Open access • Posted Content • DOI:10.1101/718007

The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data — Source link

Marina Wright Muelas, Farah Mughal, Steve O'Hagan, Philip J. R. Day ...+1 more authors

Institutions: University of Liverpool, University of Manchester

Published on: 31 Jul 2019 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Reference genes, Housekeeping gene and Gene expression profiling

SCIENTIFIC REPORTS
nature research

OPEN

The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data

Marina Wright Muelas[1*], Farah Mughal[1], Steve O'Hagan[2,3], Philip J. Day[3,4*] & Douglas B. Kell[1,5*]

We recently introduced the Gini coefficient (GC) for assessing the expression variation of a particular gene in a dataset, as a means of selecting improved reference genes over the cohort ('housekeeping genes') typically used for normalisation in expression profiling studies. Those genes (transcripts) that we determined to be useable as reference genes differed greatly from previous suggestions based on hypothesis-driven approaches. A limitation of this initial study is that a single (albeit large) dataset was employed for both tissues and cell lines. We here extend this analysis to encompass seven other large datasets. Although their absolute values differ a little, the Gini values and median expression levels of the various genes are well correlated with each other between the various cell line datasets, implying that our original choice of the more ubiquitously expressed low-Gini-coefficient genes was indeed sound. In tissues, the Gini values and median expression levels of genes showed a greater variation, with the GC of genes changing with the number and types of tissues in the data sets. In all data sets, regardless of whether this was derived from tissues or cell lines, we also show that the GC is a robust measure of gene expression stability. Using the GC as a measure of expression stability we illustrate its utility to find tissue- and cell line-optimised housekeeping genes without any prior bias, that again include only a small number of previously reported housekeeping genes. We also independently confirmed this experimentally using RT-qPCR with 40 candidate GC genes in a panel of 10 cell lines. These were termed the Gini Genes. In many cases, the variation in the expression levels of classical reference genes is really quite huge (e.g. 44 fold for GAPDH in one data set), suggesting that the cure (of using them as normalising genes) may in some cases be worse than the disease (of not doing so). We recommend the present data-driven approach for the selection of reference genes by using the easy-to-calculate and robust GC.

In a recent paper[1], we introduced the Gini index (or Gini coefficient, GC)[2–5] as a very useful, nonparametric statistical measure for identifying those genes whose expression varied least across a large set of samples (when normalised appropriately[6] to the total expression level of transcripts). The GC is a measure that is widely used in economics (e.g.[4,7–12]) to describe the (in)equality of the distribution of wealth or income between individuals in a population. However, although it could clearly be used to describe the variation in any other property between individual examples[13–16]), it has only occasionally been used in epidemiology[17–19] and in biochemistry[1,5,20–25]. Its visualisation and calculation are comparatively straightforward (Fig. 1): individual examples are ranked on the

[1]Department of Biochemistry, Institute of Integrative Biology, Faculty of Health and Life Sciences, University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK. [2]School of Chemistry, Department of Chemistry, The Manchester Institute of Biotechnology 131, Princess Street, Manchester, M1 7DN, UK. [3]The Manchester Institute of Biotechnology, 131, Princess Street, Manchester, M1 7DN, UK. [4]Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, M13 9PL, UK. [5]Novo Nordisk Foundation Centre for Biosustainability, Technical University of Denmark, 10 Building 220, Kemitorvet, 2800, Kgs. Lyngby, Denmark. *email: m.wright-muelas@liverpool.ac.uk; Philip.J.Day@manchester.ac.uk; dbk@liv.ac.uk

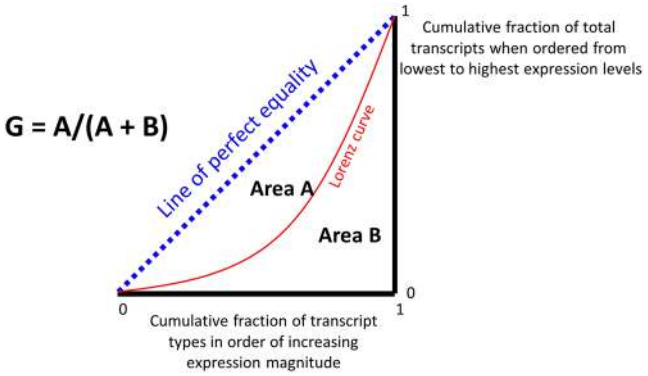

Figure 1. Graphical indication of the means by which we calculate the Gini coefficient.

abscissa in increasing order of the size of their contribution, and the cumulative contribution is plotted against this on the ordinate. The GC is given by the fractional area mapped out by the resulting 'Lorenz' curve (Fig. 1). For a purely 'socialist' system in which all contributions are equal (GC = 0), the curve joins the normalised 0,0 and 1,1 axes, while for a complete 'autocracy', in which the resource or expression is held or manifest by only a single individual (GC = 1), the 'curve' follows the two axes (0,0 → 1,0 → 1,1).

Since the early origins of large-scale nucleic acid expression profiling, especially those using microarrays[26–28], it has been clear that expression profiling methods are susceptible to a variety of more or less systematic artefacts within an experiment, whose resolution would require or benefit from some kind of normalisation (e.g.[29–39]). By this ('normalisation of the first kind'), and what is typically done, we mean the smoothing out of genuine artefacts within an array or a run, that occur simply due to differences in temperature or melting temperature or dye binding or hybridisation and cross-hybridisation efficiency (and so on) across the surface of the array. This process can in principle use reference genes, but usually exploits smoothing methods that normalise geographically local subsets of the genes to a presumed distribution.

Even after this is done, there is a second level of normalisation, that between chips or experiments, that is usually done separately, not least because it is typically much larger and more systematic, especially because of variations in the total amount of material in the sample analysed or of the overall sensitivity of the detector (much as is true of the within-run versus between-run variations observed in mass spectrometry experiments[40,41]). This kind of normalising always requires 'reference' genes whose expression varies as little as possible in response to any changes in experimental conditions. The same is true for expression profiling as performed by qPCR[42–47], where the situation is more acute regarding the choice of reference genes since primers must be selected for these a priori. Commonly, the geometric mean of the expression levels of that or those that vary the least is selected as the 'reference'. The question then arises as to which are the premium 'reference' genes to choose.

Data-driven and hypothesis-dependent science are complementary, though when a field is data-rich but hypothesis-poor, as is genomics, data-driven strategies are to be preferred[48]. Perhaps surprisingly[48], rather than simply letting the data speak for themselves, choices of candidate reference genes were often made on the basis that reference genes should be 'housekeeping' genes that would simply be assumed ('hypothesised') to vary comparatively little between cells, be involved in nominal routine metabolism and also that they should have a reasonably high expression level (e.g.[49–66]). This is not necessarily the best strategy, and there is in fact (and see below) quite a wide degree of variation of the expression of most standard housekeeping genes between cells or tissues (e.g.[53,62,65,67–79]). Indeed, Lee *et al.*[69] stated explicitly that housekeeping genes may be uniformly expressed in certain cell types but may vary in others, especially in clinical samples associated with disease.

It became obvious that an analysis of the GC of the various genes was actually precisely what was required to assess those 'housekeeping' (or any other) genes that varied least across a set of expression profiles, and we found 35 transcripts for which the GC was 0.15 or below when assessing 56 mammalian cell lines taken from a wide variety of tissues[1]. These we refer to as the 'Gini genes'. Most of these were 'novel' as they had never previously been considered as reference genes, and we noted that their Gini indices were significantly smaller (they were more stably expressed) than were those of the more commonly used reference genes[66]. However, this analysis was done on only one (albeit large) dataset of gene expression profiles. While some of the compilations (e.g.[65,80]) contain massive amounts of expression profiling data, many of these, especially the older ones, may well be of uncertain quality. Thus, especially since the GC is very prone to being raised by small numbers of large outliers, we decided for present purposes that we should compare our analyses of candidate Gini genes using a smaller but carefully chosen set of expression profiling experiments. The more modern RNA-seq (e.g.[81–85]), in which individual transcripts are simply counted digitally via direct sequencing, is seen as considerably more robust[81,86,87] and sensitive[88,89], and so we selected additional large and recent datasets that used RNA-seq in cell lines and tissues (Table 1). We note too that the precision of these digital methods (as with other, digital, single-molecule strategies[90–92]), means that the requirement for reasonably high-level expression levels is much less acute.

In a similar vein (Table 2), we selected a small number of reasonably detailed studies in which particular housekeeping genes had been proposed as reference genes.

To our knowledge, there are no large-scale studies to determine housekeeping genes in large, cell-line cohorts; the present paper serves to provide one. In addition, we include an experimental RT-qPCR analysis of a subset of the Gini genes.

Study short name	Comments	Reference
GiniGene	Study presenting novel potential housekeeping genes in cells and tissues from the HPA project cell and tissue RNA-seq data.	1
geNorm or Vandesompele	Classic set of reference genes in tissues and a means of analysing them	66
Eisenberg	Very detailed analysis of housekeeping/ reference genes in tissues using the Illumina Body Map study of RNA-seq of 16 Human Tissues. E-MTAB-513.	49
Lee	Two novel reference genes from a detailed analysis of 281 normal tissue samples from 17 different organs then compares between disease states m and cell lines.	131
Caracausi	646 expression profile data sets from 54 different human tissues.	65

Table 1. Studies used for assessing proposed stable reference genes.

Dataset short name	Comments	Reference
HPA	RNA-seq-based dataset from the Human Protein Atlas group. Two data sets available: one of 19,628 protein coding genes in 56 cell lines (HPA_C) and another of 19,613 protein coding genes in 59 tissues (HPA_T).	1,93,140
CCLE	RNA-seq-based dataset (Cancer Cell Line Encyclopedia) of 58,035 genes in 934 human cancer cell lines (downloaded from EBI Expression Atlas E-MTAB2770).	141
Klijn / Genentech	RNA-seq-based analysis of 57,711 genes in 622 human cancer cell lines (downloaded from EBI Expression Atlas E-MTAB-2706).	142
GTEx	RNA-Seq data of 46,711 genes in 53 human tissue samples from the Genotype-Tissue Expression (GTEx) project (downloaded from EBI Expression Atlas E-MTAB-5214).	143
PCAWG	RNA-Seq of 46,816 genes in 76 tissues, cancer and normal, from The International Cancer Genome Project: Pan Cancer Analysis of Whole Genomes (downloaded from EBI Expression Atlas E-MTAB-5200).	https://dcc.icgc.org/pcawg
HBM	Illumina Body Map: RNA-seq of 16 Human tissues. (downloaded from EBI Expression Atlas E-MTAB-513). Used by Eisenberg and colleagues in their analysis of housekeeping/ reference genes in tissues.	49

Table 2. Studies used for expression profiling data.

Results

The Gini Coefficient as a robust measure of gene expression stability in multiple cell-line data sets. We previously identified a number of genes in the Human Protein Atlas (HPA) cell line data set[93] with very low expression variability and thus potential for use as reference genes[1]. However, we did not compare these Gini genes to other genes that have previously been proposed as housekeeping genes. We therefore performed a similar analysis using the potential housekeeping genes we proposed in[1] as well as other reference genes proposed in other studies (Table 2) with additional large RNA-Seq cell line data sets (Table 1).

Figure 2A shows a plot of the GC of a variety of candidate Gini genes versus their median expression level in the HPA cell lines dataset set[93]. It is clear that genes we identified previously have much lower GC values in the HPA dataset than do any of the others (just two, VPS29 and CHMP2A, were also identified by Eisenberg and Levanon and another, RPL41, by Caracausi). This is not at the expense of an unusually low expression (Fig. 2A), a finding broadly confirmed when we look at the median expression levels for the CCLE dataset (Fig. 2B) and of the Klijn dataset (Fig. 2C).

Figure 3 shows the GC values for the various genes in two other datasets, viz CCLE and Klijn. Our previous Gini genes have a lower GC than that of any of the other housekeeping genes in 25 out of 38 cases in Klijn (all under 0.2) and in 26 out of 40 cases for CCLE (all under 0.22). In confirmation of this, and of the correlation found above between the median expression levels in CCLE and Klijn, the GC values are also well correlated with each other for the two datasets (Fig. 3). Thus, although the absolute numbers are slightly larger than are those for the HPA dataset (unsurprisingly, given the much larger number of examples), the trend is still very clear: the GiniGenes remain the best among those variously proposed as reference genes in a variety of large and quite independent datasets. It also suggests that variations in the total amount of mRNA are not an issue either.

Another common statistical measure, more resistant to individual outliers, is the interquartile ratio (the ratio between the 25th and 75th percentile when expression levels are ranked); by this measure too, the Gini genes that we uncovered previously stand out as being the least varying (Fig. 4 A,B). This suggests that, as a measure of gene expression stability, the GC is robust: the GiniGenes have the lowest ratio between their maximum and minimum expression values in the HPA dataset (Fig. 4C) and also the lowest interquartile ratio in their levels of expression in all three cell line data sets explored here (Fig. 4B,C) with good correlation between these two datasets.

