illumina®

BeadStudio Normalization Algorithms for Gene Expression Data

A mathematical and empirical discussion of BeadStudio normalization algorithms.

Introduction

Normalization is the process of adjusting raw microarray data to remove systematic variation of non-biological origin. BeadStudio software offers three algorithms for normalizing data generated from Illumina Gene Expression BeadChips and panel sets.

There are three major assumptions when normalizing microarray data.

1. **The effect of any systematic error will be uniform over the spatial distribution of bead types.**
2. **The majority of genes will not be differentially expressed.**
3. **Normalization should remove systematic variation while leaving biological variation intact.**

Scatter plots, box plots, and cluster analysis are all reliable methods for evaluating the overall consistency and quality of the data. Scatter plots are X versus Y plots comparing intensity values for two arrays. In general, scatter plots of signal intensities from similar tissues or replicates will display a symmetrical distribution of data points about the identity (45º) line. Because some experiments such as time-point or drug response experiments may generate asymmetrical distributions, the investigator must consider the data in the context of the study design.

Displaying scatter plots in log space usually improves the visualization of raw data, but it should be noted that log transformation can add artifacts such as curvature, increased variation for low signals, or changes in correlation among arrays. Therefore, researchers must take care not to confuse actual data features with artifacts introduced by log transformation. Note that in BeadStudio normalization and statistical computations are performed in linear space.

There are three key features to examine in scatter plots of sample data:

1. **symmetry of the distribution about the identity line**
2. **agreement between best fit line and identity line**
3. **presence of nonlinear behavior between samples.**

The investigator is looking for trends in the data such as signal saturation, distributions that do not follow the identity line, curvature, divergent scatter, extremely dispersed scatter, offset scatter, and outlier gene clusters.

Two other powerful data evaluation tools are box plots and cluster analysis. Box plots provide a quick evaluation of the magnitude and variability of signals within and between arrays and enable easy detection of outliers. Cluster analysis is a useful method for determining if non-biological variation is present in the data. Such variation may correlate with the day of experiment, total RNA preparation date, chip variability, scan order, and the day of mRNA processing. For data quality evaluation, clustering by correlation is most appropriate.

Although BeadStudio algorithms are intended to remove systematic, non-biological variation, they cannot correct data of inferior quality such as data derived from degraded mRNA or compensate for poor experimental design (e.g., confounding biological variation with systematic variation).

BeadStudio Algorithms

BeadStudio software offers three normalization algorithms—average normalization, rank invariant normalization, and cubic spline normalization—each with specific advantages for normalizing microarray data. While average normalization is recommended in most circumstances, the other algorithms may be better suited for certain situations which are discussed later in this document.

Before the researcher applies an algorithm to the raw data, a reference group must be defined. A reference group can consist of one or more arrays. The default reference group is the first array or group in the list, but any array or group can be specified as the reference. The reference group is used to calculate a virtual array. A virtual array comprises the average values from all the arrays in the reference group and is used to determine normalization parameters. Once the reference group is defined for normalization, the same group is used as the reference group during differential expression analysis.

Average Normalization

Average normalization is used to rescale intensities across multiple arrays and chips. This algorithm is an appropriate choice for experiments that employ a large number of arrays with differences in overall intensity. For average normalization, a scaling factor, S_i, is calculated by dividing the average intensity of the virtual array (μ^v) by the average intensity for all arrays in a group (μ_i).

$$S_i = \mu^v / \mu_i$$

The bead-type intensities of all arrays in the experiment are normalized by S_i.

$$I_{pi(norm)} = I_{pi} S_i$$

Each probe is indexed by p and the number of probes ranges from 1 to *n*. Each array is indexed by i and the number of arrays ranges from 1 to *m*.

Rank Invariant Normalization

Rank invariant normalization is an appropriate selection if the investigator can assume that a reasonable percentage of the genes in an experiment are not differentially expressed (i.e., rank invariant).

