Statistical Algorithms Description Document

TABLE OF CONTENTS

OVERVIEW

This document is a detailed reference guide for the *Statistical Algorithms* used in the analysis of GeneChip® expression data. The guide focuses on how they work, what calculations and approaches they comprise, and how the tunable parameters are designed. Additional references are provided for additional information.

GENECHIP® ARRAY DESIGN

It is important to understand how a GeneChip® array is designed when considering the most appropriate approaches for its analysis. A GeneChip® probe array consists of a number of probe cells where each probe cell contains a unique probe. Probes are tiled in probe pairs as a Perfect Match (PM) and a Mismatch (MM). The sequence for PM and MM are the same, except for a change to the Watson-Crick complement in the middle of the MM probe sequence. A probe set consists of a series of probe pairs and represents an expressed transcript.
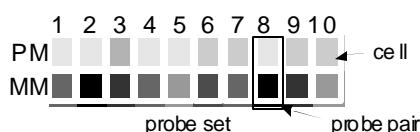
DATA OUTPUTS

The *statistical algorithms* provide the following data outputs:

Output	Descriptions
Signal	A measure of the abundance of a transcript.
Stat Pairs	The number of probe pairs in the probe set.
Stat Pairs Used	The number of probe pairs in the probe set used in the Detection call.
Detection	Call indicating whether the transcript was Present (P) or Absent (A), or Marginal (M).
Detection *p*-value	*p*-value indicating the significance of the Detection call.
Stat Common Pairs	The number of probe pairs in the probe sets from baseline and experimental arrays used in the Change call.
Change	Call indicating a change in transcript level between a baseline array and an experiment array [i.e. increase (I), decrease (D), marginal increase (MI), marginal decrease (MD), no change (NC)].
Change *p*-value	*p*-value indicating the significance of the Change call.
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 $\log_2$ ratio of 1 is the same as a Fold Change of 2.
Signal Log Ratio Low	The lower limit of the log ratio within a 95% confidence interval.
Signal Log Ratio High	The upper limit of the log ratio within a 95% confidence interval.

DATA PREPARATION

In this section we will discuss steps that occur *prior* to application of the new *statistical algorithms.*

A note about .CEL Files

The *Statistical Algorithms* begin with information contained in the .CEL file generated by Microarray Suite software. The .CEL files contain a captured image of the scanned GeneChip® array and calculations of the raw intensities for probe sets. The method for calculating individual cell intensities, thus generating the .CEL file, is not affected by the Statistical Algorithms. Therefore, it will not be discussed here.
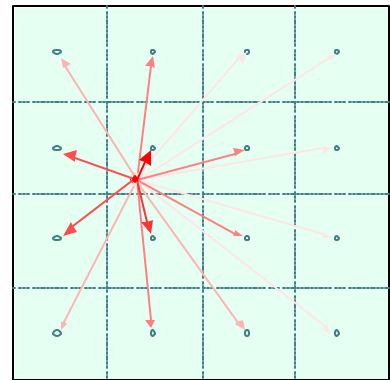
Masking

Masked probe pairs are excluded from all algorithms. For more information about "probe masking" see the Affymetrix® GeneChip® Expression Analysis Technical Manual (2001), *Section 4.1.17.*

Background Subtraction

A calculated background establishes a "floor" to be **subtracted** from each cell value.

Zone Values

- For purposes of calculating background values, the array is split up into K rectangular zones Z_k ($k = 1, …, K$, default $K = 16$).
- Control cells and masked cells are not used in the calculation.
- The cells are ranked and the lowest 2% is chosen as the background b for that zone (bZ_k).
- The standard deviation of the lowest 2% cell intensities is calculated as an estimate of the background variability n for each zone (nZ_k).

Smoothing Adjustment

To provide a smooth transition between zones, we compute distances from each cell on the chip to the various zone centers. A weighted sum is then calculated based on the reciprocal of a constant plus the square of the distances to all the zone centers. In more detail, if the distance d between the chip coordinate (x,y) and the center of the kth zone is d_k, we can calculate a weighting factor, which is related to the square of d (the relationship between w and d is illustrated in the graph next page). A small factor, *smooth*, is added to d^2 to ensure that the value will never be zero.

A GeneChip® array is divided into a number of equally spaced zones and an average background is assigned to the center of the zone, indicated by green circles. For each cell, the distance is calculated to the center of every zone. A weighting factor is then calculated as the reciprocal of the sum of a constant and the square of the distance. The colors of the arrows indicate the relative weights.
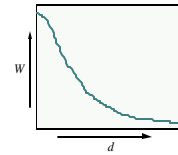
$$w_k(x,y) = \frac{1}{d_k^2(x,y) + smooth}$$
(default $smooth = 100$)

For every cell, we can use the sum of all the weights to all the zone centers to calculate the background b value to be used for cell x,y:

$$b(x,y) = \frac{1}{\sum\limits_{k=1}^{K} w_k(x,y)} \sum\limits_{k=1}^{K} w_k(x,y) bZ_k$$

Noise Correction

Now we want to compute an adjusted value that shifts intensities down by the local background. In order to do so, we must first ensure that the values would not become negative. Negative intensity values are problematic later in the calculations when log values are calculated.

For noise correction, a local noise value n based on the standard deviation of the lowest 2% of the background in that zone (nZ_k) is calculated and weighted for background values (just substitute $n(x, y)$ for $b(x, y)$ and nZ_k for bZ_k in the formula above).

Then a threshold and a floor are set at some fraction of the local noise value, so that no value is adjusted below that threshold. That is, for a cell intensity $I'(x, y)$ at chip coordinates (x, y), we compute an adjusted intensity.

$$A(x, y) = \max(I'(x, y) - b(x, y), NoiseFrac*n(x, y))$$
$$\text{where } I'(x, y) = \max(I'(x, y), 0.5)$$

NoiseFrac is the selected fraction of the global background variation.
(default *NoiseFrac* = 0.5)

EXPRESSION VALUE CALCULATION (SIGNAL)

The Signal value is calculated from the combined, background-adjusted, PM and MM values of the probe set. It represents the amount of transcript in solution.

Signal is calculated as follows:
1. Cell intensities are preprocessed for global background.
2. An ideal mismatch value is calculated and subtracted to adjust the PM intensity.
3. The adjusted PM intensities are log-transformed to stabilize the variance.
4. The biweight estimator (see Appendix I) is used to provide a robust mean of the resulting values. Signal is output as the antilog of the resulting value.
5. Finally, Signal is scaled using a trimmed mean.

