Which Reference Genes to Choose for qPCR Normalization? A Comprehensive Analysis in MCF-7 Breast Cancer Cell Line

Nityanand Jain
Rigas Stradinas Universitate
Dina Nitisa
Rigas Stradinas Universitate
Valdis Pirsko
Rigas Stradinas Universitate
Inese Cakstina (inese.cakstina@rsu.lv)
Riga Stradiņš University https://orcid.org/0000-0001-5589-0395

Research article

Keywords: MCF-7, RT-qPCR, reference genes, gene expression, sub-clones, breast cancer cell line

Posted Date: July 30th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-42293/v1

Version of Record: A version of this preprint was published on September 25th, 2020. See the published version at https://doi.org/10.1186/s12860-020-00313-x.

Abstract

Background

MCF-7 cell line remains the most extensively studied patient derived model in breast cancer research, providing pivotal results that have over the decades translated to constantly improving patient care. Many research groups, in the past have identified suitable reference genes in various breast cancer cell lines including MCF-7. However, over the course of identification of stable internal controls, a comparative analysis comprising these genes together in a single study have not been previously undertaken. Further, very little is known about how these identified reference genes are expressed in MCF-7 sub-clones and when the cell line is cultured over multiple passages (p), given the heterogenic expression behavior often associated with MCF-7 cell line. We investigated the expression dynamics of 12 previously reported and suggested endogenous reference genes using RT-qPCR, available algorithms (NormFinder, geNorm, BestKeeper etc.) and TCGA transcriptomic analysis within identically cultured two sub-clones (culture A1 and A2) of MCF-7 cell line cultured over multiple passages. Further candidate reference genes were used to normalize 4 genes of interest (2 simulated and 2 backed by qPCR data) to make an evidence-based recommendation of the least variable reference genes that could be used in MCF-7 cell line.

Results

The analysis revealed the presence of differential reference gene expression within the sub-clones of MCF-7. In culture A1, *GAPDH-CCSER2* were identified as least variable reference gene pair while for culture A2, *GAPDH-RNA28S* was recommended. However, upon validation using genes of interest, both these pairs were found to be unsuitable control pairs. Normalization with 3 genes and analyzing the combined dataset from culture A1 and A2 (supported by transcriptomic analysis), *GAPDH-PCBP1-CCSER2* triplet was found to be least variable and hence potentially more well placed to handle any expression heterogeneity that may arise within sub-clones over multiple passages.

Conclusions

The variance in expression behavior amongst sub-clones shows the need for exercising caution while selecting and using reference genes for MCF-7. Further, using same reference genes amongst sub-clones can lead to misleading results arising from inaccurate normalization. *GAPDH-PCBP1-CCSER2* triplet offers a reliable alternative to otherwise traditionally used internal controls in optimizing intra- and inter-assay gene expression differences.

1. Background

MCF-7 (Michigan Cancer Foundation 7) is one of the most commonly used patient derived breast cancer cell line, which was established in 1970 by researcher Herbert D. Soule at the Michigan Cancer

Foundation from the pleural effusion and chest wall nodule showing metastasis of breast adenocarcinoma. The number 7 in MCF-7 represents Soule's seventh attempt in which he succeeded in generating a cancer cell line [1]. The universal adoption of the cell line is evident from the fact that a simple PubMed search (Search word: MCF-7) can retrieve about 39,000 citations (from 1973 to March 2020) with about 900 citations already being reported in the first three months of 2020 alone.

In 2013, an updated approach based on IHC (immuno-histochemistry) markers was introduced by St. Gallen International Expert Consensus on Primary Therapy of Early Breast Cancer to determine subtypes of breast cancers [2]. Accordingly, luminal A subtype tumors were defined as estrogen receptor positive (ER+), progesterone receptor (PR) $\geq$ 20%, HER2 negative, Ki67 proliferation marker < 14% and if available, low recurrence risk tumors based on gene-based assays [2]. MCF-7 neoplastic cells were found to be positive for both Estrogen (ER) and Progesterone (PR) receptors along with having low metastatic activity and hence fulfilled the criteria to be classified as luminal A molecular subtype tumor cell line [3].

Cell based assays (cell lines) such as MCF-7 represents techniques that can provide more biologically meaningful information than simplified biochemical assays [4]. The key reasons for their universal adoption are lower operational costs and the ease of operation in terms of preparing and observing the cells. Further, they represent an unlimited self-replicating source that can be grown in almost infinite quantities [5] yielding unlimited amount of DNA/RNA that enables studies related to validation and downstream functional analysis.

However, MCF-7 cell line like other cell lines is also prone to certain disadvantages. It is vulnerable to genotypic and phenotypic drift during its long-term culturing [5] and is of profound concern since the cell line has been deposited in cell banks for many years now. This has risked and in some certain cases caused arising of subpopulations within the cell line. Subpopulations can cause phenotypic changes over time by the selection of specific, more rapidly growing clones within a population as demonstrated by Osborne et al. [6] and Resnicoff et al. [7] in 1987.

Over the past decades, extensive evidence that MCF-7 cells showed clonal variations has been reported depicting either differences in phenotypic traits such as estrogen/progesterone responsiveness, epidermal growth factor (EGF) expression or the ability to form tumors in syngeneic mice [8]. Also, genetic variability in the sublines and sub-clones of MCF-7 cell line on karyotypic and chromosomal levels have also been demonstrated by various researchers [9, 10, 11, 12, 13, 14, 15]. Finally, in 2016 Andre et al., illustrated variations among MCF-7 cell line obtained from the same cell bank in the same batch [16]. Presence of such heterogeneity bolsters the need for validating and cross-examining the genetic variability as well as gene expression in the cell line.

A widely accepted technique for validation of gene expression is the Reverse Transcription – quantitative Polymerase Chain Reaction (RT-qPCR). It is highly sensitive, reproducible, simple and high throughput yielding technique that can confirm gene expression differences and measure transcript abundances [17, 18]. Nevertheless, the results obtained from RT- qPCR still needs to be normalized against another data

set or set references to correct for any sampling noise such as differences in the amount of starting material in order to estimate results accurately [19].

These reference genes (previously housekeeping genes) are expressed constitutively and are required for the maintenance of basal cellular functions. In general practice, it is presumed that the endogenous reference gene represents an ideal gene that is sufficiently abundant and has stable expression across different tissues and cell lines under different experimental conditions [20]. In addition, it is assumed that the expression levels remain the same amongst biological replicate cell cultures over successive passages. However, some studies suggest that the expression of these reference genes may not be as uniform as previously thought and may also fluctuate significantly under different experimental conditions [21, 22]. Hence, it becomes imperative to validate reference genes before their use in a study as using a non-validated reference gene could lead to misleading interpretations arising from inaccurate results [20, 23].

As pointed out in the MIQE guidelines [24], normalization against a single reference gene is not recommended unless a clear evidence of its uniform expression dynamics is described for the specific experimental conditions. Many studies have been undertaken previously identifying new stable reference genes over the previous two decades for MCF-7 cell line or breast cancer as whole [25, 26, 27, 28, 29, 30, 31, 32].

However, these studies didn't undertake a detailed analysis validating the reference gene expression over multiple passages and/or, within sub-clones (biological replicate cultures). This gap in validation leaves us with a cause of serious concern especially if such cell lines are to be regarded as valid models for evaluating the behavior and development of breast cancer and validating their likely response to new drugs and therapies.

The present study, therefore, aims to fill the void by investigating the gene expression of twelve reference genes that were previously identified as stable genes by various studies [25, 26, 27, 28, 33, 34, 35, 36, 37] for MCF-7 cell line but were not accounted and studied together so as to make an evidence supported recommendation of reference genes to be used for normalization of gene of interest in routinely cultured MCF-7 cell line. Further, the study investigated for any differential reference gene expression in two sub-clones (culture A1 and A2) cultured over multiple passages (p) to further bolster the selection of appropriate reference genes.

2. Results

2.1. Curation of the Dataset and Descriptive Analysis

Three lysates were collected from each passage from both MCF-7 cultures A1 and A2 while two lysates were collected and evaluated for passage 28 (p28) and passage 31 (p31) from culture A1. A detailed selection criteria of candidate reference genes is described in Sect. 5.3. Amplification for each of the 12 reference genes produced a dataset with 900 Cq values. As shown in Table 1, *RNA18S* showed the

highest expression in both cultures A1 and A2 (Cq mean = 7.93 and 8.26 respectively), closely followed by *RNA28S* (Cq mean = 8.27 and 8.35 respectively). Both genes showed high amplification levels, appearing close to seven cycles earlier than any other gene in both cultures (*ACTB* Cq mean = 15.98 and 16.11 respectively). *CCSER2* presented the lowest expression levels in both cultures (Cq mean = 26.58 and 26.56 respectively). *GAPDH* showed the lowest standard deviation (S.D = 0.30) in culture A1 while *ACTB* showed the lowest standard deviation (S.D = 0.32) in culture A2. The largest variation between Cq values was shown by *HNRNPL* (S.D = 0.35) in culture A1, while in culture A2, *PGK1* (S.D = 0.61) showed the largest difference.

Table 1
Descriptive Statistics of the Reference Genes Cq (Quantification Cycle) Values

Gene	N *		Mean Cq ± S.D **		Minimum Cq		Maximum Cq	
	A1	A2	A1	A2	A1	A2	A1	A2
ACTB	30	45	15.98 ± 0.26	16.11 ± 0.32	15.38	15.53	16.45	16.62
GAPDH	30	45	17.13 ± 0.17	17.07 ± 0.39	16.78	16.43	17.52	17.67
RPL13A	30	45	21.03 ± 0.29	20.74 ± 0.43	20.71	20.14	21.64	21.70
PGK1	30	45	21.08 ± 0.22	21.09 ± 0.61	20.79	20.29	21.76	22.11
HSPCB	30	45	20.40 ± 0.29	20.42 ± 0.51	20.02	19.70	20.98	21.77
RNA28S	30	45	8.27 ± 0.23	8.35 ± 0.40	7.61	7.83	8.64	9.58
RNA18S	30	45	7.93 ± 0.27	8.26 ± 0.49	7.45	7.58	8.47	9.83
PUM1	30	45	23.14 ± 0.23	23.08 ± 0.39	22.73	22.35	23.68	24.23
CCSER2	30	45	26.58 ± 0.21	26.56 ± 0.39	26.04	26.03	26.95	27.32
HNRNPL	30	45	22.73 ± 0.35	22.91 ± 0.56	22.11	21.97	23.52	23.99
PCBP1	30	45	22.13 ± 0.22	22.17 ± 0.37	21.78	21.65	22.76	22.94
SF3A1	30	45	23.39 ± 0.34	23.51 ± 0.43	22.99	22.78	24.28	24.68

* N – Total number of Cq values from all triplicates of 3 lysates. ** S.D – Standard Deviation.

2.2. Coefficient of Variation (CV) and Linearization of the Cq Values

The raw Cq values are not always reliable to be used for assessing the gene expression dynamics across experimental groups [38] since they don't faithfully recapitulate the extent of variation in actual RNA quantities. Hence, the raw Cq values were converted to a linear scale using 2^{-Cq} for analysis of Coefficient of Variation (CV). The CV or Relative Standard deviation (RSD) is calculated as the ratio of the standard deviation and mean of the same expressed in terms of percentage. A lower CV% indicated

higher stability of the gene expression and vice versa [39]. It has been reported that for a heterogenous sample like breast cancer cells, a CV% lower than 50% is ideal for reference genes to be considered to have stable expression [40].

Table 2 summarizes the CVs of candidate reference genes in both cultures A1 and A2. Across both cultures, the CVs were in the range of 14.89–40.11% indicating that all the candidate reference genes were characterized by stable expression. The CV analysis revealed that *PCBP1* was the only gene to be in the top 3 stable genes (CV% = 14.89% (A1) and 24.90% (A2)) in both the cultures. Apart from *PCBP1*, the other 2 top stable genes in culture A1 were *GAPDH* (CV% = 12.35%) and *PGK1* (CV% = 14.86%), while in culture A2 those were *ACTB* (CV% = 22.98%) and *RNA28S* (CV% = 24.52%). In stark contrast, *PGK1* in culture A2 was the least stable gene (CV = 40.11%) showing variations in gene expression between the two cultures.

Table 2

Mean 2^{-Cq}, Standard Deviation 2^{-Cq} and CV% of the candidate reference genes

Candidate Reference Gene	MCF-7 Culture	Mean 2^{-Cq}	S.D. 2^{-Cq}	CV (S.D/Mean)
ACTB	A1	1.6E-05	2.9E-06	18.81%
	A2	1.4E-05	3.3E-06	22.98%
GAPDH	A1	7.0E-06	8.7E-07	12.35%
	A2	7.5E-06	2.0E-06	26.84%
RPL13A	A1	4.7E-07	8.7E-08	18.39%
	A2	5.9E-07	1.6E-07	26.71%
PGK1	A1	4.5E-07	6.7E-08	14.86%
	A2	4.9E-07	2.0E-07	40.11%
HSPCB	A1	7.4E-07	1.4E-07	19.34%
	A2	7.5E-07	2.4E-07	31.31%
RNA28S	A1	3.3E-03	5.8E-04	17.66%
	A2	3.1E-03	7.7E-04	24.52%
RNA18S	A1	4.2E-03	8.0E-04	19.17%
	A2	3.4E-03	9.9E-04	28.97%
PUM1	A1	1.1E-07	1.7E-08	15.65%
	A2	1.2E-07	2.9E-08	25.03%
CCSER2	A1	1.0E-08	1.6E-09	15.55%
	A2	1.0E-08	2.7E-09	25.65%
HNRNPL	A1	1.5E-07	3.5E-08	23.70%
	A2	1.4E-07	5.2E-08	38.21%
PCBP1	A1	2.2E-07	3.3E-08	14.89%
	A2	2.2E-07	5.5E-08	24.90%
SF3A1	A1	9.3E-08	2.0E-08	21.20%
	A2	8.7E-08	2.4E-08	27.76%

2.3. Relative Mean Changes in Expression Profiles of Selected Candidate Reference Genes

Relative mean changes in expression profiles were analyzed to study the gene expression stability and variations over successive passages in the both cultures. Passage 28 (p28) was selected as the experimental calibrator for culture A1 while Passage 25 (p25) was selected as the calibrator for culture A2, since they represented the initial investigated passages in both cultures. The fold change was then calculated as $2^{-\Delta Cq}$ with the other passages in both cultures (Figs. 1 and 2).

To determine significant relative expression changes, one-way ANOVA was used (Supplementary Table S1, see Additional file 1). In culture A1, only *CCSER2* (Fig. 2C) and *PCBP1* (Fig. 2E) showed no significant expression changes over successive passages (*ANOVA P* > 0.05). In culture A2, all the genes selected, including *CCSER2* and *PCBP1* showed significant expression changes over successive passages. Both *ACTB* (Fig. 1A) and *GAPDH* (Fig. 1B) showed significant expression differences (*ANOVA P* < 0.01) in both cultures over successive passages, providing evidence that these should be avoided to be used together, if possible, as endogenous reference genes in MCF-7 cell lines.

2.4. NormFinder Analysis of MCF-7 Sub-clones

NormFinder [41] estimates the overall variation of gene expression for each candidate reference gene and delivers a stability value, not only identifying the most stable gene but also the best control gene as shown in Fig. 3. It presents the stability of a gene as an estimate of the combined intra- and intergroup variation of the individual gene. It means that the gene with high expression stability is shown by a low standard deviation value and vice-versa.

Further, the algorithm can be used to find the pair of genes best suited to work together as reference genes as shown in Supplementary Table S2 (see Additional file 1) (only top 10 pairs for both the cultures with their pairwise standard deviations). The algorithm ranked *ACTB* and *PUM1* as the most stable pair (S.D = 0.07) in culture A1 while in culture A2 the first spot was shared by 2 pairs of genes – *RPL13A* – *SF3A1* and *GAPDH* – *SF3A1* (S.D = 0.07).

2.5. geNorm Analysis and Determination of Optimal Number of Genes Needed for Normalization of Dataset

The geNorm algorithm [20] calculates M value for each candidate reference gene based on pairwise comparison. The higher the M value, the less stable the reference gene and vice-versa (Fig. 4). geNorm uses the stepwise exclusion method of the least stable genes by calculating the average M values. It has been reported that for a heterogenous tissue, such as breast cancer, a stability M value of < 1 should be considered for candidature of reference genes. Any gene having stability value M above 1 should not be considered as candidate for reference gene [40]. There were noticeable differences in the results in both cultures. In culture A1, *ACTB-HSPCB* tied for the first position with a stability value M = 0.169 (Fig. 4A) while in culture A2, *RNA18S-RNA28S* tied for the most stable gene (M = 0.177; Fig. 4B).

geNorm can also be used to determine the optimal number of reference genes needed for an accurate estimation of normalization of expression data as shown in Fig. 5. In principle, Vn/Vn + 1 should be less than 0.15, where Vn represents the number of reference genes suitable for normalization. Although it is

suggested that 0.15 cutoff be used, it has been reported that this cutoff may not be used strictly to assess the Vn/Vn + 1, especially if Vn/Vn + 1 is 2/3. In such a case, use of three stable control genes should be enough to provide accurate results. For both the cultures A1 and A2, the V2/3 (0.00437 and 0.00468 respectively) was less than the recommended cutoff, indicating addition of a third reference gene would not make a difference in the normalization results.

2.6. BestKeeper Analysis

BestKeeper [42] was used to analyze the candidate reference genes expression stability. BestKeeper uses crossing points (CP) to decide, whether the genes are differentially expressed under the applied conditions or not. To arrive at a hierarchy of the best reference gene according to BestKeeper, different criteria have been suggested by Pfaffl [42], such as considering the i) standard deviation with crossing points (S.D ± CP; recommended cutoff ≤ 1), ii) standard deviation with changes in x-folds (S.D ± x-fold; recommended cutoff ≤ 2) and iii) coefficient of correlation (r; recommended to be as high as possible). Different authors use and report different criteria from the above-mentioned possibilities. For comparison in our study, standard deviation with crossing points was considered first for both the cultures (Fig. 6A and B).

As seen in other two previous methods (Sect. 2.4 and 2.5), BestKeeper also showed deviations in results for both cultures A1 and A2. While *GAPDH* (S.D = 0.14) and *CCSER2* (S.D = 0.18) were the top two genes in culture A1, *ACTB* (S.D = 0.28) and *RNA28S* (S.D = 0.30) took the top two spots in culture A2. Next, coefficient of correlation (r) given as Pearson's correlation by BestKeeper was evaluated to look for pairwise gene expression stability in both cultures (Fig. 7A and 7B).

The top five gene pairs with high positive coefficient of correlation (r > 0.5) in culture A1 (Fig. 7A) were *HSPCB-ACTB* (r = 0.833), *RPL13A-RNA18S* (r = 0.825), *PGK1-HSPCB* (r = 0.775), *PUM1-PCBP1* (r = 0.704) and *PUM1-SF3A1* (r = 0.598). Similarly, upon analysis of gene pairs in culture A2 (Fig. 7B), the top five gene pairs with r > 0.5 were, *PGK1-HNRNPL* (r = 0.948), *RNA28S- RNA18S* and *PGK1-GAPDH* (both pairs with r = 0.946), *PCBP1-SF3A1* (r = 0.933) and *PGK1-HSPCB* (r = 0.908).

2.7. Pairwise Comparative ΔCt Analysis of the MCF-7 Sub-clones

Comparative ΔCt [43] compares the relative expression of pairs of candidate reference genes within each sample in order to identify and rank the most stable genes. According to Silver et al., if the ΔCq value of the two genes fluctuate when analyzed in different samples, it can be concluded that, one or both genes are variably expressed [43]. Conversely, if the ΔCq value remains constant, both genes are expressed stably or are co-regulated among the samples. For a stable gene, it has been suggested that the gene should be strongly expressed, displays minimal fluctuations and is independent of expression of other genes. The ranking of the genes was based on the average Standard Deviation (S.D). The higher the S.D, the more variable the expression and less stable the gene and vice-versa (Table 3).

Table 3
Comparative ΔCt results with average Standard deviation (S.D) of genes in both cultures

Culture A1			Culture A2		
Candidate Gene	**Average S.D**	**Rank**	**Candidate Gene**	**Average S.D**	**Rank**
GAPDH	0.29	1	*RNA28S*	0.28	1
CCSER2	0.30	2	*GAPDH*	0.28	1
PCBP1	0.30	2	*PCBP1*	0.28	1
PGK1	0.31	3	*RPL13A*	0.30	2
ACTB	0.32	4	*HSPCB*	0.31	3
PUM1	0.33	5	*HNRNPL*	0.32	4
HSCPB	0.36	6	*ACTB*	0.32	4
RNA18S	0.37	7	*RNA18S*	0.32	4
RNA28S	0.37	7	*SF3A1*	0.32	4
RPL13A	0.37	7	*CCSER2*	0.33	5
HNRNPL	0.42	8	*PGK1*	0.38	6
SF3A1	0.43	9	*PUM1*	0.39	7

In culture A1, *GAPDH*, *CCSER2* and *PCBP1* were the top three genes ranked by the algorithm while in culture A2, *RNA28S*, *GAPDH* and *PCBP1* were the top three genes. It is interesting to note that in culture A1, the three top genes ranked by Comparative ΔCt (Table 3) are ranked in the same order with same rank by BestKeeper (Fig. 6A). In culture A2, only *RNA28S* has been reported in the top three by both these algorithms (Table 3; Fig. 6A).

2.8. RefFinder Analysis

RefFinder [44] is a user friendly, easy to use web based comprehensive tool (https://www.heartcure.com.au/reffinder/) that uses NormFinder, geNorm, BestKeeper and Comparative ΔCt to rank and compare candidate reference genes. It measures the geometric mean of attributed weights by these algorithms for generating an overall final ranking. The rankings from RefFinder are summarized in Table 4. The top three genes in culture A1 were reported to be *GAPDH*, *CCSER2* and *PCBP1*. For culture A2, *RNA28S*, *GAPDH* and *PCBP1* were the top three most stable genes.

Table 4
Ranking of the candidate reference genes by RefFinder for both cultures A1 and A2

Culture A1			Culture A2		
Candidate Gene	**Geomean**	**Rank**	**Candidate Gene**	**Geomean**	**Rank**
GAPDH	1.41	1	*RNA28S*	1.41	1
CCSER2	2.78	2	*GAPDH*	2.89	2
PCBP1	3.22	3	*PCBP1*	3.83	3
ACTB	3.64	4	*ACTB*	4.14	4
PGK1	3.72	5	*RPL13A*	4.43	5
HSPCB	4.86	6	*RNA18S*	4.90	6
PUM1	6.24	7	*HSPCB*	5.62	7
RNA28S	7.49	8	*CCSER2*	7.75	8
RNA18S	8.24	9	*HNRNPL*	8.24	9
RPL13A	9.49	10	*SF3A1*	8.35	10
HNRNPL	11.24	11	*PUM1*	8.49	11
SF3A1	11.74	12	*PGK1*	11.24	12

2.9. Moving Towards Selection of Reference Gene/s Based on an Integrated Approach

Coefficient of Variation (CV) is the only method where the stability of the gene is not influenced by any other genes in the study and hence can help in identifying a gene with overall high passage variance [38]. As mentioned above in the CV section, any gene having CV% greater than 50% (threshold) should be excluded from the study. Since none of our genes had CV% greater than the threshold, the next step was to evaluate the results from the five different approaches as mentioned in Sect. 2.3 to Sect. 2.8.

Based on these approaches, we came to a conclusion that for culture A1, *GAPDH* and *CCSER2* were suitable reference genes as they both were constantly ranked in top three in NormFinder (Fig. 3A), BestKeeper (Fig. 6A), Comparative ΔCt (Table 3) and RefFinder (Table 4). Further, *CCSER2* had no significant relative mean expression change over successive passages (Fig. 2C) while *GAPDH* had shown the least CV% (Table 2) in culture A1. Also, the Pearson Correlation between *GAPDH-CCSER2* ($r = 0.515$) was significant ($P = 0.004$) thereby indicating their positive correlation (Fig. 7A) and corroborating their selection. Based on a similar approach, for culture A2, *GAPDH* and *RNA28S* were selected as suitable reference genes. Interestingly, to not exclude a potential candidate for reference gene just because of strict selection of only 2 reference genes as suggested by geNorm (Fig. 5), an experimental candidate

with constant high rankings in all the algorithms for both the cultures was also chosen. *PCBP1* was identified as a strong contender in the race and hence was included in the validation of selected reference genes (*Sect. 2.10* to Sect. 2.11). *PCBP1* was one of the two genes to have no significant relative expression changes in culture A1 (Fig. 2E) and claims the top spots in NormFinder (Fig. 3), BestKeeper (Fig. 6), Comparative ΔCt (Table 3) and RefFinder (Table 4). It also maintains a constant position in top five genes per geNorm analysis for both cultures (Fig. 4A and 4B).

2.10. Generation of Data for 2 Genes of Interest (GOIs) & Validation of Reference Genes Using Normalization

To further evaluate the four selected reference genes (*GAPDH*, *RNA28S*, *CCSER2* and *PCBP1*) and their utility as reference genes, two simulated datasets were created. The genes simulated were referred to as *Gene of Interests (GOI) 1 and 2.* Both were assigned some random Cq values (triplicates for 3 lysates over 4 passages for culture A1 and over 5 passages for culture A2) as presented in Supplementary table S3 and S4 (see Additional file 1).

While assigning the Cq values for *GOI 1 (simulated to have stable gene expression)*, it was considered that the difference between Cq values doesn't exceed +/- 0.5 Cq. To randomize the Cq values and minimize human intervention, the values were generated using an online random fraction generator tool (https://onlinerandomtool.com/generate-random-fractions). The *GOI 1* was then normalized using ΔΔCt method [45] to the selected reference genes for both cultures. Significance of fold change was estimated using $P < 0.05$ (indicating significant change) from one-way *ANOVA* and respective post-hoc tests (with Bonferroni *P* value correction).

For culture A1, as suggested by different algorithms used in this study, *GAPDH – CCSER2* pair did prove its utility when it was used for normalization of the simulated *GOI 1* (Supplementary Figure S1A, see Additional file 1 and Fig. 8A). At the same time in culture A1, *GOI 1* after normalization with two other pairs namely *GAPDH – PCBP1* and *CCSER2 – PCBP1* showed no statistically significant fold changes (*ANOVA* $P > 0.05$) as seen in Fig. 8A (also see Supplementary Figure S1A, Additional file 1). In culture A2, the most stable pair suggested was *GAPDH-RNA28S* which when used for normalization of *GOI 1* showed statistically significant fold changes (*ANOVA* $P < 0.05$) for passage 25/28 and passage 25/30. Further, all other reference genes pairs in culture A2 showed statistically significant fold changes when used for normalization (Supplementary Figure S1B, see Additional file 1 and Fig. 8B). Normalization was then performed in pairs of 3 reference genes to investigate whether there is any pair that can be used for normalization as shown in Supplementary table S5 (see Additional file 1) and Fig. 8C. In culture A1, all the different combinations of reference genes possible showed successful normalization while in culture A2, only *GAPDH-PCBP1-CCSER2* triplet achieved successful normalization (Fig. 8D and Supplementary table S5, Additional file 1).

On similar lines *GOI 2 (simulated to have unstable gene expression)* was simulated and assigned random Cq values (Supplementary Table S4, see Additional file 1). This time it was considered that the difference between some Cq values should exceed +/- 0.5 Cq. The *GOI 2* was also normalized using the ΔΔCt

method [45] to the selected reference genes for both cultures. For culture A1, *GAPDH – CCSER2* pair produced statistically significant results (*ANOVA P* = 0.035) in fold change after normalization of *GOI 2*, indicating the pair may not after all be able to handle large differences in Cq values (greater than +/ 0.5 Cq) as seen in Fig. 8A and Supplementary Figure S2A (see Additional file 1). Interestingly, all the other reference gene pairs proved their utility and produced insignificant changes after normalization of *GOI 2* (*ANOVA P* > 0.05) in culture A1 (Fig. 8A and Supplementary Figure S2A, see Additional file 1). Here as well, for culture A2 *GAPDH-RNA28S* pair produced statistically significant fold changes (*ANOVA P* = 0.07), indicating it to not be a suitable pair for normalization. All pairs in culture A2 showed only statistically significant changes in one passage p25/p29 as seen in Fig. 8B (also see Supplementary Figure S2B, Additional file 1), and hence further normalization with 3 reference genes was done to look for successful normalization.

After performing further normalization of *GOI 2* for cultures A1 and A2 in pairs of three reference genes (Supplementary Table S6, see Additional file 1), our analysis showed no suitable pair of reference genes for *GOI 2* in culture A2 (Fig. 8D) but for culture A1, *GAPDH-PCBP1-CCSER2* was the only triplet to yield successful normalization (Fig. 8C).

2.11. Validation of Selected Reference Genes with Normalization of qPCR Data

To support and continue perusing the selected reference gene pairs, 2 genes of interest namely *AURKA* (Aurora Kinase A) and *KRT19* (Keratin 19) were also used. The qPCR was done, and the dataset was normalized using the reference gene pairs as shown in Supplementary Figure S3-S4 (see Additional file 1). Statistical significance and methodology were the same as in *Sect. 2.10*. The Cq range for *AURKA* was from 23.36 (min) to 24.51 (max) with mean Cq of 23.92 ± 0.39 for culture A1 while for culture A2, the Cq range was from 23.19 (min) to 24.32 (max) with mean Cq of 23.74 ± 0.31. The Cq range for *KRT19* was from 18.90 (min) to 20.13 (max) with mean Cq of 19.29 ± 0.29 for culture A1 while for culture A2, the Cq range was from 18.65 (min) to 21.19 (max) with mean Cq of 19.96 ± 0.77.

Upon normalization of *AURKA*, none of the gene pairs in culture A1 were able to normalize *AURKA* adequately (*P* < 0.05) as seen in Fig. 8A and Supplementary Figure S3A (see Additional file 1). In culture A2, however, four reference gene pairs were able to normalize the *AURKA* (Fig. 8B; Supplementary Figure S3B, see Additional File 1). For *KRT19*, in both cultures A1 and A2, no gene pair could yield successful normalization (Fig. 8A and 8B; Supplementary Figure S4A and S4B, see Additional file 1). Again, further investigations were done using genes in triplets.

For *AURKA*, in culture A1, *GAPDH-PCBP1-CCSER2* and *GAPDH-CCSER2-RNA28S* yielded successful normalization (Fig. 8C and Supplementary table S7, see Additional file 1). In culture A2, *GAPDH-PCBP1-CCSER2* and *PCBP1-CCSER2-RNA28S* pairs showed adequate normalization of AURKA (Fig. 8D and Supplementary table S7, see Additional file 1). For *KRT19*, in both cultures A1 and A2, only *GAPDH-PCBP1-CCSER2* triplet showed successful normalization (Fig. 8C and 8D; Supplementary table S8, see Additional file 1).

2.12. Combined Analysis of Dataset from MCF-7 Sub-clones A1 and A2: Analysis of MCF-7 Cell Line as Whole

From the extensive analysis provided above, the biological replicate cultures of MCF-7 (sub-clones) does not necessarily depict similar phenotypic characteristics and gene expression and hence, the reproducibility of the results among sub-clones may not be 100% efficient. Hence, to further investigate the reference gene/s which show least variance overall in the MCF-7 cell line, the dataset from both sub-clones A1 and A2 were combined and the analysis was done. CV% was determined to evaluate the stability of reference genes (Table 5). As before, comprehensive analysis was done for the combined dataset using NormFinder, geNorm, BestKeeper, Comparative ΔCt and RefFinder. The gene rankings were determined using each of these algorithms as summarized in Table 5.

Table 5
Comprehensive analysis of candidate reference gene ranking with dataset from both culture A1 and A2

CV%			NormFinder			geNorm		
Candidate Gene	CV%	Rank	Candidate Gene	Group S.D	Rank	Candidate Gene	Stability value M	Rank
PCBP1	22.13	1	*GAPDH*	0.15	1	*CCSER2*	0.23	1
ACTB	22.28	2	*PCBP1*	0.16	2	*PCBP1*	0.23	1
RNA28S	22.74	3	*RNA28S*	0.21	3	*GAPDH*	0.25	2
PUM1	22.99	4	*CCSER2*	0.21	3	*ACTB*	0.26	3
CCSER2	23.17	5	*ACTB*	0.21	3	*HSPCB*	0.28	4
GAPDH	23.62	6	*HSPCB*	0.24	4	*RNA28S*	0.29	5
SF3A1	26.49	7	*PGK1*	0.27	5	*PGK1*	0.30	6
RNA18S	27.97	8	*HNRNPL*	0.28	6	*HNRNPL*	0.31	7
RPL13A	28.11	9	*PUM1*	0.29	7	*SF3A1*	0.32	8
HSPCB	28.65	10	*RPL13A*	0.29	7	*PUM1*	0.33	9
HNRNPL	35.10	11	*RNA18S*	0.30	8	*RPL13A*	0.34	10
PGK1	35.61	12	*SF3A1*	0.30	8	*RNA18S*	0.35	11
BestKeeper			**Comparative ΔCt**			**RefFinder**		
Candidate Gene	S.D	Rank	Candidate Gene	Average S.D	Rank	Candidate Gene	Geomean	Rank
RNA28S	0.25	1	*GAPDH*	0.30	1	*GAPDH*	1.97	1
PUM1	0.26	2	*PCBP1*	0.30	1	*PCBP1*	2.00	2
ACTB	0.26	3	*CCSER2*	0.33	2	*CCSER2*	2.91	3
PCBP1	0.27	4	*RNA28S*	0.33	2	*RNA28S*	2.91	3
GAPDH	0.27	4	*ACTB*	0.33	2	*ACTB*	4.16	4
CCSER2	0.28	5	*HSPCB*	0.34	3	*PUM1*	6.34	5
SF3A1	0.32	6	*PGK1*	0.36	4	*HSPCB*	6.51	6
RPL13A	0.33	7	*HNRNPL*	0.37	5	*PGK1*	7.84	7
RNA18S	0.34	8	*PUM1*	0.38	6	*HNRNPL*	8.85	8
HSPCB	0.35	9	*SF3A1*	0.38	6	*SF3A1*	9.32	9
PGK1	0.41	10	*RPL13A*	0.38	6	*RPL13A*	9.92	10

CV%			NormFinder			geNorm		
HNRNPL	0.43	11	*RNA18S*	0.38	6	*RNA18S*	10.93	11

As reported in Table 5, CV% analysis revealed that the least variable gene is *PCBP1* (CV% = 22.13%) closely followed by *ACTB* and *RNA28S* (CV% = 22.28% and 22.74% respectively). NormFinder, Comparative ΔCt and RefFinder had almost identical rankings where *GAPDH* and *PCBP1* were constantly reported the most stable genes. Further, in Comparative ΔCt and RefFinder, both *CCSER2* and *RNA28S* came in the next most stable genes. In contrast, NormFinder ranked *RNA28S* before *CCSER2*, however the pair was still reported in top four genes. geNorm on the other hand ranked *CCSER2 – PCBP1* duo as the most stable genes (M value = 0.23). It further ranked *GAPDH* as the third most stable gene. geNorm was further used to calculate Vn/Vn + 1 to estimate the number of reference genes required for normalization of GOIs. geNorm reported that two reference genes should be used (V2/3 = 0.005 is less than the cutoff of 0.15). However, as discussed above we would recommend adding a third reference gene as well. BestKeeper, however ranked *RNA28S* and *PUM1* as the two most stable genes whilst ranking *PCBP1* as the fourth most stable gene.

Finally, BestKeeper was further used to report the Pearson's correlation (r) among genes. A high positive correlation was reported for *PCBP1-CCSER* (r = 0.757). A moderate positive correlation was reported for *PCBP1-GAPDH* and *CCSER-GAPDH* pairs (r = 0.643 and 0.666 respectively). Whilst *RNA28S* showed moderate positive correlation with *GAPDH* (r = 0.623) and *PCBP1* (r = 0.686), it portrayed a low positive correlation with *CCSER2* (r = 0.496). Two genes that were constantly ranked as the least stable (most variable) were *RNA18S* and *RPL13A* except by CV% and BestKeeper.

2.13. Transcriptomic Analysis of the Reference Genes from TCGA Database

The TCGA (The Cancer Genome Atlas) database was used for transcriptomic validation of the expression of the reference genes analysed in the study. R (via R Studio) was used to retreive the data with the help of *TCGAbiolinks* package (available on Bioconductor). Clinical, Morphological and Expression data of 1215 patients (open access) was retrieved under the project "TCGA-BRCA"(Platform - Illumina HiSeq; file.type - rsem.genes.normlised_results). Based on the PAM50 BRCA subtype classifications, data of only luminal A subtype tumor patients were analysed (others like Normal, luminal B, Undertermined, Basal etc. sub-types were removed). In the dataset, selected reference genes were looked for. Normalised Gene expression data was available for 9 of the 12 reference genes (data not available for *ACTB*, *RNA18S* and *RNA28S*) as presented in Table 6. The "*normalised_count*" is a simple transformation of the "*raw_count*" (available in the file with extension rsem.genes.results). The "*raw_count"* values are divided by the 75th percentile and then multiplied by 1000 to obtain *"normalised_count"* to make the values comparable between experiments.

Table 6
Descriptive analysis of the normalized expression count as obtained from TCGA database (Lum A BRCA)

Candidate Reference Gene	Minimum (RSEM)	Median (RSEM)	Maximum (RSEM)	Difference (Max-Min)
GAPDH	12851.20	41483.10	458078.40	445227.20
RPL13A	3437.64	14,620.59	96084.24	92646.60
PGK1	1770.74	7272.16	50,683.05	48912.31
HSPCB**	13752.21	31398.41	119980.80	106228.59
PUM1	642.36	2374.92	4575.92	3933.56
CCSER2	204.81	1141.41	4501.78	4296.97
HNRNPL	2995.01	5132.91	11451.14	8456.13
PCBP1	3732.21	7157.66	13467.68	9735.47
SF3A1	856.04	2854.47	5524.61	4668.57

** *HSPCB* was retrieved as *HSP90AB1* (HUGO gene nomenclature committee, 2020) from the database. RSEM refers to RNA-Seq by Expectation Maximization and is used by the TCGA RNASeqV2 pipeline for normalization of raw counts.

For better data visualization and understanding gene expression of the reference genes, the *"normalized_count"* were converted to the log scale using log2 (normalized_count + 1) as presented in Supplementary Figure S5. *CCSER2* is the most stable gene (from our selected reference genes) identified in the database for luminal A sub-type Breast cancer. Further, *PCBP1* is also quite stably expressed (Supplementary Figure S5 and Table 6). Interestingly, *GAPDH* which was identified as the most stable gene in both cultures was in contrast identified the least stable gene in the transcriptomic analysis.

Going ahead, to evaluate the TPM (transcripts per million), the *TCGAbiolinks* was used again to retrieve the clinical, morphological and expression data of luminal A sub-type breast cancer patients under the project "TCGA-BRCA" (Platform - Illumina HiSeq; file.type - rsem.genes.results). The *"scaled_estimate"* (estimated frequency of gene/transcripts among the total number of transcripts that were sequenced) values were obtained. TPM is obtained by multiplying *"scaled_estimate"* values by a factor of 1 million (10^6). In case of TPM, the data was available for 10 of the 12 selected candidate reference genes (except for *RNA28S* and *RNA18S*) as shown in Table 7. For better visualization and data analysis, the TPM values were also converted to logarithmic scale using log2(TPM).

Table 7
Mean and CV% of the log2TPM (transcripts per million) data as obtained from TCGA database

Candidate Reference Gene	Mean (log2 TPM)	S.D. (log2 TPM)	CV% (S.D/Mean)
ACTB	12.15	0.52	4.30%
GAPDH	11.78	0.76	6.44%
RPL13A	10.32	0.64	6.18%
PGK1	8.14	0.73	8.93%
HSPCB**	10.18	0.58	5.66%
PUM1	5.27	0.57	10.90%
CCSER2**	3.97	0.69	17.49%
HNRNPL	7.92	0.33	4.22%
PCBP1	8.52	0.33	3.96%
SF3A1	5.48	0.55	10.08%

** *HSPCB* was retrieved as *HSP90AB1* while *CCSER2* was retrieved as *FAM190B* (HUGO gene nomenclature committee, 2020) from the database.

All the selected reference genes from the database showed low variance as all the genes has S.D. (log2 TPM) less than 1. Also, all the genes showed medium to high expression level since mean (log2 TPM) were greater than 5 for all genes (except for *CCSER2* which showed lower expression levels). The log2 (TPM) ranges of all the genes are shown in Supplementary Figure S6.

2.14. Relationship Between Cq Values from RT-qPCR and log2 (TPM) of TCGA RNA-Seq

To facilitate an estimation of the relationship between the Cq values obtained from RT-qPCR for both cultures A1 and A2 and the log2 (TPM) values obtained from TCGA database, a correlation analysis was done as seen in Supplementary Figure S7. Although there exists no formal relationship or formula between Cq and Log2 (TPM), the formulas (y intercept and R^2) presented in Supplementary Figure S7 provides an estimation of the Ct for each reference gene prior to RT-qPCR experiments based on RNA-Seq data of luminal A sub-type breast cancer which may be extended to other luminal A cell lines. Pearson's correlation was also calculated between Cq values from both cultures A1 and A2 (and also combined mentioned as MCF-7) and log2 (TPM) as seen in Table 8. Statistical significance was set at $P < 0.05$. No significant correlation was found in any of the gene in both cultures when Cq values were compared with Log2 (TPM) RNA-Seq values.

Table 8
Pearson's Correlation between Cq values from culture A1 and A2 from RT-qPCR and log2 (TPM)

Candidate Reference Gene	Pearson Correlation r (Culture A1 vs log2 TPM)	Pearson Correlation r (Culture A2 vs log2 TPM)	Pearson Correlation r (MCF-7 vs log2 TPM)
ACTB	0.215 (*P* = 0.253)	− 0.028 (*P* = 0.852)	0.115 (*P* = 0.323)
GAPDH	− 0.316 (*P* = 0.088)	− 0.315 (*P* = 0.837)	0.078 (*P* = 0.504)
RPL13A	− 0.018 (*P* = 0.923)	− 0.132 (*P* = 0.385)	− 0.071 (*P* = 0.272)
PGK1	− 0.308 (*P* = 0.097)	0.002 (*P* = 0.988)	0.028 (*P* = 0.812)
HSPCB	− 0.053 (*P* = 0.780)	− 0.013 (*P* = 0.927)	− 0.004 (*P* = 0.970)
PUM1	0.229 (*P* = 0.221)	0.119 (*P* = 0.432)	0.083 (*P* = 0.478)
CCSER2	− 0.015 (*P* = 0.936)	− 0.049 (*P* = 0.747)	− 0.029 (*P* = 0.798)
HNRNPL	− 0.150 (*P* = 0.428)	0.079 (*P* = 0.601)	0.069 (*P* = 0.554)
PCBP1	0.061 (*P* = 0.745)	− 0.126 (P = 0.407)	− 0.045 (*P* = 0.702)
SF3A1	0.310 (*P* = 0.095)	0.261 (*P* = 0.082)	0.132 (*P* = 0.256)

3. Discussion

The human breast adenocarcinoma cell line, MCF-7 has been a standard model among researchers for about five decades now, serving as a laboratory tool for *in vitro* studies as well as model for investigation of key cancer driven processes that directly impact patient care and treatment plan [46]. Despite publication of extensive evidence [9, 10, 11, 12, 13, 14, 15], the genetic and phenotypic variance in MCF-7 sub-clones and subpopulations is often not accounted for in the laboratory protocols and is hence overlooked. The main reason for this overlook usually stems from assumptions that by using cells obtained from same batch and same cell bank and by standardizing protocols and limiting the number of passages, laboratories can ensure that their sub-clones will "behave" with sufficient stability and reproducibility [16].

Further, laboratories can argue that by adhering strictly to the recommendations from Good Cell Culture Practice (GCCP) [47] and employing SNP/STR cell authentication techniques, one can reproduce their results with MCF-7 sub-clones. However, this may not be necessarily enough in connection with MCF-7 cells. In fact, MCF-7 estrogen disruptor assay failed to get international validation in 2016 by US NICEATM (US National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods). The main reason cited for failure was centered on concerns regarding the inter-laboratory reproducibility of results [48].

Previously as well, the question about accuracy in reproducibility of results with MCF-7 cells has been raised [49], but no substantial proof was reported that dealt with variations in reference gene expression tendencies. The present study addresses this issue and reports a novel evidence that non-tested and vague use of common reference genes among sub-clones for qPCR normalization may lead to inaccurate results and conclusions. A pivotal aspect in any gene expression related study is the selection of the most appropriate and accurate reference gene/s. Given the widespread availability of tens of tools and algorithms including web-based platforms (RefFinder) and stand-alone or R-based algorithms (geNorm, BestKeeper, NormFinder etc.), it becomes difficult for researchers without in-depth mathematical background to choose the appropriate software based on their experimental needs.

Each of these algorithms in the past have been reported to have some advantages and some drawbacks that could render the results based on the specific experimental needs. Considering the limitations and advantages of all the algorithms, Sundaram et al [39] in their study, suggested to use an approach that involves using CV% analysis seen in Sect. 2.2 (calculated from 2-Cq linearized values), 2-ΔCq analysis (with calibrator) using *ANOVA* (or Kruskal Wallis test) seen in Sect. 2.3, to roughly estimate the stable reference genes. Further, they suggest using NormFinder (Sect. 2.4) to analyze the stability of the genes and conclude the best pair/s of reference genes (Sect. 2.9). Lastly, it was recommended following the selected pair/s further by analyzing the normalized profile of the target gene/s (gene of interest) to validate the selection (Sect. 2.10 and 2.11).

The Cq (quantification cycle) values (previously referred to as Ct – threshold cycle or Cp – crossing point) represent the number of cycles that are required for the fluorescent signal to cross the threshold. The raw values obtained from and used by the PCR machine aren't necessarily suggestive of the absolute quantity of template in terms of intergenic/inter-run comparisons, but in a simplified way, reflects the abundance of the individual template mRNA in wells [50]. As Livak et al., described, the raw Cq values are determined from a log linear plot of PCR signal vs the cycle number, thereby making Cq an exponential term rather than a linear term [45]. Ergo, for any gene expression analysis or comparisons it is necessary to convert the Cq values to a linear scale using the 2^{-Cq} method [45].

It is interesting to point that on an initial analysis of Table 1, one could argue that the mean Cq values of all genes in both cultures are very comparable to each other and hence the gene expression is stable for all genes in both cultures. However, equal raw Cq values doesn't necessarily imply that the gene expression is also uniform, it merely implies presence of nearly equal template mRNA in the wells. A 2x-3x fold significant difference between gene expression may still exist once the raw Cq values are converted to a linear scale. As shown by CV% analysis (Table 2), the variance between gene expression in both cultures is evidently present, that could otherwise have been overlooked in case raw Cq values were only considered. Further, the $2^{-\Delta Cq}$ analysis (Figs. 1 and 2) also shows significant fold change for many genes over multiple passages (Supplementary table S1, see Additional file 1). The raw Cq values can, however, give information regarding the intragroup and intergroup variation, as the gene with low variations both inter- and intra-group would have less Cq dispersion [51]. Also, the use of Cq values can be done on semi-

quantitative basis whereby a comparison simply indicates that there were fewer/more copies of gene A than gene B in a well.

Our analysis revealed that among the two biological replicate cultures (sub-clones), there are stark differences in the expression patterns and tendencies of the endogenous reference genes. *GAPDH-CCSER2* were identified as potential genes for culture A1, while *GAPDH-RNA28S* were identified for culture A2 using various algorithms. However, when they were employed for cross-normalization of genes of interests in both cultures, both gene pairs were unable to provide adequate results (Fig. 8). Addition of a third gene *PCBP1* to *GAPDH-CCSER2* pair helped to yield successful normalization for all 4 genes of interest and in both cultures A1 and A2 (except for *GOI 2* in Culture A2; this can be attributed most likely to limitation of *GOI 2* as being a simulated gene).

AURKA (Aurora Kinase A) has been known to be associated with playing a key role in centrosome duplication and chromosome segregation during mitosis [52]. Further, it has been reported at many instances to be amplified/mutated in several human cancers [53, 54, 55, 56, 57, 58]. In breast cancer, mixed evidence has been reported with some reporting its association with basal like phenotype [59, 60] whilst others suggesting it to be a marker for progression and outcome of luminal like subtype [61]. Keeping this in mind *AURKA* made a perfect candidate as gene of interest. Although when reference genes were taken in pairs of two, only *GAPDH-CCSER2* normalized *AURKA* expression in culture A2, it was adequately normalized by *GAPDH-CCSER2-PCBP1* triplet in both the cultures A1 and A2.

The other gene of interest selected was *KRT19* (Keratin 19). It belongs to the *KRT* (keratin) family which serves as important markers in RT-qPCR mediated detection of tumors in lymph nodes, peripheral blood and bone marrows of breast cancer patients [62]. *KRT19* is one of the smallest intermediate filament KRT protein [63] and has been shown to regulate breast cancer properties [64]. Using Oncomine database and RT-PCR, Saha et al. [65] evaluated expression of *KRT19* in MCF-7 and other breast cancer cell lines. They reported that *KRT19* was significantly overexpressed in MCF-7, MDA-MB-231 and SKBR3, ergo validating its choice in the present study. None of the reference gene pairs in two cultures could normalize *KRT19* when they were employed in groups of 2. However, in both the cultures, only *GAPDH-PCBP1-CCSER2* yielded successful normalization thereby proving the ability of the triplet pair to handle genes of interest.

The TCGA (The Cancer Genome Atlas Program) represents a joint venture of National Cancer Institute (NCI) and National Human Genome Research Institute (NGGRI) which began in 2006 as a pilot project with three cancer types (lung, ovarian and glioblastoma) which got expanded to present 33 tumor types encompassing a comprehensive dataset describing the molecular changes that occur in cancer [66]. Most samples in TCGA were originally aligned against the Genome Reference Consortium build GRCh37 (hg19) or the “legacy” dataset [67]. However, with advances in technology and drop in sequencing costs GDC (Genomic Data Commons – conceived by NCI) undertook harmonization effort to align the data to GRCh38 (hg38) build (“harmonized” dataset).

The workflow for generating RNA-Seq data in both legacy and harmonized dataset differs substantially [67] leading to introduction of bias between the hg19 and hg38 abundance estimates. However, Gao et

al., [67] demonstrated that there exists excellent concordance between the two workflows in relation to BRCA PAM50 subtypes. Further, they reported that relative change between conditions is preserved across all subtypes of BRCA PAM50. Hence, in the present study, the legacy dataset was accessed and analyzed for comparing the expression profiles of the reference genes.

Two different approaches were employed to analyze legacy dataset compare the gene expression, namely normalized_count (file extension rsem.genes.normalized_results) and TPM obtained from scaled_estimate (file extension rsem.genes.results). Both approaches revealed *CCSER2* to be the most stably expressed. The CV% was estimated at 17.49% (Table 7) from TCGA analysis which was between the range of 15.55% (culture A1) and 25.65% (culture A2) obtained from RT-qPCR (Table 2) in present study. *PCBP1* was also expressed quite stably with CV% estimated at 3.96% from TCGA analysis (Table 7). The RT-qPCR range for *PCBP1* obtained was from 14.89% (culture A1) and 24.90% (culture A2). Finally, CV% for the third gene in the triplet (*GAPDH-PCBP1-CCSER2*), TCGA estimates for *GAPDH* were at 6.44% while RT-qPCR range was from 12.35% (culture A1) and 26.84% (culture A2) (Table 2).

Whilst TCGA analysis of *CCSER2* and *PCBP1* supported our findings in the present study and bolsters their selection, *GAPDH* was revealed to unstably expressed, in contrast to our findings. It is important to point out that the difference in results could be attributed to various underlying limitations. Firstly, TCGA requires all malignancies in its database to be primary, untreated tumors (except cutaneous melanoma) [68]. Further, all specimens deposited were garnered from available frozen materials from different institutes thereby introducing bias in institutional biorepository collections, stemming from institutional research interests, operative protocols, or patient populations [68]. In addition, metastatic diseases or aggressive primary tumors are usually subject to neoadjuvant therapies, which make their inclusion in TCGA database difficult because of limited availability of untreated specimens [68]. However, all these limitations don't contradict the fact that TCGA remains by far the richest source of clinical and research importance especially in developing an integrated picture of commonalities, differences and emergent themes across tumors.

Despite differences in TCGA analysis and our results, use of *GAPDH* as reference gene in qPCR normalization presents a controversy of its own. *GAPDH* has been shown to have increased expression in cancers from other body regions specially from cervix, prostate, pancreas, and lungs [69, 70, 71, 72]. Furthermore, it has been reported that *GAPDH* is overexpressed in MCF-7 cells treated with estradiol [73]. Hence, many studies suggest not to use *GAPDH* as a control gene to study breast cancer or ranks *GAPDH* as the least stable gene [25, 32, 33, 73, 74, 75]. As reported by Liu et al. [25], almost half of the publications in PubMed database used *GAPDH* as a single reference gene for normalization of gene expression analyses with RT-qPCR. Even with the contradiction regarding consideration of *GAPDH* as a suitable or non-suitable candidate, attention should be paid to its selection based on the experimental conditions and study design [33].

CCSER2 is described as a "novel housekeeping gene" (nHKG) by Tilli et al. [27], who provided evidence as to its use as a reference gene in breast cancer studies. They demonstrated that expression of *CCSER2* is

expected to be like the other nHKGs they identified that can increase the assessment consistency of normalization. Indeed, *CCSER2* was ranked highly in culture A1 in our analysis as well and was confirmed by transcriptomic validation to be least variable in expression. It can be hence, used for future analysis and experiments. In addition, *PCBP1*, the most trustworthy gene in our analysis with a stable ranking across all platforms shows promising candidature as an endogenous reference gene as reported also by Jo et al [28].

Another gene that came to light in the analysis for culture A2 was *RNA28S*. The gene has not been reported much in the context for MCF-7 cell line and was added as an in-house suggestion in the present study. However, drawbacks to the use of *RNA18S* or *RNA28S* as reference genes have been reported such as absence of purified mRNA samples and their high abundance compared to target mRNA transcripts making it difficult to accurately subtract the baseline value in RT-PCR data analysis [20]. Further, the use of these genes as control genes is not suggested due to the imbalance between mRNA and rRNA fractions in these molecules [76]. In addition, it has also been shown that certain drugs and biological factors may also affect the rRNA transcription [77, 78].

ACTB, another widely used reference gene, has been previously also verified as a candidate stable reference gene for breast cancer tissue and normal tissues [32, 79]. *ACTB* and *HSPCB* were the best reference genes identified by Liu et al. [25] for ER + breast cancer cell lines including MCF-7. They further reported that *RNA18S* and *ACTB* was the best pair of genes across all breast cancer cell lines [25]. Despite the widespread acceptance of *ACTB* for normalization among a set of human breast cancer cell lines of increasing metastatic potential, limitations have been reported as well [29]. Jacob et al. [33] identified *HSPCB* as one of the suitable genes across a variety of cancer cell lines including MCF-7. Our analysis revealed contrasting results. *ACTB, HSPCB, RPL13A* and *RNA18S* were not ranked in top three in both cultures. Further, *ACTB* and *HSPCB* reported high CV% and showed significant fold changes ($2^{-\Delta Cq}$ analysis). The difference in our results from the previously reported studies is suggestive of the fact that inter-laboratory replication of results with MCF-7 cell line is not very accurate.

Nonetheless, the authors caution the readers that the results obtained in the present study must be seen in the light of some methodological limitations. An interesting observation was made regarding passage p28 from culture A2 (Figs. 1 and 2), where low Cq values were obtained for all housekeeping genes when compared to other passages in the same culture. A plausible explanation for this observation cannot be provided at this point since Good cell culturing practices were strictly adhered to and the cell culturing, RNA isolation, quality control using PCR and RT-PCR were all done by the same single operator to minimize technical errors. Further, the RNA isolation for all passages (and all lysates) for both the cultures was done in one batch together on the same day and the same is also true for RT-PCR, thereby effectively minimizing any chance of human error.

