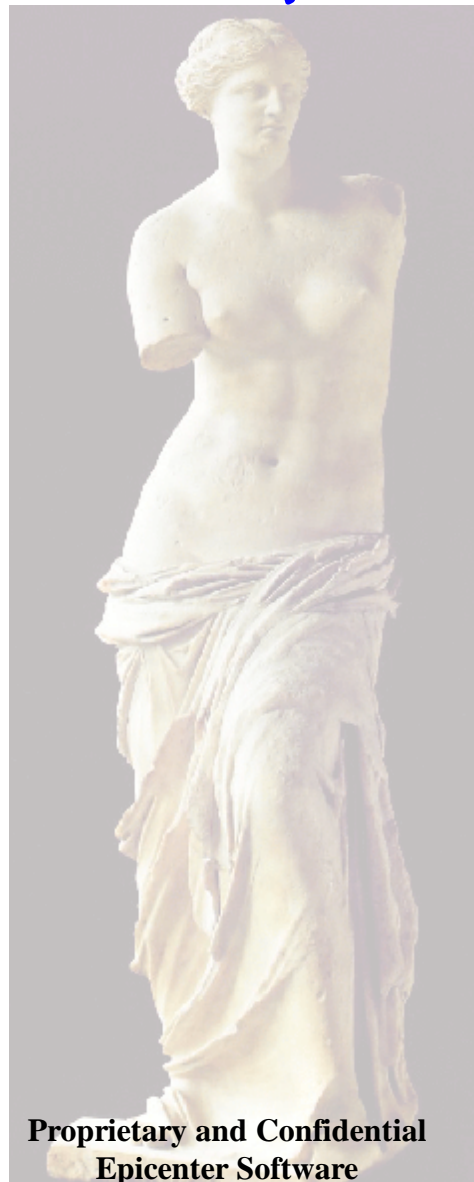


# *Genetrix*

## *QUICK START GUIDE & FAQ*



**Proprietary and Confidential  
Epicenter Software  
March 2003**



# TABLE OF CONTENTS

<b>Introduction</b> .....	1
<b>Installation</b> .....	1
<b>The Activation Key</b> .....	1
<b>The Icon Screen</b> .....	1
<b>What is a Genetrix Project?</b> .....	2
What is a Project file? .....	2

## BEFORE THE ANALYSIS

<b>How do I get expression data into Genetrix?</b> .....	4
What is probe set modeling? .....	4
Is probe set modeling always better? .....	4
How do I input expression data from an MAS .txt file? .....	4
How do I use Probe Profiler? .....	5
How do I import dChip data? .....	6
What about combining multi-chip expression data? .....	6
How do I replace expression data with updated values? .....	6
How do I add more samples to a Project? .....	6
Can I use non-Affymetrix expression data? .....	7
<b>How do I get covariate information into Genetrix?</b> .....	8
What are sample covariates? .....	8
What are gene covariates? .....	8
What sort of covariate data is allowed? .....	8
Can I use free-form text? .....	8
What about dates? .....	8
How are survival times stored? .....	8
Can I import from a text file or Excel spreadsheet? .....	8
Can I enter and edit sample covariate data directly? .....	9
What is the “Work A” covariate? .....	9
How are missing values represented? .....	9
<b>What should be done prior to analysis?</b> .....	10
<b>LABELS &amp; SETTINGS</b>	
What is a sample label? How does it differ from a sample name, or the alternate sample name? .....	10
What is a gene label? .....	10
How are display colors customized? .....	10
<b>DATA PREPROCESSING</b>	
Should the expression data be transformed prior to analysis? .....	11

How should sample replicates be handled? . . . . .	11
How should gene replicates be handled? . . . . .	12
How does Genetrix handle presence/absence data? . . . . .	12
How do I filter the genes? . . . . .	12
How do I select subsets of samples for analysis? What is the Active subset? . . . . .	12
What is the Selected subset? . . . . .	13
<b>ANALYSIS PREPROCESSING</b>	
How are survival times specified? . . . . .	13
What are Key covariates, and how are they selected? . . . . .	14
What are Key comparisons, and how are they selected? . . . . .	15

## ANALYSIS

<b>How should I approach the analysis?</b> . . . . .	16
The “Analysis Chain”	
Reducing the gene set	
Single gene or gene combination?	
Statistical vs. Biological validation	
<b>How do I do scatter plots?</b> . . . . .	18
<b>AXES</b>	
How do I specify the axes? . . . . .	18
How do I use an average value for a group of genes (or samples) for a plot axis? . . . . .	18
How do I plot fold changes? . . . . .	18
What is a “Principal Component” plot axis? . . . . .	18
Can I do multiple scatter plots? . . . . .	19
What is “Add X-X” for? . . . . .	19
In a multiple plot, can I expand a selected plot? . . . . .	19
Can I automatically scan through the plots? . . . . .	19
Can I plot standard error values? . . . . .	19
How do I get a 3-D scatter plot? . . . . .	19
Can I make it rotate? . . . . .	20
Is there a way to look at marginal distributions? . . . . .	20
Can I zoom? . . . . .	20
Can I make the X- and Y-scale the same? . . . . .	20
<b>POINTS</b>	
How can I change the size of the points? . . . . .	20
Can I add standard error bars to plot points? . . . . .	20
Can I indicate presence-absence data on the plot? . . . . .	20
Can I plot centroids (average expression for a group of points)? . . . . .	21
What are the options for coloring points? . . . . .	21
Can I highlight a subgroup of points? . . . . .	21
How do I change settings for a plot after it has been displayed? . . . . .	21
<b>INFORMATION</b>	
How do I get information on a point? . . . . .	21
How do I select a group of points? . . . . .	22
Can I label points? What if I want to change the label? . . . . .	22
Can I label centroids? . . . . .	22
What are thumbnails, and how do I get them? . . . . .	22

## ANALYSIS

Can I add a line-of-best-fit? .....	23
<b>What is “GeneScreen”?</b> .....	24
How do I request a GeneScreen? .....	24
How do I interpret the GeneScreen results? .....	24
Can I get more detail for a single gene? .....	25
Can I look at the properties of a groups of (the most significant) genes? .....	26
When would I want to save the p-values? .....	27
When would I want to run a permutation distribution? .....	27
Can I adjust for covariate data in the analysis? When might I want to? .....	27
What if I have matched pairs of samples? .....	27
Can GeneScreen consider two or more genes in combination? .....	27
How is GeneScreen used to predict outcomes? .....	27
<b>How do I display time series (ordered) data?</b> .....	29
How is the ordering specified? .....	29
What if the experiment includes two or more observations at each time point? .....	29
How do I compare two or more groups? .....	29
With more than a handful of genes, the display is just a mess of lines. Help. ....	29
How do I find genes with a certain pattern of expression? .....	29
<b>What is the similarity matrix?</b> .....	30
How is similarity measured? .....	30
The color coding is useful, but what if I want detail? .....	30
Can I select groups of cells? .....	30
How can I identify rows/columns of interest? .....	30
What happens when I “limit” the display? .....	31
Is there a way to analyze the structure in the similarity matrix? .....	31
What is Multi-Dimensional Scaling? .....	31
<b>Why cluster?</b> .....	32
What genes and samples should I use for clustering? .....	32
How do I know if the clusters mean anything? .....	32
Should I have a separate test data set? .....	34
Which clustering method should I choose? .....	34
What is the difference between supervised and unsupervised clustering? .....	34
What distance metric should I use? .....	34
<b>K-MEANS CLUSTERING</b>	
How does k-means clustering work? .....	35
<b>HIERARCHICAL CLUSTERING</b>	
How does hierarchical clustering work? .....	35
What is optimal ordering? .....	35
<b>SELF ORGANIZING MAPS</b>	
What are self organizing maps? .....	36
What does the map graphic show me? .....	36
<b>BAYESIAN CLUSTERING</b>	
What is Bayesian clustering? .....	36
<b>SUPPORT VECTORS</b>	
What are support vectors? .....	37

## META-CLUSTERING

What is meta clustering? .....	37
What is the Log? .....	38
Where do I find the Log? .....	38
How do I save images to the Log? .....	38
Can I save gene or sample lists to the Log? .....	38
Can I annotate the Log? .....	38
How can I automatically capture all output to the Log? .....	38
How can I view the Log? .....	39
How do I print Log images? .....	39
Can I share Log files with others? .....	39
How is Screen Capture different from the Log? .....	39

## SPECIFIC ANALYSES

<b>How should I approach two-group comparisons? .....</b>	<b>40</b>
How can I get fold changes? .....	40
<b>How do I use expression data for classification? .....</b>	<b>41</b>
<b>What tools are available for prognostic factor analysis? .....</b>	<b>42</b>
<b>How should I analyze time-course data? .....</b>	<b>43</b>

## INTERPRETATION

<b>How do I know what results are “real”? .....</b>	<b>44</b>
MULTIPLE TESTING	
Why is “multiple testing” a problem? .....	44
Can I adjust for multiple testing? .....	44
What is the False Discovery Rate? .....	44
RANDOMIZATION	
What does “shuffling” of samples achieve? .....	45
How can adding noise to the data be useful? .....	45
RESAMPLING	
What is the purpose of cluster resampling? .....	45
STATISTICAL VALIDATION	
When should I use an independent test set? .....	46
Why would I need a second test set? .....	46
What is leave-one-out cross-validation (LOOCV)? .....	47
BIOLOGICAL VALIDATION	
<b>What are gene Attributes? .....</b>	<b>48</b>
<b>What can Genetrix tell me about specific genes? .....</b>	<b>49</b>
How do I find a specific gene? .....	49
Where do the annotations in Gene Information come from? .....	49

How can I find other genes ‘like’ this one? . . . . .	49
How can I tell whether there are multiple probe sets targeting the same gene? . . . . .	50
What does it mean if two probe sets (same gene) have different expression values? . . .	50
Can I search the Web for more information, directly from Genetrix? . . . . .	50
<b>I have found a group of genes. Now what? . . . . .</b>	<b>51</b>
What does the table at the top of the List Genes screen tell me? . . . . .	51
Can I compare the list against a specific gene Attribute? . . . . .	51
Can I compare this list against another list? . . . . .	51
Can I get a hard copy, preferably with annotations and covariate data for each gene? . .	51
Can I save the list to the Log? . . . . .	51
When would I select Key Comparisons? . . . . .	51
What is MeanScreen? . . . . .	52
What happens if I ask to copy the list to a covariate? . . . . .	52
<b>How does Genetrix incorporate pathway information? . . . . .</b>	<b>53</b>
Can I add new pathways? . . . . .	53
What sort of information can be color-coded on gene symbols? . . . . .	53
Why are some rectangles divided into segments? . . . . .	53
Can I label genes in a pathway? . . . . .	53
What are thumbnails, and how do I get them? . . . . .	54
What is a pathway Attribute? . . . . .	54
<b>How can I map expression data to the genome? . . . . .</b>	<b>55</b>
What species are represented with a chromosomal ideogram? . . . . .	55

## Introduction

The purpose of the Quick Start Guide is to provide an abbreviated user manual that focuses on the information most needed to get a Project set up and an analysis underway. It does not attempt to describe all the features of Genetrix and many topics are covered relatively superficially. However, by focusing on core concepts and procedures, it provides the easiest and most direct path to begin to use Genetrix effectively.

The FAQs tries to focus on effective ways to use specific features of Genetrix, but does stray occasionally into more general and subjective areas of analytic strategy; be aware that advice on the relative merits of different analytic methods represent the opinions of one person.

## Installation

Genetrix has been developed for the Intel/Windows platform.

Genetrix is installed by running SETUP.EXE; this program may run automatically from the CD-ROM or you may need to select it to run it. If this is an **upgrade installation** remove the previous version before installation, through use of the Windows “Control Panel-Add/Remove Programs” facility.

The minimum/recommended hardware configuration is:

	<b>Minimum</b>	<b>Recommended</b>
Disk space	100Mb	As needed for data
Memory	512Mb	1Gb
Screen resolution	1280x1024	1280x1024 or 1600x1200
Clock speed		At least 1GHz

## The Activation Key

The first time you run Genetrix you will be asked to input an Activation key. This is a 20-character (letters plus numbers) code. Note that since the letters I and O can often be confused with 1 and 0, the key does not include any 0s or 1s and if you enter them they will be automatically converted to Os and Is. The key is not case-sensitive.

## The Icon Screen

The Icon Screen is relatively self explanatory. The first action will always be to open a Project, or to create a new Project by importing data. More than one Project can be opened, and if two or more are opened simultaneously, the Project box will list all Projects and to allow the user to switch between them.

Many functions are available for both gene data and sample data, and the icons for such pairs are

presented side-by-side. On the Icon Screen (and in many displays) the default color for depicting sample-related data is **yellow** and for gene-related data is **green**.

In the **Subsets box**, there are lists of the currently defined gene subsets (at left) and sample subsets (right). Clicking on a subset makes it the Active subset (see Data Preprocessing for description of the Active subset). Double-clicking on a subset will create a list of the genes (or samples) in the subset to view or evaluate.

At the bottom right is the **Log icon**, which may be used to view the current contents of the Genetrix Log. Ctrl-Shift-clicking on the icon toggles on or off AutoLogging. See “How do I capture analysis results?”, below, for details on the use of the Log.

## What is a Genetrix Project?

A Genetrix Project is a set of expression data, covariate information and ancillary information such as data transformation steps, labels, defined gene and sample subsets, that are stored in a single file. This file might represent the results of one or more experiment, with each experiment including multiple samples (with or without duplicates). The one thing all the samples will typically have in common will be that they address a common scientific question.

In practice, analyses may be confined to one or more subsets of the samples, but such subsetting is easily accommodated within Genetrix: the Project should include the super-set of all samples that might be needed for the analysis. Similarly, the Project will typically include all available gene information, although specific analyses may be restricted to defined gene subsets.

### What is a Project file?

After data have been imported into Genetrix they are saved as a Genetrix Project file, with suffix .gtx. Subsequent access to these data is obtained by opening this gtx file.

The advantage of a Project file is that it “freezes” the data, ensuring that analyses carried out over a period of time are based on a consistent data set, and not influenced by possible changes to the underlying database. Having a single Project file also greatly facilitates the interchange of data on a Project among collaborators.

## How many samples?

The glib answer is “more”. Given the vast amount of data that is generated from microarray experiments and the problems this usually leads to in terms of over-fitting of models and the impact of multiple testing on statistical significance levels, it is hard to imagine a microarray experiment that would not benefit from additional observations. In reality, sample size is driven more by pragmatic factors, such as cost and availability of tissues/patients.