Background-adjusted cell intensities ◊ Probe set Signal

Method

Background and Contrast Correction

Before we can proceed, we need to do the background subtraction as described in the Data Preparation section.

Ideal Mismatch (IM)

Used in Signal Calculations

The reason for including a MM probe is to provide a value that comprises most of the background cross-hybridization and stray signal affecting the PM probe. It also contains a portion of the true target signal. If the MM value is less than the PM value, it is a physically possible estimate for background, and can be directly used.

If the MM value is larger than the PM value, it is a physically impossible estimate for the amount of stray signal in the PM intensity. Instead, an idealized value can be estimated based on our knowledge of the whole probe set or on the behavior of probes in general. Specifically, we base this estimate either on the average ratio between PM and MM, or (if that measure is itself too small) a value slightly smaller than PM.

To calculate a specific background ratio representative for the probe set, we use the one-step biweight algorithm (T_{bi}), which is described in Appendix I. We find a typical log ratio of PM to MM that is simply an estimate of the difference of log intensities for a selected probe set. The biweight specific background (*SB*) for probe pair *j* in probe set *i* is:

$$SB_i = T_{bi}\left(\log_2(PM_{i,j}) - \log_2(MM_{i,j}) : j = 1, \ldots, n_i\right)$$

Throughout this text we use log base 2 exclusively.
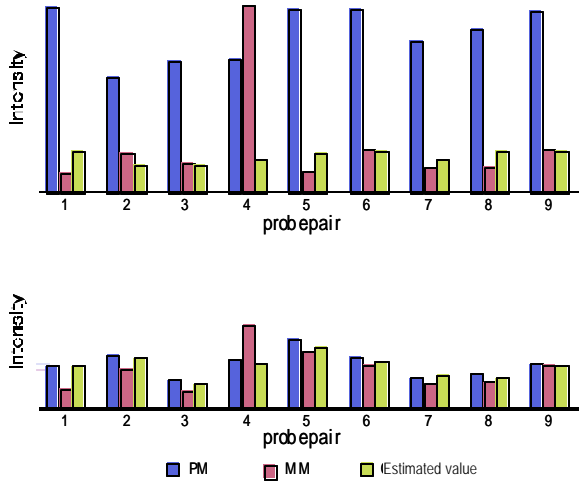
The blue bars represent PM and the red bars MM probes of a hypothetical probe set. In the top panel, most of the MM values are smaller than PM, so we use the MM directly. The yellow bars indicate the estimated value we would use if MM>PM.

For probe pair 4, the MM is larger than the PM, so it is not a useful value for estimating the stray signal component of PM. An imperfect, but useful resolution is to estimate a MM value that is typical for the probe set. For the overall probe set, the mean difference SB_i of the logs of PM and MM is large, so we can use it directly to estimate a value for MM 4. The yellow bar indicates the estimate.

In the second panel, SB_i is small, so we cannot base an accurate estimate for MM 4 on it. The best we can do is to calculate a value (indicated by the yellow bar) slightly less than PM.

If SB_i is large, then the values from the probe set are generally reliable, and we can use SB_i to construct the ideal mismatch *IM* for a probe pair if needed. If SB_i is small ($SB_i \leq contrast\tau$), we smoothly degrade to use more of the PM value as the ideal mismatch. The three cases of determining ideal mismatch *IM* for probe pair *j* in probe set *i* are described in the following formula:

Scale τ (tau) is the cutoff that describes the variability of the probe pairs in the probe set.

$$
IM_{i,j} = \begin{cases}
MM_{i,j}, & MM_{i,j} < PM_{i,j} \\[2em]
\dfrac{PM_{i,j}}{2^{(SB_i)}}, & MM_{i,j} \geq PM_{i,j} \ and \ SB_i > contrast\tau \\[2em]
\dfrac{PM_{i,j}}{2^{\left(\frac{contrast\,\tau}{1+\left(\frac{contrast\tau - SB_i}{scale\tau}\right)}\right)}}, & MM_{i,j} \geq PM_{i,j} \ and \ SB_i \leq contrast\tau
\end{cases}
$$

default *contrast* τ =0.03

default *scale* $\tau = 10$

The first case where the mismatch value provides a probe-specific estimate of stray signal is the best situation. In the second case, the estimate is not probe-specific, but at least provides information specific to the probe set. The third case involves the least informative estimate, based only weakly on probe-set specific data.

Probe Value and Signal Log Value

Given the ideal mismatch value, the formula for the probe value (PV) is fairly simple. To guarantee numerical stability, we use the formula:

$$V_{i,j} = \max(PM_{i,j} - IM_{i,j}, d) \quad \text{default } d = 2^{(-20)}$$

Now we calculate the probe value PV for every probe pair j in probeset i. n is the number of probe pairs in the probeset.

$$PV_{i,j} = \log_2(V_{i,j}), \ j = 1, \ldots, \ n_i$$

We then compute the absolute expression value for probe set i as the one-step biweight estimate (see Appendix I) of the i n adjusted probe values:

$$SignalLogValue_i = T_{bi}(PV_{i,1}, \ldots, PV_{i,n_i})$$

Scaled Probe Value

Note: the scaling (*sf*) and normalization factors (*nf*) computed in this section are reported by the software.

If the algorithm settings indicate scaling all probes sets or selected probe sets to a target we calculate a scaling factor (*sf*)

$$sf = \frac{Sc}{TrimMean\left(2^{SignalLogValue_i}, 0.02, 0.98\right)}$$

where *Sc* is the target signal (default *Sc* = 500) and the SignalLogValues in the SignalLogValue$_i$ set are the probe sets indicated in the algorithm settings. The *TrimMean* function here takes the average value of all observations after removing the values in the lowest 2% of observations and removing those values in the upper 2% of observations. If the algorithm settings indicate user defined scaling, then *sf* = user defined value.

The reported value of probe set i is:

$$ReportedValue(i) = nf * sf * 2^{(SignalLogValue_i)}$$

where *nf* = 1 for absolute analysis and is computed as follows for a comparison analysis.

If the algorithm settings indicate user defined normalization, then *nf* = user specified normalization.

Otherwise, the algorithm settings either indicate normalizing all or selected probe sets:

$$nf = \frac{TrimMean\left(SPVb_i, \ 0.02, 0.98\right)}{TrimMean\left(SPVe_i, \ 0.02, 0.98\right)}$$

where SPV$b_{[i]}$ is the baseline signal, and SPV$e_{[i]}$ is the experiment signal (scaled-only) and i defines the probe sets selected by the user.

