

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

Authors

Anne Bergstrom Lucas, Bilan Hsue,
Bahram Arezi, Paul Wolber,
Scott Happe

Agilent Technologies, Genomics
R&D, Santa Clara, California

Introduction

Gene expression profiling is 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, in addition to the ability to decipher global gene expression patterns, the capability to discover novel genes and splice variants. Data generated by the Sequencing Quality Control Consortium (SEQC) have 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 prediction models¹.

To evaluate the technical performance of both platforms, we performed a series of experiments using spike-in control transcripts to provide information on sensitivity, specificity, dynamic range and much more. The use of commercially available RNA samples and spike-ins allowed us to focus on the technical performance of the platforms, eliminating the bias linked to sample variability.

We used the exogenous RNA spike-in controls that were developed by the External RNA Controls Consortium (ERCC)². Reproducible analysis using a standard set of metrics associated with the ERCC controls was developed into an R-scripted software tool that is widely available from Bioconductor³. This "erccdashboard" software tool was used to generate the metrics used to evaluate the microarray and RNA-Seq gene expression measurement platforms.

Material and methods

In order to evaluate microarray and RNA-Seq performance, we conducted a series of experiments using the same two RNA samples from the FDA-led MAQC and SEQC projects. The use of commercial RNA samples makes it easier for any lab to reproduce these experiments and use them as a starting point to evaluate the platform and perform quality control experiments in their laboratory. The two selected samples are: Universal Human Reference RNA (UHRR, Agilent p/n 740000) and the Human Brain Reference RNA (HBRR, Ambion p/n AM6050).

In order to add positive control transcripts to this experiment, we spiked the ERCC controls into the total RNA at defined abundances and ratios, providing data that measures the technical performance of gene expression profiling methods. The ERCC consortium developed a library of 96 exogenous control transcripts from synthetic or bacterial sequences that can be added to RNA samples to perform comprehensive comparative experiments².

Two ERCC spike-in RNA transcript mixtures (Mix 1 and Mix 2) were obtained from Ambion (p/n 4456739) and are composed as follows:

- The 92 ERCC RNA control transcripts are divided into 4 different sub-pools (a - d), with each sub-pool containing 23 transcripts. The sub-pools are combined to create two mixtures, defined as Mix 1 and Mix 2 in four defined abundance ratios of 4:1, 1:2, 1:1.5, and 1:1 between the mixtures (Figure 1a).
- The 23 transcripts present in each of the four sub-pools within the two ERCC mixes span a large dynamic range of 2^{20} or 6 orders of magnitude as illustrated in Figure 1b.

These ERCC control transcript mixes are technology-independent and can be spiked into the RNA at the beginning of any gene expression analysis experiment.

For this experiment, the two ERCC spike-in RNA transcript mixtures (Mix 1 and Mix 2) were diluted 1:10, and 2 μ l of each respective spike-in mix was added to 10 μ g of either the Universal Human Reference RNA (UHRR, Agilent p/n 740000) or to the Human Brain Reference RNA (HBRR, Ambion p/n AM6050).

Four (4x) technical replicate 100 ng aliquots of total RNA from each sample were prepared for Agilent one-color gene expression microarray analysis using the Agilent Low Input Quick Amp kit

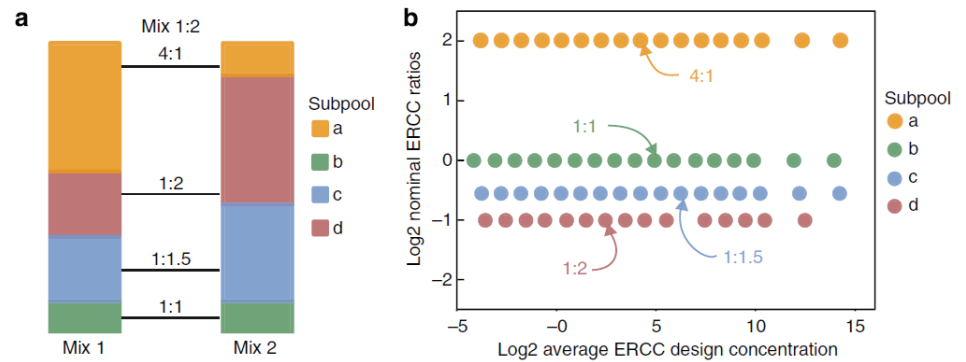


Figure 1. ERCC RNA ratio/abundance in mixtures. Schematic representation of spike-in mix preparation and final relative concentration of the different sub-pools. Figure reproduced from *Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures*, <https://www.nature.com/articles/ncomms6125.pdf>³

(p/n 5190-2305) and run on the SurePrint G3 Human Gene Expression v3 8x60K Microarrays (p/n G4851C).

