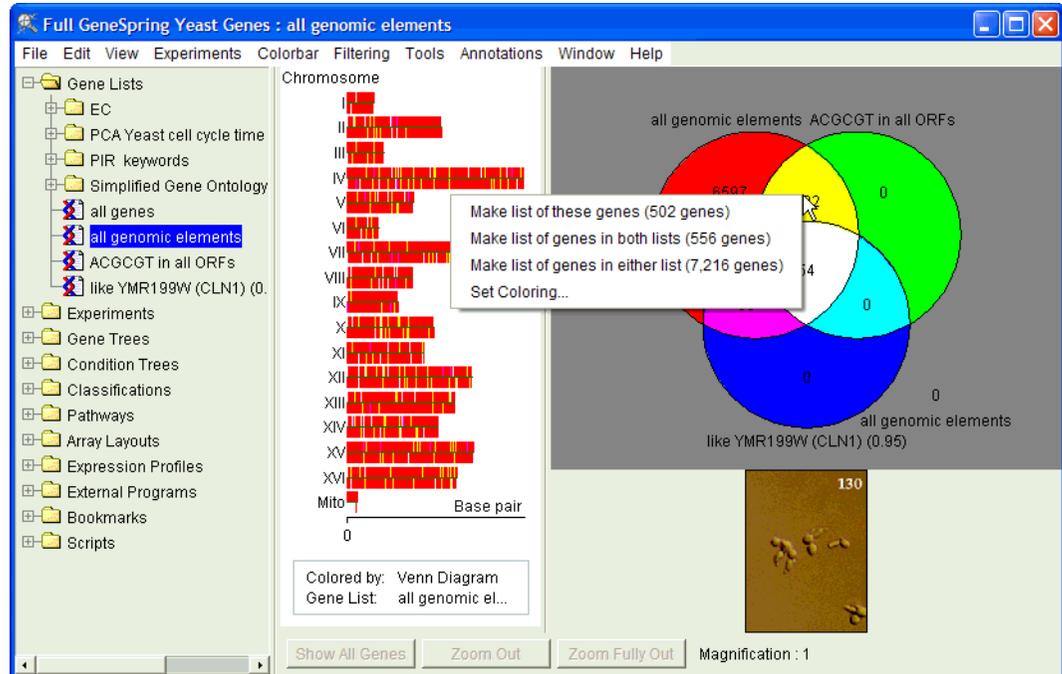


Figure 6-9 A Venn diagram with pop-up menu

3. Specify the gene list from which to obtain associated numbers and click OK.
4. Name and save your new list.

In views where lists can be ordered, such as the Ordered List and Compare Genes to Genes views, lists made from the Venn diagram are ordered according to the values associated with the lists you used to create the Venn Diagram. When more than one of these lists has values, genes are ordered according to the values of the last list added to the Venn diagram when it was created.

Making Lists from Classifications

You can generate gene lists from any classification. For example, if you have a 5-cluster k-means classification, you can view which genes are in each cluster by making a gene list from the k-means classification.

To make a gene list from a classification:

1. Right-click a classification in the Classifications folder in the navigator.
2. Select **Make Gene Lists**. GeneSpring creates a gene list folder for the classification containing one list for each cluster.

This folder appears in the Gene Lists folder in the navigator.

Making Lists from Selected Genes

This command allows you to make lists from genes you select graphically.

There are two ways to select a set of genes. If genes are grouped together in the browser, you can select a set the same way you select an area to enlarge:

1. Hold down the shift key, click on a region, and drag a rectangle across the desired area to select.
2. Release the mouse button before releasing the shift key. Selected genes appear in white.

Alternately, you can select multiple genes by clicking over their representative lines or rectangles while holding down the shift key.

3. Once you have selected all the desired genes, right-click in the genome browser and select **Make List from Selected Genes** from the pop-up menu. A New Gene List window appears.
4. Name your list and click **Save**.

For more information about this window, see “Creating and Editing Gene Lists” on page 6-2.

Creating Expression Profiles

The Creating Expression Profiles function allows you to draw a pseudo-gene to represent a hypothetical expression pattern. This function is useful if you have some idea of what gene expression pattern you are looking for, as you can simply draw a pattern and look for genes that behave similarly.

You must be in Graph view to create an expression profile. Double-click the expression profile to open the Gene Inspector for that gene.

To create an expression profile:

1. Select **Tools > Draw Expression Profile**. A new gene appears on the screen at the normalized median of your data (usually 1.0).
2. To change the shape of this gene, click on the gene and drag while holding down the control key. (On Macintosh systems, Option-click.)

To save an expression profile:

1. Double-click the expression profile to open the Gene Inspector.
2. Click **Save Expression Profile**.
3. Name the new profile and click **Save**. Your new expression profile appears in the Expression Profiles folder in the navigator.

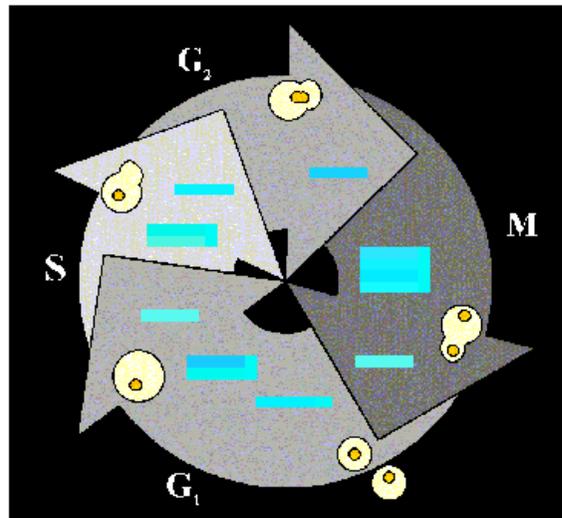
To make lists from expression profiles:

1. Double-click the expression profile to open the Gene Inspector.
2. Click **Find Similar**. A New Gene List window appears with a list of similar genes and lists.
3. Name the list and click **Save**. Your new list appears in the genome browser and in your Gene Lists folder.

Pathways

A pathway is a graphical representation of the interaction between gene products in a biological system. Genes can be superimposed on the pathway, allowing you to view their expression levels in a biological context. You can zoom in on a pathway, and move the slider to watch gene expression change over the experimental conditions.

You can draw pathways yourself or use publicly available pathways such as KEGG (Kyoto Encyclopedia of Genes and Genomes). One scenario in which a pathway can be very useful is if you are trying to identify a class of genes that are associated with a particular step or regulatory element within a pathway.



Selected Pathway:	mitosis
Colored by:	Yeast cell cycle time series (no 90 min) (Default Interpreta...
Gene List:	all genomic elements (7216)

Figure 6-10 A cell cycle pathway

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a large public database containing pathways for many organisms, all of which are available for download via FTP. To locate pathways for a specific organism, go to the following URL:

`ftp://ftp.genome.ad.jp/pub/kegg/pathways`

Folders are named using three letter abbreviations based on the Latin name of the organism, i.e., “hsa” is Homo sapiens, “mmu” is Mus musculus.

To locate generic pathways, download the “maps” folder. This folder contains reference pathways (metabolic and regulatory). These pathways contain enzymes rather than genes. When you import an organism-specific pathway, the GENE file is parsed, and for each KEGG accession number the gene names are stored (including the accession number itself, if it is not fully numeric). All EC numbers are also stored.

GeneSpring first tries to match the first gene name with the genes in the current genome. It is compared against systematic and common names. If no match is found, GeneSpring attempts to match the second name, and so on. As soon as matching genes are found (often

it is one gene, but sometimes there are more), GeneSpring stops any further processing in order to decrease the chance of false matches.

If no matches are found, GeneSpring tries to match EC numbers. Because several genes may correspond to the same EC number, in this case GeneSpring searches for matches to all EC numbers in the GENE file. This reduces the possibility of a false match.

Importing a Pathway

Before you can import a pathway, you must download it from the above FTP site.

To import a pathway:

1. Select **File > New Pathway > Import KEGG Pathway**. The Import KEGG Pathways window appears.
2. Navigate to the directory of pathways you downloaded from KEGG.
3. If desired, check the box to group pathways alphabetically into subfolders of nine pathways each. (This is useful if you want to split a window by a group of pathways.)
4. If desired, check the box to create gene lists from the imported pathways. (If you checked the box to group pathways into subfolders, the gene lists will also be grouped by subfolders.)
5. Click **OK**.
6. In the screen that appears, accept the default or enter a new name for the folder of pathways.
7. Click **OK**.

Saving pathways may take several minutes.

Adding a Gene to a Pathway

Once you have successfully imported your graphic into GeneSpring, you can place genes on top of the background image.

1. Open the appropriate Pathway in the navigator.
2. While holding down the **Ctrl** key, draw a box where you would like the gene to appear on the pathway. (Macintosh users Option-click.) The New Genes on Pathway window appears.
3. Enter the gene name, accession number, or keyword (such as a word in a gene's descriptor) and click **OK**. The gene name appears on the pathway.

If the gene name or keyword is present for more than one gene, another window appears directing you to choose a gene ID from a list. Double-click on the appropriate ID.

To remove a gene, right-click on the element and select **Delete Pathway Element**.

Finding New Genes on a Pathway

GeneSpring uses proprietary algorithms to predict the genes that fit near a selected point on a pathway. When you select a point, GeneSpring makes two lists of genes from those

currently displayed on your diagram. List A contains the two genes that appear closest to your selected point on the diagram. List B contains all other genes on the pathway.

GeneSpring examines all the genes on your currently selected gene list and finds all genes whose minimum similarity (correlation) with genes on list A is higher than their maximum similarity with genes on list B. These genes are made into a separate list for you to examine. You can place a gene from this list on the pathway (see “Adding a Gene to a Pathway” on page 6-17).

If your pathway geometry is complex, this procedure is not very useful, since it relies on screen distance only, not pathway structure or connectivity.

To find new genes on a pathway:

1. Right-click near a group of genes displayed on your pathway.
2. Choose the option **Find Genes Which Could Fit Here**. The New Gene List window appears.
3. Enter a name and destination folder and click **Save**. The new gene list is saved in your Gene Lists folder.

Pathway Commands

Right-click your Pathway in the navigator for the following options:

- **Display Pathway**—Displays the selected pathway in the genome browser.
- **Make Gene List**—Allows you to save a list of all the genes on the selected pathway.
- **Attachments**—Allows you to add a text or picture attachment to your Pathway.
- **Inspect**—Displays a listing of details such as pathway history and genome.
- **Publish to GeNet**—Uploads your information and the pathway picture to GeNet (see “Publishing Data to GeNet” on page 9-11).
- **Rename Pathway**—Allows you to rename your pathway.
- **Delete Pathway**—Deletes a pathway. A confirmation dialog box appears.
- **Export as Zip**—Allows you to export the pathway as a GeneSpring .zip file.

Regulatory Sequences

The Find Potential Regulatory Sequence window allows you to find common regulatory sequences within genes in a gene list or to search for a known sequence. It also compares the frequency of occurrence against all other gene lists in the genome.

This feature is useful for finding genes sharing similar regulatory sequences or having a particular regulatory sequence in common.

When the regulatory sequences tool compares genes to the remainder of the genome, it uses the “all genes” list. The “all genomic elements” list includes non-gene elements that are not expressed.

In GeneSpring version 4.0 and later, the sequence information is loaded automatically.

Note: To change the load automatically feature, select **Edit > Preferences > Data Files** and uncheck the Load Sequence box.

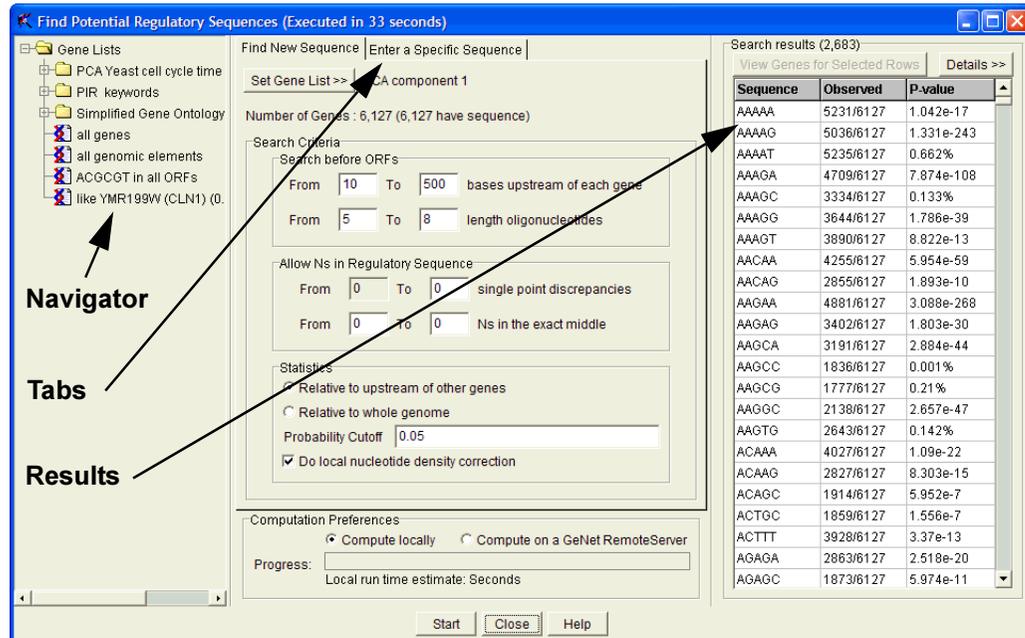


Figure 6-11 The Find Regulatory Sequences window

From this screen you can do the following:

- **Find new sequences**—This option searches for short sequences upstream of the genes in the current gene list or across the entire genome.
- **Enter a specific sequence**—This option allows you to enter a known sequence.

To find a new regulatory sequence:

1. Select **Tools > Find Potential Regulatory Sequences**. The Find Potential Regulatory Sequences window appears.
2. Click the **Find New Sequence** tab at the top of the screen.

Figure 6-12 The Find New Sequence tab

3. Select a gene list from the navigator and click **Set Gene List**.

Note: Do not choose the “all genes” or “all genomic elements” gene lists. You are already comparing your selected gene list against all other genes in the genome.

4. Enter the number of bases upstream of each gene in the **Search Before ORFs** section of the window. For example, if you enter “From 10 To 100” on a search for ACGCGT, GeneSpring searches for any part of the promoter within the region between 10 and 100. The smaller the range between these numbers, the greater the likelihood that the results are statistically significant.

Larger sequences may take longer to search. You can also search for common sequences within the ORF by using negative numbers for the bases.

5. Enter the length of the oligonucleotides to search for.
6. Enter the number of single point discrepancies allowed. This refers to a maximum number of mismatches allowed. For example, if you specify one single point discrepancy, ACGCGAT satisfies a search for ACGCGTT.
7. Enter the range of base gaps in the exact middle. This refers to the size of an allowable hole in the middle of the sequence, allowing you to look for sequences such as ACGnnnCGT, which is biologically relevant due to loops and non-binding areas.

The gap must be in the exact middle, with the longer side of odd sequences appearing before the Ns. It does not count towards the sequence length specified; hence ACGnnnCGT would be returned as an oligonucleotide of length 6.

8. Select whether the sequence is relative to the sequence upstream of other genes or relative to the whole genomic sequence. The first option is far more common.

The Probability Cutoff textbox indicates the level of significance (P-value) needed for an oligomer to be listed in the results. You can change this value.

9. Specify whether to perform the operation locally or on a remote GeNet server.

10. Click **Start**. The button becomes a **Stop** button. The progress bar lengthens as your search progresses. For very large genomes or complex search parameters, this operation may take a few minutes.

To enter a specific regulatory sequence:

1. Select **Tools > Find Potential Regulatory Sequences**. The Find Potential Regulatory Sequences window appears.
2. Click the **Enter a Specific Sequence** tab at the top of the screen.

Figure 6-13 The Enter a Specific Sequence tab

3. Select a gene list from the navigator and click **Set Gene List**.

Note: Do not choose the “all genes” or “all genomic elements” gene lists. You are already comparing your selected gene list against all other genes in the genome.

4. Enter the number of bases upstream of each gene in the **Search Before ORFs** section of the window. For example, if you enter “From 10 To 100” on a search for ACGCGT, GeneSpring searches for any part of the promoter within the region between 10 and 100. The smaller the range between these numbers, the greater the likelihood that the results are statistically significant.

Larger sequences may take longer to search. You can also search for common sequences within the ORF by using negative numbers for the bases.

5. Enter the promoter sequence in the **Sequence** box.
6. Enter the number of single point discrepancies allowed. This refers to a maximum number of mismatches allowed. For example, if you specify one single point discrepancy, ACGCGAT satisfies a search for ACGCGTT.
7. Select whether the sequence is relative to the sequence upstream of other genes or relative to the whole genomic sequence. The first option is far more common.

The Probability Cutoff textbox indicates the level of significance (P-value) needed for an oligomer to be listed in the results. You can change this value.

8. Specify whether to perform the operation locally or on a remote GeNet server.
9. Click **Start**. The button becomes a **Stop** button. The progress bar lengthens as your search progresses. For very large genomes or complex search parameters, this operation may take a few minutes.

Viewing Regulatory Sequence Search Results

The search results are displayed in the Results area of the Find Potential Regulatory Sequences window.

Click **Details** for expanded results data.

Click **View Genes for Selected Row** or double-click any sequence to view the Conjectured Regulatory Sequence window.

- **Sequence**—The nucleotide sequence of the oligomer.
- **Observed**—The number of genes in the list where the oligomer was found.
- **P-value**—The probability (P-Value) that the number of occurrences in the list came about by chance. Only nucleotide motifs with P-values below the specified probability cutoff (in this case 0.05 or 5%) are shown.
- **Random Rate**—The intrinsic probability, which is the percent of genes you would expect this specific nucleotide combination to appear upstream of, if the nucleotide sequence were strictly random (it is not, of course, but this is a good value to compare the observed probability to).
- **Observed: Other Genes**—The observed probability of this sequence motif appearing upstream of genes other than the list under inspection. If the option **Relative to sequence upstream of other genes** is selected, this becomes the probability of the observed sequence occurring relative to the genes not in the list, i.e., relative to the “all genes” list.

If the option **Relative to whole genomic sequence** is selected, this becomes the probability of one or more occurrences of the sequence based on the rate of occurrence in the entire genome.

The formula used to calculate this is:

$$1 - (1 - k/b)^n$$

where:

k = the number of occurrences in the whole sequence

b = the total number of bases

n = the length of the upstream region being searched

- **Expected**—The number of incidences in the searched gene list in which you would expect this oligomer to occur. The number for the *Expected* column is derived using the larger of the intrinsic probability and the observed probability values.
- **Single P**—this column displays the Single P value for the motif. This is the chance this particular sequence would be found if only one test was performed.
- **Tests**—The number of tests run to generate these motifs appears in the last column. This is the number of oligomers tested that were the length of the sequence motif found.

Using the Conjectured Regulatory Sequence Window

The Conjectured Regulatory Sequence window displays the common nucleotide sequence, showing the 10 bases that precede and follow it in the area near (or in) each gene where the oligomer is found. It also gives a brief description of the statistics listed in the Results box of the Find Potential Regulatory Sequences window, and allows you to modify the observed motif by removing an item, extending the promoter or making a new gene list.

Double-click one of the sequence motifs given in the Results box of the Find Potential Regulatory Sequences window to view the Conjectured Regulatory Sequence window.

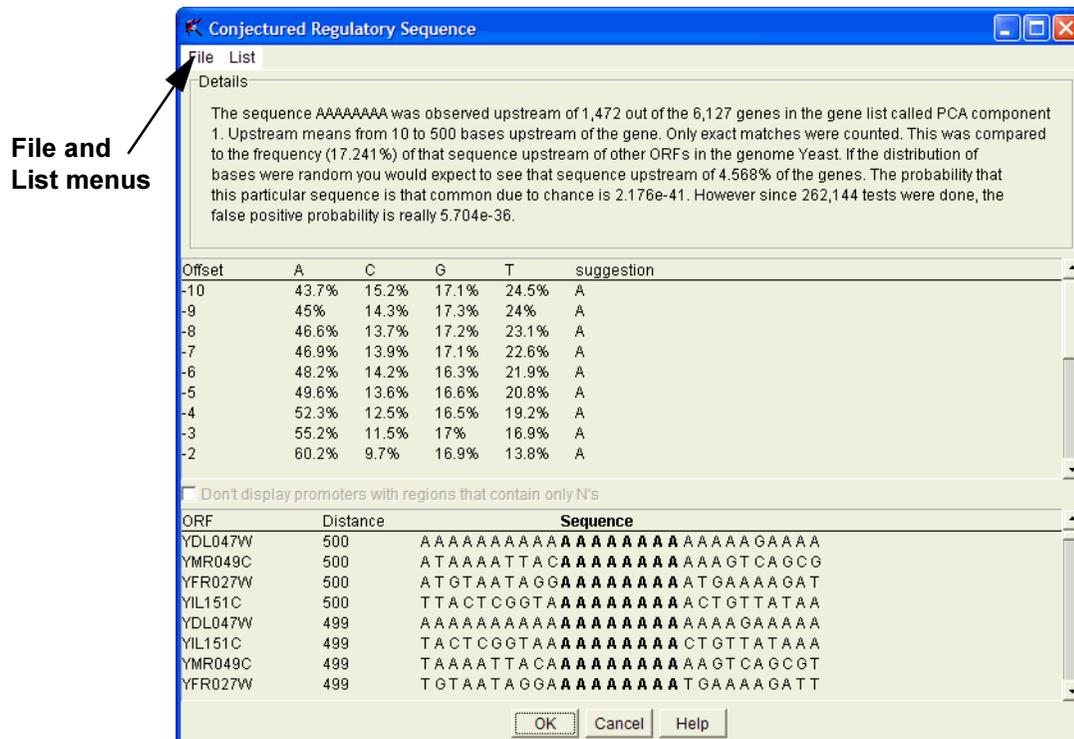


Figure 6-14 The Conjectured Regulatory Sequence window

Two menus, **File** and **List**, are located in the upper left corner of the window.

The File menu contains the following commands:

- **Print**—Prints the list in the lower half of the Conjectured Regulatory Sequence window.
- **Close**—Closes the Conjectured Regulatory Sequence window

The List menu contains the following commands:

- **Remove Item**—Removes the highlighted item and its associated sequence motif from the list matching the common sequence motif being examined.
- **Make Gene List**—Displays the new Gene List window.

When a gene list is produced based on the occurrence of a specified sequence (in this example, ACGCG in the yeast data) there is a number associated with each gene corresponding to distance of the first such sequence upstream of the ORF. The numbering

begins from first nucleotide. Zoom in on the Ordered list view or open the Gene List Inspector to view these numbers.

- **Extend Promoter**—Adds a new, longer and hopefully better promoter in the Find Potential Regulatory Sequences window.

The Conjectured Regulatory Sequence window contains the following sections:

- **Details**—Provides a general description of the common sequence motif being inspected. The details found in this box are the same numbers listed in the right-hand columns of the Results box in the Find Potential Regulatory Sequences window.
- **Offset Bases**—The middle third of the Conjectured Regulatory Sequence window contains statistics on the bases to either side of the motif. The first column contains the offset from the observed sequence. The next four columns contain the percentage of genes with that base in that position. The last column contains a suggested extension to the motif.
- **ORF**—The bottom third of the Conjectured Regulatory Sequence window contains the sequence information for the motif being inspected, as it occurs in the nucleotide sequence in the area near (or in) each gene where it is found. There are three columns of data.
 - **ORF**—Indicates the gene that the common sequence motif (given in bold, centered in the column) is upstream of.
 - **Distance**—Displays the number of bases upstream the oligomer is from the ORF associated with it in the first column. This number is the difference between the base pair number of the first base in the gene and the base pair number of the first nucleotide in the motif. It includes the distance of the promoter. This means the distance number is the difference between the promoter sequence and the ORF.
 - **Sequence**—Contains the sequence being examined, displayed in **bold**. On the left side of it are the ten bases preceding this instance of the motif, and on the right side are the 10 bases that follow it in the nucleotide sequence.

The Homology Tool

The Homology Tool automates the process of building homology tables for certain organisms. Currently, this is a limited list of organisms. Using this tool, homologies can be made between any pair of organisms that are included in both HomoloGene and UniGene. Within-genome homologies are based solely on UniGene Cluster ID or LocusLink Locus ID.

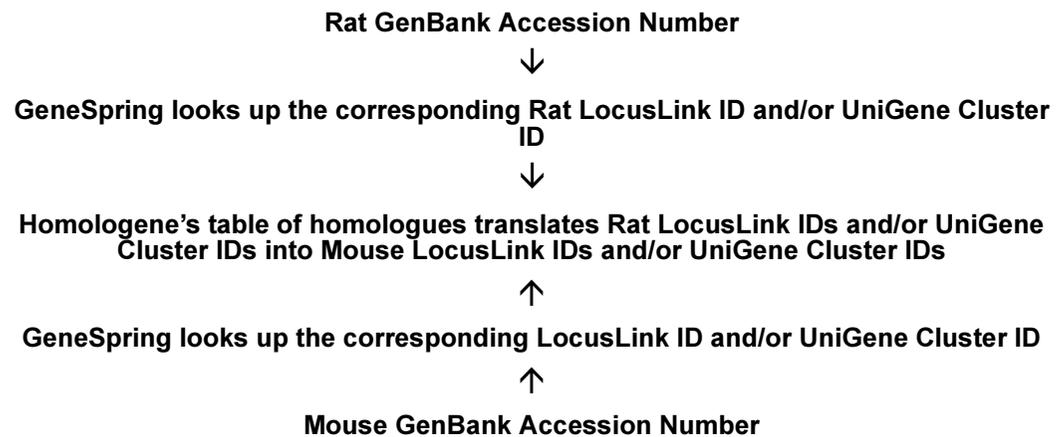
The Master Table of Genes for each genome must be in GeneSpring 5.0 format. If the selected genomes are in an earlier format, you must convert them before running the Homology Tool. For more information on the GeneSpring 5.0 Master Table of Genes format, see “The Master Gene Table File” on page 2-10.

Homologies can be created for the following organisms:

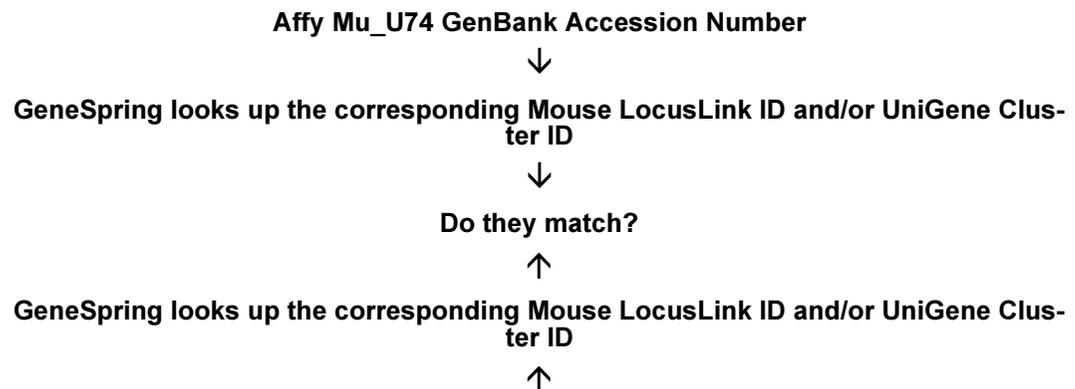
- Cow (*Bos Taurus*)
- Nematode (*Caenorhabditis elegans*)
- Sea squirt (*Ciona intestinalis*)
- Zebrafish (*Danio rerio*)

- Fruit fly (*Drosophila melanogaster*)
- Human (*Homo sapiens*)
- Mouse (*Mus Musculus*)
- Rat (*Rattus norvegicus*)
- Pig (*Sus scrofa*)
- Clawed frog (*Xenopus laevis*)
- Thale cress (*Arabidopsis thaliana*)
- Barley (*Hordeum vulgare*)
- Rice (*Oryza sativa*)
- Wheat (*Triticum aestivum*)
- Maize (*Zea mays*)

The following is an example of how GeneSpring builds a homology table between Rat GenBank Accession Numbers and Mouse GenBank Accession numbers:



Below is an example of how GeneSpring makes a homology table between Affy Mu_U74 GenBank Accession Numbers and Incyte Mouse GenBank Accession Numbers:



To use the Homology tool:

1. Select a genome from the GeneSpring navigator.

2. Select **Annotations > Build Homology Tables**. The Build Homology Tables window appears. The selected genome is displayed in the upper table on the right side of the screen.

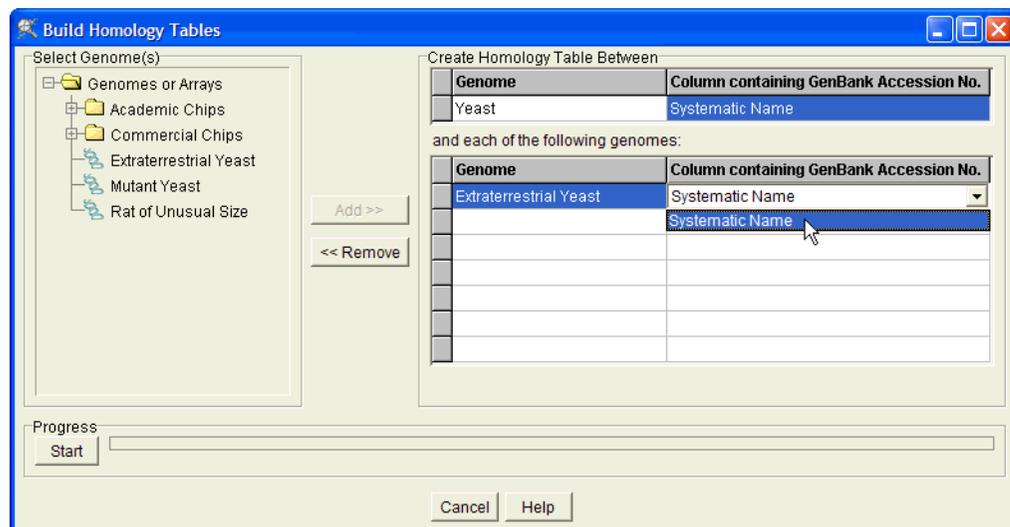


Figure 6-15 The Homology Tool

3. From the pull-down menu in the Column containing GenBank Accession No. column, select the appropriate column for the selected genome.
4. Select a second genome from the mini-navigator on the left side of the screen and click **Add**. The genome is added to the lower table on the right side of the screen.
5. From the pull-down menu next to the newly added genome, select the name of the column containing that genome's GenBank Accession Number.
6. If desired, add additional genomes using the same procedure.
7. Click **Start**. This process takes about as long to run as the GeneSpider.

If the initially selected genome does not have GenBank Accession Numbers, an error message appears. If a selected genome is not on Homologene, you will receive an error message after the Homology Tool has finished running.

8. When prompted, specify whether or not to save the UniGene Cluster IDs.

The resulting homology tables are saved in the originally selected genome's Data\Homology Tables folder.

Annotation Tools

The Annotations menu in GeneSpring allows you to update your genome, make gene lists based on annotations, and build gene ontology tables. Annotations can also be searched using the Find Gene feature in the Edit menu. See “Performing a Simple Search” on page 4-4 for details.

Updating your Master Gene Table with GeneSpider

After you have loaded a new genome, you can make sure it contains the latest information from the genome databases on the World Wide Web by using GeneSpider. To use GeneSpider, you must have GenBank accession numbers in your master gene table. GenBank accession numbers are usually added to the GenBank column of the master gene table. If you have multiple GenBank accession numbers for a single gene, they should be separated by semicolons.

For details on adding information to your master gene table see “The Master Gene Table File” on page 2-10.

Updating Annotations with GeneSpider

1. Select **Annotations > GeneSpider**. (Pre-4.1 users: Select **Tools > GeneSpider**). Choose one of four options:
 - **Update annotations from Silicon Genetics**—Retrieves gene information from the Silicon Genetics Mirror Database. The mirror database caches information from GenBank, LocusLink, and UniGene to ease the load on the NCBI server and allow you to update faster. The mirror database is updated about once every two months.
 - **Update annotations from GenBank**—Retrieve information on genes from GenBank.
 - **Update annotations from LocusLink**—Retrieve information from LocusLink.
 - **Update annotations from UniGene**—Retrieve information from UniGene.

The Update Genome window appears.

2. **If you selected Update genes from Silicon Genetics**, the window has a different appearance because more options are available. If you selected any of the other options, proceed to the next step.

To upgrade from Silicon Genetics:

- a. Check the boxes next to the annotation sources from which to retrieve data.
- b. Specify the retrieval method:
 - Select **Concatenate annotations from different sources** to retrieve annotations from all sources as a semicolon-delimited list. Exact duplicates are not retrieved. The order is fixed: GenBank, then LocusLink, then UniGene.
 - Select **Keep the highest priority annotation** to retrieve only the annotation from the highest priority source available for each gene.

- c. Proceed to step 3.
3. Select the name of the column containing GenBank accession numbers from the pull-down menu in the upper right portion of the screen.
4. To update information in places where data already exists, select the **Overwrite Existing Annotations** checkbox. If you leave this box unchecked, GeneSpring adds new information only to blank fields. When you update annotations, GeneSpring creates a back-up file of the pre-update master gene table.

Updating from Silicon Genetics or GenBank gives you the option to retrieve sequence data.

5. Click **Start** to begin updating annotations.

While the GeneSpider runs there are a number of informational fields visible.

- **Status**—this is the level of completion the GeneSpider has reached.
- **Processed**—Number of genes in the genome that the GeneSpider has finished querying the database on.
- **Found**—number of processed genes where the GeneSpider has found a useful record in the database.
- **Enhanced**—Number of genes where information has been found that was not on the master gene table.
- **Go To**—Number of genes in the genome that have not been processed.

The available options are slightly different when updating from Silicon Genetics.

The master gene table is not updated until you click **Save and Close**. This button is inactive while the GeneSpider is running. You must wait until the GeneSpider is finished, or click the stop button, before clicking **Save and Close**.

The GeneSpider Errors Window

While the GeneSpider is running, the GeneSpider Errors window may appear. This window lists any errors the GeneSpider encountered and a brief description of the problem. For example, if no match was found for some genes on your system, the Errors window displays the gene identifier and the text “Gene not found”.

The most common reason for no genes to be updated is that you did not select the annotation column containing the GenBank Accession Numbers.

Problems with the Map Location Annotations

This window appears when the information the spider has retrieved for a gene’s map location field does not meet the required criteria. This window contains a table in which you can correct the entry, and GeneSpring makes a guess as to what the entry should be when it displays the table of all the problem entries.

Which Annotations are Retrieved?

The following table describes the annotation fields retrieved by GeneSpider from the various databases:

Annotation	Silicon Genetics	GenBank	LocusLink	UniGene
Systematic name*				
Common	X	X	X	X
Map	X	X	X	X
GenBank*				
Synonym*				
EC number	X	X	X	
Description	X	X		X
Product	X	X	X	
Phenotype	X	X	X	
Function	X	X		
Keywords	X	X		
PubMedID*				
Type**				
DBId	X	X		
GO biological process †	X			
GO molecular function †	X			
GO cellular component †	X			
RefSeq	X			
UniGene***	X			
Sequence	X	X		

* Systematic name, GenBank accession number, Synonym, and PubMedID are not filled in by GeneSpring. These fields can be filled in manually by the user. (I have entered a feature request to enable the GeneSpider to fill in PubMedID from GenBank records.)

** Type is filled in by GeneSpring when it reads a genome from a GenBank (.gbk) file. The value is commonly “CDS”, but “mRNA”, “rRNA”, “terminator”, “gene”, and other GenBank feature keynames are possible entries.

*** UniGene can also be filled in by the Build Homology Tables feature. When the user requests Build Homology Tables to save UniGene cluster IDs, it replaces the UniGene column with its results, deleting any obsolete entries from the column. (This contrasts with the behavior of the GeneSpider, where overwrite will not replace existing entries with an empty entry.)

† Gene Ontology information on the Silicon Genetics Mirror server is obtained from the LocusLink database. The same information cannot be provided through Update Annotations from LocusLink because the LocusLink web page format does not indicate the organizing principles of biological process, molecular function, and cellular component.

Genome Databases

Silicon Genetics

Update Annotations from Silicon Genetics retrieves annotations from the Silicon Genetics Mirror server at info.sigenetics.com. The server downloads the complete databases for GenBank, RefSeq, LocusLink, and UniGene from NCBI. It returns the same annotations as the GeneSpiders that access GenBank, LocusLink, and UniGene, depending on the annotation sources chosen by the user in the Update genome from Silicon Genetics window. In addition, the Silicon Genetics GeneSpider fills in GO biological process, GO molecular function, GO cellular component, and RefSeq identifier fields from the LocusLink database and the UniGene cluster ID from the UniGene database.

GenBank

Update Annotations from GenBank retrieves information from the GenBank and RefSeq databases at NCBI, which both use the same record format. A sample record is shown below, with information retrieved by the GeneSpider highlighted in blue and the fields it fills in GeneSpring's master table of genes indicated on the following line. The record is organized with keywords and subkeywords. The features table is organized by feature keys.

A complete description of the format for the latest release of GenBank is available at <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>.

The GeneSpider fills in annotations from GenBank as summarized in the following table. Feature keys are found in the margin of the features table (the section between the FEATURES line and the BASE COUNT line). Qualifiers indicate information about a feature; they begin with a slash followed by the qualifier name, then an equals sign (e.g. /gene=).

Annotation	Feature Key	Qualifier
Common	CDS or gene	gene
Map	source	map or chromosome
EC number	CDS or gene	EC_number
Description	CDS or gene (also from DEFINITION line)	note
Product	CDS or gene	product
Phenotype	CDS or gene	phenotype
Function	CDS or gene	function
DBId	CDS or gene	db_xref
	keyword	

Keywords	KEYWORDS
Description	DEFINITION (also from CDS/gene note)
Sequence	following the ORIGIN line

LocusLink

Update Annotations from LocusLink reads the html source from queries to the LocusLink database at NCBI as summarized in the following table.

Annotation	LocusLink Label
Common	Official Gene Symbol or Interim Gene Symbol plus any Alternate Symbols
Map	Position, Cytogenetic, or Chromosome
EC number	EC number
Product	Product
Phenotype	Phenotype

UniGene

Update Annotations from UniGene reads the html source from queries to the UniGene database at NCBI.

The common name and description are from the line immediately following the UniGene cluster ID and the species name. If only one item is shown, it is returned as the description.

The map annotation is taken from the Cytogenetic Position under the MAPPING INFORMATION heading; if Cytogenetic Position is not given, the Chromosome number is used.

NCBI has the following requirements for automated access:

- Run retrieval scripts on weekends or between 9 PM and 5 AM ET weekdays for any series of more than 100 requests.
- Refer to the NCBI disclaimer and copyright notice at <http://www.ncbi.nlm.nih.gov/About/disclaimer.html>

See http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html for more information.)

Building a Simplified Ontology

The Build Simplified Ontology tool hierarchically groups genes into meaningful biological categories (gene lists), based on the Gene Ontology Consortium Classifications (GO). To form these groups, GeneSpring's ontology tool parses all of the annotations in the genome. It then assigns each gene to one or more ontology groups based on this analysis. The Build Simplified Ontology constructor builds over 300 biologically meaningful gene lists that can be compared, merged or browsed. These lists can be further used to annotate clusters and cross-reference new gene lists.

Note: You cannot rename these gene lists, but you can update them.

To Build a Simplified Gene Ontology list

1. Select **Annotations > Build Simplified Ontology**.
2. Enter a name for the new simplified ontology folder, or leave the default name to overwrite the existing simplified ontology list.
3. Click **OK**. The new Simplified Ontology list appears in the Gene Lists folder.

To Make Gene Lists From Properties

To create lists based on annotations, see “Making Lists from Properties” on page 6-12.

Building Homology Tables

To build a homology table, see “The Homology Tool” on page 6-24.

Statistical Analysis (ANOVA)

Statistical Analysis (ANOVA) is a filter tool that statistically compares mean expression levels between two or more groups of samples. The object is to find the set of genes for which the specified comparison shows statistically significant differences in the mean normalized expression levels as interpreted according to your current interpretation mode (logarithm, ratio or fold change) across all the groups. This comparison is performed for each gene, and the genes with the most significant differential expression (smallest p-value) are returned.

Filtering genes based on a one-sample t-test of the mean expression level across repeats or replicates versus a reference value can be done by selecting “t-test p-value” as the filter criteria in Expression Percentage Restriction.

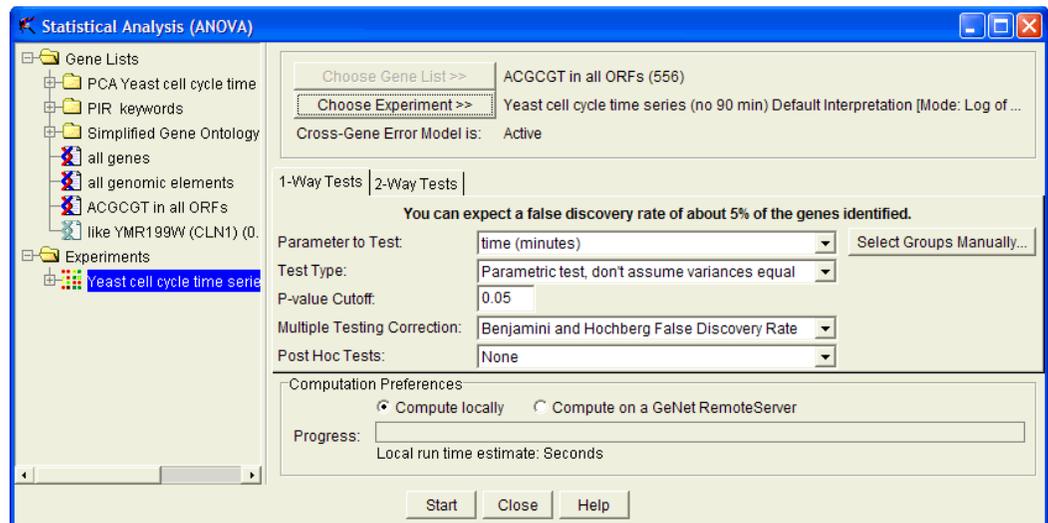


Figure 1-1 The Statistical Analysis (ANOVA) window

To perform ANOVA analysis:

1. Select **Tools > Statistical Analysis (ANOVA)**.
2. Select a gene list from the navigator and click **Choose Gene List**.
3. Select an experiment from the navigator and click **Choose Experiment**.
4. Click the tab for **1-Way Tests** or **2-Way Tests**.
5. Specify the appropriate options for your analysis.

For detailed information on the options available on each tab, see “1-Way ANOVA” on page 6-34 and “2-Way ANOVA” on page 6-43.

6. Specify whether to run this operation locally or on a GeNet Remote Server.
7. Click **Start**.

1-Way ANOVA

Use 1-Way ANOVA to filter out genes that do not vary significantly across different groups with multiple samples. This allows you to find those genes that exhibit important changes between various conditions of the experiment. This comparison is performed for each gene, and the genes with sufficiently small p-values are returned. Comparisons can be performed with parametric or non-parametric methods. The parametric comparison for two groups is known as Student's two-sample t-test. For multiple groups, this is known as one-way analysis of variance (ANOVA). You can specify whether to assume within-group variances are equal across all groups.

Calculations without the assumption of equality of variances are done using Welch's approximate t-test and ANOVA. Non-parametric comparisons are also available, corresponding to the Wilcoxon two-sample test (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups.

