Original Article

Correlation and Variation-Based Method for Identifying Reference Genes from Large Datasets

Oliver Yuan Wei Chan[1,§], Bryan Ming Hsun Keng[1,§], Maurice Han Tong Ling[2,3]

[1] Raffles Institution, Republic of Singapore
[2] Department of Zoology, The University of Melbourne, Australia
[3] School of Chemical and Biomedical Engineering, Nanyang Technological University, Republic of Singapore.
mauriceling@acm.org
[§] Equal contributors

Corresponding Author:
Maurice Han Tong Ling, School of Chemical and Biomedical Engineering, Nanyang Technological University, Republic of Singapore, Tel: +65.96669233 , Email: mauriceling@acm.org

Abstract:
Introduction: Reference genes are assumed to be stably expressed under most circumstances. Previous studies have shown that identification of potential reference genes using common algorithms, such as NormFinder, geNorm, and BestKeeper, are not suitable for microarray-sized datasets. The aim of this study was to evaluate existing methods and develop methods for identifying reference genes from microarray datasets.
Methods: We evaluated the correlation between outputs from 7 published methods for identifying reference genes, including NormFinder, geNorm, and BestKeeper, using subsets of published microarray data. From these results, seven novel combinations of published methods for identifying reference genes were evaluated.
Results: Our results showed that NormFinder's and geNorm's indices had high correlations ($R^2 = 0.987$, $P < 0.0001$), which is consistent with the findings of previous studies. However, NormFinder's and BestKeeper's indices ($R^2 = 0.489$, $0.01 < P < 0.05$) and NormFinder's coefficient of variance (CV) suggested a lower correlation ($R^2 = 0.483$, $0.01 < P < 0.05$). We developed two novel methods with high correlations with NormFinder (R^2 values of both methods were 0.796, $P < 0.0001$). In addition, computational times required by the two novel methods were linear with the size of the dataset.
Conclusion: Our findings suggested that both of our novel methods can be used as alternatives to NormFinder, geNorm, and BestKeeper for identifying reference genes from large datasets. These methods were implemented as a tool, OLIgonucleotide Variable Expression Ranker (OLIVER), which can be downloaded from http://sourceforge.net/projects/bactome/files/OLIVER/OLIVER_1.zip.

Keywords: reference standards, computing methodologies

1. Introduction

Gene expression analysis is examining the variations in gene expression by measuring DNA expression levels over time. These variations may result from many factors, such as environmental, developmental, and metabolic changes or treatments. A quantitative, real-time, polymerase chain reaction (qRT-PCR) is commonly used to quantify gene expressions (1, 2). However, qRT-PCR requires a stably expressed gene under a wide variety of

conditions (3). This stably-expressed gene, also known as a reference gene, acts as a standard to estimate the relative expression of various other genes of interest.

Candidate reference genes, which are commonly assumed to be invariant, can be identified using statistically-based algorithms, such as geNorm (4), NormFinder (5), and BestKeeper (6), or by using descriptive statistics, such as regression analyses (7). Microarrays, which usually contain thousands of probes, present a good source of data for identifying reference genes (8). Several studies have successfully identified reference genes from microarrays (9-12).

However, existing methods of identifying potential reference genes, including geNorm (4), NormFinder (5), and BestKeeper (6), are only capable of analyzing a small sample. For example, geNorm can only identify reference genes among 102 genes. Hence, for large datasets amounting to thousands of genes, such algorithms are unable to rank all genes against each other. While it has been possible to split the dataset into smaller subsets and run these algorithms on each subset before merging the results (13), comparison will be made against only a small portion of the entire dataset. This potentially results in inaccuracies by preventing ranking of all genes in the dataset against each other.

Presently, the coefficients of variance (CV) and regression parameters (14), such as gradient of gene expression, have been proposed as computationally simple methods for initial screening for potential reference genes. This reduces the potential pool of genes before using another more accurate algorithm to accurately and stringently rank the invariance of the remaining genes. Furthermore, it has been shown that there is strong correlation between a gene's CV and its ranking by the algorithm (10) but the correlation is not perfect, which may lead to discrepancies in identifying the most invariant gene. While this method is computationally feasible even for datasets containing many genes, there is a possibility of excluding genes with worse CVs that may subsequently be ranked by the second algorithm as highly invariant.

