Relative Expression Software Tool 2008

Greater Certainty in Expression Studies

By Corbett Research and M.Pfaffl (Technical University Munich)

A New Stand-alone Software for Gene Expression Analysis

REST 2008

Published: April 2008

Contents

Abstract

REST (Relative Expression Software Tool) 2008 is a standalone software tool to estimate up and down regulation for gene expression studies. The software addresses issues surrounding the measurement of uncertainty in expression ratios by using randomization and boot strapping techniques. By increasing the number of iterations from 2,000 to 50,000, hypothesis tests achieve a level of consistency on par with traditional statistical tests. New confidence intervals for expression levels also allow scientists to measure not only statistical significance of deviations, but also their likely magnitude, even in the presence of outliers. Graphical output of the data via a whisker box-plots provide a visual representation of variation for each gene that highlights potential issues such as a distribution skew.

Why REST?

Prior to REST [1], relative quantitation in qPCR was a technique which allowed the estimation of gene expression. While useful, it did not provide statistical information suitable for comparing groups of treated versus untreated samples in a robust manner.

To illustrate with an example, let us say we are testing to see if a particular mRNA is responsible for sending pain messages. We split up our patients into 2 groups; one which will be subjected to pain (such as immersion of the hand into ice water), and the other, which is our control group. Following this, we measure the quantities of targeted mRNA in both groups, relative to reference genes. Our question is: did the group subjected to pain release more mRNA than the control group?

Prior approaches are unable to properly answer this question. They may calculate an average expression value indicating whether a particular subject in one group appeared to release more or less target mRNA than another subject, but without any statistical test to determine accuracy. Due to the use of ratios in gene expression, it becomes very complex to perform traditional statistical analysis as ratio distributions do not have a standard deviation. REST 2008 overcomes these problems by using simple statistical randomization tests. Such tests may appear counter-intuitive and so it is recommended to read the discussions on randomization techniques in the 'Links' before continuing.

Note

Throughout the use of this software and manual the terms C_T (Cycle Threshold) and CP (Crossing Point) are interchangeable.

New for 2008

- **REST RGMode (See page 16)**
  A new method of input has been introduced, allowing users to copy and paste results from the Rotor-Gene software's Comparative Quantitation analysis. This is an alternative to importing standard curve and C_T results.

- **Whisker-Box Plots Exportable**
  Whisker-Box plots can now be exported by right-clicking on the graph.

- **Improved Randomisation**
  Improvements to the randomisation algorithms have been made, making confidence intervals much tighter, and p-values more accurate.

- **Handling of standard curve variation**
  REST 2008 no longer takes into account the variation of the standard curve, due to a bug which caused unnecessary widening of confidence intervals and p-values. In previous version the software would randomly pick two points from the standard curve, and calculate an efficiency based on that. However there is a situation when two points are chosen that lie close to each other on the standard curve, this can cause a bogus efficiency which adds unnecessary outliers to the random distribution. We now calculate the efficiency by determining the line of best fit for the standard curve, this efficiency is used through the randomisation process.

Reference Gene Normalisation

REST 2008 is more comprehensive than prior techniques as it takes into consideration multiple reference genes when determining expression. When estimating a sample's expression ratio, an intermediate absolute concentration value is calculated according to the following formula:

$$\text{Concentration} = \text{efficiency}^{\text{avg(controls)-avg(samples)}}$$

This formula is used to obtain mean estimates of the uncorrected absolute concentration for each gene. For a single reference gene, the gene of interest's concentration is divided by the reference gene value to obtain an expression level, as is done in the Two Standard Curve technique:

$$\text{Expression} = \text{GOIConcentration} \div \text{REF concentration}$$

For multiple reference genes, the geometric mean is taken of all the gene concentrations, since concentration estimates vary exponentially*:

$$\text{Expression} = \text{GOIConcentration} \div \text{GEOMEAN}(\text{REFConc}_1, \text{REFConc}_2, \ldots)$$

Another way to think of normalization to multiple reference genes is that the individual expressions calculated relative to each reference gene represent alternative approximations of the true expression values. To take all into account simultaneously, they are averaged using the geometric mean (since ratios are being used):

$$\text{Expression} = \text{GEOMEAN}(\text{GOIConcentration} \div \text{REFConc}_1, \text{GOIConcentration} \div \text{REFConc}_2, \ldots)$$

Since the mean concentrations of each gene do not change, they can be calculated at the beginning of the algorithm and expressed as a single value, called the '*normalisation factor*', equal to their geometric mean.[2]

** Errors in calculation of concentration occur due to linear variation in C_T values. Estimates of concentration use an equation of the form $c = A*e^{CT}$ (Where $A*e$ is the efficiency), and so vary exponentially.*

Expression Level Confidence Intervals

Previous versions of REST provide a means of determining the mean output and a P value for the likelihood of up or down regulation using a hypothesis test. Bootstrapping techniques [3] can be used to provide 95% confidence intervals for expression ratios, without normality or symmetrical distribution assumptions. While a hypothesis test provides a measure of whether there was a statistically significant result, the confidence

interval provides a range that can be checked for semantic significance. For example: Drinking cough medicine before driving may increase the chances of an accident by 1 x 10^{-6} %. While a statistical test may show the difference to be significant, it clearly poses no real threat to drivers,, when taking into account the average number of accidents a driver has in their lifetime.

