White Paper
Method Comparison between RNA-Seq and the HTG EdgeSeq™ Platform

Abstract: Next-Generation Sequencing (NGS) has emerged as a powerful tool with the ability to generate an unprecedented quantity and quality of data. Over time, RNA-Seq has established itself as the gold standard for the NGS-based quantification of gene expression but, it requires high sample input amounts, has bias caused by RNA extraction, and data analysis pipelines are time-consuming and complicated.

The data presented here show a high degree of sample-level correlation between RNA-Seq and HTG EdgeSeq gene expression profiling panels. Overall, most samples exhibited a high degree of correlation from sample to sample, and correlation was improved when both platforms used extracted RNA. Use of extracted RNA, however, caused a spike in the number of genes receiving zero counts on both platforms, suggesting that while correlation is improved, the increase in correlation comes at the expense of a loss of signal, and possibly increased Limit of Detection (LoD), for many genes.

The HTG EdgeSeq platform can be used as a competitive alternative to RNA-Seq, with several distinct advantages. HTG EdgeSeq panels require less sample input, their workflow is significantly faster, and it has a fully integrated web-based data analysis pipeline. Together, the data in this paper show that the HTG EdgeSeq gene expression profiling panels use less sample and provides researchers critical answers faster.

Introduction

Next-Generation Sequencing (NGS) has emerged as a powerful tool with the ability to generate data of unprecedented quantity and quality. What began with the amplification of a single RNA target to understand transcriptional changes has evolved with the current ability to evaluate the transcription of thousands of genes simultaneously. RNA analysis is critical for the quantification of gene expression levels that can be used to predict prognosis, and therapeutic response[1-3].

Over time, RNA-Seq has established itself as the gold standard for the NGS-based quantification of gene expression, and for good reason. RNA-Seq allows for nucleotide-level quantification of RNAs in certain cases without prior knowledge of the genome or transcriptional targets. This is because RNA-Seq can detect most transcripts in a sample using non-targeted technology[4-5].

Unfortunately, this comes with several distinct drawbacks. First, RNA-Seq requires higher amounts of input from a given sample. One reason for this is that RNA-Seq requires RNA extraction before samples can be analyzed. RNA extraction was designed to remove degraded RNA and impurities in a sample; however, it has the potential to create bias due to removal of small, partially degraded, and low abundance RNAs[6]. Second, as RNA-Seq is not targeted, sequencing output reads may be "wasted" on genes of little to no interest. Third, the RNA-Seq workflow can be time consuming, taking up to two weeks including sample preparation and sequencing time. Lastly, the bioinformatics pipelines for RNA-Seq analysis are complicated and not standardized, posing a significant problem for researchers and clinicians.

The HTG EdgeSeq technology addresses several of these limitations. First, HTG EdgeSeq assays use an extraction-free chemistry for sample preparation. This is an important improvement over conventional RNA-Seq as it eliminates the risk of RNA extraction bias, which results from the removal of small or partly degraded RNA species during the extraction process. The HTG EdgeSeq technology uses quantitative Nuclease Protection chemistry to detect RNA species of 50 nucleotides or more, meaning that short or fragmented RNAs normally removed during RNA extraction are instead measured. The extraction-free process means RNA is not lost from the sample, so less sample input is required to generate equivalent amounts of RNA. Second, the HTG EdgeSeq platform is a targeted gene expression profiling platform that only generates information for specific transcripts in a sample, meaning fewer reads are used up on sequencing RNAs of little or no interest. Third, the HTG EdgeSeq workflow can be completed in as little as 36 hours, including sample preparation and sequencing. Lastly, the HTG EdgeSeq platform employs a fully integrated web-based data analysis package, called HTG EdgeSeq Reveal, taking advantage of a fully automated and standardized bioinformatics pipeline.

