

# Uneven size distribution of mammalian genes in the number of tissues expressed and in the number of co-expressed genes

Song Liu, Chi Zhang and Yaoqi Zhou\*

Department of Physiology and Biophysics, Howard Hughes Medical Institute Center for Single Molecule Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214, USA

Received December 6, 2005; Revised and Accepted March 5, 2006

---

**Tissue specificity, the traditional predictor of gene function, has recently been used to interpret the selective pressure associated with gene architecture. In this work, we examine gene structures and their relation to the number of tissues expressed and to the number of co-expressed genes, using a recent atlas of microarray-based mouse gene expression in 55 normal tissues. We define tissue specificity and expression-pattern specificity according to the number of tissues expressed and the number of co-expressed genes, respectively. We find that, consistent with previous findings, tissue non-specific (housekeeping) genes are short in all gene regions (coding regions, intron, 5' and 3' untranslated regions). However, in contrast to previous suggestion that tissue-specific genes are long, the genes that are the most tissue-specific (expressed only in one tissue) are also short. We further show that both expression-pattern-specific and non-specific genes are long in coding and non-coding regions. The origins for short tissue-specific genes and long expression-pattern-specific genes are not clear. Genes with highly non-specific expression patterns (i.e. genes with a large number of co-expressed genes) are composed of genes that spread all tissues but are overwhelmingly enriched in the central nervous system (e.g. brain). Thus, the large sizes of these genes are possibly related to the functional complexity and/or accelerated evolutions of the central nervous system.**

---

## INTRODUCTION

There is a great interest in interpreting the relationship between genome size and mammalian complexity through the view of functional genomics (1). Partitioning genes into different expression categories on the basis of tissue-profiling data shows that housekeeping genes tend to have more compact gene structures than those of tissue-specific ones (2–7). The observation leads to several hypotheses. The ‘selection for economy’ view interprets the compactness of housekeeping genes as a way to reduce the energetic cost during transcription and translation. The ‘selection for genomic design’ view, in contrast, suggests that the long tissue-specific genes result from increasing functional and regulatory complexities (1–5). The two views focused on different genes and thus may occur simultaneously. More recently, Farh *et al.* (8) proposed that short untranslated regions (UTRs) allow housekeeping genes to avoid detrimental microRNA pairing.

These interpretations are based on tissue specificity, which classifies genes according to the number of tissues where a gene is expressed. However, it was recently found that such tissue-specific restriction of expression pattern is a poor predictor of mammal gene function (9). Instead, quantitative analysis of transcriptional co-expression is required for a more accurate assessment of mammalian gene function (i.e. co-expression leads to co-function/regulation) (9–11). This result spurs our interest in the re-examination of variation of gene sizes under the framework of transcriptional co-expression.

Co-expression networks derived from transcriptional correlation analysis have been demonstrated to follow a scale-free network topology, which is characterized by an inverse power-law relation between the population of nodes (genes) and their degrees (the number of co-expressed genes) (12–14). Protein interaction networks and transcriptional regulation networks in lower organisms (15–20) are also found to be scale-free. In a scale-free network, there exist islands with few or no

---

\*To whom correspondence should be addressed. Tel: +1 716 8292985; Fax: +1 7168292344; Email: yqzhou@buffalo.edu

**Table 1.** Average and median lengths (bp) of genes on the basis of different classification schemes

Regions	Tissue Specificity			Expression-Pattern Specificity		
	TS <sup>a</sup> (650) <sup>c</sup>	HK <sup>a</sup> (652) <sup>c</sup>	Other <sup>a</sup> (6092) <sup>c</sup>	EPN <sup>b</sup> (697) <sup>c</sup>	EPS <sup>b</sup> (695) <sup>c</sup>	Other <sup>b</sup> (6002) <sup>c</sup>
Intron (median) <sup>c</sup>	52 893 ± 5035 <sup>d</sup> (13 645)	26 248 ± 2031 (13 597)	48 928 ± 1356 (18 638)	86 392 ± 6920 (23 540)	51 284 ± 3945 (20 707)	42 271 ± 1167 (16 669)
CDS (median)	1443 ± 45 (1108)	1287 ± 44 (979)	1662 ± 18 (1318)	1715 ± 55 (1367)	1651 ± 56 (1292)	1592 ± 17 (1261)
5' UTR (median)	183 ± 11 (98)	135 ± 6 (93)	179 ± 3 (114)	196 ± 8 (133)	202 ± 10 (123)	170 ± 3 (107)
3' UTR (median)	704 ± 31 (415)	754 ± 28 (560)	918 ± 11 (656)	1056 ± 37 (798)	915 ± 34 (646)	861 ± 11 (600)

