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]*

[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

* Corresponding author:

Name: Nathan E. Lewis

Address: 9500 Gilman Drive MC 0760, La Jolla, CA 92093

E-mail: nlewisres@ucsd.edu

* Co-corresponding author:

Name: Eyleen J. O'Rourke

E-mail: ejo8b@virginia.edu

Classification:

Major: Biological Sciences

Minor: Evolutionary Biology; Biophysics and Computational Biology; and Systems Biology

Keywords: omics data, systems biology, housekeeping genes

Abstract

Two gene classes that have proven useful in understanding the phenotypic states of cells are housekeeping genes and essential genes. Housekeeping genes are often defined as stably expressed in mRNA expression experiments, as essential for cellular maintenance in functional analyses, or both. This imprecise definition can suggest that stably expressed genes are essential for cellular maintenance. Although defining whether there is a relationship between stable expression and essentiality (deleterious if not expressed) would not only aid in the design of experiment controls but could also reveal some fundamental biological principles, this question has not been formally approached. Gini coefficient has been proposed to identify housekeeping genes that we refer to as Gini genes. We use transcriptomics and functional genomics data to identify and characterize Gini genes in several human datasets, and across 12 species, that include human, chicken, and *C. elegans*. We show that Gini coefficients are highly correlated across human tissue and human cancer datasets. We also show that the Gini coefficients of Gini genes that are conserved (1:1 human orthologs) across different organisms can capture taxonomic groups such as primates. We find that essential genes tend to have lower Gini coefficients suggesting that Gini genes may also be essential. Thus, we provide here not only experimental basis for defining housekeeping genes; we also show that these genes capture organism-specific biology.

Significance

Housekeeping genes are considered to be consistently expressed across cell types due to being essential for cellular maintenance. These genes have been known to have unique evolutionary and genomic features, to be markers of organismal health, and for benchmarking gene expression experiments. Here we present the first quantitative experimental support for this definition. We further show that across species the list of housekeeping genes can vary drastically, despite being highly correlated at pathway-level. Finally, we provide a resource and computational pipeline for identifying housekeeping genes and lists of housekeeping genes for 12 different organisms.

Introduction

Analysis of large-scale "omics" data is now commonplace[1], applied to a gamut of questions[2–5] and organisms[6–13]. This situation is rife with opportunities to study the molecular bases of phenotypes and biological principles within and across organisms[11,14,15]. One key feature in all organisms is housekeeping genes. Housekeeping genes are often identified by being stably expressed in all samples/conditions (tissues, environments, cell lines, etc.)[16]. Additionally, the most pervasively used definition invokes essentiality (as in, required or necessary for cell survival)[16–20]. However, stability (similar expression across cell types and conditions) and essentiality (loss-of-function) are two very different features of a gene with different levels of regulation, that manifest at different levels of organization, and have not been shown to be related. Here, we present an experimental basis for this definition, and define housekeeping genes for several species.

Predefining housekeeping genes for an organism brings 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 (i.e. reticulocytes) and unique samples (i.e. patient biopsies) analyzed via transcriptomics[20] and quantitative real-time PCR (qRT-PCR)[21], as well as more robust ways to normalize the growing number of single-cell RNAseq studies. Indeed, one can look for expression levels for these genes, as has been done historically. However, a better list of candidates may be possible if there was a way to systemically identify a large list of these genes from transcriptomics datasets.

At the systems level, housekeeping genes can be defined as the minimal set of genes required to sustain life[22] and markers of an organism's healthy biological state[23]. At the evolutionary level, they may allow us to define organism-specific unique genomic[24–26] and evolutionary features[26–28]. Thus, knowledge of housekeeping genes can significantly contribute to explorative, basic, and translational studies. Despite these and other potential benefits, a list of housekeeping gene candidates for multiple species has not yet been produced.

Recently, we presented StanDep, 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 (tissues, cell types, cell lines, etc.)[29]. The ability of StanDep to capture housekeeping genes can be attributed to its effectiveness at capturing transcriptomic variability among different samples. Other recent efforts have been made to identify housekeeping genes [16,21,30,31]. A particularly powerful approach is a mathematical framework called GeneGini[28,30,31] 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 our work, 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. However, many questions remain about Gini genes. Do these genes retain their housekeeping status across species? Which cellular functions are they responsible for? How essential are they? Answers to these questions are central to the definition of a housekeeping gene and efforts to understand their biological relevance.

Here, we used the G_C approach to identify Gini (or housekeeping) genes in human tissue and cell lines[7,10,33–36]. We show that G_C values were highly correlated across human datasets and that the correlation was higher between datasets of similar samples (e.g. correlation between GTEx and HPA was higher than that between GTEx and CellMiner). We also applied G_C analysis to transcriptomics datasets of 12 different organisms which include those in Brawand et al. [33], humans[10,34,35,37], *C. elegans*[6], Chinese hamster tissues[38] and Chinese hamster ovary (CHO) cells. We show that the list of Gini genes may capture species relevant features and yet maintain a high across-species correlation when analyzed as GO terms. Using CRISPR-Cas9 essentiality screen of CHO[39] and cancer cell lines[40–42], and whole-animal essentiality RNAi screening of *C. elegans*, we show that essential genes tend to have lower G_C. Further, we also show that Gini genes and essential genes significantly overlap in their functions. Further, we provide a list of housekeeping genes with their Gini coefficient for each of the datasets used in this analysis. Thus, our analysis provides an experimental basis for the concept of housekeeping gene.

Results

GAPDH may not be a good candidate as a housekeeping gene

Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) is the most commonly used housekeeping gene to benchmark expression of other genes in qRT-PCR analyses. To define the appropriateness of GAPDH as a housekeeping gene, we started with previously published transcriptomics data belonging to CHO cells, hamster tissues[38], human tissues from Genotype-Tissue Expression (GTEx) project[34] and Human Protein Atlas (HPA)[10], and NCI-60 cancer cells (Klijn et al.[35], and CellMiner[43]). We calculated the G_C for these datasets and then compared the G_C of GAPDH across the 7 datasets.

Low G_C represents low variability in level of expression across tissues or samples, as would be expected for housekeeping genes. However, our analyses indicated that the G_C values for GAPDH were very different across all human and hamster datasets. For instance, Klijn et al. (i.e., NCI-60 cell lines) showed the lowest G_C value was at 14.6 percentile and hamster tissue had the highest G_C value at 57.6 percentile

(Figure 1A). The G_C values in human datasets varied by 18 percentiles while hamster and CHO data differed by 35.3 percentiles. Thus, the high variability in G_C percentiles indicate that GAPDH may not be a good candidate as a housekeeping gene as it is not as stably expressed as generally thought.

