Authors

Nancy Nabilsi
Senior Applications Scientist

Drew Cheney
Applications Scientist

Ida van Jaarsveld
R&D Team Leader (Bioinformatics)

Leendert Cloete
Bioinformatics Scientist

Davis Todt
Bioinformatics Scientist

Jennifer Pavlica
Applications Manager

Rachel Kasinskas
Director of Scientific Support & Applications

Roche Sequencing & Life Science
Wilmington, MA, USA,
and Cape Town, South Africa

KAPA RNA HyperPrep: A streamlined library preparation workflow that enables robust gene expression profiling using RNA-sequencing

RNA-sequencing is a powerful tool for molecular profiling, and the KAPA RNA HyperPrep Kit offers a streamlined solution for the construction of NGS libraries from RNA inputs of varying qualities. The results shown here demonstrate successful RNA-sequencing library construction from partially-degraded RNA, leading to superior detection of differential gene expression compared to alternative workflows.

Introduction

High-resolution RNA analysis using next-generation sequencing (RNA-seq) offers a comprehensive assessment of the transcriptome, allowing for quantification of global gene expression. The utility of RNA-seq in disease research has expanded, particularly in cancer research where molecular profiling of tumors has become increasingly informative. Multiple library preparation kits are commercially available, each using different strategies and chemistries for RNA enrichment and library construction. In this study, three RNA-seq workflows from Kapa Biosystems, Illumina, and New England Biolabs (NEB) are compared with regard to performance in tumor profiling. To evaluate the effects of workflow differences, RNA was extracted from donor-matched normal and tumor breast tissue and used as input into each RNA-seq library construction kit. The resulting libraries were compared using key library construction and sequencing metrics. Differential gene expression analysis was performed and results were verified using an independent quantitative RT-PCR (qRT-PCR) approach.

Materials and methods

Breast tissue samples

Patient-matched, fresh-frozen primary breast tumor and adjacent normal breast tissue were obtained from AMSBIO. Technical documentation stated that the tumor is a Grade 2, Stage IIb infiltrating lobular carcinoma. TNM staging (T3, N0, M0) indicated a large tumor absent of detectable lymph node involvement and distant metastasis. Following extraction and DNase treatment, total RNA was quantified using the Qubit® RNA HS Assay (ThermoFisher). RNA quality was assessed using a 2100 Bioanalyzer instrument and an Agilent® 6000 RNA Pico Kit (Agilent Technologies) (Figure 1).

Two metrics frequently used to evaluate RNA quality are the RNA Integrity Number (RIN) and the DV200 value; both values are based on the size distribution of the RNA, which is evaluated using an electrophoretic trace. The RIN score is automatically tabulated by the Agilent Expert software, and uses the ratio of intact ribosomal peaks and the presence of intermediate peaks to assign the integrity number. However, RNA extracted from archived biospecimens—such as those used in this study—typically lack distinctive ribosomal peaks, reducing the utility of the RIN score. In contrast, the DV200 value does not depend on the presence of ribosomal peaks, and is calculated as the percent of RNA molecules greater than 200 nucleotides (nt) in length. Because RNA fragments shorter than 200 nt are poor substrates for RNA-seq library construction and are unlikely to contribute to the final library, the DV200 is a more appropriate method for assessing the quality of degraded RNA samples.

Library construction

Libraries were prepared using 100 ng of input RNA. Unless indicated otherwise, library construction was performed following the manufacturer's recommendations with reagents supplied in the respective library preparation kits. The kits employed in this study and key library construction parameters are summarized in Table 1.

RNA extractions typically contain large amounts (up to 90%) of ribosomal RNA (rRNA), which is not of biological interest to most investigators. Removal of rRNA prior to RNA-seq library construction increases the economy of sequencing and improves coverage of lower-abundance transcripts. Thus, the first step in all three workflows used in this study is rRNA depletion. Under conditions that drive hybridization, the total RNA sample is incubated with probes that are complementary to rRNA sequences. Depletion of rRNA is then achieved either by the addition of RNase H to enzymatically degrade the hybridized rRNA, or by the addition of paramagnetic beads that bind to the probe/rRNA complexes and remove them from the sample when a magnetic field is applied. The rRNA depletion strategy for each workflow is listed in Table 1.

