Selecting optimal reference genes for breast cancer research

Qiang Song  
Chongqing University Three Gorges Hospital

Man Huang  
Chongqing Three Gorges University

Guicheng Wu  
Chongqing Three Gorges University

Lu Dou  
Chongqing University Three Gorges Hospital

Wenjin Zhang  
Chongqing University Three Gorges Hospital

Yang Peng  
The First Affiliated Hospital of Chongqing Medical University

Jiangen Wang  
Chongqing University Three Gorges Hospital

Chaoyi Wang  
Chongqing University Three Gorges Hospital

Qirong Zhao  
Chongqing Three Gorges University

Shaobo Wu  
Chongqing Three Gorges University

Xiaohe Luo  
Chongqing University Three Gorges Hospital

Jipan Zhang  
Southwest University

Mengyuan Wang  (mengyuanwang1@163.com)  
Chongqing Three Gorges University

Research article

Keywords: reference genes, breast cancer, qRT-PCR

DOI: https://doi.org/10.21203/rs.3.rs-91036/v1
Abstract

**Background** Quantitative reverse transcription-polymerase chain reaction (qRT-PCR) is the most sensitive technique for evaluating gene expression levels. Choosing appropriate reference genes (RGs) is critical for normalizing and evaluating changes in the expression of target genes. However, uniform and reliable RGs for breast cancer research have not been identified, limiting the value of target gene expression studies. Here, we provide a novel approach for mining RGs by using the RNA-seq dataset to identify reliable and accurate RGs that can be applied to different types of breast cancer tissues and cell lines.

**Methods** First, we compiled the transcriptome profiling data from the TCGA database involving 1217 samples to identify novel RGs and then ten genes (SF1, TARDBP, THRAP3, QRICH1, TRA2B, SRSF3, YY1, DNAJC8, RNF10, and RHOA) with relatively stable expression levels were chosen as novel candidate RGs. Additionally, six conventional RGs (ACTB, TUBA1A, RPL13A, B2M, GAPDH, and GUSB) were also selected. To determine and validate the optimal RGs we performed qRT-PCR experiments on 87 samples from 5 types of surgically excised breast tumor specimens including HR+HER2-, HR+HER2+, HR-HER2-, HR-HER2+, breast cancer after neoadjuvant chemotherapy (NAC) and their matched para-carcinoma tissues, furthermore, we also included a benign breast tumor sample. Six biological replicates were included for each tissue. Moreover, we assessed 7 breast cancer cell lines (MCF-10A, MCF-7, T-47D, MDA-MB-231, MDA-MB-468, as well as MDA-MB-231 with either CNR2 knockdown or overexpression; 3 biological replicates for each line). Five statistical algorithms (geNorm, NormFinder, ΔCt method, BestKeeper, and ComprFinder) were used to assess the stability of expression of each RG across all breast cancer tissues and cell lines.

**Results** Our results show that RG combinations SF1+TRA2B+THRAP3 and THRAP3+RHOA+QRICH1 showed stable expression in breast cancer tissues and cell lines, respectively, and that these two combinations displayed good interchangeability. Therefore, we propose that the above two combinations are optimal triplet RGs for breast cancer research.

**Conclusions** In summary, we identified novel and reliable RG combinations for breast cancer research based on a public RNA-seq dataset which lays a solid foundation for accurate normalization of qRT-PCR results across different breast cancer tissues and cells.

**Background** Quantitative reverse transcription-polymerase chain reaction (qRT-PCR) is a highly sensitive and low-cost technique that is widely used in molecular biology [1]. However, the accuracy and interpretation of its results are determined by the stability of the selected reference genes (RGs) [2]. Hence, the selection of suitable RGs is the first aim of any research system dedicated to the investigation of differential gene expression [3]. Furthermore, the simultaneous use of multiple RGs will result in more accurate data on target gene expression [2, 4].
Breast cancer is the most common malignancy in females and accounts for approximately 30% of all cancers diagnosed [5]. Based on the expression of hormone receptors (HR), including the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal-growth-factor receptor 2 (HER-2), breast cancer can be classified into four subtypes including HR+HER2-, HR+HER2+, HR-HER2+, and HR-HER2- [6]. During the course of breast cancer treatment, subtype status determines the use of neoadjuvant chemotherapy (NAC). In addition, breast disease also includes benign tumors [7]. Tumorigenesis and breast cancer metastasis are associated with gene expression changes that are most commonly detected using qRT-PCR [8]. In previous breast cancer studies commonly used RGs included beta-actin (ACTB), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), beta-glucuronidase (GUSB), ribosomal protein L13a (RPL13A), and tubulin alpha 1a (TUBA1A) [3, 8, 9]. However, research has indicated that these RGs are not consistently expressed across different tissues and experimental conditions [8, 10, 11]. Therefore, it is crucial to identify new RGs whose expression across various breast cancer tissues is more consistent.

Many novel RGs have been predicted and validated in many species and disease models, such as traumatic brain injury [12], Euscaphis konishii Hayata [13], Salix viminalis [4], Oryza sativa [14], Gentiana macrophylla [15], Homo sapiens [16], and Rhizophora apiculata [17]. However, to our knowledge, few systematic studies to validate potential RGs have been conducted for breast cancer. Available studies involved either tissues or cell lines (but not both), and the RGs concerned were not novel [3, 8, 18–20]. Considering the enormous threat breast cancer poses to human health, it is urgently necessary to identify RGs that are more relevant to a wide range of breast cancer tissues and cells across several conditions [21–23]. In this work, we hypothesized that novel RGs for breast cancer research could be identified and validated using an mRNA-seq dataset.

To this end, we employed the mRNA-seq datasets from The Cancer Genome Atlas (TCGA) to discover novel RGs. Ten relatively stably expressed genes (SF1, TARDBP, THRAP3, QRICH1, TRA2B, SRSF3, YY1, DNAJC8, RNF10, and RHOA) and six traditional RGs (ACTB, TUBA1A, RPL13A, B2M, GAPDH, and GUSB) were selected as the candidate RGs. The qRT-PCR experiments were performed on different experimental samples including six types of breast cancer tissues and seven different breast cancer cell lines. The stability of expression of these candidate RGs was calculated using geNorm [24], NormFinder [25], ΔCtmethod [26], BestKeeper [27], and ComprFinder [28]. Finally, the optimal RGs were validated and confirmed. Our study significantly improves upon previous work in breast cancer research.

**Methods**

**Breast cancer tumor**

Breast tumor and para-carcinoma tissues were supplied by the Breast Tumor Biobank of the Three Gorges Hospital Affiliated with Chongqing University. The fresh tissues were obtained from patients with written informed consent and with permission of the Three Gorges Hospital Affiliated with Chongqing University Clinical and the Laboratory Research Ethical Council. All tissues were stored frozen at -80°C.
after pathologic evaluation. We collected a total of 66 tissue samples including benign tumor tissues \((n = 6)\), as well as tissues from four subtypes of breast cancer including HR+/HER2- \((n = 6)\), HR+/HER2+ \((n = 6)\), HR-/HER2- \((n = 6)\), HR-/HER2+ \((n = 6)\) and their paired para-carcinoma tissues \((n = 6\) for each) from 24 patients who were diagnosed with breast cancer and from 6 patients who were diagnosed with breast cancer and then were treated with NAC before surgery. The para-carcinoma tissue samples had been taken from outside of the histopathological tumor border \((3\,\text{cm})\) in the same excisional biopsy specimen. The clinical patient information is shown in Additional file 1: Table S1.

