Application Note

RNA-Seq: Gene Expression Analysis

Authors

Nancy Nabilsi
Senior Applications Scientist

Ranjit Kumar
Senior Bioinformatics Scientist

Jennifer Pavlica
Applications Manager

Roche Sequencing & Life Science
Wilmington, MA, USA

Samantha Dockrall
Senior Product Support Specialist

Davis Todt
Bioinformatics Scientist

Roche Sequencing Solutions
Cape Town, South Africa

Heather Whitehorn
International Product Manager

Maryke Appel
Sr International Product Manager

Roche Sequencing Solutions
Pleasanton, CA, USA

Janez Kokošar
Senior Bioinformatician

Yolanda Darlington
Senior Product Manager

Luka Ausec
Director of Customer Success

Moses M. Feaster
Director of Commercial Strategy

Genialis, Inc.
Houston, TX, USA

Date of first publication:
October 2018

KAPA RNA HyperPrep Kits and the Genialis™ NGS data analytics platform: a qualified, streamlined RNA-Seq solution for gene expression analysis

RNA-Seq is a powerful tool for gene expression analysis. To ensure high-quality, reliable results, robust library preparation chemistry as well as qualified data analysis pipelines are needed. KAPA RNA HyperPrep Kits and the Genialis platform offer simple and complete workflow solutions for NGS-based gene expression analysis, leaving you more time to focus on biological questions.

Introduction

Next-generation sequencing (NGS) of RNA, also known as RNA-Seq, enables high-resolution and comprehensive assessment of the transcriptome, thereby allowing for the quantification of global gene expression. The utility of RNA-Seq has expanded into many areas of research, including tumor biology.[1]

Each stage of the RNA-Seq workflow has the potential to reduce or bias the intrinsic value of the biological information contained in precious NGS samples. To ensure high-quality, reliable RNA-Seq results, it is important to use efficient and robust library construction chemistry and qualified data analysis pipelines. Not all library preparation kits for RNA-Seq are equally effective in terms of RNA enrichment, cDNA synthesis, conversion of cDNA to adapter-ligated library fragments and library amplification. A plethora of data analysis algorithms and tools are available, but selecting the appropriate pipeline components and parameters often requires advanced bioinformatics expertise.

KAPA RNA HyperPrep Kits offer streamlined and flexible, stranded library preparation solutions for different RNA sample types, input amounts and sequencing applications. Efficient RNA enrichment, high conversion rates, and amplification with the low-bias KAPA HiFi enzyme typically leads to higher library complexity, fewer reads wasted on unwanted transcripts and PCR duplicates, and better coverage of low-abundance and GC-rich transcripts, as compared to reagents from other suppliers.[2]

The Genialis platform is a cloud-based suite of multi-omics computing software applications that simplify the analysis, visualization and management of NGS data. Designed with biologists in mind, the platform offers guided, visual RNA-Seq gene expression analysis and interpretation workflows, backed with automated data processing pipelines developed and qualified specifically for the KAPA RNA HyperPrep Kit portfolio.

In this study, we combined the KAPA RNA HyperPrep Kit with RiboErase (HMR) with the Genialis platform to analyze differential gene expression in a pair of matched normal and tumor breast tissue samples. Single-click tools made it simple to visualize and interrogate data, and discover differences between the KAPA chemistry and that of another supplier. To complete the workflow, the expression patterns for selected genes were confirmed by real-time PCR.

Library construction and sequencing

Samples

Donor-matched, fresh-frozen primary breast tumor and adjacent normal breast tissue were obtained from AMS Biotechnology. Technical documentation indicated the tumor to be a Grade 2, Stage IIb infiltrating globular carcinoma. TNM staging[3] (T3, N0, M0) indicated a large tumor devoid of detectable lymph node involvement and distant metastasis.

Total RNA was extracted from fresh frozen tissue and treated with DNase I, using an RNeasy® Plus Universal Mini Kit (QIAGEN®). RNA was quantified using the Qubit® RNA HS Assay (ThermoFisher). RNA quality was assessed using an Agilent® 2100 Bioanalyzer instrument and Agilent RNA 6000 Pico Kit (Agilent Technologies). Both quality metrics provided by this assay, namely the RNA Integrity Number (RIN) and DV_{200} value (% of RNA fragments with a length ≥200 nt), indicated that both RNA preparations were of medium and comparable quality (Figure 1).

RNA depletion and library construction workflow

Total RNA contains up to 90% ribosomal RNA (rRNA)[4], which is not of biological interest in most investigations. For this reason, RNA samples are typically enriched for transcripts of interest prior to library construction; to improve the coverage of lower-abundance transcripts, as well as sequencing economy. mRNA selection (with oligo-dT beads) is commonly used in gene expression analysis experiments, but results in a bias toward the 3'-portions of transcripts if input RNA is not of a high quality. Because the RNA extracted from both the tumor and normal tissues was slightly degraded, an rRNA depletion approach was selected for this study.

Duplicate libraries were prepared from 100 ng inputs of both RNA extracts, using either the KAPA RNA HyperPrep Kit with RiboErase (HMR), or the TruSeq® Stranded Total RNA with Ribo-Zero Gold kit (Illumina®). These kits employ similar overall strategies for library construction, but differ in many respects. Key similarities and differences are summarized in Table 1.

Table 1. Library construction kit comparison

Feature	KAPA RNA HyperPrep Kit with RiboErase (HMR)	TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold
Species compatibility	Human, mouse and rat	
rRNA species depleted	Cytoplasmic, mitochondrial	
Depletion technology	RNase H	Paramagnetic beads
RNA fragmentation	94°C for 4 min	
1st strand priming	Random hexamers	
Reverse transcriptase	KAPA Script	SuperScript™ II (not included)
Stranded library prep	Yes	
Cleanup beads	KAPA Pure Beads (included)	Agencourt® AMPure® XP Reagent (not included)
Library amplification enzyme	KAPA HiFi HotStart ReadyMix	TruSeq PCR Master Mix
Number of amplification cycles	13	15
Total workflow time	6.5 hours	7 hours

Refer to product documentation[5,6] for full protocol and reagent details.

The full KAPA RNA HyperPrep with RiboErase (HMR) workflow, from input RNA to sequencing-ready library, is depicted in Figure 2 on the next page.

Library QC and sequencing

After 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 an Agilent 2100 Bioanalyzer instrument and Agilent High Sensitivity DNA Kit. Results are given in Table 2 on the next page. The TruSeq workflow produced higher library yields, as a result of the two extra amplification cycles, and a higher level of residual rRNA.[7]

The eight libraries, which were prepared with single-indexed adapters, each with a different sequencing barcode, were normalized and pooled for 2 x 100 bp paired-end sequencing on an Illumina® HiSeq® 2500 instrument, using v4 chemistry.

