Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 360 (2007) 586–592
BBRC
www.elsevier.com/locate/ybbrc

Selection for the miniaturization of highly expressed genes

Shu-Wei Li, Liang Feng, Deng-Ke Niu *

MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China

Received 22 May 2007
Available online 26 June 2007

Abstract

Most widely expressed genes are also highly expressed. Based on high or wide expression, different models were proposed to explain the small sizes of highly/widely expressed genes. We found that housekeeping genes are not more compact than narrowly expressed genes with similar expression levels, but compactness and expression level are correlated in housekeeping genes (except that highly expressed *Arabidopsis* HK genes have longer intron length). Meanwhile, we found evidence that genes with high functional/regulatory complexity do not have longer introns and longer proteins. The genome design hypothesis is thus not supported. Furthermore, we found that housekeeping genes are not more compact than the narrowly expressed somatic genes with similar average expression levels. Because housekeeping genes are expected to have much higher germline expression levels than narrowly expressed somatic genes, transcription-associated deletion bias is not supported. Selection of the compactness of highly expressed genes for economy is supported.

Keywords: Intron length; Protein length; Intron number; Protein domain number; Energetic cost; Gene expression

Protein size and intron size vary considerably within each organism, but the reasons for this are not established. Early studies in diverse organisms including human, *Drosophila*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*, suggest that there are negative relationships between expression level and protein length [1–7]. Meanwhile, highly expressed genes of human and *C. elegans* were found to have short introns [3,5,6]. A hypothesis is that natural selection favors short proteins and short introns to minimize the energetic cost of gene expression [1,3,8]. In addition, transcription-associated deletion could also generate the negative relationship between germline expression level and the sizes of proteins and introns [5,6], because the majority of genes may be physically and/or ectopically expressed in germline [9–12]. On the other hand, researchers found that animal housekeeping (HK) genes have shorter proteins and shorter introns than narrowly expressed genes [13,14]. It seems that tissue-specific genes encode more complex proteins, thus require more regulatory elements that reside in intergenic spacers and introns. As HK genes are often highly expressed and tissue-specific genes are often expressed at low levels (Supplementary Table S1) [14,15], Vinogradov [14] suggested that the energetic cost hypothesis should be replaced by a genome design hypothesis.

Most recent works seem to support the genome design hypothesis. In *Drosophila*, intron length is negatively correlated with protein divergence and protein polymorphism, which indicates that long introns may play a role in regulating gene expression [16,17]. In human, multi-species conserved sequences, which may have important regulatory functions in gene expression, were found to be enriched in long introns [18]. *Dystrophin* is one of the largest and most complex genes in various organisms from human to *Drosophila*. The promoters of the *Dystrophin* gene are situated in the largest introns and the size of large introns that contain promoters is conserved [19,20]. More and more regulatory elements have been found in introns, especially in the first introns [21]. Consistently, first introns are usually longer than other introns, and the aberrantly long

* Corresponding author. Fax: +86 10 58807721.
E-mail addresses: dkniu@bnu.edu.cn, dengkeniu@hotmail.com (D.-K. Niu).

0006-291X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2007.06.085

intron is generally the first intron [22]. Vinogradov [23] analyzed the consecutive local alignments between human and mouse intronic DNA sequences, and showed that the introns of tissue-specific genes tend to have higher fraction of aligned sequences than those of HK genes. He also found that the amount of aligned intronic DNA correlates with the number of protein domains [23].

However, some studies support the energetic cost hypothesis, or disagree with the genome design hypothesis. If intron length is mainly determined by the selection to reduce the cost of transcription, genes subject to strong purifying selection should have short introns, which has been observed in the *Arabidopsis thaliana* genes expressed in pollen [24]. Later analyses of human antisense genes further support the economy selection hypothesis (i.e. selection to minimize energetic cost and/or time cost of gene expression [25]), but do not uphold the genome design hypothesis [26].

In addition, there is also evidence that regulatory elements may shape the intron size of weakly or intermediately expressed genes while the selection to reduce the cost of transcription may be dominant in highly expressed genes [27].

The above debates can be summarized to two questions. First, which genes are compact, highly expressed genes, widely expressed genes, or both of them? Second, if highly expressed genes are compact, which is the dominant force, the selection to minimize the energetic cost of gene expression or transcription-associated nonadaptive deletion bias?

Materials and methods

The gene characters were parsed from the annotated genomes downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/): *Homo sapiens* (build 36 version 1), *Mus musculus* (build 35 version 1), and *A. thaliana* (updated Nov. 04, 2005). The protein characters were estimated using the SwissPfam version 20 (ftp://ftp.genetics.wustl.edu/pub/Pfam/). In the case of alternative splicing variants, we retained the longest mRNA for analysis.

We determined the gene expression breadth, level and tissue specificity [28] in human and mouse by the microarray data, GNF GeneAtlas Version 2 [29], and those in *A. thaliana* by the microarray data from Yale Plant Genomics (http://plantgenomics.biology.yale.edu/) [30]. We also used other gene expression data to validate the results based on microarray data (See Supplementary materials and methods). The tissue specificity index (τ) is defined as:

$$\tau = \frac{\sum_i^N (1 - \frac{x_i}{x_{\max}})}{N-1}$$

where N is the number of tissue/organ samples examined, x_i is the expression level of the gene in sample i and $x_{\max}$ is the highest expression level of the gene across the N samples examined [28]. See Supplementary materials and methods for details.

As recommended [31], a gene was assumed to be expressed in a tissue if its average difference (AD) value was greater than the threshold of 200 in that tissue (Using 100, 150, 250, and 300 as the thresholds gave similar results, Supplementary Tables S2–S5 and Figs. S1–S8). We defined the HK genes as those ubiquitously expressed in all normal tissue/organ samples. The tissue/organ samples of the human and mouse microarray gene expression datasets [29] are overlapped. For example, whole brain, cerebellum, and cerebellum peduncles are listed as independent samples side by side in the gene expression data. When a stringent definition of tissue-specific genes is used (i.e. genes expressed in only one sample), only a limited number of genes are remained, giving too small a sample size to study. So narrowly expressed genes were defined as those expressed in less than 20% of total normal samples (although using 15% and 25% as the thresholds to define the narrowly expressed genes gave similar results, data not shown). For somatic cells, narrowly expressed genes in human and mouse were defined as those expressed in less than 20% of total normal samples excluding germline cells, reproductive organs, or early developmental stage (again using 15% and 25% as the thresholds gave similar results, data not shown). In *A. thaliana*, genes expressed in all the six normal organs (inflorescence, flower, cauline leaf, silique, seedling, and root) were defined as HK genes and those expressed only in one organ were defined as narrowly expressed genes. The narrowly expressed somatic genes are those narrowly expressed genes that are not expressed in inflorescence, flower, or silique.

Results

Highly expressed genes are compact, but widely expressed genes are not

HK genes are ubiquitously expressed in all tissues, so their size evolution is not related to expression breadth. We found that the compactness of HK genes is correlated with their expression levels except that highly expressed *Arabidopsis* HK genes have longer introns (Table 1).

Besides ubiquitous expression, some researchers used other additional criteria (e.g. basic cellular functions) to define HK genes more stringently [13]. In well-compiled sets of human housekeeping genes, we got stronger negative correlations (i.e. mostly have higher correlation coefficients, Table S6). Interestingly, we found that smaller samples randomly selected from the HK genes analyzed in Table 1 tend to have higher absolute values of correlation coefficients (Supplementary Fig. S9). Thus, the low absolute values of correlation coefficients in Table 1 may be attributed to large sample size, and they have the same biological significance as the higher correlation coefficients

Table 1
Spearman correlations of average expression level with various characters of housekeeping genes

	Homo sapiens			*Mus musculus*			*Arabidopsis thaliana*		
	n	*r*	*P*	*n*	*r*	*P*	*n*	*r*	*P*
Average intron length	2879	−0.232	10^{-6}	1768	−0.160	10^{-6}	7968	0.074	10^{-6}
First intron length	2879	−0.166	10^{-6}	1768	−0.086	3×10^{-4}	7968	0.058	10^{-6}
Intron number	2879	−0.216	10^{-6}	1768	−0.156	10^{-6}	7968	−0.073	10^{-6}
Protein length	2564	−0.317	10^{-6}	1742	−0.210	10^{-6}	6440	−0.128	10^{-6}
Protein domain number	2564	−0.199	10^{-6}	1742	−0.127	10^{-6}	6440	−0.084	10^{-6}

obtained from analyzing small samples. So we use the HK genes selected just by ubiquitous expression in further analyses.

To test whether functional complexity of tissue-specific expression has produced both large introns and large proteins in tissue-specific genes [14], we compared the genes

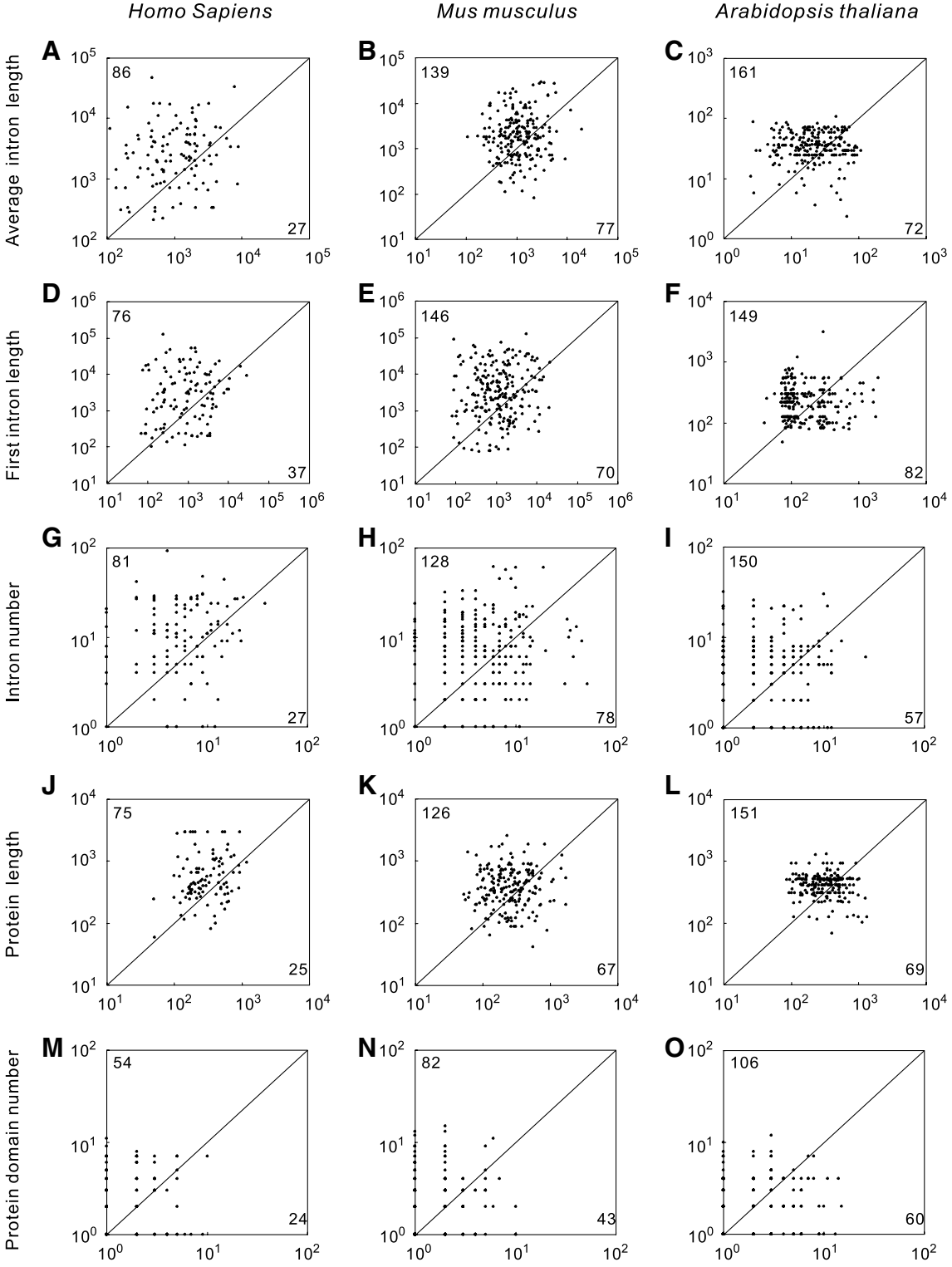

Fig. 1. Comparing housekeeping genes and narrowly expressed genes with similar average expression levels. The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: (A) 113, $P < 10^{-8}$; (B) 216, $P < 10^{-7}$; (C) 233, $P < 10^{-5}$; (D) 113, $P < 10^{-5}$; (E) 216, $P < 10^{-11}$; (F) 233, $P < 10^{-3}$; (G) 113, $P < 10^{-7}$; (H) 216, $P < 10^{-6}$; (I) 233, $P < 10^{-9}$; (J) 100, $P < 10^{-6}$; (K) 193, $P < 10^{-6}$; (L) 220, $P < 10^{-5}$; (M) 100, $P < 10^{-3}$; (N) 193, $P < 10^{-4}$; and (O) 220, $P = 0.018$.

with similar expression levels but differing greatly in expression breadth, i.e. HK genes and narrowly expressed genes with similar expression levels (for details, see Supplementary methods). Pairwise comparisons showed that the HK genes are less compact than narrowly expressed genes with similar expression levels (Fig. 1A–O), being opposite to that predicted by the genome design hypothesis [14]. The first introns contain more regulatory elements than other introns [21], but their sizes in HK genes are longer than in narrowly expressed genes with similar average expression levels (Fig. 1D–F).

