Benchmarks

The choice of reference gene affects statistical efficiency in quantitative PCR data analysis

Yi Guo[1], Michael L. Pennell[2], Dennis K. Pearl[3], Thomas J. Knobloch[4], Soledad Fernandez[5], and Christopher M. Weghorst[4]

[1]Department of Health Outcomes and Policy, College of Medicine, University of Florida, Gainesville, FL, [2]Division of Biostatistics, [3]Department of Statistics, [4]Division of Environmental Health Sciences, and [5]Center for Biostatistics, The Ohio State University, Columbus, OH

BioTechniques 55:207-209 (October 2013) doi 10.2144/000114090
Keywords: quantitative PCR; reference gene; statistical efficiency

Supplementary material for this article is available at www.BioTechniques.com/article/114090.

Quantitative polymerase chain reaction (qPCR), a highly sensitive method of measuring gene expression, is widely used in biomedical research. To produce reliable results, it is essential to use stably expressed reference genes (RGs) for data normalization so that sample-to-sample variation can be controlled. In this study, we examine the effect of different RGs on statistical efficiency by analyzing a qPCR data set that contains 12 target genes and 3 RGs. Our results show that choosing the most stably expressed RG for data normalization does not guarantee reduced variance or improved statistical efficiency. We also provide a formula for determining when data normalization will improve statistical efficiency and hence increase the power of statistical tests in data analysis.

Accurate normalization is essential for eliminating sample-to-sample variation and getting reliable results from quantitative PCR (qPCR) experiments (1). To date, the most frequently used method for normalization relies on the expression of an internal reference gene (RG). In theory, this method ensures that all of the steps in a qPCR experiment are controlled (2). However, there is a stringent requirement that the chosen RG be expressed at a relatively constant level regardless of experimental conditions. Previous studies have shown that the expression of many widely used RGs, including glyceraldehyde-3-phosphate dehydrogenase (GAPDH), β-actin (ACTB), and ribosomal subunit 18S, does not remain unchanged across tissue types and treatment conditions (3–7). Consequently, the identification of more stably expressed endogenous RGs has been recommended for normalization (8, 9).

Many software packages are available for selecting RGs, including geNorm and NormFinder. With geNorm, a gene-stability measure, which is used to rank candidate RGs, is calculated for each RG based on its pairwise expression ratios with all other RGs in the group, (10). It is recommended that the geometric mean of at least three RGs be used as a normalization factor (10). NormFinder allows direct estimation of expression variation of RGs using a model-based approach (11). These packages are powerful tools for selecting RGs based on expression stability.

However, in our experience normalization does not always reduce variation, even when the least variable RG is chosen. In this study, we investigated if and when data normalization could reduce variation by computing and comparing relative statistical efficiency (RSE), a ratio of two variances (variance with and without normalization), under different choices of RGs. The RSE not only provides a direct comparison between two variances but is also a good indicator of statistical power in qPCR data analysis (12).

Our data are from the interim analysis of a Phase I cancer prevention study (NCT01465776; http://clinicaltrials.gov) in which black raspberries (BRBs) were administered to oral cancer patients and molecular biomarkers of oral carcinogenesis were assessed for transcriptional changes (13). Pre-validated TaqMan Gene Expression Assays and Gene Expression Master Mix (Life Technologies Applied Biosystems; Grand Island, NY) were used to generate triplicate qRT-PCR data using an Applied Biosystems 7500 System and SDS software following the manufacturer's instructions. The goal of the study is to identify target genes by comparing the mRNA expression of 12 genes in tumor tissues before treatment and after treatment. Three different putative RGs were included in the qPCR experiments for normalization.

We computed estimates of RSE for the treatment effect (i.e., the change in mean expression) of BRBs, as:

$$RSE\,(Normalized\,Data, Raw\,Data) = \frac{Var_R}{Var_N} \qquad \text{[Eq. 1]}$$

where Var_R and Var_N denote the variance of the difference in mean expression (posttreatment mean – pretreatment mean) calculated from the raw and normalized data, respectively. RSEs were calculated for each RG and a normalization factor equal to the geometric mean

REPRINT WITH PERMISSION ONLY

Method summary:

Choosing the most stably expressed RG for data normalization does not guarantee reduced variance or improved statistical efficiency. We provide a formula for determining when data normalization will improve statistical efficiency and hence increase the power of statistical tests in data analysis.

Table 1. Relative statistical efficiencies: normalized Cq relative to raw Cq.

