A peer-reviewed version of this preprint was published in PeerJ on 21 August 2017.

Hoang VLT, Tom LN, Quek X, Tan J, Payne EJ, Lin LL, Sinnya S, Raphael AP, Lambie D, Frazer IH, Dinger ME, Soyer HP, Prow TW. 2017. RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. PeerJ 5:e3631 https://doi.org/10.7717/peerj.3631

RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers

Van LT Hoang [1], **Lisa N Tom** [1], **Xiu-Cheng Quek** [2,3], **Jean-Marie Tan** [1], **Elizabeth Payne** [1], **Lynlee L Lin** [1], **Sudipta Sinnya** [1], **Anthony P Raphael** [1,4], **Duncan Lambie** [5], **Ian H Frazer** [6], **Marcel E Dinger** [2,3], **H. Peter Soyer** [1], **Tarl W Prow** Corresp. [1]

[1] Dermatology Research Center, School of Medicine, Translational Research Institute, Princess Alexandra Hospital, The University of Queensland, Brisbane, Queensland, Australia

[2] Garvan Institute of Medical Research, Sydney, New South Wales, Australia

[3] St Vincent's Clinical School, University of New South Wales, Sydney, New South Wales, Australia

[4] Wellman Centre for Photomedicine, Harvard Medical School, Massachusetts General Hospital, Boston, Massachusetts, United Stated of America

[5] Department of Anatomical Pathology, Princess Alexandra Hospital, Brisbane, Queensland, Australia

[6] Diamantina Institute, Translational Research Institute, Princess Alexandra Hospital, The University of Queensland, Brisbane, Queensland, Australia

Corresponding Author: Tarl W Prow
Email address: t.prow@uq.edu.au

Identification of appropriate reference genes (RGs) is critical to accurate data interpretation in quantitative real-time PCR (qPCR) experiments. In this study, we have utilised next generation RNA-sequencing (RNA-seq) to analyse the transcriptome of a panel of non-melanoma skin cancer lesions, identifying genes, which are consistently expressed across all samples. Genes encoding ribosomal proteins were amongst the most stable in this dataset. Validation of this RNA-seq data was examined using qPCR to confirm the suitability of a set of highly stable genes for use as RGs. These genes will provide a valuable resource for the normalisation of qPCR data for the analysis of non-melanoma skin cancer.

RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers

Van L T Hoang[1], Lisa N Tom[1], Xiu-Cheng Quek[2,3], Jean-Marie Tan[1], Elizabeth Payne[1], Lynlee L Lin[1], Sudipta Sinnya[1], Anthony P Raphael[1,5], Duncan Lambie[4], Ian H Frazer[6], Marcel E Dinger[2,3], H Peter Soyer[1], Tarl W Prow[1*]

[1]Dermatology Research Center, School of Medicine, The University of Queensland, Translational Research Institute, Princess Alexandra Hospital, Brisbane, QLD 4102 Australia.
[2]Garvan Institute of Medical Research, 384 Victoria Street, Sydney, NSW, 2010, Australia.
[3]St Vincent's Clinical School, University of New South Wales, Sydney, NSW 2052, Australia.
[4]Department of Anatomical Pathology, Princess Alexandra Hospital, Brisbane, QLD 4102, Australia.
[5]Wellman Centre for Photomedicine, Massachusetts General Hospital, Harvard Medical School, Boston, USA.
[6]Diamantina Institute, The University of Queensland, Translational Research Institute, Princess Alexandra Hospital, Brisbane, QLD 4102, Australia.

*Corresponding Author
Assoc Prof Tarl Prow
Translational Research Institute
37 Kent Street, Level 5, Room 5082
Woolloongabba, QLD 4102

Email: t.prow@uq.edu.au

Abstract

Identification of appropriate reference genes (RGs) is critical to accurate data interpretation in quantitative real-time PCR (qPCR) experiments. In this study, we have utilised next generation RNA-sequencing (RNA-seq) to analyse the transcriptome of a panel of non-melanoma skin cancer lesions, identifying genes, which are consistently expressed across all samples. Genes encoding ribosomal proteins were amongst the most stable in this dataset. Validation of this RNA-seq data was examined using qPCR to confirm the suitability of a set of highly stable genes for use as RGs. These genes will provide a valuable resource for the normalisation of qPCR data for the analysis of non-melanoma skin cancer.

Key words: non-melanoma skin cancer, RNA-seq, qPCR, reference gene

Introduction

There is a growing need for biomarker identification in non-melanoma skin cancer (NMSC) in order for accurate diagnosis of early skin lesions to predict progression and patient response to novel treatments. Quantitative real-time PCR (qPCR) is an integral technique for gene expression analysis in dermatology research (Li et al. 2014; Riihila et al. 2014; Wang et al. 2014; Zou et al. 2015), due to its high sensitivity and specificity. Historically, selection of reference genes (RGs) for qPCR studies has been arbitrary, with researchers commonly selecting genes such as 18S rRNA, GAPDH, and Actin without experimental validation while making the assumption that they are stably expressed across tissues. However, in many instances these commonly used RGs exhibit tissue and treatment specific variability (Chari et al. 2010; de Jonge et al. 2007).

Validation of RGs tailored for individual experimental conditions is therefore a necessity before commencement of gene expression studies (Bustin et al. 2009). Use of a RG whose expression is variable or changes as a result of treatment conditions invariably leads to inaccurate and misleading results. It is therefore strongly recommended in the MIQE guidelines (minimum information for publication of qPCR experiments) that suitable RGs be determined for individual experimental conditions. Selecting suitable RGs is not straightforward and as a result researchers are increasingly turning to transcriptome profiling data to identify genes which are suitable for their tissue of interest.

