

# GeneChip® Expression Analysis

Data Analysis Fundamentals

# Table of Contents

	<b>Page No.</b>
<b>Introduction</b>	1
<b>Chapter 1</b>	
Guidelines for Assessing Sample and Array Quality	2
<b>Chapter 2</b>	
Statistical Algorithms Reference Guide	5
<b>Chapter 3</b>	
Single Array Analysis	13
<b>Chapter 4</b>	
Comparison Analysis	16
<b>Chapter 5</b>	
Basic Data Interpretation	20
<b>Chapter 6</b>	
Change Calculation Worksheet	23
<b>Chapter 7</b>	
The NetAffx™ Analysis Center Summary	27
<b>Chapter 8</b>	
Relevant Publications	31
<b>Appendixes</b>	
Appendix A: Glossary	33
Appendix B: GeneChip® Probe Array Probe Set Name Designations	36
Appendix C: Microarray Suite Expression Defaults	38
Appendix D: File Types	39

# Introduction

The purpose of this manual is to provide users with a comprehensive description of different terms used in GeneChip® expression analysis, to present users with information on assessing sample and array quality, and to supply instructions on how to use the Affymetrix® Microarray Suite (MAS) software to analyze expression data. This handbook is a supplement to Affymetrix manuals and does not replace them. Brief descriptions of the different sections covered in this manual are as follows:

## **Guidelines for Assessing Sample and Array Quality**

This section provides guidelines to assess array and sample quality.

## **Statistical Algorithms Reference Guide**

This chapter focuses on the new Affymetrix Statistical Algorithms used in the expression analysis of GeneChip probe arrays. It provides a basic description of the mathematical concepts behind expression measurements for both single array and comparison analysis.

## **Single Array and Comparison Analyses**

These sections provide step-by-step instructions for both single array and comparison analysis using Microarray Suite.

## **Basic Data Interpretation and Change Calculation Worksheet**

These two sections cover step-by-step instructions for sorting data and calculating false change from data generated in Microarray Suite.

## **The NetAffx™ Analysis Center Summary**

This chapter includes background information and functionality of the NetAffx Analysis Center.

## **Relevant Publications**

This chapter provides additional information and relevant publications that users might find helpful in gathering further information.

## **Appendices**

### **Appendix A: Glossary**

This appendix defines terminology used in expression analysis using GeneChip probe arrays.

### **Appendix B: GeneChip Probe Array Probe Set Name Designations**

Background information of databases from which gene sequences are derived and the definitions of the probe set extensions are also available in this section.

### **Appendix C: Microarray Suite Expression Defaults**

This appendix covers the different defaults that can be used in Microarray Suite.

### **Appendix D: File Types**

This appendix describes all the file types associated with the GeneChip microarray platform.

To learn more about Affymetrix products or technology, please visit [www.affymetrix.com](http://www.affymetrix.com).

# Chapter 1 Guidelines for Assessing Sample and Array Quality

---

The purpose of this chapter 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 performance and evaluate sample 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.

## **RNA Sample QC**

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.

260/280 absorbance readings should be measured for both total RNA and biotinylated cRNA. Acceptable 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.

For optimal results, please follow the protocols described in the Affymetrix® GeneChip® Expression Analysis Technical Manual.

## **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. Depending on the nature of the artifact, you may wish to apply an image mask (use the mouse to click and drag on the desired area, then select “Mask Cells” from the Edit menu) in order to eliminate affected probe cells from data analysis. Please contact your Field Applications Specialist (FAS) for further advice regarding image artifacts.

After scanning the probe array, the resulting image data created is stored on the hard drive of the GeneChip Analysis Suite/Microarray Suite workstation 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 Microarray Suite to determine the appropriate grid size used. 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 not aligned correctly, 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 GeneArray® Scanners calibrated to the new PMT setting (10% of maximum). For arrays scanned with GeneArray Scanners under the old PMT setting (100%), values range from 200 to 1,000. 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 GeneArray Scanner.
2. Sample quality.

Each GeneArray Scanner has a unique inherent electrical noise associated with its operation. Since a significant portion of Noise (Raw Q) is based on electrical noise, absolute Noise (Raw Q) values among scanners will vary. Arrays being compared that were scanned on the same scanner should ideally have comparable Noise (Raw Q) values.

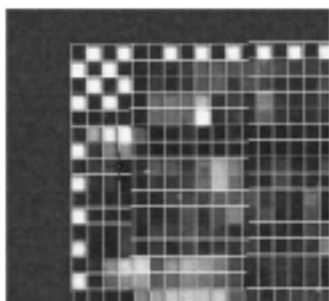
## **B2 Oligo Performance**

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

- The alternating pattern of intensities on the border
- The checkerboard pattern at each corner (Refer to Figure 1)
- The array name, located in the upper-left or upper-middle of the array (Refer to Figure 2)

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).



**Figure 1.** An example of B2 illuminating the corner and edges of the array.



**Figure 2.** The array name.

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

*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 both T3 and T7 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. Amplifying these controls with T7 RNA polymerase and biotinylated ribonucleotides will yield antisense cRNAs, which can be spiked into a hybridization cocktail and evaluated like the 20x Eukaryotic Hybridization Controls (*bioB, bioC, bioD, and cre*).

Details on poly-A preparation are described in the GeneChip Expression Analysis Technical Manual (Section 2 and Section 3, Chapter 2)

### **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 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 to 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, 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. Since the gene 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.

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. Since the gene expression assay is not biased in terms of the transcripts being amplified, 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 Genes 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 document.

### **Scaling and Normalization Factors**

Details regarding Scaling and Normalization are listed in the Affymetrix Microarray Suite User Guide Version 5.0, Appendix D. Scaling and normalization factors can be found either in the Analysis Info tab of the .chp file output or 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 for users of the HG-U133 Set is to apply the “Selected Probe Sets” method using the 100 Normalization Control probe sets.

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 2 Statistical Algorithms Reference Guide

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 1 illustrates the output of Single Array Analysis in Microarray Suite 5.0.

	Signal	Detection	Change
34820_at	2058.1	P	
160025_at	305.0	P	
160029_at	1393.2	P	

**Figure 1. 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 higher the discrimination scores are above Tau, the smaller the  $p$ -value and the more likely the transcript will be Present. The lower the discrimination scores are below Tau, the larger the  $p$ -value and the more likely the transcript will be Absent. The  $p$ -value associated with this test reflects the confidence of the Detection call.

### Detection $p$ -value

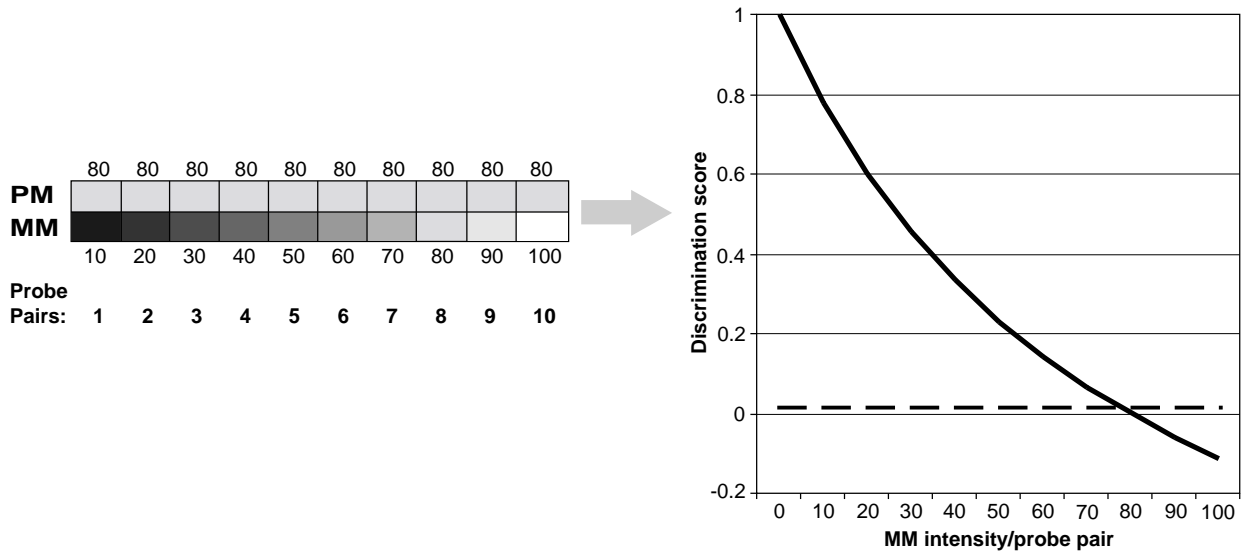
A two-step procedure determines the Detection  $p$ -value for a given probe set.