Three (3x) technical replicate 2 μ g aliquots of total RNA from each sample were prepared for RNA-Seq analysis using the Agilent SureSelect Strand-Specific RNA Library Preparation Kit (p/n G9691A/G9692A). 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 UHRR samples and 115,602,753 for HBRR samples.

Analysis – *erccdashboard* tool

In order to facilitate data analysis, we employed a Bioconductor tool developed by a team of scientists from the National Institute of Standards and Technology (NIST) called “*erccdashboard*” that produces ERCC performance metrics from expression values such as microarray signals or RNA-Seq read counts³.

Thanks to the *erccdashboard* tool, any laboratory can easily perform technical analysis of platform performance by simply adding the ERCC spike-in control transcripts to their samples before performing the experiments.

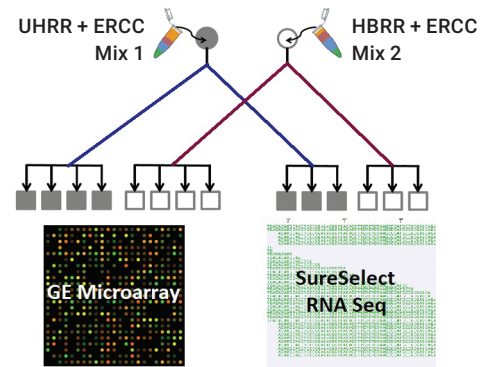


Figure 2. Schematic representation of the experimental design. There were four 100 ng technical replicates for each RNA for the microarray experiment and three 2 μ g technical replicates for each RNA for the RNA-Seq experiment. All RNA aliquots used in this experiment were derived from the same original pooled RNA samples.

Using the *erccdashboard* analysis tool, we compare the performance of the Agilent gene expression microarray and Agilent RNA-Seq platforms run with the same total RNA samples and were able to easily process data and highlight performance differences between the two different Agilent gene expression measurement technologies. More details on the analysis performed are included in the discussion of results.

The *erccdashboard* software package was developed in the R statistical scripting language and the package is

freely available from GitHub (<https://github.com/usnistgov/erccdashboard>) and Bioconductor.

Results & Discussion

All samples run both on microarray and NGS provided high-quality results and were analyzed using the *erccdashboard* tool. We focused the analysis on 4 main technical parameters:

- Dynamic range
- Sensitivity
- Specificity
- Reproducibility

Dynamic Range

Erccdashboard derived signal-abundance plots compare the dynamic range capabilities of the two different gene expression experimental platforms using the same RNA samples. In Figure 3, the points are colored by ratio sub-pool, error bars represent the standard deviation of the replicate ratio measurements, and shape represents the sample type (UHRR or HBRR). In the RNA-Seq results, ERCC controls that did not have at least one count in three libraries for either sample were not included in the signal-abundance plot.

It is interesting to note that the microarray platform has higher sensitivity at the lower end of the dynamic range as compared to the RNA-Seq data, resulting in a higher dynamic range for the microarrays as compared to the RNA-Seq data when averaging over 100M reads per indexed sample.

Sensitivity & Specificity

Sensitivity is calculated taking into consideration both estimated Limit of Detection Rates (LODR) and False Discovery Rate (FDR) cut off. This analysis demonstrates the ability of each platform to detect small changes in ERCC spike-in concentration between the UHRR and HBRR samples.

Results are summarized in Figure 4.