Use of the Gini Coefficient to find GiniGenes in an unbiased manner in cell-line data sets. Up to now, our analyses of these data sets have used a set of predefined genes to look at expression stability. We next sought to investigate whether the GC would highlight genes with high expression stability that have been reported by others or by ourselves when performing this analysis in a data-driven manner. To that end, we found 115 genes shared between the three data sets with a GC $\leq$ 0.2 (Figs. 5, 6). This value for the GC was chosen since reducing this to $\leq$0.15 meant no or very few genes were found in some data sets (e.g. no genes in the CCLE data set had a GC $\leq$ 0.15) and going above this meant the number of genes were unmanageable (e.g. 1051 genes with a GC $\leq$ 0.21 in the Klijn data set). Of the 115 genes shared between the datasets with GC $<$ 0.2, 13 were GiniGenes

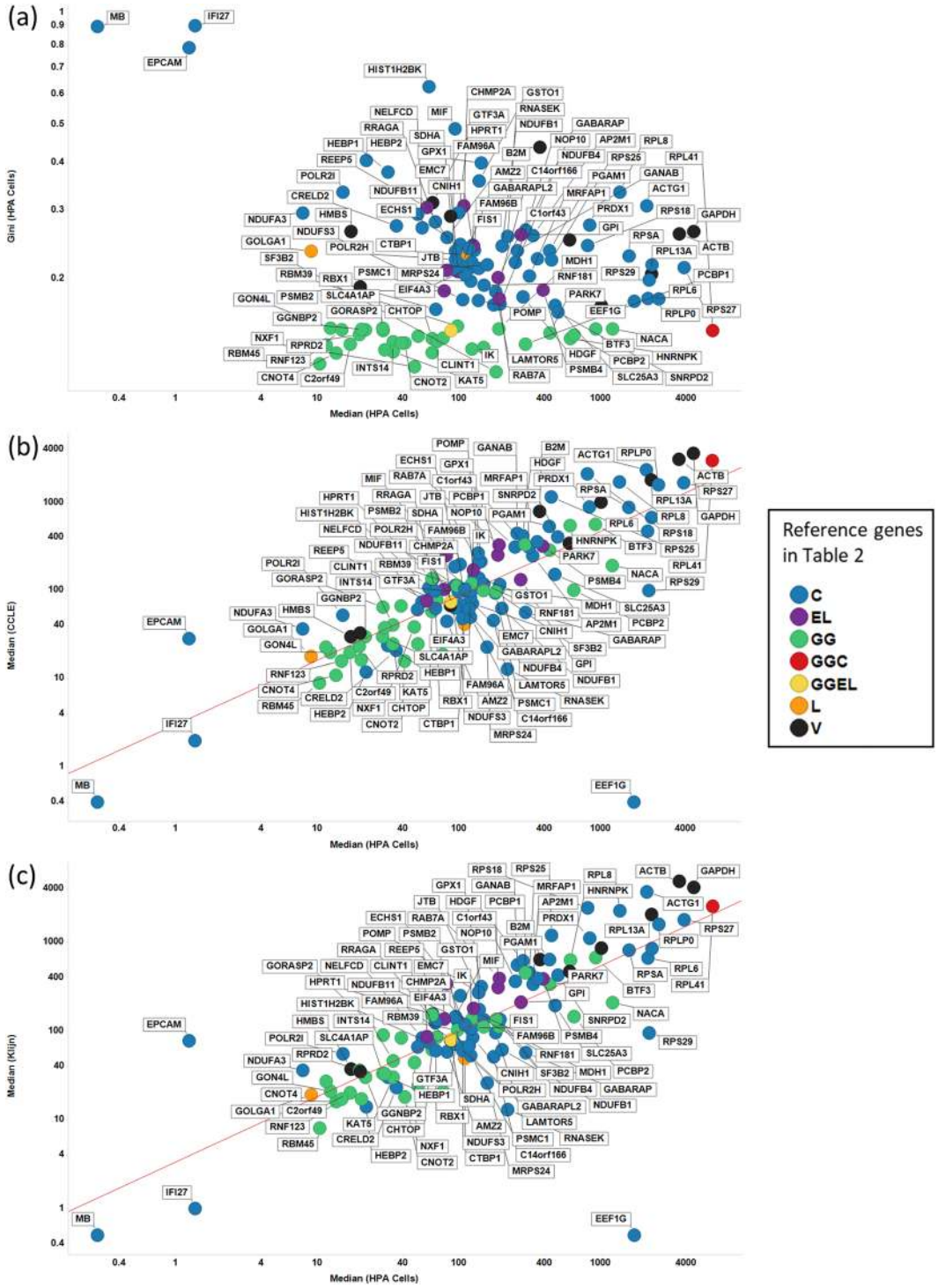

Figure 2. Gini coefficient and median expression levels of proposed reference genes in the HPA cell-line dataset. (**A**) GC versus median expression level of HPA dataset. (**B**) Median expression levels of CCLE vs HPA datasets. Line of best linear fit (in log space) shown is $y = 0.991 + 0.827 \times (r^2 = 0.606)$. (**C**) Median expression levels of CCLE vs Klijn datasets. Line of best linear fit (in log space) shown is $y = 0.998 + 0.804 \times (r^2 = 0.593)$. Colour coding: red, GeneGini reference genes; blue Eisenberg & Levanon; yellow Vandesompele; green Lee; lilac both GeneGini and Eisenberg and Levanon.

and two were housekeeping genes defined by Caracausi and colleagues (Fig. 5B). When we selected the top 20 expressing genes in each data set, only 13 of these were common across these data sets; Table 3 shows some descriptive statistics of 13 of these, with descriptive statistics of all 115 genes found in Supplementary Table S1. Of these genes, two (HNRNPK and PCBP1) are GiniGenes and one (SLC25A3) is a gene previously reported by Caracausi *et al*. Seven out of the 13 genes (HNRNPK, HNRNPC, PCBPB, SF3B1, SRSF3, EDF1 and EIF4H) here

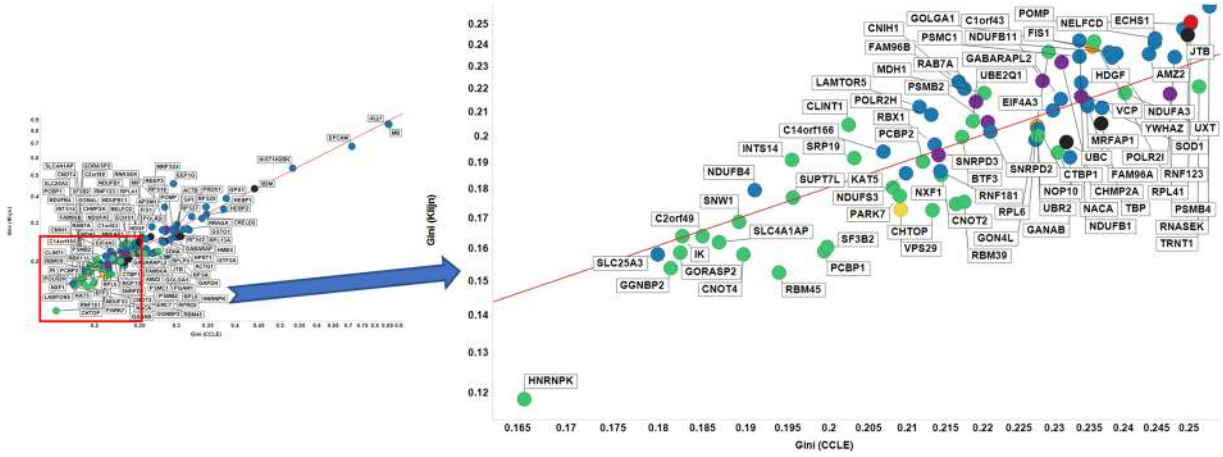

Figure 3. Gini coefficient of candidate reference genes in CCLE and Klijn/Genentech cell-line datasets. Left panel shows all proposed housekeeping genes considered in this study, with the right panel showing labels of those genes with a GC < 0.25. The line of best fit is $y = -0.171 + 0.829 \times (r^2 = 0.909)$. Colour code as in Fig. 2.

share important roles in RNA transcription, translation and stability[94–102], are implicated in a number of diseases, including cancer[94,97,103–113], and some, such as SRSF3 are essential for embryo development[114]. Given their pivotal functions, it may be unsurprising that the expression of these genes are tightly regulated across cell lines of different tissue origins, even where these are cancer cell lines. Overall, the distribution, expression stability and important functional roles of these genes suggest that these are excellent housekeeping genes across different cell types.

Of particular interest to us was finding one gene encoding a mitochondrial phosphate transporter protein (SLC25A3[115]) to be within this list of the top expressing stably expressed genes. This might seem logical since mitochondrial ATP synthesis is required by all cell types and tissues.

Figure 7 shows the robustness of the GC for the subset of 115 genes common between the three data sets studied here with a low GC (<0.2). Lower Gini coefficients correlate with lower IQR and Max:median ratios (Fig. 7: only results for the Klijn data set are shown). The range of IQR values of these genes was smaller in the larger two data sets (CCLE, 1.42–1.67; Klijn, 1.30–1.64) than in the HPA data set (1.26–1.84) suggesting the measured expression values were more stable in the larger data sets (Supplementary Table S1). This may, however, be due to a larger number of cell lines in these two large datasets (934 and 622 in CCLE and Klijn) compared with the HPA data set (56 cell lines).

Application of the Gini coefficient to human tissue RNA-Seq data sets. The results presented thus far are representative of human cell lines. Most reports in the literature regarding housekeeping genes refer to tissue expression data. This may be due to the cell lines being "dedifferentiated" with respect to the tissues from which they are derived[116].

In our previous report[1] we also analysed RNA-Seq data from tissues[93] and found 22 genes with a GC < 0.15, of which 3 (CHMP2A, VPS29 and PCBP1) were also found in cell line data with a GC < 0.15. The median expression level and GC of these and other candidate GiniGenes in this tissue data set are shown in Fig. 8. As with cell line data, the genes we previously identified (GGs, green dots in Fig. 8) have much lower GCs in this tissue data set than do any of the other candidate GiniGenes, with only two of these genes (VPS29 and CHMP2A) identified previously by Eisenberg & Levanon[49]. The low GC value of these GiniGenes is not at the expense of low expression: of the 22 GiniGenes, 13 are expressed at a median level of between 40 and 200 TPM (see Supplementary Table S2). Moreover, the GC was also representative of the variation in expression of these genes (albeit influenced to a lesser extent by outliers), as shown in Fig. 9A,B, with all GiniGenes having a GC < 0.15 and the lowest RSD (relative standard deviation), ranging from 24.096% to 28.66% and IQR (1.26 to 1.44) of this list of housekeeping genes. The expression of other housekeeping genes such as GAPDH, ACTB, RPL13A, SDHA, B2M was quite varied according to these measures. For example, the GC of GAPDH (a commonly used HKG) was 0.33, with a RSD of 72.4% and IQR of 2.24, and for ACTB (another commonly used HKG) these values were 0.29, 55.24%, and 2.11.

The median expression levels of the proposed reference genes show a similar level of correlation between the data sets as was found with the cell line data (Fig. S1A–C), and GiniGenes displayed a mid-range level of expression. The GC of the tissue GiniGenes we proposed however, tended to be higher and more variable in their GC values than in the HPA dataset (Fig. S2,A–C) suggesting that those genes may be representative of the HPA data set only. As an example, in the GTEx dataset only 28 genes had a GC < 0.2, of which the majority (17) were those reported by Caracausi and colleagues, and 7 were GiniGenes. The results here are likely influenced by the number and status (disease or normal) of the tissues analysed in the various data sets compared; for example, the GTEx data come from 53 different, normal human tissues, whereas the HPA tissue data include a mixture of disease and normal tissue samples. In addition, compared to the cell line data where hundreds (in the case of the Cancer Cell Line Encyclopedia) of cell lines were analysed, the number of tissues in these data sets was fewer than 100.

