



RefFinder: a web-based tool for comprehensively analyzing and identifying reference genes

Fuliang Xie¹ · Jianyu Wang² · Baohong Zhang¹

Received: 16 March 2023 / Revised: 7 April 2023 / Accepted: 10 April 2023 / Published online: 15 April 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Although many genes may serve as reference genes, they may cause different expression patterns by selecting different reference genes because no single gene is expressed consistently in all tested tissues of an organism under all environmental and developmental conditions. Thus, it is becoming increasingly important and necessary to identify suitable reference genes before performing gene expression analysis. Currently, there are several computational tools available for evaluating the stability of candidate reference genes. These tools are based on different statistical algorithms and may produce different rankings in stability within the same reference gene study. To date, the RefFinder is the only web-based tool available for comparing and evaluating housekeeping genes as candidates to be reference genes. In this tool, we integrated the four currently available computational programs (geNorm, NormFinder, BestKeeper, and the comparative ΔCt method) into a web-based tool for evaluating the stability and reliability of reference genes. According to the gene stability rankings derived from the four programs, we assigned an appropriate weight to each gene and calculated the geometric mean of weights for the final rankings. Aside from the overall ranking, a single program or combination of the four programs can be selected for evaluating the ranking of candidate reference genes. This tool has been widely used and validated by many research laboratories around the world. You may use this tool at <http://www.heartcure.com.au/reffinder/> or <https://blooge.cn/RefFinder/>. You can also download this algorithm program from <https://github.com/fulxie/RefFinder> and setup on your own computer. RefFinder is developed by PHP. Users can deploy it to a Php-based server (Apache + PHP) and run it.

Keywords Reference gene · Web-based tool · qRT-PCR · Gene expression

Introduction

Gene expression analysis is becoming more and more important for current biological and biomedical research. No matter what method is employed, reference genes always provide evidence that the same amount of total gene products (mRNAs or proteins) is loaded in each sample. As an optimal reference gene, its expression level should be consistent under any possible condition. Because of the nature of genes, many housekeeping genes, such as genes that encode actins, cyclophilins, elongation factor-1 α (EF-1 α), glyceraldehyde-3-phosphate dehydrogenase (GAPD),

microglobulins, ribosomal units (18S or 28S rRNA), ubiquitin (UBQ), and tubulins, have been frequently selected as reference genes (Pfaffl et al. 2004). These genes are typically needed for the maintenance of the cell structure and/or function; their expression levels are relatively stable compared to tissue-specific genes. However, recent studies have demonstrated that the transcription levels of housekeeping genes also vary in different cell types and developmental stages as well as being affected by different experimental conditions including varying treatments (Ahn et al. 2008; Jordan and Wilson 2004; Yu et al. 2020). This suggests that no perfect reference gene exists in living cells. Thus, it is necessary to evaluate and select the best reference gene(s) for an experiment before performing gene expression analysis.

To evaluate and screen reference genes for gene expression analysis, several computational programs have been developed in the past 15 years. These programs include geNorm (Vandesompele et al. 2002), NormFinder (Andersen et al. 2004), BestKeeper (Pfaffl et al. 2004), and

✉ Baohong Zhang
zhangb@ecu.edu

¹ Department of Biology, East Carolina University, Greenville, NC 27858, USA

² Guangxi University of Finance and Economics, Nanning 530031, Guangxi, China

the comparative Δ Ct method (Silver et al. 2006). Different computational programs rank the stability of housekeeping genes based on statistic endpoints and algorithms, which causes the stability score and ranking to vary among these programs. As a result, these programs may give a different ranking than others and it is hard to delineate which ranking is the best (Ahn et al. 2008). A web-based tool for comparing and analyzing the stability of housekeeping genes will significantly enhance reference gene-related studies and further enhance gene expression profiling and functional studies by using quantitative real-time PCR (qRT-PCR) and/or Northern blotting.

Currently, only a single online resource exists for comparing and evaluating housekeeping genes as candidates to be reference genes—the RefFinder tool. In this study, we describe this user-friendly tool for evaluating and screening reference genes from extensive experimental datasets. This tool integrates the currently available four major computational programs (geNorm, NormFinder, BestKeeper, and the comparative Δ Ct method) to compare and rank the tested candidate reference genes. Based on the rankings from each program, we assigned an appropriate weight to an individual gene and calculated the geometric mean of their weights for the overall final ranking.