There are discrepancies in the ranking of potential reference genes by various existing methods based on the same data (15-17), indicating that the results of a single method do not necessarily signify invariance. Many studies (10, 12) have used CV or regression parameters as a preliminary shortlist to select a certain subgroup of genes from a large dataset, after which a more complex algorithm was used to identify the most invariant genes of this subgroup. Although this reduces the number of genes to a number that is statistically manageable by stronger algorithms, such as NormFinder and geNorm, it is possible that suitable reference genes may be eliminated by the initial shortlisting (13). Even among commonly-used algorithms, such as geNorm (4), NormFinder (5) and BestKeeper (6), there are discrepancies (16, 18, 19). Therefore, caution is advised when using of a single algorithm as a benchmark.

In this study, we evaluated methods used as preliminary filters in previously-published papers that used NormFinder as a benchmark. Novel methods derived from these parameters, which were found to have high correlation with NormFinder and were suitable for identifying reference genes from large datasets, are proposed. These methods are implemented as a tool that we called OLIgonucleotide Variable Expression Ranker (OLIVER), which can be downloaded from http://sourceforge.net/projects/bactome/files/OLIVER/OLIVER_1.zip.

2. Material and Methods

2.1. Research design. This study presents a computational analysis and comparison of existing methods for identifying reference genes using publically-available, microarray datasets to derive novel methods for identifying reference genes from large datasets.

2.2. Microarray data. Five publically-available, microarray datasets were used in this study. Four of the datasets were from *E. coli* K-12, of which three were from substrain MG1655 and one was from substrain W3110. Briefly, the studies conducted with the datasets are as follows: GSE680: MG1655 grown in either aerobic or anaerobic conditions, deleted for transcriptional regulators in oxygen response, and used to validate a computational model of transcriptional and metabolic networks (20); GSE1099: aerobically grown MG1655 cells in several media with varied carbon sources, including glucose, glycerol, succinate, L-alanine, acetate, and L-proline (21); GSE1494: analysis of derivatives of strain MG1655 of wild type, fur mutant, and wild type with added $FeSO_4$, induced to overexpress *RyhB*, a noncoding RNA regulated by the fur repressor protein (22); GSE1827: W3110 cells grown aerobically and exposed to low, neutral, or high pH to study acid and base response (23). The remaining dataset (GSE 2021) was from the liver of *S. tridecemlineatus* sampled during summer, interbout arousal, and late torpor (24).

2.3. Dataset selection. The coefficient of variation (CV) of every gene was calculated as the quotient of standard deviation and arithmetic mean of its microarray data within the dataset. For the *E. coli* datasets (GSE680, GSE1099, GSE1494, GSE1827), the top 500 genes with the lowest CV from each dataset were identified. Similarly, the top 1000 genes as ranked by CV in GSE2021 were identified. These formed the large-scale datasets used for

further evaluation. From the large-scale datasets of GSE1827, GSE680, and GSE2021, 10 sets of 10 randomly-selected genes were chosen from each dataset for small-scale evaluation of the correlation of the invariant gene-finding algorithms because BestKeeper (6) can only process 10 genes per evaluation. These formed 30 small-scale datasets for evaluation.

2.4. Methods to Identify Reference Genes. We compiled eight methods for identifying reference genes and developed a software tool called OLIgonucleotide Variable Expression Ranker (OLIVER). The first method was coefficient of variation, which had been used in Chia et al. (9). Methods 2 through 5 were from Lee et al. (14), consisting of linear regression gradient, regression ratio (R^2/slope), absolute regression ratio, and product of mean and standard deviation, respectively. Methods 6 through 8 were derived from Keng et al. (12); Method 6 with product correlation, which is the absolute pairwise correlation between dataset X and the product of X and a few random samples of non-X data; Method 7 deals with ratio correlation, which is the average the absolute pairwise correlation between dataset X and the quotient of X and a few random samples of non-X data; Method 8 deals with self correlation, which is the average absolute pairwise correlation between dataset X and a few random samples of non-X data. In Methods 1, 2, 5, 6, 7, and 8, smaller values indicate greater gene stability, while, in Methods 3 and 4, larger value indicate greater gene stability.

