Housekeeping genes, revisited at the single-cell level

Yingxin Lin[1], Shila Ghazanfar[1], Dario Strbenac[1], Andy Yi-Yang Wang[1,3], Ellis Patrick[1,4], Terence P Speed[5,6], Jean Yee Hwa Yang[1,2,*], Pengyi Yang[1,2,*]

[1] School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia

[2] Charles Perkins Centre, University of Sydney, Sydney, NSW 2006, Australia

[3] Sydney Medical School, University of Sydney, NSW 2006, Australia

[4] Westmead Institute for Medical Research, University of Sydney, Westmead, NSW 2145, Australia

[5] Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Australia

[6] Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia

* To whom correspondence should be addressed. Tel: +61 2-9351-3039; Email:

pengyi.yang@sydney.edu.au

Correspondence may also be addressed to Jean Yee Hwa Yang. Tel: +61 2-9351-3012; Email:

jean.yang@sydney.edu.au

Abstract

Housekeeping genes are critical for understanding the core transcriptome and instrumental in data normalisation given their stable expression in different tissues and cells. Previous studies defined housekeeping genes using bulk transcriptome data. With recent advances in single-cell RNA-sequencing (scRNA-seq), it is now possible to identify steadily expressed genes across individual cells. Here we introduce the concept of *housekeeping index* and a framework for assessing housekeeping genes at the single-cell level using high-resolution scRNA-seq data. We apply our approach on two scRNA-seq datasets from early mammalian development and evaluate derived housekeeping genes on ten additional scRNA-seq datasets from diverse cell/tissue types.

Keywords

Housekeeping genes; Single-cell RNA-sequencing; Mixture modelling; Housekeeping features; Housekeeping index

Background

Much of the phenomenal amount of phenotypic variation across cells of a given organism can be attributed to the complex array of transcribed genes, the transcriptome, stemming from a largely invariant genetic sequence. Despite these vast differences, a subset of genes traditionally referred to as housekeeping genes are shared across cell types and tissues [1,2]. The concept of housekeeping genes is often related to the gene set required to maintain basic cellular functions and therefore is crucial to the understanding of the core transcriptome that is required to sustain [3,4] or synthesize life [5,6]. Their distinctive genomic, structural, and evolutionary properties compared to tissue-specific genes make housekeeping genes a key to understanding various aspects of transcriptomes [7–10]. Besides their biological significance, having stable expression in different tissues and cell also allow housekeeping genes to be used for normalising and removing unwanted variation [11–14] from complex experiments.

A contemporary definition of housekeeping genes is a set of genes that are stably expressed in all cells of an organism, irrespective of tissue type, developmental stage, cell cycle state, or external signals [15]. Consistent with this definition, early studies such as those by Velculescu *et al.* [16], Warrington *et al.* [17], Hsiao *et al.* [1], and Eisenberg *et al.* [9] were conducted to define such set of genes using serial analysis

of gene expression (SAGE) or cDNA microarrays. These studies provided the initial approaches to identify housekeeping genes using large-scale expression data. With the advent of biotechnologies, follow-up studies using more comprehensive data sources such as those by De Jonge *et al.* [18] and Zhu *et al.* [19], and high-throughput RNA sequencing (RNA-seq) [15,20], have extensively revised these initial housekeeping lists.

Recent advances in high-throughput ultrafast sequencing at single-cell level (scRNA-seq) offers unprecedented resolution to profile transcriptomes across individual cells [21–23]. This technology has confirmed that cells exhibit a huge amount of variation in terms of their transcriptomes [24], and can facilitate a more precise characterisation of housekeeping genes at the single-cell level compared to those defined by traditional bulk profiling either with microarray or with RNA-sequencing technologies. Compared to bulk transcriptome data that requires aggregation of millions of cells to obtain a single gene expression measure, scRNA-seq data allows, for the first time, the expression dynamics of each gene within individual cells to be monitored, and therefore enables more accurate identification of genes that are truly expressed at a steady level in individual cells across tissues and developmental stages. In light of this, several unique aspects in scRNA-seq data must be considered for identifying housekeeping genes. First, scRNA-seq data typically contains a large proportion of zeros across many genes, partially as a result of the 'dropout' events from having limited starting material [25], or as a consequence of transcriptional bursting dynamics where genes are switched on and off in different cells [26] by regulatory elements such as enhancers [27]. Furthermore, a large number of genes from scRNA-seq data exhibit bimodality or multimodality of non-zero expression values [28–30], suggesting that many of these genes may be expressed at different levels in different cells.

To leverage the power of scRNA-seq in characterising housekeeping genes, in this study, we introduce the concept of housekeeping index (HK index) and propose an analytical framework to rank genes based on various characteristics extracted from scRNA-seq data which we term 'housekeeping features'. We applied the proposed approach on two large-scale high-resolution scRNA-seq datasets generated from early human and mouse development [31,32] to identify genes stably expressed across a wide range of cell types and developmental stages. The broad coverage of these two datasets, from as early as zygotes to mature blastocysts that represent distinctive tissue precursors including trophectoderm, primitive endoderm, and epiblast [33], provides a suitable starting point for deriving housekeeping genes during

human and mouse embryogenesis. We refer to the list of housekeeping genes identified from these two datasets as "h-scHK" and "m-scHK" genes for human and mouse respectively, and collectively as "scHK" genes. We subsequently evaluated these scHK genes on ten additional independent scRNA-seq datasets generated from diverse tissues types and sequencing protocols and compared them with those previously defined using bulk microarray [9] or RNA-seq datasets [15]. Our analyses shed light on the properties of housekeeping genes and offer a new way for assessing housekeeping genes given a suitable scRNA-seq dataset with improved precision than previous approaches.

Results

A novel analytical framework for deriving housekeeping index using four housekeeping features

While the concept of housekeeping genes is associated with the minimal collection of genes that are stably expressed in all cells and tissues, given the biological and technical limitations in current transcriptome profiling studies, it may be more sensible to characterise various aspects of each gene in terms of characteristics a housekeeping gene would possess. To this end, we propose an analytical framework (Figure 1A) for deriving a housekeeping index for each gene.

Gamma-Gaussian mixture model based housekeeping features. To characterise gene expression patterns from a scRNA-seq dataset, we utilised a Gamma-Gaussian mixture model [34] to fit gene expression values across individual cells. Specifically, non-zero expression values x_i (on $\log_2$FPKM scale) of gene i across cells is modelled by a mixture of distributions comprising of a Gamma component, corresponding to cells in which the gene is expressed at a low level, and a Gaussian component, corresponding to those in which the gene is expressed at a high level. The joint density function of the mixture model is defined as follows:

$$f(x_i, \alpha_i, \beta_i, \mu_i, {\sigma_i}^2, \lambda_i) = \lambda_i \frac{{\beta_i}^{\alpha_i}}{\Gamma(\alpha_i)} {x_i}^{\alpha_i - 1} e^{-\beta_i x_i} + (1 - \lambda_i) \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x_i - \mu_i)^2}{2{\sigma_i}^2}}$$

where α_i and β_i denote the shape and rate parameters of the Gamma component, while μ_i and ${\sigma_i}^2$ denote the mean and variance of the Gaussian component, and $0 \leq \lambda_i \leq 1$ is the mixing proportion indicating the proportion of cells in the Gamma component in the fitted model for the i^{th} gene. The mixture model parameters can be estimated using the Expectation-Maximisation (EM) algorithm. In our Gamma-Gaussian mixture model setting, genes with a low mixing proportion (λ) and a small variance (σ^2) with

respect to the Gaussian component suggest, respectively, a unimodal and an invariant expression pattern across the profiled single cells and therefore more likely to be housekeeping genes.

Let us denote the percentage of zeros for a given gene i across all cells as ω_i. There are a number of reasons why the measured expression level for a given gene and cell may be zero, including technical dropout due to failure to amplify the RNA from a small amount of starting material [25], stochastic expression patterns [26], and of course if no transcription is occurring for that gene. Thus, a desired characteristic of housekeeping genes is a relatively small ω value (i.e. low proportion of zeros) observed in scRNA-seq data, since we expect these genes to be stably expressed in all cells. One confounding factor is that lowly expressed housekeeping genes may have a higher proportion of zeros than highly expressed housekeeping genes simply due to technical dropout events as opposed to the underlying biology. An approach to account for this confounding factor is to take into consideration of average expression level μ in Gaussian component of each gene such as:

$$\omega^* = \sqrt{(1-\omega)\cdot\frac{\mu-\min(\mu)}{\max(\mu)-\min(\mu)}}$$

such that we anticipate more dropout events for housekeeping genes with low expression compared to highly expressed housekeeping genes. Three of our four housekeeping features are derived from the estimated mixture model, ideally genes with small λ, σ^2, ω^* correspond to housekeeping genes.