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):

$$R = (PM - MM) / (PM + MM)$$

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 2). 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 2). If the Discrimination scores are low for the majority of the probe pairs, the calculated Detection  $p$ -value will be higher (less significant).



**Figure 2.** 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).



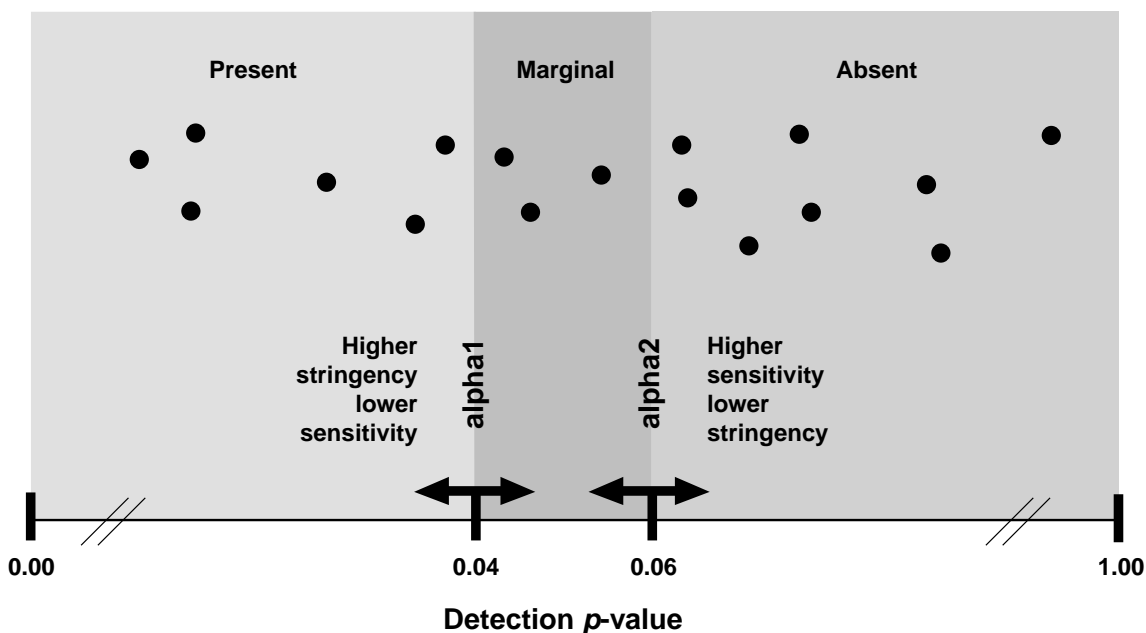
The next step toward the calculation of a Detection  $p$ -value is the comparison of each Discrimination score to the user-definable threshold Tau. Tau is a small positive number that can be adjusted to increase or decrease sensitivity and/or specificity of the analysis (default value = 0.015). The One-Sided 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.**

**Detection Call**

The user-modifiable Detection  $p$ -value cut-offs, Alpha 1 ( $\alpha_1$ ) and Alpha 2 ( $\alpha_2$ ) (See Figure 3), 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  $\alpha_1$  is assigned a Present call, and above  $\alpha_2$  is assigned an Absent call. Marginal calls are given to probe sets which have  $p$ -values between  $\alpha_1$  and  $\alpha_2$  (see Figure 3). The  $p$ -value cut-offs can be adjusted to increase or decrease sensitivity and specificity.



**Figure 3. Significance levels  $\alpha_1$  and  $\alpha_2$  define cut-offs of  $p$ -values for Detection calls. Please note that these cut-offs are for probe sets with 16–20 probe pairs.**

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. Note that a probe pair is rejected from further analysis when a Mismatch (MM) probe cell is saturated (MM = 46,000 for the 2500 GeneArray Scanner).

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

**Signal Algorithm**

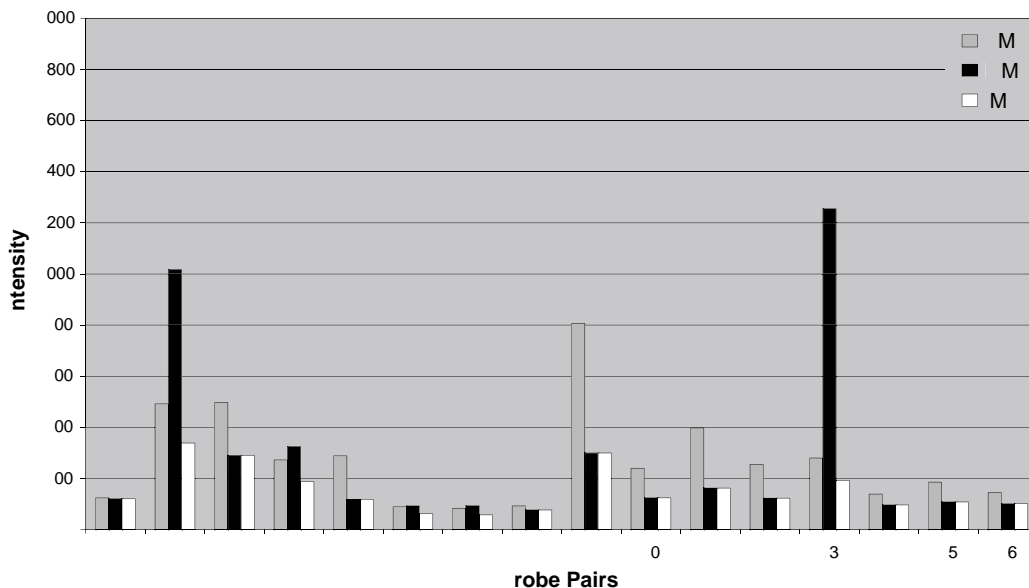
Signal is a quantitative metric calculated for each probe set, which represents the relative level of expression of a transcript. Signal is calculated using the One-Step Tukey's Biweight Estimate which yields a robust weighted mean that is relatively insensitive to outliers, even when extreme.

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 4).

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 4.** 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 5 illustrates the output of Comparison Analysis in Microarray Suite 5.0.

	Stat Common Pairs	Signal Log Ratio	Signal Log Ratio Low	Signal Log Ratio High	Change	Change <i>p</i> -value
35639_at	16	0.3	0.2	0.4	I	0.000014
1799_at	16	0.9	0.5	1.3	I	0.000015
35985_at	16	0.4	0.3	0.5	I	0.000015
34696_at	16	0.4	-0.1	0.9	I	0.000023
31356_at	16	1.8	0.8	2.8	I	0.000025
35202_at	16	0.4	0.2	0.6	I	0.000027
39651_at	16	0.4	0.3	0.5	I	0.000029
39777_at	16	0.4	0.1	0.6	I	0.000031
37610_at	16	0.4	0.2	0.5	I	0.000034
32070_at	16	0.3	0.2	0.4	I	0.000034
1581_s_at	16	0.7	0.1	1.3	I	0.000037
35283_at	16	0.5	0.3	0.6	I	0.000037

**Figure 5. 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 widens the range allowed before a change is called. 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. 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 6). Hence, the *p*-value scale is used to generate discrete change calls using thresholds. These thresholds will be described in the Change Call section.

