Genome Based Search for Reference Genes for Gene Expression Analysis in Oral Cancer: a Data Science Driven Approach

Nehanjali Dwivedi
Mazumdar Shaw Medical Foundation
Sujan K Dhar
sankhya sciences
G Charitha
narayana health
Moni Abraham Kuriakose
Mazumdar Shaw Medical Foundation
Amritha Suresh
MSMF
Manjula Das (manjula.msmf@gmail.com)
Mazumdar Shaw Medical Foundation https://orcid.org/0000-0001-6202-3919

Research article

Keywords: Mouth Neoplasms (D009062), Data Science (D000077488), Head and Neck Neoplasms (D006258), Real-Time Polymerase Chain Reaction (D060888), Gene Expression (D015870)

Posted Date: June 8th, 2019

DOI: https://doi.org/10.21203/rs.2.10145/v1

Abstract

Background Quantitative real time PCR (qPCR) remains by far the most cost-effective, fast yet sensitive technique to check the gene expression levels in various systems. The traditionally used reference genes over the years were found to be regulated heavily based on sample sources and/or experimental conditions. This paper therefore presents a data science driven -omic approach for selection of reference genes from ~60,000 candidates from The Cancer Genome Atlas (TCGA) and Broad Institute Cancer Cell Line Encyclopaedia (CCLE) for gene expression studies in head and neck squamous cell carcinoma (HNSCC). mRNA-sequencing data of 500 patient samples and 33 cell lines from publicly available databases were analysed to assess stability of genes in terms of multiple statistical measures. A final set of 12 candidate genes were studied in the Indian set of data in Gene Expression Omnibus (GEO) and validated experimentally using qPCR in 35 different types of samples from platforms like drug sensitive and resistant cell lines, normal and tumor samples, fibroblast and epithelial primary culture derived from HNSCC patients from India. Result The study lead to the choice of five most stable reference genes –TYW5, RIC8B, PLEKHA3, CEP57L1 and GPR89B across three experimental platforms. Conclusion In addition to providing a set of five most stable reference genes for future gene expression analysis experiments across different types of samples in HNSCC, the study also presents an objective framework for assessing reference genes for other disease areas as well.

Results

Statistical Analysis of RNA-sequencing data

From 56,318 genes from cell lines and 60,483 genes from patient data set 18,764 and 19,661 protein coding genes respectively were selected. Protein coding genes with non-zero expression values in at least 50% of the samples (in cell lines 16,607 and patient samples 17,477) were exclusively chosen. After assigning the genes into standard quartiles based on median expression value, 8,303 and 8,738 genes were in middle quartiles for cell line and patient datasets respectively. Clustering results of each dataset based on z-scores of CV and MAD are shown in Fig 2a and b for cell lines and patient datasets respectively. Cluster 2 from cell lines and cluster 1 from patient dataset was chosen due to minimum medoid z-score. Number of genes in the selected clusters from cell line and patient dataset were 3,893 and 4,188 respectively, with 2,744 genes common between both clusters.

To rank the genes within each cluster, we defined a combined score as average of normalized values of CV and MAD. Comparison of this score for each gene in the cell line and patient dataset shows that they are moderately correlated with $r = 0.48$ (Supplementary Fig 1).

After programmatically pruning the list of 2,744 genes based on stop-words in their GO annotation to remove DNA binding proteins or transcription factors, a list of 675 candidate reference genes was obtained, from which the top 20 candidates with least value of combined score was selected for primer design and experimental validation.

Selection of Commonly Used Reference Genes

Pubmed search yielded a total of 118 unique abstracts which were manually curated by two authors independently yielding 28 unique genes from 10 relevant articles. Two genes *RNA18SN2* (ribosomal RNA) and *MTATP6P1* (mitochondrial RNA) were not captured in TCGA/CCLE mRNA-sequencing experiments, hence omitted from further analysis. Median expression values of 26 genes when divided into quartiles in patient samples in TCGA (Fig 3a) and in cell line data sets (Fig 3b) yielded only two reference genes – *HMBS* and *TBP* in the middle quartiles (Fig 3); *GAPDH, Beta Actin (ACTB)* and *HPRT1* were also chosen for further analysis because of their extensive literature based usage not only in head and neck cancer but also in other malignancies (Supplementary table 1).

Selection of Primers

From the top 20 selected candidates from publicly available data (TCGA and CCLE), melt curve analysis (Supplementary Fig 2) gave 11 genes with a single amplicon among which 8 genes had primer efficiencies ranging from 90-110% [48]. Among the 5 commonly used reference genes 4 had acceptable range of primer efficiency (Table 1) thus making the total number of selected candidate reference genes to 12.

Expression Behaviour of Candidate Genes in Cell Lines

Candidate reference genes when analysed in CCLE dataset (Fig 4a) revealed the expression of *GAPDH* and *ACTB* to be in the 75% quartiles of median expression level which if used as reference genes will miss out most of the overexpressing genes while over-predicting the down-regulated genes. The spread of both these genes are also larger than the other genes, especially obtained from the unbiased statistical analysis, indicating variations of expression among cell lines. The trend is similar in the in-house data (Supplementary Fig 3) though not as pronounced due to small dataset (8 in-house against 33 of CCLE). As shown in Fig 4b expression pattern of the candidate genes in drug resistant Cal27 cell lines showed different level of regulation, the least being in *RIC8B* and maximum in *HPRT1*.

Expression patterns were checked in the characterized primary cultures (Supplementary Fig 4a and b). Passage numbers did not have any effect on genes like *CEP57L1* and *TYW5* whereas some genes like *VTI1A* showed huge variation (Fig 4c). Epithelial and fibroblast cells from the same patient samples expressed *CEP57L1* and *TYW5* at equal levels whereas *VTI1A* was regulated (Fig 4d).

Behaviour of Candidate Genes in Patient Samples

Analysis of effect of tumor location in 500 TCGA dataset did not reveal any variation for all the 12 candidate genes Fig 5a and b with 44 unmatched normal showed similar profiles with very high expression of *GAPDH* and *ACTB* and moderately high expression for *TBP* and *HPRT1* in TCGA dataset. However, GEO dataset of 61 Tumor samples of Indian origin threw a different light pointing out higher variation in some of the stable genes obtained from TCGA (Fig 5c). Fold change analysis on a total of 10 matched adjacent normal and tumor samples from the in-house repository showed almost similar variations for all genes (Fig 5d). All of these results indicate need of a different reference gene set in the tumor set from Indian population compared to the stable genes found in analysis of Caucasian pool from TCGA.

Stability Analysis of candidate reference genes

Stability analysis of all 12 candidate reference genes using Cq values from all patient samples (both tumor and normal), cell lines and primary culture was carried out using RefFinder tool [49]. Geometric means of ranks obtained from both algorithms was used to rank the top 5 most stable genes – *TYW5*, *PLEKHA3*, *RIC8B*, *CEP57L1* and *GPR89B* (Fig 6).

