GeneChip® Expression Analysis

Data Analysis Fundamentals

AFFYMETRIX®

Table of Contents

Foreword

Affymetrix is dedicated to helping you design and analyze GeneChip® expression profiling experiments that generate high-quality, statistically sound, and biologically interesting results. This guide provides information, resources, and tools to help you easily design and analyze experiments and maximize the value derived from your GeneChip data.

There is a diverse range of experimental objectives and uses for GeneChip microarray data, which makes the areas of experimental design and data analysis quite broad in scope. As such, there are many ways to design expression profiling experiments, as well as many ways to analyze and mine data. This guide focuses on experimental design elements, statistical tests, and biological interpretation relevant to functional genomics expression profiling experiments, including transcriptional analysis of normal biological processes, discovery and validation of drug targets, and studies into the mechanism of action and toxicity of pharmaceutical compounds.

Chapter 1 Overview of Experimental Design Strategy

The best designed microarray experiments begin with well-defined goals, anticipated technical pitfalls, and minimized cost. This design phase is critical, as overlooking these key elements can result in highly variable or un-interpretable data.

The initial task is to define the objectives of the experiment. Each experimental design should optimize the chances of answering a key hypothesis. There is a natural temptation to test all of the interesting questions in a single experiment, but this approach is dangerous, as overly complex experiments may be un-testable, meaning that the data from these experiments are not statistically powerful enough to answer all questions. In practice this is the direct result of too few replicates or too little experimental control.

It is recommended that initial experiments focus on a thorough test of a single key hypothesis which will minimize the arrays required and simplify your data analysis. Testing of more complex hypotheses is best postponed for follow up studies. This serial approach minimizes cost, maximizes statistical power, and simplifies biological interpretation. For example, in a study of the toxic effect of a drug in mice, the critical variable is dose. It may seem desirable to maximize the number of doses, minimize the number of time points, and maintain a single controlled rodent diet. However, the temptation to test many time points or a new diet at the same time may undermine the ability to statistically test the dose response.

Ideally, one would want replication to be maximized. True statistical replication means that all test variables are changed independently, one at a time. To achieve this for each new variable added to a design, the required number of arrays is multiplied. For example, to replicate five doses, a minimum of three arrays is needed to replicate each dose, or a total of fifteen arrays. If two time points are tested to represent acute and chronic reactions, thirty arrays are needed to have the same statistical power. If diet is added, sixty arrays are needed. However, if dose and time are tested first, then the maximum effective non-toxic dose and the critical time point can be determined. Then a retesting of diet is done just at that one dose and one time point. If a control dose and a single dose is used at three replicates each, and two diets are tested, only twelve arrays are required. Totaling two serial experiments, forty-two arrays are used instead of sixty to query the dose, time, and diet parameters. Evidently, interactions between these variables are not tested by serial experiments, but in general, interactions are less important than main effects. Thus, using the information from an earlier study to refine a further test is a practical way to avoid costly and complex experiments that may be difficult to execute.

Pilot microarray studies are also recommended for practical reasons. If there are any unforeseen difficulties in the acquisition of biological sample, the assay, or the data analysis, a pilot study will often find them. Refining methods after a small scale study is far cheaper and more effective than complex mathematical fixes well after the fact. Pilot studies provide a safety net if there are problems and, if there are no issues, the few early answers can be incorporated into the complete experiment. Generally, pilot studies are limited designs that focus on a single variable versus a control state. Pilot studies also provide a good estimate of the variance of gene expression, which is useful in determining how many replicates the experiment's key questions will require.

If the researcher's experience with statistics is not extensive, then enlisting the help of a statistician or consulting a good textbook on statistics is strongly recommended either before the pilot study or just after. After the pilot study, a statistician may help calculate the statistical power of the experimental design, as well as determine the analytical approach and any software that may be required. Applying statistics in these planning stages can make the entire process easier and help avoid common pitfalls.

In the following sections, some of the sources of technical variability are detailed and suggestions are provided for minimizing it. In addition, statistical methods often used for microarray analysis are discussed. While other valid methods have been applied to GeneChip microarray data, suggestions herein use simple, commonly available, statistical methods that can be found in popular software packages, such as STATA® or SASS®.

Mitigating Technical and Biological Variance

Before speculating about sources of biological variability, other non-biological sources of variability must be identified and mitigated. In the GeneChip experimental process, the sources of variability in descending order are: biological, sample preparation (total RNA isolation as well as labeling), and system (instruments and arrays). Of these, the system noise is negligible and does not need to be addressed. As a result of the standardization of the hybridization, washing, staining, and scanning, as well as the quality controls built into manufacturing processes (14), system noise is not a significant source of technical variation. However, without careful technique and planning, sample preparation can be a large, unexpected, and unnecessary source of variation.

Obviously, all equipment used in the process should be calibrated regularly to ensure accuracy. Once the equipment calibration is validated, the next consideration in controlling variability in sample preparation is the isolation of total RNA. This is an important step when preparing microarray experiments and care should be used during experimental planning to ensure that the RNA is of high quality and consistently suitable for labeling and array hybridization. Standard protocols are given in the GeneChip® Expression Analysis Technical Manual (1), though modifications to this protocol may need to be made for some tissues that are difficult to collect or have high quantities of potential contaminants.

All RNA samples should meet assay quality standards to ensure the highest quality RNA is hybridized to the gene expression arrays. Researchers should run the initial total RNA on an agarose gel and examine the ribosomal RNA bands. Non-distinct ribosomal RNA bands indicate degradation which can lead to poor dsDNA synthesis and cRNA yield.

A 260/280 absorbance reading should be obtained for both total RNA and biotinylated cRNA. Acceptable $A_{260/280}$ ratios fall in the range of 1.8 to 2.1. Ratios below 1.8 indicate possible protein contamination. Ratios above 2.1 indicate presence of degraded RNA, truncated cRNA transcripts, and/or excess free nucleotides.

Sample Quality Assessment Following Total RNA Isolation		
Action	**Expected Results**	**Comments**
Electrophorese sample through an agarose gel.	Distinct ribosomal RNA bands.	Non-distinct ribosomal RNA bands indicate degradation, which will lead to poor dsDNA synthesis and labeling.
Measure 260/280 absorbance for total RNA and biotinylated cRNA.	Ratios between 1.8 and 2.1 in TE Buffer. Ratios between 1.6 and 1.9 in H_2O.	Ratios below 1.6 indicate possible protein contamination. Ratios above 2.1 indicate presence of degraded RNA, truncated cRNA transcripts, and/or excess free nucleotides.

The quality of total RNA can also be measured by the Bioanalyzer. A good quality sample should have 18S and 28S peaks that look like the image in Figure 1. The graph should have a low baseline and sharp ribosomal peaks. A degraded sample of RNA will look similar to the image in Figure 2. A good quality sample will typically have a ratio of 28S:18S ribosomal peaks of 2:1, however, this can be sample dependent.

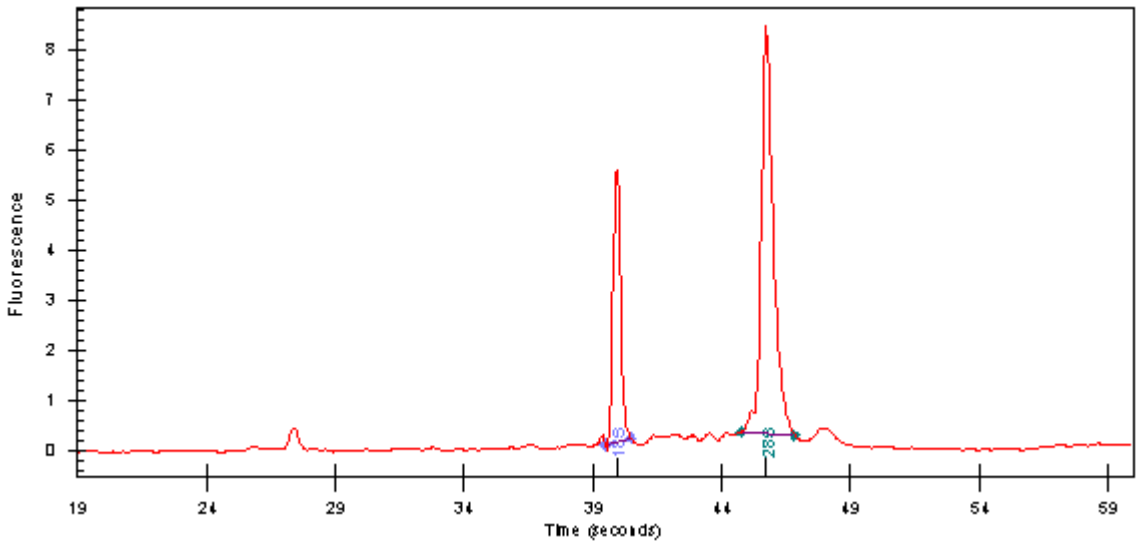


Figure 1. Good RNA sample quality measured with the Bioanalyzer.

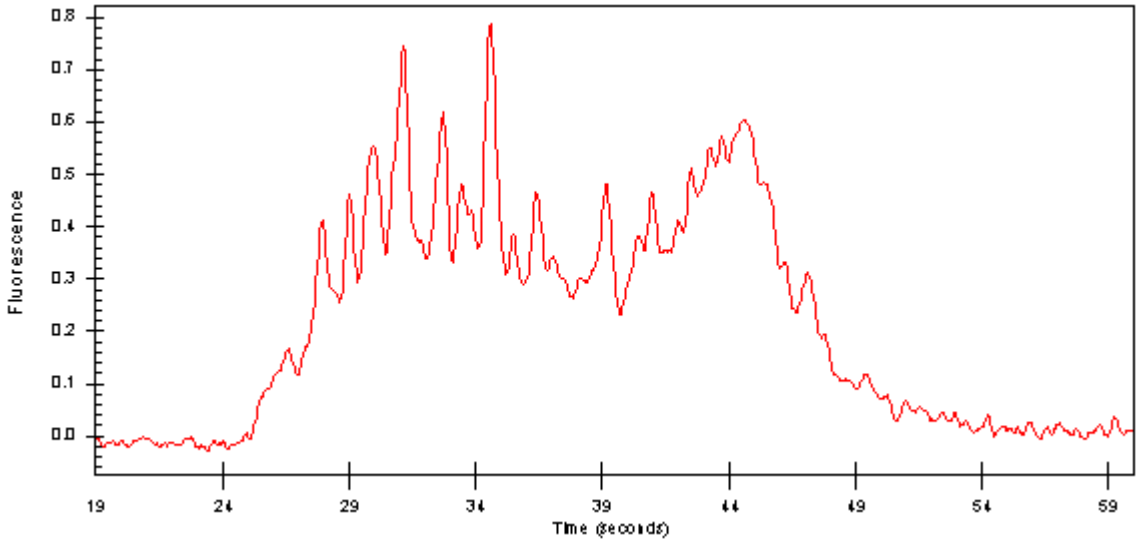


Figure 2. Degraded RNA sample quality measured with the Bioanalyzer.

Reduction in process variability is the next step to minimize variability in microarray results. A measurable source of potential variability in the labeling process is operator-to-operator differences. It is important to emphasize that care must be taken in the reverse transcription and labeling protocols to ensure consistency throughout the entire process. Common practices used to minimize such variability include processing all RNAs on the same day, using reagents from the same lots, preparing reagent master mixes, and having a single scientist responsible for all the bench manipulations. However, this is often not practical, as some experiments can be quite large and occur over an extended period of time. Thus, it is important to measure and then mitigate variations in the labeling process, both within and between the bench scientists. Such validation typically leads to the development of standard operating procedures followed by every scientist involved in a project such that each step in the process is clear and well defined.

This process validation can begin by following the sample from the isolation of the total RNA to the actual fragmentation of transcript following IVT. The use of gel electrophoresis will aid in following the sample from step to step in the assay and hybridization protocol. Gel electrophoresis can be performed after cDNA synthesis (if using poly-A mRNA as starting material), after cRNA synthesis, and after fragmentation. This will be helpful in estimating quantity and size distribution of the labeled sample. During this phase of technical evaluation, cRNA yield from a standard total RNA sample is another simple and effective method to assess consistency.

A sensitive method to assess the total process variability is to examine the correlation or concordance between data derived from standard total RNA samples, both within and between technicians. The most abundant and sensitive data point is the Signal derived from a GeneChip array. Two identical total RNA samples are labeled separately and hybridized to two arrays and the data are compared. The correlation coefficient (r) should be very high (>0.95), and the false change in Comparison Analysis between the two labeling reactions (I/D > 1 Signal Log Ratio) should be less than 1% or 2%, depending on the array used. In

addition, the change in detection from Present (P) to Absent (A) calls for the same genes should be approximately 10% or less overall. These metrics should also hold true when comparing data generated by the same or different technicians.

If the technician is not able to achieve these conditions, then potential sources of variability should be investigated. As stated above, calibration of all equipment should be done on a regular basis. In addition, seemingly subtle events may make a difference. For example, reaction times should be standardized within the recommended time window, pipetting techniques should be investigated to ensure consistency, etc. Following validation of the bench scientists' techniques for labeling and hybridizing samples, the remainder of the variability (equipment and array) should be a negligible portion of the total variability of the system.

Ultimately, variability in the final data is the important issue. Signal values are designed to be robust against noise. Before throwing out data or attempting a complex correction, check the effect on Signal. Strong biological effects can be reliably measured even in the presence of technical noise. If quality metrics and Signal data both indicate that a given array is unacceptable, the recommendation is to remove the data rather than attempt a mathematical repair. Most corrective measures are based on a theory of error, whether explicitly stated or implied. Unless the investigator is sure that the theory is a good model of reality, introducing the corrective measures may accidentally create a new class of false positives.

Every biological level of organization results in variation in gene expression so that, as a rule, biological variation will exceed technical variation in a well-controlled process (2). Unlike technical variation, these key variables are system dependent and may be more difficult to control, or may even be uncontrollable. Fortunately there are methods for handling variance, whether controllable or not.

Controlling as many variables as possible is the best option. For example, when working with a mouse model, the same gender is used, the same light/dark cycle is used, the animal is sacrificed at the same time and in the same way. What is not as obvious as these examples are seemingly innocuous changes in conditions that microarrays often detect. For example, Arabidopsis is so touch sensitive that simply spraying water on leaves triggers a suite of specialized genes (3). Heat shock, hypoxia, pH stress, and nutrient deficiencies are all examples of effects that induce gene transcription and can occur in whole animal, plant, and cell culture studies.

If similar variables are not normally controlled in an investigator's system when measuring large-scale phenotypes, the investigator may be unwittingly introducing problems of interpretation in the data set. Once again the pilot study is very helpful in this situation. Looking only at control arrays, ranking of genes by variance at every quartile of intensity can be done. That is, the lowest 25% of Signals is ranked, and then the next 25% is checked, and so on. By entering the top 100 Probe Set IDs into the Gene Ontology Analysis Tool in the NetAffx™ Analysis Center one can quickly see if a particular process is indicated. (Please refer to the "Biological Interpretation of GeneChip® Expression Data" section of this document for further explanation.) If this process is controllable, then the pilot study has been successful.

It is known that there are many factors that can not be controlled and some factors which are suspected to influence results. Controlling every possible biological factor is simply

impossible and, in studies such as human cancer, the sampling is unplanned. Fortunately, statistics offer workable solutions for many of these problems.

If a factor cannot be controlled, then randomize it. As long as replication is sufficient, random selection will dampen the factor's influence on the data. If level of control to that extent is not available, then the data can be stratified and each stratum weighted in the final analysis.

An appropriate example would be that of a liver cancer study with samples from several ethnic populations. Using the samples, the investigator intends to make a statement about the entire U.S. population. To do so, the samples can be split by ethnicity, and weighted by their frequency in the general population. As long as there is sufficient replication for each ethnicity, the weighted sample will be representative. For randomizing and representational weighting techniques, it is recommended to enlist the help of a statistician or a statistics textbook prior to beginning a full-scale experiment.

Determination of Arrays per Sample Type

One objective of a pilot study is to determine the optimal number of arrays to include in an experimental design. If this was a single measurement assay and a simple treatment vs. control test, then finding the critical number of arrays would be easy. Under the assumption of normality (please refer to the "Statistical Analysis" section of this document for further explanation) there is a standard formula based on the t-statistic. However microarrays are anything but simple, statistically speaking. Rather than a single measurement, there are thousands of measurements. Each of these thousands of genes has a different standard deviation (the normal or parametric method of estimating variance). This is why no simple answer exists to the question: how many arrays are needed for a study?

Finding the appropriate number depends on at least two considerations: the variability of the system being investigated and the minimum significant change proposed for measurement (the effect size). Significant change is the difference between means relative to the noise in the system, with the t-test being the most familiar example of a measure of effect size. However, since each gene on a microarray has a different variance no simple answer to the ideal array number is possible. Rather, each experimenter must select a threshold of significance. Ideally, data from a pilot study will help select the proper threshold, where known gene expression changes are declared significant. For planning purposes, the appropriate number of arrays is at least three and may go up to five arrays. However, the actual optimal number may vary depending on the study's samples and variance inherent in the experimental system.

For some systems, there is so much inherent variability that only a large number of arrays will allow production of statistically significant results. For example, human neural tissues often have a high level of sample-to-sample variability due to a high level of patient variability. This is the difficulty of sample collection in the operating room and the inherent sample preparation difficulties due to the high concentration of lipids. By contrast, data from cultured cells should have less variation due to the ability to tightly control the environmental conditions.

In addition, the optimal number of samples per condition may vary with the experimental condition. In an experiment investigating induced myocardial infarction, it was found that variation was increased in the experimental animals when compared with the control rats (4).

A simple method to determine the optimal number of samples per condition is to examine the coefficient of variation (CV) as a percentage of the mean value for each gene. This can be done on a continuum as depicted in the Affymetrix Technical Note on small sample preparation (5) or by sampling a number of data points from each quartile in the data (easily done in Microsoft Excel®). When this value is stabilized, that is, does not change from one biological replicate to the next, then it is unlikely that additional replicates will improve the accuracy of the samples' standard deviations, which are ultimately used to determine statistical significance in parametric statistical tests. In the example experiment summarized in Figure 3, there appears to be no significant improvement in CV between the third and fourth replicate array, which indicates a sufficient number of replicate experiments have been performed. Use of this strategy will result in n+1 number of optimal replicates, but this will serve to increase the degrees of freedom, thus increasing the power of statistical tests. If a fixed number of arrays is chosen before assessing the variance of the system, statistical estimates of variance may not be accurate due to insufficient sample size, which will decrease the accuracy of subsequent statistical tests.

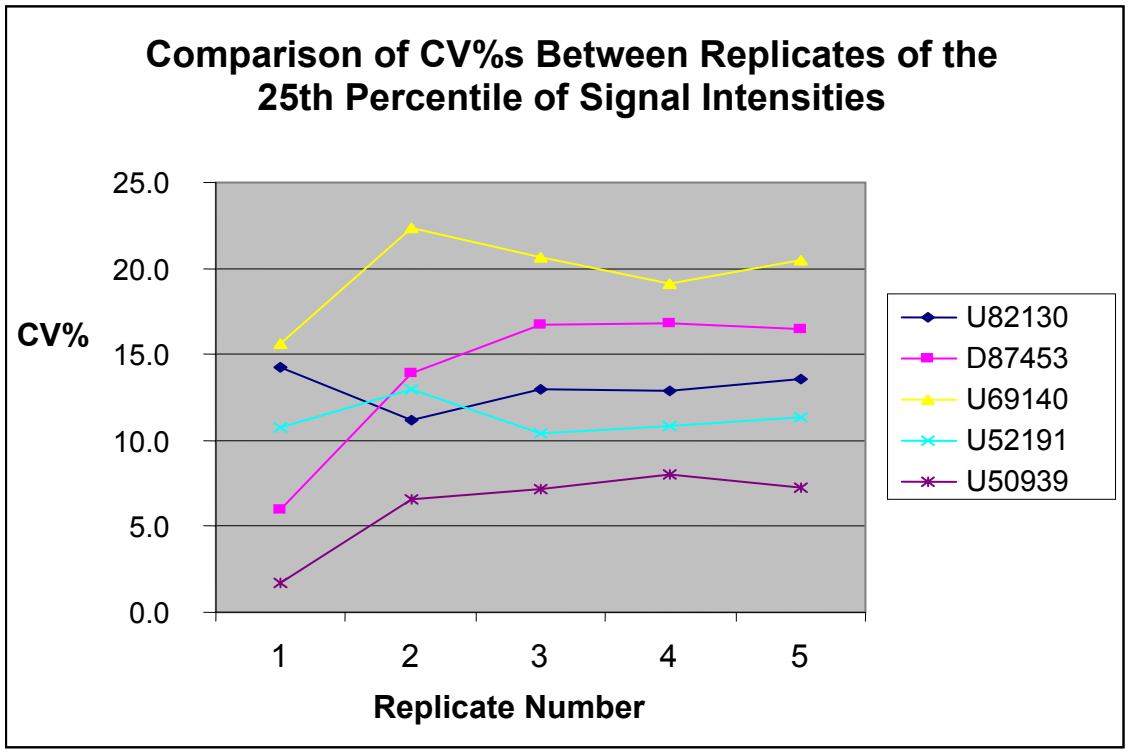


Figure 3. **Comparison of CV%s between replicates of the 25th percentile of Signal intensities.**

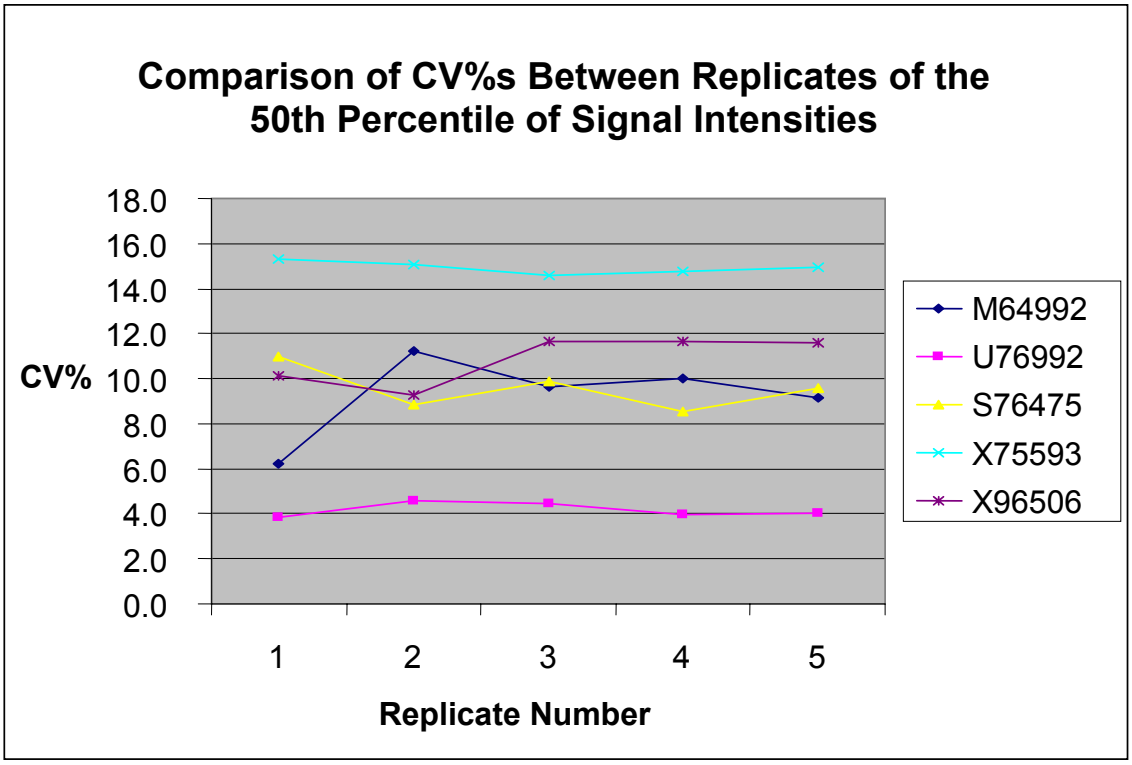


Figure 4. **Comparison of CV%s between replicates of the 50th percentile of Signal intensities.**

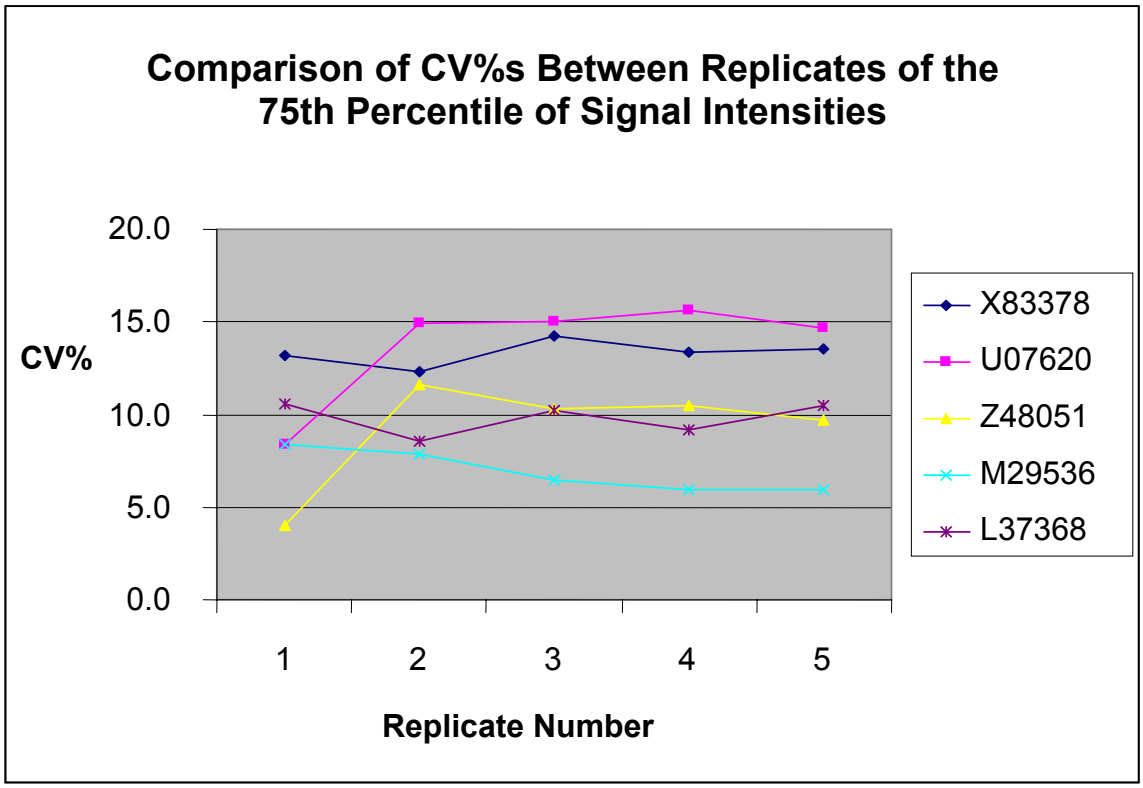


Figure 5. Comparison of CV%s between replicates of the 75th percentile of Signal intensities.

If you choose a fixed number of arrays instead of assessing the variance of the system, keep in mind that the statistical estimates of variance may not be accurate due to the insufficient sample size, which will thus decrease the accuracy of subsequent statistical tests.

Sample Pooling

At times, due to limited sample quantities, pooling of total RNA samples is necessary. It should be noted that protocols are available for labeling as little as 10 ng of sample (5). Pooling is also a common variance mitigation strategy, which, for reasons explained in this guide, is not recommended for microarrays. While pooling can be an effective way to overcome limitations in sample quantity and reduce variance, there are consequences that should be considered.

First, pooling results in irreversible loss of information. Once RNA samples are mixed there is no way to identify whether any one sample was a biological outlier. Microarrays are not extremely sensitive but measure a wide range of genes. Therefore, it is reasonable to expect that each sample may have an outlying measurement for at least a few genes. Those outlier genes may indicate a control variable which needs adjusting. Pooling averages across those outliers, so that information about system variability is lost. In addition, subsequent investigation of sample-specific attributes, such as the development of cancer in a defined age group or population, is irretrievable from a data set generated from pooled samples.

Pooling introduces a bias in microarray experiments. That is, physical mixing produces a Signal that is like the arithmetic mean of the samples. Gene expression data are best measured over several orders of magnitude, and the noise around a large Signal is greater than the noise around a small one. This means that outlier measurements are common. Since a simple average is sensitive to outliers, results from pooling will likewise be sensitive to outliers. This sensitivity is a bias because it is one sided. That is, high Signals add more noise than low Signals so the pooled signal will be biased high.

Finally, errors can be introduced when creating pools. If a researcher has clear cut class definitions, such as that seen in drug treatment studies, constructing several pools may be done safely. However, pooling is risky in studies where a classification scheme has yet to be elucidated or is often erroneous. The latter is a common problem with tumor identification. Keeping individual cancer samples separate allows a researcher to define new classifiers, which have been shown to be more subtle discriminators than histological methods as demonstrated in numerous microarray studies (6).

Even if classification is perfect, bias is minimal and no outliers exist; pools are not substitutes for replication. In the extreme example, where only a single pool exists for each treatment, pooling is especially problematic. The loss of variance measures ensures that genes are selected on the basis of changes in magnitude rather than the consistency or reliability of that change. In addition, the researcher misses those small magnitude changes that are reliable and may be biologically important.

Despite these significant concerns, pooling may be useful if applied carefully. If the amount of RNA from individual samples is very limited and pooling cannot be avoided, a researcher can benefit from statistical tests by using at least three pools for every condition being studied. For example, 30 mice are treated and three RNA pools derived from the samples from 10 mice each are created. In this way the individual idiosyncrasies of the mice are mitigated and replication is preserved. Thus, careful experimental design can alleviate some of the disadvantages of pooling. If a researcher can accept the irreversible loss of information, and the decrease in variance between pools is sufficient to separate two previously inseparable classes, then the experiment may warrant the risks of pooling.