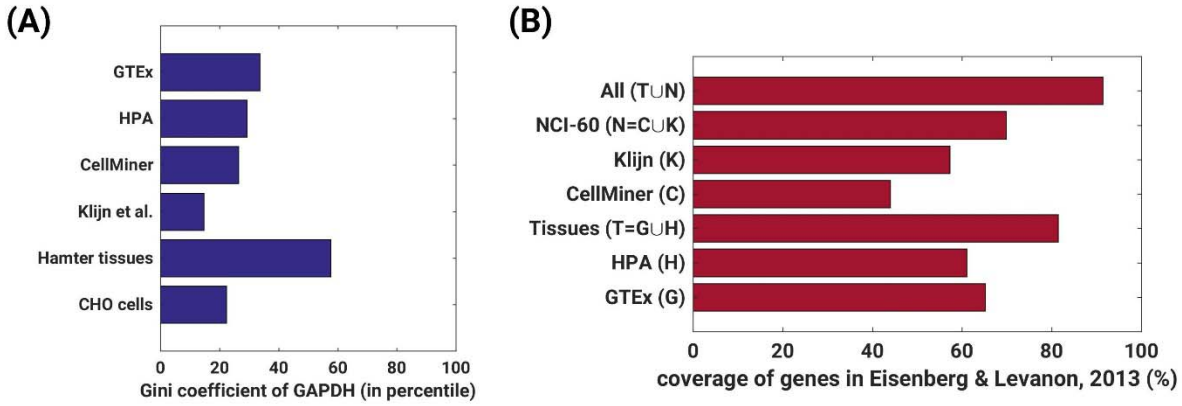


Figure 1. Analysis of previously identified housekeeping genes. (A) Glyceraldhyde 3-phosphate dehydrogenase (GAPDH) may not be a good choice for housekeeping gene. Gini coefficients were converted to percentiles (x-axis) using each of the datasets (y-axis). GAPDH has high Gini coefficient in most of the datasets. (B) Coverage of previously identified 3688 housekeeping genes[16] within the 3688 Gini genes with lowest Gini coefficients within each of the datasets.

Gini coefficient identifies consistently expressed genes across different datasets

Housekeeping genes, by definition, should have low G_C (low inequality across samples). To test whether Gini genes had low G_C, as expected for housekeeping genes, we first tested whether Gini genes are found across two datasets of the same organism. For this we compared the list of Gini genes we extracted from human tissue datasets from HPA and GTEx, and from the NCI-60 Cancer datasets from CellMiner and Klijn et al. We further compared the Gini genes to a previously published list of 3688 housekeeping genes[16]. For this analysis, we chose 3688 Gini genes for each dataset.

15687 genes were present in both human tissue datasets, 16052 genes were present in both NCI-60 cancer datasets, and 14327 genes were present in all 4 datasets. From each dataset, Gini genes were the 3688 genes that have the lowest G_C to account for different shapes of Gini coefficient distributions (Fig. S1A). Gini genes obtained from combining lists from datasets of same sample types had a coverage of 81.4% and 69.8% of the 3688 previously reported housekeeping genes, for tissue datasets and NCI-60 cancer datasets, respectively (Figure 1B). Further, the G_C for genes present in a given pair of datasets were highly correlated, and more so for datasets of the same sample types (Figure 2A). Yet, we found that for all datasets, previously identified housekeeping genes had low G_C values (Fig. S1B).

The lower accuracy of cancer datasets is expected as previously identified housekeeping genes were defined using human tissue transcriptomics. However, since both the datasets are of human cells, this is also reflected in, not only the correlation coefficient but also in the ~70% accuracy of the combined prediction from cancer datasets. In the absence of data to remove batch effects, it is difficult to control for other factors. Cross comparison of G_C of Gini genes revealed that the median of these genes was very similar (Fig. S2). These results together suggest that Gini genes for a given species will consistently have low G_C regardless of sample types or dataset being considered.