Previous observation that most tissue-specific genes have both long proteins and long introns [13,14] should be attributed to their low-level expression, rather than their functional or regulatory complexity. Thus, our results disagree with the genome design hypothesis [14].

Genes with complex expression pattern do not have longer introns and longer proteins

The tissue specificity index τ [28] measures both qualitative variations (i.e. presence/absence) and quantitative variations of expression level among tissues/organs. Obviously, τ is more representative than expression breadth for the expression complexity of a gene (Supplementary discussion and Table S7). According to the genome design hypothesis [14], the genes with complex expression patterns should have longer introns and longer proteins. However, we found that the genes with higher τ are not less compact than those with lower τ (Table 2, Supplementary Fig. S10a–o). Contrarily, the tissue specificity is negatively (weakly, but significantly) correlated with intron length, intron number, protein length, and protein domain number.

Recent evidence indicates that the genes with intermediate expression breadths may have higher functional and regulatory complexity [15]. To control for the expression level, we pairwisely compared the HK genes and intermediated expressed genes with similar expression levels (see Supplementary methods for details) and found that intermediately expressed genes are not less compact than HK genes with similar expression levels (Supplementary Table S8). Thus, the relatively longer introns and longer proteins previously found in intermediated expressed human genes [15] should be attributed to their relatively lower expression levels rather than their expression complexity.

Germline transcribed genes are not compact

The observation that highly expressed genes are compact can be explained by either the selection to minimize the energetic cost of gene expression or transcription-associated nonadaptive deletion bias. Meanwhile, fewer intron numbers may also be explained by mutational bias. Introns may be lost through a recombination between a reverse transcript and the corresponding genomic DNA [32–34]. Highly expressed genes have more potential substrates (i.e. mRNA) for reverse transcription, and thus are more likely to lose their introns.

Only the transcription-associated mutations that occurred in germline cells could accumulate in evolution. So the evolution of somatic-tissue-specific genes (STS, i.e. expressed in only one particular somatic tissue) is free from transcription-associated mutational bias. If highly expressed STS genes are more compact than weakly expressed STS genes, we could reject the hypotheses on transcription-associated deletion bias and mRNA-mediated intron losses. Unfortunately, the limited number of STS genes showed inconclusive results (data not shown).

We therefore pairwisely compared HK genes and narrowly expressed somatic genes with most similar average expression levels (see Supplementary methods for details). Although we do not know the exact germline expression levels of the HK genes and the narrowly expressed somatic genes, it is reasonable to assume that HK genes have much higher germline expression level than narrowly expressed somatic genes with the consideration of ectopical expression [9–12]. According to the transcription-associated mutation hypothesis, HK genes should be more compact than the narrowly expressed somatic genes. However, we found that the HK genes are less compact than narrowly expressed somatic genes with similar expression levels in human and mouse (Fig. 2A–O). In *A. thaliana*, HK genes have longer first intron length, higher intron number, and longer protein length than narrowly expressed somatic genes with similar expression levels (Fig. 2F, I, and L), but there are no significant differences in average intron length or protein domain number (Fig. 2C and O). Nevertheless, HK genes are not more compact than narrowly expressed somatic genes with similar expression levels. Therefore, nonadaptive transcription-associated mutational bias is not the dominant force.

Table 2
Spearman correlations of tissue specificity with gene and protein characters

	Homo sapiens			*Mus musculus*			*Arabidopsis thaliana*		
	n	*r*	*P*	*n*	*r*	*P*	*n*	*r*	*P*
Average intron length	15009	−0.079	10^{-6}	14677	−0.063	10^{-6}	16589	−0.120	10^{-6}
First intron length	15009	−0.061	10^{-6}	14677	−0.074	10^{-6}	16589	−0.130	10^{-6}
Intron number	15009	−0.081	10^{-6}	14677	−0.109	10^{-6}	16589	−0.197	10^{-6}
Protein length	13438	−0.083	10^{-6}	14730	−0.045	10^{-6}	14352	−0.061	10^{-6}
Protein domain number	13438	−0.036	$<10^{-4}$	14730	−0.020	0.013	14352	−0.036	10^{-4}

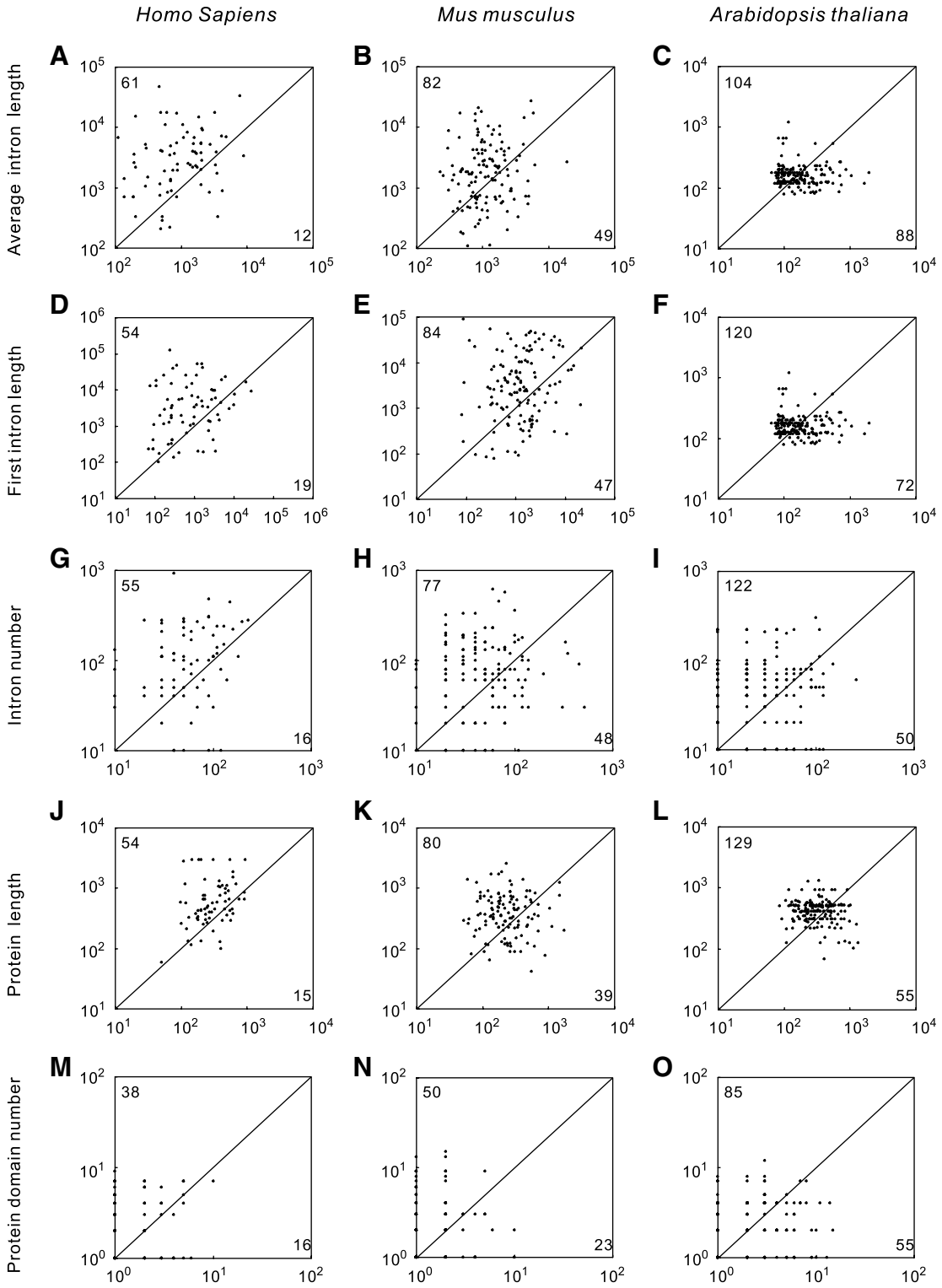

Fig. 2. Comparing housekeeping genes and narrowly expressed somatic genes with similar average expression levels. The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed somatic counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed somatic genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: (A) 73, $P < 10^{-8}$; (B) 131, $P < 10^{-4}$; (C) 192, $P = 0.608$; (D) 73, $P < 10^{-4}$; (E) 131, $P < 10^{-6}$; (F) 192, $P = 0.017$; (G) 73, $P < 10^{-5}$; (H) 131, $P < 10^{-3}$; (I) 192, $P < 10^{-7}$; (J) 69, $P < 10^{-6}$; (K) 119, $P < 10^{-4}$; (L) 184, $P < 10^{-5}$; (M) 69, $P = 0.001$; (N) 119, $P < 10^{-3}$; and (O) 184, $P = 0.161$.

Discussion

In *A. thaliana*, the highly expressed genes tend to have longer introns (Table 1, [35]), while the genes with complex expression pattern do not have longer introns (Table 2). Thus, the majority of introns in plants may not play important roles in expression complexity other than increasing expression level [36].

We successively find evidence against genome design hypothesis [14] as well as the transcription-associated mutational bias hypothesis [5,6], and support the energetic cost hypothesis [1,3,8]. By assuming the majority of mammalian intronic sequences as dispensable junks, it is easy to understand the selection for short introns to minimize energetic cost of gene expression [3]. The dominant force may be negative selection against insertions and/or positive selection for deletions [3].

But the debate will not settle down. Besides the energetic cost hypothesis, there is another (but not mutual exclusive) explanation for the negative correlation between intron length and transcription level. Genes with continuously low-level transcription may require longer introns to avoid R-loop formation than busily transcribed genes [37].

We also found that highly expressed genes are miniaturized in several other aspects including protein length, protein domain number, and intron number. Protein sequences are generally under stronger selective pressure than intron sequences. But long proteins cost much more energy than long introns because mRNA molecules are generally translated many times. Much evidence shows that highly expressed genes avoid using expensive amino acids [2,4,5]. The fact that highly expressed genes have compact proteins may be due to selection for deletions of less essential (but not necessarily neutral) amino acid residues.

Most protein domains are unlikely to be junky. We suppose that natural selection may favor gene fission and act against gene fusion in highly expressed genes. Gene fission could reduce the energetic cost of gene expression in two ways: (1) Splitting a multi-domain protein to several single-domain proteins would save the energy that was used to build the loops connecting domains. (2) Subfunctionalization of a pleiotropic protein into several small proteins with specific functions would prevent redundant expression of some regulatory domains. Consistently, majority of the short proteins less than 100 amino acids in mouse proteome are expressed in a highly tissue-restricted fashion [38]. The selection favoring gene fission and against gene fusion would also generate lower intron number in highly expressed genes.