Procedure

We are given a set of control (C_{GOI}) and sample (S_{GOI}) C_T values for the gene of interest and similarly a set of controls (C_{REF}) and sample (S_{REF}) for the reference gene. We are also given an efficiency value (e_{GOI}) for the gene of interest (GOI) and a efficiency value (e_{REF}) for the reference gene.

Let *X* be the random variable indicating the expression ratio of individual samples for the gene of interest.
Let *Y* be a list of simulated readings from *X*.
Let *n* be the size of *Y*, preferably a large value (>2000).
Let *choose()* be a function that returns a random element from a set.
Let *count()* be a function that returns the number of elements in a set.

We populate Y by randomly pairing controls and samples from the GOI and the reference gene (REF), and calculating their expression ratio:

$i \in \{1, \ldots, n\}$
$j = choose(\{1, \ldots, count(C_{GOI})\})$
$k = choose(\{1, \ldots, count(S_{GOI})\})$

We assume $count(C_{GOI}) = count(C_{REF})$ and $count(S_{GOI}) = count(S_{REF})$, since every GOI C_T must have a corresponding REF C_T.

$$Y_i = \frac{e_{GOI}^{\,C_{GOI,j} - S_{GOI,j}}}{e_{REF}^{\,C_{REF,j} - S_{REF,j}}}$$

where Y_i is a single element in the set of *Y*.

To determine confidence intervals, sort the population Y into increasing order:

$Y_{sorted} = sort(Y)$

The 95% confidence interval is defined as:

$\alpha = 0.05$

$min = Y_{sorted,\ n \times (\alpha / 2)}$
$max = Y_{sorted,\ n \times (1-\alpha/ 2)}$

Other confidence intervals can be obtained by varying a. The median of the set provides an alternative measurement of the expression ratio given by working with mean control and sample values:

$median = Y_{sorted,\ 0.5 \times n}$

Example

Say we are given the following samples, with IGF-1 our gene of interest, and GAPDH the reference gene.

Gene of Interest is IGF-1
Reference Gene is GAPDH
Efficiency = 1.01
RefEfficiency = 0.97

Index	GAPDH Control (GC)	GAPDH Sample (GS)	IGF -1 Control (IC)	IGF Sample (IS)
1	26.74	26.77	27.57	24.54
2	26.85	26.47	27.61	24.95
3	26.83	27.03	27.82	24.57
4	26.68	26.92	27.12	24.63
5	27.39	26.97	27.76	24.66
6	27.03	26.97	27.74	24.89
7	26.78	26.07	26.91	24.71
8	27.32	26.3	27.49	24.9
9		26.14		24.26
10		26.81		24.44

Randomising for a small n=10, produces the following Y:

j	k	REF_C	GOI_C	REF_S	GOI_S	Expression
6	10	27.03	27.74	26.81	24.44	8.625105575
7	8	26.78	26.91	26.3	24.9	2.938192778
1	2	26.74	27.57	26.47	24.95	5.186421266
3	1	26.83	27.82	26.77	24.54	9.480147506
6	6	27.03	27.74	26.97	24.89	7.021676066
1	7	26.74	27.57	26.07	24.71	4.675718457
6	2	27.03	27.74	26.47	24.95	4.797510275
1	2	26.74	27.57	26.47	24.95	5.186421266
1	2	26.74	27.57	26.47	24.95	5.186421266
8	6	27.32	27.49	26.97	24.89	4.844473339

Sorting yields Y_{sorted} :

Expression
2.938192778
4.675718457
4.797510275
4.844473339
5.186421266
5.186421266
5.186421266
7.021676066
8.625105575
9.480147506

To obtain a 68% confidence interval (α = 0.32), equivalent to ONE standard error interval, we examine the readings at indices 1 (~=(α/2) * (10-1)) and 8 (~=(1-α/2) * (10-1)).

$confidence_{68\%}$ = [4.675718457, 8.625105575]

For a 95% confidence interval (a = 0.05), equivalent to TWO standard error intervals, we examine the readings at indices 0 (~=(α/2) * (10-1)) and 9 (~=(1-α/2) * (10-1)).

$confidence_{95\%}$ = [2.938192778, 9.480147506]
p<0.05

With the small example, the 99.7% confidence interval (α = 0.0027) leads to the same indices 0 and 9 due to a lack of data points, leading to an identical confidence interval:

$confidence_{99.7\%}$ = [2.938192778, 9.480147506]
p<0.0027

The median is calculated as the 5th position:

median = 5.186421266

NB: While the median of even sets is traditionally taken as the average of the middle two positions, this introduces assumptions of normality on the underlying distribution. Theoretical objections can be sidestepped by always using sets that provide critical points (α = 0.5, α = 0.05, α = 0.95) at integral indices. The issue does not have a practical bearing on results, since variation between adjacent values is dominated by the effects of randomisation.