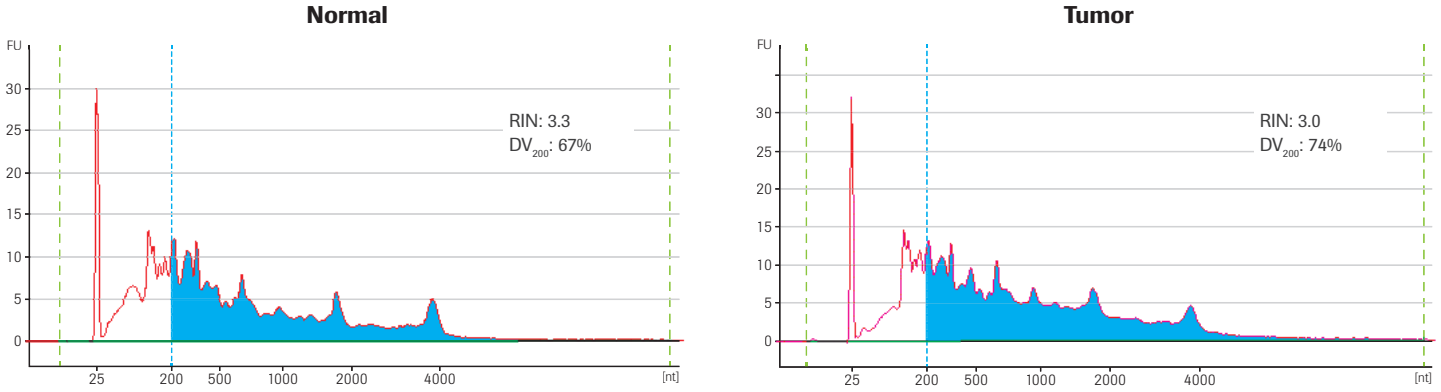

Figure 1: Quality assessment of total RNA extracts. Electropherograms of DNAse I-treated RNA extracts were generated using an Agilent RNA 6000 Pico Kit. The RNA Integrity Number (RIN) and DV_{200} value are given in the top right hand corner of each graph. Unlike the RIN, the DV_{200} value does not depend on the presence of distinct rRNA peaks, which are typically absent in RNA extracts from archived biological specimens such as these. Blue shading highlights RNA fragments ≥200 nt in length, which are suitable substrates for library construction with the kits used in this study.

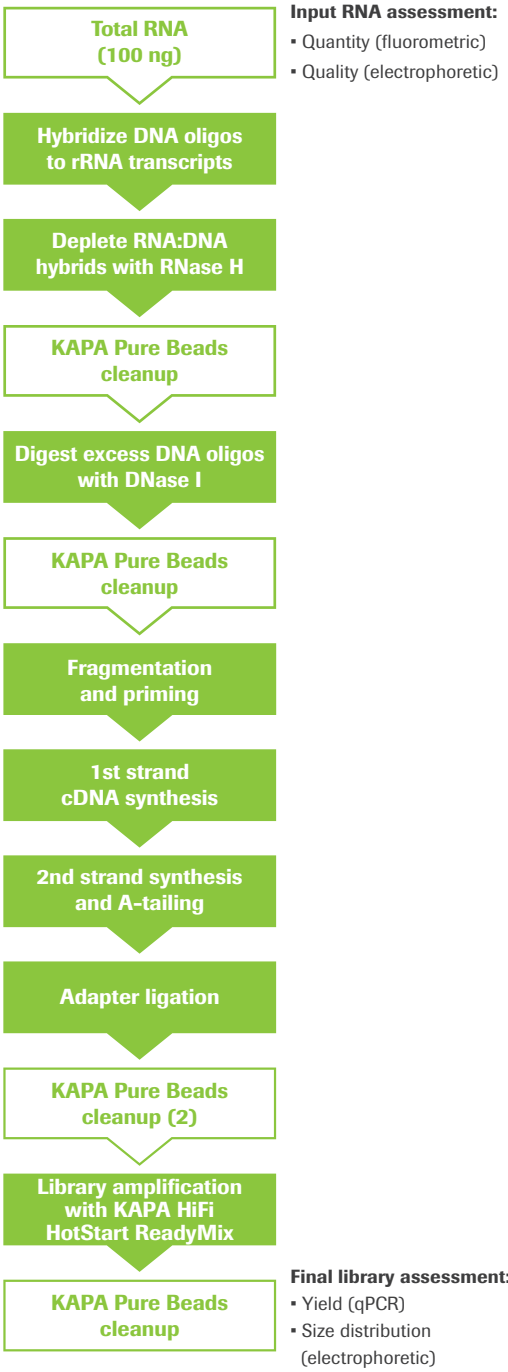

Figure 2: Overview of the KAPA RNA HyperPrep with RiboErase (HMR) RNA enrichment and library construction workflow. QC assessment was performed on input material and final libraries using the specified methods. Additional, optional QC assays that may be utilized (especially when the workflow is evaluated for the first time, or when working with degraded samples), are described in a Technical Note.[8] KAPA RNA HyperPrep Kits with RiboErase (HMR) contain all of the reagents required for rRNA depletion, cDNA synthesis and library construction (including KAPA Pure Beads for reaction cleanups), with the exception of adapters, which are available separately. The entire protocol is automation-friendly.

Table 2. Library QC metrics

Metric	KAPA		TruSeq	
	Tumor	Normal	Tumor	Normal
Final library concentration (nM)	20.9 ± 2.6	14.3 ± 2.0	82.0 ± 20.1	60.6 ± 4.4
Mean library size (bp)	369 ± 4	364 ± 5	380 ± 34	311 ± 7
Residual rRNA (%)	1.5 ± 0.39	2.0 ± 1.3	11.4 ± 0.33	21.2 ± 0.67

Data management and analysis

Gene expression application

The Genialis™ platform for gene expression analysis is hosted in the cloud, and is ubiquitously accessible via an internet browser (https://app.genialis.com/roche). After signing in, the user is greeted with a landing page which provides easy access to profile and account settings; recent data sets and data highlights; a demo video and demo data; and a quick tour of the data management and analysis workflow (Figure 3).

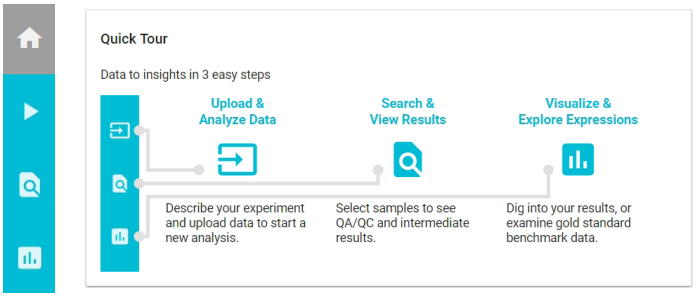

Figure 3: Genialis homepage icon bar (left) and gene expression analysis workflow (right). The application consists of four modules, represented by the four icons (from top to bottom): Home (the user dashboard), followed by *Analyze*, *Search and View Results*, and *Visualizations* (the three stages of the data management and analysis workflow). The *Analyze* module has three tabs, namely *Import Data*, *Quality Control* and *Define Experiment*. The *Search and View Results* module allows the user to easily find specific samples, and access sample history and metadata (annotations). The *Visualizations* module has five tabs: *Sample Comparison*, *Gene Expression*, *Differential Expressions*, *Venn View*, and *Heat Map*. Each of these enables the user to visualize data in different contexts; to answer different questions about the outcome of an experiment, and differences between sample types or experimental treatments. The *Visualizations* module is extremely dynamic. Plots and tables update in real-time as the user decides to include or exclude samples and/or genes of interest when interrogating the data.

Stage 1: Analyze

Raw data (compressed FASTQ files; between 17.0 and 20.4 million read pairs per sample) were imported from a local drive to Genialis using the simple drag-and-drop option on the platform's graphical interface. Alternative data upload options include:

- importing directly from BaseSpace;
- using ReSDK, an open-source, Python-based application programming interface developed by Genialis (https://resdk.readthedocs.io/en/stable/);
- file transfers via FTP or the Gene Expression Omnibus (GEO) database; or
- having data onboarded as part of Genialis' customer support service.

Once the data import was completed, autogenerated **sample names** were edited, and the appropriate raw data files were associated with each sample (sequenced library). This is an important step in the process, as the sample is the basic operational unit in the platform. All subsequent interactions with the data takes place via sample names, which are also associated with the full processing history for each sequenced library, intermediate results files and metadata.