Chapter 2 Types of Experimental Designs

As discussed previously, the first determination that must be made when creating an experimental design is how many biological replicates need to be run to produce meaningful data. The use of pilot experiments may help to determine the number of arrays potentially required for the study and may also assess whether or not biological variables are being controlled sufficiently. In addition, when planning a time-course experiment, ideal times for array hybridizations can be selected in a similar manner.

The simplest pilot experimental design is to test only one variable with a single treatment or condition against a control. Initially, the data can be collected for a small number of genes (five or more) using a quantitative PCR method. The genes selected should be those that are known to change or strongly suspected to change; and if possible, the anticipated range of expression levels is known. Thus the anticipated variance of the system can be assessed and optimal time points for the large-scale expression experiments using GeneChip® arrays can be chosen. This also provides the opportunity to refine the experimental design if necessary, prior to beginning pilot experiments with the arrays or a full-scale experiment.

Planning for data analysis is part of the experimental design. Ideally, microarrays should be treated as any other multiple endpoint analysis experiment: the biological hypothesis tested should be carefully noted as part of the prospective experimental design, the endpoints of the analysis should be specified with care; the power of the experimental design chosen should be prospectively identified, as should the analysis methods to be used.

Another important consideration is which statistical test will be used to analyze the data. The size and complexity of an experimental design will determine whether to use two-sample or multi-sample tests or whether parametric or non-parametric tests are most appropriate. Most experimental designs are essentially variations on the two types of experimental design: two-condition experimental design and multivariate experimental design. The two types of design are discussed in the following sections.

Two Condition Experimental Design

The simplest experimental design is a two-condition design, for example, normal and diseased tissue. A simple array comparison analysis can be done using the Affymetrix GeneChip® Operating Software (GCOS) software to obtain Increase/Decrease, Marginal Increase/Marginal Decrease, or No Change calls. While this data analysis approach is a good first pass, it does not take into account the variance which the experimental design is created to capture. A parametric or non-parametric (based on numerical or rank-ordered data, respectively) test of the two samples must be ultimately performed. Use of these tests makes the assumption that the minimal number of arrays per sample type has been determined as previously described. Increases in sample numbers beyond this will increase the degrees of freedom with a corresponding incremental increase in statistical power; that is, the accuracy of the variance may not change but the certainty of its accuracy is increased with increasing sample size.

An example of a standard two-design sample would be an experiment where the differences in gene expression between normal and diseased tissue are being studied. As a part of this design, pilot studies using quantitative PCR have been performed on a small selection of genes from the samples and the data used to estimate the ideal number of arrays to run for

each condition. Thus the experiments can be started and monitoring of CVs of intensities of the replicate samples can be done to determine the exact number of arrays required for each condition.

As a variation on this common, two-sample experimental design, there is also the possibility of designing a paired experiment that takes advantage of the greater statistical power found in paired-sample tests. In this case, samples are identical in all attributes except for the experimental treatment. In this type of design, the comparisons are made between the individual control and the corresponding experimental sample, and then the statistics are performed on the results of those individual comparisons within the group. While this design is extremely powerful, it is also very limited in practicality and, except in specific cases, the use of these statistics may be challenged. For example, a paired statistical analysis may be questioned when comparing data from biopsies in the same patient before and after treatment, because the patient's status (nutrition, health, age, hormonal cycles, etc.) may have changed between the first and second biopsy. However, removing a tumor from a mouse and comparing treated vs. control with *in vitro* experiments on separate sections of the same tumor may be acceptable for a paired statistical analysis.

An advantage of the GeneChip technology is the ability to add additional conditions to a study beyond a two-sample design. Using the above example of normal vs. diseased tissue, it is reasonable to assume that following discovery of some putative expression changes found by a two sample test, various experimental conditions will be added to the study to determine if these changes can be enhanced or reduced. If this is a possibility, then careful pre-planning for this contingency can save repetition of already existing data.

Multivariate Experimental Design

Multivariate experiments are powerful when the goal is to examine similarities or changes within groups. Using the example given above of normal and diseased tissues, two treatment groups are now added: A and B. The final data analysis will be comprised of four separate data groups, each with their distinct number of samples per group, as the variance can sometimes be greater in treated samples than control.

With this experimental design, there may also be opportunities to take advantage of higher-level analyses of variance, such as 2-way tests. Again, considering the above example, the need may arise to examine differences in male/female response to treatment. In this case, a sufficient number of each gender is assayed within each sample group to allow examination of differences between the various conditions, as well as find any differences between male and female responses.

Probe arrays may also be used to examine changes in gene expression over a given period of time, such as within the cell cycle. In the normal cell, the many genes involved in the cell cycle determine when and if the cell undergoes mitosis. Also built into this network are mechanisms designed to protect the body when this system fails due to mutations within one or more of the control genes, as is the case with cancerous cell growth. A GeneChip expression experiment could be designed where cell cycle data are generated in multiple arrays for each time point and then referenced to time zero and each subsequent time. This type of experimental design is fundamentally equivalent to a multi-sample design with each time point representing a discrete sample set. Multivariate analyses can be used to determine which genes are changing in relation to the other time points. However, this is a discrete

treatment of the data and other curve-fitting algorithms may be required for more sophisticated analysis.

When planning multiple-sample experiments, there are also the equivalent paired multivariate analyses, which would be subject to the same restrictions as the paired two-sample tests.

Chapter 3 Data Flow and Informatics Tools

Data storage and analysis tools are fundamental components of gene expression data generation. Affymetrix provides software that helps to facilitate the storage and flow of information throughout the experimental cycle. Figure 6 illustrates the variety of software available for each step from data storage and back-up through biological interpretation.

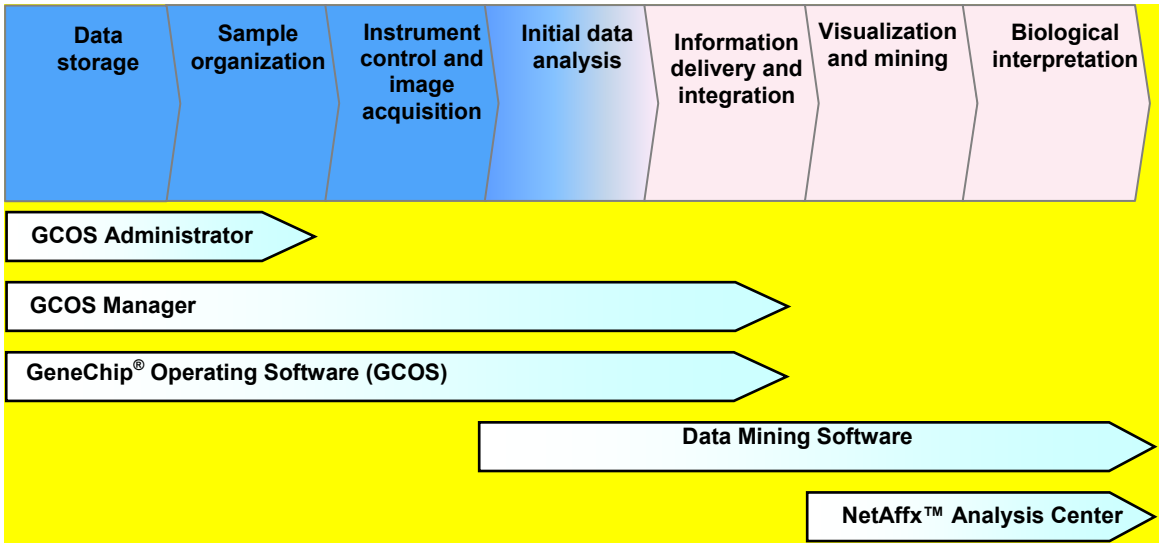


Figure 6. Software available for data storage, sample organization, instrument control, first order analysis, visualization of mining, and biological interpretation.

Software Tools

Software tools available for gene expression analysis are GCOS, GCOS Manager, GCOS Administrator, and GCOS Batch Importer.

GCOS

GCOS (GeneChip® Operating Software) provides an integrated software package for Expression data generation:

- Facilitates instrument control and data acquisition
- Provides a solution for workflow management and automation
- Performs first order data analysis

GCOS Manager

GCOS Manager provides tools to:

- Manage GeneChip microarray data in the Process and Publish databases
- Create and manage Publish databases

- Import data into the Process database
- Export experiment and analysis data
- Create usersets for analysis
- Define and manage templates for sample registration and experiment setup

GCOS Administrator

GCOS Administrator provides tools to:

- Backup (copy) a database, sample, or experiment to a compressed file format (.cab)
- Restore a database or data from a .cab file to a user-selected workstation drive or GCOS server
- Automatically backup the process database on the workstation
- Monitor available space on a workstation drive or database

GCOS Batch Importer

The GCOS Batch Importer:

- Facilitates import of Affymetrix data generated using Affymetrix® Microarray Suite (MAS) 5.X and GCOS 1.X from a different workstation

Data Hierarchy

Data are managed in GCOS by tying together a set of common experiments under a larger umbrella group called a Project. The Project is at the top of the hierarchy followed by samples, and then experiments. This hierarchy is established when an experiment is registered in GCOS.

An example of this hierarchy is illustrated in Figure 7. For the sake of simplicity, consider a cancer study with two patients: one patient with no disease, the other with cancer. Two tissue samples from each patient are taken: lung and liver. Naming of the study in GCOS would be as follows:

Project:
Cancer Study

Samples:
Normal Patient 1

Diseased Patient 2

Experiments:
Normal Lung Patient 1

Normal Liver Patient 1

Diseased Lung Patient 2

Diseased Liver Patient 2

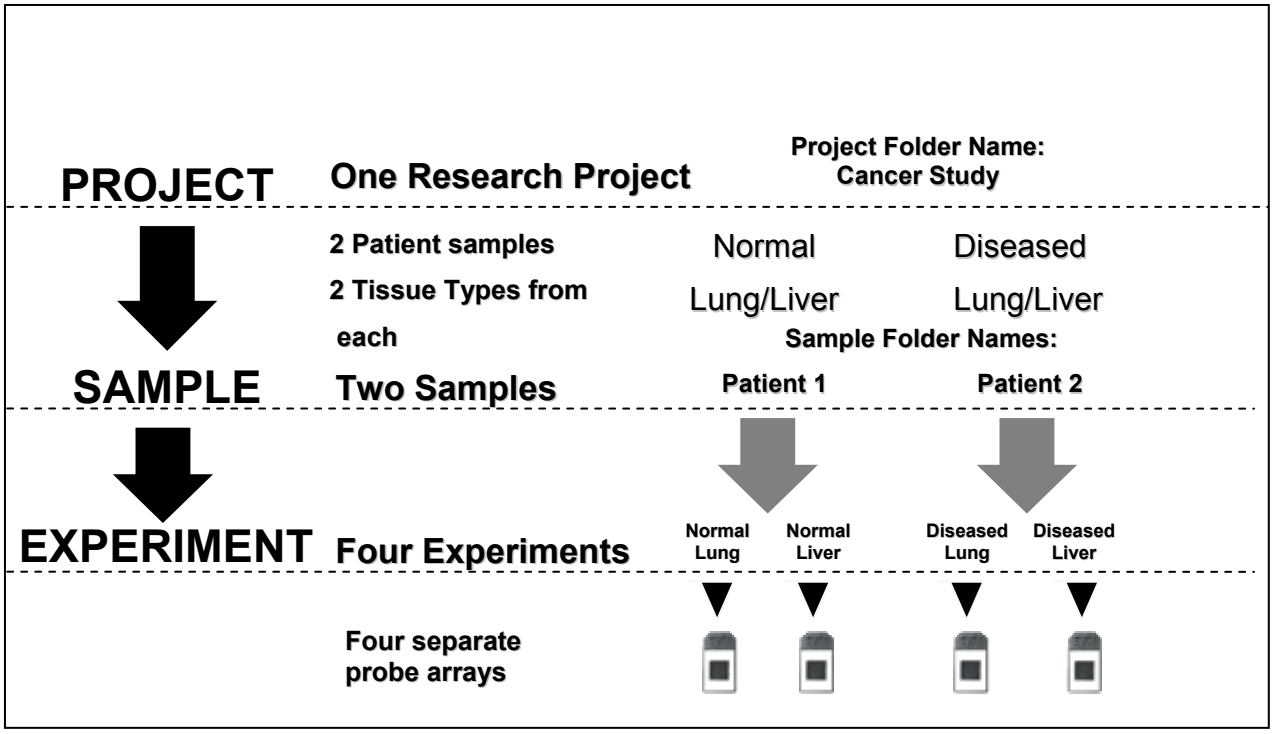


Figure 7. GCOS Naming Strategy

Registration and Data Files

A sample must be registered and an experiment defined in GCOS before processing a probe array in the fluidics station or scanning. This registration process associates a sample with a project and also allows for sample and experimental attributes to be added to the Process database. This registration process is the first component of data generation. Once the array is scanned, an image file is created called a .dat file. The software then computes cell intensity data (.cel file) from the image file. The cell intensity data is analyzed and saved as a .chp file. The .chp file contains data analysis information for each probe set on the array as well as controls. A report file (.rpt) is then created from the .chp. Figure 8 illustrates this process from registration through generating an expression report.

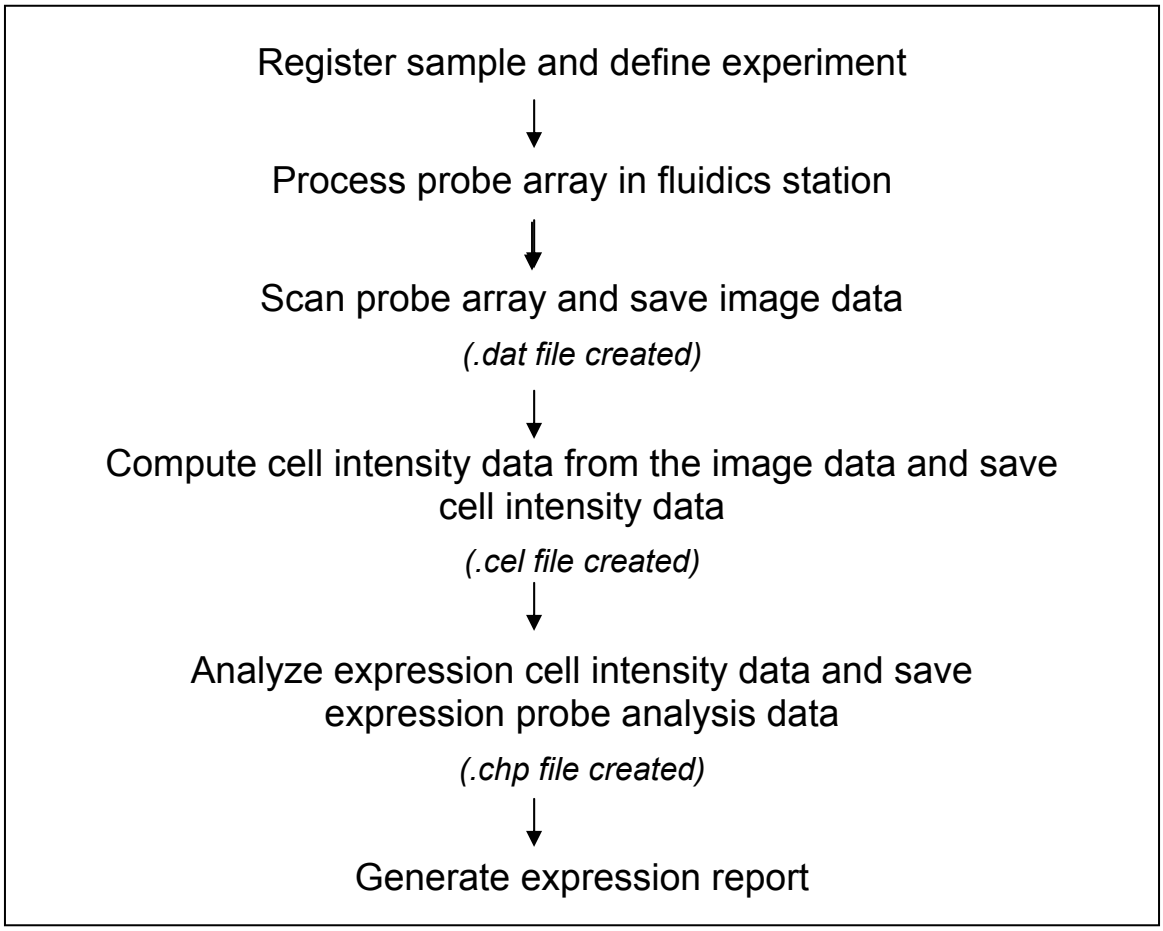


Figure 8. Experimental Data Flow Chart in GCOS

Data are organized in GCOS as follows:

Experiment Data File Name	File Extension	Description
Experiment Information File	N/A	Contains information about the experiment name, sample, and probe array type. The experiment name also provides the name for subsequent test data files generated during the analysis of the experiment.
Data File	*.dat	The image of the scanned probe array.
Cell Intensity File	*.cel	The software derives the *.cel file from a *.dat file and automatically creates it upon opening a *.dat file. It contains a single intensity value for each probe cell delineated by the grid (calculated by the Cell Analysis algorithm).
Chip File	*.chp	The output file generated from the analysis of a probe array. Contains qualitative and quantitative analysis for every probe set.
Report File	*.rpt	Text file summarizing data quality information for a single experiment. The report is generated from the analysis output file (*.chp).
Cab File	*.cab	A compressed file that is a backup copy of a process or publish database, project, sample, and/or experiment.
Data File	*.txt, *.xls	A standard format for text files. GCOS exports text in this file format. A standard format for Excel files.
Library Files	*.cif, *.cdf, *.psi	The probe information or library files contain information about the probe array design characteristics, probe utilization and content, and scanning and analysis parameters. These files are unique for each probe array type.
Fluidics Files	*.bin, *.mac	The fluidics files contain information about the washing, staining, and/or hybridization steps for a particular array format.

Chapter 4 First-Order Data Analysis and Data Quality Assessment

Single Array Analysis

This section describes a basic GeneChip® array analysis procedure that can be applied to many analysis situations. This procedure can be modified to account for specific experimental situations. It is highly recommended that, before attempting to modify this procedure, users familiarize themselves with the scaling strategies and settings involved in array analysis. More detailed information can be found in the GeneChip® Operating Software (GCOS) User Guide.

The following instructions assume that a probe array has been hybridized, washed, stained, and scanned according to the directions detailed in the Affymetrix GeneChip® Expression Analysis Technical Manual. Upon completion of the scan, the image file (.dat) is displayed in the GCOS software. After analysis of arrays, the procedures described later in this chapter can be used to assess the quality of the data generated.

These instructions relate to analyses performed in GCOS.

Data Storage

GCOS can be configured to store data on the local workstation's database or on a network accessible remote GCOS server. The default setting in GCOS is for the data to be stored in Local mode. In the GCOS user interface window, the heading of the Data Tree window pane will display: 'Data Source: Local.' In the Local mode, data are stored in the local MSDE database. See Figure 9 for an illustration of GCOS in local mode.

To register a server in GCOS, a remote GCOS server name can be entered during GCOS installation. After a server is registered, connecting to the server is performed as follows: From the menu bar, select Tools and then select Defaults. In the Defaults dialog box, select the Database tab. Choose the GCOS Server option. Experimental data will now be stored on the GCOS server. If the GCOS Server option is not selected, data are stored locally. Upon connecting to the server, the heading of the Data Tree window pane will display: 'Data Source: GCOS Server.'

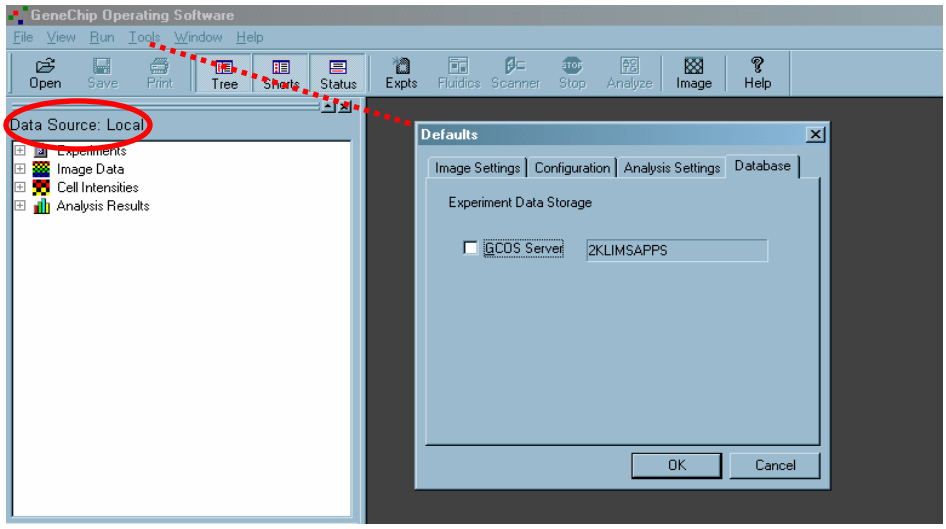


Figure 9. GCOS is set to Local Mode unless the option for the GCOS Server is selected.

Filtering Data

Filters may be applied to the experiment data. The filters determine the data that GCOS displays in the data tree, the sample history view, workflow monitor, and the instrument control dialog boxes. Filters are applied on a per user basis (identified by the logon name). Filters can be applied by selecting "Tools" and then "Filters" from the menu bar. After filters are applied, the status bar in the lower-right corner of the GeneChip Operating Software window indicates 'Filters applied.'

Quality Assessment of .dat Image

Prior to conducting array analysis, the quality of the array image (.dat file) should be assessed following the guidelines in this training manual.

Select a Scaling Strategy

These instructions use a global scaling strategy that sets the average Signal intensity of the array to a default Target Signal of 500. The key assumption of the global scaling strategy is that there are few changes in gene expression among the arrays being analyzed. This is a common strategy employed by many users, however, it should be noted that this strategy may not be appropriate for all experiments. Further discussion on scaling strategies and how to implement them can be found in Appendix E of the GeneChip Operating Software User Guide Version 1.1.

Expression Analysis Set-Up

A Single Array Analysis will create a .chp file from a .cel image file. GCOS automatically generates the .cel image file from the .dat file. To perform a Single Array Analysis, settings relating to file locations and the analysis must first be defined.

Specifying File-Related Settings

1 Select "Defaults" from the "Tools" pull-down menu.

2 Select the "Analysis Settings" tab.

2.1 Check "Prompt For Output File" to ensure display of output file name for confirmation or editing. With this option checked, GCOS will prompt for new file names for each analysis, preventing unintentional overwrite.

2.2 Check "Display Settings When Analyzing Data" to ensure display of expression settings for confirmation or editing.

Data files in GCOS by default are located in:
C:\GeneChip\Affy_Data\Data folder.

Library files in GCOS by default are located in:
C:\GeneChip\Affy_Data\Library folder.

Fluidics Protocols in GCOS by default are located in:
C:\GeneChip\Affy_Data\Protocols folder.

3 Select the "Database" tab. The tab specifies how GCOS manages experiment data, including image, cell intensity, and probe analysis data. Choose the GCOS Server option to store the experiment data on the remote GCOS server. If this option is not chosen, the experiment data are stored on the MSDE database on the workstation.

Note: The Experiment Data Storage option is only available if all windows are closed and no instruments are active.

4 Select "OK."

Expression Analysis Settings

1 Select "Expression Analysis Settings" from the "Tools" pull-down menu. The "Expression Analysis Settings" dialogue box opens.

2 Select the "Probe Array Type" to be analyzed from the drop-down menu.

3 Select the "Scaling" tab.

3.1 Select "All Probe Sets" and set "Target Signal" to 500 or to desired Target Signal.

4 Select the "Normalization" tab.

4.1 Select "User Defined" and place a "1" in the "Normalization Value" box. This ensures that no normalization procedure is applied to the data. Normalization is not necessary as the data are being scaled. Further information can be found in Appendix E of the GeneChip Operating Software User Guide Version 1.1.

5 Select the "Probe Mask" tab. This feature is used to mask user-defined probe cells.

5.1 Ensure that the "Use Probe Mask File" option is not selected.

6 Select the "Baseline" tab. For single array analysis no baseline file should be used.

6.1 Ensure "Use Baseline File Comparison" is not selected.

7 Select the "Parameters" tab.

7.1 Confirm default settings appropriate to the version of GCOS and the array being analyzed as specified in Appendix C of this training manual.

Note: These Settings should not be adjusted unless the user has advanced experience with the Affymetrix GeneChip® system.

8 Once all settings have been adjusted or confirmed select "OK" to define settings and close the dialogue box.

Performing Single Array Analysis

1. Open the file you wish to analyze (.dat or .cel) by double clicking on the file name in the data file tree. Alternatively, select "Open" from the "File" pull-down menu and select the image file you wish to analyze.

1.1. After the .dat or .cel file image is displayed, the "Analyze" button on the menu bar is activated. Click the "Analyze" button. Verify the .chp file name. The default corresponds to the name of the .dat/.cel file names. Edit the .chp file name, if necessary, and click "OK."

1.2. The alternative is to select "Analysis" from the "Run" pull-down menu.

Note: GCOS will automatically overwrite a .chp file if the filename is the same as an existing .chp file in the directory.

1.3. Verify "Expression Analysis Settings" in the subsequent pop-up window as previously set in the above Expression Analysis Settings section and select "OK" to begin analysis and generate the analysis results file (.chp).

1.4. The GCOS status window will indicate that analysis has started.

2. Once analysis is complete, generate an Expression Analysis report file (.rpt) and review the quality control metrics.

2.1. To generate the report, select "Report" from the "File" pull-down menu.

2.2. Select the appropriate analysis results file (.chp).

NOTE: Alternatively, you can highlight the appropriate .chp file in the data file tree, right click on the mouse and select "Report."

2.3. Review the quality control data as discussed in the "Guidelines for Assessing Data Quality" section.

2.3.1. Review *bioB, bioC, bioD*, and *cre* sensitivity spikes.

2.3.2. Review Percent Present determination.

2.3.3. Review housekeeping control signal output and 3'/5' ratios.

2.3.4. Review noise (Raw Q).

2.3.5. Review average background.

2.4. Return to the .chp file by closing the Report (.rpt) file, or by selecting "Window" from the Menu toolbar and select the .chp file.

NOTE: The open .chp file data are displayed in the Expression Analysis Window (EAW) and can be accessed by clicking on the Expression Analysis button in the GCOS shortcuts window.

3. Select the "Pivot" tab at the bottom of the analysis results .chp file. The Pivot table displays analysis output and descriptions for each transcript represented on the probe array. The far-left column contains the Affymetrix unique probe set identifier and the column to the far-right contains a brief description of the sequence that the probe set represents.

3.1. Display additional Pivot table columns in the analysis by selecting "Pivot Data>Absolute Results" from the "View" pull-down menu. Select the columns desired to be displayed. Columns may include "Signal," "Detection," "Detection *p*-value," "Stat Pairs," and "Stat Pairs Used."

NOTE: Values in the "Signal" column reflect intensity. The "Detection" column assigns a call of "Present," "Absent," or "Marginal" to each probe set and the "Detection *p*-value" column provides an assessment of statistical significance of each call. The "Descriptions" column provides summary information about each transcript. Right click on a transcript of interest to link to an external database for more information.

3.2. Select the "Metrics" tab at the bottom of the .chp file.

3.3. The Metrics table displays data for each distinct probe set in the .chp file. The columns displayed are similar to the Pivot table.

3.3.1. Organize the tabular data columns by right clicking at the top of the column to "Hide Column."

3.3.2. Sort by right clicking on the column header and selecting the desired sorting function.

NOTE: Refer to the section titled "Interpretation of Metrics" for recommendations on data interpretation.

3.4. Select the "Analysis Info" tab at the bottom of the analysis results or .chp file. The Analysis Information table displays experimental and sample information and algorithm settings information. This information includes Scaling or Normalization factors, Background, Raw Q, and Sample Type information.

Once a single array analysis has been completed and a .chp file generated, this file can be further utilized in a number of ways. The file can be used as a "baseline" file in a comparison analysis. The .chp file can also be published into a publish database, becoming accessible for advanced data mining software. The .chp file data can also be exported from GCOS as a text file allowing the data to be imported into third-party programs (e.g., Microsoft® Excel).

Comparison Analysis

Comparison analysis is used to compare expression profiles from two GeneChip® probe arrays of the same type. One array is designated as a baseline and the other is designated as experimental. The experimental file is analyzed in comparison to the baseline file. While the designations "experimental" and "baseline" are arbitrary, it is important to keep these designations in mind when examining the changes reported. For example, if the baseline file is derived from a treated sample and the experimental from an untreated sample, all genes activated by the treatment will have decrease calls.

Quality Assessment of .dat Image

Prior to conducting analysis of an array, the quality of the array image (.dat file) should be assessed following the guidelines found in the section "Guidelines for Assessing Data Quality."

NOTE: Single-array (or 'absolute') analyses must be previously completed and .chp files present for all samples that will be used as baseline files.

When conducting a Comparison Analysis it is important to ensure that the scaling strategy used for the Comparison Analysis is the same as that used to generate the baseline file. To examine the analysis settings of the baseline file, right click the baseline .chp file in the Data File Tree and select "Information." The following fields are of note:

TGT	Target Signal value used in both arrays should be the same. The default value is 500.
SF	Displays the scaling factor calculated. In this protocol this should NOT be 1.0000.
NF	Displays the normalization factor applied. In this protocol the value should be 1.0000, as no normalization was used.
SF Gene	Displays the Scaling strategy used. In this protocol the value should be 'All,' as the global scaling strategy was used.

Comparison Analysis Set-Up

Like the Single Array Analysis, Comparison Analysis will create a .chp file from a .cel image file. GCOS automatically generates the .cel image file from the .dat file. To perform a Comparison Analysis, settings relating to analysis must first be defined.

