



OPEN

The pseudogene problem and RT-qPCR data normalization; *SYMPK*: a suitable reference gene for papillary thyroid carcinoma

Seyed-Morteza Javadirad¹✉, Mohammad Mokhtari^{1,3}, Ghazal Esfandiarpour^{1,3} & Mohsen Kolahdouzan²

In RT-qPCR, accuracy requires multiple levels of standardization, but results could be obfuscated by human errors and technical limitations. Data normalization against suitable reference genes is critical, yet their observed expression can be confounded by pseudogenes. Eight reference genes were selected based on literature review and analysis of papillary thyroid carcinoma (PTC) microarray data. RNA extraction and cDNA synthesis were followed by RT-qPCR amplification in triplicate with exon-junction or intron-spanning primers. Several statistical analyses were applied using Microsoft Excel, NormFinder, and BestKeeper. In normal tissues, the least correlation of variation (CqCV%) and the lowest maximum fold change (MFC) were respectively recorded for *PYCR1* and *SYMPK*. In PTC tissues, *SYMPK* had the lowest CqCV% (5.16%) and MFC (1.17). According to NormFinder, the best reference combination was *SYMPK* and *ACTB* (stability value = 0.209). BestKeeper suggested *SYMPK* as the best reference in both normal ($r = 0.969$) and PTC tissues ($r = 0.958$). *SYMPK* is suggested as the best reference gene for overcoming the pseudogene problem in RT-qPCR data normalization, with a stability value of 0.319.

Differential gene expression analysis enables us to understand the patterns of complex gene networks and regulatory elements, thereby providing insight into biological processes and disease origins¹. Reverse transcription quantitative polymerase chain reaction (RT-qPCR) is the “gold standard” and the most sensitive and accurate method for comprehensively quantifying gene expression². However, owing to the context of the RT-qPCR technique and presumable human errors, several biases can obfuscate the results of these assays, leading to inaccurate data interpretations^{1,3}.

In order to obtain the highest possible RT-qPCR data accuracy, multiple levels of standardization must be carried out⁴. After routine standardization of samples and adjustment of RNA concentration, the heart of RT-qPCR fine-tuning is in the use of suitable reference genes. Numerous lines of experiments have shown the inappropriateness of commonly-used reference genes, expression of which could fluctuate in different experimental and pathological conditions^{5–11}. Additionally, heterogeneity of intratumor and interpatient expression profiles, which are common characteristics of papillary thyroid carcinoma (PTC), magnify the problem of expression fluctuation in reference genes^{12–16}.

Finally, pseudogenes, naturally existing sequences with resemblance to fully-functional genes, are another factor that is not considered with painstaking care in mRNA expression analysis. Although non-operational due to deleterious mutations or retro-transposition, these replicates of working genes must also be added to the experimental budget because they can confound RT-qPCR assay results¹⁷. As expression analysis of true-functional genes is a daily practice of molecular biologists, not addressing bias from pseudogenes in RT-qPCR quantifications can lead to great headaches¹⁸. A recent bioinformatics exploration of the standard calibration genes Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) and β -actin (*ACTB*) designated these two genes as unreliable because of contamination by their ample pseudogenes¹⁹. Consequently, pseudogenes must be considered more precisely in order to diminish the inaccuracy of RT-qPCR, which is otherwise a fundamentally accurate technique. The venerable technique of DNase treatment followed by enzyme deactivation is one solution,

¹Department of Cell and Molecular Biology and Microbiology, Faculty of Biological Science and Technology, University of Isfahan, <https://ui.ac.ir/>. ²Department of Surgery, School of Medicine, Isfahan University of Medical Sciences, <https://mui.ac.ir/>. ³These authors contributed equally: Mohammad Mokhtari and Ghazal Esfandiarpour. ✉email: javadirad@yahoo.com; sm.javadirad@sci.ui.ac.ir

Gene symbol (Genbank accession number)	The sequence of Forward and Reverse primers	Primers attachment information			Primer type	Product length (bp)	Annealing temperature (°C)
		Location of attachment (exon length, bp)	Exact nucleotide attachment on exon	Spanning Intron (length, bp)			
GAPDH (NM_001256799.3)	F:CCACTCCTCCAC CTTTGACG	Exon 7 (413)	(344–363)	Intron 7–8 (104)	Intron spanning	107	58
	R:CCACCACCC TGTTGCTGTAG	Exon 8 (271)	(19–38)				
TBP (NM_003194.5)	F:GGTTTGCTG CGGTAATCA	Exon 5 (92)	2–19	Intron 5–6 (2602)	Exon junction	100	61
	R:TGTTCTTCACTC TTGGCTCCTGT	Exon 5 (92)	Exon 5 (79–92)				
		Exon 6 (168)	Exon 6 (1–9)				
SYMPK (NM_004819.3)	F:ACGGTGCTG AGGGTCATTGA	Exon 18 (160)	135–154	Intron 18–19 (1295)	Intron spanning, Exon junction	146	60
	R:GAGGGTGGG ACTTTGTCTGTGA	Exon 19 (109)	Exon 19 (98–109)	Intron 19–20 (294)			
		Exon 20 (101)	Exon 20 (1–10)				
PYCRI (NM_006907)	F:CTTCACAGC AGCAGGCGTC	Exon 1 (553) and Exon 2 (71)	Exon 1 (540–553) Exon2 (1–5)	Intron 1–2 (554)	Intron spanning, Exon junction	122	61
	R:TCTCCTTGTGT GGGGTGTC	Exon 3 (180)	Exon 3 (18–37)	Intron 2–3 (606)			
ACTB (NM_001101.5)	F:CTGGAACGG TGAAGGTGACA	Exon 6 (744)	Exon 6 (291–310)	Not applicable	Not applicable	140	61
	R:AAGGGACTT CCTGTAACAACGC	Exon 6 (744)	Exon 6 (409–430)				
HPRT1 (NM_000194.3)	F:CGTCGTGATTAG TGATGATG	Exon 1 (174) and Exon 2 (107)	Exon 1 (168–174) Exon 2 (1–13)	Intron 1–2 (13,020)	Exon junction	123	60
	R:CGTTCAGTC CTGTCCATA	Exon 2 (107) and Exon 3 (184)	Exon 2 (98–107) Exon 3 (1–9)	Intron 2–3 (1715)			
B2M (NM_004048.3)	F:TGAGTATGCCTG CCGTGTGA	Exon 2 (279)	Exon 2 (221–240)	Intron 2–3 (627)	Intron spanning, Exon junction	97	61
	R:ATCTTCAAACCT CCATGATGCT	Exon 3 (28)	Exon 3 (17–28)	Intron 3–4 (1250)			
		Exon 4 (1271)	Exon 4 (1–10)				
GUSB (NM_000181.4)	F:CGCCCTGCCTAT CTGTATTC	Exon 5 (188)	Exon 5 (162–181)	Intron 5–6 (943)	Intron spanning	91	60
	R:TCCCCACAG GGAGTGTGTAG	Exon 6 (153)	Exon 6 (45–64)				