Recent evidence suggests that mRNA concentrations are not the major determinant factor of protein abundances [39]. In analyzing the energetic cost of protein synthesis, using mRNA concentrations as a surrogate for protein abundances is a pitfall of this study. We look forward to large-scale protein abundance data in multicellular organisms.

Acknowledgments

We thank the anonymous referee for useful comments. This research was supported by NSFC (30270695) and BNU.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2007.06.085.

References

[1] A. Coghlan, K.H. Wolfe, Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*, Yeast 16 (2000) 1131–1145.
[2] R. Jansen, M. Gerstein, Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins, Nucleic Acids Res. 28 (2000) 1481–1488.
[3] C.I. Castillo-Davis, S.L. Mekhedov, D.L. Hartl, E.V. Koonin, F.A. Kondrashov, Selection for short introns in highly expressed genes, Nat. Genet. 31 (2002) 415–418.
[4] H. Akashi, Translational selection and yeast proteome evolution, Genetics 164 (2003) 1291–1303.
[5] A.O. Urrutia, L.D. Hurst, The signature of selection mediated by expression on human genes, Genome Res. 13 (2003) 2260–2264.
[6] J.M. Comeron, Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence, Genetics 167 (2004) 1293–1304.
[7] J. Warringer, A. Blomberg, Evolutionary constraints on yeast protein size, BMC Evol. Biol. 6 (2006) 61.
[8] L.D. Hurst, G. McVean, T. Moore, Imprinted genes have few and small introns, Nat. Genet. 12 (1996) 234–237.
[9] G. Sarkar, S.S. Sommer, Access to a messenger RNA sequence or its protein product is not limited by tissue or species specificity, Science 244 (1989) 331–334.
[10] J. Chelly, J.P. Concordet, J.C. Kaplan, A. Kahn, Illegitimate transcription: transcription of any gene in any cell type, Proc. Natl. Acad. Sci. USA 86 (1989) 2617–2621.
[11] Y. Kimoto, A single human cell expresses all messenger ribonucleic acids: the arrow of time in a cell, Mol. Gen. Genet. 258 (1998) 233–239.
[12] D.-K. Niu, Low-level illegitimate transcription of genes may be to silence the genes, Biochem. Biophys. Res. Commun. 337 (2005) 413–414.
[13] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, Trends Genet. 19 (2003) 362–365.
[14] A.E. Vinogradov, Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. 20 (2004) 248–253.
[15] A.E. Vinogradov, 'Genome design' model and multicellular complexity: golden middle, Nucleic Acids Res. 34 (2006) 5906–5914.
[16] P. Haddrill, B. Charlesworth, D. Halligan, P. Andolfatto, Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content, Genome Biol. 6 (2005) R67.
[17] N. Petit, S. Casillas, A. Ruiz, A. Barbadilla, Protein polymorphism is negatively correlated with conservation of intronic sequences and complexity of expression patterns in *Drosophila melanogaster*, J. Mol. Evol. 64 (2007) 511–518.
[18] M. Sironi, G. Menozzi, G.P. Comi, R. Cagliani, N. Bresolin, U. Pozzoli, Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences, Hum. Mol. Genet. 14 (2005) 2533–2546.

[19] U. Pozzoli, G. Elgar, R. Cagliani, L. Riva, G.P. Comi, N. Bresolin, A. Bardoni, M. Sironi, Comparative analysis of vertebrate dystrophin loci indicate intron gigantism as a common feature, Genome Res. 13 (2003) 764–772.
[20] S. Neuman, M. Kovalio, D. Yaffe, U. Nudel, The *Drosophila* homologue of the dystrophin gene—introns containing promoters are the major contributors to the large size of the gene, FEBS Lett. 579 (2005) 5365–5371.
[21] J.-V. Chamary, L.D. Hurst, Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage, Mol. Biol. Evol. 21 (2004) 1014–1023.
[22] E.V. Kriventseva, M.S. Gelfand, Statistical analysis of the exon-intron structure of higher and lower eukaryote genes, J. Biomol. Struct. Dyn. 17 (1999) 281–288.
[23] A.E. Vinogradov, ''Genome design'' model: evidence from conserved intronic sequence in human-mouse comparison, Genome Res. 16 (2006) 347–354.
[24] C. Seoighe, C. Gehring, L.D. Hurst, Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction, PLoS Genet. 1 (2005) e13.
[25] J. Chen, M. Sun, L.D. Hurst, G.G. Carmichael, J.D. Rowley, Human antisense genes have unusually short introns: evidence for selection for rapid transcription, Trends Genet. 21 (2005) 203–207.
[26] J. Chen, M. Sun, J.D. Rowley, L.D. Hurst, The small introns of antisense genes are better explained by selection for rapid transcription than by ''genomic design'', Genetics 171 (2005) 2151–2155.
[27] U. Pozzoli, G. Menozzi, G.P. Comi, R. Cagliani, N. Bresolin, M. Sironi, Intron size in mammals: complexity comes to terms with economy, Trends Genet. 23 (2007) 20–24.
[28] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, O. Shmueli, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, Bioinformatics 21 (2005) 650–659.
[29] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, J.B. Hogenesch, A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. USA 101 (2004) 6062–6067.
[30] L. Ma, N. Sun, X. Liu, Y. Jiao, H. Zhao, X.W. Deng, Organ-specific expression of *Arabidopsis* genome during development, Plant Physiol. 138 (2005) 80–91.
[31] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, J.B. Hogenesch, Large-scale analysis of the human and mouse transcriptomes, Proc. Natl. Acad. Sci. USA 99 (2002) 4465–4470.
[32] G.R. Fink, Pseudogenes in yeast? Cell 49 (1987) 5–6.
[33] T. Mourier, D.C. Jeffares, Eukaryotic intron loss, Science 300 (2003) 1393.
[34] D.-K. Niu, W.-R. Hou, S.-W. Li, mRNA-mediated intron losses: evidence from extraordinarily large exons, Mol. Biol. Evol. 22 (2005) 1475–1481.
[35] X.-Y. Ren, O. Vorst, M.W.E.J. Fiers, W.J. Stiekema, J.-P. Nap, In plants, highly expressed genes are the least compact, Trends Genet. 22 (2006) 528–532.
[36] H. Le Hir, A. Nott, M.J. Moore, How introns influence and enhance eukaryotic gene expression, Trends Biochem. Sci. 28 (2003) 215–220.
[37] D.-K. Niu, Protecting exons from deleterious R-loops: a potential advantage of having introns, Biol. Direct 2 (2007) 11.
[38] M.C. Frith, A.R. Forrest, E. Nourbakhsh, K.C. Pang, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, T.L. Bailey, S.M. Grimmond, The abundance of short proteins in the mammalian proteome, PLoS Genet. 2 (2006) e52.
[39] L. Nie, G. Wu, W. Zhang, Correlation between mRNA and protein abundance in *Desulfovibrio vulgaris*: A multiple regression to identify sources of variations, Biochem. Biophys. Res. Commun. 339 (2006) 603–610.

Supplementary materials and methods

Microarray gene expression data manipulation

In total, 73 non-disease human tissue/organ samples and 61 non-disease mouse tissue/organ samples were included in this study. As recommended [1], a gene was assumed to be expressed in a tissue if its average difference (AD) value was greater than the threshold of 200 in that tissue. A gene's expression level was given by the average of the gene's expression values in all 73 human samples or in all 61 mouse samples and its expression breadth is a simple count of the samples where the gene is expressed. In defining the tissue specificity index τ of each gene, we used the $\log_2$-transformed values of gene expression levels. In analyzing tissue specificity index τ, the cut-off of 200 in AD value was not used, so the genes with expression levels lower than 200 were included without any manipulation.

For the *Arabidopsis thaliana* gene expression data from Yale Plant Genomics (http://plantgenomics.biology.yale.edu/), we assigned the expression values of the genes marked by A (absent) to be zero. To avoid negative value of $\log_2$-transformed expression level in defining the tissue specificity index τ in *A. thaliana*, the expression level was assigned to be 1 if it is below 1.

SAGE and MPSS data analysis

The Serial Analysis of Gene Expression (SAGE) data of human and mouse [2] were downloaded from NCBI SAGEmap (ftp://ftp.ncbi.nih.gov/pub/sage) [3]. The NlaIII map data were curated and converted to relative expression level values (cpm, counts per million) following the approach of Lercher et al. [4]. The resulting SAGE data include 91 libraries representing 26 human non-diseased tissues/organs, and 104 libraries representing 14 mouse non-diseased tissues/organs. HK genes were defined as those genes expressed in 24 or more human tissues/organs, or in all 14 mouse tissues/organs. Narrowly expressed genes are those expressed in no more than 5 human tissues/organs or 3 mouse tissues/organs. In calculating the tissue specificity index τ from SAGE data in human and mouse, the expression values were not $\log_2$-transformed because of too many values smaller than 1.

For *A. thaliana*, the 14 MPSS libraries [5] analyzed in reference [6] include two types of MPSS data, from Classic method or from Signature MPSS method. The Signature method was developed to reduce the bias existed in Classic data [5]. So we also selected the 11 libraries (AP1, AP3, AGM, INS, ROS, SAP, S04, S52, LES, GSE, SIS) that are sequenced using Signature method. Analyzing this 11 libraries gave similar results as the 14 libraries analyzed in reference [6] (data not shown). In calculating the tissue specificity index τ, the expression values in MPSS data were not $\log_2$-transformed because there are too many values smaller than 1.

Analyzing SAGE and MPSS data also shows that highly expressed genes are more compact than weakly expressed genes, but widely expressed genes are not more compact than narrowly expressed genes, consistent with the results from analyzing microarray data (Tables S9-S10 and Figs. S11-S12).

Intermediately expressed gene

Referring to Vinogradov [7], we defined intermediately expressed gene as those expressed in 19-66 human tissue/organ samples or 13-57 mouse tissue/organ samples.

Defining gene pairs with similar expression level

In searching HK-narrowly expressed (somatic) gene pairs (or HK-intermediately expressed gene pairs), we designate the difference in the average expression level of each gene pair as

$$D = \left| \frac{A - B}{B} \right|$$

where A is the average expression level of a HK gene and B is the average expression level of a narrowly expressed gene (or an intermediately expressed gene). The significant value of Wilcoxon signed ranks test was used to define the lower limit of D values. In human and *A. thaliana*, setting $D <$ 5% is enough to obtain gene pairs with similar average expression levels ($P > 0.05$), while in mouse, such within-pair difference must be decreased to < 1%. If two or more HK genes can be paired with one narrowly expressed gene, we selected the one that produces minimized D value. Lists of the gene pairs are available upon request.

Supplementary discussion

The expression breadth measures only the presence/absence variations of gene expression among different tissues/organs. As exemplified in Table S2, genes with the same expression breadth may differ greatly in the variations of expression levels among tissue/organ samples. The tissue specificity index τ [8] measures both the qualitative variations (i.e. presence/absence) and the quantitive variations in expression level among tissue/organ samples. Apparently, a gene quantitatively adjusted to different expression level in different tissues/organs should require more regulatory elements, e.g. gene *AT3G56310* should require more regulatory elements than gene *AT2G32930* (Table S2). According to the genome design hypothesis [9], the genes with higher values of index τ should have longer introns than those with lower τ values. So we also tested the genome design hypothesis [9] by analyzing the tissue specificity index τ. More importantly in methodology, the tissue specificity index τ is not so highly correlated with average gene expression level as the expression breadth (Supplementary Table S1). This provides us a more feasible way to study the size evolution of introns and proteins.