Expression Analysis Set-Up

1 Close any .chp files that are currently open and Select "Expression Settings" from the "Tools" pull-down menu. The "Expression Analysis Settings" dialogue box opens.

2 Select the "Probe Array Type" to be analyzed from the drop-down menu.

2.1 Select the "Scaling" tab.

3 Select "All Probe Sets" and set the appropriate "Target Signal."

3.1 Select the "Normalization" tab.

4 Select "User Defined" and place a "1" in the "Normalization Value" box.

5 Select the "Probe Mask" tab. This feature is used to mask user-defined probe cells.

5.1 Ensure that the "Use Probe Mask File" option is unchecked.

6 Select the "Baseline" tab.

6.1 Check the "Use Baseline File Comparison" option.

6.2 Click the "Browse" button.

6.3 Select the baseline .chp file.

6.4 Click the "OK" button.

7 Select the "Parameters" tab.

7.1 Confirm default settings appropriate to the version of GCOS and array being analyzed as specified in Appendix C of this training manual.

NOTE: These Settings should not be adjusted unless the user has advanced experience with the Affymetrix GeneChip® system.

8 Once all settings have been adjusted or confirmed select "OK" to define settings and close the dialogue box. One can now perform comparison analyses based upon these settings.

Performing Comparison Analysis

1 Open the designated experimental file (.dat or .cel) by double clicking in the data file tree. Alternatively, select "Open" from the "File" pull-down menu and select the experimental file.

2 Select "Analysis" from the "Run" pull-down menu. Alternatively, click the Analyze button.

2.1 Verify the .chp filename. The default corresponds to the name of the experimental .exp/.dat/.cel file names. Edit the .chp filename, if necessary, and click "OK."

NOTE: GCOS will overwrite a .chp file if the filename is the same as an existing .chp file in the directory.

2.2 Verify "Expression Analysis Settings" in the subsequent pop-up window as previously set in the above Expression Analysis Settings section and select "OK" to begin analysis and generate the .chp file.

2.3 The GCOS status window will indicate that analysis has started.

3 Once analysis is complete, generate an Expression Analysis report file (.rpt) and review the quality control metrics as described.

3.1 To generate the report, select "Report" from the "File" pull-down menu.

3.2 Select the appropriate analysis results file (.chp).

NOTE: All metrics reported in a comparison file report refer to the designated experimental file, NOT the baseline file.

3.3 Review the quality control data.

3.3.1 Review *bioB, bioC, bioD,* and *cre* sensitivity spikes.

3.3.2 Review Percent Present determination.

3.3.3 Review housekeeping control signal output 3'/5' ratios.

3.3.4 Review noise (Raw Q).

3.3.5 Review average background.

3.4 Return to the .chp file by closing the Report (.rpt) file.

NOTE: The open .chp file data is displayed in the Expression Analysis Window (EAW) and can be accessed by clicking on the Expression Analysis button in the GCOS shortcuts window.

4 Select the "Pivot" tab at the bottom of the .chp file. The Pivot table displays analysis output and descriptions for each transcript represented on the probe array. The far-left column contains the Affymetrix unique probe set identifier and the column on the far-right provides a brief description of the sequence that the probe set represents.
Display additional Pivot table columns in the analysis by selecting "Pivot Data>Comparison Results" from the "View" pull-down menu. Select the columns desired to be displayed. Suggested columns may include "Signal Log Ratio," "Change," and "Change *p*-value."
Alternatively, clicking the "Options" button in the shortcut menu and selecting the Pivot tab in the Analysis Options window will also enable column selection.
Select the "Metrics" tab at the bottom of the .chp file. The Metrics table displays data for each distinct probe set in the .chp file. Columns displayed are similar to the Pivot table.
In the Pivot table, sort data by right clicking the mouse on the column header and selecting the desired sorting function. These useful functions enable you to sort the data in ascending or descending order and to hide or unhide columns. For example, if you are interested in only those genes which are "Increasing" and have increased at a "Signal Log Ratio" of > 1, the following steps are performed:

4.1 Point the mouse cursor to the Change column header and right-click. Choose Sort Ascending. Press OK. Probe sets will be sorted in the following Change order: D, I, MD, MI, NC.

4.2 To display those probe sets with a Change call of "I," all probe sets with a Change call other than "I" need to be hidden.

4.2.1 Make sure that the scroll bar is at the top of the Pivot table page. Scroll down to the first probe set in the table with the Change call "I." Point the mouse cursor to the left column containing the probe set ID. Click the mouse to highlight the entire row.

4.2.2 Press the Shift key once and scroll down to the last probe set in the table with the Change call of "I." Point the mouse cursor to the probe set ID in the left-hand column, press the Shift key, and click the mouse. All rows, between and including those of the first and last probe set chosen, will be highlighted.

4.2.3 Select the Hide All Unselected Probe Sets button in the Shortcut menu bar. (Note: There are two Hide buttons on the shortcut menu bar. Make sure the correct one is chosen.) The unselected probe sets will be hidden. The probe sets not hidden will have the Change call of "I."

4.3 With only probe sets having the Change call of "I" displayed, now sort the Signal Log Ratio in ascending order. Point the mouse cursor on the Signal Log Ratio column header, right-click and choose Sort Ascending Order.

4.4 Then choose probe sets with Signal Log Ratio > 1 using similar operational steps as outlined in Step b) above.

NOTE: Refer to Chapter 5 for recommendations.

After the comparison analysis .chp file has been generated, this file can be further utilized in a number of ways. The .chp file can also be published into the Publish databases in MSDE (local or client mode) or GCOS Server, becoming accessible for data mining with the data mining software. The .chp file data can also be exported from GCOS as a text file allowing the data to be imported into third-party programs (e.g., Microsoft Excel).

Using the Batch Analysis Tool

Batch analysis is a way to analyze many .cel files and generate .chp files with unattended operation. Many files can be simultaneously compared to a selected baseline. Files from different experiments may also be simultaneously analyzed. It is important to select a different name for the analysis output (.chp file) otherwise batch analysis will overwrite the previous files. Either the Drag and Drop method or the Toolbar can be used to select files for batch analysis. Further details can be found in Chapter 11 of the Affymetrix GCOS User Guide Version 1.1.

NOTE: Prior to batch analysis, check the Expression Analysis settings and ensure that they are correct (i.e., select the "Baseline" tab and ensure "Use Baseline File Comparison" is unchecked).

1. Open the Batch Analysis window by selecting "Batch Analysis" from the "Run" menu. Alternatively, click on the Batch Analysis icon in the GeneChip Software section of the Shortcut bar.
2. Add files to the Batch Analysis window by:
   2.1. Dragging and Dropping each .cel or .chp file to the Batch Analysis window from the data file tree to the Batch Analysis window.

   OR

   2.2. Using the Toolbar, click the "Add" Toolbar or select "Edit>Add."

2.3. An open dialog of .cel files appears.

2.4. Select the .cel or .chp files to be analyzed.

2.5. To select all files, hold "shift" while you click on the first and last file.

2.6. To select files individually, hold "control" while selecting files.

2.7. Click open to place the files into the Batch Analysis window.

3. Verify the Output filenames.

3.1. The filename for the .chp file is listed in the Output column. If the .chp filename is already present the filename will be in red to indicate that a file is going to be overwritten.

3.2. To edit the .chp file name, double click on the output file name and type in a new name.

4. To select the baseline file, double click in the Baseline column corresponding to the .cel file being analyzed or click the .cel file and choose "Select Baseline" from the "Edit" pull-down menu.

4.1. Double click on the baseline .chp file from the dialog box.

4.2. Right clicking the baseline file and selecting "Clear Baseline" or selecting "Edit>Clear Baseline" can remove a baseline file in the batch analysis window.

5. To start the Batch Analysis, click on the Analyze button which is found immediately above the Batch Analysis window.

Guidelines for Assessing Data Quality

The purpose of this section is to help researchers establish quality control processes for gene expression analyses. To achieve this, Affymetrix has developed several controls which allow researchers to monitor assay data quality.

The following are a series of quality control parameters associated with assay and hybridization performance. Affymetrix highly encourages new users to create a running log of these parameters in order to monitor quality and potentially flag outlier samples. Evaluation of a particular sample should be based on the examination of all sample and array performance metrics.

Probe Array Image (.dat) Inspection

Inspect for the presence of image artifacts (i.e., high/low intensity spots, scratches, high regional, or overall background, etc.) on the array. Please contact your Field Applications Specialist (FAS) or 888-DNA-CHIP for further advice on image artifacts.

B2 Oligo Performance

The boundaries of the probe area (viewed upon opening the .dat/.cel file) are easily identified by the hybridization of the B2 oligo, which is spiked into each hybridization cocktail. Hybridization of B2 is highlighted on the image by the following:

1. The alternating pattern of intensities on the border

2. The checkerboard pattern at each corner (Refer to Figure 10)
3. The array name, located in the upper-left or upper-middle of the array (Refer to Figure 11)

B2 Oligo serves as a positive hybridization control and is used by the software to place a grid over the image. Variation in B2 hybridization intensities across the array is normal and does not indicate variation in hybridization efficiency. If the B2 intensities at the checkerboard corners are either too low or high, or are skewed due to image artifacts, the grid will not align automatically. The user must align the grid manually using the mouse to click and drag each grid corner to its appropriate checkerboard corner.

The B2 oligonucleotide is available as part of the GeneChip® Eukaryotic Hybridization Control Kit (P/N 900299 and 900362), and can also be ordered separately (P/N 900301).

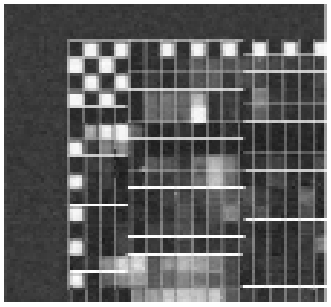


Figure 10. An example of B2 illuminating the corner and edges of the array.

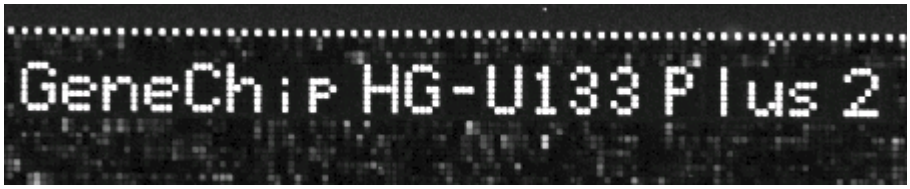


Figure 11. The array name.

After scanning the probe array, the resulting image data are stored on the hard drive of the GeneChip® Operating Software workstation, or on the process database, as a .dat file with the name of the scanned experiment. In the first step of the analysis, a grid is automatically placed over the .dat file demarcating each probe cell. One of the probe array library files, the .cif file, is used by the GeneChip® Operating Software to determine the appropriate grid size to use. Confirm the alignment of the grid by zooming in on each of the four corners and on the center of the image.

If the grid is grossly misaligned (i.e., more than two pixels off), adjust the alignment by placing the cursor on an outside edge or corner of the grid. The cursor image will change to a small double-headed arrow. The grid can then be adjusted using the arrow keys on the keyboard or by clicking and dragging the borders with the mouse.

Average Background and Noise Values

The Average Background and Noise (Raw Q) values can be found either in the Analysis Info tab of the Data Analysis (.chp) file, or in the Expression Report (.rpt) file. Although there are no official guidelines regarding background, Affymetrix has found that typical Average Background values range from 20 to 100 for arrays scanned with the GeneChip® Scanner 3000. Arrays being compared should ideally have comparable background values.

Noise (Raw Q) is a measure of the pixel-to-pixel variation of probe cells on a GeneChip array. The two main factors that contribute to noise are:

1. Electrical noise of the scanner.
2. Sample quality.

Each scanner has a unique inherent electrical noise associated with its operation. Since a significant portion of Noise (Raw Q) is electrical noise, values among scanners will vary. Array data (especially those of replicates) acquired from the same scanner should ideally have comparable Noise values.

Poly-A Controls: *lys, phe, thr, dap*

Poly-A RNA controls can be used to monitor the entire target labeling process. *Dap, lys, phe, thr*, and *trp* are *B. subtilis* genes that have been modified by the addition of poly-A tails, and then cloned into pBluescript vectors, which contain T3 promoter sequences. Amplifying these poly-A controls with T3 RNA polymerase will yield sense RNAs, which can be spiked into a complex RNA sample, carried through the sample preparation process, and evaluated like internal control genes. The GeneChip® Poly-A RNA Control Kit (P/N 900433) contains a pre-synthesized mixture of *lys, phe, thr,* and *dap*. The final concentrations of the controls, relative to the total RNA population, are: 1:100,000; 1:50:000; 1:25,000; 1:7,500, respectively. All of the Poly-A controls should be called "Present" with increasing Signal values in the order of *lys, phe, thr, dap*.

Hybridization Controls: *bioB, bioC, bioD,* and *cre*

BioB, bioC and *bioD* represent genes in the biotin synthesis pathway of *E. coli*. *Cre* is the recombinase gene from P1 bacteriophage. The GeneChip® Eukaryotic Hybridization Control Kit (P/N 900299 and 900362) contains 20x Eukaryotic Hybridization Controls that are composed of a mixture of biotin-labeled cRNA transcripts of *bioB, bioC, bioD*, and *cre,* prepared in staggered concentrations (1.5 pM, 5 pM, 25 pM, and 100 pM final concentrations for *bioB, bioC, bioD*, and *cre*, respectively).

The 20x Eukaryotic Hybridization Controls are spiked into the hybridization cocktail, independent of RNA sample preparation, and are thus used to evaluate sample hybridization efficiency on eukaryotic gene expression arrays. *BioB* is at the level of assay sensitivity (1:100,000 complexity ratio) and should be called "Present" at least 50% of the time. *BioC, bioD*, and *cre* should always be called "Present" with increasing Signal values, reflecting their relative concentrations.

The 20x Eukaryotic Hybridization Controls can be used to indirectly assess RNA sample quality among replicates. When global scaling is performed, the overall intensity for each array is determined and is compared to a Target Intensity value in order to calculate the appropriate scaling factor. The overall intensity for a degraded RNA sample, or a sample that

has not been properly amplified and labeled, will have a lower overall intensity when compared to a normal replicate sample. Thus, when the two arrays are globally scaled to the same Target Intensity, the scaling factor for the "bad" sample will be much higher than the "good" sample. However, since the 20x Eukaryotic Hybridization Controls are added to each replicate sample equally (and are independent of RNA sample quality), the intensities of the *bioB, bioC, bioD*, and *cre* probe sets will be approximately equal. As a result, the Signal values (adjusted by scaling factor) for these control probe sets on the "bad" array will be adjusted higher relative to the Signal values for the control probe sets on the "good" array.

Internal Control Genes

For the majority of GeneChip® expression arrays, β-actin and GAPDH are used to assess RNA sample and assay quality. Specifically, the Signal values of the 3' probe sets for actin and GAPDH are compared to the Signal values of the corresponding 5' probe sets. The ratio of the 3' probe set to the 5' probe set is generally no more than 3 for the 1-cycle assay. Since the Affymetrix eukaryotic expression assay has an inherent 3' bias (i.e., antisense cRNA is transcribed from the sense strand of the synthesized ds cDNA, via the incorporated T7 promoter), a high 3' to 5' ratio may indicate degraded RNA or inefficient transcription of ds cDNA or biotinylated cRNA. 3' to 5' ratios for internal controls are displayed in the Expression Report (.rpt) file. The 2-cycle assay typically gives higher 3' to 5' ratios than the 1-cycle assay, due to the additional cycle of amplification.

There are occasions when the 3' to 5' ratio of one internal control gene is normal, but the 3' to 5' ratio of another internal control gene is high. This discrepancy in 3' to 5' ratios is most likely due to a specific transcript-related or image artifact issue and is not an indication of overall sample and assay quality.

Percent Present

The number of probe sets called "Present" relative to the total number of probe sets on the array is displayed as a percentage in the Expression Report (.rpt) file. Percent Present (%P) values depend on multiple factors including cell/tissue type, biological or environmental stimuli, probe array type, and overall quality of RNA. Replicate samples should have similar %P values. Extremely low %P values are a possible indication of poor sample quality. However, the use of this metric must be evaluated carefully and in conjunction with the other sample and assay quality metrics described in this chapter.

Scaling and Normalization Factors

Details regarding Scaling and Normalization are listed in the GeneChip® Operating Software User Guide Version 1.1, Appendix E. Scaling and normalization factors can be found in the Analysis Info tab of the .chp file output and in the Expression Report (.rpt) file.

For the majority of experiments where a relatively small subset of transcripts is changing, the global method of scaling/normalization is recommended. In this case, since the majority of transcripts are not changing among samples, the overall intensities of the arrays should be similar. Differences in overall intensity are most likely due to assay variables including pipetting error, hybridization, washing, and staining efficiencies, which are all independent of relative transcript concentration. Applying the global method corrects for these variables. For global scaling, it is imperative that the same Target Intensity value is applied to all arrays being compared.

For some experiments, where a relatively large subset of transcripts is affected, the "Selected Probe Sets" method of scaling/normalization is recommended. The global approach does not make sense in this situation since the overall intensities among arrays are no longer comparable. Differences in overall intensity are due to biological and/or environmental conditions. Applying the global method skews the relative transcript concentrations. One option is to apply the "Selected Probe Sets" method using the 100 Normalization Control probe sets, which are available for the major catalog arrays.

For replicates and comparisons involving a relatively small number of changes, the scaling/normalization factors (calculated by the global method) should be comparable among arrays. Larger discrepancies among scaling/normalization factors (e.g., three-fold or greater) may indicate significant assay variability or sample degradation leading to noisier data.

Scaling/normalization factors calculated by the "Selected Probe Sets" method should also be equivalent for arrays being compared. Larger discrepancies between scaling/normalization factors may indicate significant assay or biological variability or degradation of the transcripts used for scaling/normalization, which leads to noisier data.

Chapter 5 Statistical Algorithms Reference

This chapter is a reference for the Affymetrix Statistical Algorithms used in the expression analysis of GeneChip® probe arrays. It provides the user with a basic description of the mathematical concepts behind expression measurements for either single array or comparison analysis.

The Statistical Algorithms were implemented in Affymetrix® Microarray Suite Version 5.0. Previous versions of the GeneChip® Analysis Suite and Affymetrix Microarray Suite used the Empirical Algorithms.

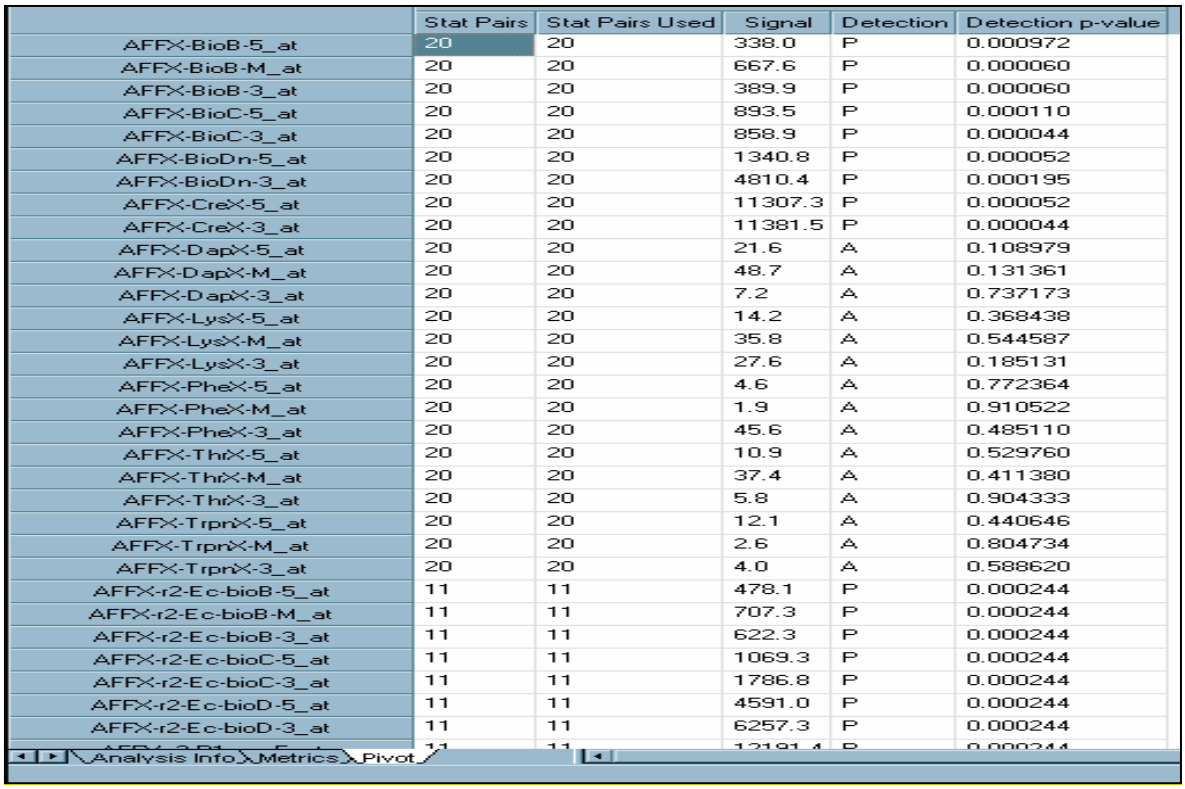
The Statistical Algorithms were developed using standard statistical techniques. The performance was validated using an experimental design called the Latin Square. In this experimental design, transcripts, naturally absent in the complex background, were spiked in at known concentrations.

Single Array Analysis

Single array analysis can be used to build databases of gene expression profiles, facilitate sample classification and transcript clustering, and monitor gross expression characteristics. In addition, the analyses provide the initial data required to perform comparisons between experiment and baseline arrays.

This analysis generates a Detection *p*-value which is evaluated against user-definable cut-offs to determine the Detection call. This call indicates whether a transcript is reliably detected (Present) or not detected (Absent). Additionally, a Signal value is calculated which assigns a relative measure of abundance to the transcript.

Figure 12 illustrates the output of Single Array Analysis in GeneChip Operating Software.



	Stat Pairs	Stat Pairs Used	Signal	Detection	Detection p-value
AFFX-BioB-5_at	20	20	338.0	P	0.000972
AFFX-BioB-M_at	20	20	667.6	P	0.000060
AFFX-BioB-3_at	20	20	389.9	P	0.000060
AFFX-BioC-5_at	20	20	893.5	P	0.000110
AFFX-BioC-3_at	20	20	858.9	P	0.000044
AFFX-BioDn-5_at	20	20	1340.8	P	0.000052
AFFX-BioDn-3_at	20	20	4810.4	P	0.000195
AFFX-CreX-5_at	20	20	11307.3	P	0.000052
AFFX-CreX-3_at	20	20	11381.5	P	0.000044
AFFX-DapX-5_at	20	20	21.6	A	0.108979
AFFX-DapX-M_at	20	20	48.7	A	0.131361
AFFX-DapX-3_at	20	20	7.2	A	0.737173
AFFX-LysX-5_at	20	20	14.2	A	0.368438
AFFX-LysX-M_at	20	20	35.8	A	0.544587
AFFX-LysX-3_at	20	20	27.6	A	0.185131
AFFX-PheX-5_at	20	20	4.6	A	0.772364
AFFX-PheX-M_at	20	20	1.9	A	0.910522
AFFX-PheX-3_at	20	20	45.6	A	0.485110
AFFX-ThrX-5_at	20	20	10.9	A	0.529760
AFFX-ThrX-M_at	20	20	37.4	A	0.411380
AFFX-ThrX-3_at	20	20	5.8	A	0.904333
AFFX-TrpnX-5_at	20	20	12.1	A	0.440646
AFFX-TrpnX-M_at	20	20	2.6	A	0.804734
AFFX-TrpnX-3_at	20	20	4.0	A	0.588620
AFFX-r2-Ec-bioB-5_at	11	11	478.1	P	0.000244
AFFX-r2-Ec-bioB-M_at	11	11	707.3	P	0.000244
AFFX-r2-Ec-bioB-3_at	11	11	622.3	P	0.000244
AFFX-r2-Ec-bioC-5_at	11	11	1069.3	P	0.000244
AFFX-r2-Ec-bioC-3_at	11	11	1786.8	P	0.000244
AFFX-r2-Ec-bioD-5_at	11	11	4591.0	P	0.000244
AFFX-r2-Ec-bioD-3_at	11	11	6257.3	P	0.000244

Figure 12. Data analysis output (.chp file) for a Single Array Analysis includes Stat Pairs, Stat Pairs Used, Signal, Detection, and the Detection *p*-value.

Detection Algorithm

The Detection algorithm uses probe pair intensities to generate a Detection *p*-value and assign a Present, Marginal, or Absent call. Each probe pair in a probe set is considered as having a potential vote in determining whether the measured transcript is detected (Present) or not detected (Absent). The vote is described by a value called the Discrimination score [R]. The score is calculated for each probe pair and is compared to a predefined threshold Tau. Probe pairs with scores higher than Tau vote for the *presence* of the transcript. Probe pairs with scores lower than Tau vote for the *absence* of the transcript. The voting result is summarized as a *p*-value. The greater the number of discrimination scores calculated for a given probe set that are above Tau, the smaller the *p*-value and the more likely the given transcript is truly Present in the sample. The *p*-value associated with this test reflects the confidence of the Detection call.

Detection *p*-value

A two-step procedure determines the Detection *p*-value for a given probe set.

1. Calculate the Discrimination score [R] for each probe pair.
2. Test the Discrimination scores against the user-definable threshold Tau.

The Discrimination score is a basic property of a probe pair that describes its ability to detect its intended target. It measures the target-specific intensity difference of the probe pair (PM-MM) relative to its overall hybridization intensity (PM+MM):

$$\mathbf{R = (PM - MM) / (PM + MM)}$$

For example, if the PM is much larger than the MM, the Discrimination score for that probe pair will be close to 1.0 (e.g., probe pair 1 in Figure 13). If the Discrimination scores are close to 1.0 for the majority of the probe pairs, the calculated Detection *p*-value will be lower (more significant). A lower *p*-value is a reliable indicator that the result is valid and that the probability of error in the calculation is small. Conversely, if the MM is larger than or equal to the PM, then the Discrimination score for that probe pair will be negative or zero (e.g., probe pairs 8, 9, and 10 in Figure 13). If the Discrimination scores are low for the majority of the probe pairs, the calculated Detection *p*-value will be higher (less significant).

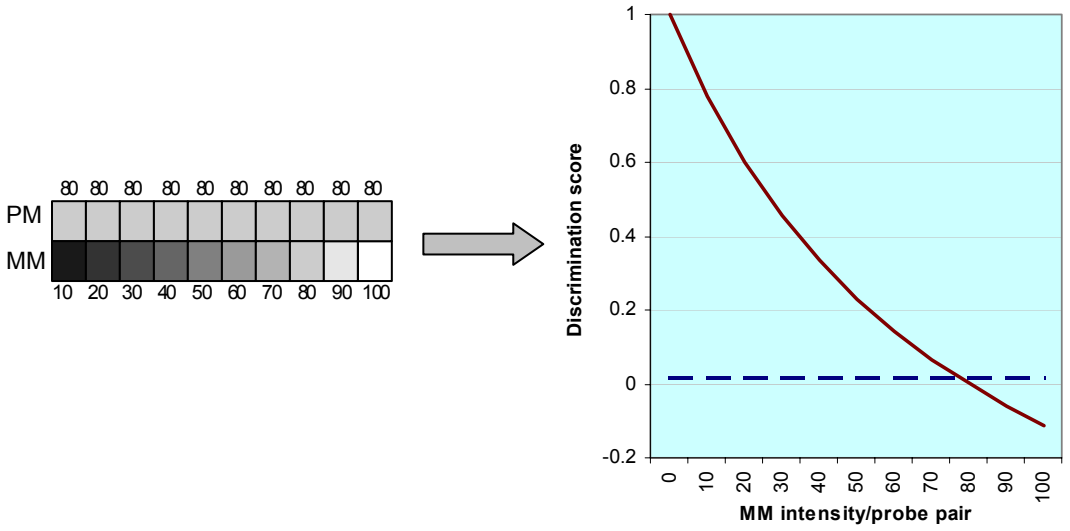


Figure 13. **In this hypothetical probe set, the Perfect Match (PM) intensity is 80 and the Mismatch (MM) intensity for each probe pair increases from 10 to 100. The probe pairs are numbered from 1 to 10. As the Mismatch (MM) probe cell intensity, plotted on the x-axis, increases and becomes equal to or greater than the Perfect Match (PM) intensity, the Discrimination score decreases as plotted on the y-axis. More specifically, as the intensity of the Mismatch (MM) increases, our ability to discriminate between the PM and MM decreases. The dashed line is the user-definable parameter Tau (default = 0.015).**

The next step toward the calculation of a Detection *p*-value is the comparison of each Discrimination score to the user-definable threshold Tau. Tau is a small positive number that can be adjusted to increase or decrease sensitivity and/or specificity of the analysis (default value = 0.015). The One-Sided Wilcoxon's Signed Rank test is the statistical method employed to generate the Detection *p*-value. It assigns each probe pair a rank based on how far the probe pair Discrimination score is from Tau.

Tunable Parameter Tip: Increasing the threshold Tau can reduce the number of false Present calls, but may also reduce the number of true Present calls. Note: Changing Tau directly influences the calculation of the Detection *p*-value. Please refer to the Tunable Parameters tech note, "Fine Tuning Your Data Analysis: Tunable Parameters of the Affymetrix® Expression Analysis Statistical Algorithms" for more information.