*F-statistic to select for equivalent expression across pre-defined experimental conditions (P)***.** We utilise the F-statistic as a housekeeping feature to select for genes in which we observe the same average gene expression across different pre-defined groups of experimental replicates, cell types, tissues, and individuals. Specifically, the F-statistic is commonly used in one-way analysis of variance testing, defined as

$$\mathrm{F-statistic} = \frac{(\sum_{k=1}^{P} n_k(\bar{x}_{k\cdot}-\bar{x}_{\cdot\cdot})^2)/(P-1)}{(\sum_{k=1}^{P}\sum_{l=1}^{n_k}(x_{kl}-\bar{x}_{k\cdot})^2)/(N-P)},$$

for N cells across P groups each with n_k cells, with dots denoting group means across the group index k and sample index l. The F-statistic measures departure from the ideal scenario of equal means across groups and we would thus expect to observe a small F-statistic associated with the experimental conditions for housekeeping genes. This observed F-statistic thus forms the fourth housekeeping feature, where we set the pre-defined class label as the associated experimental condition when available.

Taken together, genes with small λ, σ^2, ω^* and F-statistic are more likely to be housekeeping genes. We refer to these four quantities as housekeeping features. By combining these four housekeeping features, we defined for each gene a housekeeping index (HK index). Specifically, we first ranked genes in increasing order with respect to λ, σ^2, ω^* and F-statistics, respectively. Next, we rescaled the ranks of from each of the four housekeeping features to lie between 0 and 1, and defined the HK index for each gene as the average of its scaled rankings across all four housekeeping features. Thus, housekeeping genes can be selected by adjusting the HK index threshold and subsequently validated using a panel of evaluation matrices (Figure 1B). Importantly, genes can also be ranked in terms of their degree of evidence towards characteristics of housekeeping genes.

Characterising high resolution housekeeping genes at single-cell level in human and mouse

To demonstrate the proposed approach, we utilised two large-scale high-resolution scRNA-seq datasets to characterise housekeeping genes for human and mouse, respectively. Briefly, the two scRNA-seq datasets contain (i) transcriptome profiles of 1,529 individual cells derived from 88 human preimplantation embryos ranging from 3rd to 7th embryonic day [31] and (ii) transcriptome profiles of 269 individual cells derived from oocyte to blastocyst stages of mouse preimplantation development [32] (Table 1). The wide range of cell types and developmental stages captured by these two datasets provide a most suitable starting point for identifying genes stably expressed in different cell/tissue types in early human and mouse development.

Table 1. scRNA-seq datasets used for identifying scHK genes.

ID	Publication	Description	Organism	# cell	# class	Protocol
E-MTAB-3929	[31]	Early human development	Human	1529	5	SMART-Seq2
GSE45719	[32]	Early mouse development	Mouse	269	8	SMART-Seq2

We first looked at the proportion of zeros per gene across all profiled cells in the early human and mouse development scRNA-seq datasets respectively. We found that a large percentage of genes have more than 50% zero quantification across cells in both datasets (Figure 2A), suggesting most of the genes are transiently expressed during different developmental stages in both human and mouse. Nevertheless,

the mixing proportion of each gene, the variance and mean expression level from the Gaussian component from the mixture model, and the F-statistics calculated using pre-defined cell labels were different in human and mouse data (Figure 2B). This suggests there is a need to define housekeeping genes for human and mouse separately. By combining the scaled ranks of genes with respect to each housekeeping feature, the HK index distributions defined for human and mouse genes appeared to be highly comparable (Figure 2B, bottom right panel).

We derived a list of housekeeping genes for human and mouse respectively by computing the rank percentiles of HK index as well as the four housekeeping features. Genes with a HK index rank percentile above 80 as well as a reversed rank percentile above 60 for each of the four housekeeping features were included in the scHK gene list. Using this approach, we identified 1076 and 830 human scHK (h-scHK) and mouse scHK (m-scHK) genes respectively (Figure 2C). Compared to the human housekeeping genes defined previously with bulk microarray [9] (denoted as h-bHK microarray) and RNA-seq [15] (h-bHK RNA-seq), we found that our h-scHK genes have significantly smaller expression variances across individual cells (Figure 2C).

scHK genes are more robust compared to those defined by bulk transcriptome

We next investigated the reproducibility of the HK index by randomly sampling 80% of all cells and re-calculating the HK index for each sub-sample. We found the HK index to be highly reproducible in both the human and the mouse data (Figure 3A) with average Pearson correlation coefficients of 0.98 and 0.97, respectively. The HK indices also showed relatively high correlation between human and mouse (Figure 3B). Comparing human and mouse scHK genes defined in this study, there were 256 common genes, accounting for 24% of the h-scHK genes or 31% of the m-scHK genes (Figure 3C). Comparing with previously defined human housekeeping genes (Figure 3D), there were 97 common genes between our h-scHK list and those defined by microarray (9% and 18%), and 650 between our h-scHK list and those defined by bulk RNA-seq (60% and 17%). Together, these reflected a relatively low to moderate overlap amongst different housekeeping gene lists (Figure 3E).

To investigate the difference between our scHK gene list and that defined by bulk transcriptomes, we inspected a few individual genes that were defined as housekeeping genes using scRNA-seq data but not by bulk microarray or RNA-seq, and *vice versa*. We discovered that many ribosomal proteins (such as

RPL26 and *RPL36*) that were included in the scHK list but not in the bulk microarray or RNA-seq defined lists (Figure 3F) showed strong unimodal expression patterns across all cells. In contrast, genes such as *HINT1* (Histidine triad nucleotide-binding protein 1) and *AGPAT1* (1-Acylglycerol-3-Phosphate O-Acyltransferase), both of which have been reported to be differentially expressed in brain tissue [35] or malignant oesophageal tissues [36] compared to normal samples, were included in both microarray and RNA-seq defined housekeeping gene lists, but not in this study due to their bimodal expression patterns across individual cells.

Finally, we examined the expression patterns of *GAPDH* and *ACTB* (Figure 3G), genes which are commonly treated as canonical housekeeping genes for data normalisation, and observed clear bimodality in both the human and the mouse data. Consistent with previous studies [11,15,18,37], these data argue strongly against their usage as “housekeeping genes” for sample normalisation.

scHK genes exhibited stable expression across cells and developmental stages

We hypothesised that if the expression levels of the scHK genes are relatively stable, they should show relatively small expression differences across the different cell types from various biological systems. To test this in human and mouse developmental datasets, we utilised *k*-means clustering to partition cells into five and eight clusters respectively, using all genes (all expressed mRNA) or subsets of genes defined in each housekeeping gene list (i.e. h-scHK, m-scHK, h-bHK microarray and h-bHK RNA-seq) with the hypothesis that clusters arising from using housekeeping genes will exhibit lower concordance with pre-defined cell type- and tissue-specific labels (Figure 4A), thereby demonstrating consistent levels of expression across different cell and tissue types. Random subsets that contained the same number of genes as in scHK were included by sampling from either all genes or h-bHK RNA-seq list to account for the size of the gene-sets used in clustering (see Methods).

Indeed, we found that *k*-means clustering outputs using housekeeping genes derived from scRNA-seq data showed the lowest concordance to their pre-defined cell class labels (i.e. embryonic day of development or cell types) as quantified by the adjusted rand index (ARI) (Figure 4B) and the three other concordance metrics, namely Purity, Fowlkes-Mallows index (FM), and Jaccard index (Figure 4C). Together, these results demonstrate that scHK genes are stably expressed across cells and developmental stages in the two scRNA-seq datasets.

Stable expression of scHK genes generalises to ten independent scRNA-seq data

To test whether scHK genes derived from above two early mammalian development datasets are stably expressed in other cell and tissue types, we evaluated these scHK genes on ten additional datasets (Table 2) which are independent of the two scRNA-seq datasets used for identifying scHK genes. These additional datasets represent drastically different tissues and biological systems in both human and mouse, as well as different sequencing protocols and a wide range in the number of cells sequenced.

Table 2. scRNA-seq datasets used for evaluating scHK genes.