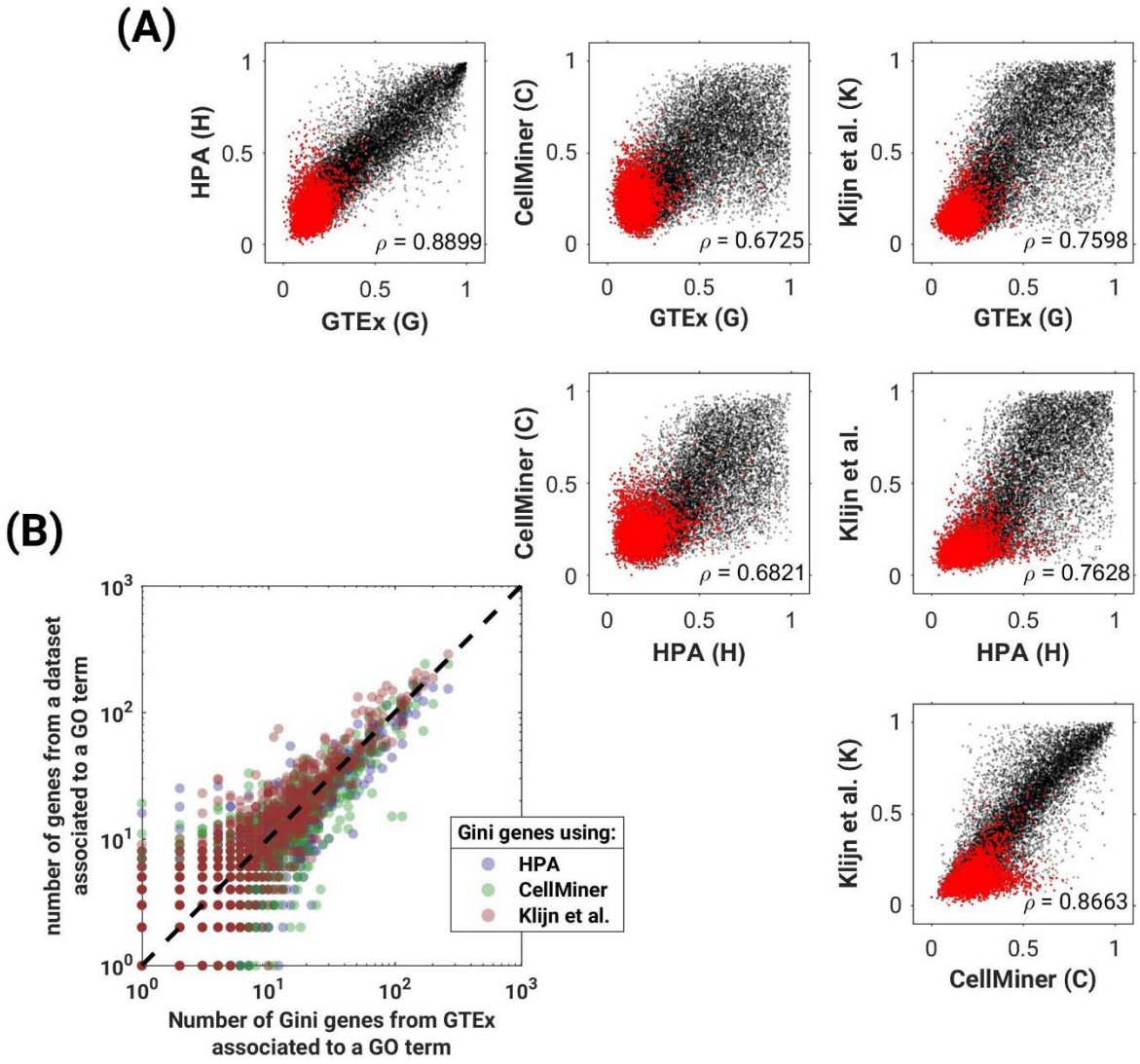


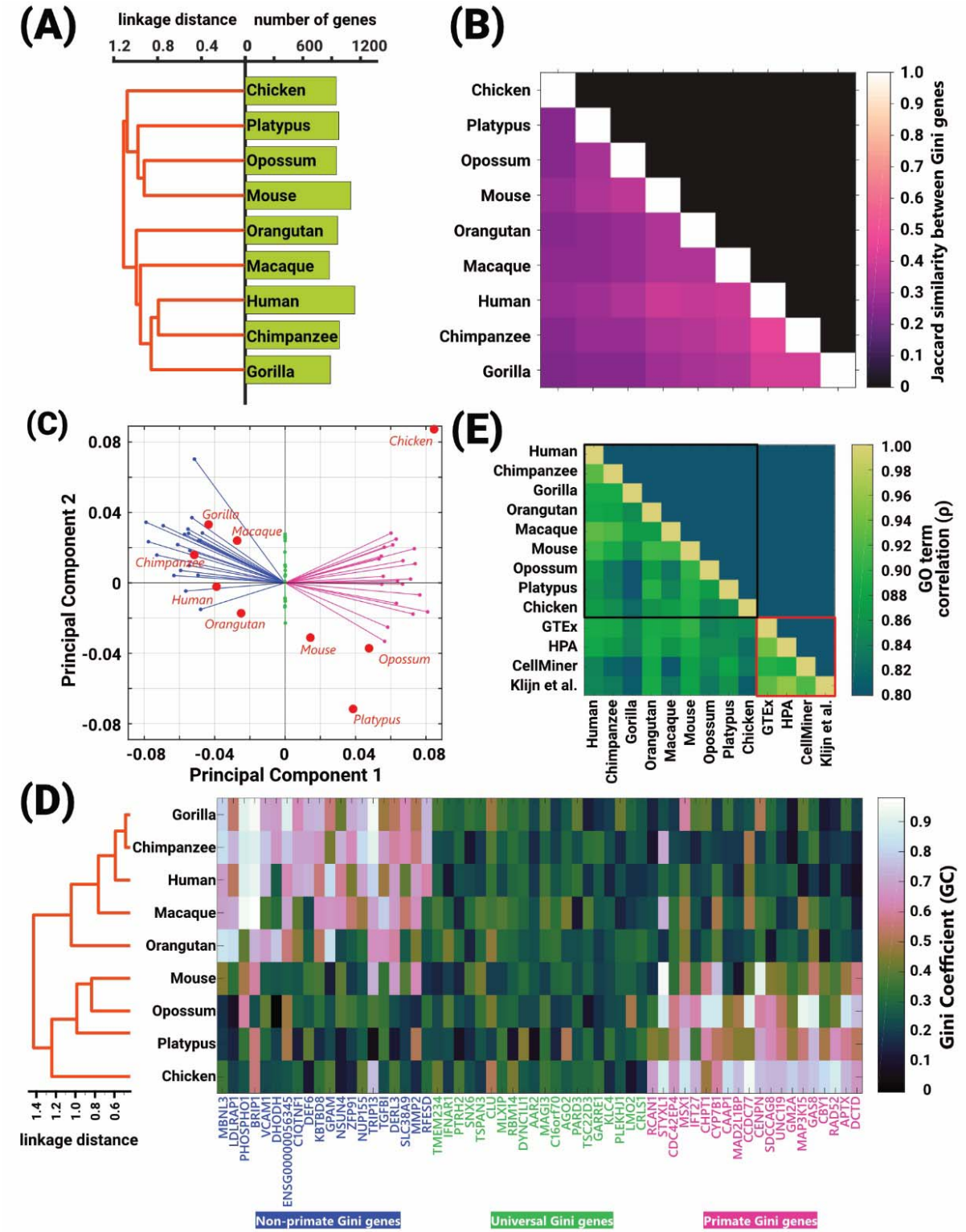
Figure 2. Gini coefficients are highly correlated for human datasets. (A) Gini coefficients of genes are highly correlated across human datasets, regardless of sample type. Datasets of sample types are more highly correlated than those of different sample types. (B) GO term coverage is highly correlated across human datasets.

Gini genes preserve organism-specific information

The correlation of G_C across datasets supports the idea that the Gini genes belong to similar pathways across datasets. To test this hypothesis, we performed GO term enrichment analysis on the Gini genes identified in the transcriptomic analyses just described above.

Across the sample types, we found 1189 GO terms enriched in at least one dataset (Fig. S3, Table S1). To minimize undesired influence from changes in the number of subject or query genes for the hypergeometric test, we analyzed all the 1189 GO terms across all datasets. Background frequency of a GO term was defined using all the genes in each dataset for which G_C was calculated. We found that coverage of these GO terms (ratio of number of Gini genes to the number in the background in a GO term) was highly correlated across datasets (ρ_{mean} = 0.93 across 6 pairs of datasets; Figure 3E (red box)). This comparison is shown for dataset pairs involving GTEx in Figure 2B. These results suggest that human Gini genes are enriched for some biological functions. However, the high correlation in enriched

158 GO terms between these datasets could also be due to datasets belonging to the same organism, i.e.
159 humans.



160
161 **Figure 3. Gini coefficients accurately capture organism-specific differences.** (A-B) Jaccard similarity between
162 Gini genes identified using organism-specific transcriptomes capture cluster containing primates. The number of
163 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 Figure 1.

Next, we adapted and applied our analysis to a previously published 9 endothermal organisms[33] dataset that includes chicken, platypus, opossum, mouse, macaque, orangutan, gorilla, chimpanzee, and humans. Since most of these organisms do not have a Gene Ontology available, we analyzed only the 1:1 orthologs across all these organisms, i.e. 5423 genes. Another advantage is also that one can control for the number of subject genes which will be same as the 5423 genes; effectively removing the influence of sample size in the statistical tests. For the purpose of this exercise, Gini genes were defined by applying a threshold at 17.5 percentile.