	Signal Log Ratio	Signal Log Ratio Low	Signal Log Ratio High	Change	Change <i>p</i> -value
i25069_s_at	-0.4	-0.5	-0.2	D	0.999979
m12303_s_at	-0.3	-0.4	-0.2	D	0.999999
m12347_t_at	-0.3	-0.3	-0.2	D	0.999997
m21496_f_at	-0.2	-0.3	-0.2	D	0.999763
m29293_s_at	-0.2	-0.2	-0.1	D	0.998697
m34173_at	-0.3	-0.4	-0.3	D	0.999505
m52867_t_at	-0.4	-0.5	-0.3	D	1.000000
n29127_rc_at	-0.4	-0.5	-0.3	D	0.997989
i07577_s_at	-0.4	-0.5	-0.4	D	1.000000
aa000380_t_at	0.4	0.2	0.6	I	0.002283
aa002704_at	0.7	0.5	0.9	I	0.000150
aa002761_s_at	0.4	0.2	0.6	I	0.000226
aa000148_s_at	0.3	0.2	0.4	I	0.001246
aa009154_t_at	0.4	0.3	0.6	I	0.000028
aa013647_s_at	0.3	0.2	0.4	I	0.002011
AA023300_at	0.2	-0.1	0.6	I	0.001140
aa023407_t_at	0.9	0.6	1.2	I	0.000088
AA408234_rc_g_at	0.5	0.3	0.6	I	0.000001

**Figure 6. 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

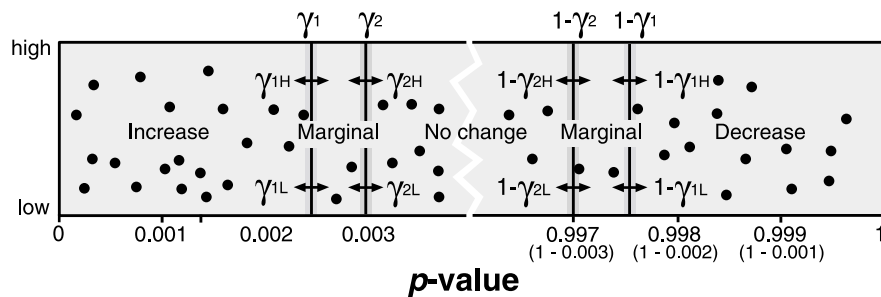
The final Change  $p$ -value described above is categorized by cutoff values called gamma1 ( $\gamma_1$ ) and gamma2 ( $\gamma_2$ ) (see Figure 7). These cut-offs provide boundaries for the Change calls: Increase (I), Marginal Increase (MI), No Change (NC), Marginal Decrease (MD), or Decrease (D).

The user does not directly set  $\gamma_1$  and  $\gamma_2$ ; rather each is derived from two user-adjustable parameters,  $\gamma_L$  and  $\gamma_H$ . In the case of  $\gamma_1$ , the two user-adjustable parameters are called  $\gamma_{1L}$  and  $\gamma_{1H}$  (defaults for probe sets with 15-20 probe pairs:  $\gamma_{1L}= 0.0025$  and  $\gamma_{1H}= 0.0025$ ), which define the lower and upper boundaries for  $\gamma_1$ . Gamma2 ( $\gamma_2$ ) is computed as a linear interpolation of  $\gamma_{2L}$  and  $\gamma_{2H}$  (defaults for probe sets with 15-20 probe pairs:  $\gamma_{2L}= 0.003$  and  $\gamma_{2H}= 0.003$ ) 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  $\gamma_{1L} = \gamma_{1H}$  and  $\gamma_{2L} = \gamma_{2H}$

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 7. 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  $\gamma$  values. The  $\gamma_1$  value is a linear interpolation of  $\gamma_{1L}$  and  $\gamma_{1H}$ . Similarly  $\gamma_2$  is derived from  $\gamma_{2L}$  and  $\gamma_{2H}$ .**

## 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 is 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)**

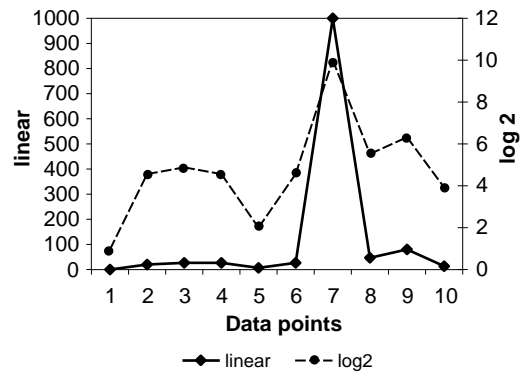
Statistical Algorithms	Empirical Algorithms
Signal	Average Difference
Detection	Absolute Call
Change	Difference Call
Signal Log Ratio	Fold Change

**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:**

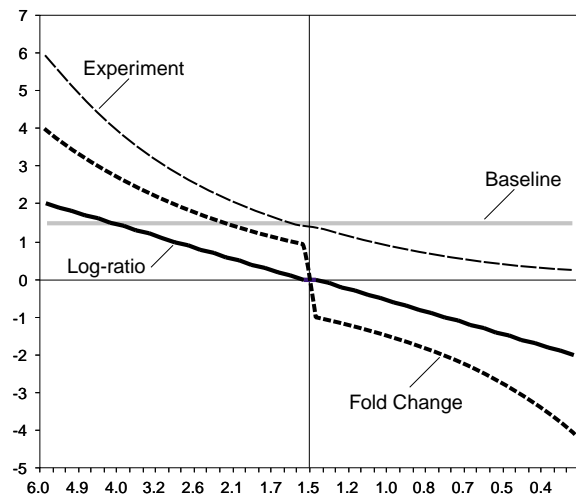
Hybridized probe intensities tend to be distributed over exponential space due to hybridization behavior that is governed by exponential functions of sequence-dependent base-pairing energetics. Thus, log transformation is an appropriate process for analyzing hybridization data. Some of the benefits are apparent in this graph where we show the same data set plotted on two scales. When the data is 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 is plotted on a Log<sub>2</sub> scale (dashed line), we can see variations in the low values as well as the very high values.



**Signal Log Ratio vs. Fold Change**

In this graph, Signal Log Ratio is compared to Fold Change in a hypothetical experiment. 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).

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<sub>2</sub>, 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:



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

## Chapter 3 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 GeneChip array analysis. More detailed information can be found in the Affymetrix Microarray Suite User Guide Version 5.0 (P/N 700293) or by contacting your Affymetrix Field Applications Specialist.

The following instructions assume that a GeneChip 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 Microarray Suite 5.0 software. After analysis of arrays, the procedures in the preceding chapters can be used to assess the quality of the data generated.

These instructions relate to analyses performed in Microarray Suite 5.0. Notes for using Microarray Suite 4.0 can be found at the end of this chapter.

### **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 Chapter 1 of this training manual.

NOTE: Refer to Chapter 1 to aid in quality assessment of the array.

### **Select a Scaling Strategy**

These instructions use a global scaling strategy that sets the average signal intensity of the array to a Target Signal of 500. The key assumption of the global scaling strategy is that there are few changes in gene expression between the arrays being analyzed. This is a common strategy used by many users, however, it should be noted that this strategy is not appropriate for all experiments. Further discussion on scaling strategies and how to implement them can be found in Appendix D of the Microarray Suite User Guide Version 5.0 or by contacting your Affymetrix Field Applications Specialist.

