Identification of Novel Reference Genes Using Multiplatform Expression Data and Their Validation for Quantitative Gene Expression Analysis

Mi Jeong Kwon¹, Ensel Oh², Seungmook Lee³, Mi Ra Roh⁴, Si Eun Kim¹, Yangsoon Lee⁵, Yoon-La Choi⁶, Yong-Ho In⁴, Taesung Park³, Sang Seok Koh⁷, Young Kee Shin¹,²*

¹Laboratory of Molecular Pathology, Department of Pharmacy, College of Pharmacy, Seoul National University, Seoul, Korea, ²Interdisciplinary Program of Bioinformatics, College of Natural Science, Seoul National University, Seoul, Korea, ³Department of Statistics, College of Natural Science, Seoul National University, Seoul, Korea, ⁴BIT center, CT&D Ltd., Seoul, Korea, ⁵LG Life Sciences, Ltd., R&D Research Park, Daejeon, Korea, ⁶Department of Pathology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea, ⁷Protein Therapeutics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea

Abstract
Normalization of mRNA levels using endogenous reference genes (ERGs) is critical for an accurate comparison of gene expression between different samples. Despite the popularity of traditional ERGs (tERGs) such as GAPDH and ACTB, their expression variability in different tissues or disease status has been reported. Here, we first selected candidate housekeeping genes (HKGs) using human gene expression data from different platforms including EST, SAGE, and microarray, and 13 novel ERGs (nERGs) (ARL8B, CTBP1, CUL1, DMT1L, FBXW2, GGBP1, LUC7L2, OAZ1, PAPOLA, SPS21, TRIM27, UBQLN1, ZNF207) were further identified from these HKGs. The mean coefficient variation (CV) values of nERGs were significantly lower than those of tERGs and the expression level of most nERGs was relatively lower than high expressing tERGs in all dataset. The higher expression stability and lower expression levels of most nERGs were validated in 108 human samples including formalin-fixed paraffin-embedded (FFPE) tissues, frozen tissues and cell lines, through quantitative real-time RT-PCR (qRT-PCR). Furthermore, the optimal number of nERGs required for accurate normalization was as few as two, while four genes were required when using tERGs in FFPE tissues. Most nERGs identified in this study should be better reference genes than tERGs, based on their higher expression stability and fewer numbers needed for normalization when multiple ERGs are required.

Citation: Kwon MJ, Oh E, Lee S, Roh MR, Kim SE, et al. (2009) Identification of Novel Reference Genes Using Multiplatform Expression Data and Their Validation for Quantitative Gene Expression Analysis. PLoS ONE 4(7): e6162. doi:10.1371/journal.pone.0006162

Editor: Jörg Hoheisel, Deutsches Krebsforschungszentrum, Germany

Received March 3, 2009; Accepted June 10, 2009; Published July 7, 2009

Funding: This study was supported by the Research Institute of Pharmaceutical Science, Korea Science and the Engineering Foundation (KOSEF) grant funded by the government of Korea (MOST) (M10529050001-06N2905-00110 and M10729000029-07N2900-02910) and a grant from the Korea Health21 R&D Project of the Ministry of Health & Welfare (A050260). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ykeeshin@snu.ac.kr

Introduction
Gene expression analysis is becoming more important in diagnostic fields as it allows for the identification of novel biomarkers relevant to diseases. Endogenous reference genes (ERGs) are widely used to normalize the mRNA level in the relative quantification to provide an accurate comparison of gene expression between different samples [1]. Traditional ERGs (tERGs), such as GAPDH and ACTB, have been used in expression studies without proper validation because of the assumption that they are expressed at constant levels across different samples and regardless of experimental treatments [2,3]. However, several reports have shown that the expression of tERGs can vary in different tissues and be regulated by experimental treatments or pathological state [2,4–10]. As the use of inappropriate ERGs in relative quantification of gene expression can result in biased expression profiles [4,11,12], the selection of proper ERGs is essential for accurate measurement especially in quantitative methods including qRT-PCR, which is a highly sensitive and accurate method [13].

Although there have been a number of previous studies aimed at finding suitable ERGs, most of them have focused on selecting the most stable genes from commonly used ERGs [14–17]. Moreover, the identification of novel ERGs (nERGs) has been based primarily on microarray data [10,18–21]. Although short-oligo microarrays such as Affymetrix GeneChips, have the advantage of being highly sensitive in detecting low abundance transcripts (nearly 3–20 transcripts per million (tpm)) [22], they have some disadvantages such as inaccurate cross hybridization between probes and transcripts, differences in hybridization efficiencies between probe sets, limited linear range of signal, and incorrect annotation of transcripts [23,24]. Therefore, an approach using only microarray data may not be sufficient to identify the most suitable ERGs. Although an ideal universal ERG may not exist [1,13], a combination of large expression data from different platforms is expected to complement the limitation of each platform [25] and allow for the identification of more suitable ERGs.

Here, we describe an algorithm for the identification of nERGs using the publicly available human gene expression datasets in addition to in-house microarray data. The expression of selected
nERGs in datasets was validated by qRT-PCR in 108 human samples and their expression stability was compared to that of tERGs.

Materials and Methods

Ethics Statement

This study was approved by the institutional review board of Samsung Medical Center. Patients’ written informed consent was not required to be obtained as the institutional review board approved the use of human tissues in this study without patients’ consent with the reason that the data were analyzed anonymously and patients could not be identified.

