Related Documents


# Table of Contents

## Chapter 1 Normalizing Options

1.1 Background Subtractions ................................................................. 1-1
1.2 Normalize to Negative Controls ...................................................... 1-2
   1.2.1 Mathematical Illustration of the Normalize to Negative Controls Method 1-2
1.3 Normalize to Control Channel Values for Each Gene ....................... 1-3
   1.3.1 Mathematical Illustration of the Normalize to a Control Channel Value for Each Gene Method ...................................................... 1-4
1.4 Normalize to Positive Controls ......................................................... 1-5
   1.4.1 Mathematical Illustration the Normalize to Positive Controls Method .... 1-5
1.5 Normalize Each Sample to Itself ...................................................... 1-6
   1.5.1 Mathematical Illustration of the Normalize Each Sample to Itself Method 1-7
1.6 Normalizing Each Sample to a Hard Number ...................................... 1-7
1.7 Normalizing Each Gene to Itself ...................................................... 1-8
   1.7.1 Mathematical Illustration of the Normalizing Each Gene to Itself Method 1-9
1.8 Normalize All Samples to Specific Samples ....................................... 1-11
   1.8.1 Required Syntax for Normalization to Specific Samples .................. 1-11
   1.8.2 Mathematical Illustration of the Normalizing Samples to a Specific Sample Method ................................................................. 1-13
1.9 Region Normalization ................................................................... 1-16
1.10 Dealing with Repeated Measurements .......................................... 1-17
   1.10.1 Single Data File ........................................................................... 1-17
   1.10.2 Mathematical Illustration of the Dealing with Repeated Measurements in a Single Data File Method .............................................. 1-17
   1.10.3 Measurement Flags ..................................................................... 1-18
1.11 Negative Signal Strengths ............................................................... 1-18
1.12 Normalization for Particular Array Types ...................................... 1-19

## Chapter 2 Experiment Parameters

2.1 Definitions of Parameters .............................................................. 2-1
   2.1.1 Data objects in the Navigator .......................................................... 2-1
   2.1.2 Parameter Values .......................................................................... 2-2
   2.1.3 A Note on Multiple Parameters ...................................................... 2-4
2.2 Parameter Display Options ............................................................. 2-8
   2.2.1 Replicate or Hidden Element .......................................................... 2-8
   2.2.2 Continuous Element ...................................................................... 2-11
   2.2.3 Non-Continuous Element (Set) ....................................................... 2-16
   2.2.4 Color Code ................................................................................... 2-20
   2.2.5 Ordering Parameter Values ............................................................ 2-25
   2.2.6 Why change the parameter display? ............................................... 2-27
2.3 Changing the Parameter Interpretations ........................................... 2-31
Chapter 3 Installing a New Genome

3.1 The Autoloader ................................................................. 3-1
3.2 Creating Folders for New Genomes ........................................ 3-2
3.3 Raw Data .............................................................................. 3-2
   3.3.1 What Data Are Necessary? ................................................ 3-2
   3.3.2 What Format do these Data Need to be in? ......................... 3-3
3.4 Genome Wizard ........................................................................ 3-11
   3.4.1 Welcome to the GeneSpring Genome Installation Wizard! panel ...... 3-11
   3.4.2 Genome Data Directory panel ........................................... 3-12
   3.4.3 The Overall Genome Properties panel ................................. 3-14
   3.4.4 The GenBank Data File panel ............................................ 3-14
   3.4.5 The Master Gene Table panel ............................................ 3-16
   3.4.6 The Genome Sequence File panel ...................................... 3-17
   3.4.7 The Additional Genetic Elements panel ............................... 3-18
   3.4.8 The Links to Web DataBases panel ..................................... 3-19
   3.4.9 The Miscellaneous Settings panel ...................................... 3-22
   3.4.10 The Finished panel ......................................................... 3-22
3.5 Installing from a Text File .............................................................. 3-23
   3.5.1 Creating Folders for New Genomes ................................... 3-23
   3.5.2 The .genomedef File ......................................................... 3-24
3.6 Example of a genome being installed ........................................... 3-29
   3.6.1 Raw data files ................................................................. 3-29
   3.6.2 The .genomedef file ........................................................ 3-32
3.7 Updating Your Information ............................................................... 3-34
   3.7.1 Options in the GeneSpider ................................................. 3-34
   3.7.2 Updating a GenBank or EMBL File .................................... 3-36
   3.7.3 Updating GenBank Information using the GeneSpider ............ 3-36
   3.7.4 Updating LocusLink Information using the GeneSpider ........... 3-36
   3.7.5 Updating UniGene Information using the GeneSpider ............. 3-37

Chapter 4 Installing a New Experiment

4.1 Some Helpful Definitions .............................................................. 4-1
4.2 Autoload Experiment ....................................................................... 4-2
   4.2.1 Headlines ............................................................................ 4-3
   4.2.2 Functions ............................................................................. 4-4
   4.2.3 Autoloader Normalizations ............................................... 4-5
   4.2.4 Remember this Format ..................................................... 4-5
   4.2.5 Load Now ........................................................................... 4-5
4.3 File Formats .................................................................................. 4-6
   4.3.1 Raw Data ............................................................................. 4-6
   4.3.2 What format does this data need to be in? ............................. 4-6
   4.3.3 Where do I put my data? .................................................... 4-13
GeneSpring Loading Data Manual

4.4 Experiment Import Wizard ................................................................. 4-15
   4.4.1 The Welcome panel of the GeneSpring Experiment Entry Wizard ... 4-15
   4.4.2 The Data File Format panel ......................................................... 4-16
   4.4.3 The Properties of Experiment Set panel ....................................... 4-18
   4.4.4 The Number of Arrays panel ....................................................... 4-18
   4.4.5 The Number of Parameters panel ............................................... 4-20
   4.4.6 The Parameter Characteristics panel ............................................ 4-21
   4.4.7 The How to Display the Parameters panel .................................... 4-22
   4.4.8 The Parameter Values panel ....................................................... 4-23
   4.4.9 The Describe your Data Files panels ......................................... 4-24
   4.4.10 The Data File Header Lines panel ............................................. 4-27
   4.4.11 The Region Normalization panel ............................................... 4-27
   4.4.12 The Gene Name panel ............................................................... 4-29
   4.4.13 The Gene Name Prefix Removal panel ...................................... 4-30
   4.4.14 The Gene Name Suffix Removal panel ...................................... 4-31
   4.4.15 The Data Column Location panel ............................................. 4-32
   4.4.16 The Control Channel Values panel ........................................... 4-33
   4.4.17 The Flags panel .................................................................. 4-34
   4.4.18 The Sample Photos panel ......................................................... 4-36
   4.4.19 The Array Photos panel ......................................................... 4-37
   4.4.20 The RT – PCR Experiments panel .......................................... 4-38
   4.4.21 The Normalizations: Negative Controls panel ............................ 4-38
   4.4.22 The Normalizations: Control Channel Values panel .................. 4-39
   4.4.23 The Normalizations: Positive Controls panel ............................. 4-40
   4.4.24 The Normalizations: Each Sample to Itself panel ....................... 4-41
   4.4.25 The Normalizations: Each Sample to a Hard Number panel ........ 4-42
   4.4.26 The Normalizations: Each Gene to Itself .................................. 4-42
   4.4.27 The Normalizations: All Samples to Specific Samples panel ....... 4-43
   4.4.28 The Graphics Specifications panel ............................................ 4-44
   4.4.29 The Finish panel ................................................................... 4-45
4.5 Installing from a Text File ............................................................... 4-46
   4.5.1 Define Your Experiment ............................................................ 4-46
   4.5.2 Define Your Parameters ............................................................ 4-47
   4.5.3 Describe your Data Files ............................................................ 4-51
   4.5.4 Data File Header Lines ............................................................. 4-52
   4.5.5 Gene Names ............................................................................ 4-53
   4.5.6 Explain to GeneSpring how to locate only the Gene Name .......... 4-53
   4.5.7 Explain to GeneSpring How to Read the Region Specifications .... 4-54
   4.5.8 Locate the Data Column ........................................................... 4-54
   4.5.9 The Control Channel Value ...................................................... 4-56
   4.5.10 Measurement Flags ............................................................... 4-57
   4.5.11 Associating a Picture with a Sample ........................................ 4-58
   4.5.12 Normalizations: Negative Controls ......................................... 4-59
   4.5.13 Normalizations: Control Channel Values .................................. 4-60
   4.5.14 Normalizations: Positive Controls ............................................ 4-61
   4.5.15 Normalizations: Each Sample to Itself ..................................... 4-62
   4.5.16 Normalizations: Each Gene to Itself ......................................... 4-62
GeneSpring Loading Data Manual

4.5.17 Normalizations: Each Sample to a Specific Sample ....................... 4-63
4.5.18 Colorbar Specifications ............................................................... 4-63
4.5.19 Graph Specifications ................................................................. 4-64
4.5.20 Examples of Installed Experiments ............................................ 4-64
4.6 Copying and Pasting Experiments .................................................. 4-71
  4.6.1 Preparation for Pasting ............................................................. 4-71
  4.6.2 Copying an Experiment or a List Out of GeneSpring .................. 4-77
4.7 Installing from a Database ............................................................. 4-78
  4.7.1 Custom Databases and GeneSpring .......................................... 4-78
  4.7.2 Adding an Experiment from a Database .................................... 4-80

Index ........................................................................................................ 1-1
Chapter 1 Normalizing Options

There are seven normalization options available in GeneSpring:

- Normalize to Negative Controls
- Normalize to Control Channel Values for Each Gene
- Normalize to Positive Controls
- Normalize Each Sample to Itself
- Normalize Each Gene to Itself
- Normalize all Samples to Specific Sample
- Region Normalization

You can follow the directions in any or all of these sections, as appropriate, to normalize your data. In a few cases, it would not make sense to apply two options together, for instance: normalizing each sample both to a positive control and across the whole sample, or normalizing each gene to itself (across all samples) and to a specific sample. The GeneSpring Experiment Wizard will only allow you to choose one of each of these. Other than those instances, you may choose any options appropriate to your data. The order the normalizations are performed in is mathematically significant. GeneSpring performs normalizations in the order listed above. Three normalizations can be applied either to samples or regions (normalize to negative controls, normalize to positive controls, and normalize each sample or region to itself) and are assumed to apply to samples unless otherwise specified. See 1.9 “Region Normalization” on page 1-16 for further information. For instructions on how to implement any of these normalizations see Chapter 4, Installing a New Experiment. There is one normalization in addition to those listed whose implementation is automatic: repeated measurements in a single data file are assumed to be repeats and will be averaged before any of the six main normalizations are implemented. See 1.10 “Dealing with Repeated Measurements” on page 1-17 for details.

1.1 Background Subtractions

When considering how to transform raw data to normalized data, the first thing that may be necessary is to subtract an estimate of background level. The background level is taken from a separate column in your data set. Typically there will be a column labeled “negative control” containing information on the background level data. The median value of the negative controls will be subtracted from the raw values for each gene before anything else is done.
1.2 Normalize to Negative Controls

If you have any genes designated as negative controls on your array (usually, you have negative controls when there is DNA from a different genome than the one you are investigating on the array), you can normalize the data using this information. This normalization removes the background from the experimental readings by giving you a general idea of the lowest amount of exposure possible for signals taken from a particular array and then subtracting this amount from your raw experimental results. The formula used is:

\[
\text{(the signal strength of gene A in sample X) - (the median signal of the negative controls in sample X)}
\]

Once you normalize to negative controls, you probably want to either normalize to positive controls or each sample to itself and then normalize each gene to itself.

### 1.2.1 Mathematical Illustration of the Normalize to Negative Controls Method

Given the raw data with negative controls:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1008</td>
<td>2060</td>
<td>1510</td>
</tr>
<tr>
<td>CLN2</td>
<td>1008</td>
<td>2060</td>
<td>510</td>
</tr>
<tr>
<td>CDC28</td>
<td>108</td>
<td>260</td>
<td>60</td>
</tr>
<tr>
<td>HSL1</td>
<td>1008</td>
<td>2060</td>
<td>510</td>
</tr>
<tr>
<td>YGP1</td>
<td>10008</td>
<td>20060</td>
<td>5010</td>
</tr>
<tr>
<td>Control 1</td>
<td>7</td>
<td>58</td>
<td>10</td>
</tr>
<tr>
<td>Control 2</td>
<td>8</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>Control 3</td>
<td>9</td>
<td>63</td>
<td>20</td>
</tr>
</tbody>
</table>
The same data normalized to negative controls:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1000</td>
<td>2000</td>
<td>1500</td>
</tr>
<tr>
<td>CLN2</td>
<td>1000</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>CDC28</td>
<td>100</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>HSL1</td>
<td>1000</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>YGP1</td>
<td>10 000</td>
<td>20 000</td>
<td>5000</td>
</tr>
<tr>
<td>Median of the Controls</td>
<td>8</td>
<td>60</td>
<td>10</td>
</tr>
</tbody>
</table>

See Chapter 4, Installing a New Experiment for how to implement this normalization option.

1.3 Normalize to Control Channel Values for Each Gene

Control Channel Values are intended to provide a baseline. Different samples can be compared to the baseline and to one another. By using these comparisons, you can determine variations caused by the particular experimental conditions you are exploring, rather than the overall sample conditions. If you have a control channel value to indicate the trust you have in your experimental data, you probably want to normalize the genes by dividing their signal strength by the control’s signal strength. The formula for this normalization option looks like this:

\[
\frac{\text{signal strength of gene A in sample X}}{\text{control channel value for gene A in sample X}}
\]

In two-color experiments the control channel is often a green signal. If you normalize to the control channel for each gene you may also want to normalize each sample to itself or to a positive control. This will provide a control for sources of variability affecting the whole chip, for example, variations in the amounts of dye added, etc. You probably do not, however, need to normalize each gene to itself or to a single control sample.
1.3.1  Mathematical Illustration of the Normalize to a Control Channel Value for Each Gene Method

Given raw data with a Control Channel:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Reference 1</th>
<th>Sample 2</th>
<th>Reference 2</th>
<th>Sample 3</th>
<th>Reference 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
<td>2000</td>
<td>1500</td>
<td>500</td>
</tr>
<tr>
<td>CLN2</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
<td>2000</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>CDC28</td>
<td>100</td>
<td>100</td>
<td>200</td>
<td>200</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>HSL1</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
<td>2000</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>YGP1</td>
<td>10 000</td>
<td>10 000</td>
<td>20 000</td>
<td>20 000</td>
<td>5000</td>
<td>5000</td>
</tr>
</tbody>
</table>

The results of normalizing to a control channel for each gene:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CLN2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CDC28</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HSL1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YGP1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

See Chapter 4, Installing a New Experiment for how to implement this normalization option.
1.4 Normalize to Positive Controls

This normalization method is intended to remove the differences in amount of exposure between samples, providing you with a baseline so different samples are comparable to one another. Positive controls give you a general idea of how well the array responded to exposure. Normalizing to positive controls will factor in this information with the experimental results you analyze. You can normalize your data with this method if you have genes designated as positive controls on your array (you usually have positive controls when there is DNA from a different genome than the one you are investigating on your array, and you added a known quantity of that DNA to your sample). The formula used to do this is:

\[
\frac{\text{the signal strength of gene A in sample X}}{\text{the median signal of the positive controls in sample X}}
\]

This normalization should not be used with normalizing each sample to itself, as they are both intended to address the same issue. After normalizing to positive controls you probably still want to normalize each gene to itself.

1.4.1 Mathematical Illustration the Normalize to Positive Controls Method

Given raw data with positive controls:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1000</td>
<td>2000</td>
<td>1500</td>
</tr>
<tr>
<td>CLN2</td>
<td>1000</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>CDC28</td>
<td>100</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>HSL1</td>
<td>1000</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>YGP1</td>
<td>10000</td>
<td>20000</td>
<td>5000</td>
</tr>
<tr>
<td>Control 1</td>
<td>5000</td>
<td>10000</td>
<td>2500</td>
</tr>
<tr>
<td>Control 3</td>
<td>2000</td>
<td>4000</td>
<td>1000</td>
</tr>
</tbody>
</table>
The results of normalizing to positive controls:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>0.5</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>CLN2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>CDC28</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>HSL1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>YGP1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

See Chapter 4, Installing a New Experiment for how to implement this normalization option.

### 1.5 Normalize Each Sample to Itself

This normalization is intended to remove the differences in amount of exposure between samples, so different samples are comparable to one another. This method makes the median of all of your measurements 1, for each sample. The formula used to do this is:

\[
\frac{\text{the signal strength of gene A in sample X}}{\text{the median of all of the measurements taken in sample X}}
\]

This normalization should not be used with normalizing to positive controls, as they are both intended to address the same issue. If you do not have either positive controls or a reference it is strongly suggested you normalize each sample to itself.

This option is also referred to as Distribution of All Genes or Global Scaling. Please refer to *GeneSpring User Manual* 5.4.3.4, “Use Distribution of All Genes” and 1.11 “Negative Signal Strengths” on page 1-18.
1.5.1 Mathematical Illustration of the Normalize Each Sample to Itself Method

Given raw data without positive controls or control channel:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN1</td>
<td>1000</td>
<td>2000</td>
<td>1500</td>
</tr>
<tr>
<td>CLN2</td>
<td>1000</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>CDC28</td>
<td>100</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>HSL1</td>
<td>1000</td>
<td>2000</td>
<td>500</td>
</tr>
<tr>
<td>YGP1</td>
<td>10,000</td>
<td>20,000</td>
<td>5000</td>
</tr>
</tbody>
</table>

The results of normalizing each sample to itself:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CLN2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CDC28</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>HSL1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YGP1</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

After Normalizing Each Sample to Itself

You can use this normalization in concert with Normalize Each Gene to Itself.

1.6 Normalizing Each Sample to a Hard Number

You would normally only use this function if you have pre-normalized data, such as data prepared with Affymetrix’s Global Scaling™. In that instance, you would want to divide all data by 2500 (or whatever number you chose to normalize by using the Affymetrix software). You will need to do this because the GeneSpring analysis algorithms assume your data is normalized to a median of 1. GeneSpring will use the following formula:

\[
\text{Normalized Signal} = \frac{\text{Signal Strength of Gene A in Sample X}}{\text{Hard Number in Sample X}}
\]

You can use this normalization in concert with Normalize Each Gene to Itself.
1.7 Normalizing Each Gene to Itself

This normalization method is intended to remove the differing intensity scales from multiple experimental readings. It normalizes each gene to itself, so the median of all of the measurements taken for that gene is one. With this normalization, you may graph a set of similar genes (defined as similar by using the correlation coefficient) and the experimental points will be graphically similar to one another. They are all on the same vertical scale, rather than the same pattern of changes on widely differing vertical levels. The formula used is:

\[
\frac{\text{(the signal strength of gene A in sample X) - the median of every measurement taken for gene A throughout all of the samples)}}{\text{(the median of every measurement taken for gene A throughout all of the samples)}}
\]

Do not use this normalization method in concert with normalizing all samples to a specific sample, as they are both intended to address the same issue. If you are using GeneSpring to do all of your normalizations, and you are not doing a two-color experiment, using this normalization method is highly recommended. This normalization option is commonly combined with either normalizing each sample to itself or normalizing to positive controls. As it is more striking mathematically to illustrate it as the second step of normalization, there are two mathematical illustrations, one following the normalization of each sample to itself, and the second following normalization to positive controls. For explanations of either of these first normalizations see 1.5 “Normalize Each Sample to Itself” on page 1-6 or 1.4 “Normalize to Positive Controls” on page 1-5.

You can specify a cutoff to prevent small and negative measurements from participating in the normalization. The cutoff is specified in terms of measurement values that have been partially normalized in previous normalization steps, so if your data has other (e.g. per-sample) normalizations, this should probably be a small number, like 0.01.

Obviously, this normalization needs more than one sample to make sense. It can be considered a synthetic control.
1.7.1 Mathematical Illustration of the Normalizing Each Gene to Itself Method

Data normalized by “Normalize Each Sample To Itself”:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CLN2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CDC28</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>HSL1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YGP1</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

The results of normalizing each gene to itself:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CLN2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CDC28</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HSL1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YGP1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Data normalized by “Normalize to Positive Controls”:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>0.5</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>CLN2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>CDC28</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>HSL1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>YGP1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

The results of normalizing each gene to itself:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CLN2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CDC28</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HSL1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YGP1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

See Chapter 4, Installing a New Experiment for how to implement this normalization option.
1.8 Normalizing All Samples to Specific Samples

This normalization option is intended to remove differing intensity scales from each sample by comparing all of the samples to one or more specific samples. The formula for this is:

\[
\frac{\text{the signal strength of gene A in sample X}}{\text{the signal strength of gene A in the control sample(s)}}
\]

Do not use this normalization method in concert with normalizing each gene to itself or normalizing to control channel values, as they are all intended to address the same issue. Unless your experiment was designed with specific control samples, it is recommended you normalize each gene to itself (i.e. to the median across all samples) rather than using this normalization method.

Only use this normalization if you have control samples for which you consider the measurements very reliable and you want all of the measurements for the other samples to be in relation to those very reliable samples. You will need normalization definitions for all your samples before you begin this.

1.8.1 Required Syntax for Normalization to Specific Samples

In this scenario you will need to use a very specific syntax to describe your samples.

If you are normalizing to a single sample, indicate the sample number in the box labeled “Enter Sample Number(s)”.

If you wish to normalize all of your samples to the mean of a set of control samples, indicate the sample numbers of the control samples. Multiple sample numbers must be separated by commas (e.g. 1,2). Ranges of sample numbers can be indicated by a dash (e.g.1-3,5).

- Example 1:
  1-3,5
  Translation: normalize all samples to the mean of samples 1, 2, 3, and 5.

Alternatively, you can normalize subsets of samples to the mean of specific subsets of control samples. Begin by listing those samples to be used as controls for a majority of the samples (as described above). For samples to be normalized to the mean of a different set of samples, add (in parentheses) a list of sample numbers for the samples to be normalized, followed by a colon, followed by a list of sample numbers for the control samples. You may specify as many of these lists as you need.

- Example 2
  1(5:4)
  Translation: normalize all samples to sample 1 (including sample 4), except for sample 5, which should be normalized to sample 4.
• **Example 3**

1(5,6:4)(7-10:7,8)
Translation: normalize all samples to sample 1 except for samples 5, 6, and 7 through 10. Sample 5 and 6 should be normalized to sample 4, and sample 7 through 10 should be normalized to the mean of samples 7 and 8.

• **Example 4**

1,2(3-5,7:3-4)(6,8-9:5)
Translation: all samples will be normalized to the arithmetic mean of samples 1 and 2, except for samples 3 through 5, and 7, which will be normalized to the average of samples 3 and 4. In addition, samples 6, 8, and 9 will be normalized to sample number 5.

• **Example 5**

The various parenthetical phrases will occur all at once, so you may place any piece in any place in the string.

(1,2:7)(7:7)(3,4:8)(8:8)(5,6:9)(9:9) is the same as
(7:7)(1,2:7)(8:8)(3,4:8)(5,6:9) is the same as
(1,2,7:7)(3,4,8:8)(5,6,9:9) is the same as
7(3,4,8:8)(5,6,9:9)
Translation: samples 1, 2, and 7 will be normalized to sample 7, samples 3, 4, and 8 will be normalized to sample 8, and samples 5,6, and 9 will be normalized to sample 9. All values for the normalized samples 7, 8, and 9 will equal one.

If you have a cutoff, then the scaling factor for this step of the normalization is computed by taking the arithmetic mean over the set of control sample measurements that have values (are not N/A) and are above the cutoff. If no such values are present for a given gene, then a special normalization is done. In this case, the cutoff value itself is used as the basis of the normalization. Any sample with a measurement level greater than or equal to the cutoff will be normalized by this factor, and any sample with measurement level less than this cutoff will be have a normalized value set to N/A. This is done in order to avoid losing data for genes that might have low measurement levels in the control group, but significantly upregulated levels in the treatment groups, without introducing artificially downregulated values.
1.8.1.1 Special cases

As an example, you might have patients, controls and drugs arranged in the following manner. There are a total of nine samples.

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug X</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Drug Y</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Drug Z</td>
<td>9</td>
<td>5</td>
</tr>
</tbody>
</table>

- To normalize the control to itself, use this syntax:
  \[(1,2,7:7)(3,4,8:8)(5,6,9:9)\]
  This will finish with sample 1 divided by raw 7, 2 divided by raw 7 and 7 divided by raw 7. All values for the normalized sample 7 will equal one.

- To normalize the control to the average of controls:
  If you want to see sample 1 divided by the raw 7, sample 2 divided by raw 7 and sample 7 divided by the average of sample 7, 8 and 9, you must use this syntax:
  \[(1,2:7)(3,4:8)(5,6:9)(7,8,9:7,8,9)\]
  This will divide sample 1 by the raw data of 7, sample 2 by the raw data of 7 and sample 7 by the average of sample 7, 8 and 9.

1.8.2 Mathematical Illustration of the Normalizing Samples to a Specific Sample Method

As an example, your experiment might be designed with three different types of tissues, 3 control samples and 6 treated samples arranged in the following manner. There are a total of nine samples.

<table>
<thead>
<tr>
<th>Tissue Type X</th>
<th>Control</th>
<th>Treated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 7</td>
<td>Sample 1</td>
<td>Sample 2</td>
</tr>
<tr>
<td>Sample 8</td>
<td>Sample 3</td>
<td>Sample 4</td>
</tr>
<tr>
<td>Sample 9</td>
<td>Sample 5</td>
<td>Sample 6</td>
</tr>
</tbody>
</table>
The results of normalizing each sample to itself:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Treated Samples</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tissue X</td>
<td>Tissue Y</td>
</tr>
<tr>
<td>CLN 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CLN2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CDC28</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>HSL1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>YGP1</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>

Samples 1, 2 and 7 are normalized to sample 7, and samples 3, 4, and 8 are normalized to sample 8, and samples 5, 6, and 9 are normalized to sample 9. Note that the normalized data for every gene in each of the three control samples will be 1.
Another way to use this normalization method requires that your experiment be designed to have a set of controls that you wish to use, en mass, as the controls for your experiment. In other words, you want to normalize all of your samples to the arithmetic mean of a set of controls.

### After Normalizing Each Sample to the Control Sample

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Tissue X</th>
<th>Tissue Y</th>
<th>Tissue Z</th>
<th>Tissue X</th>
<th>Tissue Y</th>
<th>Tissue Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cln1</td>
<td>1</td>
<td>1</td>
<td>2.5</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cln2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cdc28</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>Hsl1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ygp1</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### After Normalizing Each Sample to Itself

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Treated Samples</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cln1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cln2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cdc28</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Hsl1</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>Ygp1</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>
After normalizing each sample to itself the samples are normalized to samples to the average of the controls. Note that this allows you to analyze the variability among the controls as well as the treated samples.

**Normalizing Options Region Normalization**

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Sp. 1</th>
<th>Sp. 2</th>
<th>Sp. 3</th>
<th>Sp. 4</th>
<th>Sp. 5</th>
<th>Sp. 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN 1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CLN2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>CDC28</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HSL1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.25</td>
<td>.25</td>
<td>2.5</td>
</tr>
<tr>
<td>YGP1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

See Chapter 4, Installing a New Experiment for how to implement this normalization option and for how to define a region.

### 1.9 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 need to normalize sections of an array separately from one another. Region normalization is not a separate mathematical 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 three normalization options become:

**Normalizing to Negative Controls for a Region:**

\[
\text{Normalizing to Negative Controls for a Region:} \\
\left(\text{the signal strength of gene A in region Y of sample X}\right) - \left(\text{the median signal of the negative controls in region Y of sample X}\right)
\]

**Normalizing to Positive Controls for a Region:**

\[
\text{Normalizing to Positive Controls for a Region:} \\
\left(\text{the signal strength of gene A in region Y of sample X}\right) - \left(\text{the median signal of the positive controls in region Y of sample X}\right)
\]

**Normalizing Each Region to Itself:**

\[
\text{Normalizing Each Region to Itself:} \\
\left(\text{the signal strength of gene A in region Y of sample X}\right) - \left(\text{the median of all of the measurements taken in region Y of sample X}\right)
\]

See Chapter 4, Installing a New Experiment for how to implement this normalization option and for how to define a region.
1.10 Dealing with Repeated Measurements

1.10.1 Single Data File

Occasionally the raw experimental data in the data file for your sample will have 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 will automatically consider the measurements repeats and average all of the signal strengths together. GeneSpring will report the average to you, and it will keep track of the minimum and maximum values for each gene, but GeneSpring will not be able to access the particular values falling between the minimum and maximum values. The formula for averaging a repeated gene is:

\[
\frac{[(\text{the signal strength of gene A1}) + (\text{the signal strength of gene A2}) + ... + (\text{the signal strength of gene An})]}{N}
\]

This process is done for every gene repeated in a data file, and it is done 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 Chapter 2, Experiment Parameters for more information about repeats reported in separate data files.

1.10.2 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):

<table>
<thead>
<tr>
<th>YMR199W</th>
<th>1117</th>
</tr>
</thead>
<tbody>
<tr>
<td>YMR200W</td>
<td>1384</td>
</tr>
<tr>
<td>YMR201C</td>
<td>1101</td>
</tr>
<tr>
<td>YMR202W</td>
<td>1357</td>
</tr>
<tr>
<td>YMR203W</td>
<td>1162</td>
</tr>
<tr>
<td>YMR204C</td>
<td>1464</td>
</tr>
<tr>
<td>YMR205W</td>
<td>978</td>
</tr>
<tr>
<td>YMR199W</td>
<td>973</td>
</tr>
<tr>
<td>YMR207C</td>
<td>1618</td>
</tr>
<tr>
<td>YMR208W</td>
<td>1374</td>
</tr>
<tr>
<td>YMR209C</td>
<td>1432</td>
</tr>
<tr>
<td>YMR210W</td>
<td>1068</td>
</tr>
<tr>
<td>YMR211W</td>
<td>1568</td>
</tr>
</tbody>
</table>

GeneSpring averages all of the measurements of YMR199W to get an average signal strength of 1286. GeneSpring notices the maximum signal 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 Gene-
Spring will plot when you choose to display error bars in either the graph or the scatter plot dis-
plays. After this average has been taken, GeneSpring discards any measurements between the end
points. Hence the measurements 1313 and 1218 will be automatically discarded.