Interestingly, we found that Gini genes (Figure 3A and 3B) and correlations of Gini coefficients (Figure 3C) of 1:1 orthologs were able to cluster primate mammals from egg laying - chicken and metronome (platypus), and marsupials - mouse, and opossum. Different from the analyses of human datasets shown above, we found lower correlations among Gini values when comparing the 9 endothermal species than those found amongst human datasets (Figure 3B). Importantly, Principal Component Analysis (PCA) on these data was able to reproduce the cluster consisting of primates and the cluster consisting of other organisms (Figure 3C, dendrogram). The first two principal components accounted for 45.4% of the explained variation (Fig. S4). The first principal component determined the primate cluster; and, at the same time, revealed genes for which expression pattern remained conserved across tissues for all organisms (Figure 3D, green), primates only (Figure 3D, pink), or non-primates only (Figure 3D, blue). Interestingly, similar to human datasets, the coverage of GO terms (Table S2) associated with Gini genes was highly correlated across all the organisms (Figure 3E (black box)). These results together show that Gini genes contain important information about species-specific biology, yet higher-level features, such as GO terms, are shared by all Gini genes.

Gini genes are essential

Considering Gini genes are similar across different organisms, they are likely to also be essential for survival of these organisms. Due to technological and ethical limitations, it is difficult to test all organisms mentioned above. Therefore, we compared gene essentiality in CRIPSR-Cas9 screens with Gini genes of CHO and human cancer cell lines. For cancer cell lines, we used CRISPR guide-RNA score (log-fold change of guide-RNA) from Depmap[40,41]. For CHO cells, we used the list of genes from a published study (Table S6 for accession IDs).

Across 20 cancer cell lines, 72% of Gini genes had a negative CRISPR score; and thus, were essential. We also found that 2800 genes essential in all 20 cancer cell lines also had lower G_C, when calculated using transcriptomics data for 20 cell lines from Klijn et al. and CellMiner (Figure 4A and B). Similar results were observed for 338 essential genes in CHO (Figure 4C). Thus, suggesting that essential genes have lower G_C and are more likely to be Gini genes too.

Since gene essentiality is context and health-status dependent, we investigated the correlation between G_C and gene essentiality in a healthy living animal model. We identified Gini genes in *C. elegans* using a previously published transcriptomics dataset[6] and whole animal RNAi screen of all 1535 predicted metabolic genes in the worm (Ke W. and Drangowska-Way A. *et al*, unpublished). Three relevant

phenotypic classes were observed in RNAi-treated *C. elegans*: 1) high-confidence essential, after 5 days of incubation at 25C animals were arrested at a pre-adulthood stage in ≥5 out of the 6 independent RNAi treatments against that gene; 2) Medium confidence essential, animals did not reach adulthood in ≥3 out of the 6 independent RNAi treatments; and 3) Wild-type (Please see Supplementary Methods for details). Here, too, we found that essential genes (high confidence and medium confidence classes) had significantly lower Gini coefficients than the rest of the tested genes (Table 1, Table S4).

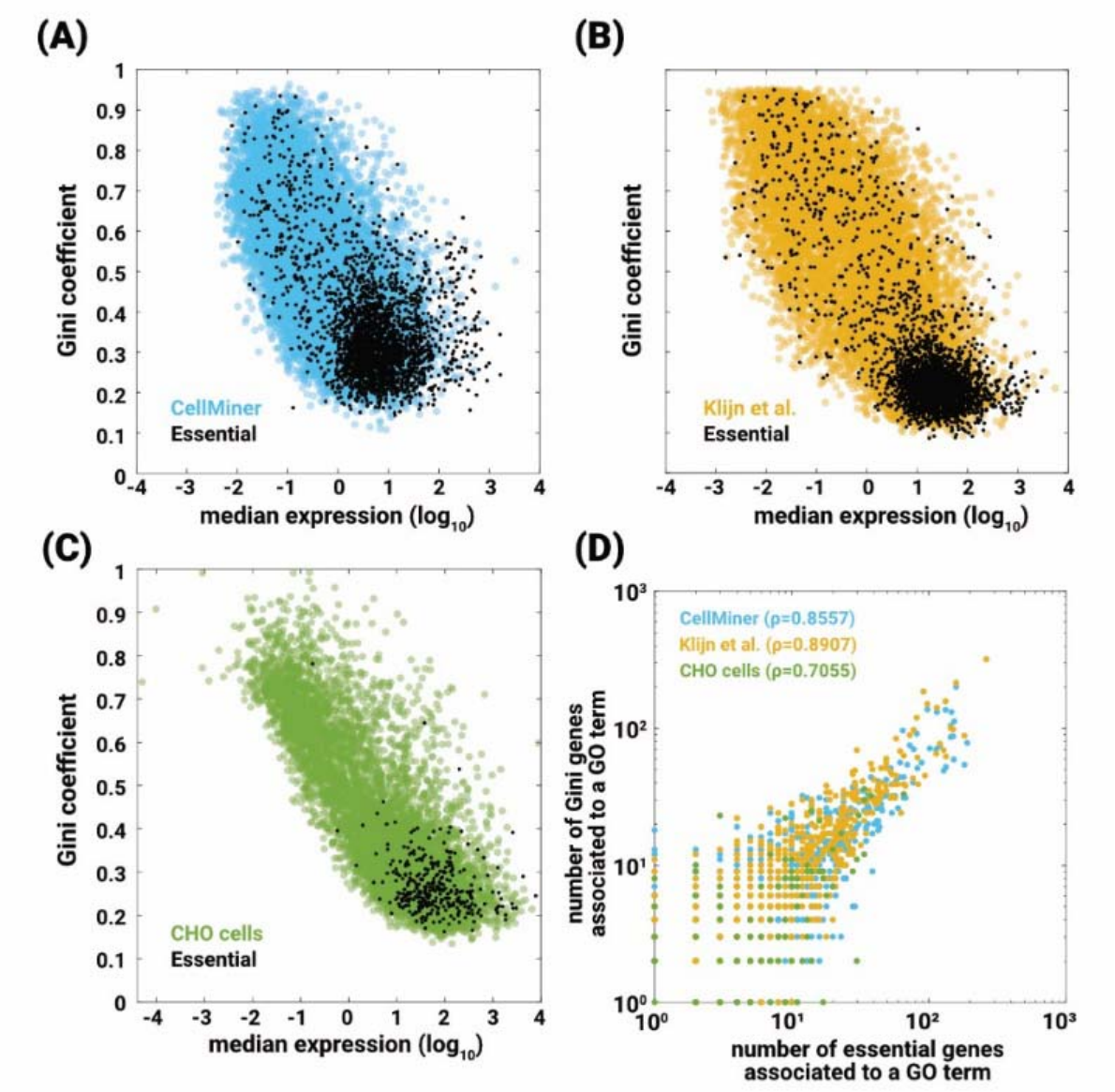


Figure 4. Gini genes are essential. Gini coefficients of essential genes compared to the complete (A) CellMiner, (B) Klijn et al. cancer datasets, and (C) CHO datasets. 2800 genes essential in 20 cell lines were extracted from DepMap[40,41], and 338 CHO essential genes (Table S5) were extracted from Kai et al.[39] (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 fewer number of essential genes in CHO.