Table 1. The information of the primers was listed. A total of eight reference genes were selected and the detailed information of corresponding primers was listed. RefSeq accession numbers of the selected reference genes, forward (F) and reverse (R) primer sequences, exact primer attachment sites at their target genes, total length of the involved exons and spanned introns, PCR product length and PCR product T_m, are listed. Intron spanning primers: primer pairs attaching within two consequential exons. Exon junction primer: primer with spanning attachment to consequential exons. *GAPDH* glyceraldehyde-3-phosphate dehydrogenase, *TBP* TATA-box binding protein, *SYMPK* symplekin, *PYCRI* pyrroline-5-carboxylate reductase 1, *ACTB* actin beta, *HPRT1* hypoxanthine phosphoribosyltransferase 1, *B2M* beta-2-microglobulin, *GUSB* glucuronidase beta.

but unfortunately such handling would lead to RNA degradation^{20,21}. A second possible solution is the usage of exon-junction (intron-spanning) primers; these have bidirectional benefits, as they not only eliminate genomic DNA but also heterogeneous nuclear RNA²². Finally, finding suitable reference genes with the lowest variation alongside an absence of pseudogenes is necessary to ensure the most accurate results.

In this study, we try to bring about a substantial improvement in the field of RT-qPCR data normalization by introducing Symplekin (*SYMPK*) as a stably expressed reference gene without pseudogenes, suitable for normalization in both PTC tumors and adjacent normal tissues (Tables 1 and 2).

Results

RNA quality and quantity. The mean A260/A280 ratios obtained for PTC tissues and their adjacent normal tissues were 1.96 ± 0.11 and 1.97 ± 0.06 , respectively. The intensity of 28srRNA bands was approximately two-fold that of 18srRNA, which indicates integrity of the extracted mRNAs.

Expression analysis of candidate reference genes. Results from basic statistical analysis of the eight candidate reference genes in both normal and PTC tissues are shown in Table 3. In normal tissues, the least CqCV% values were recorded for *PYCRI* (4.86), *SYMPK* (6.64), and *TBP* (7.19), which genes also showed the lowest MFC values (1.16, 1.21, and 1.27, respectively). In PTC tissues, the least CqCV% values were obtained for *SYMPK* (5.16), *PYCRI* (6.19), and *TBP* (6.88), as were the lowest MFC values (1.17, 1.25, and 1.24, respectively). The representation of CqCV% was illustrated in Fig. 1.

Parameter	PTC	Adjacent normal tissue	Total
Tissue number	10	10	20
Histopathology			
Classical PTC	6		
Follicular variant-PTC	4		
Age			
< 55 years old	8		
≥ 55	2		
Tumor size (cm)			
≤ 1	4		
1–2	6		
Capsular invasion			
Negative	9		
Positive	1		
Lymph-vascular invasion			
Negative	6		
Positive	4		
Lymph node metastasis			
Negative	6		
Positive	4		
TNM stage^a			
T1bN0M0	3		
T1aN1aM0	2		
T1bN1aM0	1		
T1bN1bM0	1		
T1aNxMx	1		
T1aNxM0	1		
T3N0M0	1		

Table 2. Pathological characteristics of the tissues were listed. PTC; papillary thyroid carcinoma, ^aTNM staging system is developed and is maintained by the Union for International Cancer Control (UICC), TNM stands for Tumour, Node, Metastasis.