Figure 1-2 1-Way ANOVA options

To perform one-way ANOVA:

1. From the **Parameter to Test** pull-down list, select the parameter on which to base your comparison. To select a group of parameters, see “Selecting Groups Manually” on page 6-35.
2. Select the type of test to perform. There are four testing options:
 - **Parametric test, assume variances equal**—Filter based on the results of a Student’s two-sample t-test for two groups or a one-way analysis of variance (ANOVA) for multiple groups.
 - **Parametric test, don’t assume variances equal**—Filter based on the results of an ANOVA or Welch’s approximate t-test for two groups. This is the most appropriate test for standard experiments in which the global error model is not turned on or should not be used in the analysis.
 - **Parametric test, use all available error estimates**—Filter based on the variances estimated by the cross-gene error model. If the cross-gene error model is not turned on, this test is equivalent to the Parametric test.
 - **Non-Parametric test**—Filter based on the rank of each sample, rather than the expression level. Non-parametric comparisons use the Wilcoxon two-sample rank test (also known as the Mann-Whitney U test) for two groups, and the Kruskal-Wallis test for multiple groups. This test is most successful if you have more than five replicate samples in each group.
3. Select a P-value cutoff for genes that pass the filter. P-values are the probability of a false positive, indicated by a number between zero and one.
4. Select a type of multiple testing correction. The available options are described below.

5. Select a type of post-hoc test to perform, if desired. Post-hoc tests are described in more detail below.

Selecting Groups Manually

To make groups that are defined by two or more parameters, or to make groups that correspond to a subset of the values for a given parameter, click **Select Groups Manually**.

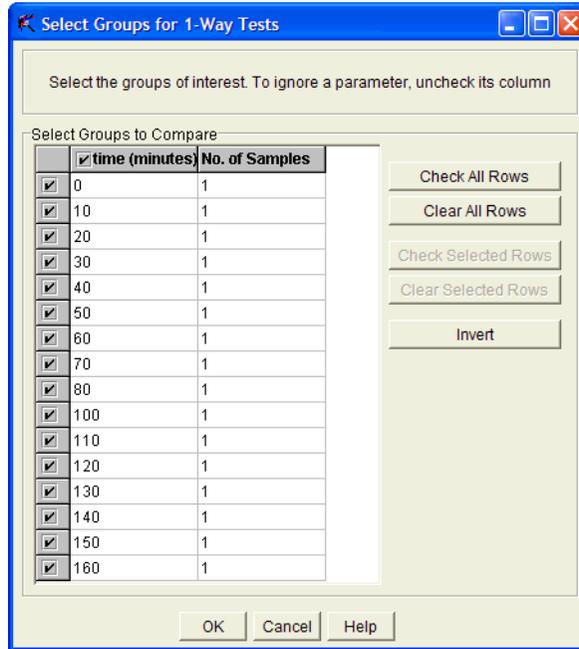


Figure 6-3 The Select Groups window for 1-way ANOVA testing

This screen features a table with a column for each experimental parameter and a row for each condition to compare.

Uncheck the box in a column’s header to ignore groupings based on that parameter. When you do this, the table is dynamically updated to reflect the change. The number of rows decreases and the number of samples associated with each condition increases.

Define the conditions to compare by checking or unchecking the box in a given row. The **Check All/Clear All** buttons allow you to check or uncheck all rows. The **Invert** button checks all unchecked rows, and clears all checked rows.

For Each Gene

GeneSpring does the following separately for each gene:

Let i index over the G groups formed by distinct levels of the comparison parameter.

Let X_{ik} be the expression values, with k running over the replicates for each situation, interpreted according to the current interpretation (ratio, log of ratio, fold change).

Let

N_i = the number of non-missing data values for each group,

$$\bar{X}_i = 1/N_i \sum_{k=1}^{N_i} X_{ik} \text{ be the group means, and}$$

$$SS_i = \sum_{k=1}^{N_i} (X_{ik} - \bar{X}_i)^2 \text{ be the within-group sum of squares.}$$

In all calculations, missing values (No Data) or (NaN) are left out of the sums, not propagated. If any of the N_i are zero, drop that parameter level from the analysis, and readjust G accordingly. If G is not at least 2, exit (p-value=1).

Parametric Test, Variances Assumed Equal

For a parametric test, with variances assumed equal, compute:

$$\bar{X} = \frac{\sum_{i=1}^G N_i \bar{X}_i}{\sum_{i=1}^G N_i} \text{ the overall mean,}$$

$$BSS = \sum_{i=1}^G N_i (\bar{X}_i - \bar{X})^2 \text{ the between-groups sum of squares,}$$

$d_1 = G - 1$ the numerator degrees of freedom,

$$BMS = BSS/d_1 \text{ the between-groups mean square,}$$

$$WSS = \sum_{i=1}^G SS_i \text{ the pooled within-group sum of squares,}$$

$$d_2 = \sum_{i=1}^G (N_i - 1) \text{ the denominator degrees of freedom,}$$

if d_2 is not greater than zero, then exit (p-value=1).

$$WMS = WSS/d_2 \text{ the within-groups mean square, and}$$

$$F = BMS/WMS \text{ the F-ratio statistic}$$

if $WMS = 0$ then make F is treated as arbitrarily large (p-value = 0).

The p-value is calculated by looking up F in the upper tail probability of an F distribution with d_1 and d_2 degrees of freedom.

Parametric Test, Variances Not Assumed Equal

For the parametric test without assuming variances equal:

First check that each group has N_i greater than or equal to 2 and SS_i greater than 0. If not, remove it from consideration and recompute G . If G is not at least 2, exit (p-value=1). (This reflects the more stringent requirements of not assuming the variances equal - if the variance estimate is pooled, replicates are only needed for at least one group, if variances are separately estimated then replicates are needed for each group.)

Then compute:

$$w_i = N_i \left(\frac{N_i - 1}{SS_i} \right) \text{ the group weights}$$

$$W = \sum_{i=1}^G w_i \text{ the sum of weights}$$

$$\tilde{X} = \frac{\sum_{i=1}^G w_i \bar{X}_i}{W} \text{ the weighted mean}$$

$$BSS = \sum w_i (\bar{X}_i - \tilde{X})^2 \text{ the between-groups sum of squares}$$

$$d_1 = G - 1 \text{ the numerator degrees of freedom}$$

$$BMS = BSS / d_1 \text{ the between-groups mean square}$$

$$Z = \frac{1}{G^2 - 1} \sum_{i=1}^G \left(1 - \frac{w_i}{W}\right)^2 / (N_i - 1)$$

$$d_2 = \frac{1}{3Z} \text{ the denominator degrees of freedom}$$

if d_2 is not greater than zero, then exit (p-value=1).

$$WMS = 1 + 2(G - 2)Z \text{ the within-group mean square}$$

$$W = BMS / WMS \text{ the test statistic}$$

The (approximate) p-value is calculated by looking up W in the upper tail probability of an F distribution with d_1 and d_2 degrees of freedom. Note that d_2 will not, in general, be an integer.

Nonparametric Analysis

For the nonparametric analysis:

Replace each X_{ik} by R_{ik} , their rank out of all of the $\{X_{ik}\}$ for the gene. Perform the same analysis as for parametric test with variances equal. P-values are approximate but asymptotically accurate.

Multiple Testing Corrections

If you rely on the nominal p-value when testing the statistical significance of group comparisons for many genes, a significant number of genes pass the filter by chance alone. For example, if you test 10,000 genes for reliable changes between groups at significance level 0.05, (assuming the tests are independent) you would expect to misidentify about 500 genes as significant, even when there is no real difference in gene expression. Even if you identify 1,000 genes showing significant behavior by this approach, half of the genes on the list appear by chance, which lessens the value of the list. Multiple testing corrections adjust the individual p-value to account for this effect.

Suppose the p-value cutoff is α and the number of genes being tested is N . The first three procedures (Bonferroni, Holm, and Westfall and Young) control the family-wise error rate (FWER), which is the overall probability of obtaining even a single false positive test to be no more than α . This is a very strong criterion, but may be so strong for large lists of genes that no genes are identified as significant. The Benjamini and Hochberg test controls the false discovery rate, defined as the proportion of genes expected to be identified by chance relative to the total number of genes called significant.

- **Bonferroni**—The Bonferroni multiple testing correction, based on Bonferroni's inequality, limits the chance of a false positive results to be no more than α by multiplying each nominal p-value by N (with a maximum of 1). This process controls the FWER, and the expected number of genes by chance is α .
- **Bonferroni step-down (Holm)**—The Holm step-down adjustment computes the most significant p-value, and whether it meets the α cutoff after multiplying by N . If that gene is found to be significant, the next-most significant gene is considered, but the gene that was found significant is removed from the multiple-testing, so the multiple-testing adjustment is now based on $N - 1$. This process is continued as long as genes pass the successive tests. This process controls the FWER, and expected number of genes by chance is α .
- **Westfall and Young permutation**—This procedure estimates the significance levels of each test by a nonparametric permutation calculation based on the distribution of the significance levels across all possible reassignments of samples to groups. For small numbers of permutations, all permutations are examined. If there are more than 1000 possible permutations, 1000 of them are selected randomly. P-values are evaluated with respect to this distribution using a step-down procedure as in the Holm procedure. This procedure controls the FWER, and the expected number of genes by chance is α . This test accounts for the dependence structure between genes, and should give a more powerful test than the Bonferroni or Holm procedure. However, the permutation process takes much longer to calculate.
- **Benjamini and Hochberg false discovery rate**—In contrast to the above procedures, the Benjamini and Hochberg procedure controls the false discovery rate (FDR), defined as the proportion of genes expected to occur by chance (assuming genes are independent) relative to the proportion of identified genes. Expected number of genes by chance is α times the number of tests found significant after applying this correction. There is no way to calculate this in advance, so the statement about the number

expected simply says expected number of genes by chance is $100\alpha\%$ of the genes identified. This procedure provides a good balance between discovery of significant genes and protection against false positives, since occurrence of the latter is held to a small proportion of the list, and is probably the best choice of multiple-testing correction for most situations.

Post-Hoc Tests

Performing a 1-way ANOVA results in a list of genes with p-values below the specified cutoff. These are the genes with significant differential expression across the specified groups. If testing has been performed across more than 2 groups, post hoc tests can be used to identify the specific groups in which significant differential expression occurs.

To perform post hoc tests, specify which post hoc test to use in the post hoc test pull-down menu. The available choices are Tukey and Student-Newman-Keuls. Pairwise comparisons between all groups are performed; group comparisons resulting in a p-value below the specified cutoff are displayed in the output. Post hoc tests can only be performed if more than 2 groups are defined by the chosen testing parameter, or if more than 2 groups are chosen manually.

Tukey

Suppose SGC with more than 2 groups has been performed (ANOVA or Kruskal-Wallis), and that some genes have passed the cutoff. The following calculations are done separately for each gene:

Let $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ be the ordered group means (ascending order). We perform pairwise group mean comparisons in the following manner:

k vs. 1, k vs. 2, ..., k vs. $k - 1$, then $k - 1$ vs. 1, $k - 1$ vs. 2, ..., $k - 1$ vs. $k - 2$, ending with 2 vs. 1.

Do not perform unnecessary tests. i.e., if there is no significant difference between a pair, do not test any “closer” pairs. Each test is performed as described below.

Parametric Tests (ANOVA)

If an ANOVA was performed, for comparing group means \bar{X}_a and \bar{X}_b , compute:

$$SE = \sqrt{\frac{WMS}{n}} \text{ if groups are of equal sizes,}$$

$$SE = \sqrt{\frac{WMS}{2} \left(\frac{1}{n_a} + \frac{1}{n_b} \right)} \text{ if unequal sizes.}$$

Where n , n_a , and n_b are the corresponding group sizes (number of samples) and WMS is the within-group mean square (from the ANOVA calculations).

$$\text{Then compute } q = \frac{\bar{X}_b - \bar{X}_a}{SE}.$$

Compare this value to the critical value $q_{\alpha,df,k}$ where df is the error degrees of freedom (from the ANOVA calculation) and k is the total number of groups. If q is larger, consider the two means significantly different.

If any of the group sizes for the gene are 1 (i.e., there are not replicate samples for every group), do not perform post-hoc tests for that gene.

Nonparametric Test (Kruskal-Wallis)

If the non-parametric option was chosen, GeneSpring performs a non-parametric Tukey test.

$$SE = \sqrt{\frac{n(nk)(nk + 1)}{12}}$$

where n is group size and k is the number of groups.

Rank order all the data and compute rank sums for each group. Order the rank sums and compute q as before, using the rank sums instead of means. Compare q to the critical value

$q_{\alpha, \infty, k}$.

Student-Newman-Keuls

All calculations here are the same as for Tukey. The only difference is in which critical value to compare q . When testing group a vs. group b, compare q to $q_{\alpha,df,p}$ where p is the number of means (inclusive) in the range being tested. For example, if comparing group 2 to group 4, $p = 3$.

For the non-parametric version, replace k with p .

Viewing Post-Hoc Test Results

After 1-way ANOVA has been performed, the 1-Way Post Hoc Testing Results window appears.

Summary by Gene

The Results Summary by Gene tab displays the mean expression level by group for each significant gene. For each gene, the coloring indicates which groups differ significantly from the others. Groups of the same color show no significant difference. Groups of different colors differ significantly from each other.

Note: Occasionally genes will pass the ANOVA cutoff, but show no significant difference across the groups (all groups the same color). This is due to the fact that ANOVA is a more powerful test than the post hoc tests.

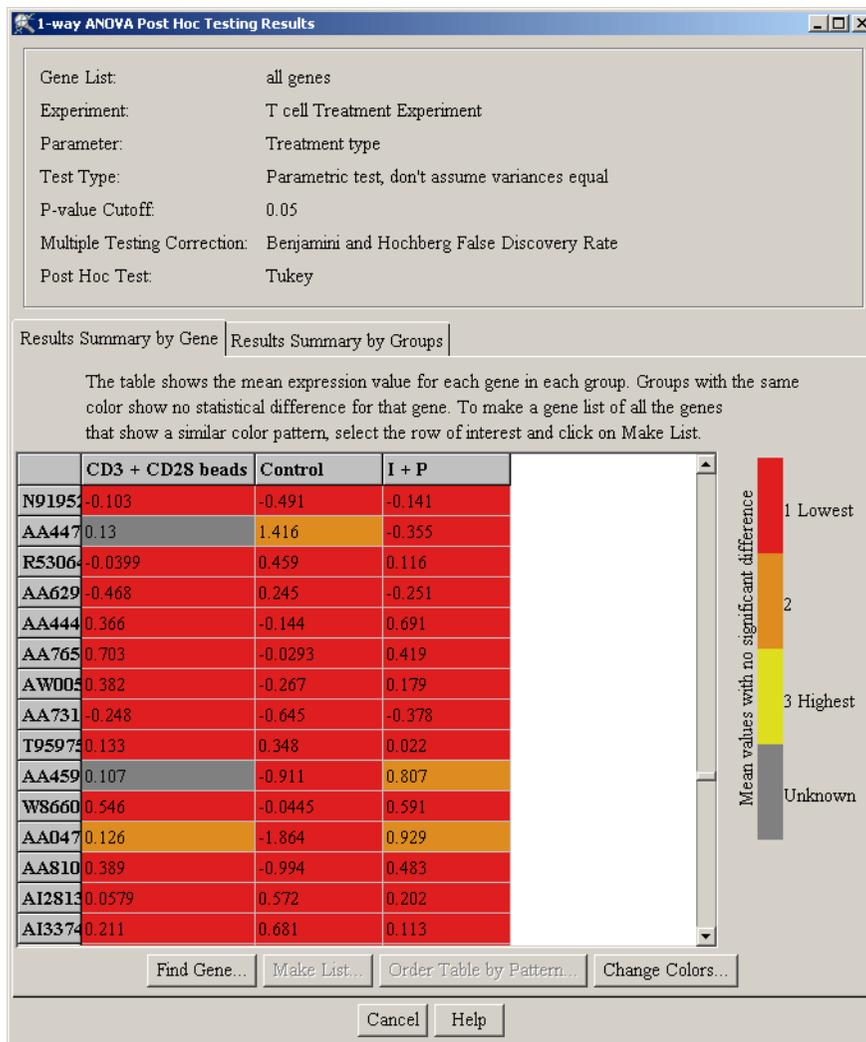


Figure 6-4 Post Hoc results summary by gene

Summary by Groups

The Results Summary by Groups tab displays a matrix, with rows and columns indexed by parameter values; each cell corresponding to a combination of groups. The numbers in the lower half of the matrix represent the number of genes that differ significantly between the groups; the numbers in the upper half are the genes which show no significant difference.

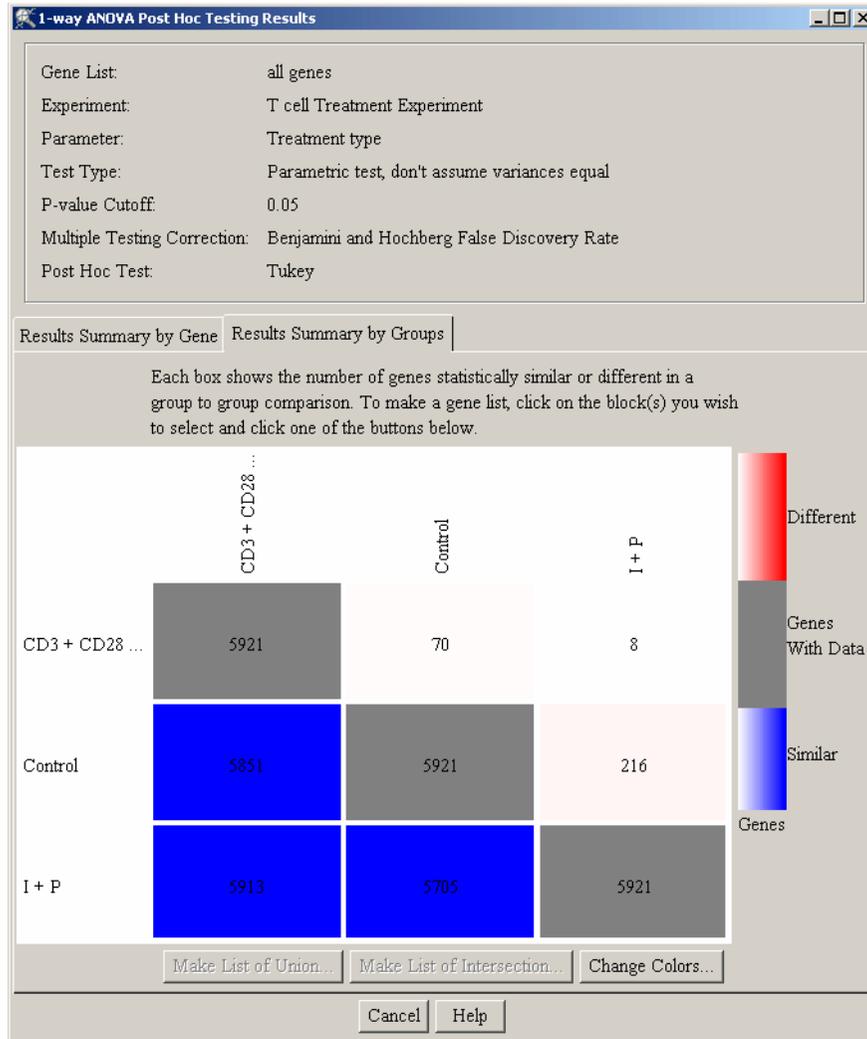


Figure 6-5 Post Hoc results summary by group

2-Way ANOVA

2-way ANOVA tests genes for significant differences across groups defined by 2 parameters. This test is appropriate to use for a 2-way design where the groups to be compared are defined by 2 parameters. A 2-way design is one that can be thought of as a matrix, with the rows indexed by the values of one parameter, and the columns indexed by the values of a second parameter. Each cell then represents the number of replicates in that particular group. Ideally, each cell should have an equal number of replicates; this is called a *balanced design*. GeneSpring can also perform 2-way ANOVA for proportional designs; you will not be able to perform 2-way ANOVA for unbalanced designs which are not proportional.

Performing a 2-way ANOVA will test for the effect of each parameter, as well as the interaction between them, simultaneously. For each gene, 3 p-values are produced, one for each parameter and one for the interaction term. Genes for which any of the p-values is less than the specified cutoff are returned. From the results window, several options for creating gene lists are available.

Figure 1-6 2-Way ANOVA options

The following options are available:

- **First Parameter to Test**—Choose a parameter
- **Second Parameter to Test**—Choose the second parameter to compare
- **Test Type**—Specify whether to perform a parametric test assuming variances equal or a non-parametric test
- **P-value Cutoff**—Default is 0.05
- **Multiple Testing Correction**—Available options are Bonferroni, Bonferroni Step-Down (Holm), Westfall & Young Permutation (slow), Benjamini & Hochberg False Discovery Rate, or None.

Selecting Groups Manually

To select groups manually, click **Select Groups Manually**.

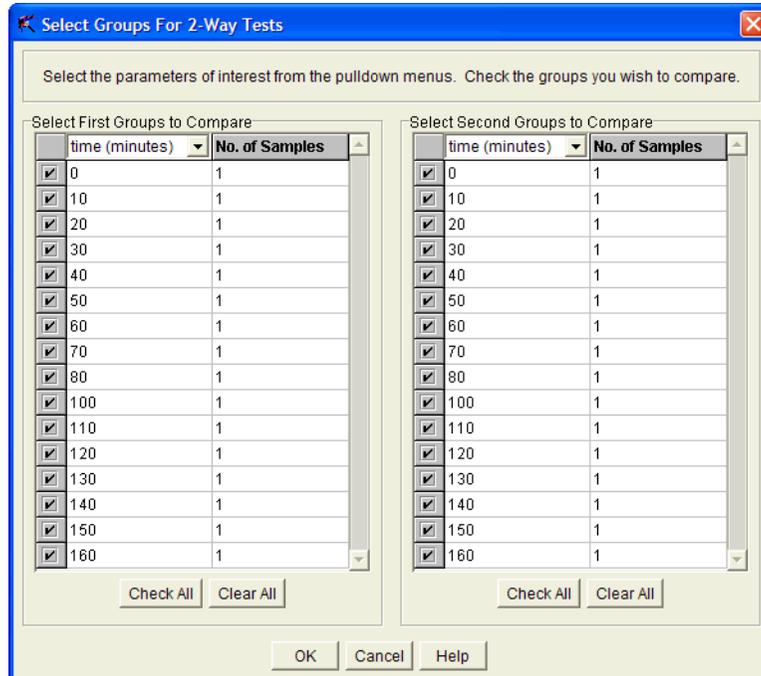


Figure 6-7 The Select Groups window for 2-way ANOVA testing

This screen features two tables. Each has one column with a pull-down menu to select the desired experimental parameter, and a row for each condition to compare.

You cannot manually define conditions for 2-way tests, but you can choose to ignore certain levels of the selected parameter by unchecking the appropriate rows. The **Check All/Clear All** buttons allow you to check or uncheck all rows.

2-Way ANOVA Results

From this screen, you can do the following:

- **Copy to Clipboard**—Copy the results in this screen to the clipboard. You can paste these results into a spreadsheet program or text editor.
- **Save Lists**—Save the results as a gene list or lists.
- **Display in Venn Diagram**—Display the results in a Venn diagram in the main GeneSpring window.

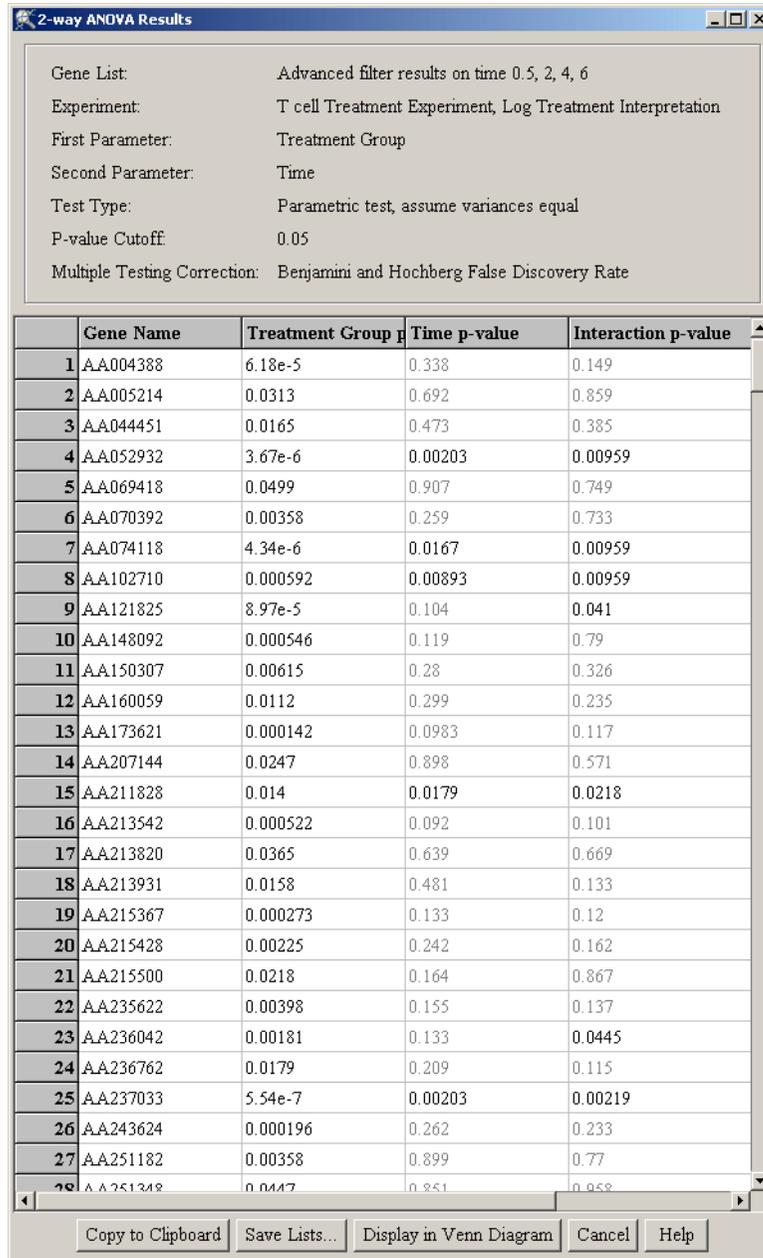


Figure 6-8 2-way ANOVA results window

Details on 2-Way ANOVA

Let A and B be the two factors (parameters) chosen by the user. Assume we are looking at a single gene and use the following notation throughout:

Factor A has a levels, indexed by i .

Factor B has b levels, indexed by j .

There are thus ab cells of data, each containing at least one value.

Standard Parametric 2-way ANOVA

This test assumes equal variances and equal or proportional replication.

Case 1 (equal replication)

All cells (groups defined by combinations of the factor levels) have the same number of replicates, say r . Thus there is a total of abr samples.

Let x_{ijk} represent the k^{th} replicate (sample) in level i of factor A and level j of factor B .

Let A_i = sum of all observations in level i of factor A

$$= \sum_i \sum_k x_{ijk}$$

Let B_j = sum of all observations in level j of factor B

$$= \sum_i \sum_k x_{ijk}$$

Let $(AB)_{ij}$ = sum of all observations in level i of factor A and level j of factor B

$$= \sum_k x_{ijk}$$

$$\text{Let } C = \frac{\left(\sum_i \sum_j \sum_k x_{ijk} \right)}{abr}$$

We can now compute the various sums of squares terms:

Total sum-of-squares:

$$SS(\text{total}) = \left(\sum_i \sum_j \sum_k x_{ijk}^2 \right) - C$$

Factor A sum-of-squares:

$$SS(A) = \frac{\sum A^2}{rb} - C$$

Factor B sum-of-squares:

$$SS(B) = \frac{\sum B_j^2}{ra} - C$$

Interaction sum-of-squares:

$$SS(AB) = \frac{\sum_i \sum_j (AB)_{ij}^2}{r} - C - SS(A) - SS(B)$$

Error sum-of-squares (a.k.a. within-group SS):

$$SS(error) = SS(total) - SS(A) - SS(B) - SS(AB)$$

Now compute the mean sums of squares:

$$MSS(A) = \frac{SS(A)}{a-1}$$

$$MSS(B) = \frac{SS(B)}{b-1}$$

$$MSS(AB) = \frac{SS(AB)}{(a-1)(b-1)}$$

$$MSS(error) = \frac{SS(error)}{abr - ab}$$

Finally, compute F-ratios:

$$F = \frac{MSS(A)}{MSS(error)}$$

$$F = \frac{MSS(B)}{MSS(error)}$$

$$F = \frac{MSS(AB)}{MSS(error)}$$

Each of these should be compared to the upper tail probability of an F distribution with numerator and denominator degrees of freedom given by the corresponding denominators above.

Case II - Proportional Replication

We first check for proportional cell sizes. Let n_{ij} = # of replicates at level i of A and level j of B , and let N be the total number of replicates in all groups = $\sum_i \sum_j n_{ij}$.

$$\text{If } n_{ij} = \frac{\left(\sum_i n_{ij}\right)\left(\sum_j n_{ij}\right)}{N} \text{ for each } n_{ij}, \text{ then we have } \textit{proportional replication}.$$

In this case, all computations are the same as before, with appropriate changes. In particular, the index k in all summations will now go from 1 to n_{ij} , instead of 1 to r .

Let A_i = sum of all observations in level i of factor A

$$= \sum_i \sum_k x_{ijk}$$

Let B_j = sum of all observations in level j of factor B

$$= \sum_i \sum_k x_{ijk}$$

Let $(AB)_{ij}$ = sum of all observations in level i of factor A and level j of factor B

$$= \sum_k x_{ijk}$$

$$\text{Let } C = \frac{\left(\sum_i \sum_j \sum_k x_{ijk} \right)}{N}$$

We can now compute the various sums of squares terms:

Total sum-of-squares:

$$SS(\text{total}) = \left(\sum_i \sum_j \sum_k x_{ijk}^2 \right) - C$$

Factor A sum-of-squares:

$$SS(A) = \sum_i \left(\frac{A_i^2}{\sum_j n_{ij}} \right) - C$$

Factor B sum-of-squares:

$$SS(B) = \sum_j \left(\frac{B_j^2}{\sum_i n_{ij}} \right) - C$$

Interaction sum-of-squares:

$$SS(AB) = \sum_i \sum_j \left(\frac{(AB)_{ij}^2}{n_{ij}} \right) - C - SS(A) - SS(B)$$

$$SS(\text{error}) = SS(\text{total}) - SS(A) - SS(B) - SS(AB)$$

Compute mean sums of squares using the following degrees of freedom:

Factor A df = $a - 1$

Factor B df = $b - 1$

Interaction (AB) df = $(a - 1)(b - 1)$

Error df = $N - ab$

Divide each SS by appropriate df to obtain mean SS. F ratios and P-values are then computed as before.

Case III - No Replicates (one per cell)

We can still perform 2-way ANOVA in this case, but it is not possible to test for an interaction effect. The calculations are essentially the same as before.

The total number of replicates is $N = ab$.

$$C = \frac{\left(\sum_i \sum_j x_{ij}\right)^2}{N}$$

$$SS(total) = \sum_i \sum_j x_{ij}^2 - C$$

$$SS(A) = \frac{\sum_i \left(\sum_j x_{ij}\right)^2}{b} - C \text{ degrees of freedom} = a - 1$$

$$SS(B) = \frac{\sum_j \left(\sum_i x_{ij}\right)^2}{a} - C \text{ degrees of freedom} = b - 1$$

$$SS(error) = SS(total) - SS(A) - SS(B) \text{ degrees of freedom} = (a - 1)(b - 1)$$

Mean sums of squares = sum of squares / degrees of freedom.

F ratios:

$$F = \frac{MSS(A)}{MSS(error)}$$

$$F = \frac{MSS(B)}{MSS(error)}$$

Compute p-values as before.

Case IV - Disproportional Replication

If a single cell is one value short of the number required for proportional replication, estimate the missing value using the following:

$$x_{ijk} = \frac{aA_i + bB_j - \sum_i \sum_j \sum_k x_{ijk}}{N + 1 - a - b}$$

where A_i, B_j are as before and N is the total number of data *including the missing value*.

If several cells are missing values, or more than one value is missing, apply this formula iteratively if necessary.

After missing values have been estimated, perform ANOVA calculations as above, but do not increase degrees of freedom. That is, error df should still be based on the original number of data points.

GeneSpring displays a warning message if missing values have been imported.

If there is a larger number of missing values (say $>\min(a,b)$), GeneSpring displays a warning and exits.

Non-Parametric Two-way ANOVA - Friedman's Test

This tests only for the effect of factor A, while controlling for factor B. To get a p-value for the other factor, reverse the factors (parameters). This does *not* test for interaction.

Case I - No replicates (one sample per cell)

Rank data within each of the b blocks separately. For each of the a levels of factor A, compute rank sums R_i .

$$\text{Then compute } \chi_r^2 = \frac{12}{ba(a+1)} \sum_{i=1}^a R_i^2 - 3b(a+1)$$

This statistic has its own distribution; however, we can approximate by computing:

$$F_F = \frac{(b-1)\chi_r^2}{b(a-1) - \chi_r^2}$$

and compare this to F with degrees of freedom $a-1$ and $(a-1)(b-1)$.

Case II - (equal) replicates in each cell

If there are n replicates within each cell, compute

$$\chi_r^2 = \frac{12}{ban^2(na+1)} \sum_{i=1}^a R_i^2 - 3b(na+1)$$

Compare this to the chi-square critical value with $a-1$ degrees of freedom.

The Filtering Menu

The Filtering menu allows you to apply a series of restrictions or filters to a gene list. These restrictions can apply to an entire experiment or interpretation, or to a single condition or sample. The filters include factors such as quality control, control strength, expression level constraints, sample to sample fold comparison, statistical group comparisons, and associated numbers restrictions. All restrictions applied to create a new list are saved in the notes.

The ability to restrict a gene list based on the behavior of its genes in experiments or in individual samples is an important quality control tool. You may want to remove genes with low precision, large error values, those that do not vary significantly across multiple samples, or those with expression levels that are too close to the background. Filtering genes also allows you to search for genes that are differentially expressed over two or more conditions.

There are eight basic filters and an Advanced Filtering option:

- Filter on Expression Level
- Filter on Fold Change
- Filter on Error
- Filter on Confidence
- Filter on Flags
- Filter on Data File
- Filter on Arbitrary File
- Filter on Gene List Numbers

To view a filtering window, select it from the **F**iltering menu.

The Basic Anatomy of a Filtering Window

Most of the Filtering windows are organized in the same way:

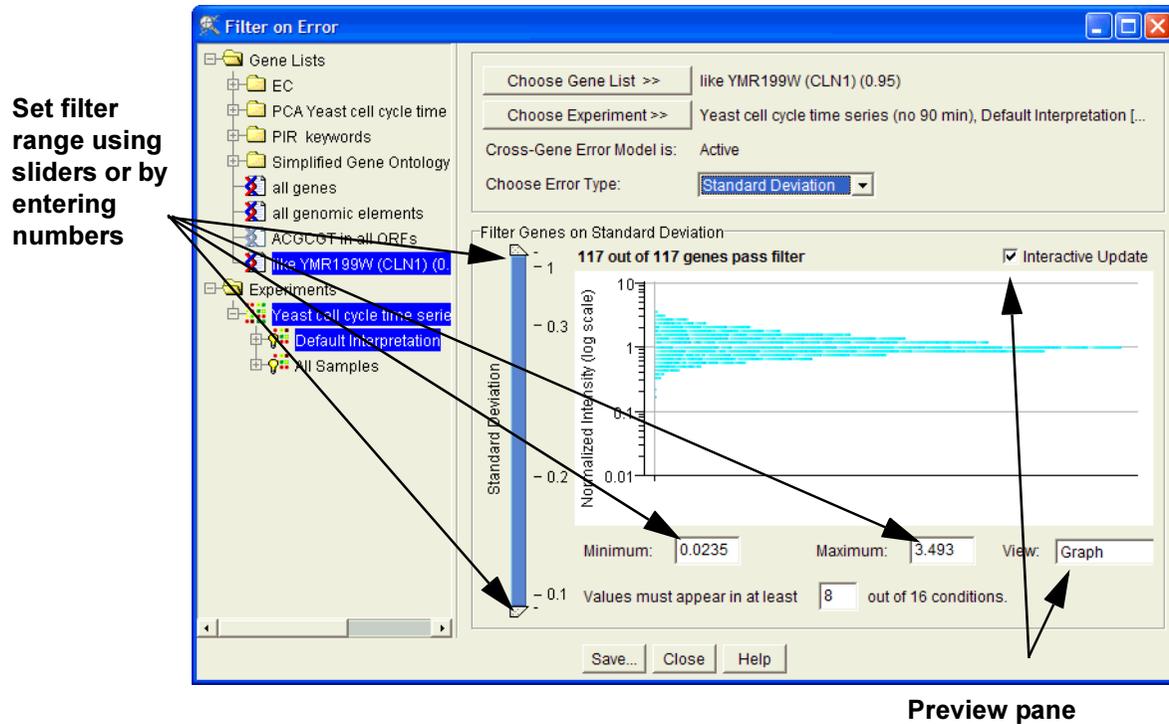


Figure 6-9 A Typical Filtering Window

Preview Pane Options

When you open a filtering window, the default view in the preview pane is based on what type of view makes the most sense for that filtering type. To link the preview display to the main GeneSpring window, select **Main Window** from the **View** menu.

The preview pane updates dynamically as you change settings for the filter. When working with large experiments, this may cause GeneSpring to respond slowly. To disable this feature, uncheck the **Interactive Update** box.

Using the Double-Ended Sliders

In filters that require you to set a minimum/maximum range, a double-ended slider appears. You can set a range either by using the sliders or by entering numbers directly in the Minimum and Maximum boxes.



Figure 6-10 A Double-Ended Slider

You can preserve the size of the specified range while changing the settings by clicking the blue bar between the sliders and dragging it to the desired position.

In some filters, the ticks on the slider are not spaced linearly or logarithmically. In these filters, the numbers are spaced so that an equal number of genes fall between each tic. This occurs since using a linear or logarithmic distribution would cause 99% of the genes to fall within three pixels of each other, making the slider impossible to use.

Note: When you enter a number in the Minimum/Maximum box, the slider is moved to that exact number. However, when you move the slider, the number shown in the Minimum/Maximum box is rounded to three digits after the decimal. At the ends of the slider this rounding may sometimes exclude the very biggest or smallest value.

Data Types for Restrictions

You can change the type of data on which to base the restriction, by choosing from a pull-down list in the applicable window. Depending on which feature you are currently using, you may have access to only some of the options in the following list.

- **Normalized Data**—Gene expression values after all normalizations have been applied. These are the default values displayed in various views and are shown in the Normalized column in the Gene Inspector. See “The Gene Inspector” on page 4-10 for details.
- **Raw Data**—Experimental data prior to application of any normalizations. This value is used as the numerator to calculate normalized values.

Note: If your computer’s default language is not English, make sure a consistent convention for decimal markers is followed.

- **Control Signal**—A value calculated from all the normalizations applied to the experiment. This value is used as the denominator to calculate normalized values.

Basic Filters

Filter on Expression Level

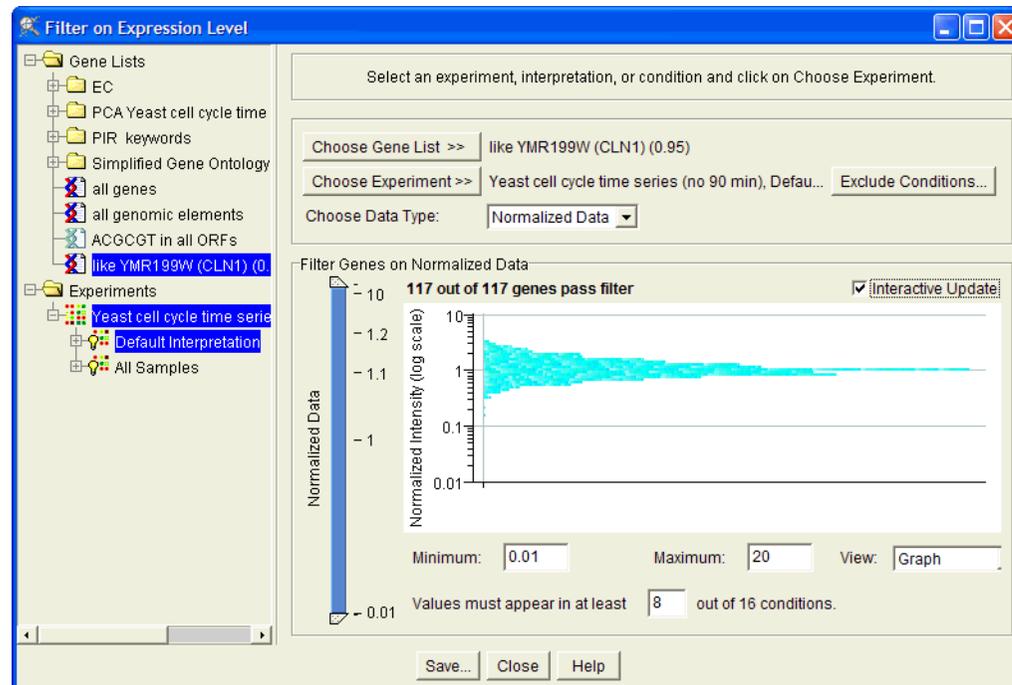


Figure 6-11 The Filter on Expression Level screen

This filter finds genes with certain values present in some of the conditions or samples in an experiment or interpretation. You can set what proportion of conditions must meet a certain threshold. For example, to eliminate genes that do not meet a specified control value at least once in the experiment, you can filter them out by setting a minimum expression value to be met in at least one condition.

To filter on Expression Level:

1. Select an experiment or condition from the navigator and click **Set Experiment**. You can also select a subset of conditions within an experiment.
2. Select the appropriate data type from the Choose Data Type menu. For more information on data types for filtering, see “Data Types for Restrictions” on page 6-53.
3. Click **Exclude Conditions...** to specify which conditions (if any) to exclude from the analysis.

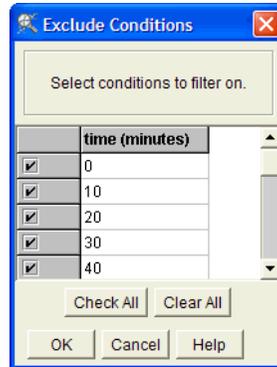


Figure 6-12 The Exclude Conditions window

By default, all conditions are selected. To exclude a condition, uncheck the box to its left. To include a condition, check the box. Click **Check All** to include all conditions, or **Clear All** to exclude all conditions.

4. Specify the following values for the filter:

- **Minimum**—the smallest gene value to allow in your list (also known as the cut-off value).
- **Maximum**—the largest gene value to allow in your list.
- **Values must appear in at least [] out of [] conditions**—the number of conditions in the total experiment where genes must meet the specified requirements. This line can refer to the whole experiment.

Filter on Fold Change

Filter on Fold Change finds genes based on a comparison of two samples or conditions. Use this tool to find fold changes in gene expression levels between two samples or conditions.