**Cell lines and related treatment**

Breast cancer cell lines T-47D, MDA-MB-231, and MDA-MB-486 were purchased from the Cell Bank of the Type Culture Collection of the Chinese Academy of Sciences (Shanghai, China). MCF-10A and MCF-7 cell lines were purchased from the American Type Culture Collection (ATCC, Manassas, USA). MDA-MB-231 and MDA-MB-486 cells were cultured in Leibovitz's L-15 Medium \((\text{L-15, Gibco, USA})\). T-47D cells were cultured in Dulbecco's Modified Eagle Medium, containing high glucose and pyruvate without glutamine \((\text{DMEM, Gibco, USA})\). MCF-10A cells were cultured in Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 \((\text{DMEM/F-12, Gibco, USA})\) and MCF-7 cells were cultured in Minimum Essential Medium supplemented with 0.01 mg/ml bovine insulin \((\text{MEM, Gibco, USA})\). Moreover, we have constructed the MDA-MB-231 cell lines overexpressing CNR2 or CNR2 knock-down using lentiviruses \((\text{Genechem, Shanghai, China})\). All culture media were supplemented with 20U/mL penicillin, 100 mg/mL streptomycin, and 10% heat-inactivated fetal bovine serum \((\text{FBS, Australia, Gibco})\). Cells were grown at \(37^\circ\text{C}\) in a humidified atmosphere including 5% \(\text{CO}_2\). At the end-point of each experiment, the final pH of the supernatant was always measured by a digital pH-meter \((\text{pH 301, HANNA Instruments, USA})\).

**Total RNA Extraction and cDNA Synthesis**

Total RNA was isolated with RNAiso Plus \((\text{Takara, Dalian, China})\) using the phenol-chloroform method. Extracted RNA was quantified using Nanodrop One \((\text{ThermoFisher, Wilmington, USA})\) and its integrity was checked on a 1% agarose gel. Only RNA samples with A260/A280 ratios between 1.9 and 2.2 and A260/A230 ratios greater than 2.0 were used for cDNA synthesis. Total RNA \((1\,\mu\text{g})\) was reverse transcribed into cDNA using random primers or an oligo dT primer using a PrimeScript RT reagent Kit with gDNA Eraser \((\text{Takara, Dalian, China})\), according to the manufacturer's protocol \([29]\). All cDNA samples were diluted 1:8 with RNase-free water and stored at \(-20^\circ\text{C}\).

**Selection of Candidate Reference Genes**

The transcriptome sequencing dataset of 1217 breast cancer samples was downloaded from the TCGA database \((\text{https://www.cureline.com/the-cancer-genome-atlas.html})\) \((\text{Fig. 1A})\). After obtaining the gene fragments per kilobase of exon model per million mapped reads \((\text{FPKM})\), transcripts which exhibited low levels \((\text{FPKM} = 0\) appearing over 61 times in 1217 transcriptome profiles, \(1217\times5\% = 60.85\)) were removed. According to the FPKM value of every gene in all transcriptome profiles \([30]\), the coefficient of variation \((\text{CV})\) \([31]\), dispersion measure \((\text{DPM})\) \([32, 33]\), and FC-5% were calculated to screen for novel RGs \((\text{Fig. 1B})\).
The CV was defined as the CV value of the 1217 FPKM values of every gene. The DPM parameter was introduced for the identification of the RGs on the Pattern Gene Finder [33]. The FC-5% was defined as the fold change between the top 5% high expression levels divided by the bottom 5% within 1217 profiles. The standard criteria of candidate RGs were relatively high expression levels and low variation according to the results from FPKM, CV, DPM, and FC-5% analyses.

Furthermore, two frequently used RGs (ACTB and GAPDH) and four RGs (GUSB, RPL13A, TUBA1A, and B2M) from previous studies were also assessed along with the novel candidate RGs. All RGs were amplified using qRT-PCR experiment for subsequent determination and validation. The probability density curves were drawn using Matlab scripts from our previous study[28]. Venn diagram analysis was performed using a webtool (http://www.omicshare.com/tools).

**Primer Design and Amplification Efficiency Analysis**

The sequences of all genes used in this study were obtained from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/). Using Primer-BLAST, primers were designed to all transcripts, with Tm values around 60 °C, GC percent 45–55%, primer lengths of 18–24 bp, and product length of 80–250 bp. Primers were analyzed with Oligo Analyzer v3.1 (https://eu.idtdna.com/calc/analyzer) to detect potential self- and hetero-dimers [34]. The primers were synthesized by the Beijing Genomics Institute (Beijing, China). Primer specificities were confirmed by melting curve analysis.

**qRT-PCR analysis**

All qRT-PCR runs were carried out in a qTower2.2 PCR System (Analytik Jena, Germany). Reaction mixtures containing 7.5 µL TB Green Premix Ex Taq II (2X, Taq RNaseH Plus), 0.3 µL ROX Reference Dye II (50X, TaKaRa, Dalian, China), 1.5 µL cDNA, 0.6 µL each of forward and reverse primers (final concentration 1 µM), and 4.6 µl nuclease-free water were prepared in MicroAmp fast optical 96-well plates (ThermoFisher, USA). Amplification conditions were set as follows: 95°C for 30 s, followed by 40 cycles of 95°C for 5 s and 60°C for 34 s. Melting curve analysis was performed from 60–95°C. Reaction mixtures containing no template were used as negative controls. All samples were analyzed with three technical replicates. To test the amplification efficiency of each paired primer, serial ten-fold dilutions (1:10^3-1:10^10 ) of the primer corresponding PCR product were used to generate a standard curve [35]. The coefficient of determination (R^2) and slope (S) values were calculated from the standard curves and primer efficiencies (E) were calculated as 10^*(1/S)-1. The qRT-PCR experiments and analyses in this study were performed according to the Minimum Information for Publication of Quantitative Digital PCR Experiments (MIQE) guidelines[36].

**Analysis of Gene Expression Stability**

The cycle threshold (Ct) results from all runs were integrated into a data matrix. Then the data matrix was evaluated by four algorithms: geNorm [24], NormFinder [25], ΔCt method [26], and BestKeeper[27]. Finally,
the gene stability values from the above four algorithms were further evaluated by the ComprFinder method [28] (Fig. 1D).

Validation of the candidate reference genes

To verify the reliability of the stable RGs, four target genes including MAPK9 and MAPK3 from the extracellular signal-regulated kinase (ERK) signal pathway, and two other vital functional genes (FAAH, encoding fatty acid amide hydrolase, and HIF1A, encoding hypoxia-inducible factor 1-alpha) were chosen for validation (Fig. 1E). These target genes play an important role in the initiation and metastasis of breast cancer [37–41]. The independent-sample t-test was performed using Microsoft Excel, and the graphs were plotted using GraphPad Prism 7. The results are presented as Mean ± standard error of the mean (SEM), * P < 0.05, ** P < 0.01. For multiple gene combinations, the geometric mean of their Ct values was calculated. The relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method. To further evaluate the internal relationship of these different types of single- or multi- RG combinations, correlation analysis was performed as previously described [28]. Additionally, correlation analysis was also performed on the p-value dataset yielded in t-test analysis under different types of normalized factors [42].

Results

Identification of candidate RGs based on a public transcriptomic database

Transcriptome sequencing data of 1217 breast cancer samples were obtained from the TCGA database. Next, 15450 unigenes that were identified after processing were evaluated by FPKM (high expression level, FPKM $\geq$ 10), CV (low variability as determined by the coefficients of variation, CV $\leq$ 40%), FC-5% (the top 5% of 1217 samples divided by the lowest 5%, FC-5% $\leq$ 5), and DPM (DPM $\leq$ 0.3) values. The results for the different statistical algorithms, shown in Fig. 2, were as follows:

(1) FPKM. A total of 4723 genes satisfied the requirement (30.57% of 15450, the blue area in Fig. 2A).

(2) CV (%). There were 2649 genes with a CV $\leq$ 40% (17.15% of 15450, the purple area in Fig. 2B) after filtering.

(3) FC-5%. This parameter allowed the identification of 2287 genes (14.80% of 15450, the green area in Fig. 2C).

