Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes

Liqing Zhang and Wen-Hsiung Li

Department of Ecology and Evolution, University of Chicago

Do housekeeping genes, which are turned on most of the time in almost every tissue, evolve more slowly than genes that are turned on only at specific developmental times or tissues? Recent large-scale gene expression studies enable us to have a better definition of housekeeping genes and to address the above question in detail. In this study, we examined 1,581 human-mouse orthologous gene pairs for their patterns of sequence evolution, contrasting housekeeping genes with tissue-specific genes. Our results show that, in comparison to tissue-specific genes, housekeeping genes on average evolve more slowly and are under stronger selective constraints as reflected by significantly smaller values of Ka/Ks. Besides stronger purifying selection, we explored several other factors that can possibly slow down nonsynonymous rates in housekeeping genes. Although mutational bias might slightly slow the nonsynonymous rates in housekeeping genes, it is unlikely to be the major cause of the rate difference between the two types of genes. The codon usage pattern of housekeeping genes does not seem to differ from that of tissue-specific genes. Moreover, contrary to the old textbook concept, we found that ~74% of the housekeeping genes in our study belong to multigene families, not significantly different from that of the tissue-specific genes (~70%). Therefore, the stronger selective constraints on housekeeping genes are not due to a lower degree of genetic redundancy.

Introduction

Housekeeping genes are genes that are always expressed in every tissue to maintain cellular functions (Watson et al. 1965). A vague, less strict definition, however, has been used in many studies (Butte, Dzau, and Glueck 2001; Duhig et al. 1998; Kagawa and Ohta 1990; Lercher, Urrutia, and Hurst 2002). For example, Lercher et al. (2002) analyzed serial analysis of gene expression (SAGE) data for 14 tissues in human and found that housekeeping genes tend to cluster on chromosomes. They defined "housekeeping genes" as genes having SAGE tags in $\geq$9 tissues of the total 14 tissues for the SAGE data, or expressed sequence tag (EST) presence in $\geq$19 tissues of the 60 tissues for the compiled EST data.

Recent large-scale gene expression studies have made it possible to examine the expression patterns of many genes at different developmental times and tissues and thus enable a more concrete description of housekeeping genes in the genomic scale. A proposed working concept of housekeeping genes has been "those genes critical to the activities that must be carried out for successful completion of the cell cycle" (Warrington et al. 2000). Several recent studies attempted to identify housekeeping genes genome-wide from a number of tissues and developmental stages. For example, Warrington et al. (2000) examined the expression levels of ~7,000 genes in 11 different human adult and fetal tissues using high-density oligonucleotide arrays, and identified 535 housekeeping genes that are turned on early in fetal development and stay on throughout adulthood in all tissues. Similarly, Hsiao et al. (2001) analyzed the expression pattern of 7,070 genes in 59 tissue samples representing 19 human tissue types, and identified 451 housekeeping genes, as expressed in all 19 tissues, among which 358 genes were also found in Warrington et al. (2000). Both studies show that housekeeping genes are not necessarily expressed at the same level across all tissues; rather, each tissue seems to have a specific expression profile of housekeeping genes. Furthermore, housekeeping genes are not necessarily the most highly expressed genes in all tissues. It is evident that the old concept and views are insufficient to reflect the nature of housekeeping genes. To better understand the role they play in the genome, more studies need to be conducted on the function, expression, and evolutionary patterns of these genes.

To date, few attempts have been made to examine how housekeeping genes evolve in general and how different they are from tissue-specific genes. A few studies have shown that broadly expressed genes tend to evolve more slowly than narrowly expressed genes (Duret and Mouchiroud 2000; Hastings 1996; Hughes and Hughes 1995). In the present study, taking advantage of the recent large-scale studies for identifying housekeeping genes, we addressed the issue of rate evolution in housekeeping genes. We used genes from Hsiao et al. (2001), because it has a larger sample of tissues than that of Warrington et al. (2000), and all data can be easily obtained from the associated Web site dedicated to an inventory of housekeeping genes. Furthermore, both studies share a majority of the housekeeping genes.

Materials and Methods

The data were downloaded from the Web site http://www.hugeindex.org/. The initial data set included 1,980 genes, with 451 housekeeping genes and 1,529 tissue-specific genes from brain, kidney, muscle, prostate, lung, liver, and vulva. We performed a BlastP search against the mouse National Center for Biotechnology Information (NCBI) RefSeq database. The best hits from mouse were used as orthologs to the human query sequences, and orthology was further verified by checking our data set with HOVERGEN (Duret, Mouchiroud, and Gouy 1994). The final data set for the subsequent analyses contains

Key words: synonymous rates, nonsynonymous rates, mutational bias, selective constraint, tissue-specific, and genetic redundancy.

E-mail: whli@uchicago.edu.