Following the final post-amplification cleanup step, library yields were quantified with the qPCR-based KAPA Library Quantification Kit for Illumina® platforms. Library size distributions were confirmed with a 2100 Bioanalyzer instrument and Agilent® High Sensitivity DNA Kit.

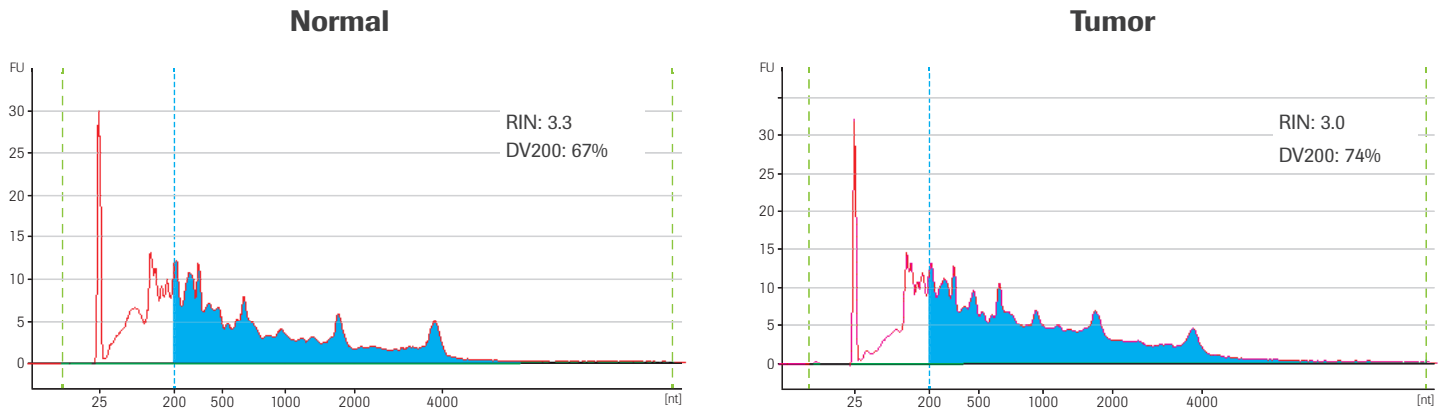

Figure 1: Input sample quality. Electrophereograms of input RNA were generated using an Agilent RNA 6000 Pico Kit. The RNA Integrity Number (RIN) and amount of material ≥200 nt (DV200) are indicated. Blue shading highlights RNA fragments ≥200 nt.

Table 1: Library construction workflow and data overview

	KAPA RNA HyperPrep Kit with RiboErase (HMR)		TruSeq® Stranded Total RNA with Ribo-Zero Gold		NEBNext Ultra Directional RNA Library Prep Kit with rRNA Depletion	
Sample RNA	Normal	Tumor	Normal	Tumor	Normal	Tumor
Input quantity (ng)	100		100		100	
Depletion method	RNase H		Paramagnetic beads		RNase H	
Fragmentation	94°C for 4 min		94°C for 4 min		94°C for 15 min	
PCR cycles	13		15		15	
Post-PCR yield (nM)	14.3 (±2)	20.9 (±2.6)	60.6 (±4.4)	82 (±20.1)	12.5 (±1.7)	20 (±2.1)
Mean library size (bp)	364 (±5)	369 (±4)	311 (±7)	380 (±34)	334 (±4)	333 (±2)
Adapter-dimer (%)	2.2 (±0.4)	1.6 (±0.4)	<1	<1	<1	<1

Sequencing and data processing

Uniquely indexed duplicate libraries from each sample and workflow were normalized and pooled for 2 x 100 bp paired-end sequencing on a HiSeq® 2500, using v4 chemistry (Illumina).