Supplementary tables

Table S1. Spearman correlations of expression level with expression breadth and tissue specificity index τ[a]

	Microarray			SAGE/MPSS		
Species	n	$r_{Level,\ Breadth}$	$r_{level,\ \tau}$	n	$r_{Level,\ Breadth}$	$r_{Level,\ \tau}$
Homo sapiens	22141	0.914	–0.353	15556	0.917	–0.758
Mus musculus	23218	0.929	–0.389	16276	0.879	–0.686
Arabidopsis thaliana	21319	0.801	–0.762	21840	0.743	–0.574

[a]See the (Supplementary) Materials and methods section in the main text for the data source and the definition of tissue specificity index τ [8]. All the correlations are significant at the 10^{-6} level.

Table S2. Spearman correlations of average expression level with various characters of housekeeping genes (AD cutoff = 100)

	Homo sapiens			*Mus musculus*		
	n	*R*	*P*	*n*	*r*	*P*
Average intron length	5808	–0.195	10^{-6}	4801	–0.214	10^{-6}
First intron length	5807	–0.119	10^{-6}	4801	–0.097	10^{-6}
Intron number	5808	–0.165	10^{-6}	4801	–0.077	10^{-6}
Protein length	5195	–0.269	10^{-6}	4744	–0.157	10^{-6}
Protein domain number	5195	–0.164	10^{-6}	4744	–0.097	10^{-6}

Table S3. Spearman correlations of average expression level with various characters of housekeeping genes (AD cutoff = 150)

	Homo sapiens			*Mus musculus*		
	n	*R*	*P*	*n*	*r*	*P*
Average intron length	3986	–0.211	10^{-6}	2906	–0.182	10^{-6}
First intron length	3985	–0.144	10^{-6}	2906	–0.081	10^{-5}
Intron number	3986	–0.196	10^{-6}	2906	–0.090	10^{-6}
Protein length	3554	–0.296	10^{-6}	2876	–0.176	10^{-6}
Protein domain number	3554	–0.182	10^{-6}	2876	–0.112	10^{-6}

Table S4. Spearman correlations of average expression level with various characters of housekeeping genes (AD cutoff = 250)

	Homo sapiens			*Mus musculus*		
	n	*R*	*P*	*n*	*r*	*P*
Average intron length	2110	–0.223	10^{-6}	1087	–0.130	2×10^{-5}
First intron length	2110	–0.166	10^{-6}	1087	–0.103	7×10^{-4}
Intron number	2110	–0.243	10^{-6}	1087	–0.157	10^{-6}
Protein length	1867	–0.352	10^{-6}	1074	–0.252	10^{-6}
Protein domain number	1867	–0.204	10^{-6}	1074	–0.151	10^{-6}

Table S5. Spearman correlations of average expression level with various characters of housekeeping genes (AD cutoff = 300)

	Homo sapiens			*Mus musculus*		
	n	*R*	*r*	*n*	*r*	*r*
Average intron length	1564	–0.249	10^{-6}	676	–0.145	2×10^{-4}
First intron length	1564	–0.186	10^{-6}	676	–0.137	3×10^{-4}
Intron number	1564	–0.239	10^{-6}	676	–0.212	10^{-6}
Protein length	1377	–0.243	10^{-6}	669	–0.294	10^{-6}
Protein domain number	1377	–0.133	6×10^{-4}	669	–0.181	2×10^{-6}

Table S6. Spearman correlations of average expression level with various characters of stringently defined human housekeeping genes

	Human housekeeping genes defined by								
	Hsiao et al. [10]			Eisenberg and Levanon [11]			Dorus et al. [12]		
	n	*r*	*P*	*n*	*r*	*P*	*n*	*r*	*P*
Average intron length	335	–0.399	10^{-6}	485	–0.169	2×10^{-4}	87	–0.489	2×10^{-6}
First intron length	335	–0.378	10^{-6}	485	–0.193	2×10^{-5}	87	–0.422	5×10^{-5}
Intron number	335	–0.322	10^{-6}	485	–0.313	10^{-6}	87	–0.489	2×10^{-6}
Protein length	325	–0.395	10^{-6}	488	–0.394	10^{-6}	84	–0.529	10^{-6}
Protein domain number	325	–0.252	4×10^{-6}	488	–0.225	10^{-6}	84	–0.298	0.006

Table S7. *Arabidopsis* gene examples illustrating the difference between expression breadth and tissue specificity index τ

Locus[a]	Inflorescence	Flower	Cauline leaf	Silique	Seedling	Root	Expression breadth	Index τ[b]
AT2G32930	258	301	238	240	249	267	6	0.03
AT3G56310	218	868	261	657	2705	83994	6	0.43
AT2G37940	74	0	77	77	0	0	3	0.60
AT3G15400	13587	0	43	0	254	0	3	0.80

[a]The expression levels were taken from http://plantgenomics.biology.yale.edu/ and manipulated as described in Supplementary methods.
[b]Index τ is a measure of tissue specificity of gene expression proposed by [8], the expression values were first $\log_2$-transformed.

Table S8. Comparison of intermediately expressed genes (IM) and housekeeping genes (HK) by Wilcoxon signed ranks test

	Homo sapiens				*Mus musculus*			
	Number of gene pairs	*P*	IM > HK (%)[a]	IM < HK (%)[b]	Number of gene pairs	*P*	IM > HK (%)[a]	IM < HK (%)[b]
Average intron length	3109	0.295	47.3	52.7	1175	0.955	48.3	51.7
First intron length	3109	2×10^{-5}	46.2	53.8	1175	0.480	48.9	51.2
Intron number	3109	10^{-43}	36.6	59.5	1175	2×10^{-5}	43.8	51.0
Protein length	2443	3×10^{-13}	43.4	56.4	1143	7×10^{-6}	44.9	54.9
Protein domain number	2443	0.034	35.8	41.0	1143	0.075	35.1	38.8

[a]Percentage of gene pairs in which the intermediately expressed gene are larger than the housekeeping gene.
[b]Percentage of gene pairs in which the housekeeping gene are larger than the intermediately expressed gene.

Table S9. Spearman correlations of average expression level with various characters of housekeeping genes (SAGE & MPSS data)

	Homo sapiens			*Mus musculus*			*Arabidopsis thaliana*		
	n	*r*	*P*	*n*	*r*	*P*	*n*	*r*	*P*
Average intron length	701	–0.248	10^{-6}	900	–0.133	6×10^{-5}	3771	0.112	10^{-6}
First intron length	701	–0.235	10^{-6}	900	–0.059	0.075	3771	0.054	0.001
Intron number	701	–0.239	10^{-6}	900	–0.158	2×10^{-6}	3771	–0.276	10^{-6}
Protein length	657	–0.317	10^{-6}	878	–0.162	10^{-6}	2937	–0.481	10^{-6}
Protein domain number	657	–0.199	10^{-6}	878	–0.132	9×10^{-5}	2937	–0.265	10^{-6}

Table S10. Spearman correlations of tissue specificity with gene and protein characters (SAGE & MPSS data)

	Homo sapiens			*Mus musculus*			*Arabidopsis thaliana*		
	n	*r*	*P*	*n*	*r*	*P*	*n*	*r*	*P*
Average intron length	13989	0.030	4×10^{-4}	13694	-2×10^{-4}	0.980	17592	–0.102	10^{-6}
First intron length	13982	–0.032	10^{-4}	13694	–0.053	10^{-6}	17592	–0.119	10^{-6}
Intron number	13989	–0.128	10^{-6}	13694	–0.168	10^{-6}	17592	–0.202	10^{-6}
Protein length	12513	0.009	0.298	13509	–0.017	0.045	14622	–0.081	10^{-6}
Protein domain number	12513	–0.003	0.712	13509	–0.010	0.238	14622	–0.053	10^{-6}

Supplementary references

[1] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, J.B. Hogenesch, Large-scale analysis of the human and mouse transcriptomes, Proc. Natl. Acad. Sci. USA 99 (2002) 4465-4470.

[2] V.E. Velculescu, L. Zhang, B. Vogelstein, K.W. Kinzler, Serial analysis of gene-expression, Science 270 (1995) 484-487.

[3] A.E. Lash, C.M. Tolstoshev, L. Wagner, G.D. Schuler, R.L. Strausberg, G.J. Riggins, S.F. Altschul, SAGEmap: a public gene expression resource, Genome Res. 10 (2000) 1051-1060.

[4] M.J. Lercher, J.V. Chamary, L.D. Hurst, Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile, Genome Res. 14 (2004) 1002-1013.

[5] J. Chen, M. Rattray, Analysis of tag-position bias in MPSS technology, BMC Genomics 7 (2006) 77.

[6] X.-Y. Ren, O. Vorst, M.W.E.J. Fiers, W.J. Stiekema, J.-P. Nap, In plants, highly expressed genes are the least compact, Trends Genet. 22 (2006) 528-532.

[7] A.E. Vinogradov, 'Genome design' model and multicellular complexity: golden middle, Nucleic Acids Res. 34 (2006) 5906-5914.

[8] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, O. Shmueli, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, Bioinformatics 21 (2005) 650-659.

[9] A.E. Vinogradov, Compactness of human housekeeping genes: selection for economy or genomic design?, Trends Genet. 20 (2004) 248-253.

[10] L.L. Hsiao, F. Dangond, T. Yoshida, R. Hong, R.V. Jensen, J. Misra, W. Dillon, K.F. Lee, K.E. Clark, P. Haverty, Z.P. Weng, G.L. Mutter, M.P. Frosch, M.E. MacDonald, E.L. Milford, C.P. Crum, R. Bueno, R.E. Pratt, M. Mahadevappa, J.A. Warrington, G. Stephanopoulos, S.R. Gullans, A compendium of gene expression in normal human tissues, Physiol. Genomics 7 (2001) 97-104.

[11] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, Trends Genet. 19 (2003) 362-365.

[12] S. Dorus, E.J. Vallender, P.D. Evans, J.R. Anderson, S.L. Gilbert, M. Mahowald, G.J. Wyckoff, C.M. Malcom, B.T. Lahn, Accelerated evolution of nervous system genes in the origin of *Homo sapiens*, Cell 119 (2004) 1027-1040.