TYW5 functions in iron binding and the biosynthesis of a hydroxywybutosine (a hyper-modified nucleoside) in tRNA by catalysing hydroxylation [50]. *RIC8B* guanine nucleotide exchange factor can activate some G-alpha proteins by changing bound GDP to free form GTP [51]. *PLEKHA3* has several biochemical functions and is involved in golgi apparatus to cell surface trafficking of protein cargo [52]. *CEP57L1* has been seen to be required for microtubule attachment to centrosomes [53-54]. *GPR89B* lastly is required for proper functioning of Golgi apparatus by maintaining the voltage dependent anion channel [55].

Discussion

The choice of reference genes becomes crucial for expression analysis of gene of interest. Levels of reference genes therefore, should remain unaltered in any given condition. The genes required for regular operations of a cell i.e. the house keeping genes were the obvious choice of reference genes. However, their expression has been shown to alter depending on different conditions. Considering only the functional aspect of the reference genes have led to erroneous picks. The functions of the chosen genes from the study also vary differently. *TYW5* functions in iron binding and the biosynthesis of a hydroxywybutosine (a hyper-modified nucleoside) in tRNA by catalysing hydroxylation [50]. *RIC8B* guanine nucleotide exchange factor can activate some G-alpha proteins by changing bound GDP to free form GTP [51]. *PLEKHA3* has several biochemical functions and is involved in golgi apparatus to cell surface trafficking of protein cargo [52]. *CEP57L1* has been seen to be required for microtubule attachment to centrosomes [53-54]. *GPR89B* lastly is required for proper functioning of Golgi apparatus by maintaining the voltage dependent anion channel [55]. Therefore, the choice of reference genes based only on their function is not sufficient. Current study thus offers an unbiased data-science based approach to shortlist reference genes. Most reliable reference genes should not be regulated across various platforms used in different gene expression analysis experiments like different cell lines, tumor and normal samples originating from different locations, various primary cultures across different passages or different drug treatment. Extreme

expressions of the reference genes can also either mask or falsely focus on differential expression of target genes. Thus, we have chosen a set of genes with moderate levels of expression across samples from various public databases by rigorous statistical analysis considering multiple parameters like CV and MAD. Moreover, they are extensively validated experimentally under different conditions as described.

The study reported here is a major improvement over similar approaches found in literature, even co-authored by two authors (AS and MAK) from this study [35]. Some of the improvements include (i) starting with an -omics pool of unbiased genes (ii) using a median-based variation parameter (MAD) in addition to the standard deviation based variation to make the analysis less susceptible to outliers often seen with patient samples (iii) using PAM clustering approach to identify a set of genes eliciting similar variations, and most importantly, (iii) extensive experimental validations using both patient and cell line datasets under various conditions like origin of tissue, inclusion of stromal tissues, drug sensitivity etc. to enhance applicability of reference genes in qPCR based analysis.

However, Fig 4 displays different level of regulation in the drug resistant cell lines and/or primary culture and Fig 5 points at a different type of HNSCC tumor in Indian population than the Caucasian population represented in TCGA [56]. Thus, though the current study displays a robust method, sequence data of various treatments of cell lines and primary cultures as well as tumor and adjacent normal samples from Indian dataset are required to find absolute set of 'invariant' reference genes, if at all.

Conclusion

The present study offers an unbiased; -omics based approach to arrive to a set of candidate genes which can be used as reference genes not only across different conditions, but also across three major platforms of research – cell lines, primary cultures and patient samples. Candidate genes, 12 in number, when checked by qPCR in 35 different systems (Fig 6 a) and subjected to RefFinder, chose 5 genes to be the most stable (Fig 6b): *TYW5*, *RIC8B*, *PLEKHA3*, *CEP57L1* and *GPR89B*. Although, a robust method for the choice of reference genes has been developed, still sequence data from Indian samples are required to come to an unbiased set of reference genes. The study therefore calls for an Indian Cancer Genome Atlas.

Methods

Gene Expression Data Acquisition

As represented in Fig 1, statistical analysis for detection of reference gene candidates was carried out based upon data generated by TCGA Research Network [36] and Broad Institute CCLE project [37]. RNA-sequencing FPKM values for a set of HNSCC patients (Project ID: TCGA-HNSC) were downloaded from NCI Genomic Data Commons Portal [38] from which the solid tumor data of 500 patients were selected. RNA-sequencing RPKM values of various cell lines were downloaded from the CCLE data portal [39] from which the data of 33 cell lines of upper aero-digestive tract origin were selected for analysis.

Expression of candidate reference genes were verified in Indian patients from gene expression datasets in GEO. Search on GEO for co-occurrence of search terms "Oral Cancer" or "Head and Neck Cancer" and "India" resulted in nine unique dataset, out of which only two datasets (GSE23558, GSE85195) reported gene expression values from Oral cancer patient samples from India [40,41]. Data in NCBI SOFT format were downloaded from the GEO portal corresponding to the above datasets. Since both datasets used the same microarray platform (Agilent 44K, GPL6480), Log_2 expression values from each dataset was merged for analysis. Altogether, both datasets had 61 tumor samples from Oral Cancer and 21 samples corresponding to precancerous lesions (Plakophoria) or normal samples. Expression data was analysed using R statistical software version 3.5.1.

Statistical Analysis of RNA-sequencingData

Protein coding genes with non-zero expression values in at least 50% of the samples were exclusively chosen for further analysis. For either cell line or patient dataset, genes were categorized on four standard quartiles based on their median expression value across samples. Genes in the two middle quartiles (Q2 and Q3) were shortlisted avoiding the extreme expression spectrum to enable capturing alteration in gene expression.

To assess stability of a gene, two measures were adopted – (i) CV = $\hat{x}$ σ x where $\hat{x}$ and σ x are mean and standard deviation of a variable *x* respectively and (ii) the normality p-value measured by the Shapiro-Wilks Test (p-value < 0.05 indicates the distribution to be away from Normal) [24]. CV, albeit most frequently used, is affected by outliers, and hence fails to be a robust measure. To address this, a third parameter –MAD = m e d i a n x - $\hat{x}$ (where $\hat{x}$ is the median of *x*) [42] was used after normalization with median. MAD is a measure of the spread of the distribution and being based on medians, is less susceptible to deviations by outliers.

However, for the patient dataset, the Normality p-value calculated by Shapiro-Wilks Test is $<10^{-4}$ for most genes, indicating that expression of none of the genes deviate from Normal distribution. Hence only CV and MAD were used as the two parameters for the study.

An ideal set of reference genes should have low or similar statistical variation (e.g. CV and MAD) across samples. Therefore, genes were clustered based on their CV and MAD values (normalized to respective z-scores) using the PAM algorithm originally proposed by Kaufman and Rousseeuw [43]. Optimal number of clusters required is calculated using the Silhouette graphical method of Rousseeuw [44]. For patient and cell line dataset, the cluster having the lowest medoid value for CV and MAD z-scores was selected, and the intersection between the two clusters was identified containing the genes having least CV and MAD values. This list was further pruned by programmatic parsing and eliminating genes based on stop words in their GO annotation such as transcription factors, nuclear receptor or other nuclear localization, DNA binding activity, response to external stimuli, translational and transcriptional activation etc. since genes with such characteristics having dependency on environmental conditions evidently are unsuitable as reference genes candidates. Top 20 genes from the pruned list with least CV and MAD values were selected for and experimental validation.