<sup>a</sup>Classified by the number of tissues expressed. There are tissue-specific genes (the most narrowly expressed, in only one tissue), housekeeping genes (the most widely expressed, in all 55 tissues) and other genes.

<sup>b</sup>Classified by the number of co-expressed genes. There are EPN genes (with ≥360 co-expressed genes), EPS genes (with ≤3 co-expressed genes) and the rest (background).

<sup>c</sup>The number of genes.

<sup>d</sup>The number in each cell is the average value ±SEM of gene structure parameters.

<sup>e</sup>The median value of gene structure parameters.

connected nodes and hubs that connect to many other nodes. A hub gene in a transcriptional regulation network is a gene that regulates many other genes, whereas a hub gene in a co-expression network is a gene with a non-specific expression pattern that correlates with the patterns of many other genes. There is no simple one-to-one correspondence between hub or island genes from different networks, although protein interaction network, transcriptional regulation network and gene co-expression network correlate with each other (21–23). Thus, we use the terms expression-pattern-non-specific (EPN) and expression-pattern-specific (EPS) genes to represent hub and island genes, respectively, in a co-expression network.

In this work, we analyze a recent landscape of mouse transcription profiles in 55 normal tissues (9) alongside their gene structures for 7394 RefSeq-defined genes. We hope to examine the size variation of genes according to both tissue specificity and expression-pattern specificity. The large data set from mouse tissue-level transcriptional profiling allows us to perform a detailed analysis.

## RESULTS AND DISCUSSION

The 7394 mouse RefSeq genes have 650 genes expressed in only one tissue and 652 genes expressed in all 55 tissues. They can represent the most narrowly (i.e. tissue-specific) and ubiquitously expressed (i.e. housekeeping) genes, respectively. The remaining 6092 genes are considered as background genes for comparison. The average and median lengths of tissue-specific and housekeeping genes are shown in Table 1 and Figure 1. The median length is more suitable than the average length to reflect the size distribution because the latter is often biased toward a few large-size genes. Overall trends, however, are essentially the same. We will discuss subsequently the results on the basis of median values, unless indicated otherwise.

Table 1 and Figure 1 show that the ubiquitously expressed genes are significantly shorter than background genes in all regions of gene structures. This is consistent with the previous finding in human that housekeeping genes are compact (4). The size difference ranges from 37% (in median size of background genes) in intron sequences to 17% in 3' UTRs.

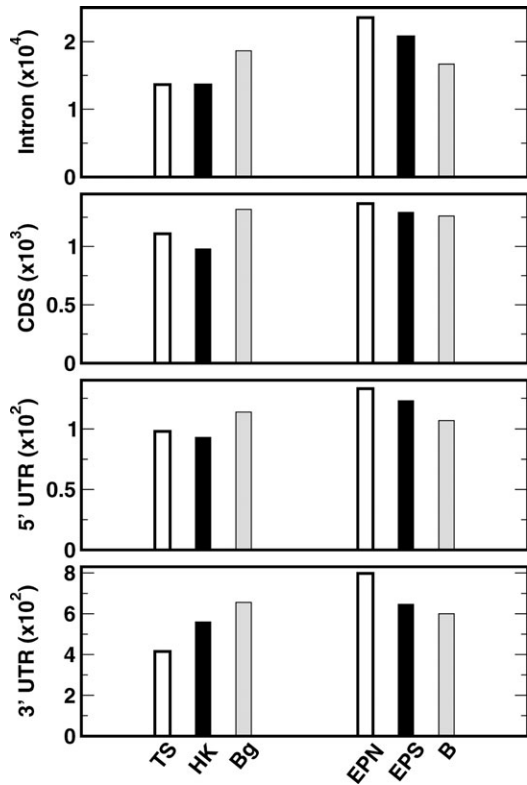
All differences are significant (Wilcoxon two-sample test,  $P < 10^{-10}$ – $P < 10^{-3}$ ).