Tissue	Gene	Mean Cq	SD	C _q CV(%)	Minimum Cq	Maximum Cq	MFC
Normal	GAPDH	26.67	4.8	18.02	18.17	34.60	1.9
	TBP	30.64	2.20	7.19	26.45	33.83	1.27
	SYMPK	30.03	1.99	6.64	26.92	32.70	1.21
	PYCR1	33.99	1.65	4.86	31.15	36.21	1.16
	ACTB	26.06	3.48	13.35	21.78	31.75	1.45
	HPRT1	31.80	3.37	10.60	26.85	37.27	1.38
	B2M	24.82	3.29	13.26	20.39	29.77	1.46
	GUSB	32.73	2.89	8.84	28.82	38.15	1.32
PTC	GAPDH	24.73	3.02	12.23	19.34	29.63	1.53
	TBP ^a	31.01	2.13	6.88	28.01	34.93	1.24
	SYMPK ^a	29.88	1.54	5.16	27.22	32.04	1.17
	PYCR1 ^a	33.66	2.08	6.19	28.84	36.10	1.25
	ACTB	25.13	2.70	10.75	21.21	29.51	1.39
	HPRT1	31.52	2.66	8.46	27.35	35.54	1.29
	B2M	24.80	2.72	10.97	20.81	30.45	1.46
	GUSB	32.65	2.38	7.31	28.69	36.27	1.26

Table 3. Basic statistics were presented. ^aSYMPK, TBP, and PYCR1 were the best according to basic analysis using Microsoft Excel 2013 spreadsheet. C_q cycle of quantification, SD standard deviation, C_qCV Cq coefficient of variation, MFC maximum fold change.

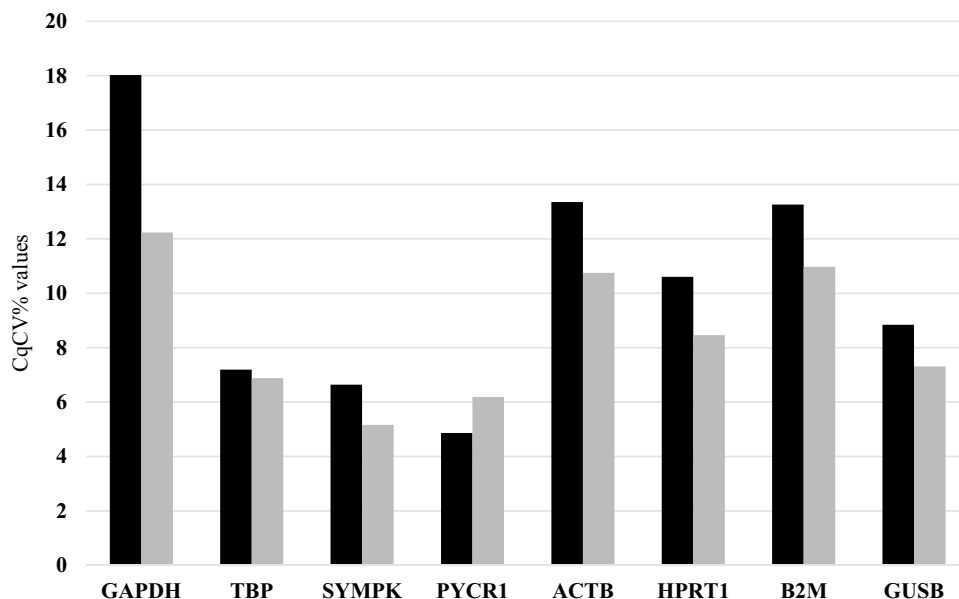


Figure 1. Basic statistical analysis representing CqCV% value of each reference gene. *PYCRI*, *SYMPK*, and *TBP*, respectively showed the lowest CqCV% values in normal tissues (black bars). The least CqCV% values were obtained for *SYMPK*, *PYCRI*, and *TBP*, respectively in PTC tissues (gray bars). The illustration confirmed pseudogene free-*SYMPK* with the lowest CqCV% value in PTC tissues.

Identification of the most stable reference genes using NormFinder. NormFinder intragroup analysis was initially used to rank the candidate reference genes from most to the least stable (Table 4). In normal tissues, *HPRT1* was the most stable reference gene, followed by *SYMPK*, *GUSB*, *ACTB*, *B2M*, *TBP*, *PYCRI*, and finally *GAPDH* (black bar, Fig. 2). In PTC tissues, *GUSB* was the most stable reference gene, followed by *SYMPK*, *ACTB*, *PYCRI*, *HPRT1*, *B2M*, *TBP*, and finally *GAPDH* (gray bars, Fig. 2). Thus, NormFinder considered *HPRT1* and *GUSB* as the best reference genes for normal and PTC tissues respectively, while *SYMPK* was the second most stable reference gene in both tissue types.

Next, intergroup analysis was performed, checking reference gene performance in PTC tissues against normal tissues. In this analysis, NormFinder declared as the single best reference genes *GUSB*, *SYMPK*, and *HPRT1* (dotted gray, Fig. 2), and the best combination of reference genes as *SYMPK* and *ACTB*, with a stability value of 0.209 (Table 4).

Identification of the most stable reference genes using BestKeeper. According to BestKeeper intragroup analysis (Table 5), the most stable gene in normal tissues was *ACTB*, with $r=0.974$. On the other hand, the most stable genes in PTC tissues were *ACTB* and *GUSB*, which tied with $r=0.966$. In normal tissues, the second and third most stable genes were *SYMPK* and *HPRT1*, with $r=0.969$ and $r=0.966$, respectively. The same ranking was observed in PTC tissues, with *SYMPK* ($r=0.958$) and *HPRT1* ($r=0.931$) having the second and third highest similarity to the BestKeeper index. Finally, PTC and normal intergroup analysis using BestKeeper ranked the candidate genes from most to least stable as follows: *ACTB*, *SYMPK*, *HPRT1*, *GUSB*, *GAPDH*, *B2M*, *PYCRI*, and *TBP* (Fig. 3).

Discussion

When performing qPCR normalization, a reference gene could be considered reliable if it is expressed consistently throughout different pathological statuses⁵. While truly constant expression might only be seen in fairy tales, a lack of confounding influence such as from pseudogenes can be achieved⁵. Omitting pseudogene signal could bring more stability to the observed expression of reference genes, hence our current approach of finding and validating genes with minimal expression variation and that lack pseudogenes^{19,23}.

