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

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

[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, [3]The Manchester Institute of Biotechnology, 131, Princess St, Manchester M1 7DN, UK. [4]Faculty of Biology, Medicine and Health, The University of Manchester M13 9PL, UK.

[5]Novo Nordisk Foundation Centre for Biosustainability, Technical University of Denmark, 10 Building 220, Kemitorvet, 2800 Kgs. Lyngby, Denmark

Emails in order of authorship: m.wright-muelas@liverpool.ac.uk, Farah.Mughal@liverpool.ac.uk, SOhagan@manchester.ac.uk , Philip.J.Day@manchester.ac.uk , dbk@liv.ac.uk

*corresponding authors: m.wright-muelas@liverpool.ac.uk and dbk@liv.ac.uk

Abstract

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.

Keywords: housekeeping genes – reference genes – Gini index – Gene Expression

Background

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 biochemistry [1, 5, 17-22]. Its visualisation and calculation are comparatively straightforward (Fig 1): individual examples are ranked on the 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 [23-25], 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. [26-36]). By this ('normalisation of the first kind'), and what is typically done, we mean the smoothing out of genuine artefacts within an arrray 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 [37, 38]). 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 [39-44], 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.

Perhaps surprisingly [45], 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. [46-63]). 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. [50, 59, 62, 64-76]). Indeed, Lee et al [66] 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 [63]. However, this analysis was done on only one (albeit large) dataset of gene expression profiles. While some of the compilations (e.g. [62, 77]) 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. [78-82]), in which individual transcripts are simply counted digitally via direct sequencing, is seen as considerably more robust [78, 83, 84] and sensitive [85, 86], 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 [87-89]), 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.

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 [90] with very low expression variability and thus potential for use as reference genes [91]. 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 [91] as well as other reference genes proposed in other studies (Table 2) with additional large RNA-Seq cell line data sets (Table *1*).

Fig 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 [90]. 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 Levenson 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).

Fig 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 and 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 and 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 ≤ 0.2 (Fig. 5, 6). This value for the GC was chosen since reducing this to ≤ 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 ≤ 0.15) and going above this meant the number of genes were unmanageable (e.g. 1051 genes with a GC ≤ 0.21 in the Klijn data set). Of the 115 genes shared between the datasets with GC <0.2, 13 were GiniGenes and two were housekeeping genes defined by Caracausi and colleagues (Fig. 5 B). 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 share important roles in RNA transcription, translation and stability [92-100], are implicated in a number of diseases, including cancer [92, 95, 101-111], and some, such as SRSF3 are essential for embryo development [112]. 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 [113]) 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 (Fig7: 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 [114].

In our previous report [1] we also analysed RNA-Seq data from tissues [90] 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 Figure 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 & Levenson [115]. 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. 9 A and 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 (Figure S1 A-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 (Figure 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.

In the case of the data set used by Eisenberg and Levanon [115], 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 (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 [116]. SNX3 has been reported to play a role in receptor recycling and formation of multivesicular bodies [117], and its dysregulation has been implicated in disorders of iron metabolism and the pathogenesis of some neurodegenerative diseases [118, 119].

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 [120] 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 [121].

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 [122]. Moreover, tissues are themselves a mixture of cell types with varying levels of gene expression in each cell type [123], 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 [122, 124].

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 [91], plus 8 of the most commonly used from the literature, including seven from [63] and one (RPL32) from [125][126], and 10 cell lines from a range of tissues (see Table 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 [127-129] that we wrote 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.

Fig 12 uses RT-qPCR data to plot the GC of the candidate reference genes analysed here versus their relative median expression level. Three GiniGenes [91] (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 Figure 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 [63], one referenced by Caracausi [130] 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 Figure 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 (Figure 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.

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 [46-63]). 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 [63]. However, the expression levels of these and other proposed HKGs have in fact been shown to vary widely between cells and tissues (e.g. [50, 59, 62, 64-76]) 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 [63], NormFinder [135], Best Keeper [136] and the comparative ΔCT finder [49]. RefFinder (http://leonxie.esy.es/RefFinder/#) and RefGenes can integrate these to enable a comparison and ranking of any tested candidate reference genes [137].

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).

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

Study short name	**Comments**	**Reference**
GiniGene	Study presenting novel potential housekeeping genes in cells and tissues from the HPA project cell and tissue RNASeq data.	[1]
geNorm or Vandesompele	Classic set of reference genes **in tissues** and a means of analysing them	[63]
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.	[46]
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.**	[62]

Table 2. Studies used for expression profiling data.

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).	[90, 91, 138]
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).	[139]
Klijn / Genentech	RNA-seq-based analysis of 57,711 genes in 622 human cancer **cell lines** (downloaded from EBI Expression Atlas E-MTAB-2706).	[140]
GTEx	RNA-Seq data of 46,711 genes in 53 human **tissue** samples from the	[141]

	Genotype-Tissue Expression (GTEx) project (downloaded from EBI Expression Atlas E-MTAB-5214).	
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).	Unpublished, may be subject to publication embargo until July 25th, 2019 https://dcc.icgc.org/pcawg
HBM	Illumina Body Map: RNA-seq of 16 Human **Tissues**. E-MTAB-513. Used by Eisenberg and colleagues in their analysis of housekeeping/ reference genes **in tissues**.	[46]

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.

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 x 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 X 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 x 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 1min. 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 [91], plus 8 of the most commonly used from the literature including seven from [63]). RNAseq data were selected for validation of stable gene expression using geNorm [63]. 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 [142]. 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® [63] 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 [127-129] 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.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The PCAWG data is under embargo until the WGS pan-cancer consortium publishes its marker paper or until July 25, 2019, whichever is earlier. Methodology papers may be published prior to this embargo, with agreement from the full scientific working group. We have been in email contact with jennifer.jennings@oicr.on.ca who asked that we advise the editor to wait until the July 25 embargo lift.

Availability of data and materials

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.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by BBSRC Project Grant BB/ P009042/1.

Authors' contributions

D.B.K. highlighted the utility of the GC as shown in [1]. M.W.M. adapted the Gini method and analyses workflows developed by S.O. from [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.

Acknowledgements

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

Legends to figures

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

Fig 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 X (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 X (r^2=0.593). Colour coding: red, GeneGini reference genes; blue Eisenberg & Levenson; yellow Vandesompele; green Lee; lilac both GeneGini and Eisenberg and Levenson.

Fig 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 + 0.829x (r^2 = 0.909). Colour code as in Fig 2.

Fig 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 X (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 X (r^2=0.930). **C.** Max:Mean vs Min expression levels in HPA data set. Colour code as in Fig 2.

Fig 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.

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

Fig. 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.

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

Fig 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 X (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 X (r^2=0.566). Colour code as in Fig 8.

Fig 10. UpSetR [143] 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 it Supplementary Table S4

Fig 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.

Fig 12. Gini coefficient and median expression levels of candidate reference genes assessed by RTqPCR. 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 Levenson; orange, Lee, yellow; black, Vandesompele; purple, Zhang and Kriegova.

Fig 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 + 0.004x (r2=0.988). Shape coding as in Fig 12.

Supplementary Fig S1. Comparison of median expression levels of proposed reference genes between tissue datasets. A. HBM vs HPA tissue datasets. Line of best linear fit (in log space) shown is $\log_{10}y$ = 0.35 +

(0.74 $\log_{10}$ (x)) (r2=0.472). B. PCAWG vs HPA tissue dataset. Line of best linear fit (in log space) shown is $\log_{10}$y = 0.46 + (0.73 $\log_{10}$ (x)) (r2=0.500). C. GTEx vs HPA Tissue. Line of best linear fit (in log space) shown is $\log_{10}$y = 0.45 + (0.68 $\log_{10}$(x)) (r2=0.429). Colour coding: blue, Caracausi reference genes; purple, Eisenberg & Levenson; green, GeneGini; yellow, both GeneGini and Eisenberg and Levenson; orange, Lee; black, Vandesompele.

Supplementary Fig S2. Comparison of Gini coefficient of proposed reference genes between tissue datasets. A. HBM vs HPA tissue datasets. Line of best linear fit (in log space) shown is log10y = -0.20 + (0.62 $\log_{10}$(x)) (r2=0.392). B. PCAWG vs HPA tissue dataset. Line of best linear fit (in log space) shown is $\log_{10}$y = -0.15 + (0.59 $\log_{10}$(x)) (r2=0.560). C. GTEx vs HPA Tissue. Line of best linear fit (in log space) shown is $\log_{10}$y = 0.22 + (0.59 $\log_{10}$(x)) (r2=0.388). Colour coding as in Fig S1.

Fig S3. Robustness of the Gini coefficient assessed experimentally by RT-qPCR using a small subset of proposed reference genes illustrated with Gini coefficient vs IQR. Left panel shows all 40 genes in Table 6, with right panel showing genes with a GC < 0.2. Colour coding: green, GeneGini reference genes; red, both GeneGini and Caracausi reference genes; yellow, GeneGini and Eisenberg and Levenson; orange, Lee, yellow; black, Vandesompele; purple, Zhang and Kriegova.

List of tables

Supplementary Table S5. Raw expression data for candidate reference genes in human cell lines by RT-qPCR.

Supplementary Table S6. Descriptive statistics and Gini coefficient data for candidate reference genes in human cell lines by RT-qPCR.

Supplementary Table S7. Extracted analyses of cell-line RNA-Seq data sets referenced in Table 2. S/A/O refers to SLC, ABC or Other respectively.

Supplementary Table S8. Extracted analyses of tissue RNA-Seq data sets referenced in Table 2. S/A/O refers to SLC, ABC or Other respectively.

Supplementary Table S9. Primer and amplicon sequences of candidate reference genes used to assess expression stability experimentally by RT-qPCR. Included are the Gini coefficient and median expression level as found in the HPA cell-line data set. S/A/O refers to SLC, ABC or Other respectively.

Supplementary Files

Supplementary File 1. KNIME workflow [127-129] that we have written to calculate descriptive statistics, including the GC, of gene expression across cell lines to assess of expression stability of candidate reference genes by RT-qPCR.

References

1. O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB: **GeneGini: assessment via the Gini coefficient of reference "housekeeping" genes and diverse human transporter expression profiles** *Cell Syst* 2018, **6**:230-244.
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.* Bologna: 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* 2012, **10**:421-443.
5. Jiang L, Tsoucas D, Yuan GC: **Assessing Inequality in Transcriptomic Data.** *Cell Syst* 2018, **6**:149-150.
6. Wagner GP, Kin K, Lynch VJ: **Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.** *Theory Biosci* 2012, **131**:281-285.
7. Wilkinson R, Pickett K: *The spirit level: why equality is better for everyone.* London: Penguin Books; 2009.
8. Kondo N, van Dam RM, Sembajwe G, Subramanian SV, Kawachi I, Yamagata Z: **Income inequality and health: the role of population size, inequality threshold, period effects and lag effects.** *J Epidemiol Community Health* 2012, **66**:e11.
9. Pickett KE, Wilkinson RG: **Income inequality and health: a causal review.** *Soc Sci Med* 2015, **128**:316-326.
10. Darkwah KA, Nortey EN, Lotsi A: **Estimation of the Gini coefficient for the lognormal distribution of income using the Lorenz curve.** *Springerplus* 2016, **5**:1196.
11. Kohler TA, Smith ME, Bogaard A, Feinman GM, Peterson CE, Betzenhauser A, Pailes M, Stone EC, Marie Prentiss A, Dennehy TJ, et al: **Greater post-Neolithic wealth disparities in Eurasia than in North America and Mesoamerica.** *Nature* 2017, **551**:619-622.
12. Nishi A, Shirado H, Rand DG, Christakis NA: **Inequality and visibility of wealth in experimental social networks.** *Nature* 2015, **526**:426-429.
13. Damgaard C, Weiner J: **Describing inequality in plant size or fecundity.** *Ecology* 2000, **81**:1139-1142.
14. Sadras V, Bongiovanni R: **Use of Lorenz curves and Gini coefficients to assess yield inequality within paddocks.** *Field Crops Res* 2004, **90**:303-310.
15. Weidlich IE, Filippov IV: **Using the gini coefficient to measure the chemical diversity of small-molecule libraries.** *J Comput Chem* 2016, **37**:2091-2097.
16. Wren JD: **Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades.** *Bioinformatics* 2016, **32**:2686-2691.
17. Ainali C, Valeyev N, Perera G, Williams A, Gudjonsson JE, Ouzounis CA, Nestle FO, Tsoka S: **Transcriptome classification reveals molecular subtypes in psoriasis.** *BMC Genomics* 2012, **13**:472.
18. Tran QN: **Improving the Accuracy of Gene Expression Profile Classification with Lorenz Curves and Gini Ratios.** *Software Tools and Algorithms for Biological Systems* 2011, **696**:83-90.
19. Jiang L, Chen H, Pinello L, Yuan GC: **GiniClust: detecting rare cell types from single-cell gene expression data with Gini index.** *Genome Biol* 2016, **17**:144.
20. Torre E, Dueck H, Shaffer S, Gospocic J, Gupte R, Bonasio R, Kim J, Murray J, Raj A: **A comparison between single cell RNA sequencing and single molecule RNA FISH for rare cell analysis.** *bioRxiv* 2017:138289.
21. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, Beqiri M, Sproesser K, Brafford PA, Xiao M, et al: **Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance.** *Nature* 2017, **546**:431-435.
22. Torre E, Dueck H, Shaffer S, Gospocic J, Gupte R, Bonasio R, Kim J, Murray J, Raj A: **Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH.** *Cell Syst* 2018, **6**:171-179 e175.

23. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: **Parallel human genome analysis - microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci* 1996, **93:**10614-10619.

24. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.

25. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW: **Microarrays: biotechnology's discovery platform for functional genomics.** *Trends Biotechnol* 1998, **16:**301-306.

26. Hoyle DC, Rattray M, Jupp R, Brass A: **Making sense of microarray data distributions.** *Bioinformatics* 2002, **18:**576-584.

27. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32 Suppl:**496-501.

28. Knight CG, Platt M, Rowe W, Wedge DC, Khan F, Day P, McShea A, Knowles J, Kell DB: **Array-based evolution of DNA aptamers allows modelling of an explicit sequence-fitness landscape.** *Nucleic Acids Res* 2009, **37:**e6.

29. Walsh CJ, Hu P, Batt J, Santos CC: **Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery.** *Microarrays (Basel)* 2015, **4:**389-406.

