

REST 2005

Greater certainty in expression studies

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

New Standalone Software for Gene Expression Analysis

Software Updates:
<http://rest-2005.gene-quantification.info>

Contact Information:
rest-2005@gene-quantification.info



REST 2005

©2005 Corbett Research Pty Ltd and Michael W. Pfaffl

All rights reserved. No parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - without the written permission of the publisher.

Products that are referred to in this document may be either trademarks and/or registered trademarks of the respective owners. The publisher and the author make no claim to these trademarks.

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

Published: December 2005

Version: 1.9.10

1 Abstract

REST 2005 is a new standalone software tool to estimate up and down-regulation for gene expression studies. The software addresses issues surrounding the measurement of uncertainty for expression ratios, by using randomisation and bootstrapping techniques. By increasing the number of iterations from 2,000 to 50,000 in this version 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 the 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 distribution skew.

2 Why Rest?

Prior to the Relative Expression Software Tool [REST], Relative Quantitation in qRT-PCR 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 with highly variable data.

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 two groups: one which will be subject to pain (such immersion of the hand in ice-cold 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 subject to pain release more target mRNA than the other, or was the perceived increase due only to chance?

Prior approaches such as relative quantitation with two standard curves [Corbett] may be insufficient to answer this question confidently for lower magnitudes of expression. While average expression values can provide an indication of whether a particular subject in one group appeared to release more or less target mRNA than another subject, a statistical test provides a robust measure of the variance in such a measurement. 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 2005 overcomes these problems by using simple statistical randomisation tests [Davidson]. Such tests can appear counter-intuitive and so it is recommended to read the discussions on randomisation techniques in the topic [Links](#) before continuing.

3 Hypothesis Test

The purpose of REST 2005 is to determine whether there is a significant difference between samples and controls, while taking into account issues of reaction efficiency and reference gene normalisation. Because the normalisation and efficiency calculations involve ratios and multiple sources of error, it would be extremely difficult to devise a traditional statistical test, and so randomisation techniques are used instead.

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

4 Reference Gene Normalisation

REST 2005 allows the researcher to take into consideration multiple reference genes when determining expression, although it still remains possible to use a single reference. When estimating a sample's expression ratio, an intermediate absolute concentration value is calculated according to the following formula [1]:

$$\text{concentration} = \text{efficiency}^{\text{avg}(\text{Controls}) - \text{avg}(\text{Samples})}$$

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

$$\text{expression} = \text{goiConcentration} \div \text{refConcentration}$$

For multiple reference genes, the geometric mean is taken of all reference gene concentrations because concentration estimates vary exponentially [Vandersompe]:

$$\text{expression} = \text{goiConcentration} \div \text{GEOMEAN}(\text{refConc}_1, \text{refConc}_2, \dots)$$

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

$$\text{expression} = \text{GEOMEAN}(\text{goiConcentration} \div \text{refConc}_1, \text{goiConcentration} \div \text{refConc}_2, \dots)$$

