Gene Expression Microarray Data Quality Control

GenomeStudio® supports easy and powerful data quality control using flexible analysis features and internal controls present on all Illumina Gene Expression products.

Introduction

Quality control (QC) of data is an important step when performing any microarray gene expression study. This technical note describes the concepts, metrics, and techniques for ascertaining data quality after all data have been collected. When used together, these techniques can help researchers identify outlier samples, the potential cause of outlier data, or decide whether a sample needs to be repeated or removed from the data set. Note that the data from the metrics and techniques described here should be considered holistically before deciding whether a sample or data set meets the expected quality standard.

Experimental design and best practices should always be taken into consideration before beginning any microarray experiment. While quality control analysis helps to identify samples for which the data characteristics are significantly different than the majority, systematic variation may be imparted by the experimental process itself and can be difficult or impossible to remove through standard normalization. Such variation may affect data in ways that are not obvious by standard QC methods, and thus may not be detected. Some important factors to take into account include, completing the experiment in as short a time frame as possible, limiting the number of batches run, balancing samples among batches and chips as much as possible, and using a single scanner. Of course there are many other factors to consider and it is import to thoroughly plan and account for as many factors as possible before the experiment begins.

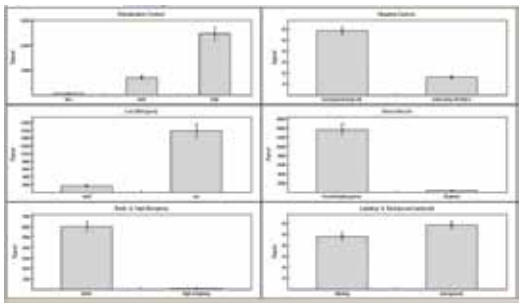
The concept of replication is an important one for gene expression studies. Herein we define **biological replicates** as a set of samples taken from a set of multiple unique individuals such that each individual contributes a given sample. For instance, if 25 samples are used from cancer type A and 25 from cancer type B, each cancer type consists of 25 biological replicates taken from a total of 50 different individuals.

A **technical replicate** is defined as a sample that has been partitioned and carried through the sample preparation process from a given point forward. For instance, a given sample that is partitioned then mixed into two different hybridization mixtures and applied to two different arrays is a technical replicate for hybridization. Similarly if a given sample is partitioned and mixed into two different IVT reactions and hybridized to two different arrays it is a technical replicate for IVT. Such technical replicates can be created for any step in the sample prep process, and they help quantify the variation due to a given set of steps in the sample preparation process.

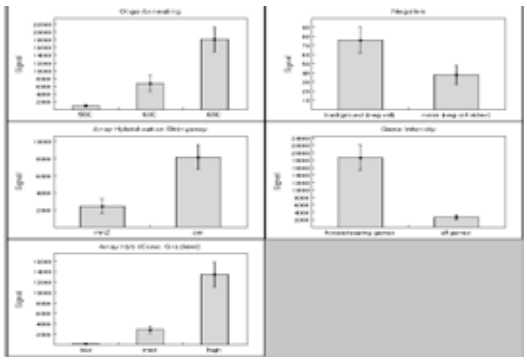
To obtain a comprehensive indication of data quality, it is necessary to consider several aspects of the data using a broad panel of statistics, plots, and metrics. It is also important to note that the QC analysis strategy uses relative comparisons rather than absolute values when examining QC metrics. Furthermore, consistency of experimental performance should be considered in the context of historical performance data.

Figure 1: Genomestudio Control Plot Examples

A: Direct Hyb Control Plots



B: WG DASL® Control Plots

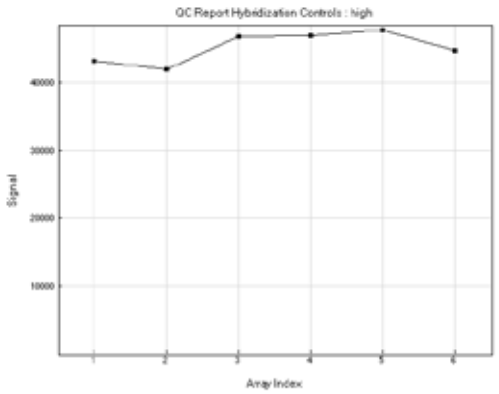


The GenomeStudio analysis software package provides several convenient QC tools to facilitate this analysis. These include plotting features such as the Control Summary Plot, Box Plots, Scatter Plots, Cluster Dendrograms, and data tables such as the Control Probe Profile and the Samples Table. Customizable tools such as the User Defined Function and filtering allow users to generate novel metrics and complex filters.

