Chapter 6

Normalization & Differential Analysis Algorithms

Topics

Introduction

This chapter describes the statistical algorithms used in expression analysis for Sentrix® arrays.

A Non-Mathematician's Guide

Normalization Methods

All methods of normalization aim to improve data by mathematically factoring out systematic errors among experimental groups so that their values can be compared. In the case of microarray experiments, systematic variation can result from variation in hybridization temperature, sample concentration, formamide concentration, etc. All forms of normalization achieve this result by making assumptions about the experimental samples and adjusting their values in a way that would factor out intensity changes arising from experimental variation without affecting changes based on true biological differences. The key to applying normalization effectively, therefore, is to understand the underlying assumptions of each method and deciding if they apply in the case of your experiment.

The sections below describe the normalization methods available in BeadStudio. For more rigorous mathematical descriptions, please see *Normalization Algorithms* on page 6-8. For the sake of simplicity, the explanations describe normalization as applied to two samples (A and B). The same principles apply when multiple samples are normalized together.

Definitions

When we speak of a sample, we refer to a single bundle on a Sentrix Array Matrix (SAM) or a single section of a Sentrix BeadChip. When we speak of a population of gene expression values, we refer to the set of all gene expression values received from a scan of a single sample. Therefore, normalization is a process by which two or more populations of gene expression values from two or more samples are adjusted for easier comparison. A scaling factor is a number by which values in one population are multiplied for the sake of normalization. For example, if a normalization technique multiplies all values in Sample B by 1.5 to normalize to Sample A, we say that a scaling factor of 1.5 was applied.

Background Method

For the DASL Assay, the negative controls consist of oligos.

This method subtracts a constant background value from each gene expression value acquired from a scanned sample. The background value is derived by averaging the signals of negative control beads built into the SAM or BeadChip. These beads or oligos contain sequences not expected to hybridize to most genomes and thus provide a measurement of non-specific hybridization, non-specific dye signal and scanner background. This method makes no biological assumptions about the samples and is thus safe to use when you have no expectations about the changes likely to exist between samples. Applying the technique allows for more quantitative assessments of fold-change differences, especially for genes with dim signals.

NOTE:

All other normalization methods described below apply background subtraction in addition to the other method-specific transformations.

Average Method

This method simply adjusts the intensities of two populations of gene expression values such that the means of the populations become equal. For example if the mean value for all genes in Sample A is 300 and the mean for Sample B is 100, all genes on Sample B will have their values scaled (multiplied) by a factor of 3 such that both populations now have a mean of 300. This method assumes that the mean expression levels of all genes should be roughly equal and similarly distributed. This assumption is generally true when samples contain large numbers of genes (such as with a whole-genome sample). However, the assumption breaks down when smaller numbers of genes are used or the samples are quite different. For example, if one had a focused sample containing a few hundred neurological genes and then used this sample to compare brain and liver sample, one would expect the brain sample to yield higher values for biological reasons; the method would therefore not apply. On the other hand, if one were comparing two brain samples, the assumption would probably apply and the method would be valid.

Cubic Spline Method

Cubic Spline normalization differs from all other methods described above in that it is non-linear. In other words, different scaling factors are applied to different parts of the population. The method first breaks the population of gene signals in each sample into a group of quantiles. If possible, the number of quantiles is chosen so that each interval contains 100 probe signals. However, the minimum number of quantiles is 15. For example, 3.3rd percentile, 10th percentile, 16.7th percentile and so on up to the 96.7th percentile. Then, for two samples to be normalized to each other, it scales the 3.3rd percentile of Sample B such that it is equal to the 3.3rd of Sample A, the 10th percentile of B to the 10th of A and so on for all quantiles. Genes whose values lie between quantiles are adjusted by interpolation of the neighboring quantiles.

The benefit of this method is that it can normalize between samples that show a non-linear relationship, such as can happen as a result of unequal sample labeling, different scan settings, etc.

The method is unnecessary when there is a linear relationship among the un-normalized signals of the populations, and in these cases cubic spline normalization should not be applied. To determine if your data has a linear relationship, you can use the Scatter Plot tool (described in Chapter 3, *Data Visualization*) to generate a scatter plot of un-normalized gene expression for the two samples.