This is reported as Signal.

Since comparison analyses are done on matched probe pairs, the individual probe pair values are also modified by this scaling factor.   The scaled probe value *SPV* is

$$SPV_{i,j} = PV_{i,j} + \log_2(nf * sf)$$

These values are used in computing the log-ratio in comparison analysis.

Performance on corrupted data

The algorithms were developed on a very clean data set, where every precaution was taken to ensure that the data was of high quality. It is important to validate the performance on poor quality data. Good performance on poor quality data will be an indication of how well the algorithms will perform under real-world situations, where the data is less than perfect. To generate data with a controlled amount of noise, several .CEL files were used and an increasing amount of data was replaced with random numbers. The random numbers were chosen uniformly within the typical intensity range (0 to 46,000). The correlation coefficient between the true concentration and the output from the newly generated corrupted data set was then plotted as a function of the proportion of uncorrupted data. When the proportion of uncorrupted data is 1 (the original data set), we see a correlation coefficient near 1.0 (even the original data set has some noise). As the proportion of good data decreases, we expect the predictive power of the data with respect to concentration to decrease. This loss of predictive power is reflected by a lower correlation coefficient.

Performance of methods against bad data (human)
(corrupt data chosen from uniform[0…46000])
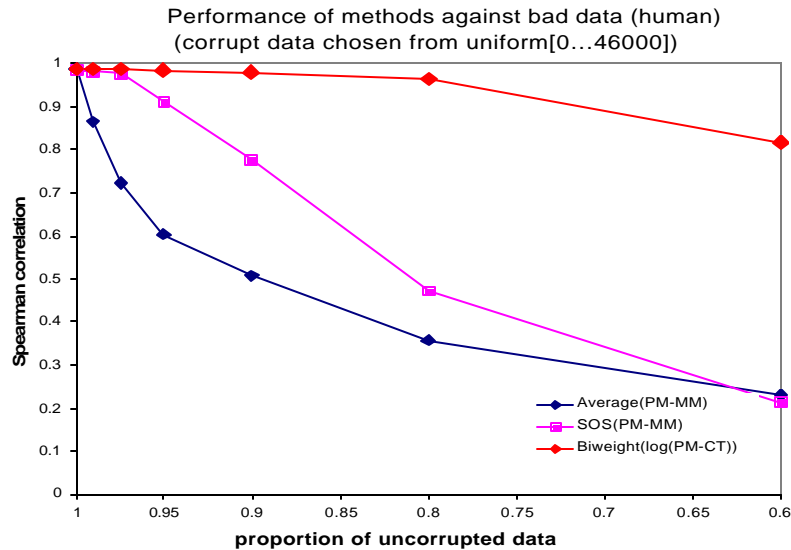
Performance against bad data (human). This graph illustrates the robustness against bad data obtained by using this analysis strategy. The data in several .CEL files from the human array HG-U95A are degraded by substituting random numbers between 0 and 46,000. The intact data set is indicated by 1 on the x-axis. As the amount of intact data decreases, the average values start deviating from the original. The data from the Signal Calculation Algorithm implemented in MAS 4.0 also starts deviating, while the biweight algorithm remains accurate even after 20% of the data were corrupted.

For averages (blue line) the correlation coefficient drops rapidly indicating that the results quickly become less accurate. This is not surprising, because an average is not robust against outliers. The AvgDiff algorithm used in MAS 4.0 (pink line) is more robust because it discards some outliers. However, this strategy loses power when the data is increasingly corrupt, because it has only a small ability to identify outliers when much of the data is corrupt. The Signal algorithm implemented in MAS 5.0 is very robust against corrupted data and the results remain well-correlated even when as much as 20% of the data is corrupt. Naturally, additional noise never improves the quality of the data, and so does degrade the results, but robustness provides a safety net against corrupted data completely destroying the utility of an array.

RELATIVE EXPRESSION VALUES (LOG RATIOS)

By doing a direct cell to cell comparison between probe sets on two arrays, systematic probe effects will be canceled out. Probe effects refer to the inherent differences in the hybridization efficiency of different probes, which is a source of variation in sampling signals from different sequences present at the same concentration, even when physically linked in the same nucleic acid polymer. Calculating the ratio of signal for the same probe on two different arrays effectively cancels the intrinsic affinity factor for that sequence.
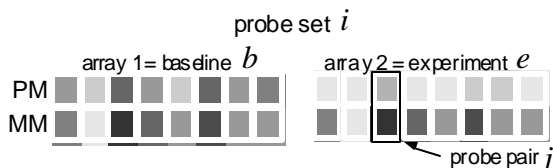
Adjusted cell intensities (Baseline and Experiment) ◊ Log Ratio, Log Ratio Low and Log Ratio High

Method
The Signal Log Ratio calculation is an extension of the Signal calculation.
1. Scale the baseline and experiment.
2. Correct for probe pair bias.
3. Calculate the signal log ratio.

The values of $\log_2$(PM-IM) plotted for a series of probe pairs over a concentration range (low concentration to the left, increasing to the right). Although the probes respond near-linearly, their affinities are slightly different. This is termed the probe affinity effect.
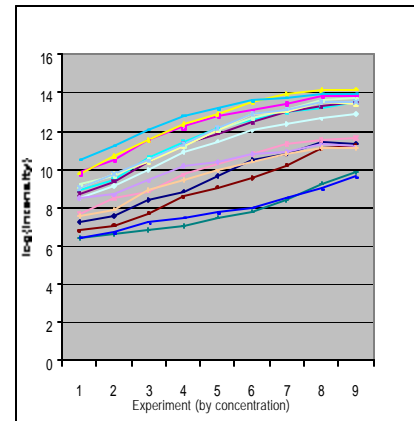
Nomenclature

Comparative experiments are always run on two arrays, one assigned baseline b and the other experiment e. Here we show the same probeset i from two arrays.

Signal Log Ratio
Once we have a scaled probe value each probe pair (SPV is calculated in the previous section), we can calculate the Signal log ratio using the biweight algorithm (see Appendix I). The probe log ratio PLR is calculated for probe pair j in probeset i on both the baseline b and experiment e arrays:

$$PLR_{i,j} = {}_e SPV_{i,j} - {}_b SPV_{i,j}$$

If we have the probe log ratios PLR we can use the biweight algorithm to calculate the SignalLogRatio (see Appendix I for a description of the biweight algorithm).

$$SignalLogRatio_i = T_{bi}(PLR_{i,1}, \ldots, PLR_{i,n_i})$$

From the biweight calculation, we can also determine the 95th confidence interval and report the results as Log Ratio Low and Log Ratio High. Thus, the relative expression for a target in two samples being compared is estimated by calculating the average of the log (ratios) for each corresponding probe pair in the probe sets.