Formal power calculations are difficult. For some methods, such as clustering, which do not involve formal hypothesis testing, no such calculations are possible. GeneScreen applies standard statistical tests, and these are directly amenable to sample size/power calculations, but the issue of multiple testing (for tens of thousands of genes) greatly complicates this. Adjusting the nominal type II error using a Bonferroni approach or something equivalent, will generally lead to sample size estimates that are too large to be practical. In addition, one doesn't usually conduct a microarray experiment in order to examine and test just one or two genes; these data sets lend themselves to detailed exploration of hundreds of genes, and it is difficult to frame a sample size/power calculation to adequately reflect this.

Two points to note: (1) The probe set modeling methods (dChip and Probe Profiler) that Genetrix uses to generate expression data from the Affymetrix .CEL files do not perform well when the number of samples is very small - that is, ten or fewer. (2) Hypothesis, as employed by GeneScreen, requires sample size (in each group being tested) greater than one: you simply cannot do a t-test of one (single) sample versus another, since these data will not provide any estimate of variance.

### **Should I run duplicates?**

Obviously, the more data the better and this applies to sample replicates. However, bearing in mind cost considerations, the better question is “Given my budget, should I run replicates of each sample - and if so what parts of the sample preparation should be replicated? - or would I get more information from expanding the study (for example, to add more patients)?”

To answer the first part first, if you run replicates, repeat as much of the experimental procedure as possible. There is relatively little value in simply doing a repeat hybridization just to get a second set of numbers (unless you have serious concerns about the performance of your microarray core). As to the second part, there is no set answer, but in general there will be more information content (and thus more potential value) in a data set that expands the number of individuals in each group than one which achieves the same sample size through replication of observations.

## How do I get expression data into Genetrix?

Genetrix is built primarily with Affymetrix data in mind, so the emphasis here will be on importing GeneChip data. Note that measurements derived from Affymetrix microarrays correspond to defined “probe sets”, which target known or putative genes. There may be more than one probe set for a given gene and generally these multiple sets hybridize to different regions of the gene sequence. Thus, most descriptions of an Affymetrix-based analysis should refer to probe sets rather than to genes, but Genetrix (and this document) adopts the short-hand nomenclature of talking about genes, except in situations where a distinction between the concept of a probe set and a gene is called for.

To obtain a single expression values from the hybridization intensities of probes in a probe set calls for selection of an algorithm to combine the values. Two fundamentally different approaches are possible: the GeneChip approach which treats data from each separately, and the probe modeling approach of dChip and Probe Profiler.

### What is probe set modeling?

In practice, not all probes in a probe set are equally informative and specific for the mRNA species of interest. MicroArray Suite derives its estimates of expression levels for each gene using data from each chip individually and this provides limited opportunity to detect the probes that perform best, within each probe set, and/or to weight the probes according to their performance. Probe set modeling attempts to derive a set of weights, using a data from many samples, that reflect each probe’s performance. Since there is no “gold standard” to refer to (no “correct” expression values that could be used to determine the best performing probes), probe set modeling evaluates probes based on their consistency across multiple samples. Thus a subset of probes that provide consistent data (increasing and decreasing in hybridization signal in concert, from sample to sample) can be assumed to be measuring a real signal, while those that vary more randomly - uncorrelated with other probes in the probe set - are likely to be responding to noise. Note that the “real signal” may or may not be the desired signal; it could equally well represent cross-hybridization.

When probes behave consistently across most samples, but atypically in a small subset, the atypical values can be flagged as probe outliers.

### Is probe set modeling always better?

Not necessarily. In particular, reliable estimation of the weights requires a minimum of around ten samples; if your Project includes fewer samples, and you don’t have an external model file that is applicable, it is probably better not to apply probe set modeling.

### How do I input expression data from an MAS .txt file?

Affymetrix’s MicroArray Suite can be used to generate expression data and presence/absence calls to write these data to a set of .txt files - one file per sample. To import such data into Genetrix, select the Delimited text icon in the Import box. Then follow the steps listed in the dialog:

- (1) *Identify the input file*, which will require specification of the number of probe sets. It is also possible to provide a data set name and a text description, but neither are essential;
- (2) *Specify the format of the input files*. For .txt files from MAS, this will generally be a file with four header lines to be skipped, and tab delimiters - the other fields in this dialog can be ignored. If you do not want to import the data from the control probe sets, these lines can be skipped over by treating them as header lines;
- (3) *Indicate how Genetrix should assign sample names*. It is recommended that you simply use the file name as the sample name. This facilitates linkage of samples back to their .CEL files, if need be, for probe-specific analysis. If the files names are cumbersome or unsuitable, they can always be replaced by an “alternative name” (see below) in analyses and output;
- (4) *Allocate columns*. Genetrix will show the first few lines of the input file. First click on the table header for the column that has the Affymetrix ID, and select Link ID (Affymetrix) from the drop down menu. Then click on the header of the expression data column, and select “Expression Data”. Finally, click on the presence/absence column header, and select “Present/absent”;
- (5) Lastly, step 6 is to *begin reading*.

Once reading has been completed, the data can be viewed through the View/Edit icon.

Note: Before exiting Genetrix, you must **Save** the data into a Project (gtx) file in order to be able to access it subsequently

### **How do I use Probe Profiler?**

The procedure for inputting expression data using Probe Profiler is as follows:

- (1) *Decide whether to use an existing model or generate a new model (set of probe weights)*

In order for probe set modeling to work effectively, the probe performance must be modeled on an adequate set of samples. This can be the **Project’s samples** (or a subset of them), or some **external reference set**. The advantage of the latter is that the probe weights used to generate expression values will be consistent across many Projects (through reference to a common “model”, or set of probe weights). The primary disadvantage is that weights for each probe in a probe set can only be reliably estimated using data in which a signal is present (that is, the gene is being expressed) in reasonable fraction of the samples. Thus a model created on samples of tissue type A will generate weights that are of limited value for tissue B - the probe sets for the genes expressed in both tissues should perform well but the expression level for genes expressed only in B will be poorly estimated.

If the choice is to use an existing model, select the file that contains the weights.

If the choice is to create a new model, select the .CEL files for the samples to be modeled (generally, these will be the samples that are to be imported to the Project). You will be asked

for a model name, the ChipSet type and a set of model parameters. The default parameter settings generally work well.

- (2) *Select any additional CEL files.* These can be processed, using the selected/created model, and then imported to the Project.
- (3) *Select the samples to be imported.*
- (4) *Select the genes to be imported.* Basically, decide whether to include the control genes.
- (5) *Select the QC covariates to import.* QC measures currently available are Noise, Brightness, Outliers, Rare Genes and Saturation; in each case, a single value per sample.

Once reading has been completed, the data can be viewed through the View/Edit icon.

Note: Before exiting Genetrix, you must **Save** the data into a Project (gtx) file in order to be able to access it subsequently

### **How do I import dChip data?**

Unlike Probe Profiler, dChip is not integrated with Genetrix. You must first execute dChip and export the expression values to a tab delimited Excel file. After that, you need to specify the ChipSet type, the samples to input and the genes (that is, with or without the control genes).

### **What about combining multi-chip expression data?**

Probe Profiler and dChip input both allow for merging of data from one probe set with another. When a sample is processed, Genetrix checks for a matching sample name among the samples already stored. A match can be exactly the same name, or one from the same extended chip set (e.g. U133A and U133B). If a match is found, Genetrix pauses before importing the data to offer four alternatives: Overwrite (replace) existing data with the new data, Skip this input sample, or Append (add) the input data as a new sample. To combine data from a multi-chip set, simply ensure that the samples have matching names and request "Overwrite". The data for any probe sets with the same ID (for example U133A probe sets that are duplicated in U133B) will be overwritten, but unique probe sets will be used to extend the list of genes.

### **How do I replace expression data with updated values?**

From the above, it should be clear that updating with new values can be achieved by importing under the same sample name, and requesting Overwrite.

### **How do I add more samples to a Project?**

When probe set modeling is not being used, or when the model used is based on samples external to the Project, the procedure is straight-forward: open the Project, then follow the same import steps as outlined above.

If the original data was created using a probe set model, based on the Project data, there is a

choice to be made: the new samples could be imported using the same model (which would now be based on a subset of samples in the Project), or a new model could be created using all samples (the original set plus the new ones).

**Can I use non-Affymetrix expression data?**

You can, but there two possible difficulties. The first is that Genetrix does not currently include a set of normalizations appropriate for two-color spotted arrays; this would need to be done first, outside Genetrix. Secondly, the input data must include a gene identifier that permits linkage to the annotation database - currently, this can be an Affymetrix ID or a LocusLink ID.

## **How do I get covariate information into Genetrix?**

Covariate information for each sample can be imported from a delimited text file, or entered interactively.

What are sample covariates?

Sample covariates are any numeric information on a sample. This could relate to the RNA sample and its processing (e.g. lab QC values), the biological sample (e.g. histological features), associated clinical information (e.g. disease code) or patient information (e.g. age, gender).

**What are gene covariates?**

Gene covariates are numeric information on a gene. In practice, these data will tend to be generated within Genetrix rather than imported.

**What sort of covariate data is allowed?**

Although all covariate information is numeric, labels can be assigned to specific numeric values and these labels will be used throughout the program (in dialogs and on output) giving the appearance of a string-valued covariate.

**Can I use free-form text?**

Samples and genes each can be given a free-format text comment of any length. This comment is not treated as a covariate, however, and is not analyzable.

**What about dates?**

Dates should be recorded as a six- or eight-digit number (MMDDYY or MMDDYYYY).

**How are survival times stored?**

There is considerable flexibility in how survival-type data (right censored data) can be stored in Genetrix. This is covered in detail below (see Survival time definition).

**Can I import from a text file or Excel spreadsheet?**

You can import from a delimited text file. If the data are in an Excel spreadsheet, save the file to a tab-delimited text file first.

To import sample covariate data, select the Delimited text icon and the Sample covariate tab. Select the input file, define its format, then select columns for the file. Columns are selected by clicking on the heading, and then choosing from the drop down menu. This menu lists all current sample covariates (which allows for importing to an existing covariate) as well as a "(New covariate)" item (to create a new covariate).

Genetrix must be able to match up the input records with the samples already in the Project. The best way to achieve this is to include a sample identifier as one column in the input file, and select this as "Sample label". Genetrix will match this label to the currently defined sample label (see below). If no "Sample label" column is selected, the incoming records are assumed to match the stored records, in the Project file, one-for-one.

When the column contains text information, Genetrix automatically assigns a distinct numeric value to different text string it finds, and adds an label that links the value to the text that it represents. Because it assigns values and adds labels in the order that new text strings are encountered, the order may not be optimal - for example, for a field that records age ranges (as text strings), the coding could end up as 1= "35-44", 2="15-24", 3="25-34". To reorder the labels to a more logical order, go to the Edit/View screen, select the Sample Covariate tab, right-click on the covariate name and select "Labels". Follow the on-screen instruction for re-ordering labels.

### **Can I enter and edit sample covariate data directly?**

The View/Edit icon provides a display of all expression and covariate data, and allows for entering/editing new values interactively.

Right-clicking on a covariate name brings up a menu which includes options to insert a new covariate, delete a covariate, change a covariate name and modify the covariate labels.

To edit a value, simple type in the new value or, if the covariate has labels assigned to values, select one of those labels from a drop-down menu.

To enter values of a single covariate for all samples, start with the first and use the Tab key to move down to each successive record.

### **What is the "Work A" covariate?**

When the Project is first created, it includes one gene and one sample covariate (called Work A). These can be used for any purpose whatever, but typically will be selected by the user to temporarily store results of analyses.

### **How are missing values represented?**

When the format of the input file is being specified, there is an option to indicate how missing values are represented in the file. This can be a special character, such as a "." or a "?", or can simply be two adjacent delimiters.

Within Genetrix, a missing value is displayed as a "?". When entering data interactively, enter a "?" to denote missing. The "?" can be used in other places that call for a numeric value, for example when defining a subset based on a covariate value - for example, SEX = ? would define the subset of all samples with SEX unknown.

## What should be done prior to analysis?

Once data have been imported, but before any analysis can begin, there are a number of actions that will generally be necessary (and others that are often desirable). These actions are of three types: *Labels and settings*, to customize the displays and output to maximize their readability and clarity; *Data preprocessing*, to prepare the data for analysis; and *Analysis preprocessing*, to “organize” the sample covariates in a way that facilitates their effective use within an analysis.

### LABELS & SETTINGS

#### What is a sample label? How does it differ from a sample name, or the alternate sample name?

We will start with the **sample name**. This is usually the name that was assigned to the sample when it was imported and will often match the file name for the associated .CEL file. It may or may not be useful to label samples on displays and outputs, depending on how descriptive, clear and succinct it is. The **alternate sample name** is intended to provide the user with a way to enter a preferred name, for display and output, without having to edit the sample name. If an alternate name is not specified for a sample, the sample name is used in its place. The alternate name may be entered in the Sample Information dialog.

The **sample label** allows for more complicated and descriptive labeling of samples. Click on the sample (yellow) Labels icon to access the dialog that is used to define the sample label. By default, the sample label is set to be the alternate sample name, which in turn defaults to be the sample name if no alternate has been entered. The user can force use of the sample name (ignoring the alternate name) or omit the sample name from the label altogether. In addition, sample covariate data can be included in the sample label. The covariate name must be entered, bracketed by “%” symbols (e.g. %Hist gp%), and more than one covariate can be specified.

#### What is a gene label?

See the description of the sample label above. The gene label is entirely analogous.

#### How are display colors customized?

Screens that show color-coded results will generally include a button that shows the current color-coding spectrum (the **Color button**). Clicking on this button brings up a screen that allows you to change the color choices. For fixed colors and group colors, clicking on a color will show a palette of alternatives.

For the Color Spectrum, there are color choices at either extreme, plus (optionally) colors placed at intermediate positions along the spectrum. Any of these colors can be changed by clicking on the small circle above the spectrum. The position of the intermediate colors can be changed by clicking and dragging them to a new location. Dragging on top of another color bar will delete an intermediate color.

Up to 10 separate color coding schemes can be defined. Back and Next button can be used to move between these schemes to modify them as required. To change from one color scheme to another, right-click on the Color button and choose from the menu.

## DATA PREPROCESSING

The data preprocessing steps may or may not apply to your Project, but at the very least you should consider each in turn to determine its relevance. You will need to Save the Project after making data preprocessing changes to ensure that these changes remain in place next time you access the Project.

### Should the expression data be transformed prior to analysis?

In general, yes. Expression values are typically quite skewed and such highly non-normal data are unsuited to many of the statistical analyses provided by Genetrix. A very simple transformation is to truncate values at 1 (to remove zero and negative values), then apply a log transformation.