Selection of Commonly Used Reference Genes

Most commonly used reference genes were shortlisted by literature based on their frequency of usage in published papers. No unique keywords were used by researchers to report studies on reference genes. Many such articles are not indexed with MeSH terms so that the subheadings can be used for disease-based search. Hence a very broad methodology was adopted in which all articles in Pubmed were searched for occurrence of any of the terms "reference gene" or "control gene" or "housekeeping gene" along with co-occurrence of the term "head and neck" or “oral” anywhere in the article. Obtained abstracts were manually curated by authors ND and SKD independently to find the relevant articles that described studies on reference genes specifically in the context of oral or head and neck cancer. The shortlisted 28 genes were run on CCLE and TCGA database for expression analysis for their segregation among four standard quartiles.

Design of Primers

Primers were designed (Table 1; supplementary table 2) using Primer Bank Harvard [45] and IDT by searching NCBI gene symbol for human species. Primers with amplicon size 100-150 base pairs and melting temperature 60-65°C were selected, and synthesized by Juniper Life Sciences, Bangalore, Karnataka, India.

Cell culture

Eleven different HNSCC cell lines were used in the study. AW13516, SCC047, HSC3, Cal27 and SCC103 were cultured in DMEM medium (Gibco, #11965092) with 10% FBS (Gibco, #10270-106) and 1X PenStrep (Gibco, #15140122). DOK required addition of 500ng/mL of hydrocortisone (Sigma, #H0888) in the basal medium, while SCC029B and SCC040 required addition of non-essential amino acids (Gibco, #111450) along with the basal medium. Cal27 resistant cell lines were cultured with appropriate drugs. Cal 27 Cis R was maintained with Cisplatin (Sigma, #P4394) at a concentration of 3.3μM, Cal 27 Dox with Docetaxel (Sigma, #01885) at a concentration of 0.2nM and Cal 27 5FU with 5-flourouracil (Sigma, #F6627) at a concentration of 6 μM in DMEM medium with 10% FBS and 1X PenStrep [46].

Patient samples

All the samples were collected after obtaining prior written consent from the patients for the study for primary cultures and RNA isolation. The project was approved by NH Ethical Committee [IRB-12/01/2009; NHH/MEC-CL-2014/216]. Study was done on

retrospective samples with inclusion criteria being a matched set of adjacent normal and tumor samples from the same patient.

Primary culture and characterization

For culturing, tissue samples were collected aseptically in RPMI-1640 (Himedia, #AT222A) and DMEM F12 media (Gibco, #11320033) with triple strength penicillin – streptomycin solution (Gibco, #15140122). The tissue was chopped and taken for explant culture with 10ng/µl human recombinant epidermal growth factor (EGF, Sigma, #E9644), N2 suppliment-1X (Gibco, #17502048), Epilife defined growth supplement (Gibco, #S0125) with 20% FBS, in RPMI-1640 media with 1X penicillin-streptomycin for MTF12 and MTE12 sample and in DMEM F12 medium for MTF05. The cells were characterized by FACS and Immuno Cyto Chemistry (ICC).

FACS Analysis

Cells with a concentration of 10^6 cells/ 100µl were washed twice in PBS, permeabilizing with 0.1% triton X-100 for 30 min and incubated with primary antibody (1:50) for 1 hour on ice. The cell types were probed with anti-Pan cytokeratin (mAb, Cell signalling technologies, #4545), anti-Fibroblast surface protein (FSP, Sigma, #F4771), and anti-alpha SMA (Sigma, #A2547). Cells were then pelleted down and washed with PBS followed by incubation with the corresponding secondary antibody with anti-rabbit Alexa 647 (Invitrogen, #A31573) or anti mouse Alexa 488 (Invitrogen, #A11029) fluorochrome. For the marker expression, cell sorting gates were established using unstained control population.

Immuno-Cyto Chemistry

Coverslips with 5000 cells were fixed in 4% Paraformaldehyde (Fisher Scientific, #F79-1) followed by permeabilization with 0.1% triton X-100 (Sigma, #T8787), probed with anti-Pan cytokeratin, alpha SMA, FSP and vimentin (Dako, #M0725). The coverslips were processed with Dako Kit (Dako, #K5007). Presence of target proteins was visualised using DAB as chromogen and the cells counter stained with Haematoxylin (Himedia, #S034) and mounted with DPX mountant (Fisher Scientific, #18404) and examined under light microscope (Nikon ECLIPSE E200).

RNA extraction and cDNA conversion

Samples were collected in RNA later (Sigma, #R0901) and processed using MN kit (Macherey Nagel, #740955). RNA extraction for primary cultures and cell lines was done using TRIzol reagent (Ambion, #15596018) [47] and quantified using NanoDrop 2000 (Thermo Fisher Scientific) and QUBIT (Thermo, #Q10210). 1µg of total RNA was used for cDNA conversion using AMV Reverse Transcriptase enzyme (NEB, #M0277S) in a 20µl reaction as per manufacturer's instructions.

qPCR [3]

qPCR was done on Roche LightCycler 480 II instrument using KAPA SyBr green Universal (Sigma, #KK4600) in triplicates in clear plates with adhesive sealers. 1µl from 1:5 diluted cDNA was used in a total of 5µl reaction volume containing SyBr mix, primers, cDNA template and water. The reaction conditions were – pre-incubation at 95°C for 10 seconds followed by the amplification for 45 cycles (95°C – 1 second; 95°C – 10 seconds; 60°C – 15 seconds and 72°C -15 seconds). For further analysis, primers with single melt curve peak were chosen (Supplementary Fig 2)

For efficiency check a two-fold five-point dilution of Cal27 Parental cDNA was used as template. Thermo primer efficiency calculator [48] was used to calculate the efficiency of primers using the equation $E= 10/^{-1/slope}$.

Data analysis

Chosen 12 reference genes were validated across 35 different samples in triplicates. Cq values thus obtained were subtracted by geometric mean of non-template control (NTC) to obtain $C_q = C_q(sample) – Geo\ mean\ C_q(NTC)$ from which the relative expression was calculated as 2^{-Cq} for each replicate. Arithmetic mean of expression values of the replicates are plotted for the chosen reference genes across different samples as depicted in results.

Abbreviations

ACTB: Beta actin aka: also known as CCLE: Cancer Cell Line Encyclopaedia cDNA: Complementary DNA *CEP57L1:* Centrosomal protein 57 like 1

CV: Coefficient of Variation Cq: Quantitation Cycle DNA: deoxy ribonucleic acid

EGF: Epidermal Growth Factor FACS: Fluorescence Activated Cell Sorting Fig: Figure

FPKM: fragments per kilobase million FSP: Fibroblast Surface Protein *GAPDH:* Glyceraldehyde-3-Phosphate Dehydrogenase GDP: Guanosine diphosphate

GEO: Gene Expression Omnibus GO: Gene Ontology *GPR89B:* G-protein coupled receptor 89B GTP: Guanosine triphosphate *HMBS:*Hydroxymethylbilane Synthase

HNSCC: Head and Neck Squamous Cell Carcinoma HPRT1:Hypoxanthine Phosphoribosyltransferase 1 IDT: Integrated DNA Technologies MAD: Median Absolute Deviation MeSH: Medical Sub Heading mRNA: Messenger RNA MTATP6P1: Mitochondria-ATP6 pseudogene 1 NCBI: National Center for Biotechnology Information

NCI: National Cancer Institute NTC: Non Template control PAM: Partitioning around Medoids PBS*:* Phosphate Buffered Saline *PLEKHA3:*Pleckstrin homology domain containing A3 qPCR: Quantitative Polymerase Chain Reaction *RIC8B:*RIC8 guanine nucleotide exchange factor B RNA: Ribonucleic acid RNA18SN2: 18S Ribosomal N2 RPKM: reads per kilobase million SMA: Smooth Muscle Actin

SOFT: Simple Omnibus Format in Text *TBP:* TATA-Box Binding Protein TCGA: The Cancer Genome Atlas tRNA: transfer RNA *TYW5:* tRNA-yW synthesizing protein 5

VTI1A: Vesicle Transport Through Interaction With T-SNAREs 1A

Declarations

Ethics approval and consent to participate

The project was approved by NH Ethical Committee [IRB-12/01/2009; NHH/MEC-CL-2014/216].