Analysis of gene expression patterns in skin lesions by whole transcriptome RNA-seq is a powerful technique for the analysis of gene expression profiles (Berger et al. 2010; Jabbari et al. 2012; Wagle et al. 2014). RNA-seq allows accurate measurement of gene expression levels with a large dynamic range of expression and high signal to noise ratio. More importantly, RNA-seq, unlike probe-based assays (such as microarrays), is able to provide an unbiased view of the transcriptome. As such, RNA-seq is an ideal strategy for identifying stably expressed genes suitable for use as qPCR RGs. The identification of stably expressed RGs in NMSC and precancerous lesions is essential to facilitate gene expression studies.

We have utilised next generation transcriptome profiling by RNA-seq on a panel of NMSC and precancerous lesions to identify a list of candidate genes which exhibit very low variability across a range of skin lesions comprising actinic keratosis (AK), intraepidermal carcinoma (IEC), squamous cell carcinoma (SCC), seborrheic keratosis (SK), basal cell carcinoma (BCC) and healthy skin. The stability of these candidate genes was validated by qPCR. Using GeNorm and Normfinder analyses, we have determined a stable combination of genes as qPCR RGs specific for skin samples. We demonstrated the importance of accurate RG selection by performing relative quantitation analysis for several targeted gene expression in healthy skin, AK and SCC lesions where normalisation were performed using either new RGs together or traditional RG GAPDH.

Materials and Methods

Patient samples

Skin lesions and healthy skin tissue samples were collected from patients at the Dermatology department in Princess Alexandra hospital. The study was approved by Metro South Human Research Ethics Committee and The University of Queensland Human Research Ethics Committee (HREC-11-QPAH-236, HREC-11-QPAH-477, HREC-12-QPAH-217, and HREC-12-QPAH-25). Written, informed consent was obtained from all patients prior to participation. Following biopsy, tissues were immersed in RNA later (Life Technologies, Carlsbad, CA) and stored at -80°C until required. All samples were sectioned and processed according to routine protocol in the Department of Anatomical Pathology located in Princess Alexandra Hospital.

RNA isolation and cDNA synthesis

RNA isolation was performed using the Qiagen RNeasy Plus Mini kit (Qiagen GmbH, Hilden, Germany). Briefly, tissue samples were cut into small pieces, and transferred into 1.5mL tube containing lysing matrix D (MP Biomedicals, Santa Ana, CA, USA) and 600uL buffer RLT containing 1% beta-mercaptoethanol and homogenised using a Fast Prep benchtop homogeniser (MP Biomedicals, Santa Ana, CA, USA). Samples were spun 5 times using setting 6.5 for 30 seconds each, and chilled on ice between spins. Lysate was removed and transferred to a fresh tube. For unfixed BCC samples embedded in OCT, 20 x 10 micron sections were cut and placed into 600 µL buffer RLT, and homogenised using a 18.5 gauge blunt needle attached to an

RNAse free syringe and resuspended 5 times. The remaining RNA extraction steps were performed as above. RNA concentration was measured using the Qubit fluorometer (Life Technologies, Carlsbad, CA) and RNA integrity determined using the 2100 Bioanalyser (Agilent Technologies, Palo Alto, CA) on RNA Pico chips. The minimum acceptable quality for RNA for analysis by qPCR was RIN >6. Complementary DNA (cDNA) was synthesised from 200 ng total RNA using the Sensifast cDNA kit (Bioline, London, UK) as per the manufacturer's instructions.

Library Preparation and RNA Sequencing

RNA-seq libraries of poly (A) RNA from 500ng total RNA obtained from AK, IEC and SCC and SK samples, were generated using the TruSeq unstranded mRNA library prep KIT for Illumina multiplexed sequencing (Illumina, San Diego, CA, USA). TruSeq stranded mRNA library prep kit was used to generate poly (A) RNA libraries for RNA obtained from normal health skin samples (Illumina, San Diego, CA, USA). Libraries were sequenced (100 bp, paired-end) on the Illumina 2000 platform and FASTQ files were analysed.

Bioinformatics pipeline

Sequencing data (~40 million reads) were checked for sequencing quality by FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adaptors and poor quality sequences were then removed using Trim Galore v0.3.7 (~6% of reads removed) (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Trimmed reads were then aligned using Tophat (V 2.1), using the unstranded protocol, against the Human Genome (hg19 build) guided with a human transcriptome generated from the GENCODE Gene annotation version 19 (23, 24). Quantification of gene expression based on counting the overlaps of mapped reads with genes annotated in the GENCODE gene annotation v19 using HTSeq (Version 0.6.1p2) (25). Read counts were then normalized and RPKM value calculated using EdgeR (26). Genes with zero read counts for all samples were discarded. The raw data can be accessed via http://skinref-dev.dingerlab.org/.

Statistical analysis for identification of RGs from RNA-Seq

The coefficient of variation (CoV) was measured by taking the standard deviation of expression value for a given gene by its mean. The maximum fold change (MFC) is the ratio of the

minimum and maximum value observed for a given gene. A pseudo count of 0.125 RPKM was added to the minimum and maximum value for the calculation of MFC.

qPCR

The primers for qPCR reactions were designed using NCBI Primer BLAST (www.ncbi.nlm.nih.gov/tools/primer-blast/) (Table 2). Primers were designed to span intron boundaries to avoid amplification of genomic DNA and to amplify all isoforms known to each gene based on the NCBI Reference Sequence Database (Refseq). Primers were synthesized by Sigma-Aldrich (Castle Hill, Australia). qPCR reactions were performed in triplicate using 1 µL diluted cDNA template in a 10 µL total volume. Reactions were performed in 384-well plate format on the ABI Viia7 Real-Time PCR system (Life Technologies, Carlsbad, CA, USA) using Sensifast SYBR Lo-rox mastermix (Bioline, London, UK). A 2-step cycling protocol was performed, comprising an initial 95 degree polymerase activation for 2 minutes, followed by 40 cycles of 95 degrees for 5 seconds, then 60 degrees for 20 seconds. The comparative Ct (ΔΔCt) method was used for data normalisation.