Transformations are specified using the Transform Icon, and then selecting from the list of options. The selected transformations are added to a list, and can be removed from the list by double-clicking on them.

It is useful to include a **log transform**, and to have it as the last step in the transformation list, since Genetrix include special programming to recognize and handle this type of transformation. In particular, when log transformation is used Genetrix will (optionally) display the untransformed expression values on the axes and labels of the graphical displays.

### How should sample replicates be handled?

They can be treated in one of three ways.

- (1) There is a specific screen, invoked by the Sample Replicates icon, that allows the user to specify a covariate that contains replicate sample data. All samples in a replicate set must be given the same (integer) value for this covariate. When this covariate is selected, to define replicates, the expression values for samples (in each sample subset) will be replaced by averages over replicate sets (including only the members of the set that are in the subset). Covariate values for a replicate set are drawn from just only of its members, which means that you must make sure the covariate data for the set is recorded for each one of its members. The standard error of an expression value, for a replicate set is calculated from the replicates.
- (2) An alternative approach to handling replicates is simply to include them as separate observations.
- (3) The third approach is to create a sample subset that include one selected member from each replicate set. Click on New in the Subset box, provide a subset name, click the "Remove all" button in the Subset box, then select the Duplicates tab on the Subset Edit screen. The Duplicates screen provides the means to flag (highlight) selected members of a replicate set which can then be moved into the subset.

## **How should gene replicates be handled?**

Affymetrix ChipSets include, for many genes, two or more probe sets with the same name and which link to the same LocusLink entry. Typically, the probes are targeted to different regions of the RNA, which can lead to different estimates of expression level for a number of reasons: different hybridization efficiencies of the probe sets, different cross-hybridization properties, recognition of different splice variants, or different susceptibilities to the effects of RNA degradation.

The options for handling replicates are the same as for sample replicates.

## **How does Genetrix handle presence/absence data?**

When expression data are imported from a MAS or dChip file, presence/absence calls are also available and may also be imported. Probe Profiler does not generate a presence/absence call. These calls can be viewed in the View/Edit display, and the information can be incorporated in several output displays (such as histograms, scatterplots and similarity matrices).

There is also a “P/A definition” icon that may be used to create or modify the presence/absence calls. It allows the user to retain the imported P/A calls (or not), and to add a further criterion based on the absolute level of expression or a comparison of expression level to its standard error (when available).

The Subset Edit screen, which is used to define gene subsets, has a tab that is specifically used to include or exclude genes based on their presence/absence data - for example, to exclude any gene that is flagged as absent in more than 80% of samples.

## **How do I filter the genes?**

We have touched on filtering in answering questions on P/A calls and gene replicates. The gene subset Edit screen can be used to create a subset of genes for analysis (a “filter”), based on a broad range of criteria. Available criteria include selection by gene name, by gene attributes, by covariate value, by measures of data quality, randomly, by P/A or standard error data and by replicate set.

The general approach is to start with a subset that may include all genes, genes in a previously created subset or with no genes (depending on the situation), and then to use the appropriate tab pages to establish the criteria to select genes that need to be added to the subset or removed from it. If you are adding to a subset, the first step is to “Highlight in Add list”; if you are removing from a subset, the first step is to “Highlight in the Subset list”. Once you have satisfied yourself that the highlighted genes are the ones you really want to add (remove), click on “Add highlighted” or “Remove highlighted” to initiate the transfer.

Note that changes to subsets are not automatically retained after exiting Genetrix. To save subset modifications, Save the Project before closing the application.

## **How do I select subsets of samples for analysis? What is the Active subset?**

Sample subsets are created in the same way as gene subsets (see above). Any defined sample subset can be selected to be the **Active sample subset**: this is the group of samples to be used in any analysis or display. In fact, when a set of samples is selected as the Active sample subset, Genetrix acts as if the other samples do not exist.

There is always an **Active gene subset** also, which is the group of genes to be used in any analysis or display.

The easiest way to select an Active subset is to click on the subset name in the list that is shown on the main Icon Screen.

### **What is the Selected subset?**

The Selected subsets (Selected gene subset and Selected sample subset) is a user-selected working subset that can be readily modified (by adding genes, or samples, to it or by taking genes/samples out) and that can be used to highlight genes (or samples) on many of the displays.

## **ANALYSIS PREPROCESSING**

The analysis preprocessing steps are optional. Time spent setting up survival time definitions, and key covariate/key comparison lists generally pays off in terms of a more efficient and informative analysis. However, you can just as easily wait until the definitions/key lists are needed, then define them.

You will need to Save the Project after making analysis preprocessing changes to ensure that these changes remain in place next time you access the Project.

### **How are survival times specified?**

The essence of a survival-time outcome is a **time** and a **sensor indicator**. The time records the date of an event (which could be death, relapse, recovery, disease onset etc.), or the date that the individual was last known to be free of the event (the censor time). The sensor indicator is used to specify whether the time is an event time or a censor time.

There are two ways to specify a survival outcome in Genetrix. The first is simply to import one covariate with a survival time (or two covariates with an “entry” and “event” date, which can be used to calculate the survival time) and a second covariate that is the indicator (taking value 1 if the time represents an event of interest and value 0 if the time is a censor time).

The second approach is to import a survival time, as above, and a separate censor time and leave it to Genetrix to determine which came first - the event or the censor time. In this situation, one or other will actually be “missing”, since a person with an event such as death will not be censored and if censored would not have a date of death. This means of defining a survival time has an advantage in situations where several events of interest are recorded and the user wants to be able to define multiple survival-type outcomes. For example, if the covariates include a date of death, date last seen alive, date of marrow relapse, and date of extramedullary relapse, it is very easy to use these dates to define a survival time (event = death; censor at date last seen), a relapse-free

survival time (event = death or marrow relapse or extramedullary relapse; censor at date last seen) or a time to relapse (event = marrow relapse or extramedullary relapse; censor at date last seen or death (e.g. from other causes, such as toxicity)).

Once a survival definition has been created it is stored in the Project file and will not need to be re-specified.

To define a survival-type outcome, first click on the Survival-times icon. Click on the New button, then:

- (1) Enter the name of this outcome
- (2) If the times need to be computed as the difference between an entry time and an event time, select the covariate that stores the entry day/date.
- (3) Select one or more covariates that record the day/dates of interest
- (4) Indicate whether the censor covariate is an indicator or records a day/date
- (5) Select the censor indicator covariate, or the censor event covariate(s), as appropriate
- (6) Specify whether the days/dates (the Entry time, the Event time and possibly the Censor time) are in the form of dates, or have already been converted to days (or weeks, months or years)

### **What are Key covariates, and how are they selected?**

Many procedures and analyses in Genetrix result in the identification of a group of samples. These could be a group that cluster together, or a group with high expression of a gene of interest or samples with specific covariate values. The obvious question for such groups is whether (and how) they might be different or unusual when compared to all other samples. Do the patients the samples come from have unusually poor prognosis? Was there anything atypical about these patients' disease? Or the patients themselves - were they older than average, have a significantly different ethnic distribution etc. These questions are highly Project-specific and depend on the nature of the covariate data that is available. However, within a given Project, it is likely that the same set of questions would be of interest each time a subset of samples is identified.

Key covariates are a user-selected list of covariates that are to be used to routinely evaluate groups of samples. Clicking on the Key covariate icon allows the user to select the covariates of interest, which can be categorical (leading to a 2xN cross-tabulation of samples in a list/not in a list against discrete values of the covariate), continuous (leading to a comparison of values for samples in the list versus not in the list - a t-test or Wilcoxon comparison) or survival (leading to a comparison of survival for patients in the list versus those not in the list). The Key covariate list is saved in the Project file and is available whenever the user wishes to examine the key properties of a group of samples.

A Key covariate analysis can be obtained directly from the Sample List screen, by clicking on the icon with the yellow key. The analysis consists of each of the requested comparisons, with associated p-values and a thumbnail graphic. Clicking on the thumbnail brings up a more detailed analysis of just that one Key covariate.

There are numerous other places in Genetrix where Key covariates are used. Whenever a Sample List is created, Genetrix automatically runs the list against the Key covariates to see which feature is most significantly associated with the list, and this feature is displayed at the top of the list as a

heading. When clustering generates sample clusters, these can be labeled in the same way - the feature that is most significantly associated with a cluster is determined and displayed. Lastly, one of the Key covariates (termed the “Primary” key covariate), can be used to create thumbnail graphics that can be superimposed on scatterplots to show the properties of groups of samples.

### **What are Key comparisons, and how are they selected?**

Key comparisons have certain parallels to Key covariates. There are user-defined lists of sample covariates that define analyses of interest, but for genes rather than samples. They may be applied to groups of genes (analogous to Key covariates and samples), but are also applicable to single genes. In fact, when multiple genes are evaluated, the expression values for the group are averaged and the vector of mean values evaluated as if it represented a single gene.

Key comparisons are defined by clicking on the Key comparison icon. Seven types of Key comparisons are available.

- *2-group* comparisons compare the distribution of expression values for one group of samples versus another.
- *k-group (distribution)* comparisons compare the distributions of expression values for k distinct group, as defined by k values of a categorical covariate
- *k-group (Box plot)* comparisons compare the median, range and interquartile range of expression values for k distinct group, as defined by k values of a categorical covariate.
- *Survival (medians)* comparisons compare survival of patients with expression values above the median value versus those with expression below the median.
- *Survival (tertiles)* comparisons compare survival of patients with expression values in the lowest, middle and highest tertiles.
- *Continuous* comparisons plot expression for the gene against a continuous covariate or against expression for another selected gene.
- *Time series* comparisons plot mean expression for an ordered (possibly time series) set of samples, as defined by a covariate, optionally grouped by a second covariate.

A Key comparison analysis can be obtained directly from the Gene List screen, by clicking on the icon with the green key. The analysis consists of each of the requested comparisons, with associated p-values and a thumbnail graphic. If the Key comparison is based on a single gene (rather than an average across multiple genes), clicking on the thumbnail brings up a more detailed analysis of just that one Key comparison.

One of the Key comparison (termed the “Primary” key comparison), can be used to create thumbnail graphics that can be superimposed on scatterplots or on pathway diagrams to show the properties of a gene or group of gene.

## How should I approach the analysis?

Gene expression data represents a unique challenge to the investigator, one that calls for special-purpose software and a blend of data management skills, statistical knowledge, machine learning heuristics, data visualization and biological insight. Genetrix is designed to facilitate the investigator in each of these areas, but the complexity of the problem and the comprehensive nature of the software can be daunting for the neophyte.

Every project will be unique in terms of the expression and covariate data available, the study aims, and the opinions of the investigator regarding the most effective analytic approach. While there is no “one-size-fits-all” solution, there are specific approaches to some standard experimental situations (comparing two groups, predicting disease outcome, following a time course etc.) that can be outlined as a ‘recommended’ framework for analysis. In addition, there are some general comments about features of Genetrix, and analytic approaches, that may be useful. The general comments are covered in this section, and more specific guidance on analytic features of Genetrix is available in the sections that follow.

### The “Analysis Chain”

A key feature of Genetrix is the way it facilitates the flow of an analysis by chaining from one display to the next. Thus, for example, a pathway schematic may include line graph thumbnail plots for selected genes; one of these could be selected (clicked on), to bring up a Key Comparison display that includes, among other analyses, a comparison of survival for patients with a high (vs. low) expression of that gene; the samples (patients) in the poor prognosis group could be selected (again, with a single click) to be evaluated using the Key Comparison dialog to see what features they have in common; those patients in a particular subgroup of the poor prognosis group could be listed; the pairwise similarity of gene expression for all genes could be determined for the samples, and this matrix of distances could be used to construct a 3-D multidimensional scaling plot showing how the samples cluster.

There is an almost infinite number of such paths through the analysis, each step directed by the curiosity and questions of the investigator in response to data presented at the previous step. Some chains of thought and investigation may lead nowhere, and the user will want to backtrack, either to start afresh, or to strike off in a new direction from some mid-point in the chain.

Genetrix keeps track of progress through this branching voyage of discovery, recording the steps in the Log (see above), and can take the user back to any intermediate step with a single menu choice.

### Reducing the gene set

Many of the challenges of gene expression microarray analysis stem from the very large number of genes being measured. The vast majority of the genes are irrelevant to any single analysis or question, and their inclusion in the analysis data set does nothing other than obscure, complicate,

slow and confuse the analysis. For the most part, we cannot exclude all the ‘irrelevant’ genes, since we do not know which ones they are, but we can still simplify the analysis by the careful exclusion of genes. A smaller gene set reduces the amount of memory and computer time needed for many analyses - in some cases making possible an analysis that would otherwise be impossible. Fewer genes means fewer tests, for example in GeneScreen (see below) which reduces the magnitude of the multi-testing problem.

Mixing genes that are known to have desirable characteristics (for example, show different distributions of expression in different sample subgroups), with “non-informative” genes will dilute out the contribution of the former group.

### **Single gene or gene combination?**

The analysis can be directed in one of two quite different directions. The first is to look for effects and relationships that apply to individual genes; the second is to look for patterns that are inherent in a group of genes. The choice will depend on the scientific question, and while each approach has merits, it is important to understand the associated limitations.

The individual gene approach has a number of advantages. The results (associations of a gene with a covariate of interest) are easy to explain, display and interpret (for the most part). Because of this, they are amenable to replication by others, and/or practical application. The methods used to identify the genes are often very straight-forward (for example, t-tests of comparisons) and familiar. The principal disadvantages are (1) that, with each gene treated separately, it is impossible to avoid massive multi-testing problems and (2) any information related to the interactions between genes is ignored.

Multivariate methods that integrate contributions from many genes simultaneously, such as clustering, can take advantage of complex interactions within the data. While these methods will generally not suffer particularly from the multi-testing problem, the risk is of over-fitting. With many genes in a single model, often involving the simultaneous estimation of many parameters, it is common for the algorithms to find solutions that fit the ‘training’ data set extremely well but which will fail when challenged with new data. In essence, the algorithm has fitted a model to the idiosyncrasies of the training data set. This necessitates the independent testing of models on a ‘test’ set of samples.

A compromise approach that takes advantage of the strengths of both methods is to apply single gene screening to select a small group of genes that are then modeled multivariately.

### **Statistical vs. Biological validation**

In many microarray experiments, finding a gene of interest amongst tens of thousands of genes can be likened to finding a needle in a haystack ... with the needle cunningly disguised to look exactly like a piece of hay. If the gene of interest behaves sufficiently differently - for example, shows a large fold change in expression in an experimental vs. control group - or if the sample size is sufficiently large, it will stand out (statistically) and be identified. More commonly, it will show up as statistically significant, but not at a level that separates it from a host of other genes with chance associations. At this point, the investigator needs to apply prior knowledge - about the disease process and/or the gene characteristics - to identify the most promising leads.