Control Summary Plots

Illumina Gene Expression BeadChips have internal control features to monitor data quality. The results of these controls can be visualized easily in GenomeStudio by selecting the Control Summary tab (Figure 1). Control data can also be exported from the Control Probe Profile and analyzed with third-party software. GenomeStudio control features are either sample-independent or sample-dependent. The sample-independent metrics make use of oligonucleotides spiked into the hybridization solution. Poor performance measured by these controls could indicate a general problem with the hybridization, washing, or staining. The sample-dependent metrics are based on measure-

Figure 2: Genomestudio Control Summary Plot



Signal intensity values (y-axis) of hybridization controls are consistently high across all six arrays/samples (x-axis) from a gene expression profiling experiment. This graph does not indicate a data quality problem.

ments from the actual sample of interest. Poor performance of these controls may indicate a problem related to the sample or labeling. Further details of the control plots in GenomeStudio are provided in Table 1, the GenomeStudio Gene Expression Module User Guide, and related Assay Guides.

Normal variations in control plot values can arise due to incidental factors such as system setup, sample origin, and BeadChip type. These factors make it difficult to determine data quality by comparison to a specified expected value for each QC metric. To minimize the influence of these factors, relative—rather than absolute—control values should be used as QC criteria. Relative comparisons of control values can be made by:

1. **Identification of outliers by comparison to current and historical data. Outlying samples for any given control metric can be quickly identified in GenomeStudio by using the control summary plot and expanding the plots to view QC values for individual samples (Figure 2). The data can also be exported and examined in other spreadsheet software, enabling further comparisons with historical data.**
2. **Comparison of samples across two or more control metrics to ensure a consistent ratio between relevant control values. This can be shown by simultaneously displaying and comparing different control plots in GenomeStudio. Alternatively, the control data can be exported and relevant values co-plotted using other software (Figure 3). For example, typical comparisons for the Direct Hyb assay include:**
   - **Housekeeping & Background. While housekeeping genes are known to fluctuate as a function of tissue type, they should be fairly consistent across arrays when from a similar sample source. Housekeeping genes should produce a higher signal than background.**
   - **Perfect Match (PM) & Mismatch (MM2). The PM probe signal is expected to be higher than the MM2 probe signal (Figure 3). Deviations in the ratio of PM to MM2 signal for a given sample could indicate a problem with specificity in the experiment.**

Table 1: Control Plots in Genomestudio

A: Direct Hyb Control Plots

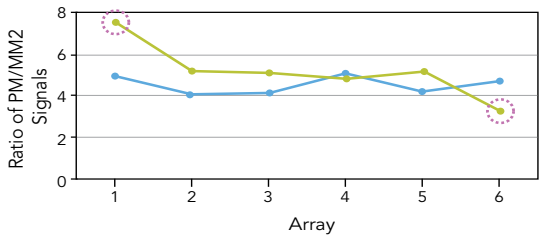
Control Metric	Expected value
Hybridization Controls*	High > Medium > Low
Low Stringency*	PM > MM2
Biotin and High Stringency*	High
Negative Controls (Background and Noise)	Low
Gene Intensity (Housekeeping and All Genes)	Higher than Background (Housekeeping > All Genes)
Labeling and Background	If used, Labeling > Background; Otherwise, Labeling ≈ Background.

B: WG DASL Control Plots

Control Metric	Expected value
Hybridization Controls*	High > Medium > Low
Contamination	One Code High, Others Low
Stringency† (Low and High)	Low: Red > Green High: Green > Red
Negative Controls (Background and Noise)	Low
Genes (Genes and Variation)	Higher than Background
Gap†	Higher than Background

*Sample-independent control metrics.
†DASL stringency and gap controls are designed against the glutaminyl-tRNA-synthetase gene (QARS) and are therefore dependent on the expression level of QARS in the sample of interest.