(4) DPM. This parameter resulted in the identification of 464 genes (3.00% of 15450, the red area in Fig. 2D).

Gene overlap between the four algorithms was identified using a Venn diagram with 4-color blocks (blue, purple, green, and red), showing that 197 genes (Fig. 2E) satisfied all four requirements. Of these 197 genes, 10 genes (SF1, TARDBP, THRAP3, QRIC1H1, TRA2B, SRSF3, YY1, DNAJC8, RNF10, and RHOA) were selected as novel candidate RGs due to their higher FPKM values and easier primers design. In addition,
GUSB, TUBA1A, RPL13A, and B2M, which previous studies suggested being stable RGs in breast cancer research, and two classical RGs, ACTB and GAPDH, were also considered. These genes were ranked based on their CV values (Table 1).

**Table 1** The summarised information of 16 potential RGs based on transcriptome data

| Gene     | FPKM  | CV     | FC-5% | DPM  | Order |
|----------|-------|--------|-------|------|-------|
| SF1      | 38.65 | 21.70% | 2.52  | 0.21 | 5     |
| TARDBP   | 20.30 | 18.51% | 2.32  | 0.18 | 1     |
| THRAP3   | 41.96 | 24.44% | 3.10  | 0.24 | 16    |
| QRICH1   | 14.67 | 25.76% | 3.11  | 0.25 | 45    |
| TRA2B    | 11.66 | 24.14% | 2.78  | 0.23 | 12    |
| SRSF3    | 39.09 | 24.08% | 2.76  | 0.23 | 11    |
| YY1      | 15.20 | 24.75% | 2.83  | 0.24 | 23    |
| DNAJC8   | 30.96 | 24.59% | 2.92  | 0.24 | 19    |
| RNF10    | 32.58 | 24.62% | 2.78  | 0.24 | 21    |
| RHOA     | 223.73| 25.60% | 3.05  | 0.25 | 40    |
| ACTB     | 1490.51| 38.06% | 5.02  | 0.36 | 1834  |
| TUBA1A   | 72.98 | 59.42% | 12.88 | 0.51 | 6728  |
| RPL13A   | 716.37| 56.46% | 8.78  | 0.49 | 6189  |
| B2M      | 625.30| 64.66% | 12.56 | 0.54 | 7483  |
| GAPDH    | 739.50| 72.26% | 11.79 | 0.59 | 8454  |
| GUSB     | 31.47 | 142.43%| 11.33 | 0.82 | 12695 |

**Primer specificity and amplification efficiency for qRT-PCR**

A total of 20 paired primers including 16 candidate RGs and 4 target genes were designed for qRT-PCR experiments. The unigene name, ENSid, gene description, primer sequences, theoretical Tm (°C), product size, primer efficiency (E), and coefficient of determination (R²) are summarized in Additional file 1: Table S2. The primer efficiency for all 20 genes ranged from 90.0% for YY1 to 105.4% for DNAJC8, and correlation coefficients varied from 0.996 (ACTB) to 0.999 (B2M, YY1). All paired primers showed adequate specificity (Additional file 1: Figure S1).

**Ct values of candidate reference genes**

The mean Ct value (average of 3 technical replicates) for every sixteen RGs are shown in Fig. 3 and Additional file 1: Table S3. The Ct values varied from 16.35 (RPL13A) to 24.57 (QRICH1) across various breast cancer tissues (Fig. 3A). The top 3 genes with low standard deviations were DNAJC8 (1.17),
The 3 most differentially expressed genes were GAPDH (2.03), B2M (1.93), and ACTB (1.91). However, the Ct values of breast cancer cell lines were overall lower than those of breast cancer tissues (Fig. 3B). A similar result of standard deviations was obtained in the breast cancer cells. To estimate the gene expression stability of these candidate RGs, more scientific algorithms will have to be introduced and used.

**Expression stability of candidate reference genes**

In this study, the qRT-PCR data matrix was analyzed using five differential algorithms: geNorm, NormFinder, BestKeeper, ΔCt method, and ComprFinder.

**geNorm analysis**

Gene expression stability was evaluated by the M value using geNorm analysis. This program determines the pairwise variation of each gene with all other analyzed genes under the same experimental conditions; the lower the M value, the higher the gene expression stability. In the breast cancer tissue group, the three most stably expressed genes (with the lowest M value) were SF1, THRAP3, and TARDBP while GAPDH, DNAJC8, and B2M were the least stably expressed genes (Table 2). In the breast cancer cell group, THRAP3, RHOA, and QRICH1 were the top three stably expressed genes, while B2M, TUBA1A, and ACTB were the least stably expressed genes (Table 3). For all samples, TARDBP was the most stably expressed gene, followed by SF1 and QRICH1. Conversely, TUBA1A, B2M, and ACTB were the least stably expressed RGs (Additional file 1: Table S4).
| Gene   | geNorm | NormFinder | BestKeeper | ΔCt method | ComprFinder |
|--------|--------|------------|------------|------------|-------------|
| SF1    | 0.369(1) | 0.018(5)   | 1.233(4)   | 0.626(4)   | 0.146(1)    |
| TRA2B  | 0.455(6) | 0.011(1)   | 1.334(10)  | 0.602(1)   | 0.152(2)    |
| THRAP3 | 0.386(2) | 0.016(3)   | 1.245(5)   | 0.624(3)   | 0.170(3)    |
| YY1    | 0.465(7) | 0.019(6)   | 1.283(7)   | 0.654(6)   | 0.193(4)    |
| RHOA   | 0.475(8) | 0.013(2)   | 1.314(8)   | 0.611(2)   | 0.200(5)    |
| RNF-10 | 0.441(5) | 0.017(4)   | 1.334(11)  | 0.637(5)   | 0.236(6)    |
| QRIC1  | 0.429(4) | 0.020(7)   | 1.281(6)   | 0.659(7)   | 0.241(7)    |
| TARDBP | 0.394(3) | 0.024(9)   | 1.183(3)   | 0.693(8)   | 0.269(8)    |
| SRSF3  | 0.514(9) | 0.022(8)   | 1.331(9)   | 0.727(9)   | 0.359(9)    |
| RPL13A | 0.615(12)| 0.037(13)  | 1.087(2)   | 0.839(12)  | 0.445(10)   |
| TUBA1A | 0.552(10)| 0.027(10)  | 1.424(13)  | 0.785(10)  | 0.513(11)   |
| DNAJC8 | 0.711(15)| 0.039(14)  | 0.992(1)   | 0.965(15)  | 0.583(12)   |
| GUSB   | 0.646(13)| 0.032(11)  | 1.344(12)  | 0.863(13)  | 0.593(13)   |
| ACTB   | 0.583(11)| 0.032(12)  | 1.529(14)  | 0.796(11)  | 0.608(14)   |
| GAPDH  | 0.675(14)| 0.046(15)  | 1.635(16)  | 0.882(14)  | 0.848(15)   |
| B2M    | 0.748(16)| 0.055(16)  | 1.576(15)  | 1.001(16)  | 0.977(16)   |
**Table 3** Gene expression stability calculated by 5 algorithms in breast cancer cell strain samples