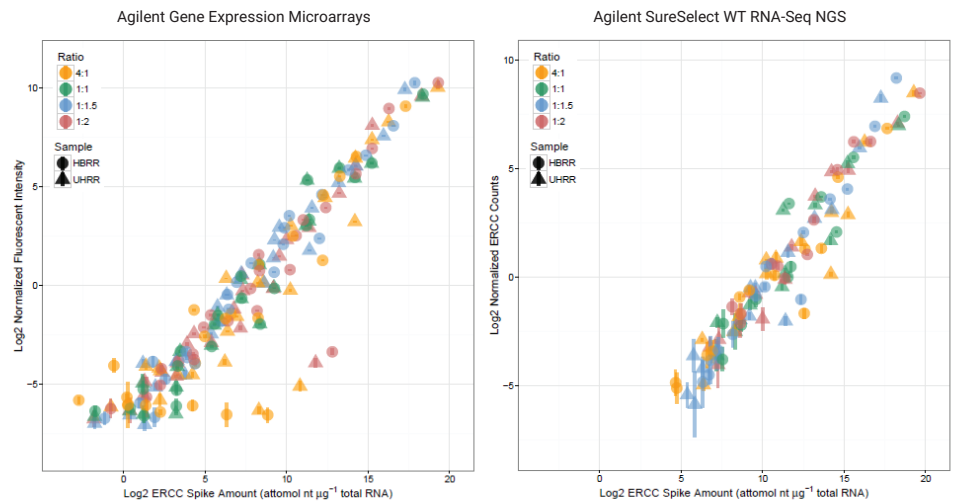
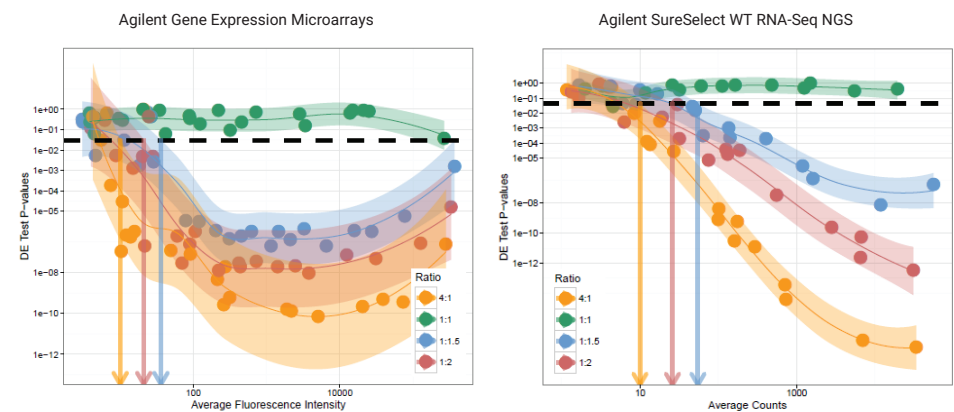


Figure 3. ERCC signal-abundance plot comparisons. While the microarray experiments are able to accurately detect the ERCC transcripts that are spiked at the lower concentrations (less than 5 attomole per microgram total RNA) as compared to the RNA-Seq measurements, there are several ERCC control transcripts that consistently give lower than expected signals (ERCC-00022, ERCC-00131, ERCC-00144) on the microarray experiments. **Note:** the linear dynamic range on these plots spans 2^{20} (microarrays) or 2^{16} (RNA-Seq), with microarrays showing higher dynamic range with respect to a whole transcriptome RNA-Seq experiment that averaged over 100M reads/sample.



RATIO	LODR ESTIMATE		90% CI LOWER BOUND		90% CI UPPER BOUND	
	Microarray	RNA Seq	Microarray	RNA Seq	Microarray	RNA Seq
4:1	10	10	7.2	6.3	13	11
1:1.5	36	55	20	30	43	62
1:2	21	26	12	12	24	32

Figure 4. Estimating Limit of Detection Rates (LODR). 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 90% confidence interval.

