frontiers in Genetics

ORIGINAL RESEARCH
published: 01 March 2019
doi: 10.3389/fgene.2019.00097

Pan-Cancer Analysis of TCGA Data Revealed Promising Reference Genes for qPCR Normalization

George S. Krasnov, Anna V. Kudryavtseva, Anastasiya V. Snezhkina, Valentina A. Lakunina, Artemy D. Beniaminov, Nataliya V. Melnikova and Alexey A. Dmitriev**

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

OPEN ACCESS

Edited by:
Yuriy L. Orlov, Institute of Cytology and Genetics (RAS), Russia

Reviewed by:
Alexey V. Pindyurin, Institute of Molecular and Cellular Biology (RAS), Russia
Vladimir Kiselev, Wellcome Trust Sanger Institute (WT), United Kingdom
Shengjie Yang, NorthShore University HealthSystem, United States

***Correspondence:**
George S. Krasnov
gskrasnov@mail.ru
Alexey A. Dmitriev
alex_245@mail.ru

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *31 October 2018*
Accepted: *29 January 2019*
Published: *01 March 2019*

Citation:
Krasnov GS, Kudryavtseva AV, Snezhkina AV, Lakunina VA, Beniaminov AD, Melnikova NV and Dmitriev AA (2019) Pan-Cancer Analysis of TCGA Data Revealed Promising Reference Genes for qPCR Normalization. Front. Genet. 10:97. doi: 10.3389/fgene.2019.00097

Quantitative PCR (qPCR) remains the most widely used technique for gene expression evaluation. Obtaining reliable data using this method requires reference genes (RGs) with stable mRNA level under experimental conditions. This issue is especially crucial in cancer studies because each tumor has a unique molecular portrait. The Cancer Genome Atlas (TCGA) project provides RNA-Seq data for thousands of samples corresponding to dozens of cancers and presents the basis for assessment of the suitability of genes as reference ones for qPCR data normalization. Using TCGA RNA-Seq data and previously developed CrossHub tool, we evaluated mRNA level of 32 traditionally used RGs in 12 cancer types, including those of lung, breast, prostate, kidney, and colon. We developed an 11-component scoring system for the assessment of gene expression stability. Among the 32 genes, *PUM1* was one of the most stably expressed in the majority of examined cancers, whereas *GAPDH*, which is widely used as a RG, showed significant mRNA level alterations in more than a half of cases. For each of 12 cancer types, we suggested a pair of genes that are the most suitable for use as reference ones. These genes are characterized by high expression stability and absence of correlation between their mRNA levels. Next, the scoring system was expanded with several features of a gene: mutation rate, number of transcript isoforms and pseudogenes, participation in cancer-related processes on the basis of Gene Ontology, and mentions in PubMed-indexed articles. All the genes covered by RNA-Seq data in TCGA were analyzed using the expanded scoring system that allowed us to reveal novel promising RGs for each examined cancer type and identify several "universal" pan-cancer RG candidates, including *SF3A1*, *CIAO1*, and *SFRS4*. The choice of RGs is the basis for precise gene expression evaluation by qPCR. Here, we suggested optimal pairs of traditionally used RGs for 12 cancer types and identified novel promising RGs that demonstrate high expression stability and other features of reliable and convenient RGs (high expression level, low mutation rate, non-involvement in cancer-related processes, single transcript isoform, and absence of pseudogenes).

Keywords: cancer, gene expression, reference genes, quantitative PCR, data normalization, RNA-Seq, TCGA, CrossHub

INTRODUCTION

Quantitative PCR (qPCR) is the most widely used technique for quantification of gene expression. qPCR is rapid, has a very high dynamic range of mRNA level quantification and provides a measurement of even small gene expression alterations in a large number of samples. The most common and convenient approach for qPCR data normalization assumes mRNA quantification of a reference gene (RG) with stable expression level between the samples under study (Huggett et al., 2005). It is a bottleneck of qPCR, and the reliability of qPCR results strongly depends on the selection of appropriate RGs. This issue becomes more acute when it comes to assessing the moderate changes in the mRNA level of target genes (<2-fold).

The problem of selecting appropriate RGs is especially crucial in cancer studies because of the presence of several molecular subtypes within a histological type and, moreover, a unique molecular portrait of each tumor (Janssens et al., 2004). Despite the fact that almost 30 years have passed since the moment when the issue of picking appropriate RGs had arisen, there is still no consensus (Janssens et al., 2004; Rubie et al., 2005; Gur-Dedeoglu et al., 2009; Ibusuki et al., 2013; Zhao et al., 2018). Many studies indicate that most frequently used RGs (*GAPDH*, *ACTB*, *B2M*, *etc.*) have a wide but limited field of applicability: they should not be illegibly used for a wide spectrum of diseases or stress conditions (Barber et al., 2005; Rubie et al., 2005; Kozera and Rapacz, 2013; Chapman and Waldenstrom, 2015). To increase the reliability of qPCR data, one should use at least two or more RGs that are not co-expressed with each other (Chapman and Waldenstrom, 2015). The most rigorous approach is to analyze a panel of 5–20 RGs and choose those with the most stable expression for a current study. Several tools have been developed for these purposes: geNorm (Vandesompele et al., 2002), NormFinder (Andersen et al., 2004), BestKeeper (Pfaffl et al., 2004). However, the vast part of researchers do not perform the analysis of RG suitability and just rely on the existing literature data concerning the object of study (Chapman and Waldenstrom, 2015).

Whole-transcriptomic data allow us to look at the problem from the other side. RNA-Seq opens up great opportunities for a complex expression analysis and identifying trends in the mRNA level changes of groups of genes between the samples. RNA-Seq data are free of bias that comes from the instability of RG expression. The most common RNA-Seq data normalization strategy is based on the assumption that the mRNA level of the majority of genes is stable. This method is implemented in popular RNA-Seq differential expression analysis packages, including edgeR [trimmed mean of M-values method, TMM; Robinson et al., 2010], DESeq2 (Love et al., 2014), and others. There are other normalization strategies: by total read count, by upper quartile or median values, FPKM/RPKM, TPM, "remove unwanted variation" (RUV) (Risso et al., 2014); as well as machine-learning approaches: RNA-Seq by Expectation-Maximization (RSEM) (Li and Dewey, 2011) and Sailfish (Patro et al., 2014). Despite the diversity of the methods, in most cases, they give rather similar results, which differ by 20–30%, with the exception of some cases when the expression of half or more of genes is changed significantly (Dillies et al., 2013; Li et al., 2015; Zyprych-Walczak et al., 2015; Evans et al., 2018).

Analysis of highly representative RNA-Seq and microarray datasets is very attractive in terms of the identifying stably expressed RGs for human (Popovici et al., 2009; Tilli et al., 2016; Chen et al., 2017; Chim et al., 2017; Hoang et al., 2017) or other organisms (Alexander et al., 2012; Carmona et al., 2017; Zhou et al., 2017). This approach is valuable for the detection of novel housekeeping gene candidates with constitutively stable mRNA level.

In 2016, Tilli et al. suggested a strategy including the large-scale screening of potential RGs from RNA-Seq data with further validation by qPCR and applied it for breast cancer (Tilli et al., 2016). The authors analyzed datasets of The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) and found that several non-traditional RGs, *CCSER2, SYMPK, ANKRD17*, as well as known RG *PUM1* demonstrated the least expression variability in breast cancer samples and normal tissues (Tilli et al., 2016). The similar approach was realized by Chen et al. for the identification of reference mRNA and miRNA suitable for human esophageal squamous cell carcinoma studies (Chen et al., 2017). It allowed authors to identify non-standard RG candidates—*DDX5*, *LAPTM4A*, *P4HB*, and *RHOA*.

TCGA is the largest resource in the field of cancer biology that is aimed at the discovery of the molecular features of various cancer types (https://cancergenome.nih.gov/). TCGA database includes genomic, transcriptomic, and epigenetic data for 33 human cancer types represented with more than 11,000 individual samples. In the present work, we analyzed TCGA transcriptome sequencing data in order to evaluate the expression stability of widely used RGs and identify novel RG candidates in 12 most common cancer types. The use of representative TCGA sample sets allows us to pay extra attention to the overall stability of mRNA level and presence of outliers, the cases of dramatic expression "blow up" or falling down in single samples. Besides the data on mRNA level, we took into account if this is a well-studied gene or not (by evaluating the number of mentions in PubMed-indexed titles/abstracts), if a gene is involved in cancer-associated biological processes like cell cycle, differentiation, and adhesion (using Gene Ontology). Additionally, we evaluated if a gene is highly mutated (using TCGA data on somatic mutations) that indicates its implication in cancer. Also, we tried to minimize the number of pseudogenes and alternatively spliced transcripts in order to improve usability: the presence of pseudogenes makes it difficult to pick up cDNA-specific primer pairs, and the presence of alternative transcripts complicates the expression analysis and may lead to flawed results. We integrated all the parameters listed above into a single scoring system. Finally, we looked for genes that demonstrate cross-tissue expression stability and may represent "universal" pan-cancer RGs.

MATERIALS AND METHODS

In the present work, we focused on TCGA data for 12 cancer types for which RNA-Seq data were available for representative

sample sets: at least 100 tumor (T) and 20 normal (N) tissue samples. The data were processed with a modified version of CrossHub (Krasnov et al., 2016), a tool for the multi-way analysis of TCGA transcriptomic and genomic data. Read counts data were downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/) and normalized using the TMM method and then recalculated for 1 million library size. The derived CPM (read counts per million) values were used as a measure of mRNA level of a gene for further expression stability analysis.

In order to assess gene expression stability, we developed a scoring system, which included several components (S_i) responsible for T-N expression level difference, expression level stability within pools of N and T samples, and correlations of mRNA level with clinical and pathological characteristics [disease stage, TNM (tumor, node, metastasis) classification, follow-up status]. Each scoring component S_i takes values from 0 to 100. All S_i are taken with different weights (W_i), which reflect the importance of component. Overall expression scoring S^{Exp} is calculated as follows:

$$S^{\mathrm{Exp}} = \left(\prod_{i=1}^{N} (S_i + \mathrm{CA}_i)^{\mathrm{W}_i} \right)^{1/\sum_{i=1}^{N} W_i}$$

where:

- CA_i is a constant summand, which is used to mitigate the impact of zero values of S_i;
- W_i is weight of a component S_i ($i = 1 \ldots \mathrm{N}$);
- N is a number of components, $N = 48$.

Values of these parameters are presented in **Table 1**.

Each individual component S_i is calculated with a common parametric formula:

$$S_i = \frac{100}{1 + \mathrm{Sq} \times \left(\frac{\max(x - \mathrm{IV}; 0)}{\mathrm{IP} - \mathrm{IV}} \right)^{\mathrm{CS}}}$$

This formula provides a (1-sigma)-like function with a customizable inflection point, tilt, and region of maximal score values. The function takes values from 0 to 100. Here:

- x is a variable to be scored (see **Table 1**).
- IV is an "ideal value." All cases with $x \leq$ IV would produce the maximum score (100). For example, S_{DP}, the component responsible for T-N expression level fold change (see **Table 1**) would be equal to 100 for any $\log_2\mathrm{FC_P}$ between -0.05 and $+0.05$ since IV = 0.05.
- IP is an "inflection point." In this point, there is the maximum decrease rate of S_i. When x is equal to IP and Sq = 1 (Sq takes these value for most S_i), the scoring component $S_i = 50$. Ideally, IP value should reflect the marginally acceptable value of x. For example, the relative standard deviation of gene expression (RelSD) values from 0 to 0.25 are appropriate, but RelSD = 0.4 … 0.5 is almost unacceptable. For the corresponding component (S_{EStD}), we chose IV = 0.1 and IP = 0.3 (see **Table 1**).
- CS is a "curve slope." The greater CS value, the stronger S_i decrease rate. Higher CS values should be assigned to more important scoring components.
- Sq is a "Squeeze," this is an auxiliary parameter. For most scoring components, it is equal to 1.

All scoring components S_i and parameters (IV, IP, CS, Sq) are presented in **Table 1**. The derived scoring functions are shown in **Figure 1**.

Two components, S_{DP} and S_{DL}, are responsible for T-N expression level difference. This is the major factor of RG suitability. S_{DP} is calculated for pooled, and S_{DL}–for paired samples. Hence, we applied the strongest scoring parameters (IV = 0.05, IP = 0.25, CS = 2.5) and assigned high weight (W = 4) for these two components. S_{DP} (or S_{DL}) would be equal to 50 if the absolute value of average $\log_2\mathrm{FC_P}$ (or $\log_2\mathrm{FC_L}$) is equal to IP = 0.25, i.e., fold change between tumor and normal is about 20%. We chose IV = 0.05–0.1 for all the components that are responsible for expression level (S_{DP}, S_{DL}, S_{DoO}, S_{DoU}, S_{DLc}, S_{EStD}, S_{EoH}, S_{EoL}). This means that 5–10% mRNA level changes are ignored.

S_{DP} and S_{DL} are calculated using the trimmed means of either CPM (pooled sample) or $\log_2\mathrm{FC_L}$ (paired samples). Only values from 10 to 90th percentiles are included. To take into account T-N expression outliers, we added two other scorings, S_{DoO} and S_{DoU}, that are responsible for the upper and lower deciles of $\log_2\mathrm{FC_L}$. For these components, we assigned increased IP value (IP = 0.7) since it is expected that $\mathrm{Abs}[\mathrm{Average}(\log_2\mathrm{FC_L})_{90-100}]$ calculated for 90–100th percentiles of $\log_2\mathrm{FC_L}$ will be much greater than such value calculated for 10–90th percentiles.

S_{EStD}, S_{EoH}, S_{EoL} are responsible for evaluating expression stability within pools of normal and tumor samples. S_{EStD} scores trimmed standard deviation of CPM values (10–90th percentiles), and S_{EoH} (or S_{EoL}) is responsible for outliers with high (or low) mRNA level (in terms of CPM). Additionally, we included scoring for average expression level (S_{EA}) and set high weight ($W = 6$) for this component in order to completely exclude genes with low mRNA level from the analysis.

Finally, we added scorings for correlations between gene expression and six clinical and pathological characteristics: pathologic TNM classification (separately for T, N, and M indexes), pathologic stage, follow-up person neoplasm cancer status and follow-up treatment success status. S_{Cr} is the component responsible for Spearman's correlation coefficient, and S_{Cp}–for correlation p-value. IV values were chosen in such a way that cases with $p > 0.25$ and $|r_s| < 0.1$ have score equals to 100. In total, each of these two components is taken 18 times: 6 clinical characteristics are analyzed for associations with CPM in tumor samples, CPM in normal samples and T-N expression fold change (paired samples). Hence, we assigned low weights—$W =$ 0.2 and 0.3 for S_{Cr} and S_{Cp}, respectively.

Besides stable and high enough expression level, an appropriate RG should also demonstrate a low mutation rate, single transcript isoform, and absence of pseudogenes in order to avoid problems with PCR priming and ensure the

TABLE 1 | Components of the scoring function.

Component	Factor	Variable (x = …)*	IV	IP	CS	Sq	CA	W	Number of times applied
EXPRESSION SCORING									
S_{DP}	T-N expression level difference (pooled samples)	Abs $(\log_2 FC_P)_{10-90}$	0.05	0.25	2.5	1	0	4	1 (all samples)
S_{DL}	T-N expression level difference (paired samples)	Abs (Average$(\log_2 FC_L)_{10-90}$)							1 (paired samples)
S_{DoO}	T-N expression level difference: outliers, overexpression	Abs (Average$(\log_2 FC_L)_{90-100}$)	0.1	0.7	2.5	1	10	1	1 (paired samples)
S_{DoU}	T-N expression level difference: outliers, underexpression	Abs (Average$(\log_2 FC_L)_{0-10}$)							1 (paired samples)
S_{DLc}	Cumulative T-N expression difference among paired samples	Average (Abs$(\log_2 FC_L)_{10-90}$)	0.1	0.5	2.5	1	5	2	1 (paired samples)
S_{EStD}	Expression level stability: standard deviation	StDev $(CPM)_{10-90}$/Average $(CPM)_{10-90}$	0.1	0.3	2	1	5	1.5	2 (all samples: normal and tumor)
S_{EoH}	Expression level stability: outliers (high expression)	$\log_2$ (Average$(CPM)_{90-100}$/Average $(CPM)_{10-90}$)	0.1	0.7	2.5	1	5	0.75	2 (all samples: normal and tumor)
S_{EoL}	Expression level stability: outliers (low expression)	$\log_2$ (Average$(CPM)_{10-90}$/Average $(CPM)_{0-10}$)							2 (all samples: normal and tumor)
S_{EA}	Average expression level	$1/\log_2 (CPM)_{10-90}$	0.07	0.15	3	1	0	6	1 (all tumor samples)
S_{Cp}	Correlations of expression with clinical parameters (*p*-values)	-$\log_2$ (*p*-value)	2	4	3	0.3	5	0.3	18 (3 × 6; 3: CPM_{10-90} all tumor samples, CPM_{10-90} all normal samples, $(\log_2 FC_L)_{10-90}$; 6: pathologic TNM classification, pathologic stage, follow-up—person neoplasm cancer status, follow-up—treatment success)
S_{Cr}	Correlations of expression with clinical parameters (r_s)	Abs (r_s)	0.1	0.25	2.5	0.3	5	0.2	18 (the same as above)
"ANTI-SCORINGS"									
S^{Mut}	Percentile of mutation rate		75	95	4	1			
$S^{Isoforms}$	Number of transcript isoforms		1	3	2	0.4			
$S^{Pseudogenes}$	Number of pseudogenes		0	2	2	0.4			

**Percentiles, which were taken into calculation, are indicated as a subscript.*

IV, ideal value; IP, inflection point; CS, curve slope; Sq, "squeeze"; CA, constant add; W, weight; Abs (…), absolute value; Average (…), mean value; CPM, counts per million, gene expression level; FC_P, ratio of the average CPM in a pool of tumor samples to the average CPM in a pool of normal samples; FC_L, ratio of CPM values between tumor and matched normal tissue (per each paired sample); StDev (…), standard deviation; r_s, Spearman's correlation coefficient.

rigorous evaluation of mRNA level. The mutation rate of a gene was assessed using TCGA data on somatic mutations. The number of transcript isoforms (per gene) was obtained from the Ensembl human genome annotation (hg38, release 88). The number of pseudogenes (per gene) was derived from psiCube (Sisu et al., 2014). Therefore, we extended the scoring system with three additional components, "anti-scorings" (**Table 1** and **Figure 1**). The resulting score S^{Final} is calculated as follows:

$$S^{Final} = S^{Exp} \cdot S^{Mut} \cdot S^{Isoforms} \cdot S^{Pseudogenes}$$

Next, we tried to find RGs that are stably expressed across multiple tissues and cancer types. For this purpose, we calculated the pan-cancer score as follows:

$$S^{Final}_{Pan-cancer} = S^{Exp\&Mut}_{Pan-cancer} \cdot S^{Isoforms} \cdot S^{Pseudogenes}$$

where:

$$S^{Exp\&Mut}_{Pan-cancer} = \left(\frac{\sum_{j=1}^{M} \left(S_j^{Exp} \cdot S_j^{Mut} + CA \right)^k}{M} \right)^{1/k}$$

where $M = 12$ (a number of cancer types analyzed); $k = -0.4$ (negative k value implies that the pan-cancer score is a harmonic mean of individual scores); $CA = 12$ (a constant add).

Finally, we assessed the involvement of a gene in cancer-related processes on the basis of Gene Ontology (GO; The Gene Ontology, 2017) data and mentions in the articles indexed by PubMed (titles and abstracts).

A RG should not be involved in cellular processes that are frequently altered in cancer. A penalty system based on GO data was developed. We evaluated the involvement of

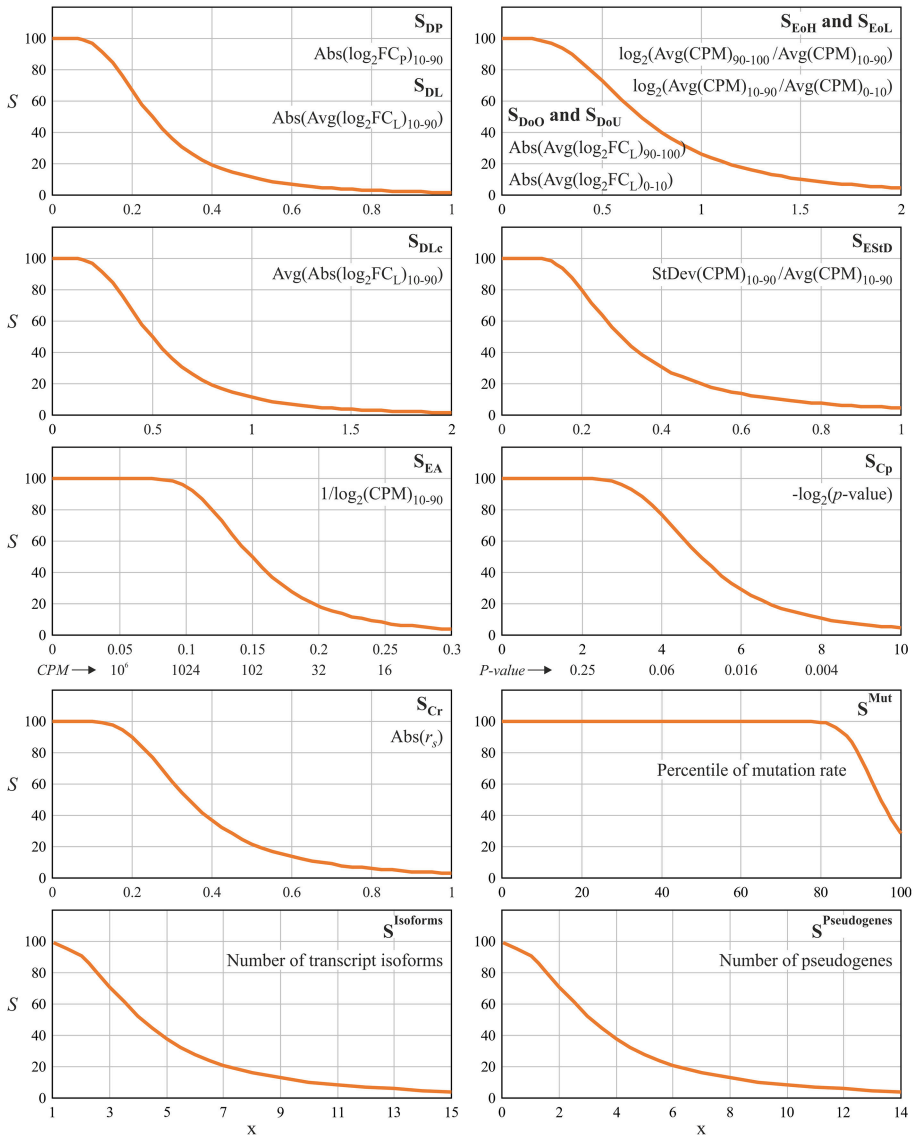

FIGURE 1 | Scoring functions used for evaluation of gene suitability for qPCR data normalization. Percentiles, which were taken into calculation, are indicated as a subscript. Abs(…), absolute value; Avg(…), mean value; CPM, counts per million, gene expression level; FC_P, ratio of the average CPM in a pool of tumor samples to the average CPM in a pool of normal samples; FC_L, ratio of CPM values between tumor and matched normal tissue (per each paired sample); StDev(…), standard deviation; r_S, Spearman's correlation coefficient.

a gene in 6 cancer-associated biological processes: cell cycle, differentiation, stress response, immune response, angiogenesis, adhesion, and cell communication. The relation of a gene to each of these processes was followed by the assignment of penalty points (from 2 to 5). Finally, these points were summed up. According to this system, a gene is penalized (1) with 5 points if its GO annotation contains at least one keyword related to cell cycle process: *cell cycle, cell division, cell growth, cell proliferation, apoptosis, apoptotic process, cell death, MAPK cascade, tumor, oncogenic, apoptotic*; (2) with 4 points if GO annotation contains a keyword related to cell differentiation: *cell differentiation, epithelial to mesenchymal transition, mesenchymal to epithelial transition, stem cell, fetal, embryonic, embryonal, embryo, gastrulation, tissue development, cellular developmental process, organ development*; (3) with 3 points for stress response related processes: *response to stress, DNA damage, DNA repair*; (4) with 2 points for inflammation and immune response: *inflammation, inflammatory, immune response, T cell activation, macrophage activation, antigen*; (5) with 2 points for angiogenesis: *angiogenesis*; (6) with 2 points for intercellular interactions: *cell communication, cell-cell signaling, cell adhesion, cell motility, cell migration*. Thus, a gene may have a maximum of $5 + 4 + 3 + 2 + 2 + 2 = 18$ penalty points.

The more accurately the gene is annotated, the more likely it is to find one of the keywords in its annotation. Therefore, GO

penalty should be normalized taking into account the number of assigned GO terms for the gene. On the other hand, the better the gene is annotated, the more extensively it is studied, and such genes represent more attractive candidates. In order to keep a balance between these two factors, we introduced normalization coefficient evaluated as the total number of GO terms (assigned for the gene) to the power of 0.3. If a gene lacked sufficient GO annotation (<3 GO terms), we assigned it 10 penalty points.

The number of PubMed-indexed articles with the mention of a gene name or its aliases was evaluated to assesses how well a gene is studied. Next, within this pool of gene-related publications, the number of cancer-related articles was also evaluated. One of the following words should be present in an article title to be treated as cancer-related: *cancer, tumor, *carcinoma, sarcoma, glioma, glioblastoma*, and other keywords.

The described components (GO and Pubmed) were not included in the main scoring and were only used for manual exclusion of cancer-associated genes. Besides, functional annotations from RefSeqGene (https://www.ncbi.nlm.nih.gov/refseq/rsg/) were added to each gene.

When revealing optimal RG pairs for each of examined cancer types, we paid special attention to the co-expression of RG candidates to avoid genes with a pronounced correlation between their mRNA levels. To implement the scoring system, we modified our previously developed CrossHub tool (the updated version can be downloaded at https://sourceforge.net/projects/crosshub/).

RESULTS

We performed the analysis of 12 cancer types from the TCGA project that have RNA-Seq data for representative sample sets: 285-1095 tumor and 19-113 matched normal tissues. These are: breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), kidney renal cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), prostate adenocarcinoma (PRAD), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), and bladder urothelial carcinoma (BLCA). For the remaining TCGA cancer types, RNA-Seq data were available only for a few normal tissue samples, and this makes it impossible to use such datasets for the discovery of reliable RGs.

First, we assessed the expression stability of a set of 32 frequently used RGs in 12 selected cancer types: *ACTB*, *ALAS1*, *B2M*, *CDKN1A*, *G6PD*, *GAPDH*, *GUSB*, *HBB*, *HMBS*, *HPRT1*, *HSP90AB1*, *IPO8*, *LDHA*, *NONO*, *PGK1*, *POP4*, *PPIA*, *PPIH*, *PSMC4*, *PUM1*, *RPL13A*, *RPL30*, *RPLP0*, *RPS17*, *RPS18*, *SDHA*, *TBP*, *TFRC*, *UBC*, *YWHAZ*, *TUBB*, *RPN1*. This set of 32 RGs was composed of commercially available RG panels: Roche "Human Reference Gene Panel, 384" (Switzerland), TATAA "Reference Gene Panel Human" (Sweden), and Bio-Rad "Reference Genes H384" (USA). In total, 31 unique genes are included in the panels, plus we added the *RPN1* gene, which was identified by us earlier as a reliable RG for lung, kidney, and colorectal cancers (Krasnov et al., 2011; Fedorova et al., 2015). Expression stability scores were calculated for each gene in each examined cancer type. The results for the top 5 genes are presented in **Table 2** and full data—in **Supplementary Table 1**. In almost each cancer type, there were 1–10 genes with expression score about 70 or more (with a theoretical maximum of 100), which can be considered as moderately high score value. PRAD and THCA demonstrated the highest number of genes with stable mRNA level-10 and 7, respectively. Only in BCLA, all the genes had scores below 70, possibly because of potential bias due to a small number of matched normal tissues (19—the smallest number among the cancer types examined). The cross-tissue analysis of 12 cancer types revealed that the most stably expressed genes were: *PUM1* ($S^{Exp} = 70$), *IPO8* ($S^{Exp} = 61$), *UBC* ($S^{Exp} = 60$), *ACTB* ($S^{Exp} = 55$), and *RPN1* ($S^{Exp} = 54$). *GAPDH*, one of the most frequently used RGs, showed one of the least stability of mRNA level—position 25 out of 32 ($S^{Exp} = 32$). According to the obtained results, *GAPDH* can be reasonably applied as a RG only in prostate and stomach adenocarcinomas. *RPN1* gene suggested by us demonstrated high expression stability score in lung, renal, colon, liver, thyroid, and prostate cancers.

Next, for each of 12 cancer types, we searched for a pair of the most suitable RGs focusing on S^{Exp} values and correlation between mRNA levels of genes in a pair. As a result, we revealed 12 optimal pairs of RGs with S^{Exp} above 65 for each gene and absence of co-expression (**Table 2** and **Supplementary Table 1**). *PUM1* came into the pair of RGs for 9 out of 12 cancer types.

It should be noted that genes with high S^{Exp} values may be inconvenient in practice because of the presence of numerous pseudogenes, alternatively spliced transcripts or a high mutation rate. Among the traditionally used RGs with high expression scores, only 3 genes met the requirements—*PUM1*, *IPO8*, and *RPN1*. These genes have no pseudogenes, one (*RPN1*), or two (*PUM1* and *IPO8*) transcript isoforms, and relatively low mutation rate in examined cancer types.

Using the expanded scoring system (**Figure 2**), in which 3 "anti-scorings" counting mutation rate, number of transcript isoforms and pseudogenes were included, we analyzed a complete list of human genes in order to reveal the most prominent pan-cancer RG candidates (**Supplementary Table 2**). Top 10 pan-cancer RG candidates included *MBTPS1*, *HNRNPA0*, *SF3A1*, *SF3B2*, *GGNBP2*, *HNRNPUL2*, *SFRS3*, *RTF1*, *CIAO1*, *TM9SF3*. All these genes had stable and high enough mRNA level and low mutation rate in most of 12 cancer types, only one annotated transcript isoform and no pseudogenes. Taking into account PubMed article search, GO annotations, and RefSeqGene information, we selected three most promising RG candidates—*SF3A1*, *CIAO1*, and *SFRS4*.

DISCUSSION

The use of inappropriate RGs leads to unreliable data and nullifies potentially high accuracy of a qPCR technique in the evaluation of differential gene expression. The search for a RG with a stable mRNA level under experimental conditions represents a separate object of research and is rarely performed during the original

TABLE 2 | Top 5 traditionally used reference genes with the highest expression scores in 12 cancer types.

Cancer type	1		2		3		4		5	
	Gene	S^{Exp}	Gene	S^{Exp}	Gene	S^{Exp}	Gene	S^{Exp}	Gene	S^{Exp}
BRCA	***UBC***	82.1	***PUM1***	75.7	*IPO8*	71.8	*RPLP0*	69.8	*RPS18*	66.2
LUAD	***UBC***	79.8	*ACTB*	76.4	***PUM1***	69.6	*RPN1*	67.9	*RPL13A*	65.5
LUSC	***UBC***	81.4	*IPO8*	72.9	*ACTB*	71.4	***PUM1***	70.7	*RPL13A*	66.3
KIRC	***NONO***	82.6	*HSP90AB1*	73.2	***RPN1***	69.7	*YWHAZ*	68.7	*PSMC4*	64.7
KIRP	***PUM1***	70.3	***PSMC4***	66.0	*PGK1*	63.2	*ALAS1*	61.7	*IPO8*	61.1
PRAD	***SDHA***	80.8	*YWHAZ*	78.4	*PSMC4*	76.2	***PUM1***	76.1	*UBC*	75.8
COAD	***PUM1***	76.9	*GUSB*	73.4	***UBC***	72.8	*ACTB*	72.0	*IPO8*	71.6
HNSC	***RPL30***	73.4	***PUM1***	72.7	*IPO8*	68.1	*ACTB*	64.2	*PSMC4*	63.1
LIHC	***RPN1***	82.3	***ACTB***	80.9	*UBC*	78.4	*PUM1*	65.7	*RPS17*	56.4
STAD	***IPO8***	71.7	***RPL30***	71.0	*GAPDH*	69.7	*RPLP0*	68.7	*PUM1*	68.1
THCA	***RPN1***	84.4	*HSP90AB1*	84.3	***PUM1***	80.0	*TUBB*	79.2	*YWHAZ*	76.0
BLCA	***SDHA***	66.3	***PUM1***	65.9	*HSP90AB1*	63.3	*RPL30*	62.2	*RPS17*	61.2
Cross-tissue	*PUM1*	70.1	*IPO8*	60.8	*UBC*	59.8	*ACTB*	54.7	*RPN1*	54.3

Optimal pairs of reference genes for each cancer type are shown in bold.

studies. RNA-Seq data of TCGA project offer a great opportunity for evaluating gene expression stability. Using our CrossHub tool, we developed a complex scoring system that allowed us to assess the suitability of 32 traditionally used RGs for qPCR data normalization in 12 cancer types characterized by high morbidity and mortality rates. The alterations of mRNA level were shown for a number of these genes, including the most frequently used *GAPDH*, in examined cancer types. The analysis across 12 cancer types revealed that *PUM1* and *IPO8* genes demonstrate the most stable expression among the 32 genes.

PUM1 (Pumilio RNA Binding Family Member 1) serves as a translational regulator of specific mRNAs by binding to their 3'-UTRs. It may be involved in translational regulation of embryogenesis, cell development, and differentiation. There are several functions that call into question its applicability as a RG. After growth factor stimulation, PUM1 binds to 3'-UTR of *CDKN1B/p27* tumor suppressor, inhibits its expression and promotes a rapid entry to the cell cycle (Kedde et al., 2010). PUM1 is capable of repressing many mitotic, DNA repair, and DNA replication factors (Lee et al., 2016). Moreover, some authors reported that *PUM1* promotes ovarian cancer proliferation, migration, and invasion (Guan et al., 2018). However, *PUM1* is identified as one of the most stably expressed genes in uterine cervical cancer (Tan et al., 2017), endometrial carcinoma (Ayakannu et al., 2015), gallbladder (Yu et al., 2015), leiomyoma (Almeida et al., 2014), breast (Ibusuki et al., 2013; Kilic et al., 2014), and non-small cell lung (Soes et al., 2013) cancers. This gene has only 2 transcript isoforms and no pseudogenes that makes it even more attractive for use as a reference one.

Recently, Tilli et al. performed a screening of breast cancer RNA-Seq datasets from the International Cancer Genome Consortium (ICGC), GEO, and TCGA repositories. Authors found that *PUM1*, along with "novel" RGs - *CCSER2*, *SYMPK*, and *ANKRD17*, had the most stable mRNA level (Tilli et al., 2016). This agrees with previous qPCR analyses of RG expression stability in breast carcinomas (Ibusuki et al., 2013; Kilic et al., 2014).

IPO8 (*importin 8*), which has 2 transcript isoforms and no pseudogenes, is the second in the cross-tissue stability list, but its mRNA level is much less stable than that of *PUM1* according to TCGA data. IPO8 mediates nuclear import of proteins with a classical nuclear localization signal. Previously, *IPO8* was found to be suitable for data normalization in endometrial (Ayakannu et al., 2015) and ovarian carcinomas (Kolkova et al., 2013), colon adenocarcinoma cell lines (Krzystek-Korpacka et al., 2016), non-small cell lung cancer (Soes et al., 2013), and other tissues and diseases: brain edema (Du et al., 2017), heart cavities (Molina et al., 2018), T cells, and neutrophils (Ledderose et al., 2011).

The *RPN1* gene (0 pseudogenes, 1 transcript isoform), which was previously suggested by us for normalization of qPCR data in LUAD, LUSC, KIRC, KIRP, and COAD (Krasnov et al., 2011; Fedorova et al., 2015), demonstrate stable expression in these cancer types as well as in PRAD, LIHC, and THCA.

The majority of the remaining genes from the set of 32 genes, even if they demonstrate stable mRNA level in certain cancer types, have many pseudogenes or high mutation rate (for example, *UBC* is above the 99th percentile in BRCA). The presence of pseudogenes is a weakness of such widely used RGs as *GAPDH* and *ACTB* (67 and 64, respectively) (Sun et al., 2012), or genes encoding ribosomal proteins, including *RPL13A* and *RPS17* (Tonner et al., 2012).

Next, we tried to find out novel reliable and convenient RGs suitable for most cancer types. As it was described above, for this purpose, we evaluated expression and mutation scorings for each examined cancer type, calculated pan-cancer scoring values given the "anti-scorings" for the number of transcript isoforms and pseudogenes, and selected the promising candidates taking into account information on functions of the genes and their involvement in carcinogenesis.

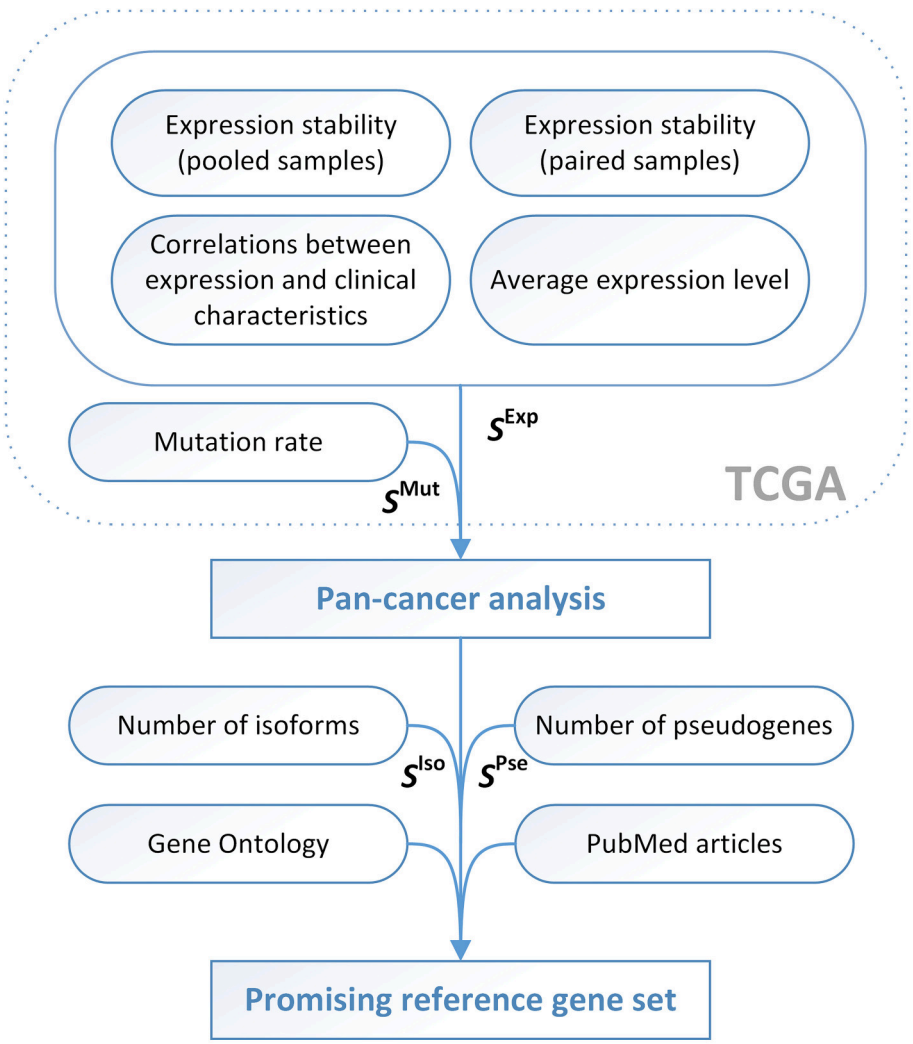

FIGURE 2 | The pipeline for identification of promising reference genes for qPCR data normalization in cancer studies. $S^{Iso} - S^{Isoforms}$, $S^{Pse} - S^{Pseudogenes}$.

Along with *SFRS4* (number 13 in the top list of "universal" reference genes), three genes that participate in pre-mRNA splicing and processing pathways (*SF3A1, SF3B2*, and *SFRS3*) are present in the top 10 of promising pan-cancer RGs. The splicing machinery (namely spliceosome) is the largest molecular machine so far described. It is composed of five small nuclear ribonucleoproteins (snRNPs U1, U2, U4, U5, and U6) and more than 100 different polypeptides (Ghigna et al., 2008). Aberrant splicing in cancer provides a way to generate alternatively spliced transcripts encoding proteins with distinct functions (Ghigna et al., 2008). There are at least two ways resulting in splicing aberrations in cancer: mutations in the affected genes, e.g., in their splice sites (*cis*-effect), and altered expression and/or activity of the elements of splicing machinery (*trans*-effect). Some of the splicing factors are known to be deregulated in cancer, by means of mRNA level alterations, mutations or posttranslational modifications (Stickeler et al., 1999; Blaustein et al., 2005; Ghigna et al., 2008). On the other hand, some of the splicing factors are considered as potential RGs. This may be explained by the complexity of the splicing machinery and various roles of its elements (David and Manley, 2010).

SF3A1 and *SF3B2* encode the subunits of splicing factors 3a and 3b. These two splicing factors together with 12S RNA unit form the U2 small nuclear ribonucleoproteins complex, which binds pre-mRNA upstream of the intron's branch site and may anchor the U2 snRNP to the pre-mRNA (Will et al., 2002). *SF3A1* is considered as a RG in sarcoma (Aggerholm-Pedersen et al., 2014), its expression was found to be stable in breast cancer (Maltseva et al., 2013), colorectal adenocarcinoma Caco-2 cells under exposure to food products (Vreeburg et al., 2011), white blood cells under treatment with growth hormone (Castigliego et al., 2010), bovine blastocysts produced by different methods (Luchsinger et al., 2014), bovine granulosa cells of dominant follicles during follicular growth and aging (Khan et al., 2016).

Considering the other splicing machinery gene, *SFRS4* (*serine and arginine rich splicing factor 4*), some authors earlier demonstrated that its mRNA level is stable in hepatocellular carcinoma (HCC) cell lines (Liu et al., 2017) and patients with alcoholic liver disease (Boujedidi et al., 2012). *SFRS4* remains stably expressed in hepatitis C virus-induced HCC, whereas *ACTB* and *GAPDH* are significantly deregulated (Waxman and Wurmbach, 2007).

CIAO1 (number 9 in the top list) is a key component of the cytosolic iron-sulfur protein assembly (CIA) complex. This is a multiprotein complex that mediates the incorporation of iron-sulfur cluster into extramitochondrial Fe/S proteins (provided by GeneCards; Stelzer et al., 2016). *CIAO1* was not previously described as a RG. Till now, there is only one article describing the possible role of the encoded protein in cancer development, namely interacting with the tumor suppressor protein WD40 (Johnstone et al., 1998). Besides this, there is almost no data on the association of this gene with cancer.

CONCLUSIONS

To reveal reliable RGs for qPCR data normalization, a comprehensive analysis of TCGA data was performed. We took into account expression stability, average mRNA level, expression correlation with clinical and pathological characteristics, number of pseudogenes and transcript isoforms, mutation rate, GO terms, and mentions of a gene in titles/abstracts of articles from PubMed. The most reliable pairs of traditionally used RGs were suggested for each of 12 examined cancer types, as well as unsuitability of some frequently used RGs was shown. Pan-cancer analysis revealed promising RG candidates with stable and sufficiently high expression level and low mutation rate across 12 cancer types. Besides, these genes have only one known transcript isoform and no pseudogenes.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or the supplementary files.

AUTHOR CONTRIBUTIONS

GK, AK, NM, and AD conceived and designed the work. GK, AK, AS, VL, AB, NM, and AD performed data analysis. GK and AD wrote the manuscript. All authors agreed with the final version of the manuscript and all aspects of the work.

FUNDING

This work was financially supported by the Russian Science Foundation, grant 17-74-20064.

ACKNOWLEDGMENTS

This work was performed using the equipment of Genome center of Engelhardt Institute of Molecular Biology (http://www.eimb.ru/rus/ckp/ccu_genome_c.php).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00097/full#supplementary-material

REFERENCES

Aggerholm-Pedersen, N., Safwat, A., Baerentzen, S., Nordsmark, M., Nielsen, O. S., Alsner, J., et al. (2014). The importance of reference gene analysis of formalin-fixed, paraffin-embedded samples from sarcoma patients-an often underestimated problem. *Transl. Oncol.* 7, 687–693. doi: 10.1016/j.tranon.2014.09.012

Alexander, H., Jenkins, B. D., Rynearson, T. A., Saito, M. A., Mercier, M. L., and Dyhrman, S. T. (2012). Identifying reference genes with stable expression from high throughput sequence data. *Front. Microbiol.* 3:385. doi: 10.3389/fmicb.2012.00385

Almeida, T. A., Quispe-Ricalde, A., Montes de Oca, F., Foronda, P., and Hernandez, M. M. (2014). A high-throughput open-array qPCR gene panel to identify housekeeping genes suitable for myometrium and leiomyoma expression analysis. *Gynecol. Oncol.* 134, 138–143. doi: 10.1016/j.ygyno.2014.04.012

Andersen, C. L., Jensen, J. L., and Orntoft, T. F. (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. doi: 10.1158/0008-5472.CAN-04-0496

Ayakannu, T., Taylor, A. H., Willets, J. M., Brown, L., Lambert, D. G., McDonald, J., et al. (2015). Validation of endogenous control reference genes for normalizing gene expression studies in endometrial carcinoma. *Mol. Hum. Reprod.* 21, 723–735. doi: 10.1093/molehr/gav033

Barber, R. D., Harmer, D. W., Coleman, R. A., and Clark, B. J. (2005). GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics* 21, 389–395. doi: 10.1152/physiolgenomics.00025.2005

Blaustein, M., Pelisch, F., Tanos, T., Munoz, M. J., Wengier, D., Quadrana, L., et al. (2005). Concerted regulation of nuclear and cytoplasmic activities of SR proteins by AKT. *Nat. Struct. Mol. Biol.* 12, 1037–1044. doi: 10.1038/nsmb1020

Boujedidi, H., Bouchet-Delbos, L., Cassard-Doulcier, A. M., Njike-Nakseu, M., Maitre, S., Prevot, S., et al. (2012). Housekeeping gene variability in the liver of alcoholic patients. *Alcohol. Clin. Exp. Res.* 36, 258–266. doi: 10.1111/j.1530-0277.2011.01627.x

Carmona, R., Arroyo, M., Jimenez-Quesada, M. J., Seoane, P., Zafra, A., Larrosa, R., et al. (2017). Automated identification of reference genes based on RNA-seq data. *Biomed. Eng. Online* 16(Suppl. 1): 65. doi: 10.1186/s12938-017-0356-5

Castigliego, L., Armani, A., Li, X., Grifoni, G., Gianfaldoni, D., and Guidi, A. (2010). Selecting reference genes in the white blood cells of buffalos treated with recombinant growth hormone. *Anal. Biochem.* 403, 120–122. doi: 10.1016/j.ab.2010.04.001

Chapman, J. R., and Waldenstrom, J. (2015). With reference to reference genes: a systematic review of endogenous controls in gene expression studies. *PLoS ONE* 10:e0141853. doi: 10.1371/journal.pone.0141853

Chen, L., Jin, Y., Wang, L., Sun, F., Yang, X., Shi, M., et al. (2017). Identification of reference genes and miRNAs for qRT-PCR in human esophageal squamous cell carcinoma. *Med. Oncol.* 34:2. doi: 10.1007/s12032-016-0860-7

Chim, S. S. C., Wong, K. K. W., Chung, C. Y. L., Lam, S. K. W., Kwok, J. S. L., Lai, C. Y., et al. (2017). Systematic selection of reference genes for the normalization of circulating RNA transcripts in pregnant women based on RNA-Seq data. *Int. J. Mol. Sci.* 18:E1709. doi: 10.3390/ijms18081709

David, C. J., and Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.* 24, 2343–2364. doi: 10.1101/gad.1973010

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics* 14, 671–683. doi: 10.1093/bib/bbs046

Du, Y., Xu, J. T., Jin, H. N., Zhao, R., Zhao, D., Du, S. H., et al. (2017). Increased cerebral expressions of MMPs, CLDN5, OCLN, ZO1 and AQPs are associated with brain edema following fatal heat stroke. *Sci. Rep.* 7:1691. doi: 10.1038/s41598-017-01923-w

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinformatics* 19, 776–792. doi: 10.1093/bib/bbx008

Fedorova, M. S., Kudryavtseva, A. V., Lakunina, V. A., Snezhkina, A. V., Volchenko, N. N., Slavnova, E. N., et al. (2015). Downregulation of OGDHL expression is associated with promoter hypermethylation in colorectal cancer. *Mol. Biol.* 49, 608–617. doi: 10.1134/S0026893315040044

Ghigna, C., Valacca, C., and Biamonti, G. (2008). Alternative splicing and tumor progression. *Curr. Genomics* 9, 556–570. doi: 10.2174/138920208786847971

Guan, X., Chen, S., Liu, Y., Wang, L. L., Zhao, Y., and Zong, Z. H. (2018). PUM1 promotes ovarian cancer proliferation, migration and invasion. *Biochem. Biophys. Res. Commun.* 497, 313–318. doi: 10.1016/j.bbrc.2018.02.078

Gur-Dedeoglu, B., Konu, O., Bozkurt, B., Ergul, G., Seckin, S., and Yulug, I. G. (2009). Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. *Oncol. Res.* 17, 353–365. doi: 10.3727/096504009788428460

Hoang, V. L. T., Tom, L. N., Quek, X. C., Tan, J. M., Payne, E. J., Lin, L. L., et al. (2017). RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. *PeerJ* 5:e3631. doi: 10.7717/peerj.3631

Huggett, J., Dheda, K., Bustin, S., and Zumla, A. (2005). Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun.* 6, 279–284. doi: 10.1038/sj.gene.6364190

Ibusuki, M., Fu, P., Yamamoto, S., Fujiwara, S., Yamamoto, Y., Honda, Y., et al. (2013). Establishment of a standardized gene-expression analysis system using formalin-fixed, paraffin-embedded, breast cancer specimens. *Breast Cancer* 20, 159–166. doi: 10.1007/s12282-011-0318-x

Janssens, N., Janicot, M., Perera, T., and Bakker, A. (2004). Housekeeping genes as internal standards in cancer research. *Mol. Diagn.* 8, 107–113. doi: 10.1007/BF03260053

Johnstone, R. W., Wang, J., Tommerup, N., Vissing, H., Roberts, T., and Shi, Y. (1998). Ciao 1 is a novel WD40 protein that interacts with the tumor suppressor protein WT1. *J. Biol. Chem.* 273, 10880–10887. doi: 10.1074/jbc.273.18.10880

Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J. A., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat. Cell Biol.* 12, 1014–1020. doi: 10.1038/ncb2105

Khan, M. I., Dias, F. C., Dufort, I., Misra, V., Sirard, M. A., and Singh, J. (2016). Stable reference genes in granulosa cells of bovine dominant follicles during follicular growth, FSH stimulation and maternal aging. *Reprod. Fertil. Dev.* 28, 795–805. doi: 10.1071/RD14089

Kilic, Y., Celebiler, A. C., and Sakizli, M. (2014). Selecting housekeeping genes as references for the normalization of quantitative PCR data in breast cancer. *Clin. Transl. Oncol.* 16, 184–190. doi: 10.1007/s12094-013-1058-5

Kolkova, Z., Arakelyan, A., Casslen, B., Hansson, S., and Kriegova, E. (2013). Normalizing to GADPH jeopardises correct quantification of gene expression in ovarian tumours - IPO8 and RPL4 are reliable reference genes. *J. Ovarian Res.* 6:60. doi: 10.1186/1757-2215-6-60

Kozera, B., and Rapacz, M. (2013). Reference genes in real-time PCR. *J. Appl. Genet.* 54, 391–406. doi: 10.1007/s13353-013-0173-x

Krasnov, G. S., Dmitriev, A. A., Melnikova, N. V., Zaretsky, A. R., Nasedkina, T. V., Zasedatelev, A. S., et al. (2016). CrossHub: a tool for multi-way analysis of The Cancer Genome Atlas (TCGA) in the context of gene expression regulation mechanisms. *Nucleic Acids Res.* 44:e62. doi: 10.1093/nar/gkv1478

Krasnov, G. S., Oparina, N. Y., Dmitriev, A. A., Kudryavtseva, A. V., Anedchenko, E. A., Kondrat'eva, T. T., et al. (2011). RPN1, a new reference gene for quantitative data normalization in lung and kidney cancer. *Mol. Biol.* 45, 211–220. doi: 10.1134/S0026893311020129

Krzystek-Korpacka, M., Hotowy, K., Czapinska, E., Podkowik, M., Bania, J., Gamian, A., et al. (2016). Serum availability affects expression of common house-keeping genes in colon adenocarcinoma cell lines: implications for quantitative real-time PCR studies. *Cytotechnology* 68, 2503–2517. doi: 10.1007/s10616-016-9971-4

Ledderose, C., Heyn, J., Limbeck, E., and Kreth, S. (2011). Selection of reliable reference genes for quantitative real-time PCR in human T cells and neutrophils. *BMC Res. Notes* 4:427. doi: 10.1186/1756-0500-4-427

Lee, S., Kopp, F., Chang, T. C., Sataluri, A., Chen, B., Sivakumar, S., et al. (2016). Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* 164, 69–80. doi: 10.1016/j.cell.2015.12.017

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* 16:347. doi: 10.1186/s12859-015-0778-7

Liu, Y., Qin, Z., Cai, L., Zou, L., Zhao, J., and Zhong, F. (2017). Selection of internal references for qRT-PCR assays of human hepatocellular carcinoma cell lines. *Biosci. Rep.* 37:BSR20171281. doi: 10.1042/BSR20171281

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Luchsinger, C., Arias, M. E., Vargas, T., Paredes, M., Sanchez, R., and Felmer, R. (2014). Stability of reference genes for normalization of reverse transcription quantitative real-time PCR (RT-qPCR) data in bovine blastocysts produced by IVF, ICSI and SCNT. *Zygote* 22, 505–512. doi: 10.1017/S0967199413000099

Maltseva, D. V., Khaustova, N. A., Fedotov, N. N., Matveeva, E. O., Lebedev, A. E., Shkurnikov, M. U., et al. (2013). High-throughput identification of reference genes for research and clinical RT-qPCR analysis of breast cancer samples. *J. Clin. Bioinforma.* 3:13. doi: 10.1186/2043-9113-3-13

Molina, C. E., Jacquet, E., Ponien, P., Munoz-Guijosa, C., Baczko, I., Maier, L. S., et al. (2018). Identification of optimal reference genes for transcriptomic analyses in normal and diseased human heart. *Cardiovasc. Res.* 114, 247–258. doi: 10.1093/cvr/cvx182

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32, 462–464. doi: 10.1038/nbt.2862

Pfaffl, M. W., Tichopad, A., Prgomet, C., and Neuvians, T. P. (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. doi: 10.1023/B:BILE.0000019559.84305.47

Popovici, V., Goldstein, D. R., Antonov, J., Jaggi, R., Delorenzi, M., and Wirapati, P. (2009). Selecting control genes for RT-QPCR using public microarray data. *BMC Bioinformatics* 10:42. doi: 10.1186/1471-2105-10-42

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rubie, C., Kempf, K., Hans, J., Su, T., Tilton, B., Georg, T., et al. (2005). Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Mol. Cell. Probes* 19, 101–109. doi: 10.1016/j.mcp.2004.10.001

Sisu, C., Pei, B., Leng, J., Frankish, A., Zhang, Y., Balasubramanian, S., et al. (2014). Comparative analysis of pseudogenes across three phyla. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13361–13366. doi: 10.1073/pnas.1407293111

Soes, S., Sorensen, B. S., Alsner, J., Overgaard, J., Hager, H., Hansen, L. L., et al. (2013). Identification of accurate reference genes for RT-qPCR analysis of formalin-fixed paraffin-embedded tissue from primary non-small cell lung cancers and brain and lymph node metastases. *Lung Cancer* 81, 180–186. doi: 10.1016/j.lungcan.2013.04.007

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics* 54:1 30 31–31 30 33. doi: 10.1002/cpbi.5

Stickeler, E., Kittrell, F., Medina, D., and Berget, S. M. (1999). Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. *Oncogene* 18, 3574–3582. doi: 10.1038/sj.onc.1202671

Sun, Y., Li, Y., Luo, D., and Liao, D. J. (2012). Pseudogenes as weaknesses of ACTB (Actb) and GAPDH (Gapdh) used as reference genes in reverse transcription and polymerase chain reactions. *PLoS ONE* 7:e41659. doi: 10.1371/journal.pone.0041659

Tan, S. C., Ismail, M. P., Duski, D. R., Othman, N. H., Bhavaraju, V. M., and Ankathil, R. (2017). Identification of optimal reference genes for normalization of RT-qPCR data in cancerous and non-cancerous tissues of human uterine cervix. *Cancer Invest.* 35, 163–173. doi: 10.1080/07357907.2017.1278767

The Gene Ontology, C. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkw1108

Tilli, T. M., Castro Cda, S., Tuszynski, J. A., and Carels, N. (2016). A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases. *BMC Genomics* 17:639. doi: 10.1186/s12864-016-2946-1

Tonner, P., Srinivasasainagendra, V., Zhang, S., and Zhi, D. (2012). Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics* 13:412. doi: 10.1186/1471-2164-13-412

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3:RESEARCH0034. doi: 10.1186/gb-2002-3-7-research0034

Vreeburg, R. A., Bastiaan-Net, S., and Mes, J. J. (2011). Normalization genes for quantitative RT-PCR in differentiated Caco-2 cells used for food exposure studies. *Food Funct.* 2, 124–129. doi: 10.1039/C0FO00068J

Waxman, S., and Wurmbach, E. (2007). De-regulation of common housekeeping genes in hepatocellular carcinoma. *BMC Genomics* 8:243. doi: 10.1186/1471-2164-8-243

Will, C. L., Urlaub, H., Achsel, T., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J.* 21, 4978–4988. doi: 10.1093/emboj/cdf480

Yu, S., Yang, Q., Yang, J. H., Du, Z., and Zhang, G. (2015). Identification of suitable reference genes for investigating gene expression in human gallbladder carcinoma using reverse transcription quantitative polymerase chain reaction. *Mol. Med. Rep.* 11, 2967–2974. doi: 10.3892/mmr.2014.3008

Zhao, H., Ma, T. F., Lin, J., Liu, L. L., Sun, W. J., Guo, L. X., et al. (2018). Identification of valid reference genes for mRNA and microRNA normalisation in prostate cancer cell lines. *Sci. Rep.* 8:1949. doi: 10.1038/s41598-018-19458-z

Zhou, Z., Cong, P., Tian, Y., and Zhu, Y. (2017). Using RNA-seq data to select reference genes for normalizing gene expression in apple roots. *PLoS ONE* 12:e0185288. doi: 10.1371/journal.pone.0185288

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Gorczak, K., Klamecka, K., Figlerowicz, M., et al. (2015). The impact of normalization methods on RNA-Seq data analysis. *Biomed Res. Int.* 2015:621690. doi: 10.1155/2015/621690

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.