Measurement of novel RGs stability

The RG stability was calculated using the geNORM algorithm (Andersen et al. 2004), which is integrated into qbase+ software (Biogazelle, Gent, Belgium) and Normfinder software (Vandesompele et al. 2002) (available as an add-in for Microsoft Word at http://moma.dk/normfinder-software).

Statistical analysis of qPCR data

Statistical tests used have been described in each figure legend. Data analysis was performed using GraphPad Prism version 5.04 for Windows (GraphPad Software, Inc., La Jolla, CA, USA).

Results

Identification of novel candidate RGs

To identify RG specific for NMSC and precancerous lesions, we first performed gene expression profiling using RNASeq on 4 healthy skin samples, 12 AK, 7 IEC, SCC lesions. As previously described, a RG should show similar expression across samples, expressed at detectable levels and not display any exceptional expression in any of the samples (de Jonge et al. 2007; Eisenberg & Levanon 2013). To identify genes that fall within these criteria, we measured the mean expression, coefficient of variation (CoV) and the maximum fold change (MFC) for each gene within the dataset. The CoV measures variability within samples and the MFC is the ratio of the maximum and minimum RPKM values for a given gene (Reads Per Kilobase of transcript per Million mapped reads). Ideally, a RG candidate should have a low CoV and MFC value and is expressed at detectable levels.

An initial shortlist of the top 100 gene candidates based on the product of CoV and MFC value for each gene is shown in Supplementary Table 1. Functional annotation on our shortlist using the DAVID algorithm (http://david.abcc.ncifcrf.gov/home.jsp) revealed that most of these stably expressed genes were ribosome proteins involved in translation (enrichment score 16.98, *P* Value < $1.8e^{-30}$) (Huang da et al. 2009) (Supplementary Table 2).

To identify RGs specific for NMSC and precancerous lesions, we then shortlisted 10 candidate genes for further validation with qPCR. These 10 candidate genes were selected based on cut-off values set lower or higher than both the mean and median values of the transcriptome (log10RPKM >1, CoV < 0.4, MFC < 5). In addition, we selected only for candidates with functions well described in literature. For instance, we chose the RPLP0 gene, whose function is not only well known for different cell and tissue types but also shown to be suitable RG for research in the differentiation of human epidermal keratinocytes (16). Finally, to distinguish mRNA from genomic DNA, we selected multi-exonic genes as candidates to aid design of primers across intron boundaries.

To demonstrate the stringency and importance of our selection process, we compared the CoV, MFC and expression value of our 10 RG candidates with three commonly used qPCR RGs in

skin - ACTB, GAPDH and HRT1, (de Kok et al. 2005). Our candidate RGs had a lower CoV and MFC compared to ACTB, HPRT1 and GAPDH (Table 1 and Figure 1).

qPCR validation of new RGs

In order to validate and extend our findings from the RNASeq analysis, we conducted qPCR on our 10 candidate RGs in addition to ACTB, GAPDH, and HPRT1 on samples derived from a diversity of skin conditions within the same disease spectrum. A total of 24 samples were tested comprising of AK (n=4), SCC (n=3), SK (n=3), BCC (n=4), IEC (n=5), and healthy skin (n=5). Results from the qPCR were analysed using GeNorm (Vandesompele et al. 2002) within Qbase+ software (Biogazelle) and Normfinder to determine the consistency of expression values among the samples for each candidate gene (Andersen et al. 2004).

Statistically, GeNorm conduct pairwise variation (V) analysis to identify genes with the least variance between samples and is denoted as the 'stability' (M) values. In general, lower M values indicate lower variance in expression value among samples and genes with M values $\leq$ 0.5 are associated with homogeneous samples. Remarkably, all of our 10 RG candidates had M values ≤ 0.5 with RPLP0, RPL7A, RPL23, RPS27A and RPL38 ranked in the top 5 genes for GeNorm M value/Stability value. In addition, to eliminate errors related to the usage of a single housekeeping gene, it is common practise to use two or more housekeeping genes. Using GeNorm, by calculating the normalization factor based on the geometric mean of multiple control genes, we identified that we need only two of our RG candidates for accurate normalization (GeNorm V, V2/3 = 0.084). V values of < 0.15 indicate acceptable stability of the RG combination, indicating no further need for additional RGs. Amongst our RG candidates, the pair of genes with optimal normalization factor was RPL38 and RPS27A, which demonstrated the lowest M values (0.257 and 0.265 respectively) (Figure 2a).

In addition, Normfinder analysis was performed for the same dataset (Andersen et al. 2004). Normfinder analysis performs estimation of both intra- and intergroup expression variation for each subgroup of samples (lesion types), with output given as a Stability Value. The most stable candidate was RPL7A, and the best combination of genes was RPL7A and RPLP0 (Figure 2b). Overall trends between GeNorm and Normfinder analyses were similar. In both formats, the

traditional RGs ACTB, GAPDH and HPRT1 were ranked as having the most variability in gene expression across the groups (increased stability values), and the genes RPLP0, RPL7A, RPL23, RPS27A and RPL38 ranked in the top 5 genes for GeNorm M value/Stability value.

To demonstrate the significance of our findings in NMSC research, we investigated the difference in expression of keratin 17 (KRT17) in AK between normalization using our candidate RGs and normalization with GAPHD (Figure 3). When using GAPDH as calibrator, there was an approximate 2-fold increase in levels of KTR17 when comparing healthy skin to AK (Figure 3a) or an approximate 3-fold increase for SCC (Figure 3b). However, the fold change was significantly higher at approximately 7-fold for AK and 12-fold for SCC, when using the combination of RPL32 and RPS27A or either one of them ($P<0.05$). There was no statistically significant difference between data normalised with RPLP0, RPS7A or a combination of the two. Overall, these results demonstrate that use of a RG that is not stably expressed can lead to inaccurate data, particularly in instances where the relative fold change is subtle.