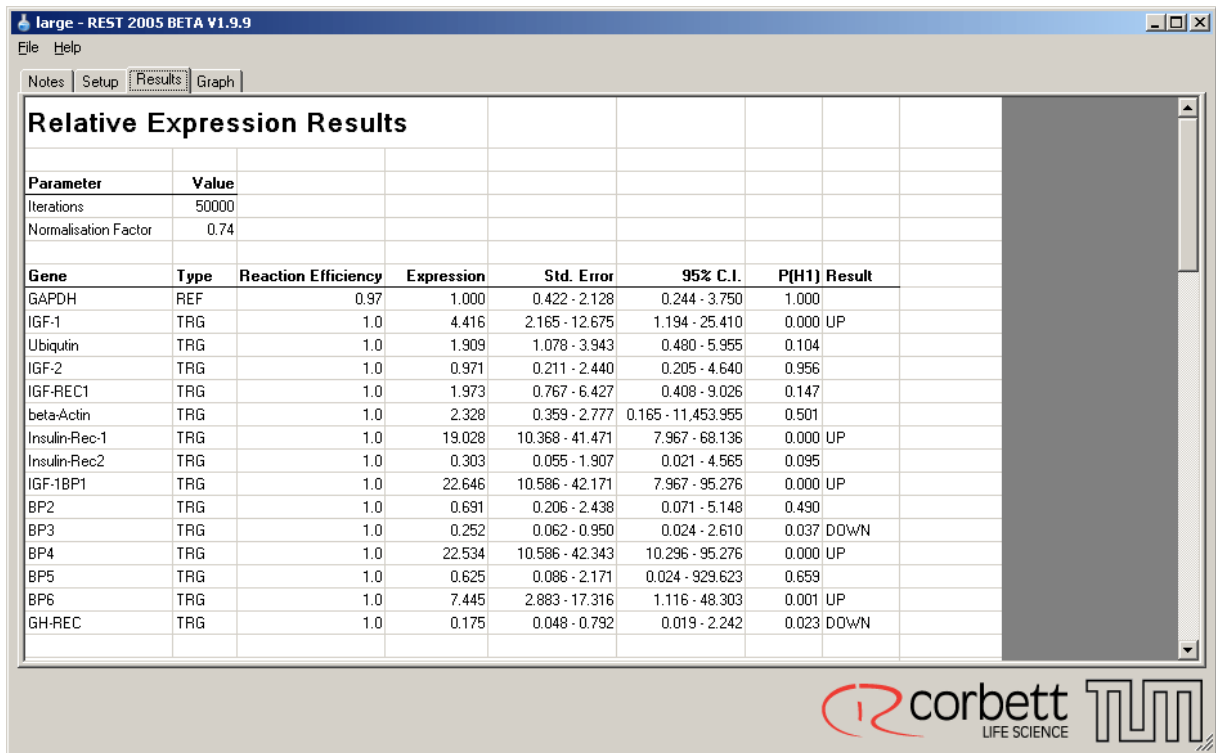
Since the mean concentrations of each gene do not change, they can be calculated at the beginning of the algorithm, and expressed as a single value, called the "*Normalisation factor*", equal to their geometric mean.

[1] *Errors in calculation of concentration occur due to linear variation in C_T values. Estimates of concentration use an equation of the form $c = A * \exp(C_T)$, and so vary exponentially.*

5 New Features

5.1 Greater Accuracy for Hypothesis Tests

The redevelopment of the REST 2005 software as a stand-alone application provides an order of magnitude of increase in performance. The speed improvements have been used to increase the number of randomisation iterations from 2,000 to 50,000, compared to earlier REST versions, increasing the accuracy and reproducibility of hypothesis tests to a level equivalent to traditional statistical tests. Figure 1 shows results from an example file, indicating the increased number of iterations.



large - REST 2005 BETA V1.9.9

File Help

Notes Setup Results Graph

Relative Expression Results

Parameter	Value
Iterations	50000
Normalisation Factor	0.74

Gene	Type	Reaction Efficiency	Expression	Std. Error	95% C.I.	P(H1) Result
GAPDH	REF	0.97	1.000	0.422 - 2.128	0.244 - 3.750	1.000
IGF-1	TRG	1.0	4.416	2.165 - 12.675	1.194 - 25.410	0.000 UP
Ubiquitin	TRG	1.0	1.909	1.078 - 3.943	0.480 - 5.955	0.104
IGF-2	TRG	1.0	0.971	0.211 - 2.440	0.205 - 4.640	0.956
IGF-REC1	TRG	1.0	1.973	0.767 - 6.427	0.408 - 9.026	0.147
beta-Actin	TRG	1.0	2.328	0.359 - 2.777	0.165 - 11.453.955	0.501
Insulin-Rec-1	TRG	1.0	19.028	10.368 - 41.471	7.967 - 68.136	0.000 UP
Insulin-Rec2	TRG	1.0	0.303	0.055 - 1.907	0.021 - 4.565	0.095
IGF-1BP1	TRG	1.0	22.646	10.586 - 42.171	7.967 - 95.276	0.000 UP
BP2	TRG	1.0	0.631	0.206 - 2.438	0.071 - 5.148	0.490
BP3	TRG	1.0	0.252	0.062 - 0.950	0.024 - 2.610	0.037 DOWN
BP4	TRG	1.0	22.534	10.586 - 42.343	10.296 - 95.276	0.000 UP
BP5	TRG	1.0	0.625	0.086 - 2.171	0.024 - 929.623	0.659
BP6	TRG	1.0	7.445	2.883 - 17.316	1.116 - 48.303	0.001 UP
GH-REC	TRG	1.0	0.175	0.048 - 0.792	0.019 - 2.242	0.023 DOWN

