DataAssist™ – Data Analysis Software for TaqMan® Real-Time PCR Data

Matt Xia, Jon Sherlock, Patricia Hegerich, Xiaoqing You, Kathy Lee, Criss Walworth, and Eugene Spier

***Abstract*—A data analysis software, DataAssist™ software, has been developed for quick analysis and interactive visualization of TaqMan® real-time PCR data. It uses the comparative C_T method (also known as the $2^{-\Delta\Delta Ct}$ method) [1] to calculate relative quantities (RQ) of gene expression for sample comparison. The software uses a refined Grubbs' outlier test to remove outlier among technical replicates, provides a metric to measure control gene stability based on the geNorm algorithm [2] to assist with endogenous control selection, and allows using single or multiple control genes for data normalization. The software provides a function-rich graphic user interface (GUI), many content-rich tables and scalable graphics charts for easy, interactive, and rapid high-throughput data analysis and visualization.**

***Index Terms*—Comparative C_T method, Gene Expression, RT-PCR, TaqMan®.**

I. INTRODUCTION

Real-time RT-PCR is widely used to quantify gene expression levels by measuring the threshold cycle (C_T), an arbitrarily placed threshold which ensures the PCR is in the exponential phase of amplification. The C_T is reversely related to the amount of target molecules in the reaction. The classic comparative C_T method can be used to calculate the expression level of the gene of interest relative to a calibrator or reference sample using the C_T data [1].

Applied Biosystems provides a large collection of TaqMan® gene expression assays that are widely used for quantitative gene expression study. We have developed a data analysis tool, DataAssist™ Software, to quickly analyze and visualize the experiment data (C_T) generated by Applied Biosystems real-time PCR instruments, especially with TaqMan® Gene Expression Assays, TaqMan® Array Plates, or TaqMan® Array 384-Well Micro Fluidic Cards.

DataAssist™ Software is a simple, yet powerful data analysis tool for sample comparison. It uses the comparative C_T method to calculate relative quantity of gene expression.

Manuscript received December 21, 2009.

Matt Xia is with Applied Biosystems Inc., 850 Lincoln Center Drive, Foster City, California, CA 94404, USA (phone: 650-638-6456; e-mail: matt.xia@lifetech.com).

J. Sherlock, P. Hegerich, X. You, K. Lee, C. Walworth, and E. Spier are with Applied Biosystems Inc., 850 Lincoln Center Drive, Foster City, California, CA 94404, USA.

First it filters outliers among technical replicates using a refined Grubbs' test, and then normalizes the C_T data using single or multiple endogenous control genes:

$$\Delta C_T = C_T \text{ gene of interest} - \text{Normalization Factor} \quad (1)$$

Normalization Factor is the arithmetic mean or geometric mean of C_T values of the selected control genes. If multiple genes are selected as controls, a gene stability measure is also calculated based on the geNorm algorithm to assist with selecting most stable control genes for data normalization [2]. The normalized ΔC_T data are used to calculate the relative gene expression fold change using a selected calibrator (reference sample):

$$\Delta\Delta C_T = \Delta C_T \text{ sample A} - \Delta C_T \text{ calibrator} \quad (2)$$

$$\text{Fold Change} = 2^{-\Delta\Delta Ct} \quad (3)$$

The fold change can also be calculated between sample groups of biological replicates, by grouping samples to biological replicates, the mean $2^{-\Delta Ct}$ of the biological replicates is used to determine the expression fold change [1]:

$$\text{Fold Change} = 2^{-\Delta Ct} \text{ group A} / 2^{-\Delta Ct} \text{ reference} \quad (4)$$

Statistical analysis is performed to provide standard deviations for gene expression comparison between samples, and p-value from t-test for comparison between biological groups.

II. METHODS AND RESULTS

DataAssist™ software was developed using Java as a standalone desktop application for Windows XP and Vista® operating systems. Java Swing was used to implement the Graphic User Interface (GUI), and the open-source Java chart library JFreeChart [3] was used to implement most charts for data visualization. The software installer for Windows was created using open source tool Nullsoft Scriptable Install System (NSIS) [4]. DataAssist™ Software is freely available at http://www.appliedbiosystems.com/DataAssist.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