Validation of the Number of Randomisations Used

A sample data tested on a larger randomisation value (n=10000) gives the following values:

confidence$_{68\%}$ = [4.121081159, 8.62510557506084]
confidence$_{95\%}$ = [2.9840236231636, 9.98446532616807]
median = 5.95072937164207

There was insufficient data to reliably calculate a 99.7% confidence interval.

For the same data set, REST calculated comparable values:

Expression = 5.927
Confidence 95% = [2.983, 9.996]
Sample up-regulated = YES (p = 0.000)

As all values in the 95% confidence interval were greater than 1, the interval is consistent with the REST P value of 0.000. The median is slightly inaccurate relative to the calculated expression, due to problems of resolution caused by permutation over a set of fixed values. While the median should therefore not be used to determine the mean expression value, it provides a useful cross-check of the confidence interval, as it is generated from the same data set. The 68% confidence interval covers roughly the same area as the standard error, but still retains a valid meaning when expanded to 95%, whereas traditional statistical methods of estimating standard error fall into negative values.

Hypothesis Test

The purpose of REST 2008 is to determine whether there is significant difference between samples and controls, while taking into account issues of reaction efficiency and reference gene normalisation. Because the normalization and efficiency calculations involve ratios and multiple sources of error, it would be extremely difficult to devise a traditional statistical test, so randomization techniques are employed.

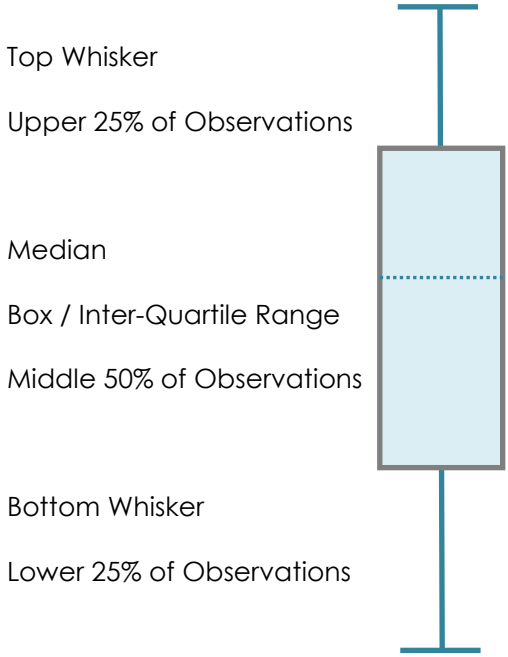
The hypothesis test *P(H1)* indicated in the results table, represents the probability of the alternate hypothesis that the difference between the sample and control groups is due only to chance. To devise a strong randomization test, we use the following randomization scenario: "if any perceived variation between samples and controls is due only to chance, then we could randomly swap values between the 2 groups and not see any greater difference than what we see between the initial groups."

The hypothesis test performs 50,000 random reallocations of samples and controls between the groups, and counts the number of times the relative expression on the randomly assigned group is greater than the sample data.

Whisker-Box Plots

In statistical applications whisker-box plots provide additional information about the skew of the data distributions that would not be available simply by plotting the sample mean. See Link (5) for further information about whisker-box plots.

To summarise, the box area in a whisker-box plot encompasses 50% of all observations, the dotted line represents the sample median and the whisker represent the outer 50% of observations as shown:



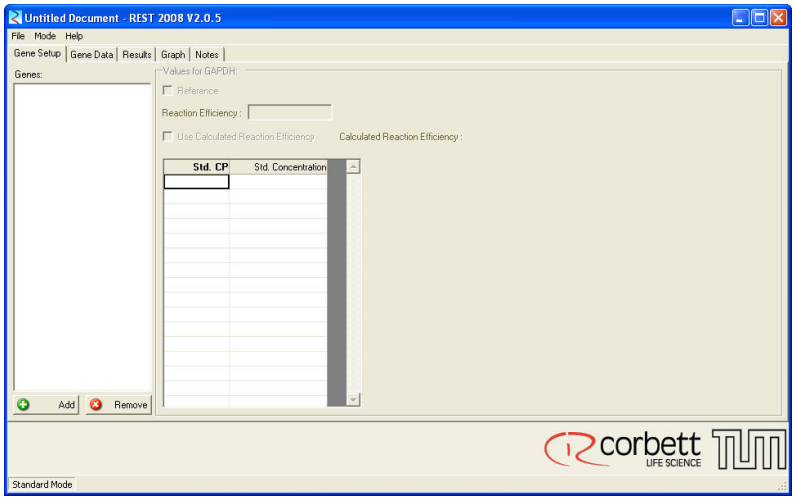
If the sample data is skewed or non-linear the tails of the data may be asymmetrical.