The purpose of this study was to compare the HTG EdgeSeq Oncology Biomarker Panel (OBP) and HTG EdgeSeq Precision Immuno-Oncology Panel (PIP) to the Illumina TruSeq RNA

Access (RNA-Seq) Assay for gene expression profiling of tumor tissue specimens. The data presented here show a high degree of sample-level correlation, suggesting that HTG EdgeSeq technology can be used as a competitive alternate to RNA-Seq with distinct advantages.

Methods

Samples

A total of 1,200 samples were tested on each of three assays: the HTG EdgeSeq PIP, HTG EdgeSeq OBP, and RNA-Seq (Illumina TruSeq RNA Access). Data generation, for both HTG EdgeSeq OBP and HTG EdgeSeq PIP, was performed at the HTG VERI/O commercial laboratory in Tucson, AZ and all RNA-Seq data were generated at Q2 Solutions, Morrisville, NC. Samples were formalin- fixed, paraffin-embedded (FFPE) tissue samples collected from patients with colorectal cancer (CRC), gastric cancer (GC), and ovarian cancer (OVC). Correlation analysis was limited to the overlapping samples that passed Quality Control metrics for each panel and to overlapping genes from each panel. See *Table 1* for a summary of samples tested, sample pass rate, and samples and genes that overlap.

Table 1. *Summary of samples, sample pass rate, sample/gene content overlap.*

Assay	Samples Tested	Sample Pass Rate	Samples Common with RNA-Seq	Genes Common with RNA-Seq
HTG EdgeSeq OBP	1,200	96.9%	1,070	2,498
HTG EdgeSeq PIP	1,200	97.3%	1,073	1,347
RNA-Seq: *Illumina TruSeq RNA Access*	1,200	91.4%		

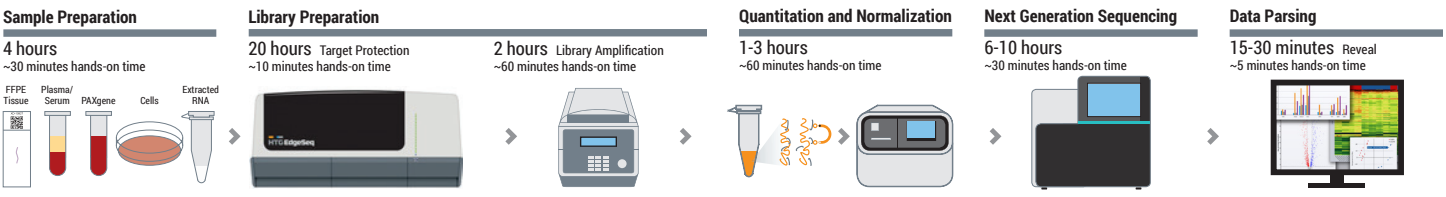
RNA-Seq Workflow

Samples for use in RNA-Seq must first go through RNA extraction. For this study, RNA extraction was done using the QIAGEN RNeasy RNA extraction kit (Qiagen; Germantown, MD). After RNA extraction, the RNA samples are fragmented in preparation for first-strand synthesis. First-strand cDNA synthesis is primed from the extracted RNA using random primers, followed by the generation of second-strand cDNA. Double-stranded cDNA undergoes end-repair, A-tailing, and ligation of adapters that included index sequences. The resulting molecules are amplified via polymerase chain reaction (PCR), their yield and size distribution are determined, and their concentrations are normalized in preparation for the enrichment step. Libraries are enriched for the mRNA fraction by positive selection using a mixture of biotinylated oligonucleotides corresponding to coding regions of the genome. Targeted library molecules are then captured via the hybridized biotinylated oligonucleotide probes using streptavidin-conjugated beads. After two rounds of hybridization/capture reactions, the enriched library molecules are amplified via PCR. Final libraries are assessed using qPCR for quantitation and Agilent TapeStation for fragment size assessment. Normalized libraries are pooled and sequenced at a plex-level appropriate to the coverage required using the Illumina HiSeq Next Generation Sequencer.