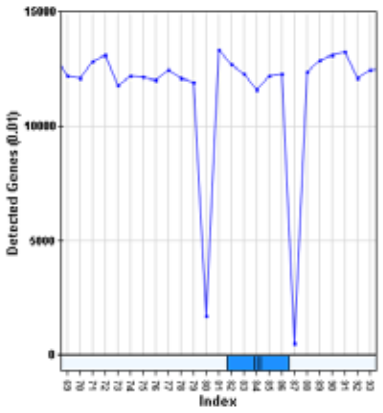
Figure 3: Comparison Across Control Metrics, PM/MM2 Ratio Line Plot



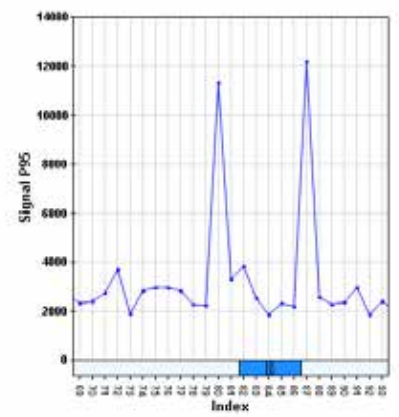
A plot of the ratio of PM/MM2 probe signals across several samples from two different BeadChips (blue and green). In the case of the blue BeadChip, all six samples have similar ratios approximately 4–5 PM/MM2. However, some arrays from the green BeadChip (circled) exhibit deviating ratios, indicating a possible difference in stringency between arrays 1 and 6.

Figure 4: Samples Table Plots

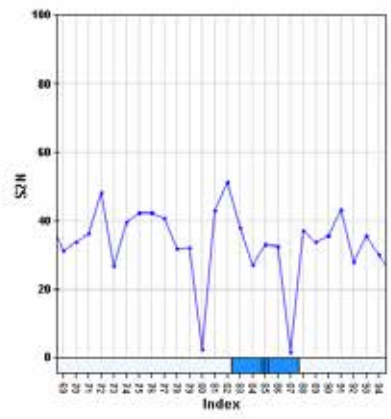
A: Number of Genes Detected for Each Sample



B: P95 Intensity for Each Sample



C: Signal-to-Noise Ratio for Each Sample



Three example line plots of control metrics that can be found in the metrics table. Samples 80 and 87 are detected as outliers in all three plots. This is an indication that these two samples may need to be repeated or excluded from further analysis.

Samples Table Metrics

Several metrics in the Samples Table allow rapid assessment of sample quality. These include the number of genes detected, the p95 intensity, and signal-to-noise ratios.

Number of Genes Detected

GenomeStudio calculates and reports a detection p-value, which represents the confidence that a given transcript is expressed above the background defined by negative control probes. This detection score determines whether a transcript on the array is called detected. A value below the user-defined p-value threshold of either 0.01 or 0.05 indicates a gene is detected. An unusually low number of detected transcripts could result from a number of causes such as high background on the array, low signal, or poor stringency. Thus, the number of detected transcripts is a good overall QC indicator. All samples on a given BeadChip, prepared from the same sample source, should have a similar number of detected transcripts (Figure 4a). However, this may not be the case with samples from highly heterogeneous tissues or from necrotic tissue.

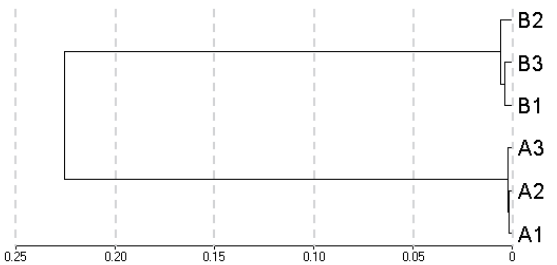
95th Intensity Percentile (p95)

This metric provides a quick way to examine the high-end intensity variation across samples (Figure 4b). Outlier arrays can be easily identified using these plots.

Signal-to-Noise Ratios (p95/p05 and p95/Background)

These metrics can be calculated using the User Defined Function () in the Samples Table, and provide a quick way to visualize the overall strength of measured signal, compared to the background (Figure 4c). This calculation should be performed on data that have not been normalized.

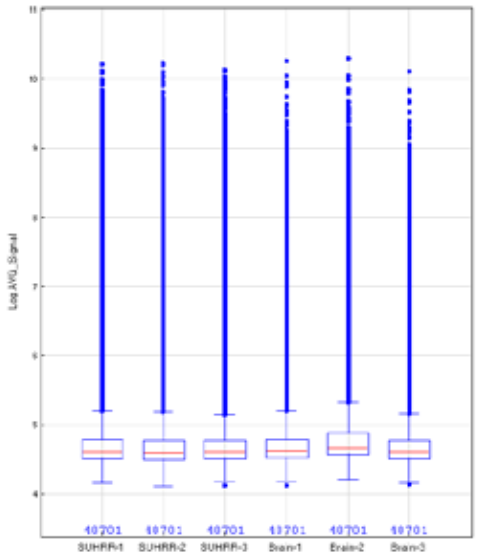
Figure 5: Dendrogram Showing Clustering Of SImilar Samples