Because REST uses randomization techniques, it draws whicker-box plots based upon the permutated expression data (Y set) rather than the raw C_T values input by the user.

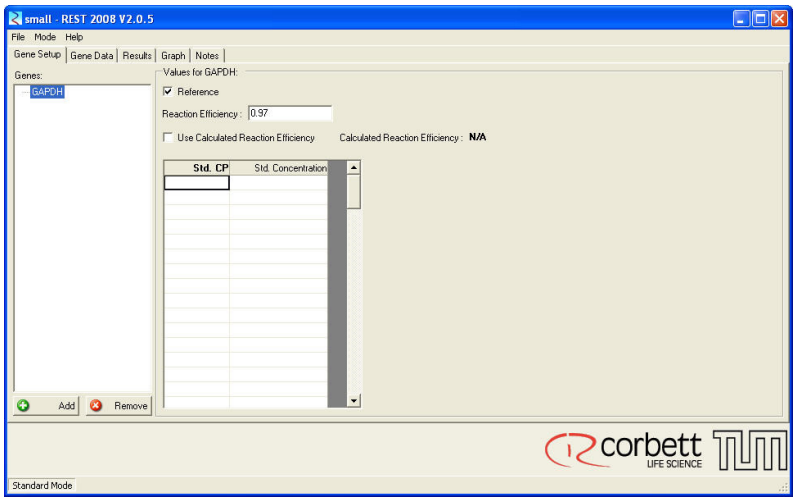
Because expression level values are ratios they will often have lopsided ratios, with greater variability on the upper tail. As ratio populations can be unpredictable, and subject to large and unseen variability, this visualization draws out characteristics of gene expression data that may otherwise go unnoticed.

User Guide

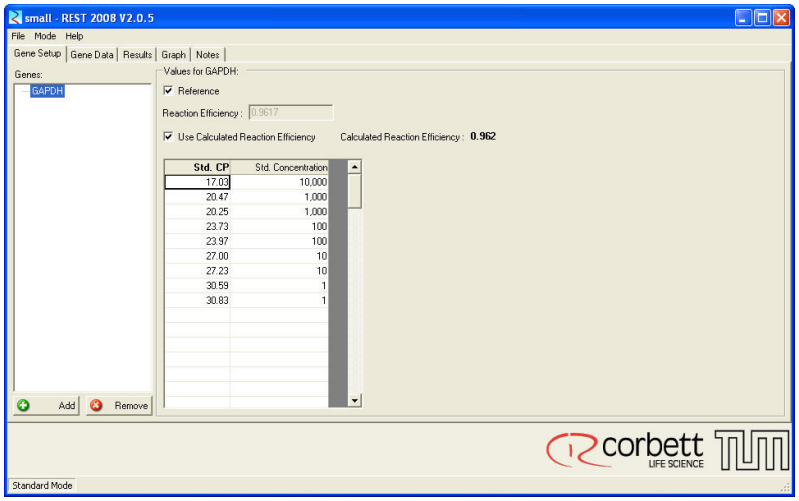
1. New format software makes data input easier. By using the 'add' button at the bottom of the window new genes can be added.



2. Once a new gene has been added to the list, use the right window to input the reaction efficiency of the data. Check the box if the gene is to be used as a reference. Input a reaction efficiency calculated independently of REST2008 for example Rotor-Gene 6000 software, Quantitation Analysis.

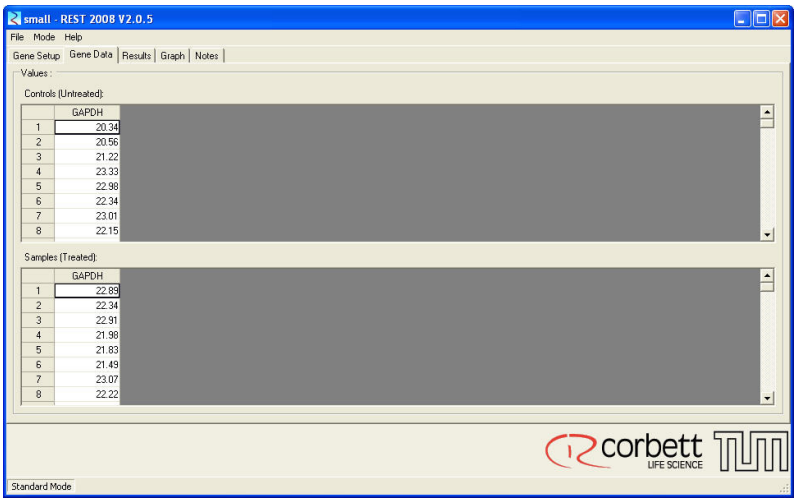


3. Or provide the CP(Crossing Point), calculated as the C_T (Cycle Threshold) in the Rotor-Gene 6000 software and concentrations of a standard curve to enable REST 2008 to calculate the reaction efficiency for that gene. The reaction efficiency does not