30. Do JH, Choi DK: **Normalization of microarray data: single-labeled and dual-labeled arrays.** *Mol Cells* 2006, **22:**254-261.

31. Steinhoff C, Vingron M: **Normalization and quantification of differential expression in gene expression microarrays.** *Brief Bioinform* 2006, **7:**166-177.

32. Dabney AR, Storey JD: **A new approach to intensity-dependent normalization of two-channel microarrays.** *Biostatistics* 2007, **8:**128-139.

33. Kreil DP, Russell RR: **There is no silver bullet--a guide to low-level data transforms and normalisation methods for microarray data.** *Brief Bioinform* 2005, **6:**86-97.

34. Rahman M, Jackson LK, Johnson WE, Li DY, Bild AH, Piccolo SR: **Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results.** *Bioinformatics* 2015, **31:**3666-3672.

35. Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YL, Sultana H, Oliver B, Harbison ST: **Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*.** *BMC Genomics* 2016, **17:**28.

36. Li X, Brock GN, Rouchka EC, Cooper NGF, Wu D, O'Toole TE, Gill RS, Eteleeb AM, O'Brien L, Rai SN: **A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data.** *PLoS One* 2017, **12:**e0176185.

37. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown N, Knowles J, Halsall A, Haselden JN, et al: **Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry.** *Nat Protoc* 2011, **6:**1060-1083.

38. Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, O'Hagan S, Knowles JD, Halsall A, HUSERMET Consortium, et al: **Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum.** *Anal Chem* 2009, **81:**1357-1364.

39. Heckmann LH, Sørensen PB, Krogh PH, Sørensen JG: **NORMA-Gene: a simple and robust method for qPCR normalization based on target gene data.** *BMC Bioinformatics* 2011, **12:**250.

40. Hruz T, Wyss M, Docquier M, Pfaffl MW, Masanetz S, Borghi L, Verbrugghe P, Kalaydjieva L, Bleuler S, Laule O, et al: **RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization.** *BMC Genomics* 2011, **12:**156.

41. Khanna P, Johnson KL, Maron JL: **Optimal reference genes for RT-qPCR normalization in the newborn.** *Biotech Histochem* 2017:1-8.

42. Ling D, Salvaterra PM: **Robust RT-qPCR data normalization: validation and selection of internal reference genes during post-experimental data analysis.** *PLoS One* 2011, **6:**e17762.

43. Sang J, Wang Z, Li M, Cao J, Niu G, Xia L, Zou D, Wang F, Xu X, Han X, et al: **ICG: a wiki-driven knowledgebase of internal control genes for RT-qPCR normalization.** *Nucleic Acids Res* 2017.

44. Vanhauwaert S, Lefever S, Coucke P, Speleman F, De Paepe A, Vandesompele J, Willaert A: **RT-qPCR gene expression analysis in zebrafish: Preanalytical precautions and use of expressed repetitive elements for normalization.** *Methods Cell Biol* 2016, **135:**329-342.

45. Kell DB, Oliver SG: **Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.** *Bioessays* 2004, **26:**99-105.

46. Eisenberg E, Levanon EY: **Human housekeeping genes, revisited.** *Trends Genet* 2013, **29:**569-574.

47. Hoerndli FJ, Toigo M, Schild A, Götz J, Day PJ: **Reference genes identified in SH-SY5Y cells using custom-made gene arrays with validation by quantitative polymerase chain reaction.** *Anal Biochem* 2004, **335:**30-41.

48. Ohl F, Jung M, Xu C, Stephan C, Rabien A, Burkhardt M, Nitsche A, Kristiansen G, Loening SA, Radonic A, Jung K: **Gene expression studies in prostate cancer tissue: which reference gene should be selected for normalization?** *J Mol Med (Berl)* 2005, **83:**1014-1024.

49. Silver N, Best S, Jiang J, Thein SL: **Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR.** *BMC Mol Biol* 2006, **7:**33.

50. de Jonge HJM, Fehrmann RSN, de Bont ESJM, Hofstra RMW, Gerbens F, Kamps WA, de Vries EGE, van der Zee AGJ, te Meerman GJ, ter Elst A: **Evidence based selection of housekeeping genes.** *PLoS One* 2007, **2:**e898.

51. Tatsumi K, Ohashi K, Taminishi S, Okano T, Yoshioka A, Shima M: **Reference gene selection for real-time RT-PCR in regenerating mouse livers.** *Biochem Biophys Res Commun* 2008, **374:**106-110.

52. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, et al: **The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.** *Clin Chem* 2009, **55:**611-622.

53. Gur-Dedeoglu B, Konu O, Bozkurt B, Ergul G, Seckin S, Yulug IG: **Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues.** *Oncol Res* 2009, **17:**353-365.

54. Li YL, Ye F, Hu Y, Lu WG, Xie X: **Identification of suitable reference genes for gene expression studies of human serous ovarian cancer by real-time polymerase chain reaction.** *Anal Biochem* 2009, **394:**110-116.

55. Thellin O, ElMoualij B, Heinen E, Zorzi W: **A decade of improvements in quantification of gene expression and internal standard selection.** *Biotechnol Adv* 2009, **27:**323-333.

56. Chervoneva I, Li Y, Schulz S, Croker S, Wilson C, Waldman SA, Hyslop T: **Selection of optimal reference genes for normalization in quantitative RT-PCR.** *BMC Bioinformatics* 2010, **11:**253.

57. 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* 2010, **399:**211-217.

58. Zampieri M, Ciccarone F, Guastafierro T, Bacalini MG, Calabrese R, Moreno-Villanueva M, Reale A, Chevanne M, Burkle A, Caiafa P: **Validation of suitable internal control genes for expression studies in aging.** *Mech Ageing Dev* 2010, **131:**89-95.

59. Casadei R, Pelleri MC, Vitale L, Facchin F, Lenzi L, Canaider S, Strippoli P, Frabetti F: **Identification of housekeeping genes suitable for gene expression analysis in the zebrafish.** *Gene Expr Patterns* 2011, **11:**271-276.

60. Jacob F, Guertler R, Naim S, Nixdorf S, Fedier A, Hacker NF, Heinzelmann-Schwarz V: **Careful selection of reference genes is required for reliable performance of RT-qPCR in human normal and cancer cell lines.** *PLoS One* 2013, **8:**e59180.

61. Oturai DB, Sondergaard HB, Bornsen L, Sellebjerg F, Christensen JR: **Identification of Suitable Reference Genes for Peripheral Blood Mononuclear Cell Subset Studies in Multiple Sclerosis.** *Scand J Immunol* 2016, **83:**72-80.

62. Caracausi M, Piovesan A, Antonaros F, Strippoli P, Vitale L, Pelleri MC: **Systematic identification of human housekeeping genes possibly useful as references in gene expression studies.** *Mol Med Rep* 2017, **16:**2397-2410.

63. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biol* 2002, **3:**RESEARCH0034.

64. Butte AJ, Dzau VJ, Glueck SB: **Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues".** *Physiol Genomics* 2001, **7:**95-96.

65. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, et al: **A compendium of gene expression in normal human tissues.** *Physiol Genomics* 2001, **7:**97-104.

66. Lee PD, Sladek R, Greenwood CM, Hudson TJ: **Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies.** *Genome Res* 2002, **12:**292-297.

67. Eisenberg E, Levanon EY: **Human housekeeping genes are compact.** *Trends Genet* 2003, **19:**362-365.

68. Dheda K, Huggett JF, Bustin SA, Johnson MA, Rook G, Zumla A: **Validation of housekeeping genes for normalizing RNA expression in real-time PCR.** *Biotechniques* 2004, **37:**112-114, 116, 118-119.

69. Barber RD, Harmer DW, Coleman RA, Clark BJ: **GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues.** *Physiol Genomics* 2005, **21:**389-395.

70. Rubie C, Kempf K, Hans J, Su T, Tilton B, Georg T, Brittner B, Ludwig B, Schilling M: **Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues.** *Mol Cell Probes* 2005, **19:**101-109.

71. Szabo A, Perou CM, Karaca M, Perreard L, Palais R, Quackenbush JF, Bernard PS: **Statistical modeling for selecting housekeeper genes.** *Genome Biol* 2004, **5:**R59.

72. Mane VP, Heuer MA, Hillyer P, Navarro MB, Rabin RL: **Systematic method for determining an ideal housekeeping gene for real-time PCR analysis.** *J Biomol Tech* 2008, **19:**342-347.

73. Teste MA, Duquenne M, François JM, Parrou JL: **Validation of reference genes for quantitative expression analysis by real-time RT-PCR in *Saccharomyces cerevisiae*.** *BMC Mol Biol* 2009, **10:**99.

74. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.

75. Kozera B, Rapacz M: **Reference genes in real-time PCR.** *J Appl Genet* 2013, **54:**391-406.

76. De Spiegelaere W, Dern-Wieloch J, Weigel R, Schumacher V, Schorle H, Nettersheim D, Bergmann M, Brehm R, Kliesch S, Vandekerckhove L, Fink C: **Reference gene validation for RT-qPCR, a note on different available software packages.** *PLoS One* 2015, **10:**e0122515.

77. Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, Burke M, Fullgrabe A, Fuentes AM, George N, et al: **Expression Atlas: gene and protein expression across multiple studies and organisms.** *Nucleic Acids Res* 2018, **46:**D246-D251.

78. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.

79. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57-63.

80. Oshlack A, Robinson MD, Young MD: **From RNA-seq reads to differential expression results.** *Genome Biol* 2010, **11:**220.

81. Xu J, Gong B, Wu L, Thakkar S, Hong H, Tong W: **Comprehensive Assessments of RNA-seq by the SEQC Consortium: FDA-Led Efforts Advance Precision Medicine.** *Pharmaceutics* 2016, **8.**

82. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol* 2016, **34:**525-527.

83. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29:**644-652.

84. Schulz MH, Zerbino DR, Vingron M, Birney E: **Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels.** *Bioinformatics* 2012, **28:**1086-1092.

85. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al: **mRNA-Seq whole-transcriptome analysis of a single cell.** *Nat Methods* 2009, **6:**377-382.

86. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al: **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.** *Cell* 2015, **161:**1202-1214.

87. Rissin DM, Walt DR: **Digital concentration readout of single enzyme molecules using femtoliter arrays and Poisson statistics.** *Nano Lett* 2006, **6:**520-523.

88. Salehi-Reyhani A, Sharma S, Burgin E, Barclay M, Cass A, Neil MAA, Ces O, Willison KR, Klug DR, Brown A, Novakova M: **Scaling advantages and constraints in miniaturized capture assays for single cell protein analysis.** *Lab Chip* 2013, **13:**2066-2074.

89. Hudecova I: **Digital PCR analysis of circulating nucleic acids.** *Clin Biochem* 2015, **48:**948-956.

90. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, et al: **A subcellular map of the human proteome.** *Science* 2017, **356**.

91. O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB: **GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles.** *Cell Syst* 2018, **6:**230-244 e231.

92. Wu Y, Zhao W, Liu Y, Tan X, Li X, Zou Q, Xiao Z, Xu H, Wang Y, Yang X: **Function of HNRNPC in breast cancer cells by controlling the dsRNA-induced interferon response.** *The EMBO Journal* 2018, **37:**e99017.

93. Bomsztyk K, Denisenko O, Ostrowski J: **hnRNP K: One protein multiple processes.** *BioEssays* 2004, **26:**629-638.

94. Makeyev AV, Liebhaber SA: **The poly (C)-binding proteins: a multiplicity of functions and a search for mechanisms.** *Rna* 2002, **8:**265-278.

95. Huo L-R, Zhong N: **Identification of transcripts and translatants targeted by overexpressed PCBP1.** *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 2008, **1784:**1524-1533.

96. Cho S-J, Jung Y-S, Chen X: **Poly (C)-binding protein 1 regulates p63 expression through mRNA stability.** *PloS one* 2013, **8:**e71724-e71724.

97. Lardelli RM, Thompson JX, Yates JR, Stevens SW: **Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing.** *Rna* 2010.

98. Kfir N, Lev-Maor G, Glaich O, Alajem A, Datta A, Sze Siu K, Meshorer E, Ast G: **SF3B1 Association with Chromatin Determines Splicing Outcomes.** *Cell Reports* 2015, **11:**618-629.

99. Effenberger KA, Urabe VK, Prichard BE, Ghosh AK, Jurica MS: **Interchangeable SF3B1 inhibitors interfere with pre-mRNA splicing at multiple stages.** *RNA* 2016, **22:**350-359.

100. He X, Zhang P: **Serine/arginine-rich splicing factor 3 (SRSF3) regulates homologous recombination-mediated DNA repair.** *Molecular Cancer* 2015, **14:**158.

101. Gallardo M, Lee Hun J, Zhang X, Bueso-Ramos C, Pageon Laura R, McArthur M, Multani A, Nazha A, Manshouri T, Parker-Thornburg J, et al: **hnRNP K Is a Haploinsufficient Tumor Suppressor that Regulates Proliferation and Differentiation Programs in Hematologic Malignancies.** *Cancer Cell* 2015, **28:**486-499.

102. Barboro P, Repaci E, Rubagotti A, Salvi S, Boccardo S, Spina B, Truini M, Introini C, Puppo P, Ferrari N, et al: **Heterogeneous nuclear ribonucleoprotein K: altered pattern of expression associated with diagnosis and prognosis of prostate cancer.** *British Journal Of Cancer* 2009, **100:**1608.

103. Park YM, Hwang SJ, Masuda K, Choi K-M, Jeong M-R, Nam D-H, Gorospe M, Kim HH: **Heterogeneous Nuclear Ribonucleoprotein C1/C2 Controls the Metastatic Potential of Glioblastoma by Regulating PDCD4.** *Molecular and Cellular Biology* 2012, **32:**4237.

104. Lee EK, Kim HH, Kuwano Y, Abdelmohsen K, Srikantan S, Subaran SS, Gleichmann M, Mughal MR, Martindale JL, Yang X, et al: **hnRNP C promotes APP translation by competing with FMRP for APP mRNA recruitment to P bodies.** *Nature structural & molecular biology* 2010, **17:**732-739.

105. Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe Nicholas M, Ule J: **Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements.** *Cell* 2013, **152:**453-466.