The study with its results aims to provide the readers and researchers an evidence-based recommendation of the most suitable reference genes in routinely cultured MCF-7 cell line with extensive research and investigations. Further, we report the possibility of existence of a heterogenous differential

behavior of the endogenous reference genes within the sub-clones of the MCF-7 cell lines. The authors suggest that more detailed and diverse studies should be undertaken to explore more about the differential expression of the endogenous reference genes. Future studies could be aimed at investigating the reference gene expression amongst other sub-populations of MCF-7 and amongst other breast cancer cell lines.

Lastly, given the widespread use of cancer cell lines not only for basic research but also for drug development and regulatory decision making, ensuring that sub-clones among the cell line are adequately standardized in their expression behavior tends to represent a challenging path forward [16]. A strategy outlined by Tilli et al. [27], whereby experiments are normalized with a panel of reference genes whose expression has been proven to be as minimally variable as possible and as robust as possible under varying conditions gains our support as well. Although geNorm recommended to use two reference genes, we would recommend that three reference genes may be employed to better overcome and handle any reference gene expression variability in the samples that may be present in the sub-clones. Triplet of *GAPDH-PCBP1-CCSER2* should provide a potential alternative to traditionally used reference genes for reference gene matrix in MCF-7 cells.

4. Conclusions

Genetic and phenotypic heterogeneity showcased by MCF-7 cell line poses conundrum with some unanswered questions. Performing reference gene analysis may not be feasible for every sub-clone and hence, based on our results, we suggest using *GAPDH-PCBP1-CCSER2* triplet on cultures that have been cultured in conditions that mimics that of our study, whilst seeing the results in the light of the limitations reported in the present study. Going forward, we encourage that studies should be undertaken especially those that can validate MCF-7 behavior in different growth conditions. We are optimistic that MCF-7 cell line will continue to contribute eminently towards improved and novel treatment modalities for breast cancer patients.

5. Materials And Methods

5.1. Culture and Seeding Conditions of MCF-7 Biological Replicate Cultures (Sub-clones)

Samples were collected from MCF-7 cell line (ATCC, HTB-22) that has been used in our laboratory for previous studies. Samples from different passages were cryopreserved during culturing at different time points. Frozen stocks were taken from MCF-7 cells grown in culturing conditions as described below. For this study, two aliquots from samples that were previously cryopreserved were thawed at the same time. Culture A1 was cryopreserved after passage 27 (p27) and culture A2 was cryopreserved after passage 24 (p24). They were then cultured in the same conditions simultaneously over multiple successive passages. From each culture, three samples were collected from each passage. Culture A1 was cultured from passage 28 (p28) till passage 32 (p32) while culture A2 was cultured from passage 25 (p25) till passage

30 (p30). Cells were cultured in DMEM/F12 (1:1) with 10% FBS (fetal bovine serum), supplemented with 1% penicillin/streptomycin (Thermo Fisher Scientific) in 37 ^{o}C, 5% CO_2, and the growth medium was replaced every 2–3 days. Cell passaging was performed using 1x TrypLE solution (Thermo Fisher Scientific). Cells were grown to 80–100% confluence in T-25cm^2 flasks. Cell count and viability was estimated using cell counting chamber (Improved Neubauer Hemocytometer). For further consecutive passages, cells were seeded at a density of 5000 cells/cm^2. Triplicates (3 samples) of 1×10^6 cells each from each passage from both cultures were taken for isolation of total RNA.

5.2. RNA Extraction and cDNA Synthesis

Total RNA was extracted using Trizol reagent (Thermo Fisher Scientific) according to manufacturer`s protocol. The concentration and quality of the RNA was assessed by Nanodrop 2000 with the mean absorption ratios A260/280 and A260/230 checked to ensure RNA purity. RNA integrity was checked using 1.8% agarose gel electrophoresis. The RNA was further examined for DNA contamination by PCR for *ACTB* and *GAPDH*. The PCR reaction was performed in the presence of both positive and negative controls. No amplified PCR product was found on the agarose gel after PCR and electrophoresis of the RNA samples (except for positive controls). The cDNA synthesis reaction was carried out using the High Capacity cDNA Reverse transcription kit (Thermo Fisher Scientific) in accordance with the manufacturer`s protocol and guidelines and was stored at -20 ^{o}C until further analysis.

5.3. Selection of Candidate Reference Gene and Primer Design

In total, 12 candidate reference genes were selected by perusing relevant literature related to breast cancer and/or MCF-7 cell line and literature that reported transcriptomic data based on TCGA (The Cancer Genome Atlas) database [25, 26, 27, 28, 33, 34, 35, 36, 37]. All selected genes are shown in Table 9. *GAPDH* and *ACTB* are two of the most used single control genes reported in more than 90% of the cases in high impact journals [80]. Further these two genes are commonly included in the commercially available cancer pathway kits like those from Qiagen, Life technologies and are also included in Oncotype DX test arrays. Life technologies kit also includes genes like *RNA18S* and *PGK1* along with several others [27]. *RNA18S-ACTB* has also been identified as an appropriate gene pair across all breast cancer cell lines [25]. *RPL13A* was described by De Jonge et al., as a novel candidate reference gene with enhanced stability among a magnitude of different cell types across varying experimental conditions [37]. *HSPCB* has been reported previously to be one of the most stable reference gene for ER + breast cancer cells [25, 33], thereby influencing its selection in the present study.

PUM 1 and *CCSER2* were identified by Tilli et al., using transcriptomic analysis across breast cancer cell lines as less variable and more accurate for research in breast cancer cell lines and tissue samples in comparison to the traditional housekeeping genes [27]. Previously, *HNRNPL* and *PCBP1* have been reported to be among the most highly expressed genes across breast cancer samples and performed better than both *ACTB* and *GAPDH* as supported by TCGA transcriptomic validation reported by Jo et al. [28]. *ATCB* and *SF3A1* were identified by using high throughput analysis of micro-assay datasets with

subsequent validation by RT-qPCR as reported by Maltseva et al. [36]. They further reported these genes as more efficient for analysis of breast cancer samples when compared with the reference gene panel provided by Oncotype DX assay. Finally. *RNA28S* was added to the selected genes as an experimental in-house suggestion.

The genes selected, as described above, have been time and again reported as the best or have been described with controversial findings with studies spanning over almost two decades from 2000 to 2019 and hence, they were never collectively investigated in MCF-7 cell line. The present study, hence, took these findings from previous studies into account while selecting the genes. The sources of gene primers and primer sequence used for the reference genes are shown in Supplementary Table S9 (see Additional file 1). The present study employed the primers that were reported before in the literature to maintain inter-reliability and inter-connectivity with the previous studies. Primers for *RNA28S* and *PGK1* were designed using Primer3Plus [81]. The melting curves of all the selected gene primers are presented in Supplementary Figures S8 to S19 (see Additional file 1). Finally, the primer sequences (designed using Primer3Plus) for genes of interest, *AURKA* and *KRT19* are also presented in Supplementary Table S9 (see Additional file 1) with their melting curves in Supplementary Figures S20-S21 (see Additional file 1).

Table 9
Description of the selected candidate Reference genes and Genes of Interest for the study

Gene Symbol	Gene Name	Molecular Function	Accession Number	Chromosomal Localisation
ACTB	β – Actin	Cytoskeleton (Contractile apparatus)	NM_001101	7p22 – p12
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase	Glycolytic enzyme	NM_002046	12p13.31
RPL13A	Ribosomal protein L13a	Ribosome subunit, translation	NM_012423	19q13.33
PGK1	Phosphoglycerate kinase 1	Glycolytic enzyme	NM_000291	Xp21.1
HSPCB	Heat Shock protein 90 kDa beta	Signal transduction, Protein folding	NM_007355	6p21.1
RNA28S	28S ribosomal RNA	Ribosome subunit, translation	NR_003287	Unknown
RNA18S	18S ribosomal RNA	Ribosome subunit, translation	NR_003286	Unknown
PUM1	Pumilio RNA binding family member 1	RNA binding protein encoding	NM_001020658	1p35.2
CCSER2	Coiled Coil Serine Rich protein 2	Microtubule binding protein encoding	NM_018999	10q23.1
HNRNPL	Heterogenous Nuclear Ribonucleoprotein L	Formation, processing & packaging of mRNA	NM_001005335	19q13.2
PCBP1	Poly (rC) Binding Protein 1	RNA binding protein encoding	NM_006196	2p13.3
SF3A1	Splicing Factor 3a Subunit 1	Spliceosome assembly & pre-mRNA splicing	NM_005877	22q12.2
AURKA*	Aurora Kinase A	Mitotic centrosomal protein kinase (controls chromosome segregation during mitosis)	NM_003600	20q13.2
KRT19*	Keratin 19	Structural molecule and constituent of cytoskeleton	NM_002276	17q21.2

*Genes that were used as gene of interest for normalization by candidate reference genes.

5.4. Real Time Quantitative PCR (RT-qPCR)

10 ng of cDNA per reaction was used for real time quantitative PCR (qPCR) using ViiA 7 RT-PCR thermocycler (Thermo Fisher Scientific). Triplicate reactions of each sample were performed using HOT FIREPol EvaGreen qPCR Supermix (Solis Biodyne) on 384 well plates (Thermo Fisher Scientific). The cycling parameters were: 95 ^{o}C for 4 min followed by 40 cycles of amplification at 95 ^{o}C for 30 sec and 58 ^{o}C for 20 sec followed by 72 ^{o}C for 30 sec with melting curve. All assays were performed with a non-template control (NTC).

5.5. Determination of the Most Stable Reference Genes from Candidate Genes

The stability of the reference genes was analyzed using CV%, as well as $2^{-\Delta Cq}$ by linearization of the raw Cq values obtained from qPCR. Further, various available algorithms like NormFinder [41], geNorm [20], BestKeeper [42], Comparative Delta Ct [43] and RefFinder [44] were used. All algorithms use different approaches to analyze the stability of gene expression as described in the respective result sections (Sect. 2.4 to 2.8). Lastly, an integrated approach was used to determine the most stable reference genes among both the biological replicate cultures. For the calculations of expression levels, it was assumed that the qPCR reaction efficiencies were equal for all reactions.

5.6. Normalisation of RT-PCR Dataset, Statistical Analysis and TCGA Transcriptomic Bioinformatics

After the selection of the most stable reference genes, two simulated RT-PCR datasets were generated (detailed in *Sect. 2.10*) and was normalized first against pair of 2 reference gene and then to the pair of 3 reference genes to analyze the possible outcomes when using the selected reference genes. Statistical significance was set at $P < 0.05$ (unless otherwise stated) and was calculated using Analysis of Variance (one-way ordinary *ANOVA* (with relevant multiple comparisons test and corrections) where data was distributed normally. For non-normally distributed data (checked using Shapiro-Wilk test), the non-parametric *ANOVA* (*Kruskal-Wallis ANOVA*) with respective corrections and multiple comparisons was used. The comparisons were made to compare the gene expression at successive passages to the gene expression at the smallest passage number. Data Management and storage along with descriptive statistics were done using MS Excel (Microsoft Office 365). The statistical analysis was done using *JASP* (Jeffery`s Amazing Statistical Program) v.0.10.2.0 and *R Studio* v3.6.3. The R based package < TCGAbiolinks > [82–84] was used to retrieve the TCGA gene expression data. The TCGA legacy archive was accessed for open access patient data. The legacy archive has unmodified copy of data that was previously stored in CGhub and TCGA data portal and uses GRCh37 (hg 19) as references. After downloading and preparing the data (GDCdownload and GDCprepare function), the data was exported to Excel for further analysis.

6. Abbreviations

MCF-7	Michigan Cancer Foundation – 7
RGs	Reference genes
RT-PCR	Real time – polymerase chain reaction
p	Passage
GOI	Gene of Interest
CV%	Coefficient of Variation %
S.D.	Standard Deviation
TPM	Transcripts per Million
BRCA	Breast Invasive Carcinoma
TCGA	The Cancer Genome Atlas
HUGO	Human Genome Organisation
Lum A	Luminal A Molecular Subtype Breast Cancer
ACTB	Actin Beta
GAPDH	Glyceraldehyde-3-phosphate Dehydrogenase
RPL13A	Ribosomal Protein L13a
PGK1	Phosphoglycerate Kinase 1
HSPCB	Heat Shock Protein 90 Alpha family Class B member 1
RNA18S	RNA, 18S ribosomal
RNA28S	RNA, 28S ribosomal
PUM1	Pumilio RNA Binding family member
CCSER2	Coiled-Coil Serine Rich protein 2
HNRNPL	Heterogenous Nuclear Ribonucleoprotien L
PCBP1	Poly(rC) Binding protein 1
SF3A1	Splicing factor 3a subunit 1
AURKA	Aurora Kinase A
KRT19	Keratin 19

7. Declarations

a. **Ethical Approval & Consent to Participate:** Not Applicable

b. **Consent for Publication:** Not Applicable

c. **Availability of Data and Materials:** The datasets generated for Genes of Interest GOI 1 and 2 are available as Supplementary Tables S3 and S4 respectively (see Additional file 1). The primer sequences for all genes studied in the study are available as Supplementary Table S9 (see Additional file 1). TCGA dataset is available for download from TCGAbiolinks package (R studio) or TCGA repository. The RT-qPCR datasets are available from the corresponding author upon reasonable request via email provided.

d. **Competing Interests:** The authors declare no competing interests in the present study. Further, neither funders nor the funding institution had a role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

e. **Funding:** The present study was funded by Riga Stradiņš University (RSU) Project Nb.5-1/127/2019.

f. **Author's Contributions:** IC, NJ and VP conceptualized the study while IC and DN were responsible for methodology. Data and Statistical analysis were done by NJ. Data curation was done by IC and NJ. Validation of the study protocol and results was done by VP, IC and NJ. Visualization was done by NJ while project supervision and funding acquisition was done by IC. Investigations and validation were done by IC, DN and VP. Original draft was prepared by NJ and IC while revisions and final editing were done by VP, DN, IC and NJ. All authors have read and approved the final manuscript for publication.

g. **Acknowledgments:** Not Applicable

h. **Authors' Information:** All Authors (NJ, DN, VP and IC) are affiliated with the Laboratory of Molecular Genetics, Institute of Oncology, Riga Stradiņš University, Dzirciema street 16, Riga, Latvia LV-1007.

8. References

1. Lee AV, Oesterreich S, Nancy E. Davidson. MCF-7 Cells—Changing the Course of Breast Cancer Research and Care for 45 Years. JNCI: Journal of the National Cancer Institute July. 2015;107:Issue 7, djv073.
2. Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: Highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer. Ann Oncol. 2013;24:2206–23.
3. Serban C, Anca MC, Marius R. The story of MCF-7 Breast Cancer Cell Line: 40 years of experience in research. Anticancer Research June. 2015;35(6):3147–54.
4. Huang SB, Chou D, Chang YH, Li KC, Chiu TK, Ventikos Y, Wu MH. Development of pneumatically driven active cover lid for multi-well microplates for use in perfusion three-dimensional cell culture. Sci Rep. 2015. 5:18352. PMID: 26669749.
5. Burdall S, Hanby A, Lansdown MR, Speirs V. Breast Cancer cell lines: friend or foe? Breast cancer Res. 2003;5:89–95.
6. Osborne CK, Hobbs K, Trent JM. Biological differences among MCF-7 human breast cancer cell lines from different laboratories. Breast Cancer Res Treat. 1987;9:111–21.

7. Resnicoff M, Medano EE, Podhajcer OL, Bravo AI, Bover L, Mardoh J Subpopulations of MCF-7 cells separated by Percoll gradient centrifugation: a model to analyze the heterogeneity of human breast cancer. PNAS. Oct 1987. 84; 20: 7295-9. PMC: 299279.
8. Baguley BC, Leung E. In breast cancer carcinogenesis, cell growth and signaling pathways. (ed Gunduz M) (InTech, 2011).
9. Seibert K, Shafie SM, Triche TJ, Whang-Peng JJ, O'Brien SJ, Toney JH, Huff KK, Lippman ME. Clonal variation of MCF-7 breast cancer cells in vitro and in athymic nude mice. Cancer Res. 1983;43:2223–39.
10. Whang-Peng J, Lee EC, Kao-Shan CS, Seibert K, Lippman M. Cytogenetic studies of human breast cancer cell lines: MCF-7 and derived variant sublines. J Natl Cancer Inst. 1983;71:687–95.
11. Butler WB, Berlinski PJ, Hillman RM, Kelsey WH, Toenniges MM. Relation of in vitro properties to tumorigenicity for a series of sublines of the human breast cancer cell line MCF-7. Cancer Res. 1986;46:6339–48.
12. Melanie N, Paul C, Julie V, Beatruce O, Lisa U, Catherine N, Daniel B, Emmanuel JP-D, Pascale C, Charles T. Genetic Variability in MCF-7 sublines: evidence of rapid genomic and RNA expression profile modifications. BMC Cancer. 2003. 3:13. PMID: 12713671.
13. Hiorns LR, Bradshaw TD, Skelton LA, Yu Q, Kelland LR, Jones BL. Variation in RNA expression and genomic DNA content acquired during cell culture. British journal of Cancer.2004. 90: 476–482.
14. Jones C, Payne J, Wells D, Delhanty JDA, Lakhani SR, Kortenkamp A. Comparative genomic hybridization reveals extensive variation among different MCF-7 cell stocks. Cancer Genet Cytogenet. 2000;117(2):153–8.
15. Bahia H, Ashman JNE, Cawkwell L, Lind M, Monson JRT, Drew PJ, Greenman J. Karyotypic variation between independently cultured strains of the cell line MCF-7 identified by multicolor fluorescence *in situ* hybridization. Int J Oncol. 2002;20:489–94.
16. Kleensang A, Vantangoli M, Odwin-DaCosta S, et al. Genetic Variability in a frozen batch of MCF-7 cells invisible in routine authentication affecting cell function. Sci Rep. 2016;6:28894. https://doi.org/10.1038/srep28994.
17. Bustin SA. Absolute quantification of mRNA using real time reverse transcription polymerase chain reaction assays. J Mol Endocrinol. 2000;25:169–93.
18. Ginzinger DG. Gene quantification using real time quantitative PCR: an emerging technology hits the mainstream. Exp Hematol. 2002;30:503–12.
19. Li Li Y, Yan H, Xu T, Qu, Baoxi W. Selection of reference genes for gene expression studies in ultraviolet B-irradiated human skin fibroblasts using quantitative real time PCR. BMC Mol Bio. 2011;p8:12.
20. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, et al. Accurate normalization of real time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002;3:Research0034.

21. Lee KS, Alvarenga TA, Guindalini C, Andersen ML, Tufik S. Validation of commonly used reference genes for sleep related gene expression studies. BMC Mol Bio. 2009;p10:45.
22. Jung M, Ramankulov A, Roigas J, Johannsen M, Ringsdorf M, Kristiansen G, Jung K. In search of suitable reference genes for gene expression studies of human renal carcinoma by real time PCR. BMC Mol Bio. 2007;p8:47.
23. Rho HW, Lee BC, Choi ES, Choi IJ, Lee YS, et al. Identification of valid gene expression studies of human stomach cancer by reverse transcription-qPCR. BMC Cancer. 2010;10:240.
24. Stephen AB, Vladimir B, Jeremy AG, Jan H, Jim H, Michael K, et al. The MIQE Guidelines: Minimum information for publication of Quantitative Real time PCR Experiments. Clin Chem. 2009;55:4.
25. Liu LL, Zhao H, Ma TF, Ge F, Chen CS, Zhang YP. Identification of valid reference genes for the normalization of RT-qPCR expression studies in human breast cancer cell lines treated with and without transient transfection. PLoS One. 2015;10(1):e0117058.
26. Kilic Y, Celebiler C, Sakizli M. Selecting Housekeeping Genes as references for normalization of quantitative PCR data in breast cancer. Published Online May 2013. Clin Transl Oncol. DOI: 10.1007/s12094-013-1058-5.
27. Tilli TM, Castro Cda S, Tuszynski JA, Carels N. A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases. BMC Genomics. 2016 Aug. 15;17(1):639. PMID:27526934.
28. Jo J, Choi S, Oh J, et al. Conventionally used reference genes are not outstanding for normalization of gene expression in human cancer research. BMC Bioinformatics. 2019;20:245.
29. Lyng MB, Laenkholm AV, Pallisgaard N, Ditzel HJ. Identification of genes for normalization of real time RT-PCR data in breast carcinomas. BMC Cancer. 2008;8:20.
30. Morse DL, Carroll D, Weberg L, Borgstrom MC, Ranger-Moore J, et al. Determining suitable internal standards for mRNA quantification of increasing cancer progression in human breast cells by real-time reverse transcriptase polymerase chain reaction. Anal Biochem. 2005;342:69–77.
31. Krasnov GS, Kudryavtseva AV, Snezhkina AV, Lakunina VA, Beniaminov AD, et al. Pan-Cancer analysis of TCGA data revealed promising reference genes for qPCR normalization. Front Genet March. 2019;10:97.
32. Gur-Dedeoglu B, Konu O, Bozkurt B, Ergul G, Seckin S, et al. Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. Oncol Res. 2009;17(8):353–65. PMID:19544972.
33. Jacob F, Guertler R, Naim S, Nixdorf S, Fedier A, Hacker NF, Heinzelmann-Schwarz V. Careful Selection of Reference Genes Is Required for Reliable Performance of RT-qPCR in Human Normal and Cancer Cell Lines. PLoS ONE. 2013;8(3):e59180.
34. Quiroz FG, Posada OM, Perez DG, Castro NH, Sarassa C, et al. Housekeeping gene stability influences the quantification of osteogenic markers during stem cell differentiation to the osteogenic lineage. Cytotechnology. 2010;62(2):109–20.
35. Balwierz A, Czech U, Polus A, Filipkowski RK, et al. Human Adipose tissue stromal vascular fraction cells differentiate depending on distinct types of media. Cell Prolif. 2008;41:441–59.

36. Maltseva DV, Khaustova NA, Fedetov NN, Matveeva EO, Lebedev AE, et al. High Throughput identification of reference genes for research and clinical RT-qPCR analysis of breast cancer samples. J Clin Bioinforma. 2013 July. 22;3(1):13. PMID: 23876162.
37. De Jonge HJM, Fehrmann RSN, De Bont ESJM, Hofstra RMW, Gerbens F, Kamps WA, et al. Evidence Based selection of housekeeping genes. PLoS ONE. 2007;2(9):e898.
38. Boda E, Pini A, Hoxha E, Parolisi R, Tempai F. Selection of reference genes for quantitative real time RT-PCR studies in mouse brain. J Mol Neurosci Humana Press Inc. 2009;37:238–53. PMID:18607772.
39. Venkat KS, Nirmal KS, Charbel M, Julien G. Optimal Use of statistical methods to validate reference gene stability in longitudinal studies. Plos One. 2019;14(7):e0219440.
40. Hellemans J, Mortier G, Paepe AD, Speleman F. Jo Vandesompele. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol. 2007 Feb;8(2):R19.
41. Andersen CL, Jensen JL, Orntoft TF. 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 research. 2004;64:5245–50.
42. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. Biotechnol Lett. 2004;26:509–15.
43. Silver N, Best S, Jiang J, Thein SL. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. BMC Mol Biol. 2006;7:33.
44. Xie F, Xiao P, Chen D, Xu L, Zhang B. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. Plant molecular biology. 2012;80(1):75–84.
45. Livak K-J, Schmittgen T-D. Analysis of Relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods. 2001;25:402–8. DOI:10.1006/meth.2001.1262.
46. Sweeney EE, McDaniel RE, Maximov PY, Fan P, Jordan VC. Models and mechanisms of acquired antihormone resistance in breast cancer: significant clinical progress despite limitations. Horm Mol Biol Clin Investig. 2012;9:143–63.
47. Coecke S, Balls M, Bowe G, Davis J, Gstrauthaler G, Hartung T, Hay R, et al. Guidance on good cell culture practice: a report of the second ECVAM task force on good cell culture practice. Altern Lab Anim. 2005;33(3):261–87.
48. NICEATM draft validation study report MCF
-7 cell proliferation test method
(Last accessed March 2020)
NICEATM draft validation study report MCF-7 cell proliferation test method.
https://ntp.niehs.nih.gov/iccvam/methods/endocrine/mcf7/mcf7-valstudyreport-19jun12-wcv2-draft.pdf. (Last accessed March 2020) (2012).

49. Ochsner SA, Steffen DL, Hilsenbeck SG, Chen ES, Watkins C, Neil. GEMS (Gene expression meta-signatures), a web resource for querying meta-analysis of expression microarray datasets: 17β estradiol in MCF-7 cells. Can Res. 2009;69(1):23–6.
50. Kosuth J, Farkasovska M, Mochnacky F, Daxnerova Z, Sevc J. Selection of Reliable reference genes for analysis of gene expression in spinal cord during rat postnatal development and after injury. Brain Sci December. 2019;10(1):6. Doi:10.3390/brainsci10010006.
51. 10.3390/ijms18051060
Piazza VG, Bartke A, Miquet JG, Sotelo AI. Analysis of Different approaches for the selection of reference genes in RT-qPCR experiments: A case study in skeletal muscle of growing mice. Int J Mol Sci. May 2017. 18(5):1060. Doi: 10.3390/ijms18051060. PMID: 28509880.
52. Wang X, Zhou Y, Qiao W, et al. Overexpression of aurora kinase A in mouse mammary epithelium induces genetic instability preceding mammary tumor formation. Oncogene. 2006;25:7148–58.
53. Miyosi Y, Iwao K, Egawa C, Noguchi S. Association of centrosomal kinase STK15/BTAK mRNA expression with chromosome instability in human breast cancers. Int J Cancer. 2001;92:370–3.
54. Gritsko TM, Domenico P, June EP, Lin Y, Mei S, Sue AS, et al. Activation and overexpression of centrosome kinase BTAK/Aurora-A in human ovarian cancer. Clin Cancer Res. 2003;9:1420–6.
55. Li D, Zhu J, Firozi PF, Abbruzzese JL, Evans DB, Cleary K, Friess H, Sen S. Overexpression of oncogenic STK15/BTAK/Aurora A Kinase in Human pancreatic cancer. Clin Cancer Res March. 2003;9(3):991–7.
56. Bischoff JR, Lee A, Yingfang Z, Kevin M, Lelia N, Brian S, Brian S, et al. A homolog of Drosophila aurora kinase is oncogenic and amplified in human colorectal cancers. EMBO J. 1998;17(11):3052–65.
57. Sen S, Hongyi Z, Ruo-Dan Z, Dong SY, Funda VL, Shigemi I, et al. Amplification/Overexpression of a mitotic kinase gene in human bladder cancer. J Natl Cancer Inst. 2002;94:1320–9.
58. Tong T, Yali Z, Jianping K, Lijia D, Yongmei S, Ming F, Zhihua L, et al. Overexpression of Aurora-A contributes to malignant development of human esophageal squamous cell carcinoma. Clin Cancer Res. 2004;10(21):7304–10.
59. Xu J, Wu X, Zhou WH, Liu AW, Wu JB, Deng JY, et al. Aurora A identifies early recurrence and poor prognosis and promises a potential therapeutic target in triple negative breast cancer. PLoS One. 2013;8(2):e56919.
60. Staff S, Isola J, Jumppanen M, Tanner M. Aurora-A gene is frequently amplified in basal-like breast cancer. Oncol Rep. 2010;23:307–12.
61. Ali HR, Dawson SJ, Blows FM, Provenzano E, Pharoah PD, Caldas C. Aurora Kinase A outperforms Ki67 as a prognostic marker in ER-positive breast cancer. Br J Cancer. 2012;106:1798–806.
62. Ignatiadis M, Xenidis N, Perraki M, Apostolaki S, Politaki E, Kafousi M, et al. Different prognostic value of cytokeratin-19 mrna-positive circulating tumor cells according to estrogen receptor and HER2 status in early stage breast cancer. J Clin Oncol. 2007;25:5194–202.