Methods and program design

Following the introduction of algorithms as described in the four computational programs (geNorm (Vandesompele et al. 2002), Normfinder (Andersen et al. 2004), BestKeeper (Pfaffl et al. 2004), and the comparative Δ Ct method (Silver et al. 2006)), we rewrote the algorithm program in PHP and then integrated them to the web. To make use more convenient, only the original Ct value from quantitative real-time PCR (qRT-PCR) is required for input on the web page (Fig. 1). Users only need to click the “submit” button and then all of the results from the four algorithms will be generated (Fig. 2). Finally, based on the rankings from each of the four algorithms, we developed a simple algorithm to present an overall ranking for testing reference genes. First,

each tested candidate gene was assigned a number from each of the four computational programs based on the stability justified by that program. The number assigned will be from 1 to N (N is the total number of tested genes) according to the rank of reference genes based on that program; one is the most stable gene according to that program, and N is the least stable gene. Then, our developed tool (RefFinder) will automatically calculate the geometric mean of each gene weight across the four algorithms and will finally re-rank these candidate reference genes based on the geometric mean of each gene. The gene with the least geometric mean is viewed as the more stable reference gene, while the gene with higher geometric mean is the less stable reference gene.

Results

Data input and gene rank

The integrated tool for analyzing reference gene expression offers users an easy-to-use interface. Users only need to copy their original Ct values from an Excel file (can be generated and exported directly from qRT-PCR) to the input box and then submit the data. Users are also allowed to input data manually according to the required data format of the tool. A comprehensive overall ranking as well as the four individual rankings from each of the four algorithms will be generated immediately (Fig. 2). To help the user understand and run this web-based tool, we also added a real dataset in this program for testing the run; this dataset was obtained from a previous study (Chen et al. 2011), in which 10 housekeeping genes (five protein-coding genes and noncoding genes) were selected to test their stability under different treatments in human breast cancer cell line MCF-7. The users just need to click the “try example,” and then they will see the raw data and analyzed results based on this program.

Features of the four tools

The Excel-based approach, geNorm can be used to determine the most stable reference genes from a set of candidate reference genes in a given cDNA sample panel (Vandesompele

Fig. 1 The interface of the integrated tool for analyzing reference gene expression

Evaluating Reference Genes Expression>>

Input your data

21.81065	24.6102	26.5362	23.36915	27.01725	18.1465	14.4691	20.0296	17.003	19.4468
21.1704	24.0964	26.02375	23.5005	26.0287	17.7986	15.0001	19.6619	16.43175	19.5778
23.4701	25.95015	27.0499	24.54845	28.30655	18.60915	16.04265	20.5171	17.3307	20.03005
19.27045	23.49115	25.0835	22.84805	24.67245	17.7206	14.336	19.8189	16.5204	19.30995
19.0253	22.8714	24.69045	22.7619	24.47635	17.8875	14.47215	19.87185	16.61655	20.05875
19.16015	22.9632	24.68925	22.5935	24.49845	18.026	14.72145	19.98605	16.76375	20.56225
20.23935	24.2292	25.4872	23.1425	25.45795	17.62315	14.73475	19.68395	16.3622	20.12155
20.6476	23.9726	25.84975	23.4667	25.92005	17.91115	15.0755	19.7871	16.47465	20.0937
22.8857	26.0722	27.2926	24.5212	27.9778	17.6749	15.2755	19.76915	16.386	20.35435
19.96615	22.7419	25.27745	22.9304	25.04025	18.04825	14.99655	20.29905	16.9748	20.3836
20.0786	22.61245	25.4461	22.79935	24.9942	17.74855	14.5316	20.155	16.67935	20.22445
20.7771	23.82425	25.7362	22.70535	25.11675	16.88815	13.50115	19.1055	15.6059	18.39635
21.58675	23.7839	26.3449	23.28645	26.0738	18.09565	15.0952	20.4421	17.02225	20.12955
22.15435	24.16015	26.665	23.533	26.52845	17.21855	14.51215	19.70135	16.02825	18.68725
24.07285	26.44245	27.4036	24.6452	29.01625	18.28	15.592	20.3794	16.7971	20.24645

Analyze Try example remove data ?

