

Real-Time StatMiner™ Software

Advanced Data Mining Software for Real-Time PCR Data Analysis

- Compatible with Applied Biosystems real-time PCR instruments and TaqMan® Arrays
- Advanced statistical and data mining tools for differential expression analysis
- Interactive and dynamic visualization
- Ideal for medium- and high-throughput real-time PCR experiments
- Functional annotation of gene lists

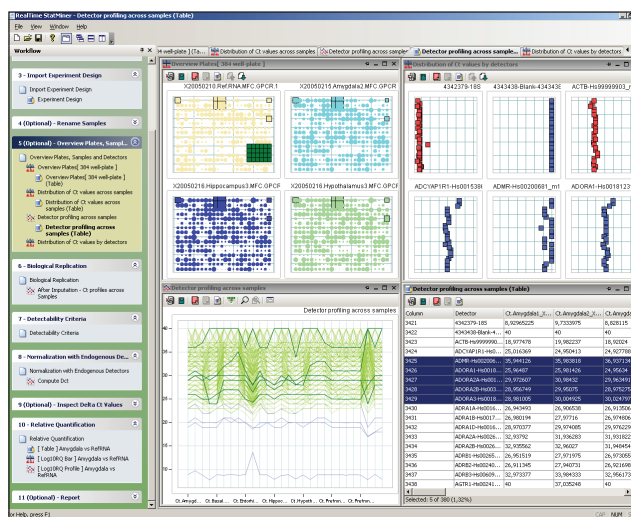


Figure 1. A powerful interactive tool for quality control and differential expression analysis of real-time PCR data designed for Applied Biosystems real-time PCR instruments.

Introduction

The need for high quality and high throughput real-time PCR gene expression analysis is growing, especially in biomarker validation and clinical research. All such studies require sophisticated data analysis methods, capable of managing the complexity of multi-gene, multi-sample projects, and returning reliable biological information.

Real-Time StatMiner™ Software from Integromics is a powerful bioinformatics tool designed to enable deeper insights into real-time PCR data from both TaqMan® Arrays and regular plate format, according to the highest quality standards established for microarray data analysis. The software

provides quantification analysis of gene expression data through a robust combination of interactive visualization, advanced statistics, and data mining tools. Real-Time StatMiner Software, which also enhances quality assurance and quality control, ensures accurate differential expression results from large datasets and multiple endogenous control genes.

With just a single click, the software imports raw, real-time PCR data or C_T values from Applied Biosystems real-time PCR instruments. The researcher is guided at every stage of the analysis process and obtains meaningful and reliable results within minutes.



Step-by-Step Procedure

Real-Time StatMiner Software requires only the following five easy steps:

1. Import raw C_T values from sequence detection system software
2. Perform quality control of biological replicates
3. Select the optimal endogenous control gene for ΔC_T computation
4. Select groups to compare and obtain relative quantification values
5. Apply the latest data mining primitives for further insights

The calculations at each step affect subsequent steps (Figure 2). The user can go back and forth within the workflow, conducting various analysis strategies and comparing results. Real-Time StatMiner Software enables the redesign and customization of the workflow and includes the procedures for each step.

Benefits of Real-Time StatMiner Software

1. Intuitive user interface with a step-by-step guide
2. High-throughput statistics powered by R and Bioconductor algorithms
3. Determination of the most stable endogenous control gene or gene set

4. Execution and interactive visualization of all analytical steps with Spotfire DecisionSite®
5. Statistics for relative quantification of gene expression
6. Extensible, modular approach and flexibility
7. Worldwide technical support

Intuitive User Interface with a Step-by-Step Guide

Real-Time StatMiner Software is available in a stand-alone version and as a plug-in to Spotfire. Both versions include a Help function that explains all software features (Figure 3), and users are guided through the easy workflow. The workflow connects them with state-of-the-art Bioconductor packages for quality control and statistical data analysis without the need for R language-related skills. The Spotfire-linked version also offers powerful graphical and reporting capabilities for statistical analysis.

High-Throughput Statistics Powered by R and Bioconductor Algorithms

Real-time PCR is a powerful tool in expression analysis that provides increased sensitivity, reproducibility, and a large dynamic range. Several variables, however, must still be controlled, including inter-sample variability, precision of technical replicates, RNA quality, and RNA concentration. Therefore, the

experimental reproducibility and systematic variability of the raw gene expression data provided by the instruments (C_T values) must be estimated, and the results must be validated.