1.10.3 Measurement Flags

Measurement flags are markers in your data set indicating whether or not any given measurement
is regarded as “Ok”, “Absent” or “Failed”. Data is assigned one of four flags.

Flags assigned when the experiment in entered into GeneSpring:

- Good Data: data is present and reliable. Marked with a “P” for passed or “O” for ok.
- Marginal Data: data is 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 will assign that measure-
  ment a “U”.

Only measurements at the “highest” available level of flag are combined and treated as replicates
in GeneSpring version 4.0. The order of flag precedence is P M U A. If one or more Ps are
present, only Ps are used, if not, and one or more Ms are present, then only Ms are used, etc. Sum-
mary statistics are collected over these cases and stored, with the corresponding flag. All other
flag data is discarded for the gene. This is done when the experiment is loaded into GeneSpring
and is not affected in any way by later user 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, which means that the
flags would not be available for other uses.

For information about measurement flags and how to load them into your experiment, please refer
to 4.4.17 “The Flags panel” on page 4-34 and 4.5.10 “Measurement Flags” on page 4-57.

1.11 Negative Signal Strengths

Some types of microarray technology report negative signal strengths. This is usually the result of
subtracting estimated background levels that are larger than the raw signal. This can happen in sit-
uations 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 signal strengths in such a
way that the 10th percentile of the signal is set equal to 0, before proceeding with the median nor-
malization. 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. You will get a warning message when you first load your data into GeneSpring if this background correction has been applied. Also, in the Gene Inspector raw signal strengths adjusted by this correction are flagged with asterisks.

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 preventing transformation problems.

1.12 Normalization for Particular Array Types

For Affymetrix or One-color experiments, you should normalize each sample to itself (as described in 1.5 “Normalize Each Sample to Itself” on page 1-6) and normalize to a single sample” (as described in 1.8 “Normalizing All Samples to Specific Samples” on page 1-11). Or, you can normalize each gene to itself (as described in 1.7 “Normalizing Each Gene to Itself” on page 1-8).

For Two-color experiments, normalize each gene to reference (as described in 1.3 “Normalize to Control Channel Values for Each Gene” on page 1-3). Then, normalize each sample to itself (as described in 1.5 “Normalize Each Sample to Itself” on page 1-6), that is not done by your scanner software.
Chapter 2  Experiment Parameters

The word parameter can apply to two slightly different concepts. There are parameters involving the samples in your experiment (also called parameter values). Sample parameters are a variable used to describe the precise condition under which each sample (or measurement) was taken. You may have many parameter values applying to a single sample (such as time, drug concentration etc.). There are also experiment parameters which can incorporate many sample parameters variables. Generally speaking, when the term parameter is used, it means an experimental parameter.

The four different ways of displaying the parameters (more information is in 2.2 “Parameter Display Options” on page 2-8) are as Continuous Elements, Non-Continuous Elements (set), Replicates, or Color Codes (color by parameter).

You will assign every parameter a default interpretation when you load its associated experiment into GeneSpring. This interpretation may be changed later within GeneSpring using the “Change Experiment Interpretation” window. This window also allows you to change the distribution of numbers on the Y-axis (and thereby the distribution of your relative intensity data) using one of four different formulas. For more information, see 2.3 “Changing the Parameter Interpretations” on page 2-31.

2.1 Definitions of Parameters

For clarity we will start with definitions of terms.

2.1.1 Data objects in the Navigator

![Figure 2-1 Data objects in the Navigator](image)

Figure 2-1 Data objects in the Navigator
• **Measurement:** The smallest unit of data used by GeneSpring, you will only see measurements as the raw values present in the upper right table in the Gene Inspector. In the Graph view this will be presented as one point on one gene’s line. (It may be easier to think of this as one spot or set of probes on one array.) A measurement is a number, such as 7.3. If you have no replicates, 1 measurement = 1 raw value = 1 spot on a chip.

• **Array:** a set of spots on a chip, typically expressed as a set of intensity measurements. An array typically has one sample on it. If you have gross slide problems, please see *GeneSpring User Manual* 3.4, “Array Layout View” for more information. If all of the interesting genes of the genome fit onto one array, then the terms array, chip and sample can be considered synonymous.

• **Sample:** The data generated from a biological object placed onto an array or set of arrays. A sample’s 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 2-1 the condition is Embryonic, day 21.

• **Interpretation:** A description of how GeneSpring displays the data for you to view. It would include a definition of applicable parameters and how the normalized numbers should be treated. This is the way a set of conditions are grouped. In Figure 2-1 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 2-1, the experiment is the NIH Rat Spinal Cord Study.

### 2.1.2 Parameter Values

• A **parameter-value** is 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”.

Even though you are intimately familiar with all of the different parameter-values involved in your experiment, sometimes it can be difficult to group those values into experiment parameters. To group your parameter-values into experimental parameters the first step is to classify your parameter-values into general groups such as time, disease type, or drug concentration. These groups will probably be your parameters. To double-check try asking yourself, “When I graph this experiment what values do I want on the x-axis?” The next thing you need to be able to do is assign one (and only one) parameter-value to every single sample in your experiment.
For example, given an experiment with the parameter-values breast cancer, kidney cancer, liver cancer, brain cancer, hepatitis A, hepatitis B, osteoporosis, arthritis, syphilis, and no disease, and given that only one of these parameters applied to any single sample in the experiment, you might be tempted to group the parameter-values in a number of different ways:

1. Type of disease: breast cancer, kidney cancer, liver cancer, brain cancer, hepatitis A, hepatitis B, osteoporosis, arthritis, and syphilis
2. Cancer: brain, breast, kidney, liver
3. Hepatitis: A, B
4. Sick: Yes, No
5. Infectious disease: Yes, No

Of these groups, only groups 4 and 5 could be used as experimental parameters in GeneSpring. Only parameters 4 and 5 can be applied to every sample. You must be able to set parameters relating to every sample in this experiment. The first three groups could be used as experimental parameters if they were expanded to include every possibility. You can add a category for healthy tissue to cover all possibilities.

1. Type of disease: breast cancer, kidney cancer, liver cancer, brain cancer, hepatitis A, hepatitis B, osteoporosis, arthritis, and syphilis
2. Type of Cancer: brain, breast, kidney, liver, and none
3. Type of Hepatitis: A, B, and none
   And/or:
4. Cancer: cancer, other disease, and no disease
5. Hepatitis: hepatitis, other disease, and no disease

A single sample can have many different parameter-values assigned to it, but it must have one, and only one, parameter-value assigned to it for each experimental parameter you are recording. If you find yourself trying to assign two different parameter-values for the same parameter to a single sample (from the example above, if one tissue sample had both osteoporosis and hepatitis A), then you have probably not defined your sample parameters sufficiently. It sometimes helps to create a table of your samples and their parameters, as shown in Figure 2-2. Also, a table like this can be pasted directly into the “Parameter Values” panel of the “New Experiment Wizard”.

As an example, your parameter names should be sufficiently specific that you can distinguish between the samples applied to different slides. For example, if the control and treated samples are applied to the each slide, and these pairs each represent samples taken from cells cultured from different durations, then “Days in culture” would be an appropriate parameter name, not “Treated”. The parameter, “Days in culture”, distinguishes samples applied to one slide from those applied to another.
Experiment Parameters

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Multiple Disease Example</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Sick (( ^{0} ))</td>
<td>no</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>3</td>
<td>Infectious Disease (( ^{0} ))</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>4</td>
<td>Type Hepatitis (( ^{0} ))</td>
<td>n</td>
<td>a</td>
<td>b</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>5</td>
<td>Disease (( ^{0} ))</td>
<td>no</td>
<td>hepatitis</td>
<td>hepatitis</td>
<td>syphilis</td>
<td>osteoporosis</td>
<td>arthritis</td>
</tr>
<tr>
<td>6</td>
<td>Time (minutes)</td>
<td>0</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>YAL001C</td>
<td>0.941667</td>
<td>0.675</td>
<td>0.95</td>
<td>0.925</td>
<td>1.166667</td>
<td>0.8</td>
</tr>
<tr>
<td>8</td>
<td>YAL002W</td>
<td>1.738318</td>
<td>0.971963</td>
<td>0.57093</td>
<td>0.635614</td>
<td>1.074766</td>
<td>0.654206</td>
</tr>
<tr>
<td>9</td>
<td>YAL003W</td>
<td>0.710966</td>
<td>0.69773</td>
<td>0.964663</td>
<td>0.679229</td>
<td>1.106667</td>
<td>0.476966</td>
</tr>
<tr>
<td>10</td>
<td>YAL004W</td>
<td>3.333333</td>
<td>2.360765</td>
<td>1.960784</td>
<td>1.019608</td>
<td>1.091968</td>
<td>0.740588</td>
</tr>
<tr>
<td>11</td>
<td>YAL005C</td>
<td>3.694501</td>
<td>3.775668</td>
<td>3.197556</td>
<td>1.274949</td>
<td>1.829921</td>
<td>0.729051</td>
</tr>
<tr>
<td>12</td>
<td>YAL007C</td>
<td>0.664667</td>
<td>0.921898</td>
<td>1.637416</td>
<td>0.962391</td>
<td>1.081987</td>
<td>0.819672</td>
</tr>
<tr>
<td>13</td>
<td>YAL009Y</td>
<td>2.355003</td>
<td>1.592006</td>
<td>2.047337</td>
<td>0</td>
<td>0.861667</td>
<td>0.566047</td>
</tr>
<tr>
<td>14</td>
<td>YAL009W</td>
<td>1.180334</td>
<td>0.495727</td>
<td>0.82906</td>
<td>0.565566</td>
<td>1.042735</td>
<td>0.547099</td>
</tr>
<tr>
<td>15</td>
<td>YAL010C</td>
<td>1.593443</td>
<td>0.95062</td>
<td>0.783526</td>
<td>0.813115</td>
<td>0.963934</td>
<td>0.544262</td>
</tr>
<tr>
<td>16</td>
<td>YAL011W</td>
<td>1.010453</td>
<td>0.780489</td>
<td>0.759682</td>
<td>0.634146</td>
<td>1.631011</td>
<td>0.801394</td>
</tr>
<tr>
<td>17</td>
<td>YAL012W</td>
<td>0.967189</td>
<td>0.844641</td>
<td>0.934379</td>
<td>0.955777</td>
<td>1.961489</td>
<td>0.848787</td>
</tr>
<tr>
<td>18</td>
<td>YAL013W</td>
<td>0.695652</td>
<td>0.956562</td>
<td>0.717391</td>
<td>0.52174</td>
<td>1.565252</td>
<td>0.634783</td>
</tr>
<tr>
<td>19</td>
<td>YAL014C</td>
<td>0.976744</td>
<td>1.067656</td>
<td>0.96322</td>
<td>0.726297</td>
<td>0.944544</td>
<td>0.637209</td>
</tr>
<tr>
<td>20</td>
<td>YAL015C</td>
<td>1.181943</td>
<td>1.212038</td>
<td>1.0342</td>
<td>1.070041</td>
<td>0.997688</td>
<td>0.746922</td>
</tr>
<tr>
<td>21</td>
<td>YAL016W</td>
<td>0.701092</td>
<td>0.945362</td>
<td>1.342701</td>
<td>5.038729</td>
<td>1.018668</td>
<td>0.865799</td>
</tr>
<tr>
<td>22</td>
<td>YAL017W</td>
<td>2.190476</td>
<td>0.44698</td>
<td>0.55102</td>
<td>0.673469</td>
<td>1.428571</td>
<td>0.829332</td>
</tr>
</tbody>
</table>

Figure 2-2 An example of a sample parameter table set up in an Excel® spreadsheet

This table shows some of the samples, parameters, and parameter-values in the fictitious Multiple Disease Example experiment and is used as an illustration for some of this chapter. As the file for this experiment is huge, only a fraction of it is shown here.

For another example, please see section 4.4.8 “The Parameter Values panel” on page 4-23. The accompanying Figure 4-13 shows 40 single samples, each with four parameters: kryptonite concentration (ppm), variety of yeast, Andromeda strain infection, and test repeat number. Each sample has the possibility of one of five kryptonite concentrations (given in ppm), with one of two varieties of yeast, either infected with the Andromeda strain or healthy, and the results of the sample were taken either from the first time the experiment was performed or the second.

You may order (or re-order) your parameter values from within GeneSpring, please refer to 2.2.5 “Ordering Parameter Values” on page 2-25.

2.1.3 A Note on Multiple Parameters

The more experimental parameters you have, the more options you have for visually querying your data. If you have samples of tissues infected with the different disease possibilities discussed earlier (breast cancer, kidney cancer, liver cancer, brain cancer, hepatitis A, hepatitis B, osteoporosis, arthritis, syphilis, and no disease) you might want to use all seven of the possible parameters discussed in 2.1 “Definitions of Parameters” on page 2-1 as the experimental 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.
The following illustrations are of a fictitious time series done on tissue samples, each infected with one of the disease possibilities listed above (see Figure 2-2). These pictures should illustrate the different visual queries you can set up by grouping the same set of experimental conditions into different parameters. There are four different parameters (not including time) used in these pictures, each one divides the eleven tissue samples into slightly different groups.

The parameters displayed are:

- “sick” with two parameter-values, yes and no.
- “infectious disease” with two parameter-values yes and no.
- “hepatitis” with three parameter-values, hepatitis A, hepatitis B, and no disease.
- “disease” with ten parameter-values, breast cancer, kidney cancer, liver cancer, brain cancer, hepatitis A, hepatitis B, osteoporosis, arthritis, syphilis, and no disease.

The following descriptions use the terms “continuous”, “non-continuous”, “replicate” and “color code” to refer to specific parameter display options. See 2.2 “Parameter Display Options” on page 2-8 for definitions of these terms.

Figure 2-3 The six parameter-values of the parameter “disease”
- Figure 2-3 shows the six parameter-values of the parameter “disease” defined as color codes, and the parameter “infectious disease” defined as non-continuous. This means two graphs are displayed side by side, the right displaying the gene lines from the tissues infected with infectious diseases, and the left displaying the gene lines of those tissues not infected by an infectious disease. Displaying these parameters in this way allows you to visually ask the questions:
  - Do infectious diseases have a similar expression profiles?
  - Is there an obvious difference between expression profiles of infectious and non-infectious diseases?
  - Do a subset of infectious or non-infectious diseases have similar expression profiles?

![GeneSpring 4.0.0 Yeast Genes: all genomic elements]

**Figure 2-4 Expression profiles of varieties of hepatitis**

- Figure 2-4 attempts to clarify any apparent similarities between the expression profiles of infectious diseases in Figure 2-3, because these similarities may be due to the fact that two of the infectious diseases studied are varieties of hepatitis. To analyze this possibility, the param-
eter “type hepatitis” was defined as non-continuous along with the parameter “infectious disease”. This creates several distinct graphs. Reading the Genome Browser from left to right:

- “n n” is a graph of non-infectious diseases and healthy tissue.
- “y a” is tissue infected with hepatitis A.
- “y b” is tissue infected with hepatitis B.
- “y n” is tissue infected with syphilis, an infectious disease.

![GeneSpring 4.0.0 Yeast Genes: all genomic elements](image)

**Figure 2-5** The parameter “sick” as a color code

- Figure 2-5 attempts to answer a slightly different question by displaying the parameter-values of the parameter “disease” as non-continuous (sets), while using the parameter “sick” as a color code. Using “sick” as a color code, causes the healthy tissue to be differently colored from all of the rest, which allows you to see if there are any major differences between the expression profiles of healthy and diseased tissues.
2.2 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 enter a new experiment in the Experiment Wizard, you will be asked which display option is most appropriate for each of your parameters. Your chosen display option will become the default display for that parameter. If you simply paste in a new experiment, all the parameters will be assigned the continuous display option. Regardless of how a parameter is entered in GeneSpring, you can change how each parameter is displayed within GeneSpring using the Experiment > Change Experiment Interpretation command. For more details on this, see 2.3 “Changing the Parameter Interpretations” on page 2-31.

2.2.1 Replicate or Hidden Element

A parameter defined as a replicate is graphically a hidden variable; no visual distinction is made based upon this parameter or its parameter-values. In fact, after the samples in the experiment are assigned parameter values, any set of those samples which have been assigned the exact same parameter-values, ignoring the value(s) given for the parameter(s) defined as a replicate, are averaged together. Defining a parameter as a replicate is the easiest way to deal with repeated samples inside GeneSpring. You can also get further insight into your data. In the fictitious Multiple Disease Example experiment, set parameter Patient# as a replicate within each “Tissue Type” group and you will see a pattern of change from tissue to tissue rather than change from patient to patient.

For example, in the Yeast Extraterrestrial Studies experiment (about one-fourth of the Excel®-formatted data table is illustrated in Figure 2-6), if the parameter “test repeat number” is defined as a replicate, then the data from sample 1 (column B) is averaged with the data from sample 21 (column V), the data from sample 2 (column C) is averaged with sample 22 (column W) and so on, for each of the twenty repeated samples. These averages will be displayed in GeneSpring as the signal strength. In the Graph or Scatter Plot displays in the Genome Browser it is possible to show the minimum and maximum of the parameter-values averaged together by showing error bars on the graphs. See Figure 2-7 for an illustration of what this looks like.
The equation used for averaging repeated samples is exactly the same one used to average repeated measurements in a raw data file. See 1.10 “Dealing with Repeated Measurements” on page 1-17 for more information. The only difference is the averaging done to repeated parameters is done after the raw data has been normalized.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast Extraterrestrial Studies</td>
<td>0.941667</td>
<td>0.575</td>
<td>0.95</td>
<td>0.925</td>
<td>1.166667</td>
<td>0.8</td>
</tr>
<tr>
<td>Kryptonite concentration (ppm)</td>
<td>1.738318</td>
<td>0.971963</td>
<td>0.57093</td>
<td>0.635514</td>
<td>1.074766</td>
<td>0.654206</td>
</tr>
<tr>
<td>Variety of yeast (*)</td>
<td>0.710966</td>
<td>0.68773</td>
<td>0.964863</td>
<td>0.679229</td>
<td>1.105696</td>
<td>0.476906</td>
</tr>
<tr>
<td>Andromeda strain infection (*)</td>
<td>3.333333</td>
<td>2.360785</td>
<td>1.960784</td>
<td>1.019608</td>
<td>1.090196</td>
<td>0.470508</td>
</tr>
<tr>
<td>Test repeat number ()</td>
<td>3.694501</td>
<td>3.775966</td>
<td>3.197556</td>
<td>1.274949</td>
<td>1.828921</td>
<td>0.725051</td>
</tr>
<tr>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
</tr>
<tr>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
</tr>
<tr>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
</tr>
<tr>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
</tr>
<tr>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
</tr>
<tr>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
<td>Healthy</td>
</tr>
</tbody>
</table>

**Figure 2-6 Yeast Extraterrestrial Studies Master Gene Table**
Figure 2-7 Yeast Extraterrestrial Studies experiment, colored by test repeat

- Figure 2-7 displays the Yeast Extraterrestrial Studies experiment with “test repeat number” defined as a color code. The error bars associated with making “yeast type” and “andromeda strain infection” replicates are also showing (to turn the error bars on or off right-click in the black genome browser and select your choice from the Error Bars submenu, please refer to Advanced Analysis Techniques Manual Chapter 8, “Global Error Models”). The gene YDL127W (PCL2) is selected, which causes its gene line to be highlighted. The error bars associated with YDL127W (PCL2), up to the minimum and maximum values averaged together to create YDL127W (PCL2)’s gene line, are also highlighted.
2.2.2 Continuous Element

A continuous variable is when each parameter-value of the experimental parameter exists on a continuum with the other parameter-values in that experimental parameter, rather than as discrete points. More loosely, each parameter-value is related to the parameter-values on either side of it and it is sensible to draw lines between them. Typically, continuous variables are numeric. This requires the parameter-values be in a particular order. GeneSpring will automatically order numerical parameters from highest to lowest, and order non-numerical parameters in alphabetical order. When graphing by a continuous parameter each parameter-value is placed on the x-axis, in order, from left to right. You can manually change this order, please refer to 2.2.5 “Ordering Parameter Values” on page 2-25 for more details.

When displaying data from a continuous parameter, adjacent data points are connected together by lines instead of being graphed as discrete points. While time series are common examples of a continuous element, they are by no means the only possible continuous parameter. Figure 2-8 is an illustration of what a graph with one continuous parameter looks like. The continuous parameter illustrated there is a drug concentration (kryptonite), not a time series.

2.2.2.1 How a continuous parameter display combines with other parameter displays

It may be helpful to see how the continuous parameter display combines with the other parameter display options. To show this you will analyze the four pictures below. They are all graphs of the fictitious Yeast Extraterrestrial Studies experiment, and they are all colored by parameter, even when there is no parameter defined as a color. When there is no parameter defined as a color code and the graph display is colored by parameter, GeneSpring colors every gene the same color. The Yeast Extraterrestrial Studies experiment has four different parameters: kryptonite concentration (given in ppm), variety of yeast, Andromeda strain infection and test repeat number. There were five parameter-values associated with kryptonite concentration: 0, 10, 20, 30, and 40 ppm. There were two possible parameter-values for infection, either the yeast cell was infected with the Andromeda strain (labeled “Andromeda strain” when graphed) or it was not (labeled “healthy”). This was done to two varieties of yeast, named A and B. All of these tests were repeated. Example parameter-values for the first seven samples in the experiment are given in Figure 2-6.
Figure 2-8 shows the parameter “kryptonite concentration” graphed as a continuous parameter (in reverse order from 40 ppm to 0 ppm, please refer to *GeneSpring User Manual 5.3.5*, “Reorder the Parameters”), and the other three parameters, “variety of yeast”, “Andromeda strain infection”, and “test repeat number” all considered replicates. As a parameter is defined as a continuous element, the genes are drawn as lines rather than discrete points. In this illustration a single gene is selected, which highlights a single gene line.
Figure 2-9 shows the parameter “kryptonite concentration” graphed as a continuous parameter, and the parameter “variety of yeast” graphed as a color code. The two parameters “Andromeda strain infection” and “test repeat number” are defined as replicates. When there is a parameter defined as a color, and one defined as continuous, there are multiple lines for each gene. There is one line per gene drawn per parameter variable defined as a color. In this example there are two gene lines drawn because the parameter “variety of yeast” has only two parameter-values. The representation of each gene is graphed as a continuous line from 40 to 0 ppms because “kryptonite concentration” is a continuous parameter. In this illustration, when a gene is selected there are two gene lines highlighted.
Figure 2-10 shows the parameter “kryptonite concentration” graphed as a continuous parameter, and the parameter “Andromeda strain infection” graphed as a non-continuous element. Both of the parameters “test repeat number” and “variety of yeast” are defined as replicates. As “Andromeda strain infection” is a parameter defined as non-continuous there are two distinct graphs, one for Andromeda-infected tissue and one for healthy tissue. Each graph repeats the same set of continuous points. These points are defined by the continuous element’s parameter-values (“kryptonite concentration”). The selected gene in this illustration is visible once in each graph.
Figure 2-11 shows the parameter “kryptonite concentration” graphed as a continuous parameter, the parameter “variety of yeast” graphed as a color, and “Andromeda strain infection” graphed as a non-continuous element. The parameter “test repeat number” is still considered a replicate. As the parameter “Andromeda strain infection” is graphed as a non-continuous element, there are two distinct graphs, both with “kryptonite concentration” as a continuous parameter. As “variety of yeast” is defined as a color, each of those two graphs have two gene lines for each gene, therefore when you select a gene, four gene lines are highlighted (two lines, each containing a break).
2.2.3 Non-Continuous Element (Set)

A non-continuous (or set) variable is when each parameter-value of the experimental parameter exists independent of each other, as discrete points. When a non-continuous element is graphed, each parameter-value is placed on the x-axis, in order, from left to right. GeneSpring will automatically order numerical parameters from highest to lowest, and order non-numerical parameters in alphabetical order. See 2.2.5 “Ordering Parameter Values” on page 2-25 if you wish non-numerical parameter-values to be graphed in a particular non-alphabetical order. When graphing by a non-continuous parameter each parameter value is placed on the x-axis, in order, from left to right.

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. In the Yeast Extraterrestrial Studies experiment the variety of yeast the experiment was done upon is defined as a non-continuous parameter, because it is meaningless to draw lines between them. 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 of these discrete graphs contain all of the parameter-values of the continuous parameter, making each of the separate graphs look like a set of parameter-values.

2.2.3.1 How a non-continuous parameter display combines with other parameter displays

It may be helpful to see how a non-continuous parameter display combines with the other parameter display options. The four pictures below are all graphs of the fictitious Yeast Extraterrestrial Studies experiment, and they are all colored by parameter, even when there is no parameter defined as a color code. (When there is no parameter defined by a color and the graph display is colored by parameter, GeneSpring colors every gene the same color, indicating they are all the results of samples under the same conditions.) The Yeast Extraterrestrial Studies experiment had four different parameters; kryptonite concentration (given in ppm), variety of yeast, Andromeda strain infection and test repeat number. There were five parameter-values associated with kryptonite concentration: 0, 10, 20, 30, and 40 ppm. There were two possible parameter-values for infection, either the yeast cell was infected with the Andromeda strain (labeled “Andromeda strain” when graphed) or it was not (labeled “healthy”). This was done to two varieties of yeast, named A and B, and all of these tests were repeated. The parameter-values for each single sample in the experiment were given in Figure 2-2.
Figure 2-12 Display using only a non-continuous element

- Figure 2-12 shows the parameter “Andromeda strain infection” graphed as a non-continuous parameter and the other three parameters, “kryptonite concentration”, “test repeat number”, and “variety of yeast” considered as replicates. As a parameter is defined as a non-continuous element, the genes are drawn as discrete points in a histogram. In this illustration a single gene, YHR136C (SPL2), is selected, which highlights two gene points, one in each histogram. There are two histograms because there are two parameter-values for “Andromeda strain infection”, Andromeda and Healthy.
Figure 2-13 shows the parameter “Andromeda strain infection” graphed as a non-continuous parameter, and the parameter “variety of yeast” graphed as a color. Two parameters, “kryptonite concentration” and “test repeat number”, are defined as replicates. When there is a parameter defined as a color, and one defined as non-continuous, there are multiple points for each gene in each histogram; for each parameter-value defined as a color there is one gene point drawn in each of the graphs defined by the parameter displayed as a set. In this example, there are two gene points drawn in each histogram because the parameter “variety of yeast” has only two parameter-values, and there are two histograms because the parameter “Andromeda strain infection” has two parameter-values.
Figure 2-14 shows the parameter “Andromeda strain infection” graphed as a non-continuous element, the parameter “kryptonite concentration” graphed as a continuous element, and “variety of yeast” graphed as a color. The parameter “test repeat number” is still considered a replicate. As the parameter “Andromeda strain infection” is graphed as non-continuous, there are two distinct graphs, both with “kryptonite concentration” as a continuous parameter. As “variety of yeast” is defined as a color each of those two graphs have two gene lines for each gene, therefore when the gene YHR136C (SPL2) is selected, four gene lines are highlighted.
2.2.4 Color Code

A color code is used for experimental parameters whose parameter-values exist independent of one another, but not unrelated to one another. When a color code is graphed, separate gene points or gene lines are drawn; one per each gene per each parameter-value defined by a color. When the browser display is colored by parameter (look under the “Colorbar” drop-down menu for options) GeneSpring will order the parameter values from top to bottom in the color bar. The parameter-value names are to the right of their color. Parameter-values are listed in alphabetic or numerical order.

A color is a category (or set of categories) used to color code the display. When coloring the browser display by parameter, each parameter-value defined as a condition is assigned a color and every measurement described by that parameter is drawn in that parameter’s color. This can be referred to as “Color by Parameter” or “Color Code”. Using this parameter display option means the browser display shows the same gene multiple times; the number of times a single gene is drawn is equal to the number of parameter-values defined as conditions. When the browser display is colored using a color option other than Color Code, it is impossible to visually distinguish which parameter-value a particular gene line or gene point represents, although separate gene lines for each parameter-value defined as a condition are still drawn. Individual patients, or strain types, are variables commonly defined as color codes (conditions) because, although they are different parameter-values, it is interesting to see them directly, visually compared to one another. It is likely the expression patterns of individual patients with the same disease are going to react in a similar way under similar conditions, often it is when the expression patterns are not similar that the results are interesting; hence the graphs of parameter-values defined as conditions are superimposed upon one another. In the Yeast Extraterrestrial Studies experiment the variable “variety of yeast” is defined as a color code in Figure 2-15.