Fold Change

Previous versions of the Affymetrix® Microarray Suite software communicated the relative signal level between the same probe set on two different arrays as a fold-change ratio that was signed depending on the direction of the change in an ordered pair. It is straight- forward to convert from the LogRatio metrix to the older Fold-Change value.

Fold Change from log ratios is calculated as follows:
Corrected formula

$$FoldChange_i = \begin{cases} 2^{SignalLogRatio_i} & SignalLogRatio_i \geq 0 \\ (-1) * 2^{-SignalLogRatio_i} & SignalLogRatio_i < 0 \end{cases}$$

SINGLE ARRAY ANALYSIS (DETECTION CALLS)

A detection call answers the question: "Is the transcript of a particular gene Present or Absent?" In this context, Absent means **that the expression level is below the threshold of detection**. That is, the expression level is not **provably** different from zero. In the case of an uncertainty, we can get a Marginal call. *It is important to note that some probe-sets are more variable than others*, and the minimal expression level **provably** different from zero may range from a small value to very large value (for a noisy probe-set). The advantage to asking the question in this way without actual expression values is that the results are easy to filter and easy to interpret. For example, we can imagine that we may only want to look at genes whose transcripts are detectable in a particular experiment.

The statistical significance, or *p*-value, of a result is the probability that the observation in a sample occurred merely by chance under the null hypothesis. The null hypothesis is that the target is absent (zero effect on the probes). For example, a *p*-value of 0.005 means that less than 5 out of 1000 probe sets for absent transcripts will be called present based on the distribution of intensity within the corresponding probe sets that is equally or less likely occur by chance. In detection, the smaller the *p*-value, the more significant the results suggesting that the gene may be present.

Raw cell intensities ◊ Absent, Present, or Marginal call plus *p*-values

Method
There are four steps to the method:
1. Remove saturated probe pairs and ignore probe pairs wherein PM ~ MM + *tau*
2. Calculate the discrimination scores. (This tells us how different the PM and MM cells are.)
3. Use Wilcoxon's rank test to calculate a significance or *p*-value. (This tells us how confident we can be about a certain result.)
4. Compare the *p*-value with our preset significance levels to make the call.

Saturation
If a mismatch cell is saturated $MM \geq 46000$, the corresponding probe pair is not used in further computations. We also discard pairs where PM and MM are within *tau* of each other.

If all probe pairs in a unit are saturated, we report the gene as detected and set the *p*-value to 0.
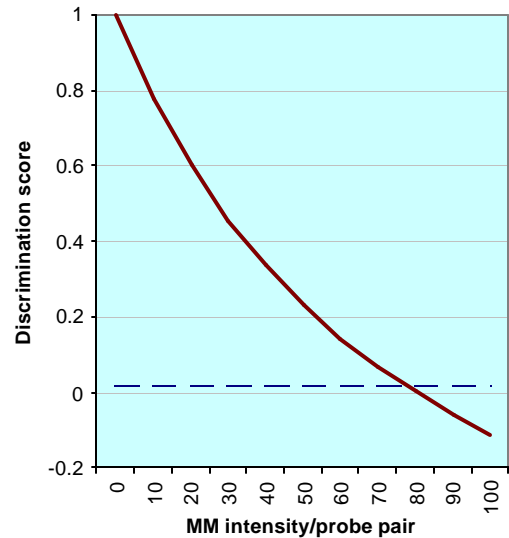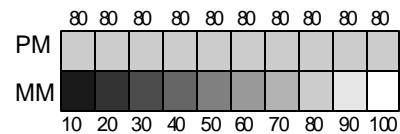
In this hypothetical probe set the PM intensity is 80 and the MM intensity for each probe pair increases from 0 to 100. The discrimination score offers a smooth function that decreases as the MM intensity increases. In other words, as the intensity of the MM increases our ability to discriminate between the PM and MM decreases. Note that the value becomes negative when MM>PM.

The broken line indicates the threshold *t*.

Discrimination Score

The discrimination score [R] is a relative measure of the difference between the PM and MM intensities. The discrimination score for the ith probe pair is:

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i}$$

We use t (default $t = 0.015$), a small threshold between 0 and 1 as a small significant difference from zero. If the median $(R_i) > \tau$, we can reject the hypothesis that PM and MM are equally hybridizing to the sample. We can make a detection call based on the strength of this rejection (the p-value).
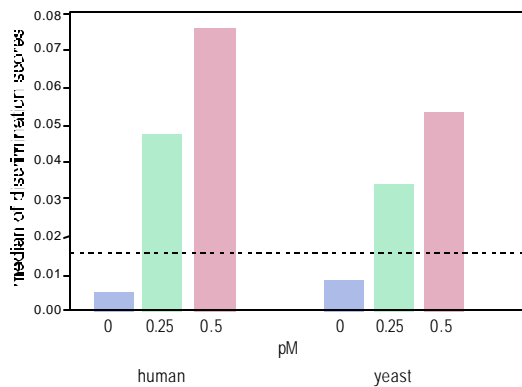
Increasing the threshold t can reduce the number of false detected calls, but may also reduce the number of true detected calls.

Computing p-values:

The one-sided Wilcoxon's Signed Rank Test (Appendix II) is used in the Call algorithms and is used to calculate the p-values for the null hypothesis:

H_0: median$(R_i - t) = 0$ versus the alternative hypothesis:

H_1: median$(R_i - t) > 0$

$$r = \frac{PM - MM}{PM + MM}$$

$$\log_2 \frac{1+r}{1-r}$$

$$= \log_2 \frac{1+\frac{PM-MM}{PM+MM}}{1-\frac{(PM-MM)}{(PM+MM)}}$$

$$= \log_2 \frac{PM+MM+PM-MM}{(PM+MM)-(PM-MM)}$$

$$= \log_2 \frac{2*PM}{2*MM}$$

$$= \log_2(PM) - \log_2(MM)$$

Determining the default t. The discrimination scores of 8960 probe pairs spiked at known concentrations on the human U95Av2 array and yeast S98 array were examined. The default t value was selected to fall between the discrimination ratio of spikes at zero and 0.25pM concentration.