Discussion

The selection of appropriate RGs is of critical importance to accurately quantify gene expression levels using qPCR. Our results concur with previous studies reporting that RNA-seq is an effective method for the identification of stably expressed transcripts for application in qPCR. Through qPCR validation, we demonstrate that transcriptome analysis by RNA-seq is a reliable strategy for identification of genes with low variability. To the best of our knowledge, this is the first study to identify suitable RGs for use in studies of pre-cancerous lesions and NMSC. Our data demonstrate that the RG candidates selected for validation are stably expressed in these lesions, showing strong stability in gene expression between different types of skin cancer lesion and healthy skin. Results suggest that our RNA-seq dataset is a valuable resource to assemble a shortlist of candidates for validation by qPCR prior to commencement of gene expression studies in NMSC and sun-damaged skin. Our RNA-seq data identified many RPL and RPS genes, which encode structural proteins associated with ribosome biosynthesis, as highly stable. This finding is in agreement with previous studies demonstrating RPL genes as some of the least variable across a wide range of cell and tissue types. In a meta-analysis of over 13,000 human gene arrays, 13 of

the top 15 genes identified were ribosomal structural proteins (de Jonge et al. 2007). The need for stability in this group of genes is logical given that ribosome biogenesis is a tightly regulated process that is critical for fundamental cellular functions including cell growth and division.

We evaluated the stability of ten potential RGs in various NMSC samples. Results showed small differences in the recommended RG combinations between GeNorm and Normfinder analysis outputs, but the overall stability of our candidate genes was shown to be consistent in both analyses (Figure 2). This effect is likely due to the way these algorithms are designed, each utilising a different method to determine the most stable gene combination. In the case of GeNorm, the algorithm uses pairwise correlation to determine stability, using the assumption that genes showing similar expression patterns are likely to also reflect mRNA (cDNA) levels. BestKeeper is another commonly used normalisation algorithm that is based on pairwise correlation (Pfaffl et al. 2004). A limitation of this type of normalisation process is genes that demonstrate co-ordinate regulation are likely to be ranked highly, even if they are not truly stable. Normfinder is an alternative algorithm that uses a mathematical model-based approach, which allows estimation of both intra- and intergroup expression variation to calculate a stability value. Due to this variability, it is a wise strategy to use more than one algorithm to confirm the most appropriate RG. In our case, there is a very small variation between the highest ranking candidates for both analysis methods. In general, any of these top ranked genes RPL38, RPL23, RPS27A, RPL7A and RPLP0 are a suitable RG for use in NMSC and precancerous lesions. By contrast, GAPDH or ACTB, which are widely used RGs are not suitable in this type of cancer as their expression is significantly different in healthy skin and different type of NMSC. This finding is in line with a recent study recommending to not use GAPDH for normalization purposes when analysing RNA expression in human keratinocytes (Beer et al. 2015).

To observe the impact of RG stability on the relative quantitation analysis, we characterized the levels of the keratin KTR17 in healthy skin, SCC, and AK lesions using either GAPDH or our most stable combination as determined by Normfinder analysis, RPS7A and RPLP0. KRT17 together with KRT16 and KRT6 are involved in keratinocyte differentiation and skin cancer (Hameetman et al. 2013). It was previously reported that there is an upregulation of intermediate filament keratins in SCC lesions compared to healthy skin (Hameetman et al. 2013; Hudson et al.

2010). In this study, we found that the comparison result was significantly altered using a different calibrator. The upregulation of KRT17 in AK and SCC lesions was even more pronounced with our candidate RGs. These results demonstrate that choosing an inappropriate RG may lead to wrong conclusions being drawn from qPCR results.

It should however be noted that despite the high stability of our candidate RGs across a range of different skin lesions, these lesions were not exposed to any treatments such as topically applied medications, which could potentially affect their expression. A literature search should be performed prior to the commencement of the study to eliminate RGs that will potentially be affected by planned clinical treatment conditions. As it is unlikely that any gene is stable across all possible experimental conditions, validation should be performed for each treatment, and in general, two or more RGs should be used to reduce the impact of any variability. For our subset of validated RGs, many are genes encoding ribosomal structural proteins. Caution should therefore be used if considering these RGs where treatment conditions which have been demonstrated to result in nucleolar stress (Nosrati et al. 2015). In this instance, selection and validation of genes with a different functional classification, such as EEF1A1 or EEF1B2, or derived from our shortlist of 100 highly stable genes would be a logical strategy.

Conclusions

In this study, we utilised whole transcriptome RNA-seq to analyse healthy skin, precancerous and lesional NMSC for the purpose of identifying reference genes, which are consistently expressed across all samples. To identify genes that fall within these criteria, we measured the mean expression, coefficient of variation and the maximum fold change for each gene within the dataset. This resulted in the identification of 100 highly stable genes. To further refine the genes specific for precancerous and NMSC lesions, we then shortlisted 10 candidate genes for further validation with qPCR. These 10 candidate genes were selected based on cut-off values set lower or higher than both the mean and median values of the transcriptome. We determined that the genes RPL38, RPL23, RPS27A, RPL7A and RPLP0, which encode structural proteins associated with ribosome biosynthesis are the most suitable reference genes for use in NMSC and precancerous lesions.

Overall, we demonstrate that transcriptome analysis by RNA seq is a reliable strategy for identification of genes with low variability. Our results concur with previous studies reporting that RNA-seq is an effective method for the identification of stably expressed transcripts for application in qPCR. To the best of our knowledge, this is the first study to identify suitable reference genes for use in studies of pre-cancerous lesions and NMSC. These genes will provide a valuable resource for the normalisation of qPCR data for the analysis of non-melanoma skin cancer.

Acknowledgements

We would like to acknowledge the cooperation and coordination between all the members involved within this multi-center study including the Dermatology Research Center, Garvan Institute of Medical Research, Department of Anatomical Pathology at the Princess Alexandra Hospital, the Diamantina Institute and assistance form those in the Institute's sequencing facility.