EST, SAGE, microarray gene expression dataset construction

EST and SAGE human gene expression data were collected from the publicly available Cancer Genome Anatomy Project (CGAP) site (http://cgap.nci.nih.gov/). Microarray data was obtained from the Cancer Genome Expression Database of LG Life Sciences, which is based on the Affymetrix HG-U133 (for the samples included in microarray data, see Table S1)[26]. A detailed description of each dataset construction is provided in the Supplementary Materials and Methods (Text S1) and shown schematically in Figure 1.

Algorithm for the identification of candidate HKGs and 13 nERGs

The methodology used to identify nERGs is outlined in Figure 1. First, to identify nERGs, we searched for candidate HKGs whose expression is detected in most tissues using 0’s proportion in EST, <0.1 in shortSAGE and <0.3 in longSAGE. 0’s prop represents 0’s proportion (number of tissues in which the gene is not expressed/total number of tissues, 0≤0’s prop≤1). Among the candidate 2,087 HKGs, 13 nERGs with the lowest CVs were further identified by selecting the genes common to all four datasets among the genes with the 400 lowest CVs (approximately 20% of candidate HKGs).

Figure 1. Flowchart of the methodology for identification of nERGs. 2,087 candidate HKGs were first identified by selecting the genes meeting the following criteria: 0’s prop <0.4 in EST, <0.1 in ShortSAGE and <0.3 in LongSAGE. 0’s prop represents 0’s proportion (number of tissues in which the gene is not expressed/total number of tissues, 0≤0’s prop≤1). Among the candidate 2,087 HKGs, 13 nERGs with the lowest CVs were further identified by selecting the genes common to all four datasets among the genes with the 400 lowest CVs (approximately 20% of candidate HKGs). doi:10.1371/journal.pone.0006162.g001
Identification and characterization of 2,087 candidate HKGs

The 2,087 candidate HKGs were first identified using 0’s proportion in each dataset (Figure 1, Table S2). According to the functional classification by FunCat [28], genes with a variety of basic cellular functions were included in this list. In particular, proteins mediating protein fate (23%) and cellular transport (21%) had the highest frequency (Figure 2A). This is in contrast to the previously reported classifications of HKGs, in which metabolic and ribosomal proteins were enriched [27,32]. We compared the frequency of genes with CpG islands in the upstream sequences of transcription start sites in HKGs relative to non-HKGs. Most HKGs (70%) were found to possess a CpG island within 1,000 bp from the transcription start site, consistent with previous studies [33,34], while fewer CpG islands were found in the upstream sequences of non-HKGs (P<0.001) (Table 2). Mean expression level of HKGs was significantly higher than that of non-HKGs in all datasets (P<0.001) (Figure 2B), also consistent with previous work [27] (for detailed description on the expression of 2,087 HKGs in 4 datasets, see Text S2). CV values of the 2,087 genes showed a poor correlation between the four datasets, whereas gene expression showed a relatively higher correlation (Table S4).

Identification and characterization of 13 nERGs

A total of 15 nERGs common to the four datasets (\(\text{ZNF207, OA2I, LUC7L2, CTBP1, TRIM27, GGBP1, ARLB, UBE2N1, POPOLA, CUL1, DMT1L, FBXW2, and SPEG}\), Table 3) were identified from 2,087 HKGs. The highest proportion (5/13) (\(\text{ZNF207, OA2I, CTBP1, POPOLA, FBXW2}\)) of genes were genes involved in cellular metabolism. CpG islands were found in the upstream region from transcription start site of all 13 nERGs (Table S3).