Gene ID	DUSP1	GAPDH	18S	NF[a]
1	5.08	10.40	11.98	16.64
2	0.26	0.12	0.16	0.16
3	3.98	1.44	2.19	2.33
4	16.06	9.75	6.36	14.07
5	1.80	2.16	3.28	2.77
6	2.62	2.13	2.87	2.93
7	14.24	6.09	3.99	7.36
8	8.66	5.48	9.49	14.02
9	3.46	1.40	1.98	2.10
10	3.76	6.75	6.00	7.22
11	0.78	0.32	0.37	0.39
12	4.99	4.95	4.94	6.56

[a]NF: normalization factor; calculated as the geometric mean of DUSP1, GAPDH, and 18S.

of the three RGs. An RSE >1 indicates that variance was reduced (i.e., statistical efficiency was improved) through normalization.

The estimated RSEs are summarized in Table 1. RSEs were >1 for all genes except genes 2 and 11. This indicates that normalization reduced variation (or improved statistical efficiency) for most of the genes. Interestingly, for genes 2 and 11 all RSEs were <1, which means that the variances of the differences in the estimated means increased after normalization ($Var_N > Var_R$). This unusual finding is due to the following property:

Property 1: Normalization increases the variance of the estimated treatment effect if

$$\rho_{ij} < \frac{1}{2}\sqrt{\frac{Var(H_j)}{Var(X_i)}} = P_{ij} \quad \text{[Eq. 2]}$$

where X_i is the raw Cq value for gene of interest i and H_j is the raw Cq value for RG j, ρ_{ij} is the correlation between X_i and H_j, and $Var(X_i)$ and $Var(H_j)$ are the variances of the raw Cq values for the gene of interest and the RG, respectively (see Supplementary Material for proof).

In Table 2, estimates of the correlations ρ_{ij}, the variances of the raw differences, and the P_{ij} values are provided for one RG (*DUSP1*). Not surprisingly, Equation 2 was satisfied only by genes 2 and 11 due to the small variances of the raw Cq values of these genes. The correlation ρ_{ij} was smaller than P_{ij} only for genes 2 and 11.

Our results show that, for the most part, normalization using RGs decreased the variance of the average change in expression and thus increased statistical efficiency and statistical power. However, the results for genes 2 and 11 demonstrated that when Property 1 holds, normalization can increase the variance of the estimated treatment effect and hence reduce statistical efficiency. This occurred even with the most stable RG in this study, *18S*, which was expressed with the least variability across experimental conditions. In addition, using the geometric mean of the three RGs as a normalization factor did not solve the problem of increased variance (Table 1). Therefore, Property 1 needs to be considered both in the design of the experiment (14) and when selecting RGs, since under certain conditions normalization could increase variation and lead to decreased statistical efficiency (or decreased statistical power).

To our knowledge, this is the first study to examine the effects of the choice of RGs on statistical efficiency. Furthermore, we present a formula for determining when normalization will reduce variance, resulting in improved statistical efficiency on an individual gene investigational basis. It is clear that the correlations between and variances of the expression levels of the RGs and the target genes need to be considered, since an unfortunate combination of these parameters can lead to increased variation, making normalization pointless. Therefore, in addition to selecting RGs based on their expression stability, it is also important to examine Property 1 (which can be easily computed using statistical software or in Microsoft Excel) when selecting RGs to avoid losing statistical power in qPCR data analysis.

Table 2. A demonstration of Property 1 using DUSP1.

Gene ID	ρ_{ij}	P_{ij}	Var_R[a]
1	0.90	0.44	34.39
2	0.87	1.39	3.43
3	0.94	0.66	15.29
4	0.97	0.50	26.94
5	0.67	0.35	53.89
6	0.82	0.54	23.09
7	0.97	0.46	31.88
8	0.94	0.49	27.52
9	0.93	0.65	15.54
10	0.86	0.41	39.78
11	0.88	0.96	7.28
12	0.90	0.41	40.47

[a]Var_R: variance of the difference in raw mean expression.

Author contributions

YG, MP, and SF contributed to the conception of the study. YG and MP were responsible for design of the study, conceived the analyses, and interpreted the findings. DP contributed to the design of the PCR experiment and assisted in the writing. TK performed the PCR experiment and assisted in the writing. CW contributed to the design of the PCR experiment and assisted in the writing.

Acknowledgments

We wish to thank David Jarjoura of the Ohio State University Center for Biostatistics for helpful comments that improved this paper. This work was supported by NIH/NCI grant 5R01CA127368. This paper is subject to the NIH Public Access Policy.

