



OPEN

Reference gene selection for transcriptional profiling in *Cryptocercus punctulatus*, an evolutionary link between Isoptera and Blattodea

Zhen Li^{1,2,5}, Xiangrui Li^{2,3,5}, Qingwen Zhang¹, Ling Yuan⁴ & Xuguo Zhou¹

The subsocial life style and wood-feeding capability of *Cryptocercus* gives us an evolutionary key to unlock some outstanding questions in biology. With the advent of the Genomics Era, there is an unprecedented opportunity to address the evolution of eusociality and the acquisition of lignocellulases at the genetic level. However, to quantify gene expression, an appropriate normalization strategy is warranted to control for the non-specific variations among samples across different experimental conditions. To search for the internal references, 10 housekeeping genes from a gut transcriptome of a wood-feeding cockroach, *Cryptocercus punctulatus*, were selected as the candidates for the RT-qPCR analysis. The expression profiles of these candidates, including *ACT*, *EF1 α* , *GAPDH*, *HSP60*, *HSP70*, *α TUB*, *UBC*, *RPS18*, *ATPase* and *GST*, were analyzed using a panel of analytical tools, including geNorm, NormFinder, BestKeeper, and comparative ΔC_T method. RefFinder, a comprehensive ranking system integrating all four above-mentioned algorithms, rated *ACT* as the most stable reference gene for different developmental stages and tissues. Expression analysis of the target genes, *Hex-1* and *Cell-1*, using the most or the least appropriate reference genes and a single or multiple normalizers signified this research. Our finding is the first step toward establishing a standardized RT-qPCR analysis in *Cryptocercus*.

Wood-feeding *Cryptocercus*: a "missing link" between cockroaches and termites. Eusociality, in which individuals surrender their own reproduction rights to care for offspring that are not their own, is a fascinating evolutionary mystery and a complex biological trait that has intrigued scientists for decades. Tracking the evolution of this complex trait, however, is not an easy task. Studies on eusocial Hymenoptera, including bees, wasps, and ants, has been greatly facilitated by the existence of intermediates between the ancestral solitary lineages and highly evolved eusocial clades¹. Such phylogenetic intermediates, however, are missing in Isoptera (termites are all eusocial) leading to a tremendous imbalance in sociogenomic research between Isopteran and Hymenopteran societies². Multiple gene sequences analysis demonstrated that subsocial wood-feeding cockroaches in the genus *Cryptocercus*, together with termites, formed a clade nested within a larger cockroach clade, suggesting that wood-feeding cockroaches may be the best model of an evolutionary intermediate between non-eusocial cockroach taxa and eusocial termites³.

Besides the close phylogenetic relationship, the genus *Cryptocercus* also possesses key attributes similar to termites, including wood-feeding capability and subsocial life style with long and complex brood care³⁻⁷. The dual lignocellulose digestion system shared by *Cryptocercus* and termites is highly efficient. Equipped with both endogenous and symbiotic enzymes, these wood-feeding Dictyoptera can convert over 90% of the recalcitrant lignocelluloses into fermentable sugars within 24 h and play a very important ecological role with respect to

¹Department of Entomology and MOA Key Lab of Pest Monitoring and Green Management, China Agricultural University, Beijing, China. ²Department of Entomology, University of Kentucky, S-225 Agricultural Science Center North, Lexington, KY 40546-0091, USA. ³State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, China. ⁴Department of Plant and Soil Sciences, KTRDC, University of Kentucky, Lexington, KY 40546, USA. ⁵These authors contributed equally: Zhen Li and Xiangrui Li email: xuguozhou@uky.edu

global forests carbon cycling and sequestration⁶. Various events have led to the separation of the ancestor group to modern *Cryptocercus*, which remains subsocial, and termites, which becomes eusocial with the evolutionary characters of division of labor, cooperative brood-care and overlapping generations⁸. *Cryptocercus*, considered a “prototermite”, is the logical and the only living intermediate, to study the evolution of eusociality in termites⁹.

Reference gene selection: an indispensable step within the MIQE guideline. Quantitative real-time polymerase chain reaction (RT-qPCR) is, by far, the most widely used and reliable method for the detection and quantification of messenger RNA (mRNA) at the transcription level. The development of RT-qPCR leads to a sensitive, cost effective, and faster measurement of gene expression in comparison to Northern blotting, and makes the accurate quantification of gene expression over a wide concentration range reliable¹⁰. In addition, RT-qPCR has been adopted to validate the results from omics and functional omics analyses^{11–13}. The accuracy of RT-qPCR, however, depends upon various factors, including the biological variability of samples and the technical factors associated with sample preparation, such as the quantity of starting material (e.g., cDNA concentration), RNA extraction, the integrity of RNA, storage conditions, and the efficacy of various reagents and enzymes. Therefore, normalization with internal controls (reference genes) whose expression levels are stable among different tissues, throughout all developmental stages, and/or under various treatments is critical for the accurate quantification of gene expression.

To ensure the reliability of research and integrity of scientific literature, to promote consistency and transparency among laboratories, and to streamline data analysis and interpretation, Bustin and colleagues (2009) proposed a set of MIQE (the Minimum Information for Publication of Quantitative Real-Time PCR Experiments) guidelines to the scientific community as a whole¹⁴. Selection of suitable reference genes is an indispensable step of the MIQE guidelines.

Historically, housekeeping genes, such as *actin* (*ACT*), *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*), and *ribosomal RNAs* (*rRNAs*)¹⁵, have been used extensively as the internal references for RT-qPCR analysis without empirical validations. Under specific experimental conditions, however, their expression may vary substantially^{16–18}. Consequently, there is a growing awareness to select suitable reference genes prior RT-qPCR analysis. This is especially true for non-model organisms, which are currently lagging behind well characterized model organisms in terms of genomic resources and empirically tested reference genes. As a result, researchers have started to embrace the MIQE guidelines and adopted the concept of using multiple rather than a single normalizers^{19–21}. In addition, both systematic and customized studies are encouraged for each organism to identify suitable reference genes^{22,23}.

Goals and objectives. The overall goal of this study is to screen for internal references for the temporal and spatial gene quantification in a wood-feeding cockroach, *C. punctulatus*. Our overarching hypothesis is that housekeeping genes represent a rich reservoir for searching the internal references for RT-qPCR analysis. To test this hypothesis, we investigated the expression profiles of ten housekeeping genes and two target genes under the temporal and spatial conditions. The candidates included *actin* (*ACT*), *elongation factor-1 α* (*EF1 α*), *glyceraldehyde 3 phosphate dehydrogenase* (*GAPDH*), *heat shock protein 60* (*HSP60*), *heat shock protein 70* (*HSP70*), *α -tubulin* (*α TUB*), *ubiquitin conjugating enzyme* (*UBC*), *ribosomal protein S18* (*RPS18*), *adenosinetriphosphatase* (*ATPase*) and *glutathione-S-transferase* (*GST*) from *C. punctulatus*. Target genes, *hexamerin-1* (*Hex-1*) and *β -1,4-endoglucanase* (*Cell-1*), play a critical role in caste differentiation and cellulose degradation^{24,25}, respectively, and serve as the positive controls. The temporal (developmental stage) and spatial (tissue type) expression profiles of these candidates were evaluated comprehensively by a panel of analytic programs, including geNorm, Normfinder, BestKeeper, and comparative ΔC_T method. Ultimately, a specific set of reference genes is recommended by RefFinder, a comprehensive ranking system integrating all four algorithms.

