The Prime Technique

Real-time PCR Data Analysis

For measuring gene expression there is only one technique: PCR. But how can it be used with maximum efficiency? This article tries to give the answer to that question.

Mikael Kubista, Institute of Molecular Genetics and TATAA Biocenter, Sweden

Radek Sindelka, Institute of Molecular Genetics, Czech Republic

Quantitative real-time PCR (QPCR) is today the prime technique to measure gene expression. [1] When properly used it offers unprecedented sensitivity, accuracy and reproducibility. But there are caveats. The target is mRNA, which must be extracted and converted to cDNA in a reverse transcription process that can proceed at highly variable yield depending on protocol. [2, 3] RNA is further rapidly degraded by nucleases abundant in biological samples. [4] To account for processing variation the expression of marker genes is normalised with appropriate endogenous control genes and technical repeats are performed to reduce confounding variance.

Typical QPCR experimental design is shown in figure 1. Studying the effect of treatment samples are collected of control and treated subjects. Each sample is divided into replicate RT reactions, which are split into replicates for QPCR. [2]

Each QPCR measurement generates a CT value, which is the number of amplification cycles required to reach a certain threshold signal level. CT values are inversely proportional to the logarithm of the initial number of target copies present in the sample. Figure 2 shows CT values in a spreadsheet. Expression of one marker gene (MG) and one reference gene (RG) was measured in 6 control and 6 treated subjects, with RT triplicates and QPCR duplicates. This gave a total of 72 samples. 36 were analyzed per run together with interplate calibrators. The treatment and the replicates are indexed in classification columns identified with labels that begin with a hatch. Another classification column indexes the sample amounts used. QPCR data are pre-processed as follows:

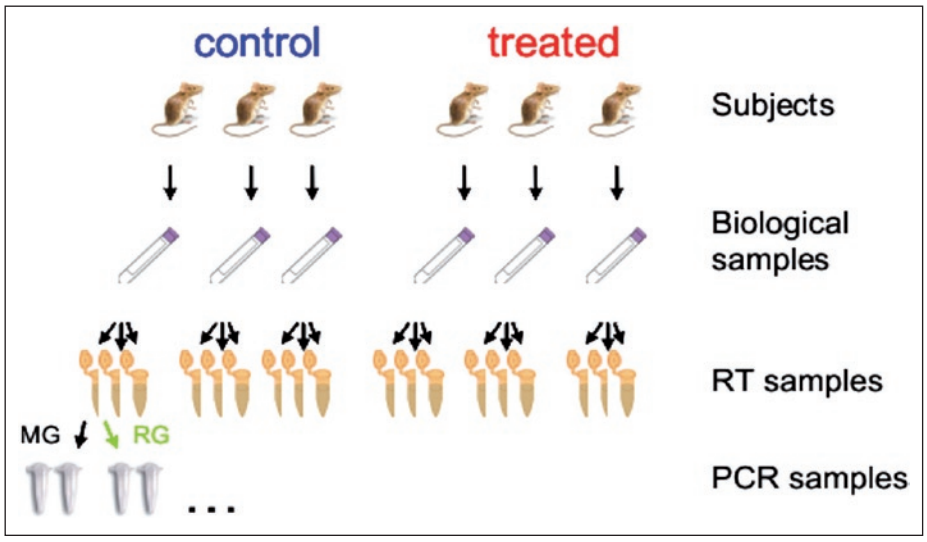


Fig. 1: Nested QPCR experimental design

Correct for Off-scale Measurements

Occasionally amplification response curves never reach threshold. Sometimes signal reaches threshold but is due to formation of aberrant products, such as primer-dimers. In either case we don't have reliable CT reading. Experiments containing such off-scale data should be analyzed with non-parametric methods that do not assume data are Normal distributed. [5] Non-parametric methods, however, are weaker than parametric in the sense that more replicates are needed to reach reliable conclusions. An alternative is to replace off-scale data with fictive CT values and use parametric tests. Fictive CT values are set to the highest CT observed for a truly positive sample, assumed to be the level of detection (LOD), plus 1. [1] This corresponds to assigning a concentration that is half of LOD to the off-scale samples. This is no more erroneous than assigning zero concentration to these samples, because we do not know that they are blank; we only know that they contain less amount of target than we are able to detect. If we are uncertain about the correction we repeat the analysis assigning CT(LOD) + 2 to the off-scale samples. If the result is virtually the same, we can be confident that the correction has negligible effect.