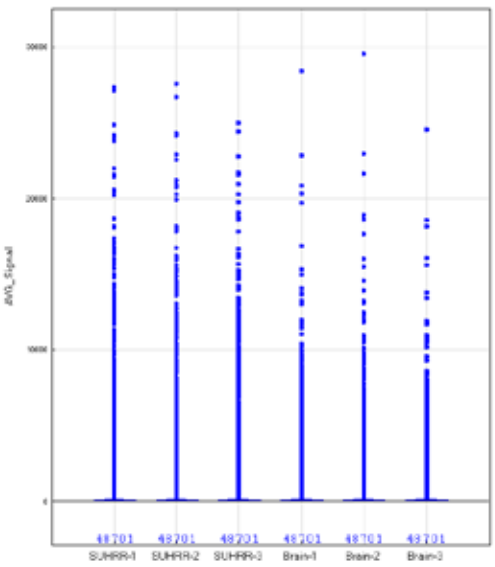
Correlations between six samples are plotted as a dendrogram. As expected, biological replicates A1–A3 and B1–B3 cluster tightly together within groups as sub-trees, and there is lower correlation between samples from different groups.

Figure 6: Box Plot Of Average Signal Across All Samples

A: Logarithmic Box Plot



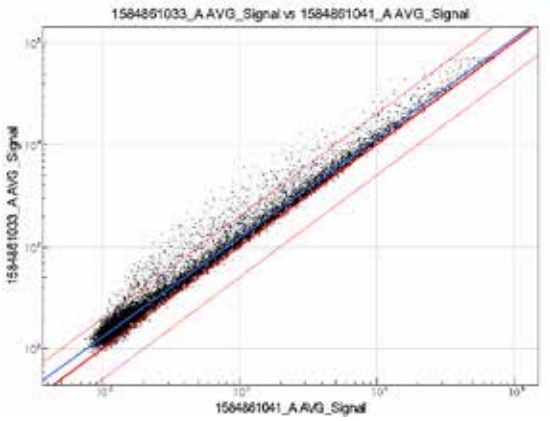
B: Avg_Signal Box Plot on Linear Axes



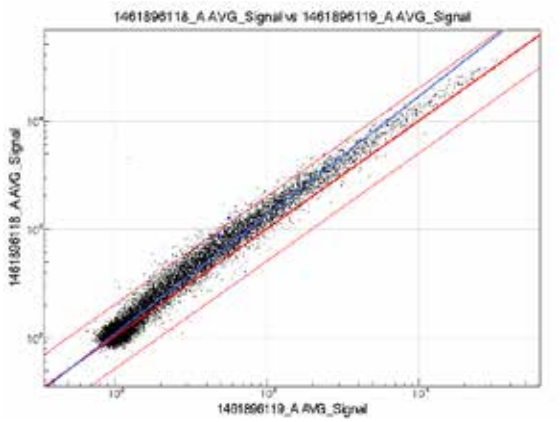
Average signal intensity of all assays is plotted for each of six samples (x-axis). A detailed view of median values is seen in (A) with a log AVG_Signal plot. A standard linear plot in (B) provides an overview of data distribution. Signal intensity values of outliers (blue points in B) are moderate and not likely to be saturated since their intensities are much less than 64,000.

Figure 7: Examples Of Data Quality Diagnosis Made By Scatter Plots

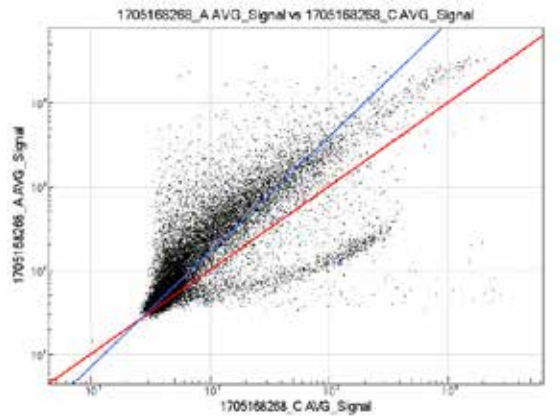
A: Saturation and Stringency Effects



B: Non-Linear Effects



C: Poor RNA Quality



Examples of diagnostic scatter plots. A variety of data features can be visualized with scatter plots. Plot A shows evidence of image saturation (see green box). Plot B shows a non-linear association between two samples. Plot C is a highly asymmetrical scatter plot, which may be indicative of poor sample quality.

