Research

Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes

Jo Vandesompele, Katleen De Preter, Filip Pattyn, Bruce Poppe, Nadine Van Roy, Anne De Paepe and Frank Speleman

Address: Center for Medical Genetics, Ghent University Hospital 1K5, De Pintelaan 185, B-9000 Ghent, Belgium.

Correspondence: Frank Speleman. E-mail: franki.speleman@rug.ac.be

Abstract

Background: Gene-expression analysis is increasingly important in biological research, with real-time reverse transcription PCR (RT-PCR) becoming the method of choice for high-throughput and accurate expression profiling of selected genes. Given the increased sensitivity, reproducibility and large dynamic range of this methodology, the requirements for a proper internal control gene for normalization have become increasingly stringent. Although housekeeping gene expression has been reported to vary considerably, no systematic survey has properly determined the errors related to the common practice of using only one control gene, nor presented an adequate way of working around this problem.

Results: We outline a robust and innovative strategy to identify the most stably expressed control genes in a given set of tissues, and to determine the minimum number of genes required to calculate a reliable normalization factor. We have evaluated ten housekeeping genes from different abundance and functional classes in various human tissues, and demonstrated that the conventional use of a single gene for normalization leads to relatively large errors in a significant proportion of samples tested. The geometric mean of multiple carefully selected housekeeping genes was validated as an accurate normalization factor by analyzing publicly available microarray data.

Conclusions: The normalization strategy presented here is a prerequisite for accurate RT-PCR expression profiling, which, among other things, opens up the possibility of studying the biological relevance of small expression differences.

Background

Gene-expression analysis is increasingly important in many fields of biological research. Understanding patterns of expressed genes is expected to provide insight into complex regulatory networks and will most probably lead to the identification of genes relevant to new biological processes, or implicated in disease. Two recently developed methods to measure transcript abundance have gained much popularity and are frequently applied. Microarrays allow the parallel analysis of thousands of genes in two differentially labeled RNA populations [1], while real-time RT-PCR provides the simultaneous measurement of gene expression in many different samples for a limited number of genes, and is especially suitable when only a small number of cells are available [2-4]. Both techniques have the advantage of speed, throughput and a high degree of potential automation compared to conventional quantification methods, such as northern-blot analysis, ribonuclease protection assay, or

competitive RT-PCR. Nevertheless, these new approaches require the same kind of normalization as the traditional methods of mRNA quantification.

Several variables need to be controlled for in gene-expression analysis, such as the amount of starting material, enzymatic efficiencies, and differences between tissues or cells in overall transcriptional activity. Various strategies have been applied to normalize these variations. Under controlled conditions of reproducible extraction of good-quality RNA, the gene transcript number is ideally standardized to the number of cells, but accurate enumeration of cells is often precluded, for example when starting with solid tissue. Another frequently applied normalization scalar is the RNA mass quantity, especially in northern blot analysis. There are several arguments against the use of mass quantity. The quality of RNA and related efficiency of the enzymatic reactions are not taken into account. Moreover, in some instances it is impossible to quantify this parameter, for example, when only minimal amounts of RNA are available from microdissected tissues. Probably the strongest argument against the use of total RNA mass for normalization is the fact that it consists predominantly of rRNA molecules, and is not always representative of the mRNA fraction. This was recently evidenced by a significant imbalance between rRNA and mRNA content in approximately 7.5% of mammary adenocarcinomas [5]. Also, it has been reported that rRNA transcription is affected by biological factors and drugs [6-8]. Further drawbacks to the use of 18S or 28S rRNA molecules as standards are their absence in purified mRNA samples, and their high abundance compared to target mRNA transcripts. The latter makes it difficult to accurately subtract the baseline value in real-time RT-PCR data analysis.

To date, internal control genes are most frequently used to normalize the mRNA fraction. This internal control - often referred to as a housekeeping gene - should not vary in the tissues or cells under investigation, or in response to experimental treatment. However, many studies make use of these constitutively expressed control genes without proper validation of their presumed stability of expression. But the literature shows that housekeeping gene expression - although occasionally constant in a given cell type or experimental condition - can vary considerably (reviewed in [9-12]). With the increased sensitivity, reproducibility and large dynamic range of real-time RT-PCR methods, the requirements for a proper internal control gene have become increasingly stringent. In this study, we carried out an extensive evaluation of 10 commonly used housekeeping genes in 13 different human tissues, and outlined a procedure for calculating a normalization factor based on multiple control genes for more accurate and reliable normalization of gene-expression data. Furthermore, this normalization factor was validated in a comparative study with frequently applied microarray scaling factors using publicly available microarray data.

Results
Expression profiling of housekeeping genes
Primers were designed for ten commonly used housekeeping genes (*ACTB*, *B2M*, *GAPD*, *HMBS*, *HPRT1*, *RPL13A*, *SDHA*, *TBP*, *UBC* and *YWHAZ*) (see Table 1 for full gene name, accession number, function, chromosomal localization, alias, existence of processed pseudogenes, and indication that primers span an intron; see Table 2 for primer sequences). Special attention was paid to selecting genes that belong to different functional classes, which significantly reduces the chance that genes might be co-regulated. The expression level of these 10 internal control genes was determined in 34 neuroblastoma cell lines (independently prepared in different labs from different patients), 20 short-term cultured normal fibroblast samples from different individuals, 13 normal leukocyte samples, 9 normal bone-marrow samples, and 9 additional normal human tissues from pooled organs (heart, brain, fetal brain, lung, trachea, kidney, mammary gland, small intestine and uterus). The raw expression values are available as a tab-delimited file (see Additional data files).

Single control normalization error
To determine the possible errors related to the common practice of using only one housekeeping gene for normalization, we calculated the ratio of the ratios of two control genes in two different samples (from the same tissue panel) and termed it the single control normalization error, E (see Materials and methods). For two ideal internal control genes (constant ratio between the genes in all samples), E equals 1. In practice, observed E values are larger than 1 and constitute the erroneous E-fold expression difference between two samples, depending on the particular housekeeping gene used for normalization. E values were calculated for all 45 two-by-two combinations of control genes and 865 two-by-two sample combinations within the available tissue panels (neuroblastoma, fibroblast, leukocyte, bone marrow and a series of normal tissues from Clontech; that is, a total of 38,925 data points) (Figure 1). In addition, the systematic error distribution was calculated by analysis of repeated runs of the same control gene. The average 75th and 90th percentile E values are 3.0 (range 2.1-3.9), and 6.4 (range 3.0-10.9), respectively.

