

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

[View the peer-reviewed version](https://doi.org/10.7717/peerj.3631) (peerj.com/articles/3631), which is the preferred citable publication unless you specifically need to cite this preprint.

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

<sup>1</sup> Dermatology Research Center, School of Medicine, Translational Research Institute, Princess Alexandra Hospital, The University of Queensland, Brisbane, Queensland, Australia

<sup>2</sup> Garvan Institute of Medical Research, Sydney, New South Wales, Australia

<sup>3</sup> St Vincent's Clinical School, University of New South Wales, Sydney, New South Wales, Australia

<sup>4</sup> Wellman Centre for Photomedicine, Harvard Medical School, Massachusetts General Hospital, Boston, Massachusetts, United States of America

<sup>5</sup> Department of Anatomical Pathology, Princess Alexandra Hospital, Brisbane, Queensland, Australia

<sup>6</sup> 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.

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

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

6 <sup>1</sup>Dermatology Research Center, School of Medicine, The University of Queensland,  
7 Translational Research Institute, Princess Alexandra Hospital, Brisbane, QLD 4102 Australia.

8 <sup>2</sup>Garvan Institute of Medical Research, 384 Victoria Street, Sydney, NSW, 2010, Australia.

9 <sup>3</sup>St Vincent's Clinical School, University of New South Wales, Sydney, NSW 2052, Australia.

10 <sup>4</sup>Department of Anatomical Pathology, Princess Alexandra Hospital, Brisbane, QLD 4102,  
11 Australia.

12 <sup>5</sup>Wellman Centre for Photomedicine, Massachusetts General Hospital, Harvard Medical School,  
13 Boston, USA.

14 <sup>6</sup>Diamantina Institute, The University of Queensland, Translational Research Institute, Princess  
15 Alexandra Hospital, Brisbane, QLD 4102, Australia.

16

17 \*Corresponding Author

18 Assoc Prof Tarl Prow

19 Translational Research Institute

20 37 Kent Street, Level 5, Room 5082

21 Woolloongabba, QLD 4102

22 Email: [t.prow@uq.edu.au](mailto:t.prow@uq.edu.au)

23

24 **Abstract**

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

33

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

35 **Introduction**

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

46

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

55

56 Analysis of gene expression patterns in skin lesions by whole transcriptome RNA-seq is a  
57 powerful technique for the analysis of gene expression profiles (Berger et al. 2010; Jabbari et al.  
58 2012; Wagle et al. 2014). RNA-seq allows accurate measurement of gene expression levels with  
59 a large dynamic range of expression and high signal to noise ratio. More importantly, RNA-seq,  
60 unlike probe-based assays (such as microarrays), is able to provide an unbiased view of the  
61 transcriptome. As such, RNA-seq is an ideal strategy for identifying stably expressed genes  
62 suitable for use as qPCR RGs. The identification of stably expressed RGs in NMSC and  
63 precancerous lesions is essential to facilitate gene expression studies.

64

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

75

## 76 **Materials and Methods**

### 77 **Patient samples**

78 Skin lesions and healthy skin tissue samples were collected from patients at the Dermatology  
79 department in Princess Alexandra hospital. The study was approved by Metro South Human  
80 Research Ethics Committee and The University of Queensland Human Research Ethics  
81 Committee (HREC-11-QPAH-236, HREC-11-QPAH-477, HREC-12-QPAH-217, and HREC-  
82 12-QPAH-25). Written, informed consent was obtained from all patients prior to participation.  
83 Following biopsy, tissues were immersed in RNA later (Life Technologies, Carlsbad, CA) and  
84 stored at -80°C until required. All samples were sectioned and processed according to routine  
85 protocol in the Department of Anatomical Pathology located in Princess Alexandra Hospital.