### **Expression Analysis Set-Up**

A single array analysis will create a .chp file from a .cel image file. Microarray Suite 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.
  - a) Check “Prompt For Output File” to ensure display of output file name for confirmation or editing. With this option checked, Microarray Suite will prompt for new file names for each analysis preventing unintentional overwrite.
  - b) Check “Display Settings When Analyzing Data” to ensure display of expression settings for confirmation or editing.
3. Select “File Locations” tab to verify:
  - a) the Location listed to the right of Probe Information is the directory containing the library files.
  - b) the Location listed to the right of Fluidics Protocols is the directory containing the fluidics protocols.
  - c) the Location listed to the right of Experiment Data is the directory containing the data files to be analyzed.

NOTE: Errors are commonly found in Microarray Suite due to incorrectly set file locations.

4. Select the “Database” tab. Select “Disk Files” mode to direct where file information will be saved.

NOTE: “Disk Files” refers to data storage on the local hard drive. “Affymetrix® LIMS” refers to storage on a dedicated server using the Affymetrix Laboratory Information Management System.

5. Select “OK.”

### **Expression Analysis Settings**

Select “Analysis Settings>Expression” from the “Tools” pull-down menu. The “Expression Analysis Settings” dialogue box opens.

1. Select the “Probe Array Type” to be analyzed from the drop-down menu.
2. Select the “Scaling” tab.
  - a) Select “All Probe Sets” and set “Target Signal” to 500 or to desired Target Signal.
3. Select the “Normalization” tab.
  - a) 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 is being scaled. Further information can be found in Appendix D of Microarray Suite User’s Guide Version 5.0.
4. Select the “Probe Mask” tab. This feature is used to mask user-defined probe cells.
  - a) Ensure that the “Use Probe Mask File” option is unchecked.
5. Select the “Baseline” tab. For single array analysis no baseline file should be used.
  - a) Ensure “Use Baseline File Comparison” is unchecked.
6. Select the “Parameters” tab.
  - a) Confirm default settings appropriate to the version of Microarray Suite 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.

7. 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 it in the data file tree. Alternatively, select “Open” from the “File” pull-down menu and select the image file you wish to analyze.
2. Select “Analysis” from the “Run” pull-down menu.
  - a) Verify the .chp file name. The default corresponds to the name of the .exp and .dat file names. Edit the .chp file name, if necessary, and click “OK.”

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

- b) 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).
  - c) The Microarray Suite status window will indicate that the analysis has started.
3. Once the analysis is complete, generate an Expression Analysis report file (.rpt) and review the quality control metrics.
  - a) To generate the report select “Report” from the “File” pull-down menu.
  - b) 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.”

- c) Review the quality control data.

NOTE: See Chapter 1 for detailed explanations.

- Review *bioB*, *bioC*, *bioD*, and *cre* sensitivity spikes.
- Review percent present determination.
- Review internal control 3'/5' ratios.
- Review noise (Raw Q).
- Review background.

- d) 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 Microarray Suite shortcuts window.

4. 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 on the far-right a brief description of the sequence that the probe set represents.

- a) 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 Call,” “Detection *p*-value,” “Stat Pairs,” and “Stat Pairs Used.”  
(\*See Notes for Microarray Suite 4.0 Users.)

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

- b) Select the “Metrics” tab at the bottom of the .chp file.
- c) The Metrics table displays data for each distinct probe set in the .chp file. Columns displayed are similar to the Pivot table.
- (1) Organize the tabular data columns by right clicking at the top of the column to “Hide Column.”
- (2) Sort by right clicking on the column header and selecting the desired sorting function.

NOTE: Refer to Chapter 5 for recommendations.

- d) 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. Information includes Scaling or Normalization factors, Background, Raw Q, and Sample Type information.

Once a single chip 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 (see Chapter 4 of this training guide). The .chp file can also be published into either the MicroDB™ or LIMS database, becoming accessible for data mining with the Affymetrix Data Mining Tool or other third-party analysis tools. The .chp file data can also be exported from Microarray Suite as a text file allowing the data to be imported into third-party programs (e.g., Microsoft® Excel).

### **Note for Microarray Suite 4.0 Users**

\*Step 4. a) in Performing Single Array Analysis.

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 “Average Difference” and “Absolute Call.”

## Chapter 4 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 the 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.

As Microarray Suite 4.0 and Microarray Suite 5.0 use different algorithms, the files being compared must be analyzed using the same version of Microarray Suite. These instructions relate to analyses performed in Microarray Suite 5.0. Notes for Microarray Suite 4.0 users can be found at the end of this section.

### **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 in Chapter 1 of this training manual.

NOTE: Refer to Chapter 1 to aid in quality assessment of the array.

### **Ensuring Consistency of Files to be Compared**

Ensure .dat and .cel files corresponding to both the designated experiment and baseline files along with the baseline .chp file are present in the data file tree. If they are not, verify that the files are in the same directory and that the directory is specified correctly, as described on page 13 of this training manual.

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 was 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 this protocol should be 500.
SF	Displays the scaling factor calculated. In this protocol this should <b>NOT</b> be 1.0000.
NF	Displays the normalization factor applied. In this protocol the value should be 1.0000, as no normalization was used.
SFGene	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. Microarray Suite automatically generates the .cel image file from the .dat file. To perform a comparison analysis, settings relating to file locations and the analysis must first be defined.

### **Expression Analysis Set-Up**

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

1. Select the “Probe Array Type” to be analyzed from drop-down menu.
2. Select the “Scaling” tab.
  - a) Select “All Probe Sets” and set “Target Signal” to 500.
3. Select the “Normalization” tab.
  - a) Select “User Defined” and place a “1” in the “Normalization Value” box.
4. Select the “Probe Mask” tab. This feature is used to mask user-defined probe cells.
  - a) Ensure that the “Use Probe Mask File” option is unchecked.

5. Select the “Baseline” tab.

- a) Check the “Use Baseline File Comparison” option.
- b) Click the “Browse” button.
- c) Select the baseline .chp file.
- d) Click the “Open” button.

6. Select the “Parameters” tab.

- a) Confirm default settings appropriate to the version of Microarray Suite 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.

7. 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.
  - a) Verify the .chp filename. The default corresponds to the name of the experimental file .exp and .dat filename. Edit the .chp filename, if necessary, and click “OK.”

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

- b) 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.
  - c) The Microarray Suite status window will indicate that the analysis has started.
3. Once the analysis is complete, generate an Expression Analysis report file (.rpt) and review the quality control metrics.
  - a) To generate the report, select “Report” from the “File” pull-down menu.
  - b) 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.

- c) Review the quality control data.

NOTE: See Chapter 1 for detailed explanations.

- Review *bioB*, *bioC*, *bioD*, and *cre* sensitivity spikes.
  - Review percent Present determination.
  - Review internal control 3’/5’ ratios.
  - Review noise (Raw Q).
  - Review background.
- d) 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 Microarray Suite 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.
  - a) 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,” “Detection,” “Detection *p*-value,” “Signal Log Ratio,” “Change,” and “Change *p*-value.” (\*See notes for Microarray Suite 4.0 Users).
  - b) 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.
  - c) 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 “Present” and have increased at a “Signal Log Ratio” of >1.

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 either the MicroDB or LIMS database, becoming accessible for data mining with the Affymetrix® Data Mining Tool or other third-party analysis tools. The .chp file data can also be exported from Microarray Suite 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 13 of the Affymetrix Microarray Suite User Guide Version 5.0.

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.
2. Add files to the Batch Analysis window by:
  - a) Dragging and Dropping each .cel or .chp file to the Batch Analysis window from the data file tree to the Batch Analysis window.