Efficiency Correction

PCR efficiency ($0 \leq E \leq 1$) depends on both the assay used and on the sample matrix. Estimations of PCR efficiency can be more or less advanced. [1] Once estimated, the measured CT values are corrected as:

$$CT_{E=100\%} = CT_E \frac{\log(1+E)}{\log(2)}$$

Variations between Runs

When all samples cannot be run in one plate correction is needed. This is done by including an identical sample, called interplate calibrator (IC), in all runs, which is analyzed for all genes.

$$CT_{plate\ norm} = CT_r - CT_r^{IC} + \frac{1}{m}\sum_{r=1}^{m} CT_{IC}$$

Sample Amount

For samples based on different starting amounts CT values are corrected as:

$$CT_{conc=1} = CT_{conc} - \log_2(conc)$$

Here is a typo! Correct equation has a plus sign. See GenEx manual (www.multid.se)

QPCR Technical Repeats

QPCR technical repeats are averaged before normalisation with reference genes.

$$CT_{QPCR\_average} = \frac{1}{n}\sum_{i=1}^{n} CT_{QPCR\_repeats}$$

Reference Genes

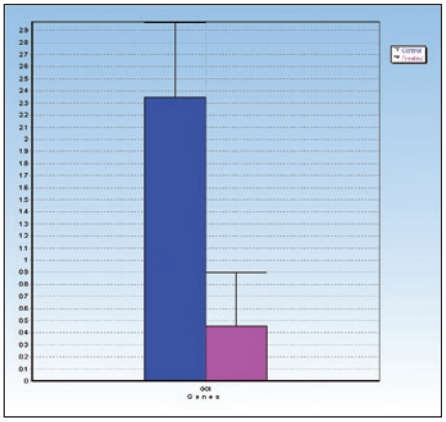
Expression of marker genes is normalised to that of reference genes:

$$CT_{MG,norm} = CT_{MG} - \frac{1}{n}\sum_{i=1}^{n} CT_{RG}$$

RT Technical Repeats

RT technical repeats are averaged after normalisation with reference genes.

$$CT_{RT\_average} = \frac{1}{n}\sum_{i=1}^{n} CT_{RT\_repeats}$$



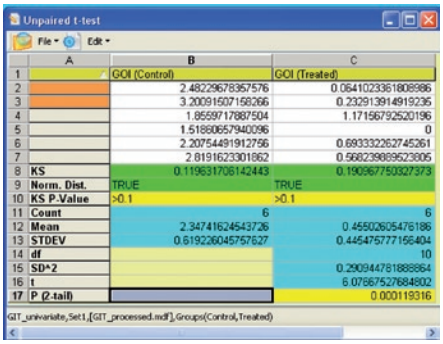
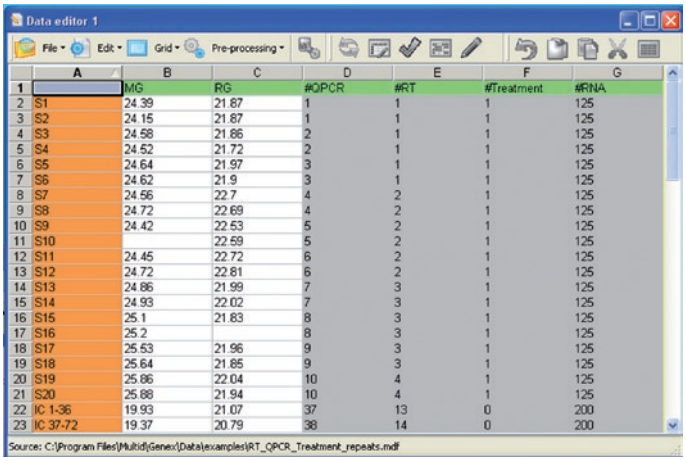


Fig. 3: Top: Bar graph showing mean with 95 % confidence interval for control and treated samples. Bottom: Result of unpaired 2-tailed t-test.

Fig. 2: Data prepared for analysis. Classification columns, identified by a hatch, index technical and biological replicates and also contain additional information needed for data pre-processing. IC is interplate calibrator.



Relative Quantities

Relative expression among samples is calculated as:

$$RQ = 2^{CT_{ref} - CT}$$

Log Scale

Data shall be Normal distributed for analysis with tests such as the t-test, linear regression, and ANOVA. Gene expression data are usually not Normal distributed when expressed as relative quantities, but often become Normal distributed by logarithmic transformation to fold differences (FD). Traditionally log base 2 is used:

$$FD = \log_2(RQ)$$

Data in figure 1 were pre-processed assuming 90 % PCR efficiency for the marker gene and 95 % efficiency for the reference gene. When averaging the technical repeats missing data were accounted for.

After pre-processing the FD's of the biological repeats were averaged and are shown in figure 3 with 95 % confidence interval. They were confirmed to be Normal distributed by the Kolmogorov-Smirnov's test, and the means were compared with the unpaired 2-sided t-test (fig. 3). This gave P = 0.00012 so we conclude that treatment most likely has effect on the expression of the marker gene. Analyzing the same data with the non-parametric Mann-Whitney's test we still find the difference significant, but with lower confidence (P_{MW} = 0.005).

