Use of Synthetic Transcript Pools to Evaluate RNA-Seq Performance

Penn Whitley, Luming Qu, Andrew Lemire, Joel Brockman, Sheila Heater, Jeff Schageman, Jian Gu, Kristi Lea, Charmaine San Jose, Natalie Hernandez, Kelli Bramlett, Diane Ilsley, Christopher Mueller and Robert Setterquist, Life Technologies/Ambion R&D, 2130 Woodward, Austin, TX, USA, 78744

INTRODUCTION

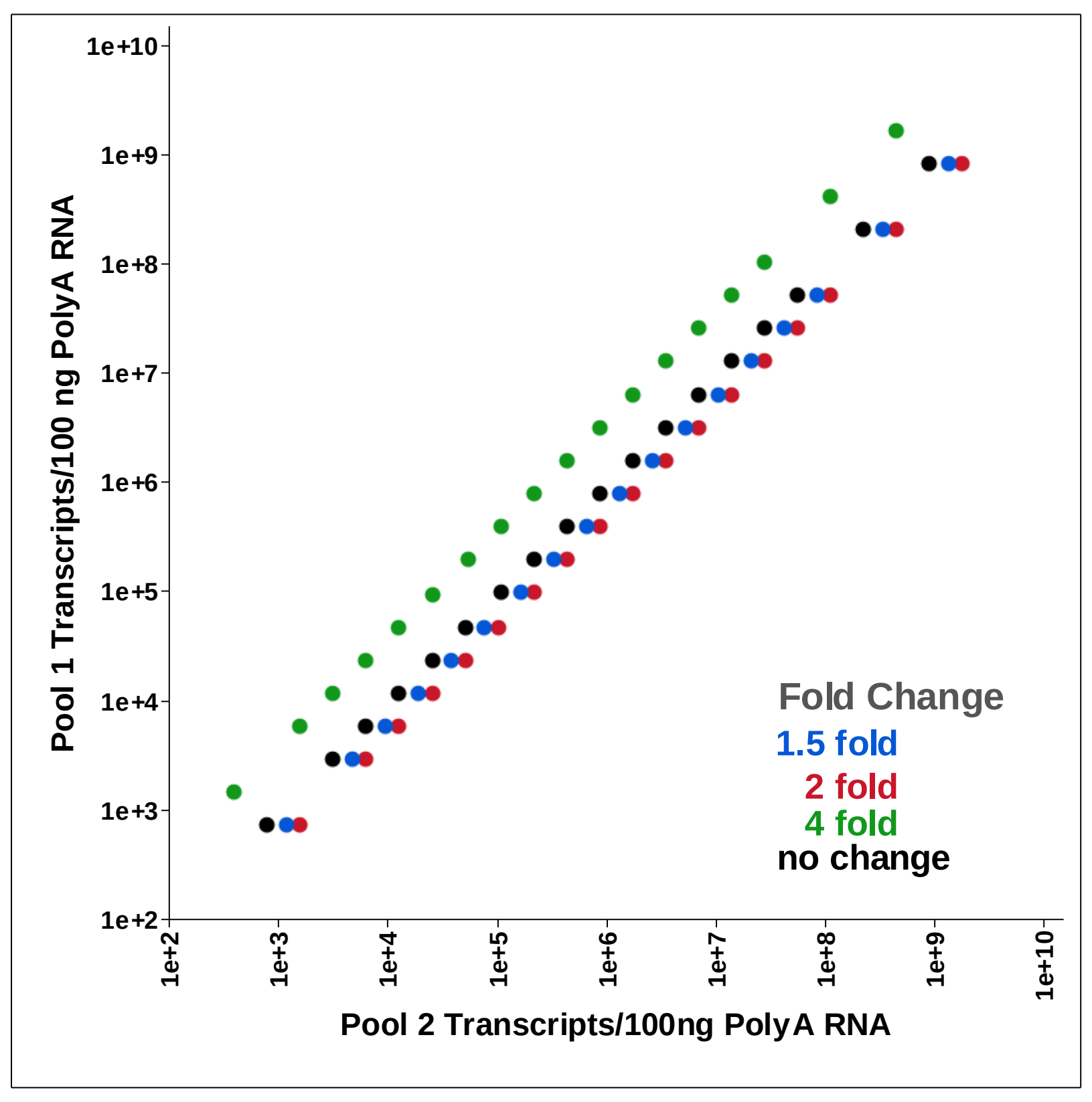
Digital RNA expression profiling with next generation sequencing (NGS) promises to revolutionize the way transcriptomics is performed. RNA-Seq and related methods can be used to not only detect differential expression of genes but also to measure isoform usage, detect mutations and measure allele-specific expression. In addition, RNA sequencing can be used as a discovery tool to characterize un-annotated regions of the genome that are transcribed. Although RNA-Seq has been used in different labs and on a variety of NGS platforms, there has not been a thorough investigation of fundamental analytical performance metrics. In order to evaluate the capabilities of RNA-Seq we have used two synthetic RNA spike pools that are added to the RNA sample prior to library preparation and sequencing. The pools consist of 92 synthetic transcripts designed by the External RNA Control Consortium (ERCC)(1). Using the SOLiD™ System for sequencing, over 350 million reads were generated from HeLa cell polyA RNA spiked with the two synthetic RNA pools. Using this sequence data we were able to precisely measure the sensitivity and dynamic range of detection for this experiment. This model system indicates that NGS offers extremely high precision and accuracy, with no attenuation at the high end of the dynamic range, as seen with analog measurement systems such as microarrays. Also, because of the two pool design used, the accuracy of differential expression measurements between samples can be observed. We have used various diagnostic methods to estimate the sensitivity and specificity of fold-change ratios for the determination of differential expression. Supporting the mission of the ERCC we report how these reagents can provide invaluable information about the performance of new RNA-Seq tools and methods, as well as, provide critical quality control metrics for individual sequencing experiments.