OR

  - a) Using the Toolbar, click the “Add” Toolbar or select “Edit>Add.”
  - b) An open dialog of .cel files appears.
  - c) Select the .cel or .chp files to be analyzed.
  - d) To select all files hold “shift” while you click on the first and last file.
  - e) To select files individually, hold “control” while selecting files.
  - f) Click open to place the files into the Batch Analysis window.
3. Verify the Output filenames.
  - a) The filename for the .chp file is listed in the Output column. If the .chp filename is already present the filename will be red to indicate that a file is going to be overwritten.
  - b) To edit the .chp file name, double click on the output file name and add by typing 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.

a) Double click on the baseline .chp file from the dialog box.

b) 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.

**Note for Microarray Suite 4.0 Users**

\*Step 4 a) in Performing Comparison Analysis.

Select “Analysis>Options...>Pivot Tab” and select the comparison analysis metrics you wish to see from the right side of the menu under Comparison Results. Recommendations include “Fold Change” and “Difference Call.”

## Chapter 5 Basic Data Interpretation

---

The use of GeneChip gene expression arrays allows interrogation of several 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. The guidelines listed below apply to Microarray Suite 5.0. Notes for Microarray Suite 4.0 users are highlighted at the end of this section.

### **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? Microarray Suite 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 greater than two 98% of the time by Wodicka *et al.* in 1997 (26). 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 Microarray Suite, 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**

NOTE: For detailed sorting instructions, please refer to Chapter 6.

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.

\* 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 Microarray Suite (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.

### **Differences in MAS 4.0**

The metrics used to sort for robust changes in MAS 5.0 are Detection, Change, and Signal Log Ratio. The equivalent metrics in MAS 4.0 are Absolute Call, Difference Call, and Fold Change, respectively.

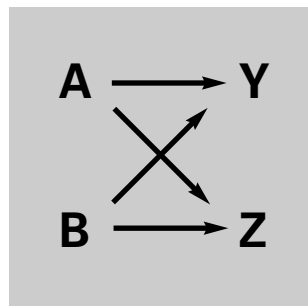
The Signal Log Ratio is essentially the log base 2 of the Fold Change. Thus, when sorting on MAS 4.0 gene expression data for significant increases, probe sets with a Fold Change value below 2.0 should be eliminated. For significant decreases, probe sets with a Fold Change value above -2.0 should be eliminated.

As with Signal Log Ratio, the Fold Change column sorts values on the magnitude 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 the Affymetrix Data Mining Tool software.



# Chapter 6 Change Calculation Worksheet

This procedure can be used to identify robust changes between two GeneChip probe arrays. These instructions relate to analyses performed in Microarray Suite 5.0. Notes for Microarray Suite 4.0 users can be found at the end of this chapter.

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 (36).

## 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 26).

## 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.  
(\*See Notes for Microarray Suite 4.0 Users.)

- a) Open the comparison .chp file in MAS 5.0, with the Pivot table view.
- b) 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.
- c) 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 1.
- d) Click on the probe set identifier, contained in the far-left column, at the top of the list.
- e) Use the mouse to scroll down the data list until the last 'P' is visible.
- f) Hold down the 'Shift' key and click on the probe set identifier corresponding to the last 'P' value.
- g) Click the "Hide unselected probe sets" button as shown in Figure 2.
- h) The number of remaining probe sets is displayed in the bottom-right of the window, as shown in Figure 3. Enter this value into the box on Line 1 of the Change Calculation Worksheet.

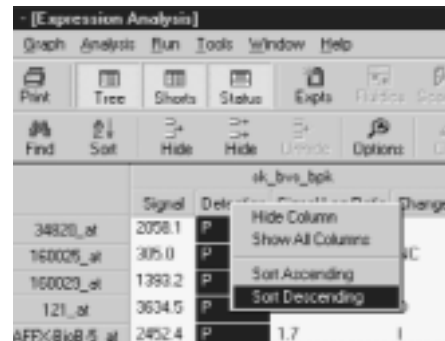


Figure 1

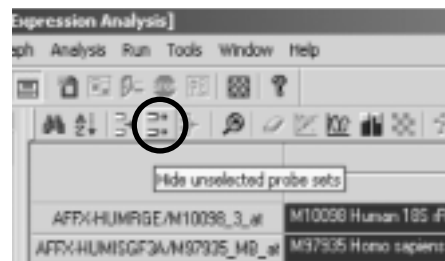


Figure 2



Figure 3

2. Calculate the number of probe sets from above list that also have a Change call of 'I.'  
 (\*\*See Notes for Microarray Suite 4.0 Users.)
  - a) 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 1.
  - b) Scroll down the list of probe sets until the first 'I' call is visible, then click on this probe set identifier.
  - c) Scroll down the list until the last 'I' call is visible, hold down the 'Shift' key and click on the corresponding probe set identifier.
  - d) Click the "Hide unselected probe sets" button as shown in Figure 2.
  - e) The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 3. 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.  
 (\*\*\*)See Notes for Microarray Suite 4.0 Users.)
  - a) 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 1.
  - b) Click on the probe set identifier at the top of the list.
  - c) 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.
  - d) Click the "Hide unselected probe sets" button as shown in Figure 2.
  - e) The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 3. 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.
  - a) Divide the number of probe sets showing significant increase (Line 3) by the number of probe sets detected (Line 1).
  - b) Multiply the above number by 100 to convert to a percentage.
  - c) 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.  
 (\*See Notes for Microarray Suite 4.0 Users.)
  - a) Open both the comparison .chp and baseline .chp files in MAS 5.0 in the Pivot table view.
  - b) 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.
  - c) 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 1.
  - d) Click on the probe set identifier contained in the far-left column at the top of the list.
  - e) Use the mouse to scroll down the data list until the last 'P' is visible in the baseline file.
  - f) Hold down the 'Shift' key and click on the probe set identifier corresponding to the last 'P' value.
  - g) Click the "Hide unselected probe sets" button as shown in Figure 2.
  - h) The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 3. 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.'  
(\*\*See Notes for Microarray Suite 4.0 Users.)
  - a) 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 1.
  - b) Click on the probe set identifier contained in the far-left column at the top of the list.
  - c) Scroll down the list until the last 'D' call is visible, hold down the 'Shift' key and click on the corresponding probe set identifier.
  - d) Click the "Hide unselected probe sets" button as shown in Figure 2.
  - e) The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 3. 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.  
(\*\*\*See Notes for Microarray Suite 4.0 Users.)
  - a) 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 1. (Note that Microarray Suite 5.0 sorts the Signal Log Ratio column on the magnitude of the Signal Log Ratio, hence, the sign of the value is ignored.)
  - b) Click on the probe set identifier at the top of the list.
  - c) 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.
  - d) Click the "Hide Unselected probe sets" button as shown in Figure 2.
  - e) The number of remaining probe sets is displayed in the bottom-right of the window as shown in Figure 3. 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.
  - a) Divide the number of probe sets showing significant decrease (Line 7) by the number of probe sets detected (Line 5).
  - b) Multiply the above number by 100 to convert to a percentage.
  - c) 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.

### **Notes for Microarray Suite 4.0 Users**

\*Step 1. The equivalent to the Detection call in Microarray Suite 4.0 is the Absolute Call.

\*\*Step 2. The equivalent to the Change call in Microarray Suite 4.0 is the Difference Call.

\*\*\*Step 3. The equivalent to the Signal Log Ratio in Microarray Suite 4.0 is the Fold Change. To identify the increases, Fold Change values  $\geq 2.0$  are required. For decreases, fold change values  $\leq -2.0$  are required.