In the *Quality Control* tab, a FastQC report is generated for every file of sequencing reads. This enables the user to review basic

sequencing metrics (e.g., number of reads, read quality and percent duplicates). If a library was sequenced in more than one lane, and is associated with multiple (pairs of) raw read files, the system automatically concatenates files before proceeding with downstream analyses.

In the final step of the *Analyze* stage, basic experimental parameters were defined in the *Define Experiment* tab. In this step, samples are arranged into **collections** for downstream analysis purposes. The four libraries (samples) generated with the KAPA workflow (KAPA_Tumor_A, KAPA_Tumor_B, KAPA_Normal_A and KAPA_Normal_B) were grouped into a collection called "KAPA RNA HyperPrep-RiboErase_Breast-Tumor-Normal". The source organism (human) and the library preparation kit (KAPA RNA HyperPrep Kit with RiboErase, HMR) were selected from drop-down lists. None of the advanced options (specification of custom adapter sequences or adapter trimming parameters) were required for this study. Once this information was completed, prompts were followed to initiate the automated data analysis pipeline, which was co-developed and qualified by Genialis and Roche bioinformatics teams (see **Appendix** for details).

The process was repeated to create a collection ("TruSeq-RiboZero Gold_Breast-Tumor-Normal") for the four samples generated with the TruSeq workflow (TS_Tumor_A, TS_Tumor_B, TS_Normal_A and TS_Normal_B). All eight samples could have been placed in the same collection and processed together.

Stage 2: Search and View Results

While data were being processed in the background, the processing status of individual samples was monitored on the *Search and View Results* page. Samples were easily found by performing a search using whole words from the collection or sample names. An automated email was received as soon as the basic analysis for each collection was completed. At this stage, a MultiQC report became available for each sample (Figure 4). This report combines statistics from FastQC (raw reads and processed reads), STAR (mapping statistics from BAM) and featureCounts (expression and quantification stats). Reports can be viewed directly or downloaded for later use.

The *Search and View Results* module is primarily designed to provide information about samples and collections. Clicking on any sample name opens a *Sample Details* page, which details all processing steps of the analysis pipeline, including parameters and tool versions. It also provides the means for viewing and editing metadata, which can be done at any time. Key metadata fields were completed to facilitate future searches.

The *Search and View Results* page also provides a segue into the *Visualizations* module. The eight samples were placed into the **sample basket** to proceed with the visualizations. In this fairly simple study, the sample basket contained all of the samples in the two collections defined for basic analysis. In more complex experiments, subsets of samples from the full complement of sequenced libraries may be selected for visualization.

Stage 3: Visualizations

The *Visualizations* module consists of five tabs with different visualization tools, each accessible with a single click. Drop-down lists and sliders allow users to toggle between different analysis parameters and/or output options. All plots generated in this module can be exported as publication-quality images.

Data visualizations are defined by the contents of the **sample basket** (see above) and **genes basket** (genes of interest), which follow the user through the different visualization tabs. Basket contents are visible (by toggling between icons at the top of the

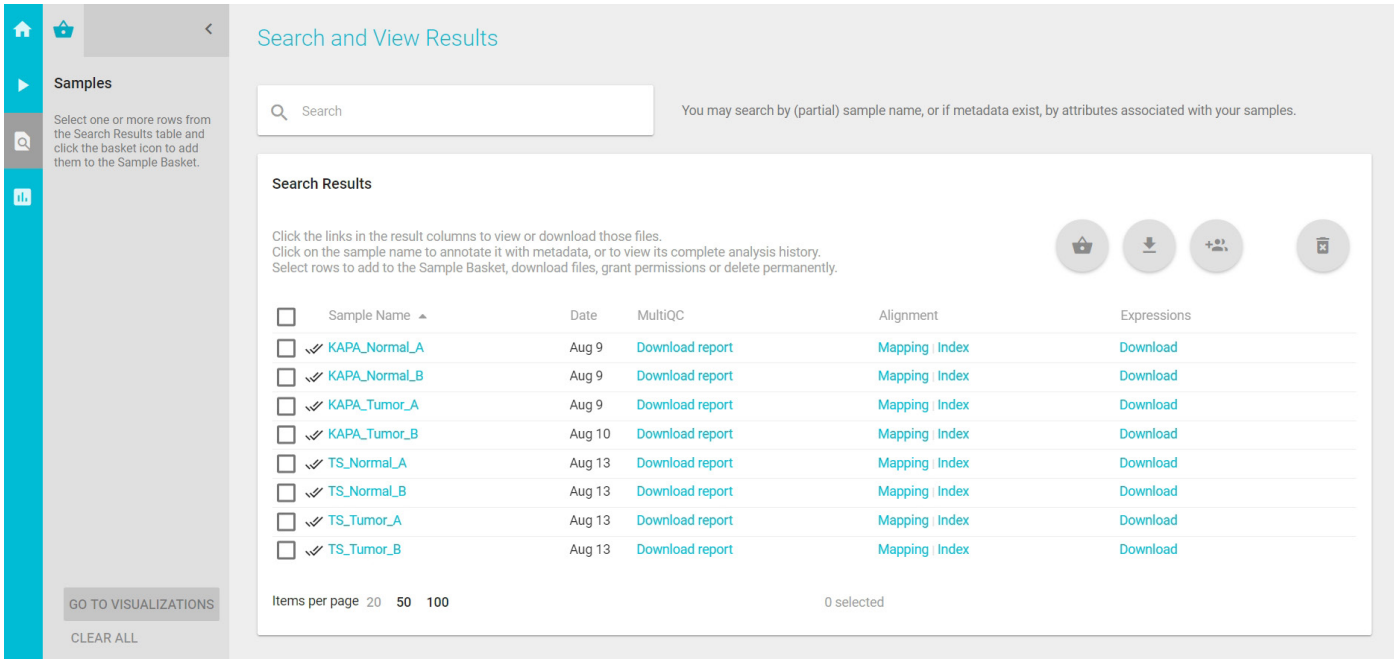

Figure 4: *Search and View Results* page view, after completion of the basic analysis and generation of MultiQC reports for the eight samples generated in this study. Buttons above the sample table allow users to select samples for visualizations, download results associated with samples, and manage permissions (for data sharing). Sample names can be also be clicked to access the *Sample Details* page, which contains sample annotations (metadata) and a detailed description of all analysis steps.

page), and may be updated at any time. Plots are automatically updated (in real time) if samples and/or genes are added or deleted, and the user may move back-and-forth between visualization tabs at will. This offers users full freedom to interrogate data in iterative cycles, without the assistance of a bioinformatics expert.

1. Sample Comparison:

The *Sample Comparison* tab provides a final layer of data QC, this time in the context of experimental design. Two tools, namely *Sample Hierarchical Clustering* and the *Principal Component Analysis (PCA) Plot*, are used to assess the consistency of results obtained from technical replicates, and whether different biological or experimental conditions yielded distinguishable results. This allows users to identify gross failure in experimental design and/or execution, and identify outliers for exclusion prior to further data exploration and interpretation.

As expected, the whole-transcriptome dendogram (Figure 5, left) revealed four distinct clusters, representing the KAPA and TruSeq Tumor and Normal samples, respectively. The two sets of normal samples clustered tightly in the *PCA Plot* (Figure 5, right), whereas the PCA suggested that a higher degree of molecular heterogeneity existed between the replicate tumor libraries generated with both of the library construction kits.