HTG EdgeSeq Workflow

HTG EdgeSeq libraries are generated by following the HTG EdgeSeq workflow *(Figure 1)*. HTG Lysis Buffer is added to lyse the FFPE tissue sample making the RNA available to subsequently bind to corresponding target-specific Nuclease Protection Probes (NPPs). Target hybridization is done on the HTG EdgeSeq processor. After the quantitative Nuclease Protection Assay is complete, samples are used as a template in individual PCR reactions with specially designed primers. These primers share common sequences that are complementary to 5'-end and 3'-end sequences of the probes and common adaptors required for cluster generation on an Illumina NextSeq sequencing platform. In addition, each tag contains a unique barcode that is used for sample identification and demultiplexing. After the PCR amplification is finished, a clean-up procedure is performed to remove unincorporated primer tags from PCR products (referred to as a library). The library concentrations are determined by quantitative PCR (qPCR), and normalized libraries are pooled and sequenced at a plex-level appropriate to the coverage required using the Illumina NextSeq Sequencer.

Figure 1. *HTG EdgeSeq workflow*



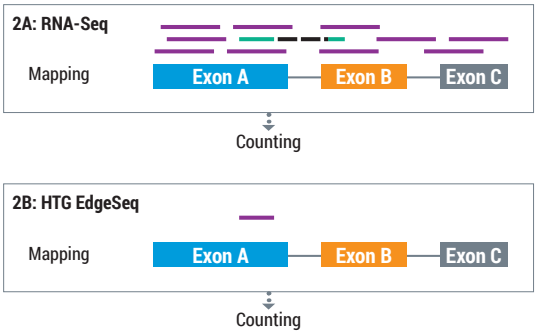


Figure 2. *(A) The RNA-Seq workflow generates multiple counts, or fragment sequences, per gene and must therefore be normalized by gene length (TPM). (B) HTG EdgeSeq workflow generated a single count per gene and is therefore only normalized to read depth (CPM).*

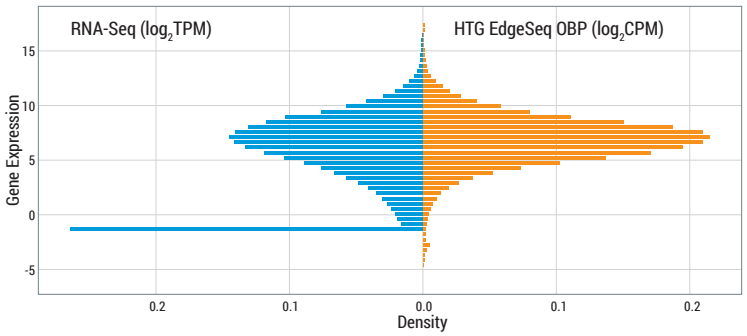


Figure 3. *Overall distribution of samples and genes between the HTG EdgeSeq OBP and RNA-Seq platform. Sample Density is plotted on the x-axis and log_2-transformed Gene Expression is plotted on the y-axis. HTG EdgeSeq data are shown in orange and RNA-Seq data are shown in blue.*

Data Output

Analysis of the RNA-Seq data must take transcript length into consideration. This is accomplished by aligning the fragmented cDNA sequences, as a single copy of target RNA, resulting in multiple output fragments, which are then aligned to the reference genome *(Figure 2A)*. These fragments, which may cover multiple exons of a single transcript, are then counted to predict the number of transcripts in the sample. The Transcripts per Million (TPM) are then calculated by normalizing the raw gene counts by the gene length in bp and then dividing by the total normalized counts of all genes across a sample multiplied by a million.

HTG EdgeSeq probes target a single location of each RNA transcript. This results in a single probe sequence per RNA transcript. The counts therefore are stoichiometrically equal to the number of transcripts in the sample and normalization by transcript length *(Figure 2B)* is not required. The HTG EdgeSeq data is transformed into Counts per Million (CPM). This is done by dividing the number of counts per gene by the number of million counts per sample.