Supplementary figures

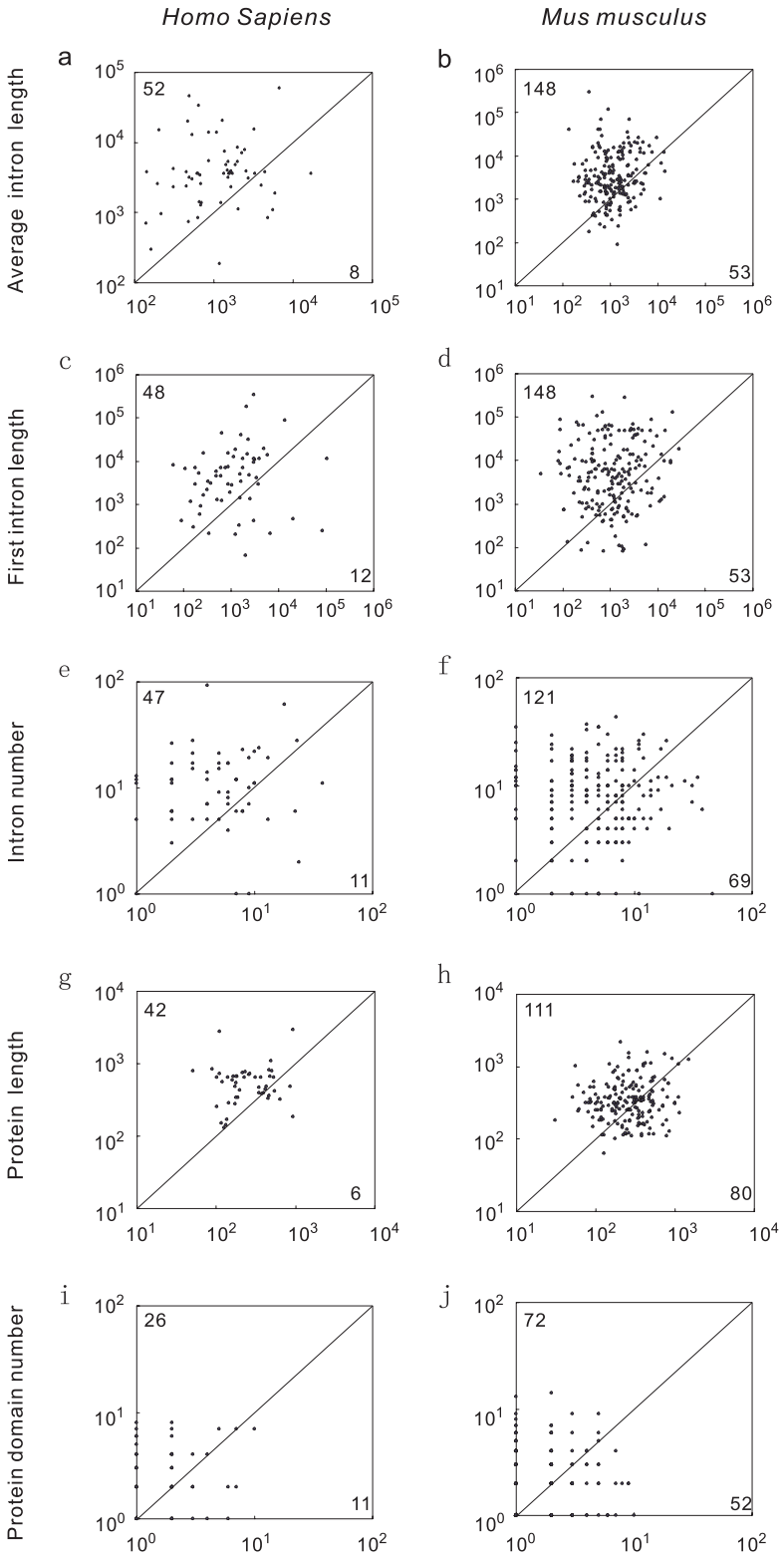

Fig. S1. Comparing housekeeping genes and narrowly expressed genes with similar average expression levels (AD cutoff = 100). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 60, $P < 10^{-6}$; b, 201, $P < 10^{-16}$; c, 60, $P < 10^{-5}$; d, 201, $P < 10^{-15}$; e, 60, $P = 10^{-5}$; f, 201, $P < 10^{-6}$; g, 48, $P < 10^{-5}$; h, 191, $P = 0.002$; i, 48, $P = 0.014$; j, 191, $P = 0.010$.

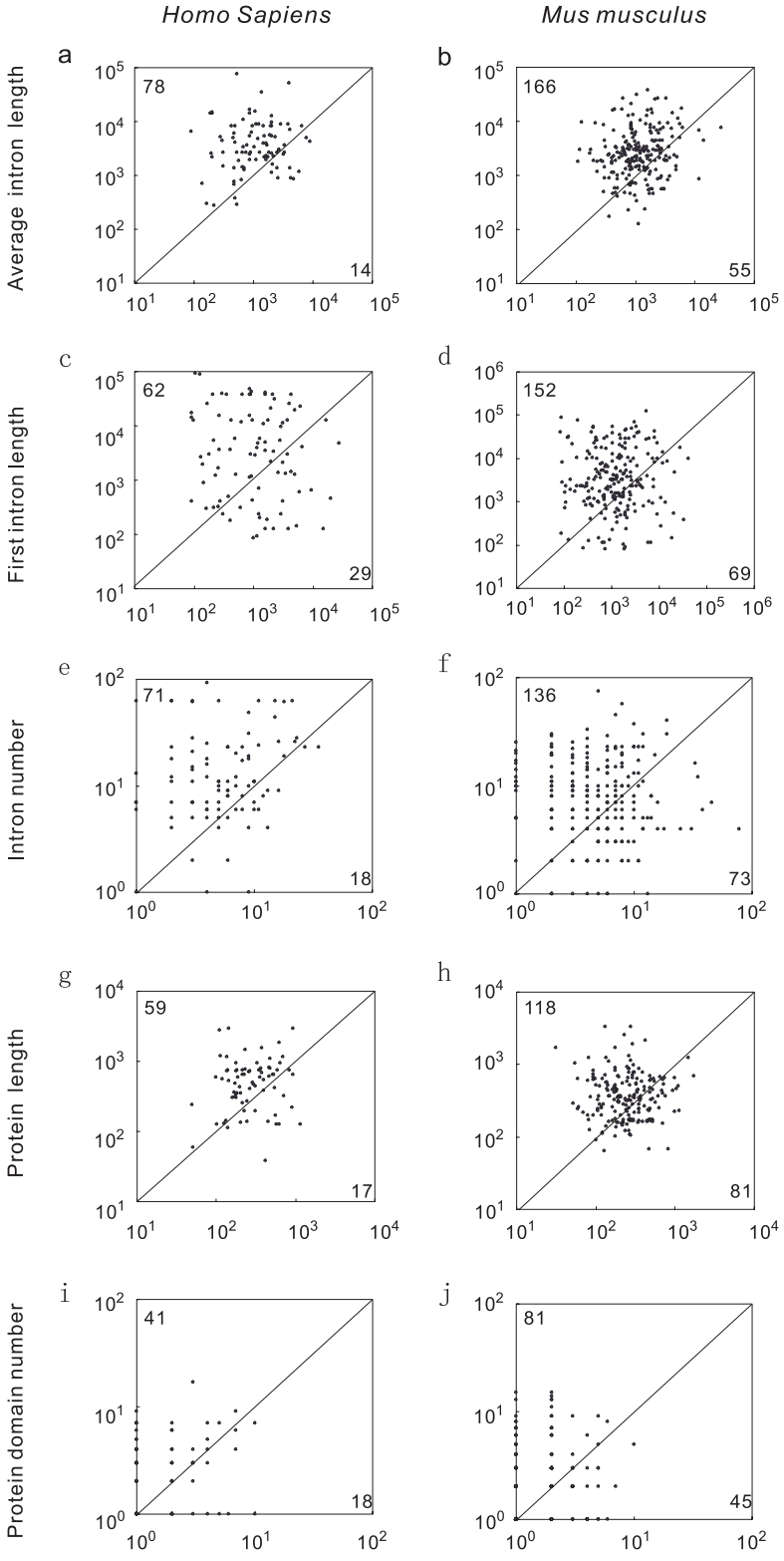

Fig. S2. Comparing housekeeping genes and narrowly expressed genes with similar average expression levels (AD cutoff = 150). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 92, $P < 10^{-9}$; b, 221, $P < 10^{-15}$; c, 92, $P = 10^{-5}$; d, 221, $P < 10^{-11}$; e, 92, $P < 10^{-8}$; f, 221, $P < 10^{-8}$; g, 76, $P = 10^{-6}$; h, 199, $P < 10^{-4}$; i, 76, $P < 10^{-3}$; j, 199, $P < 10^{-4}$.

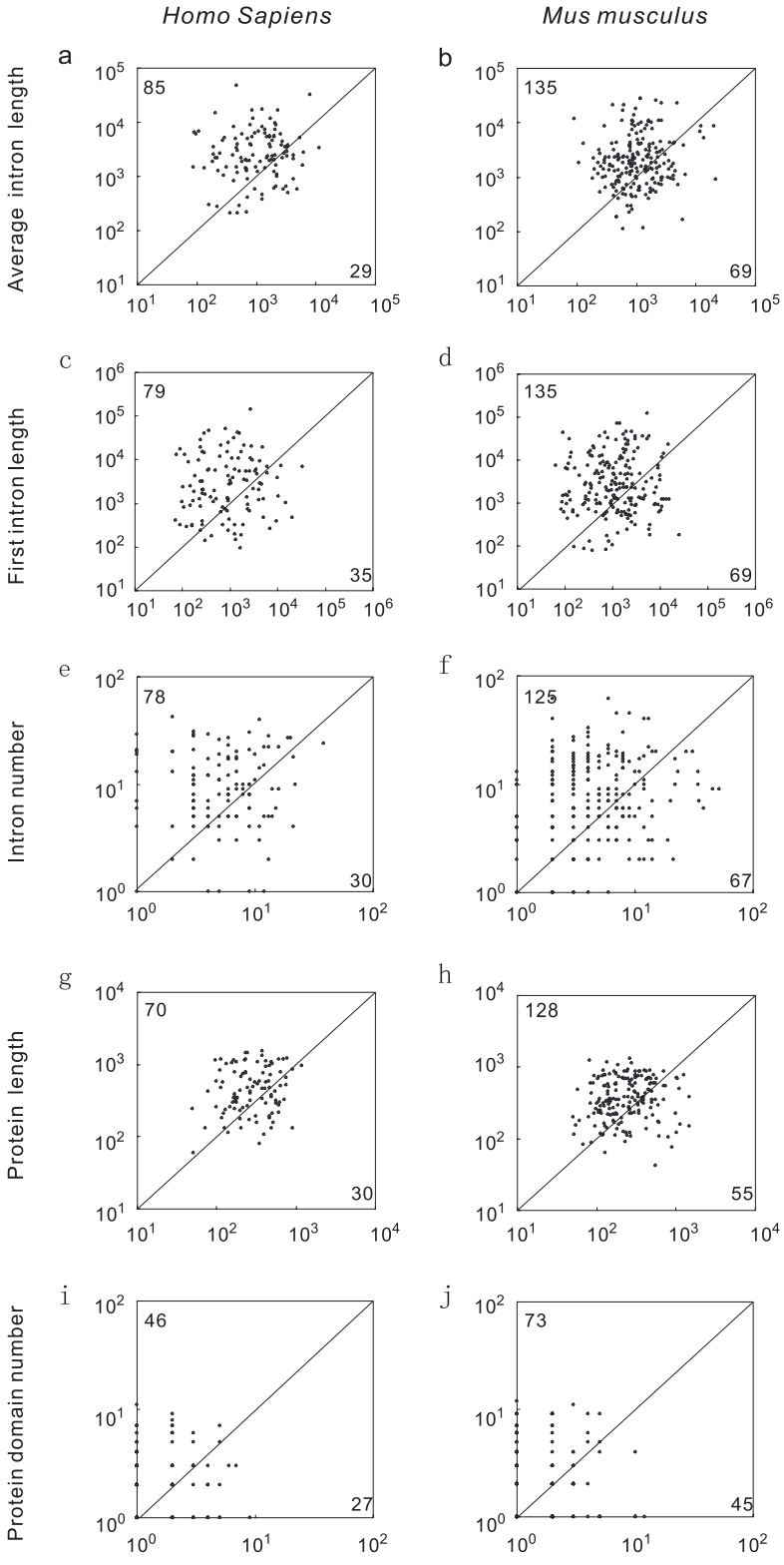

Fig. S3. Comparing housekeeping genes and narrowly expressed genes with similar average expression levels (AD cutoff = 250). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 114, $P = 10^{-8}$; b, 204, $P < 10^{-7}$; c, 114, $P < 10^{-5}$; d, 204, $P < 10^{-7}$; e, 114, $P < 10^{-6}$; f, 204, $P = 10^{-7}$; g, 100, $P = 10^{-5}$; h, 184, $P < 10^{-7}$; i, 100, $P = 0.012$; j, 184, $P = 10^{-4}$.