corbett LIFE SCIENCE TUM

Figure 1: REST 2005 Results Table

5.2 Expression Level Confidence Intervals

While previous REST publications 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 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 $1 \times 10^{-6}\%$. While a statistical test may show the difference to be significant, it clearly poses no real threat to drivers, when taking into consideration the average number of accidents a driver has in their lifetime. REST can perform its calculations based on CP and efficiency values determined by standard curve or kinetic techniques [Corbett].

Procedure

We are given a set of control (C) and sample (S) values for the gene of interest. We are given an efficiency value (e) and a normalising factor (f), based on the average expression of the reference genes.

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 (>30,000).

Let *choose()* be a function that returns a random element from a set.

We populate Y by randomly pairing controls and samples, and calculating their expression ratio:

$$i \in \{1, \dots, n\}, Y_i = e^{\text{choose}(C) - \text{choose}(S)} \div f$$

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

$$Y_{\text{sorted}} = \text{sort}(Y)$$

The 95% confidence interval is defined as:

$$\alpha = 0.05$$

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

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

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

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

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

GAPDH Control (GC)	GAPDH Sample (GS)	IGF -1 Control (IC)	IGF Sample (IS)
31.33	22.89	20.34	30.01
31.00	22.34	20.56	30.02
29.92	22.91	21.22	30.02
31.01	21.98	23.33	29.01
31.01	21.83	22.98	29.33
30.98	21.49	22.34	29.35
33.23	23.07	23.01	30.34
33.56	22.22	22.15	30.31
31.98	22.83	22.01	30.98
	22.67		29.01
	22.69		29.99
			29.94

The normalisation factor f is calculated by the following formula:

$$f = (1 + \text{refEfficiency})^{\text{avg}(\text{GC}) - \text{avg}(\text{GS})}$$

$$f = 0.7350\dots$$

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

GC _r	GI _r	Expression
31.01	29.35	4.33486199258
31.01	30.31	2.21772177376
31.0	29.01	5.45795529972
31.33	29.94	3.590147899
31.01	29.01	5.49619249906
31.0	30.98	1.37953823369
30.98	29.01	5.38227710286
31.01	29.33	4.39581287507
33.23	29.94	13.5264816955
31.01	29.99	2.77287184972

Sorting yields Y_{sorted} :

Expression

1.37953823369
 2.21772177376
 2.77287184972
 3.590147899
 4.33486199258
 4.39581287507
 5.38227710286
 5.45795529972
 5.49619249906
 13.5264816955

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

$$\text{confidence}_{68\%} = [2.21772177376, 5.49619249906]$$

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

$$\text{confidence}_{95\%} = [1.37953823369, 13.5264816955]$$

$p < 0.05$

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

$$\text{confidence}_{99.7\%} = [1.37953823369, 13.5264816955]$$

$p < 0.0027$

The median is calculated as the 5th position:

$$\text{median} = 4.33486199258$$

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 ($\alpha = 0.5$, $\alpha = 0.05$, $\alpha = 0.95$) 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

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

confidence_{68%} = [2.71540064663, 5.49619249906]
confidence_{95%} = [1.27756255737, 25.711370228]
median = 4.2747562348

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

For the same data set, REST calculated comparable values:

expression = 4.453 ± 2.36831
sample up-regulated = YES ($p = 0.001$)

As all values in the 95% confidence interval were greater than 1, the interval is consistent with the REST P value less than 0.05. 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.