Gene-stability measure and ranking of selected housekeeping genes
It is generally accepted that gene-expression levels should be normalized by a carefully selected stable internal control gene. However, to validate the presumed stable expression of a given control gene, prior knowledge of a reliable measure to normalize this gene in order to remove any nonspecific variation is required. To address this circular problem, we developed a gene-stability measure to determine the expression stability of control genes on the basis of non-normalized expression levels. This measure relies on the principle that the expression ratio of two ideal internal control genes is

Table 1

Internal control genes evaluated in this study

Symbol	Accession number	Name	Function	Localization	Pseudo-gene*	Primers†	Alias	IMAGE‡
ACTB	NM_001101	Beta actin	Cytoskeletal structural protein	7p15-p12	+	S		510455
B2M	NM_004048	Beta-2-microglobulin	Beta-chain of major histocompatibility complex class I molecules	15q21-q22	-	S		51940
GAPD	NM_002046	Glyceraldehyde-3-phosphate dehydrogenase	Oxidoreductase in glycolysis and gluconeogenesis	12p13	+	D		510510
HMBS	NM_000190	Hydroxymethyl-bilane synthase	Heme synthesis, porphyrin metabolism	11q23	-	D	Porphobilinogen deaminase	245564
HPRT1	NM_000194	Hypoxanthine phosphoribosyl-transferase 1	Purine synthesis in salvage pathway	Xq26	+	D		345845
RPL13A	NM_012423	Ribosomal protein L13a	Structural component of the large 60S ribosomal subunit	19q13	+	D	23 kDa highly basic protein	-
SDHA	NM_004168	Succinate dehydrogenase complex, subunit A	Electron transporter in the TCA cycle and respiratory chain	5p15	+	D		375812
TBP	NM_003194	TATA box binding protein	General RNA polymerase II transcription factor	6q27	-	D		280735
UBC	M26880	Ubiquitin C	Protein degradation	12q24	-	D		510582
YWHAZ	NM_003406	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	Signal transduction by binding to phosphorylated serine residues on a variety of signaling molecules	2p25	+	S§	Phospholipase A2	416026

*Presence (+) or absence (-) of a retropseudogene in the genome determined by BLAST analysis of the mRNA sequence using the high-throughput genomic sequences database (htgs) or human genome as database. †Localization of forward and reverse primer in different exons (D) or the same exon (S). ‡IMAGE cDNA clone number according to [14]. §A single-exon gene.

Table 2

Primer sequences for internal control genes

Symbol*	Forward primer	Reverse primer
ACTB	CTGGAACGGTGAAGGTGACA	AAGGGACTTCCTGTAACAATGCA
B2M	TGCTGTCTCCATGTTTGATGTATCT	TCTCTGCTCCCCACCTCTAAGT
GAPD	TGCACCACCAACTGCTTAGC	GGCATGGACTGTGGTCATGAG
HMBS†	GGCAATGCGGCTGCAA	GGGTACCCACGCGAATCAC
HPRT1	TGACACTGGCAAAACAATGCA	GGTCCTTTTCACCAGCAAGCT
RPL13A	CCTGGAGGAGAAGAGGAAAGAGA	TTGAGGACCTCTGTGTATTTGTCAA
SDHA	TGGGAACAAGAGGGCATCTG	CCACCACTGCATCAAATTCATG
UBC	ATTTGGGTCGCGGTTCTTG	TGCCTTGACATTCTCGATGGT
YWHAZ	ACTTTTGGTACATTGTGGCTTCAA	CCGCCAGGACAAACCAGTAT

*TBP primer sequences are described in [24]. †HMBS primer sequences kindly provided by E. Mensink and L. van de Locht (Nijmegen, The Netherlands).

identical in all samples, regardless of the experimental condition or cell type. In this way, variation of the expression ratios of two real-life housekeeping genes reflects the fact that one (or both) of the genes is (are) not constantly expressed, with increasing variation in ratio corresponding to decreasing expression stability. For every control gene we determined the pairwise variation with all other control genes as the standard deviation of the logarithmically transformed expression ratios, and defined the internal control gene-stability measure M as the average pairwise
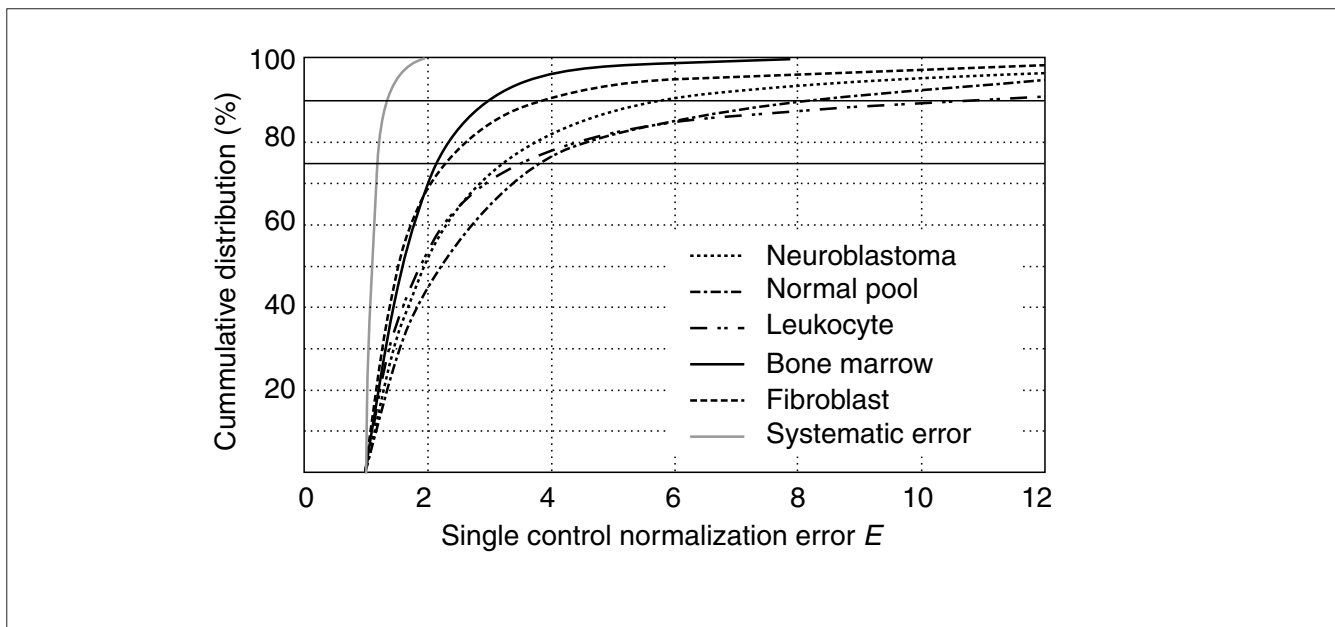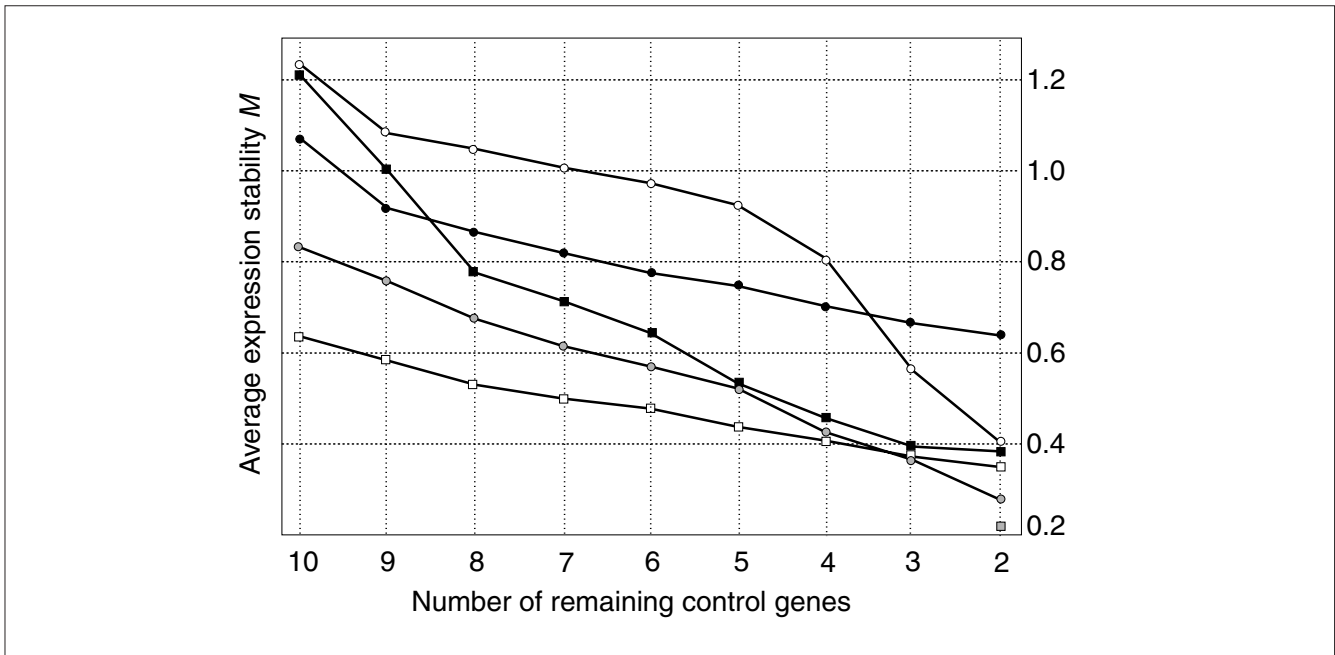
Figure 1
Single control normalization error values (*E*) were calculated as the ratio of the ratio of two control genes in two different samples (see Materials and methods), and summarized here as cumulative distribution plots for the different tissue panels, pointing at considerable variation in housekeeping gene expression.