For rank invariant normalization, a subset of probes whose rank does not change across the experiment are identified and serve to define the normalization parameters. To determine if a gene is rank invariant, the intensities from the low rank (50th percentile) and the high rank (90th percentile) are considered to be rank invariant if the absolute value of the change in relative rank (r) between the g^{th} gene in the ith sample and the gth gene in the virtual sample is less than 0.05. Here r_{gv} is the rank of gene g in the virtual array v and r_{gi} is the rank of gene g in array i.

$$\frac{|r_{gi} - r_{gv}|}{r_{gv}} < 0.05$$

Let i (1,2,3...m) enumerate all samples used in the experiment. Then for array i, normalization coefficients (a_i,b_i) can be computed using an iteratively reweighted least squares fit.

$$y_v = a_i y_i + b_i$$

Here y_v and y_i are intensity vectors of probes corresponding to the rank invariant set of probes on virtual array v and i, respectively. The Tukey bisquare weight function with the tuning constant set at 4.685 provides 95% efficiency when errors are normally distributed while maintaining protection against outliers. The standard deviation of errors is estimated using the median absolute deviation.

The rank invariant normalization algorithm normalizes intensities and subtracts background through the following equation.

$$y_i^{new} = \frac{y_i - b_i}{a_i}$$

Rank invariant normalization is more robust to outliers than the average normalization method. However, if it is known that a large percentage of the genes in an experiment may be differentially expressed, or that there is a large difference in intensity among arrays, the user should consider cubic spline normalization.

Cubic Spline Normalization

Cubic spline normalization is implemented to remove curvatures observed in scatter plots that arise from nonlinear relationships between samples or groups of samples when plotted in log space. This method initially divides the intensity distribution into a group of quantiles consisting of a similar number of gene intensities. Quantiles (q) are determined by letting

$$q_i = \frac{(i - 0.5)}{N}$$

where i equals array 1,2,3,...m and N is equal to total number of quantiles.

By letting q_i be a vector of N quantiles, then

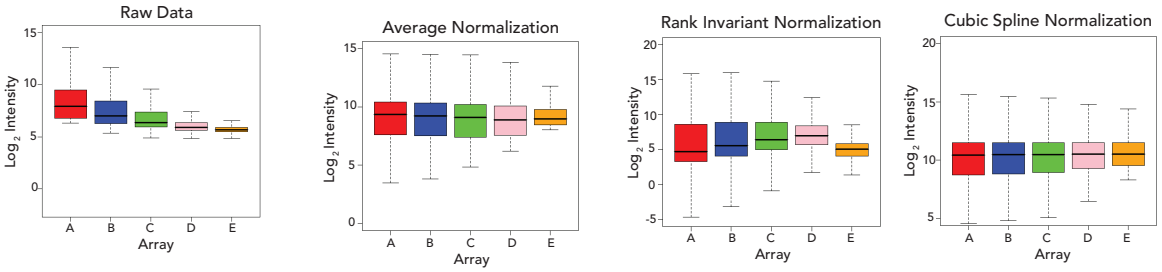
$$N = max(15, \frac{N_{probes}}{100})$$

where N_{probes} is the number of probes represented on a sample array. Using the scaling factor, the intensities from the sample array in a given quantile are adjusted to match the intensities from the virtual array. For two samples to be normalized to each other, cubic spline normalization scales the k^{th} quantile of the sample array such that it is equal to the k^{th} quantile of the virtual array. This scaling is repeated for each quantile. Genes with values that lie between quantiles are adjusted by interpolation of the neighboring quantiles. While this normalization is meant to address nonlinear relationships between samples in log space, it can also be used as a general normalization method.

Normalization of titration Data Sets in BeadStudio

For the purpose of demonstrating the results of normalization using the three algorithms available in BeadStudio, Illumina scientists analyzed two titration data sets. The first titration experiment profiled five input concentrations of human total RNA and shows the effects of normalization on data with nonlinear characteristics. The second titration data set provides a biologically relevant example of the stability of high quality data to BeadStudio normalization algorithms using four

Figure 1: Effect of Average, Rank Invariant, and Cubic Spline Normalization On Nonlinear Data



MAQC1 samples: 100% Brain total RNA, 100% Universal RNA from Stratagene, 25% Universal-75% Brain, and 75% Universal-25% Brain.