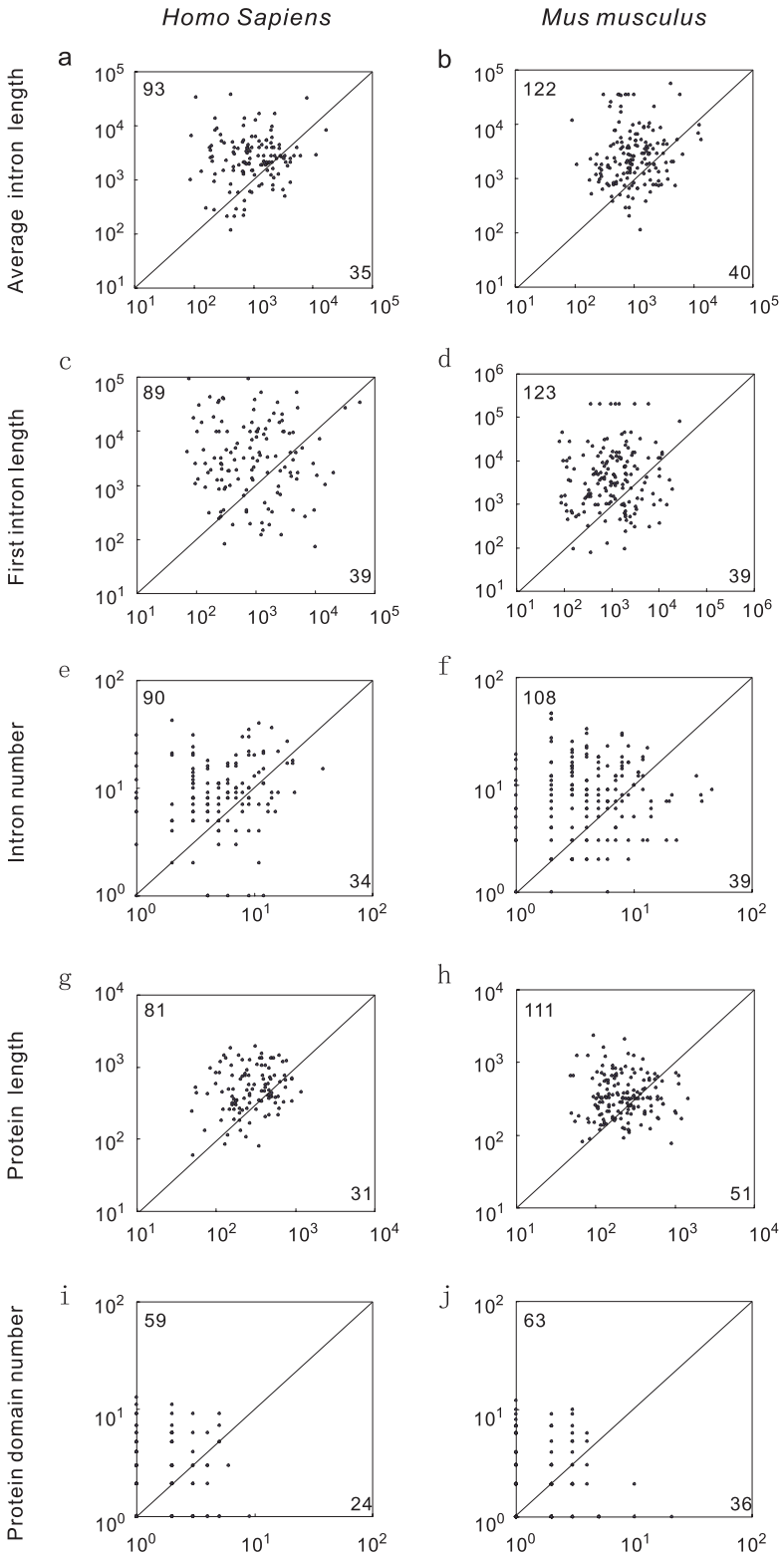

Fig. S4. Comparing housekeeping genes and narrowly expressed genes with similar average expression levels (AD cutoff = 300). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 128, $P < 10^{-8}$; b, 162, $P < 10^{-11}$; c, 128, $P < 10^{-6}$; d, 162, $P < 10^{-11}$; e, 128, $P < 10^{-7}$; f, 162, $P < 10^{-8}$; g, 112, $P < 10^{-7}$; h, 164, $P = 10^{-5}$; i, 112, $P = 10^{-5}$; j, 164, $P = 10^{-4}$.

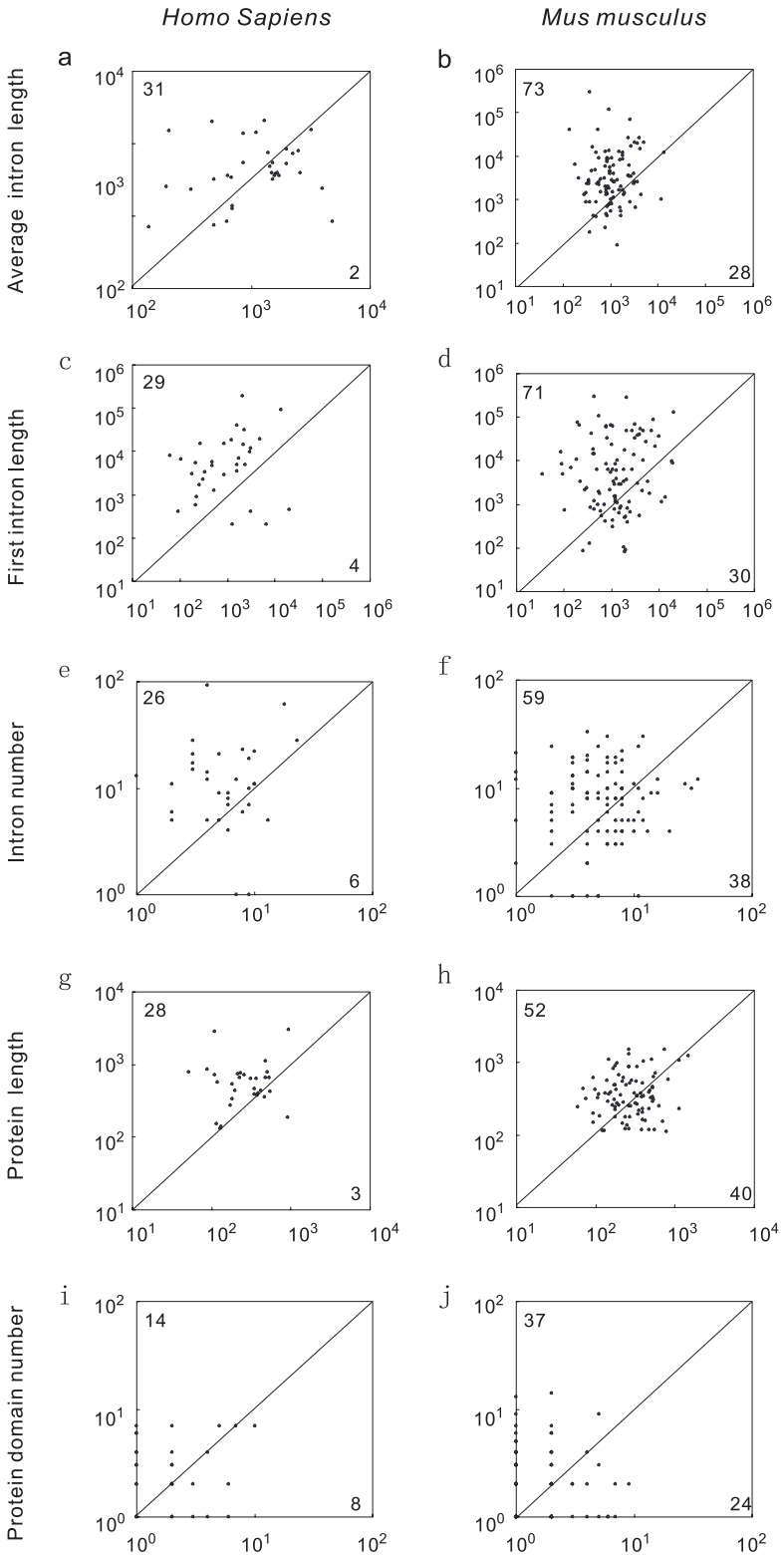

Fig. S5. Comparing housekeeping genes and narrowly expressed somatic genes with similar average expression levels (AD cutoff = 100). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed somatic counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed somatic genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 33, $P < 10^{-5}$; b, 101, $P < 10^{-9}$; c, 33, $P = 10^{-4}$; d, 101, $P < 10^{-7}$; e, 33, $P < 10^{-3}$; f, 101, $P = 0.002$; g, 31, $P = 10^{-4}$; h, 92, $P = 0.019$; i, 31, $P = 0.20$; j, 92, $P = 0.025$.

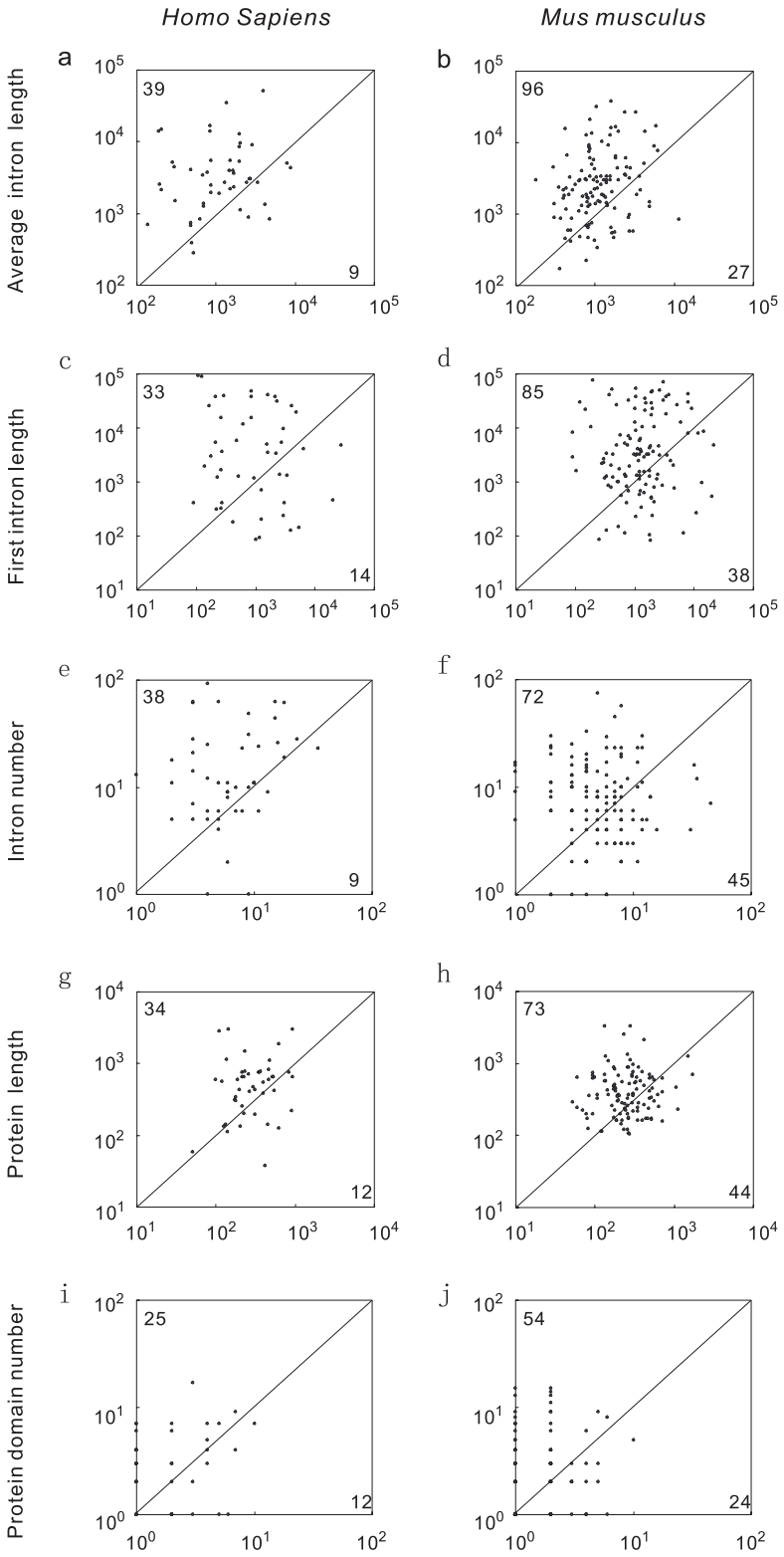

Fig. S6. Comparing housekeeping genes and narrowly expressed somatic genes with similar average expression levels (AD cutoff = 150). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed somatic counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed somatic genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 48, $P < 10^{-4}$; b, 123, $P < 10^{-10}$; c, 48, $P = 0.002$; d, 123, $P < 10^{-7}$; e, 48, $P < 10^{-4}$; f, 123, $P = 10^{-4}$; g, 46, $P < 10^{-3}$; h, 117, $P = 10^{-4}$; i, 46, $P = 0.012$; j, 117, $P < 10^{-4}$.

