A statistical test vs. a validation experiment in Gene Expression Study

Veska Noncheva, Maria Dobreva

Plovdiv University ”Paisij Hilendarski”; “University of Food Technologies”- Plovdiv
wesnon@uni-plovdiv.bg; mimi.d.d@abv.bg

Abstract : Comparative CT Method compares the Ct value of one target gene to another using the formula called $2^{-\Delta\Delta CT}$. To make this method valid, the efficiency of the target amplification (the gene of interest) and the efficiency of the reference amplification (the endogenous control) must be equal. In this article we propose to test statistical hypotheses instead to perform validation biological experiments when we want to show that the efficiencies of the target and endogenous control amplifications are approximately equal.

Keywords: *statistical hypotheses, one sample location test, gene expression.*

1. Introduction:

There are two different methods for analyzing data from quantitative real-time RT-PCR (reverse transcription followed by polymerase chain reaction): the quantitative absolute and relative quantification. Relative quantification in quantitative real-time RT-PCR is increasingly used to quantify gene expression changes [2,3,4]. It is based on the expression levels of a target gene versus a reference gene and in theory is adequate for most purposes to investigate physiological changes in gene expression levels. The units used to express relative quantities are irrelevant and the relative quantities can be compared across multiple RT-PCR experiments.

Methods for relative quantitation of gene expression allow us also to quantify differences in the expression level of a specific target gene between different biological samples. The data analysis output is expressed as a fold-change or a fold-difference of expression levels.

The aim of our research work is to find out a methodology based on statistical methods for identification genes whose expression level changes across conditions under study. Some of its applications could be to compare the levels of gene expression in a cancerous tumor and normal tissue in order to identify which genes are differentially expressed or the level of gene expression of a particular gene of interest in a chemically treated sample to be compared with the level of gene expression of the untreated sample..

The calculation method used usually for relative quantitation is called Comparative CT method. This method known also as ΔΔCT Method, uses arithmetic formulas, given below, to achieve the result for relative quantitation [1].

2. The Mathematical Model.

Biologically meaningful reporting of target mRNA copy numbers requires accurate and relevant normalization to some standard and is strongly recommended in real-time RT-PCR. Real-time RT-PCR-specific errors in the quantification of mRNA transcripts are caused by sample-to-sample variation, variation in RNA integrity, RT efficiency differences, and cDNA sample loading variation. This is especially relevant when the samples have been obtained from different individuals, different tissues, and different time courses and would result in misinterpretation. Therefore, normalization of target-gene expression levels must be performed to compensate intra- and inter-kinetic RT-PCR variations (sample-to-sample and run-to-run variations).

Because of this the relative quantitation of gene expression requires quantitation of two different genes - target gene and endogenous control.

The equation that describes the exponential amplification of PCR is:

$X_n = X_o\ (1+ E_x)^n$,

where X_n is the number of target molecules at cycle *n*, X_o is the initial number of target molecules, E_X is the efficiency of target amplification, and *n* is the number of cycles.

The threshold cycle (Ct) indicates the fractional cycle number at which the amount of amplified target reaches a fixed threshold. Thus,

$X_T = X_o\ (1+ E_x)^{Ct,X} = K_x$,

where X_T is the threshold number of target molecules, *Ct,X* is the threshold cycle for target amplification, and K_X is a constant.

The equation for the endogenous reference reaction is:

$R_T = R_o\ (1+ E_R)^{Ct,R} = K_R$,

where R_T is the threshold number of reference molecules, R_o is the initial number of reference molecules, E_R is the efficiency of reference amplification, *Ct,R* is the threshold cycle for reference amplification, and K_R is a constant.

Dividing X_T by R_T gives the following expression:

$$\frac{X_T}{R_T} = \frac{\mathrm{Xo}\ (1+E_X)^{Ct,X}}{\mathrm{Ro}\ (1+R_X)^{Ct,R}} = \frac{K_X}{K_R} = K.$$

The exact values of X_T and R_T depend on a number of factors, including reporter dye used in the probe, sequence context effects on the fluorescence properties of the probe, efficiency of probe cleavage, purity of the probe, setting of the fluorescence threshold, etc.

Note that normalization issues arise only to the extent that technical factors have sample-specific effects. Normalization takes the form of correction factors that enter into the statistical models.