ID	Publication	Description	Organism	# cell	# class	Protocol
GSE94820	[52]	Peripheral blood mononuclear cells	Human	1140	5	SMART-Seq2
GSE75748	[53]	Pluripotent stem cells and endoderm progenitors	Human	1018	7	SMARTer
GSE72056	[54]	Multicellular metastatic melanoma	Human	4645	7	SMART-Seq2
GSE67835	[55]	Adult and fetal brain	Human	466	8	SMARTer
GSE60361	[45]	Cortex and hippocampus	Mouse	3005	7	SMARTer
GSE52583	[56]	Developmental lung epithelial cells	Mouse	198	4	SMARTer
E-MTAB-4079	[57]	Mesoderm diversification	Mouse	1205	4	SMART-Seq2
GSE84133	[58]	Pancreas inter- and intra-cells	Mouse	822	13	InDrop
GSE63472	[59]	Retinal tissue	Mouse	44808	39	Drop-Seq
10x Genomics	NA	Brain (https://support.10xgenomics.com)	Mouse	~1.3m	NA	Chromium

Similar to the above, we quantified the clustering concordance with respect to each of their pre-defined cell class labels using each of the four concordance metrics (ARI, Purity, FM, and Jaccard) (Table 3 and 4). We found that on average, clustering using scHK genes gave the lowest concordance to the pre-defined cell type- and tissue-specific class labels in all tested datasets compared to those defined using bulk microarray and RNA-seq datasets. Due to the low read coverage in the mouse retinal tissue (44808 cells) and brain (1.3 millions cells) datasets, as well as the lack of pre-defined class labels in the brain dataset, we assessed the percentage of zeros of scHK genes across all cells instead of clustering (Figure 4D). We found that m-scHK genes typically have low percentage of zeros across cells. These results suggest that stable expression of scHK genes generalise to various cell/tissue types and biological

systems and expression levels of scHK genes are generally more stable than those defined by bulk transcriptome data.

Table 3. Benchmark results on human scRNA-seq datasets.

	Peripheral blood mononuclear cells Villani et al. (2017)				**hPSCs and endoderm progenitors Chu et al. (2016)**			
	All genes	h-bHK microarray	h-bHK RNA-seq	h-scHK	All genes	h-bHK microarray	h-bHK RNA-seq	h-scHK
ARI	55±8	42±3	38±4	29±6	69±5	58±5	55±6	41±3
Purity	69±7	62±2	59±1	52±5	80±4	74±3	71±5	59±3
FM	67±5	56±1	52±3	45±4	75±4	66±4	63±5	51±2
Jaccard	49±6	39±1	35±2	29±4	60±5	48±4	46±6	34±2
	Multicellular metastatic melanoma Tirosh et al. (2016)				**Adult and fetal brain Darmanis et al. (2015)**			
	All genes	h-bHK microarray	h-bHK RNA-seq	h-scHK	All genes	h-bHK microarray	h-bHK RNA-seq	h-scHK
ARI	31±5	18±2	18±1	15±1	53±7	50±3	39±4	36±3
Purity	80±5	73±1	74±1	71±1	82±3	76±4	74±3	68±2
FM	51±3	39±2	40±1	37±1	62±6	59±2	50±3	47±3
Jaccard	32±2	22±2	24±1	21±1	44±6	41±2	33±3	30±3

All indices are within the range of [0, 1] and are multiplied by 100. The lowest results from each metric in each dataset are underlined.

Table 4. Benchmark results on mouse scRNA-seq datasets.

	Cortex and hippocampus Zeisel et al. (2015)				**Developmental lung epithelial cells Treutlein et al. (2014)**			
	All genes	h-bHK microarray	h-bHK RNA-seq	m-scHK	All genes	h-bHK microarray	h-bHK RNA-seq	m-scHK
ARI	45±8	36±5	31±3	27±2	61±6	55±4	48±2	45±0
Purity	72±3	66±1	63±1	58±1	83±4	80±2	76±1	74±0
FM	55±6	49±4	44±3	41±2	72±4	68±3	62±2	60±0
Jaccard	38±6	32±4	28±2	25±1	56±5	51±3	45±2	43±0
	Mesoderm diversification Scialdone et al. (2016)				**Pancreas inter- and intra-cells Baron et al. (2016)**			
	All genes	h-bHK microarray	h-bHK RNA-seq	m-scHK	All genes	h-bHK microarray	h-bHK RNA-seq	m-scHK
ARI	54±2	43±8	49±3	32±6	37±4	22±3	23±3	18±2
Purity	66±1	62±6	65±1	59±6	89±3	78±3	76±2	72±1
FM	68±1	63±8	67±1	58±7	52±4	38±3	39±3	34±2
Jaccard	52±1	46±7	50±1	40±8	30±3	20±3	21±3	17±2

All indices are within the range of [0, 1] and are multiplied by 100. The lowest results from each metric in each dataset are underlined.

Housekeeping index correlates as expected with sequence characteristics

We further characterised our housekeeping genes by correlating the HK index and each housekeeping feature with various gene structural and conservation features extracted from various data sources. We found that the HK index correlated positively with the number of exons in a gene, gene expression, and gene conservation, and negatively with GC-content in the gene body in both human and mouse (Figure 5A). These results could also be observed by comparing the h-scHK and m-scHK genes with all genes expressed in early human and mouse datasets respectively (Figure 5B). Consistent with previous studies, we found scHK genes are more evolutionarily conserved [38] with higher phyloP scores. scHK genes also possess more exons, in agreement with previous finding [7], despite mouse genes on average having fewer exons than human genes. Both human and mouse scHK genes appeared to have a slightly lower GC-content but, similar to previously reported, the relation was relatively weak [39] (Figure 5B). Together, housekeeping genes defined in this study showed similar gene characteristics to those observed in the previous studies but the resolution of scRNA-seq data and the consistency across two mammalian species strengthened and further validated these observations.

Interactive web resource

We implemented an interactive web resource using the Shiny R application that make our approach for refining and identifying housekeeping genes universally accessible. The web resource (freely available from http://shiny.maths.usyd.edu.au/scHK) provides all key housekeeping features that were used for deriving human or mouse scHK genes and allows users to adjust the stringency of these features to tailor scHK gene list dynamically. Gene features extracted above were also incorporated for human and mouse respectively to assist interpretation of housekeeping genes. In addition, our web resource allows users to provide their own gene list for comparison in terms of gene features as well as enrichment (i.e. over-representation) with respect to the scHK gene lists using Fisher's exact test.

Discussion

Since the emergence of large-scale transcriptomic profiling around the turn of this century, the search for housekeeping genes has been a centrally important quest in modern biology. While numerous studies have categorised housekeeping genes in different organisms with varying degrees of success, the overall

concordance remains relatively low. This perhaps is due to the extreme plasticity and heterogeneity of transcriptomes in different biological systems [36,40] as well as limitations of the experimental techniques and the computational methods used for analysing transcriptome profiling data to identify housekeeping genes. By employing the latest advances in scRNA-seq, we have revisited the search for housekeeping genes. We introduced a framework in which various "housekeeping features" were defined based on the expression characteristics of each gene in scRNA-seq data. This has allowed us to generate a ranking system for the genes based on a housekeeping index derived from the housekeeping features. Using this framework, we derived a list of housekeeping genes for human and mouse, respectively, based on two comprehensive scRNA-seq datasets that cover a wide range of cell types and developmental stages in early human and mouse development.

Compared to the previously identified housekeeping gene lists in human arising from bulk microarray or RNA-seq data, we found relatively low to moderate overlap with those defined using scRNA-seq data, highlighting the distinctive opportunity of using scRNA-seq data to assess housekeeping genes. A closer inspection of a few genes (e.g. *HINT1* and *AGPAT1*) that were described as housekeeping genes in previous bulk transcriptome data revealed a clear bimodality in their expression patterns, suggesting their altered expression levels in different cells and/or states. Two commonly used data normalisation genes *GAPDH* and *ACTB* were also found to be expressed at different levels across individual cells in both human and mouse. Together, these results demonstrate the unique advantages of identifying housekeeping genes via transcriptomics analytics of single cells. Comparison of human and mouse housekeeping indexes showed high correlation. In agreement with this, housekeeping genes are evolutionarily more conserved according to phyloP scores. These together suggest the expression properties of housekeeping genes are preserved across different species.

Current efforts are under way to comprehensively characterise the transcriptome of every human cell (https://www.humancellatlas.org/), which will provide an unprecedented resolution to a large array of cells in human. Information from such resources in conjunction with our present framework for identifying housekeeping genes will provide an even more precise housekeeping gene assessment that will enrich subsequent avenues of research including biological characterisation of such genes and use of these genes for technical normalisation and standardisation. As the dynamic nature of the transcriptome is

uncovered with more resolved snapshots, we are in a unique position to interrogate housekeeping genes, identifying and characterising such genes.