Making the call

We set two significance levels a_1 and a_2 [(*)] such that

$$0 < a_1 < a_2 < 0.5$$

default $a_1 = 0.04$ (16-20 probe pairs)
default $a_2 = 0.06$ (16-20 probe pairs)

Present (detected)	$p < a_1$
Marginal	$a_1 = p < a_2$
Absent (undetected)	$p = a_2$

Significance levels a_1 and a_2 define cut-offs of p-values for making calls.

Reducing the significance level a_1 can reduce the number of false Detected calls and reduce the number of true Detected calls.

Increasing the significance level a_2 can reduce the number of false undetected calls and reduce the number of true undetected calls.

[(*)] Although in an ideal world, the distribution of probes exactly matches the statistical assumptions, in practical cases the distribution is not precise. Our thresholds therefore do not result in the exact false-positive rate specified by the alpha parameter when the target is absent. This rate will vary with extraneous factors, including the number of probe-pairs used.

COMPARATIVE ANALYSIS (COMPARISON CALLS)

A comparative call answers the question: "Does the expression level of a transcript on one chip (experiment) change (significantly) with respect to the other chip (baseline)?" The possible distinct answers are Increase, Marginally Increased, No Change detected, Marginally Decreased, and Decrease. As with detection calls, No Change **means the difference is below the threshold of detection**. That is, the difference is not **provably** different from zero. *It is important to note that some probe-sets are more variable than others*, and the minimal expression difference **provably** different from zero may range from a small value to very large value (for a noisy probe-set, or for low concentrations). Note that saturated probe pairs are excluded from the computation of comparative calls. If all probe pairs of a probe set are saturated, we report that no comparative call can be made.

We attempt to find changes in expression level by examining changes in the intensities of both PM and MM probes between experiments. The differences in PM and MM in both experiments and differences between PM and background in both experiments are examined using a non-parametric Wilcoxon rank test to look for significant differences.

Differences between PM and MM and Differences between PM and Background (Experimental and Baseline) ◊ Change call and *p*-value

Saturation

If one of the four cells (PM and MM in Baseline and PM and MM in Experiment) is saturated (PM or $MM \geq 46000$), the corresponding probe pair is not used in further computations. The number of discarded cells can be determined from the Stat Common Pairs parameter.

Quantities Used in Comparative Calls

For a probe set of n probe pairs, we form two n-dimensional vectors for comparative calls.

$$q = (q_1, \ldots, q_n) \text{ and } z = (z_1, \ldots, z_n)$$

The component q_i is the difference between Perfect Match intensity PM_i and Mismatch intensity MM_i for the ith probe pair:

$$q_i = PM_i - MM_i, \ (i = 1, \ldots, n),$$

and the component z_i is the difference between the Perfect Match intensity PM_i and background level b_i:

$$z_i = PM_i - b_i, \ (i = 1, \ldots, n)$$

b is the background and is calculated the same as for Data Preparation ($b(x,y)$) .

Using both q and z can produce better empirical call results than using only one of them.

Balancing Factors
Note: The factors computed in this section are not reported by the software.

The distributions of q and z over all probe pairs in an experiment are slightly and subtly different from each other, and are different between two experiments. We therefore provide a balancing factor for each type of data to correct some of this difference. (The distributions of q and z are also not identical to the Signal distribution, and therefore the scaling factor used for signal is not used here).

Vectors q and z have two different balancing factors:

$e[i]$ is a modified average of q_j for the probe pairs of transcript i, and is calculated as the average of all q values within three standard deviations from the average q for transcript i. We calculate a global balancing factor using the trimmed mean of e[i] over <u>all</u> transcripts i:

$$sf = \frac{Sc}{trimmedMean\left(e\left[i\right],0.02,0.98\right)}$$

If $trimmedMean(e[i],0.02,0.98) \leq 0$, the scaling factor can be calculated as:

$$sf = \frac{Sc}{trimmedMean\left(\max\left(e[i],0\right),0.02,0.98\right)}$$

These values are calculated for both the experiment (E) and the baseline (B) arrays. Now the primary balancing factor nf can be calculated:

$$nf = \frac{sfE}{sfB}$$

A second, primary balancing factor nf_2 is calculated for z. The calculations are exactly the same as for q, except only z values are used. These values are calculated for both the experiment (E) and the baseline (B) arrays to give:

$$nf_2 = \frac{sf_2E}{sf_2B}$$

The two balancing factors are combined to match the distributions of q and z over the whole signal range.

However, any calculated balancing factor is only an approximation to the true differences between the distributions. To allow for small differences between the distributions not covered by the balancing factor, we will use a range of balancing factors. We straddle the

true balancing function by using three different balancing factors *f[k]* for *q*, as well as three different balancing factors $f_2[k]$ for z ($k = 0, 1, 2$). They are defined as:

$$f[0] = nf * d \qquad f[1] = nf \qquad f[2] = \frac{nf}{d}$$

$$f_2[0] = nf_2 * d \quad f_2[1] = nf_2 \quad f_2[2] = \frac{nf_2}{d}$$

where $d \leq 1$, (default = 1.1).

<u>Special Case</u>
If the algorithm settings indicate a user defined balancing factor and the factor is not equal to 1 then, $nf = nf_2$ = user defined normalization factor * *sfE* /*sfB*

where *sfE* is the experiment *sf* and *sfB* is the baseline *sf* as described in the Scaled Probe Value section.

Wilcoxon Signed Rank Test (Appendix I)

For every unit, we can form the ($2n$)-dimensional vector *qB* in the baseline and vector *qE* in the experiment. For $k = 0, 1, 2$, three vectors were formed:

$$v[k][i] = f[k] * qE[i] - qB[i]$$

$$v[k][i + n] = C * (f_2[k] * zE[i] - zB[i])$$

$$(i = 1, \ldots, n; \ k = 0, 1, 2)$$

$$\text{default } C = 0.2$$

and three, one-sided *p*-values p[k] from the signed rank tests of the null hypothesis

$$H_0 : \text{median}(v[k]) = 0$$

versus the alternative hypothesis

$$H_1 : \text{median}(v[k]) > 0.$$

We use significance levels γ_1 and γ_2 to make calls. γ_1 is the small significance level for comparative calls of a unit (an interpolation of $\gamma_1 H$ and $\gamma_1 L$), and is 0.0025 for default values of $\gamma_1 H$ and $\gamma_1 L$. γ_2, the large significance level for comparative calls of a unit (an interpolation of $\gamma_2 H$ and $\gamma_2 L$), and is 0.003 for default values of $\gamma_2 H$ and $\gamma_2 L$. γ_1 and γ_2 should satisfy the relationship $0 < \gamma_{1L} < \gamma_{2L} < 0.5$ and $0 < \gamma_{1H} < \gamma_{2H} < 0.5$.

The γ_1 value is a linear interpolation of $\gamma_1 L$ and $\gamma_1 H$. Similarly γ_2 is derived from $\gamma_2 L$ and $\gamma_2 H$. 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. However, this feature is not used by default, because the defaults are set as $\gamma_1 L = \gamma_1 H$ and $\gamma_2 L = \gamma_2 H$
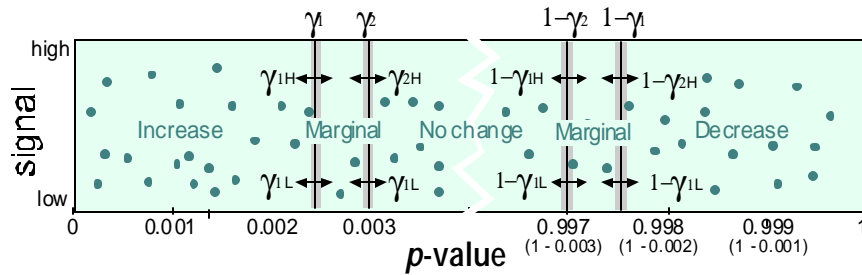
A representation of a range of *p*-values for a data set. The Y-axis is the probe set signal. The arrows on the vertical bars represent the adjustable **g** values. The **g₁** value is a linear interpolation of **g₁L** and **g₁H**. Similarly **g₂** is derived from **g₂L** and **g₂H**.

We use the "critical" *p*-value as our output *p*-value.

The critical *p*-value, *p* is the most conservative in making increase and decrease calls. It is defined by the following formula:

$$p = \max(p0, p1, p2) \quad \text{if } p0 < 0.5, \, p1 < 0.5 \text{ and } p2 < 0.5$$
$$p = \min(p0, p1, p2) \quad \text{if } p0 > 0.5, \, p1 > 0.5 \text{ and } p2 > 0.5$$
$$p = 0.5 \qquad\qquad\qquad \text{otherwise}$$

The cut off *p*-values may appear very small if you are used to the usual significance value of 0.05. This is due to two reasons:

Due to the multiple measurement problem, we require smaller *p*-values to ensure that we do not get too many false calls for the whole collection of transcripts. This is similar to a Bonferroni correction, but we determine the value empirically.

The *p*-value produced by this calculation is also an over estimate of significance, because we are using two values, PM-MM and PM-background, that are not truly independent measures. Further, the null distribution does not exactly describe the empirical situation, and so critical thresholds may change with the number of probes.

However, the purpose of the *p*-values produced here is to rank results in order of significance, so the absolute *p*-value is not important.

The margins can also be represented as follows (16-20 probe pairs):

increase
$$\begin{cases} p[0] < g_1 \\ p[1] < g_1 \\ p[2] < g_1 \end{cases}$$

marginally increase
but not increase
$$\begin{cases} p[0] < g_2 \\ p[1] < g_2 \\ p[2] < g_2 \end{cases}$$

$g_1 H = 0.0025$
$g_1 L = 0.0025$

decrease
$$\begin{cases} p[0] > 1 - g_1 \\ p[1] > 1 - g_1 \\ p[2] > 1 - g_1 \end{cases}$$

$g_2 H = 0.003$
$g_2 L = 0.003$

marginally decrease
but not decrease
$$\begin{cases} p[0] > 1 - g_2 \\ p[1] > 1 - g_2 \\ p[2] > 1 - g_2 \end{cases}$$

It is recommended to use two or more identical sample replicates to adjust pa rameters.

All comparative call thresholds should ideally be empirically determined from the false-positive rate. Set an allowed false change call rate, e.g., 0.01. Sort the *p*-values and ?nd the signi?cant level that give this rate of increasing or decreasing calls. Do not set the allowed false change call rate too small, because it will be very difficult to make increasing or decreasing calls for other experiments where these calls are reasonable.

If none of the above conditions are satisfied, we make a No-Change call.

Adjusting Parameters

- Decreasing g_1 ($\gamma_1 L$ and $\gamma_1 H$) can reduce the number of false Increase and Decrease calls, but can also reduce the number of true Increase and Decrease calls.

- Increasing g_2 ($\gamma_2 L$ and $\gamma_2 H$) can reduce the number of false No-Change-detected calls, but can also reduce the number of true No-Change-detected calls.

- Increasing the perturbation parameter d can increase the number of true No-Change-detected calls, but can also increase the number of false No-change-detected calls.

REFERENCES

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).

Hubbell, E., Liu, W.M., Mei, R. Robust Estimators for Expression Analysis. In Preparation (2002).

Liu, W.M., Mei, R., Bartell, D.M., Di, X., Webster, T.A., Ryder, T. Rank-based Algorithms for Analysis of Microarrays *Proceedings SPIE* **4266,** 56-67 (2001).

Liu, W.M., Mei, R., Di, X., Ryder, T., Hubbell, E., Dee, S., Webster, T.A., Harrington, C.A., Ho, M., Bald, J., Smeekens, S. Analysis of High Density Expression Microarrays with signed-rank call algorithms. In preparation (2001).

Wilcoxon, F. Individual Comparisons by Ranking Methods *Biometrics* **1**, 80-83 (1945).

Roderick, J.A., Little, D., Rubin, B. *Statistical Analysis with Missing Data* Wiley, New York (1987).

APPENDIX I

One-Step Tukey's Biweight Algorithm

Purpose
There are several stages in the algorithms in which we want to calculate an average. The biweight algorithm is a method to determine a robust average unaffected by outliers.
1. First, the median is determined to define the center of the data.
2. Then the distance of each data point from the median is determined. This distance is then used to determine how much each value should contribute to the average. For example, outliers that are far away from the median should contribute less to the average.

The full biweight algorithm iterates through a process of calculating the estimate and then reweighting the observations until there is no further change. We found that the first step of the biweight iteration provides the most useful increase in quality.