Assuming that efficiencies of the target and the reference are the same

$E_X = E_R = E$, we receive

$\frac{X_o}{R_o}\ (1+E)^{Ct,X-Ct,R} = K,$

$X_N\ (1+E)^{\Delta Ct} = K,$

where $X_N = X_o/R_o$, is the normalized amount of target, and $\Delta Ct = Ct,X - Ct,R$ is the difference in threshold cycles for target and reference gene. Therefore,

$X_N = K\ (1+E)^{-\Delta Ct}.$

To obtain accurate relative quantitation of an mRNA of a target gene, it is also recommended to evaluate the expression level of an endogenous control. By using an endogenous control as an active reference, you can normalize quantitation of targets for differences in the amount of total nucleic acid added to each reaction. For example, if you determine that a calibrator sample has a two-fold greater amount of endogenous control than a test sample you would expect that the calibrator sample was loaded with two-fold more cDNA than the test sample. Therefore, you would have to normalize the test sample target by two-fold to accurately quantify the fold-differences in target level between calibrator and test samples of individuals.

The final step is to divide the X_N for any sample q by the X_N for the calibrator (cb).

$\frac{X_{N,q}}{X_{N,cb}} = \frac{K\ (1+E)^{-\Delta Ct,q}}{K\ (1+E)^{-\Delta Ct,cb}} = (1+E)^{-\Delta\Delta Ct},$

where $\Delta\Delta Ct = \Delta CT,q - \Delta CT,cb.$

The calibrator is a sample that is used as the basis for comparative expression results. The calibrator sample might be an untreated sample.

For amplicons designed and optimized according to Applied Biosystems guidelines, the efficiency is close to one. Therefore, the amount of target, normalized to an endogenous reference and relative to a calibrator, is given by $2^{-\Delta\Delta Ct}$.

Note the following:

- 1 CT shows 2-fold difference in the initial amount of the template.
- A 5 cycle decrease corresponds to a 32-fold higher RT efficiency, because $2^5=32$.
- A four-fold change should yield a +2 CT difference ($4=2^2$).
- $2^{-\Delta\Delta Ct} = 2^{-(2.0)} = 0.25$ means that the test sample has 0.25 or 1/4 the amount of target RNA as the calibrator.

2.1. Assumptions of the $2^{-\Delta\Delta Ct}$ Method.

In order the calculation of the $2^{-\Delta\Delta Ct}$ to be valid the amplification efficiency of the target gene and the reference gene must be equal. Sensitive method for assessing if two amplicons (the products of the polymerase chain reaction (PCR)) have the same

efficiency is to design a new experiment and to look at how the ΔCt varies with template dilution.

The second assumption is the following. The efficiency for the amplicons must be close to one.

The gene expression dependence on both vascular risk factors (obesity, smocking) and demographic factors (age and gender) should be checked.

If all assumptions hold, the $2^{-\Delta\Delta Ct}$ method may be used to analyze the data.

Therefore, the next step in the application of the $2^{-\Delta\Delta Ct}$ method is to consider the assumptions being made. This is important as invalid assumptions will mean that the results of the $2^{-\Delta\Delta Ct}$ method are invalid.

In this paper we put the following question: *Could statistical tests determine if the $\Delta\Delta CT$ calculations are correct?*

3. Statistical hypotheses testing.

Using the $2^{-\mathbf{\Delta\Delta Ct}}$ method, the data are presented as the fold change in gene expression normalized to an endogenous reference gene and relative to the untreated control.

Usually before using the $\Delta\Delta$CT method for quantitation, biologists perform additional validation experiments to demonstrate that efficiencies of target and reference are equal.

In our work we propose an alternative approach based on a statistical method instead of validation experiments.

Suppose the CT values provided from real-time PCR instrumentation are reported from a quantitative gene expression experiment. If the efficiencies of the two PCR reactions are equal, we can rely on $\Delta\Delta$CT values and use them in our study in the area of Biology or Medicine.

Let ΔCT,cb values be calculated for each individual of a random sample of untreated individuals with size k. Let ΔCT,q be calculated for the same random sample of untreated individuals. Let us pair all different individuals. In this way we receive a sample of paired individuals with size $m=C_k^2$. Thus we receive

$$\Delta\Delta Ct_{ij} = \Delta CT,cb_i - \Delta CT,cb_j = (Ct,_{Target} - Ct,_R)_{\mathbf{Cb}i} - (Ct,_{Target} - Ct,_R)_{\mathbf{Cb}j}\text{, where } i \neq j,$$

From theoretical point of view all calculated values for the differences $\Delta\Delta Ct_{ij}$ should be close to zero. Therefore, the expectation of the random variable $\Delta\Delta$Ct must be zero.