Total RNA Titration Data

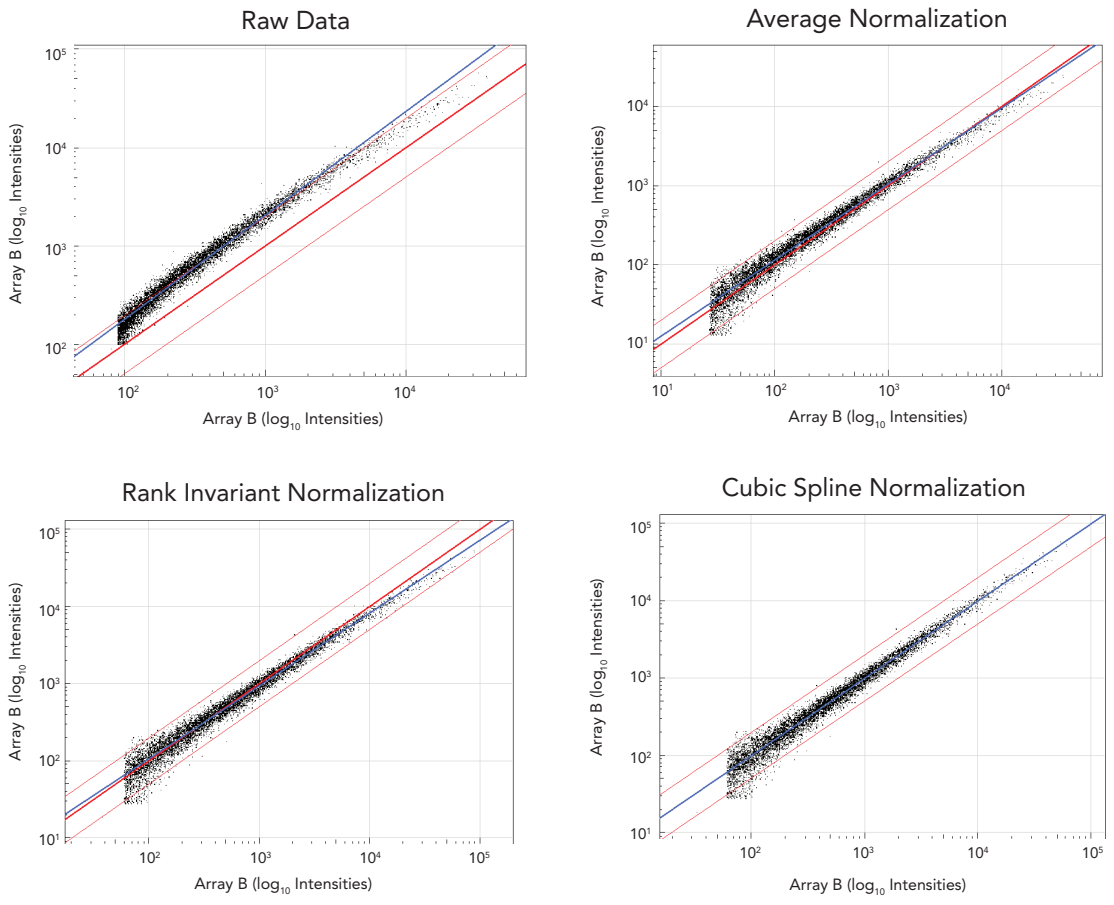
For the human total RNA titration experiment, five cRNA samples were hybridized on a single Human-6v1 BeadChip. The five samples were placed on arrays A (highest concentration) through E (lowest concentration). The box plots for the average signal intensities for all the arrays show that the intensities diminish from array A to array E (Figure 1). Although both average normalization and rank invariant normalization scale the data, average normalization scales all the arrays tighter around the mean of the arrays. Cubic spline normalization rescales all arrays so that the means are equal.

A log scatter plot displaying the intensity signals from array A versus array B demonstrates the nonlinear relationship of the raw data and reveals a significant curvature in the data set. Additionally, the raw data set is displaced from the identity line, and the best fit line is offset and skewed from the identity line. (Figure 2). When adjusting nonlinear data, cubic spline normalization achieves the greatest improvement in symmetry, and the best agreement between the identity and best fit line.