Method
The one-step biweight algorithm begins by calculating the median M for a data set with n values. In the Signal measurement, this data set consists of the log (PM-IM) probe values of a probe set with n probe pairs.
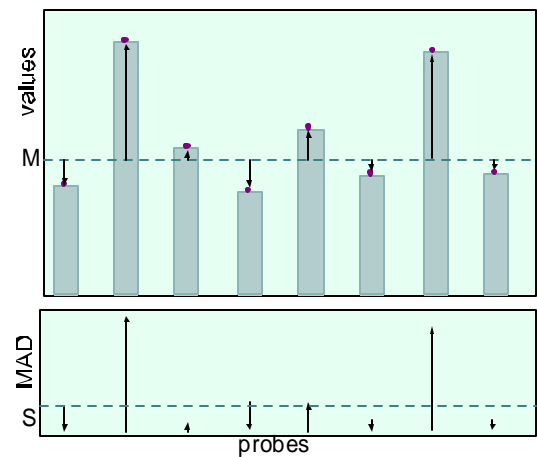
The top box represents a series of values such as intensities. The broken line represents the median M. The arrows indicate the distance of the actual values from M.

Next, we calculate the absolute distance for each data point from the median. We then calculate S, the median of the absolute distances from M. The Median Absolute Deviation, MAD, is an initial measure of spread.

The arrows are re-plotted in the bottom box. The direction of the arrows (deviation) is not important, so we determine the absolute values. The next step is to determine the Median Absolute Deviation (MAD). This is a robust measure of the spread of a set of values. Unlike standard deviation, it is not thrown off by outliers.

For each data point i, a uniform measure of distance from the center is given by:

$$u_i = \frac{x_i - M}{cS + e}, i = 1, \ldots, n$$

c is a tuning constant (default $c = 5$).

e is a very small value used to avoid zero values in the division (default $e = 0.0001$).

Weights are then calculated by the bisquare function:

$$w(u) = \begin{cases} \left(1-u^2\right)^2, & |u| \le 1 \\ 0, & |u| > 1 \end{cases}$$

For each point, the weight w is reduced by a function of its distance from the median so outliers are effectively discounted by a smooth function. When values are very far from the median, their weights are reduced to zero.

The corrected values can now be calculated by using the one-step w-estimate (a weighted mean):

$$T_{bi} = \frac{?\ w(u)x_i}{?\ w(u)}$$

"The performance of one-step-from-the-median w-estimators is essentially as good as that of fully iterated M-estimates" (Hoaglin, Mosteller, Tukey). We follow their suggestion and save computation by using this w-estimate.

The t distribution can then be used to determine a confidence interval length.

Confidence Intervals

An additional benefit of the biweight algorithm is that we can calculate confidence limits using standard statistics.
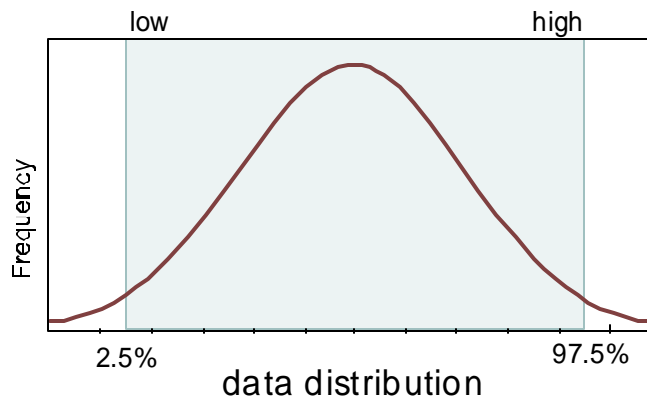
The first step is to calculate the measure of uncertainty for the biweight s_{bi}:

$$s_{bi} = \frac{\sqrt{n}\sqrt{\sum_{|u_i|<1}(x_i - T_{bi})^2(1-u^2)^4}}{\left|\sum_{|u_i|<1}\left(1-u_i^2\right)\left(1-5u_i^2\right)\right|}$$

An approximate 95% confidence interval is then computed as

$$T_{bi}\left(x_1, x_2, \ldots, x_n\right) \pm t_{df}^{(0.975)}\frac{s_{bi}}{\sqrt{n}}$$

where $t_{df}^{(0.975)}$ is the 97.5$^{\text{th}}$ percentile for the t distribution with the number of degrees of freedom set equal to $\max\left(0.7*(n\text{-}1),1\right)$.

APPENDIX II

One-sided Wilcoxon's Signed Rank Test

Wilcoxon's signed rank test possesses many good properties of nonparametric statistical methods.

1. It does not assume a normal data distribution.
2. It is robust and insensitive to outliers.
3. It can be used on raw data without much handling.

The signed rank test applies to two paired data sets. For example PM_i and MM_i probes in a probe set or $(PM_i\text{-}MM_i)_{Baseline}$ and $(PM_i\text{-}MM_i)_{Experiment}$ from paired data sets. To demonstrate signed rank test procedure, we consider two paired data sets: baseline b and experiment e with n probe pairs, $b = (b_1, ..., b_n)$ and $e = (e_1, ..., e_n)$ and calculate $d_i = e_i - b_i$ for every probe pair.
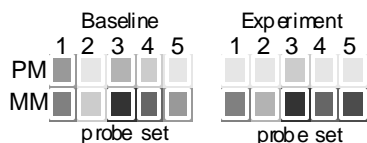
Uninformative (Tied) Probe Pairs

We first calculate the absolute differences $|d_i|$ for all pairs of data. We exclude the probe pairs whose $d_i = 0$ from further computation. If all differences are zero, we output 0.5 as the one-sided p-value.

Rank Sum

The steps for calculating the rank sum for non-zero differences are as follows:

Start values $e_i - b_i$	Convert to absolute values	Sort in increasing order	Assign ranks	Place ranks back in their original order	Their signed ranks
$d_1 = 2$	$|d_1| = 2$	$a_4 = 0.5$	$r_4 = 1.5$	$r_1 = 4.5$	$S_1 = 4.5$
$d_2 = 1$	$|d_2| = 1$	$a_5 = 0.5$	$r_5 = 1.5$	$r_2 = 3$	$S_2 = 3$
$d_3 = -2$	$|d_3| = 2$	$a_2 = 1$	$r_2 = 3$	$r_3 = 4.5$	$S_3 = -4.5$
$d_4 = 0.5$	$|d_4| = 0.5$	$a_1 = 2$	$r_1 = 4.5$	$r_4 = 1.5$	$S_4 = 1.5$
$d_5 = 0.5$	$|d_5| = 0.5$	$a_3 = 2$	$r_3 = 4.5$	$r_5 = 1.5$	$S_5 = 1.5$

Then we form the sum of positive signed ranks. In our example,
$S = s_1 + s_2 + s_4 + s_5 = 10.5$.