The most narrowly expressed genes, however, are also significantly shorter in all regions than background genes. (The exception is that the average sizes of most tissue-specific genes in intron and 5' UTRs are slightly longer than that of background genes. This is largely due to a few long genes, as demonstrated by median values.) For example, the median length of 3' UTRs for the most narrowly expressed genes is 241 bp shorter ( $P < 10^{-14}$ ) and that of protein-translated regions is 42 amino acids shorter ( $P < 10^{-5}$ ). Furthermore, the median value of tissue-specific genes in 3' UTRs (415 bp) is even 35% shorter than that of housekeeping genes (560 bp), with the  $P$ -value less than  $10^{-3}$ .

On the basis of established co-expression network (see Materials and Methods), one can classify genes according to expression-pattern specificity (i.e. the number of co-expressed genes). Of the 7394 genes, 695 genes have less than four co-expressed genes and 697 genes have 360 or more co-expressed genes. They are employed to represent EPS and EPN genes, respectively. The remaining 6002 genes are considered as background genes. The cutoff values are chosen so that the numbers of obtained EPS genes, EPN genes and background genes are similar to the corresponding numbers of housekeeping, tissue-specific and background genes, respectively.

The sizes of genes classified according to expression-pattern specificity are also shown in Table 1 and Figure 1. EPN genes are longer in every aspect of gene structures than background genes on the basis of either average or median ( $P < 10^{-10}$ – $P < 10^{-2}$ ). For example, the median size of EPN genes is longer than background genes, from 25 to 41% for non-coding regions and 8.5% for coding regions. The increase in length is even more significant if we use genes with the higher number of co-expressed genes to represent EPN genes (data not shown). Interestingly, EPS genes are also longer than background genes, although the difference is smaller than that between EPN genes and background genes. The increment is significant in non-coding regions (8–24%) and weak in coding regions.

The results described are based on a simple three-class classification of genes. Figure 2 provides a more detailed

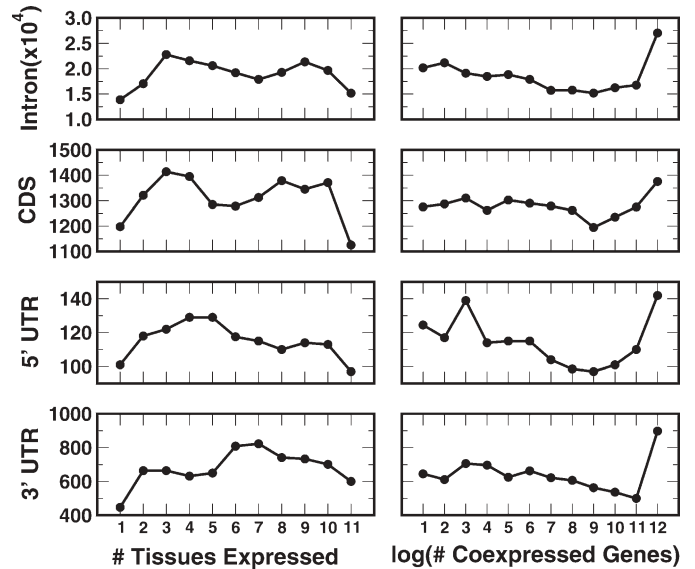


**Figure 1.** Comparison of median lengths (bp) of gene structures in genes partitioned by different criteria. From top to bottom, results are for intron, CDS, 5' UTR and 3' UTR lengths. In each figure, TS, HK and Bg (the first three bars on the left) denote the 650 tissue-specific genes expressed in only one tissue, the 652 housekeeping genes expressed in all 55 tissues and the rest of 6092 background genes. EPN, EPS and B (the next three bars) denote the 697 EPN genes with 360 or more co-expressed genes, the 695 EPS genes with three or less co-expressed genes and the rest of 6002 background genes, respectively. These figures show that both the most tissue-specific genes and housekeeping genes are compact when compared with background genes. Meanwhile, both EPS and EPN genes are longer than background genes.