| Gene   | geNorm | NormFinder | BestKeeper | ΔCt method | ComprFinder |
|--------|--------|------------|------------|------------|-------------|
| THRAP3 | 0.008(1) | 0.354(1) | 0.616(1) | 0.300(1) | 0.010(1) |
| RHOA   | 0.009(2) | 0.447(5) | 0.622(2) | 0.426(7) | 0.042(2) |
| QRICH1 | 0.013(3) | 0.544(12) | 0.664(3) | 0.507(9) | 0.111(3) |
| SF1    | 0.018(4) | 0.509(8) | 0.674(4) | 0.777(13) | 0.136(4) |
| RNF10  | 0.026(8) | 0.507(7) | 0.74(5) | 0.501(8) | 0.209(5) |
| DNAJC8 | 0.026(7) | 0.377(2) | 0.762(8) | 0.419(6) | 0.217(6) |
| GUSB   | 0.025(6) | 0.523(10) | 0.757(7) | 0.402(5) | 0.232(7) |
| YY1    | 0.027(9) | 0.495(6) | 0.773(9) | 0.353(3) | 0.254(8) |
| RPL13A | 0.038(11) | 0.393(3) | 0.784(11) | 0.639(11) | 0.267(9) |
| TARDBP | 0.019(5) | 0.539(11) | 0.744(6) | 0.318(2) | 0.268(10) |
| GAPDH  | 0.035(10) | 0.573(13) | 0.774(10) | 0.606(10) | 0.363(11) |
| TRA2B  | 0.039(12) | 0.421(4) | 0.918(12) | 0.390(4) | 0.386(12) |
| SRSF3  | 0.048(13) | 0.510(9) | 1.025(13) | 0.716(12) | 0.511(13) |
| B2M    | 0.058(14) | 0.946(15) | 1.084(14) | 0.821(14) | 0.767(14) |
| TUBA1A | 0.067(15) | 0.772(14) | 1.374(16) | 0.932(15) | 0.879(15) |
| ACTB   | 0.077(16) | 0.977(16) | 1.175(15) | 1.475(16) | 0.901(16) |

**NormFinder analysis**

Based on variance analysis to calculate the stable value of each gene, a higher NormFinder value indicates a less stably expressed gene. In the breast cancer tissue group, *TRA2B*, *RHOA*, and *THRAP3* were the most stable genes, and *DNAJC8*, *GAPDH*, and *B2M* were the most unstable genes (Table 2). In the breast cancer cell group, *THRAP3*, *DNAJC8*, and *RPL13A* were the three most stably expressed genes, while *TUBA1A*, *B2M*, and *ACTB* were the least stably expressed genes (Table 3). For all breast cancer tissue and cell samples, *THRAP3*, *RHOA*, *QRICH1* were the most stably expressed genes, and *TUBA1A*, *B2M*, *ACTB* were the least stably expressed RGs (Additional file 1: Table S4).

**BestKeeper analysis**

To further analyze the expression stability of the RGs, BestKeeper was applied, in which a lower std-value indicates a more stably expressed RG. As shown in Table 2, in the breast cancer tissue group *DNAJC8*, *RPL13A*, and *TARDBP* were the most stably expressed genes, while *ACTB*, *B2M*, and *GAPDH* were the least stably expressed genes (Table 2). In the breast cancer cell line group, *THRAP3*, *RHOA*, and *QRICH1* were the three most stably expressed genes, while *B2M*, *ACTB*, and *TUBA1A* were the least stably expressed genes (Table 3). For all samples combined, *DNAJC8*, *RPL13A*, and *TUBA1A* were the most
stably expressed genes, while \( \text{GAPDH} \), \( \text{RNF10} \), and \( \text{ACTB} \) were the least stably expressed RGs (Additional file 1: Table S4).

\( \Delta \text{Ct} \) analysis
According to the analysis results of the \( \Delta \text{Ct} \) method, \( \text{TRA2B} \), \( \text{RHOA} \), and \( \text{THRAP3} \) were the most stably expressed genes, while \( \text{DNAJC8} \), \( \text{GAPDH} \), and \( \text{B2M} \) were the least stable genes in the breast cancer tissue group (Table 2), which is consistent with the analysis according to NormFinder. In addition, in the breast cancer cell lines \( \text{THRAP3} \), \( \text{TARDBP} \), and \( \text{YY1} \) were the most stably expressed genes, while \( \text{B2M} \), \( \text{TUBA1A} \), and \( \text{ACTB} \) were the least stably expressed genes (Table 3). For all samples combined, \( \text{THRAP3} \), \( \text{RHOA} \), and \( \text{QRICH1} \) were the most stably expressed genes, while \( \text{TUBA1A} \), \( \text{B2M} \), and \( \text{ACTB} \) were the least stable RGs (Additional file 1: Table S4).

A comprehensive ranking of the four methods examined
The ComprFinder algorithm was carried out to obtain a final score (FS) which was used to rank the potential RGs. In the breast tumor tissue group, the 3 most stably expressed RGs were \( \text{SF1} \), \( \text{TRA2B} \), and \( \text{THRAP3} \) (Table 2). In the breast cancer cell lines, \( \text{THRAP3} \), \( \text{RHOA} \), and \( \text{QRICH1} \) were the most stably expressed RGs (Table 3). For all samples combined, we ranked the RGs from the highest to the lowest stability as follows: \( \text{THRAP3} > \text{RHOA} > \text{QRICH1} > \text{YY1} > \text{TRA2B} > \text{RPL13A} > \text{SF1} > \text{SRSF3} > \text{GUSB} > \text{TARDBP} > \text{DNAJC8} > \text{RNF10} > \text{GAPDH} > \text{TUBA1A} > \text{B2M} > \text{ACTB} \). Interestingly, the top 5 most stable genes (\( \text{THRAP3} \), \( \text{RHOA} \), \( \text{QRICH1} \), \( \text{YY1} \), and \( \text{TRA2B} \)) were novel RGs. In contrast, the traditionally used RGs \( \text{TUBA1A} \), \( \text{B2M} \), and \( \text{ACTB} \) were the least stably expressed RGs.

The research presented here confirmed that \( \text{THRAP3} \), \( \text{RHOA} \), \( \text{QRICH1} \), \( \text{YY1} \), and \( \text{TRA2B} \) were the most stable RGs in all samples with FS values of 0.064, 0.101, 0.122, 0.151, and 0.161, respectively (Table S4). These promising results warranted further validation of the selected RGs.

Validation of the selected genes (1): comparison of target gene profiles when using different normalized RGs
To verify the reliability of the selected RGs, the expression profiles of \( \text{MAPK3} \), \( \text{MAPK9} \), \( \text{FAAH} \), and \( \text{HIF1A} \) were determined in different breast cancer tissues and cell lines with different normalization factors (NF, i.e., the single- or multi-RG combinations). Our results indicated that \( \text{SF1} \), \( \text{TRA2B} \), and \( \text{THRAP3} \) were the top 3 stably expressed RGs in breast cancer tissues and that \( \text{THRAP3} \), \( \text{RHOA} \), and \( \text{QRICH1} \) were the top 3 stably expressed RGs in breast cancer cell lines. Moreover, five genes (\( \text{SF1} \), \( \text{TRA2B} \), \( \text{THRAP3} \), \( \text{RHOA} \), and \( \text{QRICH1} \)) were the top 5 stably expressed candidate RGs in all samples. Therefore, we considered the multi-RG combination \( \text{SF1} + \text{TRA2B} + \text{THRAP3} + \text{RHOA} + \text{QRICH1} \) as the most promising NF for breast cancer research (both in breast cancer tissues and cells). Thus, the multi-gene combinations including \( \text{SF1} + \text{TRA2B} + \text{THRAP3} + \text{RHOA} + \text{QRICH1} \), \( \text{SF1} + \text{TRA2B} + \text{THRAP3} \), \( \text{THRAP3} + \text{RHOA} + \text{QRICH1} \), \( \text{SF1} + \text{THRAP3} \), \( \text{THRAP3} + \text{RHOA} \), and single RGs including \( \text{SF1} \), \( \text{TRA2B} \), \( \text{THRAP3} \), \( \text{RHOA} \), and \( \text{QRICH1} \) were compared. In addition, \( \text{ACTB} \), \( \text{GAPDH} \), and \( \text{ACTB} + \text{GAPDH} \) were also used for comparison with the novel RGs. In total, 13 different multi-RG combinations or single RGs were assessed as NFs. For multiple gene
combinations, the geometric average of their Ct value was calculated. The relative gene expression level was calculated as $2^{-\Delta Ct}$, $\Delta Ct = \Delta (Ct_{\text{target genes}} - Ct_{\text{RGs}})$.