Detection Call

The user-modifiable Detection *p*-value cut-offs, Alpha 1 (α_1) and Alpha 2 (α_2) (See Figure 14), provide boundaries for defining Present, Marginal, or Absent calls. At the default settings, determined for probe sets with 16–20 probe pairs (defaults $\alpha_1 = 0.04$ and $\alpha_2 = 0.06$), any *p*-value that falls below α_1 is assigned a Present call, and above α_2 is assigned an Absent call. Marginal calls are given to probe sets which have *p*-values between α_1 and α_2 (see Figure 14). The *p*-value cut-offs can be adjusted to increase or decrease sensitivity and specificity.

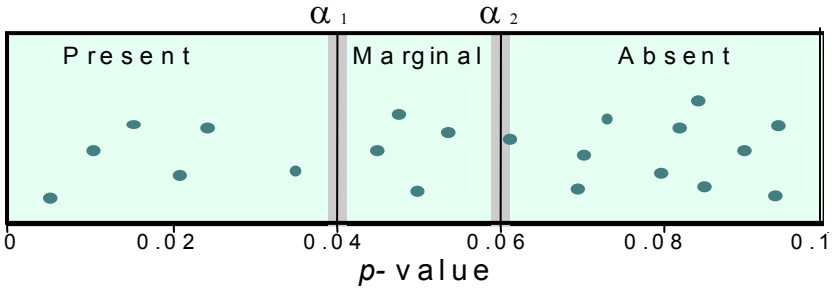


Figure 14. Significance levels α_1 and α_2 define cut-offs of *p*-values for Detection calls. Please note that these cut-offs are for probe sets with 16-20 probe pairs.

Significance levels α_1 and α_2 define cut-offs of *p*-values for Detection calls. Note that these cut-offs are for probe sets with 16–20 probe pairs.

It is important to note that prior to the two-step Detection *p*-value calculation, the level of photomultiplier saturation for each probe pair is evaluated. If all probe pairs in a probe set are saturated, the probe set is immediately given a present call.

In summary, the Detection Algorithm assesses probe pair saturation, calculates a Detection *p*-value, and assigns a Present, Marginal, or Absent call.

Signal Algorithm

Signal is a quantitative metric calculated for each probe set, which represents the relative level of expression of a transcript. Signal is calculated using the One-Step Tukey's Biweight Estimate which yields a robust weighted mean that is relatively insensitive to outliers, even when extreme.

Similar to the Detection algorithm, each probe pair in a probe set is considered as having a potential vote in determining the Signal value. The vote, in this case, is defined as an estimate of the real signal due to hybridization of the target. The mismatch intensity is used to estimate stray signal. The real signal is estimated by taking the log of the Perfect Match intensity after subtracting the stray signal estimate. The probe pair vote is weighted more strongly if this probe pair Signal value is closer to the median value for a probe set. Once the weight of each probe pair is determined, the mean of the weighted intensity values for a probe set is identified. This mean value is corrected back to linear scale and is output as Signal.

When the Mismatch intensity is lower than the Perfect Match intensity, then the Mismatch is informative and provides an estimate of the stray signal. Rules are employed in the Signal algorithm to ensure that negative Signal values are not calculated. Negative values do not make physiological sense and make further data processing, such as log transformations, difficult. Mismatch values can be higher than Perfect Match values for a number of reasons, such as cross hybridization. If the Mismatch is higher than the Perfect Match, the Mismatch provides no additional information about the estimate of stray signal. Therefore, an imputed value called Idealized Mismatch (IM) is used instead of the uninformative Mismatch (see Figure 15).

The following rules are applied:

> Rule 1: If the Mismatch value is less than the Perfect Match value, then the Mismatch value is considered informative and the intensity value is used directly as an estimate of stray signal.
>
> Rule 2: If the Mismatch probe cells are generally informative across the probe set except for a few Mismatches, an adjusted Mismatch value is used for uninformative Mismatches based on the biweight mean of the Perfect Match and Mismatch ratio.
>
> Rule 3: If the Mismatch probe cells are generally uninformative, the uninformative Mismatches are replaced with a value that is slightly smaller than the Perfect Match.
>
> These probe sets are generally called Absent by the Detection algorithm.

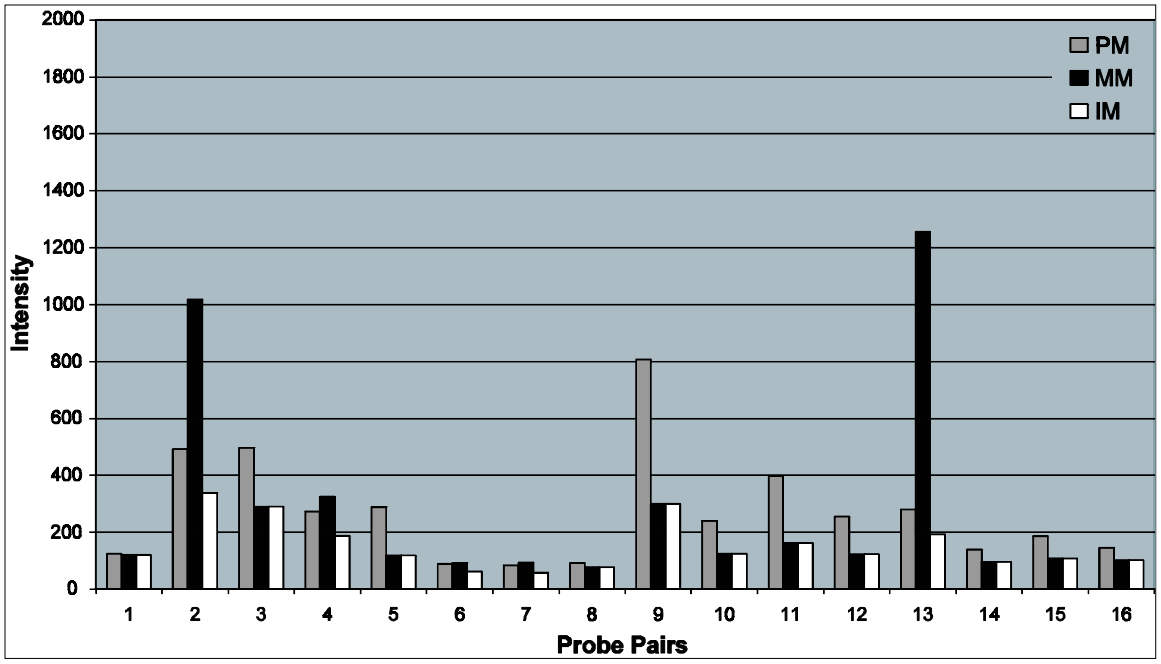


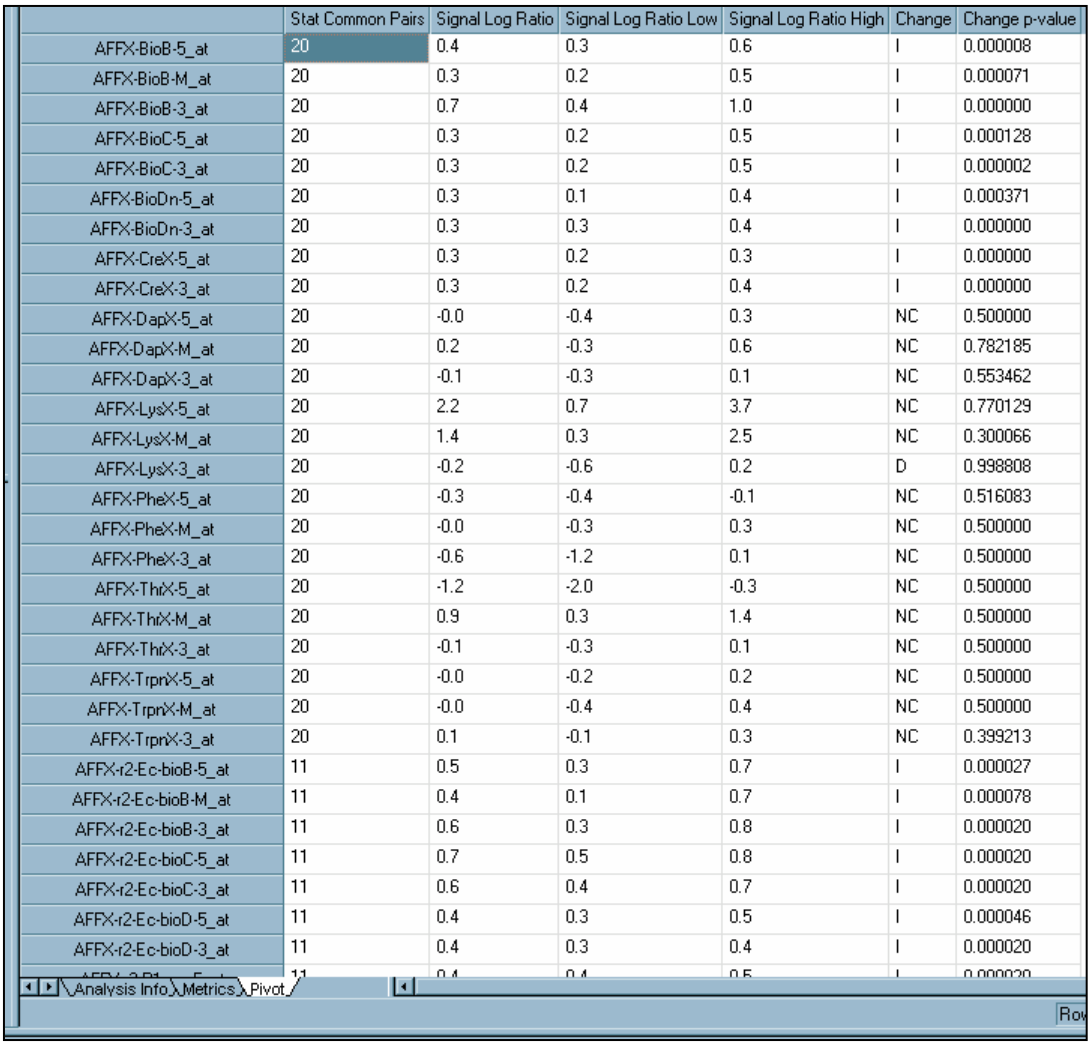
Figure 15. The grey bars illustrate the Perfect Match (PM) intensities and black bars the Mismatch (MM) intensities across a 16-probe pair probe set. The white bars, Idealized Mismatch (IM), are the intensities of the Mismatch based on the Signal rules. In this example, most of the Perfect Match intensities are higher than the Mismatch intensities and therefore Mismatch values can be used directly (e.g., probe pair 9).

When the Mismatch is larger than the Perfect Match (e.g., probe pairs 2, 4, and 13) the IM value is used instead of the Mismatch.

Comparison Analysis (Experiment versus Baseline arrays)

In a Comparison Analysis, two samples, hybridized to two GeneChip® probe arrays of the same type, are compared against each other in order to detect and quantify changes in gene expression. One array is designated as the baseline and the other as an experiment. The analysis compares the difference values (PM-MM) of each probe pair in the baseline array to its matching probe pair on the experiment array. Two sets of algorithms are used to generate change significance and change quantity metrics for every probe set. A change algorithm generates a Change *p*-value and an associated Change. A second algorithm produces a quantitative estimate of the change in gene expression in the form of Signal Log Ratio.

Figure 16 illustrates the output of Comparison Analysis in GeneChip® Operating Software (GCOS).



	Stat Common Pairs	Signal Log Ratio	Signal Log Ratio Low	Signal Log Ratio High	Change	Change p-value
AFFX-BioB-5_at	20	0.4	0.3	0.6	I	0.000008
AFFX-BioB-M_at	20	0.3	0.2	0.5	I	0.000071
AFFX-BioB-3_at	20	0.7	0.4	1.0	I	0.000000
AFFX-BioC-5_at	20	0.3	0.2	0.5	I	0.000128
AFFX-BioC-3_at	20	0.3	0.2	0.5	I	0.000002
AFFX-BioDn-5_at	20	0.3	0.1	0.4	I	0.000371
AFFX-BioDn-3_at	20	0.3	0.3	0.4	I	0.000000
AFFX-CreX-5_at	20	0.3	0.2	0.3	I	0.000000
AFFX-CreX-3_at	20	0.3	0.2	0.4	I	0.000000
AFFX-DapX-5_at	20	-0.0	-0.4	0.3	NC	0.500000
AFFX-DapX-M_at	20	0.2	-0.3	0.6	NC	0.782185
AFFX-DapX-3_at	20	-0.1	-0.3	0.1	NC	0.553462
AFFX-LysX-5_at	20	2.2	0.7	3.7	NC	0.770129
AFFX-LysX-M_at	20	1.4	0.3	2.5	NC	0.300066
AFFX-LysX-3_at	20	-0.2	-0.6	0.2	D	0.998808
AFFX-PheX-5_at	20	-0.3	-0.4	-0.1	NC	0.516083
AFFX-PheX-M_at	20	-0.0	-0.3	0.3	NC	0.500000
AFFX-PheX-3_at	20	-0.6	-1.2	0.1	NC	0.500000
AFFX-ThrX-5_at	20	-1.2	-2.0	-0.3	NC	0.500000
AFFX-ThrX-M_at	20	0.9	0.3	1.4	NC	0.500000
AFFX-ThrX-3_at	20	-0.1	-0.3	0.1	NC	0.500000
AFFX-TrpnX-5_at	20	-0.0	-0.2	0.2	NC	0.500000
AFFX-TrpnX-M_at	20	-0.0	-0.4	0.4	NC	0.500000
AFFX-TrpnX-3_at	20	0.1	-0.1	0.3	NC	0.399213
AFFX-r2-Ec-bioB-5_at	11	0.5	0.3	0.7	I	0.000027
AFFX-r2-Ec-bioB-M_at	11	0.4	0.1	0.7	I	0.000078
AFFX-r2-Ec-bioB-3_at	11	0.6	0.3	0.8	I	0.000020
AFFX-r2-Ec-bioC-5_at	11	0.7	0.5	0.8	I	0.000020
AFFX-r2-Ec-bioC-3_at	11	0.6	0.4	0.7	I	0.000020
AFFX-r2-Ec-bioD-5_at	11	0.4	0.3	0.5	I	0.000046
AFFX-r2-Ec-bioD-3_at	11	0.4	0.3	0.4	I	0.000020

Figure 16. Data analysis output (.chp file) for a Comparison Analysis includes Stat Common Pairs, Signal Log Ratio, Signal Log Ratio Low, Signal Log Ratio High, Change, and the Change *p*-value.

Before comparing two arrays, scaling or normalization methods must be applied. Scaling and normalization correct for variations between two arrays. Two primary sources of variation in array experiments are biological and technical differences. Biological differences may arise from many sources, such as genetic background, growth conditions, dissection, time, weight, sex, age, and replication. Technical variation can be due to experimental variables such as quality and quantity of target hybridized, reagents, stain, and handling error. The minimization of variation is essential, but scaling and normalization techniques provide a means to remove differences and facilitate comparison analysis.

Normalization and scaling techniques can be applied by using data from a selected user-defined group of probe sets, or from all probe sets. When normalization is applied, the intensity of the probe sets (or selected probe sets) from the experiment array are normalized

to the intensity of the probe sets (or selected probe sets) on the baseline array. When scaling is applied, the intensity of the probe sets (or selected probe sets) from the experimental array and that from the baseline array are scaled to a user-defined target intensity. In general, global scaling (scaling to all probe sets) is the preferred method when comparing two arrays.

An additional normalization factor is defined in the Robust Normalization section described below. This 'robust normalization,' which is not user-modifiable, accounts for unique probe set characteristics due to sequence-dependent factors, such as affinity of the target to the probe and linearity of hybridization of each probe pair in the probe set.

Change Algorithm

As in the Single Array Analysis, the Wilcoxon's Signed Rank test is used in Comparison Analysis to derive biologically meaningful results from the raw probe cell intensities on expression arrays. During a Comparison Analysis, each probe set on the experiment array is compared to its counterpart on the baseline array, and a Change *p*-value is calculated indicating an increase, decrease, or no change in gene expression. User-defined cut-offs (gammas) are applied to generate discrete Change calls (Increase, Marginal Increase, No Change, Marginal Decrease, or Decrease).

Robust Normalization

After scaling or normalization of the array (discussed in the Comparison Analysis overview), a further robust normalization of the probe set is calculated. Once the initial probe set normalization factor is determined, two additional normalization factors are calculated that are slightly higher and slightly lower than the original. The range by which the normalization factor is adjusted up and down is specified by a user-modified parameter called perturbation. This supplementary normalization accounts for unique probe set characteristics due to sequence dependent factors, such as affinity and linearity. More specifically, this approach addresses the inevitable error of using an average intensity of the majority of probes (or selected probes) on the array as the normalization factor for every probe set on the array. The noise from this error, if unattenuated, would result in many false positives in expression level changes between the two arrays being compared. The perturbation value directly affects the subsequent *p*-value calculation. Of the *p*-values that result from applying the calculated normalization factor and its two perturbed variants, the one that is most conservative is used to estimate whether any change in level is justified by the data. The lowest value for perturbation is 1.00, indicating no perturbation. The highest perturbation value allowed is set at 1.49. Increasing the perturbation value increases the conservativeness of the change call. For example, changes that were called Increase with a smaller perturbation value may be called No Change with a higher perturbation value. A default was established at 1.1 based on calls made from the Latin Square data set. The perturbation factor and the Latin Square data set are described in more detail in the Affymetrix Technical Notes referenced in the back of this guide.

Change *p*-value

The Wilcoxon's Signed Rank test uses the differences between Perfect Match and Mismatch intensities, as well as the differences between Perfect Match intensities and background to compute each Change *p*-value.

From Wilcoxon's Signed Rank test, a total of three, one-sided *p*-values are computed for each probe set. The most conservative value is chosen to determine the change call. That is the value that is closest to 0.5 which signifies that no change is detected. These are combined to give one final *p*-value which is provided in the data analysis output (.chp file). The *p*-value ranges in scale from 0.0 to 1.0 and provides a measure of the likelihood of change and direction. Values close to 0.0 indicate likelihood for an increase in transcript expression level in the experiment array compared to the baseline, whereas values close to 1.0 indicate likelihood for a decrease in transcript expression level. Values near 0.5 indicate a weak likelihood for change in either direction (see Figure 17). Hence, the *p*-value scale is used to generate discrete change calls using thresholds. These thresholds will be described in the Change Call section.

	Change	Change p-value
AFFX-BioB-5_at	I	0.000008
AFFX-BioB-M_at	I	0.000071
AFFX-BioB-3_at	I	0.000000
AFFX-BioC-5_at	I	0.000128
AFFX-BioC-3_at	I	0.000002
AFFX-BioDn-5_at	I	0.000371
AFFX-BioDn-3_at	I	0.000000
AFFX-CreX-5_at	I	0.000000
AFFX-CreX-3_at	I	0.000000
AFFX-DapX-5_at	NC	0.500000
AFFX-DapX-M_at	NC	0.782185
AFFX-DapX-3_at	NC	0.553462
AFFX-LysX-5_at	NC	0.770129
AFFX-LysX-M_at	NC	0.300066
AFFX-LysX-3_at	D	0.998808
AFFX-PheX-5_at	NC	0.516083
AFFX-PheX-M_at	NC	0.500000
AFFX-PheX-3_at	NC	0.500000
AFFX-ThrX-5_at	NC	0.500000
AFFX-ThrX-M_at	NC	0.500000
AFFX-ThrX-3_at	NC	0.500000
AFFX-TrpnX-5_at	NC	0.500000
AFFX-TrpnX-M_at	NC	0.500000
AFFX-TrpnX-3_at	NC	0.399213
AFFX-r2-Ec-bioB-5_at	I	0.000027
AFFX-r2-Ec-bioB-M_at	I	0.000078
AFFX-r2-Ec-bioB-3_at	I	0.000020
AFFX-r2-Ec-bioC-5_at	I	0.000020
AFFX-r2-Ec-bioC-3_at	I	0.000020
AFFX-r2-Ec-bioD-5_at	I	0.000046
AFFX-r2-Ec-bioD-3_at	I	0.000020

Analysis Info | Metrics | Pivot

Figure 17. Data analysis output (.chp file) for a Comparison Analysis illustrating Change *p*-values with the associated Increase (I) or Decrease (D) call. Increase calls have Change *p*-values closer to zero and Decrease calls have Change *p*-values closer to one.

Tunable Parameter Tip: Increasing the perturbation value can reduce the number of false changes, but may also decrease the detection of true changes. Note: Changing perturbation factor affects the calculation of the *p*-value directly.

Change Call

The final Change *p*-value described above is categorized by cutoff values called gamma1 (γ_1) and gamma2 (γ_2) (see Figure 18). These cut-offs provide boundaries for the Change calls: Increase (I), Marginal Increase (MI), No Change (NC), Marginal Decrease (MD), or Decrease (D).

The user does not directly set α_1 and α_2; rather each is derived from two user-adjustable parameters, γL and γH. In the case of γ_1, the two user-adjustable parameters are called γ_1L and γ_1H (defaults for probe sets with 15-20 probe pairs: γ_1L= 0.0025 and γ_1H= 0.0025), which define the lower and upper boundaries for γ_1. Gamma2 (γ_2) is computed as a linear interpolation of γ_2L and γ_2H (defaults for probe sets with 15-20 probe pairs: γ_2L= 0.003 and γ_2H= 0.003) in an analogous fashion.

The ability to adjust the stringency of calls associated with high and low signal ranges independently makes it possible to compensate for effects that influence calls based on low and high signals. This feature, however, is not used by default because the defaults are set as γ_1L = γ_1H and γ_2L = γ_2H

It is important to note that, like in Detection *p*-value calculation, the level of photomultiplier saturation for each probe pair is evaluated. In the computation of Change *p*-value, any saturated probe cell, either in the Perfect Match or Mismatch, is rejected from analysis. The number of discarded cells can be determined from the Stat Common Pairs parameter.

In summary, the Change algorithm assesses probe pair saturation, calculates a Change *p*-value, and assigns an Increase, Marginal Increase, No Change, Marginal Decrease, or Decrease call.

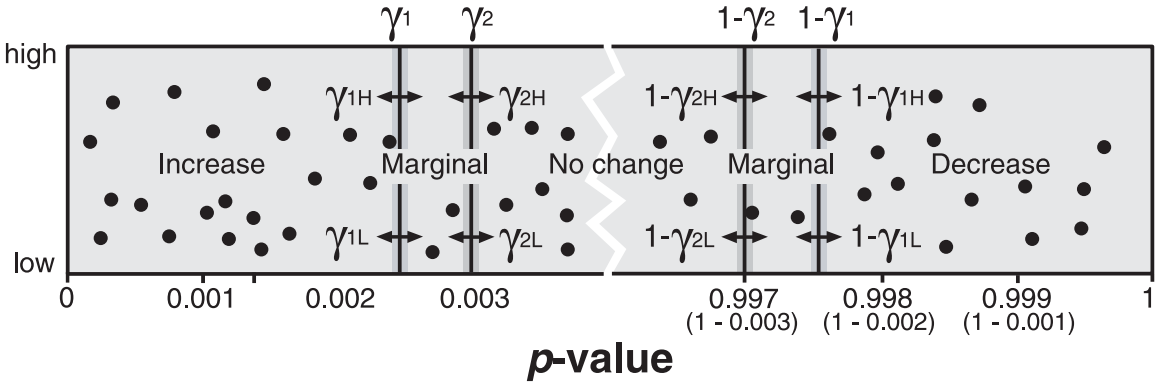


Figure 18. A representation of a range of *p*-values for a data set. The Y-axis is the probe set signal. The arrows on the vertical bars represent the adjustable γ values. The γ_1 value is a linear interpolation of γ_1L and γ_1H. Similarly γ_2 is derived from γ_2L and γ_2H.

Signal Log Ratio Algorithm

The Signal Log Ratio estimates the magnitude and direction of change of a transcript when two arrays are compared (experiment versus baseline). It is calculated by comparing each probe pair on the experiment array to the corresponding probe pair on the baseline array. This strategy cancels out differences due to different probe binding coefficients and is, therefore, more accurate than a single array analysis.

As with Signal, this number is computed using a one-step Tukey's Biweight method by taking a mean of the log ratios of probe pair intensities across the two arrays. This approach helps to cancel out differences in individual probe intensities, since ratios are derived at the probe level, before computing the Signal Log Ratio. The log scale used is base 2, making it intuitive to interpret the Signal Log Ratios in terms of multiples of two. Thus, a Signal Log Ratio of 1.0 indicates an increase of the transcript level by 2 fold and -1.0 indicates a decrease by 2 fold. A Signal Log Ratio of zero would indicate no change.

The Tukey's Biweight method gives an estimate of the amount of variation in the data, exactly as standard deviation measures the amount of variation for an average. From the scale of variation of the data, confidence intervals are generated measuring the amount of variation in the biweight estimate. A 95% confidence interval indicates a range of values, which will contain the true value 95% of the time. Small confidence intervals indicate that the data are more precise while large confidence intervals reflect uncertainty in estimating the true value. For example, the Signal Log Ratio for some transcripts may be measured as 1.0, with a range of 0.5 to 1.5 from low to high. For 95% of transcripts with such results, the true Signal Log Ratio will lie somewhere in that range. A set of noisy experiments might also report a Signal Log Ratio of 1.0, but with a range of -0.5 to 2.5, indicating that the true effect could easily be zero, since the uncertainty in the data is very large. The confidence intervals associated with Signal Log Ratio are calculated from the variation between probes, which may not reflect the full extent of experimental variation.

Terminology Comparison Table (Statistical Algorithms versus Empirical Algorithms)

Statistical Algorithms	Empirical Algorithms
Signal	Average Difference
Detection	Absolute Call
Change	Difference Call
Signal Log Ratio	Fold Change

The Logic of Logs

Quantitative changes in gene expression are reported as a Signal Log Ratio in the Statistical Algorithms as opposed to a Fold Change that was reported in the Empirical Algorithms.

The Benefit of Logs:

Hybridized probe intensities tend to be distributed over exponential space due to hybridization behavior that is governed by exponential functions of sequence-dependent

base-pairing energetics. Thus, log transformation is an appropriate process for analyzing hybridization data. Some of the benefits are apparent in Figure 19, where the same data set is plotted on two scales. When the data are plotted on a linear scale (solid) the single, high data point (7) overwhelms the graph and obscures information contained in the low values. When the same data are plotted on a Log_2 scale (dashed line), variations in the low values, as well as the very high values, are shown.

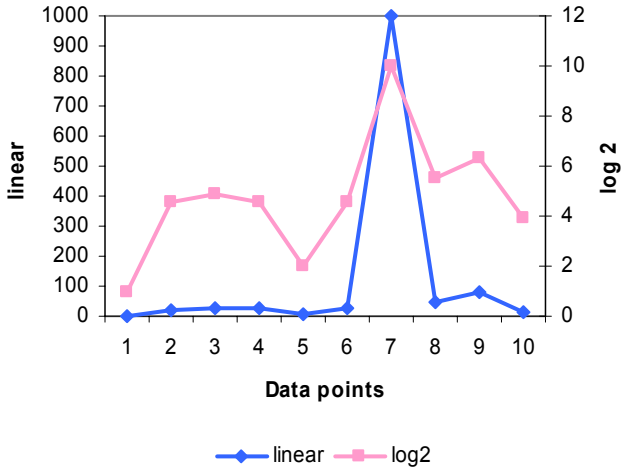


Figure 19. Illustration of the benefit of using logs. Variations in the high and low values are shown when data are plotted on a Log_2 scale.

Signal Log Ratio vs. Fold Change

Signal Log Ratio is compared to Fold Change in a hypothetical experiment in Figure 20. Baseline values were set to 1.5 and experiment values were reduced progressively from 6 to 0.375.

The X-axis illustrates the values that were decreased in the hypothetical experiment. The Y-axis represents units (e.g., Signal Log Ratio, Fold Change, or Signal for baseline and experiment).

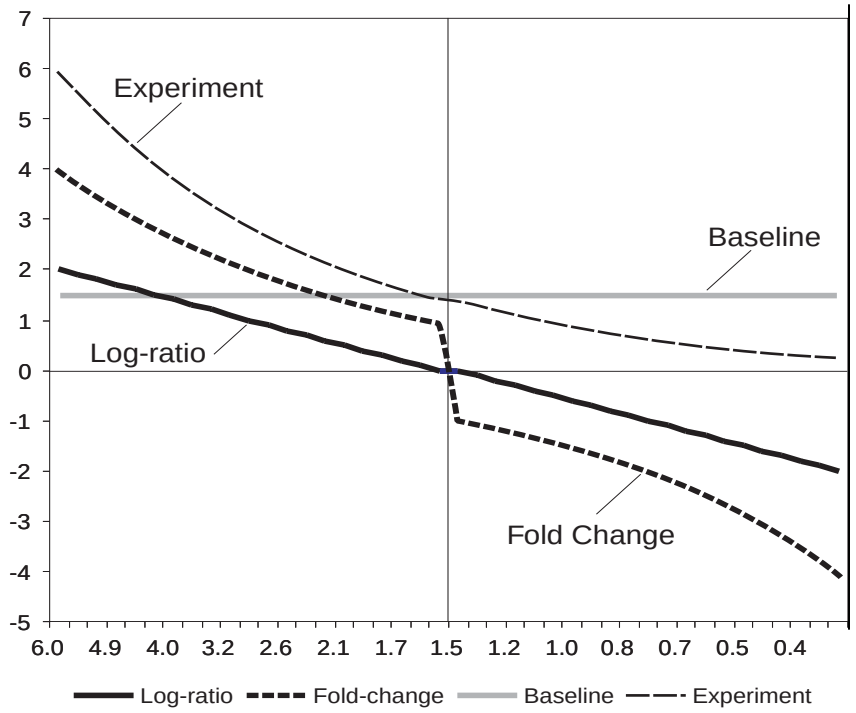


Figure 20.