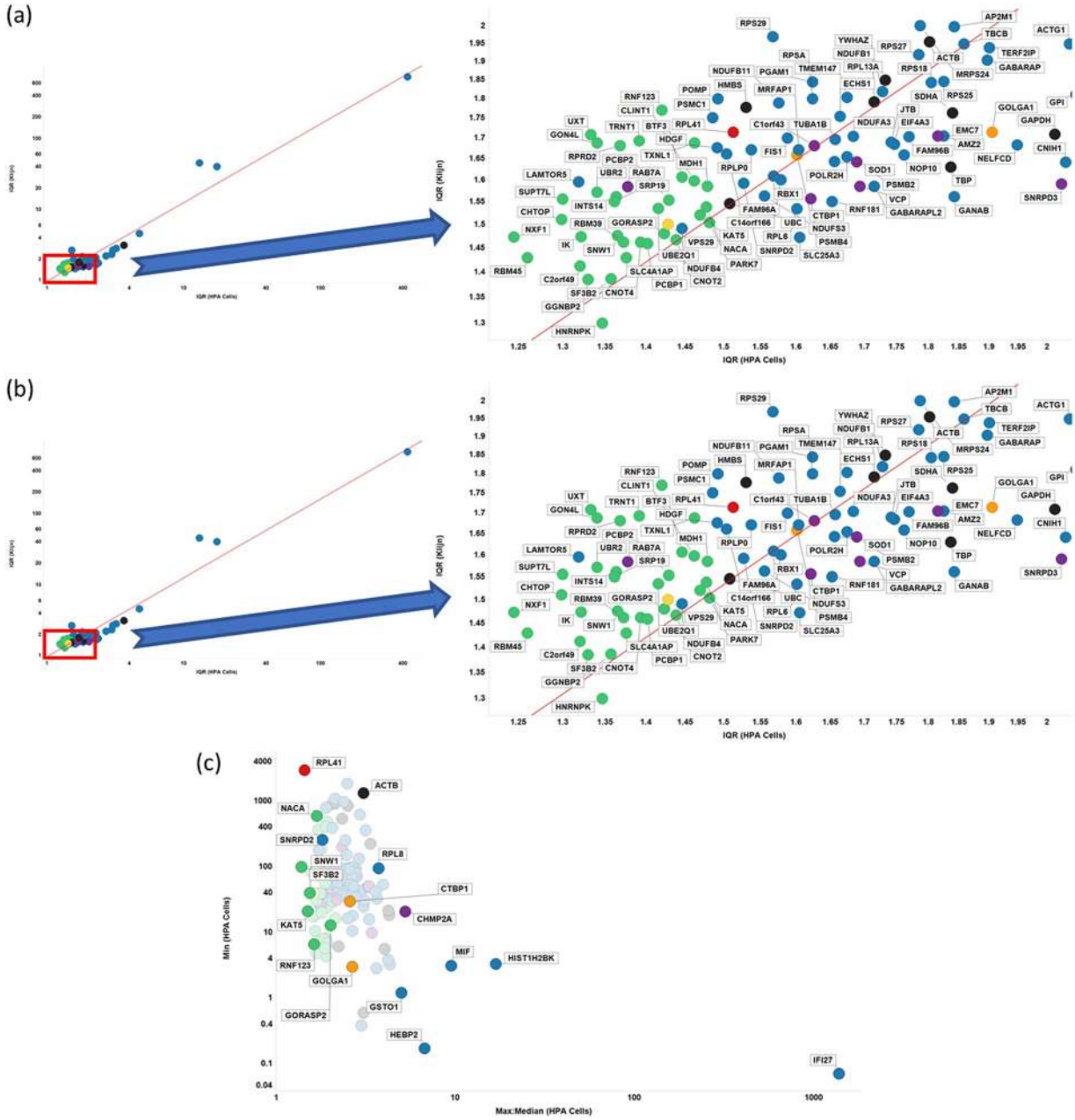

Figure 4. Robustness of the Gini coefficient. (**A**) IQR of different genes in Klijn/Genentech vs HPA cell-line dataset. Left panel shows all genes considered in this study, with right panel showing genes with IQR < 2 in both datasets. Line of best linear fit (in log space) shown is $y = 0.01 + 1.11 \times (r^2 = 0.937)$. (**B**) IQR of different genes in CCLE vs HPA cell-line dataset. Left panel shows all genes considered in this study, with right panel showing genes with IQR < 2 in both datasets. Line of best linear fit (in log space) shown is $y = 0.04 + 0.99 \times (r^2 = 0.930)$. (**C**) Min vs Max: Median expression levels in HPA data set. Colour code as in Fig. 2.

In the case of the data set used by Eisenberg and Levanon[49], viz. the Illumina Human Body Map (E-MTAB-513), 10 of the 11 housekeeping genes proposed here (which included 2 Gini Genes, CHMP2A and VPS29) had a GC ≤ 0.2 and were reasonably well expressed (with median expression levels between 50–270 TPM, see Supplementary Table S2 and Supplementary Fig. S4). This may be compared to the 5 other GGs with GC $<$ 0.2 in this data set whose expression value was lower, with median expression between 19–35 TPM. This suggests that finding suitable HKGs may be dependent on the data set itself, and the type of tissue under investigation.

We next sought to perform a more comprehensive and integrative analysis by filtering the tissue data sets to only include genes with a GC ≤ 0.2 to find common genes across these data sets with reasonable expression stability (Supplementary Table S3). As shown in Fig. 10 only 15 genes were shared between the four data sets with a GC ≤ 0.2, none of which has been reported previously as a housekeeping genes. Table 4 shows some descriptive statistics of these genes. In any case, the names of the proteins encoded by these 15 genes suggest these play important and essential roles. The median expression values of these genes varied from around 10–450 TPM, with SNX3

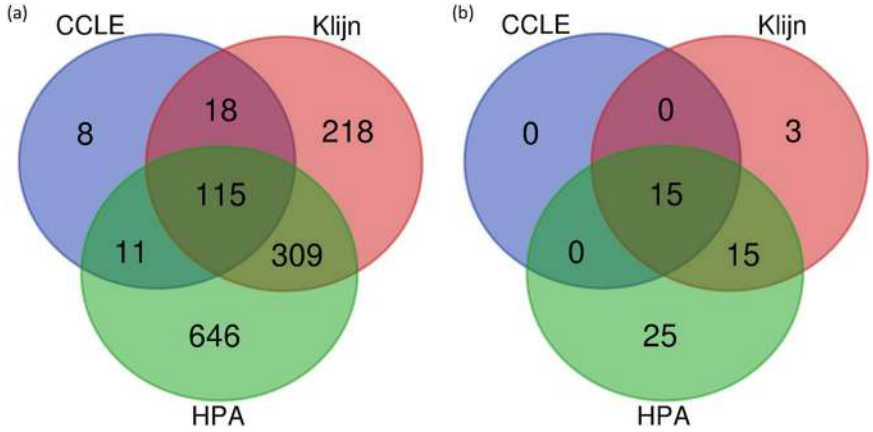

Figure 5. Shared and unique genes in HPA, CCLE and Klijn/Genentech cell-line data sets. (**A**) Genes with a GC < 0.2 **.(B)** Housekeeping genes in Table 2 with GC < 0.2.

(Sorting nexin-3 (Protein SDP3)) and COX4I1 (Cytochrome c oxidase subunit 4 isoform 1) being consistently the two highest-expressing genes.

Sorting nexins are a group of cytoplasmic and membrane-associated proteins involved in the regulation of intracellular trafficking[117]. SNX3 has been reported to play a role in receptor recycling and formation of multivesicular bodies[118], and its dysregulation has been implicated in disorders of iron metabolism and the pathogenesis of some neurodegenerative diseases[119,120].

The COX4I gene encodes the nuclear-encoded cytochrome c oxidase subunit 4 isoform 1, the terminal enzyme in the mitochondrial respiratory chain. Given the key role of the mitochondrial respiratory chain in all human cells (except red blood cells), stable expression of such a gene in all tissues may not be a surprising result. Increased RNA COX4I1 levels have been reported in sperm of an obese male rat model[121] and thus may play a role in obesity-related fertility problems, and reduced expression of this subunit leads to a reduction in mitochondrial respiration as well as sensitising cells to apoptosis[122].

The small number of genes shared between these data sets with a GC < 0.2 indicates that the data in these studies are more variable compared to cell lines alone. The cause of this variation may be due to the tissue data having been obtained from different subjects[123]. Moreover, tissues are themselves a mixture of cell types with varying levels of gene expression in each cell type[124], while cell lines are nominally clonal.

Our results suggest that in the case of RNA-seq tissue data sets, where gene expression tends to be more variable, an unbiased approach, using the Gini coefficient, may be more fruitful in the search for stably expressed genes with which to perform normalisation, than the other commonly used methods used until now[123,125].

RT-qPCR analysis of gene expression stability of some housekeeping genes in 10 cell lines. In order to illustrate the utility of the GC to find suitable housekeeping genes, we next chose to assess this experimentally by RT-qPCR using a small subset of candidate reference genes (40; top 32 genes from genes ordered by GC and expression value from[94], plus 8 of the most commonly used from the literature, including seven from[66] and one (RPL32) from[126,127], and 10 cell lines from a range of tissues (see Tables 5 and 6). We first set a Cq value (which is inversely proportional to expression level) cut-off of 32, above which no expression is observed, and subsequently used the Cq values of genes in cell lines as a relative expression level (Cq cut off/Cq value of gene). Descriptive statistics of the expression of each gene in individual cell lines were then calculated. As a final step, the median expression value of each gene in individual cell lines was used to calculate descriptive statistics, including the GC, of gene expression across these cell lines. Figure 11 illustrates a KNIME workflow[128–130] that we wrote for this purpose. The raw data and descriptive statistics extracted are provided in Supplementary Tables S5 and S6 respectively, and the KNIME analysis workflow in Supplementary File 1.

Figure 12 uses RT-qPCR data to plot the GC of the candidate reference genes analysed here versus their relative median expression level. Three GiniGenes[94] (RBM45, TRNT1 and CNOT2) had very low and variable expression. Most of the other genes analysed showed low GC values with a range of (relative) expression values; the inset in Fig. 12 shows genes with a GC < 0.2 including a mix of 35 genes: 26 GiniGenes and 6 housekeeping genes referenced by Vandesompele and colleagues[66], one referenced by Caracausi[65] and one by Lee *et al.*[131]. Two of these GiniGenes, HNRNPK and PCBP1, which we also found to be stably expressed in the cell line data suggesting these may be potential stable housekeeping genes. As shown in Fig. 13 and inset, the GC is well correlated with the % RSD.

More importantly, the GC of our GiniGenes was particularly low (Fig. 12). The low absolute magnitude reflected the fact that Cq value is based on a logarithmic scale. Various commonly used housekeeping genes (HPRT1, GAPDH, ACTB, SDHA, HMBS and B2M) displayed higher % RSDs and GC than other genes studied here in spite of their higher relative expression levels. This was also the case when inspecting the interquartile ratio against the GC of these (Fig. S3).

The above results suggest that the GC is also applicable to RT-qPCR data, with GiniGenes having good potential (as novel "housekeeping" genes) for the normalisation of such data.

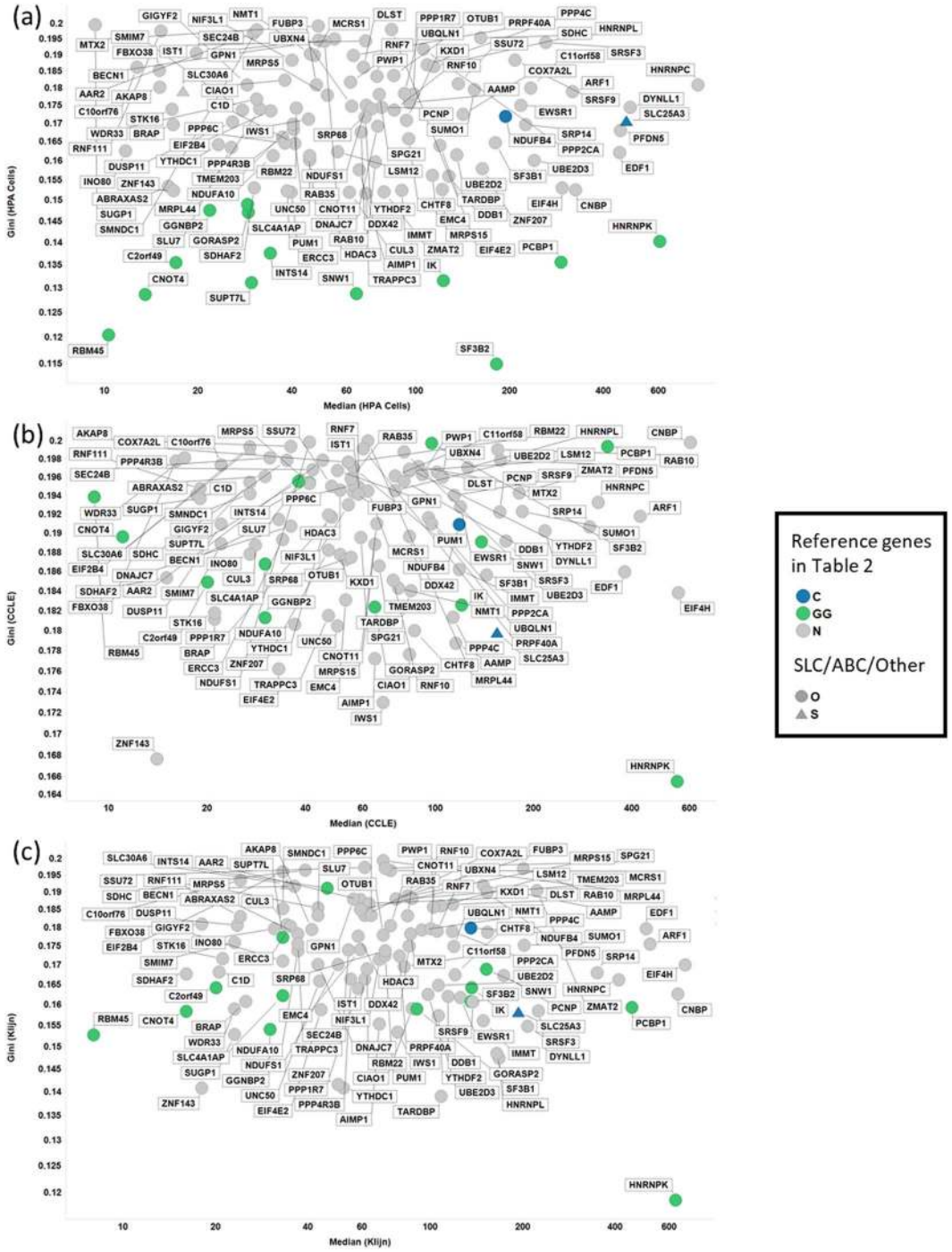

Figure 6. GC vs Median for 115 genes in. (**A)** HPA, (**B**). CCLE and **C**. Klijn/Genentech cell-line data sets. Colour coding: Blue, Caracausi; Green, GeneGini reference genes; Grey, neither. Shape coding: Circle, other; Triangle, SLC coding gene.

Discussion

Reference genes are commonly used to normalise gene expression data, so as to account for bias resulting from both biological and technical variability, and to enable quantification of gene expression changes or differences in the system under study. It is generally considered that such reference genes should come from pathways that are required for general metabolism, using only one gene per 'pathway' to avoid co-regulation which might make the gene expressions look very stable.