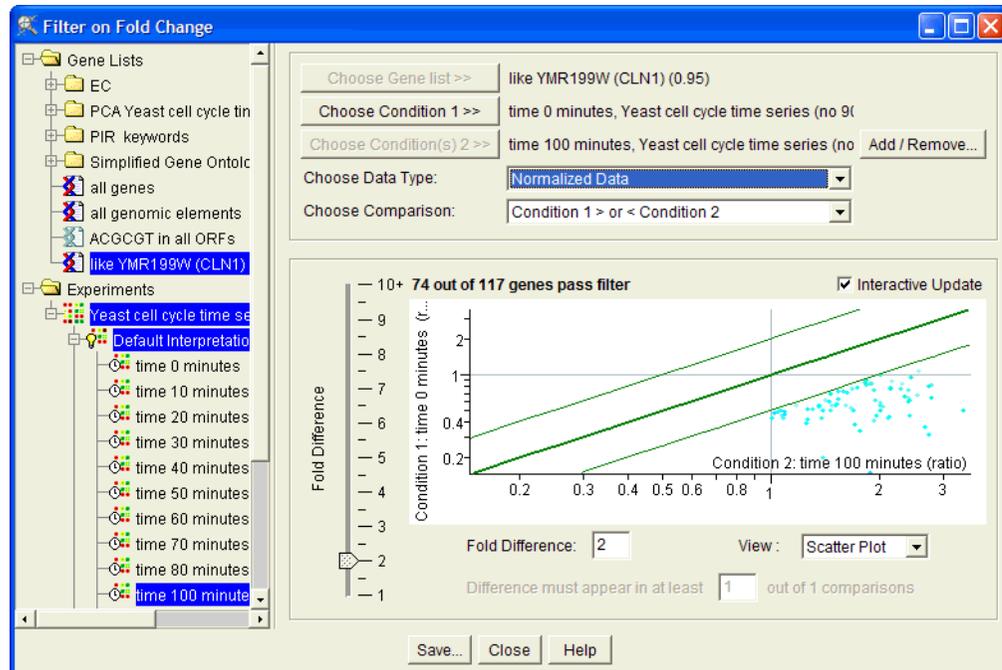


Figure 6-13 The Filter on Fold Change window

1. Select the first sample or condition and click **Choose Condition 1**.
2. Select additional samples or conditions:
 - To specify a single sample or condition, select it from the navigator and click **Choose Condition(s) 2**.
 - To select all the conditions in an experiment, select the experiment in the navigator and click **Choose Condition(s) 2**.
 - To select a pool of conditions manually (from any experiments), click **Add/Remove**. The Conditions to Filter window appears:

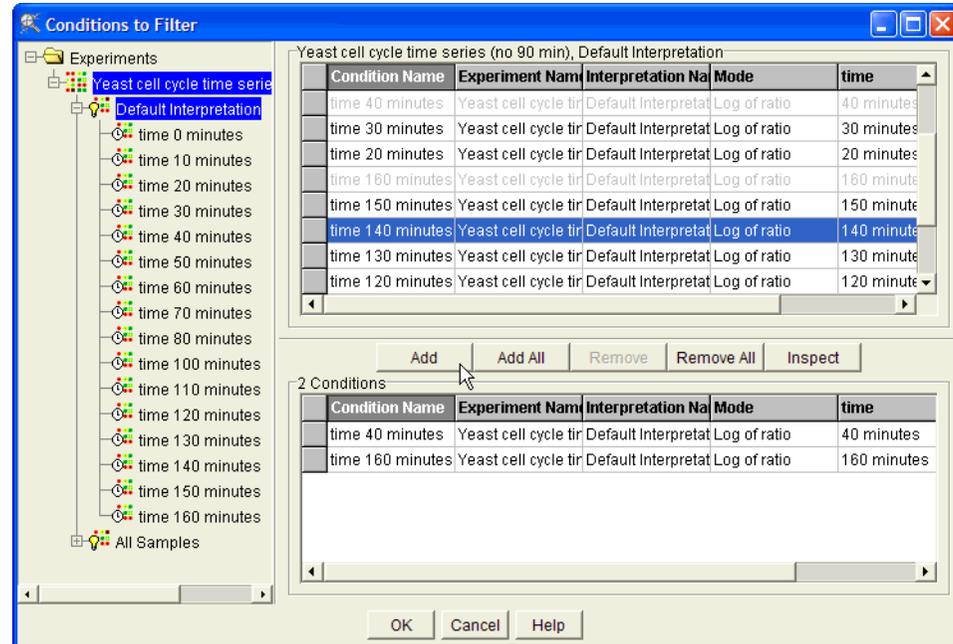


Figure 6-14 The Conditions to Filter window

To specify conditions to filter, choose an experiment in the navigator. The conditions in that experiment appear in the upper panel to the right of the navigator.

To add a condition to the filter, select it in the upper panel and click **Add**. The condition is added to the **Selected Conditions** list in the lower panel. To add all conditions from an experiment, click **Add All**.

To remove a selected condition, select it in the lower panel and click **Remove**. To remove all selected conditions, click **Remove All**.

To view a condition in the Condition Inspector, select it in either list and click **Inspect**, or double-click on a condition.

When you are done selecting conditions, click **OK**.

3. Select the appropriate data type from the **Choose Data Type** menu. For more information on data types for filtering, see “Data Types for Restrictions” on page 6-53.
4. From the **Choose Comparison** menu, choose whether you want the signal in the first sample or condition to be greater than, less than, equal to, or not equal to (greater than or less than) that in the second sample.
5. Specify a fold factor using the slider, or by entering a value in the **Fold Difference** field.
6. Enter a value in the **Difference must appear in at least [] out of [] comparisons** field.

Filter on Error

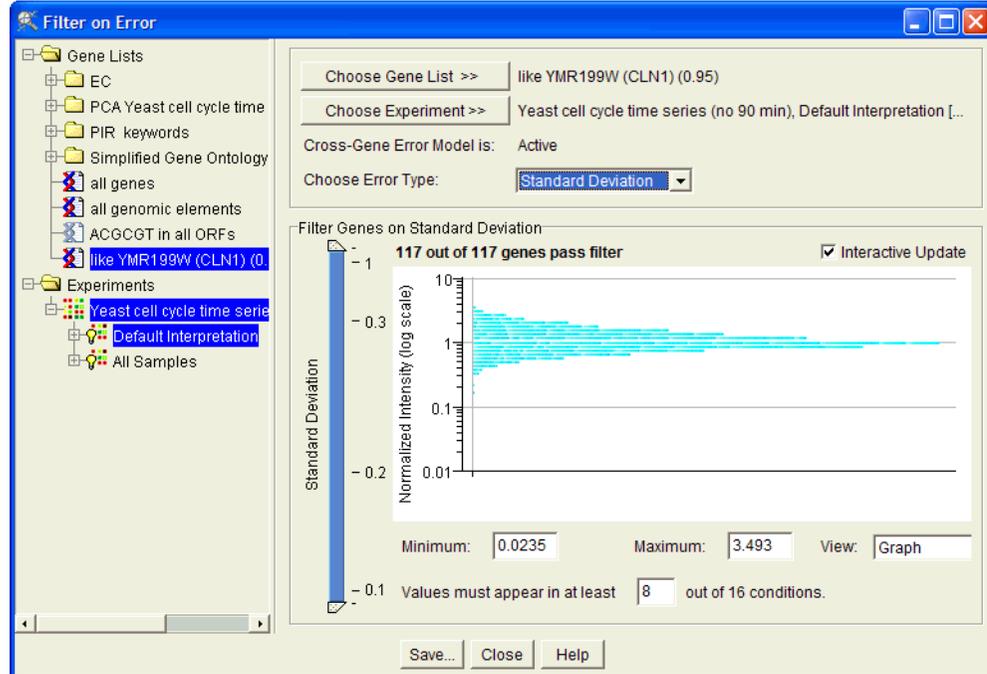


Figure 6-15 The Filter on Error window

To filter on errors:

1. Select an experiment or condition from the navigator and click **Set Experiment**. You can also select a subset of conditions within an experiment.
2. Select the error type to filter on. Available options are:
 - Standard Deviation
 - Standard Error
 - Range of Replicates
3. Specify the following values for the filter:
 - **Minimum**—the smallest gene value to allow in your list (also known as the cut-off value).
 - **Maximum**—the largest gene value to allow in your list.
 - **Values must appear in at least [] out of [] conditions**—the number of conditions in the total experiment where genes must meet the specified requirements. This line can refer to the whole experiment.

Filter on Confidence

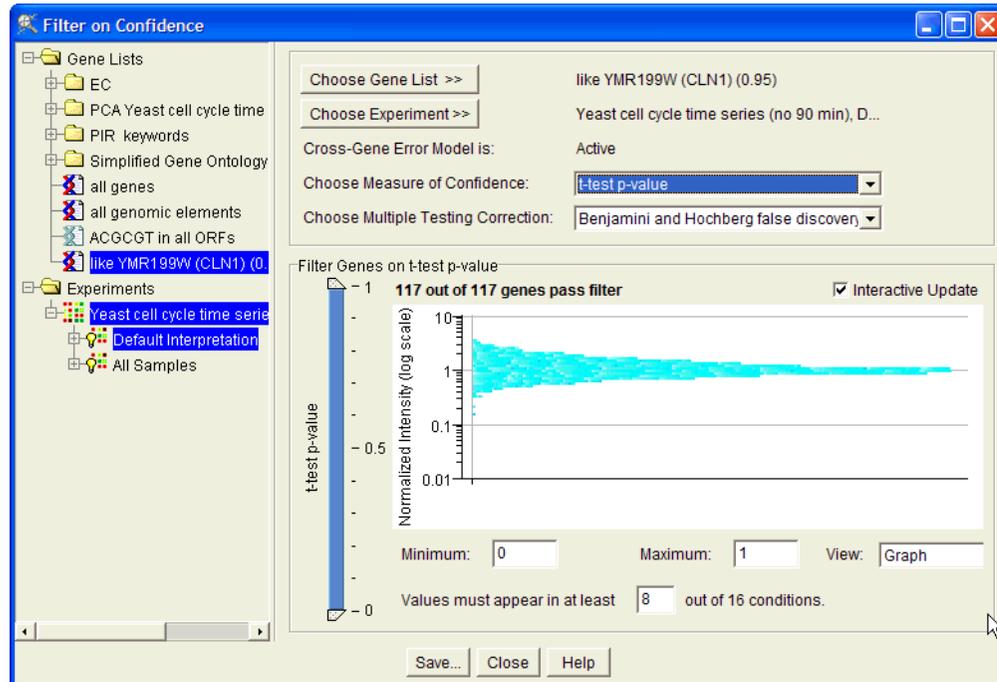


Figure 6-16 The Filter on Confidence window

To filter on confidence:

1. Select an experiment or condition from the navigator and click **Choose Experiment**. You can also select a subset of conditions within an experiment.
2. Select a measure of confidence from the Measure of Confidence menu. Available options are:
 - t-test p-value
 - Number of Replicates

For details on t-test p-values, see “The Data Table” on page 4-11.
3. Select a multiple testing correction from the Choose Multiple Testing Correction menu. Available options are:
 - Bonferroni
 - Bonferroni step down (Holm)
 - Benjamini and Hochberg False Discovery Rate
 - None
4. Specify the following values for the filter:
 - **Minimum**—the smallest gene value to allow in your list (also known as the cut-off value).
 - **Maximum**—the largest gene value to allow in your list.

- **Values must appear in at least [] out of [] conditions**—the number of conditions in the total experiment where genes must meet the specified requirements. This line can refer to the whole experiment.

Filter on Flags

GeneSpring allows you to find genes based on the data quality flags in the original data files. This option is available only if a flag column was specified in the data file when it was loaded into GeneSpring.

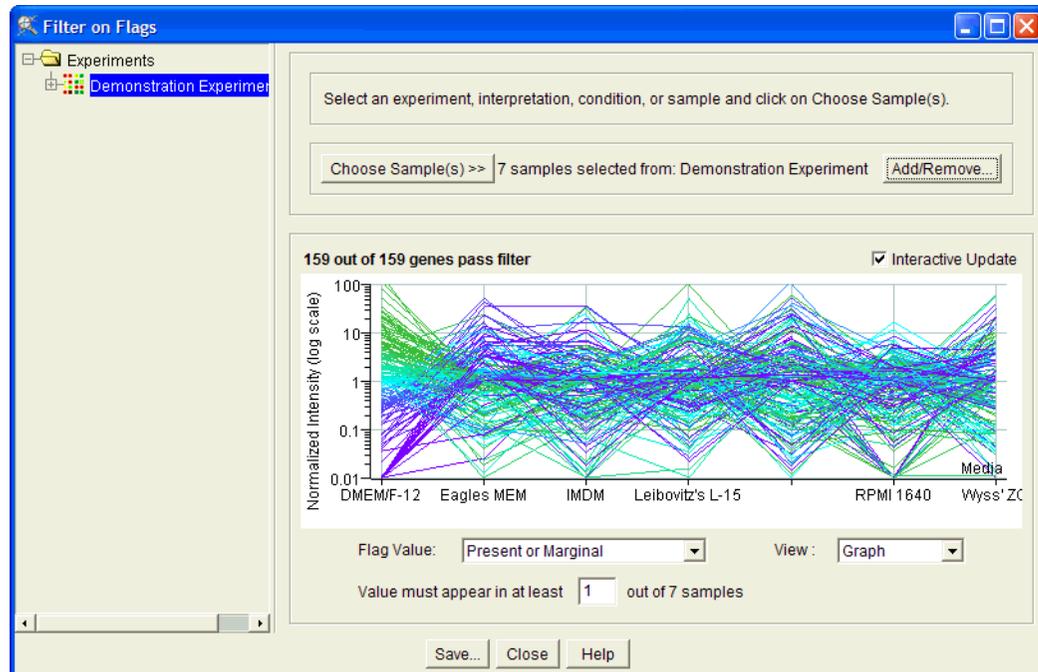


Figure 6-17 The Filter on Flags window

1. To use all the samples from an experiment, select an experiment from the navigator and click **Choose Samples**.

To select individual samples from the selected experiment, click Add/Remove. The Samples to Filter window appears.

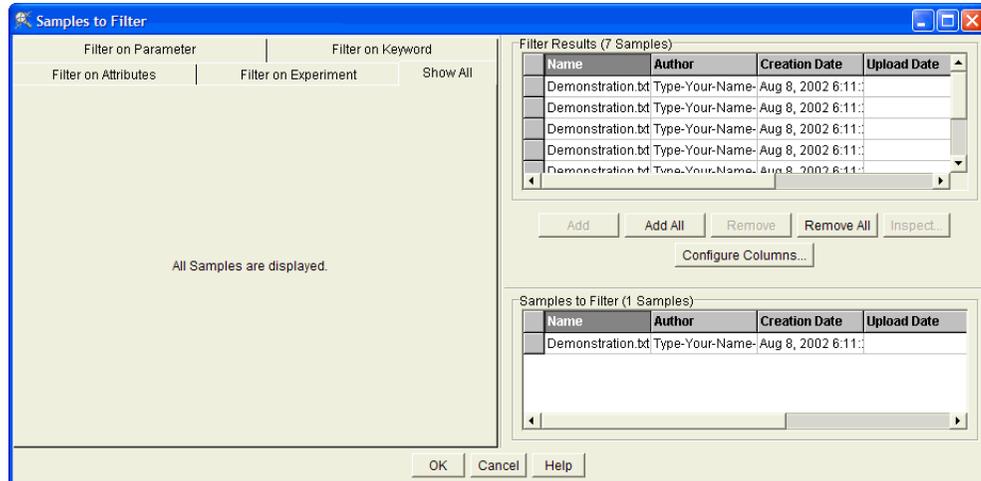


Figure 6-18 The Samples to Filter window

This window behaves exactly like the Sample Manager window. For more information, see “The Sample Manager” on page 3-23.

2. Select the desired flag value from the pull-down menu. The available options are:
 - Anything
 - Present
 - Present, Marginal
 - Present, Unknown
 - Present, Marginal, Unknown
 - Marginal
 - Absent
 - Unknown
3. Enter a value in the **Difference must appear in at least [] out of [] samples** field.

Filter on Data File

Filter on Data File allows you to filter genes based on values in a specific column of your experiment data files. For example, if you specified a flag column when you loaded your data, you can filter on Present or Marginal calls.

If your sample data files are in multiple formats, this screen appears with a separate tab for each data format. The available options on each tab are the same as the options for the standard Data File Restrictions window.

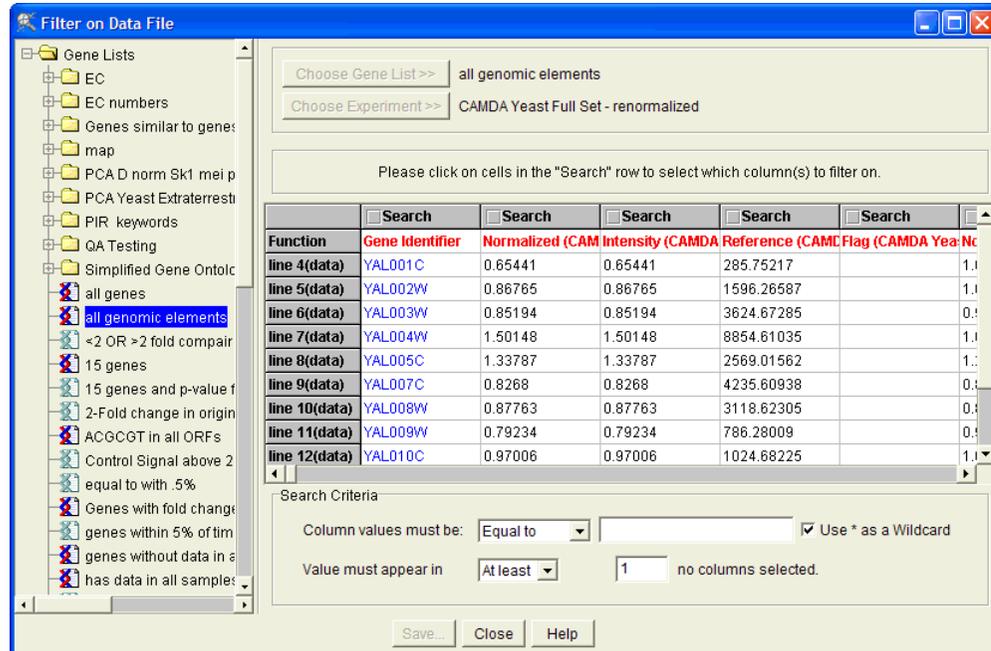


Figure 6-19 The Filter on Data File window

1. Select an experiment or interpretation from the navigator and click **Choose Experiment**.
2. To select the column or columns to search on, check the **Search** box in the header of the desired column(s) in your experiment. The column is highlighted in yellow.
3. Restrict column values by choosing a value from the **Column Values Must Be** pull-down menu and inserting a restriction value in the field provided. The available choices are:
 - Less than
 - Greater than
 - Equal to
 - Not equal to
 - Contain

For example, if you load an Affymetrix file, you can use the pull-down menu to select the **Abs/call** column and search for all entries equal to “M”. This produces a list of only marginal data.
4. Specify the number of columns in which the desired value must appear. Select an option from the **Value must appear in** pull-down menu and enter a value in the field provided. The non-editable number to the right of this box indicates the number of columns that have been selected.

If you have multiple data formats, this number reflects the total number of columns selected on all tabs.

Arbitrary File Restrictions

This option allows you to find genes based on the information in one or more columns from a selected file. You can perform only one search at a time. The selected file must have at least two columns: a column for gene identifiers and a column of some other type of data. The **Match Gene Identifier To** pull-down menu lets you specify which type of term you are using to identify each gene.

For instance, if you chose **Systematic Name or Common Name or Synonym**, GeneSpring looks for the specified identifier in any of those three columns in any of your master table of genes files. The filter then returns a list of the genes that have matching identifiers in the selected field and pass the filter in the **Search Criteria** fields.

The same search criteria are applied to every selected column; therefore all selected columns must contain the same type of information.

To perform multiple searches of columns containing different information, you must apply multiple restrictions to your data, one for each type of information.

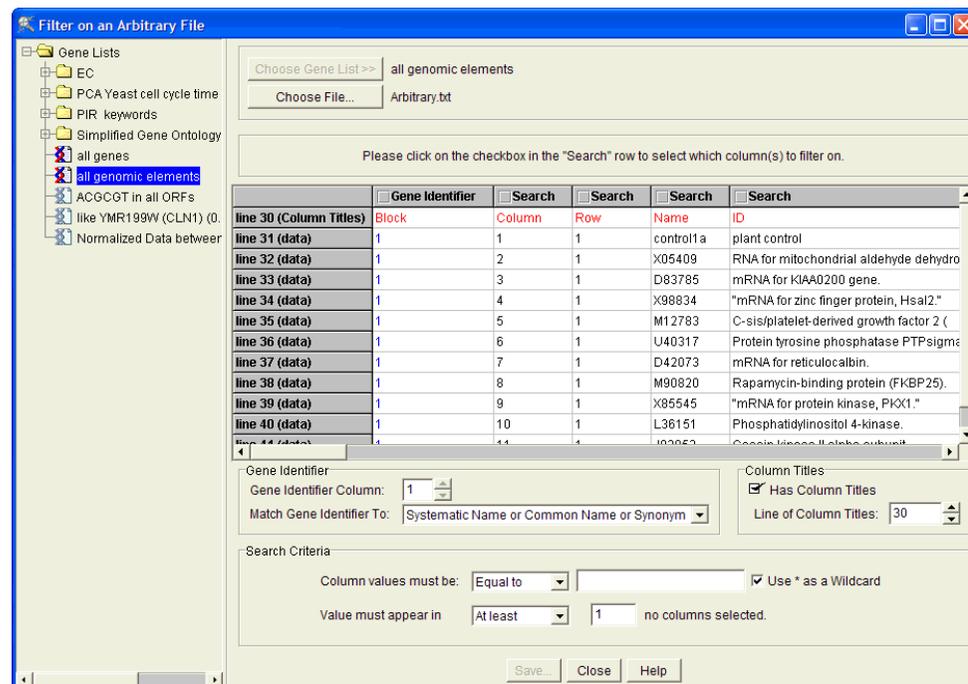


Figure 6-20 The Arbitrary File Restrictions window

1. Click **Choose File** and select the desired file from the browse menu.
This file must have a column of gene identifiers. During the loading process, GeneSpring analyzes the file to determine which column contains the Gene Identifier, and colors that column in blue. In addition, it attempts to guess whether the file has column titles, and colors that row red.
2. If GeneSpring did not select the correct column for the gene identifier, specify it in the **Column Containing Gene Identifier** field.
3. Use the **Match Gene Identifier To** menu to specify the column to which the gene identifier should be matched.

4. If the column header row chosen is incorrect, use the **First Line of Data** field to adjust the number of rows. If GeneSpring did not identify any column header row, you must first check the **Has Column Titles** box.
5. Select the column or columns in which to search by checking the **Search** box at the top of the desired column(s).
6. Restrict column values by choosing a value from the **Column values must be** pull-down menu and inserting a restriction value in the field provided. The available choices are:
 - Less than
 - Greater than
 - Equal to
 - Not equal to
 - Contain

For example, if you load an Affymetrix file, you can use the pull-down menu to select the Abs/call column and select for all entries equal to “M”. This produces a list of just the marginal data.
7. Specify the number of columns the desired value must appear in. Select an option from the **Value must appear in** pull-down menu and enter a value in the field provided. The number to the right of this box indicates the number of columns that have been selected.

Filter on Gene List Numbers

GeneSpring can filter genes according to the numbers associated with them in a gene list. When you make a new list based on a filter or similarity metric, the value used as a filter is associated with the genes on the new list. Some examples of associated numbers are correlation coefficients, p-values, fold change ratios, or in the case of a regulatory sequence search, the number of base pairs before the promoter region. Associated numbers can be found by double-clicking a gene list to bring up the Gene List Inspector.

Filtering genes by their associated numbers is helpful if you want to use this information to create a more specific list of genes. For example, you may want to find genes that are very similar to another gene (with a high correlation coefficient), or genes that are a specific distance from a promoter found using the Find Potential Regulatory Sequences tool. For details on Find Potential Regulatory Sequences see “Regulatory Sequences” on page 6-18.

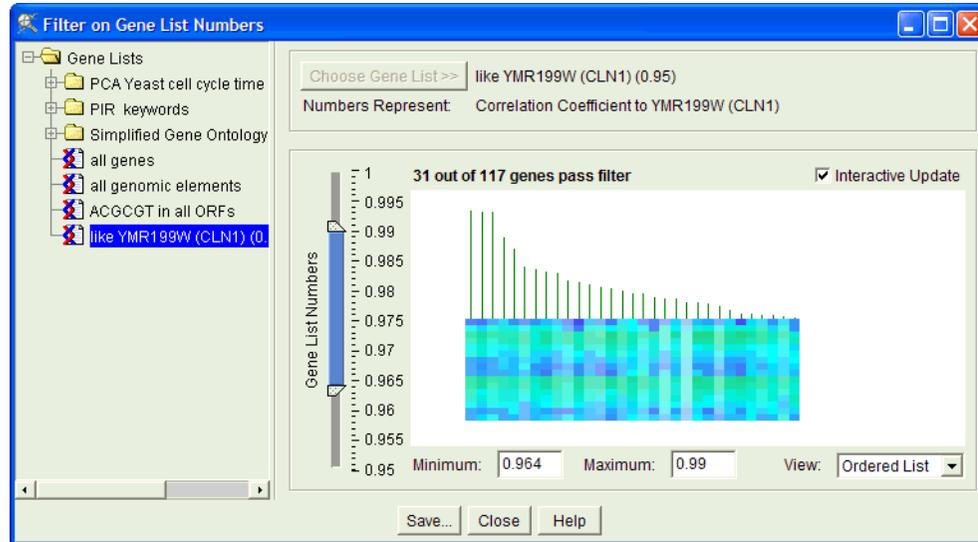


Figure 6-21 The Filter on Gene List Numbers window

To filter on gene list numbers:

1. Select a gene list with associated numbers from the navigator and click **Choose Gene List**.
2. Use the double-ended slider or enter minimum and maximum restriction values in the fields provided.

If a gene list has no associated numbers, you cannot select it for filtering in this view. For example, this filter cannot be applied to the “all genes” or “all genomic elements” lists because there are no associated values.

References

Benjamini, Y. and Hochberg, Y. (1995) “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society B*, 57, 289 -300.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2000) “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments”. *Department of Statistics Technical Report #578*, University of California, Berkeley (<http://stat-ftp.berkeley.edu/tech-reports/index.html>)

Holm, S. (1979) “A Simple Sequentially Rejective Bonferroni Test Procedure,” *Scandinavian Journal of Statistics*, 6, 65 -70.

Miller, R.G. (1981) *Simultaneous Statistical Inference*, Second Edition. New York: Springer-Verlag.

Westfall, P.H. and Young, S.S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*. New York: John Wiley & Sons, Inc.

Advanced Filtering

From the Advanced Filtering window, you can combine basic filters and analysis filters for more complex filtering operations. All of the basic filters described in the previous section are available, as well as Filter on Gene List and Filter on Annotations. In addition, you can perform Statistical Analysis (ANOVA) and Find Similar Genes operations. For more information on these operations, see “Statistical Analysis (ANOVA)” on page 6-33 and “The Find Similar Command” on page 6-6.

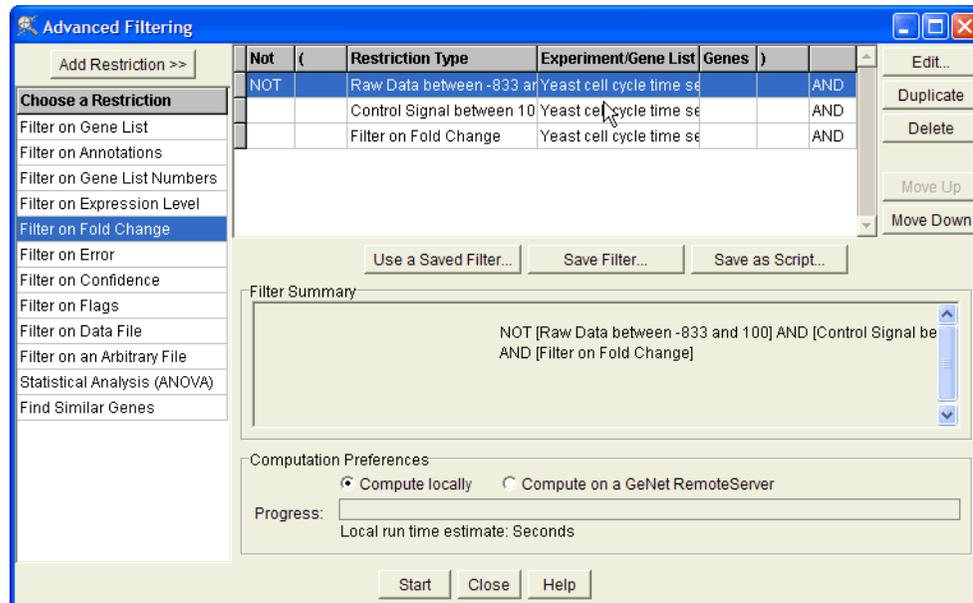


Figure 6-22 The Advanced Filtering Window

To set up an advanced filter:

1. Select a filter from the list of available filters and click **Add Restriction**. You can also add a filter by double-clicking its name in the list. The specified Filtering window appears.

Note: Each entry in the Advanced Filtering window must result in one and only one gene list. As a result, you do not have the option to run Post Hoc testing when applying 1-way ANOVA in an advanced filter, and you must select an individual gene list of interest when applying 2-way ANOVA.

2. Set up the desired filtering parameters and click **OK**. In the Advanced Filtering window, a line appears representing your filter.
3. Use the same procedure to add any additional filters.

To re-order steps, select a step in the list and click **Move Up** or **Move Down**. To insert another instance of a step, select it in the list and click **Duplicate**. To remove a step, click **Delete**.

At any time, you can view and edit an individual filtering step by double-clicking it in the Restrictions table or highlighting it and selecting **Edit**.

4. Choose from the pull-down menus in the column headers to construct the desired Boolean expression. For more information on constructing Boolean expressions, see “Creating Boolean Filters” on page 6-67.
5. Specify whether to run the script locally or on a GeNet Remote Execution Server.

If you specify Remote Execution, the preview pane in the Filtering window for each step is disabled. This is so that GeneSpring does not try to calculate the filter in real-time while you are creating it.

Note: An advanced filter using the Arbitrary File Restriction filter cannot be executed remotely.

6. Click **Start**.

Once you have created the desired filter, you can save it for future use by clicking **Save Filter** or **Save As Script**. For more information on saving filters, see “Saving Filters” on page 6-67.

Creating Boolean Filters

Not	(Restriction Type	Experiment/Gene List	Genes)	
NOT	▼	Raw Data between -833 and 833	Yeast cell cycle time series			AND
		Control Signal between 10 and 100	Yeast cell cycle time series			AND
		Filter on Fold Change	Yeast cell cycle time series			AND
	(
	((
	((((

Figure 6-23 The Restrictions table

Think of each row in the Restrictions table as a single gene list. There are no priorities between statements, so without parentheses to group statements, the order is assumed to be left-to-right. Use the pull-down parentheses menus to group restrictions together.

The **AND/OR** pull-down menu tells GeneSpring how to combine the grouped restrictions.

The **NOT** pull-down menu tells GeneSpring not to use the genes from the selected filtering step.

Saving Filters

You can save your filter either as a saved filter or as a script.

Saved Filters

Saved filters include all of the inputs to the filter and any associated information, including each restriction and its settings. They can be accessed in any genome.

When a saved filter is opened in a genome other than the one in which it was created, the genome data objects (gene lists, experiments, etc.) appear blank or undefined. You must define these fields within the new genome before running the filter. Filters that require data objects to be defined are displayed in red in the Restrictions table.

A saved filter is limited to the computer on which it was created. It cannot be sent to another user.

Saved Scripts

Filters saved as scripts save all of the current inputs as default inputs, but those inputs are not required to run the script. The script is saved in the Scripts folder in GeneSpring's navigator. It will not reconstruct the appearance of the Advanced Filtering window; instead, it runs exactly like a standard GeneSpring script. For more information on scripts in GeneSpring, see "Scripts" on page 8-2.

When you save a filter as a script, it is not limited to the computer on which it was created. This means you can send it to other GeneSpring users.

Clustering and Characterizing Data

The Clustering Window

GeneSpring's clustering algorithms are designed to divide genes or conditions into groups that have similar expression patterns. GeneSpring supports a variety of clustering methods, each designed to solve a distinct type of problem. These are useful tools to identify genes that are potentially co-regulated as well as to reveal coordinated responses to experimental treatments. To perform clustering operations, select Tools > Clustering and choose the appropriate clustering method from the menu. The following options are available:

- K-means
- Gene Tree
- Condition Tree
- Self-Organizing Map
- QT Clustering

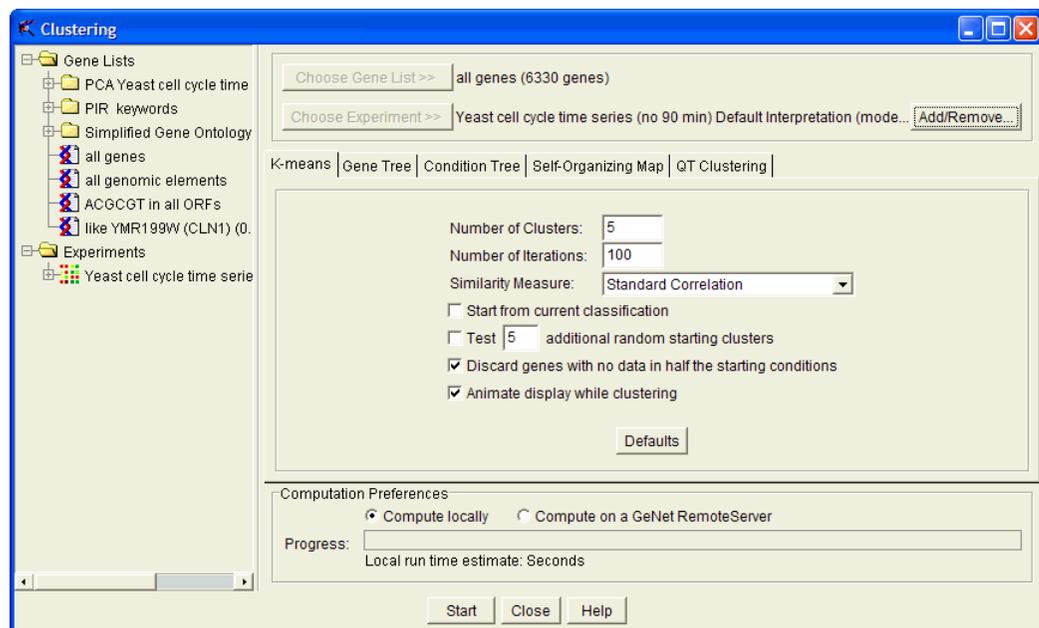


Figure 7-1 The Main Clustering Window, with the K-means tab selected

Using the Clustering Window

To perform any clustering operation, use the following steps:

1. Select a gene list from the navigator on the left side of the screen and click **Choose Gene List**.
2. If no experiment is selected, choose one from the mini-navigator and click **Choose Experiment**. Click **Add/Remove** to add or remove experiments from the list to be analyzed. For more information on adding and removing experiments, see “Add/Remove Experiments” on page 7-3.
3. Click the tab for your desired clustering method.

4. Enter settings for the clustering operation. More information on the settings available for each clustering method is available later in this chapter.
5. Specify whether to perform computation locally or on a GeNet Remote Server.
6. Click **Start**. If you are running the operation locally, its progress is indicated on the progress bar in the Computation Preferences section of the screen.

When the operation is complete, the result depends on the clustering method used. See the appropriate section of this chapter for more information.

Add/Remove Experiments

Use this window to add or remove experiments from the list to be included in the clustering operation.

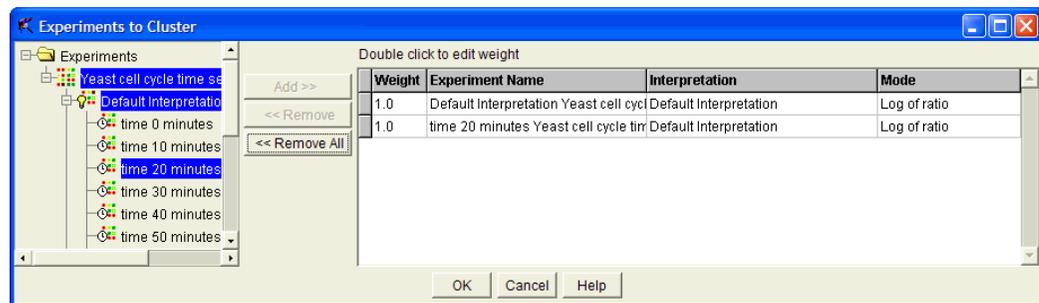


Figure 7-2 The Experiments to Cluster window

From this window you can do the following:

- To add an experiment, select it from the navigator and click **Add**.
- To remove an experiment, select it in the list and click **Remove**.
- To remove all listed experiments, click **Remove All**.
- To change an experiment's weight, double-click its number in the **Weight** column of the Experiments to Cluster window and enter a new value.

When you are done adding or removing experiments, click **OK** to return to the main clustering window.

Some Notes on Experiment Weight

Correlations of multiple experiments are performed through a weighted correlation in which you specify the weight of each experiment. You can make one experiment or experiment set more important than another. If all of the experiments or experiment sets are given the same weight, they are averaged equally.

The name of the experiment is noted directly after its relative weight. For example, you could give SampleExperiment1 a weight of 2, and Experiment2 a weight of 1. Therefore, in this example, the correlations found in the SampleExperiment1 are twice as influential in creating the tree as the correlations between the genes in the Experiment2 study.

The equation used to determine the overall correlation is:

$$\mathbf{X} = \frac{(\mathbf{Aa} + \mathbf{Bb} + \mathbf{Cc} + \dots)}{(\mathbf{a} + \mathbf{b} + \mathbf{c} + \dots)}$$

- **A** is the correlation coefficient between the gene in question in experiment 1 and the gene named in the *Experiments to Use* box, also from experiment 1.
- **a** is the weight specified for experiment 1.
- **B** is the correlation coefficient of the gene in question in experiment 2, to the gene named in the title bar, also from experiment 2.
- **b** is the weight associated with experiment 2.
- **C** is the correlation coefficient of the gene in question in experiment 3 to the gene named in the title-bar, also from experiment 3.
- **c** is the weight associated with experiment 3.
and so on.

Experiments 1, 2, 3, etc., represent all of the experiments selected in the white Correlations box. If **X** is between the minimum and maximum correlations specified in the Clustering window, the gene in question passes the correlations.

Similarity Definitions

Similarity definitions are used in several clustering types. The equations used to determine the nine types of correlations are described in detail in Appendix B, “Equations for Correlations and other Similarity Measures”.

The default correlation is the Standard Correlation, *Standard correlation* = $\mathbf{a \cdot b} / (|\mathbf{a}| |\mathbf{b}|)$.

Minimum Distance and Separation Ratios

$$\frac{\sum_{i=1}^n A_i B_i}{\sqrt{\left(\sum_{i=1}^n A_i^2 \right) \left(\sum_{i=1}^n B_i^2 \right)}}$$

To make a tree, GeneSpring calculates the correlation for each gene with every other gene in the set. Then it takes the highest correlation and pairs those two genes, averaging their expression profiles. GeneSpring then compares this new composite gene with all of the other unpaired genes.

This is repeated until all of the genes have been paired. At this point the minimum distance and the separation ratio come in to play. Both of these affect the branching behavior of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A value smaller than .001 has very little effect, because most genes are not correlated more closely than that. A higher number tends to lump more genes into a group, making the groups less specific.

Clustering Methods

Gene Tree

The classification of organisms into phylogenetic trees is a central concept to biology. Organisms sharing properties tend to be clustered together, and the location of a branch containing both organisms can be considered a measure of how similar the organisms are. You can classify genes in a similar manner—clustering those whose expression patterns are similar into nearby places in a tree. Such mock-phylogenetic trees are often referred to as gene trees.

GeneSpring can both create and display such trees. GeneSpring can also create trees of experiments, displaying the genes along one axis and the samples along the other axis. This is useful for many applications. For example, you can determine if any environmental stressors cause similar effects on the expression levels as mutant organisms do.

If you have already created or downloaded trees, open the Gene Trees folder in the navigator and select any tree for viewing.

Gene Tree Clustering Options

The following options are available for gene tree clustering:

- **Similarity Measure**—available options are:
 - Standard Correlation
 - Smooth Correlation
 - Change Correlation
 - Upregulated Correlation
 - Pearson Correlation
 - Spearman Correlation
 - Spearman Confidence
 - Two sided Spearman Confidence
 - Distance

For detailed information on these measures of similarity, see “Similarity Definitions” on page 7-4.

- **Do automatic annotation**—Specifies whether to annotate the nodes of the tree with the names of the gene lists that have similar members. Using this option can add considerable time to the tree-building process, but is usually worthwhile.

This feature becomes even more valuable once you have created a simplified ontology for the genome, as the ontological classifications can be used to label tree branches. For details on creating a simplified ontology, see “Building a Simplified Ontology” on page 6-31.

- **Only annotate with standard lists**—Specifies whether the annotations on the nodes are done with all gene lists or only the gene lists marked as standard. (This is set in the Gene List Inspector. See “Displaying a Gene List” on page 4-3 for more information.)
- **Discard genes with no data in half the starting conditions**—Discard any genes with no data in at least half the conditions in the selected experiment.

Clustering Methods

- **Merge similar branches**—Merge branches with similar results. For information on the Separation Ratio and Minimum Distance settings, see “Advanced Tree Options” on page 7-8.

Saving Gene Tree Clustering Results

When the gene tree clustering operation has completed, the Name New Gene Tree window appears.

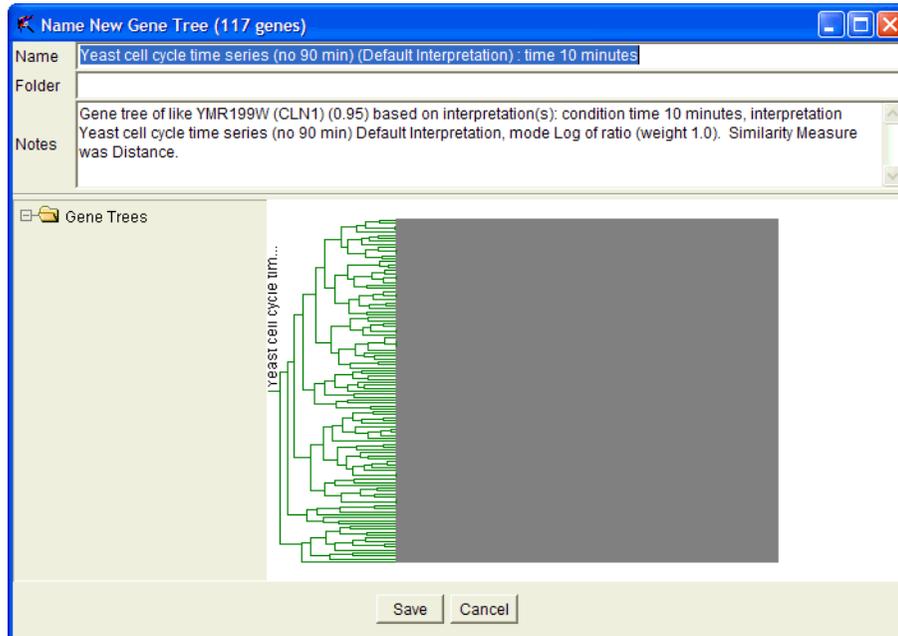


Figure 7-3 The Name New Gene Tree window

To save the new gene tree:

1. Enter a name in the **Name** field at the top of the screen. Names may not exceed 80 characters.
2. From the navigator, select a folder in which to save the new gene tree. To create a new folder, navigate to the desired parent folder and enter a new folder name in the **Folder** field.
3. Enter any additional information in the **Notes** field, if desired.
4. Click **Save**.