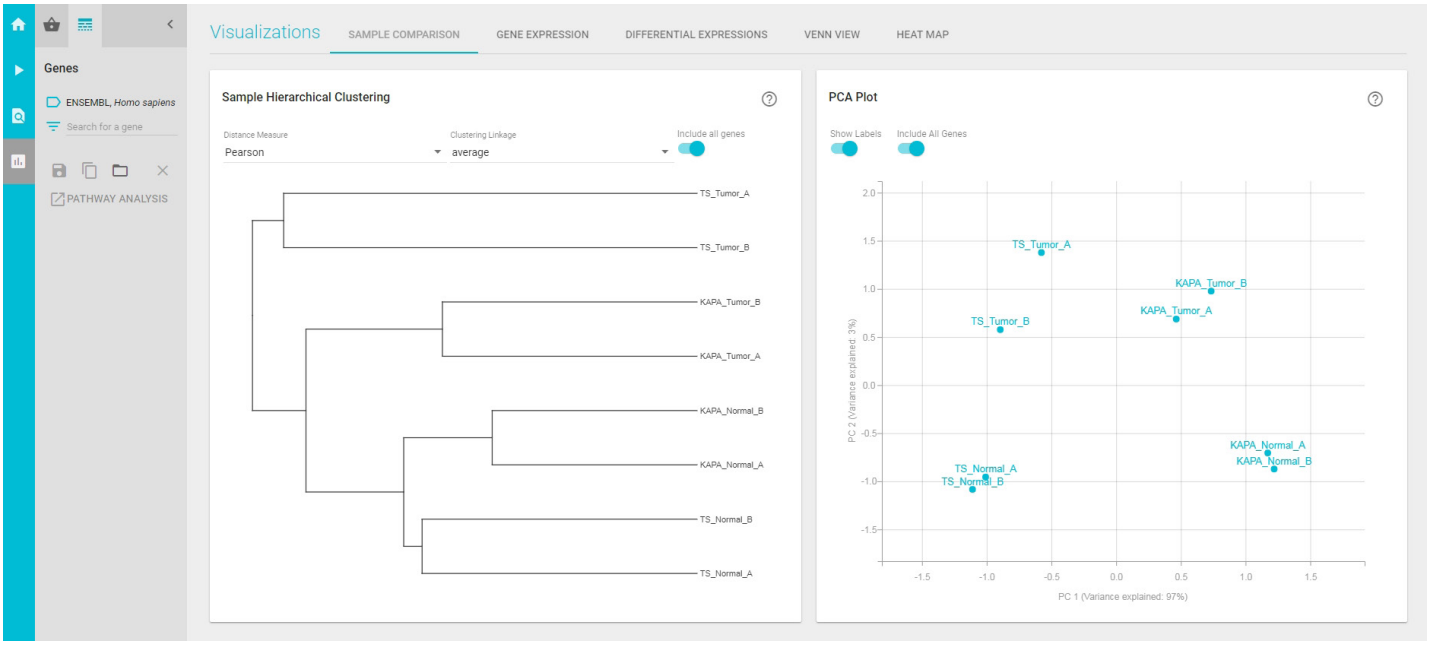

Figure 5: The *Sample Comparison* page displays the outputs of Sample Hierarchical Clustering (left) and Principal Component Analysis (right). The question mark icon in the top right hand corner links to information about each type of analysis. For the Sample Hierarchical Clustering plot, users have the option of three different distance functions (Euclidean, Pearson or Spearman) and three different linkage types (Average, Complete or Single). which may be selected from drop-down lists. Principal Component Analysis (PCA) is a mathematical approach that identifies and ranks the dimensions (principal components) that account for the largest proportion of variation within a data set. Sliders above both plots allow users to switch between outputs based on the whole transcriptome (used here), or genes in the gene basket (if defined at this stage).

2. Gene Expression:

The plots in this tab provide the first view of results on an individual gene level. Expression levels (expressed in transcripts per million, TPM) for individual genes (from the gene basket) may be viewed as box plots or bar graphs. This provides the opportunity to confirm whether genes behaved in the expected manner between experimental conditions or sample types.

For this analysis, we selected ten genes. Three of these are commonly regarded as “housekeeping” genes, whereas the other seven were randomly selected from gene sets previously shown to be associated with breast cancer.[9,10] Plots are shown in Figure 6 on the next page. Both the box plot (left) and bar graph (right) reflected the expected results for all of the selected genes, across all eight of the samples.

3. Differential Expressions:

This page provides a quick view of genes or gene sets that are up- and down-regulated within in a group of samples. This part of the *Visualizations* module provides the opportunity to really interact with the data, by creating comparisons between groups of data, changing threshold values, and selecting individual data points or groups of data points to explore further.

Two Differential Expressions (DE) groups, namely “KAPA tumor vs. normal” and “TruSeq tumor vs. normal” were created. For each, “Tumor” was entered as the *Case selection name* and “Normal” as the *Control selection name*, after which the appropriate samples were associated with each group and case. Analysis with the DESeq2 tool took several seconds. Differentially expressed genes could then be browsed, sorted, selected and saved as gene sets; and threshold parameters could be changed as desired (from the default values of 2 for fold change (up- and down-regulation) and 0.05 for false discovery rate (FDR).

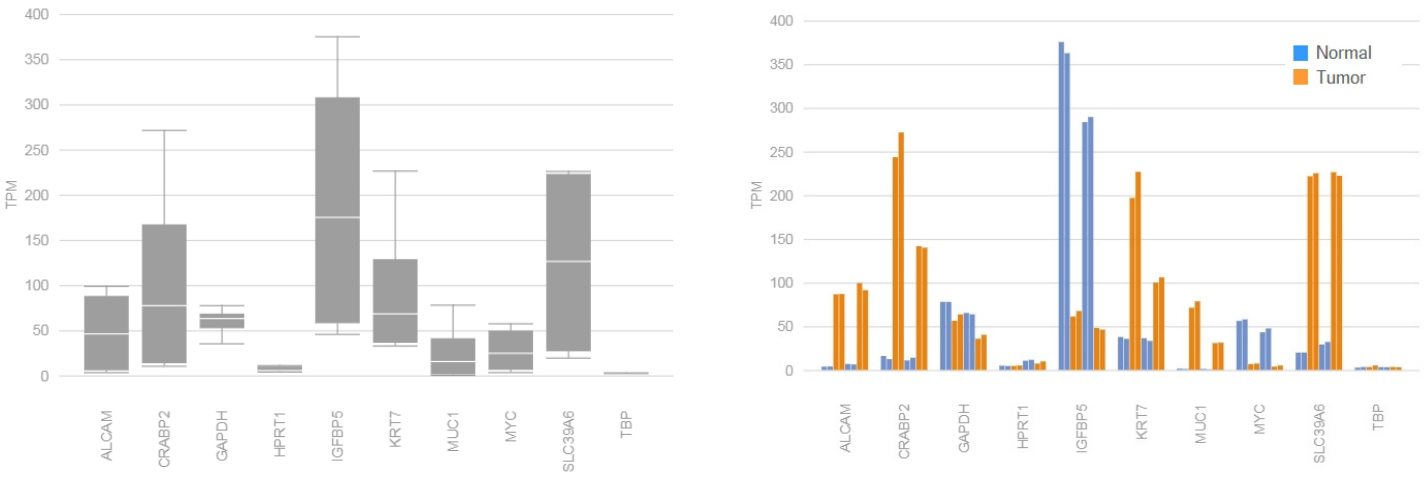

Figure 6. Expression levels for ten selected genes, visualized in the *Box Plot* (left) and *Bar Chart* (right). The *Box Plot* illustrates the distribution, central value, and variability of the expression levels of each gene, across the set of eight samples. The *Bar Chart* provides a view of the expression levels of each gene in all eight samples. The "Color by Source" option was used to color-code normal (blue) vs. tumor (orange) samples. For both plots, expression levels (y-axis) may be transformed from a linear TPM (shown here) to a $\log_2$(TPM +1) scale using a toggle. For the three housekeeping genes (*GAPDH*, *HPRT1* and *TBP*), no differential expression was observed between normal and tumor samples. *IGFBP5* and *MYC* are significantly down-regulated in breast tumor samples, whereas *ALCAM*, *CRABP2*, *KRT7*, *MUC1* and *SCL39A6* are up-regulated to different degrees. Individual genes in the selection (genes basket) may be highlighted to obtain additional information via links to external sources (e.g., ENSEMBL). Plots update in real-time as genes are added (up to a maximum of 20) or removed from the gene basket.