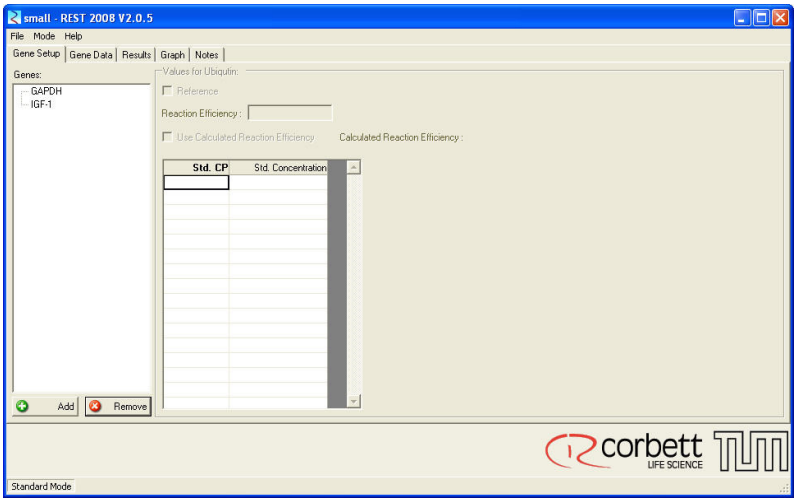


need to be performed every run, however the REST analysis tool does account for differences in reaction efficiency and therefore it must be determined for each gene product.

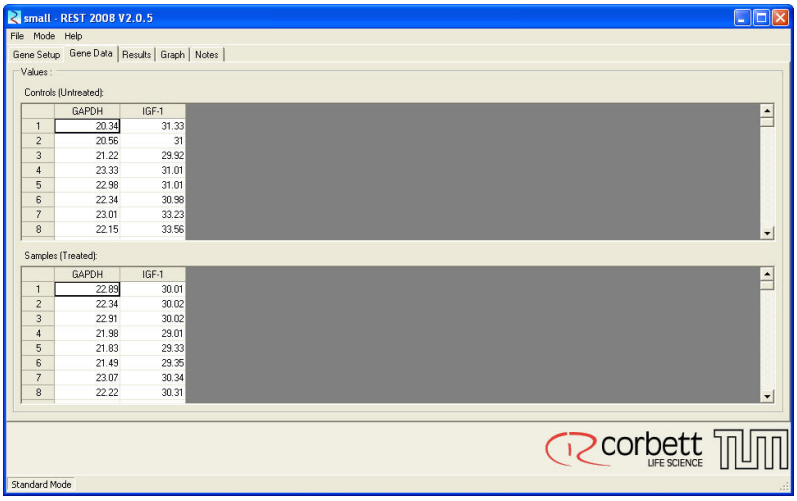
4. Move to the 'Gene Data' tab and enter the CP(C_T) of the samples into the appropriate column. The controls (untreated) should be at the top. Samples (treated) should be entered in the bottom window.



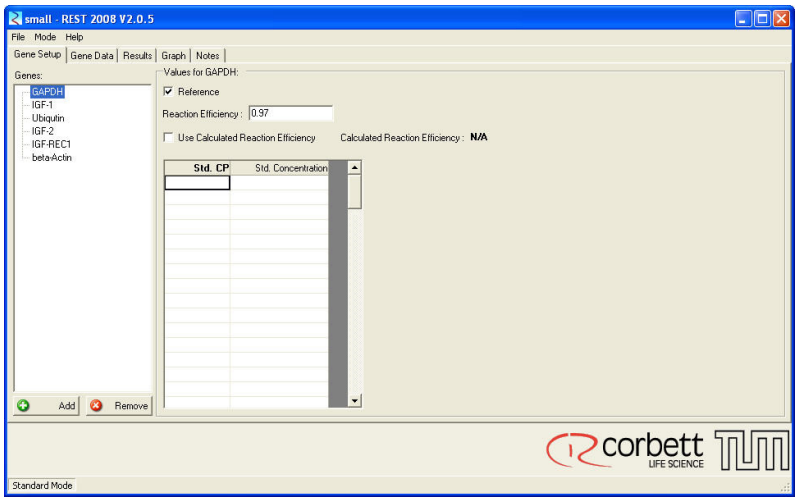
5. Enter the data for all the genes which is being analysed.



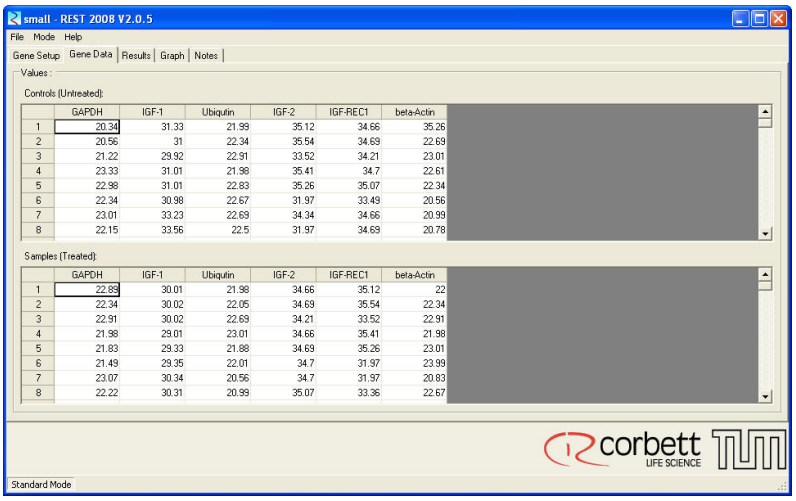
6. Enter the CP(C_T) into the table with corresponding samples in the same rows.