Thus while analysis may begin as a statistical exercise, at some point the input and guidance of the biologist becomes indispensable.

## How do I do scatter plots?

Scatter plots are available directly from the Icon screen, or from a number of other dialogs, such as the clustering output displays. Genetrix provides a very flexible scatterplot facility, in terms of selection of axes, size and color of points, number of plots displayed etc. Once a plot has been displayed, there are many ways to modify the plot, for example adding labels to points, or to interact with the plot, such identification of specific plot points of interest. Most of these features are covered in the FAQs below.

The first choice is whether to plot genes (green scatterplot icon) or samples (yellow scatterplot icon).

### AXES

#### How do I specify the axes?

There can be 2 or 3 axes (2-D or 3-D plot). When plotting genes, these axes can be expression values for a single sample, average expression values for a group of samples, expression values projected onto a principal component, or values of a gene covariate. When each point represents a sample, the axes can be expression values for a selected gene, average expression for a group of genes, projection on to a principal component, or values of a sample covariate.

Once an axis combination has been selected, it is necessary to click on “Add X-Y” to save these settings, then click on “Display”.

#### How do I use an average value for a group of genes (or samples) for a plot axis?

For the “Grouped” axis selection (average expression for a group of genes or samples), the group is defined by a covariate, a comparison operator and a value. For example, the group could be “Age > 50”. If the covariate values have labels, the group is defined using those labels. For example, the covariate could be “Hist gp”, the operator could be “=” and the value could be “Ductal”.

#### How do I plot fold changes?

When **genes are being plotted**, the expression values have been **log transformed**, and an **expression average value** (the “Grouped” selection) has been selected for an axis, this average can be replaced by a fold change (by checking the “Fold change” box). The fold change is the ratio of average expression for samples in the group to the average expression for all other samples in the Active subset.

#### What is a “Principal Component” plot axis?

If there are n samples, the expression values for each gene could in theory be plotted in n-dimensional space. We can reduce this down to a manageable 2 or 3 dimensions by selection of 2 or 3 samples, but this clearly discards a great deal of information. An alternative is to weight each sample, convert the expression values for each gene to a weighted sum across all the samples and

then use this weighted sum as a plot axis. While there are obviously an infinite number of possible weighted sums, there is only one for which the variance is a maximum: this represents the first principal component.

The second principal component is the direction in n-dimensional space along which the expression values show the second highest variance, with the important proviso that the first and second principal components be at right angles (in n-space). Higher principal components are similarly defined.

Since the principal components are chosen to “explain” as much gene-to-gene variation as possible, a significant fraction of the information content of the data can be captured by a small number of principal components (and displayed in a low-dimensional plot).

Note that the above description applies to plots of genes, but the same process can be applied to plots of samples.

### **Can I do multiple scatter plots?**

Yes. Every time you click on “Add-XY”, a new plot is specified and when “Display ” is clicked all specified plots are shown at once. A short-hand way to create multiple plots is to select multiple X-axes and/or multiple Y-axes before clicking “Add X-Y”. Thus selecting samples A,B,D for one axis and covariates P and Q for the other will create six plots (A vs. P, A vs. Q, B vs. P, B vs. Q, C vs. P and C vs. Q).

### **What is “Add X-X” for?**

Add X-X is a minor variant on Add X-Y. Instead of taking each X-axis selection in combination with each Y-axis selection (see above) to create multiple plots, every X-selection is paired with every other X-selection.

### **In a multiple plot, can I expand a selected plot?**

Yes, just Shift-click on a plot. .

### **Can I automatically scan through the plots?**

When multiple plots have been created, and one has been selected and expanded (see above) moving the cursor to the left or right arrows at the top of the display; will cause the display to cycle between all of the defined plots, at a speed that depends on how far the cursor is from the mid-point of the arrow rectangle

### **Can I plot standard error values?**

Yes. When selecting a gene or sample as an axis, click on the “SE” option. The standard error of the expression value will be used in place of the expression value itself.

### **How do I get a 3-D scatter plot?**

Select the X and Y axis as usual, then select the Z-axis tab and select the Z-axis.

### **Can I make it rotate?**

When a 3-D plot is displayed, three rectangles (colored red, green and blue). To rotate the image, move the cursor until it is within one of the rectangles. The axis of rotation will depend on which rectangle is selected; the direction of rotation depends on whether the left or right half of the rectangle is selected; and the speed of rotation depends on how far the cursor is placed from the mid-point of the rectangle.

### **Is there a way to look at marginal distributions?**

Move the cursor to the end of an axis. A small red dot will appear. Click on this dot, and a reduced-dimensional marginal plot will be displayed. Thus starting with a 3-D plot, the marginal display will be a 2-D plot using the other two axes; with a 2-D plot, a histogram of values from the other axis is shown.

### **Can I zoom?**

Yes, for 2-D plots. R-click and drag to draw a rectangle that defines the zoom area. To return to the un-zoomed (original) display, click on “Full plot” at left.

### **Can I make the X- and Y-scale the same?**

Yes, for 2-D plots. Click on “X range  $\Leftrightarrow$  Y range”. To return to the un-zoomed (original) display, click on “Full plot” at left.

## **POINTS**

### **How can I change the size of the points?**

The most direct way is to click on the up or down arrows, in the left margin. Alternatively, go into “Settings” and enter a new value for the point sizes.

### **Can I add standard error bars to plot points?**

Select the “Plot Points” tab and check the box(es) to request S.E. bars on the X- or Y-axes.

### **Can I indicate presence-absence data on the plot?**

There is an “Indicate P/A” box on the left of the plot. When that is selected, Genetrix will use a different plot symbol to depict the P/A status of X- and Y-axis values (if one or both of these are values derived from a single sample (or gene)). If the X-axis value is flagged as “absent”, a horizontal line is displayed (instead of a point); if the Y-axis value is absent, a vertical line is used; if both are absent, a plus is displayed.

Optionally, points which have X- and/or Y-axis values that are flagged as absent can be removed

from the plot. “Exclude Abs” excludes any point with either an X- or Y-axis value absent; “Exclude Abs-Abs” only excludes a point if both X- and Y-axis values are absent.

### **Can I plot centroids (average expression for a group of points)?**

On the “Plot Points” display, there is an option to plot Cluster Averages. This plots the average expression value for a group of points, where the group is defined by specified values of a covariate.

### **What are the options for coloring points?**

When defining a plot, “Symbol Colors/Sizes” presents a screen that allows for customization of the colors. The colors can be based on values of a selected covariate, the expression value for a selected sample (or selected gene, for plots of samples), or the average expression for a group of samples (genes). These values are converted to a color code using the current color palette. This palette provides a user-customizable “spectrum” of colors for coding of a continuous values and a set of “Group” colors for discrete-valued selections (for example, for color choices based on a discrete valued covariate). See “How are display colors customized”.

### **Can I highlight a subgroup of points?**

The “Symbol Colors/Sizes” display has a check box for highlighting (in white, by default) points that belong to the current “Selected” subset.

### **How do I find the point for a specific gene (sample)?**

To search for (and highlight) a single point, click “Find” on the left margin, select a gene (or sample), then after returning to the plot, select the Highlight “Point” box.

### **How do I change settings for a plot after it has been displayed?**

Click on the “Settings” button or, for color changes, on the Color spectrum button.

## **INFORMATION**

### **How do I get information on a point?**

Moving the cursor to a point will give the **gene (or sample) label** at the top of the display.

Clicking on a gene (or sample) point will display the **Gene (or Sample) Information** screen.

Right clicking on a gene point, or on a point that represents a centroid (average) for a group of genes will create a **Key Comparison** display showing the selected properties of the gene or gene cluster. Right-clicking on the centroid for a group of samples creates a **Key Covariate** display for that group. In either case, this will work only if the user has defined the Key Comparisons/ Covariates.

If a point has a label (see below), clicking on the label will show the Key Comparison (or Key Covariate), if appropriate and defined, or the Gene (Sample) Information displays.

If the point represents a **centroid**, hovering over it will display its label (at the top of the display, and will change the color of all the points that belong to the group. Ctrl-clicking on the centroid point will list all members.

### **How do I select a group of points?**

A region of the display can be selected by clicking and dragging the cursor: this region will be a rectangle, a circle or a freeform line, depending on a menu choice, on the left margin. All points in the region will be listed in a List Genes (or List Samples) dialog, which provides a host of options for evaluating the list.

### **Can I label points? What if I want to change the label?**

Before you can label a point, you must click on “Label” in the left margin to change the L-click action from “Select” (i.e. show information on) to Label. Then L-click on each point you want to label. A maximum of 100 points can be labeled. L-click on a point a second time to remove the label.

Labels can be moved by clicking and dragging them to a new location.

R-clicking on the display provides a menu of options, including the options to label all points, to label all highlighted points (if appropriate), to hide the labels or to erase all labels.

The R-click menu also allows the user to change the content and appearance of a label. The content can be changed by selecting a “Notes” or a covariate (to use in place of the Label) or by clicking on “Change label” to redefine the Label (see Sample labels in Data Preprocessing). The appearance can be change in terms of color, background color or font size.

### **Can I label centroids?**

There are many ways to label a group of genes or samples. Make sure the “Label” option has been selected (in the left margin). R-click and select “Gp: Content & Appearance”. Then select from the menu of labeling options.

Individual centroid points can be Ctrl-clicked on to show the label, or R-click to get the menu and request “Label groups” to label all centroids.

### **What are thumbnails, and how do I get them?**

Thumbnail are miniature graphics that display selected information about a gene, a group of genes, or a group of samples. It is derived from the list of selected Key Covariates (for samples) or Key Comparisons (for genes); the “primary” Key Covariate or Key Comparison is always used in a thumbnail label.

To **add a thumbnail** to a scatterplot, first make sure the “Label” option has been selected (in the

left margin). R-click and select “Gp: Content & Appearance”, then select Key Covariate (Comparison) thumbnail. If no Key Covariates (Comparisons) have been defined, clicking Apply will bring up a dialog that allows you to define the Key Covariate (Comparison); otherwise, the current “primary Key Covariate (Comparison) is selected for the thumbnail graphic. Select genes by clicking on them; select gene or sample centroids by Ctrl-clicking on them.

Thumbnails can be **moved** by clicking and dragging them to a new location.

R-clicking on the display provides a menu of options, including the options to add the thumbnail to **all points** (maximum of 100), to add a thumbnail to **all highlighted points** (if appropriate), to **hide** the thumbnails or to **erase** all thumbnails.

**Clicking on a thumbnail** brings up the full Key Covariate (or Comparison) display.

When the **cursor is placed within a Key Comparison**, information relating to its position within the thumbnail is displayed above the scatterplot. For example, if the thumbnail shows a time series, the expression value, time point and group (if any) corresponding to the position of the cursor within the thumbnail are shown.

Key Comparison thumbnails for single genes can include the **gene label**: click on “Add label as heading” in the “Label Content & Appearance” dialog.

When the Key Comparison thumbnail shows a mini-scatterplot of a selected gene against another (reference) gene, the choice of reference gene can be quickly changed by Shift-clicking on any point.

## ANALYSIS

### Can I add a line-of-best-fit?

The “Analysis” settings dialog allows for either a linear or quadratic line of best fit, with or without confidence bands.

## What is “GeneScreen”?

Gene screen is used to apply a selected statistical test to each gene individually. This test measures the significance of association of gene expression levels to values of a specified covariate. Examples might include t-tests comparing expression values for ‘treated’ samples versus ‘untreated’ samples, or the value of the expression level in predicting survival, in a Cox model.

### How do I request a GeneScreen?

A simple GeneScreen is requested by selecting the type of covariate (2-group or dichotomous), k-group or survival-type outcome, and then the type of test. For 2-group outcomes, the two groups are defined in terms of a covariate, and comparison operator and a value (or a value label). For example, the groups could be Age > 45 (versus all other samples), or Race = White (versus all other races). Note that if the comparison group (“all other races”, in the second example) is too all inclusive, and a more specific comparison is desired (e.g. White versus Black), you would need to first create a subset that included only the White and Black samples and make this the Active subset before starting GeneScreen.

For k-group outcomes, you need only select the covariate that has k distinct values. For survival outcomes, you must select a pre-defined survival-type outcome (or define a new survival-type outcome).

The GeneScreen will start when you click “Run”.

### How do I interpret the GeneScreen results?

Results of the GeneScreen are represented in four graphics and a pair of gene lists.

*Gene Lists.* The genes with the smallest p-values are listed. For most tests it is possible to distinguish between genes for which a high expression is associated with the selected outcome (the “positive” list), and those for which a low expression correlates with the outcome (the “negative” list); in these instances, two lists are created. Each list includes the gene name and its associated p-value.

Since each gene is individually tested, there can be thousands of tests performed, with the potential for (likelihood of) many false positive associations - that is, associations that are statistically significant by chance alone. The implications of this, and some approaches to handling this ‘multi-testing’ problem, are covered in detail in the section “How do I know which results are “real”?”

*Significance vs. Magnitude plot.* The largest graphic plots the p-value (on a log scale) against a measure of the magnitude of the association. This can be helpful, since a small effect that is highly significant may be of less interest than a larger effect that is perhaps less significant. How the magnitude is measured will vary depending on the type of statistical test. For 2-group comparison, the natural measure is a fold change in expression from one group to the other. For survival-type data, the measure is the relative event rate associated with a fixed (2-fold) change in

expression.

As the cursor is moved over a point on this plot, the name of the gene is selected from the Gene Lists, and the p-value and measure of magnitude of association is shown.

*Distribution of the test statistic.* All values of the test statistic are shown, as a histogram, which is compared to the theoretical distribution of the statistic. Under the null hypothesis that there is no relationship between expression and the covariate, one might assume that the observed and theoretical distributions would coincide; to the extent that they don't, this could be evidence for a presence of a group of genes for which the association with the covariate is "real".

In practice, the assumptions that underlie the theoretical distribution might not be justified. For example, the t-statistic will show the desired t-distribution, under the null hypothesis, provided the tests are independent and the data are approximately normally distributed. In practice, neither assumption is likely to be true, and the theoretical t-distribution will not be truly accurate. A more robust distribution to measure the observed values against can be obtained through generation of a permutation distribution. See "How do I know which results are "real"?"

*QQ plot.* The QQ plot provides an alternative representation of the extent to which the observed distribution of the test statistics deviate from the theoretical distribution.

*FDR plot.* The FDR plot shows the number of genes that correspond to a range of False Discovery Rates. Thus, for example, a histogram bar representing 100 genes, at an FDR of 0.10, indicates that if the user selects the 100 most significant genes, he/she can expect that 10% of them will be false positives. Bear in mind, that this assumes applicability of the comparison (theoretical).