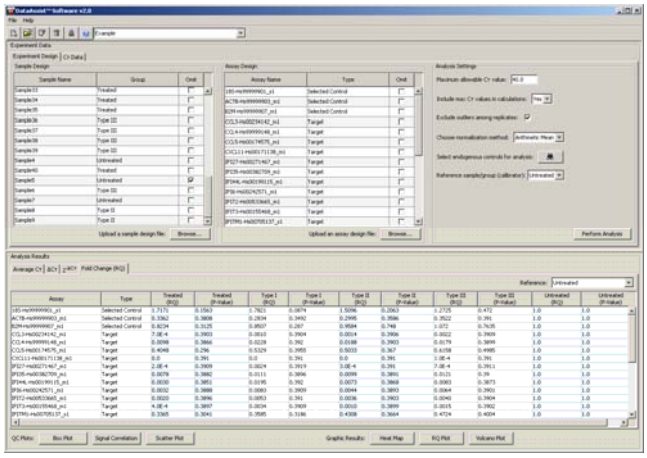

Fig. 1 DataAssist™ software main window. It allows quick and interactive experiment setup including data analysis settings and sample grouping to biological replicates.

DataAssist™ software provides a function-rich GUI for easy data importation, experiment setup, and interactive, high-throughput data analysis (Fig. 1). The calculation in DataAssist™ software is very rapid and the results are provided in content-rich tables and scalable graphics charts that can be easily exported (Fig. 2 - 9).

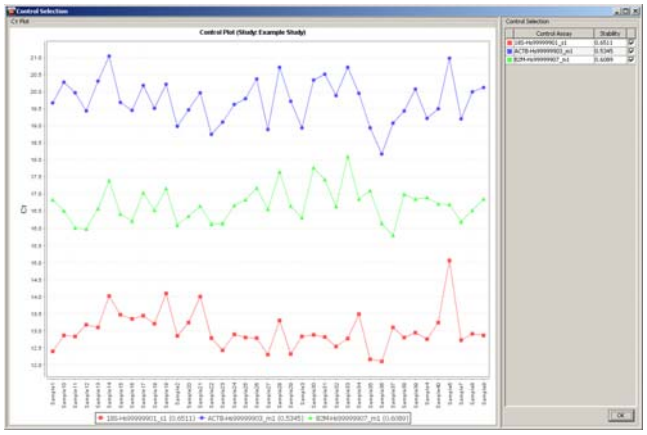

Fig. 2 Control Selection Plot and Table. The plot displays C_T values of control genes for all samples, which gives a quick overview of the expression profile of each control gene. The gene stability measure [2] is shown in the adjacent table to assist with selecting the most stable control genes for data normalization.

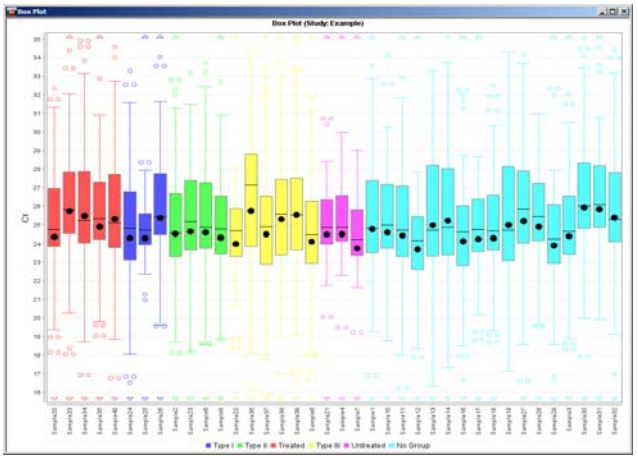

Fig. 3 Box-and-whisker Plot. It displays the overall range of C_T distribution for each sample from all genes in the experiment. The bar is colored based on the sample biological group if samples are grouped to biological replicates.

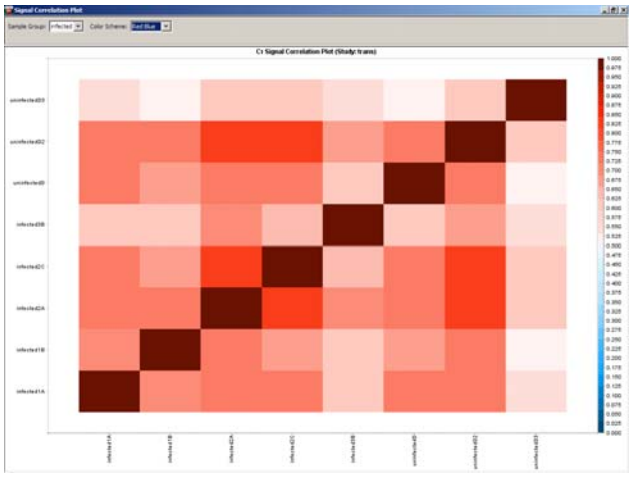

Fig. 4 C_T Signal Correlation Plot. The plot displays C_T signal correlation between samples in a selected biological group. Pearson's product moment correlation coefficient (r) is calculated for each pair of samples and displayed as a color box either in red-blue or red-green color map.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