Such reference genes are commonly referred to as 'housekeeping' genes (HKGs) because they are considered to participate in essential cellular functions, are ubiquitously expressed in all cells and tissue types, and their expression is considered to be stable[49–66]). A number of such genes have been proposed over the years, and genes such as GAPDH, ACTB, RPL13A, SDHA, B2M are frequently used in such studies[66]. However, the

Gene	Gini (HPA Cells)	Gini (CCLE)	Gini (Klijn)	Median (HPA Cells)	Median (CCLE)	Median (Klijn)	RSD (HPA Cells)	RSD (CCLE)	RSD (Klijn)	GeneGini (GG)/ GeNorm (V), Eisenberg (EL), Lee (L), Caracausi © /N (Cell Line Data sets)	Reference	S/A/O	Protein name	Uniprot ID	Role
ARF1	0.18	0.19	0.18	316.70	423.00	517.00	32.54	35.87	35.17	N	N	O	ADP-ribosylation factor 1	P84077	Essential and ubiquitous GTP-binding protein regulators of vesicular trafficking and actin remodeling.
CNBP	0.15	0.20	0.16	324.24	602.00	637.50	28.47	37.37	29.49	N	N	O	Cellular nucleic acid-binding protein	P62633	Zinc finger protein, function unclear (Pellizzoni et al. 1997), regulates protein translation and transcription (Wei 2018)
DYNLL1	0.17	0.19	0.16	485.97	215.50	224.00	30.73	34.50	28.50	N	N	O	Dynein light chain 1, cytoplasmic	P63167	Component of dynein involved in intracellular transport and motility
EDF1	0.16	0.19	0.18	449.42	379.00	502.50	29.69	33.83	34.30	N	N	O	Endothelial differentiation-related factor 1	O60869	Modulates transcription of genes involved in endothelial differentiation, also acts as a transcriptional coactivator (Cazzaniga 2018)
EIF4H	0.15	0.18	0.17	294.21	553.50	673.00	27.91	33.27	30.64	N	N	O	Eukaryotic translation initiation factor 4H	Q15056	Translation initiation factor
HNRNPC	0.18	0.19	0.17	800.62	314.50	409.50	32.96	34.41	29.97	N	N	O	Heterogeneous nuclear ribonucleoproteins C1/C2	P07910	RNA binding protein involved in regulation of RNA splicing, export, expression, stability, and translation.
HNRNPK	0.14	0.17	0.12	603.32	548.00	625.50	25.19	29.60	21.35	GG	1	O	Heterogeneous nuclear ribonucleoprotein K	P61978	Regulation of RNA transcription and translation, splicing, nuclear export, and decay
PCBP1	0.14	0.20	0.16	291.40	336.00	452.00	24.52	36.23	29.01	GG	1	O	Poly(rC)-binding protein 1	Q15365	Regulation of mRNA transcription, translation and stability
PFDN5	0.17	0.20	0.19	451.20	158.00	152.50	31.60	41.30	35.69	N	N	O	Prefoldin subunit 5	Q99471	Molecular protein folding cytosolic chaperone. Prevents misfolding of newly synthesised nascent polypeptides
SF3B1	0.16	0.19	0.15	179.26	143.00	164.00	29.26	33.89	27.01	N	N	O	Splicing factor 3B subunit 1	O75533	Essential RNA-protein complex involved in pre-mRNA splicing
SLC25A3	0.17	0.18	0.16	471.21	154.00	193.00	30.19	32.86	28.39	C	65	S	Phosphate carrier protein, mitochondrial	Q00325	Phosphate transport from cytoplasm to mitochondria, with protons.
SRP14	0.17	0.19	0.17	224.37	296.00	347.50	30.36	34.77	30.62	N	N	O	Signal recognition particle 14 kDa protein	P37108	Signal-recognition-particle assembly has a crucial role in targeting secretory proteins to the rough endoplasmic reticulum membrane. Required for elongation arrest by binding with SRP9 to the Alu domain.
SRSF3	0.19	0.19	0.15	260.33	164.00	207.00	35.17	33.97	28.60	N	N	O	Serine/arginine-rich splicing factor 3	P84103	splicing factor that promotes exon inclusion during alternative splicing. Regulatory roles in RNA metabolism and functions such as mRNA splicing and 3'end processing. Essential for embryo development

Table 3. Descriptive statistics of 13 genes common across cell-line data sets with GC < 0.2. In addition, the protein name, as well as UniProt ID and function are shown. S/A/O refers to SLC, ABC or Other respectively.

expression levels of these and other proposed HKGs have in fact been shown to vary widely between cells and tissues (e.g.[53,62,65,67–79]) and their expression has also been reported to be affected by a number of factors relating to the experiment such as cell confluence[132], pathological, experimental and tissue specific conditions[133]. As highlighted by Huggett *et al.*[134], despite the reports of the potential variability of expression of 'classic' references genes such as GAPDH and ACTB, these are still used without mention of any validation processes. Our GiniGenes are selected as reference genes through different, data-driven, criteria.

Various tools have been developed to evaluate and screen reference genes from experimental datasets; these include geNorm[66], NormFinder[135], Best Keeper[136] and the comparative ΔCT finder[52]. RefFinder (http://leonxie.esy.es/RefFinder/#) and RefGenes (https://refgenes.org/rg//) can integrate these to enable a comparison and ranking of any tested candidate reference genes[137].

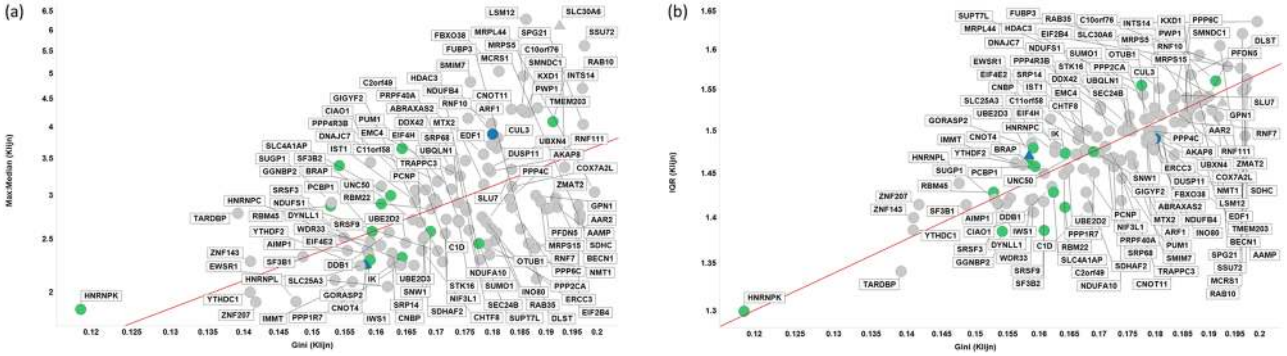

Figure 7. Robustness of GC for finding stably expressed genes using shared genes between HPA, CCLE and Klijn/Genentech cell-line data sets with GC < 0.2. Shown are the results for the Klijn/Genentech dataset. (**A**) IQR vs GC, (**B**). Max:Mean vs Min. Colour coding: Blue, Caracausi; Green, GeneGini reference genes; Grey, neither. Shape coding: Circle, other; Triangle, SLC coding gene.

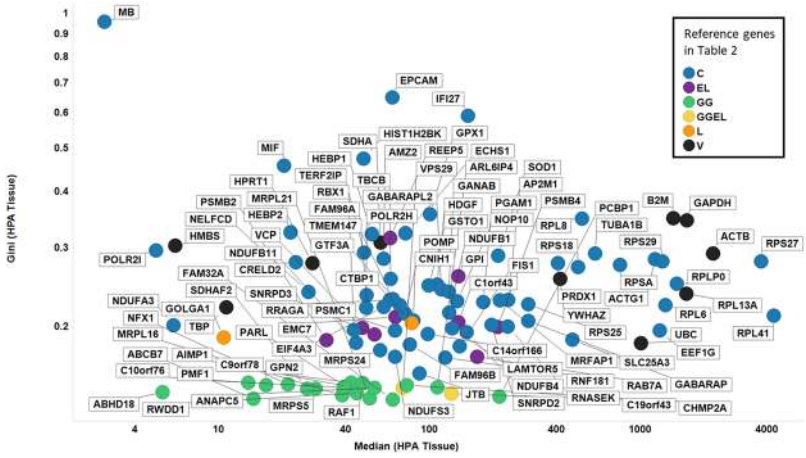

Figure 8. Gini coefficient and median expression levels of proposed reference genes in the HPA tissue dataset. Colour coding: blue, Caracausi; purple, Eisenberg and Levanon; green, GeneGini reference genes; yellow, both GeneGini and Eisenberg and Levanon; orange, Lee; black, Vandesompele.

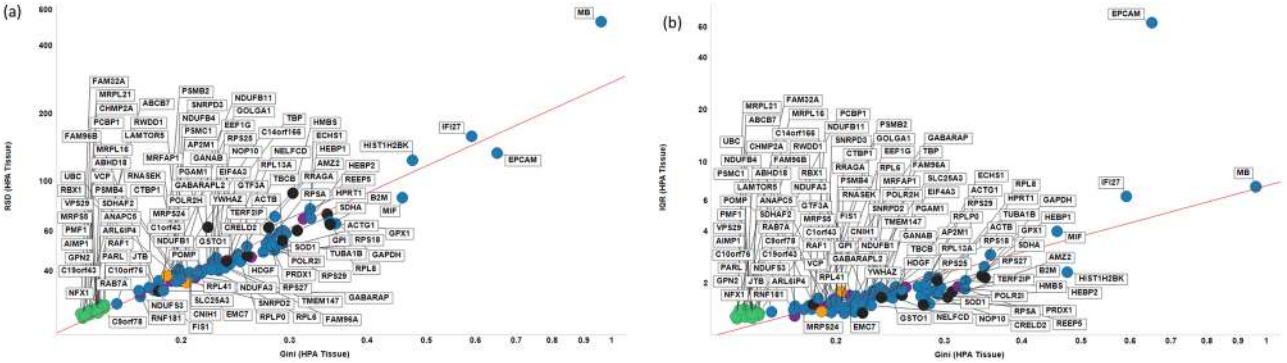

Figure 9. Robustness of the Gini coefficient in the HPA tissue data set. (**A**) RSD versus Gini coefficient of candidate reference genes. Line of best linear fit (in log space) shown is $y = 2.45 + 1.24 \times (r^2 = 0.938)$ (**B**). IQR versus Gini coefficient of candidate reference genes. Line of best linear fit (in log space) shown is $y = 0.87 + 0.96 \times (r^2 = 0.566)$. Colour code as in Fig. 8.

Gene	Gini (HPA Tissue)	Gini (GTEx)	Gini (PCAWG)	Gini (HBM)	Median (HPA Tissue)	Median (GTEx)	Median (PCAWG)	Median (HBM)	% RSD (HPA Tissue)	% RSD (GTEx)	% RSD (PCAWG)	% RSD (HBM)	Protein name	UniProt ID	Function (UniProt)
CHCHD4	0.19	0.14	0.19	0.19	13.08	17	20	25.69	36.11	27.26	40.35	35.44	Mitochondrial intermembrane space import and assembly protein 40	Q8N4Q1	Functions as a chaperone and catalyses formation of disulfide bonds in substrate proteins such as COX17, COX19 and MICU1. Required for import of small cysteine-containing proteins in the mitochondrial intermembrane space.
COPS5	0.17	0.17	0.2	0.17	45.27	19	20	20	32.46	30.76	43.24	33.27	COP9 signalosome complex subunit 5 (SGN5)	Q92905	Probable protease subunit of the COP9 signalosome complex (CSN), a complex involved in various cellular and developmental processes. The CSN complex is an essential regulator of the ubiquitin (Ubl) conjugation pathway by mediating the deneddylation of the cullin subunits of the SCF-type E3 ligase complexes, leading to decrease the Ubl ligase activity of SCF-type complexes such as SCF, CSA or DDB2. The complex is also involved in phosphorylation of p53/TP53, c-jun/JUN, IkappaBalpha/NFKBIA, ITPK1 and IRF8, possibly via its association with CK2 and PKD kinases. CSN-dependent phosphorylation of TP53 and JUN promotes and protects degradation by the Ubl system, respectively. In the complex, it probably acts as the catalytic center that mediates the cleavage of Nedd8 from cullins. It however has no metalloprotease activity by itself and requires the other subunits of the CSN complex. Interacts directly with a large number of proteins that are regulated by the CSN complex, confirming a key role in the complex. Promotes the proteasomal degradation of BRSK2.
COX4I1	0.17	0.12	0.18	0.16	447.69	123	144	94.13	33.09	22.96	37.11	28.92	Cytochrome c oxidase subunit 4 isoform 1, mitochondrial	P13073	This protein is one of the nuclear-coded polypeptide chains of cytochrome c oxidase, the terminal oxidase in mitochondrial electron transport.
IDH3G	0.16	0.18	0.17	0.18	44.67	56	60	34.75	28.6	31.58	33.02	32.45	Isocitrate dehydrogenase [NAD] subunit gamma, mitochondrial	P51553	Regulatory subunit which plays a role in the allosteric regulation of the enzyme catalyzing the decarboxylation of isocitrate (ICT) into alpha-ketoglutarate. The heterodimer composed of the alpha (IDH3A) and beta (IDH3B) subunits and the heterodimer composed of the alpha (IDH3A) and gamma (IDH3G) subunits, have considerable basal activity but the full activity of the heterotetramer (containing two subunits of IDH3A, one of IDH3B and one of IDH3G) requires the assembly and cooperative function of both heterodimers.
MAP2K2	0.2	0.17	0.18	0.17	60.91	55	58.5	30.75	36.87	31.05	34.07	31.65	Dual specificity mitogen-activated protein kinase kinase 2 (MAP kinase kinase 2) (MAPKK 2) (EC 2.7.12.2)	P36507	Catalyzes the concomitant phosphorylation of a threonine and a tyrosine residue in a Thr-Glu-Tyr sequence located in MAP kinases. Activates the ERK1 and ERK2 MAP kinases (By similarity).
MTIF3	0.18	0.17	0.19	0.19	45.15	51	55	72.63	33.81	30.61	37.88	37.82	Translation initiation factor IF-3, mitochondrial (IF-3(Mt))	Q9H2K0	IF-3 binds to the 28 S ribosomal subunit and shifts the equilibrium between 55 S ribosomes and their 39 S and 28 S subunits in favor of the free subunits, thus enhancing the availability of 28 S subunits on which protein synthesis initiation begins.
MTRF1L	0.17	0.19	0.19	0.14	7.86	11	17	17	31.84	33.29	34.42	28.57	Peptide chain release factor 1-like, mitochondrial	Q9UGC7	Mitochondrial peptide chain release factor that directs the termination of translation in response to the peptide chain termination codons UAA and UAG.
NDUFB8	0.16	0.16	0.18	0.19	143.5	39	37.5	56.63	30.35	28.76	33.91	35.07	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 8, mitochondrial	O95169	Accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I), that is believed not to be involved in catalysis. Complex I functions in the transfer of electrons from NADH to the respiratory chain. The immediate electron acceptor for the enzyme is believed to be ubiquinone.
NMT1	0.2	0.2	0.18	0.16	29.71	46	51.5	39.94	36.83	35.09	34.69	29.15	Glycylpeptide N-tetradecanoyl-transferase 1 (EC 2.3.1.97)	P30419	Enzyme catalysing transfer of myristate from CoA to proteins. Required for full expression of the biological activiteies of several N-myristoylated proteins, including the alpha subunit of the signal-transducing guanine nucleotide-binding protein (G protein) GO (GNAO1; MIM 139311)
PPID	0.16	0.17	0.17	0.19	31.11	29	32.5	44.75	29.02	32.72	34.11	33.73	Peptidyl-prolyl cis-trans isomerase D (PPIase D) (EC 5.2.1.8)	Q08752	Catalyze the cis-trans isomerization of proline imidic peptide bonds in oligopeptides and accelerate the folding of proteins. This protein has been shown to possess PPIase activity and, similar to other family members, can bind to the immunosuppressant cyclosporin A.