As shown in Fig. 4A, the expression of MAPK3 was significantly higher ($P < 0.05$) in HR + HER2- cancer tissue than in para-carcinoma tissue or benign tumor tissue when assessed by 5 or 3 multi-gene RG combinations. However, the expression pattern of MAPK3 changed when we used single or 2 multi-gene RG combinations, such as $SF1 + THRAP3$, $SF1$, $RHOA$, or $QRICH1$. Importantly, when we investigated the least stably expressed RGs ($ACTB$, $GAPDH$, or $ACTB + GAPDH$), the expression of MAPK3 was significantly changed compared with the most stably expressed RGs.

As shown in Fig. 4B, when using 5 or 3 multi-gene combinations, the expression level of the MAPK9 gene was higher in HR + HER2- cancer tissue than in para-carcinoma tissue ($P < 0.05$), while there was no significant difference between para-carcinoma tissue and benign tumor tissue. This may lead to small errors when using single or two multi-gene combinations. For example, when the less stably expressed genes $ACTB$, $GAPDH$, or $ACTB + GAPDH$, were used as the NF, the expression of MAPK9 did not show a clear expression trend compared with those of 5 or 3 multi-gene combinations.

In breast cancer cell lines, when the optimal RG combinations $SF1 + TRA2B + THRAP3 + RHOA + QRICH1$, $SF1 + TRA2B + THRAP3$, or $THRAP3 + RHOA + QRICH1$ were used for normalization, the expression of FAAH was highest in MCF-7 cells, followed by MCF-10A cells, and was least in MDA-MB-231 cells (Fig. 4C). When $ACTB$, $GAPDH$, or $ACTB + GAPDH$ were used for normalization, the expression of FAAH was not significantly different between MCF-10A and MDA-MB-231 cells.

The expression of HIF1A in breast cancer cells was higher ($P < 0.01$) in MCF-10A and MDA-MB-231 cells than in MCF-7 cells, while no significant difference was found between MCF-10A and MDA-MB-231 cells when using the 5 or 3 RG combinations ($SF1 + TRA2B + THRAP3 + RHOA + QRICH1$, $SF1 + TRA2B + THRAP3$, or $THRAP3 + RHOA + QRICH1$) for normalization (Fig. 4D). However, when $ACTB$ or $GAPDH$ (the less stably expressed RGs) were used, we found that HIF1A expression was significantly higher in MDA-MB-231 than in MCF-7 and MCF-10A cells.

The complete relative expression levels ($2^{-\Delta Ct}$) of MAPK3, MAPK9, FAAH, and HIF1A genes normalized using all 13 types of single or multiple RG combinations are listed in Additional file 1: Table S5 and Table S6.

**Validation of the selected genes (2): the relationship among different normalized RGs**

Based on the method described in our previous study [28], the relationship among different normalized RGs was explored. As shown in Additional file 1: Figure S2, there was a high correlation ($R^2$ from 0.815 to 0.979 in breast cancer tissues, and $R^2$ from 0.927 to 0.995 in breast cancer cells) between stable RGs and $SF1 + TRA2B + THRAP3 + RHOA + QRICH1$. There was also a moderate to high correlation ($R^2$ from 0.621 to 0.709 in breast cancer tissues, and $R^2$ from 0.600 to 0.916 in breast cancer cells) between unstable RGs and $SF1 + TRA2B + THRAP3 + RHOA + QRICH1$. There were few differences between the most stably
expressed RGs and the least stably expressed RGs. Therefore, we performed additional analyses of their normalized efficacy, including a correlation analysis on the \( p \)-value yielded by the \( t \)-test analysis (see Method section).

As shown in Fig. 5A, in breast cancer tissues, the normalized results using \( SF1 + TRA2 + THRAP3 \) (R\(^2\) = 0.847, \( P < 0.001 \)), \( THRAP3 + RHOA + QRICH1 \) (R\(^2\) = 0.947, \( P < 0.001 \)), \( SF1 + THRAP3 \) (R\(^2\) = 0.827, \( P < 0.001 \)), or \( THRAP3 + RHOA \) (R\(^2\) = 0.866, \( P < 0.001 \)) displayed a high correlation with \( SF1 + TRA2B + THRAP3 + RHOA + QRICH1 \) suggesting that they had extremely similar normalization capabilities. \( SF1 \), \( TRA2B \), and \( THRAP3 \) displayed a moderate correlation (R\(^2\) > 0.5), while \( RHOA \) or \( QRICH1 \) displayed a weak correlation (R\(^2\) < 0.5) with \( SF1 + TRA2B + THRAP3 + RHOA + QRICH1 \). There was a poor correlation between less stably expressed RGs (\( ACTB \), \( GAPDH \), or \( ACTB + GAPDH \)) and \( SF1 + TRA2B + THRAP3 + RHOA + QRICH1 \).

Similar results were found for the breast cancer cell lines (Fig. 5B). The complete \( p \)-value results yielded by \( t \)-test analysis are given in Additional file 1: Table S7 and Table S8.

**Discussion**

The qRT-PCR technique is one of the most valuable and reliable research tools used to quantify the expression of a target gene under different experimental conditions. Proper use of NFs is necessary to get a reliable estimate of gene expression in different types of breast cancer tissues and cells to avoid detecting variations that are not cancer-specific [43–45]. Therefore, the selection of the appropriate RGs for breast cancer research is important when using qRT-PCR to quantify gene expression. Many studies use a single endogenous control for normalization, which can influence the statistical results and may lead to erroneous data interpretation [2, 46]. In fact, in the present study no single RGs were identified that were stably expressed in all tissues or cell types across different types of breast cancer [7, 47, 48].

Theoretically, RGs should be stably expressed in all samples, and their expression levels should be unaffected by the external environment, e.g. during tumorigenesis [31, 49]. The selection and validation of RGs have to be corroborated by using a large number of samples [50, 51]. To implement this idea, in this study we collected a large number (\( n = 87 \)) of samples including 6 types of breast cancer tissues and 7 types of breast cancer cell lines. This allowed us to obtain strong results and conclusions. There was a great diversity of samples in our research for the following reasons: a) both benign and malignant tumor types were chosen, b) breast cancer samples following neoadjuvant chemotherapy were included, c) the breast cancer cell lines included overexpression and knockdown groups. With the above caveats explained, we propose that we have identified combinations of RGs that have high applicability in breast cancer research and treatment.

The target genes that were used in this study are involved in different biological processes of breast carcinogenesis and metastasis. Particularly, tumorigenesis, proliferation, apoptosis, and invasion are associated with many genes and signaling pathways. For example, genes like \( MAPK3 \) and \( MAPK9 \) encoding MAP kinases of the ERK signal pathway participate in transcription factor regulation of many biological processes [52, 53]. Recently, novel results have indicated proteins that serve important roles
during the process of cancer development. FAAH is a membrane-bound protein belonging to serine hydrolase family of enzymes that plays a significant role in the termination of signalling of fatty acid amides (FAAs), a class of bioactive lipids, both in the central nervous system and in some cancer tissues[54]. Hypoxia-inducible factors play an important role in the development of tumors, therefore the study functions of Hypoxia-inducible factors are also indispensable about cancer samples (HIF1A) [40, 55]. Therefore, to confirm the effects of these genes on the vital regulatory mechanisms in breast cancer, we compared the potential role of novel RGs (SF1, TRA2B, THRAP3, RHOA, QRIC1) vs. traditional RGs (ACTB, and GAPDH) in target gene expression normalization.