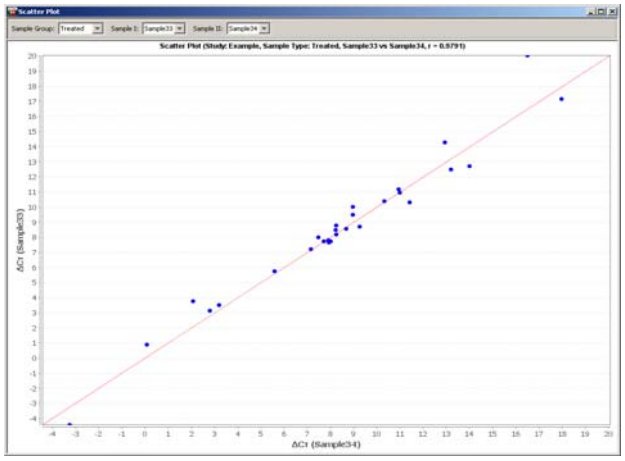

Fig. 5 Scatter Plot. It shows ΔC_T correlation between any two selected samples in a chosen biological group. Pearson's product moment correlation coefficient (r) is also calculated and included in the plot.

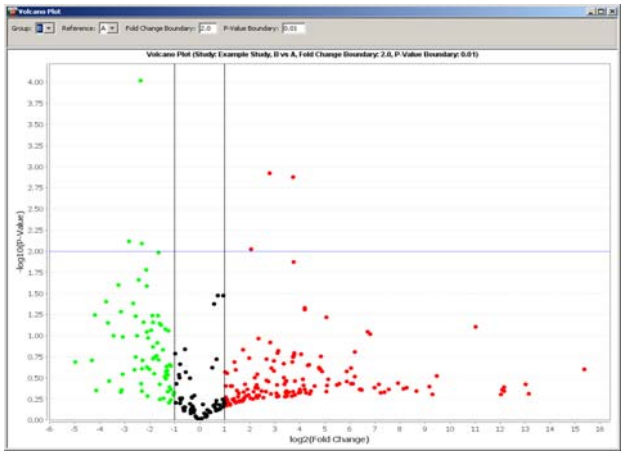

Fig. 6 Volcano Plot. It displays the fold change versus the p-value from t-test for comparison between sample groups, which gives a quick overview of the statistical significance of fold changes for all genes in the experiment. The fold change and p-value boundary can be adjusted to rearrange the genes in the plot.

Fig. 7 RQ Plot. It shows the RQ (fold change) versus Target (gene) or RQ versus Sample, as Linear, Log_{10}, or Log_2 scale. The standard deviation is also displayed as error bar for each sample on log_2 scale when no biological group is specified.

Fig. 8 Dendrogram and Heat Map. It displays the dendrograms of both sample cluster and gene cluster along with gene expression heat map. Genes and samples are clustered using hierarchical clustering [5] with average linkage, complete linkage or single linkage method. The normalized gene expression data (ΔC_T) are used to calculate the distances between samples and genes using either Pearson's correlation coefficient or Euclidean distance. The heat map can be configured as either red-blue or red-green map, with red color box representing up-regulated gene expression level, and the middle expression level can be set using the adjustable color scale on the right side.

REFERENCES

[1] Schmittgen T D, Livak K J, "Analyzing real-time PCR data by the comparative CT method" Nature Protocols 3, - 1101 - 1108 (2008).

[2] Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F, "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes" Genome Biol 2002, 3(7).

[3] JFreeChart. Available: http://www.jfree.org/jfreechart/

[4] Nullsoft Scriptable Install System (NSIS). Available: http://nsis.sourceforge.net/

[5] Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D., "Cluster analysis and display of genome-wide expression patterns" Proc. Natl. Acad. Sci. USA, vol. 95, Dec. 1998, pp. 14863–14868.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)