**Introduction**

Gene expression profiling is being widely applied in cancer research to identify biomarkers for clinical endpoint prediction such as diagnosis, prognosis, or prediction of therapeutic response. The advent of RNA deep sequencing (RNA-Seq) technologies has revolutionized eukaryotic transcriptome analysis by providing the ability to decipher global gene expression patterns beyond the limitation of microarrays, including the ability to discover novel genes and splice variants. However, data generated by the Sequencing Quality Control (SEQC) consortium has shown that RNA-Seq and microarray based gene expression models are comparable in clinical endpoint prediction performance, even when the additional information unique to RNA-Seq is included in the modeling stage.

In order to understand and characterize the reproducibility and validity of gene expression experiments, exogenous RNA spike-in controls and associated ratio performance metrics have been introduced by the External RNA Controls Consortium (ERCC). The ERCC developed a library of 98 exogenous control transcripts from synthetic, bacterially expressed, and oligo-RNA pools to serve as defined abundances and ratios, providing data that measure the technical performance of gene expression profiling methods. Scientists at the National Institute of Standards and Technology (NIST) have created a Bioinformatics software tool called “ERCCdashboard” that produces ERCC performance metrics from expression values such as microarray signals or RNA-Seq read counts. Using the ERCC spike-in control transcripts and the “ERCCdashboard” analysis tool, we can perform the aggregation of Agilent gene expression microarray and Agilent RNA-Seq platforms run with the same total RNA samples, and highlight performance differences between the two different Agilent gene expression measurement technologies.

**Experimental Design**

Two ERCC spike in RNA transcript mixtures (Mix 1 and Mix 2) were obtained from Ambion (p/n 4465719) and 2 µl of each respective spike-in was added to 10 µl of either the Universal Human Reference RNA (UHR), Agilent p/n 012500 or the Human Brain Reference RNA (HBRR) Ambion p/n AM1950 as noted in the above figure. Four (4x) technical replicate 10 ng aliquots of total RNA from each sample were prepared for Agilent One-Color Gene Expression microarray analysis using the Agilent Low Input Quick Assay kit (p/n 51980-08) and run on the SurePrint G3 Human Gene Expression v3 6x60K Microarrays (p/n G4851C). These (3x) technical replicate 2 µg aliquots of total RNA from each sample (one less technical replicate as compared to the figure above) were prepared for RNA-Seq analysis using the Agilent SureSelect Strand Specific RNA Library Preparation kit (p/n G5582A). The barcoded libraries for both the UHRR and HBRR technical replicates were pooled and prepared for paired-end multiplexed sequencing on the Illumina platform. The mean number of mapped reads per indexed library was 151,155,130 for UHR samples and 116,022,753 for HBRR samples.

**Estimating Limit of Detection Rates (LODR)**

This figure demonstrates the ability of each platform to detect small changes in ERCC spike-in concentration between the UHR and HBRR samples. Shown are the average fluorescence intensity (microarrays) or the average counts (RNA-Seq) versus the differential expression test P-values calculated by the ERCCdashboard program. The heavy black dashed line indicates the false discovery rate (FDR) cutoff of FDR = 0.01 chosen for this analysis. Colored arrows indicate the LODR estimate (average intensity or average counts) for each fold change that crosses the line indicating the upper boundary of the 95% confidence interval. LODR results and confidence intervals derived from the plots are shown in the annotation tables below the plots.

**MA Plots Demonstrate Measurement Variability**

Magnitude of fold-change (log ratio) versus average signal/count (MA) plots contain ERCC data points colored by expected ratio with error bars that represent the standard deviation of the replicate ratio measurements. Filled circles indicate ERCC ratios above the limit of detection (LODR) for the 4.1, 1.2, and 1.15 ratios. Endogenous transcript ratio measurements are shown as gray points. The estimate of RNA fraction differences between the UHR and HBRR RNA samples (\(c_1\)) with weighted standard errors is provided in an inset table and is used to adjust the nominal ERCC ratios. The nominal ratios are annotated with the light solid color lines for each ratio subpool and the adjusted ratios are annotated with the heavy dashed lines.

**Conclusions**

- The technology-independent spike-in control transcripts developed by the ERCC can serve as a truth set to benchmark the accuracy of endogenous transcript ratio measurements.
- The ERCCdashboard tool enables rapid, reproducible, and automated analysis of the ERCC spike-in control transcripts in any differential expression experiment.
- The ERCC controls demonstrate that both the Agilent One-Color Gene Expression Microarray platform and the Agilent RNA WT NGS platform can generate linear gene expression data spanning 5 orders of magnitude with the ability to detect sensitive 1.5-fold changes at levels as low as: > 38 counts with the microarray workflow
- > 55 mapped reads with the RNA-Seq workflow

---

Using ERCC Spike-In Control Transcripts Provides Confidence in Agilent Microarray and RNA-Seq Gene Expression Data

Anne Bergstrom Lucas, Namrata Joglekar, Bilan Huse, Bahram Arezi, Mary Napier, Paul Wolber and Scott Happe

Agilent Technologies Genomics R&D, Santa Clara, California