Continued

Gene	Gini (HPA Tissue)	Gini (GTEx)	Gini (PCAWG)	Gini (HBM)	Median (HPA Tissue)	Median (GTEx)	Median (PCAWG)	Median (HBM)	% RSD (HPA Tissue)	% RSD (GTEx)	% RSD (PCAWG)	% RSD (HBM)	Protein name	UniProt ID	Function (UniProt)
RTCA	0.17	0.18	0.2	0.18	26.5	24	27	33.69	30.67	35.82	42.89	33.68	RNA 3'-terminal phosphate cyclase (RNA cyclase)	O00442	Catalyzes the conversion of 3'-phosphate to a 2',3'-cyclic phosphodiester at the end of RNA. The mechanism of action of the enzyme occurs in 3 steps: (A) adenylation of the enzyme by ATP; (B) transfer of adenylate to an RNA-N3'P to produce RNA-N3'PP5'A; (C) and attack of the adjacent 2'-hydroxyl on the 3'-phosphorus in the diester linkage to produce the cyclic end product. The biological role of this enzyme is unknown but it is likely to function in some aspects of cellular RNA processing.
SELENOK	0.19	0.16	0.18	0.18	31.07	49	49	80.94	36.89	30.39	38.19	33.31	Selenoprotein K (SelK)	Q9Y6D0	Required for Ca2 + flux in immune cells and plays a role in T-cell proliferation and in T-cell and neutrophil migration (By similarity). Involved in endoplasmic reticulum-associated degradation (ERAD) of soluble glycosylated proteins (PubMed:22016385). Required for palmitoylation and cell surface expression of CD36 and involved in macrophage uptake of low-density lipoprotein and in foam cell formation (By similarity). Together with ZDHHC6, required for palmitoylation of ITPR1 in immune cells, leading to regulate ITPR1 stability and function (PubMed:25368151). Plays a role in protection of cells from ER stress-induced apoptosis (PubMed:20692228). Protects cells from oxidative stress when overexpressed in cardiomyocytes (PubMed:16962588).
SMG5	0.19	0.16	0.19	0.18	34.89	63	64	34.13	35.95	27.52	44.99	34.09	Protein SMG5 (EST1-like protein B)	Q9UPR3	Plays a role in nonsense-mediated mRNA decay. Does not have RNase activity by itself. Promotes dephosphorylation of UPF1. Together with SMG7 is thought to provide a link to the mRNA degradation machinery involving exonucleolytic pathways, and to serve as an adapter for UPF1 to protein phosphatase 2 A (PP2A), thereby triggering UPF1 dephosphorylation. Necessary for TERT activity.
SNX3	0.17	0.18	0.19	0.18	169.22	190	208.5	327.06	30.77	31.22	39.21	33.13	Sorting nexin-3 (Protein SDP3)	O60493	Phosphoinositide-binding protein required for multivesicular body formation. Specifically binds phosphatidylinositol 3-phosphate (PtdIns(P3)). Also can bind phosphatidylinositol 4-phosphate (PtdIns(P4)), phosphatidylinositol 5-phosphate (PtdIns(P5)) and phosphatidylinositol 3,5-biphosphate (PtdIns(3,5)P2) (By similarity). Plays a role in protein transport between cellular compartments. Together with RAB7A facilitates endosome membrane association of the retromer cargo-selective subcomplex (CSC/VPS). May in part act as component of the SNX3-retromer complex which mediates the retrograde endosome-to-TGN transport of WLS distinct from the SNX-BAR retromer pathway (PubMed:21725319, PubMed:24344282). Promotes stability and cell surface expression of epithelial sodium channel (ENAC) subunits SCNN1A and SCNN1G (By similarity). Not involved in EGFR degradation. Involved in the regulation of phagocytosis in dendritic cells possibly by regulating EEA1 recruitment to the nascent phagosomes (PubMed:23237080). Involved in iron homeostasis through regulation of endocytic recycling of the transferrin receptor TFRC presumably by delivering the transferrin:transferrin receptor complex to recycling endosomes; the function may involve the CSC retromer subcomplex (By similarity). In the case of Salmonella enterica infection plays arole in maturation of the Salmonella-containing vacuole (SCV) and promotes recruitment of LAMP1 to SCVs (PubMed:20482551).
SURF1	0.18	0.15	0.2	0.17	18.3	47	57.5	45.69	34.94	26.2	38.25	32.15	Surfeit locus protein 1	Q15526	Component of the MITRAC (mitochondrial translation regulation assembly intermediate of cytochrome c oxidase complex) complex, that regulates cytochrome c oxidase assembly.

Table 4. Descriptive statistics of 15 common genes across tissue data sets with a GC < 0.2. In addition, the protein name, as well as UniProt ID and function are shown.

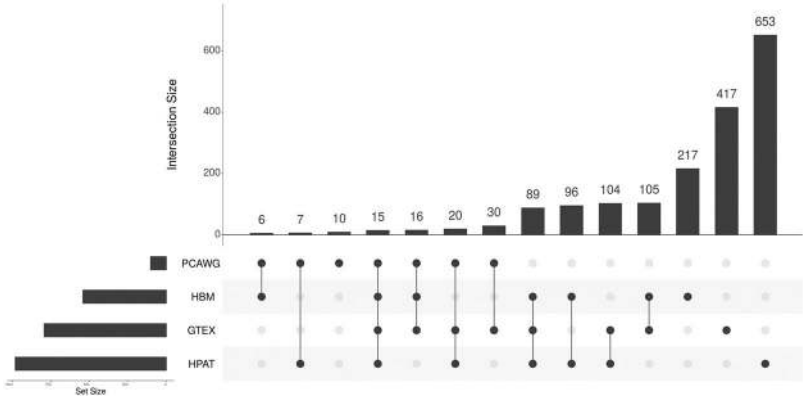

Figure 10. UpSetR[139] plot showing genes with a GC < 0.2 that are variously shared and unique across the PCAWG, HBM, GTEX and HPA tissue data sets. The data underpinning this plot can be found in Supplementary Table S4.

These tools assess expression stability of genes in different ways:

- geNorm determines gene stability through a stepwise exclusion or ranking process followed by averaging the geometric mean of the most stable genes from a chosen set. Python implementation: https://eleven.readthedocs.io/en/latest/.
- BestKeeper also uses the geometric mean but using raw data rather than copy numbers. BestKeeper[136] can be used as an Excel-based tool. It can accommodate up to 10 housekeeping genes in up to 100 biological samples. Optimal HKGs are determined by pairwise correlation analysis of all pairs of candidate genes, and the geometric mean of the top-ranking ones. http://www.gene-quantification.info.
- NormFinder measures variation, and ranks potential reference genes between study groups. NormFinder[135] has an add-in for Microsoft Excel and is available as an R programme. It recommends analysis of 5–10 candidate genes and at least 8 samples per group. https://moma.dk/normfinder-software.
- The comparative ΔCT finder requires no specialist programmes since this involves comparison of comparisons of ΔCTs between pairs of genes to find a set of genes that show least variability.
- RefGenes allows one to find genes that are stably expressed across tissue types and experimental conditions based on microarray data, and a comparison of results from geNorm, NormFinder and Best Keeper to find a set of reference genes. However, this is not a free service unless one searches for one gene at a time. Furthermore, the site for this tool is no longer available. Moreover, all these tools require the user to make a prior selection of such HKGs (introducing bias and potential errors) and most are cumbersome to understand and calculate.

We have here shown how via a simple calculation, the GC, we can find potential reference genes, and illustrated its utility in large-scale cell-line, tissue RNA-Seq data sets and RT-qPCR data. The expression of a number of classical HKGs from a number of carefully selected publications do in fact vary much more substantially between large RNA-Seq data sets, both for tissues and cell lines.

Whilst not all studies will involve large data sets such as those we have analysed here; the GC should also be of use for smaller-scale studies to select a subset of genes in a panel of cell lines or tissues relevant to the study in question.

Overall we find that (i) two of these genes, HNRNPK and PCBP1, seemed to be particularly robustly and stably expressed at reasonable levels in all cell lines studied, and (ii) a data-driven strategy based on the GC represents a useful and convenient method for normalisation in gene expression profiling and related studies.

Methods

The datasets used are described and referenced below. The data, in transcripts per million (TPM) units were downloaded from the EBI expression atlas as a .tsv file. As previously[1], the Gini Index was calculated using the **ineq** package (Achim Zeileis (2014). ineq: Measuring Inequality, Concentration, and Poverty. R package version 0.2–13. https://CRAN.R-project.org/package=ineq) in **R** (https://www.R-project.org/). These calculations were incorporated into KNIME via KNIME's R integration *R Snippet* node. A spreadsheet giving the extracted analyses is provided as Supplementary Tables (Tables S7 and S8).

Cell lines and culture conditions. A panel of 10 cell lines were grown in appropriate growth media: K562, PNT2 and T24 in RPMI-1640 (Sigma, Cat No. R7509), Panc1 and HEK293 in DMEM (Sigma, Cat No. D1145), SH-SY5Y in 1:1 mixture of DMEM/F12 (Gibco, Cat No. 21041025), J82 and RT-112 in EMEM (Gibco, Cat No. 51200–038), 5637 in Hyclone McCoy's (GE Healthcare, Cat No. SH30270.01) and PC3 in Ham's F12 (Biowest, Cat No. L0135-500). All growth media were supplemented with 10% fetal bovine serum (Sigma, Cat No. f4135) and 2 mM glutamine (Sigma, Cat No. G7513) without antibiotics. Cell cultures were maintained in T225 culture flasks (Star lab, CytoOne Cat No. CC7682-4225) kept in a 5% CO_2 incubator at 37 °C until 70–80% confluent.

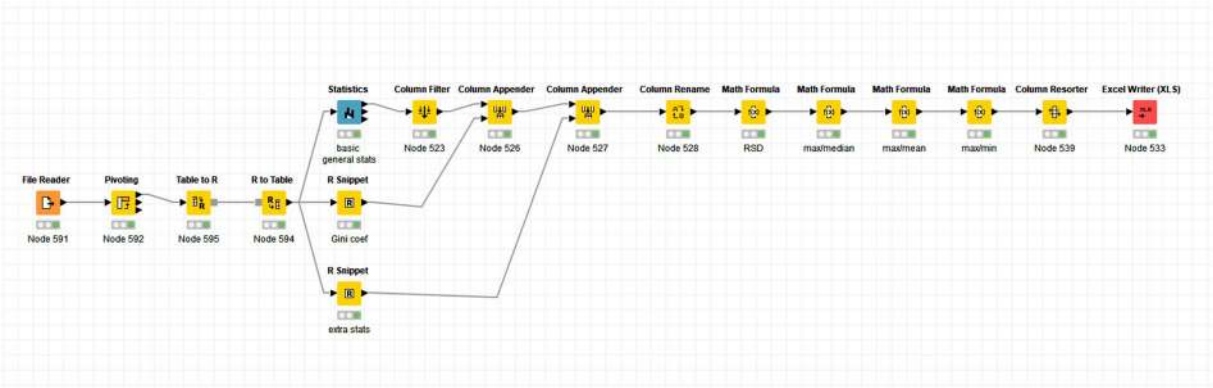

Figure 11. The KNIME workflow described here to calculate descriptive statistics and the Gini coefficient from RT-qPCR data. This workflow can be adapted for use with large RNA-Seq Data sets.

