Next-Generation Sequencing Gene Expression Analysis Using Agilent GeneSpring GX

Technical Overview

Introduction

RNA Sequencing (RNA-Seq) is one of the most commonly used next-generation sequencing (NGS) techniques in the research community today. With the massive amount of sequence data generated every day, the need to easily analyze and interpret the biological context of RNA-Seq data has become a pressing issue. The NGS gene expression analysis workflow in Agilent GeneSpring GX is designed to address this need. Figure 1 shows the three stages a typical RNA-Seq experiment goes through.

The NGS gene expression analysis workflow in GeneSpring GX allows expression profiling, differential analysis, as well as functional interpretation of processed RNA-Seq data. It also provides several visualization options. This workflow enables working with publicly available datasets from sources such as TCGA and GEO. GeneSpring GX provides tools for integration of genomics, transcriptomics, metabolomics, or proteomics data in a single view.

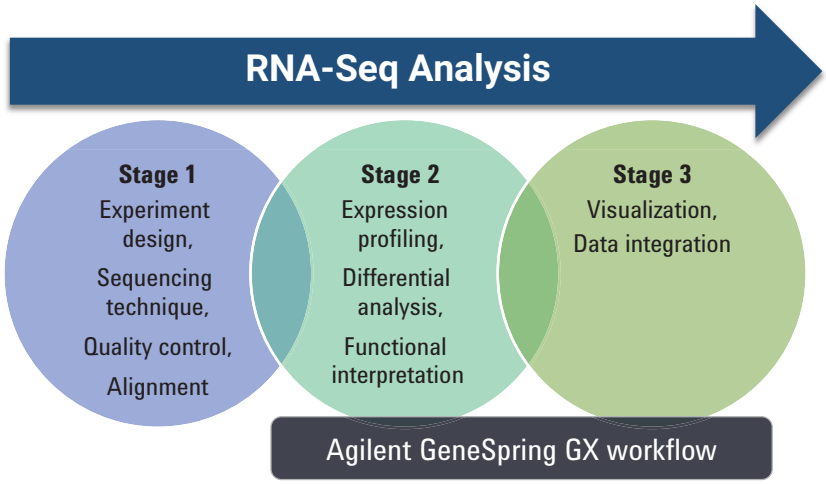

Figure 1. Different stages of a standard RNA-Seq analysis. Agilent GeneSpring GX provides analysis tools described in Stages 2 and 3.

Agilent Technologies

Workflow

The workflow enables the analysis of processed RNA-Seq data starting with a text file. The upstream processing, including alignment and quantification methods, are standardized, and can be performed in either a commercial or publicly available software tool prior to importing the data in GeneSpring GX. A researcher can quickly import the quantified and normalized data, examine its quality, and proceed with statistical analysis as well as biological contextualization.

Figure 2 shows the core features of NGS gene expression analysis in GeneSpring GX.

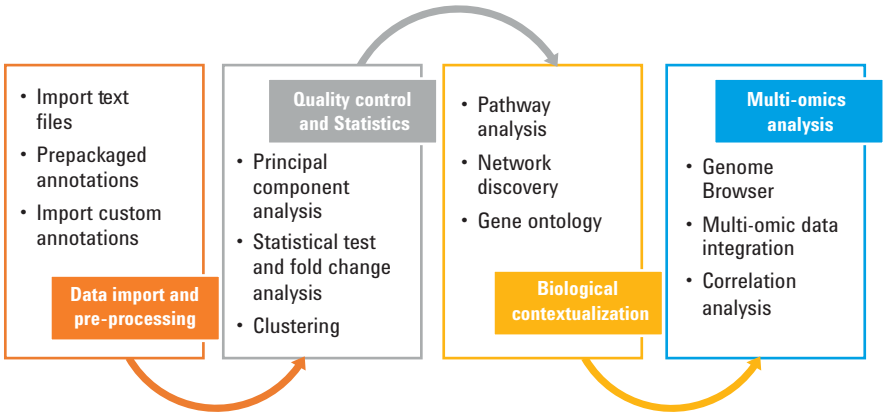

Figure 2. NGS gene expression analysis workflow in Agilent GeneSpring GX.