106. Wang H, Vardy LA, Tan CP, Loo JM, Guo K, Li J, Lim SG, Zhou J, Chng WJ, Ng SB, et al: **PCBP1 Suppresses the Translation of Metastasis-Associated PRL-3 Phosphatase.** *Cancer Cell* 2010, **18:**52-62.

107. Zhang T, Huang X-H, Dong L, Hu D, Ge C, Zhan Y-Q, Xu W-X, Yu M, Li W, Wang X, et al: **PCBP-1 regulates alternative splicing of the CD44 gene and inhibits invasion in human hepatoma cell line HepG2 cells.** *Molecular Cancer* 2010, **9:**72.

108. Liu Y, Gai L, Liu J, Cui Y, Zhang Y, Feng J: **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* 2015, **8:**7165-7172.

109. Zhang Z-Z, Shen Z-Y, Shen Y-Y, Zhao E-H, Wang M, Wang C-J, Cao H, Xu J: **HOTAIR Long Noncoding RNA Promotes Gastric Cancer Metastasis through Suppression of Poly r(C)-Binding Protein (PCBP) 1.** *Molecular Cancer Therapeutics* 2015, **14:**1162.

110. Wagener R, Aukema SM, Schlesner M, Haake A, Burkhardt B, Claviez A, Drexler HG, Hummel M, Kreuz M, Loeffler M, et al: **The PCBP1 gene encoding poly(rc) binding protein i is recurrently mutated in Burkitt lymphoma.** *Genes, Chromosomes and Cancer* 2015, **54:**555-564.

111. Ji F-J, Wu Y-Y, An Z, Liu X-S, Jiang J-N, Chen F-F, Fang X-D: **Expression of both poly r(C) binding protein 1 (PCBP1) and miRNA-3978 is suppressed in peritoneal gastric cancer metastasis.** *Scientific reports* 2017, **7:**15488-15488.

112. Jumaa H, Wei G, Nielsen PJ: **Blastocyst formation is blocked in mouse embryos lacking the splicing factor SRp20.** *Current Biology* 1999, **9:**899-902.

113. Palmieri F: **The mitochondrial transporter family SLC25: Identification, properties and physiopathology.** *Mol Asp Med* 2013, **34:**465-484.

114. Schnabel M, Marlovits S, Eckhoff G, Fichtel I, Gotzen L, Vécsei V, Schlegel J: **Dedifferentiation-associated changes in morphology and gene expression in primary human articular chondrocytes in cell culture.** *Osteoarthritis and Cartilage* 2002, **10:**62-70.

115. Eisenberg E, Levanon EY: **Human housekeeping genes, revisited.** *Trends in Genetics* 2013, **29:**569-574.

116. Cullen PJ: **Endosomal sorting and signalling: an emerging role for sorting nexins.** *Nature Reviews Molecular Cell Biology* 2008, **9:**574.

117. Naslavsky N, Caplan S: **The enigmatic endosome – sorting the ins and outs of endocytic trafficking.** *Journal of Cell Science* 2018, **131:**jcs216499.

118. Chen C, Garcia-Santos D, Ishikawa Y, Seguin A, Li L, Fegan Katherine H, Hildick-Smith Gordon J, Shah Dhvanit I, Cooney Jeffrey D, Chen W, et al: **Snx3 Regulates Recycling of the Transferrin Receptor and Iron Assimilation.** *Cell Metabolism* 2013, **17:**343-352.

119. Xu S, Nigam SM, Brodin L: **Overexpression of SNX3 Decreases Amyloid-β Peptide Production by Reducing Internalization of Amyloid Precursor Protein.** *Neurodegenerative Diseases* 2018, **18:**26-37.

120. Binder NK, Sheedy JR, Hannan NJ, Gardner DK: **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* 2015, **21:**424-434.

121. 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* 2006, **38:**283-291.

122. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM: **Gene-Expression Variation Within and Among Human Populations.** *The American Journal of Human Genetics* 2007, **80:**502-509.

123. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al: **The Genotype-Tissue Expression (GTEx) project.** *Nature Genetics* 2013, **45:**580.

124. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768.

125. Zhang X, Ding L, Sandford AJ: **Selection of reference genes for gene expression studies in human neutrophils by real-time PCR.** *BMC Mol Biol* 2005, **18:**4.

126. Kriegova E, Arakelyan A, Fillerova R, Zatloukal J, Mrazek F, Navratilova Z, Kolek V, du Bois RM, Petrek M.: **PSMB2 and RPL32 are suitable denominators to normalize gene expression profiles in bronchoalveolar cells.** *BMC Mol Biol* 2008, **31:**69.

127. Mazanetz MP, Marmon RJ, Reisser CBT, Morao I: **Drug discovery applications for KNIME: an open source data mining platform.** *Curr Top Med Chem* 2012, **12:**1965-1979.

128. Fillbrunn A, Dietz C, Pfeuffer J, Rahn R, Landrum GA, Berthold MR: **KNIME for reproducible cross-domain analysis of life science data.** *J Biotechnol* 2017.

129. O'Hagan S, Kell DB: **The KNIME workflow environment and its applications in Genetic Programming and machine learning.** *Genetic Progr Evol Mach* 2015, **16:**387-391.

130. Caracausi M, Piovesan A, Antonaros F, Strippoli P, Vitale L, Pelleri MC: **Systematic identification of human housekeeping genes possibly useful as references in gene expression studies.** *Molecular medicine reports* 2017, **16:**2397-2410.

131. Lee S, Jo M, Lee J, Koh SS, Kim S: **Identification of novel universal housekeeping genes by statistical analysis of microarray data.** *J Biochem Mol Biol* 2007, **40:**226-231.

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* 2010, **355:**76-79.

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* 2013, **92:**747-751.

134. Huggett J, Dheda K, Bustin S, Zumla A: **Real-time RT-PCR normalisation; strategies and considerations.** *Genes Immun* 2005, **6:**279-284.

135. Andersen CL, Jensen JL, Orntoft 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* 2004, **64:**5245-5250.

136. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP: **Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations.** *Biotechnol Lett* 2004, **26:**509-515.

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* 2012.

138. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al: **Proteomics. Tissue-based map of the human proteome.** *Science* 2015, **347:**1260419.

139. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.** *Nature* 2012, **483:**603-607.

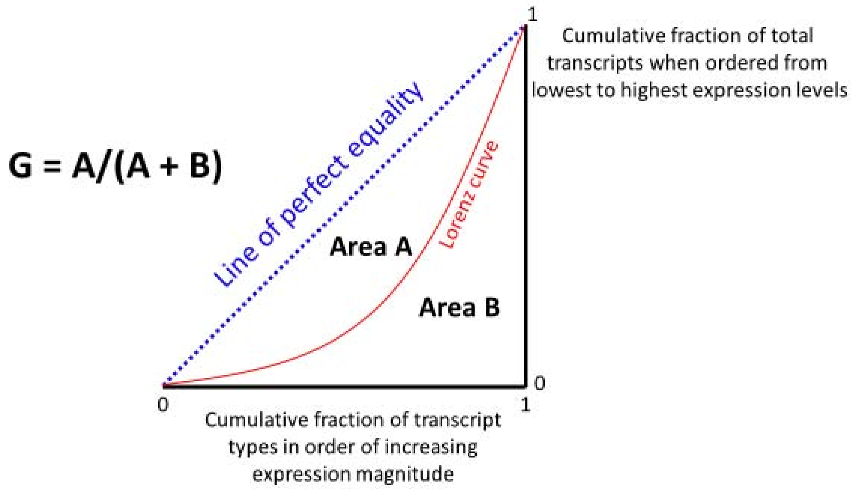
140. Klijn C, Durinck S, Stawiski EW, Haverty PM, Jiang Z, Liu H, Degenhardt J, Mayba O, Gnad F, Liu J, et al: **A comprehensive transcriptional portrait of human cancer cell lines.** *Nat Biotechnol* 2015, **33:**306-312.

141. Consortium GT: **Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.** *Science* 2015, **348:**648-660.

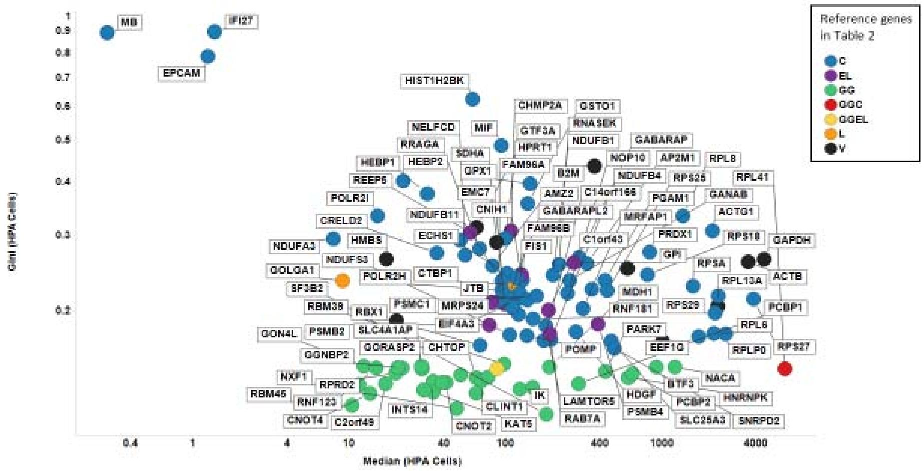
142. Livak KJ, Schmittgen TD: **Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method.** *Methods* 2001, **25:**402-408.

143. Conway JR, Lex A, Gehlenborg N: **UpSetR: an R package for the visualization of intersecting sets and their properties.** *Bioinformatics* 2017, **33:**2938-2940.

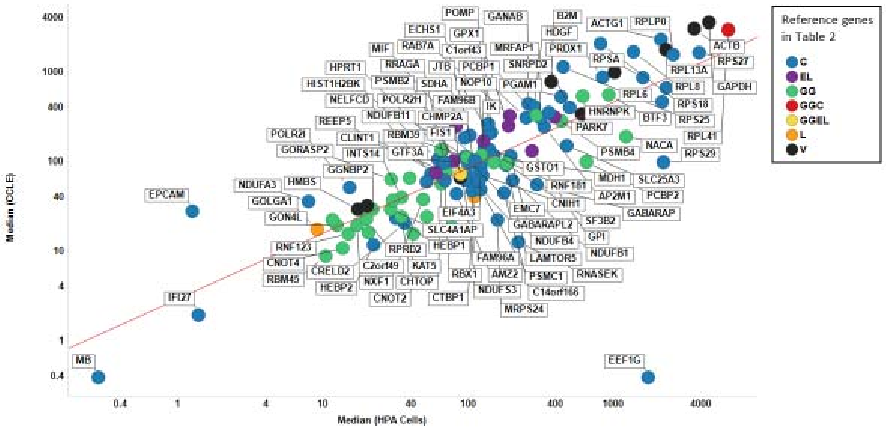
1


G = A/(A + B)
Line of perfect equality
Lorenz curve
Area A
Area B
1
0
0
1
Cumulative fraction of total transcripts when ordered from lowest to highest expression levels
Cumulative fraction of transcript types in order of increasing expression magnitude

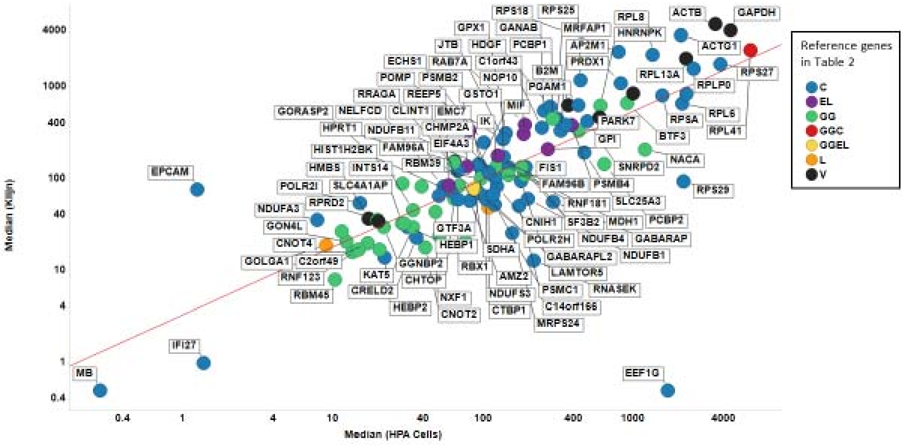
2 A.


Reference genes in Table 2
C
EL
GG
GGC
GGEL
L
V
Gini (HPA Cells)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Median (HPA Cells)
0.4
1
4
10
40
100
400
1000
4000
MB
IFI27
EPCAM
HIST1H2BK
CHMP2A
GSTO1
RNASEK
NELFCD
MIF
GTF3A
RRAGA
HPRT1
NDUFB1
GABARAP
SDHA
HEBP1
HEBP2
FAM96A
NOP10
AP2M1
RPL8
GPX1
B2M
NDUFB4
RPS25
RPL41
REEP5
EMC7
AMZ2
C14orf166
PGAM1
GANAB
POLR2I
CNIH1
GABARAPL2
MRFAP1
ACTG1
CRELD2
NDUFB11
FAM96B
PRDX1
RPS18
GAPDH
ECHS1
C1orf43
NDUFA3
HMBS
FIS1
GPI
NDUFS3
RPSA
GOLGA1
POLR2H
CTBP1
RPL13A
ACTB
SF3B2
JTB
MDH1
RBM39
PSMC1
MRPS24
RNF181
RPS29
PCBP1
RBX1
EIF4A3
PARK7
RPL6
GON4L
PSMB2
SLC4A1AP
POMP
EEF1G
RPLP0
RPS27
GGNBP2
GORASP2
CHTOP
NXF1
RPRD2
BTF3
NACA
RBM45
RNF123
INTS14
IK
LAMTOR5
HDGF
PCBP2
HNRNPK
CNOT4
C2orf49
CLINT1
CNOT2
KAT5
RAB7A
PSMB4
SLC25A3
SNRPD2

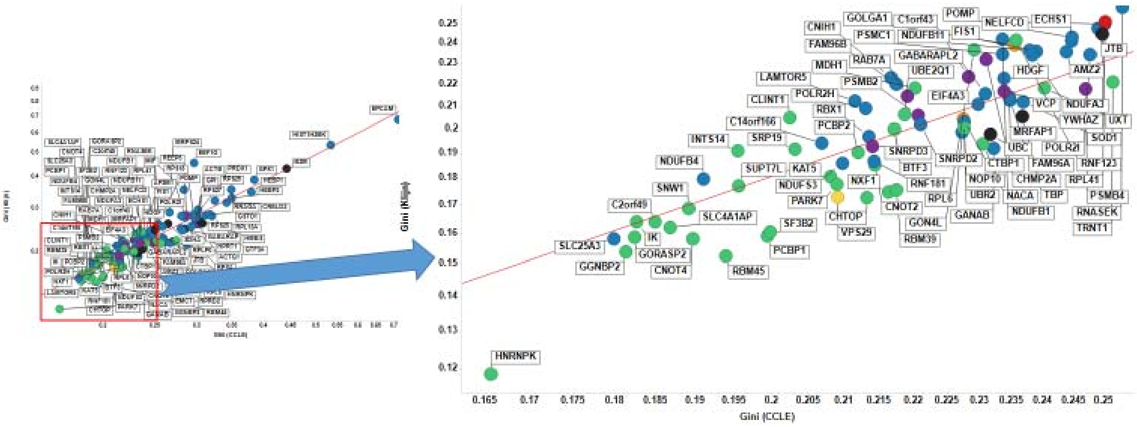
2 B.


Median (CCLE)
Median (HPA Cells)
4000
1000
400
100
40
10
4
1
0.4
0.4
1
4
10
40
100
400
1000
4000
Reference genes in Table 2
C
EL
GG
GGC
GGEL
L
V
POMP
GANAB
B2M
ACTG1
RPLP0
ECHS1
GPX1
HDGF
ACTB
MIF
RAB7A
C1orf43
MRFAP1
PRDX1
RPS27
HPRT1
RRAGA
JTB
PCBP1
SNRPD2
RPSA
RPL13A
HIST1H2BK
PSMB2
SDHA
NOP10
PGAM1
RPL8
GAPDH
NELFCD
POLR2H
FAM96B
RPL6
REEP5
NDUFB11
CHMP2A
IK
HNRNPK
BTF3
RPS18
RPS25
POLR2I
CLINT1
RBM39
FIS1
PARK7
RPL41
GORASP2
INTS14
GTF3A
PSMB4
NACA
RPS29
GGNBP2
GSTO1
NDUFA3
HMBS
MDH1
SLC25A3
EPCAM
RNF181
AP2M1
PCBP2
GOLGA1
EMC7
CNIH1
GABARAP
GON4L
EIF4A3
GABARAPL2
SF3B2
SLC4A1AP
GPI
RNF123
RPRD2
HEBP1
NDUFB4
CNOT4
FAM96A
LAMTOR5
NDUFB1
CRELD2
C2orf49
KAT5
RBX1
AMZ2
PSMC1
RNASEK
RBM45
HEBP2
NXF1
CHTOP
NDUFS3
C14orf166
IFI27
CNOT2
CTBP1
MRPS24
MB
EEF1G

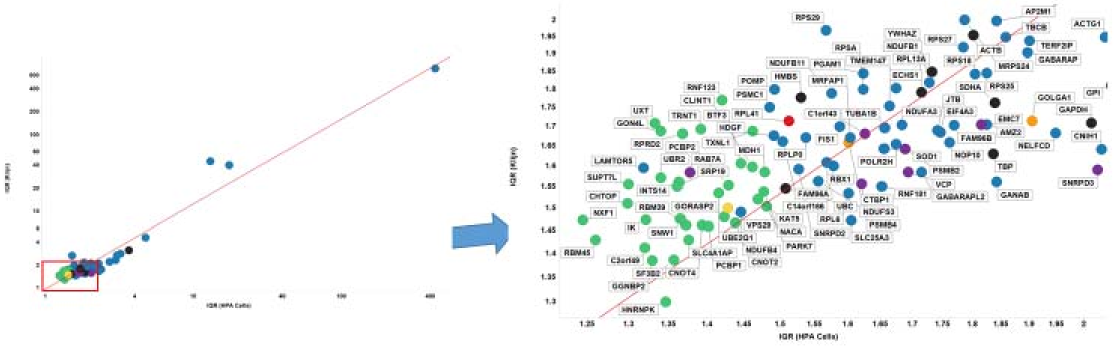
2 C.


Median (Klijn)
Median (HPA Cells)
4000
1000
400
100
40
10
4
1
0.4
0.4
1
4
10
40
100
400
1000
4000
Reference genes in Table 2
C
EL
GG
GGC
GGEL
L
V
RPS18
RPS25
RPL8
ACTB
GAPDH
GPX1
GANAB
MRFAP1
HNRNPK
JTB
HDGF
PCBP1
AP2M1
ACTG1
ECHS1
RAB7A
C1orf43
PRDX1
RPS27
POMP
PSMB2
NOP10
B2M
RPL13A
RRAGA
REEP5
GSTO1
PGAM1
RPLP0
GORASP2
NELFCD
CLINT1
EMC7
MIF
HPRT1
NDUFB11
CHMP2A
IK
PARK7
RPSA
RPL6
HIST1H2BK
FAM96A
EIF4A3
BTF3
RPL41
GPI
HMBS
INTS14
RBM39
NACA
SNRPD2
EPCAM
FIS1
POLR2I
SLC4A1AP
FAM96B
PSMB4
RPS29
RPRD2
RNF181
SLC25A3
NDUFA3
CNIH1
SF3B2
MDH1
PCBP2
GON4L
GTF3A
CNOT4
HEBP1
POLR2H
NDUFB4
GABARAP
SDHA
GABARAPL2
NDUFB1
GOLGA1
C2orf49
GGNBP2
RBX1
LAMTOR5
RNF123
KAT5
CHTOP
AMZ2
PSMC1
RNASEK
RBM45
CRELD2
NXF1
NDUFS3
HEBP2
CNOT2
CTBP1
C14orf166
MRPS24
IFI27
MB
EEF1G

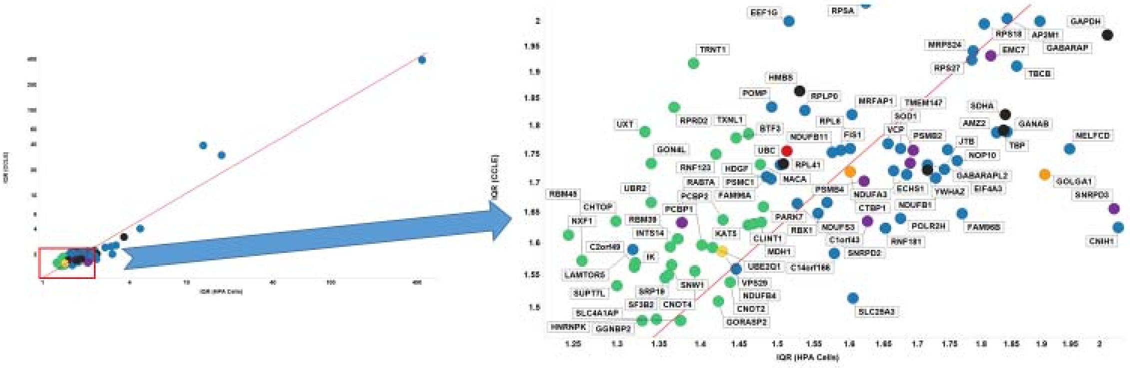
3.


Gini (Kilpi)
Gini (CCLE)
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.165
0.17
0.175
0.18
0.185
0.19
0.195
0.2
0.205
0.21
0.215
0.22
0.225
0.23
0.235
0.24
0.245
0.25
HNRNPK
GGNBP2
SLC25A3
CNOT4
GORASP2
IK
C2orf49
SNW1
NDUFB4
SLC4A1AP
RBM45
PCBP1
SF3B2
PARK7
VPS29
CHTOP
NDUFS3
SUPT7L
KAT5
NXF1
INTS14
SRP19
C14orf166
CLINT1
LAMTOR5
POLR2H
RBX1
PCBP2
PSMB2
MDH1
RAB7A
FAM96B
CNIH1
GOLGA1
PSMC1
C1orf43
NDUFB11
POMP
FIS1
NELFCD
ECHS1
GABARAPL2
UBE2Q1
EIF4A3
HDGF
AMZ2
JTB
VCP
NDUFA3
YWHAZ
UXT
MRFAP1
SOD1
UBC
POLR2I
SNRPD3
SNRPD2
BTF3
RNF181
RPL6
CNOT2
GON4L
RBM39
NOP10
CTBP1
FAM96A
RNF123
CHMP2A
RPL41
UBR2
NACA
TBP
PSMB4
GANAB
NDUFB1
RNASEK
TRNT1

4 A.

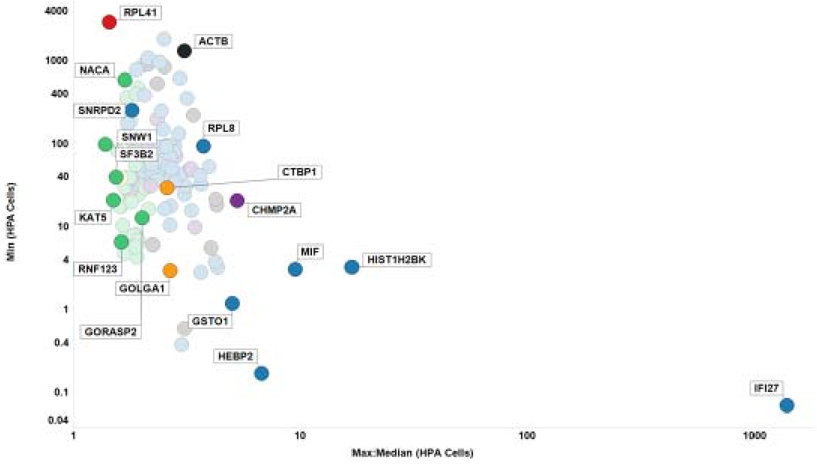

IQR (Kilpi)
IQR (HPA Cells)
1.3
1.35
1.4
1.45
1.5
1.55
1.6
1.65
1.7
1.75
1.8
1.85
1.9
1.95
2
1.25
1.3
1.35
1.4
1.45
1.5
1.55
1.6
1.65
1.7
1.75
1.8
1.85
1.9
1.95
2
HNRNPK
GGNBP2
SF3B2
CNOT4
C3orf49
RBM45
IK
SMW1
NXF1
RBM39
GORASP2
CHTOP
INTS14
SUPT7L
LAMTOR5
UBR2
RAB7A
SRP19
RPRD2
PCBP2
TXNL1
GON4L
UXT
TRNT1
BTF3
RPL41
HDGF
CLINT1
RNF123
PSMC1
POMP
HMBS
MRFAP1
NDUFB11
PGAM1
RPSA
RPS29
YWHAZ
NDUFB1
TMEM147
RPL13A
RPS27
RPS18
ACTB
TBCB
AP2M1
ACTG1
TERF2IP
GABARAP
MRPS34
ECHS1
SDHA
RPS25
GOLGA1
GPI
GAPDH
JTB
EIF4A3
NDUFA3
C1orf43
TUBA1B
FIS1
EMC7
AMZ2
FAM96B
NELFCD
CNIH1
MDH1
RPLP0
POLR2H
SOD1
NOP10
TBP
PSMB2
VCP
SNRPD3
RBX1
FAM96A
RNF181
GABARAPL2
GANAB
CTBP1
UBC
C14orf166
NDUFS3
KAT5
RPL6
PSMB4
VPS29
NACA
SNRPD2
SLC25A3
UBE2Q1
SLC4A1AP
NDUFB4
PARK7
PCBP1
CNOT2

4 B.


IQR (CCLE)
IQR (HPA Cells)
1
4
10
40
100
400
IQR (CCLE)
2
1.95
1.9
1.85
1.8
1.75
1.7
1.65
1.6
1.55
1.5
EEF1G
RPSA
TRNT1
MRPS24
RPS18
AP3M1
GAPDH
EMC7
GABARAP
RPS27
TBCB
HMBS
POMP
RPLP0
MRFAP1
TMEM147
SDHA
SOD1
UXT
RPRD2
TXNL1
BTF3
RPL8
AMZ2
GANAB
NDUFB11
FIS1
VCP
PSMB2
MELFCD
GON4L
UBC
JTB
TBP
RNF123
HDGF
RPL41
NOP10
RAB7A
PSMC1
NACA
GABARAPL2
GOLGA1
RBM45
UBR2
PCBP2
FAM98A
PSMB4
NDUFA3
ECHS1
YWHAZ
EIF4A3
SNRPD3
CHTOP
PCBP1
CTBP1
NDUFB1
NXF1
RBM39
PARK7
POLR2H
FAM96B
INTS14
KAT5
RBX1
NDUFS3
CNIH1
C2orf49
CLINT1
C1orf43
RNF181
IK
MDH1
SNRPD2
LAMTOR5
UBE2D1
C14orf166
SNW1
VPS29
SUPT7L
SRP19
NDUFB4
SF3B2
CNOT4
SLC25A3
SLC4A1AP
CNOT2
HNRNPK
GGNBP2
GORASP2
1.25
1.3
1.35
1.4
1.45
1.5
1.55
1.6
1.65
1.7
1.75
1.8
1.85
1.9
1.95
2
IQR (HPA Cells)

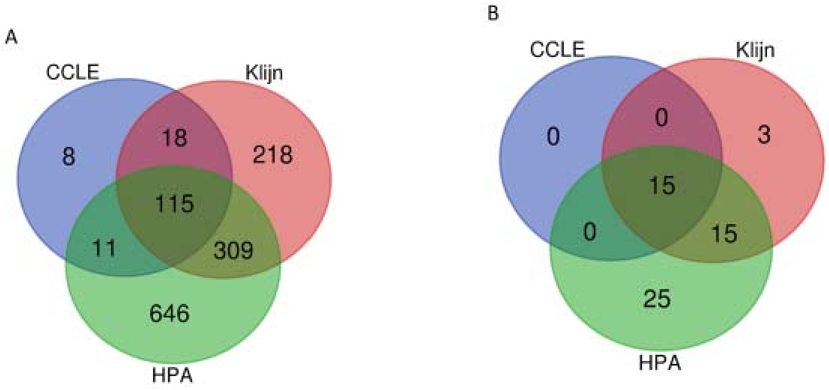
923

4 C.

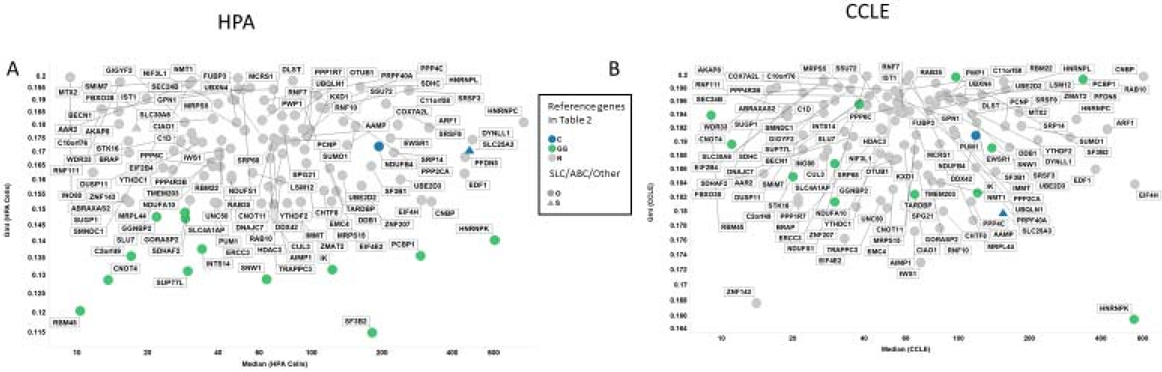

4000
1000
400
100
40
10
4
1
0.4
0.1
0.04
Min (HPA Cells)
RPL41
ACTB
NACA
SNRPD2
RPL8
SNW1
SF3B2
CTBP1
CHMP2A
KAT5
MIF
HIST1H2BK
RNF123
GOLGA1
GSTO1
GORASP2
HEBP2
IFI27
1
10
100
1000
Max:Median (HPA Cells)

924

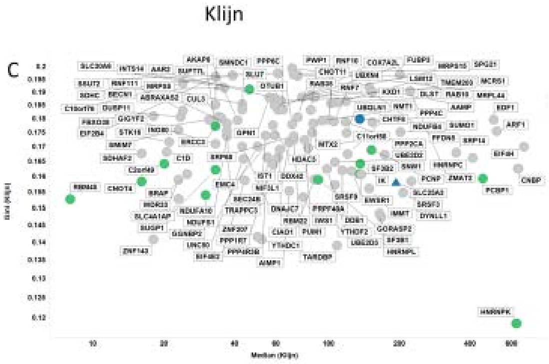
5.


A
CCLE
Klijn
HPA
8
18
218
115
11
309
646
B
CCLE
Klijn
HPA
0
0
3
15
0
15
25

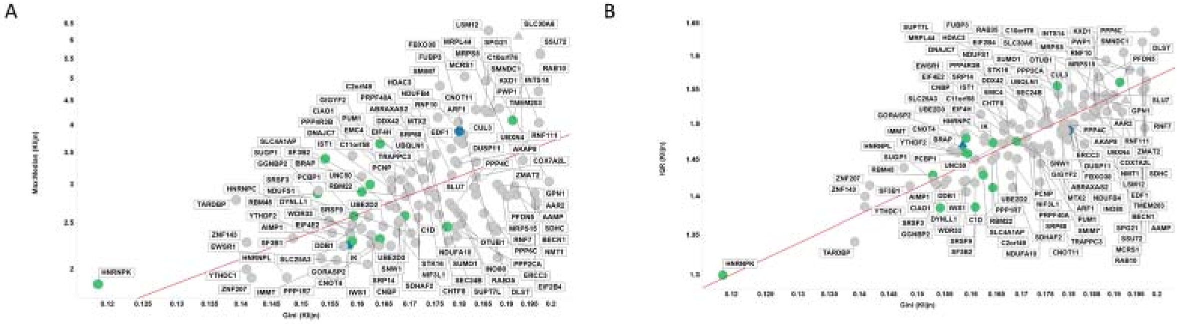
6


HPA
A
Gini (HPA Cells)
Median (HPA Cells)
10
20
40
60
100
200
400
600
RBM45
SF3B2
HNRNPK
CNOT4
SUPT7L
Reference genes in Table 2
● C
● GG
● H
SLC/ABC/Other
● O
▲ S
CCLE
B
Gini (CCLE)
Median (CCLE)
10
20
40
60
100
200
400
600
ZNF143
HNRNPK
EIF4H

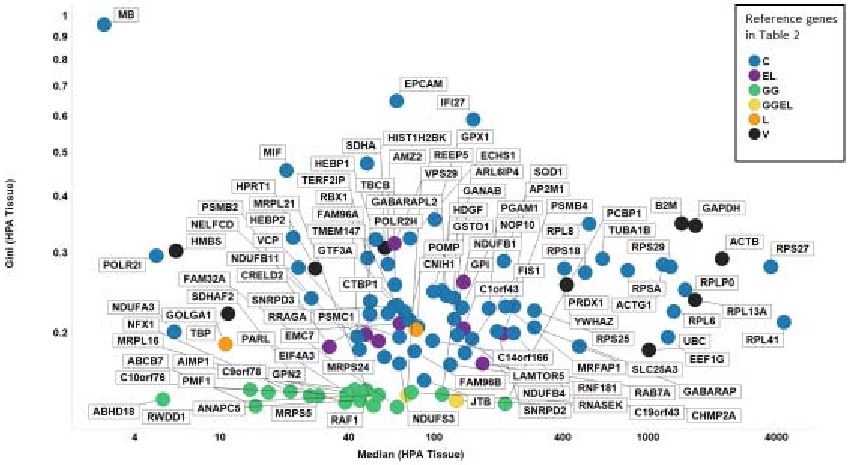
6


Klijn
C
0.2
0.195
0.19
0.185
0.18
0.175
0.17
0.165
0.16
0.155
0.15
0.145
0.14
0.135
0.13
0.125
0.12
Gini (Klijn)
10
20
40
60
100
200
400
600
Median (Klijn)
SLC30A6
INTS14
AAR2
AKAP8
SMNDC1
PPP6C
SUPT7L
SLUT
PWP1
RNF10
COX7A2L
FUBP3
MRPS15
SPG21
CNOT11
UBXN4
SSUT2
RNF111
MRPS8
OTUB1
RAB35
RNF7
LSM12
TMEM203
MCRS1
SDHC
BECN1
ABRAXAS2
CUL3
KXD1
DLST
RAB10
MRPL44
C18orf76
DUSP11
UBQLN1
NMT1
AAMP
EDF1
FBXO38
GIGYF2
CHTF8
PPP4C
EIF2B4
STK16
INO80
GPN1
NDUFB4
SUMO1
ARF1
SMIM7
ERCC3
C11orf58
PFDN5
SRP14
MTX2
PPP2CA
SDHAF2
C1D
SRP68
HDAC3
UBE2D2
EIF4H
C3orf49
SF3B2
SNW1
HNRNPC
RBM45
CNOT4
EMC4
IST1
DDX42
IK
PCNP
ZMAT2
CNBP
NIF3L1
PCBP1
BRAP
SEC24B
SRSF9
EWSR1
SLC25A3
WDR33
NDUFA10
TRAPPC3
DNAJC7
PRPF40A
SRSF3
SLC4A1AP
NDUFS1
RBM22
IWS1
DDB1
IMMT
DYNLL1
SUGP1
GGNBP2
ZNF207
CIAO1
PUM1
YTHDF2
GORASP2
ZNF143
UNC50
PPP1R7
YTHDC1
UBE2D3
SF3B1
EIF4E2
PPP4R3B
TARDBP
HNRNPL
AIMP1
HNRNPK

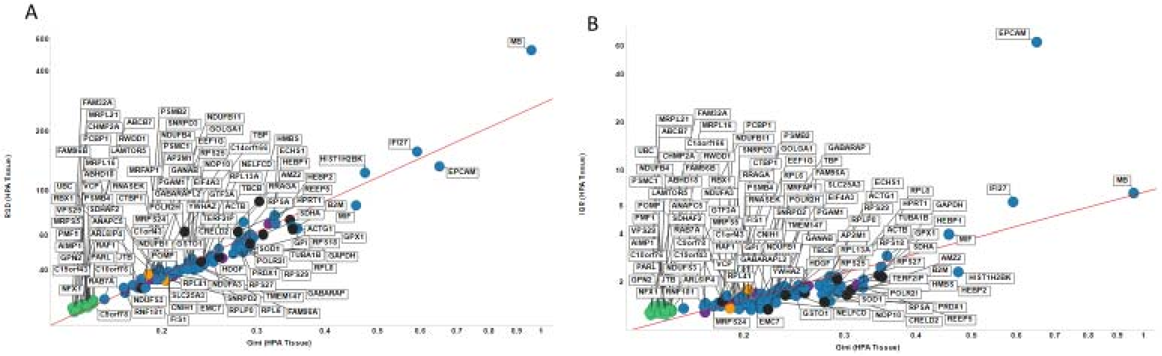
7


A
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
Max/Median (Klijn)
0.12
0.125
0.13
0.135
0.14
0.145
0.15
0.155
0.16
0.165
0.17
0.175
0.18
0.185
0.19
0.195
0.2
Gini (Klijn)
LSM12
SLC30A6
FBXO38
MRPL44
SPG21
SSUT2
FUBP3
MRPS8
C18orf76
SMIM7
MCRS1
SMNDC1
RAB10
C3orf49
HDAC3
KXD1
INTS14
GIGYF2
PRPF40A
NDUFB4
CNOT11
PWP1
CIAO1
ABRAXAS2
RNF10
ARF1
TMEM203
PPP4R3B
PUM1
DDX42
MTX2
DNAJC7
EMC4
EIF4H
SRP68
EDF1
CUL3
SLC4A1AP
UBXN4
RNF111
IST1
C11orf58
UBQLN1
SUGP1
SF3B2
TRAPPC3
DUSP11
AKAP8
GGNBP2
BRAP
PCNP
PPP4C
COX7A2L
SRSF3
PCBP1
UNC50
ZMAT2
HNRNPC
NDUFS1
RBM22
SLUT
GPN1
TARDBP
RBM45
DYNLL1
AAR2
SRSF9
UBE2D2
PFDN5
AAMP
YTHDF2
WDR33
MRPS15
SDHC
AIMP1
EIF4E2
C1D
RNF7
BECN1
ZNF143
OTUB1
PPP6C
NMT1
EWSR1
SF3B1
DDB1
HNRNPL
SLC25A3
IK
UBE2D3
NDUFA10
PPP2CA
STK16
SUMO1
INO80
SNW1
GORASP2
ERCC3
YTHDC1
CNOT4
NIF3L1
SRP14
SEC24B
RAB35
SDHAF2
EIF2B4
ZNF207
IMMT
PPP1R7
IWS1
CNBP
CHTF8
SUPT7L
DLST
HNRNPK
B
1.65
1.6
1.55
1.5
1.45
1.4
1.35
1.3
IQR (Klijn)
0.12
0.125
0.13
0.135
0.14
0.145
0.15
0.155
0.16
0.165
0.17
0.175
0.18
0.185
0.19
0.195
0.2
Gini (Klijn)
SUPT7L
FUBP3
RAB35
C18orf76
INTS14
KXD1
PPP6C
MRPL44
HDAC3
EIF2B4
SLC30A6
MRPS8
PWP1
SMNDC1
DLST
DNAJC7
NDUFS1
RNF10
SUMO1
OTUB1
PFDN5
EWSR1
PPP4R3B
STK16
PPP2CA
MRPS15
EIF4E2
SRP14
DDX42
UBQLN1
CUL3
CNBP
IST1
EMC4
SEC24B
SLC25A3
C11orf58
CHTF8
SLUT
UBE2D3
EIF4H
GPN1
GORASP2
HNRNPC
IK
AAR2
RNF7
IMMT
CNOT4
PPP4C
YTHDF2
BRAP
AKAP8
RNF111
HNRNPL
ERCC3
UBXN4
ZMAT2
SUGP1
PCBP1
UNC50
SNW1
DUSP11
COX7A2L
RBM45
GIGYF2
FBXO38
NMT1
SDHC
ZNF207
ABRAXAS2
LSM12
ZNF143
SF3B1
PCNP
MTX2
NDUFB4
EDF1
AIMP1
DDB1
UBE2D2
NIF3L1
ARF1
INO80
TMEM203
YTHDC1
CIAO1
IWS1
PPP1R7
PRPF40A
PUM1
BECN1
SRSF3
DYNLL1
C1D
RBM22
SRP68
SPG21
AAMP
GGNBP2
WDR33
SLC4A1AP
SMIM7
SSUT2
TARDBP
SRSF9
C3orf49
SDHAF2
TRAPPC3
MCRS1
SF3B2
NDUFA10
CNOT11
RAB10
HNRNPK

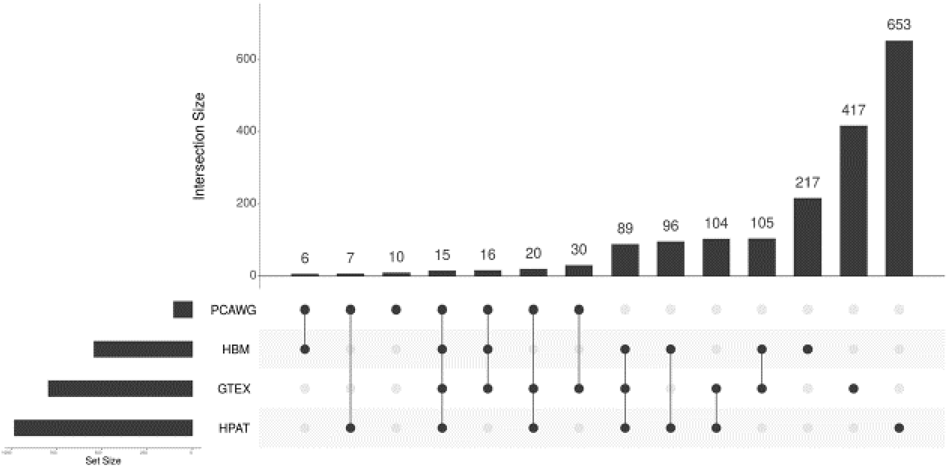
8


Reference genes in Table 2
● C
● EL
● GG
● GGEL
● L
● V
Gini (HPA Tissue)
Median (HPA Tissue)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
4
10
40
100
400
1000
4000
MB
EPCAM
IFI27
GPX1
HIST1H2BK
SDHA
MIF
HEBP1
AMZ2
REEP5
ECHS1
VPS29
ARL6IP4
SOD1
HPRT1
TERF2IP
TBCB
GANAB
AP2M1
RBX1
GABARAPL2
PSMB2
MRPL21
FAM96A
HDGF
PGAM1
PSMB4
PCBP1
B2M
GAPDH
NELFCD
HEBP2
TMEM147
POLR2H
GSTO1
NOP10
RPL8
TUBA1B
HMBS
VCP
GTF3A
POMP
NDUFB1
RPS18
RPS29
ACTB
RPS27
NDUFB11
POLR2I
CNIH1
GPI
FIS1
FAM32A
CRELD2
CTBP1
RPLP0
RPSA
SDHAF2
SNRPD3
C1orf43
NDUFA3
GOLGA1
RRAGA
PSMC1
PRDX1
ACTG1
RPL13A
NFX1
YWHAZ
RPL6
TBP
MRPL16
PARL
EMC7
RPS25
UBC
RPL41
EIF4A3
ABCB7
AIMP1
C14orf166
EEF1G
C9orf78
MRPS24
MRFAP1
SLC25A3
C10orf76
GPN2
LAMTOR5
PMF1
FAM96B
RNF181
RAB7A
GABARAP
NDUFB4
JTB
ABHD18
ANAPC5
RNASEK
SNRPD2
C19orf43
CHMP2A
RWDD1
MRPS5
RAF1
NDUFS3