86

### 87 **RNA isolation and cDNA synthesis**

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

96 RNase free syringe and resuspended 5 times. The remaining RNA extraction steps were  
97 performed as above. RNA concentration was measured using the Qubit fluorometer (Life  
98 Technologies, Carlsbad, CA) and RNA integrity determined using the 2100 Bioanalyser (Agilent  
99 Technologies, Palo Alto, CA) on RNA Pico chips. The minimum acceptable quality for RNA for  
100 analysis by qPCR was RIN >6. Complementary DNA (cDNA) was synthesised from 200 ng total  
101 RNA using the Sensifast cDNA kit (Bioline, London, UK) as per the manufacturer's instructions.  
102

### 103 **Library Preparation and RNA Sequencing**

104 RNA-seq libraries of poly (A) RNA from 500ng total RNA obtained from AK, IEC and SCC and  
105 SK samples, were generated using the TruSeq unstranded mRNA library prep KIT for Illumina  
106 multiplexed sequencing (Illumina, San Diego, CA, USA). TruSeq stranded mRNA library prep  
107 kit was used to generate poly (A) RNA libraries for RNA obtained from normal health skin  
108 samples (Illumina, San Diego, CA, USA). Libraries were sequenced (100 bp, paired-end) on the  
109 Illumina 2000 platform and FASTQ files were analysed.  
110

### 111 **Bioinformatics pipeline**

112 Sequencing data (~40 million reads) were checked for sequencing quality by FASTQC  
113 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Adaptors and poor quality  
114 sequences were then removed using Trim Galore v0.3.7 (~6% of reads removed)  
115 ([http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)). Trimmed reads were then  
116 aligned using Tophat (V 2.1), using the unstranded protocol, against the Human Genome (hg19  
117 build) guided with a human transcriptome generated from the GENCODE Gene annotation  
118 version 19 (23, 24). Quantification of gene expression based on counting the overlaps of mapped  
119 reads with genes annotated in the GENCODE gene annotation v19 using HTSeq (Version  
120 0.6.1p2) (25). Read counts were then normalized and RPKM value calculated using EdgeR (26).  
121 Genes with zero read counts for all samples were discarded. The raw data can be accessed via  
122 <http://skinref-dev.dingerlab.org/>.  
123

### 124 **Statistical analysis for identification of RGs from RNA-Seq**

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

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

129

### 130 **qPCR**

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

142

### 143 **Measurement of novel RGs stability**

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

148

### 149 **Statistical analysis of qPCR data**

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

152



## 153 **Results**

### 154 **Identification of novel candidate RGs**

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

165

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

171

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

181

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

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

186

### 187 **qPCR validation of new RGs**

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

195

196 Statistically, GeNorm conduct pairwise variation (V) analysis to identify genes with the least  
197 variance between samples and is denoted as the 'stability' (M) values. In general, lower M  
198 values indicate lower variance in expression value among samples and genes with M values  $\leq$   
199 0.5 are associated with homogeneous samples. Remarkably, all of our 10 RG candidates had M  
200 values  $\leq$  0.5 with RPLP0, RPL7A, RPL23, RPS27A and RPL38 ranked in the top 5 genes for  
201 GeNorm M value/Stability value. In addition, to eliminate errors related to the usage of a single  
202 housekeeping gene, it is common practise to use two or more housekeeping genes. Using  
203 GeNorm, by calculating the normalization factor based on the geometric mean of multiple  
204 control genes, we identified that we need only two of our RG candidates for accurate  
205 normalization (GeNorm V,  $V_{2/3} = 0.084$ ). V values of  $< 0.15$  indicate acceptable stability of the  
206 RG combination, indicating no further need for additional RGs. Amongst our RG candidates, the  
207 pair of genes with optimal normalization factor was RPL38 and RPS27A, which demonstrated  
208 the lowest M values (0.257 and 0.265 respectively) (Figure 2a).

209

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

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

218

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

230

## 231 **Discussion**

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

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

249

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

270

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

277 2010). In this study, we found that the comparison result was significantly altered using a  
278 different calibrator. The upregulation of KRT17 in AK and SCC lesions was even more  
279 pronounced with our candidate RGs. These results demonstrate that choosing an inappropriate  
280 RG may lead to wrong conclusions being drawn from qPCR results.