2.2.4.1 How a color coded parameter display combines with other parameter displays

It may be helpful to see how the color code (condition) parameter display combines with the other parameter display options. The four pictures below will demonstrate the possible combinations. They are all graphs of the fictitious Yeast Extraterrestrial Studies experiment, and are all colored by parameter. The Yeast Extraterrestrial Studies experiment had four different parameters; kryptonite concentration (given in ppm), variety of yeast, Andromeda strain infection and test repeat number. There were five parameter-values associated with kryptonite concentration: 0, 10, 20, 30, and 40 ppm. The experiment was done to two varieties of yeast, named A and B. There were two possible parameter-values for infection, either the yeast cell was infected with the Andromeda strain (labeled “Andromeda strain” when graphed) or it was not (labeled “Healthy”). All of these tests were repeated. The parameter-values for each single sample in the experiment were given in Figure 2-2.
Figure 2-15 Display using only a color code

- Figure 2-15 shows the parameter “variety of yeast” graphed as a color code, and the other three parameters, “Kryptonite Concentration”, “Andromeda strain infection”, and “test repeat number” graphed as replicates. As only one parameter is visible, and it is defined as color code, the genes are drawn as discrete points in a histogram. The number of gene points drawn per gene in that histogram is the number of parameter-values defined as conditions. In this illustration a single gene, YEL022W (GEA2), is selected, this highlights two gene points, because there are two parameter-values (A and B) in the parameter “variety of yeast”. This display in the Genome Browser has been magnified slightly.
Figure 2-16 shows the parameter “variety of yeast” graphed as a color code (also known as a condition), and the parameter “kryptonite concentration” graphed as a continuous parameter. The two parameters “Andromeda strain infection” and “test repeat number” are defined as replicates. When there is a parameter defined as a color code, and one defined as continuous, there are multiple lines for each gene; there is one line per gene drawn per parameter variable defined as a condition. In this example, there are two gene lines drawn because the parameter “variety of yeast” has only two parameter-values. The representation of each gene is graphed as a continuous line from 40 to 0 ppm because “kryptonite concentration” is a continuous parameter. In this illustration, where YEL022W (GEA2) is selected there are two gene lines highlighted.
Figure 2-17 Display using both a color code and non-continuous element

- Figure 2-17 shows the parameter “variety of yeast” graphed as a color code, and the parameter “Andromeda strain infection” graphed as a non-continuous parameter. The two parameters “kryptonite concentration” and “test repeat number” are defined as replicates. When there is a parameter defined as a color code, and one defined as a non-continuous element, there are multiple points for each gene in each histogram; for each parameter-value defined as a condition there is one gene point drawn in each of the histograms defined by the parameter displayed as a set. In this example there are two gene points drawn in each histogram because the parameter “variety of yeast” has only two parameter-values, and there are two sets because the parameter “variety of yeast” has two parameter-values. In this illustration, YEL022W (GEA2) is selected, and there are four gene points highlighted.
Figure 2-18 shows the parameter “variety of yeast” graphed as a color code, the parameter “Kryptonite concentration” graphed as a continuous element, and the parameter “Andromeda strain infection” graphed as a non-continuous element. The parameter “test repeat number” is still defined as a replicate. The parameter “Andromeda strain infection” is graphed as a non-continuous element, resulting in two distinct graphs with “Kryptonite concentration” as a continuous parameter. As “variety of yeast” is defined as a color code, each of those two graphs have two gene lines for each gene, selecting a gene, such as LDL003W (MCD1), causes four gene lines to be highlighted.
2.2.5 Ordering Parameter Values

GeneSpring will automatically order numerical parameters from lowest to highest, and order non-numerical parameters in alphabetical order.

You can change the order in which non-numeric parameters are viewed by selecting **Experiments > Change Experiment Parameters**.

![Figure 2-19 The Change Parameters panel](image)

You may only re-order non-numeric parameters However, you can tell genespring that a parameter whose values are numbers is non-numeric and re-order them. Select a column or a section of a column and click the “Set Value Order” button.
Figure 2-20 The Parameter Value Order window

You can select a single cell of those offered and move it up or down to change the order. Or, if you would prefer to have GeneSpring automatically sort the parameter values in the reverse order, you can highlight all the cells and click “Sort Descending”.

Figure 2-21 The Kryptonite Concentration parameter, sorted descendingly

For more details, please refer to GeneSpring User Manual 5.3.5, “Re-order the Parameters”.
2.2.6 Why change the parameter display?

Changing the parameter displays gives you many different visualization options. For example, given the four parameters in the Yeast Extraterrestrial Studies experiment illustrating this chapter, there are 256 different combinations of parameter visualizations. Some of these different ways of looking at your data may elucidate facts not obvious in other display options. Some of the 256-parameter display options (for the Yeast Extraterrestrial Studies experiment) are illustrated below, along with brief notes about the possible things you might learn from looking at these parameters in this way:

- Figure 2-22 Kryptonite concentration graphed as a continuous parameter, the like YHR143W-A list showing

- Figure 2-22 is the default graph view for the fictitious Yeast Extraterrestrial Studies experiment. The parameter “kryptonite concentration” is graphed as a continuous parameter (from 40 ppm to 0 ppm). The parameter “Andromeda strain infection” is graphed as a non-continuous parameter (data is either infected with the Andromeda strain or not). The parameter “variety of yeast” is graphed as a color code (and the display is therefore colored by only those two conditions). Lastly, the parameter “test repeat number” is not shown at all because it is considered a replicate parameter.
Figure 2-23 Andromeda strain infection graphed as a color code

- Figure 2-23 leaves the parameter “kryptonite concentration” as a continuous element and “test number” as a replicate element. The display of “Andromeda strain infection” is changed to a color code and the parameter “variety of yeast” is redefined as a replicate parameter, and is thus not shown at all. Displaying the parameters in this way allows you to easily see what being infected with the Andromeda strain did to the yeast cells’ gene expression.
Figure 2-24 Yeast variety A vs. yeast variety B

- Figure 2-24 leaves “kryptonite concentration” as a continuous element, “test number” as a replicate, and displays both “variety of yeast” and “Andromeda strain infection” as non-continuous element. This allows you to see distinct groups of genes with similar expression profiles based upon the combinations of the two parameters defined as sets, i.e. how yeast variety A reacted to being infected with the Andromeda strain compared to how it behaved when healthy, compared to how yeast variety B reacted to the Andromeda strain infection.
Figure 2-25 leaves “kryptonite concentration” as a continuous element, and considers both of the parameters “variety of yeast” and “Andromeda strain infection” as replicates. Coloring by the condition “test repeat number” shows the first and the second time the experiment was repeated the results were indeed very similar to each other.
2.3 Changing the Parameter Interpretations

In each experiment you can change any of the parameter interpretations from within GeneSpring. You defined the default display of each parameter when you set up the experiment. To change the parameter interpretations, use the “Change Experiment Interpretation” window. This window also allows you to change the way the y-axis is plotted.

Select **Experiment > Change Experiment Interpretation**. The “Change Experiment Interpretation” window appears.

![Change Interpretation for Yeast Extraterrestrial Studies](image)

The “Change Experiment Interpretation” window allows you to change the display option of your parameters. It is recommended that you leave the default display alone, and when you change the parameters to rename them and use the “Save As New” button. This way you will not change the default display and the next time you open GeneSpring all of the experiments’ parameters will still be displayed using their default display options.

When you “Save As New” the saved interpretations will be visible in the Navigator, please refer to Figure 2-28.

The “Change Experiment Interpretation” window lists the name of each of your parameters in separate boxes. (See Figure 2-26.) Inside each of these boxes are the four possible parameter interpretations. The number of parameters available to be changed will vary with how many parameters there are in the experiment. Compare Figure 2-26 with Figure 2-27.
To change the way a parameter is used in displays:

1. Move the cursor to the parameter box labeled with the name of the parameter whose interpretation you wish to alter.

2. Click in the circle next to the interpretation you want to use for this parameter’s display. A parameter can be displayed using only one of the four parameter display options at a time.

3. Give the interpretation a meaningful name and click “Save”.

To look at a previously saved interpretation:

1. Select the Experiments folder in the Navigator.

2. Select your current experiment to view all interpretations.

3. Select the correct interpretation from the saved interpretations (the icons with small light bulbs represent interpretations).
Figure 2-28 Interpretations in the Navigator

Typically the interpretation being used in the current view is selected along with the current experiment. For more details, please refer to *GeneSpring User Manual* Chapter 5, “Changing Your Experiment”.
Chapter 3  Installing a New Genome

In GeneSpring, a genome represents a set of genes upon which experiments are done, and contains information about all the genes in your chip or microarray setup. Generally, all the experiments and knowledge about a genome are stored in a separate directory. A GeneSpring genome need not correspond exactly to a real genome, although it usually does. For instance, a genome may contain three alleles (considered to be separate genes), whereas a particular organism would never have all three simultaneously. You can install a new genome in one of several ways: by loading a new experiment in with the Autoloader, by using the “New Genome Installation Wizard” or by creating a “.genomedef” file by hand.

3.1 The Autoloader

Loading new genomes can be a by-product of the new Autoloader feature in GeneSpring version 4.0. Please refer to 4.2 “Autoload Experiment” on page 4-2 for more details on loading experiments.

At the end of the experiment loading process GeneSpring will check to see if you have a matching Genome available. If GeneSpring cannot find a genome that matches your experiment, it will ask you to identify it.

![Select Genome box](image)

**Figure 3-1 The Select Genome box**

If you select a genome GeneSpring does not already have in memory, it will offer to create it for you. Give your new genome a name and click “Use Selected Genome”. You may get a few more screen asking if you have more data or would like the genome extended.

With GeneSpring version 4.0 and later, the only times you will need to load a Genome in the old way is when you have many subchips and the autoloader function is having difficulty processing them all into a genome.
To update genomes once they are loaded, please refer to 3.7 “Updating Your Information” on page 3-34.

3.2 Creating Folders for New Genomes

Normally, GeneSpring will create new folders for you when you use the Genome Wizard. See 3.4 “Genome Wizard” on page 3-11 for more details.

To manually create a new folder in the Genome Browser, you must go through a file management system, such as Windows Explorer®. For example, a new folder named “Mouse” has been created and placed into the data directory of GeneSpring.

Before your new Mouse folder will appear in GeneSpring Navigator you will need to create a correct “mouse.genomedef” file. A .genomedef file will contain all the information GeneSpring needs to create a folder and other data objects. Make sure you save the “.genomedef” file in the correct folder (the “Mouse” folder) after you create it. Please see 3.5.2 “The .genomedef File” on page 3-24 for details on creation.

3.3 Raw Data

3.3.1 What Data Are Necessary?

You must have a list of distinct names for all the genes you intend to work with. In addition, a genome may also have GenBank loci, sequences, alternative names, functional information, map positions, EC numbers, and so on, associated with genes. It may also include links to web-based databases. Each genome should have a distinct name, to reduce confusion.
3.3.2 What Format do these Data Need to be in?

You will generally need either a Master Gene Table or a GenBank/EMBL entry for your organism. If you use a Table of Genes containing the genes’ GenBank accession numbers, then the GenBank information associated with each gene can be automatically updated. See 3.7.3 “Updating GenBank Information using the GeneSpider” on page 3-36 for how to do this.

There are four possible formats for a Master Gene Table: “name list”, “name function”, “SGD”, and “Mapped”. The reason these formats are called “Master Gene Table” is because it is easiest to create them in spreadsheet programs, such as Microsoft Excel®, and then use the Save As command to create tab-delimited text files. Occasionally a Master Gene Table is referred to as the Table of Genes, the Master Gene List or the Array Element List.

3.3.2.1 Name List

The simplest format for a Master Gene Table is “name list”. In this format the Master Gene Table is a single column comprised of the names of the genes:

Gene1
Gene2
Gene3

Gene names with spaces in them, such as “Gene 1” are acceptable.

3.3.2.2 Name Function

The next simplest format for the Master Gene Table is “name function”. In this format the table of genes is the same as the table for “name list” except each gene may be followed by a description of its function. If you have additional information about the genes, enter it in the same row as the gene it refers to, separated from the gene name by a tab character or column separator in Microsoft Excel®. An example of this is:

Gene1  Putative Phosphokinase
Gene2
Gene3  Deletion causes 2 tails

You do not need to have information about every gene. In the example, nothing is known about Gene2, so the line after its name is left blank. If you have a list of genes and text information about them in a spreadsheet formatted as two columns with one row per gene, simply save this file as a tab-delimited text file.

3.3.2.3 SGD

A third Master Gene Table format is “SGD”. This is the format used for the list of genes in the Saccharomyces Genome Database (SGD), and is generally only relevant for yeast. As yeast comes pre-loaded in GeneSpring, details about this format are unnecessary.

3.3.2.4 Mapped

The fourth and most sophisticated Master Gene Table format is “Mapped”. Again, this format has one line per gene, with several fields separated by tabs. The first field (systematic name) must
be present; all other fields are optional. The fields are described below. When creating your Master Gene Table, these fields should be entered in the order listed here.

1. **Systematic Name**: The normal way of referring to this gene. This name must be unique. The name entered in this field can be utilized by the “Find Gene” command to find this particular gene within GeneSpring. It is recommend that the name used as the gene’s systematic name be the name which labels that gene’s raw control strength values in your experiment data files. Any of this information can be accessed when you use the “Find Gene” command.

2. **Common Name**: An alternative way of referring to this gene. The name entered in this field can be utilized by the “Find Gene” command to find this particular gene within GeneSpring. Genes are not required to have a common name, and common names do not have to be unique, although duplicated common names may lead to confusion if the common name is how the gene is referred to in the experiment files. This information can be accessed when you use the “Find Gene” command.

3. **Map**: Mapping information for this gene. Sequence position, for example, a first chromosome gene would be 1:228836..229309 inclusive, see Table 3-3 on page 3-5 for an example of mapped file with third column being the chromosome:base position. For an example of the mapped Cytogenetic position (such as 16q12.1) see Table 3-4 on page 3-6.

4. **EC number**: The EC number for this gene, if known.

5. **Description**: A description of this gene, if known. This information can be accessed when you use the “Find Gene” command.

6. **Product**: The protein product coded for by this gene, if known. This information can be accessed when you use the “Find Gene” command.

7. **Phenotype**: A description of the phenotype for this gene, if known.

8. **Function**: A description of the function of this gene product, if known.

9. **Keywords**: Keywords associated with this gene, if known. Separate keywords with semicolons. This information can be accessed when you use the “Find Gene” command.

10. **GenBank locus**: The GenBank identifier for this gene, if known. If the GenBank identifiers for your genes were not used as either their systematic or common names, then including the GenBank identifier in this field allows you to update the information about this particular gene directly from GenBank. See 3.7.3 “Updating GenBank Information using the GeneSpider” on page 3-36 for more information.

11. **Synonym**: This column allows for other names to be entered for the genes. Multiple names should be separated by semicolons (;).

12. **Sequence**: The sequence data, if known.

13. **PM**: The Public Medline accession number, if known. Multiple identifiers should be separated by semicolons (;).
14. **custom1**: Not specified. This column will not be interpreted by GeneSpring, but it is useful for some reports.

15. **custom2**: Not specified. This column will not be interpreted by GeneSpring, but it is useful for some reports.

16. **custom3**: Not specified. This column will not be interpreted by GeneSpring, but it is useful for some reports.

17. **Type**: A specific field returned by the GeneSpider.

18. **Database reference (also called DBid)**: A specific field returned by the GeneSpider.

The Mapped format allows you to link up to three different names (plus three more custom names) for the same gene. Using this method, you could query one gene using any of the data in the corresponding columns #A Systematic Name, #B Common Name, and #F Product. You can also describe genes in your overlay or do a search for a gene named in column #2 Common Name and find the corresponding accession number.

Column names are only for this example. You should not include them in your file.

![Figure 3-3 Example of what a sequenced mapped format looks like in Excel®](image)
<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>EC Number</th>
<th>Phenotype Function</th>
<th>Keywords</th>
<th>GenBank locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>luck1 (1.1.1.1)</td>
<td></td>
<td>this gene somehow causes rats to be very lucky</td>
<td>protein A</td>
<td>g764509</td>
</tr>
<tr>
<td>gene 2</td>
<td>16q21.2 (1.1.1.2)</td>
<td></td>
<td>co-produces protein B</td>
<td></td>
<td>g587439</td>
</tr>
<tr>
<td>gene 3</td>
<td>16q21.3 (1.1.1.3)</td>
<td></td>
<td>deletion causes immortal cell</td>
<td></td>
<td>g903265</td>
</tr>
<tr>
<td>gene 4</td>
<td>16q21.4 (1.1.1.4)</td>
<td></td>
<td>deletion causes immortal cell</td>
<td></td>
<td>g460349</td>
</tr>
<tr>
<td>gene 5</td>
<td>16q21.5 (1.1.1.5)</td>
<td></td>
<td>deletion causes immortal cell</td>
<td></td>
<td>g932590</td>
</tr>
<tr>
<td>gene 6</td>
<td>charm (16q42.3)</td>
<td></td>
<td>rats with gene are very cute</td>
<td>protein C</td>
<td>g238566</td>
</tr>
<tr>
<td>gene 7</td>
<td>9q11.1 (1.1.1.7)</td>
<td></td>
<td>protein D possibly involved in metabolism</td>
<td></td>
<td>g234876</td>
</tr>
<tr>
<td>gene 8</td>
<td>beauty (16q14.1)</td>
<td></td>
<td>involved in metabolism</td>
<td></td>
<td>g2939857</td>
</tr>
<tr>
<td>gene 9</td>
<td>16q76.7 (1.1.2.1)</td>
<td></td>
<td>protein D possibly involved in metabolism</td>
<td></td>
<td>g2998456</td>
</tr>
<tr>
<td>gene 10</td>
<td>16q76.8 (1.1.2.2)</td>
<td></td>
<td>deletion causes death</td>
<td></td>
<td>g2900309</td>
</tr>
<tr>
<td>gene 11</td>
<td>16q76.9 (1.1.2.3)</td>
<td></td>
<td>deletion causes death</td>
<td></td>
<td>g321197</td>
</tr>
<tr>
<td>gene 12</td>
<td>16q76.10 (1.1.2.4)</td>
<td></td>
<td>deletion causes death</td>
<td></td>
<td>g357848</td>
</tr>
<tr>
<td>gene 13</td>
<td>16q76.11 (1.1.2.5)</td>
<td></td>
<td>deletion causes death</td>
<td></td>
<td>g204968</td>
</tr>
<tr>
<td>gene 14</td>
<td>16q76.12 (1.1.2.6)</td>
<td></td>
<td>deletion causes death</td>
<td></td>
<td>g230468</td>
</tr>
<tr>
<td>gene 15</td>
<td>16q76.13 (1.1.3.1)</td>
<td></td>
<td>deletion causes death</td>
<td></td>
<td>g324166</td>
</tr>
</tbody>
</table>

**Figure 3-4** Example of what an unsequenced mapped format looks like in Excel®

**Figure 3-5** Example of the same mapped format shown in Figure 3-4 saved as a tab-delimited text file with text marks showing
A note about Figure 3-3, Figure 3-4 and Figure 3-5: you should not include the titles of the fields in your table of genes. The titles are included here only for clarity. Remember, when you are using the “mapped” format, you must include any blank fields in their appropriate columns. The gene’s systematic name should always be in the first column, its common name in the second, and its mapping information in the third column, and so on, even if the second column is completely blank because there are no common names for any of your genes.

### 3.3.2.5 GenBank or EMBL Files

If you use a single GenBank file to describe the genome, you do not have to use a Master Gene Table and therefore do not have to enter any of the information discussed in 3.3.2 “What Format do these Data Need to be in?” on page 3-3. Nor do you need a separate file to contain the sequence data (the files for sequence data are described in 3.3.2.7 “Sequence Data” on page 3-10). 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:


Generally this URL is 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.

### 3.3.2.6 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 in some extra genes. This is typically done to represent a strain slightly different from the sequenced strain. To do this you need to create a separate Master Gene Table containing all of the extra genes you wish to add. This file should be formatted using one of the four table of genes formats discussed in 3.3.2 “What Format do these Data Need to be in?” on page 3-3.

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.

To update GenBank information:

1. In GeneSpring, open the genome you wish to update.
   a. Go to File > New Genome or Array. Another menu appears. The genomes included in this submenu depend on what genomes have been loaded into your copy of GeneSpring.
b. Select the name of the genome you wish to update.

2. Go to **Tools > GeneSpider > Update genes from GenBank.**
3. Click the arrow to the right of the box labeled “Use for GenBank identifier”. A drop-down menu will appear.

   ![Figure 3-7 The “Update for GenBank identifier” box](image)

   **Figure 3-7 The “Update for GenBank identifier” box**

4. Click the name of the column in the table of genes containing the GenBank accession numbers.

5. Click the “Start” button. The GeneSpider will process GenBank’s data, displaying how far it has gotten in the box labeled “Status”.
   If you get a dialog box with an error you can click the close button on the upper right hand corner of the error messages and continue the operation.

6. Type the name of the text file you would like the new Master Gene Table saved as in the box labeled “Save gene list to”. If you save the new Master Gene Table using the same name as the current table file (in this example, ORF_table.txt) then the updated file will define this genome, rather than the previous table of genes file. If you save this updated Master Gene Table under a different file name (for example, ORF_table2.txt), then the old Master Gene
7. Click the “Save and Close” button to save the updated Master Gene Table. If, for some reason, you do not want to save, close the window by clicking the close button in the upper right hand corner. You can select the “Save and Close” button at any time during the update. The searched items will have been temporarily stored in your computer and will be visible in GeneSpring when you restart. It will go through the genes it has already updated really fast. It will take five to 30 seconds per gene depending on how much data the GeneSpider is bringing back. You may want to let this program run over your lunch hour, or for very large genomes, overnight.

3.3.2.7 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, then the sequence data should be put into a separate file and formatted using the .seq format. A severely abridged example of the yeast.seq might look like the following.

>CHR1 Chromosome I data:
CCGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGG
>CHR2 Complete DNA sequence of yeast chromosome II.
AAATAGCCCTCATGTACGTCTCCTCCAAGCCCTGTTGTCTCTTACCCGGA . . .
AGAATAGGGTAGATTGTGTAGGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGG
TGTGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGG
>CHR3 LOCUS SCCHRIII 315341 bp DNA PLN
25-NOV-1996
CCGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGGTTGTTGGG
>CHR4 LOCUS SCCHIV 315341 bp DNA PLN
25-NOV-1996

If you have multiple chromosomes, they should be named sequentially, CHR1, CHR2 and so on. If there is only one chromosome, name it CHR1.

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

3.3.2.8 Where Do I Put My Data Files?

The files should be put in the same folder within GeneSpring’s data directory. The default data directory for GeneSpring in a PC is C:\Program Files\Silicon Genetics\GeneSpring\data. In this data directory, use your file management program to create a new sub-directory to hold the new genome data. This folder is usually named after the organism you are adding, but any memorable name will suffice.

There are three possible raw data files you may have when you create a new genome.
1. You must have a Master Gene Table or a GenBank/EMBL file(s).
2. You can have sequence data in .seq format.
3. You may have a file containing extra, non-GenBank genes (if you have any). The file of extra genes should be in one of the four standard Master Gene Table formats. The three raw data files should all be placed within your new subdirectory.

3.4 Genome Wizard

Most of these panels are fairly self-explanatory. This section is intended to show the different possible appearances a panel can have, and add notes about characteristics that are not obvious. Most Wizard panels will take up most of your screen. This is to prevent any necessary boxes from being shrunk to a non-visible size. You can change the size of any panel in the usual manner of grabbing an edge with the cursor and dragging, but it is recommended you leave them at the large size. You may not see every panel presented here as you go through the Genome Wizard as the Genome Wizard will modify itself depending on your answers. The “New Genome Installation Wizard” saves your genome information as a “.genomedef” file, as explained in 3.5.2 “The .genomedef File” on page 3-24. Even if you do not plan to install genomes using a “.genomedef” file it is probably a good idea to read that section in case you want to modify an existing genome by changing the “.genomedef” file the Genome Wizard created.

In the following example a new C.Elegans genome and sequence will be loaded into GeneSpring.

3.4.1 Welcome to the GeneSpring Genome Installation Wizard! panel

![Image of Welcome to the GeneSpring Genome Installation Wizard! panel]

Figure 3-9 The “Welcome to the GeneSpring Genome Installation Wizard!” panel

This window tells GeneSpring the name of the genome you are installing. To name a genome:

1. Place the cursor in the Organism Name box.
2. Type the name of the organism as you wish it to appear in GeneSpring. This name can be anything, but a sensible, memorable name is recommended. In this example, the organism name, “Mouse” is used.

Remember, GeneSpring is case sensitive. Whatever capitalization you use here will be how the genome name appears everywhere. Click the “Next” button to move forward to the next panel.

### 3.4.2 Genome Data Directory panel

![Figure 3-10 The “Genome Data Directory” panel](image)

In this panel you can select or create a new directory. GeneSpring will bring up a default directory. If you type in the name of a non-existent directory GeneSpring will create it for you. Later you can use the Wizard to select various files and GeneSpring will copy them into this directory automatically. See 3.3 “Raw Data” on page 3-2 for the correct format of the raw data files.

To enter the directory:

1. Write the complete directory pathway in the “Specify directory” box.

If you already have a directory for the organism you named in Figure 3-9 GeneSpring will ask you to define a subdirectory. If you are starting a new species directory this will be unnecessary.

Or, if you have already created a directory as specified in 3.2 “Creating Folders for New Genomes” on page 3-2, you will need to type in or browse to find that directory.

1. Click the “Browse” button. Before you begin browsing, look at the folder to make sure you are in the folder you want.

2. Find the file directory containing your raw data files.
3. Click the directory file. This opens the directory. You should see your raw data files within this directory.

4. Click the “Save” button. This writes the pathway in the “Specify directory” box of the Genome Wizard.

When you click the “Save” button in the “Browse directory” window, the “File Name” box in the window contains the file name “[Dummy Name, leave alone]”. This is what the window is supposed to look like when you click the “Save” button. If you accidentally click one of the files within the genome’s directory, the name in the “File name” box changes. Then, when you click the “Save” button you get an error message:

![Figure 3-11 Possible error message](image)

Click the “Yes” button of this error message; this does not replace the raw data file, it simply enters the directory of the correct file into the “Specify directory” box of the Genome Wizard.

Click the “Next” button in the Genome Wizard to move to the next panel. If you press “Next” without specifying your genome directory, then GeneSpring will create a directory for you in the GeneSpring\data directory. Directories automatically created in this way are named using the name of your genome. GeneSpring will automatically copy your files into this directory. You can select **File > New Window** to see the new files.
3.4.3 The Overall Genome Properties panel

In this window you tell GeneSpring whether the genome you are entering has been sequenced, and if it has a circular genome. In the first box, click the “Yes” circle if your organism has been sequenced, otherwise leave the “No” circle selected. In the second box, click the “Yes” circle if your organism is a circular genome, like bacteria, plasmids, and viruses. If it is, GeneSpring will display it as a circle in the physical position display. Leave the default setting of “No” selected if your organism does not have a circular genome.

3.4.4 The GenBank Data File panel

While GenBank offers several different files for their complete genomes, GeneSpring can only read their .gbk files.
In this panel you tell GeneSpring if you are using a GenBank file as your data source, and if so, what the file is named. An EMBL file may be used in place of a GenBank file. For the purposes of this panel, treat the EMBL file as if it were a GenBank file; answer “Yes” to having a GenBank file and enter the file name and pathway of the EMBL file where it asks for the GenBank file name. You may need to download a GenBank file, please see 3.3.2.5 “GenBank or EMBL Files” on page 3-7.

To indicate you have a GenBank or EMBL file:

1. Select “Yes”. If you are not using a GenBank or EMBL file, leave the “No” circle selected and go on to the next panel.

2. Either type the complete file name and pathway of your GenBank/EMBL file in the “Enter filename” box, or click the “Browse” button. This brings up the browser window.

   Look at the folder listing to make sure you are in the folder you want.

3. Click the GenBank or EMBL file for this organism.

4. Click the “Open” button. This enters the complete pathway and file name of the selected file in the “Enter filename” box of the Genome Wizard.

Once you indicate you have a GenBank/EMBL file, then this panel will not let you move forward until you have entered the file name of your GenBank/EMBL file in the “Enter filename” box. When you use the “Browse” button to select the GenBank/EMBL file, click once in the Wizard panel and then click the “Next” button to go on to the next panel. If you do not use the browse feature, be very careful of spelling and capitalization errors, as GeneSpring attempts to locate the file before it allows you to progress to the next panel.
3.4.5 The Master Gene Table panel

You will not see this panel if you are using a GenBank or EMBL file for your organism.

![Master Gene Table panel](image)

Your Master Gene Table must be in a name list, name function, SGD or mapped format. Please see 3.3.2 “What Format do these Data Need to be in?” on page 3-3 for an example.

This panel tells GeneSpring what the name of your Master Gene Table is, and what format it is in. The Master Gene Table is referred to as a “Gene List” file in this panel, because the list of gene names are the most important information contained in the Master Gene Table. To enter the Master Gene Table’s file name, either type the complete pathway and file name of the Master Gene Table file, or:

1. Click the “Browse” button. A window will appear.

   Look at the folder listed to make sure you are in the folder you want.

2. In this new window, click your Master Gene Table file (for example, ORF_table.txt).

3. Click the “Open” button. This enters the filename and pathway within the “Enter GeneList Filename” box of the Genome Wizard.

The Master Gene Table file will be copied into the correct folder by GeneSpring. You will not be able to go to the next panel until a Master Gene Table file has been indicated. GeneSpring checks to make sure the file name you typed actually exists. Beware of spelling and capitalization errors because if GeneSpring cannot locate the file you indicate you will not be permitted to progress to the next panel.
3.4.6 The Genome Sequence File panel

You will not see this panel unless you indicate in the “Overall Genome Properties” panel that your genome has been sequenced, and you are not using a GenBank or EMBL file. This panel tells GeneSpring where to find the sequence data. To do this, click the “Enter Genome Sequence File Name” box and type the complete file name and pathway or:

1. Click the “Browse” button. A window will appear.
   Look at the listed folder to make sure you are in the folder you want.
2. Select the .seq file containing your organism’s sequence.
3. Click the “Open” button. This enters the file name and pathway into the “Enter Genome Sequence File Name” box of the Genome Wizard.