Data import and preprocessing

A multisample text file with an identifier and one or more sample data columns can be loaded in the NGS gene expression analysis workflow. Figure 3 shows the steps involved in data import and preprocessing.

Annotations for the genes can be added using genes and transcript models downloaded from the Agilent server through the annotations manager. Annotation manager is the central location to host all annotation databases for various organisms. GeneSpring GX also allows adding annotations from external files, and during this process, the software automatically assigns a default column mark, which, if required, can be changed manually for each annotation.

During the data import, metadata information for samples can be entered in the Experiment Grouping window. To look at the relative expression of the data for visualization, tools such as clustering and baseline transformation can be applied to the samples.

Figure 3. Data import and preprocessing in an NGS gene expression analysis workflow.

Quality control and statistics

Quality control (QC) options to examine sample and gene quality are available in the workflow browser. To examine similarity between replicates, the multipanel QC window has a Principal Component Analysis (PCA) plot, correlation values, and correlation plot. Outlier samples can be removed from further analysis using the Add/Remove option.

GeneSpring GX provides multiple filtering options including filter by expression, error, and normality. These options can then be applied to data sets that have passed sample QC. These filters are used to retain genes based on user-defined parameters for further analysis, including differential analysis.

Functional interpretation

Biological aspects of significant genes can be explored using tools such as Gene Ontology (GO), Gene Set Enrichment Analysis (GSEA), Gene Set Analysis (GSA), or Pathway analysis. To identify the enriched categories, GO analysis uses categories defined by the GO consortium with a statistical test in GeneSpring GX.

GSEA and GSA are gene set enrichment methods that use either the broad gene sets or custom gene lists in GeneSpring GX. This test identifies genes that are both biologically and statistically significant based on association to the broad gene sets and the enrichment calculation, respectively.

Figure 4 shows an NGS gene expression analysis workflow in GeneSpring GX.