There is a discontinuous transition where the experiment and the baseline converge and the fold change approaches 1 or -1. At this point (smaller changes), the fold change is less sensitive. Since we use $\log_2$, a Signal Log Ratio of 1 equals a Fold Change of 2 and a Signal Log Ratio of 2 equals a Fold Change of 4. Alternatively, use the following formula:

$$\text{Fold Change} = \begin{cases} 2^{\text{Signal Log Ratio}} & \text{Signal Log Ratio} \geq 0 \\ (-1) * 2^{-(\text{Signal Log Ratio})} & \text{Signal Log Ratio} < 0 \end{cases}$$

Basic Data Interpretation

The use of GeneChip® probe arrays allows interrogation of tens of thousands of transcripts simultaneously. One of the formidable challenges of this assay is to manage and interpret large data sets. This chapter provides users with guidelines for determining the most robust changes from a comparison analysis.

Metrics for Analysis

Which data analysis metrics should be used to determine the most significant transcripts when comparing an experimental sample to a baseline sample? GCOS provides users with both qualitative and quantitative measures of transcript performance. One standardized approach for sorting gene expression data involves the following metrics:

- Detection
- Change
- Signal Log Ratio

Detection is the qualitative measure of presence or absence for a particular transcript. A fundamental criterion for significance is the correlation of the Detection calls for a particular transcript between samples. When looking for robust increases, it is important to select for transcripts that are called "Present" in the experimental sample. When determining robust decreases, it is important to select for "Present" transcripts in the baseline sample. By following these initial guidelines, you will eliminate "Absent" to "Absent" changes, which are uninformative.

Change is the qualitative measure of increase or decrease for a particular transcript. When looking for both significant increases and decreases, it is important to eliminate "No Change" calls.

Signal Log Ratio is the quantitative measure of the relative change in transcript abundance. The Affymetrix Gene Expression Assay has been shown to identify Fold Changes of two or greater 98% of the time by Wodicka *et al.* in 1997 (15). Based on these observations, robust changes can be consistently identified by selecting transcripts with a Fold Change of >2 for increases and <2 for decreases. This corresponds to a Signal Log Ratio of 1 and -1, respectively. These value guidelines apply when performing a single comparison analysis.

NOTE: Please refer to "Introduction to Replicates" below in this chapter for exceptions.

Interpretation of Metrics

When sorting through gene expression data in GCOS, you will notice that some transcripts provide conflicting information. Here are some examples:

1. A transcript is called "Increase" but has a Signal Log Ratio of less than 1.0.
2. A transcript is called "No Change" but has a Signal Log Ratio of greater than 1.0.
3. A transcript is called "Absent" in both experimental and baseline files but is also called "Increase."

These contradictions arise due to the fact that Detection, Change, and Signal Log Ratio are calculated separately. The benefit of this approach is that transcripts can be assessed using three independent metrics.

Thus, in order to determine the most robust changes, it is crucial to use all three metrics in conjunction. The following section outlines this process.

Sorting for Robust Changes

Basic steps for determining robust increases:

1. Eliminate probe sets in the experimental sample called "Absent."
2. Select for probe sets called "Increase."*
3. Eliminate probe sets with a Signal Log Ratio of below 1.0.

Basic steps for determining robust decreases:

1. Eliminate probe sets in the baseline sample called "Absent."
2. Select for probe sets called "Decrease."*
3. Eliminate probe sets with a Signal Log Ratio of above -1.0.

NOTE: For detailed sorting instructions, please refer to "Performing Comparison Analysis" in Chapter 4.

* For those who wish to relax the Change criterion, include "Marginal Increase" and "Marginal Decrease" during selection.

"Real" Changes vs. "False" Changes

The procedures listed above can be used to determine both "Real" and "False" changes. The difference between "Real" and "False" changes lies in the relationship between the samples being compared. If the samples are different (e.g., normal vs. diseased, control vs. treated, etc.), the procedures will highlight transcripts that change significantly from the baseline sample to the experimental sample. If the samples are identical (i.e., hybridization replicates), no changes are expected. Thus, any transcripts showing significant change are false changes.

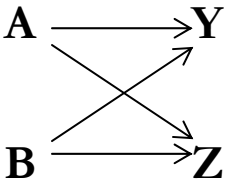
Note on Signal Log Ratio

When applying the sorting functions on Signal Log Ratio in GCOS (i.e., "Sort Ascending" and "Sort Descending"), you will notice that the column sorts on the magnitude of the Signal Log Ratio value, and not on the sign. Keep this in mind when sorting for robust changes.

Introduction to Replicates

The guidelines outlined in "Sorting for Robust Changes" above apply to a single comparison analysis. However, when biological replicates are introduced and multiple comparisons are generated, it becomes possible to relax the sorting thresholds based on consensus.

For example, here is an experiment with two sets of replicate samples consisting of two control samples (A and B) and two experimental samples (Y and Z). Performing pair-wise comparisons results in the following matrix:



This set of four analyses (A to Y, B to Y, A to Z, and B to Z) are comparison replicates. Each transcript has essentially been interrogated four times. The following is a hypothetical set of metrics for one transcript to determine whether or not it has increased:

Comparison	Detection in Exp.	Change in Exp.	Signal Log Ratio
A to Y	A	I	1.3
B to Y	P	I	1.2
A to Z	P	I	0.9
B to Z	P	I	1.2

*** Note**: "Exp." refers to the experimental sample.

Following the change guidelines for a single comparison analysis, the "Absent" call in the "A to Y" comparison would throw out this transcript. Likewise, the 0.9 Signal Log Ratio value would throw out the transcript in the "A to Z" comparison.

Overall, the transcript appears to be increasing since two of the four comparisons meet all three conditions for determining robust change and the other two comparisons meet two out of the three conditions. Based on overall consensus, we may choose to accept this transcript as a robust change.

The number of replicates to utilize and the conditions for acceptance of change are variable and up to the discretion of the user. However, the benefit of replicates in gene expression data (as with other assay data) is clear.

More advanced data analysis can be carried out in advanced data mining software.

Chapter 6 Statistical Analysis

In this section the intent is to help researchers establish a general understanding of which statistical methods may be used for advanced analysis of gene expression data. We recognize that there are a number of novel and very complex tests that are available or are being developed for analysis of large data sets, such as microarray data. However, the statistics subsequently discussed are available in common statistical software and are sufficiently robust to accurately determine statistical significance. As this is an overview, we suggest you consult one of the many statistical textbooks for the finer points of biostatistics (7).

A common early step in microarray data analysis is log transformation. Typically, log base 10 is used; however, log base 2 or natural log will work equally well. Log transformation has several important effects on the data (8). The most critical reason to log transform microarray data is that some of the error in the signal intensity measurement increases as the magnitude of signal intensity increases. That is, small numbers have less error in an absolute sense than higher numbers. Fortunately, higher numbers have roughly the same percentage error as small numbers. This roughly constant factor can be simply calculated and subtracted to normalize the data once the signals have been log transformed. There are additional effects of logging that make log transformed microarray data more closely fit statistical assumptions when applying statistical test methods. Log transformation makes data more symmetrical, one of the standard assumptions of normality. Log transformation also reduces the influence of a single measurement. Means on a log scale are more like geometric means, which are resistant to the effects of outliers, and it follows that outliers result in better estimates of variance. So, by log transforming data, common statistical methods are made more reasonable and provide more accurate insights to the biologist.

The key to using statistics when analyzing data is to determine which test is most appropriate to use, which in part is determined by the experimental design. Most common statistical methods fall into one of two categories. The first category, parametric statistics, uses the numerical data, such as the arithmetic mean, standard deviation, etc., to determine significant differences between sets of data. To utilize these statistical tests, assumptions regarding the normal distribution of the data, equality of variance among the groups, and general population normalcy must be made. These assumptions are often sufficiently satisfied to make parametric statistics extremely useful and a viable starting point for analysis. However, if data are generated from populations that do not meet these assumptions, these methods become unreliable because the mean and variance will no longer completely describe the population. This is a critical point, as parametric statistical methods essentially test for the degree of overlap of the population variance and determine the chance occurrence of this overlap when comparing differences between populations. Skewing of the data from non-normal variances will thus lead to false conclusions regarding the data set.

There are numerous statistical tests that can estimate if a population follows a normally distributed pattern. Unfortunately, these tests can be quite elaborate and are not robust enough to provide unambiguous conclusions. A simple method for examining the distribution of data is to compare the mean and median values. The mean is simply the sum of the data points within a group divided by the number of members in that group. The median is the data point that lies directly in the middle of all of the values. In other words, there are an equal number of data points on each side of this central value (i.e., the median value falls on the 50^{th} percentile of the data set).

In a normally distributed population, roughly 67% of the data fall within one standard deviation from the mean. In this same group, 95% of the data should fall within two standard deviations of the mean. If the same population is examined using the median, then the number of data points found within the 16th percentile and 84th percentile should be close to 67%. It should also be expected that 95% of the data points will fall between the 2.5th percentile and 97.5th percentile. Thus, if a sample population does follow a normal distribution, the mean and median should be of similar values with the data having a similar distribution around these values.

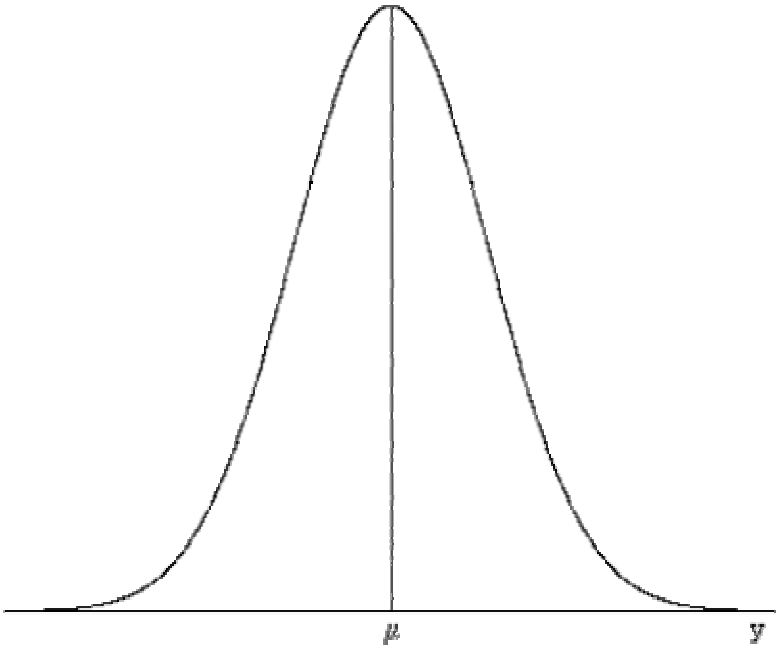


Figure 21. Image of a standard Gaussian curve.

When assessing if a population follows a normal distribution, the data should form a graph with the following characteristics.

- The graph should have a single peak at the center, which occurs at μ (the **mean**).
- The graph should be symmetrical.
- The graph should continue along the horizontal axis to infinity.
- The area under the graph should always equal one.
- Approximately 68% of the data should lie within one standard deviation (y) from the mean, 95% should lie within two standard deviations (2y), and 99.7% should lie within three standard deviations (3y).

Parametric statistics test the hypothesis that one or more treatments have no effect on the mean and variance of a chosen variable. As mentioned earlier, these tests are based on the assumption that the data are taken from a normally distributed population. However, experiments often yield data that are not consistent with these assumptions of data normality. In these cases, objects can also be tested in an ordinal, rather than an interval,

scale using a second, non-parametric approach that uses ranks of numerical data rather than the data themselves. This technique uses information about the relative sizes of observations without making any assumptions about the means and variances of the populations being tested and, thus, non-parametric methods can be used for analysis of any data set. However, if the data are normally distributed, then the parametric methods will be more powerful, that is, detect more data of significance.

In the following section, we review a series of statistical tests that can be used when analyzing GeneChip® array data. Note that the choice of tests to be used is highly dependent on experimental design. We highly encourage users of GeneChip expression arrays to plan experiments carefully to maximize the power of these tests for their projects.

Two Sample Statistical Tests

T-test

Student's t-test, often known as a simple t-test, is likely the most commonly used parametric statistical test. The t-test assesses whether the means of two groups are statistically different from each other. This statistic accomplishes this task by examining the differences between the means relative to the spread, or variance, of the data. The formula for the t-test consists of determining the ratio of the difference between the two means and the measure of the variability between the two data sets.

To test for significance, a risk level needs to be established; that is, a rate for acceptable false error. In most scientific research, this level is set at 0.05. This is considered to be the statistical determination of true differences between the two conditions tested, with a chance of being incorrect one in 20 times (a Type I error). The degrees of freedom are directly related to the number of data points and are determined by the experimental design. In the t-test, the degrees of freedom are the sum of the samples in both groups minus 2. Given the predetermined level of significance and the degrees of freedom, a t-value can be looked up in a standard table to determine whether it is large enough to be significant. If it is, you can conclude that the difference between the means for the genes in the two groups is different (even given the variability) with a set probability of false acceptance.

There are two basic versions of the t-test—the unpaired t-test (see Example 1 below) and the paired t-test (see Example 2 below)—where the data points in both groups are from two separate experimental populations. A typical experiment would be if we were comparing expression patterns of genes in two groups of patients. This is an example requiring the unpaired t-test, which is more common in most experiments involving expression data. Unpaired t-tests also have the advantage that the data do not require the two groups to be the same size.

It is possible that an experiment could be designed to examine the effect of an experimental maneuver in a single individual. An example would be measuring expression patterns of a particular gene within a single tissue sample before and after a treatment with a drug. In this case, the paired t-test would be the correct technique to use. A paired t-test is very powerful in that it compares the exact data point (e.g., a gene) between the two treatments within an experiment, and so, variations in baseline and experimental values between experiments are mitigated. It should be noted that this test should be used only in cases where the sample

population is truly paired and not in some of the more ambiguous scenarios (cell lines in the same passage, for example).

Example 1 -- Unpaired T-test

$$T = \frac{X_1 - X_2}{\sqrt{\frac{S^2 p}{N_1} + \frac{S^2 p}{N_2}}}$$

$X_1 - X_2 = \text{difference between means}$

$\sqrt{\frac{S^2 p}{N_1} + \frac{S^2 p}{N_2}} = \text{standard error}$

$S^2 p = \text{pooled variance}$

$N_1 = \text{population \# of group 1}$

$N_2 = \text{population \# of group 2}$

In the following example, the signal values for probe set X from 12 arrays are given (n1 =6 control and n2 = 6 experimental). The unpaired t-test divides the difference between the means by the square root of their pooled variance. It is then determined at what level of significance the resulting t score falls.

	Control Group	Experimental Group
Signal R1	3700	4900
Signal R2	4000	5200
Signal R3	4200	4900
Signal R4	3900	5000
Signal R5	4100	4800
Signal R6	4000	4750

Mean1 = 3983 Sum of Squares 1 = 148334 v1 = 5

Mean2 = 4925 Sum of Squares 2 = 128750 v2 = 5

$$S^2 p = \frac{148334 + 128750}{5 + 5} = 29666.8$$

$$S_{mean1-mean2} = \sqrt{\frac{29666.8}{6}} + \sqrt{\frac{29666.8}{6} = 140.63}$$

$$t = \frac{3983 - 4925}{140.63} = -6.70$$

$t_{0.05,10} = 2.228$ as $6.7 > 2.228$ then reject Ho :

P < 0.0001. The two means are not the same.

A low *p*-value for this test (less than 0.05 for example) means that there is evidence that the difference in the two means are statistically significant. If the *p*-value associated with the t-test is small (< 0.05), there is evidence that the means are significantly different at the significance level reported by the *p*-value. If the *p*-value associated with the t-test is not small (> 0.05) you conclude that there is evidence that the means are not different.

Example 2 -- Paired T-test

$$T = \frac{X_1 - X_3}{\sqrt{\frac{\sum (d_1 - d_2)^2}{n-1}}}$$

$X_1 - X_3 =$ difference between means

$$\sqrt{\frac{\sum (d_1 - d_2)^2}{n-1}} = \text{standard error}$$

$\sum (d_1 - d_2)^2 =$ the variance of the diffence scores for each individual

$n - 1 =$ the sample number minus 1

In the following example, the signal values for probe set X from 10 arrays are given (n1 =5 control and n2 = 5 experimental). The paired t-test divides the differences between the means by the Standard Error of the differences between the means. The Standard Error is the product of the standard deviation of each pair's differences divided by the square root of the number of pairs. The level of significance between which the resulting t score falls is then determined.

	Before Treatment	After Treatment
Signal R1	3700	4900
Signal R2	4000	5200
Signal R3	4200	4900
Signal R4	3900	5000
Signal R5	4100	4800
Signal R6	4000	4750

Mean1 = 3983 v1 = 5

Mean2 = 4925 v2 = 5

$$\frac{SE(d_1 - d_2)}{5} = \sqrt{228.065} = 45.61$$

$$t = \frac{941.67}{45.61} = 20.65$$

$t_{0.05,10} = 2.228$ as $20.65 > 2.228$ then reject Ho :

P < 0.0001. The differences between the means is greater than 0. If the *p*-value associated with t is low (< 0.05), then there is a difference in means across the paired observations.

Mann-Whitney Test for Independent Samples

This test is used in place of a two-sample, unpaired t-test when the data sets being compared are not normally distributed (see Example 3 below). This test derives its robustness under this condition from the fact that it does not use calculation of variance as part of the hypothesis test, but, rather, relies on rankings of the numerical values. It requires random samples of sizes n1 and n2 from two completely independent groups. The test then consists of combining the two samples into one sample of size n1 + n2. The actual observations taken from the data are replaced with their ranks. For example, in a data set with 10 (5 control and 5 experimental) samples, the highest ranking point would receive a value of 10 and the lowest a value of 1. In case of a tie, the values are given the average of the two ranks. A sum of the ranks for each group is then calculated. Continuing with our 10-sample scenario, if the values of the 5 samples from set one were all greater than those from set two, the values of the sums from set one would be 40, and set two, 15. Conversely, if the two data sets have the same distribution, then the sum of the ranks of both groups should be close to the same value. A *p*-value for the null hypothesis that the two distributions are the same can then be generated.

Example 3 -- Mann-Whitney Test

$$U = \frac{n_1(n_1 + 1)}{2} - R_1$$

n_1 = # of individuals in group 1

R_1 = sum of the ranks for group 1

In the following example, the signal values for probe set X from 11 arrays are given (6 control and 5 experimental). The Mann-Whitney ranks the combined data set of unpaired members based upon their absolute values. The ranks are separated back to their respective groups and the resulting sums of ranks are then examined using the Mann-Whitney statistic. The resulting level of significance is then reported.

	Control Group	Experiment Group	Control Rank	Experiment Rank
Signal R1	4500	3700	7	9
Signal R2	5200	3300	2	11
Signal R3	4700	4600	4	6
Signal R4	5500	3500	1	10
Signal R5	5000	3900	3	8
Signal R6	4650		2	

n1 = 6; n2 = 5; N = 11; R1 = 22; R2 = 44

Ranks of N are assigned in either lowest to highest or vice-versa.

$$U = \frac{(6)(5)+(6)(7)}{2} - 22 = 29$$

$$U' = \frac{(6)(5)+(5)(6)}{2} - 44 = 1$$

Critical values = 5 and 6

P 0.05 = 23 as 29 > 23 then reject Ho:

P = 0.01. The two groups are not the same.

If the *p*-value is low, chances are there will be little overlap between the two distributions. If the *p*-value is not low, there will be a fair amount of overlap between the two groups.

The Wilcoxon Signed-Rank Test for Paired Data

A test similar to the Mann-Whitney, the Wilcoxon Signed-Rank Test is used when each experimental subject is observed before and after a single treatment, that is, it is the non-parametric alternative to the paired t-test (see Example 4 below). This statistic consists of sorting the absolute values of the differences from smallest to largest, then assigning ranks to the absolute values regardless of sign. The sum of the ranks of the positive differences is next determined. As with the Mann-Whitney, the distribution of all possible values of the test statistic can be obtained in which the treatment has no effect. If the null hypothesis is true, the sum of the ranks of the positive differences should be similar to the sum of the ranks of the negative differences. However, if the test statistic value falls outside of this range, the null hypothesis can be rejected indicating that the treatment did indeed have some effect.

Example 4 -- Wilcoxon Signed-Rank Test

W + the value of the signed ranks

In the following example, the signal values for probe set X from 12 arrays are given. (6 control and 6 experimental). The Wilcoxon Signed-Rank Test ranks the absolute values of the difference between each pair. If the difference between a pair is equal to 0, then that value is not used any further. Also, if the difference is identical between two pairs, the average rank of the two groups is used. The Sum of the ranks is then calculated and compared to the appropriate critical values and levels of significance found in a Wilcoxon table.

	Control	Experimental	Difference	Ranks
Signal R1	4500	3700	800	4
Signal R2	3200	3300	-100	-1.5
Signal R3	4700	4600	100	1.5
Signal R4	5500	3500	2000	6
Signal R5	5000	3900	1100	5
Signal R6	4250	4400	-150	-3

$W = 12$; $n = 6$

The critical values for $n = 6$ and $\alpha = 0.063$ are 1 and 20. These can be found in almost any basic level statistics manual. Since $1 < 12 < 20$ we would accept the null hypothesis that the Control Group and Experimental Group are not significantly different.

Multivariate Statistics

One-Way Analysis of Variance

Analysis of Variance (ANOVA) is one of the most commonly used multivariate statistics. Essentially, an ANOVA employs multiple estimates of a population's variance to determine the overall variability within a multiple-group analysis (see Example 5 below). There is no restriction on the number of groups that can be analyzed by ANOVA, and it is equally valid for testing differences between two groups or among 20. In the special case where there are only two groups, ANOVA is equivalent to the t-test. Since it is also a parametric test, it has the same limitations as the t-test: the observations must follow a normal distribution, the variance in the groups must be equal, and the data points in each group must be from independent samples.

With ANOVA, there are, in fact, two estimates of variance for each group taken. The first estimate of variance is based upon the standard deviation of each group. This variance is not affected by any differences in the means of the groups being tested since this information is generated within each group. Also, the variance should not differ, as this test makes the assumption of equal variance among groups. The second population variance estimate is based upon the variability between means of each group. If these estimates of each group's variability are the same, then it is expected that the overall variance among the groups is not different. However, if there are significant differences between the means, then this will obviously lead to the possibility of a changing population variance estimate.

Once these calculations have been determined, the ANOVA requires that the population variance estimate of the means be divided by the population variance estimate of the standard deviations. Depending on the size of the resulting test statistic, a *p*-value is generated and can be used to determine significance.

At its simplest, a one-way ANOVA can be used to test the hypothesis that some variable of interest differs among groups. There are more sophisticated versions of ANOVA, (see below for descriptions): two-way ANOVA can test for differences among groups while controlling for other categorical variables, and ANCOVA (analysis of covariance) can control for continuous variables, both of which are described below.

Example 5 -- One-Way Analysis of Variance (One-Way ANOVA)

$$F = \frac{\dfrac{n_1(x_1 - x) + ...n_k(x_k - x)}{k-1}}{\dfrac{(n_1 - 1)s_1^2 .....(n_k - 1)s_k^2 k}{N-k}}$$

$$\frac{n_1(x_1 - x) + ...n_k(x_k - x)}{k-1} = \text{MSTR} = \text{treatment mean square}$$

$$\frac{(n_1 - 1)s^2 .....(n_k - 1)s^2 k}{N - k} = \text{MSE} = \text{error mean square}$$

$n_1 .....n_k$ = # of individuals in each group

$(x_1 - x)$ = the mean of each group minus the average of all groups

k = # of classes

$s_1^2 .....s_k^2$ = the group variance for each

N = total # of individuals

In the following example, the signal values for probe set X from 18 arrays are given (6 controls, 6 with disease, and 6 disease + drug). The one-way ANOVA tests whether the means of more than two groups are equal. The one-way ANOVA examines the variation among the sample means by way of a measurement of the weighted average of the squared deviations around the mean of all of the sample data (MSTR – Treatment Mean Square). This value is derived from the sum of each group divided by the total number of replicates for each group. The sum of squares for each data point divided by the total number of arrays is then subtracted. Lastly, this value is once again divided by the number of classes minus one. This estimate of variance among groups is then divided by the variation within each group (MSE – Error Mean Square). This is found in two steps. The first is by calculating the variance within each group and multiplying it by the number of replicates in each group minus one. The second step involves dividing the preceding value by the number of arrays minus the number of classes. The resulting F-value can be compared to the F table using the number of classes minus one as one degree of freedom, and the total number of arrays minus the number of classes as the other.

	Control	Disease	Disease + Drug
Replicate 1	800	700	750
Replicate 2	750	690	720
Replicate 3	845	800	870
Replicate 4	795	650	815
Replicate 5	820	780	795
Replicate 6	900	750	850
Sum	4910	4370	4800

$$MSTR = \frac{\left(\left(\frac{4910}{6}\right)+\left(\frac{4370}{6}\right)+\left(\frac{4800}{6}\right)\right)-\left(\frac{11013689}{18}\right)}{2} = \frac{27144.33}{2} = 13572.17$$

$$MSE = \frac{46166.67}{15} = 3077.778$$

$$F = \frac{MSTR}{MSE} = 4.41$$

At the *p*-value = 0.05 level and degrees of freedom (2, 15) the critical value of F is 3.68. Therefore, 4.41 > 3.68 so we would reject that the groups are equal.

Two-Way Analysis of Variance

In a one-way ANOVA, the effects of various levels or treatment conditions of one independent variable on a dependent variable are examined. That is, we are investigating changes between multiple treatment conditions. Many experimental designs can be established to test the effect that two variables may have on a data set. For example, we may examine normal vs. tumor tissues, along with the effect of two different drugs, making a total of four different sample sets. If this investigation uses male and female subjects, then a two-way ANOVA can be used to investigate differences in gene expression between the different conditions, as well as male/female differences within and between each condition (see Example 6 below). In this case, two separate ANOVAs cannot adequately examine the possible interactions that can be generated between the two variables, and, so, a two-way analysis of variance is the best methodology. A two-way ANOVA consists of three significance tests: a test of each of the two main effects and a test of the interaction of the variables.

Example 6 -- Two-Way Analysis of Variance (Two-Way ANOVA)

$$F = \frac{\frac{\frac{1}{m}\left(T_1^2 .....T_k^2\right)-\frac{\left(\sum x\right)^2}{n}}{k-1}}{\frac{\left(\sum x^2-\frac{\left(\sum x\right)^2}{n}\right)-\left(\frac{1}{m}\left(T_1^2 .....T_k^2\right)-\frac{\left(\sum x\right)^2}{n}\right)-\left(\frac{1}{k}\left(B_1^2 .....B_k^2\right)-\frac{\left(\sum x\right)^2}{n}\right)}{n-k-m+1}}$$

$$\frac{\frac{1}{m}\left(T_1^2 \ldots\ldots T_k^2\right) - \frac{\left(\sum x\right)^2}{n}}{k-1} = \text{MSTR} = \text{treatment mean square}$$

$$\frac{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right) - \left(\frac{1}{m}\left(T_1^2 \ldots\ldots T_k^2\right) - \frac{\left(\sum x\right)^2}{n}\right) - \left(\frac{1}{k}\left(B_1^2 \ldots\ldots B_k^2\right) - \frac{\left(\sum x\right)^2}{n}\right)}{n-k-m+1} = \text{MSE} = \text{error mean square}$$

m = # of blocks

k = # of classes

n = (k)(M) = total # of pieces of data

$T_{1\ldots\ldots}^2 T_k^2$ = squared sum of sample data for each treatment (class or column)

$B_{1\ldots\ldots}^2 B_k^2$ = squared sum of sample data for each block (row)

X = mean for all groups

A Two-Way Analysis of Variance is designed to test whether the variance among several groups is the same and is similar in concept to the one-way ANOVA. The major difference is that a large portion of the variation within the groups can be due to a single extraneous variable which first needs to be mathematically isolated and removed so that it will be easier to detect true differences among the groups.

In the following example, the signal values for probe set X from 18 arrays are given (6 Drug A, 6 Drug B, and 6 Drug C). The goal of the test is to compare the three drugs for their impact on expression for this probe set. The variability associated with the Individuals will be compensated for, as it will have a major effect on the overall variability within each system. The two-way ANOVA examines the variation among the sample means by way of a measurement of the weighted average of the squared deviations around the mean of all of the sample data (MSTR – Treatment Mean Square). This value is derived from the total sum of squares for each group divided by the total number of replicates for each group. However, this value is now described by three components (instead of just two for the one-way ANOVA). The third calculation involves the sum of squares for the extraneous variable as well. The sum of squares for each data point divided by the total number of arrays is then subtracted. Lastly, this value is once again divided by the number of classes minus one.

This estimate of variance among groups is then divided by the variation within each group (MSE – Error Mean Square). In this case, the sum of squares for the extraneous variable is also addressed. The resulting F-value can be compared to the F table using the number of

classes minus one as one degree of freedom, and the total number of arrays minus the number of classes, minus the number of blocks, plus 1 as the other.

	Drug A	Drug B	Drug C	Sum of Blocks
Individual 1	500	650	800	1950
Individual 2	625	700	750	2075
Individual 3	575	675	675	1925
Individual 4	600	685	795	2080
Individual 5	595	645	725	1965
Individual 6	565	650	695	1910
Sum of Treatments	3460	4005	4440	11905

$$MSTR = \frac{\left(\left(\frac{3460}{6}\right)+\left(\frac{4005}{6}\right)+\left(\frac{4440}{6}\right)\right)-\left(\frac{7873834.7}{18}\right)}{2} = \frac{80369.444468}{2} = 40184.722$$

$$MSE = \frac{15647.222}{10} = 1564.7222$$

$$F = \frac{MSTR}{MSE} = 25.68$$

At the *p*-value = 0.05 level and degrees of freedom (2, 10) the critical value of F is 4.10. Therefore, 25.68 > 4.10 so we would reject that the groups are equal.