Results

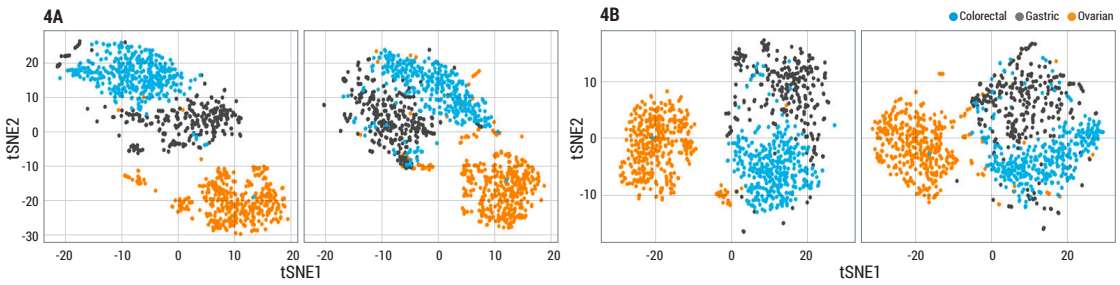
Overall Distribution Evaluation

Despite differences in platform chemistry, data processing and sample type, both platforms share the same intended use, to evaluate RNA expression in each sample. To begin to understand the correlation between the HTG EdgeSeq platform and RNA-Seq platform, a distribution of all overlapping samples and genes was generated *(Figure 3)*. This comparison included only samples tested with the HTG EdgeSeq OBP assay compared to RNA-Seq as these two assays share more overlapping genes and therefore more data points *(Table 1)*.

The expression distribution plot *(Figure 3)* shows a normal distribution of data for both platforms. The most striking difference in the distributions is the high number of genes with zero counts, or no expression, in the RNA-Seq data. While the HTG EdgeSeq data had 0.74% of genes with no detectable expression, the RNA-Seq had approximately 11.67% of genes with no detectable expression. This is measured across all common samples and genes, between the two assays, and the genes that account for the zero-expression vary from sample to sample.

The HTG EdgeSeq data and RNA-Seq data were further analyzed by comparison of sample clustering using t-Distributed Stochastic Neighbor Embedding (t-SNE). The t-SNE, a non-linear dimensionality reduction method, was used to visualize the clustering of RNA- Seq data to both HTG EdgeSeq PIP *(Figure 4A)* and HTG EdgeSeq OBP data *(Figure 4B)*. This analysis shows that the three cancer indications tested here, colorectal cancer, ovarian cancer, and gastric cluster well in both platform comparisons. Together these data show equivalent sample clustering between the methods tested here, based on the gene expression profiles.

Figure 4. *(A) HTG EdgeSeq OBP and (B) HTG EdgeSeq PIP compared to RNA-Seq using t-SNE to visualize the clustering of all common samples. Samples cluster by cancer type including colorectal cancer (blue), gastric cancer (black), and ovarian cancer (orange).*



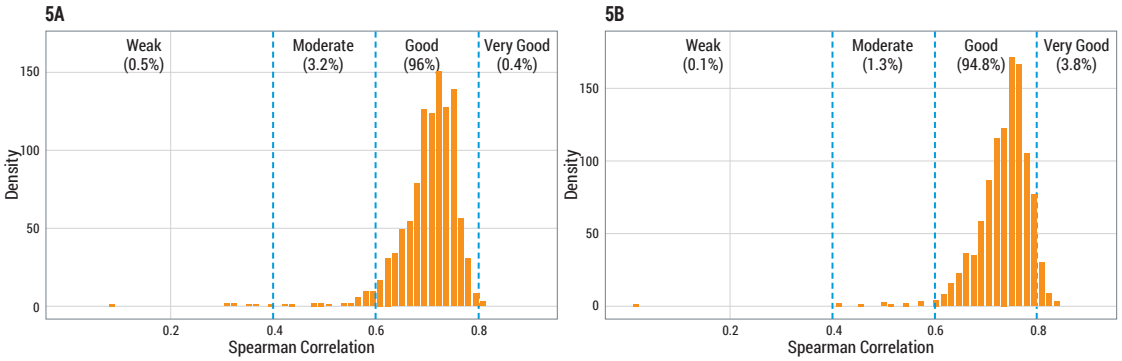


***Figure 5**. Distribution of sample-wise Spearman correlation coefficients for (A) HTG EdgeSeq OBP and (B) HTG EdgeSeq PIP as compared to RNA-Seq. The x-axis shows the value of the Spearman correlation coefficient and the y-axis shows the density of the correlation value. Moderate and weak correlations were attributed to poor sample quality.*