The FDR plot actually shows two bars for each FDR value - the upper one represents genes that have 'positive' association with outcome, the lower bar represents significantly negative associations.

For an explanation of the false discovery rate, see "How do I know which results are "real"?"

### **What are the different options for the FDR?**

There are several ways to calculate the FDR and since it is not clear which one is most appropriate for gene expression data sets, Genetrix provides a choice of methods. The simplest method is the "One Step" approach which ..... Probably a better approach is the two Step or Adaptive methods that .....

Each of these methods make an assumption that the genes are independent of each other, which is most certainly not true, with the result that more genes will be identified at a given FDR level than should be. The "Correlated stats" check box will provide an adjustment to the FDR calculations to reflect the fact that there are inter-correlations between genes in the data set. However, this adjustment is probably rather conservative, so that fewer genes will be identified at a given FDR than should be.

### **Can I get more detail for a single gene?**

You can click on a single point in the scatterplot, or on a single gene in the Gene List. Either way, this selects a gene and Genetrix displays a detailed analysis of the relationship of that gene to the selected outcome.

For **2-group** and **k-group** comparisons, the single-gene display is relatively simple: it just shows the distributions of expression values within the 2 (or k) groups and a set of test statistics.

For **survival-type** outcome, the single-gene display is a bit more complicated. Initially, a single Kaplan-Meier curve is shown, representing all samples, along with Cox regression statistics (including adjusted statistics, if the comparison was adjusted through inclusion of additional variables). The samples can be split, however, into low vs. high expressors (of the gene) by clicking on the histogram of expression values at the upper right. When the expression range is dichotomized in this way, two Kaplan-Meier curves are drawn and a logrank test statistic is calculated. The threshold for separating low from high can be moved by clicking and dragging the line. The range of expression values can be split into three (or more) divisions by clicking once again on the histogram; a dividing line can be eliminated by dragging it on top of another line.

An alternative way to split the expression values, to create groups of samples for comparison, is to let Genetrix do it. The “**Median**” button splits the expression values at an approximately median point - approximately, because it steps the dividing line across the histogram and will not split the samples within a bar. “**Tertiles**” does the same thing to create three approximately equally sized groups. “**Best split**” tries all positions for the separator to find the one that gives the smallest p-value; “**Best trend**” tries all positions for two separators to find the combination that gives the smallest p-value on a test for trend across the three groups.

Important Note: Any method of dividing the samples that allows the user to explore multiple alternatives and to choose the one with the best (or most significant) separation of survival curves is creating a multiple-testing situation that effectively invalidates the p-value. Thus this approach is fine for exploring the data and for finding ways to split the data to give maximum differences between groups, but for Heaven’s sake, don’t quote the p!

### **Can I look at the properties of a groups of (the most significant) genes?**

You can select groups of genes from any of the graphics or lists, for display in a gene list. The List Genes dialog provides a broad range of tools to evaluate and characterize the distinctive features of a group of genes (see “I have found some genes. Now what?”)

To copy the top n genes from either the positive or negative gene lists to the List Genes dialog, set N and then click on the “List” button.

To select genes in the tail of the observed distribution of statistic values (from the histogram), click and drag the mouse to draw a line through the required histogram bars. To select genes from the QQ plot, click and drag the cursor to draw a line.

To select genes from the scatterplot, click and draw to draw a rectangle surrounding the required genes.

To select genes from the FDR graphic, click on a bar. If you click on the upper (green) section of the bar, the genes from the positive gene list are selected; the lower (aqua) section selection from the negative gene list.

### **When would I want to save the p-values?**

Since the p-value (or the test statistic) measures the degree of association between expression level for a gene and the outcome, it may be a useful piece of information to include on printouts of gene lists, or to use as part of a gene label, or as a scatterplot axis, or to use to color-code a point or other symbol representing a gene. Saving GeneScreen p-values (or test statistics) to a covariate makes these values available in other parts of the program.

### **When would I want to run a permutation distribution?**

P-values derived from comparison of test statistics with the statistics' theoretical distribution are quick and easy to generate, but because there are often concerns about the appropriateness of the theoretical distribution it is advisable to at least cross check these against the more robust approach (comparing observed values against a permutation distribution). See "How do I know which results are "real"?" for a further informations on permutation distributions.

### **Can I adjust for covariate data in the analysis? When might I want to?**

Funny you should ask that. Why, of course you can, as long as a multivariate statistical method is chosen. Thus, for 2-group comparisons, you must employ logistic regression; k-group does not have a multivariate method; while the only method for survival analysis, Cox regression allows for more than one variable.

As to when you would want to, with a covariate or another gene in the model (as an adjusting variable) the GeneScreen tests measure the extent to which each gene shows independent, incremental association with the outcome variable. For example, an Cox regression (survival) GeneScreen with a "disease severity" covariate used as adjusting factor will evaluate all genes for their ability to provide additional prognostic information.

There are several ways to use a covariate. It can simply be added as an adjusting variable, and used unaltered. Alternatively, the covariate can be converted to a dichotomous (0/1) value, or log transformed. If the covariate taken k discrete values, it can be split into k-1 dichotomous covariates (0/1).

### **What if I have matched pairs of samples?**

The only method that currently supported matched analysis is the t-test. To get a matched t-test, click the "Match" tab and select the covariate that specifies the matching: each matched pair of samples must share a common value in this covariate.

### **Can GeneScreen consider two or more genes in combination?**

That is exactly what is happening when one (or more) gene is used as an adjusting variable. The selection of genes to be added as adjusting variable can be made automatically by selecting a

(forward) Stepwise analysis.

### **How is GeneScreen used to predict outcomes?**

Predictions use a leave-n-out approach. N samples are omitted, GeneScreen is run to find the gene(s) most strongly associated with outcome and these gene(s) are used to predict an value of the outcome covariate for the n samples. This procedure is then repeated with a different n samples, repeatedly, until all samples have been omitted from one analysis.

For **2-group** (logistic regression) analysis, the predicted covariate is readily determined by applying the model to the observed expression value(s) and estimating the probability of membership in each group. These probabilities are displayed on two histograms and a white line (which can be moved) sets a classification threshold. A 2x2 table gives the predicted vs. actual group membership, and the user can click on a cell in this table to see the samples falling into each group. Useful statistics such as the sensitivity and specificity of the prediction are shown. A second tab (“Gene List”) shows the genes that were used in the logistic regression model, and the frequency of their use

For **survival** analysis, the situation is more complicated because the Cox regression model does not model the survival time as a function of the covariates, but rather models the distribution of survival times against the covariates. Thus, for a new sample it is possible to use the Cox model to predict its survival distribution (that is, its expected survival curve).

Deriving a single predicted survival time from an expected curve is something of a challenge. One obvious choice is to use the median survival (the time point at which the survival probability drop below 0.50), but individuals with the best prognosis may have predicted curves that never cross 0.50. The alternative is to get away from the idea of a predicted survival time; rather to predict something like the survival probability at one year, which can be used to rank individuals from worst to best prognosis. Genetrix provides complete flexibility in this regard. The predicted survival curves for each individual are plotted, with a white line that either defines a fixed time point, or can be dragged (across to the y-axis) to define a fixed probability. Where this line intersects each survival curve will determine the single predicted value for each person. These predictions can be stored unchanged or ranked first, then stored. This same display has small squares on each line to show the actual survival time - an open square indicates a censored survival time. Placing the cursor over the square will highlight the line and show the sample name, the prediction, the expression value(s) for each gene in the Cox model (for this sample), that was used to make the prediction.

The predicted values can be plotted against the actual values (“Scatterplot”).

## How do I display time series (ordered) data?

The Line Graph is designed to display and analyze projects which include samples that have an intrinsic order, such as those that come from a time series of observations.

### How is the ordering specified?

By default, the samples are shown in the current sort order. Usually, however, the investigator will create and select a covariate that defines an ordering - for example a covariate with values 0 = 0hrs, 1= 6hrs, 2=12hr, 3=24hrs. This will define a time series order.

### What if the experiment includes two or more observations at each time point?

When a variable is selected as the Order By variable, all samples with the same value will be combined. If there is no "Group By" (see below), the mean values of samples with the same Order By covariate is shown, and if S.E. bars are requested, the standard error is calculated from these replicates.

### How do I compare two or more groups?

The Group By covariate allows the samples to be split into two or more groups, based on values of this variable. These groups are displayed as different lines (colored separately) on the line graph

### With more than a handful of genes, the display is just a mess of lines. Help.

With more than ten genes, the display gets very busy and hard to interpret. Obviously, one approach is to focus down on a minimal set of genes. Selection could be made before entering the Line Graph, or could be achieved by selecting a subset of interest within Line graph. Alternatively, you may to use a "Tiled" display, which plots each line in a separate graph.

### How do I find genes with a certain pattern of expression?

In the non-tiled display, you can either R-click on a gene (the line for the gene) or click on the Define Pattern button. The former will display a "Pattern" of expression, for each time point and within each group, and the user can click and drag any point to a new location to modify this pattern. A simple click on a point makes it a "Don't care" point: when searching for genes with a matching pattern, this time point (in this group) will be ignored. Once the pattern has been defined, click on List Neighbors, to get a list of the genes that most closely match the pattern. You will be given an opportunity to specify how many neighbors to list, and what similarity measure to use (e.g. Euclidean, correlation etc.). If the Define Pattern button is used, a "null" pattern with the same values for all time points is shown, and the user can click and drag as before to define the search pattern.

In a tiled display, R-clicking on a line is used to select genes with patterns closest to the selected line. These are shown highlighted, and may be listed using the List Neighbors button. To return to the non-highlighted display, R-click anywhere other than on a line.

# What is the similarity matrix?

## How is similarity measured?

Similarity is based on the expression data, either the numerical expression value or the presence/absence calls. For expression values, there are a range of similarity measures possible including correlation, absolute correlation and Euclidean distance. The difference between correlation and Euclidean distance is that the latter looks for genes (or samples) with similar patterns *and magnitudes* of expression, while the former is only concerned with the pattern of expression. Thus if the level of one gene controls expression of another, but the level of expression is very different for the two genes, they will have dissimilar Euclidean distance but a similar correlation. The absolute correlation is useful when a strong negative correlation is biologically relevant, as it might be if gene A was an inhibitor of gene B.

For presence/absence calls, the similarity of genes (samples) is based on the log of the odds ratio of the 2x2 table of calls for the two genes (samples).

## The color coding is useful, but what if I want detail?

There are a number of options for viewing the data that underlie a given cell.

*Summary statistic.* When the cursor is placed over a cell, the summary statistic (correlation, Euclidean distance, odds ratio etc.) for that cell will be shown above the matrix. It is also possible to display this summary statistic on the matrix by Shift-left-clicking on a cell.

*Scatterplot.* When expression data are being used, clicking on a cell will create a scatterplot for the pair of genes (samples) selected.

*2x2 table.* When presence/absence data are being analyzed, clicking on a cell creates a 2x2 table of the P/A calls for the two genes (samples).

## Can I select groups of cells?

It is possible to select more than one cell a scatterplot display. One way to do this is to Ctrl-click on a series of cells, then click on the last one; all selected pairs will be presented in a matrix of scatterplots. An alternative way is to click and drag a rectangle which will select all cells in (or partly in) the rectangle for display as scatterplots.

## How can I identify rows/columns of interest?

There are several ways to modify the matrix, the labels and the ordering to highlight and organize the display to be more informative.

*Covariate.* A covariate value (for a selected covariate) can be appended to the gene (sample) labels. This can be very useful when looking for patterns of similarity that might be expected to be related to a known characteristic (such as histology), that is recorded as a covariate.

*Highlight.* If there is a group of genes (samples) of interest, and these are represented in a subset, this subset can be highlighted on the display.

*Limit.* It is often useful to focus on a small subset of neighbors of each gene (sample) that are closest to it. The Limit option allows you to do just that (see below).

*Reorder.* If there is interest in examining the similarity within subgroups that are defined by a covariate (for example, within distinct histological subgroups) the records can be sorted according to that covariate. It is also possible to use the data to reorder the genes (samples): that is, find an order that brings the similar genes (samples) together. See Multidimensional Scaling, below.

### **What happens when I “limit” the display?**

The display can be limited so that only the closest “n” neighbors of each gene (sample) are shown; the rest are colored black. Note that each gene (sample) gets to pick its closest neighbor(s). For example if n is set to 1, gene A (sample A) will have at least one cell showing a pairwise distance to another gene (sample) - B, say, the nearest neighbor of A. Note however that it is possible for gene A to show have other neighboring distance retained, since gene A may be the nearest neighbor of genes C and D, say.

### **Is there a way to analyze the structure in the similarity matrix?**

The suggestions outlined above, under “How can I identify rows/columns of interest?” provide ways to manipulate the display to accentuate and understand any inherent structure. However, a more powerful tool that analyzes all the similarity data simultaneously is Multidimensional Scaling (see below).

### **What is Multi-Dimensional Scaling?**

The best way to understand MDS is by an example. Imagine you are given a table of ten U.S. cities showing all the distances between each pair of cities. Without knowing any geography it should be (and is) possible to use this table to position the cities on a 2-D surface so that the distance between the points match the distances in the table. The technique for finding where to put the points is MDS.

In this example, the MDS solution will exactly match the input data, and the analysis is trivial and rather uninteresting. However, if the input table measures something different (e.g. the ethnic similarity of each pair of cities) the cities placed close together will be those most ethnically similar, and the resulting pattern will provide a useful representation of the data.

MDS can also be seen as a tool for dimensionality reduction. Thus using a similarity matrix that stores a set of distances in n-dimensional space (such as would come from a measuring the Euclidean distance between samples, based on expression of a set of n genes) a 2-D MDS would find an optimal positioning of each sample in two dimensions that reflected as closely as possible the data derived from n-dimensions. Genetrix provides options for dimensionality reduction to a one-, two- or three-dimensional display. In the case of a 1-D MDS, the result is simply a reordering of the genes (or samples) to bring each gene (sample) close to its neighbors (as defined

by the distance measure).

## Why cluster?

Good question. Clustering provides a means of using inherent structure in the data to define groups of similar genes (or samples). Similarity is based on patterns of expression, typically measured for a large number of samples (genes), generally giving equal weight (or importance) to each measurement, and with a defined distance metric (see below).

In the early days of analysis of gene expression microarrays, clustering seemed an obvious solution to problems of interpreting an overwhelming amount of raw data, but with experience many investigators found that it rarely provided the sort of clear-cut, unambiguous and reproducible results they expected. That is not to say that clustering cannot be helpful; answers to FAQs below provide some advice on how to use Genetrix's clustering tools most effectively.

### What genes and samples should I use for clustering?

Before clustering you need to decide which points to cluster, and for each point what measurements will be used. Thus when clustering samples, decide which samples to include and what genes will be used to determine sample-sample distances.

Deciding on the samples to include is usually fairly straight-forward - it is usually dictated by the experiment and the purpose of the clustering. Choosing the genes to include can be more problematic. Perhaps the best advice is that more is NOT better. The computational requirements (both time and computer memory) for most clustering approaches rise steeply with increasing numbers of points being clustered. Equally important, it is highly unlikely that more than a small fraction of the tens of thousands of available genes provide information that is useful and biologically relevant to any given clustering problem, and the inclusion of all the extraneous genes greatly increases the chance that artifactual solutions will be found. Of course, identifying that small subset of relevant genes may not be easy, but certain steps immediately suggest themselves: filter genes to eliminate those that are probably not actually being expressed in the tissue under study and consider applying a GeneScreen analysis to identify genes that have statistically significant relationships to covariates of interest.

### How do I know if the clusters mean anything?

Clustering algorithms can usually be relied upon to find clusters, but are any of them meaningful? The best advice is to be aware of the limitations of clustering in general, in the context of your data; to apply a range of clustering approaches (see How do I know if the clusters mean anything?) to look for consistency; to use conventional statistical methods to filter the data prior to clustering; and to take full advantage of prior biological knowledge (about genes) and clinical/biological knowledge (about samples) to aid in interpretation of the clusters.

*Apply a range of clustering approaches.* At its simplest, this can be advice to use several methods (e.g. k-means, SOMs and hierarchical clustering) and perhaps more than one distance metric (see below) to see how consistent the patterns of clustering is. Some methods (such as k-means) use random starting points before iteratively searching for an 'optimal' clustering; while the result will be guaranteed to represent a local optimum and it may well not be a

global optimum. That is, different initializations may converge on different (local optimum) solutions. Thus another simple way to determine the validity of a set of clusters is to see the effect of restarting the clustering (with a different random initialization).

*Compare the results to non-clustering approaches.* If genes that appear to be important in defining clusters are also those that are most significant in GeneScreen analyses, the confidence is strengthened in the validity of both approaches.

*Examine the effects of minor modifications to the data.* It seems a reasonable supposition that any clustering solution that is dependent on the particular random initialization values, or is substantially altered by removal of a small proportion of the points or by minor perturbations of the expression data, should not be given much credence. The meta-clustering tool (see below) is specifically provided to re0run the clustering many times, after applying such data modifications, to see which aspect of the clustering are repeatable.

*Independently test the results.* One reason clustering methods can so readily find clusters, even when no structure is present in the data, is that they can capitalize on the noise in the data to find groupings that are 'real' for the data set being analyzed, but have no generalizability to any other data. That is, the features they use to form clusters are totally idiosyncratic to the training data set. This is classic over-fitting situation, and the most direct and powerful way to test for over-fitting is to set aside a test set of records. If the clusters found within a training data set can be used to make meaningful classifications in the test data set, the clusters clearly provide useful information.

*Examine the cluster patterns.* An effective way to examine the clusters is to plot the data in a rotating 3-dimensional scatterplot, using the first three principal components as axes, with the points color-coded according to cluster membership. The human eye is very effective at assessing degree of cohesion within clusters and the separation between them.

*Interpret the clusters using prior knowledge.* Clearly a group of genes that appear to be randomly selected is intrinsically much less interesting (and probably more likely to be artifactual) than a group that share important biological features. The same is true for clusters of samples. The cluster is of greater interest still if these biological characteristics have relevance to the experiment and/or the disease under study.

Genetrix provide a number of functions that assist the user in examining the biological features of a cluster.

Highlight points in a subset of interest. If there are genes (or samples) that are of particular interest, these can be shown on the cluster display (scatterplot or hierarchical tree).

Examine individual points in a cluster Individual points - for example, points that are misclassified, or that lie on the boundary between two clusters - can be clicked on to bring up a display that identifies the point and summarizes its known features.

Label the cluster There is a range of options for labeling groups of genes or samples (i.e. clusters). For gene clusters, perhaps the most useful is the option to ask Genetrix to scan

all Attributes of the genes (Gene Ontology codes, pathway membership, subset membership) for the Attribute that is most significantly linked to this cluster. That is, the most significant comparison, when comparing the frequency of the Attribute in the cluster (vs. all genes not in the cluster). For sample cluster, the analogous procedure is to search all the Key Covariates for the descriptive feature that best describes the cluster (again, compared to all samples not in the cluster). This could be a discrete characteristic, such as “Histology = ductal” or “Gender=male”, a continuous characteristic, such as “Age”, or an outcome such as “Poor prognosis”.

Summarize cluster properties - thumbnail labels An extension to the cluster labeling is to label groups of genes or samples with thumbnails. This display, in graphical form, a selected feature of the cluster in terms of the covariate information. Thus, for clusters of samples, the cluster can be labeled with a Kaplan-Meier survival curve thumbnail showing the survival of patients in the cluster vs. all others. An example for gene clusters would be a line graph that tracked the mean expression levels for a series of experimental conditions, separately within two or more subgroups of samples.

### **Should I have a separate test data set?**

The generalizability of a set of clusters is very difficult to establish without some sort of independent test. The simplest approach is to set aside a subgroup (perhaps a third or half of all samples, randomly selected) for the test, but this can be problematic if the investigator feels that reducing the sample size will significantly compromise the analysis. An alternative method is to apply a leave-n-out approach which is much more computationally demanding but has minimal impact on the effective sample size.

Finally, there will be situations where the interpretation of clustering will be quite clear even without a test data set. For example, if the clusters correspond to known characteristics of samples (genes) - for example, an unsupervised clustering based on expression values alone, that neatly separates one disease subtype from another - it is reasonable to assume that this clustering reflects biological reality.

### **Which clustering method should I choose?**

While each method has its own advantages, it is hard to lay down guidelines for use of one versus another. As mentioned above, there is value in trying several approaches to look for consistency.

### **What is the difference between supervised and unsupervised clustering?**

Unsupervised clustering looks for natural groupings of points (genes or samples) based only on similarities of patterns of gene expression within each group. Supervised clustering uses a known classification (represented in a covariate) to create the clusters and then applies those clusters to create decision boundaries and to classify additional points (for which the classification covariate is not known).

### **What distance metric should I use?**

The result of clustering depends very much on what metric is used for calculating the “distance”

between two points (genes or samples), in n-dimensional space. Possible metrics include correlation, absolute correlation and Euclidean distance. The difference between correlation and Euclidean distance is that the latter looks for genes (or samples) with similar patterns *and magnitudes* of expression, while the former is only concerned with the pattern of expression. Thus if the level of one gene controls expression of another, but the level of expression is very different for the two genes, they will have dissimilar Euclidean distance but a similar correlation. The absolute correlation is useful when a strong negative correlation is biologically relevant, as it might be if gene A was an inhibitor of gene B.

## K-MEANS CLUSTERING

### How does k-means clustering work?

K-means is conceptually very simple. The program randomly selects k cluster-center points then assigns every data point (gene or sample) to the cluster-center that it is closest to (using the chosen distance metric). Once all points are assigned, the cluster-centers are re-computed, this time as the actual center (mean value) of the points currently in the cluster. The distances from each point to each cluster-center are re-computed, and some points will now need to be re-assigned. This process is repeated until the iterations no longer result in change to cluster membership.

Note that the user must specify the number of clusters (k) ahead of time, although Genetrix does provide the option to specify a range of values of k, in which case each is tried in turn. Note also that since k-means starts with a random initialization, it will commonly converge to a different solution each time it is run.

## HIERARCHICAL CLUSTERING

### How does hierarchical clustering work?

The common form of hierarchical clustering, and the one implemented in Genetrix, is agglomerative - that is, it starts with every point belonging to its own cluster (of size 1) and progressively joins the clusters that are closest until there is only one cluster left (comprising all points). In this method, the resulting clusters will depend not only on the choice of distance metric for measuring the distance between two points, but also on how one defines the distance between two clusters. The choices for calculation of the distance between two clusters are: (1) determine the distance between each member of cluster A and each member of cluster B (i.e. all pairwise combinations) and average these distances; (2) find the minimum distance between any member of cluster A and any member of cluster B; (3) find the maximum distance between any member of cluster A and any member of cluster B; and (4) replace each cluster with its centroid (an average value for all cluster members) and determine the distance between centroids.

### What is optimal ordering?

Hierarchical clustering as described above is deterministic - given the same data and same choice of method, the same sequence of cluster merging will occur. However, when representing the clustering as a branching tree, as is customary, the choice of the order of the two branches at each node is arbitrary. When n points are being clustered, there are n-1 branch points and the order

can be flipped at each branch, leading to  $2^{n-1}$  possible orderings of the tree. Sometimes the arbitrary choice of ordering does not matter, but more often there is useful structural information that is lost if the tree is unordered. Optimal ordering in Genetrix applies an algorithm that searches all possible orderings for the one that minimizes the distance between adjacent points in the display.

A dramatic demonstration of the value of optimal ordering can be seen when genes and samples are clustered in the “Einstein.gtx” data set (using Euclidean metric, and minimum point distances). Compare the patterns of expression before and after the hierarchical tree is ordered (on both the gene and sample axes).

## SELF ORGANIZING MAPS

### What are self organizing maps?

In many ways, self-organizing maps (SOMs) are very like k-means. A predetermined number of nodes (or cluster-centers) are randomly placed among the data points and each data point is then assigned to the closest node. Unlike k-means nodes, the SOM nodes have a predefined geometric relationship to each other. For example, 16 nodes may be assigned to have a 4x4 grid layout; a consequence of the geometry is that the concept of ‘neighbor’ and ‘distance’ between nodes, on this grid, has meaning. Contrast this to k-means where there is absolutely no relationship between nodes (clusters).

On each iteration, the points assigned to the node *plus to a lesser extent points in neighboring nodes*, are used to move the node. Then the points are re-assigned as necessary and this process continues until convergence. If nodes had no ‘neighborhood’ and responded only to the points assigned to them, SOMs would act just like k-means. However, the result of combining data from neighboring nodes is to enforce the SOM grid geometry on the final node layout. That is, if the SOM grid is set up as rectangular, the nodes positions will converge in a way that roughly preserves this rectangular arrangement, albeit in distorted form.

### What does the map graphic show me?

The simplest interpretation of a SOM map is that it represents the SOM nodes, laid out in their original geometry (rectangular or hexagonal) with the assigned clusters attached to each node. It is possible to examine the properties of each node individually or to create a graphical display that shows selected characteristics of all nodes simultaneously. The advantage of the SOM approach (over a k-means approach) is that, since the nodes have a relationship to each other defined by the geometry, it is meaningful to look for patterns and trends across nodes, and within regions of the map. Thus, for a SOM of samples, the sample clusters in a region of the map may share certain characteristics, such as experiencing a poor therapeutic outcome.

## BAYESIAN CLUSTERING

### What is Bayesian clustering?

## **SUPPORT VECTORS**

**What are support vectors?**

## **NEAREST CENTROIDS**

**What is nearest centroids?**

**What are shrunken centroids?**

## **META-CLUSTERING**

**What is meta clustering?**

Meta-clustering is a procedure that allows the user to run clustering routines repeatedly, under circumstances that are expected to give somewhat different answers (leaving out points, different initialization, adding noise), and combine the cluster information from each run into an overall table.

# How do I capture the results of analysis?

## GENETRIX LOG

### What is the Log?

The Log is an automatically created record that charts the analytic steps during a Genetrix session and can optionally include bitmap images and text from the analysis.

### Where do I find the Log?

The Log is placed in a subfolder, within the “Log” subfolder of the Genetrix application’s folder. There is a separate folder created for each Genetrix session, so you may want to tidy up occasionally by erasing Logs that are not needed. The folder’s name includes the date of creation and a sequence number. The folder will include an XML file, which is the Log itself, and any captured screen images in separate JPG files.

### How do I save images to the Log?

There are several ways to save images to the Log.

- The *Log button* is the most direct: it captures a bitmap image of the current dialog into the Log.
- The *Capture button* provides a little more flexibility. It has functions beyond the Log (see below), but may be used to capture an image to the Log. The advantage of the Capture button is that it allows the user to specify a particular region of the image to save.
- Images may be *automatically saved* to the Log (see below).

### Can I save gene or sample lists to the Log?

When the Log button is clicked from within the List Genes dialog, Genetrix will write the gene list into the log, including optional annotations and covariate data. Similarly for samples in a Sample List.

### Can I annotate the Log?

Cntl-clicking on the Log icon displays a dialog that allows you to enter a brief text annotation to be save in the Log.

### How can I automatically capture all output to the Log?

Cntl-Shift clicking on the Log icon toggle AutoLog on and off. When AutoLog is on, the dialog bitmap image is automatically copied to the Log for each dialog visited. Note that a long analytic session, with AutoLog enabled can end up saving many bitmap images to the disk, using up a

considerable amount of disk space.

### **How can I view the Log?**

If you click on the Log icon on the main screen, or Shift-click on the smaller Log icons on individual dialogs, the GViewer will start and will display the current contents of the Log.

The dialogs visited during the session are shown in a tree structured list at left, with nodes that include graphics shown in blue. The node name is the name of the dialog, by default, but if an annotation has been provided, this is used in place of the name.

### **How do I print Log images?**

If you R-click on a node that includes a graphic image, you can select a menu choice to send the image to the printer.

### **Can I share Log files with others?**

If you R-click on a node that you can choose to convert that node and all sub-nodes to HTML format to distribute to collaborators.

## **SCREEN CAPTURE**

### **How is Screen Capture different from the Log?**

The Screen Capture button (which looks like a butterfly net) can be used to copy a bitmap image to the Log, but it has other functions as well. The image can be sent to the clipboard, directly to the printer or saved to a file. When save to a file, the degree of compression can be specified. A useful feature of Screen Capture is the ability to capture only part of the dialog. For some dialogs, there is a region or list of regions that can be selected, and for all dialogs there is the option to use the mouse to click-and-drag a rectangle to define a region.

## How should I approach two-group comparisons?

*Preamble:* Reviewing the literature, it seems as through there are almost as many ways of approaching the analysis of gene expression data as there are papers. This is partly because the field of gene expression analyses is relatively new and there has not been enough time for a consensus on optimal strategies to emerge, partly due to lack of suitable analytic software, and partly a reflection of the special needs of individual experiments and the research interests of individual statisticians.

Thus it is perhaps presumptuous to make recommendations for an analytic strategy, but we choose to do so because (1) some users of Genetrix will have had limited knowledge or experience with gene expression analysis, and will welcome guidance, and (b) we know Genetrix and its features better than anyone, and can offer advice on how best to utilize these features toward a specific goal.