2.5. Small-scale evaluation with BestKeeper, geNorm and NormFinder. BestKeeper version 1 (6), NormFinder version 0.953 (5), geNorm (4), and OLIVER Methods 1-5 were used to analyze each of the 30 small datasets of 10 randomly-selected probes. Pearson's correlation coefficient was calculated between each of the various methods' stability values to determine which methods were stably correlated to the benchmarking algorithms, i.e., BestKeeper, NormFinder, and geNorm.

2.6. Large-scale evaluation against NormFinder. From the small-scale evaluation, NormFinder was chosen as the benchmark by which to evaluate the correlation of other methods' ranking of invariant genes. OLIVER and NormFinder were used to determine the stability of each gene in the large-scale dataset. A ranking of the genes in each dataset was obtained for each method, with the lowest ranking gene having the greatest invariance. Pearson's correlations between the stability values of the respective methods with NormFinder's stability indices were calculated to determine the similarity of each method's identified potential reference genes and NormFinder's ranking. The differences between the top-ranked genes and NormFinder's rankings compared to each method also were determined.

3. Results

3.1. Correlation between existing methods. An analysis of various methods for identifying reference genes (Figure 1) suggested that NormFinder and geNorm were consistent with each other with the highest correlation ($R^2 = 0.987$). The second highest correlation occurred between BestKeeper and CV ($R^2 = 0.722$).

While the correlations between BestKeeper and CV with NormFinder (0.489 and 0.483, respectively) have p-values of less than 0.05, suggesting that the correlation was significant, the p-values were more than 0.01, indicating that the methods were not strongly correlated. These results are similar to the correlations between BestKeeper and CV with geNorm, having R^2 values of 0.459 and 0.455, respectively. The products of the mean and the standard deviation, and gradient of expression values have R^2 values of 0.140 and 0.105, respectively.

3.2. Correlation between the proposed methods and NormFinder. Table 1 shows the correlations of our three proposed methods, i.e., product correlation, ratio correlation, and self correlation, with NormFinder, for each of the five datasets. In three of the datasets, ratio correlation had a very strong correlation with NormFinder, even higher than that of CV at 0.806. Furthermore, product correlation also had a correlation with the mean value of 0.489 (P < 0.01). Therefore, ratio correlation and product correlation were used in our composite methods. Since ratio correlation had a better logarithmic fit than linear fit with NormFinder, e^{ratio} was used instead of the actual ratio correlation.

3.3. Introduction of new methods. Seven new methods were formulated to analyze the data obtained from the large datasets. These methods were derived, partially or in full, from the initial eight methods outlined in OLIVER. Ratio correlation, CV, and product correlation were selected as the methods that had the highest likelihood of producing a better correlation. These novel methods are as follows:

1. e^{ratio}, where *ratio* refers to the ratio correlation, which is OLIVER method 7
2. (e^{ratio}/ mean of e^{ratio}) + (CV / mean of CV)
3. e^{ratio}+ CV + product, where *product* refers to the product correlation and *ratio* refers to ratio correlation, which are OLIVER methods 6 and 7, respectively
4. (e^{ratio}/ mean of e^{ratio}) + (CV / mean of CV) + (product / mean of product)
5. (e^{ratio}/ minimum of e^{ratio}) + (CV / minimum of CV) + (product / minimum of product)

6. Geometric mean of e^{ratio} and CV
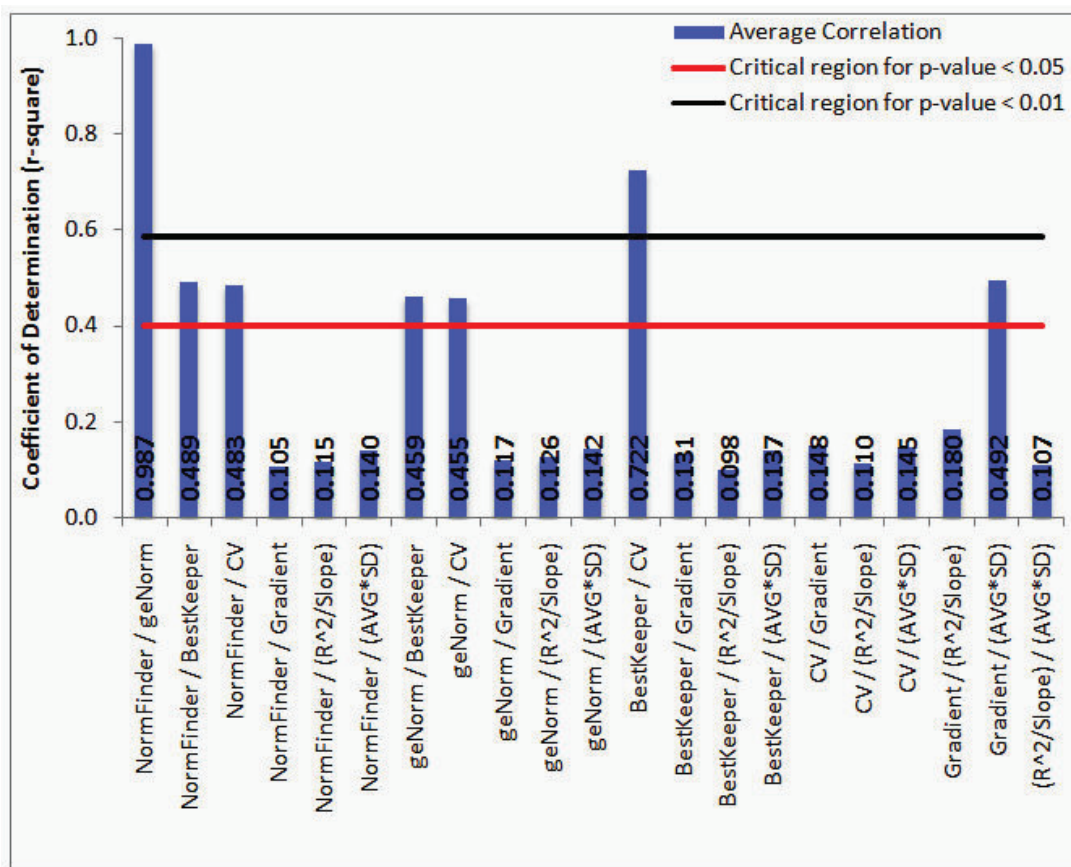7. Harmonic mean of e^{ratio} and CV

Figure 1. Correlations between methods for determining reference genes

Table 1. Correlation of our proposed methods with NormFinder

Dataset	Original methods		
	Product Correlation	Ratio Correlation	Self Correlation
GSE2021	-0.465	0.665	-0.708
GSE1827	0.705	0.935	-0.023
GSE680	0.542	0.964	-0.202
GSE1099	0.841	0.902	0.024
GSE1494	0.820	0.911	0.252
Mean value	**0.489**	**0.875**	**-0.131**

3.4. Evaluation of all methods against NormFinder. Average correlation and standard deviation of correlation between each method and NormFinder was calculated using five large datasets. Our results (Table 2) showed that Method 10 [$(e^{ratio}$ / mean of e^{ratio}) + (CV / mean of CV)] and Method 14 (Geometric mean of e^{ratio} and CV) had the strongest correlation with NormFinder (R = 0.892), which consistently outperformed CV across the five datasets, indicating that it may perform better than CV as a preliminary filter for reference genes before using NormFinder. In addition, both Methods 10 and 14 demonstrated a higher correlation with NormFinder than the correlation between NormFinder and BestKeeper (R^2 = 0.489, r = 0.700), geNorm and BestKeeper (R^2 = 0.459, r = 0.677), and CV and BestKeeper (R^2 = 0.722, r = 0.850) from the evaluation of the 30 small datasets.

Table 2. Correlation of various methods with NormFinder

OLIVER Methods	Description	Average correlation	Standard deviation of correlation
1	CV	0.806	0.235
2	Gradient	0.146	0.211
3	R^2/slope	-0.003	0.063
4	abs R^2/slope	0.061	0.125
5	AVG x SD	-0.016	0.185
6	Product correlation	0.489	0.546
7	Ratio correlation	0.875	0.120
8	Self correlation	-0.131	0.361
9	e^{ratio}	0.875	0.124
10	(e^{ratio}/ mean of e^{ratio}) + (CV / mean of CV)	0.892	0.145
11	e^{ratio}+ CV + product	0.873	0.178
12	(e^{ratio}/ mean of e^{ratio}) + (CV / mean of CV) + (product / mean of product)	0.827	0.213
13	(e^{ratio}/ minimum of e^{ratio}) + (CV / minimum of CV) + (product / minimum of product)	0.883	0.145
14	Geometric mean of e^{ratio} and CV	0.892	0.144
15	Harmonic mean of e^{ratio} and CV	0.849	0.155

Table 3. Computation time of varying sample and gene sizes

Number of Genes	Number of Samples	Time Taken (s)
1000	26	1.98
1000	130	8.04
1000	260	15.2
5000	26	9.85
5000	130	40.1
5000	260	76.2
25000	26	48.8
25000	130	200
25000	260	382
100000	26	195
100000	130	786
100000	260	1496

Both Methods 10 and 14 use CV and ratio correlation as their fundamental methods. Ratio correlation is more computationally intensive than the CV calculation because multiple correlation calculations are required to generate a ratio correlation for each gene, whereas there is only one CV calculation for each gene. Hence, we evaluated the time required to complete the ratio correlations, where 30 correlations were required for the calculation of each ratio correlation, using various combinations of the number of genes and samples (Table 3). Our test computer's specifications were as follows: Intel Core i5-2401M 2.3GHz with 4GB DDR3 RAM.

By striating our results by the number of samples, our results showed that the time required increased linearly with the number of genes. For example, using the computational time for 1000 genes for each striation of number of samples as the baseline for scaling, we expected 198 seconds for 26 samples of 100,000 genes; 804 seconds for 130 samples of 100,000 genes; and 1950 seconds for 260 samples of 100,000 genes. Our results are not significantly different from what was expected (n = 9 since the three baselines of 1000 genes were not used for statistical calculation; chi-squared P = 0.99). However, by striating our results by the number of genes, our results were significantly different from what we expected (n = 8 because the four baselines of 26 samples were not used for calculation; chi-squared P = 2×10^{-44}), suggesting that the time required to complete the ratio correlations was less for larger sample sizes than smaller sample sizes when the number of genes was kept constant. Nevertheless, the computational time taken was highly correlated ($R^2 = 0.999$) with the size of the dataset (product of the number of genes and the number of samples).

4. Discussion

4.1. Discrepancies between methods for determining reference genes. Several studies have shown discrepancies between the results of NormFinder, geNorm, and BestKeeper. Kim et al. (17) evaluated six commonly-used reference genes in rat carotid body and found that the gene rankings by these methods were similar, but not identical. This was supported by Li et al. (25), who found that in the fungus *Candida glabrata*, the ranking of reference genes was similar between geNorm and another algorithm, hkgFinder, but the ranking was different between BestKeeper and NormFinder. Together, their findings demonstrated that the three algorithms are well correlated but that they do not produce identical rankings. Ratert et al. (19) analyzed urothelial carcinoma and found that geNorm, NormFinder, and BestKeeper analyses gave different combinations of recommended reference genes for normalization. Taki and Zhang (16) found similar discrepancies in their analysis of *C. elegans* with NormFinder, delta-Ct (26), and BestKeeper. Among the top five reference genes identified by each method, only two genes, i.e., CDC-42 and TBA-1, were common between NormFinder and BestKeeper, and these genes also were common with delta-Ct. More significantly, Klie and Debener (18) found that among NormFinder, geNorm, and BestKeeper, there was no single gene that showed stable expression across all environmental conditions, and the individual rankings of each gene differed between the algorithms. Nevertheless, these three methods have all been accepted as algorithms that are able to accurately identify invariant genes.

4.2. Selecting NormFinder and CV as commonly-used benchmarks. In this study, we used NormFinder as our benchmark for evaluating the 'correctness' of simpler methods' evaluations of potential reference genes' stability. Our analysis of existing invariant gene-ranking methods (Figure 1.1) showed a high correlation between NormFinder and geNorm ($R^2 = 0.987$, P < 0.0001), while BestKeeper's correlation with NormFinder was only significant at the 95% confidence level. This was in agreement with the findings of previous studies, suggesting that NormFinder, geNorm, and BestKeeper may give similar, but not identical, results (16).

To determine the relative usage of the algorithms NormFinder, geNorm, BestKeeper, and delta-Ct in studies involving reference genes, a keyword search was conducted on PubMed (www.ncbi.nlm.nih.gov/pubmed). There were 351, 483, 239, and 114 instances of NormFinder, geNorm, BestKeeper, and delta-Ct, respectively, in the titlse or abstracts in manuscripts published before September 2013. This suggested that NormFinder and geNorm are the most widely-used methods of determining gene invariance.

Because NormFinder and geNorm are strongly correlated ($R^2 = 0.987$), we selected one as a benchmark against which our proposed methods could be assessed. NormFinder was chosen due to its ability to rank a larger number of genes. NormFinder has no fixed limits to the number of genes it can rank, and it may compute up to roughly 1000 genes in a reasonable time span. Conversely, geNorm and BestKeeper can analyze a maximum of 102 and 10 genes, respectively, as well as a maximum of 102 and 50 samples, respectively. CV is commonly used as a benchmark because it is easy to calculate and because its computational complexity scales linearly with the increasing size of the dataset, which cannot be said for NormFinder (11). Hence, we aimed to formulate methods that were better than CV for shortlisting the most invariant genes or as alternatives to NormFinder, geNorm, and BestKeeper for large datasets.

4.3. Analysis of novel methods. Our results (Table 4) from the small datasets and the large datasets suggested that Methods 2-5 (14) had low correlations with NormFinder (correlation coefficients between -0.016 and 0.157). In addition, there were larger differences in the rankings of their top-ranked genes than in NormFinder, which suggested that the method also was not viable for identifying the best candidate reference genes.

Table 4. Correlation of Methods 2-5 with NormFinder

OLIVER Methods	Description	Average Correlation		Average Top Rank Differences	
		Small Datasets	Large Datasets	NormFinder Standard	Method Standard
2	Gradient	0.0635	0.146	301.0	209.8
3	R²/slope	0.157	-0.003	182.8	279.6
4	abs R²/slope	-	0.061	283.6	334.0
5	AVG x SD	-0.00487	-0.016	377.8	510.0

Our results suggest that the correlation between CV and NormFinder (P = 0.0118) was statistically significant at the 95% confidence level, but not at the 99% confidence level. Hence, we concluded that CV may be, in certain instances, a poor filter when used prior to more computationally-intensive methods because it may exclude genes with high invariance. Our best methods, i.e., (e^{ratio}/ mean of e^{ratio}) + (CV / mean of CV) (Method 10) and geometric mean of e^{ratio} and CV (Method 14), showed strong correlations with NormFinder, indicating that they may

be suitable as preliminary filters before using NormFinder. In addition, both methods (Methods 10 and 14) demonstrated higher correlations than NormFinder and BestKeeper or NormFinder and geNorm.

4.4. Time taken for computation. We evaluated the computation time required to complete the ratio correlation for different sizes of data because this step is more computationally-intensive than the coefficient of variation calculations, and it is required by both Methods 10 and 14. Our results suggested that the computational time required to complete the ratio correlation was linearly related to the size of the data, which is the product of the number of genes and the number of samples. This suggested that Methods 10 and 14 may be suitable alternatives for identifying reference genes from large transcriptomic datasets.

4.5. Possible limitations. The numbers of samples were different in small-scale and large-scale comparisons. Since correlation is affected by sample size, this may affect the comparability of the correlations. In addition, GSE 2021 showed results that were significantly different from those of the other datasets. For example, the correlation between ratio and NormFinder was 0.665 for GSE 2021, whereas it was consistently above 0.900 for the other datasets. We hypothesized that this was because *Spermophilus* samples originating from different stages of torpor and awakening were not separated as such. Previous studies suggested that organisms' awake and torpor stages affected their gene expression levels (27). Hence, the expression levels of various genes might drastically differ between the various stages (24), affecting the probes' stability values.

5. Conclusions

In this study, we evaluated 10 different published methods for identifying reference genes and found a strong correlation between the results provided by NormFinder and geNorm. However, these two methods were not suitable for analyzing large datasets. We identified two novel methods for identifying reference genes that were capable of processing large transcriptomic datasets. In addition, the computational time required linearly proportional to the size of the dataset, suggesting that these two novel methods might be suitable alternatives for identifying reference genes in large, transcriptomic datasets. We used our methods for identifying reference genes as a tool, referred to as OLIgonucleotide Variable Expression Ranker (OLIVER), which can be downloaded from the following URL http://sourceforge.net/projects/bactome/files/OLIVER/OLIVER_1.zip.

Conflict of Interest:
There is no conflict of interest to be declared.

Authors' contributions:
This study is conceived and supervised by ML. OC and BK contributed equally in data analysis and drafting of the manuscript. ML developed OLIVER. All authors read and approved the final manuscript.

References
1. Fedrigo O, Warner LR, Pfefferle AD, Babbitt CC, Cruz-Gordillo P, Wray GA. A Pipeline to Determine RT-QPCR Control Genes for Evolutionary Studies: Application to Primate Gene Expression across Multiple Tissues. *PLoS ONE.* 2010; 5(9), e12545. PMID 20824057. dx.plos.org/10.1371/journal.pone.0012545
2. Remans T, Smeets K, Opdenakker K, Mathijsen D, Vangronsveld J, Cuypers A. Normalisation of real-time RT-PCR gene expression measurements in *Arabidopsis thaliana* exposed to increased metal concentrations. *Planta.* 2008; 227, 1343-9. PMID 18273637. dx.doi.org/10.1007/s00425-008-0706-4
3. Agabian N, Thomashow L, Milhausen M, Stuart K. Structural analysis of variant and invariant genes in trypanosomes. *American Journal of Tropical Medicine and Hygiene.* 1980; 29(Supplement 5), 1043-9. PMID 7435803. http://www.ajtmh.org/cgi/pmidlookup?view=long&pmid=7435803
4. Vandesompele J, de Preter K, Pattyn F. Accurate normalisation of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology.* 2002; 3:0034.1-0034.11. PMID 12184808. http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/12184808/
5. Andersen CL, Jensen JL, Ørntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research.* 2004; 64:5245-50. PMID 15289330. http://cancerres.aacrjournals.org/cgi/pmidlookup?view=long&pmid=15289330
6. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnology Letters*. 2004; 26(6), 509-15. PMID 15127793.
7. Kanji G: 100 Statistical tests (3rd edition). Los Angeles: *Sage Publications* 2006.
8. Wurmbach E, Yuen T, Sealfon SC. Focused microarray analysis. *Methods*. 2003; 31:306-16. PMID 14597315.
9. Chia CY, Lim CW, Leong WT, Ling MHT. High expression stability of microtubule affinity regulating kinase 3 (MARK3) makes it a reliable reference gene. *IUBMB Life.* 2010; 62(3):200-3. PMID 20087965. dx.doi.org/10.1002/iub.295
10. Heng SSJ, Chan OYW, Keng BMH, Ling MHT. Glucan biosynthesis protein G, (mdoG) is a suitable reference gene in Escherichia coli K-12. *ISRN Microbiology*. 2011; Article ID 469053. PMID 23724305. dx.doi.org/10.5402/2011/469053
11. Too HK, Ling MHT. Signal peptidase complex subunit 1 and hydroxyacyl-CoA dehydrogenase beta subunit are suitable reference genes in human lungs. *ISRN Bioinformatics*. 2012; Article ID 790452. dx.doi.org/10.5402/2012/790452
12. Keng BMH, Chan OYW, Heng SSJ, Ling, MH. Transcriptome analysis of Spermophilus lateralis and Spermophilus tridecemlineatus liver does not duggest the presence of Spermophilus-liver-specific reference genes. *ISRN Bioinformatics*. 2013; Article ID 361321. dx.doi.org/10.1155/2013/361321
13. Too IHK, Heng SSJ, Chan OYW, Keng BMH, Chia CY, Lim CWX, Leong WT, Chu QH, Ang EJG, Lin YJ, Ling MHT. Identification of Reference Genes by Meta-Microarray Analyses. In: James V. Rogers, editor. Microarrays: Principles, Applications and Technologies. New York: *Nova Science Publishers, Inc. 2014.*
14. Lee S, Jo M, Lee J, Koh SS, Kim S. Identification of novel universal housekeeping genes by statistical analysis of microarray data. *Journal of Biochemistry and Molecular Biology.* 2007; 40(2), 226-31. PMID 17394773. http://www.jbmb.or.kr/fulltext/jbmb/view.php?vol=40&page=226
15. Wang Q, Ishikaw T, Michiu T, Zhu BL, Guan DW, Maeda H. Stability of endogenous reference genes in postmortem human brains for normalization of quantitative real-time PCR data: comprehensive evaluation using geNorm, NormFinder, and BestKeeper. *International Journal of Legal Medicine*. 2012; 126(6), 943-52. PMID 23010907. dx.doi.org/10.1007/s00414-012-0774-7
16. Taki FA, Zhang B. Determination of reliable reference genes for multi-generational gene expression analysis on C. elegans exposed to abused drug nicotine. *Psychopharmacology*. 2013; 1-12. PMID 23681163. http://dx.doi.org/10.1007/s00213-013-3139-0
17. Kim I, Yang D, Tang X Carroll J. Reference gene validation for qPCR in rat carotid body during postnatal development. *BMC Research Notes*. 2011; 4(1), 440. PMID 22023793. http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/22023793/
18. Klie M, Debener T. Identification of superior reference genes for data normalisation of expression studies via quantitative PCR in hybrid roses (Rosa hybrida). *BMC Research Notes*. 2011; 4(1), 518. PMID 22123042. http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/22123042/
19. Ratert N, Meyer, HA, Jung M, Mollenkopf HG, Wagner I, Miller K et al. Reference miRNAs for miRNAome analysis of urothelial carcinomas. *PLoS ONE*. 2012; *7*(6), e39309. PMID 22745731. http://dx.plos.org/10.1371/journal.pone.0039309

20. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*. 2004; 429(6987), 92-96. PMID 15129285. dx.doi.org/10.1038/nature02456

21. Liu M, Durfee T, Cabrera JE, Zhao K, Jin DJ, Blattner FR. Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli. *Journal of Biological Chemistry*. 2005; *280*(16), 15921-7. PMID 15705577. http://www.jbc.org/cgi/pmidlookup?view=long&pmid=15705577

22. Massé E, Vanderpool CK, Gottesman S. Effect of RyhB small RNA on global iron use in Escherichia coli. *Journal of Bacteriology*. 2005; *187*(20), 6962-71. PMID 16199566.
http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/16199566/

23. Maurer LM, Yohannes E, Bondurant SS, Radmacher M, Slonczewski JL. pH regulates genes for flagellar motility, catabolism, and oxidative stress in Escherichia coli K-12. *Journal of Bacteriology*. 2005; *187*(1), 304-19. PMID 15601715. http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/15601715/

24. Williams DR, Epperson LE, Li W, Hughes MA, Taylor R, Rogers J et al. Seasonally hibernating phenotype assessed through transcript screening. *Physiological Genomics*. 2005; *24(1)*, 13–22. PMID 16249311. http://physiolgenomics.physiology.org/cgi/pmidlookup?view=long&pmid=16249311

25. Li Q, Skinner J, Bennett J. Evaluation of reference genes for real-time quantitative PCR studies in Candida glabrata following azole treatment. *BMC Molecular Biology*. 2012; *13*(1), 22. PMID 22747760. http://www.biomedcentral.com/1471-2199/13/22

26. Silver N, Best S, Jian J, Thein S. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Molecular Biology*. 2006; *7*(1), 33. PMID 17026756. http://www.biomedcentral.com/1471-2199/7/33

27. Crawford FI, Hodgkinson CL, Ivanova E, Logunova LB, Evans GJ, Steinlechner S, Loudon AS. Influence of torpor on cardiac expression of genes involved in the circadian clock and protein turnover in the Siberian hamster (Phodopus sungorus). *Physiological Genomics*. 2007; *31*(3), 521-30. PMID 17848604. http://physiolgenomics.physiology.org/cgi/pmidlookup?view=long&pmid=17848604