In our study, five algorithms were used to determine the stability of the expression of 16 candidate RGs across several different types of breast tumors and breast cancer cell lines. However, even for the same algorithm, the results varied between breast cancer tissues and cell lines. The top three genes for breast cancer tissues and cell lines were SF1 + TRA2B + THRAP3 and THRAP3 + RHOA + QRIC1, respectively, so a total of 5 RGs (SF1, TRA2B, THRAP3, RHOA, QRIC1) should be considered. However, simultaneous investigation of all five RGs would require a lot of effort.

There are still no specific literature reports prescribing how many candidate RGs should be used for qRT-PCR-dependent studies [56]. In particular, it is unknown which single or multiple gene combinations (SF1 + TRA2B + THRAP3 + RHOA + QRIC1, SF1 + TRA2B + THRAP3, THRAP3 + RHOA + QRIC1, SF1 + THRAP3, THRAP3 + RHOA, SF1, TRA2B, THRAP3, RHOA, or QRIC1) should be used. Considering that our results indicate that the single gene performances of both novel and traditional RGs, are not adequate, we propose that these types of studies should not be based on the use of single RG, even if they are top level RGs. The double gene combinations SF1 + THRAP3 and THRAP3 + RHOA (Fig. 4A-D) showed similar gene expression profiles consistent with SF1 + TRA2B + THRAP3 + RHOA + QRIC1, SF1 + TRA2B + THRAP3, and THRAP3 + RHOA + QRIC1. However, the SF1 + THRAP3 combination behaved similarly to the 5 or 3 gene combinations except for the MAPK3 and MAPK9 expression. Meanwhile, the THRAP3 + RHOA combination behaved similarly to the 5 or 3 gene combinations except for the MAPK9 expression. Therefore, considering the need for normalization accuracy, double RGs are not the best choice either.

The expression pattern of target genes was the same when 3-gene combinations or 5-gene combinations were used and they can be applied to various factors in breast cancer research. However, 3 RGs is a more manageable number for qRT-PCR experiments than 5 RGs. Therefore, we recommend that SF1 + TRA2B + THRAP3 and THRAP3 + RHOA + QRIC1 are adopted as the RG combinations for breast cancer tissue and cell line research, respectively. In the case of studies including both breast cancer tissue and cell line research, the THRAP3 + RHOA + QRIC1 combination would be optimal.

In this study, we not merely verified the use of conventional RGs but also identified and selected more appropriate novel RGs for breast cancer research. The use of a single RG is not recommended for breast cancer research. Similarly, the use of double RGs should be avoided. These findings are similar to what has been suggested in most of the studies using transcriptomic datasets [57]. Our results suggest that the recommended number of RG is at least three for breast cancer tissues or cell lines. Nevertheless,
these promising results require further verification of target genes in order to obtain more reliable data sets.

Conclusions

In this study, we tested sixteen different candidate RGs in six different breast cancer tissues and seven breast cancer cell lines, using five different statistical algorithms for evaluation. Our results indicate that $SF1 + TRA2B + THRAP3$ and $THRAP3 + RHOA + QRICH1$ are promising RG combinations for efficient gene normalization under different conditions. Furthermore, the availability of these RGs and the stability of their expression in various tumor tissues and cells will allow performing future studies focusing on genes essential for breast cancer biology, and choosing a reliable and appropriate RG combination will allow more accurate assessments of differential gene expressions in breast cancer research.

Abbreviations

RG: reference gene; HR : hormone receptors; HER-2: the human epidermal-growth-factor receptor 2; NAC: neoadjuvant chemotherapy; ACTB: beta-actin; GAPDH: glyceraldehyde-3-phosphate dehydrogenase; GUSB: beta-glucuronidase; RPL13A: ribosomal protein L13a; TUBA1A: tubulin alpha 1a; TCGA: The Cancer Genome Atlas; SF1: splicing factor 1; TARDBP: TAR DNA binding protein; THRAP3: thyroid hormone receptor associated protein 3; QRICH1: glutamine rich 1; TRA2B: transformer 2 beta homolog; SRSF3: serine and arginine rich splicing factor 3; YY1: YY1 transcription factor; DNAJC8: DnaJ heat shock protein family (Hsp40) member C8; RNF10: ring finger protein 10; RHOA: ras homolog family member A; FPKM: fragments per kilobase of exon model per million mapped reads; CV: coefficient of variation; DPM: dispersion measure; NCBI: the National Center for Biotechnology Information; MIQE: the Minimum Information for Publication of Quantitative Digital PCR Experiments; Ct: The cycle threshold; MAPK9: mitogen-activated protein kinase 9; MAPK3: mitogen-activated protein kinase 3; HIF1A: hypoxia inducible factor 1 subunit alpha; FAAH: fatty acid amide hydrolase; FS: final score; NF: normalization factors.

Declarations

Acknowledgements

We are thankful to all patients who participated in this study. We also thank the TCGA database ([https://www.cureline.com/the-cancer-genome-atlas.html](https://www.cureline.com/the-cancer-genome-atlas.html)) provides reliable data for this study.

Authors’ contributions

JPZ, QS and MH conceived and designed the study. MH and MYW collected the clinical samples. YP, GCW and CYW obtain accurate data by using the database. GCW, QRZ and JNW cultured cells. LD, WJZ and SBW extracted RNA. QS performed all qRT-PCR experiments and drafted the manuscript. JPZ
analyzed and interpreted the data, MYW, JPZ and XHL revised the manuscript. All authors read and approved the final manuscript.

**Funding**

This work was supported by the Natural Science Foundation Project of Chongqing, China (cstc2016jcyjA0338; cstc2020jcyj-msxmX0049; cstc2018jcyjAX0732; cstc2019jcyj-msxmX0823). This work was also supported in part by the National Natural Science Foundation of China (81900525) and the Doctoral Research Fund of Chongqing University Three Gorges Central Hospital (2017BSKYQDJJ06).

**Availability of data and materials**

The datasets used and/or analyzed during the current studies are available from the corresponding author on reasonable request.

**Ethics approval and consent to participate**

The study was approved by the ethics committee of the Three Gorges Hospital Affiliated with Chongqing University Clinical and the Laboratory Research Ethical Council (2020-26).

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no conflicts of interest with the contents of this article.

**Author details**

1The Center of Clinical Research of Endocrinology and Metabolic diseases in Chongqing, Chongqing University Three Gorges Hospital, School of Medicine, Chongqing University, Chongqing, China.

2Department of Central Laboratory, Chongqing University Three Gorges Hospital, School of Medicine, Chongqing University, Chongqing, China.

3Department of Breast Surgery, Chongqing University Three Gorges Hospital, School of Medicine, Chongqing University, Chongqing, China.

4Liver Center, Yu’an Branch of Department, Chongqing University Three Gorges Hospital, School of Medicine, Chongqing University, Chongqing, China.

5Department of Endocrine and Breast Surgery, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China.
6Chongqing Key Laboratory of Forage & Herbivore, Chongqing, China.

7College of Animal Science and Technology, Southwest University; Chongqing Key Laboratory of Forage & Herbivore, Chongqing, China.

8Department of Laboratory Medicine, Chongqing University Three Gorges Hospital, School of Medicine, Chongqing University, Chongqing, China.