References

Andersen CL, Jensen JL, and Orntoft TF. 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* 64:5245-5250. 10.1158/0008-5472.CAN-04-0496

Beer L, Mlitz V, Gschwandtner M, Berger T, Narzt MS, Gruber F, Brunner PM, Tschachler E, and Mildner M. 2015. Bioinformatics approach for choosing the correct reference genes when studying gene expression in human keratinocytes. *Experimental Dermatology* 24:742-747. 10.1111/exd.12759

Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C, Onofrio RC, Ziaugra L, Cibulskis K, Laine E, Barretina J, Winckler W, Fisher DE, Getz G, Meyerson M, Jaffe DB, Gabriel SB, Lander ES, Dummer R, Gnirke A, Nusbaum C, and Garraway LA. 2010. Integrative analysis of the melanoma transcriptome. *Genome Res* 20:413-427. 10.1101/gr.103697.109

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, and Wittwer CT. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* 55:611-622. 10.1373/clinchem.2008.112797

Chari R, Lonergan KM, Pikor LA, Coe BP, Zhu CQ, Chan TH, MacAulay CE, Tsao MS, Lam S, Ng RT, and Lam WL. 2010. A sequence-based approach to identify reference genes for gene expression analysis. *BMC Med Genomics* 3:32. 10.1186/1755-8794-3-32

de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ, and ter Elst A. 2007. Evidence based selection of housekeeping genes. *PLoS One* 2:e898. 10.1371/journal.pone.0000898

de Kok JB, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, Feuth T, Swinkels DW, and Span PN. 2005. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab Invest* 85:154-159. 10.1038/labinvest.3700208

Eisenberg E, and Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* 29:569-574. 10.1016/j.tig.2013.05.010

Hameetman L, Commandeur S, Bavinck JNB, Wisgerhof HC, de Gruijl FR, Willemze R, Mullenders L, Tensen CP, and Vrieling H. 2013. Molecular profiling of cutaneous squamous cell carcinomas and actinic keratoses from organ transplant recipients. *Bmc Cancer* 13. Artn 58

10.1186/1471-2407-13-58

Huang da W, Sherman BT, and Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37:1-13. 10.1093/nar/gkn923

Hudson LG, Gale JM, Padilla RS, Pickett G, Alexander BE, Wang J, and Kusewitt DF. 2010. Microarray Analysis of Cutaneous Squamous Cell Carcinomas Reveals Enhanced Expression of Epidermal Differentiation Complex Genes. *Molecular Carcinogenesis* 49:619-629. 10.1002/mc.20636

Jabbari A, Suarez-Farinas M, Dewell S, and Krueger JG. 2012. Transcriptional profiling of psoriasis using RNA-seq reveals previously unidentified differentially expressed genes. *J Invest Dermatol* 132:246-249. 10.1038/jid.2011.267

Li Y, Man X, You L, Xiang Q, Li H, Xu B, Chen Z, Zhang X, and Lian S. 2014. Downregulation of PTEN expression in psoriatic lesions. *Int J Dermatol* 53:855-860. 10.1111/ijd.12061

Nosrati N, Kapoor NR, and Kumar V. 2015. DNA damage stress induces the expression of Ribosomal Protein S27a gene in a p53-dependent manner. *Gene* 559:44-51. 10.1016/j.gene.2015.01.014

Pfaffl MW, Tichopad A, Prgomet C, and Neuvians TP. 2004. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. *Biotechnol Lett* 26:509-515.

Riihila PM, Nissinen LM, Ala-aho R, Kallajoki M, Grenman R, Meri S, Peltonen S, Peltonen J, and Kahari VM. 2014. Complement factor H: a biomarker for progression of cutaneous squamous cell carcinoma. *J Invest Dermatol* 134:498-506. 10.1038/jid.2013.346

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

Wagle N, Van Allen EM, Treacy DJ, Frederick DT, Cooper ZA, Taylor-Weiner A, Rosenberg M, Goetz EM, Sullivan RJ, Farlow DN, Friedrich DC, Anderka K, Perrin D, Johannessen CM, McKenna A, Cibulskis K, Kryukov G, Hodis E, Lawrence DP, Fisher S, Getz G, Gabriel SB, Carter SL, Flaherty KT, Wargo JA, and Garraway LA. 2014. MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. *Cancer Discov* 4:61-68. 10.1158/2159-8290.CD-13-0631

Wang F, Smith NR, Tran BA, Kang S, Voorhees JJ, and Fisher GJ. 2014. Dermal damage promoted by repeated low-level UV-A1 exposure despite tanning response in human skin. *JAMA Dermatol* 150:401-406. 10.1001/jamadermatol.2013.8417

Zou XY, Ding D, Zhan N, Liu XM, Pan C, and Xia YM. 2015. Glyoxalase I is differentially expressed in cutaneous neoplasms and contributes to the progression of squamous cell carcinoma. *J Invest Dermatol* 135:589-598. 10.1038/jid.2014.377

Table 1(on next page)

RNA-seq scoring of selected candidate reference genes and commonly used reference genes

RNA seq scoring of selected candidate reference genes and commonly used reference genes, ranked on CoV (coefficient of variation) score. Std = standard deviation, mean = mean expression value, MFC = maximum fold change. Candidates are ranked from the smallest to largest CoV values.

Table 1. RNA seq scoring of selected candidate reference genes and commonly used reference genes, ranked on CoV (coefficient of variation) score. Std = standard deviation, mean = mean expression value, MFC = maximum fold change. Candidates are ranked from the smallest to largest CoV values.