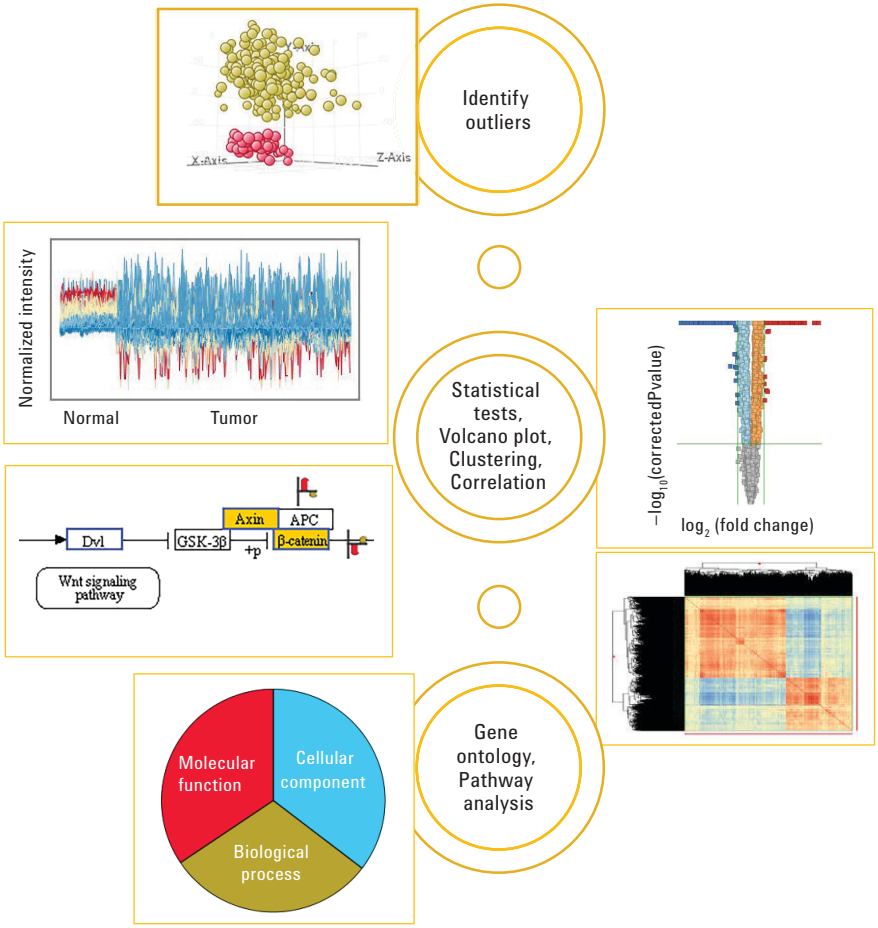

Figure 4. Agilent GeneSpring GX analysis workflow showing quality control, statistical, and biological contextualization analysis options.

Data visualization

GeneSpring GX offers several data visualization options of which Genome Browser is the most powerful tool for RNA-Seq data. A variety of plots including the genome browser, profile plot, and scatter plots can be launched from the tool bar.

All the gene associated information can be viewed side-by-side in the Genome Browser, as shown in Figure 5. The elastic genome browser can be used to zoom into one or more genomic regions of interest simultaneously. It allows pinning regions of interest, and showing them at different zoom levels.

The data tracks enable visualization of normalized values or list-associated values such as fold change and p-value.

The annotations displayed in Genome Browser can be downloaded from Annotations Manager. GeneSpring GX also allows the import of custom annotations as well as information from public databases such as UCSC, RefSeq, Ensembl, and GENCODE.

Figure 5. Data visualization in Genome Browser. A) Data track with fold change values. B) UCSC gene annotation track. C) UCSC transcript annotation track. D) GENCODE gene annotation track. E) Spreadsheet with gene annotations as well as list-associated values such as p-value and fold change value.

Pathway analysis

Single Experiment Analysis (SEA) for pathways can be used to identify significant pathways that show enrichment of a pathway based on a statistical test. A hyper-geometric p-value is calculated for each pathway. The pathways used in this analysis can be directly imported by connecting to the WikiPathways server for all of its content, and the Agilent server for KEGG and BioCyc content. Custom pathways can also be imported from file in GeneSpring GX. In the pathway view, the matched entities between the gene list used for the analysis and individual pathway are colored in yellow by default. The expression value(s) of genes can be seen in multiple ways in a pathway viewer, namely heatstrip, quilt plot, data table, and heatmap. Figure 6 shows overlapping genes colored in yellow along with a heatstrip showing the expression changes across normal and tumor sample groups for the matched genes.