Adapter and quality trimming was performed using cutadapt and trimmomatic, respectively. Reads were aligned to a hard masked version of human reference GRCh38, filtered to remove rRNA reads, and sub-sampled to the lowest common number of paired reads (14M). Gene expression was normalized and quantified using Kallisto (0.42.4). Differential gene expression analysis was performed using EdgeR. Overlap analysis was performed using VENNY2.1[1].

Validation of differential gene expression by qPCR

Validation of differential gene expression results was performed using a custom-designed RealTime ready qPCR array (Roche). The custom design includes: 3 reference genes for normalization; 5 reverse transcription controls (positive and negative); and 88 assays for transcripts identified as differentially expressed by RNA-seq. For each sample, cDNA was generated using the Transcriptor First Strand cDNA Synthesis Kit (Roche) using both random hexamers and anchored oligo-dT primers, starting with 600 ng of total RNA and following manufacturer recommendations. Following heat-inactivation, 200 ng of cDNA was combined with qPCR reagents (LightCycler® 480 Probes Master), aliquotted into the custom RealTime ready assay plate, and amplified according to the manufacturer's recommendation on the LightCycler 480 System. Duplicate plates were assayed per sample, and data was concatenated prior to analysis. Relative differential gene expression was quantified using the delta delta Cp method.

Results and discussion

Comparative library metrics

Library construction metrics: All three workflows successfully generated sufficient material for library QC, sequencing, and archiving. Similar yields were obtained for Kapa compared to NEB libraries, despite the difference in number of amplification cycles (13 and 15 cycles, respectively). Assuming similar amplification efficiencies between the two workflows, this is expected to reflect a greater diversity for Kapa libraries and translate to lower duplication rates. Illumina libraries exhibited higher yields than Kapa, consistent with using two additional cycles of amplification. Based on additional library construction QC data (data not shown), it is anticipated that higher yields from Illumina libraries may also suggest less efficient depletion of rRNA. Electrophoretic assessment of final libraries indicated that all three workflows exhibited minimal adapter-dimer formation (<3%) and produced libraries of similar sizes, although Kapa and NEB libraries were more consistent between sample types than Illumina libraries (Table 1).

Sequencing metrics: Libraries produced with the Kapa, Illumina, and NEB workflows were compared with respect to five key sequencing metrics: percent mapped reads; percent rRNA reads; uniformity of coverage; percent duplicate reads; and number of unique transcripts identified. For both samples, percent mapped reads were higher than 90% for all three workflows (Figure 2A). The Kapa and NEB workflows, which both employ enzymatic strategies for rRNA depletion, showed effective depletion of rRNA (≤2% rRNA reads), whereas the Illumina bead-based strategy resulted in up to 22% rRNA reads (Figure 2B). Better coverage uniformity, reflected by lower mean coefficient of variation (CV), was observed for Kapa libraries compared to the alternate workflows, especially for the tumor sample (Figure 2C). Kapa and Illumina workflows both outperformed NEB in regard to percent duplicate reads (Figure 2D).