References

1. Mocellin S, Rossi CR, Pilati P, Nitti D, Marincola FM: Quantitative real-time PCR: a powerful ally in cancer research. Trends Mol Med 2003, 9(5):189-195.
2. Kozera B, Rapacz M: Reference genes in real-time PCR. J Appl Genet 2013, 54(4):391-406.
3. 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, 17(1):639.
4. Ambroise V, Legay S, Guerriero G, Hausman JF, Cuypers A, Sergeant K: Selection of Appropriate Reference Genes for Gene Expression Analysis under Abiotic Stresses in Salix viminalis. Int J Mol Sci 2019, 20(17).
5. Shah KN, Faridi JS: Estrogen, tamoxifen, and Akt modulate expression of putative housekeeping genes in breast cancer cells. J Steroid Biochem Mol Biol 2011, 125(3-5):219-225.
6. Harbeck N, Gnant M: Breast cancer. Lancet 2017, 389(10074):1134-1150.
7. 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.
8. Kilic Y, Celebiler AC, Sakizli M: Selecting housekeeping genes as references for the normalization of quantitative PCR data in breast cancer. Clin Transl Oncol 2014, 16(2):184-190.
9. Yourick MR, Sandkam BA, Gammerdinger WJ, Escobar-Camacho D, Nandamuri SP, Clark FE, Joyce B, Conte MA, Kocher TD, Carleton KL: Diurnal variation in opsin expression and common housekeeping genes necessitates comprehensive normalization methods for quantitative real-time PCR analyses. Mol Ecol Resour 2019, 19(6):1447-1460.
10. Amorim LLB, Ferreira-Neto JRC, Bezerra-Neto JP, Pandolfi V, de Araujo FT, da Silva Matos MK, Santos MG, Kido EA, Benko-Iseppon AM: Cowpea and abiotic stresses: identification of reference genes for transcriptional profiling by qPCR. Plant Methods 2018, 14:88.
11. Brown AJ, Gibson S, Hatton D, James DC: Transcriptome-Based Identification of the Optimal Reference CHO Genes for Normalisation of qPCR Data. Biotechnol J 2018, 13(1).
12. Zamani A, Powell KL, May A, Semple BD: Validation of reference genes for gene expression analysis following experimental traumatic brain injury in a pediatric mouse model. Brain Res Bull 2020, 156:43-49.
13. Liang W, Zou X, Carballar-Lejarazu R, Wu L, Sun W, Yuan X, Wu S, Li P, Ding H, Ni L et al: Selection and evaluation of reference genes for qRT-PCR analysis in Euscaphis konishii Hayata based on
transcriptome data. Plant Methods 2018, 14:42.

14. Zhao Z, Zhang Z, Ding Z, Meng H, Shen R, Tang H, Liu YG, Chen L: Public-transcriptome-database-assisted selection and validation of reliable reference genes for qRT-PCR in rice. Sci China Life Sci 2020, 63(1):92-101.

15. He Y, Yan H, Hua W, Huang Y, Wang Z: Selection and Validation of Reference Genes for Quantitative Real-time PCR in Gentiana macrophylla. Front Plant Sci 2016, 7:945.

16. Molina CE, Jacquet E, Ponien P, Munoz-Guijosa C, Baczko I, Maier LS, Donzeau-Gouge P, Dobrev D, Fischmeister R, Garnier A: Identification of optimal reference genes for transcriptomic analyses in normal and diseased human heart. Cardiovasc Res 2018, 114(2):247-258.

17. Saddhe AA, Malvankar MR, Kumar K: Selection of reference genes for quantitative real-time PCR analysis in halophytic plant Rhizophora apiculata. PeerJ 2018, 6:e5226.

18. Rinnerthaler G, Hackl H, Gambenrieder SP, Hamacher F, Hufnagl C, Hauser-Kronberger C, Zehentmayer F, Fastner G, Sedlmayer F, Mlineritsch B et al.: miR-16-5p Is a Stably-Expressed Housekeeping MicroRNA in Breast Cancer Tissues from Primary Tumors and from Metastatic Sites. Int J Mol Sci 2016, 17(2).

19. Maltseva DV, Khaustova NA, Fedotov NN, Matveeva EO, Lebedev AE, Shkurnikov MU, Galatenko VV, Schumacher U, Tonevitsky AG: High-throughput identification of reference genes for research and clinical RT-qPCR analysis of breast cancer samples. J Clin Bioinforma 2013, 3(1):13.

20. 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.

21. Eisenberg E, Levanon EY: Human housekeeping genes, revisited. Trends Genet 2013, 29(10):569-574.

22. Garcia-Estevez L, Moreno-Bueno G: Updating the role of obesity and cholesterol in breast cancer. Breast Cancer Res 2019, 21(1):35.

23. Gur-Dedeoglu B, Konu O, Bozkurt B, Ergul G, Seckin S, Yulug IG: Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. Oncol Res 2009, 17(8):353-365.

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

25. 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 Res 2004, 64(15):5245-5250.

26. Lopa S, Ceriani C, Cecchinato R, Zagra L, Moretti M, Colombini A: Stability of housekeeping genes in human intervertebral disc, endplate and articular cartilage cells in multiple conditions for reliable transcriptional analysis. Eur Cell Mater 2016, 31:395-406.

27. 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(6):509-515.

28. Zhang J, Deng C, Li J, Zhao Y: Transcriptome-based selection and validation of optimal housekeeping genes for skin research in goats (*Capra hircus*). *BMC Genomics* 2020, 21(1):493.

29. Bashiardes S, Lovett M: cDNA detection and analysis. *Curr Opin Chem Biol* 2001, 5(1):15-20.

30. Mandric I, Temate-Tiagueu Y, Shcheglova T, Al Seesi S, Zelikovsky A, Mandoiu, II: Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from RNA-Seq data. *Bioinformatics* 2017, 33(20):3302-3304.

31. Pelabon C, Hilde CH, Einum S, Gamelon M: On the use of the coefficient of variation to quantify and compare trait variation. *Evol Lett* 2020, 4(3):180-188.

32. Anderson MJ, Ellingsen KE, McArdle BH: Multivariate dispersion as a measure of beta diversity. *Ecol Lett* 2006, 9(6):683-693.

33. Pan JB, Hu SC, Wang H, Zou Q, Ji ZL: PaGeFinder: quantitative identification of spatiotemporal pattern genes. *Bioinformatics* 2012, 28(11):1544-1545.

34. Chuang LY, Cheng YH, Yang CH: Specific primer design for the polymerase chain reaction. *Biotechnol Lett* 2013, 35(10):1541-1549.

35. Buh Gasparic M, Tengs T, La Paz JL, Holst-Jensen A, Pla M, Esteve T, Zel J, Gruden K: Comparison of nine different real-time PCR chemistries for qualitative and quantitative applications in GMO detection. *Anal Bioanal Chem* 2010, 396(6):2023-2029.

36. Bustin SA, Wittwer CT: MIQE: A Step Toward More Robust and Reproducible Quantitative PCR. *Clin Chem* 2017, 63(9):1537-1538.

37. Pietri E, Conteduca V, Andreis D, Massa I, Melegari E, Sarti C, Cenconetto L, Schirone A, Bravaccini S, Serra P *et al.*: Androgen receptor signaling pathways as a target for breast cancer treatment. *Endocr Relat Cancer* 2016, 23(10):R485-498.

38. Neuzillet C, Tijeras-Raballand A, de Mestier L, Cros J, Faivre S, Raymond E: MEK in cancer and cancer therapy. *Pharmacol Ther* 2014, 141(2):160-171.

39. Cajanus K, Holmstrom EJ, Wessman M, Anttila V, Kaunisto MA, Kalso E: Effect of endocannabinoid degradation on pain: role of FAAH polymorphisms in experimental and postoperative pain in women treated for breast cancer. *Pain* 2016, 157(2):361-369.

40. Sethuraman A, Brown M, Seagroves TN, Wu ZH, Pfeffer LM, Fan M: SMARCE1 regulates metastatic potential of breast cancer cells through the HIF1A/PTK2 pathway. *Breast Cancer Res* 2016, 18(1):81.

41. Deb S, Johansson I, Byrne D, Nilsson C, kConFab I, Constable L, Fjallskog ML, Dobrovic A, Hedenfalk I, Fox SB: Nuclear HIF1A expression is strongly prognostic in sporadic but not familial male breast cancer. *Mod Pathol* 2014, 27(9):1223-1230.