Condition Tree

Complex trees can be made from multiple conditions or by tightly defining the types of data to use. Select a gene list in the navigator to reduce the number of genes to be made into a tree.

Condition Tree Options

The following options are available for condition tree clustering:

- **Similarity Measure**—available options are:

- Standard Correlation
- Smooth Correlation
- Change Correlation
- Upregulated Correlation
- Pearson Correlation
- Spearman Correlation
- Spearman Confidence
- Two sided Spearman Confidence
- Distance

For more information on measures of similarity, see “Equations for Correlations and other Similarity Measures” on page B-1.

- **Merge similar branches**—Merge branches with similar results. For information on the Separation Ratio and Minimum Distance settings, see “Advanced Tree Options” on page 7-8.

Saving Condition Tree Results

When the operation is complete, the Name New Condition Tree window appears.

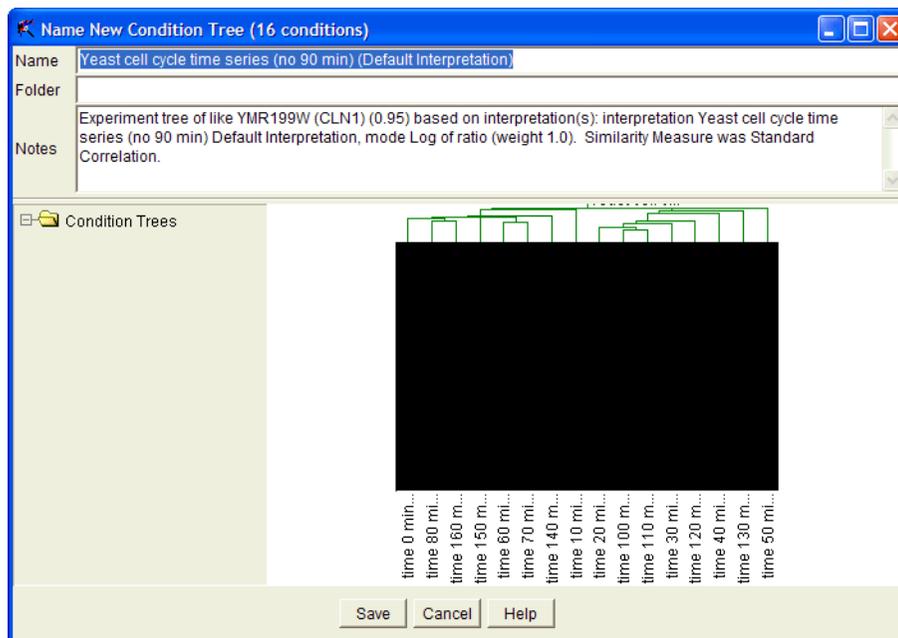


Figure 7-4 The Name New Condition Tree window

To save the new condition tree:

1. Enter a name in the **Name** field at the top of the screen. Names may not exceed 80 characters.
2. From the navigator, select a folder in which to save the new condition tree. To create a new folder, navigate to the desired parent folder and enter a new folder name in the **Folder** field.

3. Enter any additional information in the **Notes** field, if desired.
4. Click **Save**.

Advanced Tree Options

The separation ratio determines how large the correlation difference between groups of clustered genes must be for them to be considered discrete groups. This number should be between 0 and 1.

It is not usually appropriate to change separation ratio or minimum distance.

Separation Ratio

The separation ratio determines how large the correlation difference between groups of clustered genes has to be for the groups to be considered discrete groups and not be joined together.

- Increasing separation increases the ‘branchiness’ of the tree.
- Default Separation ratio is 1.0. Separation ratio can range from 0.0 to 1.0.
- At a separation ratio of 0, all gene expression profiles can be regarded as identical.

To change the maximum correlation number, enter a new value in the Separation Ratio box.

Minimum Distance

The number specified in the *Minimum distance* box determines the minimum separation considered significant between genes. This reduces meaningless structure at the base of the tree. The minimum distance deals with how far down the tree discrete branches are depicted. A higher number tends to lump more genes into a group, making the groups less specific.

- Decreasing minimum distance increases the ‘branchiness’ of the tree.
- Default minimum distance is 0.001. A value smaller than .001 has very little effect, because most genes are not correlated more closely.

To change the default minimum distance, enter a new value in the *Minimum distance* box.

References for Hierarchical Clustering

Everitt, Brian S. *Cluster Analysis* (3rd Ed.) Arnold, London, 1993, pp 62-65.

Eisen, Michael B., et. al. “Cluster analysis and display of genome-wide expression patterns” *Proc. Natl. Acad. Sci. USA*, V95, pp 14863-14868, December 1998.

k-Means Clustering

K-means clustering divides genes into groups based on their expression patterns. The goal is to produce groups of genes with a high degree of similarity within each group and a low degree of similarity between groups. Unlike self-organizing maps, k-means clustering is not designed to show the relationship between clusters. Instead, k-means clusters are constructed so that the average behavior in each group is distinct from any of the other groups. For example, in a time series experiment you could use k-means clustering to identify

unique classes of genes that are upregulated or downregulated in a time dependent manner.

GeneSpring's k-means clustering algorithm divides genes into a user-defined number (k) of equal-sized groups, based on the order in the selected gene list. It then creates centroids (in expression space) at the average location of each group of genes. With each iteration, genes are reassigned to the group with the closest centroid. After all of the genes have been reassigned, the location of the centroids is recalculated and the process is repeated until the maximum number of iterations has been reached.

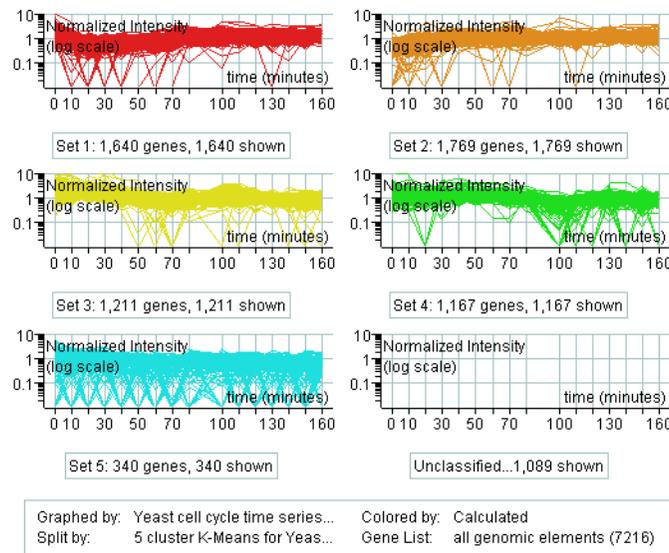


Figure 7-5 A k-means Cluster display in a Split Window

K-means clustering Options

The following options are available:

- **Number of Clusters**—The number of clusters to make
- **Number of Iterations**—The maximum number of times that each centroid is recalculated after genes are reassigned to groups with the most similar centroids.
- **Similarity Measure**—available options are:
 - Standard Correlation
 - Smooth Correlation
 - Change Correlation
 - Upregulated Correlation
 - Pearson Correlation
 - Spearman Correlation
 - Spearman Confidence
 - Two sided Spearman Confidence
 - Distance

For more information on measures of similarity, see “Similarity Definitions” on page 7-4.

Clustering Methods

- **Start from Current Classification**—Group genes using the selected classification as a starting point. Note that this option is available only if you have selected a classification. This option disables the Number of Clusters checkbox, since it automatically uses the number of classes in the current classification.
- **Test [x] Additional Random Starting Clusters**— Enter a number to make clustering as tight as possible by performing clustering several times, each time starting from a different random grouping of genes, and choosing the best result. The default value is 5.
- **Discard genes with no data in half the starting conditions**—Discard any genes with no data in at least half the conditions in the selected experiment.
- **Animate display while clustering**—Show changes in classification assignments in real time. This may slow your analysis slightly.

Saving k-means Results

When the k-means operation is complete, the Choose Classification Name window appears.

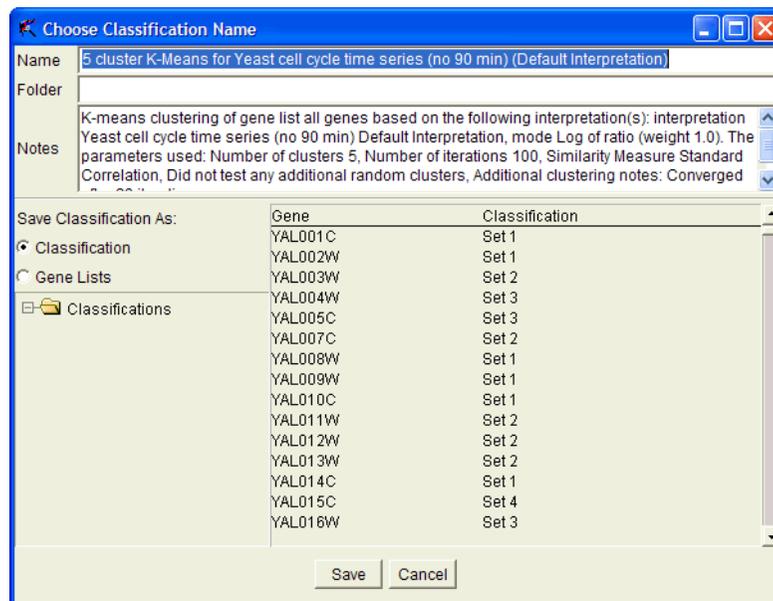


Figure 7-6 The Choose Classification Name window

To save your results:

1. Enter a name in the **Name** field at the top of the screen. Names may not exceed 80 characters.
2. To save the results as a classification, select the **Classification** radio button. To save the results as a group of gene lists, select the **Gene Lists** radio button
3. From the navigator, select a folder in which to save the new classification or gene lists. To create a new folder, navigate to the desired parent folder and enter a new folder name in the **Folder** field.
4. Enter any additional information in the **Notes** field, if desired.

5. Click **Save**.

Viewing k-means Clusters

If you use k-means clustering to produce a classification, you can view details about the classification in the Classification Inspector. For information about the Classification Inspector, see “The Classification Inspector” on page 4-22.

The easiest way to view a classification is with the Split Window feature. Right-click a classification or a gene list created with k-means clustering and select **Split Window > Both**. The genome browser splits into several smaller displays. (You can also choose vertically or horizontally.)

Self-Organizing Maps

The self-organizing map (SOM) is a clustering technique similar to k-means clustering. However, SOMs illustrate the relationship between groups by arranging them in a two-dimensional map in addition to dividing genes into groups based on expression patterns. SOMs are useful for visualizing the number of distinct expression patterns in your data and determining which of these patterns are variants of one another. SOMs were invented by Tuevo Kohonen (1991, 2000) and are used to analyze many kinds of data. Applications to gene expression analysis were described by Tamayo, et al (1999).

GeneSpring’s self-organizing map algorithm begins by creating a two-dimensional grid of nodes in the space of gene expression. In each iteration, one gene is selected and all of the nodes within a user-defined “neighborhood” are moved closer to it. This process is repeated with each gene in the selected gene list until the maximum number of iterations has been reached.

With each iteration, the “neighborhood radius” is incrementally reduced and nodes are moved by smaller and smaller amounts to produce convergence. In this way, the grid of nodes is stretched and wrapped to best represent the variability of the data, while still maintaining similarity between adjacent nodes. After the iteration is complete, genes are assigned to the nearest node, and a display grid of gene expression graphs is generated, corresponding to the initial grid of nodes.

As the iteration proceeds, the neighborhood radius decreases smoothly, so that points move more independently later in the process. The neighborhood radius is expressed in terms of Euclidean distance in grid units relative to the abstract grid of the expression patterns. (This is different from the distance between nodes in gene expression space.) For instance, point 1,2 is one unit away from 1,3.

If you make the neighborhood radius very small (less than 1) each point always moves independently, and adjacent clusters are not related. If you specify a very large neighborhood radius, initially all the nodes move toward every data point, and the grid behaves as if it is very “stiff”, with more similarity between node results, but less flexibility to explore the variations in the data.

Self-Organizing Map Options

The following options are available for SOM clustering:

Clustering Methods

- **Rows**—The number of rows in your grid. The default setting is based on the number of genes and conditions in the selected experiment(s).
- **Columns**—The number of columns in your grid. The default setting is based on the number of genes and conditions in the selected experiment(s).
- **Number of Iterations**—How many times each gene is examined. For example, if there are 10,000 genes and 60,000 iterations are specified, each gene is examined six times.
- **Neighborhood Radius**—How many nodes move toward a data point at the beginning of the iteration, and therefore how similar the profiles are for each node.
- **Discard genes with no data in half the starting conditions**—Discard any genes with no data in at least half the conditions in the selected experiment.

Note: A good way to estimate the optimum number of rows and columns is to try to predict how many distinct classes of genes are affected by the conditions in your experiment. With small data sets, the algorithm may generate a number of empty nodes. To avoid this, you might try using a smaller grid.

SOM Results

When the SOM operation is complete, the Choose Classification Name window appears.

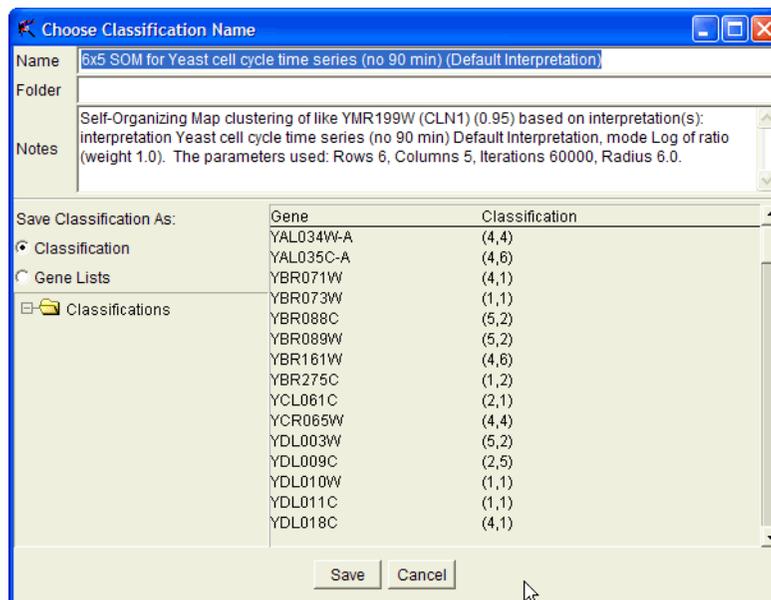


Figure 7-7 The Choose Classification Name window

To save your results:

1. Enter a name in the **Name** field at the top of the screen. Names may not exceed 80 characters.
2. To save the results as a classification, select the **Classification** radio button. To save the results as a group of gene lists, select the **Gene Lists** radio button

3. From the navigator, select a folder in which to save the new classification or gene lists. To create a new folder, navigate to the desired parent folder and enter a new folder name in the **Folder** field.
4. Enter any additional information in the **Notes** field, if desired.
5. Click **Save**.

Viewing SOMs

SOM results are most easily viewed using the Split Window feature. Each graph contains the genes associated with a SOM node. Node numbers are shown in the upper right corner of each plot.

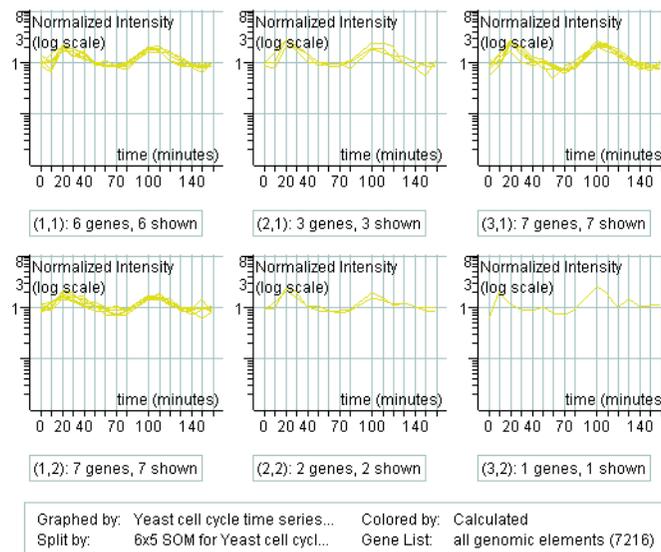


Figure 7-8 A 3x2 SOM of the “Yeast cell time series (no 90 min)” experiment

If you have selected many panels, you may want to hide the horizontal and vertical labels for easier viewing. Right-click the genome browser and select an option from the Options submenu. You can also increase your viewing space by selecting **View > Visible > Hide All**.

If you use a SOM to produce a classification, you can get details about the classification from the Classification Inspector. For information about the Classification Inspector, see “The Classification Inspector” on page 4-22. To recreate your SOM graph, click the SOM classification or folder of gene lists in the navigator and select **Split Window > Both**.

SOM References

- Kohonen, T. (1990). The Self-Organizing Map. *Proc. IEEE* 78(9):1464-1480.
- Kohonen, T. (2000). *Self-Organizing Maps* (Third Edition). Springer Verlag, Berlin.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps;

Methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci. USA* 96:2907-2912.

QT Clustering

QT clustering looks for clusters of genes such that each gene in the cluster is within a specified distance (based on a user-defined distance metric) of every other gene in the cluster. In GeneSpring, the cutoff is specified based on a correlation function, so the cutoff is the minimum allowed value: In QT clustering, the “diameter” of a cluster refers to the largest distance between any two genes in the same cluster.

QT clustering builds a cluster by starting with a single gene. The diameter at that point is 0. It then adds the gene that is closest to the starting gene. The diameter of the cluster is now equal to the distance between the two genes. It continues adding genes one at a time, always choosing the gene that will result in the smallest cluster diameter. Eventually it reaches a point where no genes can be added without the diameter growing beyond the allowed cutoff. The cluster is then complete.

The cluster obtained depends on which gene is chosen to start from. Therefore, it independently builds clusters starting from each gene in the user-selected gene list. The cluster with the most genes is kept, and is part of the final classification. All others are discarded.

There is now a single cluster. The genes in this cluster are removed from the list, and the process is begun again. A new cluster is built from every gene in the reduced gene list, the largest one is kept. This process is repeated until the number of genes in the largest cluster is smaller than a user-defined cutoff.

QT Clustering Options

The following options are available for QT Clustering:

- **Minimum Cluster Size**—The smallest allowable size for a cluster to be considered valid
- **Minimum Correlation**—The minimum correlation for any pair of genes in the same cluster
- **Similarity Measure**—available options are:
 - Standard Correlation
 - Smooth Correlation
 - Change Correlation
 - Upregulated Correlation
 - Pearson Correlation
 - Spearman Correlation
 - Spearman Confidence
 - Two sided Spearman Confidence
 - Distance

For more information on measures of similarity, see “Similarity Definitions” on page 7-4.

Saving QT Clustering Results

When the operation is complete, the Choose Classification Name window appears.

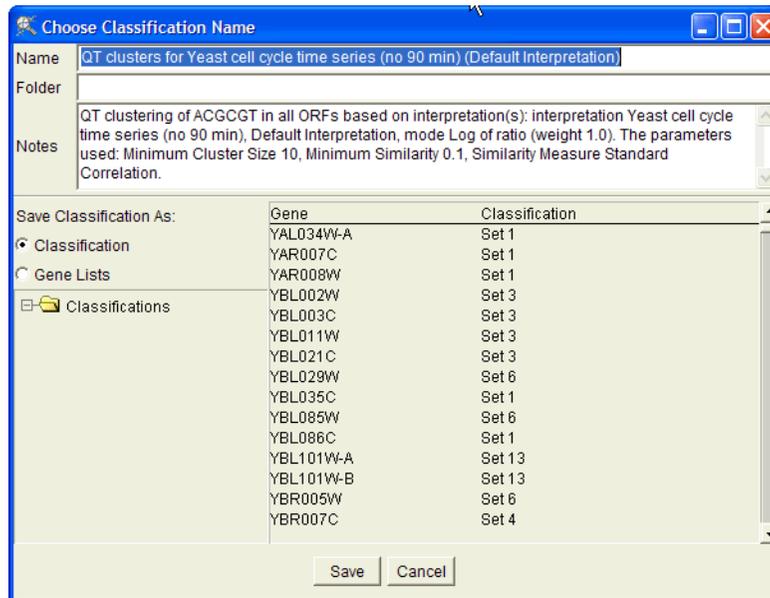


Figure 7-9 The Choose Classification Name window

To save your results:

1. Enter a name in the **Name** field at the top of the screen. Names may not exceed 80 characters.
2. To save the results as a classification, select the **Classification** radio button. To save the results as a group of gene lists, select the **Gene Lists** radio button
3. From the navigator, select a folder in which to save the new classification or gene lists. To create a new folder, navigate to the desired parent folder and enter a new folder name in the **Folder** field.
4. Enter any additional information in the **Notes** field, if desired.
5. Click **Save**.

Principal Components Analysis

Principal components analysis (PCA) is a decomposition technique that produces a set of expression patterns known as principal components. Linear combinations of these patterns can be assembled to represent the behavior of all of the genes in a given data set.

PCA is not a clustering technique. It is a tool to characterize the most abundant themes or building blocks that reoccur in many genes in your experiment. You can run PCA on genes or on conditions.

By default for PCA on Genes, PC scores are calculated by computing the standard correlation between each gene's expression profile vector and each principal component vector (eigenvector). For PCA on conditions, this means calculating the standard correlation between each condition vector and each principal component vector (eigenvector). Calculating scores this way has the advantage of scaling them to be between -1 and 1.

If you uncheck the **Report scores as correlations** box, the PC scores represent the coordinates of the genes or conditions in the system defined by the first few principal components. In other words, these scores are the values of the principal components for each gene or condition.

PCA on Genes

To run PCA on genes:

1. Select **Tools > Principal Components Analysis**.

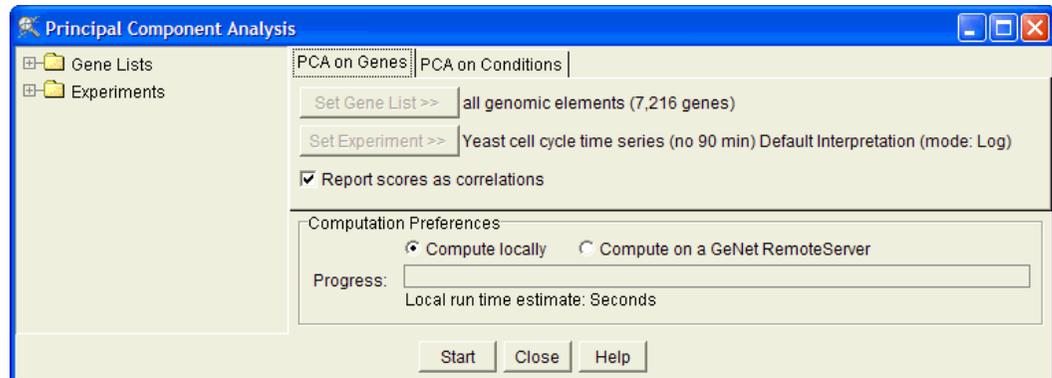


Figure 7-10 Principal Components Analysis screen

2. If it is not already selected, click the PCA on Genes tab.
3. Select a gene list from the navigator and click **Set Gene List**.
4. Select an experiment from the navigator and click **Set Experiment**.
5. Check or uncheck the **Report scores as correlations** box to specify whether to report scores as correlations or as the values of the principal components for each gene. This box is checked by default.
6. Specify whether to run the computation locally or on a GeNet Remote Server.
7. Click **Start**.

PCA on Genes Results

When the analysis is complete, the PCA Results window appears, displaying each component as a line in graph mode. The significance of each component is represented by the color of its graph line, as defined by the colorbar.

In addition, a new gene list folder appears in the GeneSpring navigator with a name that includes the experiment that you used for PCA analysis (e.g., “PCA yeast cell cycle”).

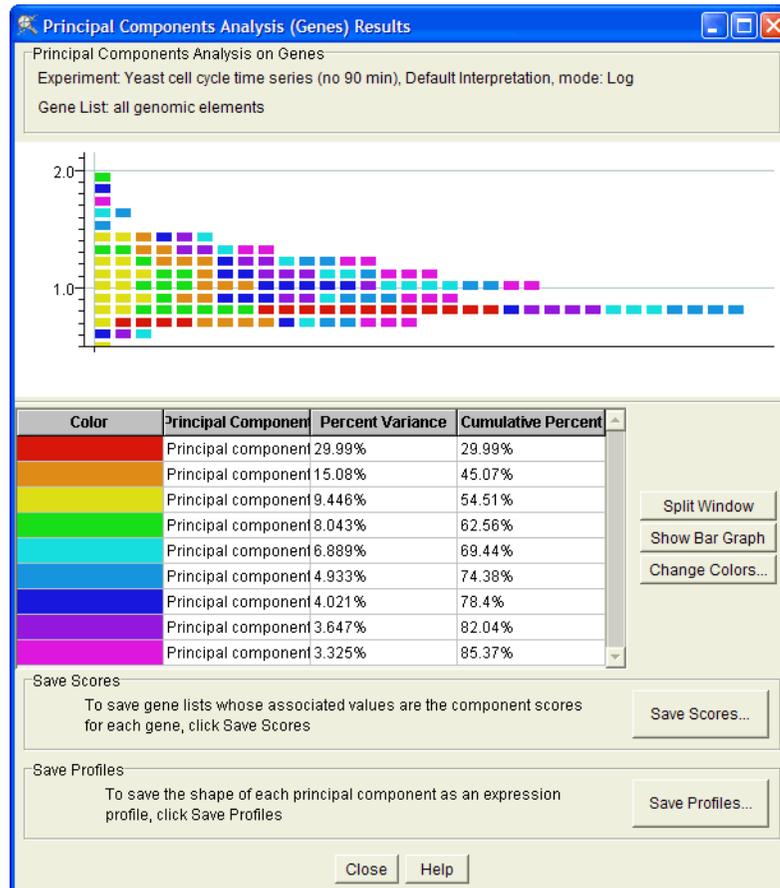


Figure 7-11 The Principal Components Analysis Results window

Double-click a component to view the Gene Inspector window, which shows the eigenvalue and explained variability in the upper-left panel.

This screen contains the following buttons:

- **Split/Unsplit Window**—toggles between the default view and splitting the graph by component.
- **Show Bar/Line Graph**—toggles between the bar and line graph views.
- **Change Colors**—allows you to change the colors used to display the components.
- **Save Scores**—save gene lists whose associated values are the component scores for each gene.
- **Save Profiles**—save the shape of each principal component as an expression profile.

PCA on Conditions

To run PCA on conditions:

1. Select **Tools > Principal Components Analysis**.

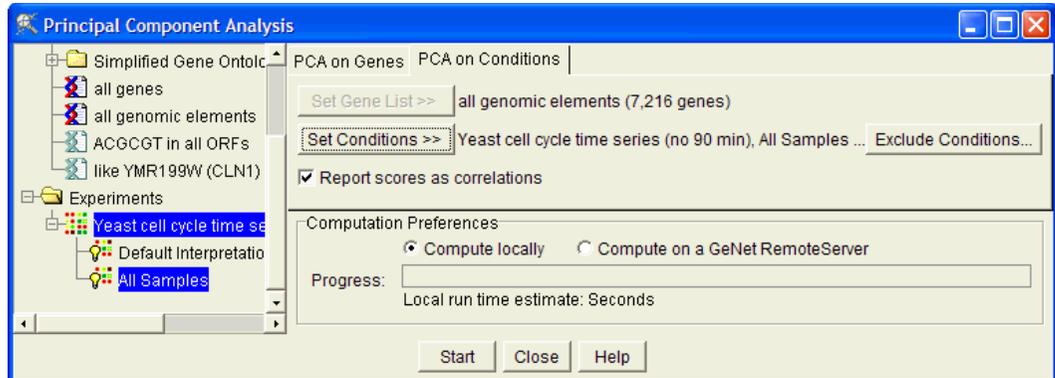


Figure 7-12 PCA on Conditions tab

2. If it is not already selected, click the PCA on Conditions tab.
3. Select a gene list from the navigator and click **Set Gene List**.
4. Select an experiment from the navigator and click **Set Conditions**.
5. Click **Exclude Conditions...** to specify which conditions (if any) to exclude from the analysis.

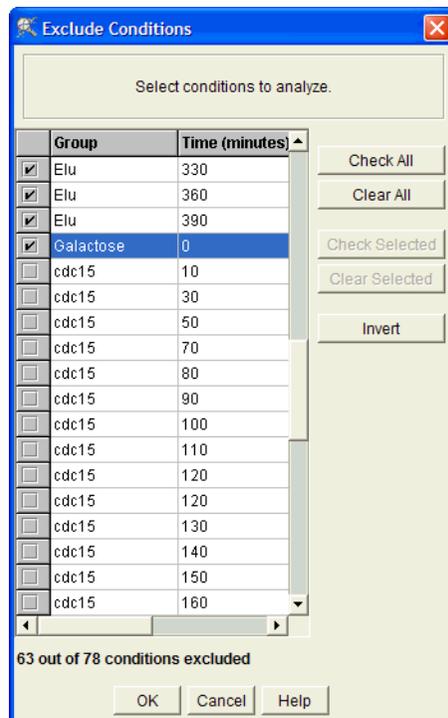


Figure 7-13 The Exclude Conditions window

By default, all conditions are selected. To exclude a condition, uncheck the box to its left. To include a condition, check the box. Click **Check All** to include all conditions, or **Clear All** to exclude all conditions. To check or clear a range of conditions, select them in the list and click **Check Selected** or **Clear Selected**.

6. Check or uncheck the **Report scores as correlations** box to specify whether to report scores as correlations or as the values of the principal components for each condition. This box is checked by default.
7. Specify whether to run the computation locally or on a GeNet Remote Server.
8. Click **Start**.

PCA on Conditions Results

When the analysis is complete, the PCA Results window appears, displaying each condition as a line in graph mode. The significance of each condition is represented by the color of its graph line, as defined by the colorbar.

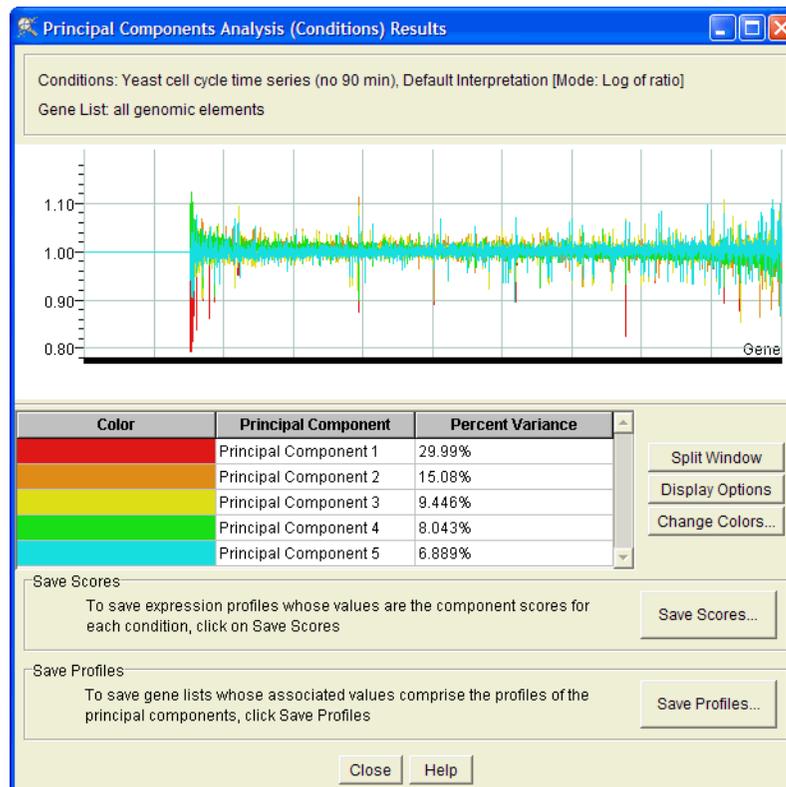


Figure 7-14 The Principal Components Analysis Results window

In the PCA results window, double-click a condition to view the Gene Inspector window, which shows the eigenvalue and explained variability in the upper-left panel.

This screen contains the following buttons:

- **Split/Unsplit Window**—toggles between the default view and splitting the graph by component.
- **Show Bar/Line Graph**—toggles between the bar and line graph views.

- **Change Colors**—allows you to change the colors used to display the components.
- **Save Scores**—save expression profiles showing the component scores.
- **Save Profiles**—save gene lists to profile the PCA components.

A second window appears that displays a condition scatter plot with the first three components on the axes. For information on using this view, see “Condition Scatter Plot” on page 4-68.

Interpreting your PCA Results

The principal components of a data set are the eigenvectors obtained from an eigenvector-eigenvalue decomposition of the covariance matrix of the data. The eigenvalue corresponding to an eigenvector represents the amount of variability explained by that eigenvector. The eigenvector of the largest eigenvalue is the first principal component. The eigenvector of the second largest eigenvalue is the second principal component and so on. Principal components which explain significant variability are displayed by GeneSpring in the Principal Components Analysis window.

There are never more principal components than there are conditions in the data.

Viewing Principal Component Loadings in a Scatter Plot

After performing principal components analysis, the genome browser displays a 3-D scatter plot in which the loadings for the first, second, and third principal components (representing the largest fraction of the overall variability) are plotted on the X, Y, and Z axes respectively. In Figure 7-15, each point represents a single gene. Its position on the Y-axis represents the loading of principal component 2. The position on the X-axis represents the loading of principal component 1.

This view is useful for selecting and making lists of genes that exhibit high levels of one or two principal components. Genes that exhibit high levels of the first principal component and low levels of the second principal component are displayed in the lower right corner of the plot, and genes exhibiting equal levels of the two components lie along the diagonal.

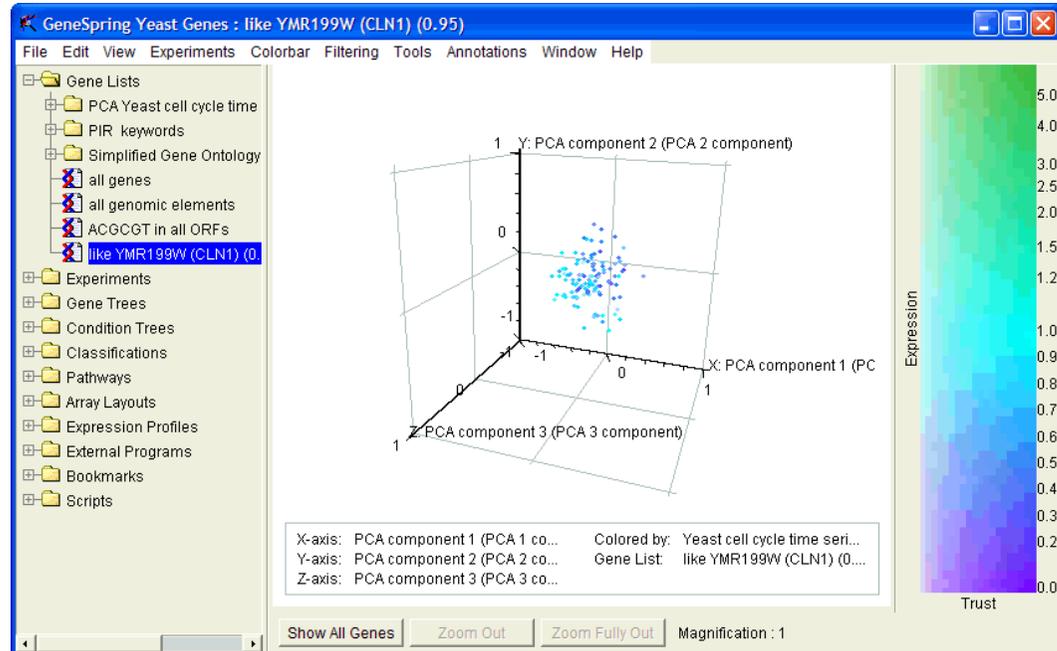


Figure 7-15 PCA Scatter Plot

You can change the components that are represented by each axis by right-clicking in the browser and selecting **Display Options**.

Regenerating the PCA Scores Scatter Plot

If you have closed the PCA scatter plot window and have saved the PCA scores as a set of gene lists or expression profiles, you can reproduce the initially displayed scatter plot by doing the following:

PCA on Genes:

1. Open **View > Scatter Plot** (or 3D Scatter Plot)
2. Right-click over the scatter plot and select **Display Options**.
3. Select the first desired gene list from the navigator and assign it to the X axis. Repeat this step for the remaining axes.
4. When you are done assigning the gene lists to the desired axes, click **OK**.

PCA on Conditions:

1. Open **View > Condition Scatter Plot**.
2. Right-click over the scatter plot and select **Display Options**.
3. Select the first desired expression profile from the navigator and assign it to the X axis. Repeat this step for the remaining axes.
4. When you are done assigning expression profiles to the desired axes, click **OK**.

Viewing Principal Components in an Ordered List

The best way to visualize the genes that exhibit the highest levels of an individual component is to use the ordered list view.

Select **View > Ordered List** and choose one of the PCA gene lists from the navigator panel. Genes exhibiting the highest levels of the selected principal component are displayed on the left side of the genome browser and have the longest lines extending upward from them. For more details, see “Ordered List View” on page 4-58.

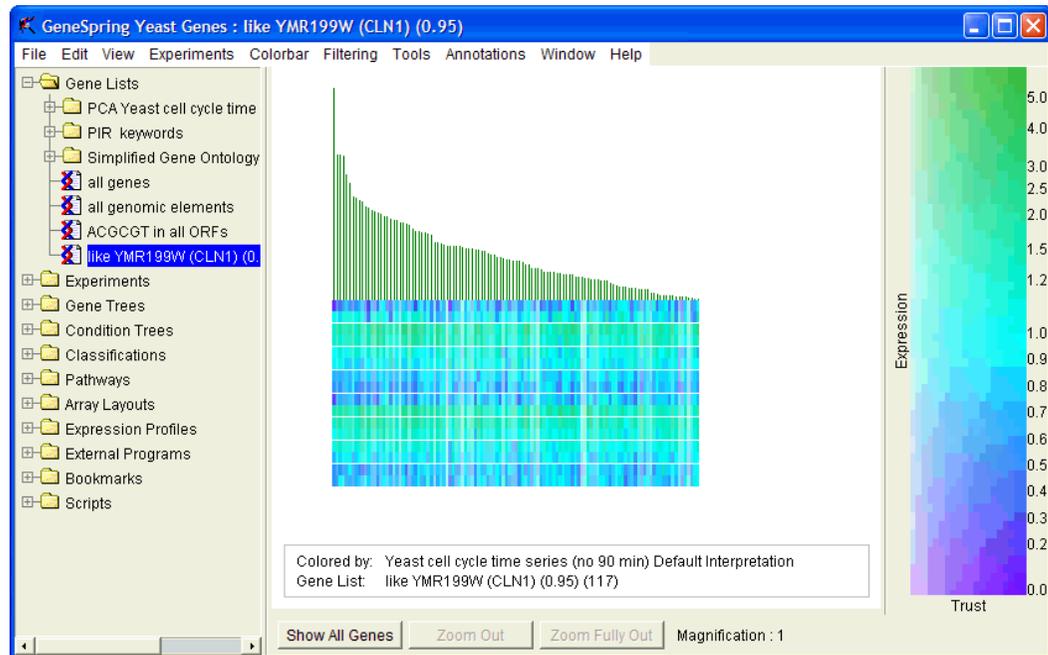


Figure 7-16 PCA in the Ordered List view

References for Principal Components Analysis

Alter O., Brown P.O., Botstein D. *Singular value decomposition for genome-wide expression data processing and modeling*. PNAS 97:10101-6 (2000) <http://www.pnas.org/cgi/content/full/97/18/10101>

Cooley, W.W. and Lohnes, P.R. *Multivariate Data Analysis* (John Wiley & Sons, Inc., New York, 1971).

Gnanadesikan, R. *Methods for Statistical Data Analysis of Multivariate Observations* (John Wiley & Sons, Inc., New York, 1977).

Neal S. Holter et al, *Fundamental patterns underlying gene expression profiles: Simplicity from complexity*. PNAS 97,8409 (2000) <http://www.pnas.org/cgi/content/abstract/97/15/8409>

Hotelling, H. *Analysis of a Complex of Statistical Variables into Principal Components*. Journal of Educational Psychology **24**, 417-441, 498-520 (1933).

Kshirsagar, A.M. *Multivariate Analysis* (Marcel Dekker, Inc., New York, 1972).

Mardia, K.V., Kent, J.T., and Bibby, J.M. *Multivariate Analysis* (Academic Press, London, 1979).

Morrison, D.F. *Multivariate Statistical Methods*, Second Edition (McGraw-Hill Book Co., New York, 1976).

Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* **6(2)**, 559 -572 (1901).

Rao, C.R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya A* **26**, 329 –358 (1964).

Raychaudhuri, S., Stuart, J.M. and Altman, R.B. *Principal components analysis to summarize microarray experiments: application to sporulation time series*. Pacific Symposium on Biocomputing (2000).

The Class Predictor

The Class Predictor is designed to predict the value, or “class”, of an individual parameter in an uncharacterized sample or set of samples. It does this in two steps.

First, the Class Predictor algorithm examines all genes in the training set individually and ranks them on their power to discriminate each class from all the others. Next it uses the most predictive genes to classify the “test set” (i.e. the set where the parameter value of interest is unknown). For example, you could attempt to diagnose the leukemia type of a leukemia patient with the Class Predictor by using expression data from patients whose leukemia type was known. You can also use the Class Predictor simply to find genes whose behavior is related to a given parameter by examining the list of predictor genes.

The list of predictor genes is assembled using Fischer’s exact test. In this method, all the measurements for a given gene are ordered according to their normalized expression levels. For each class (parameter value), the predictor places a mark in the list where the relative abundance of the class on one side of the mark is the highest in comparison to the other side of the mark. The genes that are most accurately segregated by these markers are considered to be the most predictive. A list of the most predictive genes is made for each class and an equal number of genes (lowest P-value using Fischer’s exact test) are taken from each list.

To make a prediction, the class predictor uses the k-nearest-neighbor method. It selects “k” number of samples near (as measured in Euclidean distance) the unclassified sample, and for each class, computes a P-value that is the likelihood of finding the observed number of this class within the neighborhood members by chance given the proportion of the classes in the training set. The class with the lowest P-value is assigned to the unclassified sample.

You can specify a P-value cutoff, or threshold, such that if there is not sufficient evidence in favor of a particular class, no prediction is made. The P-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.

To Use the Class Predictor

1. Select **Tools > Predict Parameter Values**. The Predict Parameter Values window appears.
2. Open the Experiments folder in the navigator and click your training set (the set of samples for which the parameters are already known). Click **Training Set**.
3. Click your test set (the set where the parameter value of interest is unknown), and click **Test Set**.
4. Open the Gene Lists folder in the navigator and click a gene list to be used in the selection process. Click **Select Genes From**.
5. Select a parameter in the **Parameter to predict** box.
6. Specify a **Number of predictor genes** to be used in the prediction.
7. Specify a **Number of neighbors**. Generally, this number should be no more than half the size of a single class, and no less than 10.