HGNC Symbol	**CoV**	**Std**	**Mean**	**MFC**
RPL9	0.291303	0.628429	2.157305	3.381323
RPL38	0.305374	0.182736	0.598401	2.999141
RPL11	0.312295	0.826112	2.645291	3.169891
RPL23	0.313737	0.49993	1.593471	3.682954
EEF1B2	0.324824	0.715224	2.20188	3.358338
RPS27A	0.3328	1.098735	3.301484	3.443064
RPL7A	0.33827	1.645695	4.865026	3.449708
RPS13	0.341127	0.695341	2.038367	3.177869
EEF1A1	0.347487	4.683542	13.47833	3.466657
RPLP0	0.385341	2.997923	7.779916	4.281834
GAPDH	*0.611902*	*2.476736*	*4.047605*	*10.97791*
HPRT1	*0.648895*	*0.004201*	*0.006474*	*26.89796*
ACTB	*0.766221*	*0.940382*	*1.227298*	*37.8592*

Figure 1

RNA-seq analysis of genes and candidate reference genes

A) Scatterplot comparing Coefficient of variation (CoV) values against mean expression values (log2) for all genes within the RNA seq dataset. Transcripts with RPKM (reads per kilobase of exon per million fragments mapped) <1 were removed (low read counts). A total of 98,756 different transcripts were measured. Each gene is represented by a single dot. Genes selected for validation to function as reference genes in non-melanoma skin cancers (NMSC) and precancerous lesions are shown in Black. Reference genes commonly used in the literature, ACTB (Red), GAPDH (Blue) and HPRT1 (Green), are also highlighted for comparisons. B) Maximum fold change score between candidate reference genes and traditional reference genes. C) Top 10 results (sorted by PValue) for GO Biological Process Term enrichment analysis conducted on first 100 genes ranked by product of their MFC and CoV score. D) Boxplot showing expression value from RNASeq experiment of 29 skin lesions of selected reference genes candidate (blue) with commonly used housekeeping genes ACTB, GAPHD and HRT1 (red). Samples with outliers value (expression < 95% interquartile range or < 5%) represented as dots.

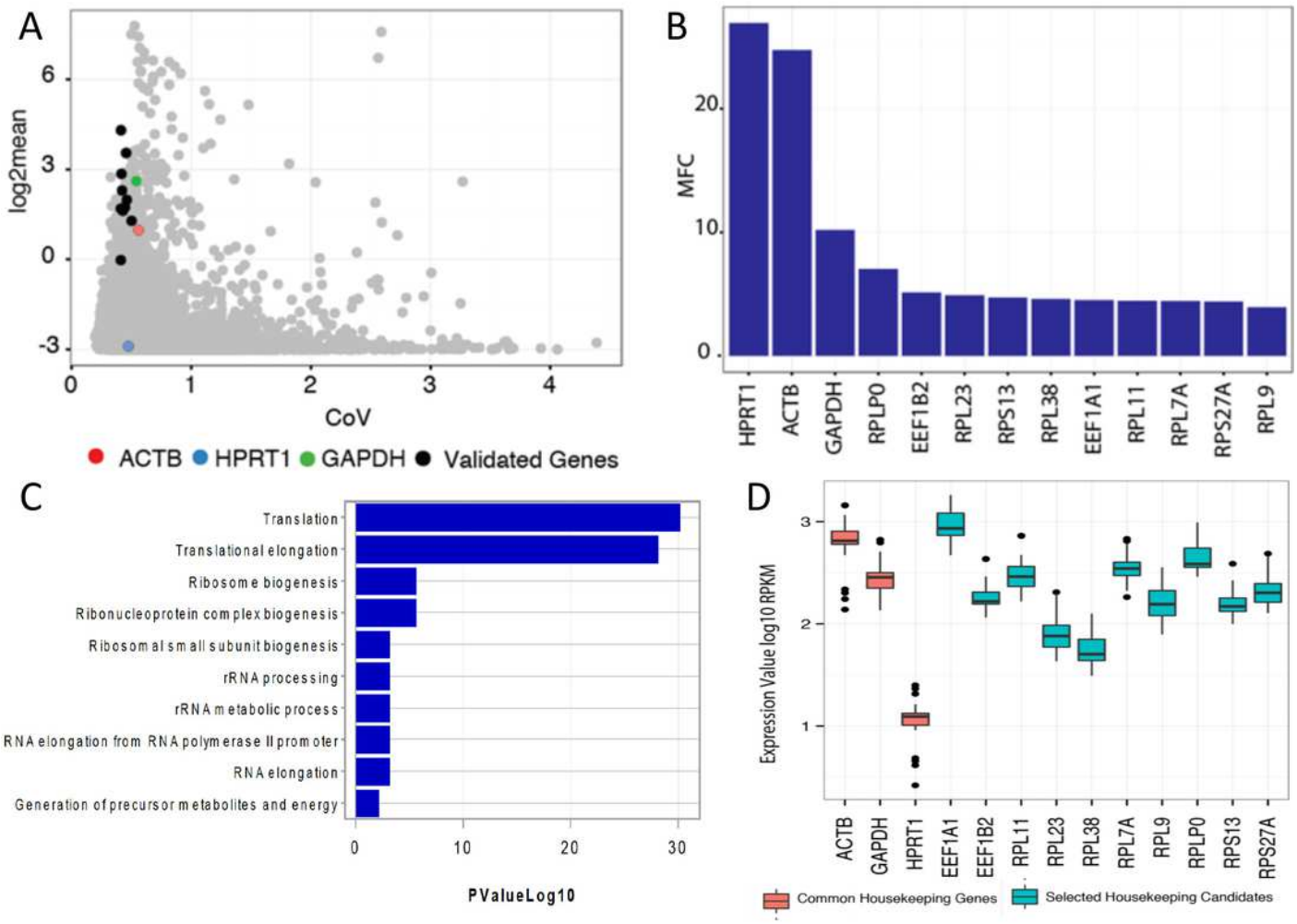

A
log2mean
CoV
ACTB HPRT1 GAPDH Validated Genes
B
MFC
HPRT1 ACTB GAPDH RPLP0 EEF1B2 RPL23 RPS13 RPL38 EEF1A1 RPL11 RPL7A RPS27A RPL9
C
Translation
Translational elongation
Ribosome biogenesis
Ribonucleoprotein complex biogenesis
Ribosomal small subunit biogenesis
rRNA processing
rRNA metabolic process
RNA elongation from RNA polymerase II promoter
RNA elongation
Generation of precursor metabolites and energy
PValueLog10
D
Expression Value log10 RPKM
ACTB GAPDH HPRT1 EEF1A1 EEF1B2 RPL11 RPL23 RPL38 RPL7A RPL9 RPLP0 RPS13 RPS27A
Common Housekeeping Genes
Selected Housekeeping Candidates