variation of a particular gene with all other control genes. Genes with the lowest *M* values have the most stable expression. Assuming that the control genes are not co-regulated, stepwise exclusion of the gene with the highest *M* value results in a combination of two constitutively expressed housekeeping genes that have the most stable expression in the tested samples. To manage the large number of calculations, we have written a Visual Basic Application (VBA) for Microsoft Excel - termed geNorm - that automatically calculates the gene-stability measure *M* for all control genes in a given set of samples (geNorm is freely available from the authors on request). The program enables elimination of the worst-scoring housekeeping gene (that is, the one with the highest *M* value) and recalculation of new *M* values for the remaining genes. Using this VBA applet, we ranked the ten control genes in the five tissue panels tested according to their expression stability (Figure 2, Table 3). In addition, the systematic variation was calculated as the pairwise variation, *V*, for repeated RT-PCR experiments on the same gene, reflecting the inherent machine, enzymatic and pipet variation.

Normalization factor calculation based on the geometric mean of multiple control genes
We concluded that in order to measure expression levels accurately, normalization by multiple housekeeping genes instead of one is required. Consequently, a normalization factor based on the expression levels of the best-performing housekeeping genes must be calculated. For accurate averaging of the control genes, we propose to use the geometric

mean instead of the arithmetic mean, as the former controls better for possible outlying values and abundance differences between the different genes. The number of genes used for geometric averaging is a trade-off between practical considerations and accuracy. It is obvious that an accurate normalization factor should not include the rather unstable genes that were observed in some tissues. On the other hand, it remains relatively impractical to quantify, for example, eight control genes when only a few target genes need to be studied, or when only minimal amounts of RNA are available. Furthermore, it is a waste of resources to quantify more genes than necessary if all genes are relatively stably expressed and if the normalization factor does not significantly change whether or not more genes are included. Taking all this into consideration, we recommend the minimal use of the three most stable internal control genes for calculation of an RT-PCR normalization factor (NF$_n$, n = 3), and stepwise inclusion of more control genes until the $(n + 1)$th gene has no significant contribution to the newly calculated normalization factor (NF$_{n+1}$). To determine the possible need or utility of including more than three genes for normalization, the pairwise variation $V_{n/n+1}$ was calculated between the two sequential normalization factors (NF$_n$ and NF$_{n+1}$) for all samples within the same tissue panel (with a_{ij}=NF$_{n,i}$ and a_{ik}=NF$_{n+1,i}$, n the number of genes used for normalization ($3 \le n \le 9$), and i the sample index; see Equations 2 and 3 in Materials and methods). A large variation means that the added gene has a significant effect and should preferably be included for calculation of a reliable

Figure 2
Average expression stability values (*M*) of remaining control genes during stepwise exclusion of the least stable control gene in the different tissue panels (black circle, neuroblastoma; white circle, normal pool; white square, bone marrow; black square, leukocyte; gray circle, fibroblast; gray square, systematic error). See also Table 3 for the ranking of the genes according to their expression stability.

normalization factor. For all tissue types, normalization factors were calculated for the three most stable control genes (that is, those with the lowest *M* value) and for seven additional factors by stepwise inclusion of the most stable remaining control gene. Pairwise variations were subsequently calculated for every series of NF_n and NF_{n+1} normalization

Table 3

Control genes ranked in order of their expression stability*

Neuro-blastoma	Fibroblast	Leukocyte	Bone marrow	Normal pool
B2M	HMBS	ACTB	ACTB	B2M
RPL13A	B2M	HMBS	B2M	ACTB
ACTB	RPL13A	HPRT1	HMBS	YWHAZ
TBP	SDHA	SDHA	TBP	RPL13A
YWHAZ	TBP	TBP	SDHA	UBC
HMBS	ACTB	RPL13A	GAPD	TBP
UBC	UBC	GAPD	HPRT1	HPRT1
SDHA	YWHAZ	B2M	YWHAZ	HMBS
HPRT1 - GAPD	HPRT1 - GAPD	UBC - YWHAZ	UBC - RPL13A	SDHA - GAPD

*Increasing from top to bottom; the two most stable control genes in each cell type, for example *HPRT1* and *GAPD* in fibroblasts, cannot be ranked in order because of the required use of gene ratios for gene-stability measurements.

factors, reflecting the effect of adding an (*n*+1)th gene (Figure 3a). It is apparent that the inclusion of a fourth gene has no significant effect (that is, low $V_{3/4}$ value) for leukocytes, fibroblasts and bone marrow. This is also illustrated by the nearly perfect correlation between NF_3 and NF_4 values, as shown for fibroblasts in Figure 3b. On the basis of these data, we decided to take 0.15 as a cut-off value, below which the inclusion of an additional control gene is not required. For neuroblastoma and the pool of normal tissues, one and two additional genes, respectively, are necessary for reliable normalization (see also Figure 3b). The high $V_{8/9}$ and $V_{9/10}$ values for the normal pool, neuroblastoma and leukocytes corroborate very well the findings obtained by stepwise exclusion of the worst-scoring control gene (Figure 2). This analysis showed an initial steep decrease in average *M* value, pointing at two aberrantly expressed control genes for leukocytes and one unstable gene for neuroblastoma and the pool of normal tissues. Furthermore, the need to include additional control genes for these last two tissue panels is in keeping with the high variation in control-gene expression, as evidenced from Figure 2.