Confidence Values

The changes in probe set values may vary greatly, from very small to very large. For example, our confidence in the change will be low if the changes are small or close to background. Large, significant changes will have high confidence. The goal is to be able to set a confidence level and give the final call based on whether or not the confidence level exceeds a certain threshold. This will allow us to only accept calls in which we have confidence, at the risk of missing small changes that may be real, but have a low confidence threshold. Conversely, we may be interested in finding as many changes as possible at the risk of including changes that may be accidental.

Fortunately, an advantage of the one-sided Wilcoxon's Signed Rank Test is that there are well-known methods for calculating p-values. The formulas are fairly complex, but not critical to understanding how the basic method works. We use two different methods: one for large probe sets and one for small probe sets.

Small Probe Sets

When the number of probe pairs n is small ($n < 12$), we can simply enumerate all the possible outcomes and compute the p-value directly. In this case, we apply signs to ranks r_i ($i = 1, 2, \ldots, n$) in every possible way, calculate the sum of positive ranks and denote this sum by S_j ($j = 1, \ldots, 2^n$).

$$p(S) = 2^{-n} \sum_{j=1}^{2^n} u\left(S_j > S\right) + 0.5u\left(S_j = S\right) \quad \begin{array}{ll} \text{if } S_j > S & u(S_j > S) = 1 \\ \text{if } S_j \leq S & u(S_j > S) = 0 \\ \text{if } S_j = S & u(S_j = S) = 1 \\ \text{if } S_j \neq S & u(S_j = S) = 0 \end{array}$$

u() is the characteristic function; when the argument is a logical expression, it is one if the argument is true, and it is zero if the argument is false; when the argument is a numeric expression, it is one if the argument is positive, and it is zero otherwise. Since these assignments of ranks are equally probable, we simply need to count the number of instances in which they are as large as our observed value.

In our example, all possible signed ranks and the sum of positive ranks S_j are listed (see table). Since the order of these ranks does not matter, we use the ascending order of their absolute values in the table and denote them by s'_j.

Random Signed Ranks for *p*-value Evaluation

J	s'_1	s'_2	s'_3	s'_4	s'_5	S'_j
1	-1.5	-1.5	-3	-4.5	-4.5	0
2	1.5	-1.5	-3	-4.5	-4.5	1.5
3	-1.5	1.5	-3	-4.5	-4.5	1.5
4	-1.5	-1.5	3	-4.5	-4.5	3
5	-1.5	-1.5	-3	4.5	-4.5	4.5
6	-1.5	-1.5	-3	-4.5	4.5	4.5
7	1.5	1.5	-3	-4.5	-4.5	3
8	1.5	-1.5	3	-4.5	-4.5	4.5
9	1.5	-1.5	-3	4.5	-4.5	6
10	1.5	-1.5	-3	-4.5	4.5	6
11	-1.5	1.5	3	-4.5	-4.5	4.5
12	-1.5	1.5	-3	4.5	-4.5	6
13	-1.5	1.5	-3	-4.5	4.5	6
14	-1.5	-1.5	3	4.5	-4.5	7.5
15	-1.5	-1.5	3	-4.5	4.5	7.5
16	-1.5	-1.5	-3	4.5	4.5	9
17	1.5	1.5	3	-4.5	-4.5	6
18	1.5	1.5	-3	4.5	-4.5	7.5
19	1.5	1.5	-3	-4.5	4.5	7.5
20	1.5	-1.5	3	4.5	-4.5	9
21	1.5	-1.5	3	-4.5	4.5	9
22	1.5	-1.5	-3	4.5	4.5	10.5
23	-1.5	1.5	-3	4.5	4.5	10.5
24	-1.5	1.5	3	-4.5	4.5	9
25	-1.5	1.5	3	4.5	-4.5	9
26	-1.5	-1.5	3	4.5	4.5	12
27	1.5	1.5	3	4.5	-4.5	10.5
28	1.5	1.5	3	-4.5	4.5	10.5
29	1.5	1.5	-3	4.5	4.5	12
30	1.5	-1.5	3	4.5	4.5	13.5
31	-1.5	1.5	3	4.5	4.5	13.5
32	1.5	1.5	3	4.5	4.5	15

The table shows the 32 possible results for the five probe ranks shown (2^5 possible ways of assigning signs to each rank).

All signed ranks above 10.5 are given a weight of 1 (there are five in Table 1) and items with signed ranks equal 10.5 are given a rank of .5 (there are four in Table 1).

In our example $p(10.5) = \dfrac{(1*5)+(0.5*4)}{32} = 0.21875$

Large Probe Sets

When the number of probe pairs n is large (in our implementation, $n \geq 12$), we use the asymptotic approximation. The statistic S' is considered to have a standard normal distribution with mean 0 and variance 1, where:

NOTE: Under certain circumstances "Fisher's permutation test" may have more power than Wilcoxon rank. We elected to use the same test across all numbers of probes to be consistent, even if it has lower power than the best possible test.

$$S' = \frac{S - \dfrac{n(n+1)}{4}}{\sqrt{n(n+1)*\dfrac{2n+1}{24} - vt}}$$

where *vt* is a term modifying the variance for ties. The formula for *vt* is:

$$vt = \sum_{k=1}^{t} \frac{b_k(b_k^2 - 1)}{48}, k = 1:t$$

where t is the number of tie groups; b_k is the number of ties in the tie group k. The one-sided *p*-value is

$$p(S) = 1 - f(S'),$$

where $f(S')$ is the standard normal cumulative distribution function.

APPENDIX III

Noise (Q) Calculation

The calculation of the Q value is provided here for completeness. It is not used anywhere in the statistical algorithm. However, since it is based on pixels it provides a useful quality measure of how well the grid was placed on the array to calculate the .CEL file. All other calculations are derived from the .CEL file.

NOTE: The noise (Q) is calculated from pixel values in the DAT file.

Q, the noise for a given probe array hybridization, is calculated by taking the average (over all the cells used in background computation) of the following value in each cell: standard deviation of the pixel intensity (*st dev$_i$*) divided by the square root of the pixel number ($\sqrt{pixel_i}$):

$$Q = \frac{1}{N}\left(\sum_i^N \frac{stdev_i}{\sqrt{pixel_i}}\right) \times SF \times NF$$

where *N* is the total number of background cells for an array, *stdev$_i$* is the standard deviation of the intensities of the pixels making up feature *i*, *pixel$_i$* is the number of pixels in feature *i*, *sf* is the scaling factor, and *NF* is the normalization factor for the analysis.

Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051
Tel: 1-888-362-2447 (1-888-DNA-CHIP)
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: +4 4 (0) 1628 552550
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

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