Correlation Evaluation

To understand how well the two platforms correlate, Spearman correlation coefficients were calculated for all samples passing QC metrics between the HTG EdgeSeq OBP *(Figure 5A)* and RNA-Seq, and the HTG EdgeSeq PIP to RNA-Seq *(Figure 5B)*. Samples were stratified into one of four categories based on their correlation: Weak (0-0.4), Moderate (0.4 to 0.6), Good (0.6 to 0.8), and Very Good (above 0.8). These data show that most of the samples tested here correlate well between all platforms, with 96% (OBP) and 94.8% (PIP) of samples showing a Spearman correlation coefficient in the "Good" category (0.6-0.8). Samples showing "Moderate" and "Weak" correlation between the two platforms, accounting for less than 4% of samples, were often identified as the same sample across platform comparisons, suggesting that these low correlation values are likely caused by poor quality samples and not poor correlation between the two platforms. Together, these data further demonstrate the comparability of the two platforms tested here.

To further illustrate the sample-wise correlation between the two platforms, three representative samples were chosen from samples found to have "Good" correlation, *(Figure 5)*. Spearman correlation coefficients were calculated for all common samples for HTG EdgeSeq OBP vs. RNA-Seq *(Figure 6A)* and HTG EdgeSeq PIP vs. RNA-Seq *(Figure 6B)*. All samples were found to have good correlation across the common genes with the average Spearman correlations of 0.68 and 0.73 for comparisons of HTG EdgeSeq OBP to RNA-Seq and HTG EdgeSeq PIP to RNA-Seq, respectively. Three samples were selected for each comparison to show representative samples with "Good" correlations. In each sample comparison, a significant number of genes show zero expression for the RNA-Seq platform, illustrated by the data points clustering at zero reads *(Figure 6*; clustered data points on left of each plot). These data points represent the genes for which there were zero counts in the RNA-Seq platform. This phenotype can be seen in the overall distribution plot *(Figure 3)*, as the spike in zero expression for the RNA-Seq data. These data suggest that these genes have expression levels below the Limit of Detection (LoD) of the RNA-Seq assay.

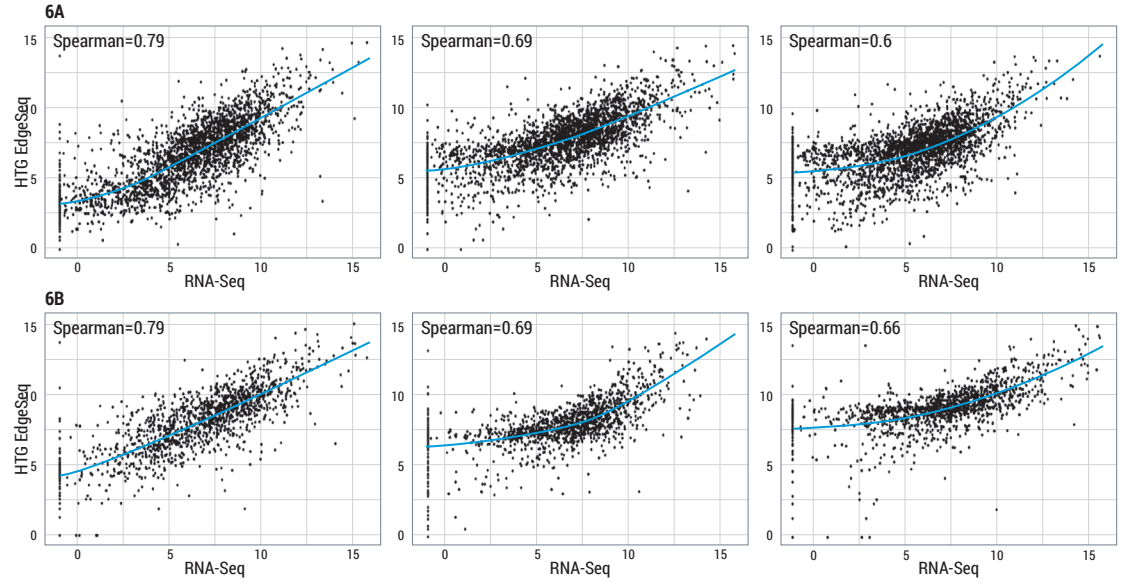


***Figure 6.** Correlation plots for individual samples. The x-axis shows RNA-Seq data in log_2TPM and the y-axis shows HTG EdgeSeq data in log_2CPM. Each data point represents a common gene between RNA-Seq and (A) HTG EdgeSeq OBP or (B) HTG EdgeSeq PIP. Both HTG EdgeSeq assays show good sample-level correlation to RNA-Seq. Clustered data points on the left of each plot show the genes with zero expression for RNA-Seq.*

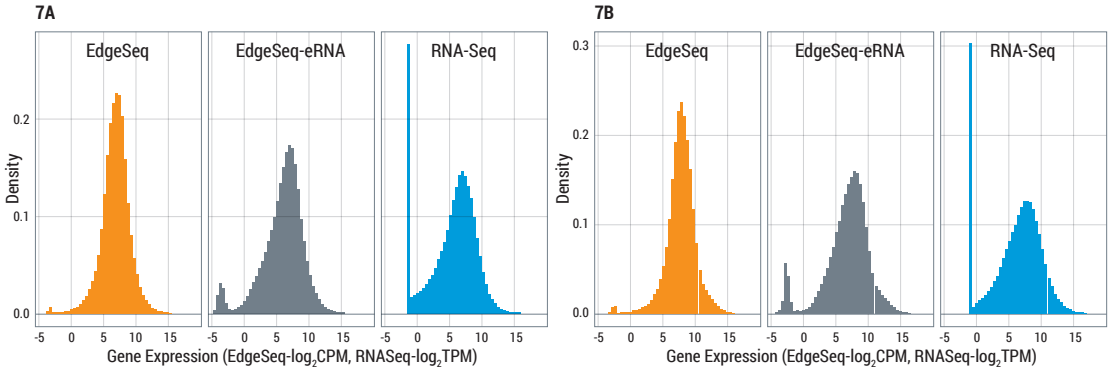


***Figure 7**. Overall distributions show zero counts increase in HTG EdgeSeq technology using extracted RNA. The x-axis shows gene expression in log_2TMP for RNA-Seq and log_2CPM for HTG EdgeSeq and the y-axis shows gene expression Density. One hundred and eighty samples were tested on HTG EdgeSeq OBP (A) and HTG EdgeSeq PIP (B) using FFPE with no extraction (HTG EdgeSeq technology; orange) extracted RNA (HTG EdgeSeq-eRNA; gray) and compared to RNA-Seq (blue).*