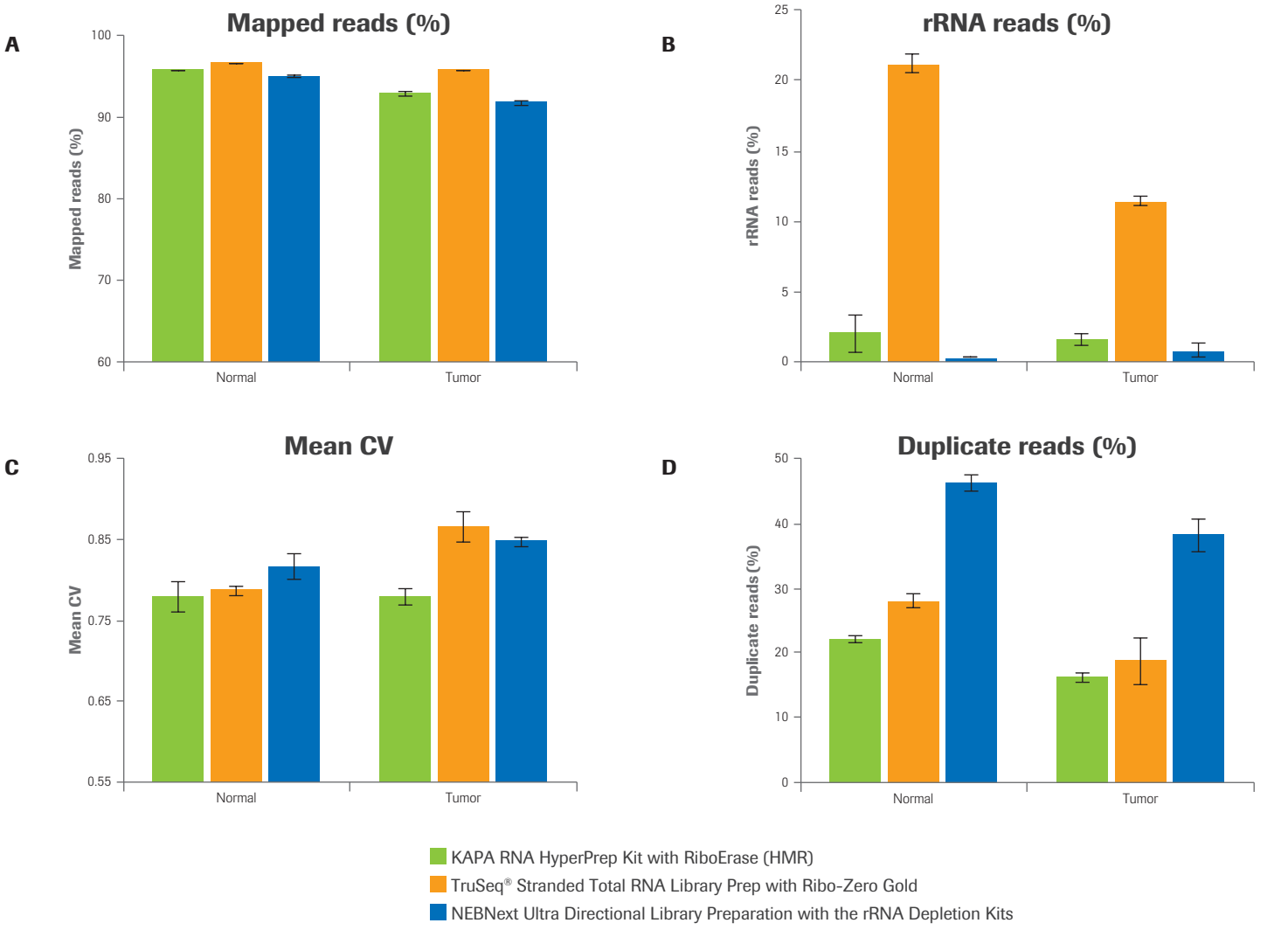

Figure 2. Workflow effects on sequencing metrics. Data was analyzed to quantify **(A)** percent mapped reads, **(B)** percent rRNA reads, **(C)** coverage uniformity (mean coefficient of variation), and **(D)** percent duplicates. Bars represent the mean from duplicate libraries and error bars indicate 0.5 of the range.

Transcript identification and differential expression: The Kapa workflow identified a greater number of unique transcripts in both tumor and normal samples than either the Illumina or NEB workflows (Figure 3A). This was expected to have implications for differential gene expression analysis. Data analyzed from Kapa libraries identified up to 50% more differentially expressed transcripts than from libraries generated with the alternative kits (Figure 3B) when the minimum fold-change was set to ≥2 at a p-value <0.05. When the statistical stringency was increased (p-value ≤0.001), the difference was even more striking; Kapa identified 5-fold more differentially expressed transcripts than alternative workflows (Figure 3B – C).

Further analysis revealed that the Kapa workflow identified the majority of differentially expressed transcripts identified by Illumina and/or NEB (69%), plus an additional 299 transcripts not found with either NEB or Illumina (Figure 3C). Notably, fewer than 5% of the transcripts identified by Illumina and/or NEB but not by Kapa are protein-coding (2 out of 41). In contrast, 58% of the transcripts identified only through the Kapa workflow are protein-coding (174 out of 299), and many of these are known to be dysregulated in breast cancer[2,3].