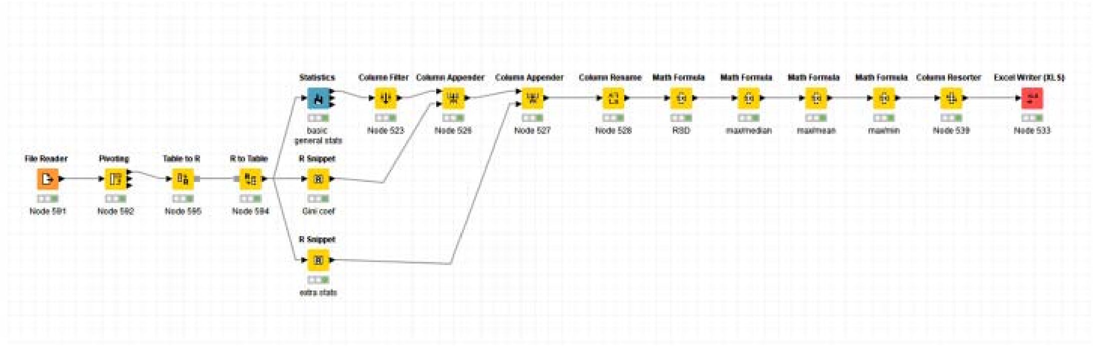
9


A
B
RSD (HPA Tissue)
IQR (HPA Tissue)
Gini (HPA Tissue)
MB
IFI27
EPCAM
HIST1H2BK

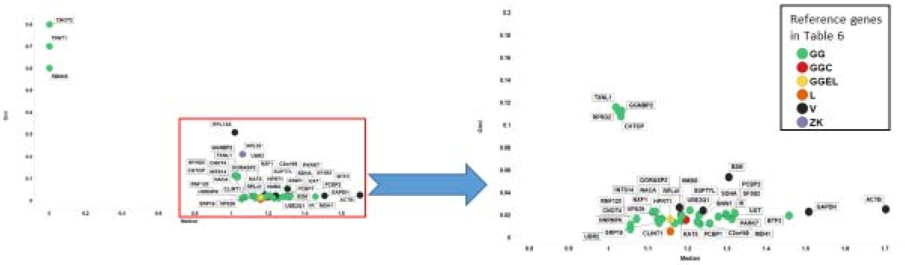
10


Intersection Size
600
400
200
0
6
7
10
15
16
20
30
89
96
104
105
217
417
653
PCAWG
HBM
GTEX
HPAT
Set Size

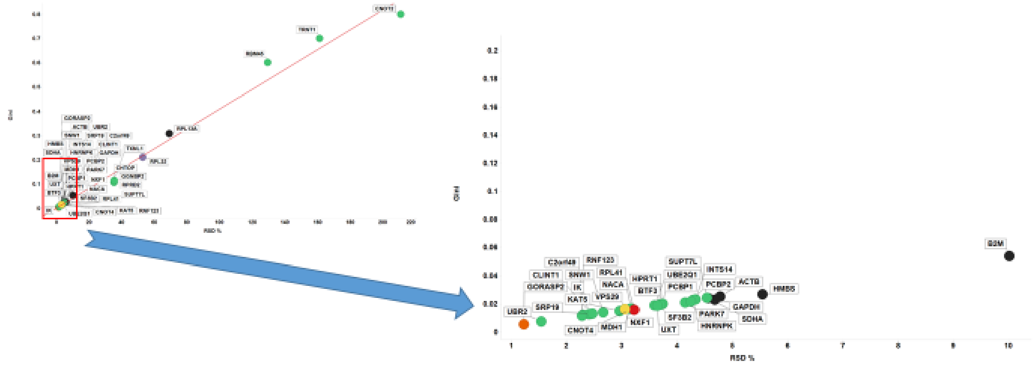
11


File Reader
Node 591
Pivoting
Node 592
Table to R
Node 595
R to Table
Node 594
Statistics
basic
general stats
R Snippet
Gini coef
R Snippet
extra stats
Column Filter
Node 523
Column Appender
Node 526
Column Appender
Node 527
Column Rename
Node 528
Math Formula
RSD
Math Formula
max/median
Math Formula
max/mean
Math Formula
max/min
Column Resorter
Node 539
Excel Writer (XLS)
Node 533

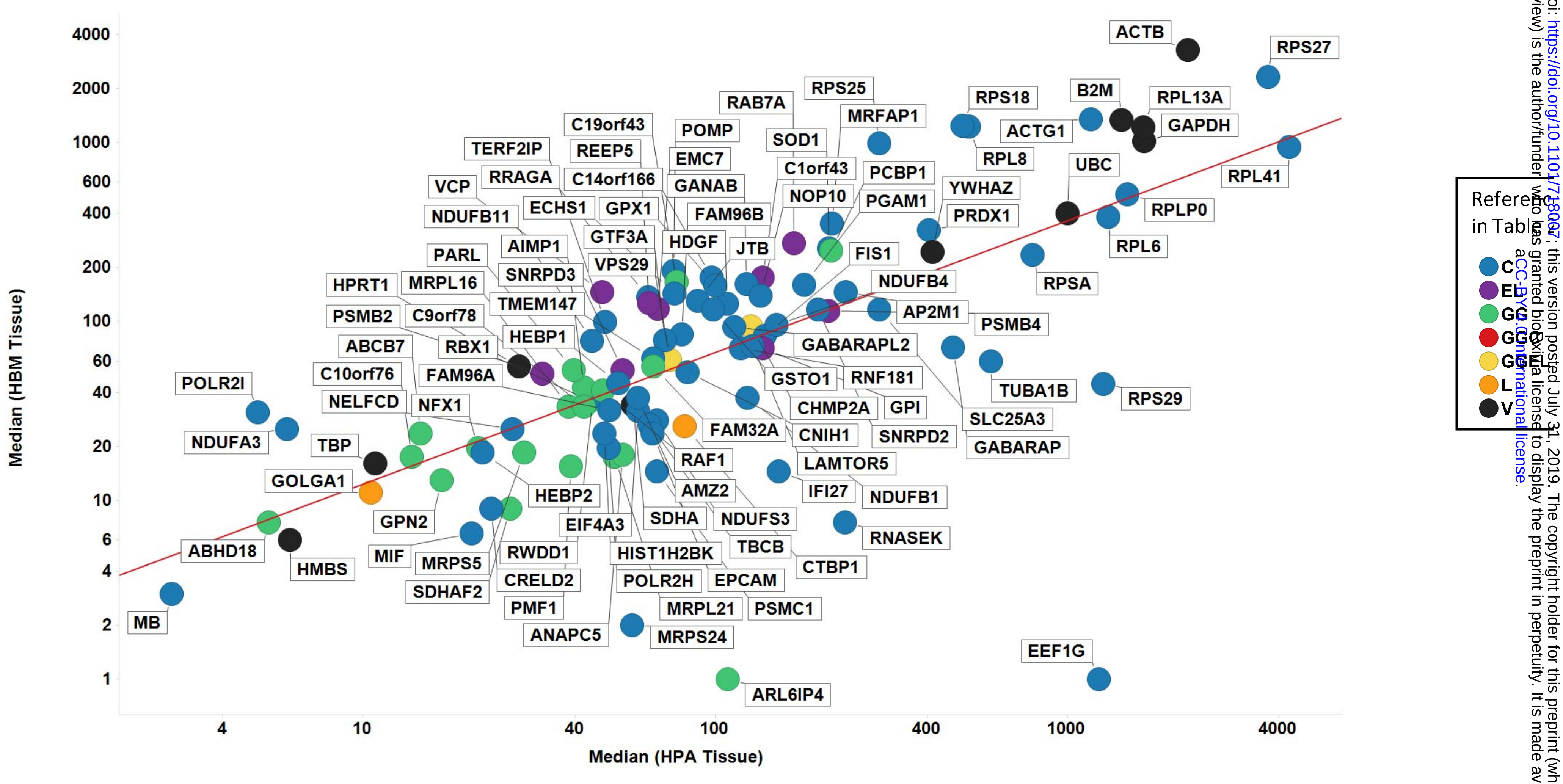
12


Reference genes in Table 6
GG
GGC
GGEL
L
V
ZK
Median
CHTOP
GORASP2
INTS14
NACA
RPL41
SUPT7L
SDHA
SF3B2
RNF123
NXF1
HPRT1
UBE2Q1
SNW1
IK
CNOT4
VPS29
UXT
HNRNPK
PARK7
BTF3
UBR2
SRP19
CLINT1
KAT5
PCBP1
C2orf49
MDH1
PCBP2
B2M
GAPDH
ACTB
RPL13A

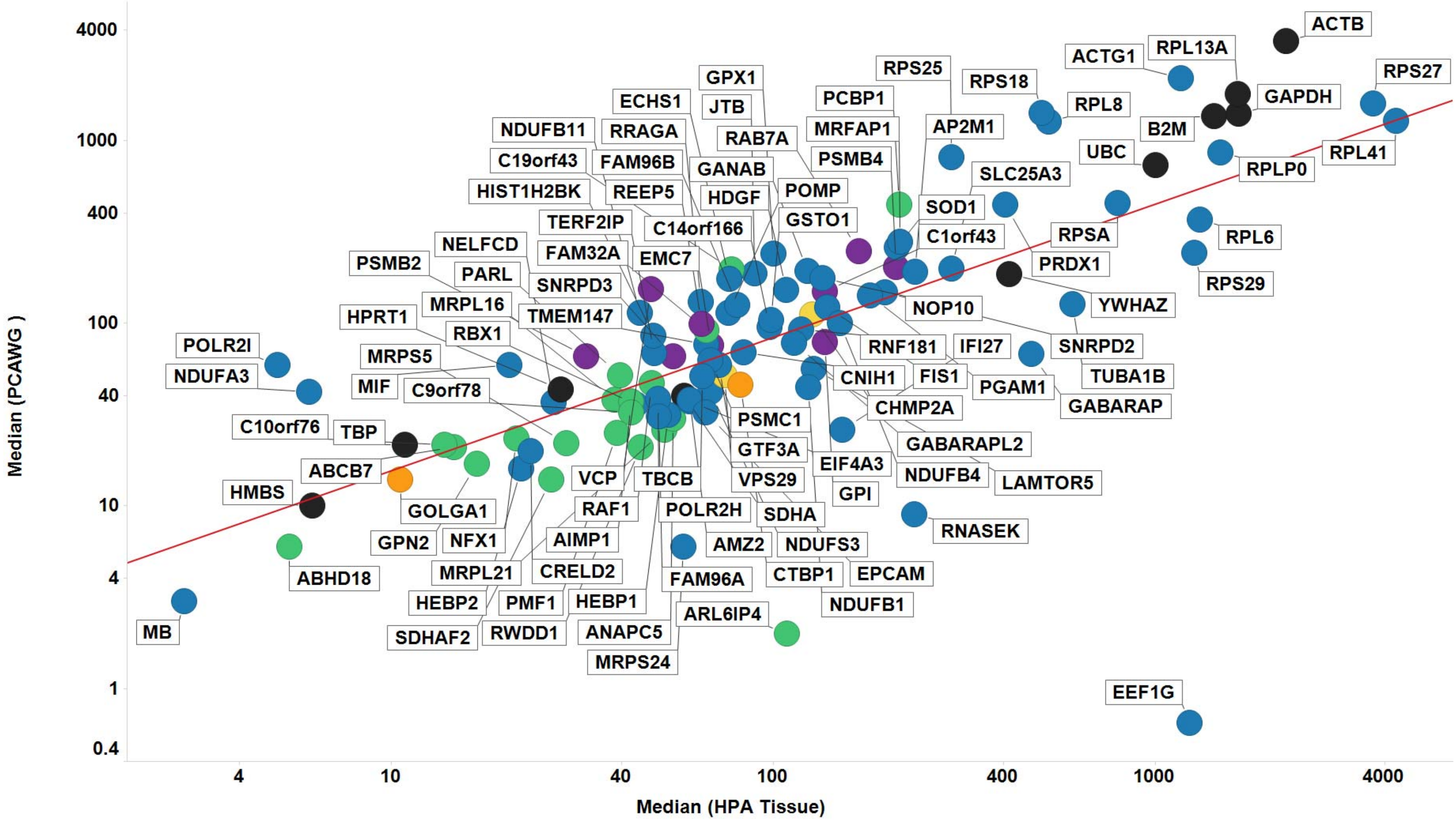
13


CNOT2
TRNT1
RPL13A
RPL32
RSD %
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1
2
3
4
5
6
7
8
9
10
B2M
C2orf49
RNF123
SUPT7L
INTS14
CLINT1
SNW1
RPL41
HPRT1
UBE2Q1
GORASP2
IK
NACA
BTF3
PCBP1
PCBP2
ACTB
HMBS
KAT5
VPS29
UBR2
SRP19
GAPDH
SF3B2
PARK7
SDHA
NXF1
CNOT4
MDH1
UXT
HNRNPK

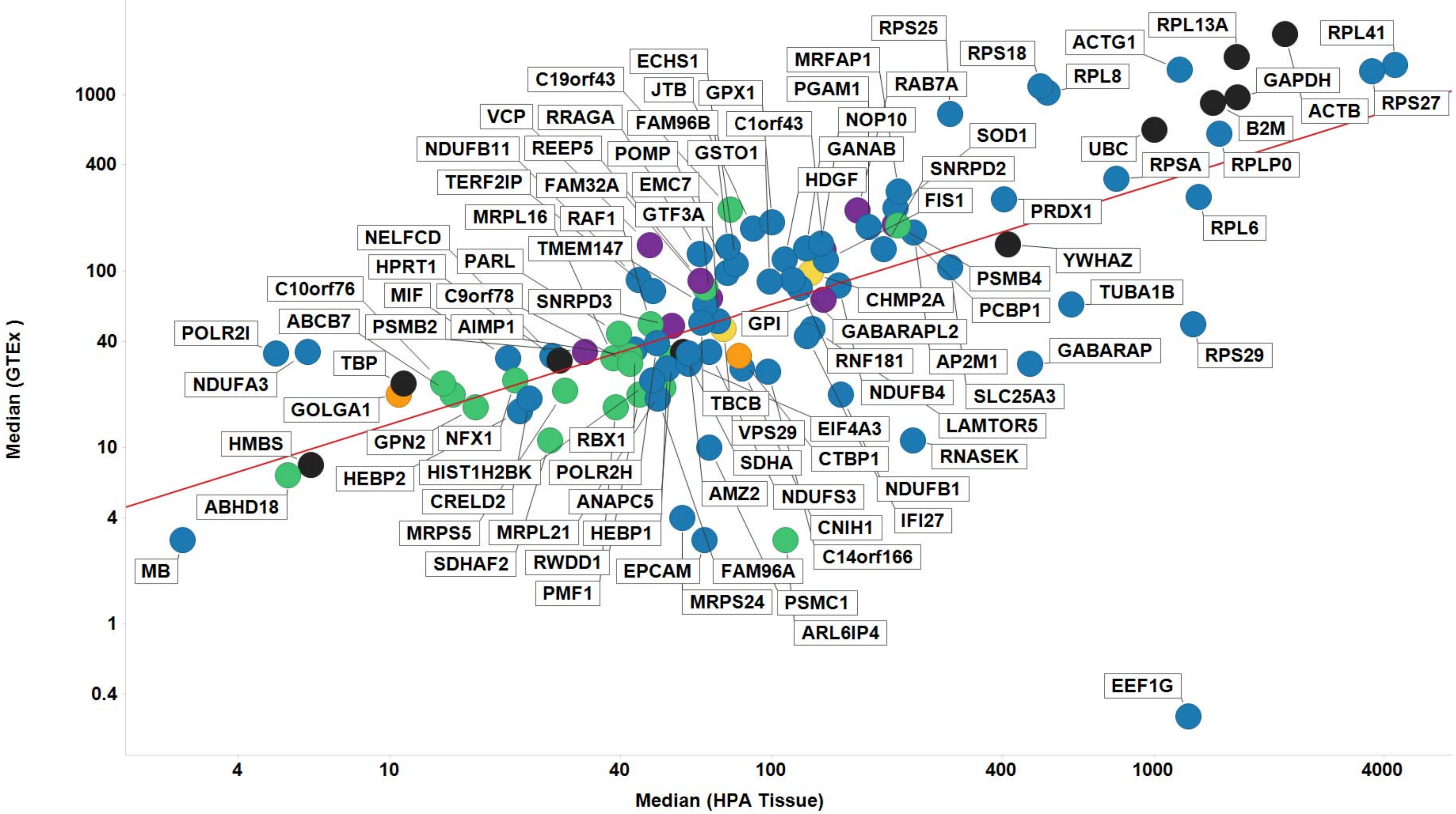
1 A


Median (HBM Tissue)
Median (HPA Tissue)
4000
2000
1000
600
400
200
100
60
40
20
10
6
4
2
1
4
10
40
100
400
1000
4000
ACTB
RPS27
RPS25
RAB7A
MRFAP1
RPS18
B2M
RPL13A
GAPDH
ACTG1
C19orf43
POMP
SOD1
TERF2IP
REEP5
EMC7
C1orf43
PCBP1
RPL8
UBC
RPL41
RRAGA
C14orf166
GANAB
NOP10
YWHAZ
VCP
ECHS1
GPX1
FAM96B
PGAM1
PRDX1
RPLP0
NDUFB11
GTF3A
HDGF
JTB
FIS1
RPL6
PARL
AIMP1
VPS29
NDUFB4
RPSA
HPRT1
MRPL16
SNRPD3
TMEM147
AP2M1
PSMB4
PSMB2
C9orf78
HEBP1
GABARAPL2
ABCB7
RBX1
C10orf76
FAM96A
GSTO1
RNF181
TUBA1B
RPS29
POLR2I
NELFCD
NFX1
CHMP2A
GPI
SLC25A3
FAM32A
CNIH1
SNRPD2
GABARAP
NDUFA3
TBP
RAF1
LAMTOR5
GOLGA1
AMZ2
IFI27
NDUFB1
HEBP2
SDHA
NDUFS3
GPN2
EIF4A3
RNASEK
MIF
RWDD1
TBCB
ABHD18
HMBS
MRPS5
HIST1H2BK
CTBP1
CRELD2
POLR2H
EPCAM
SDHAF2
PMF1
MRPL21
PSMC1
MB
ANAPC5
MRPS24
EEF1G
ARL6IP4
Referen
in Tabl
C
EL
GG
GG
GGEL
L
V

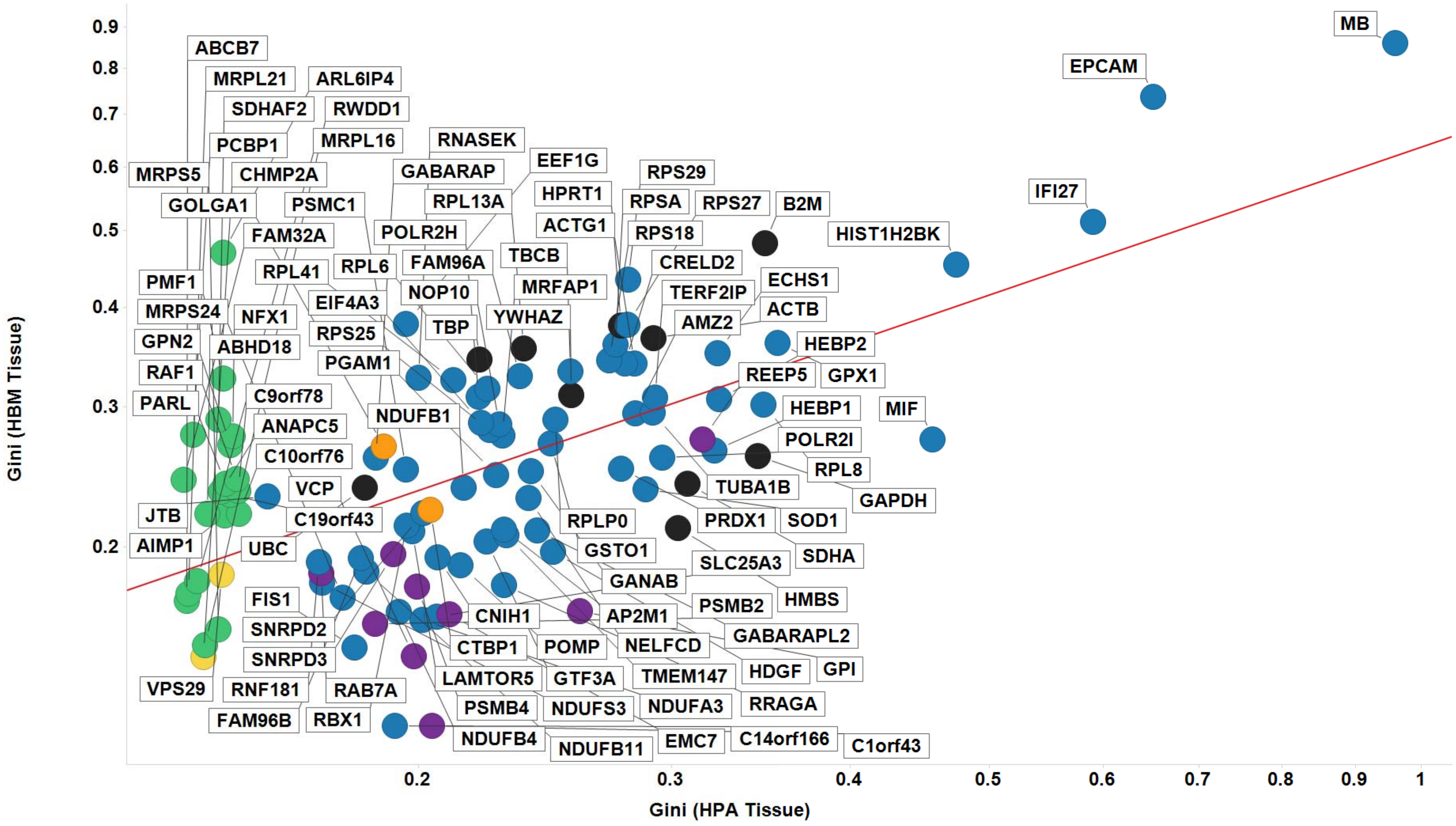
1 B


ACTB
RPL13A
ACTG1
RPS27
GAPDH
RPS25
RPS18
GPX1
PCBP1
RPL8
ECHS1
JTB
B2M
NDUFB11
RRAGA
MRFAP1
AP2M1
RAB7A
UBC
RPL41
C19orf43
FAM96B
PSMB4
GANAB
RPLP0
SLC25A3
HIST1H2BK
REEP5
POMP
HDGF
SOD1
TERF2IP
GSTO1
C14orf166
C1orf43
RPSA
RPL6
NELFCD
FAM32A
PSMB2
EMC7
PRDX1
PARL
SNRPD3
RPS29
MRPL16
YWHAZ
TMEM147
NOP10
HPRT1
RBX1
POLR2I
RNF181
IFI27
SNRPD2
MRPS5
NDUFA3
CNIH1
FIS1
TUBA1B
MIF
C9orf78
PGAM1
GABARAP
CHMP2A
C10orf76
PSMC1
TBP
GABARAPL2
GTF3A
ABCB7
EIF4A3
NDUFB4
LAMTOR5
VCP
TBCB
VPS29
HMBS
GPI
RAF1
POLR2H
SDHA
GOLGA1
RNASEK
GPN2
NFX1
AIMP1
AMZ2
NDUFS3
ABHD18
MRPL21
CRELD2
CTBP1
EPCAM
FAM96A
HEBP2
PMF1
HEBP1
NDUFB1
MB
ARL6IP4
SDHAF2
RWDD1
ANAPC5
MRPS24
EEF1G
4000
1000
400
100
40
10
4
1
0.4
Median (PCAWG)
4
10
40
100
400
1000
4000
Median (HPA Tissue)

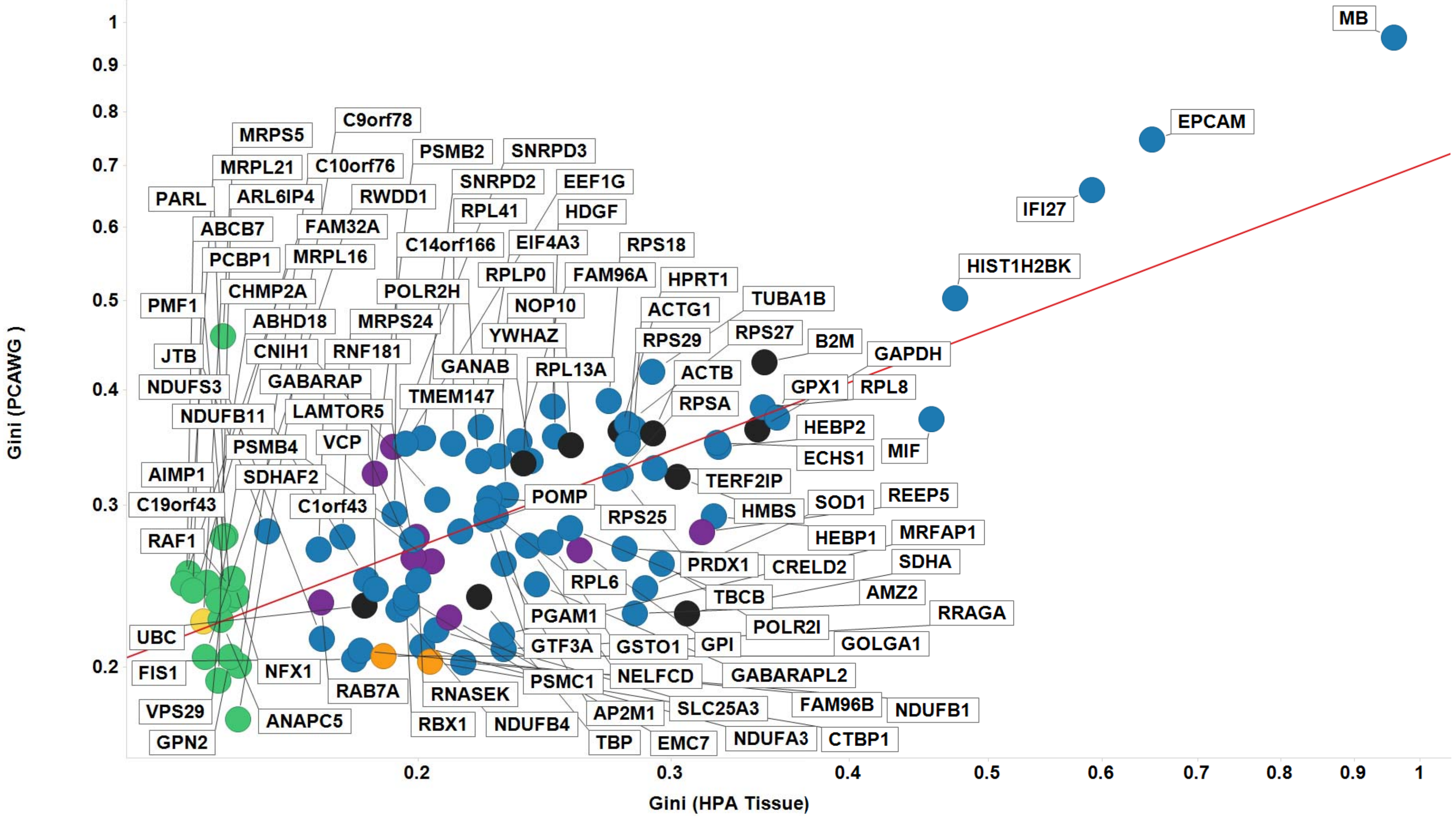
1 C


RPS25
RPL13A
RPL41
ACTG1
RPS18
ECHS1
MRFAP1
C19orf43
RPL8
GAPDH
JTB
GPX1
PGAM1
RAB7A
RPS27
ACTB
VCP
RRAGA
FAM96B
C1orf43
NOP10
B2M
SOD1
NDUFB11
REEP5
POMP
GSTO1
GANAB
UBC
RPSA
RPLP0
SNRPD2
TERF2IP
FAM32A
EMC7
HDGF
FIS1
PRDX1
MRPL16
RAF1
GTF3A
RPL6
NELFCD
TMEM147
PARL
HPRT1
YWHAZ
PSMB4
C10orf76
MIF
C9orf78
SNRPD3
TUBA1B
CHMP2A
PCBP1
ABCB7
PSMB2
AIMP1
GPI
GABARAPL2
POLR2I
RPS29
GABARAP
TBP
RNF181
AP2M1
NDUFA3
NDUFB4
SLC25A3
TBCB
GOLGA1
LAMTOR5
EIF4A3
VPS29
HMBS
GPN2
NFX1
RBX1
RNASEK
SDHA
CTBP1
HEBP2
HIST1H2BK
POLR2H
NDUFB1
AMZ2
NDUFS3
CRELD2
ANAPC5
ABHD18
IFI27
MRPS5
MRPL21
HEBP1
CNIH1
C14orf166
SDHAF2
RWDD1
EPCAM
FAM96A
MB
PMF1
MRPS24
PSMC1
ARL6IP4
EEF1G
1000
400
100
40
10
4
1
0.4
Median (GTEx)
4
10
40
100
400
1000
4000
Median (HPA Tissue)