MAQC Titration Data

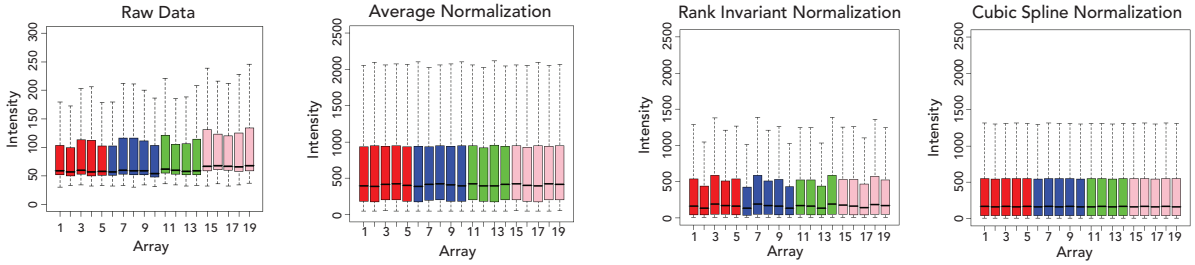
The MAQC data were analyzed to demonstrate the results of BeadStudio normalization on biologically relevant samples. The samples were assayed on four Human-6v1 BeadChips. In practice, when the data are of high quality, different normalization methods do not result in drastically different analysis outcomes. The robustness of high-quality data is illustrated with box plots, scatter plots, clustering, and differential gene expression analysis of the MAQC sample profiles (Figures 3, 4, and 5, and Table 1). The box plots show considerable similarity between average normalization and cubic spline normalization in terms of scaling the median and generating a unifrom interquartile range. In comparison, a larger degree of variability remains after rank invariant normalization. Scatter plots of the MAQC data also reveal only minor differences in the results of the three normalization algorithms, with

Figure 2: Effect of Average, Rank Invariant, and Cubic Spline Normalization on Nonlinear Data



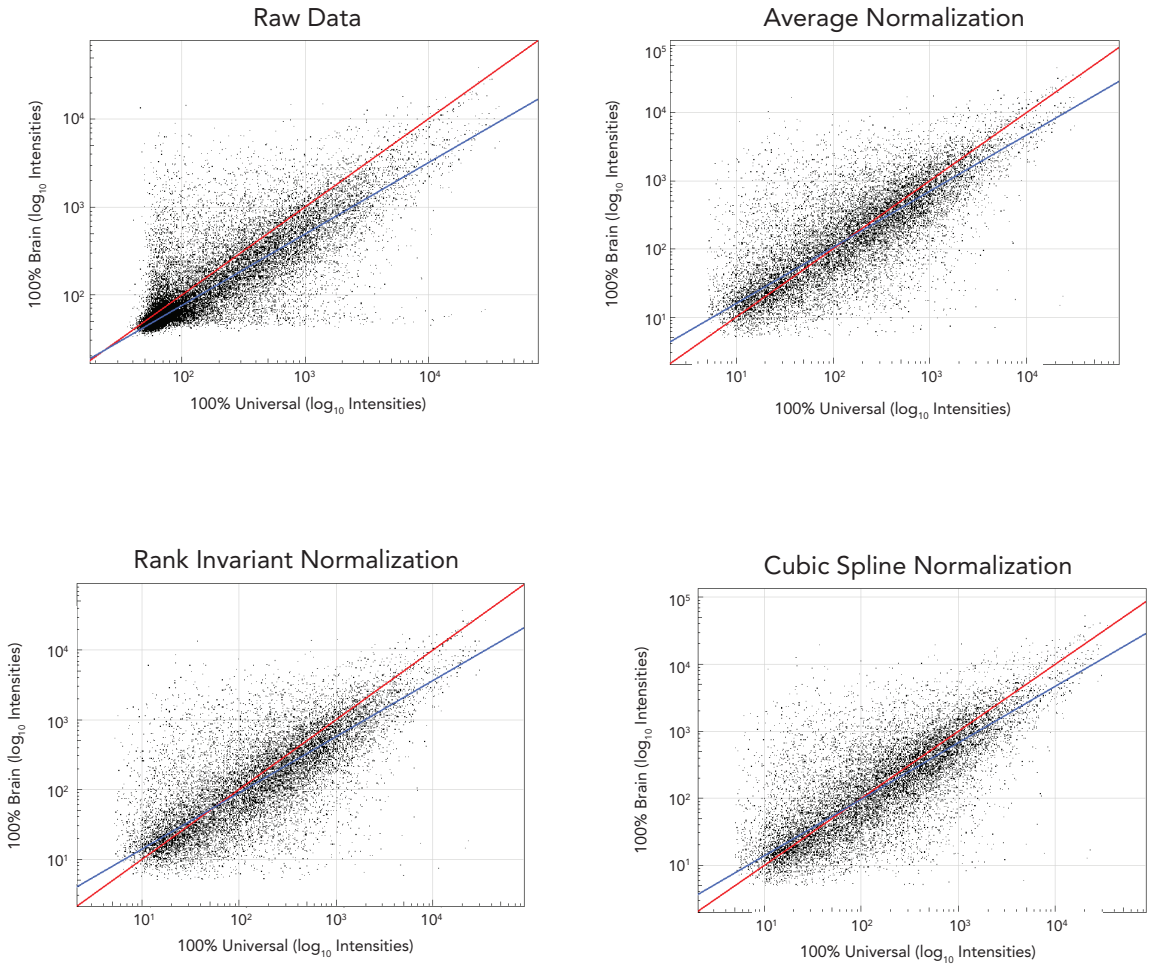
The data were filtered by the detection p-value < 0.01. The identity line and two-fold boundaries are denoted in red. The best fit line is blue.