RNA Extraction Bias

One of the most interesting findings in this study was the appearance of the spike at zero reads in the RNA-Seq data compared to the HTG EdgeSeq technology, and a slight low-end bias. We propose that this difference is caused by one of two factors, (1) differences in sample input between the two platforms or (2) differences in technology between the two platforms. To understand the contribution of sample input to the high number of zero read transcripts in the RNA-Seq platform, approximately 180 of the eRNA sample aliquots used to generate the RNA-Seq data were run on both HTG EdgeSeq assays. Overall distributions were plotted between HTG EdgeSeq OBP vs RNA-Seq, *(Figure 7A)*, and HTG EdgeSeq PIP vs RNA-Seq *(Figure 7B)* using standard HTG EdgeSeq extraction-free sample preparation (HTG EdgeSeq technology), and eRNA samples run on both HTG EdgeSeq assays (HTG EdgeSeq- eRNA), and RNA-Seq. These plots show that the proportion of zero counts in these samples increased from 0.79% to 5.15% for HTG EdgeSeq OBP, and from 0.98% to 6.53% for HTG EdgeSeq PIP when compared to the extraction-free lysates of the same samples. In addition, overall distributions appear to shift to the low expressors, a pattern consistent with the distribution seen in the RNA-Seq data. Extraction bias associated with RNA purification impacts the ability of either platform to detect low-expressing genes, suggesting the extraction-free HTG EdgeSeq technology may offer superior sensitivity. Together these data show that RNA extraction likely accounts for a loss in signal of low expressing genes, leading to the spike in zero expressing genes.

To further understand the effect of sample type to the overall correlation between the two platforms, sample-wise Spearman correlation coefficients were calculated for the same set of approximately 180 eRNA samples. The comparisons were made between HTG EdgeSeq PIP *(Figure 8A)* and HTG EdgeSeq OBP *(Figure 8B)* using standard extraction-free sample preparation (EdgeSeq), HTG EdgeSeq technology using extracted RNA (EdgeSeq_eRNA), and RNA-Seq using extracted RNA (RNA-Seq). The best correlations were observed for the comparison between EdgeSeq FFPE vs EdgeSeq_eRNA with the highest median for all correlations followed by RNA-Seq vs EdgeSeq_eRNA with the second-best correlations and EdgeSeq FFPE vs RNA-Seq with the lowest correlations *(Table 2)*. These data show that the HTG EdgeSeq platform and RNA-Seq platform correlate better when both platforms used extracted RNA for sample input, suggesting that some, but not all, of the difference in platform performance is due to the bias introduced by RNA extraction.

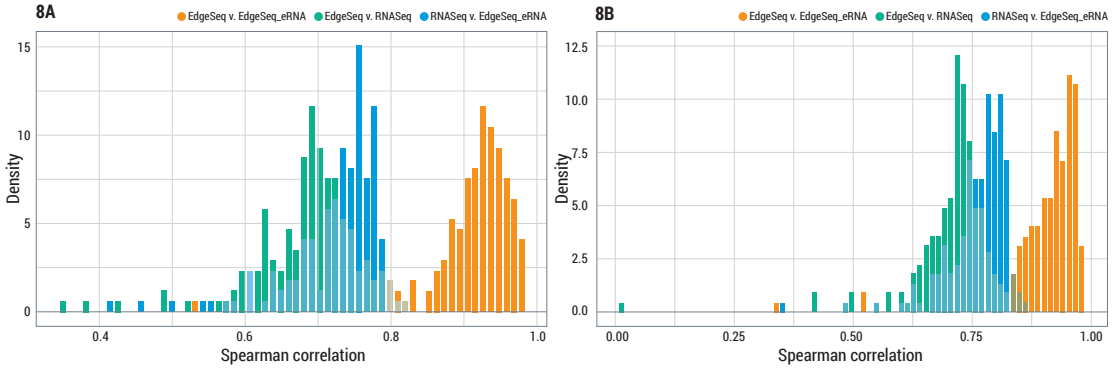


***Figure 8**: Overlapping histograms show that correlation to RNA-Seq improves (i.e., is shifted to the right) with same sample type (eRNA) is used. The x-axis shows the value of the Spearman Correlation Coefficient and the y-axis shows the Density of the correlation value comparisons were made between EdgeSeq and EdgeSeq_eRNA (orange), EdgeSeq technology and RNA-Seq (green) and EdgeSeq_eRNA and RNA-Seq (blue).*