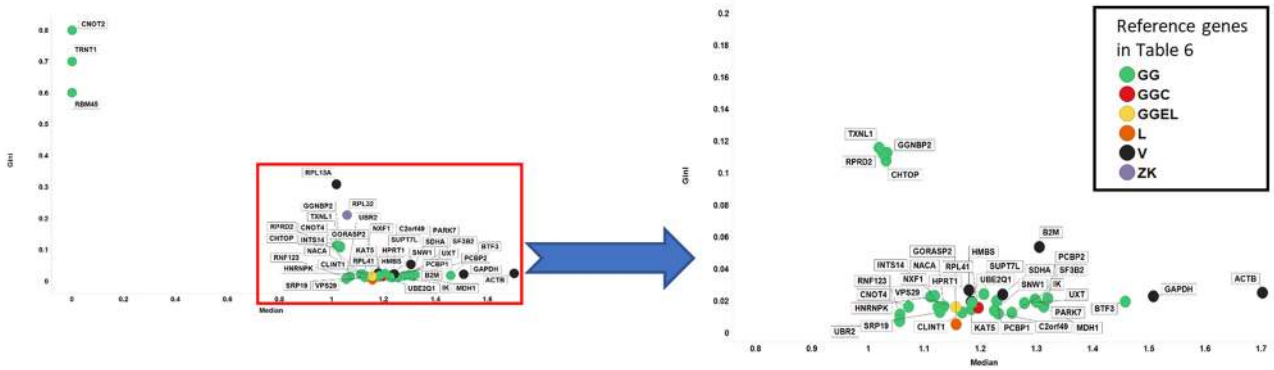

Figure 12. Gini coefficient and median expression levels of candidate reference genes assessed by RT-qPCR. Left panel shows all genes considered in this study, with right panel showing genes with GC < 0.2. Colour coding: green, GeneGini reference genes; red, both GeneGini and Caracausi reference genes; yellow, GeneGini and Eisenberg and Levanon; orange, Lee, yellow; black, Vandesompele; purple, Zhang and Kriegova.

Cell line	Tissue	Disease	Morphology	Growth mode	Media
K562	Blood	Chronic Myeloid Leukemia	Lymphoblast	Suspension	RPMI-1640
HEK293	Kidney	Immortalized cell line obtained by transfecting sheared adenovirus 5 DNA	Epithelial	Adherent	DMEM
Panc1	Pancreas	Pancreatic carcinoma of ductal origin	Epithelial	Adherent	DMEM
SH-SY5Y	Neuroblastoma	metastasis	Neuroblast	Adherent	DMEM
T24	Bladder	bladder carcinoma	Epithelial	Adherent	McCoy's 5A
J82	Bladder	Transitional cell carcinoma	Epithelial	Adherent	EMEM
RT-112	Bladder	Carcinoma	Epithelial	Adherent	RPMI-1640
5637	Bladder	Grade II carcinoma	Epithelial	Adherent	RPMI-1640
PC3	Prostate	Grade IV adenocarcinoma	Epithelial	Adherent	Ham's F12
PNT2	Prostate	Immortalized with SV40	Epithelial	Adherent	RPMI-1640

Table 5. Details of human cell lines used for the assessment of expression of candidate reference genes by RT-qPCR.

Gene Name	Uniprot	Gini (HPA Cell Lines)	GeneGini (GG)/GeNorm (V), Eisenberg & Levanon (EL), Lee (L), Caracausi (C), Zhang & Kriegova (ZK)	S/A/O	Reference
ACTB	P60709	0.26	V	O	66
B2M	P61769	0.44	V	O	66
BTF3	P20290	0.15	GG	O	1
C2orf49	Q9I8G4	0.14	GG	O	1
CHTOP	Q9Y3Y2	0.14	GG	O	1
CLINT1	Q14677	0.14	GG	O	1
CNOT2	Q9NZN8	0.14	GG	O	1
CNOT4	O95628	0.13	GG	O	1
GAPDH	P04406	0.27	V	O	66
GGNBP2	Q5SV77	0.15	GG	O	1
GORASP2	Q9H8Y8	0.15	GG	O	1
HMBS	P08397	0.26	V	O	66
HNRNPK	P04637	0.14	GG	O	1
HPRT1	P00492	0.31	V	O	66
IK	Q13123	0.13	GG	O	1
INTS14	Q96SY0	0.14	GG	O	1
KAT5	Q92993	0.13	GG	O	1
MDH1	P40925	0.15	GG	O	1
NACA	Q13765	0.15	GG	O	1
NXF1	Q9UBU9	0.12	GG	O	1
PARK7	Q99497	0.14	GG	O	1
PCBP1	Q15365	0.14	GG	O	1
PCBP2	Q15366	0.14	GG	O	1
RBM45	Q8IUH3	0.12	GG	O	1
RNF123	Q5XPI4	0.15	GG	O	1
RPL13A	P40429	0.21	V	O	66
RPL32	P62910	0.22	ZK	O	126,127
RPL41	P62945	0.15	GGC	O	1,65
RPRD2	Q5VT52	0.14	GG	O	1
SDHA	P31040	0.29	V	O	66
SF3B2	Q13435	0.11	GG	O	1
SNW1	Q13573	0.13	GG	O	1
SRP19	P09132	0.14	GG	O	1
SUPT7L	O94864	0.13	GG	O	1
TRNT1	Q96Q11	0.15	GG	O	1
TXNL1	O43396	0.14	GG	O	1
UBE2Q1	Q7Z7E8	0.14	GG	O	1
UBR2	Q8IWV8	0.14	GG	O	1
UXT	Q9UBK9	0.13	GG	O	1
VPS29	Q9UBQ0	0.15	GGEL	O	1,49

Table 6. Candidate reference genes used to assess expression stability experimentally by RT-qPCR. Included are gene name and UniProt ID, Gini coefficient as calculated using the HPA cell-line data set. S/A/O refers to SLC, ABC or Other respectively.

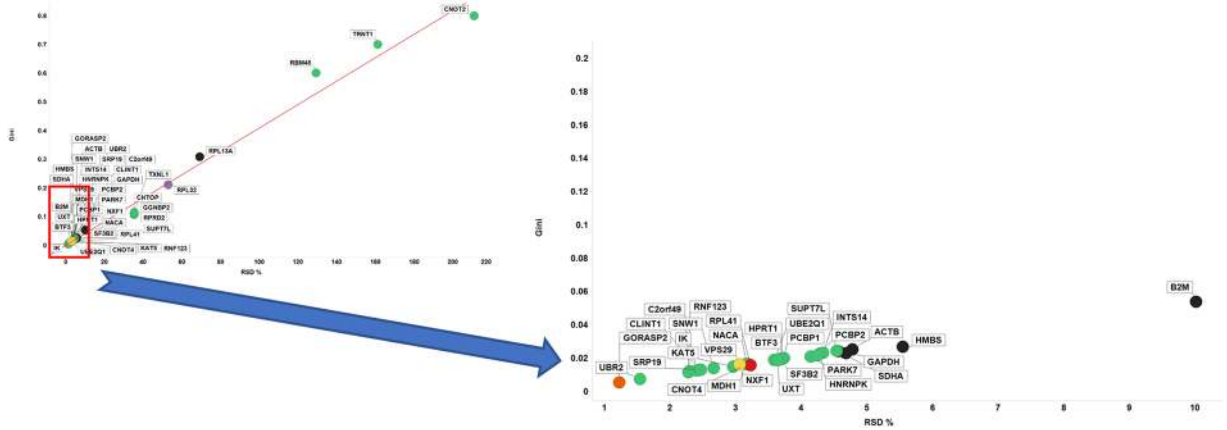

Figure 13. Robustness of the Gini coefficient in assessed experimentally by RT-qPCR using a small subset of proposed reference genes. Left panel shows Gini coefficient vs % RSD for all genes considered in this study, with right panel showing the same with genes with a GC < 0.2 and % RSD <10. Line of best linear fit shown is $y = 0.002 + 0.004 \times (r^2 = 0.988)$. Shape coding as in Fig. 12.

Harvesting Cells for RNA Extraction. Cells from adherent cell lines were harvested by removing growth media and washing twice with 5 mL of pre-warmed phosphate buffered saline (PBS) (Sigma, Cat No. D8537), then incubated in 3 mL of 0.025% trypsin-EDTA solution (Sigma Cat No. T4049) for 2–5 min at 37 °C. At the end of incubation cells were resuspended in 5–7 mL of respective media when cells appeared detached to dilute trypsin treatment. The cell suspension was transferred to 15 mL centrifuge tubes and immediately centrifuged at 300 × g for 5 min. Suspended cell lines were centrifuged directly from cultures in 50 mL centrifuge tubes and washed with PBS as above. The cell pellets were resuspended in 10–15 mL media and cell count and viability was determined using a Nexcellom Cellometer Auto 1000 Cell Viability Counter (Nexcellom Bioscience) set for Trypan Blue membrane exclusion method. Cells with >95% viability were used for downstream total RNA extraction.

RNA Extraction. Total RNA was extracted from 2–5 × 10^6 cells using the Qiagen RNeasy Mini Kit (Cat No. 74104) and DNAse treated using Turbo DNA-free kit (Invitrogen, Cat No. AM1907) according to the manufacturer's instructions. Briefly, 1 X DNA buffer was added to the extracted RNA prior to adding 2U (1 µL) of DNAse enzyme. The reaction mixture was incubated at 37 °C for 30 min and inactivated for 2 min at room temperature using DNAse inactivating reagent. The mixture was centrifuged at 10,000 × g for 1.5 min and the RNA from the supernatant was transferred to a clean tube. The RNA concentration was determined using a NanoDrop® ND-1000 spectrophotometer and further validated using an Agilent 2100 bio-analyser coupled with 2100 Expert software system. Only RNA samples with an RIN (RNA Integrity Number) between 9–10 were selected for cDNA synthesis.

Reverse Transcription and cDNA Synthesis. 1 µg of RNA was reverse transcribed into cDNA. Briefly, a 20 µL reaction was setup by adding 1 µL each of oligodT (50 µM, Invitrogen, cat No. 18418020) and dNTP mix (10 mM, Invitrogen, Cat No. 18427-013) followed by adding an appropriate volume for 1 µg of RNA. Nuclease free water (Ambion, Cat No. AM9937) was then added to make the volume up to 13 µL and incubated at 65 °C for 5 min then cooled on ice for 1 min. To initiate transcription 4 µL of 5 X first strand buffer (Invitrogen, Cat No. 1889832) and 1 µL each of 0.1 M DTT (Invitrogen, Cat No. 1907572), RNaseOUT™ (Invitrogen, Recombinant RNase Inhibitor, Cat No. 1905432) and SuperScript™ III RT (200 units/µL, Invitrogen, Cat No. 1685475) reverse transcriptase enzyme were added, mixed gently then incubated at 50 °C for 60 min followed by inactivation at 70 °C for 15 min. The cDNA was diluted 1:100 to be used in RT-qPCR experiment.

Validation of gene expression by geNorm. A set of candidate reference genes (40; top 32 genes from genes ordered by GC and expression value from[94], plus 8 of the most commonly used from the literature including seven from[66]). RNAseq data were selected for validation of stable gene expression using geNorm[66]. First, a typical qPCR protocol was prepared from a master mix for each gene to be tested per cell line in triplicate. This consisted of 10 µL/well made by adding 0.8 µL of nuclease free water (Ambion), 5 µL of LC480 SYBR Green I Master (2 X conc. Roche, Product No. 04887352001), 0.1 µL each of forward and reverse primers (20 µM) (for primer and amplicon sequences see Supplementary Table S9) and 4 µL of 1:100 diluted cDNA in a 384 well qPCR plate (Starlab Cat. No. E1042-9909-C). The no template controls (NTC) for each gene were produced by replacing cDNA with 4 µL of nuclease free water. Thermal cycling conditions used were: one cycle of 95 °C for 10 min followed by 40 cycles of 95 °C for 10 sec and 60 °C for 30 sec. qPCR was performed using Roche LightCycler LC480 qPCR platform. The fluorescence signals were measured in real time during amplification cycle (Cq) and also during temperature transition for melt curve analysis.

The mean Cq values were converted into relative values for a gene across all cell lines using ΔCq method[138]. Briefly, the lowest Cq value in a panel of cell lines for a gene was subtracted from all the values in that panel using

the equation: $R = 2^{(C_{q_{sample}} - C_{q_{control}})}$, where $C_{q_{sample}}$ is the mean Cq value obtained for a gene in each of the cell lines and $C_{q_{control}}$ is the lowest Cq value in that panel. The relative values for each gene in a panel were then obtained by applying $R = 2^{-\Delta C_q}$. These relative values were applied in geNorm Visual Basic applet for Microsoft Excel®[66] that determines the most stable reference genes from a set of genes in a given panel of cell lines.

Validation of gene expression using the Gini coefficient. To the raw RT-qPCR data a Cq value (which is inversely proportional to expression level) cut-off of 32 was set, above which no expression is observed. The Cq values of genes in cell lines were subsequently converted to a relative expression level (Cq cut off/Cq value of gene). Descriptive statistics of the expression of each gene in individual cell lines were then calculated. As a final step, the median expression value of each gene in individual cell lines was used to calculate descriptive statistics, including the GC, of gene expression across these cell lines. Figure 11 illustrates a KNIME workflow[128–130] for this purpose. The raw data and descriptive statistics extracted are provided in Supplementary Tables S5 and S6 respectively, and the KMNIME analysis workflow in Supplementary File 1.

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information Files). The original datasets used are referenced throughout and are summarised in Table 2.