Multiple Genes

The procedure can be used to compare the effect of treatment on several genes. However, such multiple testing is statistically unsound. The P-value is the probability to observe a difference that is at least as large as the measured in absence of treatment effect. It is up to the investigator to decide how low a P-value shall be taken as indicator of treatment effect. Often investigators work with 95 % confidence, which for a single test translates to P ≤ 0.05. With this criterion one out of 20 studies with no effect will turn out as a false positive. Usually this is acceptable error rate. But what happens if multiple genes are compared? If 10 genes are compared in two identical samples probability that all 10 show a differences below threshold of P = 0.05 is 0.95^10 = 0.60, or 60 %. Hence, the probability that at least one out of them gives P ≤ 0.05 is 40 %! This problem of multiple testing becomes more and more important with increasing number of genes, and has led to serious criticism of microarray gene-expression profiling studies. [6]

Multivariate Expression Profiling

The proper way to analyze data based on expression of multiple genes is with multivariate methods. [1, 7] Most common are Principal

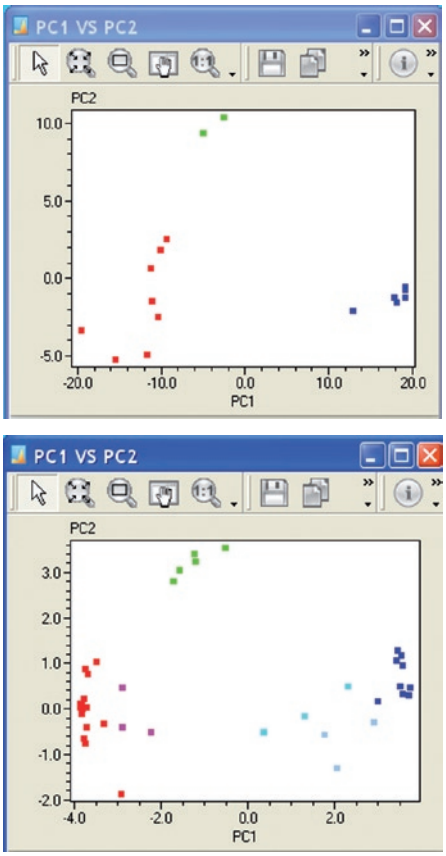


Fig. 4: PCA classification of *Xenopus laevis* expression data presented in PC1 vs. PC2 scatter plots. Top panel shows samples and bottom panel shows genes. Biological replicates are shown as independent symbols in the left panel.

Component Analysis (PCA), Hierarchical clustering, and the Self-Organised Map. With these methods multiple samples, each studied by measuring the expression of multiple genes, can be readily analyzed to identify those samples and those genes that have common expression behaviour. The data are preprocessed as above, with one additional step.

Mean Centering/Autoscaling

To remove the effect of the overall expression of the different genes the mean expression of every gene is subtracted.

$$FD_{MC} = FD - \overline{FD}$$

Hence, for mean centered data a certain increase in expression has the same significance independently of the absolute expression level of that gene. Mean-centred data are mainly used for the classification of samples.

To remove also the effect of the magnitude of the change, the data are further divided with the standard deviation:

$$FD_{AS} = \left(FD - \overline{FD}\right) / SD = FD_{MC} / SD$$

The autoscaled data are mainly used to classify genes.

Expression of sixteen genes was measured during sixteen developmental stages, ranging from the oocyte to the tadpole, of the frog *Xenopus laevis.* The data were processed as described above, though without normalisation with reference genes, since no genes with constant expression during *Xenopus laevis* development has been identified. [8] This is possible when data are scaled. [1] Top panel in figure 4 shows classification of the developmental stages based on PCA of mean-centered data and bottom panel shows classification of the genes based on autoscaled data. Both panels reveal that development goes through three distinct stages.

Acknowledgement

Material used is from TATAA Biocenter (www.tataa.com) biostatistics courses in QPCR. Data were analyzed with GenEx software from MultiD Analyses (www.multid.se). The example data are available on: www.multid.se/downloadspage.html.

References

[1] Kubista M. *et al.:* Molecular Aspects of Medicine 27, 95–125 (2006)

[2] Ståhlberg A. *et al.:* Clin. Chem. 50, 509 (2004)

[3] Ståhlberg A. *et al.:* Clin. Chem. 50, 1679 (2004)

[4] Fleige S. and Pfaffl M.W.: Molecular Aspects of Medicine 271, 26–139 (2006)

[5] Intuituve Biostatistics, Harvey Motulsky, Oxford University Press, New York, ISBN: 0-19-50-8607-4, 1995

[6] Ioannidis J.P.: Lancet 365, 454–455 (2005)

[7] Kubista M. *et al.:* European Pharmaceutical Review 56, 1 (2007)

[8] Sindelka R. *et al.:* Dev Dyn.235, 754–758 (2006)

Authors:

Mikael Kubista, Laboratory of Gene Expression, Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic and TATAA Biocenter AB

Radek Sindelka, Laboratory of Gene Expression, Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

Ales Tichopad, Physiologie Weihenstephan, Technical University Munich, Germany

Anders Bergkvist, MultiD Analyses AB, Gothenburg, Sweden

Daniel Lindh, MultiD Analyses AB, Gothenburg, Sweden

Amin Forootan, MultiD Analyses AB, Gothenburg, Sweden

▶ Contact

Mikael Kubista
TATAA Biocenter AB
Göteborg, Sweden
mikael.kubista@tataa.com