MATERIALS AND METHODS

Synthetic RNA Transcript Pool Preparation

92 plasmid stocks were obtained from the NIST ERCC group (1) and *in vitro* transcribed using standard methods. Figure 1 shows the design of pools 1 and 2. Each pool contains all 92 spikes in 4 distinct sub-pools. These sub-pools consist of 23 spikes, added to the 2 main pools at 4 fold change ratios. Each pool spans a dynamic range >6 logs.

Figure 1. *Two Pool Design*



Whole Transcriptome Library Preparation

HeLa S3 RNA samples were processed using the Ambion® Poly(A)Purist™ Kit to obtain poly(A) mRNA. Roughly 3 ng of each spike pool was added per 100ng mRNA prior to library preparation. This amount was chosen to be representative of the dynamic range of transcripts in real tissue samples. The spikes range from <1000 to >1 trillion copies in 100ng mRNA. The SOLiD™ Whole Transcriptome Analysis Kit was then used for library preparation as recommended except that the gel enrichment step was not performed. Two library replicates were performed for each HeLa/spike pool combination.

SOLiD™ System Sequencing and Analysis

Whole transcriptome libraries prepared with the SOLiD™ Whole Transcriptome Analysis Kit were amplified onto beads by emulsion PCR using recommended SOLiD™ System sequencing protocols. Enriched beads were deposited onto glass slides using the quad partition gasket (each replicate sample was deposited on 2 slide quads) and sequenced using the SOLiD™ 3 System and 50 bp reads. A total of >350 million total reads and ~94 million uniquely mapped reads per sample were generated. The Applied Biosystems Whole Transcriptome v1.0 Analysis Pipeline was used to map reads to the hg18 genome and spike sequences. The CountTags module of the Analysis Pipeline was then used to assign counts to Refseq (UCSC) and the synthetic spikes. Reads per Kilobase (RPKM) per Million mapped was calculated based on uniquely mapped reads (2).