Received: 8 August 2019; Accepted: 8 November 2019;
Published online: 29 November 2019

References

1. O'Hagan, S., Wright Muelas, M., Day, P. J., Lundberg, E. & Kell, D. B. GeneGini: assessment via the Gini coefficient of reference "housekeeping" genes and diverse human transporter expression profiles. *Cell systems* **6**, 230–244, https://doi.org/10.1016/j.cels.2018.01.003 (2018).
2. Gini, C. Concentration and dependency ratios (in Italian). English translation in: Rivista di Politica. *Economica* **87**(1997), 769–789 (1909).
3. Gini, C. *Variabilità e Mutabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. (C. Cuppini, 1912).
4. Ceriani, L. & Verme, P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J Econ Inequal* **10**, 421–443, https://doi.org/10.1007/s10888-011-9188-x (2012).
5. Jiang, L., Tsoucas, D. & Yuan, G. C. Assessing Inequality in Transcriptomic Data. *Cell systems* **6**, 149–150, https://doi.org/10.1016/j.cels.2018.02.007 (2018).
6. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* **131**, 281–285, https://doi.org/10.1007/s12064-012-0162-3 (2012).
7. Wilkinson, R. & Pickett, K. *The spirit level: why equality is better for everyone*. (Penguin Books, 2009).
8. Kondo, N. *et al*. Income inequality and health: the role of population size, inequality threshold, period effects and lag effects. *J Epidemiol Community Health* **66**, e11, https://doi.org/10.1136/jech-2011-200321 (2012).
9. Pickett, K. E. & Wilkinson, R. G. Income inequality and health: a causal review. *Soc Sci Med* **128**, 316–326, https://doi.org/10.1016/j.socscimed.2014.12.031 (2015).
10. Darkwah, K. A., Nortey, E. N. & Lotsi, A. Estimation of the Gini coefficient for the lognormal distribution of income using the Lorenz curve. *Springerplus* **5**, 1196, https://doi.org/10.1186/s40064-016-2868-z (2016).
11. Kohler, T. A. *et al*. Greater post-Neolithic wealth disparities in Eurasia than in North America and Mesoamerica. *Nature* **551**, 619–622, https://doi.org/10.1038/nature24646 (2017).
12. Nishi, A., Shirado, H., Rand, D. G. & Christakis, N. A. Inequality and visibility of wealth in experimental social networks. *Nature* **526**, 426–429, https://doi.org/10.1038/nature15392 (2015).
13. Damgaard, C. & Weiner, J. Describing inequality in plant size or fecundity. *Ecology* **81**, 1139–1142, 10.1890/0012-9658(2000)081[1139:Diipso]2.0.Co;2 (2000).
14. Sadras, V. & Bongiovanni, R. Use of Lorenz curves and Gini coefficients to assess yield inequality within paddocks. *Field Crops Res* **90**, 303–310, https://doi.org/10.1016/j.fcr.2004.04.003 (2004).
15. Weidlich, I. E. & Filippov, I. V. Using the gini coefficient to measure the chemical diversity of small-molecule libraries. *J Comput Chem* **37**, 2091–2097, https://doi.org/10.1002/jcc.24423 (2016).
16. Wren, J. D. Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades. *Bioinformatics* **32**, 2686–2691, https://doi.org/10.1093/bioinformatics/btw284 (2016).
17. LEE, W.-C. Analysis of Seasonal Data Using the Lorenz Curve and the Associated Gini Index. *International Journal of Epidemiology* **25**, 426–434, https://doi.org/10.1093/ije/25.2.426 (1996).
18. Lee, W.-C. Characterizing Exposure–Disease Association in Human Populations Using the Lorenz Curve and Gini Index. *Statistics in Medicine* **16**, 729–739, 10.1002/(SICI)1097-0258(19970415)16:7<729::AID-SIM491>3.0.CO;2-A (1997).
19. Lee, W.-C. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine* **18**, 455–471, 10.1002/(SICI)1097-0258(19990228)18:4<455::AID-SIM44>3.0.CO;2-A (1999).
20. Ainali, C. *et al*. Transcriptome classification reveals molecular subtypes in psoriasis. *BMC Genomics* **13**, 472, https://doi.org/10.1186/1471-2164-13-472 (2012).
21. Tran, Q. N. Improving the Accuracy of Gene Expression Profile Classification with Lorenz Curves and Gini Ratios. *Software Tools and Algorithms for Biological Systems* **696**, 83–90, https://doi.org/10.1007/978-1-4419-7046-6_9 (2011).
22. Jiang, L., Chen, H., Pinello, L. & Yuan, G. C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol* **17**, 144, https://doi.org/10.1186/s13059-016-1010-4 (2016).
23. Torre, E. *et al*. A comparison between single cell RNA sequencing and single molecule RNA FISH for rare cell analysis. *bioRxiv*, 138289, https://doi.org/10.1101/138289 (2017).
24. Shaffer, S. M. *et al*. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435, https://doi.org/10.1038/nature22794 (2017).
25. Torre, E. *et al*. Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. *Cell systems* **6**, 171–179 e175, https://doi.org/10.1016/j.cels.2018.01.014 (2018).
26. Schena, M. *et al*. Parallel human genome analysis - microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci.* **93**, 10614–10619 (1996).
27. Spellman, P. T. *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
28. Schena, M. *et al*. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* **16**, 301–306 (1998).
29. Hoyle, D. C., Rattray, M., Jupp, R. & Brass, A. Making sense of microarray data distributions. *Bioinformatics* **18**, 576–584 (2002).

30. Quackenbush, J. Microarray data normalization and transformation. *Nat Genet* **32**(Suppl), 496–501, https://doi.org/10.1038/ng1032 (2002).
31. Knight, C. G. *et al*. Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape. *Nucleic Acids Res* **37**, e6 (2009).
32. Walsh, C. J., Hu, P., Batt, J. & Santos, C. C. Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays (Basel)* **4**, 389–406, https://doi.org/10.3390/microarrays4030389 (2015).
33. Do, J. H. & Choi, D. K. Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol Cells* **22**, 254–261 (2006).
34. Steinhoff, C. & Vingron, M. Normalization and quantification of differential expression in gene expression microarrays. *Brief Bioinform* **7**, 166–177, https://doi.org/10.1093/bib/bbl002 (2006).
35. Dabney, A. R. & Storey, J. D. A new approach to intensity-dependent normalization of two-channel microarrays. *Biostatistics* **8**, 128–139, https://doi.org/10.1093/biostatistics/kxj038 (2007).
36. Kreil, D. P. & Russell, R. R. There is no silver bullet–a guide to low-level data transforms and normalisation methods for microarray data. *Brief Bioinform* **6**, 86–97 (2005).
37. Rahman, M. *et al*. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* **31**, 3666–3672, https://doi.org/10.1093/bioinformatics/btv377 (2015).
38. Lin, Y. *et al*. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* **17**, 28, https://doi.org/10.1186/s12864-015-2353-z (2016).
39. Li, X. *et al*. A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. *PLoS One* **12**, e0176185, https://doi.org/10.1371/journal.pone.0176185 (2017).
40. Dunn, W. B. *et al*. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* **6**, 1060–1083 (2011).
41. Zelena, E. *et al*. Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Analytical chemistry* **81**, 1357–1364 (2009).
42. Heckmann, L. H., Sørensen, P. B., Krogh, P. H. & Sørensen, J. G. NORMA-Gene: a simple and robust method for qPCR normalization based on target gene data. *BMC Bioinformatics* **12**, 250, https://doi.org/10.1186/1471-2105-12-250 (2011).
43. Hruz, T. *et al*. RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics* **12**, 156, https://doi.org/10.1186/1471-2164-12-156 (2011).
44. Khanna, P., Johnson, K. L. & Maron, J. L. Optimal reference genes for RT-qPCR normalization in the newborn. *Biotech Histochem*, 1–8, https://doi.org/10.1080/10520295.2017.1362474 (2017).
45. Ling, D. & Salvaterra, P. M. Robust RT-qPCR data normalization: validation and selection of internal reference genes during post-experimental data analysis. *PLoS One* **6**, e17762, https://doi.org/10.1371/journal.pone.0017762 (2011).
46. Sang, J. *et al*. ICG: a wiki-driven knowledgebase of internal control genes for RT-qPCR normalization. *Nucleic Acids Res*, https://doi.org/10.1093/nar/gkx875 (2017).
47. Vanhauwaert, S. *et al*. RT-qPCR gene expression analysis in zebrafish: Preanalytical precautions and use of expressed repetitive elements for normalization. *Methods Cell Biol* **135**, 329–342, https://doi.org/10.1016/bs.mcb.2016.02.002 (2016).
48. Kell, D. B. & Oliver, S. G. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26**, 99–105 (2004).
49. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569–574, https://doi.org/10.1016/j.tig.2013.05.010 (2013).
50. Hoerndli, F. J., Toigo, M., Schild, A., Götz, J. & Day, P. J. Reference genes identified in SH-SY5Y cells using custom-made gene arrays with validation by quantitative polymerase chain reaction. *Anal Biochem* **335**, 30–41 (2004).
51. Ohl, F. *et al*. Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization? *J Mol Med (Berl)* **83**, 1014–1024, https://doi.org/10.1007/s00109-005-0703-z (2005).
52. Silver, N., Best, S., Jiang, J. & Thein, S. L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol Biol* **7**, 33, https://doi.org/10.1186/1471-2199-7-33 (2006).
53. de Jonge, H. J. M. *et al*. Evidence based selection of housekeeping genes. *PLoS One* **2**, e898, https://doi.org/10.1371/journal.pone.0000898 (2007).
54. Tatsumi, K. *et al*. Reference gene selection for real-time RT-PCR in regenerating mouse livers. *Biochem Biophys Res Commun* **374**, 106–110, https://doi.org/10.1016/j.bbrc.2008.06.103 (2008).
55. Bustin, S. A. *et al*. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55**, 611–622, https://doi.org/10.1373/clinchem.2008.112797 (2009).
56. Gur-Dedeoglu, B. *et al*. Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. *Oncol Res* **17**, 353–365 (2009).
57. Li, Y. L., Ye, F., Hu, Y., Lu, W. G. & Xie, X. Identification of suitable reference genes for gene expression studies of human serous ovarian cancer by real-time polymerase chain reaction. *Anal Biochem* **394**, 110–116, https://doi.org/10.1016/j.ab.2009.07.022 (2009).
58. Thellin, O., ElMoualij, B., Heinen, E. & Zorzi, W. A decade of improvements in quantification of gene expression and internal standard selection. *Biotechnol Adv* **27**, 323–333 (2009).
59. Chervoneva, I. *et al*. Selection of optimal reference genes for normalization in quantitative RT-PCR. *BMC Bioinformatics* **11**, 253, https://doi.org/10.1186/1471-2105-11-253 (2010).
60. Wang, F., Wang, J., Liu, D. & Su, Y. Normalizing genes for real-time polymerase chain reaction in epithelial and nonepithelial cells of mouse small intestine. *Anal Biochem* **399**, 211–217, https://doi.org/10.1016/j.ab.2009.12.029 (2010).
61. Zampieri, M. *et al*. Validation of suitable internal control genes for expression studies in aging. *Mech Ageing Dev* **131**, 89–95, https://doi.org/10.1016/j.mad.2009.12.005 (2010).
62. Casadei, R. *et al*. Identification of housekeeping genes suitable for gene expression analysis in the zebrafish. *Gene Expr Patterns* **11**, 271–276, https://doi.org/10.1016/j.gep.2011.01.003 (2011).
63. Jacob, F. *et al*. Careful selection of reference genes is required for reliable performance of RT-qPCR in human normal and cancer cell lines. *PLoS One* **8**, e59180, https://doi.org/10.1371/journal.pone.0059180 (2013).
64. Oturai, D. B., Sondergaard, H. B., Bornsen, L., Sellebjerg, F. & Christensen, J. R. Identification of Suitable Reference Genes for Peripheral Blood Mononuclear Cell Subset Studies in Multiple Sclerosis. *Scand J Immunol* **83**, 72–80, https://doi.org/10.1111/sji.12391 (2016).
65. Caracausi, M. *et al*. Systematic identification of human housekeeping genes possibly useful as references in gene expression studies. *Mol Med Rep* **16**, 2397–2410, https://doi.org/10.3892/mmr.2017.6944 (2017).
66. Vandesompele, J. *et al*. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**, RESEARCH0034 (2002).
67. Butte, A. J., Dzau, V. J. & Glueck, S. B. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". *Physiol Genomics* **7**, 95–96 (2001).
68. Hsiao, L. L. *et al*. A compendium of gene expression in normal human tissues. *Physiol Genomics* **7**, 97–104, https://doi.org/10.1152/physiolgenomics.00040.2001 (2001).
69. Lee, P. D., Sladek, R., Greenwood, C. M. & Hudson, T. J. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. *Genome Res* **12**, 292–297, https://doi.org/10.1101/gr.217802 (2002).

70. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends Genet* **19**, 362–365, https://doi.org/10.1016/S0168-9525(03)00140-9 (2003).
71. Dheda, K. *et al*. Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *Biotechniques* **37**, 112–114, 116, 118–119 (2004).
72. Barber, R. D., Harmer, D. W., Coleman, R. A. & Clark, B. J. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genomics* **21**, 389–395, https://doi.org/10.1152/physiolgenomics.00025.2005 (2005).
73. Rubie, C. *et al*. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Mol Cell Probes* **19**, 101–109, https://doi.org/10.1016/j.mcp.2004.10.001 (2005).
74. Szabo, A. *et al*. Statistical modeling for selecting housekeeper genes. *Genome Biol* **5**, R59, https://doi.org/10.1186/gb-2004-5-8-r59 (2004).
75. Mane, V. P., Heuer, M. A., Hillyer, P., Navarro, M. B. & Rabin, R. L. Systematic method for determining an ideal housekeeping gene for real-time PCR analysis. *J Biomol Tech* **19**, 342–347 (2008).
76. Teste, M. A., Duquenne, M., François, J. M. & Parrou, J. L. Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*. *BMC Mol Biol* **10**, 99, https://doi.org/10.1186/1471-2199-10-99 (2009).
77. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25, https://doi.org/10.1186/gb-2010-11-3-r25 (2010).
78. Kozera, B. & Rapacz, M. Reference genes in real-time PCR. *J Appl Genet* **54**, 391–406, https://doi.org/10.1007/s13353-013-0173-x (2013).
79. De Spiegelaere, W. *et al*. Reference gene validation for RT-qPCR, a note on different available software packages. *PLoS One* **10**, e0122515, https://doi.org/10.1371/journal.pone.0122515 (2015).
80. Papatheodorou, I. *et al*. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res* **46**, D246–D251, https://doi.org/10.1093/nar/gkx1158 (2018).
81. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628, https://doi.org/10.1038/nmeth.1226 (2008).
82. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63, https://doi.org/10.1038/nrg2484 (2009).
83. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biol* **11**, 220, https://doi.org/10.1186/gb-2010-11-12-220 (2010).
84. Xu, J. *et al*. Comprehensive Assessments of RNA-seq by the SEQC Consortium: FDA-Led Efforts Advance Precision Medicine. *Pharmaceutics* **8**, https://doi.org/10.3390/pharmaceutics8010008 (2016).
85. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–527, https://doi.org/10.1038/nbt.3519 (2016).
86. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
87. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092, https://doi.org/10.1093/bioinformatics/bts094 (2012).
88. Tang, F. *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382, https://doi.org/10.1038/nmeth.1315 (2009).
89. Macosko, E. Z. *et al*. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214, https://doi.org/10.1016/j.cell.2015.05.002 (2015).
90. Rissin, D. M. & Walt, D. R. Digital concentration readout of single enzyme molecules using femtoliter arrays and Poisson statistics. *Nano Lett* **6**, 520–523, https://doi.org/10.1021/nl060227d (2006).
91. Salehi-Reyhani, A. *et al*. Scaling advantages and constraints in miniaturized capture assays for single cell protein analysis. *Lab Chip* **13**, 2066–2074, https://doi.org/10.1039/c3lc41388h (2013).
92. Hudecova, I. Digital PCR analysis of circulating nucleic acids. *Clin Biochem* **48**, 948–956, https://doi.org/10.1016/j.clinbiochem.2015.03.015 (2015).
93. Thul, P. J. *et al*. A subcellular map of the human proteome. *Science* **356**, https://doi.org/10.1126/science.aal3321 (2017).
94. Wu, Y. *et al*. Function of HNRNPC in breast cancer cells by controlling the dsRNA-induced interferon response. *The EMBO Journal* **37**, e99017, https://doi.org/10.15252/embj.201899017 (2018).
95. Bomsztyk, K., Denisenko, O. & Ostrowski, J. hnRNP K: One protein multiple processes. *BioEssays* **26**, 629–638, https://doi.org/10.1002/bies.20048 (2004).
96. Makeyev, A. V. & Liebhaber, S. A. The poly (C)-binding proteins: a multiplicity of functions and a search for mechanisms. *Rna* **8**, 265–278 (2002).
97. Huo, L.-R. & Zhong, N. Identification of transcripts and translatants targeted by overexpressed PCBP1. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1784**, 1524–1533 (2008).
98. Cho, S.-J., Jung, Y.-S. & Chen, X. Poly (C)-binding protein 1 regulates p63 expression through mRNA stability. *PloS one* **8**, e71724–e71724, https://doi.org/10.1371/journal.pone.0071724 (2013).
99. Lardelli, R. M., Thompson, J. X., Yates, J. R. & Stevens, S. W. Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *Rna* (2010).
100. Kfir, N. *et al*. SF3B1 Association with Chromatin Determines Splicing Outcomes. *Cell Reports* **11**, 618–629, https://doi.org/10.1016/j.celrep.2015.03.048 (2015).
101. Effenberger, K. A., Urabe, V. K., Prichard, B. E., Ghosh, A. K. & Jurica, M. S. Interchangeable SF3B1 inhibitors interfere with pre-mRNA splicing at multiple stages. *RNA* **22**, 350–359, https://doi.org/10.1261/rna.053108.115 (2016).
102. He, X. & Zhang, P. Serine/arginine-rich splicing factor 3 (SRSF3) regulates homologous recombination-mediated DNA repair. *Molecular Cancer* **14**, 158, https://doi.org/10.1186/s12943-015-0422-1 (2015).
103. Gallardo, M. *et al*. hnRNP K Is a Haploinsufficient Tumor Suppressor that Regulates Proliferation and Differentiation Programs in Hematologic Malignancies. *Cancer Cell* **28**, 486–499, https://doi.org/10.1016/j.ccell.2015.09.001 (2015).
104. Barboro, P. *et al*. Heterogeneous nuclear ribonucleoprotein K: altered pattern of expression associated with diagnosis and prognosis of prostate cancer. *British Journal Of Cancer* **100**, 1608, https://doi.org/10.1038/sj.bjc.6605057 (2009).
105. Park, Y. M. *et al*. Heterogeneous Nuclear Ribonucleoprotein C1/C2 Controls the Metastatic Potential of Glioblastoma by Regulating PDCD4. *Molecular and Cellular Biology* **32**, 4237, https://doi.org/10.1128/MCB.00443-12 (2012).
106. Lee, E. K. *et al*. hnRNP C promotes APP translation by competing with FMRP for APP mRNA recruitment to P bodies. *Nature structural & molecular biology* **17**, 732–739, https://doi.org/10.1038/nsmb.1815 (2010).
107. Zarnack, K. *et al*. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell* **152**, 453–466, https://doi.org/10.1016/j.cell.2012.12.023 (2013).
108. Wang, H. *et al*. PCBP1 Suppresses the Translation of Metastasis-Associated PRL-3 Phosphatase. *Cancer Cell* **18**, 52–62, https://doi.org/10.1016/j.ccr.2010.04.028 (2010).
109. Zhang, T. *et al*. PCBP-1 regulates alternative splicing of the CD44 gene and inhibits invasion in human hepatoma cell line HepG2 cells. *Molecular Cancer* **9**, 72, https://doi.org/10.1186/1476-4598-9-72 (2010).
110. Liu, Y. *et al*. Expression of poly(C)-binding protein 1 (PCBP1) in NSCLC as a negative regulator of EMT and its clinical value. *International journal of clinical and experimental pathology* **8**, 7165–7172 (2015).

111. Zhang, Z.-Z. *et al*. HOTAIR Long Noncoding RNA Promotes Gastric Cancer Metastasis through Suppression of Poly r(C)-Binding Protein (PCBP) 1. *Molecular Cancer Therapeutics* **14**, 1162, https://doi.org/10.1158/1535-7163.MCT-14-0695 (2015).
112. Wagener, R. *et al*. The PCBP1 gene encoding poly(rc) binding protein i is recurrently mutated in Burkitt lymphoma. *Genes, Chromosomes and Cancer* **54**, 555–564, https://doi.org/10.1002/gcc.22268 (2015).
113. Ji, F.-J. *et al*. Expression of both poly r(C) binding protein 1 (PCBP1) and miRNA-3978 is suppressed in peritoneal gastric cancer metastasis. *Scientific reports* **7**, 15488–15488, https://doi.org/10.1038/s41598-017-15448-9 (2017).
114. Jumaa, H., Wei, G. & Nielsen, P. J. Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20. *Current Biology* **9**, 899–902, https://doi.org/10.1016/S0960-9822(99)80394-7 (1999).
115. Palmieri, F. The mitochondrial transporter family SLC25: Identification, properties and physiopathology. *Mol Aspects Med* **34**, 465–484, https://doi.org/10.1016/j.mam.2012.05.005 (2013).
116. Schnabel, M. *et al*. Dedifferentiation-associated changes in morphology and gene expression in primary human articular chondrocytes in cell culture. *Osteoarthritis and Cartilage* **10**, 62–70, https://doi.org/10.1053/joca.2001.0482 (2002).
117. Cullen, P. J. Endosomal sorting and signalling: an emerging role for sorting nexins. *Nature Reviews Molecular Cell Biology* **9**, 574, https://doi.org/10.1038/nrm2427 (2008).
118. Naslavsky, N. & Caplan, S. The enigmatic endosome – sorting the ins and outs of endocytic trafficking. *Journal of Cell Science* **131**, jcs216499, https://doi.org/10.1242/jcs.216499 (2018).
119. Chen, C. *et al*. Snx3 Regulates Recycling of the Transferrin Receptor and Iron Assimilation. *Cell Metabolism* **17**, 343–352, https://doi.org/10.1016/j.cmet.2013.01.013 (2013).
120. Xu, S., Nigam, S. M. & Brodin, L. Overexpression of SNX3 Decreases Amyloid-β Peptide Production by Reducing Internalization of Amyloid Precursor Protein. *Neurodegenerative Diseases* **18**, 26–37, https://doi.org/10.1159/000486199 (2018).
121. Binder, N. K., Sheedy, J. R., Hannan, N. J. & Gardner, D. K. Male obesity is associated with changed spermatozoa Cox4i1 mRNA level and altered seminal vesicle fluid composition in a mouse model. *MHR: Basic science of reproductive medicine* **21**, 424–434, https://doi.org/10.1093/molehr/gav010 (2015).
122. Li, Y., Park, J.-S., Deng, J.-H. & Bai, Y. Cytochrome c oxidase subunit IV is essential for assembly and respiratory function of the enzyme complex. *Journal of Bioenergetics and Biomembranes* **38**, 283–291, https://doi.org/10.1007/s10863-006-9052-z (2006).
123. Storey, J. D. *et al*. Gene-Expression Variation Within and Among Human Populations. *The American Journal of Human Genetics* **80**, 502–509, https://doi.org/10.1086/512017 (2007).
124. Lonsdale, J. *et al*. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580, https://doi.org/10.1038/ng.2653, https://www.nature.com/articles/ng.2653#supplementary-information (2013).
125. Pickrell, J. K. *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768, https://doi.org/10.1038/nature08872, https://www.nature.com/articles/nature08872#supplementary-information (2010).
126. Zhang, X., Ding, L. & Sandford, A. J. Selection of reference genes for gene expression studies in human neutrophils by real-time PCR. *BMC Mol Biol.* **18**, 4 (2005).
127. Kriegova, E. *et al*. PSMB2 and RPL32 are suitable denominators to normalize gene expression profiles in bronchoalveolar cells. *BMC Mol Biol.* **31**, 69 (2008).
128. Mazanetz, M. P., Marmon, R. J., Reisser, C. B. T. & Morao, I. Drug discovery applications for KNIME: an open source data mining platform. *Curr Top Med Chem* **12**, 1965–1979, https://doi.org/10.2174/1568026611212180004 (2012).
129. Fillbrunn, A. *et al*. KNIME for reproducible cross-domain analysis of life science data. *J Biotechnol*, https://doi.org/10.1016/j.jbiotec.2017.07.028 (2017).
130. O'Hagan, S. & Kell, D. B. The KNIME workflow environment and its applications in Genetic Programming and machine learning. *Genetic Progr Evol Mach* **16**, 387–391, https://doi.org/10.1007/s10710-015-9247-3 (2015).
131. Lee, S., Jo, M., Lee, J., Koh, S. S. & Kim, S. Identification of novel universal housekeeping genes by statistical analysis of microarray data. *J Biochem Mol Biol* **40**, 226–231 (2007).
132. Greer, S., Honeywell, R., Geletu, M., Arulanandam, R. & Raptis, L. Housekeeping genes; expression levels may change with density of cultured cells. *Journal of Immunological Methods* **355**, 76–79, https://doi.org/10.1016/j.jim.2010.02.006 (2010).
133. Li, R. & Shen, Y. An old method facing a new challenge: Re-visiting housekeeping proteins as internal reference control for neuroscience research. *Life Sciences* **92**, 747–751, https://doi.org/10.1016/j.lfs.2013.02.014 (2013).
134. Huggett, J., Dheda, K., Bustin, S. & Zumla, A. Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun* **6**, 279–284, https://doi.org/10.1038/sj.gene.6364190 (2005).
135. Andersen, C. L., Jensen, J. L. & Orntoft, T. F. 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, https://doi.org/10.1158/0008-5472.CAN-04-0496 (2004).
136. Pfaffl, M. W., Tichopad, A., Prgomet, C. & Neuvians, T. P. 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).
137. Xie, F., Xiao, P., Chen, D., Xu, L. & Zhang, B. miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol Biol*, https://doi.org/10.1007/s11103-012-9885-2 (2012).
138. Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2-\Delta\Delta CT$ Method. *Methods* **25**, 402–408, https://doi.org/10.1006/meth.2001.1262 (2001).
139. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940, https://doi.org/10.1093/bioinformatics/btx364 (2017).
140. Uhlen, M. *et al*. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, https://doi.org/10.1126/science.1260419 (2015).
141. Barretina, J. *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607, https://doi.org/10.1038/nature11003 (2012).
142. Klijn, C. *et al*. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol* **33**, 306–312, https://doi.org/10.1038/nbt.3080 (2015).
143. Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660, https://doi.org/10.1126/science.1262110 (2015).

Acknowledgements

All authors thank the BBSRC (grant BB/P009042/1) and the Novo Nordisk Foundation (grant NNF10CC1016517) for financial support.

Author contributions

D.B.K. highlighted the utility of the G.C. as shown in reference 1. M.W.M. adapted the Gini method and analyses workflows developed by S.O. from reference 1 and performed most of the analyses that were done using KNIME. P.J.D. contributed in particular to the analysis of the housekeeping genes. F.M. performed the RT-qPCR analyses. All authors contributed to the writing and approval of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-54288-7.

Correspondence and requests for materials should be addressed to M.W.M., P.J.D. or D.B.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.