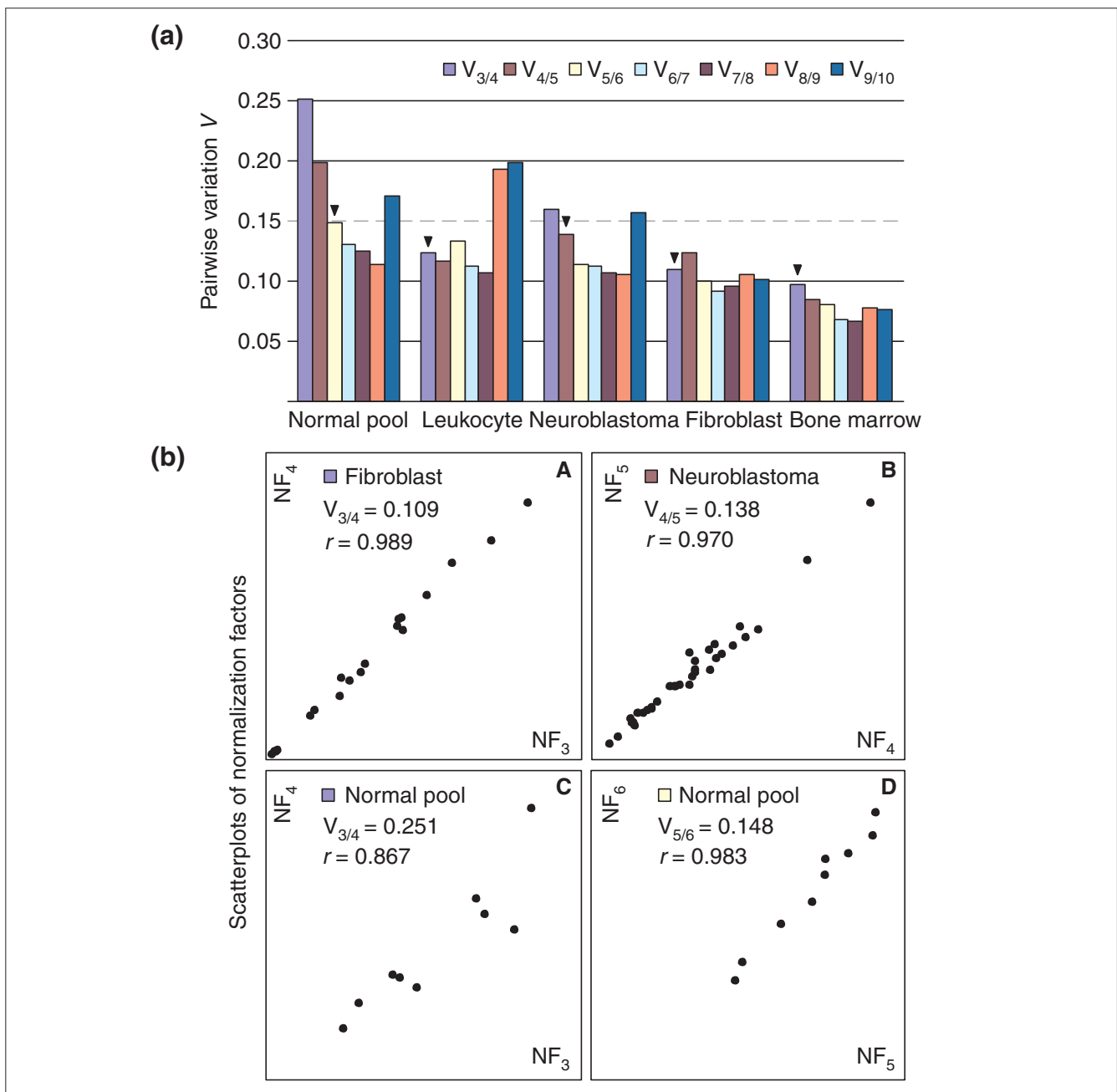## Validation of proposed real-time RT-PCR normalization factors
To assess the validity of the established gene-stability measure, that is, that genes with the lowest *M* values have indeed the most stable expression, we determined the gene-specific variation for each control gene as the variation coefficient of the expression levels after normalization. This
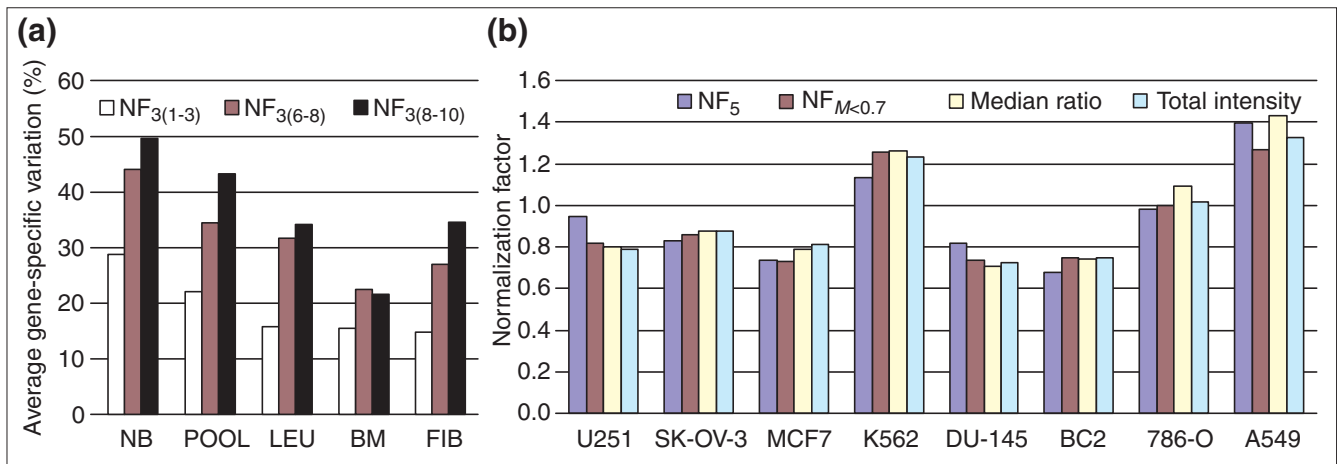
Figure 3
Determination of the optimal number of control genes for normalization. **(a)** Pairwise variation ($V_{n/n+1}$) analysis between the normalization factors NF_n and NF_{n+1} to determine the number of control genes required for accurate normalization (arrowhead = optimal number of control genes for normalization). **(b)** Selected scatterplots of normalization factors before (x-axis) and after (y-axis) inclusion of an (n + 1)th control gene (r = Spearman rank correlation coefficient). Low variation values, V, correspond to high correlation coefficients. It is clear that there is no need to include more than three, four or five control genes for fibroblast (A), neuroblastoma (B) and the normal pooled tissues (D), respectively. In contrast, panel C demonstrates that inclusion of at least a fourth control gene is required for the normal pooled tissues.

coefficient should be minimal for proper housekeeping genes. Three different normalization factors were calculated, based on the geometric mean of three genes with, respectively, the lowest ($NF_{3(1-3)}$), the highest ($NF_{3(8-10)}$), and intermediate M values ($NF_{3(6-8)}$) (as determined by geNorm). We subsequently determined the average gene-specific variation of the three genes with the most stable expression (that is, the lowest variation coefficient) for each normalization factor and within each tissue panel (Figure 4a). It is clear that the gene-specific variation in all tissue panels is by far
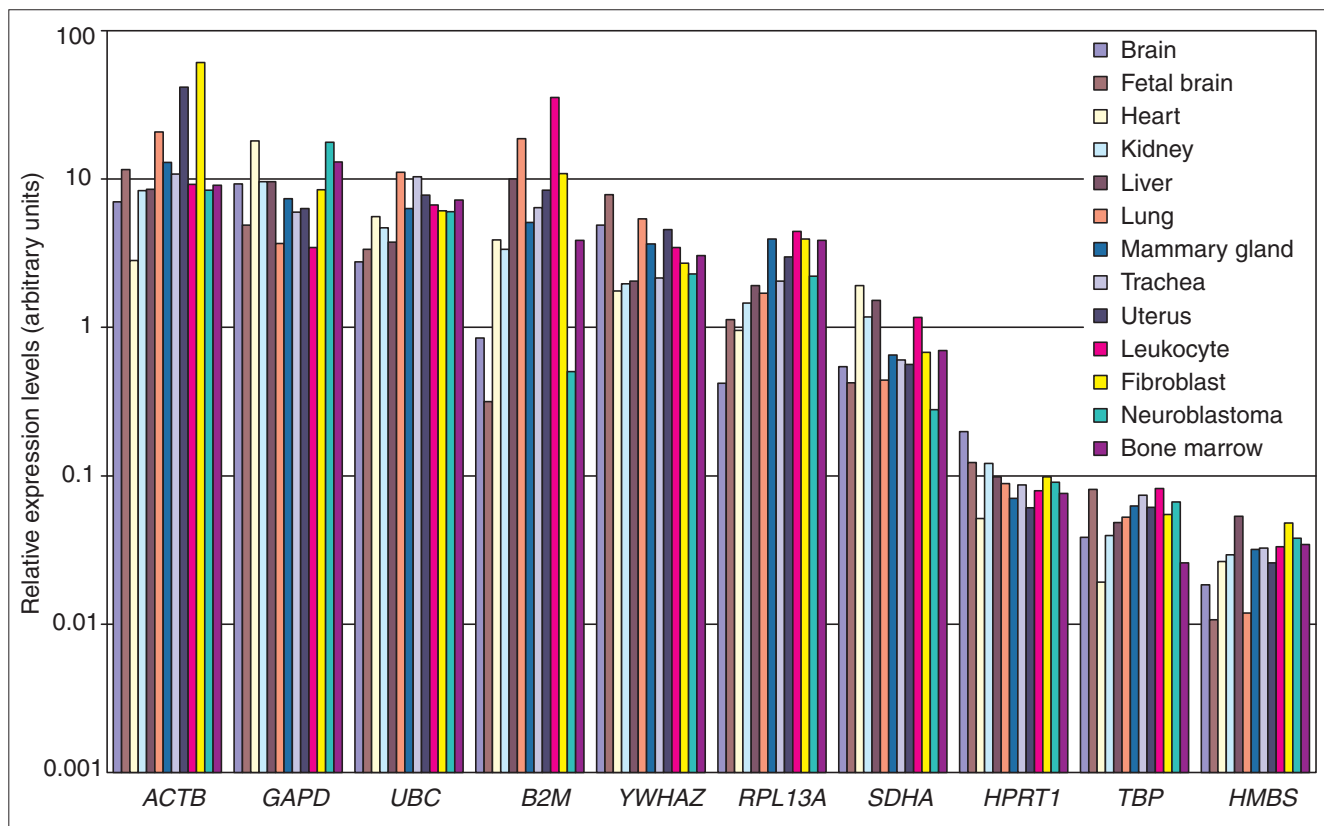
Figure 4
Validation of the gene-stability measure and the geometric averaging of carefully selected control genes for normalization. **(a)** Validation of the gene-stability measure. The average gene-specific variation (determined as coefficient of variation, in percent) for the three control genes with the smallest variation within each tissue panel after normalization with three different factors calculated as the geometric mean of the three control genes with the lowest (NF$_{3(1-3)}$), highest (NF$_{3(8-10)}$) and intermediate (NF$_{3(6-8)}$) gene stability values (as determined by geNorm). NB, neuroblastoma; POOL, normal pooled tissues; LEU, leukocytes; BM, bone marrow; FIB, fibroblasts. **(b)** Geometric averaging. Comparison of frequently applied microarray scaling factors and the proposed RT-PCR normalization factor based on the geometric mean of selected control genes (NF$_5$, geometric mean of the five control genes with the lowest M value; NF$_{M<0.7}$, geometric mean of control genes with $M < 0.7$; see Results), calculated for eight hybridizations from publicly available microarray data [14].

the smallest when the data are normalized to NF$_{3(1-3)}$. This demonstrates that the gene-stability measure effectively identified the control genes with the most stable expression. To verify that a high M value is characteristic of an unstable or differentially expressed gene, we analyzed the expression level of *MYCN* - a highly differentially expressed proto-oncogene in neuroblastoma with prognostic value [13] - together with the set of ten housekeeping genes. *MYCN* was readily identified as the most differentially expressed gene, with an M value of 6.02 compared to 2.17 for the least stable control gene (*B2M*) in neuroblastoma. It was further observed that normalization with a single control gene consistently resulted in significantly higher gene-specific variations of the other control genes (data not shown), which underscores the improvement in normalization by using multiple housekeeping genes.

To show that the associations between the best control genes are independent of cell proliferation, we analyzed the expression level of the proliferation marker *PCNA* in the neuroblastoma cancer cell lines, and determined the Spearman rank correlation coefficient between the raw expression levels of the four best housekeepers and the marker gene *PCNA*. From this analysis, it was clear that the control genes were - as expected - significantly correlated ($p < 0.001$, correlation coefficient between 0.60 and 0.76). In contrast, no correlation was observed between *PCNA* and three of the four control genes, and only a weak correlation ($p = 0.024$, coefficient = 0.43) between *PCNA* and control gene *HPRT1*. These data firmly demonstrate that the most stable control

genes (identified by the geNorm algorithm) are not *per se* linked to the state of cell proliferation of the samples.

To further validate the accuracy of geometric averaging of carefully selected control genes for normalization, the geometric means of housekeeping-gene expression levels obtained from publicly available microarray data were compared with commonly applied microarray normalization factors calculated for the same data. For this purpose, an 8,000-gene array data set [14] was chosen, containing nine of the ten control genes evaluated in this RT-PCR study. Two commonly applied microarray normalization factors (based on median ratio normalization, and total intensity normalization) [15-17] were determined for eight randomly selected hybridization sets. Subsequently, for each hybridization set, the background-corrected expression levels of nine housekeeping genes for the two fluorescence channels were imported into geNorm and ranked, as described for the RT-PCR data. As these microarray data originate from hybridizations of cell lines from various histological origin versus a reference pool of multiple cell lines, we have calculated the geometric mean of the five most stable control genes (NF$_5$) for each hybridization set, in accordance to the recommendations for reliable normalization within a heterogeneous tissue panel (see previous paragraph). Alternatively, internal control genes were excluded in a stepwise manner until the M values of the remaining genes were below 0.7 (experimental value shown to eliminate the most variable and outlying genes in this microarray dataset). Depending on the hybridization set, seven to nine genes fitted this criterion,

Figure 5
Logarithmic histogram of the expression levels of 10 internal control genes determined in 13 different human tissues, normalized to the geometric mean of 6 control genes (*GAPD*, *HPRT1*, *SDHA*, *TBP*, *UBC*, *YWHAZ*). An approximately 400-fold expression difference is apparent between the most and least abundantly expressed gene, as well as tissue-specific differences in expression levels for particular genes (for example, *B2M*).

upon which the geometric mean was calculated ($NF_{M<0.7}$). Both normalization factors (NF_5 and $NF_{M<0.7}$) were shown to be similar to the calculated microarray normalization factors (Figure 4b).

Tissue-specific housekeeping gene expression
To compare the control gene-expression levels within the heterogeneous group of all 13 tested tissues, the same set of control genes should be used for normalization. We therefore calculated the geometric mean of six control genes that were withheld from the set of ten genes after elimination of the two genes with the highest *M* value within each tissue panel (that is, *B2M*, *RPL13A*, *ACTB* and *HMBS*) (see Table 3). Given the large variety of tested tissues, this is the optimal strategy to eliminate most variation, and to allow direct comparison between the different samples. Under the assumption of equal PCR threshold cycle values for equal transcript numbers of different genes, an estimation of the transcript abundance of the various control genes can be made. Figure 5 shows that the ten tested genes belong to various abundance classes, with an approximately 400-fold expression difference between the most abundant (*ACTB*) and the rarest (*HMBS*) transcript. Although the overall

abundance of a given control gene in the different tissues is relatively similar, we clearly observe tissue-specific expression differences, for example, *B2M* expression level is 112-fold higher in leukocytes compared to fetal brain, and *ACTB* shows an expression difference of 22-fold between fibroblasts and heart tissue. It is also clear that some genes have a relatively constant expression level (for example, *UBC* and *HPRT1*) compared to the differential expression pattern of others (for example, *B2M* and *ACTB*).

Discussion
Accurate normalization of gene-expression levels is an absolute prerequisite for reliable results, especially when the biological significance of subtle gene-expression differences is studied. Still, little attention has been paid to the systematic study of normalization procedures and the impact on the conclusions. For RT-PCR, there is a general consensus on using a single control gene for normalization purposes. A comprehensive literature analysis of expression studies that were published in high-impact journals during 1999 indicated that *GAPD*, *ACTB*, 18S and 28S rRNA were used as single control genes for normalization in more than 90% of cases [11]. As

numerous studies reported that housekeeping gene expression can vary considerably [6,9-12], the validity of the conclusions is highly dependent on the applied control. Some laboratories have tried to find the optimal control gene for their experimental system, and often rRNA molecules were proposed as best references. These studies should be approached with some caution, as often only the variation in expression of the tested genes with respect to the mass loading of total RNA was assessed. As rRNA molecules make up the bulk of total RNA, they should indeed correlate very well with the total RNA mass, but that does not necessarily make them good control genes. As outlined in the introduction, total RNA and rRNA levels are not proper references, because of the observed imbalance between rRNA and mRNA fractions.

In addition to searching for a stable control gene, we aimed at determining the errors related to the common practice of single control normalization. In this study, we provide evidence that a conventional normalization strategy based on a single housekeeping gene leads to erroneous normalization up to 3.0- and 6.4-fold in 25% and 10% of the cases, respectively, with sporadic cases showing error values above 20. This analysis showed that a few control genes were unstable and significantly differentially expressed in some tissue panels, as evidenced by the decrease from 5.9 to 4.5 for the 90th-percentile single control normalization error value for neuroblastoma when the *B2M* gene is omitted (data not shown). This finding agrees with the reported differential expression of *B2M* in neuroblastoma, corresponding to the stage of differentiation of the tumor cells [18]. The error-distribution curves not only reflect the stability of expression of the applied controls, but also the sample heterogeneity within a tissue panel, as noted from the less steep curve for the heterogeneous set of normal pooled tissues compared to the other, relatively homogeneous, tissue panels. In this regard, the issue has been raised that finding proper control genes is even more important when working with tissues of different histological origin [9].

The single control normalization error values point to inherent noisy oscillations in expression levels of the control genes, a finding which has been corroborated in other large-scale studies where several thousand genes were measured in different cells or tissues by microarray analysis. No gene was found on an 8,000-feature array that did not vary by ratios of at least twofold across a panel of 60 cell lines [14], and a set of genes frequently used for normalization (including *GAPD* and *ACTB*) was found to vary in expression by 7- to 23-fold [9]. Taken together, our data and these studies clearly show that ideal and universal control genes do not exist. This warrants the search for stably expressed genes in each experimental system, and for the development of an accurate normalization strategy.

To validate the expression stability of the tested control genes without any prior assumption of a metric for standardization, we had initially measured the correlation between the raw, non-normalized expression levels of any two control genes, which should be nearly perfect for proper control genes. We observed, however, that the data range between the minimum and maximum expression levels, or any outlying value, could have a profound influence on the slope of the regression line, and consequently on the value of the correlation coefficient. This made Pearson and Spearman correlation coefficients unsuitable for this kind of analysis. We have therefore developed a new stability measure, based on the principle that the expression ratio of two proper control genes should be identical in all samples, regardless of the experimental condition or cell type, with increasing ratio variation corresponding to decreasing expression stability of one (or both) of the tested genes. The proposed standard deviation of log-transformed control gene ratios is a robust measure for the variation between two control genes, as it does not impose any requirements for normality or homoscedasticity of the data points. Furthermore, this measure is independent of the abundance difference between the genes, and is equally affected by any outlying or extreme ratio (that is, outliers for a sample with low or high overall expression, or outliers caused by an upregulated or downregulated gene have an equivalent increase in pairwise variation V). Logarithmic transformation of the ratios is required for symmetrical distribution of the data around zero, resulting in equal absolute values (but opposite signs) for a given ratio and the inverse ratio. As a result, the standard deviation of log-transformed ratios is identical to the standard deviation of log-transformed inverse ratios, which makes this measure characteristic for every combination of two genes.

Having established a robust measure to assess the variation in expression of two control genes, we subsequently defined a gene-stability measure M as the average pairwise variation between a particular gene and all other control genes. Using a VBA applet geNorm developed in-house, we ranked ten commonly used housekeeping genes belonging to different functional and abundance classes according to their expression stability in five tested tissue panels. The clear decrease of M of the remaining control genes during stepwise exclusion of the worst-scoring gene points at differences in the stability of gene-specific expression and demonstrates that the remaining genes are more stably expressed than the excluded genes. Some tissue panels show a relatively steep initial decline, which reflects the exclusion of one or more aberrantly expressed control genes (for example, *ACTB* and *HMBS* for leukocytes), as also noticed from the single control normalization error analyses (see above). The average gene stability values of the remaining genes during stepwise elimination of the least stable control genes also indicates tissue-specific differences, with bone marrow and the pool of normal tissues having the lowest and highest overall expression variation, respectively. The latter is no surprise, given the larger tissue heterogeneity in this panel.

The question of whether the observed high variation for neuroblastoma is a cancer-related phenomenon of deregulated expression is currently under further investigation. From these analyses, it is clear that there is no universal control gene suitable for all cell types. *ACTB* and *B2M* appear to be the worst-scoring genes, whereas *UBC*, *GAPD* and *HPRT1* seem to be the best overall control genes, each belonging to the four most stable genes in four out of five tested tissues. However, these generalizations should be treated with caution. *B2M* appears to be one of the least stable control genes, but is nevertheless a good choice for normalization of leukocyte expression levels. This clearly demonstrates that a proper choice of housekeeping genes is highly dependent on the tissues or cells under investigation. This is even more important when considering the differences in transcript abundance of some control genes between different tissues. The large expression differences between the tissues tested for *B2M* and *ACTB*, for instance, would definitely result in large normalization errors if they were used for standardization. Interestingly, the observed tissue-specific expression of these control genes is in keeping with their known role or function: there is high *B2M* expression in leukocytes, where it is a major cell-surface marker, and relatively low non-muscle cytoskeletal *ACTB* expression in heart tissue, which is predominantly of muscular origin.

In view of the inherent variation in expression of housekeeping genes, we recommend the use of at least three proper control genes for calculating a normalization factor, and present a procedure to determine whether or not more - and if so, how many - control genes were required for reliable normalization. This analysis clearly showed that three stable control genes sufficed for accurate normalization of samples with relatively low expression variation, whereas other tissue panels required a fourth, or even a fifth control gene to capture the observed variation.

The purpose of normalization is to remove the sampling differences (such as RNA quantity and quality) in order to identify real gene-specific variation. For proper internal control genes, this variation should be minimal or none. To validate the gene-stability measure *M* and the geNorm algorithm to identify the most stable control genes in a set of samples, we have calculated the gene-specific variation for each gene as the coefficient of variation of normalized expression levels. To this end, the raw expression values were standardized to different normalization factors, calculated as the geomean of the most, intermediate, or least stable control genes (as determined by geNorm). The rationale of this analysis is that a normalization factor based on proper internal control genes should remove all nonspecific variation. In contrast, unstable control genes cannot completely remove the nonspecific variation, and even add more variation, resulting in larger so-called gene-specific variations for the tested control genes. This analysis clearly demonstrated that most nonspecific variation was removed when the most stable control genes (as determined by geNorm) were used for normalization, which proves that the novel stability measure and strategy presented here effectively allowed the stability of gene expression in the different tissue panels to be assessed.

Further validation demonstrated that the geometric mean of carefully selected control genes is an accurate estimate of the mRNA transcript fraction, as determined by comparison with frequently applied microarray normalization factors. Although both RT-PCR normalization factors based on geometric averaging are relatively similar, the one based on at least seven control genes (that is, $NF_{M<0.7}$) is slightly more equivalent to the microarray-scaling factors. Two possible explanations can account for this observation. First, the five most stable control genes as determined by geNorm are based on only two RNA samples (that is, a Cy3-labeled reference pool, and a Cy5-labeled test sample), in contrast to the RT-PCR data, where 9 to 34 samples were used, resulting in more reliable estimation of the expression stability. Second, recent technical reports clearly state that array hybridization analyses experience considerable - often underestimated - variation and uncertainty at several levels. Accurate background fluorescence correction and spot quality assessment, among others, have been described as critical issues for reliable ratio estimation [19-21]. The higher variability associated with array hybridization results might thus explain the need for more control genes to normalize the data. Nevertheless, this study clearly showed that normalization based on the geometric mean of carefully selected control genes results in equivalent ratio estimation compared to commonly applied array scale factors, which validates its use for RT-PCR normalization. In addition, the method presented could easily be applied to normalize gene-expression levels resulting from microarray hybridization experiments, where only a limited number of genes are spotted, including some housekeeping genes.

In conclusion, we described and validated a procedure to identify the most stable control genes in a given set of tissue samples, and to determine the optimal number of genes required for reliable normalization of RT-PCR data. The strategy presented can be applied to any number or kind of genes or tissues, and should allow more accurate gene-expression profiling. This is of utmost importance for studying the biological significance of subtle expression differences, and for confirmatory and/or extended analyses of microarray results by means of RT-PCR.

Materials and methods
Sample preparation
Thirty-four neuroblastoma cell lines were grown to subconfluency according to standard culture conditions. RNA was isolated using the RNeasy Midi Kit (Qiagen) according to the manufacturer's instructions. Nine RNA samples from pooled

normal human tissues (heart, brain, fetal brain, lung, trachea, kidney, mammary gland, small intestine and uterus) were obtained from Clontech. Blood and fibroblast biopsies were obtained from different normal healthy individuals. Thirteen leukocyte samples were isolated from 5 ml fresh blood using Qiagen's erythrocyte lysis buffer. Fibroblast cells from 20 upper-arm skin biopsies were cultured for a short time (3-4 passages) and harvested at subconfluency as described [22]. Bone marrow samples were obtained from nine patients with no hematological malignancy. Total RNA of leukocyte, fibroblast and bone marrow samples was extracted using Trizol (Invitrogen), according to the manufacturer's instructions.

Real-time RT-PCR

DNase treatment, cDNA synthesis, primer design and SYBR Green I RT-PCR were carried out as described [23]. In brief, 2 μg of each total RNA sample was treated with the RQ1 RNase-free DNase according to the manufacturer's instructions (Promega). Treated RNA samples were desalted (to prevent carry over of magnesium) before cDNA synthesis using Microcon-100 spin columns (Millipore). First-strand cDNA was synthesized using random hexamers and SuperscriptII reverse transcriptase according to the manufacturer's instructions (Invitrogen), and subsequently diluted with nuclease-free water (Sigma) to 12.5 ng/μl cDNA. RT-PCR amplification mixtures (25 μl) contained 25 ng template cDNA, 2x SYBR Green I Master Mix buffer (12.5 μl) (Applied Biosystems) and 300 nM forward and reverse primer. Reactions were run on an ABI PRISM 5700 Sequence Detector (Applied Biosystems). The cycling conditions comprised 10 min polymerase activation at 95°C and 40 cycles at 95°C for 15 sec and 60°C for 60 sec. Each assay included (in duplicate): a standard curve of four serial dilution points of SK-N-SH or IMR-32 cDNA (ranging from 50 ng to 50 pg), a no-template control, and 25 ng of each test cDNA. All PCR efficiencies were above 95%. Sequence Detection Software (version 1.3) (Applied Biosystems) results were exported as tab-delimited text files and imported into Microsoft Excel for further analysis. The median coefficient of variation (based on calculated quantities) of duplicated samples was 6%.

Single control normalization error E

For any given m tissue samples, real-time RT-PCR gene-expression levels a_{ij} of n internal control genes are measured. For every combination of two tissue samples p and q, and every combination of two internal control genes j and k, the single control normalization error E was calculated (Equation 1). This is the fold expression difference between samples p and q when normalized to housekeeping gene j or k.

$(\forall\ j,k \in [1,n],\ \forall\ p,q \in [1,m],\ j \neq k$ and $p \neq q)$:

$$R_{jkpq} = \frac{a_{qj}}{a_{qk}} \cdot \frac{a_{pk}}{a_{pj}} \quad \text{(if } R < 1 \text{, then } E = R^{-1} \text{, else } E = R) \qquad (1)$$

Internal control gene-stability measure M

For every combination of two internal control genes j and k, an array A_{jk} of m elements is calculated which consist of $\log_2$-transformed expression ratios a_{ij}/a_{ik} (Equation 2). We define the pairwise variation V_{jk} for the control genes j and k as the standard deviation of the A_{jk} elements (Equation 3). The gene-stability measure M_j for control gene j is the arithmetic mean of all pairwise variations V_{jk} (Equation 4).

$(\forall\ j,k \in [1,n]$ and $j \neq k)$:

$$A_{jk} = \left\{ \log_2\left(\frac{a_{1j}}{a_{1k}}\right), \log_2\left(\frac{a_{2j}}{a_{2k}}\right), ..., \log_2\left(\frac{a_{mj}}{a_{mk}}\right) \right\} = \left\{ \log_2\left(\frac{a_{ij}}{a_{ik}}\right) \right\}_{i=1 \rightarrow m} \qquad (2)$$

$$V_{jk} = st.dev\ (A_{jk}) \qquad (3)$$

$$M_j = \frac{\sum_{k=1}^{n} V_{jk}}{n-1} \qquad (4)$$

Normalization of array data

Publicly available raw microarray data [14] were downloaded as tab-delimited files. Eight hybridization data sets were randomly selected and imported into Microsoft Excel software for further manipulation (MCF7, DU-145, 786-0, BC2, K562, A549, U251, and SK-OV-3). For each hybridization array, all spots with Cy3 or Cy5 fluorescence intensities below the average overall background level plus one standard deviation were discarded. Subsequently, a local background correction for each spot was applied. Two scale factors were calculated for each slide on the basis of median ratio normalization (median ratio set to 1) and total intensity normalization (equalized sum of fluorescence intensities for both channels). Nine housekeeping genes were identified by BLAST similarity or keyword search against the database of cDNA clones present on the array (see IMAGE clones listed in Table 1).

Additional data files

The raw expression values are available as a tab-delimited file with the online version of this paper.

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
2. Fink L, Seeger W, Ermert L, Hanze J, Stahl U, Grimminger F, Kummer W, Bohle RM: **Real-time quantitative RT-PCR after laser-assisted cell picking.** *Nat Med* 1998, **4:**1329-1333.
3. Heid CA, Stevens J, Livak KJ, Williams PM: **Real time quantitative PCR.** *Genome Res* 1996, **6:**986-994.
4. Higuchi R, Fockler C, Dollinger G, Watson R: **Kinetic PCR analysis: real-time monitoring of DNA amplification reactions.** *Biotechnology* 1993, **11:**1026-1030.
5. Solanas M, Moral R, Escrich E: **Unsuitability of using ribosomal RNA as loading control for Northern blot analyses related to the imbalance between messenger and ribosomal RNA content in rat mammary tumors.** *Anal Biochem* 2001, **288:**99-102.
6. Spanakis E: **Problems related to the interpretation of autoradiographic data on gene expression using common constitutive transcripts as controls.** *Nucleic Acids Res* 1993, **21:**3809-3819.
7. Johnson ML, Redmer DA, Reynolds LP: **Quantification of lane-to-lane loading of poly(A) RNA using a biotinylated oligo(dT) probe and chemiluminescent detection.** *Biotechniques* 1995, **19:**712-715.
8. Warner JR: **The economics of ribosome biosynthesis in yeast.** *Trends Biochem Sci* 1999, **24:**437-440.
9. Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M: **Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes.** *Physiol Genomics* 2000, **2:**143-147.
10. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E: **Housekeeping genes as internal standards: use and limits.** *J Biotechnol* 1999, **75:**291-295.
11. Suzuki T, Higgins PJ, Crawford DR: **Control selection for RNA quantitation.** *Biotechniques* 2000, **29:**332-337.
12. Bustin SA: **Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays.** *J Mol Endocrinol* 2000, **25:**169-193.
13. Maris JM, Matthay KK: **Molecular biology of neuroblastoma.** *J Clin Oncol* 1999, **17:**2264-2279.
14. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
15. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2:**418-427.
16. Hess KR, Zhang W, Baggerly KA, Stivers DN, Coombes KR: **Microarrays: handling the deluge of data and extracting reliable information.** *Trends Biotechnol* 2001, **19:**463-468.
17. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays.** *Nat Genet* 1999, **21:**10-14.
18. Cooper MJ, Hutchins GM, Mennie RJ, Israel MA: **Beta 2-microglobulin expression in human embryonal neuroblastoma reflects its developmental regulation.** *Cancer Res* 1990, **50:**3694-3700.
19. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: **Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.** *Nucleic Acids Res* 2001, **29:**2549-2557.
20. Yang MC, Ruan QG, Yang JJ, Eckenrode S, Wu S, McIndoe RA, She JX: **A statistical method for flagging weak spots improves normalization and ratio estimates in microarray.** *Physiol Genomics* 2001, **7:**45-53.
21. Brown CS, Goodwin PC, Sorger PK: **Image metrics in the statistical analysis of DNA microarray data.** *Proc Natl Acad Sci USA* 2001, **98:**8944-8949.
22. Nuytinck L, Narcisi P, Nicholls A, Renard JP, Pope FM, De Paepe A: **Detection and characterisation of an overmodified type III collagen by analysis of non-cutaneous connective tissues in a patient with Ehlers-Danlos syndrome IV.** *J Med Genet* 1992, **29:**375-380.
23. Vandesompele J, De Paepe A, Speleman F: **Elimination of primer-dimer artifacts and genomic coamplification using a two-step SYBR Green I real-time RT-PCR.** *Anal Biochem* 2002, **303:**95-98.
24. Bieche I, Laurendeau I, Tozlu S, Olivi M, Vidaud D, Lidereau R, Vidaud M: **Quantitation of MYC gene expression in sporadic breast tumors with a real-time reverse transcription-PCR assay.** *Cancer Res* 1999, **59:**2759-2765.