Mol. Biol. Evol. 21(2):236–239. 2004
DOI: 10.1093/molbev/msh010
Advance Access publication October 31, 2003
Molecular Biology and Evolution vol. 21 no. 2

Table 1
Numbers of Tissue-Specific Genes and Housekeeping Genes Studied

Tissue	Brain	Kidney	Lung	Liver	Muscle	Prostate	Vulva	Housekeeping	Total
Gene #	451	78	73	240	279	35	76	349	1,581

1,927 genes in total, with 412 housekeeping genes and 1,515 tissue-specific genes.

All protein sequences were aligned using ClustalW with the default parameters, and back-translated to their corresponding DNA sequences. The alignments were then visually inspected and modified when necessary. The number of substitutions per nonsynonymous site and the number of substitutions per synonymous site, denoted as Ka and Ks, respectively, were calculated using the maximum likelihood method implemented in the PAML package (Yang 1997). We excluded genes with Ks >1, leaving a data set of 1,581 genes in total for our final analyses (table 1).

For all measures of distances, including Ka, Ks and Ka/Ks, the Wilcoxon rank sum non-parametric test (Wilcoxon 1945) was applied to the pairwise comparison between each set of tissue-specific genes and housekeeping genes.

The effective number of codons (ENC), a measurement of codon usage bias that ranges from 20 to 61 (Wright 1990), was calculated for all genes. A gene with ENC equal to 20 uses only one type of codon for each synonymous codon set, and thus it shows the strongest codon usage bias, whereas a gene with ENC equal to 61 indicates no synonymous codon usage preference.

The conventional view on housekeeping genes has been that they are low-copy-number genes in the genome. If the notion holds, one can also attribute the slower nonsynonymous rate to less genetic redundancy in housekeeping genes. To examine this issue, we used gene families that have been compiled in the ENSEMBL database to determine the copy numbers for genes used in our study. Genes with more than one member in each species were grouped into multiple-copied gene families.

Results and Discussion

Rate Differences Between Housekeeping Genes and Tissue-Specific Genes

The average evolutionary rates of all genes are shown in table 2. Compared with tissue-specific genes (mean Ka = 0.083), housekeeping genes have on average lower nonsynonymous rates (mean Ka = 0.046; $P < 0.001$) and exhibit a smaller degree of variation. The synonymous rates are also lower in housekeeping genes (mean Ks = 0.447 vs. mean Ks = 0.492 for tissue-specific genes; P value < 0.001), but to a lesser degree (fig. 1).

Table 2
Means and Standard Deviations of Ka and Ks for Housekeeping and Tissue-Specific Genes

Tissue	Ka	Ks	Ka/Ks	ENC[a]
Brain	0.059 (0.062)	0.465 (0.136)	0.121 (0.125)	48.7 (6.5)
Kidney	0.089 (0.071)	0.517 (0.125)	0.166 (0.116)	48.3 (5.9)
Lung	0.150 (0.120)	0.540 (0.139)	0.259 (0.172)	47.9 (6.0)
Liver	0.131 (0.082)	0.546 (0.127)	0.233 (0.132)	50.6 (5.3)
Muscle	0.055 (0.059)	0.458 (0.150)	0.116 (0.112)	48.7 (6.3)
Prostate	0.066 (0.095)	0.540 (0.181)	0.108 (0.125)	48.6 (8.2)
Vulva	0.109 (0.085)	0.519 (0.152)	0.201 (0.139)	45.6 (7.5)
House-keeping	0.046 (0.063)	0.447 (0.146)	0.093 (0.114)	49.1 (6.6)

[a] ENC = effective number of codons.

The results shown here are comparable with those reported by Hastings (1994), who found that for 14 of 15 studied gene families, broadly expressed isoforms evolve slower than narrowly expressed isoforms. Duret and Mouchiroud (2000) found that genes that have ESTs presented in $\geq$16 tissues evolve more slowly in nonsynonymous sites than genes that have ESTs presented in $\leq$3 tissues. If classification of housekeeping genes is mainly based on the number of tissues in which genes are expressed, it is expected that the previous and current studies are consistent.

Possible Causes for Rate Differences

Why do housekeeping genes evolve more slowly in general than tissue-specific genes? To answer this question, we examined several factors that might affect evolutionary rates. First, the higher nonsynonymous substitution rates in tissue-specific genes could be due to higher mutation rates. However, this explanation is unlikely because if higher mutation rates are the main reason, we expect to see much higher synonymous substitution rates in tissue-specific genes than in housekeeping genes. The average synonymous rate in tissue-specific genes does not exhibit the same magnitude of increase as the average nonsynonymous substitution rate (10% increase vs. 100% increase), suggesting that mutation rate differences are not the main cause for the nonsynonymous rate difference.

Second, housekeeping genes might be under stronger selective constraints than tissue-specific genes. To compare selective constraints on these two types of genes, we calculated the Ka/Ks of these genes because Ka/Ks has commonly been used as an indicator of selective constraint. The average Ka/Ks for housekeeping genes is 0.093, whereas it increases more than twofold for lung-specific genes (mean Ka/Ks = 0.259) and liver-specific genes (mean Ka/Ks = 0.233). The Wilcoxon rank sum test shows that the average Ka/Ks is statistically higher for each of the different types of tissue-specific genes, except for prostate-specific genes, and for the pooled tissue-specific genes (table 3), indicating that, on average, tissue-specific genes are under weaker selective constraints than housekeeping genes. It is hard to know, however, to what extent selective constraint contributes to the rate difference, although previous studies have attributed the rate difference between broadly and narrowly expressed genes to be due solely to selective constraint differences (Duret and Mouchiroud 2000; Hastings 1996).

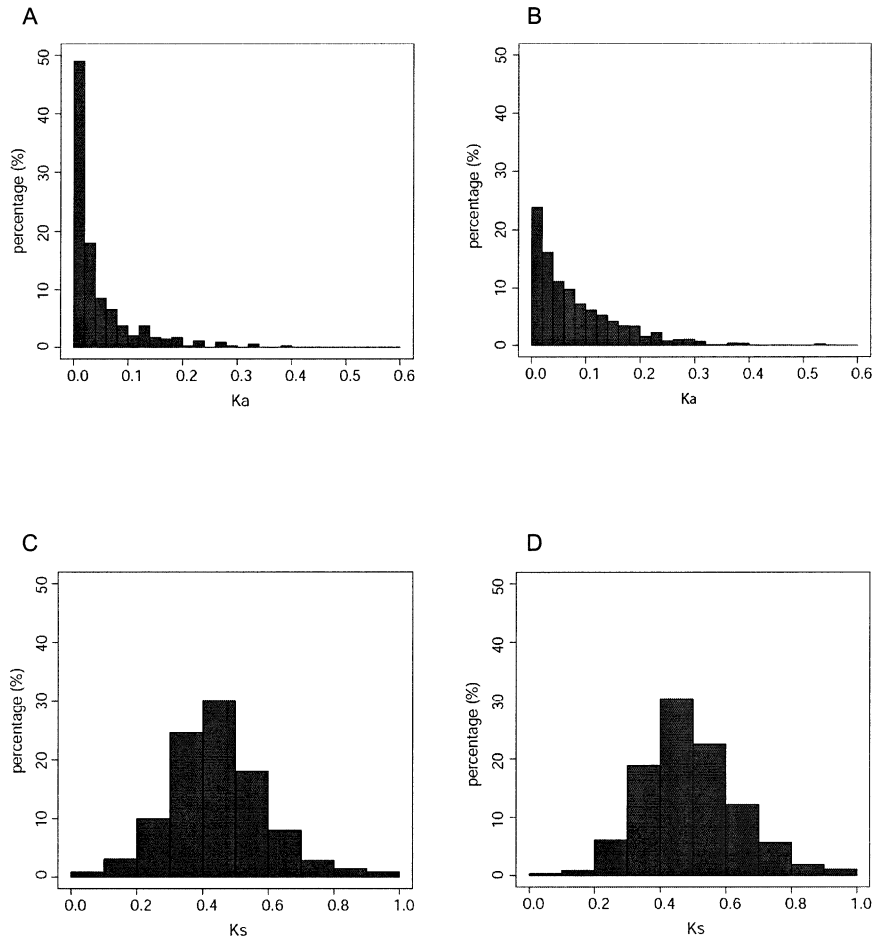

FIG. 1.—The distribution of Ka for housekeeping genes (A) and tissue-specific genes (B) and the distribution of Ks for housekeeping genes (C) and tissue-specific genes (D).

Selective constraint differences could arise from differences in gene function and expression. As housekeeping genes play a key role in the maintenance of most cells, strong purifying selection acts to preserve their normal function, whereas for tissue-specific genes, which are expressed in few tissues, the impact of a deleterious mutation is less than that of housekeeping genes. However, apart from this, could stronger selective constraints also be due to a lower degree of genetic redundancy in housekeeping genes than in tissue-specific genes? We used the copy number as a rough estimate of genetic redundancy to examine this issue. Two observations are notable: First, the percentages of genes belonging to multigene families are similar across all seven sets of tissue-specific genes (table 4), suggesting that different types of tissues harbor equivalent amounts of genetic redundancy estimated by gene copy number. Second, ~74% and ~70% of the housekeeping genes have ≥2 copies in the human and mouse genomes, respectively, and the average percentage of multigene families in tissue-specific genes is ~70% and ~69% in the human and mouse genomes, respectively. The chi-square tests show that the proportions of multigene families are not significantly different between housekeeping and tissue-specific genes ($\chi^2 = 2.13$, df = 1, $P = 0.79$ in human; $\chi^2 = 0.22$, df = 1, $P = 0.64$ in mouse). Therefore, our results show that, contrary to the traditional view, housekeeping genes harbor an amount of genetic redundancy similar to that of tissue-specific genes in the genome, and their greater selective constraint does not seem to be due to lower genetic redundancy.