Competing interests

The authors declare no competing interests.

References

1. **Bustin, S.A. and T. Nolan.** 2004. Pitfalls of quantitative real-time reverse transcription polymerase chain reaction. J. Biomol. Tech. *15*:155-166.
2. **Huggett, J., K. Dheda, S.A. Bustin, and A. Zumla.** 2005. Real-time RT-PCR normalisation; strategies and considerations. Genes Immun. *6*:279-284.
3. **Dheda, K., J.F. Huggett, S.A. Bustin, M.A. Johnson, G. Rook, and A. Zumla.** 2004. Validation of housekeeping genes for normalizing RNA expression in real-time PCR. Biotechniques *37*:112-119.
4. **Tricarico, C., P. Pinzani, S. Bianchi, M. Paglierani, V. Distante, M. Pazzagli, S.A. Bustin, and C. Orlando.** 2002. Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. Anal. Biochem. *309*:293-300.
5. **Ullmannová, V. and C. Haskovec.** 2003. The use of housekeeping genes (HKG) as an internal control for the detection of gene expression by quantitative real-time RT-PCR. Folia Biol. (Praha) *49*:211-216.
6. **Koch, I., R. Weil, R. Wolbold, J. Brockmöller, E. Hustert, O. Burk, A. Nuessler, P. Neuhaus, et al.** 2002. Interindividual variability and tissue-specificity in the expression of cytochrome P450 3A mRNA. Drug Metab. Dispos. *30*:1108-1114.
7. **Bär, M., D. Bär, and B. Lehmann.** 2009. Selection and validation of candidate housekeeping genes for studies of human keratinocytes--review and recommendations. J. Invest. Dermatol. *129*:535-537.
8. **de Jonge, H.J., R.S. Fehrmann, E.S. de Bont, R.M. Hofstra, F. Gerbens, W.A. Kamps, E.G. de Vries, A.G. van der Zee, et al.** 2007. Evidence based selection of housekeeping genes. PLoS ONE *2*:e898.
9. **Cheng, W.C., C.W. Chang, C.R. Chen, M.L. Tsai, W.Y. Shu, C.Y. Li, and I.C. Hsu.** 2011. Identification of reference genes across physiological states for qRT-PCR through microarray meta-analysis. PLoS ONE *6*:e17347.
10. **Vandesompele J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, F. Speleman.** 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3(7):RESEARCH0034.
11. **Andersen, C.L., J.L. Jensen, and T.F. Orntoft.** 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. *64*:5245-5250.
12. **Fleiss, J.L.** 1986. Design and Analysis of Clinical Experiments. John Wiley & Sons, New York.
13. **Stoner, G.D., L.S. Wang, and B.C. Casto.** 2008. Laboratory and clinical studies of cancer chemoprevention by antioxidants in berries. Carcinogenesis *29*:1665-1674.
14. **Tempelman, R.J.** 2005. Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet. Immunol. Immunopathol. *105*:175-186.

Received 30 June 2011; accepted 12 August 2013.

Address correspondence to Yi Guo, Department of Health Outcomes and Policy, College of Medicine, University of Florida. E-mail: yiguo@ufl.edu

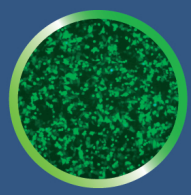
To purchase reprints of this article, contact: biotechniques@fosterprinting.com

Easy Way
for
Your Life Science Research!

Direct PCR Kit

- No DNA Purification
- No Proteinase K
- No DNA Contamination
- Save Time and Cost
- Direct PCR from Blood, Cell, Mouse tail, Tissue or Plant

Transfection Reagent



- Cationic lipid-based Reagent
- Low Cellular Toxicity
- High Efficiency
- Fast and easy to use

Cytokine ELISA Kit

BD	IFN-gamma	Resistin
BMP	IP-10	sRANKL
CNTF	Interleukin	SCF
Eotaxin	Leptin	TRAIL
EGF	MCP-1	TGF
FGF	MIP-1	TNF
G-CSF	NGF-beta	TIMP
GM-CSF	PDGF-BB	VEGF
HGF	RANTES	

Pre-Cast Gels

- Long Life EzWay™ Gel
- Speedy EzWay™ Quick Gel
- Tris-Glycine Gel
- Tricine Gel
- IEF Gel
- Zymogram Gel
- TBE Gel

biotechnology
KOMABIOTECH

for more information