Acta Pharmacologica Sinica (2010) 31: 1487–1494
 1671-4083/10 $32.00
www.nature.com/aps

Original Article

Selection of reliable reference genes for gene expression study in nasopharyngeal carcinoma

Yi GUO[1, #], Jia-xin CHEN[2, #], Shu YANG[1], Xu-ping FU[1], Zheng ZHANG[2], Ke-he CHEN[2], Yan HUANG[1], Yao LI[1], Yi XIE[1], Yu-min MAO[1, *]

[1]*State Key Laboratory of Genetic Engineering, Institute of Genetics, School of Life Science, Fudan University, Shanghai 200433, China;* [2]*Institute of Nasopharyngeal Carcinoma, the People's Hospital of Guangxi Zhuang Nationality Autonomous Region, Nanning 530021, China*

Aim: To construct a system for selecting reference genes (RGs) and to select the most optimal RGs for gene expression studies in nasopharyngeal carcinoma (NPC).
Methods: The total RNAs from 20 NPC samples were each labeled with Cy5-dUTP. To create a common control, the total RNA from 15 nasopharyngeal phlogistic (NP) tissues was mixed and labeled via reverse transcription with Cy3-dUTP. cDNA microarrays containing 14 112 genes were then performed. A mathematical approach was constructed to screen stably expressed genes from the microarray data. Using this method, three genes (*YARS*, *EIF3S7*, and *PFDN1*) were selected as candidate RGs. Furthermore, 7 commonly used RGs (*HPRT1*, *GAPDH*, *TBP*, *ACTB*, *B2M*, *G6PDH*, and *HBB*) were selected as additional potential RGs. Real-time PCR was used to detect these 10 candidate genes' expression levels and the geNorm program was used to find the optimal RGs for NPC studies.
Results: On the basis of the 10 candidate genes' expression stability level, geNorm analysis identified the optimal single RG (*YARS* or *HPRT1*) and the most suitable set of RGs (*HPRT1*, *YARS*, and *EIF3S7*) for NPC gene expression studies. In addition, this analysis determined that B2M, G6PDH, and HBB were not appropriate for use as RGs. Interestingly, ACTB was the least stable RG in our study, even though previous studies had indicated that it was one of the most stable RGs. Three novel candidate genes (*YARS*, *EIF3S7*, and *PFDN1*), which were selected from microarray data, were all identified as suitable RGs for NPC research. A RG-selecting system was then constructed, which combines microarray data analysis, a literature screen, real-time PCR, and bioinformatic analysis.
Conclusion: We construct a RG-selecting system that helps find the optimal RGs. This process, applied to NPC research, determined the single RG (*YARS* or *HPRT1*) and the set of RGs (*HPRT1*, *YARS*, and *EIF3S7*) that are the most suitable internal controls.

Keywords: Gene expression; nasopharyngeal carcinoma; reference gene; cDNA microarrays; real-time PCR

Acta Pharmacologica Sinica (2010) 31: 1487–1494; doi: 10.1038/aps.2010.115

Introduction

There is no universally accepted "best" method for choosing a set of reference genes (RGs) for normalization. It has been reported[1, 2] that RGs' expression stability may vary under different types of cancer. House-keeping genes (HKGs) cannot be used as internal controls in different tissues[3], developmental stages[4], or experimental conditions[5]. Doing so could yield conflicting results. Different experimental backgrounds should be strictly considered when determining the best RG[3, 6–8].

In gene expression studies, RGs are usually selected on the basis of previously published literature. However, these RGs are often used in very different experimental conditions and lack validation. HKGs, such as *GAPDH*, *ACTB*, and *HPRT1*, are the most widely utilized genes in the calibration of gene expression levels[6]. A growing number of studies have been reporting that the expression level of RGs may vary in different organisms or experimental conditions. Even the most commonly used HKG cannot be assumed to a universally applicable RG; its behavior in different cancers, tissues, or cells must be considered[9–11]. A set of RGs should be strictly validated before their use. Here, we report an RG selection system that could help to address such problems. nasopharyngeal carcinoma (NPC) was used as an example to test the efficiency of the system.

NPC is a prevalent tumor in Southeast Asia, particularly among the Cantonese population of southern China. Its incidence has remained high for decades[12]. Compared to other malignant tumors of the upper aerodigestive tract, NPC is

# These authors contributed equally to this work.
* To whom correspondence should be addressed.
E-mail yaoli@fudan.edu.cn
Received 2010-03-04 Accepted 2010-07-07

a special type of head and neck cancer in its epidemiology, pathology, clinical presentation, and response to treatment. The etiology of NPC involves multiple factors, including genetic susceptibility, exposure to chemical carcinogens, and Epstein-Barr virus (EBV) infection[13].

We established a strategy to determine the most suitable RGs for NPC studies. Similar studies in other forms of cancer, including prostate cancer[14], human bladder cancer[15], hepatocellular carcinoma[16], breast carcinoma[17], gastrointestinal tumor[18], and human renal cell carcinoma[19], have previously been carried out.

The goals of our study were (a) to develop an improved strategy for selecting optimal RGs in gene expression studies and (b) to investigate a panel of potential RGs with regard to their expression stability and therefore their suitability for use as RGs in NPC research. Our findings are intended to improve methods of normalization and the accuracy of gene expression data.