State-of-the-art statistics are crucial to ensure quality control and reliable analysis of raw data for filtering outliers and challenging experiments. All statistics in Real-Time StatMiner Software are implemented in the R language, which facilitates the management of large amounts of data. Because R language is also used by Bioconductor (an open source software for bioinformatics, and the largest public repository of genomic analysis data), users of this software can access Bioconductor's various data libraries [1]. For more information, visit www.bioconductor.org and www.r-project.org

Bioconductor Software Project

The Bioconductor software project for the analysis of genomic data began in 2001. The goals of the project include:

- Providing access to a wide range of powerful statistical and graphical methods for analyzing genomic data
- Facilitating the integration of biological metadata in the analysis of experimental data
- Allowing the rapid development of extensible, scalable, and interoperable software

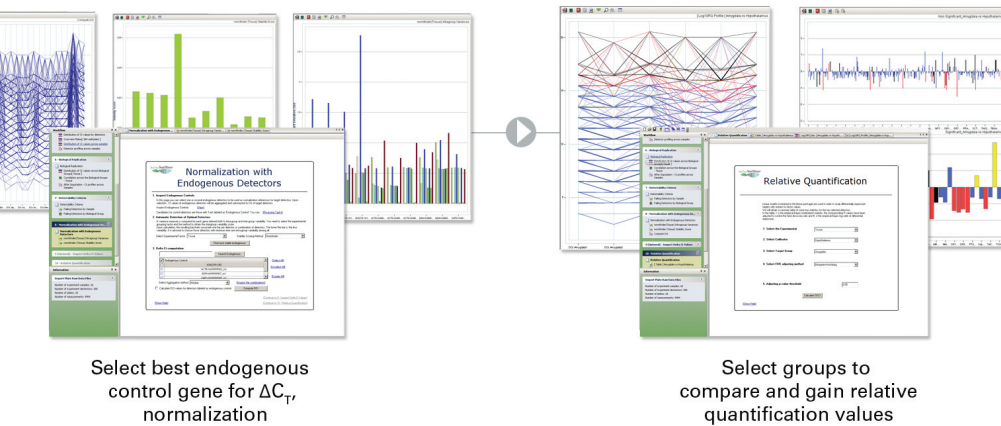


Figure 2. Real-Time StatMiner Workflow: Raw C_T Values → Differential Expression Results. Graphical visualizations facilitate the interpretation of statistically relevant results at each step of the Real-Time StatMiner workflow. Computation is performed in the R-Bioconductor environment. Quality control of sample replicates and the proper selection of endogenous controls are essential for the accurate assessment of gene expression changes between samples.

- Promoting high-quality documentation and reproducible research

Determining the Most Stable Endogenous Control Gene or Gene Set

Internal control genes (also called endogenous controls or housekeeping genes) normalize gene expression levels. Ideally, the expression of a good endogenous control gene does not vary in the tissues or cells under investigation, or in response to the experimental treatment [2]. Despite reports that expression levels of several endogenous control genes vary considerably under certain circumstances, many researchers still incorrectly assume stability for their selected genes [3–6]. Before selecting a candidate gene, it is important to ensure that no significant gene regulation occurs under the experimental conditions.

Real-Time StatMiner Software contains two commonly used mathematical strategies for determining the most stable gene(s) within a panel—the geNorm [7] and NormFinder [8] algorithms. These assessments help users select the appropriate endogenous control gene for normalizing gene expression levels in the experiment. The stability ranking of a candidate gene can be inspected by several methods. After selecting the most appropriate gene(s), the ΔC_T is

computed for each gene and the user can move forward to relative gene expression analyses.

Execution and Interactive Visualization of Analytical Steps with Spotfire DecisionSite®

Real-Time StatMiner Software can be implemented as a plug-in of Spotfire DecisionSite, a leading application for professional and interactive visualization of multidimensional data. Visualizations in DecisionSite facilitate the exploration of large data sets (e.g., the gene expression of hundreds of genes in hundreds of samples) to facilitate interactive queries, and to help enable users to focus on relevant results. After the analysis highlights characteristic expression profiles, integrating the results with other data sources, such as microarray or clinical data, is easy.

Spotfire DecisionSite contains a variety of plots and graphics, including scatter plots, profile charts, heat maps, and pie charts. The visualizations are fully interactive—the user can select genes and mark them to generate detailed views of the data. DecisionSite incorporates other data mining features, such as hierarchical clustering, K-means clustering, principal component analysis (PCA), and profile searches (e.g., to find genes quickly that have a given expression profile). The analysis results can be exported to

$\Delta\Delta C_T$ Method

The formula $2^{-\Delta\Delta C_T}$ is commonly used to calculate the amount of target normalized to an endogenous control gene and relative to a calibrator. The result is an RQ value [14]. The $\Delta\Delta C_T$ value is the CT difference between the normalized amount of sample and the normalized amount of calibrator:

$$\Delta\Delta C_T = \Delta C_{T \text{ Sample}} - \Delta C_{T \text{ Calibrator}}$$

Bayes Method

The Bayes method assumes that ΔC_T values are obtained from two sample populations. Each population comprises biological replicates from a sample group (e.g., treated or control samples); thus, the method makes inter-group comparisons. The output of the t-test is an estimation of the $\Delta\Delta C_T$ that results from comparing the two populations and accounting for biological variability. The robust t-test also produces a p-value (with false positive correction) as an assessment of the fold-change significance. P-values are easy to interpret and can be correlated with p-values obtained from previously performed microarray experiments.

