What are housekeeping genes?

Chintan J. Joshi[1], Wenfan Ke[2], Anna Drangowska-Way[2], Eyleen J. O'Rourke[2]*, Nathan E. Lewis[1,3,4,5]*

[1] Department of Pediatrics, University of California, San Diego, School of Medicine, La Jolla, CA 92093

[2] Department of Biology and Cell Biology, University of Virginia, Charlottesville, VA 22903

[3] Novo Nordisk Foundation Center for Biosustainability at the University of California, San Diego, School of Medicine, La Jolla, CA 92093

[4] Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093

[5] National Biologics Facility, Technical University of Denmark

* Corresponding authors:

Name: Nathan E. Lewis

E-mail: nlewisres@ucsd.edu

Name: Eyleen J. O'Rourke

E-mail: ejo8b@virginia.edu

Keywords: omics data, systems biology, housekeeping genes

Abstract

The concept of "housekeeping genes" has been used for four decades but remains loosely defined. Housekeeping genes are commonly described as "essential for cellular existence regardless of their specific function in the tissue or organism", and "stably expressed irrespective of tissue type, developmental stage, cell cycle state, or external signal". However, experimental support for the tenet that gene essentiality is linked to stable expression across cell types, conditions, and organisms has been limited. Here we use genome-scale functional genomic screens, bulk and single-cell sequencing technologies to test this link and optimize a quantitative and experimentally validated definition of housekeeping gene. Using the optimized definition, we identify, characterize, and provide as resources, housekeeping gene lists extracted from several human datasets, and 12 other animal species that include primates, chicken, and *C. elegans*. We find that stably expressed genes are not necessarily essential, and that the individual genes that are essential and stably expressed can considerably differ across organisms; yet the pathways enriched among these genes are conserved. Further, the level of conservation of housekeeping genes across the analyzed organisms captures their taxonomic groups, showing evolutionary relevance for our definition. Therefore, we here present a quantitative and experimentally validated definition of housekeeping genes that can contribute to better understanding of their unique biological and evolutionary characteristics.

Introduction

The concept of housekeeping genes has aided theoretical and applied biology and the study of evolution. At the grand level, housekeeping genes can be defined as the minimal set of genes required to sustain life[1]. At the practical level, they can be defined as genes stably expressed in all cells of an organism under normal conditions irrespective of tissue type, developmental stage, cell cycle state, or external signal, or as markers of an organism's healthy biological state[2]. At the evolutionary level, they may allow us to define species and higher taxa-specific genomic features[3–5] and gene functions[5–7] that may drive conservation or change. Thus, knowledge of housekeeping genes can significantly contribute to explorative, basic, and translational studies. However, despite the fundamental and translational utility of the concept of housekeeping gene, for over four decades their definition has remained axiomatic.

Housekeeping genes are often defined as being stably expressed in all cells and conditions[8]. However, expression stability is often tested in a handful of cell types and conditions, and then inferred to be stable across most cell types and conditions. Further, housekeeping genes have been proposed to be essential[2], belong to cellular maintenance pathways[8–12] and be conserved[5,6]. However, expression stability (similar expression across cell types and conditions), function (e.g., belonging to cellular maintenance), essentiality (loss-of-function is lethal), and conservation (in this context, stably expressed and essential across taxa) are four very different properties of a gene (Figure 1). No study has yet formally tested the relationships or potential linkage between these four properties. For example, it is unknown if or how the expression pattern of a gene informs its essentiality or 'housekeepingness' conservation. Thus, experimental support for the implied association between the properties defining housekeeping genes remains to be reported.

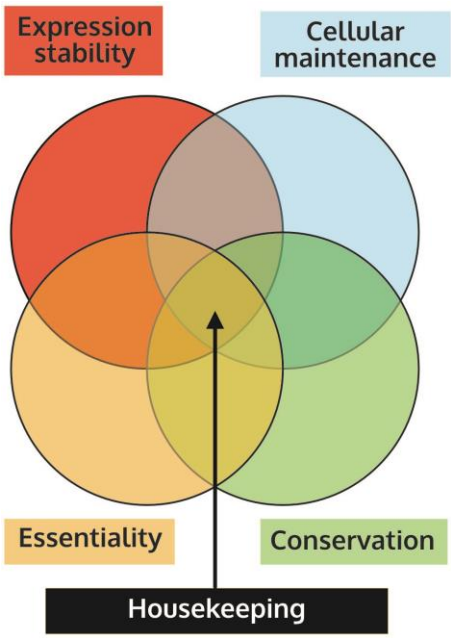


Figure 1. The relationship between the axiomatically assigned four properties of housekeeping genes remains unexplored.

Analysis of large-scale 'omics' data is now commonplace[13], applied to a gamut of questions[14–17] and organisms[18–25]. Further, the availability of single-cell transcriptomics and functional genomics studies for several species is rife with opportunities to study biological principles within and across organisms[23,26,27], including developing a more objective and experimentally validated definition of housekeeping gene.

Predefining housekeeping genes for an organism would bring several potential benefits. At the experimental level, it can save in troubleshooting for the identification and validation of mRNA expression controls in difficult cell types and unique samples (e.g., seasonal wild organisms or patient biopsies) analyzed via transcriptomics[12] and quantitative real-time PCR (qRT-PCR)[28], and it can provide more robust ways to normalize the growing number of single-cell RNAseq studies. At the more fundamental level, it is important to define the minimum number of genes necessary to sustain life and if and how these genes or their characteristics change across taxa.

Recently, we presented StanDep[29], a pipeline for constructing context-specific metabolic network models. StanDep effectively captures metabolic housekeeping genes, defined as genes expressed in most of the analyzed contexts (e.g., tissues, cell lines, etc). The capacity of StanDep to capture housekeeping genes – defined by expression stability – can be attributed to its effectiveness at capturing transcriptomic variability between different samples. Other recent efforts have been made to quantitatively define housekeeping genes [8,28,30,31]. A particularly powerful approach is a mathematical framework called GeneGini[7,30,31], which leverages the Gini coefficient (G_C) – a statistical metric quantifying inequality among groups[32]. G_C varies from 0 to 1. In economics, lower Gini coefficients mean lower income inequality. Similarly, in the framework of GeneGini, the G_C of a gene is proportional to the inequality in its expression across samples[30]. Therefore, genes with a low G_C (referred to as Gini genes here on) are stably expressed and could be considered housekeeping genes according to the property of stability. However, many questions remain about Gini genes. Which are the cellular functions carried out by Gini genes? Are Gini genes essential? Do Gini genes retain their housekeeping status across species? Answers to these questions are central to validate or reject the hypothesis that all four properties – expression stability, basic cellular function, essentiality, and conservation – are linked, and hence should be used to define housekeeping genes.

Here, we used the G_C approach to identify stably expressed genes (Gini genes) across human tissues and cell lines, and across cell types of lower organisms studied using single-cell transcriptomics[19,22,33–36]. We show that, indeed, G_C values are highly correlated across human datasets, supporting the existence of a subset of genes stably expressed across cell types and conditions. Then, we compare the properties of stability (G_C coefficient) to the property of essentiality, as defined by *in vitro* or *in vivo* functional genomics studies. Published CRISPR-Cas9 essentiality screens of Chinese-hamster ovary cells (CHO)[37] and human cell lines[38–40], and new whole-animal essentiality RNAi screening of *C. elegans* done as part of this study, show that essential genes tend to have lower G_C, suggesting that stably expressed genes are more likely to be essential. Nevertheless, not all stably expressed genes are essential. Further, genes that are stably expressed and essential are associated with GO terms enriched in basic cellular functions, and are more likely to be conserved. Therefore, although the four properties informally used to define housekeeping genes are more likely than not to associate with each other, they are not strictly linked. Thus, our analysis provides an experimentally quantitative definition of housekeeping gene, and it establishes a foundation to start asking the most interesting questions including which is the minimum set of genes necessary to sustain life, and if and how these genes change across taxa.

Results

G_C-defined housekeeping genes are stably expressed across samples of the same species

To quantify G_C across genes and test whether the G_C captures the first property of housekeeping genes (i.e., stable expression across cell types and conditions), we calculated the G_C for all genes in the Genotype-Tissue Expression (GTEx)[34] and Human Protein Atlas (HPA)[22] tissue datasets, and all genes in the NCI-60 Cancer datasets from CellMiner and Klijn et al. We used the 3688 housekeeping genes

published by Eisenberg and Levanon [8] as our benchmark gene set. A total of 15687 genes were present in HPA and GTEx, 16052 genes were present in both NCI-60 cancer datasets, and 14327 genes were present in all 4 datasets. From each dataset, we extracted the 3688 genes that had the lowest G_C, this sample-independent gene set allows us to account for the different distribution shapes and number of genes in each dataset (Fig. S1A).

The observations follow: 1) the Gini genes obtained from combining HPA and GTEx (human tissues) covered 81.4% of the 3688 Eisenberg and Levanon housekeeping genes and 2) the NCI-60 cancer datasets covered 69.8% of the 3688 housekeeping genes (**Figure 2A**). Lower accuracy in the human cancer datasets relative to the human tissue datasets is expected as Eisenberg and Levanon housekeeping genes were defined using healthy human tissue transcriptomics[8]. Nevertheless, the G_C for genes present in any given pair of datasets were highly correlated, and more so for datasets of the same sample types (healthy or cancer samples)(**Figure 2B**). Further, we found that for all datasets, Eisenberg and Levanon's housekeeping genes had low G_C values (Fig. S1B), and cross comparison of G_Cs revealed that the median G_C of these genes was very similar across samples (Fig. S2).