Geneset	Geneset definition	Number of genes in each class	Number of genes for which G_C was calculated[(a)]	Sign test (median $G_C^{geneset}$ < 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
G1 + G2 + G4		286	184	9.3050×10^{-5}
G5	Unknown[(c)]	94	58	0.347

Table 1. Whole animal essential genes in C. elegans have significantly lower Gini coefficient 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 with strong effects on health and/or development that prevented us from obtaining large enough populations of worms for quantitative analyses. These observations are in agreement with the low gene essentiality correlation p value observed for this class. (c) Untested core metabolic gene due to lack of RNAi clone or other technical limitations.

GO terms of Gini genes are highly correlated across different datasets and organisms (Figure 2B, Figure 3E). Thus, we also performed GO term analysis for essential and Gini genes in CHO and cancer cell lines. The analysis of cancer cell lines revealed that coverage of GO terms for the 2800 essential genes is highly correlated with that of same number of Gini genes identified using Klijn et al. (Figure 4D, yellow; ρ = 0.8907) and CellMiner (Figure 4D, blue; ρ = 0.8596). A similar comparison between CHO essential genes and Gini genes (at 17.5 percentile) also resulted in a high correlation of ρ = 0.9557 (Figure 4D, green). Together these results suggest that Gini genes and essential genes show the same distribution, and hence, are likely largely overlapping.

Discussion

Historically, housekeeping genes have been defined as genes that are consistently expressed across tissues, and often thought to be essential. Extending from this definition, genes qualified as "housekeeping" are extensively used in 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. Further, housekeeping genes are expected to convey important information about the needs of an organism. However, systematic investigation of whether the underlying biology supports the current definition of housekeeping genes has been mostly lacking (Supplementary Results). Thus, we extensively evaluated the claims of this definition by spanning our analysis across datasets belonging to a variety of organisms. As a result, we provide an experimentally supported list of Gini genes (Table S3) and formalize the evidence in support of the notion that we can call these Gini genes housekeeping genes as they are expressed across tissues and species. Further, we validate the notion that housekeeping genes are enriched in, though not necessarily are, essential genes.

Gini coefficient (G_C) is a statistical metric that allows one to identify inequality in gene expression across different samples. G_C has recently been shown to identify housekeeping genes in human cells[30,31]. However, it remained unclear whether housekeeping genes are "housekeeping" across species. Besides, application to other datasets and organisms, we also study the properties of G_C-identified housekeeping

genes. We referred to these genes with low G_C as Gini genes. The low G_C of these genes suggests that they are more equally, i.e. consistently, expressed across samples. Therefore, we here show that Gini genes are expressed at a wide range of expression levels, they are likely to be essential, and that they share functions across different datasets.

Scientific articles often define housekeeping genes as being required for cellular maintenance. However, they are most often identified through searching for genes with consistent expression across samples. Thus, though two important properties of housekeeping genes are: (i) they belong to cellular maintenance pathways; and hence, (ii) their functions are "essential", to the best of our knowledge, there has not been a study that quantitatively tests the basis for neither of these two implicit, and often explicit, claims. This is despite the reality that the importance of characterizing a list of essential genes has garnered significant interest[44–47]. Hence, here we quantitatively test these claims about housekeeping genes. Firstly, for the claim of essentiality, using CHO and cancer gene essentiality CRISPR-Cas9 screens, we show that the majority, but not all, of consistently expressed genes (Gini genes) are essential. Secondly, we show that there is a high correlation in GO terms derived from Gini genes from different datasets, suggesting that Gini genes are indeed coming from population of genes with similar molecular functionalities, as described by GO terms, across different datasets. However, given the vagueness of the term "cellular maintenance", it is rather arbitrary to decide whether housekeeping genes are preferentially associated to this term.

Another gap in the current understanding of housekeeping genes is whether they are "housekeeping" across species. Gini genes calculated using multi-organism datasets showed high GO term correlation across organisms, suggesting conservation of housekeeping pathways. However, this correlation across species is slightly reduced when compared to the correlation across human datasets (Figure 3E). It is possible that this reduced correlation in due to data limitations. The pathway analyses presented here were done using human GO terms; therefore, genes from any given organism were mapped to corresponding human orthologs and gene identifiers. Of course, this process eliminated many Gini genes. 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.

Until now, the concept of housekeeping was often described using a list of genes; 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 principal component containing the highest variation (39.5%) did not explain Gini values in distinctly some of the 9 organisms (Fig. S5). Gini values for large number of genes were needed to capture species-specific clusters. 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, are better described as the state of the organism. This explanation has been suggested previously[49]. To test such a hypothesis, there would be a need to prepare models of these organisms at multiple levels of regulation that could simulate and quantify an organismal phenotype. Then, one could possibly test, for example, if carbon flux across different tissues of organisms is correlated. Indeed, this means there is a need for standardized models for diverse set of organisms[50,51].

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

298 endotherms. On the other hand, even though animals seem phenotypically very different they share
299 molecular similarities that we can capture at the level of GO terms, even if not at the level of specific
300 gene IDs. Nevertheless, what is essential across environmental contexts and taxonomic groups, if
301 anything, is worth future investigation. Our study only scratches the surface of the answer to these
302 questions and shows the need for organism-specific tools and models; but not just for model organisms,
303 we need models for a diverse set of organisms. Our study suggests that analysis of the ever-increasing
304 "omics" datasets presents an opportunity for better understanding of the biological functions fundamental
305 to sustain life and drive evolution.

306 # Method

307 **Literature search**

308 We performed a literature search using Harzing's Publish or Perish 7[52] to extract the top 1000 hits from
309 Google Scholar for the query keywords: housekeeping, genes, maintenance, and required. The list of top
310 1000 papers was downloaded to an excel sheet for further analysis and visualization on MATLAB.