281

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

294

## 295 **Conclusions**

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

307

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

315

### 316 **Acknowledgements**

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

321

### 322 **References**

- 323 Andersen CL, Jensen JL, and Orntoft TF. 2004. Normalization of real-time quantitative reverse  
324 transcription-PCR data: a model-based variance estimation approach to identify genes  
325 suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res*  
326 64:5245-5250. 10.1158/0008-5472.CAN-04-0496
- 327 Beer L, Mlitz V, Gschwandtner M, Berger T, Narzt MS, Gruber F, Brunner PM, Tschachler E,  
328 and Mildner M. 2015. Bioinformatics approach for choosing the correct reference genes  
329 when studying gene expression in human keratinocytes. *Experimental Dermatology*  
330 24:742-747. 10.1111/exd.12759
- 331 Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA,  
332 Robinson J, Verhaak RG, Sougnez C, Onofrio RC, Ziaugra L, Cibulskis K, Laine E,  
333 Barretina J, Winckler W, Fisher DE, Getz G, Meyerson M, Jaffe DB, Gabriel SB, Lander  
334 ES, Dummer R, Gnirke A, Nusbaum C, and Garraway LA. 2010. Integrative analysis of  
335 the melanoma transcriptome. *Genome Res* 20:413-427. 10.1101/gr.103697.109
- 336 Bustin SA, Benes V, Garson JA, Hellems J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl  
337 MW, Shipley GL, Vandesompele J, and Wittwer CT. 2009. The MIQE guidelines:  
338 minimum information for publication of quantitative real-time PCR experiments. *Clin*  
339 *Chem* 55:611-622. 10.1373/clinchem.2008.112797
- 340 Chari R, Lonergan KM, Pikor LA, Coe BP, Zhu CQ, Chan TH, MacAulay CE, Tsao MS, Lam S,  
341 Ng RT, and Lam WL. 2010. A sequence-based approach to identify reference genes for  
342 gene expression analysis. *BMC Med Genomics* 3:32. 10.1186/1755-8794-3-32
- 343 de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van  
344 der Zee AG, te Meerman GJ, and ter Elst A. 2007. Evidence based selection of  
345 housekeeping genes. *PLoS One* 2:e898. 10.1371/journal.pone.0000898

346 de Kok JB, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, Feuth T, Swinkels DW, and  
347 Span PN. 2005. Normalization of gene expression measurements in tumor tissues:  
348 comparison of 13 endogenous control genes. *Lab Invest* 85:154-159.  
349 10.1038/labinvest.3700208

350 Eisenberg E, and Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet*  
351 29:569-574. 10.1016/j.tig.2013.05.010

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

356 10.1186/1471-2407-13-58

357 Huang da W, Sherman BT, and Lempicki RA. 2009. Bioinformatics enrichment tools: paths  
358 toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37:1-  
359 13. 10.1093/nar/gkn923

360 Hudson LG, Gale JM, Padilla RS, Pickett G, Alexander BE, Wang J, and Kusewitt DF. 2010.  
361 Microarray Analysis of Cutaneous Squamous Cell Carcinomas Reveals Enhanced  
362 Expression of Epidermal Differentiation Complex Genes. *Molecular Carcinogenesis*  
363 49:619-629. 10.1002/mc.20636

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

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

369 Nosrati N, Kapoor NR, and Kumar V. 2015. DNA damage stress induces the expression of  
370 Ribosomal Protein S27a gene in a p53-dependent manner. *Gene* 559:44-51.  
371 10.1016/j.gene.2015.01.014

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

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

378 Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, and Speleman F.  
379 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric  
380 averaging of multiple internal control genes. *Genome Biol* 3:RESEARCH0034.

381 Wagle N, Van Allen EM, Treacy DJ, Frederick DT, Cooper ZA, Taylor-Weiner A, Rosenberg  
382 M, Goetz EM, Sullivan RJ, Farlow DN, Friedrich DC, Anderka K, Perrin D, Johannessen  
383 CM, McKenna A, Cibulskis K, Kryukov G, Hodis E, Lawrence DP, Fisher S, Getz G,  
384 Gabriel SB, Carter SL, Flaherty KT, Wargo JA, and Garraway LA. 2014. MAP kinase  
385 pathway alterations in BRAF-mutant melanoma patients with acquired resistance to  
386 combined RAF/MEK inhibition. *Cancer Discov* 4:61-68. 10.1158/2159-8290.CD-13-  
387 0631