8. Specify a **Decision cutoff for P-value ratio**. The P-value cutoff is a threshold such that if there is not sufficient evidence in favor of a particular class, no prediction is made. The P-value cutoff is a ratio of the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.
9. Select **Predict Test Set** to make a prediction or **Crossvalidate Training Set** to evaluate how well the prediction rule can be used to predict the parameter values of the training set.
10. Specify whether to run this process on your local machine or a GeNet Remote Server.
11. Click **Start**.

Interpreting the Results of a Prediction

The Prediction Results window appears after you have made a prediction or validated a training set. For convenience, not all of the prediction statistics are visible until you click the **Show Details** button at the bottom of the window.

- **True Value**—the true value of the class of each sample, as calculated when the parameter for the test set is already known. Compare this with the value in the Prediction column to validate your training set.
- **Prediction**—the predicted class.
- **P-value ratio**—the P-value ratio, or the probability that the prediction was made by chance for the two classes. If you have more than two classes, the ratio is the lowest P-value divided by the next lowest P-value.
- **Class counts**—the individual class counts for each sample.
- **P-value**—probability that individual class counts were found by chance.

The Class Predictor is designed for experiments with at least 20 or so samples in each class. It is possible to use the Predictor when you have very small sample sizes if you disable the P-value cutoff function. For sample sizes of less than 5, specify 1 or 2 number of neighbors and specify 1 in the P-value cutoff field.

Find Similar Samples

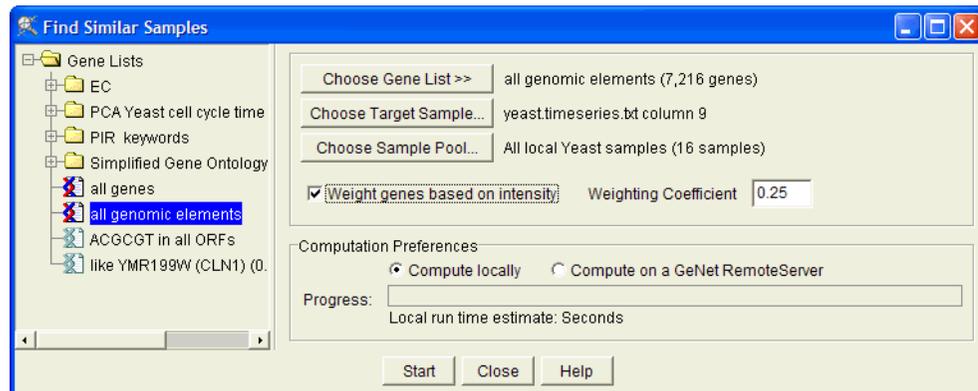


Figure 7-17 The Find Similar Samples window

There are two ways of reaching this window:

- Select **Tools > Find Similar Samples**
- From the Sample Inspector's Similar Samples tab, click the **Find Similar Samples** button

To find similar samples:

1. Select a gene list from the navigator and click **Choose Gene List**.
2. If the sample to be compared is not pre-selected, click **Choose Target Sample**. The Select Target Sample window appears. This window works similarly to the Sample Manager, except that you can select only one sample. For details on using this window, see “The Sample Manager” on page 3-23.
3. To specify the samples among which to search, click **Choose Sample Pool**. The Select Sample Pool window appears. This window works exactly like the Sample Manager. For details on using this window, see “The Sample Manager” on page 3-23.
4. If necessary, change the value in the **Weighting Coefficient** box.
If you do not wish to weight genes based on their control value, uncheck the **Weight genes based on intensity** box.
5. Specify whether to run this process locally, or on a GeNet Remote Server.
6. Click **Start**.

When the analysis is complete, the Find Similar Samples: Results window appears.

The Find Similar Samples Results Window

This window displays the results of the Find Similar Samples operation, both as a bar graph ordered by correlation and a list of samples.

Find Similar Samples

Scripts and External Programs

Scripts

What is a Script?

Scripts are time-saving tools allowing a long series of data analysis steps to be performed at once. Scripts are re-usable and can be applied to any data set. You can create your own scripts using the ScriptEditor.

Scripts in GeneSpring

Eleven predefined scripts are included with GeneSpring:

- **2-fold Expression Change**—This script makes a gene list of all genes in a selected experiment that are 2-fold overexpressed or 2-fold underexpressed in at least one condition.
- **2-fold Expression Change AND Filter on Noise NOT Input**—This script combines the 2-fold Expression Change List and Filter on Noise scripts to produce a single gene list that passes both filters but does not have any genes on the input gene list.
- **Best k-means**—Given an experiment and a gene list, this script creates four k-means classifications with three, five, eight and 15 clusters respectively and selects the classification with the highest explained variability. The selected k-means appears in a results window.
- **Clustering 2-fold Change List**—This script creates a gene list of all genes in a selected experiment that are 2-fold overexpressed or 2-fold underexpressed in at least one condition and then creates a gene tree, an condition tree, a k-means classification, and a self-organizing map.
- **Filter on Noise**—Creates a list of genes that have control strengths equal to or greater than a user-supplied cutoff in at least half of the conditions in the experiment.
- **Find List of Similar Genes**—This script makes a gene list for each of the genes in a selected experiment if there are at least five genes with similar expression profiles.
- **Pairwise Comparison**—Returns a list of genes that are two-fold overexpressed in at least one condition in an experiment at compared with a specified experiment.
- **Probe Entire Enterprise Repository for Similar Conditions (PEER-C)**—Given a condition, this script searches through GeNet for conditions similar to the input condition. If the same sample is normalized differently in different experiments, both normalizations are compared.
- **Select k-means**—Given an experiment and a gene list, this script creates two k-means classifications with the numbers of clusters specified by the use and chooses the k-means cluster with the highest explained variability as the result.
- **Send Clustering Results to GeNet**—This script creates a gene tree, condition tree, k-means classification, and self-organizing map using a list of all genes in an experiment that are 2-fold overexpressed or 2-fold underexpressed in at least one condition and automatically sends the results to GeNet.

- **Series of k-means (increments of 5)**—Generates 10 k-means classifications, each with a differing number of starting sets and returns the classification with the highest explained variability.

The Run Script Window

From this window you can execute scripts and view information about them. If you have a connection to GeNet and are using Remote Execution Servers, you can execute the script on a remote computer.

At the bottom of the screen is a button labeled **View Script**. Click this button to open a new window containing a graphical representation of the script. On this screen, click the **Details** button to view detailed information in the Script Inspector. For more information on the Script Inspector, see “The Script Inspector” on page 8-5.

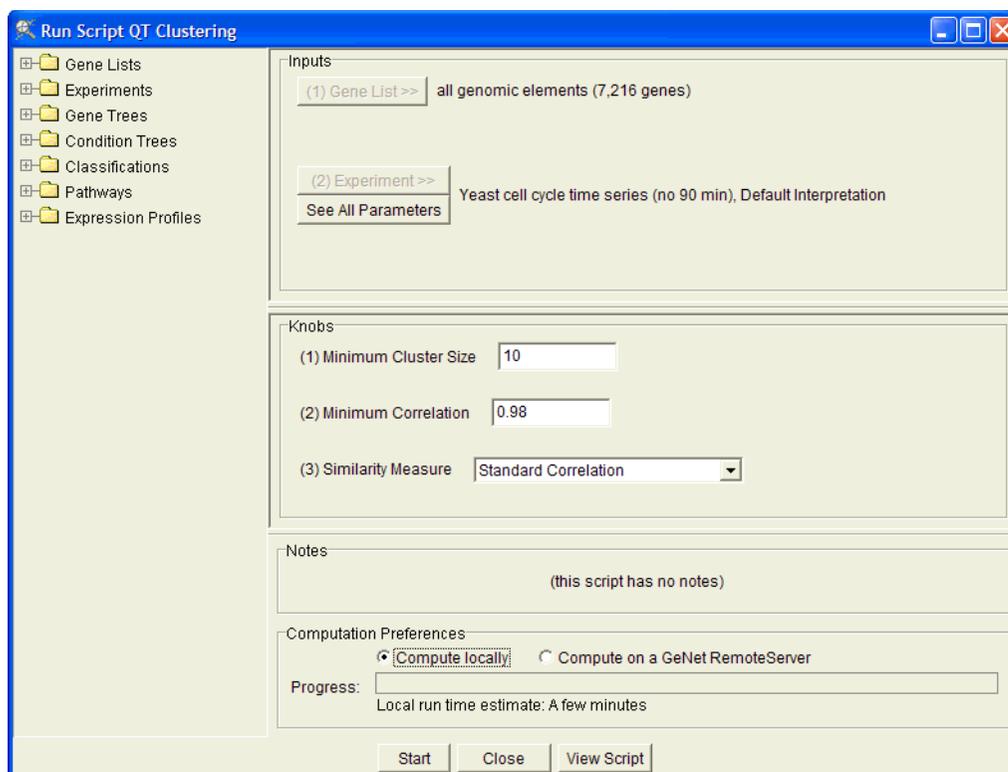


Figure 8-1 The Run Script window

To execute a script:

1. In the navigator, open the **Scripts** folder.
2. Select one of the demo scripts. The Run Script window appears, displaying various elements of the selected script.
3. Select a data object from the navigator panel and click the appropriate button in the Inputs box.

The example script in Figure 8-1 requires one data object: a condition. Some scripts need no input. Select a condition in the navigator and click **Set Condition >>**.

4. Set any specified knobs. In this particular script, you must specify a cutoff value. In other scripts, you might select a value from a pull-down menu to direct the execution of the script. The number of knobs can vary greatly, but they all appear in the Knobs box. Scroll down to make sure that all the text boxes are filled in. Not all knobs require user input.

The ScriptEditor does not recognize numbers with spaces or commas. Use periods (.) as your decimal markers.

If your script requires an array of numbers (for example, the weights associated with a complex correlation), a table appears in which you can enter these numbers. Enter one number per line. You can also paste numbers into this table in tab-delimited format from a text or excel file. The order of these numbers must match the order of the inputs that they describe.

5. Specify whether to run the script locally or on a Remote Execution Server by selecting the appropriate radio button.

If the desired server is not already in the list, select **Add New GeNet . . .** and enter the requested information in the **Edit GeNet Server** dialog that appears.

An estimate of the time required to run the script also appears in this part of the screen. These are loose estimates, which are described as follows:

- Seconds
- Less than a minute
- A few minutes
- Less than an hour
- A few hours
- Many hours...

6. Click **Start**. This button is not active until all required data has been entered.

If your script will upload data to a regulatory compliant GeNet, you will be prompted to provide an electronic signature. For more information, see “Electronic Signatures” on page 9-14.

7. When the script has finished running, the Results window appears. The appearance of the Results window varies depending on which type of script you run.
8. If the script returns a data object, name your results and click **Save**. You can also add and save information in the Notes area.

If you do not want to keep the results of your script, click **Cancel**.

Specifying Parameters for Data File Restriction

Some of the sample scripts included with GeneSpring require you to enter parameters for data file restriction. Which data file format to search, and which columns in that data file format that are to be searched is entered as a special text string in the Filter Columns Specification knob.

Use curly brackets to indicate the file format. Within the curly brackets give the column number or the column header of the column(s) to search. The order of the curly brackets must match the order of the tabs given in GeneSpring's GUI.

The first tab is the format of the majority of samples in the experiment; the second tab defines the second-most common format for that experiment, etc. If the formats describe the same number of samples, look at the GeneSpring GUI to determine the order of the formats.

Example:

```
{Column Number or Column Name1; Column Number or Column Name2} {} {}
{1;3} {} {"identifier"}
```

The Script Inspector

You can right-click any script within GeneSpring and select **Inspect** to examine that particular script. You can double-click any building block to view the building blocks that make up that block. You can also reach this screen by clicking **View** in the Run Script window and then clicking **Inspect**.

You can also edit the notes and history of your script. Click **Edit** to change the Author, Organization, Original Source, Other Software, Date and Note fields.

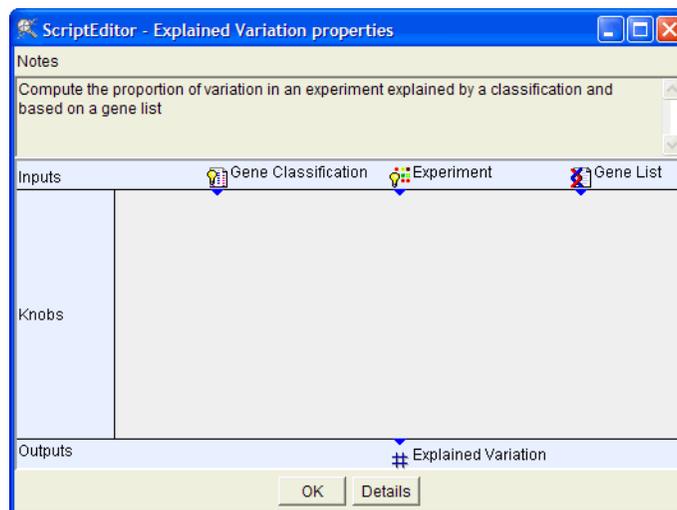


Figure 8-2 GeneSpring's Change Information window

Using the Remote Server

If you are using GeNet, and GeNet is enabled with remote execution servers and there is at least one remote server installed and configured, you can execute a script on a remote computer. The results are returned to GeneSpring when the script completes. Using a remote execution server is highly recommended when you are performing time consuming tasks such as clustering very large data sets. The remote execution server can remove the burden on your local computer while it completes the necessary computation.

To send a script to a remote execution server, select the **Compute on a GeNet Server** radio button on the Run Script window and click **Start**. Your script is sent to an available execution server and either executed immediately or placed in a queue.

To view the status of a script that may be waiting in the queue, select **Tools > Check Remote Execution Queue**. A window appears similar to the one in Figure 8-3.

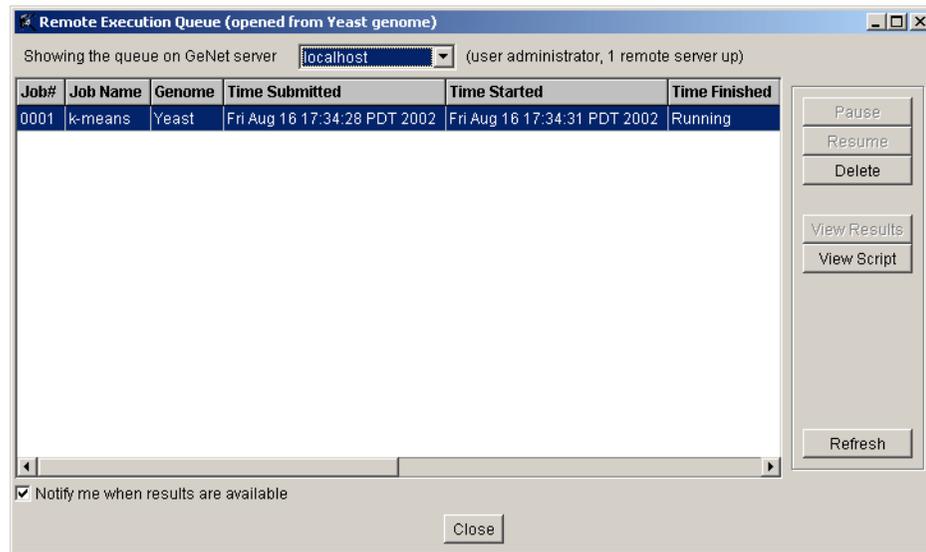


Figure 8-3 The Remote Execution Queue

The Remote Execution Queue displays the status of all jobs, pending and completed. The following information is available for each job:

- **Job#**—A unique identifier assigned by GeNet for each job.
- **Job Name**—The name of the script sent to the server.
- **Genome**—The genome from which the data to be analyzed originates.
- **Time Submitted**—The time that the user launched the script from GeneSpring.
- **Time Started**—The time that the remote execution server began executing the script. If the script is still waiting to be executed this column reads “Pending.” The column reads “Suspended” if the script was paused.
- **Time Finished**—The time that the execution server finished running the script.

To pause the execution of a pending script so that another script can run first, select the row of the desired job and click **Pause**. To resume running the script, select the desired row and click **Resume**. You cannot pause a script once it has begun executing.

This screen does not refresh automatically. To view the current status, click **Refresh**.

Each script in the queue can be viewed by clicking **View**.

Once the script has finished running on the remote server, the **View Results** button becomes available. Click this button to retrieve and save the results of your script. A series of file-dialog windows appears that correspond to each of the outputs that your script generates.

Using the ScriptEditor

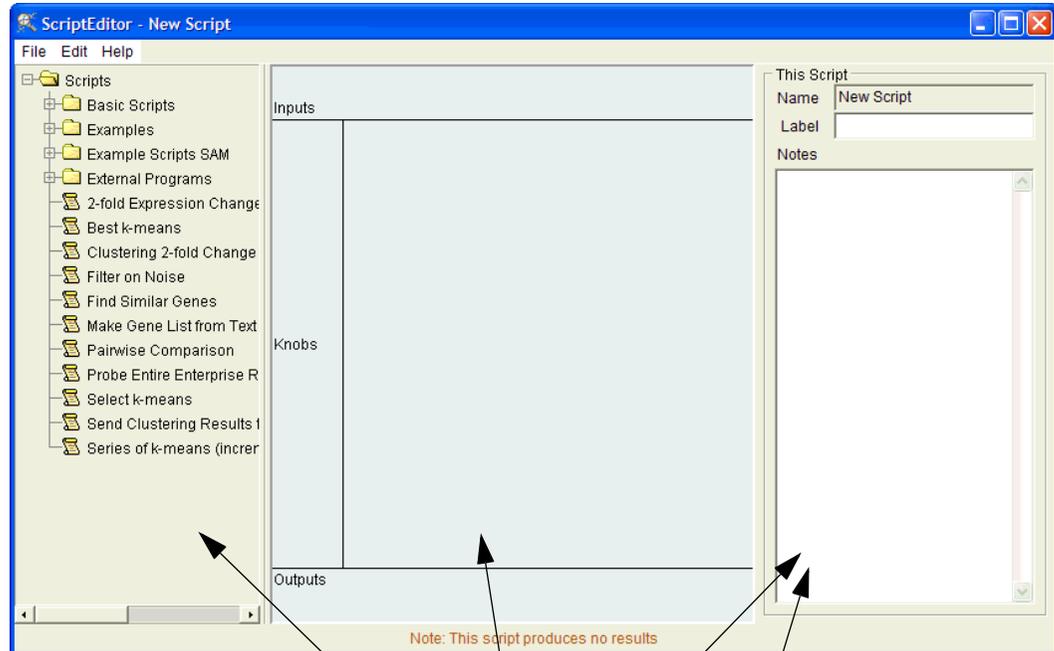
To open the script editor, select **Tools > ScriptEditor**.

ScriptEditor Concepts

- **Script building block**—Performs a simple function. This is the most basic element of a script.
- **Script**—A more complex program made up of script building blocks and/or scripts and external program building blocks.
- **Building block**—One of the building blocks used to create scripts, i.e., a script primitive, script, or external program building block.
- **External program building block**—A building block that is created for each external program defined within GeneSpring. A script with an external program building block can only be run on a version of GeneSpring with that external program installed. For more information on GeneSpring’s External Program Interface, see “External Programs” on page 8-35.
- **Socket**—Parts of a building block that send or receive a connection from another building block. Inputs and outputs are sockets.
- **Building block input**—Parts of a building block that receive information. Inputs are located on the top of a building block, and receive connections only from outputs.
- **Building block output**—Parts of a building block that send the results of the operation performed. Outputs are located on the bottom of a building block, and receive connections only from inputs.
- **Script input**—The socket at the top of a script. Script inputs are created when you drag a building block input into the “Inputs” area of the browser.
- **Script output**—The final output of a script. Script outputs are created when you drag a building block output into the “Outputs” area of the browser.
- **Knob**—A value entered when running the script.

The ScriptEditor Interface

The ScriptEditor’s main screen contains three panels: the *navigator*, *browser* and *notes*.



The ScriptEditor screen's *navigator, browser and block or notes* sections

Figure 8-1 The ScriptEditor workspace

The ScriptEditor Navigator

The navigator contains the building blocks (building blocks, scripts, external program building blocks) you use to create your scripts. Click the + (plus sign) next to a folder to open or close it.

The ScriptEditor Browser

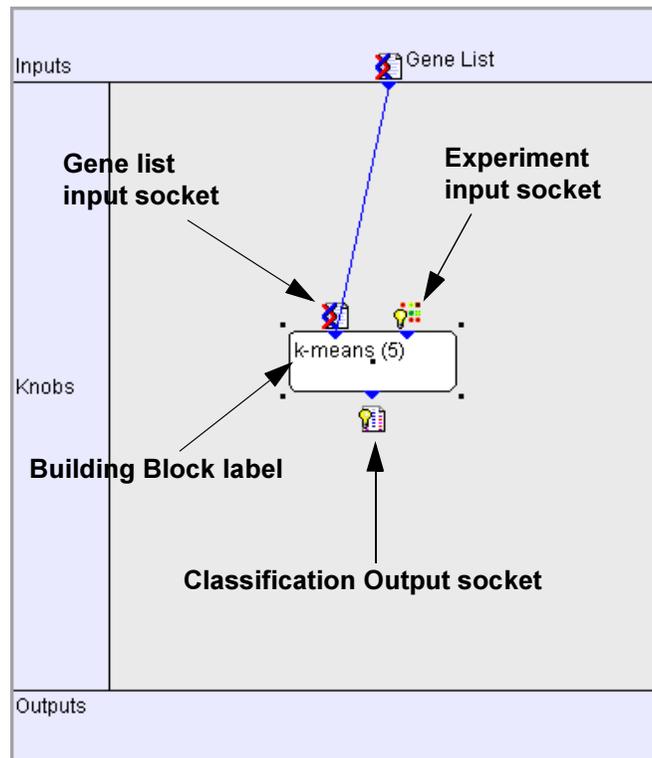


Figure 8-2 The Browser

To view an existing script or any of the building blocks:

1. Select an item in the navigator.
2. Drag the item to the browser area.
3. Click to view the selected item.

ScriptEditor Notes

The **This Script** section displays the following information about the currently selected item:

- **Name**—The name of the selected script.
- **Label**—A label for the script, distinct from the script name. You can enter a brief description of the script to make it easier to identify. The Label is displayed instead of the Script Name when the script is used as a block inside another script.
- **Notes**—You can alter these notes in any way you wish and add a nearly unlimited amount of text. These notes appear in the script and in the **Properties** box when the script is visible in GeneSpring.

If no item is selected, it displays empty boxes as it would for a new script. See Figure 8-1 for an example.

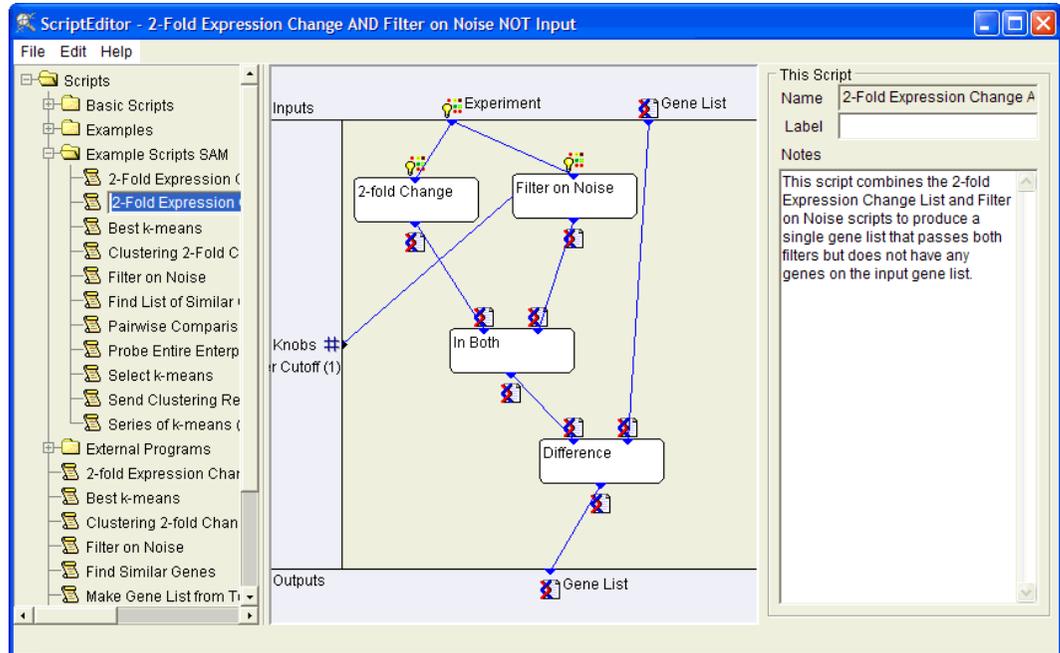


Figure 8-3 The main ScriptEditor window

The ScriptEditor Block for Building Blocks

The block section displays the following information on the currently selected item:

- **Name**—The name of the selected building block
- **Notes**—The function of the selected building block
- **Basic**—This box can display a variety of information about the selected building block. The example in Figure 8-4 shows information about the knobs associated with the selected building block. These items are different for each building block. You may see several types of input fields including pull-down menus and text boxes.

You can get context-sensitive help on any script building block by right-clicking on it and selecting **Help**. Double-clicking a building block brings up an inspector window for that building block.

In Figure 8-4, the selected item in the browser is the primitive “Filter Genes with Associated Numbers”.

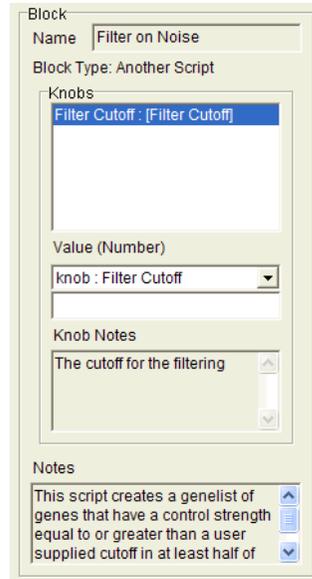


Figure 8-4 The Block section for building blocks

The Icon Legend

There are many icons in the ScriptEditor intended to help you keep track of the objects available. To view information on these icons, select **Help > Icon Legend**. You may want to leave the Icon Legend open and visible on your desktop until you are accustomed to working with the icons.

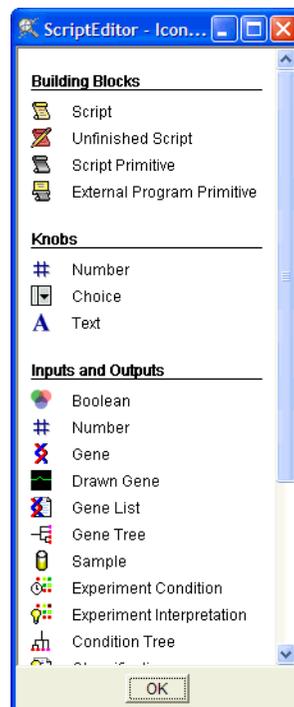


Figure 8-5 The Icon Legend

The Properties Panel

The Properties panel allows you to view many aspects of a script.

To view script properties, right-click on a script, script primitive or external program primitive in the navigator panel and select **Inspect** from the menu.

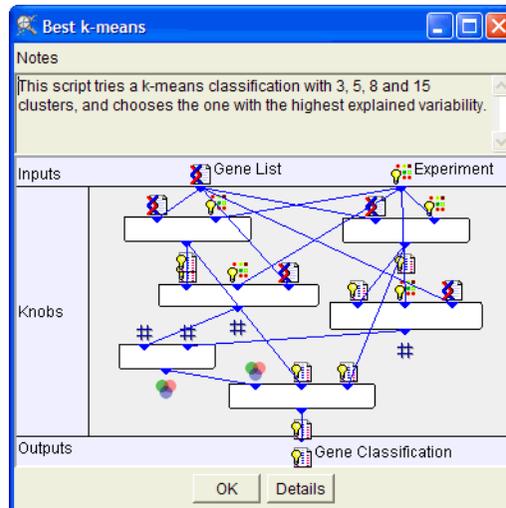


Figure 8-6 The Properties panel for a Script

Click **Details** for information on the script's author and creation date.

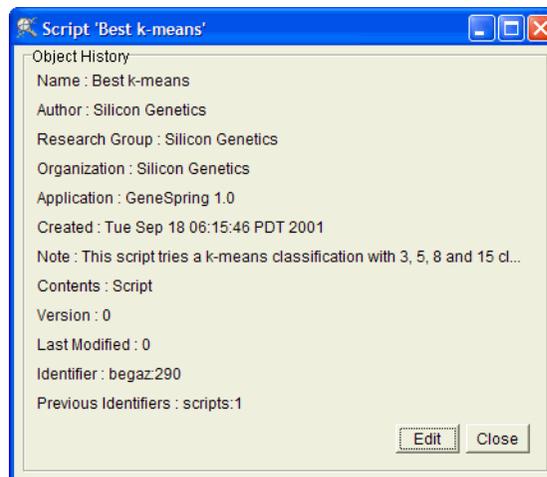


Figure 8-7 The Details (or History) panel

Click **Edit** to view the Change Information panel. You can modify the information in the white text boxes. Click **OK** to save your changes.

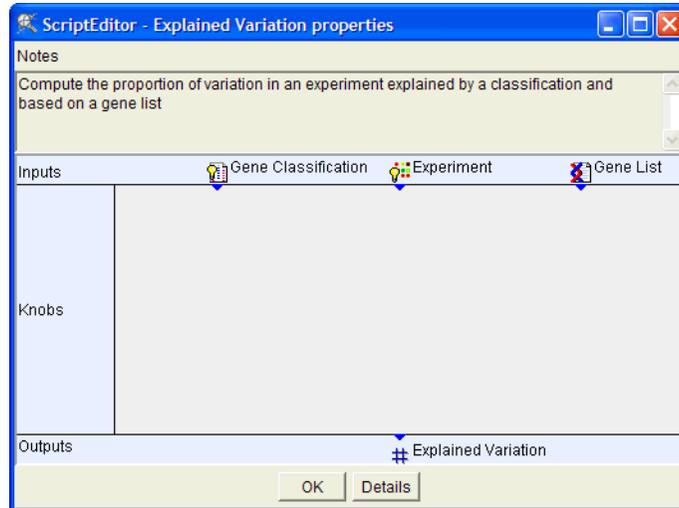


Figure 8-8 The Properties panel for a Building Block

Creating Scripts

Scripts can be created out of the most basic scripting elements, known as building blocks, as well as out of other scripts. You can combine many simple scripts to create more complex scripts.

Building Blocks

Three types of data objects can be used as building blocks in the creation of scripts:

- scripts
- script building blocks
- external program building blocks

Within a script, data is passed from one building block to another as the script runs. To create a new script, drag one or more building blocks from the navigator to the browser, then drag the cursor from the input socket of one building block to the matching output socket of another building block, or vice versa. A line appears between the two sockets.

For detailed descriptions of the available building blocks, see “Script Building Blocks” on page 8-20.

Inputs and Outputs

There are two types of inputs and outputs:

Building block inputs and outputs

Building block inputs and outputs are used to enter information into and retrieve information from a script. Building block inputs and outputs are found at the top and bottom of a given building block, respectively.

Script inputs and outputs

Script inputs are created when you drag a building block input into the “Inputs” area of the browser. Script outputs are created when you drag a building block output into the “Outputs” area of the browser.

There are three types of information produced by scripts that cannot be saved in GeneSpring. This is because there is no way to manipulate this information within GeneSpring.

- **Boolean**—A simple results box appears for this output, typically either true or false. See “Boolean” on page 8-20 for details.
- **Numbers**—A simple results box appears containing a list of numbers for this output. See “Numbers” on page 8-26 for details.
- **Sequence Information**—This information is displayed in a copy-and-pasteable table similar to the Potential Regulatory Sequences table in GeneSpring.

Knobs

Knobs allow the user running the script to enter a value when the script is run. For example, you could filter with a knob that sets the minimum normalized expression level.

Information for inputs, outputs, and knobs is entered in GeneSpring at the time you run the script. Constants required for building blocks are entered on the right side of the ScriptEditor screen, in the ScriptEditor Block area.

To create a knob:

1. Select an appropriate building block from the navigator.
2. Right-click in the Knobs area of the ScriptEditor Block and select **Make new knob** from the pop-up menu.
3. Select the appropriate variable type. The available types are:
 - Integer
 - Positive Integer
 - Number
 - Positive Number
 - Yes or No
 - Measurement Type
 - Percentage
 - Correlation
 - Name
 - Comparison
 - Chromosome Number
 - Gene Annotations
4. If desired, enter a default value for the knob in the Default Value field in the Notes area.

A Sample Script

Scripts are very simple when stripped of their fancy verbiage and icons. For example, you might want to examine your data as follows:

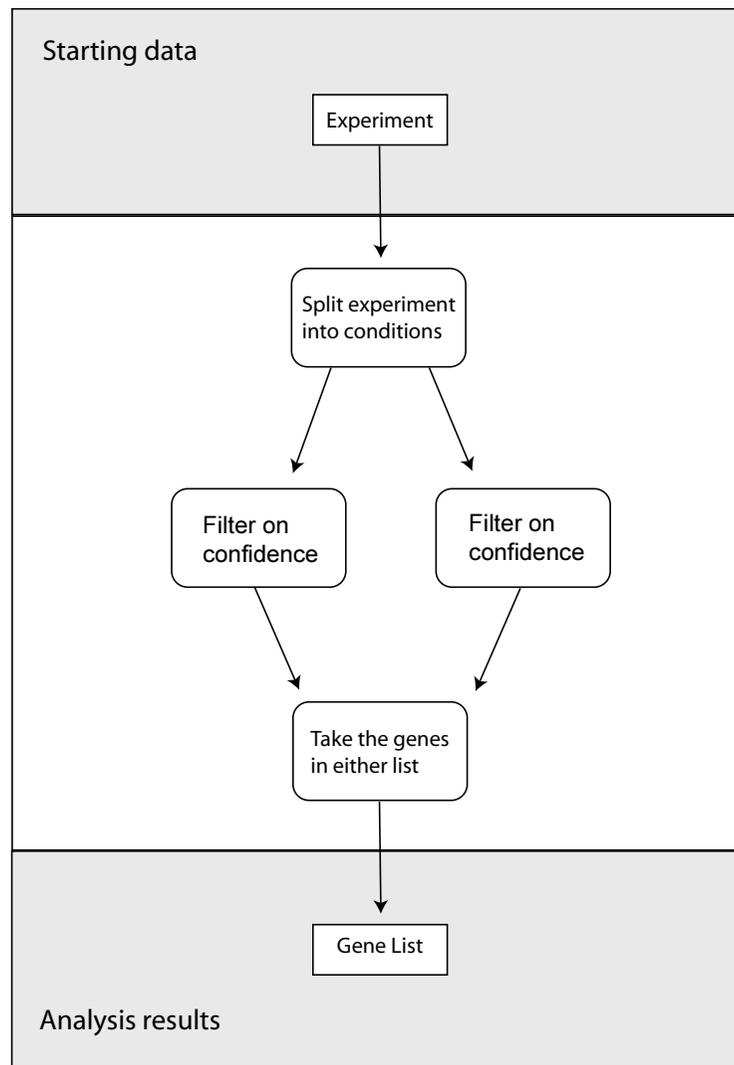


Figure 1-1 A simple flow chart of a script

In GeneSpring's Script Editor, it would look like this:

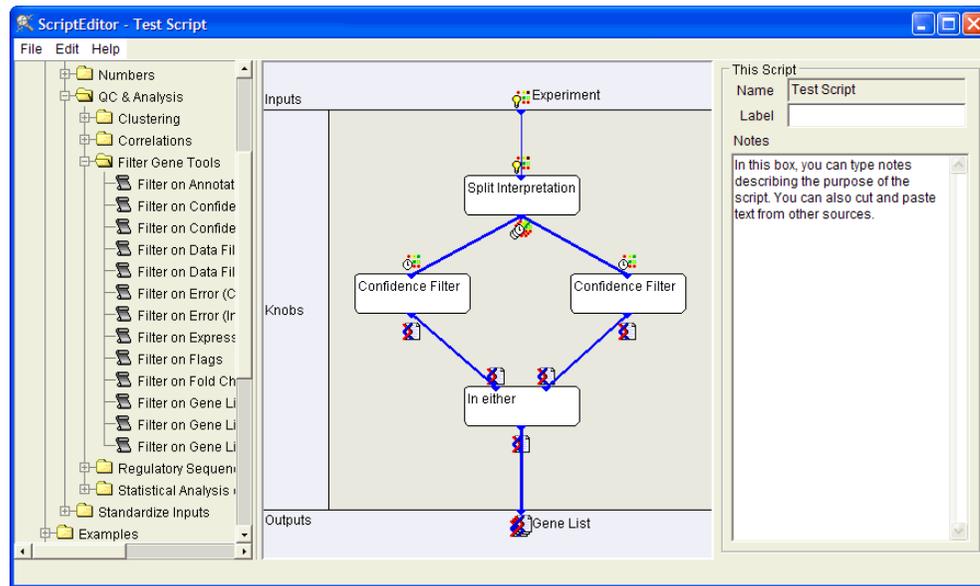


Figure 1-2 A Sample Script

Everything in the browser must be correctly connected to another object. Check the bottom of the screen for error or warning messages. If any error messages are present, the script cannot be run. You can run a script with warning messages active, but it may not function as intended.

When you are ready to save your script, see “Saving Scripts” on page 8-18.

Building Scripts

A typical script is built by arranging building blocks in the browser. Building blocks can be other scripts, script building blocks or external program building blocks.

To build a script:

1. Select a building block and click in the browser to place it. Click and drag the edges of the building block to make it larger or smaller.

Details about the building block appear in the **Block** section on the right side of the main ScriptEditor window as long as it is selected. You can click anywhere else in the browser to de-select the building block.

2. Set knobs, if necessary. These may be pull-down menus or text fields, and appear in the lower right section of the ScriptEditor window.
3. Create a line from the input (top) to the first building block.
4. Select a new building block from the folders in the navigator.
5. Connect the first building block to the second.
6. Repeat steps 2 through 6 as necessary.

7. When you are close to done, check the error messages at the bottom to make sure everything is connected properly.
8. Name and save your script. See “Saving Scripts” on page 8-18.

If you would like to create a new script immediately, select **File > New**. You can make several small scripts and join them together.

Saving a Building Block as a Script

You can save a block built from multiple other blocks as a script. To do this, right-click on the block and select **Save Block as Script**. This is useful for updating obsolete script building blocks, or if you want to use a block from a script someone else created.

Arranging Inputs and Outputs in the Browser

When you right-click over a socket icon in the browser, a pop-up menu appears. Which options appear depends on the status of the selected socket.

- **Move to Start**—This command moves the socket all the way to the left of the inputs or outputs field. It allows you to “untangle” the web of connecting lines.
- **Move to End**—This command moves the socket all the way to the right of the inputs or outputs field. It allows you to “untangle” the web of connecting lines.
- **Delete**—You will not get a warning message before the building block is removed from your potential script.

A number appears in parentheses after each socket in the input field. The numbers represent the order in which the input will be requested when the script is run.

Dynamically Naming Script Outputs

You can define output names dynamically based on the value of a script input or parameter. For example, the sample script Best k-means has two inputs, a gene list and an experiment. If you want the script output to include the name of the script, the name of the gene list, and the name of the experiment, do the following:

1. Select the appropriate script output socket.
2. In the **Make Name** field in the Notes section of the ScriptEditor, enter the following:

```
Best k-means $name$1 in $name2
```

3. Save the script.

If you run Best k-means using the gene list “ACGCGT in all ORFs” and the experiment “Extraterrestrial Yeast Study”, the output classification is automatically named “Best k-means ACGCGT in all ORFs in Extraterrestrial Yeast Study”.

The procedure is similar when basing output names on knob values, except that the format for knob values is “\$parameter\$x”, where *x* is the number of the knob. So, for example, to cause the output of the Filter on Noise sample script to include the name of the input experiment and the value of the Filter Cutoff knob, enter the following in the **Make Name** field:

```
Filter on Noise on $name$1 with Filter cutoff $param$1
```

If your input experiment is “Extraterrestrial Yeast Study” and the filter cutoff is “0.1”, the output gene list is named “Filter on Noise on Extraterrestrial Yeast Study with Filter cutoff 0.1”.

Any character that is legal in the name of a navigator object is legal in the **Make Name** field. The only illegal characters are the slash (/) and backslash (\).

Note: There is an 80-character limit for navigator object names. If the dynamically generated name of the script output is too long for GeneSpring, you will be prompted to change the name before you can save it. This may be problematic in cases where a script outputs multiple gene lists or other objects.

Saving Scripts

You can save either finished or unfinished scripts. Unfinished scripts are displayed using a different icon in the Navigator than finished scripts.

To save a script, select **File > Save**.

The first time you save a script, a **Save As** window appears in which you can specify a name for your script and a folder in which to save it. If an error message appears saying your result cannot be saved, rename the script and try again.

Folders

By default, the ScriptEditor saves scripts in the scripts folder. You can create a subfolder within this folder by right-clicking the parent folder in the Navigator and selecting **Add Folder**.

If you enter a nonexistent folder name in the **Save As** window, the ScriptEditor creates the directory for you and saves your script in it.

Moving Scripts

To move a script, click it in the navigator and drag it to the desired location or right-click it and select the **Move** command.

Warning Messages

If GeneSpring detects any problems or missing information in your script, warning or error messages may appear across the bottom of the screen.

- Warning messages are always preceded by the word *Warning* and are displayed in bright red text.
- Error messages appear in dark red text.

You can save a script when a warning or error message is active, but it may not perform as expected. You cannot run a script with an active error message.

Script Help

You can get help writing scripts as well as all other Silicon Genetics products by contacting Silicon Genetics Technical Support at 650-367-9600 or support@sigenetics.com.

To view current information on the Silicon Genetics web site, select **Help > Frequently Asked Questions**. You are directed to the following URL:

<http://www.sigenetics.com/GeneSpring/faq/index.html>

Script Building Blocks

The ScriptEditor comes with a predefined set of building blocks you can join together in various ways to build scripts. There are several categories of building blocks:

- *Boolean*
- *Boolean Select*
- *Gene List Manipulations*
- *GeNet Downloading (Default Directory)*
- *GeNet Downloading (Specified Directory)*
- *Look Up*
- *Merge-Split Groups*
- *Make Groups*
- *Select Groups*
- *Numbers*
- *Count Groups*
- *Clustering*
- *Correlations*
- *Filtering*
- *Regulatory Sequences*
- *Statistical Analysis (ANOVA)*
- *Standardize Inputs*

You can combine various building blocks to create a script. For very long or complex scripts, you may want to create several small scripts and join them together in the ScriptEditor to create the final script.

To view details on a script primitive, right-click it in the navigator and select **Properties** from the pop-up menu.

Boolean

Open **Scripts > Basic Scripts > Boolean**.

Name	Input	Output	Description
Boolean	No direct input	Boolean	Generates a <i>true</i> or <i>false</i> result. Select <i>yes</i> (true) or <i>no</i> (false) when you run the script.
Boolean AND	At least two Booleans	Boolean	Output is <i>true</i> only if both inputs are <i>true</i> .
Boolean FALSE	None	Boolean	Returns the result <i>false</i> .

Name	Input	Output	Description
Boolean NOT	One Boolean	Boolean	Output is <i>true</i> only if the input is <i>false</i> (converts true to false and false to true).
Boolean OR	Two Booleans	Boolean	Output is <i>true</i> if either input is <i>true</i> .
Boolean TRUE	None	Boolean	Returns the result <i>true</i> .

Boolean Select

Open Scripts > Basic Scripts > Boolean > Select Using Boolean.