311 **Data extraction**

312 Transcriptomic datasets were obtained from various sources (Table 2). To resolve differences in gene
313 identifiers, we mapped all to NCBI Entrez gene identifiers using BioMart, within the Ensembl website.
314 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[10]	-
	Brawand et al., 2011[33]	Converted to TPM from read per base
	GTEx[34]	
Human (NCI-60 cancer cell lines)	CellMiner[43]	-
	Klijn et al. 2015[35]	-
C. elegans (cell types)	Cao et al., 2017[6]	-
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[38]	
Chinese hamster ovaries (cell lines)	See Table S6 for accession IDs	
Essentiality screens		
Human (NCI-60 cancer cell lines) – CRISPR-Cas9	DepMap[40–42]	-

CHO cell lines – CRISPR-Cas9	Kai et al. 2020[39] and Table S5	-
C. elegans (cell types) - RNAi	Unpublished study provided by Eyleen J. O'Rourke; method described in Ke et al. 2018[53] 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[54], education[55], 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.

Financial Disclosure

This work was supported by the NIGMS (grant no. R35 GM119850, 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. Opdam, S. *et al.* A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Syst.* **4**, 318-329.e6 (2017).

2. 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).

3. 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).

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

5. 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).

6. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

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

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

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

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

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

12. 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).

13. 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.

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

15. 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).

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

17. 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.

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

19. 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).

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

21. 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.

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

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

24. 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).

25. 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.

26. 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).

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

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

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. Reinhold, W. C. *et al.* CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res.* **72**, 3499–3511 (2012).

38. 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.

39. 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.

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

41. 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).

42. 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).

43. 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.

44. Cacheiro, P. *et al.* Human and mouse essentiality screens as a resource for disease gene discovery. *Nat. Commun.* **11**, (2020).

45. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nature Reviews Genetics* vol. 19 34–49 (2018).

46. Palumbo, M. C., Colosimo, A., Giuliani, A. & Farina, L. Essentiality is an emergent property of metabolic network wiring. *FEBS Lett.* **581**, 2485–2489 (2007).

47. Bartha, I., Di Iulio, J., Venter, J. C. & Telenti, A. Human gene essentiality. *Nat. Rev. Genet.* **19**, 51–62 (2018).

48. Ballouz, S. & Gillis, J. AuPairWise: A Method to Estimate RNA-Seq Replicability through Co-expression. *PLoS Comput. Biol.* **12**, e1004868 (2016).

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

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

51. 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).

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

53. 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).

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

55. 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

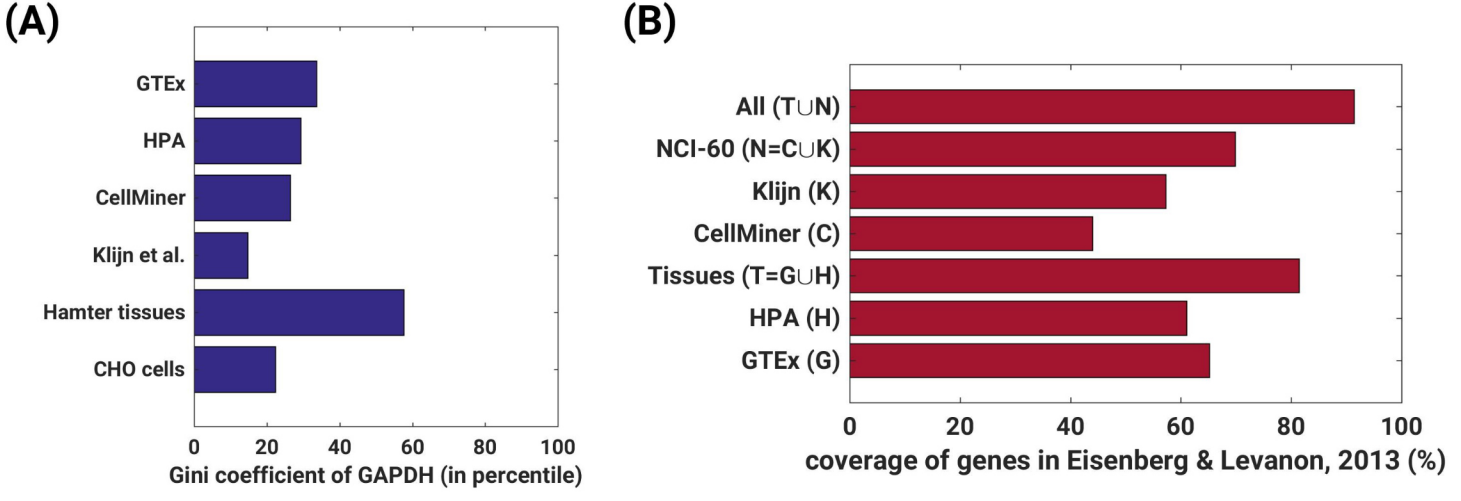
Figure 1. Analysis of previously identified housekeeping genes. (A) Glyceraldhyde 3-phosphate dehydrogenase (GAPDH) may not be a good choice for housekeeping gene. Gini coefficients were converted to percentiles (x-axis) using each of the datasets (y-axis). GAPDH has high Gini coefficient in