The advent of the next generation sequencing technologies has propelled entomological research into the Genomic Era. As the most primitive extant member of the Blattaria and the sister group of modern termites, *Cryptocercus* is the only evolutionary intermediate between cockroaches and termites. This evolutionary “missing” link represents the key species to address some major outstanding questions in biology (e.g., the evolution of eusociality). Results from this study will facilitate our efforts to (1) standardize the gene quantifications in *C. punctulatus*, (2) functionally decipher the newly sequenced and assembled *C. punctulatus* genome (unpublished data), and (3) decode the genetic basis governing the transition from solitary cockroaches to eusocial termites and the acquisition of symbiotic lignocellulolytic enzymes within woodroach-termite lineage.

Results

Validation of primer sets. The specificity of individual primer sets was evaluated using both gel electrophoresis and melting curve analyses. The banding pattern on 1% agarose gel showed a single band for candidate and target genes individually. Fluorescence data were collected for melting curve analysis, and a single peak was produced by each candidate as well as target gene. Linear regression coefficient for the reproducibility of RT-qPCR (R^2) exceeded 0.99 for all the candidate reference genes and target genes, while amplification efficiency (E%) ranged between 94.1 and 109.3%, suggesting a highly specific and efficient primer design (Table S1 and Table S2).

Optimal cDNA concentration for *GAPDH*. The correlations between the C_t value of *GAPDH* and a gradient of cDNA concentrations generated from three different tissues were shown in Fig. 1. For reproductive organs, ovary (FR) and testis (MR), there was a positive linear relationship between C_t values and cDNA concentrations ranging from 0.1 ng to 1 μ g. Similarly, a positive correlation was observed in neuron ganglion (NG) between C_t values and cDNA concentrations ranging from 0.01 ng to 1 μ g (Fig. 1). Consequently, the minimum

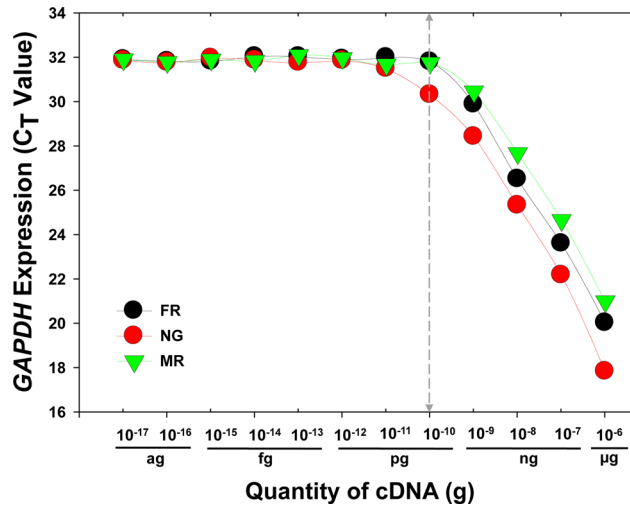


Figure 1. Optimal cDNA concentrations for RT-qPCR analysis. cDNAs synthesized from three different tissues FR (female reproductive organ, ovary), NG (neuron ganglion) and MR (male reproductive organ, testis) were subjected to a tenfold serial dilution before engaging in the subsequent RT-qPCR analysis.

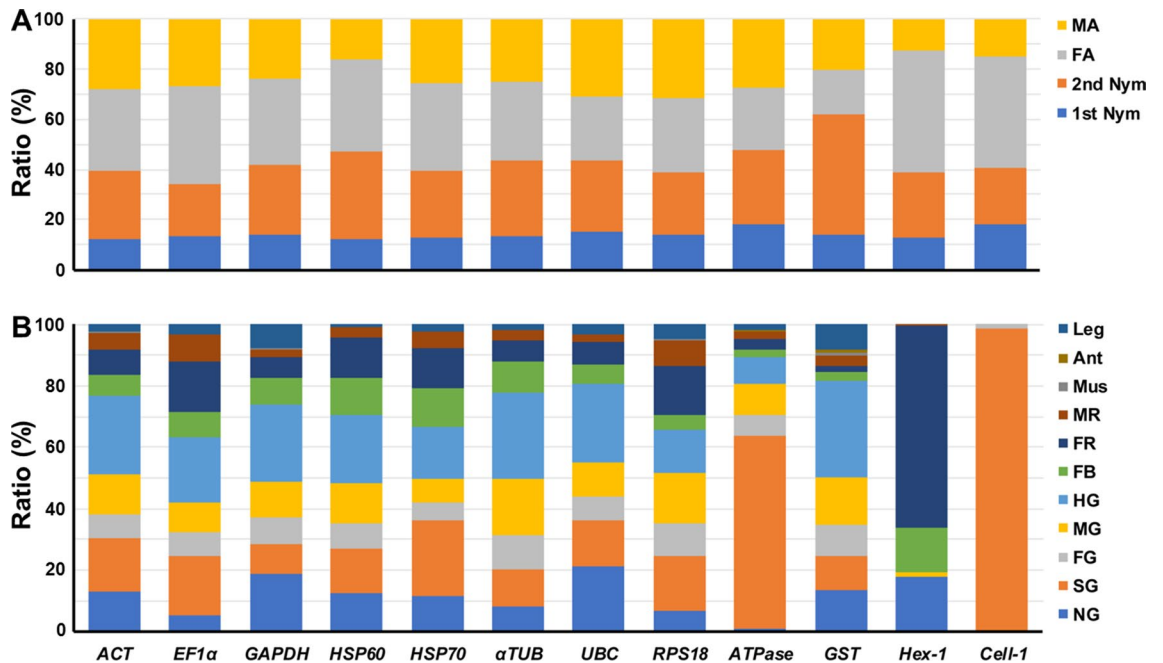


Figure 2. Relative expression ratios of candidate genes among different developmental stages and tissues. Relative expression ratio (%) is shown among (A) different developmental stages, including 1st Nym (1st nymph) and 2nd Nym (2nd nymph), MA (male adult), FA (female adult); and (B) different tissues, including NG (neuron ganglion), SG (salivary gland), FG (foregut), MG (midgut), HG (hindgut), FB (fatbody), FR (female reproductive organ, ovary), MR (male reproductive organ, testis), Mus (muscle), Leg, and Ant (antenna). The summation of a specific gene expression levels from samples of all the developmental stages or tissues are respectively regarded as 100%.

quantity of cDNAs needed for accurate quantification of *GAPDH* expression in *C. punctulatus* is approximately 0.1 ng.

Relative gene expressions among different developmental stages and tissues. Throughout different developmental stages, all candidate genes exhibited the highest expression level in adult females, and the lowest expression level in the 1st nymphs (Fig. 2A; Table S3). The results from different tissues illustrated that all candidate genes showed notably different expression patterns, especially the target genes (Fig. 2B; Table S4). *Hex-1*, a negative regulator of worker-soldier caste differentiation, exhibited significantly higher expressions