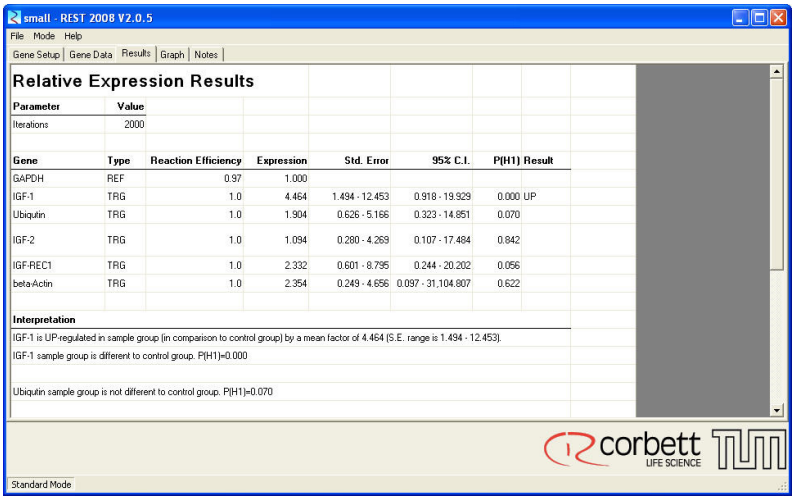
7. Add as many genes as required.



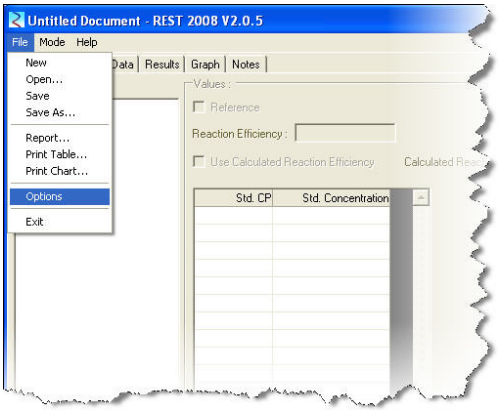
8. Enter the CP(C_T) for all the genes listed.



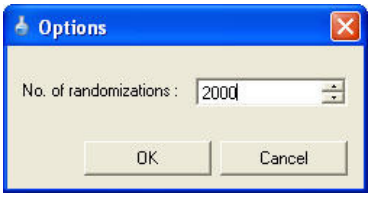
9. Move to the 'Results' Tab and the generated relative expression result will be displayed. The number of randomizations or iterations is shown at the top and can be increased to >50,000.



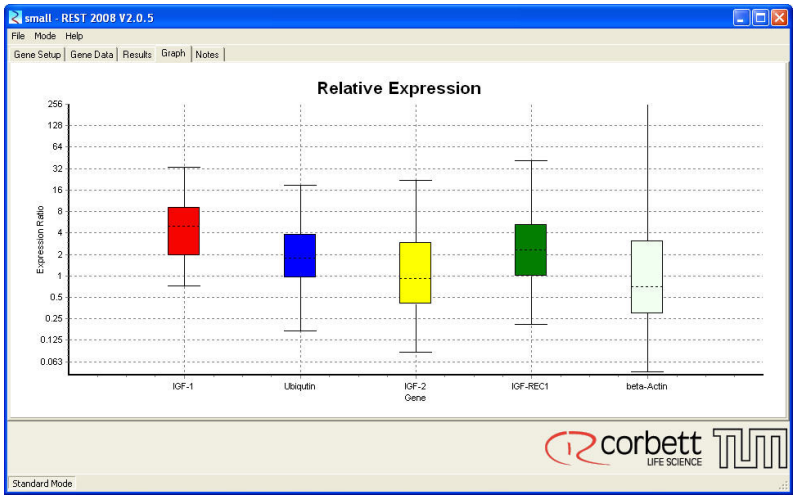
10. To increase the number of randomizations use the ‘File’ and ‘Options’.