Kruskal-Wallis

The Kruskal-Wallis (see Example 7 below) is the non-parametric equivalent to an ANOVA. It is equally valid for testing differences between two groups or among 20. In the special case where there are only two groups, the Kruskal-Wallis is equivalent to the Mann-Whitney. Because the Kruskal-Wallis test is non-parametric, there are no assumptions that need to be made regarding the distribution of the observations. All observations are ranked without regard for the group in which they are found. After the sums of each group's observations are calculated, the distributions of the ranks are then compared.

As with the ANOVA, the Kruskal-Wallis is a statistic that examines overall variation among the groups. It does not offer information about which groups are significantly different. To determine these specifics, it is necessary to run multiple pair-wise comparisons.

Example 7 -- Kruskal-Wallis

$$H = \frac{n_1(x_1 - x)^2 .....n_k(x_k - x)^2}{\frac{N(N+1)}{12}}$$

$n_1.....n_k$ = # of individuals in each group

$x_1.....x_k$ = the mean of each group's ranks

X = the mean of all groups' ranks

N = total # of individuals in all groups

$n_1(x_1 - x)^2 .....n_k(x_k - x)^2 = SSD$ = sum of squared differences of ranks

In the following example, the signal values for probe set X from 15 arrays are given (5 group 1, 5 group 2, and 5 group 3). The Kruskal-Wallis tests whether or not the ranks of more than two groups are equal. It accomplishes this by first ranking each observation regardless of group. Then the Kruskal-Wallis statistic (H) is calculated by dividing the Sum of Squared Differences (SSD) by N(N+1)/12. This is used to generate a measurement that examines how the rank of each group compares with the average rank of all the groups. Lastly, the H statistic is compared to the Chi-square distribution with k-1 degrees of freedom. If the sample size is small, then it may be necessary to use k degrees of freedom to prevent an overly conservative estimate.

	Group 1	Rank	Group 2	Rank	Group 3	Rank

Replicate 1	800	10.5	700	3	750	5.5
Replicate 2	750	5.5	690	2	720	4
Replicate 3	845	14	800	10.5	870	15
Replicate 4	795	8.5	650	1	815	12
Replicate 5	820	13	780	7	795	8.5
Sum of Ranks		51.5		23.5		45
Mean of Ranks		10.3		4.7		9

Mean of All Ranks = 8; df = k – 1 = 2

$$SSD = 5(10.3-8)^2 + 5(4.7-8)^2 + 5(9-8)^2 = 73.90$$

$$H = \frac{73.90}{\frac{15(15+1)}{12}} = 3.695$$

4.605 > 3.695 > 2.773; *p*-value 0.25 > H > .10

At the *p*-value = 0.05 level and with degrees of freedom = 2; the groups are not significantly different.

Mitigating Type I and II Errors

With any statistical test, the possibility of making an erroneous assumption is always present. The reason for having numerous tests that vary, sometimes slightly, in their approach is to minimize the likelihood of making an invalid assumption. There are two types of errors that are discussed in general statistics. The first of these errors is the false assumption that two groups are in fact significantly different (Type I error of false positive). The *p*-value that is generated as a result of the statistical tests used is a direct estimate of the chance that you will make an error of this nature. For instance, a *p*-value of 0.05 indicates that the researcher expects to make a Type I error 5% of the time.

The second type of error (Type II error false negative) deals with the false assumption that the null hypothesis is correct. In other words, groups that are significantly different will not be identified as such. It is much more difficult to detect errors of this type as the *p*-value is not an indicator. Replication and increased sample size are the most likely methodologies to understand this phenomenon.

Parametric t-tests, like all two-sample analyses, should be used only when testing populations comprised of two samples. T-tests are invalid for doing multiple comparisons in a data set

(e.g., sample 1 vs. 2, 1 vs. 3, 2 vs. 3), as the population variance is not taken into account when performing each individual test. This can easily lead to false acceptance of data. In these cases, one of the Analysis of Variance tests should be used (please refer to the "Multivariate Statistics" section of this document).

The following table describes the potential for generating an increase in false positives by performing multiple pair-wise comparisons. In the table, k = the number of groups; K = the number of comparisons that are necessary; each subsequent column represents the chosen level of significance. As multiple comparisons are done, the table shows the actual level of significance that is being met.

k	K	p = 0.1	p = 0.05	p = 0.01	p = 0.001
2	1	0.1	0.05	0.01	0.001
3	3	0.27	0.14	0.03	0.003
4	6	0.47	0.26	0.06	0.006
5	10	0.65	0.4	0.1	0.01
6	15	0.79	0.54	0.14	0.015
10	45	0.99	0.9	0.36	0.044

Multiple Comparison Corrections

After an ANOVA has been utilized to determine variation among multiple groups, it may be necessary to understand which groups are significantly different. As stated above, a strict pair-wise analysis using a standard t-test can increase the likelihood of generating a Type I error. The reason this occurs is because the level of significance used in each pair-wise test (commonly 0.05) is more likely to be met with an increase in the number of pair-wise tests used. Therefore, a number of correction statistics have been developed to address this issue.

Bonferroni Correction

This correction is a simple modification of the Student's t-test. In this case, the cut-off level of significance being used is divided by the number of means being compared. Therefore, if you are using an initial cut-off of 0.05 and there are four groups in your comparison, then to use this correction you would divide 0.05 by 4 to get a *p*-value of 0.0125 cut-off for these comparisons. The Bonferroni correction is a conservative correction that works well with a small number of data points within a group (generally, $n < 8$). As the number of data points increases, the resulting *p*-value can become increasingly conservative. For example, if you test one gene by PCR on a set of control mice and compare it to a set of treated samples you would say that the change is statistically significant if a t-test of the contrasts have a *p*-value of 0.05 or lower. For a microarray experiment you would perform one test for each probe set ID on the signals and you would then rank the 12,488 *p*-values. For 95% confidence you divide 0.05 by 12,488 to get 4.0×10^{-6}. If any *p*-value is lower than 4.0×10^{-6} it passes the Bonferroni multiple correction tests. If all genes fail then simply selecting the percentile with the lowest *p*-values will yield 125 genes in a principled manner.

Thus, use of the Bonferroni correction for mitigating false error rates in data sets with large numbers of tests (i.e., microarray data) will result in highly conservative *p*-values. However

microarray data are not perfectly normally distributed, which means that using a parametric test, like Students t or ANOVA, gives better than expected values. In those cases the stringency of Bonferroni balances the overly significant result. Furthermore Bonferroni is so easily calculated that you should at least look to see if a sufficiently large number of genes pass this threshold. Practically speaking, genes that pass the significance thresholds using the Bonferroni correction will have a very low false-error rate. Thus, in assessing the data for significant results, you can begin with this calculated threshold and empirically relax the criteria based on other indications of significance such as fold-change and biological significance.

There are other strategies for mitigation of Type I errors. As stated above, while Bonferroni is often too conservative for this purpose, it is easily applied, where other common methods, such as Westfall-Young (9), or False Discovery Rate (10), are computational intensive and generally require some programming to implement. Bonferroni, and methods like it, provides a greater confidence in the results. Microarrays are the unusual statistical case where the number of tests greatly exceeds the number of samples, so standard statistical methods for multiple comparisons are pushed to their limit. This is why non-statistical approaches must be used in conjunction with statistical methods to interpret and validate the biological importance of the data.

Chapter 7 Biological Interpretation of GeneChip® Expression Data

The last step in the analysis of gene expression data is the biological interpretation of the results, where expression profiles contribute to the functional genomics characterization of the biological system under investigation. Depending on the goals of the experiment, this step allows the testing of specific hypotheses, or the generation of new insights and hypotheses, that helps guide further research. Finding the functional relevance of expression data requires gathering and organizing a variety of additional bioinformatics associated with the sequences that show significant changes. It also involves correlating expression results with other types of data that may be gathered as part of the experiment, such as genomic, proteomic, or metabolomic data. Such integrative approaches, sometimes termed 'Systems Biology,' aim to tackle the complexity of biological systems by gathering and incorporating all the available information into one comprehensive model.

The challenges of biological interpretation and the relative paucity of tools available have made this step the true bottleneck in microarray data analysis. One fundamental difficulty is the requirement for human review and understanding of complex types of data, scattered across a variety of sources, including online databases and journal publications. While there are efforts to develop tools that would truly automate some of the biological interpretation tasks, such as knowledge mining tools and gene network modeling and prediction, most investigators rely largely on 'manual' interpretation of results, through the review of functional annotations, pathway information, and associated literature. Tools that assist in these tasks should allow for the efficient mining and organization of annotations, as well as presentation of expression data in a functional context, such as within known metabolic or signaling pathways. This section will introduce some tools useful for mining and organizing annotations, as well as visualizing expression data within functional contexts, in order to assist in the biological interpretation of expression data.

Statistical Significance vs. Biological Relevance

As discussed in previous sections, a primary utility of significance metrics, such as *p*-values generated by statistical tests, is to rank results by confidence, and help in the estimation of false positives (Type I errors) and false negatives (Type II errors). These metrics allow the scientist to tune the analysis stringency to achieve the desired balance of sensitivity and specificity, resulting in a certain amount of flexibility (and arbitrariness) when interpreting significance metrics generated by a given test. As a result, the list of 'interesting genes' generated from an expression profiling experiment may change as the analysis is refined and as additional types of data, such as functional information, are added to the analysis.

Gene expression changes are controlled through highly complex, non-linear interactions between proteins, DNA, RNA, and a variety of metabolites. Any complex biological process is likely to involve many changes at the level of gene expression. Some of these changes will be of critical importance to the biological process of interest (either causal or directly consequential), while others may be peripheral or non-essential. The degree of 'relevance' of an observed change to the biological process under study is graded. In other words, biological processes are inherently "fuzzy" in terms of the importance of the vast range of interactions and changes that can be observed with a given biological assay (Figure 22). Any

type of biometric assay that attempts to pinpoint molecular changes in a complex biological system will be faced with the issue of applying discrete categorization on a continuum.

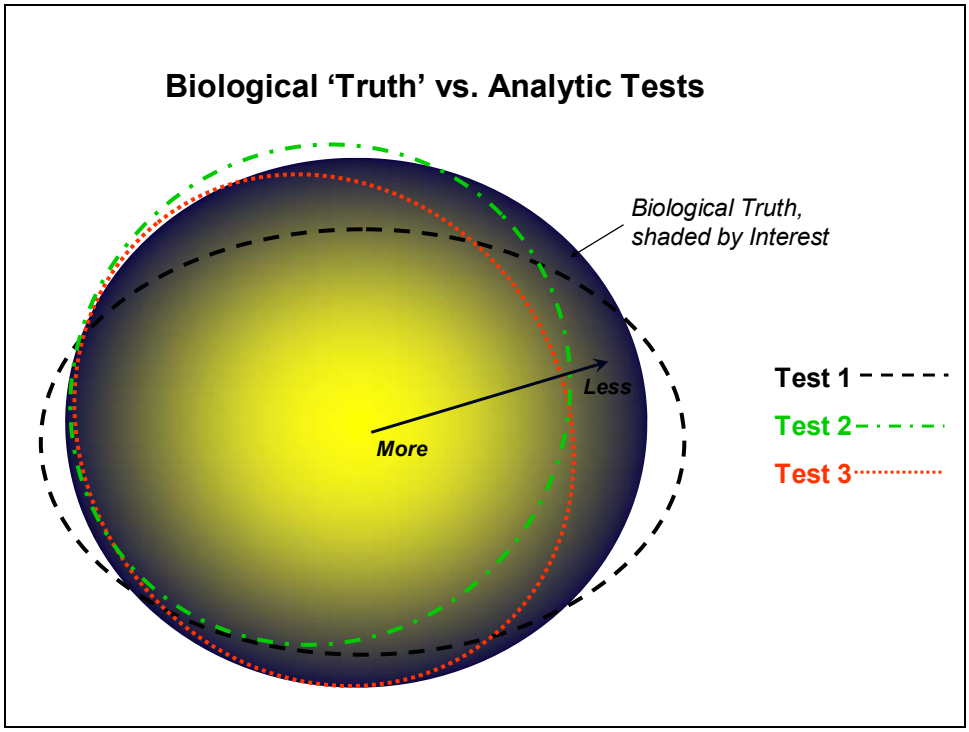


Figure 22. **The set of true molecular changes associated with a complex biological process is shown as a "fuzzy" mass, whose members are graded in terms of their relevance or "interest" to the biological process. Various analytic tests (represented by the dotted ovals) estimate the truth, with some degree of error.**

Therefore, biologically relevant gene expression changes may or may not be effectively captured with a given statistical test. The list of genes found to be significantly changing by a t-test, for example, may contain the majority of actual biological changes, but as previously explained, it will also suffer from some false positives and false negatives. Furthermore, a subset of the 'true' changes captured by the t-test may be biological noise or real biological changes that are not relevant to the process under investigation. One strategy to handle this inherent uncertainty when interpreting expression profiling data is to overlay functional information onto the statistical results, allowing biological context to help decide what is of interest and what is not.

Chapter 8 Annotation Mining Tools

While investigating the function of genes is often the most interesting and relevant part of analysis, it has also become one of the most challenging, given the amount and complexity of information. Once a list of genes exhibiting statistically significant expression patterns is generated, there is no clear method delineating how to move forward with data interpretation. For instance, while there are many databases with functional gene information, such as LocusLink, HomoloGene, RefSeq and UniGene databases, the answer to a simple question such as "What function does this gene perform?" may be in all of them, some of them, or none of them.

Affymetrix® NetAffx™ Analysis Center

The Affymetrix® NetAffx™ Analysis Center (www.affymetrix.com/analysis/) was developed to help researchers avoid this serious data bottleneck and to reach biologically meaningful results more quickly. One of the primary functions of the Analysis Center is to address the need described in the preceding section: providing comprehensive functional annotations that can be overlaid onto statistical results, enabling the creation of a list of both statistically and biologically significant genes. This online resource provides access to integrated biological annotations from a broad range of both public and Affymetrix-specific databases through one streamlined interface.

The Affymetrix Analysis Center serves as a convenient central resource for qualitative information in the Experiment Cycle (Figure 23). Rather than searching from database to database and entering information numerous times, the Analysis Center provides a single interface through which scientists can search multiple databases simultaneously. Explanations of the contents of each of the databases are provided to help determine which are most appropriate for a given search. Searches can be as broad or specific as necessary, depending on researchers' scientific goals.

Experimental Planning

The Analysis Center can be used in the experimental planning phase of the experiment cycle (see Figure 23). Researchers can initiate a query in the Analysis Center based on probe set IDs, sequence, genes, organisms, tissue, and a number of other criteria. Additionally, any or all such criteria can be searched in combination. For instance, it is possible to find enzymes expressed in Rat liver that are found on the Rat Expression Set 230 by performing a Standard Query that is illustrated in Figure 24.

Adding to this initial query, it also possible to then search for human orthologs as illustrated in Figure 25. This is done by clicking on "Show Orthologs" at the top of the results sections (Figure 26).

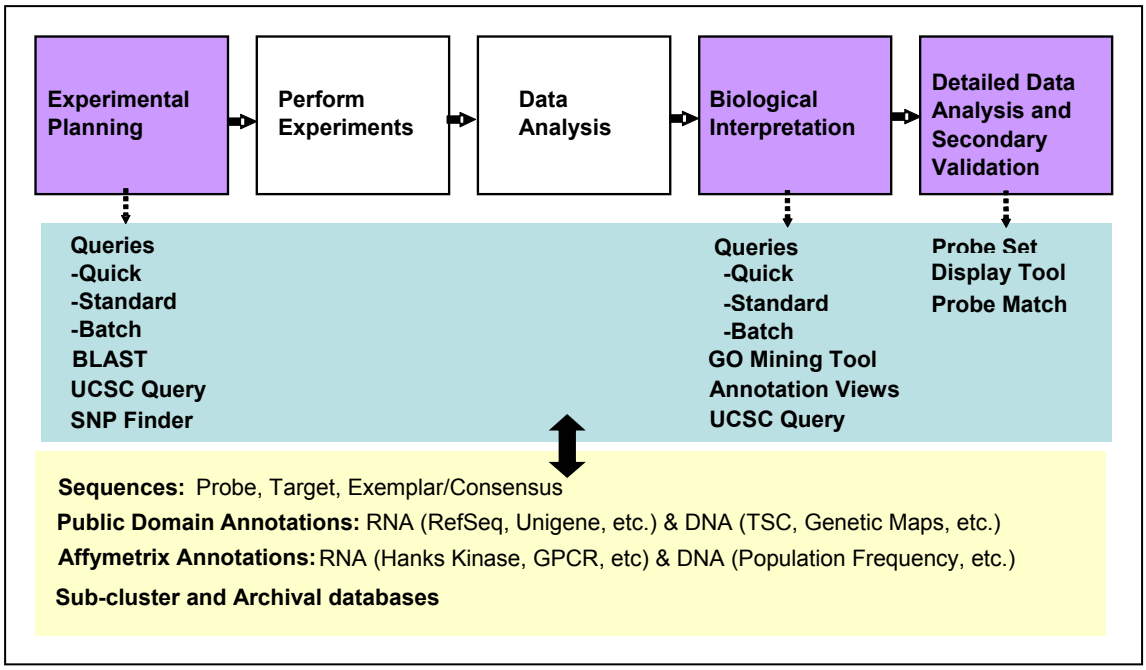


Figure 23. The Experimental Cycle. This diagram shows a schematic of the experimental cycle and aspects of the Analysis Center to use at each step.

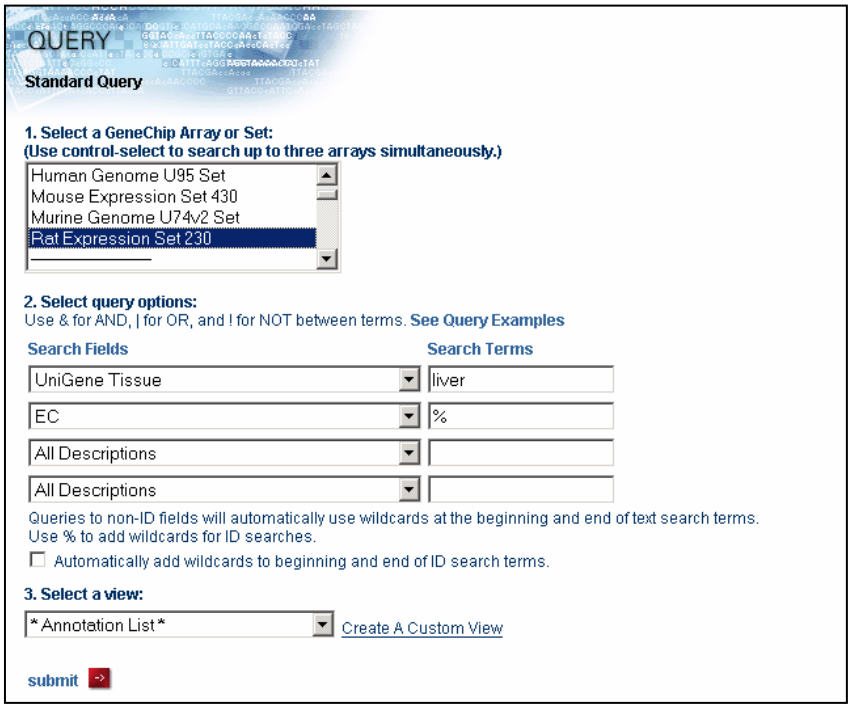
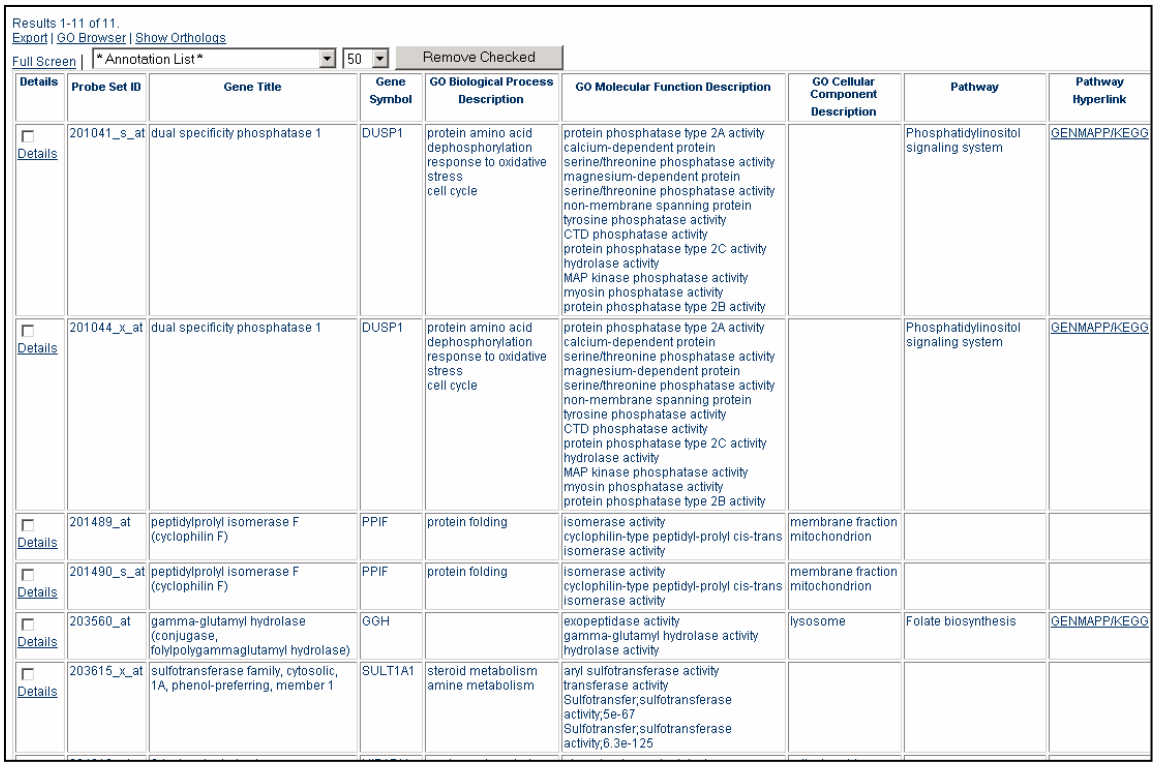


Figure 24. A Standard Query to obtain all enzymes in liver represented on the Rat Expression Set 230.



Results 1-11 of 11.
Export | GO Browser | Show Orthologs
Full Screen | * Annotation List * 50 Remove Checked

Details	Probe Set ID	Gene Title	Gene Symbol	GO Biological Process Description	GO Molecular Function Description	GO Cellular Component Description	Pathway	Pathway Hyperlink
Details	201041_s_at	dual specificity phosphatase 1	DUSP1	protein amino acid dephosphorylation response to oxidative stress cell cycle	protein phosphatase type 2A activity calcium-dependent protein serine/threonine phosphatase activity magnesium-dependent protein serine/threonine phosphatase activity non-membrane spanning protein tyrosine phosphatase activity CTD phosphatase activity protein phosphatase type 2C activity hydrolase activity MAP kinase phosphatase activity myosin phosphatase activity protein phosphatase type 2B activity		Phosphatidylinositol signaling system	GENMAPP/KEGG
Details	201044_x_at	dual specificity phosphatase 1	DUSP1	protein amino acid dephosphorylation response to oxidative stress cell cycle	protein phosphatase type 2A activity calcium-dependent protein serine/threonine phosphatase activity magnesium-dependent protein serine/threonine phosphatase activity non-membrane spanning protein tyrosine phosphatase activity CTD phosphatase activity protein phosphatase type 2C activity hydrolase activity MAP kinase phosphatase activity myosin phosphatase activity protein phosphatase type 2B activity		Phosphatidylinositol signaling system	GENMAPP/KEGG
Details	201489_at	peptidylprolyl isomerase F (cyclophilin F)	PPIF	protein folding	isomerase activity cyclophilin-type peptidyl-prolyl cis-trans isomerase activity	membrane fraction mitochondrion		
Details	201490_s_at	peptidylprolyl isomerase F (cyclophilin F)	PPIF	protein folding	isomerase activity cyclophilin-type peptidyl-prolyl cis-trans isomerase activity	membrane fraction mitochondrion		
Details	203560_at	gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)	GGH		exopeptidase activity gamma-glutamyl hydrolase activity hydrolase activity	lysosome	Folate biosynthesis	GENMAPP/KEGG
Details	203615_x_at	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	SULT1A1	steroid metabolism amine metabolism	aryl sulfotransferase activity transferase activity Sulfotransfer;sulfotransferase activity;5e-67 Sulfotransfer;sulfotransferase activity;6.3e-125			

Figure 25. Results of an ortholog search using the HG-U133 Plus 2.0 Array.

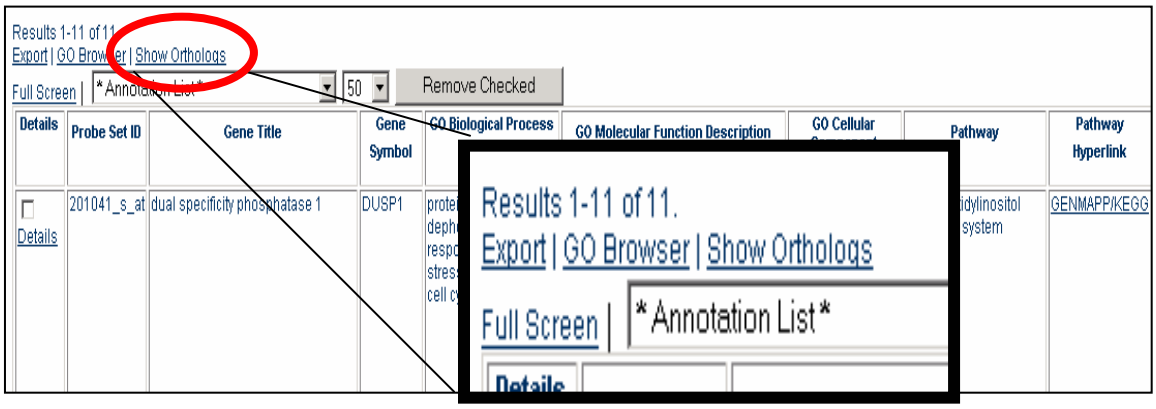


Figure 26. Zoomed-in view of the Show Orthologs link.

Additionally a list of genes or gene names can be queried using the Batch Analysis tool. Before using the Batch Query, it is imperative to format the descriptors to be included in the query. For example, a list of gene accession numbers or names should be supplied in a column of an EXCEL spreadsheet. The spreadsheet should be saved in a Text (Tab delimited) format as shown in Figure 27. Then, the list can be uploaded when conducting the Batch Query.

1001_at
1002_at
1003_at
1004_at
1005_at
1006_at
1007_at
1008_at
1009_at
1010_at
1011_s_at

Figure 27. A list of probe sets saved in Text (Tab delimited format) for Batch Query. Note that no other aspect of the data are present.

Results from a query can be customized using a predefined view. A view is created by the following steps (see Figure 28):

1. Click on the "Annotation Views" link.
2. Choose the fields of interest in the "Selected Fields" window.
3. Provide a name for the view.
4. Save the view.

The new view can be used to visualize any queries, past or present.

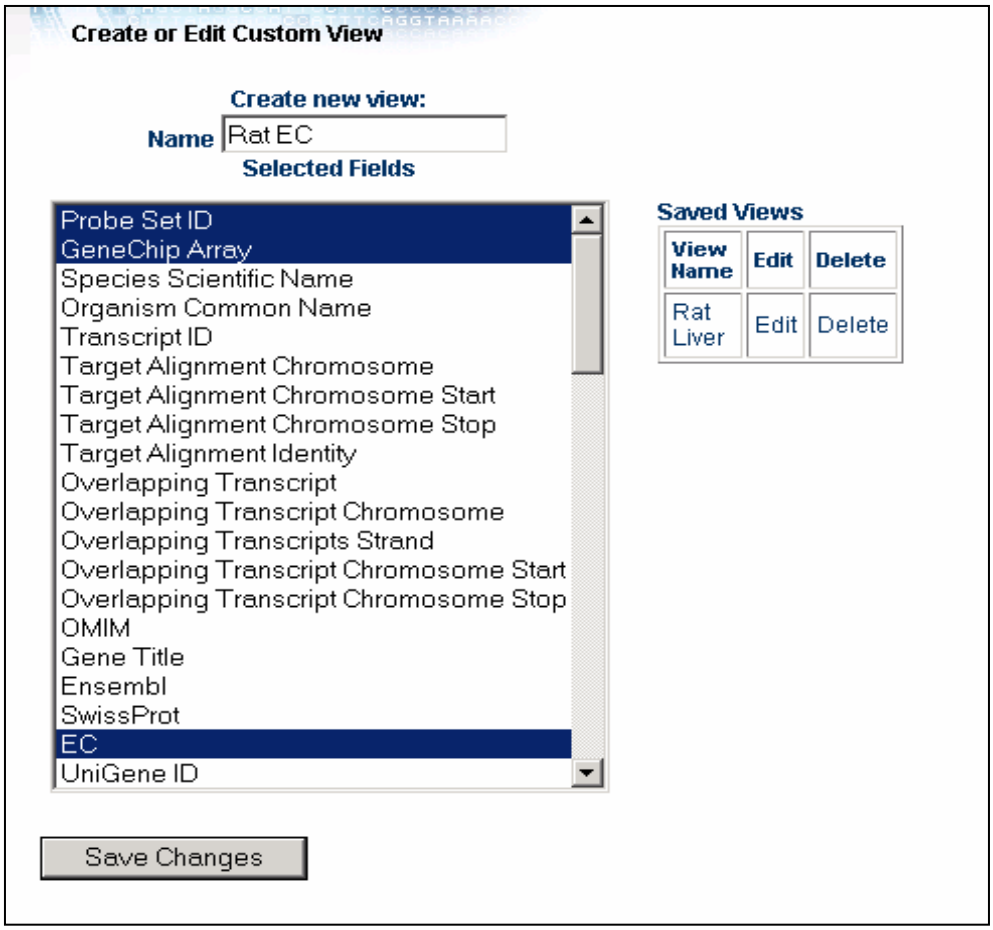


Figure 28. Annotation View Expression page used to construct a customized view.