Now we can use one sample location test to compare the location parameter of the population to constant 0. The location parameter in this task could be the mean or the median.

Therefore, using the appropriate two-sided test for statistical hypotheses testing we can make decision whether calculated $\Delta\Delta$CTs are correct.

The biological interpretation of the results is the following:

If the calculated p-value is greater than the level of significance, the efficiencies of the target and reference genes are similar and the gene expression is not influenced neither by vascular risk factors nor by demographic factors.

If the p-value is less than the level of significance, some assumptions of the $\Delta\Delta$CT method are not fulfilled.

Whether the efficiencies of target and reference are equal to 1, is still an open question.

4. Conclusion.

In relative quantification, researchers analyze changes in gene expression in a given sample relative to another reference sample. The proposed approach is useful for an investigator to analyze quantitative gene expression data using the $2^{-\Delta\Delta Ct}$ method.

To summarize the most important steps in this approach:

1. Select the gene for internal control and a random sample of untreated individuals.
2. Use Ct values and the $\Delta\Delta$CT method in order to quantify the gene expression.
3. Use the appropriate two-sided location test to ensure that some of the assumptions of the $\Delta\Delta$CT method are fulfilled.

Finally, powerful statistical methods for data analysis should be applied and new knowledge in the area of Biology and Medicine be extracted.

References:

1. Livak Kenneth J. , Thomas D. Schmittgen. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method, METHODS, 2001, V. 25, pp 402–408.
2. Livak Kenneth J, Quin F Wills, Alex J Tipping, Krishnalekha Datta, Rowena Mittal, Andrew J Goldson, Darren W Sexton, Chris C Holmes. METHODS FOR QPCR GENE EXPRESSION PROFILING APPLIED TO 1440 LYMPHOBLASTOID SINGLE CELLS, METHODS, 2013, V. 59, ISSUE 1.
3. Millera Brooke H., Zane Zeierb, Li Xid, Thomas A. Lanzd, Shibing Dengd, Julia Strathmanne, David Willoughbyf, Paul J. Kennya, John D. Elsworthg, Matthew S. Lawrencei, Robert H. Rothg, Dieter Edbauere, Robin J. Kleimand, Iaes Wahlestedtb. MicroRNA-132 dysregulation in schizophrenia has implications for both neurodevelopment and adult brain function, PNAS, Online ISSN 1091-6490, 2011.
4. Popov N., N. Madjirova, I. Minkov, T. Vachev. MICRO RNA HSA-486-3P GENE EXPRESSION PROFILING IN THE WHOLE BLOOD OF PATIENTS WITH AUTISM, MeDicAl biotechnology, 06/2012.

Резюме: *Отдавна е известно, че за трудно лечимите болести е налична генетична основа. Ето защо в последните години науката се насочва към търсенето на биологични причини, водещи до промяна на генната функция без да е на лице нарушение в структурата на гена. Генна експресия е процес, при който унаследяемата информация от ДНК се трансформира във функционален продукт. Количеството функционални продукти на ДНК, или по-точно количеството органични молекули, наричаме ниво на генна експресия. Целта е да се идентифицират микро-РНК биомаркерни молекули в серум като индикатори, използвани за молекулярна диагностика.*

Целта на работа ни е да създадем методология, която да намира значимите разлики в нивата на експресия при клинично здравите и при болните пациенти. За тези гени не трябва да има значима разлика в нивата на експресия при клинично здравите пациенти.

Има различни технологии, които измерват нивото на генна експресия в конкретен момент. Съществуват два основни метода за анализ на получените данни от RT- PCR експерименти: абсолютно количествен и относително количествен метод.

Методите за относителното количествено определяне на генната експресия ни позволяват да определим количествено разликите в нивото на експресия на гена, от който се интересуваме, от различни проби. Резултатът от анализа на данните се изразява като промяна в пъти в нивата на генната експресия.

Сравнителният CT Метод сравнява стойността Ct на един ген с друг, като използва формула $2^{-\Delta\Delta CT}$. За да бъде валиден този метод, ефективността на гена от който се интересуваме и ефективността на референтната амплификация (ендогенния контрол) трябва да бъдат от една страна равни, а от друга страна всеки от тях трябва да е равен на единица.

Коректното прилагане на метода изисква предположенията му да бъдат проверени.

В тази статия ние предлагаме за проверка на предположенията на метода $\Delta\Delta CT$ да се използва статистически критерий, вместо да се извършват скъпо струващи биологични експерименти.