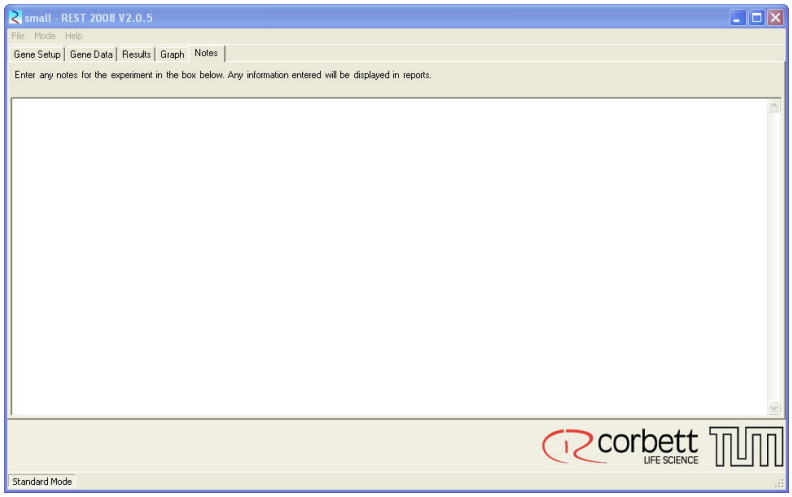
11. The number of randomizations can be increased for large data sets, however increasing can substantially slow the output of the software.



12. The ‘Graph’ tab illustrates the relative expression results data in whisker-box plots.



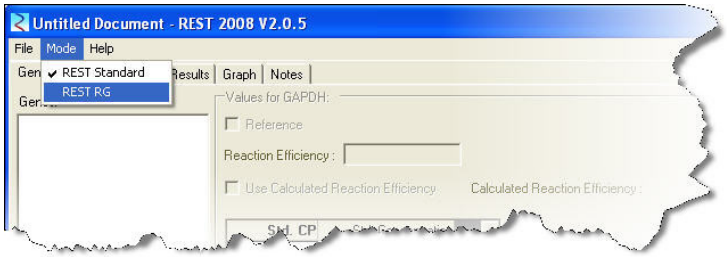
13. The ‘Notes’ tab enable any freehand notes to be made about the results, data or source.



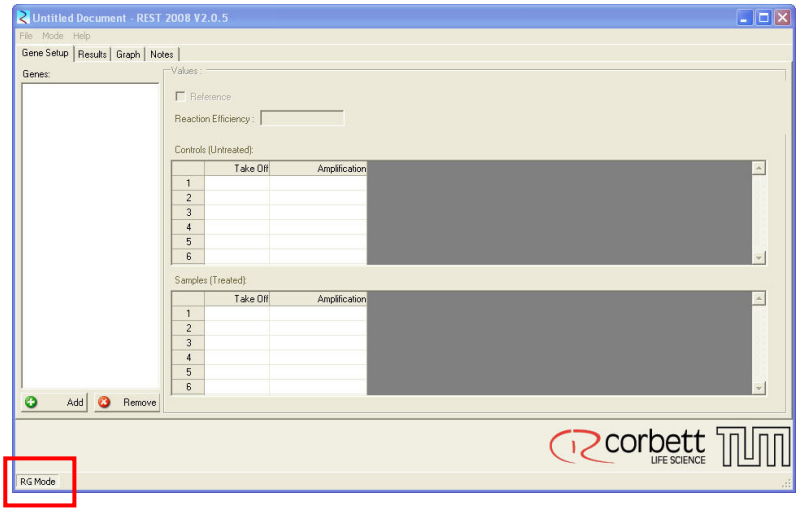
REST RG-Mode

With this latest release of REST 2008 we have included a special "RGMode" which facilitates the use of the REST with data generated from the comparative quantitation analysis tool within the Rotor-Gene software.

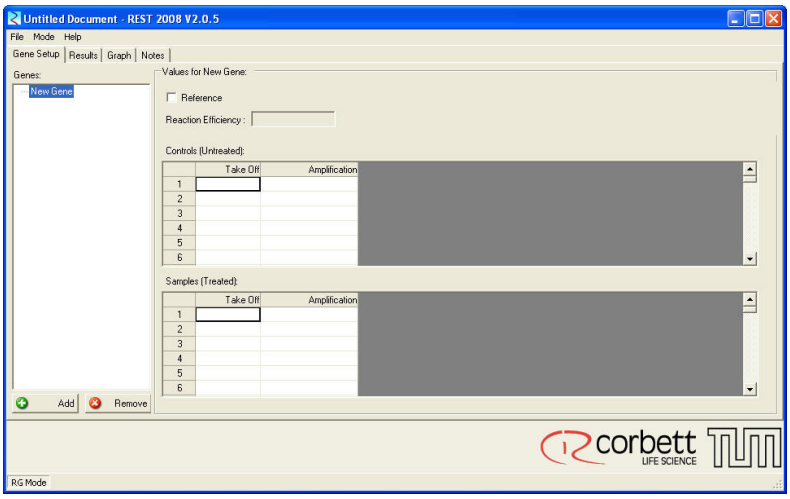
1. In the Mode menu change to "REST RG"