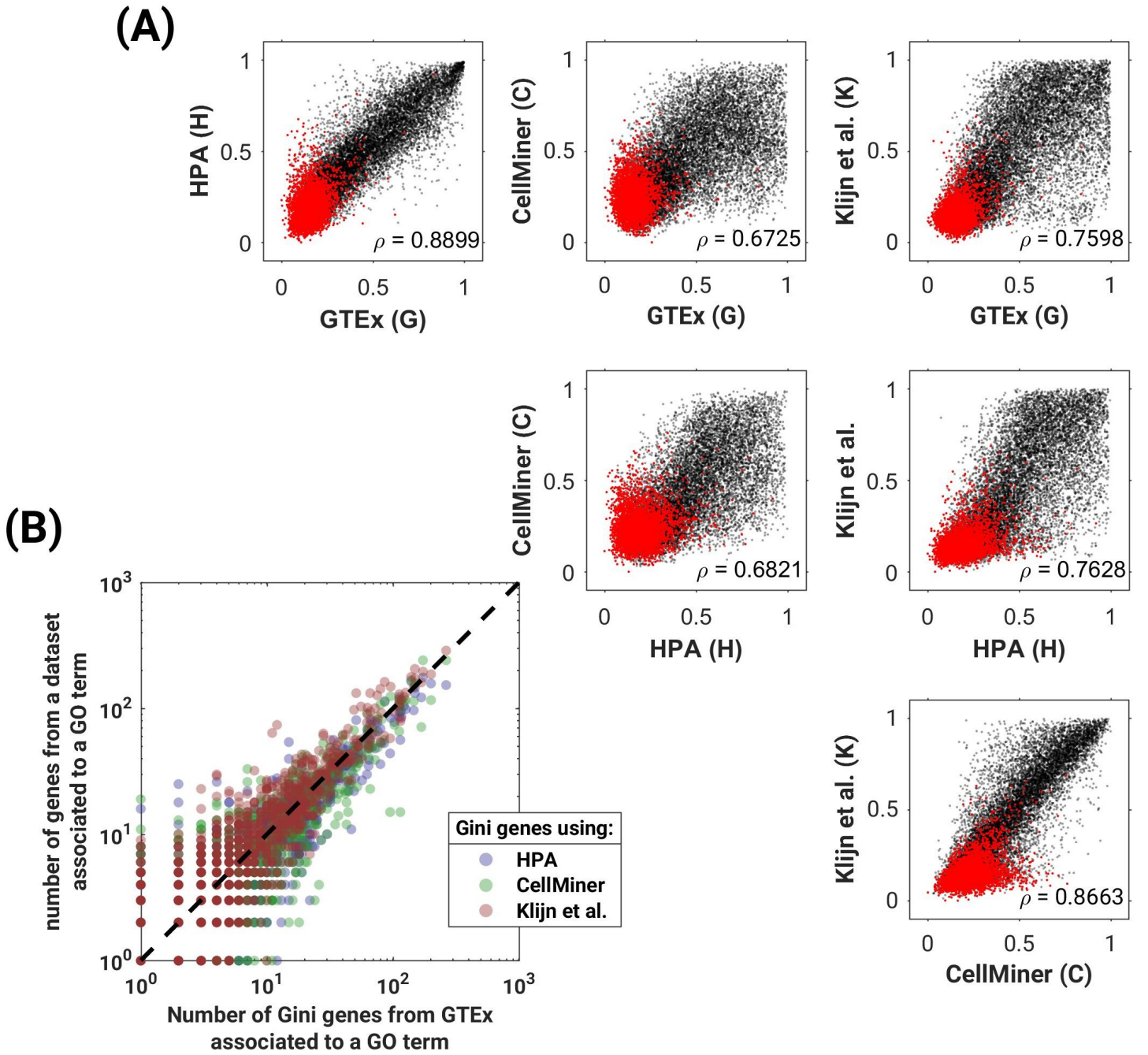
most of the datasets. (B) Coverage of previously identified 3688 housekeeping genes[16] within the 3688 Gini genes with lowest Gini coefficients within each of the datasets.

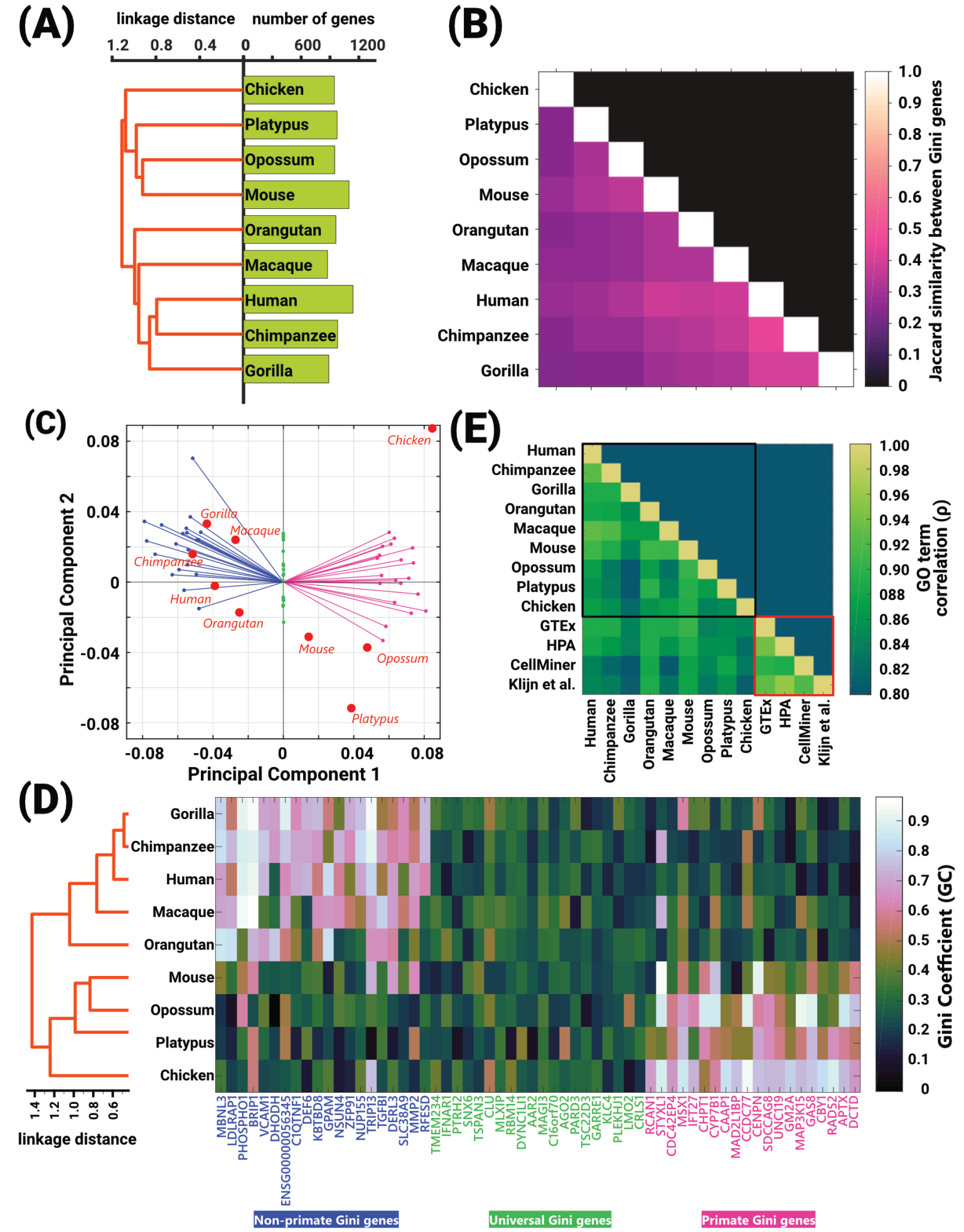
Figure 2. Gini coefficients are highly corelated for human datasets. (A) Gini coefficients of genes are highly correlated across human datasets, regardless of sample type. Datasets of sample types are more highly correlated than those of different sample types. (B) GO term coverage is highly correlated across human datasets.