The gene expression for each of the 13 nERGs showed a significant correlation between datasets (P<0.001, Table S6) with high Pearson correlation coefficients (>0.8), although the Spearman correlations of EST versus Affymetrix (0.374,
| Gene Symbol | Title | Accession number | Probe* (UPL Probe No.) | Primer | Amplicon size (bp) | PCR efficiency (dilution)** | PCR efficiency (LinRegPCR)*** |
|-------------|-------|------------------|------------------------|--------|------------------|-----------------------------|-------------------------------|
| GAPDH       |       | NM_002046        | Left 18 agccccacatcgctcagaca | 66     | 1.899            | 1.735 ± 0.048 (137)         |                               |
| ACTB        |       | NM_001101        | Left 18 ccaccacgccggaagata | 97     | 2.038            | 1.491 ± 0.034 (137)         |                               |
| B2M         |       | NM_004048        | Right 20 ccagggagtgctgctgctgct | 86     | 1.868            | 1.717 ± 0.068 (140)         |                               |
| PPIA        |       | NM_021130 Specific Probe | Left 22 cactgtcactgcaagactgag | 326    | 1.877            | 1.773 ± 0.058 (142)         |                               |
| HPRT1       |       | NM_000194        | Left 24 tcaccttgatttttgctattc | 102    | 1.800            | 1.771 ± 0.024 (143)         |                               |
| HMBS        |       | NM_000190        | Left 25 tgtgtgggaaccagctc | 92     | 1.954            | 1.431 ± 0.031 (143)         |                               |
| TBP         |       | NM_003194        | Left 18 tgtgtgggaaccagctc | 92     | 1.954            | 1.431 ± 0.031 (143)         |                               |
| H6PD        |       | NM_004285        | Left 23 ccacacagctgctgctgctgct | 74     | 1.874            | 1.832 ± 0.026 (64)          |                               |
| ZNF207      |       | NM_003457        | Left 25 tgtgtgggaaccagctc | 92     | 1.954            | 1.431 ± 0.031 (143)         |                               |
| OAZ1        |       | NM_004152        | Left 18 caccatgccgctcctaag | 67     | 2.068            | 1.498 ± 0.059 (142)         |                               |
| LUC7L2      |       | NM_016019        | Left 18 caccatgccgctcctaag | 67     | 2.068            | 1.498 ± 0.059 (142)         |                               |
| CTBP1       |       | NM_01012614      | Left 18 actgtgggtaccctgact | 86     | 2.064            | 1.651 ± 0.055 (141)         |                               |
| TRIM27      |       | NM_005510        | Left 19 caccagctgctgctgctgctgct | 71     | 1.908            | 1.693 ± 0.034 (143)         |                               |
| GPBP1       |       | NM_022913        | Left 21 caccagctgctgctgctgctgct | 71     | 1.908            | 1.693 ± 0.034 (143)         |                               |
| UBLNL1      |       | NM_013438        | Left 19 caccagctgctgctgctgctgct | 71     | 1.908            | 1.693 ± 0.034 (143)         |                               |
| ARL8B       |       | NM_018184        | Left 19 aacatgtcagggagcctggtat | 66     | 1.838            | 1.499 ± 0.074 (139)         |                               |
| PAPOLA      | Poly(A) polymerase alpha | NM_032632 | Left 21 gctggcagacactgctgctgctgct | 91     | 1.830            | 1.509 ± 0.032 (141)         |                               |
| CUL1        | Cullin 1   | NM_003592        | Left 18 gctggcagacactgctgctgctgct | 86     | 1.810            | 1.695 ± 0.027 (139)         |                               |
| DIMT1L      | DIM1 dimethyladenosine transferase 1-like (S. cerevisiae) | NM_014473 | Left 27 ttccagtggtaagatgagaactaag | 75     | 1.906            | 1.655 ± 0.037 (141)         |                               |
| FBXW2       | F-box and WD-40 domain protein 2 | NM_012164 | Left 19 cggctctgacagagtcaactc | 111    | 1.891            | 1.638 ± 0.02 (142)          |                               |
A search for genomic variation of tERGs revealed that many of the nERGs, gene copy number variations (CNVs) of the nERGs, which can affect the gene expression, were investigated by searching the Database of Genomic Variants. As shown in Table 4, only 0.206 and 0.511 were found to be located in a chromosome region where CNVs were reported, indicating their expression might be deregulated by genomic aberrations.

### Comparison of tERGs and nERGs in dataset

We compared the 13 nERGs with 13 commonly used tERGs: GADPH, ACTB, HPR1T, PPLA, B2M, TBP, HMBS, RPLP0, PGK1, GUSB, TFRC, H6PD, and ALAS1. The mean expression of the nERGs was relatively lower than the highly expressed tERGs, including GADPH, ACTB, B2M, and PPLA, in all datasets and was expressed at levels similar to those tERGs, which had lower expression levels (Figure 3, Table S7). With respect to variation, most of the tERGs showed relatively higher variation than the nERGs and the mean CV values of nERGs were significantly lower than those of tERGs (P>0.05, Table S6), which had lower expression levels (Table 4, only 0.206 and 0.511 were found to be located in a chromosome region where CNVs were reported, indicating their expression might be deregulated by genomic aberrations.

A search for genomic variation of tERGs revealed that many of them (ACTB, GADPH, PGK1, B2M, TBP, TFRC, ALAS1) were located in the genomic locus where CNVs are known, whereas only 2 genes among the nERGs were found in those regions (Table 4), suggesting that the higher expression variation of some tERGs might be due to part in CNVs.

### Validation of nERGs by qRT-PCR

To validate both the level and stability of gene expression of nERGs selected from the four datasets, the expression of 13 nERGs and 8 tERGs was measured in a total of 108 human samples, including 26 frozen tissues, 22 cancer cell lines and 60 FFPE tissues, by qRT-PCR using Taqman probes (refer to Table S1 for an explanation for why 8 tERGs among 13 tERGs were chosen for qRT-PCR). Except PPIA, a small amplicon for each gene was designed for the reliable measurement of its expression, especially in FFPE tissues where RNA from these samples is frequently degraded. When the PCR efficiency of each gene was determined using the serial dilution method, each gene was amplified at 90–100% efficiency (Table 1). The CVs of Cq values confirmed that the between-assay precision in two or three repeats was within 5% (data not shown).

First, the expression profiles of these genes in each of the 48 samples, including frozen tissues and cancer cell lines, are presented in Figure 4A. The 13 nERGs were constitutively expressed in all 48 samples. Seven tERGs showed a wide range of expression (Cq: 13.32 to 29.39), but H6PD was not widely expressed in frozen tissues and this gene was consequently excluded from subsequent calculations. The Cq values of 13 nERGs ranged from 18.90 to 28.79 (Figure 4B). tERGs could be divided into highly expressed genes (median <20 cycles) and lowly expressed genes (median >20 cycles). Highly expressed genes included B2M, PPIA, GADPH, ACTB and lowly expressed genes consisted of HPR1T, TBP and HMBS. Of the nERGs, 12 genes displayed an intermediate expression level between the highly expressed and the lowly expressed tERGs (Figure 4B). The mean expression level of nERGs was significantly lower than that of highly expressed tERGs, whereas it was significantly higher than that of lowly expressed tERGs (P<0.001, Table S9). GAPDH was the most highly expressed gene, followed by UBP1 and CUL1. OAZ1 was the gene with the weakest expression.