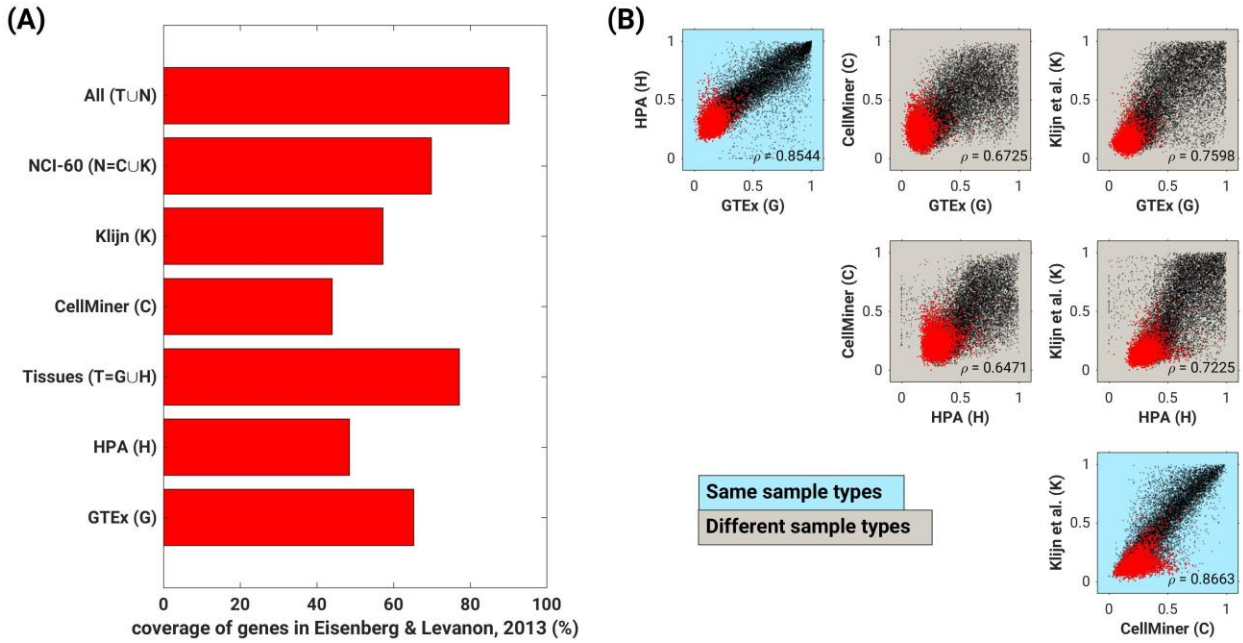


Figure 2. Gini coefficient captures the first property of housekeeping genes: stable expression. (A) Coverage of the 3688 Eisenberg and Levanon housekeeping genes[8] among the 3688 genes with lowest Gini coefficients within each dataset. **(B)** Gini coefficients of individual genes are highly correlated across human datasets regardless of sample type; however, correlations within sample types are tighter. Dots represent unique genes. Red specifically depicts housekeeping genes identified by Eisenberg and Levanon[8].

Although the Gc is informative regarding consistency of expression across samples, it does not inform about the actual level of expression of genes. To investigate whether housekeeping genes are represented across the range of expression or limited to one or other extreme, we used the GTEx[34] and HPA datasets[22] to calculate the median expression levels of the 3688 housekeeping genes previously identified by Eisenberg and Levanon[8]. We observed low median expression for the 3688 genes (**Figure 3 A & B**). Specifically, the median expression value within the 25th, 50th and 75th percentiles of the HPA dataset,

were 12.2, 22.1, and 41.5 transcripts per million (TPM), respectively; whereas within the GTEx data set, the 25th, 50th and 75th percentile values were 14.2, 24.7, and 46.4 TPM, respectively.

In summary, genes with low G_C in one sample type are the most likely to have low G_{Cs} in other sample types or datasets, suggesting that G_C captures the property of expression stability, at least within a given species. Additionally, distinct from current axiom, we found that the majority of the housekeeping genes are expressed at relatively low levels.

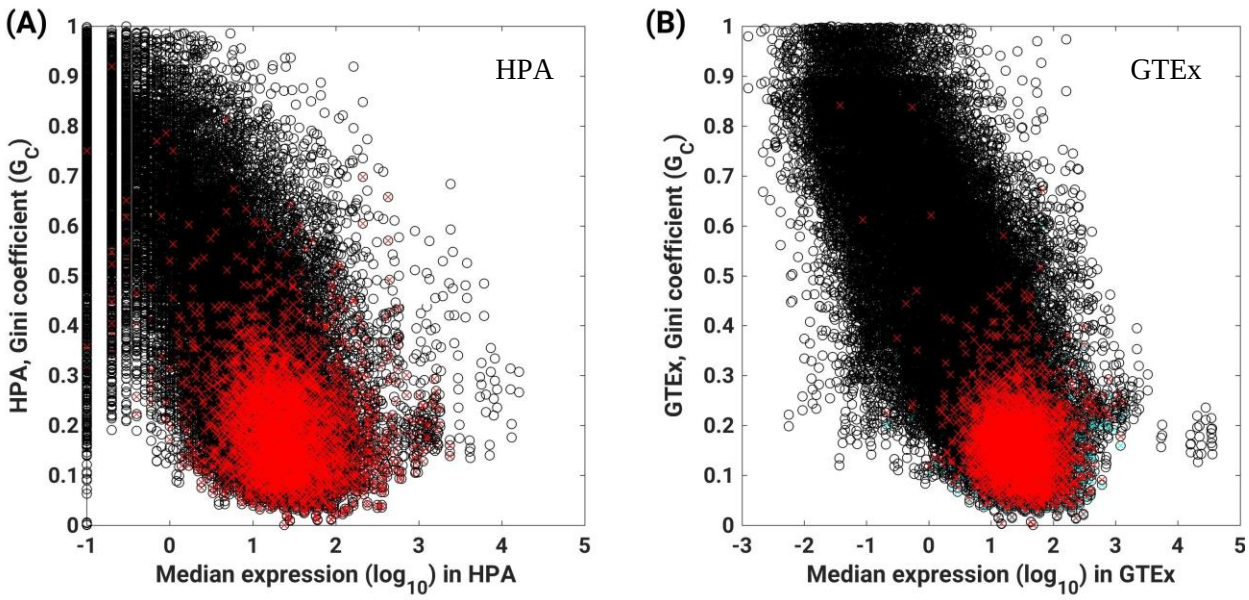


Figure 3. Housekeeping genes are expressed at relatively low levels. (A-B) Comparison of median expression levels and Gini coefficients calculated for the **(C)** HPA and (B) GTEx datasets. Dots represent unique genes. Red specifically depicts housekeeping genes identified by Eisenberg and Levanon[8].

G_C-defined housekeeping genes are enriched in cellular maintenance functions

The second property commonly associated to the concept of housekeeping genes is that these genes are involved in basic cellular maintenance functions[8–12]. To test this axiom, we performed GO term enrichment analysis on the 3688 genes with lowest G_C identified for each dataset as described above.

First, to define the baseline GO terms, we extracted all GO terms enriched in the complete HPA, GTEx, CellMiner and Klijn *et al* samples. We found 1189 different GO terms that were enriched in at least one dataset (Fig. S3, Table S1). To minimize the confounding effect of using different numbers of subject or query genes for the hypergeometric test, we defined as background frequency of a GO term the representation of the compounded 1189 GO terms among all the genes of each dataset. We found that the representation of the 1189 GO terms was highly correlated across datasets (ρ_{mean} = 0.93 across 6 pairs of datasets). Then, we performed the enrichment analysis of the 3688 genes with lowest G_C identified in each dataset (tissue and cancer cell line datasets); details about hypergeometric test can be found in the Methods section.

We found 121 GO terms commonly enriched in all four datasets. The top GO terms with highest coverage were functions classically considered basic cellular maintenance including cell cycle, regulation of mRNA stability, protein folding, protein stabilization, and protein transport, among others. By contrast,

the GO terms that were not enriched were related to stress or other context-specific responses including positive regulation of viral life cycle, response to mitochondrial depolarization, antifungal humoral response, and cellular response to misfolded protein, among others. The nature of the GO terms suggest that the GO terms most enriched among genes with low G_C relate to essential growth functions whereas the least enriched terms relate to regeneration and adaptation to stress rather than generation of new cells.

We also found sample type-specific GO terms; for instance, a GO term that was enriched in tissue data but not in cancer or vice-versa. We found 77 GO terms enriched exclusively in tissue datasets. The top 10 tissue-specific terms included positive regulation of mitochondrial translation and calcium ion transmembrane transport, branched-chain amino acid (BCAA) catabolic process, 7-methylguanosine mRNA capping, and NIK/NF-κB signaling. Excitingly, mitochondrial translation has been identified as a target for various cancers[41,42], BCAA have been identified as an essential nutrient for cancer growth[43]. Further, enrichment of NF-κB highlights how a consistently expressed pathway controls expression of tissue specific genes. Thus, at least 3 out of 4 of the GO terms we find enriched in healthy tissues but not enriched in cancer cells correspond to biological functions known to be impaired or dysregulated in several cancers, suggesting that analyses of sample-specific genes with low Gc could contribute to identify disease-specific genes that could be become targets to treat or cure these diseases. Thus, our list reinforces the notion of enrichment of cellular maintenance functions.