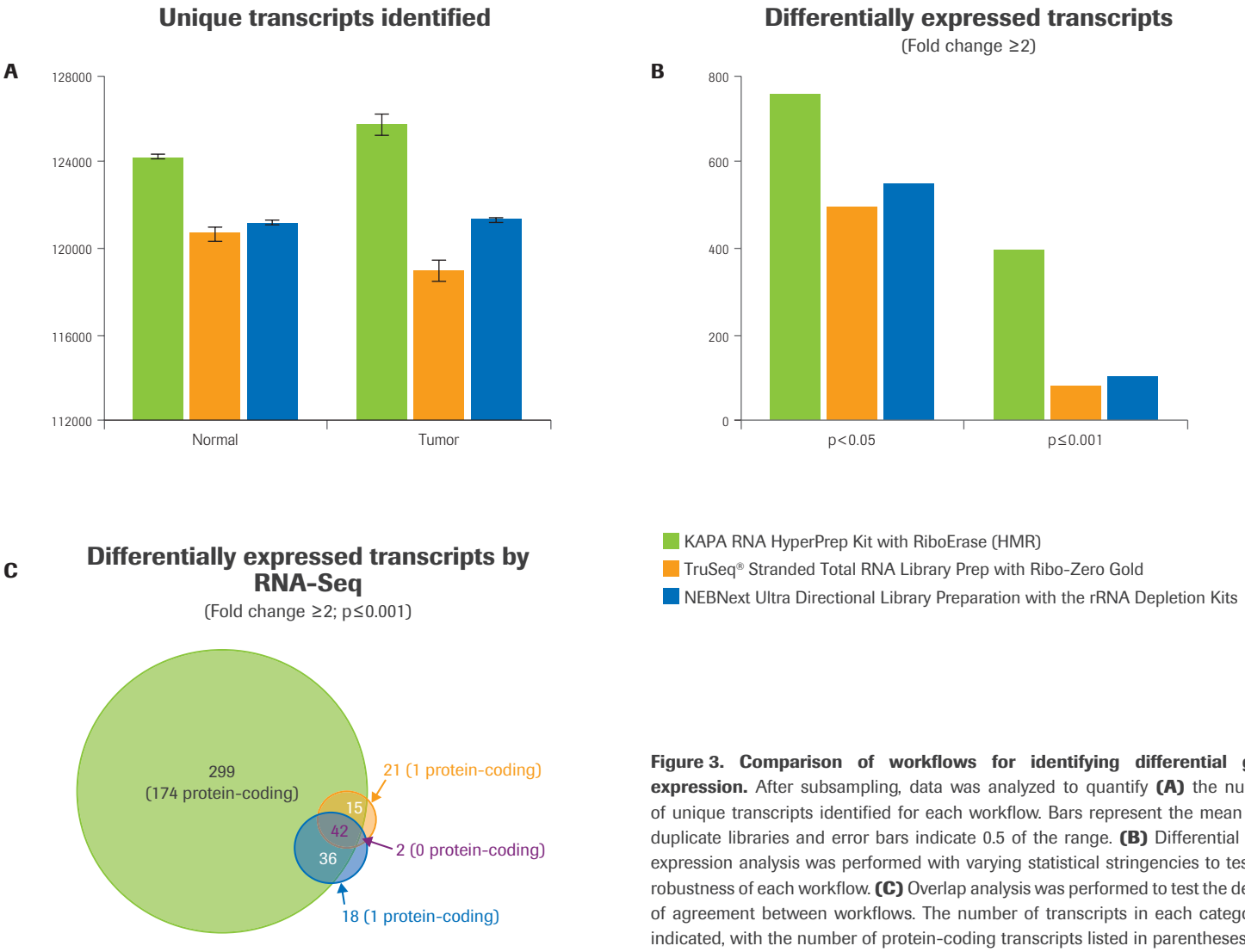

Figure 3. Comparison of workflows for identifying differential gene expression. After subsampling, data was analyzed to quantify **(A)** the number of unique transcripts identified for each workflow. Bars represent the mean from duplicate libraries and error bars indicate 0.5 of the range. **(B)** Differential gene expression analysis was performed with varying statistical stringencies to test the robustness of each workflow. **(C)** Overlap analysis was performed to test the degree of agreement between workflows. The number of transcripts in each category is indicated, with the number of protein-coding transcripts listed in parentheses.