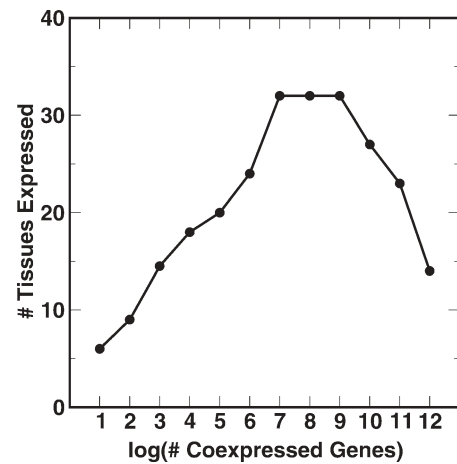
picture on how the sizes of genes vary as a function of the number of tissues expressed or as a function of the number of co-expressed genes. The curves for the size versus the number of tissues expressed can be approximately described as a bimodal distribution with three local minima at the number of tissues expressed around 1, 55 and an intermediate value, respectively. The curves for size versus the logarithm of the number of co-expressed genes have the global maximum at the highest number of co-expressed genes. As the number of co-expressed genes decreases, the size of genes experiences a sudden drop followed by a gradual increase. Thus, the overall trends for both tissue specificity and expression-pattern specificity agree with the results obtained from the simple three-class classification of genes.

Figure 3 displays the median number of tissues expressed for genes in the number of co-expressed genes (sorted in 12 bins). Clearly, the genes with the most or the least number of co-expressed genes are more tissue-specific than the genes with an intermediate number of co-expressed genes.

The result that the most tissue-specific genes are short is contrary to the suggestion that tissue-specific genes are long.



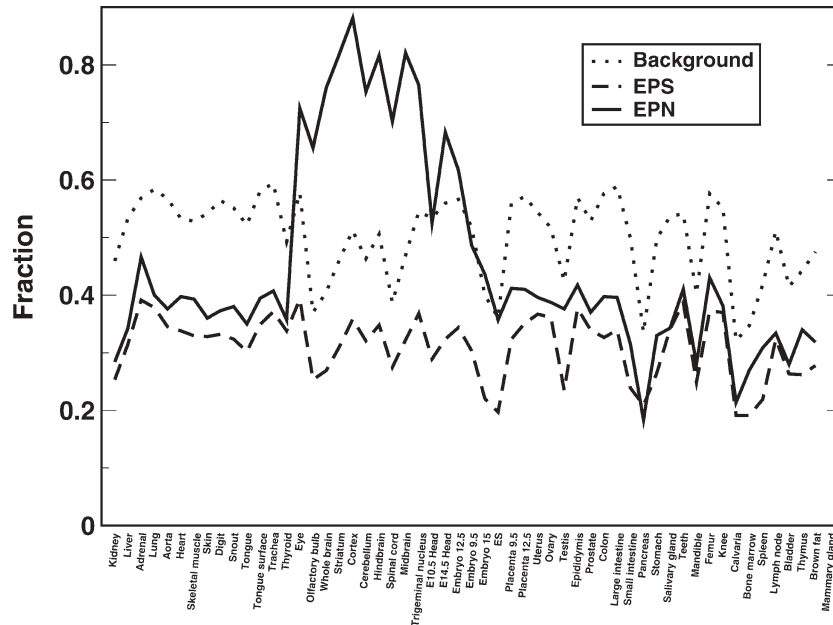
**Figure 2.** The median length (bp) of gene structures of genes as a function of their number of tissues expressed (left) and as a function of their number of co-expressed genes (right). From top to bottom, results are for intron, CDS, 5' UTR and 3' UTR, respectively. The lower bound of the number of expressed tissues for the 11 bins are 1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51, respectively. The bins for the number of co-expressed genes are sorted in the ascending order of the logarithm of the number of co-expressed genes (this bin step is 0.3). The lower bound of the number of co-expressed genes for the 12 bins are 0, 2, 4, 6, 10, 17, 29, 51, 88, 154, 269, 470, respectively.