5.3 Efficiency Error Measurement

All statistical tests in REST 2005 now include correction for variation in efficiency. If variation in efficiency is low, hypothesis tests will produce more conclusive results, and confidence bands for estimated expression will be smaller.

As all statistics are calculated using randomisation techniques, the approach for measuring standard curve error must also be stochastic, and is expressed as a challenge: If we ignore variation in the standard curve, the slope (m value) will be expressed as a constant in all equations. Say, then, we have a standard curve of six data points for the gene GAPDH that we use to estimate its efficiency. If there is no variation in the standard curve, then we could pick any two points in the curve and still measure the same gradient. If, however, there is large variation between the points, then random selection of points will greatly vary the efficiency calculated. Using a few data points, we can then simulate the random variable representing the efficiency error. The randomised efficiency value is then included in calculations instead of the slope of the line of best fit, feeding any variation in efficiency directly into the relative quantitation hypothesis tests and confidence intervals.

5.4 Whisker-Box Plots

REST 2005 replaces the bar graph visualisation in prior versions with a statistical whisker-box plot. In statistical applications, whisker-box plots provide additional information about the skew of distributions that would not be available simply by plotting the sample mean. See the link below for general information about whisker-box plots:

<http://regentsprep.org/Regents/math/data/boxwhisk.htm>

To summarise, the box area in a whisker-box plot encompasses 50% of all observations, the dotted line represents the sample median and the whiskers represent the outer 50% of observations, as shown in Figure 2:

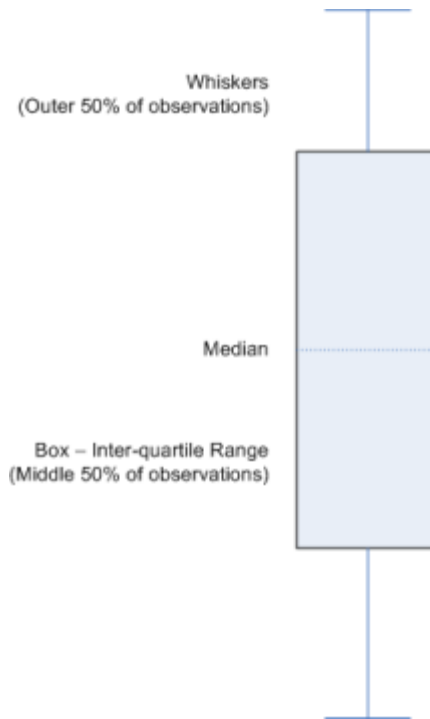


Figure 2: Typical Distribution Shown in Whisker-Box Plot

If the distribution of data is skewed, or non-linear, the tails of the data may be asymmetrical. In the example plot shown below (Figure 3), the upper 25% of observations were highly variable:



Figure 3: Whisker-Box Plot of Distribution With High Variability on Upper Tail

Because REST 2005 uses randomisation techniques, it draws whisker-box plots based upon the

permuted expression data (Y set) rather than the raw CP values input by the user. See the section [Expression Level Confidence Intervals](#) for more information on how the permuted data set is generated. Because expression values are ratios, they will often have lopsided ratios, with greater variability on the upper tail as seen above. As ratio populations can be unpredictable, and subject to large and uneven variability, this visualisation draws out characteristics of gene expression that may otherwise go unnoticed.

Figure 4 shows a Whisker-Box Plot based on example data. In the plot, while Ubiquitin has a median expression of 1.16, there is an uneven distribution of values within the inter-quartile range, whereas IGF-1 represents a more uniform distribution. β -Actin has a large spread of possible values in the upper range, while its lower bound is more stable.

By accurately representing asymmetrical variance, conclusions such as "the gene is likely to be up-regulated by a factor of 8" may still be valid in from data with large amounts of variance only on one side of the distribution.