You cannot go onto the next panel until you have entered a file name. The sequence data file will be copied by GeneSpring to the correct directory. The file you indicate in the “Enter Genome Sequence File Name” box must exist, or the Genome Wizard will not let you continue. Beware of spelling and capitalization errors as GeneSpring needs to locate the file before allowing you to progress to the next panel.
3.4.7 The Additional Genetic Elements panel

This table tells GeneSpring if you have a second table of genes. Generally a second table of genes is used if you want 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. To tell GeneSpring where to find the additional elements:

1. Click the “Yes” circle to select it. If you do not have a separate table of genes file leave the “No” circle selected and go to the next panel.

2. Either click in the “Enter Filename” box, and type the complete file name and pathway, or click the “Browse” button to select a file.
   
   Look at the listed folder to make sure you are in the correct directory.

3. Click the table of genes file containing the extra genomic information.

4. Click the “Open” button. This will insert the file information into the “Enter Filename” box.

5. Click the arrow to the right of the “Select a file format” box. A menu will appear.

Figure 3-16 The “Additional Genetic Elements” panel
Click the format used in the supplementary table of genes file. For a description of the four format options, see section 3.3.2 “What Format do these Data Need to be in?” on page 3-3.

Once you indicate you have a file containing extra genomic elements, you cannot proceed to the next panel until you have indicated a file and a file format. Beware of spelling and capitalization errors when indicating the file name and pathway, as GeneSpring checks to make sure the file you name exists before letting you go on to the next panel.

3.4.8 The Links to Web DataBases panel

Select the “Yes” circle and the “Next” button if you have databases on the World Wide Web you would like to easily access from GeneSpring.

This panel allows you to link GeneSpring directly to web-based data sources on your genes. You can create a link to a URL containing the name of one of your genes. If you would like to have any such links, select the “Yes” circle. In the “Enter number of links” box type the number of web databases you want to link the genes in this genome to. When you enter a number in this box, the
number of “Button” lines in the table below changes. 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 locus, use a percent sign (%). To use the systematic name, less anything after a dash, use the dash (-).

| Button #3 |                 |                                                     |
| Button #4 |                 |                                                     |

For each link desired, enter both a button title and a URL pointing to the correct page.

Figure 3-19 The “Links to Web Databases” panel, some options filled

If you want to place more buttons, you can change the number in the “Enter number of links” option. Then use the tab key to move through the Button Label table.

When you right-click the table in this panel of the Genome Wizard, there is no pop-up menu allowing you to cut and paste. 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 has a much higher success ratio as URLs are both spelling and case sensitive. GeneSpring will attempt 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. If this is the case you will have to skip this panel and add the web-links to the “.genomedef” file later. To add hyperlinks from GeneSpring, please see GeneSpring User Manual 6.1.6, “Web Connections”.

Chapter 3-20

Copyright 1998-2001 Silicon Genetics
For NT and Mac users, you should set the path to your usual browser, because GeneSpring cannot automatically locate the default web browser on NT or Mac machines, which may cause you trouble in this panel. To set the path to the browser:

1. Select **Edit > Preferences**.

![Figure 3-20 The Preferences window, “Browser”](image)

2. Select **Browser** from the drop-down menu.

3. 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 the right of the “Browser path” box. If you do this, a window will appear.

   The default from the Preferences box may take you into the wrong folder. You will need to look for your default browser’s files in your system directory. In a Windows NT environment your path may look something like this: “C:\Program Files\Plus!\Microsoft Internet\IEXPLORE.EXE”.

   a. Find and select the “.exe” file associated with your internet browser.

   b. Click the “Open” button in the “Browse” window. This writes the complete “.exe” file name and pathway in to the “Browser path” box of the “Preferences” window.

4. Click “OK” to close the Preferences panel. The path to your browser should be set.
3.4.9 The Miscellaneous Settings panel

This panel lets you alter the way the gene names are displayed. If you wish to force all of the systematic gene names to upper or lower case letters select the appropriate check box. It is perfectly acceptable not to select any of the check box options, and simply proceed to the next panel.

3.4.10 The Finished panel

This panel lets you alter the way the gene names are displayed. If you wish to force all of the systematic gene names to upper or lower case letters select the appropriate check box. It is perfectly acceptable not to select any of the check box options, and simply proceed to the next panel.
When you click the “Finish” button all of the answers you gave in the previous Genome Wizard panels are saved in a “.genomedef” file.

### 3.5 Installing from a Text File

The following steps are needed to load a genome. These steps are essentially the same as the questions you answer in the Genome Wizard. The specific examples and instructions given are for E. coli.

1. Open the GeneSpring data directory (typically C:\Program Files\SiliconGenetics\Gene-Spring\data), using your file management program.
2. Create a sub-directory to hold the new genome data. (In this example the folder is “Ecoli”.)
3. Copy your Master Gene Table, GenBank, or EMBL file(s) in this new directory. If you have a separate sequence file, put that in this new directory also. If you have a file containing extra genes, put that file in this new directory.
4. In the same directory, create a file describing the genome. The file name should end with “.genomedef”, such as “Ecoli.genomedef”. See 3.5.2 “The .genomedef File” on page 3-24, for what this file should contain.
5. All files within the “GeneSpring\data” directory (except those in the “cache” directory if there is one) ending in “.genomedef” are found automatically. Start GeneSpring to make sure your genome is properly loaded. You should be able to find its name by selecting File > New Genome. In this example “E. Coli” appears there.

#### 3.5.1 Creating Folders for New Genomes

To manually create a new folder in the Genome Browser, you must go through a file management system, such as Windows Explorer®. For example, a new folder named “Mouse” has been created and placed into the data directory of GeneSpring.

![Data Directory Structure](image)

**Figure 3-1 A new “Mouse” folder**

Before your new Mouse folder will appear in the Navigator you will need to create a correct “mouse.genomedef” file. A genomedef file will contain all the information GeneSpring needs to...
create a folder and other data objects. Make sure you save the “.genomedef” file in your new directory after you create it.

3.5.2   The .genomedef File

The “.genomedef” file contains a brief description of the genome. This file contains several lines, each of the form object-name space-colon-space object-value. For example: Object-name : object-value.

An example of how this actually appears in the “.genomedef” file is:

    name : e.coli

In this example “name” is the object-name and “e.coli” is the object-value. The object-value can be thought of as the answer to the question posed by the object-name. In the “.genomedef” file the order of lines is not significant, but the case (lower or upper case) of letters is significant. The spelling, especially of the object-name is also significant. Blank lines and lines beginning with the number character (#) are ignored.

3.5.2.1   Define Your Genome

This section is designed to help you create a “.genomedef” file for a particular genome, and therefore it is written as a series of questions for you to answer. There are two examples following each question. The first is the generalized form of the answer, including the generalized object-name and what sort of response constitutes a correct object-value. The second (bold-faced) example is an example of an actual answer to the question. Some of the lines the questions represent are required, and others are not, each question will be annotated accordingly. The genome, “e.coli” is used as the example throughout this section. A complete (and annotated) “.genomedef” file for the “e.coli” genome is given in 3.6 “Example of a genome being installed” on page 3-29.

1. Enter the name of your genome as you wish it to appear in GeneSpring. This line is required.

    name : the name of the genome

        name : e.coli

2. If you are using a Master Gene Table to define your genome, enter the complete file name of the file containing the Master Gene Table. This question and the next question are mutually exclusive, you must have one of them in your “.genomedef” file.

        ORFs : the complete file name of the file containing the Master Gene Table of all the genes

        ORFs : genelist.txt

3. If you are using either a GenBank file or an EMBL file to define your genome, enter the complete file name of the file describing your genome. This is necessary if you used a GenBank or EMBL file. This question and the previous question are mutually exclusive. One of the two is required.

        GenBank: the name of the GenBank/EMBL file describing this genome
Installing a New Genome

GeneSpring Loading Data Manual

Installing from a Text File

**GenBank : ecoli.gbk**

Or,

**GenBank : ecoli.ebl**

Even if you are using an EMBL file the object-name in this entry is GenBank.

4. If you have a file containing extra genes, enter the complete file name of the file containing these supplementary elements. This line is optional, but must be included in the “.genomedef” file for GeneSpring to incorporate this data.

   **nonORFs : the complete file name of the extra file containing other genomic elements than in the ORFs file**

   **nonORFs : extragenes.txt**

5. If you have a file containing the sequence data for the genome, enter the complete name of that file, including the .seq suffix. This line is optional, but must be included in the “.genomedef” file for GeneSpring to incorporate sequence data not included within a GenBank or EMBL file.

   **sequence : the name of a file containing the sequence(s) for the genome**

   **sequence : ecoli.seq**

6. If you are using a Master Gene Table to define your genome, indicate which format you used. The four Master Gene Table format options are: name list, name function, SGD, or mapped. These are also the four possible object-values for this question. See 3.3.2 “What Format do these Data Need to be in?” on page 3-3 for a description of these formats. This line is required if the ORFs line from question two was used.

   **ORFFormat : the format for the Master Gene Table specified in the ORFs line**

   **ORFFormat : mapped**

7. If you are using a supplementary table of genes file, indicate which table of genes format is used in this file. This can be one of the four table of genes format options: name list, name function, SGD, or mapped. These are also the four possible object-values for this question. See 3.3.2 “What Format do these Data Need to be in?” on page 3-3 for a description of these formats. This line is required if the nonORFs line from question four was used and the format for this file is different from the format given in response to question six.

   **nonORFFormat : the format for the file specified in the nonORFs if different from the file of ORFs**

   **nonORFFormat : name function**

8. If the genome you are entering has been sequenced, then you should answer “true” to this question. This line is optional, but if you are using a GenBank file, an EMBL file, or a “.seq” file to define your organism’s sequence, then the sequence data will not be loaded into Gene-
Spring if this line is not in the “.genomedef” file. If your organism has not been sequenced, or you do not have its sequence information available, then you do not need to enter this line in the “.genomedef” file.

\[\text{KnowGenome} : \text{set to true if the genome is sequenced, and false if not}\]

\[\text{KnowGenome} : \text{true}\]

9. If the genome you are entering is a circular genome (such as bacteria, plasmids, and viruses) then you should answer “true” to this question. This line is optional, if you do not enter it, or answer it “false” then your genome will not be plotted as a circle in the physical position display.

\[\text{CircularGenome} : \text{set to true if the genome should be plotted as a circle and false otherwise}\]

\[\text{CircularGenome} : \text{true}\]

10. Are there web-based databases you would like to be able to link to automatically? If not, skip this question. You can link to the URL of any web-based database containing the name of your gene. Each separate link should consist of one line in the “.genomedef” file. Each line should start with the phrase “GeneHypertextLinks” followed by a colon, followed by the description of the link.

The description of the link is the name of the link (the name you want to appear on a button in GeneSpring), which must be followed by a colon, not a semicolon. Any field in angle brackets (for example, <field>) will be replaced by the value of that parameter. The allowed parameters are:

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

A link will only be enabled for a particular gene if all parameters mentioned in that URL are defined for that gene.
GeneHypertextLinks: Links to external web based databases. You can have more than one of these lines; you should have one line for each link.

GeneHypertextLinks: linkname:http://www.somewhere.org&gene=<systematic>&id=<genbank>

This example should be one consecutive line beginning “GeneHypertextLinks:”, but is has been broken into separate lines to allow it to fit on this page. It should be entered into the file as single line, without carriage returns. There is no space between the semicolon following the link’s name and the associated URL.

Experiment URLs work exactly the same way, except that they begin with ExperimentHypertextLinks instead of GeneHypertextLinks, and the things in <> signs are the names of parameters. A link will only be shown in the Experiment Inspector if the experiment has parameters with names matching all fields in the URL.

In both cases, the parameter names are not case sensitive, so if an experiment has a parameter called Time, you can specify it as <time>, <Time>, or <TIME> in the URL, and they will all work.

ExperimentHypertextLinks: Links to external web based databases. You can have more than one of these lines; you should have one line for each link.

ExperimentHypertextLinks: linkname:http://www.somewhere.experimentlikemine=<systematic>&id=<time>

11. Use this line if there is a particular experiment you would like GeneSpring to automatically display in the genome browser when you open this genome. This “.genomedef” entry is optional, if it is not included GeneSpring will open the genome but not open any particular experiment when you select this genome to be displayed.

defaultExperiment: the name of the default experiment you want started when opening this genome

defaultExperiment: yeast extraterrestrial studies

The name following the object-value should be the same name given to the experiment in the name line of its “.html” file and/or it should be the name entered for the experiment in the “Properties of an Experiment Set” panel of the “New Experiment Wizard”. Both of these options are case sensitive, so make sure the spelling and capitalization is correct. See Chapter 4, Installing a New Experiment for more information about entering an experiment. If you do not know the name of any experiment done with this genome when you create it, this line can be added or modified afterwards. (Just remember to save the modified “.genomedef” file.)

12. If you work in a group that is storing data and analyses in a shared environment (usually this means that you have all of the data for the group in one file system) you will probably also want to have your own local data for each genome. A specific use of this is for gene lists (not
the genome defining Master Gene Table, but a gene list you create within GeneSpring): it is often desirable for each person to keep the gene lists they create initially separate as trial lists, and then merge them into the groups’ permanent set when they are more certain about the significance of individual lists. To store data locally, you specify (in the “.genomedef” file of each genome) a second directory to be searched for experiment data, gene lists, trees, etc. This directory is specified with the line below. This is an optional line.

```
HomeDirectory : The complete path of an extra directory to search for to find information for this genome
```

```
HomeDirectory : C:\Silicon Genetics\GeneSpring\data\Ecoli
```

Including this line means that both this directory on your local computer and the directory containing the “.genomedef” file are searched for experiment data, gene lists, classifications, and so forth. As the local directory must be indicated in the shared directory, every user in your group must keep their local directory in the same place on their local computers. In the example this place would be the C:Silicon Genetics\GeneSpring\data\Ecoli.

13. If there is a prefix (a string of characters) prepended to the start of your genes’ systematic names you can tell GeneSpring to disregard this first part of the gene name and not display it. This line is not required, and it is rarely used.

```
SystematicPrefix : a string that is often prepended to the start of gene names, and should be ignored if seen
```

```
SystematicPrefix : ecoli/
```

14. If you wish the genes’ systematic names to appear entirely in upper case letters, GeneSpring can convert them to this automatically. This line is not required, and is rarely used.

```
ForceUpperCase : set to true if you want all the names of the genes converted to upper case, set this line to false otherwise
```

```
ForceUpperCase : true
```

15. If you wish the genes’ systematic names to appear entirely in lower case letters, GeneSpring can convert them to this automatically. This line is not required, and is rarely used.

```
ForceLowerCase : set to true if you want all the names of the genes converted to lower case, set this line to false otherwise
```

```
ForceLowerCase : false
```
16. You can place any data you wish in the custom label columns.

```
Custom1Label : heading

Custom1Label : interacts with P53
```

17. You can place any data you wish in the custom label columns.

```
Custom2Label : heading

Custom2Label : molecular weight
```

18. You can place any data you wish in the custom label columns.

```
Custom3Label : heading

Custom3Label : plate and well location
```

19. If your genome has a unique identifier, such as a nickname, that would speed searching for it, enter it in this line.

```
Identifier : optional unique identifier for the whole genome

Identifier : dutch elm disease study
```

20. You can use ChromosomeNames to cause the “mito” chromosomes to be sorted separately from the remaining chromosomes.

```
ChromosomeNames :

ChromosomeNames : I;II;III;IV;V;VI;VII;VIII;IX;X;XI;XII;XIII;XIV;XV;XVI;mito
```

Make sure you save the “.genomedef” file after you create it.

### 3.6 Example of a genome being installed

#### 3.6.1 Raw data files

You can define any genome with only a list of gene names, as opposed to the more complete organism information (including mapping and sequence information) available for yeast and bacterial genomes. Hence, the examples listed here are for the E. coli genome and the human genome. E. coli is an example of an organism defined by a GenBank file, and the human genome is included as an example of an organism defined by a Master Gene Table. In this particular case the Master Gene Table is in the “mapped” format.

#### 3.6.1.1 E. coli raw data files

![Figure 3-2 The icon of the GenBank file used to define the E. coli genome](ecoli.png)
<table>
<thead>
<tr>
<th>LOCUS</th>
<th>U00096</th>
<th>4639221 bp DNA circular BCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-OCT-1997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEFINITION</td>
<td>Escherichia coli K-12 MG1655 complete genome.</td>
<td></td>
</tr>
<tr>
<td>ACCESSION</td>
<td>U00096</td>
<td></td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Escherichia coli.</td>
<td></td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Escherichia coli</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eubacteria; Proteobacteria; gamma subdivision;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Enterobacteriaceae;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Escherichia.</td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>1 (bases 1 to 4639221)</td>
<td></td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Blattner,F.R., Plunkett III,G., Bloch,C.A.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Perna,N.T.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Burland,V.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Riley,M., Collado-Vides,J., Glasner,J.D.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rode,C.K.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mayhew,G.F.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gregor,J., Davis,N.W., Kirkpatrick,H.A.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Goeden,M.A.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rose,D.J.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mau,B. and Shao,Y.</td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>The complete genome sequence of Escherichia coli K-12</td>
<td></td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Science 277 (5331), 1453-1474 (1997)</td>
<td></td>
</tr>
<tr>
<td>MEDLINE</td>
<td>97426617</td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>2 (bases 1 to 4639221)</td>
<td></td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Blattner,F.R.</td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>Direct Submission</td>
<td></td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Submitted (16-JAN-1997) Guy Plunkett III,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Laboratory of Genetics,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Wisconsin, 445 Henry Mall,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Madison, WI 53706, USA.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Email: <a href="mailto:ecoli@genetics.wisc.edu">ecoli@genetics.wisc.edu</a> Phone: 608-262-2534 Fax: 608-263-7459</td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>3 (bases 1 to 4639221)</td>
<td></td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Blattner,F.R.</td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>Direct Submission</td>
<td></td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Submitted (02-SEP-1997) Guy Plunkett III,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Laboratory of Genetics,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Wisconsin, 445 Henry Mall,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Madison, WI 53706, USA.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Email: <a href="mailto:ecoli@genetics.wisc.edu">ecoli@genetics.wisc.edu</a> Phone: 608-262-2534 Fax: 608-263-7459</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3-3** The first page of the GenBank file used to load *E. coli*. As this file is a 400+ page document, it is not shown in its entirety.
### 3.6.1.2 Human data raw files

```
Z97181  
AL022401  
AC004386  
AC004388  
Z98950  
AC004478  
AC003666  
AC002549  
AC003669  
Z82204  
AC004383  
AC004072  
AC003683  
AC003658  
Y15994  
AL009175  
AC003037  
AL008713  
AC002422  
AC002523  
M22332  
AQ409366  
AQ356884  
AQ309743  
X61295  
AQ572229  
AI683867  
AQ390430  
AQ355719  
AQ559819  
AQ557343  
AQ573089  
M54985  
AQ549999  
AI421777  
AQ420901  
AQ382430  
AQ440210  
AQ536099  
AI475350  
S67068  
AI821169  
AL045241
```

**Figure 3-4** A Master Gene Table of human genes, in “name list” format
3.6.2 The .genomedef file

If you write a “.genomedef” file you do not need to use the Genome Wizard to define your genome and vice versa; if you use the Genome Wizard to create your genome you do not need to write a separate “.genomedef” file.

3.6.2.1 An Ecoli.genomedef file

```plaintext
name : E. coli
GenBank : ecoli.gbk
defaultExperiment : time series
KnowGenome : true
CircularGenome : true
GeneHypertextLinks : ECD:http://genome4.aist-nara.ac.jp/cgi-bin/all.data.ODG?data=top&keyword-d=&gene-name=on&keyword-s=&kbpstart=0.0&kbpend=4700.0&minstart=0.0&minend=100.0
GeneHypertextLinks : GenProtEC:http://dbase.mbl.edu/genprotec/findgene.htf?G=
```

The example given here is for E. Coli. This is the entire text of the “ecoli.genomedef” file. The entries following the space-colon-space (“ : ”) are specific to the E. Coli genome.

1. “name : ”, indicates the name of the genome as it will appear in GeneSpring.
2. “GenBank : ”, indicates the name of the file containing the GenBank description of the genome.
3. “defaultExperiment : ” indicates the experiment automatically displayed when you open this genome in GeneSpring.
4. “KnowGenome : ”, indicates whether or not the genome you are loading has been sequenced. In the case of E. Coli it has, and thus the answer is true.
5. “CircularGenome : ”, is another true-false question, if the genome you are loading is circular, the answer is true.
6. The remaining lines have been broken into separate lines so they will fit on this page, but they should be put into the file as single lines without carriage returns. Each line prefaced with “GeneHypertextLinks : ” (lines 6, 8, 10 and 11) refer to an Internet source of data on particular genes for the genome being installed. In this example, the four lines refer to GenBank, ECD, GenProtEC, and the Pasteur databases.

3.6.2.2 A human.genomedef file

```plaintext
name : Human
ORFs : GeneNames.txt
ORFFormat : name list
defaultExperiment : MS
KnowGenome : false
HomeDirectory : C:\GeneSpring\data\human
GeneHypertextLinks : LocusLink:http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi?Q=,&ORG=Hs&V=0
GeneHypertextLinks : DDBJ:http://ftp2.ddbj.nig.ac.jp:8000/cgi-bin/get_entry.pl
```

The example given here is for Human. This is the entire text of the “human.genomedef” file. The entries following the space-colon-space (“ : ”) are specific to the Human genome.
The example given here is for human data. The entries following the space-colon-space (" : ") are specific to the human genome. From these specifications, the genome name will appear as “Human” in GeneSpring. The Master Gene Table is found in the GeneNames.txt and is in the ‘name list’ format.

1. “name : ”, indicates the name of the genome as it will appear in GeneSpring.

2. “ORFs : ”, indicates the name of the file containing the Master Gene Table that constitutes the description of the genome.

3. “ORFFormat : ” indicates what format the “GeneNames.txt” file is in. The answer given in the example “name list” indicates this file is only a list of gene names. See 3.3.2 “What Format do these Data Need to be in?” on page 3-3 for more information about file formats.

4. “defaultExperiment : ” indicates the experiment is automatically displayed when you open this genome in GeneSpring.

5. “KnowGenome : ”, indicates whether or not the genome you are loading has been sequenced. The human genome has not, and thus the answer is false. When it is, the answer will be true.

6. “HomeDirectory : ” use this only if you are using a shared data environment and wish to have local data as well. This line indicates the name of the local directory to be searched; in this case it is C:\GeneSpring\data\human.

7. Lines seven through fourteen are really only six lines that have been broken into separate lines so that they will fit on this page, but they should be put into the file as single lines without carriage returns. Each line that is prefaced with “GeneHypertextLinks : ” (lines 7, 9, 10, 12, and 14) refer to an Internet source of data on particular genes for the genome being installed. In this example the five lines refer to GenBank, GeneCards, UniGene, LocusLink, and the DDBJ databases.
3.7 Updating Your Information

Currently, you can update your information by using the GeneSpider to retrieve information from online databases such as GenBank or LocusLink. Please do not attempt to retrieve information from more than one database at a time, as this may cause your computer to overwrite files.

Select **Tools > GeneSpider > Update files from Silicon Genetics**.

![Figure 3-5 GeneSpider’s Silicon Genetics box](image)

### 3.7.1 Options in the GeneSpider

#### 3.7.1.1 Data Mining Information

You will need to select the column containing the GenBank locus in the drop-down menu. The GenBank locus must be included in the Master Gene Table.

Click the drop-down menu in the “Options” box to get a menu as shown in Figure 3-6. Select any option to use as your searched column.

Do not worry if your Master Gene Table includes genes without GenBank locus. When you update a genome using the GeneSpider it copies your current Master Gene Table so you do not lose any information regarding non-GenBank genes. After the GeneSpider copies this file, it updates the fields of the copied file associated with a GenBank locus. While it is doing this you may be able to see changes in the browser window reflecting the influx of new information (the “Status” box).
For details on how these fields are entered in your genome, please refer to 3.3.2.4 “Mapped” on page 3-3.

### 3.7.1.2 Overwrite Existing Information

By selecting this option, you are allowing GeneSpider and GeneSpring to overwrite a file in your current directory. When you save this information, it is saved in the “mapped” Master Gene Table format. This option is deselected by default. If you select this option, your new data will replace any old data, and you will not have to alter the file names or reset the defaults of GeneSpring to view it.

At the bottom of this window is the “Save and Close” button. If you click this button when the GeneSpider is done retrieving information, GeneSpring will save the information retrieved in the manner you selected.

If you close the GeneSpider by clicking the small “x” in the upper right corner, GeneSpring will not save the information. This can be a very useful distinction if you experience problems in the downloading process.

### 3.7.1.3 Retrieve Sequence Data

If available, the GeneSpider can also retrieve sequence data for your genome. Select the box.

### 3.7.1.4 Save to Directory

This line allows you to fill in (or paste) the exact directory and file you would like your data saved to. If you select a directory other than your current file, your old data will continue to be used by GeneSpring.

### 3.7.1.5 Running the GeneSpider

After you click start, you may get many windows like Figure 3-7. Just click on “Don’t Show Again” if you need to work on other things while the GeneSpider updates your genome.
3.7.2 Updating a GenBank or EMBL File

If you used a GenBank data file as the source for your genome you do not need to use the GeneSpider to update your genome. Instead, all you have to do is replace your old GenBank file (or EMBL file) with a new one. If you replace the GenBank/EMBL file while GeneSpring is open, you must restart GeneSpring before it can use the new file as the source of your genome data.

3.7.3 Updating GenBank Information using the GeneSpider

If you used a Master Gene Table including the GenBank locus for some or all of your genes, then the information associated with the GenBank locus can be updated from GenBank.

Select Tools > GeneSpider > Update files from GenBank.

To update the human genome from GenBank, you will get a dialog box like Figure 3-8.

A new feature in the GeneSpider is the ability to retrieve sequence and source data for Expressed Sequence Tags (EST) from GenBank. Select “Retrieve sequence data” (blank in Figure 3-8).

3.7.4 Updating LocusLink Information using the GeneSpider

Select Tools > GeneSpider > Update files from LocusLink.
3.7.5 Updating UniGene Information using the GeneSpider

The UniGene cluster number is revised and updated as UniGene improves. Unfortunately, this means GeneSpring cannot use it as a reference number, and does not save it.

Select Tools > GeneSpider > Update files from UniGene.
Once you have made sure all the information in this box is correct, you can click the Start button to initiate the update.

For details on the GeneSpider, please read the *GeneSpider Manual*. 

---

**Figure 3-10 GeneSpider’s UniGene box**

**Figure 3-11 The UniGene drop-down genome menu**
Chapter 4 Installing a New Experiment

There are five ways to load experiments described in the various sections of this chapter:

- section 4.2, Autoload Experiment
- section 4.4, Experiment Import Wizard
- section 4.5, Installing from a Text File
- section 4.6, Copying and Pasting Experiments
- section 4.7, Installing from a Database

You might want to skim through every section and decide which installation technique is best for you and your data.

Before you begin installing your new experiment you need to go through the Genome Installation Wizard to specify a new genome, if the genome for your experiment is not yet in GeneSpring so GeneSpring will correctly interpret what you are telling it. If you are not cutting and pasting data, you will need to create a folder called Experiments and place your experimental data files in that folder so they will be easy to find when you need them later in this process.

4.1 Some Helpful Definitions

- A single condition is defined as each time a numerical measurement is taken for your set of genes, i.e., the results from one array (if you are not using replicates).

- An experiment is a group of conditions you wish to have associated together under one name. This generally means they were all performed using a particular set of parameters (an interpretation).

Please refer to 2.1.1 “Data objects in the Navigator” on page 2-1 and 2.1.2 “Parameter Values” on page 2-2 for more details.

Generally, when referring to an experiment, it means a group of conditions. An experiment represents a test done on a set of genes from a particular genome under certain parameters. For more details and an example of how these data objects are represented in the Navigator of GeneSpring, please refer to 2.1.1 “Data objects in the Navigator” on page 2-1. Generally all of the experiments done on a genome are stored in the “Experiments” sub-directory of the organism’s directory, which is found within the GeneSpring/Data directory. The default pathway for this is:

   C:/Silicon Genetics/GeneSpring/Data/Genome Name/Experiments
4.2 Autoload Experiment

New in GeneSpring version 4.0 is the Autoloader. If you have unedited files from any of the standard technologies you will be able to load them into the GeneSpring in only a few minutes. It is very important to have the raw files, just as they come out of the scanner, as Genespring is programmed to find and sort by the column headers. If you have stripped out any of the columns, you will need to find your original data files and use those.

You will be a little better off if you already have the genome in GeneSpring, so you might want to load that first, if possible. This will be more important if you have subchips.

You can load files from your local hard drive as well as across networks, using shared files.

You can load data by single files or by whole folders (batches) as long as all the files have exactly the same format.

Select File > Autoload Experiment or Ctrl+O.

Figure 4-1 The File menu and the Autoload Experiment command
This will bring up the Load Experiment File window, use the browser to find the correct file and select it.

Genespring will then read the file and guess which technology created it. A dialog box will appear asking you to confirm the type of technology used. If Genespring is correct, please click yes. If not click no, and Genespring will bring up the “Set up column format” window, as shown in Figure 4-3. Double click the blue banner bar at the top of the window to expand it as large as your screen. This way you can view more columns at once. The scroll bar will also allow you to view more columns.

### 4.2.1 Headlines

Genespring will have made a guess as to which row represents your headlines (column titles). If Genespring has guessed incorrectly, you can select the entire “Column Titles” row by clicking on the title at the far left. Use the “Move Headline up” and “Move Headline down” buttons to select a new row to be the Headlines.