RESULTS

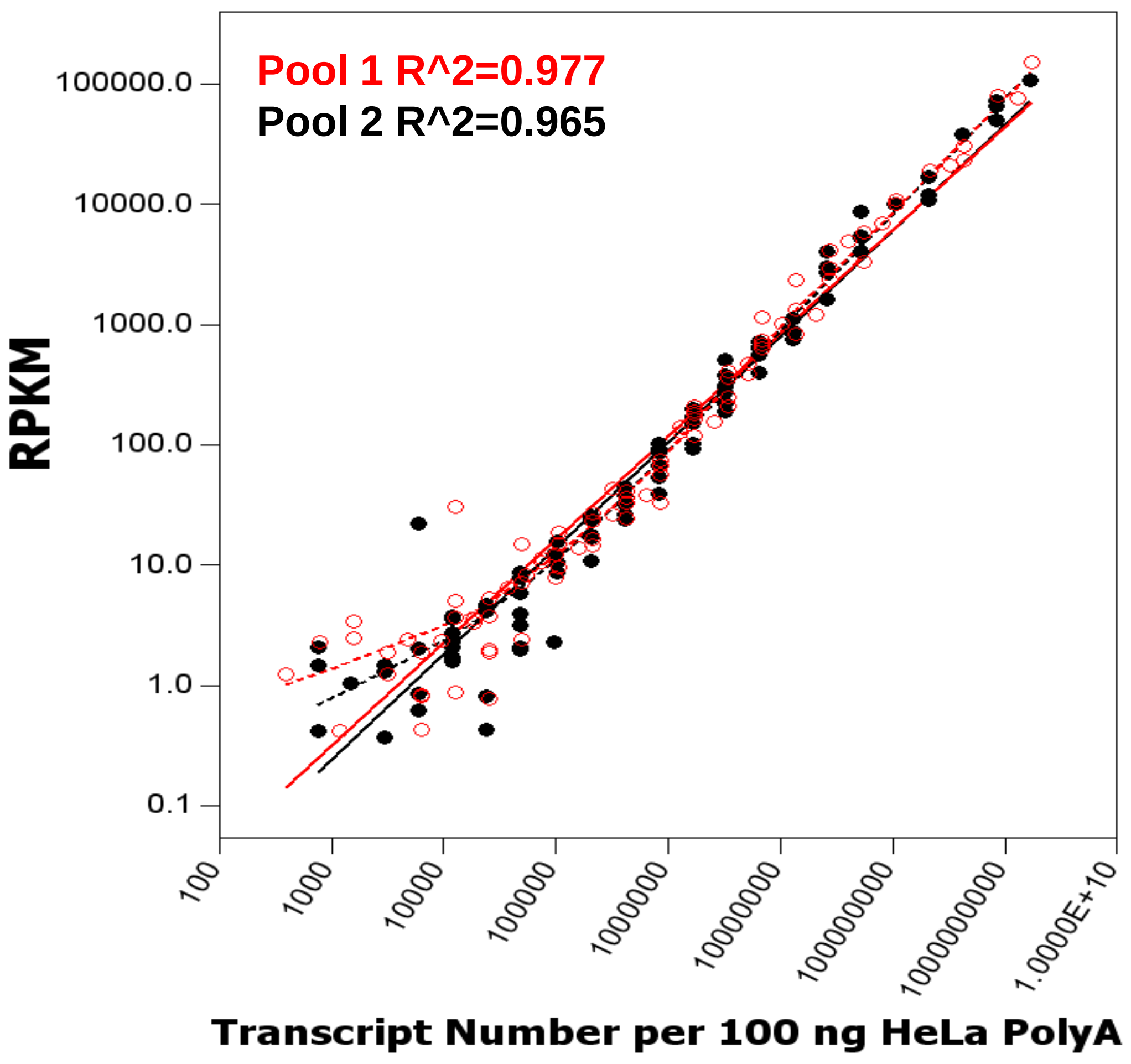


Figure 2. *Dose Response for 92 Spike-ins*

The scatterplot shows dose response data for 2 independent pools of 92 synthetic transcripts designed by the ERCC (1). The 2 pools were spiked into polyA HeLa RNA with each spike at the levels indicated on the X axis. Roughly 47 million uniquely mapped reads were generated for each pool (Pool1 and Pool 2). High linearity extends through >5 logs of dynamic range with no attenuation at the high end as is seen for analog microarray measurements where there is only 2-3 logs of linear range. Solid lines indicate linear regression fit and dashed lines are Lowess smoothing curves.