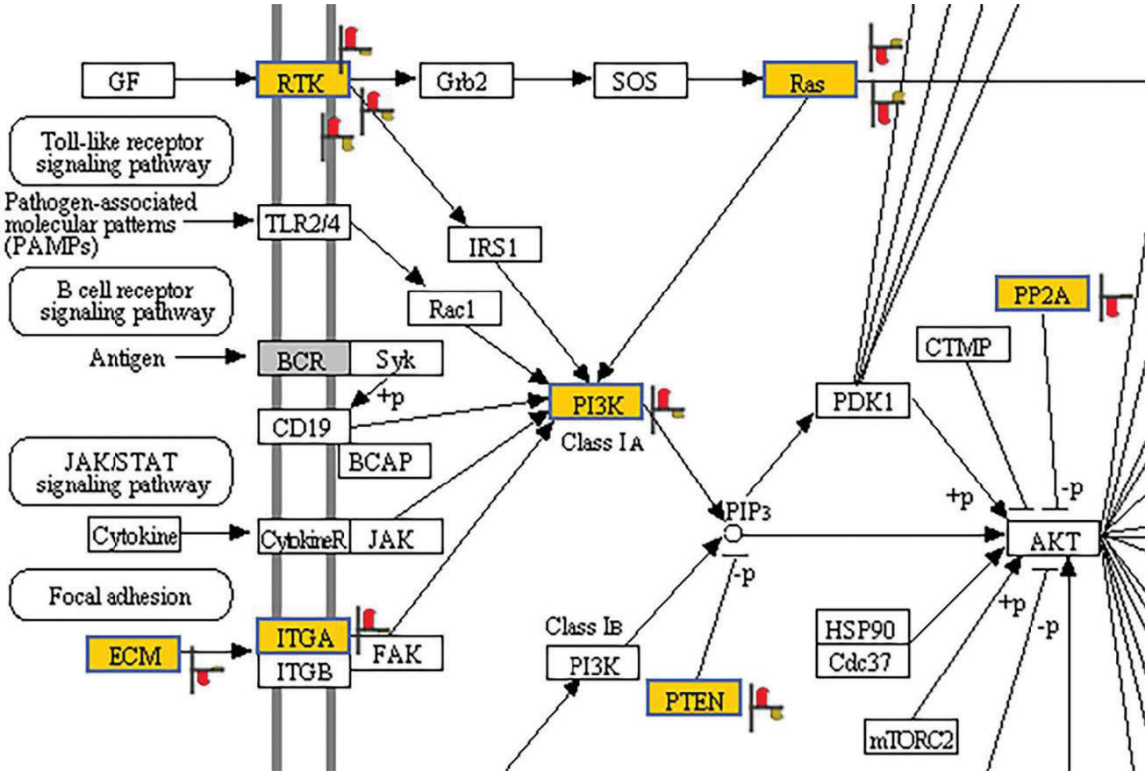

Figure 6. Enriched pathway showing genes overlapping between genes in the pathway and the genes in the list used for the analysis.

Data integration

GeneSpring GX provides multiple ways to integrate data within a single experiment or across multiple experiments. To look at the common genes or pathways across the different experiments, a Venn diagram can be created using either the gene lists or pathway lists. The Venn diagram on entity lists enables entity matching using user-specified annotations. Figure 7 shows gene overlap between four gene lists from gene expression microarray, miRNA microarray, and NGS gene expression analysis experiments.

The Genome Browser (Figure 8) also allows data integration from multiple experiments. Entity lists from one or more experiments can be dragged and dropped on the Genome Browser. For simultaneous visualization of gene level information, further annotation tracks can be added.

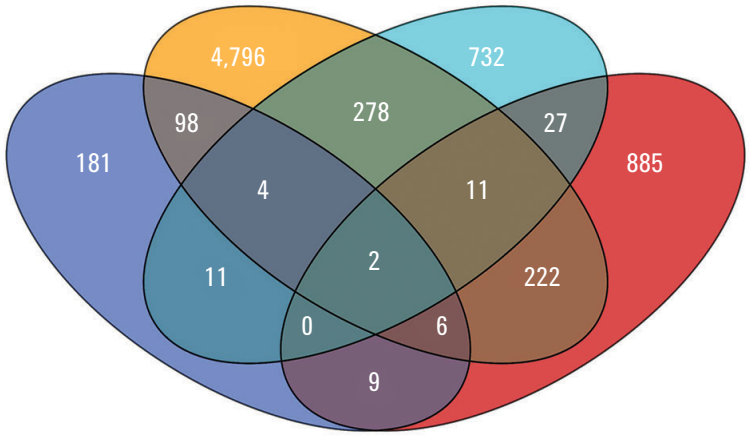

Figure 7. A Venn diagram between multiple gene lists.