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

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

394



**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.

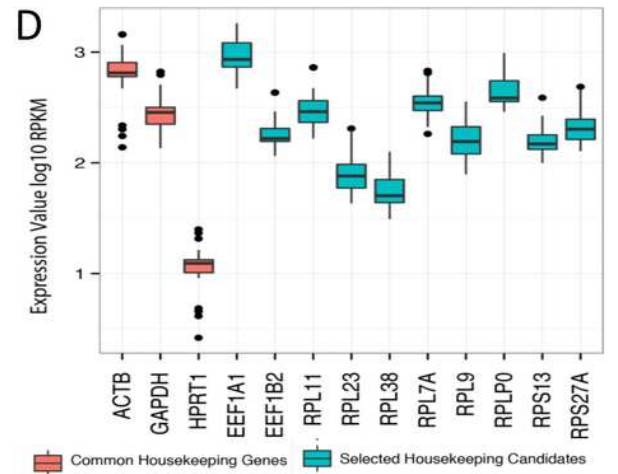
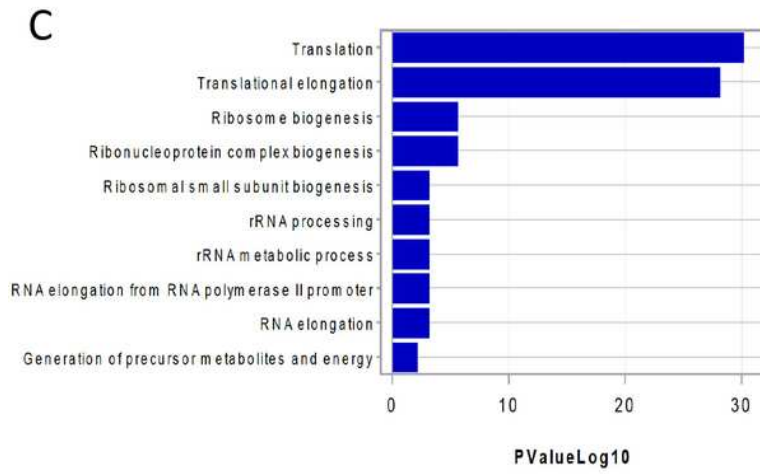
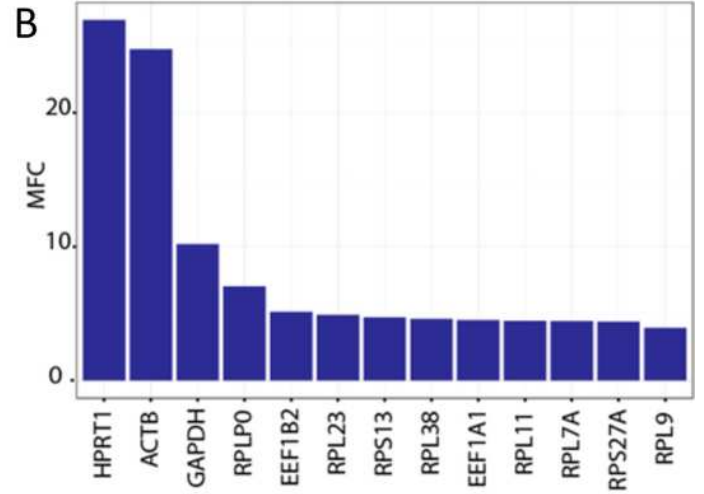
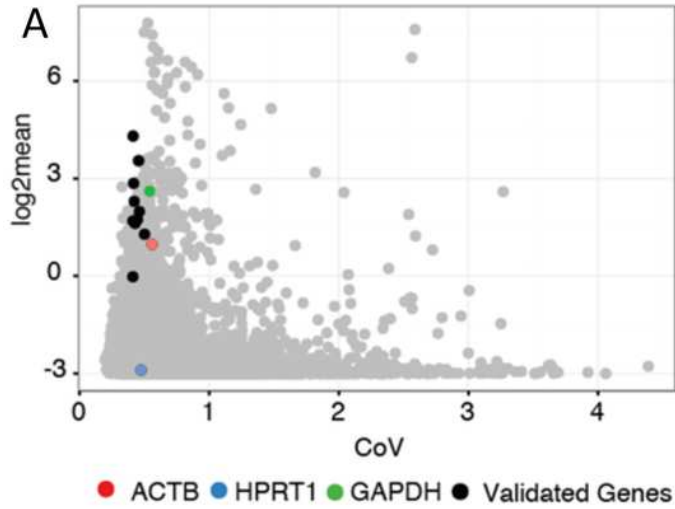
1 **Table 1.** RNA seq scoring of selected candidate reference genes and commonly used reference  
 2 genes, ranked on CoV (coefficient of variation) score. Std = standard deviation, mean = mean  
 3 expression value, MFC = maximum fold change. Candidates are ranked from the smallest to  
 4 largest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