Two-group comparisons represent some of the most common types of analyses. These include comparison of one samples from one disease versus another, one subtype versus another, one behaviour (e.g. “Metastatic”) versus another, pre- and post-treatment or exposure or other form of manipulation etc. For these comparisons, two of the most common questions are:

- (1) **What genes are differentially expressed in group A versus B?** What are the characteristics of these genes?
- (2) **Can expression data be used to separate the groups (or predict group membership)?** What is the minimal (or optimal) set of such genes? If there are covariates that are already effective at classifying the groups, can expression data be used improve that classification.

Some suggested steps for a 2-group comparison analysis are given below; little detail is provided for each step, intentionally, to avoid duplication with other sections of this document.

### DATA INPUT & PREPARATION

#### Data Input

- PP or dChip
- Data quality assessment
- Covariate data - 2-group covariate
- other data

#### Data Preprocessing

- Sort on 2-gp cov
- Xform
- PA definition
- Filter genes
- Define genes of interest → subset(s) - X
- Create pathway, if necessary
- Select cov of interest and use to define key com and key cov

## IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES

1. **GeneScreen.** A good starting point is to rank the genes according to the difference in expression values in the two groups. GeneScreen provides several statistical 2-group tests; the t-test is a reasonable choice (It is efficient and its assumption of normality is generally justified in log-transformed expression data).

- **Permutation p-values.** Check the accuracy of the reported p-values for the t-statistics by running comparing against a permutation distribution.
- *Select the most strongly associated genes.* The first place to look is at the false discovery rate (FDR) histogram. If there is a set of genes (corresponding to a particular p-value threshold) with a useful FDR rate, this may be the best set to select. Alternatively, you can choose the “n” most statistically significant genes for a positive association (high expression associated with the selected group) and/or a negative association. A fixed p-value cutoff can be used (bearing in mind the problems associated with multiple testing). Finally, a combination of strength of association (fold change) and p-value can be used when selecting a region from the scatterplot.

Irrespective of the source of the list, the properties of the selected genes can be evaluated using the following tools:

- Attribute analysis. The first thing when examining a gene list is to look at the Attribute match table shows. If any interesting Attribute is listed, you can find which genes in the list have that Attribute by clicking on the left side of the table. A more detailed Attribute analysis is available through the tree icon.
- Similarity matrix. The similarity matrix helps to address the question of which genes, predictive of group membership, are highly correlated with each other (and which are not, and thus might be independently predictive). Individual scatterplots of one gene against another can be viewed by clicking on a cell.  
**Multidimensional scaling** (MDS) is a graphical tool that uses the similarity matrix data to group genes that are mutually correlated with each other. Natural groupings may be best delineated in a 1-dimensional MDS.
- Highlight subset. If there is a gene subset previously defined to be of interest, the overlap of this subset and the list can be highlighted.
- Save to subset. It is often useful to use this list of genes to create a new subset.
- *Click on single gene.* Single genes can be selected from the GeneScreen results, either by clicking on a scatterplot point or selecting a gene from one of the drop-down lists. A detailed 2-group analysis, including histograms of expression values in each of the two groups, will be displayed for the selected gene. The Wilcoxon statistic can be checked to see whether it yields a p-value similar to the t-test; if not, this may be evidence for an unacceptable degree of non-normality in the data..

Further investigation of the gene of interest can include:

- **Gene information.** The Gene Info button provides, among other things, text gene annotation, gene Attributes and links to Internet-based databases. Buttons on the Gene Information screen provide two additional analyses:
    - Neighbors**, which will find genes with the same pattern of expression as the selected gene.
    - Key compare** which compares high-expressors of the gene with low expressors, in terms of selected sample characteristics
  - **List atypical cases.** The 2-group analysis includes two histograms of expression values, which can be used to identify and list selected cases (for example, atypical cases such as those in the “lower” expression group that have a high expression). Such cases can be examined in detail, using the **Key Covariates** function.
  - *Save p-values to a gene covariate.* This provides the option to use the p-values elsewhere in Genetrix, such as to color points on a scatterplot to reflect each gene’s association with the 2-group classification.
- 2. Scatterplot of genes.** Average expression for one group can be plotted against the average for the other to detect genes that are over- or under-expressed in one group relative to another. The points can be color-coded to reflect the p-values determined through a GeneScreen analysis to show the statistical significance of the change in expression.

If the expression data are log transformed, the same data can be displayed in the form of a fold change (Y axis) vs. average expression (X-axis).

- *Select groups of genes.* The differentially expressed genes or other regions of interest in the scatterplot can be delineated to create a list of genes. Such lists can be evaluated in the same way as described earlier (under GeneScreen).
- *Highlight groups of genes.* If there is a gene subset previously defined to be of interest, the points corresponding to that subset can be highlighted.
- *Select interesting genes.* Single genes can be selected by clicking, to bring up a gene information display (see GeneScreen).

Alternatively genes can be labeled to feature specific genes of interest (such as those more highly over- or under-expressed).

- Label with gene name. Labeling by name clearly identifies the genes.
  - Label with thumbnail. Labeling with a thumbnail superimposes informative gene-specific graphics on the scatterplot.
- 3. Pathways.** If specific pathways are of interest, either because of prior knowledge or based on the results of the analysis, this pathway can be displayed.
- *Color code with GeneScreen t-test p-values.* This highlights the genes that are differentially expressed.

- *Display thumbnails.* Labeling with a thumbnail superimposes informative gene-specific graphics on the pathway.

## PREDICTING GROUP MEMBERSHIP

The purpose of this analysis is to identify and classify genes that can separate the two groups and determine the effectiveness (specificity/specificity) of the prediction. The ‘predictive’ genes will inevitably be differentially expressed, so that this aim has much in common with that of the previous section. However, there are issues and questions, and analytic approaches that apply specifically to class prediction.

The actual group membership is known in a training set and our aim is to find genes whose expression can be used to replicate this classification for new cases. Three approaches are:

1. **GeneScreen, with predictions.** Small number of genes, multivariate, allows for adjustment.
2. **Nearest Centroid**
3. **Support Vectors**
4. **Clustering, using differentially expressed genes**
  - *Multidimensional scaling (MDS)* of samples, based on genes known to be differentially expressed. MDS using all genes would tend to group samples that have similar patterns of gene expression, not necessarily separating the two groups. However, when the MDS is based on genes that are selected to be differentially expressed in the two groups (see GeneScreen, in previous section), the MDS will inevitably cluster together samples in their respective groups. Repeating the MDS with progressively fewer genes provides a
  - *Principal Components Analysis of samples*, using subset Y of genes, color by gp
  - Hierarchical clustering



## **How do I use expression data for classification?**

**What tools are available for prognostic factor analysis?**

## **How should I analyze time-course data?**

# How do I know what results are “real”?

## MULTIPLE TESTING

### Why is “multiple testing” a problem?

When tens of thousands of genes are tested individually, many will be expected to be statistically significant by chance alone. Since 1 in a thousand tests will reach 0.001 significance by chance, there will be 40 “false positive” results (at this level of significance) in 40,000 tests. If there is one true positive, at the same significance, there is no statistical means to distinguish the true from the false positives.

### Can I adjust for multiple testing?

The simplest approach is to demand a much higher level of significance. At  $p=0.0001$ , only 4 false positives are expected; at  $p=0.00000125$ , there is a 1 in 20 (0.05) chance of a significant association; this is essentially the Bonferroni approach to multiple testing. The problem is that unless the ‘signal’ is very strong or the sample size very large, we will not usually expect to see significance at this level and many genes that are of biological interest will be discarded because they do not meet the very stringent (and conservative) significance levels created by a Bonferroni adjustment.

There are alternative adjustment methods, that are less conservative than a simple Bonferroni adjustment, and GeneScreen applies a step down multi-testing adjustment to the set of p-values generated when the permutation distribution is used. Although the step down approach is an improvement over Bonferroni, it is still less than ideal. The fundamental problem is that the problem is being framed within the context of traditional statistical thinking, which seeks to set a defined limit on the probability of seeing a **single** false positive at the threshold significance level. This approach works well for focused studies, with limited multi-testing, but is too restrictive when thousands of genes are being tested. A more logical approach is set a p-value that is realistic - that is, that is likely to identify genes of interest - and then to determine the false discovery rate.

### What is the False Discovery Rate?

The false discovery rate (FDR) is an estimate of the proportion of tests (genes identified) reaching significance at a given p-value threshold that will be false positives. Note that we are not insisting that there be a low probability (e.g. 0.05) of a single false positive, as in conventional testing; rather, we accept the inevitability that there will be false positives (for a realistic p-value), and only wish to know approximately what proportion of the significant associations are false positives (or equivalently, what proportion can be expected to be true - non-chance - associations).

## RANDOMIZATION

Randomization of the data can be used to generate distributions of statistics under a null

hypothesis, to use in the estimation of p-values.

### **What is a permutation distribution and why is it used?**

When looking for structure and associations in the data, we will often calculate a statistic (e.g. a t-statistic, comparing values in two groups) and compare this statistic to its theoretical distribution (e.g. the t-distribution) under the null hypothesis. Underlying the derivation of a theoretical distribution is a set of assumptions, which may or may not hold in practice. If there is any doubt about the appropriateness of a theoretical distribution (or, as in some situations, a formula for the theoretical distribution of a statistic is not available), the statistic can be compared to its permutation distribution. In essence, the statistic is calculated repeatedly on data that has been ‘randomized’ - randomly shuffled to remove any signal and ensure that the null hypothesis is true - and the set of statistics so-generated is used to create a permutation distribution.

In Genetrix, testing against a permutation distribution is provided as an option in GeneScreen. The theoretical distribution assumptions for the various GeneScreen tests vary, but typically include the assumption that the samples are independent, that the genes are uncorrelated and (for many tests) that the expression values follow an approximately normal distribution. If these assumptions are known (or suspected) to be untrue, a permutation distribution can be generated to provide an alternative reference for evaluating the observed statistics.

### **What does “shuffling” of samples achieve?**

When samples (with their covariate information) are shuffled with respect to the gene expression data, any systematic association of expression patterns with sample characteristics will be lost. Some associations will still be seen, through chance alone, but these will reflect the noise in the system. Thus shuffling samples (which can be achieved through the Transformation dialog) may be useful as a check on results that appear to imply structure in the data, but may be artifactual.

For example, clustering algorithms can usually be relied upon to find clusters, whether these reflect order in the data set or not. If the same clustering methods, when repeated on shuffled data, show a relative lack of structure it increases confidence in the validity of the clusters in the unshuffled data. In effect, the shuffling provides a single observation from a permutation distribution.

### **How can adding noise to the data be useful?**

It may occasionally be useful to add noise to the data to see how robust particular study results are. If results fluctuate dramatically with the addition of small amounts of noise, it would indicate that the results should be interpreted with extreme caution.

## **RESAMPLING**

### **What is the purpose of cluster resampling?**

Clustering algorithms can generally be relied upon to find clusters, even from random data. Even when structure exists in the data, the solution that a clustering method finds may be somewhat

‘arbitrary’: different starting settings could lead to different local minima (i.e. different clusters), or small changes in the data could have profound effects on the clusters. Obviously, it is useful to have a means of evaluating how robust a clustering method, for your data, and/or to determine which components of the clusters are reproducible in the face of differing start values and/or minor alterations to the data.

The Meta-clustering tool in Genetrix provides a means to re-run a clustering algorithm repeatedly, with changes in the data or changes in the methods initialization settings. The results from all the clusterings are accumulated to derive a matrix of pairwise similarities, which measure the frequency with which each pair of genes (or samples, if clustering samples) ended up in the same sample. This similarity matrix can in turn create a 2- or 3-D plot, using multidimensional scaling, with genes (samples) that cluster together in aggregate, being placed close together on the plot.

## STATISTICAL VALIDATION

### When should I use an independent test set?

With many more observations (genes expression values) than samples, it is easy to build models that are over-parameterized and **over-fitted**. To put it another way, multivariate models and clustering algorithms may appear to perform well when evaluated in terms of the ‘training’ data, but this success may be due more to modeling the noise or randomness in the training data set than to any biological reality. The only convincing proof that such models have validity outside the training data set is to apply them to new data: data that was not used to estimate parameters or establish clusters.

There is a close parallel between the over-fitting problem and the **multi-testing** problem and many people will advocate the use of an independent test set for evaluating the true significance of genes that have been selected from a GeneScreen-like approach. The argument is that the set of genes with a p-value less than 0.001, say, will include a mix of false- and true-positives, and testing this short list of genes in a second data set will serve to validate the true positives and eliminate the false positives. Indeed, a test data set may be used for this purpose, but there is a counter-argument that says that requiring a gene to pass two separate tests at 0.001 and 0.05, say, is not intrinsically different than requiring it to pass a single test at a more stringent p-value (0.0005, say), taking advantage of the greater power that comes from pooling the two data sets. Indeed, it has been shown, in some circumstances at least, that pooling the data is a more effective strategy than splitting it into train and test sets.

### Why would I need a second test set?

The test set only does its job if it is truly independent of the training. This means the test data not only cannot be used to estimate parameters, but it cannot be used to choose between alternative models. It should not be used at all, until a final model has been decided upon and fitted.

When there are enough samples to set aside a second test set, an effective strategy is to use the training set to fit the models, the first test set to help to choose between alternative models, and keep the second test set untouched until the very end, to evaluate the performance of the final, ‘best’ model.

### **What is leave-one-out cross-validation (LOOCV)?**

The problem with setting aside a test data set, which typically includes 30-50% of the samples, is that many experiments are not large enough, by reason of cost and/or limited patients/samples, to permit this. LOOCV provides an alternative approach. A test data set of size one (or some other small number) is set aside, the model is fitted to all remaining samples and then tested against the sample(s) that were left out. This process is repeated many times, typically until each of the samples has been included in the test set, and the test results accumulated across all repeats. In this way, nearly all the samples are used to define each model, all the samples get to be used as test samples, and at no time has a sample been tested against a model that it helped build. The price to be paid is that models must be fitted many times over.

### **Why would I want to leave “n” out?**

The leave-one-out cross validation can be modified to leave more samples out of the training data set. The advantage of this is that the computational burden is less and the results of testing the omitted samples (the “predictions”) can be more robust.

When **only one sample** is omitted on each iteration, the training set changes minimally from run to run, and the predictive models are highly correlated with each other. Thus while the predictions are unbiased, they are highly correlated and have a large variance. The effect of this is that the results are probably not very reproducible.