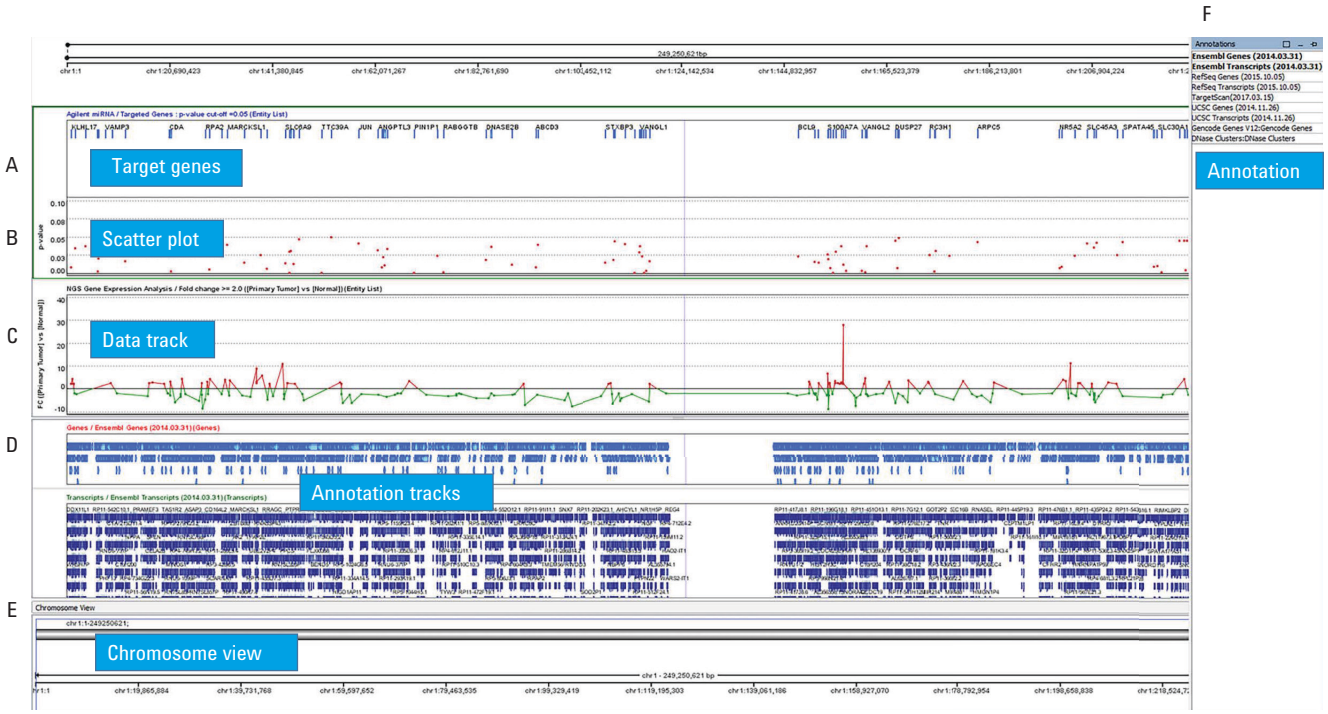

Figure 8. Genome Browser with multiple data tracks from NGS gene expression and miRNA data along with annotation tracks. A) Target genes from miRNA data. B) Scatter plot showing p-value associated with target genes. C) Data track with fold change values of significant genes from NGS Gene expression experiment. D) Ensembl gene annotation track. E) Ensembl transcript annotation track. F) Annotations available for display on Genome Browser.

Similarly, correlation analysis enables the visualization of similarity between two entity lists from the same experiment, or from two different experiments in multi-omics analysis (MOA) experiment. Figure 9 shows an example of correlation analysis between genes from NGS gene expression analysis and miRNA for a microarray experiment.