We also found 70 GO terms enriched only in the cancer datasets, suggesting that cancer cells have housekeeping genes which are different than healthy tissues. These 70 cancer-only GO terms included mitotic cell cycle and replication fork machinery, GO terms associated with the generation of new cells. Further, the overall –as opposed to exclusive– coverage of GO terms enriched in cancer datasets was higher than the overall coverage of GO terms enriched in tissue datasets (Fig. S4), supporting the notion that the oncogenic state is an overall gain of function.

In summary, the data show that low G_C is associated with cell maintenance or cell generation activities, supporting the notion that housekeeping genes defined using the G_C would predominantly be involved in basic cellular functions. Beyond, the G_C identifies as housekeeping genes, core genes that are fundamental to homeostasis.

Housekeeping genes are essential

A third property assigned to housekeeping genes is essentiality. However, this has only been validated for a handful of genes circumstantially found to be essential. Since no study has systematically tested whether essential genes are overrepresented among genes with stable expression, we here establish whether genes stably expressed (low Gc) are enriched for essential genes using functional-genomics studies in mammalian cells publicly available, as well as an *in vivo C. elegans* essentiality screen performed specifically for this study.

Essentiality in cancer cell line datasets was extracted from Depmap[38,39] as a CRISPR guide-RNA score (log-fold change of guide-RNA). For essentiality scores, we used a published functional genomics screen that identified 338 essential genes in CHO cells[37] (Table S5). The accession IDs of transcriptomic data for CHO cells are listed in Table S6. We found that the 2796 genes that were essential in all 20 cancer cell lines had lower G_C compared to all the genes combined, when calculated using transcriptomics data for the same 20 cell lines from Klijn et al. and CellMiner (Figure 4A and B). The same was true for 338 essential genes in CHO cells (Figure 4C). To facilitate comparison of GO terms between essential genes and genes with low G_C values, 2796 genes in cancer cells and 338 genes in CHO cells with lowest G_C values were chosen. This analysis of cancer cell lines revealed that coverage of GO terms for the essential genes is correlated with that of same number of low G_C genes identified using Klijn et al. (Figure 4D,

yellow; ρ = 0.8907) and CellMiner (Figure 4D, blue; ρ = 0.8557). A similar comparison between CHO essential genes and genes with low G_C also resulted in a high correlation of ρ = 0.7055 between low Gc and essentiality (Figure 4D, green). Together these results suggest that Gini genes and essential genes show the same distribution, and hence, are likely largely overlapping.

Gene essentiality is: 1) health-status dependent, 2) context-dependent, and 3) subject to buffering from higher-level cell-cell interactions. Thus, since cell lines studied here are immortalized cultured cells devoid of their physiological and multicellular context, we investigated the correlation between G_C and gene essentiality in a healthy living animal model.

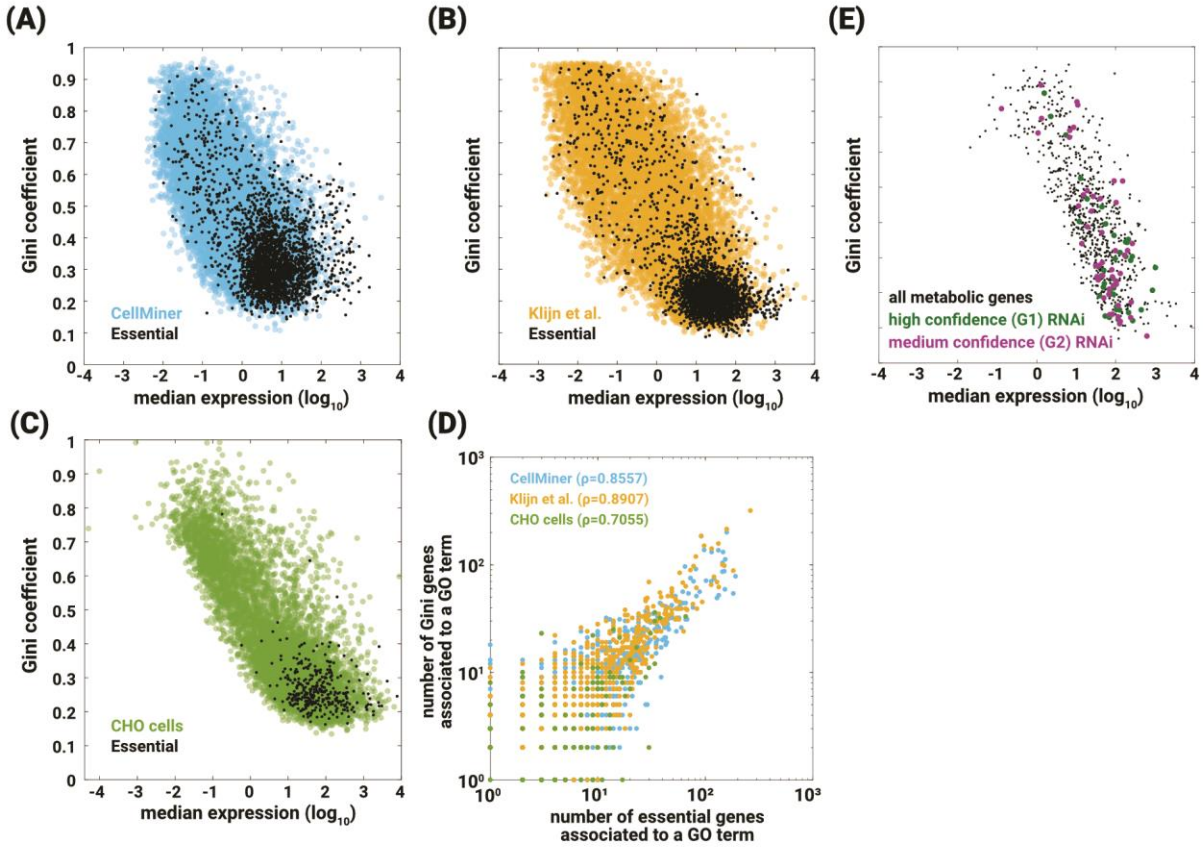


Figure 4. Gini genes are essential. Gini coefficients of essential genes (black dots) compared to the complete (A) CellMiner and (B) Klijn et al. cancer datasets, and (C) CHO datasets. The 2800 genes essential in the 20 cancer cell lines were extracted from DepMap[38,39], and the 338 CHO essential genes were extracted from Xiong et al. (D) GO term coverage of essential genes and that of Gini genes from CellMiner (blue, 0.8557), Klijn et al. (yellow, 0.8907), and CHO (green, 0.7055) are correlated. The slightly lower correlation in CHO cells is likely due to the fewer number of essential genes reported in the CHO screen. (E) High- and medium-confidence essential genes in healthy *C. elegans* have significantly lower G_C than non-essential genes ($p = 4.53 \times 10^{-5}$).

We defined the G_C for *C. elegans* genes using a whole-body single-cell transcriptomics dataset[18]. We obtained essentiality scores from our *in vivo* RNAi screen for all 1535 predicted metabolic genes in the worm (Table S4). A gene was considered essential in *C. elegans* if after rearing hatchlings on dsRNA-delivering *E. coli* for 5 days at 25ºC (*E. coli* is the standard lab food source for *C. elegans*) animals were arrested at a pre-adulthood stage (control WT animals are gravid adults at this time). Three relevant

phenotypic classes were observed in RNAi-treated *C. elegans*: 1) High-confidence essential, RNAi-treated animals arrested in ≥5 out of the 6 independent RNAi treatments against that gene; 2) Medium-confidence essential, RNAi-treated animals arrested in 3 or 4 out of the 6 independent RNAi treatments; and 3) Wild-type; see Supplementary Methods for further details. When testing the association between G_C and gene essentiality, we found that, similar to human and hamster diseased cell lines, high- and medium-confidence essential genes in healthy *C. elegans* had significantly lower G_C than the non-essential genes in the tested pool (Table 1, Table S4).

Geneset	Geneset definition	Number of genes in the class	Number of genes with G_C >0 [(a)]	Sign test (median G_C^{class} < median G_C^{all})
G1	High-confidence essential	48	38	5.808×10^{-5}
G2	Medium-confidence essential	64	49	0.0427
G1 + G2		112	87	4.5304×10^{-5}
G3	Wild-type	1095	532	0.9814
G4	Untested[(b)]	174	97	0.0335
G5	Unknown[(c)]	94	58	0.347

Table 1. *C. elegans* essential genes have significantly lower G_C than non-essential genes. (a) Numbers in this column are smaller than in column C because genes with G_C equal to zero were excluded from the analysis. (b) Untested includes genes that, possibly due to strong effects on health and/or development, prevented us from obtaining large enough populations of worms for quantitative analyses. This hypothesis, and the observations that led us to propose it, are in agreement with the low G_C – essentiality correlation p value observed for this class. (c) Untested corresponds to core metabolic genes that were not tested due to lack of RNAi clone/s or other technical limitations.

In summary, the data show that low G_C is associated with essentiality. Given the strong selective pressure on essential genes, and inferring a similar association in organisms beyond these two mammalian species and a nematode, we hypothesize that the property of essentiality may be conserved across taxa for housekeeping genes defined using the G_C.

Housekeeping genes preserve organism-specific information

One way to test whether the properties of housekeeping genes defined by having low G_C are conserved across evolution is to compare whether the genes with low G_C in one species significantly overlap with the low G_C genes in another species. Importantly, conservation is the last of the four properties commonly assigned to housekeeping genes; yet not systematically tested.

Thus, we analyzed our low G_C genes in the context of a previously published transcriptomics study of nine endothermal organisms[33] that include chicken, platypus, opossum, mouse, macaque, orangutan, gorilla, chimpanzee, and humans. Since most of these organisms do not have a Gene Ontology available, we analyzed the 1:1 orthologs across all these organisms, which corresponded to 5423 genes. Another advantage of limiting the comparison to the 5423 genes is that it removes the bias in the statistical tests otherwise introduced by the sample size.

Different from the analyses of human datasets, we found that the correlations among the nine endothermic species were lower, and the Spearman correlations of the G_C of the 1:1 orthologs distinctively cluster (Figure 5A and 5B). Gini coefficients accurately clustered non-primate mammals distinctively from

primates; thus, reinforcing the value of Gini genes not only to identify conserved housekeeping genes but also distinctive biology.

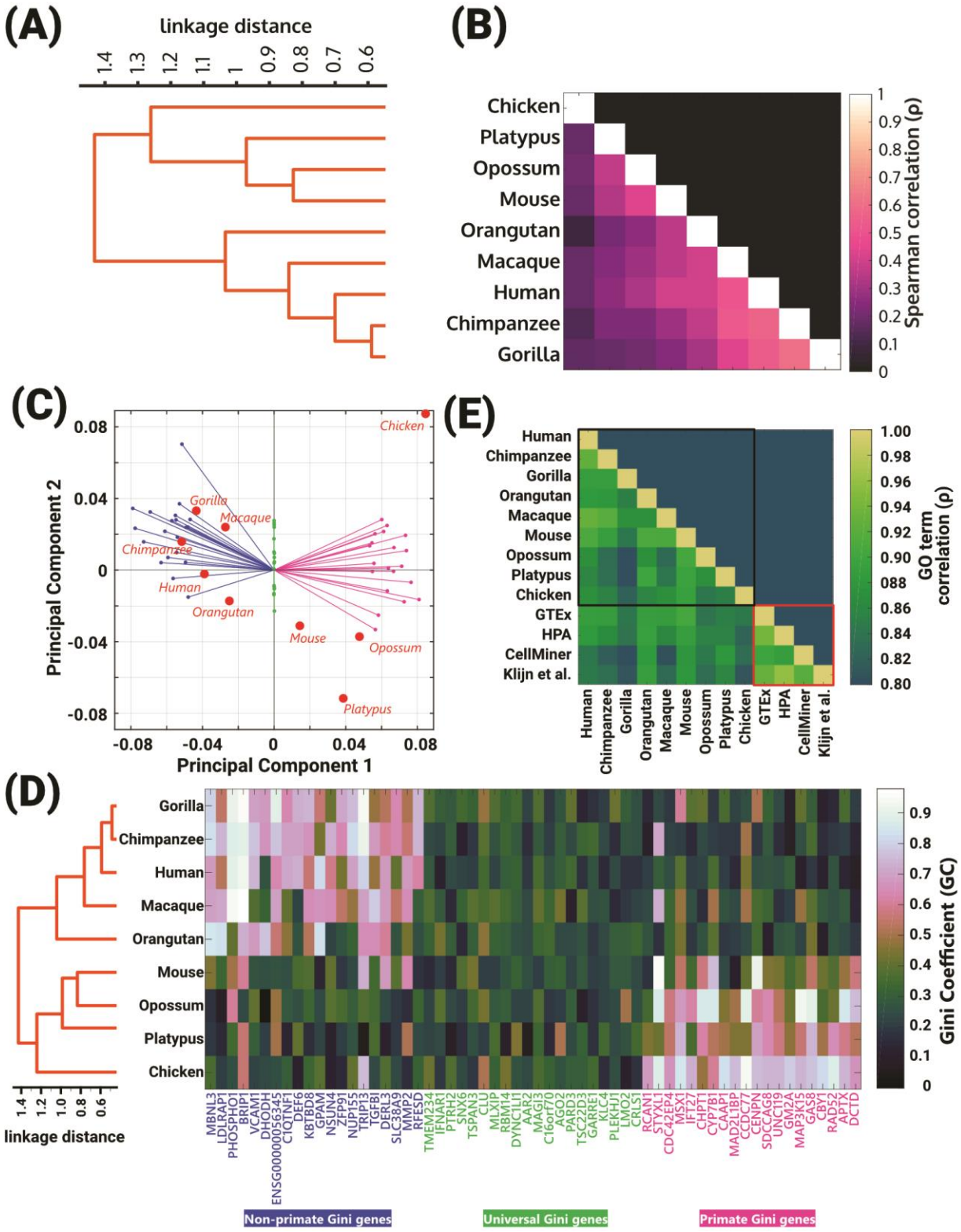


Figure 5. Gini coefficients accurately capture organism-specific differences. (A-B) Jaccard similarity between Gini genes identified using organism-specific transcriptomes capture cluster containing primates. The number of Gini genes with 1:1 orthologs in all organisms is shown using the bar plot on the right of the dendrogram. (C) Principal component 1 (PC1) also captures the cluster containing primates. Also shown are the top 20 primate Gini genes (pink), the top 20 shared Gini genes (green), and top 20 Gini genes in all non-primates (blue) using the principal component coefficients of the first principal components. (D) Correlation among Gini coefficients across different organisms reproduce cluster containing primates (left panel). The Gini coefficients of genes belonging to top 20, middle 20, and bottom 20 coefficients of PC1 are shown (right panel). Left 20 Gini genes are specific to non-primates, middle 20 Gini genes are shared, and right 20 Gini genes are specific to primates. (E) GO term coverage is highly correlated across different datasets, also shown are the GO term correlations with human datasets used in Fig. S6.

Principal Component Analysis (PCA) of G_C for 1:1 orthologs was able to reproduce the cluster consisting of primates and the cluster consisting of other organisms (Figure 5C, dendrogram) using the across taxa correlations among Gini genes. The first two principal components accounted for 45.4% of the explained variation (Fig. S5). The first principal component separated the primate cluster from the remaining species. Our clustering also revealed top candidate genes responsible (based on their G_C values) for such a clustering. The genes with stable expression across all species (Figure 5D, green), primates only (Figure 5D, pink), or non-primates only (Figure 5D, blue). Interestingly, despite the taxa-specific biology, the coverage of GO terms (Table S2) associated with the Gini genes was highly correlated across all the organisms (Figure 5E (black box)). Together, the results show that individual Gini genes contain important information about species-specific biology, yet higher-level features –such as GO terms– are shared by Gini genes across evolution.

Discussion

Historically, housekeeping genes have been defined as genes that are consistently expressed across tissues, essential, belonging to cellular maintenance, and conserved across species. Extending from this definition, genes qualified as "housekeeping" genes have been extensively used for benchmarking and normalizing gene expression results in diverse experimental settings, including qRT-PCR, bulk and single-cell transcriptomics, *in situ* hybridization, western blots, FACS, etc. This has led to usage of genes such as Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) becoming commonly used housekeeping genes. Interestingly, our analyses of GAPDH found large differences in its G_C values across different datasets (Fig. S7). Thus, we set to use Gini coefficient (G_C) as a statistical metric to identify inequality in gene expression across different samples.

The Gini coefficient (G_C) varies between 0 and 1. It quantifies consistency in expression of a gene such that genes expressed in a few conditions or sample types will have a very high G_C (closer to 1) while housekeeping genes, by definition, will have lower G_C (closer to 0). However, the threshold G_C value below which the genes can be classified as housekeeping has not yet been established. The G_C value for any given gene depends on its level of expression, but expression depends on the context (cell type, age, stressors etc). Since it is impossible to enumerate all the possible environments an organism, and hence its cells, will encounter, it is impossible to conclusively find gene expression patterns, and unambiguous G_C values to define "housekeepingness". Thus, a challenge going forward lies in identifying G_C thresholds or other formulations which can be used to calculate an exact list of housekeeping genes for a given organism. Nevertheless, our comprehensive analyses yielded experimentally supported lists of Gini genes (genes with low G_C) for different species (Table S3) and, most importantly, we here formalize, for the first time, the definition of housekeeping gene.

In addition to being stable expressed, housekeeping genes have been axiomatically defined as involved in basic cellular functions and essential. Here we use experimental data to test these notions, and show that indeed Gini genes are enriched in GO terms related to cellular maintenance, and are more likely to be essential. Interestingly, for cancer cells, neither the essential genes nor the enriched GO terms fully overlap with those of healthy tissues of the same species suggesting the definition of housekeeping gene validated here can help distinguish health from disease status and possibly aid in the identification of targets for the treatment of disease. Specifically, our comparison of GO enrichment between healthy human tissues and cancer cell lines highlighted the shift from regeneration of cells in healthy tissue to generation of new cells in cancer cells. Further, our analyses found that coverage of GO terms that were distinctly enriched in only tissue datasets was lower than those for cancer datasets (Fig. S4). Thus, rather than cancer cells undergoing dysregulation of certain housekeeping genes, our analysis suggests that cancer has a unique set of housekeeping genes that differs from the one in healthy tissues. Further, these "cancer-enriched" housekeeping genes are associated to cellular division or cellular growth. This contrasts with the notion that housekeeping genes belong to cellular maintenance; and therefore, this criterion may need to be relaxed when defining housekeeping genes. Our analyses also suggest that cellular maintenance (non-growth needs of the cell) and the growth requirements for cellular division are complementary to one another; and their distinction may not lie in the consistency of expression of a gene.

Another axiomatically accepted property of housekeeping genes is that they are "housekeeping" across species. Our analyses using multi-organism datasets showed high GO term correlation across organisms, suggesting conservation of housekeeping pathways, but no significant conservation of individual housekeeping genes. Further, even for housekeeping pathways, the correlation across species is slightly reduced when compared to the correlation across human datasets (Figure 5E). It is possible that this reduced correlation is due to data limitations. The pathway analyses presented here were done using human GO terms; therefore, genes from any given organism were mapped to the corresponding human orthologs and gene identifiers. In fact, the 1:1 ortholog-based GO term analysis used here resulted in eliminating, on average, 69% of the Gini genes from each of the 9 organisms analyzed. Therefore, availability of gene ontology beyond model organisms can provide molecular insight into species-specific biology. Nevertheless, Gc was able to capture taxa-specific biology.

Until now, the concept of housekeeping was often described using a list of genes. As discussed above, in the framework of G_C, selecting housekeeping genes would require thresholding such that genes which have a G_C below a certain value may be regarded as housekeeping genes. However, thresholding eliminates possibly meaningful information. Indeed, our PCA of genes showed that the genes with the highest variation in expression (39.5%) did not enable clustering the 9 organisms (Fig. S8), suggesting that many of the eliminated genes may have critical functions conserved across species. In this study, we also show that even though fewer GO terms were enriched in all the datasets or organisms, the coverage of GO terms that were enriched in each dataset was highly correlated across datasets. These results suggest that housekeeping functions, rather than a list of genes, better described the state of the organism. This explanation has been suggested previously[44]. To test such a hypothesis, there would be a need to generate computer-based models of these organisms that include multiple levels of regulation. Then, one could possibly test *in silico* the expression of genes across different tissues in different contexts (e.g. reduced dietary protein or infection), and infer which genes show the least variation in expression across conditions (low Gc value). Indeed, this means there is a need for functional models for diverse organisms[45,46].

Key molecular similarities likely underlie the physiological similarities between related species. By crossing Gini coefficients with CRISPR-Cas9 essentiality screens and GO terms we may have captured some of these key molecular similarities as our analysis was able to distinguish primate from non-primate endotherms. On the other hand, even though animals seem phenotypically very different they share molecular similarities that we can capture at the level of GO terms, even if not at the level of specific gene IDs. Nevertheless, what is essential across environmental contexts and taxonomic groups, if anything, is worth future investigation.

Our study only scratches the surface of the answer to practical and fundamental questions and shows the need for organism-specific tools and models; but not just for model organisms, we need models for a diverse set of organisms. Our study suggests that analysis of the ever-increasing "omics" datasets presents an opportunity for better understanding of the biological functions fundamental to sustain life and drive evolution.

Method

Literature search

We performed a literature search using Harzing's Publish or Perish 7[47] to extract the top 1000 hits from Google Scholar for the query keywords: housekeeping, genes, maintenance, and required. The list of top 1000 papers was downloaded to an excel sheet for further analysis and visualization on MATLAB.

Data extraction

Transcriptomic datasets were obtained from various sources (Table 2). To resolve differences in gene identifiers, we mapped all to NCBI Entrez gene identifiers using BioMart, within the Ensembl website. When genes did not map to an NCBI gene identifier, we discarded these genes from the analyses.

Organism (sample type)	**Data source**	**Modifications**
Transcriptomics		
Human (tissues)	HPA[22]	-
	Brawand et al., 2011[33]	Converted to TPM from read per base
	GTEx[34]	
Human (NCI-60 cancer cell lines)	CellMiner[48]	-
	Klijn et al. 2015[35]	-
C. elegans (cell types)	Cao et al., 2017[18]	-
Chicken, Platypus, Orangutan, Bonobo, Gorilla, Chimpanzee, Macaque, Mouse, Opossum (tissues)	Brawand et al., 2011[33]	Converted to TPM from read per base
Chinese hamster (tissues)	Shamie I.S.*, Duttke S.H.*, la Cour Karottki K.J., Han C.Z., Hansen A.H., Hefzi H., Xiong K., Li S., Roth S., Tao J., Lee G.M., Glass C.K., Kildegaard H.F., Benner C., Lewis N.E. A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells. bioRxiv, (2020). DOI: 10.1101/2020.10.09.334045[49]	

Chinese hamster ovaries (cell lines)	See Table S6 for accession IDs	
Essentiality screens		
Human (NCI-60 cancer cell lines) – CRISPR-Cas9	DepMap[38–40]	-
CHO cell lines – CRISPR-Cas9	Xiong et al. 2020[37] and Table S5	-
C. elegans (cell types) - RNAi	Unpublished study provided by Eyleen J. O'Rourke; method described in Ke et al. 2018[50] and Supplementary text.	-

Table 2. Data sources used for this study.

Gini Coefficient (G_C)

The G_C measures the inequality in frequency distribution of a given parameter (e.g., levels of income, income mobility[51], education[52], etc.) compared to the frequency distribution of total population[32]. For analysis of transcriptomic data, the parameter is expression of a given gene and is compared against the total gene expression is distributed across different samples[30]. The G_C is calculated as the ratio of area between the Lorenz curve and line of equality over the total area under the line of equality. The Lorenz curve is the graphical representation of the distribution of a given parameter; and is given by eqn. (1):

$$L(F(x)) = \frac{\int_{-\infty}^{x} t\, f(t)dt}{\mu} \qquad (1)$$

where μ denotes the average, $f(x)$ denotes the probability density function, and $F(x)$ denotes the cumulative distribution function. The calculation was implemented in MATLAB (R2016b), for which the code is available at GitHub (https://github.com/LewisLabUCSD/gene-gini-matlab).

Gene Ontology (GO) enrichment

Due to lack of availability of unique gene ontologies for the different organisms discussed in the study, genes of the organisms that mapped to the human ortholog genes were used to identify the respective GO term. Here, hypergeometric tests were used to check whether the number of genes associated to a GO term, in the query list, are more significant given the distribution among GO terms in the subject gene list. GO terms associated to human genes were downloaded from Gene Ontology Consortium webpage (http://current.geneontology.org/products/pages/downloads.html). All analysis was focused only on the Biological Process (P) aspect. All p-values were calculated using hypergeometric test for overrepresentation reported after correction using the Benjamini Hochberg FDR. A GO-term was considered enriched if the corrected p-values were below 0.05.

Financial Disclosure

This work was supported by the NIGMS (grant no. R35 GM119850, NEL), Novo Nordisk Foundation (NNF10CC1016517, NNF20SA0066621, NEL), a Lilly Innovation Fellows Award to CJ, and funding from the Keck Foundation (EJOR).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interest

The authors have declared that no competing interests exist.

References

1. Koonin, E. V. How many genes can make a cell: The minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* **1**, 99–116 (2000).

2. Tu, Z. *et al.* Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* **7**, 31 (2006).

3. Rach, E. A., Winter, D. R., Benjamin, A. M., Corcoran, D. L. & Ni, T. Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. *PLoS Genet* **7**, 1001274 (2011).

4. Russo, M., Natoli, G. & Ghisletti, S. Housekeeping and tissue-specific cis-regulatory elements: Recipes for specificity and recipes for activity. (2017) doi:10.1080/21541264.2017.1378158.

5. López-Maury, L., Marguerat, S. & Bähler, J. Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nature Reviews Genetics* vol. 9 583–593 (2008).

6. Zhang, L. & Li, W. H. Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. *Mol. Biol. Evol.* **21**, 236–239 (2004).

7. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).

8. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

9. Butte, A. J., Dzau, V. J. & Glueck, S. B. Further defining housekeeping, or 'maintenance,' genes Focus on 'A compendium of gene expression in normal human tissues'. *Physiol. Genomics* (2001) doi:10.1007/s10857-005-4766-0.

10. Zhu, J., He, F., Hu, S. & Yu, J. On the nature of human housekeeping genes. *Trends Genet.* **24**, 481–484 (2008).

11. Warrington, J. A., Nair, A., Mahadevappa, M. & Tsyganskaya, M. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics* **2000**, 143–147 (2000).

12. Thellin, O. *et al.* Housekeeping genes as internal standards: Use and limits. *J. Biotechnol.* **75**, 291–295 (1999).

13. Opdam, S. *et al.* A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Syst.* **4**, 318-329.e6 (2017).

14. Lewis, N. E. *et al.* Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.* **28**, 1279–1285 (2010).

15. Mardinoglu, A. *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 3083 (2014).

16. Shen, Y. *et al.* Blueprint for antimicrobial hit discovery targeting metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1082–7 (2010).

17. Kim, M. K. & Lun, D. S. Methods for integration of transcriptomic data in genome-scale metabolic models. *Comput. Struct. Biotechnol. J.* **11**, 59–65 (2014).

18. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism.

Science **357**, 661–667 (2017).

19. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017).

20. Creecy, J. P. & Conway, T. Quantitative bacterial transcriptomics with RNA-seq. *Curr. Opin. Microbiol.* **23**, 133–40 (2015).

21. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).

22. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, 1260419–1260419 (2015).

23. Barbosa-Morais, N. L. *et al.* The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science (80-.).* **338**, 1587–1593 (2012).

24. Song, Q., Ando, A., Jiang, N., Ikeda, Y. & Chen, Z. J. Single-cell RNA-seq analysis reveals ploidy-dependent and cell-specific transcriptome changes in Arabidopsis female gametophytes. *Genome Biol.* **21**, 178 (2020).

25. Ning, K. *et al.* Transcriptome profiling revealed diverse gene expression patterns in poplar (Populus × euramericana) under different planting densities. (2019) doi:10.1371/journal.pone.0217066.

26. Breschi, A. *et al.* Gene-specific patterns of expression variation across organs and species. *Genome Biol.* **17**, 151 (2016).

27. Yang, Y., Yang, Y.-C. T., Yuan, J., Lu, Z. J. & Li, J. J. Large-scale mapping of mammalian transcriptomes identifies conserved genes associated with different cell states. *Nucleic Acids Res.* **45**, 1657–1672 (2016).

28. Tilli, T. M., Da, C., Castro, S., Tuszynski, J. A. & Carels, N. A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases. (2016) doi:10.1186/s12864-016-2946-1.

29. Joshi, C. J. *et al.* StanDep: Capturing transcriptomic variability improves context-specific metabolic models. *PLoS Comput. Biol.* **16**, e1007764 (2020).

30. O'Hagan, S., Wright Muelas, M., Day, P. J., Lundberg, E. & Kell, D. B. GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles. *Cell Syst.* **6**, 230-244.e1 (2018).

31. Wright Muelas, M., Mughal, F., O'Hagan, S., Day, P. J. & Kell, D. B. The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. *Sci. Rep.* **9**, 718007 (2019).

32. Gini, C. *Measurement of Inequality of Incomes*. *Source: The Economic Journal* vol. 31 (1921).

33. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).

34. GTEx Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–5 (2013).

35. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33**, 306–312 (2015).

36. W.C., R., M., S., S., V. & J., D. CellMiner, a systems pharmacological web-application for the

NCI-60 cancerous cell-lines: Updates, data integration, and translationally relevant results. *Cancer Research* vol. 76 no pagination (2016).

37. Xiong, K. *et al.* Using targeted genome integration for virus-free genome-wide mammalian CRISPR screen. *bioRxiv* 2020.05.19.103648 (2020) doi:10.1101/2020.05.19.103648.

38. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–29 (2016).

39. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).

40. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

41. Škrtić, M. *et al.* Inhibition of Mitochondrial Translation as a Therapeutic Strategy for Human Acute Myeloid Leukemia. *Cancer Cell* **20**, 674–688 (2011).

42. Hu, B. & Guo, Y. Inhibition of mitochondrial translation as a therapeutic strategy for human ovarian cancer to overcome chemoresistance. *Biochem. Biophys. Res. Commun.* **509**, 373–378 (2019).

43. Ananieva, E. A. & Wilkinson, A. C. Branched-chain amino acid metabolism in cancer. *Curr. Opin. Clin. Nutr. Metab. Care* **21**, 64–70 (2018).

44. Zhang, Y., Li, D. & Sun, B. Do housekeeping genes exist? *PLoS One* **10**, (2015).

45. King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515-22 (2016).

46. Ganter, M., Bernard, T., Moretti, S., Stelling & Pagni, M. MetaNetX.org: A website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* **29**, 815–816 (2013).

47. Harzing, A. W. Publish or Perish. *available from http://www.harzing.com/pop.htm* http://harzing.com/pop.htm (2007).

48. Shankavaram, U. T. *et al.* CellMiner: A relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* (2009) doi:10.1186/1471-2164-10-277.

49. Shamie, I. *et al.* A Chinese hamster transcription start site atlas that enables targeted editing of CHO cells. *bioRxiv* 2020.10.09.334045 (2020) doi:10.1101/2020.10.09.334045.

50. Ke, W., Drangowska-Way, A., Katz, D., Siller, K. & O'Rourke, E. J. The ancient genetic networks of obesity: Whole-animal automated screening for conserved fat regulators. in *Methods in Molecular Biology* vol. 1787 129–146 (Humana Press Inc., 2018).

51. Shorrocks, A. Income inequality and income mobility. *J. Econ. Theory* **19**, 376–393 (1978).

52. Thomas, V., Wang, Y., Fan, X. & Bank, W. Measuring Education Inequality: Gini Coefficients of Education. *World* 1–37 (2000) doi:10.1596/1813-9450-2525.

Figure Captions

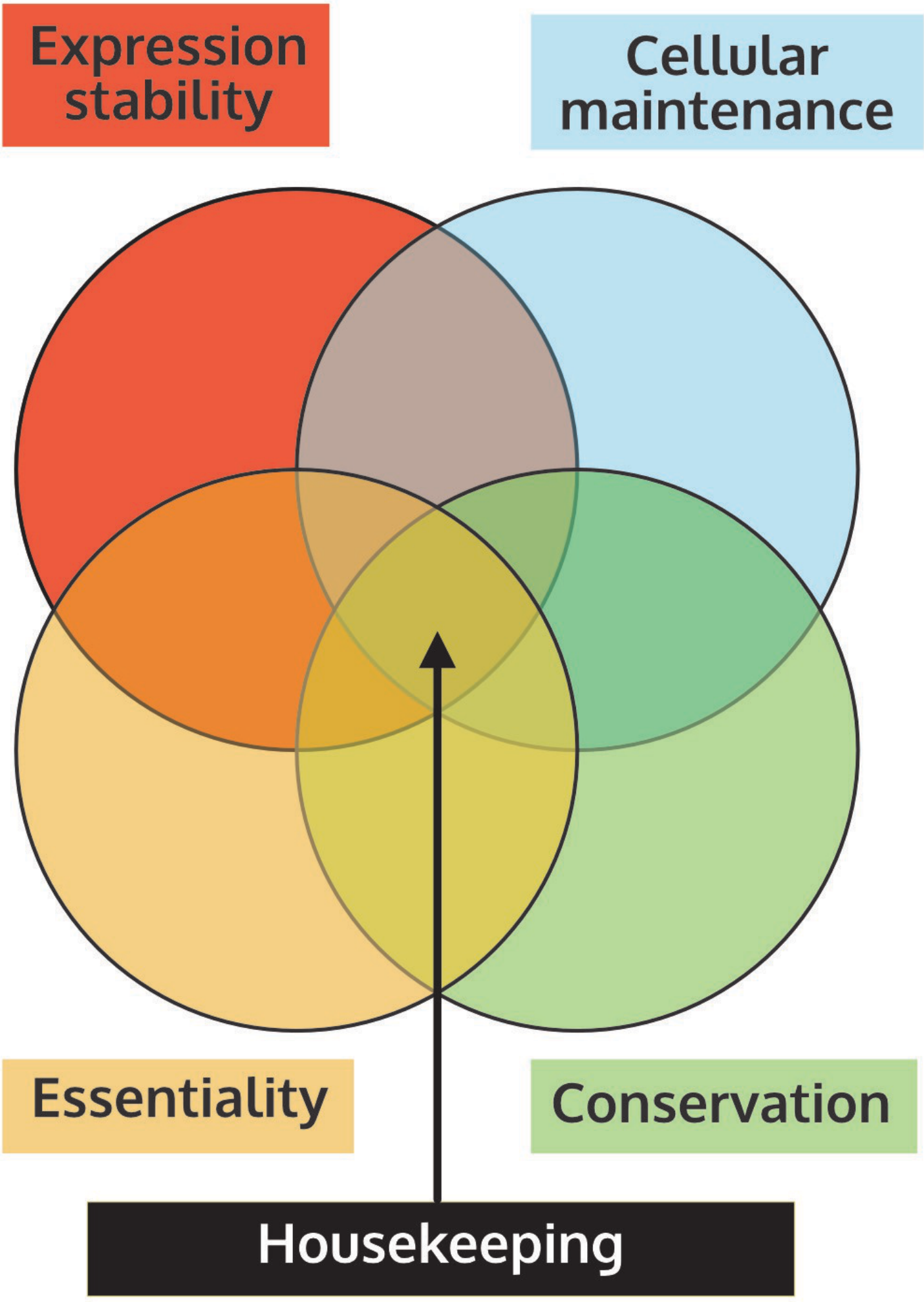
Figure 1. The relationship between the axiomatically assigned four properties of housekeeping genes remains unexplored.