Consent for publication

Publication is approved as a part of the written consent form under the project approved by NH Ethical Committee [IRB-12/01/2009; NHH/MEC-CL-2014/216].

Availability of data and material

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

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests

Funding

Current study was funded by MSMF. The funding agency did not have any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author’s contributions

SKD contributed in conceptualization, study design, data acquisition (algorithms), data interpretation and manuscript preparation. ND contributed in doing qPCR for all samples, cell culture, data analysis, data interpretation and manuscript preparation. CG

contributed in the establishment of primary cultures from patients. MAK, oral oncology surgeon contributed in getting the patient consent and samples. MD contributed in study design, conceptualization and manuscript editing. AS contributed in reviewing the manuscript.

Acknowledgements

The authors thank Joy Kuri, Haresh Dagale and Chandramani Singh for critical review of the analysis procedure and Department of Electronic Systems Engineering, IISc, Bangalore for kindly providing computing infrastructure. Cell lines used for the study were procured from various institutes – SCC029B, SCC103 and SCC040 from Dr. Susanne M Gollin, University of Pittsburgh, USA; DOK from Roswell Park Cancer Institute, Buffalo, USA; Cal27 Parental from Dr. Aditi Chatterjee, Institute of Bioinformatics, Bangalore, Karnataka, India; HSC3 from Dr. Shumpei, Tokyo Medical and Dental University, Tokyo, Japan; SCC047 from Dr. Thomas E Carey, University of Michigan, USA and AW13516 from ACTREC, Mumbai, India. All the Cal 27 resistant cell lines were a kind gift from author AS. We thank all of them for their kind contribution.

References

[1] Huggett J, Dheda K, Bustin S, Zumla A. Real-time RT-PCR normalisation ; strategies and considerations. 2005;279–84.

[2] Kozera B, Rapacz M. Reference genes in real-time PCR. J Appl Genet. 2013;54(4):391–406.

[3] Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE Guidelines : M inimum I nformation for Publication of Q uantitative Real-Time PCR E xperiments SUMMARY : 2009;622:611–22.

[4] Thellin O, Zorzi W, Lakaye B, Borman B De, Coumans B. Housekeeping genes as internal standards : use and limits. 1999;75:291–5.

[5] Connell GCO, Treadway MB, Petrone AB, Tennant CS, Lucke-wold N, Chantler PD, et al. Leukocyte Dynamics Influence Reference Gene Stability in Whole Blood : Data-Driven qRT-PCR Normalization Is a Robust Alternative for Measurement of Transcriptional Biomarkers. 2017;346–56.

[6] Li M, Rao M, Chen K, Zhou J, Song J. left and right ventricles. Gene [Internet]. Elsevier; 2017;620(February):30–5.

[7] Bamias G, Goukos D, Laoudi E, Balla IG. Comparative Study of Candidate Housekeeping Genes for Quanti fi cation of Target Gene Messenger RNA Expression by Real-Time PCR in Patients with In fl ammatory Bowel Disease. 2013;19(13).

[8] Arenas-hernandez M, Vega-sanchez R. Housekeeping gene expression stability in reproductive tissues after mitogen stimulation. BMC Res Notes [Internet]. BMC Research Notes; 2013;6(1):1. Available from: BMC Research Notes

[9] Almeida TA, Quispe-ricalde A, Montes F, Oca D, Foronda P, Hernández MM. Gynecologic Oncology A high-throughput open-array qPCR gene panel to identify housekeeping genes suitable for myometrium and leiomyoma expression analysis. Gynecol Oncol [Internet]. Elsevier Inc.; 2014; Available from: http://dx.doi.org/10.1016/j.ygyno.2014.04.012

[10] Paula A, Aline S, Damo F, Tania B, Furlanetto W. Validation of Reference Genes for Normalizing Gene Expression in Real-Time Quantitative Reverse Transcription PCR in Human Thyroid Cells in Primary Culture Treated with Progesterone and Estradiol. 2013;278–82.

[11] Kaszubowska L, Karsznia S, Damska M, Foerster J. Journal of Immunological Methods. 2015

[12] Li YI, Xiang GUIM, Liu LINLIN, Liu C, Liu FEI, Jiang DN, et al. Assessment of endogenous reference gene suitability for serum exosomal microRNA expression analysis in liver carcinoma resection studies. 2015;4683–91.

[13] Caracausi M, Piovesan A, Antonaros F, Strippoli P, Vitale L, Pelleri MC. Systematic identification of human housekeeping genes possibly useful as references in gene expression studies. 2017;2397–410.

[14] Campos RP De, Schultz IC, Mello PDA, Davies S, Gasparin MS, Paula A, et al. Cervical cancer stem like cells: Systematic review and identification of reference genes for gene expression †. 2017. 1-32 p.

[15] Wierzbicki PM, Klacz J, Rybarczyk A, Slebioda T. Identification of a suitable qPCR reference gene in metastatic clear cell renal cell carcinoma. 2014;12473–87.

[16] Romani C, Calza S, Todeschini P, Tassi RA, Zanotti L, Bandiera E, et al. Identification of Optimal Reference Genes for Gene Expression Normalization in a Wide Cohort of Endometrioid Endometrial Carcinoma Tissues. 2014;1–16.

[17] Noriega NC, Kohama SG, Urbanski HF. Microarray analysis of relative gene expression stability for selection of internal reference genes in the rhesus macaque brain. 2010

[18] Vandesompele J, Preter K De, Poppe B, Roy N Van, Paepe A De. Accurate normalization of real-time quantitative RT -PCR data by geometric averaging of multiple internal control genes. 2002;1–12.

[19] Andersen CL, Jensen JL, Ørntoft 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. 2004;5245–50.

[20] 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. 2004;509–15.

[21] Beer L, Mlitz V, Gschwandtner M, Berger T, Narzt M, Gruber F, et al. Bioinformatics approach for choosing the correct reference genes when studying gene expression in human keratinocytes. 2015;(7):742–7.

[22] Bahr SM, Borgschulte T, Kayser KJ, Lin N. Using Microarray Technology to Select Housekeeping Genes in Chinese Hamster Ovary Cells. 2009;104(5):1041–6.