Table 2(on next page)

qPCR primers

1 **Table 2.** qPCR primers

Gene	Accession number	Forward primer	Reverse primer	Amplicon size (bp)
RPL9	NM_000661.4	CTGCGTCTACTGCGAGAATGA	CACGATAACTGTGCGTCCCT	98
RPL38	NM_000999.3	GCCATGCCTCGGAAAATTG	CCAGGGTGTAAAGGTATCTGC	139
RPL11	NM_000975.3	AGAAGGGTCTAAAGGTGCGG	AGTCCAGGCCGTAGATACCA	138
RPL23	NM_000978.3	TCCAGCAGTGGTCATTCGAC	GCAGAACCTTTCATCTCGCC	117
EEF1B2	NM_001959.3	AGTATTTGAAGCCGTGTCCAG	ACATCGGCAGGACCATATTTG	144
RPS27A	NM_002954.5	ACCACTCCCAAGAAGAATAAGC	ACTTGCCATAAACACCCCAG	147
RPL7A	NM_000972.2	GGCATTGGACAGGACATCCA	AGGCACTTTCAGCCGCTTAT	114
RPS13	NM_001017.2	TCCCCACTTGGTTGAAGTTGA	AGGAGTAAGGCCCTTCTTGG	77
EEF1A1	NM_001402.5	GAAAGCTGAGCGTGAACGTG	AGTCAGCCTGAGATGTCCCT	143
RPLP0	NM_001002.3	ATCAACGGGTACAAACGAGTC	CAGATGGATCAGCCAAGAAGG	97
GAPDH	NM_002046.5	CCCACTCCTCCACCTTTGAC	TTCCTCTTGTGCTCTTGCTG	180
HPRT1	NM_000194.2	TGCTGAGGATTTGGAAAGGG	ACAGAGGGCTACAATGTGATG	115
ACTB	NM_001101.3	ACCTTCTACAATGAGCTGCG	CCTGGATAGCAACGTACATGG	148