Selecting a DE analysis automatically populates a *Volcano Plot* (Figure 7). In this plot, every dot represents a gene. A separate plot was generated for each DE group (KAPA and TruSeq). Next, all genes that were up- or down-regulated in the tumor vs. normal samples by ≥2-fold (FDR ≤0.001) were selected for each DE group. This produced:

- for the KAPA workflow, a set of 5,282 genes, of which 2,597 were up-regulated in tumor vs. normal samples, whereas the remaining 2,685 were down-regulated. All of these genes were saved as a gene set ("KAPA_all up and down_5282").
- for the TruSeq workflow, a set of 4,061 differentially expressed genes, of which 1,799 were up-regulated and 2,262 were down-regulated in tumor samples. This gene set was saved as "TruSeq_all up and down_4061".

According to this analysis, the KAPA workflow yielded 30% more differentially expressed genes than the TruSeq workflow, from a similar amount of sequencing (an average of 18.2 million read pairs per KAPA library vs. an average of 18.4 million read pairs per TruSeq library). This difference in data yield was attributed to a more efficient KAPA workflow, resulting in significantly less reads associated with residual rRNA transcripts (Table 2), and approximately 20% fewer duplicate reads (average for four libraries prepared with each workflow; data may be found in MultiQC reports.

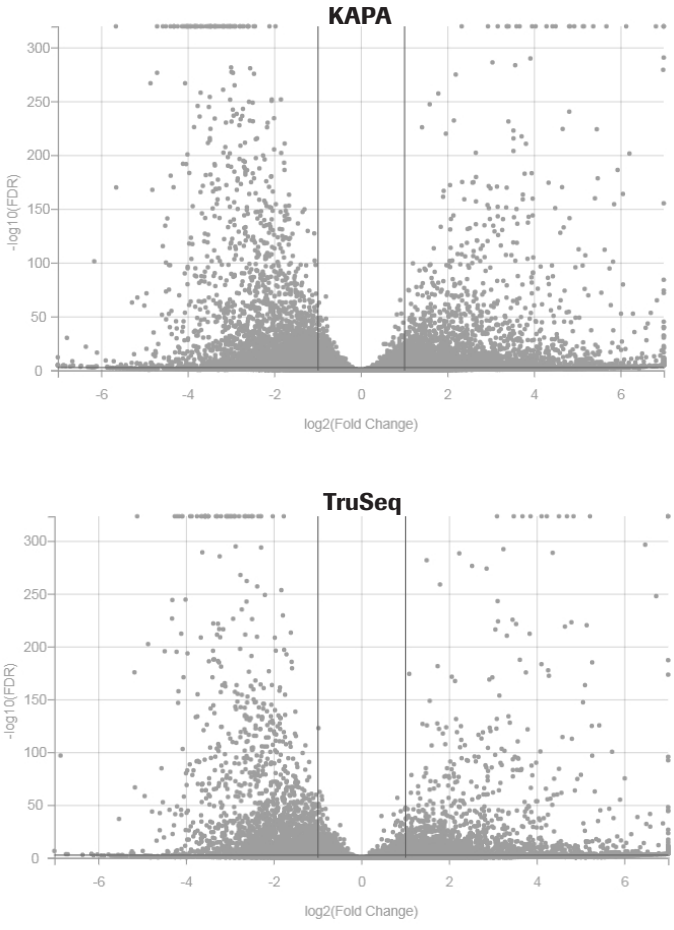

Figure 7. *Volcano Plots* for the KAPA (top) and TruSeq (bottom) Differential Expression (DE) groups. Every dot represents a gene. The statistical false discovery rate ($-\log_{10}$FDR) is plotted on the y-axis against relative fold change ($\log_2$FC, x-axis). Thus, the further from zero a gene is displayed, the greater the difference in expression level between the two conditions (x-axis) and the greater the statistical confidence (y-axis). The FC threshold is demarcated by the two darker, vertical lines in the middle of the plot, whereas the darker horizontal line (y=3 in these plots) represents the FDR threshold. These thresholds may be modified, and doing so will change the lines on the plot. Outliers (genes with a $\log_2$FC >7) are stacked by default, but may be selected with the mouse to display their actual values. Genes of interest may be selected by drawing a box around those dots with the mouse. The resulting pop-up enables the user to inspect those genes, append or overwrite them to the genes basket, or save them as a gene set. Alternatively, genes or subsets of genes may be selected from the DE group, by clicking on the group name.

4. Venn View:

To further investigate the above results, a Venn diagram was generated from the "KAPA_all up and down_5282" and "TruSeq_all up and down_4061" gene sets. Venn diagrams provide a visual manner to organize information, and to easily define relationships between subsets of data. The diagram (Figure 8 on the next page) showed that 3,718 of the differentially expressed genes were detected by both workflows, whereas 1,564 were unique to the KAPA workflow, and 343 were detected by the TruSeq workflow only.

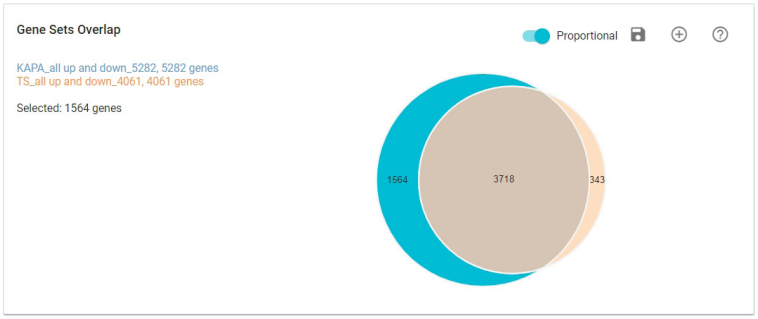

Figure 9. The *Gene Sets Overlap* card on the *Venn View* page, showing the Venn diagram. The diagram was generated by comparing all the up- and down-regulated genes (fold change ≥2, FDR ≤0.001) identified in the KAPA (blue) and TruSeq (orange) workflows. The 1,564 selected genes are unique to the KAPA workflow. New Venn diagrams are created by clicking the plus icon, and selecting gene sets from the pop-up; or previously saved Venn overlaps may be reconstituted. Up to four gene sets can be compared in a single diagram. New gene sets may be saved and further interrogated by selecting segments of the Venn diagram. Genes in selected segments can be viewed in the *Heat Map*, or functional annotation may be obtained from *Gene Ontology (GO) Enrichment Analysis* on the *Venn View* page. Pathway analysis may also be performed via a link to the *Enrichr* tool. The area of each Venn region is displayed log-proportional to the number of genes represented, but this can be toggled to choose areas of equal size.

The *Gene Ontology Enrichment Analysis* available on the *Venn View* page indicated that the 1,564 differentially expressed genes unique to KAPA workflow are strongly associated (p ≤0.01) with a number of biological processes (including regulation, adhesion, localization and metabolic processes), as well as molecular functions (regulation, transporter activity, binding, catalytic activity, and transcription factor activity/protein binding). Further investigation also revealed the "unique KAPA" gene set to include genes used in breast tumor subtyping (*CCNB1*, *EXO1* and *FGFR4*, which form part of the PAM50 classifier[9]), as well as genes associated with breast cancer survival (*BTN3A3* and *KIF3C*, which are included in the SAM264 gene classifier[10]).