From this analysis of the ERCC data, we can conclude that we need at least 10 counts on a microarray or 10 mapped reads in the RNA-Seq experiment to have 99% confidence in a gene expression ratio difference of one to four between the two RNA samples. For confidence in detecting a 1:2 ratio, we would need 21 counts on the microarray and 26 mapped reads on the RNA-Seq experiment, and for confidence in detecting a very small 1:1.5 ratio difference between the two samples, we would need 36 counts on the microarray and 55 mapped reads on the RNA-Seq experiment. As demonstrated by the ERCC spike-in controls, both Agilent gene expression measurement platforms provide high confidence in detecting very small fold-change differences between the two RNA samples.

Reproducibility

Magnitude of fold-change (log ratio) versus abundance or average signal/count (MA) plots illustrate the differences between the two RNA samples analyzed in this experiment. The MA graphs in Figure 5 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 the 1:1.5 ratios. Endogenous biological transcript ratio measurements are shown as gray points.

Because the two RNA samples tested contain different amounts of endogenous mRNA to total RNA, the estimate of mRNA fraction differences between the UHRR and HBRR RNA samples (rm) are calculated by the *erccdashboard* tool with weighted standard errors provided in an inset table. The nominal ratios are annotated with the light solid color lines for each ratio sub-pool and the adjusted ratios are annotated with the heavy dashed lines.

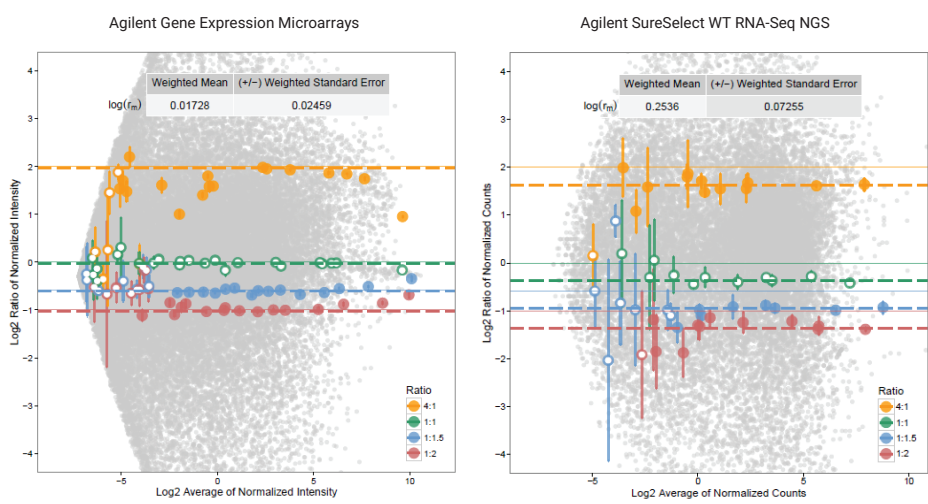
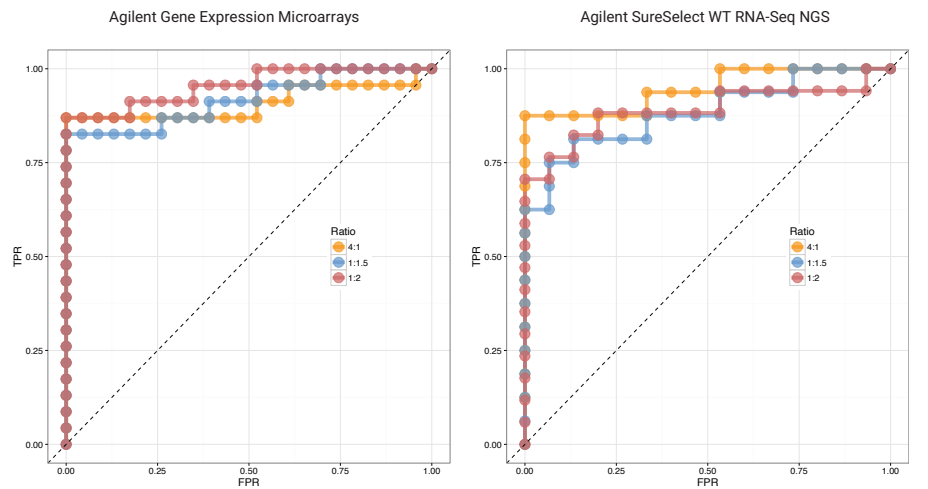


Figure 5. The magnitude versus abundance (MA) plot shows ratio measurement variability between the platforms. The x-axis represents the average gene expression abundance value in each platform (either normalized intensity or normalized counts), and the y axis shows the log₂ expression ratios between the two RNA samples. Filled symbols indicate that statistically significant ratios were generated between the two samples based on the LODR analysis in Figure 4 and open symbols indicate no significant differential expression between the two samples. As expected, the green-colored 1:1 ratio ERCC controls and other ERCC transcripts spiked at the lowest concentrations have open symbols. Color code, error bars and nominal vs normalized ratio lines highlight the performance differences between microarrays and RNA sequencing.