A current technological limitation is the 'test to destruction' of single cells, i.e. a single cell's entire transcriptome cannot be monitored over time to assess the dynamic nature of transcription. However, new technologies such as sequential fluorescence *in situ* hybridization (seqFISH) [41] and multiplexed error-robust fluorescence *in situ* hybridization (MERFISH) [42] allow the monitoring of the transcription process for up to hundreds of genes at once over time for many cells. The current limitation is the low-throughput of this technology compared to scRNA-Seq, but the scHK genes identified in the current work may lead to prioritisation of proposed housekeeping genes for further characterisation and interrogation.

Conclusions

We introduce a novel concept of housekeeping index and present an analytical framework for deriving such index from scRNA-seq data. While our newly complied housekeeping gene lists from two seminal scRNA-seq datasets have immediate utility both for understanding housekeeping gene biology and for practical applications such as data normalisation, the concept of housekeeping index and the computational framework described in this study relax the stiff definition of housekeeping genes and allow the "amount of evidence of housekeeping" to be measured for each gene. Indeed, the interactive web resource enables a more or less stringent list of genes to be selected based on their scRNA-seq expression characteristics according to different applications and purposes. Furthermore, the proposed framework can be applied in a data dependent manner to identify stably expressed genes from any given scRNA-seq dataset. This may be useful for scRNA-seq studies as defining positive control genes is often a key step in analysing such data [43,44]. Taken together, our study marks a shift in paradigm in identifying housekeeping genes at the single-cell level and extends the concept for selecting genes that are stably expressed for practical applications.

Methods

scRNA-seq data processing

Public scRNA-seq data (Table 1 and 2) were downloaded from either NCBI GEO repository or the EBML-EBI ArrayExpress repository (except the 'brain' dataset which was downloaded from

https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons). Fragments per kilobase of transcript per million (FPKM) values or counts per million (CPM) from their respective original publications were used to quantify full length gene expression for datasets generated by SMARTer or SMAART-Seq2 protocols. For the 'retinal tissue' and 'brain' datasets, log2 counts from 3'-end counting were used given their extremely low per cell read coverage. Except the 'brain' dataset, all other datasets have undergone cell-type identification using biological knowledge assisted by various clustering algorithms from their respective original publications which we retain for evaluation purposes. For each dataset, genes with more than 80% missing values (zeros) were removed and those that passed the filtering were considered as expressed in that dataset. These filtered datasets were used for all subsequent analyses.

Benchmarking of housekeeping genes

To assess the quality of the proposed housekeeping genes in this work as well as for previous studies, the *k*-means algorithm was utilised to cluster each scRNA-seq data to its pre-defined number of clusters and an array of evaluation metrics were applied to compute the concordance with respect to the pre-defined ("true") class labels. Evaluation metrics include the adjusted rand index (ARI), Purity, the Fowlkes-Mallows index (FM) and the Jaccard index.

Let $U = \{u_1, u_2, \ldots, u_P\}$ denote the true partition across P classes and $V = \{v_1, v_2, \ldots, v_K\}$ denote the partition produced from *k*-means clustering ($K = P$). Let a be the number of pairs of cells correctly partitioned into the same class by the clustering method; b be the number of pairs of cells partitioned into the same cluster but in fact belong to different classes; c be the number of pairs of cells partitioned into different clusters but belong to the same class; and d be the number of pairs of cells correctly partitioned into different clusters (Table 5).

Table 5. Confusion matrix for measuring cluster concordance with pre-defined cell class labels.

		k-means clustering output	
		# pairs in the same class	# pairs in different classes
Pre-defined class labels	# pairs in the same class	a	c
	# pairs in different classes	b	d

Then the Adjusted Rand Index [48] can be calculated as

$$\mathrm{ARI} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)};$$

the Jaccard index [49] can be calculated as

$$\mathrm{Jaccard} = \frac{a}{a+b+c};$$

the Fowlkes-Mallows index [50] can be calculated as

$$\mathrm{FM} = \sqrt{(\frac{a}{a+b})(\frac{a}{a+c})};$$

and the Purity [51] can be calculated as

$$\mathrm{Purity} = \frac{1}{N}\sum_{i} \max_{j} |u_i \cap v_j|,$$

where N is the total number of cells.

For each dataset, we calculated and compared the above four metrics using all expressed genes, housekeeping genes defined using microarray data [9], housekeeping gene defined using bulk RNA-seq data [15], and scHK genes defined in this study. In order to account for potential effects of gene list length, we also generated random subsets with the same number of genes in our scHK lists first by randomly sampling from all expressed genes in the dataset, and second by randomly sampling from the housekeeping gene list defined by bulk RNA-seq. Since the *k*-means clustering algorithm is not deterministic and the random sampling process introduces variability, the above procedure was repeated 10 times.

Housekeeping gene properties

To characterise human and mouse scHK genes more fully, we extracted gene structural features including the number of exons and percentage GC content in the gene body for human and mouse respectively, using the biomaRt [46] R package. Additionally, to characterise evolutionary conservation, phyloP scores were downloaded from the UCSC Genome Browser for mm10 and hg38 genomes. Exonic bases of each gene were determined based on GENCODE Genes release 26 for human and release 14 for mouse. The set of conservation scores for each gene was averaged to calculate a single score per gene. We assessed the concordance of housekeeping features with structural features, conservation scores, and their expression across all genes for human and mouse using Pearson correlation coefficients. We then

compared these features for genes deemed to be housekeeping genes defined in this and previous studies against all expressed genes in human and mouse, respectively.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and/or analysed during the current study are available in either the NCBI GEO repository or the EBML-EBI ArrayExpress repository (except the 'brain' dataset which was downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons) (Table 1 and 2). An interactive web resource for scHK genes is available at http://shiny.maths.usyd.edu.au/scHK.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by Australian Research Council (ARC)/Discovery Early Career Researcher Award (DE170100759) to Pengyi Yang, National Health and Medical Research Council (NHMRC)/Career Development Fellowship (1105271) to Jean Yee Hwa Yang, ARC/Discovery Project (DP170100654) grant to Pengyi Yang and Jean Yee Hwa Yang, and NHMRC/Program Grant (1054618) to Terence P Speed.

Authors' contributions

PY conceived the study with input from JYHY. All authors contributed to the design, analytics, interpretation and the direction of the study. YL and PY lead the analytics and AYW lead the curation of the datasets. All authors wrote, reviewed, edited, and approved the final version of the manuscript.

Acknowledgements

The authors thank all their colleagues, particularly at The University of Sydney, School of Mathematics and Statistics, for support and intellectual engagement. We would also like to thank Prof. Ze-Guang Han and Dr. Xianbin Su from Shanghai Jiao Tong University for informative discussion and valuable feedback.

References

1. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen R V, Misra J, et al. A compendium of gene expression in normal human tissues. Physiol. Genomics [Internet]. 2001;7:97–104. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11773596

2. Butte, Atul J., Victor J. Dzau and SBG. Further defining housekeeping, or “maintenance,” genes Focus on “A compendium of gene expression in normal human tissues.” J. Biol. Chem. [Internet]. 2002;5:2002–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21136217

3. Koonin E V. How Many Genes Can Make a Cell□: The Minimal-Gene-Set Concept. Annu. Rev. Genomics Hum. Genet. [Internet]. 2000 [cited 2017 Jun 26];1:99–116. Available from: http://www.annualreviews.org/doi/10.1146/annurev.genom.1.1.99

4. Gil R, Silva FJ, Peretó J, Moya A. Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. [Internet]. American Society for Microbiology; 2004 [cited 2017 Jun 26];68:518–37, table of contents. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15353568

5. Forster AC, Church GM. Towards synthesis of a minimal cell. Mol. Syst. Biol. [Internet]. 2006 [cited 2017 Jun 26];2:45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16924266

6. Esvelt KM, Wang HH. Genome-scale engineering for systems and synthetic biology. Mol. Syst. Biol. [Internet]. European Molecular Biology Organization; 2013 [cited 2017 Jun 26];9:641. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23340847

7. Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes. Trends Genet. 2008;24:481–4.

8. She X, Rohl CA, Castle JC, Kulkarni A V, Johnson JM, Chen R. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genomics [Internet]. BioMed Central; 2009 [cited 2017 Jun 27];10:269. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19534766

9. Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes are compact. Trends Genet. 2003;19:356–62.

10. Vinogradov AE. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. [Internet]. 2004 [cited 2017 Jun 27];20:248–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15109779

11. Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, et al. Housekeeping genes as internal standards: use and limits. J. Biotechnol. [Internet]. 1999 [cited 2017 Jun 26];75:291–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10617337

12. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. [Internet]. 2010 [cited 2017 Jun 26];11:R25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20196867

13. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. [Internet]. 2014;32:896–902. Available from: http://dx.doi.org/10.1038/nbt.2931

14. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13:539–52.

15. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. [Internet]. Elsevier Ltd; 2013;29:569–74. Available from: http://dx.doi.org/10.1016/j.tig.2013.05.010

16. Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C, et al. Analysis of human transcriptomes. Nat. Genet. [Internet]. Nature Publishing Group; 1999 [cited 2017 Jun 24];23:387–8. Available from: http://www.nature.com/doifinder/10.1038/70487

17. Warrington J a, Nair A, Mahadevappa M, Tsyganskaya M, Zhang F, Broughton RE, et al. Comparison of human adult and fetal expression and identification of 535 housekeeping / maintenance genes Comparison of human adult and fetal expression and identification of 535 housekeeping / maintenance genes. Genomics, Physiol. 2000;2:143–7.

18. de Jonge HJM, Fehrmann RSN, de Bont ESJM, Hofstra RMW, Gerbens F, Kamps WA, et al. Evidence Based Selection of Housekeeping Genes. Lichten M, editor. PLoS One [Internet]. Public Library of Science; 2007 [cited 2017 Jun 24];2:e898. Available from: http://dx.plos.org/10.1371/journal.pone.0000898

19. Zhu J, He F, Song S, Wang J, Yu J. How many human genes can be defined as housekeeping with current expression data? BMC Genomics [Internet]. BioMed Central; 2008 [cited 2017 Jun 24];9:172.

Available from: http://www.ncbi.nlm.nih.gov/pubmed/18416810

20. Ramsköld D, Wang ET, Burge CB, Sandberg R, Gao W. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. Jensen LJ, editor. PLoS Comput. Biol. [Internet]. Public Library of Science; 2009 [cited 2017 Jun 24];5:e1000598. Available from: http://dx.plos.org/10.1371/journal.pcbi.1000598

21. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods [Internet]. Nature Publishing Group; 2009 [cited 2017 Jun 24];6:377–82. Available from: http://www.nature.com/doifinder/10.1038/nmeth.1315

22. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science (80-.). [Internet]. 2014 [cited 2017 Jun 24];343. Available from: http://science.sciencemag.org/content/343/6172/776

23. Kolodziejczyk A, Kim JK, Svensson V, Marioni J, Teichmann S. The Technology and Biology of Single-Cell RNA Sequencing. Mol. Cell [Internet]. 2015 [cited 2017 Jun 24];58:610–20. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1097276515002610

24. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res. 2014;24:496–510.

25. Kharchenko P V, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat. Methods [Internet]. Nature Research; 2014 [cited 2017 Jun 24];11:740–2. Available from: http://www.nature.com/doifinder/10.1038/nmeth.2967

26. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. Science (80-.). [Internet]. 2011 [cited 2017 Jun 24];332:472–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21415320

27. Fukaya T, Lim B, Levine M. Enhancer Control of Transcriptional Bursting. Cell [Internet]. 2016 [cited 2017 Jul 20];166:358–68. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867416305736

28. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature [Internet]. 2013 [cited 2017 Jun 24];498:236–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23685454

29. McDavid A, Dennis L, Danaher P, Finak G, Krouse M, Wang A, et al. Modeling Bi-modality Improves Characterization of Cell Cycle on Gene Expression in Single Cells. Zhong S, editor. PLoS Comput. Biol.

[Internet]. 2014 [cited 2017 Jun 24];10:e1003696. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25032992

30. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol. [Internet]. BioMed Central; 2013 [cited 2017 Jun 24];14:R7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23360624

31. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell [Internet]. Elsevier; 2016 [cited 2017 Jun 24];165:1012–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27062923

32. Deng Q, Ramskold D, Reinius B, Sandberg R. Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. Science (80-.). [Internet]. 2014 [cited 2017 Jun 24];343:193–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24408435

33. Cockburn K, Rossant J. Making the blastocyst: lessons from the mouse. J. Clin. Invest. [Internet]. 2010 [cited 2017 Jun 24];120:995–1003. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20364097

34. Ghazanfar S, Bisogni AJ, Ormerod JT, Lin DM, Yang JYH. Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. BMC Syst. Biol. [Internet]. BioMed Central; 2016 [cited 2017 Jun 27];10:127. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28105940

35. Varadarajulu J, Schmitt A, Falkai P, Alsaif M, Turck CW, Martins-de-Souza D. Differential expression of HINT1 in schizophrenia brain tissue. Eur. Arch. Psychiatry Clin. Neurosci. [Internet]. 2012 [cited 2017 Jul 20];262:167–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21553311

36. Rubie C, Kempf K, Hans J, Su T, Tilton B, Georg T, et al. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. Mol. Cell. Probes [Internet]. 2005 [cited 2017 Jun 27];19:101–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15680211

37. Suzuki T, Higgins PJ, Crawford DR. Control selection for RNA quantitation. Biotechniques [Internet]. 2000 [cited 2017 Jul 21];29:332–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10948434

38. Zhang L, Li W-H. Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. Mol. Biol. Evol. [Internet]. 2004 [cited 2017 Jul 6];21:236–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14595094

39. Semon M, Mouchiroud D, Duret L. Relationship between gene expression and GC-content in

mammals: statistical significance and biological relevance. Hum. Mol. Genet. [Internet]. 2004 [cited 2017 Jul 4];14:421–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15590696

40. Arukwe A. Toxicological Housekeeping Genes: Do They Really Keep the House? American Chemical Society; 2006 [cited 2017 Jun 27]; Available from: http://pubs.acs.org/doi/abs/10.1021/es0615223

41. Shah S, Lubeck E, Zhou W, Cai L. seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus. Neuron. 2017;94:752–758.e1.

42. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science [Internet]. 2015;348:aaa6090. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25858977%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4662681

43. Bacher R, Chu L-F, Leng N, Gasch AP, Thomson JA, Stewart RM, et al. SCnorm: robust normalization of single-cell RNA-seq data. Nat. Methods [Internet]. 2017 [cited 2017 Jun 27];14:584–6. Available from: http://www.nature.com/doifinder/10.1038/nmeth.4263

44. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. [Internet]. 2016 [cited 2017 Jun 27];17:75. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0947-7

45. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-.). [Internet]. 2015 [cited 2017 Jul 6];347:1138–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25700174

46. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics [Internet]. 2005 [cited 2016 Nov 21];21:3439–40. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti525

47. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics [Internet]. 2010 [cited 2017 May 31];26:139–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19910308

48. Rand WM. Objective Criteria for the Evaluation of Clustering Methods. J. Am. Stat. Assoc. 1971;66:846–50.

49. Milligan GW, Cooper MC. A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. Multivariate Behav. Res. [Internet]. 1986;21:441–58. Available from: http://eric.ed.gov/?id=EJ344680

50. Fowlkes EB, Mallows CL. A Method for Comparing Two Hierarchical Clusterings. J. Am. Stat. Assoc. [Internet]. 1983;78:553. Available from: http://www.jstor.org/stable/2288117%5Cnhttp://www.jstor.org/stable/2288117?origin=crossref

51. Amigó E, Gonzalo J, Artiles J, Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retr. Boston. 2009;12:461–86.

52. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science (80-.). [Internet]. 2017 [cited 2017 Jul 24];356:eaah4573. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28428369

53. Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. [Internet]. 2016 [cited 2017 Jul 24];17:173. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27534536

54. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (80-.). [Internet]. 2016 [cited 2017 Jul 7];352:189–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27124452

55. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. [Internet]. 2015;112:201507125. Available from: http://www.pnas.org/content/112/23/7285.abstract

56. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature [Internet]. 2014;509:371–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4145853&tool=pmcentrez&rendertype=abstract

57. Scialdone A, Tanaka Y, Jawaid W, Moignard V, Wilson NK, Macaulay IC, et al. Resolving early mesoderm diversification through single-cell expression profiling. Nature [Internet]. 2016;535:4–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27383781%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4947525

58. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. 2016;3:346–60.

59. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.

Figure legends

Figure 1. Schematic illustration of the proposed analytic framework for deriving housekeeping index.

A. Housekeeping features extracted directly from the mixture model are coloured in blue. Those extracted from additional scRNA-seq data characteristics are in red. The overall housekeeping index are derived from the combination of all housekeeping features.
B. Housekeeping genes identified using this framework are evaluated using different metrics (see Methods for details on evaluation).

Figure 2. Characterising single cell housekeeping features.