![Figure 4-2 The Headlines in the “Set up column format” window](image)

#### 4.2.1.1 No Headlines

If your file has no headlines, deselect the check box marked “Has column titles.”
4.2.2 Functions

Once you have made sure the correct row of headlines has been selected, you can assign the various functions to those columns.

![Image showing the Set up column format window with data entries and options]

If you select any cell in the Functions row, you will get a drop-down menu as shown in Figure 4-3. Using that menu you can assign any column a specific function. You can have all the Flag and Unassigned columns you want, the others can only be used once.

You must have at least one column assigned as Gene Name and one column as Signal (raw data) before the autoloader will allow you to click on the Load Now button.
4.2.3 Autoloader Normalizations

The Autoloader will normalize your new files in ways appropriate for the technology used to create the original data files.

4.2.3.1 Replicates

If you have three or more experiments with the same samples, GeneSpring will automatically normalize each Gene to Itself. Please refer to 1.7 “Normalizing Each Gene to Itself” on page 1-8 for more details and a mathematical illustration of that normalization type.

4.2.3.2 Two Color

Two color Experiments are automatically normalized to a signal ratio.

4.2.3.3 Changing the Normalizations

You can change any of these normalizations quickly and easily in the new Normalizations window accessible in the main GeneSpring window, please refer to GeneSpring User Manual 5.4, “Experiment Normalizations”. Click the “Get Text Description” button to get a complete description of what normalizations have been applied to your experiment.

4.2.4 Remember this Format

If you have a technology that is not currently recognized by the Autoloader or have systematically stripped out the header lines from a file generated by a standard technology, you can tell GeneSpring to remember this format. This format will be added to the cache of recognized formats and GeneSpring will suggest it in the future.

4.2.4.1 Changing Remembered Formats

You cannot edit these formats. If you need to change one, you will have to build a new one.

4.2.4.2 Sharing Remembered Formats

You can share the files of Remembered Formats. Use your favorite browser or file management program to copy the file from YourLocalDrive:\Program Files\SiliconGenetics\GeneSpring\data\Genome\Experiment Formats\name.expformat
Once you have the file, you can paste it into a shared drive.

4.2.5 Load Now

When you are done, you can examine and change any of these options in GeneSpring, please refer to GeneSpring User Manual Chapter 5, “Changing Your Experiment”.
### 4.3 File Formats

You can install a new experiment in one of several ways: by using the “New Experiment Installation Wizard” (see 4.4 “Experiment Import Wizard” on page 4-15) or by creating a “.experiment” file by hand (see 4.5 “Installing from a Text File” on page 4-46). Both experiment entry methods may involve a number of corollary files.

Only one file type is necessary for installing an experiment:

- Experimental data file(s), containing the genes’ names and raw data for each sample in the experiment. Please refer to 4.3.1 “Raw Data” on page 4-6.

Other helpful files might include:

- Layout file
- Region designation files
- A file listing the positive controls
- A file listing the negative controls
- GIF or JPEG pictures to be associated with this experiment, or with particular samples within the experiment
- GIF or JPEG pictures of the Microarray plates the experiment was done on

#### 4.3.1 Raw Data

An experimental file consists of a list of gene names, a list of the raw data associated with them, and the condition or conditions during the test. In addition, an experiment may involve more than one sample, various normalization controls (such as positive and negative controls, and control channel values), pictures of the conditions during the experiment, and pictures of the array plates the experiments were done upon.

#### 4.3.2 What format does this data need to be in?

Data may be in any of the following eight formats, depending on the type of data represented.

#### 4.3.2.1 Experimental Data

You will need at least one file containing your experimental data. This file must have the gene names listed in one column, one name per line, with the experimental data reported in columns. If it were viewed in a spreadsheet it might look like this:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Signal Strength in Experiment 1</th>
<th>Control Channel Strength</th>
<th>Background Signal</th>
<th>Background Signal for the Reference</th>
<th>Experiment Flag</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLN1</td>
<td>510</td>
<td>110</td>
<td>10</td>
<td>10</td>
<td>P</td>
<td>A</td>
</tr>
<tr>
<td>MEP2</td>
<td>9</td>
<td>19</td>
<td>9</td>
<td>9</td>
<td>M</td>
<td>C</td>
</tr>
</tbody>
</table>
If created in a spreadsheet program, the file should be saved as a tab-delimited text file.

If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use comma and periods interchangeably.

GeneSpring can also read experimental data from databases via an ODBC link. Please refer to 4.7 “Installing from a Database” on page 4-78.

4.3.2.2 Pictures of the conditions during the experiment

At most there can be one picture associated with each condition. You do not need to have any pictures but they are good mnemonics, reminding you of what was happening in the experiment at the point you are viewing in GeneSpring. If you have only a few pictures, this can be very useful as GeneSpring will use the picture closest to the displayed condition. These pictures should be either GIF or JPEG files.

4.3.2.3 Pictures of the Microarray plates

At most there can be one array picture associated with each sample. They are helpful but not necessary. These pictures should be either GIF or JPEG files.

4.3.2.4 The Layout file

If you load experiments via the Experiment Wizard or the AutoLoader then you will probably never have to create your own layout file and thus you can skip this entire section. However, if you use the pasting option you may need to create the positive and/or negative control files associated with the layout file.

The layout file tells GeneSpring where to find other files associated with the experiment. If you load in experiments using a “.html” file, then you will need to create a layout file if each sample in your experiment involved more than one array, and/or if the experiment used positive or negative controls. Frequently, the same layout file can be used for more than one experiment.

There are four possible lines in a layout file. Each line is either blank or a line of the form object-name space-colon-space object-value:

Object-name : object-value

An example of this is:

IncludePosControls : false

Here “IncludePosControls” is the object-name and “false” is the object-value. The object-value can be thought of as the answer to the question posed by the object-name. In the layout file the order of lines is not significant, but the case (lower or upper case) of letters is significant. The spelling, especially of the object-name is also significant. Usually when an experiment looks like it is not installed correctly it is because of a spelling or capitalization error. Using the copy (Ctrl+C) and paste (Ctrl+V) functions will help prevent this type of error.

This section is designed to help you create a layout file for a particular experiment, rather than explaining exactly what each possible answer means. There are two examples following each
question. The first is the generalized form of the answer, including the generalized object-name and what sort of response constitutes a correct object-value. The second (bold-faced) example is an example of an actual answer to the question. A complete layout file for the fictitious “Yeast extraterrestrial studies” experiment is given at the end of this chapter. See 4.5.20.2 “An example “.experiment” file” on page 4-66.

The four possible lines in the layout file are:

1. Include this line if your experiment has positive controls. This line refers to a file listing the positive control. If you have positive controls you must have a separate file designating them. See 4.3.2.8 “The Positive and Negative Control Files” on page 4-13 for information about this file.

   PosControlFilename : the complete file name of the file listing the gene names of the positive controls, one per line

   PosControlFilename : PosControls.txt

2. Include this line if your experiment has positive controls. This line tells GeneSpring if you want to display the positive control genes in the Genome Browser with the rest of the experiment, as if they were genes from the organism you are studying. Type “true” as the object-value for this line if you wish to view the positive controls in the genome browser, and enter “false” if you do not.

   IncludePosControls : true or false

   IncludePosControls : false

3. Include this line if your experiment has negative controls. This line refers to a file listing the negative control. If you have negative controls you must have a file designating them. See 4.3.2.8 “The Positive and Negative Control Files” on page 4-13 for information about this file.

   NegControlFilename : the complete file name of the file listing the gene names of the negative controls, one per line

   NegControlFilename : NegControls.txt
4. Include this line if a sample in your experiment involved more than one array, or if there is some reason to normalize the sections of the array separately. If the genes from a sample could belong to more than one region, then the region must be noted somehow in the experimental data file (see 4.3.2.5 “The Region Designation File(s)” on page 4-9). Use this line if the region is noted as either a unique entry in its own column or if it is a suffix appended to another column’s entry. The object-value(s) in this line refer to separate files, each listing one possible region designator. See 4.3.2.5 “The Region Designation File(s)” on page 4-9 for more information. Multiple region designation files should be separated with semicolons, but not spaces.

Regions : the complete file names of the files listing the region designations, separated by semicolons

Regions : YeastRA.txt;YeastRB.txt;YeastRC.txt;YeastRD.txt

So an actual layout file as viewed in Microsoft Notepad® might look like:

```
PosControlFilename : PosControls.txt
IncludePosControls : false
NegControlFilename : NegControls.txt
Regions : YeastRA.txt;YeastRB.txt;YeastRC.txt;YeastRD.txt
```

**Figure 4-4 The Layout file**

4.3.2.5 The Region Designation File(s)

If there is more than one region to which the genes from a sample could belong, then the region must be noted somehow in the experimental data file. If the region is noted in the experimental data file as either a unique entry in its own column or as a suffix appended to another column’s entry (as is common with Affymetrix chips) then you should create separate region designation files, one for each region. In this region designation file should be one line, reading:

RegionSuffix : character or string of characters used either as a unique column entry or as a suffix. This string designates a particular region.

RegionSuffix : A
All of the entries in the region column (designated in the “.html” file or in the “Regions Normalization” panel of the Experiment Wizard) having the same suffix as the object-value indicated after one of the “RegionSuffix : ” entries are considered to be in the same region. For example, if there are four regions, A, B, C, and D there will be four region designation files, each with one of the lines:

```
RegionSuffix : A
RegionSuffix : B
RegionSuffix : C
RegionSuffix : D
```

Given a region column in the experimental data file containing these entries:

```
Gene1A
Gene2B
Gene3C
Gene4D
Gene5A
Gene6B
Gene7C
Gene8D
Gene9A
```

In this example, genes 1, 5, and 9 are all marked as in region A and could be normalized as a discrete group.
An Example:

You have experiment 1 with subchips A, B, C, Da, Dd (2 repeats for subchip D) to be compared to experiment 2 with subchips A, B, Ca, Cb, D (2 repeats for subchip C). You can load it as four samples.

<table>
<thead>
<tr>
<th>Exp 1:</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Da</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp 2:</td>
<td></td>
<td></td>
<td></td>
<td>Db</td>
</tr>
<tr>
<td>Exp 3:</td>
<td>A</td>
<td>B</td>
<td>Ca</td>
<td>D</td>
</tr>
<tr>
<td>Exp 4:</td>
<td></td>
<td></td>
<td>Cb</td>
<td></td>
</tr>
</tbody>
</table>

Table 4-1 Correct entry of repeated sub-experiments

Give experiments 1 and 2 the same parameters. Give experiments 3 and 4 the same parameters.

4.3.2.6 Entering region specifications when they are not specified in their own column or as suffixes within another column

Occasionally a region may not be designated by a unique column entry or as a suffix appended to a column entry. In this case you cannot use the Experiment Wizard to automatically read in your region designations. You will need to create a layout file for your experiment and separate region designation files. A region designation file is used to describe a region, and specifies the following information:

- How to distinguish this region from other regions.
- How to map gene names in this region to the gene names given in the list of genes defining the genome.

There are several ways regions can be distinguished. The four ways listed below are typically used separately, but can occasionally be used in combination, with each other or with the standard way to designate a region.
1. The regions are defined implicitly by the order the genes names as reported in the experimental data file. The names of the genes can be sorted in alphabetical order and used to determine whether a gene is in this region. One can specify inclusive beginning and ending genes, and any genes between them (alphabetically) will be considered part of this region. See the next option for the meaning of "UsesCommas".

   EndRegion : the last gene name in the region
   StartRegion : the first gene name in the region
   UsesCommas : false

   EndRegion : s191
   StartRegion : s001
   UsesCommas : false

2. The regions are defined implicitly by the ordered names of the genes, in a rectangular coordinate system. This is similar to the previous option, except the "names" of the genes are actually coordinates, separated by commas. In this case, a gene is only in the given region if it is between the starting and ending gene names for each dimension separated by commas. For instance:

   StartRegion : 001,100
   EndRegion : 099,199
   UsesCommas : true

3. The regions are defined explicitly by a list of gene names, and optionally a change of names. In this case, you must define a map for the region. A map can be just a list of genes, or it can be a list of names (as used in the experiment files) and the corresponding gene names (as used in gene list defining the genome). In this case, you must specify a text file describing the map (see 4.3.2.7 "How to describe a map" on page 4-12).

   Map : mapA.txt

4. The regions are defined by file name extension. The experimental data for each region is in a separate file. The file names for each sample specified in the Experiment Wizard or in the "html" file are base names, and each region adds an extension to this file name. To prevent name conflicts, this option is frequently used with the map option.

   FileNameExtension : .chipA

4.3.2.7 How to describe a map

Maps are used when you want to change gene names from the raw names (e.g. chip coordinates) into more standard gene names. They can also be used to specify a list of genes defining a region. A map file is a text file containing just two lines:

   FileName : GeneList.txt
   ChangeNames : true

The "FileName" entry specifies the name of a text file containing one line per gene. If "ChangeNames" is true, then the text file should consist of two columns (separated by a tab). The first column should be the gene names as they appear in the experiment data file; the second column
should be the gene names as they appear in the list of genes defining the genome. If “Change-Names” is false, then the text file should only have one column. In this case, the map is used only to specify what is present in a region.

4.3.2.8 The Positive and Negative Control Files

A positive control file and a negative control file are formatted in exactly the same way; their contents are different. Each file lists the control genes’ names, one name per line:

```
Control Gene Name 1
Control Gene Name 2
Control Gene Name 3
Control Gene Name 4
Control Gene Name 5
Control Gene Name 6
...
```

This list of gene names is all either file should contain. There should not be any headlines or anything else in the file, only the gene names.

Briefly, you have negative controls in your experiment when there is DNA from a different genome than the one you are investigating on the array. You are using positive controls when there is DNA from a different genome than the one you are investigating on your array, and you add a known quantity of that different DNA to your sample. For a description of the possible normalizations to be done with these controls see Chapter 1, Normalizing Options.

The names of the positive and negative controls do not need to be listed in your Master Table of Genes. If they are listed, those genes will be colored gray (not measured) in the Genome Browser because they are used in normalization not measurement.

4.3.3 Where do I put my data?

There are eight possible raw data files listed below; only the first one is necessary for loading an experiment.

You must have:

- Experimental data file(s), containing the genes’ raw data for each sample in the experiment. Please refer to 4.3.1 “Raw Data” on page 4-6.

You might have:

- A Layout file
- Region designation file(s)
- A map file
- A file listing the positive controls
- A file listing the negative controls
- GIF or JPEG pictures of the conditions during the experiment
- GIF or JPEG pictures of the Microarray plates the experiment was done on
All of the raw data files should all be placed within the “Experiment” sub-folder of the organism they pertain to. The default pathway for this directory is:

```
C:/Silicon Genetics/GeneSpring/Data/Genome Name/Experiments
```

If the defaults were changed, your version of GeneSpring may be stored elsewhere, but the end of the pathway should be identical on your computer.
4.4 Experiment Import Wizard

Most of the panels in the Experiment Import Wizard are fairly self-explanatory. This section is mainly designed to show the different possible appearances a panel can have, and add any notes about characteristics that are not obvious. The “Experiment Import Wizard” saves your experiment information as an HTML file. When you are entering a new experiment make sure the Genome Browser in the main GeneSpring window is displaying the genome the experiment refers to.

To initiate the Wizard, select File > “Experiment Import Wizard...”.

4.4.1 The Welcome panel of the GeneSpring Experiment Entry Wizard

![Image of Welcome panel](image)

Figure 4-5 The “Welcome to the GeneSpring Experiment Entry Wizard” panel

Clicking the “Help Pasting Data” button will take you to a web page with information on pasting experiments directly into GeneSpring. Pasting is very easy (if your file is set up correctly) but it is not very flexible. Please refer to 4.6 “Copying and Pasting Experiments” on page 4-71 for more information. The Experiment Wizard is very flexible, and correspondingly more complex.

The Welcome panel includes lists to remind you to create or gather your raw data files. There are five possible raw data files listed below; only the first one is necessary for loading an experiment. They should all be placed within the “Experiment” sub-folder of the relevant organisms described in 4.3.3 “Where do I put my data?” on page 4-13.
• Experimental data file(s), containing the genes’ signal strengths for each sample in the experiment
  • A file listing the positive controls
  • A file listing the negative controls
  • GIF or JPEG pictures to be associated with this experiment, or with particular samples within the experiment
  • GIF or JPEG pictures of the Microarray plates the experiment was done on

Click the “Next” button to proceed to the next panel. As you move to the next panel, a checkbox in the Wizard Navigator will change color. You can return to any of the previous panels, by clicking the check box of the panel you would like to view again. Occasionally you will get a dialog box telling you changes in a previous panel might have detrimental effects.)

4.4.2 The Data File Format panel

![Figure 4-6 The “Data File Format” panel](image)

This panel tells GeneSpring where to look for your data files, and what kind of format they will be in. There are a number of prefabricated experiment types. Choosing one of the specific types, such as “Affymetrix Text File” or “CLONTECH Text File” means some of the data file format information is already filled in for you. Choosing the “Two color experiment File” means you are using references, and the panel that asks about them will already indicate you have them. These prefabricated experiment types are included so you do not have to look at all of the possible wizard panels. If you are unsure which of the formats offered in the “What type of technology are you using?” box applies to you, then you probably want to choose the default setting, “Fully Custom…” which allows you to go through all of the wizard panels. The ten choices are illustrated below.
This manual will follow the Fully Custom screens. The screens may appear slightly differently on your screen, particularly if you have selected another “Data File Format”. Just skip the examples not applying to you. Please read carefully the screens that do.

At the moment, “Locally Accessible text files” is the only selectable option for the second drop-down menu.

Click the “Next” button to proceed to the next panel.
4.4.3 The Properties of Experiment Set panel

In the top box, enter the experiment name exactly as you want it to appear in the “Experiments” folder in the GeneSpring Navigator. This name must be unique. If the name is not unique, GeneSpring will not allow you to move on to the next panel. Enter all information carefully, as GeneSpring is spelling and case sensitive.

In the middle box, tell GeneSpring whether you want this experiment to appear in a subdirectory of the genome folder this experiment refers to. Clicking the “Yes” circle will cause another box to appear. Type in the name of any subdirectory you would like to use for this experiment. You may have more than one experiment within a folder.

In the bottom box, enter any comments or general notes you have about this experiment. These notes will be visible (and editable) in the Experiment Inspector. Please refer to GeneSpring User Manual 6.5, “Experiment and Interpretation Inspectors” for more information about that window.

Click the “Next” button to proceed to the next panel.
4.4.4 The Number of Arrays panel

This panel tells GeneSpring how many single arrays (or samples) combine to make this experiment. A single array is defined as each time a measurement is taken of your entire set of genes. If there was only a single set of measurements taken, select the “No” circle. If more than one set of measurements for your genes were taken, select the “Yes” circle. Selecting “Yes” in this panel will reveal a box to type in the number of arrays.

![New Experiment Wizard: yeast](image)

**Figure 4-9 The “Number of Arrays” panel after clicking “Yes”**

Enter the number of measurements that were taken of your gene set by typing the number in the “Number of Arrays” box. GeneSpring will not let you proceed if you click “Yes” but do not indicate how many Arrays/Samples there are.

Click “Next” to proceed to the next panel.
4.4.5 The Number of Parameters panel

![Image of the Number of Parameters panel]

**Figure 4-10 The “Number of Parameters” panel with four parameters typed in**

This panel tells GeneSpring how many parameters were used in this experiment, and what those parameters were. Briefly, a parameter is anything used to describe the condition or conditions of the experiment. A parameter consists of two or more parameter-values; for example breast cancer, lung cancer, and healthy could be parameter-values for the parameter “cancer”. For a more detailed description of parameters see Chapter 2, Experiment Parameters.

- Type the number of parameters involved in this experiment in the “Number of parameters” box. Changing the number in this box changes the number of lines given in the table below.
- Name each of your parameters in the right-hand column (labeled “Parameter Name”). You can tab forward (or use the cursor keys in some cases) to place the cursor in the next space. When you right-click this table, there is no pop-up menu allowing you to cut and paste. You can still cut and paste entries into the matrix fields by using the keyboard commands (for windows this is Ctrl+C and Ctrl+V). If you right-click one of the gray areas of this table, a pop-up menu will appear.

These pop-up menus allow you to cut and paste large sections of the table. You cannot proceed to the next panel until you have named all of your parameters. If you mis-typed the number of parameter values, just highlight over it and type in the correct number.
Select the Next button to continue.

### 4.4.6 The Parameter Characteristics panel

![Parameter Characteristics panel](image)

Figure 4-11 The “Parameter Characteristics” panel

In this panel you can define the parameters as being numbers, plotted on a log scale, and the units associated with them. In this example, Kryptonite concentration is a number with units in ppm. Use the scroll bars to view each parameter, clicking “Yes” or leaving blank items for each of the parameters set up in the previous panel. It is perfectly acceptable to answer “No” (by not selecting any options) to all of these questions.
4.4.7 The How to Display the Parameters panel

In this panel you tell GeneSpring what parameter types to use in the default interpretation. There are four possible choices as shown in Figure 4-12.

The default setting is “Denotes a non-continuous variable”, separating the data into discrete graphs viewed side by side on the screen (the non-continuous display). For more detailed information about all of these parameter displays see 2.2 “Parameter Display Options” on page 2-8.
4.4.8 The Parameter Values panel

In this panel you tell GeneSpring the parameter-values for each condition in the experiment. Initially blank, this screen has been filled in with the Parameter-values. A parameter-value is one of the possible values a variable can have. For example, in Figure 4-13, the parameter Kryptonite concentration has listed values of 0, 10, 20, 30, and 40 ppm. (For a more detailed explanation of parameters and how they can be used, please see Chapter 2, Experiment Parameters.) In the table given in the “Parameter Values” panel, each parameter you named has its own column. You must fill in every field in each column with the appropriate parameter-value for the samples named (in this example the samples are numbered 1-40) to the far left of the field. If there are more fields than fit in the panel, scroll bars will appear. You can cut and paste entries into the matrix fields by using the keyboard commands (for windows this is Ctrl+C and Ctrl+V). Pasting is highly recommend because the parameter-value entries are spelling and case sensitive. If you right-click one of the gray areas of this table, a pop-up menu will appear.

The pop-up menu resulting from right-clicking the parameter labels section of the table will say copy and paste columns. The pop-up menu resulting from right-clicking the sample labels section of the table will say copy and paste rows. The pop-up menu resulting from right-clicking the gray field in the upper left-hand corner of the table will say copy and paste all. These pop-up menus allow you to cut and paste large sections of the table. Once you have filled in every field in the table you can proceed to the next panel by clicking on the “Next” button. If there is an unfilled box, the “Next” button will remain disabled.
4.4.9 The Describe your Data Files panels

This panel tells GeneSpring where to find the experimental data file pertaining to each sample. The “Describe your Data Files” panels are large. Please double click the banner bar to expand the panel to fill your screen so you will not miss any of the possibilities.

| Sample #1 (Kryptonite concentration = 0) | YES_1.txt |
| Sample #2 (Kryptonite concentration = 10) | YES_2.txt |
| Sample #3 (Kryptonite concentration = 20) | YES_3.txt |
| Sample #4 (Kryptonite concentration = 30) | YES_4.txt |
| Sample #5 (Kryptonite concentration = 40) | YES_5.txt |
| Sample #6 (Kryptonite concentration = 0) | YES_6.txt |
| Sample #7 (Kryptonite concentration = 10) | YES_7.txt |
| Sample #8 (Kryptonite concentration = 20) | YES_8.txt |
| Sample #9 (Kryptonite concentration = 30) | YES_9.txt |
| Sample #10 (Kryptonite concentration = 40) | YES_10.txt |

Figure 4-14 Default setting for different files, same format

To begin describing your files to GeneSpring, you must select one of the options in the drop-down menu at the top of this panel. Details are provided in 4.4.9.2 “My samples are in multiple files that share a common format” on page 4-25.

Figure 4-15 File options in the “Describe your Data” panel

You have three selectable options to describe the files containing your data.
4.4.9.1 “All my samples are in one file”

First and easiest, if all of your samples are in one data file select “All my samples are in one file”. In the table at the bottom of the panel, fill in the field labeled “File Name” with the name of the text file containing your sample’s data, as in Figure 4-16. When your data is all in one file, the formats will all be the same. Be aware, as soon as you leave this panel, by clicking the “Next” button, the changes will be irrevocable. You may see the quick flutter of an error message reminding you of this.

4.4.9.2 “My samples are in multiple files that share a common format”

If your samples are in different files with exactly the same format, select the default setting, “My samples are in multiple files that share a common format”. Please see Figure 4-14.

Enter the name of the file containing the sample data for each experiment in the table. Each file should be entered in the white boxes of the column labeled “File Name” in the same row as its sample. If your data files are where GeneSpring expects them to be (i.e., in the directory set up in Figure 4-8) the names will appear in the large white box at the bottom of the screen labeled “Files present in the current data directory”. You can double click these names to insert those files into the “File Name” column. Each row will be filled in top-to-bottom order each time you double click a file name until all rows are filled. If your files are not shown in the “Files present in the current data directory” box, you may not have saved your files to the correct location. If you may need to recheck the “Properties of the Experiment Set” panel. You can select from the list of “already viewed” panels on the left side of the Wizard to view that panel again. See Figure 4-8 for an example.

If you have two files comprising a chip set you need to enter the names of both files separated by a semi-colon in the same entry blank. Please see 4.4.9.4 “If you need to put more than one file in a field” on page 4-26 for more details.

Data files have the same layout when the files for each and every sample have exactly the same number of columns, in the same order, containing the same type of data (for example, signal intensity or background readings for the experiment). Any variation, no matter how small, means your files do not have the same layout.

If all of your sample data is in the same file, and each have the same file layout, you may need to cut and paste the information into separate files or add columns to the file you already have. For example, a data file containing the signal intensities from sample 1 and sample 2 must have these results in two different columns. When this is done, the signal strength column in the data file pertaining to sample 1 is not in the same place as the column containing the signal strength for sample 2. This means the experimental data file layout for sample 1 is not the same as the layout in sample 2. An experiment reported in this way, with some, but not all of the samples in the experiment reported in the same data files cannot be considered to have the same data file layout. To tell...
GeneSpring your data is reported in this manner, answer “No” to the first two questions in the “Describe your Data” panel (the “Are all of your samples in the same data file?” question, and the “Do all the data files have the same layout?” question). Enter the name of the experimental data files containing each sample in the “File Name” column of the table. Now the table allows you to repeat a file name in multiple rows (unlike the non-repetition if you answer “Yes” to the “Do all the data files have the same layout” question). However, if you must use the same data files the same number of times, for example sample 1-4 could be named a.txt, sample 5-8 could be b.txt and 10-12 could be c.txt. To continue the same example, sample 1-4 could be a.txt, sample 5-6 could not be b.txt, sample 7-8 could not be c.txt, and sample 10-13 could be d.txt as the differing numbers of samples in each file implies a different number of columns and therefor a different layout. If you have more than one data file with differing column layouts, you will have to repeat all of the subsequent panels dealing with locating which column contains what information for each data file you name.

When you right-click the table in this panel of the Experiment Wizard, there is no pop-up menu allowing you to cut and paste. You can still cut and paste entries into the matrix fields by using the keyboard commands (for windows this is Ctrl+C and Ctrl+V). If you right-click one of the gray areas of this table, a copy and paste pop-up menu will appear. These pop-up menus allow you to cut and paste large sections of the table. Once you have filled in every field in the table you can proceed to the next panel by clicking on the “Next” button.

You may see a quick flutter of an error message if GeneSpring cannot find the correct folder in your directory. Look in the TaskBar if Genespring will not let you go to the next panel. If an error message such as “Oops... Can’t find the file:” appears use your file management system to create the correct folder and place a copy of your data file within it.

In this configuration of the “Describe your Data Files” panel, you need to click in the beige box in the “File Name” column, then double click the correct file name in the “Files present...” box. If the files names are not present in the box, please double check to make sure your files are saved in the correct folder within GeneSpring.

4.4.9.3 “My samples are in multiple files with different format”

If your samples are in various files that do not have exactly the same format, select “My samples are in multiple files with different format”.

You will not be able to continue until every field is filled and GeneSpring has verified the existence of those files.

4.4.9.4 If you need to put more than one file in a field

Occasionally when you have files of varying formats you will need to place more than one file in a sample field. To do this, you must place one file in the field in the normal fashion. Manually type in a semi-colon (;) after the file name. You must then hold down the control key while selecting the file you would like added to that same field.

You can do this with either the “My samples are in multiple files that share a common format” option or the “My samples are in multiple files with different format” option.
4.4.10 The Data File Header Lines panel

The first drop-down menu in this panel allows you to tell GeneSpring whether there are any Column titles in your experimental data files. If you do, choose “has a” line of column titles after. If you have any comment lines to discard, type the number of comment lines to be skipped the box. GeneSpring automatically skips blank lines, so you should not count blank lines among the lines to be skipped.

4.4.11 The Region Normalization panel

This panel allows you to employ region normalizations. If a sample in your experiment was performed on more than one array, or if there is some reason you want the sections on the arrays normalized individually, then you should answer, “Yes” to the question “Did each of your sample(s) use multiple arrays or sections of a single array that require separate normalization?”
You will need to enter the column of your experimental data file containing the region designation. Make sure the spelling and capitalization you enter is exactly the same as is used in the data file. (Copy and paste if you can to make sure the spelling and capitalization is identical.) If the region is the only entry in the region designation column, or if it is a suffix attached to the column’s entry, then you need to type all of the different region designators (the different suffixes or column entries defining which gene was in which region) in the “List all possible region column entries or suffixes” box. The different region designators must be separated by spaces, or else GeneSpring will read them all as one entry. After you have entered the region information click the “Next” button to proceed to the next panel. If the region designators used in your experimental data file are neither unique column entries, nor suffixes, see 4.3.2.6 “Entering region specifications when they are not specified in their own column or as suffixes within another column” on page 4-11 for how to import this information into GeneSpring. You will not be able to enter this experiment using the Wizard.