Table 2. Summary statistics of Spearman correlation coefficients

Comparison	Figure	Panel	Mean	Standard Deviation	Median
EdgeSeq vs. EdgeSeq-eRNA	8A	OBP	0.897	0.0917	0.924
RNA-Seq vs. EdgeSeq-eRNA	8A	OBP	0.762	0.0642	0.777
EdgeSeq vs. RNA-Seq	8A	OBP	0.706	0.0823	0.720
EdgeSeq vs. EdgeSeq-eRNA	8B	PIP	0.899	0.0761	0.922
RNA-Seq vs. EdgeSeq-eRNA	8B	PIP	0.724	0.0630	0.724
EdgeSeq vs. RNA-Seq	8B	PIP	0.681	0.0687	0.693

Conclusion

This paper compared the performance of two HTG EdgeSeq assays, the HTG EdgeSeq Oncology Biomarker Panel, and HTG EdgeSeq Precision Immuno-Oncology Panel, to the Illumina TruSeq RNA exome (RNA-Seq) using approximately 1,200 FFPE tumor tissue samples. This study showed slightly increased pass rates for HTG EdgeSeq technology as compared to RNA-Seq with pass rates of 96.9% and 97.3% for HTG EdgeSeq OBP and HTG EdgeSeq PIP, respectively, and 91.7% for RNA-Seq. Samples passing QC metrics showed similar overall distributions between both HTG EdgeSeq technology and RNA-Seq, with an increase in genes with zero counts (12%) in the RNA-Seq data. All assays were able to cluster samples by indication and sample-level correlations showed 94-96% of samples had Spearman correlation coefficients between 0.6 and 0.8. The few samples with correlations below 0.6 appeared to be caused by poor sample quality and not differences in HTG EdgeSeq platform and RNA-Seq platform. Correlation plots for individual samples show good overall correlation but highlight the LoD limitations for RNA-Seq. To understand whether this was attributed to sample input or differences in platform chemistry, eRNA samples were run on the HTG EdgeSeq platform. An increase in genes with zero counts in the eRNA HTG EdgeSeq data suggest that some, but not all, of the difference in platform can be attributed to RNA extraction.

The data presented here show that the HTG EdgeSeq platform can be used as a competitive alternative to RNA-Seq, with several distinct advantages. First, HTG EdgeSeq assays require less sample input due to the extraction-free sample preparation as compared to RNA extraction required for RNA-Seq. Given the challenges associated with limited clinical specimens for biomarker identification/exploratory analysis, the HTG EdgeSeq technology has the added benefit of utilizing "less for more." Second, the HTG EdgeSeq workflow is significantly faster compared to the traditional RNA-Seq workflow with the ability to provide critical answers in less time. Third, the HTG EdgeSeq workflow provides fully integrated web-based data analysis pipeline to help researchers better understand data faster. The data presented here show the compatibility of data from the two platforms making HTG EdgeSeq technology an excellent complement to traditional RNA-Seq data.

References

1: Lu YC, Yao X, Crystal JS, et al. Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions. Clin Cancer Res. 2014;20(13):3401–3410.

2: Cimino-Mathews A, Thompson E, Taube JM, et al. PD-L1 (B7-H1) expression and the immune tumor microenvironment in primary and metastatic breast carcinomas. Hum Pathol. 2016;47(1): 52–63.

3: Concha-Benavente F, Srivastava RM, Trivedi S, et al. Identification of the cell-intrinsic and extrinsic pathways downstream of EGFR and IFNy that induce PD-L1 expression in head and neck cancer. Cancer Res. 2016;76 (5):1031–1043.

4: Yan H1, Dobbie Z, Gruber SB, et al. Small changes in expression affect predisposition to tumorigenesis. Nat Genet. 2002;30(1):25–26.

5: Yuan J, Hegde PS, Clynes R, et al. Novel technologies and emerging biomarkers for personalized cancer immunotherapy. J Immunother Cancer. 2016;4:3.

6: Sultan M, Amstislavskiy V, Risch T, Schuette M, Dokel S, Ralser M, Balzereit D, Lehrach H, Yaspo ML. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics. 2014;15:675.