[23] Thomas D, Finan C, Newport MJ, Jones S. DNA entropy reveals a signi fi cant difference in complexity between housekeeping and tissue speci fi c gene promoters. Comput Biol Chem [Internet]. Elsevier Ltd; 2015;58:19–24. Available from: http://dx.doi.org/10.1016/j.compbiolchem.2015.05.001

[24] Yim AK, Wong JW, Ku Y, Qin H. Using RNA-sequencingData to Evaluate Reference Genes Suitable for Gene Expression Studies in Soybean. 2015;1–15.

[25] Carmona R, Arroyo M, José M, Quesada J, Seoane P, Zafra A. Automated identification of reference genes based on RNA seq data. Biomed Eng Online. BioMed Central; 2017;16(s1):11–33.

[26] Hoang VLT, Tom LN, Quek X, Tan J, Payne EJ, Lin LL, et al. RNA-sequencingreveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. 2017;1–15.

[27] Spiegelaere W De, Dern-wieloch J, Weigel R, Schumacher V, Vandekerckhove L, Fink C. Reference Gene Validation for RT-qPCR , a Note on Different Available Software Packages. 2015;1–13.

[28] Lallemant B, Evrard A, Combescure C, Chapuis H, Chambon G, Raynal C, et al. Reference gene selection for head and neck squamous cell carcinoma gene expression studies. 2009;10:1–10.

[29] Yigin AK, Cora T, Acar H, Kurar E, Kayis SA, Colpan B, et al. Selection of reliable reference genes for qRT-PCR analysis on head and neck squamous cell carcinomas . 2018;28(5):2014–8.

[30] Song W, Li Y, Ren M, Wang D, Li Y, Zhang T, et al. Validation of reference genes for the normalization of qRT-PCR expression studies in head and neck squamous cell carcinoma cell lines treated by different chemotherapy drugs. 2018;11(3):2430–7.

[31] Rentoft M, Hultin S, Coates PJ, Laurell G, Nylander K. Tubulin α -6 chain is a stably expressed reference gene in archival samples of normal oral tissue and oral squamous cell carcinoma. 2010;419–23.

[32] Faibish D, Suzuki M, Bartlett JD, Forsyth T. HHS Public Access. 2017;33–42.

[33] Martin JL. Validation of Reference Genes for Oral Cancer Detection Panels in a Prospective Blinded Cohort. 2016;1–7.

[34] Song W, Zhang WH, Zhang H, Li Y, Zhang Y, Yin W, et al. Cellular and Molecular Biology in oral squamous cell carcinoma cell line treated by 5 kinds of chemotherapy drugs. 2016;62(13):29–34.

[35] Palve V, Pareek M, Krishnan NM, Siddappa G, Suresh A, Kuriakose MA. A minimal set of internal control genes for gene expression studies in head and neck squamous cell carcinoma. 2018;1–12.

[36] http://cancergenome.nih.gov/; 24 April 2019.

[37] Cancer T, Line C. HHS Public Access. 2012;483(7391):603–7.

[38] https://portal.gdc.cancer.gov/; 24 April 2019.

[39] https://portals.broadinstitute.org/ccle; 24 April 2019.

[40] Ambatipudi S, Gerstung M, Pandey M, Samant T. NIH Public Access. 2013;51(2):161–73.

[41] Bhosale PG, Cristea S, Ambatipudi S, Desai RS, Kumar R, Patil A, et al. Tr a n s l a t i o n a l O n c o l o g y Chromosomal Alterations and Gene Expression Changes Associated with the Progression of Leukoplakia to Advanced Gingivobuccal Cancer. Transl Oncol [Internet]. The Authors; 2017;10(3):396–409. Available from: http://dx.doi.org/10.1016/j.tranon.2017.03.008

[42] Moncton U De, Brunswick N, Hung TL, Nam V. The Mean and Median Absolute Deviations. 2001;7177(01).

[43] Society IB. Finding Groups in Data : An Introduction to Cluster Analysis . by L . Kaufman ; P . J . Rousseuw Review by : J . E . Gentle. 2014;47(2).

[44] Rousseeuw PJ. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. 1987;20:53–65.

[45] https://pga.mgh.harvard.edu/primerbank/; 24 April 2019.

[46] Govindan SV, Kulsum S, Pandian RS, Das D, Seshadri M, Jr WH, et al. Establishment and characterization of triple drug resistant head and neck squamous cell carcinoma cell lines. 2015;3025–32.

[47] Rio DC, Jr MA, Hannon GJ, Nilsen TW. Purification of RNA Using TRIzol (TRI Reagent). 2019;2010(6):1–4.

[48] https://www.thermofisher.com/in/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/qpcr-efficiency-calculator.html; 24 April 2019.

[49] Xie F, Xiao P, Chen D. miRDeepFinder : a miRNA analysis tool for deep sequencing of plant small RNAs. 2012;75–84.

[50] Hydroxylase RNA. Expanding Role of the Jumonji C Domain as an. 2010;285(45):34503–7.

[51] Montecino N, Soto X, Guzma L, Klattenhoff C, Mellstrom B, Romo X, et al. Human Brain Synembryn Interacts With Gs a and Gq a and Is Translocated to the Plasma Membrane in Response to Isoproterenol and Carbachol. 2003;157(January):151–7.

[52] Godi A, Campli A Di, Konstantakopoulos A, Tullio G Di, Alessi DR, Kular GS, et al. FAPPs control Golgi-to-cell-surface membrane traffic by binding to ARF and PtdIns (4) P. 2004;6(5).

[53] Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, et al. Next-generation sequencing to generate interactome datasets. 2011;8(6):13–7.

[54] Manuscript A. NIH Public Access. 2015;159(5):1212–26.

[55] Maeda Y, Ide T, Koike M, Uchiyama Y, Kinoshita T. ARTICLES GPHR is a novel anion channel critical for acidification and functions of the Golgi apparatus. 2008;10(10).

[56] Dotto G, Hospitalier C, Vaudois U, Hospital MG. HHS Public Access. 2018;3(3):181–97.

Table 1

Table 1 Primer details: Primer sequences, melting temperature and primer efficiency of the shortlisted genes used in the study. The primers are arranged as per their stability (most to least stable).

HGNC Symbol	Forward primer (5′-3′)	Tm	Reverse primer (5′-3′)	Tm	Efficiency	Amplicon length (bp)
TYW5	CAGCATCAAGAGCTGCACAAA	61.5	TGTGTAGGACCATTCGTCGTG	61.8	100.97	102
PLEKHA3	ACTGTGACCTCTTAATGCAGC	60	CTCAAGCGTTGTGATGAATGTG	60.1	105.35	146
RIC8B	ATAGTGTTCAACAGTCAGATGGC	60.3	GCAAGCGCAAGTCAAAGCA	62.2	110.39	133
CEP57L1	ATGAACCATCTCAGAATTGCCAT	60	TCTCTCCAGCTCTAAACGATGAA	60.5	108.64	137
GPR89B	TCCGTGACGTTTGCATTTTCT	60.8	GCAGTAGTCGGATATTGCTCACA	62	106.12	184
STIMATE	GCTAAGGTGTGATGAGCTAGAA	62	CTCATGCAGGTCTAAGAGGAAG	62	110.09	102
PRMT9	GACCTTGCAGACTACTGGATAAA	62	CATTCCAAACCCAAGACACTAATAC	62	109.11	107
GAPDH	TCGACAGTCAGCCGCATCTTCTTT	61.2	GCCCAATACGACCAAATCCGTTGA	60.9	106	196
TBP	CCACTCACAGACTCTCACAAC	61.2	CTGCGGTACAATCCCAGAACT	61.2	96	127
VTI1A	GAAGAAGCGAAAGAACTGCTTG	60	TAGGCGATCCGTGACCTTTTA	60.6	104.94	149
ACTB	AGCCATGTACGTTGCTATCCA	58	ACCGGAGTCCATCACGATG	59	98.04	120
HPRT1	ACCCTTTCCAAATCCTCAGC	65	GTTATGGCGACCCGCAG	67	102	125