2 A

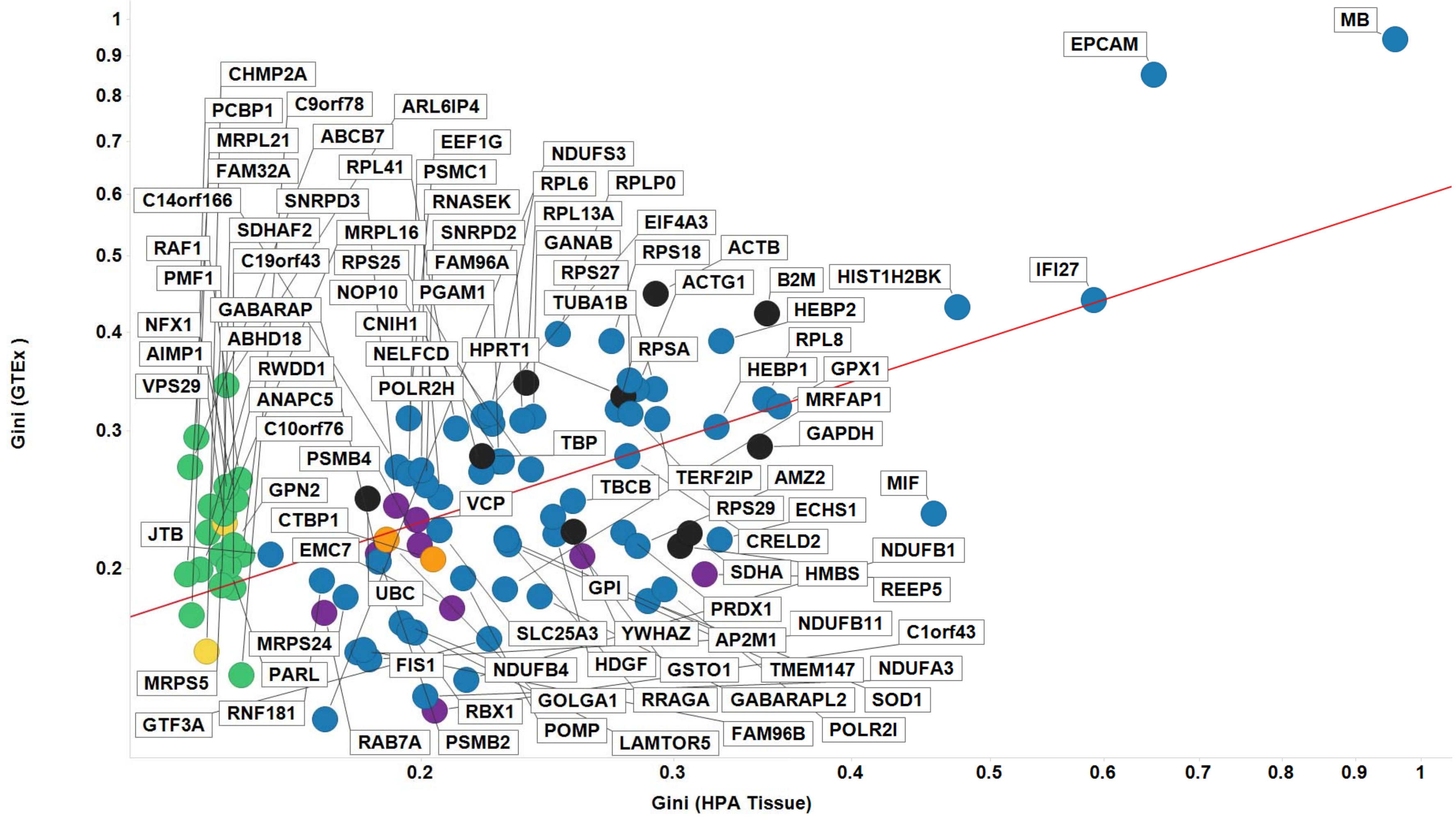

MB
ABCB7
EPCAM
MRPL21
ARL6IP4
SDHAF2
RWDD1
PCBP1
MRPL16
RNASEK
EEF1G
MRPS5
CHMP2A
GABARAP
RPS29
HPRT1
GOLGA1
PSMC1
RPL13A
RPSA
RPS27
B2M
IFI27
ACTG1
FAM32A
POLR2H
RPS18
HIST1H2BK
TBCB
RPL41
RPL6
FAM96A
CRELD2
PMF1
NOP10
MRFAP1
TERF2IP
ECHS1
MRPS24
NFX1
EIF4A3
YWHAZ
ACTB
TBP
AMZ2
GPN2
RPS25
HEBP2
ABHD18
PGAM1
RAF1
REEP5
GPX1
C9orf78
PARL
NDUFB1
HEBP1
MIF
ANAPC5
POLR2I
C10orf76
RPL8
VCP
TUBA1B
GAPDH
JTB
C19orf43
RPLP0
PRDX1
SOD1
AIMP1
UBC
GSTO1
SDHA
SLC25A3
GANAB
FIS1
HMBS
PSMB2
CNIH1
AP2M1
SNRPD2
GABARAPL2
CTBP1
POMP
NELFCD
SNRPD3
HDGF
GPI
LAMTOR5
GTF3A
TMEM147
VPS29
RNF181
RAB7A
RRAGA
PSMB4
NDUFS3
NDUFA3
FAM96B
RBX1
NDUFB4
EMC7
C14orf166
NDUFB11
C1orf43
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Gini (HBM Tissue)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Gini (HPA Tissue)

2 B



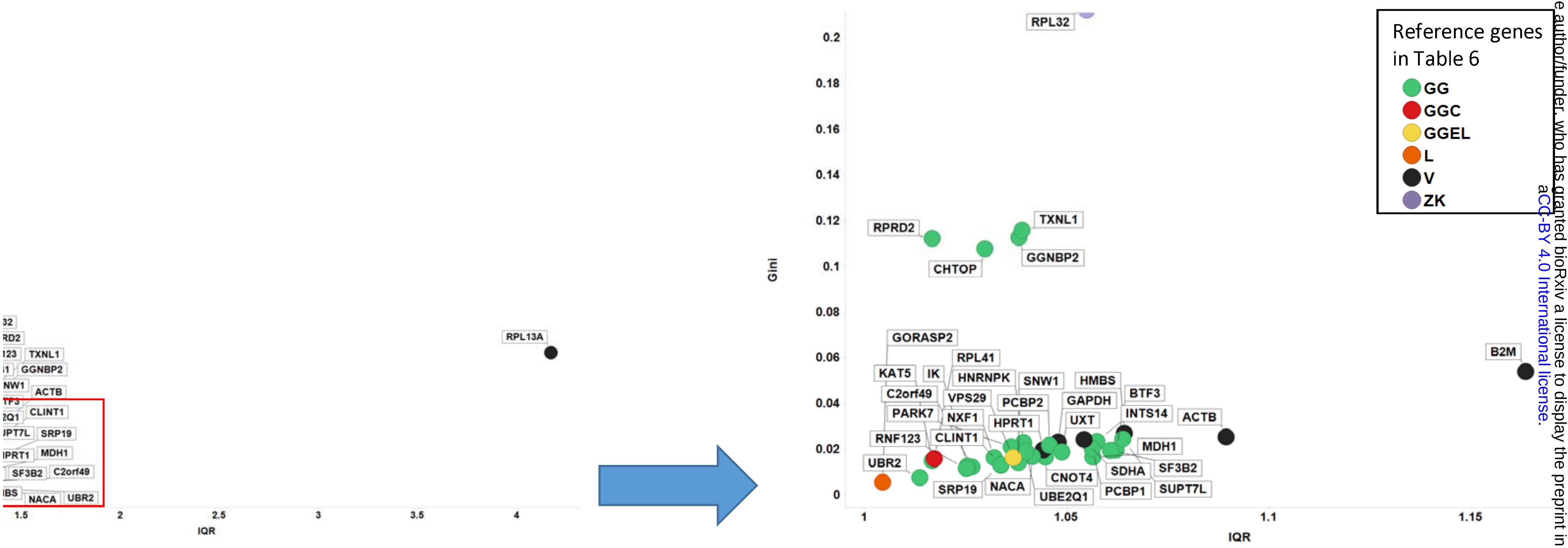
Gini (PCAWG)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Gini (HPA Tissue)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
MB
EPCAM
IFI27
HIST1H2BK
C9orf78
MRPS5
PSMB2
SNRPD3
MRPL21
C10orf76
SNRPD2
EEF1G
PARL
ARL6IP4
RWDD1
RPL41
HDGF
ABCB7
FAM32A
C14orf166
EIF4A3
RPS18
PCBP1
MRPL16
RPLP0
FAM96A
HPRT1
CHMP2A
POLR2H
TUBA1B
PMF1
NOP10
ACTG1
ABHD18
MRPS24
YWHAZ
RPS29
RPS27
B2M
JTB
CNIH1
RNF181
GANAB
RPL13A
ACTB
GAPDH
NDUFS3
GABARAP
GPX1
RPL8
TMEM147
RPSA
NDUFB11
LAMTOR5
HEBP2
PSMB4
VCP
ECHS1
MIF
AIMP1
SDHAF2
TERF2IP
C19orf43
C1orf43
POMP
REEP5
HMBS
SOD1
RPS25
RAF1
HEBP1
MRFAP1
PRDX1
CRELD2
SDHA
RPL6
TBCB
AMZ2
RRAGA
PGAM1
POLR2I
UBC
GTF3A
GSTO1
GPI
GOLGA1
FIS1
NFX1
PSMC1
NELFCD
GABARAPL2
RAB7A
RNASEK
FAM96B
NDUFB1
VPS29
ANAPC5
AP2M1
SLC25A3
RBX1
NDUFB4
GPN2
TBP
EMC7
NDUFA3
CTBP1

2 C



Gini (GTEx)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Gini (HPA Tissue)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
MB
EPCAM
IFI27
HIST1H2BK
CHMP2A
PCBP1
C9orf78
ARL6IP4
MRPL21
ABCB7
EEF1G
NDUFS3
FAM32A
RPL41
PSMC1
RPL6
RPLP0
C14orf166
SNRPD3
RNASEK
RPL13A
SDHAF2
MRPL16
SNRPD2
EIF4A3
GANAB
RAF1
C19orf43
RPS25
FAM96A
RPS18
ACTB
PMF1
NOP10
PGAM1
RPS27
ACTG1
B2M
GABARAP
TUBA1B
HEBP2
NFX1
CNIH1
ABHD18
RPSA
RPL8
AIMP1
NELFCD
HPRT1
RWDD1
HEBP1
GPX1
VPS29
POLR2H
ANAPC5
MRFAP1
C10orf76
GAPDH
TBP
PSMB4
TERF2IP
AMZ2
MIF
TBCB
GPN2
VCP
RPS29
ECHS1
CTBP1
JTB
CRELD2
NDUFB1
EMC7
SDHA
HMBS
GPI
REEP5
UBC
PRDX1
NDUFB11
C1orf43
SLC25A3
YWHAZ
AP2M1
MRPS24
FIS1
HDGF
NDUFB4
GSTO1
TMEM147
NDUFA3
PARL
MRPS5
GOLGA1
RRAGA
GABARAPL2
SOD1
RBX1
RNF181
GTF3A
POMP
FAM96B
POLR2I
RAB7A
PSMB2
LAMTOR5

3


32
RD2
123
TXNL1
1
GGNBP2
NW1
ACTB
TF3
2Q1
CLINT1
JPT7L
SRP19
IPRT1
MDH1
SF3B2
C2orf49
IBS
NACA
UBR2
RPL13A
1.5
2
2.5
3
3.5
4
IQR
RPL32
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Gini
RPRD2
TXNL1
CHTOP
GGNBP2
GORASP2
RPL41
KAT5
IK
HNRNPK
SNW1
HMBS
C2orf49
VPS29
PCBP2
GAPDH
BTF3
PARK7
NXF1
HPRT1
UXT
INTS14
ACTB
RNF123
CLINT1
MDH1
UBR2
SDHA
SF3B2
CNOT4
SRP19
NACA
PCBP1
SUPT7L
UBE2Q1
B2M
1
1.05
1.1
1.15
IQR
Reference genes
in Table 6
● GG
● GGC
● GGEL
● L
● V
● ZK