RATIO	AUC		DETECTED		SPIKED	
	Microarray	RNA Seq	Microarray	RNA Seq	Microarray	RNA Seq
4:1	0.909	0.946	23	16	23	23
1:1.5	0.919	0.883	23	16	23	23
1:2	0.955	0.890	23	17	23	23

Figure 6. ROC Curves and AUC Statistics. ROC curves and the corresponding areas under the curve (AUC) statistics change based on the discrimination of true-positive and true-negative values, where ideal performance is represented by AUC = 1 for all 23 spike-in controls spiked at a given ratio.

As highlighted both from the difference between nominal and normalized ERCC ratios and the variability in the low expressed genes, both platforms perform efficiently on genes expressed at high and average expression levels, while gene expression microarrays provide more stable and precise results measuring the low expressed genes.

Another important parameter to consider when evaluating technologies is their ability to detect real changes in expression in the samples under evaluation. We can use the true-positive and true-negative ERCC ratios to generate receiver operator characteristic (ROC) curve analysis of differential gene expression, as well as calculate the corresponding area under the curve (AUC) statistics. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the higher the AUC, the better the model is at distinguishing between patients with disease or without disease.

It is interesting to note that the microarray experiments were able to detect all 23 spike-in controls at each fold-change tested with AUCs >90%, whereas the RNA-Seq experiments at a read depth averaging >100M reads detected only 16 or 17 of the 23 spike-in controls with AUCs <90% for the smaller fold changes (1:2 fold and 1:1.5 fold) tested.

Conclusions

The technology-independent spike-in control transcripts developed by the ERCC, when included in Agilent gene expression measurement experiments, can serve as a truth set to benchmark the accuracy of endogenous transcript ratio measurements. Interpreting the data from the ERCC spike-in controls is made easier by the *erccdashboard*

tool, a free of charge analysis package that enables rapid, reproducible, and automated analysis of the ERCC spike-in control transcripts in any differential expression experiment, including both microarrays and RNA sequencing.

Including the ERCC controls in this experiment allowed us to 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 36 counts with the microarray workflow and 55 mapped reads with the RNA-Seq workflow. Both platforms show similar technical ability in discriminating gene expression changes in analyzed samples, when starting from 100 ng RNA for microarrays and 100M reads for RNA sequencing, with microarrays showing better sensitivity and specificity at detecting low expressed genes.

When performing gene expression experiments, it is important that orthogonal measurement tools provide concordant data. The use of the ERCC spike-in control transcripts and subsequent data analysis using the *erccdashboard* software in this experiment demonstrate that these two different platforms generate comparable results and highlight the strengths of each measurement tool. These results demonstrate the power of Gene Expression microarrays and the advantages of selecting the right platform for any experimental design. This enables you to plan experiments and to maximize resources and minimize costs.

Acknowledgments

The authors would like to thank Namrata Joglekar for her contributions to this application note.

References

1. Zhang *et al.*, **Comparison of RNA-seq and microarray-based models for clinical endpoint prediction** (2015) *Genome Biology* 16:133
2. Pine, *et al.* **Evaluation of the External RNA Control Consortium (ERCC) reference material using a modified Latin square design** (2016) *BMC Biotechnology*, 16:54
3. Munro, S., *et al.* **Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures** (2014) *Nature Communications*, 5:5125

www.agilent.com

For Research Use Only. Not for use in diagnostic procedures.

This information is subject to change without notice.

PR7000-2415
© Agilent Technologies, Inc. 2020
Printed in the USA, February 5, 2020
5994-1767EN