Figures

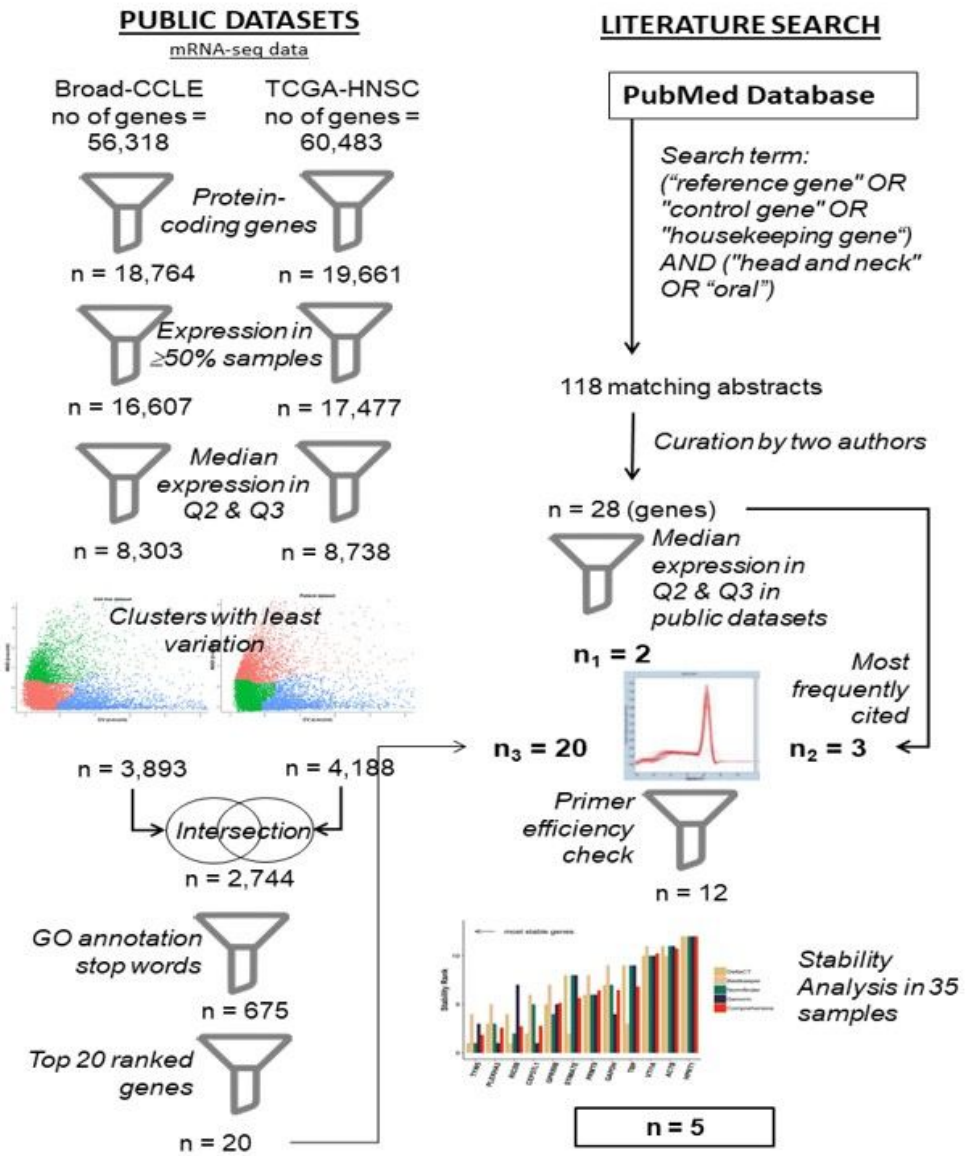
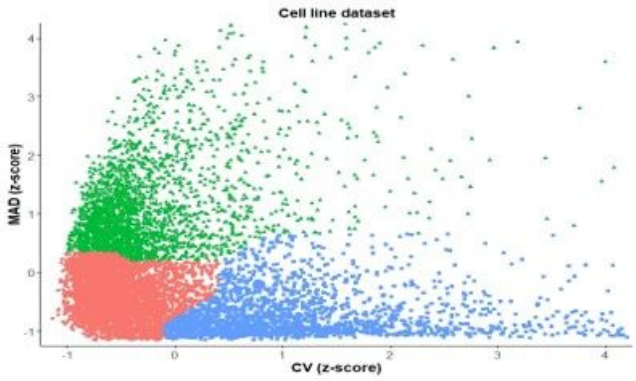


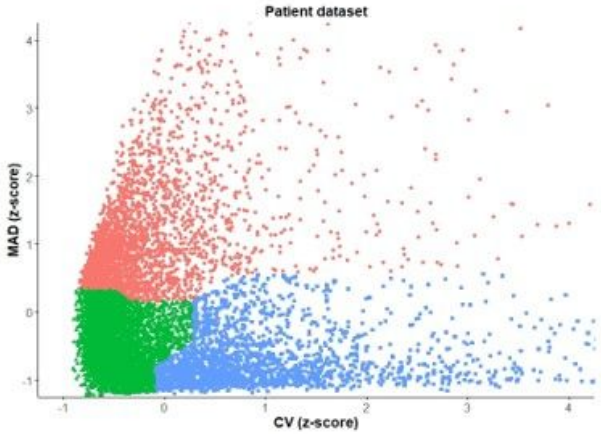
Figure 1

Work flow of the study.



(a)

	Medoid z scores	
Cluster #	CV	MAD
1	-0.31	0.92
2	**-0.50**	**-0.43**
3	0.69	-0.81



(b)

	Medoid z scores	
Cluster #	CV	MAD
1	**-0.56**	**-0.36**
2	-0.42	0.92
3	0.87	-0.87

Figure 2

Clustering analysis of cell line and patient dataset: Clustering results for (a) Broad-CCLE cell line dataset, with genes marked in pink with least values of the parameters and (b) TCGA-HNSC patient dataset with corresponding cluster marked in green.