Microsoft® Excel, Word, or PowerPoint, or to a Web page. Users can also store relevant genes in portfolio lists for comparison in subsequent projects. For more information visit www.spotfire.com

Statistics for Relative Quantification of Gene Expression

The $\Delta\Delta C_T$ method is commonly used to calculate the expression of a target gene relative to a calibrator that is normalized to an endogenous control gene [9]. (See sidebar, *$\Delta\Delta C_T$ Method*, for more details.) However, this allows only direct sample-to-sample comparisons (target vs. calibrator). The capability to group several samples according to biological criteria

(same treatment, tissue, dose, etc.) enables comparisons between groups.

To assess confidence in the fold-change value, the variability of ΔC_T values for biological replicates must be propagated to the relative quantity (RQ) value. This error propagation can be computed by various formulas and is sometimes difficult to interpret. Real-Time StatMiner Software incorporates the empirical Bayes method to moderate the standard errors of the estimated $\Delta\Delta C_T$ (see sidebar, *Bayes Method*). The result is more stable inference and improved power, especially for experiments with fewer samples [10].

Extensible, Modular Approach and Flexibility

Real-time PCR is a very powerful technique that can be applied to miRNA expression profiling, biomarker discovery, and gene copy number variation analysis. Real-Time StatMiner Software implements analytical workflows—sequential steps that guide the user from importing raw data to final data analysis—which can be easily adapted to specific user needs or experimental settings. This feature facilitates the use of the software by novices in biostatistics. On the other hand, experienced statisticians can modify the procedures or add functionality by taking advantage of the most referenced algorithms in Bioconductor.