A. Percentage of zeros per gene across individual cells profiled from scRNA-seq datasets that comprise zygotes development to tissue precursors in human and mouse respectively.
B. Fitted values of mixing proportion (λ), and variance (σ^2) and (μ) in the Gaussian component (top panels) in the mixture model for each gene in human and mouse scRNA-seq datasets. Regularised percentage of zeros, F-statistics computed from pre-defined cell class and developmental stages (bottom left panel) and HK index derived for each gene for human and mouse (bottom right panel), respectively.
C. Scatter plot showing mean expression (x-axis) and variance (y-axis) on log scale of each gene (grey circles) across profiled single cells. Open red and green circles represent housekeeping genes derived for human (h-scHK; left panel) and mouse (m-scHK; right panel) in this study whereas dark and light blue solid circles represent housekeeping genes defined previously using bulk microarray[9] and RNA-seq data[15].

Figure 3. Reproducibility and comparison of housekeeping genes defined using scRNA-seq with bulk microarray and RNA-seq.

A. Scatter plot of HK index calculated from two random sub-sampling of cells in human and mouse datasets. Mean Pearson's correlation coefficient ($\bar{r}$) were calculated from pairwise comparison of 10 repeated random sub-sampling on each dataset.

B. Scatter plot of HK index calculated from using the full set of homologous human and mouse genes.

C-E. Venn diagrams showing overlaps of housekeeping genes defined using scRNA-seq for human and mouse (C), those defined using bulk microarray and RNA-seq (D), and the overlap of all lists (E).

F. Histograms of expression patterns of example genes that are defined as h-scHK genes using scRNA-seq data but not bulk microarray or RNA-seq data (*RPL26* and *RPL36*) and vice versa (*HINT* and *AGPAT1*) across individual cells.

G. Histograms of expression patterns for *GAPDH* and *ACTB* in human and mouse (*Gapdh* and *Actb*) across individual cells.

Figure 4. Evaluation of housekeeping genes using various concordance metrics.

A. Schematic illustrating concordance of *k*-means clustering with pre-defined cell classes using a panel of four metrics including adjusted rand index (ARI), Purity, Fowlkes-Mallows index (FM), and Jaccard index.

B. Violin plot of concordance (adjusted rand index) between *k*-means clustering and pre-defined cell class labels, using all expressed genes, genes included in each housekeeping gene list, and random subsets of genes sampled from all expressed genes (random subset) or those from h-bHK RNA-seq (h-bHK RNA-seq subset) that match the size of h-scHK and m-scHK, respectively.

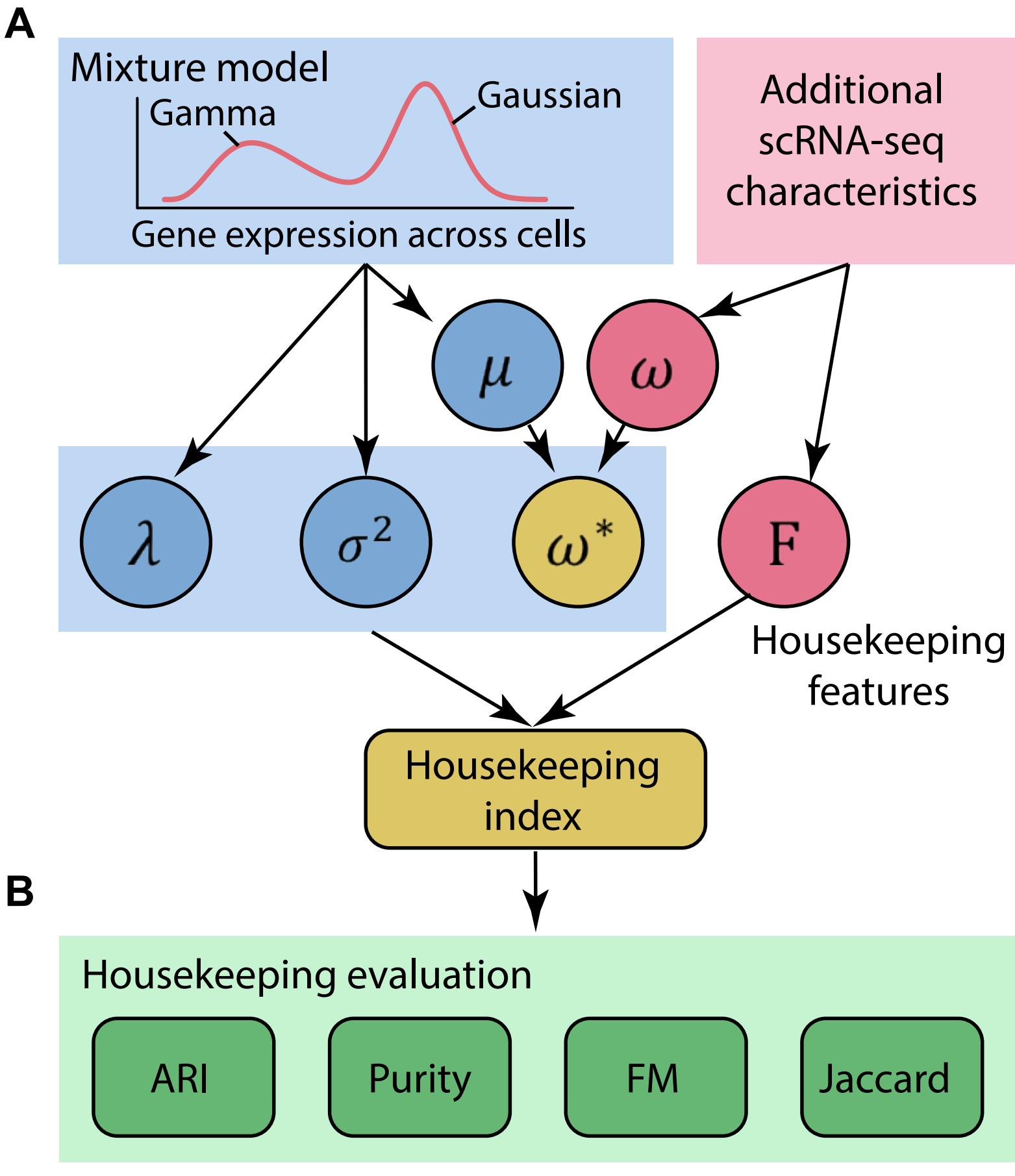
C. Barplots of concordance between *k*-means clustering and pre-defined cell class labels, using all expressed genes, genes included in each housekeeping gene list, and random subsets as in (B) for human and mouse data, respectively. Concordance is evaluated in terms of all four metrics.

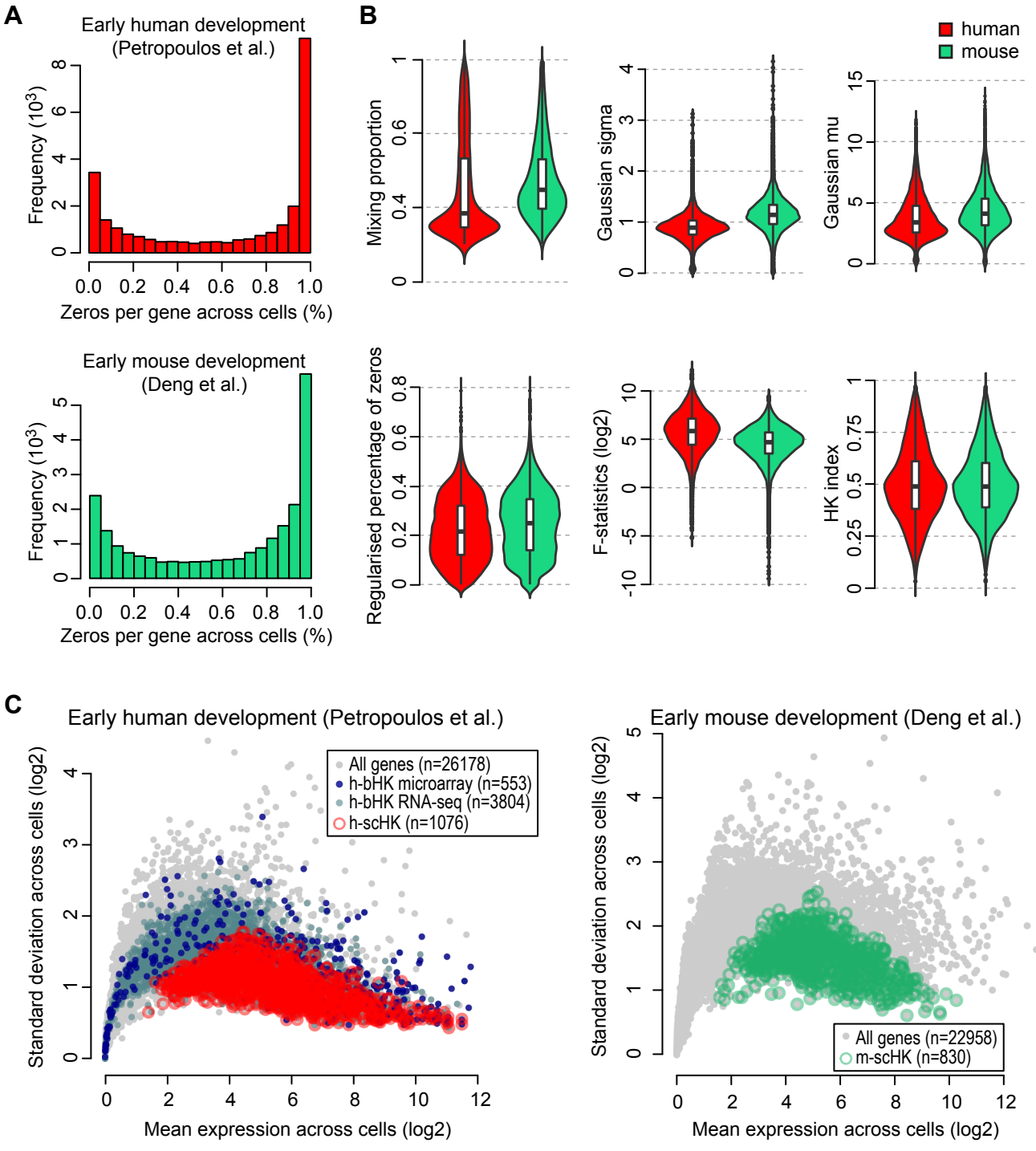
D. Percentage of zeros across cells for all expressed genes and housekeeping genes defined from h-bHK microarray, h-bHK RNA-seq, m-scHK, top-300 m-scHK, and top-100 m-scHK in mouse retinal tissue and brain datasets.