For a mathematical illustration of this normalizing option, please refer to 1.9 “Region Normalization” on page 1-16.
4.4.12 The Gene Name panel

This panel tells GeneSpring which column of your experimental data file contains the gene names, and whether the gene name is the only entry in its column. Enter the name or number of the column containing the gene name in the box labeled “Enter the gene column name”. If you are entering the column number, count the columns from left to right, starting from one. Make sure the spelling and capitalization is perfectly consistent with your file when you are entering the column names. The second question, “Does this column contain only the desired gene name without suffixes or prefixes?”, should be answered positively only if the gene name reported in the experimental data is exactly like the gene name listed in the table of genes file defining the genome. If there are prefixes, suffixes, or region designators (which are frequently noted as prefixes or suffixes in the gene column) then select “Yes” in the second question. If you do this the next two panels presented to you will be the “Gene Name Prefix Removal” panel and the “Gene Name Suffix Removal” panel. If fewer than 10% of the gene names match your current genome, you will get a warning box.
4.4.13 The Gene Name Prefix Removal panel

This panel allows you to remove prefixes from the gene names in the experimental data file, so the gene names match the gene names given in the list of genes defining the genome. If your genes do not have prefixes it is acceptable to leave the answers to both questions “No”.

You can remove one of two types of prefixes from your gene names. The first is when every gene has the same string of characters prepended to it, select the “Yes” circle for the first question, “Does the name appearing in the gene name column have a fixed unchanging prefix you want removed?”. Enter the string of characters prepended to your gene names in the “Enter fixed prefix” box that appears.

The other type of prefix is not the same for every gene it prepends, but it always ends with the same character. If this is the case, select the “Yes” circle of the second question, “Does the name appearing in the gene name column have a prefix ending in a particular character or characters?”. Enter the character marking the end of the prefix in the box labeled “Enter prefix marker character(s)”*. There may be multiple different markers indicating the end of the prefix. If this is the case, enter them all in the “Enter prefix marker character(s)” box. Do not separate multiple markers in any way, anything you use to separate the characters, including a space, will be considered a prefix marker character and be removed from the gene name, along with anything preceding it. Make sure when you are entering a set prefix or a prefix marker character you get the spelling and capitalization exactly correct.

Click the “Next” button to proceed to the next panel.
4.4.14 The Gene Name Suffix Removal panel

This panel allows you to remove suffixes from the gene names in the experimental data file, to make the gene names given there match the gene names given in the list of genes defining the genome. If your gene names do not have suffixes, it is acceptable to leave the answers to both questions “No”.

If your gene names have suffixes to remove, the suffixes can be one of two types. The first is a “set” suffix; this means every gene with a suffix has the same string of characters appended to it. Click the “Yes” circle under the question “Does the name appearing in the gene name column have a fixed, unchanging suffix you want removed?”.

The other type of suffix is not the same for every gene name it appends to, but it always starts with the same character. If this is the case, select the “Yes” circle of the second question, “Does the name appearing in the gene name column have a suffix that begins in a particular character or characters?” In the box that appears, labeled “Enter suffix marker character(s)”, enter the character marking the beginning of the suffix. There may be multiple different markers indicating the beginning of a prefix. If this is the case, enter them all in the “Enter suffix marker character(s)” box. Do not separate multiple marker characters in any way. Anything you use to separate the characters, including empty spaces, will be considered a suffix marking character and will be removed from the gene name, along with any characters following it. Make sure when you are entering a set suffix or a suffix marker character you get the spelling and capitalization exact. Click the “Next” button to proceed to the next panel.
4.4.15 The Data Column Location panel

![Figure 4-23 The “Data Column Location” panel](image)

This panel tells GeneSpring which column(s) of your experimental data files contains the genes’ raw data. Enter the name or number of the column containing raw data in the “Enter data column name” box. Make sure to use the correct spelling and capitalization for this entry. If your data file includes a column containing the background signal to be subtracted from the gene’s raw data, in the second question (“Do your data files contain a column representing background signal strength?”) select the “Yes” circle. Enter the name of this column or its number in the white “Data Background Column” on the right of Figure 4-23. Again, beware of spelling and capitalization errors. This panel will not let you proceed to the next panel until you have entered a column name or number for the raw data column for every sample (row), and for the background column (if present).
4.4.16 The Control Channel Values panel

If you have control channel values for each gene on your array then you can use this information to normalize your genes. See Chapter 1, Normalizing Options for more information regarding how this normalization works.

If you do not have a control for each gene (if you did a single-color experiment, this is probably the case) you should leave the “No” circle selected and proceed to the next panel.

If you do have control channel values, select the “Yes” circle and enter the name(s) of the column (or its number) containing the control channel signals in the “Control Channel Column” box. If your experiment took a reading of the background for the control channel values, change the selection in the bottom question to “Yes”. Then, enter the column name(s) (or number(s)) of the column containing the control channel background signal. When you enter column names make sure you use the correct spelling and capitalization.
Figure 4-25 The expanded “Control Channel Values” panel

Proceed to the next panel by clicking the “Next” button.
4.4.17 The Flags panel

If your experimental data contains a column indicating whether the experiment worked for each gene, GeneSpring can incorporate this data. Select the “Yes” circle.

- In the first column, enter the column name(s) (or number(s)) of the column(s) containing the pass-fail information in the “Flag column name” box.

- In the second column, “Passed Designator”, enter the value given in the Flag column indicating the experiment worked for any particular gene. Frequently, the designator for good data is “P” for Present/Passed or “O” for OK.

- In the third column, “Marginal Designator”, enter the value given in the Flag column indicating the experiment might have worked for any particular gene. Uncertain or marginal data is normally indicated by an “M”.

- In the fourth column, “Absent Designator”, enter the value given in the Flag column indicating the experiment did not work for any particular gene. Failed or absent data is normally indicated by an “A”.

When you are entering a column name, be sure to use the spelling and capitalization used in your experimental data file.

![Figure 4-26 Filled “Flag” columns](image)

If you have many rows and your designators are the same in every file use the “Guess the rest” button to fill down the table.

Click the “Next” button when you are ready to proceed to the next panel.
4.4.18 The Sample Photos panel

This panel tells GeneSpring if you have any pictures you wish to associate with any or all of the samples. Pictures are nice, but they are not necessary. If you do not have any, leave the “No” circle selected and proceed to the next panel.

If you have one or more pictures to associate with your sample, select the “Yes” circle. The panel will expand. If you have a picture already in the correct directory to associate with every sample, GeneSpring will display the file name(s) in the lower right-hand corner of the main window. In the table labeled “GIF File Name” enter the complete file name of the picture associated with the sample by double clicking one of the file names or typing in each file name manually. The picture must be a GIF or a JPEG file. If one of your samples does not have a picture associated with it, leave its field blank. GeneSpring will use the picture associated with the next closest sample.

The easiest way to fill in this table is to have all of your GIF or JPEG files in the experiment directory. Then the file names will appear in the white box at the bottom of the panel. Just double click on each picture in the correct order. When you right-click the “GIF File Name” table in this panel of the Experiment Wizard, there are pop-up menus allowing you to cut and paste. If you right-click one of the gray areas of this table, a pop-up menu will appear, from which you can select copy and paste options.
You can still cut and paste entries into the matrix fields by using the keyboard commands (for Windows this is Ctrl+C and Ctrl+V).

Proceed to the next panel by clicking the “Next” button.

### 4.4.19 The Array Photos panel

In this panel you tell GeneSpring if you have any pictures of the array plates used. Microarray pictures are nice, but not necessary. If you don’t have any, leave the “No” circle selected and proceed to the next panel. To associate Array Pictures with the samples, select the “Yes” circle for the question, “Do you have any pictures of the microarray plate(s)?” A table appears. In the “GIF File Name” column enter the complete name of the file containing the array picture to be specifically associated with the sample listed in the left-hand column. If you have an array picture for every sample GeneSpring will display it when you double click the picture in the lower right-hand corner of the main GeneSpring window. Array pictures must be in either GIF or JPEG format.

When you right-click the table in this panel of the Experiment Wizard, there are pop-up menus allowing you to cut and paste. You can also cut and paste entries into the matrix fields by using the keyboard commands (for Windows this is Ctrl+C and Ctrl+V).

The pop-up menu resulting from right-clicking the “GIF File Name” label, allows you to copy and paste columns. The pop-up menu resulting from right-clicking the experiment labels section of the table, allows you to copy and paste rows. The pop-up menu resulting from right-clicking the gray field in the upper left-hand corner of the table, allows you to copy and paste all.

These pop-up menus allow you to cut and paste large sections of the table. Proceed to the next panel by clicking the “Next” button.
4.4.20 The RT – PCR Experiments panel

This panel tells GeneSpring whether the data you are loading comes from a RT-PCR experiment. RT-PCR is a technology for measuring expression levels, it reports these measurements in a different form than the standard array technologies. Instead of reporting expression values it reports:

$$-\log_2(\text{expression value})$$

If you have not dealt with RT-PCR experiments or have not heard of them before, leave the “No” circle selected, and proceed to the next panel. If you are using RT-PCR technology, select the “Yes” circle.

4.4.21 The Normalizations: Negative Controls panel

This panel tells GeneSpring if you have any genes designated as negative controls on your array, and if you want to normalize your sample using this data. You typically have negative controls when there is DNA from a different genome than the one you are investigating in your array. To indicate you have negative controls to use for normalizing, select the “Yes” circle. This normalization method takes the average signal intensities for all of the negative controls and subtracts this number from the signal intensity of each gene. For more info about this normalization option, see Chapter 1, Normalizing Options. If you do not have negative controls, or do not want to normalize your sample using the data from them select the “No” circle.

Answering, “Yes” to the first question, “Do you have any genes designated as negative controls?” initiates a second question. If you are using negative controls you must have a file listing them, one gene name per line. This file should be in the same sub-directory as your experimental data. In the “Negative controls file name” box enter the name of the file listing your negative controls.
For a mathematical illustration of this normalizing option, please refer to 1.2 “Normalize to Negative Controls” on page 1-2.

### 4.4.22 The Normalizations: Control Channel Values panel

![Image](Image.png)

Figure 4-32 The “Normalizations: Control Channel Values” panel, “Yes” selected

You will only see this panel if you have already told GeneSpring your sample has control channel values for each gene. If you have control channel values for each gene to indicate the trust you have in the experimental data for each gene, you probably want to normalize the genes by dividing their signal strength by the control channel’s signal strength. If you have a background signal for either or both of these values, it is subtracted from the signal intensities before they are divided. For more information on this normalization option, see Chapter 1, Normalizing Options. If you wish to use this normalization, select the “Yes” circle. If you do not wish your data to be normalized using the control channel values leave the “No” circle selected.

If you are using your control channel values for normalization, you need to enter the minimum reference signal to be used in the normalization. This is because sometimes the control channel value is very low and would artificially inflate the noise for its gene. Indicate the minimum value you would be willing to divide a gene’s signal by in the “Minimum control channel signal strength” box.

If you are not using your control channel values for normalization, then you are using them to indicate the trustworthiness of the experimental data for each gene. Indicate the minimum value a reference must have for you to consider the data for the gene it is associated with valid in the box labeled “Minimum confidence level”.

For a mathematical illustration of this normalizing option, please refer to 1.3 “Normalize to Control Channel Values for Each Gene” on page 1-3.
4.4.23 The Normalizations: Positive Controls panel

This panel tells GeneSpring if you have any genes designated as positive controls on your array and if you want to normalize your sample using this information. You typically have positive controls when there is DNA from a different genome than the one you are investigating and you’ve added a known quantity of that DNA to your sample.

If you do not want to normalize your sample using positive controls leave the “No” circle selected.

To indicate you have positive controls for normalization, select the “Yes” circle. This normalization method takes the average signal intensities of all of the positive controls and divides each gene’s signal intensity by that number. For more information about this normalization option see Chapter 1, Normalizing Options.

If you are using positive controls you must have a file specifying what the positive controls are called, listing the gene names one per line. This file should be in the same sub-directory as your experimental data. In the “Positive controls file name” box, enter the complete name of the file listing your positive controls.

Sometimes, something will go wrong with the positive controls and you will get very low values for them, which should not be used for normalization. Indicate the minimum average value that the positive controls must have so as not to pollute a gene’s signal with noise. The default setting for the cut-off value is 10.

For a mathematical illustration of this normalizing option, please refer to 1.4 “Normalize to Positive Controls” on page 1-5.
4.4.24 The Normalizations: Each Sample to Itself panel

In this panel you tell GeneSpring if you want to normalize your data by making the median of all of your measurements 1, for each single sample in your experiment. (If you have not already performed normalizations on your data you generally want to use this normalization option.) To indicate you want to normalize each sample to itself, select the “Yes” circle. Another question will appear.

Sometimes something will go wrong with the experiment and you will get very low values for everything. In the “Enter lower cut-off value” box indicate the cut-off value. This number will be used by GeneSpring to not raise all of the signal strength values up to a median of 1 if their average is below this number.

For a mathematical illustration of this normalizing option, please refer to 1.5 “Normalize Each Sample to Itself” on page 1-6.
4.4.25 The Normalizations: Each Sample to a Hard Number panel

In this panel you tell GeneSpring if you want to normalize your samples to a value you enter. You would normally only use this function if you have pre-normalized data, such as data prepared with Affymetrix’s Global Scaling. In that instance, you would want to divide all data by 2500 (or whatever number you chose to normalize by in the Affymetrix software.) You will need to do this because the GeneSpring analysis algorithms assume your data is normalized to a median of 1.

4.4.26 The Normalizations: Each Gene to Itself panel

In this panel you tell GeneSpring if you want to normalize each gene to itself, so the median of all of the measurements taken on that gene is 1. If you are not doing a two-color experiment you generally want to do this. If you have fewer than 5 experiments, however, you probably want to normalize to a specific experiment instead.)

Sometimes something will go wrong with a measurement on a particular gene and you will get anomalously low values. Indicate the lower cutoff value for the median in the box below. This will instruct GeneSpring not to change the measurements on a given gene up to 1 if their median falls below this cutoff.

Enter lower median cutoff value: 0.01

In this panel you tell GeneSpring if you want to normalize each gene to itself, so the median of all of the measurements taken for the gene is 1. If you are not doing a two-color experiment you generally want to do this, so the default setting for this panel is to perform this normalization.

If you do not wish to employ this normalization select in the first question, the “No” circle.

If you wish to use this normalization, there is a second question. Sometimes something will go wrong with the experiments and all of the values for a particular gene are very low, in which case it will artificially inflate the noise of the gene if you normalize those values up to a median of 1.
Specify this cut-off by entering a number in the “Enter lower median cut-off value” box. The default setting for the cut-off value is 0.01.

Normalizing each gene to itself is optimal for more than five samples, as with less than five the display becomes unintuitive. Generally the better option for five samples or less is to do normalization against a particular sample.

For a mathematical illustration of this normalizing option, please refer to 1.7 “Normalizing Each Gene to Itself” on page 1-8.

4.4.27 The Normalizations: All Samples to Specific Samples panel

![Figure 4-37 The “Normalizations: All Samples to Specific Samples” panel](image)

This panel tells GeneSpring if you want to normalize each sample in the experiment to a single sample within the set. Normalizing each gene to itself is often preferable to this normalization. If you wish to normalize your data in this way, select the “Yes” circle. Another question appears. Sometimes a gene’s signal strength in the sample being normalized to is anomalously low. Enter the lowest value you are willing to use for normalizations in the “Enter lower reference cut-off value” box.

In the enter sample number box you can normalize multiple samples to several samples. You can also normalize several samples to several samples. You can normalize multiple samples to multiple different samples through a code like [1;2;3]1;2[3;4;5]3;4 which means normalize samples 1, 2 and 3 to 1 and 2, and 4, 5 and 6 to 3 and 4. Please see 1.8.1 “Required Syntax for Normalization to Specific Samples” on page 1-11 for more information regarding the syntax to use in this panel.

For a mathematical illustration of this normalizing option and several examples, please refer to 1.8 “Normalizing All Samples to Specific Samples” on page 1-11.
4.4.28 The Graphics Specifications panel

![Image: The “Graphics Specifications” panel]

**Figure 4-38 The “Graphics Specifications” panel**

4.4.28.1 Defining Trust

The upper section of this panel tells GeneSpring what the Colorbar intensity scale should be, and the relative intensity values to be graphed on the y-axis in the graph display. The intensity of the colorbar in GeneSpring indicates how reliable the data for each gene is. Indicate a raw, very reliable (a high signal strength) signal strength value, an average (a medium signal strength) value, and an unreliable (a low signal strength) value in the three boxes. Any gene with a signal strength 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 dull in color. The medium signal value gives the value for the mid-point of the colorbar, and genes with an average signal strength are colored halfway between the two color extremes.

For more information on how trust is expressed in the Genome Browser, please see the *GeneSpring User Manual* 4.1, “The Colorbar”.

4.4.28.2 Defining default x and y values

The middle section of Figure 4-38 allows you to inspect the genes’ expression profiles more closely from the Genome Browser. As GeneSpring does not graph the entire y-axis (the expression level axis), but only the portion most genes profiles fall into you will need to set the defaults for that portion. In the lower two boxes indicate the range of expression levels GeneSpring should graph. The values indicated here can be altered within GeneSpring (look in **View > Change experiment interpretation**...). Here you are simply setting the defaults.

4.4.28.3 Defining Negative Values to Zero

The bottom section in Figure 4-38 asks if you would like to force negative values to zero. Forcing all of the negative numbers to zero converts all the negative values to zero after all the normalizations have been implemented and after the genes that do not pass the Pass-Fail vote have been thrown out (this happens before any normalization is applied by GeneSpring).
4.4.29 The Finish panel

When you click the “Load Now” button all of the answers you gave in the previous Experiment Wizard panels are saved in an HTML file.

If GeneSpring is unable to load the data, you will get an error message with a list of the unrecognized genes that caused it not to load.
4.5 Installing from a Text File

This is possibly the most tedious and unforgiving of the experiment loading methods. However, it is necessary to be at least slightly familiar with the methods, as you will need to change the “.experiment” file (or re-enter your experiment through another method) when you need to make changes to the experiment. Hopefully, with the new features in GeneSpring User Manual Chapter 5, “Changing Your Experiment”, entering or altering your “.experiment” file will soon be obsolete.

Generally, an “.experiment” file is a text file describing where the data file(s) are, what their format is, what the parameters for the experiment are, and what normalizations need to be done. You can also specify pictures to be associated with the files, and various other things. Each line in an “.experiment” file is either blank or a line of the form object-name space-colon-space object-value:

```
Object-name : object-value
```

An example of this is:

```
name : Yeast extraterrestrial studies
```

Obviously, “name” is the object-name and “Yeast extraterrestrial studies” is the object-value. The object value can be thought of as the answer to the question posed by the object-name. In the “.experiment” file the order of lines is not significant, but the case (lower or upper case) of letters is significant. The spelling, especially of the object-name is also significant. Usually, when an experiment looks like it is not installed correctly it is because of a spelling or capitalization error.

Due to the complexity of the information contained in the “.experiment” file, this section is designed to help you create a “.experiment” file for a particular experiment, rather than explaining exactly what each possible answer means. There are two examples following each question. The first is the generalized form of the answer, including the generalized object-name and what sort of response constitutes a correct object-value. The second (bold-faced) example is an example of an actual answer to the question. A fictitious experiment, “Yeast extraterrestrial studies”, is used as the example experiment throughout this chapter. A complete “.experiment” file for the “Yeast extraterrestrial studies” experiment is given in this chapter. (Please refer to 4.5.20.2 “An example “.experiment” file” on page 4-66.) There are eighteen sections and thirty eight questions which must be answered in their presented order.

### 4.5.1 Define Your Experiment

1. Enter the name of your experiment or samples as you wish it to appear in the GeneSpring menu system.

```
name : Your experiment name here
```

```
name : Yeast extraterrestrial studies
```
2. How many samples are there in the experiment you have just named? A sample is defined as each time a numerical measurement is taken for your entire set of genes.

   Experiments : The number of samples

   Experiments : 40

3. How many different parameters were taken? A parameter is used to describe the condition (or conditions) in the experiment. See Chapter 2, Experiment Parameters for a more thorough description of parameters.

   Parameters : The number of parameters

   Parameters : 4

4. Name the parameters:

   Parameter#Name : Name of the indicated parameter

   Make sure to name each of the parameters enumerated in question 3.

   Parameter1Name : Kryptonite concentration
   Parameter2Name : Variety of yeast
   Parameter3Name : Test repeat number
   Parameter4Name : Andromeda Strain infection

4.5.2 Define Your Parameters

In number 4 of section 4.5.1, Define Your Experiment you named and numbered each parameter. They will be referred to by their number for the remainder of this example. For reasons of brevity, the questions in this section are all phrased in reference to parameter 1, but you should answer each question for every parameter enumerated in question 4.

5. If there are units associated with parameter 1, name them.

   Parameter#Units : name of the units associated with the indicated parameter

   If a parameter does not have a unit name associated with it, either do not enter the line “Parameter#Units : ” for the parameter without units, or enter the object-name “Parameter#Units” and the space-colon-space, but leave the name of the units (the object-value) blank.

   Parameter1Units : ppm
   Parameter2Units :
   Parameter3Units :
   Parameter4Units :
6. Is parameter 1 defined by a number, i.e. are the parameter values associated with parameter 1 numbers? If the answer is yes, enter “true” after “Parameter1IsNumber:” and if the answer is no, enter “false”.

   Parameter#IsNumber : enter either true or false
   Parameter1IsNumber : true
   Parameter2IsNumber : false
   Parameter3IsNumber : true
   Parameter4IsNumber : false

7. This question is only applicable to those parameters defined by a number. (I.e. for those parameters for whom the answer to question 6 is true.) Would you like the number defining parameter 1 graphed on a logarithmic scale? If this answer is yes, enter “true” as the object-value following “Parameter1IsLogarithmic”. If the answer is no, either do not enter the “Parameter1IsLogarithmic:” line, or type “false” as the object-value. The answer to this question is automatically false if a number does not define the parameter in question.

   Parameter#IsLogarithmic : enter either true or false
   Parameter1IsLogarithmic : false
   Parameter2IsLogarithmic : false
   Parameter3IsLogarithmic : false
   Parameter4IsLogarithmic : false

8. Of the following four choices, choose the most appropriate display for parameter 1. (You may alter your choice within GeneSpring, the display you are indicating here will simply be the default display). See Chapter 1, Normalizing Options for more details about each of these display options.

   • Parameter 1 is continuous. This means when you are graphing the data by this parameter the data points will be connected together by lines instead of being graphed as discrete points. Follow “Parameter1IsContinuous” with true if this is how you wish the parameter to be graphed. If one of the other possibilities seems more correct for parameter 1, either enter “false” as the object-value, or do not include the line beginning with “Parameter1IsContinuous”.

   Parameter#IsContinuous : either true or false
   Parameter1IsContinuous : true
   Parameter2IsContinuous : false
   Parameter3IsContinuous : false
   Parameter4IsContinuous : false
• Parameter 1 is a category (or set of categories) and you wish to color code the display by their membership. If this is the display you wish for parameter 1, answer the object-name lines, “Parameter1IsContinuous”, “Parameter1IsSet”, and “Parameter1IsRepeat” all with the object-value “false”.

This is the case for parameter 2 in the Yeast cancer time series experiment.

• Parameter 1 is a replicate parameter by which you do not wish to distinguish information graphically. Follow “Parameter1IsRepeat” with the object-value “true” if this is how you wish this parameter to be graphed. If one of the other possible parameters interpretations is correct for parameter 1, either enter “false” as the object-value, or do not include the line beginning with “Parameter1IsRepeat”.

Parameter#IsRepeat : either true or false

Parameter1IsRepeat : false
Parameter2IsRepeat : false
Parameter3IsRepeat : true
Parameter4IsRepeat : false

• You wish to use parameter 1 to separate the data into discrete graphs viewed next to each other on the same screen. This is a non-continuous parameter. Follow “Parameter1IsSet” with the object-value “true” if this is how you wish this parameter to be displayed. If one of the other possibilities seems more correct for parameter 1, either enter “false” as the object-value, or do not include the line beginning with “Parameter1IsSet”.

Parameter#IsSet : either true or false

Parameter1IsSet : false
Parameter2IsSet : false
Parameter3IsSet : false
Parameter4IsSet : true

9. Enter the number or label applicable to each sample, as it is associated with parameter 1. This is where you tell GeneSpring what each condition means, as far as each parameter is concerned.

Parameter#Experiment# : either a value or a name associated with both the parameter indicated and the sample indicated.

For each parameter you must indicate a label to associate with every condition.

Parameter1Experiment1 : 0
Parameter1Experiment2 : 10
Parameter1Experiment3 : 20
Parameter1Experiment4 : 30
Parameter1Experiment5 : 40
Parameter1Experiment6 : 0
Parameter1Experiment7 : 10
Parameter1Experiment8 : 20
Parameter1Experiment9 : 30
Parameter1Experiment10 : 40
Parameter1Experiment11 : 0
Parameter1Experiment12 : 10

Parameter2Experiment1 : A
Parameter2Experiment2 : A
Parameter2Experiment3 : A
Parameter2Experiment4 : A
Parameter2Experiment5 : A
Parameter2Experiment6 : B
Parameter2Experiment7 : B
Parameter2Experiment8 : B
Parameter2Experiment9 : B
Parameter2Experiment10 : B
Parameter2Experiment11 : A
Parameter2Experiment12 : A

Parameter3Experiment1 : Test 1
Parameter3Experiment2 : Test 1
Parameter3Experiment3 : Test 1
Parameter3Experiment4 : Test 1
Parameter3Experiment5 : Test 1
Parameter3Experiment6 : Test 1
Parameter3Experiment7 : Test 1
Parameter3Experiment8 : Test 1
Parameter3Experiment9 : Test 1
Parameter3Experiment10 : Test 1
Parameter3Experiment11 : Test 1
Parameter3Experiment12 : Test 1
Parameter3Experiment13 : Test 1
Parameter3Experiment14 : Test 1
Parameter3Experiment15 : Test 1
Parameter3Experiment16 : Test 1
Parameter3Experiment17 : Test 1
Parameter3Experiment18 : Test 1
Parameter3Experiment19 : Test 1
Parameter3Experiment20 : Test 1
Parameter3Experiment21 : Test 2
Parameter3Experiment22 : Test 2
Parameter3Experiment23 : Test 2
Parameter3Experiment24 : Test 2

Parameter4Experiment1 : healthy
Parameter4Experiment2 : healthy
Parameter4Experiment3 : healthy
Parameter4Experiment4 : healthy
Parameter4Experiment5 : healthy
Parameter4Experiment6 : healthy
Parameter4Experiment7 : healthy
Parameter4Experiment8 : healthy
Parameter4Experiment9 : healthy
Parameter4Experiment10 : healthy
Parameter4Experiment11 : Andromeda strain
Parameter4Experiment12 : Andromeda strain
Parameter4Experiment13 : Andromeda strain
...

In order to illustrate how to write all four of the possible parameter displays, the Yeast extraterrestrial study is a fairly large experiment, with many samples, as well as many parameters. This makes the entry for question 9 extremely long. You may well have a much smaller and less complex set of notations to write down.

**4.5.3 Describe your Data Files**

10. Are all of your samples in the same data file? If so enter this:

    DataFileName : complete name of the file containing your experimental data

    DataFileName : array.txt

If even one of your experiment’s samples are in a separate file from the rest, you must specify a separate file name for each sample.

    Experiment#FileName : complete name of the file containing the data from the sample indicated

    Experiment1FileName : 1A0.txt
    Experiment2FileName : 1A10.txt
    Experiment3FileName : 1A20.txt
    Experiment4FileName : 1A30.txt
    Experiment5FileName : 1A40.txt
    Experiment6FileName : 1B0.txt
    Experiment7FileName : 1B10.txt
    Experiment8FileName : 1B20.txt
    Experiment9FileName : 1B30.txt
    Experiment10FileName : 1B40.txt
    Experiment11FileName : 1AndromedaA0.txt
    Experiment12FileName : 1AndromedaA10.txt
    Experiment13FileName : 1AndromedaA20.txt
    Experiment14FileName : 1AndromedaA30.txt
    Experiment15FileName : 1AndromedaA40.txt
    Experiment16FileName : 1AndromedaB0.txt
### 4.5.4 Data File Header Lines

If you have more than one data file, and they have different column layouts, then you must answer these questions for *every* experiment/sample data file you have.