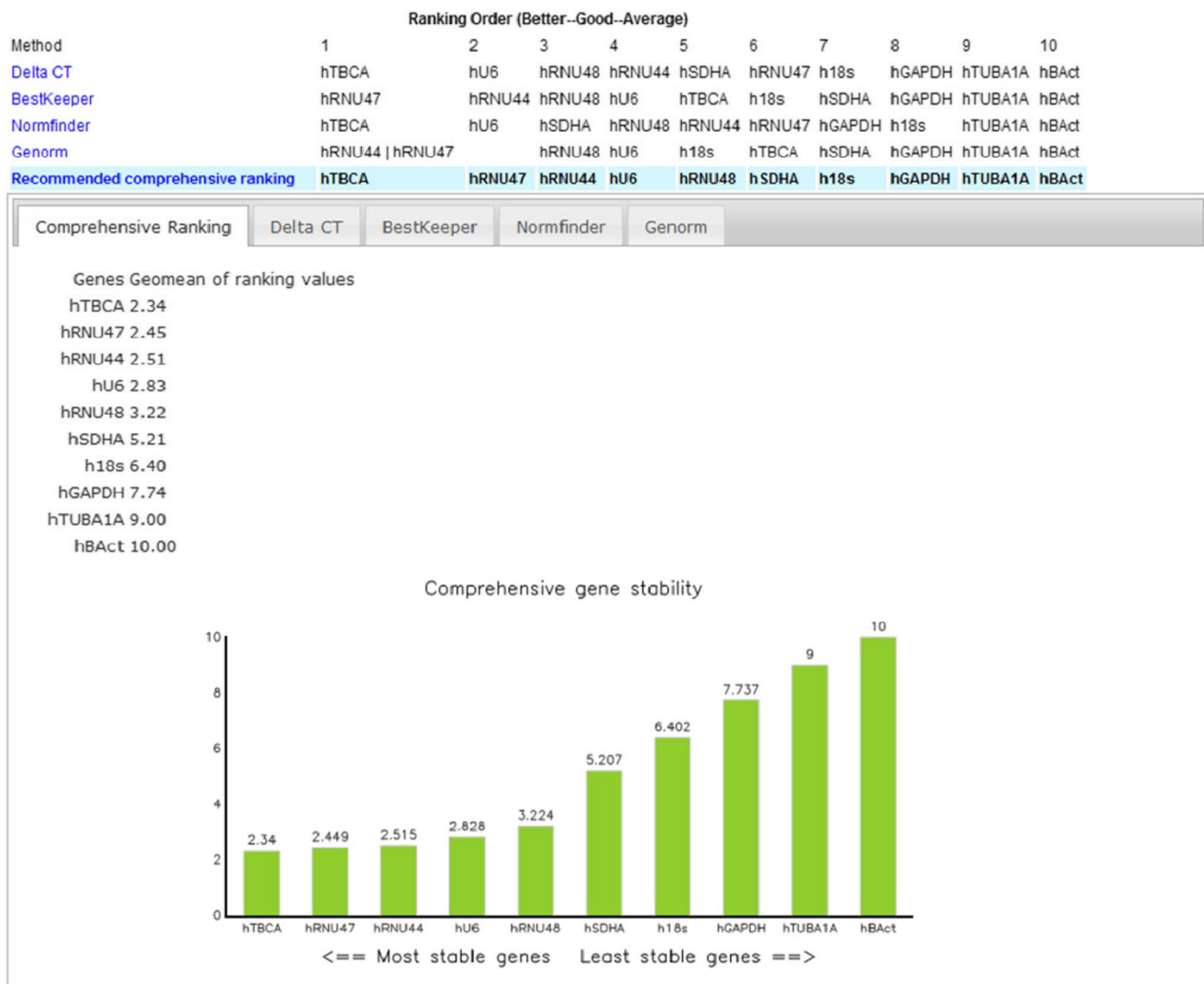


Fig. 2 Results of analyzing reference gene expression by using the integrated tool

et al. 2002). By calculating the average pairwise variation of a particular gene with all other genes, all candidate reference genes are ranked based on average expression stability value (M value) from the most stable to the most variable genes. Stepwise exclusion of the least stable gene with the highest M value is iteratively performed to rank the tested genes until the two most stable genes cannot be further ranked (Vandesompele et al. 2002). NormFinder employed a model-based strategy to identify stable expressed genes among a set of candidate normalization genes (Andersen et al. 2004). Based on a mathematical model of gene expression, NormFinder estimates not only the overall variation of the candidate reference genes but also the variation between subgroups of samples. System error induced when using a gene can also be evaluated, due to the fact that NormFinder allows a direct measure for the estimated expression variation (Andersen et al. 2004). Unlike geNorm, NormFinder

assesses the expression stability of each candidate independently. BestKeeper is also an Excel-based tool which determines the best suited reference genes out of candidate genes using pairwise correlation analysis of candidate reference genes (Pfaffl et al. 2004). Correlations of the expression levels of all candidate genes are firstly estimated by BestKeeper and then the highly correlated ones are combined into an index. Finally, three indicators including standard deviation, percent covariance, and power of the candidates, are calculated to help users determine the best reference genes (Pfaffl et al. 2004). The comparative ΔCt method avoids the need to accurately quantify input RNA and instead uses ΔCt comparisons between genes (Silver et al. 2006). Using an algorithm similar to geNorm, the comparative ΔCt method assesses the most stable reference genes by comparing the relative expression of “pairs of genes” within each tissue sample or each treatment. The mean of standard deviation

values derived from a comparison between a particular reference gene and any other candidate is calculated as the gene stability indicator. The lower the arithmetic mean, the more stable the gene (Silver et al. 2006).

Although these four tools are widely used in the selection of reference genes, to the best of our knowledge, there is no report of comparison on these four tools (Xie et al. 2012). It is hard to determine which one is the best due to the individual specific algorithm. Using assigned weights on each reference gene according to ranking from each of the four programs, we calculated the geometric mean of the weights from each ranking and made a comprehensive ranking from a general point of view. More importantly, we provided a web-based tool for public use, that will significantly enhance the studies of gene expression and functional analysis.

Discussion

Gene expression analysis is one of the most popular experiments which is widely used in many fundamental and applied biological and biomedical research, including genetics and toxicology studies, particularly during the past-genome sequencing era (Liu and Zhang 2022). During gene expression analysis, in almost all studies, an internal standard, termed a reference gene, is needed to normalize mRNA levels between different samples. The mRNA levels can be altered due to several factors, including pipetting errors and RNA quality. Unlike the PCR efficacy that will cause the fold changes to differ, selecting the wrong reference gene may result in a wrong conclusion. Thus, a good reference gene should be expressed at a consistent level and would not be altered among different developmental and environmental conditions. To meet this criterion, an appropriate reference gene is generally involved in regulating basic cell function (such as transcription elongation factor) and/or maintaining cell structures (such as actin); these genes are commonly referred to as housekeeping genes. Unfortunately, there is no perfect reference gene because the expression profiles of all genes are affected by developmental or environmental changes; the only difference is that some changes are large, and some changes are small. Thus, screening the most reliable reference genes has become an important study when performing a gene expression analysis. The common methods and experimental steps for screening and selecting an appropriate reference gene are (1) selecting housekeep genes pool (usually 10–16 genes) and designing primers following the primer designer criteria, such as avoiding four nucleotides in a row, no primer dimer or hairpin structure; (2) treating the samples at different environmental and developmental conditions; (3) running RT-PCR and qRT-PCR. During this procedure, you need to be very careful to add the same amount of mRNAs as possible because of no reference gene as an

internal standard; and (4) comparing each candidate gene at different conditions and selecting the most stable gene as the reference gene for the following study. Among all these steps, comparing the stability of each gene may be the most difficult part because of qRT-PCR generating so much raw data. To solve this problem, several laboratories have developed several computation programs to compare each candidate gene. Among these programs, geNorm (Vandesompele et al. 2002), NormFinder (Andersen et al. 2004), BestKeeper (Pfaffl et al. 2004), and the comparative Δ Ct method (Silver et al. 2006) are four frequently used programs for studying reference genes. These four programs rank a reference gene based on different statistic endpoints and algorithms. Thus, each one has different advantages and may rank the reference genes with a slight difference. To solve this problem and allow people better rank each individual reference gene, we develop a new computational program called RefFinder. This is the first program to integrate the four most commonly used computational programs to rank a set of reference genes. Thus, the rank of reference genes is more reliable. More importantly, RefFinder is designed on a web platform and is easier to use for people without any bioinformatics skills. Currently, RefFinder has been widely used and validated by hundreds of research laboratories around the world for studying, comparing, and screening the most reliable reference genes for various research purposes (Bansal et al. 2015; Hazarika et al. 2023; Kochhar et al. 2022; Taki et al. 2014; Taki and Zhang 2013; Wang et al. 2013; Zhang et al. 2012, 2023).

Acknowledgements We would like to thank the authors of the four programs (geNorm, Normfinder, BestKeeper, and the comparative Δ Ct method) for their agreement to allow us usage of their programs. We also greatly appreciate various web servers, such as <http://www.heartcure.com.au/reffinder/> and <https://blooge.cn/RefFinder/>, to host the web tool for public users. RefFinder can be downloaded from <https://github.com/fulxie/RefFinder> and set up on your own computer and/or your own website.

Authors' contributions Fuliang Xie designed the web tool and wrote the manuscript; Jianyu Wang performed the data analysis; Baohong Zhang formed the concepts and revised the manuscript. All authors approved this manuscript.

Data Availability Not applicable

Declarations

Conflict of interest The authors declare no competing interests.

References

- Ahn K, Huh JW, Park SJ, Kim DS, Ha HS, Kim YJ, Lee JR, Chang KT, Kim HS (2008) Selection of internal reference genes for SYBR green qRT-PCR studies of rhesus monkey (*Macaca mulatta*) tissues. *BMC Mol Biol* 9:78

- Andersen CL, Jensen JL, Orntoft TF (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
- Bansal R, Mittapelly P, Cassone BJ, Mamidala P, Redinbaugh MG, Michel A (2015) Recommended reference genes for quantitative PCR analysis in soybean have variable stabilities during diverse biotic stresses. *PLoS one* 10:e0134890
- Chen D, Pan X, Xiao P, Farwell MA, Zhang B (2011) Evaluation and identification of reliable reference genes for pharmacogenomics, toxicogenomics, and small RNA expression analysis. *J Cell Physiol* 226:2469–2477
- Hazarika A, Nongkhlaw B, Mukhopadhyay A (2023) Identification of stable reference genes in peripheral blood mononuclear cells from type 2 diabetes mellitus patients. *Sci rep* 13:486
- Jordan MA, Wilson L (2004) Microtubules as a target for anticancer drugs. *Nat Rev Cancer* 4:253–265
- Kochhar P, Dwarkanath P, Ravikumar G, Thomas A, Crasta J, Thomas T, Kurpad AV, Mukhopadhyay A (2022) Placental expression of RNU44, RNU48 and miR-16-5p: stability and relations with fetoplacental growth. *Eur J Clin Nutr* 76:722–729
- Liu W, Zhang B (2022) The landscape of genome sequencing and assembling in plants. *Funct Integr Genom* 22:1147–1152
- Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. *Biotechnol Lett* 26:509–515
- Silver N, Best S, Jiang J, Thein SL (2006) Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol Biol* 7:33
- Taki FA, Zhang B (2013) Determination of reliable reference genes for multi-generational gene expression analysis on *C. elegans* exposed to abused drug nicotine. *Psychopharmacology* 230:77–88
- Taki FA, Abdel-Rahman AA, Zhang B (2014) A comprehensive approach to identify reliable reference gene candidates to investigate the link between alcoholism and endocrinology in Sprague-Dawley rats. *PLoS One* 9:e94311
- Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paep A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 3:RESEARCH0034
- Wang M, Wang Q, Zhang B (2013) Evaluation and selection of reliable reference genes for gene expression under abiotic stress in cotton (*Gossypium hirsutum* L.). *Gene* 530:44–50
- Xie F, Xiao P, Chen D, Xu L, Zhang B (2012) miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol Biol* 80(1):75–84. <https://doi.org/10.1007/s11103-012-9885-2>
- Yu J, Su Y, Sun J, Liu J, Li Z, Zhang B (2020) Selection of stable reference genes for gene expression analysis in sweet potato (*Ipomoea batatas* L.). *Mol and Cell Probes* 53:101610
- Zhang Y, Chen D, Smith MA, Zhang B, Pan X (2012) Selection of reliable reference genes in *Caenorhabditis elegans* for analysis of nanotoxicity. *PLoS One* 7:e31849
- Zhang G, Zhang Z, Wan Q, Zhou H, Jiao M, Zheng H, Lu Y, Rao S, Wu G, Chen J, Yan F, Peng J, Wu J (2023) Selection and validation of reference genes for RT-qPCR analysis of gene expression in *Nicotiana benthamiana* upon single infections by 11 positive-sense single-stranded RNA viruses from Four Genera. *Plants* 12:857

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.