**With larger n**, the models and predictions are less correlated, the variance of the predicted values is reduced and the user can have more confidence that the results of a repeat experiment will not be too different. The disadvantage of a larger n is that each model is based on a smaller training set, and at some point this reduced sample size can have a significant adverse effect on the variance of the test-sample estimates. A rule of thumb advocated by some is to set n equal to 10% of the full data set.

## **BIOLOGICAL VALIDATION**

Typically, while statistical analyses go a long way towards organizing, displaying, simplifying and analyzing the data, a pure statistical approach rarely provides clear, unambiguous and definitive answers. Given the large number of genes studied, the complexity of biological systems, and inherent limitations of the gene expression microarray technologies, it is too much to expect that a purely numerical approach will suffice. Fortunately, no experiment is conducted in isolation - there is a vast amount of information available on genes, gene interactions and pathways that can be used to place expression data into a biological context.

A large part of Genetrix is devoted to providing tools that the investigator can use to integrate biological knowledge into the analytic process. The following sections describe these tools.

## What are gene Attributes?

Gene Attributes are gene characteristics or annotations that place genes into discrete descriptive categories. The most extensive categorical descriptions are the **gene ontology** (GO) annotations that have been assigned to many genes, and are available to Genetrix through the LocusLink annotation database. The three main gene ontology subgroups are Molecular Function, Biological Process and Cellular Localization.

The full gene ontology can be represented as a tree, with general descriptors being subdivided into progressively more specific terms. A gene may be assigned a GO code from a terminal node (leaf) of this tree, or if it cannot be assigned to any of the highly-specific descriptors of the leaves (because it has a function that is more accurately represented by an intermediate node, or because the precise specificity of the gene has yet to be established) it may be assigned to an intermediate node. Because of the hierarchical nature of the ontology, a gene can be characterized by the node to which it belongs or, for broader categorization, to any earlier node along its branch. Thus, for example, melatonin receptor 1A has been assigned a GO code for “Melatonin receptor”, which is a branch of “G-protein coupled receptor”. Going further back the tree, it is more generally characterized as a “Transmembrane receptor”, a “Receptor” and a “Signal Transducer”. In Genetrix, genes assigned to a node are considered to have that node Attribute, as well as the Attributes of all higher level (broader) categorizations.

In addition the GO Attributes, membership in a pathway is treated by Genetrix as an Attribute, and the user can choose to have membership in selected gene subsets treated as Attributes.

### How do I find the Attributes for a specific gene?

The Gene Information dialog will tabulate the Attributes for a selected gene (assuming the annotation file includes such information for this gene: not all genes are fully annotated).

### How do I find genes that have a given Attribute?

There are two ways to approach this.

In Gene Information, you can double click on the Attribute listed for a given gene to select that Attribute and then click on “List” to see all other genes (in the Active Gene subset) with the same Attribute. If more than one Attribute is selected in this way, you can choose to list genes that have all the selected Attributes or genes with any of the selected Attributes.

The full Attribute tree can be displayed either from the Icon screen, or from the List Genes dialog. In either instance, the genes belonging to a node can be listed.

### How do I determine the Attributes of a group of genes?

Genetrix determines the Attribute properties of a group of genes by constructing and testing a set of 2x2 tables for every Attribute. This 2x2 table divides the genes into (1) all those in the group vs. those (in the Active subset) that are not in the group, and (2) those with the Attribute vs. those not having the Attribute. For each 2x2 table (and thus each Attribute) there is an **odds ratio** that

measures the degree of association between group membership and possession of the Attribute and a **p-value** that reflects the statistical significance of the association.

Genetrix can use this list of odds ratios and associated p-values to find the Attributes that are most strongly over-represented in the group (compared to genes not in the group). In the List Genes dialog, this process is automatic: whenever this dialog opens to display a gene list, the Attributes most significantly (or, optionally, those with the highest p-values) associated with the list are shown as a header. The best match is shown for each major subtype of Attribute (GO Molecular Function, Biologic Process and Cellular Localization, Pathway and Subset). The List Genes dialog also includes an icon which can be used to show the full Attribute tree, with all the odds ratios and p-values.

A very similar approach can be used to label gene clusters. In this instance, only one label is desired: the most significant (highest OR) from all Attribute subtypes (GO Molecular Function, Biologic Process and Cellular Localization, Pathway and Subset) is selected.

### **Can I define my own Attributes?**

An Attribute can be thought of as a characteristic and a list of genes with that characteristic, which is essentially what a gene subset is. The user can check a box in the Gene Subset Select/Edit dialog to have a defined subset treated as an Attribute.

# What can Genetrix tell me about specific genes?

## How do I find a specific gene?

The Gene Information dialog provides the ability to search for a gene. The Gene Information dialog can be reached directly from the Icon screen, but is also reachable from numerous places in Genetrix - basically anywhere that individual genes are represented, such as points on a scatterplot or genes in a pathway schematic.

Once in the Gene Information screen, you can enter a search string next to the Find button, then click on Find. The search string can be text in the gene name, the LocusLink ID, the affymetrix ID or the gene symbol.

## Where do the annotations in Gene Information come from?

The annotations are derived from LocusLink, plus limited data from the U.C. Santa Cruz Golden Path Web site. The data provided (for homo sapiens) include:

- Gene symbol
- Gene name
- Gene annotation
- Location in the genome
- Attributes
  - ▶ Gene Ontology codes
  - ▶ Pathway membership

## How can I find other genes ‘like’ this one?

It depends what you mean by ‘like’.

*Genes with similar names* are most readily found by first sorting the gene list by name so that, for example, the “Interleukins” will appear consecutively in the data display.

*Genes duplicates* (those with the same name and LocusLink ID) are automatically flagged in Gene Information: the number of duplicates in the database (for the current gene) is shown on a button at the bottom of the display (unless there are no duplicates) and clicking on this button will create a Gene List of the full set.

*Genes with the same Attributes* or combinations of Attributes as a given gene can be obtained directly through the Gene Information Screen. When a gene is selected, its Attribute set is shown in a table. Double clicking on any term in that table selects it, and displays a “List” button. Clicking on that button creates a gene List of all genes with that Attribute. It is possible to select a second (or more) Attribute, and then request a list of all genes with both the Attributes or all genes with either of the selected Attributes.

*Genes with a similar pattern of expression* can be obtained by clicking on the “Neighbors” button.

### **How can I tell whether there are multiple probe sets targeting the same gene?**

When a gene is selected in the Gene Information dialog, the number of probe sets that match that single gene is shown on a button at the bottom of the dialog (If the number is just one, the button does not appear).

### **What does it mean if two probe sets (same gene) have different expression values?**

It could mean a number of things. Generally two probe sets targeted at the same gene will have been designed to hybridize to different regions of the gene. Factors which could affect the signal strength include:

- The probe set might target different splice variants.
- A probe set that was directed at more 3' regions would be more susceptible to the effects of RNA degradation.
- One probe set might experience greater cross-hybridization
- One probe set might have 'better' probe sequences that yield a stronger signal.

### **Can I search the Web for more information, directly from Genetrix?**

The Gene Information dialog provides a set of links to selected Internet databases. The user can readily modify and extend this set to include additional links.

## **I have found a group of genes. Now what?**

### **What does the table at the top of the List Genes screen tell me?**

Whenever a Gene List is created it is compared to all other genes (Active genes not in the list). The results of this comparison are presented in a table that shows, for each of the major Attribute subgroups, the Attribute that is most significantly different in the list (vs. not in list). Optionally, this analysis can find the Attribute with the largest odds ratio. By default, the analysis will ignore Attributes that are not present in at least  $n=2$  genes but this can be changed to any value of  $n$ .

### **Can I compare the list against a specific gene Attribute?**

The Attribute match table (see above) provides a summary of the best matches to the list, but if you want to look at the complete list of Attributes, click in the Attribute icon (the tree).

### **Can I compare this list against another list?**

There are several ways to cross reference one list of genes against another. For all of them, the initial step must be to store list A as a gene subset. Then when list B is generated ...

The genes in the list A can be highlighted (shown in red) in the Gene List (list B).

If the list A has been flagged as an Attribute, the degree of association of list A and B can be seen in an Attribute analysis (see above).

The overlap between two lists can be determined if both are gene subsets. In the Gene Subset Select dialog, select each subset and click on Venn Add. This creates a Venn diagram which shows how many genes the lists have in common and, by clicking on the Venn segments, the genes in one list but not the other or in both lists can be copied into a new Gene List.

### **Can I get a hard copy, preferably with annotations and covariate data for each gene?**

Click on "Copy Gene List to ... File".

### **Can I save the list to the Log?**

Click on the Log icon.

### **When would I select Key Comparisons?**

Key Comparisons uses sample covariate data to examine properties of a gene or group of genes. Examples include a comparison of survival for samples with high vs. low expression, and comparison of expression levels in males vs. females. Since, for groups of genes the expression level is always the *average* expression of genes in the group, a Key Comparison analysis would be appropriate if the genes in the group had similar patterns of expression. (If not, the average expression of the group might not be very meaningful).

## **What is MeanScreen?**

MeanScreen averages the expression values for all gene in the list to create a centroid vector, and applies a selected GeneScreen statistical test to this vector.

## **What happens if I ask to copy the list to a covariate?**

Genetrix will write a fixed value (specified by the user) into a selected covariate (which can be a new covariate) for all genes in the list. The covariate value for all other genes will be unaltered or can be initialized to a different value. There are many analytic and visualization functions of Genetrix that take advantage of information in covariates and this process allows the information in a Gene List to be transferred into a covariate.

## **What do the Gene List icons do differently from the same icons on the main screen?**

For most, the only difference is that only genes in the list are passed to the analytic or visualization function. This is true for Pathways, GeneScreen, Chromosomal Mapping, Line Graphs, Expression Matrix and Similarity Matrix.

Other icons that appear both on the Gene List and main screen include:

*Key Comparisons.* On the main (icon) screen, this icon is used to select Key Comparison; in Gene List the icon runs a key Comparison based on gene in the list.

*Attribute tree.* On the main (icon) screen, this icon is displays the Attribute tree; in Gene List the icon requests an analysis to show the relative representation of genes in the list, for each Attribute.

*Scatterplot.* On both the main screen and the Gene List dialog the Scatterplot icon request a scatterplot that includes all Active genes. The only difference is that in Gene List, the genes in the list are copied into a temporary subset, which is made the Selected subset, so that they can be readily highlighted in any scatterplot that the user creates.

## How does Genetrix incorporate pathway information?

Genetrix includes a set of pathway schematics that can be displayed, with gene expression or gene covariate data represented in color-coded or graphic thumbnail form. In addition, pathway membership is recorded as an “Attribute” of a gene, so that evaluations of the properties of groups of genes automatically includes tests of association between the gene list and membership in each of the pathways.

### What pathways are available?

The available pathways are those derived from KEGG (mostly metabolic pathways), two adapted from the medical literature (Kohn ), and a number of purpose-built signal transduction pathways.

### Can I modify a pathway?

Modification of a pathway requires changes to the schematic (a .jpg format file) and/or a text file (.COORD extension) that maps each rectangle on the schematic to one or more LocusLink IDs. While neither task is particularly difficult, it can be time consuming and Genetrix does not provide any tools to assist in the process. However, a pathway drawing tool is in development and, when completed, will greatly facilitate changes to existing pathways.

### Can I add new pathways?

To add a new pathway, a new schematic (.jpg) and coordinate (.COORD) file must be created. In addition, the “formatted\_paths.txt” file must be edited to list the new pathway. Genetrix does not currently provide any tools to assist in this process. However, a pathway drawing tool is in development and, when completed, will greatly facilitate the creation of new pathways.

### What sort of information can be color-coded on gene symbols?

There are two options. The expression value for the gene, for a selected sample, can be used to define a color for the rectangle that represents a pathway gene. A button is available to scan through each sample in order. Alternatively, the value of a gene covariate can be used to determine the gene rectangle color. This covariate could, for example, be a p-value (derived from GeneScreen) that measured the significance of association between expression of the gene and some outcome of interest.

### Why are some rectangles divided into segments?

A single rectangle could be divided into two or more segments (each colored separately) if (1) several related genes are represented by one rectangle (e.g. “Integrins”), or (2) if there are multiple probe sets for the one gene (that is, multiple probe sets with the same LocusLink ID).

### Can I label genes in a pathway?

Before you can label a point, you must click on “Label” in the left margin to change the L-click

action from “Select” (i.e. show information on) to Label. Then L-click on each gene you want to label. A maximum of 100 genes can be labeled. L-click on a gene a second time to remove the label. Labels can be moved by clicking and dragging them to a new location.

R-clicking on the display provides a menu of options, including the options to label all genes, to label all highlighted genes (if appropriate), to hide the labels or to erase all labels.

The R-click menu also allows the user to change the content and appearance of a label. The content can be changed by selecting a “Notes” or a covariate (to use in place of the Label) or by clicking on “Change label” to redefine the Label (see Sample labels in Data Preprocessing). The appearance can be change in terms of color, background color or font size.

### **What are thumbnails, and how do I get them?**

Thumbnail are miniature graphics that display selected information about a gene. It is derived from the list of selected Key Comparisons; the “primary” Key Comparison is always used in a thumbnail label.

To **add a thumbnail** to a pathway, first make sure the “Label” option has been selected (in the left margin). R-click and select “Content & Appearance”, then select Key Comparison thumbnail. If no Key Comparisons have been defined, clicking Apply will bring up a dialog that allows you to define the Key Comparison; otherwise, the current “primary Key Comparison is selected for the thumbnail graphic. Select genes by clicking on them. Thumbnails can be **moved** by clicking and dragging them to a new location.

R-clicking on the pathway provides a menu of options, including the options to add the thumbnail to **all points** (maximum of 100), to add a thumbnail to **all highlighted points** (if appropriate), to **hide** the thumbnails or to **erase** all thumbnails.

**Clicking on a thumbnail** brings up the full Key Comparison display.

When the **cursor is placed within a Key Comparison**, information relating to its position within the thumbnail is displayed above the pathway. For example, if the thumbnail shows a time series, the expression value, time point and group (if any) corresponding to the position of the cursor within the thumbnail are shown.

Key Comparison thumbnails can include the **gene label**: click on “Add label as heading” in the “Label Content & Appearance” dialog.

When the Key Comparison thumbnail shows a mini-scatterplot of a selected gene against another (reference) gene, the choice of reference gene can be quickly changed by Shift-clicking on any gene.

### **What is a pathway Attribute?**

Membership in a pathway is treated as an attribute of a gene, analogous to having a specific gene ontology code. Thus, when genes are analyzed in terms of shared attributes, there are assessed according to whether they belong to a common pathway - or share a gene ontology code - with

greater frequency than would be expected by chance.

## **How can I map expression data to the genome?**

**What species are represented with a chromosomal ideogram?**