11. Does your data file have one or more headlines not containing experimental data?

```
Headlines : number of headlines in the data file
```

If your data files all use different layouts, but all of them have the same number of headlines, you may use the general object-name given above, rather than entering the number of headlines for each data file. If you have more than one data file, with different numbers of headlines use the object-name given below. If you are doing this, make sure to indicate the number of headlines for every sample.

```
Experiment#Headlines : number of headlines in the data file of the experiment indicated
```

<table>
<thead>
<tr>
<th>Experiment</th>
<th>File Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment17</td>
<td>1AndromedaB10.txt</td>
</tr>
<tr>
<td>Experiment18</td>
<td>1AndromedaB20.txt</td>
</tr>
<tr>
<td>Experiment19</td>
<td>1AndromedaB30.txt</td>
</tr>
<tr>
<td>Experiment20</td>
<td>1AndromedaB40.txt</td>
</tr>
<tr>
<td>Experiment21</td>
<td>2A0.txt</td>
</tr>
<tr>
<td>Experiment22</td>
<td>2A10.txt</td>
</tr>
<tr>
<td>Experiment23</td>
<td>2A20.txt</td>
</tr>
<tr>
<td>Experiment24</td>
<td>2A30.txt</td>
</tr>
<tr>
<td>Experiment25</td>
<td>2A40.txt</td>
</tr>
<tr>
<td>Experiment26</td>
<td>2B0.txt</td>
</tr>
<tr>
<td>Experiment27</td>
<td>2B10.txt</td>
</tr>
<tr>
<td>Experiment28</td>
<td>2B20.txt</td>
</tr>
<tr>
<td>Experiment29</td>
<td>2B30.txt</td>
</tr>
<tr>
<td>Experiment30</td>
<td>2B40.txt</td>
</tr>
<tr>
<td>Experiment31</td>
<td>2AndromedaA0.txt</td>
</tr>
<tr>
<td>Experiment32</td>
<td>2AndromedaA10.txt</td>
</tr>
<tr>
<td>Experiment33</td>
<td>2AndromedaA20.txt</td>
</tr>
<tr>
<td>Experiment34</td>
<td>2AndromedaA30.txt</td>
</tr>
<tr>
<td>Experiment35</td>
<td>2AndromedaA40.txt</td>
</tr>
<tr>
<td>Experiment36</td>
<td>2AndromedaB0.txt</td>
</tr>
<tr>
<td>Experiment37</td>
<td>2AndromedaB10.txt</td>
</tr>
<tr>
<td>Experiment38</td>
<td>2AndromedaB20.txt</td>
</tr>
<tr>
<td>Experiment39</td>
<td>2AndromedaB30.txt</td>
</tr>
<tr>
<td>Experiment40</td>
<td>2AndromedaB40.txt</td>
</tr>
</tbody>
</table>
4.5.5 Gene Names

12. Which column of your data file contains the gene name?

GeneColumn : number of the column the gene name is found in

    GeneColumn : 1

If your data files all have a different column layout, but all of them have the gene name in the same column, you may use the general object-name given above, rather than entering the column number of the gene name for each data file. If you have more than one data file with different column layouts, and they have different columns containing the gene name, use the object-name given below. If you are doing this, make sure to indicate the column containing the gene name for every sample.

    Experiment#GeneColumn : number of the column the gene name is found in, for the experiment indicated

    Experiment1GeneColumn : 2
    Experiment2GeneColumn : 3
    Experiment3GeneColumn : 2

... 

4.5.6 Explain to GeneSpring how to locate only the Gene Name

These questions are only applicable if the column containing the gene name contains other notations as well, notations not occurring in the list of genes defining the genome. If column containing the gene names in your data file(s) only contains the gene name as it appears in the table of genes file or the GenBank/EMBL file defining this genome, skip these two questions and do not enter the lines associated with them in your ".experiment" file.

13. GeneSpring can remove a set suffix from a gene name. A set suffix is a fixed string of characters which appear frequently at the end of your genes.

RemoveGeneSuffix : exact suffix you wish removed from the gene name

    RemoveGeneSuffix : _at

14. GeneSpring can remove the entire notation following a slash (/), including the slash itself. To do this, enter "true" as the object-value. To ignore this ability, thus leaving the gene name alone either enter "false" as the object-value after "RemoveSlash : " or do not include this line in your ".experiment" file.

RemoveSlash : either true or false

    RemoveSlash : true
4.5.7 **Explain to GeneSpring How to Read the Region Specifications**

Skip these questions, and their associated entries in the “.experiment” file, if the samples in your experiment did not involve multiple arrays or sections of arrays needing to be normalized separately.

15. If your experiment used multiple arrays, or sections of arrays, needing to be normalized separately, indicate to GeneSpring which column of your data file indicates the region of the array, and/or which array a particular gene reading came from.

   RegionColumn : number of the column the region specification is found in

   **RegionColumn : 1**

If your data files all have a different column layout, but all of them have the region specification in the same column, you may use the general object-name given above, rather than entering the column number of the region specification for each data file. If you have more than one data file with different column layouts, and they have different columns containing the region specification, use the object-name given below. If you are doing this, make sure to indicate the column containing the region specification for every sample.

   Experiment#RegionColumn : number of the column the region specification is found in, for the experiment indicated

   **Experiment1RegionColumn : 1**
   **Experiment2RegionColumn : 2**
   **Experiment3RegionColumn : 1**
   . . .

4.5.7.1 **The required layout file for Region Specifications**

16. If you have region specifications you must have a layout file. (See 4.3.2.4 “The Layout file” on page 4-7 for everything this file can or should contain.) Tell GeneSpring where to find this file:

   Layout : complete name of the layout file

   **Layout : AffyYeastLayout4.txt**

4.5.8 **Locate the Data Column**

17. Which column of your data file contains the raw data reading for Sample 1?

   Experiment#IntensityColumn : number of the column containing the raw data for the sample indicated

   **Experiment1IntensityColumn : 4**
   **Experiment2IntensityColumn : 9**
   **Experiment3IntensityColumn : 14**
   **Experiment4IntensityColumn : 19**
If your data is all in the same file you will have to indicate the raw data column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the raw data. On the other hand, if you have separate data files, with the same column containing the raw data you may use the general object-name given below, rather than entering the column number of the raw data for each file.

Experiment5IntensityColumn : 24  
Experiment6IntensityColumn : 29  
Experiment7IntensityColumn : 34  

If your data is all in the same file you will have to indicate the background reading column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the background data. On the other hand, if you have separate data files, with the same column containing the background data you may use the general object-name given below, rather than entering the column number of the background data for each file.

IntensitColumn : number of the column containing the signal intensity data

IntensityColumn : 7

If your data file has a column indicating the background signal, tell GeneSpring which column contains that information. If your data does not have a background reading, skip this question, and the associated “.experiment” file entry.

Experiment1IntensityBackColumn : 5  
Experiment2IntensityBackColumn : 10  
Experiment3IntensityBackColumn : 15  
Experiment4IntensityBackColumn : 20  
Experiment5IntensityBackColumn : 25  
Experiment6IntensityBackColumn : 30  
Experiment7IntensityBackColumn : 35  

If your data is all in the same file you will have to indicate the background reading column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the background data. If, on the other hand, you have separate data files, with the same column containing the background data you may use the general object-name given below, rather than entering the column number of the background data for each file.

IntensityBackColumn : number of the column containing the background reading

IntensityBackColumn : 8
4.5.9 The Control Channel Value

These questions only apply if your sample has a control channel, which is generally only applicable to two color experiments, such as Incyte or Sentini experiments. If your data does not have control channel values, skip this section and the associated “.experiment” file entries.

19. If your data has control channel values, which column of your data file gives the reference value? If your data does not have control channel values, skip this question, and the associated “.experiment” file entry.

Experiment#ReferenceColumn : number of the column containing the control channel values for the experiment indicated

<table>
<thead>
<tr>
<th>Experiment</th>
<th>ReferenceColumn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment1</td>
<td>6</td>
</tr>
<tr>
<td>Experiment2</td>
<td>11</td>
</tr>
<tr>
<td>Experiment3</td>
<td>16</td>
</tr>
<tr>
<td>Experiment4</td>
<td>21</td>
</tr>
<tr>
<td>Experiment5</td>
<td>26</td>
</tr>
<tr>
<td>Experiment6</td>
<td>31</td>
</tr>
<tr>
<td>Experiment7</td>
<td>36</td>
</tr>
<tr>
<td>. . .</td>
<td>. . .</td>
</tr>
</tbody>
</table>

If your data is all in the same file you will have to indicate the reference column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the control channel values. On the other hand, if you have separate data files with the same column containing the control channel values, you may use the general object-name given below, rather than entering the column number for the control channel values in each file.

ReferenceColumn : number of the column containing the control channel values

ReferenceColumn : 9

20. If your data includes the control channel’s background signal, which column of your data file contains that information? If your data does not have control channel values, skip this question, and the associated “.experiment” file entry.

Experiment#ReferenceBackColumn : number of the column containing the control channel’s background signals for the sample indicated

<table>
<thead>
<tr>
<th>Experiment</th>
<th>ReferenceBackColumn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment1</td>
<td>7</td>
</tr>
<tr>
<td>Experiment2</td>
<td>12</td>
</tr>
<tr>
<td>Experiment3</td>
<td>17</td>
</tr>
<tr>
<td>Experiment4</td>
<td>22</td>
</tr>
<tr>
<td>Experiment5</td>
<td>27</td>
</tr>
<tr>
<td>Experiment6</td>
<td>32</td>
</tr>
<tr>
<td>Experiment7</td>
<td>37</td>
</tr>
<tr>
<td>. . .</td>
<td>. . .</td>
</tr>
</tbody>
</table>
If your data is all in the same file you will have to indicate the control channel background column for each experiment, illustrated above. This is also true if you have two or more data files with different columns containing the control channel’s background values. If, on the other hand, you have separate data files, with the same column containing the control channel’s background values, you may use the general object-name given below, rather than entering the column number of the control channel’s background values for each file.

```
ReferenceBackColumn : number of the column containing the control channel’s background values
```

### 4.5.10 Measurement Flags

21. If your data file has a notation (flag) indicating whether or not the experiment worked for each gene, indicate which column contains this information. If your data does not include this information, skip this question, and the associated “.experiment” file entries.

```
Experiment#OkColumn : number of the column saying whether or not the experiment indicated worked for each gene
```

```
Experiment1OkColumn : 8
Experiment2OkColumn : 13
Experiment3OkColumn : 18
Experiment4OkColumn : 23
Experiment5OkColumn : 28
Experiment6OkColumn : 33
Experiment7OkColumn : 38
```

If your data is all in the same file you will have to indicate the “experiment worked” column for each sample, illustrated above. This is also true if you have two or more data files with different columns containing the “experiment worked” information. If, on the other hand, you have separate data files, with the same column containing the “experiment worked” notation, you may use the general object-name given below, rather than entering the column number of the reference’s background values for each file.

```
OkColumn : number of the column saying whether or not the experiment worked for each gene
```

```
OkColumn : 11
```
22. If you have a column indicating whether or not your experiment worked, what is the designation used in this column to indicate the experiment worked? (Often this is just a letter, such as P for Present or Passed.) If you do not have an “experiment worked” column, skip this question and the associated “.experiment” entry.

StatusOkString : the value, letter or word indicating the sample is ok to use

StatusOkString : P

You can have more than one entry indicating the status. If you were not sure if your experiment recorded P for passed or O for OK, place both in the line, separated by vertical bars.

You might also have a designation for Marginal or Questionable data. (Often this is just a letter, such as M for Marginal.)

StatusMarginalString : the value, letter or word indicating the sample is of marginal quality

StatusMarginalString : M|Q

You might also have a designation for Failed or Absent data. (Often this is just a letter, such as A for Absent.)

StatusFailedString : the value, letter or word indicating the sample is absent

StatusFailedString : F|A

4.5.11 Associating a Picture with a Sample

Pictures are nice, but they are not necessary. If you don’t have any, skip this section and the associated “.experiment” file entries.

23. If you have any pictures you wish to associate with any or all of the samples use the line given below to tell GeneSpring where to find the picture. If you do not have a picture to associate with every sample, GeneSpring will display the picture associated with the next closest sample with an associated picture.

Experiment#Image : the complete file name of the file containing the picture to associate with the indicated file

If you have a picture associated with every sample this section of your “.experiment” file should look similar to this:

Experiment1Image : yeastpict1A0.gif
Experiment2Image : yeastpict1A10.gif
Experiment3Image : yeastpict1A20.gif
Experiment4Image : yeastpict1A30.gif
Experiment5Image : yeastpict1A40.gif
Experiment6Image : yeastpict1B0.gif
If you have only one picture to associate with the entire experiment being described in your “.experiment” file, the picture entry should look similar to this one:

Experiment1Image : happy_yeast_picture.gif

If you have some pictures to associate with some but not all points in your sample the picture entries in your “.experiment” file should look similar to these:

Experiment1Image : yeastpict1A.gif
Experiment6Image : yeastpict1B.gif
Experiment11Image : yeastpict1AndromedaA.gif
Experiment16Image : yeastpict1AndromedaB.gif
Experiment21Image : yeastpict2A.gif
Experiment26Image : yeastpict2B.gif
Experiment31Image : yeastpict2AndromedaA.gif
Experiment36Image : yeastpict2AndromedaB.gif

4.5.12 Normalizations: Negative Controls

24. Do you have any genes designated as negative controls on your array? You have negative controls when there is DNA from a different genome than the one you are investigating on the array. Entering “true” as the object-value of the line given below means you have negative controls, and you want GeneSpring to normalize your samples using the negative control values. This normalization method takes the average signal intensities for all of the negative controls and subtracts this number from the signal intensity of each gene. For more info about this normalization option, see Chapter 1, Normalizing Options. If you do not have negative controls, or do not want to normalize your samples using the data from them, either do not enter the “NormalizeNegControl : ” line, or type “false” as the object-value.

NormalizeNegControl : either true or false

NormalizeNegControl : false

4.5.12.1 The required layout file for negative controls

25. If you do not have negative controls or are not using them to normalize your data, skip this question and the associated “.experiment” file entry. If you are using negative controls you must have a layout file. (See 4.3.2.4 “The Layout file” on page 4-7 for what this file can or should contain.) There are two normalization options requiring you to have a layout file. They both use this line to tell GeneSpring where to find the layout file. You should only have one layout file, and you should only enter the line, “Layout : name of layout file”, once. You may have entered this file already, please refer to 4.5.7.1 “The required layout file for Region Specifications” on page 4-54.

Layout : complete name of the layout file

Layout : AffyYeastLayout4.txt
4.5.13 Normalizations: Control Channel Values

If you do not have control channel values, skip these questions and the associated “.experiment” file entries.

26. If you have a control channel value for each gene to indicate the trust you have in the experimental data for each gene you probably want to normalize the genes by dividing their signal strength by the control channel’s signal strength. If you have a background signal for either or both of these values, it is subtracted from the signal intensities before they are divided. For more information on this normalization option, see Chapter 1, Normalizing Options. If you wish to use this normalization, enter “true” as the object-value in the line illustrated below. If you do not have control channel values, or you do not wish your data to be normalized using the control channel values, either do not enter the line “NormalizeToReference : “, or enter “false” as the object-value in that line. Control channels generally apply to two color experiments.

\[
\text{NormalizeToReference : either true or false}
\]

\[
\text{NormalizeToReference : true}
\]

27. If you do not have control channel values, skip this question and the associated “.experiment” file entry. Sometimes the control channel value is very low and would artificially inflate the noise for its gene, indicate the minimum value you would be willing to divide a gene’s signal by:

\[
\text{NormalizeMinControl : the minimum signal value to be used as a reference value for normalization purposes}
\]

\[
\text{NormalizeMinControl : 10}
\]

If you do not enter this line in your “.experiment” file and you do have control channel values, GeneSpring will automatically use the value given here, 10, as the default cut-off value.

28. If you have control channel values for your experiment, but the column containing the “raw data” has already been normalized using this information (for example, your data is reported in ratio form), you can tell GeneSpring this, using the line illustrated below. If you have the raw data from both the gene and its control it is suggested you let GeneSpring perform your normalization, rather than using this option. For example, Incyte data is reported in what they call “ratio” form, but the ratio reported is not actually the gene’s signal divided by its control; in this case it would probably be better to use the raw signal and control values and let GeneSpring perform the normalization. If you want to go ahead and use previously normalized data as your raw data, you should still tell GeneSpring in which column(s) the control signals are located.

\[
\text{UseReferenceAsStrength : enter true or false}
\]

\[
\text{UseReferenceAsStrength : false}
\]
4.5.14 Normalizations: Positive Controls

29. Do you have any genes designated as positive controls on your array? You typically have positive controls when there is DNA from a different genome than the one you are investigating on your array, and you added a known quantity of that DNA to your sample. Entering “true” as the object-value of the line given below means you have positive controls, and you want GeneSpring to normalize your experiment using the positive control values. This normalization method takes the average signal intensities of all of the positive controls and divides each gene’s signal intensity by that number, for more information about this normalization option see Chapter 1, Normalizing Options. If you do not want to normalize your experiment using positive controls, either do not enter the “NormalizePosControl : ” line, or type “false” as the object-value.

NormalizePosControl : either true or false

NormalizePosControl : true

4.5.14.1 The required layout file for positive controls

30. If you do not have positive controls or if you are not using them to normalize your data, skip this question and the associated “.experiment” file entry. If you are using positive controls you must have a layout file, and a file specifying what the positive controls are, this second file must have the gene names of the positive controls written in a list, one gene per line. See section 4.3.2.4 “The Layout file” on page 4-7 for more information about these files. Specify the complete file name of the layout file with the line below.

Layout : complete name of the layout file, the file name can be anything, with or without spaces

Layout : AffyYeastLayout4.txt

There are two normalization options requiring you to have a layout file; both use the same line to tell GeneSpring where to find the file. You should only have one layout file, and you should only enter the line, “Layout : name of layout file”, once. You may have already entered this file, please refer to 4.5.7.1 “The required layout file for Region Specifications” on page 4-54.

31. If you do not have positive controls or are not using them to normalize your data, skip this question and the associated “.experiment” file entry. Sometimes something will go wrong with the positive controls and you will get very low values for all of them, which you will not want to use for normalization purposes. Indicate the minimum average the positive controls must have such that dividing each genes’ signal strength by the average of the positive controls will not artificially inflate the noise of the genes.

NormalizeMinRange : indicate the minimum average allowable for the positive controls

NormalizeMinRange : 10

The number indicated in the example (10) is the default cut-off value. If you do not enter this line, this is the cutoff value GeneSpring will use.
4.5.15 Normalizations: Each Sample to Itself

32. Do you want to normalize your data by making the median of all of your measurements 1, for each sample in your experiment? (If you have not already preformed normalizations on your data you generally want to use this normalization option.) For more information about this normalization option, see Chapter 1, Normalizing Options.

   NormalizeNoControl : either true or false
   
   NormalizeNoControl : true

33. If you are not normalizing each sample to itself, skip this question and the associated "experiment" file entry. Sometimes something will go wrong with the experiment and you will get very low values for everything. Indicate the cut-off value by telling GeneSpring not to raise all of the signal strength values up to a median of 1 if their average is below this number:

   NormalizeMinRange : Specify the cut-off value
telling GeneSpring not to raise all of the signal strength values up to a median of 1 if the average signal strength is below this number

   NormalizeMinRange : 10

   The number indicated in the example (10) is the default cut-off value. If you do not enter this line, this is the cutoff value GeneSpring will use.

4.5.16 Normalizations: Each Gene to Itself

34. Do you want to normalize each gene to itself, so the median of all the measurements taken for the gene is one? See Chapter 1, Normalizing Options for more information about this option. If you are not doing a two-color experiment you generally want to do this.

   NormalizeEachGene : either true or false
   
   NormalizeEachGene : true

35. Skip this question and the associated entry if you are not normalizing each gene to itself. Sometimes something will go wrong with the samples and all of the values for a particular gene are very low, in which case GeneSpring will artificially inflate the noise of the gene if you normalize those values up to a median of one. To specify where this cut-off is, type the line below in the "experiment" file:

   NormalizeMinMedian : the numerical cut-off value
below which you will not normalize a gene to itself

   NormalizeMinMedian : 0.01

   The number indicated in the example (0.01) is the default cut-off value. If you do not enter this line, this is the cutoff value GeneSpring will use.
4.5.17 Normalizations: Each Sample to a Specific Sample

36. Do you want to normalize each sample to one sample within the experiment? If so, enter the number of the sample, counting from zero as the object-value in the line below. Silicon Genetics does not recommend using this normalization option, unless you have very specific reasons as described in Chapter 1, Normalizing Options.

   NormalizeToExperiment : true or false
   NormalizeToExperiment : 0

4.5.18 Colorbar Specifications

37. The intensity of the colorbar in GeneSpring indicates how reliable the data for each gene is. 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. The default values are specified in the example. If you do not indicate a high, medium, and low values specifically, then the values GeneSpring will automatically use to determine the color bar are:

   SignalHigh : a high number, this indicates high confidence in the data
   SignalMedium : a medium number, this indicates average confidence in the data
   SignalLow : a low number, this indicates low confidence in the data

   SignalHigh : 500
   SignalMedium : 150
   SignalLow : 50

These numbers are arbitrary. They are intended to be general benchmarks, not hard boundaries.
4.5.19  Graph Specifications

The values indicated here can be altered within GeneSpring, you are simply setting the default values here.

38. To allow you to inspect the genes’ expression profiles closely, GeneSpring does not graph the entire y-axis (the expression level axis), but only the portion most genes profiles fall into. Indicate the range of expression levels GeneSpring should graph.

\[
\begin{align*}
\text{LowerBound} : & \text{ Indicate the lowest expression level to graph on the y-axis} \\
\text{UpperBound} : & \text{ Indicate the highest expression level to graph on the y-axis}
\end{align*}
\]

LowerBound : 0
UpperBound : 5.0

A lower bound of 0 and an upper bound of 5 are the default settings of GeneSpring.

4.5.20  Examples of Installed Experiments

4.5.20.1  An Example of Raw data files

The experimental data file for the Yeast Extraterrestrial Studies Experiment is too long to show in its entirety, but here is the data from the first 10 genes listed in the experimental data files. The first table shows the file as a tab delineated text file. There are 41 columns in this file, one for the gene name, and one for each experimental measurement taken. The second table shows the same data (for the first 10 samples) as it appears in Microsoft Excel®, which is a much easier format for a human to look at this data in.
## Table 4-2

The first 10 genes of the experimental data file for the experiment “Yeast Extraterrestrial Studies”, as a tab-delimited text file

<table>
<thead>
<tr>
<th>experiment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>YFL039C</td>
<td>830</td>
<td>1117</td>
<td>1384</td>
<td>1101</td>
<td>1357</td>
<td>1162</td>
<td>1464</td>
<td>1384</td>
<td>973</td>
<td>1618</td>
</tr>
<tr>
<td></td>
<td>1432</td>
<td>1374</td>
<td>1068</td>
<td>1568</td>
<td>1313</td>
<td>1638</td>
<td>1648</td>
<td>973</td>
<td>1162</td>
<td>1384</td>
</tr>
<tr>
<td></td>
<td>1117</td>
<td>1384</td>
<td>1101</td>
<td>1357</td>
<td>1384</td>
<td>1464</td>
<td>978</td>
<td>973</td>
<td>1618</td>
<td>1432</td>
</tr>
<tr>
<td></td>
<td>1068</td>
<td>1568</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YFL039C</td>
<td>1282</td>
<td>1218</td>
<td>1496</td>
<td>1431</td>
<td>1847</td>
<td>1255</td>
<td>1402</td>
<td>1496</td>
<td>1160</td>
<td>2531</td>
</tr>
<tr>
<td></td>
<td>2072</td>
<td>1765</td>
<td>1133</td>
<td>1726</td>
<td>1950</td>
<td>1756</td>
<td>1879</td>
<td>1160</td>
<td>1255</td>
<td>1496</td>
</tr>
<tr>
<td></td>
<td>1218</td>
<td>1496</td>
<td>1431</td>
<td>1847</td>
<td>1496</td>
<td>1402</td>
<td>1177</td>
<td>1160</td>
<td>2531</td>
<td>2072</td>
</tr>
<tr>
<td></td>
<td>1133</td>
<td>1726</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YFL039C</td>
<td>1446</td>
<td>1431</td>
<td>1776</td>
<td>1762</td>
<td>2003</td>
<td>1387</td>
<td>1551</td>
<td>1776</td>
<td>1354</td>
<td>4074</td>
</tr>
<tr>
<td></td>
<td>2387</td>
<td>2065</td>
<td>1246</td>
<td>2430</td>
<td>2662</td>
<td>2372</td>
<td>2643</td>
<td>1354</td>
<td>1387</td>
<td>1776</td>
</tr>
<tr>
<td></td>
<td>1431</td>
<td>1776</td>
<td>1762</td>
<td>2003</td>
<td>1776</td>
<td>1551</td>
<td>1337</td>
<td>1354</td>
<td>4074</td>
<td>2387</td>
</tr>
<tr>
<td></td>
<td>1246</td>
<td>2430</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YER148w</td>
<td>715</td>
<td>971</td>
<td>722</td>
<td>874</td>
<td>886</td>
<td>900</td>
<td>662</td>
<td>722</td>
<td>788</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td>786</td>
<td>757</td>
<td>805</td>
<td>986</td>
<td>817</td>
<td>951</td>
<td>898</td>
<td>788</td>
<td>900</td>
<td>722</td>
</tr>
<tr>
<td></td>
<td>971</td>
<td>722</td>
<td>874</td>
<td>886</td>
<td>722</td>
<td>662</td>
<td>805</td>
<td>788</td>
<td>437</td>
<td>786</td>
</tr>
<tr>
<td></td>
<td>805</td>
<td>986</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YER148w</td>
<td>756</td>
<td>1231</td>
<td>1054</td>
<td>1057</td>
<td>1084</td>
<td>1134</td>
<td>1093</td>
<td>1054</td>
<td>983</td>
<td>834</td>
</tr>
<tr>
<td></td>
<td>1116</td>
<td>1064</td>
<td>955</td>
<td>1686</td>
<td>1395</td>
<td>1659</td>
<td>1375</td>
<td>983</td>
<td>1134</td>
<td>1054</td>
</tr>
<tr>
<td></td>
<td>1231</td>
<td>1054</td>
<td>1057</td>
<td>1084</td>
<td>1054</td>
<td>1093</td>
<td>972</td>
<td>983</td>
<td>834</td>
<td>1116</td>
</tr>
<tr>
<td></td>
<td>955</td>
<td>1686</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YER148w</td>
<td>448</td>
<td>937</td>
<td>1051</td>
<td>921</td>
<td>681</td>
<td>891</td>
<td>1156</td>
<td>1051</td>
<td>681</td>
<td>1190</td>
</tr>
<tr>
<td></td>
<td>725</td>
<td>773</td>
<td>773</td>
<td>1321</td>
<td>1290</td>
<td>1683</td>
<td>1115</td>
<td>681</td>
<td>891</td>
<td>1051</td>
</tr>
<tr>
<td></td>
<td>937</td>
<td>1051</td>
<td>921</td>
<td>681</td>
<td>1051</td>
<td>1156</td>
<td>508</td>
<td>681</td>
<td>1190</td>
<td>725</td>
</tr>
<tr>
<td></td>
<td>773</td>
<td>1321</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YER022w</td>
<td>65</td>
<td>48</td>
<td>40</td>
<td>69</td>
<td>81</td>
<td>74</td>
<td>77</td>
<td>40</td>
<td>67</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>81</td>
<td>71</td>
<td>73</td>
<td>52</td>
<td>65</td>
<td>63</td>
<td>67</td>
<td>74</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>40</td>
<td>69</td>
<td>81</td>
<td>40</td>
<td>77</td>
<td>82</td>
<td>67</td>
<td>33</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YER022w</td>
<td>116</td>
<td>76</td>
<td>79</td>
<td>124</td>
<td>124</td>
<td>124</td>
<td>151</td>
<td>101</td>
<td>79</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>123</td>
<td>129</td>
<td>102</td>
<td>83</td>
<td>70</td>
<td>111</td>
<td>137</td>
<td>151</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>76</td>
<td>79</td>
<td>124</td>
<td>124</td>
<td>79</td>
<td>101</td>
<td>129</td>
<td>137</td>
<td>51</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>129</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YER022w</td>
<td>390</td>
<td>376</td>
<td>308</td>
<td>398</td>
<td>439</td>
<td>451</td>
<td>348</td>
<td>308</td>
<td>390</td>
<td>311</td>
</tr>
<tr>
<td></td>
<td>361</td>
<td>380</td>
<td>301</td>
<td>387</td>
<td>385</td>
<td>425</td>
<td>310</td>
<td>390</td>
<td>451</td>
<td>308</td>
</tr>
<tr>
<td></td>
<td>376</td>
<td>308</td>
<td>398</td>
<td>439</td>
<td>308</td>
<td>348</td>
<td>390</td>
<td>390</td>
<td>311</td>
<td>361</td>
</tr>
<tr>
<td></td>
<td>301</td>
<td>387</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use comma and periods interchangeably.
This was a very simple experiment, there were no pictures taken of the experimental conditions, there were no pictures taken of the arrays, there were no regions, or positive or negative controls, and the genes were reported in exactly the same format as was used in the list of genes defining the genome, thus there was no need for a map file.