Figure 3: Effect of Average, Rank Invariant, and Cubic Spline Normalization on High-Quality Data



Scale differs from raw to normalized data because normalized data were filtered by detection p-value < 0.05.

Figure 4: Effect of Average, Rank Invariant, and Cubic Spline Normalization on High-Quality Data



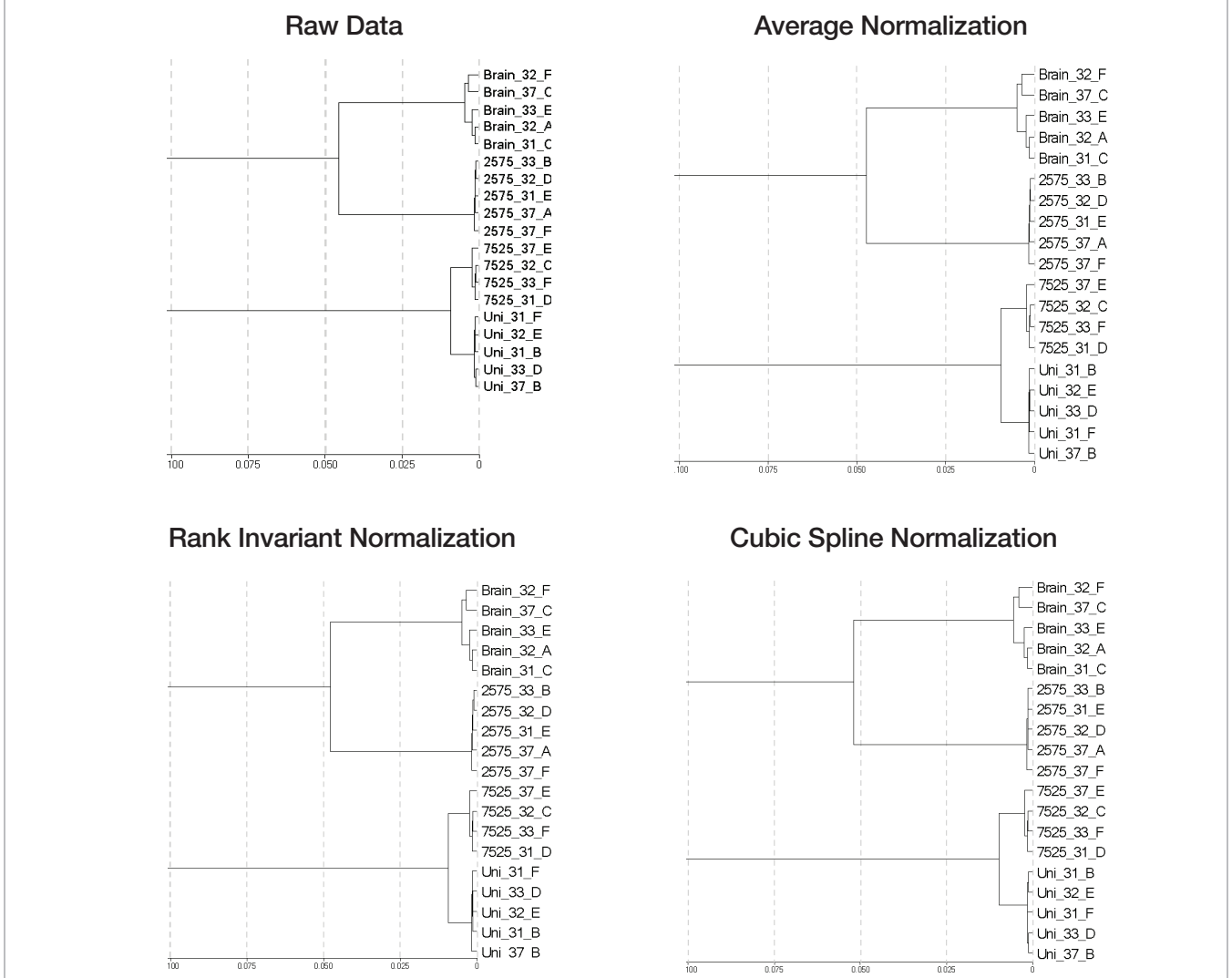
The data were filtered by the detection p-value < 0.01. The identity line and two-fold boundaries are denoted in red. The best fit line is blue.