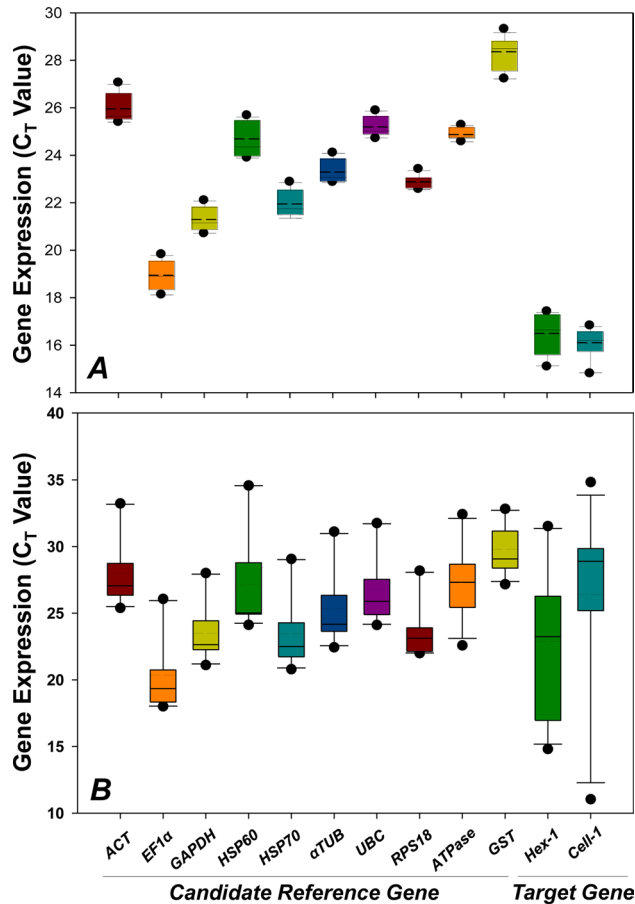


Figure 3. Variability of candidate genes at mRNA level among different experimental conditions. BoxPlots of (A) different developmental stages and (B) different tissues were generated from raw C_T values of ten candidate reference genes and two target genes examined by RT-qPCR analysis. The plots denote median, upper and lower quartiles, and 10th and 90th percentile of data. Dashed lines within bars denote the means.

in the ovary (FR) and fat body (FB). *Cell-1*, a highly conserved endogenous endoglucanases, resided predominantly in the salivary gland (SG). These results demonstrated that the expression profile of housekeeping genes, although relatively stable in comparison to target genes, could vary among different developmental stages and tissues, signifying the importance and necessity for the selection of suitable reference genes.

Stability analysis. Based on the C_T values and BoxPlot analysis (SigmaPlot 10.0), the dispersal of expressions in candidate reference genes displayed range, extreme values and outliers (Fig. 3A,B). Among them, the expression profiles of *ATPase*, *RPS18*, *UBC*, and *αTUB* were relatively stable throughout different developmental stages (Fig. 3A), whereas *RPS18*, *GAPDH*, *UBC*, *HSP70*, *ACT* and *αTUB* were relatively stable across different tissues (Fig. 3B).

geNorm calculates M-value (stability value) for each candidate reference gene and genes with a lower M-value (below the threshold value of 1.5) were considered stable. For different developmental stages, *αTUB* was the most stable candidate with the lowest M value, while *ACT* was the most stable reference gene among tissues (Table 1). BestKeeper calculates the SD and r value of each reference gene. Genes with a SD value < 1.0 and r value > 0.9 are considered stable. Candidate with the lowest SD and the highest r values was identified as the most stable reference gene. *GAPDH* was the most stable candidate throughout developmental stages, while *RPS18* was the one among different tissues (Table 1). NormFinder calculates gene stability through an ANOVA -based algorithm and genes showing the lowest stability values (below the threshold value of 1) are consider stable. *GAPDH* and *EF1α* were the most stable candidates for different developmental stages and tissues, respectively (Table 1). The comparative ΔC_T method also ranks the stability of reference gene through a stability value, in which genes with a lower stability values were considered with a higher level of stability. As a result, *ACT* and *HSP70* were the most stable candidates throughout developmental stages, while *ACT* was also the most stable reference gene among tissues (Table 1).

Finally, RefFinder provides the most comprehensive ranking by integrating the geomean of stability values derived from all four analytic tools. For developmental stages, the rank of candidates from the most to the least stable was *ACT* $>$ *HSP70* $>$ *GAPDH* $>$ *αTUB* $>$ *UBC* $>$ *EF1α* $>$ *HSP60* $>$ *GST* $>$ *ATPase* $>$ *RPS18*, while, for different tissues, it was *ACT* $>$ *UBC* $>$ *EF1α* $>$ *HSP70* $>$ *αTUB* $>$ *RPS18* $>$ *GAPDH* $>$ *GST* $>$ *ATPase* $>$ *HSP60* (Fig. 4).

| geNorm Ranking | Developmental stage | | Tissue | | | |
|-------------------------|---------------------|-----------------|---------------|-----------------|-------------|--------------|
| | Gene name | M-value | Gene name | M-value | | |
| 1 | <i>αTUB</i> | 0.300 | <i>ACT</i> | 0.892 | | |
| 2 | <i>GAPDH</i> | 0.307 | <i>UBC</i> | 0.991 | | |
| 3 | <i>HSP70</i> | 0.316 | <i>αTUB</i> | 1.018 | | |
| 4 | <i>ACT</i> | 0.329 | <i>HSP70</i> | 1.038 | | |
| 5 | <i>UBC</i> | 0.371 | <i>EF1α</i> | 1.043 | | |
| 6 | <i>RPS18</i> | 0.387 | <i>RPS18</i> | 1.123 | | |
| 7 | <i>ATPase</i> | 0.408 | <i>GAPDH</i> | 1.144 | | |
| 8 | <i>EF1α</i> | 0.453 | <i>ATPase</i> | 1.599 | | |
| 9 | <i>HSP60</i> | 0.551 | <i>HSP60</i> | 1.630 | | |
| 10 | <i>GST</i> | 0.717 | <i>GST</i> | 1.709 | | |
| BestKeeper Ranking | Developmental stage | | | Tissue | | |
| | Gene name | SD | [r] | Gene name | SD | [r] |
| 1 | <i>GAPDH</i> | 0.37 | 0.977 | <i>RPS18</i> | 0.56 | 0.975 |
| 2 | <i>αTUB</i> | 0.40 | 0.984 | <i>GAPDH</i> | 0.74 | 0.959 |
| 3 | <i>ACT</i> | 0.43 | 0.999 | <i>ATPase</i> | 0.91 | 0.889 |
| 4 | <i>HSP70</i> | 0.45 | 0.996 | <i>UBC</i> | 0.96 | 0.981 |
| 5 | <i>UBC</i> | 0.31 | 0.893 | <i>ACT</i> | 0.99 | 0.995 |
| 6 | <i>EF1α</i> | 0.47 | 0.898 | <i>EF1α</i> | 1.03 | 0.982 |
| 7 | <i>RPS18</i> | 1.24 | 0.977 | <i>HSP70</i> | 1.14 | 0.984 |
| 8 | <i>ATPase</i> | 0.51 | 0.043 | <i>αTUB</i> | 1.19 | 0.988 |
| 9 | <i>GST</i> | 0.59 | 0.598 | <i>HSP60</i> | 1.98 | 0.987 |
| 10 | <i>HSP60</i> | 0.66 | 0.772 | <i>GST</i> | 1.52 | 0.837 |
| Normfinder Ranking | Developmental stage | | Tissue | | | |
| | Gene name | Stability Value | Gene name | Stability Value | | |
| 1 | <i>GAPDH</i> | 0.036 | <i>EF1α</i> | 0.145 | | |
| 2 | <i>αTUB</i> | 0.038 | <i>ACT</i> | 0.166 | | |
| 3 | <i>HSP70</i> | 0.088 | <i>HSP70</i> | 0.190 | | |
| 4 | <i>ACT</i> | 0.145 | <i>UBC</i> | 0.206 | | |
| 5 | <i>UBC</i> | 0.230 | <i>αTUB</i> | 0.212 | | |
| 6 | <i>RPS18</i> | 0.231 | <i>RPS18</i> | 0.492 | | |
| 7 | <i>EF1α</i> | 0.236 | <i>GAPDH</i> | 0.598 | | |
| 8 | <i>ATPase</i> | 0.238 | <i>ATPase</i> | 0.813 | | |
| 9 | <i>HSP60</i> | 0.258 | <i>HSP60</i> | 0.816 | | |
| 10 | <i>GST</i> | 0.491 | <i>GST</i> | 1.235 | | |
| Comparative ΔCt Ranking | Developmental stage | | Tissue | | | |
| | Gene name | Stability Value | Gene name | Stability Value | | |
| 1 | <i>ACT</i> | 0.46 | <i>ACT</i> | 0.90 | | |
| 2 | <i>HSP70</i> | 0.46 | <i>UBC</i> | 1.00 | | |
| 3 | <i>αTUB</i> | 0.46 | <i>αTUB</i> | 1.02 | | |
| 4 | <i>GAPDH</i> | 0.47 | <i>HSP70</i> | 1.04 | | |
| 5 | <i>UBC</i> | 0.57 | <i>EF1α</i> | 1.04 | | |
| 6 | <i>EF1α</i> | 0.57 | <i>RPS18</i> | 1.13 | | |
| 7 | <i>HSP60</i> | 0.70 | <i>GAPDH</i> | 1.15 | | |
| 8 | <i>GST</i> | 0.87 | <i>ATPase</i> | 1.60 | | |
| 9 | <i>ATPase</i> | 1.08 | <i>HSP60</i> | 1.67 | | |
| 10 | <i>RPS18</i> | 1.26 | <i>GST</i> | 1.71 | | |

Table 1. Ranking of candidate reference genes. The bold indicates the gene expression is unstable under specific experimental conditions.

The optimal number of reference genes. To search for the optimal number of reference genes, geNorm calculates all pairwise variations under each experimental condition (Fig. 5). Based on Vandesompele and colleagues²⁶, a V_n/V_{n+1} threshold value of 0.15 suggests that the addition of “N + 1” reference gene is not necessary, i.e., “N” number of references genes is sufficient to normalize qRT-PCR results. For developmental stages, $V_{2/3}$ was lower than 0.15, indicating that *ACT* and *HSP70* were sufficient for the accurate normalization (Fig. 5).

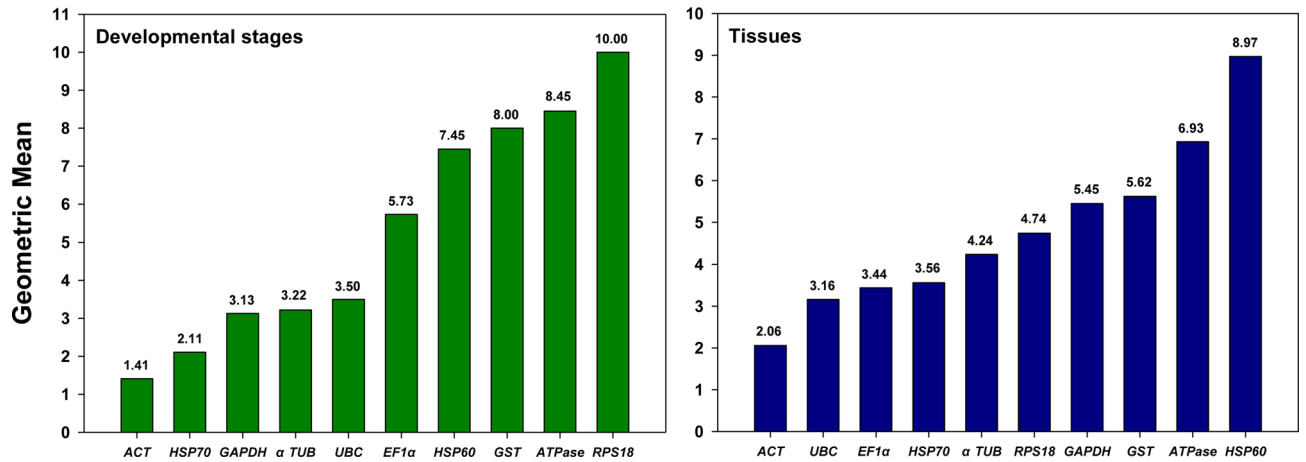


Figure 4. Expression stability of candidate reference genes. The geometric means of the expressional stability were comprehensively calculated by RefFinder for candidate reference genes in different developmental stages and tissues, and the lower value of geometric mean denotes higher expressional stability.

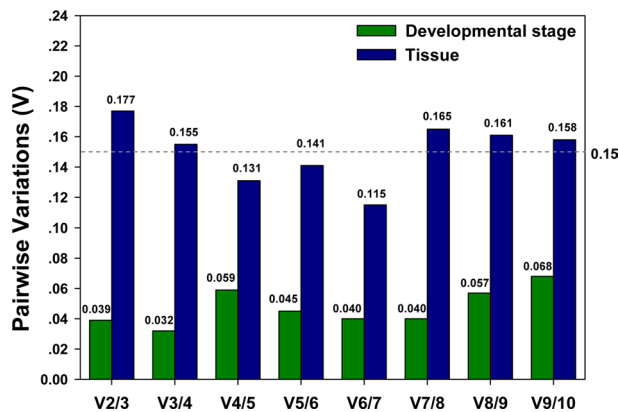


Figure 5. Pairwise variation analysis by geNorm. Optimal number of reference genes required for accurate normalization of target transcript expressions among different developmental stages and different tissues were determined by geNorm analysis based on pairwise variations of $V_{n/n+1}$.

For tissues, however, the first V value less than the threshold was at $V_{4/5}$, suggesting that *ACT*, *UBC*, *EF1 α* and *HSP70* were the best combination for the precise normalization (Fig. 5).

Validation of selected reference genes with target genes *Hex-1* and *Cell-1*. The expression profiles of *Hex-1* and *Cell-1*, the target genes, were evaluated to validate the recommended reference genes under different biotic conditions. Across different developmental stages, the expression profile of *Hex-1* was similar when normalized to the most stable reference gene *ACT* and the recommended multi-gene normalizer (*ACT* and *HSP70*). The expression of *Hex-1* was significantly different when it was normalized to the least stable reference gene *RPS18* (Fig. 6). Specifically, the expression of *Hex-1* was significantly underestimated in the 1st nymphs.

Among different tissues, similar expression profiles of *Cell-1* were observed when *Cell-1* was normalized to the most stable reference gene *ACT*, the recommended multi-gene normalizer (*ACT*, *UBC*, *EF1 α* and *HSP70*), and the least stable gene *HSP60*. Although the expression profiles were similar, *Cell-1* expressions in both salivary gland and foregut were overestimated, especially when *HSP60* was used as the normalizer (Fig. 6).

Discussion

Selection of candidate reference genes. It is unrealistic to find a “universal” normalizer showing constant expression level across all experimental conditions. In this study, expressions of candidate reference genes varied, more or less, among different developmental stages and tissues. Changes in C_t values ≥ 1.0 represent \geq twofold changes in gene expression level, i.e., small variability in C_t values could have drastic impact on target gene expression²⁷. Consequently, selection and validation of genes exhibiting a relative low variability under specific experimental conditions is a critical step toward accurate gene quantification study.