### 4.5.20.2 An example “.experiment” file

This is the complete .experiment file for the “Yeast Extraterrestrial Studies” Experiment.

```plaintext
name : Yeast Extraterrestrial Studies

Experiments : 40

Parameters : 4

Parameter1Name : Kryptonite concentration
Parameter2Name : Variety of yeast
Parameter3Name : Test repeat number
Parameter4Name : Andromeda Strain infection

Parameter1Units : ppm
Parameter2Units : 
Parameter3Units : 
Parameter4Units : 

Parameter1IsNumber : true
Parameter2IsNumber : false
Parameter3IsNumber : true
Parameter4IsNumber : false

Parameter1IsLogarithmic : false
Parameter2IsLogarithmic : false
Parameter3IsLogarithmic : false
```

<table>
<thead>
<tr>
<th>SAMPLE</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>YFL039C</td>
<td>830</td>
<td>1117</td>
<td>1384</td>
<td>1101</td>
<td>1357</td>
<td>1162</td>
<td>1464</td>
<td>1384</td>
<td>973</td>
</tr>
<tr>
<td>YFL039C</td>
<td>1282</td>
<td>1218</td>
<td>1496</td>
<td>1431</td>
<td>1847</td>
<td>1255</td>
<td>1402</td>
<td>1496</td>
<td>1160</td>
</tr>
<tr>
<td>YFL039C</td>
<td>1446</td>
<td>1431</td>
<td>1776</td>
<td>1762</td>
<td>2003</td>
<td>1387</td>
<td>1551</td>
<td>1776</td>
<td>1354</td>
</tr>
<tr>
<td>YER148w</td>
<td>715</td>
<td>971</td>
<td>722</td>
<td>874</td>
<td>886</td>
<td>900</td>
<td>662</td>
<td>722</td>
<td>788</td>
</tr>
<tr>
<td>YER148w</td>
<td>756</td>
<td>1231</td>
<td>1054</td>
<td>1057</td>
<td>1084</td>
<td>1134</td>
<td>1093</td>
<td>1054</td>
<td>983</td>
</tr>
<tr>
<td>YER148w</td>
<td>448</td>
<td>937</td>
<td>1051</td>
<td>921</td>
<td>681</td>
<td>891</td>
<td>1156</td>
<td>1051</td>
<td>681</td>
</tr>
<tr>
<td>YER022w</td>
<td>65</td>
<td>48</td>
<td>40</td>
<td>69</td>
<td>81</td>
<td>74</td>
<td>77</td>
<td>40</td>
<td>67</td>
</tr>
<tr>
<td>YER022w</td>
<td>116</td>
<td>76</td>
<td>79</td>
<td>124</td>
<td>124</td>
<td>151</td>
<td>101</td>
<td>79</td>
<td>137</td>
</tr>
<tr>
<td>YER022w</td>
<td>390</td>
<td>376</td>
<td>308</td>
<td>398</td>
<td>439</td>
<td>451</td>
<td>348</td>
<td>308</td>
<td>390</td>
</tr>
</tbody>
</table>

**Table 4-3** The experimental data of the first 10 samples from the first 10 genes in the Yeast Extraterrestrial Studies Experiment shown in a Microsoft Excel® table
Parameter4IsLogarithmic : false
Parameter1IsContinuous : true
Parameter2IsContinuous : false
Parameter3IsContinuous : false
Parameter4IsContinuous : false
Parameter1IsRepeat : false
Parameter2IsRepeat : false
Parameter3IsRepeat : true
Parameter4IsRepeat : false
Parameter1IsSet : false
Parameter2IsSet : false
Parameter3IsSet : false
Parameter4IsSet : true

Parameter1Experiment1 : 0
Parameter1Experiment2 : 10
Parameter1Experiment3 : 20
Parameter1Experiment4 : 30
Parameter1Experiment5 : 40
Parameter1Experiment6 : 0
Parameter1Experiment7 : 10
Parameter1Experiment8 : 20
Parameter1Experiment9 : 30
Parameter1Experiment10 : 40
Parameter1Experiment11 : 0
Parameter1Experiment12 : 10
Parameter1Experiment13 : 20
Parameter1Experiment14 : 30
Parameter1Experiment15 : 40
Parameter1Experiment16 : 0
Parameter1Experiment17 : 10
Parameter1Experiment18 : 20
Parameter1Experiment19 : 30
Parameter1Experiment20 : 40
Parameter1Experiment21 : 0
Parameter1Experiment22 : 10
Parameter1Experiment23 : 20
Parameter1Experiment24 : 30
Parameter1Experiment25 : 40
Parameter1Experiment26 : 0
Parameter1Experiment27 : 10
Parameter1Experiment28 : 20
Parameter1Experiment29 : 30
Parameter1Experiment30 : 40
Parameter1Experiment31 : 0
Parameter1Experiment32 : 10
Parameter1Experiment33 : 20
Parameter1Experiment34 : 30
Parameter1Experiment35 : 40
Parameter1Experiment36 : 0
Parameter1Experiment37 : 10
Parameter1Experiment38 : 20
Parameter1Experiment39 : 30
Parameter1Experiment40 : 40

Parameter2Experiment1 : A
Parameter2Experiment2 : A
Parameter2Experiment3 : A
Parameter2Experiment4 : A
Parameter2Experiment5 : A
Parameter2Experiment6 : B
Parameter2Experiment7 : B
Parameter2Experiment8 : B
Parameter2Experiment9 : B
Parameter2Experiment10 : B
Parameter2Experiment11 : A
Parameter2Experiment12 : A
Parameter2Experiment13 : A
Parameter2Experiment14 : A
Parameter2Experiment15 : A
Parameter2Experiment16 : B
Parameter2Experiment17 : B
Parameter2Experiment18 : B
Parameter2Experiment19 : B
Parameter2Experiment20 : B
Parameter2Experiment21 : A
Parameter2Experiment22 : A
Parameter2Experiment23 : A
Parameter2Experiment24 : A
Parameter2Experiment25 : A
Parameter2Experiment26 : B
Parameter2Experiment27 : B
Parameter2Experiment28 : B
Parameter2Experiment29 : B
Parameter2Experiment30 : B
Parameter2Experiment31 : A
Parameter2Experiment32 : A
Parameter2Experiment33 : A
Parameter2Experiment34 : A
Parameter2Experiment35 : A
Parameter2Experiment36 : B
Parameter2Experiment37 : B
Parameter2Experiment38 : B
Parameter2Experiment39 : B
Parameter2Experiment40 : B

Parameter3Experiment1 : Test 1
Parameter3Experiment2 : Test 1
Parameter3Experiment3 : Test 1
Parameter3Experiment4 : Test 1
Parameter3Experiment5 : Test 1
Parameter3Experiment6 : Test 1
Parameter3Experiment7 : Test 1
Parameter3Experiment8 : Test 1
Parameter3Experiment9 : Test 1
Parameter3Experiment10 : Test 1
Parameter3Experiment11 : Test 1
Parameter3Experiment12 : Test 1
Parameter3Experiment13 : Test 1
Parameter3Experiment14 : Test 1
Parameter3Experiment15 : Test 1
Parameter3Experiment16 : Test 1
Parameter3Experiment17 : Test 1
Parameter3Experiment18 : Test 1
Parameter3Experiment19 : Test 1
Parameter3Experiment20 : Test 1
Parameter3Experiment21 : Test 2
Parameter3Experiment22 : Test 2
Parameter3Experiment23 : Test 2
Parameter3Experiment24 : Test 2
Parameter3Experiment25 : Test 2
Parameter3Experiment26 : Test 2
Parameter3Experiment27 : Test 2
Parameter3Experiment28 : Test 2
Parameter3Experiment29 : Test 2
Parameter3Experiment30 : Test 2
Parameter3Experiment31 : Test 2
Parameter3Experiment32 : Test 2
Parameter3Experiment33 : Test 2
Parameter3Experiment34 : Test 2
Parameter3Experiment35 : Test 2
Parameter3Experiment36 : Test 2
Parameter3Experiment37 : Test 2
Parameter3Experiment38 : Test 2
Parameter3Experiment39 : Test 2
Parameter3Experiment40 : Test 2

Parameter4Experiment1 : healthy
Parameter4Experiment2 : healthy
Parameter4Experiment3 : healthy
Parameter4Experiment4 : healthy
Parameter4Experiment5 : healthy
Parameter4Experiment6 : healthy
Parameter4Experiment7 : healthy
Parameter4Experiment8 : healthy
Parameter4Experiment9 : healthy
Parameter4Experiment10 : healthy
Parameter4Experiment11 : Andromeda strain
Parameter4Experiment12 : Andromeda strain
Parameter4Experiment13 : Andromeda strain
Parameter4Experiment14 : Andromeda strain
Parameter4Experiment15 : Andromeda strain
Parameter4Experiment16 : Andromeda strain
Parameter4Experiment17 : Andromeda strain
Parameter4Experiment18 : Andromeda strain
Parameter4Experiment19 : Andromeda strain
Parameter4Experiment20 : Andromeda strain
Parameter4Experiment21 : healthy
Parameter4Experiment22 : healthy
Parameter4Experiment23 : healthy
Parameter4Experiment24 : healthy
Parameter4Experiment25 : healthy
Parameter4Experiment26 : healthy
Parameter4Experiment27 : healthy
Parameter4Experiment28 : healthy
Parameter4Experiment29 : healthy
Parameter4Experiment30 : healthy
Parameter4Experiment31 : Andromeda strain
Parameter4Experiment32 : Andromeda strain
Parameter4Experiment33 : Andromeda strain
Parameter4Experiment34 : Andromeda strain
Parameter4Experiment35 : Andromeda strain
Parameter4Experiment36 : Andromeda strain
Parameter4Experiment37 : Andromeda strain
Parameter4Experiment38 : Andromeda strain
Parameter4Experiment39 : Andromeda strain
Parameter4Experiment40 : Andromeda strain

DataFileName : extraterrestrial.txt

Headlines : 1

ExperimentGeneColumn : 1

Experiment1IntensityColumn : 1
Experiment2IntensityColumn : 3
Experiment3IntensityColumn : 4
Experiment4IntensityColumn : 5
Experiment5IntensityColumn : 6
Experiment6IntensityColumn : 7
Experiment7IntensityColumn : 8
Experiment8IntensityColumn : 9
Experiment9IntensityColumn : 10
Experiment10IntensityColumn : 11
Experiment11IntensityColumn : 12
Experiment12IntensityColumn : 13
Experiment13IntensityColumn : 14
Experiment14IntensityColumn : 15
Experiment15IntensityColumn : 16
Experiment16IntensityColumn : 17
Experiment17IntensityColumn : 18
Experiment18IntensityColumn : 19
Experiment19IntensityColumn : 20
Experiment20IntensityColumn : 21
Experiment21IntensityColumn : 22
Experiment22IntensityColumn : 23
Experiment23IntensityColumn : 24
Experiment24IntensityColumn : 25
Experiment25IntensityColumn : 26
Experiment26IntensityColumn : 27
Experiment27IntensityColumn : 28
Experiment28IntensityColumn : 29
Experiment29IntensityColumn : 30
Experiment30IntensityColumn : 31
Experiment31IntensityColumn : 32
Experiment32IntensityColumn : 33
Experiment33IntensityColumn : 34
Experiment34IntensityColumn : 35
Experiment35IntensityColumn : 36
Experiment36IntensityColumn : 37
Experiment37IntensityColumn : 38
Experiment38IntensityColumn : 39
Experiment39IntensityColumn : 40
Experiment40IntensityColumn : 41

NormalizeNoControl : true
NormalizeMinRange : 10
NormalizeToEachGene : true
NormalizeMinMedian : 0.1

SignalHigh : 500
SignalMedium : 150
SignalLow : 50

LowerBound : 0
UpperBound : 5
4.6 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. This is a very quick, but somewhat inflexible function of GeneSpring.

4.6.1 Preparation for Pasting

You should have normalized data in an Excel® file (Figure 4-41) or saved as tab-delimited text. (Figure 4-42). 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.

1. Name
   - First line must be the unique name of the experiment.

2. Parameters
   - The second line must be the first parameter (you may have as many parameters as you want, but you must have at least one).

<table>
<thead>
<tr>
<th>The five parameters for this experiment</th>
<th>The parameter values for the third sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Disease Example</td>
<td></td>
</tr>
<tr>
<td>Sick (Y)</td>
<td>no y y y y y</td>
</tr>
<tr>
<td>Infectious Disease (Y)</td>
<td>n y y y n n</td>
</tr>
<tr>
<td>Type Hepatitis (Y)</td>
<td>n a b n n n</td>
</tr>
<tr>
<td>Disease (Y)</td>
<td>no hepatitis hepatitis syphilis osteoporosis arthritis</td>
</tr>
<tr>
<td>Time (minutes)</td>
<td>0 10 20 30 40 50</td>
</tr>
<tr>
<td>YAL001C</td>
<td>0.941667 0.575 0.95 0.925 1.166667 0.8</td>
</tr>
</tbody>
</table>

   Figure 4-40 Example of parameter arrangements and values

   - In the first column is the name of the parameter.
   - Subsequent columns have values for parameter in that sample.
   - Each parameter must have units in parentheses in the same column as the name. For example, the parameter “time” would be immediately followed by (minutes). If your parameters have no units you must follow the name with an empty set of parentheses, or GeneSpring will not recognize it as a parameter.
   - As a default, GeneSpring assumes that the parametric values to follow are numeric and to be displayed in numerical order. If the parametric values for a parameter are non-numeric, immediately after the unit-indicating parentheses (empty if no units), enter an asterisk (*). There should be a space between right parenthesis and the asterisk (*). This tells GeneSpring to expect non-numeric parametric values and then treat the data appropriately.
• The default setting for interpretation of parameters is as a continuous element, please see 2.2.2 “Continuous Element” on page 2-11 for details. To have the parameters treated differently, enter the following codes just after the parentheses:
  • S — means the data will be interpreted as a non-continuous element, also known as a discrete element. Please see 2.2.3 “Non-Continuous Element (Set)” on page 2-16 for details.
  • C — data will be colored by the different parametric values assigned automatically by GeneSpring. In Figure 4-42 each column would get a different color as time values 0-160. Please see 2.2.4 “Color Code” on page 2-20 for details.
  • R — data will be interpreted as a replicate (not shown). Please see 2.2.1 “Replicate or Hidden Element” on page 2-8 for details.
• Of course, you can just enter all parameters with the default (no code after the parentheses) and change the interpretation later from within GeneSpring, please see GeneSpring User Manual 5.5, “Change Experiment Interpretation”.
• For example, for the parameter tissue type, a non-continuous non-numeric parameter, the first column might look like:
  tissue type() *S.

If you have no parameters give it arbitrary (but meaningful) names so you will be able to distinguish each sample from those in other columns.

3. Data
• There can only be one gene per line.
• The name of gene must be in the first column.
• The following columns are data points for each parameter.
Figure 4-41 Example of a correctly formatted Excel® sheet

<table>
<thead>
<tr>
<th>Experiment Name</th>
<th>First Parameter Name with units</th>
<th>Parameter Values</th>
<th>Normalized Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast cell cycle time series (no 90 min)</td>
<td>time (minutes)</td>
<td>control</td>
<td>control</td>
</tr>
<tr>
<td>YAL001C</td>
<td>0</td>
<td>0.941666722</td>
<td>0.575900894</td>
</tr>
<tr>
<td>YAL002W</td>
<td>1.733817728</td>
<td>107</td>
<td>0.919625714</td>
</tr>
<tr>
<td>YAL003W</td>
<td>0.710962891</td>
<td>1764.5</td>
<td>0.687302535</td>
</tr>
<tr>
<td>YAL004W</td>
<td>3.394501162</td>
<td>245.5</td>
<td>0.286784531</td>
</tr>
<tr>
<td>YAL006C</td>
<td>3.694501162</td>
<td>245.5</td>
<td>2.368784531</td>
</tr>
<tr>
<td>YAL007C</td>
<td>0.684667289</td>
<td>518.5</td>
<td>0.9219008</td>
</tr>
<tr>
<td>YAL008W</td>
<td>2.355028583</td>
<td>84.5</td>
<td>1.692307714</td>
</tr>
<tr>
<td>YAL009W</td>
<td>1.188034296</td>
<td>117</td>
<td>0.495726526</td>
</tr>
<tr>
<td>YAL010C</td>
<td>1.59342678</td>
<td>152.5</td>
<td>0.950196714</td>
</tr>
<tr>
<td>YAL011W</td>
<td>1.010452986</td>
<td>143.5</td>
<td>0.780487835</td>
</tr>
<tr>
<td>YAL012W</td>
<td>0.967189729</td>
<td>701</td>
<td>0.884950793</td>
</tr>
<tr>
<td>YAL013W</td>
<td>0.995052115</td>
<td>0.895052115</td>
<td>0.895052115</td>
</tr>
<tr>
<td>YAL014C</td>
<td>0.976744175</td>
<td>279.5</td>
<td>1.687654908</td>
</tr>
<tr>
<td>YAL015C</td>
<td>1.181942582</td>
<td>365.5</td>
<td>1.212838398</td>
</tr>
<tr>
<td>YAL016W</td>
<td>0.781092362</td>
<td>583.5</td>
<td>0.945382297</td>
</tr>
<tr>
<td>YAL017W</td>
<td>2.190476179</td>
<td>147</td>
<td>0.448979586</td>
</tr>
</tbody>
</table>

Figure 4-42 Example of a correctly formatted tab-delimited file
4.6.1.1 Most common mistakes in pasting

- forgetting the title
- not using parentheses
- not having parameters
- using unnormalized data
- 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
  (If your computer is set for a non-English language that typically uses commas for decimal markers, GeneSpring will recognize this. If, for example, your computer is set for French, the comma will be recognized as a decimal marker. You cannot use comma and periods interchangeably.)

4.6.1.2 Pasting your Experiment into GeneSpring

If you have not already, give your experiment a unique name. If it turns out it is not a unique name, then GeneSpring will append a number on the end to distinguish it from other experiments of the same name.

You can copy (Ctrl+C) all or part of a correctly set up Excel® or tab-delimited file.

In the main GeneSpring window, go to Edit > Paste > Paste Experiment.
GeneSpring will automatically update the window, regardless of which “View” you are in. Larger files may take longer to paste, depending on your system.

WARNING! Some computers will 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 Java virtual machine update.

GeneSpring will bring up a new Choose Experiment Name box, with the current name of the experiment already in the Name text box.
GeneSpring will take you back to main window with your new experiment already on display. From here, you can alter the normalizations with Experiment > Change Normalizations command or alter the interpretation with the Experiment > Change Interpretation command.
4.6.2 Copying an Experiment or a List Out of GeneSpring

Choose an experiment or a gene list from the Navigator. When you choose to copy and experiment, please be aware you will copy only the gene list currently selected. If you want to copy all the genes in your currently-viewed experiment, please right-click over the “All genes” list and select “Display List” before you begin to copy.

In the main GeneSpring window, select **Edit > Copy > Copy Experiment**.

Your data will be saved to the clipboard. From there you can paste your experiment or gene list into Microsoft® Notepad, Microsoft® Word or Microsoft® Excel.

When you paste the gene list will be sorted into the order presented in the “Ordered List” view.
4.7 Installing from a Database

4.7.1 Custom Databases and GeneSpring

You can load experiments into GeneSpring from your company’s database. To do this you will need to set up a .database file prior to starting the New Experiment Wizard.

4.7.1.1 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.

4.7.1.2 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 have to 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 have to 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 will figure out 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 need to do is install an ODBC driver specific to the type of database you will be using.

4.7.1.3 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

4.7.1.4 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 will be 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.

4.7.1.5 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.

4.7.1.6 Databases and GeneSpring

Experimental data is not always stored on the researcher’s desktop in simple text files. Sometimes the data is 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 then load in the data using the Experiment Wizard almost exactly as you would if they were text files (see 4.7.2.3 “Entering your Prepared Database into GeneSpring” on page 4-84). 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 statements. Silicon Genetics can provide these for GATC and help write these for individual databases. This only needs to be done once. Afterwards, the customer simply chooses the database and GeneSpring will get data from it. Normalization and other options can also be set for a database.

4.7.2 Adding an Experiment from a Database

Make sure you have a database. Any database software can be used to produce a database. First you must make sure that GeneSpring will be able to see your database. Your database’s creator should have done this already. If they have, you can skip down to 4.7.2.2 “Connect your Database to GeneSpring” on page 4-82.

1. Go to the control panel of your computer.
2. Select ODBC Data Sources. A new window will come up.

To make a new ODBC source:

1. Go to the system DSN
2. Click “Add”, which will bring up a new window
3. Select the correct type of database from the scrollable list. This will bring up a new panel.

4. Give the experiment a name. This is the name GeneSpring will use, so please remember that GeneSpring is case sensitive.
5. Click the “Select...” button to browse for the correct database.
   Normally you will need to browse into a new computer (server) to access the database.

6. Now there will be a new entry in the list of databases.

4.7.2.1 Test to make sure your ODBC connection is working

1. From Excel go to the “Tools” menu.
2. Select “Get External Data”.
3. Select “New Database Query...”. Look for your database in the presented list.

4.7.2.2 Connect your Database to GeneSpring

A database specification file must be set up. This is a plain text file, in a subdirectory of the main GeneSpring data directory entitled “Databases”. The text file should have the extension .database. This file will tell GeneSpring how to contact the customer’s database. The file contains several lines. Each line contains the name of a parameter you should set, followed by a colon, then followed by the value you want to set the parameter to.

The purpose of this file is to tell GeneSpring how to read this file 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 titled “database”.
2. Create a .database file. This file has specific requirement of what must be in it, but the items can be in any order.
   - jdbc : odbc : NameofDatabase
   - ExperimentTableName : SampleName
     If the index and gene name are separate, you will need more than one table.
     this should be a one word name. Case sensitivity depends on the database.
   - ExperimentTableIndex : which column contains the experiment number
   - GeneColumn : column number with names in it
   - IntensityColumn : should contain actual results
   - debug : true
     When true it will show what commands are sent to the database when you use the Experiment Wizard.
3. Arranging your Parameters

You need to make an SQL command that will get the parameters in all samples. You can use MicroSoftQuery in Excel to generate SQL commands.

- From Excel go to the “Tools” menu.
- Select “Get External Data”.
- Select “New Database Query...”.
- Make sure you tell it you want to edit in MicroSoftQuery.

GeneSpring wants:

1. Experiment ID.
2. Another experiment ID (must be unique).
3. Other parameters, Heading from tables, name of column. Double click headings to change the name if you want.

Button at the top of the query box says SQL. Click it to get SQL statements.

SQL Get experiment and indexes: SQL statements
(this needs to be on one unbroken line, do not use word wrap in your text editor.

Still missing from your experiment is:

- the default normalizations.
- specifications for Display Options
- specifications for Table Headings

Figure 4-50 Sample .database file

...
4.7.2.3 Entering your Prepared Database into GeneSpring

Using the Experiment Wizard, select the “Get Everything from the Database” option.

Figure 4-51 The “Data File Format” panel

The majority of the remaining Experiment Wizard panels will be filled in automatically.

If you left the debug setting for true an extra window will open up. When the query boxes come put these will contain actual SQL commands.

GeneSpring will have to go back to the database to get information every time you restart the program. If this takes too long, you might consider right-clicking over the correct database icon and selecting the save to disk option.

All commands in the .experiment files can also be added to the .database file.
4.7.2.4 Entering more Complicated Data from a Database

You can link various tables together in SQL. This typically requires a proficient user of databases, please 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 Table 4-4.

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Parameter Name</th>
<th>Parameter Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>elephants</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>elephants</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>daises</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 4-4 Sample table of mixed-up parameters

In Table 4-4 you do not have parameters in the individual columns. All parameters tables should have an associated sample number somewhere.

If you use a GATC database, you will have to re-link all the sample numbers to the parameter numbers. In that case, you need to 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.
Index

A
adding extra genes 3-7
affine background correction 1-18
All Samples to Specific Samples 4-63
API 4-78
Array Element List see Master Gene Table
Array Photos 4-37
Autoloader
column titles 4-3
headlines 4-3
B
background signal 4-55
C
Change Experiment Interpretation 2-31
CLI 4-79
Color by Parameter 2-20
color code parameter 4-49
colorbar 4-63
Common Name 3-4
conditions 4-1
constant value. see hard number
continuous parameter 4-48
Control Channel Background Column 4-33, 4-56
Control Channel Values 4-33, 4-56, 4-60
minimum value 4-60
pre-normalized data 4-60
Copying and Pasting data 4-71
D
Data Column Location 4-32, 4-54
data directory 3-10, 4-13
Data File Format 4-16
Data File Header Lines 4-27, 4-52
Data Import Wizard
Experiment 4-15
Genome 3-11
data location 4-13
database 4-78
DBMS 4-78
decimal markers 4-65, 4-74
Describe your Data Files 4-24, 4-51
Display Parameters 4-48

E
Each Gene to Itself 4-62
minimum average 4-62
Each Sample to Itself 4-62
minimum average 4-62
EC Number 3-4
experiment installation files 4-6
experiment interpretation
changing 2-31
saving 2-31
Experiment Name 4-46
experiment parameter 2-1
condition 2-20
example 2-3
multiple 2-4
parameter-value 2-2
Experiment Wizard 4-15
experimental data file 4-6
Export data
by copying 4-77
Expressed Sequence Tags 3-36
expression values
determining 1-1
F
Finish 4-45
Flags 1-18, 4-34, 4-57
G
GATC 4-79
GenBank locus 3-4
Gene Name 4-29, 4-53
Gene Name Prefix Removal 4-30, 4-53
Gene Name Suffix Removal 4-31, 4-53
GeneSpider
GenBank 3-36
LocusLink 3-36
non-GenBank genes 3-34
Save and Close 3-35
Sequence Data 3-35
Silicon Genetics 3-34
UniGene 3-37
Graphics Specifications 4-44
Guess the rest 4-35
H
hard number 1-7
headlines 4-52
How to Display the Parameters 4-22
I
Import data
   by pasting 4-71
installation files 4-6
interpretations 2-31
K
Keywords 3-4
L
layout file 4-7
   negative controls 4-59
   positive controls 4-61
   region specifications 4-54
M
Mapped format 4-12
   Common Name 3-4
   custom 3-5
   EC Number 3-4
   function 3-4
   GenBank locus 3-4
   gene list formats 3-3
   Keywords 3-4
   Map 3-4
   phenotype 3-4
   Protein Product 3-4
   Public Medline accession number 3-4
   sample 3-5, 3-6
   sequence 3-4
   Systematic Name 3-4
Mapping information 3-4
Master Gene List 3-3
Master Gene Table
   gene list formats 3-3
measurement flags 1-18, 4-34, 4-57
N
name function 3-3
   gene list formats 3-3
   name list 3-3
   gene list formats 3-3
Negative Controls 4-59
negative signal strengths 1-18
   non-continuous parameter 4-49
normalization options
   Control Channel Values 4-39
   Distribution of All Genes 1-6
   Global Scaling 1-6
   Normalizations by All Samples to a
      Specific Sample 4-43
   Normalizations by Each Gene to Itself 4-42
   Normalizations by Each Sample to a
      Hard Number 4-42
   Normalizations by Each Sample to
      Itself 4-41
   Normalizations by Negative Controls 4-38
   Normalizations by Positive Controls 4-40
   Normalize Each Gene to Itself 1-8
   Normalize Each Sample to a Hard
      Number 1-7
   Normalize Each Sample to Itself 1-6
   Normalize to Control Channel Values
      for Each Gene 1-3
   Normalize to Negative Controls 1-2
   Normalize to Positive Controls 1-5
   Normalizing All Samples to Specific
      Samples 1-11
   Region Normalization 1-16
   normalization techniques 1-1
   Normalization to Specific Samples 1-11
   Number of Arrays 4-18, 4-47
   Number of Parameters 4-20, 4-47
O
ODBC 4-78
P
parameter
   display 4-22
   non-numeric 4-71
   numeric 2-11, 4-71
Parameter Characteristics 4-21, 4-47
Parameter Interpretations 2-31
Parameter names 4-47
Parameter Values 4-23, 4-49
Parameters
  category 4-49
  color code 4-49
  continuous 4-48
  discrete 4-49
  display instructions 4-47
  non-continuous 4-49
  numbers 4-48
  replicate 4-49
  set 4-49
  units 4-47

parameters
  non-numeric 2-16

Pass/fail column. see Flags

pasting data 4-15

Phenotype 3-4

Pictures 4-58

Positive Controls 4-61
  minimum average 4-61

Properties of Experiment Set 4-18, 4-46

Protein Product 3-4

raw data 4-6

Reference Value Normalization Setup. see Control Channel Values

References Values. see Control Channel Values

region designation file 4-11

Region Normalization 1-16, 4-27
  multiple arrays 4-54

replicate parameter 4-49

RT-PCR Experiments 4-38

Sample Photos 4-36, 4-58

SGD 3-3
  gene list formats 3-3

SQL 4-78

Syntax 1-11

Systematic Name 3-4

Table of Genes see Master Gene Table

W
  web databases 3-19
    special character 3-20

Welcome panel 4-15

Wizard Panels
  Array Photos 4-37
  changing panels manually 4-16
  Control Channel Values 4-33, 4-39
  Data Column Location 4-32
  Data File Format 4-16
  Data File Header Lines 4-27
  Describe your Data Files 4-24
  Finish 4-45
  Flags 4-34
  Gene Name 4-29
  Gene Name Prefix Removal 4-30
  Gene Name Suffix Removal 4-31
  Graphics Specifications 4-44
  How to Display the Parameters 4-22
  Normalizations by All Samples to a Specific Sample 4-43
  Normalizations by Each Gene to Itself 4-42
  Normalizations by Each Sample to Itself 4-41
  Normalizations by Negative Controls 4-38
  Normalizations by Positive Controls 4-40
  Normalizations Each Sample to a Hard Number 4-42

Number of Arrays 4-18

Number of Parameters 4-20

Parameter Characteristics 4-21

Parameter Values 4-23

Properties of Experiment Set 4-18

Region Normalization 4-27

RT-PCR Experiments 4-38

Sample Photos 4-36

Welcome 4-15

Y
  y-axis 4-64
Document History for GeneSpring Loading Data

Update from 3.3 to 3.5
Update section 1.6 Normalizing to Specific Experiments
Improved index markers
Language change to “Master Gene Table”

Update from 3.2.2 to 3.3
Affine transformations, section 1.8.2
Updated graphics
Added section on normalizing multiple experiments

Update from 3.0 to 3.2.2
Revised section 3.4.5 “The Gene List File panel”
Changed hierarchy under section 3.2.1 What Data Are Necessary? to improve cross-references
Updated graphics in Genome Wizard
Updated graphics and revised Experiment Wizard

Update from 3.0 to 3.2
Added section for Mac Users
Added section 3.1 Creating folders for new Genomes
Moved Cutting and pasting to end of Chapter 4 to prevent confusion.
Added introduction to Chapter 4 to prevent confusion.
New features in GeneSpider