**Figure 3.** The median number of tissues expressed as a function of the number of co-expressed genes, which is binned as in Figure 2.

The discrepancy is likely due to our focus on the most tissue-specific genes (genes only expressed in one tissue), whereas others employ a looser definition of tissue-specific genes. The latter may be long because there is a peak in gene sizes at an intermediate tissue specificity (e.g. 10–20 tissues).

Are EPN and/or EPS genes dominated by genes in specific tissues or with specific functions? Figure 4 shows the fractions of EPN, EPS and background genes in different tissues. Both background genes and EPS genes spread evenly in all tissues.



**Figure 4.** The fraction of 697 EPN genes, 695 EPS genes and 6002 background genes in each of 55 normal mouse tissues expressed. Note that the sum of the fractions exceed 100% of EPN, EPS or background genes because genes may be expressed in multiple tissues.

EPN genes, however, are composed of genes that spread in all tissues but are overwhelmingly enriched in the central nervous system (e.g. brain). Thus, EPN genes might perform functional roles or be involved in the biological processes in the nervous system. This suggests that the large sizes of EPN genes may be related to the functional complexity and/or accelerated evolutions of mammal nervous systems (24,25). Clearly, more studies are needed to confirm this hypothesis. Work in this area is in progress.

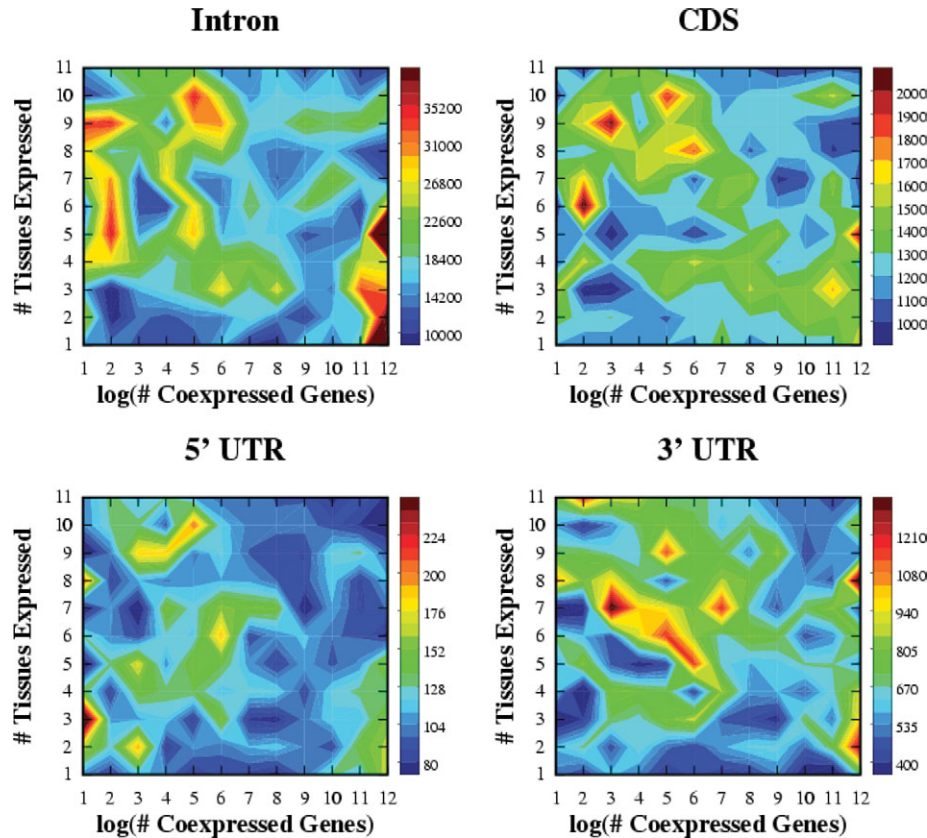
To further address the relation between the size of genes, tissue specificity and expression-pattern specificity, we plot the length distribution of all 7394 mouse RefSeq genes as a function of both tissue specificity and expression-pattern specificity, with the sizes of gene structures color coded from red for long genes to blue for short ones (Fig. 5). If tissue specificity were a determining factor for gene sizes (i.e. more tissue-specific genes are longer), one would expect a rainbow of blue close to the top and red around the bottom across different bins in the number of co-expressed genes. This pattern, however, is not observed. In fact, long genes (red or orange) more frequently appear on two sides (EPS and EPN genes), often regardless of tissue specificity. (It should be noted that because of the logarithmic scale, the first six bins for the number of co-expressed genes are made of 2043 genes with less than 17 co-expressed genes. That is, they all can be classified as EPS genes.) Short genes (blue) are frequently located at the bottom (the most tissue-specific genes) and in the middle (between bins 8 and 10 for the number of genes expressed). The overall trends observed in a single variable (either tissue specificity or expression-pattern specificity, as shown in Figure 2) are somewhat reproduced here. However, the local pattern is far more complex. For example, in the coding regions and 3' UTRs (Fig. 5, top right and bottom right), it appears that long genes are