Name	Input	Output	Description
Select Boolean	Three Booleans	Boolean	Selects the second boolean input if the first input is <i>true</i> and selects the third boolean input if the first input is <i>false</i> .
Select Condition	One Boolean, two conditions	Condition	Selects the first condition if <i>true</i> , the second condition if <i>false</i> .
Select Condition Tree	One Boolean, two condition trees	Condition tree	Selects the first tree if <i>true</i> , the second tree if <i>false</i> .
Select Experiment	One Boolean, two experiment interpretations	Experiment interpretation	Selects the first interpretation if <i>true</i> , the second if <i>false</i> .
Select Gene	One Boolean, two genes	Gene	Selects the first gene if <i>true</i> , the second if <i>false</i> .
Select Gene Classification	One Boolean, two classifications	Classification	Selects the first classification if <i>true</i> , the second if <i>false</i> .
Select Gene List	One Boolean, two gene lists	Gene list	Selects the first gene list if <i>true</i> , the second if <i>false</i> .
Select Gene Tree	One Boolean, two gene trees	Gene tree	Selects the first tree if <i>true</i> , the second if <i>false</i> .
Select Number	One Boolean, two numbers	Number	Selects the first number if <i>true</i> , the second if <i>false</i> .
Select Sequence	One Boolean, two sequences	Sequence	Selects the first sequence if <i>true</i> , the second if <i>false</i> .

Gene List Manipulations

Open Scripts > Basic Scripts > Gene List Manipulations.

Name	Input	Knobs	Description
All Genes	None	None	Outputs the list of all genes

Script Building Blocks

Name	Input	Knobs	Description
All Genomic	None	None	Outputs a list of all genomic elements
Count Genes in Gene List	At least one gene list	None	Outputs the number of genes in the input list(s).
Gene List Difference	Two gene lists	None	Outputs a list of the genes that are in the first gene list, but not the second.
Gene List Intersection	Two gene lists	None	Outputs a list of the genes that are in both input lists.
Gene List Inversion	Two gene lists	None	Outputs a list of the genes that are in either input list.
Gene List Union	Two gene lists	None	Outputs a list of the genes that are in either input list.
In all Gene Lists	One gene list group	None	Outputs a list of the genes in all of the input lists.
In at Least One	One gene list group	None	Outputs a list of the genes in at least one of the input lists.
Merge Gene List Group	At least one gene list group	Percentage, Comparison	Outputs a list of the genes in the specified proportion of the input lists.
In a Number of Gene Lists	Array of gene lists	Number of gene lists from the array a gene must be in, Comparison (>, <, =, <=, >=)	Outputs a list of genes that appear in at least (or other comparison) the specified number of gene lists.
In a Percentage of Gene Lists	Array of gene lists	Percentage, Comparison	Outputs a list of genes that appear in a specified proportion of the gene lists.
Sort Gene List	One gene list	none	Outputs the gene list sorted in descending order based on its associated numbers.

GeNet Downloading (Default Directory)

Note that you will need to login to GeNet before using a script to download data from GeNet. Open **Scripts > Basic Scripts > GeNet > GeNet Downloading > Default Directory**.

Name	Inputs	Knobs	Description
Download a Gene List from GeNet	None	Gene List name	Outputs a specified gene list retrieved from GeNet.
Download an Experiment from GeNet	None	Experiment Name	Outputs a specific experiment retrieved from GeNet.
Download all Gene Lists from GeNet	None	None	Outputs a group of gene lists retrieved from GeNet.

Name	Inputs	Knobs	Description
Download All Experiments from GeNet	None	None	Outputs a group of experiment interpretations retrieved from GeNet.

GeNet Downloading (Specified Directory)

Note that you will need to login to GeNet before using a script to download data from GeNet. Open **Scripts > Basic Scripts > GeNet > GeNet Downloading > Specified Directory**.

Name	Inputs	Knobs	Description
Download All Gene Lists from Directory in GeNet	None	Directory Name	Outputs an array of gene lists retrieved from the specified folder in GeNet.
Download All Experiments from Directory in GeNet	None	Directory Name	Outputs an experiment or array of experiments retrieved from the specified folder in GeNet. This allows you to hard-code a reference to an experiment or folder of experiments in your script.

GeNet Publishing (Default Directory)

Note that you will need to login to GeNet before using a script to autoload data to GeNet. Open **Scripts > Basic Scripts > GeNet Publishing > Default Directory**.

Name	Input	Knobs	Description
Send Classification to GeNet	One classification	None	Publishes a classification to your default directory in GeNet. No output to GeneSpring.
Send Experiment to GeNet	One experiment interpretation	None	Publishes an experiment interpretation to your default directory in GeNet. No output to GeneSpring.
Send Condition Tree to GeNet	One condition tree	None	Publishes an condition tree to your default directory in GeNet. No output to GeneSpring.
Send Gene List to GeNet	One gene list	None	Publishes a gene list to your default directory in GeNet. No output to GeneSpring.
Send Gene Tree to GeNet	One gene tree	None	Publishes a gene tree to your default directory in GeNet. No output to GeneSpring.

GeNet Publishing (Specific Directory)

Note that you will need to login to GeNet before using a script to autoload data to GeNet. Open **Scripts > Basic Scripts > GeNet Publishing > Specific Directory**.

Name	Input	Knobs	Description
Send Classification to Directory in GeNet	One classification	Directory	Publishes a classification to a chosen directory in GeNet. No output to GeneSpring.
Send Experiment to Directory in GeNet	One experiment interpretation	Directory	Publishes an experiment interpretation to a chosen directory in GeNet. No output to GeneSpring.
Send Condition Tree to Directory in GeNet	One condition tree	Directory	Publishes an condition tree to a chosen directory in GeNet. No output to GeneSpring.
Send Gene List to Directory in GeNet	One gene list	Directory	Publishes a gene list to a chosen directory in GeNet. No output to GeneSpring.
Send Gene Tree to Directory in GeNet	One gene tree	Directory	Publishes a gene tree to a chosen directory in GeNet. No output to GeneSpring.

Look Up

Open **Scripts > Basic Scripts > Look Up**.

Name	Input	Description
Is Gene in Gene List	One gene and one gene list	Return true if the gene list contains the input gene.
Number Associated with Gene in Condition	One Gene and one condition	Return the number (0 if none) associated with a gene in a condition. There is a knob for Type.
Number Associated with Gene in Gene List	At least one gene and one gene list	Return the number (0 if none) associated with a gene in a condition.

Merge-Split Groups

Open **Scripts > Basic Scripts > Merge-Split Groups**.

Name	Input	Description
Merge Genes	One gene group	Outputs a list containing all lists in the group input.
Merge Genes and Numbers	One gene group and one number group	Outputs a list of genes.
Split Classification	One classification	Splits the classification into a group of gene lists.

Name	Input	Description
Split Conditions	One experiment interpretation	Splits the experiment interpretation into a group of conditions.
Split Gene List	One gene list	Splits the gene list into a collection of individual genes.
Split Gene List With Numbers	At least one gene list	Splits the gene list into a group of genes and an associated group of numbers.

Make Groups

Open **Scripts** > **Basic Scripts** > **Merge-Split Groups** > **Make Groups**

Name	Input	Description
Make Classification Group	A folder of classifications	Produces a group of classifications.
Make Experiment Group	A folder of experiments	Outputs a group of interpretations. There is one knob, to select whether to include all the interpretations or only the defaults.
Make Gene List Group	A folder of gene lists	Outputs a group of gene lists.
Make Gene Tree Group	A folder of gene trees	Outputs a group of gene trees.

Select Groups

Open **Scripts** > **Basic Scripts** > **Merge-Split Groups** > **Select Groups**.

Name	Input	Description
Filter Boolean Group	Two Boolean groups	If true, pass the second argument for each Boolean through the corresponding first argument
Filter Condition Group	One Boolean group and one condition group	If true, pass the second argument for each Boolean through the corresponding first argument
Filter Experiment Group	One Boolean group and one experiment interpretation	If true, output an experiment interpretation for each Boolean in the first argument.
Filter Condition Tree Group	One Boolean group and one condition tree group	If true, output an condition tree for each Boolean in the first argument.
Filter Gene Classification Group	One Boolean group and one classification group	If true, output a classification for each Boolean in the first argument.
Filter Gene Group	One Boolean group and one gene group	If true, output a gene for each Boolean in the first argument.

Script Building Blocks

Name	Input	Description
Filter Gene List Group	One Boolean group and one gene list group	If true, output a gene list for each Boolean in the first argument.
Filter Gene Tree Group	One Boolean group and one gene tree group	If true, output a gene tree for each Boolean in the first argument.
Filter Number Group	One Boolean group and one number group	If true, output a number for each Boolean in the first argument.
Filter Sequence Group	One Boolean group and one sequence group	If true, output a sequence for each Boolean in the first argument.

Numbers

Open **Scripts** > **Basic Scripts** > **Numbers**.

Name	Input	Knobs	Description
Compare 1 Number	One number	Comparison, Number	Compare a number to another number and output a Boolean.
Compare 2 Numbers	Two numbers	Comparison	Compares two numbers and outputs a Boolean.
Number	None	Number	Output the number specified by a knob.
Number Add	Two numbers	None	Add two numbers together and output the result.
Number Divide	Two numbers	None	Divide one number by another and output the result.
Number Multiply	Two numbers	None	Multiply two numbers and output the result.
Number Subtract	Two numbers	None	Subtract the second number from the first number and output the result.
Number Log	Number	none	Output the log of the number specified in the input.
Sum of Numbers in Group	A group of numbers	none	Outputs the sum of the numbers in the input group.

Count Groups

Open Scripts > Basic Scripts > Numbers > Count Groups.

Name	Input	Knobs	Description
Count Conditions in Group	Array of Conditions	None	Determine the number of objects in the specified array and output the result.
Count Experiments in Group	Array of Experiments	None	Determine the number of objects in the specified array and output the result.
Count Gene Lists in Group	Array of Gene Lists	None	Determine the number of objects in the specified array and output the result.
Count Sequences in Group	Array of sequences	None	Determine the number of objects in the specified array and output the result.

Clustering

Open Scripts > Basic Scripts > QC & Analysis > Clustering.

Name	Input	Knobs	Description
Build Condition Tree	At least one gene list and one experiment interpretation	Correlation type, Separation ratio, Minimum distance	Outputs a condition tree
Build Gene Tree	One gene list, one experiment interpretation	Similarity measure, Merge similar branches, Discard bad, Separation ratio, Minimum distance, Do automatic annotation, Use standard	Outputs a gene tree. NOTE: See "Using the Remote Server" on page 8-5 for important details.
Explained Variation	At least one classification, one experiment interpretation, and one gene list.	None	Computes the proportion of variation in an experiment interpretation explained by a classification and a gene list. Output is a number between zero and one inclusive (i.e., 0.14567 is 14.567% explained variability).
Find Predictor Genes	One experiment interpretation and one gene list	Parameter name, number of genes	Outputs a list of genes that are good at predicting a given parameter in an experiment.

Script Building Blocks

Name	Input	Knobs	Description
K-means	One gene list, one experiment interpretation	Number of groups, Similarity measure, Maximum iterations, Additional tries, and Discard bad.	Outputs a k-means classification
K-means with Starting Classification	One gene list, one experiment interpretation, and one classification	Similarity measure, Number of iterations, and Discard bad.	Outputs a k-means clustering starting from an existing classification.
Self-Organizing Map	One gene list, one experiment interpretation	Iterations, Discard bad, Rows, Columns, and Radius.	Outputs self-organizing map.

Correlations

Open **Scripts** > **Basic Scripts** > **QC & Analysis** > **Correlations**.

Name	Input	Knobs	Description
Condition Correlation	Two conditions, one gene list	Correlation	Compare two conditions looking at only those genes in the specified gene list, and output a p-value.
Find Similar Conditions in Experiment	One condition, one experiment, one gene list	Correlation, Cut-off value	Compare the specified condition to every condition in the specified experiment, using only the genes in the specified gene list. Output an array of conditions and an associated array of p-values.
Find Similar Genes	One gene, one gene list, one or more experiment interpretations, one numeric value	Similarity measure, Minimum correlation, Maximum correlation	Outputs a list of genes whose expression profiles are correlated to a specified gene over the conditions of an interpretation.
Find Similar Genes (with custom input)	One gene, one gene list, one or more experiment interpretations, two numeric values, one or more conditions	Similarity measure, Minimum correlation, Maximum correlation, Conditions usage	Outputs a list of genes whose expression profiles are correlated to a specified gene over the conditions of an interpretation,
Find Similar Samples	One gene list, one target sample, one sample pool	Apply weights, Weighting coefficient	Outputs a list of samples correlated to a specified sample.

Name	Input	Knobs	Description
Gene Correlation	Two genes, one experiment	Correlation	Compare the two genes determine their similarity with respect to the selected experiment. Output a p-value.
Gene List Similarity p-Value	2 gene lists	None	Calculate the similarity between two gene lists, using the All Genes list as the Universe. Output a number representing the probability that the intersection between the two lists could be due to chance. Note that there is no multiple testing correction applied by this script.
Gene List Similarity p-Value, Specified Universe	3 gene lists	None	Calculate the similarity between two gene lists, using the specified gene list as the Universe. Output a number representing the probability that the intersection between the two lists could be due to chance. Note that there is no multiple testing correction applied by this script.

Filtering

Open Scripts > Basic Scripts > QC & Analysis > Filter Gene Tools.

Name	Input	Knobs	Description
Filter on Annotations	Gene list	The string to be searched	Outputs a list of genes whose annotations contain a specified text string.
Filter on Confidence (Condition Input)	One condition	Measure of confidence, Multiple testing correction, Minimum, Maximum	Filters on t-test p-value or number of replicates using a condition. Outputs a gene list.
Filter on Confidence (Interpretation Input)	One experiment interpretation	Measure of confidence, Multiple testing correction, Minimum, Maximum, Minimum # of conditions	Filters on t-test p-value or number of replicates using an experiment interpretation. Outputs a gene list.
Filter on Data File (Condition Input)	One condition	Filter method, Filter text, Use * as wildcard, Must appear in, # of samples, Filter columns specification	Filters on a data file using a condition. Outputs a gene list.

Script Building Blocks

Name	Input	Knobs	Description
Filter on Data File (Interpretation Input)	One experiment interpretation	Filter method, Filter text, Use * as wildcard, Must appear in, # of samples, Filter columns specification	Filters on a data file using an experiment interpretation. Outputs a gene list.
Filter on Error (Condition Input)	One condition	Error type, Minimum, Maximum	Filters on errors using a condition. Outputs a gene list.
Filter on Error (Interpretation Input)	One interpretation	Error type, Minimum, Maximum, Minimum # of conditions	Filters on errors using an experiment interpretation. Outputs a gene list.
Filter on Expression Level	One condition input	Data type, Minimum, Maximum, Minimum # of conditions	Outputs a gene list containing the genes that have a measurement relative to a cutoff.
Filter on Flags	One or more samples	Flag value, Minimum # of samples	Filters on flags and outputs a gene list.
Filter on Fold Change	One condition, one condition group	Data type, Comparison, Fold difference, Must appear in	
Filter on Gene List	Gene list	None	Outputs the input gene list.
Filter on Gene List Numbers	Gene list	Cutoff, Comparison	Produces a gene list (from an existing gene list) containing the genes whose associated number meets the specified criteria.
Filter on Gene List Numbers (In Range)	Gene list	Upper bound, Lower bound	Produces a gene list (from an existing gene list) containing genes whose associated number meets the specified criteria.

Regulatory Sequences

Open Scripts > Basic Scripts > QC & Analysis > Regulatory Sequence Search.

Name	Input	Knobs	Description
Find Genes with Specific Regulatory Sequence	One sequence	From Base, To Base, Maximum errors	Output a list of the genes that contain the input regulatory sequence.

Name	Input	Knobs	Description
Find Regulatory Sequences	One gene list	From Base, To Base, Minimum Length, Maximum Length, Minimum Errors, Maximum Errors, Minimum Interior N's, Relative Genomic, P-value Cut-off.	Outputs a list of regulatory sequences upstream of the genes in the input list.

Statistical Analysis (ANOVA)

Open Scripts > Basic Scripts > QC & Analysis > Statistical Analysis.

Name	Input	Knobs	Description
1-way ANOVA	One gene list, one experiment interpretation	Groups specification, Test type, P-value cut-off, Multiple testing correction	See "1-Way ANOVA" on page 6-34 for details.
1-way ANOVA with Post Hoc Tests	One gene list, one experiment interpretation	Groups specification, Test type, P-value cut-off, Multiple testing correction, Post hoc tests	See "1-Way ANOVA" on page 6-34 for details.
2-way ANOVA	One gene list, one experiment interpretation	Groups specification (1st parameter), Groups specification (2nd parameter), Test type, P-value cutoff, Multiple testing correction	See "2-Way ANOVA" on page 6-43 for details.
2-way ANOVA (Specific Result)	One gene list, one experiment interpretation	Groups specification (1st parameter), Groups specification (2nd parameter), Test type, P-value cutoff, Multiple testing correction, Restriction	See "2-Way ANOVA" on page 6-43 for details.

Standardize Inputs

Open Scripts > Basic Scripts > Standardize Inputs.

Name	Input	Knobs	Description
Use Specific Experiment	None	Experiment name	Specify a single local experiment by name. This is useful as an input to another block.
Use Specific Experiment Folder	None	Subfolders, Experiment folder name	Specify a single local experiment folder by name. This is useful as an input to another block.

Script Building Blocks

Name	Input	Knobs	Description
Use Specific Gene List	None	Gene list name	Specify a single local gene list by name. This is useful as an input to another block.
Use Specific Gene List Folder	None	Subfolders, Gene list folder name	Specify a single local gene list folder by name. This is useful as an input to another block.

Scripts to External Programs

External program building blocks are points of contact for other programs. External program building blocks are not editable, but can be used in any script. If no sample external program building blocks appear in your ScriptEditor's navigator, click the File Access link to download the samples from <http://www.sigenetics.com/cgi/SiG.cgi/Products/GeneSpring/extProgs.smf>.

Place the .jar file in the folder `..GeneSpring\data\Programs\`.

You may have more external program building blocks, as GeneSpring and ScriptEditor will create an external program primitive for every external program in your GeneSpring. You may not have any external program building blocks in the ScriptEditor if you are working on an older version on GeneSpring, or if you do not have any external programs.

Open **Building Blocks > External Programs**.

Name	Input	Description
Load Classification from File	A filename	Runs an external program to load and output a classification from a file on disk.
Load Experiment from File	A filename	Runs an external program to load and output an experiment from a file on disk.
Load Gene List from File	A filename	Runs an external program to load and output a gene list from a file on disk.
Load Gene List with Numbers from File	A filename	Runs an external program to load and output a gene list with associated numbers from a file on disk.
Load Tree from File	A filename	Runs an external program to load and output a gene tree from a file on disk.
Save Classification to File	One classification	Runs an external program to save a classification to disk. A save data window appears and prompts you to enter a filename.
Save Experiment to File	One Experiment	Runs an external program to save a classification to disk. A save data window appears and prompts you to enter a filename.
Save Gene List to File	One gene list	Runs an external program to save a gene list to disk. A save data window appears and prompts you to enter a filename.

Name	Input	Description
Save Gene List with Numbers to File	One gene list	Runs an external program to save a gene list with associated numbers to disk. A save data window appears and prompts you to enter a filename.
Save Tree to File	One gene tree	Runs an external program to save a gene tree to disk. A save data window appears and prompts you to enter a filename.

Scripts and External Programs

External programs are listed in the ScriptEditor under the navigator's external programs folder. In this folder you will find an external program building block for each external program defined in GeneSpring. Ten example external program building blocks are provided.

There are two significant types of external program building blocks:

- **Load data object to file** —the *loading* external programs make a specified data object from outside GeneSpring available for use in a script.
- **Save data object to file** —the *saving* external programs output a result from a script to your local hard drive. It does not save to GeneSpring or GeNet, as there are already building blocks that perform that function.

GeneSpring checks for new, changed or deleted external programs each time it is started.

External programs you have used in scripts cannot be deleted. If you have shared a script with a colleague, make sure your version of GeneSpring has all the same external programs as your colleague.

If an external program is not present in the local version of GeneSpring that is being used to run the script, a message announces that script is corrupted. Corrupted scripts cannot be run in GeneSpring.

For details on how to create external programs, please refer to the Silicon Genetics FAQs and “External Programs” on page 8-35.

External Programs

The GeneSpring External Program interface allows you to run external analysis programs from within GeneSpring. These programs can be useful when your research calls for a type of analysis that GeneSpring does not perform. The external program interface is also useful for parsing and pre-formatting data for use in another application.

When you launch an external program from within GeneSpring, the data that is displayed in the genome browser is sent to the external program as standard input. When the external program runs, GeneSpring recognizes the standard output generated by the external program and displays it in the genome browser.

To run an external program, double-click its name in the GeneSpring navigator, or right-click on it and select **Run**.

In earlier versions of GeneSpring it was necessary to manually create a .programdef file for each external program. In GeneSpring 5.1, this process is automated through the GeneSpring interface.

The New External Program Window

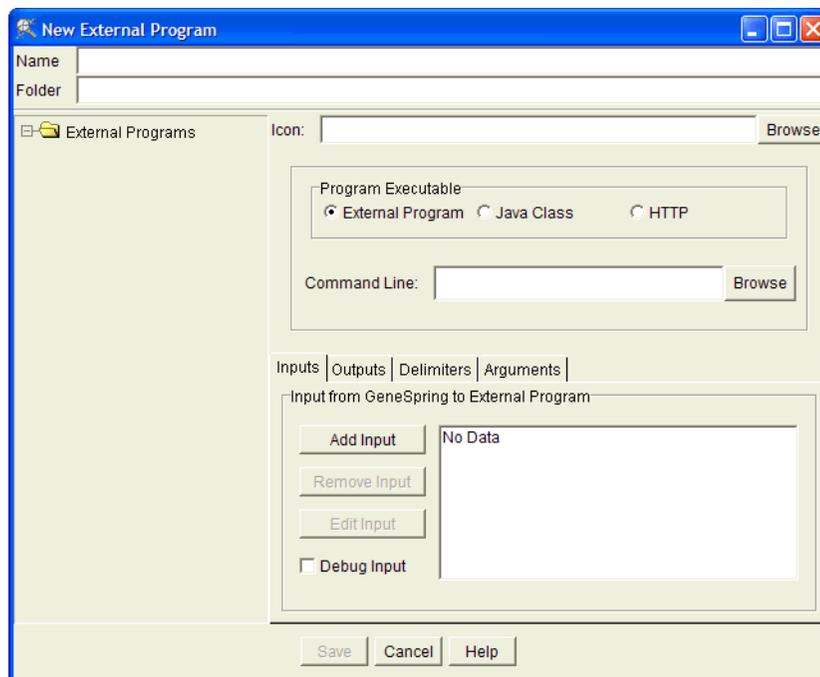


Figure 8-3 The New External Program window

From this window, you can specify the inputs, outputs, program type, and other information about an external program. Any program capable of receiving standard input can be run directly from Genespring.

To install a new external program, you must provide the following information:

External Programs

- **Name**—The name of the program as it will appear in the GeneSpring navigator. Choose a descriptive name that will be easy to remember.
- **Folder**—The folder in which to save the external program definition. To save the experiment in an existing folder, navigate to that folder in the directory browser in the lower left portion of the screen. The selected folder appears in the Folder field. To save in a new subfolder, navigate to the desired parent folder and enter a name for the new folder in the **Folder** field.
- **Program Executable**—The type of program (External program, Java class, or HTTP link).
- **Command Line**—The complete pathname required to run the program, i.e.,
c:\Program Files\Perl\ScriptCompletionHaiku.pl or http://ecoli.sample.com/analysis.cgi.
- **Inputs**—Data input for the external program. For more information, see “The Inputs Tab” on page 8-36.
- **Outputs**—Output of the external program. For more information, see “The Outputs Tab” on page 8-37.
- **Delimiters**—Delimiters to separate multiple outputs. For more information, see “The Delimiters Tab” on page 8-39.
- **Arguments**—Necessary command line arguments for the external program. For more information, see “The Arguments Tab” on page 8-39.

The Inputs Tab

The input is what GeneSpring sends to the external program. On this tab, specify any necessary input data for the external program. You can add as many inputs as you like.

To add a program input:

1. Click **Add Input**. The Choose Type of Input screen appears.

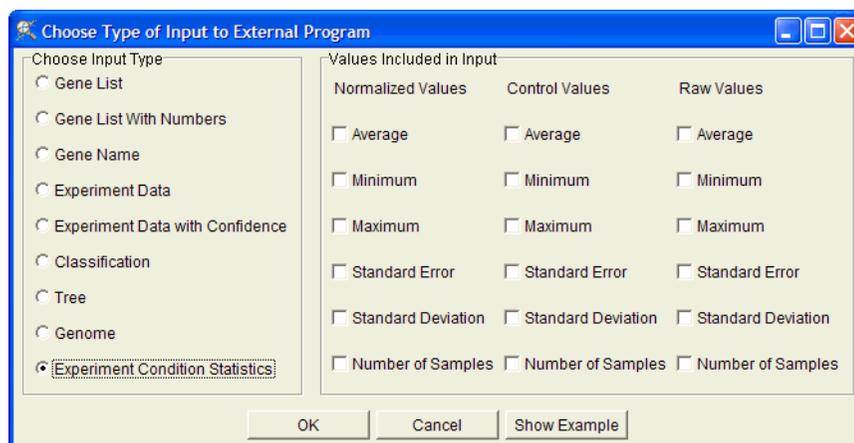


Figure 8-4 The Choose Input Type window

2. Select the type of input to send to the external program. The available choices are:
 - **Gene List**—A tab-delimited list of systematic names, one per line

- **Gene List With Numbers**—A tab-delimited list of systematic names and associated numbers, one pair per line
- **Gene Name**—A single systematic name
- **Experiment Data**—Normalized experiment data, one line per gene, one column per experiment, with header lines for the experiment name and each parameter. Only genes in the currently selected gene list are sent.
- **Experiment Data with Confidence**—Normalized experiment data, one line per gene, two lines per experiment (one for normalized data and one for control values), with header lines for the experiment name and each parameter. Only genes in the currently selected gene list are sent.
- **Classification**—A tab-delimited list of systematic names and the name of the associated classification group, one pair per line. Only genes in the currently selected gene list and classification are sent.
- **Tree**—A hierarchical tree in XML format
- **Genome**—An XML representation of the genome, which will be automatically saved to disk during Import to GeneSpring
- **Experiment Condition Statistics**—One gene per line, one column per statistical quantity per condition, with labels for the statistics across the top in the form `{N,R,C}_{AVERAGE, MIN, MAX, STDERR, STDDEV, N}`

Note: The Values Included in Input panel is active only if you selected the Experiment Condition Statistics radio button.

Click **Show Example** for an example of the selected output. Examples can be viewed as plain text, hex code, or as a spreadsheet. Non-displayable characters such as tabs are displayed as boxes.

3. Click **OK** to return to the New External Program window.

If desired, check the **Debug Input** box. This option writes to the console window when input is sent from GeneSpring to the external program.

To edit an existing input type, select it in the list box and click **Edit Input**. To remove an input type, select it in the list box and click **Remove Input**.

The Outputs Tab

The output is what GeneSpring receives from the external program. On this tab, specify the desired output of the external program. You may add as many outputs as you like. If the external program does not send any data back to GeneSpring, you do not need to enter anything in this tab.

To add an external program output:

1. Click **Add Output**. The Choose Type of Output window appears.

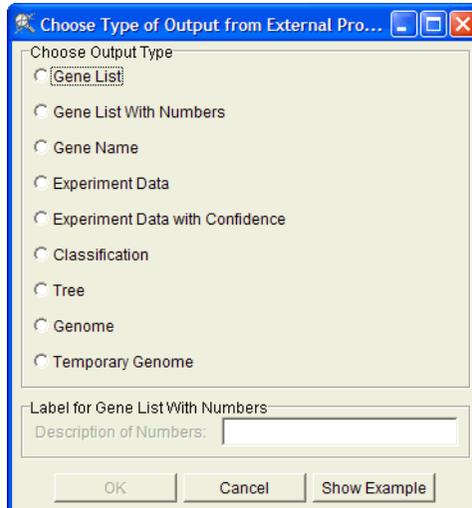


Figure 8-5 The Choose Type of Output window

2. Specify the type of output GeneSpring will receive from the external program. Available options are:
 - **Gene List**—A tab-delimited list of systematic names, one per line
 - **Gene List With Numbers**—A tab-delimited list of systematic names and associated numbers, one pair per line
 - **Gene Name**—A single systematic name
 - **Experiment Data**—Normalized experiment data, one line per gene, one column per experiment, with header lines for the experiment name and each parameter
 - **Experiment Data with Confidence**—Normalized experiment data, one line per gene, two lines per experiment (one for normalized data and one for control values), with header lines for the experiment name and each parameter
 - **Classification**—A tab-delimited list of systematic names and the name of the associated classification group, one pair per line
 - **Tree**—A hierarchical tree in XML format
 - **Genome**—An XML representation of the genome, which will be automatically saved to disk during Import to GeneSpring
 - **Temporary Genome**—Identical to the Genome format, except that it is not saved to disk

Click **Show Example** for an example of the selected output. Examples can be viewed as plain text, hex code, or as a spreadsheet. Non-displayable characters such as tabs are displayed as boxes.

3. If the output is a gene list with numbers, enter a label to describe what the numbers represent in the **Description of Numbers** field. This label appears in the Gene List Inspector as the title of the column of numbers.
4. Click **OK** to return to the New External Program window.

If desired, check the **Debug Output** box. This option writes to the console window when output is sent from the external program to GeneSpring.

To edit an existing output type, select it in the list box and click **Edit Output**. To remove an output type, select it in the list box and click **Remove Output**.

The Delimiters Tab

This tab is necessary only if your external program has multiple inputs or outputs.

Figure 8-6 The Delimiters Tab

Both GeneSpring and the external program need to know when a new data type is being sent. Certain characters are used to indicate this new data type. This is usually the ASCII 255 character, however, your program may require a different delimiter.

By default, GeneSpring uses ASCII 255 as the data type delimiter. To enter a custom delimiter, uncheck the **Use ASCII 255 as delimiter** box and enter the desired delimiter in the **Use custom delimiter string** text box. This may be a character or a string.

Some external programs may look for ASCII 255 to indicate that the data has finished being sent. If this is the case, check the **Terminate Last Input to Program with ASCII 255** box.

The Arguments Tab

Command line arguments are a way of providing extra information to the external program. For example, if the external program can perform one of three clustering methods, a command line argument might tell the external program which clustering method to use.

This tab is optional, since only some external programs require command line arguments.

Figure 8-7 The Arguments tab

This tab contains a table of name-value pairs displaying the name of the argument and its default value.

External Programs

To enter a new argument:

1. Click **Add Argument**. In the table to the right, a new line appears.
2. Replace the text `name1` and `value1` with the appropriate argument name and value, i.e., `-v` and `all`.

Some arguments may not have values. In this case, enter only the argument name, and delete the sample text `value1` and leave the Default Value field blank.
3. Select the separator between the the argument name and the argument value. This is either `=` or `:`.
4. Some versions of Windows will not correctly match up argument name/value pairs, and will read values as new arguments. To avoid this, check the **Fill in missing argument values** box and enter a filler term in the provided text box. This filler term should be something the external program will ignore, or a character that doesn't occur normally, such as ASCII 255.
5. Click **Save**.

To remove an argument, select its line in the table and click **Remove Argument**.

Running an External Program

1. Right-click the program in the GeneSpring navigator and click Run.
2. If your program takes the data from a tree or a classification as input, be sure these are selected and visible as well.
3. Open the external program folder in the navigator panel and click the program to run.

Examples

External Program Interface Example: SAS™ for Windows

This example demonstrates how to use GeneSpring's external program interface. The External Program Interface exports GeneSpring experimental data, runs a SAS™ program to analyze it, and brings the results back into GeneSpring for display. This example was developed with Windows 2000 using SAS™ version 8. It should work with earlier versions of Windows, but earlier versions of SAS™ require some modifications.

This particular example sets up an interface to the SAS™ procedure FASTCLUS to do gene clustering. You will need to create two text files with a text editor such as Microsoft NotePad™. These files are `Runsas.bat` and `Fastclus.sas`. These are each described below.

1. Create a batch file called `runsas.bat`. This batch file takes the standard input from GeneSpring, stores it in a file, executes SAS™, and passes the results back to GeneSpring via standard output. The program `cat.exe` simply copies standard input into standard output. If you do not have something equivalent on your system, `cat.exe` can be downloaded from the Silicon Genetics website.

Place the following text in the batch file:

```
@echo off
set infile=%2
```

```

set outfile=%3
cat.exe > %2
set SASROOT=C:\PROGRA~1\SASINS~1\SAS\V8
%SASROOT%\SAS %1.sas -nologo -config %SASROOT%\SASV8.CFG
cat.exe < %3
del %1.lst %1.log %2 %3

```

2. Create a text file called `fastclus.sas`. This batch file runs PROC FASTCLUS, specifying 5 clusters. In PROC IMPORT, the `datarow=3` command skips the first two lines of the exported data, which contain the dataset name and one parameter. If you have more than one parameter, adjust the data-row value accordingly.

PROC EXPORT puts a header line on the return data set listing the variable names, and GeneSpring displays an error message and skips this line (unless you have a gene named VAR1, in which case you should rename VAR1 to something else in your application).

Place the following text in the batch file:

```

filename infile "%sysget(infile)";
filename outfile "%sysget(outfile)";

proc import datafile=infile DBMS=TAB out=experiment replace;
datarow=3;
getnames=no;
run;

proc fastclus data=experiment maxclusters=5 maxiter=50
out=clusters(keep=var1 cluster);
id var1;
run;

proc export data=clusters outfile=outfile DBMS=TAB replace;

run;

```

3. Once you have saved the batch file, open File > New External Program from the GeneSpring menu and do the following:
 1. Enter “SAS FASTCLUS” in the Name field.
 2. Leave the Folder field blank. The external program is saved in the External Programs folder by default.
 3. Select the **External Program** radio button.
 4. Enter the following in the **Command Line** field:

```

runsas.bat fastclus expt.txt clus.txt

```
 5. On the Inputs tab, click **Add Input** and select the **Experiment Data** radio button.
 6. On the Outputs tab, click **Add Output** and select the **Experiment Data with Confidence** radio button.

7. Click **Save**.

Your external program should now appear in the GeneSpring Navigator in the “External Programs” folder.

The External Program Inspector

The External Program Inspector allows you to view details of an existing external program. To open the External Program Inspector, right-click an external program from the GeneSpring navigator and select **Inspect**.

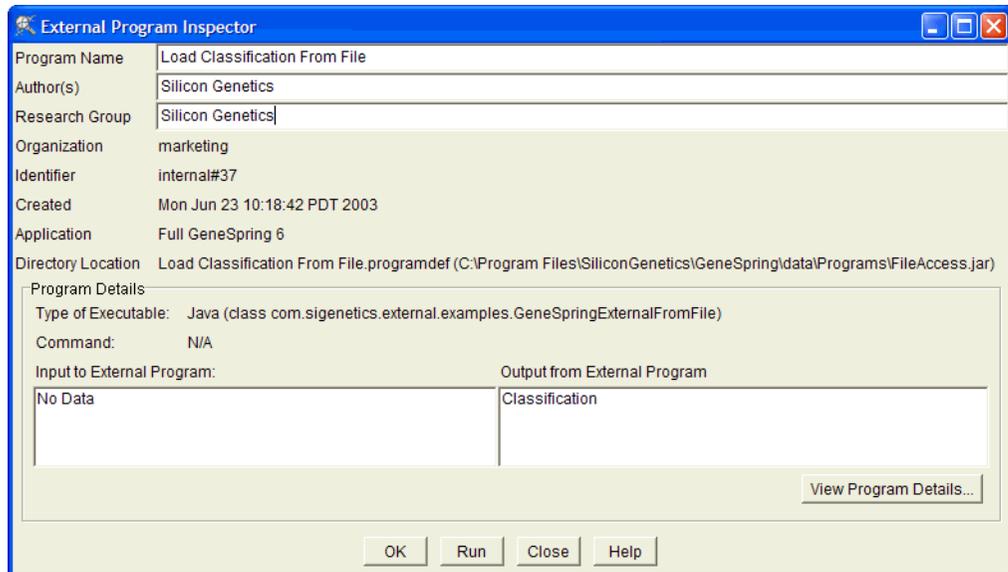


Figure 8-8 The External Program Inspector

The top portion of the screen displays the following information:

- Program Name
- Author(s)
- Research Group
- Organization
- Identifier
- Created
- Application
- Directory Location

You can edit the Program Name, Author, and Research Group fields. To modify these, enter the new information and click **OK**.

The Program Details panel contains the following information:

- **Type of Executable**—The executable type, i.e., Java class, external program, or URL
- **Command**— The command line, if the executable is an external program
- **Input to External Program**—The inputs defined for the external program
- **Output from External Program**—The outputs defined for the external program

If a program cannot be edited, such as the external programs included with GeneSpring, the button in the lower right portion of the screen is labeled **Program Details**. Click this to view additional details about the program. You cannot make any changes from the View Additional Program Details window.

If the program can be edited, the button reads **Edit Program**. Click **Edit Program** to view the Edit Program window. This window is identical to the Create New External Program window. See “The New External Program Window” on page 8-35 for more information.

Data Formats for the External Program Interface

You can choose what data and what format data is sent and received, according to the table below. All formats are optionally terminated in a special termination character, ascii 255. The formats are generally text files.

You can also send or receive multiple data objects. For example, your program might want to receive both the currently selected gene list and the currently selected experiment; or it might send a new genome to GeneSpring, followed by an experiment for that genome.

Format	Input	Output	Num	Description	Example
No data	Yes	Yes	0	No data. Typically used for one way communication.	
Gene List	Yes	Yes	1	List of gene names, one per line	L20294 M89777 X95403 M63630
Gene List (with numbers)	Yes	Yes	2	List of gene names, one per line, with each gene followed by a tab and an associated number	L20294 1 M89777 3 X95403 3.566 M63630 0
Gene Name	Yes	Yes	3	The name of one gene	L20294
Experiment Data	Yes	Yes	4	Experimental data. One line per gene, one column per experiment, with a header line for each parameter.	Tup1 deletion experiment time (minutes) 1 2 YPR1 0.88 1.09 YGR1 1.81 1.63 YNL1 0.52 1.18

External Programs

Format	Input	Output	Num	Description	Example
Experiment data with confidence	Yes	Yes	5	As above, except two columns per experiment. The first column has normalized data, the second has confidence values.	Tup1 deletion experiment time (minutes) 1 control 2 control YPR1 8 43 9 70 YGR1 1 7.3 3 7 YNL1 5 4.9 8 49
Classification	Yes	Yes	6	One gene per line, followed by tab, followed by name of classification	146 set3 158 unclassified 159 set5 170 set3 171 set3 181 set1
Tree	Yes	Yes	7	Hierarchical	<TREE DISTANCE=0.6 TITLE=a> YMR199W YPL256C <TREE DISTANCE=0.1 TITLE=b> YAL001C YAL002W </TREE> <TREE DISTANCE=0.2 TITLE=c> YAL019W YAL017W </TREE> </TREE>
Genome	Yes	Yes	8	XML description of genome. Hyperlinks and sequence are optional.	<GENOME CIRCULAR="false"> <NAME>Rat</NAME> <HYPERLINKS> %GenBank;http://www.ncbi.nlm.nih.gov/... \$PubMed;http://www.ncbi.nlm.nih.gov:80/... </HYPERLINKS> <MAPPED_FILE> GAD65 GAD2 4.1.1.15 glutamic acid decarboxylase ... pre-GAD67 GAD67 4.1.1.15 glutamic acid decarboxylase </MAPPED_FILE> <SEQUENCE> >CHR1 Chromosome I data: CCACACCACACCCACACACCCACAC... </SEQUENCE> </GENOME>

Format	Input	Output	Num	Description	Example
Temporary Genome	Yes	Yes	9	Same as above, but genome is not saved to disk. No data for it can be saved, and all records of the genome disappear when the window is closed.	Same as above

Exporting GeneSpring Data

Saving Images

You can save a GeneSpring image and import it into a graphics or other program, where you can polish and format it for publication. GeneSpring saves images of pathways, Venn diagrams, the genome browser, and the colorbar as .pct and .png files, which can be imported into Microsoft PowerPoint, Word, Publisher, Excel, CorelDRAW, and Adobe Illustrator among other programs.

Saving a Genome Browser Image

1. Display the image to save in the genome browser.
2. Select **File > Save Image** and choose **Browser**. The Export Options window appears.
3. Choose an output format for the image. The available formats are PICT or PNG. PICT is a vector based graphic format. PNG is a bitmap based format.

There are size limits to both formats. PICT files cannot be larger than 450x450 inches. PNG files have a soft limit based on available memory. If the estimated amount of memory to produce the PNG file is greater than the memory use setting in your Preferences file, a warning dialog appears. In this case, you can attempt to save the PNG anyway, but if there is not sufficient memory, nothing is saved.

PNG images are automatically saved at your current screen resolution. For a higher-resolution image, save in PICT format.

4. Choose an image size from the Page Size pull-down menu. You have the following options:
 - **Scale to Fit**—calculates the best page size in order to display the graphic and all specified labels. In some cases, this option will specify a page size larger than the maximum. In this case, you must choose another option.
 - **Original Image Size**—lets you save the image exactly as it appears in the genome browser.
 - **Original Aspect Ratio**—allows you to change the image size, but maintain the original width-to-height ratio displayed in the genome browser.
 - **US Letter**—8.5 by 11 inches.
 - **US Legal**—8.5 x 14 inches
 - **A4**—8.3 x 11.7 inches
 - **3 Foot by 5 Foot Poster**—3 ft. by 5 ft.
 - **Custom**—allows you to save to any size up to 450 inches by 450 inches.
5. Choose a margin size. If you choose **Custom**, enter the appropriate percentage in the **Enter Percentage** box.
6. Choose a page orientation - either landscape or portrait.
7. Specify whether to show labels, and if so, which labels. Your options are:
 - Show Horizontal Labels
 - Show Vertical Labels

- Use Rotated Text for Vertical Labels
- Force all Text to Show

You can also specify the text size and font for the labels.

8. Specify the color scheme to use. You can choose either your current color scheme or any of the presets in the pull-down menu.
9. Click **Save**. A Save As window appears.
10. Choose a directory, enter a file name and click **Save**.

You may need to save your file as a large custom size, such as 150x150 inches, to ensure all data are included in the saved image. Images are saved as vector graphics, which are expandable. Data that are too small to view in the genome browser are saved in most cases, and reappear when you expand the image.

Note: Images containing a very large number of genes can require an exceptional amount of memory. The fewer genes included in an image, the smaller the image file.

To Save the Colorbar or Venn Diagram

1. Display the colorbar or Venn diagram to save in the display window.
2. Select **File > Save Image** and choose Colorbar or Venn Diagram. A Save As window appears.
3. Choose a directory and file name and click **Save**.

To Save the Entire Window

Windows—Press the **Alt** and **Print Screen** keys simultaneously to copy a picture of the current active window. Paste the image into any program that accepts graphics and save it.

Macintosh—Press **⌘-Shift-4-Caps Lock** simultaneously. The cursor changes to a bull's-eye. Click on a GeneSpring window to save the image as a file on your hard drive called "Picture". You must rename this file, otherwise it is overwritten each time you repeat this procedure.

To Save the Entire Screen

Windows—Press the **Print Screen** key to save an image of your entire computer screen. Paste the image into any program that accepts graphics and save it.