To examine whether differences in selective constraints have any effect on codon usage, we calculated the ENC values for all genes studied. The average ENCs are

Table 3
Wilcoxon Rank Sum Tests Between Housekeeping Genes and Tissue-Specific Genes for the Evolutionary Rates

	Ka		Ks		Ka/Ks	
Tissue	w[a]	p[b]	w[a]	p[b]	w[a]	p[b]
Brain	94351.5	***	84604.5	0.034*	93476.5	***
Kidney	20206.5	***	17784.5	***	19762.5	***
Lung	21278	***	17398	***	21133.5	***
Liver	70093.5	***	59657	***	69259.5	***
Muscle	58240.5	***	50586	0.2	58285.5	***
Prostate	7004.5	0.076	7871	0.002**	6624.5	0.2
Vulva	20367.5	***	16659	***	20209	***
Pooled	291542	***	254560	***	288751	***

[a] w = Wilcoxon rank sum.
[b] *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

Table 4
Percentage of Genes Having More than or Equal to Two Copies in Human and Mouse Genome

Tissue	Brain	Kidney	Lung	Liver	Muscle	Prostate	Vulva	Tissue-average	Housekeeping
Human	70.0	64.4	61.0	70.2	72.5	75.8	65.7	69.7	73.8
Mouse	69.1	70.5	75.0	67.8	67.2	74.3	67.7	68.9	70.3

49.1 and 48.8 for housekeeping and tissue-specific genes, respectively (table 2). Therefore, even though housekeeping genes are ubiquitously expressed, they do not show stronger codon usage bias. Together with the observation that housekeeping genes are not necessarily the highly expressed genes in a specific tissue (Hsiao et al. 2001; Warrington et al. 2000), our result suggests that a gene's expression level is a more important factor than expression breadth in determining the gene's codon usage preference.

Although it is not clear how many housekeeping genes are needed for the normal function of an organism, the evolution of this subset of genes does provide us a glance at the evolution of these important genes. More studies are needed to better characterize housekeeping genes. For example, how to define housekeeping genes in terms of their function and expression? How are housekeeping genes organized in the genome? Is there a minimum number of housekeeping genes for the normal function of cells, and if there is, does this number vary with the complexity of the organism? Finally, how different are housekeeping genes from tissue-specific genes in expressional, functional, and evolutionary perspective?

Acknowledgments

This study was supported by National Institutes of Health (NIH) grants GM30998 and GM66104. We thank Jing Yang for comments.

Literature Cited

Butte, A. J., V. J. Dzau, and S. B. Glueck. 2001. Further defining housekeeping, or "maintenance," genes focus on "A compendium of gene expression in normal human tissues". Physiol. Genomics **7**:95–96.

Duhig, T., C. Ruhrberg, O. Mor, and M. Fried. 1998. The human surfeit locus. Genomics **52**:72–78.

Duret, L., and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68–70.

Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN, a database of homologous vertebrate genes. Nucleic Acids Res. **22**:2360–2365.

Hastings, K. E. M. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J. Mol. Evol. **42**:631–640.

Hsiao, L.-L., F. Dangond, T. Yoshida, et al. (23 co-authors). 2001. A compendium of gene expression in normal human tissues. Physiol. Genomics **7**:97–104.

Hughes, A. L., and M. K. Hughes. 1995. Self peptides bound by HLA class I molecules are derived from highly conserved regions of a set of evolutionarily conserved proteins. Immunogenetics **41**:257–262.

Kagawa, Y., and S. Ohta. 1990. Regulation of mitochondrial ATP synthesis in mammalian cells by transcriptional control. Int. J. Biochem. **22**:219–229.

Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. **31**:180–183.

Warrington, J. A., A. Nair, M. Mahadevappa, and M. Tsyganskaya. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. Physiol. Genomics **2**:143–147.

Watson, J. D., N. H. Hopkins, J. W. Roberts, J. A. Steitz, and A. M. Weiner. 1965. Molecular biology of the gene, vol. 1. p 704. Benjamin/Cummings, Menlo Park, Calif.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. Biometrics **1**. 80–83.

Wright, F. 1990. The 'effective number of codons' used in a gene. Gene **87**:23–29.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

Naruya Saitou, Associate Editor

Accepted August 26, 2003