Before moving to the last *Visualizations* tool (the *Heat Map*, see below), the 1,564 "unique KAPA" genes were saved as a gene list. At this stage, we returned to the *Differential Expressions* page, and selected both the KAPA and TruSeq DE groups to generate a *Differential Expressions Comparison Plot*, on which the 1,564 genes were highlighted (Figure 10, top). As expected, all of the genes fell outside the area defined by a $-1 \geq \log_2 FC \geq 1$ on both axes. The majority of genes also fell off the x=y diagonal, indicating that abundances differed between the two DE groups.

The *Differential Expressions Comparison Table* was used to sort the 1,564 "unique KAPA" genes in different ways to learn more about differences in gene abundances between the two workflows. In the excerpt shown in bottom half of Figure 10, genes were sorted in ascending order based on their $\log_2 FC$ value for the TruSeq DE group. This produced a group of eleven genes that were detected in the KAPA samples, but not in any of the TruSeq libraries. Further investigation (using the link-out to ENSEMBL) revealed that nine of these eleven genes contain regions of high GC content and/or GC-repeats. The above findings were consistent with previous observations that KAPA RNA HyperPrep Kits offer improved coverage of GC-rich and low-abundance genes.[2]

5. Heat Map:

Quantitative differences in expression levels of selected genes in individual samples are plotted in the *Expression Heat Map*. This allows the user to obtain a visual overview of the transcriptome landscape of different biological or experimental conditions.

In a recently published paper, 110 putative genes were identified in 33 breast cancer risk loci, using the Hi-C technique.[11] We thought it would be interesting to see how many of these genes were detected in the libraries prepared for this study. A gene set comprising the genes from the paper was defined,

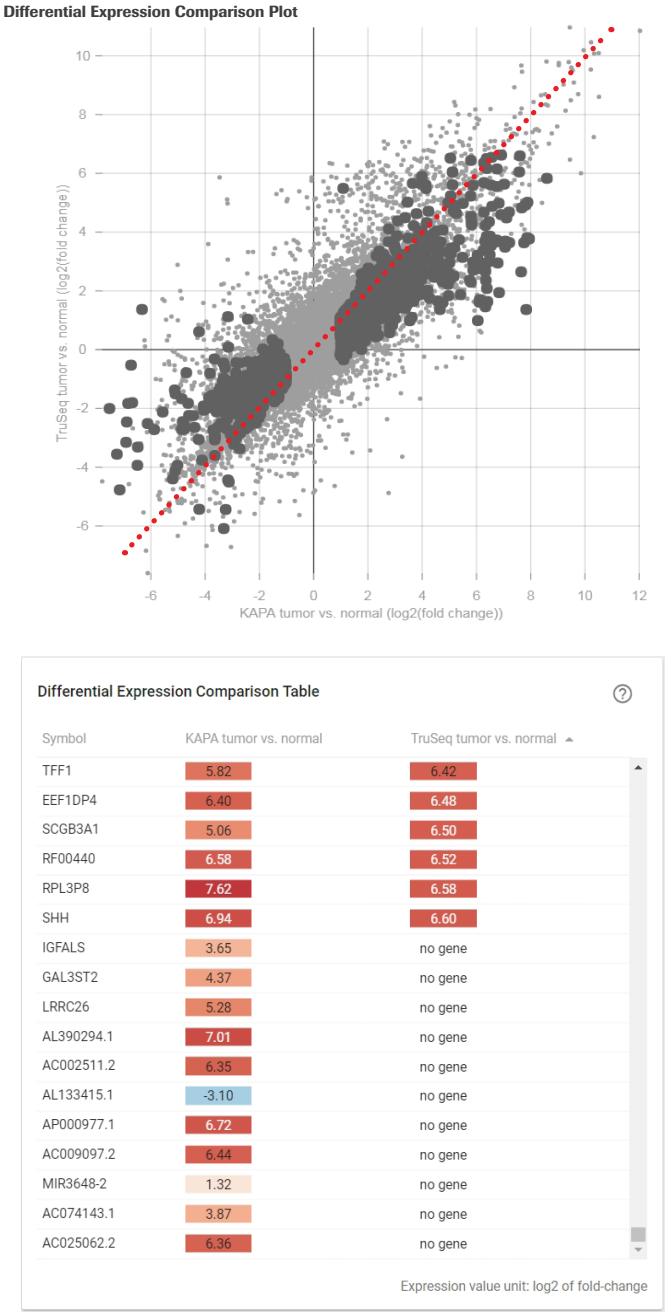

Figure 10. *Differential Expressions Comparison Plot (top) and an excerpt from the Differential Expressions Comparison Table (bottom)*, for the KAPA vs. TruSeq DE groups, with the 1,564 "unique KAPA" genes highlighted. Every dot in the comparison plot represents a gene, allowing the user to simultaneously compare the relative fold change of all the genes in the transcriptome. The Pearson correlation for this plot was 0.85. Genes that fall along the red x = y diagonal display similar abundance patterns in both DE groups, which is not the case for the majority of the highlighted genes. Actual $\log_2 FC$ values for the 1,564 selected genes may be obtained from the *Differential Expression Comparison Table* depicted below the plot. Values may be sorted by any column, and are shaded using a color scale (similarly to a heat map) to facilitate detection of genes with markedly different abundance patterns. This excerpt of the table shows the eleven genes that were not detected in the TruSeq workflow.

after eliminating entries that could not be found in the ENSEMBL database. These genes were highlighted in the volcano plots for the KAPA and TruSeq DE groups, respectively (not shown). Inspection of the *DE Comparison Table* for each DE group revealed about a dozen genes from the list that were not detected in the KAPA and/or TruSeq libraries. These genes were deleted from the list, to yield a final set of 80 genes, which were used to generate the *Heat Map* (Figure 11).

For several of the 80 genes (e.g., *SNX32* and *CDCA7*, positioned at top and bottom ends of the heat map), technical replicates returned inconsistent results. Notwithstanding this experimental variation, the heat map divides the genes from the paper into four noteworthy groups:

- genes that have a higher abundance in tumor vs. normal samples for both workflows (most of the genes in block A);
- genes that have a higher abundance in normal vs. tumor samples for both workflows (block B);
- genes that show different abundance profiles based on workflow, rather than tissue type (most genes in the two C blocks); and
- genes that have a higher abundance in KAPA tumor samples only (block D).

The expression patterns in blocks A and B were attributed to biological variation between tumor and normal tissues, whereas the patterns in blocks C and D were likely the result of experimental factors, including biases introduced during RNA enrichment and/or library construction. Further investigation of select genes in blocks C and D again confirmed that the KAPA chemistry offers better coverage of genes with a high overall GC-content, or that contain GC-rich motifs.

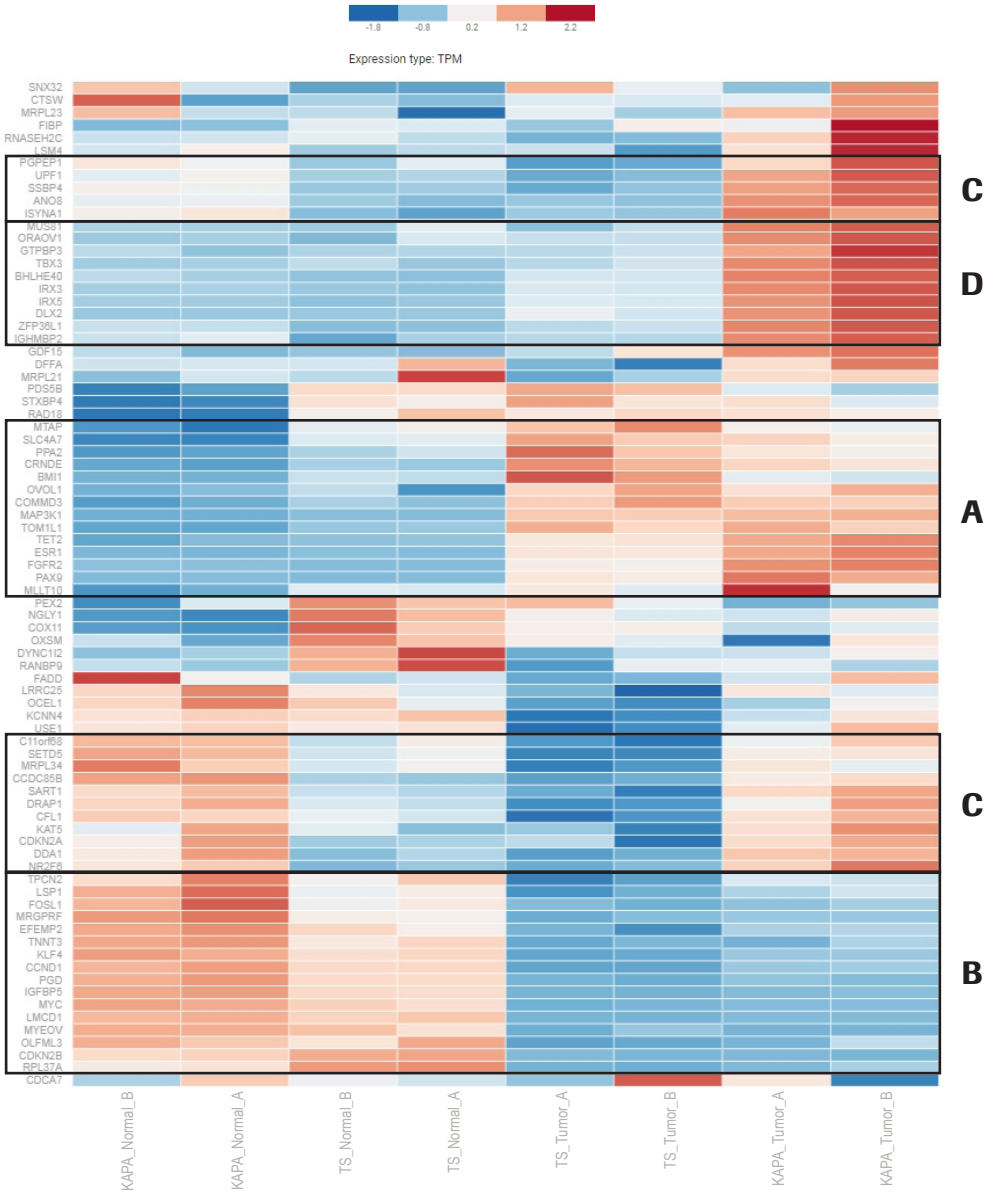

Figure 11. *Expression Heat Map* for the eight libraries sequenced in this study, defined by 80 putative genes associated with 33 breast cancer risk loci. Hierarchical clustering in this plot is based on Euclidean distance, which is applicable regardless of the data transformation approach used, and robust with respect to non-normal data distributions. Different row-wise data transformations, including Z-score (default; used here), $\log_2$ or Z-score of $\log_2$, are available. Selecting a different transformation method will recompute the clustering and modify the color scales accordingly. The heat map updates automatically when genes are added/removed from the gene basket. Gene names may be displayed or not, and additional information is revealed when hovering with the mouse over any cell.

Confirmation of results by RT-qPCR

Gene expression analysis data and insights obtained by RNA-Seq are often confirmed by an orthogonal method, such as micro-arrays or reverse transcription quantitative PCR (RT-qPCR).

Target selection

In this study, 66 genes shown to be differentially expressed between tumor and normal libraries generated with the KAPA workflow ($-1 \geq \log_2 FC \geq 1$, FDR ≤ 0.001) were randomly selected from the KAPA DE group, based on the availability of a RealTime ready qPCR Assay (Roche). The selected genes represent a wide range of fold changes and confidence levels based on the RNA-Seq data, as indicated in the volcano plot shown in Figure 12.

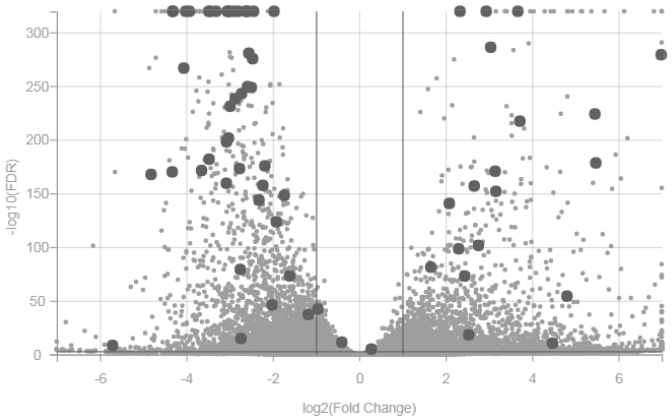

Figure 12. Genes selected for confirmation by RT-qPCR. The 66 genes selected based on availability of Roche RealTime ready qPCR Assays, represent a range of fold changes and confidence (FDR or p values) in the KAPA DE group.

RT-qPCR protocol

cDNA was generated from the same tumor and normal RNA extracts used for RNA-Seq library construction, using the Transcriptor First Strand cDNA Synthesis Kit (Roche). Total RNA (600 ng) was used as input, and reverse transcription was performed with both random hexamers and anchored oligo-dT primers. After heat-inactivation, 200 ng of cDNA was combined with a qPCR master mix (Roche LightCycler® 480 Probes Master), aliquotted into a RealTime ready assay plate, and amplified according to standard recommendations, using the LightCycler® 480 System (Roche). Assays were performed in duplicate, and relative differential gene expression values were calculated using the "ΔΔCp" method.[12]

Log_2-transformed fold changes obtained from the RT-qPCR assay was plotted against the log_2-transformed fold changes from the RNA-Seq analysis (Figure 13). The R^2 value of 0.77 indicated a good correlation between the fold changes obtained by RNA-Seq vs. RT-qPCR.

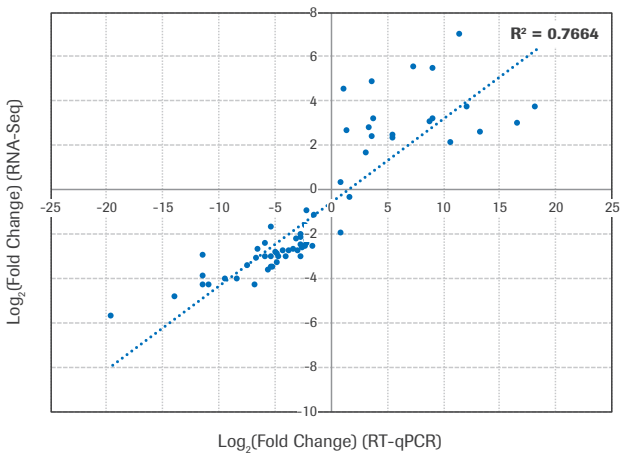

Figure 13. Confirmation of RNA-Seq results by RT-qPCR. Hydrolysis probe-based RT-qPCR was performed using Roche RealTime ready qPCR Assays, as described above.

Conclusions

In this study, replicate RNA-Seq libraries were prepared from breast cancer and matched normal tissue samples of fair quality, using the KAPA RNA HyperPrep Kit with RiboErase (HMR), and the TruSeq® Stranded Total RNA with Ribo-Zero Gold kit (Illumina®). Sequencing data was analyzed with the Genialis™ platform. The platform's intuitive user interface, data management capabilities, versatile visualization tools, and links to external genomics resources made it simple to assess the integrity of experimental results, before interrogating the data in different contexts; without any bioinformatics expertise.