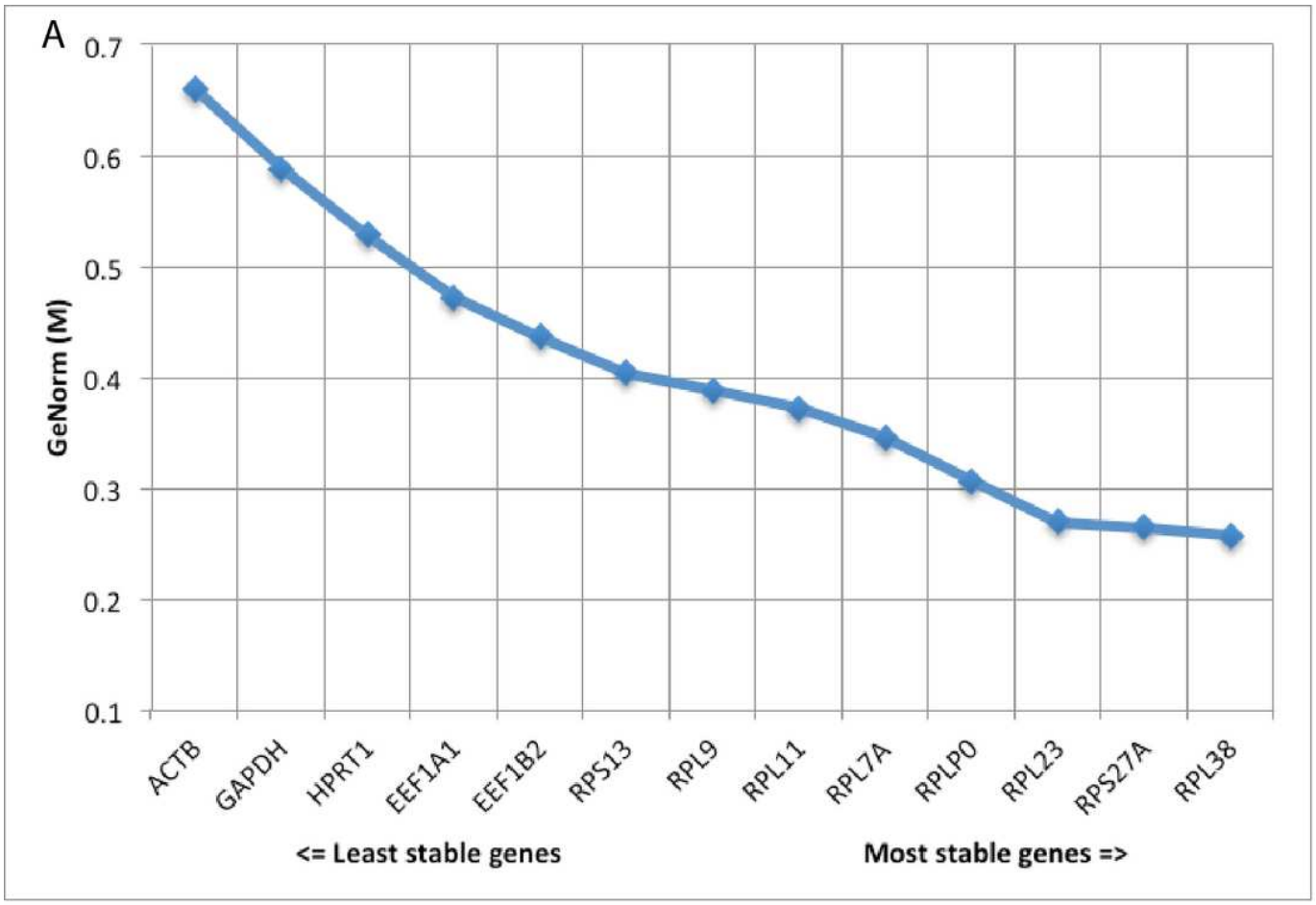
2

Figure 2

Comparison of expression stability using GeNorm and Normfinder

A) Average expression stability of reference targets (GeNorm). GeNorm M value, an indicator of gene expression stability, was determined using the GeNorm algorithm. Decreasing values correlate with smaller variations in gene expression levels across lesion groups AK, SCC, SK, BCC, IEC, and healthy skin. B) Average expression stability of reference targets (Normfinder). Stability values were determined for each gene using the Normfinder algorithm. Decreasing values correlate with smaller variations in gene expression levels across lesion groups AK, SCC, SK, BCC, IEC, and healthy skin.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2331v1 | CC BY 4.0 Open Access | rec: 2 Aug 2016, publ: 2 Aug 2016


A
GeNorm (M)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
ACTB
GAPDH
HPRT1
EEF1A1
EEF1B2
RPS13
RPL9
RPL11
RPL7A
RPLP0
RPL23
RPS27A
RPL38
<= Least stable genes
Most stable genes =>

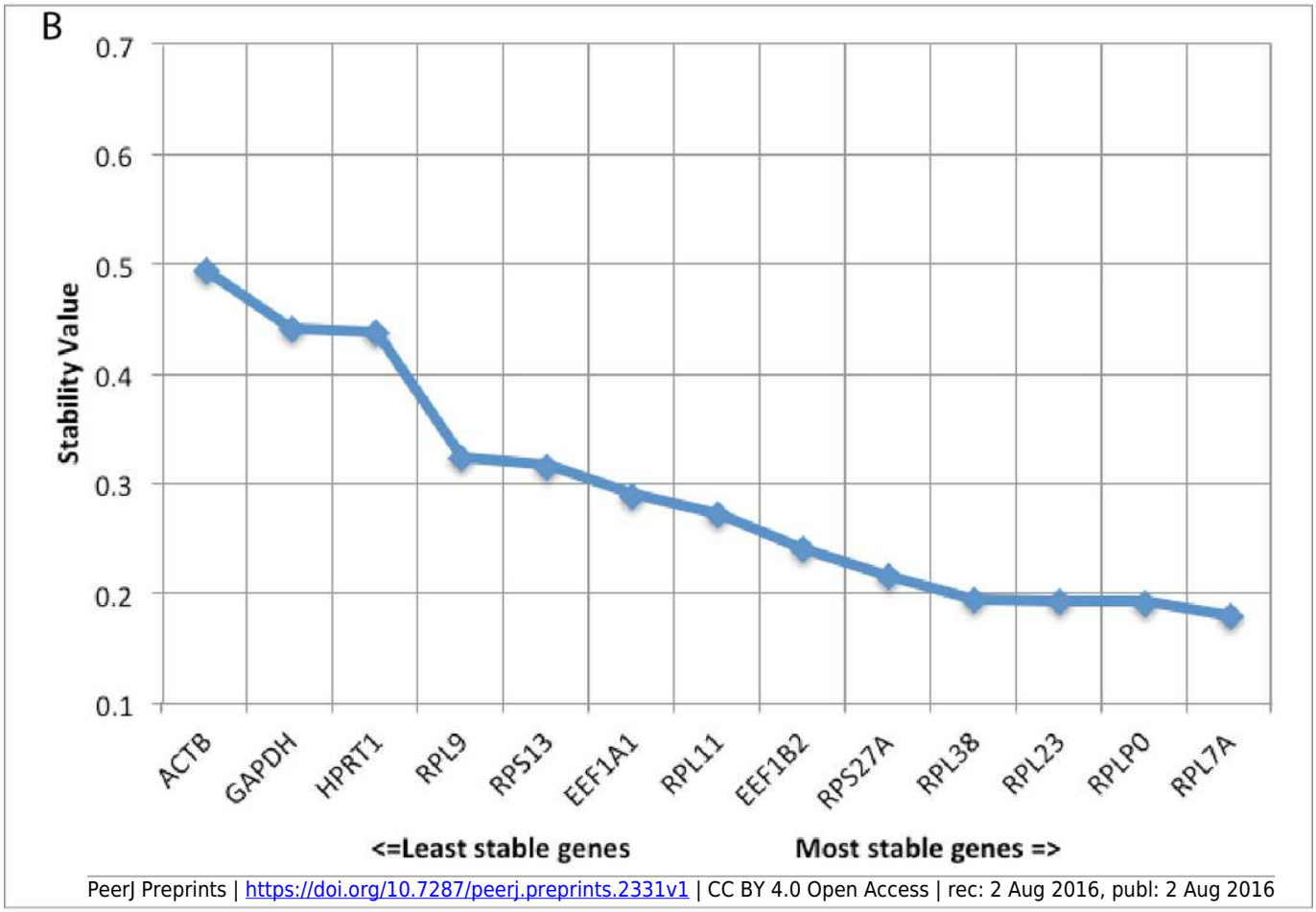

B
Stability Value
0.7
0.6
0.5
0.4
0.3
0.2
0.1
ACTB
GAPDH
HPRT1
RPL9
RPS13
EEF1A1
RPL11
EEF1B2
RPS27A
RPL38
RPL23
RPLP0
RPL7A
<=Least stable genes
Most stable genes =>

Figure 3

KRT17 levels in precancerous and lesional NMSC

Comparison of relative quantitation analysis of KRT17 levels in AK (a) and SCC (b) lesions using either RPS7A/RPLP0 or GAPDH as the reference gene relative to healthy skin. Data are presented as mean ± SEM, n = 3, * indicates $P < 0.05$; one-way ANOVA and Turkey post-test.