Macintosh—Press **⌘-Shift-3** simultaneously to save an image of your entire computer screen. The image is saved as a file on your hard drive called "Picture".

Saving Pictures and Printing

You can print an image of the genome browser, the genome browser with the colorbar, or the display window. Such images can be useful for reports or handouts. Use a high-resolution color printer to print GeneSpring images.

Printing the Genome Browser and/or Colorbar

1. Select the **File > Print Image** command.
2. Choose from the following options:
 - Browser
 - Browser and Colorbar
 - Colorbar
3. Select a printer and click **OK**.

Printing the Display Window

Windows:

1. Hold the **Alt** and **Print Screen** keys down simultaneously. This copies a picture of the active window only.
2. Paste into any program that accepts graphics.
3. Print.

Macintosh:

1. Hold the **Command-Shift-4-Caps Lock** keys down simultaneously. The cursor changes to a bull's-eye.
2. Release the keys and use the mouse to click on the window. This creates a screenshot of your window (you will hear the sound of a snapshot). The screenshot is saved on your hard drive with the name "Picture".
3. Open the picture and print.

Exporting Gene Lists

You can make gene lists and annotated gene lists available to another application. An annotated list includes functional descriptions, as well as standard deviation, standard error and other information associated with the gene list.

Dragging Lists out of GeneSpring

You can drag a gene list out of the navigator to your desktop or directly into Microsoft Excel. If dragged to the desktop, the gene list is saved as a .zip file. If dragged into Excel, it appears in columnar format. The resulting list contains only the gene identifiers and associated values.

Dragging and dropping a gene list does not produce the same list as does the Copy Annotated Gene List function.

Copying Gene Lists

There are two methods for copying a gene list.

Method 1:

1. Select the gene list to copy from the Gene Lists folder in the navigator.
2. Select **Edit > Copy > Copy Gene List**.
3. Paste the list into another application, such as a spreadsheet program.

Method 2:

1. Open the Gene List Inspector. (Double-click a gene list or right-click and select **Inspect**.)
2. Click **Copy to Clipboard**.
3. Paste the list into a new application.

Copying Annotated Gene Lists

1. Select the gene list in the Gene List folder in the navigator.
2. Select **Edit > Copy > Copy Annotated Gene List**. A menu appears.
3. Choose an experiment interpretation from the **Copy based on interpretation** pull-down menu. (See “Experiment Interpretations” on page 3-39 for information on experiment interpretations.)
4. Choose options on the Copy Annotated Gene List window by checking or unchecking the boxes.
5. Click **Copy to Clipboard**.
6. Paste the list into another application.

Saving Annotated Gene Lists

1. Select a gene list from the Gene List folder in the navigator.

2. Select **Edit > Copy > Copy Annotated Gene List**. A menu appears.

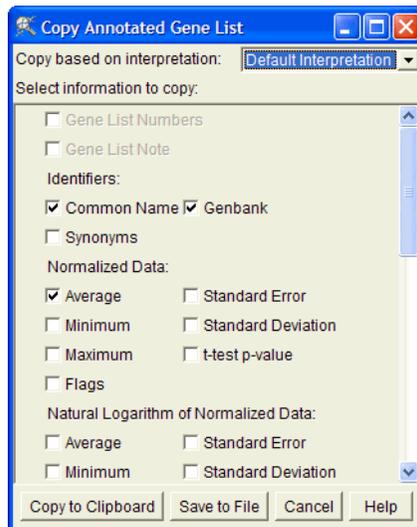


Figure 9-1 The Copy Annotated Gene List window

3. Choose the experiment interpretation from the **Copy based on interpretation** pull-down menu. (See “Experiment Interpretations” on page 3-39 for information on experiment interpretations.)
4. Click **Save to File**.
5. Choose a name and location to save your gene list.

The resulting text file can be opened in any program that accepts tab-delimited text, such as spreadsheet and word processing programs.

Annotation Options

Your options for copying and saving information with an annotated gene list are listed in the Copy Annotated Gene List window. Descriptions of these items can be found by clicking **Help**. The type and amount of information listed varies depending on your genome and the way that genome was loaded into GeneSpring.

The Systematic Name is always saved in the first column of a gene list.

General

Gene List Numbers—The values (if any) that GeneSpring has associated with this gene list. This column appears only if you have associated values. See “Adding an Associated Number Restriction” on page -10 for details on the types of numbers GeneSpring attaches to gene lists.

Gene List Note—Any notes attached to a gene list.

Identifiers

- **Common Name**—A non-systematic way of referring to a gene.
- **Synonyms**—Other names entered for your gene list.
- **GenBank**—A gene’s GenBank Accession Number, if known.

Normalized Data

- **Average**—The mean of any normalized replicates in the experiment.
- **Minimum**—The minimum normalized signal values for each gene.
- **Maximum**—The maximum normalized signal values for each gene.
- **Flags**—Flags associated with each gene
- **Standard Error**—The standard error of the normalized values for each gene.
- **Standard Deviation**—The standard deviation (the square root of the variance) of the raw data values for each gene.
- **t-test p-value**—The statistical test of differential expression for a specific condition.

Natural Logarithm of Normalized Data

- **Average**—The mean of any normalized replicates in the experiment.
- **Minimum**—The minimum normalized signal values for each gene.
- **Maximum**—The maximum normalized signal values for each gene.
- **Standard Error**—The standard error of the normalized values for each gene.
- **Standard Deviation**—The standard deviation (the square root of the variance) of the normalized values for each gene.

Raw Data

- **Average**—The mean of any raw data replicates in the experiment.
- **Minimum**—The minimum raw data signal values for each gene.
- **Maximum**—The maximum raw data signal values for each gene.
- **Standard Error**—The standard error of the raw data values for each gene.
- **Standard Deviation**—The standard deviation (the square root of the variance) of the raw data values for each gene.

Control Value

- **Average**—The mean of any control value replicates in the experiment.
- **Minimum**—The minimum control value signal values for each gene.
- **Maximum**—The maximum control value signal values for each gene.
- **Standard Error**—The standard error of the control values for each gene.
- **Standard Deviation**—The standard deviation (the square root of the variance) of the control values for each gene.

Annotations

- **Map Position**—A gene's mapping information.
- **Chromosome**—The chromosome on which a gene is located, if known.
- **User Notes**—Any additional notes you may have associated with a gene.
- **EC**—A gene's EC (Enzyme Commission) number, if known.
- **Description**—A gene's description, if known.

- **Product**—The protein product coded for by a gene, if known.
- **Phenotype**—A description of a gene's phenotype, if known.
- **Function**—A description of the function of a gene's product, if known.
- **Keywords**—Keywords associated with a gene, if known.
- **PubMed ID**—A gene's PubMed identifier.
- **Custom Field 1, Custom Field 2, Custom Field 3**—Any information you choose to place here for your own use.
- **Type**—The feature type from the GenBank file.
- **DB id**—A reference used to identify a gene within GeNet.
- **GO Biological Process**—The Gene Ontology Biological Process classification
- **GO Molecular Function**—The Gene Ontology Molecular Function classification
- **GO Cellular Component**—The Gene Ontology Cellular Component classification
- **RefSeq**—The gene's NCBI Reference Sequence project identifier.
- **UniGene**—The gene's UniGene cluster identifier.

Exporting MAGE-ML Data

MAGE-ML (Microarray Gene Expression Markup Language) is a markup language based on XML and designed to describe and communicate information about microarray experiments. Using MAGE-ML, your experimental data can include information about microarray designs, manufacturing information, and experiment setup and execution information as well as gene expression data and analysis results. Certain publications and laboratories require results to be published to Array Express or Gene Expression Omnibus in MAGE-ML format.

Array Express is a new public repository for microarray based gene expression data. Incyte provided funding for its creation, and it is now funded by EMBL and EBI. Array Express accepts only data submitted via MIAMExpress (a web-based submission interface) or via FTP in MAGE-ML format.

MAGE-ML is extremely broad in its definitions. As a result, EBI/Array Express has defined its own version of MAGE-ML. Array Express will only accept submissions that are both MIAME compliant and conform to their MAGE-ML standard.

GeneSpring currently supports export only of MAGE-ML data. It does not fully support export of data in a format that is applicable for ArrayExpress submission.

To export an experiment in MAGE-ML format:

1. Right-click an experiment in the GeneSpring navigator and select **Export as MAGE-ML**. The MAGE-ML Export window appears.

Figure 9-1 The MAGE-ML Export window

2. Enter all necessary information in the fields provided.

For EBI compliance you must include the following information:

- Choose an Experiment Design Type (also required for MIAME compliance)
- Check *only* the **Experimental Parameters**, **Sample Attributes**, and **Raw Data Files** boxes
- Fill in all of the Contact Information, including your lab's EBI accession number. If you do not specify an accession number, your experiment's accession number is used by default.

Note: If you do not have an EBI accession number, you must contact EBI to obtain one.

By default, GeneSpring's MAGE-ML export feature specifies the appropriate options to produce EBI-compliant MAGE-ML data. You have the option of including more information than EBI requires, but if you do so, your data will not be accepted by Array Express.

3. Click **OK**. The Choose Output Directory window appears.

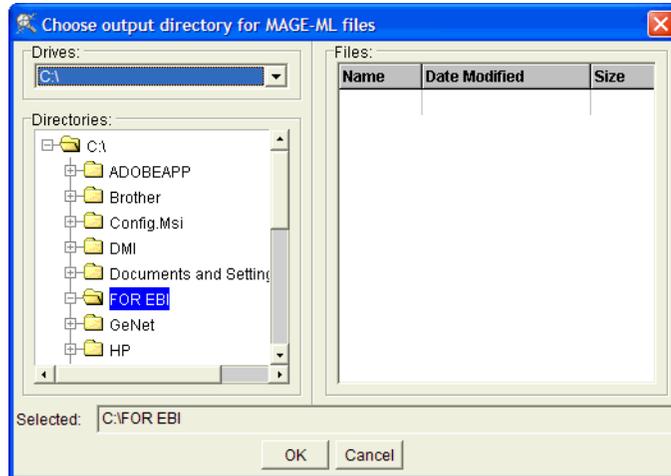


Figure 9-2 The MAGE-ML Choose Output Directory window

4. Select the directory in which the MAGE-ML export files will be saved.

An EBI-compliant MAGE-ML experiment must contain at least the following three files:

- The XML file describing the experiment
- The MAGE-ML.dtd file, which defines the MAGE-ML format
- A plain text file containing the raw experimental data

Your export may also include raw data files and/or sample or array images.

Note: Make sure the Systematic names of your genes are the same identifiers used by the chip manufacturer. This eases mapping from the vendor MAGE-ML file to the experiment MAGE-ML file.

5. Click **OK**. The MAGE-ML files are saved to the directory you specified.

Publishing Data to GeNet

GeNet is a scalable workspace for GeneSpring users that streamlines microarray research at large or multicampus organizations. It provides facilities for robust data analysis, collaborative workflow management, automated research procedures, and secure data administration. GeNet scales to meet the demands of both high-throughput sample volumes and increasing numbers of users.

You can publish any data object from GeneSpring to GeNet.

Uploading Data Objects to GeNet

1. Select the upload method. There are two methods for uploading data objects to GeNet:

Method 1: Position your cursor over a data object in the navigator you want to upload and right-click. Select **Upload to GeNet** from the pop-up menu.

Method 2: Select the desired object in the navigator and drag it into the appropriate GeNet folder (displayed in *italics*).

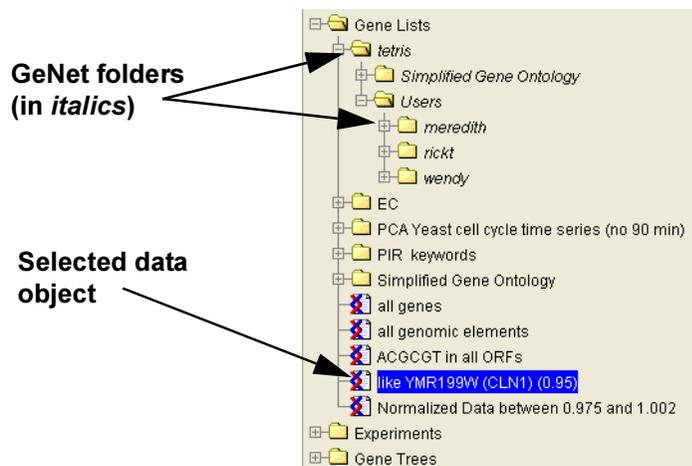


Figure 9-3 GeNet folders in the GeneSpring navigator

2. When the GeNet Upload window appears, enter a destination directory (or accept the default). To create a new destination directory, enter a name.

To browse for a directory, click **Change . . .** and select the desired directory.

3. Click **Start** to begin uploading to GeNet.

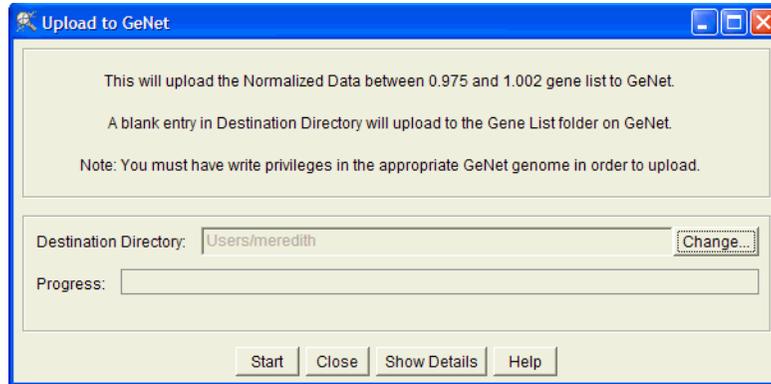


Figure 9-4 The Upload to GeNet window

Note: If you are uploading to a regulatory compliant GeNet server, you are prompted to enter your password in the electronic signature dialog. For more information, see “Electronic Signatures” on page 9-14.

The upload status box notifies you when the upload is complete.

If you are having trouble uploading, ask your administrator to confirm that you have access to the target directory on GeNet.

Deleting Data Objects from GeNet

You can delete data objects of which you are the owner in GeNet using the GeneSpring interface. Right-click the object in the GeneSpring navigator and select **Delete**. A confirmation dialog appears.

Click **Yes** to continue.

Note: If you are deleting objects from a regulatory compliant GeNet server, you are prompted to enter your password in the electronic signature dialog. For more information, see “Electronic Signatures” on page 9-14.

Uploading Genomes to GeNet

The Bulk Upload feature allows you to upload entire genomes and large amounts of data to GeNet at once. To perform a bulk upload, select **File > Bulk Upload to GeNet**.

Note: If you are uploading to a regulatory compliant GeNet server, you are prompted to enter your password in the electronic signature dialog. For more information, see “Electronic Signatures” on page 9-14.

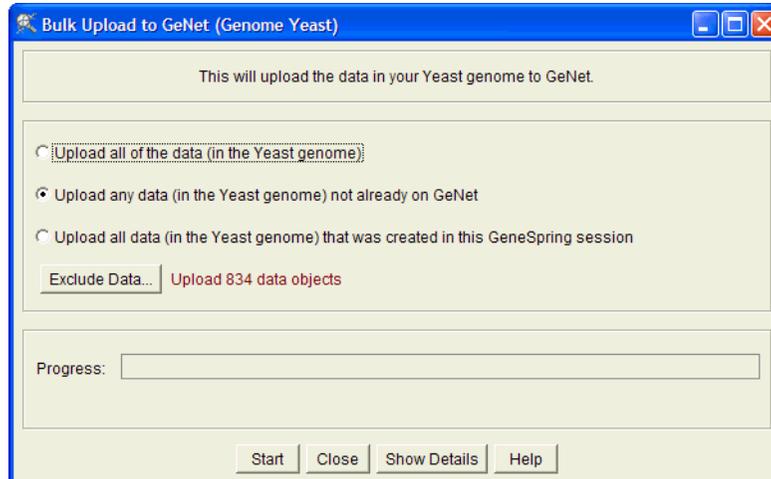


Figure 9-5 The Bulk Upload window

Users who do not have administrator access to GeNet have the following options:

- **Upload all of the data**—Upload all data in the current genome to GeNet. This option creates duplicates on GeNet of any data already uploaded.
- **Upload any data not already on GeNet**—Upload all data in the current genome that is not already present on GeNet.
- **Upload all data that was created in this GeneSpring session**—Upload all data in the current genome that was created during the current GeneSpring session.

Users who are logged into GeNet as an administrator from GeneSpring have the following additional options:

- **Upload no data objects**—Do not upload any data. This option appears only if the active genome already exists on GeNet.
- **Upload the genome**—If the active genome does not already exist on GeNet, upload the genome and all associated data. This option appears only if the active genome does not already exist on GeNet.

In addition, administrators have the option to upload data directly into the root directory, or into their own default directory. In the Destination Directory section of the screen, select “root” to upload the experiments in your GeneSpring Experiments folder to the GeNet Experiments folder. Select “default” to upload the experiments to your GeNet user directory.

The Exclude Data Window

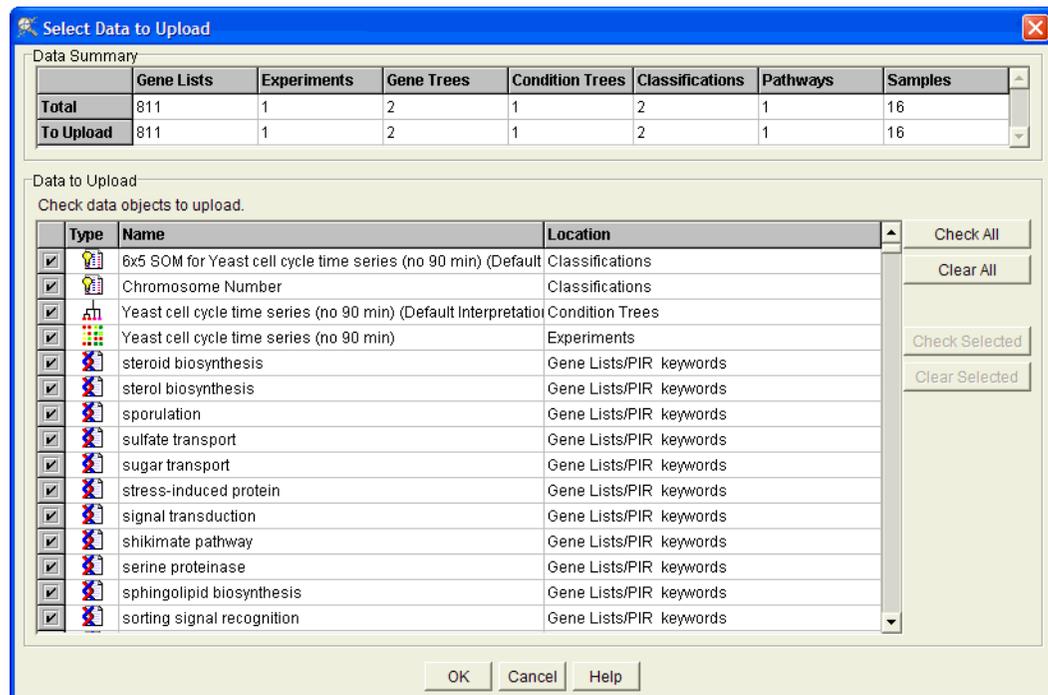


Figure 9-6 The Exclude Data window

Click the Exclude Data button to view a summary of all the data to be uploaded. From this screen, you can select data objects to be excluded from the bulk upload. Only checked items are uploaded. To exclude a file or data object, uncheck the box next to it in the table.

Electronic Signatures

If you are interacting with a regulatory compliant GeNet server, you must provide an electronic signature each time you upload a data object to GeNet or delete an object from a GeNet folder.

When you perform an action requiring a signature, the Electronic Signature dialog appears.

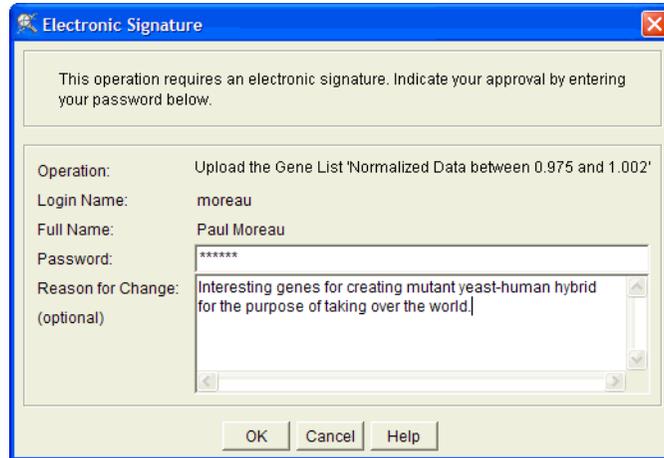


Figure 9-7 The Electronic Signature dialog

Your GeNet password serves as your electronic signature. Enter your password in the field provided. Optionally you may enter a brief text message describing the purpose of your action.

For additional information about electronic signatures and other regulatory compliance features in GeNet, see the GeNet Administration Manual or the GeNet User Manual.

Appendix A

Installing from a Database

Custom Databases and GeneSpring

You can load experiments into GeneSpring from your company's database. To do this you must set up a `dataloader.xml` file prior to starting GeneSpring.

Databases

A database is an organized collection of information. Essentially, it is a collection of records. In database terms, a record consists of all the useful information you can gather about a particular item. Each little bit of information making up a record is called a field. An example of a non-computerized database would be your address book. Each record represents one of your contacts, and each record consists of many fields such as name, address, number, and so on.

Computer databases automatically keep records organized and enable you to search for or pull out particular records based on any field in the record. The software allowing you to create and maintain databases is called a *Database Management System*, or DBMS. In database terminology, a *file* is called a *table*. Each *record* in the file is called a *row*, and each *field* is called a *column*.

A relational database is the most common type of database in client/server systems. Simply stated, in this type of database, relationships are established between tables based on common information.

Open Database Connectivity

Open Database Connectivity (ODBC) is an Application Programming Interface (API) allowing a programmer to abstract a program from a database. When writing code to interact with a database, you usually must add code that talks to a particular database using a proprietary language. If you want your program to talk to Access, Fox and Oracle databases, you must code your program with three different database languages. This can be a very difficult or time consuming task.

This is where ODBC enters the picture. When programming to interact with ODBC you only need to speak the ODBC language (a combination of ODBC API function calls and the SQL language). The ODBC Manager determines how to contend with the type of database you are targeting. Regardless of the database type you are using, all of your calls will be to the ODBC API. All you must do is install an ODBC driver specific to the type of database you are using.

Structured Query Language

Structured Query Language (SQL) is a standard language for defining and accessing relational databases. All of the major database servers used in client/server applications work with SQL. It is a query language designed to extract, organize and update information in relational databases. Each database vendor has its own particular dialect. These dialects are similar to one another, but different enough that programmers must pay close attention to which RDBMS is being used. The most important dialects of SQL are ANSI/ISO SQL, IBM DB2, SQL Server, Oracle, Ingres, and ODBC.

SQL uses statements to get work done. Examples of some of these statements are:

- SELECT
- INSERT
- DELETE
- UPDATE
- DECLARE
- OPEN
- CLOSE
- CREATE
- PREPARE
- DESCRIBE

SQL Call Level Interfaces

When a Call Level Interfaces (CLI) is used, a program requests database services by calling special SQL interface routines rather than embedding SQL statements directly into the program. There are two distinct types of CLIs. First, each DBMS vendor provides its own unique API for its database. The vendor-specific API is usually the most efficient way to access the database, but each vendor's API is unique. As a result, if you decide to write programs that use a vendor API, you lock yourself into using that vendor's DBMS. However, your programs are efficient as possible.

The second type of CLI is a standard or open API which is supported by more than one database vendor. Several open database APIs are available, one of which is ODBC. ODBC is a standard CLI for accessing SQL databases from Windows.

The Genetic Analysis Technology Consortium

The Genetic Analysis Technology Consortium (GATC) was formed in an attempt to standardize the rapidly growing field of array-based genetic analysis. The consortium was created to provide a unified technology platform to design, process, read and analyze DNA-arrays.

The goal of the GATC is to make micro-arrays broadly available and provide a technology platform that allows investigators to use components from multiple vendors.

Databases and GeneSpring

Experimental data are not always stored on the researcher's desktop in simple text files. Sometimes the data are stored on a relational database. GeneSpring can save and load all types of data to an SQL database through ODBC.

Experimental data can be loaded from a database simply by telling GeneSpring which table(s) contain the data and which columns contain the experimental index. You load in the data using the Experiment Wizard almost exactly as you would if they were text files (see "Entering your Database into GeneSpring" on page A-23). The only difference is you enter experiment identifiers instead of file names, and SQL table columns instead of tab-delimited column headers.

Parameters describe what the database knows about each sample. Different databases have different ways of storing parameters, so they must be retrieved by explicit SQL state-

Custom Databases and GeneSpring

ments. Silicon Genetics can provide these for GATC and help write these for individual databases. This needs to be done only once. Afterwards, the customer simply chooses the database and GeneSpring retrieves data from it. Normalization and other options can also be set for a database.

Adding an Experiment from a Database

This section describes the process of making your database visible to GeneSpring. The database administrator should have done this already. If they have, you can skip down to “Connecting your Database to GeneSpring” on page A-6.

Creating a New ODBC Source

1. Select **Start > Control Panel**.
2. Open **Administrative Tools**.
3. Open **Data Sources (ODBC)**. A new window, **The ODBC Data Source Administrator**, appears.
4. Go to the **system DSN**
5. Click **Add** for a new **Create New Data Source** window.
6. Select the correct type of database from the scrolling list. A new panel appears.
7. Give the experiment a name. Remember that experiment names in GeneSpring are case sensitive.
8. Click **Select** to browse for the correct database. Usually you will connect to a new computer (server) to access the database.
9. The new entry appears in the list of databases.

Testing Your ODBC Connection

1. Open **Excel**.
2. Select **Data > Get External Data**.
3. Select **New Database Query**. Look for your database in the presented list.

Connecting your Database to GeneSpring

The following section describes the use of the database configuration file. You must customize this file for your system before running GeneSpring. You can have any number of different configuration files

The purpose of this file is to tell GeneSpring how to read the database as if it were a simple text file. It pulls the data together and places it in columns recognized by GeneSpring. Column names and sample name references are entered in the Experiment Wizard as normal.

1. Using your file management software, create a new folder in the data directory of GeneSpring called `Databases`.
2. Create a file named `dataloader.xml`. Instructions for setting up this file are provided in the section below, *Configuration File Reference*.
3. Issue a SQL command to retrieve the parameters in all samples. Use `MicrosoftQuery` in Excel to generate SQL commands.
 - a. In Excel open the **Tools** menu.
 - b. Select **Get External Data**.
 - c. Select **New Database Query**.
 - d. Make sure you specify what to edit in `MicrosoftQuery`.

Configuration File Reference

The following section contains a list of the tags used in the database configuration file (`dataloader.xml`). The configuration file is in XML format and uses tags enclosed in angle brackets much like an HTML document.

In such a document, an *element* consists of a tag enclosed in angle brackets, and usually includes a closing tag. For example, the top-level element of this configuration file is the `ExternalDatabaseConfiguration` element. This element consists of opening and closing tags, i.e.:

```
<ExternalDatabaseConfiguration>
...
</ExternalDatabaseConfiguration>
```

An element's *contents* are tags or text nested between the current element's opening and closing tags. The following are examples of elements with contents:

```
<PhysicalDatabase>
  <UserName>BioMan</UserName>
</PhysicalDatabase>
```

In the above example, the `<UserName>` element (including its own contents and its closing tag) is the contents of the `<PhysicalDatabase>` element. The username `BioMan` is the contents of the `<UserName>` element.

Attributes are values defined within the opening tag of the element itself. In the following example, name is an attribute, and "dbname" is the value of the <PhysicalDatabase> element:

```
<PhysicalDatabase name="dbname">
```

An element may have any number of attributes or contents. An empty element (an element that has attributes but no contents) can be closed within the opening tag by adding a slash at the end, i.e., <tag value="example" />.

Tag Reference Table

This table provides a list of all available tags for the database configuration file. For more detailed information on the use of each tag, see "Tag Definitions" on page A-9.

The Element column lists each of the available tags. The Contents column lists the type of contents that tag can contain (i.e., plain text, or the names of the tags it can contain). The Attributes column lists the attributes that tag can contain. The Allowed In column lists the tags between which the current element is allowed.

Element	Contents	Attributes
<ExternalDatabaseConfiguration>	<GeneralConfiguration> <Database>	n/a
<GeneralConfiguration>	<LoadClass> <ProcessedDataListFile>	n/a
<Database>	<PhysicalDatabase> <TechnologyType> <Header> <GenomeNames> <GetSampleIDs> <GetSampleAttributes> <GetFile> <GetRawData>	name, icon
<LoadClass>	plain text	n/a
<ProcessedDataListFile>	plain text	n/a
<PhysicalDatabase>	<UserName> <Password> <URL> <Prefetch>	name
<TechnologyType>	n/a	name
<Header>	<Author> <Research_Group> <Organization>	n/a
<GenomeNames>	<GenomeMappingSpec>	n/a
<UserName>	plain text	n/a
<Password>	plain text	n/a
<URL>	plain text	n/a

Connecting your Database to GeneSpring

Element	Contents	Attributes
<Prefetch>	plain text	n/a
<Author>	plain text	n/a
<Research_Group>	plain text	n/a
<Organization>	plain text	n/a
<GetSampleIDs>	<DatabaseQuery> <DataDirectory> <FileNameMask> <IDFromFileName> <JavaQuery>	location
<GetSampleAttributes>	<DatabaseQuery> <JavaQuery>	cacheable numeric
<GetFile>	<DatabaseQuery> <JavaQuery>	type location deleteAfterwards mimeType
<GetRawData>	<DatabaseQuery> <Format>	n/a
<GenomeMappingSpec>	n/a	targetName sourceName baseDirectory
<DatabaseQuery>	SQL command	useGenome- Name db
<DataDirectory>	plain text	n/a
<FileNameMask>	plain text	n/a
<IDFromFileName>	<RegexMatch> <DatabaseQuery>	n/a
<RegexMatch>	plain text	n/a>
<JavaQuery>	n/a	class extraArgs

Element	Contents	Attributes
<Format>	<GeneColumn> <Headlines> <SignalColumn> <NormalizedColumn> <ReferenceColumn> <SignalBackgroundColumn> <ReferenceBackgroundColumn> <ExperimentWorkedColumn> <ExperimentWorkedDesignation> <ExperimentAbsentDesignation> <ExperimentMarginalDesignation> <RegionColumn> <TreatNoSignalAsInvalid> <LowerBoundOnSignalColumn> <UpperBoundOnSignalColumn> <StandardDeviationSignalColumn> <ColumnHeaderLine>	type
<GeneColumn>	plain text	n/a
<Headlines>	plain text	n/a
<SignalColumn>	plain text	n/a
<NormalizedColumn>	plain text	n/a
<ReferenceColumn>	plain text	n/a
<SignalBackgroundColumn>	plain text	n/a
<ReferenceBackgroundColumn>	plain text	n/a
<ExperimentWorkedColumn>	plain text	n/a
<ExperimentWorkedDesignation>	plain text	n/a
<ExperimentAbsentDesignation>	plain text	n/a
<ExperimentMarginalDesignation>	plain text	n/a
<RegionColumn>	plain text	n/a
<TreatNoSignalAsInvalid>	plain text	n/a
<LowerBoundOnSignalColumn>	plain text	n/a
<UpperBoundOnSignalColumn>	plain text	n/a
<StandardDeviationSignalColumn>	plain text	n/a
<ColumnHeaderLine>	plain text	n/a

Tag Definitions

<ExternalDatabaseConfiguration>

The top-level element defining the entire database configuration. This element contains all of the other tags.

Connecting your Database to GeneSpring

Contents: <GeneralConfiguration>, <Database>

Attributes: n/a

Usage:

```
<ExternalDatabaseConfiguration>
...
</ExternalDatabaseConfiguration>
```

Notes: Required, can appear only once in the configuration file

<GeneralConfiguration>

The element containing all of the general configuration options for the database.

Contents: <LoadClass>, <ProcessedDataListFile>

Attributes: n/a

Usage: <GeneralConfiguration>...</GeneralConfiguration>

Notes: Required, can appear only once in the configuration file

<Database>

The element containing the specifics of the source or sources of sample data, whether it is in a database or a directory of flat files. You must have one Database section for each source to which you will connect. The icon attribute is optional, and allows you to specify a graphical image to represent the database.

Contents: <PhysicalDatabase>, <TechnologyType>, <Header>, <Genome-Names>, <GetSampleIDs>, <GetSampleAttributes>, <GetFile>, <GetRawData>

Attributes: name (required), icon

Usage:

```
<Database name="Affymetrix Database" icon="/usr/local/graphics/icon.gif">
...
</Database>
```

Notes: Required, can appear multiple times

<LoadClass>

Loads the driver that connects to the database. In some cases you may want to use a JDBC driver written in Java which must be instantiated at startup. You can specify any number of these drivers. Any class you specify, however, must be in your CLASSPATH. This element is optional. If you are using a default driver, this is not necessary, but if you are using a specific driver, you must specify it here.

Contents: plain text

Attributes: n/a

Usage: <LoadClass>sun.jdbc.odbc.JdbcOdbcDriver</LoadClass>

Notes: Optional, frequently used

<ProcessedDataListFile>

This setting specifies where the database will save the list of samples that have been uploaded.

Contents: plain text

Attributes: n/a

Usage:

```
<ProcessedDataListFile>/usr/database/ProcessedList.txt</ProcessedListData-File>
```

Notes: Required

<PhysicalDatabase>

You must have one Physical Database tag for each physical SQL database to which you will connect. The “name” element specifies the database name for any <Database-Query> tags that occur within the <Database> element. If you are loading data from flat files, you may not use the <PhysicalDatabase> element.

Contents: <UserName>, <Password>, <URL>, <Prefetch>

Attributes: name

Usage: <PhysicalDatabase name=”dbname”>...</PhysicalDatabase>

Notes: Required to retrieve files from a SQL database

<UserName>

This element specifies the username for logging into the sample database.

Contents: plain text

Attributes: n/a

Usage: <UserName>Ndege MacKenzie</UserName>

Notes: Required for <PhysicalDatabase>

<Password>

This element specifies the password for logging into the sample database.

Contents: plain text

Attributes: n/a

Usage: <Password>dbPassword</Password>

Notes: Required for <PhysicalDatabase>

<URL>

Specifies the physical address of the database or directory from which you will retrieve sample data.

Contents: plain text

Attributes: n/a

Usage: <URL>jdbc:odbc:database</URL>

Notes: Required for <PhysicalDatabase>

<Prefetch>

This is an optional, rarely-used element that allows you to specify how many rows to retrieve from the database during the prefetch process. This may be useful in certain cases where there are performance issues. This element is contained by <PhysicalDatabase>.

Contents: plain text

Attributes: n/a

Usage: <Prefetch>20</Prefetch>

Notes: Optional, rarely used

<TechnologyType>

Sample special field technology to be set for each sample uploaded. This identifies the chip or technology used for the sample.

Contents: n/a

Attributes: name

Usage: <TechnologyType name="Affymetrix"/>

Notes: Optional

<Header>

This element specifies header fields to be set for each sample uploaded from the current database.

Contents: <Author>, <Research_Group>, <Organization>

Attributes: n/a

Usage: <Header>...</Header>

Notes: Required for <PhysicalDatabase>

<Author>

Specifies the sample author to be designated in the header field for each sample being uploaded.

Contents: plain text

Attributes: n/a

Usage: <Author>Juanita Nguyen</Author>

Notes: Required for <Header>

<ResearchGroup>

Specifies the research group to be designated in the header field for each sample being uploaded.

Contents: plain text

Attributes: n/a

Usage: <ResearchGroup>Discovery Central</Research Group>

Notes: Required for <Header>

<Organization>

Specifies the organization to be designated in the header field for each sample being uploaded.

Contents: plain text

Attributes: n/a

Usage: <Organization>Cures R Us</Organization>

Notes: Required for <Header>

<GenomeNames>

This setting allows you to associate samples with genomes. One database may have several genomes. Within this element there must be at least one <GenomeMappingSpec> tag.

Contents: <GenomeMappingSpec>

Attributes: n/a

Usage: <GenomeNames>...</GenomeNames>

Notes: Required, can appear only once in a configuration file

<GenomeMappingSpec>

This element specifies the name of the genome in your sample source, the target genome on GeneSpring to upload it into, and the base directory to use for <GetSampleIDs> or <GetFile> elements. This tag must appear at least once. However, if the “useGenomeName” attribute is set to false in the <DatabaseQuery> tag for <GetSampleIDs>, this tag must appear *only* once. The attribute values for this tag are as follows:

- **targetName**—The genome on GeneSpring into which the samples will be uploaded
- **sourceName**—The name of the database or directory from which to retrieve samples
- **baseDirectory**—The directory to use for <GetSampleIDs> or <GetFile> elements if no data directory is specified

Contents: n/a

Attributes: targetName, sourceName, baseDirectory

Usage:

```
<GenomeMappingSpec targetName="Yeast" sourceName="YeastDB"
baseDirectory="/" />
```

Please note in the above example that the `baseDirectory` attribute value is `"/"`. The second slash (`/`) takes the place of the `</GenomeMappingSpec>` tag.

Notes: Required

<GetSampleIDs>

This element specifies the location from which to upload samples. There are three accepted values for the “location” attribute:

- `database`—perform a database search
- `directory`—locate files in a directory
- `java`—upload files based on the result of a Java call

These values are case-insensitive.

Contents: `<DatabaseQuery>`, `<DataDirectory>`, `<FileNameMask>`, `<IDFromFileName>`, `<JavaQuery>`

Attributes: `location`

Usage: `<GetSampleIDs location="database">...</GetSampleIDs>`

Notes: Required for `<PhysicalDatabase>`

<GetSampleAttributes>

Specifies parameters for retrieving attributes associated with samples. These can include the name, value, or units of the sample attribute, and possibly a flag specifying whether the attribute is numeric. The “cacheable” attribute defines whether to cache sample attribute values for previously-uploaded samples. This can greatly improve performance for uploads from external databases. This will not affect automatic upload performance, since in this case sample attributes are already retrieved only once. Acceptable values are “true” and “false”.

The “numeric” attribute indicates whether the retrieved values should be considered numeric. Acceptable values are “yes”, “no”, or “guess”. This attribute is used only for database queries. If the Java query option is used, this setting is overridden.

Contents: `<DatabaseQuery>`, `<JavaQuery>`

Attributes: `cacheable`, `numeric`

Usage: `<GetSampleAttributes cacheable="true" numeric="guess">...</GetSampleAttributes>`

Notes: Optional

<GetFile>

Specifies parameters for retrieving associated files. This element has four attributes.

The “type” attribute specifies the type of file to be retrieved. This attribute is required, and is case-insensitive. There are five possible values:

- Sample Image—a picture or pictures of the biological sample
- Array Image—a picture or pictures of the scanned array(s)
- CEL File—an Affymetrix CEL file (actually stored as a general attachment with MIME type application/x-AffyCELFile)
- Raw Data File—a raw data file or files
- attachment—a general attachment

The “location” attribute specifies the location of the files to be retrieved. This attribute is required, and is case-insensitive. There are four possible values:

- database—returns the contents of the file (typically a blob)
- file—returns a file pathname
- URL—returns a URL
- java—returns `com.sigenetics.ext.database.getFile`

The “deleteAfterwards” attribute specifies whether to delete the file once it has been imported. Accepted values are “true” and “false”. This attribute applies only if the “location” attribute is set to “file”. This attribute is optional. If not specified, its value defaults to “false”.

The “mimeType” attribute specifies the MIME type of the file or files being retrieved. Any valid MIME type is an acceptable value. This attribute is optional.

Contents: `<DatabaseQuery>`, `<JavaQuery>`

Attributes: `type`, `location`, `deleteAfterwards`, `mimeType`

Usage: `<GetFile type="Sample Image" location="database" deleteAfterwards="true" mimeType="image/gif">...</GetFile>`

Notes: Optional

<GetRawData>

This option specifies how to retrieve the actual sample data. This may come from either a database or a raw file (or files). The raw file itself may have been a file downloaded from a database or extracted from a Java class.

If the data is located in a database, use `<DatabaseQuery>` to retrieve it. If it is in a file or directory of files, it is interpreted as a tab-delimited file, and you must specify the file format using the `<Format>` tag.

Contents: `<DatabaseQuery>`, `<Format>`

Attributes: n/a

Usage: `<GetRawData>...</GetRawData>`

Notes: Required

<DatabaseQuery>

This element allows you to enter a SQL query that produces a list of sample identifiers, attributes, or other data based on the provided genome name. If “useGenomeName” is

Connecting your Database to GeneSpring

true, the SQL query is passed the sourceName specified in the current <GenomeMappingSpec> tag. Use the “db” attribute to specify the database to query.

Accepted values for “useGenomeName” are “true” and “false”.

The data retrieved by this option varies depending on which tag contains it, as follows:

- <GetSampleIDs>—a list of sample identifiers
- <GetSampleAttributes>—three columns (or a multiple of three columns, in which case each set of three is considered independently). These three columns are:
 - sample attribute value
 - sample attribute name
 - sample attribute units

Each row represents one attribute. If there is more than one set of three columns, for each set of three, each row represents an attribute.

- <GetFile>—if the “location” attribute is set to “database”, returns two or three columns:
 - the data
 - the filename
 - the mime type (if present, this overrides the mimeType specified in <GetFile>)

Each row in the result represents a file to be loaded.

Contents: SQL command

Attributes: useGenomeName, db

Usage: <DatabaseQuery useGenomeName=”true” db=”dbname”>select ID from Experiments where Experiments.chipType=?</DatabaseQuery>

Notes: Required if the location attribute value in <GetSampleIDs> is “database”.

<DataDirectory>

If samples are contained in flat files rather than a database, this setting specifies the directory in which sample files are located. If a directory is not specified or does not begin with an absolute path, the baseDirectory attribute of the <GenomeMappingSpec> tag is used.

Contents: plain text

Attributes: n/a

Usage: <DataDirectory>/usr/share/affy</DataDirectory>

Notes: Optional

<FileNameMask>

FileNameMask is applied to all files in DataDirectory to filter the FileNames. If the DataDirectory is not specified or does not begin with an absolute path, the baseDirectory attribute of the current<GenomeMappingSpec> section is used. If baseDirectory does not begin with an absolute path, the current user directory is used.

Contents: plain text

Attributes: n/a

Usage: `<FileNameMask>*/AffyChipID*.chip</FileNameMask>`

Notes: Required for retrieving data from flat files in a directory

<IDFromFileName>

This allows you to generate sample IDs directly from file names. If sample IDs are generated using `<Regexpmatch>`, only one genome should be specified in the `<GenomeNames>/<GenomeMappingSpec>` tags. The result of the `<RegexpMatch>` on the file names provides the sample IDs. If you are using `<DatabaseQuery>` instead, the file names are passed as arguments to the specified SQL query.

Contents: plain text

Attributes: n/a

Usage: `<IDFromFileName>...</IDFromFileName>`

Notes: Optional

<RegexpMatch>

When using `<IDFromFileName>`, use either this tag *or* `<DatabaseQuery>`, but not both.