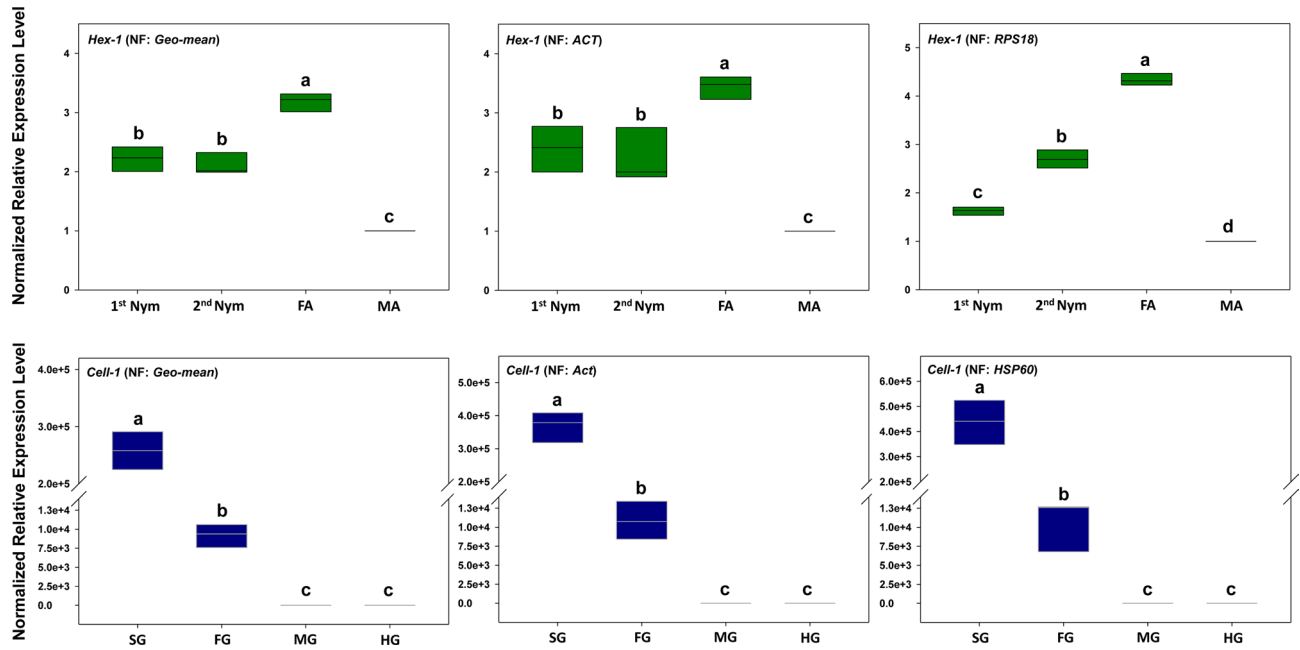


Figure 6. Comparative RT-qPCR analysis of target gene expressions based on different reference genes. Among different developmental stages (**UPPER**), transcriptional profiles of target genes *Hex-1* were determined with the recommended multi-gene normalizer (*ACT* + *HSP70*), the single best endogenous reference gene *ACT* and the single worst endogenous reference gene *RPS18*, respectively. Among different tissues (**LOWER**), expression patterns of *Cell-1* were determined with the recommended multi-gene normalizer (*ACT* + *UBC* + *EF1 α* + *HSP70*), the single best endogenous reference gene *ACT* and the worst endogenous reference gene *HSP60*, respectively. Different letters denote significant expression differences among the three normalizers using one-way ANOVA test ($p < 0.05$) by SPSS (IBM SPSS Statistics 20).

A suitable reference gene should have consistent transcription in all types of cell/tissue types at specific testing conditions, and the transcription of such gene should not be regulated by either internal or external factors²⁸. Additionally, the expression level (C_t value) of target and reference genes should be comparable to ensure that all transcripts are subject to the same kinetic interactions during qRT-PCR²⁶. Otherwise, the expression of a highly abundant internal reference (e.g., ribosomal proteins with significant lower C_t values) can mask the subtle, but potentially biologically relevant, changes in the expression of target genes²⁹. Although the number of reference gene selection publications has been steadily increased for the past decade, the average number of reference genes been tested was 9.53¹⁵. In this study, we selected ten housekeeping genes, which have a track record of being used as the internal controls, as the reference gene candidates. Target genes, *hexamerin-1* (*Hex-1*) and β -1,4-*endoglucanase* (*Cell-1*) are of primary importance for caste differentiation and cellulose degradation research. The expression levels of target and candidate reference genes were comparable, with C_t values ranging between 16 and 25 using cDNAs generated from the whole body of *C. punctulatus* adults.

Previous studies have demonstrated the significant impacts of tissue/cell types and developmental stages on the stability of reference gene expression, in some case, even greater than treatments^{30–33}. Here, we empirically examined the temporal and spatial stability of these candidate genes, and recommended different sets of reference genes for tissue/cell types and developmental stages, respectively.

Stability assessment. Although the underlying algorithms employed by each analytical tool are different, they all focus on the variance in C_t values of each reference gene across treatments³⁴. In this study, reference genes recommended by the four analytical tools exhibit some discrepancies, albeit share some commonalities. For different developmental stages, *GAPDH* was rated as the most stable reference gene by both BestKeeper and Normfinder, whereas *α TUB* and *ACT* were the top choice by geNorm and comparative ΔC_t method. Similarly, *GAPDH* was the reference gene of choice in a few lepidopterans, including the silkworm *Bombyx mori*, *Chilo suppressalis*, the pink stem borer *Sesamia inferens*, and the oriental leafworm moth *Spodoptera litura*^{35–38}, and optimal reference gene for profiling of seasonal and labor-specific gene in Western honey bee, *Apis mellifera*¹⁶. *ACT* was also considered the most stable reference gene in the western corn rootworm, *Diabrotica virgifera virgifera*, the striped rice stem borer, *C. suppressalis* and the Jackfruit borer, *Diaphania caesalis*^{35,36,39}. However, the least stably expressed candidate in *C. punctulatus*, *RPS18*, showed the highest stability in the pink spotted lady beetle, *Coleomegilla maculate*, the housefly, *Musca domestica* and *A. mellifera*^{16,40,41}.

For tissues, both geNorm and comparative ΔC_t method ranked *ACT* as the most stable reference gene, while *RPS18* and *EF1 α* were, respectively, recommended by BestKeeper and Normfinder. Robledo and colleagues³⁴ used a set of empirical data evaluated the accuracy of BestKeeper, Normfinder, geNorm, and comparative ΔC_t method. Authors suggested that NormFinder, complemented with the descriptive statistics calculated by BestKeeper, offers

the most reliable recommendation. In this study, NormFinder selected *GAPDH* and *EF1 α* as the most stable reference genes, respectively, for developmental stages and tissues (Table 1). Indeed, *EF1 α* has been picked as the most stable reference genes across different tissues in many insects, such as bed bug, *Cimex lectularius*, bumble bee, *Bombus lucorum*, diamondback moth, *Plutella xylostella* and oriental armyworm, *Mythimna separata*^{19,42–44}.

The commonality and discrepancies displayed here confirm the notion that no universal reference genes exist for all contexts and reference gene selection and validation is crucial for accurate quantification of gene expression under specific experimental conditions. Without these studies, single un-validated endogenous controls can have profound impacts on data analysis and lead to questionable interpretation^{16,18,19,45,46}. In this study, the expression of *Hex-1* was significantly underestimated in the 1st nymphs when the least stable instead of the most stable and recommended reference genes was used to normalize target gene expression. Similarly, *Cell-1* expressions in both salivary gland and foregut were overestimated when we elected the least stable instead of the most stable and recommended reference genes (Fig. 6). This is consistent with other validation studies that compared the use of stable vs unstable reference genes in the estimation of the target gene expression, in which normalization to unstable reference genes led to over- or under-estimated expressions in the target genes^{47–49}.

Optimal number of reference genes: single vs multiple normalizers. Besides stability, the number of reference genes used for normalization in a specific experiment can impact RT-qPCR analysis as well. Suzuki and colleagues reported that over 90% of the RNA transcription analysis published in peer-reviewed journals used a single housekeeping gene as reference⁵⁰. Housekeeping genes, such as *GAPDH*, *ACT*, and *RPS18*, have been used extensively as the single reference gene without empirical validation, however, many of these reference genes showed substantial variations at expression level under different experimental conditions^{17,51–53}. In fact, as the pool expanded, the chance of these “generic” candidates to be the reference gene of choice decreases³⁴. Since the introduction of MIQE guidelines in 2009, researchers have grown more receptive to adopt multiple rather than a single reference gene in RT-qPCR analysis. Despite changes in perception, the implementation of these guidelines has been challenging. The average number of reference genes used in peer-reviewed publications between 2010 and 2015 remained 1.23, in which 13% of the studies used more than a single reference gene³⁴.

The optimal number of reference genes in a specific study is suggested by geNorm based on the calculation of normalization factors (NFs) in parallel samples. Pairwise variation ($V_{n/n+1}$) is obtained from NF ratios between N and $N + 1$ reference genes. The minimum $V_{n/n+1}$ on a U-shape curve composed by all the $V_{n/n+1}$ represents the most stable NF that can be obtained among all the reference genes in a specific sample set. The number “ N ” corresponds to the optimal number of reference genes that are needed for the most accurate data normalization²⁶. In this study, geNorm showed that all the V values were below the threshold among different developmental stages, with $V_{3/4}$ had the lowest pairwise variation value of 0.032. However, we elected to recommend two reference genes instead of three as the optimal number because $V_{2/3}$ value of 0.039 was equally low and far more practical and economical. Similarly, although $V_{6/7}$ (0.115) predicted the best number of reference genes for different tissues, four was the number of choice for the same set of reasons ($V_{4/5} = 0.131$; Fig. 5).

Interestingly, it seems that more samples involved in the experiment (4 developmental stages vs 11 tissues) demand a higher number of reference genes (2 vs 4) for accurate normalization. A plausible explanation for this phenomenon is that when more samples were added into the analysis, $V_{n/n+1}$ would be slower to reach the minimum value due to the introduction of more unstable factors. Consequently, there is no fixed number of internal controls for gene expression studies. The optimal number of reference genes for accurate normalization can be influenced by $V_{n/n+1}$, sample size, and practicality/feasibility.

cDNA concentration. The other factor which can impact the accuracy of RT-qPCR analysis is the initial concentration of cDNA template. In RT-qPCR, fluorescence is positively correlated with the amount of amplified product, suggesting the C_t value is cDNA concentration-dependent. In this study, the optimal range of cDNA concentration to precisely quantify *GAPDH* expression was between 0.1 ng and 1 μ g for reproductive and neuron tissues. When cDNA was less than 0.1 ng, the expression of tested genes (C_t value) did not correlate with the quantity of cDNA template, which meant no changes could be detected. Although 0.1 ng–1 μ g is specifically for *GAPDH*, accurate quantification of gene expression depends on the optimal range of cDNA concentration, i.e., the quality and quantity of cDNA template can directly impact the accuracy of RT-qPCR analysis.

Materials and methods

Ethics statement. Woodroaches were collected from rotting logs on the grounds of Mountain Lake Biological Station, Giles Co., Virginia (latitude 37.364, longitude 80.519). No specific permits were required for the described field studies.

Colony maintenance. The collected woodroaches were maintained at the University of Kentucky in a ten-gallon glass aquarium under complete darkness and provisioned with brown rotted pine at 20 ± 1 °C with limited humidity. The identity of *Cryptocercus* species was determined by a combination of morphological traits and a molecular marker, 12S rRNA. Based on the diagnostic nucleic acid sites embedded in the amplified 12S rRNA fragments, collected *Cryptocercus* were identified as *C. punctulatus*⁵⁴.

Sample preparation. *Cryptocercus punctulatus* colonies were acclimated in the laboratory for two weeks before they were subjected to the sample preparation. *Cryptocercus punctulatus* colony typically contains a pair of reproductives (adult male and female) and different-sized nymphs.

For developmental stages, we collected four 1st nymphs (1st Nym), three 2nd nymphs (2nd Nym) and one adult male (MA) and one adult female (FA) to represent respective developmental stages within a colony. A total of three colonies were used in this experiment, and each colony represented a biological replication.

For different tissues, leg (Leg), antenna (Ant), muscle (Mus), neuron ganglion (NG), salivary gland (SG), foregut (FG), midgut (MG), hindgut (HG), fatbody (FB), ovary (FR), and testis (MR) were individually dissected from *C. punctulatus* adults. Before dissection, *C. punctulatus* were surface sterilized in 70% ethanol for 1 min and followed by rinsing in sterile water for 30 s. *Cryptocercus punctulatus* adults were dissected under a binocular microscope in 10 mM phosphate buffered saline (PBS, pH 7.8), and respective tissues were snap frozen in liquid nitrogen and stored at -80 °C. Dissected individual tissue samples from three same-sex adults were pooled to represent one tissue type in one biological replication. A total of three biological replications were carried out for this experiment.

Total RNA extraction and cDNA synthesis. *Cryptocercus punctulatus* whole body or dissected tissues were snap frozen in liquid nitrogen, and then ground to powder using a mortar and pestle. To preserve the integrity of RNA, the grinding process was carried out in liquid nitrogen. The resultant ground up powder (≤ 30 mg) was transferred to a 1.5 ml microcentrifuge tube for RNA extraction using a SV Total RNA Isolation Kit (Promega, Madison, WI, USA) according to the manufacturer's instruction. DNA contamination was eliminated by the DNAase treatment for 15 min. Quality and quantity of total RNA was measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher, USA). cDNA was synthesized using the resultant total RNA as the template and M-MLV transcriptase (Grand Island, NY, USA). Samples without reverse transcriptase were used as the negative controls to make sure there was no contamination of DNA.

Selection of candidate reference genes and design of RT-qPCR primers. The selection of candidate reference genes in this study has followed three criteria: (1) they must be housekeeping genes, which are constitutively expressed in all cells/tissue types and maintain basic cellular functions; (2) they have been used historically/extensively as internal references for gene quantification studies in other organisms; and (3) they are presented in a *C. punctulatus* transcriptome (unpublished data). Based on these criteria, we selected ten housekeeping genes, *actin* (*ACT*), *elongation factor-1 α* (*EF1 α*), *glyceraldehyde 3 phosphate dehydrogenase* (*GAPDH*), *heat shock protein 60* (*HSP60*), *heat shock protein 70* (*HSP70*), *α -tubulin* (*α TUB*), *ubiquitin conjugating enzyme* (*UBC*), *ribosomal protein S18* (*RPS18*), *adenosinetriphosphatase* (*ATPase*) and *glutathione-S-transferase* (*GST*), as the candidates with accession numbers from JQ686945 to JQ686954, respectively. Target genes, *hexamerin-1* (*Hex-1*) and *β -1,4-endoglucanase* (*Cell-1*), were extracted from the same transcriptome (unpublished data) with accession numbers JQ686955 and JQ686956, respectively.

Primers were designed by Primer3 (SimGene.com) (Supplementary Table S2), synthesized and diluted to a working concentration of 10 μ M. RT-qPCR reactions were run in triplicate on a Bio-Rad MyiQ™ Single-Color Real-Time PCR Detection System (BioRad, Hercules, CA). The thermal cycling profile included an initial denaturation step at 95 °C for 5 min, followed by 40 cycles of 95 °C for 15 s, annealing at 53 °C for 45 s, and concluded by an extension step at 72 °C for 30 s. Samples were run on 1% agarose gel, and then run with the dissociation protocol for melting curve analysis to check the specificity of each individual primer sets. In addition, amplification efficiency (E%) and correlation coefficient (R^2) were determined based on the standard curves generated from a tenfold serial dilution of cDNAs.

Optimal cDNA concentration for RT-qPCR analysis. cDNAs from ovary (FR), neuron ganglion (NG) and testis (MR), respectively, were quantified using a Smart Spec Plus spectrophotometer (Bio-Rad, Hercules, CA). A tenfold serial dilution was carried out to generate a cDNA concentration gradient ranging from 10^{-6} to 10^{-17} g. After RT-qPCR, C_t (Threshold Cycle, which is the number of cycles required for the fluorescent signal to exceed the threshold line of background level) values of *GAPDH* transcripts corresponding to a gradient of cDNA concentrations were analyzed, and the optimal range of cDNA concentrations was determined.

Stability analysis. Relative expression level of the ten candidate reference genes and the two target genes were calculated by $2^{-\Delta C_t}$ method⁵⁵. The relative expression levels of candidate reference genes across different developmental stages and tissues were analyzed using one-way ANOVA with SPSS Statistics 17.0 (SPSS Inc., Chicago, IL, USA). The means were compared by Tukey test, if the data fit homoscedasticity, and Games-Howell test were performed if not. Specifically, throughout different developmental stages, Tukey test was used for *EF1 α* , *GAPDH*, *HSP70*, *α TUB*, *UBC*, *GST* and *Hex-1*, while Games-Howell test was carried out for *ACT*, *HSP60*, *RPS18*, *ATPase* and *Cell-1*. Relative expression of all the candidate reference genes across different tissues was analyzed using Games-Howell test. The dispersion of C_t values was assessed using a Box Plot.

The expression profiles of the candidate reference genes and target genes under different biotic conditions (developmental stages and tissues) were evaluated individually using a panel of analytic tools, including geNorm²⁶, BestKeeper⁵⁶, Normfinder⁵⁷ and the comparative ΔC_t method⁵⁸. For geNorm, each reference gene is evaluated by calculating the pairwise variation with all other genes to determine the gene-stability value, M^{26} . BestKeeper ranks the reference genes based on the standard deviation (SD) of C_t value and the repeated pairwise correlation analyses of all the candidate genes⁵⁶. Instead of measuring the overall stability, Normfinder selects reference genes based on the possible intra- and inter- group variation across different samples⁵⁷. The comparative ΔC_t method ranks the reference genes by comparing relative expression of "pairs of genes" within each sample, and the stability of the candidates was obtained according to the repeatability of the gene expression differences among different samples⁵⁸. The final composite ranking of stability, however, was provided by RefFinder⁵⁹ (<http://150.216.56.64/referencegene.php>). RefFinder, a web-based analysis tool, assigns an appropriate weight

of the four above mentioned analytical tools to an individual gene and calculates the geometric mean of their weights for the overall ranking.

Relative expression of the target genes, *Hex-1* and *Cell-1*, was calculated using $\Delta\Delta C_t$ method⁶⁰. Differences in their expression using an array of normalization factors were compared according to one-way ANOVA with Tukey test.

Received: 5 September 2017; Accepted: 26 November 2020

Published online: 17 December 2020

References

- Toth, A. L. & Rehan, S. M. Molecular evolution of insect sociality: an eco-evo-devo perspective. *Ann. Rev. Entomol.* **62**, 419–442 (2017).
- Scharf, M. E., Zhou, X. Termite sociogenomics: a growing field. *Correspondence to Nature Reviews Genetics*. [http://www.nature.com/nrg/journal/v6/n4/corres/nrg1575_fs.html] (2005).
- Lo, N. *et al.* Evidence from multiple gene sequences indicates that termites evolved from wood-feeding cockroaches. *Curr. Biol.* **10**, 801–804 (2000).
- Kitade, O. Comparison of symbiotic flagellate faunas between termites and a wood-feeding cockroach of the genus *Cryptocercus*. *Microbes Environ.* **19**, 215–220 (2004).
- Todaka, N. *et al.* Phylogenetic analysis of cellulolytic enzyme genes from representative lineages of termites and a related cockroach. *PLoS ONE* **5**, e8636 (2010).
- Ohkuma, M. Termite symbiotic system: efficient bio-recycling of lignocellulose. *Appl. Microbiol. Biotechnol.* **61**, 1–9 (2003).
- Kinjo, Y. *et al.* Parallel and gradual genome erosion in the *Blattabacterium* endosymbionts of *Mastotermes darwiniensis* and *Cryptocercus* Wood Roaches. *Genome Biol. Evol.* **10**, 1622–1630 (2018).
- Thorne, B. L. Evolution of eusociality in termites. *Annu. Rev. Ecol. Syst.* **28**, 27–54 (1997).
- Klass, K. D., Nalepa, C. A. & Lo, N. Wood-feeding cockroaches as models for termite evolution (Insecta: Dictyoptera): *Cryptocercus* vs. *Perisphaeria boleiriana*. *Mol. Phylogenet. Evol.* **46**, 809–817 (2008).
- Setiawan, A. N. & Lokman, P. M. The use of reference gene selection programs to study the silencing transformation in a freshwater eel *Anguilla australis*: a cautionary tale. *BMC Mol. Biol.* **11**, 75 (2010).
- Wang, X. *et al.* De nova characterization of microRNAs in oriental fruit moth *Grapholita molesta* and selection of reference genes for normalization of microRNA expression. *PLoS ONE* **12**, e0171120 (2017).
- Zhou, X., Wheeler, M. M., Oi, F. M. & Scharf, M. E. RNA interference in the termite *Reticulitermes flavipes* through ingestion of double-stranded RNA. *Insect Biochem. Mol. Biol.* **38**, 805–815 (2008).
- Li, S. *et al.* The genomic and functional landscapes of developmental plasticity in the American cockroach. *Nat. Commun.* **9**, 1008 (2018).
- Bustin, S. A. *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* **55**, 611–622 (2009).
- Chapman, J. R. & Waldenström, J. With reference to reference genes: a systematic review of endogenous controls in gene expression studies. *PLoS ONE* **10**, e0141853 (2015).
- Moon, K., Lee, S. H. & Kim, Y. H. Validation of quantitative real-time PCR reference genes for the determination of seasonal and labor-specific gene expression profiles in the head of Western honey bee, *Apis mellifera*. *PLoS ONE* **13**, e0200369 (2018).
- Thellin, O. *et al.* Housekeeping genes as internal standards: use and limits. *J. Biotechnol.* **75**, 291–295 (1999).
- Freitas, F. C. P. *et al.* Evaluation of reference genes for gene expression analysis by realtime quantitative PCR (qPCR) in three stingless bee species (Hymenoptera: Apidae: Meliponini). *Sci. Rep.* **9**, 17692 (2019).
- Li, K. *et al.* Identification and validation of reference genes for RT-qPCR normalization in *Mythimna separata* (Lepidoptera: Noctuidae). *BioMed Res. Int.* **2018**, 1828253 (2018).
- Zhang, L. *et al.* Selection of reference genes for gene expression analysis of plant-derived microRNAs in *Plutella xylostella* using qRT-PCR and ddPCR. *PLoS ONE* **14**, e0220475 (2019).
- Deng, Y. *et al.* Screening and validation of reference genes for RT-qPCR under different honey bee viral infections and dsRNA treatment. *Front. Microbiol.* **11**, 1715 (2020).
- Hruz, T. *et al.* RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genom.* **12**, 156 (2011).
- Gutierrez, L. *et al.* The lack of a systematic validation of reference genes: a serious pitfall undervalued in reverse transcription-polymerase chain reaction (RT-PCR) analysis in plants. *Plant Biotechnol. J.* **6**, 609–618 (2008).
- Zhou, X., Oi, F. M. & Scharf, M. E. Social exploitation of hexamerin: RNAi reveals a major caste-regulatory factor in termites. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 4499–4504 (2006).
- Zhou, X., Wheeler, M. M., Oi, F. M. & Scharf, E. Inhibition of termite cellulases by carbohydrate-based cellulase inhibitors: Evidence from *in vitro* biochemistry and *in vivo* feeding studies. *Pestic. Biochem. Phys.* **90**, 31–41 (2008).
- Vandesompele, J. *et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, research0034.1-research0034.11 (2002).
- Ferguson, B. S., Nam, H., Hopkins, R. G. & Morrison, R. F. Impact of reference gene selection for target gene normalization on experimental outcome using real-time qRT-PCR in adipocytes. *PLoS ONE* **5**, e15208 (2010).
- Nygaard, A. B., Jørgensen, C. B., Cirera, S. & Fredholm, M. Selection of reference genes for gene expression studies in pig tissues using SYBR green qPCR. *BMC Mol. Biol.* **8**, 67 (2007).
- Dheda, K. *et al.* Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *Biotechniques* **37**, 112–114 (2004).
- Jacobsen, A. V., Yemaneab, B. T., Jass, J. & Scherbak, N. Reference gene selection for qPCR is dependent on cell type rather than treatment in colonic and vaginal human epithelial cell lines. *PLoS ONE* **9**, e115592 (2014).
- Rocha-Martins, M., Njaine, B. & Silveira, M. S. Avoiding pitfalls of internal controls: Validation of reference genes for analysis by qRT-PCR and Western blot throughout rat retinal development. *PLoS ONE* **7**, e43028 (2012).
- Uddin, M. *et al.* Age-related changes in relative expression stability of commonly used housekeeping genes in selected porcine tissues. *BMC Res. Notes* **4**, 441 (2011).
- Fernandes, J. M. O., Mommens, M., Hagen, O., Babiak, I. & Solberg, C. Selection of suitable reference genes for real-time PCR studies of Atlantic halibut development. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **150**, 23–32 (2008).
- Robledo, D. *et al.* Analysis of qPCR reference gene stability determination methods and a practical approach for efficiency calculation on a turbot (*Scophthalmus maximus*) gonad dataset. *BMC Genom.* **15**, 648 (2014).
- Rodrigues, T. B. *et al.* Validation of reference housekeeping genes for gene expression studies in western corn rootworm (*Diabrotica virgifera virgifera*). *PLoS ONE* **9**, e109825 (2014).