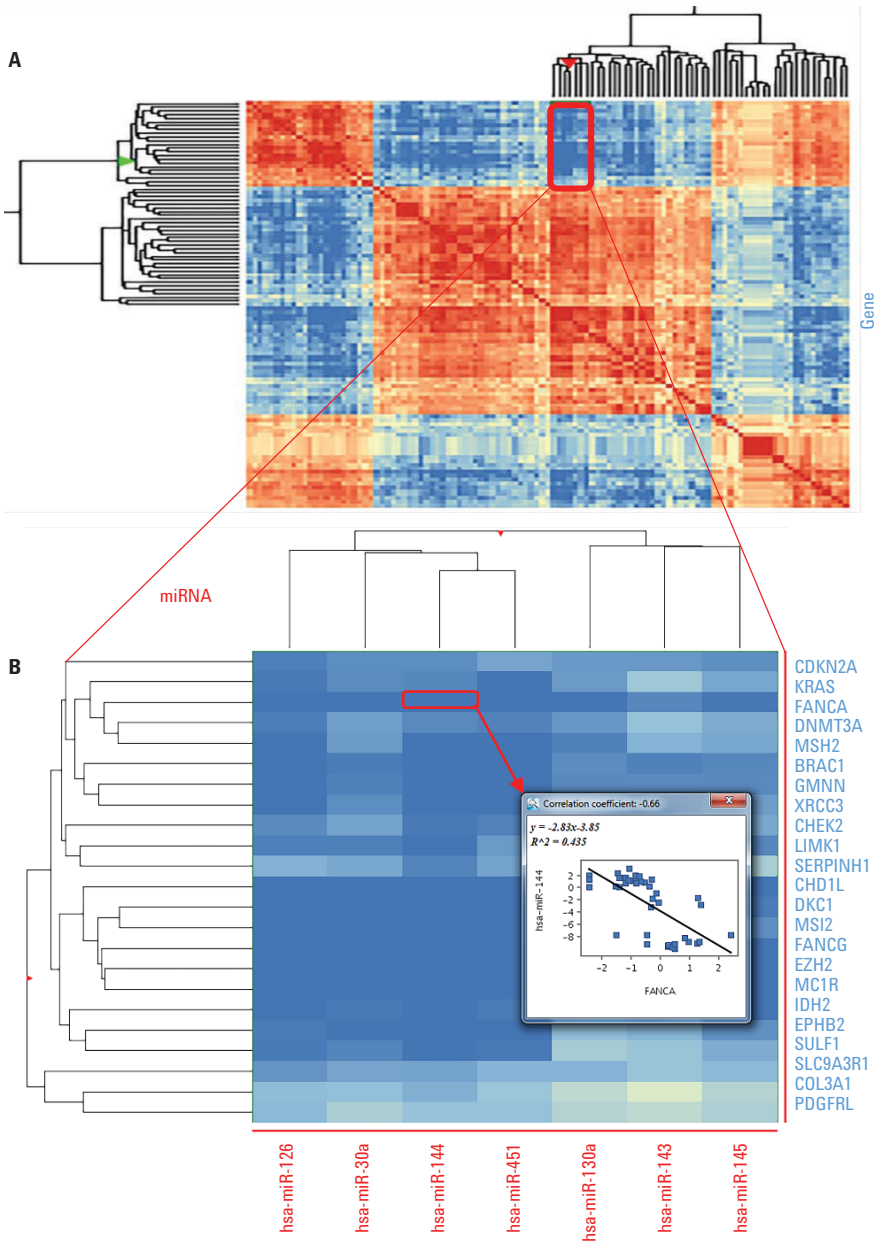

Figure 9. A) Correlation analysis between mRNA and miRNA data. B) Zoomed view, showing the correlation coefficient value of a gene and miRNA.

The pathway analysis in a multi-omics experiment enables simultaneous visualization of information from two experiments on a curated pathway. Figure 10 shows a significant pathway with genes from NGS gene expression analysis and the miRNA targets from a microarray experiment that were found to be involved.

The quilt plots seen in the view is from NGS gene expression analysis. Since the miRNA target genes do not have signal values associated with them, a quilt plot was not drawn. The analysis results, such as p-value or fold change values, can also be overlaid in this view.