Results from any query are easily downloaded by using the Export command (see Figure 29). The results are saved as a .tsv file which can be opened in EXCEL. It is important to note that results that are downloaded from a query or GO Mining Tool are limited to a 3000 probe set maximum.

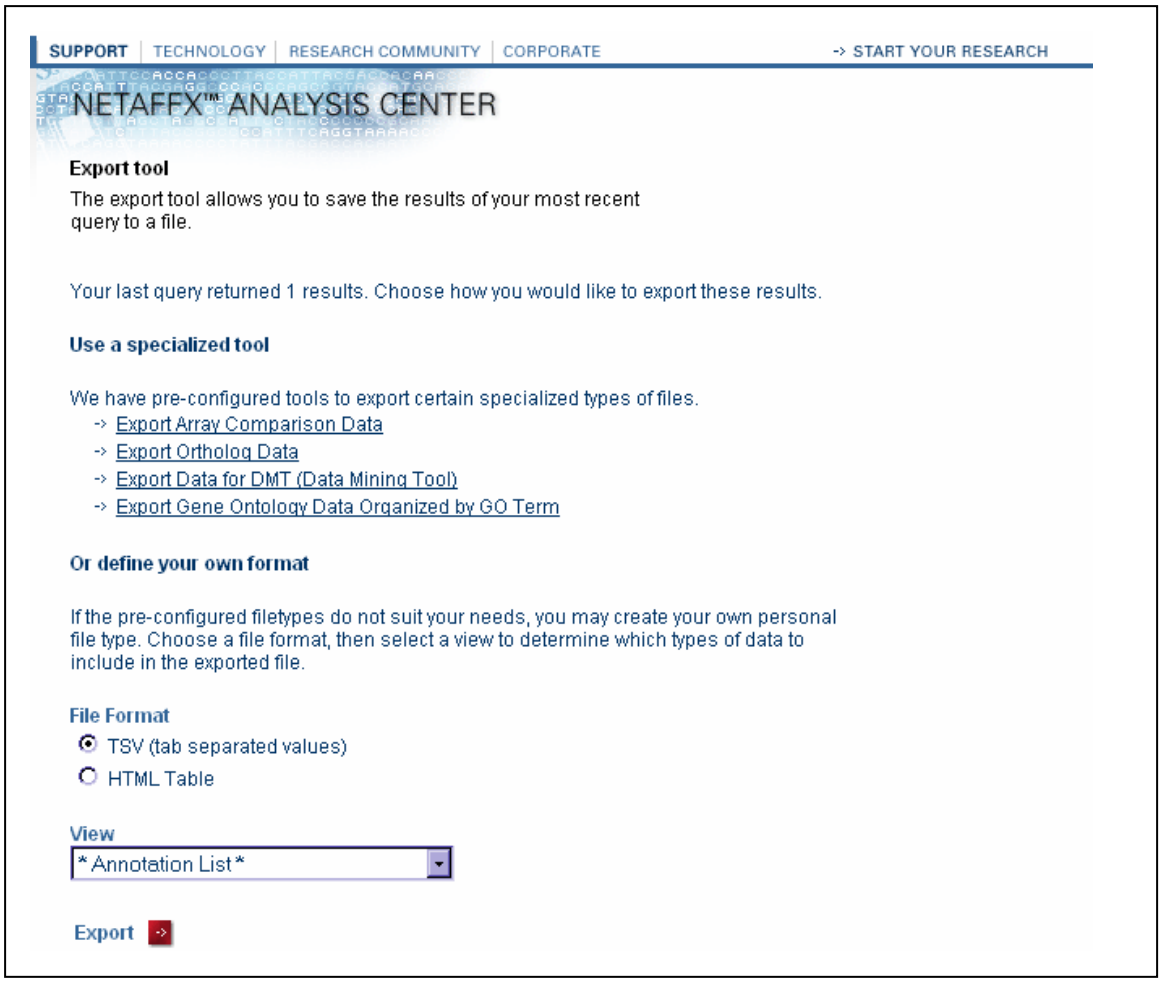


Figure 29. Example of the Export window.

Biological Interpretation

Biological interpretation of results is typically executed once a final gene list is determined from the initial first order analysis. Specific probe sets or groups of probe sets can be searched using Quick Query or Batch Query, respectively.

To search for individual probe sets, enter the entire probe set ID, complete with suffix (e.g., 1001_at), in a Quick query. If 1001 is entered, the results may contain any probe set possessing the moniker 1001 in any of its accompanying accession numbers.

Interrogation of multiple probe sets is done through the Analysis Center Batch Query Tool as discussed above. A list of probe set IDs alone should be supplied in a column of an EXCEL spreadsheet. The spreadsheet should be saved in a Text (Tab delimited) format as described in the previous section. The list can then be uploaded into Batch Query.

The Gene Ontology (GO) Mining Tool can also be used to effectively illustrate results from a query. After obtaining results that are 3000 probe sets or less from a Quick, Standard, or Batch Query, the GO Mining Tool option is made available above the table of results. Employing the GO Mining Tool option will illustrate the data as a multi-branched graph.

The branches of the graph connect similar GO functions. Figure 30 shows a GO Mining Tool graph.

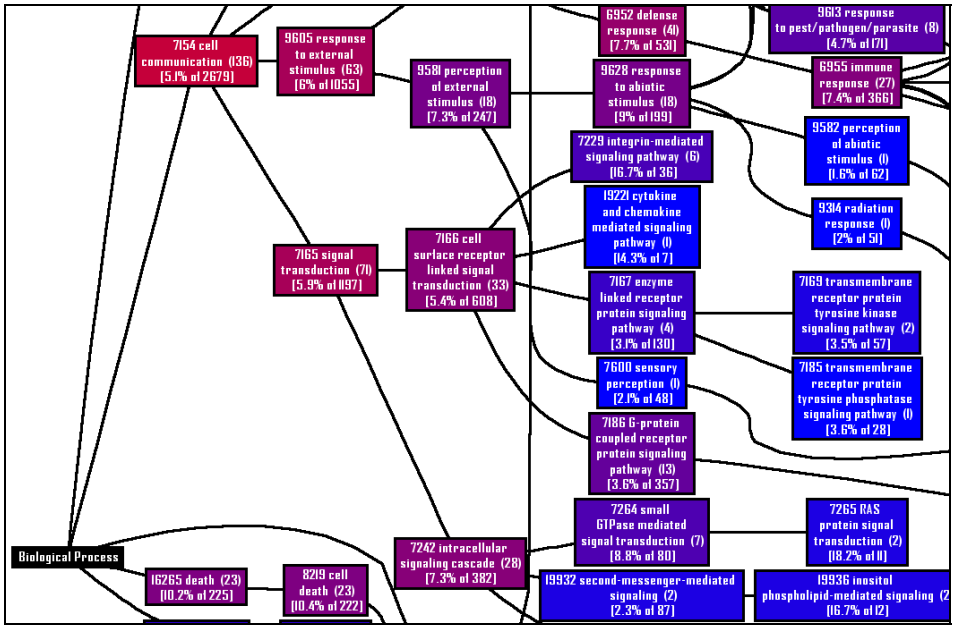


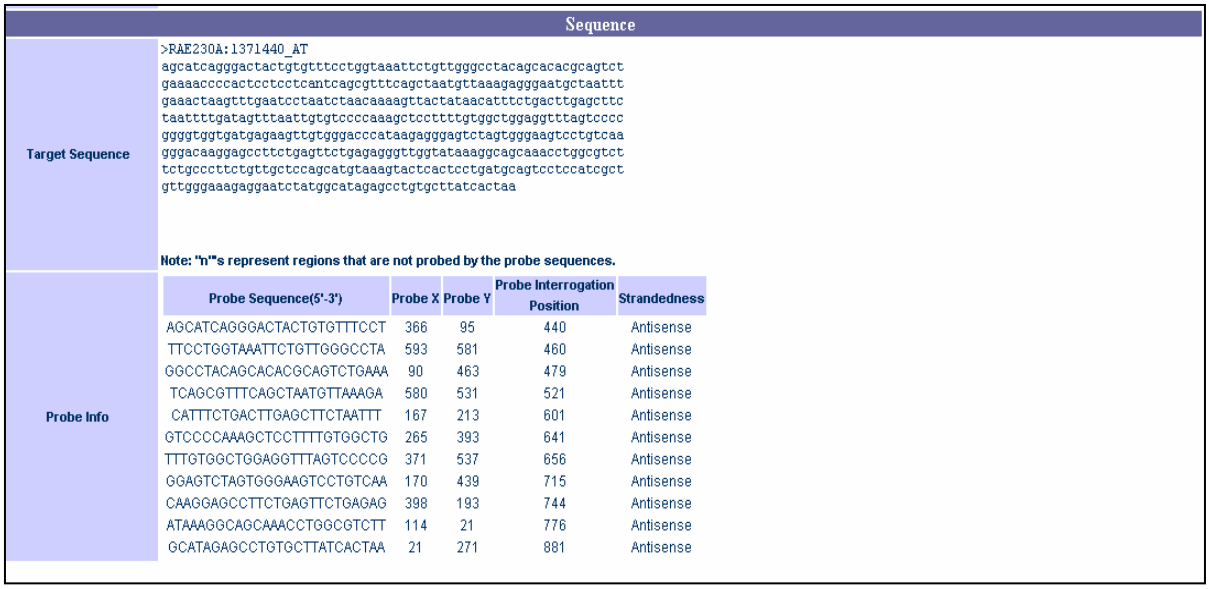
Figure 30. Query results further illustrated using the Gene Ontology Mining Tool. Members of a tree node as well as trees originating at a given node can be obtained using this graphical aid also.

Queries are saved, combined, or deleted using the Query History part of the Analysis Center. Once a query is performed it is saved automatically and the five most recent queries are bookmarked on the initial Getting Started page.

Detailed Data Analysis and Secondary Validation

The Analysis Center is also useful for detailed data analysis of GeneChip microarray data results. By doing detailed data analysis it is possible to design a more refined, suitable filter for a specific data set. For example, data can be filtered in GCOS using arbitrary criteria such as a Signal Log Ratio of 1.0 or greater. Those results can be uploaded into the Analysis Center using Batch Query. The query, for example, may return information that highlights a functional group, such as nucleotide metabolism. Upon further investigation, it may be noted that most of the genes associated with nucleotide metabolism are upregulated. This may lead to subsequent questions about other genes involved in nucleotide metabolism that were not present on the original list. A Quick, Standard, or Batch Query can be used to determine which probe sets represent other nucleotide metabolism genes and, consequently, determine whether those genes were also up regulated, albeit slightly less than a Signal Log Ratio of 1.0. Based on this detailed data analysis it may be necessary to relax the stringency of the quantitative analysis and then reevaluate results.

The Analysis Center also serves as a valuable resource when planning further experiments to verify array results. Technologies such as quantitative PCR are often used for this purpose. If quantitative PCR is used, then it is recommended to select the primers in the same sequence region that was used for the probe array design. Designing the probe based on the Target Sequence provided assures that the same transcript variant is assayed in both technologies. The target sequence can be obtained from the Full Record of a given probe set as shown below in Figure 31:



Sequence

Target Sequence

>RAE230A:1371440_AT
agcatcagggactactgtgtttcctggtaaattctgttgggcctacagcacacgcagtct
gaaaaccccactcctcctcantcagcgtttcagctaatgttaaagagggaatgctaattt
gaaactaagtttgaatcctaatctaacaaaagttactatataacatttctgacttgagcttc
taattttgatagtttaattgtgtccccaaagctccttttgtggctggaggtttagtcccc
ggggtggtgatgagaagttgtgggacccataagagggagtctagtgggaagtcctgtcaa
gggacaaggagccttctgagttctgagagggttggtataaaggcagcaaacctggcgtct
tctgcccttctgttgctccagcatgtaaagtactcactcctgatgcagtcctccatcgct
gttgggaaagaggaatctatggcatagagcctgtgcttatcactaa

Note: "n"s represent regions that are not probed by the probe sequences.

Probe Info

Probe Sequence(5'-3')	Probe X	Probe Y	Probe Interrogation Position	Strandedness
AGCATCAGGGACTACTGTGTTTCCT	366	95	440	Antisense
TTCCTGGTAAATTCTGTTGGGCCTA	593	581	460	Antisense
GGCCTACAGCACACGCAGTCTGAAA	90	463	479	Antisense
TCAGCGTTTCAGCTAATGTTAAAGA	580	531	521	Antisense
CATTTCTGACTTGAGCTTCTAATTT	167	213	601	Antisense
GTCCCCAAAGCTCCTTTTGTGGCTG	265	393	641	Antisense
TTTGTGGCTGGAGGTTTAGTCCCCG	371	537	656	Antisense
GGAGTCTAGTGGGAAGTCCTGTCAA	170	439	715	Antisense
CAAGGAGCCTTCTGAGTTCTGAGAG	398	193	744	Antisense
ATAAAGGCAGCAAACCTGGCGTCTT	114	21	776	Antisense
GCATAGAGCCTGTGCTTATCACTAA	21	271	881	Antisense

Figure 31. Target and probe sequences of probe set 1371440_at as depicted at the bottom of the Full record. The Full record of a probe set is obtained by choosing the Details option for a probe set.

Pathway Analysis and Modeling

For most biologists using expression analysis in functional genomics studies, the desired analysis end point is a new or an improved pathway model of the biological process of interest. Pathway pictures are a useful and powerful way to summarize the network of molecular interactions that make up the biological process under study. Placing global gene expression data within the context of a pathway image enables you to identify affected pathways.

One highly useful tool for biological pathway analysis of expression data, GenMAPP, was developed by the Conklin Laboratory at UCSF (http://www.genmapp.org). GenMAPP, which is a freeware program, and its associated files, can be downloaded from the web site. GenMAPP allows the application of quantitative data, such as, but not limited to, expression signals, as colors to pathway elements on a pathway map. In most cases, the pathway images must be developed based on prior knowledge (GenMAPP has basic drawing functions). Gene Ontology groups and data from KEGG have been used to generate MAPP files for human and mouse, but a considerable amount of manual work may be required to develop

an image of a specific pathway of interest. The premise behind its development is to provide a tool that helps expand and refine pathway knowledge through mapping of gene expression data. As more scientists use it and publish better MAPP files, it becomes more useful and powerful as an analysis tool. Figure 32 shows an example of pathway images with gene expression data overlaid as colors (cell cycle controls in maturing neutrophils). Red indicates an increase while blue indicates a decrease; the median Signal Log Ratio (base 2) is shown to the right of each box. Note the step-wise change between day 4 and day 6, including downstream genes from p53 and CDK 4, 6.

In addition to GenMAPP, some sophisticated commercial products are emerging, such as Visual Cell from Gene Network Sciences (www.gnsbiotech.com), which provides an environment for creating large and highly complex graphic representations of cellular and molecular processes, and the integration of a variety of biological data, including RNA expression.

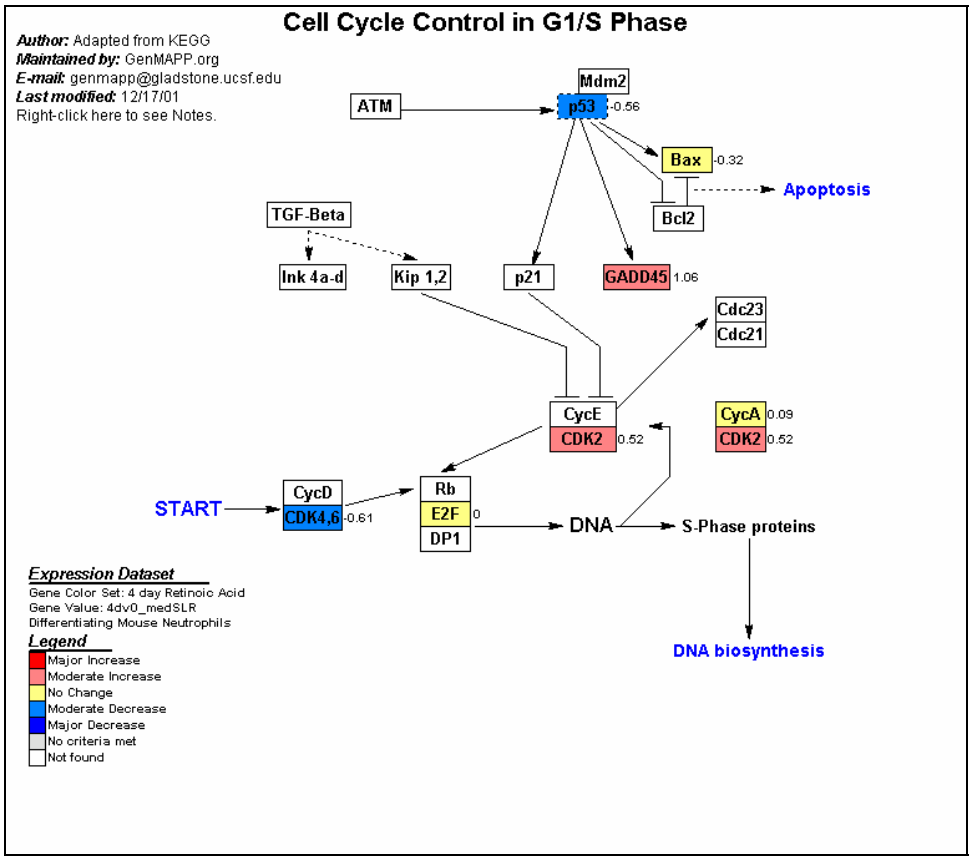


Figure 32. RNA Levels of Cell Cycle Genes Visualized with GenMAPP (maturing neutrophils).

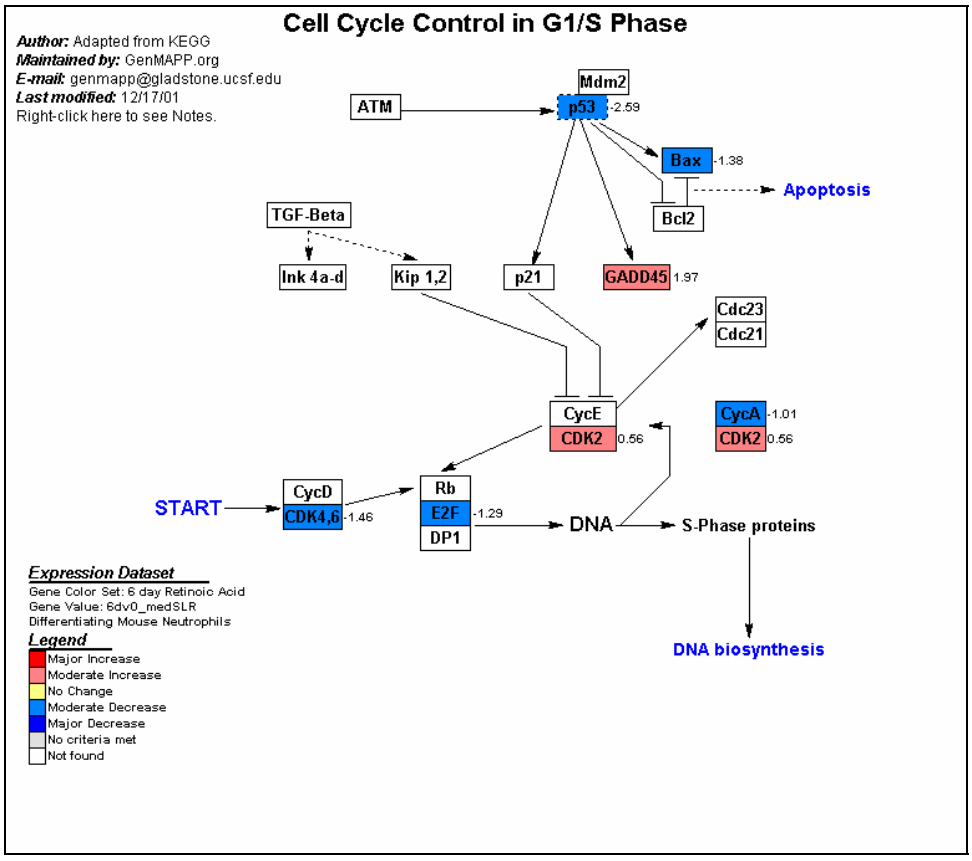


Figure 33. **Cell Cycle Control in G1/S Phase. Example of output from GenMAPP: two images of the G1/S phase transition cell cycle control pathway in maturing neutrophils at 4 days, then 6 days after treatment with retinoic acid. Note the small down regulation in p53 and CDK 4, 6 at 4 days, followed by greater change at 6 days, as well as downstream down regulation of Bax and E2F.**

Analysis of Promoter Sequences of Regulated Transcripts

Another area of follow-up is to analyze promoter sequences of regulated transcripts to identify elements that may be involved in transcriptional regulation. There are two primary directions that can be taken. The first involves looking for previously characterized elements in the promoters of transcripts that appear to be regulated by a known transcription factor. For example, Figure 33 shows a series of transcripts that appear to show an immediate-early response profile to stimulation with retinoic acid (through a nuclear receptor) in developing mouse neutrophils. The promoters of these transcripts may contain a Retinoic Acid Response Element (RARE). Indeed the gene MAD is known to contain a RARE in its promoter and shows a typical immediate-early response to the signal. Known sequence motifs can be located in promoter sequences using a variety of sequence search and alignment tools, such as those provided by the NCBI. Promoter sequences for many well-characterized genes are available in public databases, such as GenBank and RefSeq, and their annotations continue to improve with time. In addition, the growing Eukaryotic Promoter

Database (EPD) is a dedicated, though currently limited, resource on promoter information of eukaryotes (http://www.epd.isb-sib.ch).

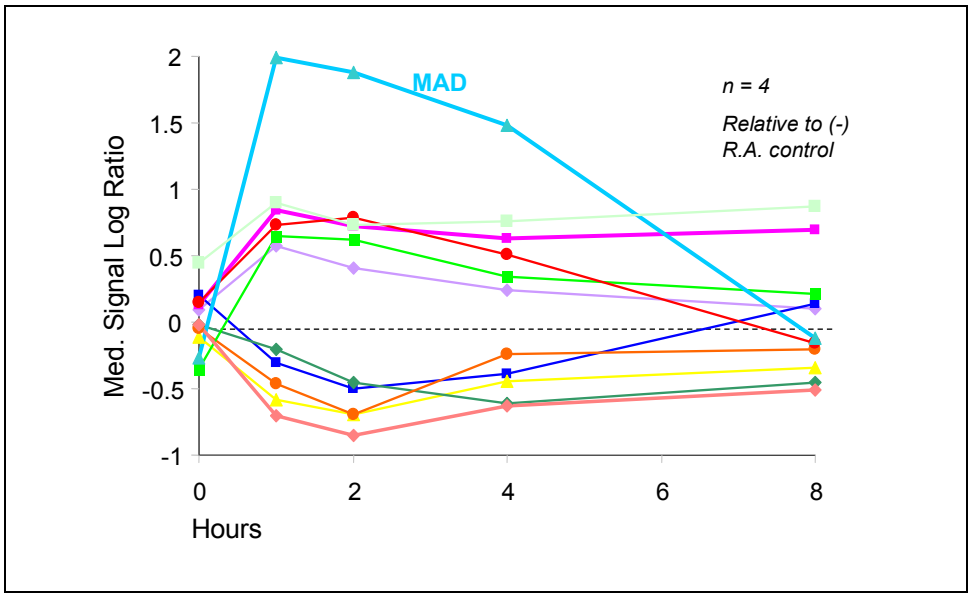


Figure 34. Transcripts Showing 'Immediate–Early' Change in Response to Retinoic Acid

Transcripts showing 'Immediate Early' change in response to retinoic acid (positive Median Signal Log Ratios indicate an increase relative to control; negative indicate a decrease. Signal Log Ratio is base 2). MAD, produced by a gene known to have a Retinoic Acid Response Element (RARE) in its promoter, serves as a positive control. The remaining 10 transcripts may have such elements, a hypothesis that can be tested by searching the promoters of these transcripts.

The second direction of follow-up on promoter sequences is searching for novel motifs in the promoters of transcripts that appear to be co-regulated (13). This may be a more challenging task, especially if the putative transcription factor is unknown. This direction of research will continue to expand as more is understood about transcription factors and their binding sites on gene promoters. In addition to the EPD, a useful resource is TRANSFAC: the Transcription Factor Database (http://transfac.gbf.de/TRANSFAC/).

Appendix A: Glossary

- MAS 4.0-Specific Terms (Empirical Algorithms)

- MAS 5.0-Specific Terms (Statistical Algorithms)

- Absolute Analysis: The qualitative analysis of a single array to determine if a transcript is Present, Absent, or Marginal.

- Array: A collection of probes on glass encased in a plastic cartridge.

- Average Difference: A quantitative relative indicator of the level of expression of a transcript ($\sum$(PM-MM)/pairs in the average).

- Background: A measurement of signal intensity caused by auto-fluorescence of array surface and non-specific binding of target/stain molecules (SAPE).

- Baseline Array: An array used for normalization purposes during comparison analysis. Also see Comparison Analysis.

- Biweight Estimate: An estimate of the central value of a sample used by the Affymetrix® Statistical Algorithms.

- Change: A qualitative call indicating an Increase (I), Marginal Increase (MI), No Change (NC), Marginal Decrease (MD), or Decrease (D) in transcript level between a baseline array and an experiment array.

- Change *p*-value: A *p*-value indicating the significance of the Change call. The change *p*-value measures the probability that the expression levels of a given probe set differ in two different arrays when the *p*-value is close to 0.5, they are likely to be the same. When the *p*-value is close to 0, the expression level in the experiment array is higher than that of the baseline array. When the *p*-value is close to 1, the expression level in the experiment arrays is lower than that of the baseline.

- Chip: See Array.

- Comparison Analysis: The analysis of an experimental array compared to a baseline array.

- Decision Matrix: An algorithm that examines a collection of metrics used to determine the status of a hybridized transcript.

- Detection: A qualitative measurement indicating if a given transcript is detected (Present), not detected (Absent), or marginally detected (Marginal).

- Detection *p*-value: A *p*-value indicating the significance of the Detection call. A Detection *p*-value measures the probability that the discrimination scores of all probe pairs in the probe set are above a certain level (Tau), and that the target is likely to be Present.

- Discrimination Score [R]: The relative difference between a Perfect Match and its Mismatch (R=(PM-MM)/(PM+MM)).

- Empirical Algorithms: The algorithms utilized by GeneChip® Analysis Suite and Microarray Suite 4.0 based on empirical data generated by Affymetrix.

- Experimental Array: An array that is used in comparison analysis to be compared against a baseline array to detect changes in expression.

- Feature: A single square-shaped probe cell on an array (another term for probe cell). A feature ranges in size from 8 to 50 microns depending on the array type.

- Hybridization Controls: Controls added to the sample before hybridization to the array (refer to Chapter 1 for more information).

- Idealized Mismatch: A value used in place of the Mismatch intensity when Rules 2 and 3 are used in the Signal Algorithm (refer to Chapter 2 for more information on Rules in the Statistical Algorithms).

- Latin Square: An experimental design used to monitor the ability to detect a transcript accurately over a range of concentrations. It also allows the statistical analysis of patterns and variability in repeated measurements in a systematic fashion.

- Mask: Filter used during synthesis of a GeneChip® array that exposes discreet areas of a wafer to ultraviolet light.

- Metric: The calculated answer of mathematical equations used by the GeneChip® algorithms.

- Mismatch Probe (MM): A 25-mer oligonucleotide designed to be complementary to a reference sequence except for a single, homomeric (nucleotide mismatch that contains the complementary base to the original) base change at the 13th position. Mismatch probes serve as specificity controls when compared to their corresponding Perfect Match probes.

- Noise: The result of small variations in digitized signals in the scanner as it samples the probe array surface and is measured by examining the pixel-to-pixel variations in signal intensities.

- Non-parametric Test: A statistical test without the assumption of a particular distribution of the data, also known as a distribution-free test.

- Normalization: Adjusting an average value of an experimental array equal to that of the baseline array so that the arrays can be compared (refer to Algorithms description for more information).

- *p*-value: The probability that a certain statistic is equal or more extreme to the observed value when the null hypothesis is true. The null hypothesis is that the two samples are the same.

- Parametric Test: A statistical test that assumes the data are sampled from a population following a Gaussian or normal distribution.

- Perfect Match Probe (PM): A 25-mer oligonucleotide designed to be complementary to a reference sequence. The probe sequence that is complementary to the sequence to be hybridized.

- Perturbation: The range by which the normalization factor is adjusted up or down by the user.

- Photolithography: The process used to manufacture probe arrays in conjunction with combinatorial chemistry through a series of cycles. Using light, photolabile protecting groups are removed from linkers bound to the glass substrate (wafer) to enable nucleoside phosphoramidite addition in specific deprotected locations. Each light exposure and subsequent phosphoramidite addition is equal to one cycle. Typically, probe arrays are synthesized in about 80 cycles.

- Probe: A 25-mer oligonucleotide synthesized *in situ* on the surface of the array using photolithography and combinatorial chemistry. Hybridization to probes provides intensity data used in both Empirical and Statistical algorithms.

- Probe Array Tiling: The spatial organization of probe array features into probe pairs and sets.

- Probe Cell: A single square-shaped feature on an array containing probes with a unique sequence. A probe cell ranges in size from 18 to 50 microns per side depending on the array type (refer to Figure 35).

- Probe Pair: Two features within a probe set (refer to Figure 35). Each probe of a probe pair is designed to differ only at the nucleotide base interrogation position. The probe pair is designed to detect a Perfect Match (PM) and a Mismatch (MM).

- Probe Set: A collection of probe pairs which interrogates the same sequence, or set of sequences. A probe set typically contains between 11 to 20 probe pairs (refer to Figure 35).

- SAPE: Streptavidin-phycoerythin dye used to bind the biotin. In the GeneChip® Expression Assay, the biotinylated nucleotides are incorporated into the cRNA during the *in vitro* transcription (IVT) reaction.

- Scaling: Adjusting the average intensity or signal value of every array to a common value (target intensity) in order to make the arrays comparable.

- Signal: A quantitative measure of the relative abundance of a transcript.

- Signal Log Ratio: The change in expression level for a transcript between a baseline and an experiment array. This change is expressed as the $\log_2$ ratio. A Signal Log Ratio of 1 is the same as a Fold Change of 2.

- Signal Log Ratio High: The upper limit of the Signal Log Ratio within a 95% confidence interval.

- Signal Log Ratio Low: The lower limit of the Signal Log Ratio within a 95% confidence interval.

- Single Array Analysis: See Absolute Analysis.

- Spike Controls: Controls that are added to the sample before cDNA synthesis (refer to Chapter 1 for more information).

- Stat Pairs: The number of probe pairs in the probe set.

- Stat Common Pairs: The number of common probe pairs on two arrays (experiment versus baseline) after saturation across the probe set is determined.

- Stat Pairs Used: The number of probe pairs in the probe set used in the Detection call.

- Statistical Algorithms: The algorithms contained in Microarray Suite Version 5.X and GCOS 1.X. This algorithm was developed using standard statistical methods.

- Tau: A user-definable threshold used to determine the detection call.

- Target: The sample applied as labeled (biotinylated), fragmented cRNA to a GeneChip® probe array for hybridization.

- Wafer: The glass substrate onto which probes are synthesized during the manufacturing of probe arrays.

- Wilcoxon's Signed Rank Test: A non-parametric pair-wise comparison test. This test is used to determine the Detection and Change calls for analysis.

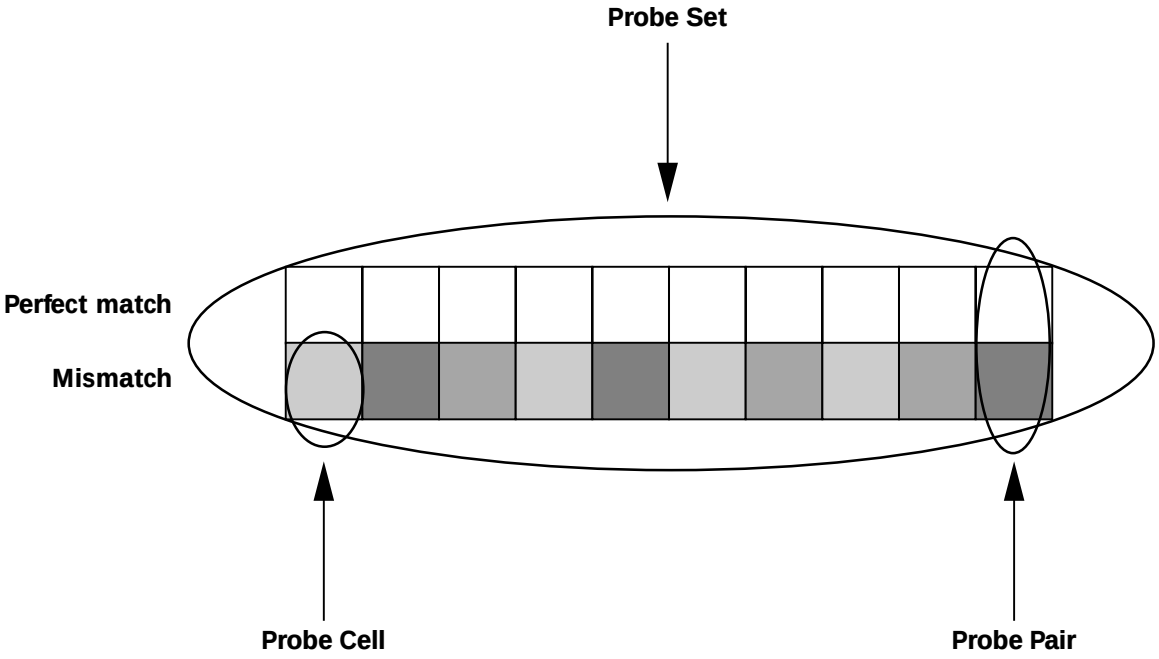


Figure 35. Image of a probe set, which includes 10 probe pairs.

Appendix B: GeneChip® Probe Array Probe Set Name Designations

In addition to the _at ("antisense target") and _st ("sense target") probe set name designations, there are other designations that reflect special characteristics of a particular probe set based on probe design and selection criteria. These designations are listed below.

Probe Set Name Designations Prior to HG-U133 Set:

_f_at (sequence family):

Probe set that corresponds to sequences for which it was not possible to pick a full set of 16-20 unique and/or shared similarity-constrained probes. Some probes in this set are similar (e.g., polymorphic) but not necessarily identical to other gene sequences. Some family members overlap a portion of the probe set. Family members can be singleton or an Affymetrix designated group of sequences.

--- --- --- --- ---	12345_f_at probes
----------	transcript #1
----------	transcript #2
----------	transcript #3
-X------X------	transcript #4 (w/polymorphisms)

_s_at (similarity constraint):

Probe set that corresponds to a small number of unique genes (<5%) that share identical sequence. Probes were chosen from the region that is common to these genes. Group members can be singleton or a group of sequences. For _s probe sets, there is not enough unique sequence to design a separate _at probe set.

--- --- --- ---	23456_s_at probes
----------	transcript #5
----------	transcript #6
----------	transcript #7

_g_at (common groups):

Probes chosen in region of overlap. To differentiate from an _s group, the sequences are represented as singletons (_at probe sets either have the same probe set ID number or the preceding probe set ID number) on the same probe array as well. In other words, for _g probe sets, there is enough unique sequence to design a separate _at probe set.

--- --- ---	34567_at probes
----------	transcript #8

_r_at (rules dropped):

Designates sequences for which it was not possible to pick a full set of unique probes using Affymetrix' probe selection rules. Probes were picked after dropping some of the selection rules.

_i_at (incomplete):

Designates sequences for which there are fewer than the required numbers of unique probes specified in the design.

_b_at (ambiguous probe set):

All probe selection rules were ignored. Withdrawn from GenBank.

_l_at (long probe set):

Sequence represented by more than 20 probe pairs.

Probe Set Name Designations for HG-U133 Set and HG-U133A 2.0

These are the only probe set extensions used in these designs

_s_at:

Designates probe sets that share all probes identically with two or more sequences. The _s probe sets can represent shorter forms of alternatively polyadenylated transcripts, common regions in the 3' ends of multiple alternative splice forms, or highly similar transcripts. Approximately 90% of the _s probe sets represent splice variants. Some transcripts will also be represented by unique _at probe sets.

_x_at:

Designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridization were dropped in order to design the _x probe sets. These probe sets share some probes identically with two or more sequences and therefore, these probe sets may cross-hybridize in an unpredictable manner.

Probe Set Name Designations for HG-U133 Plus 2.0

These are the only probe set extensions used in the HG-U133 Plus 2.0 Array

Original content

_s_at:

Designates probe sets that share all probes identically with two or more sequences. The _s probe sets can represent shorter forms of alternatively polyadenylated transcripts, common regions in the 3' ends of multiple alternative splice forms, or highly similar transcripts. Approximately 90% of the _s probe sets represent splice variants. Some transcripts will also be represented by unique _at probe sets.

_x_at:

Designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridization were dropped in order to design the _x probe sets. These probe sets share some probes identically with two or more sequences and therefore, these probe sets may cross-hybridize in an unpredictable manner.

"Plus" content

_a_at:

Designates probe sets that recognize alternative transcripts from the same gene (a subset of the _s probe sets as described under HG-U133 Set).

_s_at:

Designates probe sets with common probes among multiple transcripts from different genes.

_x_at:

Designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridization were dropped in order to design the _x probe sets. These probe sets share some probes identically with two or more sequences and, therefore, these probe sets may cross-hybridize in an unpredictable manner.

At the time of the HG-U133 Set design the "_a" probe set was not defined. The non-unique probe set type, "_a", was introduced with the Mouse Expression Set 430 to indicate probe sets that recognize alternative transcripts from the same gene. Probe sets with common probes among multiple transcripts from separate genes are annotated with the "_s" suffix. For consistency, the names of existing probe sets with the "_s" suffix were not changed between the HG-U133 Set and the HG-U133 Plus 2.0 and HG-U133A 2.0 Arrays. The new (Plus) content on the HG-U133 Plus 2.0 Array incorporates both the "_a" and "_s" probe sets.

Probe Set Name Designations for Mouse Set 430, Mouse 430 2.0 Arrays, Rat Set 230, and Rat 230 2.0 Array

These are the only probe set extensions used in these designs

_a_at:

Designates probe sets that recognize alternative transcripts from the same gene.

_s_at:

Designates probe sets with common probes among multiple transcripts from different genes.

_x_at:

Designates probe sets where it was not possible to select either a unique probe set or a probe set with identical probes among multiple transcripts. Rules for cross-hybridization were dropped in order to design the _x probe sets. These probe sets share some probes identically with two or more sequences and, therefore, these probe sets may cross-hybridize in an unpredictable manner.

Appendix C: Expression Default Settings

GCOS 1.0 Expression Analysis Default Settings

Default Parameter	16-20 probe pairs/probe set	11-15 probe pairs/probe set	
		18µm feature size	11µm feature size
Alpha1	0.04	0.05	0.05
Alpha2	0.06	0.065	0.065
Tau	0.015	0.015	0.015
Gamma1L	0.0025	0.0045	0.002
Gamma1H	0.0025	0.0045	0.002
Gamma2L	0.003	0.006	0.002667
Gamma2H	0.003	0.006	0.002667
Perturbation	1.1	1.1	1.1

MAS 5.0 Expression Analysis Default Settings

Default Parameter	# probe pairs/probe set	
	16-20	11-15
Alpha1	0.04	0.05
Alpha2	0.06	0.065
Tau	0.015	0.015
Gamma1L	0.0025	0.0045
Gamma1H	0.0025	0.0045
Gamma2L	0.003	0.006
Gamma2H	0.003	0.006
Perturbation	1.1	1.1

Appendix D: Change Calculation Worksheet

This procedure can be used to identify robust changes between two GeneChip® probe arrays. These instructions relate to analyses performed in GeneChip® Operating Software.

If the samples hybridized to the two arrays are derived from separate samples, this procedure will identify probe sets showing significant change and serves as a useful starting point for further data analysis. If the two samples are derived from the same hybridization cocktail, this procedure will identify false changes. According to the Affymetrix specification, the false change observed should be no more than 2%. This value is based on observations reported by Wodicka *et al.* in 1997 (15).

Data Preparation

1. Choose the two data sets that you wish to analyze.
2. Conduct a single array analysis of the baseline data set as described in Chapter 3 of this manual.
3. Conduct a comparison analysis of the experiment data set using the previous data set as the baseline as described in Chapter 4 of this manual. Ensure that the scaling strategy used in step 2 is also used in step 3.
4. Record the file names of the baseline and experiment in the appropriate spaces on the Change Calculation Worksheet (see page 101).

Calculate Increases

The first step of this procedure is to calculate the number of significant increases.

1 Calculate the number of probe sets that have a Detection call of 'P' in the Experiment file.

1.1 Open the comparison .chp file in GCOS, with the Pivot table view.

1.2 Display additional Pivot table columns in the analysis by selecting "Pivot Data>Absolute Results" from the "View" pull-down menu. Ensure that the Detection, Change and Signal Log Ratio Columns are displayed.

1.3 Sort the data on the Detection column in descending order by right-clicking on the Detection column heading and selecting "Sort Descending" from the pop-up menu as shown in Figure 36.

1.4 Click on the probe set identifier, contained in the far-left column, at the top of the list.

1.5 Use the mouse to scroll down the data list until the last 'P' is visible.

1.6 Hold down the 'Shift' key and click on the probe set identifier corresponding to the last 'P' value.

1.7 Click the "Hide unselected probe sets" button as shown in Figure 37.

1.8 The number of remaining probe sets is displayed in the bottom-right of the window, as shown in Figure 38. Enter this value into the box on Line 1 of the Change Calculation Worksheet.

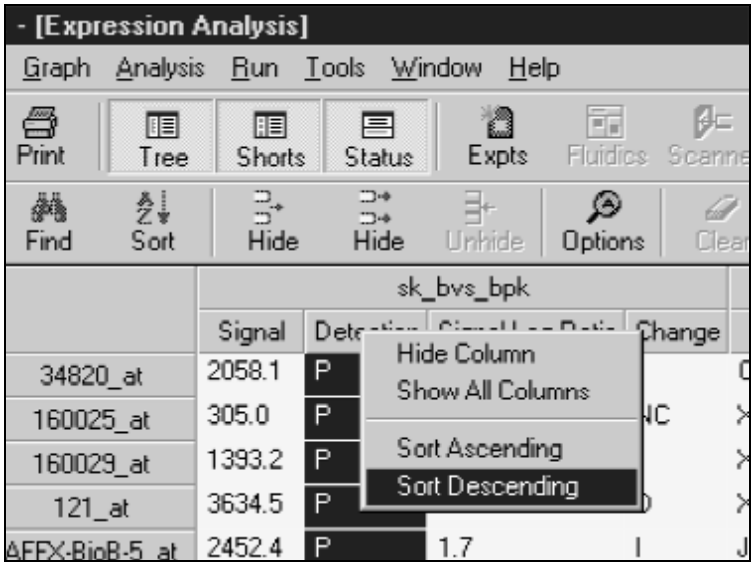


Figure 36.

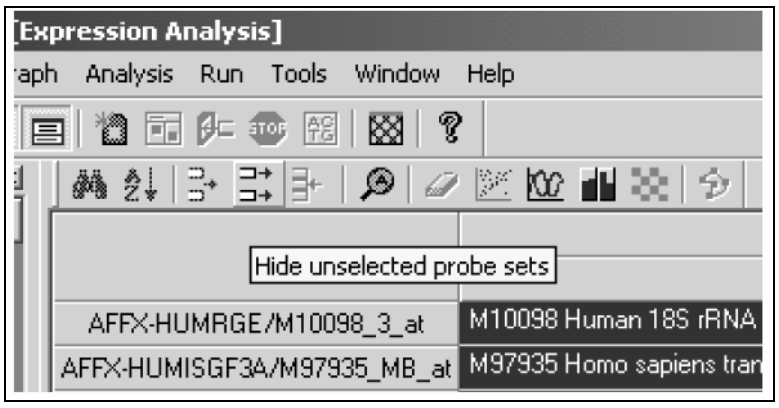


Figure 37.

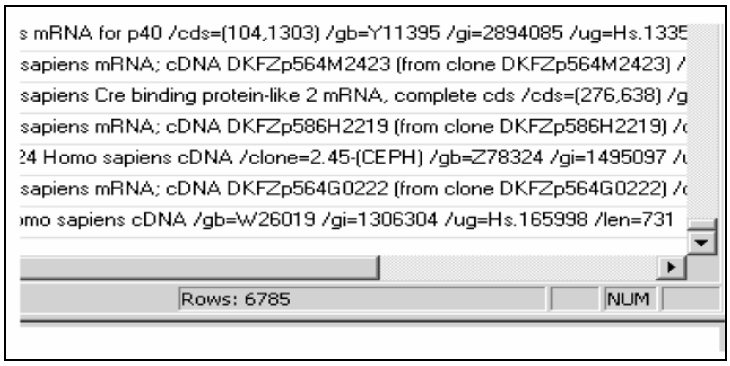


Figure 38.

2 Calculate the number of probe sets from above list that also have a Change call of 'I.'

2.1 After performing step 1 of the Increase calculation, sort the data on the Change column in ascending order, by right-clicking the Change column heading and selecting "Sort Ascending" from the pop-up menu as shown in Figure 36.

2.2 Scroll down the list of probe sets until the first 'I' call is visible, then click on this probe set identifier.

2.3 Scroll down the list until the last 'I' call is visible, hold down the 'Shift' key and click on the corresponding probe set identifier.

2.4 Click the "Hide unselected probe sets" button as shown in Figure 37.

2.5 The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 38. Enter this value into the box on Line 2 of the Change Calculation Worksheet.

3 Calculate the number of probe sets from the above list that also have a Signal Log Ratio of 1.0 or greater.

3.1 After performing step 2 of the Increase calculation, sort the data on the Signal Log Ratio column in descending order by right-clicking the Signal Log Ratio column heading and selecting "Sort Descending" from the pop-up menu as shown in Figure 36.

3.2 Click on the probe set identifier at the top of the list.

3.3 Scroll down the list until the last Signal Log Ratio value (equal to 1.0) is visible, hold down the 'Shift' key and click on the corresponding probe set identifier.

3.4 Click the "Hide unselected probe sets" button as shown in Figure 37.

3.5 The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 38. Enter this value into the box on Line 3 of the Change Calculation Worksheet.

4 Calculate the number of probe sets that have increased as a percentage of the probe sets detected.

4.1 Divide the number of probe sets showing significant increase (Line 3) by the number of probe sets detected (Line 1).

4.2 Multiply the above number by 100 to convert to a percentage.

4.3 Enter the value in the box on Line 4 of the Change Calculation Worksheet.

Calculate Decreases

The next part of this procedure is to calculate the number of significant decreases.

1 Calculate the number of probe sets that have a Detection call of 'P' in the Baseline file.

1.1 Open both the comparison .chp and baseline .chp files in GCOS in the Pivot table view.

1.2 Display Pivot table columns in the analysis by selecting "Pivot Data>Absolute Results" from the "View" pull-down menu. Ensure that the Detection, Change, and Signal Log Ratio columns are displayed.

1.3 Sort the data on the Detection column of the baseline file in descending order by right-clicking the Detection column heading and selecting "Sort Descending" from the pop-up menu as shown in Figure 36.

1.4 Click on the probe set identifier contained in the far-left column at the top of the list.

1.5 Use the mouse to scroll down the data list until the last 'P' is visible in the baseline file.

1.6 Hold down the 'Shift' key and click on the probe set identifier corresponding to the last 'P' value.

1.7 Click the "Hide unselected probe sets" button as shown in Figure 37.

1.8 The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 38. Enter this value into the box on Line 5 of the Change Calculation Worksheet.

2 Calculate the number of probe sets from the above list that also have a Change call of 'D.'

2.1 After performing step 1 of the Decrease calculation, sort the data on the Change column of the comparison file in ascending order by right-clicking the Change column heading and selecting "Sort Ascending" from the pop-up menu as shown in Figure 36.

2.2 Click on the probe set identifier contained in the far-left column at the top of the list.

2.3 Scroll down the list until the last 'D' call is visible, hold down the 'Shift' key and click on the corresponding probe set identifier.

2.4 Click the "Hide unselected probe sets" button as shown in Figure 37.

2.5 The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 38. Enter this value into the box on Line 6 of the Change Calculation Worksheet.

3 Calculate the number of probe sets from above list that also have a Signal Log Ratio of -1.0 or less.

3.1 After performing step 2 of the Decrease calculation, sort the data on the Signal Log Ratio column of the comparison file in descending order by right-clicking the Signal Log Ratio column heading and selecting "Sort Descending" from the pop-up menu as shown in Figure 36. (Note that GCOS sorts the Signal Log Ratio column on the magnitude of the Signal Log Ratio, hence, the sign of the value is ignored.)

3.2 Click on the probe set identifier at the top of the list.

3.3 Scroll down the list until the last Signal Log Ratio value equal to -1.0 is visible, hold down the 'Shift' key, and click on the corresponding probe set identifier.

3.4 Click the "Hide Unselected probe sets" button as shown in Figure 37.

3.5 The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 38. Enter this value into the box on Line 7 of the Change Calculation Worksheet.

4 Calculate the number of probe sets that have decreased, as a percentage of the probe sets detected.

4.1 Divide the number of probe sets showing significant decrease (Line 7) by the number of probe sets detected (Line 5).

4.2 Multiply the above number by 100 to convert to a percentage.

4.3 Enter the value into the box on Line 8 of the Change Calculation Worksheet.

Calculate Total Percentage Change

Finally, add the Percentage Increase (Line 4) to the Percentage Decrease (Line 8) and place the sum into the box on Line 9 of the Change Calculation Worksheet.

If the two samples being compared are from the same hybridization cocktail, the value in Line 9 should be less than 2.0. If this is not the case, it is likely that the arrays were not analyzed using the same scaling strategy. The data should be re-analyzed paying particular attention to ensure that the scaling strategy is identical for all analyses performed before contacting your Affymetrix Field Applications Specialist for further consultation.

Appendix E: Change Calculation Worksheet for GeneChip® Operating Software

Experiment File name: _____

Baseline File name: _____

Increases

Number of probe sets with Detection of 'P' in Experiment:	Line 1
Number of probe sets from Line 1 that have a Change call of 'I' :	Line 2
Number of probe sets from Line 2 that have a Signal Log Ratio of >1:	Line 3
% Increase (Line 3 divided by Line 1)*100:	Line 4

Decreases

Number of probe sets with Detection of 'P' in Baseline:	Line 5
Number of probe sets from Line 5 that have a Change call of 'D' :	Line 6
Number of probe sets from Line 6 that have a Signal Log Ratio of <-1:	Line 7
% Decrease (Line 7 divided by Line 5) x 100:	Line 8

Total Changes

Total % Change (Line 4 + Line 8):	% Line 9

Appendix F: Statistical Analysis Flow

The Statistical Analysis Flow questions and diagram (see Figure 39) provides important considerations before starting a statistical analysis of gene expression data.

1. Are my observations approximately normally distributed? If yes, then use a parametric test. If not, use a non-parametric test.
2. If using a parametric test, are variances between groups approximately equal? If yes, proceed. If not, apply weighting techniques.
3. How many groups will I be examining? If two, go to step 4. If more than two, use an ANOVA or Kruskal-Wallis test, which can be utilized to examine overall group variability.
4. If my number of groups equals two, are the observations paired or unpaired? If they are paired, use a paired Student's t-test or Wilcoxon Signed-Rank test. If they are unpaired, then use an unpaired Student's t-test or a Mann-Whitney.
5. What if I have used an ANOVA and I want to examine which groups are different? In this case, you would want to use a multiple comparisons test. It will be necessary to correct for Type I error by either using a Bonferroni correction, or some other type of correction.
6. Retrieve all relevant annotation information from the NetAffx Analysis Center.
7. Do all significant genes require follow up? If yes, go to step 8. If no, filter using functional information and return to step 6. Researchers must "draw the line" with regard to the list of "interesting" genes by integrating other types of biological data with the statistical analysis.
8. Are all significant genes well annotated? If no, identify further experiments and research. If yes, integrate other data types (cell biology, biochemistry, proteomics, histology, etc.).
9. Are high-level biological patterns and gene groups present? If no, identify further experiments and research. If yes, prepare research findings for publication and then identify further experiments and research.

Statistical Analysis Flow Diagram

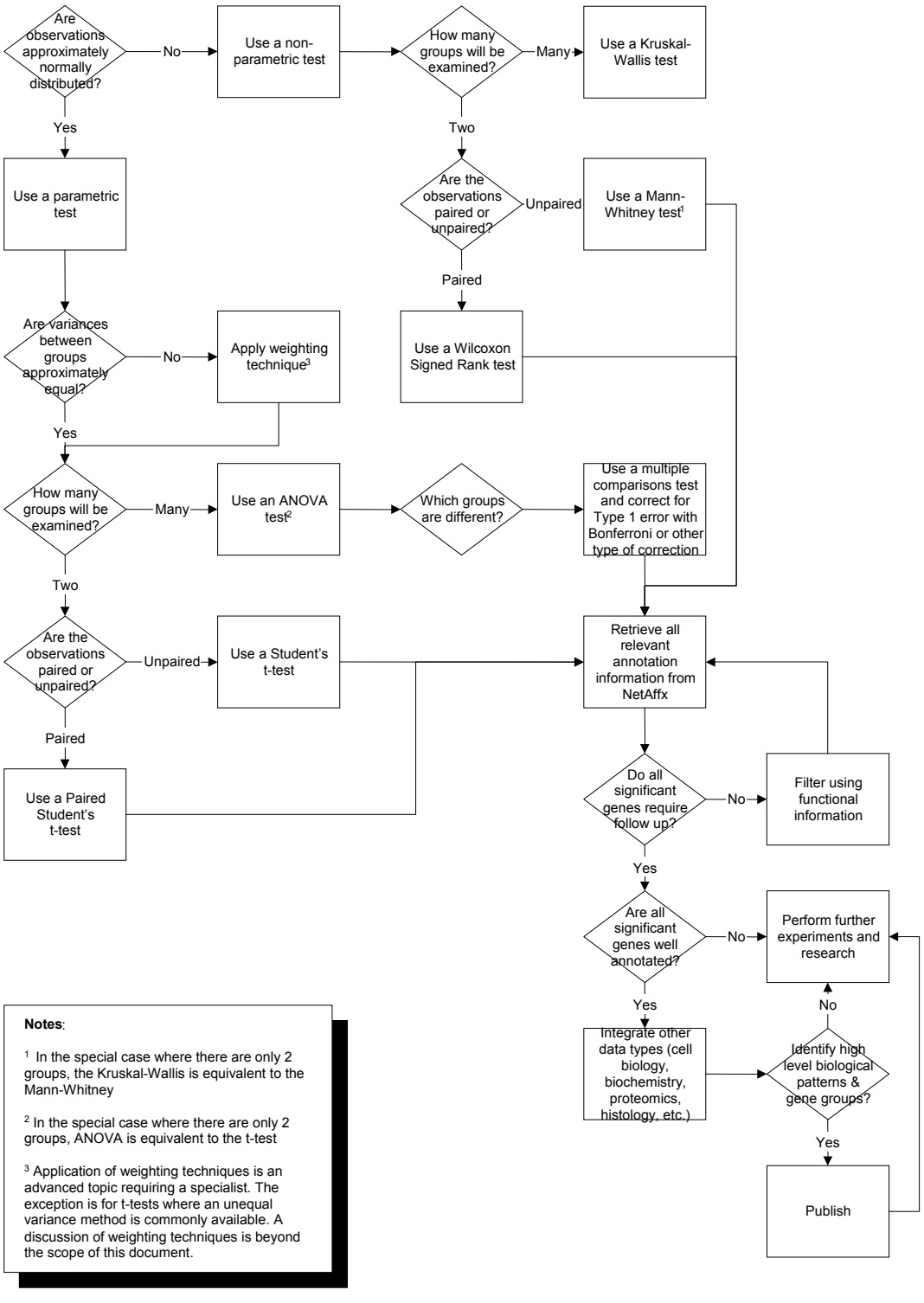


Figure 39. The statistical analysis flow diagram is one representation of many possibilities.

Appendix F: References

1. Affymetrix, Inc. "GeneChip® Expression Analysis Technical Manual" http://www.affymetrix.com/support/technical/manual/expression_manual.affx (August 2002).
2. Coller, H. A. et al. "Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion" *Proceedings of the National Academy of Sciences of the United States of America* 97(7): 3260-5 (2000).
3. Braam, J. "The Arabidopsis TCH Genes: Regulated in Expression by Mechanotransduction?" *Plant Tolerance to Abiotic Stresses in Agriculture: Role of Genetic Engineering. NATO Advance Research Workshop Proceedings* J.H. Cherry et al. (eds): 29-37 (2000).
4. Jin, Hongkui M. D. et al. "Effects of Early Angiotensin-Converting Enzyme Inhibition on Cardiac Gene Expression after Acute Myocardial Infarction" Circulation 103(5): 736-742 (2001).
5. Affymetrix, Inc. "GeneChip® Eukaryotic Small Sample Target Labeling Assay Version II" http://www.affymetrix.com/support/technical/technotes/smallv2_technote.pdf (2003).
6. Golub, T. R. et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" *Science* 286(5439): 531-537 (1999).
7. Motulsky, Harvey. *Intuitive Biostatistics.* ISBN: 0195086074 *Oxford University Press*, New York (October 1995).
8. Terry Speed's Microarray Data Analysis Group. "Always log spot intensities and ratios" http://www.stat.berkeley.edu/users/terry/zarray/Html/log.html (2000).
9. Westfall, Peter H., and Young, S. Stanley. "Resampling-Based Multiple Testing: Examples and Methods for *P*-Value Adjustment" ISBN 0-471-55761-7 *John Wiley and Sons, Inc.*, Somerset, NJ (1993).
10. "Thresholding of statistical maps in functional neuroimaging using the false discovery rate" *Neuroimage* 15(4): 870-8., PMID: 11906227 [PubMed - indexed for MEDLINE] (April 2002).
11. Lian et al. *Blood* 100(9): 3209-3220 (2002).
12. Jansen et al. *Genome Research* 12: 37-46 (2002).
13. Halfon, M. S. et al. *Physiol Genomics* 10(3): 131-143 (2002).
14. Affymetrix, Inc. "Manufacturing Quality Control and Validation Studies of GeneChip® Arrays" http://www.affymetrix.com/support/technical/technotes/manufacturing_quality_technote.pdf (2002).
15. Wodicka, L. *et al.* A Genome-Wide Expression Monitoring in *Saccharomyces Cerevisiae. Nature Biology* **15**, 1359-1367 (1997).

AFFYMETRIX, INC.
3380 Central Expressway
Santa Clara, CA 95051 USA
Tel:1 -888-DNA-CHIP (1-888-362-2447)
Fax:1 -408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd
Voyager, Mercury Park,
Wycombe Lane, Wooburn Green
High Wycombe HP10 0HH
United Kingdom
Tel:+44 (0) 1628 552550
Fax:+44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

AFFYMETRIX JAPAN K.K.
Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel:+81 -(0)3-5730-8200
Fax:+81 -(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com

For research use only.

Not for use in diagnostic procedures.

Part No. 701190 Rev. 4