RPKM Detection Threshold	# of Synthetic Transcript Molecules Detected Above Threshold	# of transcripts per 100 ng polyA RNA	Complexity Ratio	mRNA Copies/Cell Detection
1	4,500	9.E+10	1:20,000,000	0.015 copies/cell
5	28,000	9.E+10	1:3,300,000	0.09 copies/cell
10	63,000	9.E+10	1:1,500,000	0.2 copies/cell

Table 1. *Detection Sensitivity*

Detection rates for 3 RPKM thresholds was estimated based on the average least squares fit from figure 1. The number of mRNA molecules in 100ng polyA was calculated based on an average transcript length of 2 Kb. The complexity ratio is simply the # of mRNA molecules divided by the # of spikes detected. The detection rate expressed as copies per cell was then estimated based on an average transcript number of 300,000/cell (3).

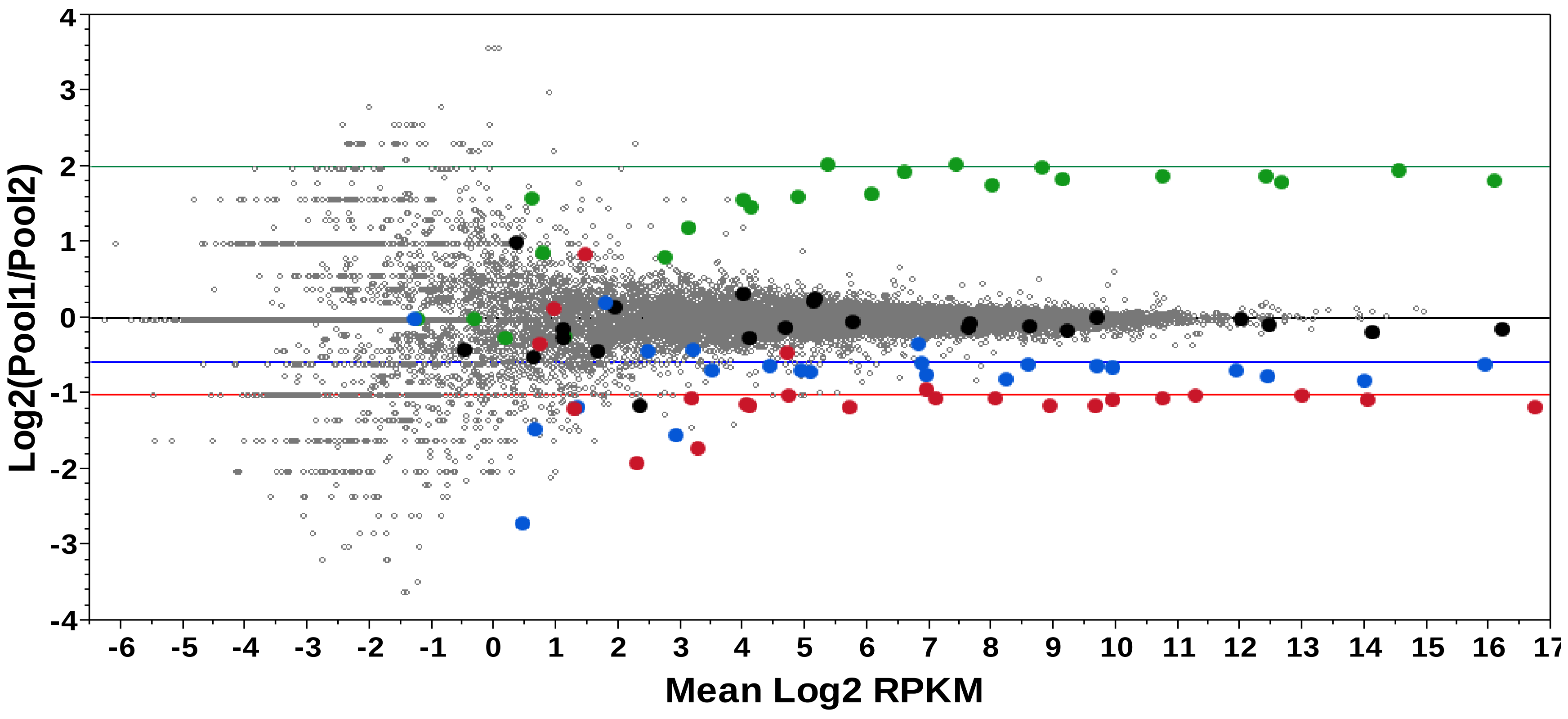


Figure 3. *MA Plot Comparing Spike Pool Ratios*

The MA Plot shows ratio-metric performance of the 2 ERCC pools. The pools are divided equally into 4 sub-pools of 23 spikes each. These sub-pools are designed to evaluate 1.5, 2 and 4 fold change, as well as, no change between the 2 ERCC pools. As can be seen the different ratios are performing as predicted throughout most of the dynamic range. The spikes are designed to extend through 20 log2 units. The spikes are very closely approximating expectation with at least 18 log2 units of dynamic range. As was seen in fig. 6, no high end saturation was observed for the ratios, as with array data (4). Green=4 fold, red=2 fold, blue=1.5 fold, black=no change, grey=RefSeq Genes. Corresponding colored lines indicate expected Log2 values for the 2 pools.

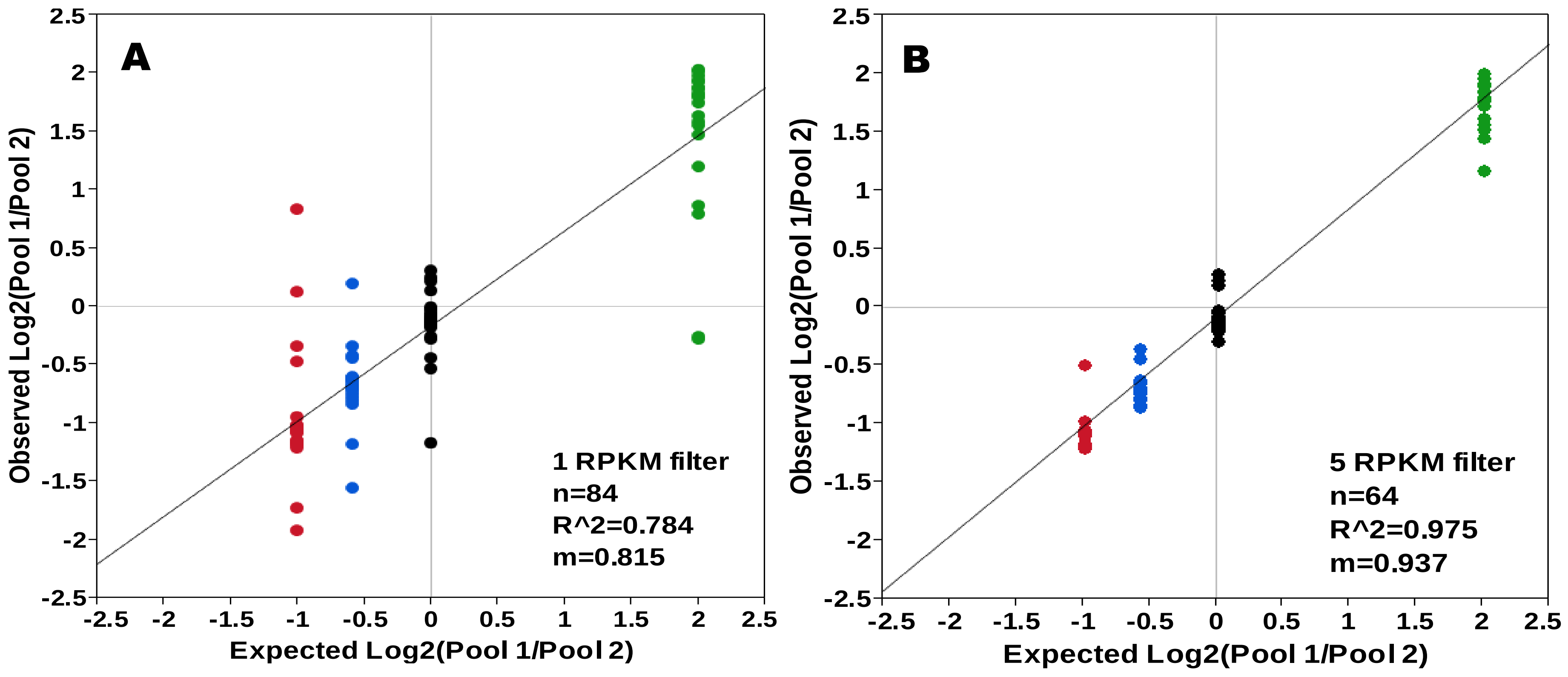
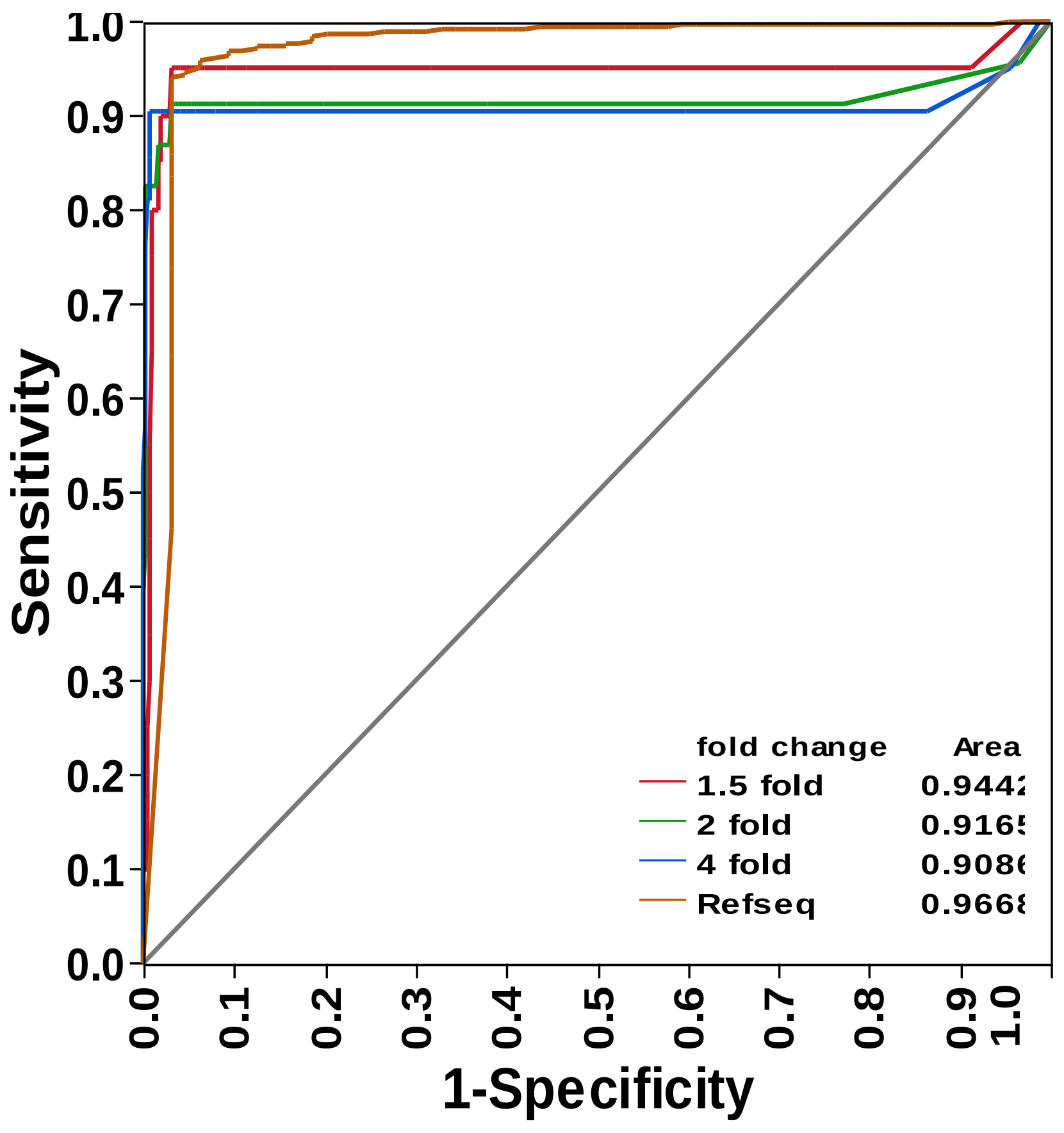


Figure 4. ***Expected vs. Observed fold change estimates with two RPKM thresholds***

The log ratio of pool 1 vs. pool 2 was calculated for all 92 spikes across the two pools. The observed values were then compared with the expected log fold change values based on the concentration of each spike in the two pools. A. Acceptable accuracy (R^2=0.784, slope=0.815) was found when both samples passed a filter of >1 RPKM. B. Accuracy improved dramatically (R^2=0.975, slope=0.937) after using a 5 RPKM threshold. As previously suggested (2) values of ~ 1 RPKM provide a sensible threshold for reliable detection and may also be applied to filter datasets used for differential expression analysis.

Figure 5. ***Receiver Operator Characteristics (ROC) Analysis***



A ROC analysis was carried out to assess the performance of RNA-Seq in determining differential expression. We used the estimated log2 fold change ratio as a diagnostic rule for determining differential expression as in (4). Since only spike-in transcripts are differentially expressed in this experiment it is easy to assess the true positive rate (TP) for detection of differential expression. The false positive rate (FP) was estimated using Refseq as an unchanging control group (derived from the parental HeLa sample). The ROC curves are created by plotting the TP and FP rates for all possible Log fold change values. Area under the curve (AUC) is then used to assess performance of each nominal fold-change group (1.5, 2, 4-fold) and for all groups combined (referred to as Refseq in legend). All AUC values exceed .9, indicating extremely high sensitivity and specificity for this experiment. A 1 RPKM filter was applied for this analysis.

CONCLUSIONS

As RNA sequencing becomes more common a better understanding of the performance of these methods will be increasingly important. Synthetic spikes have proven to be useful tools for evaluating expression analysis platforms like microarrays and this study demonstrates their usefulness with RNA-Seq. We have shown detection levels of less than .015 transcript copies/cell with fewer than 50 million uniquely mapped reads. This is roughly 70X more sensitive than current microarray technology. We have also demonstrated >5 logs of actual dynamic range with almost no compression relative to known input concentrations. Using a variety of traditional diagnostic methods we also shown that RNA sequencing results in very accurate estimates of differential expression at the transcript level. We envision these spikes becoming a standardized tool for library quality assessment and sequencing performance.

REFERENCES

1. External RNA Controls Consortium (2005) Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics 6:150.
2. Mortazavi A et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5(7):621.
3. Hashimoto H et al. (2009) High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. *PLoS One* 4:e4108.
4. Cope LM et al. (2003) A Benchmark for Affymetrix GeneChip Expression Measures. Bioinformatics 1(1):1-10.

TRADEMARKS/LICENSING

For Research Use Only. Not intended for animal or human therapeutic or diagnostic use.

Life Technologies • 5791 Van Allen Way • Carlsbad, CA 92008 • www.lifetech.com