63. Wu Y-J, Rheinwald JG. A new small (40 kd) keratin filament protein made by some cultured human squamous cell carcinomas. Cell. 1981;25:627–35.
64. Ju J-H, Yang W, Lee K-M, Oh S, Nam K, Shim S, et al. Regulation of cell proliferation and migration by keratin19 induced nuclear import of early growth response-1 in breast cancer cells. Clin Cancer Res. 2013;19:4335–46.
65. Saha SK, Choi HY, Kim BW, Dayem AA, Yang GM, Kim KS, Yin YF, Cho SG. KRT19 directly interacts with β-catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties. Oncogene Jan. 2017;19(3):332–49. 36(.
66. Hutter C, Zenklusen J-C. The Cancer Genome Atlas: Creating Lasting value beyond its Data. Cell April. 2018;173(2):283–5.
67. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, Akbani R, et al. Before and After: Comparison of Legacy and harmonized TCGA Genomic Data Common's data. Cell Systems July. 2019;9(1):24–34. e10.
68. Cooper L-AD, Demicco E-G, Saltz JH, Powell RT, Rao A, Lazar A-J. PanCancer insights from the Cancer Genome Atlas: the pathologist's perspective. J Pathol April. 2018;244(5):512–24.
69. Kim JW, Kim SJ, Han SM, Paik SY, Hur SY, et al. Increased glyceraldehyde-3-phosphate dehydrogenase gene expression in human cervical cancers. Gynceol Oncol. 1998;71:266–9.
70. Rondinelli RHEDE, Tricoli JV. Increased glyceraldehyde-3-phosphate dehydrogenase gene expression in late pathological stage human prostate cancer. Prostate Cancer Prostatic Dis. 1997;1:66–72.
71. Schek NHBL, Finn OJ. Increased glyceraldehyde-3-phosphate dehydrogenase gene expression in human pancreatic adenocarcinoma. Cancer Res. 1988;48:6354–9.
72. Tokunaga K, Nakamura Y, Sakata K, Fujimori K, Ohkubo M, et al. Enhanced expression of a glyceraldehyde-3-phosphate dehydrogenase gene in human lung cancers. Cancer Res. 1987;47:5616–9.
73. Révillion F, Pawlowski V, Hornez L, Peyrat JP. Glyceraldehyde-3-phosphate dehydrogenase gene expression in human breast cancer. Eur J Cancer. 2000;36:1038–42.
74. McNeill RE, Miller N, Kerin MJ. Evaluation and validation of candidate endogenous control genes for real-time quantitative PCR studies of breast cancer. BMC Mol Biol. 2007;8:107.
75. De Kok JB, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, et al. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. Lab Invest. 2005;85:154–9.
76. Solanas M, Moral R, Escrich E. Unsuitability of using ribosomal RNA as loading control for Northern blot analyses related to the imbalance between messenger and ribosomal RNA content in rat mammary tumors. Anal Biochem. 2001;288:99–102.
77. Johnson ML, Redmer DA, Reynolds LP. Quantification of lane to lane loading of poly(A) RNA using a biotinylated oligo(dT) probe and chemi-luminescent detection. Biotechniques. 1995;19:712–5.

78. Spanakis E. Problems related to the interpretation of autoradiograph data on gene expression using common constitutive transcripts as controls. Nuclei Acids Res. 1993;21:3809–19.
79. Majidzadeh -AK, Esmaeili R, Abdoli N. TFRC and ACTB as the best reference genes to quantify Urokinase Plasminogen Activator in breast cancer. BMC Res Notes. 2011;4:215. 10.1186/1756-0500-4-215.
80. Suzuki T, Higgins PJ, Crawford DR. Control selection for RNA quantification. Biotechniques. 2000;29:332–7.
81. Andreas U, Harm N, Xiangyu R, Ton B, Rene G, Jack AML. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 2007;35:W71–4.
82. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data Nucleic Acids Research. May 2016. 44(8): e71.
83. Silva TC, et al. TCGA Workflow: Analyze cancer genomics and epi-genomics data using Bioconductor packages. F1000Research 5 (2016).
84. Mounir M, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. PLoS computational biology. 2019. 15.3: e1006701.
85. **Additional File/Material**.
86. **File Name**: Additional File 1.
87. **File Format**. PDF (Adobe Acrobat) – Additional File 1.pdf.
88. **Title of Data**. Supplementary Tables S1-S9 and Figures S1-S21.
89. **Description of Data**: Supplementary Tables S1-S9 and Figures S1-S21.

Figures

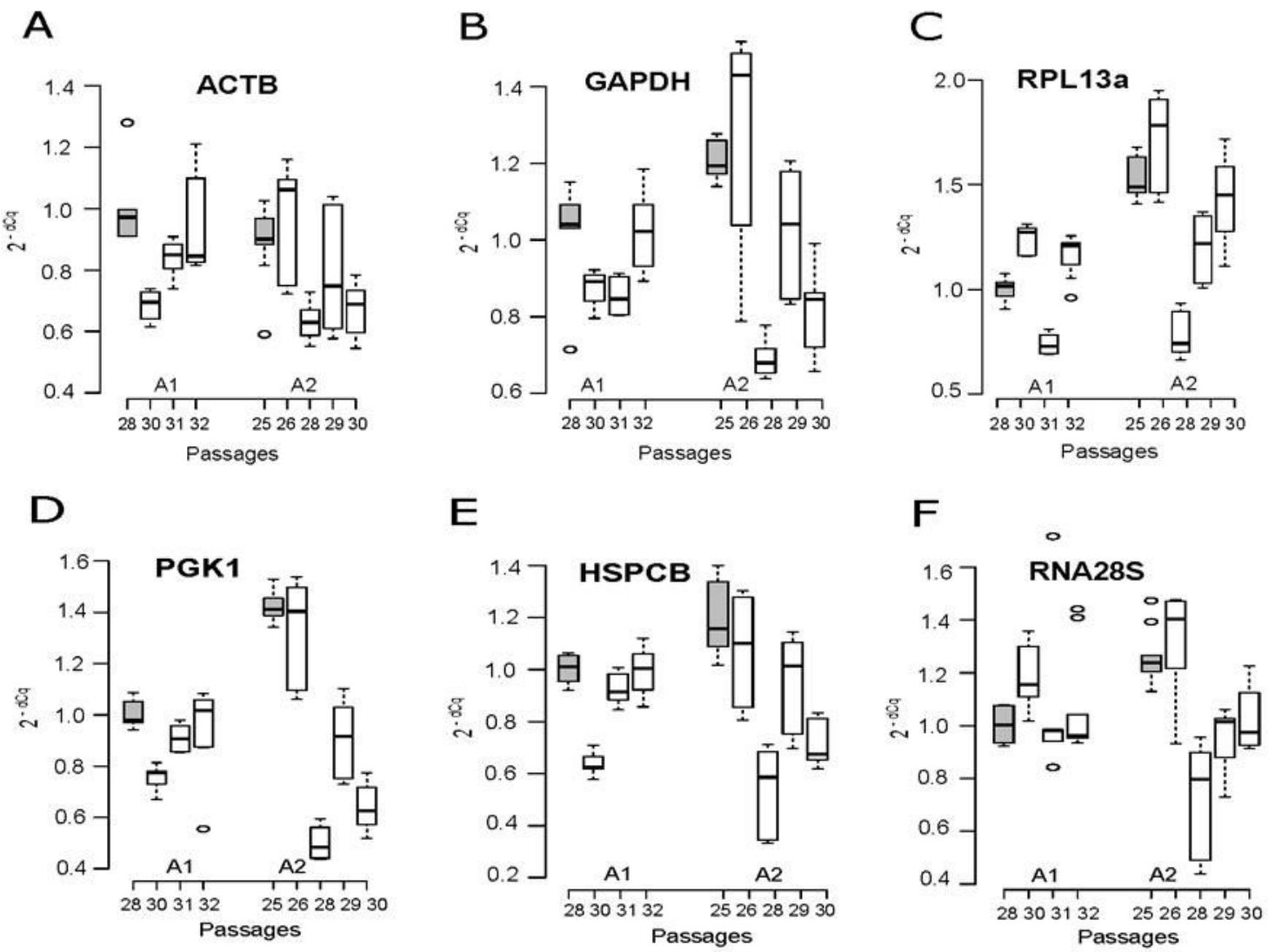

Figure 1

Relative expression change of selected reference genes in both cultures A1 and A2 Legend: Relative change of expression of the candidate reference genes (A) ATCB (B) GAPDH (C) RPL13A (D) PGK1 (E) HSPCB and (F) RNA28S over successive passages in both cultures A1 and A2. The expression change was measured as 2-ΔCq with the experimental calibrators marked as gray boxplots (p28 in culture A1 and p25 in culture A2).

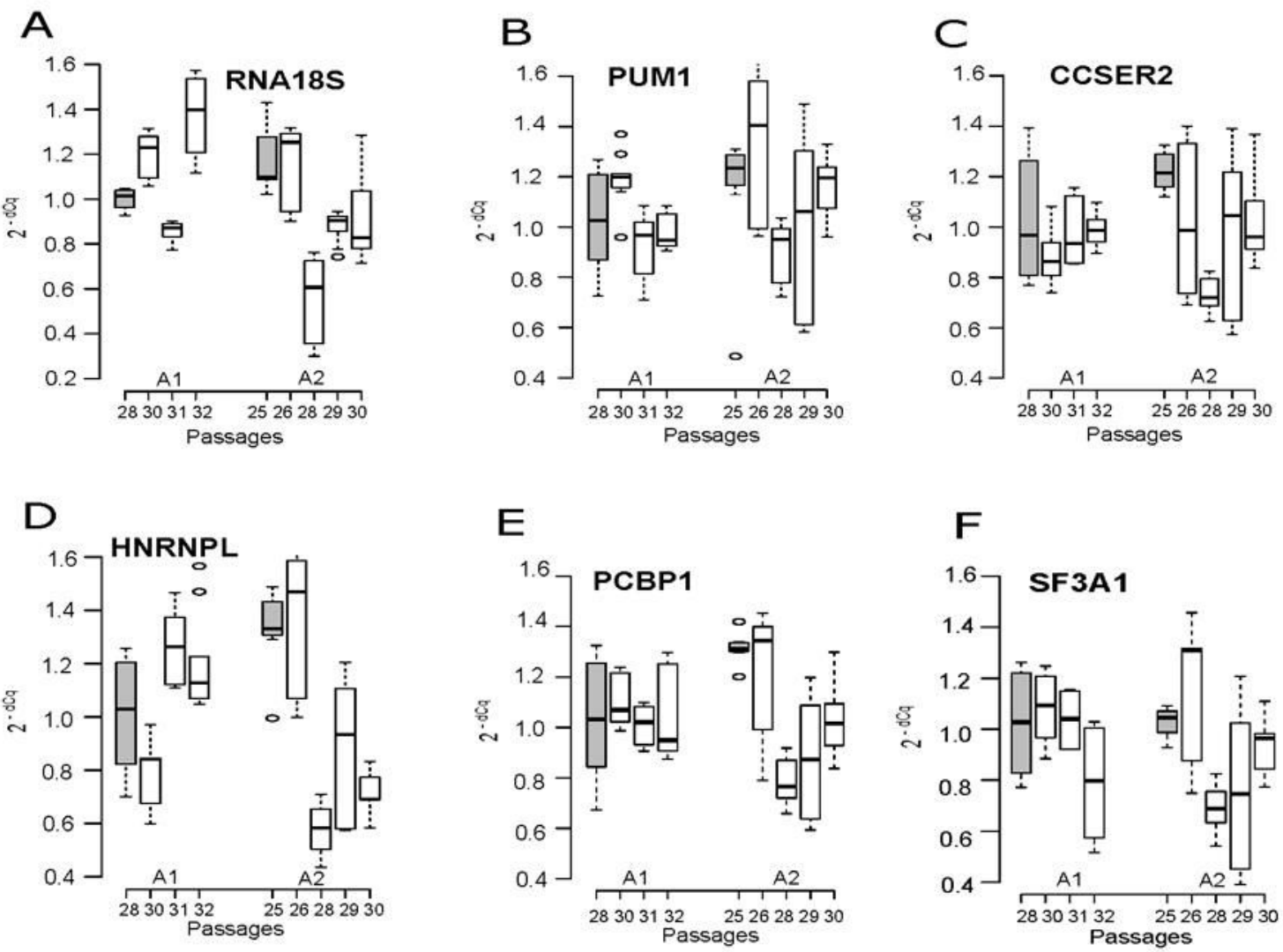

Figure 2

Relative expression change of selected reference genes in both cultures A1 and A2 Legend: Relative change of expression of the candidate reference genes (A) RNA18S (B) PUM1 (C) CCSER2 (D) HNRNPL (E) PCBP1 and (F) SF3A1 over successive passages in both cultures A1 and A2. The expression change was measured as 2-ΔCq with the experimental calibrators marked as gray boxplots (p28 in culture A1 and p25 in culture A2).

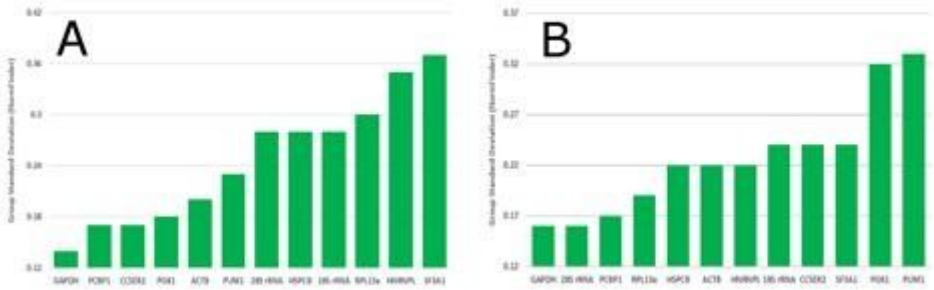

Figure 3

NormFinder analysis of reference genes in both cultures A1 and A2 Legend: NormFinder analysis for both cultures (A) A1 and (B) A2 shown as group standard deviations (S.D). 28S rRNA refers to RNA28S while 18S rRNA refers to RNA18S.