Data Visualization and Analysis tools

Sample Clustering

An overall view of experiment performance can be gained by clustering samples using the correlation metric. Dendrograms based on this metric are useful for identifying outlying samples. Samples that have the most similar expression profiles, as determined by correlation value, are clustered together. The x-axis position of the node linking two samples on a cluster dendrogram is equal to 1-correlation value (r). Therefore, samples joined with a node positioned closer to 0 are more similar and have higher correlation values (Figure 5). Technical replicates (e.g., hybridization or labeling) are expected to be the most similar and to be clustered together. Biological replicates from the same cellular origin would also be expected to cluster together.

However, in some cases, when samples in an experiment show few differences in gene expression levels, samples may not cluster according to their biological category. This issue can arise when samples lack sufficient differences to segregate into discrete clusters and may not necessarily indicate a problem with the microarray process.

Box Plots

Box plots are useful tools to quickly visualize the variation within an array and between arrays (Figure 6a). The key features of a box plot are the median (red line), 75th percentile (top edge of box), 25th percentile (bottom edge of box), 95th percentile (top edge of vertical line), 5th percentile (bottom edge of vertical line), and the outlier points (above and below the vertical lines).

Samples showing an abnormal distribution of signal intensities may indicate a general problem. This can be assessed by using non-normalized data to create a new log AVG_Signal intensity subcolumn with the User Defined Function. Generating a box plot of the log AVG_Signal intensities, rather than absolute signal intensities, ensures a more even representation of data (Figure 6a).

Saturation can be quickly assessed in GenomeStudio by creating box plots of the AVG_Signal across all individual samples. Transcript intensities are measured in fluorescence units. Illumina's BeadChip readers, including the iScan System, iScanSQ and BeadArray™ Reader, have a dynamic range of 0—65,535 fluorescence units. Saturation results in a decreased sensitivity to true differential expression. The fluorescence measurements of all samples can be quickly scanned in GenomeStudio by performing an analysis without any normalization and using the box plot or scatter plot function to plot the AVG_signal across all individual samples (Figure 6b). High intensity values across a large number of transcripts typically indicate that the scan factor settings were set too high and that the BeadChip may need to be rescanned with lower setting values.

Scatter Plots

Sample clustering provides an overall indication of sample similarity but does not identify differences between individual transcripts. Therefore, it is also useful to create scatter plots comparing transcript expression level data between two samples. Technical replicates should show the fewest number of up or down regulated genes. Biological replicates should also exhibit similar transcript levels. In titration experiments performed by Illumina scientists, correlation values (r^2) greater than 0.99 in linear space are typically achieved for hybridization replicates. Scatter plots can be used to diagnose several different data quality issues such as image saturation, inconsistent stringency, large differences in signal, and poor sample quality (Figure 7).

Images

If control metrics identify potential problems with a particular sample, it is often useful to review the actual images of the corresponding scan, using the GenomeStudio image viewer. Viewing images may identify areas of poor hybridization signal, high background, or localized artifacts. Due to the large number of redundant features on all BeadArray products, these artifacts are unlikely to affect sample results unless large areas of the BeadChip are affected.

Biological Significance

In addition to monitoring the standard metrics described above, it is important to account for known biologically derived controls. For example, well studied transcripts may be expected to be up or down regulated by a given treatment. The search functions in GenomeStudio enable researchers to confirm that such known transcripts show the expected patterns of expression.

Conclusion

Using GenomeStudio to analyze control measurement data provides a basis for assessing the validity of the results of a gene expression microarray experiment. All of these metrics should be analyzed holistically to determine an overall measure of assay performance. Major problems are normally apparent in more than one QC metric, and small variations within an individual metric may be insignificant when interpreted in the context of the entire QC data set. An expected range of QC values for a given system setup can be established by monitoring these QC values over different experiments and over time. As a best practice, run the report generation function of the GenomeStudio GX Module and record the control values after every experiment. Having a baseline of the normal range of QC metric values permits easier identification of problems if they arise. Observing normal patterns for controls can give researchers more confidence in results.

References

1. Oeser S, Baker S, Chudin E, Kuhn K, McDaniel TK, Methods for Assessing Microarray Performance. Technological Innovations in Life Sciences, Volume 1.
2. mRNA Expression Profiling Data Sheet (PDF) http://www.illumina.com/Documents/products/datasheets/datasheet_mRNA_expression.pdf.
3. Whole-Genome DASL Assay for Expression Profiling in FFPE Samples Data Sheet (PDF) http://www.illumina.com/Documents/products/datasheets/datasheet_whole_genome_dasl.pdf.

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

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA