Systematic identification of human housekeeping genes possibly useful as references in gene expression studies

MARIA CARACAUSI, ALLISON PIOVESAN, FRANCESCA ANTONAROS, PIERLUIGI STRIPPOLI, LORENZA VITALE and MARIA CHIARA PELLERI

Department of Experimental, Diagnostic and Specialty Medicine (DIMES), Unit of Histology, Embryology and Applied Biology, University of Bologna, I-40126 Bologna, Italy

Received November 15, 2016; Accepted March 16, 2017

DOI: 10.3892/mmr.2017.6944

Abstract. The ideal reference, or control, gene for the study of gene expression in a given organism should be expressed at a medium-high level for easy detection, should be expressed at a constant/stable level throughout different cell types and within the same cell type undergoing different treatments, and should maintain these features through as many different tissues of the organism. From a biological point of view, these theoretical requirements of an ideal reference gene appear to be best suited to housekeeping (HK) genes. Recent advancements in the quality and completeness of human expression microarray data and in their statistical analysis may provide new clues toward the quantitative standardization of human gene expression studies in biology and medicine, both cross- and within-tissue. The systematic approach used by the present study is based on the Transcriptome Mapper tool and exploits the automated reassignment of probes to corresponding genes, intra- and inter-sample normalization, elaboration and representation of gene expression values in linear form within an indexed and searchable database with a graphical interface recording quantitative levels of expression, expression variability and cross-tissue width of expression for more than 31,000 transcripts. The present study conducted a meta-analysis of a pool of 646 expression profile data sets from 54 different human tissues and identified actin γ 1 as the HK gene that best fits the combination of all the traditional criteria to be used as a reference gene for general use; two ribosomal protein genes, *RPS18* and *RPS27*, and one aquaporin gene, POM121 transmembrane nucleporin C, were also identified. The present study provided a list of tissue- and organ-specific genes that may be most suited for the following individual tissues/organs: Adipose tissue, bone marrow, brain, heart, kidney, liver, lung, ovary, skeletal muscle and testis; and also provides in these cases a representative, quantitative portrait of the relative, typical gene-expression profile in the form of searchable database tables.

Correspondence to: Dr Lorenza Vitale, Department of Experimental, Diagnostic and Specialty Medicine (DIMES), Unit of Histology, Embryology and Applied Biology, University of Bologna, Via Belmeloro 8, I-40126 Bologna, Italy
E-mail: lorenza.vitale@unibo.it

Key words: gene expression, housekeeping gene, reference gene, human tissues, transcriptome mapping

Introduction

The quantitative study of gene expression in terms of the amount of RNA produced by a certain gene in a given biological condition is fundamental to our understanding of gene structure and function. Molecular laboratory techniques used to quantitatively measure RNA expression levels include northern blot analysis, reverse transcription-polymerase chain reaction (RT-PCR), expression microarrays and, recently, RNA sequencing (RNA-Seq). These techniques typically require a form of normalization of the measured RNA expression level of a gene to account for the potentially different RNA input quantities used in the assay. The best way to do this is to relate the transcripts to the number of templates (DNA strands) creating them. Owing to difficulties in obtaining these parameters from the same samples, several methods have been proposed over the years and are commonly used to relate the RNA amount of a given molecular species to one or more reference RNAs that are assumed to be expressed at a constant level in the cell type under consideration (1). More specifically, the ideal reference (or control) gene for the study of gene expression in a given organism should: i) Be expressed at a medium-high level so it can be easily detected; ii) be expressed at a constant/stable level in different cell types and within the same cell type undergoing different treatments; and iii) maintain these features through as many different tissues as possible within the organism (that is, ubiquitously expressed). These features would maximize the usefulness of the genes in the expression studies (2).

From a biological point of view, the theoretical requirements for an ideal reference gene appear to be best suited to the housekeeping (HK) genes, a large class of genes that are constitutively expressed, subjected to low levels of regulation in different conditions and perform biological actions that are fundamental for the basic functions of the cell (1). Their fundamental roles also mean that they tend to be expressed in high levels, confirming their suitability as reference genes.

Since the 1980s, several human genes have been widely used as 'classic' reference genes based on their fulfilling of the aforementioned requirements, as assessed typically by northern blot analysis (3), and this set of genes was seamlessly transferred for use in RT-PCR analyses in the 1990s (4). However, in a situation in which there was only preliminary knowledge of the human genome, the choice of these genes could be only anecdotal, among the limited pool of the genes known at the time. Following the widespread use of expression microarray techniques in the early 2000s (5), along with the initial sequencing and characterization of the human genome, it became theoretically possible to study and select HK genes by the systematic analysis of transcriptomes. This possibility was readily exploited in certain initial studies (6,7), which demonstrated that common control genes used in human studies, including the most popular glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), actin β (*ACTB*) and β_2-microglobulin (*B2M*) (3,8,9), actually exhibited considerable variability in expression within and across microarray data sets, and in certain cases this was confirmed by quantitative RT-PCR (RT-qPCR) analysis (10). The main conclusion was that the choice of a reference gene should be suited to the specific investigated tissue. In the following years, the problem of selecting a human HK gene by exploiting the availability of transcriptome-scale data was addressed by several studies and remains under debate (11).

The present study considered whether recent advancements in the quality and completeness of human expression microarray data, along with developments in statistical analysis, may be able to provide new clues toward the quantitative standardization of human tissue gene expression studies, cross- and within-tissue. A general framework is presented for choosing reference genes that may be useful in gene expression studies on normal human tissues and organs; the present study also addresses certain previous assumptions and provides an approach that is based on the Transcriptome Mapper (TRAM; http://apollo11.isto.unibo.it/software/TRAM) tool, which can overcome a number of problems associated with cross-platform analysis (such as, probe assignment to locus, intra- and inter-sample normalization and scaled quantile statistics) (12,13). TRAM can integrate data from hundreds or thousands of complete microarray data sets and provide unique combinations of features that are particularly suited to allow the choice of reference genes based on the three properties aforementioned. TRAM calculates a quantitative measurement for a consensus mean-expression value for tens of thousands of human transcripts expressed in a specific tissue or organ, thus allowing for a precise estimation of the intensity of its expression in terms of a percentage of the mean expression value in the pool of analyzed transcriptomes and the choice of a reference gene expressed at medium-high or high level (12,14). In addition, TRAM provides the standard deviation (SD) from the mean (normalized as the percentage of the mean value) for the mean expression value of a given locus, thus allowing for the selection of genes that may have more stable expression values in a variety of different samples and/or experimental platforms that have been investigated for a given tissue/organ. Finally, TRAM is able to integrate data from numerous sources, allowing verification of the consistency of the first two features through a wide range of different tissues within the organism studied (12).

The present study conducted a meta-analysis of 646 data sets that were obtained from different studies associated with 54 different normal human tissues and organs, using various experimental platforms. This meta-analysis produced results for 35,131 individual loci, including known genes and expressed sequence tag (EST) clusters, in the form of a database that may be extensively queried by freely combining a number of criteria, thus identifying the best intersection of moderate-high level of expression, low expression-value variability and expression in a large number of tissues. Results from the present study demonstrated that the human actin γ 1 (*ACTG1*) gene may potentially be used as a general reference gene for human cross-tissue studies and that specific genes are most suited for individual within-tissue studies. An enrichment analysis in functional classes for the identified HK genes is also presented.

Materials and Methods

Database search. To retrieve data sets that have been derived from normal adult human tissues, a systematic search of the gene expression data repository Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo) (15) was performed for any available sample series associated with pools of normal human tissues/organs, choosing *Homo sapiens* as the organism. The query used was: 'Homo sapiens [ORGANISM] AND tissue* [TI] OR organ* [TI]'. These selection criteria led to the generation of a pool of samples that included all of the main human organs and tissues, which served as a reference set already partially used by the authors of the present study (14). The searches were performed up to May 2013, and search results were then filtered using inclusion and exclusion criteria as explained in the *Data set selection* subsection below.

Data set selection. Inclusion criteria of data sets used in the present analysis were: i) Experiments were carried out on whole organs or tissues; ii) organs/tissues were obtained from individuals exhibiting normal phenotypes; iii) samples used in the studies were obtained from adults; and iv) the availability of the raw or pre-processed data. Exclusion criteria included: i) Studies using exon arrays (which hamper data elaboration by TRAM, owing to an exceedingly high number of data rows) or platforms using probes that are split into several different arrays for each sample, which hampers intra-sample normalization; ii) platforms that examine an atypical number of genes (that is, <5,000 or >60,000); and iii) data that is derived from cell lines, pathological or treated tissue, or children or fetal tissues.

A quantitative transcriptome map was obtained by linearizing values from each data set that were provided as logarithms. If only raw files (such as CEL files) were available in the GEO database, they were converted into pre-processed data using the AltAnalyze version 2.0 program (http://www.altanalyze.org) (16). Tables I and II summarize the data sets used to build the integrated transcriptome map from different human tissues/organs. The mean number of retrieved data sets was 22.2±6.6 for the group of the main tissues/organs (Table I) and 6.5±5.2 for the group of the minor tissues/organs

(Table II) since there were a lower number of representatives of certain of them owing to the limited availability of data in the repositories.

TRAM analysis. The TRAM tool (12) allows gene expression data to be imported in a tab-delimited text format. It also allows data integration by decoding probe-set identifiers to gene symbols using UniGene data parsing (17), normalizing data from multiple platforms using intra-sample and inter-sample normalization (that is, scaled quantile normalization) (13), and creates tables of expression values for each transcript.

A directory (folder) was created that contains the whole pool of all the sample data sets (n=646) associated with 54 different human tissues/organs that were retrieved and downloaded from the GEO database. This set included 629 samples described previously (14), which were in addition to the following GEO samples: GSM2829, GSM2856, GSM18792, GSM18793, GSM18951, GSM18952, GSM39975, GSM39980, GSM39992, GSM39994, GSM39999, GSM40001, GSM44671, GSM52565, GSM175935, GSM790822 and GSM790831. These samples were pre-processed according to the TRAM tool guide to be ready for import and processing in TRAM. During the pre-processing step, TRAM allows the linearization of data sets when provided as logarithms. All pre-processed samples of the whole pool were then imported as pool B in the TRAM table, and the entire set of analyses permitted by TRAM was performed as described in detail in the tool guide, using default parameters as previously described (12). The significance of the over-/underexpression of single genes was determined within pool B by running TRAM in 'Map' mode (12) with a segment window of 12,500 bp and a minimum number of one over-/underexpressed gene in that window. This window size is <20% of the 67 kb mean size of a human protein-coding gene, as determined by searching the GeneBase database (18); therefore, significant over-/underexpression of a segment [at $q<0.05$, where q is the P-value corrected for false discovery rate (12)] almost always corresponds to that of a single gene. When the segment window contains >1 gene, significance is maintained if the expression value of the over-/underexpressed gene prevails over the others.

To study tissues or organs individually, 10 representative biological conditions were selected (Table I) and the data were exported for the pool of samples relative to each tissue/organ from the TRAM table 'Values B' and reimported in TRAM as pool A to allow comparison between the specific tissue/organ (pool A) and the whole pool (pool B).

TRAM version 1.2.1, 2015 human version, was used in the present study and is freely available at http://apollo11.isto.unibo.it/software. The complete set of TRAM results from the import and analysis of the 646 data sets is not currently distributed owing to its very large size (25 GB); only the final results are available. However, the complete set of results may be regenerated locally by running the automated import and analysis of the data sets in the aforementioned pool B folder into the TRAM 1.2.1 software. Briefly, gene expression values were assigned to individual loci using UniGene, data were subjected to intra-sample normalization as a percentage of the mean value and then to inter-sample normalization by scaled quantile. The value for each locus within each biological condition is the mean of all available values for that locus. The median value of whole genome gene expression was used to determine the percentiles of expression for each gene. Only mapped genes or EST clusters with an assigned gene name or UniGene code, respectively, that have start and end genomic coordinates, and have a non-empty raw intensity value were selected for the analysis.

Further improvements were made for the present study by adding the total number of biological samples (microarrays) from which each gene expression value is derived, in addition to the number of data points, as certain experimental platforms used to assess expression levels for a sample may contain a variable number of microarray spots, each with a different probe, which generate multiple expression data points for the same gene. These improvements are available (upon request) as a dedicated script and are to be fully integrated into the next version of TRAM (TRAM 1.3), which is due to be released in 2017. Since a different number of microarray spots may be available on a platform for a given gene, these new features offer the possibility to normalize the quantity of information that is available for a gene based on the actual number of distinct biological samples that provide measured expression values for that gene.

To create transcriptome maps, TRAM does not consider probes for which expression values are not available, assuming that an expression level has not been measured. Furthermore, the software gives 95% of the minimum positive value present in a sample to those expression values ≤0 to obtain meaningful numbers when it is required to obtain a ratio between values in pool A and pool B. If it is assumed that, in these cases, the expression level is too low to be detected under the experimental conditions used, then this transformation may be useful to highlight differential gene expression.

HK gene search. The predicted genes that behave as HK genes were determined, as they are mainly involved in fundamental cellular functions and are ubiquitously and constitutively expressed in all tissues (19-21). A search for HK genes in the transcriptome maps was performed using the following parameters: i) An expression value ≥300, which in TRAM is given as a percentage of the mean expression value in a sample in order to select genes that are expressed at least threefold above the mean value and are therefore expressed at an easily appreciable level; ii) a SD ≤30, expressed as a percentage of the mean value to identify genes with a low expression variation among different samples; and iii) a sample number ≥80% of the total number of samples for each analyzed pool to select commonly expressed HK genes (for example, ≥80% indicates 16 out of 20 samples for adipose tissue). When a very low number of suitable HK genes were identified by these criteria, the parameters were relaxed to ≥150 mean expression value, ≤45% SD and ≥65% of samples in the pool with a measured value for the gene (Table I).

To select the HK genes with the best overall features to be proposed as reference genes, the genes identified as fitting the described criteria were first arranged in descending order of expression value, followed by ascending order of SD% and finally by descending order of sample number. An ascending rank number was assigned for each sorting criterion, the mean among these three ranks was calculated and the lowest mean rank was considered to correspond to the gene with the

Table I. Summary of the samples used in the meta-analysis for 10 specific tissues/organs.

Sample	ID	n	Data points	Genes	HK[a]	Value	Samples % (n)	SD%
Whole pool[b]	1-646	646	21,840,330	35,131	27	≥150	≥65 (420)	≤45
Adipose tissue	1-20	20	943,480	25,722	363	≥300	≥80 (16)	≤30
Bone marrow	46-60	15	548,322	26,097	222	≥300	≥80 (12)	≤30
Brain	61-84	24	507,799	21,603	174	≥150	≥65 (16)	≤45
Heart	150-166	17	452,683	34,486	56	≥150	≥65 (11)	≤45
Kidney	167-183	17	405,285	33,816	65	≥150	≥65 (11)	≤45
Liver	184-216	33	971,110	35,090	49	≥150	≥65 (21)	≤45
Lung	217-249	33	817,145	34,872	49	≥150	≥65 (21)	≤45
Ovary	281-304	24	788,447	34,853	157	≥150	≥65 (16)	≤45
Skeletal muscle	403-418	16	324,382	27,160	77	≥300	≥80 (13)	≤30
Testis	495-517	23	745,430	34,853	227	≥150	≥65 (15)	≤45

[a]Please refer to the text for a description of the selection criteria used to identify HK genes according to expression values, number of samples and SD. [b]Whole pool contains all sample data sets related to the 54 different human tissues/organs listed in Tables I and II. ID, identifier used in the present study; HK, number of housekeeping genes retrieved; value, expression value; Samples % (n), the percentage of samples in each dataset in which the HK genes fulfiled the selection criteria; n, total sample number; SD%, standard deviation from the mean expression value of a given locus expressed as a percentage.

overall best fit to the three criteria. For tissue-specific analysis, a fourth criterion was added by calculating the ascending rank of the absolute deviation from 1 of the ratio between the mean expression value of the gene in the considered tissue and in the whole pool of 646 samples, respectively (lowest rank considered the best). This fourth criterion selected for genes with the most similar expression values in the whole pool and each tissue-specific pool, suggesting a particularly stable expression level. The mean rank was then calculated for all four criteria.

Functional analysis. The hypothesis that the most suitable HK genes identified in the analysis could be enriched for particular functional classes was tested using the web tool FuncAssociate version 3.0 (http://llama.mshri.on.ca/funcassociate) (22).

Results

Human HK genes for general use. Using the search criteria detailed in the Methods section of the present study, several genes that were present in ≥65% of the whole pool list of 646 samples from 54 human tissues/organs were identified. A total of eight genes were identified that best fulfilled the criteria to be proposed as reference genes (Table III), and the HK gene that best fit the combination of all the traditional criteria to be used as a general reference gene was *ACTG1*.

Human HK genes for individual tissues. A total of 10 human tissues/organs were selected for a systematic search for the best suitable references genes within the respective biological type. By searching with the four criteria aforementioned, several genes were identified that had a mean expression value ≥300 (or ≥150, for searches performed with less stringent criteria), with an SD≤30 (or SD≤45) and with an expression value measured in ≥80% (or ≥65% with relaxed criteria) of the samples within different tissues/organs (Tables IV-VII). The eight known genes with the lowest mean rank scored for the selected criteria are listed in Table IV (adipose tissue and bone marrow), Table V (brain, heart and skeletal muscle), Table VI (kidney, liver and lung) and Table VII (ovary and testis). The complete gene name corresponding to each gene symbol listed in Tables III-VII is provided in Table VIII, along with the number of times that each gene is represented in a different pool among the 11 pools analyzed in Tables III-VII.

Analysis of the associated function. Using the FuncAssociate web tool, the identified HK genes were revealed to be significantly enriched in certain functional classes. Of the genes identified as suitable for human tissue-wide use listed in Table III, there were statistically significant enrichments [with adjusted P-value (P-adj)≤0.05] only in the Gene Ontology categories: Ribosome (P-adj=0.003), translation termination (P-adj=0.038), translation elongation (P-adj=0.045) and cellular protein complex disassembly (P-adj=0.05).

A significant enrichment in numerous Gene Ontology categories was also identified by pooling in a unique list all 63 genes (Table IX) identified for the 10 specific tissues/organs (Table IV-VII); all enrichments were associated with basic cellular components, molecular functions and/or biological processes. The biological process with the highest statistical significance (P-adj<0.001) was RNA catabolic processes and translation (data not shown).

Discussion

Following the diffusion of expression microarray technology, a number of attempts have been made to use microarray data to perform a systematic analysis of the features of gene expression to identify HK genes that may be best suited as reference genes. Early attempts suffered from the limited number of samples available for analysis in addition to a lack of choice of computational biology techniques to analyze them, in

Table II. Samples used in the meta-analysis for the whole pool in addition to all samples listed in Table I.

Sample	ID	n
Adrenal gland	21-30	10
Aorta	31-34	4
Appendix	35-38	4
Bladder	39-41	3
Blood	42-45	4
Breast	85-89	5
Bronchus	90-95	6
Ciliary ganglion	96-99	4
Colon	100-105	6
Connective	106	1
Dental pulp	107-108	2
Dorsal root ganglia	109-130	22
Esophagus	131-141	11
Fallopian tube	142-148	7
Gall bladder	149	1
Lymph node	250-266	17
Mammary gland	267-272	6
Oral mucosa	273-280	8
Pancreas	305-315	11
Parathyroid	316-318	3
Penis	319-324	6
Pericardium	325-326	2
Pharyngeal mucosa	327-334	8
Pituitary gland	335-351	17
Prostate	352-382	31
Salivary gland	383-402	20
Skin	419-431	13
Small intestine	432-437	6
Smooth muscle	438-441	4
Spinal cord	442-465	24
Spleen	466-483	18
Stomach	484-488	5
Synovial membrane	489-494	6
Thymus	518-532	15
Thyroid	533-554	22
Tongue	555-564	10
Tonsil	565-578	14
Trachea	579-592	14
Trigeminal ganglion	593-612	20
Urethra	613-620	8
Uterus	621-628	8
Vagina	629-637	9
Vena cava	638	1
Vulva	639-646	8

ID, identifier used in the present study; n, total sample number.

particular for cross-platform analysis (6). Similar investigations were also conducted using the EST database (23) and, in recent years, RNA-Seq data (24,25). However, the generation of ESTs is subject to certain biases depending on the level of ability for an mRNA to be cloned during the generation of the cDNA EST libraries; therefore, although these data are useful to identify expressed sequences, they are less useful for quantitative analysis. RNA microarrays and RNA-Seq are the two main types of high-throughput technologies used to assess gene expression (26). Although RNA-Seq is considered to be more sensitive and has a broader dynamic range than RNA microarrays (27), in large comparative studies these two methods have produced comparable results in terms of gene expression profiling (27-29). Microarrays remain an accurate tool for measuring the levels of gene expression (29) and continue to provide useful data-mining resources. The approach of the present study takes advantage of the large number of previous transcriptomic studies that were performed with microarray technology and stored in publicly available databases, and also of the results provided in the form of a list of genes and the corresponding expression values.

The diverse origin of the data, in terms of different investigated individuals, different experimenters and different experimental platforms in the field of microarray analysis, provided a richness in the context of an analysis such as in the present study. That is, following data integration, the final results were not affected by systematic biases that may be linked to the particular samples or experimenters/platforms involved in the generation of the data, and they are likely to best represent the actual 'mean' status for a gene (12), compared with works based only on the original data obtained through a single platform (6). In addition, the approach of the present study exploited the combination of: i) Automated reassignment of probes to the corresponding genes by the updated UniGene data embedded in TRAM; ii) intra- and inter-sample normalization, including the scaled-quantile method that allows for comparison among platforms with a highly different number of probes; and iii) elaboration and representation of gene expression values in linear form within an indexed, searchable database, with a graphical interface recording quantitative levels of expression (mean expression values), expression variability (SD) and cross-tissue expression of more than 31,000 transcripts. These features represent a clear advancement in comparison with other meta-analyses that were based on published microarray data and were also aimed at identifying human reference genes (30,31), particularly considering that several studies on the subject were conducted in years when there was a reduced availability of samples and/or the experimental platforms were less complete (32).

The meta-analysis in the present study was performed on a pool of 646 data sets from 54 different human whole tissues/organs, and excluded analyses of individual cell types as the whole organ/tissue includes a vast number of cell types in its structure [as discussed by Fagerberg *et al* (33)], and also due to the requirement of selecting a representative set of samples as a result of the very long elaboration time for each analysis. The *ACTG1* gene was identified as the HK gene that best fit the selection criteria for use as a general reference gene in the study of human gene expression. This gene proved to be statistically significantly over-expressed in the transcriptome map, according to the described criteria (34), with the following features: A very high mean expression value (3,453.7, indicating an expression level of ~35-fold

Table III. A list of the eight genes that best fulfill the criteria proposed for use as a reference in gene expression studies across all 646 pool B samples of human tissues and organs that were examined.

Mean rank	Gene[a]	Chromosome	Value	n	SD%
2.8	*ACTG1*[b]	17	3,453.7	513	37.8
5.5	*RPS18*	6	4,933.5	472	41.4
5.8	*POM121C*	7	349.7	425	28.0
6.5	*MRPL18*	6	226.6	546	39.4
6.5	*TOMM5*	9	273.6	426	38.2
7.0	*YTHDF1*	20	213.0	546	39.1
7.3	*TPT1*	13	5,941.7	508	43.1
8.0	*RPS27*	1	4,355.6	513	44.1

[a]Please see Table VIII for a complete list of gene definitions. [b]*ACTG1* was significantly overexpressed in the transcriptome map (q=0.03; where q is the P-value corrected for false discovery rate). Value, expression value; SD%, standard deviation expressed as percentage of the mean expression value for the locus.

Table IV. A list of the eight genes that best fulfill the selection criteria proposed for a reference in gene expression studies of human adipose tissue and bone marrow.

A, Adipose tissue									
Mean rank	Gene[a]	Chromosome	Value A	Value B	A/B	Sample count A	Sample count B	SD% A	SD% B
25.0	*RPL6*	12	2,207.0	2,069.0	1.1	20	592	11.6	89.5
27.5	*RPS25*	11	2,189.7	2,113.3	1.0	20	568	14.2	75.2
27.5	*SOD1*	21	1,273.0	1,231.8	1.0	20	592	12.1	67.9
33.5	*RNASEK*	17	912.6	881.1	1.0	20	335	11.3	41.7
35.8	*GABARAP*	17	1,112.7	1,185.6	0.9	20	580	12.7	72.6
41.3	*ACTG1*	17	3,821.3	3,453.7	1.1	20	513	15.2	37.8
43.8	*GABARAPL2*	16	591.6	591.9	1.0	20	592	12.7	69.8
45.0	*MRFAP1*	4	1,583.2	1,499.0	1.1	20	480	17.6	64.2
B, Bone marrow									
13.0	*RPL41*	12	5,739.4	5,838.5	1.0	13	568	18.5	45.1
17.8	*RPLP0*	12	3,659.6	3,341.5	1.1	13	508	16.5	44.5
17.8	*RPS27*	1	4,494.8	4,355.6	1.0	13	513	19.5	44.1
23.3	*TUBA1B*	12	2,893.1	2,478.6	1.2	13	508	16.4	81.1
24.3	*RPSA*	3	2,016.1	2,036.8	1.0	13	575	20.6	80.1
25.5	*SLC25A3*	12	918.5	956.7	1.0	13	592	17.6	73.5
26.5	*ACTG1*	17	3,728.3	3,453.7	1.1	13	513	20.3	37.8
30.0	*EEF1G*	11	2,713.6	2,686.9	1.0	13	587	22.6	48.3

[a]Please see Table VIII for a complete list of gene definitions. Value A, mean expression values of the gene in the pool A, including the sample related to the specific tissue; value B, mean expression values of the gene in the whole pool B consisting of all the 646 samples; A/B, ratio between values A and B; SD%, standard deviation expressed as percentage of the mean expression value (A or B) for the locus.

in comparison with the mean expression value of all the genes in each sample, set as equal to 100) and an SD% of 37.8%; the number of samples in which a measure for this gene was available was 513 out of 646 (79.4%). Notably, this well-characterized gene, whose coding sequence appears to be completely characterized (35), encodes for a cytoplasmic form of actin that is known to be ubiquitously expressed in human cells, but is different from the actin β (*ACTB*) that is routinely used as a reference gene. According to the data of the present study, the commonly used reference genes *ACTB* and *GAPDH* had excellent features in terms of high expression value and diffuse expression in human cells. However,

Table V. A list of the eight genes that best fulfill the selection criteria proposed for a reference in gene expression studies of human brain, heart and skeletal muscle.

Mean rank	Gene[a]	Chromosome	Value A	Value B	A/B	Sample count A	Sample count B	SD% A	SD% B
A, Brain									
9.3	*NDUFB4*	3	590.9	545.7	1.1	20	551	20.6	110.0
9.8	*NDUFB1*	14	546.7	586.0	0.9	22	592	22.7	59.9
17.5	*GSTO1*	10	345.9	370.6	0.9	22	592	22.0	75.5
28.0	*AMZ2*	17	299.5	276.1	1.1	20	467	29.3	99.5
28.8	*POLR2I*	19	302.8	254.7	1.2	22	568	23.0	79.2
29.5	*NDUFA3*	19	365.0	280.9	1.3	20	551	19.7	59.0
30.0	*RRAGA*	9	463.0	350.9	1.3	22	592	26.5	45.1
30.5	*POMP*	13	283.5	338.2	0.8	20	546	24.0	106.0
B, Heart									
10.5	*MIF*	22	896.7	845.6	1.1	11	527	42.1	66.5
12.5	*ECHS1*	10	567.8	571.0	1.0	15	592	43.2	100.3
13.0	*FAM96A*	15	338.6	289.7	1.2	11	480	35.0	63.8
13.5	*NOP10*	15	474.8	428.5	1.1	13	546	41.4	48.0
13.5	*TBCB*	19	319.8	290.6	1.1	15	592	38.8	67.9
14.8	*RRAGA*	9	292.4	350.9	0.8	15	592	37.0	45.1
15.0	*IFI27*	14	510.4	441.9	1.2	15	592	42.2	103.1
15.5	*MB*	22	6,845.8	745.2	9.2	15	592	30.6	260.2
C, Skeletal muscle									
4.3	*RPL41*	12	4,995.8	5,838.5	0.9	16	568	17.3	45.1
8.3	*PRDX1*	1	921.2	1,096.5	0.8	16	592	18.5	71.7
10.8	*RPL8*	8	2,065.6	2,069.0	1.0	16	513	25.2	44.1
11.3	*C14orf166*	14	576.0	505.7	1.1	16	546	19.1	57.0
11.8	*JTB*	1	714.0	606.9	1.2	16	592	20.6	61.6
11.8	*RPS29*	14	2,751.4	2,551.2	1.1	16	592	25.4	57.9
13.0	*SNRPD2*	19	516.4	524.7	1.0	16	592	21.9	57.5
14.5	*NOP10*	15	497.5	428.5	1.2	16	546	19.2	48.0

[a]Please see Table VIII for a complete list of gene definitions. Value A, mean expression values of the gene in the pool A, including the sample related to the specific tissue; value B, mean expression values of the gene in the whole pool B consisting of all the 646 samples; A/B, ratio between these value A and value B; SD%, standard deviation expressed as percentage of the mean expression value (A or B) for the locus.

they have an SD% almost double that of *ACTG1*. Owing to the high similarity between *ACTB* and *ACTG1* (91% identity with no gaps between their coding sequences, as determined by standard BLASTN analysis; data not shown), probes and primers need to be accurately selected to specifically identify the desired form of RNA.

Among the HK genes identified to be best suited as general reference genes for human studies, two ribosomal protein genes, *RPS18* and *RPS27*, and one aquaporin gene, POM121 transmembrane nucleporin C (*POM121C*), were identified. *ACTG1* and *RPS27* were also included in the top 20 HK human genes across 42 human tissues in a previous study based on a single platform (36), however, the data sets from the present study were not included in the present meta-analysis since it was not possible to derive the expression values as linear numbers for each microarray channel from the deposited data. Notably, the eight genes listed in Table III were also classified at the transcript and protein level as 'expressed in all tissues' in the Human Protein Atlas (33), further supporting the results of the present study regarding them as the most generally suitable reference genes.

In accordance with the relevance of the *ACTG1* gene in cross-tissue analysis, this gene is also present in the greatest number of lists (n=5) of the 10 tissue-specific genes best fulfilling criteria to be used as reference genes (Tables IV-VII). Ribosomal proteins were another notable example of known classes of general HK genes that are well represented in several human tissues. Although there is a clear,

Table VI. A list of the eight genes that best fulfill the selection criteria proposed for a reference in gene expression studies of human kidney, liver and lung.

Mean rank	Gene[a]	Chromosome	Value A	Value B	A/B	Sample count A	Sample count B	SD% A	SD% B
A, Kidney									
6.5	*PGAM1*	10	688.8	741.8	0.9	14	556	29.6	64.0
7.5	*NOP10*	15	413.9	428.5	1.0	12	546	31.9	48.0
8.0	*FIS1*	7	340.7	329.0	1.0	12	546	29.6	52.9
9.0	*GPX1*	3	459.9	496.7	0.9	15	592	34.2	66.9
10.3	*GANAB*	11	266.7	260.0	1.0	15	587	35.1	49.5
11.0	*NDUFB11*	X	343.8	294.9	1.2	12	551	25.0	67.7
12.8	*HEBP1*	12	288.6	249.2	1.2	12	551	27.7	53.9
13.0	*HDGF*	1	444.3	432.2	1.0	14	568	40.5	58.4
B, Liver									
8.5	*HEBP2*	6	335.1	362.5	0.9	31	592	35.8	79.2
11.8	*NDUFS3*	11	289.3	325.3	0.9	31	592	36.1	60.8
12.3	*POLR2H*	3	162.0	160.1	1.0	33	641	32.4	45.2
12.5	*MRPS24*	7	438.5	445.0	1.0	23	426	42.0	57.9
13.3	*FAM96B*	16	287.1	325.2	0.9	29	546	37.0	47.9
13.3	*GTF3A*	13	349.5	310.3	1.1	22	501	38.6	62.0
13.3	*HIST1H2BK*	6	216.3	186.6	1.2	28	568	30.3	72.7
13.8	*CRELD2*	22	297.3	279.9	1.1	24	467	39.5	52.7
C, Lung									
7.3	*RBX1*	22	283.7	315.4	0.9	29	551	36.6	57.8
7.5	*RRAGA*	9	289.7	350.9	0.8	31	592	34.9	45.1
11.5	*LAMTOR5*	1	253.1	344.6	0.7	33	646	34.1	55.7
11.8	*CNIH1*	14	207.4	250.4	0.8	31	592	32.3	63.4
12.0	*EPCAM*	2	252.8	256.7	1.0	31	592	39.0	218.7
12.8	*EIF4A3*	17	271.2	340.0	0.8	25	563	37.8	47.1
13.5	*ACTG1*	17	3,067.3	3,453.7	0.9	27	513	43.9	37.8
14.0	*FAM96B*	16	277.0	325.2	0.9	24	546	39.2	47.9

[a]Please see Table VIII for a complete list of gene definitions. Value A, mean expression values of the gene in the pool A, including the sample related to the specific tissue; value B, mean expression values of the gene in the whole pool B consisting of all the 646 samples; A/B, ratio between these value A and value B; SD%, standard deviation expressed as percentage of the mean expression value (A or B) for the locus.

expected prevalence among the identified loci of genes encoding for basic cell structure (such as genes encoding for cytoskeletal components) and function (such as genes encoding for transcription and translation, reduction-oxidation metabolism and signaling proteins), it is worth noting that specific members of the same gene family involved in these processes may be identified in one particular tissue/organ and not in the others. The vast majority of genes that may be more suitable as reference genes for individual tissues (Tables IV-VII) are still typical HK genes, with the clear exception of the tissue-specific myoglobin gene in the heart (Table V).

The approach used in the present study allows for the systematic search for ideal reference genes and, at the same time, made available a 'consensus' reference gene-expression profile for 10 human tissues/organs. From this particular point of view, the presented results are less systematic than other previous attempts conducted in the case of the brain (34) and heart (14). Only the samples belonging to experiments in which a series of normal human tissues were analyzed have been included here, without searching for any single samples recorded for a given tissue in any type of available experiment (for example, comparisons between normal and pathological samples). However, the present data have been obtained through an improvement of the search algorithm for HK genes and may still offer interesting hints to their biological specificity in the transcriptome of these tissues. The same approach may be applied to data sets deriving from cell lines or pathological samples.

Table VII. A list of the eight genes that best fulfill the selection criteria proposed for a reference in gene expression studies of human ovary and testis.

Mean rank	Gene[a]	Chromosome	Value A	Value B	A/B	Sample count A	Sample count B	SD% A	SD% B
A, Ovary									
10.0	*ACTG1*	17	3,547.3	3,453.7	1.0	17	513	30.8	37.8
20.0	*AP2M1*	3	387.9	375.5	1.0	17	513	32.0	51.8
22.5	*MIF*	22	885.0	845.6	1.0	17	527	36.5	66.5
24.3	*NELFCD*	20	240.2	239.5	1.0	17	522	31.3	118.6
24.5	*RNF181*	2	331.3	356.5	0.9	16	426	24.8	52.5
28.0	*PSMC1*	14	367.9	421.4	0.9	17	513	29.3	48.8
30.0	*TMEM147*	19	295.9	282.8	1.0	22	592	35.3	47.2
30.3	*TERF2IP*	16	323.5	344.6	0.9	22	592	34.7	69.9
B, Testis									
18.5	*TUBA1B*	12	2,774.3	2,478.6	1.1	18	508	30.9	81.1
21.0	*FAM96B*	16	339.6	325.2	1.0	19	546	28.0	47.9
22.8	*RPL8*	8	1,637.9	2,069.0	0.8	18	513	27.9	44.1
23.0	*RPS18*	6	4,434.5	4,933.5	0.9	16	472	34.9	41.4
24.5	*ACTG1*	17	4,320.0	3,453.7	1.3	18	513	27.7	37.8
29.3	*RPS27*	1	5,444.6	4,355.6	1.3	18	513	32.5	44.1
31.5	*TBCB*	19	343.1	290.6	1.2	21	592	29.1	67.9
32.3	*RPL41*	12	4,676.1	5,838.5	0.8	20	568	36.1	45.1

[a]Please see Table VIII for a complete list of gene definitions. Value A, mean expression values of the gene in the pool A, including the sample related to the specific tissue; value B, mean expression values of the gene in the whole pool B consisting of all the 646 samples; A/B, ratio between these value A and value B; SD%, standard deviation expressed as percentage of the mean expression value (A or B) for the locus.

Systematic analysis aimed to evaluate if the genes identified as possible reference genes (Tables IV-VII) were significantly enriched in a particular class of genes confirmed to be involved in the most basic biological processes; in particular, in the metabolism of the informational macromolecules (nucleic acids and proteins). This therefore justifies their tendency to constitutive, stable and almost universal expression, which was also observed in a previous analysis that ranked genes by combining the average expression level and its SD in a single score (37). The biological peculiarity of HK genes was also highlighted by a significant difference in complexity between HK and tissue-specific gene promoters, as revealed by DNA entropy analysis (38).

While the present study was in progress, an article on the topic was published that suggested that a 'universal' human HK gene does not exist and provided a list of suitable reference genes for individual tissues/organs (11). However, the method employed by that study was different from the approach of the present study in a number of relevant aspects. In particular, the results of the previous study were originally obtained by combining the lists of genes retrieved from studies performed using heterogeneous techniques (such as microarray, EST or RNA-Seq analysis), were analyzed using the logic of classification, previous judgments concerning the suitability of certain genes as HK/reference genes were accepted and then these lists were combined. This approach of combining the lists of results was also used by Chang *et al* (39) and, following ranking, by Shaw *et al* (40). By contrast, the present study re-elaborated and normalized original raw data, and generated a fully quantitative analysis of human gene expression, which may explain certain differences in the results obtained by the algorithm used in the current study. Conversely, certain shared general conclusions were highlighted, including the necessity to calibrate the search criteria for HK genes according to cell/tissue type; however, the general analysis of the present study can still identify certain general-purpose genes with acceptable criteria that may be proposed as reference genes. Several previous studies have demonstrated that the results provided by the TRAM tool were highly reliable, having been confirmed by RT-qPCR experiments for several diverse human tissues, demonstrating a correlation coefficient (r) between TRAM and RT-qPCR data of r=0.98 for brain (34), r=0.99 for hippocampus (41) and r=0.98 for heart (14). However, it is commonly accepted that additional experimental studies may be required to verify that the identified candidate reference gene is suitable for the actual biological condition investigated (42). Additional studies are in progress to verify if the HK profile identified in normal tissues may be applicable to aneuploid cells, in particular for systematic analysis of trisomy 21 cells (43), in light of the fact that all of the best reference genes identified by that study are not

Table VIII. Gene symbols, corresponding descriptions and the number of recurrences (n) in the Tables for the genes listed in Tables III-VII.

Gene symbol	n	Gene description
ACTG1	**5**	**Actin γ 1**
AMZ2	1	Archaelysin family metallopeptidase 2
AP2M1	1	Adaptor-related protein complex 2, μ1 subunit
C14orf166	1	Chromosome 14 open reading frame 166
CNIH1	1	Cornichon family AMPA receptor auxiliary protein 1
CRELD2	1	Cysteine rich with EGF like domains 2
ECHS1	1	Enoyl-CoA hydratase, short chain 1
EEF1G	1	Eukaryotic translation elongation factor 1γ
EIF4A3	1	Eukaryotic translation initiation factor 4A3
EPCAM	1	Epithelial cell adhesion molecule
FAM96A	1	Family with sequence similarity 96 member A
FAM96B	**3**	**Family with sequence similarity 96 member B**
FIS1	1	Fission, mitochondrial 1
GABARAP	1	GABA type A receptor-associated protein
GABARAPL2	1	GABA type A receptor associated protein like 2
GANAB	1	Glucosidase II α subunit
GPX1	1	Glutathione peroxidase 1
GSTO1	1	Glutathione S-transferase ω1
GTF3A	1	General transcription factor IIIA
HDGF	1	Heparin binding growth factor
HEBP1	1	Heme binding protein 1
HEBP2	1	Heme binding protein 2
HIST1H2BK	1	Histone cluster 1 H2B family member k
IFI27	1	Interferon α-inducible protein 27
JTB	1	Jumping translocation breakpoint
LAMTOR5	1	Late endosomal/lysosomal adaptor, MAPK and MTOR activator 5
MB	1	Myoglobin
MIF	**2**	**Macrophage migration inhibitory factor (glycosylation-inhibiting factor)**
MRFAP1	1	Morf4 family associated protein 1
MRPL18	1	Mitochondrial ribosomal protein L18
MRPS24	1	Mitochondrial ribosomal protein S24
NDUFA3	1	NADH:ubiquinone oxidoreductase subunit A3
NDUFB1	1	NADH:ubiquinone oxidoreductase subunit B1
NDUFB4	1	NADH:ubiquinone oxidoreductase subunit B4
NDUFB11	1	NADH:ubiquinone oxidoreductase subunit B11
NDUFS3	1	NADH:ubiquinone oxidoreductase core subunit S3
NELFCD	1	Negative elongation factor complex member C/D
NOP10	**3**	**NOP10 ribonucleoprotein**
PGAM1	1	Phosphoglycerate mutase 1
POLR2H	1	RNA polymerase II subunit H
POLR2I	1	RNA polymerase II subunit I
POM121C	1	POM121 transmembrane nucleoporin C
POMP	1	Proteasome maturation protein
PRDX1	1	Peroxiredoxin 1
PSMC1	1	Proteasome 26S subunit, ATPase 1
RBX1	1	Ring-box 1
RNASEK	1	Ribonuclease K
RNF181	1	Ring finger protein 181
RPL6	1	Ribosomal protein L6
RPL8	**2**	**Ribosomal protein L8**

Table VIII. Continued.

Gene symbol	n	Gene description
RPL41	**3**	**Ribosomal protein L41**
RPLP0	1	Ribosomal protein lateral stalk subunit P0
RPS18	**2**	**Ribosomal protein S18**
RPS25	1	Ribosomal protein S25
RPS27	**3**	**Ribosomal protein S27**
RPS29	1	Ribosomal protein S29
RPSA	1	Ribosomal protein SA
RRAGA	**3**	**Ras related GTP binding A**
SLC25A3	1	Solute carrier family 25 member 3
SNRPD2	1	Small nuclear ribonucleoprotein D2 polypeptide
SOD1	1	Superoxide dismutase 1
TBCB	**2**	**Tubulin folding cofactor B**
TERF2IP	1	TERF2 interacting protein
TMEM147	1	Transmembrane protein 147
TOMM5	1	Translocase of outer mitochondrial membrane 5
TPT1	1	Tumor protein, translationally-controlled 1
TUBA1B	**2**	**Tubulin α1b**
YTHDF1	1	YTH N6-methyladenosine RNA binding protein 1

The genes with a number of recurrences (n) >1 are shown in bold. AMPA, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid; CoA, coenzyme A; EGF, epidermal growth factor; GTP, guanosine 5'-triphosphate; MAPK, mitogen-activated protein kinase; Morf4, mortality factor 4 (pseudogene); MTOR, mechanistic target of rapamycin; TERF2, telomeric repeat binding factor 2.

located on chromosome 21 and that a very small portion of this chromosome appears to be associated with the basic features of Down syndrome (DS) (44). In this regard, consulting the published differential transcriptome map comparing acute megakaryoblastic leukemia (AMKL) cells from children with DS (DS AMKL) and euploid megakaryocyte cells (euploid MK) (45) reveals promising values for *ACTG1* (DS AMKL/euploid MK gene expression ratio=1.08) and *POM121C* (DS AMKL/euploid MK gene expression ratio=1.01); the ratios range between 0.47 and 3.55 (45) for the other genes listed in Table III.

A previous study using RNA-Seq data to identify reference genes across multiple human tissues focused mainly on low SD and so proposed a list of 11 genes with exceptionally low variability (1). The corresponding values have been checked by TRAM analysis in the present study, which confirmed their expression across multiple tissues, with a generally low SD (although not exceptional in the data of the current study). However, these identified genes had low expression values, all in the range of 150-400 in terms of a percentage of the mean value. This was recognized by Kwon *et al* (46), who selected low variability as the leading parameter, as have other studies, whereas the approach of the present study was aimed at finding genes with the best combination of criteria, considering that a high expression value may be advantageous in practice for the usability of a reference gene. Finally, it should be noted that other studies have frequently used very different and original approaches to the problem, using computational classifiers (47), the controlled vocabulary of Medical Subject Headings (48) and Gene Ontology classifications (29).

It may finally be noted that several of the available analysis tools are aimed at determining the best reference genes for normalization of gene expression data, and are largely based on the evaluation of the minimal variation of the expression level of a given gene among different conditions. BestKeeper (49) and GeNorm (50) are commonly used for screening the reference genes to perform an accurate normalization of RT-qPCR data, whereas NormFinder (51) may be used to evaluate reference genes for normalization of RT-qPCR and microarray experiments. Regarding NormFinder, a direct comparison with the tool employed in the present study is not possible, since NormFinder requires the same number of measured values for each gene and the TRAM algorithm does not have this limitation. In addition, NormFinder evaluation is based only on the analysis of variation among different samples, whereas the TRAM approach integrates this parameter with the level of expression (at least medium-high) and with the highest possible number of samples in which each gene is measured; all are essential parameters by which to search for a suitable reference gene, thus allowing the identification of genes with the overall best fit to the three criteria. Finally, although NormFinder has the ability to identify the ideal combination of biologically independent genes for each tissue, this analysis requires the creation of a matrix with two groups of data deriving from two different conditions, which in the case of TRAM can be either one of the ten tissues analyzed or the whole pool. In the whole pool sample, the difference in the number of measures for each gene increases, so the exclusion of a large number of measures to create a matrix of data does not make the

Table IX. A list of all 63 genes identified for the 10 specific tissues and organs presented in Tables IV-VII.

Number	Gene
1	*ACTG1*
2	*AMZ2*
3	*AP2M1*
4	*C14orf166*
5	*CNIH1*
6	*CRELD2*
7	*ECHS1*
8	*EEF1G*
9	*EIF4A3*
10	*EPCAM*
11	*FAM96A*
12	*FAM96B*
13	*FIS1*
14	*GABARAP*
15	*GABARAPL2*
16	*GANAB*
17	*GPX1*
18	*GSTO1*
19	*GTF3A*
20	*HDGF*
21	*HEBP1*
22	*HEBP2*
23	*HIST1H2BK*
24	*IFI27*
25	*JTB*
26	*LAMTOR5*
27	*MB*
28	*MIF*
29	*MRFAP1*
30	*MRPS24*
31	*NDUFA3*
32	*NDUFB1*
33	*NDUFB11*
34	*NDUFB4*
35	*NDUFS3*
36	*NELFCD*
37	*NOP10*
38	*PGAM1*
39	*POLR2H*
40	*POLR2I*
41	*POMP*
42	*PRDX1*
43	*PSMC1*
44	*RBX1*
45	*RNASEK*
46	*RNF181*
47	*RPL41*
48	*RPL6*
49	*RPL8*
50	*RPLP0*
51	*RPS18*
52	*RPS25*
53	*RPS27*
54	*RPS29*
55	*RPSA*
56	*RRAGA*
57	*SLC25A3*
58	*SNRPD2*
59	*SOD1*
60	*TBCB*
61	*TERF2IP*
62	*TMEM147*
63	*TUBA1B*

direct comparison between NormFinder and TRAM results possible.

In conclusion, the present study provided, to the best of our knowledge, the first systematic analysis to quantitatively combine all of the traditional criteria aimed at identifying the HK genes that are best suited to be reference genes for the study of human gene expression. Several genes were identified and proposed to be suitable in cross-tissue studies, and certain genes were proposed as references for tissue/organ-specific studies. The wealth of data generated by this approach may also provide a representative portrait of typical gene-expression profiles for several human tissues and organs in the form of searchable database tables and suggested that currently uncharacterized transcripts, even EST clusters, may be worthy of further investigation as strong candidates to represent HK genes, or tissue-specific genes, expressed in high levels in human cells.

Acknowledgements

The authors wish to thank sincerely the Fondazione Umano Progresso, Milano, Italy for its fundamental support to our research on trisomy 21 and to the present study. The authors are pleased to thank Mrs. Vittoria Aiello and Mr. Massimiliano Albanese (Washington, DC, USA) for having undertaken an international initiative in support of our research in addition to all the donors contributing to this initiative listed at http://www.massimilianoalbanese.net/ds-research/?lang=en. The authors thank all the other people that have kindly contributed through individual donations to support part of the fellowships, in addition to computer hardware, that allowed the present study to be performed. In particular, we are grateful to Matteo and Elisa Mele (Bologna, Italy), to the Costa family, 'Gruppo Arzdore', 'Parrocchia di Dozza' the community of Dozza and 'Associazione Turistica Pro Loco di Dozza' (Dozza, Bologna, Italy), in addition to Mrs. Rina Bini (Fermo, Italy). The authors are grateful to Mrs. Kirsten Welter for her kind and expert revision of the manuscript. M.C.'s fellowship was funded by a donation from Fondazione Umano Progresso and by a grant from Fondazione Del Monte di Bologna e Ravenna (Bologna,

Italy). A.P. was funded by The Department of Experimental, Specialty and Diagnostic Medicine (University of Bologna, Bologna, Italy) and the Fondazione Umano Progresso. F.A. was funded by 'Gruppo Arzdore' and the Natali family (Petriolo, Macerata, Italy), in memory of Leonardo Natali. M.C.P. was funded by a donation from Fondazione Umano Progresso and by donations following the international fundraising initiative by Vittoria Aiello and Massimiliano Albanese.

References

1. Eisenberg E and Levanon EY: Human housekeeping genes, revisited. Trends Genet 29: 569-574, 2013.
2. Casadei R, Pelleri MC, Vitale L, Facchin F, Lenzi L, Canaider S, Strippoli P and Frabetti F: Identification of housekeeping genes suitable for gene expression analysis in the zebrafish. Gene Expr Patterns 11: 271-276, 2011.
3. Finnegan MC, Goepel JR, Hancock BW and Goyns MH: Investigation of the expression of housekeeping genes in non-Hodgkin's lymphoma. Leuk Lymphoma 10: 387-393, 1993.
4. Lion T: Current recommendations for positive controls in RT-PCR assays. Leukemia 15: 1033-1037, 2001.
5. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M and Johnston R: An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Res 29: E41-E41, 2001.
6. Warrington JA, Nair A, Mahadevappa M and Tsyganskaya M: Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiol Genomics 2: 143-147, 2000.
7. Lee PD, Sladek R, Greenwood CM and Hudson TJ: Control genes and variability: Absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res 12: 292-297, 2002.
8. Perfetti V, Manenti G and Dragani TA: Expression of housekeeping genes in Hodgkin's disease lymph nodes. Leukemia 5: 1110-1112, 1991.
9. Bhatia P, Taylor WR, Greenberg AH and Wright JA: Comparison of glyceraldehyde-3-phosphate dehydrogenase and 28S-ribosomal RNA gene expression as RNA loading controls for northern blot analysis of cell lines of varying malignant potential. Anal Biochem 216: 223-226, 1994.
10. Barber RD, Harmer DW, Coleman RA and Clark BJ: GAPDH as a housekeeping gene: Analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiol Genomics 21: 389-395, 2005.
11. Zhang Y, Li D and Sun B: Do housekeeping genes exist? PLoS One 10: e0123691, 2015.
12. Lenzi L, Facchin F, Piva F, Giulietti M, Pelleri MC, Frabetti F, Vitale L, Casadei R, Canaider S, Bortoluzzi S, *et al*: TRAM (Transcriptome Mapper): Database-driven creation and analysis of transcriptome maps from multiple sources. BMC Genomics 12: 121, 2011.
13. Piovesan A, Vitale L, Pelleri MC and Strippoli P: Universal tight correlation of codon bias and pool of RNA codons (codonome): The genome is optimized to allow any distribution of gene expression values in the transcriptome from bacteria to humans. Genomics 101: 282-289, 2013.
14. Caracausi M, Piovesan A, Vitale L and Pelleri MC: Integrated transcriptome map highlights structural and functional aspects of the normal human heart. J Cell Physiol 232: 759-770, 2017.
15. Barrett T and Edgar R: Gene expression omnibus: Microarray data storage, submission, retrieval and analysis. Methods Enzymol 411: 352-369, 2006.
16. Emig D, Salomonis N, Baumbach J, Lengauer T, Conklin BR and Albrecht M: AltAnalyze and DomainGraph: Analyzing and visualizing exon expression data. Nucleic Acids Res 38 (Web Server issue): W755-W762, 2010.
17. Lenzi L, Frabetti F, Facchin F, Casadei R, Vitale L, Canaider S, Carinci P, Zannotti M and Strippoli P: UniGene Tabulator: A full parser for the UniGene format. Bioinformatics 22: 2570-2571, 2006.
18. Piovesan A, Caracausi M, Ricci M, Strippoli P, Vitale L and Pelleri MC: Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank. DNA Res 22: 495-503, 2015.
19. Butte AJ, Dzau VJ and Glueck SB: Further defining housekeeping, or 'maintenance,' genes Focus on 'A compendium of gene expression in normal human tissues'. Physiol Genomics 7: 95-96, 2001.
20. Tu Z, Wang L, Xu M, Zhou X, Chen T and Sun F: Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics 7: 31, 2006.
21. Pilbrow AP, Ellmers LJ, Black MA, Moravec CS, Sweet WE, Troughton RW, Richards AM, Frampton CM and Cameron VA: Genomic selection of reference genes for real-time PCR in human myocardium. BMC Med Genomics 1: 64, 2008.
22. Berriz GF, King OD, Bryant B, Sander C and Roth FP: Characterizing gene sets with FuncAssociate. Bioinformatics 19: 2502-2504, 2003.
23. Lü B, Yu J, Xu J, Chen J and Lai M: A novel approach to detect differentially expressed genes from count-based digital databases by normalizing with housekeeping genes. Genomics 94: 211-216, 2009.
24. Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J and Teufel A: RNA-Seq Atlas-a reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics 28: 1184-1185, 2012.
25. Chen M, Xiao J, Zhang Z, Liu J, Wu J and Yu J: Identification of human HK genes and gene expression regulation study in cancer from transcriptomics data analysis. PLoS One 8: e54082, 2013.
26. Costa Ade F and Franco OL: Insights into RNA transcriptome profiling of cardiac tissue in obesity and hypertension conditions. J Cell Physiol 230: 959-968, 2015.
27. Zhao S, Fung-Leung WP, Bittner A, Ngo K and Liu X: Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 9: e78644, 2014.
28. Guo Y, Sheng Q, Li J, Ye F, Samuels DC and Shyr Y: Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. PLoS One 8: e71462, 2013.
29. Zhang Y, Akintola OS, Liu KJ and Sun B: Membrane gene ontology bias in sequencing and microarray obtained by housekeeping-gene analysis. Gene 575: 559-566, 2016.
30. Lee S, Jo M, Lee J, Koh SS and Kim S: Identification of novel universal housekeeping genes by statistical analysis of microarray data. J Biochem Mol Biol 40: 226-231, 2007.
31. Zhu J, He F, Song S, Wang J and Yu J: How many human genes can be defined as housekeeping with current expression data? BMC Genomics 9: 172, 2008.
32. de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ and ter Elst A: Evidence based selection of housekeeping genes. PLoS One 2: e898, 2007.
33. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, *et al*: Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics 13: 397-406, 2014.
34. Caracausi M, Vitale L, Pelleri MC, Piovesan A, Bruno S and Strippoli P: A quantitative transcriptome reference map of the normal human brain. Neurogenetics 15: 267-287, 2014.
35. Casadei R, Piovesan A, Vitale L, Facchin F, Pelleri MC, Canaider S, Bianconi E, Frabetti F and Strippoli P: Genome-scale analysis of human mRNA 5' coding sequences based on expressed sequence tag (EST) database. Genomics 100: 125-130, 2012.
36. She X, Rohl CA, Castle JC, Kulkarni AV, Johnson JM and Chen R: Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genomics 10: 269, 2009.
37. Popovici V, Goldstein DR, Antonov J, Jaggi R, Delorenzi M and Wirapati P: Selecting control genes for RT-QPCR using public microarray data. BMC Bioinformatics 10: 42, 2009.
38. Thomas D, Finan C, Newport MJ and Jones S: DNA entropy reveals a significant difference in complexity between housekeeping and tissue specific gene promoters. Comput Biol Chem 58: 19-24, 2015.
39. Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, Huang CL and Hsu IC: Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS One 6: e22859, 2011.
40. Shaw GT, Shih ES, Chen CH and Hwang MJ: Preservation of ranking order in the expression of human Housekeeping genes. PLoS One 6: e29314, 2011.
41. Caracausi M, Rigon V, Piovesan A, Strippoli P, Vitale L and Pelleri MC: A quantitative transcriptome reference map of the normal human hippocampus. Hippocampus 26: 13-26, 2016.

42. Cheng WC, Chang CW, Chen CR, Tsai ML, Shu WY, Li CY and Hsu IC: Identification of reference genes across physiological states for qRT-PCR through microarray meta-analysis. PLoS One 6: e17347, 2011.
43. Strippoli P, Pelleri MC, Caracausi M, Vitale L, Piovesan A, Locatelli C, Mimmi MC, Berardi AC, Ricotta D, Radeghieri A, *et al*: An integrated route to identifying new pathogenesis-based therapeutic approaches for trisomy 21 (Down Syndrome) following the thought of Jérôme Lejeune. Sci Postprint 1: e00010, 2013.
44. Pelleri MC, Cicchini E, Locatelli C, Vitale L, Caracausi M, Piovesan A, Rocca A, Poletti G, Seri M, Strippoli P and Cocchi G: Systematic reanalysis of partial trisomy 21 cases with or without Down syndrome suggests a small region on 21q22.13 as critical to the phenotype. Hum Mol Genet 25: 2525-2538, 2016.
45. Pelleri MC, Piovesan A, Caracausi M, Berardi AC, Vitale L and Strippoli P: Integrated differential transcriptome maps of acute megakaryoblastic leukemia (AMKL) in children with or without down syndrome (DS). BMC Med Genomics 7: 63, 2014.
46. Kwon MJ, Oh E, Lee S, Roh MR, Kim SE, Lee Y, Choi YL, In YH, Park T, Koh SS and Shin YK: Identification of novel reference genes using multiplatform expression data and their validation for quantitative gene expression analysis. PLoS One 4: e6162, 2009.
47. Chiang AW, Shaw GT and Hwang MJ: Partitioning the human transcriptome using HKera, a novel classifier of housekeeping and tissue-specific genes. PLoS One 8: e83040, 2013.
48. Ersahin T, Carkacioglu L, Can T, Konu O, Atalay V and Cetin-Atalay R: Identification of novel reference genes based on MeSH categories. PLoS One 9: e93341, 2014.
49. Pfaffl MW, Tichopad A, Prgomet C and Neuvians TP: Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper-Excel-based tool using pair-wise correlations. Biotechnol Lett 26: 509-515, 2004.
50. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A and Speleman F: Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3: RESEARCH0034, 2002.
51. Andersen CL, Jensen JL and Ørntoft TF: Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res 64: 5245-5250, 2004.