42. Du L, Liu F, Liu K, Yao X, Risacher SL, Han J, Guo L, Saykin AJ, Shen L, Alzheimer's Disease Neuroimaging I: Identifying diagnosis-specific genotype-phenotype associations via joint multitask sparse canonical correlation analysis and classification. *Bioinformatics* 2020, 36(Supplement_1):i371-i379.
43. Gao Z, Deng W, Zhu F: Reference gene selection for quantitative gene expression analysis in black soldier fly (Hermetia illucens). *PLoS One* 2019, 14(8):e0221420.

44. Hu J, Wang Z, Liao BY, Yu L, Gao X, Lu S, Wang S, Dai Z, Zhang X, Chen Q *et al*.: Human miR-1228 as a stable endogenous control for the quantification of circulating microRNAs in cancer patients. *Int J Cancer* 2014, 135(5):1187-1194.

45. Hu Y, Fu H, Qiao H, Sun S, Zhang W, Jin S, Jiang S, Gong Y, Xiong Y, Wu Y: Validation and Evaluation of Reference Genes for Quantitative Real-Time PCR in Macrobrachium Nipponense. *Int J Mol Sci* 2018, 19(8).

46. Gao D, Kong F, Sun P, Bi G, Mao Y: Transcriptome-wide identification of optimal reference genes for expression analysis of Pyropia yezoensis responses to abiotic stress. *BMC Genomics* 2018, 19(1):251.

47. McLoughlin KJ, Pedrini E, MacMahon M, Guduric-Fuchs J, Medina RJ: Selection of a Real-Time PCR Housekeeping Gene Panel in Human Endothelial Colony Forming Cells for Cellular Senescence Studies. *Front Med (Lausanne)* 2019, 6:33.

48. Nikishin DA, Filatov MA, Kiseleva MV, Bagaeva TS, Konduktorova VV, Khramova YV, Malinova IV, Komarova EV, Semenova ML: Selection of stable expressed reference genes in native and vitrified/thawed human ovarian tissue for analysis by qRT-PCR and Western blot. *J Assist Reprod Genet* 2018, 35(10):1851-1860.

49. Panina Y, Germond A, Masui S, Watanabe TM: Validation of Common Housekeeping Genes as Reference for qPCR Gene Expression Analysis During iPS Reprogramming Process. *Sci Rep* 2018, 8(1):8716.

50. Pombo MA, Ramos RN, Zheng Y, Fei Z, Martin GB, Rosli HG: Transcriptome-based identification and validation of reference genes for plant-bacteria interaction studies using Nicotiana benthamiana. *Sci Rep* 2019, 9(1):1632.

51. Rice J, Roberts H, Rai SN, Galandiuk S: Housekeeping genes for studies of plasma microRNA: A need for more precise standardization. *Surgery* 2015, 158(5):1345-1351.

52. Pradhan N, Parbin S, Kar S, Das L, Kirtana R, Suma Seshadri G, Sengupta D, Deb M, Kausar C, Patra SK: Epigenetic silencing of genes enhanced by collective role of reactive oxygen species and MAPK signaling downstream ERK/Snail axis: Ectopic application of hydrogen peroxide repress CDH1 gene by enhanced DNA methyltransferase activity in human breast cancer. *Biochim Biophys Acta Mol Basis Dis* 2019, 1865(6):1651-1665.

53. Lu H, Guo Y, Gupta G, Tian X: Mitogen-Activated Protein Kinase (MAPK): New Insights in Breast Cancer. *J Environ Pathol Toxicol Oncol* 2019, 38(1):51-59.

54. Tripathi RKP: A perspective review on fatty acid amide hydrolase (FAAH) inhibitors as potential therapeutic agents. *Eur J Med Chem* 2020, 188:111953.

55. Wang Y, Chen Y, Bao L, Zhang B, Wang JE, Kumar A, Xing C, Wang Y, Luo W: CHD4 Promotes Breast Cancer Progression as a Coactivator of Hypoxia-Inducible Factors. *Cancer Res* 2020, 80(18):3880-3891.
56. Jin Y, Liu F, Huang W, Sun Q, Huang X: Identification of reliable reference genes for qRT-PCR in the ephemeral plant Arabidopsis pumila based on full-length transcriptome data. *Sci Rep* 2019, 9(1):8408.

57. Krasnov GS, Kudryavtseva AV, Snezhkina AV, Lakunina VA, Benaminov AD, Melnikova NV, Dmitriev AA: Pan-Cancer Analysis of TCGA Data Revealed Promising Reference Genes for qPCR Normalization. *Front Genet* 2019, 10:97.

**Figures**

**Figure 1**

The workflow of this study (A) The gene expression profiles of 1217 breast cancer (BC) samples were obtained from TCGA public database; (B) Four indexes including FPKM, CV, DPM, and FC-5% were used to select candidate RGs and Venn diagram analysis identified the RGs common to these indexes; (C) qRT-PCR experiments were carried out on various BC tissues, including benign tumors, and BC cell lines, other types P&C, adjacent tissues (P) and cancer tissues (C). Six biological repetitions in each tissue group (66 samples in total) and three repetitions for each BC cell line including 21 samples) were used in this study. ov, over expression; kd, knockdown. (D) Candidate RGs were identified and evaluated by 5 public algorithms including geNorm, NormFinder, BestKeeper, ΔCt method, and ComprFinder; (E) The selected RGs were validated using 4 target genes. The capability and difference between all types of RGs were tested by correlation analysis.
Figure 2

Probability density curve of FPKM, CV, DPM and FC-5% in 15458 unigenes (A-D) The y-axes indicate the probability values for all 15457 genes. The candidate reference genes (RGs) were preliminarily screened by four criteria including FPKM ≥ 10, CV ≤ 0.4, DPM ≤ 0.3, and FC-5% ≤ 4. (E) The 197 RGs that met various criteria were identified by Venn diagram analysis, and among these 10 novel candidate RGs were selected.

Figure 3

Distribution of Ct values for 16 candidate reference genes Boxplot of absolute Ct value of the 16 candidate genes in breast cancer tissue samples (A) and cell lines (B). Boxes indicated median (Q2) and
first and third quartiles (Q1 and Q3, respectively) and whiskers corresponded to the minimum and maximum values.

**Figure 4**

Relative expression levels normalized by 13 types of single or multiple gene combinations of RGs. The relative expression levels of 2 target genes including MAPK3 (A) and MAPK9 (B) were determined in breast cancer (BC) tissue samples (C_HR+HER2, P_HR+HER2, and Benign Tumor), while the relative expression...
levels of 2 different target genes, FAAH (C) and HIF1A (D), were determined in BC cell lines (MCF-7, MCF-10A, and MDA-MD-231). Relative expression levels of each target gene were normalized to the most stable single RGs or multiple RG combinations (SF1+THRAP3+TRA2B+RHOA+QRICH1, SF1+THRAP3+TRA2B, THRAP3+RHOA+QRICH1, SF1+THRAP3, TRA2B+RHOA, SF1, THRAP3, TRA2B, RHOA, and QRICH1) and the least stable single RGs or multiple RG combinations (ACTB, GAPDH, and ACTB+GAPDH). The error bars represent the SEM, and the independent sample t-test in any two levels, *P <0.05, **P <0.01, n=6 for each BC tissue level, n=3 for each BC cell strain.

Figure 5

Heat map of correlation coefficients of detection efficiency based on different normalized RGs Two target genes were detected in 66 breast cancer (BC) tissue samples (A) and two target genes were detected in 21 BC cell strain samples (B) and were normalized by different types of RGs. The number in each color block is the correlation coefficient (R-value), and the number below each color block is the P-value of the corresponding R-value.

Supplementary Files

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

- Meltingcurvesforthe12candidateRGsand4targetgenes.pdf
- Heatmapofcorrelationcoefficientsofrelativelevels.png
- TableS1.docx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
• TableS5.xlsx
• TableS6.xlsx
• TableS7.xlsx
• TableS8.xlsx