If all gene signals from the samples are plotted against each other and show a generally linear relationship (such as in the left plot in Figure 6-1), the cubic spline normalization should not be applied.

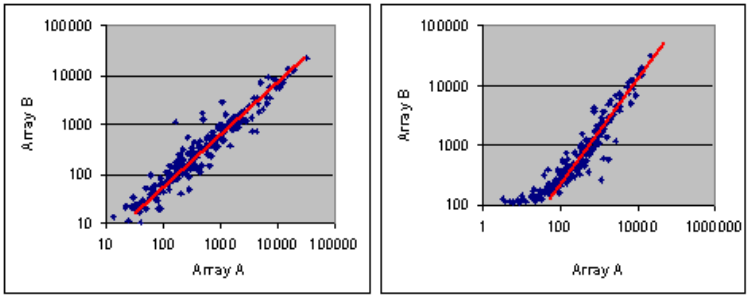

Figure 6-1 Sample A & Sample B

If there is a distortion in the linear relationship (as in the right plot in Figure 6-1), cubic spline may be applied and has the potential to correct the distortion.

Hyb Controls Method

This method works like the 'Average' method above, except, instead of using the signal of all genes to calculate the scaling factor between samples, it uses signals of positive control probes. These probes are included in every sample and hybridize to corresponding labeled oligonucleotides contained in our standard hybridization buffers. The advantage of this approach is that it allows a signal-based normalization of samples while making no biological assumptions about the similarity of the samples. However, due to differences between the dyes used in the control oligonucleotides and those used in the RNA labeling, as well as pipette errors, the hyb controls may be imperfect proxies for the genes in your sample. Also, the Hyb Controls method does not allow correction for differences arising from sample labeling. For these reasons, this method should be applied with caution.

Rank-Invariant Method

For most types of expression experiments, this is our most highly recommended normalization method. Like the Average method above, Rank-Invariant normalization uses a linear scaling of the populations being compared. However, unlike with averaging, the scaling factor is determined not by an average of all genes, but by only rank-invariant genes. 'Rank-invariant' genes are those whose expression values show a consistent order relative to other genes in the population. For example, a gene that is the 200th brightest gene in Sample A and 203rd in Sample B would be considered rank-invariant and would be used to arrive at the normalization factor; a gene that goes from 200th to 10,000th would not be rank-invariant and would not be used. This method is much more resistant to outliers than straight averaging is and generally gives better results. However, as with averaging, if samples are very different in their behaviors, the underlying assumption of rank-invariance (the existence of a subpopulation of genes whose expression is constant across samples showing consistent ranks) will not be true and the method should not be applied.

Due to the DASL Assay's oligo-directed nature, the assumption of similar behavior between samples is often not true. Although rank normalization is preferred for similar samples, the degree of similarity depends on: 1) gene expression in the samples analyzed; and 2) the genes chosen for the oligo pool. Illumina recommends examining un-normalized DASL data in scatter plots before choosing a normalization method for further analysis.

Normalization Algorithms

For all algorithms, normalization is computed with respect to a mathematically calculated “virtual” sample that represents averaged probe intensities across a group of samples. In the cases of spline and rank invariant normalizations, the virtual sample is computed based on the content of the reference group. If there is no reference group, the first group in the list of groups displayed in the Experiment pane is used for group analysis. For SAM/BeadChip analysis, the virtual sample is computed based on the content of the first alphanumeric entry in the upper-left area of the Matrix pane. For the hyb controls and average methods, all samples in the experiment are averaged to produce the virtual sample. A detailed description of normalization algorithms follows.

Background The background value is derived by averaging the signals of built-in negative control Bead types. Outliers are removed using the median absolute deviation method.

Average Sample intensities are simply scaled by a factor equal to the ratio of average intensity of virtual sample to the average intensity of the given sample. Background is subtracted prior to the scaling.

Cubic Spline

The method is similar to the one proposed by Workman et al.[1] The normalization uses quantiles of sample intensities to fit smoothing B-splines.

Let $q_i = \frac{(i-0.5)}{N}, i = 1, 2, \ldots, N$ be a vector of *N* quantiles ($N = max\left(15, \frac{N_{probes}}{100}\right)$). Here, N_{probes} is the number of probes represented on an sample.