located mostly in the diagonal region of the plane of the number of expressed tissues versus the logarithm of the number of co-expressed genes. Such a complex pattern suggests that gene sizes may not be simply related to either tissue specificity or expression-pattern specificity alone or both. Different gene regions may be subject to different types of selective pressure. It should be emphasized, however, that this result is subject to the limitation of the data set available. As shown in Supplementary Material, Figure S1, the 7394 genes are distributed in 132 grids (Fig. 5). In each grid, the number of genes ranges from 8 to 248 (129 of 132 grids have at least 10 genes). Thus, it is necessary to further investigate the relation between gene sizes, tissue specificity and expression-pattern specificity when more data are available.

In summary, we find that the size distribution of mammal genes is not uniform in either tissue specificity or expression-pattern specificity. The trend is reasonably clear if a single parameter is used. We find that in addition to house-keeping genes, the most tissue-specific genes are also short. Moreover, on average, genes with the most co-expressed genes and the least co-expressed genes are long. Thus, the use of tissue specificity allows us to locate compact genes, whereas expression-pattern specificity (or non-specificity) detects genes that are long. This suggests that it is necessary to combine tissue restriction and quantitative information of expression data (9) to understand the selective pressures in genome evolution.

## METHODS

We use the compendium of mouse expression profiles for 21 622 presumed distinct and confidently detected mRNA



**Figure 5.** The median lengths (bp) of genes of different regions (intron, top left; coding regions, top right; 5' UTR, bottom left; and 3' UTR, bottom right) as a function of the number of tissues expressed and the logarithm of the number of co-expressed genes. The sizes of genes are color coded from red (long) to blue (short). Long genes (red or orange) more frequently appear at two sides (EPS and EPN genes), sometimes, regardless of tissue specificity. Short genes (blue) are frequently located at the bottom (tissue specific genes) and in the middle (between bin 8 and bin 10 for the number of genes expressed). Bins for the number of tissues expressed and for the number of co-expressed genes are defined in Figure 2.

probes across 55 normal tissues by custom-built DNA oligonucleotide microarrays (9). These probes, characterized by expression intensity above 99% of negative-control spots to rule out experimental noise, provide rigorous information about expression pattern of associated genes. Using the mRNA probe location file provided by Zhang *et al.*, we extract the gene structure from UCSC genome browser (<http://genome.ucsc.edu/>). Only confidently detected probes that represent characterized genes with annotation by RefSeq database (<http://www.ncbi.nlm.nih.gov/RefSeq/>) are used. In case that multiple probes match the same RefSeq gene, only one set is arbitrarily selected to represent its gene expression (7). This leads to a total of 7394 mouse genes with curated RefSeq structure information.

The transcriptional co-expression network is derived from the pairwise correlations between genes' expression profiles (12). In such a network, each vertex represents a distinct gene, and two vertices are linked by an edge if the absolute correlation between their expression profiles is higher than a certain cutoff. Following Zhou *et al.* (26,27), we employ a conservative cutoff of 0.6. The degree  $k$  of a vertex (i.e. gene) is the number of edges linked to its co-expressed genes. The resulted co-expression network has a scale-free behavior (Supplementary Material, Fig. S2) because the

distribution of degrees,  $P(k)$ , decays in a power-law [ $P(k) \sim k^{-\gamma}$ ] (15). A scale-free network indicates that there is a relatively large number of EPS genes with few co-expressed genes and a relatively small number of EPN genes, which are co-expressed with many other genes (12,15).

## SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

## ACKNOWLEDGEMENTS

We wish to thank two anonymous referees for invaluable comments that led to a substantial improvement of the manuscript. We thank Dr Yang Zhang and Yi Xing for their helpful discussions and comments. This work was supported by NIH (R01 GM 966049 and R01 GM 068530), a grant from HHMI to SUNY Buffalo, and by the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Y.Z. was also supported in part by a two-base grant from National Science Foundation of China.

*Conflict of Interest statement.* There is no conflict of interest.

## REFERENCES

- Vinogradov, A.E. (2004). Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? *Curr. Opin. Genet. Dev.*, **14**, 620–626.
- Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. and Kondrashov, F.A. (2002). Selection for short introns in highly expressed genes. *Nat. Genet.*, **31**, 415–418.
- Urrutia, A.O. and Hurst, L.D. (2003). The signature of selection mediated by expression on human genes. *Genome Res.*, **13**, 2260–2264.
- Eisenberg, E. and Levanon, E.Y. (2003). Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.
- Vinogradov, A.E. (2004). Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.*, **20**, 248–253.
- Lehner, B. and Fraser, A.G. (2004). Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet.*, **20**, 468–472.
- Cohen-Gihon, I., Lancet, D. and Yanai, I. (2005). Modular genes with metazoan-specific domains have increased tissue specificity. *Trends Genet.*, **21**, 210–213.
- Farh, K.K.-H., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B. and Bartel, D.P. (2005). The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science*, **310**, 1817–1821.
- Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A., Mitsakakis, N., Mohammad, N., Robinson, M.D., Zirngibl, R., Somogyi, E. *et al.* (2004). The functional landscape of mouse gene expression. *J. Biol.*, **3**, 21.
- Jonathan, B.W. (2004). Co-regulation of mouse genes predicts function. *J. Biol.*, **3**, 19.
- Holmes, C. and Brown, S.D. (2004). All systems go for understanding mouse gene function. *J. Biol.*, **3**, 20.
- Jordan, I.K., Mario-Ramrez, L., Wolf, Y.I. and Koonin, E.V. (2004). Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.*, **21**, 2058–2070.
- Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–390.
- Petti, A.A. and Church, G.M. (2005). A network of transcriptionally coordinated functional modules in *Saccharomyces cerevisiae*. *Genome Res.*, **15**, 1298–1306.
- Barabasi, A.L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
- Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H. and Gerstein, M. (2004). Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.*, **73**, 1051–1087.
- Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
- Yu, H., Greenbaum, D., Lu, H.X., Zhu, X. and Gerstein, M. (2004). Genomic analysis of essentiality within protein networks. *Trends Genet.*, **20**, 227–231.
- Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
- Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. and Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.
- Ge, H., Liu, Z., Church, G.M. and Vida, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.*, **29**, 482–486.
- Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.-D.J., Hao, T., Berriz, G.F., Bertin, N., Huang, J., Chuang, L.-S. *et al.* (2005). Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 801–865.
- Maciag, K., Altschuler, S.J., Slack, M.D., Krogan, N.J., Emili, A., Greenblatt, J.F., Maniatis, T. and Wu, L.F. (2006). Systems-level analyses identify extensive coupling among gene expression machines. *Mol. Syst. Biol.*, msb4100045–E1.
- Dorus, S., Vallender, E.J., Evans, P.D., Anderson, J.R., Gilbert, S.L., Mahowald, M., Wyckoff, G.J., Malcom, C.M. and Lahn, B.T. (2004). Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell*, **119**, 1027–1040.
- Sironi, M., Menozzi, G., Comi, G.P., Cagliani, R., Bresolin, N. and Pozzoli, U. (2005). 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**, 2533–2546.
- Zhou, X.H., Kao, M.C.J. and Wong, W.H. (2002). Transitive functional annotation by shortest path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.
- Zhou, X.J., Kao, M.-C.J., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E. and Wong, W.H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol.*, **23**, 238–243.