# Change Calculation Worksheet for Microarray Suite 5.0

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 'T' :  Line 2

Number of probe sets from Line 2 that have a Signal Log Ratio of  $\geq 1$ :  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  $\leq -1$ :  Line 7

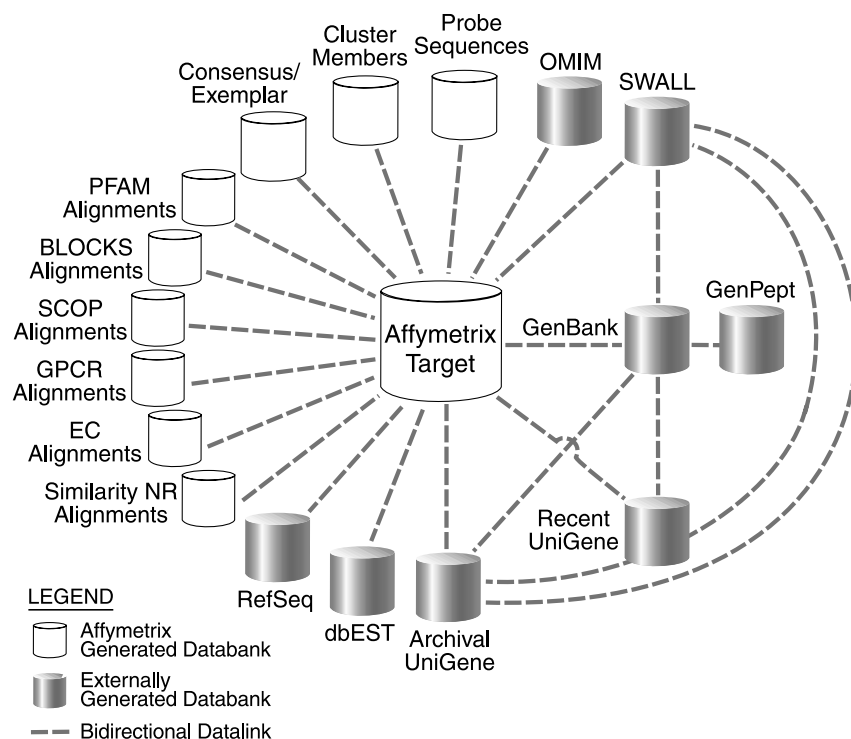
% Decrease (Line 7 divided by Line 5) \*100:  Line 8

## **Total Changes**

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

## Chapter 7 The NetAffx™ Analysis Center Summary

The NetAffx™ Analysis Center ([www.netaffx.com](http://www.netaffx.com)) is an online resource that allows researchers to correlate their GeneChip array results to a catalog of array design and annotation information. The NetAffx Analysis Center uses the SRS (Sequence Retrieval System) data and application integration platform.



**Figure 1**

This useful tool enables you to access product-specific biological annotations from both the public domain and Affymetrix (Figure 1). Specifically, you may link from target sequences to the information represented in PFAM, BLOCKS, SCOP, Similarity NR, and EC. These annotations provide further structural and functional information, helping you to draw biologically relevant conclusions about your experimental results.

For instructions on using the NetAffx Analysis Center for downstream analysis, please refer to the NetAffx Interactive Tutorial in the NetAffx Help Center.

### **Analysis Center**

The NetAffx Analysis Center is a comprehensive resource of functional annotations and public database information integrated with the probe sets. Now researchers can access detailed array content, including target and probe sequences. The NetAffx Analysis Center is now available to anyone who completes a short registration form.

Flexible query capabilities are provided to help you retrieve biological information for probe sets from both public and proprietary data. Unless otherwise noted, public data representations are updated on the site once every three months.

A new interactive Probe Set Display tool allows users to visualize information about probe alignments. It is currently available for HG-U133 Set and HG-U95 Set. To explore an interactive display graphic, you will need to download and install an SVG viewer from Adobe ([www.adobe.com](http://www.adobe.com)). For more information about our Probe Set Display tool, please refer to the user's guide.

### **Download Center**

This enables you to efficiently access the data represented on Affymetrix GeneChip catalog arrays. You may download consensus, exemplar, target (SIF), and probe sequences and incorporate this information into your internal data analysis pipelines.

## **Brief Information on the Databases Available on the NetAffx Analysis Center**

### **dbEST**

dbEST is a database for Expressed Sequence Tags (ESTs). More information about dbEST may be found at [www.ncbi.nlm.nih.gov/dbEST/](http://www.ncbi.nlm.nih.gov/dbEST/).

### **Domains\_PFAM (Affymetrix internal)**

Domains\_PFAM contains detailed alignment data associated with the computational annotation of protein domains represented in the PFAM database using the HMMer program. PFAM entries are derived from seed alignments largely generated through human curation. More information about PFAM may be found at [pfam.wustl.edu](http://pfam.wustl.edu).

### **Domains\_BLOCKS (Affymetrix internal)**

Domains\_BLOCKS contains detailed alignment data associated with the computational annotation of protein domains represented in the BLOCKS database. BLOCKS entries represent domains or motifs from multiply aligned, ungapped segments in the most highly conserved regions of proteins. More information about BLOCKS may be found at [www.blocks.fhcrc.org](http://www.blocks.fhcrc.org).

### **Families\_GPCR (Affymetrix internal)**

Families\_GPCR contains alignments to families of G protein coupled receptors as organized by SWISS-PROT. The alignments are generated by scoring against SAM-T99 derived HMM models. The GPCR classification list may be found at [www.expasy.ch/cgi-bin/lists?7tmrlist.txt](http://www.expasy.ch/cgi-bin/lists?7tmrlist.txt)

### **Families\_SCOP (Affymetrix internal)**

Families\_SCOP contains detailed alignment data associated with the computational prediction of structural classification based on protein sequence similarity to representative sequences from the SCOP database. SCOP is the Structural Classification of Proteins database containing a hierarchical representation of classes, folds, super families, families and individual proteins. Predictions are based on the creation of individual sub-family models using the SAM program and T-99-derived methods for HMM model generation. More information about SCOP may be found at [scop.mrc-lmb.cam.ac.uk/scop](http://scop.mrc-lmb.cam.ac.uk/scop).

### **Families\_EC (Affymetrix internal)**

Families\_EC contains detailed alignment data associated with the computational identification of homology to enzymes using the SAM-T99 method for HMM model generation. The Enzyme Commission (EC) classification scheme contains a hierarchical representation based on broad enzymatic classes, sets of substrates and cofactors/reagents. Enzyme commission (EC) numbers and associated pathway data is available via hypertext links to the Kyoto Encyclopedia of Genes and Genomes (KEGG). Information on KEGG can be found at [www.genome.ad.jp/kegg/](http://www.genome.ad.jp/kegg/). More information about EC may be found at [www.chem.qmw.ac.uk/iupac/jcbl/](http://www.chem.qmw.ac.uk/iupac/jcbl/) or on a page at Rockefeller University at [prowl.rockefeller.edu/enzymes/enzymes.htm](http://prowl.rockefeller.edu/enzymes/enzymes.htm).

### **GenBank**

GenBank is a public database of genetic sequences and annotations maintained by the National Center for Biotechnology Information (NCBI). You can access the web page at [www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/).

Summary of available information:

- Brief description of sequence includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding).
- Publications by the authors of the sequence that discuss the data reported in the record with links to corresponding MEDLINE records.
- Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features (promoter, 5'UTR, 3'UTR etc).
- Protein and DNA sequence in FASTA format.

### **GenPept**

GenPept is a database of translated protein coding sequence that is copied from the GenBank translation information. It is a duplication of the GenBank protein translation information. GenPept is the sequence format most appropriate to perform similarity searches.

### **LocusLink**

LocusLink provides curated gene sequences and descriptive information about genetic loci. More information about LocusLink can be found at [www.ncbi.nlm.nih.gov/LocusLink](http://www.ncbi.nlm.nih.gov/LocusLink).