Different statistical parameters could affect the interpretation of reference gene stability. The first factor that came to our attention was CqCV%, as it is more comprehensive than mean and SD. It must be mentioned that CqCV% or distribution frequency is a statistical factor that measures the dispersion of the data and the repeatability and precision of the experiment²⁴. In this study, *SYMPK* and *PYCRI* showed the lowest CqCV% in both tissues, meaning that these two genes had the lowest dispersion around the mean value. MFC is another basic statistical factor that indicates the dispersion, the variation, and the difference between minimum and maximum Cq. In the same manner, *SYMPK* and *PYCRI* showed the lowest MFC in both tissues.

According to the NormFinder algorithm, the best single reference gene was *GUSB* (stability value = 0.301); however, *ACTB* and *SYMPK* were suggested as the best combination, with a stability value of 0.209. Interestingly, despite fundamental differences in the analytical software, the BestKeeper algorithm suggested exactly the same two genes as the most stable. This aligns well with the basic statistical analysis, in which *SYMPK* also

Analysis type	Tissue	Candidated genes	Stability value
Intragroup	Normal	GAPDH	1.941
		TBP	1.030
		<u>SYMPK</u>	<u>0.704</u>
		PYCR1	1.305
		ACTB	0.747
		<u>HPRT1</u>	<u>0.667</u>
		B2M	1.013
		<u>GUSB</u>	<u>0.745</u>
	PTC	GAPDH	1.416
		TBP	1.295
		<u>SYMPK</u>	<u>0.458</u>
		<u>PYCR1</u>	<u>0.754</u>
		<u>ACTB</u>	<u>0.524</u>
		HPRT1	0.856
B2M		1.237	
GUSB		0.421	
Intergroup	Normal vs. PTC	GAPDH	0.698
		TBP	0.551
		<u>SYMPK^a</u>	<u>0.319</u>
		PYCR1	0.436
		ACTB	0.395
		<u>HPRT1^a</u>	<u>0.337</u>
		B2M	0.432
		<u>GUSB^a</u>	<u>0.301</u>
		Best combination of two genes	SYMPK and ACTB

Table 4. NormFinder analysis was presented. NormFinder analysis was used to rank the candidate reference genes. NormFinder considered *HPRT1* and *GUSB* as the best reference genes for normal and PTC tissues respectively, while *SYMPK* was the second most stable reference gene in both tissue types. Intergroup analysis indicated *SYMPK* and *ACTB*, with stability value of 0.209, as the best combination of reference genes. ^aGUSB, SYMPK, and HPRT1 were the best according to the Normfinder algorithm.

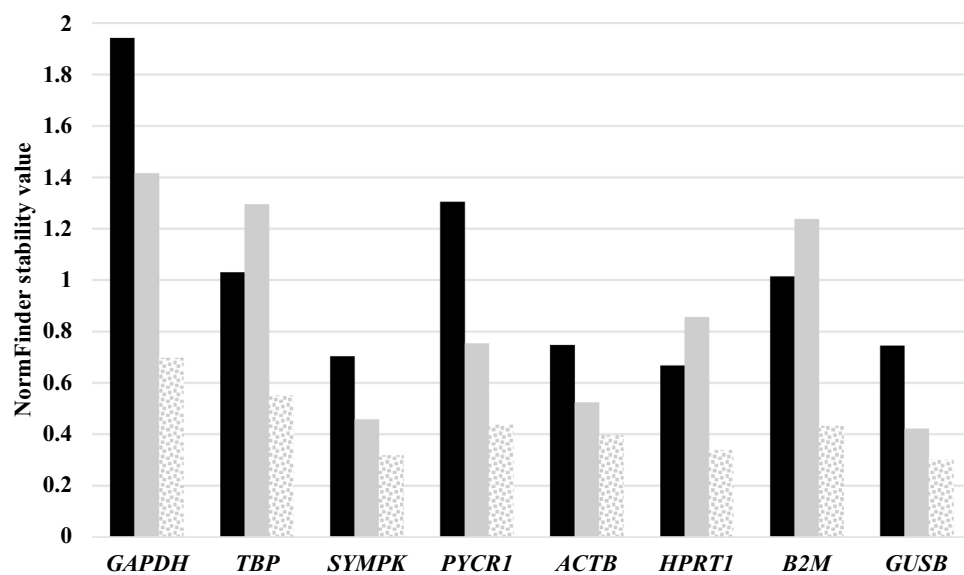


Figure 2. Intragroup and Intergroup Stability Value by NormFinder. *HPRT1* in normal tissues (black bar), and *GUSB* in PTC tissues (gray bars), were the most stable reference genes. *SYMPK* was the second most stable reference gene in both normal and PTC tissues. *SYMPK* with the narrowest margin, after *GUSB*, became a good candidate in Intergroup analysis (dotted gray).

Analysis type	Tissue	Symbol gene	Coefficient of correlation (r)
Intragroup	Normal	GAPDH	0.950
		TBP	0.883
		<u>SYMPK</u>	<u>0.969</u>
		PYCR1	0.844
		<u>ACTB</u>	<u>0.974</u>
		<u>HPRT1</u>	<u>0.966</u>
		B2M	0.918
		GUSB	0.933
	PTC	GAPDH	0.816
		TBP	0.666
		<u>SYMPK</u>	<u>0.958</u>
		PYCR1	0.831
		<u>ACTB</u>	<u>0.966</u>
		<u>HPRT1</u>	<u>0.931</u>
Intergroup	Normal vs. PTC	GAPDH	0.903
		TBP	0.771
		<u>SYMPK^a</u>	<u>0.963</u>
		PYCR1	0.807
		<u>ACTB^a</u>	<u>0.969</u>
		<u>HPRT1^a</u>	<u>0.951</u>
		B2M	0.847
		GUSB	0.941

Table 5. BestKeeper analysis was presented. BestKeeper intragroup analysis showed *ACTB* and *ACTB/ GUSB* as the most stable genes in normal and PTC tissues respectively. The second most stable gene was *SYMPK* according to the BestKeeper index. Finally, PTC and normal intergroup analysis ranked the candidate genes from most to least stable as follows: *ACTB*, *SYMPK*, *HPRT1*, *GUSB*, *GAPDH*, *B2M*, *PYCR1*, and *TBP*. ^a*ACTB*, *SYMPK*, and *HPRT1* were the best according to the BestKeeper algorithm.

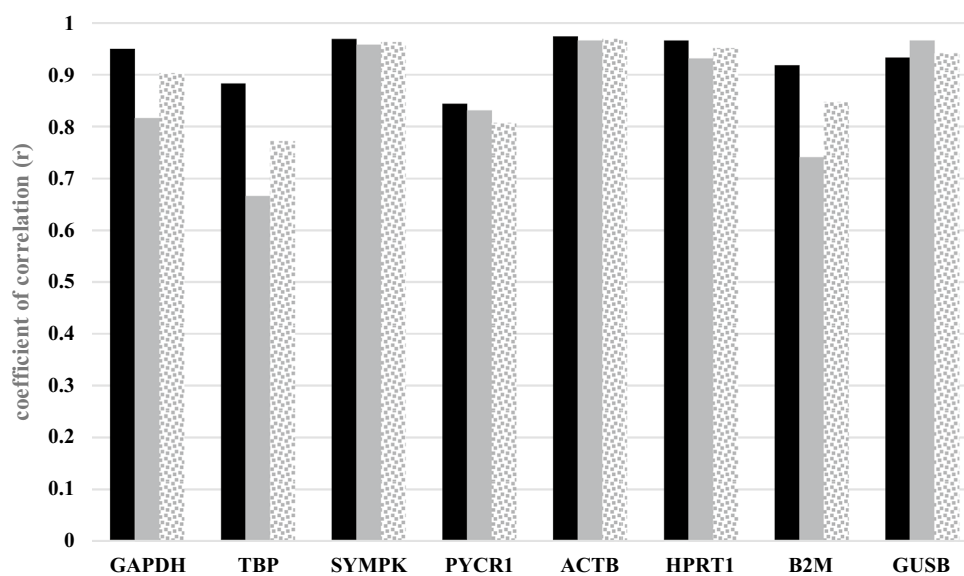


Figure 3. Intragroup and intergroup stability value by BestKeeper. Normal tissues (black bars), PTC tissues (gray bars), intergroup analysis (gray dotted bars).

had the best MFC value and CqCV%, and thus the aforementioned predictions from microarray data were in accordance with our labwork.

To date, very few studies have investigated reference genes in thyroid cell lines and human tissues. A previous study of six candidate reference genes (*ACTB*, *B2M*, *HPRT1*, *GAPDH*, *SDHA*, and *YWHAZ*) in seven goiter and seven normal tissues met with failure when intragroup analysis was done²⁵. However, they declared *SDHA* and *ACTB* as the most stable reference genes based on NormFinder intergroup analysis²⁵. In a same manner, this study found *ACTB* alone was not a well-qualified reference gene for intragroup analysis, but *SYMPK/ACTB* in combination had advantages in terms of basic statistics. Unfortunately, *ACTB* suffers from considerable pseudogene bias, and thus its reliability as a reference gene is dubious. Another study of 45 tissue samples from Iranian patients (15 PTC, 15 paired normal and 15 goiter tissues) checked twelve genes (*GAPDH*, *ACTB*, *HPRT1*, *TBP*, *B2M*, *PPIA*, *18SrRNA*, *HMBS*, *GUSB*, *PGK1*, *RPLP0*, and *PGM1*)²⁶. They declared consistency between the different algorithms of NormFinder, BestKeeper, and GeNorm because all three suggested *GUSB* and *HPRT1* as the most stably expressed genes in all thyroid tumours. In line with this study, we found *GUSB* to be a stable reference gene, but *GUSB* also has numerous pseudogenes that can obfuscate RT-qPCR data normalization. *HPRT1* was not considered a suitable reference gene in our study as it was only the third most stable reference gene according to both NormFinder and BestKeeper.

Considering the limitations described above, the ideal combination of *SYMPK/ACTB*, despite its stability value of 0.209, could not in fact be appropriate for RT-qPCR normalization. Pseudogene contamination of *ACTB* and also *GUSB* put these two genes out of the running^{19,27–29}. As pseudogene can cause falsely lower Cq values due to undesirable mRNA amplification, therefore *ACTB* cannot be helpful beside *SYMPK*. To avoid the problem of pseudogenes, the next most stable gene—*SYMPK*, with a stability value of 0.319—is the most suitable reference.

According to our knowledge, *SYMPK* was suggested as the second suitable reference gene only in one previous study; albeit after *CCSER2* gene³⁰. *CCSER2/SYMPK* inauguration was launched after transcriptome and microarray data analysis in breast cancer cell lines and tissues. Unfortunately we found *CCSER2* with a pseudogene (*LOC100127962*) sufficiently enough to be ruled out of this study. As a final point, *SYMPK* with the lack of pseudogenes and being presented only as a single isoform within cells showed mastery over the other competitor and it was selected in this study.