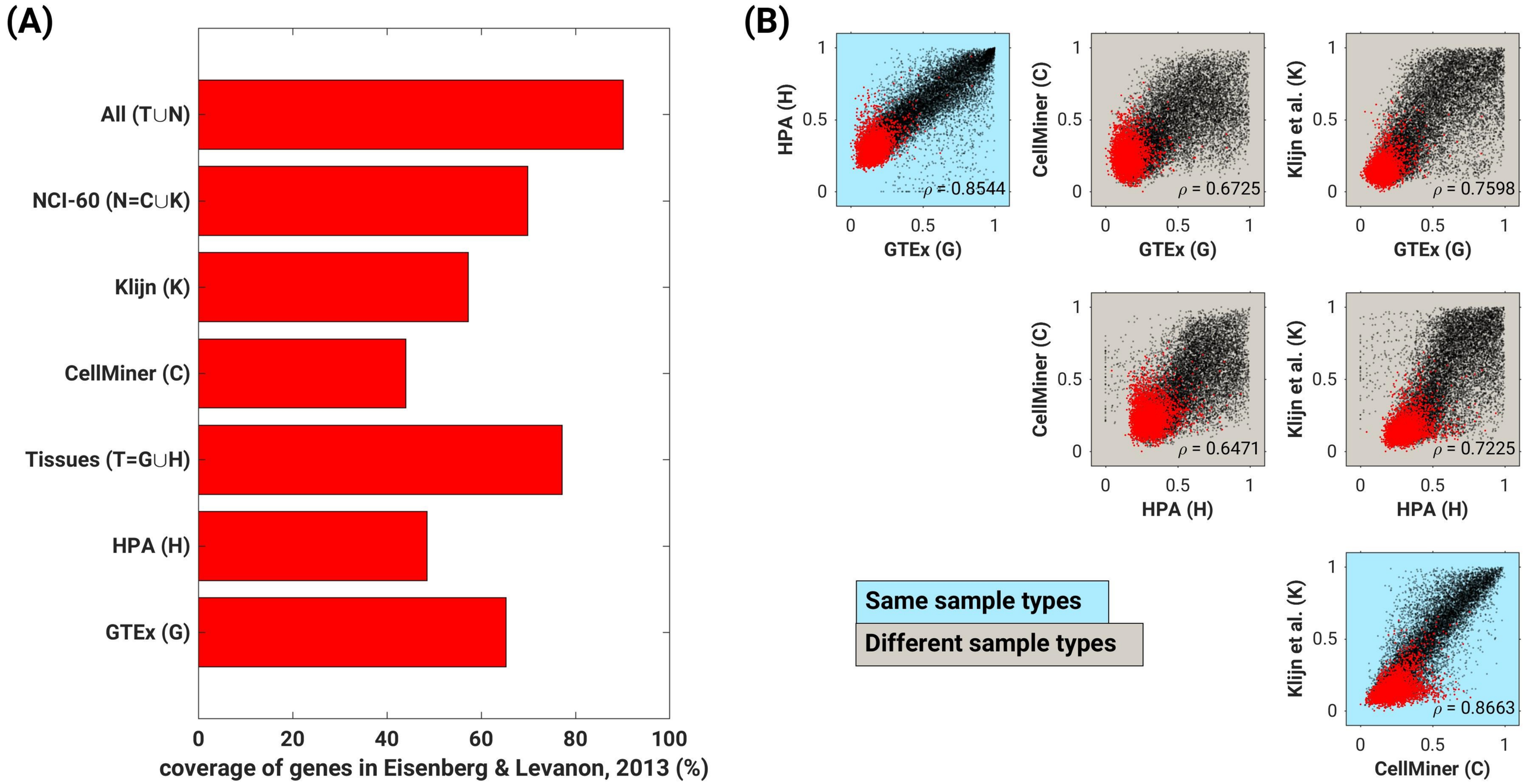
Figure 2. Gini coefficient captures the first property of housekeeping genes: stable expression. (A) Coverage of the 3688 Eisenberg and Levanon housekeeping genes[8] among the 3688 genes with lowest Gini coefficients within each dataset. **(B)** Gini coefficients of individual genes are highly correlated across human datasets regardless of sample type; however, correlations within sample types are tighter. Dots represent unique genes. Red specifically depicts housekeeping genes identified by Eisenberg and Levanon[8].

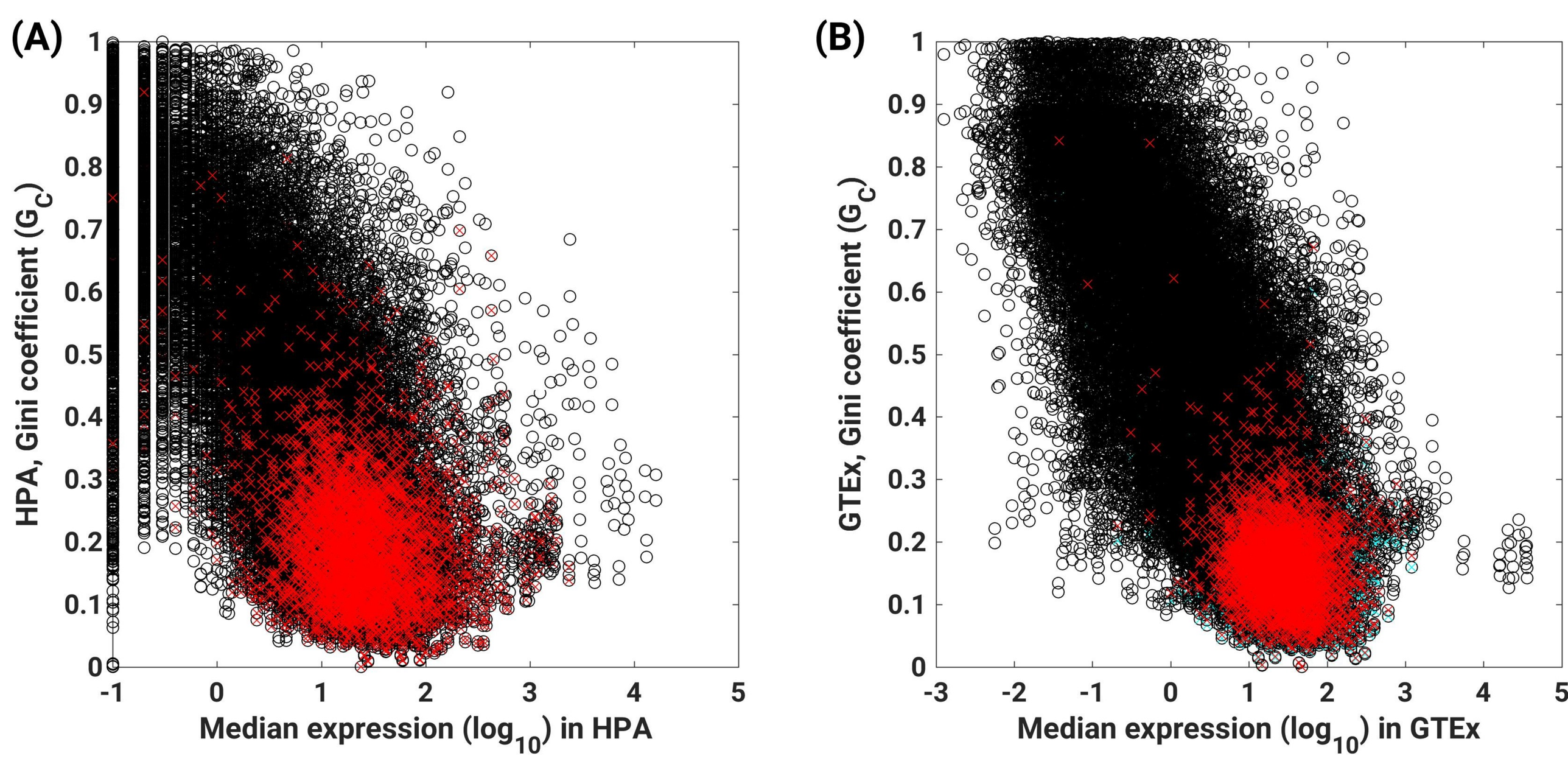
Figure 3. Housekeeping genes are expressed at relatively low levels. (A-B) Comparison of median expression levels and Gini coefficients calculated for the **(C)** HPA and (B) GTEx datasets. Dots represent unique genes. Red specifically depicts housekeeping genes identified by Eisenberg and Levanon[8].

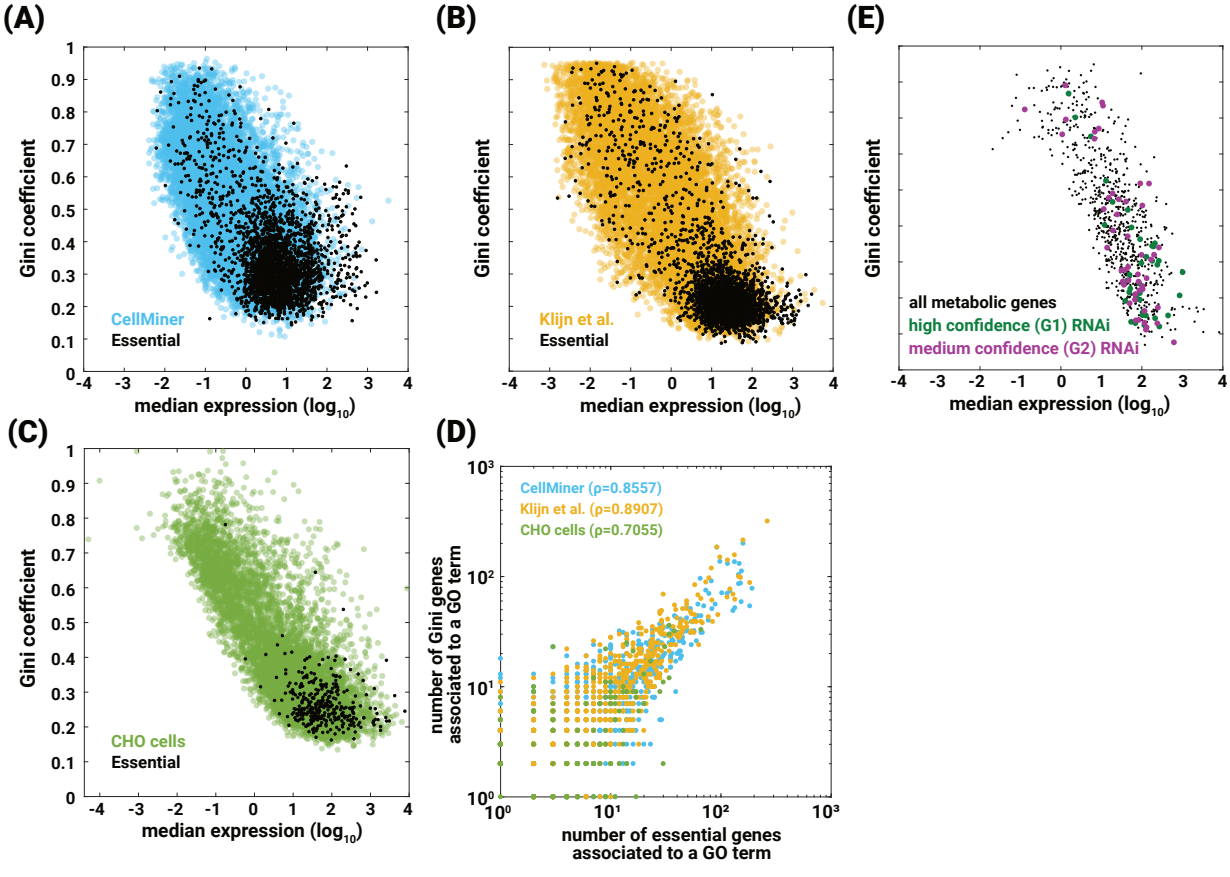
Figure 4. Gini genes are essential. Gini coefficients of essential genes (black dots) compared to the complete (A) CellMiner and (B) Klijn et al. cancer datasets, and (C) CHO datasets. The 2800 genes essential in the 20 cancer cell lines were extracted from DepMap[38,39], and the 338 CHO essential genes were extracted from Xiong et al.[37] (D) GO term coverage of essential genes and that of Gini genes from CellMiner (blue, 0.8557), Klijn et al. (yellow, 0.8907), and CHO (green, 0.7055) are correlated. The slightly lower correlation in CHO cells is likely due to the fewer number of essential genes reported in the CHO screen. (E) High- and medium-confidence essential genes in healthy *C. elegans* have significantly lower G_C than non-essential genes ($p = 4.53 \times 10^{-5}$).

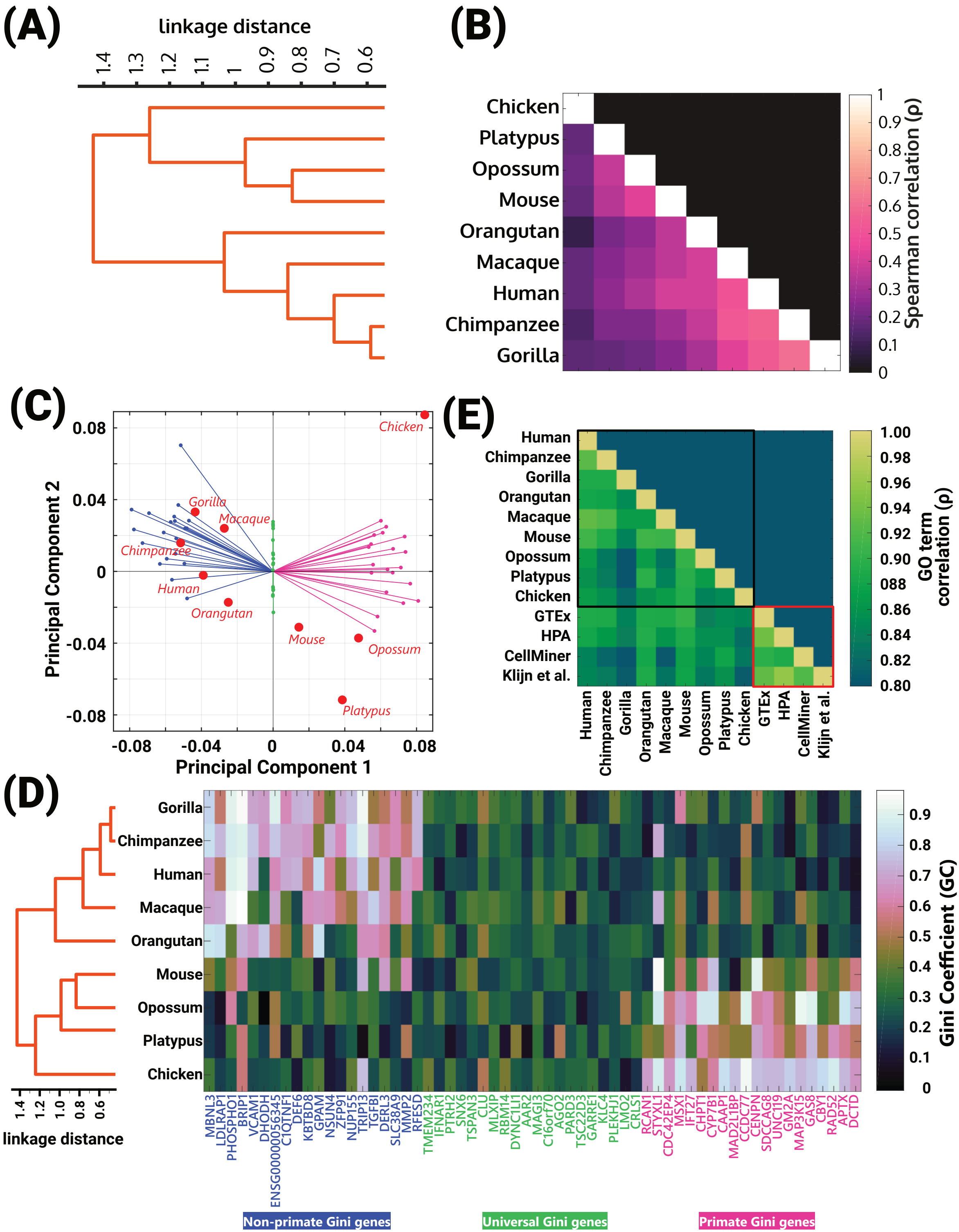
Figure 5. Gini coefficients accurately capture organism-specific differences. (A-B) Jaccard similarity between Gini genes identified using organism-specific transcriptomes capture cluster containing primates. The number of Gini genes with 1:1 orthologs in all organisms is shown using the bar plot on the right of the dendrogram. (C) Principal component 1 (PC1) also captures the cluster containing primates. Also shown are the top 20 primate Gini genes (pink), the top 20 shared Gini genes (green), and top 20 Gini genes in all non-primates (blue) using the principal component coefficients of the first principal components. (D) Correlation among Gini coefficients across different organisms reproduce cluster containing primates (left panel). The Gini coefficients of genes belonging to top 20, middle 20, and bottom 20 coefficients of PC1 are shown (right panel). Left 20 Gini genes are specific to non-primates, middle 20 Gini genes are shared, and right 20 Gini genes are specific to primates. (E) GO term coverage is highly correlated across different datasets, also shown are the GO term correlations with human datasets used in Fig. S6.


Expression stability
Cellular maintenance
Essentiality
Conservation
Housekeeping


(A)
All (T∪N)
NCI-60 (N=C∪K)
Klijn (K)
CellMiner (C)
Tissues (T=G∪H)
HPA (H)
GTEx (G)
0
20
40
60
80
100
coverage of genes in Eisenberg & Levanon, 2013 (%)
(B)
HPA (H)
GTEx (G)
ρ = 0.8544
CellMiner (C)
GTEx (G)
ρ = 0.6725
Klijn et al. (K)
GTEx (G)
ρ = 0.7598
CellMiner (C)
HPA (H)
ρ = 0.6471
Klijn et al. (K)
HPA (H)
ρ = 0.7225
Same sample types
Different sample types
Klijn et al. (K)
CellMiner (C)
ρ = 0.8663
0
0.5
1


(A)
HPA, Gini coefficient (G_C)
Median expression ($\log_{10}$) in HPA
(B)
GTEx, Gini coefficient (G_C)
Median expression ($\log_{10}$) in GTEx


(A)
Gini coefficient
median expression (log10)
CellMiner
Essential
(B)
Gini coefficient
median expression (log10)
Klijn et al.
Essential
(E)
Gini coefficient
median expression (log10)
all metabolic genes
high confidence (G1) RNAi
medium confidence (G2) RNAi
(C)
Gini coefficient
median expression (log10)
CHO cells
Essential
(D)
CellMiner (ρ=0.8557)
Klijn et al. (ρ=0.8907)
CHO cells (ρ=0.7055)
number of Gini genes associated to a GO term
number of essential genes associated to a GO term


(A)
linkage distance
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
(B)
Chicken
Platypus
Opossum
Mouse
Orangutan
Macaque
Human
Chimpanzee
Gorilla
Spearman correlation (ρ)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(C)
Principal Component 2
Principal Component 1
0.08
0.04
0
-0.04
-0.08
Chicken
Gorilla
Macaque
Chimpanzee
Human
Orangutan
Mouse
Opossum
Platypus
(E)
Human
Chimpanzee
Gorilla
Orangutan
Macaque
Mouse
Opossum
Platypus
Chicken
GTEx
HPA
CellMiner
Klijn et al.
GO term correlation (ρ)
1.00
0.98
0.96
0.94
0.92
0.90
088
0.86
0.84
0.82
0.80
(D)
Gorilla
Chimpanzee
Human
Macaque
Orangutan
Mouse
Opossum
Platypus
Chicken
linkage distance
1.4
1.2
1.0
0.8
0.6
Gini Coefficient (GC)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
MBNL3
LDLRAP1
PHOSPHO1
BRIP1
VCAM1
DHODH
ENSG00000056345
C1QTNF1
DEF6
KBTBD8
GPAM
NSUN4
ZFP91
NUP155
TRIP13
TGFBI
DERL3
SLC38A9
MMP2
RFESD
TMEM234
IFNAR1
PTRH2
SNX6
TSPAN3
CLU
MLXIP
RBM14
DYNC1LI1
AAR2
MAGI3
C16orf70
AGO2
PARD3
TSC22D3
GARRE1
KLC4
PLEKHJ1
LMO2
CRLS1
RCAN1
STYXL1
CDC42EP4
MSX1
IFT27
CHPT1
CYP7B1
CAAP1
MAD2L1BP
CCDC77
CENPN
SDCCAG8
UNC119
GM2A
MAP3K15
GAS8
CBY1
RAD52
APTX
DCTD
Non-primate Gini genes
Universal Gini genes
Primate Gini genes