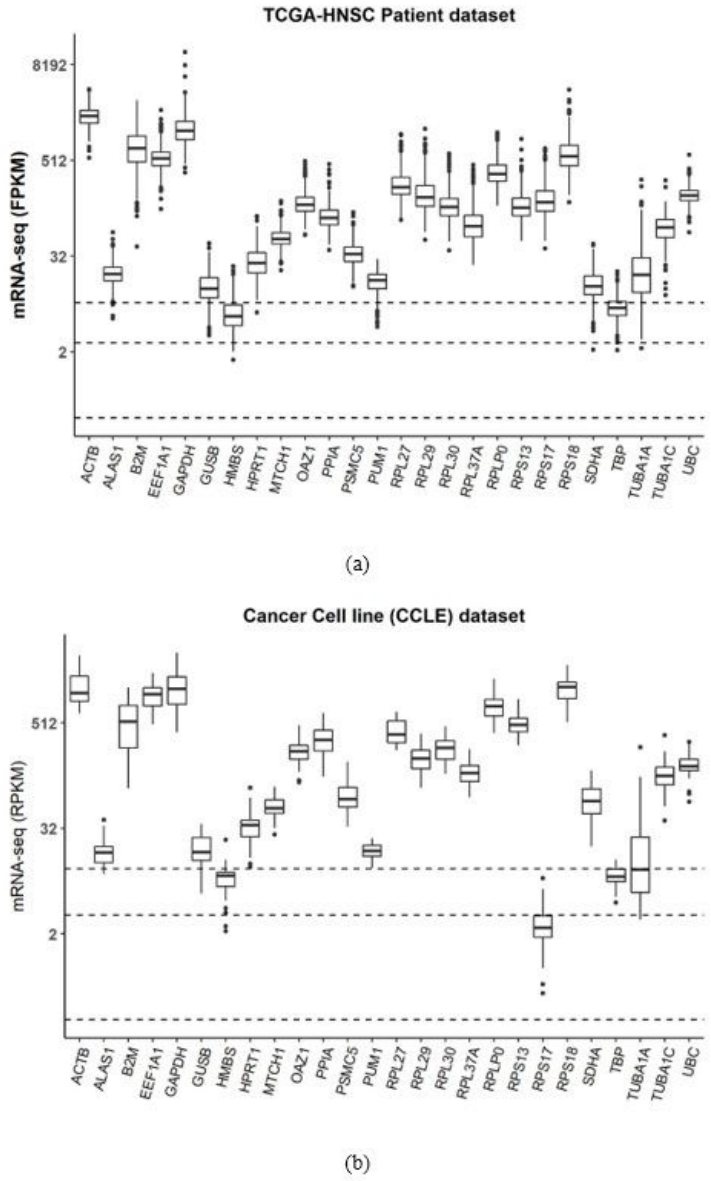


Figure 3

Expression of the commonly used reference genes in literature: Expression in (a) TCGA-HNSC patient dataset of 500 and (b) CCLE cell line dataset of 33. Dashed horizontal lines from bottom represent 25%, 50% and 75% quartiles of median expression levels of genes respectively. As seen from the figure, only HMBS and TBP genes have median expression levels within the 25-75% quartiles in both datasets.

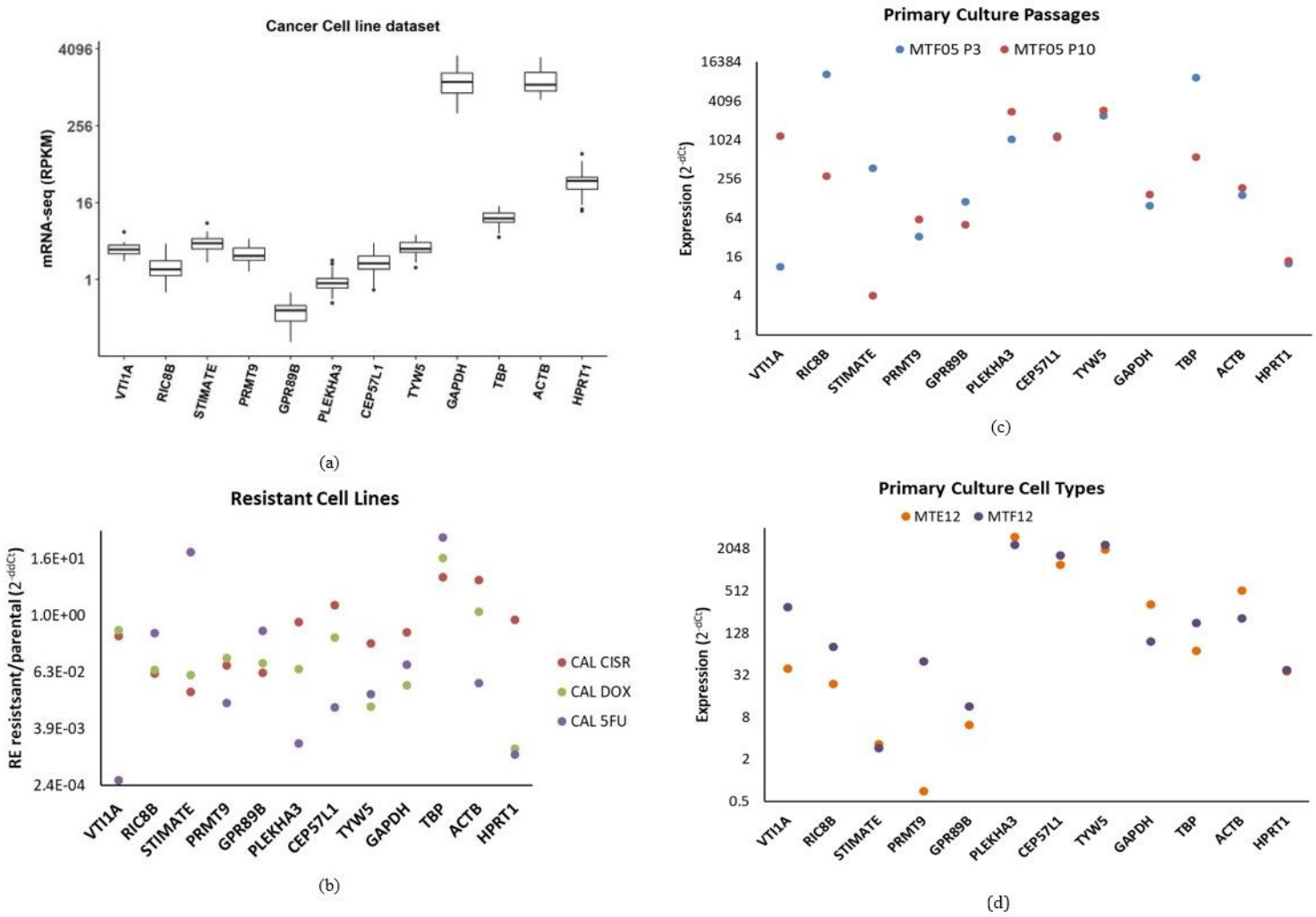


Figure 4

Expression of candidate reference genes in various platforms: (a) Expression of the candidate reference genes in CCLE cell line dataset of 33 (b) Relative expression of drug resistant CAL 27 cell line over parental (c) Expression in Primary culture passage P3 and P10 and (d) Expression in epithelial cells (MTE12) and fibroblasts (MTF12) from the same patient samples.