Conclusion

Synchronous results of BestKeeper and NormFinder suggested the combination of *SYMPK* and *ACTB* for RT-qPCR data normalization, but confounding from *ACTB* pseudogenes takes it out of the running and leaves *SYMPK* as the champion.

Subjects and methods

Reference gene selection. A total of eight reference genes were selected, with six based on literature review: *SYMPK*³⁰, Beta-2-microglobulin (*B2M*)²⁵, *GAPDH*^{26,31,32}, TATA-box binding protein (*TBP*)^{26,33}, *ACTB*^{26,34}, and Hypoxanthine phosphoribosyltransferase 1 (*HPRT1*)^{26,34}. The other two candidate reference genes were Pyrroline-5-carboxylate reductase 1 (*PYCRI*), which promotes cell proliferation in different human neoplasms but surprisingly does not fluctuate in PTC tumors (unpublished data), and β -glucuronidase (*GUSB*), selected based on a recently published study²⁶.

Reference gene validation using microarray data. Our selected reference genes were validated using the GSE3678 microarray dataset, which contains seven PTC samples and seven adjacent normal tissues. The dataset was checked for log scale and the quality of the data was assessed. Quantile normalization of the dataset was checked and principal component analysis (PCA) was performed (Supplementary Fig. 1). Finally, a list of genes with minimal variation was extracted and the presence of our selected reference genes in that list was confirmed. The microarray dataset was analyzed with an R program (script attached).

Exon-junction primer design. Exon-junction primer design was undertaken in order to eliminate pseudo-amplification of genomic DNA and/or heterogeneous nuclear RNA³⁵. Seven pairs of exon-junction primers were designed using Beacon Designer 8.1 (Premier Biosoft International, Palo Alto, CA, USA). To do this, the exact positions of introns were extracted from the Ensemble Genome Browser, and the exact nucleotide sequences of exons were imported into Beacon Designer 8.1 separately for each of our seven mRNAs. At the final step, the software was asked to design primers that span selected exon(s) in order to specifically amplify mature mRNAs of the desired genes. Primers that did not pass the exon-spanning requirement were rejected and not included in the study. For the eighth candidate gene, *GUSB*, primer sequences were taken directly from a recently published study²⁶. Secondary structures of the primers were re-checked using Oligo7, and to insure the selective amplification of mature mRNAs, specificity was confirmed using NCBI-primer BLAST. Melting temperatures for all pairs of primers were validated using gradient PCR (Sinaclon Bioscience, Tehran, Iran). The complete details of the designed primers, including exon-junction information, are listed in Table 1.

Tissue acquisition. A written informed consent was obtained before tissue acquisition. The study was approved by the Iran National Committee for Ethics in Biomedical Research Review Board and the Ethics Committee of Research Institute for Endocrine Sciences, Isfahan University of Medical Sciences, Isfahan, Iran (IR.Ul.REC.1398.058). PTC and adjacent normal tissues were obtained from patients undergoing thyroidectomy at the Alzahra and Sina hospitals, Isfahan province, Iran. Tissues were immediately submerged in adequate RNAlater RNA Stabilization Reagent (Qiagen, Hilden, Germany) and incubated at 4 °C for 24 h according to the manufacturer's instruction. If needed, samples were stored at – 80 °C. According to the postoperative pathologi-

cal and histological examinations, of the 20 thyroid tissue samples, ten were PTC and ten were adjacent normal thyroid tissues, each with the approximate size of 0.5 cm. Histopathological examinations were done by either the operating hospital or a third-party laboratory. The staging of the tumors was determined by specialized pathologists according to the 7th edition of the American Joint Committee on Cancer Tumor-Node-Metastasis (TNM) staging system. However, because of the restricted number of other thyroid neoplasm tissues in Iran, the data presented here are restricted to PTC tissues. We had only one anaplastic thyroid cancer sample, which was not included in the study. Patients' information and tumour pathological characteristics were presented in Table 2.

RNA extraction and quantification. Total RNA was extracted from RNAlater-treated tissues using a one-step RNA extraction reagent (BIOBASIC, Canada), according to the manufacturer's instructions. The concentration of the isolated RNA was measured using a NanoDrop oneC spectrophotometer (ThermoScientific, Waltham, MA, USA), and quality was determined using the A260/A280 and A260/A230 ratios. The integrity of extracted RNA was confirmed by 1.0% agarose gel electrophoresis (Invitrogen).

Complementary DNA (cDNA) synthesis. DNase I (ThermoScientific, Germany) was applied to eliminate residual genomic DNA according to the manufacturer's instruction. The ThermoScientific RevertAid Reverse Transcriptase kit was used to reverse transcribe 1 µg of total RNA according to the manufacturer's instruction. The reaction was done in a total volume of 20 µL.

Reverse transcription quantitative polymerase chain reaction (RT-qPCR). RT-qPCR reactions were performed in a Bio-Rad Chromo4 machine using SYBR Green RealQ Plus 2× Master Mix (Ampliqon, Odense, Denmark). The RT-qPCR reaction was as follows: one cycle of enzyme activation and initial denaturation at 95 °C for 15 min, then 40 cycles of 95 °C for 30 s, specific primer annealing temperature for 30 s, and 72 °C for 30 s. A plate reading was carried out after each cycle. All RT-qPCR reactions were run in triplicate, and a non-template control (NTC) was used for each run.

Melt curve analysis. The specificity of the RT-qPCR was evaluated using melt curve analysis through the gradual increase of temperature with a transition rate of 1 °C (from 55 to 95 °C). After each temperature increase, a plate reading was performed on the green channel. The derivative of fluorescence change over temperature (y axis) was plotted against the temperature (°C, x axis).

Statistical analysis. Microsoft Excel 2013 was used to calculate the mean C_q , standard deviation (SD), correlation of variation ($C_q CV\%$, $C_q CV\% = SD/mean \times 100\%$), minimum C_q , maximum C_q , and maximum fold change (MFC) (Table 1). MFC was calculated by dividing the maximum C_q by minimum C_q ; this value is an estimate that represents the distribution of the C_q , with a lower MFC value indicating a narrower distribution of a reference gene C_q ²⁵. NormFinder³⁶ and BestKeeper²⁴ were used to validate the appropriateness of reference genes. NormFinder ranks reference genes based on their stability value, with lower stability value indicating a more stable reference gene³⁶. BestKeeper creates its reference gene index using pair-wise correlations based on the average C_q values, SD, and CV²⁴. This algorithm considers the gene with the highest Pearson's correlation coefficient (r) as the most stable reference gene.

Ethical approval. The study was approved by the Iran National Committee for Ethics in Biomedical Research Review Board and the Ethics Committee of Research, Institute for Endocrine Sciences, Isfahan University of Medical Sciences, Isfahan, Iran (IR.UI.REC.1398.058). All experiments involving participants were performed in accordance with the seventh edition of the Helsinki declaration and the research protocol was first submitted to the Iran National Committee for Ethics in Biomedical Research Review Board and the Ethics Committee of Research, Institute for Endocrine Sciences, Isfahan University of Medical Sciences. Informed consent, including the aims, methods, and sources of funding, institutional affiliations of the researchers, the anticipated benefits and potential risks of the study, and post-study provisions, was obtained for each participant. We also informed each participant of their right to refuse.

Data availability

The dataset analyzed during the current study is available in the NCBI-Gene Expression Omnibus repository, [<https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1732>].

Received: 2 March 2020; Accepted: 14 October 2020

Published online: 27 October 2020

References

1. Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, 1. <https://doi.org/10.1186/gb-2002-3-7-research0034> (2002).
2. Nolan, T., Hands, R. E. & Bustin, S. A. Quantification of mRNA using real-time RT-PCR. *Nat. Protoc.* **1**, 1559–1582. <https://doi.org/10.1038/nprot.2006.236> (2006).
3. Ginzinger, D. G. Gene quantification using real-time quantitative PCR. *Exp. Hematol.* **30**, 503–512. [https://doi.org/10.1016/s0301-472x\(02\)00806-8](https://doi.org/10.1016/s0301-472x(02)00806-8) (2002).
4. Javadirad, S. M., Hojati, Z., Ghaedi, K., Nasr Esfahani, M. H. & Abbasy, B. Evaluation of the most stable reference genes in testicular tissue of the men with Azoospermia. *J. Isfahan Med. School (I.U.M.S)* **33**, 2407–2416 (2016).

5. Bustin, S. A. *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* **55**, 611–622. <https://doi.org/10.1373/clinchem.2008.112797> (2009).
6. Radonic, A. *et al.* Guideline to reference gene selection for quantitative real-time PCR. *Biochem. Biophys. Res. Commun.* **313**, 856–862. <https://doi.org/10.1016/j.bbrc.2003.11.177> (2004).
7. Arukwe, A. Toxicological housekeeping genes: do they really keep the house?. *Environ. Sci. Technol.* **40**, 7944–7949. <https://doi.org/10.1021/es0615223> (2006).
8. Selvey, S. *et al.* Beta-actin: an unsuitable internal control for RT-PCR. *Mol. Cell. Probes* **15**, 307–311. <https://doi.org/10.1006/mcpr.2001.0376> (2001).
9. Tanic, N., Perovic, M., Mladenovic, A., Ruzdijic, S. & Kanazir, S. Effects of aging, dietary restriction and glucocorticoid treatment on housekeeping gene expression in rat cortex and hippocampus—evaluation by real time RT-PCR. *J. Mol. Neurosci.* **32**, 38–46. <https://doi.org/10.1007/s12031-007-0006-7> (2007).
10. Thellin, O. *et al.* Housekeeping genes as internal standards: use and limits. *J. Biotechnol.* **75**, 291–295. [https://doi.org/10.1016/s0168-1656\(99\)00163-7](https://doi.org/10.1016/s0168-1656(99)00163-7) (1999).
11. Tricarico, C. *et al.* 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. [https://doi.org/10.1016/s0003-2697\(02\)00311-1](https://doi.org/10.1016/s0003-2697(02)00311-1) (2002).
12. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337. <https://doi.org/10.1038/nature12624> (2013).
13. Ramon, Y. C. S. *et al.* Clinical implications of intratumor heterogeneity: challenges and opportunities. *J. Mol. Med. (Berl)* **98**, 161–177. <https://doi.org/10.1007/s00109-020-01874-2> (2020).
14. Martin-Pardillos, A. *et al.* The role of clonal communication and heterogeneity in breast cancer. *BMC Cancer* **19**, 666. <https://doi.org/10.1186/s12885-019-5883-y> (2019).
15. Losic, B. *et al.* Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat Commun* **11**, 291. <https://doi.org/10.1038/s41467-019-14050-z> (2020).
16. Chmielik, E. *et al.* Heterogeneity of thyroid cancer. *Pathobiol. J. Immunopathol. Mol. Cell. Biol.* **85**, 117–129. <https://doi.org/10.1159/000486422> (2018).
17. Harrison, P. M., Echols, N. & Gerstein, M. B. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res* **29**, 818–830. <https://doi.org/10.1093/nar/29.3.818> (2001).
18. Chen, S. M., Ma, K. Y. & Zeng, J. Pseudogene: lessons from PCR bias, identification and resurrection. *Mol. Biol. Rep.* **38**, 3709–3715. <https://doi.org/10.1007/s11033-010-0485-4> (2011).
19. Sun, Y., Li, Y., Luo, D. & Liao, D. J. Pseudogenes as weaknesses of ACTB (Actb) and GAPDH (Gapdh) used as reference genes in reverse transcription and polymerase chain reactions. *PLoS ONE* **7**, e41659. <https://doi.org/10.1371/journal.pone.0041659> (2012).
20. Roth, R., Madhani, H. D. & Garcia, J. F. Total RNA isolation and quantification of specific RNAs in fission yeast. *Methods Mol. Biol.* **1721**, 63–72. https://doi.org/10.1007/978-1-4939-7546-4_6 (2018).
21. Paska, C., Barta, I., Drozdovszky, O. & Antus, B. Elimination of bacterial DNA during RNA isolation from sputum: bashing bead vortexing is preferable over prolonged DNase treatment. *PLoS ONE* **14**, e0214609. <https://doi.org/10.1371/journal.pone.0214609> (2019).
22. Kuang, J., Yan, X., Genders, A. J., Granata, C. & Bishop, D. J. An overview of technical considerations when using quantitative real-time PCR analysis of gene expression in human exercise research. *PLoS ONE* **13**, e0196438. <https://doi.org/10.1371/journal.pone.0196438> (2018).
23. Chapman, J. R. & Waldenstrom, J. With reference to reference genes: a systematic review of endogenous controls in gene expression studies. *PLoS ONE* **10**, e0141853. <https://doi.org/10.1371/journal.pone.0141853> (2015).
24. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. 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. <https://doi.org/10.1023/b:bile.0000019559.84305.47> (2004).
25. Weber, R. *et al.* Validation of reference genes for normalization gene expression in reverse transcription quantitative PCR in human normal thyroid and goiter tissue. *Biomed. Res. Int.* **2014**, 198582. <https://doi.org/10.1155/2014/198582> (2014).
26. Razavi, S. A. *et al.* Validation of reference genes for normalization of relative qRT-PCR studies in papillary thyroid carcinoma. *Sci. Rep.* **9**, 15241. <https://doi.org/10.1038/s41598-019-49247-1> (2019).
27. Sargent, C. A., Chalmers, I. J., Leversha, M. & Affara, N. A. A rearrangement on chromosome 5 of an expressed human beta-glucuronidase pseudogene. *Mamm Genome* **5**, 791–796. <https://doi.org/10.1007/BF00292015> (1994).
28. Speleman, F. *et al.* Localization by fluorescence in situ hybridization of the human functional beta-glucuronidase gene (GUSB) to 7q1121 → q1122 and two pseudogenes to 5p13 and 5q13. *Cytogenet. Cell Genet.* **72**, 53–55. <https://doi.org/10.1159/000134161> (1996).
29. Ng, S. Y. *et al.* Evolution of the functional human beta-actin gene and its multi-pseudogene family: conservation of noncoding regions and chromosomal dispersion of pseudogenes. *Mol. Cell. Biol.* **5**, 2720–2732. <https://doi.org/10.1128/mcb.5.10.2720> (1985).
30. Tilli, T. M., Castro Cda, S., Tuszynski, J. A. & Carels, N. A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases. *BMC Genomics* **17**, 639. <https://doi.org/10.1186/s12864-016-2946-1> (2016).
31. Jo, J. *et al.* Conventionally used reference genes are not outstanding for normalization of gene expression in human cancer research. *BMC Bioinformatics* **20**, 245. <https://doi.org/10.1186/s12859-019-2809-2> (2019).
32. Barber, R. D., Harmer, D. W., Coleman, R. A. & Clark, B. J. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics* **21**, 389–395. <https://doi.org/10.1152/physiolgenomics.00025.2005> (2005).
33. Lisowski, P., Pierzchala, M., Goscik, J., Pareek, C. S. & Zwierzchowski, L. Evaluation of reference genes for studies of gene expression in the bovine liver, kidney, pituitary, and thyroid. *J. Appl. Genet.* **49**, 367–372. <https://doi.org/10.1007/BF03195635> (2008).
34. Chantawibul, S., Anuwong, A. & Leelawat, K. Validation of appropriate reference genes for gene expression studies in human thyroid gland using real-time RT-PCR. *J. Med. Assoc. Thailand* **95**(Suppl 3), S36–S40 (2012).
35. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinf.* **13**, 134. <https://doi.org/10.1186/1471-2105-13-134> (2012).
36. Andersen, C. L., Jensen, J. L. & Orntoft, T. F. 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. *Can. Res.* **64**, 5245–5250. <https://doi.org/10.1158/0008-5472.CAN-04-0496> (2004).

Acknowledgements

This study has been conducted in Isfahan University of Iran and was supported financially by the Departments of Research, Technology and Graduate Offices (Grant number: A\10979\97). Authors sincerely thank the volunteers for their participation.

Author contributions

S.-M.J.: Conception, design, assembly of data, data analysis, interpretation, financial supports, drafting the manuscript, revising it critically for important intellectual content, and final approval of the manuscript. M.M.: Conception, design, collection, and/or assembly of data, data analysis, interpretation, and drafting of the manuscript. G.E.: Conception, design, collection, and/or assembly of data, data analysis, interpretation, and drafting of the manuscript. M.K.: Conception, Surgical, and pathological approval Drafting the manuscript and final approval of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-75495-7>.

Correspondence and requests for materials should be addressed to S.-M.J.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020