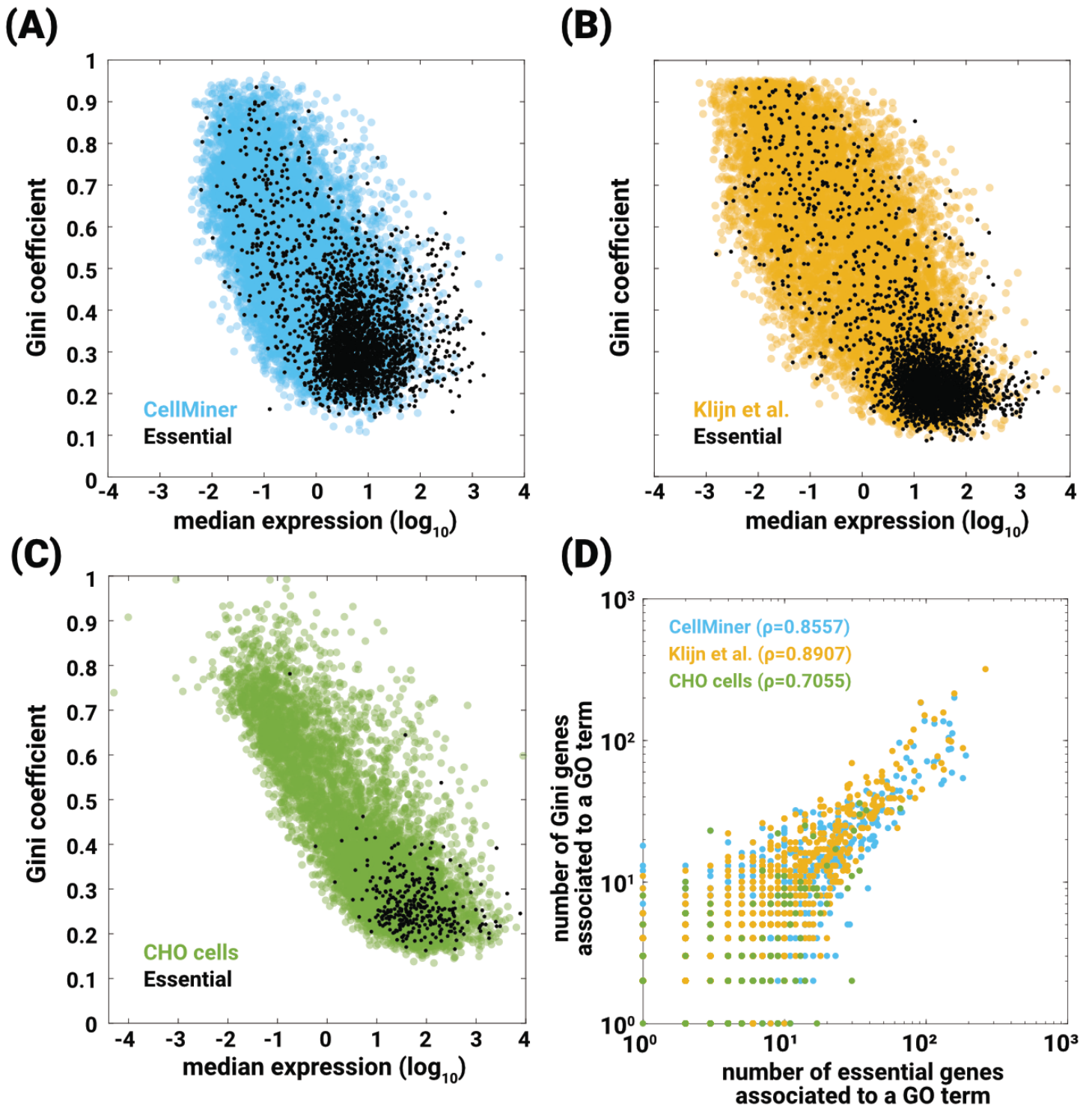
Figure 3. Gini coefficients accurately capture organism-specific differences. (A-B) Jaccard similarity between Gini genes identified using organism-specific transcriptomes capture cluster containing primates. The number of Gini genes with 1:1 orthologs in all organisms is shown using the bar plot on the right of the dendrogram. (C) Principal component 1 (PC1) also captures the cluster containing primates. Also shown are top 20 (pink), middle 20 (green), and bottom 20 (blue) 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). Top 20 Gini genes are specific to primates, middle 20 are universal Gini genes, and bottom 20 are specific to non-primates. (E) GO term coverage is highly correlated across different datasets, also shown are the GO term correlations with human datasets used in Figure 1.

Figure 4. Gini genes are essential. Gini coefficients of essential genes compared to the complete (A) CellMiner, (B) Klijn et al. cancer datasets, and (C) CHO datasets. 2800 genes essential in 20 cell lines were extracted from DepMap[40,41], and 338 CHO essential genes were extracted from Kai et al.[39] (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 fewer number of essential genes in CHO.


(A)
GTEx
HPA
CellMiner
Klijn et al.
Hamter tissues
CHO cells
0
20
40
60
80
100
Gini coefficient of GAPDH (in percentile)
(B)
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 (%)


(A)
HPA (H)
GTEx (G)
ρ = 0.8899
CellMiner (C)
GTEx (G)
ρ = 0.6725
Klijn et al. (K)
GTEx (G)
ρ = 0.7598
CellMiner (C)
HPA (H)
ρ = 0.6821
Klijn et al.
HPA (H)
ρ = 0.7628
Klijn et al. (K)
CellMiner (C)
ρ = 0.8663
(B)
number of genes from a dataset associated to a GO term
Number of Gini genes from GTEx associated to a GO term
Gini genes using:
HPA
CellMiner
Klijn et al.


(A)
linkage distance
number of genes
1.2
0.8
0.4
0
600
1200
Chicken
Platypus
Opossum
Mouse
Orangutan
Macaque
Human
Chimpanzee
Gorilla
(B)
Chicken
Platypus
Opossum
Mouse
Orangutan
Macaque
Human
Chimpanzee
Gorilla
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Jaccard similarity between Gini genes
(C)
0.08
0.04
0
-0.04
-0.08
Principal Component 2
Chicken
Gorilla
Macaque
Chimpanzee
Human
Orangutan
Mouse
Opossum
Platypus
-0.08
-0.04
0
0.04
0.08
Principal Component 1
(E)
Human
Chimpanzee
Gorilla
Orangutan
Macaque
Mouse
Opossum
Platypus
Chicken
GTEx
HPA
CellMiner
Klijn et al.
1.00
0.98
0.96
0.94
0.92
0.90
088
0.86
0.84
0.82
0.80
GO term correlation (ρ)
(D)
Gorilla
Chimpanzee
Human
Macaque
Orangutan
Mouse
Opossum
Platypus
Chicken
1.4
1.2
1.0
0.8
0.6
linkage distance
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
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Gini Coefficient (GC)


(A)
Gini coefficient
median expression (log10)
CellMiner
Essential
(B)
Gini coefficient
median expression (log10)
Klijn et al.
Essential
(C)
Gini coefficient
median expression (log10)
CHO cells
Essential
(D)
number of Gini genes associated to a GO term
number of essential genes associated to a GO term
CellMiner (ρ=0.8557)
Klijn et al. (ρ=0.8907)
CHO cells (ρ=0.7055)