Materials and methods

Tissue information, RNA isolation, and reverse transcription

Poorly differentiated squamous NPC tissue samples were obtained from 37 NPC patients with consent, before treatment, at the Institute of Nasopharyngeal Carcinoma, The People's Hospital of Guangxi Zhuang Nationality Autonomous Region, Nanning, China. In addition, 20 nasopharyngeal phlogistic (NP) tissues were obtained from patients without NPC at the same hospital. All specimens were reviewed by an otorhinolaryngologic pathologist. All fresh tissues were snap-frozen in liquid nitrogen and stored at -80 °C until use.

Each specimen was ground into a fine powder in a 10 cm ceramic mortar (RNase-free), and then RNA was isolated following the manufacturer's recommendation (Biostar, Shanghai). RNA was resuspended in Milli-Q H_2O. RNA concentrations were determined by spectrophotometry using a GeneQuant II (Pharmacia Biotech). The integrity of the RNA was verified by 1% agarose gel electrophoresis. Following the manufacturer's instructions, total RNA was reverse transcribed with Oligo (dT) primers (Takara) and a Superscript III RNase H-Reverse Transcriptase kit (Invitrogen). The resultant synthesized cDNA was stored at -20 °C.

Gene expression profiles

Fluorescent cDNA probes were prepared through reverse transcription of the isolated total RNA and then purified. The total RNAs from 15 nasopharyngeal phlogistic tissues were mixed and labeled with Cy3-dUTP by reverse transcription. This mixture is the common control pool in our experiments. RNA samples from 20 NPC patients were separately labeled with Cy5-dUTP. 20 cDNA microarrays containing 14112 genes were then performed, following the manufacturer's instructions. The microarrays were scanned to detect emission from Cy5 and Cy3 by ScanArray 4000 (Packard, Biochip Technologies) at two wavelengths, 635 nm and 532 nm, respectively. The acquired images were analyzed using GenePix Pro 3.0 software. The intensities at the two wavelengths for each spot indicate the quantity of bound Cy3-dUTP and Cy5-dUTP. Ratios of Cy5 to Cy3 were computed by GenePix Pro 3.0 using the median ratio method. Overall intensities were normalized using the Lowess method. The specifics of the 20 cDNA microarray (BioStarH-141s, Biostar, Shanghai) experiments, including probe preparation, microarray hybridization, image detection and data normalization, were carried out as in our previous reports[20, 21].

Potential reference gene selection in microarray data

An equation based on the coefficient of variation was constructed to evaluate the data resulting from the 20 gene expression profiles. In Equation 1, for two genes x and y, a_{ix} and a_{iy} stand for their ratio from microarray i. Matrix A_{xy} consists of m elements that are $\log_2$-transformed ratios a_{ix}/a_{iy}, which are calculated from the m=20 microarrays (Equation 1). We defined the pairwise variation V_{xy} for the two genes x and y as the standard deviation of the matrix A_{xy} (Equation 2). The gene stability measure S_x for gene x was the arithmetic mean of all pairwise variations V_{xy} (Equation 3).

$$(\forall x, y \in [1, n],\ x \neq y \text{ and } m = 20):$$

$$A_{xy} = \left\{ \log_2\left(\frac{a_{1x}}{a_{1y}}\right), \log_2\left(\frac{a_{2x}}{a_{2y}}\right), \cdots, \log_2\left(\frac{a_{mx}}{a_{my}}\right) \right\} = \left\{ \log_2\left(\frac{a_{ix}}{a_{iy}}\right) \right\}_{i=1\to m} \quad (1)$$

$$V_{xy} = STDEV(A_{xy}) = \sqrt{\frac{m\sum A_{xy}{}^2 - \left(\sum A_{xy}\right)^2}{m(m-1)}} \quad (2)$$

$$S_x = \frac{\sum_{y=1}^{n} V_{xy}}{n-1} \quad (3)$$

Primer design and real-time quantitative PCR

PCR primer sequences were based on the NCBI database (National Center for Biotechnology Information, [http://www.ncbi.nlm.nih.gov/]) and Ensembl database ([http://www.ensembl.org/]) and were designed using Primer Premier 5. BLAST analysis against genomic DNA was performed, using both databases, to test the specificity of the primers. Information about the primers is listed in Table 1. These primers were then used in quantitative real-time PCR, with the ultimate purpose of determining the appropriate RGs. In the real-time PCR experiments, ten housekeeping genes (*YARS*, *EIF3S7*, *PFDN1*, *HPRT1*, *GAPDH*, *TBP*, *ACTB*, *B2M*, *G6PDH*, and *HBB*) were replicated in triplicate using Sybr Green Mastermix on the ABI Prism 7900 Sequence Detection System (Applied Biosystems, Foster City, CA). Each reaction contained cDNA corresponding to 20 ng of RNA along with 400 nmol/L primers in a final volume of 10 μL. A reaction mixture without cDNA template was used as a negative control. For PCR, reaction mixtures were initially incubated at 95 °C for 15 min and then subjected to 40 cycles of 95 °C for 15 s and 60 °C for 1 min. Melting curve analysis was performed on each sample to ensure that a single amplicon was amplified in the reaction. For each different pair of primers, efficiency of real-

Table 1. Information on the primers used for real-time PCR.

Gene	Accession No	Gene name	Forward primer sequence [5′→3′] Reverse primer sequence [5′→3′]	Position in cDNA	Amplicon size (bp)	E	R^2
GAPDH	NM_002046	Glyceraldehyde-3-phosphate dehydrogenase	GAA GGT GAA GGT CGG AGT C GAA GAT GGT GAT GGG ATT TC	2nd Exon 4th Exon	226	98.4	0.984
G6PDH	NM_000402	Glucose-6-phosphate dehydrogenase	ATC GAC CAC TAC CTG GGC AA TTC TGC ATC ACG TCC CGG A	6th Exon 7th/8th Exon*	191	100.3	0.971
HPRT1	NM_000194	Hypoxanthine phosphoribosyltransferase 1	GAC CAG TCA ACA GGG GAC AT AAC ACT TCG TGG GGT CCT TTT C	4th Exon 7th Exon	195	100.2	0.998
TBP	NM_003194	TATA box binding protein	TTC GGA GAG TTC TGG GAT TGT A TGG ACT GTT CTT CAC TCT TGG C	3rd Exon 5th/6th Exon*	227	99.6	0.95
ACTB	NM_001101	Beta-actin	CAT CGA GCA CGG CAT CGT CA TAG CAC AGC CTG GAT AGC AAC	3rd Exon 4th Exon	211	101.2	0.99
B2M	NM_004048	Beta-2-microglobulin	GGC TAT CCA GCG TAC TCC A ACG GCA GGC ATA CTC ATC T	1st/2nd Exon* 2nd Exon	247	104.0	0.998
HBB	NM_000518	Hemoglobin, beta	TGA GCT GCA CTG TGA CAA G TTA GTG ATA CTT GTG GGC CAG	2nd Exon 3rd Exon	175	97.2	0.999
YARS	NM_003680	Tyrosyl-tRNA synthetase	TGA GCA CGT GTT TGT GAA G AAG ATG CTG GGC TGG CTA	12th Exon 13th Exon	211	98.2	0.997
PFDN1	NM_002622	Prefoldin subunit 1	CTC GCA GAC ATA CAG ATT GA CGC TCC AGG TAG GAC TTT T	2nd Exon 4th Exon	224	95.4	0.991
EIF3S7	NM_003753	Eukaryotic translation initiation factor 3, subunit 7	TAT TGA CCT TAT TGT CCG TTG T TCA GAT CCA GCC AGC AAA G	12th Exon 13th Exon	231	98.2	0.972

E is a measure of real-time PCR reaction efficiency and R^2 indicates reproducibility of the reaction.

time PCR (E), slope values, and correlation coefficients (R^2) were determined using serial 1:5 dilutions of template cDNA. E was calculated with the equation ($E=(10^{[-1/slope]}-1)\times100$).

GeNorm analysis

The software program geNorm, version 3.4, was used to compare the stability of candidate reference genes. It is a visual basic application for Microsoft Excel ([http://medgen.ugent.be/~jvdesomp/genorm/])[22]. The expression stability (M) was calculated as the standard deviation of the logarithmically transformed expression ratios[23]. This M value is the average pairwise variation of one particular gene compared to all the other control genes. Genes with the lowest M value had the most stable transcription[28]. The least stable gene, which is the one with the highest M value, was automatically excluded from the next calculation round. The cut-off value of M was set at 1.5[14, 24–27].

GeNorm also calculated the optimal set of RGs through producing normalization factors (NF_n), which are based on the geometric mean of the n most stable reference genes' expression levels[23, 29]. For example, a normalization factor (NF_2) was calculated as the geometric mean of the two most stably expressed genes' expression levels. The procedure then continued to produce NF_3 to NF_n by stepwise inclusion of additional RGs. The optimal number of reference genes was chosen by calculating the pairwise variation parameter V[23, 30], which is defined as the pairwise variation between 2 sequential normalization factors for each normalization factor and its predecessor. If the additional gene has a significant effect, it will have a large variation and should be included in the RGs set[14].

Results

RNA quality

All RNA was separately extracted from 37 NPC and 20 NP tissues. From these, 20 NPC and 15 NP tissues were used to perform cDNA microarray studies. The other 17 NPC and 5 NP tissues were prepared for real-time PCR experiments, for which the RNA concentrations are listed (Table 2). RNA concentrations and purity were assessed using agarose gel electrophoresis and by determining the absorbance ratio of 260 nm to 280 nm (mean±SD, 1.91±0.03).

Potential reference genes selection

An equation (see Materials and Methods) was constructed to evaluate the data resulting from the 20 gene expression profiles. The most stably expressed genes are listed in Table 3. In addition, all HKGs [31] in the microarrays were analyzed. The distribution of the HKGs' *S* scores is shown in Figure 1A. The three HKGs (YARS, EIF3S7, and PFDN1) with the lowest S values were selected as the candidate RGs.

MEDLINE was searched for Medical Subject Heading (MeSH) terms "nasopharyngeal neoplasm" and "RT-PCR." We evaluated the entire set of 151 articles resulting from this search and found 53 articles could be used (Figure 1B). Four genes were selected as candidate RGs: *GAPDH* (27 times; 50.9%), *ACTB* (18 times; 34.0%), *HBB* (3 times; 5.7%),and *HPRT1* (2 times; 3.8%).

Table 2. RNA concentrations of NPC and NP samples for real-time PCR.

Samples	RNA (μg/μL)	Samples	RNA (μg/μL)	Samples	RNA (μg/μL)	Samples	RNA (μg/μL)
NP1	9.48	NPC6	11.68	NPC12	6.32	NPC17	8.40
NP2	5.94	NPC7	6.68	NPC13	4.86	NPC18	5.28
NP3	6.20	NPC8	6.39	NPC14	4.96	NPC19	5.04
NP4	4.94	NPC9	6.09	NPC15	6.05	NPC20	3.16
NP5	9.87	NPC10	6.27	NPC16	7.75	NPC21	6.61
		NPC11	6.78			NPC22	4.46

Table 3. The most stably expressed genes from microarray data. √ means the gene is HKG, while × means not.

Gene name	Symbol	HKG	Accession No
Tyrosyl-tRNA synthetase	YARS	√	NM_003680
Prefoldin subunit 1	PFDN1	√	NM_002622
Eukaryotic translation initiationfactor 3, subunit D	EIF3S7	√	NM_014989
E2F transcription factor 4	E2F4	√	NM_001950
Tetratricopeptide repeat domain 1	TTC1	√	NM_003314
HCLS1 associated protein X-1	HAX1	√	NM_001018837
Glutathione peroxidase 7	GPX7	×	NM_003753
Regulating synaptic membrane exocytosis 1	RIMS1	×	NM_015696
COMM domain containing 2	COMMD2	×	NM_016094
Component of oligomeric golgi complex 1	COG1	×	NM_018714
Peptidylprolyl isomerase (cyclophilin)-like 3	PPIL3	×	NM_032472
FAD-dependent oxidoreductase domain containing 1	H17	×	NM_017547
General transcription factor IIH, polypeptide 4	GTF2H4	×	NM_001517
Methylmalonic aciduria (cobalamin deficiency) cblA type	MMAA	×	NM_172250
SMAD family member 2	SMAD2	×	NM_001003652
Plexin domain containing 2	PLXDC2	×	NM_032812
RUN and SH3 domain containing 1	RUSC1	×	NM_001105203
Inhibitor of growth family, member 1	ING1	×	NM_005537
polymerase (DNA directed) kappa	POLK	×	NM_016218
Regulator of cohesion maintenance, homolog B	APRIN	×	NM_015032

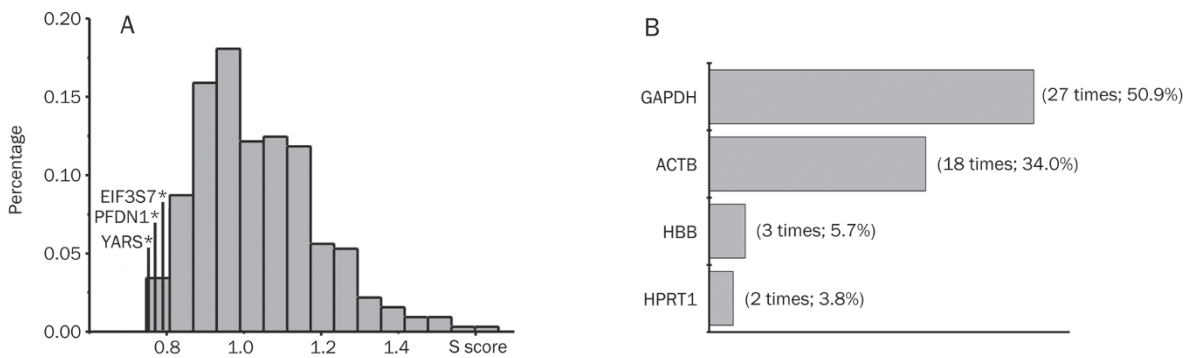

Figure 1. Selecting potential reference genes. According to gene expression levels, all housekeeping genes in cDNA microarrays were ranked by *S* score. The distribution of the housekeeping genes' S scores is shown here. This distribution includes the S scores of the three genes (*YARS*, *EIF3S7*, and *PFDN1*) with the lowest S scores. In previously published NPC studies, the 4 genes used most frequently as internal control genes were *GAPDH*, *ACTB*, *HBB*, and *HPRT1*.

Expression levels of candidate RGs

Real-time PCR was performed to detect the expression levels of the 10 candidate RGs in 17 NPC tissues and 5 NP tissues. Ct values of candidate RGs were between 16 and 30, which is a

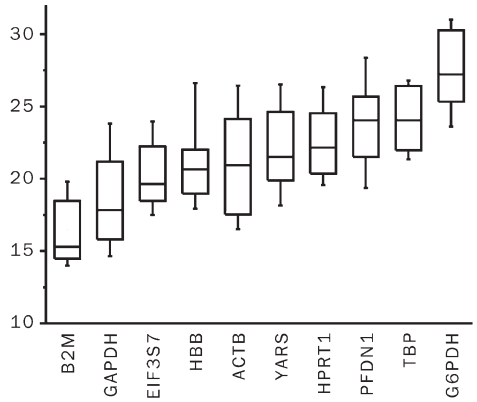

Figure 2. Expression levels of candidate reference genes in NPC and NP samples. Values are given as real-time PCR cycle threshold numbers (Ct values). Boxes represent the lower and upper quartiles with medians; bars represent the ranges for the data.

wide expression range for real-time PCR experiments (Figure 2).

GeNorm analysis

Gene expression stability was analyzed by geNorm[22, 23], which calculated the measure of gene expression stability (*M*) based on the average pairwise variation of all studied genes.

The gene with the lowest M value was considered the most stable[23, 32]. The genes studied, in order from the most stable to the least stable, were *YARS* and *HPRT1*, *EIF3S7*, *GAPDH*, *TBP*, *PFDN1*, *ACTB*, *B2M*, *G6PDH*, and *HBB* (Figure 3A).

Furthermore, the optimal number of reference gene sets was evaluated by comparing the pairwise variation between sequential normalization factors (NFs) (Figure 3B). The results show that the inclusion of the fourth gene had about the same effect (V=0.298) on the NF as the inclusion of the third gene (V=0.287) had. There is no significant improvement in the normalization factor to be gained from using more than three genes (*HPRT1*, *YARS*, and *EIF3S7*).

Discussion

In our strategy (Figure 4), there are two main steps to creating the candidate RG pool. The first step involves screening genes using microarray data. We established a mathematical basis for screening genes based on microarray data and calculated every gene's *S* score using the relevant equations. The *S* score for a gene represents the stability of that gene's expression level. The lower the *S* value is, the more stably the gene is expressed. In those genes with relatively low *S*, we preferred HKGs to be candidate RGs because they are a group of genes that are required for the maintenance of basal cellular function and are constitutively expressed in all cells. The second main step in our strategy involves screening published literature to identify frequently used control genes. Such a two-step approach to select candidate RGs could provide more genes' information and identify optimal RGs with greater accuracy to help finding optimal RGs. We validated those candidate genes

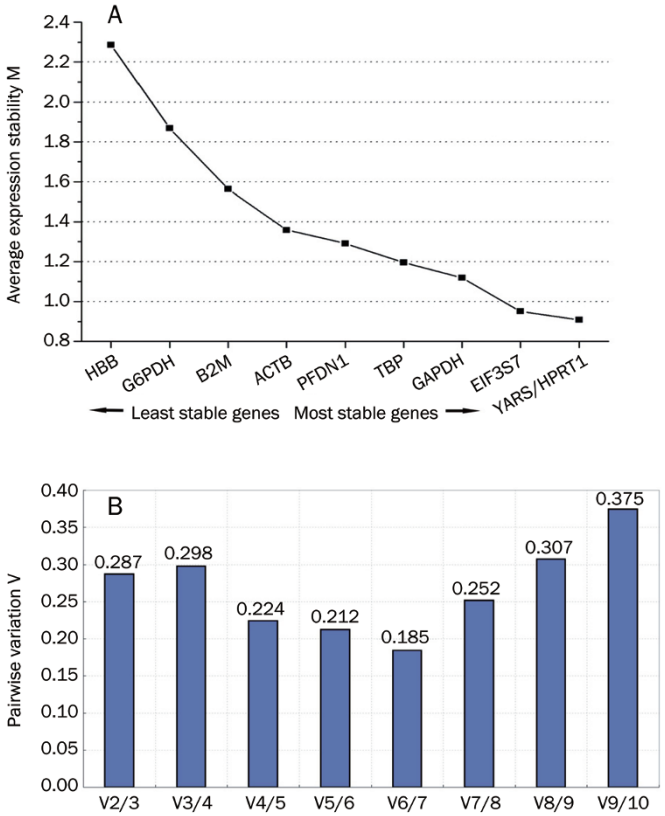

Figure 3. GeNorm analysis of 10 candidate genes. (A) The stability parameter M, which was calculated for each gene in every calculation round, is plotted on the Y axis. The X axis shows the 10 genes ranked according to their expression stability. (B) geNorm calculated NF from leastwise 2 genes to determine the optimal number of reference genes. V is defined as the pairwise variation between 2 sequential NF_n and NF_{n+1}.

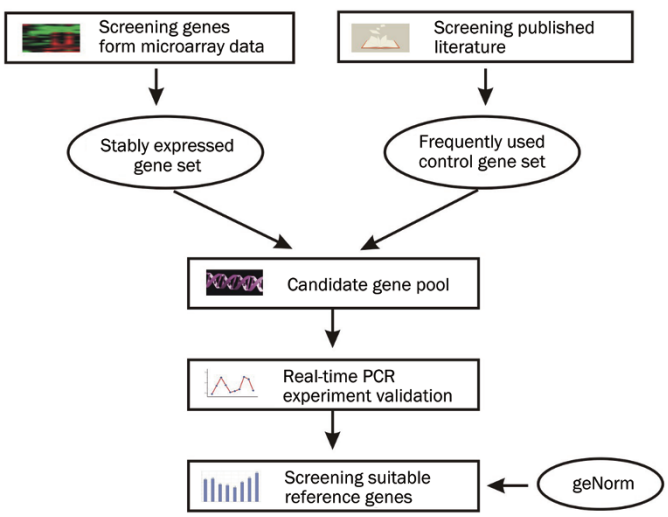

Figure 4. The strategy of selecting reference genes.

using the results of real-time PCR. Then geNorm, a widely available program, was used to screen the best RGs from the

real-time PCR data. The GeNorm program was used to validate RGs by determining the most stably expressed candidate gene based on each gene's average expression stability (*M*). A normalization factor (NF) was then generated by calculating the geometric mean of the most stable RGs[23].

Potential reference genes selection

The more stable the expression of a gene is, the lower that gene's *S* score. When n genes were ranked by *S* score, the most unstable gene with the highest *S* score was then excluded for the next round because this gene may cause significant bias in the results. Then, a new *S* score of each of the remaining n-1 genes was calculated in the second round. This procedure was repeated until just two genes remained. Ultimately, every gene in the gene expression profile was ranked according to the stability of its expression. Three HKGs (*YARS*, *EIF3S7*, and *PFDN1*) with the lowest *S* value were selected as the candidate RGs. The literature does not report previous use of these genes as RGs.

We performed a MEDLINE search to identify RGs used in previously published NPC studies . The results of this search showed that no standard set of RGs for NPC expression studies currently exists. Four genes (*GAPDH*, *ACTB*, *HBB*, and *HPRT1*) were then selected as candidate RGs. In addition, we selected another three HKGs (*TBP*, *B2M*, and *G6PDH*) that are frequently used as RGs in other cancer studies[14, 15, 33–37]. Then, the candidate RGs were narrowed down to ten HKGs. Such an approach to select candidate RGs could provide more genes' information to help identify optimal RGs.

Expression levels of candidate RGs

Real-time PCR was then performed to find the expression level of the 10 candidate RGs. In gene expression studies, it is better to use RGs with similar expression levels to normalize the target genes. In our study, Ct values of candidate RGs were between 16 and 30, which is a wide expression range for real-time PCR experiments (Figure 2). *GAPDH*, *EIF3S7*, and *ACTB* were highly expressed, with Ct values below 21 cycles. *YARS*, *PFDN1*, *HPRT1*, and *TBP* were expressed at relatively low levels and had Ct values over 21 cycles. This set of genes could be an effective reference for target genes with a large range of gene expression levels.

GeNorm analysis

GeNorm calculated the measure of gene expression stability (*M*). Following the procedure specified by geNorm, we converted Ct values to linear quantities by $2^{-\Delta Ct}$ using the highest expressed sample as a calibration sample. *M* is defined as the average pairwise variation of one gene compared to each of the others. Genes were ranked from the most stable to the least stable as follows: *YARS* and *HPRT1*, *EIF3S7*, *GAPDH*, *TBP*, *PFDN1*, *ACTB*, *B2M*, *G6PDH*, and *HBB* (Figure 3A). NPC expression research required *M* values of 1.5 or less[14, 24–27]. B2M, G6PDH, and HBB were not suitable for NPC expression research because their *M* values were above the 1.5 threshold. *YARS* and *HPRT*, both of which had the lowest *M* value of the group, were identified as the two most stably expressed genes.

In our studies, *HPRT1* and *YARS* were identified by geNorm[23] as the most appropriate internal control genes of 10 candidate RGs. Furthermore, 5 other genes (*EIF3S7*, *GAPDH*, *TBP*, *PFDN1*, and *ACTB*) were also shown to be appropriate RGs for NPC studies (Figure 3A). The three least stable genes (*B2M*, *G6PDH*, and *HBB*) were all beyond the cut-off value. Although these three genes have commonly been used as RGs in other cancers, they should be avoided as internal control genes in NPC. Interestingly, the results showed that *ACTB* was not as stable as predicted. It had the lowest *M* value of seven qualified RGs. To the best of our knowledge, *ACTB* was one of the most commonly used RGs in previously published NPC studies. According to the articles evaluated, *ACTB* was used 18 times (34.0%), which was more frequently used than other genes, second to *GAPDH*, which was used 27 times (50.9%). The relatively high variability of *ACTB* expression further demonstrates the problems of relying on commonly used RGs. Similar erroneous normalizations had been reported by other researchers[14, 23]. These examples indicate that the absence of a thorough validation process for RGs could result in imprecise normalization.

Furthermore, NFs were calculated for the two most stably expressed genes by stepwise inclusion of a less stable RG. The results showed there is no significant improvement in the normalization factor when using more than three genes (*HPRT1*, *YARS*, and *EIF3S7*). This result indicates that the three-gene set is adequate for the normalization in NPC gene expression studies. We strongly recommended that if conditions permit, this three-gene set should be used instead of the single RG for data normalization.

Advantages and limits to the RG-selecting system

The conventional approach to finding RGs involves only screening earlier studies. There are 2 obvious defects in this approach. The first defect is as noted above, a lack of any systematic comparison of particular experimental contexts. Simply using conventionally used RGs may result in a bias of normalization. The second defect is that if no RGs appear in the prior published research for the cancer or species to be studied, then one is left with no method for choosing the right RGs. Our strategy helps to remedy both of these defects by including microarray data analysis as part of the method for choosing RGs. Furthermore, the present study is the first systematic comparison of the effectiveness of a set of potential RGs for nasopharyngeal carcinoma research. No previous studies have tried to validate optimal RGs in NPC. One possible reason this validation has not been done previously is that NPC and nasopharyngeal phlogistic specimens are very difficult to acquire because (a) most NPC are prevalent in Southeast China; and (b) the size of the cancer tissue is very small.

When we compared the cDNA microarray screen results to the geNorm results, we found that *HPRT1*, one of the best control genes, was not identified in the screen of the microarray data. The failure to identify it as microarray candidate RGs

may have biological and methodological reasons. One reason might be that many genes in the microarray such as ribosomal protein genes were stably expressed. Although *HPRT1* was not selected as candidate gene, it was still in the pool of stably expressed genes with low *T* values. Another reason it was not identified might be the fact that, to calculate *T* values, more than 10 000 genes were used to correct for each other, but only a few candidate genes were used in geNorm. Furthermore, real-time PCR data were thought to be more accurate than cDNA microarray data. Methodological differences between cDNA microarray and real-time PCR were also responsible for the bias. We can conclude from this result that the literature screen is an important complement to the microarray screening strategy.

We also reported here three new HKGs (*YARS*, *PFDN1*, and *EIF3S*), which were all shown to be appropriate RGs for NPC gene expression studies. These genes were identified directly from the 20 microarray data sets. The current study is the first report of these genes as potential RGs in NPC. The current study also shows that the strategy of screening microarray data for suitable candidate RGs is feasible. The process for finding candidate RGs shown in this study may help researchers who wish to find RGs for a new species, tissue, or cell type that has no record of proven RGs.

Acknowledgements

This research was supported by grant 30860081 from the National Natural Science Foundation of China. This research was also supported by the Guangxi Science and Technology Key Project (No 0719006-2-4).

Author contribution

Yi GUO and Jia-xin CHEN designed the strategy of the research presented in this article. NPC tissues were collected by Jia-xin CHEN, and biology experiments were carried out by Yi GUO. Yi GUO, Jia-xin CHEN, Shu YANG, and Xu-ping FU performed microarray data and RT-PCR analysis. Yi GUO drafted the manuscript. Zheng ZHANG, Ke-he CHEN, Yan HUANG, Yao LI, Yi XIE, and Yu-min MAO helped to revise the manuscript. Yao LI organized all the results and provided advice in the preparation of the manuscript.

References

1 Bustin SA. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. J Mol Endocrinol 2002; 29: 23–39.
2 Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol 2000; 25: 169–93.
3 Radonic A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A. Guideline to reference gene selection for quantitative real-time PCR. Biochem Biophys Res Commun 2004; 313: 856–62.
4 Bilodeau-Goeseels S, Schultz GA. Changes in the relative abundance of various housekeeping gene transcripts in *in vitro*-produced early bovine embryos. Mol Reprod Dev 1997; 47: 413–20.
5 Zhong H, Simons JW. Direct comparison of GAPDH, beta-actin, cyclophilin, and 28S rRNA as internal standards for quantifying RNA levels under hypoxia. Biochem Biophys Res Commun 1999; 259: 523–6.
6 Huggett J, Dheda K, Bustin S, Zumla A. Real-time RT-PCR normalisation; strategies and considerations. Genes Immun 2005; 6: 279–84.
7 Radonic A, Thulke S, Bae HG, Muller MA, Siegert W, Nitsche A. Reference gene selection for quantitative real-time PCR analysis in virus infected cells: SARS corona virus, Yellow fever virus, Human Herpesvirus-6, Camelpox virus and Cytomegalovirus infections. Virol J 2005; 2: 7.
8 Bogaert L, Van Poucke M, De Baere C, Peelman L, Gasthuys F, Martens A. Selection of a set of reliable reference genes for quantitative real-time PCR in normal equine skin and in equine sarcoids. BMC Biotechnol 2006; 6: 24.
9 Goidin D, Mamessier A, Staquet MJ, Schmitt D, Berthier-Vergnes O. Ribosomal 18S RNA prevails over glyceraldehyde-3-phosphate dehydrogenase and beta-actin genes as internal standard for quantitative comparison of mRNA levels in invasive and noninvasive human melanoma cell subpopulations. Anal Biochem 2001; 295: 17–21.
10 Schmittgen TD, Zakrajsek BA. Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. J Biochem Biophys Methods 2000; 46: 69–81.
11 de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, *et al*. Evidence based selection of housekeeping genes. PLoS ONE 2007; 2: e898.
12 Agulnik M, Siu LL. State-of-the-art management of nasopharyngeal carcinoma: current and future directions. Br J Cancer 2005; 92: 799–806.
13 Tao Q, Chan AT. Nasopharyngeal carcinoma: molecular pathogenesis and therapeutic developments. Expert Rev Mol Med 2007; 9: 1–24.
14 Ohl F, Jung M, Xu C, Stephan C, Rabien A, Burkhardt M, *et al*. Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization? J Mol Med 2005; 83: 1014–24.
15 Ohl F, Jung M, Radonic A, Sachs M, Loening SA, Jung K. Identification and validation of suitable endogenous reference genes for gene expression studies of human bladder cancer. J Urol 2006; 175: 1915–20.
16 Gao Q, Wang XY, Fan J, Qiu SJ, Zhou J, Shi YH, *et al*. Selection of reference genes for real-time PCR in human hepatocellular carcinoma tissues. J Cancer Res Clin Oncol 2008; 134: 979–86.
17 Lyng MB, Laenkholm AV, Pallisgaard N, Ditzel HJ. Identification of genes for normalization of real-time RT-PCR data in breast carcinomas. BMC Cancer 2008; 8: 20.
18 Kidd M, Nadler B, Mane S, Eick G, Malfertheiner M, Champaneria M, *et al*. GeneChip, geNorm, and gastrointestinal tumors: novel reference genes for real-time PCR. Physiol Genomics 2007; 30: 363–70.
19 Jung M, Ramankulov A, Roigas J, Johannsen M, Ringsdorf M, Kristiansen G, *et al*. In search of suitable reference genes for gene expression studies of human renal cell carcinoma by real-time PCR. BMC Mol Biol 2007; 8: 47.
20 Li Y, Li Y, Tang R, Xu H, Qiu M, Chen Q, *et al*. Discovery and analysis of hepatocellular carcinoma genes using cDNA microarrays. J Cancer Res Clin Oncol 2002; 128: 369–79.
21 Wei Q, Li Y, Chen L, Zhang L, He X, Fu X, *et al*. Genes differentially expressed in responsive and refractory acute leukemia. Front Biosci 2006; 11: 977–82.
22 Van Zeveren AM, Visser A, Hoorens PR, Vercruysse J, Claerebout E, Geldhof P. Evaluation of reference genes for quantitative real-time PCR in Ostertagia ostertagi by the coefficient of variation and geNorm

approach. Mol Biochem Parasitol 2007; 153: 224–7.
23 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: RESEARCH0034.
24 Botteldoorn N, Van Coillie E, Grijspeerdt K, Werbrouck H, Haesebrouck F, Donne E, *et al*. Real-time reverse transcription PCR for the quantification of the mntH expression of Salmonella enterica as a function of growth phase and phagosome-like conditions. J Microbiol Methods 2006; 66: 125–35.
25 Spinsanti G, Panti C, Lazzeri E, Marsili L, Casini S, Frati F, *et al*. Selection of reference genes for quantitative RT-PCR studies in striped dolphin (Stenella coeruleoalba) skin biopsies. BMC Mol Biol 2006; 7: 32.
26 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.
27 Mamo S, Gal AB, Bodo S, Dinnyes A. Quantitative evaluation and selection of reference genes in mouse oocytes and embryos cultured *in vivo* and *in vitro*. BMC Dev Biol 2007; 7: 14.
28 Nailis H, Coenye T, Van Nieuwerburgh F, Deforce D, Nelis HJ. Development and evaluation of different normalization strategies for gene expression studies in Candida albicans biofilms by real-time PCR. BMC Mol Biol 2006; 7: 25.
29 von Smolinski D, Leverkoehne I, von Samson-Himmelstjerna G, Gruber AD. Impact of formalin-fixation and paraffin-embedding on the ratio between mRNA copy numbers of differently expressed genes. Histochem Cell Biol 2005; 124: 177–88.
30 Heckmann LH, Connon R, Hutchinson TH, Maund SJ, Sibly RM, Callaghan A. Expression of target and reference genes in Daphnia magna exposed to ibuprofen. BMC Genomics 2006; 7: 175.
31 Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends Genet 2003; 19: 362–5.
32 Robinson TL, Sutherland IA, Sutherland J. Validation of candidate bovine reference genes for use with real-time PCR. Vet Immunol Immunopathol 2007; 115: 160–5.
33 Zhang X, Ding L, Sandford AJ. Selection of reference genes for gene expression studies in human neutrophils by real-time PCR. BMC Mol Biol 2005; 6: 4.
34 Goossens K, Van Poucke M, Van Soom A, Vandesompele J, Van Zeveren A, Peelman LJ. Selection of reference genes for quantitative real-time PCR in bovine preimplantation embryos. BMC Dev Biol 2005; 5: 27.
35 Lee PD, Sladek R, Greenwood CM, Hudson TJ. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res 2002; 12: 292–7.
36 Wong TS, Kwong DL, Sham JS, Wei WI, Kwong YL, Yuen AP. Quantitative plasma hypermethylated DNA markers of undifferentiated nasopharyngeal carcinoma. Clin Cancer Res 2004; 10: 2401–6.
37 To EW, Chan KC, Leung SF, Chan LY, To KF, Chan AT, *et al*. Rapid clearance of plasma Epstein-Barr virus DNA after surgical treatment of nasopharyngeal carcinoma. Clin Cancer Res 2003; 9: 3254–9.