Verification of differential gene expression: Pre-designed qPCR assays were used for hydrolysis probe-based detection of differentially expressed transcripts (custom panel design, Figure 4). The custom panel includes: 3 reference genes for normalization; 5 reverse transcription controls (positive and negative); and 88 assays selected based on RNA-seq data. Targets include 10 assays targeting transcripts identified as differentially expressed by multiple workflows (Kapa, Illumina, and/or NEB), and 78 assays targeting transcripts identified as differentially expressed by KAPA RNA HyperPrep only. Of the 88 assays, 81 (92%) generated high-confidence Cp data for both replicates of the tumor and normal samples. These 81 transcripts, which included 71 "Kapa-only" targets, were further analyzed for differential expression. All 10 of the transcripts identified as differentially expressed by multiple RNA-seq workflows exhibited ≥2-fold change by qPCR in the same direction; 7 were down-regulated, and 3 were up-regulated (Figure 5). Of the 71 targets identified only with the Kapa workflow, 62 transcripts (87%) exhibited ≥2-fold change in the expected direction by qPCR. An additional 7 targets trended in the same direction as the RNA-seq data, but did not meet the 2-fold minimum change by qPCR. Only 2 targets (3%) showed no numerical change between normal and tumor samples. Overall, these qPCR results offer independent verification that the differentially expressed transcripts identified only by the Kapa workflow reflect measurable changes in gene expression, and are not RNA-seq artifacts.

It is worth noting that four of the transcripts identified as differentially expressed only by the Kapa workflow are part of the PAM50 classifier, a clinical tool used to assign breast cancers to one of four intrinsic subtypes and often used as a prognostic indicator. The differentially expressed PAM50 transcripts were FOXA1, MLPH, MYC and SLC39A6[2]. Additional "Kapa-only" transcripts are part of the SAM264 gene classifier set that represents genes associated with breast cancer patient survival. These genes include ALCAM, AQP3, CDH1, CRABP2, IGFBP5, KRT18, KRT7, KRT8, MUC1, S100A1, and S100P[3]. Thus, the Kapa workflow accurately detects the differential expression of genes that are relevant to disease research.

LAMB2	CDH1	TACSTD2	S100A1	S100A1 (2)	CD63	CALM1	PGM1	PBX1	ACTB	ETFB	ERBB3
CD164	RASD1	FOS	SLC39A6	WFDC2	CCL21	CLDN4	CD63	EGR1	TIMP2	ALCAM	HIST1H1B
RHOA	MT2A	TSC22D3	FBN1	ORM1	KRT18	COL15A1	ADAMTS1	S100P	HIST3H2BB	ATP2B4	KRT8
ID2	KRT7	MYC	CLDN3	GUSB	HSPG2	B2M	AHNAK	GPC3	EFEMP1	NDRG1	FSTL1
VWF	EPAS1	HIST2H2BE	MMD	SERPINF1	MGST3	AC137834	LAMC1	EMP1	CAT	EHD2	SLP1
TIMP3	FOXA1	PNPLA2	FTL	ITGB1	PTRF	HBA2	CRABP2	HSPB1	SOCS3	AGPAT2	TKT
AQP3	RARRES2	AZGP1	COL4A1	KRT19	MUC1	GSN	CD36	XBP1	G0S2	RBP4	PON3
CFD	LPL	FABP4	PIP	GAPDH	TBP	HPRT1	RT+	RT+	RT+	RT-	RT-

≤1 x 10^{-7} 0.001

Figure 4. RealTime ready custom panel design. This panel was designed to test 88 up- and down-regulated transcripts across a range of fold changes (≥2) and p-values (0.001 to 9.02E^{-39}) identified by RNA-seq. The coloring reflects the p-value for each target measured by RNA-seq. The 10 targets highlighted in white are differentially expressed transcripts that were identified by multiple workflows. The remaining 78 targets were identified by Kapa only. Housekeeping genes and controls for the reverse transcription reaction are in grey.