We further investigated the expression of the 13 nERGs by qRT-PCR in 60 FFPE tissues to test whether the nERGs could be used in such tissues showing the significant degradation of mRNA. Except DMT1L, the expression of all genes was measurable in all 60 samples. The Cq range for tERGs was 19.95 to 33.02 and for nERGs was 23.33 to 31.58 (Figure 4B). DMT1L was not amplified in 5 samples and therefore was excluded from further expression stability analysis. The expression pattern in the FFPE tissues was similar to that of previous 48 samples (26 frozen tissues and 22 cancer lines) despite the discrepancy in sample types. Remarkably, PPLA expression which was detected at high level in frozen tissues/cell lines was observed at markedly decreased level in FFPE tissues. This observation might be due to the long amplicon size of PPLA (326 bp), whereas the amplicon size of other genes is small ranging from 60 to 110 bp (Table 1), indicating that small size of amplicon is required for the detection of gene expression in FFPE tissues in which RNA is frequently degraded.

### Gene expression stability of nERGs

We first assessed the gene expression stability (detailed in Text S1) in 48 samples, including 26 frozen tissues and 22 cell lines based on qRT-PCR using two programs, geNorm and NormFinder. All genes tested displayed relatively high expression stability with low M values (<0.9), which were below the default limit of 1.5 in geNorm (Table 5a). GPBP1 and CUL1 were identified as the two most stable genes. B2M was the least stable gene and had the highest M value (0.886), followed by ACTB (0.843), HMBS (0.815), and GADPH (0.793). When calculated by NormFinder, TBP and...
Figure 2. Characterization of candidate HKGs identified in this study. (A) The functional distribution of candidate HKGs was classified by FunCat. Among the 2,087 UniGene clusters, a total of 1,605 UniGene clusters which have GO terms under biological processes were classified according to their major functional categories using FunCat (version 2.0) by mapping the GO terms to FunCat categories. A total of 1,318 UniGene clusters were classified by FunCat. The number of UniGene clusters belonging to each category is presented in A. In some cases, UniGene clusters mapped to two or more FunCat categories. (B) Comparison of gene expression between candidate HKGs and non-HKGs in each dataset. Box and Whisker plots provide a simple description of a distribution of values by depicting the 25th and 75th percentile values as the bottom and top of a box, respectively. The Y axis represents the natural logarithm transformed mean gene expression levels. The median expression values of HKGs and non-HKGs are marked by horizontal lines in the boxes and the values are provided to the right of each box. *P<0.001.

doi:10.1371/journal.pone.0006162.g002
PAPOLA were the two most stable genes (i.e. having the lowest S values) (Table 5a). Similar to the results from geNorm, tERGs including B2M, ACTB, GAPDH, HMBS and HPRT1, were found to have less stable expression than nERGs. The mean gene expression stability values in nERGs by geNorm and NormFinder were significantly lower than those in tERGs in both 48 samples and 60 FFPE tissues (P<0.05, Tables S10). Both analyses demonstrated that most nERGs showed relatively higher expression stability compared to tERGs, suggesting that nERGs are more suitable for normalization. Moreover, even when gene expression stability was analyzed with relative expression level calculated by PCR efficiencies using the LinRegPCR program, most of nERGs showed a more stable expression than tERGs (data not shown).

Pearson correlation analysis revealed the higher concordance of both M and S values with CVs in the EST and shortSAGE (Table 6) than in the Affymetrix. High correlation between M and S (Pearson correlation: 0.953, P<0.001) was also observed, indicating that both analyses produced similar results.

Consistently with frozen tissues and cell lines, both gene expression stability values demonstrated that most nERGs with low stability values are superior to tERGs in terms of expression stability in FFPE tissues (Table 5b), proving the usefulness of our nERGs as reliable measurements of gene expression in those tissues. GPBP1 and PAPOLA were the two least variable genes in geNorm and NormFinder were the top two ranked genes in NormFinder. However, in the analysis by each tissue, FBXW2, TRIM27, and CUL1 showed high stability values in breast, ovary, and stomach, respectively (Table 7), suggesting that they have high expression variation in each tissue. Also, S values by NormFinder in the ovary and stomach FFPE tissues were calculated based on the combination of intra- and inter-group variations between normal and tumor samples. The relatively high S values of TRIM27 in the ovary and CUL1 in the stomach suggest that their expression might be regulated in specific tumors compared to their normal tissues.

The optimal number of ERGs for normalization was determined using geNorm. In both the 48 human frozen and cell line samples and 60 FFPE tissues, the optimal number of nERGs required for normalization was fewer than when using tERGs (Figure 5). Four tERGs and three nERGs were calculated as the optimal number of ERGs needed in the 48 samples when using a V of 0.15 as the cut-off value [1]. In the FFPE samples, V2/3 was under 0.15 when using nERGs, suggesting that only two genes are sufficient for optimal normalization, whereas four of seven tERGs were necessary for accurate normalization. This indicates that fewer ERGs are required for optimal normalization when using our nERGs rather than using tERGs.

Discussion

In the present study, we identified nERGs in human samples using a comparative analysis of different large datasets of human gene expression profiles, while previous attempts to identify nERGs that are superior to tERGs were limited to the analysis of a microarray [10,18–21] or EST data [35]. Candidate HKGs were first selected and included 2,087 genes, which is a larger number of genes than previously identified by other groups [27,32,36]. Their characteristics, including high levels of expression [27,36] and the prevalence of CpG islands in the promoter regions [33,34], were in line with previous studies based on smaller numbers of HKGs, reflecting that “real” HKGs showing constitutive expressions in all tissues are enriched in our list. Thus, this list can be used as a reliable source for the study of HKGs.

The 13 nERGs further identified from candidate HKGs showed relatively lower CV and lower expression than most of the tERGs in all datasets. These findings were further confirmed by qRT-PCR using frozen tissues and cell lines. Generally, the expression of 13 nERGs was lower than the highly expressed tERGs and higher or similar to the weakly expressed tERGs. The expression stability values of the nERGs calculated by both programs also demonstrated that nERGs are generally more stably expressed than tERGs. Although there were slight differences in their rankings between the two programs, PAPOLA, CUL1, TBP, LUC7L2, GPBP1 and TRIM27 were found to be the genes with the most stable expression in 48 samples. The observation that TBP is one of the most stable genes is not surprising because relatively lower variability of TBP among traditional ERGs was already expected in datasets including EST and ShortSAGE (Figure 3). On the other hand, our data further supported the unsuitability of the most commonly used ERGs, like GAPDH, ACTB, for normalization, in line with previous works [7,8,14].

Our nERGs were also successfully validated by qRT-PCR in FFPE tissues. Despite the usefulness of archival FFPE tissue specimens in conjunction with clinical data, frequent degradation of RNA from FFPE tissues has been regarded as an obstacle in the gene expression analysis of those samples [37]. The expression of nERGs was measurable with reliable Cp values in all 60 FFPE tissues and most of the nERGs outperformed tERGs with respect to expression stability.
Furthermore, the expression level of target gene can be calculated relative to the expression level of one or multiple nERGs using standard curve or comparative Ct(Cp) method in quantitative gene expression analyses, including qRT-PCR. The most suitable ERG or ERGs in the designed experiment can be selected among 13 nERGs or combination of tERGs and 13 nERGs based on gene expression stability values calculated by the geNorm and/or NormFinder program. Recently developed PCR array, high throughput gene expression measurement using qRT-PCR, also requires more suitable ERGs than conventional tERGs for accurate quantification of gene expression [38]. Recently, normalization using the geometric mean of multiple ERGs has been considered to be more accurate for normalization [1], especially in situations when no optimal ERG has been identified.

Table 3. nERGs identified from four datasets.

| UniGene cluster | Gene Symbol | Gene Title | EST | SHORT SAGE | LONG SAGE | Affymetrix | Gene Ontology | Biological Process | Molecular Function |
|-----------------|-------------|------------|-----|------------|-----------|------------|---------------|---------------------|--------------------|
| Hs.446427       | OAZ1        | Ornithine decarboxylase antizyme 1 | 673.68 | 62.71 | 0.069 | 576.39 | 55.94 | 0 | 444.92 | 41.01 | 0 | 1860.9 | 22.07 | Polyamine biosynthesis | Ornithine decarboxylase inhibitor activity |
| Hs.9589         | UBQLN1      | Ubiquilin 1 | 111.34 | 60.19 | 0.31 | 75.93 | 61.3 | 0.036 | 71.77 | 49.75 | 0.111 | 919.32 | 26.81 | Kinase binding |
| Hs.444279       | GPBP1       | GC-rich promoter binding protein 1 | 132.98 | 56.92 | 0.24 | 63.11 | 61.65 | 0.001 | 80.72 | 51.79 | 0.111 | 746.56 | 26.6 | |
| Hs.208597       | CTBP1       | C-terminal binding protein 1 | 136.74 | 48.53 | 0.13 | 213.99 | 62.0 | 0 | 112.96 | 50.6 | 0 | 481.51 | 24.72 | Negative regulation of cell proliferation; Protein phosphorylation; Viral genome replication | Protein C-terminus binding; Transcription factor binding |
| Hs.253726       | PAPOLA      | Poly(A) polymerase alpha | 216.15 | 65.36 | 0.17 | 118.24 | 58.02 | 0 | 89.5 | 50.88 | 0 | 451.23 | 27.68 | mRNA polyadenylation | RNA binding |
| Hs.250009       | ARL8B       | ADP-ribosylation factor-like 8B | 132.27 | 55.79 | 0.39 | 134.21 | 61.55 | 0 | 59.19 | 55.14 | 0.111 | 418.28 | 26.81 | Chromosome segregation | α-tubulin binding; β-tubulin binding; GDP binding; GTP binding; GTPase activity |
| Hs.242458       | SPG21       | Spastic paraplegia 21 (autosomal recessive, Mast syndrome) | 120.35 | 59.44 | 0.31 | 76.41 | 57.83 | 0.036 | 73.3 | 49.12 | 0 | 415.64 | 29.35 | Antigen receptor-mediated signaling pathway | CD4 receptor binding |
| Hs.530118       | LUC7L2      | LUC7-like 2 (S. cerevisiae) | 132.76 | 59.55 | 0.17 | 74.65 | 65.41 | 0 | 57.21 | 50.39 | 0.111 | 386.79 | 22.9 | |
| Hs.500775       | ZNF207      | Zinc finger protein 207 | 233.29 | 62.27 | 0.03 | 165.68 | 56.77 | 0 | 154.32 | 52.88 | 0.111 | 358.69 | 18.38 | Regulation of transcription, DNA-dependent | Transcription factor activity; Zinc ion binding |
| Hs.533222       | DIMT1L      | Dimethyladenosine transferase 1-like (S. cerevisiae) | 129.17 | 69.14 | 0.57 | 42.41 | 60.55 | 0.071 | 36.87 | 44.13 | 0.111 | 164.72 | 28.3 | |
| Hs.440382       | TRIM27      | Tripartite motif containing 27 | 155.33 | 68.54 | 0.17 | 80.66 | 63.06 | 0 | 67.84 | 45.41 | 0.111 | 163.6 | 26.3 | Cell proliferation; Spermatogenesis | Metal ion binding; Transmembrane receptor protein tyrosine kinase activity |
| Hs.146806       | CUL1        | Culin 1 | 120.27 | 57.5 | 0.2 | 69.33 | 65.76 | 0.036 | 76.43 | 55 | 0.111 | 156.47 | 27.78 | Cell cycle arrest; G1/S transition of mitotic cell cycle; Induction of apoptosis by intracellular signals; Negative regulation of cell proliferation | Protein binding |
| Hs.494985       | FBXW2       | F-box and WD-40 domain protein 2 | 97.45 | 68.65 | 0.37 | 40.27 | 58.47 | 0 | 23.54 | 51.61 | 0.111 | 69.32 | 28.36 | Proteolysis | Protein binding; ubiquitin conjugating enzyme activity; ubiquitin-protein ligase activity |

Mean: Mean gene expression, CV: Coefficient of Variation (%), 0's P: 0’s proportion, GO terms were searched in the Gene Ontology site (http://www.geneontology.org/). doi:10.1371/journal.pone.0006162.t003
analysis using FFPE tissues. The superiority of nERGs over tERGs is based in their lower expression stability. Use of ERGs with expression levels similar to target genes is recommended so that the comparisons fall on the same linear scale [39]. Therefore, nERGs, with relatively lower expression, rather than GAPDH or ACTB showing high expression, can be better candidates for normalization of a wide range of genes, including weakly expressed genes. This is significant given that the majority of transcripts in human tissues are expressed in low abundance [36].

Remarkably, we observed no significant correlation between the stability values calculated from qRT-PCR data and CV in the Affymetrix data, which is in contrast to the significant correlation between stability values and CV in the EST and SAGE dataset. This suggests that EST or shortSAGE may be better sources for exploring nERGs rather than microarrays, which have traditionally been used as a source for screening ERG and supports our initial hypothesis that the microarray might not be a good source for ERGs.

Furthermore, nERGs identified here might be used as reference for relative measurements of gene amplification, which is a frequent genetic alteration leading to unregulated gene expression in cancer [43]. As the relatively constant expression of these genes in both normal and tumor tissues provides the possibility that these genes are located in a chromosomal region in which no genetic alterations are found in human tumors, we investigated the genomic CNVs of nERGs using publicly available databases. Most nERGs, except OAZ1 and DMT1L, were located in genomic regions where CNVs were not reported, whereas many tERGs were located in regions with CNVs. The relatively lower expression stability of OAZ1 and DMT1L, as well as tERGs like GAPDH and ACTB, might be explained by these genomic aberrations. However, the suitability of nERGs as a reference for the relative measurement of gene amplification remains to be further investigated and validated through experiments. Meanwhile, even genes with genomic variations can be used for the normalization in gene expression, provided that their expression is not affected by genomic aberrations.

In conclusion, we have identified a set of candidate HKGs and nERGs based on a comparative analysis of EST, SAGE, and Affymetrix datasets. This is the first study using three different platforms to identify nERGs, and most of the 13 ERGs identified in these datasets were demonstrated to be more stably expressed than tERGs through the validation using qRT-PCR in large human samples including cell lines, frozen tissues and FFPE tissues. Moreover, these nERGs were expressed at relatively lower levels than most commonly used high expressing tERGs, making them more suitable for normalization of transcripts from a wide range of expression levels. We have also shown that fewer ERGs are required for accurate normalization using nERGs than using tERGs, especially in FFPE tissues when the use of multiple ERGs is required.
Figure 3. Comparison of gene expression and CV between nERGs and tERGs in each dataset. (A) Comparison of gene expression between 13 nERGs and 13 tERGs. (B) Comparison of CV between 13 nERGs and 13 tERGs. Empty squares represent nERGs identified in this study and circles represent the tERGs.
doi:10.1371/journal.pone.0006162.g003
Figure 4. The distribution of expression levels of 13 nERGs and 7 tERGs determined by qRT-PCR using Taqman probes in human samples. (A) The distribution of mRNA levels of tested ERGs in 48 samples, including frozen tissues and cancer cell lines. (B) The mRNA levels of tERGs (red) and nERGs (blue) in Cp values over all 48 samples (left) and 60 FFPE tissues (right). Values are given as “Crossing point” (Cp) values. All measurements of qRT-PCR were repeated three times for frozen tissues and cell lines and twice for FFPE tissues and mean “crossing point” (Cp) values of repeats were calculated. Box and Whisker plots provide a simple description of the distribution of values by depicting the 25th and 75th percentile values as the bottom and top of the box, respectively. The median value is marked by a line within the box and the minimum and maximum values are depicted by error bars, or whiskers, protruding from the box.

doi:10.1371/journal.pone.0006162.g004
Table 5. nERGs and tERGs ranked according to their expression stability, as calculated by the two programs, geNorm and NormFinder, based on qRT-PCR data in 48 frozen tissues/cell lines and 60 FFPE tissues.

(a) 48 frozen tissues/cell lines

| Gene Symbol | Average expression stability M | Stability value S |
|-------------|-------------------------------|------------------|
| GPBP1       | 0.496                         | TBP              |
| CUL1        | 0.536                         | PAPOLA           |
| TBP         | 0.548                         | LUC7L2           |
| LUC7L2      | 0.565                         | CUL1             |
| TRIM27      | 0.585                         | FBXW2            |
| CTBP1       | 0.608                         | CTBP1            |
| UBQLN1      | 0.623                         | DIMT1L           |
| DIMT1L      | 0.637                         | PPIA             |
| PPIA        | 0.661                         | UBQLN1           |
| OAZ1        | 0.682                         | OAZ1             |
| ZNF207      | 0.709                         | ARL8B            |
| ARL8B       | 0.731                         | SPG21            |
| SPG21       | 0.749                         | ZNF207           |
| HPRT1       | 0.77                          | HPRT1            |
| GAPDH       | 0.793                         | HMBS             |
| HMB5        | 0.815                         | GAPDH            |
| ACTB        | 0.843                         | ACTB             |
| B2M         | 0.888                         | B2M              |

(b) 60 FFPE tissues

| Gene Symbol | Average expression stability M | Stability value S |
|-------------|-------------------------------|------------------|
| GPBP1       | 0.496                         | ARL8B            |
| PAPOLA      | 0.28                          | PAPOLA           |
| ARL8B       | 0.437                         | CTBP1            |
| CTBP1       | 0.454                         | LUC7L2           |
| FBXW2       | 0.528                         | FBXW2            |
| OAZ1        | 0.545                         | PAPOLA           |
| TRIM27      | 0.555                         | UBQLN1           |
| UBQLN1      | 0.567                         | TRIM27           |
| PPRIA       | 0.383                         | TRIM27           |
| RTB         | 0.398                         | HPRT1            |
| ARL8B       | 0.494                         | HMB5             |
| SPG21       | 0.502                         | ZNF207           |
| ZNF207      | 0.502                         | ZNF207           |
| GBPDH       | 0.568                         | GAPDH            |
| B2M         | 0.715                         | B2M              |
| ACTB        | 0.737                         | ACTB             |
| B2M         | 0.815                         | B2M              |

Low average expression stability value M and stability value S indicate the high expression stability.
doi:10.1371/journal.pone.0006162.t005

Table 6. Correlation between gene expression stability of nERGs and tERGs from qRT-PCR data and CV from each dataset.

| 48 frozen tissues/cell lines | EST-M   | EST-S   | ShortSAGE-M | ShortSAGE-S | LongSAGE-M | LongSAGE-S | Affy-M  | Affy-S | M-S    |
|-----------------------------|---------|---------|-------------|-------------|------------|------------|---------|--------|-------|
| Pearson                     | 0.676   | 0.792   | 0.659       | 0.75        | 0.427      | 0.561      | 0.039   | 0.017  | 0.953 |
| P value                     | <0.001  | <0.001  | <0.001      | <0.001      | 0.01       | 0.869      | 0.944   | <0.001 |
| Spearman                    | 0.589   | 0.605   | 0.277       | 0.268       | 0.092      | 0.105      | 0.424   | 0.357  | 0.955 |
| P value                     | 0.006   | 0.005   | 0.237       | 0.254       | 0.701      | 0.661      | 0.063   | 0.123  | <0.001|
| 60 FFPE tissues             | EST-M   | EST-S   | ShortSAGE-M | ShortSAGE-S | LongSAGE-M | LongSAGE-S | Affy-M  | Affy-S | M-S    |
| Pearson                     | 0.623   | 0.626   | 0.656       | 0.737       | 0.481      | 0.672      | 0.243   | 0.335  | 0.852 |
| P value                     | 0.004   | 0.004   | 0.002       | <0.001      | 0.037      | 0.002      | 0.317   | 0.161  | <0.001|
| Spearman                    | 0.663   | 0.596   | 0.515       | 0.502       | 0.374      | 0.567      | 0.521   | 0.583  | 0.841 |
| P value                     | 0.002   | 0.008   | 0.024       | 0.03        | 0.115      | 0.013      | 0.022   | 0.009  | <0.001|

M: average expression stability calculated by the geNorm program, S: stability value calculated by the NormFinder program. For 48 samples, 20 reference genes including 13 novel genes and 7 classical genes, were used in the correlation analysis. For 60 FFPE tissues, 19 reference genes excluding DIMT1L were included in the analysis.
doi:10.1371/journal.pone.0006162.t006
**Table 7.** nERGs and tERGs ranked according to their expression stability, as calculated by the two programs, geNorm and NormFinder, based on qRT-PCR data in each tissue type of FFPE tissues.

| Gene Symbol | Average expression stability M | Gene Symbol | Stability value S |
|-------------|-------------------------------|-------------|-------------------|
| TBP         | 0.186                         | LUC7L2      | 0.080             |
| CTBP1       | 0.199                         | UQBLN1      | 0.138             |
| PAPOLA      | 0.231                         | CUL1        | 0.153             |
| OAZ1        | 0.245                         | TBP         | 0.173             |
| ARL8B       | 0.265                         | SPG21       | 0.190             |
| GPBP1       | 0.298                         | ARL8B       | 0.190             |
| LUC7L2      | 0.307                         | SPG21       | 0.190             |
| ZNF207      | 0.317                         | CUL1        | 0.153             |
| B2M         | 0.335                         | HMB5        | 0.240             |
| SPG21       | 0.356                         | PAPOLA      | 0.246             |
| HMB5        | 0.374                         | ZNF207      | 0.305             |
| TRIM27      | 0.389                         | HPRT1       | 0.253             |
| PPIA        | 0.417                         | ACTB        | 0.316             |
| ACTB        | 0.438                         | PPIA        | 0.317             |
| HPRT1       | 0.464                         | HMB5        | 0.390             |
| FBXW2       | 0.489                         | FBXW2       | 0.447             |
| GAPDH       | 0.530                         | GAPDH       | 0.570             |
| Best combination of two genes | ARB1L8B and SPG21 | 0.043 | Best combination of two genes | CTBP1 and UQBLN1 | 0.066 |

*Gene expression stability S by NormFinder was calculated as an estimate of the combined intra and intergroup variation between normal and tumor tissues. For ovary tissues (n = 33), 10 normal and 23 tumor tissues were included and 17 stomach tissues, including normal (n = 8) and tumor (n = 9) tissues, were used in the analysis. Low average expression stability M and stability value S indicate the high expression stability.

doi:10.1371/journal.pone.0006162.t007

**Supporting Information**

**Text S1** Supplementary materials and methods
Found at: doi:10.1371/journal.pone.0006162.s001 (0.09 MB DOC)

**Text S2** Expression of 2,087 candidate HKGs in the four datasets
Found at: doi:10.1371/journal.pone.0006162.s002 (0.04 MB DOC)

**Table S1** List of 567 samples including 13 tissue types in the HG-U133 array used in this study
Found at: doi:10.1371/journal.pone.0006162.s003 (0.07 MB DOC)

**Table S2** A list of 2,087 candidate housekeeping genes
Found at: doi:10.1371/journal.pone.0006162.s004 (0.86 MB XLS)

**Table S3** Human frozen tissues and cancer cell lines used in qRT-PCR
Found at: doi:10.1371/journal.pone.0006162.s005 (0.08 MB DOC)

**Table S4** Correlation of gene expression and CV of 2,087 candidate HKGs between the four datasets
Found at: doi:10.1371/journal.pone.0006162.s006 (0.05 MB DOC)

**Table S5** CpG islands analysis in the upstream region from transcription start site of 13 nERGs
Found at: doi:10.1371/journal.pone.0006162.s007 (0.04 MB DOC)

**Table S6** Correlation of gene expression and CV of 13 nERGs between the four datasets
Found at: doi:10.1371/journal.pone.0006162.s008 (0.05 MB DOC)

**Table S7** tERGs used in this study
Found at: doi:10.1371/journal.pone.0006162.s009 (0.05 MB DOC)

**Table S8** Comparison of CV between nERGs and tERGs in the dataset
Found at: doi:10.1371/journal.pone.0006162.s010 (0.04 MB DOC)

**Table S9** Comparison of Cp values between nERGs and tERGs in qRT-PCR
Found at: doi:10.1371/journal.pone.0006162.s011 (0.04 MB DOC)
Variable V defines the pair-wise variation between two sequential normalization factors containing an increasing number of genes. For example, V2/3 indicates the variation of the normalization factor of two genes in relation to three genes. A large V indicates that the added gene should be included for calculation of the normalization factor. 0.15 was proposed as a cut-off value, below which the inclusion of an additional reference gene is not required. Pair-wise variation analysis to determine the number of ERGs required for accurate normalization was performed in 48 samples including human frozen tissues and cell lines (A), 60 FFPE tissues (B), 33 ovary FFPE tissues (C) and 17 stomach FFPE tissues (D). For each case, the analysis was done for total ERGs, tERGs and nERGs.

doi:10.1371/journal.pone.0006162.g005
Table S10 Comparison of gene expression stability values between nERGs and tERGs

Found at: doi:10.1371/journal.pone.0006162.s012 (0.04 MB DOC)

Acknowledgments

We are very grateful to Dr. Soomyung Paik (Division of Pathology, National Surgical Adjuvant Breast and Bowel Project, Pittsburgh, PA) and Dr. Jonathan R. Pollack (Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA) for thoughtful suggestions and critical comments.

Author Contributions

Conceived and designed the experiments: MJK YKS. Performed the experiments: MJK SEK. Analyzed the data: MJK EO SL MRR YHL. Contributed reagents/materials/analysis tools: YL YLC SSK. Wrote the paper: MJK. Gave advice on the statistical analysis: TP.