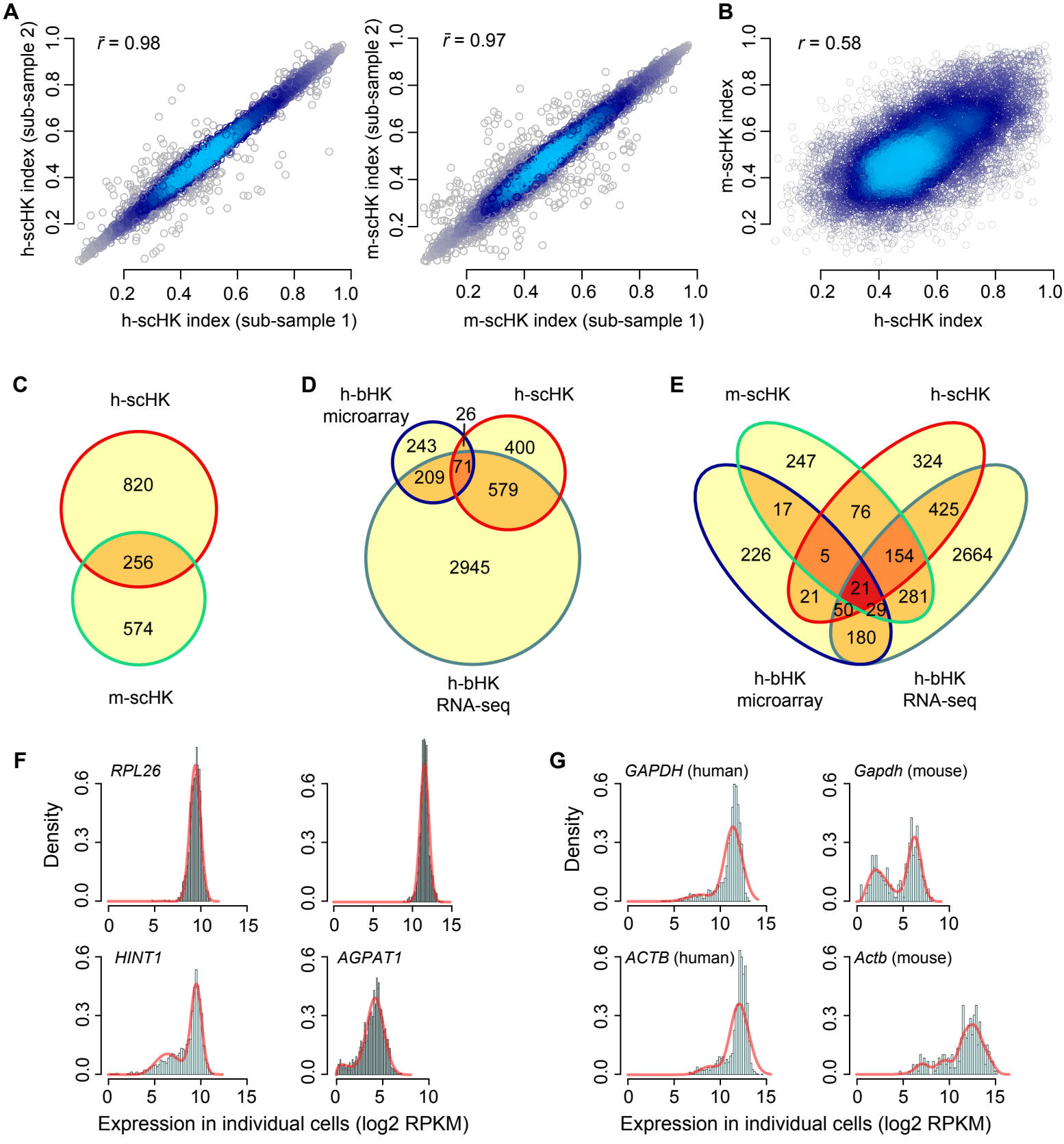
Figure 5. Characterisation of housekeeping gene properties.

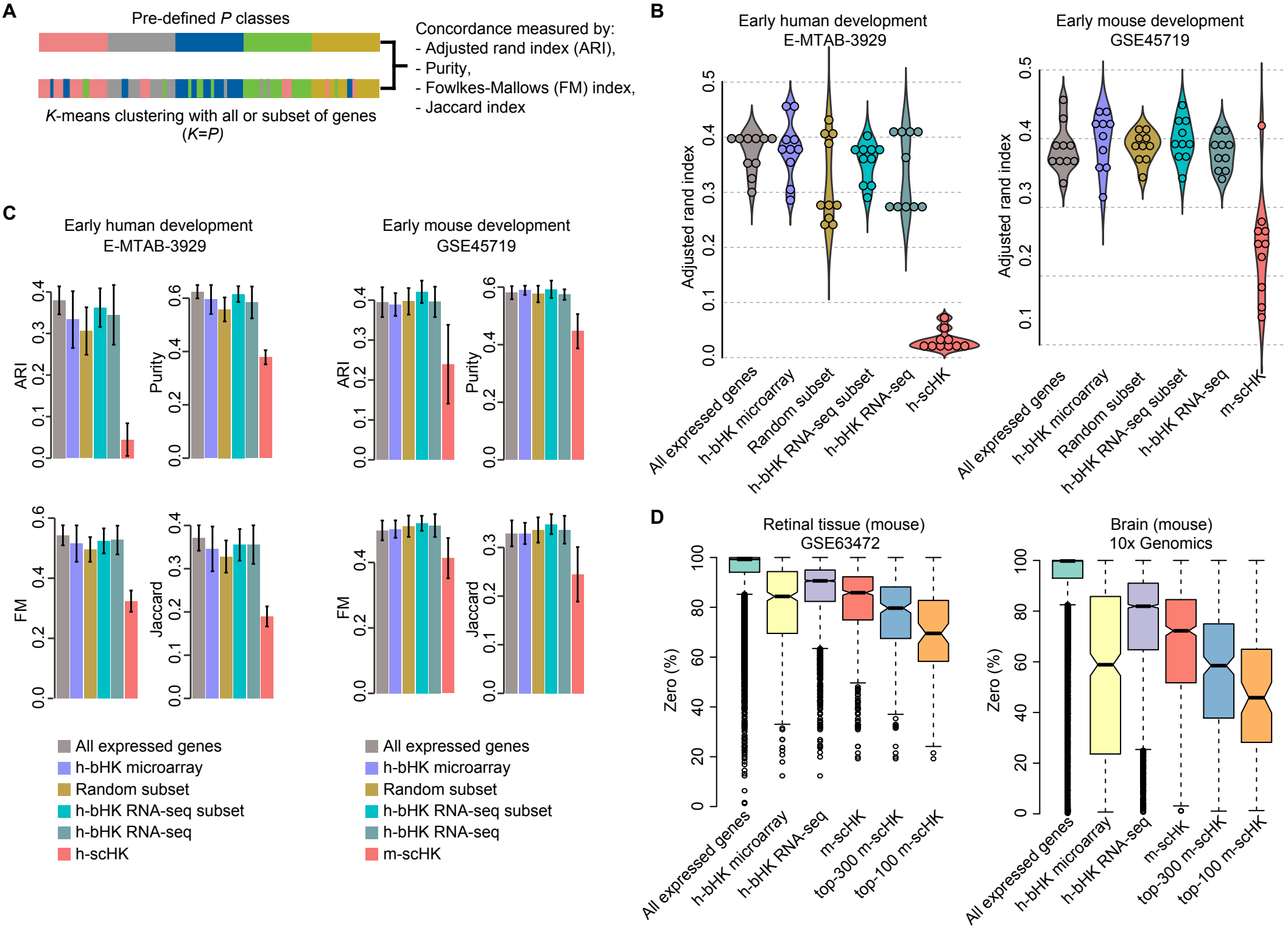
A. Pearson correlation analyses of human and mouse housekeeping features with respect to genomic structural and evolutional gene features.

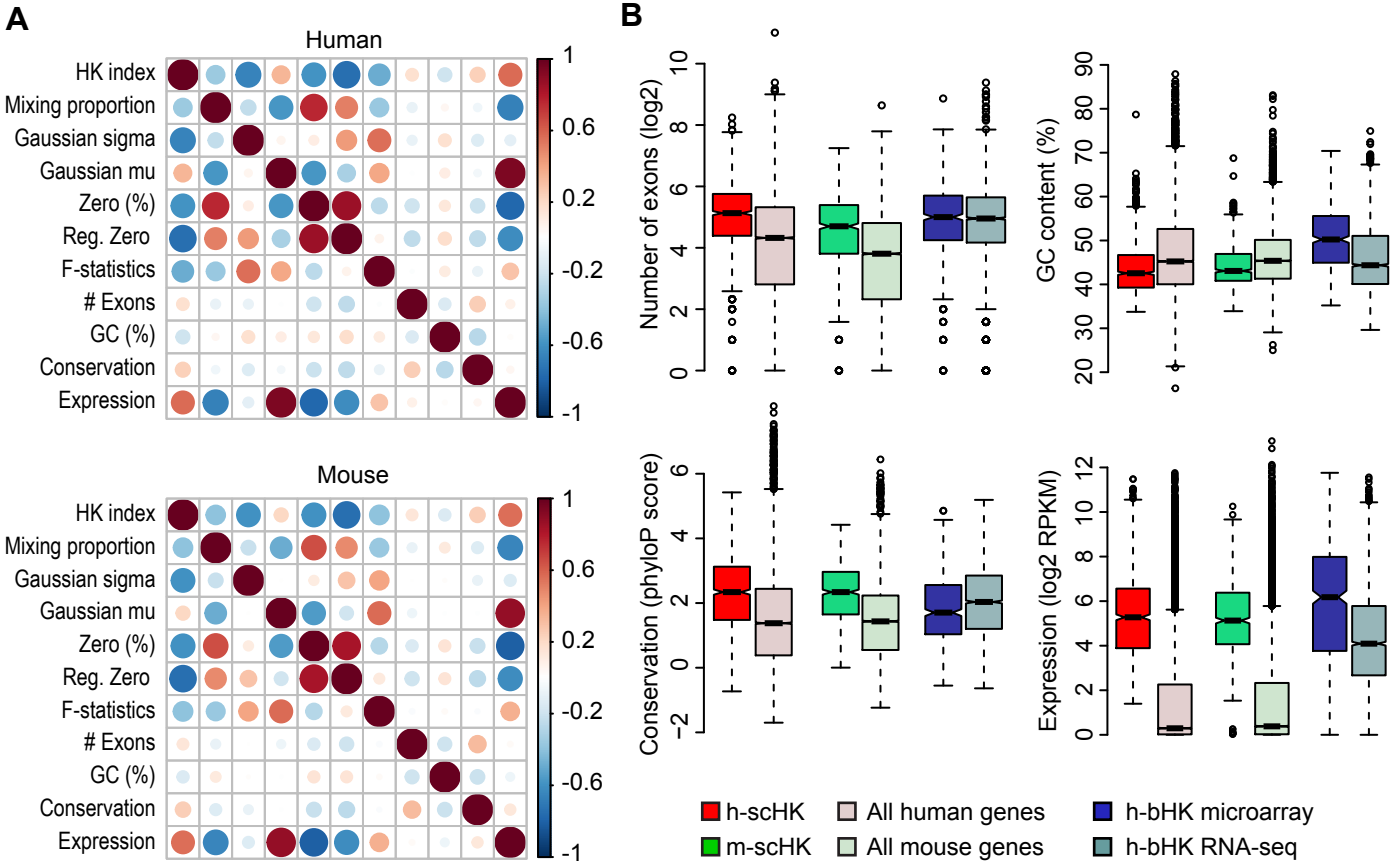
B. Boxplots of individual gene features (number of exons, GC content, conservation score, and expression level) for human and mouse scHK genes, housekeeping genes defined previously using bulk microarray and RNA-seq, and all genes (i.e. all expressed mRNA) in early human and mouse development data.


A
Mixture model
Gamma
Gaussian
Gene expression across cells
Additional scRNA-seq characteristics
μ
ω
λ
σ^2
ω^*
F
Housekeeping features
Housekeeping index
B
Housekeeping evaluation
ARI
Purity
FM
Jaccard


A
Early human development
(Petropoulos et al.)
Frequency (10^3)
Zeros per gene across cells (%)
Early mouse development
(Deng et al.)
B
human
mouse
Mixing proportion
Gaussian sigma
Gaussian mu
Regularised percentage of zeros
F-statistics (log2)
HK index
C
Early human development (Petropoulos et al.)
All genes (n=26178)
h-bHK microarray (n=553)
h-bHK RNA-seq (n=3804)
h-scHK (n=1076)
Standard deviation across cells (log2)
Mean expression across cells (log2)
Early mouse development (Deng et al.)
All genes (n=22958)
m-scHK (n=830)


A
$\bar{r}$ = 0.98
h-scHK index (sub-sample 2)
h-scHK index (sub-sample 1)
$\bar{r}$ = 0.97
m-scHK index (sub-sample 2)
m-scHK index (sub-sample 1)
B
r = 0.58
m-scHK index
h-scHK index
C
h-scHK
820
256
574
m-scHK
D
h-bHK microarray
h-scHK
26
243
400
71
209
579
2945
h-bHK RNA-seq
E
m-scHK
h-scHK
247
324
17
76
425
226
5
154
2664
21
21
281
50
29
180
h-bHK microarray
h-bHK RNA-seq
F
RPL26
HINT1
AGPAT1
Density
Expression in individual cells (log2 RPKM)
G
GAPDH (human)
Gapdh (mouse)
ACTB (human)
Actb (mouse)
Density
Expression in individual cells (log2 RPKM)


A
Pre-defined P classes
K-means clustering with all or subset of genes (K=P)
Concordance measured by:
- Adjusted rand index (ARI),
- Purity,
- Fowlkes-Mallows (FM) index,
- Jaccard index
B
Early human development
E-MTAB-3929
Early mouse development
GSE45719
Adjusted rand index
All expressed genes
h-bHK microarray
Random subset
h-bHK RNA-seq subset
h-bHK RNA-seq
h-scHK
m-scHK
C
Early human development
E-MTAB-3929
Early mouse development
GSE45719
ARI
Purity
FM
Jaccard
All expressed genes
h-bHK microarray
Random subset
h-bHK RNA-seq subset
h-bHK RNA-seq
h-scHK
m-scHK
D
Retinal tissue (mouse)
GSE63472
Brain (mouse)
10x Genomics
Zero (%)
All expressed genes
h-bHK microarray
h-bHK RNA-seq
m-scHK
top-300 m-scHK
top-100 m-scHK


A
Human
HK index
Mixing proportion
Gaussian sigma
Gaussian mu
Zero (%)
Reg. Zero
F-statistics
Exons
GC (%)
Conservation
Expression
1
0.6
0.2
-0.2
-0.6
-1
Mouse
HK index
Mixing proportion
Gaussian sigma
Gaussian mu
Zero (%)
Reg. Zero
F-statistics
Exons
GC (%)
Conservation
Expression
1
0.6
0.2
-0.2
-0.6
-1
B
Number of exons (log2)
0
2
4
6
8
10
GC content (%)
20
30
40
50
60
70
80
90
Conservation (phyloP score)
−2
0
2
4
6
Expression (log2 RPKM)
0
2
4
6
8
10
12
h-scHK
m-scHK
All human genes
All mouse genes
h-bHK microarray
h-bHK RNA-seq