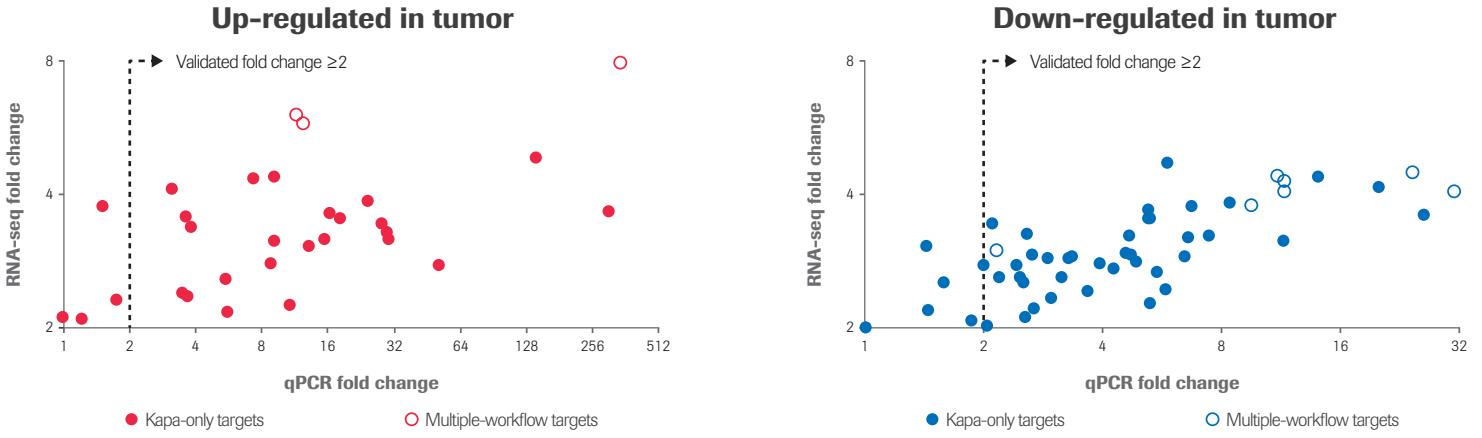

Figure 5. qPCR validation of differential gene expression. RealTime ready custom qPCR arrays were used to validate gene expression profiles obtained by RNA-seq. Scatter plots compare fold change values obtained by RNA-seq to values measured by qPCR. Each circle represents the average of two replicate measurements for a unique transcript. Filled circles indicate transcripts that were identified as differentially expressed by only the Kapa workflow; open circles indicate transcripts identified as differentially expressed by multiple workflows.

Conclusions

The KAPA RNA HyperPrep Kit with RiboErase (HMR) is a streamlined, robust workflow for the construction of RNA-seq libraries using partially degraded RNA samples. Library QC metrics and sequence data indicated successful and reproducible library construction with higher complexity, more efficient rRNA depletion, and more balanced base coverage than alternative workflows. Additionally, the Kapa workflow identified a greater number of differentially expressed transcripts. Verification of these findings by qPCR confirms that KAPA RNA HyperPrep Kit with RiboErase (HMR) is a powerful tool for molecular profiling.

References

1. Oliveros, J.C. (2007 – 2015) Venny. An interactive tool for comparing lists with Venn's diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html)
2. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al: Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009, 27 (8): 1160-1167. 10.1200/JCO.2008.18.1370.
3. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van den Rijn M, Jeffrey SS, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA.* 2001;98:10869–10874.

For more information about Roche RNA-seq products and solutions, please visit: **sequencing.roche.com/RNA-seq**

Published by:

Roche Sequencing and Life Science
9115 Hague Road
Indianapolis, IN 46256

sequencing.roche.com

MC-US-07780 APP111002 A520 4/20