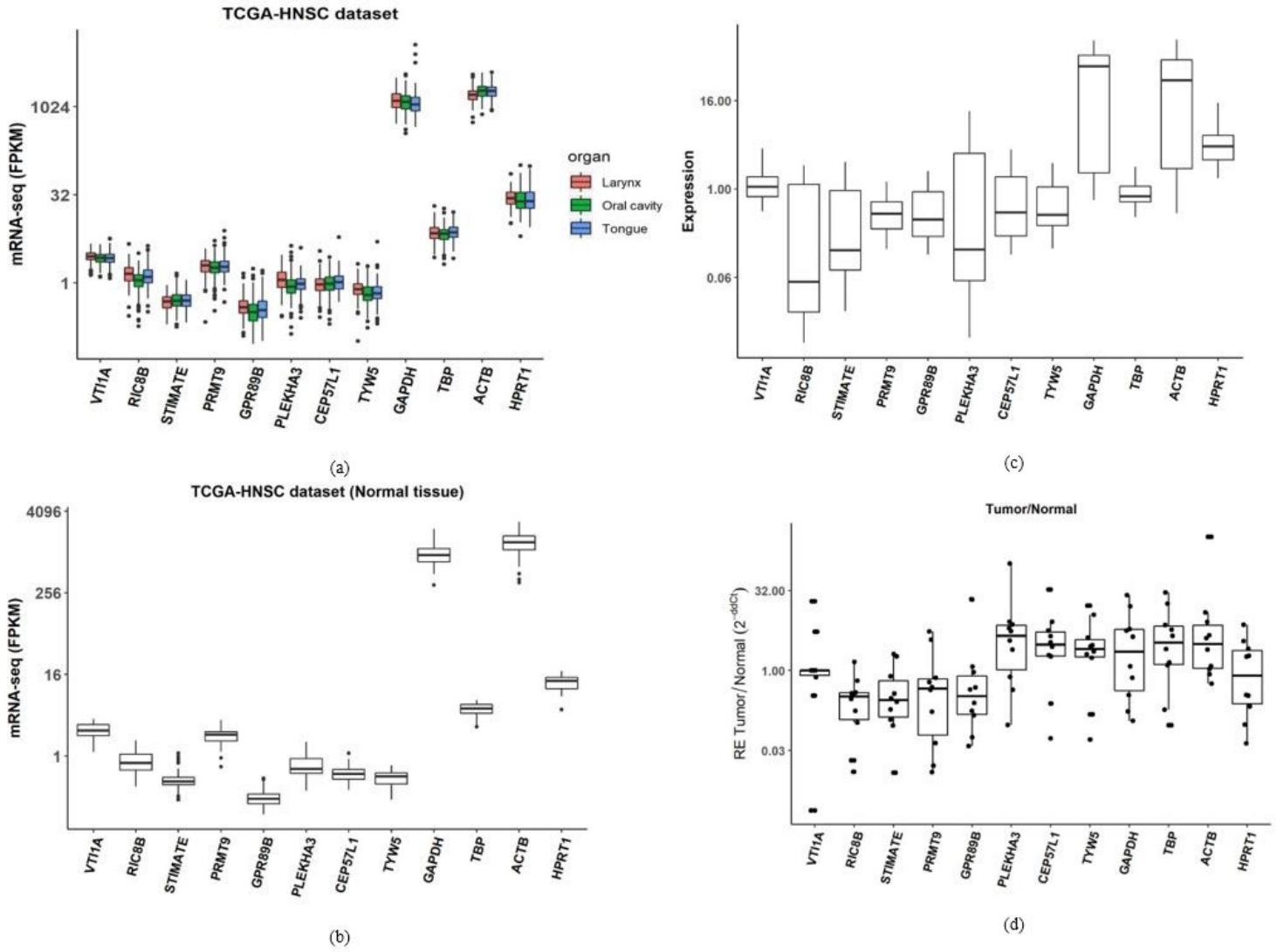


Figure 5

Expression of the candidate reference genes in patient samples: (a) TCGA tumor samples n=500 (b) normal sample of n=44 (c) GEO tumor samples of n=61 and (d) Relative expression of tumor over matched normal samples of n=10.

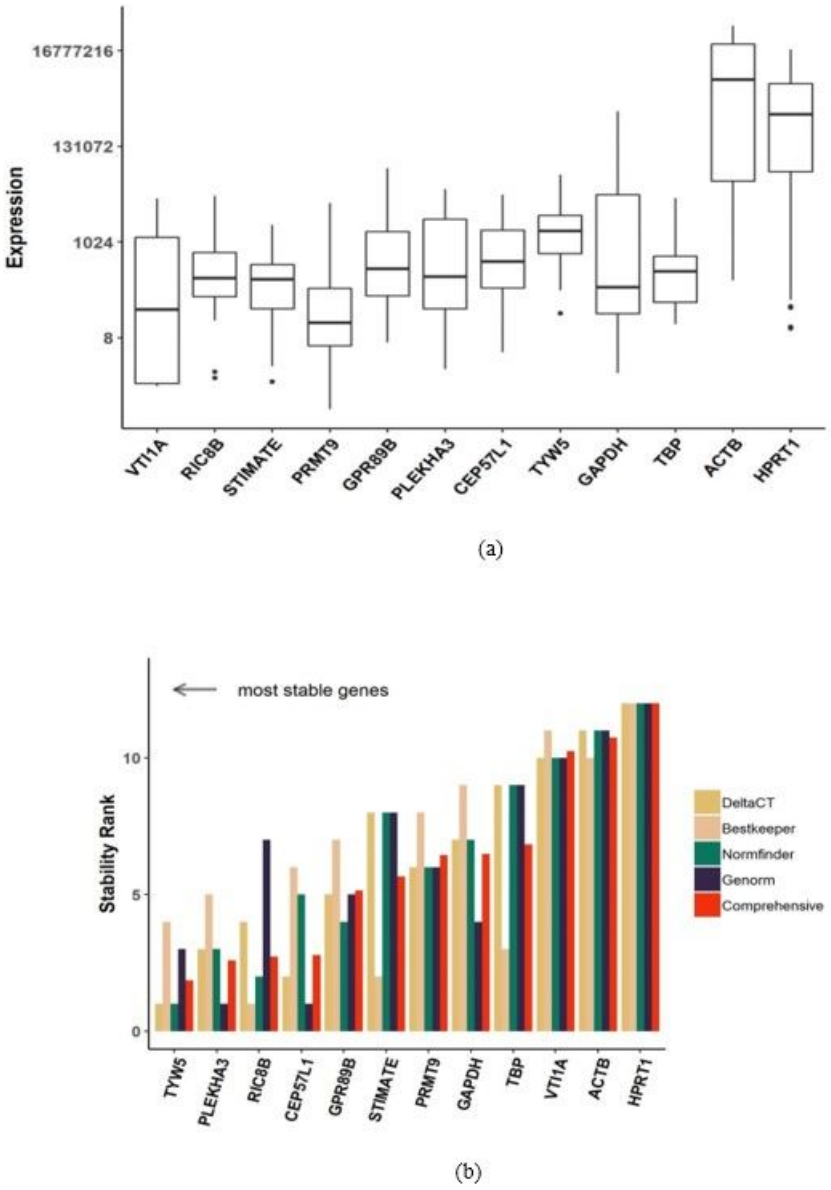


Figure 6

Stability analysis of candidate reference genes: Stability of the candidate reference genes by (a) qPCR analysis in 35 systems (b) as analysed by RefFinder.

Supplementary Files

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

- Supplementarydata.pdf