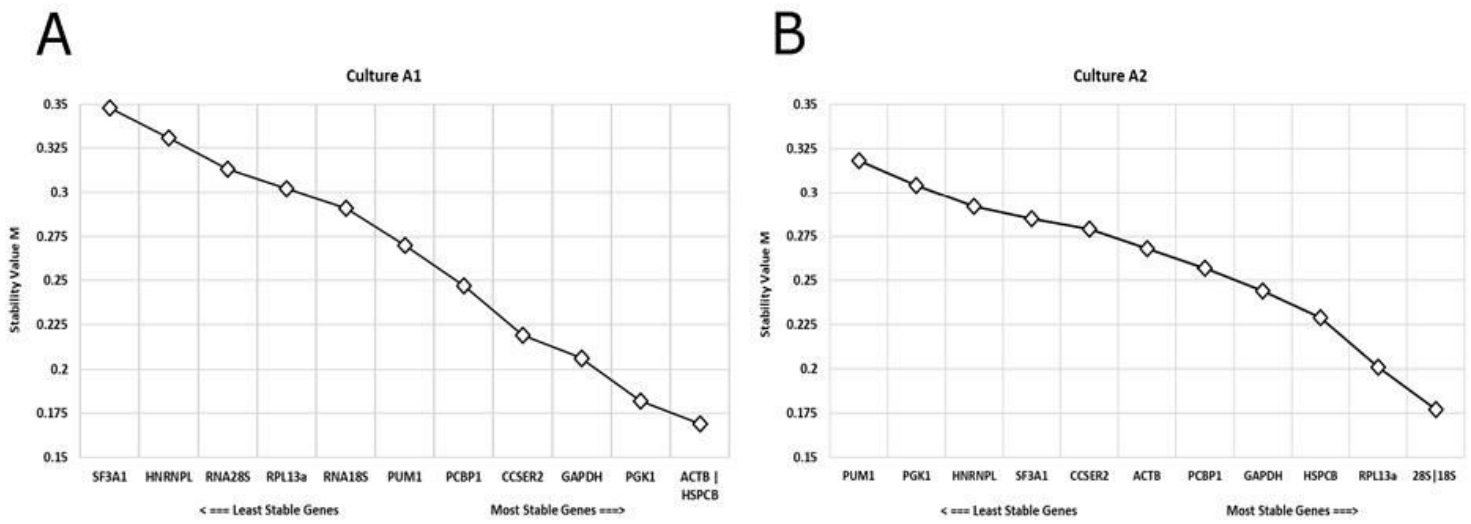

Figure 4

geNorm analysis of reference genes in both cultures A1 and A2 Legend: geNorm analysis of the selected candidate reference genes in both cultures (A) A1 and (B) A2 indicated as M values (stability values) on the Y-axis and genes on the X-axis. The lower the stability value, the more stable the gene and vice-versa. M value of less than 1 is considered appropriate for candidature as reference gene. 18S and 28S refers to RNA28S and RNA18S respectively.

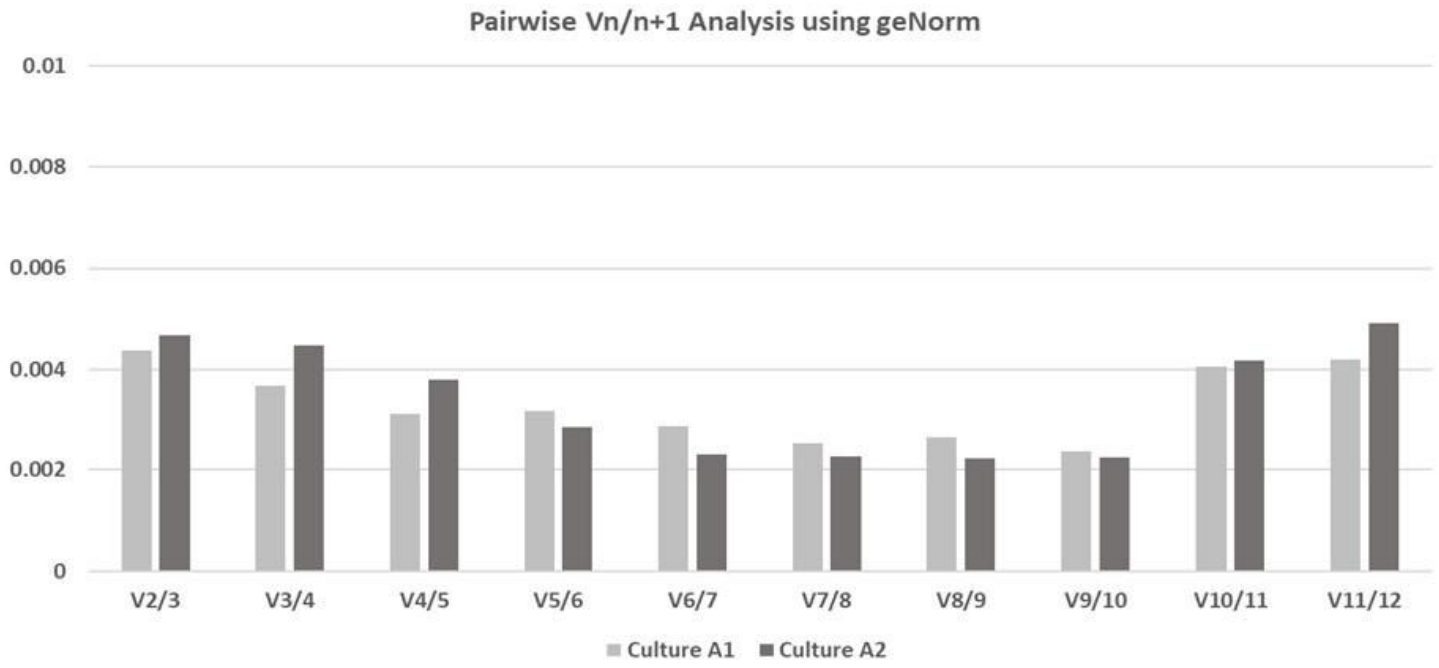

Figure 5

Determination of optimal number of reference genes needed for normalization (geNorm) for both cultures Legend: Pairwise Vn/n+1 analysis done using geNorm. The recommended cutoff is the lowest Vn/n+1 below the threshold of 0.15. In our case, both the cultures had V2/3 less than 0.15, indicating addition of a third reference gene would not affect normalization results.

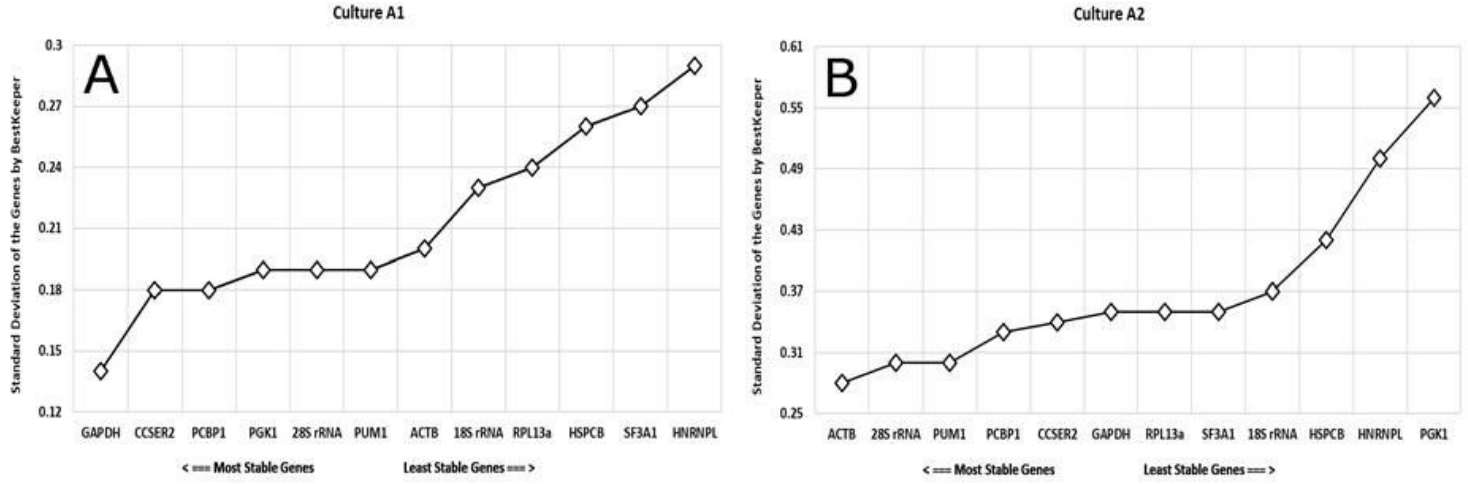

Figure 6

BestKeeper analysis of reference genes in both cultures A1 and A2 Legend: BestKeeper results for standard deviation with crossing points (S.D ± CP) on the Y axis with the selected candidate reference

genes on the X-axis for (A) culture A1 and (B) culture A2. The most stable genes are considered to have a S.D. as close as possible to 1 with not greater than 1. 18S rRNA and 28S rRNA refers to RNA18S and RNA28S respectively.

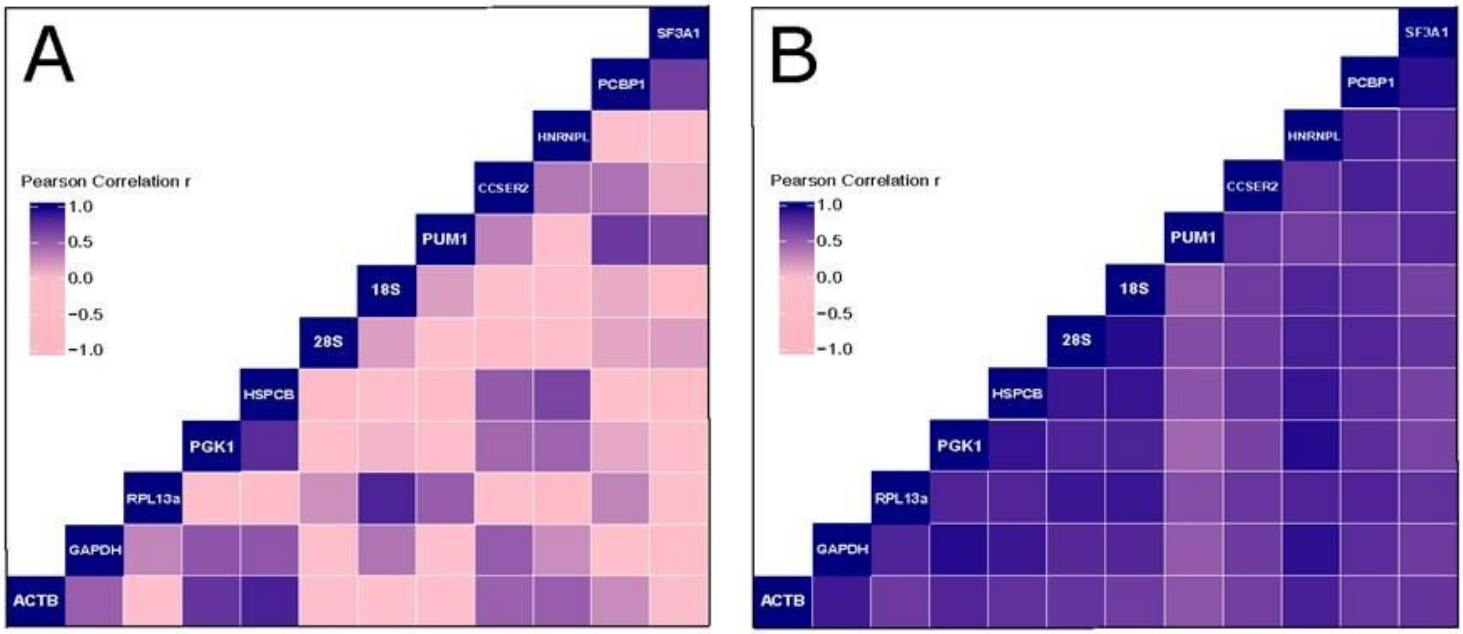

Figure 7

Pearson's correlation as determined by BestKeeper in both cultures A1 and A2 Legend: Pearson correlation of the pairwise gene expression stability as obtained from BestKeeper algorithm. The correlation was assessed for all candidate reference genes in both (A) culture A1 and (B) culture A2. The darker the blue, the higher the positive correlation between the genes as seen in the legend. 18S and 28S refers to RNA18S and RNA28S respectively.

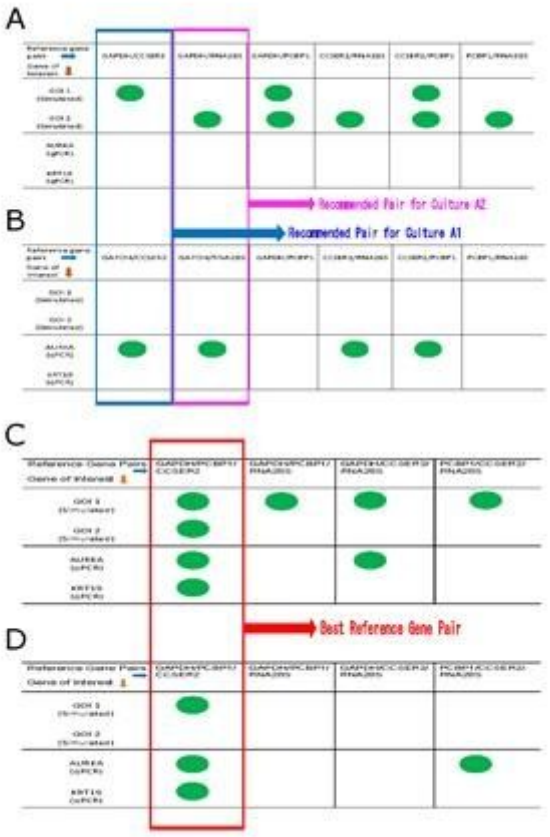

Figure 8

Summary of results of normalization of genes of interest in both cultures A1 and A2 Legend: Summary of normalization of GOI 1, GOI 2, AURKA and KRT19 by different recommended reference genes pairs. Green circles indicate successful normalization of the gene of interest by respective reference gene pair. (A) Normalization of the four genes of interest by selected reference genes in pairs of 2 in culture A1. (B) Normalization of the four genes of interest by selected reference genes in pairs of 2 in culture A2. (C) Normalization of the four genes of interest by selected reference genes in pairs of 3 in culture A1. (D) Normalization of the four genes of interest by selected reference genes in pairs of 3 in culture A2. Blue box indicates normalization by GAPDH-CCSER2 pair (recommended pair in culture A1) while Pink box indicates normalization by GAPDH-RNA28S pair (recommended pair in culture A2). Red box indicates the best reference gene triplet (GAPDH-PCBP1-CCSER2).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.pdf