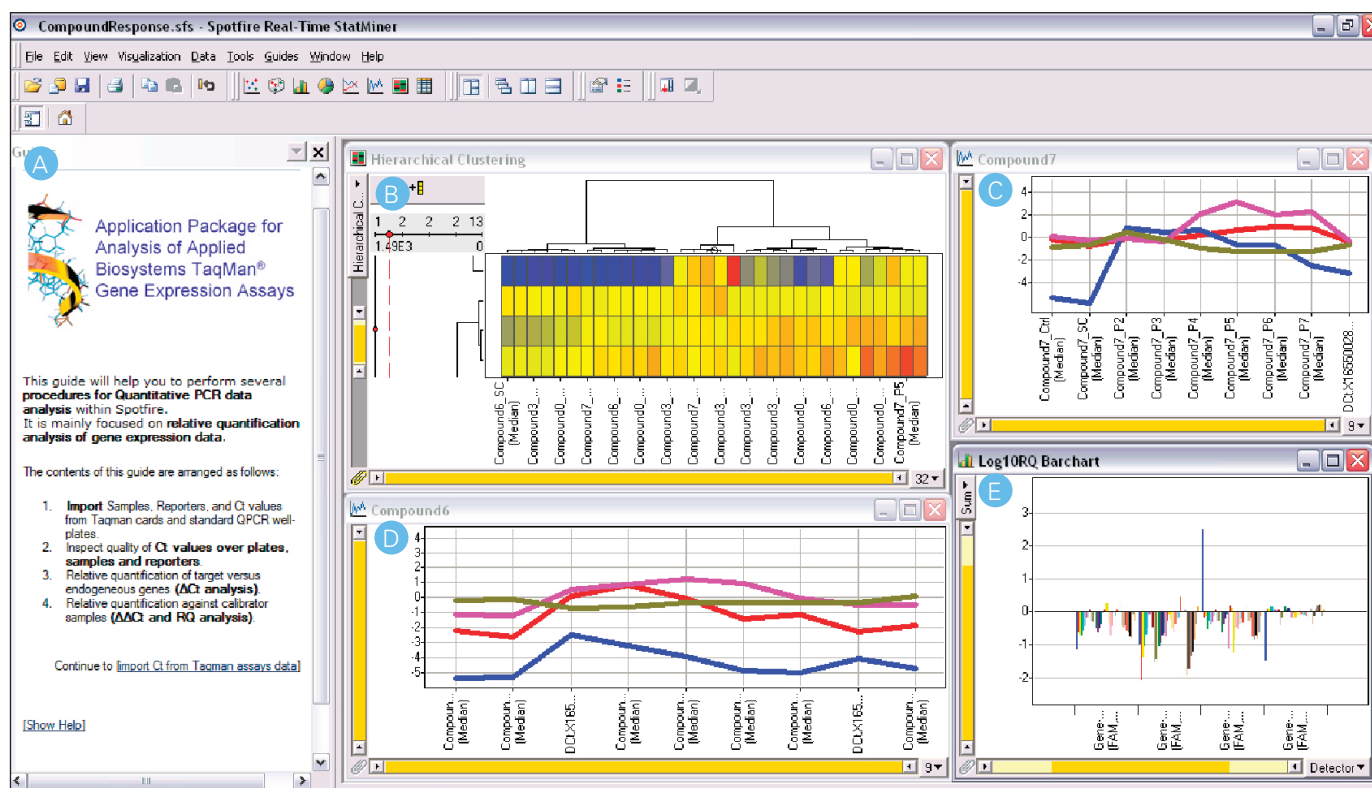


Figure 3. Real-Time StatMiner Software User Interface. Panel A shows the navigator frame that guides the user step-by-step through the application (an optional Help function is available), and enables the non-expert to run data analysis within minutes. Interactive visualizations are displayed in the central area (Panels B and D). In this dataset, changes in gene expression, resulting from various compound treatments, were investigated in cells. Genes that show similar expression profiles are grouped by hierarchical clustering (Panel B). The relative expression values of several genes in response to various compounds can be displayed as profiles (Panel C, D) or bar charts (Panel E).

Worldwide Technical Support

Unlike existing freeware, Integromics™ provides worldwide support for customers of Real-Time StatMiner Software. For more information, contact: **technical.support@integromics.com**

Scientific Contributors

Astrid Ferlinz, Alexander Sartori, and Jihad Skaf at Applied Biosystems, Applera Deutschland GmbH, Darmstadt, Germany, and Foster City, USA.

References

1. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics *Genome Biol* **(5)**:10:R80.
2. Application Note: Using TaqMan® Endogenous Control Assays to Select an Endogenous Control for Experimental Studies, Applied Biosystems 127AP08-01 (http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_042279.pdf)
3. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E (1999) Housekeeping genes as internal standards: use and limits. *J Biotechnol* **75**:291–295.
4. Suzuki T, Higgins PJ, Crawford DR (2000) Control selection for RNA quantitation. *Biotechniques* **29**:332–337.
5. Bustin SA (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* **25**:169–193.
6. Huggett J, Dheda K, Bustin S, Zumla A (2005) Real-time RT-PCR normalization; strategies and considerations. *Genes Immun* **6(4)**:279–284.
7. Vandesompele J, De Preter K and Pattyn F et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3(7)**:1–12.
8. Andersen CL, Ledet-Jensen J, Ørntoft T (2004) Normalization of real-time quantitative RT-PCR data: a model based variance estimation approach to identify genes suited for normalization - applied to bladder and colon-cancer data-sets. *Cancer Res* **64**:5245–5250.
9. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method. *Methods* **25(4)**:402–408.
10. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3(1)**: Article 3.

Product Literature

For more information on Real-Time StatMiner Software, visit <http://www.integromics.com/StatMiner.php>

For an example of how Real-Time StatMiner Software is used in the laboratory, please attend the recorded webinar, “RT-PCR Data Analysis with Integromics Real-Time StatMiner Powered by Spotfire,” available at www.Integromics.com

Ordering Information

For ordering information and pricing,
please contact Integromics at
sales@integromics.com

Description	P/N
Real-Time StatMiner™ License (stand-alone; commercial use)	4398086
Real-Time StatMiner™ License (stand-alone; academic use)	4398087
Real-Time StatMiner™ License (Spotfire-based*; commercial use)	4397503
Real-Time StatMiner™ License (Spotfire-based*; academic use)	4397504

* For existing Spotfire customers

Note: The above are all 3-year licenses.

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

NOTICE TO PURCHASER: LIMITED LICENSE

Applied Biosystems, AB (Design), and Applera are registered trademarks of Applera Corporation or its subsidiaries in the US and/or certain other countries.

DecisionSite and Spotfire are registered trademarks of Spotfire, Inc.

StatMiner and Integromics are trademarks of Integromics, SL.

TaqMan is a registered trademark of Roche Molecular Systems, Inc.

All other trademarks are the sole property of their respective owners.

©2007 Applied Biosystems. All rights reserved.

Printed in the USA, 12/2007 Publication: 117DA04-01



Headquarters

850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.345.5224
www.appliedbiosystems.com

International Sales

For our office locations please call the division
headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm