Post-Sequencing Quality Control Metric for the HTG EdgeSeq PATH Assay

Introduction

This document describes a statistical method used to identify post-sequencing Quality Control (QC) sample failures for the HTG EdgeSeq PATH Assay. These QC failures typically manifest as count data for which all probes in the assay have approximately equivalent values and do not appear to represent biologic information. These QC failures indicate potential sample preparation or HTG EdgeSeq processor issues and should be removed from subsequent analysis.

Methods

Study Design

Standard performance is established by a baseline study; the results from which establish expected performance. The recommended sample size for establishing representative "good" (expected) performance for a specific study is at least forty-eight (48) samples representative of the biologic study of interest. The current QC metric used at HTG Molecular Diagnostics, Inc. (HTG), is based on 48 individual formalin-fixed, paraffin-embedded (FFPE) tissue lysates from individual luminal A breast cancer (n = 10), non-small cell lung cancer, NSCLC (n = 10), melanoma (n = 10), diffuse large B-cell lymphoma, DLBCL (n = 10), and colorectal cancer (n = 8) samples as well as 16 replicates of three cell line lysates (MCF-7, A549, SUDHL6).

For best results in establishing the baseline, each reaction product should be individually tagged/barcoded, individually cleaned-up, and sequenced with balanced quantitation, preferably in one sequencing run.

Statistical Approach

The data phenomena that is being measured is non-differential, or uniform distribution of probe-level counts within a sample/sequencing well. With the diversity of gene probes found in HTG's commercially available assays, this profile would not be expected in a biological sample.

This data pattern is captured by the difference between the third and first quartiles of the distribution; the sufficient statistic called the Interquartile Range (IQR). Conceptually, in a sample with biologically driven signal, there is an expectation of a significant difference in the average signal of the lowest quartile of signal (genes with no to low expression) and the third quartile (genes with moderate to high expression).

Each sample/well assessment is represented by a single IQR. A sample manifesting a QC failure will have a statistically small IQR compared to non-QC failures. Since each sample/tissue type will provide a different profile of gene expression, data from the baseline study are used to establish the expected value (mean, or E) and variance, V, that is then used as fixed values for evaluation of subsequent samples. The goodness-of-fit (GOF) test statistic,

$$\text{GOF test statistic} = \frac{(O_i - E)^2}{V}$$

Where $O_i = Q3 - Q1$, is the IQR for sample i for fixed E and V estimated from the baseline study. E is the mean from non-QC failures from the baseline sample, $E = \sum_{j=1}^{48} IQR_j$ and $V = \sum_{j=1}^{48} (IQR_j - E)^2$. This GOF test statistic value follows a χ^2 distribution with one degree of freedom and the p-value can be expressed as,

$$1 - \chi_1^2(\text{test statistic}) < 0.05$$

Samples with p-values less than 0.05 are considered QC failures. Note that GOF tests are modeling null hypotheses that test "do the sample values come from the same distribution as the distribution that were used to establish the E and V values." P-values that are greater than 0.05 suggest that the new sample might not be drawn from the same baseline distribution.

Example Data and Analysis

When estimating E and V for the GOF test to test future samples it is important that possible QC failures be removed before estimating E and V. One simple way to qualitatively identify QC failures in the baseline study is to graph within sample expression (counts) after transforming to $\log_2$ Counts Per Million ($\log_2$(CPM)):

$$log_2 CPM = \log_2\left(\frac{r_{gi} + 0.5}{\mathrm{R}_i + 1}\ x10^6\right)$$

Where, r_{gi} is the number of sequence reads for each probe (g) and sample (i) and R_i is the number of mapped reads (sum of probe level counts) for each sample (i). The $\log_2$(CPM) counts for each probe are plotted for each sample, as demonstrated in the left side of Figure 1. QC failures, as shown in red, have very little difference in expression and typically are represented in a line across probes (note that some housekeeping genes (HK) will have high expression even in QC failures, and HTG recommends removing them before estimating E and V from the baseline study counts). The resulting samples that qualitatively demonstrate the pattern typical with QC failures are removed and remaining non-QC failures from the baseline study can then be used to estimate E and V values that will be used for identifying QC failures in subsequent studies. The Mean IRQ table below shows the typical difference between the GOF test statistic values in QC failures and non-QC failures.

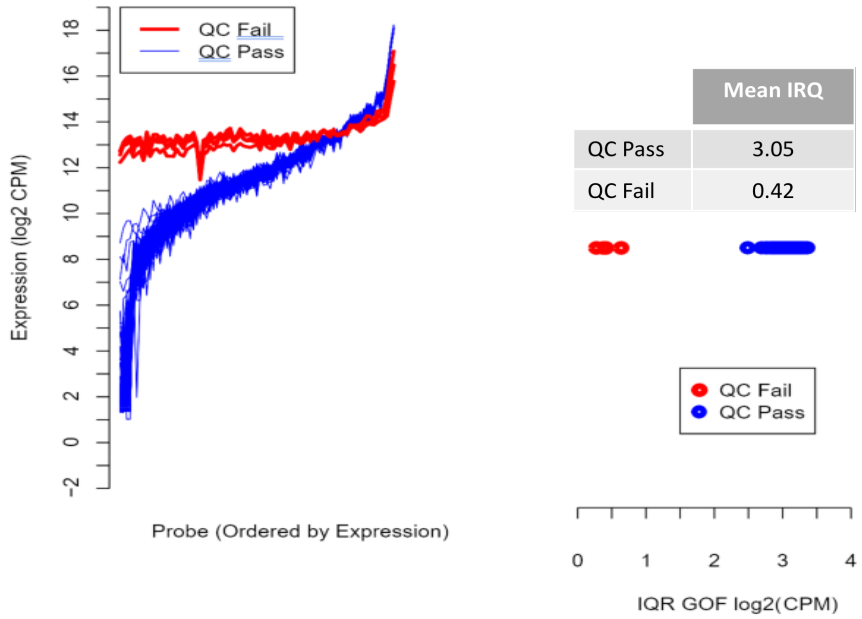

Figure 1. Two exemplar plots for qualitative (left side) and quantitative (right side) evaluation of the GOF test statistic for identification of non-biologic expression used to identify QC failures.

HTG Molecular

HTG Molecular Diagnostics, Inc.
3430 E Global Loop, Tucson, AZ 85706, USA
Call 1-877-289-2615 Fax 1-520-547-2837
info@htgmolecular.com or feedback@htgmolecular.com
Service and Support 1-877-507-3259
Orders orders@htgmolecular.com