| <b>HGNC<br/>Symbol</b> | <b>CoV</b>      | <b>Std</b>      | <b>Mean</b>     | <b>MFC</b>      |
|------------------------|-----------------|-----------------|-----------------|-----------------|
| 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                  | <i>0.611902</i> | <i>2.476736</i> | <i>4.047605</i> | <i>10.97791</i> |
| HPRT1                  | <i>0.648895</i> | <i>0.004201</i> | <i>0.006474</i> | <i>26.89796</i> |
| ACTB                   | <i>0.766221</i> | <i>0.940382</i> | <i>1.227298</i> | <i>37.8592</i>  |

# Figure 1

## RNA-seq analysis of genes and candidate reference genes

A) Scatterplot comparing Coefficient of variation (CoV) values against mean expression values ( $\log_2$ ) 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, GAPDH and HRT1 (red). Samples with outliers value (expression  $< 95\%$  interquartile range or  $< 5\%$ ) represented as dots.



**Table 2** (on next page)

qPCR primers

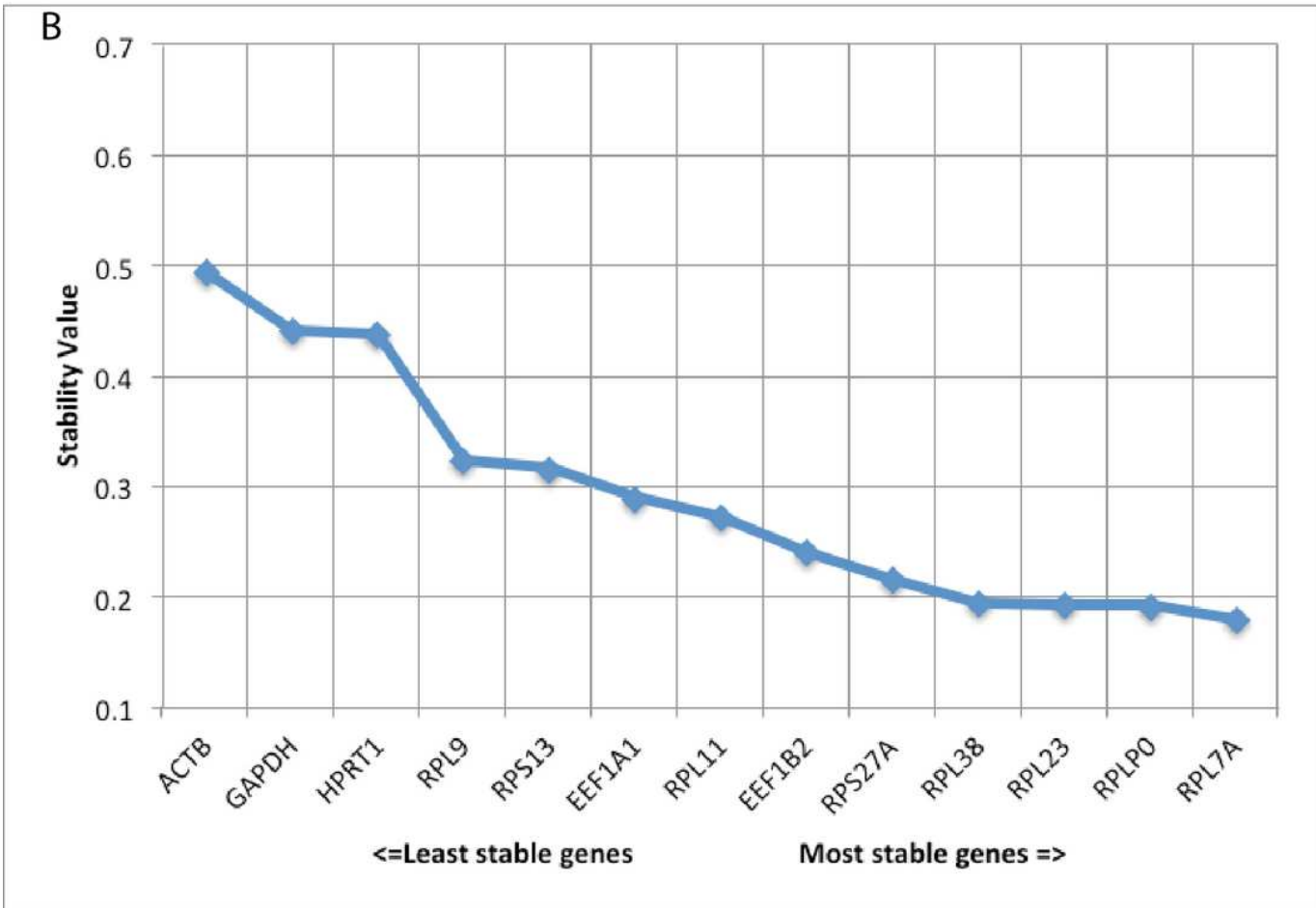
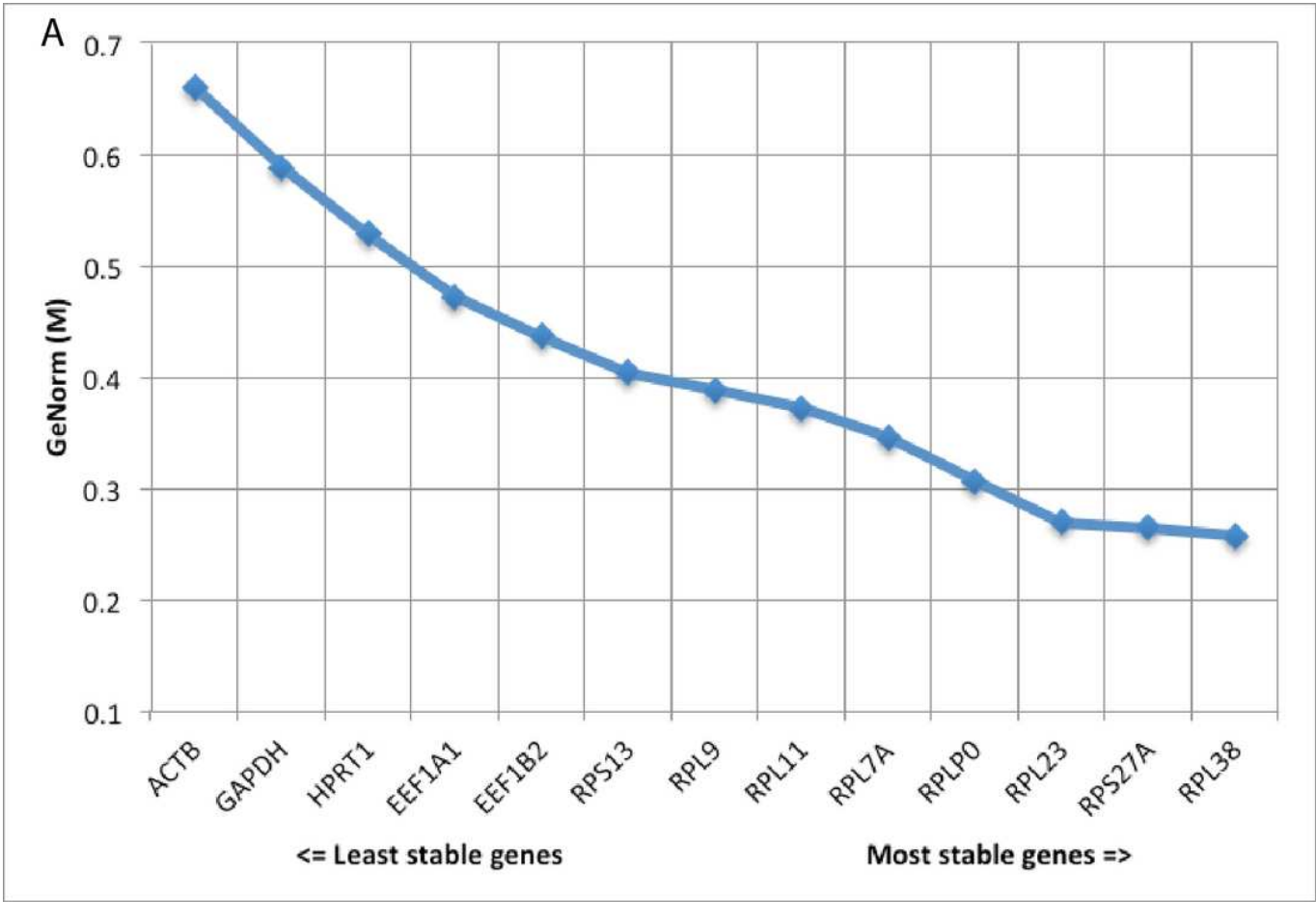
1 **Table 2.** qPCR primers