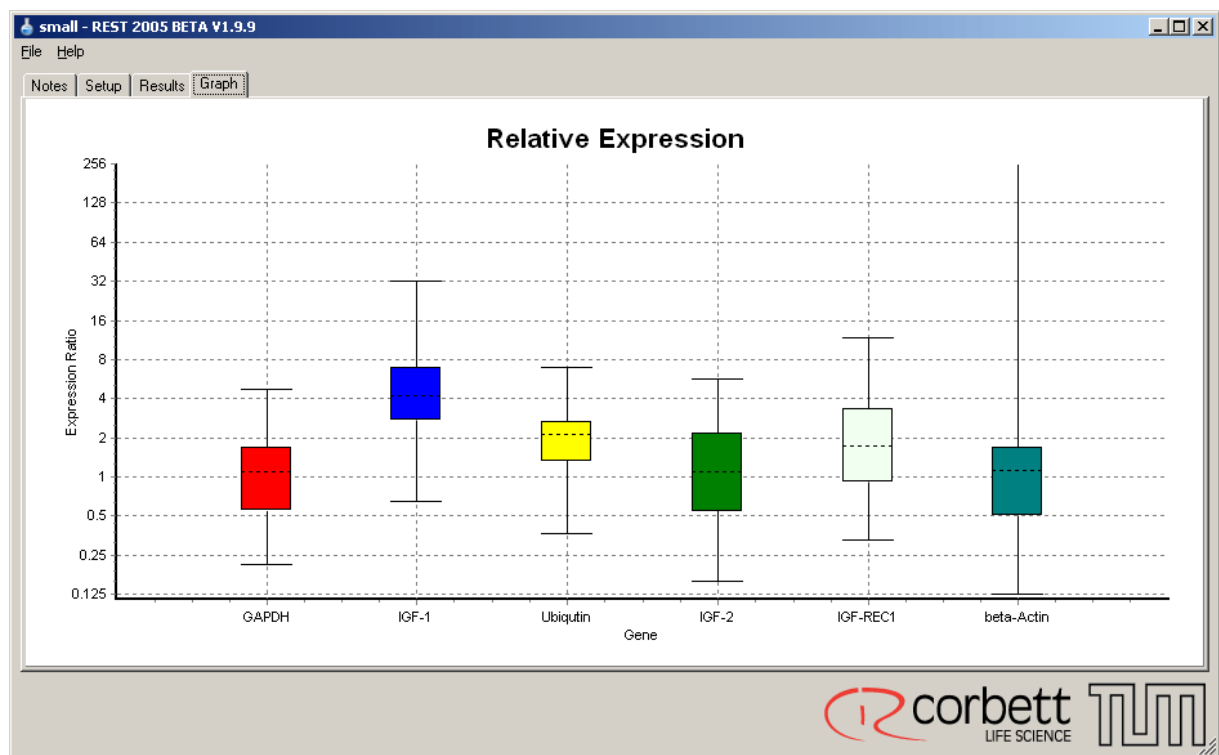


Figure 4: Whisker-Box Plots for 6 Analysed Genes

6 References

[Pfaffl] 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)

[Vandesompele] 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)

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

[Corbett] Corbett Research Pty Ltd: Rotor-Gene 6 Software User Guide (Australia, Sydney, 2004)

7 Links

This reference provides a good introduction to the philosophy of randomised tests:

<http://ordination.okstate.edu/permute.htm>

This reference provides an online interactive example of the test:

<http://www.bioss.ac.uk/smart/unix/mrandt/slides/frames.htm>

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>

A description of Whisker-Box Plots:

<http://regentsprep.org/Regents/math/data/boxwhisk.htm>

8 Contact Information

Please find more information in the **HELP manual** of the REST 2005 software (by pressing **F1**) or at the respective web page on the www.gene-quantification.info pages:

<http://rest-2005.gene-quantification.info/>

Obtain software updates to REST 2005 here:

<http://rest.gene-quantification.info/>

If you have further questions or comments to improve the software, your suggestions are always welcome. Please contact us at this address:

rest-2005@gene-quantification.info?subject=REST-2005

Corbett Life Science Pty Ltd:

<http://www.corbettlifescience.com/>