Summary of available information:

- Official gene symbol and link to the Human Genome Nomenclature Database.
- Locus information: alternate gene symbols, links to the corresponding UniGene and OMIM records.
- Map information: chromosomal and cytogenetic location, STS markers associated with the locus, links to the NCBI Map Viewer.
- Links to GenBank, GenPept, PFAM records.
- Gene Ontology categories and links to appropriate PubMed records.

### **OMIM**

OMIM, Online Mendelian Inheritance in Man, is a catalog of human genes and genetic disorders. More information can be found at [www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM).

Summary of available information:

- Detailed description of the gene and its protein product.
- Summary of the literature and large set of links to the corresponding MEDLINE records.

### **Pathways**

Pathways contains mappings of signaling, metabolic, and biosynthetic pathways from [www.genmapp.org](http://www.genmapp.org) mapped to Affymetrix GeneChip probe sets. A link to the GenMAPP web site enables the user to pull down the GenMAPP software for examining pathways curated by the Conklin group at the Gladstone Institute at UCSF. Future releases of the NetAffx web site will include HTML documents depicting these pathways.

### **PFAM**

PFAM contains a large collection of multiple-sequence alignments and Hidden Markov Models covering many common protein domains. More information on this database can be found at [pfam.wustl.edu](http://pfam.wustl.edu).

### **Protein\_Summary**

Protein\_Summary contains the summary results of homology modeling of the translated peptide sequences associated with a probe set. The current databank contains annotations on the publicly annotated protein coding regions (CDS) of known full-length sequences. Sequence similarity is determined through several approaches as follows: Protein domains are identified using the HMMer program to search the PFAM database and by using position-specific weight matrices to search the BLOCKS database. A hidden Markov model is a previously trained statistical model for an ordered sequence of symbols such as bases or amino acids. It functions as a state machine that generates a symbol each time a transition is made from one state to the next. HMMs can function as probabilistic models for multiple sequence alignments where all possible combinations of matches, mismatches, and gaps are used to generate alignment in a series of sequences or may model periodic patterns in a single sequence. HMMs have been found in the exons of a gene or families of similar protein structure. Structural family prediction is based on hidden Markov models representing each SCOP structural sub-family using the SAM program. Enzyme classifications (EC) and associations with pathways are also obtained by using hidden Markov model searching. G protein coupled receptor (GPCR) classifications are obtained using hidden Markov model searching. General sequence similarity is obtained using the BLASTP program to search the non-redundant protein database (nr).

### **RefSeq (Reference Sequence Project)**

RefSeq is a non-redundant set of reference sequences including constructed genomic contigs, mRNAs, and proteins. It is a stable reference point for mutation analysis, gene expression studies, and polymorphism discovery. More information can be found at [www.ncbi.nlm.nih.gov/LocusLink/refseq.html](http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html).

Records are classified as follows:

- (NT\_#####) constructed genomic contigs
- (NM\_#####) curated mRNAs
- (NP\_#####) curated proteins
- (NC\_#####) chromosomes
- (XM\_#####) model mRNAs corresponding to genomic contig
- (XP\_#####) model proteins corresponding to genomic contig

**Similarity\_NR**

Similarity\_NR contains detailed alignment data associated with the annotation of homologous protein sequences determined by sequence similarity searching using the BLASTP program against the non-redundant protein database (nr) from the National Center for Biotechnology Information (NCBI). More information about the BLAST family of programs and the non-redundant protein database can be found at [www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/).

**Swiss\_Prot**

Swiss\_Prot is a curated protein sequence database that provides a high level of annotations, a minimal level of redundancy and a high level of integration with other databases. More information can be found at [www.expasy.org/sprot/](http://www.expasy.org/sprot/).

Summary of available information:

- Annotation information: description of protein function, domain structure, post-translational modifications, variants, etc.
- Extensive links to MEDLINE records.

**UniGene**

UniGene provides a non-redundant set of gene-oriented clusters. More information can be obtained at [www.ncbi.nlm.nih.gov/UniGene](http://www.ncbi.nlm.nih.gov/UniGene).

Summary of available information:

- Expression pattern (tissue-specific expression).
- Similarity to proteins in model organisms.
- Links to UniGene, Locus Link, dbEST, HomoloGene records and mapping information.



## Chapter 8 Relevant Publications

---

- Affymetrix Technical Note: Fine Tuning Your Data Analysis. (2001).
- Affymetrix Technical Note: New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays. (2001).
- Alevizos, I. *et al.* Oral cancer *in vivo* gene expression profiling assisted by laser capture microdissection and microarray analysis. *Oncogene* **20**, 6196-204 (2001).
- Bumol, T.F., Watanabe, A.M. Genetic information, genomic technologies, and the future of drug discovery. *Journal of the American Medical Associations* **285**, 551-555 (2001).
- Cao, S.X., Dhahbi, J.M., Mote, P.L., Spindler, S.R. Genomic profiling of short- and long-term caloric restriction effects in the liver of aging mice. *Proceedings of the National Academy of Sciences of the USA* **98**, 10630-10635 (2001).
- Carter, T.A. *et al.* Chipping away at complex behavior: Transcriptome/phenotype correlations in the mouse brain. *Physiology & Behavior* **73**, 849-57 (2001).
- Cho, R.J. *et al.* Transcriptional regulation and function during the human cell cycle. *Nature Genetics* **27**, 48-54 (2001).
- Cutler, D.J. *et al.* High-throughput variation detection and genotyping using microarrays. *Genome Research* **11**, 1913-25 (2001).
- Dong, D. *et al.* Flexible Use of High-Density Oligonucleotide Arrays for Single-Nucleotide Polymorphism Discovery and Validation. *Genome Research* **11**, 1418-1424 (2001).
- Electronic Textbook*, StatSoft, Inc. (1984-2000). [www.statsoft.com/textbook/stathome](http://www.statsoft.com/textbook/stathome).
- Fodor, S.P.A. Genes, Chips and the Human Genome. *FASEB Journal* **11**, A879 (1997).
- Gerhold, D. *et al.* Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays. *Physiological Genomics* **5**, 161-170 (2001).
- Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).
- Harrington, C., Rosenow, C., Retief, J. Monitoring gene expression using DNA microarrays. *Current Opinion in Microbiology* **3**, 285-291 (2000).
- Hoaglin, D.C., Mosteller, F., Tukey, J.W. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York (2000).
- Hollander, M., Wolfe, D.A. *Nonparametric Statistical Methods* (second edition). John Wiley & Sons, New York (1999).
- Hsiao, L.L. *et al.* A Compendium of Gene Expression in Normal Human Tissues Reveals Tissue-Selective Genes and Distinct Expression Patterns of Housekeeping Genes *Physiological Genomics*. *In press* (2001).
- Jin, H. *et al.* Effects of Early Angiotensin-Converting Enzyme Inhibition on Cardiac Gene Expression After Acute Myocardial Infarction. *Circulation* **103**, 736 (2001).
- Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics* **24**, 381-386 (2000).
- Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860 - 921 (2001).
- Lee, C.K., Weindruch, R., Prolla, T.A. Gene-expression profile of the aging brain in mice. *Nature Genetics* **25**, 294-297 (2000).
- Lee, C.K., Klopp, R.G., Weindruch, R., Prolla, T.A. Gene Expression Profile of Aging and Its Retardation by Caloric Restriction. *Science* **285**, 1390-1393 (1999).
- Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genetics Chipping Forecast* **21**, 20-24 (1999).
- Liu, W.M. *et al.* Analysis of high density expression microarrays with signed-rank call algorithms. *In Preparation* (2001).
- Liu, W.M. *et al.* Rank-based algorithms for analysis of microarrays. *Proceedings SPIE* **4266**, 56-67 (2001).
- Ly, D.H., Lockhart, D.J., Lerner, R.A., Schultz, P.G. Mitotic misregulation and human aging. *Science* **287**, 2486-2492 (2000).