36. Teng, X., Zhang, Z., He, G., Yang, L. & Li, F. Validation of reference genes for quantitative expression analysis by real-time RT-PCR in four lepidopteran insects. *J. Insect Sci.* **12**, 60 (2012).
37. Sun, M., Lu, M., Tang, X. & Du, Y. Exploring valid reference genes for quantitative real-time PCR analysis in *Sesamia inferens* (Lepidoptera: Noctuidae). *PLoS ONE* **10**, e0115979 (2015).
38. Lu, Y. *et al.* Identification and validation of reference genes for gene expression analysis using quantitative PCR in *Spodoptera litura* (Lepidoptera: Noctuidae). *PLoS ONE* **8**, e68059 (2013).
39. Wang, Z. *et al.* Identification and evaluation of reference genes for normalization of gene expression in developmental stages, sexes, and tissues of *Diaphania caesalis* (Lepidoptera, Pyralidae). *J. Insect Sci.* **20**, 6 (2020).
40. Yang, C. *et al.* Selection of reference genes for RT-qPCR analysis in a predatory biological control agent, *Coleomegilla maculata* (Coleoptera: Coccinellidae). *Sci. Rep.* **5**, 18201 (2015).
41. Zhong, M. *et al.* Selection of reference genes for quantitative gene expression studies in the house fly (*Musca domestica* L.) using reverse transcription quantitative real-time PCR. *Acta Biochim. Biophys. Sin.* **45**, 1069–1073 (2013).
42. Mamidala, P., Rajarapu, S. P., Jones, S. C. & Mittapalli, O. Identification and validation of reference genes for quantitative real-time polymerase chain reaction in *Cimex lectularius*. *J. Med. Entomol.* **48**, 947–951 (2011).
43. Hornáková, D., Matoušková, P., Kindl, J., Valterová, I. & Pichová, I. Selection of reference genes for real-time polymerase chain reaction analysis in tissues from *Bombus terrestris* and *Bombus lucorum* of different ages. *Anal. Biochem.* **397**, 118–120 (2010).
44. Fu, W. *et al.* Exploring valid reference genes for quantitative real-time PCR analysis in *Plutella xylostella* (Lepidoptera: Plutellidae). *Int. J. Biol. Sci.* **9**, 792–802 (2013).
45. VanGuilder, H. D., Vrana, K. E. & Freeman, W. M. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* **44**, 619–626 (2008).
46. Bustin, S. A. & Nolan, T. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J. Biomol. Tech.* **15**, 155–166 (2004).
47. Kosir, R. *et al.* Determination of reference genes for circadian studies in different tissues and mouse strains. *BMC Mol. Biol.* **11**, 60 (2010).
48. Anderson, K. C. & Elizur, A. Hepatic reference gene selection in adult and juvenile female Atlantic salmon at normal and elevated temperatures. *BMC Res. Notes* **5**, 21 (2012).
49. Ju, X. H. *et al.* Selection of reference genes for gene expression studies in PBMC from Bama miniature pig under heat stress. *Vet. Immunol. Immunop.* **144**, 160–166 (2011).
50. Suzuki, T., Higgins, P. J. & Crawford, D. R. Control selection for RNA quantitation. *Biotechniques* **29**, 332–337 (2000).
51. Gutierrez, L., Mauriat, M., Pelloux, J., Bellini, C. & Van Wuytswinkel, O. Towards a systematic validation of references in real time RT-PCR. *Plant Cell* **20**, 1734–1735 (2008).
52. Deindle, E., Boengler, K., van Royen, N. & Schaper, W. Differential expression of GADPH and beta3-actin in growing collateral arteries. *Mol. Cell Biochem.* **236**, 139–146 (2002).
53. Glare, E. M., Divjak, M., Bailey, M. J. & Walters, E. H. Beta-Actin and GADPH housekeeping gene expression in asthmatic airways is variable and not suitable for normalising mRNA levels. *Thorax* **57**, 765–770 (2002).
54. Kambhampati, S. & Smith, P. T. PCR primers for the amplification of four insect mitochondrial gene fragments. *Insect Mol. Biol.* **4**, 233–236 (1995).
55. PrimerDesign Ltd. (2012) GeNorm™ housekeeping gene selection kit handbook.
56. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. 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 (2004).
57. Andersen, C. L., Jensen, J. L. & Ørntoft, T. F. 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 (2004).
58. Silver, N., Best, S., Jiang, J. & Thein, S. L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* **7**, 33 (2006).
59. Xie, F., Sun, G., Stiller, J. W. & Zhang, B. Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database. *PLoS ONE* **6**, e26980 (2011).
60. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, e45 (2001).

Acknowledgements

Authors are grateful to Christine Nalepa (North Carolina State University) for providing *C. punctulatus* and for her comments on an earlier draft. Authors are also thankful to anonymous reviewers for their constructive comments and suggestions. This research was supported by a hatch project from USDA-NIFA (Accession No. 0220839; Project No. KY008053) and a pilot project from Kentucky Tobacco Research and Development Center (KTRDC), University of Kentucky. The information reported in this paper is part of a project of the Kentucky Agricultural Experiment Station and is published with the approval of the Director. These agencies had no role in study design, data collection/analysis, manuscript preparation, or the decision to publish.

Author contributions

Z.L. and X.L. conceived the experiments; Z.L. and X.Z. designed the study; Z.L. and X.Z. analyzed the data; Z.L. drafted the manuscript, Q.Z., L.Y., and X.Z. revised the manuscript. All authors read and approved the final version of manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-79030-6>.

Correspondence and requests for materials should be addressed to X.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020