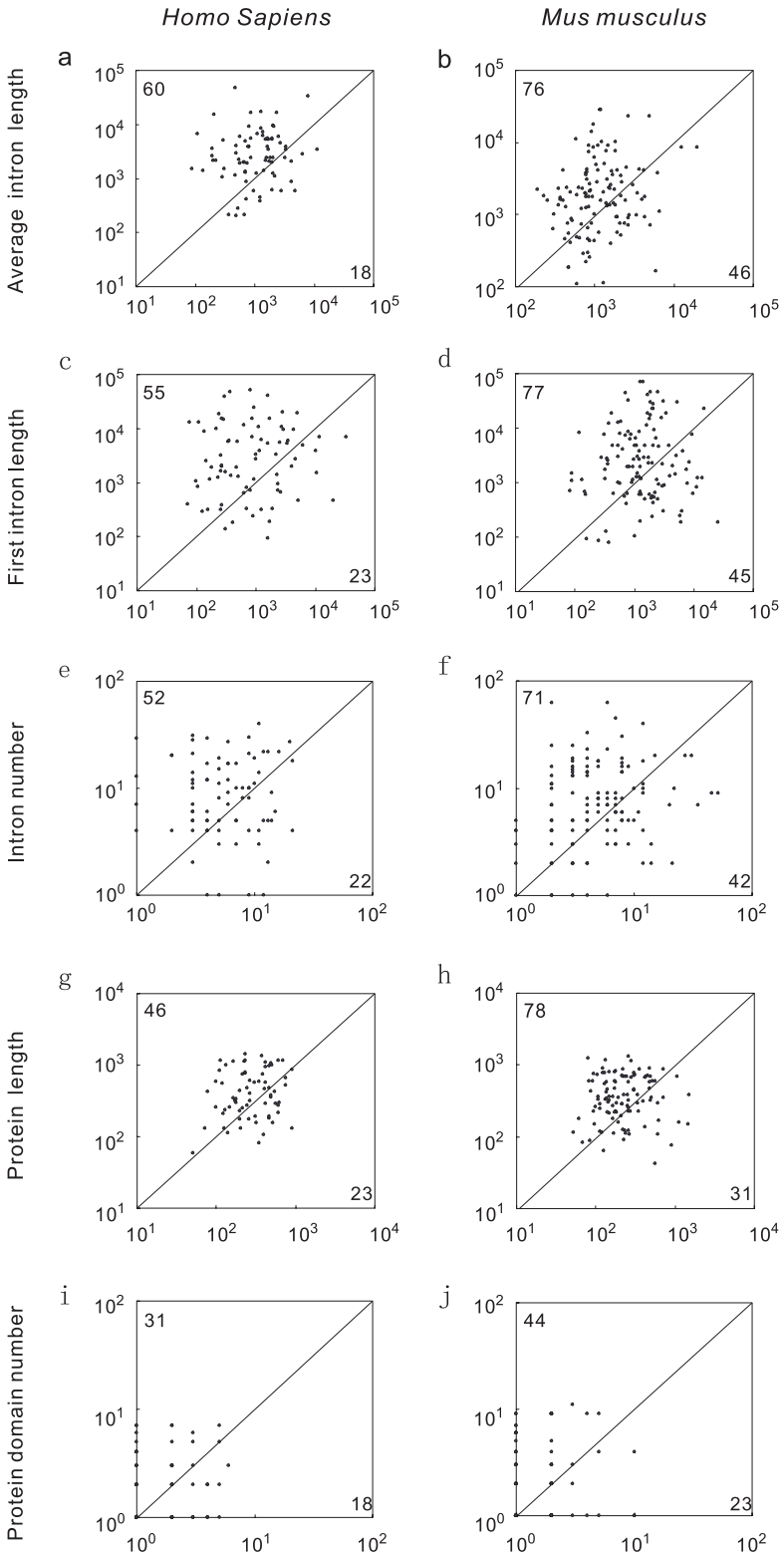

Fig. S7. Comparing housekeeping genes and narrowly expressed somatic genes with similar average expression levels (AD cutoff = 250). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed somatic counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed somatic genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 78, $P < 10^{-6}$; b, 122, $P < 10^{-3}$; c, 78, $P < 10^{-4}$; d, 122, $P < 10^{-3}$; e, 78, $P < 10^{-3}$; f, 122, $P = 10^{-3}$; g, 69, $P = 10^{-3}$; h, 109, $P < 10^{-5}$; i, 69, $P = 0.069$; j, 109, $P = 10^{-3}$.

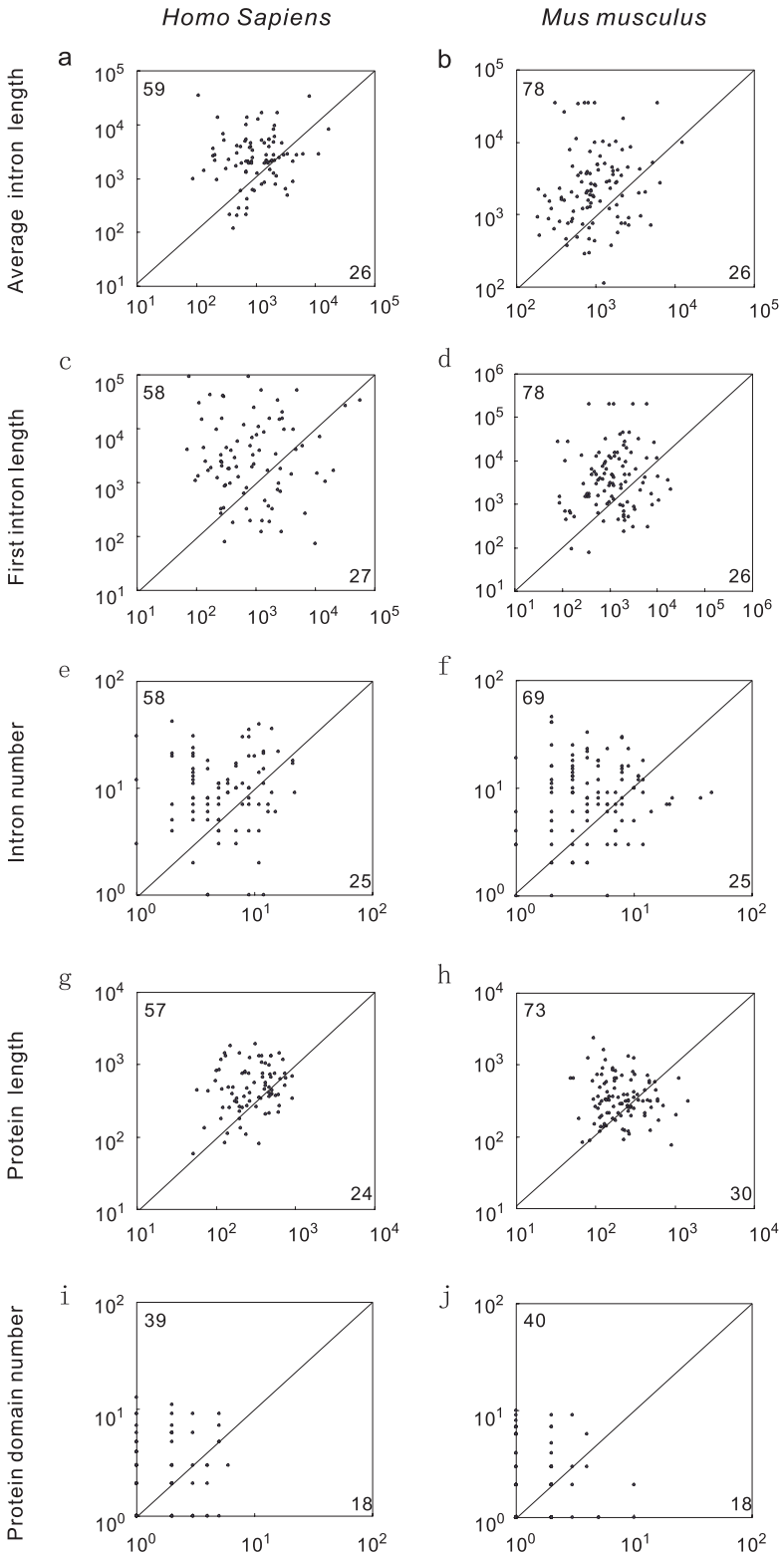

Fig. S8. Comparing housekeeping genes and narrowly expressed somatic genes with similar average expression levels (AD cutoff = 300). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed somatic counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed somatic genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 85, $P < 10^{-5}$; b, 104, $P < 10^{-7}$; c, 85, $P = 10^{-4}$; d, 104, $P = 10^{-7}$; e, 85, $P = 10^{-5}$; f, 104, $P < 10^{-6}$; g, 81, $P < 10^{-5}$; h, 103, $P < 10^{-4}$; i, 81, $P < 10^{-3}$; j, 103, $P < 10^{-3}$.

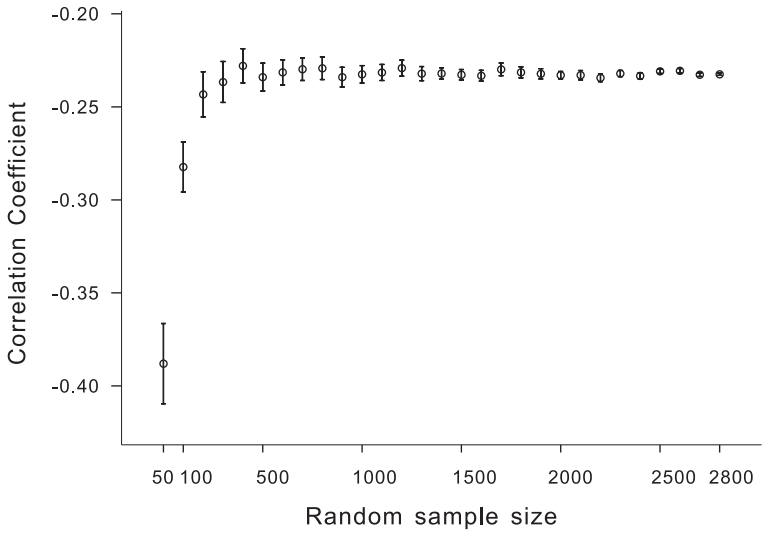

Fig. S9. Dependence of the correlation coefficients (between average expression level and average intron length) on the sample size of human housekeeping genes. From 2879 housekeeping genes selected only by ubiquitous expression, we randomly selected a series of samples to perform Spearman correlation analysis. Random sample size range from 50, 100, 200, 300 to 2800, and for each sample size, the dataset was selected 100 times. For sample size of 50, 100 and 200, we discarded 67, 29 and 10 samples with *P* value > 0.05. Error bars represent 95% confidence interval of the group mean.