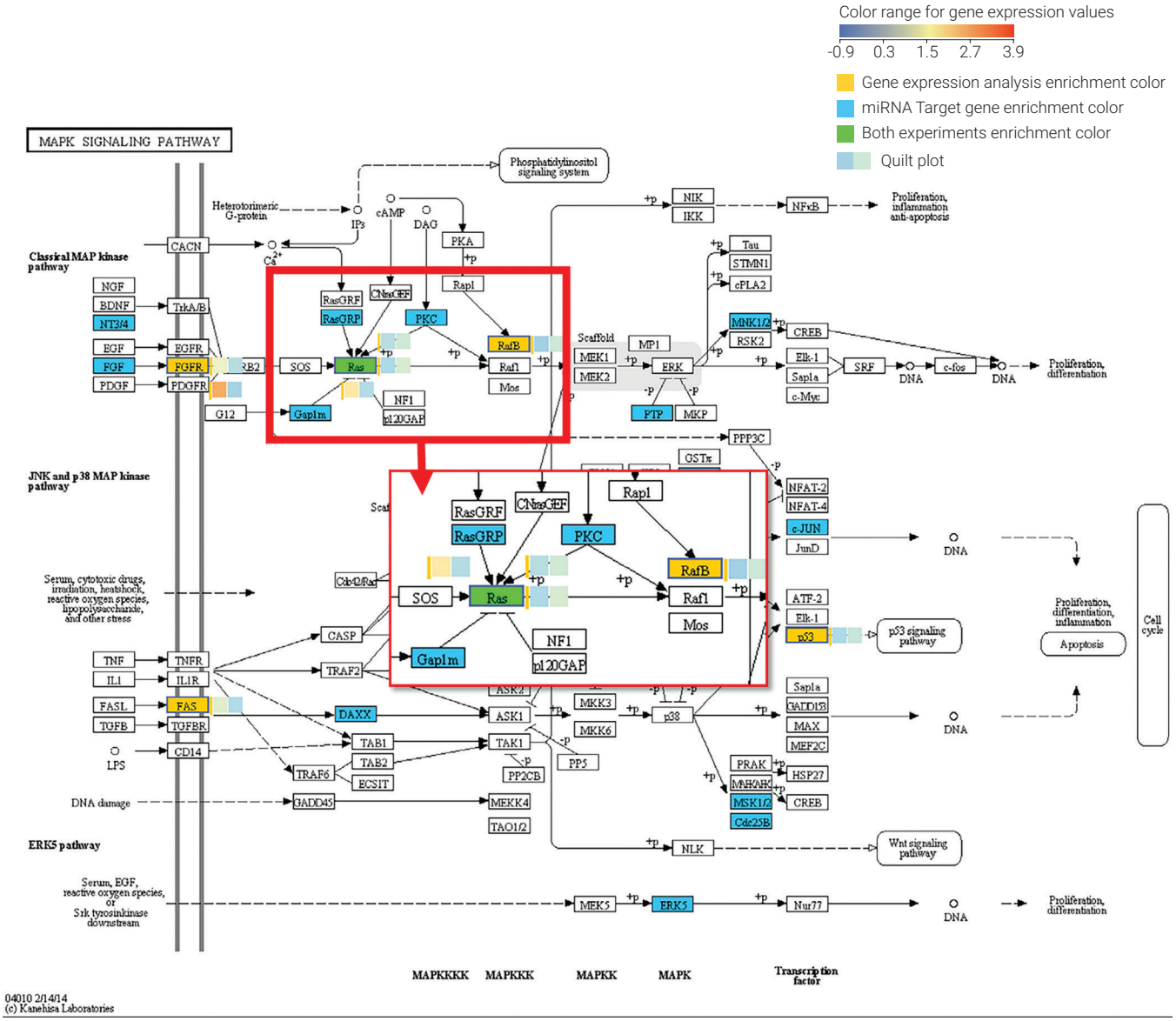

Figure 10. Pathway analysis in multi-omics analysis experiment with data overlay of mRNA and miRNA target genes.

Conclusion

Agilent GeneSpring GX's NGS gene expression analysis workflow offers a simplistic approach to analyze and interpret RNA-Seq data. A combination of statistical tests and visualization tools available in GeneSpring GX make the results interpretation process easier for a researcher. The GeneSpring suite enables seamless integration to support the growing interest in the research community in integrating either inter-genomic or multi-omic data. The ability to work with genomics, transcriptomics, metabolomics, as well as proteomics data in one platform brings all the major omics analyses under one umbrella, enabling easy interpretation of biological information from different disciplines.

www.agilent.com/chem

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

This information is subject to change without notice.

Published in the USA, September 25, 2017
5991-8333EN

Agilent Technologies