For each sample, we compute its vector of quantile intensities. Similarly, we compute quantiles for the "virtual" averaged sample after background subtraction. Cubic B-spline is computed and used for interpolation. For points with intensities ranked outside the $[q_1, q_N]$ interval, we use linear extrapolation rather than spline to avoid nonlinear effects outside the region of interpolation.

Hyb Controls

Let *k = 1 ..., N* enumerate all samples used in the experiment. Then for sample *k*, normalization coefficients (a_k,b_k) are computed using iteratively re-weighted least-squares fit $y_v = a_k y_k + b_i$. Here, y_v, y_k are vectors of intensities of probes corresponding to hybridization controls on virtual and sample *k*, respectively. Tukey bisquare weight function with tuning constant 4.685 provides 95% efficiency when errors are normally distributed, while maintaining protection against outliers. Standard deviation of errors is estimated using median absolute deviation. Normalized intensities are computed with $y_k^{new} = \frac{y_k - b_k}{a_k}$, and then background is subtracted. For further information on the use of hyb controls, see the System Manual *System Controls* appendix for your specific product.

---

1. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 2002 Aug 30;3(9):research0048. PMID: 12225587 [PubMed - indexed for MEDLINE]

Rank Invariant

This method is exactly the same as *Hyb Controls* on page 6-9, except it uses a rank invariant set of probes between a given sample and a virtual sample instead of hybridization controls. The rank invariant set is found as follows: we start by considering probes with intensities ranked between LowRank=50th percentile and HighRank=90th percentile. If the probe's relative rank change $\frac{|r_x - r_y|}{r_y} \leq 0.05$, then the probe is considered to be rank invariant. If less than 2% of all probes in the region are identified as rank invariant, then LowRank is gradually decreased until it reaches 25th percentile.

Differential Expression Algorithm

All algorithms compare a group of samples (referred to as the condition group) to a reference group. The comparison is done using the following error models:

- Illumina custom
- Mann-Whitney
- T-test

Illumina Custom

This model assumes that target signal intensity (I) is normally distributed among replicates corresponding to some biological condition. The variation has three components: sequence specific biological variation (σ_{bio}), nonspecific biological variation (σ_{neg}), and technical error (σ_{tech}).

$$I = N(\mu, \sigma)$$
$$\sigma = \sqrt{\sigma_{tech}^2 + \sigma_{neg}^2 + \sigma_{bio}^2}$$
$$\sigma_{tech} = a + b < I >$$

Variation of nonspecific signal σ_{neg} is estimated from the signal of negative control sequences (using median absolute deviation). For σ_{tech}, we estimate two sets of parameters (a_{ref}, b_{ref}) and (a_{cond}, b_{cond}) for reference and condition groups respectively.

We estimate σ_{tech} using iterative robust least squares fit which reduces influence of highly variable genes. This implicitly assumes that the majority of genes do not have high biological variation among replicates. When this assumption does not hold we overestimate technical error by some averaged biological variation.

When groups contain biological replicates, we produce p-values using the following approach:

$$\sigma_{ref} = \max(s_{ref}, a_{ref} + b_{ref} I_{ref})$$
$$\sigma_{cond} = \max(s_{cond}, a_{cond} + b_{cond} I_{cond})$$
$$p = Z\left(\frac{|I_{cond} - I_{ref}|}{\sqrt{\frac{\sigma_{ref}^2 + \sigma_{neg(ref)}^2}{N_{ref}} + \frac{\sigma_{cond}^2 + \sigma_{neg(cond)}^2}{N_{cond}}}}\right)$$

where s_{ref} and s_{cond} are standard deviations of probe signals.

NOTE:

N_{ref} and N_{cond} denote the number of samples in the reference and condition groups respectively.

We consider that standard deviations exceeding σ_{tech} reflects biological variation. However, we assume that estimates smaller than σ_{tech} are caused by random errors. Therefore, we use the larger of two estimates. Usage of σ_{neg} provides regularization for low abundance targets. Z is two-sided tail probability of standard normal distribution.

When reference and conditions groups contain one sample each, we can neither estimate sequence specific biological variation nor sample processing variation. Instead, we can only assess σ using bead type variation. Therefore, we penalize for that by a factor of 2.5 applied to parameter b. This factor was determined empirically from examination of real sample data.

$$p = Z\left(\frac{|I_{cond} - I_{ref}|}{\sqrt{(a_{ref} + 2.5 b_{ref} I_{ref})^2 + \sigma_{neg(ref)}^2 + (a_{cond} + 2.5 b_{cond} I_{cond})^2 + \sigma_{neg(cond)}^2}}\right)$$

DASL™ Assay

In DASL mode, this factor is 3.