average and cubic spline normalization showing the closest similarity in performance (Figure 4).

Hierarchical clustering shows that the four MAQC sample groups partition distinctly according to biological variation and do not show appreciable differences with regard to composition and correlation as a function of normalization (Figure 5). The cluster consisting of the universal total RNA samples shows the highest variability in sample order. However, the correlation within this group is extremely tight and small differences in correlation results within a rearrangement of the samples in this cluster.

Finally, differential analyses were conducted between the MAQC pure brain and universal RNA groups to determine the impact of applying different normalizations. The Illumina Custom error model with false discovery rate (FDR) correction was applied to data sets after normalizing by each algorithm separately. The resulting lists of genes

Figure 5: BeadStudio Dendrograms of MAQC Samples Before and After Normalization



AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGAT
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAG

Table 1: Analysis of BeadStudio Normalization on Pure Brain and Pure Universal MAQC Titration Samples

Normalization Algorithm	Number of Differentially Expressed Genes ($p < 0.001$)	Analysis	% Overlap in the number of genes
Average	12,489	Average versus Rank Invariant	88%
Rank Invariant	12,720	Rank Invariant versus Cubic Spline	92%
Cubic Spline	12,671	Cubic Spline versus Average	96%

were analyzed to determine the extent of overlap among the various normalizations (Table 1). Overall, there is considerable overlap among gene lists, illustrating that conclusions drawn from differential expression analysis using high-quality data should not change drastically as a function of normalization.

Conclusion

The results of these analyses indicate that average normalization and cubic spline normalization algorithms perform most similarly when the data are of high quality. Further, rank invariant normalization seems to be more conservative with regard to the magnitude of the adjustment it makes to the data. When the data are of high quality, different normalizations will likely not lead to large differences in the results of statistical analysis, class discovery, or class prediction methods. In general, Illumina recommends using average normalization; however, when there are nonlinear associations among arrays, cubic spline normalization is a more appropriate choice.

References

1. MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, et al. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotech 24(9):1151-1161.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2011 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, HiScan, TruSeq, and Eco are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 470-2007-005 Current as of 03 January 2011.

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA