In: Microarrays: Principles, Applications and Technologies ISBN: 978-1-62948-669-7
Editor: James V. Rogers

Chapter 6

IDENTIFICATION OF REFERENCE GENES BY META-MICROARRAY ANALYSES

Issac HK Too[1], Sean SJ Heng[2], Oliver YW Chan[2], Bryan MH Keng[2], Ching-Yee Chia[3], Claudia WX Lim[3], Wan-Ting Leong[3], Qinghao Chu[2], Ernest JG Ang[2], Yongjie Lin[2] and Maurice HT Ling[4,5]*

[1]Department of Biological Sciences, National University of Singapore, Singapore
[2]Raffles Institution, Singapore
[3]Nanyang Girls' High School, Singapore
[4]Department of Zoology, The University of Melbourne, Australia
[5]School of Chemical and Biomedical Engineering,
Nanyang Technological University, Singapore

ABSTRACT

The expression levels of reference genes used in gene expression studies are assumed to not change under most circumstances. However, a number of studies have demonstrated that genes theoretically assumed to be stably expressed were found to vary under experimental conditions. In addition, previous studies have also reported that stably expressed genes in an organ, may not be stably expressed in other organs or in a different organism, suggesting the need to identify reference genes for each organ and each organism. Due to its ability to analyze the expression of thousands of genes in an experiment, microarrays present a suitable resource for the analysis and identification of reference genes. We present four cases on practical applications of microarrays whereby multiple published microarray data sets were examined to identify suitable reference genes using coefficient of variation (CV) and NormFinder. Our results suggest that microtubule affinity-regulating kinase 3 (MARK3) is a suitable reference gene for mouse liver, 40S ribosomal protein S29 (Rps29) is a suitable reference gene for mouse testes and pancreas, signal peptidase complex subunit 1 (SPCS1) and hydroxyacyl-CoA

* Corresponding author: mauriceling@acm.org.

dehydrogenase beta subunit (HADHB) are suitable reference genes for human lungs, and glucan biosynthesis protein G (mdoG) is a suitable reference gene for *Escherichia coli*. Further analysis suggests that the identified reference genes are involved in fundamental biochemical processes. This supports the theoretical basis and previous studies that housekeeping genes, on the whole, are generally stably expressed. However, our results also suggest that certain housekeeping genes that are stably expressed in one tissue or one organism may not be stably expressed in different tissues or organisms, supporting the need to identify reference genes for each tissue and organism.

INTRODUCTION

In the context of molecular cell biology, it is crucial to examine the variation in gene expression over the time in order to study the functions of individual genes. This provides valuable insight on the effects of differential gene expression on cell division, as well as biological responses in various conditions, such as developmental phases and effects of treatments. Profiling of gene expression by the mean of quantitative real-time polymerase chain reaction (qRT-PCR), Northern blot and DNA microarray analysis [1] have been well established. However, a number of variables, such as cell type, mRNA extraction and handling techniques, and analytical quantification approaches may result in different gene expression measurements [2]. This may affect analytical accuracy [3]. In order to address these variations, normalization usually involving a group of calibrating genes is employed in many gene expression quantification studies [3].

Housekeeping genes are generally used as calibrating genes in gene expression studies [3] due to their theoretical stable expression [4]. The ideal set of housekeeping genes should be expressed stably and consistently across all samples [5]. However, several studies have suggested that universal reference genes are relatively difficult to identify [6-9]. This corroborates several studies demonstrating that genes that are originally considered invariable in terms of expression may vary under different experimental conditions [10-12]. For an accurate comparison of DNA expression in different samples, it is necessary to use verified reference genes, such as GAPDH (glyceraldehyde-3-phosphate dehydrogenase) or UBQ (ubiquinone) [13], for normalization or determine new reference genes for each experimental system with varying external stimuli [6, 14]. However, it has been demonstrated that the expression of GAPDH [15] and UBQ [16] varied in some conditions. This indicates that GAPDH or UBQ may not be useful reference genes in different experimental platforms. Previous studies showed that ACTB (beta actin) and SDHA (succinate dehydrogenase complex, subunit A) were effectively used as reference genes in breast tumor studies [17], while IPO8 (importin 8) was stably expressed in human lung specimens. Another study also showed that EIF2A (eukaryotic translation initiation factor 2A, 65kDa) and PPIB (peptidylprolyl isomerase B, also known as cyclophilin B) were stably expressed in mouse liver and adrenal gland [18]. Other studies also demonstrated that EEF1A1 (eukaryotic translation elongation factor 1 alpha 1) expressed consistently in human cervical tissues [19], while B2M (beta-2-microglobulin) and RPL29 (ribosomal protein L29) were well used as reference genes in human stomach tissue studies [20]. High expression stability showed by RPL32 (ribosomal protein L32), GAPDH, POLR2A (DNA directed RNA polymerase II polypeptide A, 220kDa), TBP (TATA box binding protein), PGK1 (phosphoglycerate kinase

1) and RPL4 (ribosomal protein L4) are demonstrated as the most frequently observed reference genes in rodent and human heart gene expression studies [21], while RPS18 (ribosomal protein S18) was stably expressed in head and neck squamous cell carcinoma [22]. In addition, GUSB (glucuronidase, beta) was generally used as reference gene in human ovarian studies [23], while RPS4X (ribosomal protein S4, X-linked) and RPL13A (ribosomal protein L13a) were useful in quantifying the gene expression in larvae of flatfish [24] and osteoarthritic canine articular tissues [3] respectively. This suggests that expression of different housekeeping genes may be inconsistent among different species or organs.

To date, the selection of reference genes can be achieved by few mathematical methods, such as geNorm [25], BestKeeper© [26] and NormFinder [27] to evaluate gene expression stability in qRT-PCR [3]. BestKeeper© [26] and geNorm [25] are based on pairwise correlation has been shown to be sensitive to co-regulated genes [28]. On the other hand, NormFinder is less sensitive to co-regulated genes as it took account of the variations across subgroups [27]. However, these tools are usually applicable in a small sample size cDNA panel to determine the most stable reference genes [29]. Microarray is commonly used in relative quantification of gene expression because it affords speed and high-throughput [30, 31]. Due to its ability to analyze the expression of thousands of genes in a single experiment, the microarray presents a suitable resource for the analysis and identification of reference genes [13].

In this chapter, we present four cases of identifying reference genes by analyzing multiple microarray datasets using coefficient of variance (CV), followed by NormFinder analysis [27] whenever possible. In each dataset, the arithmetic mean and standard deviation were calculated for each gene and the CV was determined as a quotient of arithmetic mean and standard deviation. CV was used as a filter to select reference genes that showed low variation in expression. In the first and second cases, datasets from the different platform were used and direct comparison of probes by NormFinder [27] was not possible due to differences in probe sets. In the first case, we identified microtubule affinity-regulating kinase 3 (MARK3) as a suitable reference gene for mouse liver. In the second case, we identified 40S ribosomal protein S29 (Rps29) as a suitable reference gene for mouse testes and pancreas. In the third and fourth cases, datasets from the same platform were used, which enabled direct comparison. In the third case, we renormalized the entire set of microarray datasets and identified signal peptidase complex subunit 1 (SPCS1) and hydroxyacyl-CoA dehydrogenase beta subunit (HADHB) as suitable reference genes in human lungs. In the fourth case, we identified glucan biosynthesis protein G (mdoG) as suitable reference gene for 2 sub-strains of Escherichia coli K-12, MG1655 and W3110. Further analysis suggests that the identified reference genes are involved in fundamental biochemical processes.

Case 1: Finding Reference Genes for Mouse Liver Using Non-comparable Datasets [32]

Seven data sets of gene expression in the mouse liver under different conditions were obtained from Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo), National Center for Biotechnology Information (NCBI). These data sets used different platforms. Briefly, these data sets originated from the following studies:

1. LDL receptor-deficient mice fed on either a low fat diet or a high fat, western-style diet for 12 weeks with 12,548 genes located (GDS279)
2. Mice injected intraperitoneally with a cytokine and examined at 4 hours with 12,558 genes located (GDS280)
3. Analysis of animals fed a diet containing metformin, glipizide, rosiglitazone, or soy isoflavone extract and compared to hepatic gene expression profiles produced by long-term caloric restriction with 12,593 genes located (GDS1808)
4. Analysis of day 13.5 and 15.5 p38alpha-deficient hematopoietic mice with 45,101 genes located (GDS2693)
5. Comparison of embryonic day 13.5 and 15.5; analysis of livers from NMR-1 females fed human and chimpanzee diets for 2 weeks with 45,101 genes located (GDS3221)
6. Analysis of livers in C57BL/6 male mice maintained on a high-fat high-calorie diet supplemented with 0.04% resveratrol, a compound in red wine that extends lifespan of diverse species with 27,397 genes located (GDS2413)
7. Analysis of livers at 13 weeks following treatment with chemicals that test positive in two-year rodent cancer bioassays with 45,102 genes located (GDS2497).

A threshold of less than 0.1 CV was used to select stably expressed genes in each data set. The number of genes with CV ranging from <0.05 to <0.1 from each dataset were tabulated in Table 1. With GDS279 having the smallest number of gene probes compared to other data sets, it had limited the possible number of genes with CV <0.1 common in all data sets. Dataset GDS2497 has over 45102 probes, which resulted in the high number of genes having a CV lower than <0.1.

Despite the fact that there were close to 40,000 genes with CV <0.1 in datasets like GDS2497, GDS2693 and GDS3221, the expression level of many genes was not consistent under the different conditions, resulting in the small number of 39 genes with the CV of <0.1 present across the 7 datasets. Nephroblastoma Overexpressed gene (Nov), a gene that is essential for cell growth, was expressed with a low CV of 0.052 under the conditions of being p38alpha-deficient and hematopoietic (GDS2693), but under the conditions of being on different diets for 12 weeks (GDS279), it was expressed with a high CV of 0.561.

Analysis of the 7 datasets had identified 39 genes that were most stably expressed, with the criteria of having CV lesser than 0.1 across each of the conditions as shown in Table 2. Our results indicated that Microtubule Affinity-Regulating Kinase 3 (MARK3) was the most invariantly expressed gene with a CV of <0.08 - the lowest CV across all conditions, followed by 8 other genes with a CV of <0.09; namely Damage specific DNA Binding protein 1(Ddb1); D-fructose-1, 6-bisphosphate 1- phosphohydrolase 1 (Dnajc8); Fbp1, Ribosomal Protein L5 (Rpl5); Ribosomal Protein L10 (Rpl10), Ribosomal Protein L17 (Rpl7); Serine/arginine Repetitive Matrix 1(Srrm1) and Ubiquitin-Like 5 (Ubl5).

Of these 9 genes, at least 5 genes (MARK3, Ddb1, Rpl5, Rpl7, Srrm1 and Ubl5) are considered common housekeeping genes. Rpl5, Rpl10 and Rpl7 are ribosomal proteins, and ribosomal protein genes are constitutively expressed in all cell types and are essential for the biogenesis of new ribosomes for the synthesis of proteins, suggesting that ribosomal proteins are more likely to be stably expressed because any increase or decrease of the gene level will result in an abnormal amount of ribosomal level, which in turn may result in mutated or

disabled cells as shown by Thorrez et al., [33]. In addition to coding for a ribosomal protein, Srrm1 is also involved in numerous pre-mRNA processing events [34]. Undoubtedly, varying availability of proteins in the translational process may result in variability in translational efficiency [35]. Taken together, it is likely that genes coding for proteins that are involved in translation of mRNA are more likely to be constitutively expressed at a constant level.

Although it has been shown that some housekeeping genes are stably expressed [36], other commonly used housekeeping genes like GAPDH was shown to have high fluctuations in the gene expression level. Our results show that CV of gene expression of GAPDH is 10% or more. Hence, GAPDH is not on our list of good housekeeping genes (Table 2). On the other hand, Mark3 appears to be stably expressed in mouse liver, suggesting that it is a suitable reference gene for mouse liver.

Table 1. Number of genes with coefficient of variation (CV) under values of 0.05, 0.06, 0.07, 0.08, 0.09 and 0.1 in their respective datasets

	CV Threshold					
	<0.05	**<0.06**	**<0.07**	**<0.08**	**<0.09**	**<0.1**
GDS279	1	10	37	86	169	310
GDS280	302	940	2060	3479	4928	6177
GDS1808	1	8	40	130	283	524
GDS2413	85	97	112	124	136	146
GDS2497	34075	39761	42263	43445	44018	44362
GDS2693	29494	35617	39607	41889	43020	43606

Table 2. Specific genes and their CV values

Genes	**CV**	**Genes**	**CV**
MARK3	<0.08	Eif2s3x	<0.1
4833439L19Rik	<0.09	Epn1	<0.1
9530068E07Rik	<0.09	H2-K1	<0.1
Ddb1	<0.09	Hdgf	<0.1
Dnajc8	<0.09	Mark2	<0.1
Fbp1	<0.09	Myo1b	<0.1
Rpl5	<0.09	Rpl3	<0.1
Rpl10	<0.09	Pcid2	<0.1
Rpl7	<0.09	Pex14	<0.1
Srrm1	<0.09	Pgm2	<0.1
Ubl5	<0.09	Pigs	<0.1
2410166I05Rik	<0.1	Prdx1	<0.1
Ahcyl1	<0.1	Rps3	<0.1
Ap3b1	<0.1	Rps3a	<0.1
Arpc5	<0.1	Stard3	<0.1
Bscl2	<0.1	Sucla2	<0.1
Cd151	<0.1	Trpc4ap	<0.1
Cox7c	<0.1	Zdhhc9	<0.1

Cstf1	<0.1	Zwint	<0.1
Cyb5	<0.1		

Case 2: Finding Reference Genes for Mouse Testes and Pancreas Using Non-Comparable Datasets [37]

Fifteen datasets were obtained from GEO, NCBI; of which, 8 were from mice testes and 7 were from mice pancreas. Briefly, the studies conducted with the datasets were as follows:

1. Analysis of mice testis from day 1 to adult to determine transcripts expressed specifically by male germ cells late in post-meiotic spermatogenesis (GDS410)
2. Spermatogenesis time course generated from mice testis collected from birth through adulthood to provide insight into genes implicated in maturation, maintenance, and function of testis and the integrated process of spermatogenesis (GDS607)
3. Analysis of CD4+ automimmune T and CD45+ hematopoeitic cells from non-obese diabetic (NOD) and transgenic mice with the BDC2.5 T cell receptor (TCR) from a diabetogenic T cell (GDS828)
4. Analysis of pancreatic islets from transgenic mice after treatment with 200 mg/kg cyclophosphamide (CY) at various time points up to 3 days (GDS1056)
5. Expression profiling of male germ cells lacking the telomerase RNA component, Terc, or the DNA repair/telomere binding protein, Ku86, or both to identify processes affected by dysfunctional telomeres (GDS1248)
6. Analysis of testes from Sertoli cell-selective androgen receptor knockouts (SCARKO) between postnatal day 8 and 20 to provide insight into molecular consequences of selective absence of androgen action in Sertoli cells at onset of spermatogenesis (GDS1367)
7. Analysis of the pancreas from 3-week-old transgenic mcie that express a diabetogenic T cell receptor on the NOD or NOD.scid genetic background (GDS1533)
8. Analysis of the pancreas from transcription factor Mist1 knockouts (Mist1KO) injected with caerulein to identify genes that may contribute to susceptibility and initiation of pancreatitis (GDS1731)
9. Expression profiling of testes from *Mus musculus* subspecies castaneus, domesticus, musculus, and a strain from Iran (GDS1884)
10. Analysis of transgenic pancreatic islet beta cells after activation of Myc for 24 hours or 21 days followed by deactivation for up to 6 days to elucidate the molecular basis of Myc-driven tumorigenesis in vivo (GDS2025)
11. Expression profiling of normal whole testes at various stages of development from gestation day 11 to postnatal day 2 (GDS2098)
12. Analysis of the epididdymis of animals at various ages from gestational day 12 to post-natal day 2 to provide insight into the molecular events involved in testicular development (GDS2202)
13. Analysis of ribonucleoprotein (RNP) and polysome fractions of prepuberal and adult testes to provide insight into the translation state of individual mRNAs as spermatogenesis proceeds (GDS2277)

14. Analysis of MIN6 pancreatic beta-cells treated for up to 24 hours with various concentrations of human islet amyloid polypeptide (GDS2945)
15. Analysis of pancreatic tissues from NMRI animals from embryonic day 12.5 to 16.5 to provide insight into the molecular mechanisms underlying the development of the pancreas (GDS2950).

One hundred probes with the lowest CV from each dataset identified and grouped based on the organ used, namely testes and pancreas. Non-unique probes in each organ list were isolated, indicating that they were the least variant genes for that platform and organ, supported by several datasets. Common non-unique probes, indicating low CV in both organs and supported by multiple data sets in each organ, were identified. The average CV and percentile rank were calculated to allow for comparison.

A total of 5 genes were recorded as being shown to have a stable variation: Rps7 (40S ribosomal protein S7), Rps28 (40S ribosomal protein S28), Rps29 (40S ribosomal protein S29), Rpl31 (60S ribosomal protein L31), Rpl37a (60S ribosomal protein L37a) and MARK3. The gene with the lowest CV and percentile values was Rps29 (Table 3).

All of the 5 genes identified to have a low variance were ribosomal proteins. In addition, the least variable gene from the commonly used housekeeping gene set was the ribosomal protein, Rpl13a. Thus, it can be concluded from the results that genes encoding for ribosomal proteins appear to be the most stably expressed. This coincides with results from previous findings also indicating that genes coding for ribosomal proteins have the least variance [33]. Ribosome has the vital role in cell maintenance, growth, survival and function. Cells with an abnormal level of ribosomes may result in cell mutation or death [33]. This suggests that most cells should possess comparable levels of ribosomes; hence, explaining the low expression variance of genes coding for ribosomal proteins.

Case 3: Finding Reference Genes for Human Lung Using Comparable Datasets [38]

Fourteen microarray data sets studying human lung epithelial from GEO were used. All of which employed Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) containing 54,676 probes. This allows for comparisons across different data sets. The fourteen data sets were:

1. Bronchial epithelial cells at 4 and 24 hours following treatment with respiratory syncytial virus (GDS2023)
2. Airway epithelial cells of phenotypically normal smokers (GDS2486 and GDS2491)
3. Airway epithelia from healthy individuals 7 and 14 days following injury by epithelial denudation (GDS2495)
4. Epithelial and mesothelial lung cell lines at various time points up to 7 days after exposure to asbestos (GDS2604)
5. Twenty-eight day air-liquid cultured airway epithelial cells (GDS2615)
6. Analysis of resveratrol-treated lung carcinoma A549 cells (GDS2966).

7. Analysis of house dust mite (HDM) extract exposed H292 bronchial epithelial cells (GDS3003)
8. Bronchial epithelial cells exposed to cigarette smoke from a typical full flavor brand for up to 24 hours (GDS3493)
9. Bronchial epithelial cells exposed to cigarette smoke from a typical light flavor brand for up to 24 hours with (GDS3494)
10. Lung adenocarcinoma CL1-5 cells overexpressing Claudin-1 (CLDN1) (GDS3510)
11. Comparison of non-small cell lung cancer histological subtypes: adenocarcinomas (AC) and squamous cell carcinomas (SCC) (GDS3627)
12. Analysis of normal lung WI-38 fibroblasts exposed to various concentrations of the carcinogen benzo[a]pyrene-diol epoxide (BPDE) (GDS3706)
13. Analysis of A549 epithelial cells treated for up to 72 hours with TGF-beta (GDS3710).

The data sets were normalized by arithmetic mean transformation and Z transformation to construct parallel and comparable data sets based on the method described in [39]. In each dataset, the intensity value for each probe ($Probe_{Initial}$) was used to compute the average probe intensity for each dataset ($\mu_{InitialProbe}$). By assuming the standard arithmetic mean of all the probes as 1000 for the overall combined microarray data sets ($\mu_{Assumed}$), the correction factor for the average probe intensity for each data sets was constructed as a quotient of $\mu_{Assumed}$ and $\mu_{InitialProbe}$, i.e. Correction Factor = $\mu_{Assumed}/\mu_{InitialProbe}$. The transformed intensity of each original probe ($Probe_{Transformed}$) is then calculated using the equation, $Probe_{Transformed}$ = $Probe_{Initial}$ × Correction Factor. The $Probe_{Transformed}$ was then used to construct Z-score by the equation, $Z_{score} = (Probe_{Transformed} - \mu_{Assumed})/ SD_{Dataset}$, where $SD_{Dataset}$ is the standard deviation of the original dataset from which the probe ($Probe_{Initial}$) originates. The Z-scores for each probe across different original data sets would then be comparable.

By constructing the CV values for all the probes of the chosen reference genes from the fourteen data sets, the mean CV values for each reference gene was generated. Using the Z-transformed dataset, 10% of the probes with the lowest CV were isolated and filtered. In this case, the CV would be the quotient of standard deviation of the Z-score to the arithmetic mean of the Z-score. For example, if geneA is represented in the microarray as 3 probes and all 3 probes are found in list of 10% lowest CVs, geneA is retained. On the other hand, if geneB is represented in the microarray as 3 probes and one of the probe is not in the list of 10% lowest CV, geneB is removed from the list. Only probes with gene symbols were included in the list. The filtered list was then analyzed with NormFinder [27] using the normalized microarray data as expression values. Due to the limitation that NormFinder has for analyzing zero and negative values, the normalized expression values of the fourteen data sets, which were lower than 0.0001 or close to zero value, were transformed into the absolute values or 0.0001, respectively.

Table 3. The 5 identified invariant genes, and their CV and percentile values for both endocrine glands

Gene	Average CV	Average Percentile	Minimum CV	Maximum CV	Minimum Percentile	Maximum Percentile
Rps7	0.188608	0.218980	0.007615	1.075255	0.000619	0.985346

Rps29	0.136192	0.133532	0.006590	0.357252	0.000266	0.777627
Rps28	0.161726	0.230025	0.009225	0.603931	0.000088	0.916240
Rpl31	0.306197	0.295305	0.005310	1.181695	0.000220	0.965566
Rpl37a	0.192236	0.229087	0.009193	0.459498	0.000442	0.740711

Table 4. Stability of housekeeping genes generated by NormFinder [27].

Gene Symbols	NormFinder Stability Index	Rank
SPCS1	0.326	1
HADHB	0.355	2
EIF2A	0.691	371
PPIB	0.862	630
PPIB	0.867	634
RPL13A	0.651	305
RPL13A	0.811	548
RPL13A	0.832	587
RPL13A	0.911	689
RPL13A	0.950	730
RPL32	0.739	454
RPL4	0.723	428
RPL4	0.789	518
RPL4	0.805	541
RPS4X	0.777	499
RPS4X	0.838	592

By extracting the pool of Z-transformed probes intensities with the lowest 10% CV, the remaining 5,458 probes (out of 54,676 probes) were further selected to eliminate undefined genes. The selected 743 genes were input in NormFinder [27] to generate the stability value for each gene as a direct measure of the estimated expression variation; genes were then ranked accordingly. SPCS1 and HADHB were found to have the lowest CV values, demonstrated the lowest stability index, and were subsequently ranked as first and second, respectively.

By extracting the 10% probes with the lowest CV values from the transformed Z-score data, 5,458 probes were extracted from the original 54,676 probes. After removing gene probes without a specific gene name, the total probes number left was 2,213. Among these 2,213 probes, 743 genes fell within the lowest 10% CV, which were encoded by 932 probes. The analysis result showed that among the chosen 20 housekeeping genes from previous studies [3, 17, 19, 20, 22, 23, 32, 40-42], only EIF2A, PPIB, RPL4, 60S ribosomal protein L13a (RPL13A), 60S ribosomal protein L32 (RPL32) and 40S ribosomal protein S4 (RPS4X) were found in the lowest 10% CV subset. Expression stability analysis using NormFinder [27] showed that signal peptidase complex subunit 1 (SPCS1) and hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase, beta subunit (HADHB) had the highest stability of rank 1 and 2 with a stability index of 0.326 and 0.355 respectively, followed by 71 genes with a stability index of 0.360. EIF2A, PPIB, RPL4, RPL13A, RPL32 and RPS4X had the rank ranging from 305 to 730 (Table 4). The mean Z-score of the probes within the lowest 10% CV ranges from -0.4073 to 11.7692. The mean Z-scores of HADHB and SPCS1 were 1.8419 (61.8 percentile) and 2.8516 (75.8 percentile) respectively.

The standard deviation of the CV values and NormFinder stability index were 0.077 and 0.265, respectively. Due to these values varying by more than 3 times the standard deviation, homoscedasticity (constant variance) was not assumed. Thus, Spearman's rank correlation coefficient was carried out to determine the correlation of stability index by NormFinder and CV values, which showed that the sum of d2 is 85109486 (Spearman's rank correlation coefficient = 0.369) and the p-value was 1.79×10-31. Since the p-value was less than 0.01, the null hypothesis is rejected, indicating that there is correlation between the NormFinder stability index and CV values.

Our results suggested that SPCS1 and HADHB were considerably stable among the datasets analyzed and can be used as suitable calibrating genes in human lung studies for both healthy and perturbed tissues. The mean Z-scores of HADHB and SPCS1 were at the 61.8 percentile and 75.8 percentile, respectively, suggesting that these genes can be used as suitable reference genes for genes with average to high expression. Further studies are needed to determine their validity for genes with low expression.

The SPCS1 gene is located on chromosome 3 at 3p21.1 and functions as signal peptidase complex in the signal sequences cleavage of most secretory and membrane proteins [43]. Most secretory proteins required to be translocated into endoplasmic reticulum (ER) membrane in order to allow the proteins to fold and assemble in a proper way before they are transported to the Golgi apparatus. Proteins that fail to assemble or fold into their native state will be translocated back across ER membrane to cytoplasm and undergo degradation [44]. The signal sequences on secretory proteins must be cleaved during protein synthesis and SPCS1 plays a role in cleavage of signal sequences of most secretory proteins to enable their translocation across the ER membrane [43].

HADHB is located on chromosome 2 at 2p23 and is involved in mitochondrial beta-oxidation of long chain fatty acids [45]. A previous study [46] had shown that different organs have different tendencies in fatty acid distribution, as liver is served as the largest reservoir for fatty acids followed by brain and lung. Since lung is one of the reservoirs for fatty acid, HADHB, a lipase, will be likelihood expressed more stably in the lung in order to degrade fatty acids and generate ATP. However, these two genes have not been used as reference genes in gene quantification studies.

Case 4: Finding Reference Gene for *Escherichia Coli* Using Comparable Datasets [47]

Four data sets of *E. coli* K-12 using Affymetrix *E. coli* Antisense Genome Array (GPL199) were obtained from Gene Expression Omnibus, National Centre for Biotechnology Information in which 3 were from *E. coli* K-12 sub-strain MG1655 and 1 from sub-strain W3110. Briefly, the studies conducted with the data sets are as follows:

1. MG1655 grown in either aerobic or anaerobic conditions, deleted for transcriptional regulators in oxygen response, used to validate a computational model of transcriptional and metabolic networks (GDS680)
2. MG1655 cells cultured aerobically in several media with varied carbon sources, including glucose, glycerol, succinate, L-alanine, acetate, and L-proline (GDS1099)

3. Analysis of derivatives of strain 1655: wild-type, fur mutant, and wild-type with added $FeSO_4$, induced to overexpress RyhB, a noncoding RNA regulated by the Fur repressor protein (GDS1494)
4. W3110 cells exposed to acidic, neutral, or alkaline pH in order to study acid and base responses (GDS1827).

Table 5. Weighted mean CV and NormFinder stability index of 39 invariant genes across 3 data sets (MG1655)

Gene Symbol	Gene Name	Weighted CV values	Weighted NormFinder Stability Index
mdoG	glucan biosynthesis protein G	0.099	0.082
dapA	dihydrodipicolinate synthase	0.106	0.090
crp	DNA-binding transcriptional dual regulator	0.106	0.102
hslV	peptidase component of the HslUV protease	0.111	0.105
mrdB	cell wall shape-determining protein	0.101	0.114
fucU	L-fucose mutarotase	0.107	0.117
yjgP	LPS transport (lptF)	0.105	0.117
yigC	3-octaprenyl-4-hydroxybenzoate decarboxylase	0.126	0.117
sun	16S rRNA m(5)C967 methyltransferase, S-adenosyl-L-methionine-dependent	0.130	0.119
gor	glutathione oxidoreductase	0.117	0.126
hflB	ATP-dependent metalloprotease	0.127	0.130
yqiB	predicted dehydrogenase	0.125	0.134
murG	N-acetylglucosaminyl transferase	0.124	0.134
yrbG	predicted calcium/sodium:proton antiporter	0.122	0.134
yejK	nucleotide associated protein	0.120	0.141
yfgA	cytoskeletal protein required for MreB assembly	0.118	0.142
hflX	putative GTPase HflX	0.105	0.142
spoT	bifunctional (p)ppGpp synthetase II/ guanosine-3',5'-bis pyrophosphate 3'-pyrophosphohydrolase	0.117	0.143
holC	DNA polymerase III, chi subunit	0.134	0.144
xerD	site-specific tyrosine recombinase	0.114	0.146
tolB	periplasmic protein	0.115	0.146
yheS	fused predicted transporter subunits of ABC superfamily: ATP-binding components	0.110	0.146
ntpA	dihydroneopterin triphosphate pyrophosphatase	0.118	0.147
yabB	conserved protein, MraZ family	0.115	0.148
lolA	chaperone for lipoproteins	0.117	0.153
yggD	predicted DNA-binding transcriptional regulator	0.116	0.153
pnp	polynucleotide phosphorylase/polyadenylase	0.110	0.155
yrbB	ABC transporter maintaining OM lipid asymmetry, cytoplasmic STAS component	0.123	0.156
rnc	RNase III	0.117	0.157
xerC	site-specific tyrosine recombinase	0.138	0.160
rfaF	ADP-heptose:LPS heptosyltransferase II	0.120	0.161
yigP	conserved protein, SCP2 family	0.122	0.164
gyrB	DNA gyrase, subunit B	0.126	0.164
nagC	DNA-binding transcriptional dual regulator,	0.132	0.165

	repressor of N-acetylglucosamine		
nrdR	Conserved protein	0.118	0.168
hemD	uroporphyrinogen III synthase	0.108	0.169
pheT	phenylalanine tRNA synthetase, beta subunit	0.124	0.171
frr	ribosome recycling factor	0.129	0.173
cls	cardiolipin synthase 1	0.129	0.181

For 4,631 genes, the top 10% with the lowest CV from each dataset were listed. The intersection between the 3 MG1655 datasets (GDS680, GDS1099, and GDS1494) was identified and analysed using NormFinder [27] to rank the stability of these genes. A weighted stability index for each gene was then calculated from the NormFinder stability index, and an average determined from the NormFinder stability indices multiplied by sample number.

A threshold of less than 10% CV was used to select the stably expressed genes across the three data sets GDS 680, 1099 and 1494 (MG1655). A total of 39 genes with consistently low variance were identified (Table 5) with the weighted CV values ranging from 0.099 to 0.138. mdoG was found to be most stable with both the lowest weighted CV value and weighted NormFinder Stability Index for MG1655. In GDS 1827 (W3110), mdoG was the most stable in the dataset, with a CV of 0.088 and NormFinder Stability Index of 0.078. The highest CV in GDS 1827 is 0.791 for hslV (peptidase component of the HslUV protease). Although it is not a well-established reference gene, our results suggest that mdoG may be a suitable reference gene across both *E. coli* strains W3110 and MG1655 and that mdoG may also be suitable for use as reference genes in other strains of *E. coli* K-12.

The mdoG gene has been shown to be involved in the formation of the β-1,6 glucose linkage [48] and in the periplasmic release of newly synthesized osmoregulated periplasmic glucans [49, 50] needed for the bacterial cell wall. Thus, it is plausible that the expression of mdoG is needed during binary fission. As *E. coli* divides rapidly, constant synthesis of the cell wall is needed, necessitating constitutive mdoG expression.

CONCLUSION

The accuracy of most quantitative gene expression studies relied on the expressional stability of using the reference genes [5]. Previous studies [3, 22, 40, 51] had shown that not all commonly-used housekeeping genes such as GAPDH and beta-actin can be utilized as their gene expression may vary under different conditions [3] and the likelihood of suitable reference genes across many organisms and organs is low [7]. In addition, a recent study [52] examined the published microarray datasets from the livers of two closely related species of squirrels at comparable physiological states did not find suitable liver-specific reference genes. This suggests a need to identify reference genes for each organism and tissue. High-throughput transcriptome profiling technologies, such as microarray, presents a suitable resource for the analysis and identification of reference genes [13].

The advantage of CV over other more complex methods, such as NormFinder [27], is its capability to analyze as large a number of samples as required as the number of calculations increases proportionally to the sample size, resulting in linear complexity. NormFinder uses residual analysis between sample subgroup variation and the overall variation of the

expression dataset to evaluate the variation contributed by each gene in the entire dataset [27]. Thus, the computational complexity of NormFinder increases exponentially as the number of samples increase; hence, NormFinder is only able to work with a small number of genes within reasonable time and computational resources. Given the advantageous ability of CV to process large amounts of data such as those derived from microarrays, it is plausible that CV can be used as a weaker filter for a broad category of genes with low expression variation, followed by stronger statistical analysis by NormFinder [27] to identify suitable reference genes.

CV does not take into consideration systematic variations such as those introduced by inaccuracy in sample preparation [53]. Thus, CV can only be used as a categorical filter of gene expression into a striation of classes. NormFinder is employed as the fine diagnostic tool to further determine the stability of gene expression within the class of genes with low expression variation to identify the most stably expressed gene. In the four case studies mentioned above, the identification of reference genes in multiple microarray datasets was achieved by analyzing the CV using either comparable or non-comparable microarray datasets on eukaryotic or prokaryotic biological samples. In the first case study, genes with the CV value of less than 0.1 were selected for further analysis while only the top 10% of genes that showed the lowest CV were chosen for downstream NormFinder [27] ranking. A threshold of 0.1 CV or 10% of lowest CV were set in these case studies. In Cases 3 and 4, the threshold functioned as a basic filter to limit the data for further processing by NormFinder [27]. By setting the threshold at 0.1 or extracting data only from the lowest 10% CV, many of the data were filtered out and left with the genes that showed low expression variation and able to be processed by NormFinder [27]. However, in the studies of comparable microarray datasets (Case 3 and 4) or different organs (Case 2), the threshold setting may exclude genes that showed low variation in “Dataset A” but not in “Dataset B” or in “Organ A” but not in “Organ B”, or vice versa. The cross analysis of the data using the threshold in these scenarios will give rise to the identification of genes with the highest stability among the datasets. Nonetheless, in each dataset, there might be other genes that show lowest deviation in the gene expression. This could be one of the disadvantages that resulted from using CV to cross analyze the data from different datasets or different organs.

It is plausible that the results of CV and NormFinder [27] analysis may be correlated as both methods had been used to identify reference genes within a dataset [3]. Our results from Case 3 suggested that CV and NormFinder [27] stability were significantly correlated to each other (p-value = 1.79×10-31). However, the strength of this correlation is difficult to establish as the significance in p-value did not indicate the correlation strength. At the same time, the comparison between CV, NormFinder [27], geNorm [25] and BestKeeper© [26] requires a more comprehensive study using a datasets spanning across different tissues and organisms.

In all the cases presented, we utilized CV as a filter to screen all the genes of differential expression to reduce the number of genes/probes to be processed by NormFinder. However, it is possible that some stably-expressed genes may be excluded at this stage. Hence, using a 2-step process of CV and NormFinder may not be optimal. Due to this limitation of CV processing, a novel statistical tool with statistical stringency (measured by specificity and sensitivity) as NormFinder [27], geNorm [25] and BestKeeper© [26] but at the same time can extensively process the data as CV is neccessary. This tool will allow the processing all the

data within the microarray sets in a larger scale, thereby taking account of all the available data. This is likely to be statistically stronger than a 2-step approach of CV-NormFinder.

A possible way to achieve increased statistical stringency be expanding on the functions of NormFinder, which is usually used to process small sample size [27, 29, 54]. Hence, in this proposed extended version of NormFinder, the raw data in the microarray dataset are separated into to a few small manageable parts to run, for instance, “Run #1” and “Run #2” (Figure 1). In each run, NormFinder will generate a list of stability indices for the genes or probes. In the two lists generated, one gene or probe sample is overlapped across the two runs (Figure 1). By merging or normalizing the stability index for each particular probe in two runs, a new list of NormFinder stability indices that consist of all the genes or probe samples in both runs can be constructed. This process can be repeated to merge multiple NormFinder runs across large datasets using sets of overlapping samples, resulting in an overall comparison of expressional stability across entire genome.

Another possible method for microarray scale analysis might be the use of pairwise correlation of expression values in the entire dataset. Keng et al., [52] reported that CV/NormFinder-identified stably expressed genes demonstrated significantly less expressional correlation to other randomly selected genes. This agrees with a recent study by Ling et al., [55] suggesting that orthologous genes demonstrate higher degree of expressional correlation compared to random pairs of genes. Moreover, pairwise correlation of expression values may be the basis of a tool to identify reference genes from large datasets.

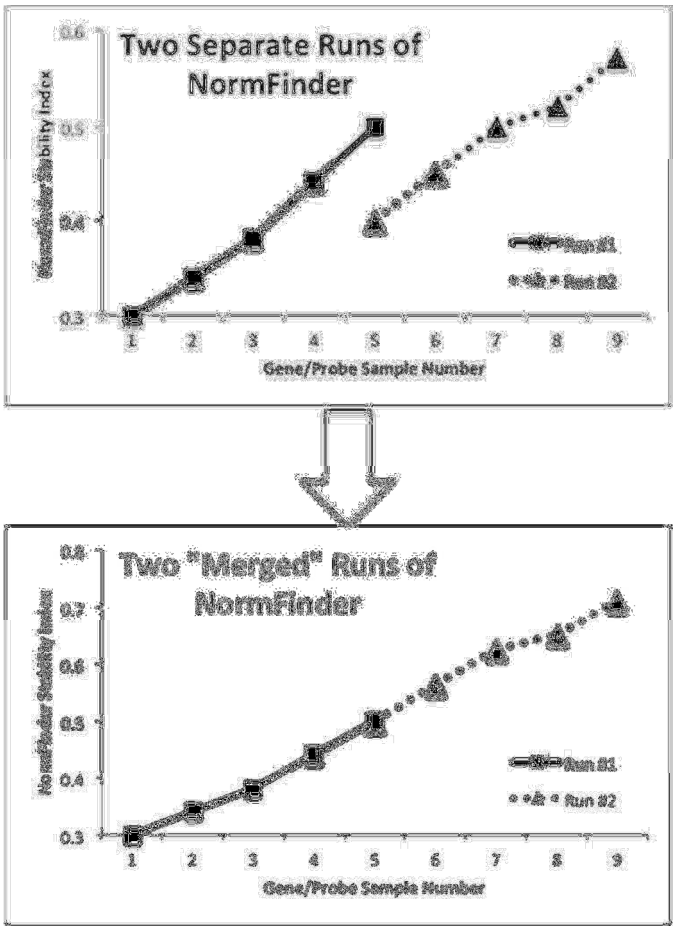

Figure 1. Two runs are performed on the same microarray dataset that has been fragmented into smaller manageable sample size to be analyzed by NormFinder. Each run generates a list of stability index for the gene/ probe sample. By merging or normalizing the stability index of the overlapped gene (Gene number "5"), a new list of stability index can be constructed that compromises all genes or probe samples within the microarray dataset.

In this chapter, we are not suggesting that a standardized method should be employed in the process of analyzing and identifying a suitable reference gene. Instead, more comprehensive alternatives should be incorporated with a higher tolerance rate in sample size to allow an overall comparison of the gene expression in order to select the most appropriate housekeeping genes in each case.

ACKNOWLEDGMENTS

The authors will like to thank the following colleagues for their comments on this manuscript: JSH Oon (University of Queensland, Australia), QMA Xie (University of Queensland, Australia), PCK Au (CSIRO, Australia), CC Goh (University of Virginia, USA), and YZ Koh (University of Portsmouth, UK). SSH, OYC, BMK, CC, CWL, WL, QC, EJA,

and YL contributed to this work as part of the Science Mentorship Programme under Gifted Education Branch, Ministry of Education, Singapore.

REFERENCES

[1] Chey, S; Claus, C; Liebert, UG. Validation and application of normalization factors for gene expression studies in rubella virus-infected cell lines with quantitative real-time PCR. *J Cell Biochem*, 2010, 110(1), 118-128.

[2] Bustin, SA; Nolan, T. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J Biomolecular Tech*, 2004, 15(3), 155-166.

[3] Maccoux, LJ; Clements, DN; Salway, F; Day, PJ. Identification of new reference genes for the normalisation of canine osteoarthritic joint tissue transcripts from microarray data. *BMC Mol Biol*, 2007, 8, 62.

[4] Airoldi, EM; Huttenhower, C; Gresham, D; Lu, C; Caudy, AA; Dunham, MJ, et al. Predicting cellular growth from gene expression signatures. *PLoS Comput Biol*, 2009, 5(1), e1000257.

[5] Gubern, C; Hurtado, O; Rodriguez, R; Morales, JR; Romera, VG; Moro, MA; et al. Validation of housekeeping genes for quantitative real-time PCR in in-vivo and in-vitro models of cerebral ischaemia. *BMC Mol Biol*, 2009, 10:57.

[6] Czechowski, T; Stitt, M; Altmann, T; Udvardi, MK; Scheible, WR. Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol*, 2005, 139(1), 5-17.

[7] Dundas, JB; Ling, MHT. Reference genes for measuring mRNA expression. *Theory in Biosci*, 2012, 131, 215-223.

[8] Jain, M; Nijhawan, A; Tyagi, AK. Validation of housekeeping genes as internal control for studying gene expression in rice by quantitative real-time PCR. *Biochem Biophys Res Commun*, 2006, 345, 646-651.

[9] Nicot, N; Hausman, JF; Hoffmann, L. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *J Exp Botany*, 2005, 56, 2907-2914.

[10] Gibson, UE; Heid, CA; Williams; PM. A novel method for real time quantitative RT-PCR. *Genome Res*, 1996, 6, 995-1001.

[11] Sturzenbaum, SR; Kille, P. Control genes in quantitative molecular biological techniques: the variability of invariance. *Comp Biochem Physiol B Biochem Mol Biol* 2001, 130, 281-289.

[12] Takle, GW; Toth, IK; Brurberg, MB. Evaluation of reference genes for real-time RT-PCR expression studies in the plant pathogen Pectobacterium atrosepticum. *BMC Plant Biol*, 2007, 7, 50.

[13] Noriega, NC; Kohama, SG; Urbanski, HF. Microarray analysis of relative gene expression stability for selection of internal reference genes in the rhesus macaque brain. *BMC Mol Biol*, 2010, 11, 47.

[14] 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, 1343-1349.

[15] Glare, EM; Divjak, M; Bailey, MJ; Walters, EH. beta-Actin and GAPDH housekeeping gene expression in asthmatic airways is variable and not suitable for normalizing mRNA levels. *Thorax*, 2002, 57(9), 765-770.

[16] Gutierrez, L; Mauriat, M; Guénin, S; Pelloux, J; Lefebvre, JF; Louvet, R; 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*, 2008, 6, 609-618.

[17] Gur-Dedeoglu, B; Konu, O; Bozkurt, B; Ergul, G; Seckin, S; Yulug, IG. Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. *Oncol Res*, 2009, 17(8), 353-365.

[18] Kosir, R; Acimovic, J; Golicnik, M; Perse, M; Majdic, G; Fink, M; et al. Determination of reference genes for circadian studies in different tissues and mouse strains. *BMC Mol Biol,* 2010, 11, 60.

[19] Shen, Y; Li, Y; Ye, F; Wang, F; Lu, W; Xie, X. Identification of suitable reference genes for measurement of gene expression in human cervical tissues. *Anal Biochem*, 2010, 405(2), 224-229.

[20] Rho, HW; Lee, BC; Choi, ES; Choi, IJ; Lee, YS; Goh, SH. Identification of valid reference genes for gene expression studies of human stomach cancer by reverse transcription-qPCR. *BMC Cancer*, 2010, 10, 240.

[21] Brattelid, T; Winer, LH; Levy, FO; Liestol, K; Sejersted, OM; Andersson, KB. Reference gene alternatives to Gapdh in rodent and human heart failure gene expression studies. *BMC Mol Biol*, 2010, 11, 22.

[22] Lallemant, B; Evrard, A; Combescure, C; Chapuis, H; Chambon, G; Raynal, C; et al. Reference gene selection for head and neck squamous cell carcinoma gene expression studies. *BMC Mol Biol*, 2009, 10, 78.

[23] Li, YL; Ye, F; Hu, Y; Lu, WG; Xie, X. Identification of suitable reference genes for gene expression studies of human serous ovarian cancer by real-time polymerase chain reaction. *Anal Biochem*, 2009, 394(1), 110-116.

[24] Infante, C; Matsuoka, MP; Asensio, E; Canavate, JP; Reith, M; Manchado, M. Selection of housekeeping genes for gene expression studies in larvae from flatfish using real-time PCR. *BMC Mol Biol*, 2008, 9, 28.

[25] Vandesompele, J; De Preter, K; Pattyn, F; Poppe, B; Van Roy, N; De Paepe, A; et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*, 2002, 3(7), RESEARCH0034.

[26] 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. *Biotechnol Lett*, 2004, 26(6), 509-515.

[27] Andersen, CL; Jensen, JL; Orntoft, 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 Res*, 2004, 64(15), 5245-5250.

[28] He, JQ; Sandford, AJ; Wang, IM; Stepaniants, S; Knight, DA; Kicic, A; et al. Selection of housekeeping genes for real-time PCR in atopic human bronchial epithelial cells. *Eur Respir J*, 2008, 32(3), 755-762.

[29] Ren, S; Zhang, F; Li, C; Jia, C; Li, S; Xi, H; et al. Selection of housekeeping genes for use in quantitative reverse transcription PCR assays on the murine cornea. *Mol Vis*, 2010, 16, 1076-1086.

[30] De, RK; Ghosh, A. Interval based fuzzy systems for identification of important genes from microarray gene expression data: Application to carcinogenic development. *J Biomed Inform*, 2009, 42(6),1022-1028.

[31] Galiveti, CR; Rozhdestvensky, TS; Brosius, J; Lehrach, H; Konthur, Z. Application of housekeeping npcRNAs for quantitative expression analysis of human transcriptome by real-time PCR. *RNA*, 2010, 16(2), 450-461.

[32] Chia, CY; Lim, CW; Leong, WT; Ling, MH. High expression stability of microtubule affinity regulating kinase 3 (MARK3) makes it a reliable reference gene. *IUBMB Life*, 2010, 62(3), 200-203.

[33] Thorrez, L; Van Deun, K; Tranchevent, LC; Van Lommel, L; Engelen, K; Marchal, K; et al. Using ribosomal protein genes as reference: a tale of caution. *PLoS One*, 2008, 3(3), e1854.

[34] Le Hir, H; Maquat, LE; Moore, MJ. Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions. *Genes Dev*, 2000, 14, 1098-1108.

[35] Meng, Z; Jackson, NL; Choi, H; King, PH; Emanuel, PD; Blume, SW. Alterations in RNA-binding activities of IRES-regulatory proteins as a mechanism for physiological variability and pathological dysregulation of IGF-IR translational control in human breast tumor cells. *J Cell Physiol*, 2008, 217, 172-183.

[36] Coulson, DTR; Brockbank, S; Quinn, JG; Murphy, S; Ravid, R; Irvine, GB; et al. Identification of valid reference genes for the normalization of RT qPCR gene expression data in human brain tissue. *BMC Mol Biol*, 2008, 9, 46.

[37] Chu, QH; Lin, YJ; Ang, EJG; Ling, MHT. Identification of transcriptional invariant genes in mouse endocrine glands from microarray data. *Proceedings of the 16th Youth Science Conference*, 2010, Singapore.

[38] Too, IH; Ling, MH. Signal peptidase complex subunit 1 (SPCS1) and hydroxyacyl-CoA dehydrogenase beta subunit (HADHB) are suitable reference genes in human lungs. *ISRN Bioinformatics*, 2012, 2012, Article ID 790452.

[39] Cheadle, C; Vawter, MP; Freed, WJ; Becker, KG. Analysis of microarray data using Z score transformation. *J Mol Diagn*, 2003, 5(2), 73-81.

[40] Nguewa, PA; Agorreta, J; Blanco, D; Lozano, MD; Gomez-Roman, J; Sanchez, BA; et al. Identification of importin 8 (IPO8) as the most accurate reference gene for the clinicopathological analysis of lung specimens. *BMC Mol Biol*, 2008, 9, 103.

[41] Liu, DW; Chen, ST; Liu, HP. Choice of endogenous control for gene expression in nonsmall cell lung cancer. *Eur Respir J*, 2005, 26, 1002-1008.

[42] Saviozzi, S; Cordero, F; Lo Iacono, M; Novello, S; Scagliotti, GV; Calogero, RA. Selection of suitable reference genes for accurate normalization of gene expression profile studies in non-small cell lung cancer. *BMC Cancer*, 2006, 6, 200.

[43] Kalies, KU; Hartmann, E. Membrane topology of the 12- and the 25-kDa subunits of the mammalian signal peptidase complex. *J Biol Chem*, 1996, 271(7), 3925-3929.

[44] Swanton, E; Bulleid, NJ. Protein folding and translocation across the endoplasmic reticulum membrane. *Mol Membr Biol*, 2003, 20(2), 99-104.

[45] Aoyama, T; Wakui, K; Orii, KE; Hashimoto, T; Fukushim, Y. Fluorescence in situ hybridization mapping of the alpha and beta subunits (HADHA and HADHB) of human mitochondrial fatty acid beta-oxidation multienzyme complex to 2p23 and their evolution. *Cytogenet Cell Genet*, 1997, 79(3-4), 221-224.

[46] Fu, Z; Attar-Bashi, NM; Sinclair, AJ. 1-14C-linoleic acid distribution in various tissue lipids of guinea pigs following an oral dose. *Lipids*, 2001, 36(3), 255-260.

[47] Heng, SSJ; Chan, OYW; Keng, BMH; Ling, MHT. Glucan biosynthesis protein G (mdoG) is a suitable reference gene in Escherichia coli K-12. *ISRN Microbiol*, 2011, 2011, Article ID 469053.

[48] Loubens, I; Debarbieux, L; Bohin, A; Lacroix, JM; Bohin, JP. Homology between a genetic locus (mdoA) involved in the osmoregulated biosynthesis of periplasmic glucans in Escherichia coli and a genetic locus (hrpM) controlling pathogenicity of Pseudomonas syringae. *Mol Microbiol*, 1993, 10(2), 329-340.

[49] Bohin, JP. Osmoregulated periplasmic glucans in Proteobacteria. *FEMS Microbiol Lett*, 2000, 186(1), 11-19.

[50] Page, F; Altabe, S; Hugouvieux-Cotte-Pattat, N; Lacriox, JM; Robert-Baudouy, J; Bohin, JP. Osmoregulated periplasmic glucan synthesis is required for Erwinia chrysanthemi pathogenicity. *J Bacteriol*, 2001, 183(10), 3134-3141.

[51] Meldgaard, M; Fenger, C; Lambertsen, KL; Pedersen, MD; Ladeby, R; Finsen, B. Validation of two reference genes for mRNA level studies of murine disease models in neurobiology. *J Neurosci Methods*, 2006, 156(1-2), 101-110.

[52] Keng, BMH; Chan, OYW; Hen, SSJ; Ling, MHT. Transcriptome analysis of Spermophilus lateralis and Spermophilus tridecemlineatus liver does not suggest the presence of Spermophilus-liver-specific reference genes. *ISRN Bioinformatics*, 2013, 2013, Article ID 361321.

[53] Sauer, U; Preininger, C; Hany-Schmatzberger, R. Quick and simple: quality control of microarray data. *Bioinformatics*, 2005, 21, 1572-1578.

[54] Silberberg, G; Baruch, K; Navon, R. Detection of stable reference genes for real-time PCR analysis in schizophrenia and bipolar disorder. *Anal Biochem*, 2009, 391, 91-97.

[55] Ling, MHT; Ban, Y; Wen, H; Wang, SM; Ge, SX. Conserved expression of natural antisense transcripts in mammals. *BMC Genomics*, 2013, 14(1), 243.