- McDonald, M.J., Rosbash, M. Microarray Analysis and Organization of Circadian Gene Expression in *Drosophila*. *Cell* **107**, 567-578 (2001).
- Motulsky, H. Intuitive Biostatistics. Oxford University Press, New York (1995).
- Notterman, D.A., Alon, U., Sierk, A.J., Levine, A.J. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research* **61**, 3124-30 (2001).
- Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the USA* **98**, 15149-15154 (2001).
- Little, R. and Rubin, D. Statistical analysis with missing data. Wiley & Sons, New York (1987).
- Stamey, T.A. *et al.* Molecular genetic profiling of Gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia. *Journal of Urology* **166**, 2171-7 (2001).
- Warrington, J.A., Nair, A., Mahadevappa, M., Tsyganskaya, M. Comparison of human adult and fetal expression and identification of 535 housekeeping / maintenance genes. *Physiological Genomics* **2**, 143-147 (2000).
- Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics* **1**, 80-83 (1945).
- Winzeler, E.A. *et al.* Functional Characterization of the *S. cerevisiae* Genome by Gene deletion and Parallel Analysis. *Science* **285**, 901-906 (1999).
- Wodicka, L. *et al.* A Genome-Wide Expression Monitoring in *Saccharomyces Cerevisiae*. *Nature Biotechnology* **15**, 1359-1367 (1997).

# Appendix A: Glossary

NOTE:

▲ = 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 ( $\sum(\text{PM-MM})/\text{pairs}$  in the average).

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

**Baseline Array:** An array designated as the baseline when being analyzed in comparison analysis with which the experimental array is compared to detect changes in expression. For example, if the baseline file is derived from a treated sample and the experiment from an untreated sample, all genes activated by the treatment will have decrease calls.

◆ **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 probe set in two different arrays are the same or not. 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.

**Comparative 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 the 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 ( $R=(\text{PM-MM})/(\text{PM+MM})$ ).

▲ **Empirical Algorithms:** The algorithms contained in 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 to the baseline array to detect changes in expression. For example, if the baseline file is derived from an untreated sample and the experiment from a treated sample, all genes activated by the treatment will have increase calls.

**Feature:** A single square-shaped probe cell on an array (another term for probe cell). A feature ranges in size from 18 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 discrete 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 that the data sampled is from a population that follows 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 1).

**Probe Pair:** Two features within a probe set (refer to Figure 1). 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 1).

**SAPE:** Streptavidin-phycoerythrin 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 5.0. 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.

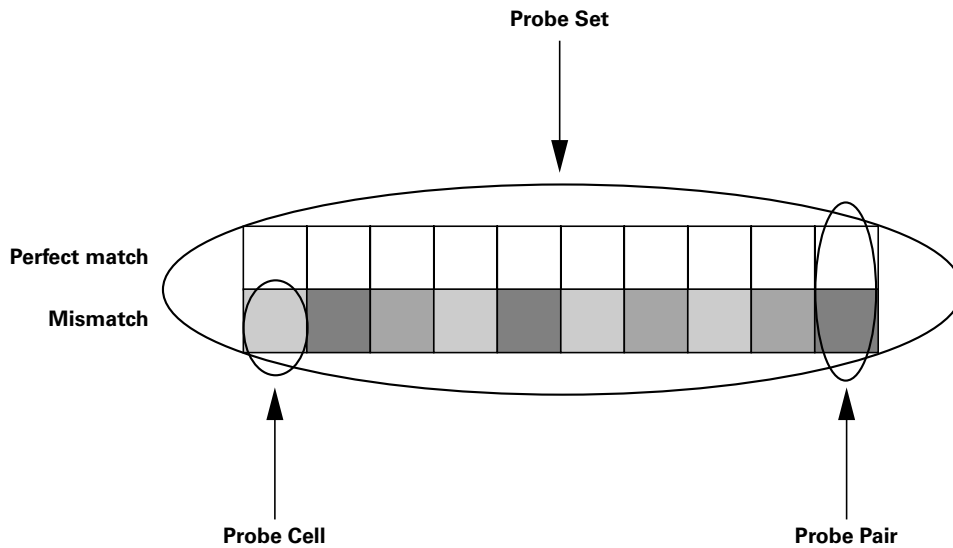


Figure 1

# 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
--- --- ---
----- 34568_g_at probes
----- transcript #9

```

`_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 (These are the only probe set extensions used in the HG-U133 Set)**

**\_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 that share some probes identically with two or more sequences. Rules for cross-hybridization were dropped in order to design the **\_x** probe sets.

**Probe Set Name Designations for Rat 230 Set (These are the only probe set extensions used in the Rat 230 Set)**

**\_a\_at:**

Designates probe sets that recognize multiple alternative transcripts from the same gene (on HG-U133 these probe sets have an “**\_s**” suffix).

**\_s\_at:**

Designates probe sets that share 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. Therefore, these probe sets may cross-hybridize in an unpredictable manner with other sequences.

**Probe Set Name Designations for Mouse 430 Set (These are the only probe set extensions used in the Mouse 430 Set)**

**\_a\_at:**

Designates probe sets that recognize multiple alternative transcripts from the same gene (on HG-U133 these probe sets have an “**\_s**” suffix).

**\_s\_at:**

Designates probe sets that share 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. Therefore, these probe sets may cross-hybridize in an unpredictable manner with other sequences.

## Appendix C: Microarray Suite Expression Defaults

### MAS 5.0 Expression Analysis Default Settings

Parameter	# Probe Pairs/Probe Set	
	16-20	11
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

### MAS 4.0 Expression Analysis Default Settings

Parameter	Value
SDT Multiplier	4.0*
Ratio Threshold	1.50
Ratio Limit	10.00
Pos/Neg Min	3.0
Pos/Neg Max	4.0
Pos Ratio Min	0.33
Pos Ratio Max	0.43
Avg Log Ratio Min	0.90
Avg Log Ratio Max	1.30
STP	3.0
CT Multiplier	<compute>
% Change Threshold	80
Inc/Dec Min	3.0
Inc/Dec Max	4.0
Inc Ratio Min	0.33
Inc Ratio Max	0.43
Dpos-Dneg Ratio Min	0.20
Dpos-Dneg Ratio Max	0.30
Avg Log Ratio Change Min	0.90
Avg Log Ratio Change Max	1.30

\* The default SDT Multiplier value is 4.0 for antibody-stained arrays. The default SDT Multiplier value for non-antibody-stained arrays is 2.0.



# Appendix D: File Types

## File Types in Microarray Suite

Experiment Data File Name	File Extension	Description
Experiment Information File	*.exp	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.
Report File	*.rpt	The report generated from the analysis output file (*.chp).
Experiment Information File	*.tif	A standard file format for graphic images. The Microarray Suite software exports graphic images in this file format.
Data File	*.txt, *.xls	A standard format for text files. The Microarray Suite software 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.

**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](http://www.affymetrix.com)

For research use only.  
Not for use in diagnostic procedures.

Part No. 701190 Rev. 3

©2002-2003 Affymetrix, Inc. All rights reserved. Affymetrix®, GeneChip®, , , , HuSNP®, Jaguar®, EASI™, MicroDB™, GenFlex®, Flying Objective™, CustomExpress™, CustomSeq™, and NetAffix™ are trademarks owned or used by Affymetrix, Inc. Products may be covered by one or more of the following patents and/or sold under license from Oxford Gene Technology: U.S. Patent Nos. 5,445,934; 5,744,305; 6,261,776; 6,291,183; 5,700,637; 5,945,334; 6,346,413; and 6,399,365; and EP 619 321; 373 203 and other U.S. or foreign patents. GeneArray® is a registered trademark of Agilent Technologies, Inc.