A DiffScore for a probe is computed as:

$$DiffScore = 10\,\mathrm{sgn}(\mu_{cond} - \mu_{ref})\log_{10}(p)$$

For the gene, DiffScores of corresponding probes are averaged. In addition, concordance between probes is reported.

In DASL, p-values are generated for red and green channels independently. These are averaged and the final p-value is generated from the distribution of the average of two independent uniform (0, 1) variables. If direction of intensity change is different for red and green signals, then the larger of p-values is replaced by 1 - p-value prior to averaging.

Concordance is defined as $\frac{|n_u - n_d|}{|n_u + n_d|}$ where n_u is the number of probes showing upregulated signal and n_d is the number of probes showing downregulated signal.

Mann-Whitney

This implementation produces exact p-value if:

$$\min(N_{ref}, N_{cond}) < 3$$

OR

$$\min(N_{ref}, N_{cond}) < 9 \quad AND \quad \max(N_{ref}, N_{cond}) < 13$$

Otherwise, normal approximation with continuity correction is used. Differential scores are computed as described for the Illumina Custom model (page 6-11).

T-test When either the reference group or a condition group contains at least two samples, variance is estimated across replicate samples. Otherwise, variance is estimated from bead-to-bead variation*. We use t-test with the assumption of equal variance.

** Variance computed from bead-to-bead variation may significantly underestimate total variance. We recommend using Illumina Custom model in this case.*

Differential scores are computed the same way as described for the Illumina Custom model (see page 6-7).

Detection Score

Detection scores are computed using negative control signals. Because Illumina's whole genome BeadChips contain large numbers of negative controls (1,000 - 2000), while its focused array and DASL products contain fewer negative controls, different algorithms are used for each type of product.

Whole Genome BeadChips

For whole genome BeadChips, the detection algorithm uses a large number of negative control probes.

Instead of using parametric assumptions, gene signals are ranked relative to the distribution of signals of the negative controls.

DetectionScore = R / N, where R is the rank of the gene signal relative to negative controls and N is number of negative controls. For groups containing multiple samples, the following modification is used. Let m be the number of samples in the group. On the i^{th} sample, the g^{th} gene signal is converted to a Z value and the average Z value across all m samples is computed.

$$Z_g = \frac{1}{m}\sum_i Z_g^i$$

$$Z_g^i = \frac{I - \mu^i_{neg}}{\sigma^i_{neg}}$$

Here μ^i_{neg} and σ^i_{neg} are the mean and standard deviation of signals of the negative controls on the i[th] sample. I is the signal from gene g. The same transformation that is applied to I is also applied to the signals of negative controls. Detection Scores are computed based on the rank of the Z value of a gene relative to the Z values of the negative controls.

Focused Array & DASL Products

Since these products typically contain small numbers of negative controls (20 - 40), their signals (with outliers removed using median absolute deviation) are modeled by normal distribution. The detection score for the probe with intensity I_{probe} is given by:

$$1 - Z\left(\frac{|I_{probe} - \mu_{neg}|}{\sigma_{neg}}\right)$$

Here, μ_{neg} is the average intensity of negative controls and σ_{neg} is the standard deviation of their signals. Z is the one-sided tail probability of standard normal distribution. For the gene represented by N probes we use:

$$1 - Z\left(\sqrt{N}\frac{|I_{gene} - \mu_{neg}|}{\sigma_{neg}}\right)$$

When experimental group contains M replicate samples, the average Z value of $Z_1,\ldots,Z_M$, computed for each sample independently, is assumed to follow a normal distribution

$$N\left(0, \sqrt{\frac{1 + (M-1)r}{M}}\right)$$

where r is the average correlation coefficient of signals of negative controls.

Averaging is done across all pairs of different samples.

In DASL mode, detection p-values (1 - DetectionScore) in red and green channels are computed independently. Their average is assumed to follow distribution of the average of two independent variables distributed uniformly on the interval (0,1). The p-value is generated from that distribution and converted to DetectionScore as 1-pvalue.

NOTES