Contents: plain text

Attributes: n/a

Usage: `<RegexpMatch>AffyChipID(.*)\.chip</RegexpMatch>`

Notes: Optional

<JavaQuery>

Allows you to use a Java class to return an array of identifiers. You specify a command such as:

```
<JavaQuery class="com.pharma.database.getParameters"
extraArgs="Blah"/>
```

and it creates an instance of `com.pharma.database.getParameters` using the default constructor. That class should implement `com.sigenetics.ext.database.GetAttributes`. Then for each attribute, a function is called with arguments of the database identifier, the database genome name, and the extra argument. The return value is an array of `com.sigenetics.ext.database.Attribute` objects, each with `name`, `value`, `units` and `isNumeric` fields.

Contents: n/a

Attributes: class, extraArgs (optional)

Usage: `<JavaQuery class="com.pharma.database.getIDs" extraArgs=""/>`

Notes: Optional, applies only to `<GetSampleIDs>`, `<GetSampleAttributes>`, and `<GetFile>`.

<Format>

Specifies the format of raw data to be retrieved. If this data is in a known format, you can specify it using the “type” attribute. Currently supported types are:

- Incyte Internet Download
- Incyte
- Affymetrix
- Affymetrix Pivot Table
- AtlasImage
- GenePix Results
- Imagene
- ScanArray
- QuantArray

Important: Affymetrix Pivot files with more than one sample per file are not supported.

If you are retrieving data from a directory containing data in multiple formats, any files that do not match the format you specify here will be ignored.

If data are being retrieved from a database as a set of columns using a SQL query, a known format type may not be used and must be explicitly defined using the format described below.

If your data are not in a standard format, you must define the format using the available tags. Columns can be specified either as a number (first column=1) or header. If columns are specified by the header, and data is retrieved from a database using a SQL query, make sure the headers retrieved in the SQL query exactly match the headers specified here. It is a good idea to write your SQL queries as follows:

```
select column1 as "column1" from table_name where ...
```

If a column is not used, you can omit the line or enter -1 (" " for strings).

Contents: <GeneColumn>, <Headlines>, <SignalColumn>, <NormalizedColumn>, <ReferenceColumn>, <SignalBackgroundColumn>, <ReferenceBackgroundColumn>, <ExperimentWorkedColumn>, <ExperimentWorkedDesignation>, <ExperimentAbsentDesignation>, <ExperimentMarginalDesignation>, <RegionColumn>, <TreatNoSignalAsInvalid>, <LowerBoundOnSignalColumn>, <UpperBoundOnSignalColumn>, <StandardDeviationSignalColumn>, <ColumnHeaderLine>

Attributes: type

Usage: <Format type="Affymetrix"/> **or** <Format>...</Format>

Notes: Required

<GeneColumn>

Specifies which column in the sample data contains the gene identifier. This tag is used only if your data are in a nonstandard format.

Contents: plain text

Attributes: n/a

Usage: <GeneColumn>1</GeneColumn>

Notes: Required if data type is not specified in the <Format> tag.

<Headlines>

Number of header lines to skip at the top before further processing. This can usually be determined automatically if the columns are specified by header. This tag does not apply when sample data are retrieved from a database.

Contents: plain text

Attributes: n/a

Usage: <Headlines>0</Headlines>

Notes: Optional

<SignalColumn>

Specifies the column containing the raw signal data.

Contents: plain text

Attributes: n/a

Usage: <SignalColumn>31</SignalColumn>

Notes: Required if data type is not specified in the <Format> tag.

<NormalizedColumn>

Specifies the column containing normalized data.

Contents: plain text

Attributes: n/a

Usage: <NormalizedColumn>30</NormalizedColumn>

Notes: Optional, rarely used

<ReferenceColumn>

Specifies the column containing raw reference data. This is typically present in two-color experiments.

Contents: plain text

Attributes: n/a

Usage: <ReferenceColumn>32</ReferenceColumn>

Notes: Optional

<SignalBackgroundColumn>

Specifies the column containing the background signal to be subtracted from the main signal before further processing.

Contents: plain text

Attributes: n/a

Usage: `<SignalBackgroundColumn>-1</SignalBackgroundColumn>`

Notes: Optional, rarely used

<ReferenceBackgroundColumn>

Specifies the column containing the background signal to be subtracted from the reference signal before further processing.

Contents: plain text

Attributes: n/a

Usage: `<ReferenceBackgroundColumn>-1</ReferenceBackgroundColumn>`

Notes: Optional, rarely used

<ExperimentWorkedColumn>

Specifies the column containing a flag or flags indicating success of the measurement.

Contents: plain text

Attributes: n/a

Usage: `<ExperimentWorkedColumn>33</ExperimentWorkedColumn>`

Notes: Optional, see also `<ExperimentWorkedDesignation>`, `<ExperimentAbsentDesignation>`, `<ExperimentMarginalDesignation>`

<ExperimentWorkedDesignation>

Specifies the flag in the column specified by `<ExperimentWorkedColumn>` that indicates that the measurement worked well.

Contents: plain text

Attributes: n/a

Usage: `<ExperimentWorkedDesignation>P</ExperimentWorkedDesignation>`

Notes: Optional, see `<ExperimentWorkedColumn>`

<ExperimentAbsentDesignation>

Specifies the flag in the column specified by `<ExperimentWorkedColumn>` that indicates that the measurement did not work well.

Contents: plain text

Attributes: n/a

Usage: `<ExperimentAbsentDesignation>A</ExperimentAbsentDesignation>`

Notes: Optional, see `<ExperimentWorkedColumn>`

<ExperimentMarginalDesignation>

Specifies the flag in the column specified by <ExperimentWorkedColumn> that indicates that the measurement worked only marginally.

Contents: plain text

Attributes: n/a

Usage: <ExperimentMarginalDesignation>M</ExperimentMarginalDesignation>

Notes: Optional, see <ExperimentWorkedColumn>

<RegionColumn>

Specifies the column that indicates regions to be normalized separately.

Contents: plain text

Attributes: n/a

Usage: <RegionColumn>-l</RegionColumn>

Notes: Optional, rarely used

<TreatNoSignalAsInvalid>

Specifies whether a signal of “0” should be treated as blank. If no value is specified for this tag, it defaults to “no”. Accepted values are “no” and “yes”.

Contents: plain text

Attributes: n/a

Usage: <TreatNoSignalAsInvalid>no</TreatNoSignalAsInvalid>

Notes: Optional, rarely used

<LowerBoundOnSignalColumn>

When an error model is known, specifies a lower bound on the signal value.

Contents: plain text

Attributes: n/a

Usage: <LowerBoundOnSignalColumn>-l</LowerBoundOnSignalColumn>

Notes: Optional, rarely used

<UpperBoundOnSignalColumn>

When an error model is known, an upper bound on the signal value.

Contents: plain text

Attributes: n/a

Usage: <UpperBoundOnSignalColumn>-l</UpperBoundOnSignalColumn>

Notes: Optional, rarely used

<StandardDeviationSignalColumn>

When an error model is known, specifies the standard deviation of the signal value.

Contents: plain text

Attributes: n/a

Usage: <StandardDeviationSignalColumn>-1</StandardDeviationSignalColumn>

Notes: Optional, rarely used

<ColumnHeaderLine>

Specifies the row containing the header names. Usually this can be determined automatically.

Contents: plain text

Attributes: n/a

Usage: <ColumnHeaderLine>-1</ColumnHeaderLine>

Notes: Optional, rarely used

Entering your Database into GeneSpring

Prepared Databases

In the main GeneSpring window, select `File > Get Data from Database`.

The majority of the remaining Experiment Wizard panels are filled in automatically.

If you left the debug setting ‘true’ an additional window appears. When the query boxes appear these will contain actual SQL commands.

GeneSpring must re-query the database each time you restart the program. If this takes too long, you can right-click the appropriate database icon and select the ‘save to disk’ option.

All commands in the *.experiment* files can also be added to the *.database* file.

More Complicated Databases

You can link various tables together in SQL. This typically requires a proficient user of databases, check with the person who built your database if you have questions.

There are many ways to enter and organize data within databases. If the data organization in your database is confusing, you might want to make separated tables for your data or part of your data. For example you could make a separate table just for parameters, like the table below:

Sample 1	Parameter Name	Parameter Value
1	elephants	2
2	elephants	34
2	daises	30

In this table there are no parameters in the individual columns. All parameters tables should have an associated sample number.

If you use a GATC database, you must re-link all the sample numbers to the parameter numbers. In that case you must define an SQL. In that case, you must define a SQL line to get those parameters, for example:

```
SQLgetParameters : select
```

This should retrieve values of and names of the parameter.

Appendix B

Equations for Correlations and other Similarity Measures

Measures of Similarity

Many advanced analysis techniques are based on measures of gene similarity. Similarity or “nearness” between genes is usually based on the correlation between the expression profiles of the two genes.

GeneSpring offers nine choices of similarity measures. Each can be selected from a pull-down list appearing in the Clustering and Filtering windows. See , “Clustering and Characterizing Data” and “The Filtering Menu” on page -51 respectively.

Each measure takes two expression patterns and produces a number representing how similar the two genes are. Most of the measures of similarity are correlation measures, and their value varies from -1 (exactly opposite) to 1 (the same). For a measure of distance, the result varies from 0 (the same) to infinity (different). For confidences, the result varies from 0 (no confidence) to 1 (perfect confidence). Both distance and confidence are actually measures of dissimilarity (small means close and large means far away). These are each transformed to measures of similarity by GeneSpring in ways detailed below.

If one expression value for a particular sample for either gene is missing, that sample is not considered in the calculation.

The notation used to describe the formulas:

- *Result* : the result of the calculation for genes **A** and **B**.
- *n* : the number of samples being correlated over.
- **a** : the vector $(a_1, a_2, a_3 \dots a_n)$ of expression values for gene **A**.
- **b** : the vector $(b_1, b_2, b_3 \dots b_n)$ of expression values for gene **B**.

Normal mathematical notation for vectors are used. In particular:

- $\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$
- $|\mathbf{a}| = \text{square root}(\mathbf{a} \cdot \mathbf{a})$

Common Correlations

Standard Correlation

Standard correlation measures the angular separation of expression vectors for Genes **A** and **B** around zero. As almost all normalized values for genes are positive, you find mostly positive correlations between genes when you use the Standard correlation. This metric is designed to answer the question “do the peaks match up?” or to put it another way, “are the two genes expressed in the same samples?” Since these questions are the most frequent questions a biologist is trying to get answered, GeneSpring calls it “Standard correlation”. It is important to note, what mathematicians and statisticians refer to as “correlation” usually refers to the Pearson correlation. The “Standard correlation” would be called “Pearson correlation around zero” by mathematicians and statisticians.

This is how to compute a Standard correlation:

Standard correlation = $\mathbf{a} \cdot \mathbf{b} / (|\mathbf{a}| |\mathbf{b}|)$

Or, in summation notation:

$$\frac{\sum_{i=1}^n a_i b_i}{\sqrt{\left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right)}}$$

Figure 1-1 Summation Notation for the Standard Correlation

Pearson Correlation

The Pearson correlation is very similar to the Standard correlation, except that it measures the angle of expression vectors for genes **A** and **B** around the mean of the expression vectors (for example, the mean of the expression values constituting the profiles for Gene **A** and Gene **B**). Generally the mean of the expression vectors is positive since expression values are based on concentrations of mRNA. Using the Pearson correlation you get more negative correlations than from the Standard correlation (for example, you find more genes that behave opposite to each other, because of where you put the baseline—at zero almost all gene values are above it, at 1 there are a fair amount that read below the baseline).

For data normalized to an overall level of 1 (as with all normalizations that GeneSpring performs) the Pearson correlation gives you almost the same correlations as the Standard correlation when they are both performed on the logarithms of the genes’ expression values.

To compute a Pearson Correlation:

1. Calculate the mean of all elements in vector **a**.
2. Subtract that value from each element in **a**.

Common Correlations

3. Do the same for **b**.

$$\text{Pearson Correlation} = \mathbf{A} \cdot \mathbf{B} / (|\mathbf{A}| |\mathbf{B}|)$$

Or, in summation notation:

$$\frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\left(\sum_{i=1}^n (A_i - \bar{A})^2 \right) \left(\sum_{i=1}^n (B_i - \bar{B})^2 \right)}}$$

Figure 1-2 Summation Notation for the Pearson Correlation

Spearman Correlation

The Spearman correlation is a nonparametric correlation similar to the Pearson correlation except it replaces the data for Gene **A** and **B** with the ranks of the data (i.e. the lowest measurement for a gene becomes 1, the second lowest 2, and so forth). Spearman correlation calculates the correlation of the ranks for Genes **A** and **B**'s expression data around the mean of the ranks, using the same formula as Pearson correlation. In the Spearman correlation only the order of the data are important, not the level, therefore extreme variations in expression values have less control over the correlation. If there are ties in the data, then all of the tied values are assigned the average of the ranks, e.g. if the 5th, 6th and 7th lowest values are tied, all three datapoints are assigned a rank of 6.

To compute a Spearman correlation:

1. Order all the elements of vector **a**.
2. Use this order to assign a rank to each element of **a**.
3. Make a new vector **a'** where the i^{th} element in **a'** is the rank of a_i in **a**.
4. Now make a vector **A** from **a'** in the same way as **A** was made from **a** in the Pearson Correlation.
5. Similarly, make a vector **B** from **b**.

$$\text{Spearman correlation} = \mathbf{A} \cdot \mathbf{B} / (|\mathbf{A}| |\mathbf{B}|)$$

Spearman Confidence

Spearman confidence is a measure of similarity, not a correlation. Spearman confidence is one minus the p-value for the statistical test that the Spearman correlation is zero versus the alternative that is larger than zero. There is a high Spearman confidence value if there is a high Spearman correlation and a low p-value, meaning there is a low probability to find a correlation this high. This measure is very similar to looking for large Spearman

correlation values, but it takes into account the number of sub-experiments in your experiment set.

To compute a Spearman confidence:

If r is the value of the Spearman correlation as described in “Spearman Correlation” on page -4, then:

Spearman confidence = $1 - (\text{probability you would get a value of } r \text{ or higher by chance.})$

Two-Sided Spearman Confidence

Two-sided Spearman confidence is again a measure of similarity but not a correlation. It is very similar to the Spearman confidence discussed in “Spearman Confidence” on page -4, except it is based on the two-sided test of whether the Spearman correlation is either significantly greater than zero or significantly lower than zero. There is a high Two-sided Spearman confidence value if the absolute value of the Spearman correlation is large and has a small p-value, meaning there is a low probability to find a correlation with absolute value this large.

This “similarity” measure is really good for answering the question “What genes behave similarly to a specific gene, and at the same time, what genes behave opposite to a specific gene?”. It should probably not be used for the advanced clustering algorithms (such as k-means and hierarchical clustering) because the genes with high two-sided confidence values are really a mixture of similar and dissimilar genes.

To compute a Two-sided Spearman confidence:

If r is the value of the Spearman correlation as described above, then:

Two-sided Spearman confidence = $1 - (\text{probability you would get a Spearman correlation of } |r| \text{ or higher, or } -|r| \text{ or lower, by chance.})$

Distance

Distance is not a correlation at all, but a measurement of dissimilarity. Distance is based on the measurement of euclidean distance between the expression profile for gene **A** (defined by its expression values for each point in n -dimensional space, where n is the number of experimental points (conditions) with data in your experiment) and the expression profile for gene **B**. This is more formally known as the euclidean metric. To standardize this difference GeneSpring divides by the square root of the number of conditions.

To compute a euclidean distance:

Distance = $|\mathbf{a}-\mathbf{b}| / \text{square root of } n$

Since distance is a measure of dissimilarity, the distance (d) is converted when needed to a similarity measure $1/(1+d)$.

Or, in summation notation:

Common Correlations

$$D = \sqrt{\sum_{i=1}^n (A_i - B_i)^2}$$

The next three metrics should only be used to look at special cases. They are all modified versions of the Standard correlation. Using these three metrics only makes sense when your data are in a sequence, such as “before” and “after”, a time series, or a drug series. The sequence does not have to be continuous, but it must have an order. If your experiment is set up with an experimental point taken at each of “before”, “after”, and “control” the following correlations will not make sense applied to your data.

Smooth Correlation

To compute a Smooth correlation:

Make a new vector **A** from **a** by interpolating the average of each consecutive pair of elements of **a**. Insert his new value between the old values. Do this for each pair of elements that would be connected by a line in the graph screen. Do the same to make a vector **B** from **b**.

$$\text{Smooth correlation} = \mathbf{A} \cdot \mathbf{B} / (|\mathbf{A}| |\mathbf{B}|)$$

$$\text{similarity between gene A and B} = \frac{((7 \cdot 3) + (2.5 \cdot 3.5) + (1.5 \cdot 5)(2.5 \cdot 8))}{((\sqrt{7^2 + 2.5^2 + 1.5^2 + 2.5^2})(\sqrt{3^2 + 3.5^2 + 5^2 + 8^2}))}$$

Experiment	1	2	3	4	5
Gene A	10	4	1	2	3
Gene B	2	4	3	7	9
Gene C	2	8	6	7	8
Between Experiments		1 and 2	2 and 3	3 and 4	4 and 5
Gene A		7	2.5	1.5	2.5
Gene B		3	3.5	5	8
Gene C		5	7	6.6	7.5

Change Correlation

The Change correlation looks for the opposite of what the Smooth correlation looks for. The change correlation only looks at the change in expression level of adjacent points. However, it is also very similar to the Standard correlation, in that it measures the angular separation of expression vectors for genes A and B around zero (i.e. in comparison to zero). However, instead of using the expression values in each experimental point to create the expression vector for gene A, it is based on an arctangent transformation of the ratio between adjacent pairs of experimental points. It uses these to create the expression vector.

This correlation looks for instances where gene A and gene B are changing at the same time. Using the arctangent makes a measure of change that is less sensitive to outliers than using the ratio directly.

To compute a Change correlation:

1. Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**.
2. Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\text{atan}(a_{i+1}/a_i) - \pi/4$.
3. Do the same to make a vector **B** from **b**.

$$\text{Change correlation} = \mathbf{A} \cdot \mathbf{B} / (|\mathbf{A}| |\mathbf{B}|)$$

Upregulated Correlation

The Upregulated correlation is very similar to the Change correlation, except that it only considers positive changes. All negative values for the arc tangent transform of the ratio are set to zero. This emphasizes only periods when new RNA is being synthesized.

To compute an Upregulated correlation:

1. Make a new vector **A** from **a** by looking at the change between each pair of elements of **a**.
2. Do this for each pair of elements that would be connected by a line in the graph screen. The value created between two values a_i and a_{i+1} is $\max(\text{atan}(a_{i+1}/a_i) - \pi/4, 0)$.
3. Do the same to make a vector **B** from **b**.

$$\text{Upregulated correlation} = \mathbf{A} \cdot \mathbf{B} / (|\mathbf{A}| |\mathbf{B}|)$$

Number of Samples Required to do Analyses

For a gene to be included in an analysis listed below, it must have been assigned an expression value in at least half of the total number of conditions in the experiment. Each gene must also have the minimum number of measurements listed in the chart below:

	k-means	SOM	Trees	Find Similar
Standard Correlation	2	N/A	2	2
Distance	1	2	1	1
Smooth Correlation	2	N/A	2	2
Change Correlation	3	N/A	3	3
Unregulated Correlation	3	N/A	3	3
Pearson Correlation	3	N/A	3	3
Spearman Correlation	3	N/A	3	3
Spearman Confidence	5	N/A	5	5
Two-Sided Spearman Confidence	5	N/A	5	5

For details on each of these clustering techniques, see , “Clustering and Characterizing Data”. For Find Similar, see “The Find Similar Command” on page -6.

Technical Details for the Predictor

Gene Selection

In order to select genes for use in the predictor, all genes are examined individually and ranked on their power to discriminate each class from all others, using the information on that gene alone. For each gene, and each class, all possible cutoff points on gene expression level for that gene are considered to predict class membership either above or below that cutoff. Genes are scored on the basis of the best prediction point for that class. The score function is the negative natural logarithm of the p-value for a hypergeometric test (Fisher's exact test) of predicted versus actual class membership for this class versus all others.

A combined list containing the most discriminating genes for each class is produced as the predictor list. Each class is examined in turn, and the gene with the highest score for that class is added to the list, if it is not already on the list. Then genes with the next highest scores for each class are added. This is continued in rotation among the classes until the specified number of predictor genes is obtained. If you save the list of predictor genes as a Gene List, the best prediction score of the gene among the classes for which it would have been added to the list is saved as the attached number on the list.

Classifying the Test Samples

Based on the selected genes, classifications are then predicted for the independent test data, using the k-nearest-neighbors rule. A sample in the independent set is classified by finding the (user specified) k nearest neighbors of the sample among the training set samples, based on Euclidean distance between the normalized expression ratio profiles of the samples. The class memberships of the neighbors are examined, and the new sample is assigned to the class showing the largest relative proportion among the neighbors after adjusting for the proportion of each class in the training set.

Decision Threshold

P-values are computed for testing the likelihood of seeing at least the observed number of neighborhood members from each class based on the proportion in the whole training set. The class with the smallest p-value is given as the predicted class. The column labeled "P-value ratio" is the ratio of the p-value for the best class to that of the second-best class. The predictor will make a prediction if this ratio is less than the "P-value Cutoff" specified on the initial panel, and will not make a prediction if the ratio is above this cutoff. Setting the p-value cutoff to 1 will force the algorithm to always make a prediction but may result in more actual prediction errors.

References for the Predictor

Cover, T.M. and Hart, P.E. (1967) "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, IT-13, 21-27.

Duda, R. O. and Hart, P. E. (1973) Pattern Classification and Scene Analysis, Wiley, New York.

Golub, T.R. et. al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" Science, v286, pp 531-537 (1999)

Glossary

Array A set of spots on a chip, typically expressed as a set of intensity measurements. An array generally has one sample. If all of the interesting genes fit onto one array, the terms array, chip and sample can be considered synonymous.

Array Layout A synthetic picture of genes on arrays. The Array Layout view can be used to check for gross slide related problems

Chip The measurements from a glass slide containing DNA samples for microarray analysis.

Classification A grouping of genes by k-means or SOM clustering that is stored in the Classifications folder.

Cluster A collection of genes that have been grouped according to a certain criteria, such as similar mean expression values.

Colorbar The rectangle on the far right of the main GeneSpring screen. The intensity of the colorbar in GeneSpring indicates the reliability of the data for each gene. Indicate a raw signal strength value to be considered very reliable (a high signal strength) value, an average (a medium signal strength) value, and an unreliable (a low signal strength) value. Any gene with a signal strength (control) above the value indicated as a high signal strength will be colored using the brightest color appropriate, any gene with a signal strength below the value given for unreliable data will be almost black in color. The medium signal value gives the value for the mid-point of the color bar, and genes with a medium signal strength are colored halfway between the two color extremes.

Condition A grouping of one or more samples.

Condition Tree A dendrogram used to show the relationships between the expression levels of conditions.

Control An experiment data set that provides a comparison or contrast to experimental results.

Control Strength The quantity the raw value is divided by to get the normalized value. (see also expression strength)

Data Objects Any downloadable or uploadable items in GeneSpring, such as genomes, gene lists, classifications, etc.

Dendrogram A diagram showing hierarchical relationships, based on similarity between elements, for example, similarity of gene expression levels.

Experiment A group of conditions associated together under one name. This generally means they were all performed using a particular set of parameters.

Experimental Parameter A variable used to describe the condition or conditions during an experiment. A set of parameter values defines a single experimental parameter. When the word “parameter” is used alone, it usually refers to an experimental parameter.

Experiment Interpretation Tells GeneSpring how to treat and display your experiment parameters and how normalized values should be treated.

Experiment Specification Area The area under the genome browser that indicates which (if any) sub-experiment is being displayed, e.g. a particular time point in a time series experiment.

Expression Production of mRNA through transcription of a DNA gene sequence.

Expression Level The amount of mRNA produced by a given gene under specific conditions.

Expression Profile Lines representing gene profiles that you draw in the genome browser. You can then search for genes matching that profile.

External Program Analysis programs outside GeneSpring which can be launched from within GeneSpring. Data from GeneSpring is sent to the program and output from the program is recognized by GeneSpring. These programs are kept in the External Programs folder.

Folders The yellow icons denoting the various directories where data are stored, e.g., Gene Lists folder, Experiments folder, etc.

Gene List A list of genes based on some criteria.

Gene Tree Dendrograms used as a method of showing relationships between the expression levels of genes over a series of conditions.

Genome The set of all genes on a chip or array.

Genome Browser The area of a GeneSpring window containing a visual representation of genes.

Main Screen The first GeneSpring window that appears after you open a genome, such as the default yeast genome window that appears after initially starting the program.

Measurement The smallest “unit” of data recognized by GeneSpring. These raw values can be seen in the upper right table in the Gene Inspector.

Menu Pull-down options that allow you to perform tasks in GeneSpring. The main menu can be found at the top the main GeneSpring window (Windows) or at the top of your screen (Macintosh).

Navigator The left panel of GeneSpring windows containing data organized into folders.

Normalize The use of statistical methods to eliminate systematic variation in microarray experiments that can influence measured gene expression levels.

Panel Section of a window or screen.

Pathways A pathway is a graphical representation of the interaction between gene products in a biological system. Genes can be superimposed on the pathway, allowing you to view their expression levels in a biological context.

Parameter-Value One of the possible values assigned to a variable. For example, in the equation:

$X = \{1, 2, 3 \text{ or } 4\}$ “X” is the experimental parameter and the numbers 1, 2, 3 or 4 are each a different parameter-value of “X”. A more pertinent example is the parameter values breast cancer, kidney cancer, liver cancer, brain cancer, and no cancer could all be different parameter values for the experimental parameter “cancer”.

Parameters

Color Code is similar to a discrete parameter, except you would expect points on a graph with

the same parameters other than this one to be at the same horizontal position. Colors would then be typically used to distinguish these points. Typical examples are the same as for non-continuous parameters. This may be referred to as category.

Continuous Parameter is a numerical parameter for which interpolation makes sense. Graphs using this parameter are line graphs. If there are no continuous parameters in an experiment, then histograms are shown instead of line graphs. A typical example of a continuous parameter is time, or drug concentration. Continuous parameters can optionally be made logarithmic for display purposes.

Non-continuous Parameter is a (possibly numerical) parameter for which drawing lines between points does not make sense, but you still wish to graph it along the horizontal axis. Typical examples of such parameters are drug type, strain of the organism under study, or tissue type. GeneSpring will typically display smaller graphs side by side in the genome browser. This may also be referred to as discrete.

Replicate is not interpreted by GeneSpring. Instead, it is considered a tracking identifier. Sub-experiments that have all parameters (other than the “Replicate” parameter) the same are considered repeats. These are visually represented on graphs by taking the median of the data values and plotting error bars. Typical examples of such parameters are database identifiers, and individual organism names.

Pop-up Menu A list of options that appears from a sub-menu or by right-clicking (Option-click for Macintosh).

Replicate Can be multiple spots on the same array representing the same gene (also referred to as a copy), the same sample in more than one array or a biological replicate - that is equivalent samples taken from more than one organism. A parameter defined as a replicate is graphically a hidden variable; no visual distinction is made based upon this parameter or its parameter values.

Regulatory Sequence The sequence upstream of a given gene to which regulatory enzymes bind, determining the amount of expression of a particular gene.

Sample The measurements taken from one or more chips containing a single liquid sample, or the data generated from a biological object placed onto an array or set of arrays.

Slider A horizontal scrollbar at the bottom of the GeneSpring window that changes the display of genes from one sub-experiment to another, e.g., in a time series experiment, the slider moves the displayed genes across the different time periods.

t-test T-tests calculate p-values which measure the significance of differential gene expression in each condition.

Trust A measure of reliability of the data.

Two-Color Experiment An experiment where a control is used.

Variable A factor such as a disease, drug concentration, patient name, pipette number, time, the strain of organism tested, or who performed the experiment, etc. These variables allow you to look for meaningful patterns in you data and deal sensibly with replicate experiments.

Index

Symbols

.layout file 2-12
examples 2-13

Numerics

3D scatter plot view 4-49
X, Y, and Z axes 4-50, 4-69

A

adding extra genes 2-7
advanced search 4-4
affine background correction 5-13, 5-20
Affymetrix data
normalizing 5-17
animation controls 4-71
secondary 4-71
annotations, updating 6-27
API A-2
arbitrary file restrictions 6-63
array layout view 4-60
attributes 3-35

B

bar graph view 4-39
blocks view 4-36
bookmarks 4-26
browser display
picture 4-37
Build Simplified Ontology 6-31

C

change correlation 6-10, B-6
Change Experiment Interpretation 3-39
changing experiment parameters 3-32, 3-35
changing experimental data range 4-32

class predictor 7-24
classification inspector 4-22
classifications
color by 4-34
CLI A-3
clustering
k-means 7-8
similarity definitions B-2
Color
by Parameter 3-32
color 1-20
background 1-20
changing defaults 1-19
selected 1-20
structure 1-19
trust 4-30
color by classifications 4-34
color by expression 4-30
color by parameter 4-33
color by secondary experiment 4-35
color by significance 4-32
color by venn diagram 4-32
color code 3-31
color options 4-30
coloring scheme 4-30
column assignments, default 3-12
column editor 3-9
advanced options 3-11
assigning columns 3-9
column assignments 3-10
compare genes to genes view 4-64
condition inspector 4-18
condition scatter plot 4-68
Conjectured Regulatory Sequence 6-23
continuous element 3-31
copying and pasting experiments 3-17
copying gene lists 9-5
correlation commands B-3
correlation equations
change correlation B-6

- distance 6-11
- hange correlation 6-10
- Pearson correlation 6-11
- smooth correlation 6-10, B-6
- Spearman confidence 6-11
- Spearman correlation 6-11
- standard 6-10
- standard correlation B-3
- two-sided Spearman confidence 6-11
- upregulated correlation 6-11
- correlations 6-10
 - weighted 7-3
- creating new parameters 3-33, 3-36
- cytogenetic band markers 4-5, 4-6

D

- data format 2-10
- data loading 1-8
- data types
 - restrictions 6-53
- database
 - JDBC driver 1-18
- databases A-2
 - installing GeneSpring from A-2
- DBMS A-2
- decimal markers 8-4
- default colors, changing 4-35
- default column assignments 3-12
- default normalizations 3-21
 - one-color experiments 3-21
 - pre-normalized data 3-21
 - two-color experiments 3-21
- default normalizations, applying 5-4
- deleting objects from GeNet 9-12
- deleting parameters 3-33, 3-36
- display elements
 - hide all 4-71
 - show all 4-71
 - show/hide 4-71
- display options 4-25
 - 3D scatter plot 4-49
 - array layout 4-60
 - bar graph 4-39
 - blocks view 4-36
 - bookmarks 4-26
 - color 4-30
 - color by classification 4-34

- color by expression 4-30
- color by parameter 4-33
- color by secondary experiment 4-35
- color by significance 4-32
- color by venn diagram 4-32
- condition scatter plot 4-69
- error bars 4-28
- graph by genes view 4-65
- graph view 4-37
- legend 4-28
- linked windows 4-25
- no color 4-34
- ordered list 4-58
- pathway view 4-62
- physical position 4-43
- scatter plot 4-45
- split windows 4-25
- tree view 4-56
- vertical axis 4-27
- distance 6-11
- divide signal by control channel 5-7
- downregulated color
 - changing 1-19
- drag and drop
 - gene lists 9-5

E

- electronic signatures 8-4, 9-14
- EMBL files 2-6
 - adding genes to 2-7
- equations
 - overall correlation 7-3
- error bars 4-28
- error model 3-44
 - technical details 3-46
- Euclidian metric B-5
- exclude data window 9-14
- experiment inspector 4-16
 - interpretations 4-17
 - normalizations 4-18
 - parameters 4-17
- experiment interpretation
 - changing 3-39
 - Fold change 3-42
 - log ratio 3-41
 - vertical axis 3-40
- experiment normalizations window 5-2

- experiment parameter 3-35
 - condition 3-31
 - multiple 3-30
 - parameter-value 3-29
- experiment parameters 3-29
 - changing 3-32, 3-35
 - color code 3-31
 - continuous element 3-31
 - creating new 3-33, 3-36
 - deleting 3-33, 3-36
 - display options 3-30
 - hidden elements 3-31
 - importing 3-32, 3-35
 - non-continuous elements 3-31
 - replacing 3-33, 3-36
 - values 3-29
- experimental data range 4-32
- experiments
 - copying and pasting 3-17
 - creating new 3-16
 - inspecting 4-16
- export data
 - by copying 3-20
- exporting
 - gene lists 9-5
- external program interface 8-35
- external programs 8-34, 8-35
 - arguments 8-39
 - creating new 8-35
 - delimiters 8-39
 - inputs 8-36
 - inspecting 8-42
 - outputs 8-37
 - running 8-40
 - scripts and 8-32

F

- FDA compliance
 - electronic signatures 9-14
- files
 - .layout 2-13
- Filter Genes
 - Data File Restriction 6-53
 - Filter on Data File 6-61
 - restricting data types 6-61
- filter on fold change 6-55
- filtering

- gene list numbers 6-64
- find genes 4-4
- Find Potential Regulatory Sequence 6-18
- find similar
 - minimum number B-8
- find similar command 6-6
- find similar genes 4-12
- flag
 - values 3-9
- flags 1-8, 3-11, 5-19
- formula notation B-2

G

- GATC A-3
- GenBank files 2-6
 - adding genes to 2-7
- gene inspector 4-10
 - control 4-11
 - description 4-11
 - normalized 4-11
 - notes 4-13
 - raw 4-11
 - save profile 4-13
 - Student's t-test 4-12
 - t-test p-value 4-11
 - web connections 4-13
- gene list editor
 - filtering methods 6-3
- gene list inspector 4-20
- gene lists
 - copying 9-5
 - creating 6-2
 - dragging 9-5
 - editing 6-2
 - exporting 9-5
 - filtering methods 6-3
 - find similar 6-6
- gene similarity B-2
- genes
 - find similar 4-12
- GeneSpider 6-27
 - lists from annotations 6-12
 - updating master gene table with 6-27
- GeneSpring navigator 1-13
- GeNet
 - deleting from 9-12
 - upload to 9-11

GeNet database 1-5
genomes
 creating 2-2, 2-9
 creating from experiment data 2-9
 definition 1-6
 opening, opening genomes 1-16
graph by genes view 4-65
graph view 4-37

H

help
 ScriptEditor 8-18
help menu
 about 1-5
 SiG on the web 1-4
 system monitor 1-5
 version notes 1-4
hidden elements 3-31
homology tool 6-24
housekeeping genes 5-11

I

icon legend 8-11
import data
 by pasting 3-17
import data command 3-3
importing parameters 3-32, 3-35
inspect
 condition 4-18
 experiment 4-16
 gene 4-10
inspectors
 classification 4-22
 external program 8-42
 gene list 4-20
 script 8-5
installing from CD 1-2
installing from the Web 1-2
interpretations 3-39
interpreted data
 definition 1-7
IUPAC-IUB ambiguity codes 4-5, 4-6

J

JDBC driver 1-18

K

k-means
 minimum number B-8
k-means clustering 7-8

L

legend 4-28
license key, obtaining 1-3
linked windows 4-25
list inspector 4-20
Lists
 from annotations 6-12
 Regulatory Sequences 6-23
 Venn Diagram 6-13
load sequence
 command 4-43
loading data 1-8
loading experiments 3-3
 new experiment checklist 3-7
 the Define File Format window 3-3
 the Merge Files window 3-5
 the Required Sample Attributes window
 3-6
 the Select Corresponding Files window 3-4
 the Select Files window 3-3

M

macintosh tips 1-13
magnification 4-71
managing samples 3-23
master gene table 2-10, 6-27
mathematical notation B-2
measurement flags 5-19
 Abs/Call 3-40
memory 1-3
merge files 3-5
minimum distance 7-4
missing expression values B-2
MySampleAttributes.xml file 3-35

N

navigator 1-13, 4-71
negative control strengths 5-20
new parameters 3-33, 3-36
no color 4-34
nodes 7-11
non-continuous elements 3-31
normalization types 5-6
normalizations 5-2
 add step 5-3
 affine background correction 5-13
 Affymetrix data 5-17
 all samples to specific samples 5-13
 applying default 5-4
 data transformation 5-6
 divide by specific samples 5-13
 divide signal by control channel 5-7
 dye incorporation 5-8
 dye swap 5-7
 edit step 5-3
 gene to itself 5-15
 intensity dependent 5-8
 median polishing 5-15
 negative control 5-7
 negative control strengths 5-20
 per chip 5-10
 per gene 5-13
 per spot 5-10
 per-chip 5-10
 per-spot 5-7
 pre-normalized data 5-12
 real time PCR transform 5-6
 region 5-12, 5-17
 remove step 5-3
 re-order steps 5-4
 repeated measurements 5-18
 reserve control channel 5-8
 set measurements 5-6
 start with pre-normalized values 5-6
 to constant value 5-12
 to median 5-15
 to median or percentile 5-10
 to positive control genes 5-11
 transform from log to linear 5-7
 two-color microarray 5-17
normalizations window 5-2
normalizations, default 3-21

normalized data
 definition 1-7

O

ODBC A-2
one-color experiments 4-31
ontology, building 6-31
ordered list view 4-58
over-expressed color
 changing 1-19

P

panning 4-2
Parameter Interpretations
 fold change (+100% is 1,-50% is -1) 3-42
 log ratio 3-41
 ratio 3-41
 ratio of signal/control 3-41
Parameters
 non-numeric 3-31, 3-34
 numeric 3-31, 3-34
 order 3-34
parameters 3-29
 changing 3-32, 3-35
 color code 3-31
 continuous element 3-31
 creating new 3-33, 3-36
 definition 1-7
 deleting 3-33, 3-36
 display options 3-30
 hidden elements 3-31
 importing 3-32, 3-35
 non-continuous elements 3-31
 non-numeric 3-18
 numeric 3-18
 replacing 3-33, 3-36
 values 3-29
Pathway view 6-16
 adding new elements 6-17
pathway view 4-62
Pearson correlation 6-11
per gene normalizations 5-13
percent explained variability 4-23
per-chip normalizations 5-10
per-spot normalizations 5-7
physical position view 4-41

picture
 secondary 4-71
predictor 7-24
preferences 1-18
 browser 1-21
 color 1-18
 data directory 1-18
 data files 1-18
 database 1-18
 default correlation 1-23
 default font 1-24
 default genome 1-18
 disk cache size 1-22
 firewall 1-21
 gene labels 1-21
 GeNet 1-22
 license manager 1-22
 memory 1-22
 misc 1-23
 remote execution 1-23
 restrict gene list searches 1-23
 system 1-22
 text color 1-20
 web browser defaults 1-21
principal components analysis 7-16
print
 trees with labels 4-55
printing 9-4
publish to GeNet 9-12
 exclude data 9-14

R

raw data
 definition 1-7
region normalization 5-17
regions 1-7, 3-11
regulatory compliance 8-4
 electronic signatures 9-14
Regulatory Sequence 6-18
 Expected 6-22
 Observed 6-22
 P-value 6-22
 Random Rate 6-22
 Sequence 6-22
 Single P 6-22
 Tests 6-22
remote server 8-5

replacing parameters 3-33, 3-36
replicates
 definition 1-7
required sample attributes 3-6
restrict data types
 Control Signal 6-53
 Normalized Data 6-53
 Raw Data 6-53
restricting data types 6-61
restrictions
 arbitrary 6-63
 data types 6-53
 filter on fold change 6-55
R-values 7-4

S

sample attributes 3-35
sample manager 3-23
 filter on attributes 3-26
 filter on experiment 3-25
 filter on keyword 3-27
 filter on parameter 3-25
 filtering methods 3-25
saving images 9-2
scatter plot
 condition 4-68
scatter plot view 4-45
 vertical/horizontal axes 4-46
scatter plot view, 3D 4-49
script building blocks 8-7, 8-20
 ANOVA 8-31
 boolean 8-20
 boolean select 8-21
 clustering 8-27
 correlations 8-28
 count groups 8-27
 filtering 8-29
 gene list manipulations 8-21
 GeNet downloading 8-22, 8-23
 GeNet publishing 8-23
 input 8-7
 look up 8-24
 make groups 8-25
 merge-split groups 8-24
 numbers 8-26
 output 8-7
 select groups 8-25

- statistical analysis 8-31
- script inputs
 - delete 8-17
 - move to end 8-17
 - move to start 8-17
- script output
 - boolean 8-14
 - numbers 8-14
 - sequence information 8-14
- script outputs
 - delete 8-17
 - move to end 8-17
 - move to start 8-17
- script primitives 8-7
- ScriptEditor 8-7
 - blocks 8-16
 - browser 8-9
 - building blocks 8-13, 8-20
 - change information 8-12
 - concepts 8-7
 - help 8-18
 - icon legend 8-11
 - inputs 8-13
 - knobs 8-14
 - move to end 8-17
 - move to start 8-17
 - navigator 8-8
 - notes 8-9
 - outputs 8-13
 - saving scripts 8-18
 - support 8-18
 - warning messages 8-18
- scripts 8-2, 8-3, 8-7
 - 2-fold expression change 8-2
 - best k-means 8-2
 - blocks 8-16
 - building blocks 8-13, 8-20
 - change information 8-12
 - clustering 2-fold change list 8-2
 - definition 8-2
 - filter on noise 8-2
 - find similar genes 8-2
 - help 8-18
 - input 8-7
 - inputs 8-13
 - inspecting 8-5
 - knobs 8-7, 8-14
 - output 8-7
 - outputs 8-13
 - pairwise comparison 8-2
 - parameters for data file restriction 8-4
 - PEER-S 8-2
 - predefined 8-2
 - remote server 8-5
 - running 8-3
 - saving 8-18
 - select k-means 8-2
 - send clustering results to GeNet 8-2
 - series of k-means 8-3
 - sockets 8-7
 - support 8-18
 - to external programs 8-32
 - warning messages 8-18
- search
 - simple 4-4
- secondary picture 4-71
- select a gene(s)
 - deselect a gene 6-15
- selecting genes 4-4
- self-organizing maps 7-8
- separation ratio 7-8
- simple search 4-4
- simplified ontology 6-31
- smooth correlation 6-10, B-6
- SOM
 - Euclidean distance 7-11
 - minimum number B-8
- Spearman confidence 6-11
- Spearman correlation 6-11
- split windows 4-25
 - classification 4-35
- spreadsheet view 4-67
- SQL A-2
- standard correlation 6-10, B-3
- starting GeneSpring 1-3
- system requirements 1-2

T

- text color 1-20
- tick spacing 4-28
- tree view 4-52
 - labels 4-55
 - magnifying 4-55
 - printing 4-55
 - viewing gene names in 4-56

- viewing nodes 4-55
- viewing parameters in 4-56
- viewing subtrees 4-52
- trees
 - comparing genes in nodes 4-52
 - minimum distance 7-4
 - minimum number B-8
 - R-values 7-4
 - separation ratio 7-8
- trust 4-30
- t-test 4-12
- two-color experiments 4-30
- two-sided Spearman confidence 6-11

U

- under-expressed color
 - changing 1-19
- Update annotations 6-27
- updating annotations 6-27
- updating master gene table 6-27
- upload to GeNet 9-12
- uploading to GeNet 9-11
- upregulated color
 - changing 1-19
- upregulated correlation 6-11

V

- version notes 1-4
- vertical axis 4-27
- view gene details 4-10
- views
 - 3D scatter plot 4-49
 - array layout 4-60
 - bar graph 4-39
 - blocks 4-36
 - compare genes to genes 4-64
 - graph 4-37
 - graph by genes 4-65
 - ordered list 4-58
 - pathway 4-62
 - physical position 4-41
 - scatter plot 4-45
 - spreadsheet 4-67
 - trees 4-52

W

- web databases 2-4
 - special character 2-4
- web links 4-13
- wizards
 - new genome 2-2