Using the suite of tools offered by the Genialis software, we were able to confirm that the KAPA workflow:

- offers more efficient depletion of rRNA transcripts prior to library construction (% reads associated with residual rRNA was 7.5- to 10-fold lower than for the TruSeq workflow).
- produced higher-quality libraries (the average % duplicate reads for the four KAPA libraries was approximately 20% lower than for the four TruSeq libraries).
- yielded approximately 30% more genes that were up- or down-regulated by ≥2-fold in tumor vs. normal samples (FDR ≤0.001), as compared to the TruSeq workflow, from the same amount of sequencing.
- better preserved genes with a high GC-content or GC-rich sequence motifs, including genes previously associated with breast tumor subtyping, risk and survival.
- produced gene expression results that correlated well with RT-qPCR results, for the subset of 66 genes included in the confirmation experiment.

As demonstrated in this study, Roche and Genialis offer a qualified, simple and complete workflow solution for NGS-based gene expression profiling, from RNA to analysis, consisting of the following components:

- KAPA RNA HyperPrep Kits, with an mRNA capture module, or the KAPA RiboErase (HMR) module for the selective depletion of rRNA, globin transcripts (from blood samples), and/or any other transcripts of choice (with user-supplied depletion oligos[13]). KAPA RNA HyperPrep Kits contain all the components for cDNA synthesis and library construction, including KAPA Pure Beads for reaction cleanups, and KAPA HiFi HotStart ReadyMix for library amplification.
- KAPA Adapters, which are available separately.
- KAPA Library Quantification Kits for qPCR-based library quantification prior to normalization and pooling for multiplexed sequencing.
- Reagents for RT-qPCR, including the Transcriptor First Strand cDNA Synthesis Kit for cDNA synthesis, RealTime ready qPCR Assays and the LightCycler® 480 Probes Master for hydrolysis-based qPCR.
- The LightCycler® 480 System for high-throughput library quantification and RT-qPCR.
- The Roche-qualified gene expression application on the Genialis platform.

Benefits of the complete Roche-Genialis gene expression analysis solution include the following:

- A reduction in overall time from RNA to analysis. Streamlined, automation-friendly KAPA RNA HyperPrep Kits, combined with an automated, pre-configured bioinformatics pipeline make it possible to complete experiments in days, rather than weeks or months.
- Improved sequencing economy. Highly efficient KAPA RNA enrichment and library construction chemistry results in fewer reads associated with unwanted transcripts and PCR duplicates, and better preservation of "difficult" (e.g., GC-rich) content. This translates to the detection of more differentially expressed genes, and improved sensitivity.
- Flexibility and control over your data. The intuitive, cloud-based Genialis software, offers real-time visualization tools for biologists with limited bioinformatics expertise, while facilitating data management and sharing.
- Peace of mind from using a qualified workflow, with integrated support from Roche and Genialis.

References

1. Han Y, Gao S, Muegge K, et al. Advanced Applications of RNA Sequencing and Challenges. *Bioinformatics and Biology Insights* 2015,9(S1):29 – 46 doi: 10.4137/BBI.S28991.
2. Roche. Roche Sample Prep Solutions for RNA-Seq. 2018. Accessed October 2018.
3. Donovan CA, Giuliano AE. Evolution of the Staging System in Breast Cancer. *Ann Surg Oncol.* 2017,24:3469. doi: 10.1245/s10434-017-6035-8.
4. Lodish H, Berk A, Zipursky SL, et al. *Molecular Cell Biology.* 4th edition. New York: W. H. Freeman; 2000. Section 11.6, Processing of rRNA and tRNA. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21729/. Accessed September 2018.
5. Roche. KAPA RNA HyperPrep Kit with RiboErase (HMR), Illumina® platforms. Technical Data Sheet. 2017. Accessed October 2018.
6. Illumina. TruSeq Stranded Total RNA Reference Guide. 2017. Accessed October 2018.
7. 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. This analysis was performed prior to the development of the Roche pipeline on Genialis.
8. Roche. Sequencing Solutions Technical Note: How To… Prepare libraries from degraded RNA inputs with the KAPA RNA HyperPrep Kit with RiboErase (HMR) for whole transcriptome sequencing. 2018. Accessed October 2018.
9. Parker JS, Mullins M, Cheang MCU, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol*. 2009,27(8):1160. doi: 10.1200/JCO.2008.18.1370.
10. Jenssen TK, Kuo, WP, Stokke T, et al. Associations between gene expressions in breast cancer and patient survival. *Hum. Genet.* 2002,111:411. doi: 10.1007/s00439-002-0804-5.
11. Baxter JS, Leavy OC, Dryden NH, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nature Communications* 2018,9:1028. doi: 10.1038/s41467-018-03411-9.
12. Livak KJ, Schmittgen TD. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* 2001,25(4):402. doi: 10.1006/meth.2001.1262.
13. Roche. KAPA RiboErase (HMR) Kits offer a flexible technology for selective transcript depletion prior to library construction for whole transcriptome analysis. 2018. Accessed October 2018.
14. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013,29:15. doi: 10.1093/bioinformatics/bts635.
15. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014,30(7):923. doi: 10.1093/bioinformatics/btt656.

Appendix: Qualified Genialis™ gene expression analysis pipeline for KAPA RNA-Seq library preparation kits

The gene expression analysis pipeline described in Table A1 was co-developed by Genialis and Roche bioinformatics teams, for complete RNA-to-analysis sequencing workflows. The pipeline has been qualified for stranded libraries generated with the KAPA RNA HyperPrep Kit, with RiboErase (HMR) or RiboErase (HMR) Globin, or the KAPA mRNA Capture module. Support for the older-generation portfolio of KAPA Stranded RNA-Seq Library Preparation Kits can also be provided. Roche and Genialis are committed to continued collaboration, to expand features of the gene expression application, and provide future support for additional sequencing applications.

Table A1. Data analysis tools and specifications

Process	Program/Algorithm and version	Description/parameters/comments
Adapter removal and quality trimming (single- or paired-end reads)	BBDuk (BBMap 37.90)	▪ A selection of Illumina adapters is already available on the platform. Should these not suffice, a user can add their own adapter sequences upon data upload. ▪ Parameters: minlength (20); k (23); hammingdistance (1); ktrim (r); mink (11); qtrim (r); trimq (30)
Alignment	STAR[14] (2.5.4b)	▪ Maps to reference genomes, which are already available on the platform (*Homo sapiens*, *Rattus norvegicus* and *Mus musculus*; all ENSEMBL version 92). ▪ Default parameters
Rate of rRNA and globin mRNA depletion	Seqtk (1.2-r94) STAR (2.5.4b)	▪ Sub-sampling of trimmed reads and subsequent alignment to globin and rRNA reference sequences using STAR
Gene expression quantification	featureCounts (1.6.0)[15]	▪ Uses annotations of respective genome versions. ▪ A custom script (expression_fpkm_tpm.R) is used to calculate normalized expression values (FPKM and TPM). Users can optionally select DESeq2 (now) or EdgeR (soon) to run differential expression analysis. ▪ Parameters: strand-specific read counting with featureCounts parameters set to match the ENSEMBL-derived GTF file.

Additional bioinformatics pipelines, tools and applications are available for advanced users. Please contact Genialis for more details.

Published by:

Roche Sequencing Solutions, Inc.
4300 Hacienda Drive
Pleasanton, CA 94588

sequencing.roche.com

SEQ100361 10/2018