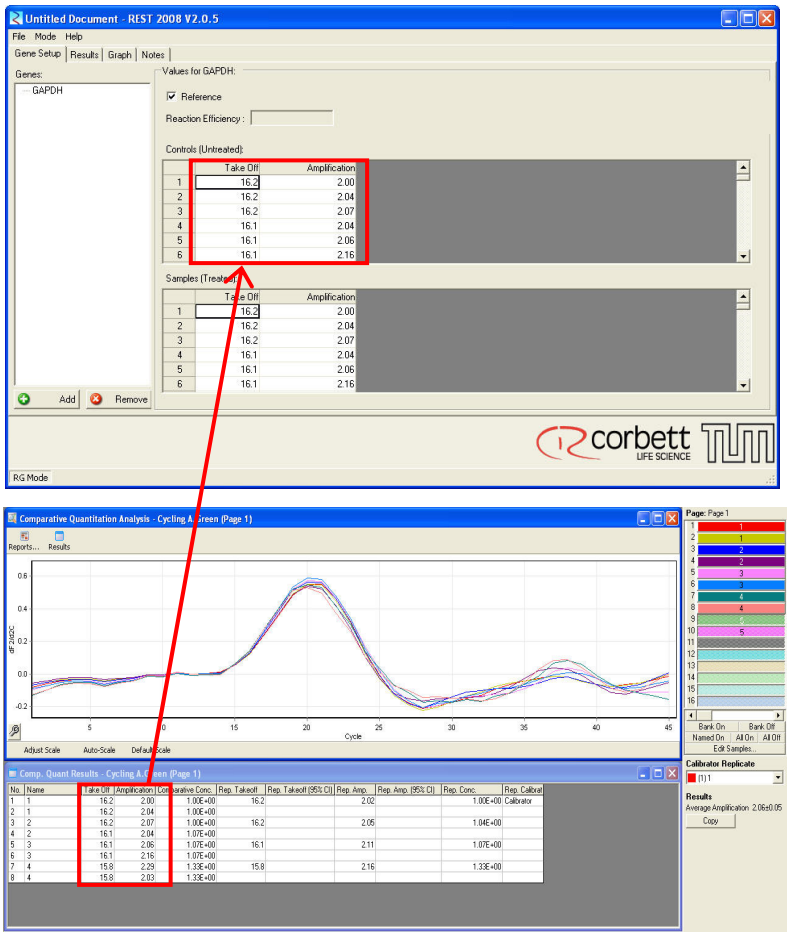
2. The Format of the screen will change and RGMode will appear in the bottom left of the window.



3. Add a new gene as described for the standard mode.



4. Enter the data from the Comparative Quantitation analysis tool.



Procedure

The REST-RG algorithm is almost identical to the standard algorithm however there are some differences in how it uses these new inputs.

The standard REST algorithm takes in two inputs, *Efficiency*, and C_Ts.

Efficiency = -1 ÷ *m*

where *m* is the gradient calculated from the standard curves line of best fit.

or alternatively

Efficiency = User inputted efficiency

where the user enters a specific efficiency to be used.

C_Ts = set of C_T values for all samples.

The standard algorithm uses *Efficiency* and C_{TS} to calculate P-values , and to calculate confidence intervals.

When we use REST-RG mode exactly the same algorithms are used however

Efficiency = average(*Amplification*) - 1

C_{TS} = *Take Off*

The average amplification of all genes is calculated, and used as the input to the algorithm.

Note: Minus 1 is included since *Amplification* lies on a scale from [1- 2] whereas *Efficiency* lies on a scale of [0-1].

In addition the Take Off values are substituted in place of the C_{TS}, and used in the randomisation algorithm.

References

[1] M. W. Pfaffl, G. W. Horgan & Leo Dempfle: "Relative Expression Software Tool (REST) for group-wise comparison and statistical analysis of relative expression results in Real-Time PCR" (Nucleic Acids Research 2002 May 1; 30(9): E36)

[2] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe & F. Speleman: "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes" (Genome Biology 2002, 3:research0034.1-0034.11)

[3] A.C. Davidson & D.V. Hinkley: Bootstrap Methods and their Application (ISBN 0-521-57391-2, Cambridge University Press 2002)

Links

This reference explains the development of the original REST methodology
(1) *Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR*
Michael W. Pfaffl, Graham W. Horgan and Leo Dempfle
http://rest.gene-quantification.info

(2) This reference provides a good introduction to the philosophy of randomised tests:
http://ordination.okstate.edu/permute.htm

(3) This reference provides an online interactive example of the test:
http://www.bioss.ac.uk/smart/unix/mrandt/slides/frames.htm

(4) This reference provides more detailed descriptions on how to carry out traditional tests, such as determination of confidence intervals and hypothesis testing using bootstrapping and randomisation:
http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html

(5) A description of Whisker-Box Plots:
http://regentsprep.org/Regents/math/data/boxwhisk.htm

Contact Information

Obtain software updates to REST 2008 here:
http://rest.gene-quantification.info/

If you have further questions or comments to improve the software, your suggestion are always welcome. Please contact us at this address:
rest-2008@gene-quantification.info

Corbett Life Science Pty Ltd:
http://www.corbettlifescience.com/