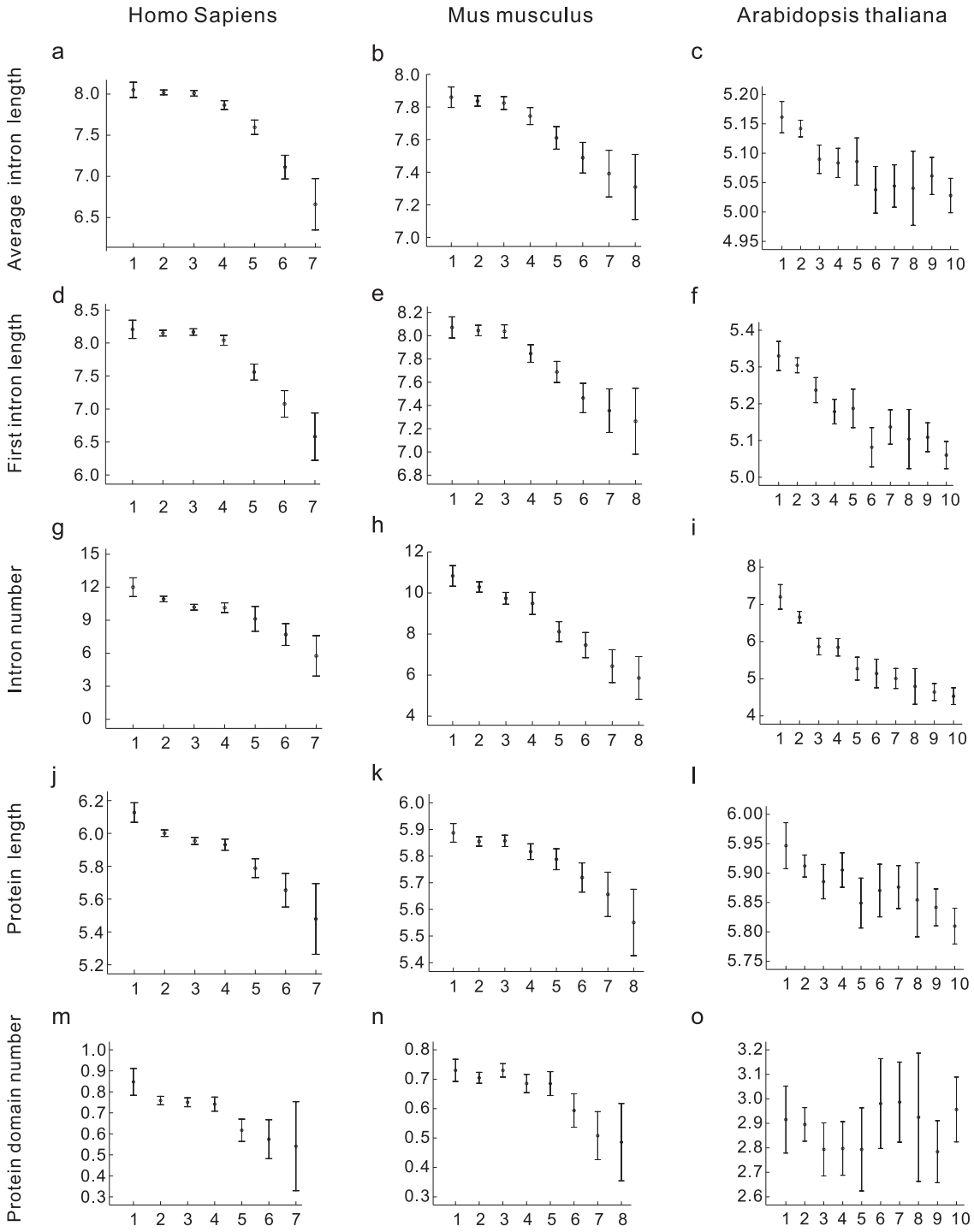

Fig. S10. Relationship between gene characters and tissue specificity index. Genes were grouped according to the value of tissue specificity index τ and labeled on the X axis, while bars show the mean of gene character (log-transformed) with 95% confidence interval. Spearman test analyses of these data are present in Table 2.

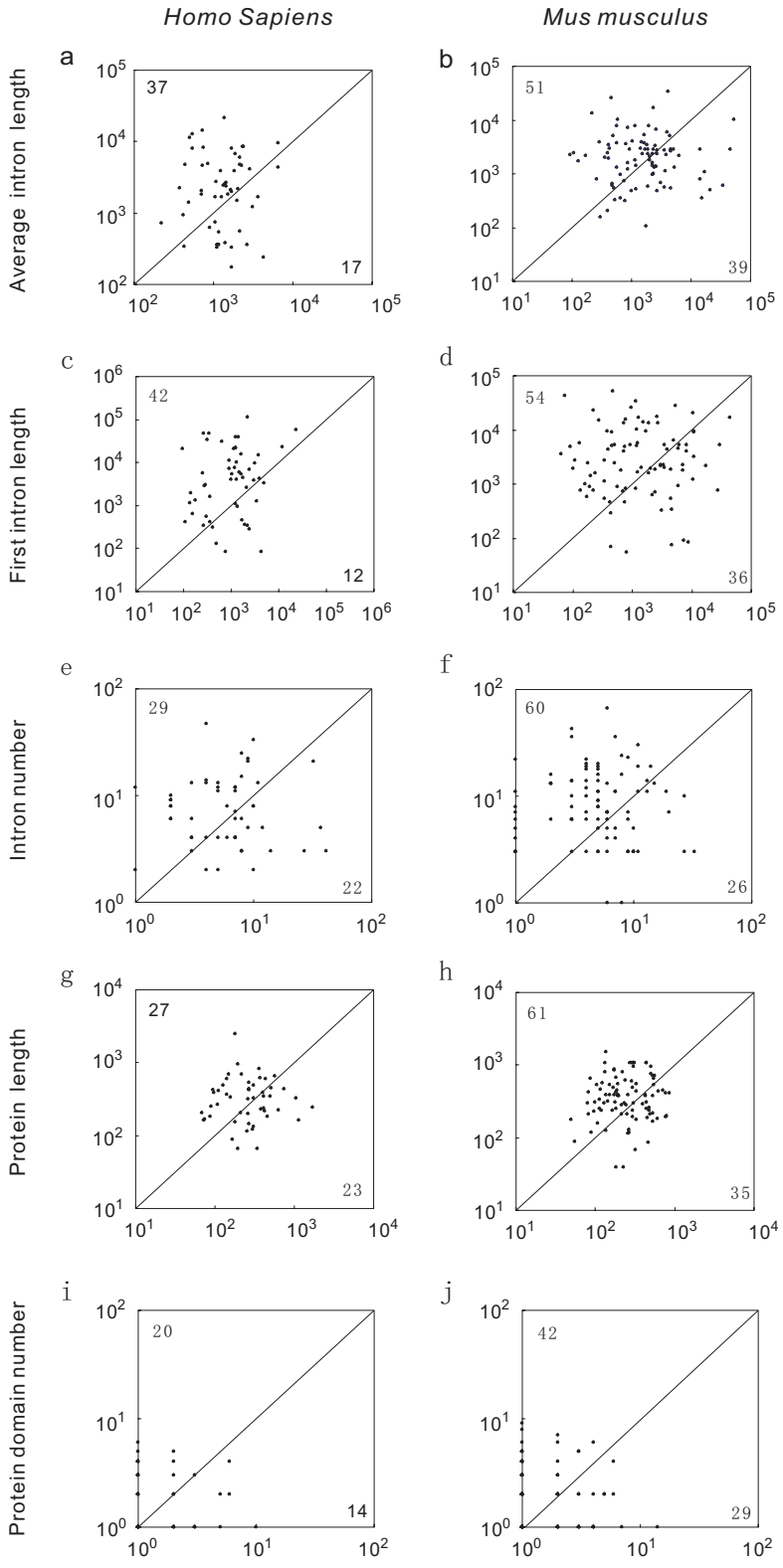

Fig. S11. Comparing housekeeping genes and narrowly expressed genes with similar average expression levels (SAGE data). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 54, $P = 10^{-3}$; b, 90, $P = 0.310$; c, 54, $P = 10^{-6}$; d, 90, $P = 0.030$; e, 54, $P = 0.130$; f, 90, $P < 10^{-4}$; g, 50, $P = 0.265$; h, 96, $P = 10^{-3}$; i, 50, $P = 0.297$; j, 96, $P = 0.194$.

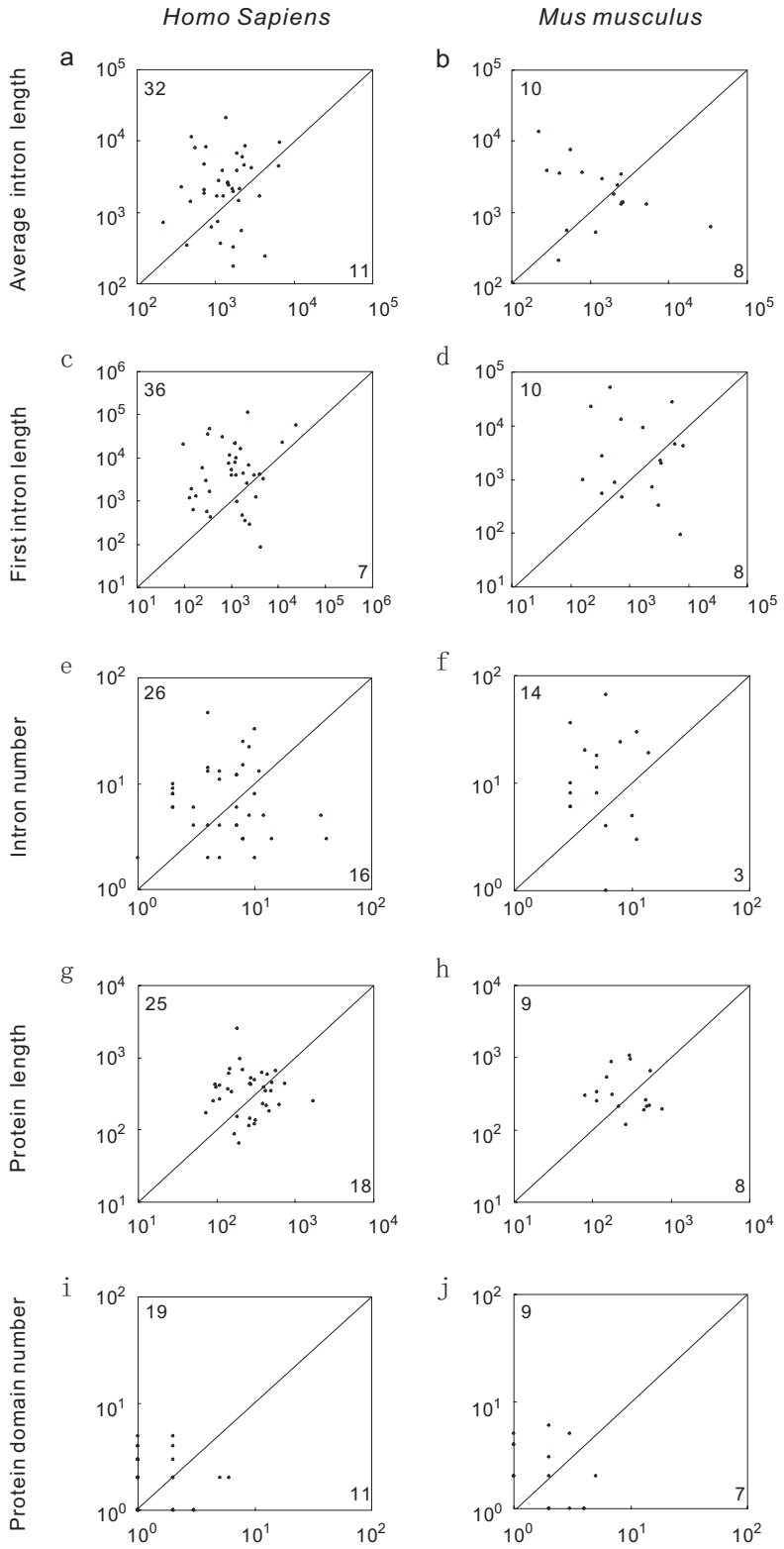

Fig. S12. Comparing housekeeping genes and narrowly expressed somatic genes with similar average expression levels (SAGE data). The Y axis represents housekeeping genes, while the X axis shows their narrowly expressed somatic counterparts. The numbers of dots above (marked at the top left corner) and below (marked at the bottom right corner) the right angle bisector intuitively illustrate the comparison between housekeeping genes and narrowly expressed somatic genes. Meanwhile, we performed Wilcoxon signed ranks test to determine the significance of the differences. The number of gene pairs and the significant levels are: a, 43, $P = 10^{-3}$; b, 18, $P = 0.59$; c, 43, $P < 10^{-5}$; d, 18, $P = 0.35$; e, 43, $P = 0.055$; f, 18, $P = 0.005$; g, 43, $P = 0.046$; h, 17, $P = 0.69$; i, 43, $P = 0.10$; j, 17, $P = 0.64$.