| <b>Gene</b> | <b>Accession number</b> | <b>Forward primer</b>      | <b>Reverse primer</b>     | <b>Amplification size (bp)</b> |
|-------------|-------------------------|----------------------------|---------------------------|--------------------------------|
| RPL9        | NM_000666<br>1.4        | CTGCGTCTACTGCGAGAA<br>TGA  | CACGATAACTGTGCGTC<br>CCT  | 98                             |
| RPL38       | NM_000999<br>9.3        | GCCATGCCTCGGAAAATT<br>G    | CCAGGGTGTAAGGTAT<br>CTGC  | 139                            |
| RPL11       | NM_000977<br>5.3        | AGAAGGGTCTAAAGGTG<br>CGG   | AGTCCAGGCCGTAGATA<br>CCA  | 138                            |
| RPL23       | NM_000977<br>8.3        | TCCAGCAGTGGTCATTCG<br>AC   | GCAGAACCTTTCATCTC<br>GCC  | 117                            |
| EEF1B2      | NM_001959<br>9.3        | AGTATTTGAAGCCGTGTC<br>CAG  | ACATCGGCAGGACCATA<br>TTTG | 144                            |
| RPS27A      | NM_002959<br>4.5        | ACCACTCCCAAGAAGAA<br>TAAGC | ACTTGCCATAAACACCC<br>CAG  | 147                            |
| RPL7A       | NM_000977<br>2.2        | GGCATTGGACAGGACAT<br>CCA   | AGGCACTTTCAGCCGCT<br>TAT  | 114                            |
| RPS13       | NM_001017<br>7.2        | TCCCCACTTGGTTGAAGT<br>TGA  | AGGAGTAAGGCCCTTCT<br>TGG  | 77                             |
| EEF1A1      | NM_001409<br>2.5        | GAAAGCTGAGCGTGAAC<br>GTG   | AGTCAGCCTGAGATGTC<br>CCT  | 143                            |
| RPLP0       | NM_001009<br>2.3        | ATCAACGGGTACAAACG<br>AGTC  | CAGATGGATCAGCCAAG<br>AAGG | 97                             |
| GAPDH       | NM_002049<br>6.5        | CCCACTCCTCCACCTTTG<br>AC   | TTCCTCTTGTGCTCTTGC<br>TG  | 180                            |
| HPRT1       | NM_000199<br>4.2        | TGCTGAGGATTTGGAAA<br>GGG   | ACAGAGGGCTACAATGT<br>GATG | 115                            |
| ACTB        | NM_001110<br>1.3        | ACCTTCTACAATGAGCTG<br>CG   | CCTGGATAGCAACGTAC<br>ATGG | 148                            |

## 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.





# 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  $\pm$  SEM,  $n = 3$ , \* indicates  $P < 0.05$ ; one-way ANOVA and Turkey post-test.

